City Research Online

City, University of London Institutional Repository

Citation: Koluman, C. (2024). Cognitive Modelling with Burst Learning and Targeted Constrained Search. (Unpublished Doctoral thesis, City, University of London)

This is the accepted version of the paper.

This version of the publication may differ from the final published version.

Permanent repository link: https://openaccess.city.ac.uk/id/eprint/33436/

Link to published version:

Copyright: City Research Online aims to make research outputs of City, University of London available to a wider audience. Copyright and Moral Rights remain with the author(s) and/or copyright holders. URLs from City Research Online may be freely distributed and linked to.

Reuse: Copies of full items can be used for personal research or study, educational, or not-for-profit purposes without prior permission or charge. Provided that the authors, title and full bibliographic details are credited, a hyperlink and/or URL is given for the original metadata page and the content is not changed in any way.

City Research Online: http://openaccess.city.ac.uk/ publications@city.ac.uk

City, University of London

Doctoral Thesis

Cognitive Modelling with Burst Learning and Targeted Constrained Search

Author:
Mustafa Can Koluman

Supervisors:
Dr. Christopher Child
First Supervisor
Dr. Tillman E. Weyde
Second Supervisor

A thesis submitted in fulfilment of the requirements for the degree of Doctor of Philosophy

in the

Research Centre for Adaptive Computer Systems and Machine Learning
Department of Computer Science

July 11, 2024

Declaration of Authorship

I, Mustafa Can KOLUMAN, declare that this thesis titled, "Cognitive Modelling with Burst Learning and Targeted Constrained Search" and the work presented in it are my own. I confirm that:

- This work was done wholly or mainly while in candidature for a research degree at this University.
- Where any part of this thesis has previously been submitted for a degree or any other qualification at this University or any other institution, this has been clearly stated.
- Where I have consulted the published work of others, this is always clearly attributed.
- Where I have quoted from the work of others, the source is always given. With the exception of such quotations, this thesis is entirely my own work.
- I have acknowledged all main sources of help.
- Where the thesis is based on work done by myself jointly with others, I have made clear exactly what was done by others and what I have contributed myself.

Signed: MCK

Date: July 11, 2024

Abstract

This work investigates nonrational iterative learning and searching in a stochastic setting, where the nature of the stochasticity is unknown. Such problems are difficult because at each iteration, the decision making model strives to make the best decision and simultaneously develops its representation of the underlying stochasticity. Outside of a nonrational context, Q-learning or stochastic approximation provide well-known methods for solving such problems subject to restrictions on the speed of learning rate decay and with the use of an infinite time horizon.

The nonrational context proposed here departs from the usual Q-learning approaches by stipulating that the learning rate decays exponentially. Additionally, a search technique named **C**onstrained **S**ingle **U**nconstrained **D**ouble perturbation stochastic approximation (CSUD) is introduced. CSUD comprises a probabilistic hybrid of double- and single-sided simultaneous perturbation stochastic approximation, and is able to constrain not only input updates but also input perturbations. Using performance criteria targeting loss functions and input constraints, a nonrational CSUD search strategy is developed, in the sense of producing not globally unique but only satisficing outcomes.

Normal versus **v**entro**m**edial **p**re**f**rontal **c**ortex (vmPFC) impaired results reported in the **I**owa **G**ambling **T**ask (IGT) are used to calibrate with CSUD search, a series of single-state exponential learning rate decay Q-learning models, culminating in the burst learning model, where the learning rate can be reset via an 'emotion' mediated signal. The key results obtained from the automatic calibration of the Q-learning models consist of: (1) high learning rate decay produces vmPFC impaired behaviour, and (2) for Q-learning models to match human IGT outcomes, exploration must be very high. The presence of high exploration is validated in corresponding human IGT outcomes by introducing an entropy based exploration index (EI). Four different Q-learning architectures including ε-Greedy and Boltzmann exploration are considered, and it is found that no single exploration architecture can alone adequately explain human exploration.

Finally, the performance of nonrational CSUD in tuning a (rational) artificial neural network (ANN) is assessed. For a complex network, nonrational search strategy validation accuracy exceeds random search tuners, but lags behind that of Gaussian-mixture Bayesian tuners.

Published and Submitted Works

- Koluman, C., Child, C., & Weyde, T. (2019). Modelling Emotion Based Reward Valuation with Computational Reinforcement Learning. In A. K. Goel, C. M. Seifert, & C. Freksa (Eds.). Cognitive Science Society
- Koluman, C., Child, C., & Weyde, T. (N.A.). Learning Rate Decay in Q-learning Models Decision Making under VMF Impairment [Unpublished, submitted to Cognitive Computation, submission active since 2020]

"The reason for valuing choice is both conditional and relative. It is conditional because the value of my response as a predictor of my future satisfaction depends on the nature of the question, my capacities of discernment, and the conditions under which my response is elicited. It is relative because it also depends on the reliability of the available alternative means for selecting the outcomes in question."

Scanlon (2000)

"In a world where not all risks are known, statistics and logic are not sufficient - additional tools, such as heuristics, are needed."

Gigerenzer and Gaissmaier (2015)

Contents

Declaration of Authorship **i**

Abstract **ii**

Published and Submitted Works **iv**

Contents **vi**

List of Figures **xii**

List of Tables **xvi**

Acknowledgements **xix**

List of Abbreviations **xx**

List of Symbols **xxi**

1 Introduction **1**

1.1 Criticism, Limitations, and Considerations 8

1.2 Structure 9

1.3 Summary of Contributions 10

2 Literature Review **11**

2.1 Psychology 11

2.2 Semantics: Rational, Nonrational, and Irrational 13

2.3 Emotion and Neuroscience 16

2.3.1 Emotion 16

2.3.2 Neuroscience 18

2.4 Cognitive Iterative Decision Making Models 21

2.4.1 The Expectancy Valence Model 21

2.4.2 The Prospect Valence Model 23

2.4.3 The Outcome-Representation Learning Model 25

2.5 Rational Iterative Learning Models 28
2.5.1 N-Armed Bandits 28
2.5.2 Dynamic and Approximate Dynamic Programming . . 31
2.5.3 Gaussian Mixture Models 33
2.6 Emotion in Iterative Learning 35

3 Learning from Repeated Sampling 38
3.1 General Framework 39
3.2 General Framework Specification 40
3.3 Nonrational Search of Parameter and Hyper-Parameter Spaces 42
3.4 Rational Search of Parameter and Hyper-Parameter Spaces 43
3.5 Switching between Nonrational and Rational Contexts 45

4 The Iowa Gambling Task 48
4.1 The IGT and its Variants 48
4.1.1 The Original IGT 49
Description 49
Aim 49
Implementation Variations 50
Current Implementation 50
4.1.2 The Re-shuffled IGT 51
Description 51
Aim 51
Implementation 52
Current Implementation 52
4.1.3 The Random IGT 52
Description 52
Aim 53
Implementation 53
Current Implementation 53
4.1.4 Additional Investigated IGT Variants 53
Reversed IGT 54
Soochow Gambling Task (SGT) 54
4.1.5 IGT Environments Not Considered 55
4.2 Analysis of IGT Human Behaviour 56

4.2.1 Cumulative Aggregated Measures 56
Computational Model Benchmarks 61
4.2.2 Blocked Itemised Measures 62
4.2.3 The Exploration Index 65

5 **IGT Modelling with Reinforcement Learning and Exponential Learning Rate Decay: Introduction** **68**
5.1 Single state Q-learning with Exponential Learning Rate Decay 69
5.1.1 Exponential Learning Rate Decay 71
5.2 Software agent implementations 71
5.2.1 The Boltzmann Agent 72
5.2.2 The ε-Greedy Agent 73
5.2.3 The Adaptive ε-Greedy Agent 74
5.2.4 The Decaying ε-Greedy Agent 74
5.3 Search strategy implementations 75
5.3.1 Search Criteria Specification 75
5.3.2 Loss Function Specification 77

6 **The Original, Re-Shuffled, and Random IGT Environments with Simple Reinforcement Learning Modelling via CSUD** **80**
6.1 General Methodology 81
6.2 Joint Search of the Original, Re-shuffled, and Random IGT environments 82
6.3 ε-Greedy Agent Results 83
6.3.1 ε-Greedy Agent Discussion 98
6.4 Boltzmann Agent Results 100
6.4.1 Boltzmann Agent Discussion 112
6.5 Adaptive ε-Greedy Agent Results 113
6.5.1 Adaptive ε-Greedy Agent Discussion 123
6.6 Decaying ε-Greedy Agent Results 125
6.6.1 Decaying ε-Greedy Agent Discussion 135
6.7 Summary 135

7 **Reversed IGT with Simple Reinforcement Learning Modelling via CSUD** **140**
7.1 Search of the Reversed IGT environment 141
7.2 Search Results: Reversed IGT 142
7.3 Discussion: Reversed IGT 154

8 The SGT Environment with Simple Reinforcement Learning Modelling via CSUD 157
8.1 Search of the SGT environment 158
8.2 Search Results: SGT 158
8.3 Discussion: SGT 168

9 Reinforcement Learning: The Iowa Gambling Task with Discount Rate and Trace Decay 171
9.1 The Amended Single State Q-learning Model, ARA(κ) 173
9.2 Software agent implementations 174
9.3 Methodology: Joint Search of the Original, Re-shuffled, and Random IGT environments 175
9.4 CSUD Search Results 177
9.5 CSUD Search Discussion 195

10 Reinforcement Learning: Iowa Gambling Task with Burst Learning 199
10.1 Single State Exponential Learning Rate Decay Q-learning with Burst Learning 200
10.1.1 Emotions 200
10.1.2 Learning Rate Decay 201
10.1.3 Behaviour Intervention 202
10.1.4 Two Emotion Single State On-Policy Q-learning 203
10.1.5 vmPFC Impairment 206
10.2 Software agent implementations 207
10.3 Methodology: Joint Search of the Original, Re-shuffled, and Random IGT environments 208
10.4 CSUD Search Results 209
10.5 Discussion: Burst Learning Model 234

11 Future Directions 239

12 Constrained Perturbations Stochastic Search 244
12.1 The SPSA Framework 245
12.1.1 Unconstrained Double-Sided SPSA 246
12.2 Unconstrained Single-Sided SPSA 247
12.2.1 Taylor Expansions 251
12.2.2 The Bias of the Gradient Estimate 253

12.2.3 Almost Sure Convergence of the Parameter Estimate . 256
12.3 CSUD 260
12.3.1 Background 262
12.3.2 CSUD Assumptions 263
12.3.3 Bias of the CSUD Gradient Estimate 265
12.3.4 Almost sure convergence of CSUD 266
12.4 CSUD Search Strategy 270
12.4.1 Search Problem Specification 270
12.4.2 CSUD Search Strategy Assumptions 272
12.4.3 Bias of the CSUD Search Strategy Gradient Estimate . . 275
12.4.4 Almost sure convergence of the CSUD Search Strategy 275
12.5 Discussion 277

13 Application: Neural Network Hyper-parameter Tuning 280
13.1 ANN Tuning 280
13.2 The Training, Validation, and Tuning Pipeline 282
13.2.1 ANN Training 283
13.2.2 ANN Validation and Tuning 284
13.3 CSUD Search Strategy Implementation 285
13.4 Methodology 285
13.4.1 Hyper-Parameter Mapping 288
13.5 Results 289
13.6 Discussion 296

14 Conclusion 299

A IGT Yield Structures 302
A.1 Original IGT 302
A.2 Re-shuffled IGT 303
A.3 Random IGT 303
A.4 Reversed IGT 304
A.5 Soochow Gambling Task (SGT) 305

B Deck Looping Behavior 306

C Learning Rate Decay Grid Construction 309

D CSUD Search Strategy a.s. Convergence: Error Term Bounds 310

Bibliography 314

List of Figures

4.1 5-block 20-draw mean fraction of good decks $\bar{f}_G^H$ for IGT environments with both healthy control and vmPFC impaired participants. 63

4.2 5-block 20-draw mean fraction of good decks $\bar{f}_G^H$ results for IGT environments with only healthy control participants. . . . 64

4.3 5-block 20-draw good versus bad decks exploration index **EI** results for IGT environments with both healthy control and vmPFC impaired participants. 66

4.4 5-block 20-draw good versus bad decks exploration index **EI** results for IGT and SGT environments with healthy control participants. 67

6.1 ε-Greedy agent CSUD iterations. 84

6.2 ε-Greedy agent 2D contours showing learning decay λ and exploration ε effects. 88

6.3 ε-Greedy agent 3D contours with learning decay λ, exploration ε, and initial learning rate α_1. 90

6.4 ε-Greedy Agent 20-draw blocks comparison at CSUD search matches. 91

6.5 ε-Greedy Agent 20-draw exploration index (EI) comparison at CSUD search matches. 93

6.6 ε-Greedy agent comparison of repeated simulation outcomes to human IGT results. 95

6.7 Boltzmann agent CSUD iterations. 101

6.8 Boltzmann agent 2D contours showing learning decay λ and exploration τ effects. 104

6.9 Boltzmann agent 3D contours for learning decay λ, exploration τ, and initial learning rate α_1. 105

6.10 Boltzmann agent 20-draw blocks comparison at CSUD search matches. 106

6.11 Boltzmann agent 20-draw blocks exploration index (EI) comparison at CSUD search matches. 107
6.12 Boltzmann agent comparison of repeated simulation outcomes to human IGT results. 108
6.13 Boltzmann agent exploration temperature τ and action selection probabilities at IGT completion. 111
6.14 Adaptive ε-Greedy agent CSUD iterations. 114
6.15 Adaptive ε-Greedy agent 2D contours showing learning decay λ and exploration τ effects. 116
6.16 Adaptive ε-Greedy agent 3D contours with learning decay λ, exploration τ, and initial learning rate α_1. 117
6.17 Adaptive ε-Greedy agent 20-draw blocks comparison at CSUD search matches. 118
6.18 Adaptive ε-Greedy agent 20-draw blocks exploration index (EI) comparison at CSUD search matches. 119
6.19 Adaptive ε-Greedy agent comparison of repeated simulation outcomes to human IGT results. 120
6.20 Adaptive ε-Greedy agent 2D contours showing exploration temperature τ effects during the IGT. 122
6.21 Decaying ε-Greedy agent CSUD iterations. 125
6.22 Decaying ε-Greedy agent 2D contours showing learning decay λ and exploration decay ν effects. 128
6.23 Decaying ε-Greedy agent 3D contours with learning decay λ, exploration decay ν, and initial exploration ε_1. 129
6.24 Decaying ε-Greedy agent 20-draw blocks comparison at CSUD search matches. 130
6.25 Decaying ε-Greedy agent 20-draw blocks exploration index (EI) comparison at CSUD search matches. 131
6.26 Decaying ε-Greedy agent comparison of repeated simulation outcomes to human IGT results. 132
6.27 Decaying ε-Greedy agent 2D contours showing exploration decay ν effects during the IGT. 134

7.1 Reversed IGT CSUD searches. 144
7.2 Reversed IGT CSUD verification grid search 2D contours. . . . 147
7.3 Reversed IGT CSUD verification grid search 3D contours. . . . 149
7.4 Reversed IGT agent 20-draw blocks comparison. 150

7.5 Reversed IGT agent 20-draw blocks exploration index (EI) comparison. 151
7.6 Reversed IGT. Comparison of repeated agent simulation outcomes to human IGT results. 152

8.1 SGT CSUD searches. 160
8.2 SGT CSUD verification grid search 2D contours. 163
8.3 SGT CSUD verification grid search 3D contours. 164
8.4 SGT agent 20-draw blocks comparison. 165
8.5 SGT agent 20-draw blocks exploration index (EI) comparison. 166
8.6 SGT. Comparison of repeated agent simulation outcomes to human IGT results. 167

9.1 ARA(κ) joint original, re-shuffled and random IGT CSUD search matches. 179
9.2 ε-Greedy Agent ARA(κ) CSUD verification grid search 2D contours. 181
9.3 Boltzmann Agent ARA(κ) CSUD verification grid search 2D contours. 182
9.4 ARA(κ) ε-Greedy agent CSUD search grid verification 3D contours. 184
9.5 ARA(κ) Boltzmann agent CSUD search grid verification 3D contours. 185
9.6 ARA(κ) agent 20-draw blocks comparison. 187
9.7 ARA(κ) agent 20-draw block exploration index (EI) comparison. 188
9.8 ε-Greedy ARA(κ) agent. Comparison of repeated simulation outcomes to human IGT results. 189
9.9 ARA(κ) Boltzmann agent. Comparison of repeated simulation outcomes to human IGT results. 190
9.10 ARA(κ) Boltzmann agent exploration temperature τ and action selection probabilities at IGT task completion. 194
9.11 Random IGT outcomes for control subjects. 197

10.1 2EmST agent joint original, re-shuffled and random IGT CSUD search matches. 212
10.2 2EmST ε-Greedy Agent CSUD verification grid search 2D contours. 215

10.3 2EmST Boltzmann agent CSUD verification grid search 2D contours. 216
10.4 2EmST ε-Greedy agent CSUD verification grid search 3D contours. 219
10.5 2EmST Boltzmann agent CSUD verification grid search 3D contours. 220
10.6 2EmST ε-Greedy agent CSUD search grid verification 3D contours for $\alpha_1 = 0.65$. 222
10.7 2EmST agent 20-draw blocks comparison. 223
10.8 2EmST agent 20-draw block exploration index (EI) comparison. 224
10.9 Exploration index (EI) for mean fraction of good decks $\bar{f}_G$. . . 225
10.10 2EmST ε-Greedy agent Comparison of repeated simulation outcomes to human IGT results. 226
10.11 2EmST Boltzmann agent. Comparison of repeated simulation outcomes to human IGT results. 227
10.12 2EmST ε-Greedy agent per period mean learning rate $\bar{\alpha}_t$ progression. 232
10.13 2EmST Boltzmann agent per period mean learning rate $\bar{\alpha}_t$ progression. 233

12.1 CSUD behaviour with a single input loss function $L(\theta)$. 262

13.1 Fashion MNIST convolutional network architecture. 286
13.2 Model A tuning passes and validation accuracy. 290
13.3 Fashion MNIST model A pairwise comparison matrix of mean validation accuracy with p-values using the DSCF criterion. . 291
13.4 Model B tuning passes and validation accuracy. 293
13.5 Fashion MNIST model B pairwise comparison matrix of mean validation accuracy with p-values using the DSCF criterion. . 295

B.1 Normal agents deck looping versus exploration at mean fraction of cards per deck at 100th draw. 306
B.2 vmPFC Impaired agents deck looping versus exploration at mean fraction of cards per deck at 100th draw. 307
B.3 Steingroever et al. (2018), healthy subjects, mean fraction of cards per deck at 100th draw. 308

List of Tables

4.1 The original IGT decks. Theoretical infinite horizon mean net yields per draw and deck. 51

4.2 The SGT decks. Theoretical infinite horizon mean net yields per draw and deck. 55

4.3 Original IGT environment. 100th draw cumulative good deck means. 58

4.4 Re-shuffled IGT environment. 100th draw cumulative good deck means. 59

4.5 Random IGT environment. 100th draw cumulative good deck means. 59

4.6 Reversed IGT environment. 100th draw cumulative good deck means. 60

4.7 Soochow (SGT) environment. 100th draw cumulative good deck means. 60

4.8 IGT and SGT mean fraction of good deck $\bar{f}_G^H$ ranges used for comparing agent and literature results. 61

5.1 Tunable hyper-parameters by software agent. 76

6.1 Search Methodology. Joint original, re-shuffled, random IGT hyper-parameter CSUD search criteria. 82

6.2 ε-Greedy agent CSUD search matches after 1000 iterations. . . 85

6.3 ε-Greedy agent CSUD verification. Hyper-parameter grid search criteria for joint original, re-shuffled, and random IGT. 87

6.4 ε-Greedy agent np-M/ANOVA analysis of mean fraction of good decks $\bar{f}_G$ with learning decay λ as factor. 97

6.5 Boltzmann agent CSUD minimum loss search matches after 1000 iterations. 102

6.6 Boltzmann agent CSUD verification. Hyper-parameter grid search criteria for joint original, re-shuffled, random IGT. . . . 103

6.7 Boltzmann agent np-M/ANOVA analysis of mean fraction of good decks $\bar{f}_G$ with learning decay λ as factor. 110
6.8 Adaptive ε-Greedy agent CSUD minimum loss search matches after 1000 iterations. 115
6.9 Adaptive ε-Greedy agent CSUD verification. Hyper-parameter grid search criteria for joint original, re-shuffled, and random IGT. 115
6.10 Adaptive ε-Greedy agent np-M/ANOVA analysis of mean fraction of good decks $\bar{f}_G$ with learning decay λ as factor. . . . 121
6.11 Decaying ε-Greedy agent CSUD minimum loss search matches after 1000 iterations. 126
6.12 Decaying ε-Greedy agent CSUD verification. Hyper-parameter grid search criteria for joint original, re-shuffled, and random IGT. 127
6.13 Decaying ε-Greedy agent np-M/ANOVA analysis of mean fraction of good decks $\bar{f}_G$ with learning decay λ as factor. . . . 133

7.1 Search Methodology: Joint reversed IGT hyper-parameter CSUD search criteria by agent. 142
7.2 Reversed IGT CSUD search matches after 1000 iterations. . . . 145
7.3 CSUD Verification. Reversed IGT hyper-parameter grid search criteria by software agent. 146
7.4 Agent Reversed vs. Learning Decay λ np-M/ANOVA analysis of mean fraction of good decks $\bar{f}_G$ with learning decay λ as factor. 153

8.1 Search Methodology: SGT hyper-parameter CSUD search criteria by agent. 158
8.2 SGT minimum loss CSUD search matches after 1000 iterations. 161
8.3 CSUD Verification. SGT hyper-parameter grid search criteria by agent. 162

9.1 ARA(κ) CSUD Search Criteria. 176
9.2 ARA(κ) CSUD search matches. 178
9.3 ARA(κ) CSUD verification grid search configurations. 180
9.4 ARA(κ) ε-Greedy agent np-M/ANOVA analysis of mean fraction of good decks $\bar{f}_G$ with learning decay λ as factor. 192

9.5 ARA(κ) Boltzmann agent np-M/ANOVA analysis of mean fraction of good decks $\bar{f}_G$ with learning decay λ as factor. . . . 193

10.1 Two emotion learning and response, abstracted from Rolls (2013) 201
10.2 Agent variants, activation thresholds, associated emotions and learning rate index i behaviour. 205
10.3 2EmST agent CSUD search criteria. 208
10.4 2EmST agent joint original, re-shuffled and random IGT CSUD search matches. 210
10.5 2EmST agent joint original, re-shuffled and random IGT CSUD search grid verification configurations. 213
10.6 2EmST ε-Greedy agent np-M/ANOVA analysis of mean fraction of good decks $\bar{f}_G$ with behaviour as factor. 229
10.7 2EmST Boltzmann agent np-M/ANOVA analysis of mean fraction of good decks $\bar{f}_G$ with behaviour as factor. 231

12.1 Gradient bias by gradient approximation method 255

13.1 Could SPSA be used for ANN training? 284
13.2 Tuning and training algorithm configuration settings. 287
13.3 Evaluation stage configuration. 288

Acknowledgements

I am deeply indebted to my supervisors for their invaluable insights and direction. I would also like to thank my wife for her support, patience, and for putting up with the long hours required; and my mother for her interest and steadfast encouragement.

List of Abbreviations

2EmST	**2** **E**motion **S**toic **T**hreshold
a.s.	**a**lmost **s**urely
ADAM	**Ad**aptive **M**oment **E**stimation
ADP	**A**pproximate **D**ynamic **P**rogramming
ANN	**A**rtificial **N**eural **N**etwork
ARA	**A**ction **R**esponse **A**ction
cp-SPSA	**c**onstrained **p**erturbations **S**imultaneous **P**erturbations **S**tochastic **A**pproximation
CSUD	**C**onstrained **S**ingle **U**nconstrained **D**ouble cp-SPSA
DP	**D**ynamic **P**rogramming
EI	**E**xploration **I**ndex
EV	**E**xpectancy **V**alence Model
IGT	**I**owa **G**ambling **T**ask
i.o.	**i**nfinitely **o**ften
NFL	**N**o **F**ree **L**unch Theorems
OFC	**O**rbitofrontal **C**ortex
ORL	**O**utcome-**R**epresentation **L**earning Model
PV	**P**rospect **V**alence Model
RL	**R**einforcement **L**earning
SE	**S**tandard **E**rror
SGT	**S**oochow **G**ambling **T**ask
SPSA	**S**imultaneous **P**erturbations **S**tochastic **A**pproximation
TTB	**T**ime-**T**o-**B**ound
vmPFC	**V**entro**m**edial **P**refrontal **C**ortex

List of Symbols

α	statistical significance level, learning rate (context dependent)
α_1, α_t	initial learning rate, learning rate at time t
δ_t	temporal difference error at t
Δ	difference operator
Δ_t	random perturbation vector at t
ϵ	measurement error
ε	exploration, measurement error (context dependant)
ε_1, ε_t	initial exploration, exploration at t
Θ, $\hat{\Theta}$	hyper-parameter input, hyper-parameter input estimate
γ	discount rate
κ	trace decay
λ, λ_N, λ_{vmPFC}	learning (rate) decay, normal, vmPFC impaired
μ	mean
μ_t	perturbation step size at iteration t
ν	exploration decay
Ψ	parameter input
σ^2	variance
τ	exploration (temperature)
a	action
$1/B$	emotion activation threshold
D	attenuation
f_G	fraction of good decks
$\bar{f}_G$, $\bar{f}_G^H$	mean fraction of good decks for software agents, for humans
$Q_t(a)$	Q-value at time t from action a
t	time or iteration index
x_t^a	net yield x at time t from action a
$\{\mathscr{M}_t\}$	a martingale process
$\mathbb{N}_0$	the set of natural numbers including 0
$\mathbb{R}$	the set of real numbers
$\mathbb{Z}$	the set of integers
$\mathbb{Z}_k$	the (k^{th}) input constraint set
$g(\cdot)$, $\hat{g}(\cdot)$	gradient generating function, its approximation
$L(\cdot)$	unobserved loss
$M(\cdot)$	reductive performance statistic generating function
$R(\cdot)$	objective function in parameters and hyper-parameters
$Y(\cdot)$	observed stochastic loss

Chapter 1

Introduction

This work proposes two new approaches in computational technology; these new approaches are inspired by the role of emotion in decision making and by heuristic decision making theory. On a mathematical and algorithmic level, single-state model-free reinforcement learning (RL) Q-learning learning variants with exponentially decaying learning rate, and a constrained perturbations, gradient driven automatic model tuner are proposed. Methodologically, human Iowa Gambling Task (IGT) (Bechara et al., 1994) outcomes are used to automatically calibrate a series of Q-learning models, where this calibration is conducted using **C**onstrained **S**ingle **U**nconstrained **D**ouble stochastic approximation (CSUD) as the model tuner; the stochastic approximation technique introduced in chapter 12.

The fundamental concepts are introduced in chapter 3. Chapters 5 to 10 present the calibration and discussion of the single-state Q-learning variants. It is shown that high exponential learning rate decay reproduces the performance of **ventromedial prefrontal cortex** (**vmPFC**) impaired human subjects in the Iowa Gambling Task (IGT) (Bechara et al., 1994). Interestingly, it is also shown that for Q-learning models to match the IGT outcomes of normal and vmPFC impaired human subjects, exploration must be very high. The high exploration finding is discussed in the context of the No Free Lunch theorems, which state that in the space of all problems and all algorithms, no algorithm is better than random search (Wolpert & Macready, 1997). It is proposed that in the presence of incomplete information and uncertainty, algorithmic exploration can be interpreted in a heuristic context as a defence against algorithmic specificity.

The IGT and its variants employed in this work are discussed in detail in chapter 4. Bechara et al., 1994 introduces the IGT as a clinical psychological test capable of identifying ventromedial prefrontal cortex (vmPFC) impairment, which leads to forward looking planning deficits (Goel et al., 1997). The original IGT, proposed by Bechara et al., 1994, consists of a four deck card game played with virtual money. The length of the task consists of 100 card draws from any deck, but this length is unknown to the participant. Each card draw yields a fixed reward with a random punishment (fine), but the net yield, that is $reward - fine$ may be positive or negative. Unknown to the participants, two good decks produce on average positive net yields, while the remaining bad decks produce on average negative net yields. The participants must determine which decks are the good decks; however initially, the bad decks look good and this further complicates discovery of the good decks. Variants of the IGT discussed here retain the 4 deck structure but differ in the incidence and frequency of the gains and losses. Appendix A presents the IGT yield structures used in this work.

Chapter 12 presents in detail the constrained perturbations, gradient driven automatic model tuner, which is named **c**onstrained **s**ingle **u**nconstrained **d**ouble perturbation stochastic approximation (CSUD), and comprises a probabilistic hybrid of double-sided (Spall, 1992) and single-sided (Chen et al., 1999) simultaneous perturbation stochastic approximation (SPSA). Chapter 13 then presents a short comparison of CSUD versus well-established industrial-scale artificial neural network (ANN) tuners such as asynchronous successive halving algorithm (ASHA) based approaches (L. Li et al., 2020).

This work possesses interdisciplinary character and draws from cognitive science, psychology, decision making, optimisation theory, and computer science, in particular from generative machine learning, where the term generative is used to indicate computational technologies with generalised models, which are then trained and tuned via interaction with data to produce outcome behaviours with desirable qualities. For example in the context of Q-learning or reinforcement learning (RL) models, such desirable behaviour may comprise an optimised choice selection policy (Tsitsiklis, 1993). While this work has interest in human behaviour, the focus here is to introduce computational approaches inspired by human behaviour. Of course, it is of interest to ask if the resulting algorithms would then in turn receive support from human behaviour? That is, could human behaviour indeed result from such algorithms being employed at some level by the organism itself?

In recent years both in decision making theory and psychology, there has been interest to find support for the type of optimisation and trade-offs, which occur in computational reinforcement learning (RL) (O'Doherty et al., 2006; Olschewski et al., 2024). Such efforts investigate support for RL models by employing fitted models, where a model and its settings are estimated by fitting reinforcement learning model parameters using maximum likelihood (Wilson & Collins, 2019).

The journey from observed outcomes to (fitted) model parameters is much more perilous, involves long inference chains, for example, such as in Bayesian hierarchical estimation (Piray et al., 2019), and consequently requires much stronger statistical assumptions, typically consisting of parametric distributions. Further, given the RL exploitation-exploration trade-off, when fitting outcome data to RL model parameters, it is sometimes difficult to derive sensible exploration and learning rate combinations, with a-priori constraints on learning rate and exploration being applied (Daw, 2011). In constrast with generative modelling, it is easier to assess outcomes because such outcomes consist of behavioural choice metrics. Hence, generative models can be compared on the basis of simulated, or real-life, performance. In this work, models are assessed on the basis of simulation outcomes.

In fitted modelling, however, comparing model performance is not straightforward. Due to limitations of sample size and subject patience, model performance can only be compared in the context of low volume small sample realizations. Model comparison assessment measures inevitably compare the same outcome data across multiple models, and this raises the risk of over-fitting, leading to multiple mitigation techniques such as the Bayes Information Criterion (BIC), the confusion matrix, or others discussed in Wilson and Collins, 2019. The current work proposes an alternative generative workflow, where the researcher models hypotheses in terms of a loss function (CSUD), assessed via outcome realizations. Such a workflow makes it easier to search through relevant RL model setting combinations subject to the prior information introduced by any loss function. Section 5.3.2 presents an example of how to specify such a loss function. By using such loss function constructs, rapid prototyping of the model (algorithm) space may be achieved, and the resulting information may be used to inform on any physical experiment design.

Statistically CSUD, as a probabilistic hybrid of double- and single-sided SPSA, has the properties of a robust algorithm, requiring finiteness of the

1^{st} and 2^{nd} loss function error moments; the loss function error, in fact, may possess a non-zero mean. CSUD searches through the model space by using random perturbations. Unlike in SPSA, however, these perturbations may be constrained without affecting the statistical robustness of the method. Consequently, CSUD offers the possibility of clean model space decomposition.

Another difficulty, which is unavoidable in RL is the selection of the exploration method. In psychology and the decision making literature, the exploration methodology of choice is the Boltzmann, or soft-max, rule, where exploration is conducted in proportion to some value metric such that advantageous choices are chosen more frequently (Erev et al., 2010). Boltzmann exploration, however, by normalising choice selection, produces unimodal choice selection distributions, and such an a-priori imposition may not always be justifiable. Indeed, while it is not undertaken here, one must question when a Gaussian likelihood function is implemented in the presence of a Boltzmann rule, as to what extent model fitting results reflect a-priori imposed regularised outcomes. When model fit errors look nice, the researcher may assume that the model is a good fit, but it is possible that the obtained fit had been already imposed a-priori. In this work given the key criterion, as noted for example in section 6.4 Table 6.7, where normal and vmPFC impaired patients should produce outcomes where the null hypothesis of an equal outcome fails to be rejected; this expected result does not obtain under Boltzmann exploration. This results raises the question as to whether and to what extent the Boltzmann rule can actually model human exploration. Here, four competing exploration methods are assessed: ε-Greedy, Boltzmann, adaptive ε-Greedy, and decaying ε-Greedy exploration. Based on generative simulations, no single exploration model emerges as a clear description of human exploration in the IGT.

From an RL algorithmic perspective, for exploration to work towards the discovery of an optimal policy, the sampled outcome process must be ergodic; that is given an infinite time horizon, it must be possible to sample all outcomes eventually. Psychological tests of human exploration schemes tend to be ergodic by design as such experiments typically involve two choices one of which must be sampled (Erev et al., 2010). Further, human exploration is decomposed into directed exploration and fully random exploration (Wilson et al., 2014). Elsewhere it is suggested that human exploration may be inherently randomised due to transcoding errors during learning (Findling et al., 2019). Such results on human exploration do indeed suggest that the simple exploration schemes, including the popular Boltzmann rule, employed here may not adequately model human exploration. In the burst learning model in Chapter 10, however, it is demonstrated for example in Fig. 10.10 and Table 10.6 that the burst learning ε-Greedy exploration scheme does produce simulation results, which (1) satisfy the finding that high exponential learning rate decay produces vmPFC impairment, and (2) support the normal and vmPFC impaired human results on the original and re-shuffled IGT variants (Bechara et al., 1994; Fellows & Farah, 2005). Hence, even if none of the simple exploration schemes used here fully capture human exploration, with the burst learning model, it appears that ε-Greedy exploration may produce a simple catch-all representation for exploration with systematic and random components.

Comparing models with differing exploration strategies remains challenging. Often the immediate effect of an exploration setting is not immediately obvious. For example, in the case of the Boltzmann rule, it is not easy to see how varying the temperature would influence actual outcomes. In section 4.2.3, this work introduces an entropy based measure called the Exploration Index (EI). This exploration index yields a value of 100 for random search, and 0 for fully deterministic search. EI comprises, however, a measure of implied exploration, in the sense that exploration is estimated based on the subject's choices over a specific time interval. For example, if a subject systematically chose each one of available 5 options over a period of 10 trials, then the EI would produce a score of 100. In other words, fully systematic exploration is also capable of producing a score of 100, which is obtained from random search. In sum, while the EI cannot distinguish between a-priori directed and random exploration, it is helpful in terms of scoring the amount of exploration in terms of choice variability in the actual observed outcomes.

Related to the question of exploration are the issues of decisions from risk versus decisions from experience. It has been observed that human beings appear more risk averse towards rare adverse outcomes from description than towards rare adverse outcomes from experience (uncertainty) (Hertwig et al., 2004). In this work, such a distinction has not been coded into RL model behaviour. In all Q-learning simulations, software agents start with 0 initialised Q-values, and must then build up a value representation of outcomes. For example, it might have been possible to code adverse outcome risk as negative initial Q-values; this was not investigated here, but forms an interesting investigative possibility.

The present study also differs from psychological experiments on risk and uncertainty in one other key aspect, and that is the use of the IGT. In Hertwig et al., 2004 for example, subject choices are limited to two options, the discovery of which are carefully controlled to produce laboratory test conditions. The IGT presents a substantially different environment: the subject has four choices, the subject is essentially being tricked and must see through this; finally the subject does not know when the task ends (Bechara et al., 1994). In this work, from a probabilistic perspective, a choice task with only two options was deemed to be too simple an environment: since the probabilities are normalised, working out the likelihood of one choice is sufficient to solve the a choice problem. An environment with multiple choices was desired to develop a better understanding of probabilistic choice. Further, the IGT provides a task substantially closer to living experiences, where effective learning involves disambiguation of multiple influences; such a task was deemed interesting for training and testing the proposed Q-learning and automatic tuning algorithms. Finally, high exponential learning rate decay can act as a frequency filter, attenuating the value feedback of any choice after a certain number of iterations (section 10.1.2), and this could therefore in an RL context generate vmPFC type behaviour. Based on this supposition, this work proposes as a heuristic model an exponential learning rate decay Q-learning architecture, and the IGT with its vmPFC subject outcomes provides a suitable calibration platform. Such calibration was performed automatically via CSUD, and then verified via small grid-searches.

In the discussion of any algorithm, it remains a challenge to develop a verbal vocabulary to adequately describe the rationale and operation of the algorithm. The development of such a verbal vocabulary inevitably involves compromises. The vocabulary in this work is no exception. With this caveat in mind, some important verbal concepts are introduced next.

Here the term **rational** describes probabilistically astute and infinitely lived *oracle* based models, which devise structures to resolve imperfections resulting from uncertainty and incomplete information. Such optimisers are commonly used in engineering, however, where they are not referred to as being rational. Gigerenzer and Gaissmaier, 2015 have suggested the term

nonrational or **heuristic** to refer to models, which use simple local approximations to deal with limitations of information, time, and lack of certainty. For example in bounded rationality (Lorkowski & Kreinovich, 2018), the actor foregoes optimization, and instead *satisfices*, a term which describes a choice strategy of choosing the first alternative fulfilling selection criteria (Simon, 1956). Nonrational or heuristic decision making models may use domain specific selection shortcuts (Gigerenzer & Gaissmaier, 2015). For example, in order to catch a fly ball, a baseball player adjusts his running speed towards the ball while looking at the ball at a constant angle.

The term **iterative learning optimisers** refers to tools capable of learning and optimising over time. Problems requiring such tools have been typically formulated in engineering control theory or in economic social planning. In the form of artificial neural networks and reinforcement learning, such tools are also fast becoming the mainstay of decision making automation in machine learning (Kochenderfer, 2015).

Here iterative learners are said to optimise directly a set of behaviour input variables, called **parameters**, subject to a set of performance tuning variables called **hyper-parameters**. The parameters and hyper-parameters are estimated in two separate stages per learning iteration. Further in each such iteration, iterative learners also approximate the objective functions, which form the basis of parameter and hyper-parameter selection.

In this work, the term **nonrational** describes a class of iterative decision making models, which comprise a subset of the bounded rationality and heuristic decision making ecosystem. The nonrational models proposed relax the typical assumptions of rational optimisation in two specific ways: (1) by inducing a sub-infinite time horizon, and (2) by using targeted exploratory search.

A sub-infinite time horizon is induced via exponential learning rate decay, which in and of itself does not comprise a new concept in the optimisation literature, but the use of which cannot lead to theoretical statistical convergence guarantees obtained for example in Tsitsiklis, 1993. CSUD is used to produce targeted exploratory search; however, perturbation constraints and custom loss function targets (which may be unattainable), lead to a gradual modification of statistical SPSA convergence guarantees from global uniqueness to a local best result. This gradual transformation from a rational to a nonrational satisficing context is discussed in propositions 12.2.1, 12.3.1 and 12.4.1 in chapter 12.

1.1 Criticism, Limitations, and Considerations

This work attempts to port key rational decision making concepts into a nonrational context. One might ask, "Why attempt this when rational models provide well-established solutions?"

In a rational context, it is only in the limit and subject to regularity assumptions that iterative learners can be guaranteed to optimally converge (Hall et al., 2014; Yin & Kushner, 2003). Among other conditions, a finite mean and variance are required. By construction, the central tendency and deviation are generally well-behaved, however, this good behaviour may obscure any poor, or extraordinary results, which may obtain in the relevant probability distribution's tail. Further, from an individual's or a software agent's point of view, when faced with a limited number of choice trials with entry, execution, and exit costs and benefits, the concept of probability of success is difficult to enumerate. Given such difficulties in population as well as individual experiential sampling, it might make more sense to forego rational optimisation, and instead conduct flexible iterative searches, which can be varied over decision iterations according to some customised rules. Such approaches are discussed in Volz and Hertwig (2016) for single-period problems. The proposed CSUD search technique can be seen as a multi-period version of such searches.

Finally, in the (generative) IGT simulations in chapters 6 to 10, a representative agent architecture is used. Further, to calibrate representative agent behaviour, normal and vmPFC impaired mean human IGT outcomes are employed. In contrast in psychology, reinforcement learning (RL) model fitting aims to fit for each individual participant the most plausible model settings for that particular individual's observed data (Daw, 2011). The use of a representative agent may be critiqued not only as being a rational hold-over, but also as being an unrealistic individual modelling approach. Both critiques raise valid points. When a representative agent is used with quadratic loss for example, the abstract form (3.1) can be implemented as recursive least squares, where the learning rate at iteration t is $1/t$ (Powell, 2011, pp. 422-423). Here, however, as proposed in section 3.5, the usage of exponential learning rate decay can be used to switch into a nonrational context. While individual modelling is an interesting approach, the aim here is to investigate RL model behaviour, and for that purpose using a representative agent provides the best approach for analysing model and IGT interaction, as any differences in individual simulation outcomes are solely driven by IGT (data) stochasticity.

1.2 Structure

Chapter 2 presents the literature review. Chapter 3 introduces the mathematical treatment of nonrational learning from repeated sampling, and presents the general learning framework, which separates the iterative learning task into choice selection via behavioural parameters, and tuning via performance hyper-parameters.

This is followed in chapter 4 by a discussion of the Iowa Gambling Task (IGT). Chapter 4 quantifies the human behavioural outcomes, which are used to calibrate software agents.

Using exponential learning rate decay, chapter 5 develops a nonrational single-state Q-learning model with the initial learning rate, learning rate decay, and exploration as tunable hyper-parameters. Chapter 6 to chapter 8 present IGT applications, where software agents are calibrated using the simple model introduced in chapter 5.

Chapter 9 presents a discount rate and trace decay augmented nonrational RL model with IGT applications. Subsequently chapter 10 presents the nonrational burst learning RL model, where an emotion signal may reset the exponentially decaying learning rate.

Chapter 11 presents future directions suggested by the application of the CSUD search tool to IGT environments.

Next, chapter 12 introduces the theoretical foundation for CSUD and the CSUD search strategy. This chapter also compares and contrasts the rational and nonrational approaches to iterative decision making. This technical CSUD chapter is followed by a further application using CSUD search to tune a convolutional Fashion MNIST (Xiao et al., 2017) **artificial neural network** (ANN). Finally chapter 14 summarises and presents conclusions.

1.3 Summary of Contributions

- Nonrational modelling is achieved via exponential learning rate decay, loss function driven searches, and the proposed CSUD algorithm, which probabilistically hybridises double- and single-sided SPSA, while allowing for update and perturbation constraints.

- Using single-state Q-learning models and IGT variants, it is shown that vmPFC impairment can be modelled by high exponential leaning rate decay. Further it is found that for simulated Q-learning model results to match corresponding human outcomes, exploration must be high.

- The entropy based Exploration Index (EI) is introduced for comparing exploration resulting from different model simulation, and human IGT outcomes.

- It is shown that Boltzmann exploration cannot account for original and re-shuffled variant IGT human outcomes, whereas ε-Greedy exploration can do so.

- In a nonrational context, it is shown that CSUD search can produce results consistent with the "satisficing" criterion (Simon, 1956).

Chapter 2

Literature Review

The literature review is divided into sections, each covering a separate component of this work's research context.

2.1 Psychology

As discussed in the introduction, the main contextual considerations relating to the psychology literature in decision making consist of risk versus experience (uncertainty) (Erev et al., 2010; Hertwig et al., 2004), the role of exploration (Findling et al., 2019; O'Doherty et al., 2006; Wilson et al., 2014), the individual fitting and plausibility testing of RL model parameters (Daw, 2011; Piray et al., 2019); and appropriate experiment design (Wilson & Collins, 2019).

The major contextual points have been already covered in the introduction. However, the concept of "rare events," experiment design, and RL specific considerations need to be discussed further. Hertwig et al., 2004 define a rare event as one that has probability at or below 0.2, and additionally study rare events with probabilities of 0.1 and 0.025. The authors find that a-priori description of risk structure leads to overweighting, while experiential discovery leads to under-weighting of the probability of the rare event. In the IGT, which is used in the RL simulations here, subject instructions do not include a description of the risk structure; subjects are only told that some card decks produce net yields worse than others (Bechara et al., 2000). In this sense, the IGT constitutes an experiential task, where the participant faces uncertainty in terms of deck net yield means and probabilistic yield structures. In the Soochow Gambling Task (SGT) (Lin et al., 2009), where the rare event gain or loss frequency is 0.2, human participants generally perform worse than random search; this result supports the findings in Hertwig et al. Further in the SGT, as discussed in chapter 8, the RL models are challenged as well, displaying during grid search verification, at varying exploration settings, intersecting outcome valuation contours. Interestingly, these intersections appear to occur at or near the vicinity of corresponding human IGT outcomes, raising the possibility that the heuristic humans are using has evolved well to extract rare event information.

Regarding decision making experiment design, Olschewski et al., 2024 highlight the current gap between decisions from risk versus decisions from experience experiments under controlled conditions, and realistic decision making problems. In general in the laboratory environment, a choice between two options must be made, and the subject faces clearly defined signals, even if some signals occur rarely. The authors call for a need to develop more realistic test scenarios. The IGT and its variants, albeit unwittingly, fulfil such a need. In the IGT, the subject must choose from four card decks, and each choice may yield a gain and a loss. Further, the original IGT attempts to trick the subject into believing that the on average negative net yield bad decks are actually good (Bechara et al., 1994). The original, re-shuffled, and random IGT variants exhibit identical steady state net yields, but reveal differing human behaviour between normal and vmPFC impaired subjects. This discrepancy suggests each card deck exhibits card sequencing effects. This in turn suggests that IGT participants may develop card deck value representations based on only a small number of samples. Here it is proposed that in RL modelling such card sequencing effects may be achieved by an exponentially decaying learning rate.

As is the case here, in the psychology decision making literature, computational RL models are generally implemented as single-state Q-learning constructs (Daw, 2011). In such models, the key parameters consist of the learning rate and exploration. The learning rate determines the contribution of the current outcome to the overall value representation, whereas exploration insures continued sampling of all choices. Unlike the approach discussed in Daw, 2011, where the learning rate is constant, here the learning rate exhibits exponential decay as defined in (3.8). While it is difficult to estimate from individual outcome data a variable learning rate using maximum likelihood, the RL models proposed here pair an initial learning rate with learning rate decay, and therefore produce only one additional parameter to be estimated. A test of the simulation results obtained here via maximum likelihood model fitting to human outcomes constitutes an interesting future prospect. However, the focus in this work is on generative modelling with models being assessed via simulation outcomes.

2.2 Semantics: Rational, Nonrational, and Irrational

Precise understanding of the words *rational*, *nonrational*, and *irrational* is required. This precision is needed since the word rational can have different meanings in different contexts. The Oxford Dictionary defines rational as "based on or in accordance with reason or logic, able to think sensibly or logically."[1] This is a good definition for everyday usage, but does not give voice to the *disillusionment* inherent in post-modern thinking regarding reason and rationality. For example, Alexander (2013) coins the phrase "the dream of reason,"[2] and describes it as "the image of rationally perfected life in thought, but of course not a reflection of 'real,' material life alone."[3]

According to Alexander (2013), the dream of reason is driven by "**logical positivism**", which believes that "any thought worth thinking could be reduced to rational and eventually mathematical propositions."[4] It is in this sense that Von Neumann and Morgenstern (1944) used *rational* to refer to a person who acts to optimise a utility function, and who is capable of dealing with risk, where the term risk is used to indicate that choice probabilities are known or can be reliably estimated. Von Neumann and Morgenstern's approach laid the foundation in economics for later, more complex multi-period models with increasing modelling sophistication, using for example, stochastic difference or differential equations (Lucas & Sargent, 1981), dynamic programming (Samphantharak & Townsend, 2013), or robust control (Hansen & Sargent, 2008). Of the three mentioned techniques, robust control allows for divergence from an underlying true model, and formulates a mathematical notion for "good enough." Such a notion of "good enough" also provides the rationale for the CSUD search strategy.

Robust control makes allowance for the main philosophical criticism of rational decision making, namely that it has become a proscriptive ideology. Indeed post-modern defenders of rationality de-emphasize the proscriptive,

[1] *rational*. Concise Oxford English Dictionary, 2008.

[2] Alexander, 2013, p. 10.

[3] ibid., p. 10.

[4] ibid., p. 10.

and instead highlight the descriptive aspect of the concept. In this view, rationality is normative and describes what should be or could be. The purpose of rationality is to approximate reality as closely as is possible (Wedgwood, 2017). In a decision making context, it does not matter that one uses a fantasy or made-up model, which might be underpinned by unrealistic or unverifiable assumptions. What matters is whether the model can be successful in its respective problem domain. When rationality is viewed from this perspective, similarities between rational and heuristic approaches become evident: both approaches make approximations, and both approaches focus on specialised problem domains.

Gigerenzer and Gaissmaier (2011) define a heuristic as "a strategy that ignores part of the information, with the goal of making decisions more quickly, frugally, and/or accurately than more complex methods." So by definition, a heuristic strives for some form of simplicity. For lack of a better term, heuristic models are also referred to as *nonrational* (Gigerenzer & Gaissmaier, 2015), or have also been described under the term bounded rationality (Simon, 1956).

How does one classify a model as being rational or nonrational? This difficult question has motivated Gigerenzer (2016) to present the case for a "rational theory of heuristics." In this work, nonrational modelling is clearly defined by an exponentially decaying learning rate (in Q-learning modelling) and a satisficing search strategy (CSUD).

A unified decision making theory must also be willing to admit the *irrational*. This is necessary to allow for decision making processes to be driven by substance abuse, pathology, erroneous beliefs, or simply for instituting a paradigm shift. Smith (2020) argues that irrationality cannot be eradicated, and it must instead be accepted as the dual of rationality. There is "continuous movement between the two poles of rationality and irrationality."[5] Hegelian dialectic, where thesis is followed by anti-thesis, which is then followed by synthesis, already contains the seeds of the irrational. For example, how is the evolution of an anti-thesis possible without irrational rebellion?[6] However, there is also a pernicious type of irrationality "of knowing what the best thing to do is while instead doing the opposite."[7]

[5] Smith, 2020, p. 14.

[6] Consider for example The Rebel by Camus (1962).

[7] Smith, 2020, p. 273.

Here it is proposed that models of decision making should be able to accommodate irrational behaviour. In general, the class of iterative learning models is capable of inducing irrational behaviour by learning rate modification. For example, exponential learning rate decay can induce computational convergence prior to statistical convergence. Alternatively too big a learning rate could overshoot any optimum, or consistently diverge from it. In such models the learning rate controls the contribution of new information to the already existing information. Further, the proposed CSUD constrained search specification can restrict behavioural parameter and performance hyper-parameter ranges, and consequently define input value ranges where rational or irrational behaviour may be obtained.

The concept of a decision making theory must be broad enough to be able to model good, poor, and realistic decision making. In particular, with iterative learning models, one needs approaches that acknowledge finiteness, incomplete information, can move from a poor to a good decision, and that can alternate between poor and good decision making.

Consider for example, the earnest transformative learning experience shared in Smith (2020), "In the writing of this [his] book, mostly between 2016 and 2018, I [the author, Smith] quit drinking, I bought a Fitbit and a blood-pressure monitor, I closed my Facebook account (a plague on humanity worse than any drug), I finally committed to being fully honest with everyone in my life, and I got my long-sloppy finances in order. I pulled myself together, wised up: finally carried out the 'impossible syllogism' and realized I've got only a finite amount of time to do everything I want to do. I got rational, in my limited and relative way."

Smith's experience summarizes the human experience with the *rational*: an excess of the rational brings about an excess of the irrational, yet when rationality is applied with an understanding of its limitations, decision making improves. This is what this work strives to achieve, and this perhaps is what Gigerenzer (2016) wishes to achieve with a theory of "rational heuristics."

The terms rational, nonrational, and irrational may be ill-suited to describe decision making. An alternate classification system consisting of the terms *causal*, *correlational*, and *acausal* may make more sense. The primary goal of any decision is to affect a causal relationship between decision and outcome. Hence one strives for decisions, which produce causal results.

In the absence of sufficient accumulated information, a correlational decision, where there is high correlation between decision and outcome could be the next best option. When one knows nothing, or when one aims to disassociate decision and outcome, an acausal approach, such as a random selection, or doing what one wants anyhow, comes to mind. The cognitive science driven decision making models discussed in section 2.4, as well as this work, can account for poor as well as good decisions, and provide a more rounded, humanistic decision making approach.

2.3 Emotion and Neuroscience

The Q-learning model IGT calibrations are inspired by the role of the vmPFC in emotion valuation and consequently in decision making. Section 2.3.1 below presents a short summary of major emotion models, and reviews aspects of emotion considered interesting from a decision making point of view. This is followed by a discussion on the role of the vmPFC.

2.3.1 Emotion

A unified decision making theory should have some mechanism for addressing non-verbal reasoning such as emotion (Plutchik, 2001) or emotional intelligence (Goleman, 2005). But what role could emotions play in decision making? If heuristics are the Swiss army knife or adaptive toolbox of decision making,[8] then emotions are the Swiss army knife of heuristics. When it comes to decision making, emotion plays a key and heuristic role. When people reason, they "take different scenarios apart and . . . perform a cost / benefit analysis of each of them."[9] However, this process takes time, and without emotion, making a decision will "[a]t best, . . . take an inordinately long time."[10]

Expanding on the Swiss army knife role of emotions, Böhm and Pfister (2008) propose that emotions provide four contributions to decision making: information, speed, relevance, and commitment. Emotions motivate one to search for information, to decide as quickly as possible, to focus on relevant details, and to persist in the face of uncertainty.

[8] Gigerenzer and Gaissmaier, 2015, p. 912.

[9] Damasio, 2006, pp. 170-171.

[10] Damasio, 2006, p. 172.

From a psycho-evolutionary viewpoint, Plutchik (2001) proposes that the aim of emotions is to activate behaviours to ensure achievement of an equilibrium or target. From a behavioural and neuroscience perspective, Rolls (2013, Ch. 2, p. 4) defines emotions as, "states elicited by rewards and punishers, that is, by instrumental reinforcers," where an instrumental reinforcer is any cognitive input, which can influence stimulus-response associations.

There exist many additional definitions and theories of emotion. For the sake of completeness, some of the major emotion theories are briefly introduced. Based on facial expressions, Ekman (1992, p. 550) proposes that there are six basic emotions, consisting of "happiness, surprise, fear, sadness, anger, and disgust combined with contempt." Ekman's model has been very popular in computer science and automated facial emotion recognition. However, the link between facial expressions and emotions remains contested (Heaven, 2020). Based on cross-cultural surveys, Plutchik (2001) develops a complex emotion classification system known as the emotion circumplex, where emotions are grouped into bipolar opposites and additionally vary in intensity. Another well-known classification based system is the cognitive Ortony et al. (1990, OCC) model, where emotions are generated by cognitions elicited from outcome perceptions, taking into account consequences, attribution, and attraction. The OCC model can generate and classify many multiples of emotions. Finally, the pleasure, arousal, and dominance (PAD) model generates, according to Floyd, 1997, emotions along "three independent bipolar dimensions: pleasure-unpleasant, arousal-unaroused, and dominant-submissive."[11]

The above models all form good candidates for modelling emotion. This work, however, focuses on decision making, and for that reason, employs the Rolls (2013) and Plutchik (2001) definitions, which readily lend themselves to decision making and iterative learning. The Plutchik (2001) definition suggests that evolutionary processes have produced behaviour adaptation, that is, learning mechanisms, where emotions guide learning behaviour. This approach is employed in the burst learning model in chapter 10. The Rolls (2013) definition on the other hand produces a link between behavioural reinforcement learning and computational reinforcement learning, by creating a conceptual framework where outomes produce emotions,

[11] Floyd, 1997, p. 85.

which accrue over time and are valued in the sense of reward and punishment to produce behavioural adaptations.

While irrationality and emotion are introduced as important concepts in decision making, this does not imply that there is a causal link between emotion and irrationality. Nor can it be assumed that rational decision making is superior to emotion mediated decision making.

The notion that emotion is irrational or inferior to rational thinking resulted from combining Descartes' dualistic separation of mind and body (Damasio, 2001) with the subsequent James-Lange theory, which attributed emotions to the body (Rolls, 2013, Ch. 2, pp. 32-35). This implication coupled with the mind-over-body belief produced a connotation that emotion is inferior to rational thinking. However, this view is not supported by scientific and clinical research. Damasio (2006) states, "[w]hen emotion is entirely left out of the reasoning picture, . . . , reason turns out to be even more flawed than when emotion plays bad tricks on our decisions."[12]

Neuroscience, where the link between emotion and decision making is being forged, is considered next.

2.3.2 Neuroscience

The connection between emotion, intelligence, and decision making, has developed from work in clinical human and animal studies (Ernst & Paulus, 2005; Hornak et al., 2003; Kringelbach, 2005; Rolls, 2000). In their discussion of emotional intelligence and the brain, Hogeveen et al. (2016), identify in humans the **ventromedial prefrontal cortex** (**vmPFC**), the insula, the anterior cingulate cortex, and the amygdala as the key brain areas where emotion generation and processing takes place. They further propose that as the four mentioned areas are additionally implicated in general intelligence, emotional and general intelligence are intertwined, with emotional intelligence itself "measuring individual differences in one's ability to integrate emotions into cognitive operations (e.g., using reward valuations to drive the items held in working memory, using inhibitory control to regulate aversive emotional reactions)."[13] The ventromedial prefrontal cortex (vmPFC) is also known as the orbitofrontal cortex (OFC) and has been abbreviated in the initial IGT literature as VMF (Bechara et al., 1994). Krawczyk (2002, pp.

[12] Damasio, 2006, p. xviii.

[13] Hogeveen et al., 2016, p. 703.

633-635) provides a detailed discussion of the terms VMF and OFC. This work retains the use of the term vmPFC throughout.

From a decision making perspective, reward valuation and working memory are of interest. Reward (and punishment) valuation permits value aggregation and comparison across alternative choices, while working memory implicitly suggests a limit on input information.

Ever since the tragic yet remarkable case of Phineas Gage[14] in 1848 (Teles, 2020), in clinical terms, vmPFC impairment patients generally present with good intellectual ability, however, are unable to engage in decisions requiring forward planning tasks, with this deficit leading to a serious decline in personal and professional relationships (Damasio, 2006; Goel et al., 1997). Such patients perform well on the usual battery of psychological tests (e.g. the Wisconsin Card Sorting Test), and can even acknowledge or verbalize poor decisions, but they cannot learn from any poor outcomes (Bechara et al., 1994; Damasio, 2006).

In what has proved to be a pivotal contribution, Bechara et al. (1994) introduce a clinical test, now known as the **Iowa Gambling Task**, or **IGT**, which is capable of identifying vmPFC impairment. The premise of the test is simple: the participant faces four card decks, two of which produce on average winning net yields, while the other two produce on average losing net yields. Each net yield consists of a positive payout and a negative penalty, that is a reward and a punishment. At each turn, the participant must draw a single card from any deck, and observe the net yield achieved. The participant is tasked with winning as much as is possible, and faces two unknowns: (1) the net yield evolution of each deck, and (2) the length of the task. The participant is expected to discover the good decks.

The assessment tool provided by the IGT has lead to the development of a large area of research, which has produced further IGT variants, has applied these variants to wider patient populations, and has proposed models and heuristics, which may explain observed IGT behaviour. Bechara et al. (2000) discuss a non-stationary IGT version, where card rewards and punishments change every 10 turns. Lin et al. (2009) discuss a variant called

[14] Phineas Gage was a railroad worker who as a result of an accidental explosion suffered, and yet survived, a one meter long iron bar impalement to his skull, damaging his frontal lobes. He is the first known case presenting with vmPFC impairment. Teles (2020) presents a detailed account of Phineas Gage's case.

the Soochow Gambling Task (SGT), where deck yields are adjusted to produce expected value and gain-loss frequency effects observed in healthy IGT participants. The IGT has been applied to substance abuse (Ahn et al., 2014; Wood et al., 2005), gambling addiction (Ciccarelli, 2017), mental illness (Sevy et al., 2007), neurological illness (Busemeyer & Stout, 2002), as well as older adult populations (Wood et al., 2005). Using the IGT, Fellows and Farah (2003, 2005) investigate reversal learning, where an initially learned stimulus-response association is extinguished and re-learned. The large body of work on the IGT has also produced a large healthy participant data set available for research (Steingroever et al., 2015).

The IGT has generated a large body of work focusing on decision making. The exact physiological mechanisms driving normal and vmPFC impaired emotion valuation outcomes, however, continue to be debated. The somatic marker hypothesis (SMH) discussed in Bechara et al. (1997) and Damasio (1998), proposes an involuntary feedback mechanism where a physical or virtual body sensation is associated with a particular emotion. This feedback mechanism develops with decision making experience and can pre-empt or influence decisions. The SMH has been criticised for its complexity and proximity to the James-Lange theory of emotions, with reversal learning, which has been introduced above, proposed as an alternative (Dunn et al., 2006; Maia & McClelland, 2004, 2005). It has been suggested that vmPFC impairment in turn leads to reversal learning impairment (Fellows & Farah, 2003). Therefore vmPFC lesions leading to vmPFC impairment provide a direct and simpler alternative explanation to the SMH (Fellows & Farah, 2005).

In relation to decision making, the debate on the physiological mechanism of vmPFC impairment continues (Reimann & Bechara, 2010). Nevertheless the discourse around the IGT has contributed to the development of many cognitive decision making models. Section 2.4 presents key cognitive decision making models. Emotions are not explicitly formulated in these models. These models, however, have a distinctly different flavour than the rational models surveyed in section 2.5. In particular, the formulation of cognitive models allows for decision making shifts, which could be attributed to the result of vmPFC driven emotion valuation.

2.4 Cognitive Iterative Decision Making Models

This section surveys cognitive iterative decision making models, which can be considered as having nonrational tendencies as discussed in section 2.2. The models presented do not require infinite knowledge or lifespan to guarantee a good decision, and may even produce a poor decision.

The following sections aim to unify notation as much as is possible. Greek lettering is used for commonly occurring hyper-parameters. Time is always represented as discrete time, and α is always used to denote the learning rate. While this approach introduces some divergence from the respectively cited authors' notations, it is hoped that this unified notation will make it easier to compare and contrast the various models discussed.

2.4.1 The Expectancy Valence Model

Busemeyer and Stout (2002) discuss, among other models, a version of the **Expectancy Valence** (**EV**) model, where valence results from reward and punishment scoring, and decisions are made on the basis of expected valence, which is discovered over time. Choices are probabilistic and use a Boltzmann rule[15], with the choice with the highest expected valence having the highest probability of being chosen. The EV model produces a probabilistic choice strategy subject to three critical performance determining hyper-parameters: (1) the attention weight a, which determines the relative contribution of the reward and punishment to valence v_t, (2) the update rate (learning rate) α, which determines the relative contribution of the current valence score, and (3) the iteration dependent sensitivity c, which adjusts the greediness of probabilistic choice with higher sensitivity leading to increasing the probability of choosing the highest valence. In the EV model, sensitivity can increase or decrease over time, leading to exploration decrease or increase respectively.

[15] To be precise, choice probabilities are determined by the softmax activation function, which is a normalised version of the Boltzmann distribution. The term "Boltzmann rule", however, is commonly used and this practice is continued here without loss of any generality as the two concepts only differ by the specification of a normalisation constant. Please see Bishop (2006, p. 198 and p. 387) for details.

The formulation of the EV model with cost c_t and reward r_t arising from choice i at iteration t thus becomes

$$v_t = (1-a)r_t - ac_t \tag{2.1a}$$

$$E_t\left[v_t^i\right] = \begin{cases} \alpha v_t + (1-\alpha)E_{t-1}\left[v_{t-1}^i\right], & \text{if } v_t \text{ from choice i} \\ E_{t-1}\left[v_{t-1}^i\right], & \text{otherwise} \end{cases} \tag{2.1b}$$

$$Prob_{t+1}\left[i\right] = \frac{exp\left(E_t\left[v_t^i\right]s_t\right)}{\sum_i exp\left(E_t\left[v_t^i\right]s_t\right)}, \quad s_t = (t/10)^g \tag{2.1c}$$

where v_t^i is called the valence of choice i, and E_t denotes the conditional expectation. Note that the learning rate α, the attention weight a, and the sensitivity g do not change over time. Note that $0 < a, \alpha < 1$ and $g \in \mathbb{R}$.

In the IGT literature, costs are denoted in negative numbers. However, to simplify comparison with future models, here the convention is adopted that costs are positive and hence must be subtracted from rewards. The notation here reflects this convention. In symbols $\forall i, t,\ r_t > 0$ and $c_t > 0$.

Is the EV model described in (2.1) rational or nonrational? Is it causal, correlational, or acausal? The EV model has strong structure regarding the evolution of valence v_t, and on that basis, it appears causal. However, due to the probabilistic choice mechanism (2.1c), it is considered to be correlational.

Given the expectations terms in (2.1b), the EV model appears rational. It is however nonrational. This stems from the paradoxical application of stochasticity and its resolution via the expectations operator. Firstly, the nature of stochasticity is not very clear. For example, does it arise from cost c_t and reward r_t, or is it a result of measurement error built into valence determination in (2.1a), or both? Secondly, if expectations can be computed, then the relevant probability density distributions must be known ex-ante. But if this were the case, the problem, rationally speaking, could be formulated as an n-armed bandit problem, introduced in section 2.5.1, as bandit problems comprise the best way to choose among i alternatives when means are conditionally known. On the other hand, if expectations must be learned over time, then (2.1b) with a constant learning rate α cannot theoretically lead to expectations convergence.

Hence, the expectations operator in (2.1b) is interpreted as an abridged shorthand for a limited ability to filter uncertainty, and on that basis, the EV model is considered as being nonrational. The primary strength of the EV model is in its formulation of performance hyper-parameters: potential loss aversion via the attention weight a, a potentially limited learning horizon via the learning rate α, and intertemporal exploration effects via sensitivity g.

With the help of these three performance hyper-parameters, the EV model can generate a range of choices, poor as well as good. Using maximum likelihood, Busemeyer and Stout (2002) fit Iowa Gambling Task (IGT) outcomes generated by healthy, Huntington and Parkinson's affected participants to the EV model. The IGT outcomes are scored by the percentage of cards chosen from good, that is on average positive net yield, decks. The experiment results show that the percentage of cards chosen from the good decks increase throughout the 100 turns of the test for healthy controls, and to a smaller extent for Parkinson's patients. However, for Huntington's patients, the percentage of cards chosen from the good decks declines. The EV model receiving support from the Huntington's poor decision making outcomes in the IGT is hyper-parametrised with a negative sensitivity value, and a relatively high learning rate, which in combination lead to increasing exploration and high emphasis on the most recent outcome. In sum, the EV model is a nonrational model with some performance hyper-parameters leading to configurations, which can generate poor as well as good decisions.

2.4.2 The Prospect Valence Model

Ahn et al. (2008) present results from fitting IGT and Soochow Gambling Task (SGT) outcomes to Expectancy and Prospect Valence model variants. They present a large study, whose full scope goes beyond this review. Here only **Prospect Valence** (**PV**) model variants are reported. The primary difference between the Expectancy and Prospect Valence models is the introduction of the prospect utility function, attributed largely to Kahneman and Tversky (1979).

The prospect utility function can be seen as a non-linear scoring filter, and captures not only loss aversion as noted in the EV model above in section 2.4.1, but also accommodates gain-loss frequency effects. Gain-loss frequency effects refer to a heuristic where a gain or loss, which occurs frequently is commensurately emphasized more in relation to a gain or loss of a larger amount occurring rarely.

Retaining notational conventions, including the convention that $r_t, c_t > 0$, the Prospect Valence model is summarised for cost c_t and reward r_t arising from choice i at iteration t as

$$v_t = \begin{cases} (r_t - c_t)^b, & \text{if } r_t - c_t \geq 0 \\ a \left| r_t - c_t \right|^b, & \text{otherwise} \end{cases} \tag{2.2a}$$

$$E_t\left[v_t^i\right] = \begin{cases} E_{t-1}\left[v_{t-1}^i\right] + \alpha\, d_t^i \left(v_t^i - E_{t-1}\left[v_{t-1}^i\right]\right), & \text{if Rescorla-Wagner} \\ \alpha E_{t-1}\left[v_{t-1}^i\right] + d_t^i\, v_t^i, & \text{if decay-reinforcement} \end{cases} \tag{2.2b}$$

$$d_t^i = \begin{cases} 1, & \text{if } v_t \text{ from choice } i \\ 0, & \text{otherwise} \end{cases} \tag{2.2c}$$

$$Prob_{t+1}\left[i\right] = \frac{exp\left(E_t\left[v_t^i\right] s_t\right)}{\sum_i exp\left(E_t\left[v_t^i\right] s_t\right)}, \quad s_t = \begin{cases} (t/10)^g, & \text{if iteration dependent} \\ 3^g - 1, & \text{if iteration independent} \end{cases} \tag{2.2d}$$

where v_t^i is called the valence of choice i, and E_t denotes the conditional expectation. Note that the learning rate α, the attention weight a, utility shape b, and the sensitivity g do not change over time. Note that $0 < \alpha < 1$ and $a, b \in \mathbb{R}$. When $b \to 0$, the functional shape expressed in (2.2a) increasingly becomes step-like. Further when iteration dependent, $g \in \mathbb{R}$; however, when iteration independent $g \in [0, 5]$. In the iteration independent case, when $g = 0$, $3^g - 1 = 0$, and the softmax choice rule weights each choice i equally leading to fully random choice selection.

The Rescorla-Wagner expected utility update rule (Rescorla & Wagner, 1972) in (2.2b) is also used in (2.1b) above. The Rescorla-Wagner rule describes the well-known parameter update form employed in many branches of stochastic approximation (Spall, 2003, pp. 23-30). In contrast to Rescorla-Wagner updating, the decay-reinforcement update rule, discussed in Erev and Roth (1998), increases emphasis on the most recent outcome while geometrically discounting past outcomes.

Choice in the PV model is also conducted via the Boltzmann, or softmax, rule (2.2d). The iteration dependent sensitivity rule has already been discussed above in the EV model. The iteration independent sensitivity rule optionally implements exploration, which remains constant throughout iterations.

The PV model exhibits in the expectation term, the same difficulties, which have already discussed with the EV model. If expectations are known, then better rational formulations exist. If expectations are not known, then a constant learning rate cannot theoretically guarantee expectations convergence. Hence, as with the EV model, the PV model is classified as a nonrational correlational model.

The main contribution of the PV model to nonrational decision making is the addition of the gain-loss frequency parameter b. The gain-loss frequency is an important heuristic aiming at capturing the simple observation that it is hard, perhaps impossible, to learn the population density function of rare events. In that context, decision makers will score with higher emphasis more frequently occurring gains or losses.

Ahn et al. (2008) fit the discussed PV model variants to IGT and SGT outcomes obtained from healthy participants, and conclude that the PV model provides better fits and prediction quality than the EV model.

2.4.3 The Outcome-Representation Learning Model

Haines et al. (2018) present a decision making model, referred to as the **outcome-representation learning** model (**ORL**), which addresses not only expected value assessment and gain-loss frequency, but also choice perseveration and reversal learning. Choice perseveration is a term used to describe the exploitation versus exploration trade-off, and also to refer to the related win-stay/lose-shift heuristic (Worthy & Maddox, 2014). Reversal learning, as previously discussed, refers to unlearning a previously learned response and re-learning in its place an alternative response. Reversal learning is engaged when the initial choice stops being advantageous (Fellows & Farah, 2003).

Owing to its integration of four different heuristics, the ORL model is complex and consists of six equations. However, the main contribution of the ORL model is not its combination of four heuristics, but rather its use of distinct learning rates for positive and negative yields to model loss aversion. The use of dual learning rates is motivated by results from neuroscience indicating that positive and negative outcomes may be processed by different receptors (Cox et al., 2015).

Retaining notational conventions, including the convention that for any choice i, $r^i_t, c^i_t > 0$. First define for any choice i with cost c^i_t and reward r^i_t at iteration t

$$x_t^i = r_t^i - c_t^i \tag{2.3a}$$

$$\alpha = \begin{cases} \alpha_+, & \text{if } x_t^i \geq 0 \\ \alpha_-, & \text{otherwise} \end{cases} \tag{2.3b}$$

$$\alpha' = \begin{cases} \alpha_-, & \text{if } x_t^i \geq 0 \\ \alpha_+, & \text{otherwise} \end{cases} \tag{2.3c}$$

$$C = \begin{cases} 1, & \text{if } count(i) = 1 \\ count(i) - 1, & \text{otherwise} \end{cases} \tag{2.3d}$$

where x_t^i in (2.3a) denotes net yield, and the learning rate α in (2.3b) may take two distinct values α_+, or α_- depending on whether the net yield is 0 or more, or negative respectively. α' in (2.3c) is used for scoring reversal learning in (2.4c) below, and reverses the aggregation logic in (2.3b). Note that (2.4c) below applies to all actions j, which have not been selected. Hence, the reversal of the learning rates in (2.3c) provides a mechanism for reducing the usage count contribution of actions, which were previously advantageous. C in (2.3d) is a constant, which reflects the number of choices, which were not chosen, and which is used in calculating reversal learning cost.

Next the outcome-representation learning model is summarised for i choices with net yield x_t^i at iteration t as

$$E_t\left[v_t^i\right] = \alpha x_t^i + (1-\alpha)E_{t-1}\left[v_{t-1}^i\right] \tag{2.4a}$$

$$E_t\left[f_t^i\right] = \alpha \cdot sgn\left(x_t^i\right) + (1-\alpha)E_{t-1}\left[f_{t-1}^i\right] \tag{2.4b}$$

$$E_t\left[f_t^j\right] = \alpha' \cdot -sgn\left(x_t^i\right)/C + (1-\alpha')E_{t-1}\left[f_{t-1}^j\right], \quad j \neq i \tag{2.4c}$$

$$p_t^i = \begin{cases} 3^{-g}, & \text{if } i \\ 3^{-g}p_{t-1}^i, & \text{otherwise} \end{cases} \tag{2.4d}$$

$$V_t^i = E_t\left[v_t^i\right] + dE_t\left[f_t^i\right] + fp_t^i \tag{2.4e}$$

$$Prob_{t+1}\left[i\right] = \frac{exp\left(V_t^i\right)}{\sum_i exp\left(V_t^i\right)} \tag{2.4f}$$

where the performance determining hyper-parameters consist of the net gain learning rate, $0 < \alpha_+ < 1$, the net loss learning rate, $0 < \alpha_- < 1$, choice perseverance decay, $g \in [0, 5]$, frequency effects aggregation weight, $d \in \mathbb{R}$, and choice perseverance aggregation weight, $f \in \mathbb{R}$.

(2.4a) reflects standard net yield aggregation seen in reinforcement learning. (2.4b) and (2.4c) use the sign of the net yield x_t^i to generate a usage score based on choice frequency. The usage score increases for a repeatedly selected net gain, and decreases for a repeatedly selected net loss (2.4b). The usage scores for all alternatives foregone equally increase for any choice yielding a net loss, and equally decrease for any choice yielding a net gain (2.4c). Choice perseverance in (2.4d) is specified as a simple trace decay. When a choice is selected, its trace decay weight is reset to the highest level.

(2.4e) provides aggregation of the expected value (2.4a), the frequency and learning reversal effects (2.4b)-(2.4c), and the choice perseverance effect (2.4d). Finally, choice is affected via the basic Boltzmann rule in (2.4f).

In the nomenclature introduced, the ORL also comprises a nonrational correlational model. If the time horizon could be extended to infinity, subject to some stochastic process and learning rate decay restrictions, then in a rational iterative learning setting (2.4a) would be sufficient to generate a clear choice strategy based on unconditional net yield means. However, the reality of human existence, knowledge and time constraints make the infinite view impossible.

The one criticism facing the ORL model is that it is highly structured and complex. The model was specifically built to generate and assess observed human IGT outcomes. In spite of this criticism, when applied to human data sets, the five performance hyper-parameters, α_+, α_-, g, d, and f appear to have statistically significant effects, and produce good one-step-ahead prediction and simulation results (Haines et al., 2018).

The ORL model concludes the review of cognitive iterative learning and decision making models. The RL model with learning rate decay, presented in chapter 5, is also capable of generating limited expectations and loss aversion in terms of greedy choice, gain-loss frequency effects, choice perseveration and reversal learning. These effects are achieved with three performance hyper-parameters as opposed to five in the ORL. In chapter 5, the pathways for achieving the discussed heuristic effects are different. Further, owing to the lower number of hyper-parameters, the RL (Q-learning) model is not able to attribute heuristic performance on a per parameter basis.

Finally, as a motivating factor for the development of the CSUD algorithm, the use of Parameter Space Partitioning (PSP) employed, in the IGT outcome comparison of the EV and PV models (Steingroever et al., 2013), is mentioned. PSP can map the performance hyper-parameter value ranges associated with any classification scheme (Pitt et al., 2006). For example, in the IGT, the selection of more cards from the good decks could be a classification criterion. PSP is a Monte-Carlo based search method, where in order to establish performance ranges, hyper-parameter space is sampled statistically. CSUD is also a search method, but contracts grid search via the use of a loss function gradient.

2.5 Rational Iterative Learning Models

The models reviewed in this section have their origins in statistics, engineering, and control theory. In the respective literatures, the word *rational* is rarely used to describe these models. Here, the word rational is used to highlight that these classes of models have strong probabilistic and temporal assumptions or requirements, which would appear unrealistic from a strictly heuristic perspective. Whilst heuristically unrealistic, the main aim of the below models is to precisely engage in a type of reasoning or decision making, which humans cannot do. It will also be seen that model assumptions, which make engineering sense, may in a social science context lead to unintended complications such as using infinity to smooth out time horizon constraints.

2.5.1 N-Armed Bandits

Given i stochastic net yield streams $\{x_t^i\}$, n-armed or multi-armed bandit problems aim to learn the process with the best yield. The term bandit originally referred to a single lever slot machine with a negative net yield (Sutton & Barto, 2018, p. 18). Bandit problems seek to balance exploitation, using what is known, with exploration, searching for new information. The aim in bandit problems is to minimise loss, known as regret, which given T iterations, measures the loss arising from the difference between the best and the selected choices. Bubeck and Cesa-Bianchi (2012) discuss bandit problems with various implementations of regret. They define pseudo-regret, alternatively known as total expected regret (Kuleshov & Precup, 2014), as

$$L_T = T \max_i Ex^i - \sum_{t=1}^{T} Ex^{i(t)} \tag{2.5}$$

where L_T denotes loss after T iterations, $max_i Ex^i$ denotes the choice with highest expected yield, and $Ex^{i(t)}$ is the expected net yield of the actual choice in iteration t. That is the second term in (2.5) is the sum of the expected net yields of the T choices actually made.

In specifying (2.5), $\{x_t^i\}$ are assumed to be independently and identically distributed (i.i.d.) with finite means and variances. This in turn means that the $\{x_t^i\}$ comprise stationary and ergodic processes. Stationarity indicates that yield distributions do not change over time, while ergodicity means that repeated measurements will ensure that one will eventually have visited sufficient outcomes so as to be able to characterise each yield process accurately.

The simple bandit problem with strong assumptions, as in (2.5), proves surprisingly difficult to solve, and constitutes a major research area in decision making theory (Guha et al., 2010). Bubeck and Cesa-Bianchi (2012) discuss versions of the bandit problem where the strong assumptions underlying (2.5) are relaxed, for example as in adversarial bandits, where the net yield processes are set by an adversary. Gittins and Glazebrook (2011) focus on Markovian bandits, where yields are generated by Markov processes, which relax the i.i.d. assumption above.

Bandit problems remain difficult to solve because (1) iterations are limited, (2) when expectations are not known they must be discovered, and (3) when expectations are known a forecast for the next iteration must be constructed. Further difficulties arise when the maximum iteration budget T is less than the number of choices $count(i)$. In such cases, some of the choices must be ignored, but which ones should be ignored?

One solution to the pseudo-regret problem posed in (2.5) has been proposed by Auer et al. (2002) and is referred to as the upper confidence bounds (UCB) algorithm. According to Kuleshov and Precup (2014) in the UCB, each choice is initially chosen once. Following this initial directed exploration period, given i choices with actualised net yields x_t^i per iteration, at iteration T the best choice would be described by

$$i_T = arg \max_i \left(\frac{1}{n_i} \sum_{t=1}^{T-1} x_t^i + \sqrt{\frac{2lnT}{n_i}} \right) \tag{2.6}$$

where n_i denotes the number of times choice i has been chosen up to iteration T, $T > 1$, and $\sum_i n_i = T - 1$. The first term on the right hand side of (2.6) denotes estimated sample net yield while the second term constitutes an exploration premium, and increments relatively more, the sample means of less frequent choices. Note that the exploration premium is independent of actual observed yields, and is only a function of the current iteration T and the number of times choice i has been chosen, n_i. The exploration premium term is derived from the Chernoff-Hoeffding bound and estimates the bounds of a one-sided confidence interval, which would contain the *unobserved* true net yield for choice i with "overwhelming probability."[16]

While humans are unlikely to consciously perform in real-time, the mathematical operations driving bandit problem solutions, Costa et al. (2019) test rhesus monkey performance in an n-armed bandit task, where novel options requiring exploration are introduced over time. They find that two key subcortical regions are involved in the exploitation versus exploration trade-off. Exploration involves the amygdala while exploitation involves the ventral striatum. In corresponding simulations, the authors use a partially observed Markov decision making process (POMDP), which produces a choice strategy with an exploration premium, in principle similar to the second term in (2.6); however, encapsulated in Boltzmann activation as seen in (2.4f). This raises the question whether the monkeys, at a biological level, use a computational approximation comparable to a bandit algorithm.

In sum, while at a higher mental level real-time conscious use of bandit algorithms remains unlikely, it is possible that evolutionary mechanisms have lead to the development of an automatic sub-cortical scoring mechanism, which handles exploration of novelty.

Despite their complexities, bandit algorithms likely comprise the best approach for making the best choice among ergodic net yield processes. Note that due to the presence of exploration, bandit algorithms are considered to be rational correlational decision making algorithms.

[16] Auer et al., 2002, p. 237.

2.5.2 Dynamic and Approximate Dynamic Programming

There are two major differences between n-armed bandits and the approaches discussed in this section: (1) bandit problems focus on the use of a specific loss function called regret, and (2) dynamic (DP) and approximate dynamic programming (ADP) approaches are capable of outputting for the problem in question, a complex action or choice selection profile called a **policy**, typically denoted by π.

The decision making space is broken into learning iterations (i.e. time), states of the world, outcomes, and possible choices (actions). The purpose of the DP or ADP algorithm, is to score choice outcomes over time and states, and then produce an optimal policy of actions, π^*. Enumeration of time, state, outcome, and choice space, however can lead to the exponentially explosive proliferation of possibilities, a problem known as the the curse of dimensionality. Approximate dynamic programming is an effort to mitigate the curse of dimensionality by means of techniques, which bring about dimensionality reduction (Bertsekas, 2012; Powell, 2011). Reinforcement learning (RL), a machine learning technique (Sutton & Barto, 2018), can be seen as an ADP technique, where scoring is conducted of the basis of a stochastic net yield stream $\{x_t^i\}$, measured in response to the choices taken. In DP, ADP, and RL problems, the scoring (objective optimisation) function can itself be learned over time.

Given a sequence of loss functions $\{L_t\}$, the fundamental decision making problem is specified as selecting policies π to minimise expected loss over the planning horizon

$$\min_{\pi} E\left[\sum_{t=0}^{T} \gamma^t L_t\left(s_t, a_t\right)\right] \tag{2.7}$$

where s_t, a_t denote the state, and action (choice) at iteration t. L_t is an outcome scoring function. γ^t denotes a discount rate, which reduces future outcome scores, that is, indicates a preference for the present over the future. It should be noted that both action and state could be vectorised. However, for ease of exposition, scalar notation is employed. Finally note the expectations operator, an implication of which is that expectations of stochastic output variables can be taken, and that transitions from one state of the world into the next can be enumerated probabilistically.

In general (2.7) can be expressed iteratively in the Bellman equation form as a forward looking value equation

$$V_t\left(s_t\right) = \min_{a_t}\left(L_t\left(s_t, a_t\right) + \gamma E_t\left[V_t\left(s_{t+1}\right)\right]\right) \tag{2.8}$$

where E_t indicates conditional expectations. (2.8) assumes that the production of s_t constitutes a Markov process. Given assumptions on state transition dynamics, (2.8) can be solved backwards from the final period T. When the discount rate $\gamma < 1$, then (2.8) can also be solved for an infinite time horizon. (2.8) seems to imply that one can forecast the future relatively well, and that one paradoxically knows the future before one knows the present. However, (2.8) is better seen as a strategic plan answering the question, "which actions must be taken to achieve the lowest cost outcome in T periods from now?" For example, given aviation congestion risks, Kochenderfer (2015, pp. 249-276) presents an ADP application for automated airborne collision avoidance.

Q-learning initially proposed by Watkins (1989) constitutes an approximation to (2.8), where one can step forward through time (Kochenderfer, 2015, p. 122). Similar to the arms of a bandit, and the best choice bandit algorithm presented in (2.6), in Q-learning, the world is divided into state-action pairs, which are scored, and at each iteration, the best possible action is chosen. Given states s_t, s_{t+1}, and action a_t, Q-learning is formulated as

$$Q_{t+1}\left(s_t, a_t\right) = \alpha_t\left(L_t\left(s_t, a_t\right) + \gamma \min_{a_{t+1}} Q_t\left(s_{t+1}, a_{t+1}\right)\right) + \left(1 - \alpha_t\right) Q_t\left(s_t, a_t\right) \tag{2.9}$$

where the minimisation term proposes that once state s_{t+1} is observed, the action with the least approximated accumulated loss should be chosen prior to updating accumulated loss scores.

Tsitsiklis (1993) shows that under certain regularity assumptions, when $T \to \infty$, Q-learning converges to an optimal policy $\pi^* = \{a_t, \ldots, a_T\}$. Note that just like bandit algorithms, the Q-learning algorithm also requires some form of exploration. This is typically implemented as Boltzmann exploration as seen in (2.4f), or as ϵ-Greedy exploration as will be discussed in section 5.2.2.

Note that (2.9) is similar in structure to (2.1b), the Rescorla-Wagner branch of (2.2b), and (2.4a). The EV, PV, and ORL models discussed in sections 2.4.1, 2.4.2, and 2.4.3 respectively, share the theme of approximating the central tendency of a decision making value by stepping forward in time.

Q-learning and RL techniques have attracted the attention of neuroscientists as plausible computational mechanisms underlying choice selection in humans and primates. For example, Schultz et al. (1997) propose that dopamine release encodes the temporal difference error, that is the discrepancy between predicted and realised Q-values. Further serotonin release modulates the discount rate γ, acetylcholine modulates the learning rate α_t, and noradrenaline controls exploration (Doya, 2002). These neuro-transmitter correlates suggest that both computational bandit and RL models may have biological implementations.

DP requires a model or knowledge of state transition probabilities. In both ADP and RL, knowledge of the state transition and the value function (2.8) is not required. Provided sufficient samples have been obtained, the state transition and value functions can both be iteratively estimated from simulated or realised outcomes. RL methods such a Q-learning (2.9) sample both time and states, and comprise "model-free"[17] approaches. In the proposed nomenclature, DP, ADP, and RL methods remain as rational correlational. It could be argued that DP methods in particular with their strong probabilistic assumptions could be considered as rational causal methods. However in stochastic environments, the use of the term causal should be qualified either with a confidence interval or a test statistic.

2.5.3 Gaussian Mixture Models

Gaussian mixture models can be seen as an application of non-parametric statistics, where the parameters describing a statistical distribution are allowed to go to infinity. Gaussian refers to population density functions, which have a Gaussian kernel

$$k(x) = \frac{1}{\sqrt{2\pi}} e^{-x^2/2} \tag{2.10}$$

where, by construction

$$\int k(x)dx = 1, \quad E\left[k(x)\right] = 0, \quad 0 < Var\left[k(x)\right] < \infty \tag{2.11}$$

[17] Kochenderfer, 2015, pp. 121-124.

that is, $k(x)$ is a normalized, mean zero, finite variance process, where the functional form in (2.10) places some restrictions on tail decay behaviour (Wasserman, 2006).

A (suitably normalised) Gaussian mixture model consists of multiple (parametrised) Gaussian kernels, which in principle can even be infinite in number. Gaussian mixture models can therefore by construction represent multi-modal distributions, and in that manner address one of the shortcomings of central limit theorems, which asymptotically produce unimodal distributions (McKean, 2014).

In general given iteration t, and letting $K_t(x_t|\mu_{ti}, \sigma_{ti}^2)$ denote a suitably normalized Gaussian kernel with mean μ_{ti} and variance σ_{ti}^2, the Gaussian mixture is formulated as

$$pdf_t(x_t) = \sum_i v_{ti} K_t \left(x_t | \mu_{ti}, \sigma_{ti}^2 \right) \tag{2.12}$$

where v_{ti} are the mixing weights, $\sum_i v_{ti} = 1$, $v_{ti} \in [0, 1]$, and scalar notation is used for simplicity. In practice (2.12) is solved using the Expectation Maximization algorithm (Bishop, 2006, pp. 430-455).

Gaussian mixture models (GMM) have been used in artificial neural networks for automated hyper-parameter optimisation (Nogueira, 2014–; Stander & Craig, 2002). Agostini and Celaya (2010) apply GMM techniques to Q-learning with continuous state-action spaces. They formulate joint Gaussian probability density functions (pdfs) consisting of scalar action, vector state, and scalar Q-value components. The Gaussian pdfs are combined in a variable unit mixture model and Q-value updates for state-action tuples are derived from the incrementally estimated (marginal) means of the Gaussian mixture model. Using GMMs allows calculation of Q-value variances, which in turn contribute to exploration, consisting of a variance-contributed and a fully random term. Pinto (2021) presents a GMM driven deep Q-learning model, which subject to tuning, can learn from a single pass through data.

Mori et al. (2022) train a neural network to output GMM parameters for predicting future behavioural states of C. elegans, a type of nematode (worm). Next they use the estimated GMM together with virtual nematodes in a reinforcement learning setting, where the virtual nematodes are induced to move towards a goal point. Finally, they extract the successful virtual behaviour policy, apply it to real nematodes, and find that the policy can successfully direct real nematode behaviour. That is, the authors are able to summarise and predict nematode behaviour with a GMM. This shows that a GMM, when augmented with reinforcement learning, can predict for any given stimulus, its expected response. Hence a GMM can provide predictive decision making output.

Evolving Gaussian mixture models (GMMs) of the form shown in (2.12) can be seen in a Bayesian context with each iteration producing an updated posterior pdf. GMMs are considered to be rational correlational models. Of the rational models discussed, GMMs encapsulate the highest amount of prior information. This prior information is in the form of initial mixing coefficients, means, covariance structure, and tail decay behaviour.

This is more prior information than what would be required in a classical Bayesian setting. However in practice, the multi-modal form and control of tail behaviour is extremely useful in modelling results of machine learning simulations, where decision criterion output often exhibits multiple optima, and due to simulation limitations, tail data may be scarce. When tail data is scarce, fitting different GMMs, or more generally, kernel mixture models, can produce sensitivity analyses regarding tail decay effect on hypothesis testing.

2.6 Emotion in Iterative Learning

Because emotion is a complex phenomenon (Rolls, 2013), the introduction of emotion into computational decision making constitutes a difficult undertaking. In section 2.3.1 major emotion models, which have been automated, have already been discussed. Here proposals are considered, where a rational model produces some intermediate output, which in turn is used for emotion synthesis. The resulting emotions may then lead to amended choices.

Broekens et al. (2015) present joy-distress and hope-fear labelled emotions, which are synthesized from the reinforcement learning value function (2.8). However, the synthesized emotions are not input into the decision making problem. Their emotion synthesis approach is reviewed here, as it provides a good introduction. They specify at iteration t with net yield x_t

$$J(s_{t-1}, a_{t-1}, s_t) = (x_t + V(s_t) - V(s_{t-1}))\,(1 - Prob(s_{t-1}, a_{t-1}, s_t)) \tag{2.13a}$$

$$H(s_t) = max\left(-V(s_t), 0\right) \tag{2.13b}$$

where $J > 0$ denotes joy, $J < 0$ represents distress, $H \geq 0$ is the hope-fear dimension. The term $(1 - Prob(s_{t-1}, a_{t-1}, s_t))$ in (2.13a) is the complement of the last observed transition probability, and scales the (undiscounted) temporal difference error $(x_t + V(s_t) - V(s_{t-1}))$. For example, a high (positive) temporal difference error and low state transition probability will generate higher joy due to the realisation of a better than probabilistically anticipated outcome. Hope in (2.13b) is constrained to be high when the value function is negative. Hence in the Broekens et al. (2015) model, hope increases with adversity.

The Broekens et al. (2015) model in (2.13) highlights two key difficulties faced in computational emotion synthesis. Firstly when multiple emotions are generated from transformations of a single or very few feedback signals, this could in principle lead to identifiability problems. For example, if all emotions were synthesized from linear transformations of the same value signal, essentially only a single emotion continuum would exist. Secondly, emotion labels, such as happy, sad, angry, and so on are anthropomorphisms, and therefore largely arbitrary. The same model could be developed with alternative terminologies.

Antos and Pfeffer (2011) present a utility function based model where utility across goals is summed. The value of each goal is in turn determined by three components: relative importance, priority, and degree of achievement. Their model labels five emotions: hope, fear, boredom, anger, and sadness, each activated by a specific activation function. Emotion activation modulates goal priorities and indirectly influences utility maximisation. In 5-armed bandit simulations, the authors find that the emotion agent's accumulated net yield is only surpassed by an all-knowing and an optimal index agent, both of which possess substantial additional environmental knowledge. In sum, their emotion heuristic approach achieves good results with minimum environmental knowledge.

Moerland et al. (2018) survey 52 papers, published between 1998 and 2016, modelling emotion modulated reinforcement learning agents, where emotion is synthesised from net yields, from internal appraisals such as the temporal difference error, from homoeostatic drives such as energy level, or from hard-wired emotion mechanics. Emotion activation modulates decision making via net yield, state, action, or hyper-parameter alterations.

Given the difficulty with emotion labelling, and the possibility of acceptable alternative terminologies, this work uses emotion labels sparingly. Chapter 5 initially presents a reinforcement model, without labelled emotion; however, which strictly speaking under the Moerland et al. system could be considered as a hyper-parameter modulated emotion model. Chapter 10 then extends this initial model to the two emotion burst learning model.

While dealing with emotion in computational decision making is challenging, emotions can act as decision making heuristics. Further, as humans are emotional beings, computational use of emotion, be it synthetic or imitative, can be useful in establishing affective connections in human interaction (Damiano et al., 2012). The models discussed here fall into the nonrational causal (in a mechanical automation sense), or nonrational correlational categories.

Chapter 3

Learning from Repeated Sampling

This chapter presents a learning from repeated sampling framework, which can be rational context compliant; however, which is extended to formalise nonrational contexts. Two extensions producing a nonrational context are suggested: (A) an exponentially decaying learning rate, and (B) a custom loss function.

Being able to learn by means of repeated sampling requires that the sampled outcomes provide accurate and relevant information relative to the decision making control variables. Or, if that is not the case, that at least such accurate and relevant information can be obtained *eventually* as a result of following a learning process. This chapter initially assumes that such dynamics, as stipulated in rational models, are achieved. Another important consideration, as discussed in section 2.5.1, where multi-armed bandits are reviewed, is how to allocate existing resources to assess (sample) N potential courses of action. Hence, the basic problem in repeated sampling is to assess N options in T iterations, where T is as small as possible, even when N is large. Throughout this chapter it is assumed that estimation must be done sequentially across time.

In general, this is a difficult problem. However, in some cases, one of the most effective solutions is the $1/N$ heuristic, where for example, in an investment context $1/N$ of funds are allocated to N assets. Zhou and Palomar (2020) show that over a 10 year period, the $1/N$ heuristic exhibits one of the lowest maximum drawdown (MDD) outcomes, where maximum drawdown refers to the distance between an asset portfolio's previous peak and the subsequent trough. While MDD is a forensic, retrospective measure, the $1/N$ heuristic was only bettered by portfolio rules where funds were allocated in inverse proportion to volatility. However such inverse volatility rules require substantially more computational resources, and they also require that samples across time are collected.

3.1 General Framework

The approach employed in this work consists of using a known loss function $Y(\cdot)$ to score an intractable objective function $R(\cdot)$. The framework introduced here belongs to the class of multi-stage estimator frameworks. Such a technique is used in Widrow and Hoff (1960), whose work has formed the basis of the back-propagation algorithm (LeCun et al., 2012; Rumelhart & McClelland, 1987; Rumelhart et al., 1986). A similar two-stage iterative bootstrapping strategy is employed by the expectation-maximisation algorithm (Bishop, 2006, pp. 435-441). Further, the gradient approximation technique CSUD proposed later in chapter 12 adds to the class of Simultaneous Perturbation Stochastic Approximation (SPSA) (Spall, 1992) algorithms. On a conceptual level, CSUD loosens typical rational model guardrails, which guarantee via assumptions that the result desired from the outset is actually achieved. The desire to remove such guardrails may appear unusual from a rational view point; however, it is believed that loosening rational assumptions may assist in exploring model input space.

In the proposed framework, analytical derivatives are not required. The loss function $Y(\cdot)$ and the objective function $R(\cdot)$ may each have its own selection logic. The loss function $Y(\cdot)$ scores the objective function $R(\cdot)$ indirectly by means of a reductive performance statistic $M(\cdot)$, which is sampled from repeated applications of the chosen objective function, and is subsequently fed into the loss function. Problem space inputs are decoupled into behavioural parameter and performance hyper-parameters. The loss function may be arbitrarily specified in performance parameters, and then these performance parameters may be tuned via the approximated loss gradient.

Under appropriate assumptions, the framework proposed is fully rational model compliant. When such assumptions are presented, always the simplest case is presented. In a rational context, the proposed technologies have been implemented in computational artificial neural network tuning (Liaw et al., 2018). Here such tuning technologies are formally expressed as multi-stage estimators. The aim here is to present a theoretical framework, which may unify rational and nonrational modelling, and which can help assess the exploitation versus exploration performance of rational and nonrational learning models.

3.2 General Framework Specification

At iteration t, let Ψ_t and Θ_t denote behavioural parameters and performance tuning hyper-parameters respectively. Let $R_t(\Psi_t, \Theta_t) : (\mathbb{R}^q, \mathbb{R}^p) \to \mathbb{R}^o$, $(p, q, o \geq 1, 2, \cdots)$, be a vector valued stochastic function, mapping the $(p+q)$-dimensional parameter and hyper-parameter inputs (controls) into o-dimensional outputs. Ideally R_t is to be optimised jointly with respect to Ψ_t and Θ_t. However, the structure of R_t is such that joint optimisation is not analytically feasible. For example, R_t may be linear in the parameters but non-linear in the hyper-parameters making derivation of an analytical form impossible.

One could optimise R_t directly via Monte-Carlo techniques. Such an approach would be more susceptible to the curse of dimensionality originating from the size of the input space, the size of the output space, and the need to take multiple samples (Powell, 2011, pp. 112-113). Further, joint optimisation of behavioural and performance inputs presents attribution challenges arising from mixing behaviour and performance effects on outcomes.

The method proposed here presents a structured alternative to direct Monte Carlo sampling. By assessing parameter and hyper-parameter effects separately, with the help of a flexible performance measure $M_t(\cdot)$ and scoring criterion $Y_t(\cdot)$, the problem dimensionality is reduced, and this aids in gaining better understanding of underlying process dynamics.

At iteration t, with parameters Ψ_t, given hyper-parameters $\hat{\Theta}_t$, performance measure $M_t(\cdot)$, and loss function $Y_t(\cdot)$, the general problem is formulated as

$$Y_t(\cdot) \equiv Y_t\left(M_t\left(\left\{\underset{\Psi_{it}}{sel}\, R_t\left(\Psi_{it}, \hat{\Theta}_t\right)\right\}_{i=1}^{N}\right)\right) \tag{3.1a}$$

$$\hat{\Theta}_{t+1} = \hat{\Theta}_t - \alpha_t \nabla_{\hat{\Theta}_t} Y_t(\cdot) \tag{3.1b}$$

where *sel* is a selection operator; for example, *sel* may be 'maximise', 'minimise', 'median', or 'take top 5 percent.' Further, $\hat{\Theta}_t$ is the current loss minimising hyper-parameter (vector) estimate, the term $\{sel_{\Psi_{it}}\, R_t\left(\Psi_{it}, \hat{\Theta}_t\right)\}_{i=1}^{N}$ denotes a sequence of outputs derived from N applications of selecting $R_t(\cdot)$

with respect to Ψ_{it} given $\hat{\Theta}_t$, and $\nabla_{\hat{\Theta}_t} Y(\cdot)$ denotes the gradient of $Y_t(\cdot)$ with respect to $\hat{\Theta}_t$.

(3.1) specifies a repeated sampling learning algorithm, where learning is summarised in the value-evolution of parameters Ψ and given hyper-parameters $\hat{\Theta}$.

At each iteration t, (3.1a) is solved sequentially from the innermost to the outermost criterion. That is, one first simulates the selection of $R_t(\cdot)$ N times to get an output sequence. Next, the performance measurement criterion $M_t(\cdot)$ reduces the repeated selection outputs to a performance criterion. For example, one may wish to compute the mean of the outputs. Finally, the performance criterion result is scored by the loss function $Y_t(\cdot)$, which produces via the update equation (3.1b) the next iteration hyper-parameter candidates $\hat{\Theta}_{t+1}$.

The update rule (3.1b) is well-known in stochastic approximation and has been studied for cases when analytic gradients are available or must be approximated (Kushner, 2010; Robbins & Monro, 1951; Spall, 2003). Indeed (3.1b) is one of the most frequently encountered update strategies in computational algorithms. For example while not discussed here, (3.1b) is used in back-propagation (LeCun et al., 2012). With respect to gradient approximation (3.1b), this work proposes CSUD (Constrained Single Unconstrained Double) simultaneous perturbations stochastic approximation), which is discussed in detail in chapter 12.

Up to now, (3.1) has been discussed from a mathematical and computational point of view, using for conciseness, the equation rather than algorithmic form. One might naturally ask whether iterating (3.1) goes to anywhere useful? By useful, one is suggesting iterative convergence to an optimal parameter and hyper-parameter combination (Ψ^*, Θ^*). That is, given (3.1) with CSUD, can it be asserted that

$$\lim_{t \to \infty} \Psi_t \to \Psi^* \quad \text{and} \quad \lim_{t \to \infty} \hat{\Theta}_t \to \Theta^* \tag{3.2}$$

where Ψ^* and Θ^* denote optimal parameter and hyper-parameter settings respectively?

In any rational model, the convergence criteria (3.2) would be achieved via corresponding assumptions. Section 3.4 below presents some simple assumptions for achieving such optimality results. The full set of assumptions and indicative proofs, however, are discussed in chapter 12.

```
initialise $Scores$;
while $iterating$ do
    $t \leftarrow$ iteration counter;
    $\hat{\Theta}_t \leftarrow$ GET;
    initialise $M_t$;
    for $i = 1$ to $N$ do
        $M_t \leftarrow$ ACCUMULATE $sel_{\Psi_{it}} R_t \left(\Psi_{it}, \hat{\Theta}_t\right)$)
    end
    $M_t \leftarrow$ REDUCE $M_t$;
    $\hat{\Theta}_t \leftarrow$ UPDATE USING $Y_t(M_t)$;
    $Scores \leftarrow$ YIELD $Y_t(M_t)$;
    $t{+}{+}$;
    UPDATE $iterating$;
end
```

Algorithm 1: Conditional Sequential Optimisation

3.3 Nonrational Search of Parameter and Hyper-Parameter Spaces

The short answer as to whether optimal parameter and hyper-parameter results can be guaranteed in a nonrational setting is, "no, they cannot." However, this does not mean that optimal results cannot be achieved. It only means that one cannot theoretically guarantee the existence of such results. In a nonrational context, (3.1) is simply seen as a means of getting answers to stipulated search criteria. It remains the search algorithm operator's task, to assess the search results.

One of the consequences of (3.1a) is that joint selection of Ψ_t and Θ_t is replaced with sequential conditional selection. Given any hyper-parameter estimate $\hat{\Theta}_t$, Ψ_t is optimised, $\hat{\Theta}_t$ is updated; then the cycle is repeated, guided by the prior structure imposed on $\hat{\Theta}_t$ via the performance statistic and the form of the loss function.

Hence, the selection, or optimisation, problem in (3.1) is really of the form shown in Algorithm 1, where the selection of parameters is contingent on having selected some hyper-parameters first. In general, this is the usually followed sequential practice, for example, when assessing performance of artificial neural networks (A. Li et al., 2019; L. Li et al., 2018).

For repeated samples $i = 1 \ldots N$, it should be noted that accumulation of parameter Ψ_{it} selection outcomes can produce variations due to the stochastic nature of $R_t(\Psi_{it}, \hat{\Theta}_t)$. This work does not discuss any considerations or resolution methods for dealing with parameter variation. In the context of reinforcement learning for example, Tsitsiklis (1993) discusses conditions for achieving a stationary policy, which constitutes a generalised notion of Ψ^*.

3.4 Rational Search of Parameter and Hyper-Parameter Spaces

This section provides an example of the rational requirements for achieving (3.2) with (3.1) and CSUD, that is, for converging to optimal parameters and hyper-parameters as learning iterations go to infinity. CSUD is discussed in detail Chapter 12, which also presents propositions and proofs. However, the salient features are introduced here without any loss of generality.

Recall that the inner selection loop (3.1a) has input and output value dimensions consisting of $R_t(\Psi_t, \Theta_t) : (\mathbb{R}^q, \mathbb{R}^p) \to \mathbb{R}^o$, $(p, q, o \geq 1, 2, \cdots)$, and that the selection of $R_t(\cdot)$ is sampled N times.

Suppose that for any given hyper-parameter set $\hat{\Theta}_t$, N samples of $sel_{\Psi_{it}} R_t\left(\Psi_{it}, \hat{\Theta}_t\right)$ produce the solution set $\Omega^{\hat{\Theta}_t}_{\Psi^*}$. Assume that performance measure mapping reduces the solution set into r-dimensional performance criteria. That is, $M(\Omega^{\hat{\Theta}_t}_{\Psi^*}) : \mathbb{R}^{No} \to \mathbb{R}^r$. Further, assume the performance criteria m can be represented in simplified form as a time invariant mean process, consisting of a central estimate $\hat{m}(\hat{\Theta}_t)$ and a random error term ε

$$m = \hat{m}(\hat{\Theta}_t) + \varepsilon, \quad m \text{ has } r \text{ elements}, \quad \varepsilon \sim i.i.d \tag{3.3a}$$
$$E\left[\varepsilon\right] = 0 \tag{3.3b}$$
$$E|\varepsilon_j^2| < \infty, \quad E\varepsilon_j^3 = 0, \quad E|\varepsilon_j^4| < \infty, \quad j \in \{1, \cdots, r\} \tag{3.3c}$$
$$E\left[\varepsilon\varepsilon'\right] = \Sigma = \begin{bmatrix} \sigma_1^2 & & \\ & \ddots & \\ & & \sigma_r^2 \end{bmatrix} \tag{3.3d}$$

where some restrictions on higher order moments (3.3c) have been imposed, whilst allowing for heteroscedasticity (3.3d).

Finally given $\hat{\Theta}_t$, formulate loss in quadratic form as

$$\Upsilon(\hat{\Theta}_t) = \left[m(\hat{\Theta}_t) - m^*\right]' S \left[m(\hat{\Theta}_t) - m^*\right] \tag{3.4}$$

where $m(\hat{\Theta}_t)$ is an r-dimensional performance measurement, m^* is a performance measurement target, and S is an $r\ x\ r$ diagonal matrix of weights.

Using (3.3a), the quadratic loss function in (3.4) can be expanded as,

$$\begin{aligned} Y(\hat{\Theta}_t) &= \left[\hat{m}(\hat{\Theta}_t) - m^* + \varepsilon\right]' S \left[\hat{m}(\hat{\Theta}_t) - m^* + \varepsilon\right] && (3.5a) \\ &= \left[\hat{m}(\hat{\Theta}_t) - m^*\right]' S \left[\hat{m}(\hat{\Theta}_t) - m^*\right] + \varepsilon' S \varepsilon + \ 2\varepsilon' S \left[\hat{m}(\hat{\Theta}_t) - m^*\right] && (3.5b) \\ &= L(\hat{\Theta}_t) + \epsilon(\hat{\Theta}_t) && (3.5c) \end{aligned}$$

where

$$\begin{aligned} L(\hat{\Theta}_t) &= \left[\hat{m}(\hat{\Theta}_t) - m^*\right]' S \left[\hat{m}(\hat{\Theta}_t) - m^*\right] && (3.6a) \\ \epsilon(\hat{\Theta}_t) &= \varepsilon' S \varepsilon + \ 2\varepsilon' S \left[\hat{m}(\hat{\Theta}_t) - m^*\right] \ . && (3.6b) \end{aligned}$$

Note that the error term $\epsilon(\hat{\Theta}_t)$ includes a quadratic component, and therefore has non-zero mean. That is, $E\left[\epsilon(\hat{\Theta}_t)\right] = tr(S\Sigma) < \infty$.

Therefore given the assumptions in (3.3), the stochastic loss function in (3.5c) produces biased loss estimates. However, this bias is time invariant.

As noted in Chapter 12, CSUD is a type of simultaneous perturbations stochastic approximations (SPSA, Spall, 1992) algorithm, where gradient estimates are calculated from two separate loss measurements, $Y(\Theta)$ and $Y(\Theta')$. Note that $E\left[\epsilon(\Theta) - \epsilon(\Theta')\right] = 0$. It follows that CSUD with quadratic loss, as in (3.4), does produce asymptotically unbiased gradient, and consequently asymptotically consistent and unbiased hyper-parameter estimates Θ^*, which minimise loss. The expectation of loss remains biased. However, this bias is time invariant, and therefore does not affect loss rankings.

Assumptions have been presented, under which it can shown that hyper-parameter estimates $\hat{\Theta}_t$ converge to the least loss generating set Θ^*. Convergence arguments have been outlined. It should be noted that even for a case with assumptions as simple as those presented here, proving such convergence is a non-trivial task. Chapter 12 provides a detailed roadmap for such proofs.

While the quadratic wrapper function (3.4) provides some search regularization, it may no longer be possible to guarantee that a global optimum, in the sense of finding a unique parameter and hyper-parameter (Ψ^*, Θ^*) tuple. All one will be able to say is that Ψ^* fulfils minimum loss criteria subject to Θ^*.

The convergence difficulties illustrated for (3.2) are not unique to the general framework in (3.1), or to the properties of CSUD. In any rational framework, proving iterative convergence requires assumptions, some of which may be unverifiable, or costly to verify. Rational convergence theory is not able to provide a threshold number of iterations after which convergence is guaranteed. Consequently, the boundary between rational and nonrational versions of iterative learning approaches is blurred, and often relies on the algorithm operator, additional metadata, or budgetary considerations for when to practically stop learning.

3.5 Switching between Nonrational and Rational Contexts

It is proposed that the learning rate α_t determines whether the recursive update rule (3.1b) operates in a rational or nonrational context.

The recursive update rule (3.1b) is central to gradient driven iterative learning. Intuitively, it is easy to see that (3.1b) is a first order difference equation, with a steady state for some Θ^* where $\nabla_{\Theta^*} Y(\cdot) = 0$ (the minimum, or a minimum, of (3.1a)).

Note that (3.1b) is used to search (3.1a) for any minima. Therefore α_t cannot be constant, as a constant α_t would produce perpetual oscillations about any minimum. Further, one can surmise that at a steady state α_t can be any value; however, in order to reduce oscillations around Θ^*, α_t must decay over time. Finally at or near $\alpha_t = 0$, computational convergence can be forced, without necessarily knowing that a steady state has been reached. In sum given (3.1a) and (3.1b), achieving theoretical guarantees of convergence to Θ^* requires (1) that α_t must decay; yet (2) that there exist some speed limits, which are imposed on the decay of α_t, so that α_t does not decay too quickly or too slowly.

Note that with a deterministic one-dimensional quadratic (squared) loss function, with 1^{st} and 2^{nd} analytical derivatives available, by setting the learning rate equal to the inverse Hessian, $\alpha_t = \nabla^2_{\hat{\Theta}_t} Y(\cdot)^{-1}$, convergence can be achieved in a single iteration.

With multi-dimensional stochastic loss functions, using the inverse Hessian as the learning rate α_t promises fastest theoretical convergence. Computation of the inverse Hessian is difficult and forms an important area of research in stochastic approximation. For example, Zhu et al. (2020) present an LBL factorisation of the Hessian with the resulting approximated inverse Hessian leading to a reduction in per iteration computational costs from

$O(p^3)$ to $O(p^2)$, where p is the dimensionality of the hyper-parameter vector $\hat{\Theta}_t$. The authors' reduction in computational costs arises from the dimensionality reduction induced via LBL. From a rational perspective therefore setting the learning rate equal to the inverse Hessian provides the fastest iterative pathway towards Θ^* subject to Hessian approximation and inversion constraints.

Alternatives exist to using resource intensive Hessian approximation. A well-known alternative rational approach is to set α_t equal to a scalar, subject to the following conditions (Spall, 2003, pp. 105 - 108)

$$\alpha_t > 0, \quad \lim_{t \to \infty} \alpha_t = 0, \quad \sum_{t=0}^{\infty} \alpha_t = \infty, \quad \sum_{t=0}^{\infty} \alpha_t^2 < \infty. \tag{3.7}$$

When the conditions in (3.7) are fulfilled, it can be shown that (3.1b) asymptotically converges to Θ^*. Hence (3.7) defines conditions in which the learning rate can be seen as being in a rational context. For example given $t \in \{0, 1, \dots\}$, $\alpha_t = 1/(1+t)$ fulfils these conditions.

With scalar α_t, when compared to the use of the inverse Hessian, the number of iterations to convergence by necessity increases. Powell (2011, pp. 419 - 452) discusses deterministic and stochastic step sizes, where reducing the number of step sizes to convergence is a key consideration.

Powell (2011, p. 427) also presents exponentially decaying learning rates and shows that such decay rules do not fulfil the convergence criteria in (3.7). Note that a constant learning rate $\alpha_t = \bar{\alpha} > 0$, as is frequently employed in neural network training, also fails to satisfy (3.7).

The context, where learning rate decay rules do not satisfy (3.7), is defined in this work as being nonrational. Hence when a learning rate sequence fulfils (3.7), the model using such a decay pattern is considered to be in a rational context. In contrast, when the learning rate decay profile does not fulfil (3.7), it is said that the model is operating in a nonrational context.

Heuristics constitute a key hallmark of a model operating in a nonrational context. For example, Riedmiller and Braun (1993) propose RPROP for the training of artificial neural networks. Rather than reducing the learning rate α_t in advance, RPROP uses an initially constant learning rate and relies on detecting suspected over-shoots of Θ^*, based on which overshoots the learning rate is heuristically adjusted. **Adam** (**Ada**ptive **M**oment Estimation) proposed in Kingma and Ba (2014), uses a related strategy of scaling a constant learning rate by using gradient moment metadata, in effect turning a constant scalar learning rate into a scaled per gradient element learning rate. With the exception of a heuristic to guard against division by zero, Adam can be seen in a rational context.

The nonrational context in this work uses decaying learning rates of the form

$$\alpha_t = \Lambda \alpha_{t-1} \tag{3.8}$$

where $\Lambda \in (0, 1]$. To generate a decaying learning rate sequence $\{\alpha_t\}$, (3.8) is used to construct an exponential decay sequence with:

$$\Lambda = e^{-\lambda} \tag{3.9a}$$

$$\alpha_t = e^{-\lambda} \alpha_{t-1} \tag{3.9b}$$

where $\lambda \in [0, \infty)$ is the decay factor. When $\lambda = 0$, the learning rate is constant. Note that for any λ, $\left(1 - e^{-\lambda}\right) * 100$ represents the constant per period percentage decay of the learning rate. Further, with (3.9b), given any initial learning rate α_1 for $t > 0$, α_t will decay at the same rate λ. This turns out to be helpful in increasing resistance to initial learning rate effects.

Given (3.9)

$$\lim_{t \to \infty} \sum \alpha_t = \alpha_1 / (1 - \Lambda) < \infty \tag{3.10}$$

where α_1 is the initial learning rate; it is in this sense that exponential learning rate decay violates (3.7). (3.10) implies that for some planning horizon T, learning rate decay λ, and initial learning rate α_1, one obtains $\sum^T \alpha_t \approx \alpha_1 / (1 - \Lambda)$, which shows that exponential learning rate decay induces a forward planning horizon limit T. Here, a limited forward planning horizon is seen a hallmark of nonrational systems. Note that such a nonrational limit can be applied directly by simply setting T, the number of learning iterations.

Chapter 4

The Iowa Gambling Task

This chapter introduces the Iowa Gambling Task (IGT), a card selection task, and its variations. Human IGT results are used to calibrate single-state Q-learning applications of (3.1). The term IGT is used most generally to refer to all IGT variants and derivatives. When required by the context, however, the term IGT is used to refer to Iowa Gambling Task (Bechara et al., 1994) variants, and the term SGT is used to refer to the Soochow Gambling Task (Lin et al., 2009).

Section 4.1 introduces the IGT, its variants, asociated summary yield structures, and reports key human performance benchmarks relevant for bootstrapping IGT model behaviour. Section 4.2 reports in further detail literature analysis of human IGT outcomes. Section 4.2 also develops the performance analysis vocabulary, which is used in assessing Q-learning modelling results. Appendix A presents the draw-by-draw IGT yield structures used in this work.

4.1 The IGT and its Variants

The IGT is now presented in more detail. Since its inception in 1994, there have been many variants of the IGT.

The IGT and its variants involve the use of virtual money. This work implicitly assumes that virtual money is capable of producing some emotion involvement. Tsampallas et al., 2023 study a virtual money based roulette game, and conclude that virtual money driven gambling may produce some emotional reaction; however, they also note that their study does not assess the difference in emotion reactions to virtual versus real money gambling tasks.

In this work, the IGT variants are broadly classified as belonging to one of the original, re-shuffled, random, reversed, or Soochow categories. In the existing research literature, variants belonging to the same category have not always been implemented in identical manner. As far as it is known, implementation specific differences exhibited in the reported research literature will be highlighted in the below discussion.

4.1.1 The Original IGT

Description

The classification *original IGT* refers to the initial implementation of the IGT reported in Bechara et al. (1994). The original IGT is the first clinical test capable of successfully identifying vmPFC impairment in test subjects. As discussed in section 2.3.2, vmPFC impairment refers to a complex condition affecting the ventromedial prefrontal cortex, leading to forward planning deficits.

In the original IGT, the participant is loaned virtual money to play a card game, and is told to maximize profit inclusive of the loan repayment. The participant is presented with four card decks: A, B, C, and D. In each turn, the participant draws a single card from any deck. The participant then receives a fixed reward, and occasionally has to pay a fine. If the participant runs out of virtual money, additional loans are available.

While unknown to the participant, the decks and the game are structured as follows. Decks C and D are good decks, give low fixed rewards, low fines, and on average yield net gains. A and B are bad decks, yield high rewards, but even higher losses, and on average produce a net loss. Each bad deck starts with a misleading sequence of eight cards, where the player initially receives positive net gains. However, each bad deck, misleading sequence is subsequently followed by high fines, causing the player, on further selections from the same deck, to lose all gains and move into debt. The game stops after 100 turns, when the dealer announces the end. However, while playing, the participant does not know when the game will end.

Aim

The aim is to assess if participants are able to discover the good, low risk, on average net positive yield decks and choose accordingly. In the IGT, a score of more than 50 draws from the good decks is defined as a normative pass (Fellows & Farah, 2005).

Implementation Variations

Bechara et al. (1994, p. 9), replicated in Appendix A.1, present the original IGT payout sheet with rewards and fines consisting of 40 entries per deck. In Bechara et al. (2000), if any participant uses up all 40 entries for one deck, they must then choose from the remaining decks. Therefore, any participant cannot choose more than 80 cards from the good, or the bad, decks. When a participant realises that only a few cards are left in a particular deck, such a limit could potentially influence their exploration strategy.

In contrast, Steingroever et al. (2018) devise a computerised version, where subjects are able to choose more than 40 cards from each deck.[1] When comparing payoff structures between Bechara et al. (1994) and Steingroever et al. (2018), the fine structure varies for deck C. In Steingroever et al. (2018) deck C fine is always 50; however in Bechara et al. (1994, p. 9), this deck has the value range $\{25, 50, 75\}$. Steingroever et al. (2018) also give participants a task performance related bonus, a condition not specified in the original Bechara et al. (1994) specification.

Current Implementation

In the implementation here, deck C penalties use the original fine values consisting of $\{25, 50, 75\}$. Further for each deck, a steady state is induced. This steady state is achieved by looping the end of each deck to the beginning of the deck. This formalises the manner in which more than 40 cards can be sequentially drawn from the same deck, while retaining payout sheet ordering. As the IGT only lasts for 100 turns, it is believed that this looping strategy will not provide participants with added opportunities for tracking and memorising card locations with the best net payoff locations.

The potential to select more than 40 cards from the same deck removes any binding exploration versus exploitation constraints. Deck looping and its effects are considered in Appendix B.

[1] This is not mentioned in the paper, but can be seen by looking at the associated open data set available at Steingroever et al. (2015). Given a pool of 40 outcomes for each card deck, for each deck draw, a card is randomly selected from the corresponding card pool. It is not discussed whether random selection is with or without replacement.

Type	Bad Decks (A, B)	Good Decks (C, D)
Mean Net Yield	-25	25
At 100th draw:	Good deck choices	Cumulative Net Yield
	20	-1500
	50	0
	80	1500

TABLE 4.1
The original IGT decks. Theoretical infinite horizon mean net yields per draw and deck.

With the introduction of a steady-state, one can quantify the infinite horizon (rational) net yield properties of each card deck. Table 4.1 presents the infinite horizon mean net yields for the original IGT decks. It can be seen that the minimum normative pass criterion of choosing more than 50 cards from the good decks implies non-zero mean net yield. The re-shuffled, and random IGT environments, which are discussed below have identical long-term mean net yields. The rational infinite horizon solution illustrates that at infinity, the individual deck card sequencing differences among the original, re-shuffled, and random IGT variants in terms of the net mean yield disappear.

4.1.2 The Re-shuffled IGT

Description

The classification *re-shuffled IGT* refers to the implementation of the IGT by Fellows and Farah (2003, 2005) used to investigate reversal learning effects. As discussed in section 2.3.2, vmPFC impairment can be explained in terms of loss of reversal learning, that is, the ability to unlearn a previously learned response.

Aim

As in section 4.1.1 above, the aim is for the participants to discover the good, low risk, on average positive yield decks and choose accordingly. As before, the normative pass criterion, that is achieving a *normal* behavioural result, requires more than 50 draws from the good decks.

Implementation

In the original IGT, the first 8 cards of each deck, when selected in sequence, produce positive net yields. In the re-shuffled deck version, Fellows and Farah (2005) move the first 8 cards of each original deck to the end. Further in the original bad deck B, cards with payout indices 11 and 14 are switched. This re-shuffle removes the initial, misleading net positive yield conditioning sequence in the bad decks A and B. As a result of the re-shuffles, players experience, in all decks, rewards and fines relatively quickly. The details of the re-shuffled decks can be found in Fellows and Farah (2005, p. 59), and are also noted in Appendix A.2.

Current Implementation

The Fellows and Farah (2005) re-shuffles as discussed above are fully implemented. However, as with the implementation of the original IGT in 4.1.1, the end of each deck is looped to the beginning of the deck. As before, this introduces the ability to choose more than 40 cards from the same deck, and also induces a steady state, which has the infinite horizon payout structure shown in Table 4.1.

4.1.3 The Random IGT

Description

Both the original IGT and the re-shuffled IGT aim to test performance effects, which arise from how the cards in each deck are ordered. In the original IGT, the bad decks A and B each start with a misleading sequence where positive net yields are achieved. The test subject is fooled into thinking that the bad decks are good. In contrast in the re-shuffled IGT, the test subject is exposed immediately within the first 1-11 cards to the true and negative mean net yield nature of decks A and B. Both the original and re-shuffled IGT test sequential learning with exploration and exploitation, where the key learning effects may occur in the early learning iterations.

The question arises as to whether IGT behaviour changes when the sequential draw order is eliminated. Such randomized deck environments are employed in Horstmann et al. (2012) and Steingroever et al. (2018). The data for both studies is available in Steingroever et al. (2015) and via Haines et al. (2018).

Aim

As in sections 4.1.1 and 4.1.2 above, the aim is for the participants to discover the good, low risk, on average positive net yield decks and choose accordingly. As before, the normative pass criterion remains the same.

Implementation

It is not clear whether Horstmann et al. (2012) or Steingroever et al. (2018) implement randomisation with or without replacement. Both Horstmann et al. (2012) and Steingroever et al. (2018) present participants with a performance related bonus. In Horstmann et al. (2012), task duration is revealed to participants. In the framework here, it is assumed that task duration is unknown, hence for random environment comparisons, only the Steingroever et al. (2018) data is used.

Current Implementation

The Bechara et al. (1994) original IGT environment payout sheet is implemented as random draws without replacement. For each deck, the randomised payout pool is initially of size 40. When the last payout has been issued, the pool size returns to 40.

The end of each deck is looped to the beginning of the deck. As before, this introduces the ability to choose more than 40 cards from the same deck, and induces a steady state, which leads to the infinite horizon payout structure shown in Table 4.1.

4.1.4 Additional Investigated IGT Variants

The original, re-shuffled, and random IGT environments discussed in sections 4.1.1, 4.1.2, and 4.1.3 respectively constitute the primary focus in this work. With these three variants, the efficacy of outcome driven sequential learning can be investigated. Bechara et al. (2000), Chiu et al. (2008), and Lin et al. (2009) have also introduced additional IGT variants for investigating the effect of the timing, the progression, and the incidence of rewards and fines.

Chapter 7 presents RL model fitting to human results using IGT variant EFGH (Bechara et al., 2000), which is here referred to as the reversed IGT, and is introduced below. Chapter 8 presents RL model fitting to human results using the Soochow Gambling Task (SGT) (Chiu et al., 2008; Lin et al., 2009). The SGT is also described below.

Reversed IGT

In this work, version EFGH is referred to as the *reversed* IGT variant, where decks E and G are the good decks, producing frequent high fines with less frequent but higher rewards; and, F and H are the bad decks with frequent low fines but even lower less frequent rewards. The reversed IGT attempts to distinguish "hypersensitivity to reward" from "insensitivity to punishment."[2] The reversed IGT environment payout sheet is available at Bechara et al. (2000, p. 2193), and is also replicated in Appendix A.4. Chapter 7 presents an application with the reversed IGT environment implemented with looped deck structure.

The reversed IGT could test for loss aversion as in the EV (2.1a), PV (2.2a) models, or distinct reward and fine learning rates as in the ORL (2.3b) model. However, the application presented here in chapter 7 focuses instead on being able to identify the good decks.

Soochow Gambling Task (SGT)

As the original IGT became a well-known test and more patient populations were tested, it emerged that healthy controls had a hard time identifying bad deck B as being bad. In this context, Chiu et al. (2008) and Lin et al. (2009) discuss gain-loss frequency effects, and propose the Soochow Gambling Task (SGT) as an explanation for observed normal participant deck B behaviour.

The original IGT environment attempts to mislead the participant into seeing the bad decks as being good by using an initial sequence of positive net yields. In the SGT, the bad decks are not hidden by misleadingly good initial net yields, but by high frequency losses and low frequency gains. The good decks have high frequency losses with low frequency gains, but yield per 10 consecutive deck draws a net gain. The bad decks have high frequency gains with low frequency losses, but yield per 10 consecutive deck draws a net loss. The payoff sheet for the SGT can be found at Chiu et al. (2008, p. 15) or in Appendix A.5.

[2] Bechara et al., 2000, p. 2190.

Type	Bad Decks (A, B)	Good Decks (C, D)
Mean Net Yield	-50	50
At 100th draw:	Choices from good decks	Cumulative Net Yield
	20	-3000
	50	0
	80	3000

TABLE 4.2
The SGT decks. Theoretical infinite horizon mean net yields per draw and deck.

Chiu et al. (2008) and Lin et al. (2009) find that normal SGT participants on average choose cards from the bad decks A and B, and are not able to work out that in the long term, the good decks C and D indeed produce positive net yields. The authors conclude that frequency effects lead the participants to focus on immediate gain, and undervalue the probability of losses. In relation to rare events, these conclusions are mirrored in Hertwig et al., 2004, where under decisions from experience, that is uncertainty, rare events are probabilistically undervalued.

From a rational statistical perspective, it takes longer to learn a rare event distribution. In the case of the SGT, participants have to learn four rare event distributions. Hence, the SGT presents a challenging signal extraction problem, which humans may not be able to solve in 100 turns. Table 4.2 presents the SGT steady state net yield characteristics.

4.1.5 IGT Environments Not Considered

In addition to the above described IGT variants, the IGT literature also discusses adversarial environments. Such adversarial environments possess non-stationary reward and fine distributions. Bechara et al., 2000, p. 2194-2195 present an adversarial IGT environment consisting of 60 cards per deck. After every 10 cumulative same-deck draws, rewards increase and losses worsen; that is, good deck net yields increase, but the bad deck net yields worsen.

This adversarial IGT environment has been studied in mental illness by Premkumar et al. (2008), in ageing by Wood et al. (2005), and in addiction by Ahn et al. (2014) and Fridberg et al. (2010). This work, however, does not consider any adversarial IGT variants. Such variants do not substantially alter the fundamental problem structure posed in the original, re-shuffled, random, and reversed IGT; and in the SGT environments.

4.2 Analysis of IGT Human Behaviour

The analysis of IGT outcomes focuses on cumulative aggregated measures as well as blocked itemised measures. Cumulative aggregated measures assess the total number of good cards at trial completion, and attempt to summarise general performance as being normal or vmPFC impaired.

Blocked itemised measures assess individual deck, or good deck, performance in 10- or 20-block draw segments, and aim to examine exploration versus exploitation behaviour. Brand et al. (2007) administer the IGT to healthy subjects, and use repeated measurements MANOVA (multivariate analysis of variance) with 20-draw blocks as within-subjects factors to assess choices from good minus bad decks as the outcome. They find statistically significant differences ($p < .001$), indicating that per block response behaviour shifted towards the good decks as the IGT trial progressed.

Not all studies report all measures, not all measures are applied to all patient populations, and only some raw test data is available. This makes it difficult to use the same measures when assessing software agent replication of normal and vmPFC impaired human IGT variant results. However, good comparisons for cumulative aggregate measures can be obtained and cumulative aggregated outcomes will constitute the main reporting measure here. Additionally some results involving blocked itemised measures will be reported as well.

The main human IGT results, used for agent benchmarking and calibration, are presented next.

4.2.1 Cumulative Aggregated Measures

IGT task outcomes are typically assessed at the end of the trial, after 100 draws, by looking at the number of cards chosen from the good decks. Fellows and Farah (2005) state that the normative pass criterion for this measure is more than 50 cards chosen from the good decks.

Another cumulative criterion, which has been reported in the literature, is the number of cards chosen from the good decks minus the number of cards chosen from the bad decks (Bechara et al., 1994; Brand et al., 2007).

This work reports cumulative aggregated results in terms of the *fraction of cards chosen from the good decks*, abbreviated as f_G. For example, 1 indicates that all cards were chosen from the good decks. After 100 draws, at the end of an IGT trial, a result of more than 0.5 indicates normal human behaviour, whereas a result of 0.5 or less indicates vmPFC impairment.

Note that the fraction of good decks measure f_G used here can be transformed into the good decks and good decks minus bad decks measures used in the literature according to

$$\begin{aligned} f_G &= 0.5 + \frac{cards\ good - cards\ bad}{200} \\ cards\ good + cards\ bad &= 100 \end{aligned} \tag{4.1}$$

where fraction of cards chosen from the good decks $f_G \in [0, 1]$. (4.1) is useful for transforming reported variance, standard deviation, or standard error values.

The remainder of this section presents fraction of cards chosen from the good decks f_G results extracted from the IGT literature with the extraction sources and method noted on a per IGT environment basis.

Table 4.3 reports means and standard errors for the good decks, good - bad decks, and fraction of good decks f_G measures obtained from the literature using the original IGT environment, where the bad decks begin with misleading positive net yields. Note that the three measures reported in Table 4.3 are related as shown in (4.1). As not all studies reported in this table use the same outcomes measures, outcome measures are reported in each study's preferred format followed by the fraction of good decks f_G measure used in this work. With the exception of Fellows and Farah (2005), comparison of the healthy controls with vmPFC impaired subjects indicates that in the original IGT environment at task termination, healthy controls achieve a normative pass, while vmPFC impaired participants fail. Note that the vmPFC impaired Fellows and Farah (2005) results straddle the normative pass criterion, and will be viewed as indicative of a fail. As the original test data for Bechara et al. (2000), Bechara et al. (1994, 1998), and Fellows and Farah (2005) was not available, the graphical presentations are converted into numerical format using pixel matching.

Table 4.4 reports means and standard errors for the good decks, and fraction of good decks f_G measures obtained from the literature using the

Subjects	Study[a]	N	Good Deck Cards	(Good-Bad) Deck Cards	Mean fraction of good decks $\bar{f}_G^H$
Controls	Bechara et al. (1994)	44		37.00 ± 3.00	0.69 ± 0.015
	Bechara et al. (1998)	21	62.10 ± 3.17		0.62 ± 0.032
	Bechara et al. (2000)[b]	20		17.38 ± 3.79	0.59 ± 0.019
	Fellows and Farah (2005)	14	62.86 ± 2.32		0.63 ± 0.023
vmPFC Impaired	Bechara et al. (1994)	6		-25.10 ± 11.1	0.37 ± 0.055
	Bechara et al. (1998)	9	39.70 ± 3.49		0.40 ± 0.035
	Bechara et al. (2000)[b]	10		-9.66 ± 5.53	0.45 ± 0.028
	Fellows and Farah (2005)	9	50.00 ± 2.00[c]		0.50 ± 0.020[c]

[a] All values pixel match Computed.

[b] Results reported in 20 draw blocks. Calculation of 100 draw values here assume no inter-block covariance. This is may be incorrect, and the aggregated standard error may have been either under- or over-estimated.

[c] vmPFC impaired participants straddle the pass point, indicative of a fail.

TABLE 4.3
Original IGT environment. 100th draw cumulative good deck means ± SE.[a] Controls pass, while vmPFC impaired participants fail.[c]

re-shuffled IGT environment, where the bad deck misleading positive net yield sequences are re-shuffled to the end of the decks. Comparison of the healthy controls with vmPFC impaired subjects indicates that in the re-shuffled IGT environment at task termination, both healthy controls and vmPFC impaired subjects achieve a normative pass. As the original test data Fellows and Farah (2005) was not available, the graphical presentations are converted into numerical format using pixel matching.

Table 4.5 reports the mean and standard error for the fraction of good decks f_G measure obtained from the literature using the random IGT environment, where no card sequencing effects exist. Only data for normal

Subjects	Study[a]	N	Good Deck Cards	Mean fraction of good decks $\bar{f}_G^H$
Controls	Fellows and Farah (2005)	17	72.30 ± 3.74	0.72 ± 0.038
vmPFC Impaired	Fellows and Farah (2005)	9	66.76 ± 7.84	0.67 ± 0.078

[a]Pixel Match Computed.

TABLE 4.4
Re-shuffled IGT environment. 100th draw cumulative good deck means ± SE.[a] Both controls and vmPFC impaired participants pass.

Subjects	Study[a]	N	Mean fraction of good decks $\bar{f}_G^H$
Controls	Steingroever et al. (2018)	70	0.53 ± 0.023

[a]Computed from longitudinal data available at Steingroever et al. (2015).

TABLE 4.5
Random IGT environment. 100th draw cumulative good deck means ± SE.[a] Healthy (controls) pass.

participants is available. Results indicate that in the random IGT environment at task termination, healthy controls achieve a normative pass. f_G is directly computed from the longitudinal dataset, reported in Steingroever et al. (2015).

Table 4.6 reports means and standard errors for the fraction of good decks f_G measure obtained from the literature using the reversed IGT environment, where the good decks exhibit high fines with even higher rewards. Comparison of the healthy controls with vmPFC impaired subjects indicates that in the reversed IGT environment at task termination, healthy controls achieve a clear pass, while vmPFC impaired subjects straddle the pass point, and the vmPFC impaired mean $\bar{f}_G$ will be viewed as indicative of a fail. As the original test data Bechara et al. (2000) was not available, the graphical presentations are converted into numerical format using pixel matching. Further, as results were reported in 20-draw blocks, cumulative 100 draw values needed to be re-calculated. Re-calculation of cumulative 100 draw values assumes no inter-block covariance.

Table 4.7 reports the mean and standard error for the fraction of good decks f_G measure obtained from the literature using the **SGT** environment,

Subjects	Study[ab]	N	(Good-Bad) Deck Cards	Mean fraction of good decks $\bar{f}_G^H$
Controls	Bechara et al. (2000)	20	35.12 ± 4.87	0.68 ± 0.0243
vmPFC Impaired	Bechara et al. (2000)	10	1.58 ± 5.97	0.51 ± 0.0298

[a] Pixel Match Computed.

[b] Results reported in 20 draw blocks. Calculation of 100 draw values here assume no inter-block covariance. This is may be incorrect, and the aggregated standard error may have been either under- or over-estimated.

TABLE 4.6
Reversed IGT environment. 100th draw cumulative good deck means ± SE.[a] Healthy controls pass, while vmPFC impaired participants straddle the pass point, indicative of a fail.

Subjects	Study[ab]	N	Good Deck Cards	Mean fraction of good decks $\bar{f}_G^H$
Controls	Chiu et al. (2008)	48	40.13 ± 2.11	0.40 ± 0.0211

[a] Pixel Match Computed.

[b] Results reported per deck. Calculation of aggregated good deck values here assume no inter-deck covariance. This is may be incorrect, and the aggregated standard error may have been either under- or over-estimated.

TABLE 4.7
Soochow (SGT) environment. 100th draw cumulative good deck means ± SE.[a] Healthy controls fail the SGT. No published results for vmPFC impaired SGT outcomes have been found.

where the good decks exhibit frequent high fines with rare but even higher rewards. At task termination, healthy controls fail. As the original test data for Chiu et al. (2008) was not available, the graphical presentations are converted into numerical format using pixel matching. Further, results were reported per deck. Re-calculation of aggregated good deck values assume no inter-deck covariance.

IGT Variant	Rule	Behaviour	Mean	Range	Pass[c]
Original	The minimum and maximum of $\bar{f}_G^H$ from Table 4.3	Normal	0.64[b]	0.59 to 0.69[a]	✓
		vmPFC Impaired	0.44[b]	0.37 to 0.50	–
Re-shuffled	$\bar{f}_G^H$ ± 2 SEs from Table 4.4	Normal	0.72	0.64 to 0.80	✓
		vmPFC Impaired	0.67	0.51 to 0.83	✓
Random	$\bar{f}_G^H$ ± 2 SEs from Table 4.5	Normal[d]	0.53	0.49 to 0.58	✓
Reversed	$\bar{f}_G^H$ ± 2 SEs from Table 4.6	Normal	0.68	0.63 to 0.72	✓
		vmPFC Impaired	0.51	0.45 to 0.57	?
Soochow	$\bar{f}_G^H$ ± 2 SEs from Table 4.7	Normal[d]	0.40	0.36 to 0.44	–

[a] The Steingroever et al. (2015) dataset also contains original IGT environment results by Maia and McClelland (2004), which are not used due to the presence of an introspective questionnaire, which was administered after the 1st 20 draws and every 10 draws thereafter. It is noted that Maia and McClelland (2004) results produce a mean fraction of good decks $\bar{f}_G^H$ of 0.61, which remains inside the minimum and maximum values in Table 4.3.

[b] Computed as weighted averages from Table 4.3.

[c] Pass refers to a mean fraction of good decks score greater than 0.5, $\bar{f}_G^H > 0.5$. ✓ indicates a pass, – a fail, and ? an inconclusive result.

[d] No published results for vmPFC impaired outcomes have been found.

TABLE 4.8
IGT and SGT mean fraction of good deck $\bar{f}_G^H$ ranges used for comparing agent and literature results.

Computational Model Benchmarks

Table 4.8 shows the mean fraction of good deck $\bar{f}_G$ ranges used to compare computationally modelled IGT software agent results with the IGT literature results. For the original IGT environment, the literature reports a broad range of values. Therefore the minimum and maximum of the reported (human) mean fraction good deck values $\bar{f}_G^H$ from Table 4.3 is used. For the re-shuffled, random, reversed, and Soochow deck environments, with one study each per environment, the reported (human) mean fraction good decks $\bar{f}_G^H \pm 2$ standard errors from Tables 4.4, 4.5, 4.6, and 4.7 respectively are used.

Software agent IGT behaviour results are compared to the subject trial outcome values shown in Table 4.8. For the original IGT, the control and vmPFC impaired subject match ranges are [0.59, 0.69], and [0.37, 0.50] respectively. For the re-shuffled IGT, the control and vmPFC impaired subject match ranges are [0.64, 0.80] and [0.51, 0.83] respectively. For the random IGT, the control subject match range consists of [0.49, 0.58]. For the reversed IGT, the control and vmPFC impaired subject match ranges are [0.63, 0.72] and [0.45, 0.57] respectively. For the SGT, the control subject match range consists of [0.36, 0.44].

4.2.2 Blocked Itemised Measures

Blocked itemised measures consist of those, which assess individual deck, or a combined deck selection, over a specified number of draws. For example, Bechara et al. (2000) assess original and reversed IGT outcomes over five draw blocks consisting of draws 1-20, 21-40, 41-60, 61-80, 81-100 for good decks minus bad decks for normal and vmPFC impaired subjects. Fellows and Farah (2005) construct the same measure for good decks to assess original and re-shuffled IGT outcomes for normal and vmPFC impaired subjects. For normal subjects, Chiu et al. (2008) report results for the SGT using this 5-block 20-draw measure for good decks. Finally, using the data from Steingroever et al. (2015), this measure is constructed for the fraction of good decks f_G for the random IGT environment outcomes reported in Steingroever et al. (2018).

Blocked itemised measure outcomes are reported with repeated samples ANOVA in Chiu et al. (2008) and Fellows and Farah (2005); and repeated samples M/ANOVA in Bechara et al. (2000). Due to lack of individual data, M/ANOVA results for the studies noted above cannot be re-tested. However, for visual reference, the 5-block, 20-draw behaviour for the mean fraction of cards chosen from the good decks $\bar{f}_G^H$ for the original, re-shuffled, reversed, random, and SGT environments are presented.

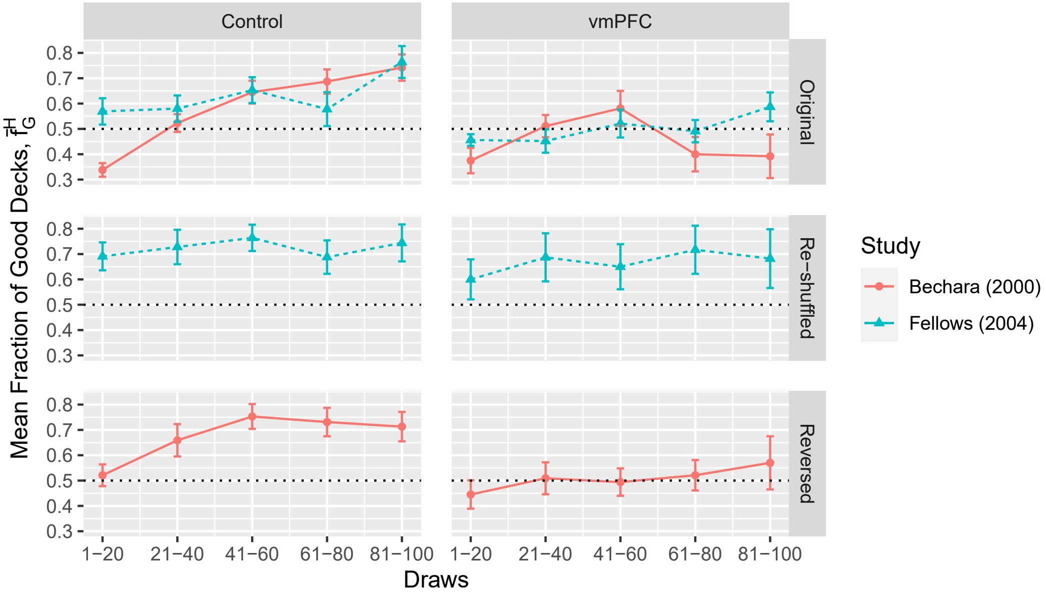


FIGURE 4.1: 5-block 20-draw mean fraction of good decks $\bar{f}_G^H$ results for IGT environments with both healthy control and vmPFC impaired participants. The dotted horizontal line at 0.5 indicates the pass point. In the original IGT environment, healthy subjects increase choices from the good decks, while vmPFC impaired participants fail to do so. However, both healthy and vmPFC impaired subjects pass the re-shuffled IGT variant at every 20-draw block. Further comments in the text.

Fig. 4.1 presents 5-block 20-draw mean fraction of cards chosen from the good decks $\bar{f}_G^H$ results for IGT environments with blocked data for both healthy and vmPFC impaired subjects. Error bars represent ±1 standard errors. Due to the unavailability of source data in Bechara et al. (2000) and Fellows and Farah (2005), numeric values are derived from pixel matching. The dotted horizontal line at 0.5 indicates the normative pass threshold. In the original IGT environment, healthy subjects increase choices from the good decks, while vmPFC impaired participants fail to do so. In the original IGT environment, from draw 40 onwards, healthy subjects exhibit a mean fraction of good decks $\bar{f}_G^H$ above the pass mark of 0.5. Both healthy and vmPFC impaired subjects pass the re-shuffled IGT variant at every 20-draw block. Healthy subjects pass the reversed IGT at every 20-draw block. In contrast vmPFC impaired subjects' performance at the reversed IGT straddles the pass mark; the vmPFC impaired subjects appear to have learned in the last draw block 81-100, but this cannot be determined conclusively. Block data is used as a visual aid to indicate shifts towards the mean fraction of good

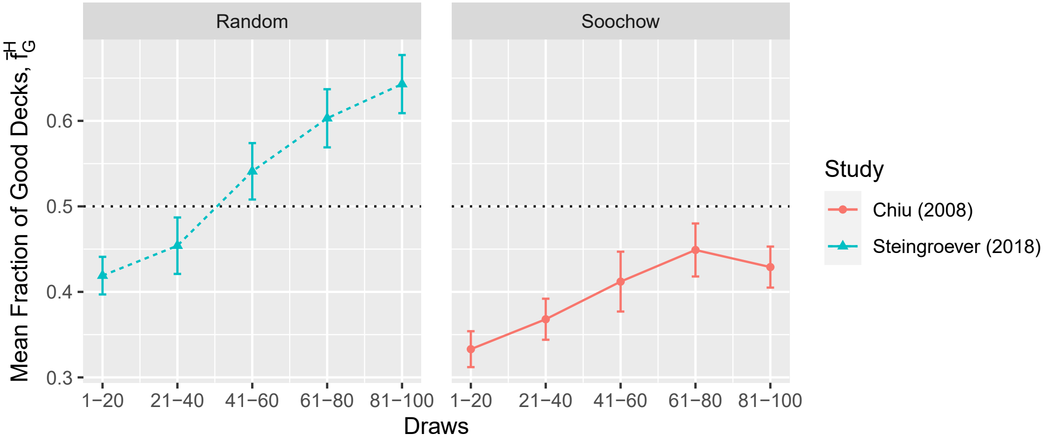


FIGURE 4.2: 5-block 20-draw mean fraction of good decks $\bar{f}_G^H$ results for IGT environments with only healthy control participants. The dotted horizontal line at 0.5 indicates the pass point. In the random IGT environment, healthy subjects increase choices from the good decks. However, healthy subjects fail the SGT variant at every 20-draw block. Further comments in the text.

decks over trial draws.

Fig. 4.2 presents 5-block 20-draw mean fraction of cards chosen from the good decks $\bar{f}_G^H$ results for IGT environments with blocked data for only healthy subjects. Error bars represent ±1 standard errors. For Steingroever et al. (2018), numerical values are derived from the longitudinal data set in Steingroever et al. (2015). Due to the unavailability of source data in Chiu et al. (2008), numeric values are derived from pixel matching. The dotted horizontal line at 0.5 indicates the normative pass threshold. In the random IGT environment, healthy subjects increase choices from the good decks, and achieve passes above 0.5 from draw 41 onwards. However, healthy subjects fail the SGT variant at every 20-draw block.

Blocked itemised measures were devised to help answer the exploitation versus exploration question, and to see if one can conclusively ascertain movement towards good decks over task duration. Based on Figures 4.1 and 4.2 with a block size of 20-draws, shifts can be observed towards the good decks among healthy participants for the original, reversed, and random IGT environments. However, at 20-block resolution the re-shuffled IGT environment does not display any clear shift towards the good decks for healthy and vmPFC impaired participants. The vmPFC impaired reversed IGT and normal SGT results appear to show in 20-draw blocks a pattern, which could be seen as a mild shift towards the good decks. These observed behavioural differences will be important in agent calibrations.

4.2.3 The Exploration Index

To further assess exploration versus exploitation, a measure referred to as the exploration index (EI) is defined. The proposed exploration index aims to quantify the exploration versus exploitation dilemma, and establishes a rating based on the number of available and actual choices made over a period of time. Given N choices over an assessment block of b periods, the index rates exploring over a uniform distribution at 100 (that is, at frequency $1/N$ for N choices), and never exploring at 0.

In the IGT, there are four choices but only one choice can be made per draw. Therefore, the minimum block size must be 4. However, in line with the results in section 4.2.2, exploration index results will be reported using a block value of $b = 20$. The exploration index is reported on a per-deck basis for the random IGT presented in Steingroever et al. (2018) from the longitudinal data set in Steingroever et al. (2015), and for the SGT reported in Chiu et al. (2008) from pixel match computations. A version of the exploration index is also reported for good versus bad decks for all of the IGT environments discussed in 4.2.2.

Given N total choices over b periods, with all choices available at each period and only one choice selected at any one period, the entropy based exploration index is defined as

$$\mathbf{EI} = 100\frac{\sum_{i=1}^{N} f\left(N_i/b\right)}{log(1/N)}, \quad f\left(N_i/b\right) = \begin{cases} N_i/b \cdot log\left(N_i/b\right), & \text{if } N_i > 0 \\ 0, & \text{otherwise} \end{cases} \tag{4.2}$$

where $b \geq N$, N_i is the number of selections of choice i, and N_i/b is the observed frequency of choice i over the sampling block b. Note that if $N_i/b = 1/N$ for all i, then $\mathbf{EI} = 100$, and this indicates full exploration. On the other hand, given that a choice selection for every period is required, $EI = 0$ indicates that for all periods, the same choice was selected; this means no exploration, which in turn implies full exploitation of some choice. Note the the EI does not distinguish between directed or random exploration.

Further when applied to the IGT, the exploration index does not indicate whether the subject or software agent is shifting towards the good decks.

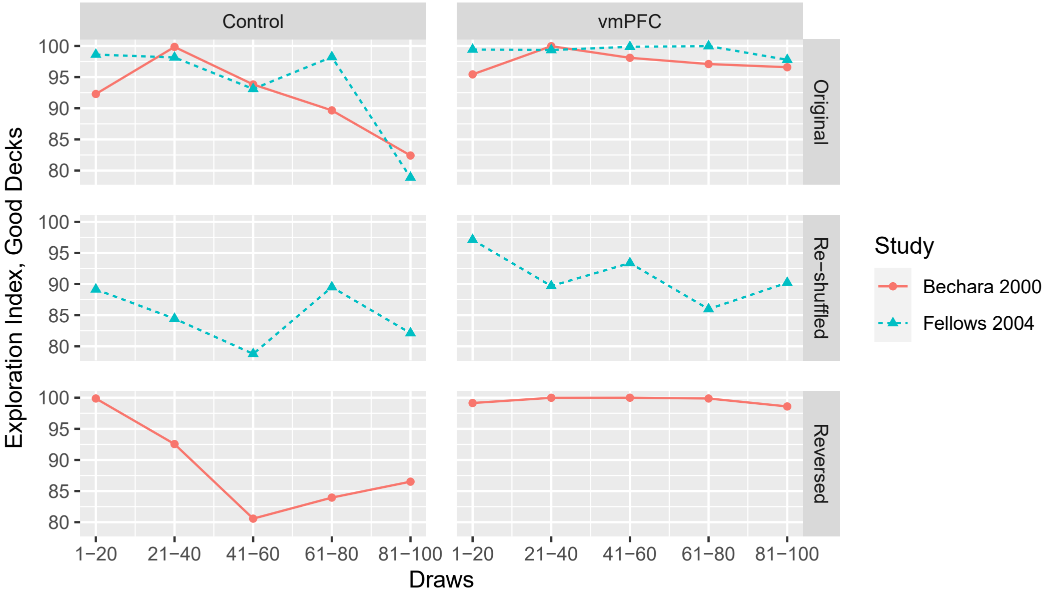


FIGURE 4.3: 5-block 20-draw good versus bad decks exploration index **EI** results for IGT environments with both healthy control and vmPFC impaired participants. For the original, re-shuffled, and reversed environments, healthy control subjects exhibit a reduction of the exploration index over draw blocks. Further comments in the text.

Also, in the sense of (4.2), exploitation does not imply having shifted towards the good decks, it only means decreased choice variance, which could have resulted from exploiting any of the good or bad decks. (4.2) purely indicates whether exploration is increasing or decreasing over time, and can be used in conjunction with the blocked fraction of good decks measure presented in section 4.2.2. Finally, (4.2) is descriptive of an exploration strategy implied by an observed outcome, but this implied exploration strategy may be different from the one employed by the actor. (4.2) is used to rank the implied exploration strategies of IGT results.

Fig. 4.3 shows 5-block 20-draw good versus bad decks exploration index **EI** results for IGT environments with both healthy control and vmPFC impaired participants. For the original, re-shuffled, and reversed environments, healthy control subjects exhibit a reduction of the exploration index over draw blocks. In contrast vmPFC impaired patients fail to reduce implied exploration for the original and reversed IGT environments, while lagging behind in the re-shuffled IGT environment.

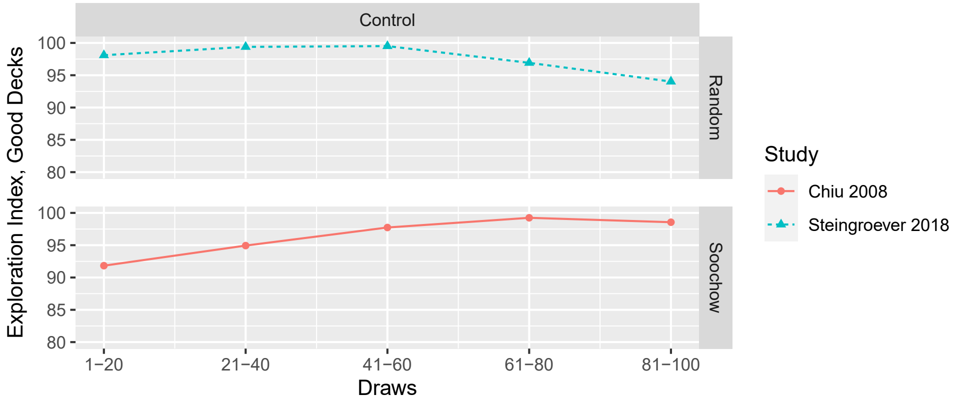


FIGURE 4.4: 5-block 20-draw good versus bad decks exploration index **EI** results for IGT and SGT environments with healthy control participants. For the random IGT environment, healthy control subjects exhibit a reduction of the exploration index over draw blocks. In the SGT (Soochow), healthy controls appear to show an increase in exploration.

Fig. 4.4 shows 5-block 20-draw good versus bad decks exploration index **EI** results for IGT and SGT environments with healthy control participants. For the random IGT environment, healthy control subjects exhibit a reduction of the exploration index over draw blocks. In the SGT (Soochow), healthy controls appear to show an increase in exploration.

In sum, exploration index **EI** results support the notion that exploration during the IGT decreases over draw blocks in healthy subjects, and remains unresponsive, or slow to respond in vmPFC impaired subjects. The SGT environment healthy subject outcomes provide the exception, where over draw blocks, exploration seems to increase. This could be because of the difficulties in sampling from rare event distributions, leading to a prolonged exploration phase.

Chapter 5

IGT Modelling with Reinforcement Learning and Exponential Learning Rate Decay: Introduction

This chapter proposes a (nonrational context) single-state Q-learning model with exponential learning rate decay for IGT modelling. The single state approach removes the need for tracking different states of the world, which for example, would need to be tracked if different rewards and costs obtained under different states. The IGT versions studied here do not posses such complex yield structures. Therefore stateful Q-learning can safely be simplified.

Additionally CSUD (Constrained Single Unconstrained Double SPSA) search is employed to discover Q-learning model hyper-parametrisations, which in software agents lead to behaviour consistent with that exhibited by human IGT participants.

Section 5.1 defines the single-state Q-learning model, which is used to solve the IGT in a nonrational learning context. Section 5.2 presents the reinforcement learning agent implementations with differing exploration strategies. The well-known Boltzmann and ε-Greedy RL agents are reviewed in sections 5.2.1 and 5.2.2 respectively. The lesser known value difference based exploration agent (Tokic, 2010) is introduced in section 5.2.3. Finally, section 5.2.4 presents, as far as is known, a novel ε-Greedy agent variant with exponential exploration rate decay.

Section 5.3 illustrates how the proposed CSUD search strategy can be employed as a hyper-parameter tuner used for discovering model hyper-parametrisations, leading to human IGT outcomes. Rather than using traditional statistical techniques, such as linear regression or maximum likelihood for determining the value of exponential learning rate decay and other hyper-parameters supported by a particular outcome data set, section 5.3 proposes a search strategy, implemented via CSUD, which is discussed in detail in chapter 12. The aim of the CSUD search strategy is to tune model hyper-parameters so as to produce agent outcomes conformant with performance matching criteria from Table 4.8. Based on the search considerations discussed in section 3.4, section 5.3 implements a search strategy where the loss scoring function includes a linear-quadratic loss form, chained (joint) search criteria, and a structure, which can be flexed to allow for standard error uncertainties.

5.1 Single state Q-learning with Exponential Learning Rate Decay

The rational n-armed bandit problem discussed in section 2.5.1 constitutes a good starting point. Accordingly, any IGT environment is modelled as a single state, four deck, environment with four actions. Here Q-learning is not implemented as initially proposed by Watkins (1989), where the current contribution to the Q-factors uses off-policy updating, so that in any state, the agent estimates action contribution values by choosing the best known action in that state. Instead, on-policy value function updates are applied as suggested by Sutton and Barto (2018, p. 32). While Watkins' approach approximates (weighted) present value, Sutton and Barto's approximation can be seen as estimating (weighted) yield.

Both approaches originate from the same stochastic approximation function class (Tsitsiklis, 1993). Given the single state, and given that yield risk (variance) is not considered, a deck with the highest yield also produces the highest present value. By not requiring an off-policy term, Sutton and Barto's approach reduces computational complexity. In line with Occam's razor, the less complex Sutton and Barto approach is chosen. Next terminology and the model are specified.

An *action* a consists of choosing a card from a particular deck. The net payoff for action a at iteration t, x_t^a, is the difference between the realized reward and fine. That is, $x_t^a = r_t^a - c_t^a$, where r_t^a and c_t^a denote rewards and costs obtained at iteration t with action a.

From any (software) agent's perspective, the net payoff stream $\{x_t^a\}$ is unknown. It is assumed that, the unknown net payoff streams are independently and identically distributed (i.i.d.) over time and decks. Consequently, net payoff streams are stationary and fulfil the requirements of a Markov process.

In the IGT, as discussed in section 4.1.1, the agent does not know when the task ends. Therefore, the agent must iteratively approximate a net-yield stream value function from the present until the game ends.

Given any action a, let $Q(a)$ be an unknown value function. Let $Q_t(a)$ denote the t^{th} iteration. Let $Q_t(a)$ represent, for action a, the value of the net payoff stream from iteration t onwards. Then the computational estimation problem is writen as

$$Q_t(a) = \alpha_t x_t^a + (1 - \alpha_t)\, \gamma Q_{t-1}(a) \tag{5.1a}$$

$$= \gamma Q_{t-1}(a) + \alpha_t \left(x_t^a - \gamma Q_{t-1}(a)\right) \tag{5.1b}$$

where the term $Q_{t-1}(a)$ *is used to forecast the future value* of action a, and γ represents the discount rate. This notational convention is used to compactly represent the agent's actual decision making sequence at iteration t, which consists of using the existing known Q-values $Q_{t-1}(a)$ to choose action a, and then observe outcome x_t^a, leading to Q-value updates.

(5.1a) represents the standard on-policy Q-learning form, where the learning rate, α_t, re-weights the relative contribution of the current net payoff to the (discounted) value function. (5.1b) shows the stochastic gradient form where $x_t^a - \gamma Q_{t-1}(a)$ is the gradient approximation and α_t is the stepsize.

As the learning rate α_t approaches 1, the last net payoff contributes increasingly more to the value function. (5.1a) states that when estimating the current value, the agent uses a mixture of the current payoff and $\gamma Q_{n-1}(a)$, its best, last-known approximation of the updated value function.

The parameter $\gamma \in [0, 1]$ is the discount rate, and, when less than 1, indicates a preference for current over future rewards. The discount rate is used to compute the present value of a yield stream. However, with a single state, estimating mean yields is sufficient and mathematically simpler. Furthermore in this section, it is assumed that the length of the card game, although unknown, is not long enough to create a preference for present rewards. Therefore here the discount γ rate is set as $\gamma = 1$.

5.1.1 Exponential Learning Rate Decay

Let $\{\alpha_t\}_1^\infty$ be a sequence of learning rates. Section 3.5 has indicated that exponentially decaying learning rates constitute a nonrational form of learning and cannot produce theoretical guarantees of Q-value convergence. An exponentially decaying learning rate sequence $\{\alpha_t\}$, could get sufficiently close to 0 prior to some final period T, and in that manner effectively curtail learning. However, this does not mean that exponentially decaying learning rates will not produce well-performing Q-value functions.

For review purposes, the exponential learning rate decay form (3.9b) is repeated below

$$\alpha_t = e^{-\lambda}\alpha_{t-1} \tag{3.9b}$$

where $\lambda \in [0, \infty)$ is the decay factor. Next set $\gamma = 1$. Then using (3.9b) to recursively expand (5.1a) yields

$$Q_t(a) = \prod_{k=1}^{t}\left(1 - e^{-\lambda(t-k)}\alpha_1\right) Q_0(a) + \alpha_1 \sum_{k=1}^{t} e^{-\lambda(k-1)} \prod_{m=1}^{t-k}\left(1 - e^{-\lambda(t-m)}\alpha_1\right) x_k^a \tag{5.2}$$

where $\alpha_1 \in (0, 1]$ is the initial learning rate and λ is the learning rate decay. After setting $Q_0(a) = 0$, only the second term of (5.2) remains.

With the first term in (5.2) set to 0, the effects of the initial learning rate α_1, and learning rate decay λ on Q-value evolution can be better understood. In the second term the initial learning rate α_1, is bounded above by 1, and the decay factor, λ is bounded below by 0. Therefore it can be seen that α_1 and λ influence $Q_t(a)$ in opposing directions. Everything else being equal, increasing the initial learning rate increases net yield value attribution, while increasing the decay factor reduces net yield value attribution. The recursively expanded form (5.2) suggests that, when $Q_0(a) = 0$, the decay factor λ, is the most dominant parameter.

5.2 Software agent implementations

In principle, software agents learning the IGT exhibit unconstrained maximising behaviour. At any iteration t, the agent picks the deck with the highest attributed value

$$Q_t^* = \max_a \; Q_t(a), \; a \in \text{IGT Decks} \tag{5.3}$$

However, this rule on its own is overly greedy, and suffers from a significant drawback. Once the maximization in (5.3) eliminates a deck, the agent would never re-draw from the eliminated deck, even if the eliminated deck would have later produced positive net yields.

Thus (5.3) on its own indicates sequential choice shortcomings similar to the "hot stove" effect discussed in Denrell and March, 2001. (5.3) however produces choice selections, which are very strict when compared with the "hot stove" effect; with (5.3) choices are eliminated not on a probabilisitc but on a permanent basis. Therefore a stochastic exploration rule must be introduced to mitigate the shortcomings of the overly greedy selection rule (5.3). That is, (5.3) must be augmented with an exploration rule to ensure continuing exploration of deck selection.

Note that software agents are not varied to exhibit any individual differences. In that sense, software agents in this study comprise representative agents. Any agent can be thought of as a projection filter solely assessing the uncertainty inherent in the IGT task. If the same agent hyperparametrisation produces substantially different outcomes on re-runs of a particular IGT task, then such differences will need to be explained.

5.2.1 The Boltzmann Agent

Implementations of Boltzmann exploration have already been presented in cognitive models in (2.1c), (2.2d), and (2.4f). The Boltzmann agent formulation here is closest to (2.4f). However, an iteration invariant temperature term τ is added

$$Q_{t,B}^*(a) = a \text{ with probability } \frac{exp\left(Q_t(a)/\tau\right)}{\sum_{i \in \{A,B,C,D\}} exp\left(Q_t^i/\tau\right)} \tag{5.4}$$

where $\tau > 0$. As $\tau \rightarrow 0$, (5.4) becomes increasingly greedy, and the action with the highest Q-value becomes more likely to be chosen. In contrast for high τ, the deck selection probability gap between high and low Q-values is decreased.

Boltzmann exploration constitutes an indirect and proportional approach. Provided Q-values are learned correctly, higher Q-valued actions are chosen more frequently, and this leads to reduction of suboptimal action sampling.

However, the requirement that Q-values must be learned correctly also forms the Achilles heel of Boltzmann exploration. It will be seen that in a nonrational context, for sufficiently high learning rate decay λ, Boltzmann agents cannot learn correct Q-values, and are subsequently unable to recover from suboptimal selection policies.

Interestingly, in a rational context, it can be shown that even a Boltzmann agent with decaying exploration may exhibit suboptimal sampling behaviour. Cesa-Bianchi et al. (2017) show that with monotonically decaying inverse temperature sequences $\{1/\tau_t\}$, Boltzmann exploration does not exhibit optimal regret behaviour. The authors derive a near optimal version of Boltzmann exploration, however only for cases when duration T is known a-priori. Cesa-Bianchi et al. (2017) report that with n-armed bandits, both Boltzmann and ε-Greedy exploration strategies tend to oversample from the suboptimal arms.

5.2.2 The ε-Greedy Agent

The ε-Greedy agent employs direct exploration and therefore, unlike the Boltzmann agent, is more resistant to incorrectly learned Q-values. Further the ε-Greedy agent is computationally simpler. In this work, a constant exploration variant is formulated

$$Q^*_{t,\varepsilon} = \begin{cases} Q^*_t, & \text{with probability } 1-\varepsilon, \\ a \in \{A,B,C,D\}, & \text{with probability } \varepsilon \text{, choose } a \text{ randomly} \end{cases} \tag{5.5}$$

where ε is typically a small number and indicates the probability of exploration.

Note that regardless of learned Q-values, exploration always occurs with probability ε, and therefore suboptimal decks continue to be sampled, even after they have been discovered to be suboptimal.

Note that in a rational context, (5.5) is always suboptimal in terms of regret. However, in a nonrational context with learning rate decay λ, (5.5) provides stronger potential for recovery from incorrectly learned Q-values.

One of the considerations in the IGT is whether deck exploration decreases over time. To accommodate such considerations, two further software agents with decreasing exploration are formulated. These are the adaptive (Tokic, 2010), and exponentially decaying exploration ε-Greedy agents.

5.2.3 The Adaptive ε-Greedy Agent

This agent is discussed in detail in Tokic (2010), where it is called the ε-Greedy VDBE-Boltzmann agent, with VDBE standing for *value-difference based exploration*. This agent employs a mixture of the Boltzmann agent and ε-Greedy agent features. For simplicity, in this work, this agent is called the adaptive ε-Greedy agent. Salient agent features are presented below, where the author's original notation has been slightly altered.

$$Q_{t,\tau,\beta} = \begin{cases} Q_t^*, & \text{with probability } 1 - \varepsilon_t, \\ a \in \{A, B, C, D\}, & \text{with probability } \varepsilon_t \text{, choose } a \text{ randomly} \end{cases} \tag{5.6a}$$

$$f(a,\tau) = \frac{1 - e^{-b/\tau}}{1 + e^{-b/\tau}}, \quad b = |\alpha_t \left(x_t^a - Q_{t-1}(a)\right)| \tag{5.6b}$$

$$\varepsilon_{t+1} = \beta f(a,\tau) + (1-\beta)\varepsilon_t \tag{5.6c}$$

where $\tau > 0$ is the temperature, $0 \leq \beta < 1$ is the exploration adaptation parameter, and $|\alpha_t \left(x_t^a - Q_{t-1}(a)\right)|$ is the absolute value of the temporal difference error between the actual net yield and forecasted Q-value.

(5.6b) introduces a scaled system "shock" such that $0 \leq f(a,\tau) < 1$. This shock increases towards 1 with larger temporal difference errors. (5.6c) is the linear combination of this scaled shock with existing exploration ε_t. If the scaled shock is consistently above current exploration, then exploration increases; otherwise exploration decreases. In a nonrational context, (5.6) remains susceptible to getting stuck after learning incorrect Q-values. This is because with high learning rate decay, the exploration update mechanism cannot generate accurate exploration updates.

5.2.4 The Decaying ε-Greedy Agent

The decaying exploration ε-Greedy agent constitutes an example of a fully nonrational (heuristic) agent. In this agent type, exploration begins at a set value, and then decays exponentially with each iteration. This agent is used to model decision making approaches, where after some initial learning period, a decision must be produced regardless of the consequences. This agent is formulated as

$$Q_{t,\varepsilon_1,\nu} = \begin{cases} Q_t^*, & \text{with probability } 1 - \varepsilon_t, \\ a \in \{A, B, C, D\}, & \text{with probability } \varepsilon_t \text{, choose } a \text{ randomly} \end{cases} \tag{5.7a}$$

$$\varepsilon_{t+1} = e^{-\nu}\varepsilon_t \tag{5.7b}$$

where ν is the exploration decay factor, $\nu \geq 0$. Note that initial exploration ε_1 must be given. Furthermore when $\nu = 0$, (5.7) reduces to (5.5).

5.3 Search strategy implementations

Section 5.2 introduces (agent) implementations of the inner selection function $R(\cdot)$ in (3.1a) for solving IGT environments with single state Q-learning subject to given model hyper-parameters. Here an implementation is discussed for solving the companion equation (3.1b), searching for the best hyper-parameter settings fulfilling specified search criteria. Search strategy and criteria specification are presented, which are similar to the quadratic loss example (3.5) provided in section 3.4.

Table 5.1 presents a summary of agent specific tunable hyper-parameters. Note that the Boltzmann and ε-Greedy agents each have three tunable; and the adaptive and decaying ε-Greedy agents each have four tunable hyper-parameters. All agents require the initial learning rate α_1 and learning rate decay λ, but differ in exploration implementation. The Boltzmann and ε-Greedy agents each have one exploration hyper-parameter: τ defined in (5.4) and ε defined in (5.5) respectively. In the adaptive ε-Greedy agent, τ and β modulate exploration and are defined in (5.6b) and (5.6c). Finally in the decaying ε-Greedy agent initial exploration ε_1 and exploration decay ν defined in (5.7b) influence exploration.

5.3.1 Search Criteria Specification

Section 3.4 discusses the general approach. Within the IGT context, the fundamental building block of search criteria is the mean fraction of good decks

Agent	Boltzmann	ε-Greedy	Adaptive ε-Greedy	Decaying ε-Greedy
Initial Learning Rate α_1	✓	✓	✓	✓
Learning Rate Decay λ	✓	✓	✓	✓
Exploration	✓ (τ)	✓ (ε)	✓ (τ)	✓ (ε_1)
Exploration Adaptation β	–	–	✓	–
Exploration Decay ν	–	–	–	✓

TABLE 5.1
Tunable hyper-parameters by software agent.

$\bar{f}_G$ measure and its standard error, both introduced in section 4.2.1 along with human performance results, denoted as $\bar{f}_G^H$, presented in tables 4.3 to 4.8. The mean fraction of good decks $\bar{f}_G^H$ resides in the unit interval, and this makes it easier to discern to what extent the search strategy discovered hyper-parametrisations are able to achieve the IGT normative pass criterion regardless of the respective cumulative net payoff magnitude differences among IGT variations.

Searches seek hyper-parameter combinations capable of simultaneously matching multiple human IGT environment outcomes. In particular, it is hypothesized that learning rate decay λ is a key parameter for matching normal versus vmPFC impaired human IGT performance. This hypothesis is motivated by the observation that in the original IGT with the misleading initial bad deck card sequences, vmPFC impaired human subjects: (a) cannot recover from having learned an incorrect policy, but (b) perform at par with normal human subjects in the re-shuffled IGT, where the initial misleading sequence has been moved to the end. In the context of (5.1a) and (5.2), such a result can be approximated by a fast decaying learning rate, which leads to a very short learning horizon.

Θ is used to indicate the hyper-parameter vector. For example, using the Boltzmann agent, $\Theta_N = (\alpha_1, \lambda_N, \tau)$, and $\Theta_{vmPFC} = (\alpha_1, \lambda_{vmPFC}, \tau)$ are specified to define the hyper-parametrisation of normal, and vmPFC impaired Boltzmann agents. Next, for given Boltzmann agent hyper-parameter vectors $\hat{\Theta}_N$ and $\hat{\Theta}_{vmPFC}$, the mean fraction of good decks obtained by the normal and vmPFC impaired agents are denoted as $\bar{f}_{G,\hat{\Theta}_N}$ and $\bar{f}_{G,\hat{\Theta}_{vmPFC}}$ respectively.

For any IGT environment, using human performance outcomes as targets, search criteria can be represented as squared loss target deviations

$$\left(\bar{f}_{G,\hat{\Theta}_N} - \bar{f}^H_{G,N}\right)^2, \qquad \left(\bar{f}_{G,\hat{\Theta}_{vmPFC}} - \bar{f}^H_{G,vmPFC}\right)^2 \tag{5.8}$$

where mean fraction of good decks $\bar{f}_G$ is a scalar, N indicates normal, $vmPFC$ denotes vmPFC impaired behaviour, and H indicates corresponding human outcome targets. In sum, the central loss measure indicated in (5.8) consists of the squared loss of the difference between simulated agent and target human mean fraction of good deck outcomes for normal (control) and vmPFC impaired behaviours.

5.3.2 Loss Function Specification

With the vocabulary defined in (5.8), a linear-quadratic loss function is specified with standard error penalisation and with multiple, chained search criteria. The loss function presented below is used to generate the results for chapter 6. Chapters 7 and 8 use similarly constructed loss functions.

Based on the availability of targeting data, one wishes to search for parameter combinations $\hat{\Theta}_N$ and $\hat{\Theta}_{vmPFC}$, which will produce target matches as follows: in the original and re-shuffled IGT environments for normal and vmPFC impaired human subjects, and in the random IGT environment for normal human subjects.

For example, let $\Theta = (\alpha_1, \lambda_N, \lambda_{vmPFC}, \dots)$ be the hyper-parameters. Let $Env = \{Or, Re, Rn\}$ denote the original, re-shuffled, and random IGT environments respectively. Let $\mathbb{B} = \{N, vmPFC\}$ denote the set of behaviours. For each $j \in Env$ and $k \in \mathbb{B}$, let $\bar{f}^{H,j}_{G,k}$, $f^{H,j,+}_{G,k}$ and $f^{H,j,-}_{G,k}$ be the mean, maximum and minimum fraction of good decks respectively of the corresponding human match ranges available from table 4.8. Then, for example for the original IGT environment with normal participants, from table 4.8 the following values are obtained: $\bar{f}^{H,Or}_{G,N} = 0.64$, $f^{H,Or,+}_{G,N} = 0.69$, and $f^{H,Or,-}_{G,N} = 0.59$.

As illustrated above, the human match targets used in this work consist of average IGT outcome ranges. Hence software agents are targeted to match not individual human IGT, but averaged human IGT outcomes. This decision is in part driven by the lack of available data, and in part by the question of whether and under what circumstances average, that is, thighly clustered behaviours may obtain.

More generally, the loss function is formulated as

$$Y(\Theta) = \sum_{\substack{j\in\\ Env}} \sum_{\substack{k\in\\ Env\cap\mathbb{B}}} A_{jk} + D_{jk},$$

$$D_{jk} = \begin{cases} \left(\bar{f}^{j}_{G,\Theta_k} - \bar{f}^{H,j}_{G,k}\right)^2, & \text{if } f^{H,j,-}_{G,k} \leq \bar{f}^{j}_{G,\Theta_k} \leq f^{H,j,+}_{G,k} \\ -\bar{f}^{j}_{G,\Theta_k} + B_{jk}, & \text{if } \bar{f}^{j}_{G,\Theta_k} < f^{H,j,-}_{G,k} \\ \bar{f}^{j}_{G,\Theta_k} + C_{jk}, & \text{if } \bar{f}^{j}_{G,\Theta_k} > f^{H,j,+}_{G,k} \end{cases} \tag{5.9}$$

where A_{jk} is a regularisation term accounting for the standard error of simulated fraction of good deck results, B_{jk}, C_{jk} are intercept terms, and $k \in Env \cap \mathbb{B}$ is shorthand for the behaviours available to the IGT environment. For example, according to table 4.8, in the original IGT environment, both normal and impaired behaviours are available, whereas in the random IGT environment only the normal behaviour is available.

The terms A_{jk}, B_{jk}, C_{jk} expand as

$$A_{jk} = \begin{cases} SE^{+} - f^{H,j,+}_{G,k} & \text{if } SE^{+} > f^{H,j,+}_{G,k} \quad SE^{+} = \bar{f}^{j}_{G,\Theta_k} + 2 * SE(\bar{f}^{j}_{G,\Theta_k}) \\ f^{H,j,-}_{G,k} - SE^{-} & \text{if } SE^{-} < f^{H,j,-}_{G,k} \quad SE^{-} = \bar{f}^{j}_{G,\Theta_k} - 2 * SE(\bar{f}^{j}_{G,\Theta_k}) \end{cases} \tag{5.10a}$$

$$B_{jk} = \left(f^{H,j,-}_{G,k} - \bar{f}^{H,j}_{G,k}\right)^2 + f^{H,j,-}_{G,k} \tag{5.10b}$$

$$C_{jk} = \left(f^{H,j,+}_{G,k} - \bar{f}^{H,j}_{G,k}\right)^2 - f^{H,j,+}_{G,k} \tag{5.10c}$$

Note that (5.10a) constitutes a linear penalty term equal to the positive difference between the human mean fraction of good decks outcome bound and the agent mean fraction of good decks performance bound, where the agent mean fraction of good decks performance bound is calculated as the agent mean fraction of good decks plus twice its standard error.

Strictly speaking (5.9) has discontinuities at $f^{H,j,-}_{G,k}$ and $f^{H,j,+}_{G,k}$, at which points the loss function switches from quadratic to linear form. While these simple discontinuities can be managed analytically, it is not done so here. Also in the strict sense, as seen later in chapter 12 that these discontinuities would violate the rational convergence derivation for CSUD. However, the loss construction remains such that the computational aspects of CSUD are not affected, and therefore, the CSUD algorithm is able to computationally manage these discontinuities.

Further note that (5.9) is simple enough that it can be solved via computation of analytical derivatives, with the discontinuities being handled. That is, (5.9) can be adapted to standard stochastic gradient descent, and this would be computationally cheaper. In this work however, (5.9) is only a search strategy driving the second, tuning stage for the IGT choice problem encapsulated in the Q-learning agents (5.4) to (5.7). Consequently, the additional randomness injected via hyper-parameter Θ perturbations in the CSUD search strategy can be helpful in achieving better exploration of the underlying problem.

From a rational perspective, Lorraine et al. (2020) show that within the context of tuning artificial neural networks with weights w and hyper-parameters Θ, the analytical derivative version of the approach presented here can be proven to produce optimal tuning results via the implicit function theorem. Their work is closest to what is generally proposed in (3.1); more specifically here via the IGT Q-learning agents (5.4)-(5.7) and the search strategy (5.9). Other than the use of gradient approximation via CSUD, this proposal also differs from that of Bengio (2000) and Lorraine et al. (2020), in that a flexible, tuning targeting mid-layer is introduced. That is to say, while here the mean fraction of good decks is used for tuning, one could also substitute another performance measure without changing the general flow of the approach and implementation proposed in this work.

Chapter 6

The Original, Re-Shuffled, and Random IGT Environments with Simple Reinforcement Learning Modelling via CSUD

This chapter uses the simple single state reinforcement model (5.1) with discount rate $\gamma = 1$. With exponential learning rate decay, and when the initial Q-values are initialised to 0, then (5.1) can be unrolled to solely consist of the second term of (5.2)

$$Q_t(a) = \alpha_1 \sum_{k=1}^{t} e^{-\lambda(k-1)} \prod_{m=1}^{t-k} \left(1 - e^{-\lambda(t-m)}\alpha_1\right) x_k^a \tag{6.1}$$

where t is an iteration index, a is an action, x_k^a denotes a net yield entity, $\alpha_1 \in (0, 1]$ is the initial learning rate and λ is the learning rate decay. Note that for given net yield entities, Q-valuation at iteration t is solely influenced by the initial learning rate α_1 and learning rate decay λ. With exploration and any additional agent specific hyper-parameters, Q-valuation forms the basis of agent learning and decision making.

In any given iteration, the nonrational CSUD search strategy sets problem hyper-parameters, which then produce an IGT mean fraction of good decks result $\bar{f}_G$. The mean fraction of good decks result in turn is scored by the CSUD loss function, which evaluates deviations from human outcome ranges. The CSUD loss scores are used to update the problem hyper-parameters for the next iteration.

It will be seen that learning rate decay λ proves to be the critical parameter for mimicking vmPFC impaired behaviour.

6.1 General Methodology

Simulation based methodologies are employed for discovering and verifying software agent results, which match human IGT behaviour in terms of mean fraction of cards chosen from the good decks $\bar{f}_G$. The discovery and verification process consists of three stages.

In the first stage, the CSUD search strategy is used to discover reinforcement learning hyper-parameters values, which produce human performance match candidates. In the second stage, a local grid search is constructed around the CSUD discovered hyper-parameter values. This grid search helps to establish performance topology in a localised neighbourhood. Finally, IGT results are re-simulated using the discovered hyper-parameter candidates and a percentage measure for consistently replicating human performance outcomes is derived. Additionally, non-parametric multiple ANOVA (np-M/ANOVA) tests are performed for the effects of learning rate decay across different IGT environments.

CSUD is an iterative search algorithm with asymptotic convergence. Such algorithms are typically run subject to fixed search iterations or to a pre-determined change in the loss threshold. Here, a fixed iteration budget is used. A fixed iteration budget is in line with the approach of de-emphasizing infinity. In hyper-parameter tuning, when the search iteration budget is fixed, at the end of the iterations, either the result of the last iteration, or the result associated with the lowest loss can be chosen (Liaw et al., 2018). Here the *lowest* loss choice from among the *best* match selection is chosen. In a search for matches to normal and vmPFC impaired human behaviour for the original, re-shuffled, and random IGT environments, the best match would consist of matching five human outcome zones: normal original, re-shuffled, and random; and vmPFC impaired original and re-shuffled. Compared to the maximum likelihood fitting and verification process for example as discussed in Wilson and Collins, 2019, this final stage takes the place of parameter recovery, with the difference that re-simulated outcomes are directly used to perform statistical tests.

Agent	Boltzmann	ε-Greedy	Adaptive ε-Greedy	Decaying ε-Greedy
Hyper-parameter				
Initial learning rate α_1	0.01 - 0.99	0.01 - 0.999	0.01 - 1.0	0.05 - 0.99
Normal learning rate decay λ_N	0.0001 - 0.22	0.03 - 0.30	0.03 - 0.13	0.03 - 0.17
vmPFC impaired learning rate decay λ_{vmPFC}	0.22 - 1.2	0.25 - 1.2	0.12 - 1.2	0.2 - 1.2
Temperature τ	0.5 - 500		0.01 - 50	
Exploration ε		0.05 - 0.70	*1.0*[ac]	0.5 - 1.0[a]
Exploration adaptation β			*0.25*[bc]	
Exploration decay ν				0.002 - 0.02
CSUD Iterations	1000	1000	1000	1000
Gradient Samples	5	1	1	5

IGT length	Q-learning samples
100	750

[a] Exploration ε refers to initial exploration ε_1.

[b] Exploration adaptation β value set as recommended in Tokic (2010) to $1/actions$.

[c] These parameters are fixed and do not vary.

TABLE 6.1
Search Methodology. Joint original, re-shuffled, random IGT hyper-parameter CSUD search criteria.

6.2 Joint Search of the Original, Re-shuffled, and Random IGT environments

The original, re-shuffled, and random IGT environments are conceptually closely related. All three environments use the same payoff sheet with variations arising from how cards are sequenced within each deck. The original IGT discussed in 4.1.1 uses in the bad decks, initially misleading card sequences making the bad decks look good. The test subject is then expected to discover the genuinely good decks producing on average positive net yields. The re-shuffled IGT presented in 4.1.2 re-orders the original IGT deck cards so that the bad decks are immediately identifiable. The random

IGT discussed in 4.1.3 simply randomises (without replacement) the original IGT decks, thereby eliminating any specific card sequencing effects.

The human outcome differences across these three IGT environments originate from subject health status and card sequencing attributes. For that reason, a joint search across all three environments is conducted, subject to available human comparison data to look for hyper-parameter combinations, which will produce matching software agent performance.

Table 6.1 summarises software agent CSUD search hyper-parameter constraints and attributes. Hyper-parameters, which *do not* vary are italicised. In summary, the fixed hyper-parameters consist of initial exploration and exploration adaptation for the adaptive ε-Greedy agent. In general, broad parameter search ranges are used for the initial learning rate and exploration, while constrained ranges are applied for normal and vmPFC impaired learning rate decay. One might argue that by limiting learning rate decay ranges, prior information is being injected, and this guides the search towards a desired result. Indeed, this is precisely what the CSUD search strategy aims to do, that is, to see whether a specific search hypothesis produces any results. By construction, CSUD search loss is minimised to the extent that search criteria are fulfilled. It is in this sense that the CSUD search strategy can be seen as a contraction of the grid search space.

Because CSUD is a stochastic, gradient driven search technique, sometimes a single gradient evaluation is not sufficient to produce a reliable gradient estimate. Under such circumstances, for the same set of hyper-parameter values, multiple gradient samples may be obtained and then averaged (Spall, 1992). The 'Gradient Samples' entry in Table 6.1 indicates if any gradient sampling was employed.

6.3 ε-Greedy Agent Results

In this section, a very detailed report of the ε-Greedy agent IGT simulation results will be provided. This detail is provided as an introduction to the simulation outcome analysis tools.

Simulations produce a high volume of data. When there is a high volume of data, it is generally preferred to use consolidated summary statistics. Here however, outcomes with specific behavioural implications are of

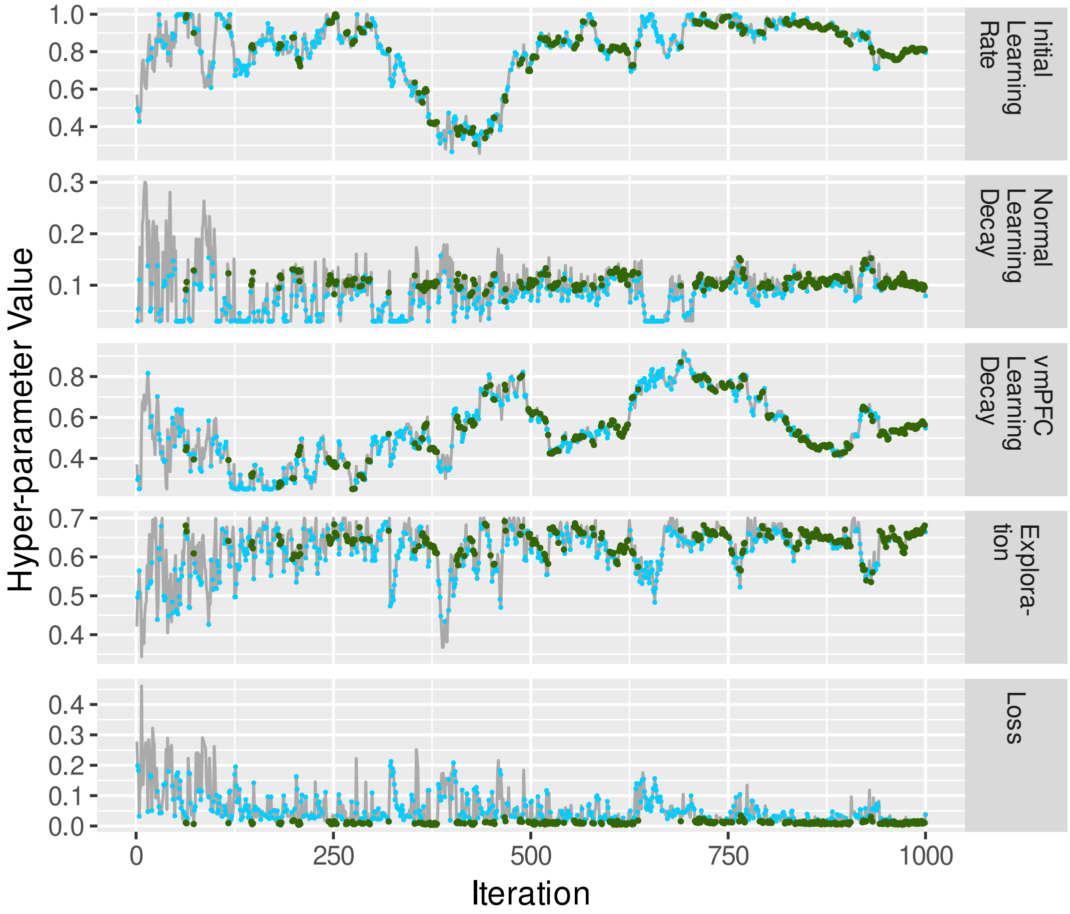


FIGURE 6.1: ε-Greedy agent CSUD iterations. Green points indicate (α_1, λ_N, λ_{vmPFC}, ε) hyper-parameter tuples, which produce search matches for all available IGT environment and behaviour combinations.

interest. Therefore, a good balance between summary statistics and data diversity becomes important. The primary summary statistic used is the mean fraction of cards chosen from the good decks $\bar{f}_G$. Data diversity is shared via visual plots, which can incorporate large amounts of numeric information.

Table 6.2 and Fig. 6.1 present CSUD search results in tabular and graphic forms respectively. This is then followed by 2D and 3D grid search verification plots in Fig. 6.2 and Fig. 6.3 respectively, reflecting grid search results in a neighbourhood expanded about the selected CSUD hyper-parameter values. Finally, a simple consistency analysis is presented. In general, the ε-Greedy agent provides excellent matching for human $\bar{f}_G^H$ outcomes. However, the selected exploration hyper-parameter value appears very high and counter-intuitive, and is further considered in section 6.3.1, where ε-Greedy

	Minimum Loss	Range	Mean	Median	Standard Error
Loss	0.004	0.00404- 0.0346	0.0105	0.0101	0.00025
Initial learning rate α_1	0.417	0.307 - 0.999	0.817	0.843	0.0095
Normal learning rate decay λ_N	0.104	0.0685 - 0.153	0.108	0.105	0.00072
vmPFC impaired learning rate decay λ_{vmPFC}	0.449	0.25 - 0.870	0.540	0.530	0.0075
Exploration ε	0.627	0.535 - 0.691	0.642	0.647	0.0016
Matched environments	For normal human behaviour: original, re-shuffled, random. For vmPFC impaired human behaviour: original, re-shuffled.				
Match count	290 of 1000 iterations				

TABLE 6.2
ε-Greedy agent CSUD search matches after 1000 iterations. The highlighted minimum loss column shows selected agent hyper-parameters. Light gray indicates minimum loss and the associated initial learning rate α_1. Dark-gray, mid-gray, and light blue indicate minimum loss associated normal learning rate decay λ_N, vmPFC impaired learning rate decay λ_{vmPFC}, and exploration ε respectively.

agent results are discussed.

Figure 6.1 shows ε-Greedy agent CSUD iteration progression. Green points indicate initial learning rate, normal learning decay, vmPFC impaired learning decay, and exploration, that is $(\alpha_1, \lambda_N, \lambda_{vmPFC}, \varepsilon)$ hyper-parameter tuples, which produce agent performance matching normal human mean fraction of good decks $\bar{f}_G$ outcomes for the original, re-shuffled, and random; and vmPFC impaired human mean fraction of good decks $\bar{f}_G$ outcomes for the original and re-shuffled IGT environments. The match ranges can be found in Table 4.8. It is seen that as the iterations progress, the number of matches increases.

Table 6.2 depicts the ε-Greedy agent CSUD search matches after 1000 iterations. The search produces 290 hyper-parameter value sets at which normal human behaviour outcomes are matched for the original, re-shuffled, and random decks, and vmPFC impaired human behaviour outcomes are matched for the original and re-shuffled decks. In the coloured ‘Minimum Loss’ column, Table 6.2 shows that the minimum loss hyper-parameter initial learning rate, normal learning decay, vmPFC impaired learning decay, and exploration $(\alpha_1, \lambda_N, \lambda_{vmPFC}, \varepsilon)$ tuple discovered by the search is (0.417, 0.104, 0.449, 0.627) respectively. It is observed that in the ε-Greedy agent Q-learning model, the variable leading to a switch from normal versus vmPFC impaired type human IGT outcome behaviour is the learning rate decay. Normal human behaviour is matched when learning decay is low at λ_N = 0.104, while vmPFC impaired human behaviour is matched when learning decay is high at λ_N = 0.449.

Table 6.2 shows that the initial learning rate α_1 varies from 0.307 to 0.999. That is, behavioural matches obtain within a large range, indicating that for purposes of pivoting between normal and vmPFC impaired behaviour, the initial learning rate α_1 is not a determining hyper-parameter. Exponential learning rate decay leads to proportional per period decay regardless of the initial learning rate. Therefore it is believed that model design with exponential learning rate decay is responsible for the low influence of the initial learning rate.

Table 6.2 reveals that an exploration range from 0.535 to 0.691 is associated with matched normal and vmPFC impaired, human original, re-shuffled, and random IGT outcomes. That is, in order for the ε-Greedy agent to achieve matched human outcomes, exploration needs to be very high; at least at 0.535 (53.5%), and at the minimum loss selection at 0.627 (62.7%). Usual experiment design would set this exploration hyper-parameter between 0.01 and 0.15. It is concluded that, compared to normal and vmPFC impaired subjects, the ε-Greedy agent may obtain in the original, re-shuffled, and random IGT environments, results superior to those achieved by human subjects. Indeed, this conjecture is verified in the below grid search plots. The question of why a human decision maker may be using exceptionally high exploration is discussed in section 6.3.1 in relation to the No Free Lunch theorems (Wolpert & Macready, 1997) and directed versus random exploration (Wilson et al., 2014).

Next CSUD verification grid searches are presented. Table 6.3 reports the localised grid search specification used to verify ε-Greedy agent CSUD results. Initial tests indicated that for the initial learning rate and exploration, a small grid of four points was sufficient. The learning (rate) decay grid generally consists of 20 points, and is constructed to include λ_N = 0.014 and λ_{vmPFC} = 0.449, both of which come from the minimum loss CSUD matches. Appendix C provides the construction method of the learning decay grid.

Fig. 6.2 and 6.3 present 2D and 3D views of the selected CSUD search hyper-parameters in the context of a neighbourhood grid search. As the 2D

Hyper-parameter	Grid Points
Initial learning rate α_1	0.01, 0.417, 0.6, 0.999
Learning rate decay λ[a]	0.104, 0.449
Exploration ε	0.1, 0.535, 0.627, 0.691
IGT length	Q-learning samples
100	750

[a] The learning rate decay grid is constructed around the CSUD reported values. Appendix C provides the construction method.

TABLE 6.3
ε-Greedy agent hyper-parameter grid search criteria for joint original, re-shuffled, and random IGT.

local search plots in Fig. 6.2 below show, the lower learning rate decay of $\lambda_N = 0.104$ implies a 10% per iteration decrease in the learning rate, while the higher learning rate decay of $\lambda_{vmPFC} = 0.449$ implies a 36% per iteration decrease in the learning rate. The higher learning rate decay is sufficiently high so that in the original IGT environment, once the initial misleading sequence of 8 draws is completed, the learning rate has decreased so much that new information no longer accurately updates Q-value accruals. In turn, this inaccurate update of Q-values leads to reproduction of vmPFC impaired original IGT human behaviour.

Fig. 6.2 depicts 2D contour lines which show the effect of learning rate decay λ and exploration ε on the mean fraction cards chosen from the good decks, $\bar{f}_G$. The dark and light gray zones indicate normal and vmPFC impaired human outcome match ranges for the original, re-shuffled, and random IGT environments. For the random IGT environment, only normal human outcome match ranges are available. Match range derivation is explained in Table 4.8. The lower x-axis indicates learning rate decay. Using the conversion formula $(1 - e^{-\lambda}) * 100$, the upper x-axis translates the lower x-axis learning rate decay value into a per period learning rate decay percentage, a measure that makes more intuitive sense.

In Fig. 6.2, at the CSUD minimum loss initial learning rate $\alpha_1 = 0.417$, when exploration $\varepsilon = 0.627$, and with normal learning decay $\lambda_N = 0.104$ and vmPFC impaired learning decay $\lambda_{vmPFC} = 0.449$, the ε-Greedy agent reproduces human IGT outcomes in the original, re-shuffled, and random environments. These results are indicated by the solid black contours. That

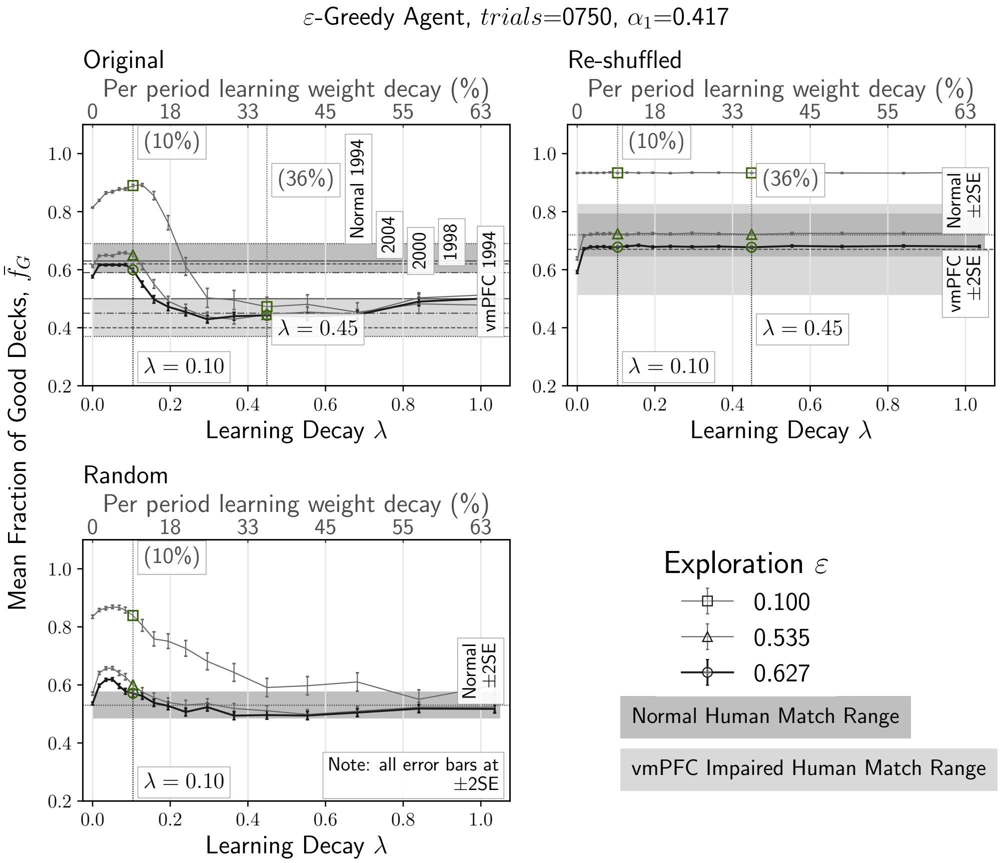


FIGURE 6.2: ε-Greedy agent 2D contours showing learning decay λ and exploration ε effects. The dark and light gray zones indicate normal and vmPFC impaired human outcome match ranges for the original, re-shuffled, and random IGT environments. Learning decay variation reproduces human IGT outcomes, while exploration variation is responsible for matching human IGT outcome performance ranges.

is, the normal configured agent achieves a normative pass in the original, re-shuffled, and random IGT environments; while the vmPFC impaired configured agent fails the original but passes the re-shuffled IGT environment. In general, as learning decay increases, agent performance degrades and becomes conformant with vmPFC impaired performance.

Also note that at $\varepsilon = 0.627$, CSUD minimum loss exploration is very high. In a rational context such a high level of exploration could be interpreted as poor model fit. As rational models are typically geared towards the exploitation of a central tendency, however, in the present nonrational context, the interpretation of high exploration is not so straightforward,. Here results are presented, and section 6.3.1 will then interpret these results in a nonrational context in light of reported hyper-parameter values and their interactions.

The 2D grid search verification contour plots show that multiple alternative solutions are available, which satisfy search criteria. This is in line with the (290) multiple matches, which were found in the CSUD search. However, the surprising observation from the 2D grid search verification plots is that at exploration $\varepsilon = 0.100$, the ε-Greedy agent produces mean fraction of good deck $\bar{f}_G$ values, which are higher than 85%, that is much superior to the results produced by human IGT participants. Indeed, the agent can only reproduce human IGT outcomes at a very high exploration rate in the range of $0.535 \leq \varepsilon \leq 0.691$, with the minimum loss exploration rate being $\varepsilon = 0.627$. The discussion in section 6.3.1 however, will make a case for high exploration as being indicative of a robust search strategy.

For the ε-Greedy agent, Fig. 6.3 depicts 3D contour plots, which in addition to learning decay λ and exploration ε, show the effect of the initial learning rate α_1. The minimum loss CSUD solution is annotated with $\alpha_1 = 0.417$, $\varepsilon = 0.627$, ♦ : $\lambda_N = 0.104$ and ▼ : $\lambda_{vmPFC} = 0.449$. Both learning decay λ and exploration ε effects across different initial learning rate values α_1 retain the characteristics discussed in Fig. 6.2.

For all IGT environments, Fig. 6.3 shows that the initial learning rate α_1 has very little effect on the mean fraction of cards chosen from the good decks $\bar{f}_G$. As is observed, in the direction of the initial learning rate α_1 axis, the 3D plot surfaces are horizontal with respect to the fraction of good decks chosen, indicating very little influence. However, a notable initial learning rate α_1 effect occurs with very low learning decay λ and high initial learning rate α_1, leading to the triangular-shaped, draped areas visible in the back of each IGT environment plot.

These triangular shaped areas summarise the technical difficulties that occur in iterative learning at the beginning of the learning process, when the learning rate is very high, leading to a strong contribution towards Q-values and when the agent is initially learning incorrect responses. In such a scenario, a high learning rate with low learning decay leads to large incorrect contributions to the Q-values. Under such conditions, reducing the initial learning rate or increasing learning rate decay can lead to performance gains in terms of the mean fraction of cards chosen from the good decks $\bar{f}_G$.

Note that the ε-Greedy agent CSUD matches for human behaviour occur with learning decay matches of ♦ : $\lambda_N = 0.104$ and ▼ : $\lambda_{vmPFC} =$

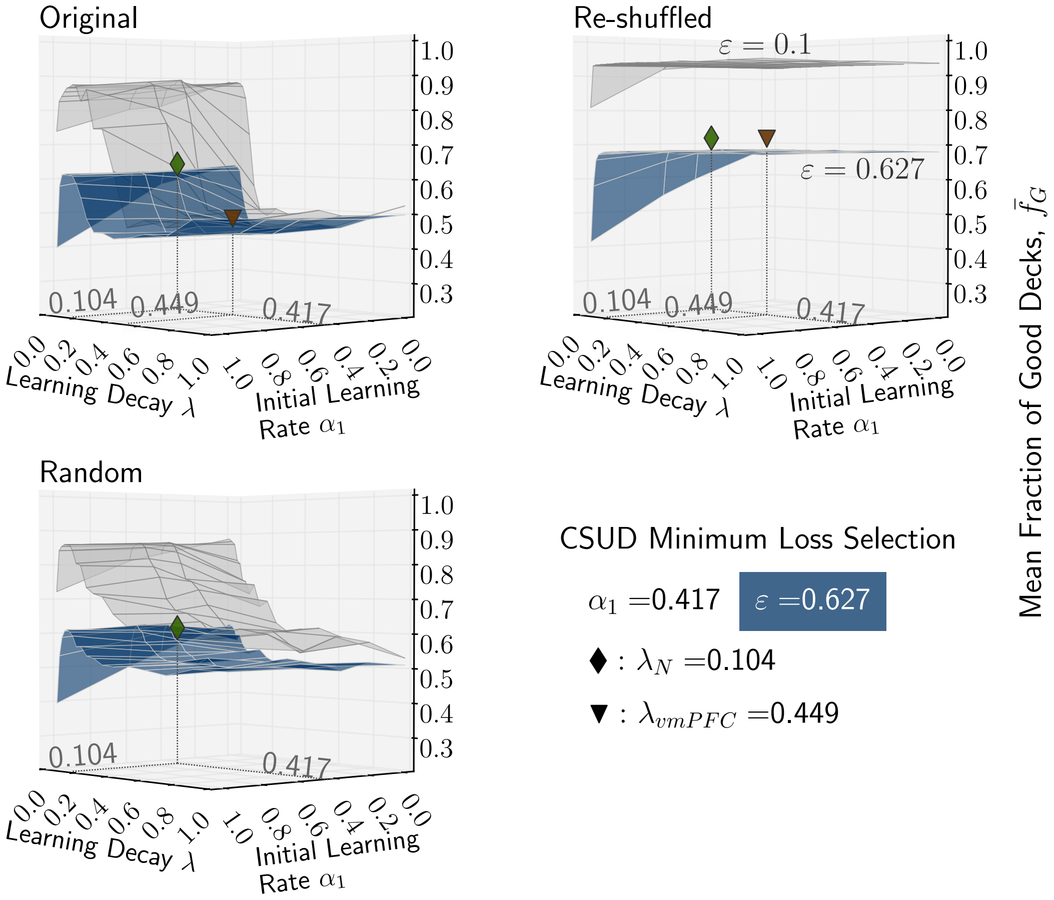


FIGURE 6.3: ε-Greedy agent 3D contours with learning decay λ and exploration ε, but focusing on initial learning rate α_1 effects. Learning decay and exploration variation mirror the contours in Fig. 6.2. The initial learning rate α_1 shows a small technical effect at very high initial learning rates, but otherwise exerts no determining effect.

0.449 for normal and vmPFC impaired subject type behaviour respectively. Both learning decay matches occur after the initial triangularly shaped non-stationarity zone, where incorrect learning pre-dominates. As human results indicate that human subjects would be able to negotiate the non-stationarity zone, this zone is not considered to be of interest in the search outcomes.

The IGT literature includes 20-draw blocked analysis of mean fraction of good decks $\bar{f}_G$. For the original, re-shuffled, and random IGT environments, Figs. 4.1 and 4.2 summarise these results. 20-draw blocked analysis aims to

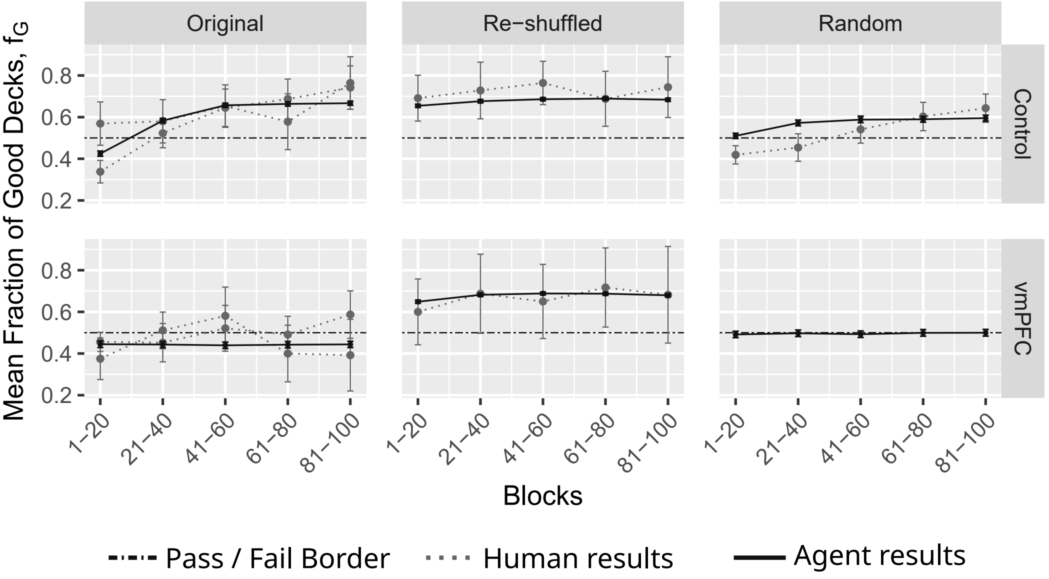


FIGURE 6.4: ε-Greedy Agent 20-draw blocks comparison at CSUD search matches $\alpha_1 = 0.417, \lambda_N = 0.104, \lambda_{vmPFC} = 0.449, \varepsilon = 0.627$. Human results in light gray. Agent results in dark gray, averaged from 750 samples. All error bars at $\pm 2SE$. When error bars are taken into account agent and human 20-draw block performance appears relatively similar. Details in text below.

assess whether exploration decreases during the 100-draw long task, because it is hypothesized participants incorporate what they learn from previous draws and increasingly switch to exploitation. The expectation is that healthy subjects are able to switch from exploration to exploitation while vmPFC impaired subjects are not.

Fig. 6.4 shows 20-draw blocked analysis for the ε-Greedy agent at 20-draw blocks comparison with minimum loss CSUD search matches at $\alpha_1 = 0.417, \lambda_N = 0.104, \lambda_{vmPFC} = 0.449$, and $\varepsilon = 0.627$. For comparison purposes, human results are reproduced in light gray. Agent results, averaged from 750 samples, are in dark gray. All error bars are at $\pm 2SE$ (standard errors). Light gray human subject results have larger standard error bands due to having been obtained from smaller samples as indicated in tables 4.3 to 4.5, where sample sizes range between 6 and 70.

In general, across the original, re-shuffled, and random IGT environments, it is observed that normal ($\lambda_N = 0.104$) and vmPFC impaired ($\lambda_N = 0.449$) parametrised ε-Greedy agents exhibit 20-draw block mean fraction of good decks $\bar{f}_G$ progression respectively similar to human outcomes when $\pm 2SE$ bars are taken into account. That is, for the original IGT environment, the normal ε-Greedy agent, similar to human outcomes, appears to switch from exploration to exploitation, while the vmPFC impaired agent is unable to do so. For the re-shuffled IGT environment, at 20-draw block resolution no exploration versus exploitation effects are discernible. For the random IGT environment, only healthy subject comparison data is available. Agent and human results, however, do appear to differ slightly, especially in the beginning phases of blocks $1-20$, and $21-40$.

In the random IGT environment, if with human data, two-sided confidence bands with 69 degrees of freedom ($samples = 70$) are constructed, then at significance level $\alpha = 99\%$, t-value = 2.6490, for blocks $1-20$ and $21-40$, the null hypothesis that agent and human mean fraction of good decks are equal would be rejected. If the significance level is increased to further decrease the risk of type I error, then at significance level $\alpha = 99.9\%$, t-value = 3.4372, one would fail to reject this null hypothesis. Exploration versus exploitation switch considerations will be revisited after presenting exploration index (EI) comparisons in Fig. 6.5. The exploration index (EI) introduced in section 4.2.3 is a measure of *implied* exploration. The theoretical limits for the exploration index (EI) are 100 for full exploration and 0 for full exploitation.

Fig. 6.5 presents 20-draw EI comparisons at the minimum loss CSUD solution consisting of $\alpha_1 = 0.417$, $\lambda_N = 0.104$, $\lambda_{vmPFC} = 0.449$, and $\varepsilon = 0.627$. Human results are in dotted light gray, whereas agent results appear in solid dark gray, and are averaged from 750 samples. In general, agent results mirror human outcomes. For normal behaviour, the original, re-shuffled, and random IGT environment EI values decrease over 20-draw blocks. For vmPFC impaired behaviour, EI values only decrease for the re-shuffled IGT environment. Except for normal behaviour re-shuffled IGT environment results, agent responses qualitatively look like a smoothed version of the respective human responses.

However, it is possible that the modelled agent and underlying human decision making dynamics partially differ, and therefore lead to divergence regarding normal re-shuffled IGT outcomes. As noted in Table 4.4, the normal re-shuffled IGT human study consists of 17 participants. It is also possible that a small human sample size is leading to increased variation. Further studies could shed light on human exploration index behaviour.

Most significantly, Figs. 6.4 and 6.5 show that over 20-draw blocks, learning rate decay, in general, exerts a strong offsetting influence on exploration.

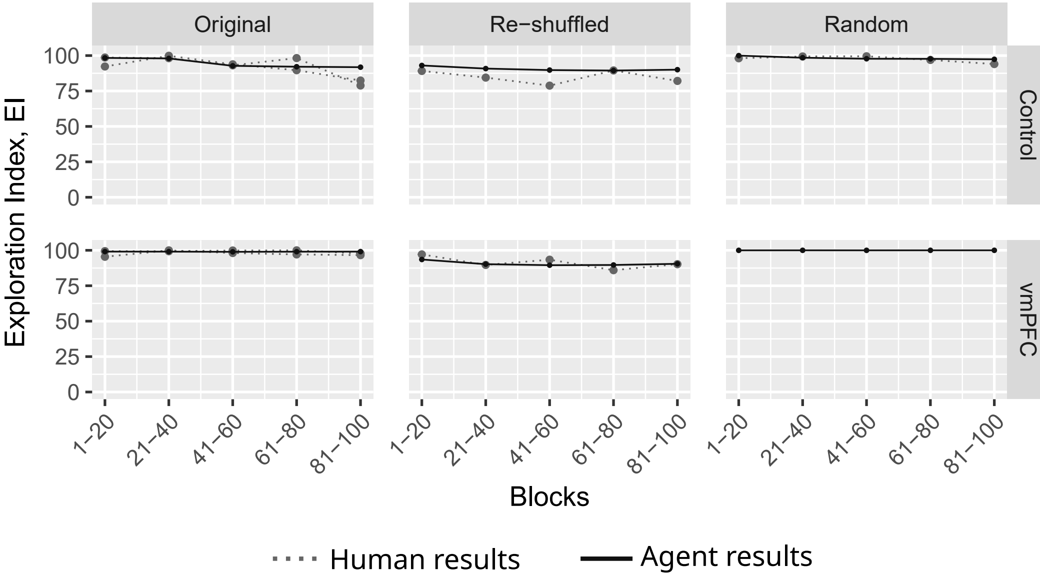


FIGURE 6.5: ε-Greedy Agent 20-draw exploration index (EI) comparison at CSUD search matches $\alpha_1 = 0.417, \lambda_N = 0.104, \lambda_{vmPFC} = 0.449, \varepsilon = 0.627$. Human results in dotted light gray. Agent results in solid dark gray, averaged from 750 samples. Human subject and agent exploration index responses appear relatively similar except for normal behaviour in the re-shuffled IGT environment. Details in text below.

The ε-Greedy agent, despite high constant exploration at $\varepsilon = 0.627$, does show a tapering response both in mean fraction of good decks $\bar{f}_G$, and in the exploration index in the normal original, re-shuffled, and random, and in the vmPFC impaired re-shuffled configurations. The observed tapering is much smoother that what would be expected from 62.7% exploration.

In other words, learning rate decay, by reducing Q-value attribution, leads over time to a decrease in implied exploration, which is measured by the exploration index (EI). The time horizon, in which learning rate decay "freezes"[1] learning, and consequently produces a decrease in implied exploration, depends on the value of learning rate decay λ. For the ε-Greedy agent, at $\lambda_N = 0.104$, the learning rate decreases by 10% per iteration, and this leads to the observed tapered responses in implied exploration as measured by the Exploration Index (EI). On the other hand, at $\lambda_{vmPFC} = 0.449$, the learning rate decreases by 36% per iteration, at which rate IGT environment card sequencing effects appear. For the original and random decks, given vmPFC impaired settings, implied exploration (EI) stays close to 100,

[1] The term "freezing" was suggested by a reviewer for Koluman et al. (2019).

suggesting that agent learning was frozen prior to the end of the first 20-draw block (since all Q-values are initialised to 0, there must have been very little choice differentiating Q-value accruals). However, by design the re-shuffled IGT reveals deck characteristics within the first 20-draw block, and this leads to a tapered response.

One of the applications of simulation based cognitive computing is to hypothesize about the computational model, which may be underlying a given decision making problem. For example, Doya (2002) presents such a review centring on TD(κ) reinforcement learning models applied to neurotransmitter effects. Among others, Maia and McClelland (2005) consider the exploration versus exploitation trade-off in the IGT.

The 20-draw block analysis indicates that exploration versus exploitation in the IGT could be mediated via two computational pathways: (a) a direct exploration pathway (which for the CSUD hyper-parametrised ε-Greedy agent is constant), and (b) an indirect pathway driven by learning rate decay, which creates a learning freeze after a set number of iterations, and thereby leads to decreasing implied exploration. With the ε-Greedy agent with constant direct exploration, learning rate decay λ appears to provide a determining indirect influence on exploration.

For the original, re-shuffled, and random IGT environments, the jitter plots in Fig. 6.6 assesses the human outcome match performance of repeated simulations of the discovered minimum loss CSUD solution at $\alpha_1 = 0.417$, $\lambda_N = 0.104$, $\lambda_{vmPFC} = 0.449$, and $\varepsilon = 0.627$. On the horizontal axis, in addition to the CSUD discovered exploration $\varepsilon = 0.627$, exploration rates of $\varepsilon = 0.10$ and $\varepsilon = 0.535$ are also listed. As listed in Table 6.3, these additional exploration rates have been used to investigate agent behaviour at further exploration values to provide context for the CSUD minimum loss discovered agent hyper-parameter values.

The dashed horizontal lines indicate the minimum and maximum of the to be matched human IGT outcome ranges. The red squares mark sample means ± 2 standard errors. The green coloured dots represent matches to respective human IGT outcome ranges. The green coloured text reports the number of samples, and in parenthesis, the percentage of human range matches achieved. The blue coloured dots indicate a normative pass, that is a mean fraction of good decks value above 0.50, hypothesized to be associated with normal learning decay $\lambda_N = 0.104$. The blue coloured text

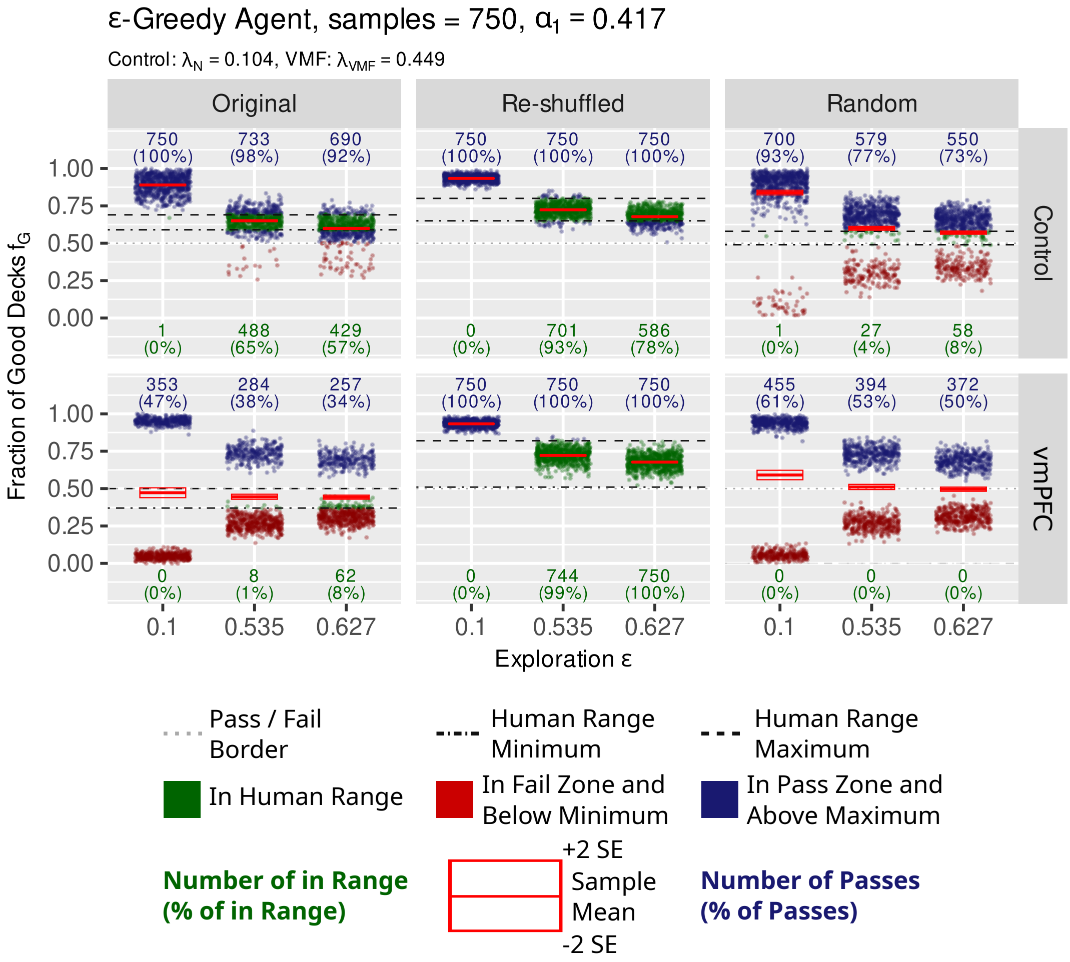


FIGURE 6.6: ε-Greedy agent comparison of repeated simulation outcomes to human IGT results. At $\varepsilon = 0.627$ with CSUD minimum loss hyper-parameter values, the ε-Greedy agent achieves the highest human range matches for the re-shuffled IGT. The vmPFC impaired ε-Greedy agent tends towards bi-modal outcomes for the original and random IGT. Full details are in the text.

reports the number of samples, and in parenthesis, the percentage of normative pass matches achieved. The red dots represent a normative fail, that is a mean fraction of good decks value of 0.50 or lower, hypothesized to be associated with vmPFC impaired learning decay $\lambda_{vmPFC} = 0.449$.

Concerning human IGT outcome range matches, Fig. 6.6 shows that given 750 simulated samples at the CSUD minimum loss hyper-parameters, that is at $\alpha_1 = 0.417$, $\lambda_N = 0.104$, $\lambda_{vmPFC} = 0.449$, and $\varepsilon = 0.627$, the ε-Greedy agent achieves in the re-shuffled IGT, 78% matches with normal and 100% matches with vmPFC impaired behaviour configurations. The agent also achieves 57% matches with original IGT normal learning decay. However, for the original IGT vmPFC impaired behaviour configuration, and the random IGT normal behaviour configuration, the agent only achieves 8% of matches. There is no human outcome match data available for the random IGT vmPFC impaired configuration.

The ε-Greedy agent does not achieve a high level of human IGT outcome matches across all ranges reported in Table 4.8. However, the agent performs better in matching the respective normative pass outcomes reported in Table 4.8. In the re-shuffled IGT, the ε-Greedy agent achieves a 100% normative pass match for the normal and vmPFC impaired configurations. In the original IGT, the agent achieves 92% and 34% normative pass matches for the normal and vmPFC impaired configurations respectively. Note that the original IGT vmPFC impaired configuration, 34% normative pass match is equivalent to a 66% normative fail match. In the random IGT normal configuration, the agent achieves a 73% normative pass match.

Fig. 6.6 reveals that both learning decay λ and exploration ε interact with IGT environment card sequencing effects to produce different density (jitter) plots for the fraction of good decks f_G. For normal learning decay with $\lambda_N = 0.104$, the original and re-shuffled IGT fraction of good decks density appears unimodal and relatively symmetric as indicated by the red bar, which shows the mean ± 2 SEs. However, as exploration increases, symmetry decreases in favour of a left-hand tail. In contrast, the random IGT appears bi-modal, with the modes tending towards the match ranges as exploration increases. For vmPFC impaired learning decay with $\lambda_{vmPFC} = 0.449$, the re-shuffled IGT fraction of good decks f_G appears unimodal symmetric, whereas the original and random IGT present as bi-modal, asymmetric, and with the modes tending towards 0.5 as exploration increases.

Finally, non-parametric multi-variate analysis of variance (np-M/ANOVA) is performed using the npmv R package (Bathke et al., 2008; Burchett et al., 2017). The np-M/ANOVA method assesses the response of multiple variables to a single factor with multiple levels. The np-M/ANOVA analysis asks the question: given initial learning rate $\alpha_1 = 0.417$ and exploration $\varepsilon = 0.627$, do learning rate decay $\lambda_N = 0.104$ and $\lambda_{vmPFC} = 0.449$ parametrisations lead to a statistically significant difference in mean fraction of good decks $\bar{f}_G$ for the original, re-shuffled, and random IGT environments? Hence the single factor of interest is learning rate decay and the responses are the mean fraction of good decks for the original, re-shuffled, and random IGT environments. The np-M/ANOVA design is balanced with 750

Test Variant	Test Statistic	df1	df2	p-Value	Subset Results
Original ∣ Re-Shuffled ∣ Random vs. Learning Decay λ					At α = 0.01, the null hypotheses of learning decay factor equality is rejected. Only equality of the re-shuffled response cannot be rejected.
ANOVA Type[a]	71.02	2.984	4470.435	0	
Original ∣ Random vs. Learning Decay λ					At α = 0.01, the null hypotheses of equal original and random, original only, and random only responses are rejected.
ANOVA Type	108.824	1.995	2988.442	0	
Wilks Lambda	113.288	2.000	1497.000	0	
Re-Shuffled vs. Learning Decay λ					
ANOVA Type	0.258	1.000	1498	0.611	Single response variable, no subsets.
Wilks Lambda	0.258	1.000	1498	0.611	

[a] Wilks Lambda could not be computed due a singular rank matrix.

TABLE 6.4
ε-Greedy agent np-M/ANOVA analysis of mean fraction of good decks $\bar{f}_G$ with learning decay λ as factor. At significance level $\alpha = 0.01$ Mean fraction of good decks $\bar{f}_G$ responses are statistically significantly different, except for the re-shuffled IGT environment.

samples per cell.

The results are reported in Table 6.4. The most comprehensive null hypothesis is no multivariate response to any factor levels. This null hypothesis is tested in the first row of the table. The p-value of 0 indicates that the null hypothesis is strongly rejected globally. The right-most column of Table 6.4 summarises subset responses. At significance level $\alpha = 0.01$, factor level equality is rejected. Further equality of mean fraction of good decks $\bar{f}_G$ for the following response variable subsets is rejected: original and re-shuffled and random, re-shuffled and random, original and random, original and re-shuffled, original, and finally random.

Only equality of mean fraction of good decks $\bar{f}_G$ for the re-shuffled IGT environment fails to be rejected. That is to say, the normal and vmPFC impaired ε-Greedy agent mean fraction of good deck results mirror, from a statistical hypothesis testing perspective, corresponding human results for the re-shuffled IGT environment, where both normal and vmPFC impaired humand subjects achieve a mean fraction of good decks pass. The second and third rows of the table provide further insight into the subset results. The third row shows that re-shuffled mean fraction of good decks $\bar{f}_G$ response to learning decay factor variation produces a p-value of 0.611, indicating failure to reject the null hypothesis of equality, thereby corroborating the visual result in Fig. 6.6. The np-M/ANOVA results also verify that the ε-Greedy agent at CSUD selected minimum loss hyper-parameter values does indeed replicate expected human behavioural results in terms of IGT outcomes for the original, re-shuffled, and random IGT environments.

6.3.1 ε-Greedy Agent Discussion

High learning rate decay appears to be the central mechanism driving the ε-Greedy agent's ability to replicate human IGT outcome results for the original, re-shuffled, and random IGT environments for healthy and vmPFC impaired subjects. As the 750 sample grid search verification results in Fig. 6.6 reveal, at the selected minimum loss CSUD hyper-parametrisation, the agent achieves variable consistency in matching human performance ranges. The agent's human outcome match consistency improves with respect to the IGT normative pass criterion. The statistical analysis summarised in Table 6.4 shows that with high learning rate decay as a proxy for vmPFC impairment, the agent qualitatively replicates human IGT outcome results in statistically significant terms at a significance level of $\alpha = 0.01$.

The ε-Greedy agent represents a very special decision making arrangement in that nominal exploration is constant, large, and never decreases. However, as exploration index (EI) results in Fig. 6.5 indicate, high learning rate decay effectively leads to a decrease in implied exploration. It has also been revealed that to approximate human IGT performance ranges, exploration has to be very high at $\varepsilon = 0.627$.

The question arises as to what might be the aim of a decision making strategy with very high exploration? The No Free Lunch (NFT) theorems (Wolpert & Macready, 1997) state that in the space of all problems and all algorithms, no algorithm can outperform random search. Accordingly, it is proposed that high exploration could be a mitigation strategy aiming to address limited information, finite time, and limited opportunities, which may all be expressed as algorithmic specificity.

Another plausible explanation to high exploration is offered in Wilson et al., 2014, who find that in choice tasks with a longer horizon, human subjects exhibit higher exploration driven not by random but by directed exploration, which increases sampling from more informative options. In their study, the authors compare choice selection tasks with durations of 5 and 10 periods respectively (including 4 periods of training), with the longer duration task outcomes exhibiting high exploration geared towards the more informative option. In these 5 and 10 period duration tasks, however, task length was communicated to the participants. In the IGT, task duration consists of 100 periods but is not known by the participants. Given the relatively longer length of the IGT, however, it is possible that the participants start to act as if the IGT is a long duration task and accordingly explore the informative choices more. Such behaviour has been indicated, for example, with respect to the (original) IGT disadvantageous deck B (Lin et al., 2007), and discussed in terms of frequency-gain effects, where a rare event is undervalued. The undervaluing a rare event has also been noted in Hertwig et al., 2004. It is possible, however, that the increased draws from deck B stem not from undervaluing, but from information seeking about the rare event; such information can only be found by increasing the number of draws.

The original, re-shuffled, and random IGTs do not constitute a neutral environment. The original and re-shuffled IGTs use card sequencing effects to respectively disguise or reveal deck net yields during the first 8 rounds of the task. The random IGT, by design does not have card sequencing effects; but contains frequency and loss effects as have been discussed in the EV (2.1), PV (2.2), and ORL (2.4) models. Frequency effects refer to events, which occur rarely, making estimation of the central tendency difficult. Loss effects refer to the desire to avoid negative outcomes and chase large payoffs, even if deck cards would produce, by means of regular small penalties, a net loss despite high infrequent payoffs.

In Fig. 6.6, both normal and vmPFC impaired random deck IGT configurations produce bi-modal distributions with comparatively extreme outcomes for low exploration. At $\varepsilon = 0.1$, 93% of agent simulations achieve a high normative pass (blue dots), while 7% of simulations end in a low normative fail. At $\varepsilon = 0.627$, 73% of agent simulations achieve a normative pass, while 27% of simulations end in a normative fail. However, at $\varepsilon = 0.627$, the normative pass or fail jitter plot masses (modes) get closer to each other, leading to comparatively moderate outcomes: while the passes are not as high, the fails are also not as low.

Hence, a difference between population level and individual decision making is observed. For the random IGT, the agent population benefits at $\varepsilon = 0.1$, but a few individual agents are much worse off. At $\varepsilon = 0.627$, the population is worse off in terms of the central tendency as indicated by the red bar, but individual agents who fail mainly do so above a mean fraction of good decks $\bar{f}_G$ value of 0.25.

The IGT could encapsulate a decision making problem where individual mistakes can be very costly and irrecoverable. In a population, such as human beings, where individuals are valued highly, it could be speculated that to mitigate limited decision making resources, evolutionary or behavioural tendencies may therefore have adopted to produce as high individual exploration as the population can tolerate. As noted above, NFT theorems (Wolpert & Macready, 1997) state that in the space of all search algorithms and search problems, no algorithm will perform universally better than random search. Hence, one approach to dealing with complex search problems is by increasing randomness to reduce search algorithm specificity.

6.4 Boltzmann Agent Results

Table 6.5 and Fig. 6.7 present Boltzmann agent CSUD search results in tabular and graphic forms respectively. Boltzmann agent CSUD searches could not achieve the full search match consisting of matching normal human IGT outcome original, re-shuffled, and random environment ranges; and vmPFC impaired human IGT outcome original and re-shuffled environment ranges. CSUD searches produced very few full search match candidates, and the minimum loss selection chosen from these candidates in turn failed full grid search verification.

A search budget of 1000 iterations was used and 5 gradient samples for each gradient evaluation per iteration were employed. After 1000 iterations, only 9 full-match candidates were found. The minimum loss candidate was chosen from among these nine matches as per the search methodology. However during grid search verification, the CSUD full search minimum loss candidate only fulfilled partial match conditions consisting of, for normal and vmPFC impaired human IGT outcome ranges, the original and re-shuffled environments. A simultaneous match to normal human random

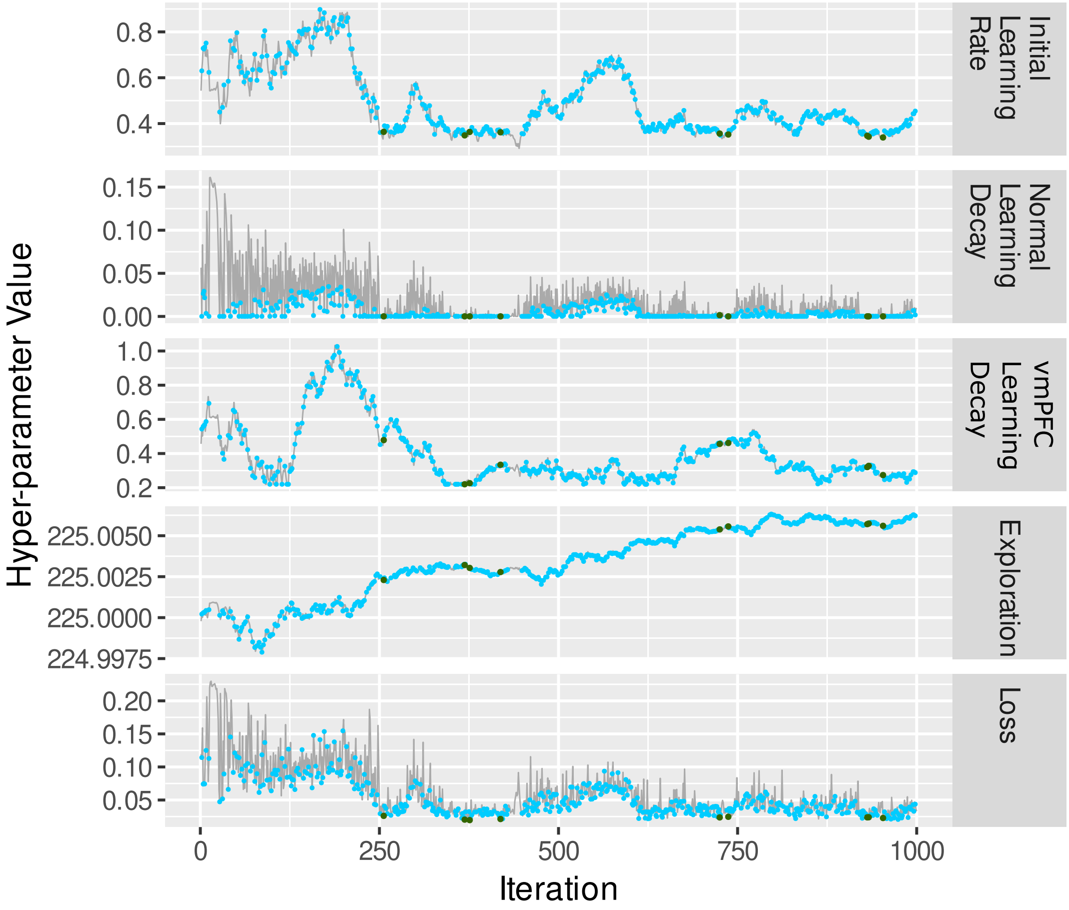


FIGURE 6.7: Boltzmann agent CSUD iterations. Green and blue points indicate $(\alpha_1, \lambda_N, \lambda_{vmPFC}, \tau)$ hyper-parameter tuples, which produce full and partial CSUD search matches respectively. Partial search matches fulfil normal and vmPFC impaired human IGT outcome ranges for the original and re-shuffled IGT environments. Details in text.

IGT outcomes could not be found. Here, these partial match results are reported.

Fig. 6.7 highlights CSUD search matches for full and partial matches, which are coloured green and light blue respectively. CSUD search finds many matches (459 of 1000 iterations) where CSUD hyper-parameter selections generate agent results, which lie within normal and vmPFC impaired human outcome ranges for the original and re-shuffled IGT environments. These light blue coloured zones constitute partial search matches. The green coloured dots (9 of 1000) represent full matches, which in addition to the partial match defined above, match normal human random IGT outcome ranges. The numeric results in Table 6.5 come from the CSUD full match

	Minimum Loss	Range	Mean	Median	Standard Error
Loss	0.0196	0.0196 - 0.0261	0.0229	0.0237	0.000703
Initial learning rate α_1	0.364	0.339 - 0.364	0.353	0.352	0.00306
Normal learning rate decay λ_N	1.0e^{-4}	1.0e^{-4} - 0.00151	2.56e^{-4}	1.0e^{-4}	0.000156
vmPFC impaired learning rate decay λ_{vmPFC}	0.226	0.22 - 0.479	0.344	0.328	0.0333
Temperature τ	225.002	225.002 - 225.005	225.004	225.005	0.000494
Matched environments	Full CSUD and partial grid search verified matches as discussed in text.				
Match count	9 of 1000 iterations (Full CSUD matches).				

TABLE 6.5
Boltzmann agent CSUD minimum loss search matches after 1000 iterations. The highlighted minimum loss column shows selected agent hyper-parameters. Light gray indicates minimum loss and the associated initial learning rate α_1. Dark-gray, mid-gray, and light blue indicate minimum loss associated normal learning rate decay λ_N, vmPFC impaired learning rate decay λ_{vmPFC}, and exploration τ respectively.

set. However, during grid search verification, the CSUD full match set could only generate the partial matches as defined by the light blue range in Fig. 6.7.

Table 6.5 highlights the CSUD minimum loss hyper-parameter combination, which for initial learning rate α_1, normal learning decay λ_N, vmPFC impaired learning decay λ_{vmPFC}, and temperature τ is at ($\alpha_1 = 0.364, \lambda_N = 1.0e^{-4}, \lambda_{vmPFC} = 0.226, \tau = 225.002$). Note that the selected normal learning decay λ_N is at minimum loss at $1.0e^{-4}$, and that exploration temperature τ does not fluctuate much. The contribution of exploration temperature τ to decision making is not easy to discern, and the associated Boltzmann action probabilities are provided later.

It is noted that as with the ϵ-Greedy agent, normal and vmPFC impaired behaviour configurations are driven by learning decay λ. Normal learning decay λ_N is very close to 0 and at the lower constraint boundary. This suggests that the normal Boltzmann agent may have a constant learning rate solution. Here, on the basis that a learning decay rate of $1.0e^{-4}$ is already very low, a constant learning rate Boltzmann agent variant is not discussed. The question considered next is why the Boltzmann agent CSUD search might

Hyper-parameter	Grid Points
Initial learning rate α_1	0.05, 0.364, 0.66, 0.99
Learning rate decay λ[a]	0.0001, 0.226
Temperature τ	5, 75, 225, 425
IGT length	Q-learning samples
100	750

[a] The learning rate decay grid is constructed around the CSUD reported values. Appendix C provides the construction method.

TABLE 6.6
Boltzmann agent CSUD verification. Hyper-parameter grid search criteria for joint original, re-shuffled, random IGT.

not discover full match hyper-parameter settings, which can be replicated in grid search verification.

Table 6.6 reports the localised grid search specification used to verify Boltzmann agent CSUD results. Fig. 6.8 and Fig. 6.9 present 2D and 3D views respectively of the selected search hyper-parameters in the context of the neighbourhood grid search presented in Table 6.6.

Fig. 6.8 presents visual impressions qualitatively similar to those obtained from the ε-Greedy agent searches. As before, the lower x-axis indicates learning rate decay. Using the conversion formula $(1 - e^{-\lambda}) * 100$, the upper x-axis translates the lower x-axis learning rate decay value into a per period learning rate decay percentage. The contour line associated with CSUD minimum loss hyper-parametrisation is coloured in black. Learning rate decay λ is responsible for determining normal ($\lambda_N = 1.0e^{-4}$) versus vmPFC impaired ($\lambda_{vmPFC} = 0.226$) agent behaviour. Increased temperature τ leads to a vertical downward shift of the mean fraction of good decks $\bar{f}_G$ contours, and facilitates inclusion into the normal (dark gray) and vmPFC impaired (light gray) human match ranges.

Fig. 6.8 indicates why, subject to the search criteria, the Boltzmann agent is not able to achieve a match in all tested IGT environments. In particular, it is important note that for a given initial learning rate α_1, the Boltzmann agent cannot obtain a simultaneous match in the original and random IGT environments for the respective normal human outcome ranges. At $\lambda_N = 1e^{-4}$, the original IGT environment match is at the lower human

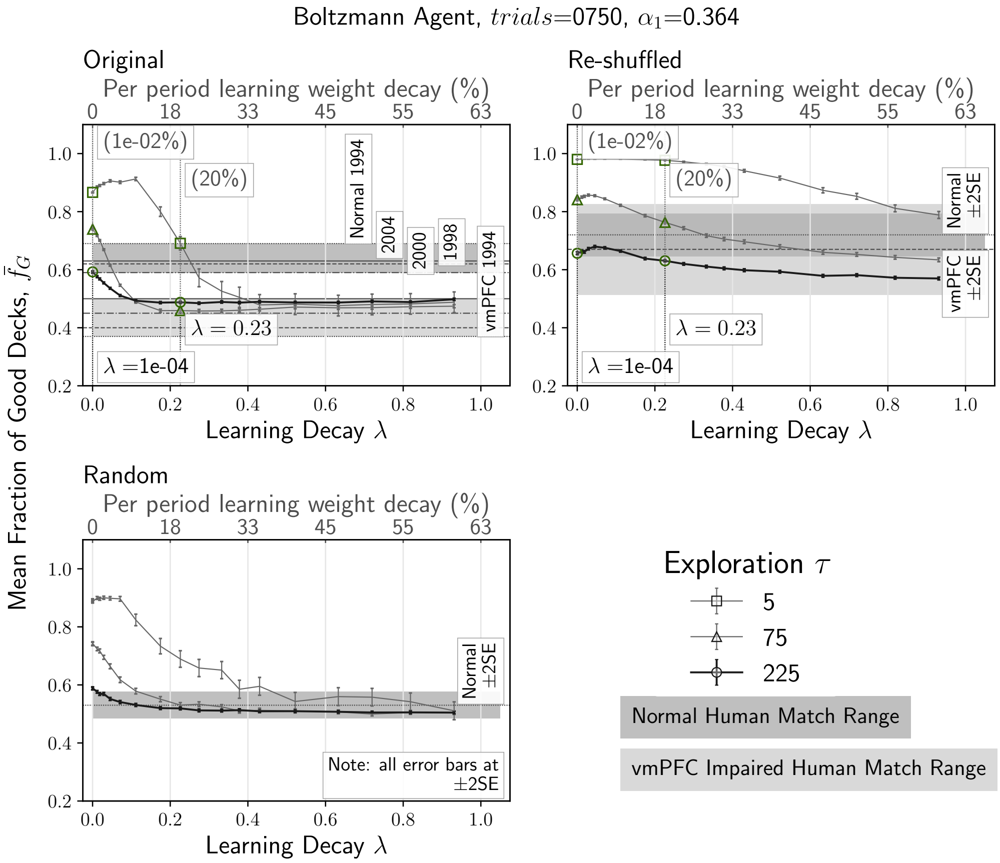


FIGURE 6.8: Boltzmann agent 2D contours showing learning decay λ and exploration τ effects. The dark and light gray zones indicate normal and vmPFC impaired human outcome match ranges for the original, re-shuffled, and random IGT environments. Learning decay variation reproduces human IGT outcomes, while exploration variation is responsible for matching human IGT outcome performance ranges.

performance boundary. However, at $\lambda_N = 1e^{-4}$, agent random IGT performance is slightly above the corresponding human outcome match range maximum.

Fig. 6.9 reveals that for the Boltzmann agent, the initial learning rate α_1 has some influence on decision making outcome. However, as with the ε-Greedy agent, this effect is most noticeable at high learning rates and low learning decay rates, leading to initial non-stationarity effects consisting of a region, where increasing learning rate decay improves mean fraction of good decks outcomes. The visual signature of these initial non-stationarity effects consists of the draped over area especially prominent in the original, and to a lesser extent in the random IGT environment outcome contour

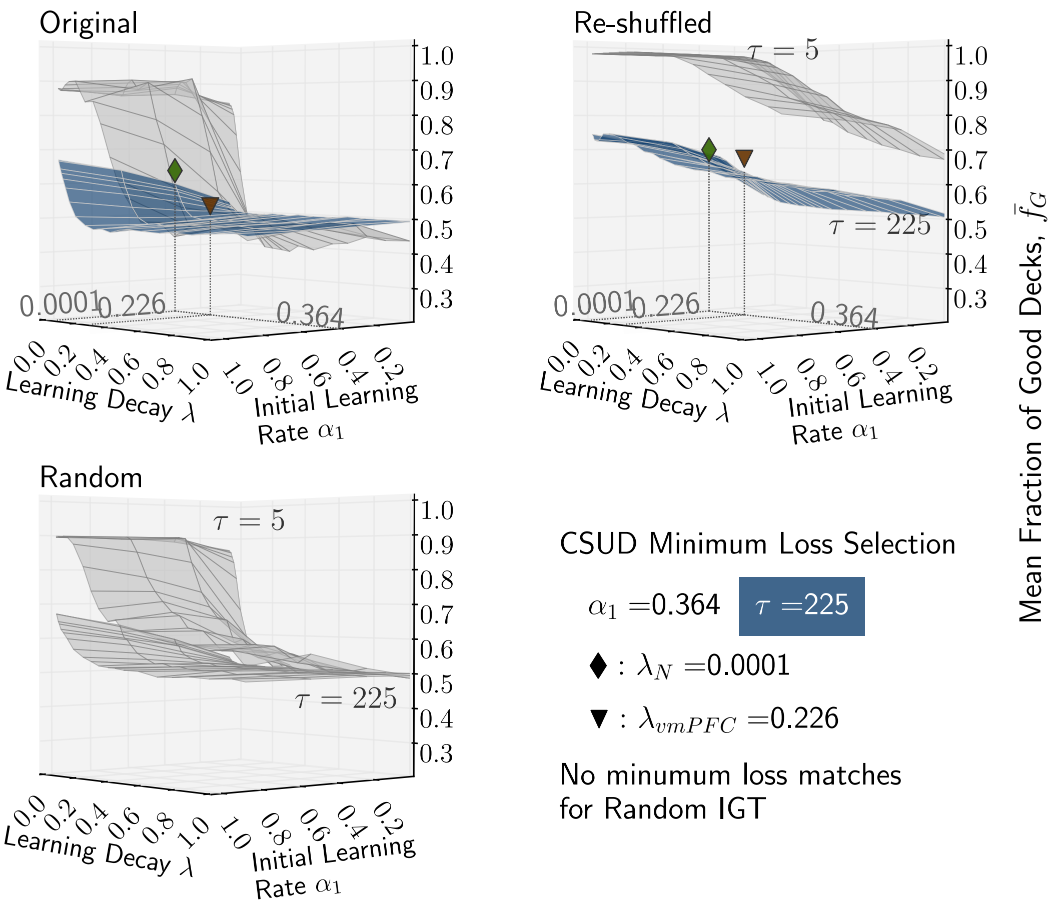


FIGURE 6.9: Boltzmann agent 3D contours for learning decay λ and exploration τ, focusing on initial learning rate α_1 effects. Learning decay and exploration variation mirror the contours in Fig. 6.8. The initial learning rate α_1 shows a small technical effect at very high initial learning rates, but otherwise exerts no determining effect.

plots.

While grid search results indicate that there are indeed multiple matching solutions in addition to the one discovered by CSUD, none of these solutions appear capable of producing a match across all of considered IGT environments and human behaviours. Hence, the simple Boltzmann agent implementation does not appear capable of achieving the desired number of simultaneous human outcome matches. The matching misses are quite close, and matches could potentially be obtained by increasing the range of the catchment zone. However, such mitigating approaches are not explored here so as to keep the methodology presented in Table 4.8 consistent.

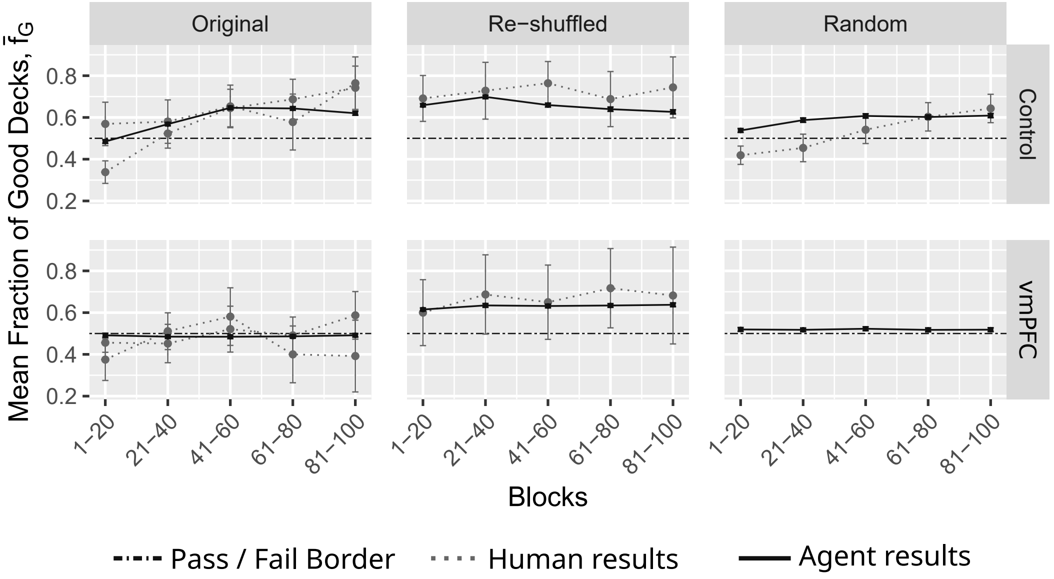


FIGURE 6.10: Boltzmann agent 20-draw blocks comparison at CSUD search matches $\alpha_1 = 0.364, \lambda_N = 1e^{-4}, \lambda_{vmPFC} = 0.226, \tau = 225.002$. Human results in dotted light gray. Agent results in solid dark gray, averaged from 750 samples. All error bars at $\pm 2SE$. When error bars are taken into account agent and human 20-draw block performance appears relatively similar. Details in text below.

Further, increasing the catchment zone would not alter the result that the Boltzmann agent exhibits difficulty in achieving a match across all IGT environments and human behaviours as measured in terms of distance to the corresponding human means.

Fig. 6.10 and Fig. 6.11 show Boltzmann agent mean fraction of good decks $\bar{f}_G$ and exploration index (EI) behaviour respectively in 20-draw blocks for normal and vmPFC impaired behaviour. Agent outcomes are plotted in solid dark gray. Corresponding human 20-draw block outcomes are depicted by the dotted light gray lines. The dash-dotted line indicates the normative pass point of 0.5. In terms of mean fraction of good decks, in Fig. 6.10, when human performance ±2 standard error (SE) is taken into account, Boltzmann agent 20-draw block performance is relatively similar but shows small deviations for normal behaviour for blocks 1-20 and 21-40 in the random IGT environment where agent performance is better than human performance.

Fig. 6.11 depicts the Boltzmann agent exploration index (EI) results. Corresponding human 20-draw block outcomes are depicted by the dotted light

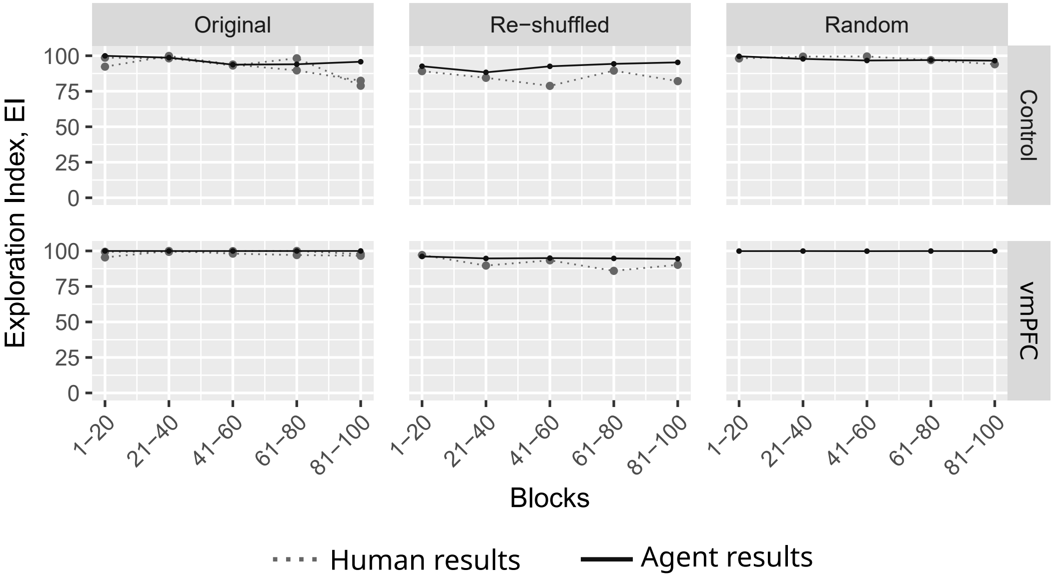


FIGURE 6.11: Boltzmann agent 20-draw blocks exploration index (EI) comparison at CSUD search matches $\alpha_1 = 0.364, \lambda_N = 1e^{-4}, \lambda_{vmPFC} = 0.226, \tau = 225.002$. Human results in dotted light gray. Agent results in solid dark gray, averaged from 750 samples. Human subject and agent exploration index responses indicate high index values, which are higher for the agent. This is because high Boltzmann exploration approximates random search. Details in text below.

gray lines. The Boltzmann agent exhibits mixed (qualitative) success in matching human exploration index profiles. When exhibiting vmPFC impaired behaviour, the Boltzmann agent exploration index shows similar trends for the original and re-shuffled IGT environments. However when exhibiting normal behaviour, the Boltzmann agent indicates a higher level of exploration than the corresponding human values in the latter draw blocks of the original and re-shuffled IGT environments.

Fig. 6.12 assesses the percentage of fraction of good decks f_G matches achieved for 750 repeated IGT simulations at the selected CSUD values. The green coloured jitter plots indicate matches to the corresponding human ranges. Blue jitter plots indicate any normative passes outside of the human match ranges, whereas red jitter plots indicate normative failures ($f_G \leq 0.50$) outside of human match ranges. The red central line and bars indicate the mean and ±2 SEs. At CSUD selection $\alpha_1 = 0.364, \lambda_N = 1^e{-}4, \lambda_{vmPFC} = 0.226, \tau = 225$, the Boltzmann agent achieves for normal configuration, 55%, 57%, and 43% human range matches for the original, re-shuffled, and random IGT environments respectively. For vmPFC impaired configuration,

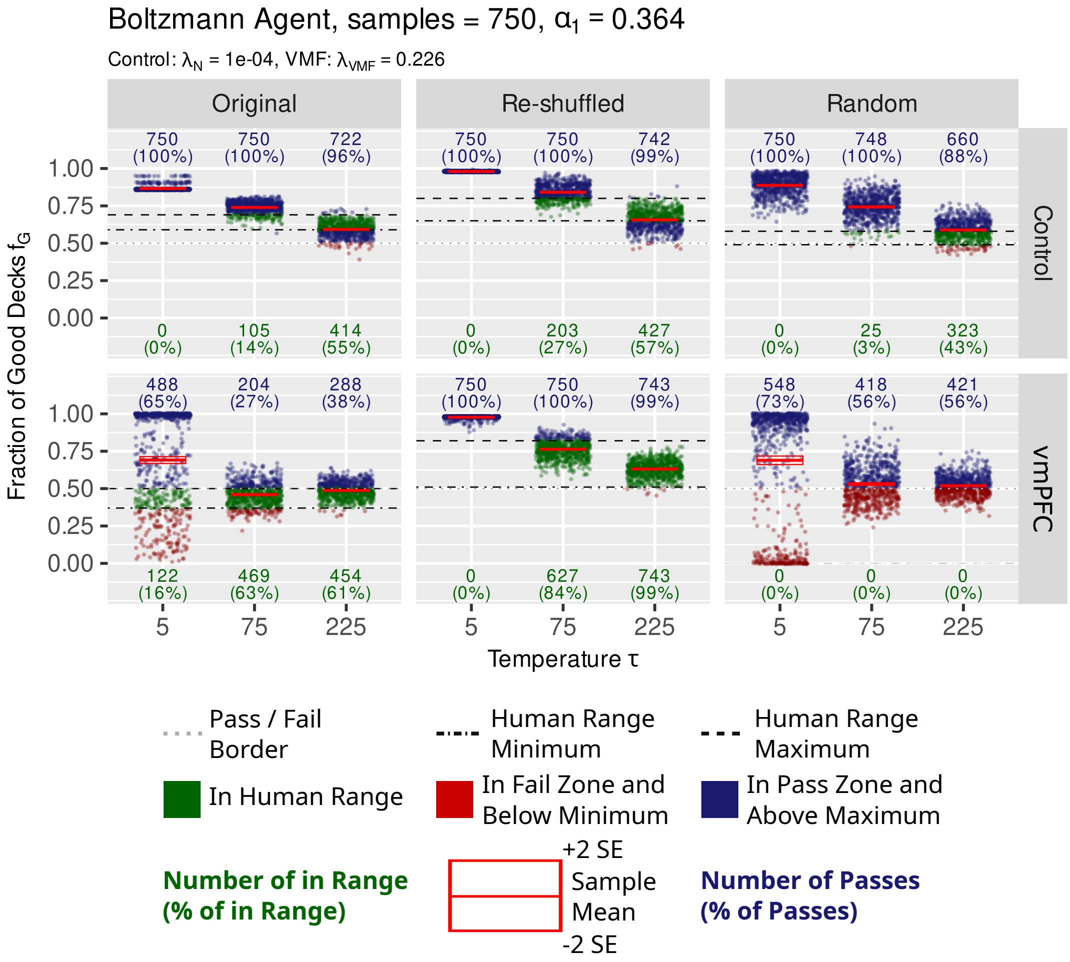


FIGURE 6.12: Boltzmann agent comparison of repeated simulation outcomes to human IGT results. At $\tau = 225$ and reported CSUD minimum loss hyperparameter values, the Boltzmann agent achieves the highest matches for the re-shuffled IGT environment. Full details are in the text.

the agent achieves 61% and 99% human range matches for the original and re-shuffled IGT environments respectively.

When compared with the corresponding ε-Greedy agent results in Fig. 6.6, the Boltzmann agent results in Fig. 6.12 produce an overall better conformance to human match ranges. In particular, the ε-Greedy agent exhibits a stronger tendency towards bi-modal fraction of good decks outcomes, which exhibit low probabilistic mass at the corresponding simulation means. That is, in the ε-Greedy agent plots, there are a lot of areas, where there are few jitter plot dots around the red lines. Put differently, the Boltzmann agent is by design resistant to the polarising effect of learning rate decay. This is a well-known design feature of the Boltzmann agent, and also one of the reasons why Boltzmann agents remain popular: They are designed to produce smooth unimodal probabilistic decision making.

In Fig. 6.12, it is noted that as exploration increases from $\tau = 5$ to $\tau = 225$, agent mean fraction of good decks performance decreases across all simulations. That is, the entire jitter plot tends to shift down as exploration increases. This observation reminds of the ε-Greedy agent result that in order to attain human match ranges, exploration needs to be high. The second point to note is that, while the Boltzmann agent is resistant to the polarising effect of learning rate decay, increasing exploration alone, is not enough to escape poor learning. This is because at high learning rate decay, even if exploration leads to a positive result, such a result can no longer contribute sufficiently to overturn aggregated Q-values.

Table 6.7 presents the results of the np-M/ANOVA analysis with original, re-shuffled, random IGT mean fraction of good deck $\bar{f}_G$ output as response variables, and with $\lambda_N = 1e^{-4}$ and $\lambda_{vmPFC} = 0.226$ as the learning rate decay factor values. For the Boltzmann agent at statistical significance level $\alpha = 0.01$, learning rate decay exerts a significant effect for all possible response variable combinations: original and re-shuffled and random, original and random, original and re-shuffled, re-shuffled and random, original, re-shuffled, and random. However, in order to replicate human results, one would have expected learning rate decay not to have a significant effect for the mean fraction of good decks results for the re-shuffled IGT environment. Hence, on the basis of these results, the Boltzmann agent replicates most human results, but is unable to support the key result, which expects that with the re-shuffled IGT environment, normal ($\lambda_N = 1e^{-4}$) and vmPFC impaired ($\lambda_{vmPFC} = 0.226$) behaviour should not lead to statistically significantly different mean fraction of good decks outcomes.

Finally, some insight is provided into what an exploration temperature of $\tau = 225$ (or, to be precise $\tau = 225.002$) means in practice. Fig. 6.13 shows Boltzmann agent action (deck) selection probabilities at completion of the IGT. The dash-dotted line at 0.25 highlights the uniform probability selection threshold. Probabilistic selection of 0.25 for each deck implies that 50% of the cards have been chosen from the good decks, and therefore leads to a normative fail at the maximum fail level of mean fraction of good decks $\bar{f}_G = 0.5$. In other words, given the high Boltzmann exploration figure of $\tau = 225$, it is probable that some random draw realizations may achieve mean fraction of good decks $\bar{f}_G \geq 0.5$ The jitter plots summarise the range

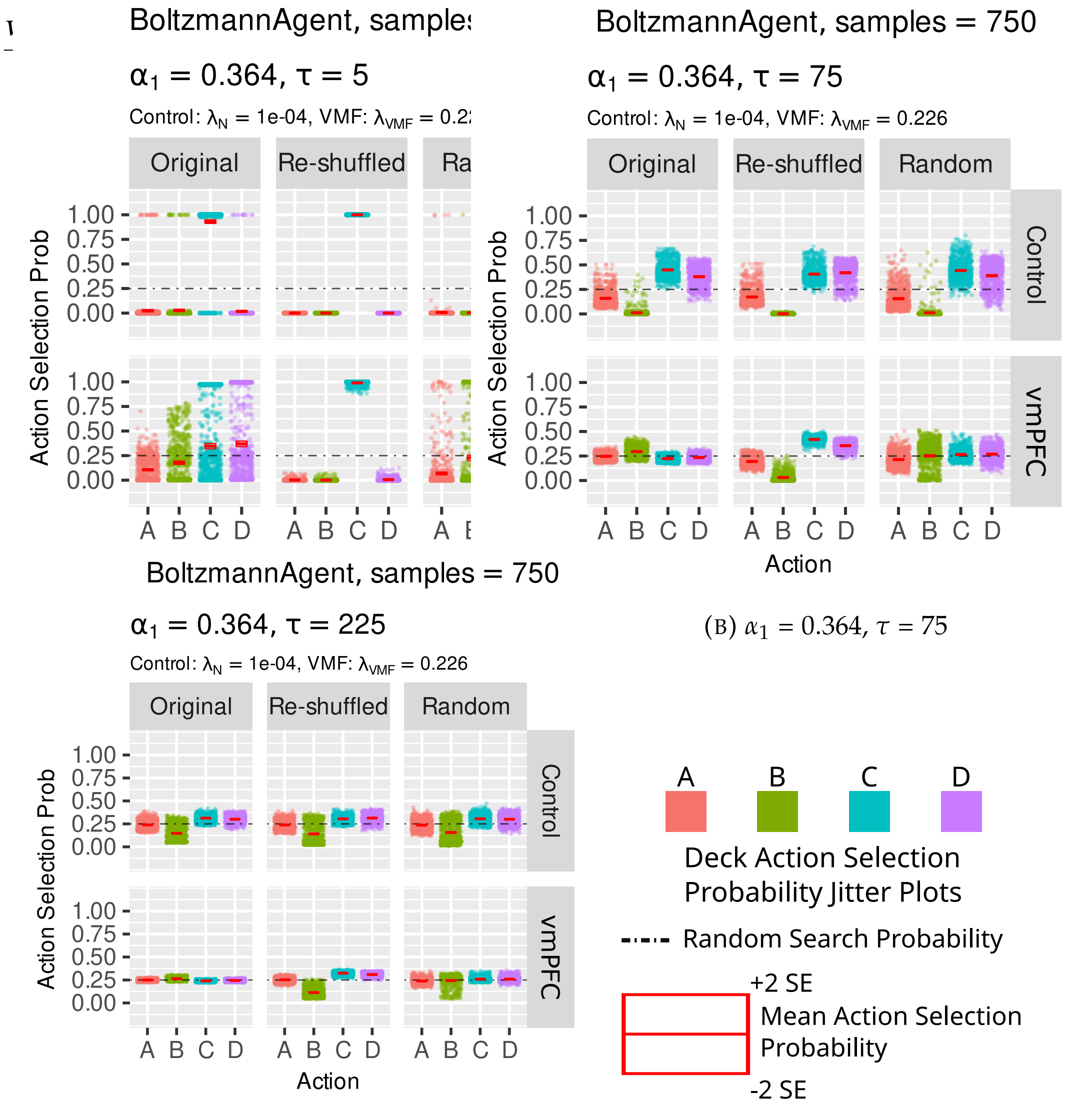
Test Variant	Test Statistic	df1	df2	p-Value	Subset Results
Original | *Re-Shuffled* | *Random vs. Learning Decay* λ					
ANOVA Type[a]	543.823	2.774	4154.857	0	At $\alpha = 0.01$, the null hypotheses of learning decay factor equality is rejected for all response variable combinations with $\lambda_N = 1e^{-4}$ and $\lambda_{vmPFC} = 0.226$ as normal and vmPFC impaired factors respectively.

[a] Wilks Lambda could not be computed due a singular rank matrix.

TABLE 6.7
Boltzmann agent np-M/ANOVA analysis of mean fraction of good decks $\bar{f}_G$ with learning decay λ as factor. At statistical significance level $\alpha = 0.01$, mean fraction of good decks $\bar{f}_G$ responses are statistically significantly different, *even* for the re-shuffled IGT environment. The Table only presents test-statistic values for the joint response test, with $p - value = 0 \leq 0.01$.

of action selection probabilities exhibited by the simulation population of $n = 750$. The red line and boxes represent the mean action selection probability and the ±2 SE range. Jitter plot means have not been normalised but provide sufficient indication of probabilistic effect.

The CSUD discovered solution at exploration temperature $\tau = 225$, in Fig. 6.13c shows that for normal behaviour, mean good deck (C and D) card selection probabilities are above 0.25, and mean bad deck (A and B) card selection probabilities are below 0.25. For vmPFC impaired behaviour, as predicted in the re-shuffled IGT environment, good deck selection probabilities are above 0.25; but in the original IGT environment, bad decks A and B exhibit selection probabilities above 0.25 (thereby increasing the probability of a normative fail). At $\tau = 75$, as Fig. 6.13b shows, the action selection probabilities of good decks are further increased. However, at $\tau = 5$, Fig. 6.13a reveals a shift in behaviour, where good deck selection probability is predominantly attributed to deck C. This may be because deck D produces regular low rewards and seldom very high fines (1/25 chance), while deck C produces regular low rewards with occasional low fines (1/5 chance). Therefore probabilistically speaking, at low exploration, the agent may treat deck D as if it were a bad deck. In general, for normal behaviour



(B) $\alpha_1 = 0.364$, $\tau = 75$

(C) $\alpha_1 = 0.364$, $\tau = 225$

(D) Legend

FIGURE 6.13: Boltzmann agent exploration temperature τ and action selection probabilities at IGT completion. Simulation sample size $n = 750$. As exploration temperature increases from $\tau = 5$ to $\tau = 225$, normal behaviour mean action selection probabilities shift towards 0.25, the random search probability.

configuration, as exploration temperature τ increases, deck selection probabilities approach 0.25, the random search threshold. For vmPFC impaired behaviour, however, deck sequencing effects appear to influence deck selection probabilities differently. For example, at low exploration temperature $\tau = 5$, vmPFC impaired agent mean action selection probabilities for decks C and D are above 0.25. However, this is not the case at $\tau = 75$ or $\tau = 225$. In contrast with the re-shuffled deck, vmPFC impaired agents on average continue to select from the good decks as exploration temperature τ increases.

6.4.1 Boltzmann Agent Discussion

The Boltzmann agent produces mixed results.

As a decision making entity, the Boltzmann agent's main contribution is to make a probabilistic action selection, where actions which are valued more highly, have a higher probability of being chosen. In this sense, the Boltzmann agent is considered to be a rational agent and an efficient explorer: actions which are valued more highly have a higher probability of being chosen, and exploration is in proportion to the value of an action. Thus exploration is probabilistically geared towards selections, which have the higher aggregated Q-values. As long as the learning rate has an appreciable effect, Q-value updates produce changes in action selection probabilities. Since action selection probabilities are normalized, the Boltzmann agent tends to exploit statistical central tendencies. Because of its emphasis on central tendencies, the Boltzmann agent would be more appropriately called an efficient exploiter than an efficient explorer.

The IGT performance of the simple Boltzmann agent partially replicates human IGT outcomes, but also raises some questions as to whether the Boltzmann agent can adequately capture human decision making. As noted in Fig. 6.8 the Boltzmann agent is unable to simultaneously match all tested IGT environments and behaviours, only matching original and re-shuffled outcomes for normal and vmPFC impaired configurations, while missing normal random IGT outcomes. Further, for the re-shuffled IGT environment, while normal and vmPFC impaired configurations produce matches, the np-M/ANOVA analysis in Table 6.7 shows that the normal λ_N and the vmPFC impaired λ_{vmPFC} learning decay rates produce mean fraction of good decks $\bar{f}_G$ values, where the null hypothesis of equal outcomes is rejected. The a priori expectation however is that in statistical terms, this null hypothesis should fail to be rejected.

Despite these variations, as Fig. 6.12 indicates, the Boltzmann agent in general produces results where high learning rate decay, that is modelled vmPFC impairment, leads to normative fails in the original and random IGT environments. Also as Fig. 6.11 shows, the exploration index (EI) presentation of the Boltzmann agent appears qualitatively similar to those produced by human outcomes.

Finally, a CSUD limitation is noted, which limitation is a known issue in simultaneous perturbations stochastic approximation (SPSA). From Fig. 6.7, note that exploration τ varies very little. While this can result from a flat gradient, in this case, the result obtains from a discrepancy between the scale of perturbations and that of the temperature parameter, which in relation to the remaining parameters has a much larger range, reaching from 0.5 to 500 (see Table 5.1). In contrast, remaining parameters approximately range over the unit interval. Spall (2003, Ch. 7, p. 189) has suggested such a scale issue can be dealt with by remapping so that all parameters have similar ranges. Here the approach of using a constant perturbation scaling vector has been employed. Such scaling retains standard SPSA asymptotic guarantees.[2] However, in practice increasing the scale of perturbations can reduce the gradient even further, since scaled perturbations would appear both in the numerator and denominator of (12.13). Based on 2D and 3D grid search plots, it is believed that the reduced movement in temperature τ does not impact adversely on the presented results.

6.5 Adaptive ε-Greedy Agent Results

The adaptive ε-Greedy agent has been introduced by Tokic (2010), and is discussed in section 5.2.3. The adaptive ε-Greedy agent provides a good platform for testing exploitation versus exploration effects. Unlike the constant exploration ε-Greedy agent or the proportionate exploration Boltzmann agent, the adaptive ε-Greedy agent formulates exploration, which responds to the temporal difference error. Consequently exploration can vary from 0 (no exploration) to 1 (full exploration) in response to selected action outcomes.

Fig. 6.14 and Table 6.8 present adaptive ε-Greedy agent CSUD search results in graphic and tabular forms respectively. Note that the Tokic hyper-parametrisation requires an initial exploration value, which is set to $\varepsilon_1 = 1$. Further, (5.6c) requires a mixing hyper-parameter β, called influence, used for updating exploration. Influence β scales the contribution of exploration adjustment $f(a, \tau)$ and current exploration ε_t to next period's exploration ε_{t+1}. Following Tokic, $\beta = 1/norm(actions) = 0.25$ is used. This leaves the following four free hyper-parameters for estimation: initial learning rate α_1, normal learning decay λ_N, vmPFC impaired learning decay λ_{vmPFC}, and

[2] This can be shown by an extension of the proofs in chapter 12; but is not an extension discussed in this work.

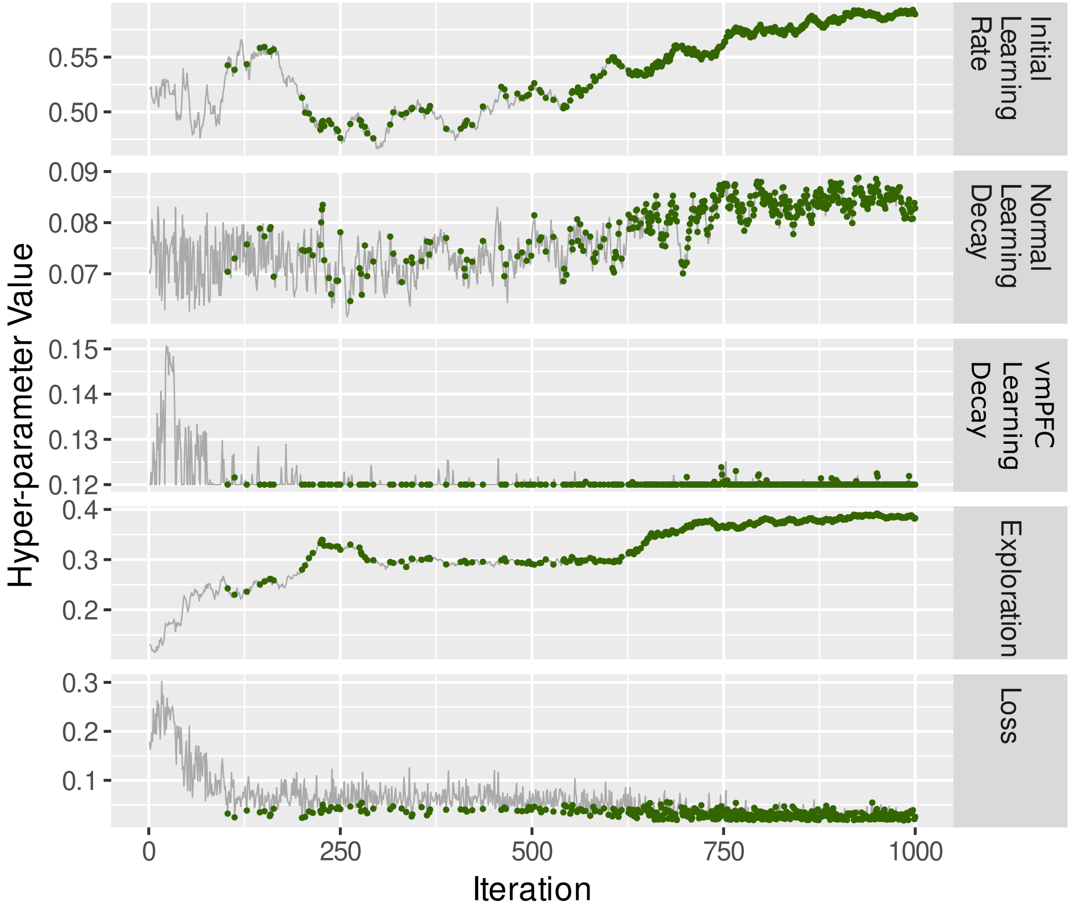


FIGURE 6.14: Adaptive ε-Greedy agent CSUD iterations. Green points indicate $(\alpha_1, \lambda_N, \lambda_{vmPFC}, \tau)$ hyper-parameter tuples, which produce full CSUD search matches. Many full CSUD search matches are obtained, increasingly so as search iterations advance.

exploration temperature τ. The CSUD minimum loss solution for these parameters is reported in Table 6.8.

Fig. 6.14 indicates that over 1000 CSUD iterations, full matches (i.e., the green dots) increase towards later iterations. Hence CSUD traversal of the hyper-parameter space is indeed proceeding in the direction of minimising loss.

Table 6.9 presents the CSUD grid search verification configuration. Fig. 6.15 and Fig. 6.16 present the 2D and 3D contours respectively generated from a localised grid search around the CSUD discovered hyper-parameters as indicated in Table 6.9.

	Minimum Loss	Range	Mean	Median	Standard Error
Loss	0.0180	0.0180 - 0.0552	0.0313	0.0314	0.000419
Initial learning rate α_1	0.591	0.476 - 0.593	0.559	0.572	0.00150
Normal learning rate decay λ_N	0.082	0.0647 - 0.0888	0.0813	0.0827	0.000245
vmPFC impaired learning rate decay λ_{vmPFC}	0.120	0.12 - 0.124	0.120	0.12	0.0000196
Temperature τ	0.383	0.230 - 0.392	0.356	0.373	0.00174
Matched environments	For normal human behaviour: original, re-shuffled, random. For vmPFC impaired human behaviour: original, re-shuffled.				
Match count	405 of 1000 iterations				

TABLE 6.8
Adaptive ε-Greedy agent CSUD minimum loss search matches after 1000 iterations. The highlighted minimum loss column shows selected agent hyper- parameters. Light gray indicates minimum loss and the associated initial learning rate α_1. Dark-gray, mid-gray, and light blue indicate minimum loss associated normal learning rate decay λ_N, vmPFC impaired learning rate decay λ_{vmPFC}, and exploration τ respectively.

Hyper-parameter	Grid Points
Initial learning rate α_1	0.01, 0.3, 0.591, 0.999
Learning rate decay λ^a	0.082, 0.120
Exploration Temperature τ	0.1, 0.383, 1, 24
IGT length	Q-learning samples
100	750

[a] The learning rate decay grid is constructed around the CSUD reported values. Appendix C provides the construction method.

TABLE 6.9
Adaptive ε-Greedy agent CSUD verification. Hyper-parameter grid search criteria for joint original, re-shuffled, and random IGT.

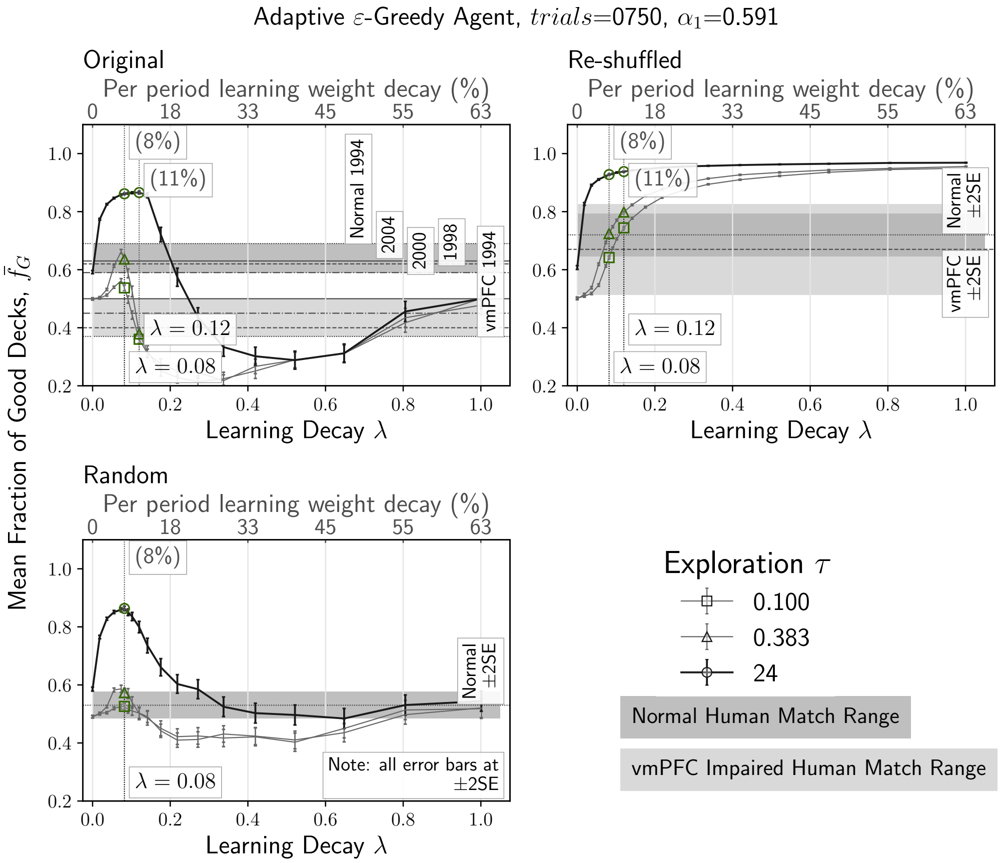


FIGURE 6.15: Adaptive ε-Greedy agent 2D contours showing learning decay λ and exploration τ effects. The dark and light gray zones indicate normal and vmPFC impaired human outcome match ranges for the original, re-shuffled, and random IGT environments. Learning decay variation reproduces human IGT outcomes, while exploration variation leads to contour shifts.

From Fig. 6.15, note that as in the case of the ε-Greedy and Boltzmann agents, learning rate decay λ is the key hyper-parameter for inducing normal versus vmPFC impaired IGT behaviour. As learning rate decay increases, exploration contours shift from the dark gray normal match zone through to the light gray vmPFC impaired behaviour match zone. Similarly changes in exploration temperature produce a vertical shift in the 2D mean fraction of good deck $\bar{f}_G$ contours.

For the adaptive ε-Greedy agent, initial exploration $\varepsilon_1 = 1$ and this indicates random search. Exploration temperature τ determines how quickly the agent will depart from random search. For low τ, the agent remains closer to random search, and for high τ, the agent becomes an increasingly specific exploiter. At the CSUD discovered exploration temperature

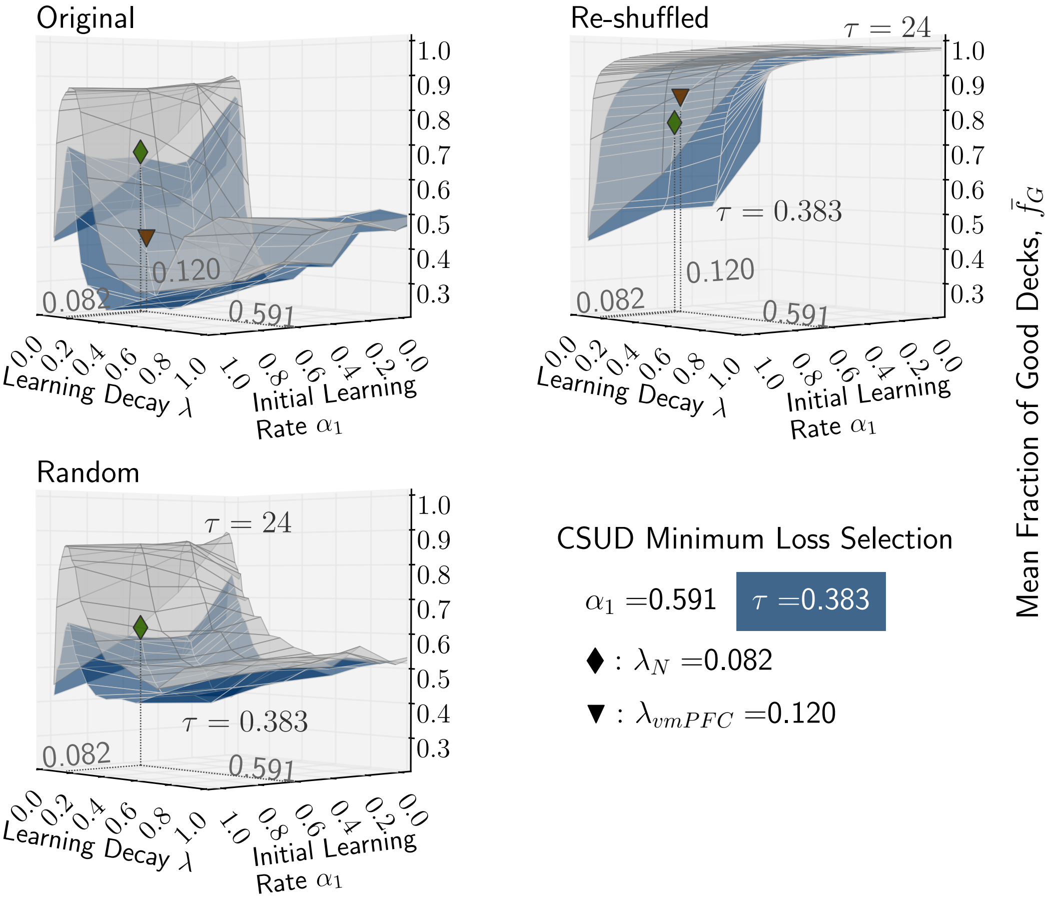


FIGURE 6.16: Adaptive ε-Greedy agent 3D contours with learning decay λ and exploration τ, but focusing on initial learning rate α_1 effects. Learning decay and exploration variation mirror the contours in Fig. 6.15. The initial learning rate α_1 shows a technical effect at very high initial learning rates, but otherwise exerts no determining effect.

$\tau = 0.383$, the black contour line with the green triangles indicates within range matches for the normal original, normal re-shuffled, vmPFC impaired re-shuffled, and on-border matches for vmPFC impaired original, normal random IGT behaviour and environment combinations. Finally note that as learning rate decay increases, the transition from normal to vmPFC impaired behaviour is very sudden, with $\lambda_N = 0.082$ and $\lambda_{vmPFC} = 0.120$.

The 3D visualisation in Fig. 6.16 indicates that mean fraction of good decks $\bar{f}_G$ surfaces exhibit increased complexity when compared with the corresponding ε-Greedy and Boltzmann figures 6.8 and 6.9 respectively.

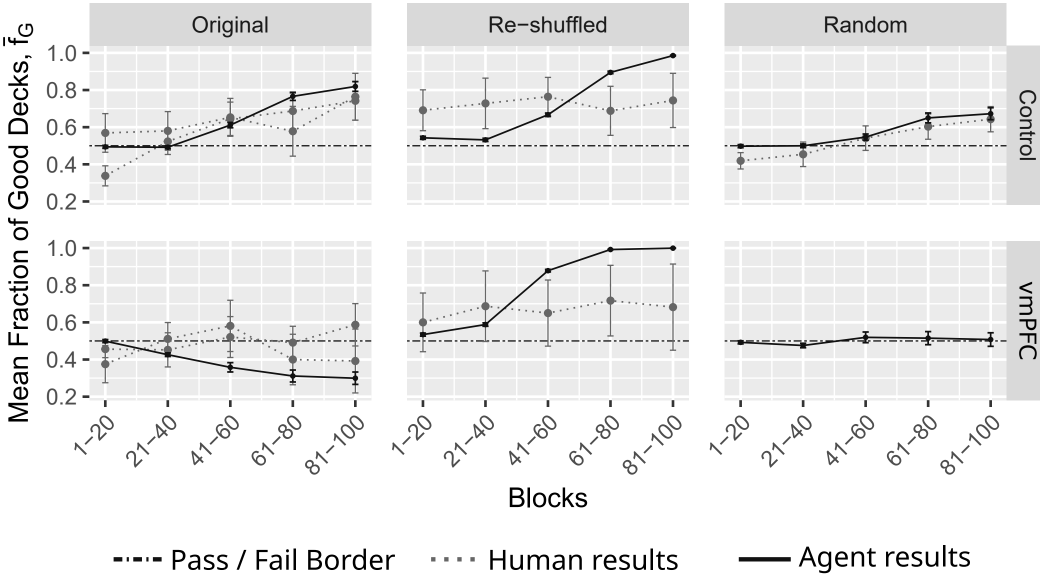


FIGURE 6.17: Adaptive ε-Greedy agent 20-draw blocks comparison at CSUD search matches $\alpha_1 = 0.591, \lambda_N = 0.082, \lambda_{vmPFC} = 0.120, \tau = 0.383$. Human results in light gray dotted lines.[3] Agent results in dark gray solid lines, averaged from 750 samples. All error bars at $\pm 2SE$. Even when error bars are taken into account, agent and human 20-draw block performances generally differ.

However, the influence of the initial learning rate α_1 remains relatively modest, except for an increase in the initial non-stationarity (the draped-over surface) zone.

Fig. 6.17 and Fig. 6.18 demonstrate that human versus adaptive ε-Greedy agent comparative 20-draw blocked results show differences from human behaviour. In particular, agent behaviour is more exploitative than corresponding human behaviour. This is most visible in Fig. 6.18 for the re-shuffled IGT, where the agent exploration index (EI) decreases markedly from IGT draw 40 onwards.

Specifically with normal behaviour configuration, Fig. 6.17 shows that the adaptive ε-Greedy agent achieves higher mean fraction of good decks $\bar{f}_G$ in the final two draw blocks, 61-80 and 81-100, across all IGT environments except for the vmPFC impaired original case, with the difference being most pronounced for the re-shuffled IGT case. When ±2 SE bars are taken into account, in the re-shuffled environment, mean fraction of good decks $\bar{f}_G$ outcomes for blocks 61-80 and 81-100 lie outside the SE catchment areas.

[3] As a reminder human results in are initially presented in Fig. 4.1 and Fig. 4.2.

[4] As a reminder human results in are initially presented in Fig. 4.3 and Fig. 4.4.

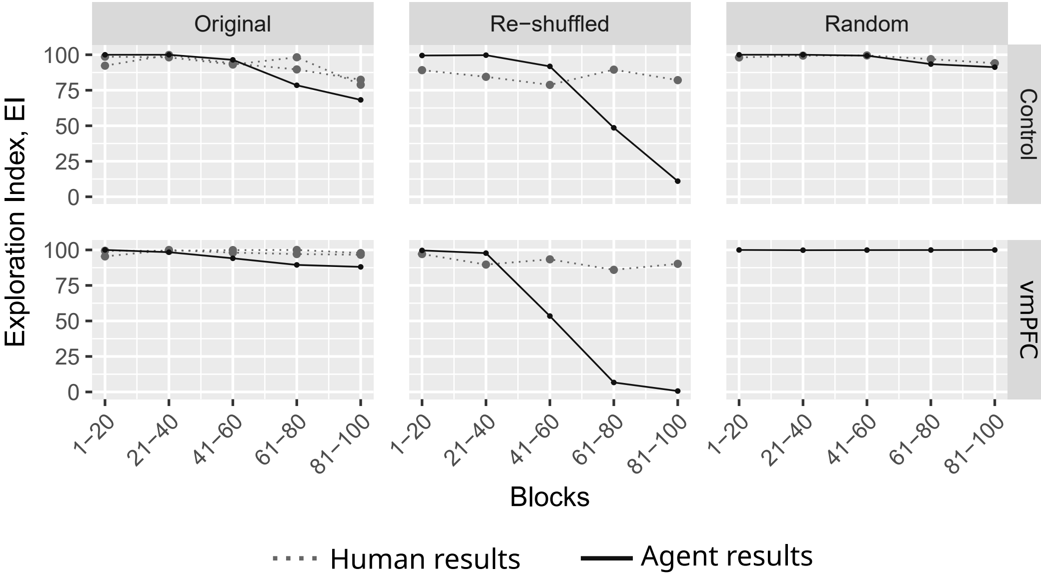


FIGURE 6.18: Adaptive ε-Greedy agent 20-draw blocks exploration index (EI) comparison at CSUD search matches $\alpha_1 = 0.591, \lambda_N = 0.082, \lambda_{vmPFC} = 0.120, \tau = 0.383$. Human results in dotted light gray lines.[4]Agent results in solid dark gray lines, averaged from 750 samples. Human subject and agent exploration index responses show comparative differences especially for the re-shuffled IGT environment. Details in text below.

Fig. 6.18 indicates that the design of the adaptive ε-Greedy agent enables the agent to drive down exploration. With the exception of the normal behaviour random IGT environment case, agent exploration index (EI) is substantially lower in draw blocks 61-80 and 81-100, noticeably so for the re-shuffled IGT environment, where EI decreases to approximately 10 and 0 for the normal and vmPFC impaired configurations respectively, indicating very high exploitation. Compared to ε-Greedy agent figures 6.4 and 6.5; and Boltzmann agent figures 6.10 and 6.11, it is noted that the corresponding adaptive ε-Greedy agent results show large variations from the comparable human benchmark results.

Fig. 6.19 provides a jitter plot density summary for the fraction of good decks f_G outcomes obtained at the CSUD selected minimum-loss hyperparameter values for 750 samples of the original, re-shuffled, and random IGT environments for normal (control) and vmPFC impaired configurations. Green dots mark outcomes inside human performance ranges. Blue dots mark additional normative pass results, whereas red dots mark additional normative pass fails. Green numbers give total matches out of 750

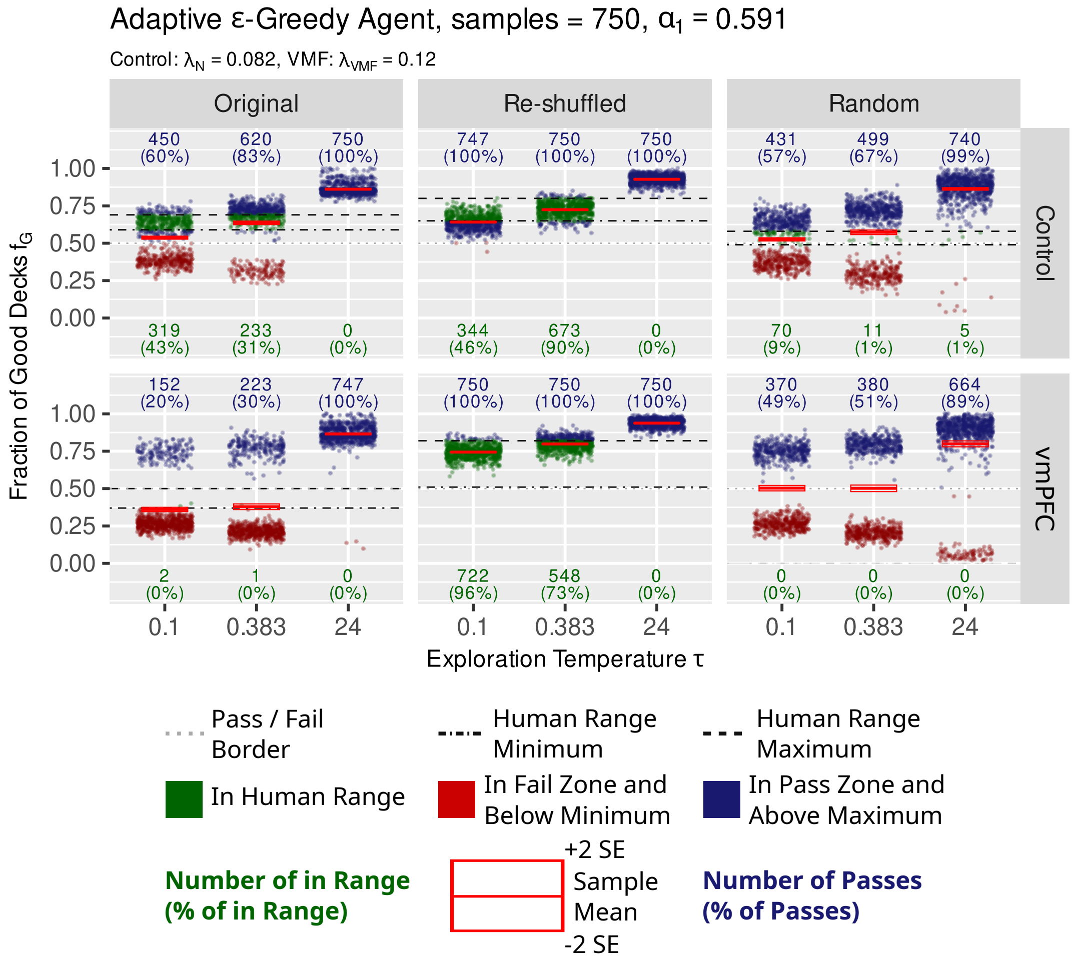


FIGURE 6.19: Adaptive ε-Greedy agent comparison of repeated simulation outcomes to human IGT results. At τ = 0.383 and reported CSUD minimum loss hyper-parameter values, the adaptive ε-Greedy agent achieves the highest matches for the re-shuffled IGT environment and otherwise does not perform well. Full details are in the text.

samples, and the values in brackets indicate percentages matched. A high percentage matched value is indicative of predictive simulation success. Finally the red bars and the box indicate central tendency in terms of the mean and ±2 SEs (standard errors).

Fig. 6.19 indicates that at exploration temperature $\tau = 0.383$, adaptive ε-Greedy agent simulations achieve the best results in the re-shuffled IGT, with 90% and 73% matches for the normal and vmPFC impaired configurations respectively. In the remaining comparable original normal and vmPFC impaired, and random normal configurations, the adaptive ε-Greedy agent does not perform well. This low simulation fidelity is due to a tendency

Test Variant	Test Statistic	df1	df2	p-Value	Subset Results
Original ∣ Re-Shuffled ∣ Random vs. Learning Decay λ					
ANOVA Type[a]	372.438	2.861	4286.479	0	At α = 0.01, the null hypotheses of learning decay factor equality is rejected. Only equality of the random response cannot be rejected.
Original ∣ Re-shuffled vs. Learning Decay λ					
ANOVA Type[a]	635.344	1.942	2908.666	0	At α = 0.01, the null hypotheses of equal original and re-shuffled, original only, and re-shuffled only responses are rejected.
Random vs. Learning Decay λ					
ANOVA Type	2.916	1.000	1498	0.088	Single response variable, no subsets.
Wilks Lambda	2.916	1.000	1498	0.088	

[a]Wilks Lambda could not be computed due a singular rank matrix.

TABLE 6.10
Adaptive ε-Greedy agent np-M/ANOVA analysis of mean fraction of good decks $\bar{f}_G$ with learning decay λ as factor. At significance level $\alpha = 0.01$, mean fraction of good decks $\bar{f}_G$ responses are statistically significantly different, *even* for the re-shuffled IGT environment.

towards a bimodal outcome distribution with low or no density within the respective human performance ranges.

Table 6.10 shows the results of the np-M/ANOVA analysis assessing the effect of learning decay as a factor in contributing to normal versus vmPFC impaired behaviour. The joint multi-variate response of original, re-shuffled, and random mean fraction of good decks $\bar{f}_G$ outcomes to learning decay as a factor is significant at significance level $\alpha = 0.01$. The joint multi-variate response of original and re-shuffled, univariate original, and univariate re-shuffled fraction of good decks $\bar{f}_G$ outcomes also exhibit significantly different factor responses at significance level $\alpha = 0.01$. In contrast, the univariate random IGT environment response to learning decay as a factor does not appear to be significantly different at $\alpha = 0.01$, with univariate random vs. learning decay np-M/ANOVA revealing a p-value of 0.088.

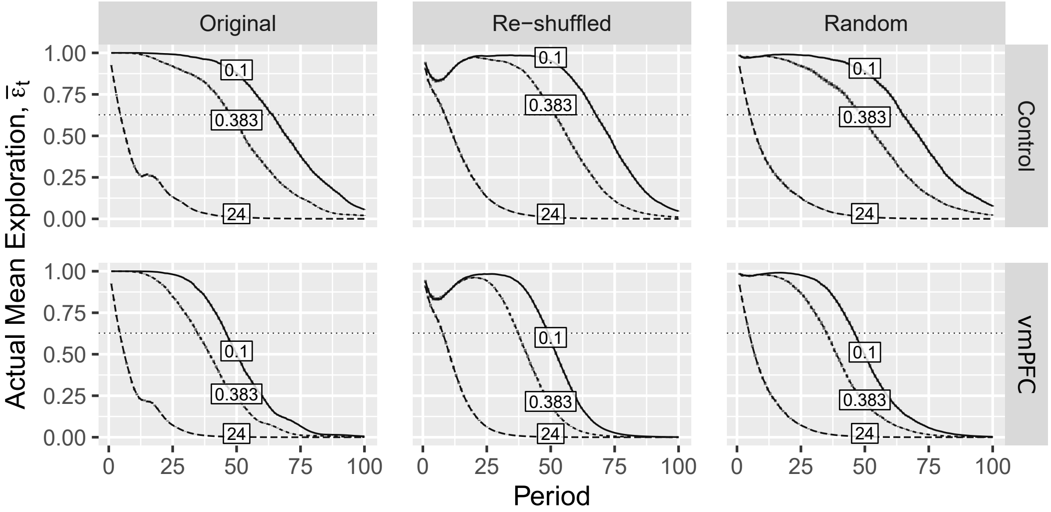


FIGURE 6.20: Adaptive ε-Greedy agent 2D contours showing exploration temperature τ effects. In general, as τ increases, exploitation increases at a faster rate. The dotted line marks ε = 0.627, the CSUD discovered minimum-loss exploration value for the ε-Greedy agent results presented in section 6.3. The stepwise appearance are due to the presence of small $\pm 2SE$ bands, indicating low sample variation.

As it was the case with the Boltzmann agent, the adaptive ε-Greedy agent np-M/ANOVA results reject the expected null hypothesis of no learning rate decay λ factor effect for the re-shuffled IGT environment. However, based on human outcome data it is expected that for the re-shuffled IGT environment, the null hypothesis of no learning rate decay λ factor effect would have been failed to be rejected. That is for the re-shuffled IGT environment, the learning rate decay factor settings of λ_N = 0.082 and λ_{vmPFC} = 0.12 should not have produced at significance level α = 0.01, a statistically significant mean fraction of good decks $\bar{f}_G$ effect.

Finally, Fig. 6.20 presents exploration temperature τ contours for actual mean exploration $\bar{\varepsilon}$ achieved over the duration of IGT tasks. Per period mean exploration $\bar{\varepsilon}_t$ is computed from 750 simulation samples for each IGT environment and behaviour configuration. The plot includes ±2 SE error bars. However, at 750 samples, the magnitude of the error bars is relatively modest.

In general, as τ increases, exploitation increases. The dotted line marks ε = 0.627, the CSUD discovered minimum-loss exploration value for the ε-Greedy agent results presented in section 6.3. Hence the dotted line helps to contextualise per period mean exploration $\bar{\varepsilon}_t$ versus constant exploration ε.

Fig. 6.20 shows that at the CSUD discovered minimum-loss exploration temperature $\tau = 0.383$, for normal (control) cases, per period mean exploration $\bar{\varepsilon}_t$ is above constant exploration $\varepsilon = 0.627$ for approximately the first half (50 periods) of the IGT task, thereafter decaying rapidly towards 0 indicating a strong shift towards exploitation in the second 50 periods. For vmPFC impaired behaviour, however, exploration decays more rapidly, and exploitation relative to the constant exploration mark starts approximately by period 37. In relation to the re-shuffled IGT, and to a smaller extent the random IGT environments, note that periods up to 25 exhibit regions of increasing per period mean exploration $\bar{\varepsilon}_t$.

6.5.1 Adaptive ε-Greedy Agent Discussion

The adaptive ε-Greedy agent is theoretically interesting because it is able to increase and decrease exploration in response to the temporal difference error. In theory, this behaviour should allow the agent to swiftly shift from exploration to exploitation, while also being resistant to proportional exploration (Boltzmann agent) induced central tendency focused vision.

With CSUD searches using the mean fraction of good decks $\bar{f}_G$, measured cumulatively at the end of the IGT at the 100th draw, the adaptive ε-Greedy agent achieves 405 full environment and behaviour matches in 1000 search iterations. That is, CSUD produces 405 agent hyper-parametrisations, which produce agent outcomes residing within respective human outcome catchment zones for normal original, re-shuffled, random, and vmPFC impaired original and re-shuffled IGT environments.

This high number of full CSUD matches initially looks promising, however, the 20-draw blocked mean fraction of good decks performance of the agent indicates large differences from corresponding human outcomes. Compared to human outcomes, agent exploration index (EI) exhibits differing per 20-draw block exploration decay patterns. The theoretical innovation, which allows the agent to adjust exploration, also leads to differences from human outcomes. This suggests that regarding the IGT, if humans do decrease exploration in response to learning, they do not do this using the adaptive ε-Greedy algorithm.

In general, the adaptive ε-Greedy agent does well with solving IGT tasks. As discussed above however, the adaptive ε-Greedy agent results display characteristics, which are quite different from those displayed by human benchmark data. Additionally, this agent is more sensitive to learning rate decay than either the ε-Greedy or Boltzmann agent. In summary, the adaptive ε-Greedy agent appears to be an unlikely candidate for modelling human IGT behaviour.

Note that from the exploration index (EI) results in Fig. 6.18 unlike human benchmarks, the adaptive ε-Greedy agent is quite successful at decreasing exploration. The per period mean exploration $\bar{\varepsilon}_t$ results in Fig. 6.20 support this finding. The original IGT environment case in Fig. 6.15 shows that, when it comes to learning rate decay λ, the adaptive ε-Greedy agent appears to exhibit a tipping point beyond which agent performance in terms of mean fraction of good decks $\bar{f}_G$ rapidly degrades.

When learning rate decay is present, the second term in (5.6b), which is $b = |\alpha_t \left(x_t^a - Q_{t-1}(a)\right)|$ can become very small, even when the temporal difference error $x_t^a - Q_{t-1}(a)$ is large. According to (5.6c), this in turn leads to small and decreasing per period exploration ε_t.

The adaptive ε-Greedy agent has a hard time dealing with decreasing learning rates, and may work best with a constant learning rate, a decision making case, which is not discussed here. Learning rate decay λ periodically decreases the initial learning rate α_1, and this leads to weakening of the exploration adjustment signal. Hence, high learning rate decay contributes to the agent acting as if it has completed learning, and subsequent decreases in exploration amplify exploitation. If the agent learns the correct solution, it may surpass per block human performance. But if the agent has learned the incorrect solution, then as indicated in the vmPFC impaired original IGT case of Fig. 6.17, the agent produces worse than human results.

Finally, the adaptive ε-Greedy agent np-M/ANOVA results, like the Boltzmann agent results, reject for the re-shuffled environment the null hypothesis of no learning rate decay λ factor effects.

In summary, while of strong theoretical value, it is not believed that the adaptive ε-Greedy agent can reflect human IGT behaviour. In the rest of this work, this agent will not be considered any further.

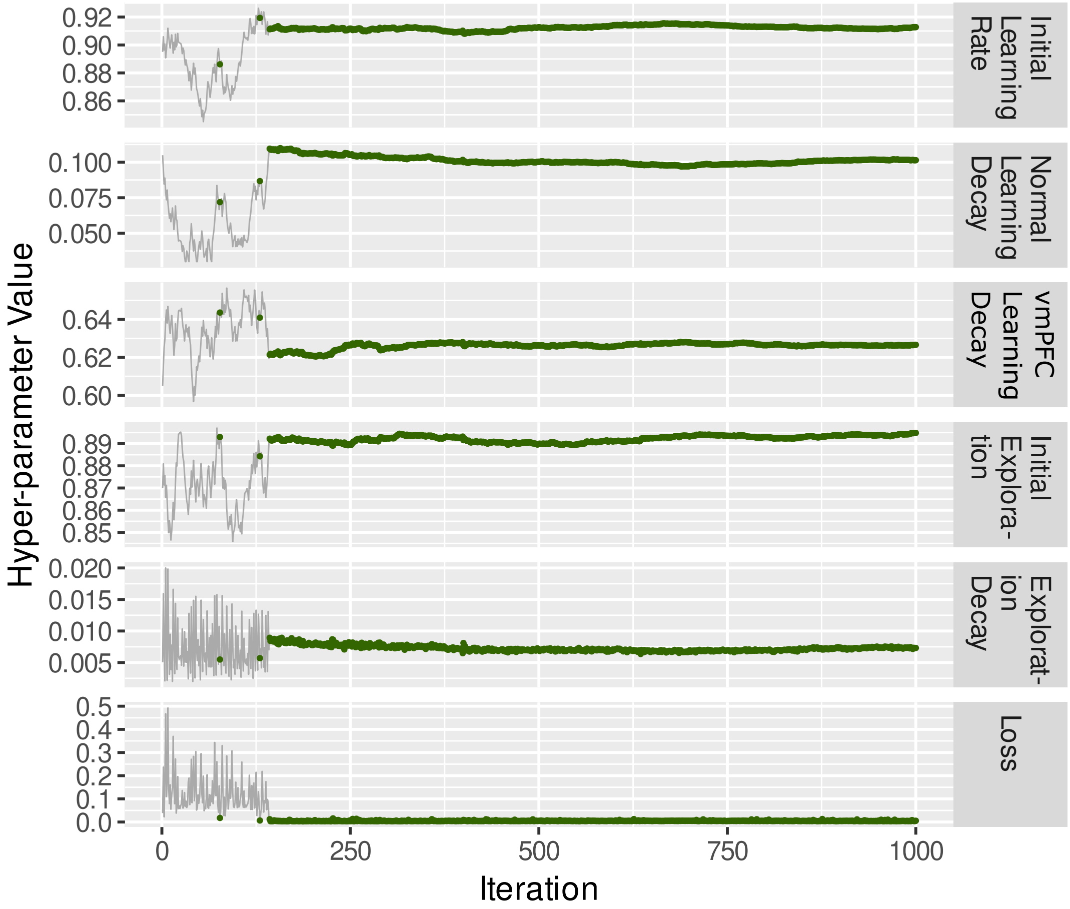


FIGURE 6.21: Decaying ε-Greedy agent CSUD iterations. Green points indicate $(\alpha_1, \lambda_N, \lambda_{vmPFC}, \varepsilon_1, \nu)$ hyper-parameter tuples, which produce full CSUD search matches.

6.6 Decaying ε-Greedy Agent Results

Section 5.2.4 has introduced the decaying ε-Greedy agent formulation. The decaying ε-Greedy agent exhibits heuristic exploration decay by using, just like learning rate decay λ, an exponential decay paradigm. Decaying ε-Greedy agent hyper-parameters consist of the initial learning rate α_1, normal learning decay λ_N, vmPFC impaired learning decay λ_{vmPFC}, initial exploration ε_1, and exploration decay ν. Note that the same exploration decay ν value is used in both normal and vmPFC impaired behaviours.

Fig. 6.21 and Table 6.11 present the CSUD search results. With 5 repeated gradient samples, the decaying ε-Greedy agent CSUD hyper-parameter search converges relatively quickly given the 1000 iteration search budget,

	Minimum Loss	Range	Mean	Median	Standard Error
Loss	0.00276	0.00276 - 0.0176	0.00506	0.00474	$5.21e^{-5}$
Initial learning rate α_1	0.911	0.886 - 0.919	0.912	0.912	$6.00e^{-5}$
Normal learning rate decay λ_N	0.106	0.0719 - 0.110	0.101	0.100	$1.02e^{-4}$
vmPFC impaired learning rate decay λ_{vmPFC}	0.622	0.620 - 0.644	0.626	0.626	$6.13e^{-5}$
Initial Exploration ε_1	0.891	0.884 - 0.895	0.892	0.892	$4.96e^{-5}$
Exploration Decay ν	0.00842	0.00548 - 0.00893	0.00724	0.00713	$1.50e^{-5}$
Matched environments	For normal human behaviour: original, re-shuffled, random. For vmPFC impaired human behaviour: original, re-shuffled.				
Match count	860 of 1000 iterations				

TABLE 6.11
Decaying ε-Greedy agent CSUD minimum loss search matches after 1000 iterations. The highlighted minimum loss column shows selected agent hyper- parameters. Light gray indicates minimum loss and the associated initial learning rate α_1. Dark-gray, mid-gray indicate minimum loss associated normal learning rate decay λ_N and vmPFC impaired learning rate decay λ_{vmPFC} respectively. Light blue highlights minimum loss initial explocation ε_1 and exploration decay ν.

and achieves 860 full search matches. In Fig. 6.21 green points indicate initial learning rate, normal learning decay, vmPFC impaired learning decay, initial exploration, and exploration decay, that is $(\alpha_1, \lambda_N, \lambda_{vmPFC}, \varepsilon_1, \nu)$ hyper-parameter tuples, which produce agent performance matching mean fraction of good decks $\bar{f}_G$ outcomes for the normal behaviour configuration original, re-shuffled, and random; and vmPFC impaired configuration original and re-shuffled IGT environments.

Table 6.12, Fig. 6.22 and Fig. 6.23 present CSUD grid search verification configuration, 2D, and 3D contour mean fraction of good deck $\bar{f}_G$ plots respectively.

At initial learning rate $\alpha_1 = 0.911$ and initial exploration $\varepsilon_1 = 0.891$, Fig. 6.22 presents 2D mean fraction of good deck $\bar{f}_G$ contours obtained at four distinct exploration decay ν values. As previously, the dark and light gray zones represent normal and vmPFC impaired match zones respectively. The derivation of the match zones is discussed in Table 4.8. The

Hyper-parameter	Grid Points
Initial learning rate α_1	0.1, 0.33, 0.66, 0.911
Learning rate decay λ[a]	0.106, 0.622
Initial Exploration ε_1	0.25, 0.5, 0.891, 1
Exploration Decay ν	0.002, 0.00842, 0.015, 0.05
IGT length	Q-learning samples
100	750

[a] The learning rate decay and exploration decay grids are constructed around the two values above. Appendix C provides the construction method.

TABLE 6.12
Decaying ε-Greedy agent CSUD verification. Hyper-parameter grid search criteria for joint original, re-shuffled, and random IGT.

black coloured line represents the CSUD minimum-loss exploration decay at $\nu = 0.00842$. The legend on the right additionally indicates that when $\nu = 0.00842$, then at the IGT termination, final exploration is at 0.3871.

It is noted that in Fig. 6.22 at tuple ($\alpha_1 = 0.911, \lambda_N = 0.106, \lambda_{vmPFC} = 0.622, \varepsilon_1 = 0.891, \nu = 0.00842$), the black coloured line exhibits matches in all of the marked dark and light gray zones, indicating a full match to human IGT outcomes in all cases. Further it is noted that as observed in the (constant exploration) ε-Greedy and Boltzman agent results, matching human outcomes once more requires high exploration; here, in the form of high initial exploration ε_1.

At initial learning rate $\alpha_1 = 0.911$, Fig. 6.23 illustrates 3D mean fraction of good deck $\bar{f}_G$ contours obtained at initial exploration values $\varepsilon_1 = 0.25$ and $\varepsilon_1 = 0.891$. Fig. 6.23 assesses learning decay λ and exploration decay ν interactions, while holding initial learning rate constant at $\alpha_1 = 0.911$. The blue coloured 3D contour marks the CSUD discovered initial exploration value $\varepsilon_1 = 0.891$. The green diamond and red inverted triangular shapes mark CSUD discovered normal and vmPFC impaired learning decay rates respectively.

The 3D mean fraction of good decks $\bar{f}_G$ surfaces reveal that learning rate decay continues to induce normal versus vmPFC impaired behaviour. Exploration decay ν itself produces some effects, which depend on initial exploration ε_1 and the specific IGT environment. Increasing exploration decay

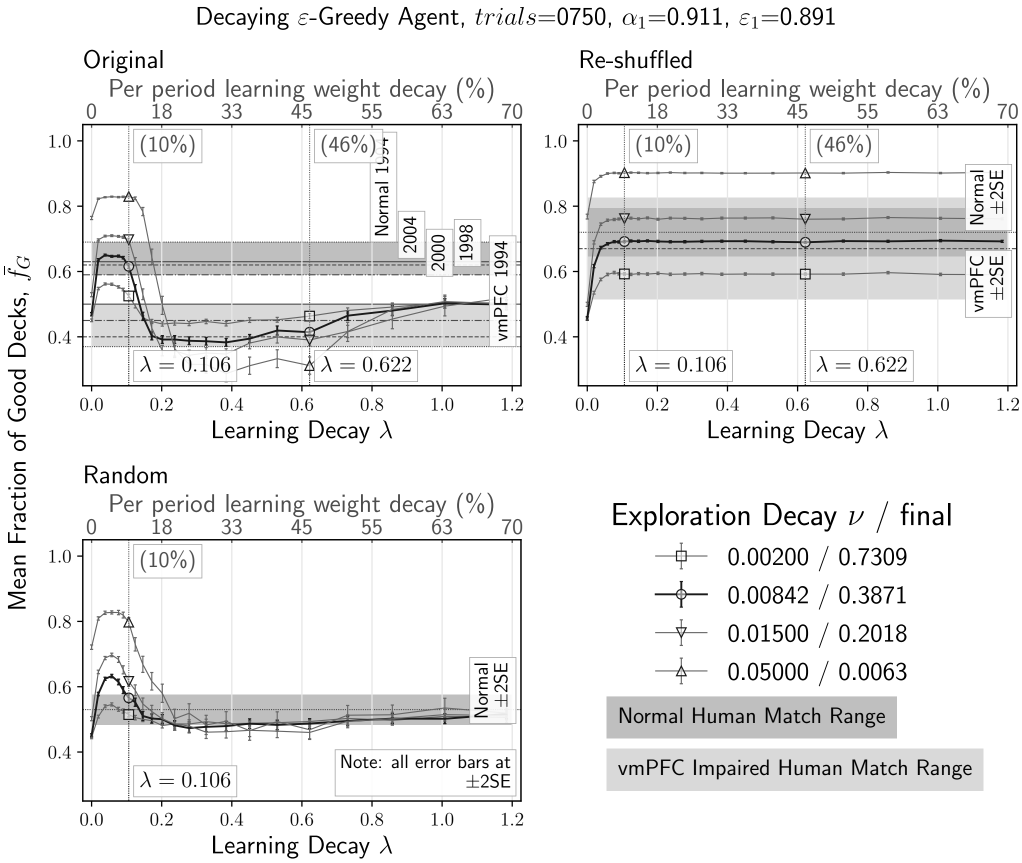


FIGURE 6.22: Decaying ε-Greedy agent 2D contours showing learning decay λ and exploration decay ν effects at $\alpha_1 = 0.911$ and $\varepsilon_1 = 0.891$. The dark and light gray zones indicate normal and vmPFC impaired human outcome match ranges for the original, re-shuffled, and random IGT environments. Learning decay variation reproduces human IGT outcomes, while increasing exploration decay produces upwards contour shifts due to exploration decreasing faster over time.

leads to an increase in mean fraction of good decks. This effect is more pronounced at higher initial exploration, and for the original and random IGT environments, at lower learning decay λ. The initial learning rate $\alpha_1 = 0.911$ is high, and an initial non-stationarity effect is notable, with low learning rate decay, towards the rear of the mean fraction of good decks $\bar{f}_G$ contours, in the area where learning decay λ is close to zero. In this non-stationary zone, due to the high initial learning rate, increasing learning rate decay initially leads to an increase in the mean fraction of good decks.

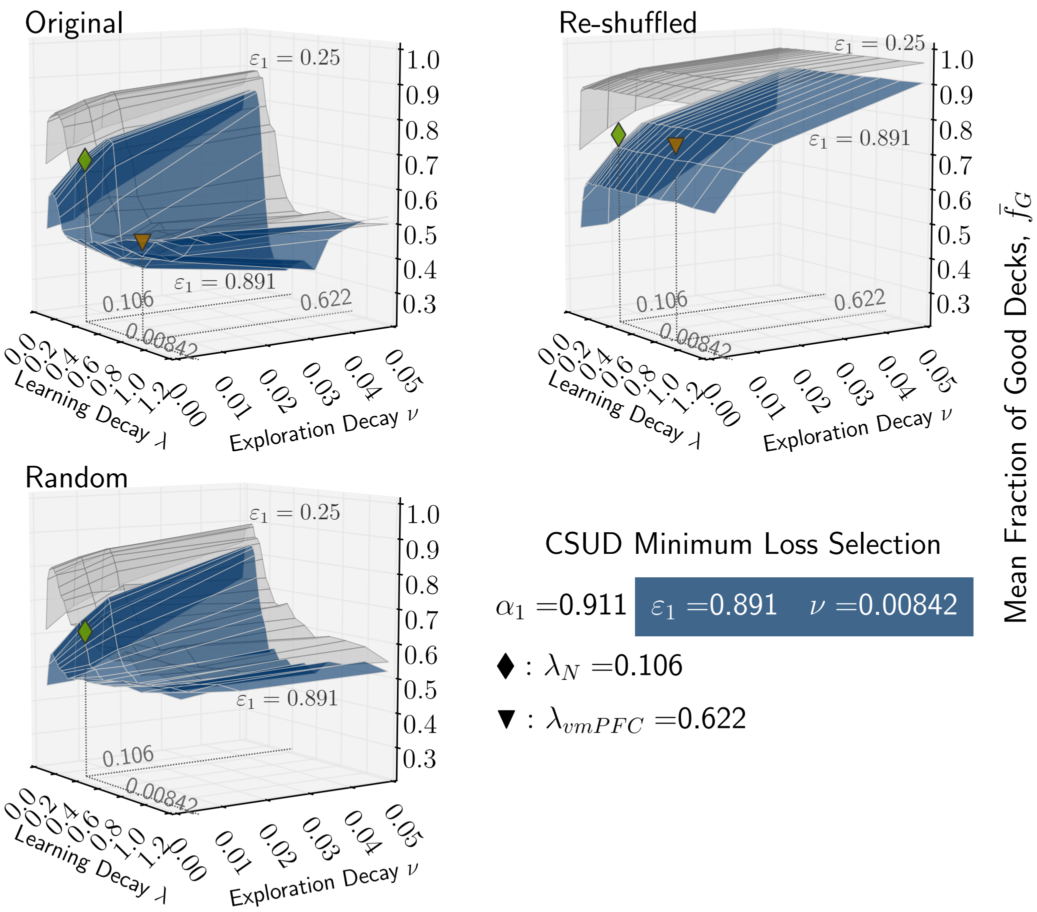


FIGURE 6.23: Decaying ε-Greedy agent 3D contours with learning decay λ, exploration decay ν, and initial exploration ε_1 at initial learning rate $\alpha_1 = 0.911$. Both learning decay and exploration decay show $\bar{f}_G$ influences in all IGT environments. However, learning decay shows a stronger effect.

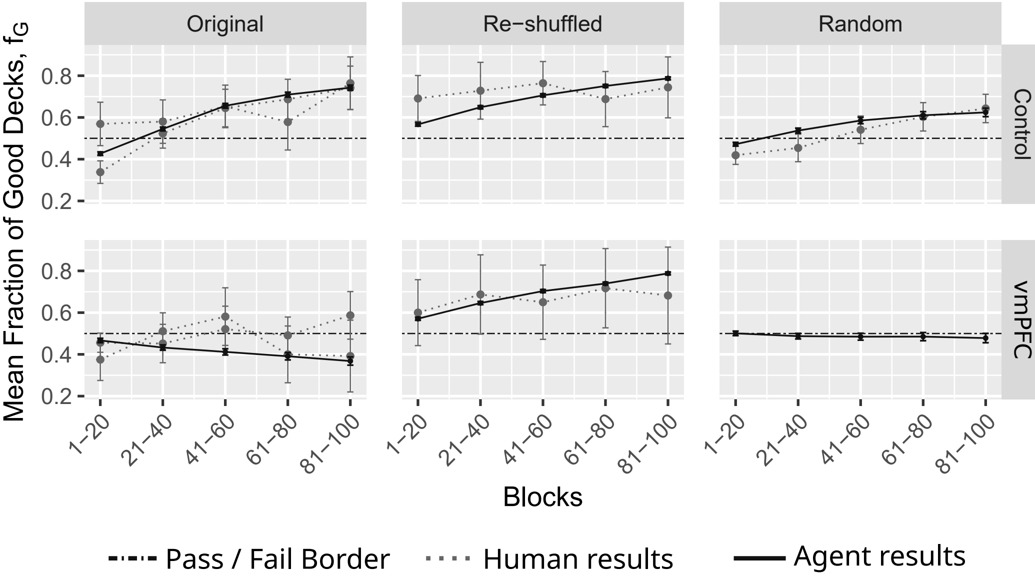


FIGURE 6.24: Decaying ε-Greedy agent 20-draw blocks comparison at CSUD search matches $\alpha_1 = 0.911, \lambda_N = 0.106, \lambda_{vmPFC} = 0.622, \varepsilon_1 = 0.891, \nu = 0.00842$. Human results in dotted light gray. Agent results in solid dark gray, averaged from 750 samples. All error bars at $\pm 2SE$. When error bars are taken into account agent and human 20-draw block performance appears relatively similar. Details in text below.

Fig. 6.24 and Fig. 6.25 present 20-draw block results for mean fraction of good decks $\bar{f}_G$ and the exploration index (EI) respectively. In Fig. 6.24, the dotted gray lines represent human benchmarks while the solid dark gray lines show agent results. The dash-dotted line indicates $\bar{f}_G = 0.50$, above which a normative pass is achieved. It is expected that as the IGT advances from block 1-20 towards 81-100, for normal behaviour, mean fraction of good decks $\bar{f}_G$ increases and then levels out. In contrast for vmPFC impaired behaviour, $\bar{f}_G$ increases and levels out for the re-shuffled case, while decreasing or staying the same for the original IGT environment. The decaying ε-Greedy agent when taking human $\bar{f}_G$ benchmark outcomes with ± 2 SEs into account, matches the trends exhibited for the human benchmarks in all behaviour and environment cases except for the vmPFC impaired re-shuffled case, where the agent displays a steeper increasing trend.

In Fig. 6.25 the dotted gray lines represent human benchmarks while the solid dark gray lines show agent results in relation to the exploration index (EI). The results mirror those in Fig. 6.24, that is, human benchmark

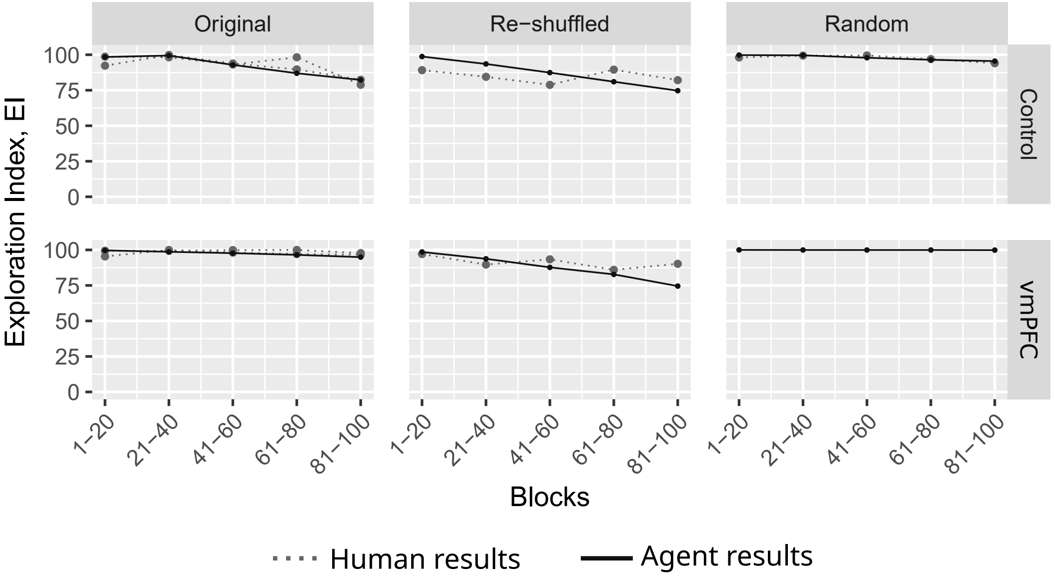


FIGURE 6.25: Decaying ε-Greedy agent 20-draw blocks exploration index (EI) comparison at CSUD search matches $\alpha_1 = 0.911, \lambda_N = 0.106, \lambda_{vmPFC} = 0.622, \varepsilon_1 = 0.891, \nu = 0.00842$. Human results in dotted light gray. Agent results in solid dark gray, averaged from 750 samples. Human subject and agent exploration index responses appear relatively similar except for vmPFC impaired behaviour in the re-shuffled IGT environment. Details in text below.

EI trends are matched well in all cases except for the vmPFC impaired re-shuffled case. This agent human outcome match discrepancy in the respective vmPFC impaired re-shuffled environment outcomes suggest that vmPFC impaired humans may not be exhibiting the exponential exploration decay heuristic. However, as reported in Table 4.3, the vmPFC impaired human population is very small ranging from 6 to 10 subjects. On the other hand, the agent population is at $n = 750$. Therefore, it would be difficult to make a definitive assessment as confidence bands have not been computed for the non-linear exploration index (EI) transforms.

Fig. 6.26 provides a jitter plot density summary for the fraction of good decks f_G outcomes obtained at the CSUD selected minimum-loss hyper-parameter values ($\alpha_1 = 0.911, \lambda_N = 0.106, \lambda_{vmPFC} = 0.622, \varepsilon_1 = 0.891$) for 750 samples of the original, re-shuffled, and random IGT environments for normal (control) and vmPFC impaired configurations. Green dots mark outcomes inside human performance ranges. Blue dots mark additional normative pass results, whereas red dots mark additional normative pass fails. Green numbers give total matches out of 750 samples, and the values in

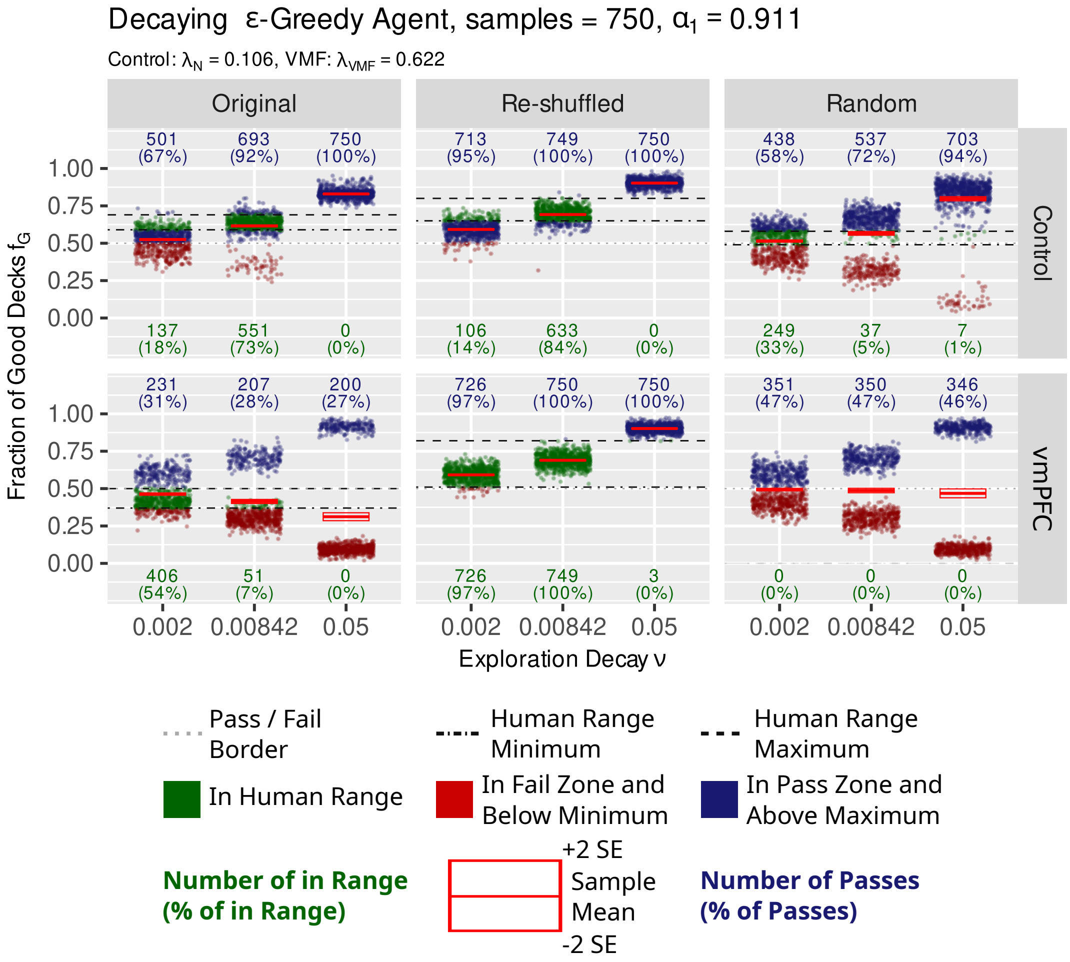


FIGURE 6.26: Decaying ε-Greedy agent comparison of repeated simulation outcomes to human IGT results. At $\nu = 0.00842$ and reported CSUD minimum loss hyper-parameter values, the decaying ε-Greedy agent achieves 633 and 749 full matches for the normal and vmPFC impaired configurations respectively. Full details are in the text.

brackets indicate percentages matched. A high percentage matched value is indicative of predictive simulation success. Finally the red bars and box indicate central tendency in terms of the mean and ±2 SEs (standard errors).

Fig. 6.26 indicates that at exploration decay $\nu = 0.00842$, decaying ε-Greedy agent simulations achieve the best results in the re-shuffled IGT, with 84% and 100% matches for the normal and vmPFC impaired configurations respectively. In the remaining configurations, the decaying ε-Greedy agent achieves for the normal original case 73% matches, while only achieving 7% and 5% actual matches for the vmPFC impaired original and normal random configurations respectively. In the vmPFC impaired original and

Test Variant	Test Statistic	df1	df2	p-Value	Subset Results
Original ∣ Re-Shuffled ∣ Random vs. Learning Decay λ					
ANOVA Type[a]	110.633	2.969	4447.856	0	At α = 0.01, the null hypotheses of learning decay factor equality is rejected. Only equality of the re-shuffled response cannot be rejected.
Original ∣ Random vs. Learning Decay λ					
ANOVA Type	171.039	1.982	2968.573	0	At α = 0.01, the null hypotheses of equal original and random, original only, and random only responses are rejected.
Wilks Lambda	186.696	2.000	1497.000	0	
Re-Shuffled vs. Learning Decay λ					
ANOVA Type	1.698	1.000	1498	0.193	Single response variable, no subsets.
Wilks Lambda	1.698	1.000	1498	0.193	

[a] Wilks Lambda could not be computed due a singular rank matrix.

TABLE 6.13
Decaying ε-Greedy agent np-M/ANOVA analysis of mean fraction of good decks $\bar{f}_G$ with learning decay λ as factor. At significance level α = 0.01 mean fraction of good decks $\bar{f}_G$ responses are statistically significantly different, except for the re-shuffled IGT environment.

random normal cases, where low matches are achieved, note that the red bars indicate that simulation means lie within the respective human outcome catchment areas marked at the maximum by the dashed, and at the minimum, by the dot-dashed lines. However, the low match cases display bi-modal outcome distributions leading to very low mass in the catchment areas.

Table 6.13 depicts decaying ε-Greedy agent np-M/ANOVA results considering the factor effect of normal and vmPFC impaired learning rate decay $\lambda_N = 0.106$ and $\lambda_{vmPFC} = 0.622$ relative to the mean fraction of good decks $\bar{f}_G$. The test statistic and degrees of freedom (df1 and df2) columns report the details of the relevant test variant against which the reported p-value result is obtained. The results show that at significance level $\alpha = 0.01$, the joint (multivariate) factor effect of learning rate decay across the original, re-shuffled, and random IGT environments is significant. Further, the learning

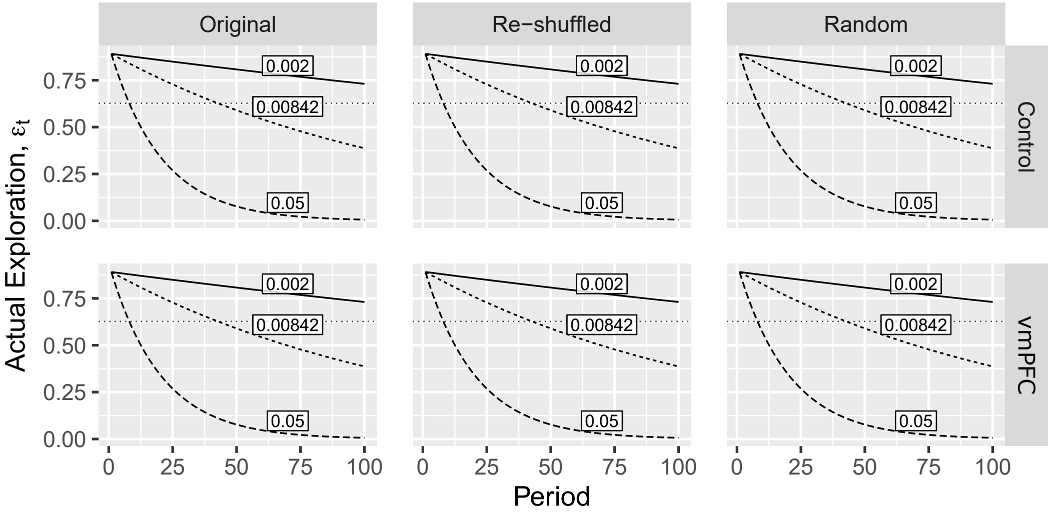


FIGURE 6.27: Decaying ε-Greedy agent 2D contours showing exploration decay ν effects during the IGT. In general, as ν increases, exploitation increases at a faster rate. The dotted line marks $\varepsilon = 0.627$, the CSUD discovered minimum-loss exploration value for the ε-Greedy agent results presented in section 6.3.

rate decay factor effect fails to produce a statistically significant effect for the re-shuffled IGT environment alone; this result is in conformance with corresponding human IGT results.

Finally given initial exploration $\varepsilon_1 = 0.891$, Fig. 6.27 presents how exploration decay ν affects per period exploration ε_t. In general, as ν increases, exploration decay increases. The dotted line marks $\varepsilon = 0.627$, the CSUD discovered minimum-loss exploration value for the ε-Greedy agent results presented in section 6.3. Hence the dotted line helps to contextualise per period decaying exploration ε_t versus constant exploration ε.

Note that at CSUD selected exploration decay $\nu = 0.00842$, in relation to constant exploration at $\varepsilon = 0.627$, decaying exploration is above this value for approximately the first IGT half (apprx. 43 periods), while decreasing below constant exploration in approximately the second IGT half. In that manner,on average, over the course of the 100 draws, exploration decay with $\nu = 0.00842$ appears to replicate the results the constant exploration ε-Greedy agent achieves with $\varepsilon = 0.627$.

6.6.1 Decaying ε-Greedy Agent Discussion

The decaying ε-Greedy agent presents primarily good results. However, these results are not very different from those obtained by the simpler constant exploration ε-Greedy agent presented in section 6.3. This raises the question of whether in general there exist constant exploration counterparts to the exponential exploration decay model employed here.

In relation to minimising CSUD search loss in 1000 iterations, the decaying ε-Greedy search delivers the lowest minimum loss at $loss_{min} = 0.00276$. In relation to CSUD grid search verification, the decaying ε-Greedy agent delivers at hyper-parameter values $\alpha_1 = 0.911$, $\lambda_N = 0.106$, $\lambda_{vmPFC} = 0.622$, $\varepsilon_1 = 0.891$, and $\nu = 0.00842$, mean fraction of good decks $\bar{f}_G$ matches for all behaviour and IGT environment data, for which human IGT outcome comparables exist.

In relation to 20-draw blocked data, this agent achieves good matches as well. Further for np-M/ANOVA results with normal learning decay λ_N and vmPFC impaired learning decay λ_{vmPFC} as factors and at significance level $\alpha = 0.01$, the decaying ε-Greedy achieves a joint statistically significantly different result, while failing to achieve as expected a statistically significantly different result for the re-shuffled IGT.

As with all other ε-Greedy based agents, CSUD verification fraction of good decks f_G jitter plots display bi-modal densities with little or no mass in human IGT outcome catchment zones, especially for the normal original and random, and vmPFC impaired original IGT cases. This tendency could be an artefact of exploring all possible alternatives subject to learning decay.

6.7 Summary

A joint CSUD search is conducted across the original, re-shuffled, and random IGT environments, to discover mean fraction of good decks $\bar{f}_G$ matches for normal and vmPFC impaired human behaviours, modelled in software agents by initial learning rate α_1, exploration, and normal and vmPFC impaired learning decay λ_N and λ_{vmPFC} respectively.

The results reported here are based on aggregated human IGT outcomes. With the exception of the random IGT (Steingroever et al., 2015), individual human IGT outcome data was not available. As noted in chapter 1 when combined with the human $\bar{f}_G^H$ targeting outcomes, a representative agent and squared loss, CSUD resembles recursive least squares. From this perspective that is, thinking of CSUD as recursive least squares, one might ask to what extent CSUD derived Q-learning hyper-parameters are useful at explaining individual human behaviour? This question, however, is beyond the scope of this work. The primary purpose of this work is not to explain human behaviour, but to develop nonrational computational technologies inspired by human behaviour. This work, however, provides some benchmark results, which may in future work be fitted against individual IGT outcome data to ascertain whether observed simulation results do indeed obtain in humans.

From an algorithmic perspective, this work focuses on the key Q-learning hyper-parameters involved in learning over time and in the exploitation versus exploration trade-off. Grid search results show that given exponential learning rate decay, original, re-shuffled and random IGT $\bar{f}_G$ contours decompose very nicely into onion-layered surfaces. Although it is not reported here, such a decomposition does not obtain with linear learning rate decay. The onion layer decomposition effectively minimises hyper-parameter interaction. The main contribution of the initial learning rate α_1 is to produce initial non-stationarities. Exploration produces vertical $\bar{f}_G$ contour shifts, whereas learning rate decay λ determines normal versus vmPFC behaviour. Grid search results verify that with this onion layered decomposition, CSUD does indeed achieve its algorithmic objective of tuning Q-learning hyper-parameters to minimise target deviations. It is also clear from 2D and 3D contour plots that the CSUD solution is not unique. The non-uniqueness of CSUD could be taken as an indication that the hyper-parameters in question may vary across individuals, however, that there exists some general key tendencies.

Here two key results are presented: (1) regardless of agent type, increasing learning rate decay leads to vmPFC impaired agent behaviour, and (2) human exploration in the IGT is shown to be very high both in terms of ex-ante agent exploration and the ex-post implied exploration index (EI).

From an algorithmic perspective either for RL or CSUD, both learning rate decay and exploration remain indispensable components of operation. As it has been noted (Ljung, 1978; Spall, 1992; Tsitsiklis, 1993; Yin & Kushner, 2003), in iterative learning such as RL or stochastic approximation, for theoretical guarantees of convergence, the learning rate *must* decay, however, subject to decay speed limits. A constant learning rate such as in the

EV (2.1b), the PV (2.2b), or the ORL (2.4b) does not satisfy this prerequisite; nor however, does the exponentially decaying learning rate used here. Intuitively, a constant learning rate leads to non-degrading oscillations about the optimum, and in a stochastic context, this could possibly lead to divergence. The rational learning requirements apply to a broad range of problems. It would be of interest to fit human choice problem outcome data, such as IGT outcomes, to RL models with rational and exponentially decaying learning rates.

Regarding exploration, a further result in the IGT literature states that normal and vmPFC impaired subjects do not produce statistically significant group effects with respect to re-shuffled IGT outcomes, while producing a corresponding significant group effect with respect to original IGT outcomes (Fellows & Farah, 2005, pp. 60-61). This result is assessed using np-M/ANOVA (non-parametric multivariate analysis of variance) with the cumulative mean fraction of good decks measure as the response and, normal and vmPFC impaired learning rate decay as factors. It is found that the human subject result analogue can only be obtained with the ε-Greedy and decaying ε-Greedy agents.

One interesting consideration at the outset of this work was to see whether universal normal and vmPFC impaired learning decay rates could be obtained. Such universal rates were hypothesized to remain the same across different agents and IGT environments. However, in search results such a common value set could not be discovered. It was found instead that CSUD discovered normal and vmPFC impaired learning decay rates varied among agents, while remaining relatively stable across the original, re-shuffled, and random IGT environments. All of these three IGT environments have identical long term net yield structures. As agents only differ by exploration implementations, however, normal and vmPFC impaired learning decay rates may be sensitive to underlying model stochasticity. A direct comparison of randomness and corresponding levels of learning rate decay has not been undertaken here and remains outside of the scope of this current work.

Four reinforcement learning software agents are considered, of which the Boltzmann and adaptive ε-Greedy agents can be considered as rational, as they include sophisticated probabilistic modelling leading in the Boltzmann agent to proportionate probabilistic exploration, and in the case of the adaptive ε-Greedy agent to temporal difference based exploration scaling. The remaining two agents, consisting of the ε-Greedy and decaying ε-Greedy agents, comprise heuristic agents, with the former agent exhibiting constant and the latter agent exhibiting constant decay exploration. Literature references for the single-state (exponential decay) decaying exploration agent used here have not been found. However, in multi-state environments, a version of the decaying ε-Greedy agent, where exploration decays in proportion to the number of state visits has been discussed for example in Powell, 2011, p. 466.

In general, the heuristic or nonrational agents perform better than the rational agents in the following sense: minimum CSUD loss and ability to match all human results at CSUD discovered hyper-parameters. Both the ε-Greedy and decaying ε-Greedy agents produce mean fraction of good decks $\bar{f}_G$ results, which for normal behaviour original, re-shuffled, and random; and vmPFC impaired behaviour original and re-shuffled IGT environments produce results in the corresponding IGT human outcome match zones. Further, treating normal learning decay λ_N and vmPFC impaired learning decay λ_{vmPFC} as factors at statistical significance level $\alpha = 0.01$, the original IGT environment produces behaviour driven statistically significantly different $\bar{f}_G$ outcomes, while the re-shuffled environment fails to do so.

In contrast, the rational agents struggle to produce $\bar{f}_G$ performance matches for all behaviour and IGT environment cases. The Boltzmann agent CSUD hyper-parametrisation fails to match normal random IGT environment outcomes, while the adaptive ε-Greedy agent CSUD hyper-parametrisation obtains a match where the vmPFC impaired original and normal random outcomes are just at the respective match zone boundaries. Also, both rational agents fail to reject the null hypothesis of no learning decay factor effect in the re-shuffled IGT.

Only the Boltzmann agent, however, generates unimodal fraction of good decks f_G jitter plots. In the original and random IGT environments especially with vmPFC impaired behaviour, ε-Greedy agent variants may produce bi-modal jitter plots, which may display little or no mass inside human IGT outcome match zones. This effect may result from a combination of learning rate decay with equal-incidence exploration. When agents learn correct responses, high exploration leads to an additional low performance cluster, and when agents learn incorrect responses, high exploration likewise leads to an additional high performance cluster. These high exploration induced individual performance variations produce bi-modal jitter plots, where the central tendency is computed to be in the IGT human catchment areas, however, with little or no actual individual software agent mass inside catchment zones.

20-draw block exploration index (EI) plots suggest that the ε-Greedy and decaying ε-Greedy agents come closest to matching corresponding human outcomes. In 1000 CSUD iterations, the decaying ε-Greedy agent achieves 860 full matches while the ε-Greedy agent only achieves 290 full matches. Therefore in relation to the cumulative end-of-task mean fraction of good decks measure used in CSUD, having heuristic exploration decay appears to improve search outcomes. However, Fig. 6.27 reveals that, the two ε-Greedy models may be quite similar, when considering average exploration. In fact, the decaying ε-Greedy model appears to exhibit average exploration around the $\varepsilon = 0.627$ value employed by the constant exploration ε-Greedy model. Between these two heuristic models, the constant exploration model remains the simpler alternative.

It is unclear why the heuristic models perform better. This could be due to the statistical properties of the original, re-shuffled, and random IGT payouts, where some decks exhibit low frequency realisation of the determining payouts, making it difficult to accurately develop mean net payout representations. Chapters 7 and 8, further explore this possibility by looking at the reversed IGT and SGT environments.

Chapter 7

Reversed IGT with Simple Reinforcement Learning Modelling via CSUD

The reversed IGT (Bechara et al., 2000, p. 2193) has been introduced in section 4.1.4. To review, the reversed IGT environment consists of four decks E, F, G, and H, where decks E and G are the good decks, producing frequent high fines with less frequent but higher rewards providing a positive net yield; and F and H are the bad decks with frequent low fines but even lower less frequent rewards providing a negative net yield. Over the course of 100 turns, a duration unknown to the participants, the participants must discover the on average positive net yield decks E and G. While the original, re-shuffled, and random IGT environments produce a steady stream of rewards with occasional fines, the reversed IGT environment produces a steady stream of fines, with occasional rewards. However, for decks E and G, the occasional rewards produce on average net positive yields.

The reversed IGT draw-by-draw yield structure is shown in Appendix A.4. It wil be noted that the reversed IGT more closely resembles gambling activities, where in each period a constant entry cost must be incurred to gain admission to the possibility of a large payout. Bechara et al., 2000 report that the aim of the reversed IGT, that is of making punishments constant, is to assess contributions of insensitivity to punishment and hypersensitivity to reward in vmPFC impaired subject outcomes. This work reports on the reversed IGT, because in a value maximisation context, the constant accrual of costs makes it more difficult to determine the best on average positive net yield decks. It is of interest to assess the effect of such a more difficult signal extraction problem on the values of the initial learning rate, exploration, and learning rate decay.

The adaptive ε-Greedy agent is no longer considered. This is on the basis of the poor 20-draw block mean fraction of good deck $\bar{f}_G$ and emotion index results presented in Fig. 6.17 and Fig. 6.18. Further, based on the np-M/ANOVA results reported in Table 6.10, the adaptive ε-Greedy agent cannot replicate the key IGT literature result that for the re-shuffled IGT, there is no statistically significant factor (group) effect arising from normal learning rate λ_N and vmPFC impaired learning rate λ_{vmPFC} as behavioural factors, when using mean fraction of good decks as the response (Fellows & Farah, 2005, pp. 60-61). Due to these two observations, the adaptive ε-Greedy agent, while exhibiting a very interesting rational technology, does not appear to be capable of producing human analogue results.

With the remaining ε-Greedy, Boltzmann, and decaying ε-Greedy architectures, it will be investigated to what extent learning rate decay λ and exploration influence software agent ability to achieve human reversed IGT outcomes for normal and vmPFC impaired behaviours. All discussed agents will continue to use base model (6.1) introduced in chapter 6. The general methodology remains the same as in section 6.1.

7.1 Search of the Reversed IGT environment

Table 7.1 summarises software agent CSUD search parameter constraints and attributes. As in chapter 6, broad parameter search ranges are used for the initial learning rate and exploration, while smaller ranges are employed for normal and vmPFC impaired learning rate decay.

It is found from preliminary searches that the ranges for normal and vmPFC impaired learning rate decay, λ_N and λ_{vmPFC} respectively, must be set carefully, so as to avoid outcomes where λ_N and λ_{vmPFC} produce simulation results, which do not reside in the corresponding human IGT outcome catchment zones. It is believed that this odd behaviour results from the use of a loss function where normal and vmPFC impaired loss are added together; this addition can be seen in the more complex multi-environment search loss specification (5.9), where losses across environments and behaviours (normal, vmPFC impaired) are aggregated additively. Here, a simple version of (5.9) with a single environment and two behaviours is used.

Agent	Boltzmann	ε-Greedy	Decaying ε-Greedy
Hyper-parameter			
Initial learning rate α_1	0.01 - 0.99	0.01 - 0.999	0.05 - 0.99
Normal learning rate decay λ_N	0.0001 - 0.25	0.03 - 0.30	0.03 - 0.32
vmPFC impaired learning rate decay λ_{vmPFC}	0.25 - 1.2	0.25 - 1.2	0.32 - 1.2
Temperature τ	0.5 - 500		
Exploration ε		0.05 - 0.70	0.5 - 1.0[a]
Exploration decay ν			0.002 - 0.02
CSUD Iterations	1000	1000	1000
Gradient Samples	5	1	3

IGT length	Q-learning samples
100	750

[a] Exploration ε refers to initial exploration ε_1.

TABLE 7.1
Search Methodology: Joint reversed IGT hyper-parameter CSUD search criteria by agent.

By limiting learning rate decay ranges, prior information in the sense of range boundaries is injected into the search query. In this case, the notion that λ_N and λ_{vmPFC} must be distinct and that the former must be less than the latter. By construction, search loss is minimised to the extent that search criteria are fulfilled. It is in this sense that the CSUD search strategy can be seen as a contraction of a grid search over a constrained space.

Because CSUD is a stochastic, gradient driven search technique, sometimes a single gradient evaluation is not sufficient to produce a reliable gradient estimate. The 'Gradient Samples' entry in Table 7.1 indicates if any gradient sampling was employed.

7.2 Search Results: Reversed IGT

Searches are conducted via CSUD. The reinforcement learning layer is implemented using the Boltzmann, ε-Greedy, and decaying ε-Greedy agents.

For the reversed IGT environment, cumulative and 20-draw block human mean fraction of good decks $\bar{f}_G^H$ data is available for both normal and vmPFC impaired subjects. Additionally 20-draw block exploration index (EI) values can be computed (see chapter 4). The CSUD searches will aim to find hyper-parameter combinations, which can produce simultaneous performance matches for two human IGT outcomes: normal reversed and vmPFC impaired reversed cumulative mean fraction of good decks results.

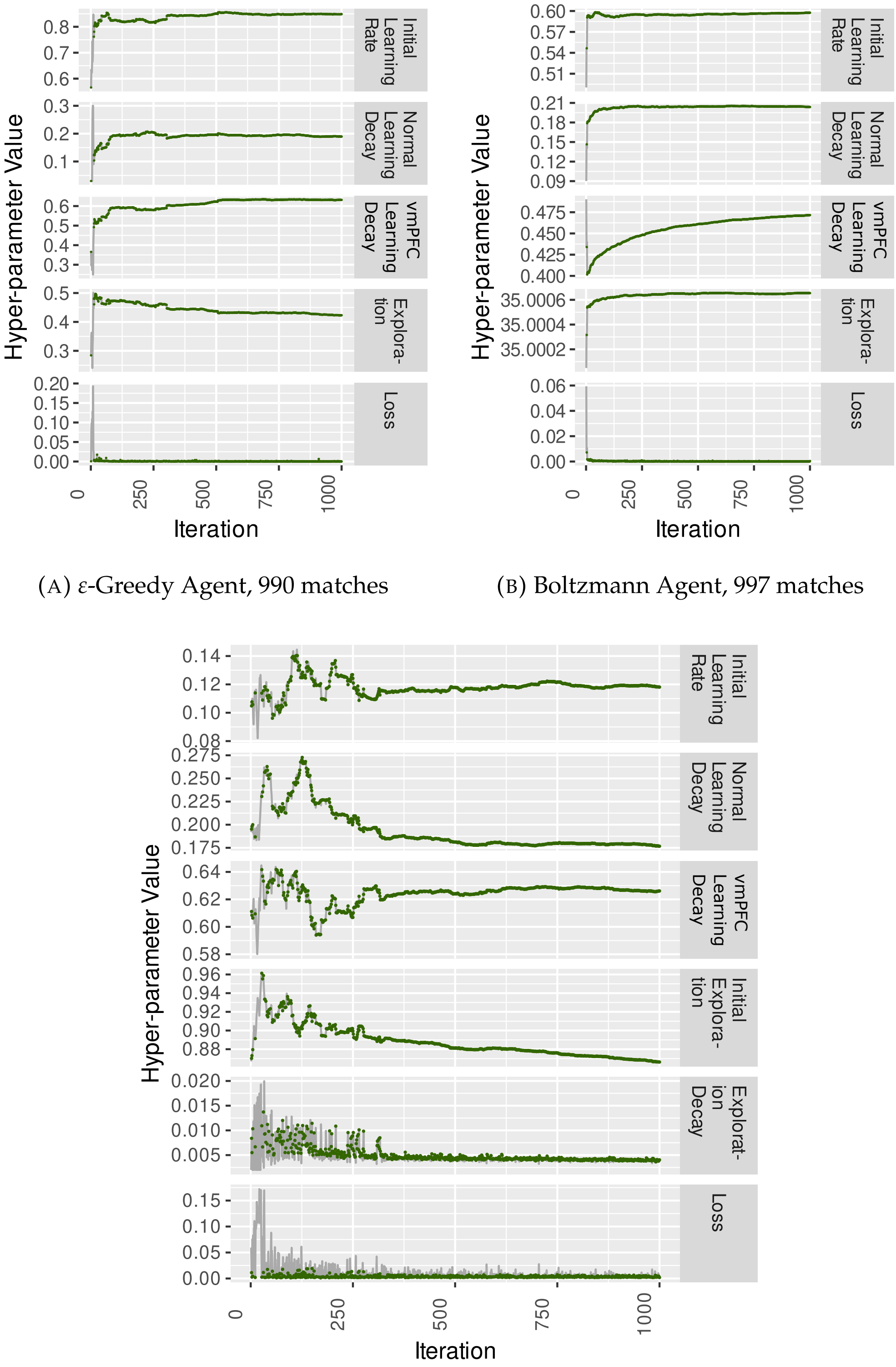
For ease of comparison, the ε-Greedy, Boltzmann, and decaying ε-Greedy agent results are presented side-by-side.

Fig. 7.1 and Table 7.2 present CSUD search results in graphic and tabular forms respectively.

Table 7.2 indicates that after 1000 CSUD iterations, all agents have achieved a large number of matches within the normal and vmPFC impaired reversed IGT outcome human match zones, which are noted in Table 4.8. For the reversed IGT environment, the (rational) Boltzmann agent has the lowest minimum loss performance, while the (heuristic) decaying ε-Greedy has the highest minimum loss performance.

As discussed in section 6.7, prior to empirical results, one consideration in this work was whether universal normal and vmPFC impaired learning decay rates, reflecting perhaps some unknown organic rule and therefore applying to all agents and environments, could be obtained. It was found, however, in chapter 6 that such a universal value set could not be obtained across different agents. The results in Table 7.2 further suggest that at least with the CSUD search methodology, such a universal normal and vmPFC learning rate decay value set cannot be found across IGT environments with substantially different net yield structures.

Intuitively the lack of discovering across considered IGT environments a universal normal and vmPFC impaired learning decay rate can be seen as arising from the combination of the strong loss assessment of the CSUD employed squared loss function and the fact that exponential learning rate decay λ acts as a sampling frequency band-pass filter, attenuating sampling after a certain search iteration. Consequently, the reversed IGT environment has a substantially different net yield structure, and there is no a-priori reason to expect for this structure to be discovered at the same sampling frequency as that is applied to the normal, re-shuffled, and random IGTs via the corresponding learning decay rates.



(A) ε-Greedy Agent, 990 matches

(B) Boltzmann Agent, 997 matches

(C) Decaying ε-Greedy Agent, 752 matches

FIGURE 7.1: Reversed IGT CSUD searches. Green dots indicate matches. All agents achieve a large number of matches. But at 1000 iterations search convergence might not yet have been achieved. Details in text.

Agent	ε-Greedy	Boltzmann	Decaying ε-Greedy
Hyper-parameter			
Initial Learning Rate α_1	0.848	0.596	0.121
Normal Learning Decay λ_N	0.193	0.205	0.226
vmPFC Impaired Learning Decay λ_{vmPFC}	0.633	0.466	0.594
Exploration	$\varepsilon = 0.431$	$\tau = 35.0$	$\varepsilon_1 = 0.905$ $\nu = 0.00609$
Minimum Loss	8.12e^{-7}	4.64e^{-7}	0.00125
Matches	990	997	752

TABLE 7.2
Reversed IGT CSUD search matches after 1000 iterations. Minimum loss column shows selected agent hyper-parameters.

Fig. 7.1 depicts CSUD iteration results for agent hyper-parameters and loss. Green dots indicate agent hyper-parameter combinations, where agent results for normal and vmPFC impaired learning decay λ_N and λ_{vmPFC} lie within normal and vmPFC impaired human reversed IGT environment outcome ranges. Such a result is referred to as having achieved a match to human IGT outcomes.

Rather than obtain a global minimum, conceptually CSUD search aims to satisfice, that is produce one or more suitable search result candidates. As noted in Proposition 12.4.1, only if a global optimum already exists in the constrained search space, does the CSUD search strategy theoretically guarantee convergence to a global minimum loss outcome; however, only in terms of the performance statistic ($\bar{f}_G$) and not in terms of the performance statistic generating (Q-learning) hyper-parameters. If search budgets permit, then further searches at different iterations, or with different initial conditions may be undertaken to assess whether the initially obtained results correspond to those with a global minimum. In such endeavours however, care must be taken as complex searches with complex loss functions over the constrained space may not exhibit a global minimum.

Based on the slow convergence paths of vmPFC impaired learning rate decay λ_{vmPFC} and initial exploration ε_1 for the Boltzmann and decaying ε-Greedy agents respectively, Fig. 7.1 indicates that at 1000 iterations, hyper-parameter searches for these two agents may not yet have converged. Fig. 7.1b shows that for the Boltzmann agent at 1000 iterations, vmPFC impaired

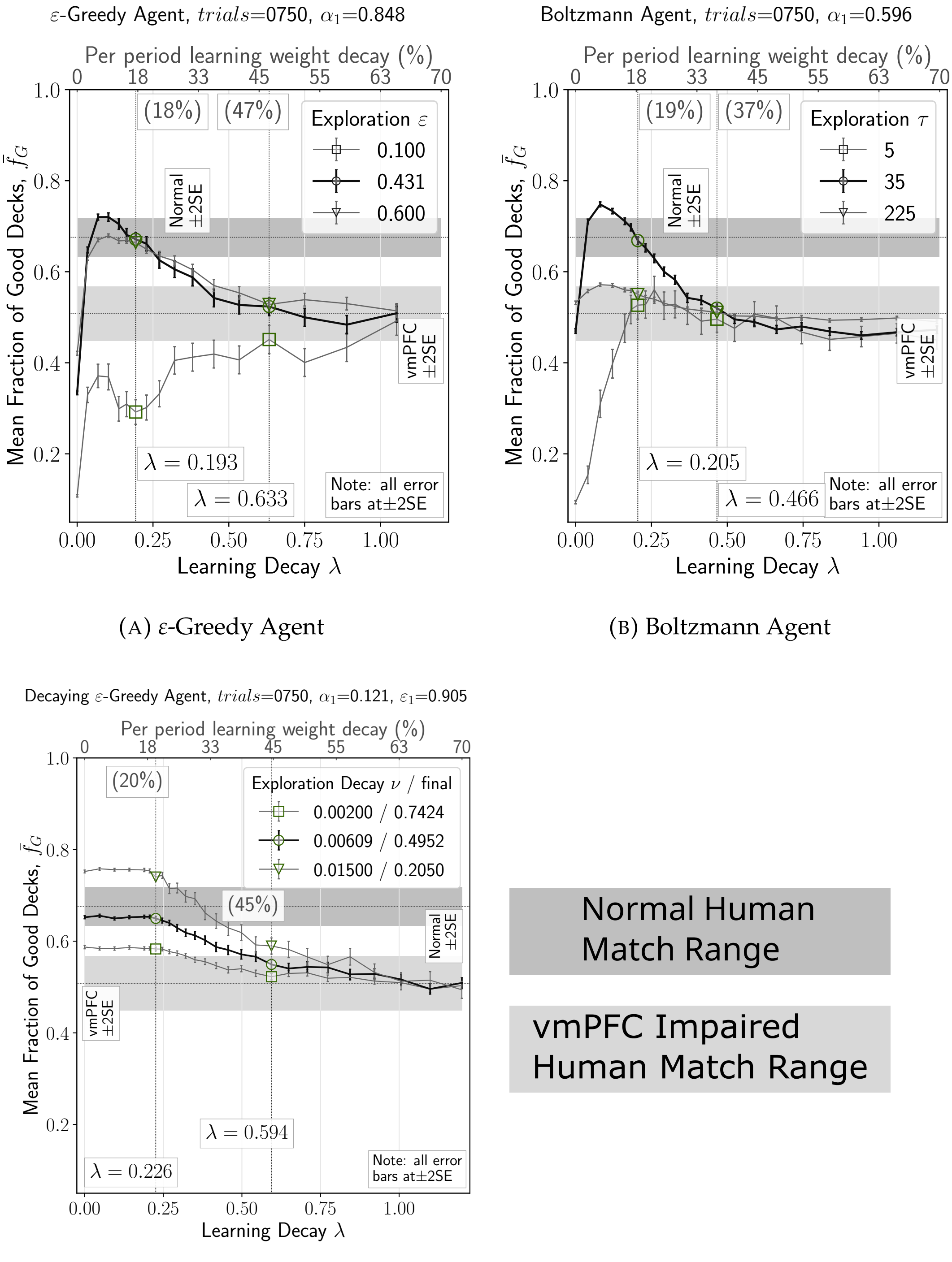
Agent	ε-Greedy	Boltzmann	Decaying ε-Greedy
Hyper-parameter			
Initial Learning Rate α_1	0.1, 0.5, 0.848, 0.999	0.05, 0.33, 0.596, 0.99	0.01, 0.121, 0.45, 0.9
Normal Learning Decay λ_N[a]	0.193	0.205	0.226
vmPFC Impaired Learning Decay λ_{vmPFC}[a]	0.633	0.466	0.594
Exploration	ε = 0.1, 0.431, 0.6, 0.8	τ = 5, 35, 75, 225	ε_1 = 0.25, 0.5, 0.905, 1 ν = 0.002, 0.00609, 0.015, 0.05
IGT length	Q-learning samples		
100	750		

[a] The learning rate decay and exploration decay grids are constructed around the two values above. Appendix C provides the construction method.

TABLE 7.3
CSUD Verification. Reversed IGT hyper-parameter grid search criteria for the ε-Greedy, Boltzmann, and decaying ε-Greedy agents.

learning decay λ_{vmPFC} appears be increasing albeit at a lower trend. Fig. 7.1c shows that for the decaying ε-Greedy agent, initial exploration still appears to be on a decreasing trend. However, keeping in line with the nonrational idea of a limited search budget, which here is set to 1000 iterations, only these results are reported here. Despite these unconverged searches, at 1000 iterations, all agents achieve a high number of matches out of 1000 iterations, with the Boltzmann agent achieving 997 out of 1000 matches. From a nonrational CSUD search perspective, all that is needed is to achieve such search matches.

Table 7.3 presents agent grid search verification configurations. Fig. 7.2 and Fig. 7.3 show 2D and 3D CSUD grid search verification contours. In Fig. 7.2, dark and light gray zones represent normal and vmPFC impaired human IGT outcome match areas respectively. Solid black contours show response to learning decay λ at CSUD selected minimum loss hyper-parameter values. In terms of mean fraction of good decks $\bar{f}_G$, Fig. 7.2 indicates that for all agents, increasing learning rate continues to lead to outcomes, which are consistent with vmPFC impaired behaviour. That is for all agents, there is a lower learning decay rate λ_N and a higher learning decay rate λ_{vmPFC} associated with normal and vmPFC impaired behaviour respectively.



(A) ε-Greedy Agent

(B) Boltzmann Agent

(C) Decaying ε-Greedy Agent

(D) Legend

FIGURE 7.2: Reversed IGT CSUD verification grid search 2D contours. Dark and light gray zones represent normal and vmPFC impaired human IGT outcome match areas respectively. Solid black contours show response to learning decay λ at CSUD selected minimum loss hyper-parameter values.

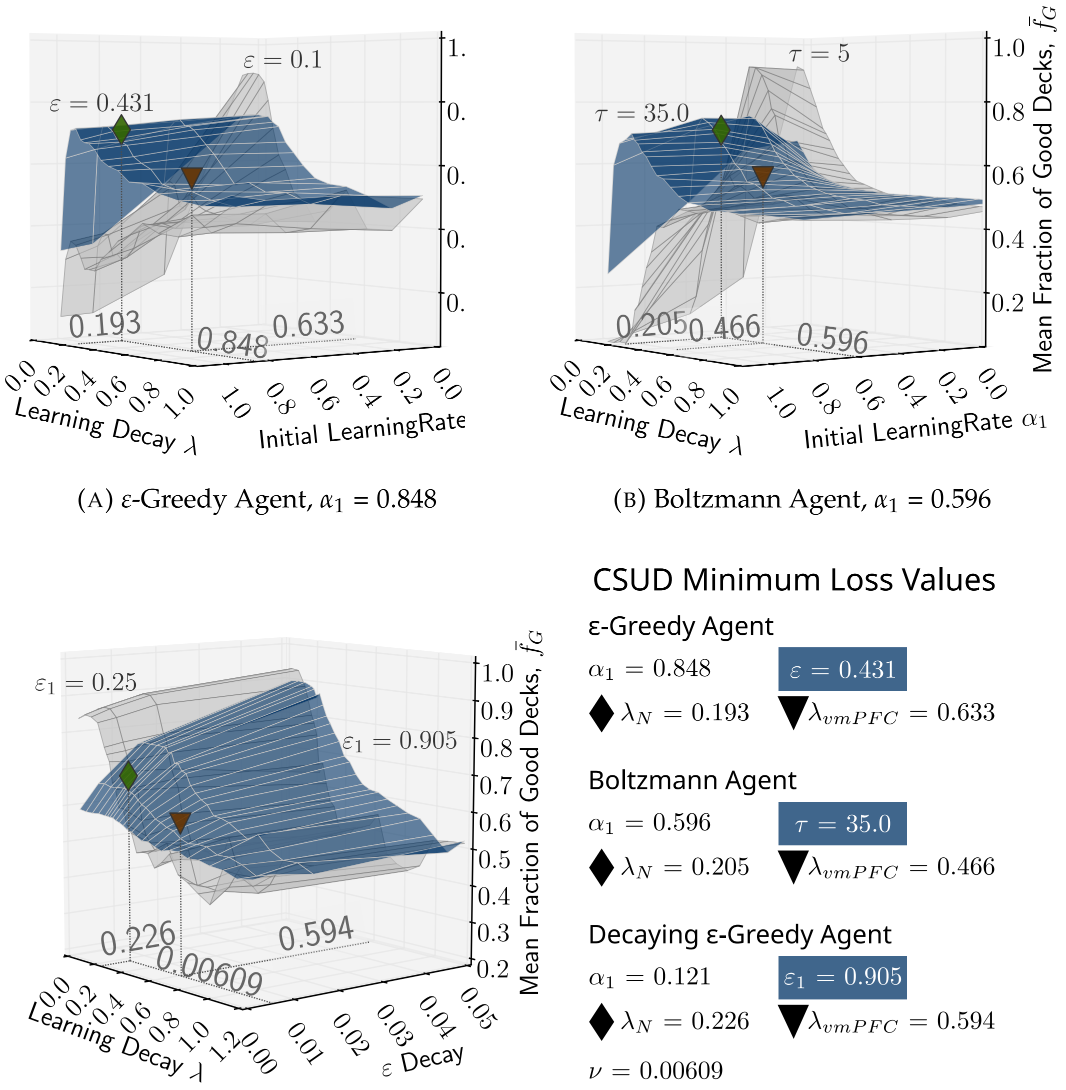
However, for all agents, the role of exploration becomes more complex. For the ε-Greedy and Boltzmann agents, unlike in the original, re-shuffled, random IGT environment joint search results in chapter 6, low exploration no longer guarantees higher than normal human outcomes.

Fig. 7.2a and Fig. 7.2b show that in the reversed IGT environment for the ε-Greedy and Boltzmann agents, low exploration leads to $\bar{f}_G$ outcomes, which remain below the human dark gray normal IGT outcomes. Fig. 7.2c shows that at initial exploration $\varepsilon_1 = 0.905$, increasing exploration decay ν shifts $\bar{f}_G$ contours upwards as seen in chapter 6. However, all of the contours depicted in Fig. 7.2c may produce solution candidates in the dark and light gray match zones.

In such circumstances, where many solutions exist, note that by design CSUD produces solutions, which attempt to minimise deviations from the corresponding human mean fraction of good deck outcome means. This is noted for example in Fig. 7.2.

For the ε-Greedy and Boltzmann agents, Fig. 7.3a and Fig. 7.3b respectively illustrate that low exploration in combination with a low initial learning rate may produce mean fraction of good decks $\bar{f}_G$ outcomes, which exceed the normal reversed IGT outcome human match zones. At low exploration, for example at $\varepsilon = 0.1$, or $\tau = 5$, holding learning decay λ constant, and increasing the initial learning rate leads to a strong decline in mean fraction of good decks $\bar{f}_G$. Hence in the reversed IGT environment, both the ε-Greedy and Boltzmann agents exhibit some interaction between exploration and the initial learning rate. However, as indicated by the lack of a similar decreasing slope in blue coloured surfaces, when exploration is higher then this interaction appears to stop.

For the decaying ε-Greedy agent, with learning decay λ and exploration decay ν on the horizontal axes and at initial learning rate $\alpha_1 = 0.121$, Fig. 7.3c presents a different view. Note that even when initial learning rate is held constant at $\alpha_1 = 0.121$, exploration behaviour is more complex in the sense that the lower exploration contour with initial exploration $\varepsilon_1 = 0.25$ crosses through the blue coloured surface associated with the CSUD discovered minimum loss hyper-parameter values consisting of ($\alpha_1 = 0.121$, $\lambda_N = 0.226$, $\lambda_{vmPFC} = 0.594$, $\varepsilon_1 = 0.905$, $\nu = 0.00609$). The decaying ε-Greedy agent 3D visualisation results also support the notion of more complex agent exploration behaviour in the reversed IGT.



(A) ε-Greedy Agent, $\alpha_1 = 0.848$

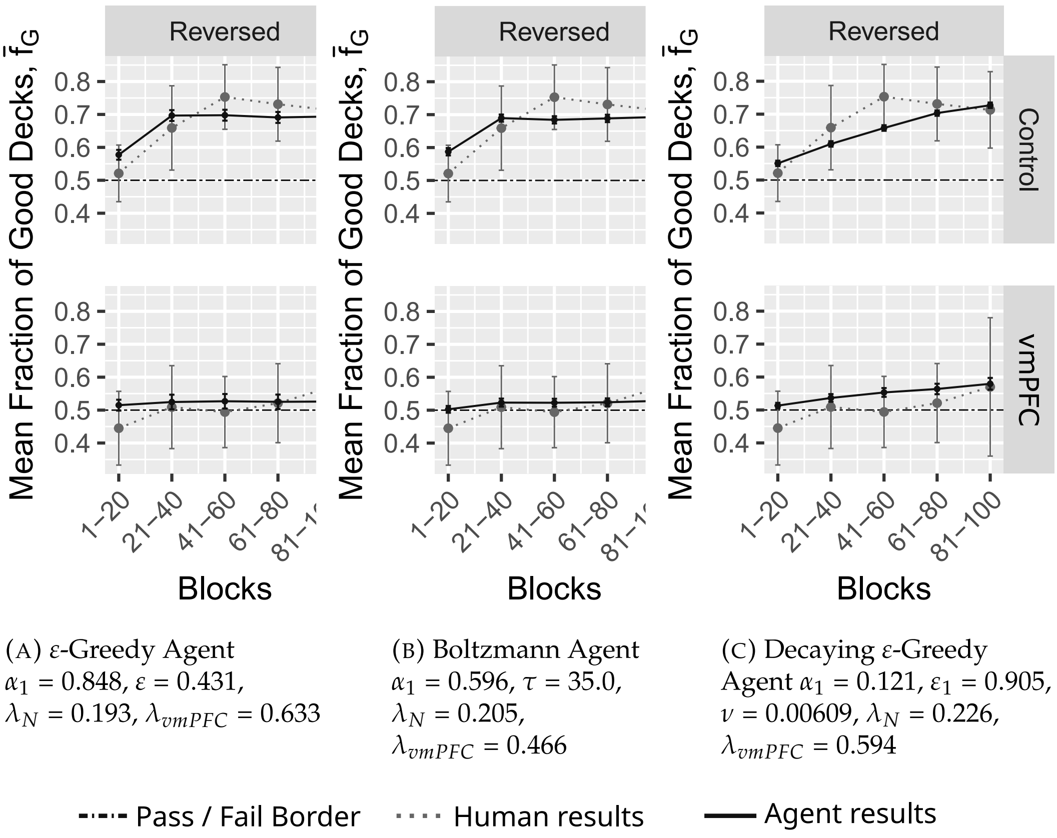
(B) Boltzmann Agent, $\alpha_1 = 0.596$

(C) Decaying ε-Greedy Agent, $\alpha_1 = 0.121$

(D) Legend

FIGURE 7.3: Reversed IGT CSUD verification grid search 3D contours. Blue coloured surfaces show response to learning decay λ at CSUD selected minimum loss hyper-parameter values. The diamond and inverted triangular shapes mark CSUD minimum loss normal and vmPFC impaired learning decay rates respectively.

Fig. 7.4 and Fig. 7.5 compare 20-draw block agent results to corresponding human outcomes. Fig. 7.4 compares mean fraction of good decks $\bar{f}_G$ results achieved by agents and humans. Human results are in light gray. Agent results appear in dark gray, averaged from 750 samples. All error bars are at ±2SE. When error bars are taken into account, agent and human 20-draw block performance appears relatively similar. For normal behaviour, Fig. 7.4a and Fig. 7.4b show that the ε-Greedy and Boltzmann



(A) ε-Greedy Agent $\alpha_1 = 0.848$, $\varepsilon = 0.431$, $\lambda_N = 0.193$, $\lambda_{vmPFC} = 0.633$

(B) Boltzmann Agent $\alpha_1 = 0.596$, $\tau = 35.0$, $\lambda_N = 0.205$, $\lambda_{vmPFC} = 0.466$

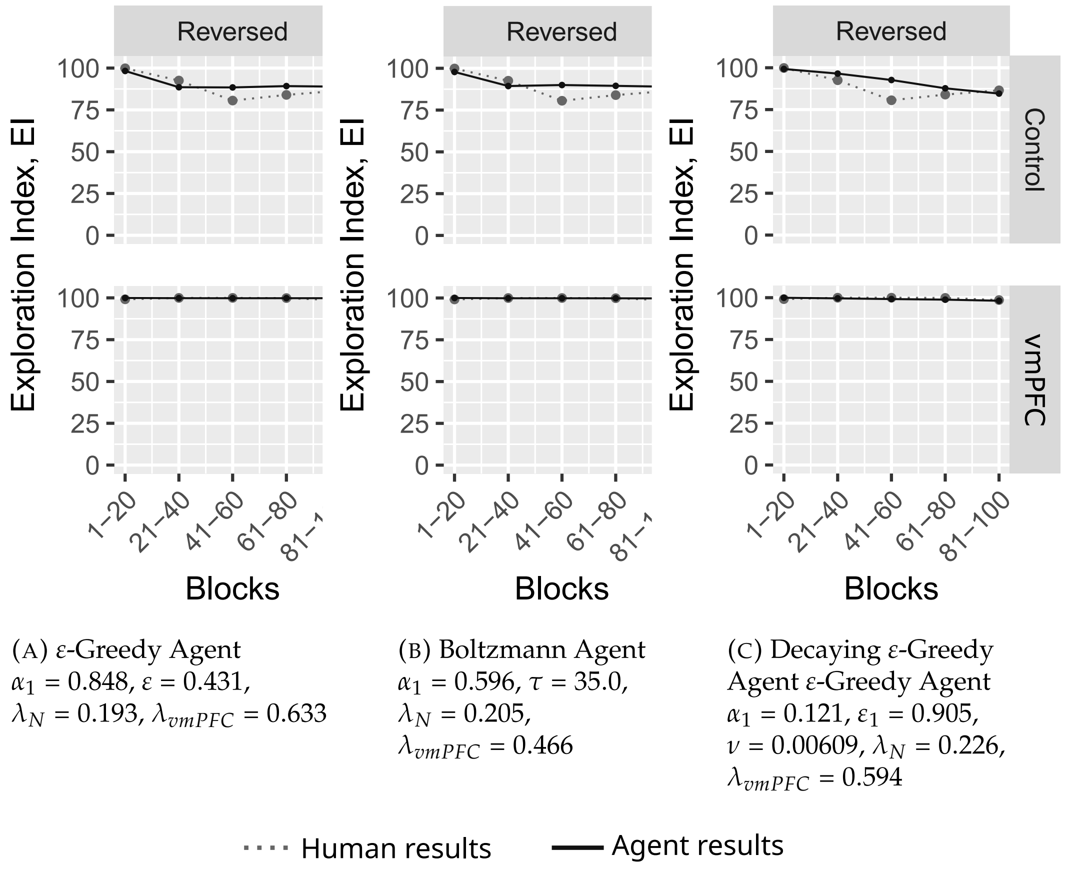
(C) Decaying ε-Greedy Agent $\alpha_1 = 0.121$, $\varepsilon_1 = 0.905$, $\nu = 0.00609$, $\lambda_N = 0.226$, $\lambda_{vmPFC} = 0.594$

(D) Legend

FIGURE 7.4: Reversed IGT agent 20-draw blocks comparison at CSUD minimum loss search matches noted in Table 7.2. Human results in dotted light gray. Agent results in solid dark gray, averaged from 750 samples. All error bars at ±2SE. When error bars are taken into account agent and human 20-draw block performance appears relatively similar. Details in text below.

agents match human per block $\bar{f}_G$ results more precisely than the decaying ε-Greedy agent, which matches the general trend.

Fig. 7.5 compares exploration index (EI) results achieved by agents and humans. At full exploitation, the exploration index reduces to 0. Human results are in light gray. Agent results appear in dark gray, averaged from 750 simulation samples. Human and agent exploration index responses appear relatively similar. However, for normal behaviour, Fig. 7.5a and Fig. 7.5b show that the ε-Greedy and Boltzmann agents match human per block exploration index (EI) results more precisely than the decaying ε-Greedy agent, which only matches the general trend. For normal behaviour, the exploration index does not drop under 75, indicating that agents do not switch to full exploitation. For vmPFC impaired behaviour, the exploration index



(A) ε-Greedy Agent
$\alpha_1 = 0.848$, $\varepsilon = 0.431$,
$\lambda_N = 0.193$, $\lambda_{vmPFC} = 0.633$

(B) Boltzmann Agent
$\alpha_1 = 0.596$, $\tau = 35.0$,
$\lambda_N = 0.205$,
$\lambda_{vmPFC} = 0.466$

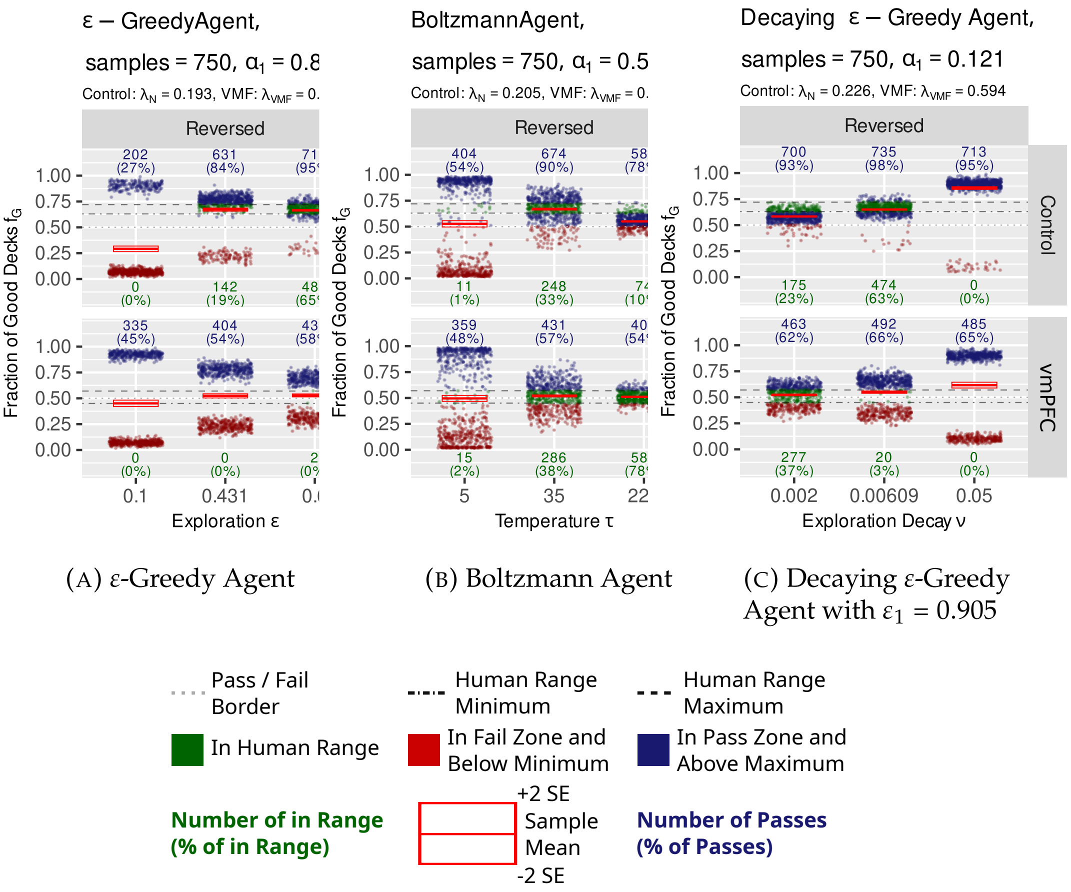
(C) Decaying ε-Greedy Agent ε-Greedy Agent
$\alpha_1 = 0.121$, $\varepsilon_1 = 0.905$,
$\nu = 0.00609$, $\lambda_N = 0.226$,
$\lambda_{vmPFC} = 0.594$

(D) Legend

FIGURE 7.5: Reversed IGT agent 20-draw blocks exploration index (EI) comparison at CSUD minimum loss search matches noted in Table 7.2. Human results in dotted light gray. Agent results in solid dark gray, averaged from 750 samples. All agents match the decreasing trend in normal human (control) EI and replicate the vmPFC human impaired outcome exhibited lack of EI reduction. Details in text below.

(EI), that is implied exploration, remains close to 100, indicating that the agents cannot learn to exploit the good decks.

Fig. 7.6 provides agent jitter plot density summary for the fraction of good decks f_G outcomes obtained at the CSUD selected minimum loss agent hyper-parameter values for 750 simulation samples of the reversed IGT environments for normal (control) and vmPFC impaired configurations. Green dots mark outcomes inside human performance ranges. Blue dots mark additional normative pass results, whereas red dots mark additional normative pass fails. Green numbers give total matches out of 750 samples, and the values in brackets indicate percentages matched. A high percentage



(A) ε-Greedy Agent

(B) Boltzmann Agent

(C) Decaying ε-Greedy Agent with $\varepsilon_1 = 0.905$

(D) Legend

FIGURE 7.6: Reversed IGT. Comparison of repeated agent simulation outcomes to human IGT results. CSUD minimum loss exploration values are $\varepsilon = 0.431, \tau = 35, \nu = 0.00609$ for the ε-Greedy, Boltzmann, decaying ε-Greedy agents respectively. ε-Greedy based agents exhibit a tendency towards bi-modal f_G. Due to this bi-modal tendency, agents do not achieve many simulation results, which are inside human match zones. Details are in the text.

matched value is indicative of predictive simulation success. The dashed and dash-dotted horizontal lines indicate the maximum and minimum respectively of the human match range. Finally the red bars and box indicate central tendency in terms of the mean and ±2 SEs (standard errors).

Fig. 7.6 shows that agents do not achieve many simulation results inside the respective human match zones. At CSUD selected exploration $\varepsilon = 0.431$, the ε-Greedy agent achieves 19% normal matches and no vmPFC impaired matches. At CSUD selected exploration temperature $\tau = 35$, the Boltzmann agent, which had the lowest CSUD search loss, achieves 33% normal and 38% vmPFC impaired matches. At CSUD selected initial exploration

Test Variant	Test Statistic	df1	df2	p-Value	Subset Results
ε-Greedy Agent, $\lambda_N = 0.193, \lambda_{vmPFC} = 0.633$					At $\alpha = 0.01$, the null hypotheses of Learning Decay factor equality is rejected.
ANOVA Type[a]	44.275	1.000	1498	0	
Boltzmann Agent, $\lambda_N = 0.205, \lambda_{vmPFC} = 0.466$					At $\alpha = 0.01$, the null hypotheses of Learning Decay factor equality is rejected.
ANOVA Type[a]	576.055	1.000	1498	0	
Decaying ε-Greedy Agent, $\lambda_N = 0.226, \lambda_{vmPFC} = 0.594$					At $\alpha = 0.01$, the null hypotheses of Learning Decay factor equality is rejected.
ANOVA Type[a]	128.535	1.000	1498	0	

[a] Wilks Lambda produces identical results.

TABLE 7.4
Agent Reversed vs. Learning Decay λ np-M/ANOVA analysis of mean fraction of good decks $\bar{f}_G$ with learning decay λ as factor. At significance level $\alpha = 0.01$ Mean fraction of good decks $\bar{f}_G$ responses are statistically significantly different.

$\varepsilon_1 = 0.905$ and exploration decay $\nu = 0.00609$, the decaying ε-Greedy agent achieves 63% normal matches and 3% vmPFC impaired matches. For ε-Greedy agent variants, this may in part be a result of the tendency towards bi-modal agent fraction of good decks f_G outcome densities, which have little or no mass at the distribution mean. Interestingly for any agent, increasing exploration leads towards a density migration towards the normal and vmPFC impaired human match zones. In contrast, as exploration decreases, for example as seen in Fig. 7.6b, as τ reduces from 225 to 5, even the Boltzmann agent moves towards a bi-modal f_G outcome density. Also note that at $\tau = 225$ vmPFC impaired simulations achieve 587 matches inside the corresponding human outcome catchment zone. In contrast, at $\tau = 35$, vmPFC impaired Boltzmann simulations achieve only 268 in catchment zone matches. The primary reported Boltzmann results use $\tau = 35$ since exploration is constrained to be identical across normal and vmPFC configured agents, and $\tau = 35$ performs better overall across both configurations.

Table 7.4 shows np-M/ANOVA effect analysis for reversed IGT mean fraction of good decks $\bar{f}_G$ with normal learning decay λ_N and vmPFC impaired learning decay λ_{vmPFC} as factors. Test variants can be thought of as non-parametric versions of the F-test, with a test statistic and two degrees of freedom. These three quantities are then assessed to derive the p-value. The test statistics are discussed in Burchett et al., 2017.

For all agents, at significance level $\alpha = 0.01$, the effect of learning decay as a factor is significant. That is, in the models proposed here, learning decay is instrumental in generating normal versus vmPFC impaired behaviour. The findings are discussed next.

7.3 Discussion: Reversed IGT

The hallmark of the reversed IGT task is that while the subject must identify the good decks E and G, which on average give net positive yields, they must do so in face of a per draw constant and high loss of 100 (imaginary) dollars. In contrast, the bad decks F and H produce a lower per draw constant loss of 50 dollars with even lower positive yields. Therefore, the subject must not only identify the good decks but also disambiguate for each good deck the loss and the gain signals. This type of task is difficult for human beings, and this difficulty has been modelled via loss aversion in the expectancy valence (section 2.4.1), via risk aversion in the prospect valence (section 2.4.2), and via separate loss and gain signal learning rate weighting in the outcome-representation model (section 2.4.3). In all three models, however, the common theme is that humans react differently to loss and gain.

An alternative explanation regarding frequency-gain effects may be provided by Hertwig et al., 2004, who find that in decisions from experience, that is under uncertainty, low frequency events are probabilistically underweighted. In the reversed IGT, good deck E and bad deck H exhibit with probability 0.10 (rare) gains of 1250 and 250 respectively. The under-weighting of these events may have an effect on cumulative task end $\bar{f}_G$ outcomes. However, as the reversed IGT also contains low frequency gain decks, a clear indication of probabilistic under-weighting cannot be obtained at the aggregated good versus bad decks and cumulative task end outcome measures used in this work.

The three Q-learning agents discussed in this chapter lack the signal extraction sophistication of the EV, PV, or ORL models. Thus the presence of learning decay with a maximisation criterion and the increased experience of negative yields produce a harder signal extraction problem. This is because due to learning rate decay, any deck yield streams, which initially produce a lower net yield will be established as "bad decks," even when they produce better results later on in the IGT. Based on slower CSUD iteration convergence as noted in Fig. 7.1, and complex exploration interactions as noted in Fig. 7.2 and Fig. 7.3, it would appear that the reversed IGT is harder to solve for the ε-Greedy, Boltzmann, and decaying ε-Greedy agents. This is possibly due to the low frequency of reward, leading to the accrual of negative Q-values, which in combination with learning rate decay, may produce a delay in reflecting any positive updates from decks E and G, leading to a lag, or inability, in learning the good decks.

The general result that high learning decay leads to vmPFC impaired behaviour is retained. For the reversed IGT environment, the Boltzmann agent, in terms of minimum CSUD loss and simulation outcome matches, produces the overall best results. In comparison, the decaying ε-Greedy agent, which had performed best in the original, re-shuffled, and random IGT environments, performs relatively poorly in that it produces the highest CSUD minimum loss. That is, the decaying ε-Greedy agent does poorly in terms of the CSUD squared distance measure from corresponding human IGT outcome means. The decaying ε-Greedy agent jitter plot in Fig. 7.5c, however, shows that the decaying ε-Greedy agent achieves better simulation verification results than the constant exploration ε-Greedy agent. Given the 1000 iteration limited CSUD budget, the decaying ε-Greedy agent searches do not appear to have fully converged, and it may be that the agent's poor performance regarding individual matches inside corresponding match zones is a side-effect of this iteration limit.

However, a case could also be made in terms of nonrational (heuristic) versus rational learners. It is possible that the decaying ε-Greedy agent's constant exploration decay is not best suited for solving the reversed IGT. The 2D and 3D exploration contours in Fig. 7.2 and Fig. 7.3 indicate complex exploration behaviour, which may be missed by the decaying exploration heuristic.

As with the original, re-shuffled, and random IGT environments, in the reversed IGT environment as shown in Fig. 7.6, fraction of good decks f_G

jitter density plots with exploration modulation on the horizontal axis reveal a tendency towards bi-modal f_G densities with little or no mass in the human outcome catchment areas. This effect is increased in 'greedy' exploration parameter configurations, such as in Fig. 7.6a with $\varepsilon = 0.1$, in Fig. 7.6b with $\tau = 5$, and in Fig. 7.6c with $\nu = 0.05$. At these hyper-parameter values, some agents produce very high f_G, while others produce very low fraction of good decks f_G. Hence, being greedy may not be the best survival strategy for solving the reversed IGT, and this in turn may explain the reason for requiring high exploration to match corresponding human outcomes. At higher exploration, the bi-modal effect is reduced and f_G outcomes coagulate around the human outcome catchment zones, especially for normal behaviour. In other words, higher exploration appears to forego extreme high and low outcomes, in favour of a mean centred outcome, where the overall risk of normative failure is lower. This risk smoothing effect of high exploration can also be observed in the original, re-shuffled, and random IGT jitter plot outcomes reported in chapter 6.

Chapter 8

The SGT Environment with Simple Reinforcement Learning Modelling via CSUD

The SGT (Soochow Gambling Task) has been introduced in section 4.1.4. To recap, the SGT environment consists of four decks A, B, C and D, where decks C and D are the good decks. These good decks produce high frequency losses with low frequency rewards; yielding per 10 draws a net positive gain. Decks A and B are the bad decks with low frequency fines and high frequency rewards, but yielding per 10 draws a net loss.

The SGT draw-by-draw yield structure is depicted in Appendix A.5. In the SGT, the determining events, that is the rewards in good decks C and D, and the losses in bad decks A and B, consistently happen rarely, making it challenging to determine per deck mean net yields. As noted in chapter 1, the SGT models both uncertain rewards and uncertain fines as rare events occurring with probability 0.2.

As Table 4.8 indicates, normal human subjects *do not* pass the SGT, scoring a mean fraction of good decks value $\bar{f}_G^H = 0.40$, with a $\pm 2SE$ range of $0.36 \leq \bar{f}_G^H \leq 0.44$. Note that the normal human mean fraction of good decks outcome remains below $\bar{f}_G = 0.5$, the score that could be achieved via pure random search. In that sense, normal humans must be using a deck selection strategy, which performs worse than random search. This paradoxical non-optimal human behaviour is also noted in Hertwig et al., 2004, where under decisions from experience, that is uncertainty, rare events are probabilistically undervalued. In the context of the SGT, probabilistically undervalued events lead to non-avoidance of high but rare fines, and could lead to the observed low, non-passing mean fraction of good deck $\bar{f}_G$ outcomes.

Agent	Boltzmann	ε-Greedy	Decaying ε-Greedy
Hyper-parameter			
Initial learning rate α_1	0.01 - 0.99	0.01 - 0.999	0.01 - 0.999
Normal learning rate decay λ_N	0.0001 - 0.3	0.03 - 0.30	0.03 - 0.32
Temperature τ	0.5 - 500		
Exploration ε		0.05 - 0.95	0.5 - 1.0[a]
Exploration decay ν			2.0e^{-8} - 0.03
CSUD Iterations	1000	1000	1000
Gradient Samples	1	1	3
IGT length	Q-learning samples		
100	750		

[a] Exploration ε refers to initial exploration ε_1.

TABLE 8.1
Search Methodology: SGT hyper-parameter CSUD search criteria by agent.

8.1 Search of the SGT environment

Table 8.1 summarises software agent CSUD search parameter constraints and attributes. As in chapters 6 and 7, broad parameter search ranges are used for the initial learning rate and exploration, while a smaller range is employed for normal learning rate decay. By limiting the learning rate decay range, it is intended to inject prior information, which captures the previous results of low learning rate decay leading to normal IGT behaviour. Since only normal human data is available for the SGT, learning rate decay is constrained to a range consistent with previously achieved normal learning rate decay values.

By construction, CSUD loss is minimised to the extent that search criteria are fulfilled, and in this sense, CSUD search can be seen as a contraction of a grid search space.

8.2 Search Results: SGT

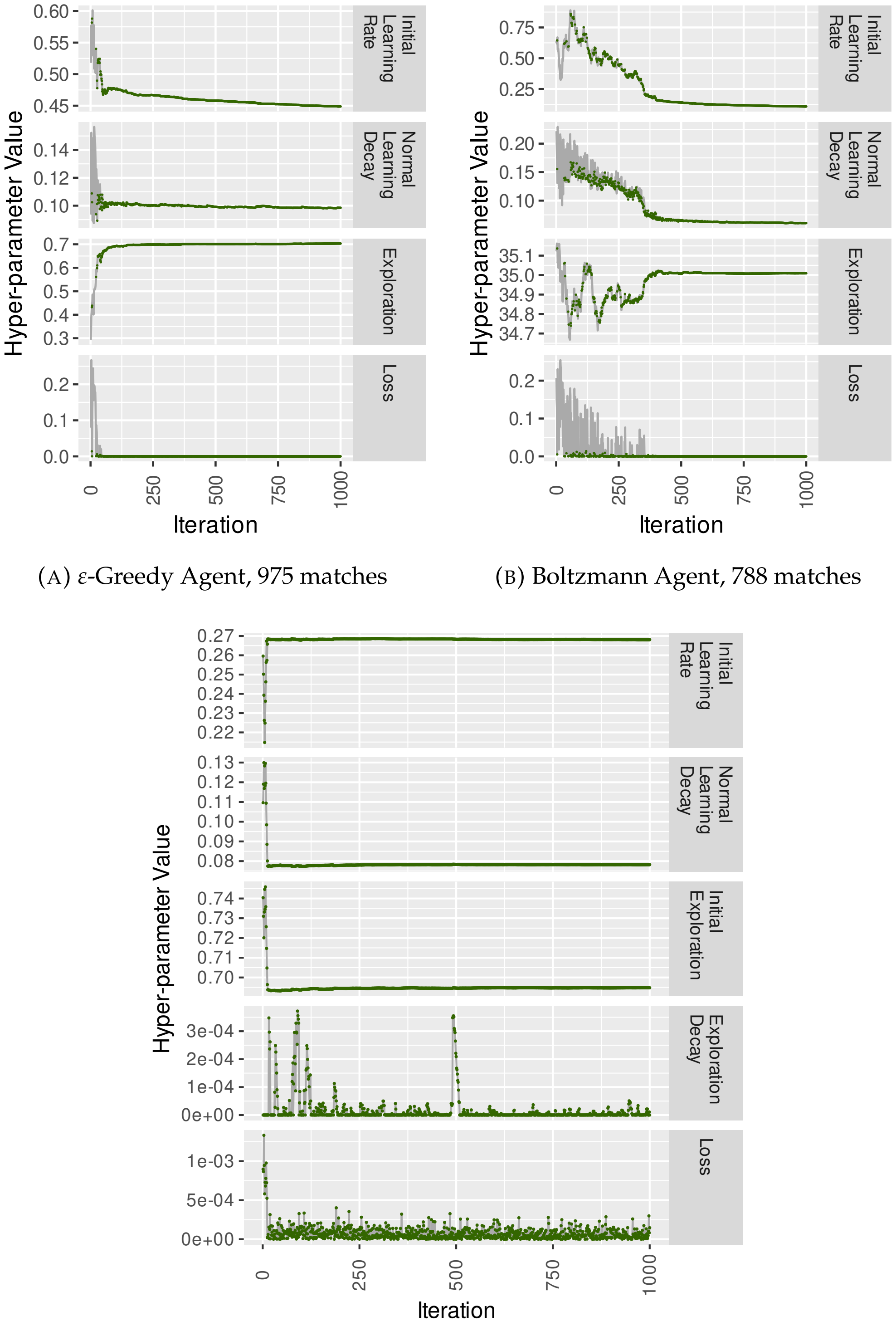
The results for the searches introduced above in section 8.1 are now presented. Each search is conducted via CSUD. The reinforcement learning layer is implemented using the Boltzmann, ε-Greedy, and decaying ε-Greedy agents discussed in sections 5.2.1, 5.2.2, and 5.2.4 respectively. As only normal human mean fraction of good decks $\bar{f}_G^H$ data is available, CSUD searches will aim to find hyper-parameter combinations, which produce performance matches for normal human IGT outcomes.

For ease of comparison, results are presented for the ε-Greedy, Boltzmann, and decaying ε-Greedy agents side-by-side. Fig. 8.1 and Table 8.2 present CSUD search results in graphic and tabular forms respectively.

Fig. 8.1a shows that in CSUD searches, the ε-Greedy agent initially has wide search space traversal for all hyper-parameters, before settling down.

Fig. 8.1b shows that the Boltzmann agent has wide initial traversal for the initial learning rate and learning rate decay. However, Boltzmann agent exploration temperature τ does not change much. It is believed this is due to a parameter scale factor consideration. That is, the scale of perturbations is not large enough to induce sizeable variations in exploration temperature. However, fundamental agent conclusions are not affected in so far as the 2D grid search contours in Fig. 8.2b suggest that the range $5 \leq \tau \leq 200$ should be able to produce for some (normal) learning rate decay λ_N, mean fraction of good deck $\bar{f}_G$ outcomes residing in the human catchment zone. Hence at least for the SGT searches, if the τ perturbation scale issue is corrected, one would expect that multiple λ_N values are found for multiple τ values. In CSUD search the telltale sign of such a situation is when across search iterations, τ and λ_N settle at multiple locations. It would be interesting to see if under such circumstances, another estimation method perhaps maximum likelihood could produce definitively unique λ_N and τ estimates.

Fig. 8.1c shows that the decaying ε-Greedy agent search settles relatively quickly. Exploration decay ν exhibits, along with loss, some dispersion, however the range of this dispersion is small. In general, all agents display relatively good (qualitative) convergence at the 1000 iteration cut-off point. This indicates that given the respective hyper-parameter constraints reported in Table 8.1, the agents are able to consistently produce, within the iteration budget, (cumulative) mean fraction of good decks $\bar{f}_G$ outcomes, which lie within the respective human catchment areas.



(A) ε-Greedy Agent, 975 matches

(B) Boltzmann Agent, 788 matches

(C) Decaying ε-Greedy Agent, 1000 matches

FIGURE 8.1: SGT CSUD searches. Green dots indicate matches defined as agent cumulative mean fraction of good decks $\bar{f}_G$ residing in corresponding human catchment zones. All agents achieve a large number of matches. At 1000 iterations, agents display overall good search convergence. Details in text.

Agent	ε-Greedy	Boltzmann	Decaying ε-Greedy
Hyper-parameter			
Initial Learning Rate α_1	0.450	0.189	0.268
Normal Learning Decay λ_N	0.099	0.073	0.0780
Exploration	ε = 0.702	τ = 34.997	ε_1 = 0.693 ν = $2e^{-8}$
Minimum Loss	6.66e^{-8}	1.24e^{-8}	2.00e^{-9}
Matches	975	788	1000

TABLE 8.2
SGT minimum loss CSUD search matches after 1000 iterations.

Table 8.2 indicates that given the 1000 iteration search budget, the decaying ε-Greedy agent produces 1000 matches and delivers the lowest minimum loss CSUD search score. In comparison, the ε-Greedy and Boltzmann agents achieve 975 and 788 matches respectively. For the decaying ε-Greedy agent, 3 repetition gradient averaging is employed. For all agents, normal learning decay λ_N achieves similar values. The initial learning rate α_1 stretches over a broader range, however remaining under 0.5. Exploration related hyper-parameters remain high, but are lower than those observed in the original, re-shuffled, random, and reversed IGT environments. Exploration decay ν is close to zero, indicating constant exploration, close to that discovered for the ε-Greedy agent (ε_1 = 0.693 versus ε = 0.702).

Table 8.3 displays agent grid search configurations for CSUD search results verification. As in previous application chapters, the CSUD search discovered minimum loss agent hyper-parameter values are used to construct a small search grid around these values.

Fig. 8.2 and Fig. 8.3 show 2D and 3D CSUD grid search verification contours. In Fig. 8.2, the dark zone represents the normal human IGT outcome match area. Solid black contours show response to learning decay λ at CSUD selected minimum loss hyper-parameter values.

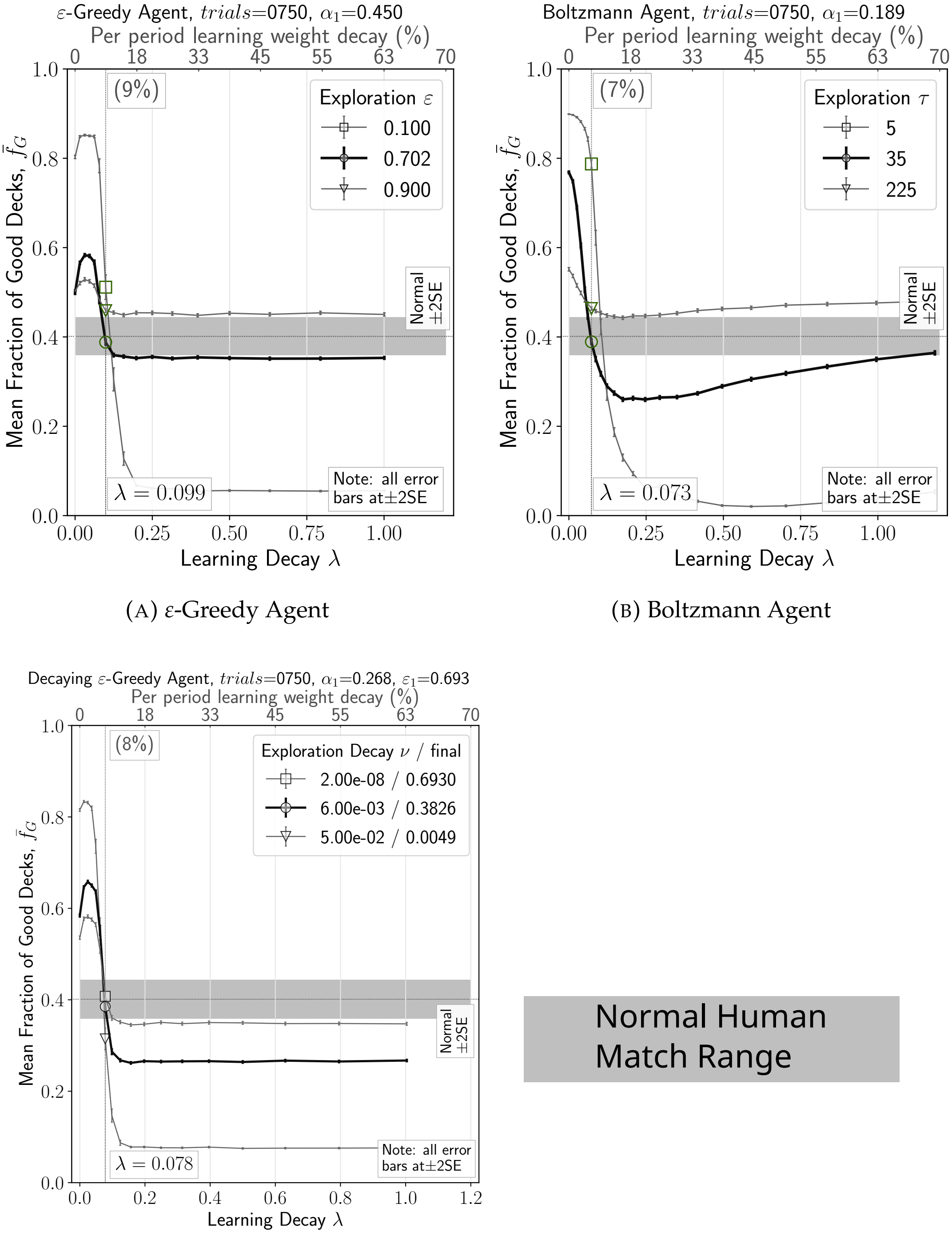
In terms of mean fraction of good decks $\bar{f}_G$, Fig. 8.2 contours for all agents cross over each other, and this indicates more complex exploration and learning decay interactions than seen in other IGT environments. Note that with respect to the dark gray catchment zones, CSUD has not identified any solutions with lower exploration. For example, in Fig. 8.2a, visual

Agent	ε-Greedy	Boltzmann	Decaying ε-Greedy
Hyper-parameter			
Initial Learning Rate α_1	0.1, 0.450, 0.75, 0.9	0.05, 0.189, 0.5, 0.9	0.01, 0.268, 0.45, 0.9
Normal Learning Decay λ_N[a]	0.099	0.073	0.0780
Exploration	$\varepsilon = 0.1, 0.4, 0.702, 0.9$	$\tau = 5, 34.997, 75, 225$	$\varepsilon_1 = 0.25, 0.45, 0.693, 0.9$ $\nu = 2e^{-8}, 0.006, 0.015, 0.05$
IGT length	Q-learning samples		
100	750		

[a] The learning rate decay and exploration decay grids are constructed around the two values above. Appendix C provides the construction method.

TABLE 8.3
CSUD Verification. SGT hyper-parameter grid search criteria for the ε-Greedy, Boltzmann, and decaying ε-Greedy agents.

inspection indicates a match solution with $\varepsilon = 0.1$ and $\lambda_N \approx 0.1$. Similar results are observed for the remaining two agents. Other solutions may possibly have been discovered at longer iterations or with different initial starting points. However, it is noted that at the selected agent normal learning rate decay λ_N values, 2D exploration contours are close to or go through an inflexion point and have high change rates; these two conditions create more challenging searches. In general, rational search methods would not consider optima at or near inflection points. Nonrational CSUD, however, is able to tune through such locations and provide insight as to the configuration of primitive decision making hyper-parameters, whose values in this instance produce mean fraction of good decks outcome solutions, which perform worse than pure random search.



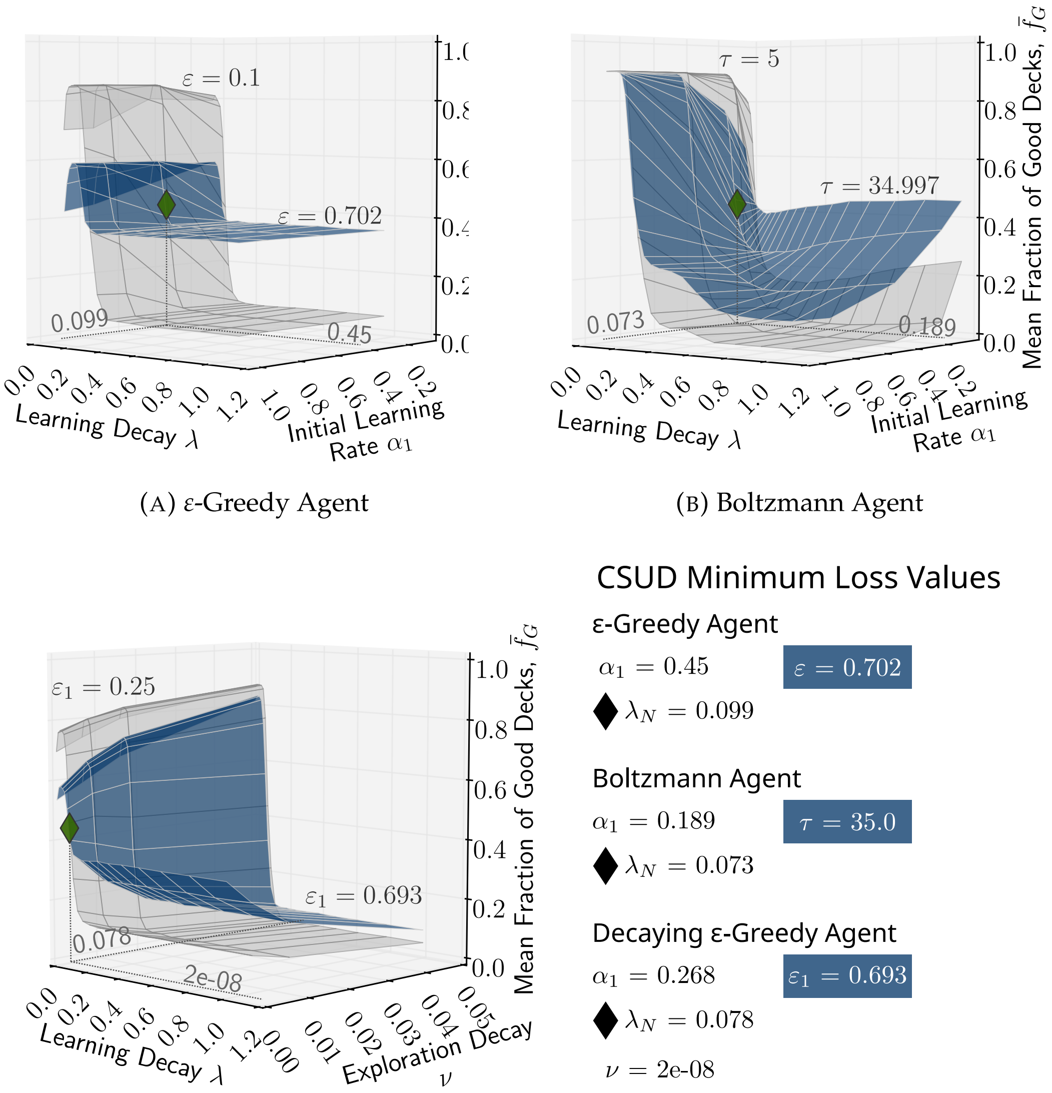
(A) ε-Greedy Agent

(B) Boltzmann Agent

(C) Decaying ε-Greedy Agent

(D) Legend

FIGURE 8.2: SGT CSUD verification grid search 2D contours. The dark gray zone represents the normal human IGT outcome match area. Solid black contours show response to learning decay λ at CSUD selected minimum loss hyper-parameter values. Human outcome range matches occur in areas with high slope. vmPFC impaired human SGT outcomes have not been reported in the literature. Details in text.



(A) ε-Greedy Agent

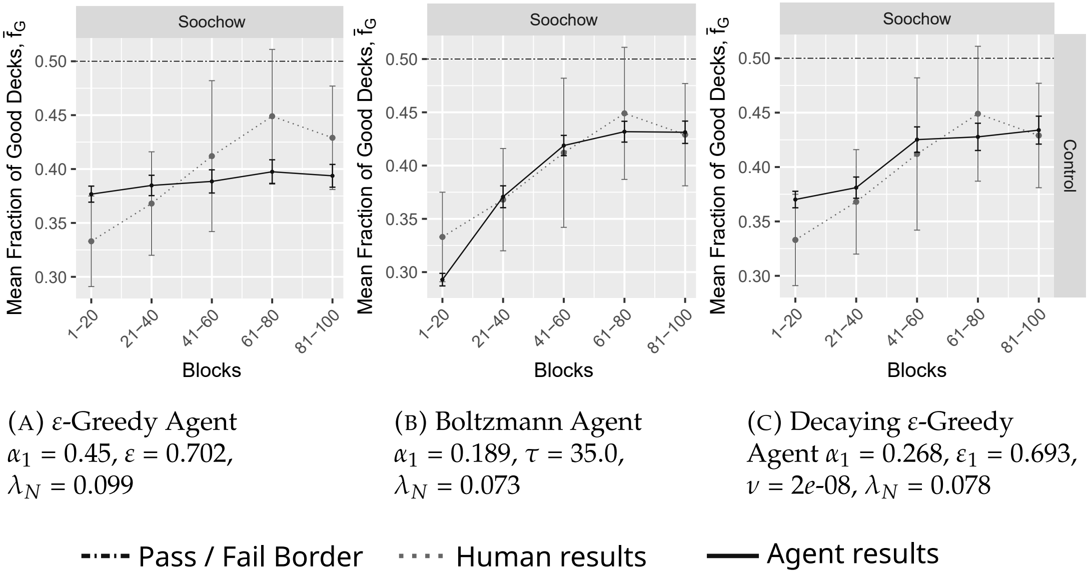
(B) Boltzmann Agent

(C) Decaying ε-Greedy Agent, $\alpha_1 = 0.268$

(D) Legend

FIGURE 8.3: SGT CSUD verification grid search 3D contours. Blue coloured surfaces show response to learning decay λ at CSUD selected minimum loss hyper-parameter values. The diamond shape marks the CSUD minimum loss normal learning decay rate λ_N. vmPFC impaired human SGT outcomes have not been reported in the literature. Details in text.

Fig. 8.3 shows 3D CSUD verification contours. Blue coloured surfaces show response to learning decay λ at CSUD selected minimum loss hyper-parameter values. The diamond shape marks the CSUD minimum loss normal learning decay rate $\lambda_N = 0.078$. Fig. 8.3b and Fig. 8.3c indicate that when exploration is allowed to vary, complex mean fraction of good decks



(A) ε-Greedy Agent α_1 = 0.45, ε = 0.702, λ_N = 0.099

(B) Boltzmann Agent α_1 = 0.189, τ = 35.0, λ_N = 0.073

(C) Decaying ε-Greedy Agent α_1 = 0.268, ε_1 = 0.693, ν = 2e-08, λ_N = 0.078

(D) Legend

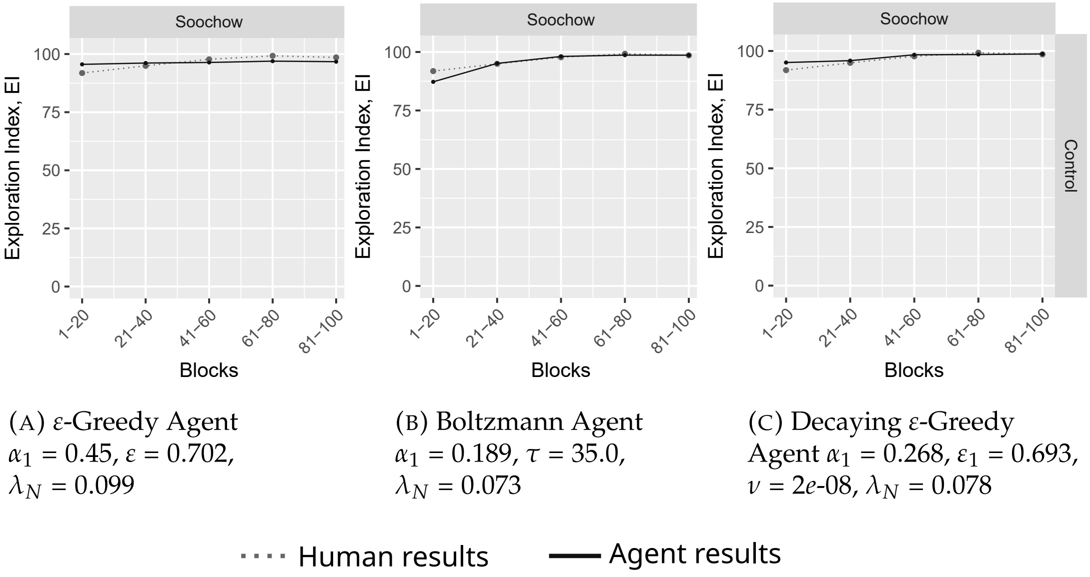
FIGURE 8.4: SGT agent 20-draw blocks comparison at CSUD minimum loss search matches noted in Table 8.2. Human results in dotted light gray. Agent results in solid dark gray, averaged from 750 samples. All error bars at ±2SE. When error bars are taken into account agent and human 20-draw block performance appears to provide good matches. Details in text below.

$\bar{f}_G$ surfaces result. Such complex surfaces are not observed with the previously discussed original, re-shuffled, random, or reversed IGT environments, suggesting that the SGT is more difficult to solve. Further, replicating human performance is more difficult as well in the sense of the matching human outcome generating agent hyper-parameters being resident in surface zones of high curvature (high slope).

Fig. 8.3b indicates interaction between learning decay λ and the initial learning rate α_1 as exploration temperature τ increases. Similarly Fig. 8.3c shows increasing interaction between learning decay λ and exploration decay ν as initial exploration ε_1 increases.

Fig. 8.2 and Fig. 8.3 suggest that for all agents, the human performance zone is located in a mean fraction of good decks $\bar{f}_G$ range where the 2D and 3D surfaces cross over. With the data available however, it is not possible to ascertain whether this observation is significant or coincidental.

Fig. 8.4 and Fig. 8.5 show 20-draw blocked mean fraction of good decks $\bar{f}_G$ and exploration index (EI) outcomes respectively. Human reference results are in light gray dotted lines, and agent results, averaged from 750



(A) ε-Greedy Agent $\alpha_1 = 0.45$, $\varepsilon = 0.702$, $\lambda_N = 0.099$

(B) Boltzmann Agent $\alpha_1 = 0.189$, $\tau = 35.0$, $\lambda_N = 0.073$

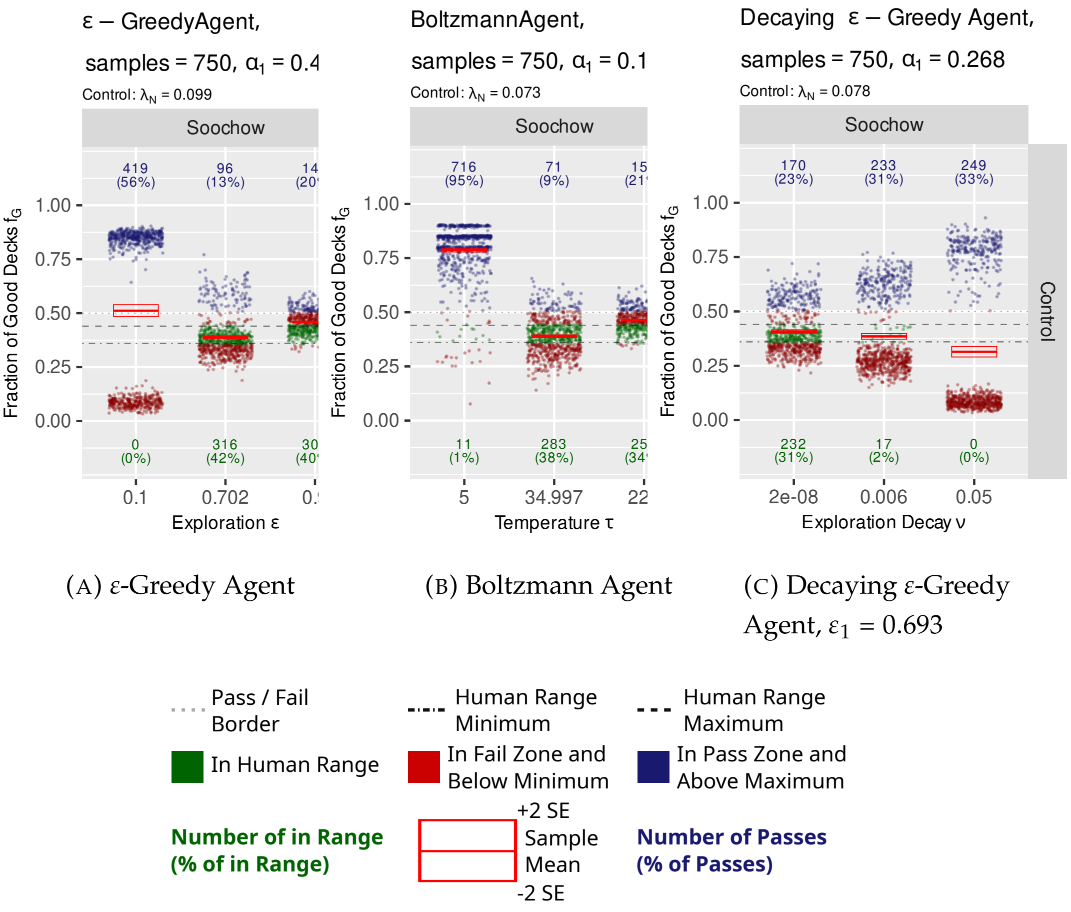
(C) Decaying ε-Greedy Agent $\alpha_1 = 0.268$, $\varepsilon_1 = 0.693$, $\nu = 2e\text{-}08$, $\lambda_N = 0.078$

(D) Legend

FIGURE 8.5: SGT agent 20-draw blocks exploration index (EI) comparison at CSUD minimum loss search matches noted in Table 8.2. Human results in dotted light gray. Agent results in solid dark gray, averaged from 750 samples. Human subject and agent exploration index responses appear relatively similar. Further details in text below.

samples, appear in dark gray solid lines.

Fig. 8.4, where the dash-dotted line at $\bar{f}_G = 0.5$ indicates the IGT normative pass threshold, shows that all agent 20-draw block $\bar{f}_G$ results remain within the human reference $\pm 2SE$ catchment zones. However, Fig. 8.4b and Fig. 8.4c show that Boltzmann and decaying ε-Greedy agents respectively better mirror human reference points, indicating support for varying exploration throughout the IGT. Agent exploration indices (EI) in Fig. 8.5 support the above finding. More importantly, exploration index (EI) values remain high, close to 100 throughout the IGT, indicating that while exploration varies, it does not move towards full exploitation as the IGT progresses.



(A) ε-Greedy Agent

(B) Boltzmann Agent

(C) Decaying ε-Greedy Agent, $\varepsilon_1 = 0.693$

(D) Legend

FIGURE 8.6: SGT. Comparison of repeated agent simulation outcomes to human IGT results. CSUD minimum loss exploration values are $\varepsilon = 0.702, \tau = 34.997, \nu = 2e^{-8}$ for the ε-Greedy, Boltzmann, decaying ε-Greedy agents respectively. ε-Greedy based agents exhibit a tendency towards bi-modal f_G for high exploration. At CSUD selected exploration values, ε-Greedy, Boltzmann and decaying ε-Greedy agents achieve 42%, 38%, and 31% respectively of simulation results inside human match zones. Full details are in the text.

Fig. 8.6 provides agent jitter plot density summary for the fraction of good decks f_G outcomes obtained at the CSUD selected minimum loss agent hyper-parameter values for 750 simulation samples of the SGT environment for normal (control) behaviour. Green dots mark outcomes inside the human performance range. Blue dots mark additional normative pass results, whereas red dots mark additional normative pass fails. Green numbers give total human performance zone matches out of 750 samples, and the values in brackets indicate percentages matched. A high percentage matched value is indicative of predictive simulation success. The dashed and dash-dotted horizontal lines indicate the maximum and minimum respectively of the human match range. Finally the red bars and box indicate central tendency in terms of the mean and ± 2 SEs (standard errors).

Fig. 8.6 shows that agents do not achieve many simulation results, which are inside human match zones. At CSUD selected exploration $\varepsilon = 0.702$, the ε-Greedy agent achieves 42% normal matches. At CSUD selected exploration temperature $\tau = 34.997$, the Boltzmann agent achieves 38% normal matches. At CSUD selected initial exploration $\varepsilon_1 = 0.693$ and exploration decay $\nu = 2e^{-8}$, the decaying ε-Greedy agent achieves 31% normal matches. Interestingly for any agent, increasing exploration leads towards a density migration towards the normal human match zones. As exploration decreases as seen in Fig. 8.6b, as τ reduces from 225 to 5, the Boltzmann agent does not move towards a bi-modal f_G outcome density, but exhibits some banding.

8.3 Discussion: SGT

The SGT embodies a complex decision making environment, where humans on average, fail to make the correct decisions subject to the SGT task limitation of 100 turns. The results show that agents exhibit hyper-parameter ranges, where they can produce decisions superior to those achieved by normal humans. However, only the rational Boltzmann agent retains unimodal mean fraction of good decks $\bar{f}_G$ densities.

The SGT (Soochow Gambling Task) results indicate that, this task is generally more difficult to solve. Note for example in Fig. 8.6a how the ε-Greedy agent at low exploration $\varepsilon = 0.1$, produces strongly bi-modal outcomes, where agents cluster either high above or very low below the human catchment zone. As seen in Fig. 8.6b, even the rational Boltzmann agent at $\tau = 5$ produces skewed outcomes with a heavy tail towards worse outcomes. Agents are capable of producing better than human mean fraction of good decks results. The agents also exhibit, however, hyper-parameter ranges where human SGT outcome results are matched. Such ranges are associated with mild exponential learning rate decay of around 7%-10% per period, or a normal learning decay λ_N range of approximately $0.073 - 0.1$.

For the original, re-shuffled, and random IGT environment simulation results in chapter 6, the ε-Greedy, Boltzmann, and decaying ε-Greedy agents exhibit normal learning rate decay λ_N values of $\lambda_N = 0.10$, $\lambda_N = 1e^{-04}$, $\lambda_N = 0.106$ respectively. Hence the ε-Greedy and decaying ε-Greedy agents appear to exhibit for the original, re-shuffled, random, and SGT cases relatively similar numerical values for λ_N. The Boltzmann agent λ_N value, however, is considerably lower for the original, re-shuffled, and random deck simulations.

It is possible that decision making hyper-parameter values such as the learning decay rate λ_N are themselves calibrated during the decision making process. The question of whether universal decision making hyper-parameter values exist, or whether such hyper-parameters are themselves calibrated during decision making would provide an interesting extension of the current work. It is possible that in a heuristic decision making context, with learning rate decay imposed event sampling cut-off, learning rate decay somehow reacts to the underlying sampled process frequency dynamics, for example as in a frequency domain context.

As in previous IGT environments, learning rate decay λ plays a key role in mean fraction of good decks $\bar{f}_G$ outcomes. However, it is not possible to obtain clear visual support, as in Fig. 7.2 with the reversed IGT, for the proposition that matching human outcomes requires high exploration. This is because as seen in Fig. 8.2, the SGT mean fraction of good decks $\bar{f}_G$ contours cross over, and there is no human vmPFC impaired data, which would further help clarify the role of high exploration.

As Fig. 8.2a and Fig. 8.2c indicate, the ε-Greedy and decaying ε-Greedy agents produce similar solutions, and this is due to the very low exploration decay value $\nu = 2e^{-8}$, coupled with the closeness of ε-Greedy agent exploration $\varepsilon = 0.702$ to decaying ε-Greedy agent initial exploration $\varepsilon_1 = 0.693$.

The adaptive ε-Greedy and decaying ε-Greedy agents are not considered any longer. The adaptive ε-Greedy agent appeared only in chapter 6 as a rational agent capable of adjusting exploration in response to the temporal difference error.

The decaying ε-Greedy agent, on the other hand, is not considered any further because it performs very close to the simpler constant ε-Greedy model. The decaying ε-Greedy agent presents good results for replicating human IGT outcomes. These results, however, obtain at relatively low exploration decay ν values with $\nu = 0.00842$, $\nu = 0.00609$, and $\nu = 2e^{-8}$ for the original IGT and its variants, the reversed IGT, and the SGT environments respectively.

It would appear that the decaying ε-Greedy agent presents as a truly heuristic model, with the exploration parameter ν acting as a way to slightly dampen exploration so as to improve on the results of the constant exploration ε-Greedy variant. So the decaying ε-Greedy agent acts more as a fitted model than as a generative model, with ν values making sure that good fits are obtained. The constant ε-Greedy agent model, however, achieves relatively good results and can be seen as a simple generative model. It is on this conceptual basis that, the decaying ε-Greedy model is omitted in further discussions in favour of its simpler alternative. With relevant theoretical work, the decaying ε-Greedy agent may yet become a generative heuristic alternative. Such efforts, however, remain beyond the scope of this work.

Chapter 9

Reinforcement Learning: The Iowa Gambling Task with Discount Rate and Trace Decay

Up to now, it has been argued that in the IGT, the initial learning rate α_1, learning rate decay λ, and exploration sufficiently capture single-state Q-learning decision making dynamics. It was also argued that the 100-draw IGT was not long enough to generate a preference for immediate over future yields. Therefore the discount rate γ was set to $\gamma = 1$.

Further, it was assumed that agents do not employ any persistence dynamics attributed to specific yield streams. That is, in their Q-value attributions, the agents only considered action net yield at face value, but did not apply any additional weighting to actions deemed to produce favourable outcomes. One way to achieve such additional action tracking is through trace decay, where frequently chosen actions exert an additional influence on Q-values. Trace decay can also be used to track multi-period action attribution of yields. Here trace decay is denoted by κ. This chapter introduces the discount rate γ and trace decay κ in order to asses to what extent the earlier assumptions regarding their omission may be justified.

In iterative learning, the discount rate $0 < \gamma < 1$ is used to reduce the perceived value of future unrealized net yield streams. In short, the discount rate induces a preference for the present. It is important to recall that test subjects in the IGT variants tested here do not know the length of the task. From the perspective of these participants, a discount rate γ less than 1 could be used to capture various reasons for losing interest in the task as it progresses, where this loss of interest is mathematically formulated as decreasing value attribution of future outcomes. Alternatively, it could be that the sheer uncertainty over the length of the task induces a preference for immediate rewards, perhaps in the sense of the participant using a high discount rate to affect a single period look-ahead. Finally on a technical note, it is very difficult to solve finite duration reinforcement learning problems when the end period is unknown, and as in the case of the IGT there is no other absorbing state. One may for example need to establish a probability for the length of the task. One easy approach, however, when task end is unknown is to drive the problem into infinity in the presence of a discount rate less than 1. In reinforcement learning (for an infinite horizon), a discount rate less than one is an additional requirement for convergence and the discovery of optimal policies, see for example Szepesvári (2010, p. 9).

Unlike the forward-looking discount rate, trace decay κ, looks backwards and is its intended use consists of accounting for gestation lag related outcome attribution. One shortcoming of (discrete time) iterative learning is the difficulty in attributing to previous actions any outcomes, which require more than one iteration to affect yield. Trace decay creates a mechanism for capturing such actions, which exhibit a gestation lag. In such cases trace decay is referred to as an eligibility trace (Sutton & Barto, 2018, p. 287).

Trace decay can also be used as a mathematical technique for instituting choice perseveration, a phenomenon where having chosen an option leads to it being chosen again regardless of its value. The dynamics (9.1) of the model introduced here, differ from those in Miller et al., 2019, where choice perseverance is indeed action independent, and where choice perseverance and action value attribution are modelled separately and then combined linearly via a controller unit. The model below also differs from the trace decay implementation in Sutton and Barto, 2018, p. 292, where eligibility trace updates include a value function contribution, that is eligibility traces are value dependent. In the single state Q-value IGT modelling here, however, trace decay mathematics are implemented so as to produce a hybrid of the two above views: (1) as in Sutton and Barto, trace decay directly influences Q-learning without any controller logic, but (2) as in Miller et al., a choice perseverance component can be said to be present as the incrementing of a trace decay solely depends on whether an action has been chosen but not its value.

Such an approach is adopted for the sake of keeping a simple model, which can algorithmically capture any potential multi-period dynamics. When learning is not vmPFC impaired, then in principle the model below can lead to an increase in the Q-values of good, and to a decrease in the Q-values of bad decks; the trace decay will function more like an eligibility trace. In the case of vmPFC impairment, when bad decks are more frequently chosen, increments in Q-values should then lead to the bad actions having high Q-values and thereby affect choice perseverance. That is, the model here is expected to magnify good as well as poor choice selection. Further it is noted that low trace decay κ and low discount rate γ reduce the effective decision making horizon, while high trace decay κ and high discount rate γ increase it. Finally, the original, re-shuffled, and random IGT environments employed here do not themselves contain any multi-period effects. Nevertheless, it is interesting to see if any agents would behave as if such effects were present.

9.1 The Amended Single State Q-learning Model, ARA(κ)

In the classical SARSA(κ) implementation (Sutton & Barto, 2018, pp. 303-307), each state-action (s, a) pair has its own trace decay (eligibility trace). Here as there is only a single state, trace decay, in the hybrid sense, is applied to each individual action.

SARSA itself denotes the sequential process state-action → response → state-action, where the initial and consequential state-action pair may be different. With only a single state, the SARSA mnemonic reduces to ARA, action → response → action. Further in the ARA implementation in this chapter, the initial (pre-response) action is persisted as the post-response action. That is unlike SARSA, in the present ARA implementation, the initial pre-response and forecasted post-response actions are identical and equal to the pre-response action, which is the action the agent will take, and may be the Q-value maximum or an exploratory action.

Hence, the current ARA implementation uses hybrid trace decay, and also in formulating next period's Q-values, employs a discounted and traced forecast resulting from the one-period-ahead persisted pre-response action.

The model in this chapter will be referred to as ARA(κ), where κ indicates (hybrid) trace decay. The literature has typically used λ for trace decay, in particular for eligibility traces. However in this work, λ indicates learning rate decay, and therefore the use of κ is retained to symbolise trace decay.

Given the pre-response action $\tilde{a}$, and any action a, let $Q(a)$ be an unknown value function, let $e(a)$ denote the eligibility trace for action a, and let $Q_t(a)$, $e_t(a)$ denote the t^{th} iterations. Then the computational ARA(κ) model is written as

$$\delta_t = x_t^{\tilde{a}} + (\gamma - 1)\, Q_{t-1}(\tilde{a}) \tag{9.1a}$$

$$e_{t-1}(\tilde{a}) = e_{t-1}(\tilde{a}) + 1 \tag{9.1b}$$

$$Q_t(a) = Q_{t-1}(a) + \alpha_t \delta_t e_{t-1}(a) \tag{9.1c}$$

$$e_t(a) = \gamma \kappa e_{t-1}(a) \tag{9.1d}$$

where γ is the discount rate, δ_t is a version of the temporal difference error, $x_t^{\tilde{a}}$ is the net yield following the pre-response action $\tilde{a}$, $Q_{t-1}(\tilde{a})$ is used to generate the Q-value *forecast* of the pre-response action $\tilde{a}$; that is, it is assumed that the same action if chosen, would produce the same Q-value contribution.

(9.1b) indicates that if an action is frequently selected, then its trace decay is incremented frequently, and this in turn leads to larger Q-value accruals. Note that, the trace decay increment is 1 and thus independent of the actual action value. Also note that the time index on both the left and right hand sides of (9.1b) is $t-1$. This convention is used here to indicate that (9.1b) is always applied prior to (9.1d). (9.1d) shows that in general the trace for each action decays every iteration. Therefore, seldom selected actions lead to smaller Q-value accruals.

9.2 Software agent implementations

The only agents considered here are the ε-Greedy and Boltzmann agents as presented in sections 5.2.2 and 5.2.1 respectively. The ε-Greedy agent has constant exploration ε, and the Boltzmann agent specifies exploration temperature τ, which controls how Q-values contribute to individual action selection probabilities.

The adaptive and decaying ε-Greedy agents are no longer considered. The rationale for these omissions is discussed in section 8.3. To review, the adaptive ε-Greedy agent obtains good cumulative mean fraction of good decks results. Due to its ability to substantially reduce exploration, however, as noted in Fig. 6.17 and Fig. 6.18, in simulation outcomes, this agent did not exhibit the 20-draw block mean fraction of good deck $\bar{f}_G$ and exploration index (EI) properties of corresponding human IGT outcome cases, especially for the re-shuffled IGT environment. Moreover the adaptive ε-Greedy agent could not replicate the key human IGT outcome result that for the re-shuffled IGT environment, normal and vmPFC impaired behaviour configurations should not exert a statistically significant factor effect on the cumulative mean fraction of good decks.

The decaying ε-Greedy agent, on the other hand, with its fixed exploration exponential decay heuristic, can accommodate residual exploration, and does match human 20-draw block IGT mean fraction of good deck and exploration index results. These results, however, obtain at relatively low exploration decay ν values with $\nu = 0.00842$, $\nu = 0.00609$, and $\nu = 2e^{-8}$ for the original IGT and its variants, reversed IGT, and SGT environments respectively. Since exploration decay is very small, the decaying ε-Greedy model is omitted in favour of its simpler alternative the (constant exploration) ε-Greedy variant.

9.3 Methodology: Joint Search of the Original, Reshuffled, and Random IGT environments

The general methodology remains as discussed in section 6.1. A joint CSUD (hyper-parameter) search is conducted of the original, re-shuffled, and random IGT environments, looking for a hyper-parameter configuration, which may lead to mean fraction of good decks $\bar{f}_G$ outcomes residing within respective human catchment ranges for normal and vmPFC impaired IGT outcome behaviour. This CSUD search is conducted on a per agent basis.

Note that as discussed in section 7.2, a joint search for all IGT environments is not undertaken. A joint CSUD hyper-parameter search for the original IGT, its variants, the reversed, and SGT environments has been attempted but not found to produce hyper-parameter combinations, which could be verified in subsequent simulations. Further it is believed that the

Agent Hyper-parameter	Boltzmann	ε-Greedy
Discount rate γ	0.15 - 0.85	0.5 - 0.99
Trace decay κ	0.15 - 0.75	0.25 - 0.99
Initial learning rate α_1	0.01 - 0.6	0.01 - 0.999
Normal learning rate decay λ_N	1.0e^{-8} - 0.22	0.03 - 0.30
vmPFC impaired learning rate decay λ_{vmPFC}	0.22 - 0.8	0.25 - 1.2
Exploration	$\tau = 0.5 - 500$	$\varepsilon = 0.05 - 0.70$
CSUD Iterations	2000	1000
Gradient Samples	4	1
IGT length	Q-learning samples	
100	750	

TABLE 9.1
ARA(κ) Search Methodology. Joint original, re-shuffled, random IGT hyper-parameter CSUD search criteria.

learning rate decay λ parameter is sensitive to the underlying IGT environment yield frequency distribution characteristics, and that on that basis, searches are limited to IGT environment combinations with similar frequency dynamics, with the related original, re-shuffled, and random IGT environments providing the largest common yield characteristics search set.

Therefore, for each agent architecture, CSUD search loss is minimised for hyper-parameter configurations which produce a match in five outcome cases: normal original, normal re-shuffled, normal random, vmPFC impaired original, and vmPFC impaired re-shuffled. The hyper-parameters CSUD search investigates are the discount rate γ, trace decay κ, the initial learning rate α_1, normal learning rate decay λ_N, vmPFC impaired learning rate decay λ_{vmPFC}, and exploration.

By design, (per agent) CSUD loss is minimised to the extent that search criteria are fulfilled. It is in this sense that CSUD can be seen as a contraction of a grid search space. Initially hyper-parameter ranges were broad, and where applicable, identical accross models. However, preliminary CSUD searches indicated that search direction could be further focused by limiting the ranges of γ, κ, α_1, λ_N, and λ_{vmPFC} on a per agent basis. Table 9.1 summarises the final software agent CSUD search hyper-parameter constraints and attributes.

Because CSUD is a stochastic, gradient driven search technique, sometimes a single gradient evaluation is not sufficient to produce a reliable gradient estimate. Under such circumstances, for the same set of hyper-parameter values, multiple gradient samples maybe obtained and then averaged (Spall, 1992). The 'Gradient Samples' entry in Table 9.1 indicates that gradient sampling was employed for the Boltzmann but not for the ε-Greedy agent.

It is not certain why the Boltzmann agent required gradient sampling to produce high quality CSUD search induced hyper-parameter updates. It is thought that this result obtains because the Boltzmann agent produces tightly clustered mean fraction of good decks outcomes, and such tight clustering increases the sensitivity of gradient calculations.

The CSUD search results are verified by running 750 repeated software agent simulations in a small grid search around the CSUD discovered hyper-parameter values. The grid search verification results are then analysed and discussed with respect to software agent outcomes.

9.4 CSUD Search Results

For ease of comparison, results for the ε-Greedy and Boltzmann agents are presented side-by-side. Table 9.2 and Fig. 9.1 present CSUD search results in tabular and graphic forms respectively.

Table 9.2 shows that as in the simple reinforcement learning model results in chapter 6, normal IGT outcomes are associated with low learning decay, $\lambda_N = 0.080$ and $\lambda_N = 0.00291$ for the ε-Greedy and Boltzmann agents respectively. In contrast, vmPFC impaired IGT outcomes are associated with high learning decay, $\lambda_{vmPFC} = 0.270$ and $\lambda_{vmPFC} = 0.402$ for the ε-Greedy and Boltzmann agents respectively.

Exploration with $\varepsilon = 0.612$ and $\tau = 225$ remains high for both the ε-Greedy and Boltzmann agents respectively. That is, as in chapters 6 and 8, for agents to match human IGT outcomes, exploration must be high. The initial learning rate α_1 for either agent is lower than the corresponding value in the simple reinforcement learning model. For example with ARA(κ), the ε-Greedy agent CSUD search produces $\alpha_1 = 0.137$, while the corresponding CSUD search in the simple model in chapter 6 produces $\alpha_1 = 0.417$.

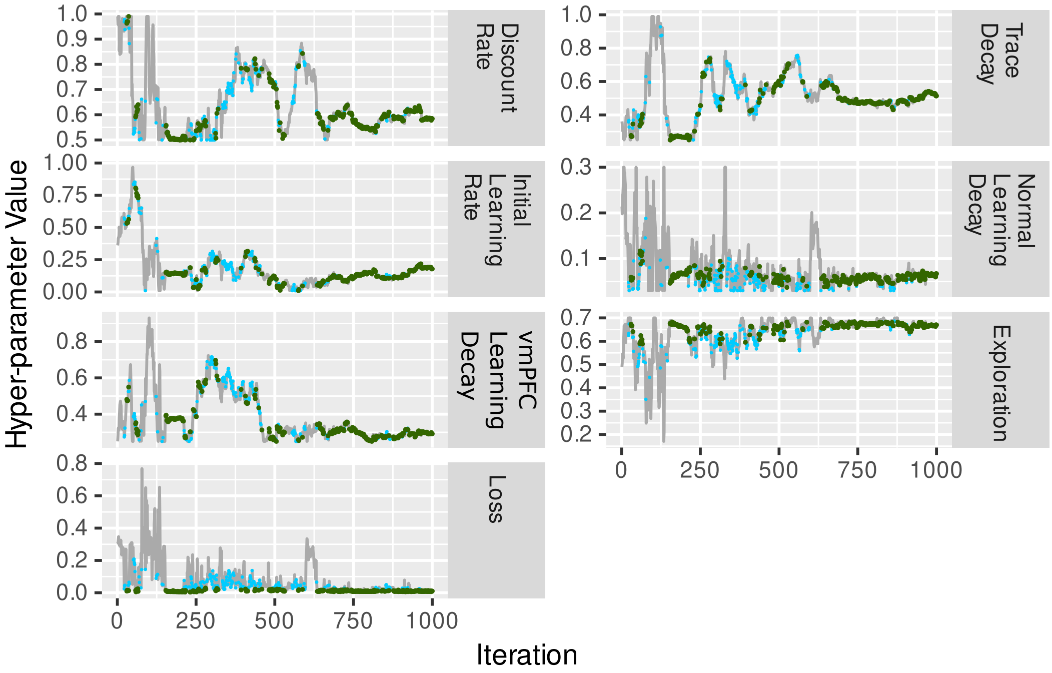
Agent Hyper-parameter	ε-Greedy	Boltzmann
Discount rate γ	0.514	0.206
Trace decay κ	0.264	0.626
Initial Learning Rate α_1	0.137	0.290
Normal Learning Decay λ_N	0.080	0.00291
vmPFC Impaired Learning Decay λ_{vmPFC}	0.270	0.402
Exploration	$\varepsilon = 0.612$	$\tau = 225$
Minimum Loss	0.00386	0.0222
Matches	389	16 / 951[a]

[a] 16 full CSUD matches. 951 partial matches. Partial matches include normal and vmPFC impaired behaviour, original and re-shuffled decks. Full matches consist of partial matches plus the normal behaviour random IGT.

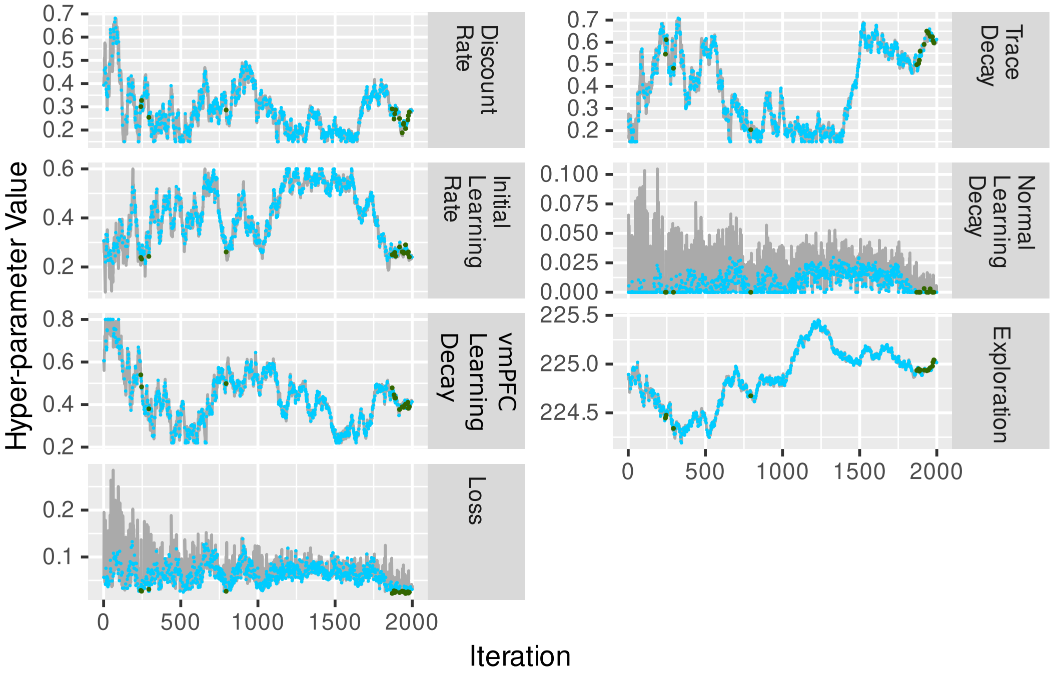
TABLE 9.2
ARA(κ) joint original, re-shuffled and random IGT CSUD search matches.

Discount rate γ and trace decay κ values differ between agents. The trace decay rule in (9.1d) indicates that in calculating trace decay, the product $\gamma\kappa$ is used. Compared to the Boltzmann agent, the ε-Greedy agent has a relatively higher discount rate γ and relatively lower trace decay κ. For example, CSUD search delivers a Boltzmann agent, which with $\gamma = 0.206$ is highly focused on the present. Note however that (9.1d) indicates, eligibility traces use the product $\gamma\kappa$; this product is quite similar at 0.135696 and 0.128956 for the ε-Greedy and Boltzmann agents respectively.

Table 9.2 and Fig. 9.1 indicate that the Boltzmann agent requires more computational effort for achieving convergence, and as in chapter 6, has difficulty in matching all IGT outcome cases. Boltzmann agent CSUD search produces 16 full matches over 2000 iterations with 4 gradient averages, while the corresponding ε-Greedy agent search produces 389 full matches over 1000 iterations with no gradient averaging. However, the Boltzmann agent does produce 951 partial matches consisting of matching the normal original and re-shuffled; and the vmPFC impaired original and re-shuffled IGT environments. To see whether more full matches could be obtained, the Boltzmann agent iteration budget was increased to 4000, but this did not lead to an increase in the number of full matches. Here only the 2000 iteration CSUD search Boltzmann agent results are reported.



(A) ARA(κ) ε-Greedy Agent



(B) ARA(κ) Boltzmann Agent

FIGURE 9.1: ARA(κ) joint original, re-shuffled and random IGT CSUD search matches. Green dots indicate full matches. Blue dots indicate partial matches as noted in Table 9.2. The Boltzmann agent is slower to converge and at 2000 iterations, search convergence might not yet have been achieved. Details in text.

Agent Hyper-parameter	ε-Greedy	Boltzmann
Discount rate γ	0.2, 0.514, 0.9	0.206, 0.5, 0.9
Trace decay κ	0.264, 0.5, 0.9	0.2, 0.626, 0.9
Initial Learning Rate α_1	0.137, 0.33, 0.66, 0.9	0.1, 0.290, 0.66, 0.9
Learning Decay λ[a]	0.080, 0.270	0.00291, 0.402
Exploration	$\varepsilon = 0.1, 0.5, 0.612, 0.75$	$\tau = 5, 75, 225, 425$
IGT length	Q-learning samples	
100	750	

[a] The learning rate decay and exploration decay grids are constructed around the two values above. Appendix C provides the construction method.

TABLE 9.3
ARA(κ) joint original, re-shuffled and random IGT CSUD verification grid search configurations.

Table 9.3 presents the CSUD grid search verification configurations. Note that adding two additional parameters, the discount rate γ and trace decay κ with 3 grid points each leads to a 9-fold increase of the grid search.

For the ε-Greedy and Boltzman agents, Fig. 9.2 and Fig. 9.3 respectively show 2D contours resulting from CSUD verification grid search at CSUD selected minimum loss discount rate γ and trace decay κ values.

In Fig. 9.2 and Fig. 9.3, dark and light gray zones represent normal and vmPFC impaired human IGT outcome match areas respectively. Solid black contours show response to learning decay λ at CSUD selected minimum loss hyper-parameter values. In terms of mean fraction of good decks $\bar{f}_G$, Fig. 9.2 and Fig. 9.3 indicate that for both agents, increasing learning rate decay continues to lead to outcomes, which are consistent with vmPFC impaired behaviour. For both agents, there is a lower learning decay rate λ_N and a higher learning decay rate λ_{vmPFC} associated with normal and vmPFC impaired behaviour respectively. Additionally, to achieve simulation outcomes in corresponding human match zones, it is noted that exploration must be high with $\varepsilon = 0.612$ and $\tau = 225$ for the ARA(κ) ε-Greedy and ARA(κ) Boltzmann agents respectively.

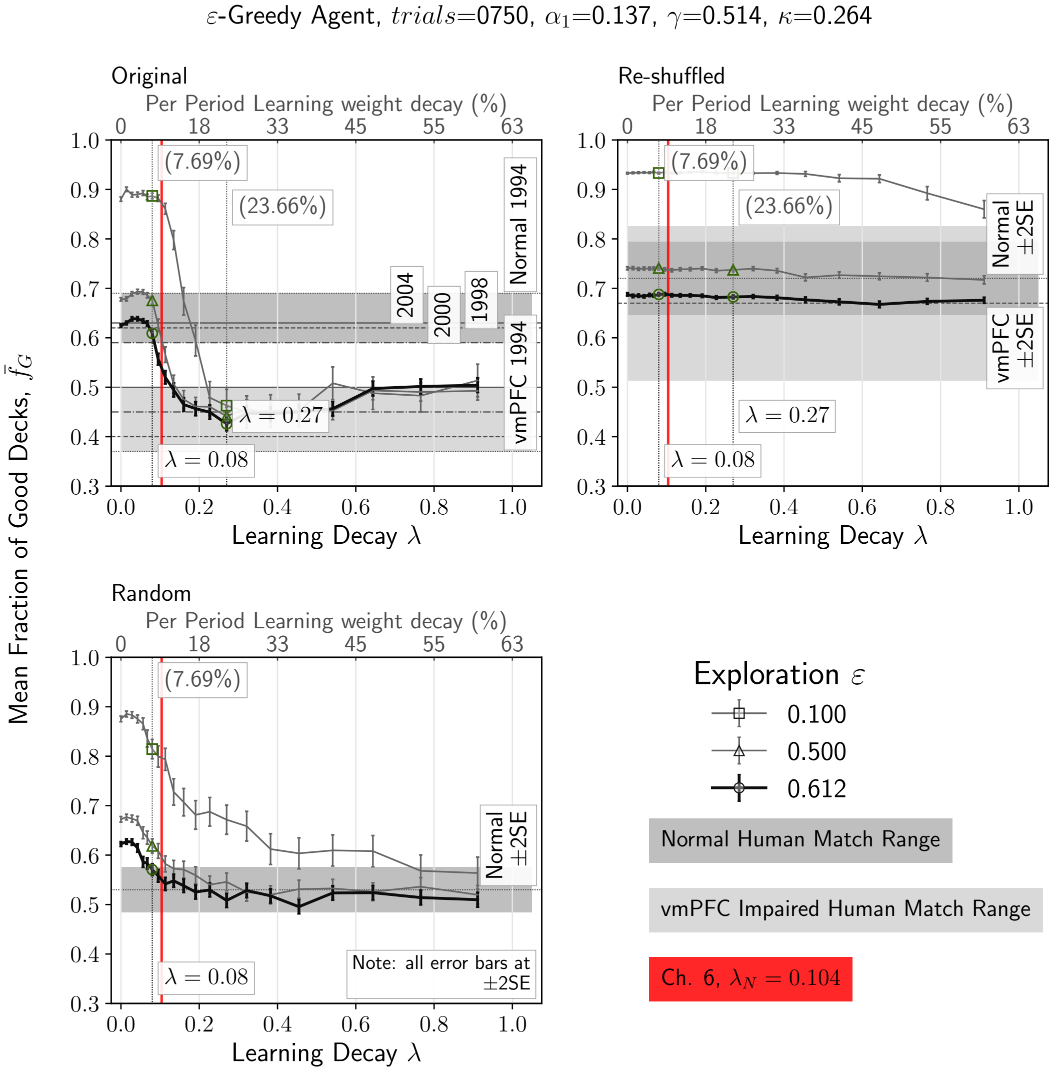


FIGURE 9.2: ε-Greedy Agent ARA(κ) CSUD verification grid search 2D contours. The dark gray zone represents the normal human IGT outcome match area. The light gray zone represents the vmPFC impaired human IGT outcome match area. Solid black contours show response to learning decay λ at CSUD selected minimum loss hyper-parameter values consisting of $\gamma = 0.514$, $\kappa = 0.264$, $\alpha_1 = 0.137$, $\varepsilon = 0.612$, $\lambda_N = 0.08$, $\lambda_{vmPFC} = 0.27$. The solid red line provides an approximate comparison between chapter 6 simple RL and the current ARA(κ) models. It is noted that the ARA(κ) ε-Greedy agent would not achieve a match in the original IGT at the chapter 6 reported normal learning rate decay value of $\lambda_N = 0.104$. The solid red line provides an indication that it does not appear possible to find universal full match CSUD minimum loss hyper-parameter values across different agent architectures. The general results of chapter 6 continue to apply. Just as in Fig. 6.2, increasing learning rate leads to vmPFC impaired behaviour, and human match zones require high exploration. Due to initial learning rate differences, however, a direct overlay comparison of this plot with Fig. 6.2 is not possible.

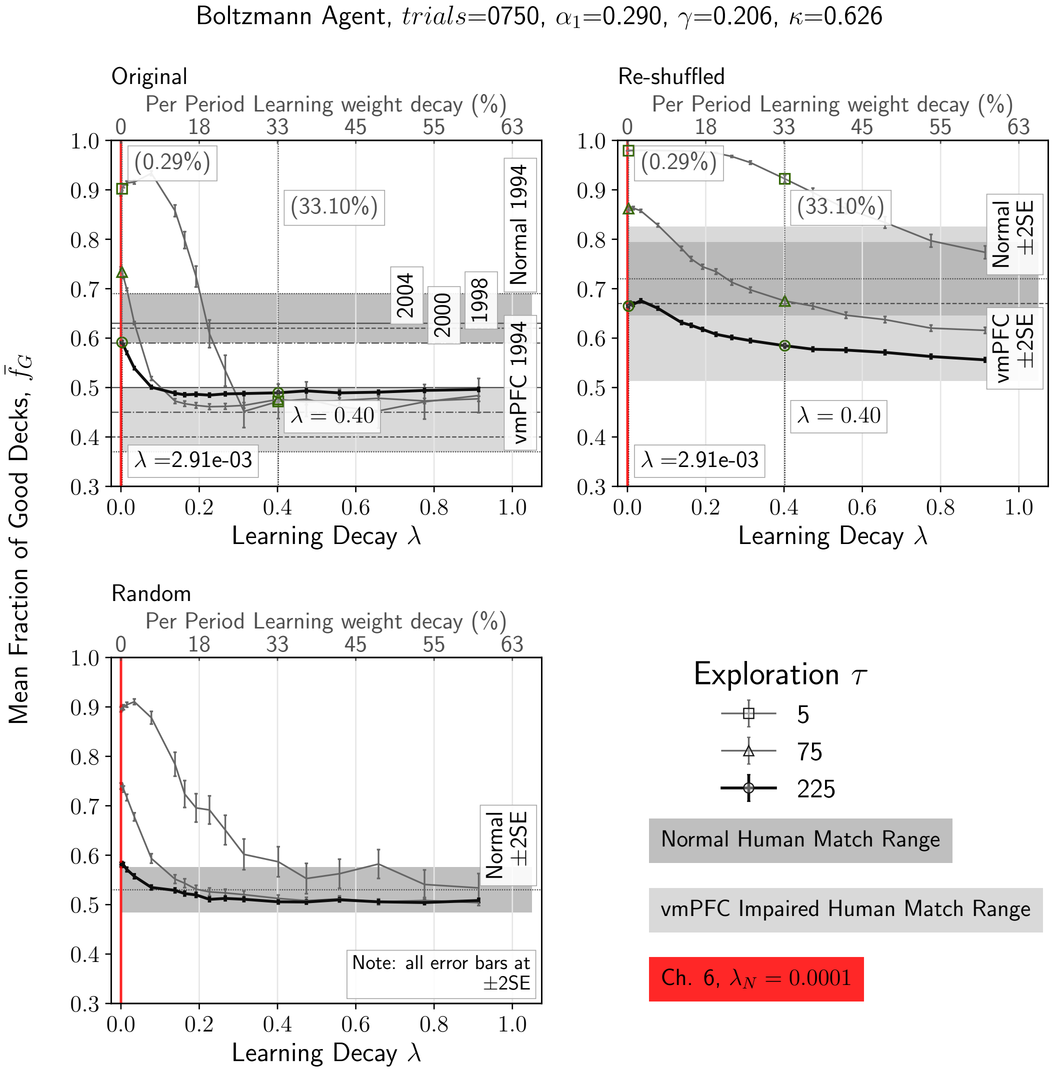


FIGURE 9.3: Boltzmann Agent ARA(κ) CSUD verification grid search 2D contours. The Boltzmann agent does not obtain hyper-parameter matches for the random IGT. The dark gray zone represents the normal human IGT outcome match area. The light gray zone represents the vmPFC impaired human IGT outcome match area. Solid black contours show response to learning decay λ at CSUD selected minimum loss hyper-parameter values consisting of γ = 0.206, κ = 0.626, α_1 = 0.290, τ = 225, λ_N = 2.91e^{-03}, λ_{vmPFC} = 0.40. The solid red line provides an approximate comparison between chapter 6 simple RL and the current ARA(κ) models. It is noted that the ARA(κ) Boltzmann agent would not achieve a match in the random IGT at the chapter 6 reported normal learning rate decay value of λ_N = 0.0001. The solid red line provides an indication that it does not appear possible to find universal full match CSUD minimum loss hyper-parameter values across different agent architectures. The general results of chapter 6 continue to apply. Just as in Fig. 6.8, increasing learning rate leads to vmPFC impaired behaviour, and human match zones require high exploration. Due to initial learning rate differences, however, a direct overlay comparison of this plot with Fig. 6.8 is not possible.

Fig. 9.3 shows that like the simple Boltzmann agent, the ARA(κ) Boltzmann agent has difficulty matching all IGT outcome cases. The CSUD driven search produces a partial match indicated by the annotated circles on the solid black contour lines. However, even the partial matches obtained are close to the match boundaries for the normal original, vmPFC impaired original, and normal re-shuffled cases.

The question arises as to how the CSUD discovered minimum loss hyper-parameters compare across the chapter 6 simple RL model and the current ARA(κ) model. The ARA(κ) model has the additional discount rate γ and trace decay κ hyper-parameters. Also a comparison of Table 9.2 and Table 6.2 reveals that across the two models agents exhibit differing initial learning rates. The ARA(κ) ε-Greedy agent has $\alpha_1 = 0.137$, while the simple ε-Greedy agent has $\alpha_1 = 0.417$. The ARA(κ) Boltzmann agent has $\alpha_1 = 0.290$, while the simple Boltzmann agent has $\alpha_1 = 0.364$. Of greater interest are the normal learning decay λ_N values. In particular, one could ask whether common λ_N values exist in light of the observation that λ_N values appear numerically close on a per agent basis. The ARA(κ) ε-Greedy agent has $\lambda_N = 0.08$, while the simple ε-Greedy agent has $\lambda_N = 0.104$. The ARA(κ) Boltzmann agent has $\lambda_N = 2.91e^{-03}$, while the simple Boltzmann agent has $\lambda_N = 1.0e^{-04}$. The solid red vertical line in Fig. 9.2 and Fig. 9.3 marks the simple agent normal learning rate decay λ_N values for the simple ε-Greedy and Boltzmann agents respectively. It can be seen that at $\lambda_N = 0.104$ the ARA(κ) ε-Greedy agent cannot produce a corresponding simulation match residing in the normal human original IGT outcome catchment zone. Similarly at $\lambda_N = 0.0001$ the ARA(κ) Boltzmann agent cannot produce a corresponding simulation match residing in the normal human random IGT outcome catchment zone. Such "misses" illustrate the difficulties encountered in searching for universal learning rate decay parameter values across different agent architectures. As noted above, however, regardless of model induced hyper-parameter value differences, the general observations remain intact. Learning rate decay increase leads to vmPFC impaired behaviour and matching human IGT outcomes requires high exploration.

Fig. 9.4 and Fig. 9.5 depict ARA(κ) agent 3D contour plots, which at CSUD selected minimum loss discount rate γ and trace decay κ, show the effects of the initial learning rate α_1, learning decay λ and exploration.

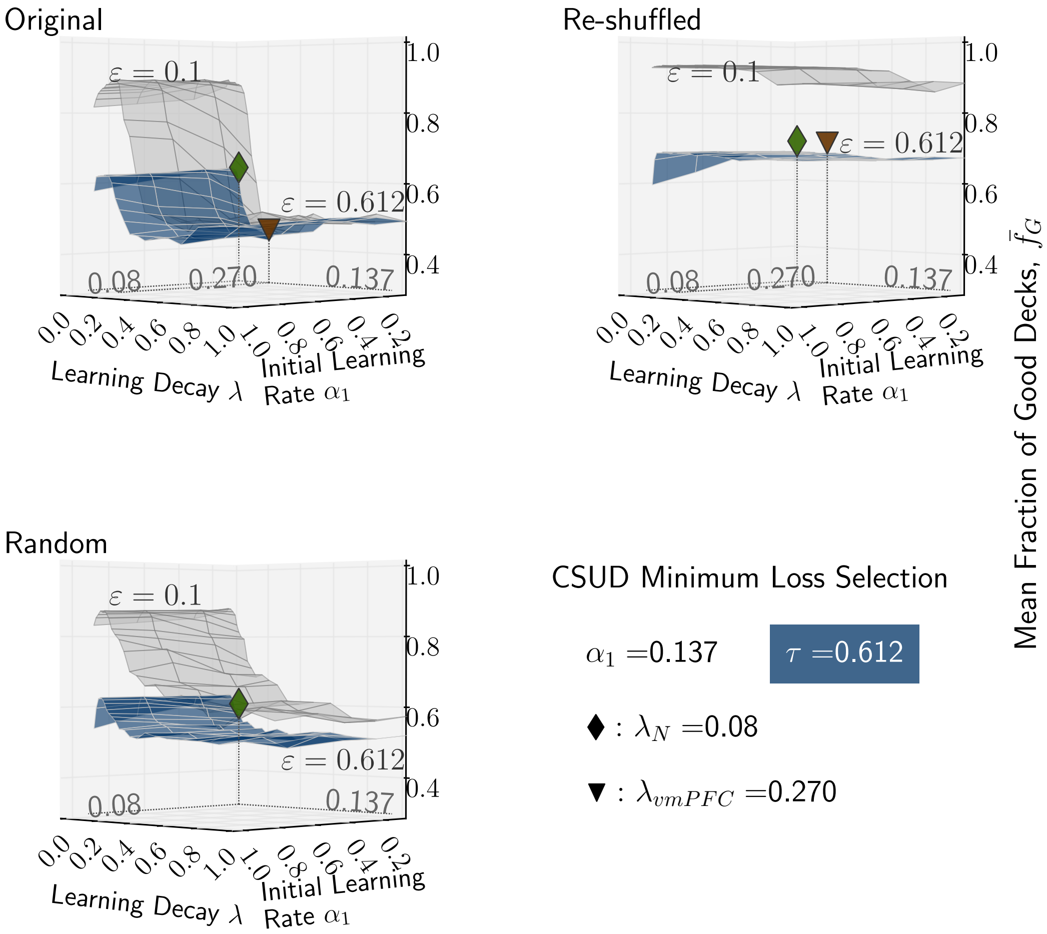


FIGURE 9.4: ARA(κ) ε-Greedy Agent, γ = 0.514, κ = 0.264, CSUD search grid verification 3D contours. Blue coloured surfaces show response to learning decay λ at CSUD selected minimum loss hyper-parameter values. The diamond and inverted triangular shapes mark CSUD minimum loss normal and vmPFC impaired learning decay rates respectively.

Blue coloured surfaces show response to learning decay λ at CSUD selected minimum loss hyper-parameter values. The minimum loss CSUD solutions are annotated on the graphs, with the diamond ♦ and the inverted triangle ▼ highlighting the normal and vmPFC impaired solutions respectively.

For all IGT environments and for the ARA(κ) ε-Greedy and ARA(κ) Boltzmann agents respectively, Fig. 9.4 and Fig. 9.5 show that the initial learning rate α_1 has little effect on the mean fraction of cards chosen from the good decks $\bar{f}_G$. As noted in the simple reinforcement learning agents, an initial learning rate effect α_1 occurs at very low learning decay λ and high initial learning rate α_1, leading to the dome-shaped areas visible in the rear of some

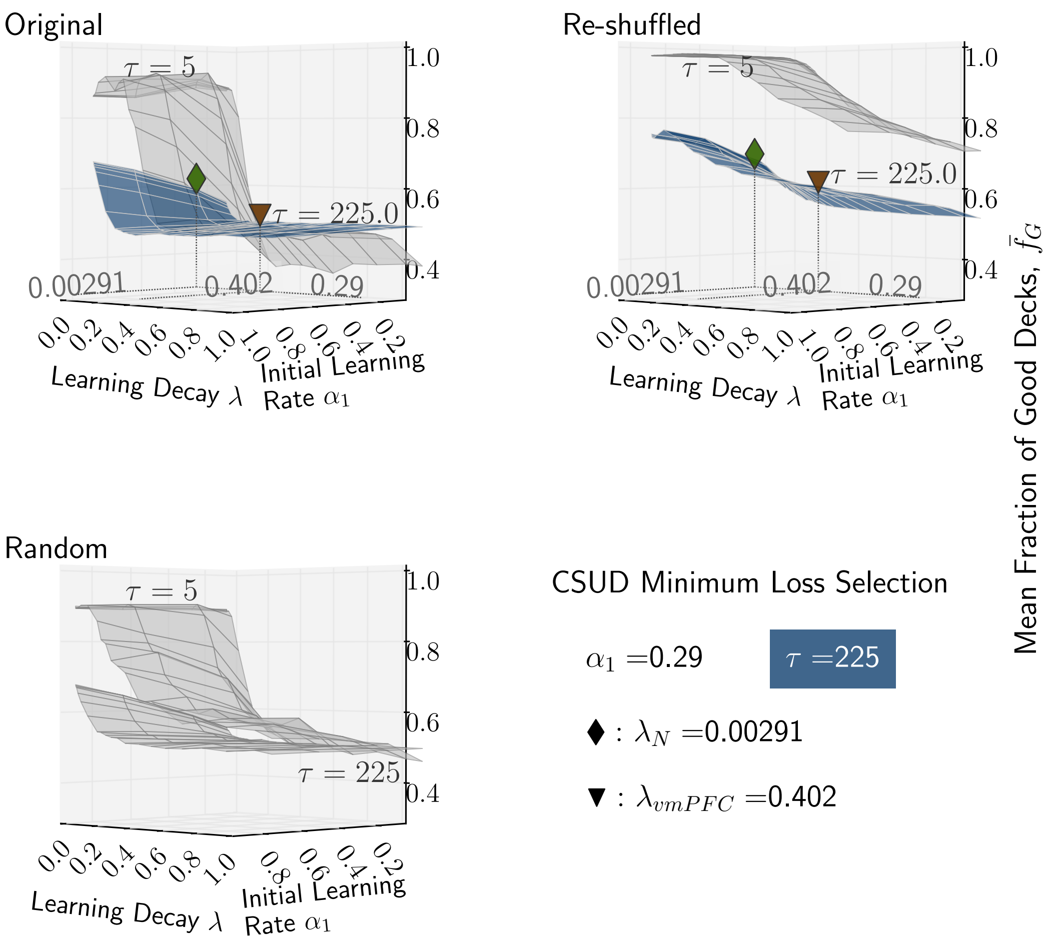


FIGURE 9.5: ARA(κ) Boltzmann agent, $\gamma = 0.206$, $\kappa = 0.626$, CSUD search grid verification 3D contours. Blue coloured surfaces show response to learning decay λ at CSUD selected minimum loss hyper-parameter values. The diamond and inverted triangular shapes mark CSUD minimum loss normal and vmPFC impaired learning decay rates respectively.

IGT environment plots, for example as is notable for the ARA(κ) Boltzmann agent Fig. 9.5 $\tau = 5$ 3D contour, in the area where learning rate decay λ is low and the initial learning rate α_1 is high.

In Fig. 9.4 and Fig. 9.5, increases in exploration lead to downward shifts of the mean fraction of good decks $\bar{f}_G$ surfaces. The blue 3D surfaces indicate that to match human IGT outcomes, ARA(κ) agents must have high exploration.

In general, at CSUD selected minimum loss discount rate γ and trace decay κ, ARA(κ) agents retain similar behavioural mean fraction of good decks $\bar{f}_G$ 2D and 3D response contours, which were initially noted in Fig. 6.2 and Fig. 6.3 for the simple ε-Greedy; and Fig. 6.8 and Fig. 6.9 for the simple

Boltzmann agents respectively.

ARA(κ) 20-draw block agent behaviour retains the same features discussed in section 6.3, Fig. 6.4 and Fig. 6.5; and section 6.4, Fig. 6.10 and Fig. 6.11 for the simple ε-Greedy and Boltzmann agents respectively.

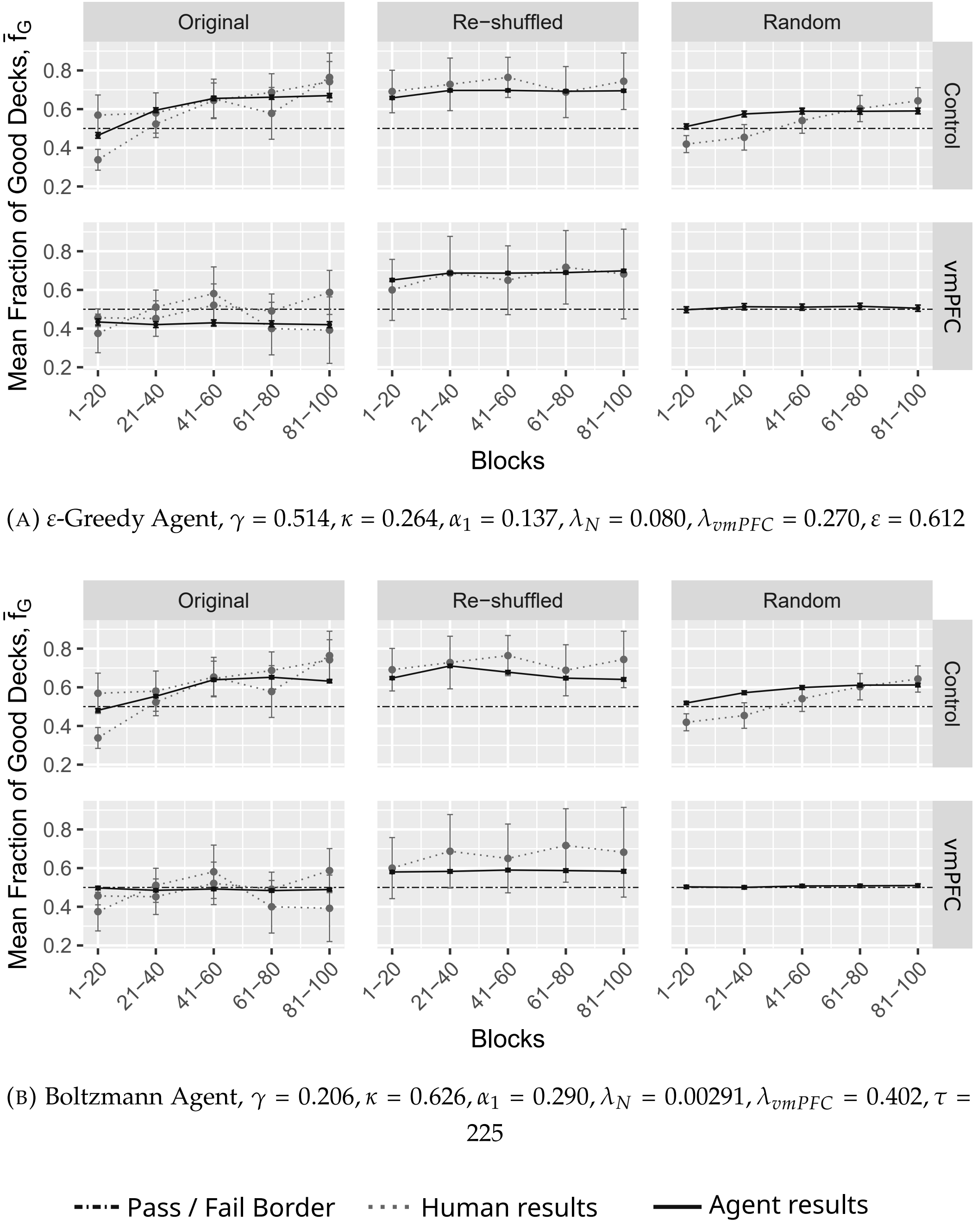
However for completeness, ARA(κ) agent 20-draw block mean fraction of good decks $\bar{f}_G$ and exploration index (EI) results are presented in Fig. 9.6 and Fig. 9.7 respectively.

In terms of mean fraction of good decks $\bar{f}_G$ in Fig. 9.6, when human performance ± 2 standard error (SE) is taken into account, ARA(κ) agent 20-draw block performance is generally within human ranges but shows small deviations for normal behaviour for blocks 1-20 and 21-40 in the random IGT environment where, agent performance is better.

The exploration index (EI) measure in Fig. 9.7b indicates that the ARA(κ) Boltzmann Agent produces in the normal and vmPFC impaired re-shuffled IGT, higher 20-draw block EI values than those exhibited by the human reference data.

In contrast, as Fig. 9.7a shows, the ARA(κ) ε-Greedy agent attains comparatively closer exploration index (EI) values.

Fig. 9.8 and Fig. 9.9 depict success of simulation outcomes, and provide agent jitter plot density summary for the fraction of good decks f_G outcomes obtained at the CSUD selected minimum loss agent hyper-parameter values for 750 simulation samples for normal (control) and vmPFC impaired configurations. Green dots mark outcomes inside human performance ranges. Blue dots mark additional normative pass results, whereas red dots mark additional normative pass fails. Green numbers give total matches out of 750 samples, and the values in brackets indicate percentages matched. A high percentage matched value is indicative of predictive simulation success. The dashed and dash-dotted horizontal lines indicate the maximum and minimum respectively of the respective human match range. Finally the red bars and box indicate central tendency in terms of the mean and ± 2 SEs (standard errors).

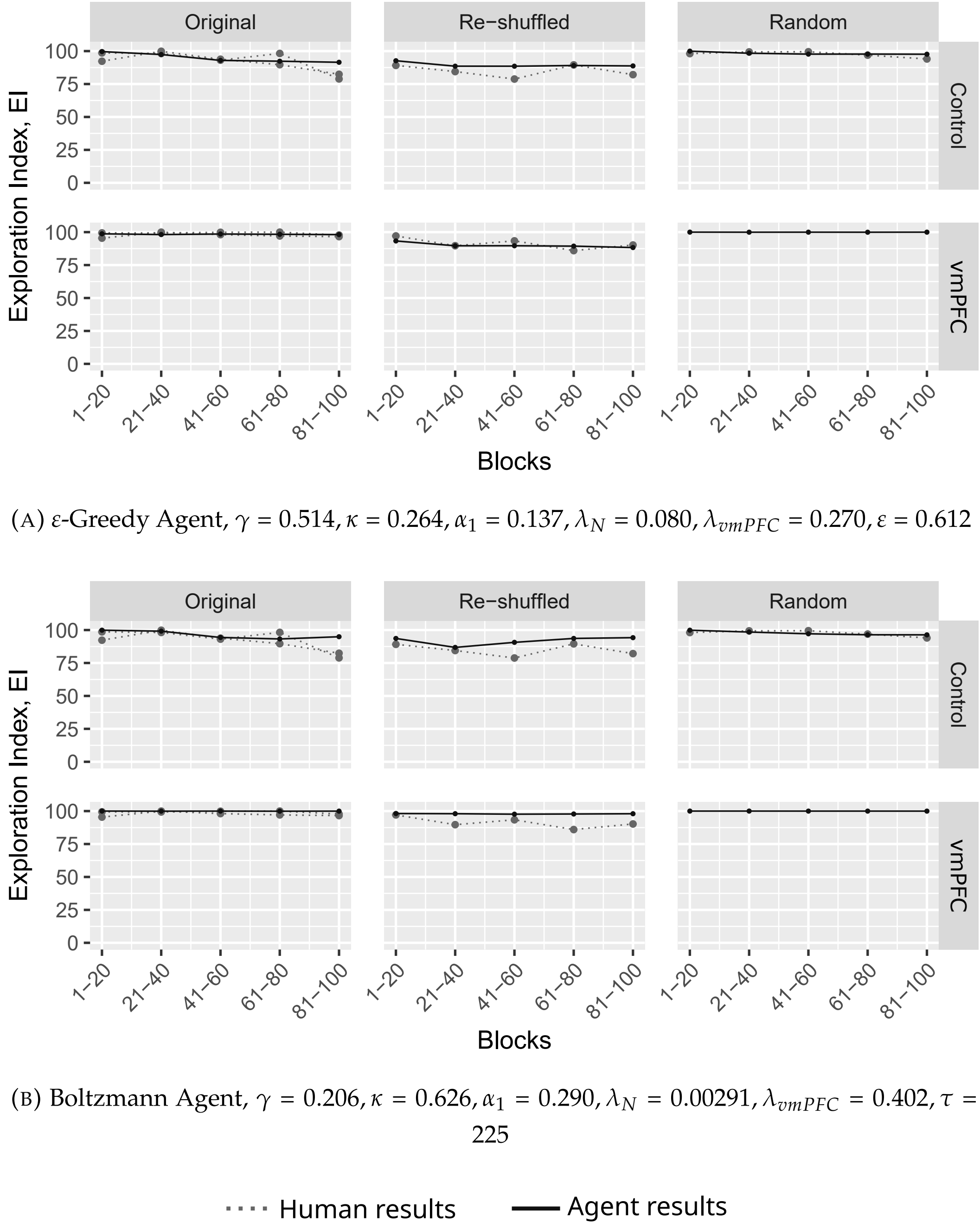


(A) ε-Greedy Agent, $\gamma = 0.514, \kappa = 0.264, \alpha_1 = 0.137, \lambda_N = 0.080, \lambda_{vmPFC} = 0.270, \varepsilon = 0.612$

(B) Boltzmann Agent, $\gamma = 0.206, \kappa = 0.626, \alpha_1 = 0.290, \lambda_N = 0.00291, \lambda_{vmPFC} = 0.402, \tau = 225$

(C) Legend

FIGURE 9.6: ARA(κ) agent 20-draw blocks comparison at CSUD search values as indicated above. Human reference results in dotted light gray. Agent results in solid dark gray, averaged from 750 samples. All error bars at $\pm 2SE$. Human vmPFC outcomes show variation, whereas vmPFC configured agent outcomes appear flat, suggesting the absence of a per 20-draw block learning effect. When $\pm 2SE$ human error bars are taken into account, however, agent 20-draw block performance resides within human $\pm 2SE$ ranges.



(A) ε-Greedy Agent, $\gamma = 0.514, \kappa = 0.264, \alpha_1 = 0.137, \lambda_N = 0.080, \lambda_{vmPFC} = 0.270, \varepsilon = 0.612$

(B) Boltzmann Agent, $\gamma = 0.206, \kappa = 0.626, \alpha_1 = 0.290, \lambda_N = 0.00291, \lambda_{vmPFC} = 0.402, \tau = 225$

(C) Legend

FIGURE 9.7: ARA(κ) agent 20-draw block exploration index (EI) comparison at CSUD search values as indicated above. Human reference results in light gray. Agent results in dark gray, averaged from 750 samples. Human subject and agent exploration index responses appear relatively similar except in the re-shuffled IGT, where Boltzmann agent EI appears higher.

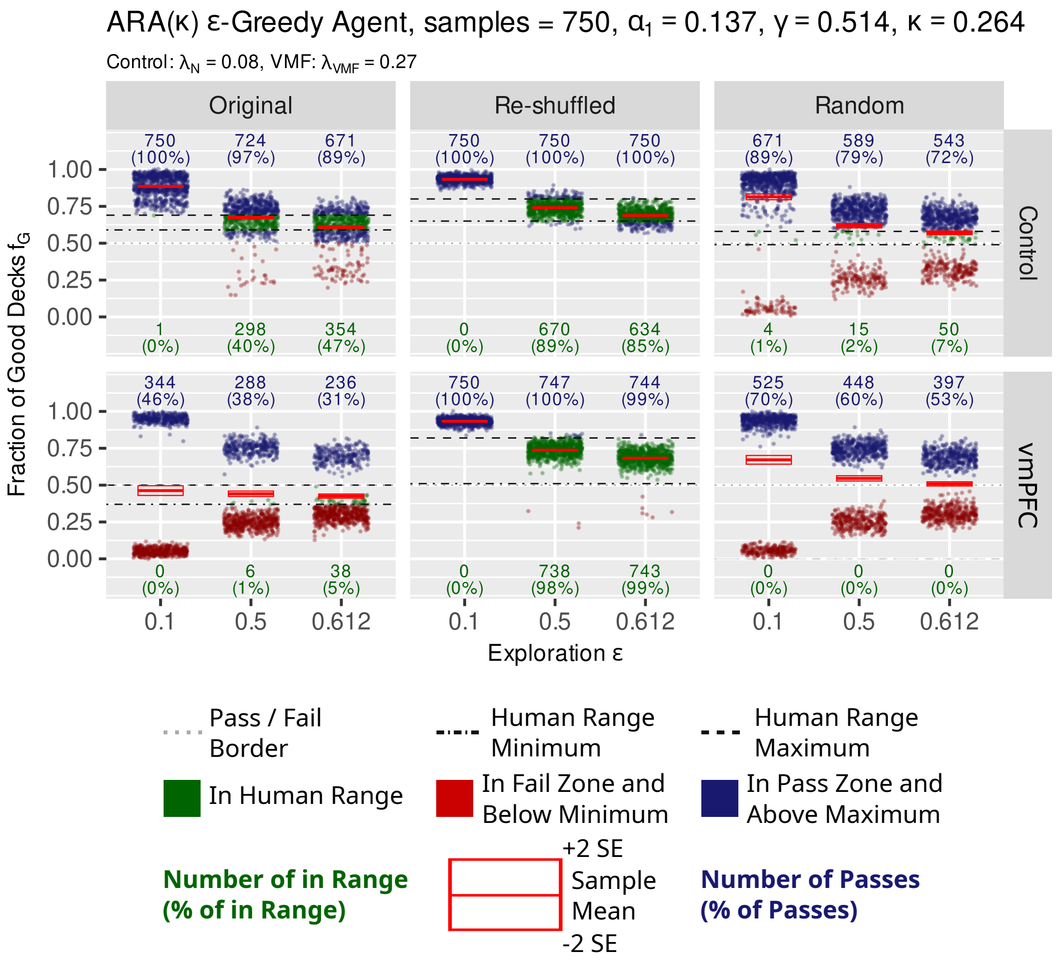


FIGURE 9.8: ε-Greedy ARA(κ) agent, CSUD minimum at $\varepsilon = 0.612$. Comparison of repeated simulation outcomes to human IGT results. At $\varepsilon = 0.612$ and remaining reported CSUD minimum loss hyper-parameter values as noted above, the ARA(κ) ε-Greedy agent generally achieves high matches for the presented IGT cases. Full details are in the text.

Fig. 9.8 shows that the ARA(κ) ε-Greedy agent behaves similar to the simple ε-Greedy agent. Comparison of Fig. 9.8 with Fig. 6.6 reveals that the addition of discount rate γ and trace decay κ has not changed the bi-modal jitter plot densities observed especially for the normal random, vmPFC impaired original and random cases. For the ARA(κ) ε-Greedy and simple ε-Greedy agents with CSUD minimum loss exploration at $\varepsilon = 0.627$ and $\varepsilon = 0.612$ respectively, it is noted that the ARA(κ) ε-Greedy agents exhibits a slight decrease in human match zone values across normal and vmPFC impaired behaviours and all IGT cases. For example, the simple ε-Greedy agent achieves 429, 586, and 58 matches for the normal original, re-shuffled, and random environments respectively. In contrast, the ARA(κ) ε-Greedy

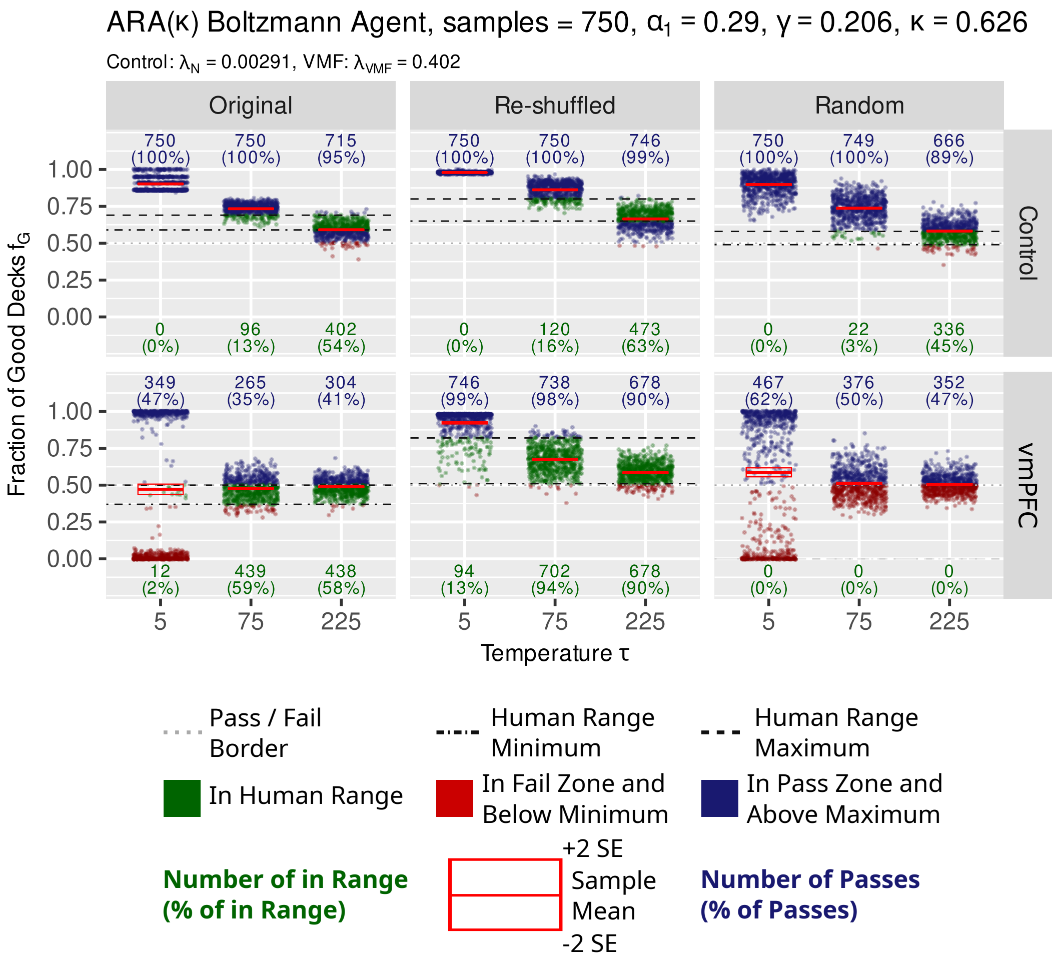


FIGURE 9.9: Boltzmann ARA(κ) agent, CSUD minimum at $\tau = 225$. Comparison of repeated simulation outcomes to human IGT results. At $\tau = 225$ and remaining reported CSUD minimum loss hyper-parameter values as noted above, the ARA(κ) Boltzmann agent generally achieves high matches for the presented IGT cases. Full details are in the text.

agent achieves slightly reduced values with 354, 634, and 50 matches for the normal original, re-shuffled, and random environments respectively. Further, the simple ϵ-Greedy agent achieves 62, and 750 matches for the normal original and re-shuffled environments respectively. In contrast, the ARA(κ) ϵ-Greedy agent achieves slightly reduced values with 38 and 743 matches for the normal original and re-shuffled environments respectively.

In general the ARA(κ) ϵ-Greedy agent achieves the highest simulation matches in the re-shuffled IGT environment with 85% and 99% matches for the normal and vmPFC impaired behaviour configurations respectively. The normal original case achieves 47% matches. However the vmPFC impaired normal and normal random cases only achieve 5% and 7% matches respectively, with these two cases showing strongly bi-modal densities with very little mass in the human catchment zones. Therefore overall the discount rate γ and trace decay κ appear to have a very small, and possibly negligible effect in the RL formulations of the IGT.

Further Fig. 9.9 shows that the ARA(κ) Boltzmann agent behaves similar to the simple Boltzmann agent. The addition of discount rate γ and trace decay κ produces a bi-modal jitter plot for the vmPFC impaired original case, but otherwise does not substantially alter the remaining bi-modal jitter plot densities. In general the ARA(κ) Boltzmann agent achieves the highest simulation matches in the re-shuffled IGT environment with 63% and 90% matches for the normal and vmPFC impaired behaviour configurations. The normal and vmPFC impaired original cases achieve 54% and 58% matches respectively. The normal random case achieves 45% matches. The ARA(κ) Boltzmann agent overall produces more matches inside IGT human outcome catchment zones. However, as the red coloured agent mean $\pm 2SE$ boxes indicate, despite higher in-zone matches, Boltzmann agent means appear in general near catchment zone boundaries.

A comparison of Fig. 9.8 with $\varepsilon = 0.612$ and Fig. 9.9 with $\tau = 225$ shows with respect to the original and random IGT environments that the Boltzmann agents achieves more individual outcome placements inside corresponding human catchment zones than does the ε-Greedy agent. Regarding outcome means however for the normal random IGT, the ε-Greedy simulation $\bar{f}_G$ narrowly lies in the corresponding human catchment zone, whereas the $\bar{f}_G$ of the Boltzmann agent narrowly misses the corresponding human catchment zone. Given that in Fig. 9.1, the Boltzmann agent hyper-parameter search appears unconverged, one might ask if the Boltzmann agent could achieve random IGT normal configuration outcomes inside the corresponding catchment zone with a longer search horizon. This was attempted for the Boltzmann agent with 4000 search iterations. The CSUD search after 4000 iterations, however, could not produce any in-catchment-zone normal random IGT $\bar{f}_G$ outcomes, and on that basis this approach was then abandoned. As noted in section 9.5, it would be possible to get Boltzmann agent normal random IGT outcome $\bar{f}_G$ matches by increasing the size of the corresponding catchment zone. Such an approach, however, is not undertaken in this work.

Table 9.4 and Table 9.5 present np-M/ANOVA results for the ARA(κ) ε-Greedy and Boltzmann agents respectively. Test variants can be thought of

Test Variant	Test Statistic	df1	df2	p-Value	Subset Results
Original ∣ Re-Shuffled ∣ Random vs. Learning Decay λ					
ANOVA Type[a]	92.139	2.979	4463.136	0	At $\alpha = 0.01$, the null hypotheses of learning decay factor equality is rejected. Only equality of the re-shuffled response cannot be rejected.
Re-Shuffled vs. Learning Decay λ					
ANOVA Type	1.07	1.000	1498	0.301	Single response variable, no subsets.
Wilks Lambda	1.07	1.000	1498	0.301	

[a] Wilks Lambda could not be computed due a singular rank matrix.

TABLE 9.4
ARA(κ) ε-Greedy agent np-M/ANOVA analysis of mean fraction of good decks $\bar{f}_G$ at $\gamma = 0.514$, $\kappa = 0.264$, $\alpha_1 = 0.137$, $\varepsilon = 0.612$, with learning decay λ as factor. $\lambda_N = 0.08$, $\lambda_{vmPFC} = 0.27$. At significance level $\alpha = 0.01$, mean fraction of good decks $\bar{f}_G$ responses are statistically significantly different, except for the re-shuffled IGT environment.

as non-parametric versions of the F-test, with a test statistic and two degrees of freedom. These three quantities are then assessed to derive the p-value. The test statistics are discussed in Burchett et al., 2017.

The np-M/ANOVA tests check to see whether learning rate decay λ as a factor leads to a switch from normal to vmPFC impaired behaviour. The learning decay factors are normal learning decay λ_N and vmPFC impaired learning decay λ_{vmPFC}. The null hypothesis is that there is no learning rate decay value induced factor effect. At significance level 0.01, a p-value less than 0.01 leads to the rejection of this null, whereas p-values greater than 0.01 indicate that the null hypothesis of no factor effects cannot be rejected. Based on human IGT mean fraction of good deck $\bar{f}_G$ results, the expectation is that for the original decks, high learning rate decay leads to vmPFC impaired behaviour in the original but *not* in the re-shuffled IGT environment. While there is no human outcome data for the vmPFC impaired random IGT case, a test is also included, based on simulation results, for the random IGT environment. Human data does not exist to validate the random IGT results; however, based on simulation results, expected human outcomes can be hypothesised for the random IGT environment.

At the CSUD discovered minimum loss hyper-parameter values as listed

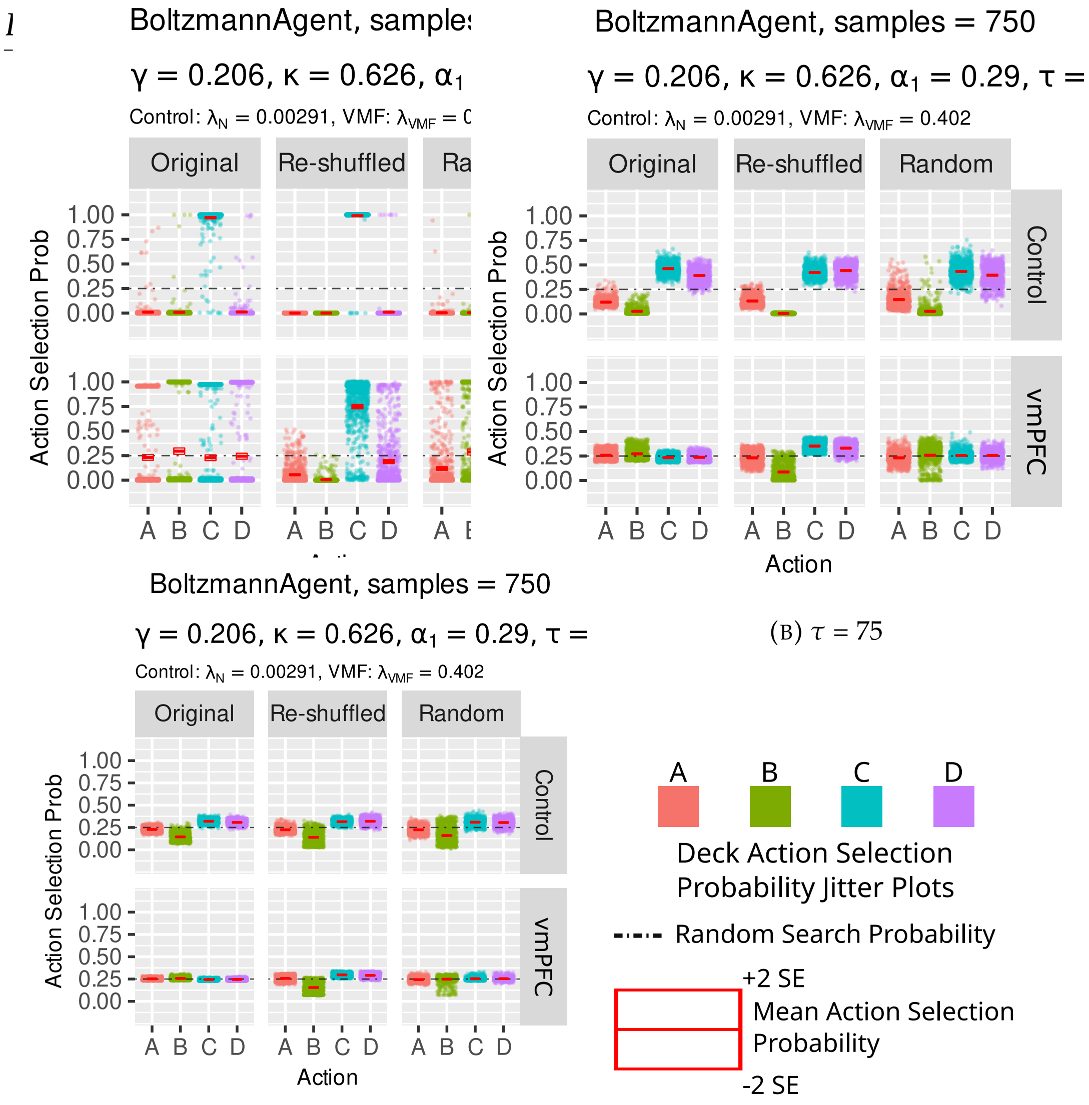
Test Variant	Test Statistic	df1	df2	p-Value	Subset Results
Original | Re-Shuffled | Random vs. Learning Decay λ					At α = 0.01, the null hypotheses of learning decay factor equality is rejected for all response variable subsets
ANOVA Type[a]	918.78	2.915	4366.213	0	

[a] Wilks Lambda could not be computed due a singular rank matrix.

TABLE 9.5
ARA(κ) Boltzmann agent np-M/ANOVA analysis of mean fraction of good decks $\bar{f}_G$ at γ = 0.206, κ = 0.626, α_1 = 0.290, τ = 225, with learning decay λ as factor. λ_N = 0.00291, λ_{vmPFC} = 0.402. At significance level α = 0.01, mean fraction of good decks $\bar{f}_G$ responses are statistically significantly different.

in the respective tables, Table 9.4 and Table 9.5 show that at significance level α = 0.01, normal versus vmPFC impaired learning rate decay (λ_N versus λ_{vmPFC}) produces a statistically significant joint difference in mean fraction of good decks $\bar{f}_G$. However, as Table 9.4 indicates, only for the ARA(κ) ε-Greedy agent, does one fail to reject the null hypothesis of no factor effect for the re-shuffled IGT environment. That is, the ARA(κ) ε-Greedy agent reproduces statistically expected human re-shuffled IGT outcomes, whereas the ARA(κ) Boltzmann agent does not. Both agent results indicate that if vmPFC impaired subjects are given the random IGT, then there should be statistically significantly different mean fraction of good decks results between the normal and vmPFC impaired subjects.

Fig. 9.10 displays ARA(κ) Boltzmann agent action (deck) selection probabilities at CSUD discovered minimum loss hyper-parameter values for exploration (temperature) τ = 5, 75, and 225. At IGT completion and with τ = 225, the CSUD selected minimum loss exploration value, with normal behaviour configuration λ_N = 0.00291, the agent chooses from the good decks C and D at probabilities above 0.25, while bad deck choice probabilities are under 0.25. Bad deck B has the lowest selection probability in all three IGT environments. However, in the vmPFC impaired behaviour configuration with λ_{vmPFC} = 0.402, in the original IGT, deck B selection probability is above 0.25. At τ = 5, the agent is too greedy, and for normal behaviour, focuses primarily on deck C. At τ = 75, the agent for normal behaviour achieves selection probabilities with the good decks C and D being clearly favoured.



(B) $\tau = 75$

(C) $\tau = 225$

(D) Legend

FIGURE 9.10: ARA(κ) Boltzmann agent at CSUD selected minimum loss with $\gamma = 0.206$, $\kappa = 0.626$, $\alpha_1 = 0.290$. Exploration temperature τ and action selection probabilities at IGT task completion. Simulation sample size $n = 750$. At $\tau = 225$, mean action selection probabilities for the good decks C and D are above 0.25. Note that per deck individual human results with $n = 70$ are only available for the random IGT, however, have not been included in this plot, as corresponding human data for the original and re-shuffled IGTs are not available.

However, as the 2D contours in Fig. 9.3 indicate, at $\tau = 75$, the agent is already achieving better than human performance across all behaviour and IGT environment cases, in particular for the normal random IGT case.

9.5 CSUD Search Discussion

From an infinite horizon perspective, the original, re-shuffled, and random IGT environments appear identical and all exhibit the same mean net yields shown in Table 4.1. However, as the results in chapter 6 and here indicate, the three IGT environments exhibit sequencing effects, which lead to variations in the human mean fraction of good decks $\bar{f}_G^H$ achieved, especially with respect to normal and vmPFC impaired behaviour.

The simple and ARA(κ), ε-Greedy and Boltzmann agent reinforcement learning models discussed and implemented in chapter 6 and here, differ by two additional parameters, the discount rate γ and trace decay κ. The addition of the discount rate and trace decay does not alter previous findings regarding mean fraction of good decks $\bar{f}_G$ outcomes. In particular, increasing learning rate decay λ continues to lead to vmPFC impaired behaviour and matching human results continues to require high exploration.

At the low learning rate decay λ_N, normal configured agents pass the original, re-shuffled, and random IGTs. However, at the higher learning rate decay λ_{vmPFC}, vmPFC impaired configured agents fail the original, but pass the re-shuffled IGT. Software agents qualitatively achieve this result at multiple $\bar{f}_G$ outcomes, including those, which exceed normal human performance. To match human performance agents must have high exploration.

Further for given discount γ and trace decay κ values, ARA(κ) agent mean fraction of good decks $\bar{f}_G$ 2D and 3D contours as well as jitter plots present with similar visual characteristics as those observed in the corresponding simple RL action value agent plots in chapter 6. This indicates that γ and κ do not appear to have much interaction with λ and exploration at least for the original, re-shuffled, and random IGT environments.

In this work, exponential learning rate decay is used to effectively induce a finite learning horizon. A finite learning horizon constitutes a plausible explanation for the sequencing effect driven mean fraction of good decks $\bar{f}_G$ outcome differences observed in original, re-shuffled and random IGT environments.

Considering the fraction of good decks f_G jitter plot density results in Fig. 9.8 and Fig. 9.9, in conjunction with the np-M/ANOVA results in Tables 9.4 and 9.5 for the ARA(κ) ε-Greedy and Boltzmann agents respectively, it is noted that neither software agent provides an entirely convincing explanation of human decision making. However, it is believed that both agents do represent learning decay λ effects well, while neither agent fully captures the nature of human exploration. Further, agent original and re-shuffled IGT, vmPFC impaired configuration, 20-draw block results are flatter than corresponding human results, indicating that vmPFC impaired humans exhibit some learning, which the agents do not appear to capture.

At CSUD discovered exploration $\varepsilon = 0.612$, the ARA(κ) ε-Greedy agent displays bi-modal f_G jitter plot densities for the vmPFC impaired original and normal random cases. These bi-modal densities have very little probabilistic mass inside the indicated human match zones. However, as the respective red bars indicate, the respective mean fraction of good decks $\bar{f}_G$ outcomes do fall within the human catchment zones. It is possible that the combination of learning rate decay λ and constant exploration ε effects leads to the development of bi-modal fraction of good decks f_G densities. Whether agent choices converge towards the good or bad decks, constant exploration still brings in enough choices from the complement decks, and this leads to bi-modal densities.

In terms of the np-M/ANOVA assessment, which uses ranks to determine whether behaviour and IGT environment cases exhibit significantly different means, the ARA(κ) ε-Greedy agent does achieve conformance with expected human IGT outcomes.

These results offer good insight into the difficulty of solely assessing based on central tendency measures. Unfortunately, assessing potential bi-modal human response patterns requires a large sample of individual human data, which is not available. Fellows and Farah (2005, p. 60, Fig. 4) provide a breakdown for vmPFC impaired subjects for the original and re-shuffled IGT environments. vmPFC impairment is a rare condition however, and their participant sample size is 9; this makes it difficult to develop comparative density plots.

Fig. 9.11 depicts jitter density plot for fraction of good decks f_G^H achieved by normal human subjects taking the random IGT with individual participant data available from Steingroever et al. (2015). Green dots mark outcomes inside human performance ranges. Blue dots mark additional normative pass results, whereas red dots mark additional normative pass fails. Green numbers give total matches out of 70 samples, and the values in

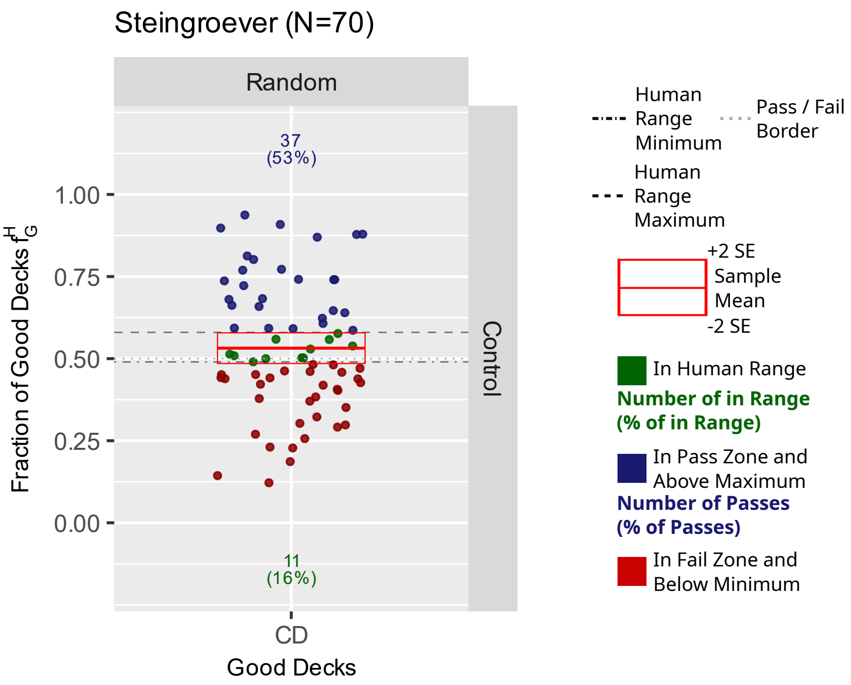


FIGURE 9.11: Random IGT fraction of good deck f_G^H outcomes for control subjects. Data from Steingroever et al. (2015). Only 11% of participants achieve outcomes placed in the $\pm 2SE$ catchment area denoted by the dashed top and dash-dotted bottom lines. Details in text.

brackets indicate percentage matched. The dashed and dash-dotted horizontal lines indicate the maximum and minimum respectively of the human match range. Finally the red bars and box indicate central tendency in terms of the mean and ±2 SEs (standard errors).

In Fig. 9.11, note that only 11% of participants placed inside the $\pm 2SE$ catchment area. Therefore, it is possible that catchment areas of $\pm 2SE$ are too narrow. However, here this possibility is not investigated further.

At 70 samples, the jitter plot in Fig. 9.11 itself does not indicate a tendency towards bi-modality, and in shape is similar to the corresponding Fig. 9.9 $\tau = 225$ ARA(κ) Boltzmann agent jitter plot, however with wider dispersal. The wider f_G^H dispersal suggests that at 70 participants, the resulting f_G^H distribution is unlikely to have converged to a normal distribution. Hence, the use of np-M/ANOVA, which does not rely on normality for statistical significance testing, appears to be a correct choice.

At $\tau = 225$, the ARA(κ) Boltzmann agent does not exhibit bi-modality, and if catchment areas are widened, it could match all available human behaviour, IGT environment $\bar{f}_G$ outcome ranges. However, in np-M/ANOVA, the Boltzmann agent fails to reject the hypothesis of no learning decay factor effect for the mean fraction of good decks $\bar{f}_G$ in the re-shuffled IGT environment. Widening the catchment areas would not address this divergence from expected human behaviour, where it is expected that the re-shuffled IGT normal and vmPFC impaired mean fraction of good deck $\bar{f}_G$ outcomes do not produce a statistically significant difference.

In sum, neither of the two ARA(κ) agents considered here achieves a decisive match on the basis of non-aggregated, individual performance and aggregated np-M/ANOVA analysis. It is believed that this is because neither model adequately expresses human exploration; whereas ε-Greedy exploration is too loose, Boltzmann exploration is too precise. The underlying mathematical representation of human exploration behaviour may lie somewhere between these two exploration alternatives.

Chapter 10

Reinforcement Learning: Iowa Gambling Task with Burst Learning

This chapter develops a simple model for burst learning. The term *burst learning* is used to describe an iterative learning scenario, where the learning rate may suddenly increase, leading to increased contributions from current or future projected outcomes. Therefore an increased learning rate introduces the capability to overwrite a previously learned response.

A variable learning rate is not a new concept. In rational iterative learning as well as in stochastic search, a variable learning rate, for example, can be implemented as an inverse Hessian approximation (Zhu et al., 2020), as a deterministic rule, or in response to knowledge accumulation (Powell, 2011, pp. 419-452).

Here, a heuristic approach is taken to learning rate variation. The exponential learning rate decay model (5.2) with $Q_0(a) = 0$, introduced in chapter 5, is capable of generating human IGT outcomes. At lower learning decay λ_N, the agents match normal human IGT outcomes, and at higher learning decay λ_{vmPFC}, the agents match vmPFC impaired IGT outcomes.

As discussed in section 2.3, the vmPFC, or orbitofrontal cortex (OFC), is implicated in emotion mediated outcome valuation. Here, it is proposed that the default learning rate is always decaying, but that via emotion mediation, a decayed learning rate may be *reset* to a higher learning rate. In doing so, a theoretical model is presented that provides an alternative to infinite horizon, continuous learning rational models, such as the ones discussed in section 2.5.

In the model proposed here, the learning rate may decay so quickly as to freeze learning within a few iterations. However, outcomes may elicit emotions, which reset the learning rate, thus allowing the agent to re-learn, but once again, for a limited number of iterations. The interplay between emotion signals, which reset the learning rate, and high default learning rate decay creates sequences of short learning episodes. This phenomenon is called *burst learning*.

The term *burst learning* been employed in a psychological context to refer to directed episodes of "focused learning" (Kunitani, 2016). The term "bursting" is also used in neuroscience, where it refers to episodes of multiple neuron spikes, and has been employed in the modelling of forward-pass neural networks to generate learning benefits (Ohta et al., 2022).

In the current context, the notion of bursting is applied via emotion signals to modulate a decaying learning rate. When the decaying learning rate is reset, learning benefits are obtained. The use of an emotion signal to regulate a decaying learning rate constitutes as far as is known a novel approach. It should be noted, however, that the model proposed here can be developed without reliance on emotion labels. It is believed, however, that using emotion labels may be justified by the role of emotion in decision making as discussed in section 2.3.1 and section 2.3.2.

As in the previous chapter, first the decision making model is presented, and then the reinforcement learning agent implementations are introduced. This is followed by IGT applications, and then a discussion.

10.1 Single State Exponential Learning Rate Decay Q-learning with Burst Learning

The burst learning heuristic has three key components: emotions, default learning rate behaviour, and the intervention logic, which resets the learning rate. Each key component is introduced in turn.

10.1.1 Emotions

Abstracting from Rolls (2013), a simple two emotion system is defined. In this system, within target results produce happiness, while out-of target results produce anger. The experience of happiness results in default learning rate decay behaviour. The software agent only alters its default behaviour

Outcome	Emotion	Behaviour
On-target	Happiness	Default behaviour
Off-target	Anger	Intervention

TABLE 10.1
Two emotion learning and response, abstracted from Rolls (2013)

when it experiences anger. That is, anger leads to resetting the learning rate. Table 10.1 summarizes two emotion learning, and the resulting agent behaviour.

10.1.2 Learning Rate Decay

Let $\alpha_1 \in [0,1]$ be the initial learning rate, where the notation [] indicates a closed interval. A class of bounded exponential decay learning weights are defined, such that the upper bound is the initial learning rate α_1, and the lower bound is an attenuated fraction of the initial learning rate, α_1/D with $D > 1$. The term time-to-bound, TTB, is used to indicate the time required to go from α_1 to α_1/D.

Under exponential decay, for any decay factor $\lambda \in (0, \infty^+)$, the time-to-bound TTB required to reach α_1/D can be computed as

$$TTB = \frac{lnD}{\lambda} \tag{10.1}$$

Note that, the initial learning rate α_1 does not affect the time-to-bound, which is solely expressed in terms of the attenuation factor D and the decay factor (learning rate decay) λ.

For example, $D = 2$ would correspond to the half-life of the initial learning rate. If $\alpha_1 = 0.5$ and $\lambda = 0.25$, then it would take $ln2/0.5 = 2.8$ periods for the learning rate to decay from 0.5 to 0.25. If $D = 100$ and $\lambda = 0.5$, then it would take $ln100/0.5 = 9.2$ periods for the initial learning rate to reduce by 100-fold. For example, it would take 9.2 periods for an initial learning rate of 0.5 to decay to 0.005. Time is represented in discrete iterations, and therefore, TTB is always rounded to the *nearest* integer. This rounding operation is denoted as $Round(TTB)$.

Given the initial learning rate α_1, learning rate decay λ, and attenuation D, define an indexed learning rate sequence $\{\alpha_{t(i)}\}_{i\in\mathbb{Z}}$ where

$$\alpha_{t(i)} = \begin{cases} \alpha_1, & \text{if } i \leq 0 \\ e^{-i\lambda}\alpha_1, & \text{if } 0 < i < Round(TTB) \\ \alpha_1/D, & \text{otherwise} \end{cases} \tag{10.2}$$

and $\mathbb{Z}$ denotes the set of integers.

Note that $\alpha_{t(i)}$ as defined in (10.2) above does not fulfil the theoretical convergence criteria described in (3.7) that would be required for convergence guarantees for the general iterative update rule (3.1b). Intuitively this is easy to see, as (10.2) consists of the concatenation of exponentially decaying and constant learning rates, neither of which satisfy convergence criteria (3.7) on their own.

It would be straightforward to modify the lower α_1/D bound in (10.2) so that $\alpha_{t(i)}$ satisfies convergence criteria. For example for $i \geq Round(TTB)$, one can set

$$\alpha_{t(i)} = 1/t - 1/floor(TTB) + \alpha_1/D. \tag{10.3}$$

Provided there exists a time period t after which the learning rate is no longer *reset*, then such a sequence can be shown to satisfy convergence criteria (3.7). This proof is not shown here, however, intuitively it can be seen that such a proof is driven by the $1/t$ term in (10.3). The remaining terms in (10.3) are constant, finite and bounded, and on that basis do not affect convergence dynamics as $t \to \infty$.

Such bounded convergent learning rates as in (10.3) have been employed by the author in experimental studies, but are not reported in this work, as the results do not differ much from the simpler decay mechanics reported here in (10.2). Since the IGT only lasts for 100 periods, and is administered on a one-shot learning basis, the use of bounded convergent learning rates did not much change below findings, which use the bounded but 'non-convergent' learning rate schedule presented in (10.2).

10.1.3 Behaviour Intervention

Default behaviour consists of incrementing the learning rate index i by 1 per learning iteration. This default behaviour is denoted by $i{+}{+}$. As the middle branch of (10.2) shows, under default agent behaviour, when learning rate decay λ and attenuation D are high, the incremented learning rate can decay very quickly and effectively lead to truncation of learning. This default behaviour may be modified on the basis of elicited emotions.

The behaviour cycle consists of four stages: assess, act, report, and prepare ($AARP$). Emotions are 'broadcast' into a global buffer accessible to any behavioural stage. The assess and act stages approximately correspond to the critic and actor respectively in an actor-critic reinforcement learning framework. However, in addition to value function updates, the assessment stage here also populates the global emotion buffer. The agent's internal state remains private, but in the reporting stage, the agent has the option to disclose the emotion buffer. The preparation stage is where the agent may engage in additional set-up such as default preparatory behaviours.

The proposed architecture is general and permits a rich set of interactions. However, in the present study, to investigate emotion and learning rate interaction, behaviour is simplified as follows: the emotion buffer only contains a single emotion, and is cleared at the beginning of each decision making cycle. Emotions simply consist of labels, and hence, can be seen as just being on or off.

10.1.4 Two Emotion Single State On-Policy Q-learning

Chapter 9 shows that the addition of discount rate γ and trace decay κ does not lead to changes in initial learning rate α_1, learning decay λ, and exploration responses. Therefore in this chapter, rather than ARA(κ) agents, the simple reinforcement learning agents introduced in chapter 5 are employed.

The use of single state on-policy (action-value) Q-learning (5.1) is retained as suggested by Sutton and Barto (2018, p. 32). With discount rate $\gamma = 1$ and indexed learning rate $\alpha_{t(i)}$, model (5.1) becomes

$$Q_t(a) = \alpha_{t(i)} x_t^a + \left(1 - \alpha_{t(i)}\right) Q_{t-1}(a) \tag{10.4}$$

where a denotes action, $\alpha_{t(i)}$ is defined in (10.2), and $i \in \mathbb{Z}$ is an index. x_t^a indicates the net yield for action a.

Note that (10.4) is general enough to accommodate different indexed learning rates for each card deck, or could even be generalised to have separate gain and loss learning rates. Such differences could for example be achieved by simply letting the index vary across decks, gains, and losses. The notation above, however, is geared towards the simplest case where the indexed learning rate remains the same across all decks, and instead of gains and losses, only net yield is considered. The approach here produces a simpler model and facilitates an architectural comparison with previously presented models.

Next define the one-step temporal difference error as

$$\delta_t \equiv (x_t^a - Q_{t-1}(a)) \tag{10.5}$$

where $Q_{t-1}(a)$ constitutes the agent's best forecast of the value of action a at time t.

Given (10.4) and (10.5), define the one-step Q-difference error as

$$\Delta Q_t(a) \equiv Q_t(a) - Q_{t-1}(a) = \alpha_{t(i)}\delta_t \tag{10.6}$$

where Δ indicates the difference operator.

All-or-nothing emotions are elicited via threshold activation criteria defined as the ratio of the current Q-difference error to the last-achieved Q-value, which also forms the best one period ahead forecast. The emotion activation threshold is defined in terms of a fraction involving the current learning rate $\alpha_{t(i)}$. Using Equation (10.6), specify

$$\frac{\Delta Q_t(a)}{Q_{t-1}(a)} = \frac{\alpha_{t(i)}\delta_t}{Q_{t-1}(a)} \substack{<\\=\\>} \frac{\alpha_{t(i)}}{B} \tag{10.7}$$

where $\substack{<\\=\\>}$ denotes a three way switch consisting of "less than, equal to, or greater than," and $B > 1$ is a scaling term applied to the current learning rate $\alpha_{t(i)}$. Intuition for B is presented after the introduction of the next equation.

For computational convenience, the threshold boundary condition in Equation (10.7) is further simplified as,

$$\frac{\delta_t}{Q_{t-1}(a) + \xi} \substack{<\\=\\>} 1/B \tag{10.8}$$

where $B > 1$, and $\xi > 0$ is a small computational guard to deal with the case when the denominator $Q_{t-1}(a)$ is 0. Looking at (10.8), one can think of $1/B$ as defining the fraction threshold for the activation paths of the three way switch. For example, if $B = 2$, then $1/B = 0.5$, and (10.8) can be interpreted

Variant	Emotions	Learning Rate $\alpha_{t(i)}$ Index Behaviour
Tempered	$input >= 1/B \rightarrow Happy$	i++, Default behaviour
	$< \ 1/B \rightarrow Angry$	i = 0, re-learn, i++
Stoic	$input >= -1/B \rightarrow Happy$	i++, Default Behaviour
	$< \ -1/B \rightarrow Angry$	i = 0, re-learn, i++
Buffered	$input > 1/B \rightarrow Happy$	i++, Default Behaviour
	$< -1/B \rightarrow Angry$	i = 0, re-learn, i++
	$otherwise \rightarrow NOP$	i = i

TABLE 10.2
Agent variants, activation thresholds, associated emotions and learning rate index i behaviour for input δ_t/Q_{t-1}. Details in text.

as, "if the temporal difference error is less than, equal to, or more than half the action value." Regarding the value of the computational guard against division by 0, in this chapter $\xi = 1e^{-8}$ is employed.

Using equation (10.8), three agent temperaments are defined: tempered, stoic, and buffered. For each temperament, Table 10.2 presents associated activation thresholds, emotions, and learning rate index behaviour. As Table 10.2 shows, δ_t/Q_{t-1} is the assessment criterion based on which emotions are emitted. The activation thresholds are $1/B$, $-1/B$, and the complement of the interval $[-1/B, 1/B]$ for the tempered, stoic, and buffered agents respectively.

The tempered agent is only happy when the temporal difference error as a fraction of $Q_{t-1}(a)$ is equal to or above $1/B$. For example, when $1/B = 0.50$, then the agent is only happy when the temporal difference error gain is at or above 50% of the most recently experienced Q-value. Therefore the tempered agent is only happy when positive gains are achieved. Setting $1/B = 0.50$ produces a high threshold for the three way switch, which controls the behavioural pathways noted in Table 10.2. One can think of this high threshold as inducing behavioural inertia so that agents are only "motivated" to switch when there is a relatively high (unanticipated) temporal difference error.

With $-1/B = -0.50$, the stoic agent is happy when the temporal difference error loss is at or less than 50% of most recently experienced Q-value. In short, the stoic agent can tolerate some disappointment and still remain happy. Finally, the buffered agent is happy or angry depending on whether the temporal difference error gain or loss is above or below a certain percentage, for example 50%, of the last experienced Q-value. The buffered agent does not react when the input δ_t/Q_{n-1} is in the closed interval $[-1/B, 1/B]$. Computationally speaking, this non-reactance is denoted with *NOP* (no operation), a term borrowed from assembly mnemonics. In this context, *NOP* is not interpreted as a separate emotion, but describes a state where sub-activation threshold emotion is present.

In all three agent temperaments, *happy* does not produce an intervention and implicitly leads to default agent behaviour. The learning rate index increments by 1, and this leads to a decay of the learning rate in the next period, at a speed set by learning rate decay λ. In contrast, when *angry*, the agent resets its learning rate back to the initial learning rate α_1. The agent then re-computes, in the current period, the value function using the initial learning rate. Next, as per the default behaviour, the learning rate index is incremented, and this leads to a decrease of the learning rate for the next period. For the buffered agent, in the *NOP* case, the learning rate remains unchanged.

10.1.5 vmPFC Impairment

vmPFC impairment is modelled as the inability to assess emotions, with this inability leading to perpetual continuation of default learning rate decay behaviour.

That is, when a vmPFC impaired configured agent, for example the stoic agent, experiences the *angry* emotion, the agent is no longer able to reset the learning rate back to the initial learning rate α_1. Instead, the learning rate continues to decrease at decay rate λ towards the lower learning rate bound α_1/D. Once this lower bound is attained, the learning rate remains at the lower bound.

These dynamics imply that emotion responses are necessary for normal behaviour, and if emotion responses are not present then vmPFC impaired behaviour results.

10.2 Software agent implementations

Only the ε-Greedy and Boltzmann agents, introduced in sections 5.2.2 and 5.2.1 respectively, are considered. The ε-Greedy agent has constant exploration ε, and the Boltzmann agent defines exploration temperature τ, which controls how Q-values contribute to individual action probabilities.

The adaptive and decaying ε-Greedy agents are no longer considered. The rationale for these omissions is discussed in section 8.3 and section 9.2. To briefly review, the adaptive ε-Greedy agent does not do well replicating human 20-draw block and normal versus vmPFC impaired human M/ANOVA mean fraction of good deck results. The decaying ε-Greedy agent on the other hand produces IGT outcome results close to the simpler constant exploration ε-Greedy agent, leading to the retention of the simpler variant.

Standard ε-Greedy and Boltzmann agent behaviours introduced in sections 5.2.2 and 5.2.1 are augmented by the stoic agent emotion response patterns presented in Table 10.2. As discussed in section 10.1, the *happy* emotion response leads to learning rate decay towards the lower learning rate bound α_1/D. The *angry* emotion response, however, leads to a reset to the initial learning rate α_1, and the re-assessment of the relevant Q-value at the higher learning rate, after which the learning rate is once more decremented for the next iteration. When the agent emits the *angry* emotion, the agent has a chance to learn from the missed expected target, assessed in relation to a percentage threshold in terms of the ratio of the temporal difference error to the last relevant Q-value.

Methodology, results, and the discussion are presented next. Only simulations using **stoic** emotion agents are presented with the emotion and learning rate dynamics shown in Table 10.2. Tempered and buffered emotion agent variants have been tested, however, these agents presented similar results. While tempered and buffered emotion agent variants have different behavioural pathways, it is possible that the high learning rate decay specifications lead to results similar to those obtained by the stoic emotion variant. In the interest of brevity, here only stoic emotion agent results are reported.

To review the agent learning problem, given original, re-shuffled, and random IGT environments, the stoic emotion agents choose cards from one

Agent	Boltzmann	ε-Greedy
Hyper-parameter		
Attenuation D	2 - 1000	2 - 1000
Emotion activation threshold $1/B$	0.5	0.5
Initial learning rate α_1	0.01 - 0.99	0.01 - 0.45
Learning rate decay λ	0.08 - 0.75	0.1 - 0.8
Exploration	$\tau = 0.5 - 500$	$\varepsilon = 0.05 - 0.90$
CSUD Iterations	4000	4000
Gradient Samples	2	2
IGT length	Q-learning samples	
100	750	

TABLE 10.3
2EmST agent search methodology. Joint original, re-shuffled, random IGT hyper-parameter CSUD search criteria.

of the four card decks, A, B, C, and D; and are expected to discover good decks C and D, and choose accordingly.

As in chapter 9, it is believed that the learning rate decay λ parameter is sensitive to the underlying IGT environment yield frequency distribution characteristics; on that basis, searches in this chapter are limited to IGT environment combinations with similar frequency dynamics, with the related original, re-shuffled, and random IGT environments providing the largest common yield characteristics search set.

The prefix *2EmST* is used to differentiate the two-emotion stoic ε-Greedy and Boltzmann agents from the ARA(κ) agents discussed in chapter 9 and the simple RL agents discussed in chapter 6.

10.3 Methodology: Joint Search of the Original, Re-shuffled, and Random IGT environments

The general methodology remains as discussed in section 6.1. The hyper-parameters CSUD search investigates are attenuation D, emotion activation threshold $1/B$, the initial learning rate α_1, learning decay λ, and exploration. Table 10.3 summarises software agent CSUD search parameter constraints and attributes. Note that the emotion activation threshold is constrained at

$1/B = 0.5$. For each agent, CSUD is run for 4000 iterations and two gradient samples are averaged in each iteration.

A joint CSUD (hyper-parameter) search is conducted of the original, re-shuffled, and random IGT environments, looking for a hyper-parameter configuration, which may lead to mean fraction of good decks $\bar{f}_G$ outcomes residing within respective human catchment ranges for normal and vmPFC impaired behaviour.

As indicated in section 10.1.5, normal versus vmPFC impaired behaviour is induced by disabling the emotion response described in Table 10.2. Therefore CSUD search in effect traverses two separate Q-learning models for each agent, the emotion enabled model for normal behaviour and the emotion disabled model for vmPFC impaired behaviour. CSUD search loss is minimised for hyper-parameter configurations which produce a match in five outcome cases: normal original, normal re-shuffled, normal random, vmPFC impaired original, and vmPFC impaired re-shuffled. Searches in this chapter are limited to IGT environment combinations with similar frequency dynamics, with the related original, re-shuffled, and random IGT environments providing the largest common yield characteristics search set.

In sum, this chapter uses CSUD search to assess a common hyper-parameter set controlling the behaviour of two related models, which differ by the inclusion, or exclusion, of learning rate reset dynamics. The CSUD results are then verified by running 750 repeated software agent simulations in a small grid search around the CSUD discovered minimum loss hyper-parameter values. Grid search verification results are analysed and software agent performance is discussed.

10.4 CSUD Search Results

For ease of comparison, results for the two emotion stoic threshold (2EmST) ε-Greedy and Boltzmann agents are presented side-by-side. Table 10.4 and Fig. 10.1 present CSUD search results in tabular and graphic forms respectively.

Table 10.4 shows that after 4000 search iterations, the 2EmST ε-Greedy agent achieves 725 full matches. In contrast, the 2EmST Boltzmann agent only achieves 5 full matches, while achieving 2230 partial matches.

Agent Hyper-parameter	ε-Greedy	Boltzmann
Attenuation D	100.372	100.012
Emotion Activation Threshold $1/B$	0.5	0.5
Initial Learning Rate α_1	0.150	0.420
Learning Decay λ	0.176	0.102
Exploration	ε = 0.665	τ = 225.002
Minimum Loss	0.00767	0.0233
Matches	725	5 / 2230[a]

[a] 5 full CSUD matches. 2230 partial matches when excluding normal random IGT behaviour. Partial matches include normal and vmPFC impaired behaviour, original and re-shuffled decks.

TABLE 10.4
2EmST agent joint original, re-shuffled and random IGT CSUD search matches.

A full search match consists of achieving mean fraction of good decks $\bar{f}_G$ outcomes residing in the respective five human outcome catchment zones: normal original, normal re-shuffled, normal random, vmPFC impaired original, and vmPFC impaired re-shuffled. In a partial match, such as that exhibited by the 2EmST Boltzmann agent, the agent fails to achieve matches for the normal random IGT outcome case.

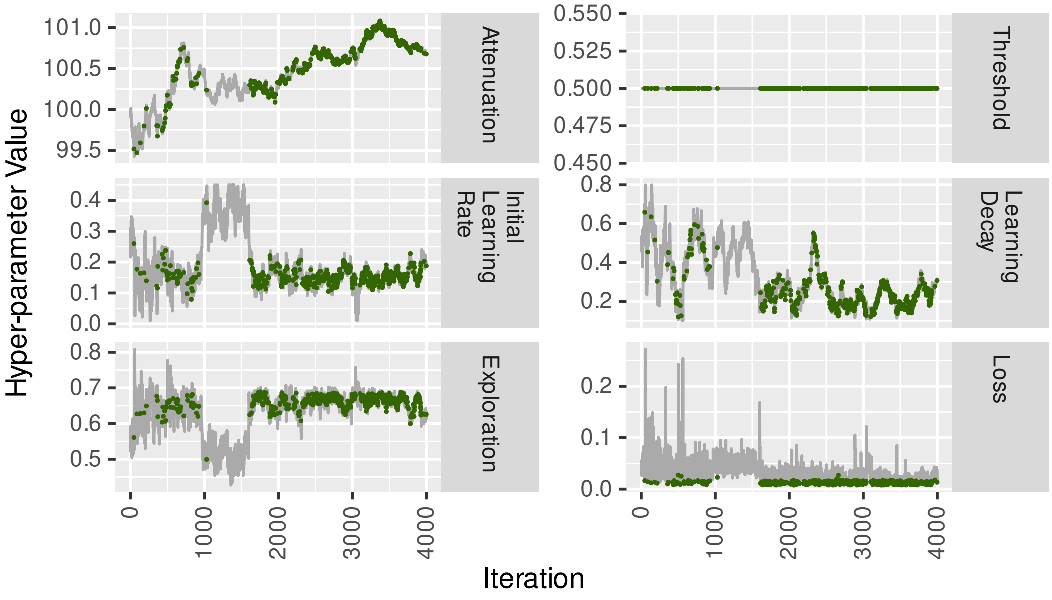
Exploration with $\varepsilon = 0.665$ and $\tau = 225.002$ remains high for both the 2EmST ε-Greedy and Boltzmann agents respectively. The initial learning rate α_1 for the 2EmST ε-Greedy agent is lower than the corresponding value in the simple reinforcement learning model. With 2EmST, the ε-Greedy agent CSUD search produces $\alpha_1 = 0.150$, while the corresponding CSUD search in the simple model in chapter 6 produces $\alpha_1 = 0.417$. This difference possibly originates because the added behavioural complexities of the emotion agent do not do as well with initial non-stationarities, occurring at the beginning of the IGT; consequently the CSUD search algorithm is forced to consider lower initial learning rate values.

Fig. 10.1a and Fig. 10.1b show the outcome of 4000 iterations of CSUD loss minimising hyper-parameter searches for the 2EmST ε-Greedy and Boltzmann agents respectively. Green dots indicate full matches, where as blue dots represent partial matches. Fig. 10.1a shows that the ε-Greedy agent exhibits, with the exception of attenuation D, strong traversal of hyper-parameter space. Further, the clustering of the green dots indicates that there is indeed a specific parameter range which satisfy the search criteria towards which the search is converging.

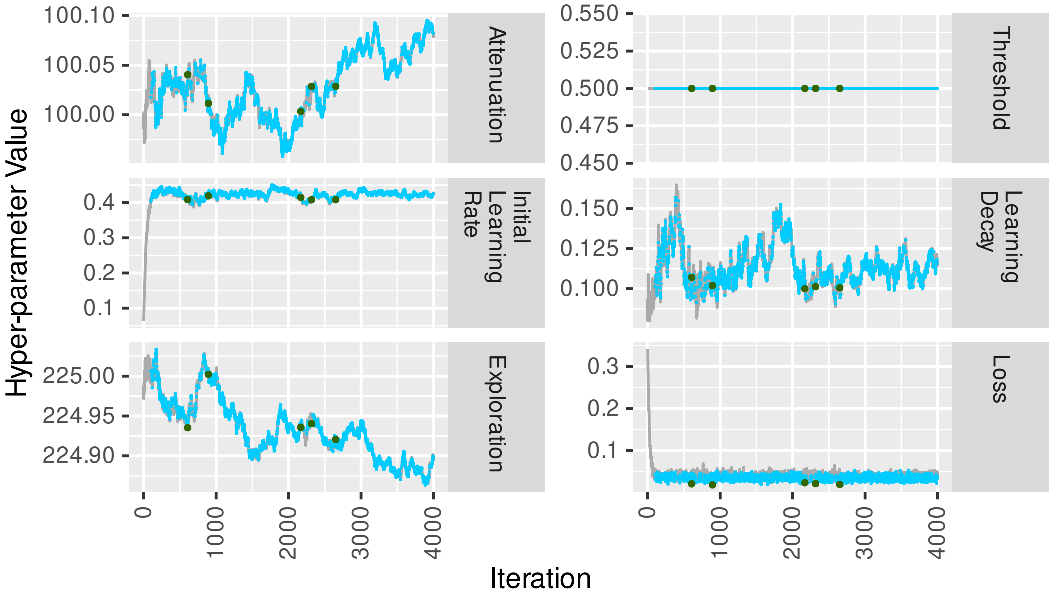
In contrast, Fig. 10.1b shows that the corresponding CSUD search for the Boltzmann agent exhibits only good hyper-parameter range traversal for the initial learning rate α_1. Attenuation D and exploration τ show little range traversal, while learning decay λ shows some range traversal.

As indicated by the sparsity of green dots within the iteration budget, the Boltzmann agent does not show a tendency towards discovery of a full search match. However, the Boltzmann agent does discover a large number of partial matches consisting of in-zone mean fraction of good deck $\bar{f}_G$ matches for the normal and vmPFC impaired, original and re-shuffled IGT environments. It is possible that with different initial hyper-parameter values, larger iteration budget, or mapping of hyper-parameter ranges, the Boltzmann agent would achieve better results. However, the results exhibited here are consisted with the difficulties Boltzmann agents in previous chapters have exhibited, as indicated by Fig 9.1b and Fig 6.7 for the ARA(κ) and simple Boltzmann agents respectively. It is believed that the difficulties seen here are an inherent by-product of the Boltzmann agent exploration specification.

Also Boltzmann exploration in Fig 9.1b does not change much, and it might be asked whether the Boltzmann agent could have achieved better match results had there been wider hyper-parameter space traversal for exploration τ. As mentioned previously, the low traversal result for τ is due to a perturbation scale effect in CSUD. Fig. 10.3 2D grid search results indicate that it is unlikely that a value of τ exists at which the (2EmST) Boltzmann agent would achieve $\bar{f}_G$ matches across all catchment zones. On that basis, the τ perturbation scale and resulting low range traversal issue is not addressed any further.



(A) 2EmST ε-Greedy Agent, 725 full matches



(B) 2EmST Boltzmann Agent, 5 full, 2230 partial matches

FIGURE 10.1: 2EmST agent joint original, re-shuffled and random IGT CSUD search matches. Green dots indicate full matches. Blue dots indicate partial matches as noted in Table 10.4. The Boltzmann agent is slower to converge and at 4000 iterations, search convergence might not yet have been achieved. The Boltzmann agent CSUD search minimum loss match extracted from the few full matches did not produce full matches in subsequent CSUD verification simulations.

Agent	ε-Greedy	Boltzmann
Hyper-parameter		
Attenuation D^{a}	100.372	100.012
Emotion Activation Threshold $1/B$	0.5	0.5
Initial Learning Rate α_1	0.05, 0.150, 0.35, 0.65	0.05, 0.15, 0.420, 0.65
Learning Decay λ^{b}	0.176, 0.5	0.102, 0.5
Exploration	$\varepsilon = 0.10, 0.5, 0.665, 0.75$	$\tau = 5, 75, 225, 475$
IGT length	Q-learning samples	
100	750	

[a] Attenuation grid elements are: 2, 4, 6, 11, 20, 35, 62, a, 197, 349, 620, 1100, where a is replaced as noted above.

[b] The learning rate decay and exploration decay grids are constructed around the two values above. Appendix C provides the construction method.

TABLE 10.5
2EmST agent joint original, re-shuffled and random IGT CSUD search grid verification configurations.

CSUD grid search verification results are presented next. Based on CSUD grid search verification, the burst learning model implemented with two emotion stoic activation threshold (2EmST) agents, displays from a decision theoretic point of view, some significant results, which are especially noticeable with the 2EmST ε-Greedy agent. To summarise, the burst learning model results produce very good effect decomposition among the three key decision making hyper-parameters, the initial learning rate α_1, learning rate decay λ, and exploration ε or τ. The results are now discussed in some detail.

Table 10.5 presents the 2EmST agent CSUD verification grid search configurations. In CSUD searches the emotion activation threshold is constrained to $1/B = 0.5$. Also to conform with agent specifications discussed in section 10.2, emotion use ability is not altered during grid searches.

Fig. 10.2 and Fig. 10.3 show respectively the 2EmST ε-Greedy and Boltzmann agent CSUD verification 2D mean fraction of good decks $\bar{f}_G$ contours obtained at IGT completion. For the 2EmST ε-Greedy agent the CSUD verification grid search 2D contours are displayed at CSUD selected minimum loss attenuation $D = 100.372$ and initial learning rate $\alpha_1 = 0.150$. For the 2EmST Boltzmann agent the CSUD verification grid search 2D contours are displayed at CSUD selected minimum loss attenuation $D = 100.012$ and initial learning rate $\alpha_1 = 0.420$.

In Fig. 10.2 and Fig. 10.3, dark and light gray zones represent normal and vmPFC impaired human IGT outcome match areas respectively. Solid and dash-dotted contours show learning decay λ response of normal (emotion on) and vmPFC impaired (emotion off) configured agents respectively at CSUD selected minimum loss hyper-parameter values. The black contours represent the CSUD minimum loss exploration contour at $\varepsilon = 0.665$. The gray coloured contours present mean fraction of good decks $\bar{f}_G$ behaviour at alternative grid exploration values. Each learning rate decay λ grid point is reported with $\pm 2SE$ error bars.

Most significantly, in terms of mean fraction of good decks $\bar{f}_G$, Fig. 10.2 and Fig. 10.3 indicate that for both agents normal behaviour, 2D (solid line) exploration contours in all IGT environments flatten and the influence of learning rate decay λ is strongly reduced. Previously in chapters 6 and 9, such an effect was only noticeable in the re-shuffled IGT environment. This flattening effect is more noticeable in Fig. 10.2 with the ε-Greedy agent. This result implies that in burst learning and with normal behaviour, the original, re-shuffled, and random IGT environment card sequencing effects no longer play an outcome determining role.

In the case of vmPFC impaired behaviour however, with both agents, the 2D dotted line exploration contours retain the sinusoid like shapes previously noted in chapters 6 and 9. Note that the sinusoid shapes are most prominent in the vmPFC impaired original case, while being squashed to a line in Fig. 10.2 for the 2EmST ε-Greedy agent vmPFC impaired re-shuffled IGT environment case. This result implies that in burst learning and with vmPFC impaired behaviour, card sequencing, in the the original, re-shuffled, and random IGT environments, does lead to sequencing effects, modulated by learning rate decay λ and the IGT environment itself.

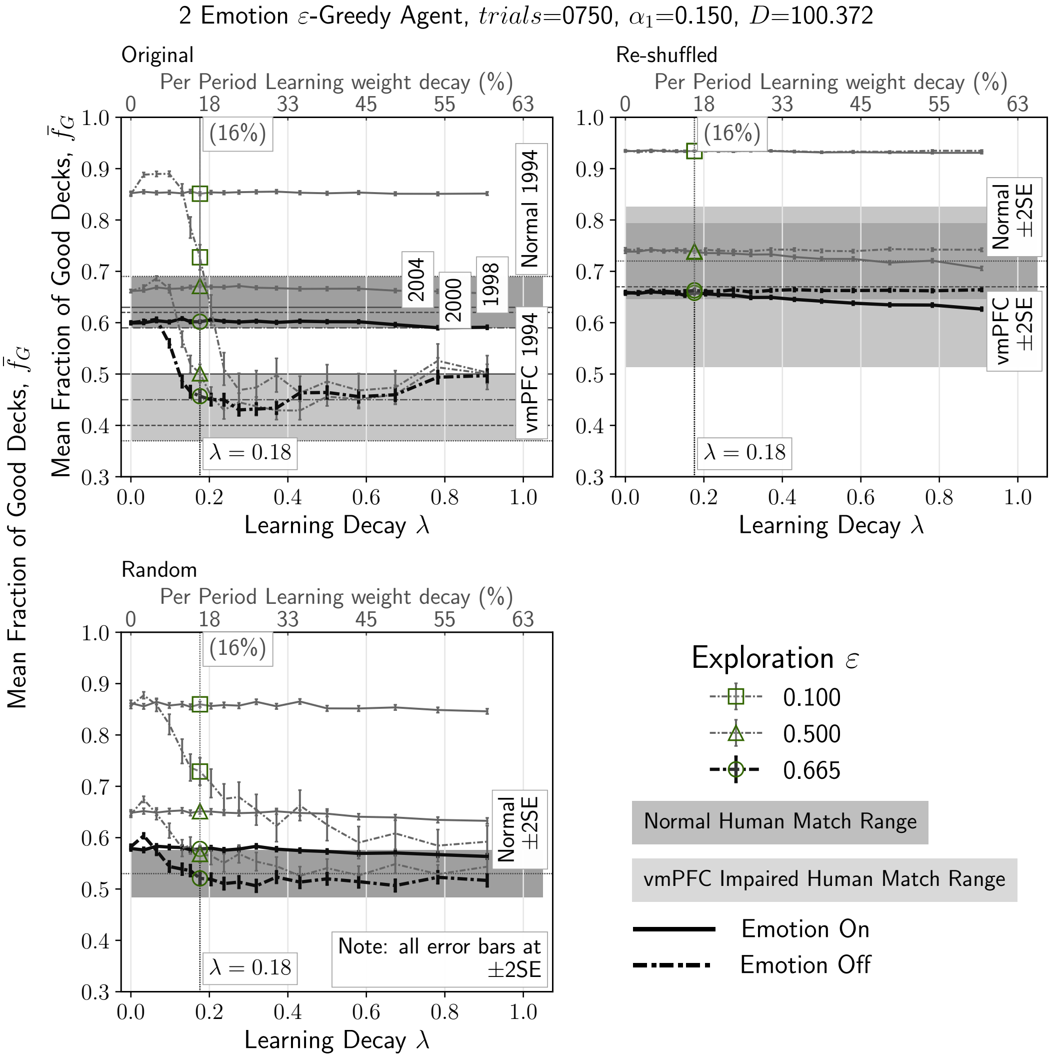


FIGURE 10.2: 2EmST ε-Greedy agent CSUD verification grid search 2D contours. Normal and vmPFC impaired behaviours are depicted by solid and dotted lines respectively. For normal behaviour, burst learning produces significant decoupling of learning decay and exploration responses. The dark gray zone represents the normal human IGT outcome match area. The light gray zone represents the vmPFC impaired human IGT outcome match area. Solid black contours show response to learning decay λ at CSUD selected minimum loss hyper-parameter values. Details in text.

Regarding exploration, for either agent and behaviour configuration, increasing exploration leads to a downward shift of the corresponding 2D contours. However, as full exploration is equivalent to random search, which would yield an expected result of $\bar{f}_G = 0.5$, there may be a lower limit to this downward shift; indeed the limiting point of increasing exploration appears to be $\bar{f}_G = 0.5$.

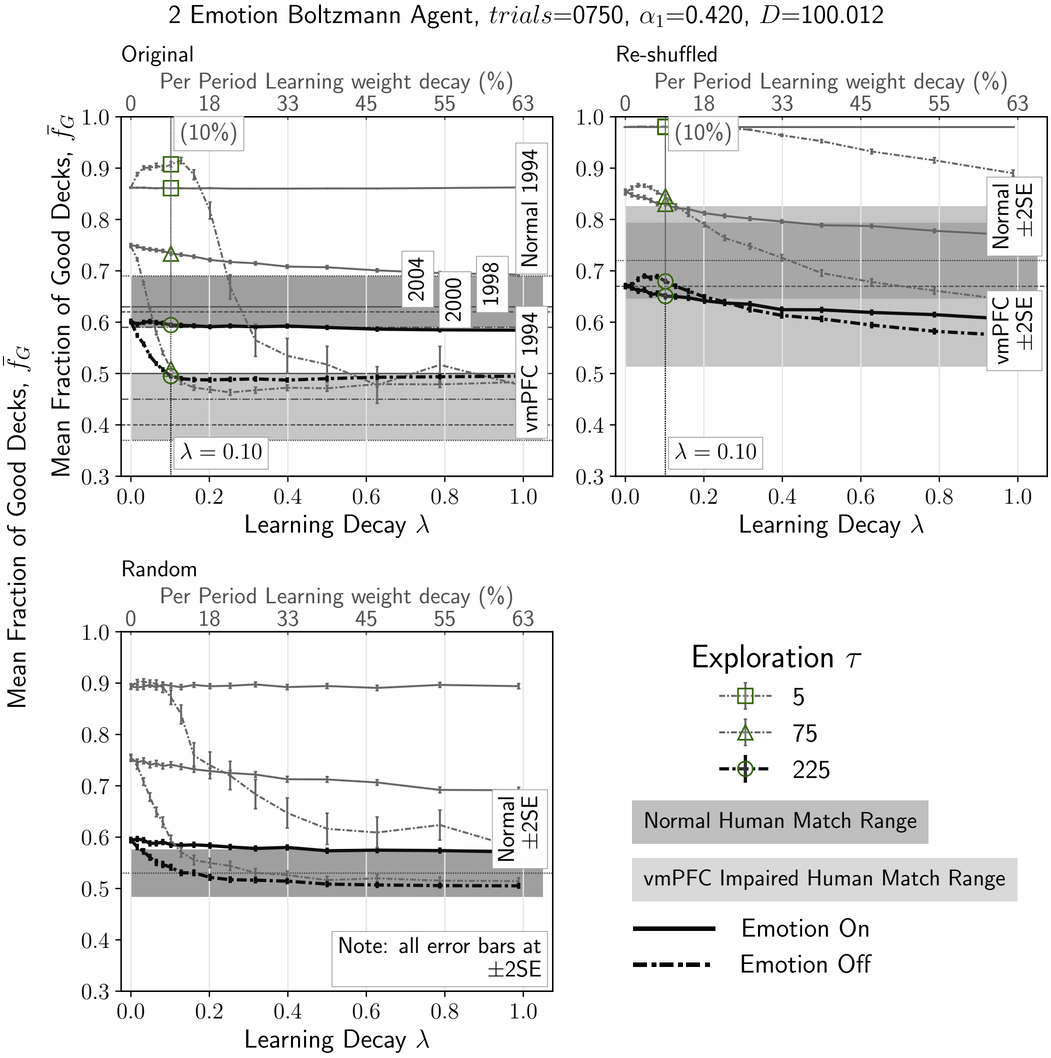


FIGURE 10.3: 2EmST Boltzmann agent CSUD verification grid search 2D contours. Normal and vmPFC impaired behaviours are depicted by solid and dotted lines respectively. For normal behaviour, burst learning produces significant decoupling of learning decay and exploration responses. The dark gray zone represents the normal human IGT outcome match area. The light gray zone represents the vmPFC impaired human IGT outcome match area. Solid black contours show response to learning decay λ at CSUD selected minimum loss hyper-parameter values. Details in text.

Since each IGT environment by construction consists of 50% good, and 50% bad decks and the 50/50 mean as noted in Table 4.1 is 0 net yield, such a limiting point appears plausible (by construction).

The 2D contour view provides an opportunity to discuss the most prominent initial learning rate effect, which consists of a technical non-stationarity effect increasing as learning decay increases from 0. Fig. 10.2 and Fig. 10.3, with vmPFC impaired behaviour, with exploration $\varepsilon = 0.100$ and $\tau = 5$ for ε-Greedy and Boltzmann agents respectively, depict a portion of the respective sinusoidal dotted 2D vmPFC impaired contours which rises above the corresponding solid 2D normal behaviour contours as learning rate decay increases from $\lambda = 0$ towards $\lambda = 0.15$. As the initial learning rate α_1 increases, the amplitude of this sinusoidal region, where the vmPFC impaired dotted 2D contour rises above the corresponding normal solid 2D contour, increases. This increased amplitude effect encapsulates the primary consequences of increasing the initial learning rate α_1.

Therefore in the burst learning model, there is a region of parameter space, consisting of high initial learning rate α_1 and low learning decay λ, where for vmPFC impaired behaviour, strong initial learning rate α_1 and exploration interactions take place. From a decision theoretic perspective, this effect is not taken into consideration here because there is no support for such a potential outcome in human IGT data. That is, original IGT vmPFC impaired human outcomes do not exceed corresponding normal human outcomes.

It has been well established that, high initial learning rate leads to nonstationarities, especially, when starting learning with Q-values, which have been initialised at arbitrary values (such as 0). In such cases, by reducing the unduly high initial learning rate α_1, increasing learning rate decay λ would naturally lead to outcome improvements. Here focus remains on learning rate decay behaviours, which obtain when the initial learning rate has been specified so as to minimise any nonstationarity effects. Indeed, as Fig. 10.2 and Fig. 10.3 indicate, CSUD searches approximating human IGT outcomes discover hyper-parameter solutions, where learning rate decay is such that the selected points on the vmPFC impaired dotted contours do not lie on the increasing portion of the curve. This suggests that if the computational model presented here describes accurately mathematical human IGT behaviour, then vmPFC impaired humans are able to mitigate initial potential nonstationarities by selecting a relatively low initial learning rate.

Abstracting for a moment from human IGT outcome zones, in the burst learning model, initial learning rate α_1, learning rate decay λ, and exploration effects are sufficiently decoupled for normal behaviour cases, effectively letting hyper-parameter tuning produce the desired mean fraction of good decks $\bar{f}_G$ target. These primary decision making theoretic hyper-parameters are also sufficiently decoupled in the vmPFC impaired behaviour case, and likewise allow hyper-parameter tuning through to a specific performance target.

Considering agents' ability to match all human IGT outcome cases, it is noted that for the 2EmST ε-Greedy agent, as the black 2D contours placed inside the dark and light gray zones indicate in Fig. 10.2, all outcome cases are matched; however for normal behaviour, such matches lie close to the boundaries of the respective catchment zones. In Fig. 10.3 the 2EmST Boltzmann agent achieves a partial search match consisting of the normal and vmPFC impaired, original and re-shuffled IGT environments; however, except for the vmPFC impaired re-shuffled case, remaining matches are on catchment zone boundaries. The normal random catchment zone is only narrowly missed.

It is possible that overall matches are improved by increasing catchment zone sizing. However, such a possibility is not pursued any further here. The construction of the catchment zones employed is described in section 4.2.1 and Table 4.3 to Table 4.8.

Fig. 10.4 and Fig. 10.5 depict 2EmST agent CSUD verification grid search 3D contours. Green and blue coloured surfaces show normal and vmPFC impaired behaviour response respectively to learning decay λ and attenuation D at CSUD selected minimum loss hyper-parameter values. The diamond ⧫ and inverted triangular ▼ shapes mark CSUD minimum loss normal and vmPFC impaired learning decay and attenuation rates respectively.

Regarding the initial learning rate α_1, it has already been established in previous chapters that due to exponential learning rate decay, the initial learning rate does not have much influence. As the 3D contours in Fig 10.4 and Fig 10.5 illustrate, this result does not change. Further, the 2EmST agent CSUD 3D grid search verification plots reinforce the results of the corresponding 2D plots in Fig 10.2 and Fig 10.3.

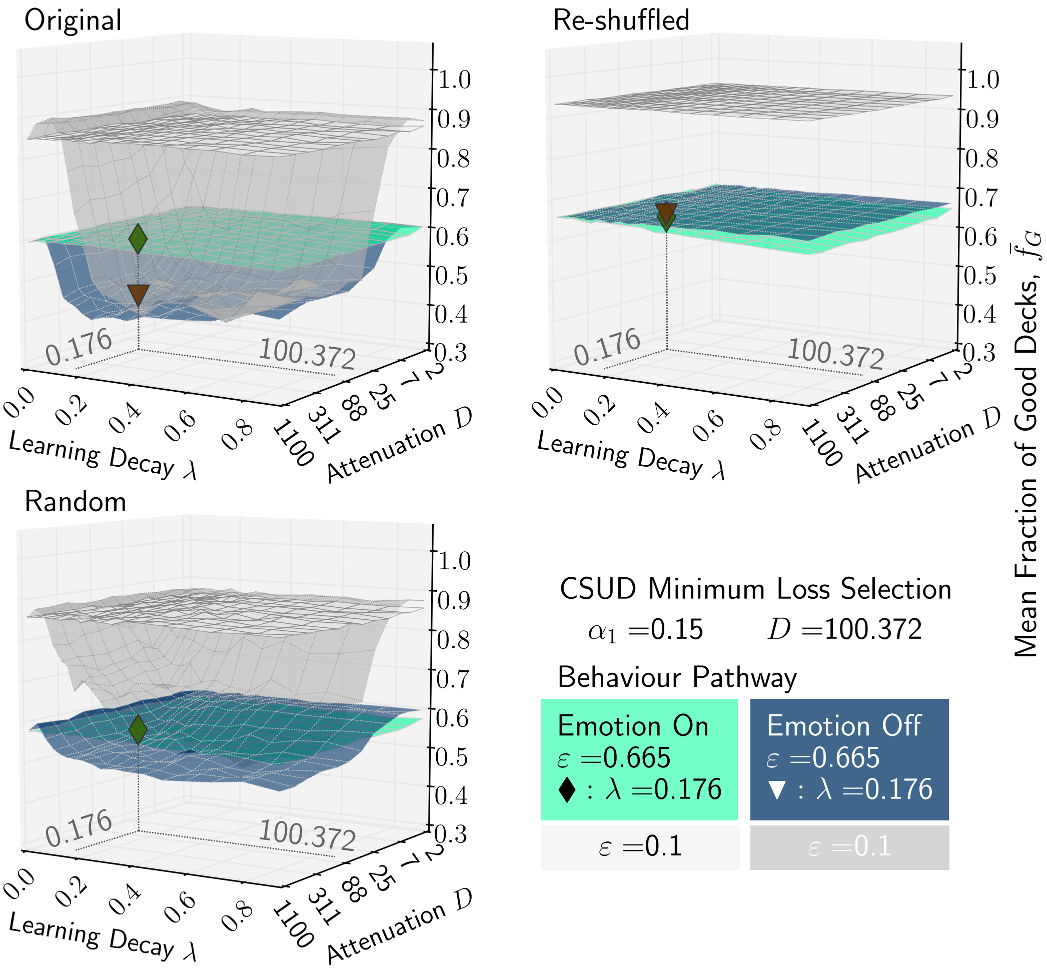


FIGURE 10.4: 2EmST ε-Greedy agent CSUD verification grid search 3D contours. Green and blue coloured surfaces show normal and vmPFC impaired behaviour response respectively to learning decay λ and attenuation D at CSUD selected minimum loss hyper-parameter values. The diamond ♦ and inverted triangular ▼ shapes mark CSUD minimum loss normal and vmPFC impaired learning decay and attenuation rates respectively. When emotion is 'On' corresponding normal human IGT outcomes are matched. When emotion is 'Off' corresponding vmPFC impaired human IGT outcomes are matched. The grey 3D contours show agent emotion behaviour response at ε = 0.1.

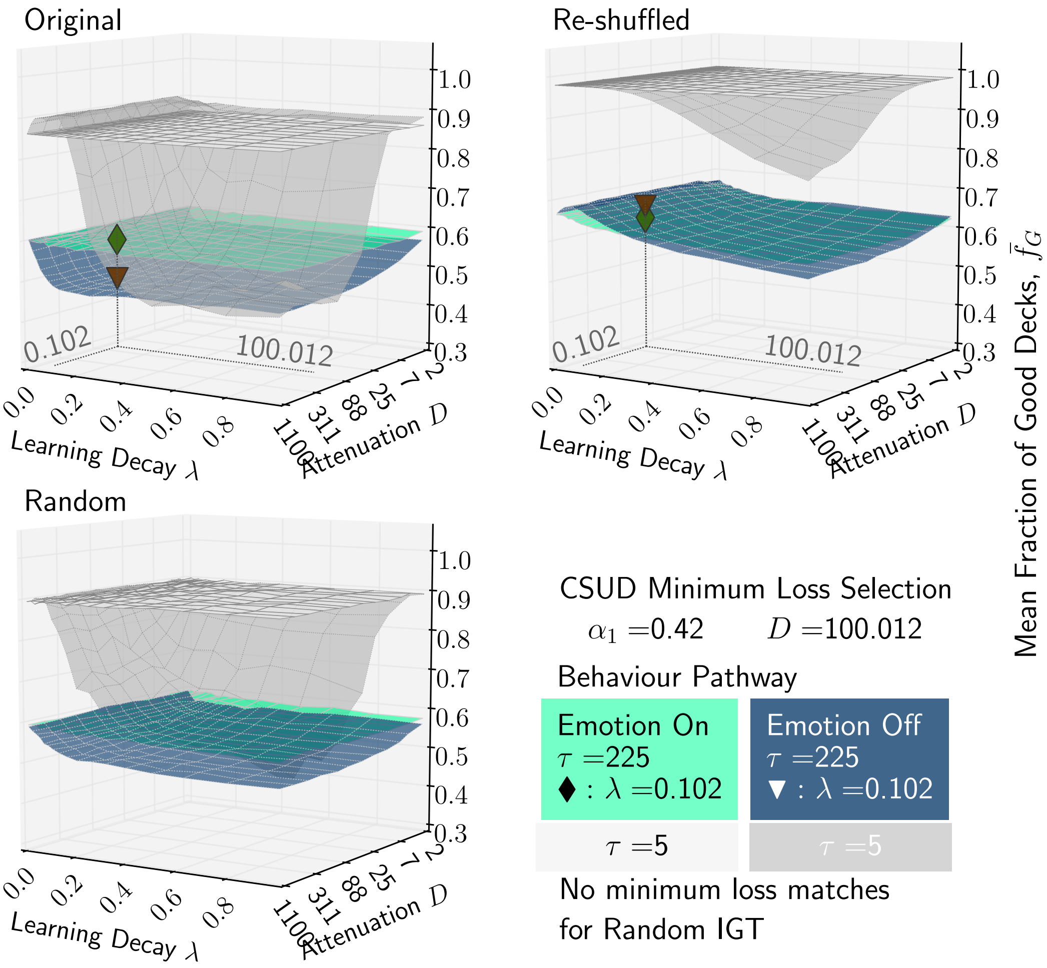


FIGURE 10.5: 2EmST Boltzmann agent CSUD verification grid search 3D contours. Green and blue coloured surfaces show normal and vmPFC impaired behaviour response respectively to learning decay λ and attenuation D at CSUD selected minimum loss hyper-parameter values. The diamond ♦ and inverted triangular ▼ shapes mark CSUD minimum loss normal and vmPFC impaired learning decay and attenuation rates respectively. When emotion is ′On′ corresponding normal human IGT outcomes are matched. When emotion is ′Off′ corresponding vmPFC impaired human IGT outcomes are matched. The grey 3D contours show agent emotion behaviour response at $\tau = 5$.

The 3D plots in Fig. 10.4 and Fig 10.5 show that for both the 2EmST ε-Greedy and Boltzmann agents, as indicated by the relatively flat normal behaviour green coloured 3D surfaces, increasing learning decay λ and attenuation D has little effect on the mean fraction of good decks $\bar{f}_G$ outcomes.

In contrast, for the original and random IGT environments with vmPFC impaired behaviour, the blue coloured 3D surfaces present a bowl like response with reduced $\bar{f}_G$ outcomes as learning decay λ and attenuation D increase. By design in the re-shuffled IGT environment, this bowl like response is substantially mitigated.

For low exploration at $\varepsilon = 0.1$ and $\tau = 5$, the gray 3D contours for the 2EmST ε-Greedy and Boltzmann agents in Fig. 10.4 and Fig. 10.5 respectively, indicate that lower exploration increases mean fraction of good decks $\bar{f}_G$ outcomes. However, in the original and random IGT with vmPFC impaired behaviour, this leads to deeper bowls in response to increased learning rate decay and attenuation.

Fig. 10.6 shows the effect of increasing the learning rate. In this case, increasing the learning rate from the CSUD discovered value $\alpha_1 = 0.15$ to $\alpha_1 = 0.65$ for the 2EmST ε-Greedy agent with vmPFC impaired behaviour, leads to increased nonstationarity effects on the far edges of the blue and gray bowl shaped 3D surfaces.

Fig. 10.7 and Fig. 10.8 present 2EmST agent 20-draw block behaviour for mean fraction of good decks $\bar{f}_G$ and exploration index (EI) respectively. In both figures, human reference and agent results appear in light gray dotted lines and dark gray solid lines respectively. Agent results are averaged from 750 samples. All error bars are at $\pm 2SE$.

Fig. 10.7 shows that agent 20-draw block mean fraction of good decks $\bar{f}_G$ lie within the $\pm 2SE$ of corresponding human IGT outcomes for the normal and vmPFC impaired, original and re-shuffled IGT environments. However, for the normal random IGT case for both agents, in blocks 1-20 and 21-40, agent $\bar{f}_G$ lies above $\pm 2SE$ of corresponding human values. The agent curves appear parabolic and asymptotic, while the corresponding human curves have a sigmoid shape. This indicates that possibly neither the Boltzmann nor the ε-Greedy agents fully capture the nature of human exploration.

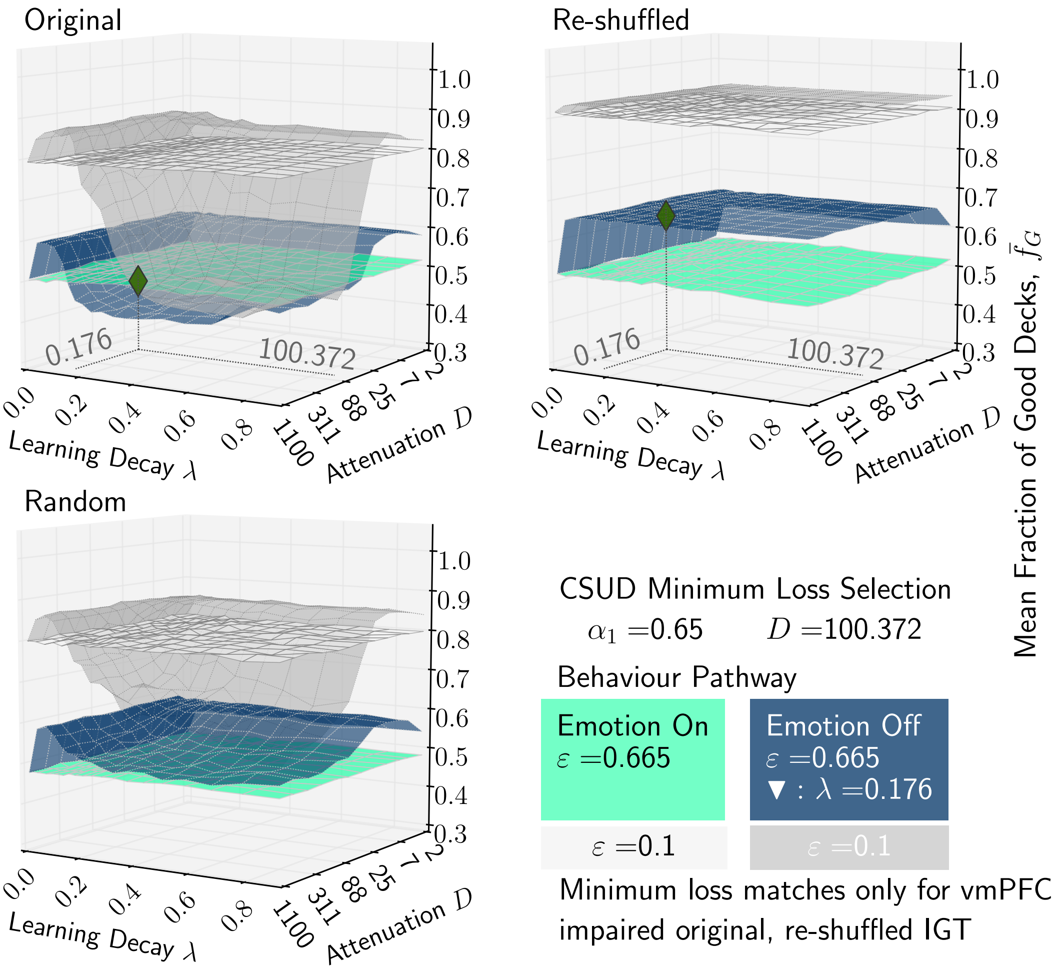
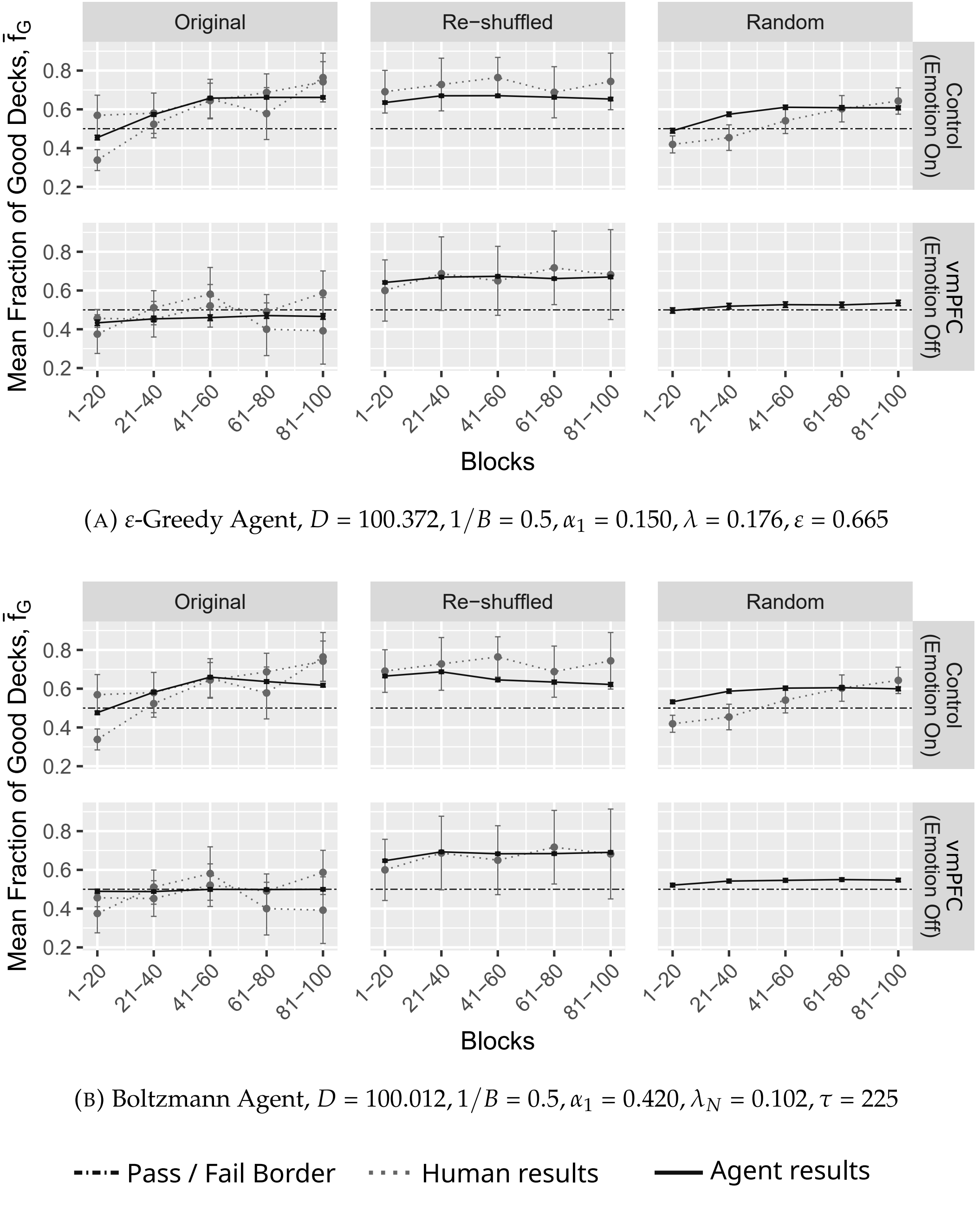


FIGURE 10.6: 2EmST ε-Greedy agent CSUD search grid verification 3D contours for α_1 = 0.65. Green and blue coloured surfaces show normal and vmPFC impaired behaviour response respectively to learning decay λ and attenuation D at CSUD selected minimum loss hyper-parameter values. The inverted triangular ▼ shape marks CSUD minimum loss vmPFC impaired learning decay and attenuation rates respectively. When emotion is 'Off' corresponding vmPFC impaired human IGT outcomes are matched. The grey 3D contours show agent emotion behaviour response at ε = 0.1. Compared to α_1 = 0.15 in Fig 10.4, at high initial learning rate α_1 = 0.65, the edges of the vmPFC impaired behaviour surfaces reveal increased nonstationarity effects. At α_1 = 0.65, CSUD minimum loss matches are only achieved for the vmPFC impaired original and re-shuffled IGT environments.

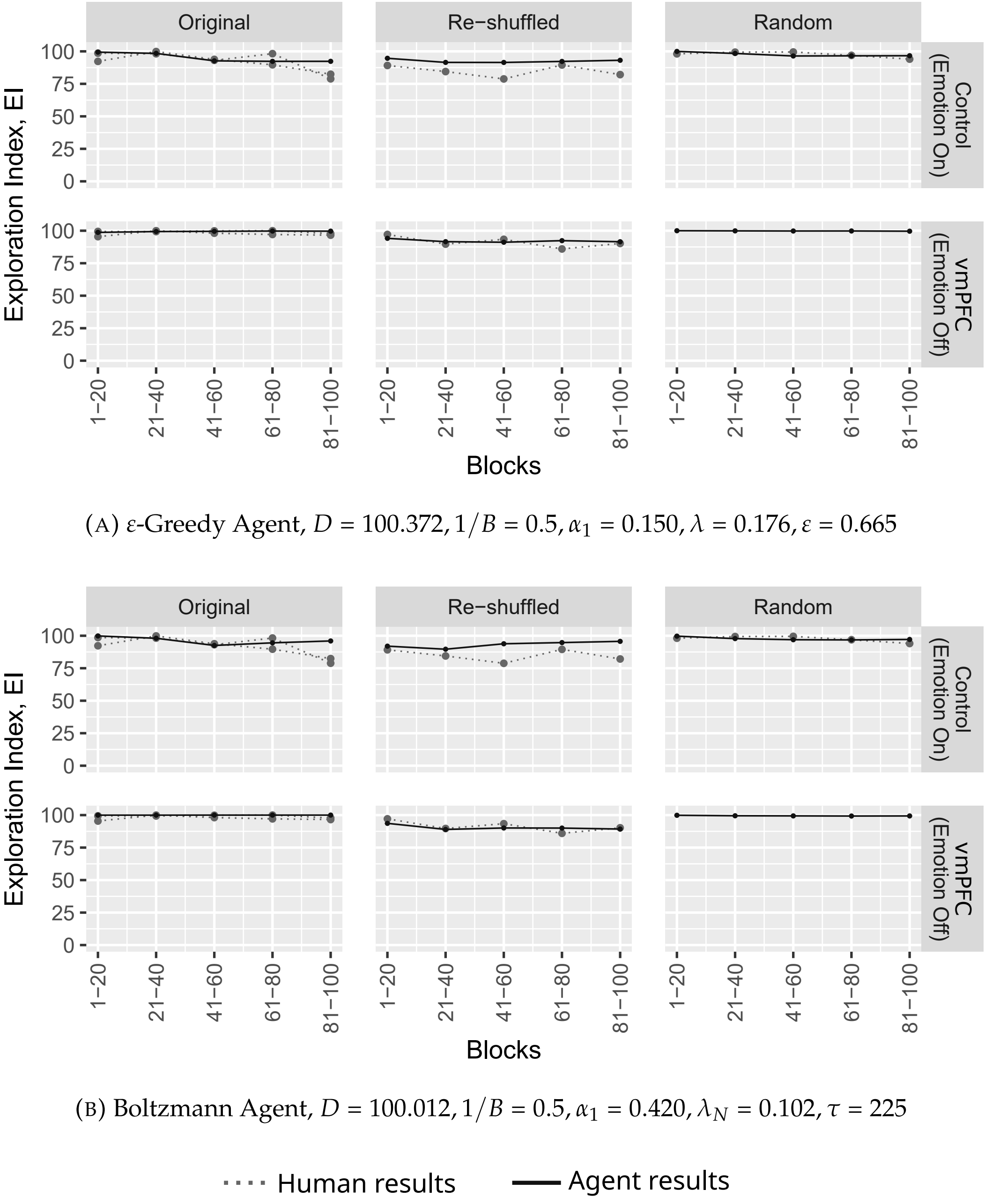


(A) ε-Greedy Agent, $D = 100.372, 1/B = 0.5, \alpha_1 = 0.150, \lambda = 0.176, \varepsilon = 0.665$

(B) Boltzmann Agent, $D = 100.012, 1/B = 0.5, \alpha_1 = 0.420, \lambda_N = 0.102, \tau = 225$

(C) Legend

FIGURE 10.7: 2EmST agent 20-draw blocks comparison at CSUD search values as indicated above. Human reference results in dotted light gray. Agent results in solid dark gray, averaged from 750 samples. All error bars at $\pm 2SE$. When error bars are taken into account agent and human 20-draw block performance appears relatively similar, except for the random IGT. Details in text.



(A) ε-Greedy Agent, $D = 100.372, 1/B = 0.5, \alpha_1 = 0.150, \lambda = 0.176, \varepsilon = 0.665$

(B) Boltzmann Agent, $D = 100.012, 1/B = 0.5, \alpha_1 = 0.420, \lambda_N = 0.102, \tau = 225$

(C) Legend

FIGURE 10.8: 2EmST agent 20-draw block exploration index (EI) comparison at CSUD search values as indicated above. Human reference results in dotted light gray. Agent results in solid dark gray, averaged from 750 samples. Human subject and agent exploration index responses appear relatively similar except in the normal re-shuffled IGT, where agent EI appears higher. Details in text.

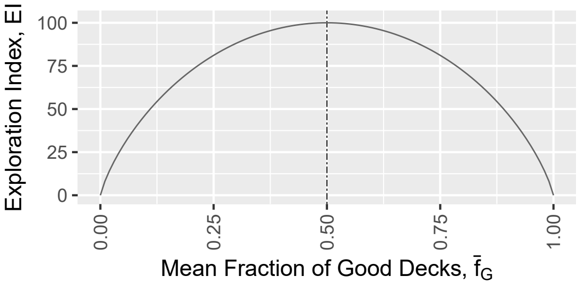


FIGURE 10.9: Exploration index (EI) for mean fraction of good decks $\bar{f}_G$. The solid line indicates the exploration index, while the dash dotted line highlights $\bar{f}_G = 0.5$. Note that $EI = 100$ at $\bar{f}_G = 0.5$.

Fig. 10.8 shows 2EmST agent 20-draw block exploration index (EI) outcomes. The exploration index (EI) profiles presented in Fig. 10.8 are similar to those obtained in Fig. 6.5 and Fig. 6.11 for the simple reinforcement agents; and in Fig. 9.7 for the ARA(κ) agents. Therefore the inclusion of the model reset mechanism, which normal configured 2EmST agent employ, does not appear to be associated with a change in exploration index (EI) behaviour. Recall that the exploration index (EI) measured implied exploration based on realized choices, and that learning rate decay could thus have an indirect effect on the exploration index via the initial learning rate. Such an theoretical effect, however, is not exhibited in Fig. 10.8.

The exploration index (EI) ranges from 0 to 100, with 0 indicating full exploitation and 100 indicating full exploration (i.e., random search). In the sense of the No Free Lunch theorems (Wolpert & Macready, 1997), the exploration index measures search algorithm specificity, with 0 indicating a fully specific response, and 100 indicating random search. As no specific algorithm can provide universal solutions over all algorithms and search problems, an increased EI can be seen as an attempt to build hybrid search algorithms, which attempt to increase search space generalisation by increasing randomness.

The exploration index (EI) is not a linear scale, and is described in more detail in section 4.2.3. For two selection options, such as fraction of good decks versus fraction of bad decks, however, EI can be readily visualised and is presented in Fig. 10.9. With reference to Fig. 10.9, note that based on Fig. 10.8a and Fig. 10.8b neither the ε-Greedy nor the Boltzmann agent respectively can fully capture human exploration.

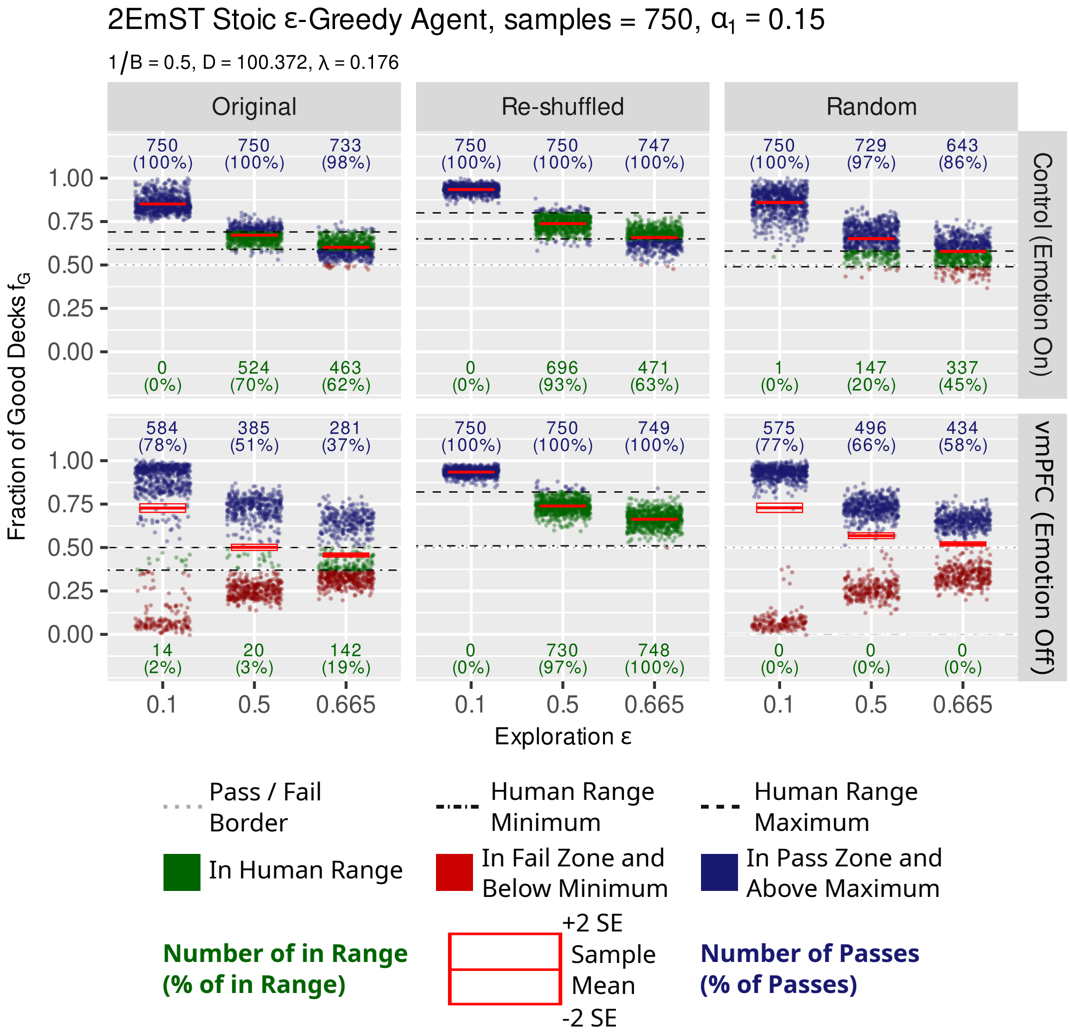


FIGURE 10.10: 2EmST ε-Greedy agent, CSUD minimum at $\varepsilon = 0.665$. Comparison of repeated simulation outcomes to human IGT results. 2EmST ε-Greedy agent jitter plots reveal that normal behaviour (emotion on) configuration nolonger exhibits bi-modality. Full details are in the text.

This is noted in the respective normal original and re-shuffled IGT environments at block 81-100, where human reference EI values are lower indicating a greater reduction in exploration than that evidenced in agent results.

Fig. 10.10 and Fig. 10.11 depict success of 2EmST agent simulation outcomes, and provide agent jitter plot density summary for the fraction of good decks f_G outcomes obtained at the CSUD selected minimum loss agent hyper-parameter values from 750 simulated samples for normal (control, emotion on) and vmPFC impaired (emotion off) configurations. Green dots mark outcomes inside human performance ranges. Blue dots mark additional normative pass results, whereas red dots mark additional normative

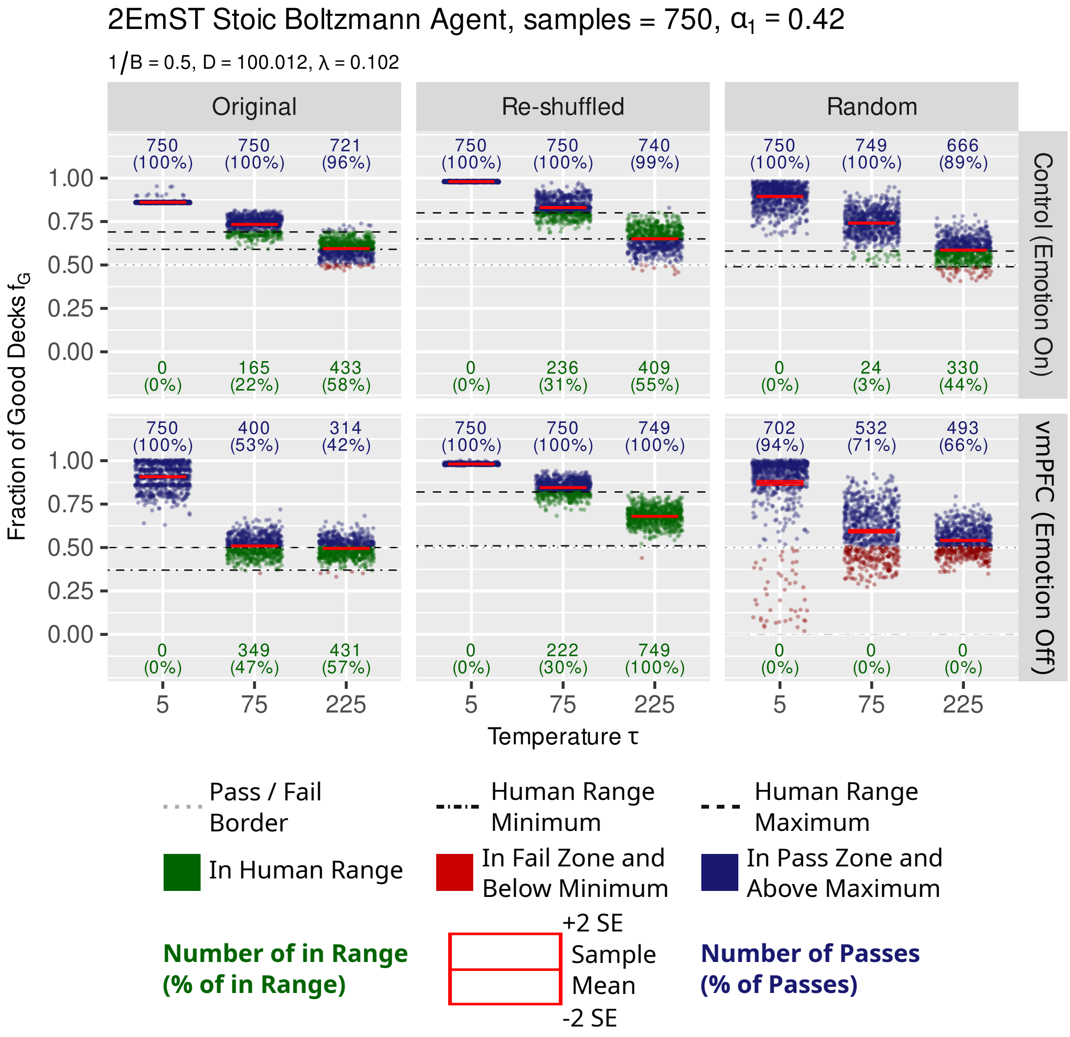


FIGURE 10.11: 2EmST Boltzmann agent, CSUD minimum at $\tau = 225$. Comparison of repeated simulation outcomes to human IGT results. Full details are in the text.

fails. Green numbers give total matches out of 750 samples, and the values in brackets indicate percentages matched. A high percentage matched value is indicative of predictive simulation success. The dashed and dash-dotted horizontal lines indicate the maximum and minimum respectively of the human match range. Finally the red bars and box indicate central tendency in terms of the mean and ±2 SEs (standard errors).

For 2EmST agents, with normal behaviour configured agents capable of resetting to the initial learning rate α_1, the ε-Greedy agent as noted in Fig. 10.10 exhibits a significant shift in behaviour when compared with results obtained with the simple and ARA(κ) reinforcement learning agents. Unlike previous ε-Greedy agents, the 2EmST ε-Greedy agent with normal behaviour does not display bi-modal fraction of good decks f_G outcomes.

In Fig. 9.8 for example, the normal behaviour ARA(κ) ε-Greedy agent with CSUD loss minimising exploration at $\varepsilon = 0.612$ displays for the fraction of good decks f_G, a heavy tail in the original, and a bi-modal density in the random IGT environments. In contrast in Fig. 10.10 the normal behaviour 2EmST ε-Greedy agent, with CSUD loss minimising exploration at $\varepsilon = 0.665$, displays for f_G, a clustered (unimodal) density both in the original and random IGT environments. This result is significant because it indicates that the current model with learning rate resetting and an ε-Greedy agent provides an alternative to the Boltzmann architecture for achieving unimodal f_G jitter plot outcomes.

As Fig. 10.10 shows, the vmPFC impaired behaviour 2EmST ε-Greedy agent presents with reduced bi-modality when compared in Fig. 6.6 and Fig. 9.8 to the simple and ARA(κ) ε-Greedy agents respectively. At $\varepsilon = 0.665$, this leads to increased matches in the respective human IGT outcome catchment zones. It is believed that this is a consequence of the 2EmST model requiring lower learning decay for achieving vmPFC impaired behaviour.

As Fig. 10.11 shows, the vmPFC impaired behaviour 2EmST Boltzmann agent presents with some differences when compared in Fig. 6.12 and Fig. 9.9 to the simple and ARA(κ) Boltzmann agents respectively. Both the simple and ARA(κ) Boltzmann agents display at $\tau = 5$, in the original and random environments, heavy tailed or bi-modal jitter plots. In the 2EmST Boltzmann agent, the strength of this tendency is diminished. As mentioned above, this observation is possibly a consequence of the 2EmST model requiring lower learning decay for achieving vmPFC impaired behaviour.

Test Variant	Test Statistic	df1	df2	p-Value	Subset Results
Original | Re-Shuffled | Random vs. Behaviour					At $\alpha = 0.01$, the null hypotheses of behaviour factor equality is rejected. Equality of the re-shuffled | random, re-shuffled, and random subset responses cannot be rejected.
ANOVA Type	79.856	2.984	4469.406	0	
Wilks Lambda	88.976	3.000	1496.000	0	
Re-Shuffled | Random vs. Behaviour					
ANOVA Type[a]	2.065	2	2995.551	0.127	Joint response is not significant, no subsets.
Re-Shuffled vs. Behaviour					
ANOVA Type	1.927	1.000	1498	0.165	Single response variable, no subsets.
Wilks Lambda	1.927	1.000	1498	0.165	
Random vs. Behaviour					
ANOVA Type	2.203	1.000	1498	0.138	Single response variable, no subsets.
Wilks Lambda	2.203	1.000	1498	0.138	

[a]Wilks Lambda could not be computed due a singular rank matrix.

TABLE 10.6
2EmST ε-Greedy agent np-M/ANOVA analysis of mean fraction of good decks $\bar{f}_G$ at $D = 100.372$, $1/B = 0.5$, $\alpha_1 = 0.150$, $\lambda = 0.176$, $\varepsilon = 0.665$, with behaviour (normal, vmPFC impaired) as factor. At significance level $\alpha = 0.01$, mean fraction of good decks $\bar{f}_G$ responses are statistically significantly different, except for the re-shuffled | random, re-shuffled, and random IGT environment subsets.

Table 10.6 and Table 10.7 present np-M/ANOVA results for the 2EmST ε-Greedy and Boltzmann agents respectively. The np-M/ANOVA tests check to see whether behaviour configuration as normal (emotion on) or vmPFC impaired (emotion off) produces a statistically significant joint mean fraction of good decks $\bar{f}_G$ difference across the original, re-shuffled, and random IGT environments, and their subset combinations. Test variants can be thought of as non-parametric versions of the F-test, with a test statistic and two degrees of freedom. These three quantities are then assessed to derive the p-value. The test statistics are discussed in Burchett et al., 2017.

The null hypothesis is that there is no emotion induced factor effect. At significance level 0.01, a p-value less than 0.01 leads to the rejection of this null, whereas p-values greater than 0.01 indicate that the null hypothesis of no factor effects cannot be rejected.

Based on human IGT mean fraction of good deck $\bar{f}_G^H$ results, the expectation is that for the original decks, a statistically significantly different $\bar{f}_G$ is found for normal (emotion on) versus vmPFC impaired (emotion off) behaviour. On the other hand, the expectation is that for the re-shuffled decks, the null hypothesis of equal $\bar{f}_G$ cannot be rejected with respect to normal and vmPFC impaired behaviour. While there is no human outcome data for the vmPFC impaired random IGT case, based on simulation results, a test is also included for the random IGT environment. Human data to validate vmPFC configured random IGT simulation results does not exist; however, one can hypothesize expected human outcomes for the random IGT environment. In particular, the ε-Greedy 2EmST burst learning model predicts that should a normal versus vmPFC impaired random IGT variant test be conducted with human participants, then there should be no significant factor effect due to vmPFC impairment at $\alpha - 0.01$.

For the ε-Greedy agent, Table 10.6 reveals at significance level $\alpha = 0.01$ a statistically significant mean fraction of good decks $\bar{f}_G$ joint response across all three, the original, re-shuffled, and random IGT environments. Subsetted responses, indicate that the null hypothesis of a statistically significant behaviour driven $\bar{f}_G$ response fails to be rejected for the re-shuffled and random, re-shuffled only, and random only IGT environments. In terms of human benchmarks, rejecting the null with respect to the original IGT, and failure to reject the null with respect to the re-shuffled IGT environment is the expected response. Additionally, the results predict that if human IGT decision-making is adequately explained by the 2EmST ε-Greedy model, then an ANOVA looking at human mean fraction of good decks $\bar{f}_G^H$ for normal, vmPFC impaired behaviour differences for the random IGT should not yield statistically significant results.

Table 10.7, showing np-M/ANOVA results for the 2EmST Boltzmann agent with behaviour (normal, vmPFC impaired) as factor, reveals at significance level $\alpha = 0.01$, a statistically significant mean fraction of good decks $\bar{f}_G$ joint response across all three, the original, re-shuffled, and random IGT environments. Additionally all sub-setted response combinations also produce statistically significantly different outcomes. This indicates that the Boltzmann agent does not match expected human outcomes. For the re-shuffled IGT environment, one would have expected failure to reject the

Test Variant	Test Statistic	df1	df2	p-Value	Subset Results
Original | Re-Shuffled | Random vs. Behaviour					At $\alpha = 0.01$, the null hypotheses of behaviour factor equality is rejected for all response variable subsets
ANOVA Type[a]	438.663	2.776	4158.388	0	

[a] Wilks Lambda could not be computed due a singular rank matrix.

TABLE 10.7
2EmST Boltzmann agent np-M/ANOVA analysis of mean fraction of good decks $\bar{f}_G$ at $D = 100.012$, $1/B = 0.5$, $\alpha_1 = 0.420$, $\lambda = 0.102$, $\tau = 225$, with behaviour (normal, vmPFC impaired) as factor. At significance level $\alpha = 0.01$, joint and all sub-setted mean fraction of good decks $\bar{f}_G$ responses are statistically significantly different.

null hypothesis of no factor effect; however, this is not the case.

Fig. 10.11, at $\tau = 225$ for the re-shuffled IGT, shows that the Boltzmann agent achieves a high number of control and vmPFC impaired matches. The np-M/ANOVA results however reveal statistically significantly different $\bar{f}_G$ of the respective jitter plot clusters. Therefore while in the re-shuffled IGT case the Boltzmann agent visually places normal and vmPFC impaired agent $\bar{f}_G$ inside the respective human catchment zones, statistically speaking the two distinct behaviour result means do not appear to come from the same distribution. The 2EmST Boltzmann agent's divergent np-M/ANOVA results may be due to the agent's exploration architecture, which enforces tight distributions due to Q-value proportional exploration.

Fig. 10.12 and Fig. 10.13 depict 2EmST agent per period average learning rate $\bar{\alpha}_t$ progression for normal and vmPFC impaired original, re-shuffled, and random IGT environments. The solid black lines represent mean learning rate $\bar{\alpha}_t$ in IGT period t. The red vertical pinhead lines indicate for normal (control) behaviour configured agents, the maximum and minimum range of learning rate values at period t across all agents, with the minimum marked by a dark red pinhead. The blue solid line indicates smoothed trends. The top dashed line marks the initial learning rate α_1. The bottom dash-dotted line marks the learning rate bound α_1/D, and the dotted vertical line marks the time-to-bound TTB. For convenience, the numeric α_1, α_1/D, TTB values are depicted in green in the vmPFC impaired, random IGT case grid panel.

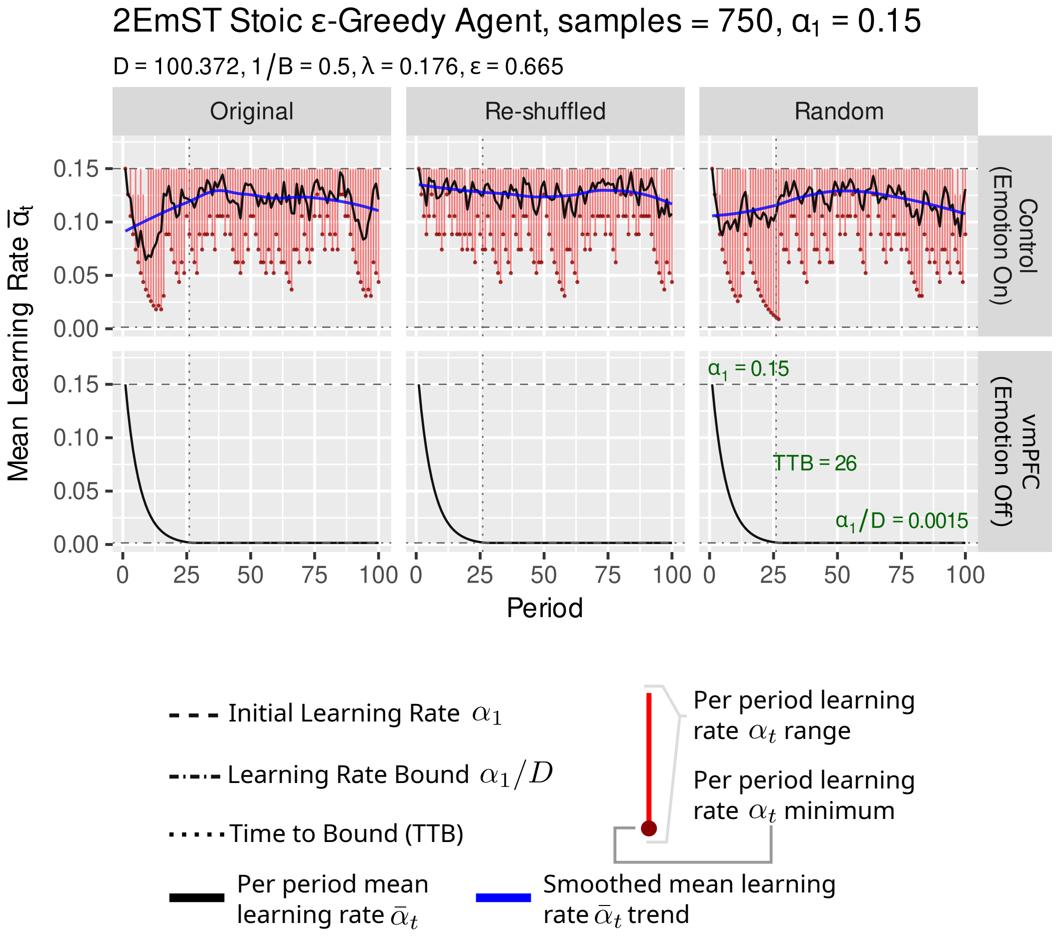


FIGURE 10.12: 2EmST ε-Greedy agent per period mean learning rate $\bar{\alpha}_t$ progression for normal and vmPFC impaired original, re-shuffled, and random IGT environments. Red saw-tooth patterns indicate burst learning. Full details are in the text.

In Fig. 10.12 and Fig. 10.13 for 2EmST ε-Greedy and Boltzmann agents respectively, for normal configured 2EmST agents, the red saw-tooth patterns indicate that in any IGT period t, there are agents with decaying learning rate, as well as agents engaging in emotion mediated learning rate re-setting. That is, the saw-tooth patterns indicate episodes of burst learning. In general, the dark solid mean learning rate $\bar{\alpha}_t$ line appears above the mid-way point of the red pinhead lines, indicating that at any one period, there are more learning rate re-setting than learning rate decaying agents.

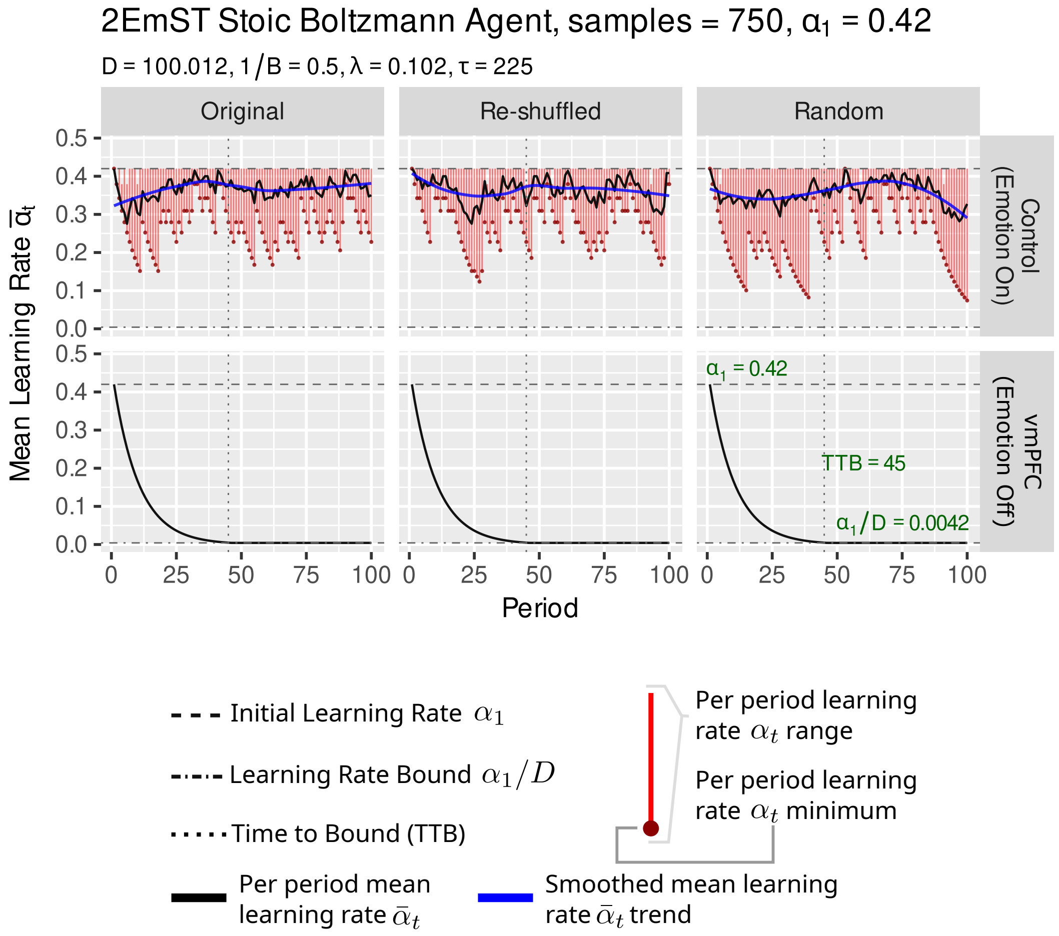


FIGURE 10.13: 2EmST Boltzmann agent per period mean learning rate $\bar{\alpha}_t$ progression for normal and vmPFC impaired original, re-shuffled, and random IGT environments. Red saw-tooth patterns indicate burst learning. Full details are in the text.

However the blue smoothed trends indicate that, with the exception of the 2EmST Boltzmann agent normal configured original IGT case, the remaining behaviour and IGT environment cases show a decrease in mean learning rate $\bar{\alpha}_t$ with increasing periods; this indicates that over time, the number of learning rate decaying agents is increasing. That is, as normal behaviour 2EmST agents learn the respective IGT environment, their Q-value representation improves and temporal difference errors, which exceed the stoic emotion agent activation threshold, decrease.

As Fig. 10.12 and Fig. 10.13 for 2EmST ε-Greedy and Boltzmann agents respectively indicate, vmPFC impaired 2EmST agents do not exhibit emotion mediated learning rate re-setting, and consequently, the respective learning rates decline exponentially from the initial learning rate α_1 to the lower bound α_1/D in time-to-bound TTB periods. For both agents, the initial learning rate declines approximately by 100 fold. However, the 2EmST ε-Greedy agent is configured with higher learning decay $\lambda = 0.176$, and this is why this agent has a shorter time-to-bound of only 26 periods, which is approximately a quarter of the IGT task duration.

10.5 Discussion: Burst Learning Model

The burst learning model has been presented with reference to emotion triggered signals. Strictly speaking the proposed error correction heuristics in Table 10.2 do not require emotion labels. It is hoped, however, that the use of emotion labels is helpful in drawing attention to the possible role of this computational architecture in explaining the contributions of emotion and vmPFC impairment in decision making.

The burst learning model is implemented with (10.4), (10.5), and (10.8) with stoic emotion activation threshold behaviour as described in Table 10.2. The implementation is completed using ε-Greedy and Boltzmann exploration architectures introduced in sections 5.2.2 and 5.2.1 respectively.

Of the emotion activation thresholds presented in Table 10.2, the stoic and buffered emotion activation strategies have the lowest activation threshold ($-1/B$) for engaging the learning rate reset mechanism. However with a two-behaviour response pathway, the stoic agent possesses the simpler behavioural pathway. In contrast, in order to achieve learning decay, the tempered emotion activation strategy always requires a positive temporal difference error. At $1/B = 0.5$, this would require that the temporal difference error is consistently equal to or more than half the current aggregated action value estimate $Q_{t-1}(a)$. Consequently, the tempered strategy would lead to comparatively more emotion mediated learning rate resetting with a slower learning rate decay profile, leading to a larger difference between normal and vmPFC impaired configured agents. Here the stoic strategy is preferred over the tempered strategy, because the former strategy should in theory make it more difficult to obtain normal and vmPFC impaired behavioural configuration differences. That is, if stoic threshold emotion activation produces normal and vmPFC impaired behaviour configuration IGT simulation outcome differences, so would the tempered threshold emotion activation strategy. Further, the tempered strategy does not admit of any loss potential, and it is believed that human decision making models should have some built-in tolerance for adverse outcomes.

Using CSUD search as a model hyper-parameter tuner, the burst learning model agent implementations are calibrated with the original, re-shuffled, and random IGT environments to discover agent hyper-parameter configurations capable of generating simulated IGT outcomes matching respective normal and vmPFC impaired behaviour human outcomes.

Then grid searches are conducted around the CSUD discovered minimum loss 2EmST agent hyper-parameter values. Grid search IGT simulated mean fraction of good decks $\bar{f}_G$ outcomes are used to assess agent performance and agent ability to generate human IGT outcomes in terms of iterative decision making primitives, the learning rate, learning rate decay, and exploration.

Section 7.2 and section 9.3 discuss reasons for foregoing in this work a comprehensive joint CSUD search of all discussed IGT environments; namely, such searches have been attempted but have not produced a full match solution consisting of hyper-parameter combinations yielding agent IGT outcomes in corresponding human IGT outcome match zones. It is believed that the lack of such a joint solution is driven by underlying IGT process yield distribution characteristics, which affect learning rate decay λ selection. It has been noted that learning rate decay λ acts as a band-pass filter in the frequency domain leading to learning cut-off after a certain number of iterations. Hence if certain IGT environments required more iterations than others to be learned well, then the same learning rate decay λ value may not apply to all IGT environments. The original, re-shuffled, and random IGT environments however, are closely related and only differ by the manner in which cards are sequenced in each deck. Given that only one card may be selected at each IGT period, these three IGT environments present an ideal test base for assessing learning rate decay, which may affect a ‘learning freeze’ after a certain number of periods.

Assessment of the original, re-shuffled, and random IGT environments with simple, ARA(κ) ε-Greedy and Boltzmann agents in chapters 6 and 9 respectively produce the decision making results, which consist of (1) learning rate decay λ leads to vmPFC impaired behaviour, (2) for agents to match respective human IGT outcomes, exploration must be high, and (3) at high values, the initial learning rate α_1 produces a technical non-stationarity effect.

In the simple and ARA(κ) reinforcement learning models, two different learning rate decay values λ_N and λ_{vmPFC} are used for generating normal and vmPFC impaired behaviour respectively. In these models, the pathology causing vmPFC impairment somehow leads to an increase of learning rate decay from λ_N to λ_{vmPFC}.

In burst learning, under normal configured behaviour and in the absence of an emotion trigger, the learning rate decays exponentially; however an emotion mediated signal may lead to a reset of the decayed learning rate. Therefore burst learning proposes a model, where vmPFC impairment results not from *ex-machina* factors but from the disruption of a specific *emotion* mediated pathway.

In the absence of an emotion signal, (1) the learning rate cannot be reset, and instead (2) continues to decay as per model dynamics at a fast rate and to a lower bound; the combination of these two effects lead to vmPFC impairment. So from an iterative learning, and decision theoretic point of view, the burst learning model proposes that learning rate decay, combined with the inability to reset to the initial learning rate, provides the impetus for poor decision making as exhibited by vmPFC impaired human IGT participants. Using solely decision making primitives, the burst learning model proposes and captures well a computational approach, and decision making explanation, in terms of the initial learning rate α_1 and learning rate decay λ for modelling vmPFC impairment. The burst learning model presented here provides a mathematical nonrational formulation of decision making and learning mediated by emotion, where the concept of burst learning found in psychology (Kunitani, 2016) and neuroscience (Ohta et al., 2022) is applied in a novel manner using emotion mediated exponential learning rate decay.

However, when exploration is considered as a decision making primitive, the results in chapter 6, chapter 9, and here with constant exploration ε-Greedy and proportional Boltzmann exploration architectures, indicate that neither agent architecture adequately captures the nature of human exploration. For example, as the mean fraction of good decks $\bar{f}_G$ 2D contour plots in Fig. 10.2 and Fig. 10.3; and, the CSUD grid search verification plots in Fig. 10.10 and Fig. 10.11 indicate, burst learning 2EmST agents must have very high exploration in order to produce mean fraction of good decks $\bar{f}_G$ outcomes, which lie within corresponding human IGT outcome catchment zones.

Further as 2EmST agent 20-draw block mean fraction of good deck $\bar{f}_G$ and exploration index (EI) figures, Fig. 10.7 and Fig. 10.8 respectively reveal, agents are unable to fully mirror human reference characteristics across all behaviour and IGT environment cases. In particular, systematic $\bar{f}_G$ deviations appear in blocks 1-20, 21-40, and possibly 41-60 in the healthy, random IGT case. In Fig. 10.7, the healthy, random IGT case human 20-draw block mean fraction of good decks $\bar{f}_G$ profiles have a sigmoid shape whereas the corresponding agent results produce a parabolic segment like shape. For the exploration index (EI), 2EmST agent 20-draw block EI values are consistently overestimated in the normal re-shuffled IGT case. These variations do not appear to affect the key result regarding the effect of emotion mediated learning rate decay in generating normal versus vmPFC impaired IGT outcome matches.

Further as Fig. 10.8 reveals, in terms of the exploration index (EI) measure, the observed 20-draw block mean fraction of good decks $\bar{f}_G$ differences do not transform into large exploration index value differences. While agent human exploration behaviour differences do not appear to influence emotion mediated learning rate decay effects, such exploration behaviour differences do suggest that human exploration in the IGT is using a different approach than those employed by either the ε-Greedy or Boltzmann agent architectures.

The jitter plots summarising 750 repeated simulation f_G outcomes in Fig. 10.10 show that in the 2EmST ε-Greedy agent, burst learning leads to increased human IGT outcome matches for the normal behaviour, original and random IGT cases, and also for the vmPFC impaired, original case. Compared to Fig. 6.6 and Fig. 9.8 for simple and ARA(κ) ε-Greedy agents respectively, as Fig. 10.10 shows, the burst learning specification also leads to removal of bi-modality for the normal, and reduction of bi-modality for the vmPFC impaired cases. This shows that with burst learning the heuristic constant exploration ε-Greedy agent can achieve, under normal behaviour, the same unimodal f_G outcome characteristics as those exhibited by the rational Boltzmann agent noted in Fig. 10.11. There is, however, a significant difference. In the case of the 2EmST Boltzmann agent, this unimodal f_G outcome is a direct consequence of the exploration achitecture. In the case of the 2EmST ε-Greedy agent, however, the unimodal f_G outcome comprises an emergent behaviour arising from the interaction of the agent with the decision making task.

Unfortunately individual human reference data is not available for jitter plot comparison. However, Fig. 9.11 shows control human jitter plot reference data for the random IGT. Note that in general, the human jitter plot demonstrates wider dispersion than that observed for either agent at normal behaviour with the random IGT environment. In sum, it appears that the 2EmST exploration architectures produce simulation results, which do not quite match corresponding human IGT characteristics. As mentioned above, these agent human exploration differences do not appear to influence emotion mediated learning rate decay effects exhibited in the presented burst learning model. The exploration difference results, however, indicate that further research into the nature of human exploration is needed.

The key human IGT statistical result (ANOVA, Neuman-Keuls) indicates that when grouped into control and vmPFC impaired categories, for the original IGT, the vmPFC impaired group had significantly worse $\bar{f}_G^H$, while for the re-shuffled IGT, no statistically significant $\bar{f}_G^H$ difference existed (Fellows & Farah, 2005). With the burst learning model, as well as with the simple and ARA(κ) reinforcement learning models, in np-M/ANOVA tests only the ε-Greedy agent variant could match this result. In the case of Boltzmann agent variants, np-M/ANOVA returned a significant factor (group) effect between normal (control) and vmPFC impaired behaviour configured agent re-shuffled IGT $\bar{f}_G$ outcomes.

Possibly the divergence of the Boltzmann agent from the established human reference results originates from its proportional exploration architecture. It appears that proportional exploration leads to tight dispersion in behaviour and IGT environment case fraction of good deck f_G outcomes; it is this tight dispersion that may lead to the observed np-M/ANOVA result mismatch.

In sum, the burst learning model produces good results with respect to the effects of the initial learning rate α_1, learning rate decay λ, and learning rate bursts α_t, as depicted in Fig. 10.12 and Fig. 10.13. However, agent exploration architectures appear to produce simulation results, which only partially match observed human results.

Chapter 11

Future Directions

This chapter indicates future directions suggested by the results obtained in chapters 5 to 10. In these chapters, the CSUD search strategy was used to target human IGT outcomes to calibrate nonrational Q-learning models, where the to be calibrated hyper-parameters primarily consisted of the initial learning rate, learning rate decay, and exploration.

In CSUD search when hyper-parameters do not share the same scale, that is the same order of magnitude, search space traversal may suffer. This effect has been observed for example in the Boltzmann agent temperature τ, or in the 2EmST agent attenuation D hyper-parameters. Further, as seen in grid search verification, when there are multiple solutions, CSUD will tend to gravitate towards one of them. This work however did not consider mitigating these CSUD search issues. The rationale for this omission was provided ex-post, as 2D and 3D grid search plots did not reveal any adverse affects, which may have arisen from the lack of space traversal in the CSUD discovered τ or D values. While grid search results indicated the presence of multiple hyper-parameter solutions, CSUD performed well at approximating the targeted human $\bar{f}_G^H$ values. While there could be some technical adjustments to CSUD itself, discussion on these will be relegated to chapter 12.

In this chapter, decision theoretic aspects are of interest, especially with respect to modelling human exploration or human learning heuristics. The most surprising result originating from this work has been the finding that human exploration in the IGT, as modelled by ε-Greedy and Boltzmann agents, is quite high. This finding of high exploration holds in ex-ante exploration hyper-parameter values as well as in ex-post exploration index (EI) results. Human reference exploration index (EI) values depicted in Fig. 4.3 and Fig. 4.4 corroborate high exploration findings.

For example, these figures indicate that for control (normal) subjects, the exploration index for the original and re-shuffled IGT environments at draw block 81-100 is around 80-85, and for the random IGT environment around 93. Fig. 10.9 shows that this implies that at the final draw block, approximately 62-70% of cards are drawn from the good decks by control IGT participants. For vmPFC impaired IGT participants, who fail the original IGT environment, at block 81-100, the exploration index is approximately at 97.5, indicating in combination with the vmPFC impaired original case in Fig. 4.1, that around 50% of the cards have been selected from the good decks. Hence, in the case of human controls, the IGT mean fraction of good deck $\bar{f}_G^H$ results do not indicate a very strong trend towards exploitation. vmPFC impaired human subjects, who do not learn well, appear to exhibit a trend focused on random search.

Such a high exploration result is verified by the EI measure, however, it cannot be explained by the Q-learning models presented here, as excepting the reverse IGT results, all Q-learning models produce lower than human exploration options where they can exceed corresponding human $\bar{f}_G^H$ performance benchmarks. Wilson et al., 2014 provide an argument for directed exploration and Findling et al., 2019 suggest learning noise as additional factors in high exploration behaviour. Also if humans did use directed exploration to learn more about the lesser known choices, and if rare events themselves are probabilistically undervalued under experience (Hertwig et al., 2004), then rare events such as in the SGT, decks B and D in the original IGT, decks E and H in the reversed IGT may all lead to increased exploration. Except the SGT, remaining IGT environments mix rare and frequent events, and along with the aggregated task-end cumulative measure of the mean fraction of good decks $\bar{f}_G$, this makes it very difficult to break down exploration into directed versus random components. It would be interesting to devise an iterative learning experience, subject to learning decay conditions, with more than two choice options, where exploration behaviour can be tested more accurately. Here it is suggested that each exploration model considered here cannot on its own adequately reflect human exploration behaviour. The question therefore remains as to how human exploration changes for example over successive IGT draws?

More research is needed to establish the nature of exploration in human-analogue machine decision making. For example, does choice exploration produce a focusing effect about the outcome central tendency such as $\bar{f}_G$?

Does exploration decay? If exploration decay were to be used as a decision making hyper-parameter, would this hyper-parameter be expected to have different behavioural implications regarding vmPFC impairment? Further, what if there is a 'boredom effect', or a directed effort (Wilson et al., 2014) increasing exploration, thereby counteracting actual exploration decay. Then for example the measured exploration decay based here on the cumulative end of task $\bar{f}_G$ may have been understated in this work. Fig. 4.4 shows that for normal SGT participants 20-draw block EI does indeed increase as the task progresses. It would be interesting to develop a model where such exploration responses can be differentiated.

It is possible that in the context of the 'No Free Lunch' theorems, humans use exploration as a defence against algorithm specificity, or to aid in generalised learning. Such an approach could explain why human exploration remains high. In general, many questions regarding human exploration, and the algorithmic modelling of human exploration remain open; more research is needed.

In contrast, learning rate, or step size, decay has been well studied, and (3.7) provides conditions typically required for ensuring that an iterative learning problem with decaying learning rate converges to an optimum. In this work however, agent modelling uses exponential learning rate decay, which does not fulfil such theoretical convergence criteria.

The burst learning model offers an alternative pathway for achieving theoretical convergence, whilst retaining the exponential decay induced learning freeze. There are two possible pathways that could be explored, but neither has been explored in this work. The first and harder pathway is to develop a theoretical convergence proof based on the cyclic nature of burst learning with learning rate decay presented in (10.2). The second and easier approach is to swap out the lower bound α_1/D in (10.2) with a corresponding $1/t$ dependent decay pattern as discussed in section 10.1.2. Both of these approaches provide interesting opportunities for exploring the connections between the rational and nonrational schools of thought.

One of the themes in this work has been time constraints in human decision making, which are known to exist but unknown as to when the constraints will become binding; consequently how to therefore effectively make decisions when resources such as time or opportunities are limited. Such constraints are embedded in the IGT. Here exponential learning rate decay has been used to induce finiteness. Is this really how humans deal with limited resources? This question is also waiting to be answered.

In engineering, infinity is one of the most useful concepts for producing smooth, convergent outcomes. Is the successful use of infinity in scientific endeavours a domain specific benefit, originating from the locally time invariant presentation of natural order? For example, a ray of light may be conceptualised as travelling in a straight line, unless of course it is being bent by a strong gravitational force. In conceptualising the natural world, with assumptions about the existence of a to be discovered *ground truth*, statistical assessment, in terms of a central tendency with asymptotic distributions, makes sense.

The problems and statistics of human endeavours, however, remain as yet quite unclear. For example, do individual human decision makers engage in any central tendency based optimisation, or is such optimisation only an artefact observed in aggregated behavioural data? Do individual human decision makers instead engage in threshold based targeting in relation to a cumulative density function, with fall-back thresholds in case of target misses? In such a case, humans would continually be setting goals, recovering, and re-setting goals as needed. Such a strategy could be helpful when time constraints and lack of information make it difficult or costly to formulate the corresponding rational decision making strategy.

The burst learning model can capture such re-targeting mechanics, as a learning rate reset is capable of leading to an over-write of the current solution. Similarly, burst learning should be able to navigate non-stationary environments subject to good targeting guidelines.

When do humans optimise, and when do they, in the words of Herbert A. Simon, "satisfice?" Even with generalised models such as reinforcement learning, much is yet to be understood about how to apply these models to human behaviour.

A few final interesting notes remain. On a technical note, burst learning could provide a computationally cheap alternative to Hessian, or Hessian approximation, driven loss searches. That is, the burst learning results presented here indicate that a heuristically varying (emotion mediated) learning rate (or step-size), may offer in comparison to Hessian methods, a relatively simple and rapid gradient approximation alternative. More research on burst learning methods is needed.

The result that high learning rate decay in the presented Q-learning models leads to vmPFC impairment poses further questions regarding the use of exponentially decaying learning rates. In chapter 10, CSUD search produces an attenuation parameter of about $D = 100$ for both agents. As noted in section 10, this indicates that it will take 9.2 iterations for the learning rate to diminish by 100-fold. Does this number of 9.2 choice iterations have relevance regarding the size of working memory? Or, does it in some way relate to the amount of trials needed to extract information from well-behaved ergodic processes? These provide good future directions as well.

In any case, it is hoped that the Q-learning modelling discussed in this work has provided useful computational techniques for the nonrational modelling of iterative choice

Chapter 12

Constrained Perturbations Stochastic Search

The CSUD search algorithm is based on simultaneous perturbations stochastic approximation (SPSA), initially proposed by Spall (1992), where double symmetric randomised perturbations are employed. SPSA variants have been developed with single-sided (Chen et al., 1999), point (Spall, 1997), repeated measurements (Abdulla & Bhatnagar, 2006), and non-stochastic perturbations (Bhatnagar et al., 2013). Further work on SPSA discusses update constraints (Spall, 2003, Ch. 7), asymptotic distributions (Hernández & Spall, 2019), and approximated Hessian driven variable learning rates (Zhu et al., 2020). CSUD itself is a probabilistic hybrid of double- and single-sided SPSA. In CSUD however, both updates and perturbations are constrained without invalidating the perturbation restrictions required for statistical convergence of SPSA.

The intended general purpose and use of SPSA algorithms is for stochastic optimisation. CSUD however is employed as a search strategy, which uses a loss function to tune and evaluate another hyper-parametrised model. The loss function provides a loss value, which can be used for ranking purposes. A search is not necessarily expected to produce a unique best result, which would however typically be expected in the stochastic optimisation.

CSUD loss is used to score the hyper-parametrisation of another model. Since such hyper-parameters may need to be bounded over a certain range, not only hyper-parameter updates, but also hyper-parameter perturbations are constrained. The search method proposed here is most generally called constrained perturbations SPSA, or **cp-SPSA** for short. In terms of nomenclature however, cp-SPSA is not specific enough. As noted above, SPSA has many variants. As discussed in Bhatnagar et al. (2013), it is helpful to denote a specific SPSA variant by its key innovation. In this context, the cp-SPSA algorithm specifically implemented here is called *Constrained Single Unconstrained Double* perturbations, or **CSUD** for short.

It is not always immediately apparent why a hyper-parameter search needs to be conducted in a bounded space. From a procedural perspective, bounded hyper-parameter regions can help divide the search space into smaller, manageable regions. Or, such bounds can highlight specific areas of interest. For example in chapters 6 to 9, bounds are used on normal and vmPFC impaired learning rate decay, λ_N and λ_{vmPFC} respectively to focus on ranges of interest. Finally, some hyper-parameter values may be illegal and may be restricted on that basis. For example, for ARA(κ) agents discussed in chapter 9, the discount rate γ may not be less than or equal to 0, or greater than or equal to 1. Since inclusive bound constraints are used in CSUD, as indicated in Table 9.1 the ARA(κ) ε-Greedy and Boltzmann agent discount rates are limited to inclusive ranges of $0.5 - 0.99$ and $0.15 - 0.85$ respectively.

Adding hyper-parameter constraints produces increased computational overhead, as at each iteration, hyper-parameter updates and perturbations need to be checked for any constraint violations, with any such violations then being mitigated. It is proposed that CSUD helps to assess interesting areas of parameter space, and while constraints introduce additional computational overhead, their use makes it easier to prototype and assess model specifications.

This chapter first introduces double- and single-sided SPSA. Then a standard proof of optimal convergence is provided for (unconstrained) single-sided SPSA under a stringent set of assumptions. This initial proof will set-up vocabulary and provide a basis for a similar optimal convergence proof for CSUD, where some of the initial assumptions will be relaxed. Finally, conditions will be provided for local optimal convergence for the CSUD search strategy, initially introduced in section 3.4.

12.1 The SPSA Framework

The simultaneous perturbations stochastic approximation (SPSA) framework is introduced. First the SPSA loss function and the input update equation is discussed. Next the gradient estimators are reviewed for double-sided

SPSA. This is then followed by the development of *unconstrained* single-sided SPSA.

Let Θ be a input vector of dimension p. Let $Y(\Theta)$ be a stochastic loss function, which can be expressed as

$$Y(\Theta) = L(\Theta) + \epsilon \tag{12.1}$$

where $Y(\cdot)$ is the observed loss, $L(\cdot)$ is the unknown loss function, and ϵ is the observation, or measurement error. Using the observed loss $Y(\cdot)$, the aim is to find a unique input Θ^* such that $L(\Theta)$ is minimised, assuming such a minimum exists. If such a minimum exists, it is known as the root of the function.

Standard gradient descent methods search for the function root by using analytical gradients. However, when the functional form of the observed or underlying loss is not known, analytical gradients cannot be derived. Therefore minimum root discovery cannot be conducted via analytical gradients. SPSA aims to solve the minimum root discovery problem by using a gradient descent approach, but with gradient approximations derived from randomly perturbed input vectors.

Section 12.2 below presents a simple set of assumptions, under which given observed stochastic loss $Y(\cdot)$, a unique minimum may be found. The key steps of the argument are introduced here. For the moment, assume that somehow the unknown loss gradient can be approximated. Let $\partial L(\Theta)/\partial\Theta \equiv g(\Theta)$ be the true but unobserved gradient of $L(\Theta)$. Therefore, one wishes to find Θ^* such that $g(\Theta^*) = 0$. Let $t \in \{0, 1, 2, \dots\}$ indicate an iteration counter. Let $\hat{g}_t(\Theta)$ denote the gradient approximation at iteration t, and let α_t denote step size (learning rate). Then the iterative input update rule is defined as

$$\hat{\Theta}_{t+1} = \hat{\Theta}_t - \alpha_t \hat{g}_t(\hat{\Theta}_t). \tag{12.2}$$

The premise of SPSA is that under certain conditions, as $t \to \infty$, $\hat{g}_t(\cdot) \to g_t(\cdot)$ and consequently $\hat{\Theta}_t \to \Theta^*$.

12.1.1 Unconstrained Double-Sided SPSA

Unconstrained double-sided SPSA is introduced and discussed in detail in Spall (1992, 2003). The major innovation of double-sided SPSA, over the finite differences gradient approximation method by Kiefer and Wolfowitz

(1952) is a reduction in the number of per-iteration loss measurements for gradient approximation. Given an input vector of dimension p, the finite differences method requires $2p$ measurements per iteration. In contrast, double-sided SPSA only needs to take 2 loss measurements per iteration; that is, the number of loss measurements are constant and do not depend on the size of the input vector. Furthermore, double-sided SPSA exhibits a comparatively good convergence rate along with a comparatively low gradient approximation bias, which vanishes asymptotically.

In unconstrained double-sided SPSA, the gradient approximation rule is defined as

$$\hat{g}_{ti}^{D}(\hat{\Theta}_t) = \frac{Y(\hat{\Theta}_t + \mu_t\Delta_t) - Y(\hat{\Theta}_t - \mu_t\Delta_t)}{2\mu_t\Delta_{ti}} \tag{12.3}$$

where $\hat{g}_{ti}^{D}(\hat{\Theta}_t)$ represents the double-sided gradient approximation at the t^{th} iteration for the i^{th} element of the input vector estimate $\hat{\Theta}_t$, μ_t is the perturbation step size, and Δ_t is a random perturbation vector, subject to the restrictions discussed in Spall (1992, 2003), and summarized below in Assumption 12.2.6. Finally Δ_{ti} denotes the perturbation corresponding to the i^{th} input vector element at iteration t.

12.2 Unconstrained Single-Sided SPSA

Spall (1998) discusses the notion of single-sided SPSA. Chen et al. (1999) present a single-sided SPSA variant with input estimate constraints. This section discusses unconstrained single-sided SPSA, where all input dimensions are varied simultaneously using single-sided perturbations. The resulting gradient approximation rule can be expressed as

$$\hat{g}_{ti}^{S}(\hat{\Theta}_t) = \frac{Y(\hat{\Theta}_t + \mu_t\Delta_t) - Y(\hat{\Theta}_t)}{\mu_t\Delta_{ti}} \tag{12.4}$$

where $\hat{g}_{ti}^{S}(\hat{\Theta}_t)$ represents the single-sided gradient approximation at the t^{th} iteration for the i^{th} element of the input vector estimate $\hat{\Theta}_t$, μ_t is the perturbation step size, and Δ_t is a random perturbation vector. Finally Δ_{ti} denotes the perturbation corresponding to the i^{th} input vector element at iteration t.

The bias and convergence properties of unconstrained single-sided SPSA are now presented. The proof below uses classical techniques, and is believed to be easier to follow than that presented in Chen et al. (1999). However, in order to achieve ease of exposition, strict assumptions are employed. The approach below also differs in notation. Where Taylor expansions are required, multi-index notation is used, along with results from Folland (1990, 2020), which results allow for reducing Taylor expansion continuity requirements to twice continuously differentiable.

It is believed that strict assumptions can be used at this stage without any loss of generality, and at the same time to facilitate introduction of notation and vocabulary, which will be employed later in sections 12.3 and 12.4 for analysis of CSUD and the CSUD search strategy respectively.

The assumptions required for asymptotic convergence of unconstrained single-sided SPSA are now introduced.

Assumption 12.2.1 (Loss Topology). $\Theta \in \mathbb{R}^p$. $L(\Theta) : \mathbb{R}^p \to \mathbb{R}$ *is strictly convex and at least of class* C^2 *(two times continuously differentiable). (A result from Folland, 1990 is used so that the approximation of the remainder of a second order Taylor expansion only requires existence of the second derivative.) Further, let* Θ^* *be the minimum, and* $B(\Theta^*, r)^p \subset \mathbb{R}^p$ *be an open ball of radius* r*, where* $0 < r < \infty$*. Then* $\forall\, \Theta \in B(\Theta^*, r)^p$, $L(\Theta)$ *is bounded, and has bounded derivatives. In general* $\forall\, \Theta \in \mathbb{R}^p$, $L(\Theta)$ *is bounded below at* $L(\Theta^*)$.

Assumption 12.2.1 states that the unknown loss function is well-behaved and has a unique minimum. If at some point inside the bounding ball, true loss became infinite, it would not be possible to calculate a numeric gradient. Therefore $L(\Theta)$ must be bounded, and posses bounded derivatives so that gradient approximations can be computed; so that gradient approximations converge to the true gradient.

Assumption 12.2.2 (Measurement Errors). $\epsilon \sim$ *i.i.d.*, $E|\epsilon| < \infty$, $E\epsilon^2 = \sigma_\epsilon^2 < \infty$.

That is, measurement errors are independently and identically distributed (i.i.d.); and have finite mean and variance. Note that 0 mean measurement error is not required. *i.i.d.* distributions are required however, as these make proof dynamics easier.

Assumption 12.2.3 (Iterate Dynamics). $\sup\limits_t \|\hat{\Theta}_t\| < \infty$ *almost surely.* $\hat{\Theta}_t \to \Theta^*$ *infinitely often.*

Assumption 12.2.3 is of theoretical value when conducting a generalised proof. It is a reminder of an obvious practical insight, namely that input estimates must themselves remain bounded in each iteration. If input estimates $\hat{\Theta}_t$ start exploding, then something must be wrong in the update equation (12.2). The second statement indicates that different input iteration paths, for example originating from different initial values, should converge to the same unique minimum.

Assumption 12.2.4 (Mean ODE Dynamics). *Let t denote time. Let $g(\Theta)$ be the gradient of $L(\Theta)$. By Assumption 12.2.1, $g(\Theta)$ is continuous. Let $Z(t)$ be a differentiable function. Then as $t \to \infty$, the differential equation $dZ(t)/dt = -g(Z(t))$ converges to a fixed point at Θ^*.*

Assumption 12.2.4 follows from the fact that at the minimum, the gradient is 0. Therefore if $Z(t)$ is used to search through $g(\cdot)$, then at the fixed point with $dZ(t)/dt = -g(Z(t)) = 0$, the minimum is achieved and the search stops. Note that this result is established with respect to the *unknown* loss function gradient.

Assumptions 12.2.1 to 12.2.4 establish the theoretical requirements under which SPSA, when seen as an optimiser, could discover the minimum of the *unknown* loss function. However, $L(\Theta)$ cannot be observed directly. Only $Y(\Theta)$ can be sampled, and therefore additional conditions are needed to ensure that any measurement error or stochasticity does not countermand true loss topology. Additionally over iterations and input updates, it must also be possible to approximate the true gradient. The remaining assumptions achieve this result.

Assumption 12.2.5 (Step Sizes, α_t and μ_t). *$\forall$ iterations t, the input update step size (learning rate) $\alpha_t > 0$, the perturbation step $\mu_t > 0$. Further, $\lim_{t\to\infty} \alpha_t = 0$, $\lim_{t\to\infty} \mu_t = 0$, $\sum_{t=0}^{\infty} \alpha_t = \infty$, but $\sum_{t=0}^{\infty} \alpha_t^2/\mu_t^2 < \infty$. The update-step, perturbation-step ratio boundedness condition is critical for achieving asymptotic behaviour.*

Assumption 12.2.5 embodies the most important aspect of SPSA learning, where learning refers to the correct discovery of the unique minimum Θ^*. With each input update iteration (12.2), the learning rate α_t captures what is being learned from the current gradient estimate. As the input iterate $\hat{\Theta}_t$ approaches the minimum Θ^*, ideally input updates as well as input perturbations must be diminished, as this would reduce oscillations around the minimum.

An approximation of the unknown loss function Hessian would provide an input update step with the best information, and a similar technique could be used for scaling perturbation steps. However such computations are costly. Assumption 12.2.5 indirectly implies decay limits for update and perturbation step sizes, α_t and μ_t respectively. The purpose of these decay limits is to induce discovery of the minimum as update iterations go to infinity.

Assumption 12.2.5 cannot induce by itself discovery of the unknown loss function minimum. The manner in which inputs are perturbed is critical to SPSA success.

Assumption 12.2.6 (Perturbations)**.** *Let Δ_t be a p-dimensional perturbation vector at iteration t. Then $\forall$ t and $\forall$ $i \in p$, $\Delta_{ti} \sim$ i.i.d. and symmetric about* 0*. Further $|\Delta_{ti}| < \infty$, $E(\Delta_{ti}) = 0$, $E(\Delta_{ti}^2) < \infty$, $E(1/\Delta_{ti}) < \infty$, and $E(1/\Delta_{ti}^2) < \infty$. Consequently, note that $Prob(\Delta_{ti} = 0) = 0$ (Spall, 1992).*

Assumption 12.2.6 presents some unusual implications. Most importantly, while perturbation mean must be 0, the perturbation probability distribution must have *no mass* at the mean. Therefore the perturbations *must not* come from uni-modal distributions. Spall (2003, Ch. 7) suggests unit Bernoulli, segmented uniform, or U-shaped distributions.

Finally, restrictions must be specified on the interactions between observed loss $Y(\cdot)$ and perturbations Δ.

Assumption 12.2.7 (Stochastic Interactions)**.** *Define input history $\mathscr{I}_t \equiv \{\hat{\Theta}_0, \ldots, \hat{\Theta}_t\}$, errors $\epsilon_t^+ \equiv \epsilon(\hat{\Theta}_t + \mu_t\Delta_t)$ and $\epsilon_t \equiv \epsilon(\hat{\Theta}_t)$. Assume $E_{\mathscr{I}_t,\Delta_t}(\epsilon_t^+ - \epsilon_t) =_{a.s.}$ 0, where $E_{\mathscr{I}_t,\Delta_t}(\cdot)$ is the conditional expectation given input history and the current perturbation; that is $E_{\mathscr{I}_t,\Delta_t}(\cdot) \equiv E(\cdot \mid \mathscr{I}_t, \Delta_t)$. Perturbation Δ_t is independent of $\mathscr{I}_t$ and therefore, $E_{\mathscr{I}_t}(\epsilon_t^+ - \epsilon_t) =_{a.s.}$ 0. Further when $Y_t(\cdot) = L_t(\cdot) + \epsilon_t$, then*

$$\forall t, i: \; E\left[\left(\frac{Y_t(\cdot)}{\Delta_{ti}}\right)^2\right] = E\left[\frac{1}{\Delta_{ti}^2}\right] E\left[Y_t(\cdot)^2\right] < \infty \tag{12.5a}$$

$$\begin{aligned} \forall t, i \neq h: \; & E\left[\frac{Y_t(\cdot)^2}{\Delta_{ti}\Delta_{th}}\right] \\ &= E\left[\frac{1}{\Delta_{ti}}\right] E\left[\frac{1}{\Delta_{th}}\right] E\left[Y_t(\cdot)^2\right] < \infty. \end{aligned} \tag{12.5b}$$

(12.5a) *and* (12.5b) *ensure that the estimator $\hat{g}_t^S(\hat{\theta}_{ti})$ in* (12.4) *has a finite 2^{nd} moment.*

Assumption 12.2.7 essentially states that using perturbed inputs to generate observed loss does not add any systematic bias into the observed results, and additionally does not cause probabilistic instability in the sense of unbounded second moments.

It should be noted that perturbation step sizes μ_t and the distribution of perturbations Δ_t are determined by the analyst, and can be constructed to satisfy all assumptions. In cases where SPSA is applied to an unknown loss function however, it is generally not possible to verify any assumptions relating to $Y(\cdot)$ and $L(\cdot)$. In such cases, the quality of the SPSA results should be assessed using empirical output, with output evidence being checked for violations of assumptions.

12.2.1 Taylor Expansions

Taylor expansions play a key role in developing bias and convergence properties. Using multi-index notation, relevant Taylor expansions are defined using Folland (1990, 2020), and then briefly discussed.

For a vector of dimension p, let β be a multi-index defined as

$$\begin{gathered} \beta = (\beta_1, \beta_2, \dots, \beta_p), \quad \beta_i \in \mathbb{N}_0^p \\ |\beta| = \sum_{i=1}^{p} \beta_i, \quad \beta! = \prod_{i=1}^{p} \beta_i! \end{gathered} \tag{12.6}$$

where $\mathbb{N}_0^p$ is the p dimensional set of natural numbers including 0. Further, for any vector $\mathbf{z} \in \mathbb{R}^p$, $\mathbf{z} = (z_1, \dots, z_p)$, define $\mathbf{z}^\beta = (z_1^{\beta_1}, \dots, z_p^{\beta_p})$. Then, for any $\Theta \in \mathbb{R}^p$, partial derivatives may be expressed as

$$\partial^\beta Y(\Theta) = \partial_1^{\beta_1} \dots \partial_p^{\beta_p} = \frac{\partial^{|\beta|} Y(\Theta)}{\partial\theta_1^{\beta_1} \dots \partial\theta_p^{\beta_p}} \,. \tag{12.7}$$

For example, let $p = 3$ and $\beta = (1, 1, 0)$, then

$$\partial^{(1,1,0)} Y(\Theta) = \frac{\partial^2 Y(\Theta)}{\partial\theta_1 \partial\theta_2} \,. \tag{12.8}$$

For any $\Theta \in \mathbb{R}^p$, using multi-index notation and assuming $L(\cdot)$ is at least of class C^2, the below Taylor expansions can be defined

$$\begin{aligned} L(\Theta_t + \mu_t \Delta_t) &= \sum_{|\beta| \le 2} \mu_t^{\beta} \frac{\partial^{\beta} L(\Theta_t)}{\beta!} \Delta_t^{\beta} + R_2(\mu_t \Delta_t) \\ &= L(\Theta_t) + \mu_t \sum_{|\beta|=1} \partial^{\beta} L(\Theta_t) \Delta_t^{\beta} + \frac{1}{2} \mu_t^2 \sum_{|\beta|=2} \partial^{\beta} L(\Theta_t) \Delta_t^{\beta} \\ &+ R_2(\Theta_t, \mu_t \Delta_t) \end{aligned} \tag{12.9a}$$

$$\begin{aligned} L(\Theta_t - \mu_t \Delta_t) &= \sum_{|\beta| \le 2} (-\mu_t)^{\beta} \frac{\partial^{\beta} L(\Theta_t)}{\beta!} \Delta_t^{\beta} + R_2(-\mu_t \Delta_t) \\ &= L(\Theta_t) - \mu_t \sum_{|\beta|=1} \partial^{\beta} L(\Theta_t) \Delta_t^{\beta} + \frac{1}{2} \mu_t^2 \sum_{|\beta|=2} \partial^{\beta} L(\Theta_t) \Delta_t^{\beta} \\ &+ R_2(\Theta_t, -\mu_t \Delta_t) \end{aligned} \tag{12.9b}$$

where the $R(\cdot)$ denotes the remainder, $|\beta| = p$ is an indexing short-hand, and p is the dimension of the input vector. For example, with two inputs (i.e. $p = 2$), $|\beta| = 2$ expands to $\beta \in \{(2,0), (0,2), (1,1)\}$.

From Folland (1990, p. 235), the remainder term in multi-index notation is expanded as,

$$\begin{aligned} &R_2(\Theta_t, \mathbf{h}) = \\ &\sum_{|\beta|=2} \mathbf{h}^{\beta} \int_0^1 (1 - \nu) \left[\partial^{\beta} L(\Theta_t + \nu \mathbf{h}) - \partial^{\beta} L(\Theta_t) \right] d\nu \end{aligned} \tag{12.10}$$

where $\nu \in (0,1)$, $\mathbf{h} = \pm \mu_t \Delta_t$. Further, from Folland (1990), it is established that

$$|R_2(\Theta_t, \mathbf{h})| \le \frac{\mathbb{M}}{6} \mu_t^3 \left(\sum_i |\Delta_{ti}| \right)^3 \tag{12.11}$$

where $0 < \mathbb{M} < \infty$.

Note that if $L(\cdot)$ is quadratic additive, then the remainder $R_2(\Theta_t, \mathbf{h})$ in (12.10) becomes 0, and consequently the 2nd order Taylor expansions in (12.9a) and (12.9b) become exact. In non-quadratic cases, by Assumption 12.2.5, $\mu_t \to 0$ and by Assumption 12.2.6, $|\Delta_{ti}| < \infty$. Consequently, the remainder in equation (12.11) vanishes as $t \to \infty$.

12.2.2 The Bias of the Gradient Estimate

By definition $\hat{g}_t^S(\cdot)$ in (12.4) is a stochastic function estimate. Given the input vector $\hat{\Theta}_t$ at iteration t, it is necessary to show that for the i^{th} element $\hat{\theta}_{ti}$, the bias

$$\lim_{t\to\infty} \mathscr{B}_{ti}^{\hat{g}_t^S} = \lim_{t\to\infty} E_{\mathscr{T}_t}\left[\hat{g}_{ti}^S(\hat{\Theta}_t) - g_i(\hat{\Theta}_t)\right] = 0. \tag{12.12}$$

That is, asymptotically speaking, the gradient approximation approaches the true gradient.

Using the definition of the loss function in (12.1) together with the single-sided gradient approximation (12.4), and taking expectations yields

$$\begin{aligned}
&E_{\mathscr{T}_t}\left[\hat{g}_{ti}^S(\hat{\Theta}_t)\right] \\
&= E_{\mathscr{T}_t}\left[\frac{L(\hat{\Theta}_t + \mu_t\Delta_t) - L(\hat{\Theta}_t)}{\mu_t\Delta_{ti}} + \frac{(\epsilon^+ - \epsilon)}{\mu_t\Delta_{ti}}\right]
\end{aligned} \tag{12.13a}$$

$$\begin{aligned}
&= E_{\mathscr{T}_t}\left[\frac{\mu_t \sum_{|\beta|=1} \partial^\beta L(\hat{\Theta}_t)\Delta_t^\beta}{\mu_t\Delta_{ti}}\right] + \\
&E_{\mathscr{T}_t}\left[\frac{\frac{1}{2}\mu_t^2 \sum_{|\beta|=2} \partial^\beta L(\hat{\Theta}_t)\Delta_t^\beta + R_2(\hat{\Theta}_t, \mu_t\Delta)}{\mu_t\Delta_{ti}}\right]
\end{aligned} \tag{12.13b}$$

where Assumption 12.2.7 and the Taylor expansion in (12.9a) have been used.

Expanding the multi-index terms in (12.13b), simplifying and re-arranging

$$\begin{aligned}
E_{\mathscr{T}_t}\left[\hat{g}_t^S(\theta_{ti})\right] = E_{\mathscr{T}_t}&\left[\frac{\partial L}{\partial\theta_{ti}} + \frac{1}{\Delta_{ti}}\sum_{j\neq i}\frac{\partial L}{\partial\theta_{tj}}\Delta_{tj}\right] + \\
&\mu_t E_{\mathscr{T}_t}\left[\frac{1}{2}\frac{\partial^2 L}{\partial\theta_{ti}^2}\Delta_{ti} + \sum_{j\neq i}\frac{\partial^2 L}{\partial\theta_{ti}\partial\theta_{tj}}\Delta_{tj} + \right. \\
&\left.\frac{1}{2\Delta_{ti}}\sum_{j\neq i}\frac{\partial^2 L}{\partial\theta_{tj}^2}\Delta_{tj}^2 + \frac{1}{\Delta_{ti}}\sum_{j\neq i}\sum_{h\neq i}\frac{\partial^2 L}{\partial\theta_{tj}\partial\theta_{th}}\Delta_{tj}\Delta_{th}\right] \\
&+ E_{\mathscr{T}_t}\left[\frac{R_2(\hat{\Theta}_t, \mu_t\Delta_t)}{\mu_t\Delta_{ti}}\right].
\end{aligned} \tag{12.14}$$

Applying expectations to the perturbation terms, expanding $\partial L/\partial\theta_{ti}$ as $g_i(\hat{\Theta}_t)$ further simplifying, and re-arranging, the bias at iteration t is found as

$$\mathscr{B}_{ti}^{\hat{g}_t^S} = E_{\mathscr{F}_t}\left[\hat{g}_{ti}^S(\hat{\Theta}_t) - g_i(\hat{\Theta}_t)\right] = \frac{\mu_t}{2} E_{\mathscr{F}_t}\left[\frac{1}{\Delta_{ti}}\right] \sum_{j\neq i} \frac{\partial^2 L}{\partial\theta_{tj}^2} E_{\mathscr{F}_t}\left[\Delta_{tj}^2\right] + E_{\mathscr{F}_t}\left[\frac{R_2(\hat{\Theta}_t, \mu_t\Delta_t)}{\mu_t\Delta_{ti}}\right]. \tag{12.15}$$

As (12.15) shows, the single-sided SPSA gradient estimate is biased. Computing the specific magnitude of the bias for iteration t requires knowledge of the (unknown) loss function $L(\cdot)$, the perturbation step size μ_t, and the distribution of Δ_t.

However, with Assumptions 12.2.1 and 12.2.6, which invoke boundedness of 2nd order loss derivatives, and existence of 2nd order perturbation moments respectively, a further result may be derived

$$\lim_{t\to\infty} \mathscr{B}_{ti}^{\hat{g}_t^S} = 0 \tag{12.16}$$

for all θ_{ti}.

Using Assumptions 12.2.1 and 12.2.6, for some finite bounds d, $d_{[-1]}$, $d_{[2]}$ $l_{[2]} > 0$, it is established that

$$|\Delta_i| \leq d, \quad \left|E_{\mathscr{F}_t}\left[\frac{1}{\Delta_{ti}}\right]\right| \leq d_{[-1]}, \quad \left|E_{\mathscr{F}_t}\left[\Delta_{ti}^2\right]\right| \leq d_{[2]}, \text{ and } \quad \left|\frac{\partial^2 L}{\partial\Theta_t^2}\right| \leq l_{[2]}. \tag{12.17}$$

Then the first term in (12.15) is bounded by

$$-\mu_t\frac{\mathbb{B}_1}{2} \leq \frac{\mu_t}{2} E_{\mathscr{F}_t}\left[\frac{1}{\Delta_{ti}}\right] \sum_{j\neq i} \frac{\partial^2 L}{\partial\theta_{tj}^2} E_{\mathscr{F}_t}\left[\Delta_{tj}^2\right] \leq \mu_t\frac{\mathbb{B}_1}{2} \tag{12.18}$$

where $\mathbb{B}_1 = (p-1)d_{[-1]}l_{[2]}d_{[2]}$. Therefore by (12.18), as $\mu_t \to 0$, the first term in (12.15) vanishes to 0.

SPSA Method	**Bias of i$^{\text{th}}$ gradient element at k$^{\text{th}}$ iteration**	**Asymptotic Bias**
Double-Sided	$\frac{1}{\mu_t} E_{\mathscr{T}_t}\left[\frac{1}{\Delta_{ti}}\right] E_{\mathscr{T}_t}\left[R_2(\hat{\Theta}_t, \mu_t \Delta_t) - R_2(\hat{\Theta}_t, -\mu_t \Delta_t)\right]$	0
Single-Sided	$E_{\mathscr{T}_t}\left[\frac{1}{\Delta_{ti}}\right] \left(\frac{\mu_t}{2} \sum_{j \neq i} \frac{\partial^2 L}{\partial \theta_{tj}^2} E_{\mathscr{T}_t}\left[\Delta_{tj}^2\right] + \frac{1}{\mu_t} E_{\mathscr{T}_t}\left[R_2(\hat{\Theta}_t, \mu_t \Delta_t)\right]\right)$	0

TABLE 12.1: Gradient bias by gradient approximation method

To show convergence to 0 of the remainder term in (12.15), the bound in (12.11) is used to establish

$$\begin{aligned}
-\mu_t^2 \frac{\mathbb{M}\mathbb{B}_2}{6} \quad &\leq -\mu_t^2 \frac{\mathbb{M}}{6} \frac{\left(\sum_i |\Delta_{ti}|\right)^3}{\Delta_{ti}} \leq \frac{R_2(\hat{\Theta}_t, \mu_t \Delta_t)}{\mu_t \Delta_{ti}} \\
&\leq \mu_t^2 \frac{\mathbb{M}}{6} \frac{\left(\sum_i |\Delta_{ti}|\right)^3}{\Delta_{ti}} \leq \quad \mu_t^2 \frac{\mathbb{M}\mathbb{B}_2}{6}
\end{aligned} \tag{12.19a}$$

$$\left| \left(E_{\mathscr{T}_t}\left[\frac{R_2(\hat{\Theta}_t, \mu_t \Delta_t)}{\mu_t \Delta_{ti}}\right] \right)_{-\mu_t^2 \frac{\mathbb{M}\mathbb{B}_2}{6}}^{\mu_t^2 \frac{\mathbb{M}\mathbb{B}_2}{6}} \right| \leq \mu_t^2 \frac{\mathbb{M}\mathbb{B}_2}{3} \tag{12.19b}$$

where $\mathbb{B}_2 = p^3 d^2$ and $0 < \mathbb{M} < \infty$. (12.19a) states that all occurrences of the remainder term must be bounded. Therefore as indicated by (12.19b), the expectation of the remainder term is also bounded. If $\mu_t \to 0$, then $\mu_t^2 \frac{\mathbb{M}\mathbb{B}_2}{3} \to 0$, and the remainder term in (12.15), that is, the left term in (12.19b), is enveloped to 0.

In sum, the enveloping results in (12.18) and (12.19b) show that asymptotically, $\hat{g}_t^S(\hat{\Theta}_t)$ is an unbiased estimator of $g(\hat{\Theta}_t)$, in the sense that $\forall\ i$,

$\mathscr{B}_{ti}^{\hat{g}_t^S} = E_{\mathscr{F}_t}\left[\hat{g}_{ti}^S(\hat{\Theta}_t) - g_i(\hat{\Theta}_t)\right] \rightarrow 0$. Further, the bounds on the expanded remainder term in (12.19a) indicate that as with double-sided SPSA (Spall, 1992, 1997), in single-sided SPSA, $\mathscr{B}_{ti}^{\hat{g}_t^S} = O(\mu_t^2)$.

Table 12.1 presents the bias properties of the double-sided and single-sided estimators of the gradient $g(\cdot)$. It is seen that SPSA gradient estimators can be asymptotically unbiased. However, at any iteration t, the double-sided gradient has a smaller bias than the single-sided SPSA gradient estimates.

12.2.3 Almost Sure Convergence of the Parameter Estimate

This section shows that $\lim\limits_{t\to\infty} \hat{\Theta}_t =_{a.s.} \Theta^*$ (almost surely convergence).

Chen et al. (1999) present proofs for the asymptotic properties of single-sided SPSA with relatively relaxed assumptions. For example, Chen et al. (1999) reduce the loss function triple differentiability requirement found in Spall (1992) to only twice continuous differentiability, as is also done here via Folland (1990).

However, Chen et al. (1999) trade-off such relaxations for additional proof complexity. Here, the general approach discussed among others in Ljung (1978) is followed, and a simpler proof is presented, where notation is used, which closely mirrors computational quantities.

To start, equation (12.2) is rewritten as

$$\begin{aligned}
\hat{\Theta}_{t+1} &= \hat{\Theta}_t - \alpha_t \hat{g}_t^S(\hat{\Theta}_t) && \text{(12.20a)}\\
&= \hat{\Theta}_t - \alpha_t \Big(\hat{g}_t^S(\hat{\Theta}_t) + g(\hat{\Theta}_t) - g(\hat{\Theta}_t)\\
&\qquad + E_{\mathscr{F}_t}\left[\hat{g}_t^S(\hat{\Theta}_t)\right] - E_{\mathscr{F}_t}\left[\hat{g}_t^S(\hat{\Theta}_t)\right]\Big) && \text{(12.20b)}\\
&= \hat{\Theta}_t - \alpha_t g(\hat{\Theta}_t) - \alpha_t \mathscr{B}_t^{\hat{g}_t^S} - \alpha_t \xi_t && \text{(12.20c)}
\end{aligned}$$

where the bias and error terms are defined as

$$\begin{aligned}
&\mathscr{B}_t^{\hat{g}_t^S} = E_{\mathscr{F}_t}\left[\hat{g}_t^S(\hat{\Theta}_t) - g(\hat{\Theta}_t)\right] = E_{\mathscr{F}_t}\left[\hat{g}_t^S(\hat{\Theta}_t)\right] - g(\hat{\Theta}_t),\\
&\xi_t = \hat{g}_t^S(\hat{\Theta}_t) - E_{\mathscr{F}_t}\left[\hat{g}_t^S(\hat{\Theta}_t)\right].
\end{aligned} \qquad \text{(12.21)}$$

Intuitively, one can see that a steady state solution to (12.20c) exists at Θ^* where $g(\Theta^*) = 0$. Provided the bias and error terms are bounded and tend to 0, given Assumptions 12.2.3 and 12.2.4, the iteration in (12.20c) will converge to Θ^*. The behaviour of the bias and error terms in (12.20c) are now formalised.

Proposition 12.2.1 (Convergence)**.** *Given Assumptions 12.2.1 to 12.2.7 and the iteration rule* (12.20c)*, as $t \to \infty$, $\hat{\Theta}_t \to \Theta^*$ a.s.*

Proof. Consider the N period forward-shifted representation of (12.20c)

$$\hat{\Theta}_{t+N} - \hat{\Theta}_t = -\sum_{j=0}^{N-1} \alpha_{t+j} g(\hat{\Theta}_{t+j}) - \sum_{j=0}^{N-1} \alpha_{t+j} \mathscr{B}_{t+j}^{\hat{g}^S_{t+j}} - \sum_{j=0}^{N-1} \alpha_{t+j} \xi_{t+j} \tag{12.22}$$

where $N - 1 \geq t$. It is necessary to show that $\lim\limits_{t\to\infty} \hat{\Theta}_{t+N} - \hat{\Theta}_t =_{a.s.} 0$. In the below discussion, N is a shorthand for $N > t$.

The Bias Term:

From Assumption 12.2.5, (12.19a) (boundedness) and (12.19b) (a.s. convergence), it follows that

$$\lim_{t\to\infty} \sum_{j=0}^{N-1} \alpha_{t+j} \mathscr{B}_{t+j}^{\hat{g}^S_{t+j}} = \lim_{t\to\infty} \alpha_t \mathscr{B}_t^{\hat{g}^S_t} + \cdots + \lim_{t\to\infty} \alpha_{t+N-1} \mathscr{B}_{t+N-1}^{\hat{g}^S_{t+N-1}} =_{a.s.} 0 \tag{12.23}$$

The Error Term:

From Assumptions 12.2.6 and 12.2.7, with history $\{\mathscr{T}_t\}$ given, $\{\alpha_t \xi_t\}$ forms an independent, mean zero sequence. Define $\mathscr{M}_t = \sum_{j=0}^{t-1} \alpha_j \xi_j$, and note that $E_{\mathscr{T}_t}[\mathscr{M}_{t+1}] = \mathscr{M}_t$. Therefore $\{\mathscr{M}_t\}$ forms a martingale.

For any $\mathbf{v} \in \mathbb{R}^p$, let $\|\mathbf{v}\|$ denote the Euclidean norm, $\|\mathbf{v}\| \equiv \sqrt{\sum_i v_i^2}$.

Following Yin and Kushner (2003, Ch. 5, Theorem 2.1), Doob's martingale inequality is invoked. For any $\eta > 0$, $\eta \in \mathbb{R}$

$$Prob\left(\sup_{N \geq j \geq t} \|\mathscr{M}_j - \mathscr{M}_t\| \geq \eta\right) \leq \frac{1}{\eta^2} E\left[\|\mathscr{M}_N - \mathscr{M}_t\|^2\right]. \tag{12.24}$$

Since $\mathscr{M}_N - \mathscr{M}_t = \sum_{j=0}^{N-1} \alpha_{t+j}\xi_{t+j}$, the right hand side of (12.24) can be written as

$$\frac{1}{\eta^2} E\left[\left\|\sum_{j=0}^{N-1} \alpha_{t+j}\xi_{t+j}\right\|^2\right] =$$

$$\frac{1}{\eta^2} E\left[\sum_{i=1}^{p}\left(\sum_{j=0}^{N-1} \alpha_{t+j}\xi_{(t+j)i}\right)^2\right] = \frac{1}{\eta^2} E\left[\sum_{i=1}^{p} e_i' A e_i\right] \tag{12.25a}$$

$$= \frac{1}{\eta^2}\sum_{i=1}^{p} tr\left(AE\left[e_i e_i'\right]\right) = \frac{1}{\eta^2}\sum_{i=1}^{p}\sum_{j=0}^{N-1} \alpha_{t+j}^2 E\xi_{(t+j)i}^2$$

$$\leq \frac{1}{\eta^2}\sum_{i=1}^{p}\sum_{j=t}^{\infty}\alpha_j^2 E\xi_{ji}^2 \leq \frac{1}{\eta^2}\sum_{i=1}^{p}\sum_{j=t}^{\infty}\alpha_j^2 E\left[\left(\hat{g}_{ji}^S(\hat{\Theta}_j)\right)^2\right] \tag{12.25b}$$

$$\leq \frac{p\left(\mathbb{L}+2\mathbb{E}\right)\mathbb{D}}{\eta^2}\sum_{j=t}^{\infty}\frac{\alpha_j^2}{\mu_j^2} \tag{12.25c}$$

where for notational ease

$$a = \begin{bmatrix} \alpha_t \\ \vdots \\ \alpha_{(N-1)} \end{bmatrix}, \ A = aa',$$

$$e_i = \begin{bmatrix} \xi_{ti} \\ \vdots \\ \xi_{(N-1)i} \end{bmatrix}, \ E[e_i e_i'] = \begin{bmatrix} E\xi_{ti}^2 & & \\ & \ddots & \\ & & E\xi_{(N-1)i}^2 \end{bmatrix} \tag{12.26}$$

and, the following properties have been used

$$\begin{aligned} E\left[\xi_{ji}\xi_{j'i}\right] &= E\left[E_{\mathscr{T}_{j'}}\left[\xi_{ji}\xi_{j'i}\right]\right] \\ &= E\left[\xi_{ji}E_{\mathscr{T}_{j'}}\left[\xi_{j'i}\right]\right] = 0, \quad t \leq j < j' \leq N-1 \end{aligned} \tag{12.27a}$$

$$\begin{aligned} E\xi_{ji}^2 &= E\left[\hat{g}_{ji}^S(\hat{\Theta}_t)^2\right] - E\left[\left(E_{\mathscr{T}_j}\left[\hat{g}_{ji}^S(\hat{\Theta}_t)\right]\right)^2\right] \\ &\leq E\left[\hat{g}_{ji}^S(\hat{\Theta}_t)^2\right] \end{aligned} \tag{12.27b}$$

$$\begin{aligned}
&E\left[\hat{g}^S_{ji}(\hat{\Theta}_t)^2\right] \\
&\quad = E\left[\left(\frac{L(\hat{\Theta}_j + \mu_j\Delta_j) - L(\hat{\Theta}_j) + \epsilon^+_j - \epsilon_j}{\mu_j\Delta_{ji}}\right)^2\right] \\
&\quad \leq \frac{1}{\mu_j^2}\left(\mathbb{L} + 2\mathbb{E}\right)\mathbb{D}
\end{aligned} \tag{12.27c}$$

$$\begin{aligned}
&E\left[\left(L(\hat{\Theta}_j + \mu_j\Delta_j) - L(\hat{\Theta}_j)\right)^2\right] \leq \mathbb{L} \\
&E\epsilon^2 \leq \mathbb{E}, \quad E\left[\left(\frac{1}{\Delta_{ji}}\right)^2\right] \leq \mathbb{D}
\end{aligned} \tag{12.27d}$$

where $0 < \mathbb{D}, \mathbb{E}, \mathbb{L} < \infty$ are bounds. Regarding $\mathbb{L}$, from Assumption 12.2.6 the perturbations are bounded. Assumption 12.2.1 implies that for any arbitrary Θ and $r > 0$, the loss function $L(\Theta)$ is bounded on $B(\Theta, r)$. Therefore on $B(\Theta, r)$, all events of the squared loss term in (12.27d) are always bounded, and consequently, its expectation is also bounded.

Using equations (12.24) and (12.25a) to (12.25c)

$$\begin{aligned}
&\sum_{t=0}^{\infty} Prob\left(\sup_{N \geq j \geq t} \|\mathscr{M}_j - \mathscr{M}_t\| \geq \eta\right) \\
&\qquad\qquad \leq \frac{p\left(\mathbb{L} + 2\mathbb{E}\right)\mathbb{D}}{\eta^2}\sum_{t=0}^{\infty}\sum_{j=t}^{\infty}\frac{\alpha_j^2}{\mu_j^2} \quad < \quad \infty
\end{aligned} \tag{12.28}$$

since from Assumption 12.2.5, as $t \to \infty$, $\sum_{j=t}^{\infty} \alpha_j^2/\mu_j^2 \to 0$.

Given the result in (12.28), the Borel-Cantelli Lemma (Durrett, 2019, Ch. 5, Theorem 2.3.1) states that $Prob(\sup_{N \geq j \geq t} \|\mathscr{M}_j - \mathscr{M}_t\| \geq \eta \; i.o.) = 0$, where $i.o$ denotes infinitely often. Consequently as $t \to \infty$, $\sup_{N \geq j \geq t} \|\mathscr{M}_j - \mathscr{M}_t\| \to_{a.s} 0$. That is, the error term converges to 0 almost surely.

The Mean ODE term:

The convergence of the mean ODE term $\sum_{j=0}^{N-1} \alpha_{t+j} g(\hat{\Theta}_{t+j})$ requires some additional conditions on N. Yin and Kushner (2003), Ljung (1978), and Metivier and Priouret (1984) all provide mean ODE convergence presentations, which work well in the present context. The full arguments are not repeated here. In outline, looking at the first part of (12.22), one can see that the steady state for $\hat{\Theta} = \hat{\Theta} - \sum_{j=0}^{N-1} \alpha_{t+j} g(\hat{\Theta}_{t+j})$ can only be reached if for some time indices $t^* > t$, $g(\hat{\Theta}_{t^*}) = 0$. By definition, this can only happen at $\hat{\Theta} = \Theta^*$ since $g(\Theta^*) = 0$. Given any initial condition $\hat{\Theta}_0 \neq \Theta^*$, $g(\cdot)$ provides corrective direction and magnitude updates, $\sum_{t=0}^{\infty} \alpha_t = \infty$ ensures continuing updates, and $\alpha_t \to 0$ leads to convergence.

□

12.3 CSUD

Theoretical conditions are now presented under which CSUD converges to the roots of a loss function. Given the strictly convex loss function assumption employed in section 12.2, in principle, constraints are not necessary. Even if the input iterate increases temporarily, eventually it must converge.

In practice, however, it becomes desirable to guard against large input iterate deviations, or to restrict input value ranges. For example, computational numerical range is limited, and underflow or overflow may occur. Furthermore, when measurement error variance overwhelms perturbation step size μ_t, this can lead to incorrect gradient approximations, which can cause divergence. Therefore, in practice, it is common to apply truncation constraints on the input updates.

Constraining input updates is sufficient to address the issues discussed above. However, there are also circumstances, where it is necessary to apply constraints to mask out potentially illegal input values, for example when an input value must be between 0 and 1. Finally constraining both input updates and perturbations creates well-defined partitions of input space. Such well-defined partitions can help focus on a region of interest, or allow one to reduce the globally strictly convex loss function requirement to a loss function with local strict convexity in the constrained partition.

The CSUD input update and gradient approximation rules are presented next. The CSUD input update rule is specified as

$$\hat{\Theta}_{t+1} = \hat{\Theta}_t - \alpha_t \hat{g}_t^C(\hat{\Theta}_t) - \alpha_t Z_{tk} \tag{12.29}$$

where Z_{tk} is a corrective term (Yin & Kushner, 2003, Ch. 5), which ensures that $\hat{\Theta}_{t+1} \in \mathbb{Z}_k$, with $\mathbb{Z}_k$ denoting the constrained input space. That is, (12.29) contains a projection constraint which is applied to input updates. The subscript k on $\mathbb{Z}_k$ indicates that multiple constraint sets, subject to terms discussed after Assumption 12.3.1, may be present. In other words, it is possible to run CSUD in multiple partitioned constraint sets $\mathbb{Z}_k$ of a larger input space.

Using indicator functions, the gradient estimator for the i^{th} element of $\hat{\Theta}_t$ is defined as

$$\hat{g}^{C}_{ti}(\hat{\Theta}_t) = \mathbf{1}^{S+}_{ti}\hat{g}^{S+}_{ti}(\hat{\Theta}_t) + \mathbf{1}^{S-}_{ti}\hat{g}^{S-}_{ti}(\hat{\Theta}_t) + \mathbf{1}^{D}_{ti}\hat{g}^{D}_{ti}(\hat{\Theta}_t) \tag{12.30}$$

where $\hat{g}^{D}_{ti}(\hat{\Theta}_t)$ is the gradient approximation for double-sided SPSA defined in (12.3), and

$$\hat{g}^{S+}_{ti}(\hat{\Theta}_t) = \left[\frac{Y(\hat{\Theta}_t + \mu_t\Delta_t) - Y(\hat{\Theta}_t)}{\mu_t\Delta_{ti}}\right],$$

$$\hat{g}^{S-}_{ti}(\hat{\Theta}_t) = \left[\frac{Y(\hat{\Theta}_t) - Y(\hat{\Theta}_t - \mu_t\Delta_t)}{\mu_t\Delta_{ti}}\right], \tag{12.31}$$

$$\mathbf{1}^{S+}_{ti} = \begin{cases} 1 & \text{if } \hat{\Theta}_{ti} - \mu_t\Delta_{ti} \notin \mathbb{Z}_k \\ 0 & \text{otherwise} \end{cases}$$

$$\mathbf{1}^{S-}_{ti} = \begin{cases} 1 & \text{if } \hat{\Theta}_{ti} + \mu_t\Delta_{ti} \notin \mathbb{Z}_k \\ 0 & \text{otherwise} \end{cases}$$

$$\mathbf{1}^{D}_{ti} = \begin{cases} 1 & \text{if } \hat{\Theta}_{ti} \pm \mu_t\Delta_{ti} \in \mathbb{Z}_k \\ 0 & \text{otherwise} \end{cases} \tag{12.32}$$

where $\mathbb{Z}_k$ indicates the input constraint set. In words, if at the t^{th} iteration double-sided perturbations remain in the constraint set, then a double sided gradient is computed for the i^{th} element of the input vector. Otherwise, a single-sided gradient approximation with perturbations in the direction opposite to the violated boundary is used.

Figure 12.1 provides a visual illustration of the operation of the CSUD gradient approximator (12.30) for a single input loss function. The grey area marked $\mathbb{Z}_k$ represents the input constraint. Note that the constraint applies in the horizontal direction. However, for visual clarity, the constraint zone is extended vertically as well. At point B, marked green, where a double-sided perturbation would remain inside the constraint, the double-sided perturbation is chosen. At points A (blue) or C (red), where a double double-sided perturbation would exceed the constraint, a single-sided perturbation in the

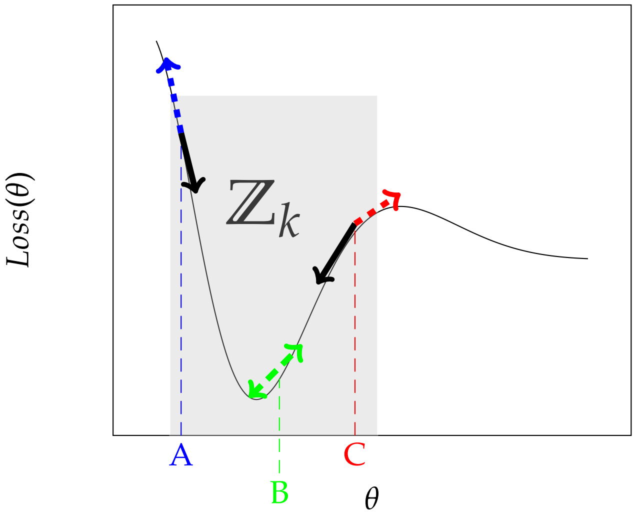


FIGURE 12.1: CSUD behaviour with a single input loss function $L(\theta)$. $\mathbb{Z}_k$ indicates the constraint zone. At point **A**, the dashed blue line indicates that a double sided perturbation would violate the constraint. Accordingly, to remain in the constraint zone $\mathbb{Z}_k$, a single-sided gradient in the opposite direction, indicated by the solid black line is calculated. Point **C** is evaluated similarly. At point **B**, the constraint $\mathbb{Z}_k$ is not binding, and a double sided perturbation is used in approximating the gradient.

opposite direction (marked black) is applied.

12.3.1 Background

Note that constraining either side of the double-sided perturbation defined in (12.3), would immediately violate the restrictions imposed by Assumption 12.2.6 on the perturbations. In particular, perturbations must be symmetrically distributed about 0, with 0 mean, but with no mass on a 0 outcome. It could be assumed that as $t \to \infty$, perturbation constraints would cease to be binding, and consequently an initially asymmetric perturbation distribution would converge to a symmetric one. However, even with such an assumption, any unilaterally constrained (double-sided) perturbation interval would become unbalanced, and lead to an over or underestimation of the gradient. So, in addition to a distributional assumption, for any constrained perturbations, it would be necessary to computationally rebalance the denominator of the gradient estimator in (12.3).

Such a re-balancing is of course feasible by scaling (dampening) the denominator of a constrained-perturbation double-sided gradient estimate down from $2\mu_t\Delta_{ti}$ to the appropriate sized interval. However, due to the inherent symmetry of double-sided SPSA, this would imply that the incoming input vector was not $\hat{\Theta}_t$, but another slightly altered input.

Such a set-up has been simulated elsewhere, where a dampening constant $d < 1$ was arbitrarily picked and applied to constrained double-sided gradient estimates. Surprisingly, once the dampening factor was tuned, acceptable results were achieved. However, it is suspected that the most effective value of this dampening constant may depend on the underlying estimation problem.

In order to avoid further theoretical, computational, and estimation complexities, the simpler alternative presented here was developed. As shown in (12.30), (12.31), and (12.32), this simple alternative consists of using constrained single and unconstrained double sided perturbations.

12.3.2 CSUD Assumptions

The theoretical requirements for the convergence of CSUD to minimum loss in the constrained input zone are not very different from those of unconstrained single-sided SPSA. The CSUD convergence assumptions are now presented.

Assumption 12.3.1 (Constrained Loss Topology)**.** *Let $\{\mathbb{Z}_k\}$ be a countable collection of compact subsets of $\mathbb{R}^p$, and $L(\Theta) : \mathbb{Z}_k \to \mathbb{R}$ is strictly convex and at least of class C^2 (two times continuously differentiable). Further, $L(\Theta)$ is bounded, has bounded derivatives, and has a unique minimum Θ^{k*} on $\mathbb{Z}_k$. In general $\forall\ \Theta \in \mathbb{Z}_k$, $L(\Theta)$ is bounded below at $L(\Theta^{k*})$. Assume that there is at least one such subset $\mathbb{Z}_k$.*

Assumption 12.3.1 requires that the loss function be strictly convex over the constrained input space $\mathbb{Z}_k$, which is being investigated. This could in theory lead to the use of loss functions, which are not strictly convex outside of the constrained input space. If multiple constraint sets $\mathbb{Z}_k$ are present and parameter space partitioning is used, then $\bigcap_k \mathbb{Z}_k = \emptyset$ must hold. This is required so that each partitioned parameter space can be identified uniquely.

Assumption 12.3.2 (Measurement Errors)**.** *Same as Assumption 12.2.2.*

Assumption 12.3.3 (Iterate Dynamics)**.** *The second part of Assumption 12.2.3 is retained. However, since any constraint sets $\{\mathbb{Z}_k\}$ are by definition bounded, the finite norm assumption is no longer needed.*

Assumption 12.3.4 (Mean ODE Dynamics)**.** *The prerequisites of Assumption 12.2.4 apply. However, it is now required that for any $\Theta \in \mathbb{Z}_k$, as $t \to \infty$, the differential equation $dZ(t)/dt = -g(Z(t))$ moves towards point Θ^{k*}.*

In short, the imposition of constraint sets is addressed by requiring that a local minimum exists in any constraint set, and that a path to this minimum exists from any point *inside* the constraint set. However, note that there no longer exists any assurance that a constraint set minimum Θ^{k*} is also the global minimum over the unconstrained domain of the loss function. By implication it is no longer required that $g(\Theta^{k*}) = 0$. Hence, the mean ODE dynamics may create a persistent drift, absorbed however, by the constraint boundaries.

For example in Fig. 12.1, the constrained minimum is also the global loss minimum, but it is easy to see that by shifting the gray constraint zone to the left or to the right, one can achieve a constrained minimum, where this is not the case.

The weakening of the previous assumptions from section 12.2 seems odd from a stochastic optimisation point of view, where one would always want to find the global minimum. However, it helps to prepare for a shift from stochastic optimisation towards stochastic search, where search costs may influence the amount of resources available towards finding a global minimum. In such cases, one may instead be willing to look for the second-best, third-best, or the first-available minimum.

Incidentally, if the loss function domain is fully partitioned into separate search zones, one could attempt to find the global minimum by only sampling a subset of the search zones, perhaps guided by a prior on the global minimum eligibility of each sample search zone. Such rational needle-in-the-haystack search problems are well-known and formulated using radial basis functions, see for example Abedinia and Amjady (2016). CSUD could be generalised in this direction; however this is not discussed in this work.

Assumption 12.3.5 (Step Sizes, α_t and μ_t)**.** *Same as Assumption 12.2.5.*

Assumption 12.3.6 (Perturbations)**.** *Same as Assumption 12.2.6.*

Assumption 12.3.7 (Stochastic Interactions)**.** *In addition to Assumption 12.2.7, let $\epsilon_t^- \equiv \epsilon(\hat{\Theta}_t - \mu_t \Delta_t)$. Assume $E_{\mathcal{F}_t,\Delta_t}(\epsilon_t - \epsilon_t^-) =_{a.s.} 0$, $E_{\mathcal{F}_t,\Delta_t}(\epsilon_t^+ - \epsilon_t^-) =_{a.s.} 0$, $E_{\mathcal{F}_t}(\epsilon_t - \epsilon_t^-) =_{a.s.} 0$, and $E_{\mathcal{F}_t}(\epsilon_t^+ - \epsilon_t^-) =_{a.s.} 0$, the latter two being implied by the independence of Δ_t.*

Since CSUD applies single- as well as double-sided perturbations, Assumption 12.3.7 also requires that double-sided perturbations do not add any systematic bias to observed loss. There is also a need for an additional assumption on constraint behaviour.

Assumption 12.3.8 (Constraint Behaviour). *On any constraint set $\mathbb{Z}_k$ with initial condition $\hat{\Theta}_{k0} \in \mathbb{Z}_k$, on subsequent updates for given $\hat{\Theta}_{ti}$, the measure of the event ($\hat{\Theta}_{ti} + \mu_t \Delta_{ti} \notin \mathbb{Z}_k \wedge \hat{\Theta}_{ti} - \mu_t \Delta_{ti} \notin \mathbb{Z}_k$) is zero.*

Assumption 12.3.8 is needed for proving unbiasedness of the constrained gradient estimator. Effectively, it is stipulated that when a single-sided perturbation is applied in either direction, then the resulting input in one of the directions must remain inside the constraint zone. In practice, one can construct $\mathbb{Z}_k$ and pick $\hat{\Theta}_{k0}$ so that this assumption holds.

12.3.3 Bias of the CSUD Gradient Estimate

It is necessary to show that the bias of the i$^{\text{th}}$ component of the constrained gradient estimate is asymptotically zero, $\mathscr{B}_{ti}^{\hat{g}_t^C} = \lim_{t\to\infty} E_{\mathscr{T}_t}\left[\hat{g}_{ti}^C(\hat{\Theta}_t) - g_i(\hat{\Theta}_t)\right] = 0$. Since the indicator function constructs (12.32) are independent of the gradient approximators in (12.31) and (12.3)

$$\begin{aligned} E_{\mathscr{T}_t}\left[\hat{g}_{ti}^C(\hat{\Theta}_t)\right] &= E_{\mathscr{T}_t}\left[\mathbf{1}_{ti}^{S+}\right] E_{\mathscr{T}_t}\left[\hat{g}_{ti}^{S+}(\hat{\Theta}_t)\right] \\ &\quad + E_{\mathscr{T}_t}\left[\mathbf{1}_{ti}^{S-}\right] E_{\mathscr{T}_t}\left[\hat{g}_{ti}^{S-}(\hat{\Theta}_t)\right] \\ &\quad + E_{\mathscr{T}_t}\left[\mathbf{1}_{ti}^{D}\right] E_{\mathscr{T}_t}\left[\hat{g}_{ti}^{D}(\hat{\Theta}_t)\right] \qquad (12.33a) \\ &= w_{ti}^{S+} E_{\mathscr{T}_t}\left[\hat{g}_{ti}^{S+}(\hat{\Theta}_t)\right] + w_{ti}^{S-} E_{\mathscr{T}_t}\left[\hat{g}_{ti}^{S-}(\hat{\Theta}_t)\right] \\ &\quad + w_{ti}^{D} E_{\mathscr{T}_t}\left[\hat{g}_{ti}^{D}(\hat{\Theta}_t)\right] \qquad (12.33b) \end{aligned}$$

where for any event $e \in \{S+, S-, D\}$, the weight w_{ti}^e is defined as the probability of that event at iteration t for element i, that is, $E_{\mathscr{T}_t}\left[\mathbf{1}_{ti}^e\right] = Prob(e_{ti}|\mathscr{T}_t) \equiv w_{ti}^e$. Note that by Assumption 12.3.8, $\sum_e w_{ti}^e = 1$.

Apply Taylor expansions to the right-hand side gradient approximation terms in (12.33b), simplify and re-collect terms to get

$$E_{\mathscr{T}_t}\left[\hat{g}_{ti}^C(\hat{\Theta}_t)\right] - E_{\mathscr{T}_t}\left[g_i(\hat{\Theta}_t)\right] \sum_e w_{ti}^e =$$

$$\frac{\mu_t\left(w_{ti}^{S+}-w_{ti}^{S-}\right)}{2}E_{\mathscr{F}_t}\left[\frac{1}{\Delta_{ti}}\right]\sum_{j\neq i}\frac{\partial^2 L}{\partial\theta_{tj}^2}E_{\mathscr{F}_t}\left[\Delta_{tj}^2\right]+$$
$$\frac{2w_{ti}^{S+}+w_{ti}^{D}}{2}E_{\mathscr{F}_t}\left[\frac{R_2(\hat{\Theta}_t,\mu_t\Delta_t)}{\mu_t\Delta_{ti}}\right]-$$
$$\frac{2w_{ti}^{S-}+w_{ti}^{D}}{2}E_{\mathscr{F}_t}\left[\frac{R_2(\hat{\Theta}_t,-\mu_t\Delta_t)}{\mu_t\Delta_{ti}}\right] \tag{12.34}$$

$$E_{\mathscr{F}_t}\left[\hat{g}_{ti}^{C}(\hat{\Theta}_t)-g_i(\hat{\Theta}_t)\right]=O(\mu^2),\quad \lim_{t\to\infty}\mu_t=0 \tag{12.35}$$

where in going from (12.34) to (12.35), $\sum_e w_{ti}^e=1$ has been used as implied by Assumption 12.3.8. Furthermore the remainder approximation of Folland (1990) is used, and the previous bound results from (12.19a) and (12.19b) have been applied.

In sum, with the updated Assumptions 12.3.1 to 12.3.8, it can be shown that each component of the CSUD gradient estimator $\hat{g}_t^C(\hat{\Theta}_t)$ is asymptotically unbiased.

12.3.4 Almost sure convergence of CSUD

Given a constraint set $\mathbb{Z}_k$ and an initial stating point $\hat{\Theta}_{k0}\in\mathbb{Z}_k$, it is necessary to show that $\hat{\Theta}_t\rightarrow_{a.s}\Theta^{k*}$. Following Yin and Kushner (2003, Ch. 5.1), the N-period shifted version of the one-period constrained update equation (12.29) is rewritten as

$$\hat{\Theta}_{t+N}-\hat{\Theta}_t=-\sum_{j=0}^{N-1}\alpha_{t+j}g(\hat{\Theta}_{t+j})-\sum_{j=0}^{N-1}\alpha_{t+j}\mathscr{B}_{t+j}^{\hat{g}_{t+j}^{C}}$$
$$-\sum_{j=0}^{N-1}\alpha_{t+j}\xi_{t+j}-\sum_{j=0}^{N-1}\alpha_{t+j}Z_{k,t+j} \tag{12.36}$$

where $N-1\geq t$, and the bias and error terms are respectively defined as

$$\begin{aligned}\mathscr{B}_t^{\hat{g}_t^C}&=E_{\mathscr{F}_t}\left[\hat{g}_t^C(\hat{\Theta}_t)-g(\hat{\Theta}_t)\right]=E_{\mathscr{F}_t}\left[\hat{g}_t^C(\hat{\Theta}_t)\right]-g(\hat{\Theta}_t),\\ \text{and }\xi_t&=\hat{g}_t^C(\hat{\Theta}_t)-E_{\mathscr{F}_t}\left[\hat{g}_t^C(\hat{\Theta}_t)\right].\end{aligned} \tag{12.37}$$

Using the techniques introduced in Proposition 12.2.1, almost sure convergence of the CSUD input estimate is now demonstrated.

Proposition 12.3.1 (CSUD Convergence)**.** *Given Assumptions 12.3.1 to 12.3.8 and the iteration rule* (12.29)*, if* $\Theta^{k*} \in \mathbb{Z}_k$ *and* $g(\Theta^{k*}) = 0$*, then as* $t \to \infty$*, for any such* $\mathbb{Z}_k$*,* $\hat{\Theta}_t \to \Theta^{k*}$ *a.s.*

Proof. Consider the N period forward-shifted representation shown in (12.36). Given any conforming $\mathbb{Z}_k$, it is necessary to show that $\lim\limits_{t\to\infty} \hat{\Theta}_{t+N} - \hat{\Theta}_t =_{a.s.} 0$. This result implies that $\hat{\Theta}_t \to_{a.s} \Theta^{k*}$. In the below discussion, N is a shorthand for $N > t$.

The Mean ODE term: Yin and Kushner (2003) provides a mean ODE convergence presentation, which works well in this context. The argument is not repeated here in detail. The overview argument presented in Proposition 12.2.1 continues to hold here, as long as $\Theta^{k*} \in \mathbb{Z}_k$ *and* $g(\Theta^{k*}) = 0$.

The Constraint Projection Term: The final term in (12.36) represents accrued corrections, which have accumulated to ensure that $\hat{\Theta}_j \in \mathbb{Z}_k$ for $j = t+1, \cdots, t+N$. By Assumption 12.3.1, Θ^{k*} lies in $\mathbb{Z}_k$. Therefore $\lim\limits_{t\to\infty} \sum_{j=0}^{t} \alpha_t Z_{m,t} < \infty$, and consequently $\lim\limits_{t\to\infty} \sum_{j=0}^{N-1} \alpha_{t+j} Z_{m,t+j} = 0$. In other words over time, the first (mean ODE) term in (12.36) will move the input estimate towards Θ^{k*}, therefore eventually, the correction terms will tend to 0. Hence as long as $\Theta^{k*} \in \mathbb{Z}_k$ *and* $g(\Theta^{k*}) = 0$, it follows that,

$Prob\left(\lim\limits_{t\to\infty} \sum_{j=0}^{N-1} \alpha_{t+j} Z_{m,t+j} = 0\right) = 1.$

Comments: In the mean ODE term and constraint projection term arguments above, it is assumed that $\Theta^{k*} \in \mathbb{Z}_k$ *and* $g(\Theta^{k*}) = 0$. This is necessary for the application of standard proof arguments. However, the possibility of gradient drift, where $g(\Theta^{k*}) \neq 0$ remains, in which case the standard arguments no longer produce the desired results. This possibility is not addressed any further here, however, will be commented on in section 12.5. The remainder of this proof, which uses standard approaches, is not affected by any gradient drift.

The Bias Term:

From (12.35), each bias term in the sum asymptotically converges to 0. Hence, it is established that $Prob\left(\lim\limits_{t\to\infty} \sum_{j=0}^{N-1} \alpha_{t+j} \mathscr{B}_{t+j}^{\hat{g}^C_{t+j}} = 0\right) = 1.$

The Error Term:

From Assumptions 12.3.6 and 12.3.7 given $\{\mathscr{T}_t\}$, $\{\alpha_t \xi_t\}$ forms an independent, mean zero sequence. Define $\mathscr{M}_t = \sum_{j=0}^{t-1} \alpha_j \xi_j$, and note that $E_{\mathscr{T}_t}\left[\mathscr{M}_{t+1}\right] = \mathscr{M}_t$. Therefore $\{\mathscr{M}_t\}$ forms a martingale. Once more, following Yin and Kushner (2003, Ch. 5, Theorem 2.1), Doob's martingale inequality is invoked. For any $\eta > 0, \eta \in \mathbb{R}$

$$Prob\left(\sup_{N\geq j\geq t} \|\mathscr{M}_j - \mathscr{M}_t\| \geq \eta\right) \leq \frac{1}{\eta^2} E\left[\|\mathscr{M}_N - \mathscr{M}_t\|^2\right]. \tag{12.38}$$

Developing the right hand side of (12.38) as in Proposition 12.2.1 yields

$$\frac{1}{\eta^2} E\left[\left\|\sum_{j=0}^{N-1} \alpha_{t+j}\xi_{t+j}\right\|^2\right] \leq \frac{1}{\eta^2}\sum_{i=1}^{p}\sum_{j=t}^{\infty} \alpha_j^2 E\left[\left(\hat{g}_{ji}^{C}(\Theta_j)\right)^2\right] \tag{12.39a}$$

$$\leq \frac{1}{\eta^2}\sum_{i=1}^{p}\sum_{j=t}^{\infty} \alpha_j^2 E\Bigg[\Big(\mathbf{1}_{ji}^{S+}\hat{g}_{ji}^{S+}(\Theta_j) + \mathbf{1}_{ji}^{S-}\hat{g}_{ji}^{S-}(\Theta_j) + \mathbf{1}_{ji}^{D}\hat{g}_{ji}^{D}(\Theta_j)\Big)^2\Bigg] \tag{12.39b}$$

$$\leq \frac{1}{\eta^2}\sum_{i=1}^{p}\sum_{j=t}^{\infty} \alpha_j^2 \sum_{e} v_{ji}^{e} E\left[\left(\hat{g}_{ji}^{e}(\hat{\Theta}_j)\right)^2\right] \tag{12.39c}$$

where $e \in \{S+, S-, D\}$. Result (12.39c) follows from (12.39b) since the indicator functions are independent of the gradient approximators, $\left(\mathbf{1}_{ji}^{e}\right)^2 = \mathbf{1}_{ji}^{e}$, $E\left[\mathbf{1}_{ji}^{e}\right] = Prob(e_{ji}) \equiv v_{ji}^{e}$, and given (12.32) and Assumption 12.3.8, for distinct events $e_{ji}, e'_{ji} \in \{S_{ti}^{+}, S_{ti}^{-}, D_{ti}\}$, it is the case that $E\left[\mathbf{1}_{ji}^{e}\mathbf{1}_{ji}^{e'}\right] = Prob(\mathbf{1}_{ji}^{e} \cap \mathbf{1}_{ji}^{e'}) = 0$.

Next the expectation terms in (12.39c) are evaluated and the enveloping result is established. Define bounds such that

$$\begin{aligned} E\left[\hat{g}_{ji}^{S+}(\hat{\Theta}_j)^2\right] &\\ &= E\left[\left(\frac{L(\hat{\Theta}_j + \mu_j\Delta_j) - L(\hat{\Theta}_j) + \epsilon_j^{+} - \epsilon_j}{\mu_j\Delta_{ji}}\right)^2\right] \\ &\leq \frac{1}{\mu_j^2}\left(\mathbb{L} + 2\mathbb{E}\right)\mathbb{D}, \end{aligned} \tag{12.40a}$$

$$\begin{aligned}
&E\left[\hat{g}_{ji}^{S-}(\hat{\Theta}_j)^2\right] \\
&\quad = E\left[\left(\frac{L(\hat{\Theta}_j) - L(\hat{\Theta}_j - \mu_j\Delta_j) + \epsilon_j - \epsilon_j^-}{\mu_j\Delta_{ji}}\right)^2\right] \\
&\quad \leq \frac{1}{\mu_j^2}\left(\mathbb{L} + 2\mathbb{E}\right)\mathbb{D},
\end{aligned} \tag{12.40b}$$

$$\begin{aligned}
&E\left[\hat{g}_{ji}^{D}(\hat{\Theta}_j)^2\right] \\
&\quad = E\left[\left(\frac{L(\hat{\Theta}_j + \mu_j\Delta_j) - L(\hat{\Theta}_j - \mu_j\Delta_j) + \epsilon_j^+ - \epsilon_j^-}{2\mu_j\Delta_{ji}}\right)^2\right] \\
&\quad \leq \frac{1}{\mu_j^2}\left(\mathbb{L} + 2\mathbb{E}\right)\mathbb{D},
\end{aligned} \tag{12.40c}$$

$$\begin{aligned}
&E\left[\left(L(\hat{\Theta}_j \pm \mu_j\Delta_j) - L(\hat{\Theta}_j)\right)^2\right] \leq \mathbb{L}, \\
&\quad E\left[\left(L(\hat{\Theta}_j + \mu_j\Delta_j) - L(\hat{\Theta}_j - \mu_j\Delta_j)\right)^2\right] \leq \mathbb{L},
\end{aligned} \tag{12.40d}$$

$$E\epsilon^2 \leq \mathbb{E}, \quad E\left[\left(\frac{1}{\Delta_{ji}}\right)^2\right] \leq \mathbb{D} \tag{12.40e}$$

where $0 < \mathbb{D}, \mathbb{E}, \mathbb{L} < \infty$ are bounds. Regarding the construction of $\mathbb{L}$, from Assumption 12.3.6 the perturbations are bounded. Assumption 12.3.1 implies that for any arbitrary $\Theta \in \mathbb{Z}_k$ and radius $r > 0$, the loss function $L(\Theta)$ is bounded for the open ball $B(\Theta, r)$. Therefore on $B(\Theta, r)$, all events of the squared loss terms in (12.40d) are always bounded, and consequently, the expectations are also bounded.

Using the results in (12.40a) to (12.40e) in equation (12.39c), and the fact that from Assumption 12.3.8, $\sum_e v_{ji}^e = 1$, simplification yields

$$\begin{aligned}
&\sum_{t=0}^{\infty} Prob\left(\sup_{N \geq j \geq t} \|\mathscr{M}_j - \mathscr{M}_t\| \geq \eta\right) \\
&\qquad\qquad \leq \frac{p\left(\mathbb{L} + 2\mathbb{E}\right)\mathbb{D}}{\eta^2}\sum_{t=0}^{\infty}\sum_{j=t}^{\infty}\frac{\alpha_j^2}{\mu_j^2} < \infty
\end{aligned} \tag{12.41}$$

since from Assumption 12.3.5, as $t \to \infty$, $\sum_{j=t}^{\infty} \alpha_j^2/\mu_j^2 \to 0$.

Given the result in equation (12.41), as before the Borel-Cantelli Lemma Durrett (2019, Ch. 5, Theorem 2.3.1) states that $Prob(\sup_{N \geq j \geq t}\|\mathscr{M}_j - \mathscr{M}_t\| \geq \eta \; i.o.) = 0$, where *i.o* denotes infinitely often. Consequently as $t \to \infty$,

$\sup_{N \geq j \geq t} \| \mathscr{M}_j - \mathscr{M}_t \| \rightarrow_{a.s} 0$. That is, the error term converges to 0 almost surely.

□

12.4 CSUD Search Strategy

The theoretical requirements for using CSUD as a search strategy (tool) are developed here. For providing search scores, a quadratic loss function is used of the form

$$Y(\Theta) = [m(\Theta) - m^*]' \, S \, [m(\Theta) - m^*] \tag{12.42}$$

where $m(\Theta)$ is an r-dimensional performance measurement, m^* is a performance measurement target, and S is an $r \ x \ r$ diagonal matrix of weights. The form of the loss function can be varied. However, quadratic loss functions are easy to specify, twice differentiable, and have no Taylor expansion remainder, improving at least in theory asymptotic performance.

Also note that by design, the loss function (12.42) is strictly convex with respect to the performance measure $m(\Theta)$. However, no claims are made regarding loss function convexity with respect to the underlying inputs Θ, which form the real area of interest. In fact, for two different inputs Θ and Θ', it is possible to have equal performance measures $m(\Theta) = m(\Theta')$; in such a case both Θ and Θ' would yield equal search scores. In a context where multiple underlying inputs produce the same search scores, input constraints, such as those employed by CSUD help to keep the search problem manageable and assist in focusing on specific input value ranges of interest.

12.4.1 Search Problem Specification

Let $R(\Psi, \Theta) : (\mathbb{R}^q, \mathbb{R}^p) \rightarrow \mathbb{R}^o$, $(p, q, o \geq 1, 2, \cdots)$, be a vector valued stochastic function, where the function input space has been partitioned into optimising parameters Ψ, and performance tuning hyper-parameters Θ.

The aim is to optimise R jointly with respect to the parameters Ψ and hyper-parameters Θ. Unfortunately the structure of the function R is such that

$$\underset{\Psi, \Theta}{opt} \, R(\Psi, \Theta) \tag{12.43}$$

is too difficult or too costly to solve jointly. For example, R may have such a form that it is linear, or log-linear in the parameters, but non-linear in the hyper-parameters.

In (12.43), the term *opt* indicates an optimisation operator such as *max* or *min*. Suppose that the input space of the function R can be partitioned such that for any fixed $\bar{\Theta}$, $\max_{\Psi} R(\Psi, \bar{\Theta})$ can be computationally optimised. Further suppose that a mapping function M can be constructed. This mapping function maps multiple realizations of $\max_{\Psi} R(\Psi, \bar{\Theta})$ to a statistical performance measure with at least a mean and variance. Then, both Θ and Ψ can be optimised by using a (two-stage) minimax strategy with a wrapper loss function $Y(\Theta)$

$$\min_{\Theta} Y\left(M\left(\left\{\max_{\Psi} R\left(\Psi, \Theta\right)\right\}_1^N\right)\right) \tag{12.44}$$

where the performance statistic M is computed using N repetitions.

(12.44) represents the general search problem formulation. To the author's knowledge, Widrow and Hoff (1960) initially proposed the technique of using a tractable loss function to assess an intractable objective function. This technique was later generalized in Rumelhart and McClelland (1987) and Rumelhart et al. (1986), to devise the back-propagation algorithm, which to date remains the primary method for training artificial neural networks (LeCun et al., 2012). While the CSUD search strategy shares mathematical similarities with back-propagation, the notable differences are (a) analytical derivatives are not used, and (b) the inner function has its own maximisation objective.

Recall that $M(\cdot)$ is a mapping function, which is applied to repeated measurements of the objective function $R(\cdot)$. For notational convenience, let $m(\Theta)$ denote the output of the mapping function $M(\cdot)$, which in expanded notation would include repeated sampling and maximisation with respect to Ψ. The notation $m(\Theta)$ facilitates focusing on hyper-parameter loss minimisation. Then it is straightforward to note that (12.42) represents a quadratic loss function implementation of the more general form (12.44). The theoretical requirements for (12.42) to produce asymptotically convergent minimum loss search score results are formalised next.

12.4.2 CSUD Search Strategy Assumptions

Assumption 12.4.1 (Search Strategy Topology). *$\Theta \in \mathbb{R}^p$, $m(\Theta) \in \mathbb{R}^r$. Let $\{\mathbb{Z}_k\}$ be a countable collection of compact subsets of $\mathbb{R}^r$ such that $\bigcap_k \mathbb{Z}_k = \emptyset$. The search function $L(m(\Theta)) : \mathbb{R}^p \to \mathbb{R}^r \to \mathbb{R}$ is strictly convex in m and at least of class C^2 (two times continuously differentiable). For any $\Theta^k \in \mathbb{Z}_k$, the measurement function $m(\Theta^k)$ is bounded; $L(m(\Theta^k))$ is bounded, and has bounded derivatives.*

Further, let m^* be the unique minimum of $L(m)$. Let there be at least one $\Theta^{k*} \in \mathbb{Z}_k$ such that $\|m(\Theta^{k*}) - m^*\|$ is minimized. Assume that there is at least one such subset $\mathbb{Z}_k$.

Assumption 12.4.1 indicates that it is no longer guaranteed that the loss function $L(m(\Theta))$ has a unique minimum in Θ. However, it is required that the loss function must be constructed in such a way that in any chosen constraint set with $\Theta^k \in \mathbb{Z}_k$, the measurements $m(\Theta^k)$ must produce a distance measure to the loss function minimum m^*. This distance measure can then in turn be used for ranking and selection. One can think of m^* as the measurement target.

Assumption 12.4.2 (Performance Mapping, Errors). *Given some fixed $\bar{\Theta} \in \mathbb{Z}_k$ and the associated behaviour solution set $\Omega^{\bar{\Theta}}_{\psi^*}$, let $M(\cdot)$ be a reductive mapping, which transforms the output of the behaviour solution set into r-dimensional performance criteria, $M\left(\left\{R\left(\Omega^{\bar{\Theta}}_{\psi^*}, \bar{\Theta}\right)\right\}\right) : \mathbb{R}^{|\Omega^{\bar{\Theta}}_{\psi^*}|_o} \to \mathbb{R}^r$. Assume that the mapped performance criteria output of $M(\{\cdot\})$ can be equivalently represented in simplified form by*

$$m = \hat{m}(\Theta) + \varepsilon, \quad \varepsilon \sim i.i.d, m \text{ is of size } r \tag{12.45a}$$

$$E\left[\varepsilon\right] = 0 \tag{12.45b}$$

$$\begin{aligned} &E|\varepsilon_i^2| < \infty, \quad E\varepsilon_i^3 = 0, \\ &\quad E|\varepsilon_i^4| < \infty, \quad i \in \{1, \cdots, r\} \end{aligned} \tag{12.45c}$$

$$E\left[\varepsilon\varepsilon'\right] = \Sigma = \begin{bmatrix} \sigma_1^2 & & \\ & \ddots & \\ & & \sigma_r^2 \end{bmatrix}. \tag{12.45d}$$

Furthermore, since for any t, the iterate $\hat{\Theta}_t$, and also its perturbations $\hat{\Theta}_t \pm \mu_t\Delta_t$ are by design bounded, then $\hat{m}(\cdot)$ is bounded. Consequently $\hat{m}(\hat{\Theta}_t)$ and $\hat{m}(\hat{\Theta}_t \pm \mu_t\Delta_t)$ have defined and finite moments up to and including order 4.

Assumption 12.4.2 states that each estimated performance criterion can be represented as the sum of a central estimate and a random error term, which may be heteroscedastic, that is, the variance of each performance criterion may differ.

Assumption 12.4.2 is strict but reasonable. Regarding the errors ε, given (12.45b), in (12.45c) symmetry is imposed about 0 and finite kurtosis is assumed. If for all $\hat{\Theta}_t \in \mathbb{Z}_k$ with positive measure, $\hat{m}(\hat{\Theta}_t)$ is bounded, then its moments up to any order would also be bounded on $\mathbb{Z}_k$. Additionally, cross-correlation of performance criteria is not permitted. The algorithm operator designs the reductive mapping $M(\cdot)$. Given that $M(\cdot)$ uses repeated measurements, it can be argued that the output form hypothesised in (12.45a) is consistent with the mean obtained from the operation of a central limit theorem. Here however, this issue is not considered any further.

The *i.i.d.* assumption in (12.45a) implies that in any iteration t, $E_{\mathscr{T}_t}\left[\varepsilon_t\right] = 0$ and that $E_{\mathscr{T}_t}\left[\varepsilon_t \varepsilon_t'\right] = \Sigma$.

Assumption 12.4.3 (Iterate Dynamics)**.** *Same as Assumption 12.3.3.*

Assumption 12.4.4 (Mean ODE Dynamics)**.** *Let t denote time. Let $g(m(\Theta))$ be the gradient of $L(m(\Theta))$. By Assumption 12.4.1, $g(m(\Theta))$ is continuous in m. Let $Z(t)$ be a differentiable function in m. Then as $t \to \infty$, the differential equation $dZ(t)/dt = -g(Z(t))$ converges towards a fixed point at m^* such that $\|m(\Theta^{k*}) - m^*\|$ is minimised for some $\Theta^{k*} \in \mathbb{Z}_k$.*

Note that Assumption 12.4.4 does not produce a fixed point result in the traditional sense. However, it is required that the iterate of interest $\hat{\Theta}_t$ will produce at least one outcome such that its measurement mapped distance from m^* is minimised. This approach is employed in robustness techniques (Hansen & Sargent, 2008), where finite deviations from an optimum are accommodated.

Assumption 12.4.5 (Step Sizes, α_t and μ_t)**.** *Same as Assumption 12.3.5.*

Assumption 12.4.6 (Perturbations)**.** *Same as Assumption 12.3.6.*

Assumption 12.4.7 (Stochastic Interactions)**.** *Same as Assumption 12.3.7.*

Assumption 12.4.8 (Constraint Behaviour)**.** *Same as Assumption 12.3.8.*

It is also assumed that the maximization step in (12.44) has solutions. This requirement is added as a reminder that the search strategy must receive useful information to ensure meaningful search results.

Assumption 12.4.9 (Existence of Maxima). *Given a constraint set $\mathbb{Z}_k$, for any $\bar{\Theta} \in \mathbb{Z}_k$, repeated applications of $\max_{\Psi} R\left(\Psi, \bar{\Theta}\right)$ produces a $\bar{\Theta}$-contingent optimised behaviour solution set $\Omega_{\psi^*}^{\bar{\Theta}}$.*

Repeated applications of the maximisation in Assumption 12.4.9 can be seen as mini-batches. Because $R(\cdot)$ is stochastic, repeated optimisation of Ψ for given $\bar{\Theta}$ help to establish a central tendency for the value of Ψ. In general, mini-batch size would depend on the variance of $R(\cdot)$. The higher the variance, the larger a batch size would be required.

Given Assumption 12.4.2 and (12.45), the quadratic loss function in (12.42) can be expanded as

$$\begin{aligned} Y(\Theta) &= \left[\hat{m}(\Theta) - m^* + \varepsilon\right]' S \left[\hat{m}(\Theta) - m^* + \varepsilon\right] && (12.46a)\\ &= \left[\hat{m}(\Theta) - m^*\right]' S \left[\hat{m}(\Theta) - m^*\right] + \epsilon(\Theta) && (12.46b)\\ &= L(\Theta) + \epsilon(\Theta) && (12.46c) \end{aligned}$$

where $\epsilon(\Theta) = \varepsilon' S\varepsilon + 2\varepsilon' S \left[\hat{m}(\Theta) - m^*\right]$.

There are two important observations regarding (12.46c). Firstly, note that the canonical SPSA loss function decomposition is achieved, as initially shown in (12.1), consisting of observed loss, unobserved loss, and additive error. This decomposition allows one to focus on hyper-parameter Θ search.

Secondly the error term $\epsilon(\Theta)$ includes a quadratic component, and therefore has a non-zero finite mean. $E\left[\epsilon(\Theta)\right] = tr(S\Sigma) < \infty$. However, in the SPSA / CSUD context, for errors $\epsilon(\Theta)$ and $\epsilon(\Theta')$ from any two draws $Y(\Theta)$ and $Y(\Theta')$, $E\left[\epsilon(\Theta) \; - \epsilon(\Theta')\right] = 0$. Furthermore Assumption 12.4.2 on errors ε, ensures that Assumption 12.4.7 on stochastic interactions continues to hold.

Given some initial $\hat{\Theta}_0 \in \mathbb{Z}_k$, (12.44) is solved step-wise iteratively in two successive stages

$$\begin{aligned} Stage1 : M\left(\left\{\max_{\Psi} R\left(\Psi, \hat{\Theta}_t\right)\right\}_1^N\right) &\rightarrow \hat{m}(\hat{\Theta}_t) + \varepsilon_t && (12.47a)\\ Stage2 : \hat{g}_t^C(\hat{\Theta}_t) &\rightarrow \hat{\Theta}_{t+1} = \hat{\Theta}_t - \alpha_t \hat{g}_t^C(\hat{\Theta}_t) - \alpha_t Z_k. && (12.47b) \end{aligned}$$

At each iteration t given $\hat{\Theta}_t$, stage 1 obtains N samples of $\max_{\Psi} R(\Psi, \hat{\Theta}_t)$, and then computes the corresponding performance statistics. Stage 2 implements the loss function in (12.46a) and the corresponding CSUD gradient and update equations (12.29) to (12.32). When the iteration budget is completed, the final $(\Psi, \hat{\Theta})$ tuple is extracted as solution candidates.

12.4.3 Bias of the CSUD Search Strategy Gradient Estimate

As before, it is necessary to show that the bias of the i$^{\text{th}}$ component of the constrained gradient estimate is asymptotically zero, $\mathscr{B}_{ti}^{\hat{g}_t^C} = \lim_{t\to\infty} E_{\mathscr{F}_t}\left[\hat{g}_{ti}^C(\hat{\Theta}_t)\right.$ - $\left.g_i(\hat{\Theta}_t)\right] = 0$. Taylor expansions of the loss function in (12.46b) are used, and the expanded gradient formula (12.33b) from section 12.3.3 is applied.

Note that the use of Taylor expansions in the context of (12.46b) is subject to some considerations in practice. In particular, the intermediate targeting measurement function $\hat{m}(\Theta)$ must be differentiable in Θ. This is likely to be the case for summary statistics such as the mean, but may not apply to non-parametric statistics such as the median. In future works, Assumption 12.4.1 could be re-written to reflect this. However, here all topology related assumptions have been kept in the same format for narrative simplicity.

Given Assumptions 12.4.2 and 12.4.7, the conditional expectations of error difference terms

$$\frac{(\epsilon^+ - \epsilon)}{\mu_t \Delta_{ti}}, \quad \frac{(\epsilon - \epsilon^-)}{\mu_t \Delta_{ti}}, \text{ and } \quad \frac{(\epsilon^+ - \epsilon^-)}{2\mu_t \Delta_{ti}} \tag{12.48}$$

evaluate to 0. Then arguments (12.34) and (12.35) from section 12.3.3 are applied. These arguments show that each component of the CSUD search strategy gradient estimator $\hat{g}_t^C(\hat{\Theta}_t)$ remains asymptotically unbiased.

12.4.4 Almost sure convergence of the CSUD Search Strategy

Using the techniques introduced in Proposition 12.2.1 and 12.3.1, almost sure convergence is now outlined of the CSUD search strategy hyper-parameter iterate $\hat{\Theta}_t$ to a non-unique result $\Theta^{k*} \in \mathbb{Z}_k$.

Proposition 12.4.1 (CSUD Search Strategy Convergence). *Given Assumptions 12.4.1 to 12.4.9, quadratic loss* (12.46), *and the iteration rule* (12.47b), *given any initial $\hat{\Theta}_0 \in \mathbb{Z}_k$, if $g(m(\Theta^{k*})) = 0$, then as $t \to \infty$, $\hat{\Theta}_t \to \Theta^{k*}$ a.s., where however Θ^{k*} may not be unique.*

Proof. The arguments in Proposition 12.3.1 continue to apply. The N period shifted representation of the update equation in (12.47b) is used. The N

period shifted equation continues to have the form shown in (12.36) and (12.37).

For the mean ODE, bias, and constraint projection terms, the arguments and comment in Proposition 12.3.1 apply without any modifications. However, the argument for the error terms is altered as follows.

The Error Term: Using the arguments in Propositions 12.2.1 and 12.3.1, which lead to (12.25b) and (12.39c), it is established that

$$\frac{1}{\eta^2}E\left[\left\|\sum_{j=0}^{N-1}\alpha_{t+j}\xi_{t+j}\right\|^2\right] \leq \frac{1}{\eta^2}\sum_{i=1}^{p}\sum_{j=t}^{\infty}\alpha_j^2\sum_e v_{ji}^e E\left[\left(\hat{g}_{ji}^e(\hat{\Theta}_j)\right)^2\right] \tag{12.49}$$

where $e \in \{S+, S-, D\}$ and $Prob(e_{ji}) \equiv v_{ji}^e$.

Next the expectation terms in (12.49) are evaluated, and then the enveloping result is established. For any two independent $\hat{\Theta}^a, \hat{\Theta}^b \in \mathbb{Z}_k$ and $e \in \{S+, S-, D\}$, consider the commonly occurring gradient approximation term

$$\begin{aligned}
&E\left[\left(\frac{Y(\hat{\Theta}^a) - Y(\hat{\Theta}^b)}{\mu_j\Delta_{ji}}\right)^2\right] \\
&\quad = \frac{1}{\mu^2}E\left[\left(\frac{1}{\Delta_{ji}}\right)^2\right]E\left[\left(L(\hat{\Theta}^a) - L(\hat{\Theta}^b) + \epsilon^a - \epsilon^b\right)^2\right] \\
&\quad \leq \frac{1}{\mu^2}E\left[\left(\frac{1}{\Delta_{ji}}\right)^2\right]E\left[\left(L(\hat{\Theta}^a) - L(\hat{\Theta}^b)\right)^2 + \left(\epsilon^a - \epsilon^b\right)^2\right] \\
&\quad \leq \frac{2}{\mu^2}\left(\mathbb{L} + \mathbb{E}\right)\mathbb{D}
\end{aligned} \tag{12.50a}$$

$$\text{where } E\left[\left(\frac{1}{\Delta_{ji}}\right)^2\right] \leq \mathbb{D} \tag{12.50b}$$

and $0 < \mathbb{D}, \mathbb{E}, \mathbb{L} < \infty$ are bounds. The details for the derivation of $\mathbb{E}$ and $\mathbb{L}$ are in Appendix D.

Using the result (12.50a) in (12.49), and the fact that from Assumption 12.4.8, $\sum_e v_{ji}^e = 1$, simplification yields

$$\sum_{t=0}^{\infty} Prob\left(\sup_{N\geq j\geq t}\|\mathscr{M}_j - \mathscr{M}_t\| \geq \eta\right)$$

$$\leq \frac{2p\left(\mathbb{L}+\mathbb{E}\right)\mathbb{D}}{\eta^2}\sum_{t=0}^{\infty}\sum_{j=t}^{\infty}\frac{\alpha_j^2}{\mu_j^2} \quad < \quad \infty \qquad (12.51)$$

since from Assumption 12.4.5, as $t \to \infty$, $\sum_{j=t}^{\infty} \alpha_j^2/\mu_j^2 \to 0$.

Given the result in (12.51), as $t \to \infty$, $\sup_{N \geq j \geq t} \|\mathscr{M}_j - \mathscr{M}_t\| \to_{a.s} 0$. That is, the error term converges to 0 almost surely.

Non-uniqueness of Θ^{k*}: The standard proof approach so far has shown that the CSUD search strategy can asymptotically approximate the gradient of the unknown loss function. This gradient drives the loss function component $\hat{m}(\hat{\Theta}_t)$ towards $\hat{m}(\Theta^{k*})$ such that $\|\hat{m}(\Theta^{k*}) - m^*\|$ is minimised. By Assumption 12.4.1 the uniqueness of m^* is required, however the uniqueness of Θ^{k*} is not guaranteed.

□

12.5 Discussion

This chapter presents asymptotic convergence proofs of the input iterates $\hat{\Theta}_t$ for unconstrained single-sided SPSA, CSUD, and the CSUD search strategy.

In the asymptotic proofs, the employed standard modern techniques rely on the properties of martingales. The techniques employed originate from automated root finding and are best suited for strictly convex functions over the unconstrained input domain.

In automated root finding problems, one looks for the function root, which is an optimum at which the gradient is 0. Starting from an initial input value $\hat{\Theta}_0$, the function is traversed according to a gradient driven update rule. In such cases, standard convergence theory states that automated traversal will eventually locate the function root.

In unconstrained single-sided SPSA, as in all other SPSA algorithms, the main innovation is that in addition to automated function traversal towards the root, the gradient itself is being approximated over time. Hence over iterations, as the gradient quality improves, so does traversal towards the function root.

In practice, unconstrained single-sided SPSA, along with all other SPSA derivatives, must be carefully calibrated initially (Spall, 2003, pp. 190-191). Otherwise in practice, due to the instabilities introduced by traversing a function using an approximated gradient, especially in the early iterations, the input iterate might diverge and lead to a numerical overflow or underflow, or to large oscillations which increase convergence time.

One approach to dealing with SPSA instability is presented in Bhatnagar et al. (2013, pp. 52-65), who employ deterministic perturbations. Alternatively, as Spall (2003, p. 195) indicates, input iterate constraints may be used. In this context, constraining input iterates constitutes an elegant way of helping the function root search through any instabilities, while reducing the need for costly calibration.

CSUD constrains not only input iterates, but also the input perturbations used in gradient approximation. This would help in circumstances where perturbations are too large, or lead to function evaluation at illegal values. Additionally, CSUD only uses single-sided SPSA in the event of a binding perturbation constraint, otherwise using double-sided SPSA. As noted in Table 12.1, double-sided SPSA has smaller bias than single-sided SPSA. Therefore when perturbation constraints are seldom binding, CSUD will approximate double-sided SPSA efficiency.

The most significant consequence of constraining inputs is that additional assumptions are required regarding loss function topology. In particular, for standard martingale convergence proofs to work with constrained inputs, one must at least assume that the function root lies inside the constrained domain. However, in CSUD, where additionally perturbations are being constrained, it is additionally necessary that the initial input iterate must be selected from within the constraint zone. In practice this added requirement is easy to fulfil.

There is a further subtlety, which the CSUD convergence proposition 12.3.1 does not fully address. Assumption 12.3.1 introduces the notion of constraint sets $\mathbb{Z}_k$, asserts that the loss function is strictly convex for any $\Theta^k \in \mathbb{Z}_k$, and further requires the existence of a local minimum Θ^{k*}.

While it is not immediately obvious, these requirements define a problem, which differs from the automated root finding problem. In particular, the local minimum Θ^{k*} may occur at the constraint boundary possibly with non-zero gradient. A non-zero gradient impacts the outline of the *mean ODE term* component of proposition 12.3.1. In proposition 12.2.1 for the convergence of unconstrained single-sided SPSA, the *mean ODE term* proof relies on the discovery of a 0 gradient. In CSUD the discovery of such a 0 gradient is no longer guaranteed. However, any input iterate update drift caused by a non-zero gradient would be absorbed by the constraint. Currently, the proofs presented here do not explicitly deal with this eventuality. Improving the proofs to address this shortcoming could be an interesting future task. In practice, if a constrained CSUD iterate exhibits decaying bounces towards the lower or upper constraint limit, then the analyst should consider moving the impinged bound.

The CSUD search strategy departs even further from the automated root finding context. As with CSUD, due to constraints, the possibility exists that a non-zero gradient may obtain at constraint boundaries. Further, the CSUD search strategy uses a performance measurement target to tune the hyper-parameters of another model. Therefore, it is possible that there are not any hyper-parameter combinations, which fully satisfy the performance measurement target, in which case a non-zero gradient may continue to exert hyper-parameter update drift even when not at a constraint boundary. The *mean ODE term* proof in proposition 12.4.1 does not fully deal with this issue, and addressing the potential effects of such residual update drift constitutes another interesting area of future research. In practice, if the CSUD search strategy hyper-parameter trajectory shows a gradual creep in one or multiple parameters, while maintaining decreasing or constant loss, then one may suspect an update drift situation. However, if the drifting hyper-parameter(s) are associated with similar search strategy loss rankings, then such hyper-parameter combinations can be regarded as equally ranked candidates.

Finally, it is also possible for the CSUD search strategy to produce distinct hyper-parameter outcomes, each of which have similar performance measures, and consequently similar loss function rankings. In other words, the CSUD search strategy cannot guarantee a unique optimal result, as for example is expected from the canonical automated root finding exercise. The proof of proposition 12.4.1 extends the standard proof technology to its limit. There may be better, more comprehensive, or easier proof approaches. This once more could be an interesting area of future research. For example, it would be interesting to see whether the proofs can be carried out using a version of the burst learning model presented in chapter 10.

Chapter 13

Application: Neural Network Hyper-parameter Tuning

This chapter presents a case study of the CSUD search strategy as an artificial neural network (ANN) hyper-parameter tuner. ANN hyper-parameter tuning is a well known problem with both industrial and commercial applications, and has received a considerable amount of attention in computing literature. Well-known and high-performing ANN tuning techniques exist. This chapter investigates how the proposed implementation of the CSUD search strategy fares in an ANN tuning context.

This chapter is presented as a self-contained unit. First an introduction to ANN tuning is presented, followed by a brief review of the literature. Then using a small neural network model with the Fashion MNIST data set (Xiao et al., 2017), the CSUD search strategy implementation is presented and discussed.

13.1 ANN Tuning

Due to truth table compression, which produces prediction uncertainty (Widrow & Hoff, 1960), average artificial neural network (ANN) prediction and consequently validation accuracy cannot reach 100% (in non-training data). The question arises as to what level of validation accuracy is acceptable, how one can most quickly attain the desired level of validation accuracy, and to what extent can the chosen model be fine-tuned reliably and reproducibly? These questions lie at the core of the ANN training, validation, and tuning pipeline.

As ANNs can be very large, with thousands, sometimes millions or billions of network weights and many hyper-parameters, the main challenge in ANN hyper-parameter tuning is to mitigate the dimensional cost of a pure grid search. There are four general mitigation strategies: (1) distributional, (2) progressive resource allocation, (3) genetic, and (4) gradient based.

Using the expected improvement criterion to assess the contribution of hyper-parameter variations, Bergstra et al. (2011) present Gaussian process, and tree-structured Parzen hyper-parameter optimizers. They parallelise the iterative search process by using place-holder or replacement values for performance threshold returns. Snoek et al. (2012) discuss Bayesian hyper-parameter optimisation, a distributional method where hyper-parameters are directly drawn from an iteratively evolving Gaussian process.

As a pure resource allocation tuner, L. Li et al. (2018) introduce Hyperband, where the number of hyper-parameter configurations are seen as bandit arms. Using a trade-off between the number of bandit arms and the number of arm-pulls, Hyperband develops pay-off and stopping criteria, then progressively reduces the number of configurations, which eventually converge to the best hyper-parameter configuration.

Distributional tuning methods can be coupled together with resource allocation methods. For example, Bayesian hyper-parameter tuning can be coupled with a resource allocation scheduler like the Asynchronous Successive Halving Algorithm ASHA (L. Li et al., 2020). ASHA supports industrial scale parallelisation, early termination, and can be used in settings where hyper-parameter configurations exceed parallel processing capacity. ASHA parallelises the Successive Halving Algorithm, where for a small number of configurations, resource capped trials are executed. After each execution batch, a small number of top performing configurations are carried forward with an increased resource cap. Eventually, the best performing configuration gets to be tested with the most resources. Combining pure random search with ASHA can provide a lower computational cost alternative to Bayesian optimisation.

In terms of genetic algorithm driven ANN tuners, A. Li et al. (2019) introduce Population Based Training (PBT), a genetic search algorithm with early termination and check-pointing support applied to a collection (i.e., population) of hyper-parameter varied networks. PBT also supports selection criteria which may not be differentiable. Parker-Holder et al. (2020) present PB2, a variant of PBT, where hyper-parameters are perturbed not randomly but by Gaussian process bandits. Their approach minimises the number of hyper-parameter perturbations, which do not lead towards viable outcomes. Here PBT is used, as it is the conceptually simpler alternative.

In gradient based hyper-parameter optimisation (Bengio, 2000), hyper-parameters are updated via the implicit function theorem, using gradient information derived from a model scoring function. Lorraine et al. (2020) extend Bengio's approach to potentially millions of hyper-parameters. Gradient based approaches are attractive because a high quality gradient provides a shortcut through the hyper-parameter grid towards the optimum. However, the scoring function must be differentiable and possess at least one optimum over the hyper-parameter grid. In practice, hyper-parameter scoring is usually conducted using validation loss, with the validation loss gradient driving the hyper-parameter updates. Further, the ANN validation loss gradient is computed using automatic differentiation, which is possible when known differentiable activation and loss functional forms are used.

The CSUD search strategy can be seen as a gradient approximation driven hyper-parameter optimiser, where the search scoring function is evaluated at randomly perturbed hyper-parameter values, with the gradient being approximated and asymptotically achieved. From this perspective, one can immediately see that compared to automatic differentiation gradient methods, the CSUD search strategy will present with increased computational complexity, as every gradient approximation requires two distinct hyper-parametrised ANN evaluations. However, the random perturbations nature of CSUD, may lead to interesting hyper-parametrised ANN configurations, that might have been ordinarily missed. Further, in cases where automatic differentiation is too difficult, the CSUD search strategy makes it possible to conduct gradient based ANN tuning.

13.2 The Training, Validation, and Tuning Pipeline

In general, tuning an ANN is a complex undertaking with critical sequential steps. Firstly, every hyper-parameter configured ANN to be tuned must be trained to a high enough standard. Secondly when training is completed, ANN performance must be measured using data not seen during training. That is, the validation error must be computed. The validation loss gradient must be computed for gradient based tuners from a single validation error measurement, and in the case of CSUD, from two validation error measurements. In turn, this gradient is used to compute hyper-parameter updates; then the cycle is repeated. Since ANN training is time and resource intensive, and since ANN tuning may require the training of many alternative ANNs, it is noted that ANN tuning is a costly process.

13.2.1 ANN Training

ANN training is by far the costliest step, and efforts to speed up training whilst reducing the risk of over-training have received much attention. Adam by Kingma and Ba (2014) is generally considered to be a fast ANN training algorithm, which is also resistant to overtraining. It belongs to the class of iterative update stochastic approximation algorithms, where parameter updates make use of scaled gradient inputs. In Newton-Raphson like approaches, gradients are scaled by the loss function inverse Hessian or its approximation (Spall, 2003, pp. 27-30). However, direct computation or approximation of the inverse Hessian is costly and can produce instabilities (Bishop, 2006, pp. 249-256). In Adam, the incoming gradient is smoothed and also normalised. The smoothing and normalisation work together to (1) achieve gradient scaling, and (2) apply bounds to the update magnitudes (Kingma & Ba, 2014, pp. 2-3). Hence, Adam avoids the need for costly inverse Hessian computations, but retains the advantage of gradient scaling. This makes Adam comparatively fast to converge.

Could CSUD, or SPSA be used to train an ANN? In principle, the answer to this question is yes. Gradient smoothing, for example as used in Adam, has been discussed among others by Spall and Cristion (1994), who present an unnormalised gradient smoothing technique for SPSA. Zhu et al. (2020) discuss the computational cost of the Hessian and introduce an inverse Hessian approximation technique, which reduces computational cost from $O(p^3)$ to $O(p^2)$, where p denotes the number of parameters. They report real-data results with an empirical loss function and the airfoil self-noise data set, where their SPSA implementation for training a small ANN achieves lower loss than Adam. They do not discuss whether in their implementation, SPSA updates all parameters of the network, or just the output layer parameters, with back-propagation being used for hidden layer updates. Also their study does not present any validation accuracy results,

Method	Epochs	Training Loss	Validation Loss	Training Accuracy	Validation Accuracy
Adam	20	0.2305	0.2431	0.9150	0.9097
SPSA	20	0.8530	0.8790	0.6882	0.6758
Adam	50	0.1785	0.2289	0.9328	0.9192
SPSA	50	0.7487	0.7618	0.7286	0.7248
Adam	1000	0.1011	0.3058	0.9619	0.9169
SPSA	1000	0.5374	0.5601	0.8076	0.7998

TABLE 13.1
Could SPSA be used for ANN training? Fashion MNIST ANN training reference benchmarks. Adam outperforms unconstrained double-sided SPSA at 20, 50, and 1000 iterations.

making it difficult to assess how their lower training loss translates into prediction accuracy.

Using Fashion MNIST data, the use of unconstrained double-sided SPSA (12.1.1) is tested for training by forward pass, all weights of an ANN with 30,573 trainable weights. As Table 13.1 indicates at 20, 50, and 1000 training epochs, compared to Adam, unconstrained double-sided SPSA remains slow to converge, with Adam achieving lower loss and higher accuracy for training and validation respectively. Therefore, this chapter retains the use of Adam for ANN training.

13.2.2 ANN Validation and Tuning

The most important aspect of ANN validation is the use of data, which has not been presented to the ANN during training. ANN validation therefore is conducted with unseen data, typically reporting validation loss or validation accuracy. It is possible to design a custom validation score function, however, validation loss is used here.

Statistically evaluating a completed hyper-parameter tuning pass against any alternatives is important but has not received much consideration in the literature. This area is addressed by proposing an evaluation methodology based on pairwise non-parametric testing of a repeated performance metric, for example, validation accuracy. The use of the Dwass-Steele-Critchlow-Fligner all-pairs test is proposed. While the proposed method is resource intensive, it yields results, which can be statistically tested for the null hypothesis of performance equality, thereby providing insight into the relative efficacy of each alternative hyper-parameter configured ANN.

The next section describes the CSUD search strategy implementation used for ANN tuning.

13.3 CSUD Search Strategy Implementation

The CSUD search strategy implementation used here differs slightly from the one presented in section 12.4.

Specifically, finite delays to the learning rate α_t and perturbation step μ_t decay schedules are permitted. Such a delay is conducted via the use of *hold* and *span* operations, where *hold* indicates the number of delays, and *span* indicates the duration of each delay in CSUD iterations. For example, $hold = 2$, $span = 2$ implies that given a learning rate sequence $\{\alpha_1, \alpha_2, \alpha_3, \dots\}$, the learning rate in iterations 1 and 2 will be α_1, and in iterations 3 and 4, α_2; then in subsequent iterations will continue from α_3 onwards as usual. The perturbation step μ_t is likewise synchronized. This modification improves search speed. Since the altered learning rate and perturbation step sequences are of finite length, proposition 12.4.1 can be updated relatively easily; however, this is not demonstrated here.

Validation loss is measured by sparse categorical cross entropy, a loss function which deviates from the quadratic loss function (12.42) employed by the CSUD search strategy in section 12.4. This change in loss function may require re-working of the error structure in Assumption 12.4.2, and possibly further assumption updates. Once more, here these issues are not addressed formally. Instead, the current chapter focuses on empirical results.

13.4 Methodology

The Fashion MNIST (Xiao et al., 2017) data set is used with a convolutional neural network architecture, which is illustrated in Fig. 13.1. Two models are considered, model A and model B, the former without and the latter with L2 regularisation, which applies to the network loss function a

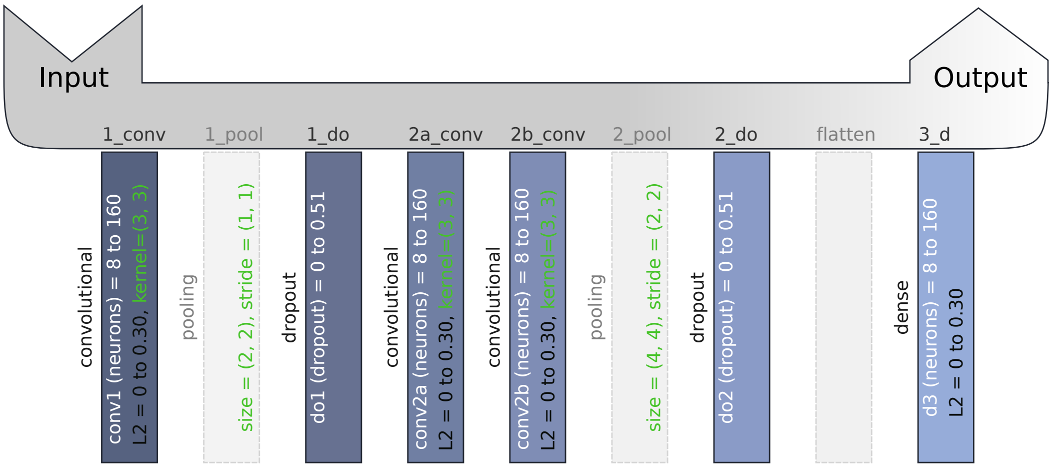

Input
Output
1_conv
1_pool
1_do
2a_conv
2b_conv
2_pool
2_do
flatten
3_d
convolutional
conv1 (neurons) = 8 to 160
L2 = 0 to 0.30, kernel=(3, 3)
pooling
size = (2, 2), stride = (1, 1)
dropout
do1 (dropout) = 0 to 0.51
convolutional
conv2a (neurons) = 8 to 160
L2 = 0 to 0.30, kernel=(3, 3)
convolutional
conv2b (neurons) = 8 to 160
L2 = 0 to 0.30, kernel=(3, 3)
pooling
size = (4, 4), stride = (2, 2)
dropout
do2 (dropout) = 0 to 0.51
dense
d3 (neurons) = 8 to 160
L2 = 0 to 0.30

FIGURE 13.1: Convolutional network architecture for use with Fashion MNIST. Coloured layers contain model A and B tunable hyper-parameters, which are in white typeface. Additional model B hyper-parameters are in dark gray. Light gray layers are not tunable. In convolutional and dense layers, the number of neurons ranges from 8 to 160. In computations, integer neuron numbers are mapped to the unit interval. Dropout range is from 0 to 0.51. L2 regularization scaling is from 0 to 0.30.

quadratic regularisation penalty based on the number of trainable network weights.

In convolutional neural network (ANN) tuning, frequently the learning rate and convolution window related parameters are tuned. However, at 28x28 pixels, Fashion MNIST image data resolution is relatively small, and the tuning opportunities for convolution window size, especially for a multi-layer architecture such as the one used here are limited. Further, Adam's automatic gradient scaling makes initial selection of the ANN learning rate resilient to over or under specification. Therefore, here learning rate or convolution window parameters are not tuned. Further, the number of (maximum) training epochs are fixed at 20, a duration at which Adam achieves good results.

The goal here is to investigate the effects of layer size, dropout, and L2 regularisation on ANN validation accuracy, and tuning parameters have been selected accordingly. As Fig. 13.1 indicates, model A has six tunable hyper-parameters: the number of neurons for three convolutional and one dense layer, 1_*conv*, 2*a*_*conv*, 2*b*_*conv*, and 3_*d*; and two dropout settings, 1_*do*, and 2_*do*. Model B has ten tunable hyper-parameters: all those listed under model A, plus L2 regularisation applied to model A's convolutional and dense layers, 1_*conv*_*l*2, 2*a*_*conv*_*l*2, 2*b*_*conv*_*l*2, and 3_*d*_*l*2.

Tuning Algorithm	Criterion	Settings	Passes
ASHA Bayes	min loss_val	$samples$ = 50	3
ASHA Random	min loss_val	$samples$ = 50	3
PBT	min loss_val	$population$ = 24	3
CSUD	final loss_val	$iterations$ = 60, α_1 = 0.01, $hold$ = 2, $span$ = 2	3
Training Algorithm	**Criterion**	**Settings**	**Epochs**
Adam	min loss_trn	α = 0.001	20

TABLE 13.2
Tuning and training algorithm configuration settings.

In convolutional and dense layers, the number of neurons to be tuned ranges from 8 to 160. In tuning searches, the 8 to 160 integer range is mapped to the unit interval. Tunable dropout ranges from 0 to 0.51. For L2 regularisation, the L2 scaling factor is between 0 and 0.30. In Fig. 13.1, tunable model A hyper-parameters are in white. Additional model B hyper-parameters are in dark gray. Non-tuned hyper-parameters appear in green.

Having presented the ANN architecture and tunable hyper-parameters, the tuning and assessment procedure, consisting of two stages, is discussed next. Firstly the tuning stage runs the to be evaluated hyper-parameter tuning algorithm. Four tuning algorithms are considered: ASHA random, ASHA Bayes, population based training (PBT), and CSUD. ASHA denotes use of the resource scheduler, the asynchronous successive halving algorithm introduced in section 13.1. At the end of a tuning pass, performance is assessed via validation loss. Each tuning algorithm is run 3 times, yielding 3 tuning passes each for the ASHA random, ASHA Bayes, PBT, and CSUD tuners. Table 13.2 and Table 13.3 provide all tuning and evaluation stage configuration settings.

The ASHA random, ASHA Bayes, and PBT tuners all employ resource scheduling techniques, and minimise validation loss over their respective tuning budgets. CSUD also minimises validation loss over its iteration budget. However, in CSUD the final budgeted iteration hyper-parameter combination is chosen by default. This is because, in general, CSUD is expected to converge asymptotically, and it is assumed that as iterations progress, hyper-parameter combinations will lead to networks with lower validation

Evaluation Statistic	Criterion	Settings	Sample Size
DSCF[a] pairwise testing	acc_val	p-value = 0.01	25

Training Algorithm	Criterion	Settings	Epochs
Adam	min loss_trn	α = 0.001	20

[a] DSCF stands for the Dwass-Steele-Critchlow-Fligner all-pairs test (Konietschke et al., 2015).

TABLE 13.3
Evaluation stage configuration.

loss. In practice, however, stochasticity or oscillations could slow down CSUD convergence, or lead to an overshoot. Hence, the strategy of naively assuming that the last iteration will also yield the optimised choice may not always be appropriate.

As Table 13.3 depicts, in the second, the performance evaluation stage, the hyper-parameter architecture selected in each tuning pass is trained for 25 separate times with randomised initial weights. Then the results are used to generate a sampling distribution for validation accuracy (the assessment metric), which is then evaluated using non-parametric pairwise equality tests. The Dwass-Steele-Critchlow-Fligner all-pairs test (DSCF, Konietschke et al., 2015) is used to assess among all tuning passes, statistically significant validation accuracy differences at a p-value of 0.01.

The tuning stage and data generation for the evaluation stage were implemented in Ray-Tune (Liaw et al., 2018), on three separate computers, each equipped with an NVIDIA RTX 3060 card, however without using Ray-Tune's distributed features to farm-out a single task across multiple computers. Thus parallelisation was limited solely to the Tensorflow (Abadi et al., 2015) features provided for use with a single graphics card.

13.4.1 Hyper-Parameter Mapping

As Fig. 13.1 indicates, layer size ranges from 8 to 160 neurons, while dropout ranges from 0 to 0.51, and L2 regularisation ranges from 0 to 0.30. The scale difference in layer size against dropout, or L2 regularisation, is quite large. In order to apply similar perturbation scales in CSUD for all hyper-parameters, the layer size range is mapped to the unit interval, with 0 indicating 8 and 1 indicating 160.

13.5 Results

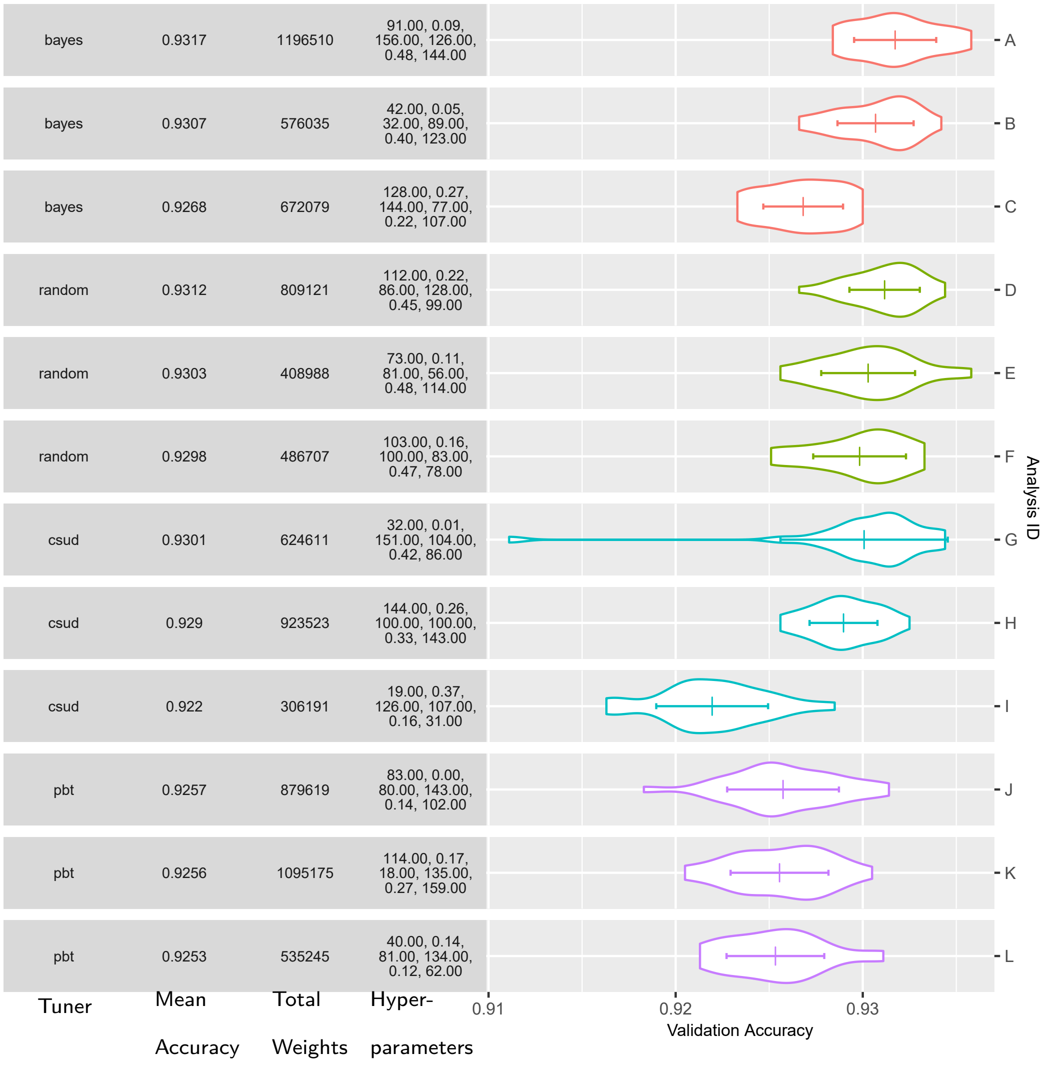
Fig. 13.2 shows ANN hyper-parametrisations discovered for model A after three distinct tuning passes of each tuning algorithm, and displays validation accuracy violin plots produced consequent to 25 training sessions for each tuning pass discovered hyper-parametrisation. The columns on the left display from left to right, tuning algorithm, mean validation accuracy, total trainable weights, and hyper-parameter configuration.

The hyper-parameters column display order is 1_conv, 1_do, $2a\_conv$, $2b\_conv$, 2_do, and 3_d, with layer abbreviations as discussed in Fig. 13.1.

The violin plots on the right present validation accuracy densities corresponding to 25 training sessions with mean validation accuracy $\pm 1SE$ marked in each violin plot. Results are grouped by tuning algorithm, and then in order of descending mean validation accuracy.

Fig. 13.2 reveals that for the Fashion MNIST ANN model A, compared to CSUD and PBT (population based training), the ASHA Bayes and ASHA random algorithms produce in general higher validation accuracy clusters.

PBT displays hyper-parametrisations with lowest dropout for both dropout layers. The CSUD hyper-parametrisation with analysis ID G displays a violin plot with a long left tail, indicating that this hyper-parametrisation may be more sensitive to initial training weight values. This sensitivity may have been caused by the combination of a small initial convolutional layer $1\_conv = 32$ and a first dropout layer with very low dropout $1\_do = 0.01$. The most consistent performance in terms of high mean validation accuracy is produced by the ASHA random tuner, followed by the ASHA Bayes, CSUD and PBT tuners.



Tuner	Mean Accuracy	Total Weights	Hyper-parameters	Analysis ID
bayes	0.9317	1196510	91.00, 0.09, 156.00, 126.00, 0.48, 144.00	A
bayes	0.9307	576035	42.00, 0.05, 32.00, 89.00, 0.40, 123.00	B
bayes	0.9268	672079	128.00, 0.27, 144.00, 77.00, 0.22, 107.00	C
random	0.9312	809121	112.00, 0.22, 86.00, 128.00, 0.45, 99.00	D
random	0.9303	408988	73.00, 0.11, 81.00, 56.00, 0.48, 114.00	E
random	0.9298	486707	103.00, 0.16, 100.00, 83.00, 0.47, 78.00	F
csud	0.9301	624611	32.00, 0.01, 151.00, 104.00, 0.42, 86.00	G
csud	0.929	923523	144.00, 0.26, 100.00, 100.00, 0.33, 143.00	H
csud	0.922	306191	19.00, 0.37, 126.00, 107.00, 0.16, 31.00	I
pbt	0.9257	879619	83.00, 0.00, 80.00, 143.00, 0.14, 102.00	J
pbt	0.9256	1095175	114.00, 0.17, 18.00, 135.00, 0.27, 159.00	K
pbt	0.9253	535245	40.00, 0.14, 81.00, 134.00, 0.12, 62.00	L

FIGURE 13.2: Model A tuning passes and validation accuracy from 25 repeated training sessions. Configuration on the left. Violin plots with mean validation accuracy $\pm 1SE$ on the right. Bayesian and random tuners perform well. Hyper-parametrisations with larger total number of weights have higher mean validation accuracy. Details in text.

Are the differences in mean validation accuracy statistically significant? Fig. 13.3 shows the model A pairwise comparison matrix of mean validation accuracy with p-value results using the DSCF all-pairs tests (Konietschke et al., 2015). Rows and columns are grouped by tuning algorithm. Each tuning algorithm grouping contains 3 passes with randomly initialised weights. Each cell contains p-values resulting from pairwise comparison of equality of mean validation accuracy obtained from 25 separate training runs conducted at the respective hyper-parametrisations reported in Fig. 13.2.

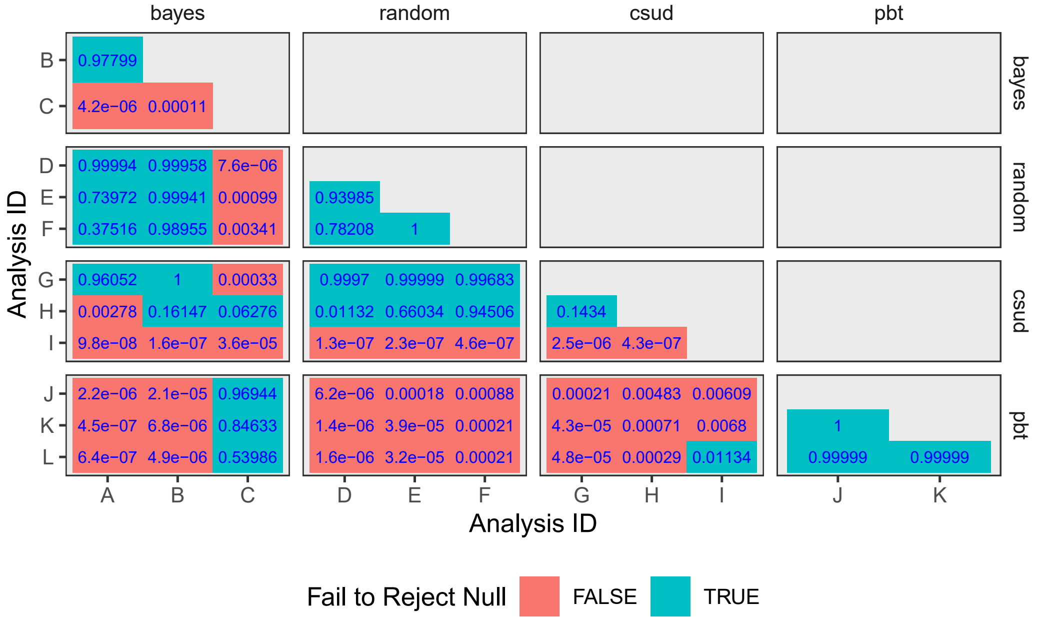


FIGURE 13.3: Fashion MNIST model A pairwise comparison matrix of mean validation accuracy with p-values using the DSCF criterion. Rows and columns are grouped by tuning algorithm. P-values obtained from pairwise comparison of mean validation accuracy using 25 separate training runs. Cells with p-values less than 0.01, where the null hypothesis of equality of mean validation accuracy is rejected, are coloured in red. For PBT and ASHA random, the null hypothesis of equal mean validation accuracy could not be rejected within each group at a significance level of 0.01. Green cells indicate CSUD tuning results show some overlap with ASHA random and ASHA Bayes. Further details in text.

Cells with p-values less than 0.01, where the null hypothesis of equality of mean validation accuracy is rejected, are coloured in red.

Results indicate that PBT and random tuning groups produce ANN hyper-parameter configurations where the null hypothesis of equal mean validation accuracy could not be rejected *within* each group at a significance level of 0.01. Tuning algorithm cross-group comparisons show that based on pairwise comparisons of mean validation accuracy at a significance level of 0.01, the ASHA random, ASHA Bayes, and CSUD algorithms show some overlap.

At significance level 0.01, ASHA Bayes versus ASHA random and CSUD versus ASHA random cross-group comparisons do not exhibit statistically significantly different mean validation accuracy results for 6 of 9 possible combinations. CSUD versus ASHA Bayes cross-group comparisons do not exhibit statistically significantly different mean validation accuracy results for 4 of 9 possible combinations. In contrast, PBT tuned model A hyper-parametrisations yield mean validation accuracy outcomes, which in cross-group comparison terms are largely statistically significantly different.

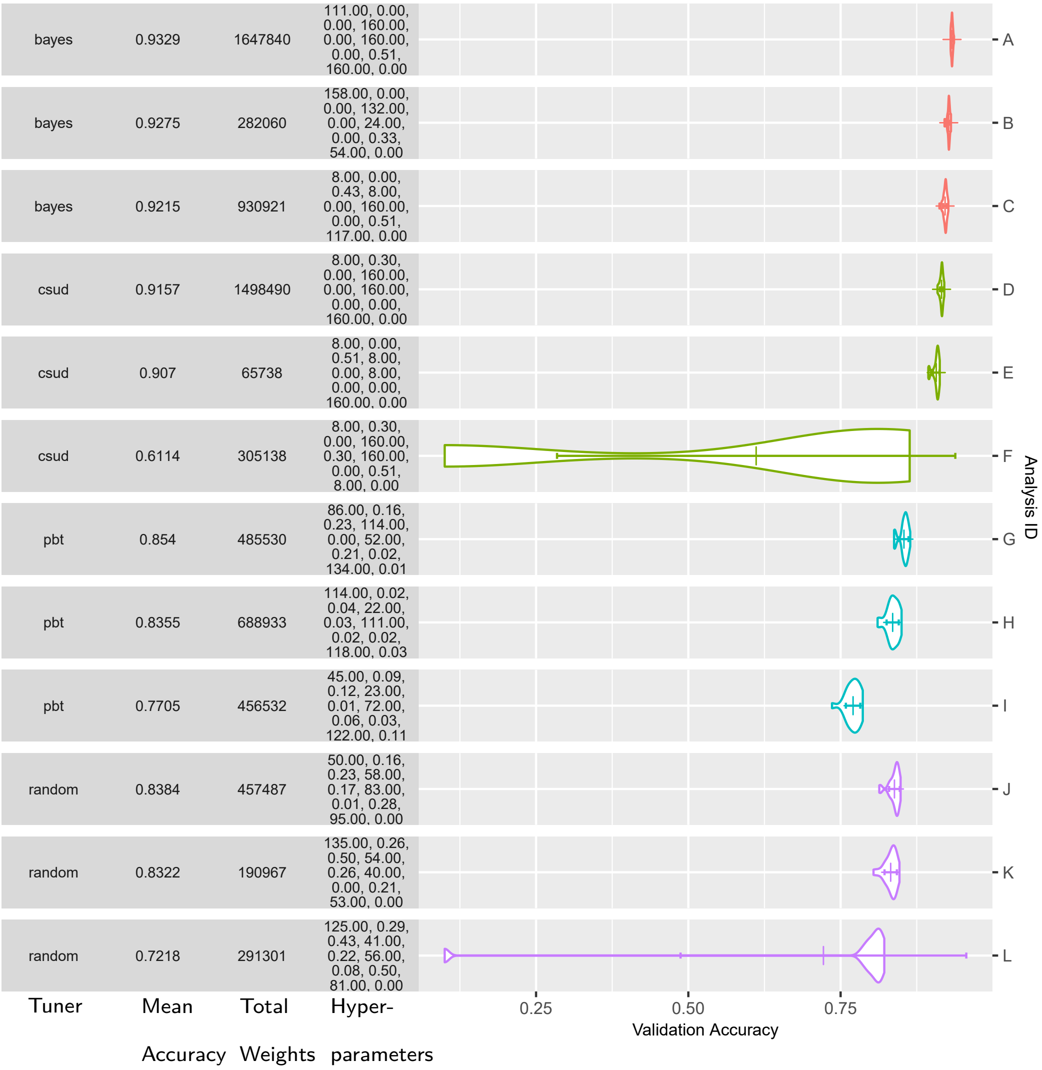
Fig. 13.4 shows model B ANN hyper-parametrisations discovered after three distinct tuning passes of each tuning algorithm, and displays validation accuracy violin plots produced consequent to 25 training sessions for each tuning pass discovered hyper-parametrisation. The columns on the left display from left to right, tuning algorithm, mean validation accuracy, total trainable weights, and hyper-parameter configuration.

The hyper-parameters column display order is 1_conv, 1_conv_L2, 1_do, $2a\_conv$, $2a\_conv\_L2$, $2b\_conv$, $2b\_conv\_L2$, 2_do, 3_d, 3_d_l2, with layer abbreviations as discussed in Fig. 13.1. Therefore the only difference between Fashion MNIST models A and B is the addition of L2 regularisation error, applied to the three convolutional layers 1_conv, $2a\_conv$, $2b\_conv$, and to the dense pre-output layer 3_d.

In model B, the application of dropout and L2 regularisation within the same ANN architecture may not be the best design. Typically, either dropout or L2 regularisation may be applied, as the two methods could create adverse interactions, leading to poor learning. However, it is investigated to what extent the presented tuning algorithms would be able to negotiate this type of challenging architecture.

In Fig. 13.4, the violin plots on the right present validation accuracy densities corresponding to 25 training sessions with mean validation accuracy $\pm 1SE$ marked in each violin plot. Results have been grouped by tuning algorithm, and then in order of descending mean validation accuracy.

Note that with the complex Fashion MNIST model B, dropout versus L2 regularisation ANN architecture, the ASHA Bayes and CSUD tuners, with the exception of CSUD analysis ID F, produce higher mean validation accuracy and tighter violin plot clusters.



Tuner	Mean Accuracy	Total Weights	Hyper-parameters
bayes	0.9329	1647840	111.00, 0.00, 0.00, 160.00, 0.00, 160.00, 0.00, 0.51, 160.00, 0.00
bayes	0.9275	282060	158.00, 0.00, 0.00, 132.00, 0.00, 24.00, 0.00, 0.33, 54.00, 0.00
bayes	0.9215	930921	8.00, 0.00, 0.43, 8.00, 0.00, 160.00, 0.00, 0.51, 117.00, 0.00
csud	0.9157	1498490	8.00, 0.30, 0.00, 160.00, 0.00, 160.00, 0.00, 0.00, 160.00, 0.00
csud	0.907	65738	8.00, 0.00, 0.51, 8.00, 0.00, 8.00, 0.00, 0.00, 160.00, 0.00
csud	0.6114	305138	8.00, 0.30, 0.00, 160.00, 0.30, 160.00, 0.00, 0.51, 8.00, 0.00
pbt	0.854	485530	86.00, 0.16, 0.23, 114.00, 0.00, 52.00, 0.21, 0.02, 134.00, 0.01
pbt	0.8355	688933	114.00, 0.02, 0.04, 22.00, 0.03, 111.00, 0.02, 0.02, 118.00, 0.03
pbt	0.7705	456532	45.00, 0.09, 0.12, 23.00, 0.01, 72.00, 0.06, 0.03, 122.00, 0.11
random	0.8384	457487	50.00, 0.16, 0.23, 58.00, 0.17, 83.00, 0.01, 0.28, 95.00, 0.00
random	0.8322	190967	135.00, 0.26, 0.50, 54.00, 0.26, 40.00, 0.00, 0.21, 53.00, 0.00
random	0.7218	291301	125.00, 0.29, 0.43, 41.00, 0.22, 56.00, 0.08, 0.50, 81.00, 0.00

FIGURE 13.4: Model B tuning passes and validation accuracy from 25 repeated training sessions. Configuration on the left. Violin plots with mean validation accuracy ±1SE on the right. Bayesian and CSUD tuners perform well while random and PBT tuners, which rely on pure randomisation do less well.

In contrast the PBT and ASHA random tuners, which solely rely on random search, produce substantially lower mean validation accuracy results.

Further note that the best performing ASHA Bayes ANN hyper-parametrisations all favour dropout over L2 regularisation. A review of ASHA Bayes analysis IDs A, B, and C reveals that for all cases, and in all layers, L2 regularisation is set to 0. Also dropout is used sparingly with analysis ID B and C only exhibiting high stage 2 dropout 2_d of 0.51 and 0.33 respectively. Finally, note that ASHA Bayes tuned architectures tend to favour layer sizes at range boundaries. ASHA Bayes analysis ID A produces the largest network with 1, 647, 840 trainable weights, and with the highest mean validation accuracy of 0.9329. Almost all layers for case analysis ID *A* have the layer maximum of 160 neurons.

CSUD tuner produced ANN hyper-parametrisations demonstrate greater hyper-parameter and mean validation accuracy variability. All CSUD hyper-parametrisations exhibit boundary range values. Layer neurons are either at 8 or 160, L2 regularisation is either 0 or 0.30, and dropout is either 0 or 0.51. CSUD tuning passes consisted of 60 iterations. The above results suggest that possibly 60 iterations is not sufficient to achieve convergence, and that instead boundary range oscillations are being obtained. It could also be that for the complex model B architecture, range boundary hyper-parameter values give better results. These possibilities are not discussed further here.

CSUD analysis ID *F* produces a violin plot with large variation and a mean validation accuracy of only 0.6114. The hyper-parametrisation for this case reveals mixed use of dropout and L2 regularisation. For example $1\_conv\_L2 = 0.30$ and $2a\_conv\_L2 = 0.30$, while $2\_do = 0.51$. Further the initial convolution layer 1_conv, and the dense pre-output layer 3_d consist of only 8 neurons. It appears this ANN architecture may be quite sensitive to the initial random ANN weights, leading to wide variations in training, and consequently in mean validation accuracy.

CSUD analysis IDs *D* and *E* produce opposing models, with case *D* favours L2 regularisation, while case *E* favours dropout. CSUD case*E* produces, with 65, 738 trainable weights, the smallest neural network. Smaller ANNs typically have a smaller memory footprint and faster inference speed. CSUD case *E* provides an interesting model for circumstances, where inference speed and mean validation accuracy trade-offs are acceptable.

The PBT and ASHA random tuners yield ANN hyper-parametrisations primarily with inside-boundary values; with mixed dropout and L2 regularisation. The resulting models produce comparatively lower mean validation accuracy outcomes, suggesting that the ASHA Bayes and CSUD tuners produce hyper-parametrisations with commensurately improved mean validation accuracy. It is believed the comparatively improved performance of the ASHA Bayes and CSUD tuners results from the use of additional distributional information for the former, and validation loss gradient for the latter tuner.

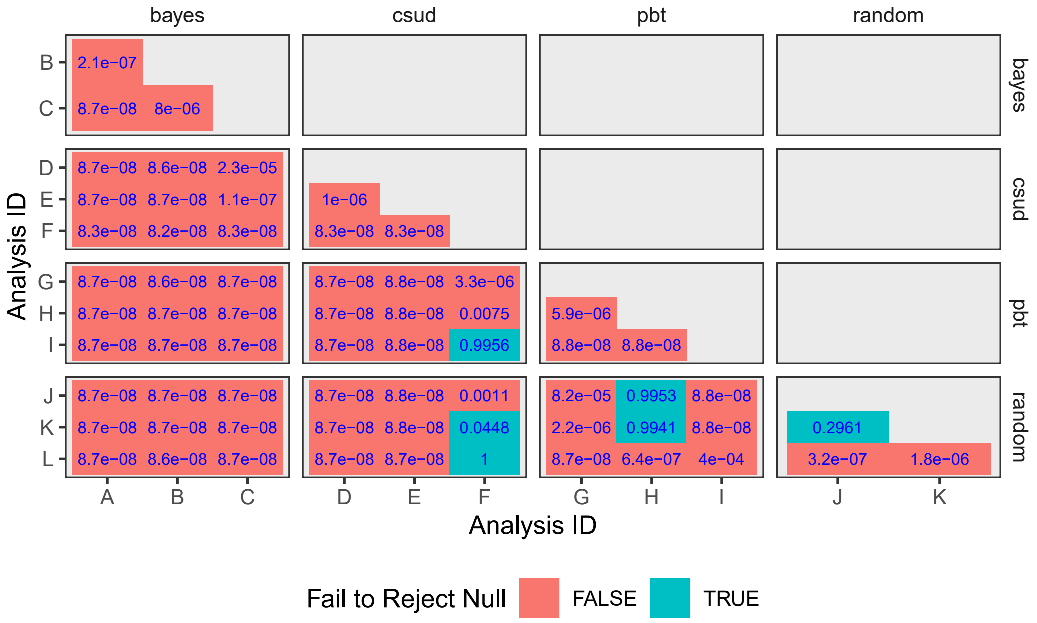


FIGURE 13.5: Fashion MNIST model B pairwise comparison matrix of mean validation accuracy with p-values using the DSCF criterion. Rows and columns are grouped by tuning algorithm. P-values obtained from pairwise comparison of mean validation accuracy, using 25 separate training runs. Cells with p-values less than 0.01, where the null hypothesis of equality of mean validation accuracy is rejected, are coloured in red. Both cross-group and intra-group pairwise mean validation accuracy comparisons generally show statistically significant differences at significance level 0.01. Details in text.

For the more complex model B Fashion MNIST ANN, both the ASHA Bayes and CSUD tuners offer an improvement over undirected random search.

Are the differences in mean validation accuracy statistically significant? Fig. 13.5 shows model B pairwise comparison matrix of mean validation accuracy with p-value results using the DSCF all-pairs tests (Konietschke et al., 2015). Rows and columns are grouped by tuning algorithm. Each tuning algorithm grouping contains 3 passes with randomly initialised weights. Each cell contains p-values resulting from pairwise comparison of equality of mean validation accuracy obtained from 25 separate training runs conducted at the respective hyper-parametrisations reported in Fig. 13.4. Cells with p-values less than 0.01, where the null hypothesis of equality of mean validation accuracy is rejected, are coloured in red.

Fig. 13.5 indicates that both *cross-group* and *intra-group* pairwise mean validation accuracy comparisons generally show statistically significant differences at significance level 0.01. Fig. 13.4 shows that the outlier CSUD case

F has large standard error. Consequently one fails to reject the null hypothesis of equal mean validation accuracy between CSUD case *F* and PBT case *I*; and CSUD case *F* and ASHA random cases *K*, or *L*. One also fails to reject the null hypothesis of equal mean validation accuracy between PBT case *H* and ASHA random cases *J*, or *K*; and between ASHA random case *J* and *K*.

However given the above exceptions, the remaining pairwise mean validation accuracy comparisons support the notion that in general the tuners produce hyper-parametrisations with mean validation accuracy outcomes, which are statistically significantly different. Hence, with the more complex model B Fashion MNIST architecture, even repeated applications of the same tuner may produce ANN configurations, which produce statistically significantly different mean validation accuracy, implying that these configurations can be regarded as distinct possibilities.

Furthermore in general, mean validation accuracy results statistically significantly differ among the different tuners, with ASHA Bayes producing the highest, and CSUD producing the second highest mean validation accuracy outcomes.

13.6 Discussion

Two Fashion MNIST ANN architectures are considered, model A and model B, differing by the addition of L2 regularisation loss to the latter. The presence of both dropout and L2 regularisation make model B more difficult to tune with respect to validation loss.

ASHA Bayes, ASHA random, CSUD, and PBT tuners are employed. For each tuning algorithm, model A or model B architectures are tuned three times. Subsequently, the resulting ANN hyper-parametrisations are trained 25 times with random initial weights, and then after training, validation accuracy is assessed.

Looking at mean validation accuracy, it is found that ASHA random performs well for the simple model A, but poorly for the more complex model B, in the sense that under identical resource constraints as indicated in Table 13.2, subsequent validation accuracy results are worse for model B. The PBT tuner performance slightly improves in relative terms for model B. In contrast ASHA Bayes performs well in tuning both models A and B, while CSUD performance improves as model complexity increases. However, for both model A and B, CSUD validation accuracy and hyper-parametrisation results show relatively higher variation.

Unlike ASHA random and PBT, which use resource management with random search, both ASHA Bayes and CSUD employ additional information, which increases algorithm specificity; this increased specificity appears to help in achieving good tuning results with the more complex model B. ASHA Bayes uses iteratively evolving Gaussian mixtures, which increasingly focus on hyper-parameter regions associated with lower validation loss. The CSUD search strategy uses iteratively improving validation loss gradient approximations to drive hyper-parameter updates in the direction of lower validation loss.

Between ASHA Bayes and the CSUD implementation however, based on the results, ASHA Bayes appears to perform better in terms of mean validation accuracy, especially as model complexity increases. It is proposed this is because of the more greedy exploitation behaviour of ASHA Bayes, which tends to produce more condensed posterior distributions. In contrast, as discussed in section 12.4, the CSUD search strategy specification may not lead to discovery of a unique optimised hyper-parameter selection. Under such circumstances of non-uniqueness, given that CSUD approximates both the loss gradient and the hyper-parameter updates, it may take longer to converge; this lengthened convergence lag may in the short run, lead to increased variation in hyper-parameter selections as has been observed.

Due to its convergence lag, it is possible that CSUD will provide more comprehensive coverage than ASHA Bayes in the sense of visiting a larger area of the hyper-parameter search space.

When compared to ASHA Bayes, the implementation of CSUD is more computationally intensive due to requiring two loss function measurements per tuning iteration for gradient approximation. Another cp-SPSA variant called constrained measurement reuse with single-sided perturbations (CMR), derived from Algorithm 1 (SPSA2-1UR, measurement reuse) introduced in Abdulla and Bhatnagar (2006) has been tested elsewhere. CMR reduces the number of loss function measurements from two to only one measurement per iteration after the initial iteration.

However, in terms of asymptotic efficiency, the measurement reuse SPSA2-1UR algorithm (Abdulla & Bhatnagar, 2006) and by extension CMR, perform worse than single-sided and double-sided SPSA, both of which are components of CSUD. Despite hyper-parameter range constraints, CMR while requiring less computation time, has not performed better than the CSUD tuner presented here. It is suspected this may be due to CMR's decreased asymptotic efficiency. Hence, reducing the computational demands of CSUD, while retaining or improving its tuning outcomes could benefit from further research.

In sum, the CSUD search strategy shows some promise as an ANN tuner, but more research is needed to determine in which specific area and in what manner it may be best employed.

Chapter 14

Conclusion

This thesis contributes to the field of nonrational computational technologies. Inspired by human Iowa Gambling Task (IGT) outcomes, and the role of the vmPFC and emotion in human decision making, this work uses such IGT outcomes to automatically calibrate and assess nonrational Q-learning models. This automatic calibration is conducted via CSUD search, a nonrational search technology additionally developed in this work.

The concept of nonrationality is expressed via two different pathways. In Q-learning, nonrationality is formulated as exponential learning rate decay coupled with a finite but unknown time horizon. In CSUD search, on the other hand, nonrationality is modelled as a departure from stochastic optimisation leading to a search algorithm, which satisfices in the sense of foregoing theoretical guarantees of finding a global minimum.

Simulation based IGT Q-learning modelling is conducted, where the key model parameters consisting of the initial learning rate, learning rate decay, and exploration are automatically calibrated by CSUD search so as to produce simulated mean fraction of good deck $\bar{f}_G$ outcomes residing in corresponding normal versus vmPFC impaired human outcome catchment zones. Q-learning exploration is modelled using a variety of techniques including the ε-Greedy and Boltzmann rules.

Simulation results produce two major findings, (a) high learning rate decay leads to VMF impairment commensurate outcomes, and (b) to match human outcomes exploration must be high. A further, more detailed key result is demonstrated with respect to the joint assessment of the original, re-shuffled, and random IGT environments. Based on jitter plot assessment and np-M/ANOVA analysis, it is shown that neither ε-Greedy nor Boltzmann exploration appear to provide a convincing view of human exploration. In jitter plots, ε-Greedy software agents show bi-modal f_G densities, however, with little or no mass on $\bar{f}_G$. In contrast, Boltzmann agents achieve unimodal f_G densities, with some mass at $\bar{f}_G$, while failing to achieve an in-catchment zone $\bar{f}_G$ match for the random IGT. Based on jitter plots, ε-Greedy exploration appears too dispersed, while Boltzmann exploration appears too concentrated.

For ε-Greedy exploration, simulation f_G outcome np-M/ANOVA normal versus VMF impaired factor analysis produces, as expected from corresponding human results, a failure to reject the null hypothesis of no factor effects in the re-shuffled IGT. A corresponding failure to reject this factor effect null, however, does not obtain with Boltzmann exploration. Finally, the chapter 10 burst learning model is able to produce unimodal f_G jitter plots for normal configured ε-Greedy agents in the original and random IGT environments. This appears to suggest that ε-Greedy exploration, at least with respect to the IGT, proposes a better representation of human exploration.

As has been noted in the cognitive models in section 2.4 and elsewhere (Daw, 2011; Findling et al., 2019), Boltzmann exploration is commonly employed in psychological models of decision making. The results here question why Boltzmann exploration, other than for being rationally pleasant, is actually used in decision making.

The simulation methodology employed here, however, differs from the maximum likelihood based individual fitting techniques used in psychology (Piray et al., 2019; Wilson & Collins, 2019). The methodology employed here uses representative agents, which are calibrated using averaged human outcomes. It would be interesting to see to what extent the current results can be achieved in human driven verification studies, where for example learning rate decay and ε-Greedy exploration are assessed via maximum likelihood. If, for example, human exploration had directed and random components (Wilson et al., 2014), then one might expect ε-Greedy exploration to be able to provide a generalist catch-all exploration implementation.

It is believed that the key results reported above provide evidence of heuristic learning strategies. An exponentially decaying learning rate provides a learning freeze, in effect enforcing a finite time horizon. On the other hand, in the context of the No Free Lunch theorems (Wolpert & Macready, 1997), high exploration can be seen as a way to maintain some generalised search capability without specifically committing to a particular search algorithm. Hence, high exploration can be a rule of thumb to mitigate incomplete probabilistic information. Wilson et al., 2014 state that directed exploration may be used increase learning from the more informative options. In a statistical context, as is well known in Monte Carlo sampling, the only way to assess the tail of a distribution is by drawing more samples. In a finite choice context, such a strategy would manifest itself as high exploration.

From an algorithmic perpective, the nonrational burst learning model is interesting because the re-setting learning rate mechanics may provide an alternative to the more computationally expensive but accurate inverse Hessian driven learning rates (step sizes). This potential of the burst model, however, needs additional research.

CSUD and CSUD search provide further exciting computational methods for assessing regions of a model space while ensuring that the assessed area is constrained to avoid numerical under- or overflow errors, illegal values, or overlap of the to be assessed input value spaces. Compared to analytical gradient descent, CSUD is computationally more costly as gradient approximation requires two loss evaluations, and input and perturbation constraints must be checked at each iteration. The use of nonrational CSUD search is tested in an ANN tuning contest with mixed performance outcomes. It appears that the use of CSUD to embed gradient information may produce models, which perform better than those derived via PBT or ASHA random search; but this result is only obtained with a more complex model. ASHA Bayes tuning, however, produces results, which are better than those of CSUD search tuning. The further study of CSUD search in ANN tuning could be of interest.

Finally, it would be of interest to investigate the properties of models, which use burst learning style learning rate re-setting for both parameter optimisation and hyper-parameter tuning. While this is a natural extension of the current work, it has not been attempted.

This brings the readers to the end of this work. The author thanks the readers for their interest and patience, and hopes that some of the ideas presented here will prove to be of value.

Appendix A

IGT Yield Structures

The yield structures for the original, re-shuffled, reversed IGT, and SGT are presented below. Rewards are coloured black, whereas fines appear in red boxes.

A.1 Original IGT

Decks	A	B	C	D
Reward	100	100	50	50
Order				
1				
2				
3	150		50	
4				
5	300		50	
6				
7	200		50	
8				
9	250	1250	50	
10	350		50	250
1 1				
12	350		25	
13			75	
14	250	1250		
15	200			
16				
17	300		25	
18	150		75	
19				
20			50	250
21		1250		
22	300			
23				
24	350		50	
25			25	
26	200		50	
27	250			
28	150			
29			75	250
30			50	
3 1	350			
32	200	1250		
33	250			
34			25	
35			25	250
36				
37	150		75	
38	300			
39			50	
40			75	

Source: Bechara et al., 1994, p. 9.

A.2 Re-shuffled IGT

Decks	A	B	C	D
Reward	100	100	50	50
Order				
1	250	1250	50	
2	350		50	250
3		1250		
4	350		25	
5			75	
6	250			
7	200			
8				
9	300		25	
10	150		75	
11				
12			50	250
13		1250		
14	300			
15				
16	350		50	
17			25	
18	200		50	
19	250			
20	150			
21			75	250
22			50	
23	350			
24	200	1250		
25	250			
26			25	
27			25	250
28				
29	150		75	
30	300			
31			50	
32			75	
33				
34				
35	150		50	
36				
37	300		50	
38				
39	200		50	
40				

Source: Fellows and Farah, 2005, p. 59, described in *Tasks*.

A.3 Random IGT

The random IGT is implemented as a randomised (without replacement) version of A.1. When the random pool for a deck is deleted it is re-initialised.

A.4 Reversed IGT

Decks	E	F	G	H	
Fine	100	50	100	50	
Order					
1					
2			350		
3	1250				
4		25	250		
5		50			
6			300		
7			200		
8		75		250	
9		25	150		
10		75			
11	1250	50			
12					
13		25	350		
14					
15			250		
16		25			
17		75	200		
18			150		
19					
20		75	300	250	

Decks	E	F	G	H	
Reward	100	50	100	50	
Order					
21	1250				
22			300		
23					
24		25	350		
25		75			
26		50	150		
27			200		
28			250		
29		75			
30		25		250	
31			150		
32			200		
33	1250		350		
34		50		250	
35		50			
36					
37		25	200		
38			350		
39		75			
40		50			

Source: Bechara et al., 2000, p. 2193.

A.5 Soochow Gambling Task (SGT)

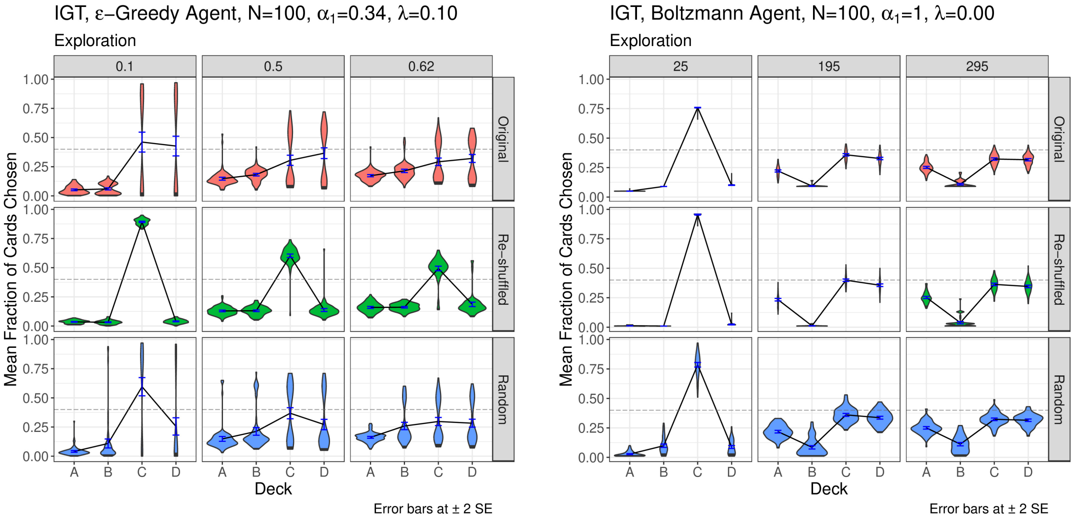
Decks	A	B	C	D
Fixed Yield	200	100	200	100
Order				
1				
2				
3				
4				
5	1250	750	1250	750
6				
7				
8				
9				
10	1250	750	1250	750
11				
12				
13				
14				
15	1250	750	1250	750
16				
17				
18				
19				
20	1250	750	1250	750

Decks	E	F	G	H
Fixed Yield	200	100	200	100
Order				
21				
22				
23				
24				
25	1250	750	1250	750
26				
27				
28				
29				
30	1250	750	1250	750
31				
32				
33				
34				
35	1250	750	1250	750
36				
37				
38				
39				
40	1250	750	1250	750

Source: Adapted from Chiu et al., 2008, Table 1. For comparison purposes with the other IGT environments, the table above reports both rewards and fines. Note that the random beneficial or adverse outcome occurs with probability 0.2.

Appendix B

Deck Looping Behavior



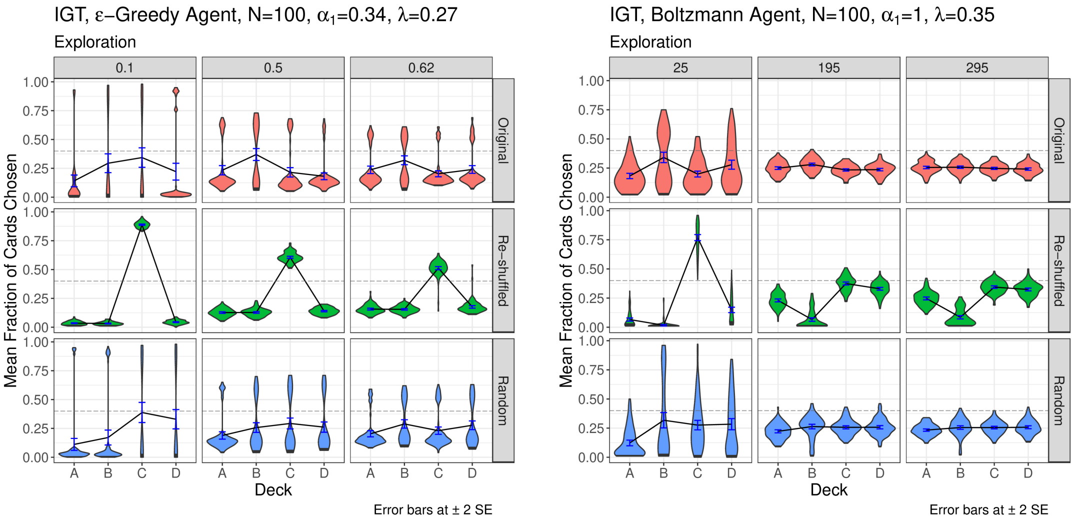
(A) ε-Greedy Agent

(B) Boltzman Agent

FIGURE B.1: Normal agents deck looping versus exploration at mean fraction of cards per deck at 100th draw. The dashed line represents the 40th card. In general, as exploration increases, deck looping decreases on average, but especially for the ε-Greedy case, some agents continue to loop. As expected from the soft-max constraint, Boltzmann agent violin plots approximate a unimodal symmetric distribution, while ε-Greedy agent violin plots tend to bi-modality and skewness for most decks and IGT variants.

In this appendix, Fig. B.1 to Fig. B.3 depict observed deck looping behaviour for simple ε-Greedy and Boltzmann software agents, and for normal (control) random IGT subjects.

Fig. B.1 depicts normal behaviour configured software agent deck looping behaviour. Fig. B.2 depicts vmPFC impaired behaviour configured software agent deck looping behaviour. In general, in software agents, as exploration increases, deck looping decreases.



(A) ε-Greedy Agent

(B) Boltzman Agent

FIGURE B.2: vmPFC Impaired agents deck looping versus exploration at mean fraction of cards per deck at 100th draw. The dashed line represents the 40th card. In general, as exploration increases, deck looping decreases on average, but especially for the ε-Greedy case, some agents continue to loop. As expected from the soft-max constraint, Boltzmann agent violin plots approximate a unimodal symmetric distribution, while ε-Greedy agent violin plots tend to bimodality and skewness for most decks and IGT variants.

In Fig. B.3 in the random IGT variant in Steingroever et al. (2018), as indicated by the violin plot mass extending above the dashed line at 0.40, individual healthy subjects exhibit looping behaviour in decks B, C, and D.

Comparison of violin plot shapes shows that in the random IGT variant, human violin plot behaviour is somewhere between that of the ε-Greedy and Boltzmann agents with unimodal shapes skewed in the direction of looped decks. Further research is needed to characterise software agent and human subject violin plot behaviour across the different IGT variants.

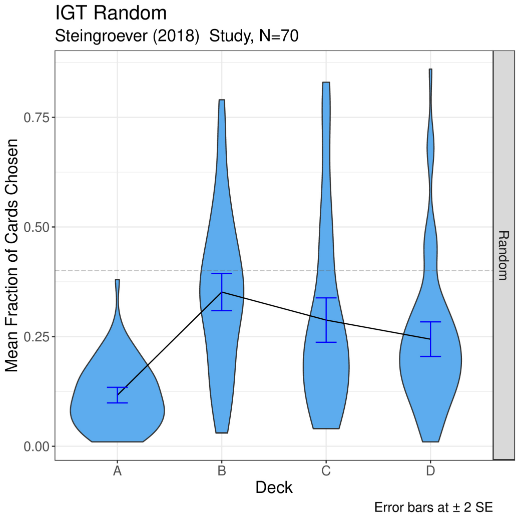


FIGURE B.3: Steingroever et al. (2018), healthy subjects, mean fraction of cards per deck at 100th draw. The dashed line represents the 40th card. On average, for each deck, the mean fraction of cards chosen is below 0.40. However, with the exception of Deck A, skewed unimodal violin plots indicate that some proportion of individual subjects exhibit deck looping. Deck looping is most notable for Deck B.

Appendix C

Learning Rate Decay Grid Construction

Given (3.9b), one wishes to construct a finite size learning rate decay λ grid such that the grid passes though λ_a and λ_b, with f samples before λ_a, g samples between λ_a and λ_b, and h samples after λ_b. A grid is generated by decrementing or incrementing λ proportionally. This operation produces a grid, which is dense at areas of rapid change in the mean fraction of good decks $\bar{f}_G$ metric. This task is achieved with the following Scala algorithm.

```scala
def sampleFromIntervalWithPowerDecay(x1: Double, x2: Double, pre: Int, steps: Int, post: Int): Vector[Double] = {
  require (x1 > 0 && x2 > 0, "x1 $x1 and x2 $x2 must be positive")
  require(x2 > x1, s"x2 $x2 must be greater than x1 $x1")
  def increase(v: Double, by: Double): Double = v*by
  def decrease(v: Double, by: Double): Double = v/by
  val pStep = math.pow(x2/x1, 1d/steps)
  var buffer = x1
  val seq2 = Seq(x1) ++ (for (_ <- 1 to steps + post) yield {
    buffer = increase(buffer, pStep)
    buffer
  })
  buffer = x1
  val seq1 = for (_ <- 1 to pre) yield {
    buffer = decrease(buffer, pStep)
    buffer
  }
  (seq1.sorted ++ seq2).toVector
}
sampleFromIntervalWithPowerDecay(a, b, f, g, h)
```

Appendix D

CSUD Search Strategy a.s. Convergence: Error Term Bounds

For any two independently drawn $\hat{\Theta}^a, \hat{\Theta}^b \in \mathbb{Z}_k$ and $e \in \{S+, S-, D\}$, consider the expectation term $E\left[\left(L(\hat{\Theta}^a) - L(\hat{\Theta}^b) + \epsilon^a - \epsilon^b\right)^2\right]$. It is necessary to establish some bounds with reference to the loss function in (12.46). The expectation term is expanded to get

$$\begin{aligned} E\Big[&\left(L(\hat{\Theta}^a) - L(\hat{\Theta}^b)\right)^2 + \left(\epsilon^a - \epsilon^b\right)^2 \\ &+2\left(L(\hat{\Theta}^a) - L(\hat{\Theta}^b)\right)\left(\epsilon^a - \epsilon^b\right)\Big] \end{aligned} \tag{D.1a}$$

$$\begin{aligned} L(\hat{\Theta}^a) &= \left(m(\hat{\Theta}^a) - m^*\right)' S \left(m(\hat{\Theta}^a) - m^*\right), \\ L(\hat{\Theta}^b) &= \left(m(\hat{\Theta}^b) - m^*\right)' S \left(m(\hat{\Theta}^b) - m^*\right) \end{aligned} \tag{D.1b}$$

$$\begin{aligned} \epsilon^a &= \varepsilon^{a'} S \varepsilon^a + 2\varepsilon^{a'} S \left(m(\hat{\Theta}^a) - m^*\right), \\ \epsilon^b &= \varepsilon^{b'} S \varepsilon^b + 2\varepsilon^{b'} S \left(m(\hat{\Theta}^b) - m^*\right). \end{aligned} \tag{D.1c}$$

The cross-term in (D.1a) is considered first. Since $L(\cdot)$ and ϵ are independent, the cross term evaluates to

$$\begin{aligned} &E\left[L(\hat{\Theta}^a) - L(\hat{\Theta}^b)\right] E\left[\epsilon^a - \epsilon^b\right] \\ &\qquad = E\left[L(\hat{\Theta}^a) - L(\hat{\Theta}^b)\right]\left(tr(S\Sigma) - tr(S\Sigma)\right) = 0. \end{aligned} \tag{D.2}$$

Next consider the expectation of the squared error difference term in (D.1a): $E\left[\left(\epsilon^a - \epsilon^b\right)^2\right] = E\left[\epsilon^{a^2} + \epsilon^{b^2}\right] - 2E\left[\epsilon^a\right] E\left[\epsilon^b\right] < E\left[\epsilon^{a^2} + \epsilon^{b^2}\right]$, since ϵ^a and ϵ^b are independent and $E\left[\epsilon^a\right] = E\left[\epsilon^b\right] = tr(S\Sigma) > 0$. The independence of ϵ^a and ϵ^b follows from the error structure in (D.1c), $\varepsilon \sim$ *i.i.d.* Further $\hat{\Theta}^a$, $\hat{\Theta}^b \sim$ *independently* given the construction of perturbations in Assumption

12.4.6. Next expand the squared error terms

$$\begin{aligned}\epsilon^{a^2}+\epsilon^{b^2} &= \left(\varepsilon^{a'}S\varepsilon^{a}+2\varepsilon^{a'}S\left(m(\hat{\Theta}^{a})-m^{*}\right)\right)^{2}\\ &\quad+\left(\varepsilon^{b'}S\varepsilon^{b}+2\varepsilon^{b'}S\left(m(\hat{\Theta}^{b})-m^{*}\right)\right)^{2}\end{aligned}\tag{D.3a}$$

$$\begin{aligned}&=\left(\varepsilon^{a'}S\varepsilon^{a}\right)^{2}+4\varepsilon^{a'}S\varepsilon^{a}\varepsilon^{a'}S\left(m(\hat{\Theta}^{a})-m^{*}\right)\\ &\quad+\left(\varepsilon^{b'}S\varepsilon^{b}\right)^{2}+4\varepsilon^{b'}S\varepsilon^{b}\varepsilon^{b'}S\left(m(\hat{\Theta}^{b})-m^{*}\right)\\ &+4\left(\varepsilon^{a'}S\left(m(\hat{\Theta}^{a})-m^{*}\right)\right)^{2}+4\left(\varepsilon^{b'}S\left(m(\hat{\Theta}^{b})-m^{*}\right)\right)^{2}\end{aligned}\tag{D.3b}$$

$$=\sum_{i=1}^{r}s_i^2\left(\varepsilon_i^{a^4}+\varepsilon_i^{b^4}\right)+\sum_{i=1}^{r}\sum_{j\neq i}^{r}s_is_j\left(\varepsilon_i^{a^2}\varepsilon_j^{a^2}+\varepsilon_i^{b^2}\varepsilon_j^{b^2}\right)\tag{D.3c}$$

$$\begin{aligned}&+4\sum_{i=1}^{r}s_i^2\varepsilon_i^{a^3}\left(m_i(\hat{\Theta}^{a})-m_i^{*}\right)+4\sum_{i=1}^{r}s_i^2\varepsilon_i^{b^3}\left(m_i(\hat{\Theta}^{b})-m_i^{*}\right)\\ &\quad+4\sum_{i=1}^{r}\sum_{j\neq i}^{r}s_is_j\varepsilon_i^{a^2}\varepsilon_j^{a}\left(m_j(\hat{\Theta}^{a})-m_j^{*}\right)\\ &\qquad+4\sum_{i=1}^{r}\sum_{j\neq i}^{r}s_is_j\varepsilon_i^{b^2}\varepsilon_j^{b}\left(m_j(\hat{\Theta}^{b})-m_j^{*}\right)\end{aligned}\tag{D.3d}$$

$$\begin{aligned}&+4\sum_{i=1}^{r}s_i^2\varepsilon_i^{a^2}\left(m_i(\hat{\Theta}^{a})-m_i^{*}\right)^{2}+4\sum_{i=1}^{r}s_i^2\varepsilon_i^{b^2}\left(m_i(\hat{\Theta}^{b})-m_i^{*}\right)^{2}\\ &\quad+4\sum_{i=1}^{r}\sum_{j\neq i}^{r}s_is_j\varepsilon_i^{a}\varepsilon_j^{a}\left(m_i(\hat{\Theta}^{a})-m_i^{*}\right)\left(m_j(\hat{\Theta}^{a})-m_j^{*}\right)\\ &\qquad+4\sum_{i=1}^{r}\sum_{j\neq i}^{r}s_is_j\varepsilon_i^{b}\varepsilon_j^{b}\left(m_i(\hat{\Theta}^{b})-m_i^{*}\right)\left(m_j(\hat{\Theta}^{b})-m_j^{*}\right).\end{aligned}\tag{D.3e}$$

In (D.3b) to (D.3e), re-group and colour code as follows: black for error terms, blue and gray for vanishing terms, and red for non-negative moment-bounded cross-terms.

Taking expectations and using Assumption 12.4.2, the gray coloured cross-terms and blue coloured 3rd order error terms vanish to yield

$$E\left[\left(\epsilon^{a}-\epsilon^{b}\right)^{2}\right]\leq E\left[\epsilon^{a^2}+\epsilon^{b^2}\right]=2\mathbb{E}\tag{D.4}$$

where

$$\mathbb{E}=tr\left(S^2\tilde{\Sigma}+S^2\Sigma\tilde{\mathbb{M}}\right)+\sum_{i=1}^{r}\sum_{j\neq i}^{r}s_is_j\sigma_i^2\sigma_j^2<\infty\tag{D.5a}$$

$$
\begin{aligned}
S^2 &= \begin{bmatrix} s_1^2 & & \\ & \ddots & \\ & & s_r^2 \end{bmatrix}, \\
\tilde{\Sigma} &= \begin{bmatrix} \sigma_1^4 & & \\ & \ddots & \\ & & \sigma_r^4 \end{bmatrix}, \quad \Sigma = \begin{bmatrix} \sigma_1^2 & & \\ & \ddots & \\ & & \sigma_r^2 \end{bmatrix}
\end{aligned}
\tag{D.5b}
$$

$$
\begin{aligned}
\tilde{\mathbb{M}} &= \begin{bmatrix} \tilde{m}_1 & & \\ & \ddots & \\ & & \tilde{m}_r \end{bmatrix}, \\
E\left[\left(m_i(\hat{\Theta}^a) - m_i^*\right)^2\right], E&\left[\left(m_i(\hat{\Theta}^b) - m_i^*\right)^2\right] \leq \tilde{m}_i < \infty.
\end{aligned}
\tag{D.5c}
$$

Finally consider the squared loss terms. Since the loss functions in (D.1b) are non-negative

$$
E\left[\left(L(\hat{\Theta}^a) - L(\hat{\Theta}^b)\right)^2\right] \leq E\left[L(\hat{\Theta}^a)^2 + L(\hat{\Theta}^b)^2\right]. \tag{D.6}
$$

Next expand the right hand side of (D.6) to get

$$
\begin{aligned}
&L(\hat{\Theta}^a)^2 + L(\hat{\Theta}^b)^2 \\
&\qquad = \left(\sum_{i=1}^{r} s_i \left(m_i^a - m_i^*\right)^2\right)^2 + \left(\sum_{i=1}^{r} s_i \left(m_i^b - m_i^*\right)^2\right)^2 \\
&\qquad = \sum_{i=1}^{r} s_i^2 \left(m_i^a - m_i^*\right)^4 + \sum_{i=1}^{r} s_i^2 \left(m_i^b - m_i^*\right)^4 \\
&\qquad\quad + \sum_{i=1}^{r}\sum_{j\neq i}^{r} s_i s_j \left(\left(m_i^a - m_i^*\right)^2 \left(m_j^a - m_j^*\right)^2\right. \\
&\qquad\qquad\qquad \left. + \left(m_i^b - m_i^*\right)^2 \left(m_j^b - m_j^*\right)^2\right).
\end{aligned}
\tag{D.7}
$$

Taking expectations of (D.7), and using Assumption 12.4.2

$$
E\left[L(\hat{\Theta}^a)^2 + L(\hat{\Theta}^b)^2\right] \leq 2tr\left(S^2\tilde{\mathbb{M}}\right) + 2\sum_{i=1}^{r}\sum_{j\neq i}^{r} s_i s_j \tilde{m}_i \tilde{m}_j = 2\mathbb{L} \tag{D.8}
$$

where

$$\tilde{\tilde{\mathbb{M}}} = \begin{bmatrix} \tilde{\tilde{m}}_1 & & \\ & \ddots & \\ & & \tilde{\tilde{m}}_1 \end{bmatrix},$$

$$E\left[\left(m_i(\hat{\Theta}^a) - m_i^*\right)^4\right], E\left[\left(m_i(\hat{\Theta}^b) - m_i^*\right)^4\right] \leq \tilde{\tilde{m}}_i < \infty, \quad \text{(D.9)}$$

and S^2 and $\tilde{m}_i$ are from (D.5b) and (D.5c) respectively. It has now been determined that

$$E\left[\left(L(\hat{\Theta}^a) - L(\hat{\Theta}^b) + \epsilon^a - \epsilon^b\right)^2\right] \leq 2\mathbb{L} + 2\mathbb{E}. \quad \text{(D.10)}$$

Bibliography

Abadi, M., Agarwal, A., Barham, P., ... Zheng, X. (2015). TensorFlow: Large-Scale Machine Learning on Heterogeneous Systems [Software available from tensorflow.org]. https://www.tensorflow.org/

Abdulla, M., & Bhatnagar, S. (2006). SPSA Algorithms with Measurement Reuse. *Proceedings of the 2006 Winter Simulation Conference*, 320–328. https://doi.org/10.1109/WSC.2006.323089

Abedinia, O., & Amjady, N. (2016). Short-term Load Forecast of Electrical Power System by Radial Basis Function Neural Network and New Stochastic Search Algorithm. *International Transactions on Electrical Energy Systems*, *26*(7), 1511–1525. https://doi.org/10.1002/etep.2160

Agostini, A., & Celaya, E. (2010). Reinforcement Learning with a Gaussian Mixture Model. *The 2010 International Joint Conference on Neural Networks (IJCNN)*, 1–8. https://doi.org/10.1109/IJCNN.2010.5596306

Ahn, W.-Y., Busemeyer, J. R., Wagenmakers, E.-J., & Stout, J. C. (2008). Comparison of Decision Learning Models Using the Generalization Criterion Method. *Cognitive Science*, *32*(8), 1376–1402. https://doi.org/10.1080/03640210802352992

Ahn, W.-Y., Vasilev, G., Lee, S.-H., ... Vassileva, J. (2014). Decision-making in Stimulant and Opiate Addicts in Protracted Abstinence: Evidence from Computational Modeling with Pure Users. *Frontiers in Psychology*, *5*, 849. https://doi.org/10.3389/fpsyg.2014.00849

Alexander, J. C. (2013). *The Dark Side of Modernity*. Polity.

Antos, D., & Pfeffer, A. (2011). Using Emotions to Enhance Decision-making. *Proceedings of the Twenty-second International Joint Conference on Artificial Intelligence*, *1*, 24. http://ijcai.org/papers11/contents.php

Auer, P., Cesa-Bianchi, N., & Fischer, P. (2002). Finite-time Analysis of the Multiarmed Bandit Problem. *Machine Learning*, *47*(2-3), 235–256. https://doi.org/10.1023/A:1013689704352

Bathke, A. C., Harrar, S. W., & Madden, L. V. (2008). How to Compare Small Multivariate Samples Using Nonparametric Tests. *Computational Statistics and Data Analysis*, *52*(11), 4951–4965. https://doi.org/10.1016/j.csda.2008.04.006

Bechara, A., Damasio, H., Tranel, D., & Damasio, A. R. (1997). Deciding Advantageously before Knowing the Advantageous Strategy. *Science*, *275*(5304), 1293–1295. https://doi.org/10.1126/science.275.5304.1293

Bechara, A., Tranel, D., & Damasio, H. (2000). Characterization of the Decision-making Deficit of Patients with Ventromedial Prefrontal Cortex Lesions. *Brain*, *123*(11), 2189–2202.

Bechara, A., Damasio, A. R., Damasio, H., & Anderson, S. W. (1994). Insensitivity to Future Consequences Following Damage to Human Prefrontal Cortex. *Cognition*, *50*(13), 7–15. https://doi.org/10.1016/0010-0277(94)90018-3

Bechara, A., Damasio, H., Tranel, D., & Anderson, S. W. (1998). Dissociation Of Working Memory from Decision Making within the Human Prefrontal Cortex. *Journal of Neuroscience*, *18*(1), 428–437. https://doi.org/10.1523/JNEUROSCI.18-01-00428.1998

Bengio, Y. (2000). Gradient-Based Optimization of Hyperparameters. *Neural Computation*, *12*(8), 1889–1900. https://doi.org/10.1162/089976600300015187

Bergstra, J., Bardenet, R., Bengio, Y., & Kégl, B. (2011). Algorithms for Hyper-Parameter Optimization. In J. Shawe-Taylor, R. S. Zemel, & P. L. Bartlett (Eds.), *NIPS'11: Proceedings of the 24th International Conference on Neural Information Processing Systems*. Curran Associates Inc.

Bertsekas, D. P. (2012). *Dynamic Programming and Optimal Control* (4th ed.). Athena Scientific.

Bhatnagar, S., Prasad, H. L., & Prashanth, L. A. (2013). *Stochastic Recursive Algorithms for Optimization: Simultaneous Perturbation Methods* (Vol. 434). Springer.

Bishop, C. M. (2006). *Pattern Recognition and Machine Learning*. Springer.

Böhm, G., & Pfister, H.-R. (2008). The Multiplicity of Emotions: A Framework of Emotional Functions in Decision Making. *Judgment and Decision Making*, *3*(1), 5–17.

Brand, M., Recknor, E. C., Grabenhorst, F., & Bechara, A. (2007). Decisions under Ambiguity and Decisions under Risk: Correlations with

Executive Functions and Comparisons of Two Different Gambling Tasks with Implicit and Explicit Rules. *Journal of Clinical and Experimental Neuropsychology*, *29*(1), 86–99. https://doi.org/10.1080/13803390500507196

Broekens, J., Jacobs, E., & Jonker, C. M. (2015). A Reinforcement Learning Model of Joy, Distress, Hope and Fear. *Connection Science*, *27*(3), 215–233.

Bubeck, S., & Cesa-Bianchi, N. (2012). Regret Analysis of Stochastic and Nonstochastic Multi-armed Bandit Problems. https://doi.org/10.48550/ARXIV.1204.5721

Burchett, W. W., Ellis, A. R., Harrar, S. W., & Bathke, A. C. (2017). Nonparametric Inference for Multivariate Data: The R Package npmv. *Journal of Statistical Software*, *76*(4), 1–18. https://doi.org/10.18637/jss.v076.i04

Busemeyer, J. R., & Stout, J. C. (2002). A Contribution of Cognitive Decision Models to Clinical Assessment: Decomposing Performance on the Bechara Gambling Task. *Psychological Assessment*, *14*(3), 253–262. https://doi.org/https://dx.doi.org/10.1037/1040-3590.14.3.253

Camus, A. (1962). *The Rebel*. Penguin Books.

Cesa-Bianchi, N., Gentile, C., Lugosi, G., & Neu, G. (2017). Boltzmann Exploration Done Right. *31st Conference on Neural Information Processing Systems (NIPS 2017)*, 1–15. https://arxiv.org/pdf/1705.10257.pdf

Chen, H. F., Duncan, T. E., & Pasik-Duncan, B. (1999). A Kiefer-Wolfowitz Algorithm with Randomized Differences. *IEEE Transactions on Automatic Control*, *44*(3), 442–453. https://doi.org/10.1109/9.751340

Chiu, Y.-C., Lin, C.-H., Huang, J.-T., . . . Hsieh, J.-C. (2008). Immediate Gain Is Long-Term Loss: Are There Foresighted Decision Makers in the Iowa Gambling Task? *Behavioral and Brain Functions*, *4*(1), 13–13. https://doi.org/10.1186/1744-9081-4-13

Ciccarelli, M. (2017). Decision Making, Cognitive Distortions and Emotional Distress: A Comparison between Pathological Gamblers and Healthy Controls. *Journal of Behavior Therapy and Experimental Psychiatry*, *54*, 204–210. https://doi.org/10.1016/j.jbtep.2016.08.012

Costa, V. D., Mitz, A. R., & Averbeck, B. B. (2019). Subcortical Substrates of Explore-Exploit Decisions in Primates. *Neuron*, *103*(3), 533–545. https://doi.org/10.1016/j.neuron.2019.05.017

Cox, S. M. L., Frank, M. J., Larcher, K., … Dagher, A. (2015). Striatal D1 and D2 Signaling Differentially Predict Learning from Positive and Negative Outcomes. *NeuroImage*, *109*, 95–101. https://doi.org/10.1016/j.neuroimage.2014.12.070

Damasio, A. R. (1998). The Somatic Marker Hypothesis and the Possible Functions of the Prefrontal Cortex. In A. C. Roberts, T. W. Robbins, & L. Weiskrantz (Eds.). Oxford University Press. https://doi.org/10.1093/acprof:oso/9780198524410.003.0004

Damasio, A. R. (2001). Descartes Error Revisited. *Journal of the History of the Neurosciences*, *10*(2), 192–194. https://doi.org/10.1076/jhin.10.2.192.7250

Damasio, A. R. (2006). *Descartes' Error: Emotion, Reason and the Human Brain*. Vintage.

Damiano, L., Dumouchel, P., & Lehmann, H. (2012, October). Should Empathic Social Robots Have Interiority? In S. S. Ge, O. Khatib, J.-J. Cabibihan, … M.-A. Williams (Eds.), *4th international conference on social robotics* (pp. 268–277, Vol. 7621). Springer. https://doi.org/10.1007/978-3-642-34103-8_27

Daw, N. D. (2011). Trial-by-trial Data Analysis Using Computational Models. Oxford University Press. https://doi.org/10.1093/acprof:oso/9780199600434.003.0001

Denrell, J., & March, J. G. (2001). Adaptation as Information Restriction: The Hot Stove Effect. *Organization science (Providence, R.I.)*, *12*(5), 523–538. https://doi.org/10.1287/orsc.12.5.523.10092

Doya, K. (2002). Metalearning and Neuromodulation. *Neural Networks*, *15*(4), 495–506. https://doi.org/10.1016/S0893-6080(02)00044-8

Dunn, B. D., Dalgleish, T., & Lawrence, A. D. (2006). The Somatic Marker Hypothesis: A Critical Evaluation. *Neuroscience and Biobehavioral Reviews*, *30*(2), 239–271. https://doi.org/10.1016/j.neubiorev.2005.07.001

Durrett, R. (2019). *Probability: Theory and Examples* (Fifth, Vol. 49). Cambridge University Press.

Ekman, P. (1992). Are There Basic Emotions? *Psychological Review*, *99*(3), 550–553.

Erev, I., Ert, E., Roth, A. E., … Lebiere, C. (2010). A Choice Prediction Competition: Choices from Experience and from Description. *Journal of*

Behavioral Decision Making, *23*(1), 15–47. https://doi.org/10.1002/bdm.683

Erev, I., & Roth, A. E. (1998). Predicting How People Play Games: Reinforcement Learning in Experimental Games with Unique, Mixed Strategy Equilibria. *The American Economic Review*, *88*(4), 848–881.

Ernst, M., & Paulus, M. P. (2005). Neurobiology of Decision Making: A Selective Review from a Neurocognitive and Clinical Perspective. *Biological Psychiatry*, *58*(8), 597–604. https://doi.org/10.1016/j.biopsych.2005.06.004

Fellows, L. K., & Farah, M. J. (2003). Ventromedial Frontal Cortex Mediates Affective Shifting in Humans: Evidence from a Reversal Learning Paradigm. *Brain*, *126*(8), 1830–1837. https://doi.org/10.1093/brain/awg180

Fellows, L. K., & Farah, M. J. (2005). Different Underlying Impairments in Decision-making Following Ventromedial and Dorsolateral Frontal Lobe Damage in Humans. *Cerebral Cortex*, *15*(1), 58–63. https://doi.org/10.1093/cercor/bhh108

Findling, C., Skvortsova, V., Dromnelle, R., . . . Wyart, V. (2019). Computational Noise in Reward-guided Learning Drives Behavioral Variability in Volatile Environments. *Nature Neuroscience*, *22*(12), 2066–2077. https://doi.org/10.1038/s41593-019-0518-9

Floyd, M. F. (1997). Pleasure, Arousal, and Dominance: Exploring Affective Determinants of Recreation Satisfaction. *Leisure Sciences*, *19*(2), 83. https://doi.org/10.1080/01490409709512241

Folland, G. B. (1990). Remainder Estimates in Taylor's Theorem. *The American Mathematical Monthly*, *97*(3), 233. https://doi.org/10.2307/2324693

Folland, G. B. (2020). Higher-Order Derivatives and Taylors Formula in Several Variables [Accessed: 2020-11-26]. https://sites.math.washington.edu/~folland/Math425/taylor2.pdf

Fridberg, D. J., Queller, S., Ahn, W.-Y., . . . Stout, J. C. (2010). Cognitive Mechanisms Underlying Risky Decision-making in Chronic Cannabis Users. *Journal of Mathematical Psychology*, *54*(1), 28–38. https://doi.org/10.1016/j.jmp.2009.10.002

Gigerenzer, G. (2016). Towards a Rational Theory of Heuristics. In R. Frantz & L. Marsh (Eds.). Palgrave Macmillan. https://doi.org/10.1057/9781137442505_3

Gigerenzer, G., & Gaissmaier, W. (2011). Heuristic Decision Making. *Annual Review of Psychology*, *62*(1), 451–482. https://doi.org/10.1146/annurev-psych-120709-145346

Gigerenzer, G., & Gaissmaier, W. (2015). Decision Making: Nonrational Theories. In J. D. Wright (Ed.), *International encyclopedia of the social & behavioral sciences* (Second Edition, pp. 911–916). Elsevier. https://doi.org/10.1016/B978-0-08-097086-8.26017-0

Gittins, J. C., & Glazebrook, R. W. K. D. (2011). *Multi-armed Bandit Allocation Indices* (2nd). John Wiley & Sons.

Goel, V., Grafman, J., Tajik, J., … Danto, D. (1997). A Study of the Performance of Patients with Frontal Lobe Lesions in a Financial Planning Task. *Brain*, *120*(10), 1805–1822. https://doi.org/10.1093/brain/120.10.1805

Goleman, D. (2005). *Emotional Intelligence: Why It Can Matter More Than IQ* (10th anniversary). Bantam Books.

Guha, S., Munagala, K., & Shi, P. (2010). Approximation Algorithms for Restless Bandit Problems. *Journal of the ACM*, *58*(1), 1–50. https://doi.org/10.1145/1870103.1870106

Haines, N., Vassileva, J., & Ahn, W.-Y. (2018). The OutcomeRepresentation Learning Model: A Novel Reinforcement Learning Model of the Iowa Gambling Task. *Cognitive Science*, *42*(8), 2534–2561. https://doi.org/10.1111/cogs.12688

Hall, P., Heyde, C. C., Birnbaum, Z. W., & Lukacs, E. (2014). *Martingale Limit Theory and Its Application*. Elsevier Science. https://books.google.co.uk/books?id=gqriBQAAQBAJ

Hansen, L. P., & Sargent, T. J. (2008). *Robustness*. Princeton University Press.

Heaven, D. (2020). Why Faces Don't Always Tell the Truth about Feelings. *Nature (London)*, *578*(7796), 502–504. https://doi.org/10.1038/d41586-020-00507-5

Hernández, K., & Spall, J. C. (2019). Generalization of a Result of Fabian on the Asymptotic Normality of Stochastic Approximation. *Automatica*, *99*, 420–424. https://doi.org/https://doi.org/10.1016/j.automatica.2018.10.017

Hertwig, R., Barron, G., Weber, E. U., & Erev, I. (2004). Decisions from Experience and the Effect of Rare Events in Risky Choice. *Psychological Science*, *15*(8), 534–539. https://doi.org/10.1111/j.0956-7976.2004.00715.x

Hogeveen, J., Salvi, C., & Grafman, J. (2016). Emotional Intelligence: Lessons from Lesions. *Trends in Neurosciences*, *39*(10), 694–705. https://doi.org/10.1016/j.tins.2016.08.007

Hornak, J., Bramham, J., Rolls, E. T., … Polkey, C. E. (2003). Changes in Emotion after Circumscribed Surgical Lesions of the Orbitofrontal and Cingulate Cortices. *Brain*, *126*(7), 1691–1712. https://doi.org/10.1093/brain/awg168

Horstmann, A., Villringer, A., & Neumann, J. (2012). Iowa Gambling Task: There is More to Consider Than Long-term Outcome. Using a Linear Equation Model to Disentangle the Impact of Outcome and Frequency of Gains and Losses. *Frontiers in Neuroscience*, *6*, 61–61. https://doi.org/10.3389/fnins.2012.00061

Kahneman, D., & Tversky, A. (1979). Prospect Theory: An Analysis of Decision under Risk. *Econometrica*, *47*(2), 263. https://www.proquest.com/scholarly-journals/prospect-theory-analysis-decision-under-risk/docview/214665840/se-2

Kiefer, J., & Wolfowitz, J. (1952). Stochastic Estimation of the Maximum of a Regression Function. *The Annals of Mathematical Statistics*, *23*(3), 462–466. https://doi.org/10.1214/aoms/1177729392

Kingma, D. P., & Ba, J. (2014). Adam: A Method for Stochastic Optimization. https://doi.org/10.48550/ARXIV.1412.6980

Kochenderfer, M. J. (2015). *Decision Making under Uncertainty: Theory and Application*. MIT Press.

Koluman, C., Child, C., & Weyde, T. (2019). Modelling Emotion Based Reward Valuation with Computational Reinforcement Learning. In A. K. Goel, C. M. Seifert, & C. Freksa (Eds.). Cognitive Science Society.

Koluman, C., Child, C., & Weyde, T. (N.A.). Learning Rate Decay in Q-learning Models Decision Making under VMF Impairment [Unpublished, submitted to Cognitive Computation, submission active since 2020].

Konietschke, F., Placzek, M., Schaarschmidt, F., & Hothorn, L. A. (2015). nparcomp: An R Software Package for Nonparametric Multiple Comparisons and Simultaneous Confidence Intervals. *Journal of Statistical Software*, *64*(1), 1–17. https://doi.org/10.18637/jss.v064.i09

Krawczyk, D. C. (2002). Contributions of the Prefrontal Cortex to the Neural Basis of Human Decision Making. *Neuroscience and Biobehavioral Reviews*, *26*(6), 631–664. https://doi.org/10.1016/S0149-7634(02)00021-0

Kringelbach, M. L. (2005). The Human Orbitofrontal Cortex: Linking Reward to Hedonic Experience. *Nature Reviews Neuroscience*, *6*(9), 691–702. https://doi.org/10.1038/nrn1747

Kuleshov, V., & Precup, D. (2014). Algorithms for multi-armed bandit problems. https://doi.org/10.48550/ARXIV.1402.6028

Kunitani, C. (2016). Right-sized Surety Training. *SC Magazine*, *27*(2). https://go.exlibris.link/bb3MtWFs

Kushner, H. (2010). Stochastic Approximation: A Survey. *WIREs Computational Statistics*, *2*(1), 87–96. https://doi.org/10.1002/wics.57

LeCun, Y. A., Bottou, L., Orr, G. B., & Müller, K.-R. (2012). Efficient BackProp. In G. Montavon, G. Orr, & K.-R. Müller (Eds.), *Neural Networks: Tricks of the Trade* (2nd, pp. 9–48, Vol. 7700). Springer. https://doi.org/10.1007/978-3-642-35289-8_3

Li, A., Spyra, O., Perel, S., … Gupta, P. (2019). A Generalized Framework for Population Based Training. *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, 1791–1799. https://doi.org/10.1145/3292500.3330649

Li, L., Jamieson, K., Rostamizadeh, A., … Talwalkar, A. (2020). A System for Massively Parallel Hyperparameter Tuning. In I. Dhillon, D. Papailiopoulos, & V. Sze (Eds.), *Proceedings of machine learning and systems* (pp. 230–246, Vol. 2).

Li, L., Jamieson, K., DeSalvo, G., … Talwalkar, A. (2018). Hyperband: A Novel Bandit-Based Approach to Hyperparameter Optimization. *Journal of Machine Learning Research*, *18*(185), 1–52. http://jmlr.org/papers/v18/16-558.html

Liaw, R., Liang, E., Nishihara, R., … Stoica, I. (2018). Tune: A Research Platform for Distributed Model Selection and Training. https://doi.org/10.48550/ARXIV.1807.05118

Lin, C.-H., Chiu, Y.-C., & Huang, J.-T. (2009). Gain-loss Frequency and Final Outcome in the Soochow Gambling Task: A Reassessment. *Behavioral and Brain Functions*, *5*(1), 45–45. https://doi.org/10.1186/1744-9081-5-45

Lin, C.-H., Chiu, Y.-C., Lee, P.-L., & Hsieh, J.-C. (2007). Is Deck B a Disadvantageous Deck in the Iowa Gambling Task? *Behavioral and Brain Functions*, *3*(1), 16. https://doi.org/10.1186/1744-9081-3-16

Ljung, L. (1978). Strong Convergence of a Stochastic Approximation Algorithm. *Annals of Statistics*, *6*(3), 680–696. https://doi.org/10.1214/aos/1176344212

Lorkowski, J., & Kreinovich, V. (2018). *Bounded Rationality in Decision Making Under Uncertainty: Towards Optimal Granularity* (1st Ed., Vol. 99). Springer International Publishing.

Lorraine, J., Vicol, P., & Duvenaud, D. (2020). Optimizing Millions of Hyperparameters by Implicit Differentiation. In S. Chiappa & R. Calandra (Eds.), *International conference on artificial intelligence and statistics* (pp. 1540–1552, Vol. 108). http://proceedings.mlr.press/v108/lorraine20a/lorraine20a.pdf

Lucas, R. E., & Sargent, T. J. (1981). *Rational Expectations and Econometric Practice: Volume 1*. University of Minnesota Press.

Maia, T. V., & McClelland, J. L. (2004). A Reexamination of the Evidence for the Somatic Marker Hypothesis: What Participants Really Know in the Iowa Gambling Task. *Proceedings of the National Academy of Sciences of the United States of America*, *101*(45), 16075–16080.

Maia, T. V., & McClelland, J. L. (2005). The Somatic Marker Hypothesis: Still Many Questions but No Answers. *Trends in Cognitive Sciences*, *9*(4), 162–164. https://doi.org/10.1016/j.tics.2005.02.006

McKean, H. (2014). *Probability: The Classical Limit Theorems*. Cambridge University Press.

Metivier, M., & Priouret, P. (1984). Applications of a Kushner and Clark Lemma to General Classes of Stochastic Algorithms. *IEEE Transactions on Information Theory*, *30*(2), 140–151. https://doi.org/https://doi.org/10.1109/TIT.1984.1056894

Miller, K. J., Shenhav, A., & Ludvig, E. A. (2019). Habits Without Values. *Psychological review*, *126*(2), 292–311. https://doi.org/10.1037/rev0000120

Moerland, T. M., Broekens, J., & Jonker, C. M. (2018). Emotion in Reinforcement Learning Agents and Robots: A Survey. *Machine Learning*, *107*(2), 443–480. https://doi.org/10.1007/s10994-017-5666-0

Mori, K., Yamauchi, N., Wang, H., . . . Iino, Y. (2022). Probabilistic Generative Modeling and Reinforcement Learning Extract the Intrinsic Features of Animal Behavior. *Neural Networks*, *145*, 107–120. https://doi.org/10.1016/j.neunet.2021.10.002

Nogueira, F. (2014–). Bayesian Optimization: Open Source Constrained Global Optimization Tool for Python. https://github.com/fmfn/BayesianOptimization

O'Doherty, J. P., Dayan, P., Daw, N. D., ... Dolan, R. J. (2006). Cortical Substrates for Exploratory Decisions in Humans. *Nature*, *441*(7095), 876–879. https://doi.org/10.1038/nature04766

Ohta, M., Asabuki, T., & Fukai, T. (2022). Intrinsic Bursts Facilitate Learning of Lévy Flight Movements in Recurrent Neural Network Models. *Scientific reports*, *12*(1), 4951. https://doi.org/10.1038/s41598-022-08953-z

Olschewski, S., Luckman, A., Mason, A., ... Konstantinidis, E. (2024). The Future of Decisions From Experience: Connecting Real-World Decision Problems to Cognitive Processes. *Perspectives on Psychological Science*, *19*(1), 82–102. https://doi.org/10.1177/17456916231179138

Ortony, A., Clore, G. L., & Collins, A. (1990). *The Cognitive Structure of Emotions*. Cambridge University Press.

Parker-Holder, J., Nguyen, V., & Roberts, S. J. (2020). Provably Efficient Online Hyperparameter Optimization with Population-Based Bandits. *Proceedings of the 34th International Conference on Neural Information Processing Systems*, 17200–17211.

Pinto, R. (2021). Fast Reinforcement Learning with Incremental Gaussian Mixture Models. *2021 International Joint Conference on Neural Networks (IJCNN)*, 1–8. https://doi.org/10.1109/IJCNN52387.2021.9533632

Piray, P., Dezfouli, A., Heskes, T., ... Daw, N. D. (2019). Hierarchical bayesian inference for concurrent model fitting and comparison for group studies. *PLoS computational biology*, *15*(6), e1007043. https://doi.org/10.1371/journal.pcbi.1007043

Pitt, M. A., Kim, W., Navarro, D. J., & Myung, J. I. (2006). Global Model Analysis by Parameter Space Partitioning. *Psychological Review*, *113*(1), 57–83. https://doi.org/10.1037/0033-295X.113.1.57

Plutchik, R. (2001). The Nature of Emotions: Human Emotions Have Deep Evolutionary Roots, a Fact That May Explain Their Complexity and Provide Tools for Clinical Practice. *American Scientist*, *89*(4), 344–350. http://www.jstor.org/stable/27857503

Powell, W. B. (2011). *Approximate Dynamic Programming: Solving the Curses of Dimensionality* (2nd). Wiley.

Premkumar, P., Fannon, D., Kuipers, E., … Kumari, V. (2008). Emotional Decision-making and Its Dissociable Components in Schizophrenia and Schizoaffective Disorder: A Behavioural and MRI Investigation. *Neuropsychologia*, *46*(7), 2002–2012. https://doi.org/10.1016/j.neuropsychologia.2008.01.022

Reimann, M., & Bechara, A. (2010). The Somatic Marker Framework As a Neurological Theory of Decision-making: Review, Conceptual Comparisons, and Future Neuroeconomics Research. *Journal of Economic Psychology*, *31*(5), 767–776. https://doi.org/10.1016/j.joep.2010.03.002

Rescorla, R. A., & Wagner, A. R. (1972). A Theory of Pavlovian Conditioning: Variations on the Effectiveness of Reinforcement and Nonreinforcement. In A. H. Black & W. F. Prokasy (Eds.). Appleton-Century-Crofts.

Riedmiller, M., & Braun, H. (1993). A Direct Adaptive Method for Faster Backpropagation Learning: The RPROP Algorithm. *IEEE International Conference on Neural Networks*, *1*, 586–591. https://doi.org/10.1109/ICNN.1993.298623

Robbins, H., & Monro, S. (1951). A Stochastic Approximation Method. *The Annals of Mathematical Statistics*, *22*(3), 400–407. https://doi.org/https://doi.org/10.1214/aoms/1177729586

Rolls, E. T. (2000). The Orbitofrontal Cortex and Reward. *Cerebral Cortex*, *10*(3), 284–294. https://doi.org/10.1093/cercor/10.3.284

Rolls, E. T. (2013). *Emotion and Decision-making Explained*. Oxford University Press. https://doi.org/10.1093/acprof:oso/9780199659890.001.0001

Rumelhart, D. E., Hinton, G. E., & Williams, R. J. (1986). Learning Representations by Back-propagating Errors. *Nature*, *323*(6088), 533–536. https://doi.org/https://doi.org/10.1038/323533a0

Rumelhart, D. E., & McClelland, J. L. (1987). Learning Internal Representations by Error Propagation. MIT Press.

Samphantharak, K., & Townsend, R. (2013). Risk and Return in Village Economies. *NBER Working Paper Series*, 19738–n/a. https://doi.org/10.3386/w19738

Scanlon, T. M. (2000). *What We Owe to Each Other*. Belknap Press of Harvard University Press.

Schultz, W., Dayan, P., & Montague, P. R. (1997). A Neural Substrate of Prediction and Reward. *Science (American Association for the Advancement of Science)*, *275*(5306), 1593–1599. https://doi.org/10.1126/science.275.5306.1593

Sevy, S., Burdick, K. E., Visweswaraiah, H., . . . Bechara, A. (2007). Iowa Gambling Task in Schizophrenia: A Review and New Data in Patients with Schizophrenia and Co-occurring Cannabis Use Disorders. *Schizophrenia Research*, *92*(1), 74–84. https://doi.org/10.1016/j.schres.2007.01.005

Simon, H. A. (1956). Rational Choice and the Structure of the Environment. *Psychological Review*, *63*(2), 129–138. https://doi.org/10.1037/h0042769

Smith, J. E. H. (2020). *Irrationality: A History of the Dark Side of Reason*. Princeton University Press.

Snoek, J., Larochelle, H., & Adams, R. P. (2012). Practical Bayesian Optimization of Machine Learning Algorithms. In F. Pereira, C. Burges, L. Bottou, & K. Weinberger (Eds.), *Advances in Neural Information Processing Systems* (Vol. 25). Curran Associates, Inc. https://proceedings.neurips.cc/paper/2012/file/05311655a15b75fab86956663e1819cd-Paper.pdf

Soanes, C., & Stevenson, A. (Eds.). (2008). *Rational* (11th rev.). Oxford University Press.

Spall, J. C. (1992). Multivariate Stochastic Approximation Using a Simultaneous Perturbation Gradient Approximation. *IEEE Transactions on Automatic Control*, *37*(3), 332–341. https://doi.org/https://doi.org/10.1109/9.119632

Spall, J. C. (1997). A One-measurement Form of Simultaneous Perturbation Stochastic Approximation. *Automatica*, *33*(1), 109–112. https://doi.org/10.1016/S0005-1098(96)00149-5

Spall, J. C. (1998). An Overview of the Simultaneous Perturbation Method for Efficient Optimization [Accessed: 2020-12-09]. *Johns Hopkins APL Technical Digest*, *19*(4), 482–492. https://www.jhuapl.edu/Content/techdigest/pdf/V19-N04/19-04-Spall.pdf

Spall, J. C. (2003). *Introduction to Stochastic Search and Optimization: Estimation, Simulation, and Control*. Wiley-Interscience. https://doi.org/https://doi.org/10.1002/0471722138

Spall, J. C., & Cristion, J. A. (1994). Nonlinear Adaptive Control Using Neural Networks: Estimation with a Smoothed Form of Simultaneous Perturbation Gradient Approximation. *Statistica Sinica*, *4*(1), 1–27.

Stander, N., & Craig, K. J. (2002). On the Robustness of a Simple Domain Reduction Scheme for Simulation-based Optimization. *Engineering Computations*, *19*(4), 431–450. https://doi.org/10.1108/02644400210430190

Steingroever, H., Fridberg, D. J., Horstmann, A., … Wagenmakers, E.-J. (2015). Data from 617 Healthy Participants Performing the Iowa Gambling Task: A Many Labs Collaboration. *Journal of Open Psychology Data*, *3*(e5). https://openpsychologydata.metajnl.com/articles/10.5334/jopd.ak

Steingroever, H., Pachur, T., Smíra, M., & Lee, M. D. (2018). Bayesian Techniques for Analyzing Group Differences in the Iowa Gambling Task: A Case Study of Intuitive and Deliberate Decision-makers. *Psychonomic Bulletin & Review*, *25*(3), 951–970. https://doi.org/10.3758/s13423-017-1331-7

Steingroever, H., Wetzels, R., & Wagenmakers, E.-J. (2013). A Comparison of Reinforcement Learning Models for the Iowa Gambling Task Using Parameter Space Partitioning. *The Journal of Problem Solving*, *5*(2). https://doi.org/10.7771/1932-6246.1150

Sutton, R. S., & Barto, A. G. (2018). *Reinforcement Learning: An Introduction* (Second). The MIT Press.

Szepesvári, C. (2010). *Algorithms for Reinforcement Learning* (Vol. 9). Morgan & Claypool. https://doi.org/https://doi.org/10.2200/S00268ED1V01Y201005AIM009

Teles, R. V. (2020). Phineas Gage's Great Legacy. *Dementia & Neuropsychologia*, *14*(4), 419–421. https://doi.org/10.1590/1980-57642020dn14-040013

Tokic, M. (2010). Adaptive ε-Greedy Exploration in Reinforcement Learning Based on Value Differences. In R. Dillmann, J. Beyerer, U. D. Hanebeck, & T. Schultz (Eds.), *Ki 2010: Advances in artificial intelligence* (pp. 203–210). Springer. https://doi.org/https://doi.org/10.1007/978-3-642-16111-7_23

Tsampallas, V., Renshaw-Vuillier, L., Charles, F., & Kostoulas, T. (2023). Emotions and Gambling: Towards a Computational Model of Gambling Experience. In E. André, M. Chetouani, D. Vaufreydaz, … A. Vinciarelli (Eds.). ACM. https://doi.org/10.1145/3610661.3616126

Tsitsiklis, J. N. (1993). Asynchronous Stochastic Approximation and Q-learning. *Proceedings of 32nd IEEE Conference on Decision and Control*, *1*, 395–400. https://doi.org/10.1109/CDC.1993.325119

Volz, K. G., & Hertwig, R. (2016). Emotions and Decisions: Beyond Conceptual Vagueness and the Rationality Muddle. *Perspectives on Psychological Science*, *11*(1), 101–116. https://doi.org/10.1177/1745691615619608

Von Neumann, J., & Morgenstern, O. (1944). *Theory of Games and Economic Behavior* (60th anniversary). Princeton University Press. Retrieved July 4, 2022, from http://www.jstor.org/stable/j.ctt1r2gkx

Wasserman, L. (2006). *All of Nonparametric Statistics*. Springer Verlag. https://doi.org/10.1007/0-387-30623-4

Watkins, C. (1989). *Learning from Delayed Rewards* [Unpublished doctoral dissertation]. King's College, London, UK. http://www.cs.rhul.ac.uk/~chrisw/new%5C_thesis.pdf

Wedgwood, R. (2017). *The Value of Rationality*. Oxford University Press.

Widrow, B., & Hoff, M. E. (1960). Adaptive Switching Circuits. *IRE Wescon Convention*.

Wilson, R. C., & Collins, A. G. E. (2019). Ten Simple Rules for the Computational Modeling of Behavioral Data. *eLife*, *8*. https://doi.org/10.7554/eLife.49547

Wilson, R. C., Geana, A., White, J. M., … Cohen, J. D. (2014). Humans Use Directed and Random Exploration to Solve the Explore-Exploit Dilemma. *Journal of Experimental Psychology*, *143*(6), 2074–2081. https://doi.org/10.1037/a0038199

Wolpert, D. H., & Macready, W. G. (1997). No Free Lunch Theorems For Optimization. *IEEE Transactions on Evolutionary Computation*, *1*(1), 67–82. https://doi.org/https://dx.doi.org/10.1109/4235.585893

Wood, S., Busemeyer, J., Koling, A., … Davis, H. (2005). Older Adults as Adaptive Decision Makers: Evidence From the Iowa Gambling Task. *Psychology and Aging*, *20*(2), 220–225. https://doi.org/10.1037/0882-7974.20.2.220

Worthy, D. A., & Maddox, W. T. (2014). A Comparison Model of Reinforcement-learning and Win-stay-lose-shift Decision-making processes: A Tribute to W.K. Estes. *Journal of Mathematical Psychology*, *59*, 41–49. https://doi.org/10.1016/j.jmp.2013.10.001

Xiao, H., Rasul, K., & Vollgraf, R. (2017). Fashion-MNIST: A Novel Image Dataset for Benchmarking Machine Learning Algorithms. https://doi.org/10.48550/ARXIV.1708.07747

Yin, G. G., & Kushner, H. J. (2003). *Stochastic Approximation and Recursive Algorithms and Applications* (Second). Springer. https://doi.org/10.1007/b97441

Zhou, R., & Palomar, D. P. (2020). A Theoretical Basis for Practitioners Heuristic 1/N and Long-Only Quintile Portfolio. *ICASSP 2020 - 2020 IEEE International Conference on Acoustics, Speech and Signal Processing*, 8434–8438. https://doi.org/10.1109/ICASSP40776.2020.9053772

Zhu, J., Wang, L., & Spall, J. C. (2020). Efficient Implementation of Second-order Stochastic Approximation Algorithms in High-dimensional Problems. *IEEE Transactions on Neural Networks and Learning Systems*, *31*(8), 3087–3099. https://doi.org/10.1109/TNNLS.2019.2935455