Reinforcement Learning, Bit by Bit

Other titles in Foundations and Trends® in Machine Learning

Conformal Prediction: A Gentle Introduction
Anastasios N. Angelopoulos and Stephen Bates
ISBN: 978-1-63828-158-0

Introduction to Riemannian Geometry and Geometric Statistics: From Basic Theory to Implementation with Geomstats
Nicolas Guigui, Nina Miolane and Xavier Pennec
ISBN: 978-1-63828-154-2

Graph Neural Networks for Natural Language Processing: A Survey
Lingfei Wu, Yu Chen, Kai Shen, Xiaojie Guo, Hanning Gao, Shucheng Li, Jian Pei and Bo Long
ISBN: 978-1-63828-142-9

Model-based Reinforcement Learning: A Survey
Thomas M. Moerland, Joost Broekens, Aske Plaat and Catholijn M. Jonker
ISBN: 978-1-63828-056-9

Divided Differences, Falling Factorials, and Discrete Splines: Another Look at Trend Filtering and Related Problems
Ryan J. Tibshirani
ISBN: 978-1-63828-036-1

Risk-Sensitive Reinforcement Learning via Policy Gradient Search
Prashanth L. A. and Michael C. Fu
ISBN: 978-1-63828-026-2

Reinforcement Learning, Bit by Bit

Xiuyuan Lu
DeepMind
lxlu@deepmind.com

Benjamin Van Roy
DeepMind
benvanroy@deepmind.com

Vikranth Dwaracherla
DeepMind
vikranthd@deepmind.com

Morteza Ibrahimi
DeepMind
mibrahimi@deepmind.com

Ian Osband
DeepMind
iosband@deepmind.com

Zheng Wen
DeepMind
zhengwen@deepmind.com

now
the essence of knowledge
Boston — Delft

Foundations and Trends® in Machine Learning

Published, sold and distributed by:
now Publishers Inc.
PO Box 1024
Hanover, MA 02339
United States
Tel. +1-781-985-4510
www.nowpublishers.com
sales@nowpublishers.com

Outside North America:
now Publishers Inc.
PO Box 179
2600 AD Delft
The Netherlands
Tel. +31-6-51115274

The preferred citation for this publication is

X. Lu *et al.*. *Reinforcement Learning, Bit by Bit*. Foundations and Trends® in Machine Learning, vol. 16, no. 6, pp. 733–865, 2023.

ISBN: 978-1-63828-255-6

Foundations and Trends® in Machine Learning
Volume 16, Issue 6, 2023

Editorial Board

Editorial Scope

Topics

Foundations and Trends® in Machine Learning publishes survey and tutorial articles in the following topics:

- Adaptive control and signal processing
- Applications and case studies
- Behavioral, cognitive and neural learning
- Bayesian learning
- Classification and prediction
- Clustering
- Data mining
- Dimensionality reduction
- Evaluation
- Game theoretic learning
- Graphical models
- Independent component analysis
- Inductive logic programming
- Kernel methods
- Markov chain Monte Carlo
- Model choice
- Nonparametric methods
- Online learning
- Optimization
- Reinforcement learning
- Relational learning
- Robustness
- Spectral methods
- Statistical learning theory
- Variational inference
- Visualization

Information for Librarians

Foundations and Trends® in Machine Learning, 2023, Volume 16, 6 issues. ISSN paper version 1935-8237. ISSN online version 1935-8245. Also available as a combined paper and online subscription.

Contents

1 Introduction **2**
1.1 Data Efficiency . . . 3
1.2 Information Versus Computation . . . 3
1.3 Preview . . . 5

2 Environments and Agents **6**
2.1 Agent-Environment Interface . . . 7
2.2 Probabilistic Framework and Notation . . . 9
2.3 Rewards . . . 10
2.4 Prototypical Examples . . . 11

3 Elements of Agent Design **13**
3.1 Sources of Uncertainty . . . 15
3.2 Agent State . . . 16
3.3 Information . . . 19
3.4 Learning and Prioritization . . . 23

4 Cost-Benefit Analysis **28**
4.1 Agent Policy and Regret . . . 28
4.2 Information Gain . . . 31
4.3 The Information Ratio . . . 33
4.4 A Regret Bound . . . 34
4.5 Examples . . . 37

5 Retaining Information **51**
5.1 Epistemic State 51
5.2 Environment Proxies 54

6 Seeking Information **61**
6.1 Learning Targets 61
6.2 Exploration, Exploitation, and the Information Ratio 63
6.3 Information-Directed Sampling 64

7 Computational Examples **72**
7.1 A Practical Information Seeking Agent 73
7.2 Epistemic Neural Networks 74
7.3 Information Seeking via Variance-IDS 78
7.4 Variance-IDS with General Value Functions 80

8 Closing Remarks **87**

Acknowledgements **89**

Appendices **90**

A Analysis of Thompson Sampling with an Episodic MDP **91**

B Analysis of IDS in an Episodic Environment **107**

C Convexity and Support of Value-IDS **111**

D Relation Between Information Gain and Variance **114**

E Implementation and Computation **118**

References **124**

Reinforcement Learning, Bit by Bit

Xiuyuan Lu[1], Benjamin Van Roy[2], Vikranth Dwaracherla[3], Morteza Ibrahimi[4], Ian Osband[5] and Zheng Wen[6]

[1]*DeepMind, USA; lxlu@deepmind.com*
[2]*DeepMind, USA; benvanroy@deepmind.com*
[3]*DeepMind, USA; vikranthd@deepmind.com*
[4]*DeepMind, USA; mibrahimi@deepmind.com*
[5]*DeepMind, USA; iosband@deepmind.com*
[6]*DeepMind, USA; zhengwen@deepmind.com*

ABSTRACT

Reinforcement learning agents have demonstrated remarkable achievements in simulated environments. Data efficiency poses an impediment to carrying this success over to real environments. The design of data-efficient agents calls for a deeper understanding of information acquisition and representation. We discuss concepts and regret analysis that together offer principled guidance. This line of thinking sheds light on questions of *what information to seek*, *how to seek that information*, and *what information to retain*. To illustrate concepts, we design simple agents that build on them and present computational results that highlight data efficiency.

Xiuyuan Lu, Benjamin Van Roy, Vikranth Dwaracherla, Morteza Ibrahimi, Ian Osband and Zheng Wen (2023), "Reinforcement Learning, Bit by Bit", Foundations and Trends® in Machine Learning: Vol. 16, No. 6, pp 733–865. DOI: 10.1561/2200000097.

1

Introduction

"*Other learning paradigms are about minimization; reinforcement learning is about maximization.*"

The statement quoted above has been attributed to Harry Klopf, though it might only be accurate in sentiment. The statement may sound vacuous, since minimization can be converted to maximization simply via negation of an objective. However, further reflection reveals a deeper observation. Many learning algorithms aim to mimic observed patterns, minimizing differences between model and data. Reinforcement learning is distinguished by its open-ended view. A reinforcement learning agent learns to improve its behavior over time, without a prescription for eventual dynamics or the limits of performance. If the objective takes nonnegative values, *minimization* suggests a well-defined desired outcome while *maximization* conjures pursuit of the unknown. Indeed, Klopf (1982) argued that, by focusing on *minimization* of deviations from a desired operating point, then-prevailing theories of homeostasis were too limiting to explain intelligence, while a theory centered around heterostasis *could* by allowing for *maximization* of open-ended objectives.

1.1 Data Efficiency

In reinforcement learning, the nature of data depends on the agent's behavior. This bears important implications on the need for data efficiency. In supervised and unsupervised learning, data is typically viewed as static or evolving slowly. If data is abundant, as is the case in many modern application areas, the performance bottleneck often lies in model capacity and computational infrastructure. This holds also when reinforcement learning is applied to simulated environments; while data generated in the course of learning does evolve, a slow rate can be maintained, in which case model capacity and computation remain bottlenecks, though data efficiency can be helpful in reducing simulation time. On the other hand, in a real environment, data efficiency often becomes the gating factor.

Data efficiency depends on what information the agent seeks, how it seeks that information, and what it retains. This tutorial offers a framework that can guide associated agent design decisions. This framework is inspired in part by concepts from another field that has grappled with data efficiency. In communication, the goal is typically to transmit data through a channel in a way that maximizes throughput, measured in bits per second. In reinforcement learning, an agent interacts with an unknown environment with an aim to maximize reward. An important difference that emerges is that bits of information serve as means to maximizing reward and not ends. As such, an important factor arising in reinforcement learning concerns weighing costs and benefits of acquiring particular bits of information. Despite this distinction, some concepts from communication can guide our thinking about information in reinforcement learning.

1.2 Information Versus Computation

Communication was a particularly active area of research at the turn of the twentieth century, with an emphasis on scaling up power generation to enable transmission of analog signals over increasing distances. At the time, encoding and decoding was handled heuristically. In the 1940s, following Shannon's maxim of "information first, then computation,"

the focus shifted to understanding what is possible or impossible. This initiative introduced the *bit*[1] as a unit of information and established fundamental limits of communication. The maxim encouraged understanding possibilities and deferred the study of computation. Design of encoding and decoding algorithms that attain fundamental limits arrived in the 1960s, with practical implementations emerging in the 1990s. It is fair to say that this thread of research formed a cornerstone for today's connected world (Jha, 2016).

Reinforcement learning seems to have followed an opposite maxim: "computation first, then information." Beginning with heuristic evolution of algorithmic ideas such as temporal-difference learning (Witten, 1976; Witten, 1977; Sutton, 1988; Watkins, 1989) and actor-critic architectures (Witten, 1977; Barto *et al.*, 1983; Sutton, 1984), followed by demonstrated promise (Tesauro, 1992; Tesauro, 1994), over the last decades of the twentieth century, much effort was directed toward computational methods, with little regard to data-efficiency (Bertsekas and Tsitsiklis, 1996; Sutton and Barto, 2018; Bertsekas, 2019). The past decade has experienced a great deal of further innovation, with an emphasis on scaling up computations and environments, leading to reinforcement learning agents that have produced impressive results in simulated environments and attracted enormous interest (Mnih *et al.*, 2015; Schrittwieser *et al.*, 2020). However, data efficiency presents an impediment to the transfer of this success to real environments. Unlike communication, *information* has not been the focus in these lines of research. Questions that are central to data efficiency, such as what information an agent should acquire and the cost of gathering that information, have mostly been ignored.

While much of the focus has been on developing heuristics and scaling up computation, there is a few notable exceptions. The work of Hutter (2007) aims to design a "universal" agent, building on ideas such as Solomonoff's universal prior while putting aside any computational consideration. It remains unclear, though, how this line of thinking may offer a path towards designing practical, data-efficient agents. There is also a body of work that aims to address data efficiency and derive

[1] Originally termed the *binary digit*, then the *binit*, before the *bit*.

sample complexity bounds in stylized environments including bandits and Markov decision processes (Kearns and Singh, 2002; Brafman and Tennenholtz, 2003; Jaksch *et al.*, 2010; Azar *et al.*, 2017; Jiang *et al.*, 2017; Jin *et al.*, 2020). However, methods considered in this line of work are not sufficiently scalable to address real, complex environments. The generality of our theoretical framework for thinking about information and data efficiency accommodates reasoning about scalable agent designs. This serves our ultimate goal of designing practical, data-efficient agents for real applications.

1.3 Preview

In this tutorial, we present a framework for studying costs and benefits associated with information. As we will explain, this can guide how agents represent knowledge and how they seek and retain new information. In particular, the framework sheds light on the questions of *what information to seek*, *how to seek that information*, and *what information to retain*.

We begin in Section 2 with a formalism for studying agents and environments. We present a simplified version of the DQN agent (Mnih *et al.*, 2013; Mnih *et al.*, 2015) and an ensemble-DQN agent (Osband *et al.*, 2016; Osband *et al.*, 2019) as examples. Then, in Section 3, we discuss conceptual elements arising in the design of practical agents that can operate effectively in complex environments, with particular emphasis on informational considerations. By interpreting the DQN and ensemble-DQN agents through this lens, we illustrate abstract concepts and highlight sources of inefficiency. In Section 4, we study a regret bound that applies to all agents and provides insight into design trade-offs. We also illustrate insights offered by the bound when used to study particular classes of environments and agents. As discussed in Sections 5 and 6, this bound can be used to think about how to design agents that seek and retain the right information. In Section 7, we present scalable agent designs. Computational results reported in Section 7 serve to illustrate concepts covered in the tutorial and demonstrate their practical applicability.

Acknowledgements

Our thinking about the relation between information and sequential decision was shaped by an earlier collaboration with Dan Russo, which focused on bandit environments. Tor Lattimore offered many helpful comments on an earlier draft. Johannes Kirschner provided valuable feedback during the review process that helped improve the monograph significantly. Chao Qin's careful study of value-IDS revealed technical limitations, as captured by his insightful result that we cite and discuss at the end of Section 4. We also would like to thank Chao Qin for the helpful discussion with him on the "optimism conjecture" in the Thompson sampling analysis. The monograph also benefited from discussions with and feedback from Dilip Arumugam, Seyed Mohammad Asghari, Andy Barto, Dimitri Bertsekas, Adithya Devraj, Shi Dong, Abbas El Gamal, Yanjun Han, Mike Harrison, Geoffrey Irving, Anmol Kagrecha, Ayfer Ozgur, Warren Powell, Doina Precup, Omar Rivasplata, David Silver, Satinder Singh, Rich Sutton, David Tse, John Tsitsiklis, and Tsachy Weissman.

Appendices

A

Analysis of Thompson Sampling with an Episodic MDP

In this appendix, we provide an analysis of Thompson sampling for the episodic MDP described in Section 4.5.2. We start by establishing useful concentration inequalities in Section A.1. We then propose an optimism conjecture in Section A.2 and support it through empirical simulations. Finally, in Section A.3, we establish a regret bound for Thompson sampling in the environments described in Section 4.5.2, assuming that the optimism conjecture holds. Note that all the entropy and mutual information terms in this section are measured in nats.

A.1 Information and Concentration

Lemma A.1. If p and $\hat{p}$ are independent and identically beta-distributed random variables with parameters $\alpha \geq 1$ and $\beta \geq 1$ then, for all $c > 0$,

$$\mathbb{P}(\sqrt{c\mathbb{I}(p;b)} - |p - \hat{p}| \leq 0) \leq 2e^{-c/6},$$

where $b \sim \text{Bernoulli}(p)$ conditioned on p.

Proof. A real-valued random variable X is said to be σ^2-sub-Gaussian if $\mathbb{E}[\exp(\lambda(X - \mathbb{E}X))] \leq \exp(\lambda^2\sigma^2/2)$ for all λ. We first prove that $p - \hat{p}$ is $\frac{1}{2(\alpha+\beta)}$-sub-Gaussian. Since p and $\hat{p}$ are i.i.d. from $\text{Beta}(\alpha, \beta)$,

thus $\mathbb{E}[p - \hat{p}] = 0$. Moreover, from Theorem 4 of Elder (2016), for $p \sim \text{Beta}(\alpha, \beta)$, $p - \mathbb{E}[p]$ is $\frac{1}{4(\alpha+\beta)+2}$-sub-Gaussian. Consequently, $\mathbb{E}[p] - \hat{p} = \mathbb{E}[\hat{p}] - \hat{p} = -(\hat{p} - \mathbb{E}[\hat{p}])$ is also $\frac{1}{4(\alpha+\beta)+2}$-sub-Gaussian. Since, $p - \mathbb{E}[p]$ and $\mathbb{E}[p] - \hat{p}$ are also independent, we have

$$p - \hat{p} = (p - \mathbb{E}[p]) + (\mathbb{E}[p] - \hat{p})$$

is $\frac{1}{2(\alpha+\beta)+1}$-sub-Gaussian. Since $\frac{1}{2(\alpha+\beta)+1} < \frac{1}{2(\alpha+\beta)}$, $p - \hat{p}$ is also $\frac{1}{2(\alpha+\beta)}$-sub-Gaussian.

Consequently, from the sub-Gaussian tail bound, we have

$$\begin{aligned}
&\mathbb{P}(\sqrt{c\mathbb{I}(p;b)} - |p - \hat{p}| \leq 0) \\
&= \mathbb{P}\left(|p - \hat{p}| \geq \sqrt{c\mathbb{I}(p;b)}\right) \\
&\leq 2\exp\left(-\frac{c\mathbb{I}(p;b)}{2 \cdot \frac{1}{2(\alpha+\beta)}}\right) = 2\exp\left(-c\mathbb{I}(p;b)(\alpha+\beta)\right).
\end{aligned}$$

From Lemma 10 of Lu (2020), for $p \sim \text{Beta}(\alpha, \beta)$ with $\alpha \geq 1$ and$\beta \geq 1$,

$$\mathbb{I}(p;b) \geq \frac{1}{6(\alpha+\beta)}.$$

Thus, we have

$$\mathbb{P}(\sqrt{c\mathbb{I}(p;b)} - |p - \hat{p}| \leq 0) \leq 2\exp\left(-c\mathbb{I}(p;b)(\alpha+\beta)\right) \leq 2e^{-c/6}.$$

□

Lemma A.2. For any positive integer N, if $p_1, \ldots, p_N$ are independent beta-distributed random variables with parameters greater than or equal to 1 and, for each n, $\hat{p}_n$ is independent and distributed identically with p_n then, for all $\delta \in (0, 1)$,

$$\mathbb{E}\left[\min_{n \in \{1,\ldots,N\}} \left(\sqrt{6\mathbb{I}(p_n;b_n)\ln\frac{2N}{\delta}} - |p_n - \hat{p}_n| + \delta\right)\right] \geq 0,$$

where $b_n \sim \text{Bernoulli}(p_n)$ conditioned on p_n for $n = 1, \ldots, N$.

Proof. For any $c > 6\ln(2)$, we have $2e^{-c/6} < 1$, thus

$$\begin{aligned}
&\mathbb{P}\left(\min_{n\in\{1,\dots,N\}}\left(\sqrt{c\mathbb{I}(p_n;b_n)}-|p_n-\hat{p}_n|\right)\leq 0\right)\\
&=1-\mathbb{P}\left(\min_{n\in\{1,\dots,N\}}\left(\sqrt{c\mathbb{I}(p_n;b_n)}-|p_n-\hat{p}_n|\right)\geq 0\right)\\
&=1-\prod_{n=1}^{N}\mathbb{P}\left(\sqrt{c\mathbb{I}(p_n;b_n)}-|p_n-\hat{p}_n|\geq 0\right)\\
&=1-\prod_{n=1}^{N}\left(1-\mathbb{P}\left(\sqrt{c\mathbb{I}(p_n;b_n)}-|p_n-\hat{p}_n|\leq 0\right)\right)\\
&\leq 1-\left(1-2e^{-c/6}\right)^{N} \quad \text{from Lemma A.1 and } 2e^{-c/6}<1\\
&\leq 1-1+2Ne^{-c/6} \quad \text{from Bernoulli's inequality}\\
&=2Ne^{-c/6}.
\end{aligned}$$

On substituting $c = 6\ln\frac{2N}{\delta} > 6\ln(2)$,

$$\mathbb{P}\left(\min_{n\in\{1,\dots,N\}}\left(\sqrt{6\mathbb{I}(p_n;b_n)\ln\frac{2N}{\delta}}-|p_n-\hat{p}_n|\right)\leq 0\right)\leq\delta$$

To simplify the notation, we define

$$h=\min_{n\in\{1,\dots,N\}}\left(\sqrt{6\mathbb{I}(p_n;b_n)\ln\frac{2N}{\delta}}-|p_n-\hat{p}_n|\right).$$

Notice that since $p_n, \hat{p}_n \in [0,1]$, thus $h \geq -1$ always holds. Also, from the above results, $\mathbb{P}(h \leq 0) \leq \delta$. Therefore,

$$\begin{aligned}
\mathbb{E}[h] &= \mathbb{E}[h|h\geq 0]\mathbb{P}(h\geq 0)+\mathbb{E}[h|h<0]\mathbb{P}(h<0)\\
&\overset{(a)}{\geq}\mathbb{E}[h|h<0]\mathbb{P}(h<0)\overset{(b)}{\geq}-\delta,
\end{aligned}$$

where (a) holds since $\mathbb{E}[h|h \geq 0]\mathbb{P}(h \geq 0) \geq 0$, and (b) holds since $\mathbb{E}[h|h < 0] \geq -1$ and $\mathbb{P}(h < 0) \leq \mathbb{P}(h \leq 0) \leq \delta$. Consequently, $\mathbb{E}[h + \delta] \geq 0$, that is

$$\mathbb{E}\left[\min_{n\in\{1,\dots,N\}}\left(\sqrt{6\mathbb{I}(p_n;b_n)\ln\frac{2N}{\delta}}-|p_n-\hat{p}_n|+\delta\right)\right]\geq 0.$$

□

Lemma A.3. Assume that p is a beta-distributed random variable with $\alpha \geq 1$ and $\beta \geq 1$, and b is drawn from Bernoulli(p) conditioned on p. For $\delta = 1/2, 1/3, 1/4, \ldots$, define $q = \lceil p/\delta \rceil$. Then we have

$$\mathbb{I}(p; b) \geq \mathbb{I}(q; b) \geq \mathbb{I}(p; b) - \max\left\{3, \ln\left(\frac{2}{\delta}\right)\right\}\delta.$$

Proof. Notice that conditioning on p, q is deterministic, and q and b are independent. Consequently,

$$\mathbb{I}(p; b) = \mathbb{I}(p, q; b) \geq \mathbb{I}(q; b).$$

Let $f(\cdot)$ denote the probability density function of p. To simplify exposition, we use $\tilde{q}_i$ to denote $\mathbb{P}(b = 1 | q = i\delta)$ for $i = 1, \ldots, 1/\delta$. Note that

$$\tilde{q}_i = \mathbb{P}(b = 1 | q = i\delta) = \mathbb{E}\left[p | q = i\delta\right] = \frac{\int_{(i-1)\delta}^{i\delta} p f(p) dp}{\int_{(i-1)\delta}^{i\delta} f(p) dp} \quad \forall i = 1, 2, \ldots, 1/\delta.$$

With some algebraic manipulation, we can show that

$$\begin{aligned}
\mathbb{H}(b|q) &= \sum_{i=1}^{1/\delta} \mathbb{P}(q = i\delta) \left(\tilde{q}_i \ln \frac{1}{\tilde{q}_i} + (1 - \tilde{q}_i) \ln \frac{1}{1 - \tilde{q}_i}\right) \\
&= \sum_{i=1}^{1/\delta} \int_{(i-1)\delta}^{i\delta} p f(p) dp \ln \frac{1}{\tilde{q}_i} + \int_{(i-1)\delta}^{i\delta} (1 - p) f(p) dp \ln \frac{1}{1 - \tilde{q}_i} \\
&= \sum_{i=1}^{1/\delta} \int_{(i-1)\delta}^{i\delta} \left(p \ln \frac{1}{\tilde{q}_i} + (1 - p) \ln \frac{1}{1 - \tilde{q}_i}\right) f(p) dp.
\end{aligned}$$

Thus,

$$\begin{aligned}
\mathbb{I}(p; b) - \mathbb{I}(q; b) &= \mathbb{H}(b|q) - \mathbb{H}(b|p) \\
&= \sum_{i=1}^{1/\delta} \int_{(i-1)\delta}^{i\delta} \left(p \ln \frac{p}{\tilde{q}_i} + (1 - p) \ln \frac{1 - p}{1 - \tilde{q}_i}\right) f(p) dp \\
&= \sum_{i=1}^{1/\delta} \int_{(i-1)\delta}^{i\delta} \mathbf{d}_{\mathrm{KL}}\left(p \| \tilde{q}_i\right) f(p) dp,
\end{aligned}$$

where, with some abuse of notation, we use $\mathbf{d}_{\mathrm{KL}}(p \| \tilde{q}_i)$ to denote a shorthand for $\mathbf{d}_{\mathrm{KL}}\left(\mathrm{Bern}(p) \| \mathrm{Bern}(\tilde{q}_i)\right)$.

Without loss of generality, we assume that $\alpha \leq \beta$ (the other case is symmetric). Obviously, $\forall i = 1, 2, \ldots, 1/\delta$, we have $(i-1)\delta \leq \tilde{q}_i \leq i\delta$. Moreover, since $1 \leq \alpha \leq \beta$, we have $\tilde{q}_{1/\delta} \leq 1 - \delta/2$. This is because for $\text{Beta}(\alpha, \beta)$ with $1 \leq \alpha \leq \beta$, $\text{Beta}(\alpha, \beta)$ is either a uniform distribution, or a uni-modal distribution with mode less than or equal to 0.5. Hence, $f(p)$ is strictly decreasing on interval $[1-\delta, 1]$, and hence $\tilde{q}_{1/\delta} \leq 1 - \delta/2$. Consequently, for $i = 2, \ldots, 1/\delta$, we have

$$\mathbf{d}_{\text{KL}}(p\|\tilde{q}_i) \overset{(a)}{\leq} \frac{p^2}{\tilde{q}_i} + \frac{(1-p)^2}{1-\tilde{q}_i} - 1 = \frac{(p-\tilde{q}_i)^2}{\tilde{q}_i(1-\tilde{q}_i)} \overset{(b)}{\leq} \frac{\delta^2}{\frac{3}{8}\delta} = \frac{8}{3}\delta < 3\delta,$$

where (a) follows from Theorem 1 of Dragomir *et al.* (2000), and (b) follows from $|p - \tilde{q}_i| \leq \delta$, and $\delta \leq \tilde{q}_i \leq 1 - \delta/2$ for $i \geq 2$. Specifically, for $\delta \leq \tilde{q}_i \leq 1 - \delta/2$, we have

$$\tilde{q}_i(1-\tilde{q}_i) \geq \frac{\delta}{2}\left(1 - \frac{\delta}{2}\right) \geq \frac{\delta}{2}\left(1 - \frac{1}{4}\right) = \frac{3}{8}\delta,$$

where the second inequality follows from $\delta \leq \frac{1}{2}$.

We now consider the case when $i = 1$ and bound $\mathbf{d}_{\text{KL}}(p\|\tilde{q}_1)$ for $p \in (0, \delta]$. Notice that for $p \in (0, \delta]$, we have

$$\begin{aligned}
\mathbf{d}_{\text{KL}}(p\|q_1) &\overset{(a)}{\leq} \max\left\{\mathbf{d}_{\text{KL}}(0\|q_1), \mathbf{d}_{\text{KL}}(\delta\|q_1)\right\} \\
&= \max\left\{\ln\frac{1}{1-\tilde{q}_1}, \delta\ln\frac{\delta}{\tilde{q}_1} + (1-\delta)\ln\frac{1-\delta}{1-\tilde{q}_1}\right\} \\
&\overset{(b)}{\leq} \max\left\{2\delta, \delta\ln\frac{\delta}{\tilde{q}_1}\right\},
\end{aligned}$$

where (a) follows from $p \in (0, \delta]$, and (b) follows from $\ln\frac{1-\delta}{1-\tilde{q}_1} \leq 0$ and

$$\ln\frac{1}{1-\tilde{q}_1} \leq \ln\frac{1}{1-\delta} \leq \ln\left(1 + \frac{\delta}{1-\delta}\right) \leq \frac{\delta}{1-\delta} \leq 2\delta,$$

where the last inequality follows from $\delta \leq 1/2$. We now derive a lower bound on $\tilde{q}_1$. Let $F(\cdot; \alpha, \beta)$ denote the CDF of $\text{Beta}(\alpha, \beta)$, then we have

$$\begin{aligned}
\tilde{q}_1 &= \frac{1}{F(\delta;\alpha,\beta)}\int_0^\delta \frac{x^\alpha(1-x)^{\beta-1}}{B(\alpha,\beta)}dx \\
&= \frac{B(\alpha+1,\beta)}{B(\alpha,\beta)F(\delta;\alpha,\beta)}\int_0^\delta \frac{x^\alpha(1-x)^{\beta-1}}{B(\alpha+1,\beta)}dx \\
&= \frac{B(\alpha+1,\beta)F(\delta;\alpha+1,\beta)}{B(\alpha,\beta)F(\delta;\alpha,\beta)} \\
&\overset{(a)}{=} \frac{\alpha}{\alpha+\beta}\left[1-\frac{\delta^\alpha(1-\delta)^\beta}{\alpha\int_0^\delta x^{\alpha-1}(1-x)^{\beta-1}dx}\right] \\
&\overset{(b)}{\geq} \frac{\alpha}{\alpha+\beta}\left[1-\frac{\delta^\alpha(1-\delta)^\beta}{\alpha\int_0^\delta x^{\alpha-1}(1-\delta)^{\beta-1}dx}\right] = \frac{\alpha}{\alpha+\beta}\delta,
\end{aligned}$$

where $B(\cdot,\cdot)$ is the beta function. Note that (a) follows from $B(\alpha+1,\beta) = B(\alpha,\beta)\frac{\alpha}{\alpha+\beta}$, and

$$F(\delta;\alpha+1,\beta) = F(\delta;\alpha,\beta) - \frac{\delta^\alpha(1-\delta)^\beta}{\alpha B(\alpha,\beta)},$$

and (b) follows from $1-x \geq 1-\delta > 0$ and $\beta \geq 1$, and hence $(1-x)^{\beta-1} \geq (1-\delta)^{\beta-1}$.

Combining the above results, we have $\mathbf{d}_{\mathrm{KL}}(p\|q_1) \leq \max\left\{2, \ln\frac{\alpha+\beta}{\alpha}\right\}\delta$. Since $\mathbf{d}_{\mathrm{KL}}(p\|q_i) < 3\delta$ for $i \geq 2$, we then have

$$\mathbf{d}_{\mathrm{KL}}(p\|q_i) \leq \max\left\{3, \ln\frac{\alpha+\beta}{\alpha}\right\}\delta \quad \forall i = 1,2,\ldots,1/\delta \text{ and } \forall p \in ((i-1)\delta, i\delta].$$

This implies that

$$\mathbb{I}(p;b) - \mathbb{I}(q,b) \leq \max\left\{3, \ln\frac{\alpha+\beta}{\alpha}\right\}\delta.$$

On the other hand, from Lemma 10 and Lemma 11 in Lu (2020), we have

$$\begin{aligned}
\mathbb{I}(p;b) = &\frac{\alpha}{\alpha+\beta}\left(\psi(\alpha+1) - \ln\alpha\right) \\
&+ \frac{\beta}{\alpha+\beta}\left(\psi(\beta+1) - \ln\beta\right) - \left(\psi(\alpha+\beta+1) - \ln(\alpha+\beta)\right),
\end{aligned}$$

where ψ is the digamma function. From the digamma inequalities $\ln(x+0.5) \leq \psi(x+1) \leq \ln(x) + \frac{1}{2x}$ for $x > 0$ (Lemma 11 of Lu,

2020), we have $\psi(\alpha+1)-\ln\alpha \le \frac{1}{2\alpha}$, $\psi(\beta+1)-\ln\beta \le \frac{1}{2\beta}$, and $\psi(\alpha+\beta+1)-\ln(\alpha+\beta) > 0$. Consequently, we have $\mathbb{I}(p;b) < \frac{1}{\alpha+\beta}$. Thus we have

$$\mathbb{I}(p;b) - \mathbb{I}(q;b) \le \mathbb{I}(p;b) < \frac{1}{\alpha+\beta}.$$

Combining the above results, we have

$$\mathbb{I}(p;b) - \mathbb{I}(q,b) \le \min\left\{\max\left\{3, \ln\frac{\alpha+\beta}{\alpha}\right\}\delta, \frac{1}{\alpha+\beta}\right\}.$$

Finally, note that

$$\begin{aligned}
&\min\left\{\max\left\{3, \ln\frac{\alpha+\beta}{\alpha}\right\}\delta, \frac{1}{\alpha+\beta}\right\} \\
&\overset{(a)}{\le} \max\left\{3\delta, \min\left\{\delta\ln(\alpha+\beta), \frac{1}{\alpha+\beta}\right\}\right\} \\
&\overset{(b)}{\le} \max\left\{3\delta, \max_{x\ge 2}\min\left\{\delta\ln x, \frac{1}{x}\right\}\right\} \\
&\overset{(c)}{=} \max\left\{3\delta, \min_{x\ge 2}\max\left\{\delta\ln x, \frac{1}{x}\right\}\right\} \\
&\overset{(b)}{\le} \max\left\{3\delta, \delta\ln\left(\frac{2}{\delta}\right)\right\} = \max\left\{3, \ln\left(\frac{2}{\delta}\right)\right\}\delta,
\end{aligned}$$

where (a) follows from $\alpha \ge 1$, (b) follows from $\alpha+\beta \ge 2$, (c) follows from $\max_{x\ge 2}\min\left\{\delta\ln x, \frac{1}{x}\right\} = \min_{x\ge 2}\max\left\{\delta\ln x, \frac{1}{x}\right\}$ for $x \ge 2$ (note that $\delta \le 1/2$), and (d) follows by choosing $x = 2/\delta$ in $\max\left\{\delta\ln x, \frac{1}{x}\right\}$. □

Lemma A.4. For any positive integer N, let $p_1, \ldots, p_N$ be independent, beta-distributed random variables such that $p_n \sim \text{Beta}(\alpha_n, \beta_n)$ with $\alpha_n > 1$ and $\beta_n > 1$ for each n. Moreover, for each n, $\hat{p}_n$ is independent and distributed identically with p_n, and $b_n \sim \text{Bernoullil}(p_n)$ conditioned on p_n. Then, for all $\delta = 1/2, 1/3, 1/4, \ldots$,

$$\mathbb{E}\left[\min_{n\in\{1,\ldots,N\}}\left(\sqrt{6\mathbb{I}(q_n;b_n)\ln\frac{2N}{\delta}} - |q_n-\hat{q}_n| + 2\delta \right.\right.$$
$$\left.\left. + \sqrt{6\max\left\{3, \ln\left(\frac{2}{\delta}\right)\right\}\delta\ln\frac{2N}{\delta}}\right)\right] \ge 0,$$

where $q_n = \delta\lceil p_n/\delta\rceil$ and $\hat{q}_n = \delta\lceil \hat{p}_n/\delta\rceil$ are quantized approximations.

Proof. Notice that $q_n - \delta < p_n \le q_n$ and $\hat{q}_n - \delta < \hat{p}_n \le \hat{q}_n$. We now prove that $|q_n - \hat{q}_n| \le |p_n - \hat{p}_n| + \delta$. Without loss of generality, assume that $q_n \ge \hat{q}_n$ (the other case is symmetric), then we have

$$|q_n - \hat{q}_n| = q_n - \hat{q}_n \overset{(a)}{\le} q_n - \hat{p}_n \overset{(b)}{<} p_n + \delta - \hat{p}_n \le |p_n - \hat{p}_n| + \delta,$$

where (a) follows from $\hat{p}_n \le \hat{q}_n$, and (b) follows from $q_n < p_n + \delta$. Thus, we have $-|q_n - \hat{q}_n| + \delta \ge -|p_n - \hat{p}_n|$.

On the other hand, we have

$$\begin{aligned}
&\sqrt{6\mathbb{I}(q_n; b_n) \ln \frac{2N}{\delta}} + \sqrt{6 \max\left\{3, \ln\left(\frac{2}{\delta}\right)\right\} \delta \ln \frac{2N}{\delta}} \\
&\ge \sqrt{6\left(\mathbb{I}(q_n; b_n) + \max\left\{3, \ln\left(\frac{2}{\delta}\right)\right\} \delta\right) \ln \frac{2N}{\delta}} \\
&\ge \sqrt{6\mathbb{I}(p_n; b_n) \ln \frac{2N}{\delta}},
\end{aligned}$$

where the last inequality follows from Lemma A.3. Combining the above results, we have

$$\begin{aligned}
&\sqrt{6\mathbb{I}(q_n; b_n) \ln \frac{2N}{\delta}} - |q_n - \hat{q}_n| + 2\delta + \sqrt{6 \max\left\{3, \ln\left(\frac{2}{\delta}\right)\right\} \delta \ln \frac{2N}{\delta}} \\
\ge &\sqrt{6\mathbb{I}(p_n; b_n) \ln \frac{2N}{\delta}} - |p_n - \hat{p}_n| + \delta.
\end{aligned}$$

Then, the result of this lemma directly follows from Lemma A.2. □

A.2 Optimism

The following conjecture concerns the optimistic behavior of Thompson sampling for the "ring" MDPs considered in Section 4.5.2. Our analysis in this appendix assumes that the conjecture holds.

Conjecture A.2.1. Under the "ring" MDPs considered in Section 4.5.2, for any episode ℓ and any time $t = \ell\tau, \ldots, (\ell+1)\tau - 1$ within the episode,

$$\mathbb{E}[V_{\tau,\rho}(S_t)|P_{\ell\tau}] \le \mathbb{E}[\hat{V}_\ell(S_t)|P_{\ell\tau}].$$

Note that Conjecture A.2.1 always holds with equality for $t = \ell\tau$ and $t = (\ell+1)\tau - 1$, $\forall \ell = 0, 1, \ldots$. To understand why, note that for $t = \ell\tau$, $S_{\ell\tau} = S_0$ is deterministic, and $V_{\tau,\rho}$ and $\hat{V}_\ell$ are i.i.d. conditioned on $P_{\ell\tau}$. Thus,

$$\mathbb{E}[V_{\tau,\rho}(S_{\ell\tau})|P_{\ell\tau}] = \mathbb{E}[V_{\tau,\rho}(S_0)|P_{\ell\tau}] = \mathbb{E}[\hat{V}_\ell(S_0)|P_{\ell\tau}] = \mathbb{E}[\hat{V}_\ell(S_{\ell\tau})|P_{\ell\tau}].$$

Note that the reward model r is deterministic. Consequently, $V_{\tau,\rho}(s) = \hat{V}_\ell(s) = \max_{a\in\mathcal{A}} r(s, a, S_0)$ for all $s \in \mathcal{S}_{\tau-1}$. Thus, by definition, for $t = (\ell+1)\tau - 1$, we have $S_t \in \mathcal{S}_{\tau-1}$ and

$$\mathbb{E}[V_{\tau,\rho}(S_t)|P_{\ell\tau}] = \mathbb{E}[\hat{V}_\ell(S_t)|P_{\ell\tau}] = \mathbb{E}\left[\max_{a\in\mathcal{A}} r(S_t, a, S_0)\right].$$

We leave the proof of Conjecture A.2.1 for $\ell\tau < t < (\ell+1)\tau - 1$ in this "ring" example for future work. In the remainder of this section, we provide some numerical results suggesting that Conjecture A.2.1 holds.

It is worth pointing out that Conjecture A.2.1 might not hold in more general problems. In particular, if the prior distribution admits *generalization* of transition probabilities across state-action pairs – for example, if ρ is *correlated* across state-action pairs – this conjecture may fail to hold.

A.2.1 Numerical Verification

We now provide numerical verification of Conjecture A.2.1. Note that Conjecture A.2.1 states that for any episode ℓ and any time $t = \ell\tau, \ldots, (\ell+1)\tau - 1$,

$$\mathbb{E}[\hat{V}_\ell(S_t) - V_{\tau,\rho}(S_t)|P_{\ell\tau}] \geq 0. \tag{A.1}$$

We numerically verify this conjecture as follows: we sweep over $M = 5, 10, 20$ and $\tau = 3, 8, 10, 20, 30$. For each (M, τ) pair, we rerun the Thompson sampling algorithm on the "ring" MDP with state space $\mathcal{S} = \{0, \ldots, M-1\} \times \{0, \ldots, \tau - 1\}$ for 50 times, and each time we run for 300 episodes. Then, we numerically test Conjecture A.2.1 every three episodes. Moreover, when we test this conjecture, we test it for every time $t = \ell\tau, \ldots, (\ell+1)\tau - 1$. Thus, in total we test Conjecture A.2.1 for $3 \times 50 \times 100 \times (3 + 8 + 10 + 20 + 30) = 1,065,000$ times.

In particular, we test if the left-hand side of (A.1) is non-negative, as well as if it is strictly positive in some cases. The former indicates if this conjecture holds, while the latter indicates if this conjecture holds with strict inequality in some cases. The following procedure illustrates how we compute a point estimate of the left-hand side of Equation (A.1), as well as its standard error, at a given episode ℓ based on the Monte-Carlo simulation. Specifically, for each round of Monte-Carlo simulation $i = 1, 2, \ldots, L$:

1. sample transition models ρ, $\hat{\rho}$ i.i.d. from $P_{\ell\tau}$
2. compute $V_{\tau,\rho}$, the optimal state value function under ρ
3. compute $\hat{V}_\ell$ and $\hat{\pi}$, which are respectively the optimal state value function and an optimal policy under $\hat{\rho}$
4. for all $t = \ell\tau, \ldots, (\ell+1)\tau - 1$, compute ν_t, the state distribution of S_t, under policy $\hat{\pi}$ and transition model ρ
5. finally, compute $d^i_{\ell,t} = \sum_s \nu_t(s) \left[\hat{V}_\ell(s) - V_{\tau,\rho}(s)\right]$ for all $t = \ell\tau, \ldots, (\ell+1)\tau - 1$.

We carry out $L = 10,000$ Monte-Carlo simulations, and compute the point estimate of the left-hand side of Equation (A.1) and its standard error according to

$$\bar{d}_{\ell,t} = \frac{1}{L}\sum_{i=1}^{L} d^i_{\ell,t} \quad \text{and} \quad \texttt{stderr}_{\ell,t} = \frac{1}{L}\sqrt{\sum_{i=1}^{L}\left(d^i_{\ell,t} - \bar{d}_{\ell,t}\right)^2}.$$

For any $\kappa > 0$, we define the upper confidence bound (UCB) and the lower confidence bound (LCB) parameterized by κ as

$$\begin{aligned}\texttt{UCB}_{\ell,t}(\kappa) &= \bar{d}_{\ell,t} + \kappa \cdot \texttt{stderr}_{\ell,t} \\ \texttt{LCB}_{\ell,t}(\kappa) &= \bar{d}_{\ell,t} - \kappa \cdot \texttt{stderr}_{\ell,t}\end{aligned}$$

We report the fractions of cases for which $\texttt{UCB}_{\ell,t}(\kappa) < 0$ or $\texttt{LCB}_{\ell,t}(\kappa) > 0$ for a wide range of κ, and compare it with the Gaussian benchmark $1 - \Phi(\kappa)$, where $\Phi(\cdot)$ is the cumulative distribution function of the standard normal distribution $N(0,1)$. Intuitively, a negative UCB suggests that

Table A.1: Numerical verification of Conjecture A.2.1

κ	frac. $\texttt{UCB}_{\ell,t}(\kappa) < 0$	frac. $\texttt{LCB}_{\ell,t}(\kappa) > 0$	$1 - \Phi(\kappa)$
1	6.31315%	45.29878%	15.86553%
1.5	2.49822%	34.76901%	6.68072%
2	0.80826%	26.56451%	2.27501%
2.5	0.20451%	20.3923%	0.62097%
3	0.04085%	15.8954%	0.13499%
3.5	0.00516%	12.53643%	0.02326%
4	0.00075%	10.02516%	0.00317%
4.5	0%	8.10714%	0.00034%
5	0%	6.62685%	0.00005%
5.5	0%	5.4477%	0%
6	0%	4.50235%	0%

the conjecture does not hold, while a positive LCB suggests that the conjecture holds with strict inequality. The results are summarized in Table A.1.

The experiment results suggest that Conjecture A.2.1 holds in the "ring" MDP. In particular, for each chosen κ, the fraction of negative UCBs is much smaller than the benchmark $1 - \Phi(\kappa)$. Moreover, they also suggest that there are cases where the conjecture holds with strict inequality. In particular, for each chosen κ, the fraction of positive LCBs is much larger than the benchmark.

A.3 Regret

The following results pertain to application of Thompson sampling to the "ring" episodic MDP described in Section 4.5.2.

Lemma A.5. Assume that Conjecture A.2.1 holds. Then, for all integers $m \geq 2$ and times t,

$$\mathbb{E}[V_*(H_t) - Q_*(H_t, A_t) - g(\delta)\tau^2]_+^2 \leq 6\tau^3 \ln \frac{2\mathcal{SA}}{\delta} \mathbb{I}(\chi; H_{t:(\ell+1)\tau} | P_t),$$

where $\delta = 1/m$, $g(\delta) = 3\delta + \sqrt{6 \max\left\{3, \ln\left(\frac{2}{\delta}\right)\right\} \delta \ln \frac{2\mathcal{SA}}{\delta}}$, and χ is a quantized approximation of ρ for which $\chi(s+1|s,a) = \delta\lceil \rho(s+1|s,a)/\delta \rceil$ and $\chi(s-1|s,a) = 1 - \chi(s+1|s,a)$.

Proof. Time t resides in episode $\ell = \lfloor t/\tau \rfloor$. Recall that $\hat{\rho}_\ell$ is the observation probability function sampled at the start of episode ℓ. Let $\hat{\chi}_\ell$ be the corresponding quantization, for which $\hat{\chi}_\ell(s+1|s,a) = \delta\lceil \hat{\rho}_\ell(s+1|s,a)/\delta \rceil$ and $\hat{\chi}_\ell(s-1|s,a) = 1 - \hat{\chi}_\ell(s+1|s,a)$. Also note that in this problem, since $P_t = \mathbb{P}(\cdot|H_t)$, S_t is conditionally deterministic given P_t; thus, conditioning on H_t is equivalent to conditioning on (P_t, S_t) and conditioning on P_t. Let

$$I_t = \mathbb{I}(\chi(S_t+1|S_t,A_t); A_t, S_{t+1}|P_t = P_t).$$

Note that

$$\mathbb{I}(\chi(S_t+1|S_t,A_t); A_t, S_{t+1}|P_t = P_t) = \mathbb{I}(\chi; A_t, S_{t+1}|P_t = P_t).$$

By the chain rule of mutual information, $\mathbb{I}(\chi; H_{t:(\ell+1)\tau}|P_t = P_t) = \sum_{k=t}^{(\ell+1)\tau} \mathbb{E}[I_k|P_t]$. Recall that $\mathcal{S} = \mathcal{S}_0 \cup \cdots \cup \mathcal{S}_{\tau-1}$ and $|\mathcal{S}_0| = \cdots = |\mathcal{S}_{\tau-1}| = M$.

By Lemma A.4,

$$\begin{aligned}
&\mathbb{E}\left[|\hat{\rho}_\ell(S_t+1|S_t,A_t) - \rho(S_t+1|S_t,A_t)| \,\middle|\, P_t\right] \\
\le &\mathbb{E}\left[|\hat{\chi}_\ell(S_t+1|S_t,A_t) - \chi(S_t+1|S_t,A_t)| \,\middle|\, P_t\right] + \delta \\
\le &\mathbb{E}\left[\sqrt{6I_t \ln\frac{2\mathcal{S}\mathcal{A}}{\delta}}\middle| P_t\right] + \underbrace{3\delta + \sqrt{6\max\left\{3, \ln\left(\frac{2}{\delta}\right)\right\}\delta \ln\frac{2\mathcal{S}\mathcal{A}}{\delta}}}_{g(\delta)},
\end{aligned}$$

where we define $g(\delta) = 3\delta + \sqrt{6\max\left\{3, \ln\left(\frac{2}{\delta}\right)\right\}\delta \ln\frac{2\mathcal{S}\mathcal{A}}{\delta}}$ to simplify the exposition. It follows that, at time $t = (\ell+1)\tau - 1$,

$$\begin{aligned}
&\mathbb{E}[\hat{V}_\ell(S_t) - Q_{\tau,\rho}(S_t,A_t)|P_t] \\
= &\mathbb{E}\left[\max_{a\in\mathcal{A}} \sum_{s'\in\mathcal{S}} \hat{\rho}_\ell(s'|S_t,a) r(S_t,a,s') - \sum_{s'\in\mathcal{S}} \rho(s'|S_t,A_t) r(S_t,A_t,s')|P_t\right] \\
= &\mathbb{E}\left[\sum_{s'\in\mathcal{S}} (\hat{\rho}_\ell(s'|S_t,A_t) - \rho(s'|S_t,A_t)) r(S_t,A_t,s')\middle| P_t\right] \\
\le &\frac{1}{2}\mathbb{E}\left[\sum_{s'\in\mathcal{S}} |\hat{\rho}_\ell(s'|S_t,A_t) - \rho(s'|S_t,A_t)| \,\middle|\, P_t\right]
\end{aligned}$$

$$
\begin{aligned}
&= \mathbb{E}\left[\left|\hat{\rho}_\ell(S_t+1|S_t,A_t) - \rho(S_t+1|S_t,A_t)\right| \Big| P_t\right] \\
&\le \mathbb{E}\left[\sqrt{6I_t \ln\frac{2\mathcal{S}\mathcal{A}}{\delta}}\Big| P_t\right] + g(\delta).
\end{aligned}
$$

Similarly, at times $t = \ell\tau, \ell\tau+1, \ldots, (\ell+1)\tau - 2$,

$$
\begin{aligned}
&\mathbb{E}[\hat{V}_\ell(S_t) - Q_{\tau,\rho}(S_t,A_t)\Big| P_t] \\
&= \mathbb{E}\Bigg[\max_{a\in\mathcal{A}} \sum_{s'\in\mathcal{S}} \hat{\rho}_\ell(s'|S_t,a)(r(S_t,a,s') + \hat{V}_\ell(s')) \\
&\qquad\qquad - \sum_{s'\in\mathcal{S}} \rho(s'|S_t,A_t)(r(S_t,A_t,s') + V_{\tau,\rho}(s'))\Big| P_t\Bigg] \\
&= \mathbb{E}\Bigg[\sum_{s'\in\mathcal{S}} \hat{\rho}_\ell(s'|S_t,A_t)(r(S_t,A_t,s') + \hat{V}_\ell(s')) \\
&\qquad\qquad - \sum_{s'\in\mathcal{S}} \rho(s'|S_t,A_t)(r(S_t,A_t,s') + V_{\tau,\rho}(s'))\Big| P_t\Bigg] \\
&= \mathbb{E}\Bigg[\sum_{s'\in\mathcal{S}} \hat{\rho}_\ell(s'|S_t,A_t)(r(S_t,A_t,s') + \hat{V}_\ell(s'))) \\
&\qquad\qquad - \sum_{s'\in\mathcal{S}} \rho(s'|S_t,A_t)(r(S_t,A_t,s') + \hat{V}_\ell(s'))\Big| P_t\Bigg] \\
&\qquad\qquad + \mathbb{E}\left[\sum_{s'\in\mathcal{S}} \rho(s'|S_t,A_t)(\hat{V}_\ell(s') - V_{\tau,\rho}(s'))\Big| P_t\right] \\
&= \mathbb{E}\left[\sum_{s'\in\mathcal{S}} (\hat{\rho}_\ell(s'|S_t,A_t) - \rho(s'|S_t,A_t))(r(S_t,A_t,s') + \hat{V}_\ell(s'))\Big| P_t\right] \\
&\qquad\qquad + \mathbb{E}\left[\hat{V}_\ell(S_{t+1}) - V_{\tau,\rho}(S_{t+1})|P_t\right] \\
&\le \frac{\tau}{2}\mathbb{E}\left[\sum_{s'\in\mathcal{S}} \left|\hat{\rho}_\ell(s'|S_t,A_t) - \rho(s'|S_t,A_t)\right| \Big| P_t\right] \\
&\qquad\qquad + \mathbb{E}\left[\hat{V}_\ell(S_{t+1}) - V_{\tau,\rho}(S_{t+1})|P_t\right] \\
&= \tau\mathbb{E}\left[\left|\hat{\rho}_\ell(S_t+1|S_t,A_t) - \rho(S_t+1|S_t,A_t)\right| \Big| P_t\right]
\end{aligned}
$$

$$+\mathbb{E}\left[\hat{V}_\ell(S_{t+1}) - V_{\tau,\rho}(S_{t+1})|P_t\right]$$
$$\leq \tau\mathbb{E}\left[\sqrt{6I_t \ln\frac{2\mathcal{SA}}{\delta}}\Big|P_t\right] + g(\delta)\tau + \mathbb{E}\left[\hat{V}_\ell(S_{t+1}) - Q_{\tau,\rho}(S_{t+1}, A_{t+1})|P_t\right].$$

It follows that

$$\begin{aligned}
&\mathbb{E}[\hat{V}_\ell(S_t) - Q_{\tau,\rho}(S_t, A_t)|P_t] \\
\leq &\sum_{k=t}^{(\ell+1)\tau-1}\left(\tau\mathbb{E}\left[\sqrt{6I_k \ln\frac{2\mathcal{SA}}{\delta}}\Big|P_t\right] + g(\delta)\tau\right) \\
\leq &\tau^{3/2}\sqrt{6\sum_{k=t}^{(\ell+1)\tau-1}\mathbb{E}[I_k|P_t]\ln\frac{2\mathcal{SA}}{\delta}} + g(\delta)\tau^2 \\
\leq &\tau^{3/2}\sqrt{6\mathbb{I}(\chi; H_{t:(\ell+1)\tau}|P_t \leftarrow P_t)\ln\frac{2\mathcal{SA}}{\delta}} + g(\delta)\tau^2.
\end{aligned}$$

Further, under Conjecture A.2.1, for all episode ℓ and time $t = \ell\tau, \ldots, (\ell+1)\tau - 1$,

$$\begin{aligned}
\mathbb{E}[V_*(H_t) - Q_*(H_t, A_t)|P_{\ell\tau}] &= \mathbb{E}[V_{\tau,\rho}(S_t) - Q_{\tau,\rho}(S_t, A_t)|P_{\ell\tau}] \\
&\leq \mathbb{E}[\hat{V}_\ell(S_t) - Q_{\tau,\rho}(S_t, A_t)|P_{\ell\tau}],
\end{aligned}$$

where the last inequality follows from Conjecture A.2.1. Therefore, we have

$$\begin{aligned}
&\mathbb{E}[V_*(H_t) - Q_*(H_t, A_t) - g(\delta)\tau^2|P_{\ell\tau}] \\
\leq &\tau^{3/2}\sqrt{6\ln\frac{2\mathcal{SA}}{\delta}}\mathbb{E}\left[\sqrt{\mathbb{I}(\chi; H_{t:(\ell+1)\tau}|P_t \leftarrow P_t)}\Big|P_{\ell\tau}\right],
\end{aligned}$$

which further implies that

$$\begin{aligned}
&\mathbb{E}[V_*(H_t) - Q_*(H_t, A_t) - g(\delta)\tau^2] \\
\leq &\tau^{3/2}\sqrt{6\ln\frac{2\mathcal{SA}}{\delta}}\mathbb{E}\left[\sqrt{\mathbb{I}(\chi; H_{t:(\ell+1)\tau}|P_t \leftarrow P_t)}\right].
\end{aligned}$$

Since the right-hand side of the above inequality is positive, we then have

$$\begin{aligned}&\mathbb{E}[V_*(H_t) - Q_*(H_t, A_t) - g(\delta)\tau^2]_+ \\ \leq\ &\tau^{3/2}\sqrt{6\ln\frac{2\mathcal{SA}}{\delta}}\mathbb{E}\left[\sqrt{\mathbb{I}(\chi; H_{t:(\ell+1)\tau}|P_t \leftarrow P_t)}\right].\end{aligned}$$

Therefore, we have

$$\begin{aligned}&\mathbb{E}[V_*(H_t) - Q_*(H_t, A_t) - g(\delta)\tau^2]_+^2 \\ \leq\ &6\tau^3\ln\frac{2\mathcal{SA}}{\delta}\left(\mathbb{E}\left[\sqrt{\mathbb{I}(\chi; H_{t:(\ell+1)\tau}|P_t \leftarrow P_t)}\right]\right)^2 \\ \leq\ &6\tau^3\ln\frac{2\mathcal{SA}}{\delta}\mathbb{E}\left[\mathbb{I}(\chi; H_{t:(\ell+1)\tau}|P_t \leftarrow P_t)\right] \\ =\ &6\tau^3\ln\frac{2\mathcal{SA}}{\delta}\mathbb{I}(\chi; H_{t:(\ell+1)\tau}|P_t).\end{aligned}$$

□

Finally, we prove the following theorem based on Lemma A.5 and Theorem 4.3, under Conjecture A.2.1.

Theorem A.6. Assume Conjecture A.2.1 holds. Then, for all integers $m \geq 2$ and times t,

$$\begin{aligned}&\mathrm{Regret}(T|\pi_{\mathrm{TS}}) \\ \leq\ &\tau^2\sqrt{6\mathcal{SA}T\ln\left(\frac{1}{\delta}\right)\ln\left(\frac{2\mathcal{SA}}{\delta}\right)} + \left[3\delta + \sqrt{6\max\left\{3, \ln\left(\frac{2}{\delta}\right)\right\}\delta\ln\left(\frac{2\mathcal{SA}}{\delta}\right)}\right]\tau^2 T \\ =\ &\mathcal{O}\left(\tau^2\sqrt{\log\left(\frac{1}{\delta}\right)\log\left(\frac{\mathcal{SA}}{\delta}\right)}\left[\sqrt{\mathcal{SA}T} + T\sqrt{\delta}\right]\right),\end{aligned}$$

where $\delta = 1/m$.

Proof. By Lemma A.5, we first prove that $\Gamma_{\tau,\epsilon,t} \leq 6\tau^4\ln\frac{2\mathcal{SA}}{\delta}$ for

$$\epsilon = g(\delta)\tau^2 = \left[3\delta + \sqrt{6\max\left\{3, \ln\left(\frac{2}{\delta}\right)\right\}\delta\ln\frac{2\mathcal{SA}}{\delta}}\right]\tau^2.$$

From Lemma A.5, we have

$$\mathbb{E}[V_*(H_t) - Q_*(H_t, A_t) - \epsilon]_+^2 \leq 6\tau^3\ln\frac{2\mathcal{SA}}{\delta}\mathbb{I}(\chi; H_{t:(\ell+1)\tau}|P_t).$$

Also, note that in this case, the environment $\mathcal{E}$ determines the proxy $\tilde{\mathcal{E}}$ and hence the target χ, and consequently

$$\begin{aligned}&\mathbb{I}(\chi;\mathcal{E}|P_t) - \mathbb{I}(\chi;\mathcal{E}|P_{t+\tau})\\ =&\mathbb{H}(\chi|P_t) - \mathbb{H}(\chi|P_{t+\tau}) = \mathbb{I}(\chi;H_{t:t+\tau}|P_t) \geq \mathbb{I}(\chi;H_{t:(\ell+1)\tau}|P_t),\end{aligned}$$

where $\mathbb{H}$ is the entropy function in nats. Consequently, by definition of $\Gamma_{\tau,\epsilon,t}$, we have

$$\begin{aligned}\Gamma_{\tau,\epsilon,t} &= \frac{\mathbb{E}[V_*(H_t) - Q_*(H_t,A_t) - \epsilon_t]_+^2}{(\mathbb{I}(\chi;\mathcal{E}|P_t) - \mathbb{I}(\chi;\mathcal{E}|P_{t+\tau}))/\tau}\\ &= \frac{\mathbb{E}[V_*(H_t) - Q_*(H_t,A_t) - \epsilon_t]_+^2}{\mathbb{I}(\chi;H_{t:t+\tau}|P_t)/\tau}\\ &\leq \frac{\mathbb{E}[V_*(H_t) - Q_*(H_t,A_t) - \epsilon_t]_+^2}{\mathbb{I}(\chi;H_{t:(\ell+1)\tau}|P_t)/\tau} \leq 6\tau^4 \ln\frac{2\mathcal{SA}}{\delta}.\end{aligned}$$

Then, by Theorem 4.3, we have

$$\mathrm{Regret}(T|\pi_{\mathrm{TS}}) \leq \sqrt{\mathbb{I}(\chi;\mathcal{E})\sum_{t=0}^{T-1}\Gamma_{\tau,\epsilon,t}} + T\epsilon \leq \sqrt{6\tau^4 T\mathbb{H}(\chi)\ln\frac{2\mathcal{SA}}{\delta}} + T\epsilon.$$

Note that $\mathbb{H}(\chi) \leq \mathcal{SA}\ln\left(\frac{1}{\delta}\right)$, we have

$$\begin{aligned}&\mathrm{Regret}(T|\pi_{\mathrm{TS}})\\ \leq&\tau^2\sqrt{6T\mathcal{SA}\ln\left(\frac{1}{\delta}\right)\ln\left(\frac{2\mathcal{SA}}{\delta}\right)} + \left[3\delta + \sqrt{6\max\left\{3,\ln\left(\frac{2}{\delta}\right)\right\}\delta\ln\left(\frac{2\mathcal{SA}}{\delta}\right)}\right]\tau^2 T\\ =&\mathcal{O}\left(\tau^2\sqrt{\ln\left(\frac{1}{\delta}\right)\ln\left(\frac{\mathcal{SA}}{\delta}\right)}\left[\sqrt{\mathcal{SA}T} + T\sqrt{\delta}\right]\right).\end{aligned}$$

This concludes the proof. □

B

Analysis of IDS in an Episodic Environment

Consider the environment and agent described in Section 4.5.3. Let $\tilde{\mathcal{E}} = \mathcal{E}$ and let the epistemic state indicate the value of $r_{\tau-1}$ or that it has not been observed. The value-IDS agent in Section 4.5.3 selects actions by optimizing

$$\min_{\nu \in \Delta_{\mathcal{A}}} \frac{\mathbb{E}\left[V_*(S_t) - Q_*(S_t, \tilde{A}_t) | X_t\right]^2}{\mathbb{I}(\pi_*(\cdot|S_t); \tilde{A}_t, Q_*(S_t, \tilde{A}_t) | X_t \leftarrow X_t)}$$

where $\tilde{A}_t$ is drawn from ν.

We will bound the τ-information ratio. Since $\mathcal{E}$ determines $\chi = \pi_*$, the τ-information ratio simplifies to

$$\Gamma_{\tau,t} = \frac{\mathbb{E}[V_*(S_t) - Q_*(S_t, A_t)]^2}{(\mathbb{H}(\pi_*|P_t) - \mathbb{H}(\pi_*|P_{t+\tau}))/\tau}.$$

To do this, we will find a uniform bound $\overline{\Gamma}$ on the conditional information ratio,

$$\tilde{\Gamma}_{\tau,t} = \frac{\mathbb{E}[V_*(S_t) - Q_*(S_t, A_t)|X_t]^2}{\mathbb{E}[\mathbb{H}(\pi_*|P_t \leftarrow P_t) - \mathbb{H}(\pi_*|P_{t+\tau} \leftarrow P_{t+\tau})|X_t]/\tau} \equiv \frac{\Delta_t^2}{I_t}.$$

If $\tilde{\Gamma}_{\tau,t} \leq \overline{\Gamma}$ for all t, then

$$\begin{aligned}
&\mathbb{E}[V_*(S_t) - Q_*(S_t, A_t)]^2 \\
&\leq \mathbb{E}\left[\mathbb{E}[V_*(S_t) - Q_*(S_t, A_t)|X_t]^2\right] \\
&= \mathbb{E}\left[\tilde{\Gamma}_{\tau,t}\mathbb{E}[\mathbb{H}(\pi_*|P_t \leftarrow P_t) - \mathbb{H}(\pi_*|P_{t+\tau} \leftarrow P_{t+\tau})|X_t]/\tau\right] \\
&\leq \overline{\Gamma}\left(\mathbb{H}(\pi_*|P_t) - \mathbb{H}(\pi_*|P_{t+\tau})\right)/\tau,
\end{aligned}$$

and so $\Gamma_{\tau,t} \leq \overline{\Gamma}$.

Let's consider an episode where the agent is still uncertain about the environment. Otherwise, the agent is done learning and the conditional information ratio will be zero. Let us overload $\tilde{\Gamma}_s$ to denote the conditional τ-step information ratio for convenience for $s = 0, \ldots, \tau - 2$, with the intention that $\tilde{\Gamma}_s = \tilde{\Gamma}_{\tau,t}$ given $S_t = s$ and $P_t = \text{null}$. Similarly, we use Δ_s and I_s to denote the corresponding regret and information gain. Further, let $\nu_{a,s}$ denote the probability that value-IDS selects action a given $S_t = s$ and $P_t = \text{null}$.

Let $\overline{r}_{\tau-2} = 0$ and $\overline{r}_s = \max\{r_{s+1}, \ldots, r_{\tau-2}\}$ for $0 \leq s \leq \tau - 3$ denote the maximum exit rewards starting from state $s+1$ to state $\tau - 2$. Then, for $0 \leq s \leq \tau - 2$, we have

$$\begin{aligned}
Q_*(s, 0) &= r_s, \\
Q_*(s, 1) &= \begin{cases} 1 & \text{w.p. } \frac{1}{2}, \\ \overline{r}_s & \text{w.p. } \frac{1}{2}. \end{cases}
\end{aligned}$$

Let $\Delta_{a,s} = \mathbb{E}[V_*(s) - Q_*(s, a)]$. We have for states $s = 0, \ldots, \tau - 2$

$$\Delta_{0,s} = \tfrac{1}{2}(1 - r_s) + \tfrac{1}{2}(\overline{r}_s - r_s)_+, \text{ and } \Delta_{1,s} = \tfrac{1}{2}(r_s - \overline{r}_s)_+.$$

The information gain in the denominator of value-IDS is zero for action 0, and 1 bit for action 1. Thus, value-IDS selects probabilities $(\nu_{0,s}, \nu_{1,s})$ that minimizes

$$\min_{\nu'_{0,s}, \nu'_{1,s}} \frac{(\nu'_{0,s}\Delta_{0,s} + \nu'_{1,s}\Delta_{1,s})^2}{\nu'_{1,s}}.$$

Note that

$$\begin{aligned}
\tfrac{1}{\nu'_{1,s}}(\nu'_{0,s}\Delta_{0,s} + \nu'_{1,s}\Delta_{1,s})^2 &= \tfrac{1}{\nu'_{1,s}}((1 - \nu'_{1,s})\Delta_{0,s} + \nu'_{1,s}\Delta_{1,s})^2 \\
&= (\Delta_{1,s} - \Delta_{0,s})^2\nu'_{1,s} + \tfrac{\Delta^2_{0,s}}{\nu'_{1,s}} + 2\Delta_{0,s}(\Delta_{1,s} - \Delta_{0,s}).
\end{aligned}$$

Thus, the minimizer

$$\nu_{1,s} = \min\left(\frac{\Delta_{0,s}}{(\Delta_{1,s} - \Delta_{0,s})_+}, 1\right).$$

Also note that $\nu_{1,s} < 1$ if and only if

$$2\Delta_{0,s} < \Delta_{1,s} \quad \Leftrightarrow \quad r_s > \frac{2 + \overline{r}_s}{3}.$$

Now, we show inductively that $\tilde{\Gamma}_s \leq \frac{\tau}{8}$ for all $0 \leq s \leq \tau - 2$.

For the base case $s = \tau - 2$, note that the average τ-step information gain is equal to $\nu_{1,s}/\tau$. Thus,

$$\tilde{\Gamma}_s = \frac{\Delta_s^2}{\nu_{1,s}/\tau} = \begin{cases} \tau(1 - r_s)(2r_s - \overline{r}_s - 1) & \nu_{1,s} < 1 \\ \frac{\tau}{4}\left((r_s - \overline{r}_s)_+\right)^2 & \nu_{1,s} = 1. \end{cases}$$

For $s = \tau - 2$, $\overline{r}_s = 0$. We have $(1 - r_s)(2r_s - 1) \leq \frac{1}{8}$, and $\frac{1}{4}r_s^2 \leq \frac{1}{9}$ if $\nu_{1,s} = 1$, since $\nu_{1,s} = 1$ implies that $r_s \leq \frac{2}{3}$. Thus, the base case holds.

Our induction hypothesis is that $\tilde{\Gamma}_{s'} \leq \frac{\tau}{8}$ for all $s + 1 \leq s' \leq \tau - 2$. Now let's consider state $1 \leq s < \tau - 2$. Let $\overline{s} = \min\{s' : s < s' \leq \tau - 2, \nu_{1,s'} < 1\}$ and if the set is empty, let $\overline{s} = \text{null}$. We have three cases to consider.

(i) If $\overline{s}$ is null, then $\nu_{1,s'} = 1$ for all $s' = s + 1, \ldots, \tau - 2$. Thus, the average information gain $I_s = \nu_{1,s}/\tau$. Then, similar to the base case,

$$\tilde{\Gamma}_s = \frac{\Delta_s^2}{\nu_{1,s}/\tau} = \begin{cases} \tau(1 - r_s)(2r_s - \overline{r}_s - 1) & \nu_{1,s} < 1 \\ \frac{\tau}{4}\left((r_s - \overline{r}_s)_+\right)^2 & \nu_{1,s} = 1. \end{cases}$$

Now, $(1 - r_s)(2r_s - \overline{r}_s - 1) \leq \frac{1}{8}(1 - \overline{x}_s)^2 \leq \frac{1}{8}$. When $\nu_{1,s} = 1$, we have $\frac{1}{4}\left((r_s - \overline{r}_s)_+\right)^2 \leq \frac{(1-\overline{r}_s)^2}{9} \leq \frac{1}{9}$ since $\nu_{1,s} = 1$ implies that $r_s \leq \frac{2+\overline{r}_s}{3}$. Therefore, we have $\tilde{\Gamma}_s \leq \frac{\tau}{8}$.

(ii) If $\overline{s}$ is not null and $\nu_{1,s} = 1$, then the average information gain $I_s = I_{\overline{s}}$. Further, we have

$$\Delta_s = \Delta_{1,s} = \tfrac{1}{2}(r_s - \overline{r}_s),$$

and

$$\begin{aligned}\Delta_{\overline{s}} &= \nu_{0,\overline{s}}\Delta_{0,\overline{s}} + \nu_{1,\overline{s}}\Delta_{1,\overline{s}} \\ &= \left(1 - \frac{\Delta_{0,\overline{s}}}{\Delta_{1,\overline{s}} - \Delta_{0,\overline{s}}}\right)\Delta_{0,\overline{s}} + \frac{\Delta_{0,\overline{s}}}{\Delta_{1,\overline{s}} - \Delta_{0,\overline{s}}}\Delta_{1,\overline{s}} \\ &= 2\Delta_{0,\overline{s}} = 1 - r_{\overline{s}}.\end{aligned}$$

Since $\overline{r}_s \geq r_{\overline{s}}$ by definition, we have $\Delta_s \leq \frac{1}{2}\Delta_{\overline{s}}$. Therefore, $\tilde{\Gamma}_s \leq \frac{1}{4}\tilde{\Gamma}_{\overline{s}}$, and by the induction hypothesis, $\tilde{\Gamma}_s \leq \frac{\tau}{32}$.

(iii) If $\overline{s}$ is not null and $\nu_{1,s} < 1$, then the average information gain $I_s = \nu_{1,s} I_{\overline{s}}$. Since,

$$\tilde{\Gamma}_s = \frac{\Delta_s^2}{I_s} = \frac{\Delta_s^2}{\nu_{1,s} I_{\overline{s}}} = \frac{\Delta_s^2}{\nu_{1,s}\Delta_{\overline{s}}^2}\Gamma_{\overline{s}},$$

we will upper bound $\frac{\Delta_s^2}{\nu_{1,s}\Delta_{\overline{s}}^2}$ and then apply the inductive hypothesis. We have

$$\frac{\Delta_s^2}{\nu_{1,s}\Delta_{\overline{s}}^2} = \frac{(1-r_s)(2r_s - \overline{r}_s - 1)}{(1-r_{\overline{s}})^2} \leq \frac{(1-r_s)(2r_s - \overline{r}_s - 1)}{(1-\overline{r}_s)^2}.$$

Write $y = 1 - \overline{r}_s$ and define $\alpha = \frac{1-r_s}{1-\overline{r}_s}$. We have

$$\frac{(1-r_s)(2r_s - \overline{r}_s - 1)}{(1-\overline{r}_s)^2} = \frac{\alpha y(y - 2\alpha y)}{y^2} = \alpha(1-2\alpha) \leq \frac{1}{8}.$$

Therefore, $\frac{\Delta_s^2}{\nu_{1,s}\Delta_{\overline{s}}^2} \leq \frac{1}{8}$, and $\tilde{\Gamma}_s \leq \frac{1}{8}\tilde{\Gamma}_{\overline{s}}$. By the induction hypothesis, $\tilde{\Gamma}_s \leq \frac{\tau}{64}$.

Therefore, $\tilde{\Gamma}_s \leq \frac{\tau}{8}$ for all $0 \leq s \leq \tau - 2$. This implies that for any exit rewards $r_0, \ldots, r_{\tau-2} \in [0, 1)$, the τ-information ratio $\Gamma_{\tau,t} \leq \frac{\tau}{8}$ for all t. By Theorem 4.3, this implies a regret bound

$$\text{Regret}(T) \leq \sqrt{\tfrac{1}{8}\tau T \mathbb{H}(\pi_*)} = \sqrt{\tfrac{1}{8}\tau T}.$$

C

Convexity and Support of Value-IDS

The following result and proof are based on an analysis from Russo and Van Roy (2014a) and Russo and Van Roy (2018).

Theorem C.1. For all vectors $\alpha, \beta \in \Re^N$ and functions $\psi : \Re^N \mapsto \Re$ of the form $\psi(\nu) = (\nu^\top \alpha)^2 / \nu^\top \beta$, ψ is convex on $\{\nu \in \Re^N : \nu^\top \beta > 0\}$, and there exists a vector $\nu^* \in \Delta_N$ such that $|\{n : \nu_n^* > 0\}| \leq 2$ and $\psi(\nu^*) = \min_{\nu \in \Delta_N} \psi(\nu)$.

Proof. Consider a function $\phi : \Re^2 \mapsto \Re$ given by $\phi(x) = x_1^2 / x_2$. We have

$$\nabla_x \phi(x) = \begin{bmatrix} 2x_1/x_2 \\ -x_1^2/x_2^2 \end{bmatrix} \quad \text{and} \quad \nabla_x^2 \phi(x) = \begin{bmatrix} 2/x_2 & -2x_1/x_2^2 \\ -2x_1/x_2^2 & 2x_1^2/x_2^3 \end{bmatrix}.$$

If $x_2 > 0$ then the $\text{tr}(\nabla_x^2 \phi(x)) = 4x_1^2/x_2^4 > 0$ and $|\nabla_x^2 \phi(x)| = 0$. It follows that ϕ is convex. For any $\gamma \in [0, 1]$ and $\nu, \overline{\nu} \in \Re^N$ such that $\nu^\top \beta > 0$ and $\overline{\nu}^\top \beta > 0$,

$$\begin{aligned}\psi(\gamma\nu + (1-\gamma)\overline{\nu}) =& \phi\left(\left[\begin{array}{c}(\gamma\nu + (1-\gamma)\overline{\nu})^\top\alpha\\(\gamma\nu + (1-\gamma)\overline{\nu})^\top b\end{array}\right]\right)\\ =& \phi\left(\gamma\left[\begin{array}{c}\nu^\top\alpha\\\nu^\top\beta\end{array}\right] + (1-\gamma)\left[\begin{array}{c}\overline{\nu}^\top\alpha\\\overline{\nu}^\top b\end{array}\right]\right)\\ \leq& \gamma\phi\left(\left[\begin{array}{c}\nu^\top\alpha\\\nu^\top\beta\end{array}\right]\right) + (1-\gamma)\phi\left(\left[\begin{array}{c}\overline{\nu}^\top\alpha\\\overline{\nu}^\top\beta\end{array}\right]\right)\\ =& \gamma\psi(\nu) + (1-\gamma)\psi(\overline{\nu}).\end{aligned}$$

Hence, ψ is convex.

Let $\nu^* \in \arg\min_{\nu\in\Delta_N}\psi(\nu)$ and $\zeta(\nu) = (\nu^\top\alpha)^2 - \psi(\nu^*)\nu^\top\beta$. Note that, for all $\nu \in \Delta_N$,

$$\zeta(\nu) = (\nu^\top\alpha)^2 - \psi(\nu^*)\nu^\top\beta \geq (\nu^\top\alpha)^2 - \psi(\nu)\nu^\top\beta = 0,$$

and $\zeta(\nu^*) = 0$, implying $\arg\min_{\nu\in\Delta_N}\psi(\nu) \subseteq \arg\min_{\nu\in\Delta_N}\zeta(\nu)$. Let $\overline{\nu} \in \arg\min_{\nu\in\Delta_N}\zeta(\nu)$. Then, $\zeta(\overline{\nu}) = 0$ and

$$\psi(\overline{\nu}) = \frac{(\overline{\nu}^\top\alpha)^2}{\overline{\nu}^\top\beta} = \frac{\zeta(\overline{\nu}) + \psi(\nu^*)\overline{\nu}^\top\beta}{\overline{\nu}^\top\beta} = \psi^*,$$

implying $\arg\min_{\nu\in\Delta_N}\psi(\nu) \supseteq \arg\min_{\nu\in\Delta_N}\zeta(\nu)$. It follows that

$$\arg\min_{\nu\in\Delta_N}\psi(\nu) = \arg\min_{\nu\in\Delta_N}\zeta(\nu).$$

To complete the proof, we will establish that there exists $\nu^\dagger \in \arg\min_{\nu\in\Delta_N}\zeta(\nu)$ with at most two positive components. If $\overline{\nu}$ has two or fewer positive components, we are done, we will treat the case where $\overline{\nu}$ has more than two positive components. Note that $\nabla_\nu\zeta(\nu) = 2\alpha\nu^\top\alpha - \psi(\nu^*)\beta$. By the KKT conditions, $\overline{\nu} \in \arg\min_{\nu\in\Delta_N}\zeta(\nu)$ if and only if there exists a vector $d \geq 0$ and a scalar c such that

$$2\alpha\overline{\nu}^\top\alpha - \psi(\nu^*)\beta - d + c\mathbf{1} = 0 \qquad \text{and} \qquad d^\top\overline{\nu} = 0.$$

These conditions can equivalently be written as

$$\overline{\nu}^\top(2\alpha\overline{\nu}^\top\alpha - \psi(\nu^*)\beta + c\mathbf{1}) = 0.$$

Let $\overline{n} = \arg\max_{n:\overline{\nu}_n>0}\overline{\nu}_n$ and $\underline{n} = \arg\min_{n:\overline{\nu}_n>0}\overline{\nu}_n$. Let $\gamma \in [0,1]$ be such that

$$\overline{\nu}^\top\alpha = \gamma\overline{\nu}_{\overline{n}}\alpha_{\overline{n}} + (1-\gamma)\overline{\nu}_{\underline{n}}\alpha_{\underline{n}},$$

and let $\tilde{\nu} = \gamma\overline{\nu}_{\overline{n}}\mathbf{1}_{\overline{n}} + (1-\gamma)\overline{\nu}_{\underline{n}}\mathbf{1}_{\underline{n}}$. Note that $\tilde{\nu}^\top\alpha = \overline{\nu}^\top a$ and

$$\tilde{\nu}^\top(2\alpha\tilde{\nu}^\top\alpha - \psi(\nu^*)\beta + c\mathbf{1}) = 0,$$

since $\text{support}(\tilde{\nu}) \subset \text{support}(\overline{\nu})$. It follows that $\tilde{\nu} \in \arg\min_{\nu\in\Delta_N} \zeta(\nu)$. Since $\tilde{\nu}$ has two positive components, the result follows. □

This result implies that the objective minimized by each version of value-IDS is convex and that the minimum can be attained by randomizing between at most two actions. To see why, consider the objective of the basic version:

$$\min_{\nu\in\Delta_{\mathcal{A}}} \frac{\mathbb{E}\left[V_{\pi_\chi}(H_t) - Q_{\pi_\chi}(H_t, \tilde{A}_t)|X_t\right]^2}{\mathbb{I}(\chi; \tilde{A}_t, \tilde{Y}_{t+1}|X_t \leftarrow X_t)}.$$

Shortfall and information gain can be rewritten as

$$\begin{aligned}&\mathbb{E}\left[V_{\pi_\chi}(H_t) - Q_{\pi_\chi}(H_t, \tilde{A}_t)|X_t\right]\\ &= \sum_{a\in\mathcal{A}} \nu(a)\mathbb{E}\left[V_{\pi_\chi}(H_t) - Q_{\pi_\chi}(H_t, \tilde{A}_t)|X_t, \tilde{A}_t = a\right],\end{aligned}$$

and

$$\begin{aligned}\mathbb{I}(\chi; \tilde{A}_t, \tilde{Y}_{t+1}|X_t \leftarrow X_t) &\overset{(a)}{=} \mathbb{I}(\chi; \tilde{Y}_{t+1}|X_t \leftarrow X_t, \tilde{A}_t) + \mathbb{I}(\chi; \tilde{A}_t|X_t \leftarrow X_t)\\ &\overset{(b)}{=} \mathbb{I}(\chi; \tilde{Y}_{t+1}|X_t \leftarrow X_t, \tilde{A}_t)\\ &= \sum_{a\in\mathcal{A}} \nu(a)\mathbb{I}(\chi; \tilde{Y}_{t+1}|X_t \leftarrow X_t, \tilde{A}_t = a),\end{aligned}$$

where (a) follows from the chain rule of mutual information and (b) follows from the fact that $\chi \perp \tilde{A}_t|X_t$. Without loss of generality, let $\mathcal{A} = \{1, \ldots, |\mathcal{A}|\}$. Then, letting

$$\alpha_a = \mathbb{E}\left[V_{\pi_\chi}(H_t) - Q_{\pi_\chi}(H_t, \tilde{A}_t)|X_t, \tilde{A}_t = a\right],$$

and

$$\beta_a = \mathbb{I}(\chi; \tilde{Y}_{t+1}|X_t \leftarrow X_t, \tilde{A}_t = a),$$

Theorem C.1 confirms our assertion about convexity of the value-IDS objective and existence of a 2-sparse optimal solution.

D

Relation Between Information Gain and Variance

All the mutual information terms are in nats in this section.

Lemma D.1. Let $X_t = (Z_t, S_t, P_t)$ be the agent state at timestep t, $\tilde{A}_t$ be a random action sampled from some distribution ν that only depends on X_t, $Q_\dagger$ be a vector of GVFs with dimension n, $\tilde{Y}_{t+1} = Q_\dagger(H_t, \tilde{A}_t) + W_{t+1}$ be a pseudo-observation of $Q_\dagger$ where W_{t+1} is some zero-mean random noise, and π_χ be a target policy. If each component of $Q_\dagger$ has a span of at most M_1 and each component of W_{t+1} has a span of at most M_2,

$$\begin{aligned}&\mathbb{I}(\pi_\chi(\cdot|S_t); \tilde{A}_t, \tilde{Y}_{t+1}|X_t \leftarrow X_t)\\ &\geq \frac{2}{n(M_1+M_2)^2}\mathbb{E}\left[\mathrm{tr}\left(\mathrm{Cov}\left[\mathbb{E}\left[Q_\dagger(H_t, \tilde{A}_t)|X_t, \tilde{A}_t, \pi_\chi(\cdot|S_t)\right]\middle|X_t, \tilde{A}_t\right]\right)|X_t\right].\end{aligned}$$

Proof. We have

$$\begin{aligned}&\mathbb{I}(\pi_\chi(\cdot|S_t); \tilde{A}_t, \tilde{Y}_{t+1}|X_t \leftarrow X_t)\\ &= \mathbb{I}(\pi_\chi(\cdot|S_t); \tilde{Y}_{t+1}|X_t \leftarrow X_t, \tilde{A}_t) + \mathbb{I}(\pi_\chi(\cdot|S_t); \tilde{A}_t, |X_t \leftarrow X_t)\\ &\overset{(a)}{=} \mathbb{I}(\pi_\chi(\cdot|S_t); \tilde{Y}_{t+1}|X_t \leftarrow X_t, \tilde{A}_t)\\ &= \sum_{\tilde{a}\in\mathcal{A}} \nu(\tilde{a})\mathbb{I}(\pi_\chi(\cdot|S_t); \tilde{Y}_{t+1}|X_t \leftarrow X_t, \tilde{A}_t = \tilde{a})\end{aligned}$$

where (a) follows from $\pi_\chi(\cdot|S_t) \perp \tilde{A}_t|X_t$. Then,

$$
\begin{aligned}
&n\mathbb{I}(\pi_\chi(\cdot|S_t); \tilde{Y}_{t+1}|X_t \leftarrow X_t, \tilde{A}_t = a) \\
&\overset{(a)}{\geq} \sum_{i=1}^{n} \mathbb{I}(\pi_\chi(\cdot|S_t); \tilde{Y}_{t+1,i}|X_t, \tilde{A}_t = a) \\
&= \sum_{i=1}^{n} \mathbb{E}\Bigg[\mathbf{d}_{\mathrm{KL}}\Bigg(\mathbb{P}(\tilde{Y}_{t+1,i} \in \cdot|X_t, \pi_\chi(\cdot|S_t), \tilde{A}_t = a) \\
&\qquad\qquad \| \, \mathbb{P}(\tilde{Y}_{t+1,i} \in \cdot|X_t, \tilde{A}_t = a)\Bigg)|X_t, \tilde{A}_t = a\Bigg] \\
&\overset{(b)}{\geq} \frac{2}{(M_1+M_2)^2} \sum_{i=1}^{n} \mathbb{E}\Bigg[\Bigg(\mathbb{E}\left[\tilde{Y}_{t+1,i}|X_t, \pi_\chi(\cdot|S_t), \tilde{A}_t = a\right] \\
&\qquad\qquad - \mathbb{E}\left[\tilde{Y}_{t+1,i}|X_t, \tilde{A}_t = a\right]\Bigg)^2 |X_t, \tilde{A}_t = a\Bigg] \\
&= \frac{2}{(M_1+M_2)^2} \sum_{i=1}^{n} \mathrm{Var}\left[\mathbb{E}\left[\tilde{Y}_{t+1,i}|X_t, \tilde{A}_t = a, \pi_\chi(\cdot|S_t)\right]|X_t, \tilde{A}_t = a\right],
\end{aligned}
$$

where (a) follows from chain rule and mutual information being non-negative, and (b) follows from Pinsker's inequality with span of each component of $\tilde{Y}_{t+1}$ being at most $M_1 + M_2$. Since each component of $Q_\dagger(\cdot,\cdot)$ has a span of atmost M_1 and each component of W_{t+1} has a span of at most M_2, each component of $\tilde{Y}_{t+1} = Q_\dagger(H_t, \tilde{A}_t) + W_{t+1}$ has a span of at most $M_1 + M_2$.

$$
\begin{aligned}
&\mathbb{I}(\pi_\chi(\cdot|S_t); \tilde{A}_t, \tilde{Y}_{t+1}|X_t \leftarrow X_t) \\
&= \sum_{\tilde{a}\in\mathcal{A}} \nu(\tilde{a}) \mathbb{I}(\pi_\chi(\cdot|S_t); \tilde{Y}_{t+1}|X_t \leftarrow X_t, \tilde{A}_t = \tilde{a}) \\
&\geq \frac{2}{n(M_1+M_2)^2} \sum_{\tilde{a}\in\mathcal{A}} \nu(\tilde{a}) \sum_{i=1}^{n} \mathrm{Var}\left[\mathbb{E}\left[\tilde{Y}_{t+1,i}|X_t, \tilde{A}_t = \tilde{a}, \pi_\chi(\cdot|S_t)\right]|X_t, \tilde{A}_t = \tilde{a}\right] \\
&= \frac{2}{n(M_1+M_2)^2} \sum_{i=1}^{n} \mathbb{E}\left[\mathrm{Var}\left[\mathbb{E}\left[\tilde{Y}_{t+1,i}|X_t, \tilde{A}_t, \pi_\chi(\cdot|S_t)\right]|X_t, \tilde{A}_t\right]\right] \\
&= \frac{2}{n(M_1+M_2)^2} \sum_{i=1}^{n} \mathbb{E}\left[\mathrm{Var}\left[\mathbb{E}\left[Q_\dagger(H_t, \tilde{A}_t)_i|X_t, \tilde{A}_t, \pi_\chi(\cdot|S_t)\right]|X_t, \tilde{A}_t\right]\right] \\
&= \frac{2}{n(M_1+M_2)^2} \mathbb{E}\left[\mathrm{tr}\left(\mathrm{Cov}\left[\mathbb{E}\left[Q_\dagger(H_t, \tilde{A}_t)|X_t, \tilde{A}_t, \pi_\chi(\cdot|S_t)\right]|X_t, \tilde{A}_t\right]\right)\Big|X_t\right].
\end{aligned}
$$

□

Lemma D.2. Let $X_t = (Z_t, S_t, P_t)$ be the agent state at timestep t, $\tilde{A}_t$ be a random action sampled from some distribution ν over actions, $Q_\dagger$ be a vector of GVFs with dimension n, and $\tilde{Y}_{t+1} = Q_\dagger(H_t, \tilde{A}_t) + W_{t+1}$ be a pseudo-observation of $Q_\dagger$ where W_{t+1} is some zero-mean random noise. If each component of $Q_\dagger$ has a span of at most M_1 and each component of W_{t+1} has a span of at most M_2,

$$\mathbb{I}(Q_\dagger(H_t, \tilde{A}_t); \tilde{A}_t, \tilde{Y}_{t+1} | X_t \leftarrow X_t) \geq \frac{2}{n(M_1+M_2)^2} \mathbb{E}\left[\operatorname{tr}\left(\operatorname{Cov}\left[Q_\dagger(H_t, \tilde{A}_t) | X_t, \tilde{A}_t\right]\right) | X_t\right].$$

Proof. We have

$$\begin{aligned}
&\mathbb{I}(Q_\dagger(H_t, \tilde{A}_t); \tilde{A}_t, \tilde{Y}_{t+1} | X_t \leftarrow X_t) \\
&= \mathbb{I}(Q_\dagger(H_t, \tilde{A}_t); \tilde{A}_t | X_t \leftarrow X_t) + \mathbb{I}(Q_\dagger(H_t, \tilde{A}_t); \tilde{Y}_{t+1} | X_t \leftarrow X_t, \tilde{A}_t) \\
&\overset{(a)}{\geq} \mathbb{I}(Q_\dagger(H_t, \tilde{A}_t); \tilde{Y}_{t+1} | X_t \leftarrow X_t, \tilde{A}_t) \\
&= \sum_{a \in \mathcal{A}} \nu(a) (\mathbb{I}(Q_\dagger(H_t, \tilde{A}_t); \tilde{Y}_{t+1} | X_t \leftarrow X_t, \tilde{A}_t = a),
\end{aligned}$$

where (a) follows from mutual information being non-negative. Then,

$$\begin{aligned}
&n\mathbb{I}(Q_\dagger(H_t, \tilde{A}_t); \tilde{Y}_{t+1} | X_t \leftarrow X_t, \tilde{A}_t = a) \\
&\overset{(a)}{\geq} \sum_{i=1}^{n} \mathbb{I}(Q_\dagger(H_t, \tilde{A}_t)_i; \tilde{Y}_{t+1,i} | X_t \leftarrow X_t, \tilde{A}_t = a) \\
&= \sum_{i=1}^{n} \mathbb{E}\Big[\mathbf{d}_{\mathrm{KL}}\Big(\mathbb{P}(\tilde{Y}_{t+1,i} | Q_\dagger(H_t, \tilde{A}_t)_i, X_t, \tilde{A}_t = a) \\
&\qquad\qquad \| \mathbb{P}(\tilde{Y}_{t+1,i} | X_t, \tilde{A}_t = a)\Big) | X_t, \tilde{A}_t = a\Big] \\
&\overset{(b)}{\geq} \frac{2}{(M_1+M_2)^2} \sum_{i=1}^{n} \mathbb{E}\Big[\Big(\mathbb{E}\left[\tilde{Y}_{t+1,i} | Q_\dagger(H_t, \tilde{A}_t)_i, X_t, \tilde{A}_t = a\right] \\
&\qquad\qquad - \mathbb{E}\left[\tilde{Y}_{t+1,i} | X_t, \tilde{A}_t = a\right]\Big)^2 | X_t, \tilde{A}_t = a\Big] \\
&= \frac{2}{(M_1+M_2)^2} \sum_{i=1}^{n} \operatorname{Var}\left[Q_\dagger(H_t, \tilde{A}_t)_i | X_t, \tilde{A}_t = a\right],
\end{aligned}$$

where (a) follows from the fact that mutual information between two random vectors is not less than mutual information between any of there

components, and (b) follows from Pinsker's inequality with span of each component of $\tilde{Y}_{t+1}$ being at most $M_1 + M_2$. Since each component of $Q_\dagger(\cdot,\cdot)$ has a span of atmost M_1 and each component of W_{t+1} has a span of at most M_2, each component of $\tilde{Y}_{t+1} = Q_\dagger(H_t, \tilde{A}_t) + W_{t+1}$ has a span of at most $M_1 + M_2$. Therefore,

$$\begin{aligned}
&\mathbb{I}(Q_\dagger(H_t, \tilde{A}_t); \tilde{A}_t, \tilde{Y}_{t+1} | X_t \leftarrow X_t) \\
&\geq \frac{2}{n(M_1+M_2)^2} \sum_{a\in\mathcal{A}} \nu(a) \sum_{i=1}^{n} \mathrm{Var}\left[Q_\dagger(H_t, \tilde{A}_t)_i | X_t, \tilde{A}_t = a\right] \\
&= \frac{2}{n(M_1+M_2)^2} \sum_{i=1}^{n} \mathbb{E}\left[\mathrm{Var}\left[Q_\dagger(H_t, \tilde{A}_t)_i | X_t, \tilde{A}_t\right] | X_t\right] \\
&= \frac{2}{n(M_1+M_2)^2} \mathbb{E}\left[\mathrm{tr}\left(\mathrm{Cov}\left[Q_\dagger(H_t, \tilde{A}_t) | X_t, \tilde{A}_t\right]\right) | X_t\right].
\end{aligned}$$

□

E

Implementation and Computation

This section provides the details and parameters for our computational experiments. We specify the agent through the agent state, and the action selection policy $\pi(\cdot|X_t)$. For the most part, these are both explained in the main body of our monograph. However, our next subsections expand on these implementational details.

E.1 Agent state update

Agent state dynamics are determined by the update rules (f_{algo}, f_{alea}, f_{epis}) introduced in Equations (3.1), (3.2), and (3.3). For the most part, these updates are already described in Section 7.2.1, but we use this appendix to spell out some more of the details, particularly in regards to the epistemic update.

Algorithmic state The algorithmic state is null $Z_t = \emptyset$ for both IDS and ϵ-greedy action selection. For Thompson sampling the algorithmic state $Z_t = Z_{t_k}$ is resampled uniformly from the set of epistemic indices for the relevant ENN at the end of each episode. This fully specifies the algorithmic update for all our experiments, and so we will not address this further.

Aleatoric state All of the agents considered are 'feed-forward' variants of DQN with aleatoric state given by the current observation $S_t = O_t$. This fully specifies the aleatoric update for all our experiments, and so we will not address this further.

Epistemic state All agents' epistemic state are given $P_t = (\theta_t, B_t)$. Here θ_t are parameters of an ENN f and B_t is a FIFO experience replay buffer. Further, all of our experiments are specialized to the case where f represents a (potentially general) value function over $\mathcal{A}$ finite actions. In all of our experiments we learn via minibatch SGD, according to Equation (7.2).

For each experiment, we can therefore fully specify the epistemic update through the ENN f, the loss function $\ell(\theta; f, \theta^-, z, (s, a, r, s')) \to \mathbb{R}$, the initial parameters θ_0, the SGD update procedure, n_{batch}, n_{index}. For the replay buffer in each experiment we set a minimum replay size equal to n_{batch} and a maximum replay size of 10,000.

E.2 Action selection

Action selection via ϵ-greedy (Mnih *et al.*, 2013) and Thompson sampling (Osband *et al.*, 2019) are relatively straightforward, and have been covered extensively in prior work. In this subsection we expand on the sample-based implementation of IDS that approximates Equations (6.6) and (6.7). Note that, since we know the solution has support on at most two actions in $\mathcal{A}$, we can approximately optimize the objective by effectively searching over a probability grid for each *pair* of actions in $\mathcal{A}$. In all of our experiments we search with granularity of $\frac{1}{100}$ in each action probability.

In all of our experiments, we use a simple sample-based approach to approximating the shortfall and variance in Equations (6.6) and (6.7). The first step is to generate n_{IDS} samples from the ENN given the agents epistemic state P_t. The expected shortfall is then calculated simply as the average of the shortfall for each action, for each sample. The variance-based information gain is approximated through the sample variance, as detailed below.

(a) **Learning Target = Optimal Action.** We use the same n_{IDS} samples from the ENN to approximate the information gain in Equation (6.6). In our experiments we only perform experiments with action-value learning targets, and so we can simplify the exposition significantly. Let $Q_1, .., Q_{n_{\text{IDS}}} \in \mathbb{R}^{\mathcal{A}}$ be samples of the action value generated by the agent, $\mathcal{Q}_a := \{Q_n \mid a \in \arg\max_\alpha Q_n(S_t, \alpha)\}$, $\overline{Q}_a := \frac{1}{|\mathcal{Q}_a|} \sum_{q \in \mathcal{Q}_a} q$ and $\overline{Q} = \frac{1}{n_{\text{IDS}}} \sum_{n=1}^{n_{\text{IDS}}} Q_n$. Then the approximate information gain used in our experiments can be written:

$$\mathbb{E}\left[\text{tr}\left(\text{Cov}\left[\mathbb{E}[Q_\dagger(H_t, \tilde{A}_t)|X_t, \tilde{A}_t, \pi_\chi(\cdot|S_t)])|X_t, \tilde{A}_t\right]\right)|X_t\right] \simeq \frac{1}{n_{\text{IDS}}} \sum_{a=1}^{n_{\mathcal{A}}} |\mathcal{Q}_a| \left(\overline{Q}_a - \overline{Q}\right)^2. \quad \text{(E.1)}$$

(b) **Learning Target = GVF.** In a similar manner, we reuse the same n_{IDS} samples from the ENN to approximate the information gain in Equation (6.7). For a problem with general value function $Q_\dagger \in \mathbb{R}^d$, let $Q_1, .., Q_{n_{\text{IDS}}} \in \mathbb{R}^{\mathcal{A}}$ be samples of the action value generated by the agent and $\overline{Q}$ the sample mean. The approximate information gain used in our experiments is then:

$$\mathbb{E}\left[\text{tr}\left(\text{Cov}\left[Q_\dagger(H_t, \tilde{A}_t)|X_t, \tilde{A}_t\right]\right)|X_t\right] \simeq \frac{1}{n_{\text{IDS}}} \sum_{i=1}^{d} \sum_{j=1}^{n_{\text{IDS}}} \left(Q_{i,j} - \overline{Q}_i\right)^2. \quad \text{(E.2)}$$

E.3 Parameters for each experiment

In this section we list the settings used to generate the results in Section 7. It is our intention to provide some elements of our agents and code for opensource.

7.2.1, 7.3.1, and `bsuite` IDS implementation We use the exact same settings for the experiments in these sections:

- ENN = ensemble of 20 50-50-MLPs with matched prior functions initialized according to JAX standards (Osband *et al.*, 2018).

- Loss $\ell = \ell^{Q,\gamma}$ (Equation (7.1)) with $\gamma = 0.99$.
- Action selection via Equation (E.1) with $n_{\text{IDS}} = 40$.
- SGD update according with ADAM with learning rate $\alpha = 0.001$.
- $n_{\text{batch}} = 128$ for RL tasks and $n_{\text{batch}} = 1$ for bandit task.

7.3.2 Sparse bandit Same settings as 7.2.1 except for the ENN and loss function designed to encode the prior knowledge:

- ENN = ensemble of 20 logits over N possible rewarding arms.
- Loss ℓ of the cross-entropy on the posterior probabilty of the observation given the rewarding arm.
- Action selection via Equation (E.1) with $n_{\text{IDS}} = 40$, given the knowledge of how to convert logits to associated action values.
- Vanilla SGD update with learning rate $\alpha = 0.1$.

7.4 Variance-IDS with general value functions These settings are almost identical to 7.3.2, but using a different action selection and with specialized ENNs:

- ENN = ensemble of 20 logits over N possible rewarding arms (rewarding states in the chain problem).
- Loss ℓ of the cross-entropy on the posterior probabilty of the observation given the ENN logits.
- Action selection via Equation (E.2) with $n_{\text{IDS}} = 40$, given the knowledge of how to convert logits to associated action values.
- Vanilla SGD update with learning rate $\alpha = 0.1$.

E.4 `bsuite` Report

The *Behaviour Suite for Core Reinforcement Learning*, or `bsuite` for short, is a collection of carefully-designed experiments that investigate core capabilities of a reinforcement learning (RL) agent. The aim of the `bsuite` project is to collect clear, informative and scalable problems that capture key issues in the design of efficient and general learning algorithms and study agent behaviour through their performance on these shared benchmarks. This report provides a snapshot of agent performance on `bsuite2019`, obtained by running the experiments from `github.com/deepmind/bsuite` (Osband *et al.*, 2020).

E.4.1 Agent Definition

We compare the performance of three agents as outlined in Section 7.2.1. In each case, the agents learn with an ensemble ENN formed of 20 50-50-MLPs and identical learning rules. The only difference is in the action selection 'planner':

- **egreedy**: ϵ=5% greedy action selection, essentially DQN (Mnih *et al.*, 2013).
- **TS**: Thompson Sampling, aka *bootstrapped DQN* (Osband *et al.*, 2016).
- **IDS**: Information Directed Sampling, with 40 samples per step.

E.4.2 Summary Scores

Each `bsuite` experiment outputs a summary score in $[0, 1]$. We aggregate these scores according to the `bsuite` analysis notebook available at https://github.com/deepmind/bsuite/blob/main/bsuite/analysis/results.ipynb.

E.4.3 Results Commentary

Figures E.1 and E.2 present the aggregate and individual performances of ϵ-greedy, TS, and IDS agents on `bsuite`. These results show a strong signal that action selection via IDS and TS can greatly outperform that of ϵ-greedy in domains where exploration is crucial. At least in the current collection of `bsuite` tasks, IDS and TS perform similarly

overall. We also see some evidence that ϵ-greedy action selection is more robust to changes in scale, but less robust to noise.

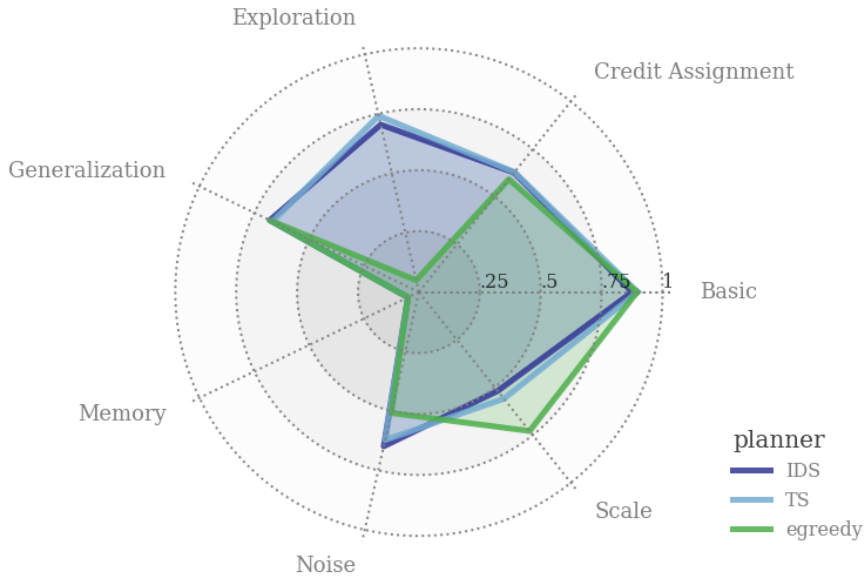

Figure E.1: Snapshot of agent behaviour.

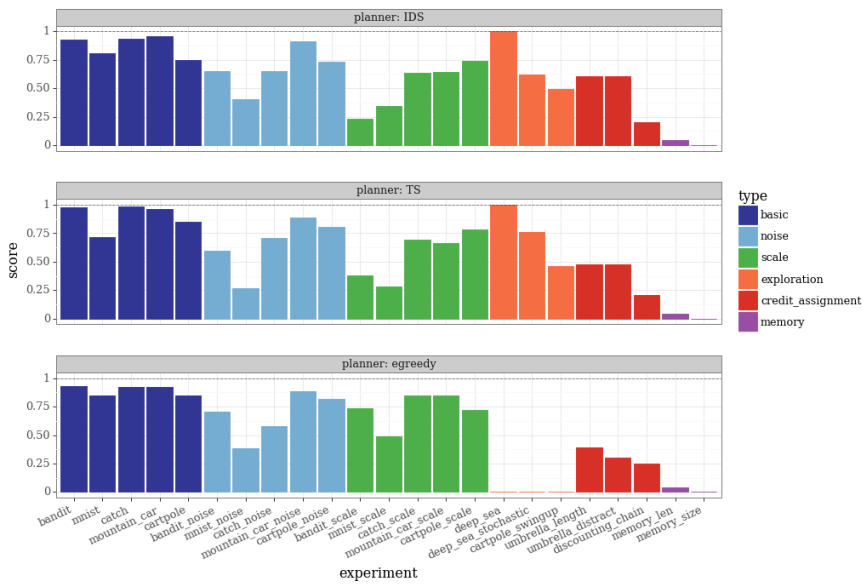

Figure E.2: Score for each `bsuite` experiment.

References

Agrawal, S. and R. Jia. (2017). "Optimistic posterior sampling for reinforcement learning: worst-case regret bounds". In: *Advances in Neural Information Processing Systems*. Ed. by I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett. Vol. 30. Curran Associates, Inc. 1184–1194.

Anscombe, F. J. and R. J. Aumann. (1963). "A definition of subjective probability". *Annals of mathematical statistics*. 34(1): 199–205.

Arumugam, D. and B. Van Roy. (2021). "Deciding What to Learn: A Rate-Distortion Approach". arXiv: 2101.06197 [cs.LG].

Auer, P., N. Cesa-Bianchi, and P. Fischer. (2002). "Finite-time analysis of the multiarmed bandit problem". *Machine learning*. 47(2): 235–256.

Azar, M. G., I. Osband, and R. Munos. (2017). "Minimax Regret Bounds for Reinforcement Learning". In: *Proceedings of the 34th International Conference on Machine Learning*. Ed. by D. Precup and Y. W. Teh. Vol. 70. *Proceedings of Machine Learning Research*. PMLR. 263–272.

Barto, A. G., R. S. Sutton, and C. W. Anderson. (1983). "Neuronlike adaptive elements that can solve difficult learning control problems". *IEEE transactions on systems, man, and cybernetics*. SMC-13(5): 834–846.

Bertsekas, D. P. (2019). *Reinforcement learning and optimal control*. Athena Scientific Belmont, MA.

Bertsekas, D. P. and J. N. Tsitsiklis. (1996). *Neuro-dynamic programming*. Athena Scientific.

Brafman, R. I. and M. Tennenholtz. (2003). "R-Max - a General Polynomial Time Algorithm for near-Optimal Reinforcement Learning". *Journal of Machine Learning Research*. 3: 213–231.

Bubeck, S. and N. Cesa-Bianchi. (2012). "Regret analysis of stochastic and nonstochastic multi-armed bandit problems". *arXiv preprint arXiv:1204.5721*.

Bubeck, S., O. Dekel, T. Koren, and Y. Peres. (2015). "Bandit Convex Optimization:$\sqrt{T}$ Regret in One Dimension". In: *Conference on Learning Theory*. PMLR. 266–278.

Bubeck, S. and R. Eldan. (2016). "Multi-scale exploration of convex functions and bandit convex optimization". In: *Conference on Learning Theory*. PMLR. 583–589.

Bubeck, S. and M. Sellke. (2020). "First-Order Bayesian Regret Analysis of Thompson Sampling". In: *Algorithmic Learning Theory*. PMLR. 196–233.

Burda, Y., H. Edwards, D. Pathak, A. Storkey, T. Darrell, and A. A. Efros. (2019). "Large-Scale Study of Curiosity-Driven Learning". In: *ICLR*.

Chang, H. S., M. C. Fu, J. Hu, and S. I. Marcus. (2005). "An adaptive sampling algorithm for solving Markov decision processes". *Operations Research*. 53(1): 126–139.

Coulom, R. (2006). "Efficient selectivity and backup operators in Monte-Carlo tree search". In: *International conference on computers and games*. Springer. 72–83.

Cover, T. M. and J. A. Thomas. (2006). *Elements of Information Theory*. John Wiley and Sons.

Daswani, M., P. Sunehag, and M. Hutter. (2013). "Q-learning for history-based reinforcement learning". In: *Asian Conference on Machine Learning*. PMLR. 213–228.

Daswani, M., P. Sunehag, M. Hutter, *et al.* (2014). "Feature reinforcement learning: state of the art". In: *Sequential decision-making with big data: papers from the AAAI-14 workshop*. Association for the Advancement of Artificial Intelligence.

Devraj, A. M., K. Xu, and B. Van Roy. (2021). "A Bit Better? Quantifying Information for Bandit Learning". arXiv: arXiv:2102.09488.

Dong, S., T. Ma, and B. Van Roy. (2019). "On the performance of Thompson sampling on logistic bandits". In: *Conference on Learning Theory*. PMLR. 1158–1160.

Dong, S. and B. Van Roy. (2018). "An Information-Theoretic Analysis for Thompson Sampling with Many Actions". In: *Advances in Neural Information Processing Systems*. Ed. by S. Bengio, H. Wallach, H. Larochelle, K. Grauman, N. Cesa-Bianchi, and R. Garnett. Vol. 31. Curran Associates, Inc. 4157–4165.

Dragomir, S., M. Scholz, and J. Sunde. (2000). "Some upper bounds for relative entropy and applications". *Computers & Mathematics with Applications*. 39(9-10): 91–100.

Duff, M. O. (2003). "Optimal learning: Computational procedures for Bayes-adaptive Markov decision processes". *PhD thesis*. University of Massachusetts, Amherst.

Dwaracherla, V., X. Lu, M. Ibrahimi, I. Osband, Z. Wen, and B. Van Roy. (2020). "Hypermodels for Exploration". In: *International Conference on Learning Representations*.

Dwaracherla, V. and B. Van Roy. (2021). "Langevin DQN". arXiv: arXiv:2002.07282.

Elder, S. (2016). "Bayesian adaptive data analysis guarantees from subgaussianity".

Engel, Y., S. Mannor, and R. Meir. (2003). "Bayes meets Bellman: The Gaussian process approach to temporal difference learning". In: *Proceedings of the 20th International Conference on Machine Learning (ICML-03)*. 154–161.

Engel, Y., S. Mannor, and R. Meir. (2005). "Reinforcement learning with Gaussian processes". In: *Proceedings of the 22nd international conference on Machine learning*. 201–208.

Flennerhag, S., J. X. Wang, P. Sprechmann, F. Visin, A. Galashov, S. Kapturowski, D. L. Borsa, N. Heess, A. Barreto, and R. Pascanu. (2020). "Temporal Difference Uncertainties as a Signal for Exploration". *arXiv preprint arXiv:2010.02255*.

Frankel, A. and E. Kamenica. (2019). "Quantifying information and uncertainty". *American Economic Review*. 109(10): 3650–80.

Ghavamzadeh, M., S. Mannor, and J. Pineau. (2015). "Bayesian Reinforcement Learning: A Survey". *Foundations and Trends® in Machine Learning*. 8(5-6): 359–483. URL: http://dx.doi.org/10.1561/2200000049.

Gittins, J. (1974). "A dynamic allocation index for the sequential design of experiments". In: North Holland. 241–266.

Gittins, J. C. and D. M. Jones. (1979). "A dynamic allocation index for the discounted multiarmed bandit problem". *Biometrika*. 66(3): 561–565.

Grimm, C., A. Barreto, S. Singh, and D. Silver. (2021). "The Value Equivalence Principle for Model-Based Reinforcement Learning". In: *Advances in Neural Information Processing Systems*. Vol. 33. MIT Press.

Howard, R. A. (1966). "Information value theory". *IEEE Transactions on systems science and cybernetics*. 2(1): 22–26.

Hutter, M. (2007). "Universal Algorithmic Intelligence: A Mathematical Top→Down Approach". In: *Artificial General Intelligence*. Ed. by B. Goertzel and C. Pennachin. *Cognitive Technologies*. Berlin: Springer. 227–290. URL: http://www.hutter1.net/ai/aixigentle.htm.

Jaksch, T., R. Ortner, and P. Auer. (2010). "Near-optimal Regret Bounds for Reinforcement Learning". *Journal of Machine Learning Research*. 11(51): 1563–1600.

Jha, A. (2016). "Without Claude Shannon's information theory there would have been no internet". URL: https://www.theguardian.com/science/2014/jun/22/shannon-information-theory.

Jiang, N., A. Krishnamurthy, A. Agarwal, J. Langford, and R. E. Schapire. (2017). "Contextual Decision Processes with low Bellman rank are PAC-Learnable". In: *Proceedings of the 34th International Conference on Machine Learning*. Ed. by D. Precup and Y. W. Teh. Vol. 70. *Proceedings of Machine Learning Research*. 1704–1713.

Jin, C., Z. Allen-Zhu, S. Bubeck, and M. I. Jordan. (2018). "Is Q-Learning Provably Efficient?" In: *Advances in Neural Information Processing Systems*. Ed. by S. Bengio, H. Wallach, H. Larochelle, K. Grauman, N. Cesa-Bianchi, and R. Garnett. Vol. 31. Curran Associates, Inc. 4863–4873.

Jin, C., Z. Yang, Z. Wang, and M. I. Jordan. (2020). "Provably efficient reinforcement learning with linear function approximation". In: *Proceedings of Thirty Third Conference on Learning Theory*. Ed. by J. Abernethy and S. Agarwal. Vol. 125. *Proceedings of Machine Learning Research*. PMLR. 2137–2143.

Kakade, S. and J. Langford. (2002). "Approximately Optimal Approximate Reinforcement Learning". In: *Proceedings of the Nineteenth International Conference on Machine Learning (ICML 2002)*. Ed. by C. Sammut and A. Hoffman. San Francisco, CA, USA: Morgan Kauffman. 267–274.

Kearns, M. and S. Singh. (2002). "Near-Optimal Reinforcement Learning in Polynomial Time". *Machine Learning*. 49: 209–232.

Kingma, D. and J. Ba. (2015). "Adam: A Method for Stochastic Optimization". *Proceedings of the International Conference on Learning Representations*.

Kirschner, J. (2021). "Information-Directed Sampling – Frequentist Analysis and Applications". *PhD thesis*. ETH Zurich.

Kirschner, J., T. Lattimore, and A. Krause. (2020a). "Information directed sampling for linear partial monitoring". arXiv: 2002.11182.

Kirschner, J., T. Lattimore, C. Vernade, and C. Szepesvári. (2020b). "Asymptotically Optimal Information-Directed Sampling". arXiv: 2011.05944.

Klopf, A. H. (1982). *The Hedonistic Neuron: A Theory of Memory, Learning, and Intelligence*. Hemisphere Publishing Corporation.

Kocsis, L. and C. Szepesvári. (2006). "Bandit based Monte-Carlo planning". In: *European conference on machine learning*. Springer. 282–293.

Lai, T. L. (1987). "Adaptive treatment allocation and the multi-armed bandit problem". *The Annals of Statistics*: 1091–1114.

Lai, T. L. and H. Robbins. (1985). "Asymptotically efficient adaptive allocation rules". *Advances in applied mathematics*. 6(1): 4–22.

Lattimore, T. and A. György. (2020). "Mirror Descent and the Information Ratio". arXiv: 2009.12228.

Lattimore, T. and C. Szepesvári. (2019). "An information-theoretic approach to minimax regret in partial monitoring". In: *Conference on Learning Theory*. PMLR. 2111–2139.

Littman, M., R. S. Sutton, and S. Singh. (2002). "Predictive Representations of State". In: *Advances in Neural Information Processing Systems*. Ed. by T. Dietterich, S. Becker, and Z. Ghahramani. Vol. 14. MIT Press. 1555–1561.

Liu, F., S. Buccapatnam, and N. Shroff. (2018). "Information directed sampling for stochastic bandits with graph feedback". In: *Proceedings of the AAAI Conference on Artificial Intelligence*. Vol. 32. No. 1.

Lu, X. (2020). "Information-directed sampling for reinforcement learning". *PhD thesis*. Stanford University.

Lu, X. and B. Van Roy. (2019). "Information-Theoretic Confidence Bounds for Reinforcement Learning". In: *Advances in Neural Information Processing Systems*. Ed. by H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, and R. Garnett. Vol. 32. Curran Associates, Inc. 2461–2470.

McCallum, R. A. (1995). "Instance-based utile distinctions for reinforcement learning with hidden state". In: *Machine Learning Proceedings 1995*. Elsevier. 387–395.

Mnih, V., K. Kavukcuoglu, D. Silver, A. Graves, I. Antonoglou, D. Wierstra, and M. Riedmiller. (2013). "Playing Atari With Deep Reinforcement Learning". In: *NIPS Deep Learning Workshop*. URL: http://arxiv.org/abs/1312.5602.

Mnih, V., K. Kavukcuoglu, D. Silver, A. A. Rusu, J. Veness, M. G. Bellemare, A. Graves, M. Riedmiller, A. K. Fidjeland, G. Ostrovski, *et al.* (2015). "Human-level control through deep reinforcement learning". *Nature*. 518(7540): 529–533.

Nikolov, N., J. Kirschner, F. Berkenkamp, and A. Krause. (2019). "Information-Directed Exploration for Deep Reinforcement Learning". In: *Proceedings of the International Conference on Learning Representations (ICLR)*. arXiv: 1812.07544 [cs.LG].

O'Donoghue, B. (2018). "Variational Bayesian reinforcement learning with regret bounds". *arXiv preprint arXiv:1807.09647*.

O'Donoghue, B., I. Osband, R. Munos, and V. Mnih. (2018). "The uncertainty Bellman equation and exploration". In: *International Conference on Machine Learning*. 3836–3845.

Osband, I., J. Aslanides, and A. Cassirer. (2018). "Randomized prior functions for deep reinforcement learning". In: *Advances in Neural Information Processing Systems*. 8617–8629.

Osband, I., C. Blundell, A. Pritzel, and B. Van Roy. (2016). "Deep exploration via bootstrapped DQN". In: *Advances In Neural Information Processing Systems 29*. 4026–4034.

Osband, I., Y. Doron, M. Hessel, J. Aslanides, E. Sezener, A. Saraiva, K. McKinney, T. Lattimore, C. Szepesvári, S. Singh, B. Van Roy, R. Sutton, D. Silver, and H. van Hasselt. (2020). "Behaviour Suite for Reinforcement Learning". In: *International Conference on Learning Representations*.

Osband, I., D. Russo, and B. Van Roy. (2013). "(More) Efficient Reinforcement Learning via Posterior Sampling". In: *Advances in Neural Information Processing Systems*. Ed. by C. J. C. Burges, L. Bottou, M. Welling, Z. Ghahramani, and K. Q. Weinberger. Vol. 26. Curran Associates, Inc. 3003–3011.

Osband, I. and B. Van Roy. (2017). "Why is Posterior Sampling Better than Optimism for Reinforcement Learning?" In: *Proceedings of the 34th International Conference on Machine Learning*. Ed. by D. Precup and Y. W. Teh. Vol. 70. *Proceedings of Machine Learning Research*. International Convention Centre, Sydney, Australia: PMLR. 2701–2710.

Osband, I., B. Van Roy, D. J. Russo, and Z. Wen. (2019). "Deep Exploration via Randomized Value Functions". *Journal of Machine Learning Research*. 20(124): 1–62.

Oswald, J. V., S. Kobayashi, J. Sacramento, A. Meulemans, C. Henning, and B. F. Grewe. (2021). "Neural networks with late-phase weights". In: *International Conference on Learning Representations*.

Ouyang, Y., M. Gagrani, A. Nayyar, and R. Jain. (2017). "Learning Unknown Markov Decision Processes: A Thompson Sampling Approach". In: *Advances in Neural Information Processing Systems*. Ed. by I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett. Vol. 30. Curran Associates, Inc. 1333–1342.

Powell, W. B. and I. O. Ryzhov. (2012). *Optimal learning*. John Wiley & Sons.

Qin, C. (2023). "A Note on Reinforcement Learning, Bit by Bit".

Russo, D. and B. Van Roy. (2014a). "Learning to optimize via information-directed sampling". *Advances in Neural Information Processing Systems*. 27: 1583–1591.

Russo, D. and B. Van Roy. (2014b). "Learning to optimize via posterior sampling". *Mathematics of Operations Research*. 39(4): 1221–1243.

Russo, D. and B. Van Roy. (2016). "An information-theoretic analysis of Thompson sampling". *The Journal of Machine Learning Research*. 17(1): 2442–2471.

Russo, D. and B. Van Roy. (2018). "Learning to optimize via information-directed sampling". *Operations Research*. 66(1): 230–252.

Russo, D. and B. Van Roy. (2020). "Satisficing in Time-Sensitive Bandit Learning". arXiv: 1803.02855 [cs.LG].

Russo, D. J., B. Van Roy, A. Kazerouni, I. Osband, and Z. Wen. (2018). "A Tutorial on Thompson Sampling". *Foundations and Trends® in Machine Learning*. 11(1): 1–96. URL: http://dx.doi.org/10.1561/2200000070.

Ryzhov, I. O., W. B. Powell, and P. I. Frazier. (2012). "The Knowledge Gradient Algorithm for a General Class of Online Learning Problems". *Operations Research*. 60(1): 180–0195.

Schaul, T., D. Horgan, K. Gregor, and D. Silver. (2015). "Universal Value Function Approximators". In: *Proceedings of the 32nd International Conference on Machine Learning*. Ed. by F. Bach and D. Blei. Vol. 37. *Proceedings of Machine Learning Research*. Lille, France: PMLR. 1312–1320.

Schrittwieser, J., I. Antonoglou, T. Hubert, K. Simonyan, L. Sifre, S. Schmitt, A. Guez, E. Lockhart, D. Hassabis, T. Graepel, T. Lillicrap, and D. Silver. (2020). "Mastering Atari, go, chess and shogi by planning with a learned model". *Nature*. 588(7839): 604–609.

Strens, M. J. A. (2000). "A Bayesian Framework for Reinforcement Learning". In: *ICML*. Vol. 2000. 943–950.

Sutton, R. S. (1984). "Temporal Credit Assignment in Reinforcement Learning". *PhD thesis*. University of Massachusetts, Amherst.

Sutton, R. S. (1988). "Learning to predict by the methods of temporal differences". *Machine learning*. 3(1): 9–44.

Sutton, R. S. (1992). "Gain adaptation beats least squares". In: *Proceedings of the 7th Yale workshop on adaptive and learning systems*. Vol. 161168.

Sutton, R. S. and A. G. Barto. (2018). *Reinforcement learning: An introduction*. MIT press.

Sutton, R. S., J. Modayil, M. Delp, T. Degris, P. M. Pilarski, A. White, and D. Precup. (2011). "Horde: A scalable real-time architecture for learning knowledge from unsupervised sensorimotor interaction". In: *The 10th International Conference on Autonomous Agents and Multiagent Systems-Volume 2*. 761–768.

Tang, H., R. Houthooft, D. Foote, A. Stooke, O. Xi Chen, Y. Duan, J. Schulman, F. DeTurck, and P. Abbeel. (2017). "#Exploration: A Study of Count-Based Exploration for Deep Reinforcement Learning". In: *Advances in Neural Information Processing Systems*. Ed. by I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett. Vol. 30. Curran Associates, Inc. 2753–2762.

Tesauro, G. (1992). "Practical issues in temporal difference learning". *Machine learning*. 8(3): 257–277.

Tesauro, G. (1994). "TD-Gammon, a self-teaching backgammon program, achieves master-level play". *Neural computation*. 6(2): 215–219.

Thompson, W. R. (1933). "On the likelihood that one unknown probability exceeds another in view of the evidence of two samples". *Biometrika*. 25(3/4): 285–294.

Thompson, W. R. (1935). "On the theory of apportionment". *American Journal of Mathematics*. 57(2): 450–456.

Van Seijen, H., M. Fatemi, J. Romoff, R. Laroche, T. Barnes, and J. Tsang. (2017). "Hybrid Reward Architecture for Reinforcement Learning". In: *Advances in Neural Information Processing Systems*. Ed. by I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett. Vol. 30. Curran Associates, Inc. 5392–5402.

Veeriah, V., M. Hessel, Z. Xu, J. Rajendran, R. L. Lewis, J. Oh, H. P. van Hasselt, D. Silver, and S. Singh. (2019). "Discovery of Useful Questions as Auxiliary Tasks". In: *Advances in Neural Information Processing Systems*. Ed. by H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, and R. Garnett. Vol. 32. Curran Associates, Inc. 9310–9321.

Vlassis, N., M. Ghavamzadeh, S. Mannor, and P. Poupart. (2012). "Bayesian reinforcement learning". *Reinforcement learning: State of the Art*: 359–386.

Watkins, C. J. C. H. (1989). "Learning from delayed rewards". *PhD thesis*. King's College, Cambridge.

Welling, M. and Y. W. Teh. (2011). "Bayesian Learning via Stochastic Gradient Langevin Dynamics". In: *Proceedings of the International Conference on Machine Learning*.

Witten, I. H. (1976). "Learning to Control". *PhD thesis*. University of Essex.

Witten, I. H. (1977). "An adaptive optimal controller for discrete-time Markov environments". *Information and Control*. 34(5): 286–295.

Zimmert, J. and T. Lattimore. (2019). "Connections between mirror descent, Thompson sampling and the information ratio". *arXiv preprint arXiv:1905.11817*.