MASTER THESIS

Evaluating Reinforcement-Learning-based Sepsis Treatments via Tabular and Continuous Stationary Distribution Correction Estimation

written at the

Institute for Information Systems Engineering,
Department of Computer Science,
Technical University of Vienna

under the guidance of

Prof. Clemens Heitzinger

by

Richard Weiss
student number: 11805800
Viktoriagasse 7/2/22, 1150 Vienna

Vienna, December 19, 2024

Statutory Declaration

I declare under oath that I wrote this master's thesis independently and without outside help, that I did not use any sources or aids other than those specified, or that I identified the passages taken literally or analogously as such.

Vienna, December 19, 2024

Richard Weiss

Contents

1 Introduction **2**

2 Methods **4**
2.1 Measure Theory . . . 4
2.2 Optimization Theory . . . 5
2.2.1 Convex Optimization Theory . . . 7
2.2.2 Linear Optimization Theory . . . 13
2.3 Reproducing Kernel Hilbert Spaces . . . 14
2.4 Matrix Theory . . . 16
2.4.1 Perron-Frobenius Theorem . . . 16
2.4.2 Derivatives . . . 16
2.4.3 Block Matrices . . . 18
2.5 ODE Lemma . . . 18
2.6 Extended Delta Method . . . 19
2.6.1 Hadamard Differentiability . . . 19
2.6.2 Confidence Intervals . . . 22

3 Reinforcement Learning **25**
3.1 General . . . 25
3.2 Bellman Operators and Equations . . . 29
3.3 Bellman Linear Programs . . . 37
3.4 Classical Off-Policy Evaluation . . . 38
3.5 Stationary DIstribution Correction Estimation . . . 39

4 Algorithms **43**
4.1 Summary . . . 43
4.1.1 Tabular Case . . . 43
4.1.2 Continuous Case . . . 45
4.2 Tabular stationary DIstribution Correction Estimation . . . 45
4.3 Dual stationary DIstribution Correction Estimation . . . 46
4.3.1 Objectives . . . 47
4.3.2 The Dual Variable . . . 48
4.3.3 Implementation . . . 48
4.3.4 Convergence . . . 49
4.4 Generalized stationary DIstribution Correction Estimation . . . 50
4.4.1 Objectives . . . 50
4.4.2 Implementation . . . 51
4.4.3 Convergence . . . 52
4.5 Gradient stationary DIstribution Correction Estimation . . . 52
4.5.1 Objectives . . . 53
4.5.2 Implementation . . . 54
4.5.3 Convergence . . . 54

4.6 Confidence Interval DIstribution Correction Estimation . . . 63
4.6.1 Embedded Q-LP . . . 63
4.6.2 Generalized Estimating Equations . . . 65
4.6.3 Confidence Interval Derivation . . . 67
4.6.4 Confidence Interval Calculation . . . 71

5 Environments **74**
5.1 General . . . 74
5.2 Boyan Chain . . . 75
5.3 OpenAI Gym . . . 75
5.4 Medical . . . 76

6 Numerical Results **78**
6.1 Boyan Chain . . . 78
6.2 OpenAI Gym . . . 79
6.3 Medical Application . . . 81

7 Conclusions and Future Work **85**

Bibliography **87**

8 Appendix **90**
8.1 Additional Numerical Results . . . 90

Abstract

This work presents the results of state-of-the-art offline behavior agnostic policy evaluation algorithms based on stationary distribution correction estimation, evaluated within a healthcare setting using data from the `AmsterdamUMCdb`. We firstly, present the theory of these algorithms. This includes the introduction of four tabular estimators and a revision of the well known DualDICE, GenDICE, and GradientDICE. All algorithms are implemented in a modular open source Python library. In order to evaluate the efficacy of the algorithms, they are tested in the environments `BoyanChain` as well as the OpenAI Gym applications `FrozenLake`, `Taxi`, and `Cartpole`. The continuous state space algorithms DualDICE, GenDICE, and GradientDICE are run directly on the healthcare dataset. Additionally, the state space of healthcare applications is clustered in order to perform policy evaluation in the tabular setting. Our analysis provides a comprehensive examination of the practical functioning of all estimators, elucidating the underlying theory and the connections between the results and the theory.

1 Introduction

Due to the rapid development of *Machine Learning* (ML) in recent years, there has been a growing demand for ML algorithms to solve complex real-world applications. A subfield of ML called *Reinforcement Learning (RL)* provides a versatile theoretical framework for optimal decision making [38]. This is done by formulating a theoretical *environment* called *Markov Decision Process (MDP)*, characterized by a set of *states*, *actions*, and *rewards* and their interactions. They provide an abstract mathematical reflection of their real-world application counterparts, which may come from fields such as games [27], robotics [3], or conversational systems [15, 22, 1]. The goal is to train a *policy*, which executes actions in certain states of the environment, leading to a new state and reward. The quality of the policy is measured in terms of the *policy value (PV)*, which is the expectation of an exponential average of the rewards along an episode.

In many cases, the use of simulators for these MDPs is essential to facilitate a rapid and straightforward interaction between the agent training the policy and the environment. These can be constructed from a physical model or a dataset. However, this sort of simplification introduces a *sim-to-real gap*, which can lead to a misrepresentation of the original environment. In many cases, this makes it questionable, whether the behavior learned in the simulator can safely be transferred to the real world [36]. Especially applications, such as recommendation [23], education [24], autonomous driving [16, 18], and healthcare [30, 7, 8, 39, 19], where deploying a new policy can be expensive and risky, call for policy optimization and evaluation algorithms, which use an environment that most accurately represents the original task, ideally without the use of a simulator.

An approach, having recently gained high popularity, uses a limited and fixed dataset of samples describing an MDP, which models the application. Of the related contributions so far in *policy optimization (PO)* [33, 21, 20, 25, 13] and *policy evaluation (PE)* [34, 44, 41, 43, 13, 28], we will be using `NeuralDualDice` [34], `NeuralGenDice` [44], and `NeuralGradientDice` [41] in our work.

We refer to algorithms as *online*, if they are allowed to use an environment, with which they can easily interact, by starting an episode and giving it actions in a state and receiving the next state and reward. Removing this commodity limits us to the *offline* setting. These concepts are similar to notions of *on-policy (OnP)* and *off-policy (OffP)*. In the former setting, we gather data on an *evaluation policy* for an MDP by executing its own actions in an environment, while the latter uses a possibly different *behavior policy* for data collection [38]. Furthermore, classical off-policy methods need full knowledge, not only for the evaluation policy, but also for the behavior policy [38]. Moreover, data collection must be performed by a single behavior policy, whereas real-world data is most often gathered by a mix of multiple behavior policies. Should an algorithm be limited to samples generated by a single or multiple behavior policies, and not require any explicit knowledge of its distribution, then it is called *behavior agnostic*.

In healthcare, certain sensory data is gathered from patients during the treatments, carried out by human clinicians. In this publication, we use `AmsterdamUMCdb`[1] [39]. Thus, the original

[1] https://amsterdammedicaldatascience.nl/amsterdamumcdb

task of reproducing or even improving such treatments via RL is inherently offline. Furthermore, the distribution describing the mix of various clinician's behavior policies, can only be approximated at best [8], which motivates the use of behavior agnostic algorithms.

In this work, we apply the offline behavior agnostic algorithms `NeuralDualDice` [34], `NeuralGenDice` [44], and `NeuralGradientDice` [41] to estimate the value of a policy, which was trained to treat critically ill septic patients. The foundation of these three algorithms is the assertion that the policy value can be expressed in terms of the expectation of the reward taken with respect to the policy's stationary distribution. The stationary distribution can be expressed through a uniquely solvable system of linear equations or an eigenvalue problem, depending on whether the setting is discounted or undiscounted, respectively. Each algorithm has its own loss function, which is constructed from the equations and certain regularizers. The algorithms approximate a saddle point of their aforementioned loss function through the application of gradient descent and ascent. Because conventional approaches use clustering on the states of the dataset in order to provide the possible use of tabular algorithms [7], we also explore this approach, followed by our tabular policy evaluation methods `TabularVafe`, `TabularDice`, `TabularDualDice`, and `TabularGradientDice`, that are based on the same theory as the three above.

The policies and their respective datasets come from Bologheanu et al. [8]. Please refer to our GitHub repository[2] for details regarding the implementation.

In order to gain insight into the algorithm's practical behavior, we test them on various well established environments retaining certain selected properties. These include the tabular environment Boyan Chain [11], also used by Zhang et al. [41]. It is an environment with a scalable state space, where all the transition dynamics and rewards are known. This enables the comparison of our approximate solutions to an analytical one. For the continuous algorithms, we use one-hot-encoding to embed the state space of Boyan Chain. We further reinforce the credibility of our algorithms by additionally running on some famous environments from OpenAI Gym[3].

[2] https://github.com/MrWhiteRichard/dice_rl_sepsis.git
[3] https://gymnasium.farama.org/

2 Methods

2.1 Measure Theory

Consider a set X. Denote the space of probability measures, measures and signed measures on X into $\mathbb{R}$, respectively, by $\mathcal{P}(X) \subseteq \mathcal{M}(X) \subseteq \mathcal{S}(X)$. We can interpret a signed measure $S \in \mathcal{S}(X)$, a measure $M \in \mathcal{M}(X)$, and a probability measure $P \in \mathcal{P}(X)$, as a linear functional [14]. For any integrable $f : X \to \mathbb{R}$, write

$$Sf \doteq \int_X f \, \mathrm{d}S, \quad Mf \doteq \int_X f \, \mathrm{d}M, \quad \text{or} \quad Pf \doteq \mathbb{E}_P[f]. \tag{2.1}$$

For a class $\mathcal{H}$ of functions on X, consider the space of bounded linear functionals on $\mathcal{H}$ equipped with the uniform norm

$$L_\infty(\mathcal{H}) \doteq \{H : \mathcal{H} \to \mathbb{R} \text{ linear} \mid \|H\|_{L_\infty(\mathcal{H})} < \infty\}, \quad \text{where} \quad \|H\|_{L_\infty(\mathcal{H})} \doteq \sup_{\ell \in \mathcal{H}} |H\ell|,$$
$$\text{and} \quad B(\mathcal{H}, S) \doteq \left\{H \in L_\infty(\mathcal{H}) \mid H \text{ uniformly } \|\cdot\|_{L_2(S)}\text{-continuous}\right\}.$$

By interpreting signed measures as linear functionals, we can define the space of signed measures, measures and probability measures on X, bounded by $\|\cdot\|_{\mathcal{H}}$,

$$\mathcal{S}(X;\mathcal{H}) \doteq \{H \in \mathcal{S}(X) \mid \|H\|_{L_\infty(\mathcal{H})} < \infty\},$$
$$\mathcal{M}(X;\mathcal{H}) \doteq \{H \in \mathcal{M}(X) \mid \|H\|_{L_\infty(\mathcal{H})} < \infty\},$$
$$\mathcal{P}(X;\mathcal{H}) \doteq \{H \in \mathcal{P}(X) \mid \|H\|_{L_\infty(\mathcal{H})} < \infty\}.$$

Let $X_\mathrm{D} \subseteq \mathbb{R}^m$ be discrete and $X_\mathrm{C} \subseteq \mathbb{R}^m$ continuous. Denote the sub sets of X_D by $\mathcal{P}(X_\mathrm{D})$, the Borel-sets on X_C by $\mathcal{B}(X_\mathrm{C})$, the counting measure by μ, and the m-dimensional Lebesgue measure by λ. Then, for P we will denote its Radon-Nikodym derivative with the corresponding lower case letter, and vice versa, i.e.,

$$p = \frac{\mathrm{d}P}{\mathrm{d}\mu} \quad \text{and} \quad \forall A \in \mathcal{P}(X_\mathrm{D}) : \; P(A) = \int_A p \, \mathrm{d}\mu = \sum_{x \in A} p(x),$$
$$p = \frac{\mathrm{d}P}{\mathrm{d}\lambda} \quad \text{and} \quad \forall B \in \mathcal{B}(X_\mathrm{C}) : \; P(B) = \int_B p \, \mathrm{d}\lambda = \int_B p(x) \, \mathrm{d}x.$$

Let $\Delta(X)$ be the set of probability density functions on X, i.e., the above Radon-Nikodym derivatives, depending on whether $X \subseteq \mathbb{R}^m$ is continuous or discrete. For a given $q \in \Delta(X)$, we write the set of distributions, absolutely continuous with respect to q, as

$$\Delta_q(X) \doteq \{p \in \Delta(X) \mid p \ll q\}.$$

Sometimes, q will have finite support $\mathrm{supp}(q) = \{x_1, \ldots, x_n\}$. Define the set of probability vectors as

$$\Delta^n \doteq \left\{ \vec{p} \in \mathbb{R}^n \;\middle|\; \vec{p} \geq 0, \; \sum_{i=1}^n p_i = 1 \right\}.$$

Then, we have

$$\Delta_q(X) = \left\{ \sum_{i=1}^{n} p_i \mathbb{1}_{x_i=x} \;\middle|\; \vec{p} \in \Delta^n \right\}. \tag{2.2}$$

Now, consider a *dataset* $\mathcal{D} = (x_i)_{i=1}^n \subset X$, where the samples are taken by distribution $p^{\mathcal{D}} \in \Delta(X)$. We define the *empirical distribution* $\hat{p}^{\mathcal{D}} \in \Delta_{p^{\mathcal{D}}}(X)$ via

$$\hat{p}^{\mathcal{D}}(x) \doteq \frac{1}{n} \sum_{i=1}^{n} \mathbb{1}_{x_i=x}. \tag{2.3}$$

Note that, since the samples $x_1, \ldots, x_n$ are random variables, the empirical distribution $\hat{p}^{\mathcal{D}}$ is a random function on X. Since $\operatorname{supp}(\hat{p}^{\mathcal{D}}) = \{x_1, \ldots, x_n\}$, we can apply (2.2) to $\hat{p}^{\mathcal{D}}$. This holds even if there are samples in $\mathcal{D}$ that occur more than once.

2.2 Optimization Theory

The DICE algorithms all require additional optimization techniques to regular RL algorithms. To this end Nachum and Dai have published a review of optimization theory specifically for DICE [32]. This section provides a far more detailed and rigorous review of the necessary background theory.

Consider the *general optimization problem*

$$\begin{aligned} &\min_{x \in \mathbb{R}^n} f(x) \quad \text{subject to} \quad g(x) = 0 \quad \text{and} \quad h(x) \leq 0, \\ &\qquad\qquad \text{where} \quad f : \mathbb{R}^n \to \mathbb{R}, \quad g : \mathbb{R}^n \to \mathbb{R}^\ell, \quad h : \mathbb{R}^n \to \mathbb{R}^{k-\ell}. \end{aligned} \tag{2.4}$$

We can summarize this entire problem in two different ways:

1. For a subset $C \subseteq \mathbb{R}$ we define the *indicator function*

$$\delta_C : \mathbb{R} \to \{0, \infty\}, \quad \delta_C(x) \doteq \begin{cases} 0, & \text{if } x \in C, \\ \infty, & \text{else.} \end{cases}$$

It is easy to verify that the initial formulation of the optimization problem can be restated in terms of the *summarized objective*

$$\min_{x \in \mathbb{R}^n} P(x) \doteq f(x) + \sum_{i=1}^{\ell} \delta_{\{0\}}(g_i(x)) + \sum_{i=1}^{k-\ell} \delta_{\mathbb{R}_{\leq 0}}(h_i(x)).$$

2. We introduce the *Lagrangian function*

$$L : \begin{cases} \mathbb{R}^n \times \mathbb{R}^\ell \times \mathbb{R}^{k-\ell}_{\geq 0} \to \mathbb{R}, \\ x \mapsto f(x) + \lambda^\top g(x) + \mu^\top h(x), \end{cases}$$

and consider the Lagrangian formulation

$$\min_{x \in \mathbb{R}^n} P(x) \doteq \max_{\lambda \in \mathbb{R}^\ell} \max_{\mu \in \mathbb{R}^{k-\ell}_{\geq 0}} L(x, \lambda, \mu).$$

We can check that both definitions of P are the same.

- If x satisfies all constraints, $g_i(x) = 0$ and $h_i(x) \leq 0$, then $L(x, \lambda, \mu)$ is maximized when taking $\lambda = \mu = 0$, and its value will be $f(x)$.
- If x violates some constraint, $g_i(x) \neq 0$ or $h_i(x) > 0$, then $L(x, \lambda, \mu) \to \infty$ for $\lambda_i \to \operatorname{sgn} g_i(x)\infty$ or $\mu_i \to \infty$, respectively.

This motivates us to consider a similar formulation to the Lagrangian

$$\max_{\lambda \in \mathbb{R}^\ell} \max_{\mu \in \mathbb{R}^{k-\ell}_{\geq 0}} D(\lambda, \mu) \doteq \min_{x \in \mathbb{R}^n} L(x, \lambda, \mu).$$

We call the initial problem the *primal* and this one the *dual optimization problem*. Consequently, D is named the *dual function*. A very important result in duality theory is Weak Duality, provided in Proposition 2.2.1.

Proposition 2.2.1 (Weak Duality)**.** *Consider the general optimization problem* (2.4)*. Then*

$$\max\{D(\lambda, \mu) \mid \lambda \in \mathbb{R}^k,\ \mu \in \mathbb{R}^\ell_+\} \leq \min\{f(x) \mid x \in \mathbb{R}^n,\ g(x) = 0,\ h(x) \leq 0\}.$$

Proof. Let x^* be optimal with respect to the primal problem, i.e.,

$$x^* = \arg\min\{f(x) \mid x \in \mathbb{R}^n,\ g(x) = 0,\ h(x) \leq 0\}.$$

This leads to

$$D(\lambda, \mu) \leq L(x^*, \lambda, \mu) = f(x^*) + \lambda^\top g(x^*) + \mu^\top h(x^*) \leq f(x^*).$$

For the last inequality we used

$$\begin{aligned} g(x^*) = 0 &\implies \lambda^\top g(x^*) = 0, \\ \mu \geq 0,\ h(x^*) \leq 0 &\implies \mu^\top h(x^*) \leq 0. \end{aligned}$$

□

We will assume that for all $i = 1, \ldots, m$, where $m \leq n$, we add $x_i \geq 0$ to our general optimization problem. Since we could simply include these inequalities via h, this does not add any extra expressiveness. However, in some cases it turns out to be convenient to separate these simple constraints.

With Lemma 2.2.2 we provide a criterion that allows us to check, whether two optimization problems are dual to one another.

Lemma 2.2.2. *Let L_P and L_D be the Lagrangians of some optimization problems (P) and (D), respectively, where*

$$L_P(x, (\lambda, \mu)) = L_D((\mu, \lambda), x) \quad \text{for all} \quad x \in \mathbb{R}^m_{\geq 0} \times \mathbb{R}^{n-m}, \quad \lambda \in \mathbb{R}^\ell, \quad \mu \in \mathbb{R}^{k-\ell}_{\geq 0}.$$

Then (D) is the dual of (P).

Proof. The dual function of (P) reads

$$D_{\mathrm{P}}(\lambda, \mu) = \min_{x \in \mathbb{R}^m_{\geq 0} \times \mathbb{R}^{n-m}} L_{\mathrm{P}}(x, (\lambda, \mu)) = \min_{x \in \mathbb{R}^m_{\geq 0} \times \mathbb{R}^{n-m}} L_{\mathrm{D}}((\mu, \lambda), x) = P_{\mathrm{D}}((\mu, \lambda)),$$

where P_{D} is the summarized objective of (D).

□

2.2.1 Convex Optimization Theory

This section summarizes the elements of convex analysis [10] which are important for this work. We say that a function $f : \Omega \to \overline{\mathbb{R}}$ is *convex* on a domain $\Omega \subseteq \mathbb{R}^n$ if, for arbitrary $0 \leq \alpha \leq 1$ and $x_1, x_2 \in \Omega$, we have

$$f(\alpha x_1 + (1-\alpha)x_2) \leq \alpha f(x_1) + (1-\alpha) f(x_2).$$

We call $f_* : \Omega_* \to \overline{\mathbb{R}}$ the *convex conjugate* or *Legendre-Fenchel transformation* of f, where

$$f_*(y) \doteq \sup_{x \in \Omega} \langle x, y \rangle - f(x) \quad \text{and} \quad \Omega_* = \{y \in \mathbb{R}^n \mid f_*(y) < \infty\}.$$

Remark 2.2.3. Let $f : \Omega \to \overline{\mathbb{R}}$ be a convex and differentiable function on $\Omega = \mathbb{R}^n$ and $f_* : \Omega_* \to \overline{\mathbb{R}}$ its conjugate. Assume that f' is invertible.

By setting the derivative to zero,

$$0 \overset{!}{=} \partial\Big(\langle x, y \rangle - f(x)\Big) = y - \partial f(x), \quad \text{we get} \quad f_*(y) = \big(\langle x, y \rangle - f(x)\big)\Big|_{x=(\partial f)^{-1}(y)}.$$

■

Example 2.2.4. Consider a symmetric positive definite matrix $A \in \mathbb{R}^{n \times n}$, a vector $b \in \mathbb{R}$ and a constant $c \in \mathbb{R}$. Then

$$f : \begin{cases} \mathbb{R} \to \overline{\mathbb{R}} \\ x \mapsto \frac{1}{2} x^\top A x + b^\top x + c \end{cases} \quad \text{has conjugate} \quad f_* : \begin{cases} \mathbb{R} \to \overline{\mathbb{R}} \\ y \mapsto \frac{1}{2}(y-b)^\top A^{-1}(y-b) - c. \end{cases}$$

We use Remark 2.2.3 and calculate

$$\nabla f(x) = Ax + b, \quad \text{so} \quad (\partial f)^{-1}(y) = A^{-1}(y-b)$$

and

$$f_*(y) = (A^{-1}(y-b))^\top y - \frac{1}{2}(A^{-1}(y-b))^\top A(A^{-1}(y-b)) - b^\top (A^{-1}(y-b)) - c.$$

■

Example 2.2.5. Let p and q be *Hölder conjugates*, i.e., $p, q > 1$ and $1/p + 1/q = 1$. Then

$$f : \begin{cases} \mathbb{R} \to \overline{\mathbb{R}} \\ x \mapsto \frac{1}{p}|x|^p \end{cases} \quad \text{has conjugate} \quad f_* : \begin{cases} \mathbb{R} \to \overline{\mathbb{R}} \\ y \mapsto \frac{1}{q}|y|^q. \end{cases}$$

Recall that Young's inequality for products reads

$$xy \leq \frac{1}{p}|x|^p + \frac{1}{q}|y|^q,$$

where equality holds iff $|x|^p = |y|^q$. From this, it immediately follows that

$$\sup_{x \in \mathbb{R}} xy - \frac{1}{p}|x|^p = \frac{1}{q}|y|^q.$$

Furthermore,

$$f : \begin{cases} \mathbb{R}^n \to \overline{\mathbb{R}} \\ x \mapsto \frac{1}{p}\|x\|_p^p \end{cases} \quad \text{has conjugate} \quad f_* : \begin{cases} \mathbb{R}^n \to \overline{\mathbb{R}} \\ y \mapsto \frac{1}{q}\|y\|_q^q. \end{cases}$$

Using Hölder's and Young's inequality, we calculate

$$\sup_{x\in\mathbb{R}^n} \langle x, y\rangle - \frac{1}{p}\|x\|_p^p \leq \sup_{x\in\mathbb{R}^n} \|x\|_p\|y\|_q - \frac{1}{p}\|x\|_p^p = \sup_{\xi\in\mathbb{R}} \xi\|y\|_q - \frac{1}{p}|\xi|^p = \frac{1}{q}\|y\|_q^q.$$

The inequality becomes an equality for linear independent $|x|^p$ and $|y|^q$ and $\|x\|_p^p = \|y\|_q^q$ and the last equality follows from the above.

■

Another useful insight is that building convex conjugates is additive, as stated in Lemma 2.2.6.

Lemma 2.2.6. *Let $f : A \to \overline{\mathbb{R}}$ and $g : B \to \overline{\mathbb{R}}$ be convex functions. Then*

$$h : \begin{cases} A \times B \to \overline{\mathbb{R}} \\ x \mapsto f(x_1) + g(x_2) \end{cases} \quad \textit{has conjugate} \quad h_* : \begin{cases} A_* \times B_* \to \overline{\mathbb{R}} \\ x \mapsto f_*(x_1) + g_*(x_2). \end{cases}$$

Proof. Let us firstly prove that $(A \times B)_* = A_* \times B_*$. For all $x_1 \in A$ and $x_2 \in B$ we have

$$\max\{f_*(x_1), g_*(x_2)\} < \infty \iff h_*(x) < \infty.$$

Secondly, we calculate

$$h_*(y) = \sup_{x\in A\times B} \langle x, y\rangle - h(x) = \sup_{\substack{x_1\in A \\ x_2\in B}} \langle x_1, y_1\rangle + \langle x_2, y_2\rangle - f(x_1) - f(x_2) = f_*(y_1) + g_*(y_2).$$

□

Example 2.2.7. Consider the sets $A \subseteq \mathbb{R}^n$ and $B \subseteq \mathbb{R}^m$. By Lemma 2.2.6,

$$f : \begin{cases} \mathbb{R}^{nm} \to \overline{\mathbb{R}} \\ x \mapsto \delta_{A\times B}(x) \\ \qquad = \delta_A(x_1) + \delta_B(x_2) \end{cases} \quad \text{has conjugate} \quad f_* : \begin{cases} \mathbb{R}^{nm} \to \overline{\mathbb{R}} \\ y \mapsto \delta^*_{A\times B}(y) \\ \qquad = \delta^*_A(y_1) + \delta^*_B(y_2). \end{cases}$$

Now, fix a domain $\Omega \subseteq \mathbb{R}^n$ and $a \in \Omega$. Then

$$f : \begin{cases} \Omega \to \overline{\mathbb{R}} \\ x \mapsto \delta_{\{a\}}(x) \end{cases} \quad \text{has conjugate} \quad f_* : \begin{cases} \Omega \to \overline{\mathbb{R}} \\ y \mapsto \langle a, y\rangle, \end{cases}$$

since

$$\begin{aligned} \{\langle x, y\rangle - f(x) \mid x = a\} &= \{\langle x, y\rangle\}, \\ \text{and} \quad \{\langle x, y\rangle - f(x) \mid x \neq a\} &= \{-\infty\}. \end{aligned}$$

In particular, for $a = 0$, we get that

$$f : \begin{cases} \Omega \to \overline{\mathbb{R}} \\ x \mapsto \delta_{\{0\}}(x) \end{cases} \quad \text{has conjugate} \quad f_* : \begin{cases} \Omega \to \overline{\mathbb{R}} \\ y \mapsto 0 = \delta_\Omega(y). \end{cases}$$

Finally,

$$f : \begin{cases} \mathbb{R}^n \to \overline{\mathbb{R}} \\ x \mapsto \delta_{\mathbb{R}^n_{\geq 0}}(x) \end{cases} \quad \text{has conjugate} \quad f_* : \begin{cases} \mathbb{R}^n \to \overline{\mathbb{R}} \\ y \mapsto \delta_{\mathbb{R}^n_{\leq 0}}(y), \end{cases}$$

since we can use the first identity and case $n = 1$, which follows from

$$\{xy - f(x) \mid x \in \mathbb{R}_{\geq 0}\} = \{xy \mid x \in \mathbb{R}_{\geq 0}\} = \begin{cases} \mathbb{R}_{\geq 0}, & \text{if } y > 0, \\ \mathbb{R}_{\leq 0}, & \text{if } y < 0, \\ \{0\}, & \text{else.} \end{cases}$$

$$\text{and} \quad \{xy - f(x) \mid x \in \mathbb{R}_{<0}\} = \{-\infty\}.$$

■

Example 2.2.8. Let $f : \mathbb{R}^n \to \overline{\mathbb{R}}$ be a convex function, $a \in \mathbb{R}^n$ and $b > 0$. Then

$$g : \begin{cases} \mathbb{R}^{nm} \to \overline{\mathbb{R}} \\ x \mapsto \langle a, x \rangle + bf(x) \end{cases} \quad \text{has conjugate} \quad g_* : \begin{cases} \mathbb{R}^{nm} \to \overline{\mathbb{R}} \\ y \mapsto bf_*\left(\frac{y-a}{b}\right), \end{cases}$$

since

$$\begin{aligned} g_*(y) &= \sup_{x \in \mathbb{R}^n} \langle x, y \rangle - g(x) = \sup_{x \in \mathbb{R}^n} \langle x, y - a \rangle - bf(x) \\ &= b\Big(\sup_{x \in \mathbb{R}^n} \langle x, (y-a)/b \rangle - f(x) \Big) = bf_*\left(\frac{y-a}{b} \right). \end{aligned}$$

■

Since the supremum is taken over a family of functions affine in y, the convex conjugate is convex. Another useful property is that it can be specified in a differential equation, up to an additive constant.

Lemma 2.2.9. *Let $f : \Omega \to \overline{\mathbb{R}}$ be a convex and differentiable function on $\Omega = \mathbb{R}$ and $f_* : \Omega_* \to \overline{\mathbb{R}}$ its conjugate. Assume that f' is invertible and its inverse $(f')^{-1}$ is differentiable. Then*

$$f_*' = (f')^{-1}.$$

Proof. Let x_y be a critical point of the function $x \mapsto xy - f(x)$. Since f is convex, the supremum is achieved at $g(y) \doteq (f')^{-1}(y) = x_y$. Therefore, we get

$$f_*(y) = x_y y - f(x_y) = g(y)y - f(g(y)).$$

Differentiating with respect to y leaves us with

$$\begin{aligned} f_*'(y) &= g'(y)y + g(y) - f'(g(y))g'(y) \\ &= g'(y)y + g(y) - yg'(y) \\ &= (f')^{-1}(y). \end{aligned}$$

□

If there exists $x \in \Omega$ such that $f(x) < \infty$ and $f(x) > -\infty$ for all $x \in \Omega$, then f is a *proper function*. When $\{x \in \Omega \mid f(x) > \alpha\} \neq \emptyset$ for any $\alpha \in \mathbb{R}$, then f is said to be *lower semi-continuous*. In case a function fulfills all of these conditions, we can apply the Fenchel–Moreau Theorem 2.2.10. It implies that all the examples we have seen so far can be reversed.

Theorem 2.2.10 (Fenchel–Moreau)**.** *Consider a function $f : \Omega \to \overline{\mathbb{R}}$ on a domain $\Omega \subseteq \mathbb{R}^n$. If $f \equiv \pm\infty$ or f is proper, lower semi-continuous and convex, then $f = f^{**}$, i.e., for any $x \in \Omega$, we have*

$$f(x) = \max_{y \in \Omega_*} \langle x, y \rangle - f_*(y) = f^{**}(x).$$

f-Divergence

Here, we review some material from Nachum and Dai [32].

Let p and q be probability distributions with $p \ll q$. A common way of characterizing their deviation is using the *Kullbach-Leibler (KL) divergence*

$$D_{\mathrm{KL}}(p \parallel q) \doteq \mathbb{E}_p \left[\log \frac{p}{q} \right].$$

We will call a convex function $f : \mathbb{R}_{\geq 0} \to \overline{\mathbb{R}}$ *divergence conform* iff

$$\forall t > 0 : \; f(t) < \infty, \quad f(0) = \lim_{t \downarrow 0} f(t) \quad \text{and} \quad f(1) = 0.$$

Using such a function, we can generalize the KL divergence and define the f-divergence as

$$D_f(p \parallel q) \doteq \mathbb{E}_q \left[f\left(\frac{p}{q} \right) \right].$$

We get back the KL divergence by choosing $f(x) = x \log x$, because then

$$D_f(p \parallel q) = \mathbb{E}_q \left[\frac{p}{q} \log \frac{p}{q} \right] = \mathbb{E}_p \left[\log \frac{p}{q} \right] = D_{\mathrm{KL}}(p \parallel q).$$

Another prominent member of the group of f-divergences is the χ^2-divergence D_{χ^2}, where $f(x) = (x-1)^2$, so

$$D_{\chi^2}(p \parallel q) \doteq \mathbb{E}_q \left[\left(\frac{p}{q} - 1 \right)^2 \right].$$

Applying Jensen's inequality to f, we get that the f-divergence is non-negative,

$$D_f(p \parallel q) \geq f\left(\mathbb{E}_q \left[\frac{p}{q} \right] \right) = f(\mathbb{E}_p[1]) = f(1) = 0.$$

Later on, we are going to need some other properties of the f-divergence as a function in the first argument. We summarize them in Lemma 2.2.11.

Lemma 2.2.11. *Let f be divergence conform and fix a distribution q. Consider the f-divergence $D_f(\cdot \parallel q)$ as a function on*

$$\Omega \doteq \{x : \mathbb{R} \to \mathbb{R} \mid 0 \leq x \ll q\}.$$

This function is convex with conjugate

$$D_f^*(y \parallel q) = \mathbb{E}_q[f_* \circ y].$$

Proof. Consider $x, y \in \Omega$ and $0 \leq \alpha \leq 1$. Since f is convex, we can use the monotonicity of the expected value and get

$$\begin{aligned} D_f(\alpha x + (1-\alpha)y \parallel q) &= \mathbb{E}_{z\sim q}\left[f\left(\alpha\frac{x(z)}{q(z)} + (1-\alpha)\frac{y(z)}{q(z)}\right)\right] \\ &\leq \mathbb{E}_{z\sim q}\left[\alpha f\left(\frac{x(z)}{q(z)}\right) + (1-\alpha)f\left(\frac{y(z)}{q(z)}\right)\right] \\ &= \alpha D_f(x \parallel q) + (1-\alpha)D_f(y \parallel q). \end{aligned}$$

The convex conjugate can be calculated as

$$\begin{aligned} D_f^*(y \parallel q) &= \sup_{x\in\Omega}\langle x, y\rangle - D_f(x \parallel q) \\ &= \sup_{x\in\Omega}\mathbb{E}_{z\sim q}\left[\frac{x(z)y(z)}{q(z)}\right] - \mathbb{E}_{z\sim q}\left[f\left(\frac{x(z)}{q(z)}\right)\right] \\ &= \mathbb{E}_{z\sim q}\left[\sup_{0\leq x\ll q(z)}\frac{x}{q(z)}y(z) - f\left(\frac{x}{q(z)}\right)\right] \\ &= \mathbb{E}_{z\sim q}\left[\sup_{\xi\geq 0}\xi y(z) - f(\xi)\right] \\ &= \mathbb{E}_{z\sim q}[f_*(y(z))]. \end{aligned}$$

□

Example 2.2.12. Assume the χ^2-divergence $D_{\chi^2}(\parallel p)$ to be a function on Ω, just like in Lemma 2.2.11, then

$$D_{\chi^2}^*(y \parallel p) = \mathbb{E}_{z\sim p}\left[y(z) + \frac{y(z)^2}{4}\right].$$

We calculate the convex conjugate of $f(x) = (x-1)^2$. Build the derivative and its inverse respectively,

$$f'(x) = 2(x-1) \quad \text{and} \quad (f')^{-1}(y) = \frac{y}{2} + 1.$$

Now, use Lemma 2.2.9 to get

$$f^*(y) = \left(\frac{y}{2}+1\right)y - \left(\frac{y}{2}+1-1\right)^2 = \frac{y^2}{2} + y - \frac{y^2}{4} = y + \frac{y^2}{4}.$$

■

Example 2.2.13. Consider the KL divergence $D_{\mathrm{KL}}(\parallel q)$, but only as a function on probability measures absolutely continuous with respect to q, i.e., $\Omega = \Delta_q$, then

$$D_{\mathrm{KL}}^*(y \parallel q) = \log \mathbb{E}_q[\exp y].$$

Suppose, the domain of our probability measures is finite. Then, similar to 2.2, we can take

$$\Delta_q = \left\{ x \in \mathbb{R}^n \;\middle|\; 0 \leq x \ll q,\ \sum_{i=1}^n x_i = 1 \right\}.$$

Since this simplex is compact, the supremum in the definition of the convex conjugate is obtained at some $x \in \Delta_q$, i.e.,

$$D^*_{\mathrm{KL}}(y \parallel q) = \sup_{\xi \in \Omega} \langle \xi, y \rangle - D_{\mathrm{KL}}(\xi \parallel q) = \sum_{i=1}^{n} x_i \left(y_i - \log \frac{x_i}{q_i} \right).$$

To find x, we consider the Lagrangian

$$L(x, \lambda) \doteq \sum_{i=1}^{n} x_i \left(y_i - \log \frac{x_i}{q_i} \right) + \lambda \left(\sum_{i=1}^{n} x_i - 1 \right).$$

We set its partial derivatives to zero,

$$0 \stackrel{!}{=} \frac{\partial}{\partial x_i} L(x, \lambda) = \left(y_i - \log \frac{x_i}{q_i} \right) - x_i \left(\frac{1}{x_i / q_i} \frac{1}{q_i} \right) + \lambda = y_i - \log \frac{x_i}{q_i} - 1 + \lambda.$$

This yields

$$\log \frac{x_i}{q_i} = y_i - 1 + \lambda \quad \text{and} \quad x_i = q_i e^{y_i - 1 + \lambda}.$$

Notice, since $e^{y_i - 1 + \lambda} > 0$, we have $0 \leq x \ll q$. Furthermore, we must choose λ, such that

$$1 = \sum_{i=1}^{n} x_i = \sum_{i=1}^{n} q_i e^{y_i} / e^{1-\lambda} \quad \text{and} \quad 1 - \lambda = \log \sum_{i=1}^{n} q_i e^{y_i}.$$

Finally, substituting $\log \frac{x_i}{q_i}$, we get

$$\begin{aligned} D^*_{\mathrm{KL}}(y \parallel q) &= \sum_{i=1}^{n} x_i (y_i - (y_i - 1 + \lambda)) = (1 - \lambda) \sum_{i=1}^{n} x_i \\ &= \log \sum_{i=1}^{n} q_i e^{y_i} = \log \mathbb{E}_{z \sim q}[\exp y(z)]. \end{aligned}$$

■

Lastly, we will define the ϵ-ball for some $\epsilon > 0$ around some probability distribution $q \in \Delta(\Omega)$ with respect to the f-divergence, as

$$B^f_\epsilon(q) \doteq \{ p \in \Delta_q \mid D_f(p \parallel q) \leq \epsilon \}. \tag{2.5}$$

If we chose $q \doteq \hat{p}_n$, the empirical distribution from (2.3), we can write the f-divergence as

$$D_f(p \parallel \hat{p}_n) = \sum_{i=1}^{n} f \left(\frac{p_i}{1/n} \right) \frac{1}{n} = \frac{1}{n} \sum_{i=1}^{n} f(n p_i). \tag{2.6}$$

Fenchel Optimization

Consider functions $f : X \to \overline{\mathbb{R}}$ and $g : Y \to \overline{\mathbb{R}}$ as well as a linear map $A : X \to Y$. Define the Fenchel optimization problems

$$\min_{x \in X} J_{\mathrm{P}}(x) \doteq f(x) + g(Ax), \quad \max_{y \in Y} J_{\mathrm{D}}(y) \doteq -f_*(-A_* y) - g_*(y).$$

Not only does *Weak Duality* hold, under mild conditions, we even get *Strong Duality*,

$$\min_{x\in X} J_{\mathrm{P}}(x) \geq \max_{y\in Y} J_{\mathrm{D}}(y), \tag{2.7}$$

$$\min_{x\in X} J_{\mathrm{P}}(x) = \max_{y\in Y} J_{\mathrm{D}}(y). \tag{2.8}$$

One can also show that if f_*' is well-defined, by using the optimal solution y^* of the dual, we can get the optimal solution of the primal through $x^* = f_*'(-A_*y^*)$. More generally, the set of all primal solutions is $\partial f_*(-A_*y^*) \cap A^{-1}\partial g_*(y^*)$.

We can explain why the first one of these problems is primal and the second one is dual by defining a Lagrangian

$$L(x,y) \doteq f(x) + \langle Ax, y\rangle - g_*(y).$$

This results in

$$\begin{aligned}\sup_{y\in\Omega_*} L(x,y) &= \sup_{y\in\Omega_*} f(x) + \langle Ax,y\rangle - g_*(y)\\ &= f(x) + \sup_{y\in\Omega_*} \langle y, Ax\rangle - g_*(y)\\ &= f(x) + g(Ax),\end{aligned} \qquad \begin{aligned}\inf_{x\in\Omega} L(x,y) &= \inf_{x\in\Omega} f(x) + \langle Ax,y\rangle - g_*(y)\\ &= -\sup_{x\in\Omega} \langle x, -A_*y\rangle - f(x) - g_*(y)\\ &= -f_*(-A_*y) - g_*(y).\end{aligned}$$

Therefore, Strong Duality (2.8) leads to *Lagrange Duality*

$$\inf_{x\in\Omega}\sup_{y\in\Omega_*} L(x,y) = \sup_{y\in\Omega_*}\inf_{x\in\Omega} L(x,y). \tag{2.9}$$

2.2.2 Linear Optimization Theory

Consider the optimization problem, also called *Linear Programm (LP)*,

$$\min f(x) = \langle c, x\rangle$$

$$\begin{aligned}\text{s.t.}\quad b_{\leq \ell} - A_{\leq \ell}x &= g(x) = 0,\\ b_{>\ell} - A_{>\ell}x &= h(x) \leq 0,\\ x_1,\dots,x_m &\geq 0,\end{aligned}$$

$$\begin{aligned}x \in \mathbb{R}^m_{\geq 0}\times\mathbb{R}^{n-m}, \quad y = (\lambda,\mu) \in \mathbb{R}^\ell \times \mathbb{R}^{k-\ell}_{\geq 0},\\ A \in \mathbb{R}^{k\times n}, \quad b\in\mathbb{R}^k, \quad c\in\mathbb{R}^n.\end{aligned}$$

The Lagrangian for this optimization problem reads

$$L(x,y) = c^\top x + y^\top(b - Ax) = b^\top y + x^\top(c - A^\top y),$$

and hence, the dual function

$$D(y) = \min_{x\in\mathbb{R}^m_{\geq 0}\times\mathbb{R}^{n-m}} L(x,y) = \begin{cases} b^\top y, & \text{if} \quad \begin{aligned}(c-A^\top y)_{\leq m} &\geq 0,\\ (c-A^\top y)_{>m} &= 0,\end{aligned}\\ -\infty, & \text{else.}\end{cases}$$

Optimizing the dual function leads to the dual optimization problem

$$\max \tilde{f}(y) = \langle b, y \rangle$$

$$\begin{aligned} \text{s.t.} \quad (A^\top)_{>m} y - c_{>m} &= \tilde{g}(y) = 0, \\ (A^\top)_{\leq m} y - c_{\leq m} &= \tilde{h}(y) \leq 0, \\ y_{\ell+1}, \ldots, y_k &\geq 0. \end{aligned}$$

We now want to apply Strong Duality (2.8) from Convex Optimization to Linear Optimization. To this end, we reformulate our primal LP as a Fenchel program with

$$f(x) = \langle c, x \rangle + \delta_{\mathbb{R}^m_{\geq 0} \times \mathbb{R}^{k-m}}(x) \quad \text{and} \quad g(x) = \delta_{\{b_{\leq m}\} \times [b_{>m}, \infty)}(x).$$

Building conjugates via Example 2.2.8 and 2.2.7 and Lemma 2.2.6, we get

$$\begin{aligned} f_*(y) &= \delta^*_{\mathbb{R}^m_{\geq 0} \times \mathbb{R}^{k-m}}(y - c) \\ &= \delta^*_{\mathbb{R}^m_{\geq 0}}(y_{\leq m} - c_{\leq m}) + \delta^*_{\mathbb{R}^{k-m}}(y_{>m} - c_{>m}) \\ &= \delta_{\mathbb{R}^m_{\leq 0}}(y_{\leq m} - c_{\leq m}) + \delta_{\{0\}}(y_{>m} - c_{>m}) \\ &= \delta_{\mathbb{R}^m_{\leq 0} \times \{0\}}(y - c), \end{aligned}$$

$$\begin{aligned} g_*(y) &= \delta^*_{\{b_{\leq m}\}}(y_{\leq m}) + \delta^*_{[b_{>m}, \infty)}(y_{>m}) \\ &= \langle b_{\leq m}, y_{\leq m} \rangle + \langle b_{>m}, y_{>m} \rangle \\ &\quad + \delta_{\mathbb{R}^m}(y_{\leq m}) + \delta_{\mathbb{R}^{k-m}_{\leq 0}}(y_{>m}) \\ &= \langle b, y \rangle + \delta_{\mathbb{R}^m \times \mathbb{R}^{k-m}_{\leq 0}}(y), \end{aligned}$$

because

$$\delta^*_{[b_{>m}, \infty)}(y_{>m}) = \left(\delta_{\mathbb{R}^{k-m}_{\geq 0}}(y_{>m} - b_{>m})\right)_* = \left(\delta^*_{\mathbb{R}^{k-m}_{\leq 0}}(y_{>m} - b_{>m})\right)_* = \langle b_{>m}, y_{>m} \rangle + \delta_{\mathbb{R}^{k-m}_{\leq 0}}(y_{>m}).$$

If we now make the switch $y \to -y$, we get that the dual Fenchel optimization problem is a reformulation of our dual LP.

2.3 Reproducing Kernel Hilbert Spaces

The well known *reproducing kernel Hilbert spaces (RKHS)* provide a theoretical framework to characterize neural networks [4, 5]. Since in RL algorithms we often assume that our function classes are instances of neural networks, we can use RKHS in their theoretical analysis.

Another use case of RKHS involves the construction of a metric, the *maximum mean discrepancy (MMD)*, to distinguish between probability distributions [29]. Mousavi et al. [28] consider parameterizations of the stationary distribution correction d^π and its application to the backwards Bellman oprator $\mathcal{P}^\pi_*$, whose support is a subset of the dataset $\mathcal{D}$. In this way, the can find explicit equations for the MMD between these parameterizations and minimize them, thus approximating the solution to the backwards Bellman equations (3.9). Since the focus of this work is on DICE, we do not pursue this approach.

Let $\mathcal{H}$ be a Hilbert space over $\mathbb{F}$ with scalar product $\langle \cdot, \cdot \rangle_\mathcal{H}$. Denote the algebraic and continuous dual space of $\mathcal{H}$, respectively, by

$$\mathcal{H}^* \doteq \{T : \mathcal{H} \to \mathbb{F} \mid T \text{ linear}\} \quad \text{and} \quad \mathcal{H}' \doteq \{T \in \mathcal{H}^* \mid \|T\|_{\mathcal{H}^*} < \infty\},$$
$$\text{where} \quad \|T\|_{\mathcal{H}^*} \doteq \sup_{f \in \mathcal{H} \setminus \{0\}} \frac{|Tf|}{\|f\|_\mathcal{H}} \quad \text{for} \quad T \in \mathcal{H}^*.$$

Now, let X be a set and consider the set of functions $\mathcal{H} \subseteq \mathbb{F}^X$. Define the *evaluation functionals*

$$E_x : \begin{cases} \mathcal{H} \to \mathbb{F} \\ f \mapsto f(x), \end{cases} \qquad x \in X.$$

If $\mathcal{H}$ together with $\langle , \rangle_{\mathcal{H}}$ is a Hilbert space and all evaluation functionals are bounded, we call $\mathcal{H}$ a *reproducing kernel Hilbert space (RKHS)* on X.

For every $x \in X$, the evaluation functional E_x is linear and bounded, hence they are all continuous. Therefore, by Riesz' representation theorem,

$$\exists! k_x \in \mathcal{H} : \quad \forall f \in \mathcal{H} : \quad E_x(f) = \langle f, k_x \rangle_{\mathcal{H}}, \quad \|k_x\|_{\mathcal{H}} = \|E_x\|_{\mathcal{H}^*}.$$

From this, we can construct the *reproducing kernel for* $\mathcal{H}$,

$$k : \begin{cases} X \times X \to \mathbb{F} \\ (x, y) \mapsto k_y(x). \end{cases}$$

By the Cauchy-Schwarz inequality, we have

$$|f(x)| = |\langle f, k(\cdot, x) \rangle_{\mathcal{H}}| \leq \|f\|_{\mathcal{H}} \|k(\cdot, x)\|_{\mathcal{H}}. \tag{2.10}$$

We say that the kernel function k is bounded by $K < \infty$ iff

$$\sup_{x \in X} \|k(\cdot, x)\|_{\mathcal{H}} \leq K.$$

We now want to list some important properties of RKHS with Propositions 2.3.1, 2.3.3, 2.3.4 and Theorem 2.3.5. The proofs can be found in any textbook on RKHS [29, 2, 6].

Proposition 2.3.1. *Let $\mathcal{H}$ be an RKHS on the set X with kernel k. Then $(k_x)_{x \in X}$ spans a dense sub space of $\mathcal{H}$, i.e.*

$$\overline{\operatorname{span}\{k(\cdot, x)\}_{x \in X}} = \mathcal{H}.$$

Remark 2.3.2. We can rewrite the kernel with the scalar product, yielding

$$k(x, y) = k_y(x) = \langle k_y, k_x \rangle_{\mathcal{H}} = \langle k(\cdot, y), k(\cdot, x) \rangle_{\mathcal{H}}. \tag{2.11}$$

Proposition 2.3.1 implies that we can represent a function $f \in \mathcal{H}$ as a countable sum

$$f(x) = \sum_{i=1}^{\infty} u_i k(x, x_i), \quad \text{where} \quad (u_i)_{i \in \mathbb{N}} \subset \mathbb{F}, \quad (x_i)_{i \in \mathbb{N}} \subset X.$$

Let $(v_i)_{i \in \mathbb{N}}$ be the coefficients of another $g \in \mathcal{H}$, w.l.o.g. with respect to the same $(x_i)_{i \in \mathbb{N}}$. Then the inner product of f and g can be represented as

$$\langle f, g \rangle_{\mathcal{H}} = \sum_{i,j=1}^{\infty} u_i v_j \langle k(\cdot, x_i), k(\cdot, x_j) \rangle = \sum_{i,j=1}^{\infty} u_i v_j k(x_i, x_j).$$

■

Call a function $k : X \times X \to \mathbb{F}$ positive semi-definite, if for every set $\{x_1, \ldots, x_n\} \subset X$ of n distinct elements, the matrix $(k(x_i, x_j))_{i,j=1}^n$ is positive semi-definite. Proposition 2.3.3 and 2.3.4 and Theorem 2.3.5 describe the relationship between these functions and RKHS.

Proposition 2.3.3. *Let $\mathcal{H}_1$ and $\mathcal{H}_2$ be RKHS on X with kernels k_1 and k_2, respectively. If $k_1 \equiv k_2$, then $\mathcal{H}_1 = \mathcal{H}_2$ and $\| \cdot \|_{\mathcal{H}_1} \equiv \| \cdot \|_{\mathcal{H}_2}$.*

Proposition 2.3.4. *Let $\mathcal{H}$ be an RKHS on X with kernel k. Then k is positive semi-definite.*

Theorem 2.3.5. *Let X be a set and $k : X \times X \to \mathbb{F}$ a positive semi-definite function. Then there exists an RKHS $\mathcal{H}$ on X with k as its reproducing kernel.*

2.4 Matrix Theory

2.4.1 Perron-Frobenius Theorem

Since one may view a Markov Decision Process as a graph, which is traversed according to the initial state distribution, transition dynamics and policy, we will occasionally require the well known Perron-Frobenius Theorem 2.4.1.

For a matrix $A = (a_{i,j})_{i,j=1}^n \in \mathbb{R}^{n\times n}$, consider the graph $G = (V, E)$, where $V = \{1, \dots, n\}$ and $(i, j) \in E$ iff $a_{i,j} \neq 0$. We call A *irreducible* iff G is *strongly connected*, i.e., one can reach any vertex $j \in V$ from any other vertex $i \in V$ by just traveling along edges in E in the forwards direction.

Theorem 2.4.1 (Perron-Frobenius theorem)**.** *Let $A \in \mathbb{R}^{n\times n}$ be irreducible with non-negative components. Then the spectral radius r of A is a positive eigenvalue of A, the* Perron-Frobenius eigenvalue*. It is also simple, i.e., both left- and right-eigenspaces of r are one dimensional. Additionally, there exist a left-eigenvector v and right-eigenvector w for r whose components are all non-negative. If A even has positive components, then all eigenvalues $\lambda \neq r$ of A have $|\lambda| < r$.*

2.4.2 Derivatives

In order to identify stationary points or to apply *stochastic gradient (SG)* methods, it is necessary to calculate a gradient of some objective function. In order to facilitate the calculation of multivariate derivatives, we present a series of elementary identities.

Remark 2.4.2. For $x, y \in \mathbb{R}^n$, we get partial derivatives

$$\frac{\partial}{\partial x_k}\langle x, y\rangle = \frac{\partial}{\partial x_k}\sum_{i=1}^n x_i y_i = y_k \quad \text{and} \quad \partial_x\langle x, y\rangle = y^\top.$$

For $A \in \mathbb{R}^{n\times m}$ with rows $a_1^\top, \dots, a_n^\top$, we therefore have

$$\partial_x Ax = \partial_x(\langle a_i, x\rangle)_{i=1}^n = (\partial_x\langle a_i, x\rangle)_{i=1}^n = (a_i^\top)_{i=1}^n = A.$$

Now, let $A \in \mathbb{R}^{n\times n}$ and

$$g : \mathbb{R}^n \to \mathbb{R}^{2n}, \quad g(x) \doteq \begin{pmatrix} Ax \\ x \end{pmatrix} = \begin{pmatrix} A \\ I \end{pmatrix} x \quad \text{and} \quad f : \mathbb{R}^{2n} \to \mathbb{R}, \quad f(y_1, y_2) \doteq \langle y_1, y_2\rangle.$$

We further get the derivatives

$$\partial_y f(y_1, y_2) = \begin{pmatrix} \partial_{y_1}\langle y_1, y_2\rangle & \partial_{y_2}\langle y_1, y_2\rangle \end{pmatrix} = \begin{pmatrix} y_2^\top & y_1^\top \end{pmatrix} \quad \text{and} \quad \partial_x g(x) = \begin{pmatrix} A \\ I \end{pmatrix}.$$

Therefore, according to the chain rule,

$$\begin{aligned} \partial_x\langle Ax, x\rangle = \partial_x(f \circ g)(x) &= \partial_y f(g(x))\partial_x g(x) \\ &= \begin{pmatrix} x^\top & (Ax)^\top \end{pmatrix}\begin{pmatrix} A \\ I \end{pmatrix} = x^\top A + x^\top A^\top = x^\top(A + A^\top) \end{aligned}$$

In particular, if A is symmetric, i.e., $A = A^\top$, we get

$$\partial_x \|x\|_A^2 = \partial_x\langle Ax, x\rangle = 2x^\top A.$$

This means that for $d \in \mathbb{R}^n$,

$$\partial_x \|x\|^2_{\ell_2(d)} = \partial_x \langle \operatorname{diag}(d)x, x\rangle = 2x^\top \operatorname{diag}(d) \quad \text{and} \quad \partial_x \|x\|_2^2 = 2x^\top.$$

■

Remark 2.4.3. Let a matrix $A \in \mathbb{R}^{n\times n}$ be positive definite. It is injective, because for any $x \in \mathbb{R}^n$,

$$Ax = 0 \implies x^\top Ax = 0 \implies x = 0.$$

The latter follows from positive definiteness. Since matrices represent linear operators between finite dimensional vector spaces, A is also invertible.

■

Remark 2.4.4. Consider a matrix $A \in \mathbb{R}^{m\times n}$, a vector $b \in \mathbb{R}^m$, and a symmetric, positive definite matrix $D \in \mathbb{R}^{m\times m}$. Let A^+ be the pseudo-inverse of A. It is commonly known that A^+b solves the

$$\text{linear fitting problem:} \quad \min_{x\in\mathbb{R}^n} \|Ax - b\|_2^2 \quad \text{or equivalently,} \tag{2.12}$$

$$\text{Gaussian normal equations:} \quad A^\top Ax = A^\top b. \tag{2.13}$$

We assume that the columns of A are linearly independent. Then, $A^\top A$ is positive definite, hence, by Remark 2.4.3, invertible. The solution to (2.13) and (2.12) becomes unique and

$$A^+ = (A^\top A)^{-1} A^\top.$$

Now, we pose this minimization problem with respect to the norm $\|\cdot\|_D$. Building derivatives according to Remark 2.4.2, we get

$$\nabla_x \frac{1}{2}\|Ax - b\|_D^2 = A^\top D^\top (Ax - b) = A^\top DAx - A^\top Db.$$

The existence of the inverse of $A^\top DA$ can be argued analogously to before, just by replacing the norms. Setting the gradient to zero, yields

$$x = A_D^+ b, \quad \text{where} \quad A_D^+ \doteq (A^\top DA)^{-1} A^\top D.$$

We define

$$P \doteq AA_D^+ = A(A^\top DA)^{-1} A^\top D.$$

Now, P is a projection onto the range of A, since for any $y = Ax$,

$$Py = A(A^\top DA)^{-1} A^\top DAx = Ax = y.$$

It is even an orthogonal projection with respect to the scalar product induced by D, since it is self-adjoint,

$$\langle x, Py\rangle_D = \langle Dx, Py\rangle = \langle P^\top Dx, y\rangle = \langle D^\top A(A^\top DA)^{-\top} A^\top Dx, y\rangle = \langle DPx, y\rangle = \langle Px, y\rangle_D.$$

Therefore, the Hilbert space projection theorem yields

$$Pb = \arg\min \left\{ \|y - b\|_D^2 \;\middle|\; y \in \operatorname{ran}(A) \right\} = \arg\min \left\{ \|Ax - b\|_D^2 \;\middle|\; x \in \mathbb{R}^n \right\}.$$

■

2.4.3 Block Matrices

When applying SG methods on multiple parameterizations simultaneously, we can collect their parameters in a single vector. Reformulating the objective yields a block matrix. To work with the determinant and potential inverse of such a block matrix, we list the well known Lemma 2.4.5.

Lemma 2.4.5. *Let A, B, C and D be matrices, and consider a block matrix*

$$M = \begin{pmatrix} A & B \\ C & D \end{pmatrix}.$$

Define the Schur complements $M/A \doteq D - CA^{-1}B$ *and* $M/D \doteq A - BD^{-1}C$. *Then*

$$\det(M) = \det(A)\det(M/A) \tag{2.14}$$
$$= \det(D)\det(M/D). \tag{2.15}$$

In case A and M/A, or D and M/D are invertible, i.e., M is invertible,

$$M^{-1} = \begin{pmatrix} A^{-1} + A^{-1}B(M/A)^{-1}CA^{-1} & -A^{-1}B(M/A)^{-1} \\ -(M/A)^{-1}CA^{-1} & (M/A)^{-1} \end{pmatrix} \tag{2.16}$$
$$= \begin{pmatrix} (M/D)^{-1} & -(M/D)^{-1}BD^{-1} \\ -D^{-1}C(M/D)^{-1} & D^{-1} + D^{-1}C(M/D)^{-1}BD^{-1} \end{pmatrix}. \tag{2.17}$$

2.5 ODE Lemma

The ODE Lemma 2.5.7 serves as a powerful tool to prove convergence of SG methods [9].

In this section we will briefly cover some basic concepts from ODE theory to understand Lemma 2.5.7 better and help apply it.

Let $G \subseteq \mathbb{R}^n$ be a simply connected domain and $f : G \to \mathbb{R}^n$ be Lipschitz continuous. A point $y_0 \in G$ is called *equilibrium* iff $f(y_0) = 0$. Define the set of all equilibrium points

$$\mathcal{E} \doteq \{y_0 \in G \mid f(y_0) = 0\}.$$

Denote by y_{y_0} the solution to the autonomous ODE $y'(t) = f(y(t))$ with $y(0) = y_0$. It is called *asymptotically stable* if it is *stable* and *attractive*, which means, respectively,

$$\forall \epsilon > 0\ \exists \delta > 0 \text{ s.t. } \forall \tilde{y_0} \in B_\delta(y_0):\ y_{y_0} \text{ exists on } [0,\infty) \text{ and } \|y_{y_0}(t) - y_{\tilde{y_0}}(t)\| < \epsilon \text{ for all } t \geq 0,$$
$$\text{and} \quad \exists \delta > 0 \text{ s.t. } \forall \tilde{y_0} \in B_\delta(y_0):\ y_{y_0} \text{ exists on } [0,\infty) \text{ and } \|y_{y_0}(t) - y_{\tilde{y_0}}(t)\| \xrightarrow{t\to\infty} 0.$$

Theorem 2.5.1. *Consider the function $f(y) = Ay$, where $A \in \mathbb{R}^{n\times n}$. Then $y^* = \vec{0}$ is a stable equilibrium iff all eigenvalues λ of A fulfill $\Re(\lambda) \leq 0$ and if $\Re(\lambda) = 0$, then λ is semi-simple, i.e., its algebraic and geometric multiplicity coincide. Also, $y^* = \vec{0}$ is an asymptotically stable equilibrium iff all eigenvalues λ of A fulfill $\Re(\lambda) < 0$.*

A function $L : G \to \mathbb{R}$ is called *(strict) Ljapunov function* for f iff it is continuously differentiable and for every y_0, where y_{y_0} is not constant, $L \circ y_{y_0}$ is (strictly) monotonically decreasing.

Lemma 2.5.2. *Consider a continuously differentiable function $L : G \to \mathbb{R}$. It is a Ljapunov-function or even a strict one, respectively, if*

$$\forall y \in G:\ \langle \nabla L(y), f(y)\rangle \leq 0 \quad \textit{and} \quad \forall y \in G \setminus \mathcal{E}:\ \langle \nabla L(y), f(y)\rangle < 0.$$

Theorem 2.5.3 (Ljapunov's method)**.** *Let $L : G \to \mathbb{R}^n$ be a Ljapunov function and y^* an equilibrium. If y^* is a strict minimum of V, then y^* is stable. If y^* is also isolated in $\mathcal{E}$ and V is a strict Ljapunov function, then y^* is asymptotically stable.*

Consider the iteration

$$y_{t+1} \doteq y_t + \alpha_t(h(y_t) + M_{t+1}), \quad \text{where} \quad t \in \mathbb{N},$$

for $y_t \in \mathbb{R}^n$, step sizes $\alpha_t \in (0, 1]$, random vectors $(M_t)_{t\in\mathbb{R}}$, and the function $h : \mathbb{R}^n \to \mathbb{R}^n$.

Assumption 2.5.4 (Lipschitz and ODEs)**.** The function h is Lipschitz continuous and there exists a function $h_\infty : \mathbb{R}^n \to \mathbb{R}^n$ such that

$$\lim_{r\to\infty} \frac{h(ry)}{r} = h_\infty(y) \quad \text{for all} \quad y \in \mathbb{R}^n.$$

Furthermore, let the origin $\vec{0} \in \mathbb{R}^n$ be an asymptotically stable equilibrium for the ODE

$$y'(t) = h_\infty(y(t)) \quad \text{for all} \quad t \geq 0$$

and let $y^* \in \mathbb{R}^n$ be the unique globally asymptotically stable equilibrium for the ODE

$$y'(t) = h(y(t)) \quad \text{and} \quad t \geq 0.$$

Assumption 2.5.5 (bounded martingale difference sequence)**.** The sequence $(M_t)_{t\in\mathbb{N}}$ is a martingale difference sequence with respect to the filtration $\mathcal{F}_t \doteq \sigma(y_i, M_i)_{i=1}^t$ and for any initial condition $y_0 \in \mathbb{R}$,

$$\mathbb{E}[\|M_{t+1}\|_2^2 \mid \mathcal{F}_t] = \mathcal{O}(\|y_t\|_2^2 + 1).$$

Assumption 2.5.6 (Robbins-Monro)**.** The Robbins-Monro conditions for $(\alpha_t)_{t\in\mathbb{N}} \subset \mathbb{R}_{>0}$ read

$$\sum_{t=1}^{\infty} \alpha_t = \infty \quad \text{and} \quad \sum_{t=1}^{\infty} \alpha_t^2 < \infty. \tag{2.18}$$

Now, we can finally formulate the ODE Lemma 2.5.7 by Borkar et al. [9].

Theorem 2.5.7 (ODE Lemma)**.** *Under Assumptions 2.5.4, 2.5.5 and 2.5.6, for any initial condition $y_0 \in \mathbb{R}^n$, we have*

$$y_t \xrightarrow{t\to\infty} y^* \quad \textit{almost surely.}$$

2.6 Extended Delta Method

2.6.1 Hadamard Differentiability

The objective of this section is to present a more abstract notion of differentiability, which will enable us to apply an extended version of the well-known Delta Method to functionals $T : \mathcal{P}(X) \subset \mathcal{M}(X) \to \mathbb{R}$. The content is based on van der Vaart and Wellner [40].

Let $\mathcal{X}$ and $\mathcal{Y}$ be normed spaces over $\mathbb{R}$. A function $f : \mathcal{X}_f \subseteq \mathcal{X} \to \mathcal{Y}$ is called *Hadamard differentiable* at $x_0 \in \mathcal{X}_f$ tangentially to $\mathcal{X}_{\partial f_{x_0}} \subseteq \mathcal{X}$ with *Hadamard derivative* $\partial f_{x_0} : \mathcal{X}_{\partial f_{x_0}} \to \mathcal{Y}$

at x_0 iff ∂f_{x_0} is continuous and linear and for all $x \in \mathcal{X}_{\partial f_{x_0}}$ and series $(t_n)_{n\in\mathbb{N}} \subset \mathbb{R}\setminus\{0\}$ and $(x_n)_{n\in\mathbb{N}} \subset \mathcal{X}$, where

$$t_n \xrightarrow{n\to\infty} 0, \quad \|x_n - x\|_{\mathcal{X}} \xrightarrow{n\to\infty} 0, \quad \text{and} \quad x_0 + t_n x_n \in \mathcal{X}_f,$$

we have

$$\frac{f(x_0 + t_n x_n) - f(x_0)}{t_n} \xrightarrow{n\to\infty} \partial f_{x_0}(x).$$

In order to highlight the difference between classically and Hadamard differentiable functions, we aim to prove Lemmas 2.6.1 and 2.6.2. To this end, for a sub set $K \subseteq \mathcal{X}_{\partial f_{x_0}}$, define

$$r_K(t) \doteq \sup\left\{ \left\| \frac{f(x_0 + tx) - f(x_0)}{t} - \partial f_{x_0}(x) \right\|_{\mathcal{Y}} \;\middle|\; x \in K,\ x_0 + tx \in \mathcal{X}_f \right\}.$$

Lemma 2.6.1. *Hadamard differentiability is equivalent to*

$$\forall K \subset \mathcal{X}_{\partial f_{x_0}} \text{ compact}: \ r_K(t) \xrightarrow{t\to 0} 0. \tag{2.19}$$

Proof. Firstly, we assume that f is Hadamard differentiable and show (2.19). We show that (2.19) holds for any $(t_n)_{n\in\mathbb{N}} \to 0$. For every $n \in \mathbb{N}$, consider $(x_{n,m})_{m\in\mathbb{N}}$, with $x_{n,m} \in K$ and $x_0 + t_n x_{n,m} \in \mathcal{X}_f$, such that

$$\left\| \frac{f(x_0 + t_n x_{n,m}) - f(x_0)}{t_n} - \partial f_{x_0}(x_{n,m}) \right\|_{\mathcal{Y}} \xrightarrow{m\to\infty} r_K(t_n).$$

Let $(\epsilon_n)_{n\in\mathbb{N}} \downarrow 0$ and choose $(m_n)_{n\in\mathbb{N}}$, such that

$$r_K(t_n) \le \left\| \frac{f(x_0 + t_n x_{n,m}) - f(x_0)}{t_n} - \partial f_{x_0}(x_{n,m}) \right\|_{\mathcal{Y}} + \epsilon_n \quad \text{for all} \quad n \in \mathbb{N} \quad \text{and} \quad m \ge m_n.$$

Because K is compact, there exists a sub sequence $(x_n)_{n\in\mathbb{N}}$ of $(x_{n,m_n})_{n\in\mathbb{N}} \subset K$, converging against some $x \in K \subset \mathcal{X}_{\partial f_{x_0}}$. We can now apply the triangle inequality and the continuity of ∂f_{x_0}, to get

$$\begin{aligned} r_K(t_n) &\le \left\| \frac{f(x_0 + t_n x_n) - f(x_0)}{t_n} - \partial f_{x_0}(x_n) \right\|_{\mathcal{Y}} + \epsilon_n \\ &\le \left\| \frac{f(x_0 + t_n x_n) - f(x_0)}{t_n} - \partial f_{x_0}(x) \right\|_{\mathcal{Y}} + \|\partial f_{x_0}(x_n) - \partial f_{x_0}(x)\|_{\mathcal{Y}} + \epsilon_n \xrightarrow{n\to\infty} 0. \end{aligned}$$

Secondly, we assume (2.19) and show Hadamard differentiability. Let $K \subset \mathcal{X}_{\partial f_{x_0}}$ be compact, such that $x \in K$ and $x_n \in K$ for almost every $n \in \mathbb{N}$. Again,

$$\begin{aligned} &\left\| \frac{f(x_0 + t_n x_n) - f(x_0)}{t_n} - \partial f_{x_0}(x) \right\|_{\mathcal{Y}} \\ &\quad \le \|\partial f_{x_0}(x_n) - \partial f_{x_0}(x)\|_{\mathcal{Y}} + \left\| \frac{f(x_0 + t_n x_n) - f(x_0)}{t_n} - \partial f_{x_0}(x_n) \right\|_{\mathcal{Y}} \\ &\quad \le \|\partial f_{x_0}(x_n) - \partial f_{x_0}(x)\|_{\mathcal{Y}} + r_K(t_n) \xrightarrow{n\to\infty} 0. \end{aligned}$$

□

Lemma 2.6.2. *Frechét differentiability is equivalent to*

$$\forall K \subset \mathcal{X}_{\partial f_{x_0}} = \mathcal{X} \text{ bounded}: \; r_K(t) \xrightarrow{t\to 0} 0. \tag{2.20}$$

Proof. Recall that f is Frechét differentiable in x_0 iff

$$\|f(x_0+x) - f(x_0) - \partial f_{x_0}(\xi)\|_{\mathcal{Y}} = o(\|\xi\|_{\mathcal{X}}), \quad \|\xi\|_{\mathcal{X}} \to 0, \quad x_0+\xi \in \mathcal{X}_f.$$

This means that for all $C > 0$ there exists an $\epsilon > 0$, such that for all $\xi \in B_\epsilon(0)$ with $x_0+\xi \in \mathcal{X}_f$, we have

$$\|f(x_0+\xi) - f(x_0) - \partial f_{x_0}(\xi)\|_{\mathcal{X}} \leq C\|\xi\|_{\mathcal{X}}.$$

Since we can bound the supremum $r_K(t)$ in (2.20) by taking $B_\varepsilon^{\mathcal{X}}(0) \supseteq K$ with $\varepsilon = \sup_{x\in K}\|x\|_{\mathcal{X}}$, instead of K, we can assure w.l.o.g. that K is of such form. It follows that for all bounded $K \subseteq X$ and $\delta > 0$ there exists a $t_0 > 0$ such that for all $t \in (0, t_0)$ and $x \in K$ with $x_0 + tx \in \mathcal{X}_f$, we have

$$\left\| \frac{f(x_0+tx) - f(x_0)}{t} - \partial f_{x_0}(x) \right\|_{\mathcal{X}} \leq \delta,$$

which we can see by considering the substitution, w.l.o.g. $K \neq \{0\}$,

$$C = \delta / \sup_{x\in K}\|x\|_{\mathcal{X}}, \quad t_0 = \epsilon / \sup_{x\in K}\|x\|_{\mathcal{X}}, \quad \text{and} \quad \xi = tx.$$

The other direction follows by taking the same substitution, but with the unit ball $K \doteq B_1^{\mathcal{X}}(0)$. □

Corollary 2.6.3. *Every Frechét differentiable function $f : \mathcal{X}_f \subseteq \mathcal{X} \to \mathcal{Y}$ is also Hadamard differentiable tangentially to $\mathcal{X}_{\partial f_{x_0}} = \mathcal{X}$. The reverse is true if the unit ball $B_1^{\mathcal{X}}(0)$ in $\mathcal{X}$ is compact with respect to the norm topology, i.e., $\mathcal{X}$ is finite dimensional.*

Proof. We consider the characterizations from Lemmas 2.6.2 and 2.6.1. Since every compact set $K \subseteq \mathcal{X}$ is bounded, the first claim is immediate. If $B_1^{\mathcal{X}}(0)$ is compact and $K \subseteq \mathcal{X}$ is bounded, $B_\varepsilon^{\mathcal{X}}(0) \supseteq K$ with $\varepsilon = \sup_{x\in K}\|x\|_{\mathcal{X}}$ is compact and we can bound the supremum $r_K(t)$ in (2.20) by using the $B_\varepsilon^{\mathcal{X}}(0)$ instead of K.

□

Note that the normed space $\mathcal{M}(X)$ is not finite dimensional. By Corollary 2.6.3, the notion of Frechét differentiability is not powerful enough to support our original endeavor. Thus, we extend the well known Delta Method for classically differentiable functions, to Hadamard differentiable functions. For further details, please refer to Vaart and Wellner [40].

Theorem 2.6.4 (Informal Extended Delta Method)**.** *Let $f : \mathcal{X}_f \subseteq \mathcal{X} \to \mathcal{Y}$ be Hadamard differentiable at θ tangentially to $\mathcal{X}_{\partial f_{x_0}}$. Let $(r_n)_{n\in\mathbb{N}} > 0$ be a sequence with $r_n \xrightarrow{n\to\infty} \infty$. Let $(X_n)_{n\in\mathbb{N}}$ be a sequence of random variables and X a single random variable into $\mathcal{X}_f$ and $\mathcal{X}_{\partial f_{x_0}}$, respectively, such that*

$$r_n(X_n - \theta) \xrightarrow{n\to\infty} X \quad \text{almost surely.}$$

Then

$$r_n(f(X_n) - f(\theta)) = \frac{f(\theta + r_n^{-1} r_n(X_n - \theta)) - f(\theta)}{r_n^{-1}} \xrightarrow{n\to\infty} \partial f_\theta(X) \quad \text{in distribution.}$$

If ∂f_θ is defined and continuous on all of $\mathcal{X}$, then

$$r_n(f(X_n) - f(\theta)) - \partial f_\theta(r_n(X_n - \theta)) \xrightarrow{n\to\infty} 0 \quad \text{in probability.}$$

2.6.2 Confidence Intervals

In this section, we apply the Extended Delta Method 2.6.4 to obtain some asymptotic confidence intervals, as described in Theorem 2.6.8. The results are taken from the work of Duchi et al. [14].

Call $T^{(1)} : X \times \mathcal{P}(X) \to \mathbb{R}$ an *influence function* of T iff we have $\mathbb{E}_{P^{\mathcal{D}}}[T^{(1)}(\cdot; P^{\mathcal{D}})] = 0$ and

$$\partial T_{P^{\mathcal{D}}}(Q - P^{\mathcal{D}}) = \int_X T^{(1)}(x; P^{\mathcal{D}}) \, \mathrm{d}(Q - P^{\mathcal{D}})(x) \quad \text{for all} \quad Q \in \mathcal{M}(X).$$

The law $\mathbb{P}_G$ of a Borel-measurable random variable G is called *tight* iff

$$\forall \epsilon > 0 : \ \exists K \text{ compact} : \ \mathbb{P}_G(K) \geq 1 - \epsilon.$$

Recall that, since the samples in $\mathcal{D}$ are random variables on X, the empirical measure $\hat{P}^{\mathcal{D}}$ is a random measure. Let $\mathcal{H} \subset L_2(P^{\mathcal{D}})$ be a class of functions. We say that $\mathcal{H}$ is $P^{\mathcal{D}}$*-Donsker* iff there exists some tight Borel-measurable $G \in L_\infty(\mathcal{H})$, such that

$$\left\| \sqrt{n} \left(\hat{P}^{\mathcal{D}} - P^{\mathcal{D}} \right) - G \right\|_{L_\infty(\mathcal{H})} \xrightarrow{n \to \infty} 0 \quad \text{in probability.}$$

We say that $\mathcal{H}$ has an L_2-integrable *envelope* $C : X \to \mathbb{R}_{\geq 0}$ iff

$$C \in L_2(P^{\mathcal{D}}) \quad \text{and} \quad \forall \ell \in \mathcal{H} : \ \ell \leq C \quad \text{almost surely.}$$

Lemma 2.6.5. *Let $\mathcal{H} \doteq \{\ell(\cdot; \theta)\}_{\theta \in \mathcal{F}_\theta}$ be a set of functions, C_ℓ-Lipschitz continuous in θ, with $C_\ell \in L_2(P^{\mathcal{D}})$ and compact $\mathcal{F}_\theta$. Then $\mathcal{H}$ is $P^{\mathcal{D}}$-Donsker with L_2-integrable envelope.*

Assumption 2.6.6 (Smoothness of f-divergence)**.** The function $f : \mathbb{R}_{\geq 0} \to \mathbb{R} \cup \{\infty\}$ divergence conform, i.e.,

$$\forall t > 0 : \ f(t) < \infty, \quad f(0) = \lim_{t \downarrow 0} f(t) \quad \text{and} \quad f(1) = 0.$$

It is also three times differentiable in a neighborhood of 1 as well as

$$f'(1) = 0 \quad \text{and} \quad f''(1) = 2.$$

Remark 2.6.7. Notice that Assumption 2.6.6 restricts the set of f-divergences we can use severely. By differentiating $x \log x$ twice, one can easily see that the KL divergence does not meet the requirements imposed by Assumption 2.6.6,

$$\frac{\mathrm{d}}{\mathrm{d}x}\big(x \log x\big) = \log x + x\frac{1}{x} = \log x + 1 \quad \text{and} \quad \frac{\mathrm{d}^2}{\mathrm{d}x^2}\big(x \log x\big) = \frac{1}{x}. \tag{2.21}$$

Nonetheless, by starting with an arbitrary $f : \mathbb{R}_{\geq 0} \to \mathbb{R} \cup \{\infty\}$, which is divergence conform and three times differentiable in a neighborhood of 1, we can enforce the missing conditions by instead taking

$$\frac{2}{f''(1)} \big(f(x) - f'(1)(x - 1) \big). \tag{2.22}$$

For the KL divergence, the modified version would then be based on

$$f(x) \doteq 2x \log x - 2(x - 1). \tag{2.23}$$

■

Theorem 2.6.8 (General Asymptotic Coverage)**.** *Let Assumption 2.6.6 hold for f. Let $\mathcal{H}$ be a $P^{\mathcal{D}}$-Donsker class of functions with L_2-integrable envelope. Let the limit G of $\sqrt{n}(\hat{P}^{\mathcal{D}} - P^{\mathcal{D}})$ take values inside $B(\mathcal{H}, P^{\mathcal{D}}) \subset \mathcal{M}(X;\mathcal{H})$. Assume that $T : \mathcal{P}(X) \to \mathbb{R}$ is Hadamard differentiable at $P^{\mathcal{D}}$ tangentially to $B(\mathcal{H}, P^{\mathcal{D}})$ with influence function $T^{(1)}(\cdot; P^{\mathcal{D}})$ and $\partial T_{P^{\mathcal{D}}}$ is continuous and defined on the whole of $\mathcal{M}(X;\mathcal{H})$. If $0 < \operatorname{Var}_{P^{\mathcal{D}}}[T^{(1)}(\cdot; P^{\mathcal{D}})] < \infty$, then*

$$\lim_{n\to\infty} \mathbb{P}\left(T(P^{\mathcal{D}}) \in \left\{T(P) \;\middle|\; P \in B^f_{\xi/n}(\hat{P}^{\mathcal{D}})\right\}\right) = \mathbb{P}(\chi^2_1 \le \xi).$$

Proof sketch. We want to apply the extended delta method from Theorem 2.6.4 to

$$f = T, \quad r_n = \sqrt{n}, \quad X_n = \hat{P}^{\mathcal{D}}, \quad \theta = P^{\mathcal{D}} \quad \text{and} \quad X = G.$$

We can rewrite

$$\begin{aligned} T(\hat{P}^{\mathcal{D}}) &= T(P^{\mathcal{D}}) + \mathbb{E}_{\hat{P}^{\mathcal{D}}}[T^{(1)}(\cdot; P^{\mathcal{D}})] + \kappa(\hat{P}^{\mathcal{D}}), \\ &\text{where} \quad \kappa(P) \doteq T(P) - T(P^{\mathcal{D}}) - \mathbb{E}_P[T^{(1)}(\cdot; P^{\mathcal{D}})]. \end{aligned}$$

Because

$$\partial T_{P^{\mathcal{D}}}(\hat{P}^{\mathcal{D}} - P^{\mathcal{D}}) = \int_X T^{(1)}(x; P^{\mathcal{D}}) \, \mathrm{d}(\hat{P}^{\mathcal{D}} - P^{\mathcal{D}})(x) = \mathbb{E}_{\hat{P}^{\mathcal{D}}}[T^{(1)}(\cdot; P^{\mathcal{D}})] - \underbrace{\mathbb{E}_{P^{\mathcal{D}}}[T^{(1)}(\cdot; P^{\mathcal{D}})]}_{0},$$

we can apply the Extended Delta-Method from Theorem 2.6.4 and get

$$\begin{aligned} \sqrt{n}\kappa(\hat{P}^{\mathcal{D}}) &= \sqrt{n}\left(T(\hat{P}^{\mathcal{D}}) - T(P^{\mathcal{D}}) - \mathbb{E}_{\hat{P}^{\mathcal{D}}}[T^{(1)}(\cdot; P^{\mathcal{D}})]\right) \\ &= \sqrt{n}\left(T(\hat{P}^{\mathcal{D}}) - T(P^{\mathcal{D}}) - \partial T_{P^{\mathcal{D}}}(\sqrt{n}(\hat{P}^{\mathcal{D}} - P^{\mathcal{D}})\right) \xrightarrow{n\to\infty} 0 \quad \text{in probability.} \end{aligned}$$

By [14, Lemma 16], this convergence even holds uniformly over $B^f_{\xi/n}(\hat{P}^{\mathcal{D}})$, i.e.,

$$\forall \epsilon > 0 : \quad \limsup_{n\to\infty} \mathbb{P}\left(\sqrt{n} \sup\left\{|\kappa(P)| \;\middle|\; P \in B^f_{\xi/n}(\hat{P}^{\mathcal{D}})\right\} \ge \epsilon\right) = 0.$$

By definition of κ,

$$\begin{aligned} \sup_{P\in B^f_{\xi/n}(\hat{P}^{\mathcal{D}})} (T(P) - T(P^{\mathcal{D}})) &\le \sup_{P\in B^f_{\xi/n}(\hat{P}^{\mathcal{D}})} \kappa(P) + \sup_{P\in B^f_{\xi/n}(\hat{P}^{\mathcal{D}})} \mathbb{E}_P[T^{(1)}(\cdot; P^{\mathcal{D}})], \\ \sup_{P\in B^f_{\xi/n}(\hat{P}^{\mathcal{D}})} \mathbb{E}_P[T^{(1)}(\cdot; P^{\mathcal{D}})] &\le \sup_{P\in B^f_{\xi/n}(\hat{P}^{\mathcal{D}})} (-\kappa(P)) + \sup_{P\in B^f_{\xi/n}(\hat{P}^{\mathcal{D}})} (T(P) - T(P^{\mathcal{D}})). \end{aligned}$$

Applying uniform convergence, we obtain

$$\begin{aligned} &\left|\sqrt{n} \sup_{P\in B^f_{\xi/n}(\hat{P}^{\mathcal{D}})} \left(T(P) - T(P^{\mathcal{D}})\right) - \sqrt{n} \sup_{P\in B^f_{\xi/n}(\hat{P}^{\mathcal{D}})} \mathbb{E}_P[T^{(1)}(\cdot; P^{\mathcal{D}})]\right| \\ &\qquad \le \sqrt{n} \sup_{P\in B^f_{\xi/n}(\hat{P}^{\mathcal{D}})} |\kappa(P)| \xrightarrow{n\to\infty} 0 \quad \text{in probability.} \end{aligned}$$

By [14, Theorem 9],

$$\begin{aligned} &\sqrt{n} \sup_{P\in B^f_{\xi/n}(\hat{P}^{\mathcal{D}})} \mathbb{E}_P[T^{(1)}(\cdot; P^{\mathcal{D}})] - \left(\sqrt{n}\mathbb{E}_{\hat{P}^{\mathcal{D}}}[T^{(1)}(\cdot; P^{\mathcal{D}})] + \sqrt{\xi \operatorname{Var}_{\hat{P}^{\mathcal{D}}}[T^{(1)}(\cdot; P^{\mathcal{D}})]}\right) \\ &\qquad \xrightarrow{n\to\infty} 0 \quad \text{in probability.} \end{aligned}$$

By the Central Limit Theorem,

$$\sqrt{n}\big(\mathbb{E}_{\hat{P}^{\mathcal{D}}}[T^{(1)}(\cdot;P^{\mathcal{D}})] - \underbrace{\mathbb{E}_{P^{\mathcal{D}}}[T^{(1)}(\cdot;P^{\mathcal{D}})]}_{0}\big) \xrightarrow{n\to\infty} N\left(0, \mathrm{Var}_{P^{\mathcal{D}}}[T^{(1)}(\cdot;P^{\mathcal{D}})]\right).$$

The sample variance is a consistent estimator, i.e.,

$$\mathrm{Var}_{\hat{P}^{\mathcal{D}}}[T^{(1)}(\cdot;P^{\mathcal{D}})] \xrightarrow{n\to\infty} \mathrm{Var}_{P^{\mathcal{D}}}[T^{(1)}(\cdot;P^{\mathcal{D}})].$$

Putting it all together yields

$$\begin{aligned}
\mathbb{P}\left(T(P^{\mathcal{D}}) \le \sup_{P\in B^f_{\xi/n}(\hat{P}^{\mathcal{D}})} T(P)\right) &= \mathbb{P}\left(0 \le \sqrt{n} \sup_{P\in B^f_{\xi/n}(\hat{P}^{\mathcal{D}})} \big(T(P) - T(P^{\mathcal{D}})\big)\right) \\
&\xrightarrow{n\to\infty} \mathbb{P}\left(0 \le \sqrt{\xi\,\mathrm{Var}_{P^{\mathcal{D}}}[T^{(1)}(\cdot;P^{\mathcal{D}})]} + N\left(0, \mathrm{Var}_{P^{\mathcal{D}}}[T^{(1)}(\cdot;P^{\mathcal{D}})]\right)\right) \\
&= \mathbb{P}\left(-\sqrt{\xi} \le N(0,1)\right).
\end{aligned}$$

By a symmetric argument on $-T(P^{\mathcal{D}})$, we get

$$\mathbb{P}\left(\inf_{P\in B^f_{\xi/n}(\hat{P}^{\mathcal{D}})} T(P) \le T(P^{\mathcal{D}})\right) \xrightarrow{n\to\infty} \mathbb{P}\left(N(0,1) \le \sqrt{\xi}\right).$$

□

3 Reinforcement Learning

3.1 General

For a set X, let $\Delta(X)$ be the set of probability distributions on X. We consider a *Markov Decision Process (MDP)* $(S, A, R, T, d_0, \gamma)$.

- S and A are the set of *states* and *actions*, respectively.
- $R : S \times A \times S \to \Delta(\mathbb{R})$ is the *reward function* that assigns a state-action pair a reward distribution, but we are often going to treat $R(s, a, s')$ as a random variable and write $r(s, a) \doteq \mathbb{E}_{s' \sim T(s,a)}[R(s, a, s')]$. We assume that the reward function is bounded almost surely.
- The *transition probability* distributions are given via $T : S \times A \to \Delta(S)$ and we write $T(s' \mid s, a)$ for the probability of transitioning into state $s' \in S$ when choosing action $a \in A$ in state $s \in S$.
- The *initial state distribution* is $d_0 \in \Delta(S)$.
- The *discount factor* $\gamma \in (0, 1]$, that is, we allow for our MDP to be discounted or even undiscounted in certain cases.

For the sake of brevity, when we want to make claims for the state-space S and state-action-space $S \times A$ at the same time, we simply use $\Omega = S$ or $\Omega = S \times A$. In case the state-action-space $S \times A$ is finite, we fix some global enumeration. We write the column vector

$$\vec{f} \doteq (f(s, a))_{(s,a) \in S \times A} \quad \text{for} \quad f : S \times A \to \mathbb{R}.$$

Also, for any operator A on these functions, we write $\vec{A}$ for the operator on these column vectors. In particular, this means that if A is linear, $\vec{A}$ will be a matrix. Define the multiplication operator

$$Dw \doteq dw \quad \text{for} \quad w : S \times A \to \mathbb{R}, \quad \text{where} \quad d : S \times A \to \mathbb{R}.$$

Any sub or super scripts that d may have will get carried over to D. Note, that for a finite state-action-space $S \times A$, the matrix version of the operator D will be a diagonal matrix

$$\vec{D} = \operatorname{diag}(\vec{d}).$$

For convenience, we define

$$d_0^\pi(s_0, a_0) \doteq \pi(a_0 \mid s_0) d_0(s_0) \quad \text{and} \quad T^\pi(s', a' \mid s, a) \doteq \pi(a' \mid s') T(s' \mid s, a) \quad \text{for} \quad (s_0, a_0) \in S \times A \quad \text{and} \quad (s, a), (s', a') \in S \times A.$$

Assumption 3.1.1 (MDP ergodicity)**.** A finite MDP is said to be *ergodic* if for any policy, starting from any state, it is possible to reach any other state (including itself) within a finite number of steps with non-zero probability.

The *policy value (PV)* ρ^π of π is defined as

$$\rho^\pi \doteq \lim_{H\to\infty} \frac{1}{\sum_{t=0}^{H}\gamma^t}\mathbb{E}_{(s_0,a_0)\sim d_0^\pi,\ (s_{t+1},a_{t+1})\sim T^\pi(s_t,a_t)}\left[\sum_{t=0}^{H}\gamma^t R(s_t,a_t)\right].$$

In the discounted case, the policy value ρ^π reduces to

$$\rho^{\pi,\gamma} \doteq (1-\gamma)\mathbb{E}_{(s_0,a_0)\sim d_0^\pi,\ (s_{t+1},a_{t+1})\sim T^\pi(s_t,a_t)}\left[\sum_{t=0}^{\infty}\gamma^t r(s_t,a_t)\right],$$

where as the undiscounted case yields

$$\begin{aligned}\rho^{\pi,1} &\doteq \lim_{H\to\infty} \mathbb{E}_{(s_0,a_0)\sim d_0^\pi,\ (s_{t+1},a_{t+1})\sim T^\pi(s_t,a_t)}\left[\frac{1}{H}\sum_{t=0}^{H-1} r(s_t,a_t)\right]\\ &= \lim_{H\to\infty} \mathbb{E}_{(s_0,a_0)\sim d_0^\pi,\ (s_{t+1},a_{t+1})\sim T^\pi(s_t,a_t)}[r(s_H,a_H)].\end{aligned}$$

The *(state) value function* and *(state) action value function* of π are denoted V^π and Q^π, respectively. In the discounted case, the state- and action-value functions of π are

$$\begin{aligned}V^{\pi,\gamma}(s) &\doteq \mathbb{E}_{a_t\sim\pi(s_t),\ s_{t+1}\sim T(s_t,a_t)}\left[\sum_{t=0}^{\infty}\gamma^t r(s_t,a_t)\,\middle|\, s_0=s\right],\\ Q^{\pi,\gamma}(s,a) &\doteq \mathbb{E}_{(s_{t+1},a_{t+1})\sim T^\pi(s_t,a_t)}\left[\sum_{t=0}^{\infty}\gamma^t r(s_t,a_t)\,\middle|\, \begin{matrix}s_0=s,\\ a_0=a\end{matrix}\right],\end{aligned}$$

in the undiscounted case, however,

$$\begin{aligned}V^{\pi,1}(s) &\doteq \mathbb{E}_{\substack{a_t\sim\pi(s_t),\\ s_{t+1}\sim T(s_t,a_t)}}\left[\sum_{t=0}^{\infty}(r(s_t,a_t)-\rho^\pi)\,\middle|\, s_0=s\right],\\ Q^{\pi,1}(s,a) &\doteq \mathbb{E}_{(s_{t+1},a_{t+1})\sim T^\pi(s_t,a_t)}\left[\sum_{t=0}^{\infty}(r(s_t,a_t)-\rho^\pi)\,\middle|\, \begin{matrix}s_0=s,\\ a_0=a\end{matrix}\right].\end{aligned}$$

The *stationary distribution* d^π *of* π is defined as

$$\begin{aligned}d^\pi(s,a) &\doteq \lim_{H\to\infty} \frac{1}{\sum_{t=0}^{H}\gamma^t}\mathbb{E}_{(s_0,a_0)\sim d_0^\pi,\ (s_{t+1},a_{t+1})\sim T^\pi(s_t,a_t)}\left[\sum_{t=0}^{H}\gamma^t \mathbb{1}_{s_t=s,\ a_t=a}\right]\\ d^\pi(s) &\doteq \lim_{H\to\infty} \frac{1}{\sum_{t=0}^{H}\gamma^t}\mathbb{E}_{(s_0,a_0)\sim d_0^\pi,\ (s_{t+1},a_{t+1})\sim T^\pi(s_t,a_t)}\left[\sum_{t=0}^{H}\gamma^t \mathbb{1}_{s_t=s}\right]\end{aligned}$$

In the discounted case, the stationary distribution d^π of π reduces to

$$\begin{aligned}d^{\pi,\gamma}(s,a) &= (1-\gamma)\mathbb{E}_{(s_0,a_0)\sim d_0^\pi,\ (s_{t+1},a_{t+1})\sim T^\pi(s_t,a_t)}\left[\sum_{t=0}^{\infty}\gamma^t \mathbb{1}_{s_t=s,\ a_t=a}\right],\\ d^{\pi,\gamma}(s) &= (1-\gamma)\mathbb{E}_{(s_0,a_0)\sim d_0^\pi,\ (s_{t+1},a_{t+1})\sim T^\pi(s_t,a_t)}\left[\sum_{t=0}^{\infty}\gamma^t \mathbb{1}_{s_t=s}\right],\end{aligned}$$

where as the undiscounted case yields

$$\begin{aligned} d^{\pi,1}(s,a) &= \lim_{H\to\infty} \mathbb{E}_{(s_0,a_0)\sim d_0^\pi,\ (s_{t+1},a_{t+1})\sim T^\pi(s_t,a_t)} \left[\frac{1}{H}\sum_{t=0}^{H-1} \mathbb{1}_{s_t=s,\ a_t=a}\right] \\ &= \lim_{H\to\infty} \mathbb{P}_{(s_0,a_0)\sim d_0^\pi,\ (s_{t+1},a_{t+1})\sim T^\pi(s_t,a_t)}(s_H = s,\ a_H = a), \\ d^{\pi,1}(s) &= \lim_{H\to\infty} \mathbb{E}_{(s_0,a_0)\sim d_0^\pi,\ (s_{t+1},a_{t+1})\sim T^\pi(s_t,a_t)} \left[\frac{1}{H}\sum_{t=0}^{H-1} \mathbb{1}_{s_t=s}\right] \\ &= \lim_{H\to\infty} \mathbb{P}_{(s_0,a_0)\sim d_0^\pi,\ (s_{t+1},a_{t+1})\sim T^\pi(s_t,a_t)}(s_H = s). \end{aligned}$$

For simplicity, we abbreviate the statements

$$\begin{aligned} (s_0 = s,\ a_0 = a)_t &:\iff s_0 = s, \quad a_0 = a, \quad (s_t,a_t)\sim T^\pi(s_{t-1},a_{t-1}), \\ (s_0 = s)_t &:\iff s_0 = s, \quad a_0\sim\pi(s_0), \quad (s_t,a_t)\sim T^\pi(s_{t-1},a_{t-1}), \\ (*)_t^\pi &:\iff s_0\sim d_0, \quad a_0\sim\pi(s_0), \quad (s_t,a_t)\sim T^\pi(s_{t-1},a_{t-1}), \\ (*)_t &:\iff s_0\sim d_0, \quad a_0\sim\pi(s_0), \quad s_t\sim T(s_{t-1},a_{t-1}), \\ &\text{and everywhere} \quad \forall\tau = 1,\dots,t:\ (s_\tau,a_\tau)\sim T^\pi(s_{\tau-1},a_{\tau-1}). \end{aligned}$$

Now, we define the expected reward and occupancy at some time $t\in\mathbb{N}$ as

$$\begin{aligned} r_t^\pi(s,a) &\doteq \mathbb{E}_{(s_0=s,\ a_0=a)_t}[r(s_t,a_t)], & d_t^\pi(s,a) &\doteq \mathbb{P}_{(*)_t^\pi}(s_t = s,\ a_t = a), \\ r_t^\pi(s) &\doteq \mathbb{E}_{(s_0=s)_t}[r(s_t,a_t)], & d_t^\pi(s) &\doteq \mathbb{P}_{(*)_t}(s_t = s). \end{aligned}$$

Notice that

$$\begin{aligned} r_0^\pi(s,a) &= r(s,a), & d_0^\pi(s,a) &= d_0(s)\pi(a\mid s), \\ r_0^\pi(s) &= \mathbb{E}_{a\sim\pi(s)}[r(s,a)], & d_0^\pi(s) &= d_0(s). \end{aligned}$$

We can also recover the functions Q^π and V^π,

$$Q^\pi(s,a) = \sum_{t=0}^{\infty}\gamma^t r_t^\pi(s,a) \quad\text{and}\quad V^\pi(s) = \sum_{t=0}^{\infty}\gamma^t r_t^\pi(s), \tag{3.1}$$

as well as the stationary distribution

$$d^\pi = \lim_{H\to\infty}\sum_{t=0}^{H}\gamma^t d_t^\pi \Big/ \sum_{t=0}^{H}\gamma^t. \tag{3.2}$$

The discounted and undiscounted setting, respectively, yield

$$d^{\pi,\gamma} = (1-\gamma)\sum_{t=0}^{\infty}\gamma^t d_t^\pi, \quad d^{\pi,1} = \lim_{H\to\infty}\frac{1}{H}\sum_{t=0}^{H-1} d_t^\pi = \lim_{H\to\infty} d_H^\pi.$$

For the sake of brevity, when we want to make claims for Q^π and V^π at the same time, we simply use $f^\pi = Q^\pi$ or $f^\pi = V^\pi$. This also applies to using none or other sub and super scripts. Also notice that we overload r_t^π, d_t^π and d^π as functions on states as well as state-action pairs. In case statements are meant for both cases, we will omit the arguments.

Lemma 3.1.2 tells us something about the connection of the stationary distribution on the state- and state-action-space.

Lemma 3.1.2. *For any $s \in S$ and $a \in A$, we have*

$$r_t^\pi(s) = \mathbb{E}_{a\sim\pi(s)}[r_t^\pi(s,a)] \quad \text{and} \quad V^\pi(s) = \mathbb{E}_{a\sim\pi(s)}[Q^\pi(s,a)],$$

as well as, for any $s' \in S$ and $a' \in A$,

$$d_t^\pi(s') = \int_A d_t^\pi(s',a')\ \mathrm{d}a', \qquad d_t^\pi(s',a') = d_t^\pi(s')\pi(a' \mid s')$$
$$\text{and} \quad d^\pi(s') = \int_A d^\pi(s',a')\ \mathrm{d}a', \qquad d^\pi(s',a') = d^\pi(s')\pi(a' \mid s').$$

Proof. All the equations, except the last two, follow directly from the definitions. For the last two equations, we can use the ones before and (3.2).

□

The functions Q^π, V^π and d^π are useful for policy evaluation, because they allow us to describe the policy value ρ^π in a simpler manner, as given in Lemma 3.1.3.

Lemma 3.1.3. *For $0 < \gamma < 1$, we have*

$$\rho^\pi = (1-\gamma)\mathbb{E}_{(s_0,a_0)\sim d_0^\pi}[Q(s_0,a_0)] = (1-\gamma)\mathbb{E}_{s_0\sim d_0^\pi}[V(s_0)]. \tag{3.3}$$

For $0 < \gamma \leq 1$, we have

$$\rho^\pi = \mathbb{E}_{(s,a)\sim d^\pi}[r(s,a)]. \tag{3.4}$$

Proof. The first equations follow directly from the definition and Lemma 3.1.2. For the last equation, we apply the theorem of total expectation and get

$$\begin{aligned}
\rho_\pi &= \lim_{H\to\infty} \mathbb{E}_{(*)_H}\left[\sum_{t=0}^{H}\gamma^t r(s_t,a_t) \Big/ \sum_{t=0}^{H}\gamma^t\right] \\
&= \lim_{H\to\infty} \sum_{t=0}^{H}\gamma^t \mathbb{E}_{(*)_t}[r(s_t,a_t)] \Big/ \sum_{t=0}^{H}\gamma^t \\
&= \lim_{H\to\infty} \sum_{t=0}^{H}\gamma^t \sum_{s\in S}\sum_{a\in A} r(s,a)\mathbb{P}_{(*)_t}(s_t = s,\ a_t = a) \Big/ \sum_{t=0}^{H}\gamma^t \\
&= \sum_{s\in S}\sum_{a\in A} r(s,a) \lim_{H\to\infty} \sum_{t=0}^{H}\gamma^t d_t^\pi(s,a) \Big/ \sum_{t=0}^{H}\gamma^t \\
&= \sum_{s\in S}\sum_{a\in A} r(s,a) d^\pi(s,a) \\
&= \mathbb{E}_{(s,a)\sim d^\pi}[r(s,a)].
\end{aligned}$$

□

We can also give a more high level proof of Lemma 3.1.3. It does not use any explicit definition of Q^π or d^π. Instead, we can simply treat them as solutions to the Bellman equations (3.5) or (3.7) and (3.9), respectively.

Alternative proof of Lemma 3.1.3. We multiply the backwards Bellman equations (3.9) with Q^π and use the fact from Lemma 3.2.3 that $\mathcal{P}_*^\pi$ is the adjoint operator of $\mathcal{P}^\pi$,

$$\langle Q^\pi, d^\pi \rangle = \langle Q^\pi, (1-\gamma) d_0^\pi + \gamma \mathcal{P}_*^\pi d^\pi \rangle = (1-\gamma)\langle Q^\pi, d_0^\pi \rangle + \gamma \langle \mathcal{P}^\pi Q^\pi, d^\pi \rangle.$$

We subtract the last summand, and apply the forwards Bellman equations (3.5) or (3.7),

$$\begin{aligned}\rho^\pi &= (1-\gamma)\langle Q^\pi, d_0^\pi \rangle + \rho^\pi \mathbb{1}_{\gamma=1} \\ &= \langle Q^\pi - \gamma \mathcal{P}^\pi Q^\pi, d^\pi \rangle + \langle \rho^\pi \mathbb{1}_{\gamma=1}, d^\pi \rangle \\ &= \langle \mathcal{B}^\pi Q^\pi + \rho^\pi \mathbb{1}_{\gamma=1} - \gamma \mathcal{P}^\pi Q^\pi, d^\pi \rangle = \langle r, d^\pi \rangle.\end{aligned}$$

□

3.2 Bellman Operators and Equations

For functions $Q, d : S \times A \to \mathbb{R}$, we define the *expected Bellman operator* and its adjoint as

$$\begin{aligned}\mathcal{P}^\pi Q(s,a) &\doteq \int_{S\times A} Q(s',a') T^\pi(s',a' \mid s,a)\, \mathrm{d}s'\, \mathrm{d}a' \\ &= \mathbb{E}_{(s',a')\sim T^\pi(s,a)}[Q(s',a')], && \text{where} \quad (s,a) \in S \times A, \\ \mathcal{P}_*^\pi d(s',a') &\doteq \int_{S\times A} d(s,a) T^\pi(s',a' \mid s,a)\, \mathrm{d}s\, \mathrm{d}a, && \text{where} \quad (s',a') \in S \times A.\end{aligned}$$

We can also make these definitions for $V, d : S \to \mathbb{R}$,

$$\begin{aligned}\mathcal{P}^\pi V(s) &\doteq \int_A \int_S V(s') T(s' \mid s,a)\, \mathrm{d}s' \pi(a \mid s)\, \mathrm{d}a \\ &= \mathbb{E}_{a\sim\pi(s),\ s'\sim T(s,a)}[V(s')], && \text{where} \quad s \in S, \\ \mathcal{P}_*^\pi d(s') &\doteq \int_A \int_S d(s) \pi(a \mid s) T(s' \mid s,a)\, \mathrm{d}s\, \mathrm{d}a, && \text{where} \quad s' \in S.\end{aligned}$$

Also, define the operators

$$A \doteq I - \gamma \mathcal{P}_*^\pi \quad \text{and} \quad A_* \doteq I - \gamma \mathcal{P}^\pi.$$

Note, that $\mathcal{P}^\pi$ is a linear operator on $C(\Omega)$. In case the state-action-space $S \times A$ is finite, we write

$$\vec{\mathcal{P}}^\pi \doteq (T^\pi(s',a' \mid s,a))_{(s,a),(s',a')\in S\times A} \quad \text{and} \quad \vec{\mathcal{P}}_*^\pi \doteq (\vec{\mathcal{P}}^\pi)^\top.$$

It turns out that some of the properties of V^π, Q^π and d^π get preserved when applying $\mathcal{P}^\pi$ and $\mathcal{P}_*^\pi$, respectively. We state them collectively in the following Lemma 3.2.1.

Lemma 3.2.1. *For any $s \in S$ and $a \in A$, we have*

$$\mathcal{P}^\pi r_t^\pi(s) = \mathbb{E}_{a\sim\pi(s)}[\mathcal{P}^\pi r_t^\pi(s,a)] \quad \textit{and} \quad \mathcal{P}^\pi V^\pi(s) = \mathbb{E}_{a\sim\pi(s)}[\mathcal{P}^\pi Q^\pi(s,a)],$$

as well as, for any $s' \in S$ and $a' \in A$,

$$\begin{aligned}\mathcal{P}_*^\pi d_t^\pi(s') &= \int_A \mathcal{P}_*^\pi d_t^\pi(s',a')\, \mathrm{d}a', && \mathcal{P}_*^\pi d_t^\pi(s',a') = \mathcal{P}_*^\pi d_t^\pi(s') \pi(a' \mid s'), \\ \textit{and} \quad \mathcal{P}_*^\pi d^\pi(s') &= \int_A \mathcal{P}_*^\pi d^\pi(s',a')\, \mathrm{d}a', && \mathcal{P}_*^\pi d^\pi(s',a') = \mathcal{P}_*^\pi d^\pi(s') \pi(a' \mid s').\end{aligned}$$

Proof. We use the corresponding identities from Lemma 3.1.2, without $\mathcal{P}^\pi$. Firstly,

$$\begin{aligned}\mathbb{E}_{a\sim\pi(s)}[\mathcal{P}^\pi r_t^\pi(s,a)] &= \mathbb{E}_{a\sim\pi(s)}\left[\mathbb{E}_{(s',a')\sim T^\pi(s,a)}[r_t^\pi(s',a')]\right]\\ &= \mathbb{E}_{a\sim\pi(s),\ s'\sim T(s,a)}[r_t^\pi(s')]\\ &= \mathcal{P}^\pi r_t^\pi(s),\end{aligned}$$

and

$$\begin{aligned}\mathbb{E}_{a\sim\pi(s)}[\mathcal{P}^\pi Q^\pi(s,a)] &= \mathbb{E}_{a\sim\pi(s)}\left[\mathbb{E}_{(s',a')\sim T^\pi(s,a)}[Q^\pi(s',a')]\right]\\ &= \mathbb{E}_{a\sim\pi(s),\ s'\sim T(s,a)}[V^\pi(s')]\\ &= \mathcal{P}^\pi V^\pi(s).\end{aligned}$$

Secondly,

$$\begin{aligned}\int_A \mathcal{P}_*^\pi d_t^\pi(s',a')\ \mathrm{d}a' &= \int_A\int_{S\times A} d_t^\pi(s,a)T^\pi(s',a'\mid s,a)\ \mathrm{d}s\ \mathrm{d}a\ \mathrm{d}a'\\ &= \int_A\int_S d_t^\pi(s)\pi(a\mid s)T(s'\mid s,a)\ \mathrm{d}s\ \mathrm{d}a\underbrace{\int_A \pi(a'\mid s')\ \mathrm{d}a'}_{1} = \mathcal{P}_*^\pi d_t^\pi(s')\end{aligned}$$

and

$$\begin{aligned}\mathcal{P}_*^\pi d_t^\pi(s')\pi(a'\mid s') &= \int_A\int_S d_t^\pi(s)\pi(a\mid s)T(s'\mid s,a)\ \mathrm{d}s\ \mathrm{d}a\pi(a'\mid s')\\ &= \int_{S\times A} d_t^\pi(s,a)T^\pi(s',a'\mid s,a)\ \mathrm{d}s\ \mathrm{d}a = \mathcal{P}_*^\pi d_t^\pi(s',a').\end{aligned}$$

For the last two equations, we can use the two before, the linearity and continuity of the $\mathcal{P}_*^\pi$, and (3.2).

□

Lemma 3.2.2. *For all $t\in\mathbb{N}$ it holds that*

$$\begin{aligned} r_{t+1}^\pi &= \mathcal{P}^\pi r_t^\pi, & r_t^\pi &= (\mathcal{P}^\pi)^t r_0^\pi,\\ d_{t+1}^\pi &= \mathcal{P}_*^\pi d_t^\pi, & d_t^\pi &= (\mathcal{P}_*^\pi)^t d_0^\pi.\end{aligned}$$

Proof. Firstly,

$$\begin{aligned}\mathcal{P}^\pi r_{t-1}^\pi(s,a) &= \sum_{s'\in S}\sum_{a'\in A} r_{t-1}^\pi(s',a')T^\pi(s',a'\mid s,a)\\ &= \sum_{s'\in S}\sum_{a'\in A}\mathbb{E}_{(s_0=s',\ a_0=a')_{t-1}}[r(s_{t-1},a_{t-1})]T^\pi(s',a'\mid s,a)\\ &= \mathbb{E}_{(s_0=s,\ a_0=a)_t}[r(s_t,a_t)]\\ &= r_t^\pi(s,a).\end{aligned}$$

Secondly, applying the theorem of total probability,

$$\begin{aligned} d_t^\pi(s') &= \mathbb{P}_{(*)_t}(s_t=s')\\ &= \sum_{s\in S}\sum_{a\in A}\mathbb{P}_{(*)_t}(s_t=s'\mid s_{t-1}=s,\ a_{t-1}=a)\mathbb{P}_{(*)_t}(a_{t-1}=a\mid s_{t-1}=s)\mathbb{P}_{(*)_t}(s_{t-1}=s)\\ &= \sum_{s\in S}\sum_{a\in A} T(s'\mid s,a)\pi(a\mid s)d_{t-1}^\pi(s)\\ &= \mathcal{P}_*^\pi d_{t-1}^\pi(s').\end{aligned}$$

Now, we just apply $\mathbb{E}_{a\sim\pi(s)}$ or multiply with $\pi(a' \mid s')$ and use Lemmas 3.1.2 and 3.2.1 to also get the result for only states or state-action pairs, respectively.

□

For some of the proofs later on we are going to need some more technical properties of $\mathcal{P}^\pi$ and $\mathcal{P}^\pi_*$, which we summarize in the following Lemma 3.2.3.

Lemma 3.2.3. *The expected Bellman operator $\mathcal{P}^\pi$ and $\mathcal{P}^\pi_*$ have the following properties:*

1. *On the Hilbert space $L_2(\Omega)$, $\mathcal{P}^\pi_*$ is the adjoint operator of $\mathcal{P}^\pi$.*
2. *The operator norms on $L_\infty(\Omega)$ and $L_1(\Omega)$, respectively, are*
$$\|\mathcal{P}^\pi\|_{L_\infty(\Omega)} = 1 \quad \text{and} \quad \|\mathcal{P}^\pi_*\|_{L_1(\Omega)} = 1.$$
3. *Given a probability distribution $d \in \Delta(\Omega)$ and $p \geq 1$, we have*
$$\|\mathcal{P}^\pi\|_{L_\infty(d)} = 1 \quad \text{and} \quad \|\mathcal{P}^\pi\|_{L_p(\mathcal{P}^\pi_* d),\, L_p(d)} = 1.$$
In case $d > 0$, it also holds that $\|D^{-1}\mathcal{P}^\pi_* D\|_{L_1(d)} = 1$.
4. *The spectral radius of $\mathcal{P}^\pi$, $\mathcal{P}^\pi_*$ and $D^{-1}\mathcal{P}^\pi_* D$ is 1.*

Proof. 1. Using Fubini's theorem, we calculate

$$\begin{aligned}
\langle \mathcal{P}^\pi Q, d\rangle &= \int_{S\times A} \left(\int_{S\times A} Q(s',a') T^\pi(s',a' \mid s,a)\, \mathrm{d}s'\, \mathrm{d}a' \right) d(s,a)\, \mathrm{d}s\, \mathrm{d}a \\
&= \int_{S\times A} Q(s',a') \left(\int_{S\times A} d(s,a) T^\pi(s',a' \mid s,a)\, \mathrm{d}s\, \mathrm{d}a \right) \mathrm{d}s'\, \mathrm{d}a' = \langle Q, \mathcal{P}^\pi_* d\rangle, \\
\langle \mathcal{P}^\pi V, d\rangle &= \int_S \left(\int_A \int_S V(s') T(s' \mid s,a)\, \mathrm{d}s' \pi(a \mid s)\, \mathrm{d}a \right) d(s)\, \mathrm{d}s \\
&= \int_S V(s') \left(\int_A \int_S d(s) \pi(a \mid s) T(s' \mid s,a)\, \mathrm{d}s\, \mathrm{d}a \right) \mathrm{d}s' = \langle V, \mathcal{P}^\pi_* d\rangle.
\end{aligned}$$

2. Using the triangle inequality, we calculate

$$\|\mathcal{P}^\pi f\|_{L_\infty(\Omega)} = \sup_{\omega\in\Omega} \left| \mathbb{E}_{\omega'\sim T^\pi(\omega)}[f(\omega')] \right| \leq \sup_{\omega\in\Omega} \mathbb{E}_{\omega'\sim T^\pi(\omega)} |f(\omega')| \leq \|f\|_{L_\infty(\Omega)}.$$

Applying it again, this time together with Fubini's theorem,

$$\begin{aligned}
\|\mathcal{P}^\pi_* d\|_{L_1(S\times A)} &= \int_{S\times A} \left| \int_{S\times A} d(s,a) T^\pi(s',a' \mid s,a)\, \mathrm{d}s\, \mathrm{d}a \right| \mathrm{d}s'\, \mathrm{d}a' \\
&\leq \int_{S\times A} |d(s,a)| \underbrace{\left(\int_{S\times A} T^\pi(s',a' \mid s,a)\, \mathrm{d}s'\, \mathrm{d}a' \right)}_{1} \mathrm{d}s\, \mathrm{d}a = \|d\|_{L_1(S\times A)}, \\
\|\mathcal{P}^\pi_* d\|_{L_1(S)} &= \int_S \left| \int_A \int_S d(s) \pi(a \mid s) T(s' \mid s,a)\, \mathrm{d}s\, \mathrm{d}a \right| \mathrm{d}s' \\
&\leq \int_S |d(s)| \underbrace{\left(\int_A \int_S T(s' \mid s,a)\, \mathrm{d}s' \pi(a \mid s)\, \mathrm{d}a \right)}_{1} \mathrm{d}s = \|d\|_{L_1(S)}.
\end{aligned}$$

Consider $f \equiv 1$ and $d \geq 0$ for equality, respectively.

3. Similarly to before, we calculate,

$$\|\mathcal{P}^\pi f\|_{L_\infty(d)} = \operatorname*{ess\,sup}_{\omega\in\Omega} \left|\mathbb{E}_{\omega'\sim T^\pi(\omega)}[f(\omega')]\right| \leq \operatorname*{ess\,sup}_{\omega\in\Omega} \mathbb{E}_{\omega'\sim T^\pi(\omega)}|f(\omega')| \leq \|f\|_{L_\infty(d)}$$

and use Jensen's inequality on $\mathcal{P}^\pi$ and monotonicity of the inner product with a non-negative second factor, to calculate

$$\|\mathcal{P}^\pi f\|^p_{L_p(d)} = \langle|\mathcal{P}^\pi f|^p, d\rangle \leq \langle\mathcal{P}^\pi|f|^p, d\rangle = \langle|f|^p, \mathcal{P}^\pi_* d\rangle = \|f\|^p_{L_p(\mathcal{P}^\pi_* d)}.$$

Consider $f \equiv 1$ for equality. Then, using Fubini's theorem again,

$$\begin{aligned}\|D^{-1}\mathcal{P}^\pi_* Dw\|_{L_1(d)} &= \int_\Omega \left|\frac{1}{d(\omega')}\int_\Omega d(\omega)w(\omega)T^\pi(\omega' \mid \omega)\,\mathrm{d}\omega\right| d(\omega')\,\mathrm{d}\omega' \\ &\leq \int_\Omega \Bigg(\underbrace{\int_\Omega T^\pi(\omega' \mid \omega)\,\mathrm{d}\omega'}_{1}\Bigg)|w(\omega)|d(\omega)\,\mathrm{d}\omega = \|w\|_{L_1(d)}.\end{aligned}$$

Consider $w \geq 0$ for equality.

4. It is well known that the spectrum of an operator is contained within the ball with its operator norm as the radius. This bounds the spectral radius from above by 1. Now, by definition of $\mathcal{P}^\pi$, the constant function 1 is an eigenvector with eigenvalue 1. It is also well known that the spectrum of an operator is identical to that of its adjoint. Therefore, the spectral radius of $\mathcal{P}^\pi_*$ is even equal to 1. Finally, note that similar operators share the same spectrum and spectral radius.

□

For $Q : S \times A \to \mathbb{R}$, $V : S \to \mathbb{R}$ and $d : \Omega \to \mathbb{R}$, define the *forwards Bellman operator* $\mathcal{B}^\pi$ and *backwards Bellman operator* $\mathcal{T}^\pi$ by

$$\begin{aligned}\mathcal{B}^\pi Q(s,a) &\doteq \begin{cases} r^\pi_0(s,a) + \gamma\mathcal{P}^\pi Q(s,a) & \text{for} \quad \gamma \in (0,1), \\ r^\pi_0(s,a) - \rho^\pi + \mathcal{P}^\pi Q(s,a) & \text{for} \quad \gamma = 1, \end{cases} \\ \mathcal{B}^\pi V(s) &\doteq \begin{cases} r^\pi_0(s) + \gamma\mathcal{P}^\pi V(s) & \text{for} \quad \gamma \in (0,1), \\ r^\pi_0(s) - \rho^\pi + \mathcal{P}^\pi V(s) & \text{for} \quad \gamma = 1, \end{cases} \\ \mathcal{T}^\pi d &\doteq (1-\gamma)d^\pi_0 + \gamma\mathcal{P}^\pi_* d \quad \text{for} \quad \gamma \in (0,1],\end{aligned}$$

where $s \in S$ and $a \in A$. We will also use the notation

$$\mathcal{B}^{\pi,\gamma}_\rho f = r^\pi_0 - \rho + \gamma\mathcal{P}^\pi f, \quad \text{so} \quad \mathcal{B}^\pi = \begin{cases} \mathcal{B}^{\pi,\gamma}_0 & \text{if} \quad 0 < \gamma < 1, \\ \mathcal{B}^{\pi,1}_{\rho^\pi} & \text{if} \quad \gamma = 1. \end{cases}$$

The following two Lemmas 3.2.4 and 3.2.5 are both Bellman equations. They have in common that they describe the state-(action-)function and stationary distribution by means of a time shift. The first, better known result, uses states and actions after going forward in time by applying $\mathcal{P}^\pi$. Accordingly, the latter uses previous states and actions together with $\mathcal{P}^\pi_*$.

Lemma 3.2.4 (forwards Bellman equations)**.** *Let $\gamma \in (0,1)$, then Q^π and V^π are fixed points of the Bellman operator $\mathcal{B}^{\pi,\gamma}$, i.e., solution to the* (discounted) forwards Bellman equations*,*

$$Q = r_0^\pi + \gamma \mathcal{P}^\pi Q, \qquad \text{where} \quad Q : S \times A \to \mathbb{R}, \tag{3.5}$$
$$V = r_0^\pi + \gamma \mathcal{P}^\pi V, \qquad \text{where} \quad V : S \to \mathbb{R}. \tag{3.6}$$

Let $\gamma = 1$, then ρ^π, Q^π and V^π are solutions to the (undiscounted) forwards Bellman equations*,*

$$Q = r_0^\pi - \rho + \mathcal{P}^\pi Q, \qquad \text{where} \quad \rho \in \mathbb{R}, \quad Q : S \times A \to \mathbb{R}, \tag{3.7}$$
$$V = r_0^\pi - \rho + \mathcal{P}^\pi V, \qquad \text{where} \quad \rho \in \mathbb{R}, \quad V : S \to \mathbb{R}. \tag{3.8}$$

Proof. We first show the claim for state-action-values,

$$\begin{aligned}
&Q^\pi(s,a) \\
&= \mathbb{E}_{(s_{t+1},a_{t+1})\sim T^\pi(s_t,a_t)} \left[\sum_{t=0}^{\infty} (\gamma^t r(s_t,a_t) - \rho^\pi \mathbb{1}_{\gamma=1}) \;\middle|\; \begin{matrix} s_0 = s, \\ a_0 = a \end{matrix} \right] \\
&= \mathbb{E}_{(s_t,a_t)\sim T^\pi(s_{t-1},a_{t-1})} \left[(r(s_0,a_0) - \rho^\pi \mathbb{1}_{\gamma=1}) + \gamma \sum_{t=1}^{\infty} (\gamma^{t-1} r(s_t,a_t) - \rho^\pi \mathbb{1}_{\gamma=1}) \;\middle|\; \begin{matrix} s_0 = s, \\ a_0 = a \end{matrix} \right] \\
&= r(s,a) - \rho^\pi \mathbb{1}_{\gamma=1} \\
&\qquad + \gamma \mathbb{E}_{(s_1,a_1)\sim T^\pi(s_0,a_0)} \left[\mathbb{E}_{(s_{t+1},a_{t+1})\sim T^\pi(s_t,a_t)} \left[\sum_{t=1}^{\infty} (\gamma^{t-1} r(s_t,a_t) - \rho^\pi \mathbb{1}_{\gamma=1}) \right] \;\middle|\; \begin{matrix} s_0 = s, \\ a_0 = a \end{matrix} \right] \\
&= r(s,a) - \rho^\pi \mathbb{1}_{\gamma=1} \\
&\qquad + \gamma \mathbb{E}_{(s',a')\sim T^\pi(s,a)} \left[\mathbb{E}_{(\tilde{s}_{t+1},\tilde{a}_{t+1})\sim T^\pi(\tilde{s}_t,\tilde{a}_t)} \left[\sum_{t=0}^{\infty} (\gamma^t r(\tilde{s}_t,\tilde{a}_t) - \rho^\pi \mathbb{1}_{\gamma=1}) \;\middle|\; \begin{matrix} \tilde{s}_0 = s', \\ \tilde{a}_0 = a' \end{matrix} \right] \right] \\
&= r(s,a) - \rho^\pi \mathbb{1}_{\gamma=1} + \gamma \mathbb{E}_{(s',a')\sim T^\pi(s,a)} [Q^\pi(s',a')] \\
&= r(s,a) - \rho^\pi \mathbb{1}_{\gamma=1} + \gamma \mathcal{P}^\pi Q^\pi(s,a).
\end{aligned}$$

By applying the expectation $\mathbb{E}_{a\sim\pi(s)}$ on both sides and using Lemma 3.2.1, we get the claim for the state-value function.

□

Lemma 3.2.5 (backwards Bellman equations)**.** *Let $\gamma \in (0,1]$, then d^π is a fixed point of the backwards Bellman operator $\mathcal{T}^\pi$, i.e., a solution to the* backwards Bellman equations

$$d = (1-\gamma) d_0^\pi + \gamma \mathcal{P}_*^\pi d, \quad \text{where} \quad d : \Omega \to \mathbb{R}. \tag{3.9}$$

In case $\gamma = 1$, we need to add the normalization constraint

$$\int_\Omega d(\omega) \, \mathrm{d}\omega = 1 \quad \text{and} \quad d \geq 0.$$

Proof. Because $\mathcal{P}_*^\pi$ is linear and bounded by Lemma 3.2.3, we get

$$\begin{aligned} d^\pi &= \lim_{H\to\infty} \sum_{t=0}^{H} \gamma^t d_t^\pi \Big/ \sum_{t=0}^{H} \gamma^t \\ &= \lim_{H\to\infty} d_0^\pi \Big/ \sum_{t=0}^{H} \gamma^t + \lim_{H\to\infty} \gamma \sum_{t=1}^{H} \gamma^{t-1} \mathcal{P}_*^\pi d_{t-1}^\pi \Big/ \sum_{t=0}^{H} \gamma^t \\ &= (1-\gamma) d_0^\pi + \gamma \mathcal{P}_*^\pi \Bigg(\underbrace{\lim_{H\to\infty} \sum_{t=1}^{H} \gamma^{t-1} d_{t-1}^\pi \Big/ \sum_{t=0}^{H} \gamma^t}_{d^\pi} \Bigg). \end{aligned}$$

□

It remains to be discussed, whether Q^π, V^π and d^π really exist, i.e., they are well defined, and if they are unique as fixed points of the corresponding Bellman operators.

Remark 3.2.6. For $\gamma < 1$, we can verify that $\mathcal{B}^\pi$ and $\mathcal{T}^\pi$ are γ-contractions with respect to $\|\cdot\|_{L_\infty(\Omega)}$ and $\|\cdot\|_{L_1(\Omega)}$, respectively. For $f_1, f_2 : \Omega \to \mathbb{R}$ and $d_1, d_2 : \Omega \to \mathbb{R}$, we calculate

$$\begin{aligned} \|\mathcal{B}^\pi f_1 - \mathcal{B}^\pi f_2\|_{L_\infty(\Omega)} &= \|\gamma \mathcal{P}^\pi (f_1 - f_2)\|_{L_\infty(\Omega)} \leq \gamma \|\mathcal{P}^\pi\|_{L_\infty(\Omega)} \|f_1 - f_2\|_{L_\infty(\Omega)}, \\ \|\mathcal{T}^\pi d_1 - \mathcal{T}^\pi d_2\|_{L_1(\Omega)} &= \|\gamma \mathcal{P}_*^\pi (d_1 - d_2)\|_{L_1(\Omega)} \leq \gamma \|\mathcal{P}_*^\pi\|_{L_1(\Omega)} \|d_1 - d_2\|_{L_1(\Omega)}. \end{aligned}$$

By Lemma 3.2.3, the operator norms falls away. By Banach's fixed point theorem, Q^π, V^π and d^π are fixed points of $\mathcal{B}^\pi$ and $\mathcal{T}^\pi$, respectively. ■

Remark 3.2.7. A more constructive alternative to Banach's fixed point theorem uses the Bellman equations (3.5), (3.6), and (3.9), and Neumann series. We start by rewriting the Bellman equations as

$$(I - \gamma \mathcal{P}^\pi) f = r_0^\pi \quad \text{and} \quad (I - \gamma \mathcal{P}_*^\pi) d = (1-\gamma) d_0^\pi.$$

According to Lemma 3.2.3,

$$\|\gamma \mathcal{P}^\pi\|_{L_\infty(\Omega)} = \gamma < 1 \quad \text{and} \quad \|\gamma \mathcal{P}_*^\pi\|_{L_1(\Omega)} = \gamma < 1.$$

Therefore, we can construct the Neumann series expansions for the resolvents at γ,

$$(I - \gamma \mathcal{P}^\pi)^{-1} = \sum_{t=0}^{\infty} \gamma^t (\mathcal{P}^\pi)^t \quad \text{and} \quad (I - \gamma \mathcal{P}_*^\pi)^{-1} = \sum_{t=0}^{\infty} \gamma^t (\mathcal{P}_*^\pi)^t.$$

We apply this and use Lemma 3.2.2, as well as (3.1) and (3.2), respectively, to get

$$\begin{aligned} f &= \sum_{t=0}^{\infty} \gamma^t (\mathcal{P}^\pi)^t r_0^\pi = \sum_{t=0}^{\infty} \gamma^t r_t^\pi = f^\pi, \\ \frac{1}{1-\gamma} d &= \sum_{t=0}^{\infty} \gamma^t (\mathcal{P}_*^\pi)^t d_0^\pi = \sum_{t=0}^{\infty} \gamma^t d_t^\pi = \frac{1}{1-\gamma} d^\pi. \end{aligned}$$

■

Remark 3.2.8. For the undiscounted setting $\gamma = 1$, the Bellman operators are not a contraction and the Neumann series expansion does not converge any more.

Once f^π is fixed, ρ^π becomes a unique solution to the rest of the undiscounted forwards Bellman equations (3.7) or (3.8), respectively. Any $\rho \in \mathbb{R}$, such that $f^\pi = \mathcal{B}_\rho^{\pi,1} f^\pi$, causes

$$0 = f^\pi - f^\pi = \mathcal{B}^\pi f^\pi - \mathcal{B}_\rho^{\pi,1} f^\pi = \rho - \rho^\pi.$$

The backwards Bellman equations (3.9) from Lemma 3.2.5 reduce to the eigenvalue problem

$$\mathcal{P}_*^\pi d = d, \quad \text{where} \quad \int_\Omega d(\omega)\, \mathrm{d}\omega = 1 \quad \text{and} \quad d \geq 0.$$

Therefore, the question is, whether $\mathcal{P}_*^\pi$ has eigenvalue 1 with a real-valued eigenvector. By Lemma 3.2.3, the spectral radius of $\mathcal{P}_*^\pi$ is equal to 1. For a finite state-action-space, we can use the Perron-Frobenius Theorem 2.4.1 on $\vec{\mathcal{P}}_*^\pi$. In order for the matrix to be positive and irreducible, we require Assumption 3.1.1. If so, d^π exists and is unique as solution to this eigenvalue problem. For an infinite state-action-space, the theory gets more complicated and beyond the scope of this work [26].

■

We can use the Bellman equations (3.5), (3.6), and (3.9), from Lemmas 3.2.4 and 3.2.5 to derive some simple error bounds for the policy value and Q^π, V^π and d^π, respectively, in terms of the according Bellman error. Unfortunately, the bound is only sharp for very low discount factor γ.

Proposition 3.2.9. *Let $\gamma < 1$. Then, for any $Q : S \times A \to \mathbb{R}$, $V : S \to \mathbb{R}$, and $d : S \times A \to \mathbb{R}$, we have*

$$\begin{aligned}
\left|\mathbb{E}_{(s_0,a_0)\sim d_0^\pi}[Q(s_0,a_0)] - \rho^\pi\right| &\leq \|Q^\pi - Q\|_{L_\infty(d_0^\pi)}, \\
\left|\mathbb{E}_{s_0\sim d_0^\pi}[V(s_0)] - \rho^\pi\right| &\leq \|V^\pi - V\|_{L_\infty(d_0^\pi)}, \\
\left|\mathbb{E}_{(s,a)\sim d}[r(s,a)] - \rho^\pi\right| &\leq \|d^\pi - d\|_{L_1(\Omega)} \|r\|_{L_\infty(\Omega)},
\end{aligned}$$

and

$$\begin{aligned}
\|Q^\pi - Q\|_{L_\infty(d_0^\pi)} &\leq \frac{1}{1-\gamma} \|\mathcal{B}^\pi Q - Q\|_{L_\infty(d_0^\pi)}, \\
\|V^\pi - V\|_{L_\infty(d_0^\pi)} &\leq \frac{1}{1-\gamma} \|\mathcal{B}^\pi V - V\|_{L_\infty(d_0^\pi)}, \\
\|d^\pi - d\|_{L_1(\Omega)} &\leq \frac{1}{1-\gamma} \|\mathcal{T}^\pi d - d\|_{L_1(\Omega)}.
\end{aligned}$$

Proof. For the first inequalities, we use Lemma 3.1.3 and calculate

$$\begin{aligned}
\left|\mathbb{E}_{(s_0,a_0)\sim d_0^\pi}[Q(s_0,a_0)] - \rho^\pi\right| &= \left|\mathbb{E}_{(s_0,a_0)\sim d_0^\pi}[Q^\pi(s_0,a_0) - Q(s_0,a_0)]\right| \\
&\leq \|Q^\pi - Q\|_{L_1(d_0^\pi)} \leq \|Q^\pi - Q\|_{L_\infty(d_0^\pi)}, \\
\left|\mathbb{E}_{s_0\sim d_0^\pi}[V(s_0)] - \rho^\pi\right| &= \left|\mathbb{E}_{s_0\sim d_0^\pi}[V^\pi(s_0) - V(s_0)]\right| \\
&\leq \|V^\pi - V\|_{L_1(d_0^\pi)} = \|V^\pi - V\|_{L_\infty(d_0^\pi)}, \\
\left|\mathbb{E}_{(s,a)\sim d}[r(s,a)] - \rho^\pi\right| &= \left|\int_{S\times A} r(s,a)(d^\pi(s,a) - d(s,a))\, \mathrm{d}s\, \mathrm{d}a\right| \\
&\leq \|d^\pi - d\|_{L_1(\Omega)} \|r\|_{L_\infty(\Omega)}.
\end{aligned}$$

According to Remark 3.2.6, the Bellman operators $\mathcal{B}^\pi$ and $\mathcal{T}^\pi$ are γ-contractions with respect to $\|\cdot\|_{L_\infty(\Omega)}$ and $\|\cdot\|_{L_1(\Omega)}$, where f^π and d^π are fixed points, respectively. Together with the triangle inequality, this leads to

$$\begin{aligned}
\|f^\pi - f\|_{L_\infty(d_0^\pi)} &= \|\mathcal{B}^\pi f^\pi - f\|_{L_\infty(d_0^\pi)} \leq \|\mathcal{B}^\pi f^\pi - \mathcal{B}^\pi f\|_{L_\infty(d_0^\pi)} + \|\mathcal{B}^\pi f - f\|_{L_\infty(d_0^\pi)} \\
&\leq \gamma \|f^\pi - f\|_{L_\infty(d_0^\pi)} + \|\mathcal{B}^\pi f - f\|_{L_\infty(d_0^\pi)}, \\
\|d^\pi - d\|_{L_1(\Omega)} &= \|\mathcal{T}^\pi d^\pi - d\|_{L_1(\Omega)} \leq \|\mathcal{T}^\pi d^\pi - \mathcal{T}^\pi d\|_{L_1(\Omega)} + \|\mathcal{T}^\pi d - d\|_{L_1(\Omega)} \\
&\leq \gamma \|d^\pi - d\|_{L_1(\Omega)} + \|\mathcal{T}^\pi d - d\|_{L_1(\Omega)}.
\end{aligned}$$

□

So far, it has become clear that the discounted case is usually much easier than the undiscounted. It is also more general.

Remark 3.2.10. Let d^π and $\mathcal{P}^{\pi,\gamma}$ be the stationary distribution and expected Bellman operator for some discounted MDP. Using the backwards Bellman equations (3.9), we can express the expected Bellman operator $\mathcal{P}^{\pi,1}$ for an equivalent undiscounted MDP, i.e., the stationary distributions are the same.

Writing down the backwards Bellman equations for both settings, we have

$$(1-\gamma) d_0^\pi + \gamma \mathcal{P}_*^{\pi,\gamma} d^\pi = d^\pi \overset{!}{=} \mathcal{P}_*^{\pi,1} d^\pi.$$

Let λ be the Lebesgue measure on $S \times A$. Because d^π is a distribution, using (2.1), we get

$$\lambda d^\pi = \int_{S\times A} d^\pi(s,a) \, \mathrm{d}s \, \mathrm{d}a = 1.$$

Hence, we define

$$\mathcal{P}_*^{\pi,1} \doteq (1-\gamma) d_0^\pi \lambda + \gamma \mathcal{P}_*^{\pi,\gamma}.$$

Now, we want to calculate the adjoint operator $\mathcal{P}^{\pi,1}$. We start with

$$\begin{aligned}
\langle Q, d_0^\pi \lambda d \rangle &= \int_{S\times A} Q(s,a) d_0^\pi(s,a) \int_{S\times A} d(s',a') \, \mathrm{d}s' \, \mathrm{d}a' \, \mathrm{d}s \, \mathrm{d}a \\
&= \int_{S\times A} \left(\int_{S\times A} Q(s,a) d_0^\pi(s,a) \, \mathrm{d}s \, \mathrm{d}a \right) d(s',a') \, \mathrm{d}s' \, \mathrm{d}a' = \langle \mathbb{E}_{d_0^\pi}[Q], d \rangle.
\end{aligned}$$

Since building adjoint operators is linear, using (2.1), we have

$$\mathcal{P}^{\pi,1} = (1-\gamma) d_0^\pi + \gamma \mathcal{P}^{\pi,\gamma}.$$

In case the state-action-space is finite, we have

$$\vec{\mathcal{P}}^{\pi,1} = (1-\gamma) \vec{1} (\vec{d}_0^\pi)^\top + \gamma \vec{\mathcal{P}}^{\pi,\gamma}.$$

This means that when transitioning from (s,a) to (s',a') in the equivalent undiscounted MDP, we transition, according to the transition probability $T^\pi(s',a' \mid s,a)$ of the original discounted MDP with probability γ and re-spawn at an initial state-action pair, according to the initial state-action-distribution $d^\pi(s,a)$ of the original discounted MDP, with probability $1-\gamma$.

■

3.3 Bellman Linear Programs

There are two types of Bellman equations that follow a somewhat similar structure, using $\mathcal{P}^\pi$ and $\mathcal{P}^\pi_*$, respectively, raises the question of whether there is some way to unify them. Indeed, the theory of Linear Programming allows this, provided that the state-action-space is finite. We provide the result in the following Lemma 3.3.1.

Lemma 3.3.1 (Q-LP)**.** *Let* $0 < \gamma < 1$*. Then, the* (discounted) primal Q-LP

$$\begin{aligned} \rho^{\pi,\gamma} = \min_{Q:S\times A\to\mathbb{R}} & (1-\gamma)\mathbb{E}_{(s_0,a_0)\sim d_0^\pi}[Q(s_0,a_0)] \\ s.t. \quad & \forall (s,a)\in S\times A:\ Q(s,a) \geq r(s,a) + \gamma \mathcal{P}^\pi Q(s,a), \end{aligned} \tag{3.10}$$

has the (discounted) dual Q-LP

$$\begin{aligned} \rho^{\pi,\gamma} = \max_{d:S\times A\to\mathbb{R}_{\geq 0}} & \mathbb{E}_{(s,a)\sim d}[r(s,a)] \\ s.t. \quad & \forall (s,a)\in S\times A:\ d(s,a) = (1-\gamma)d_0^\pi(s,a) + \gamma \mathcal{P}^\pi_* d(s,a). \end{aligned} \tag{3.11}$$

They have unique solutions $Q^{\pi,\gamma}$ *and* $d^{\pi,\gamma}$*, respectively.*

Let $\gamma = 1$*. Then, the* (undiscounted) primal Q-LP

$$\begin{aligned} \rho^{\pi,1} = \min_{Q:S\times A\to\mathbb{R},\ \rho\in\mathbb{R}} & \rho \\ s.t. \quad & \forall (s,a)\in S\times A:\ Q(s,a) = r(s,a) - \rho + \mathcal{P}^\pi Q(s,a), \end{aligned} \tag{3.12}$$

has the (undiscounted) dual Q-LP

$$\begin{aligned} \rho^{\pi,1} = \max_{d:S\times A\to\mathbb{R}} & \mathbb{E}_{(s,a)\sim d}[r(s,a)] \\ s.t. \quad & \forall (s,a)\in S\times A:\ d(s,a) = \mathcal{P}^\pi_* d(s,a), \\ & \sum_{(s,a)\in S\times A} d(s,a) = 1. \end{aligned} \tag{3.13}$$

They have solutions $(Q^{\pi,1}, \rho^{\pi,1})$ *and* $d^{\pi,1}$*, respectively, of which* $\rho^{\pi,1}$ *is unique.*

Proof. Let $0 < \gamma < 1$ and $L_{\mathrm{P}}(Q,d)$ and $L_{\mathrm{D}}(d,Q)$ be the Lagrangian of the primal and dual Q-LP (3.10) and (3.11), respectively. By Lemma 3.2.3, $\mathcal{P}^\pi_*$ is the adjoint of $\mathcal{P}^\pi$. Thus, the conditions of Lemma 2.2.2 hold,

$$\begin{aligned} L_{\mathrm{P}}(Q,d) &= (1-\gamma)\mathbb{E}_{(s_0,a_0)\sim d_0^\pi}[Q(s_0,a_0)] + \langle d, r + \gamma\mathcal{P}^\pi Q - Q\rangle \\ &= (1-\gamma)\langle Q, d_0^\pi\rangle + \langle r, d\rangle + \gamma\langle \mathcal{P}^\pi Q, d\rangle - \langle Q, d\rangle \\ &= \mathbb{E}_{(s,a)\sim d}[r(s,a)] + \langle Q, (1-\gamma)d_0^\pi + \gamma\mathcal{P}^\pi_* d - d\rangle = L_{\mathrm{D}}(d,Q). \end{aligned}$$

Consider an arbitrary Q, which is feasible for the primal LP, i.e., $Q \geq \mathcal{B}^\pi Q$. Since $\mathcal{B}^\pi$ is a monotonic γ-contraction, we can apply Banach iteration to Q and get

$$Q \geq \mathcal{B}^\pi Q \geq (\mathcal{B}^\pi)^2 Q \geq (\mathcal{B}^\pi)^3 Q \geq \cdots \geq \lim_{t\to\infty} (\mathcal{B}^\pi)^t Q = Q^{\pi,\gamma}.$$

Because we are minimizing the objective, $Q^{\pi,\gamma}$ is the optimal solution.

The dual constraints are exactly the backwards Bellman equations (3.9). Their solution $d^{\pi,\gamma}$ is unique in the discounted setting.

Now, let $\gamma = 1$ and $L_{\mathrm{P}}((\rho, Q), d)$ and $L_{\mathrm{D}}(d, (\rho, Q))$ be the Lagrangian of the primal and dual Q-LP (3.12) and (3.13), respectively. By Lemma 3.2.3, $\mathcal{P}_*^\pi$ is the adjoint of $\mathcal{P}^\pi$. Like before,

$$\begin{aligned} L_{\mathrm{P}}((\rho, Q), d) &= \rho + \langle d, r - \rho + \mathcal{P}^\pi Q - Q \rangle \\ &= \langle r, d \rangle + \langle Q, \mathcal{P}_*^\pi d \rangle - \langle Q, d \rangle + \rho - \langle \rho, d \rangle \\ &= \mathbb{E}_{(s,a)\sim d}[r(s,a)] + \langle Q, \mathcal{P}_*^\pi d - d \rangle + \rho(1 - \langle 1, d \rangle) = L_{\mathrm{D}}(d, (\rho, Q)). \end{aligned}$$

The primal and dual constraints are exactly the forwards and backwards Bellman equations (3.7) and (3.9) respectively. Due to the objective of the undiscounted primal Q-LP (3.12), $\rho^{\pi,1}$ must be the unique solution.

□

Notice that applying Strong Duality yields Lemma 3.1.3. The primal constraints can be viewed as a relaxation of the forwards Bellman equations (3.5), since the equality is replaced by an inequality, while at the same time introducing a minimization to compensate for the loss of information. Since the dual constraints are exactly the backwards Bellman equations from Lemma 3.2.5, we establish a duality between these two notions.

One might wonder, if there is a reasonable V-LP, which would yield V^π as its optimal solution [32, 31]. Indeed, there is, but the solution is V^{π_*}, where π_* is the optimal policy. Therefore, it is more related to policy optimization rather than -evaluation. Therefore, we will not discuss it further.

3.4 Classical Off-Policy Evaluation

We review the basics of the classical approach for OPE [38]. Consider an MDP with finite horizon H and trajectory

$$\tau = (s_0, a_0, s_1, a_1, \dots s_{H-1}, a_{H-1}, s_H).$$

The probability of obtaining τ via evaluation policy π is

$$\Pr_\pi[\tau] \doteq d_0(s_0) \prod_{t=0}^{H-1} \pi(a_t \mid s_t) T(s_{t+1} \mid s_t, a_t).$$

Now, introduce a behavior policy b. Analogous to Assumption 3.5.1, we require $\pi \ll b$. Since the initial state distribution d_0 and transition probabilities T stay the same, we get the *importance sampling quotient (ISQ)*

$$q_\tau \doteq \frac{\Pr_\pi[\tau]}{\Pr_b[\tau]} = \prod_{t=0}^{H-1} q_t, \quad \text{where} \quad q_t \doteq \frac{\pi(a_t \mid s_t)}{b(a_t \mid s_t)}.$$

We can use the ISQ to restate the policy value as

$$\rho^\pi = \mathbb{E}_\pi[G_\tau] = \mathbb{E}_b[G_\tau q_\tau], \quad \text{where} \quad G_\tau \doteq \sum_{t=0}^{H-1} \gamma^t R(s_t, a_t, s_{t+1}).$$

Let $(\tau_i)_{i=1}^n$ be trajectories, sampled by using the behavior policy b. Since we have access to both π and b, we also get the evaluation- and behavior probabilities, respectively,

$$\left((\pi(a_{t,i} \mid s_{t,i}))_{t=0}^{H-1}\right)_{i=1}^n \quad \text{and} \quad \left((b(a_{t,i} \mid s_{t,i}))_{t=0}^{H-1}\right)_{i=1}^n.$$

Using this, we can calculate the corresponding returns $(G_{\tau_i})_{i=1}^n$ and ISQ $(q_{\tau_i})_{i=1}^n$. Then, approximate ρ^π via

$$\hat{\rho}^\pi_{\text{SIS}} \doteq \frac{1}{n}\sum_{i=1}^n G_{\tau_i} q_{\tau_i} \quad \text{or} \quad \hat{\rho}^\pi_{\text{WIS}} \doteq \frac{1}{\sum_{i=1}^n q_{\tau_i}} \sum_{i=1}^n G_{\tau_i} q_{\tau_i}.$$

They are called *simple (SIS)* and *weighted importance sampling (WIS)* estimators, respectively. They are consistent due to the law of large numbers. In order to see consistency of the second one, we expand the fraction with $1/n$ and use the fact that $\mathbb{E}_b[q_\tau] = 1$.

A major drawback of these estimators is that in many applications we do not have direct access to the behavior policy b. Even worse, the variance of $\hat{\rho}^\pi_{\text{SIS}}$ increases exponentially as $H \to \infty$. This is known as the curse of the horizon [42]. To see this, rewrite the ISQ as

$$q_\tau = \exp \sum_{t=0}^{H-1} \log q_t.$$

By the Central Limit Theorem, we have

$$\sum_{t=0}^{H-1} \log q_t \approx N(-H\mu, H\sigma^2), \quad \text{where} \quad \mu \doteq \mathbb{E}[\log q_t] \quad \text{and} \quad \sigma^2 \doteq \text{Var}[\log q_t].$$

This means that q_τ is asymptotically $\log N(-H\mu, H\sigma^2)$ with variance $e^{H\sigma^2} - 1$. The algorithms of the subsequent section apply IS differently and circumvent these issues.

3.5 Stationary DIstribution Correction Estimation

In many applications we are limited to data that was collected independently of any RL algorithm. On top of that, we are often not informed on the distribution of the data. In a more concrete manner, we are provided with a dataset of experience $\mathcal{D} = ((s_{0,i}, s_i, a_i, r_i, s'_i))_{i=1}^n$, where samples are drawn according to

$$s_0 \sim d_0,\ (s,a) \sim d^{\mathcal{D}},\ r \sim R(s,a,s'),\ s' \sim T(s,a), \quad \text{or, for short,} \quad (s_0, s, a, r, s') \sim p^{\mathcal{D}}.$$

We assume no prior knowledge of the distributions d_0, $d^{\mathcal{D}}$, R and T. Motivated by the lack of knowledge of our behavior policy and its stationary distribution $d^{\mathcal{D}}$, we call methods that work within this setting *behavior agnostic*. Even though we do not know $d^{\mathcal{D}}$ explicitly, we assume that $\mathcal{D}$ provides enough data, to include all the states and actions that we would visit under the evaluation policy π, as stated in Assumption 3.5.1.

Assumption 3.5.1. The stationary distribution d^π of the evaluation policy π is absolutely continuous with respect to the distribution $d^{\mathcal{D}}$ of the dataset $\mathcal{D}$, i.e., $d^\pi \ll d^{\mathcal{D}}$ or

$$d^\pi(s,a) > 0 \implies d^{\mathcal{D}}(s,a) > 0 \quad \text{for all} \quad (s,a) \in S \times A.$$

Many of these methods start off by a marginalized version of *importance sampling (IS)*. Using Assumption 3.5.1, applying IS to d^π in (3.4) from Lemma 3.1.3, we get

$$\rho^\pi = \mathbb{E}_{(s,a)\sim d^{\mathcal{D}}}[r(s,a) w_{\pi/\mathcal{D}}(s,a)], \quad \text{where} \quad w_{\pi/\mathcal{D}} = d^\pi / d^{\mathcal{D}}. \tag{3.14}$$

For any $(s,a) \in S \times A$, such that $d^{\mathcal{D}}(s,a) = 0$, we leave $w_{\pi/\mathcal{D}}(s,a)$ undefined. This works as long as we multiply it by zero every time we use it.

Approximating this *stationary distribution correction (SDC)* $w_{\pi/\mathcal{D}}$ by some $\hat{w}_{\pi/\mathcal{D}}$ is referred to as *(stationary) distribution correction estimation (DICE)*. We can use the empirical density $\hat{p}^{\mathcal{D}}$ to further approximate ρ^π by a simple Monte Carlo estimate or, since $\mathbb{E}_{(s,a)\sim d^{\mathcal{D}}}[w_{\pi/\mathcal{D}}(s,a)] = 1$, by a weighted Monte Carlo estimate, respectively,

$$\hat{\rho}^\pi_{\mathrm{S}} \doteq \mathbb{E}_{(s_0,s,a,r,s')\sim\hat{p}^{\mathcal{D}}}[r(s,a)\hat{w}_{\pi/\mathcal{D}}(s,a)] = \frac{1}{n}\sum_{i=1}^{n} r_i \hat{w}_{\pi/\mathcal{D}}(s_i,a_i), \tag{3.15}$$

$$\hat{\rho}^\pi_{\mathrm{W}} \doteq \frac{\mathbb{E}_{(s_0,s,a,r,s')\sim\hat{p}^{\mathcal{D}}}[r(s,a)\hat{w}_{\pi/\mathcal{D}}(s,a)]}{\mathbb{E}_{(s,a)\sim\hat{d}^{\mathcal{D}}}[\hat{w}_{\pi/\mathcal{D}}(s,a)]} = \frac{1}{\sum_{i=1}^{n}\hat{w}_{\pi/\mathcal{D}}(s_i,a_i)}\sum_{i=1}^{n} r_i \hat{w}_{\pi/\mathcal{D}}(s_i,a_i). \tag{3.16}$$

Let $\mathcal{P}^\pi \doteq \mathbb{E}_{T^\pi}$ and $\mathcal{P}^\pi_*$ denote the *expected Bellman operator* and its adjoint. Lemmas 3.2.4 and 3.2.5, the *forwards* and *backwards Bellman equations*, state that Q^π or V^π and d^π are fixed points of the *forwards* and *backwards Bellman operators* $\mathcal{B}^\pi$ and $\mathcal{T}^\pi$, respectively.

We want to formulate *modified backwards Bellman equations* for the stationary distribution correction $w_{\pi/\mathcal{D}}$. To this end, define

$$\mathcal{T}^\pi_d w \doteq D^{-1}\mathcal{T}^\pi D w = (1-\gamma)D^{-1}d^\pi_0 + \gamma D^{-1}\mathcal{P}^\pi_* D w,$$

for $\gamma \in (0,1]$, and $d : S \times A \to \mathbb{R}_{>0}$, where $w : S \times A \to \mathbb{R}$, as well as

$$Dw(s,a) = d(s,a)w(s,a) \quad \text{for} \quad (s,a) \in S \times A.$$

Now, we can define the *modified Bellman operator* $\mathcal{T}^\pi_{\mathcal{D}} \doteq \mathcal{T}^\pi_{d^{\mathcal{D}}}$. For any $(s,a) \in S \times A$, such that $d^{\mathcal{D}}(s,a) = 0$, we leave $\mathcal{T}^\pi_{\mathcal{D}} w(s,a)$ undefined.

Lemma 3.5.2 (modified backwards Bellman equations)**.** *Let $\gamma \in (0,1]$, then $w_{\pi/\mathcal{D}}$ is a solution to the modified backwards Bellman equation for $w : S \times A \to \mathbb{R}$:*

$$D^{\mathcal{D}} w = \mathcal{T}^\pi D^{\mathcal{D}} w = (1-\gamma)d^\pi_0 + \gamma \mathcal{P}^\pi_* D^{\mathcal{D}} w \quad \textit{and} \quad w = \mathcal{T}^\pi_{\mathcal{D}} w. \tag{3.17}$$

In case $\gamma = 1$, we need to add the normalization constraint

$$\mathbb{E}_{(s,a)\sim d^{\mathcal{D}}}[w(s,a)] = 1.$$

Proof. The first equation follows from the backwards Bellman equations (3.9) and the fact that $d^\pi = D^{\mathcal{D}} w_{\pi/\mathcal{D}}$. For the second equation, note that $w_{\pi/\mathcal{D}}$ is not defined iff $\mathcal{T}^\pi_{\mathcal{D}} w_{\pi/\mathcal{D}}$ is not defined. Otherwise, we can apply $(D^{\mathcal{D}})^{-1}$ to the first equation.

By definition of $w_{\pi/\mathcal{D}}$,

$$\mathbb{E}_{(s,a)\sim d^{\mathcal{D}}}[w_{\pi/\mathcal{D}}(s,a)] = \int_{S\times A} d^{\mathcal{D}}(s,a) w_{\pi/\mathcal{D}}(s,a)\, \mathrm{d}s\, \mathrm{d}a = \int_{S\times A} d^\pi(s,a)\, \mathrm{d}s\, \mathrm{d}a = 1.$$

□

Remark 3.5.3. For $\gamma < 1$, $w_{\pi/\mathcal{D}}$ is the unique solution to the modified backwards Bellman equations (3.17). For any $d : S \times A \to \mathbb{R}_{\geq 0}$, we can verify that $\mathcal{T}^\pi_d$ is a γ-contraction with respect to $\|\cdot\|_{L_1(d)}$. We calculate

$$\|\mathcal{T}^\pi_d w_1 - \mathcal{T}^\pi_d w_2\|_{L_1(d)} = \|\gamma D^{-1}\mathcal{P}^\pi_* D(w_1 - w_2)\|_{L_1(d)} \leq \gamma \|D^{-1}\mathcal{P}^\pi_* D\|_{L_1(d)} \|w_1 - w_2\|_{L_1(d)}.$$

Note, that although $\mathcal{T}_d^\pi$ and D^{-1} are not always well defined, since there may be $(s,a) \in S \times A$, such that $d(s,a) = 0$, this problem can be ignored within the norm $\|\cdot\|_{L_1(d)}$. By Lemma 3.2.3, the operator norm falls away.

■

Remark 3.5.4. For $\gamma = 1$, $w_{\pi/\mathcal{D}}$ is the unique solution to the modified backwards Bellman equations (3.17) given one of two conditions.

1. We include the constraint $w \geq 0$ and the stationary distribution is unique [44]. To see this, notice that $d \doteq D^{\mathcal{D}} w \geq 0$ is a distribution, because the normalization constraint leads to

$$\int_{S\times A} d(s,a)\ \mathrm{d}s\ \mathrm{d}a = \int_{S\times A} w(s,a) d^{\mathcal{D}}(s,a)\ \mathrm{d}s\ \mathrm{d}a = \mathbb{E}_{(s,a)\sim d^{\mathcal{D}}}[w(s,a)] = 1.$$

Additionally,

$$d = D^{\mathcal{D}} w = \mathcal{T}^\pi D^{\mathcal{D}} w = \mathcal{T}^\pi d = \mathcal{P}_*^\pi d.$$

This means that d is a stationary distribution, satisfying Assumption 3.5.1. Since we assume that it is unique, we get $d = d^\pi$ and hence

$$w = d/d^{\mathcal{D}} = d^\pi / d^{\mathcal{D}} = w_{\pi/\mathcal{D}}.$$

2. The state-action space is finite and Assumption 3.1.1 is satisfied [41]. Now, $\vec{D}^{\mathcal{D}}\vec{w}$ is an eigenvector of $\vec{\mathcal{P}}_*^\pi$ with eigenvalue 1. According to the backwards Bellman equations (3.9) this also holds for $\vec{d}^\pi$. By Assumption 3.1.1, our MDP is ergodic, so $\vec{\mathcal{P}}_*^\pi$ is a non-negative irreducible matrix. Hence, we can apply the Perron-Frobenius Theorem 2.4.1 to $\vec{\mathcal{P}}_*^\pi$. In this case, it says that the eigenspace of the eigenvalue 1 is one-dimensional and we get a scalar $\alpha \in \mathbb{R}$, such that

$$\vec{D}^{\mathcal{D}}\vec{w} = \alpha \vec{d}^\pi.$$

We use the fact that d^π is a distribution and the normalization constraint, to show that

$$\alpha = \alpha\langle \vec{d}^\pi, \vec{1}\rangle = \langle \vec{1}, \alpha\vec{d}^\pi\rangle = \langle \vec{1}, \vec{D}^{\mathcal{D}}\vec{w}\rangle = \langle \vec{D}^{\mathcal{D}}\vec{1}, \vec{w}\rangle = \langle \vec{D}^{\mathcal{D}}, \vec{w}\rangle = 1.$$

Hence, $\vec{D}^{\mathcal{D}}\vec{w} = \vec{d}^\pi$, i.e., $w = w_{\pi/\mathcal{D}}$.

■

We can derive a simple error bound for the policy value ρ^π and stationary distribution correction $w_{\pi/\mathcal{D}}$, analogous to Lemma 3.2.9.

Proposition 3.5.5. *Let $\gamma < 1$ and $d^{\mathcal{D}} > 0$. Then, for any $w : S \times A \to \mathbb{R}$,*

$$\left|\mathbb{E}_{(s,a)\sim d^{\mathcal{D}}}[r(s,a)w(s,a)] - \rho^\pi\right| \leq \|w_{\pi/\mathcal{D}} - w\|_{L_1(d^{\mathcal{D}})} \|r\|_{L_\infty(d^{\mathcal{D}})},$$

and

$$\|w_{\pi/\mathcal{D}} - w\|_{L_1(d^{\mathcal{D}})} \leq \frac{1}{1-\gamma} \|\mathcal{T}_{\mathcal{D}}^\pi w - w\|_{L_1(d^{\mathcal{D}})}.$$

Proof. For the first inequality, we use (3.14) and calculate

$$\begin{aligned}\left|\mathbb{E}_{(s,a)\sim d^{\mathcal{D}}}[r(s,a)w(s,a)] - \rho^\pi\right| &= \left|\mathbb{E}_{(s,a)\sim d^{\mathcal{D}}}[r(s,a)(w_{\pi/\mathcal{D}}(s,a) - w(s,a))]\right| \\ &\leq \|w_{\pi/\mathcal{D}} - w\|_{L_1(d^{\mathcal{D}})} \|r\|_{L_\infty(d^{\mathcal{D}})}.\end{aligned}$$

According to Remark 3.5.3, the modified backwards Bellman operator $\mathcal{T}_{\mathcal{D}}^{\pi}$ is a γ-contraction with respect to $\|\cdot\|_{L_1(d^{\mathcal{D}})}$. By the modified backwards Bellman equations (3.17), $w_{\pi/\mathcal{D}}$ is a fixed point. Together with the triangle inequality, this leads to

$$\begin{aligned}\|w_{\pi/\mathcal{D}} - w\|_{L_1(d^{\mathcal{D}})} &= \|\mathcal{T}_{\mathcal{D}}^{\pi} w_{\pi/\mathcal{D}} - w\|_{L_1(d^{\mathcal{D}})} \leq \|\mathcal{T}_{\mathcal{D}}^{\pi} w_{\pi/\mathcal{D}} - \mathcal{T}_{\mathcal{D}}^{\pi} w\|_{L_1(d^{\mathcal{D}})} + \|\mathcal{T}_{\mathcal{D}}^{\pi} w - w\|_{L_1(d^{\mathcal{D}})} \\ &\leq \gamma \|w_{\pi/\mathcal{D}} - w\|_{L_1(d^{\mathcal{D}})} + \|\mathcal{T}_{\mathcal{D}}^{\pi} w - w\|_{L_1(d^{\mathcal{D}})}.\end{aligned}$$

□

4 Algorithms

This chapter describes the theoretical background of the algorithms used to generate our numerical results. This includes

- `TabularVafe`, estimating the state-action value function Q^π with an approximate linear equation system of the forwards Bellman equations (3.5),
- `TabularDice`, estimating the stationary distribution correction $w_{\pi/\mathcal{D}}$ with an approximate linear equation system of the modified backwards Bellman equations (3.17),
- `TabularDualDice`, setting the gradient of the primal objective (4.3) of DualDICE to zero,
- `TabularGradientDice`, setting the gradient of the primal objective (4.5) of GradientDICE to zero,
- `NeuralDualDice`, performing stochastic gradient descent and ascent on the dual objective (4.6) of DualDICE,
- `NeuralGenDice`, performing stochastic gradient descent and ascent on the dual objective (4.7) of GenDICE,
- `NeuralGradientDice`, performing stochastic gradient descent and ascent on the dual objective (4.8) of GradientDICE,
- `NeuralCoinDice`, using theorem 4.6.10 and stochastic gradient descent and ascent to obtain approximations of policy value confidence intervals.

Using the methods described above, we give detailed derivations and discuss the theoretical significance of the hyperparameters. We also provide convergence and consistency proofs along with assumptions that support the theoretical justification of the algorithms.

4.1 Summary

This section provides a compact description of the algorithms, sufficient to interpret the numerical results at a high level. Further details are given in the subsequent sections.

We divide this summary section into two parts, where the state-action space $S \times A$ is finite and infinite, i.e., the *Tabular Case* and the *Continuous Case*, respectively.

4.1.1 Tabular Case

We can rewrite the modified Bellman equations (3.17) as

$$\begin{aligned} (1-\gamma)\vec{d}_0^{\pi} &= (I - \gamma \vec{\mathcal{P}}_*^{\pi})\vec{D}^{\mathcal{D}}\vec{w}_{\pi/\mathcal{D}} && \text{for} \quad 0 < \gamma < 1, \\ \vec{w}_{\pi/\mathcal{D}} &= (\vec{D}^{\mathcal{D}})^{-1}\vec{\mathcal{P}}_*^{\pi}\ \vec{D}^{\mathcal{D}}\vec{w}_{\pi/\mathcal{D}}, \quad \langle \vec{d}^{\mathcal{D}}, \vec{w}_{\pi/\mathcal{D}} \rangle = 1 && \text{for} \quad \gamma = 1. \end{aligned} \tag{4.1}$$

The second equation is only defined for $(s,a) \in S \times A$, such that $d^{\mathcal{D}}(s,a) > 0$. Accordingly, it is no problem that some diagonal elements of $\vec{D}^{\mathcal{D}}$ are zero and $(\vec{D}^{\mathcal{D}})^{-1}$ is not invertible.

For $\gamma < 1$ and from Assumption 3.5.1, we get unique solvability, assuming that $d^{\mathcal{D}} > 0$, as discussed in Remarks 3.2.6 and 3.2.7. For $\gamma = 1$ and Assumption 3.1.1, we can apply Lemma 3.2.3 and the Perron-Frobenius Theorem 2.4.1 to $\vec{\mathcal{P}}^{\pi}_{*}$.

Replacing by Law of Large Numbers estimates (4.9), (4.10), (4.11), and (4.12), we get some approximate modified Bellman equations. Using (4.13), (4.14), (4.15), and (4.16), we can write these as

$$\begin{aligned} (1-\gamma)\bar{d}_0^{\pi} &= (\bar{D}^{\mathcal{D}} - \gamma \bar{\mathcal{P}}^{\pi}_{1,*})\hat{w}_{\pi/\mathcal{D}} && \text{for} \quad 0 < \gamma < 1, \\ \hat{w}_{\pi/\mathcal{D}} &= (\bar{D}^{\mathcal{D}})^{-1}\bar{\mathcal{P}}^{\pi}_{1,*}\ \hat{w}_{\pi/\mathcal{D}}, \quad \langle \bar{d}^{\mathcal{D}}, \hat{w}_{\pi/\mathcal{D}} \rangle = n && \text{for} \quad \gamma = 1. \end{aligned} \tag{4.2}$$

Now, the primal objectives for DualDICE [34], GenDICE [44], and GradientDICE [41], respectively, are

$$\min_{v:S\times A\to\mathbb{R}} J_{\text{Dual}}(v), \qquad \min_{w:S\times A\to\mathbb{R}_{\geq 0}} J_{\text{Gen}}(w), \qquad \min_{w:S\times A\to\mathbb{R}} J_{\text{Gradient}}(w),$$

where

$$J_{\text{Dual}}(v) \doteq (1-\gamma)\mathbb{E}_{(s_0,a_0)\sim d_0^{\pi}}[v(s_0,a_0)] + \mathbb{E}_{(s,a)\sim d^{\mathcal{D}}}[\phi_*(\gamma \mathcal{P}^{\pi} v(s,a) - v(s,a))], \tag{4.3}$$

$$J_{\text{Gen}}(w) \doteq D_{\phi}(D^{\mathcal{D}} w \parallel \mathcal{T}^{\pi} D^{\mathcal{D}} w) + \frac{\lambda}{2}\left(\mathbb{E}_{(s,a)\sim d^{\mathcal{D}}}[w(s,a)] - 1\right)^2, \tag{4.4}$$

$$J_{\text{Gradient}}(w) \doteq \frac{1}{2}\|\mathcal{T}^{\pi}_{\mathcal{D}} w - w\|^2_{L_2(d^{\mathcal{D}})} \quad + \frac{\lambda}{2}\left(\mathbb{E}_{(s,a)\sim d^{\mathcal{D}}}[w(s,a)] - 1\right)^2. \tag{4.5}$$

The objectives (4.3) and (4.5) are unconstrained. Since they are all quadratic, setting their gradient to zero yields a linear equation system.

Choose $\phi \doteq \frac{1}{2}(\cdot)^2$ and perform the same replacements as before. This gives us the respective approximate linear equation systems

$$-\bar{A}_{1/2}\bar{A}_{1/2}^{\top}\hat{v} = (1-\gamma)\bar{d}_0^{\pi} \quad \text{and} \quad \hat{w}_{\pi/\mathcal{D}} = -(\bar{A}_{1/2}(\bar{D}^{\mathcal{D}})^{-1/2})^{\top}\hat{v},$$

$$\left(\bar{A}_1^{\top}(\bar{D}^{\mathcal{D}})^{-1}\bar{A}_1 + \frac{\lambda}{n}\bar{d}_0^{\pi}(\bar{d}_0^{\pi})^{\top}\right)\hat{w}_{\pi/\mathcal{D}} = (1-\gamma)\bar{A}_1^{\top}(\bar{D}^{\mathcal{D}})^{-1}\bar{d}_0^{\pi} + \lambda \bar{d}_0^{\pi}.$$

For $0 < \gamma < 1$, we also perform *value function estimation (VAFE)*, by using the approximate version of the forwards Bellman equations (3.5),

$$\hat{r} = (I - \gamma\hat{\mathcal{P}}^{\pi})\hat{Q}^{\pi}, \quad \hat{\rho}^{\pi} = (1-\gamma)\langle \hat{d}_0^{\pi}, \hat{Q}^{\pi}\rangle.$$

Some practical considerations lead us to two flaws in our algorithms, which both occur due to possible inadequacies of the underlying dataset. For each of these, we employ a heuristic, whose negative influence decreases as the quality of the dataset increases.

- It may not always be possible, to gather samples for all $(s,a) \in S \times A$ and assure that $\bar{D}^{\mathcal{D}}$ is invertible. There may appear $(s,a) \in S \times A$, where $\bar{d}^{\mathcal{D}}(s,a) = 0$. According to Assumption 3.5.1, the corresponding row and column inside $\bar{\mathcal{P}}^{\pi}$ should also be zero. By manually defining $0/0$, we can still work in this situation.

  However, in case Assumption 3.5.1 is not satisfied, we can project into the subspace $\mathbb{R}^{|\operatorname{supp}(\bar{d}^{\mathcal{D}})|} \leq \mathbb{R}^{|S\times A|}$, solve using our estimator, and then embed back into the original space. All further steps never use any values of our stationary distribution correction estimate $\hat{w}_{\pi/\mathcal{D}}(s,a)$, where $\bar{d}^{\mathcal{D}}(s,a) = 0$. Therefore, it does not matter anyways, what values we set them to. We chose -1 for error handling.

- We cannot guarantee that the Perron-Frobenius theorem 2.4.1 holds for the approximating matrix in the eigenvalue problem (4.2). Thus, we chose an eigenpair whose eigenvalue is closest to one and take the absolute value of the eigenvector.

4.1.2 Continuous Case

The objectives (4.3), (4.4), and (4.5) have expectations inside non-linear functions. This prevents us from performing SGD directly. However, applying Fenchel-Rockerfeller duality 2.2.10, respectively, yields [34, 44, 41]

$$\max_{v:S\times A\to\mathbb{R}} \min_{w:S\times A\to\mathbb{R}} J_{\text{Dual}}(v,w), \qquad \min_{w:S\times A\to\mathbb{R}_{\geq 0}} \max_{v:S\times A\to\mathbb{R},\ u\in\mathbb{R}} J_{\text{Gen}}(w,v,u),$$
$$\min_{w:S\times A\to\mathbb{R}} \max_{v:S\times A\to\mathbb{R},\ u\in\mathbb{R}} J_{\text{Gradient}}(w,v,u),$$

where

$$J_{\text{Dual}}(v,w) \doteq \mathbb{E}_{(s_0,s,a,r,s')\sim p^{\mathcal{D}}}\Big[L(v,w;s_0,s,a,s') + \phi(w(s,a))\Big], \tag{4.6}$$
$$J_{\text{Gen}}(w,v,u) \doteq \mathbb{E}_{(s_0,s,a,r,s')\sim p^{\mathcal{D}}}\Big[L(v,w;s_0,s,a,s') + N(w,u;s,a) - \frac{1}{4}v(s,a)^2 w(s,a)\Big], \tag{4.7}$$
$$J_{\text{Gradient}}(w,v,u) \doteq \mathbb{E}_{(s_0,s,a,r,s')\sim p^{\mathcal{D}}}\Big[L(v,w;s_0,s,a,s') + N(w,u;s,a) - \frac{1}{2}v(s,a)^2\Big], \tag{4.8}$$

$$L(v,w;s_0,s,a,s') \doteq (1-\gamma)\mathbb{E}_{a_0\sim\pi(s_0)}[v(s_0,a_0)] + w(s,a)\left(\gamma\mathbb{E}_{a'\sim\pi(s')}[v(s',a')] - v(s,a)\right),$$
$$\phi(x) \doteq \frac{1}{p}|x|^p, \quad p>1, \quad N(w,u;s,a) \doteq \lambda\Big(u(w(s,a)-1) - \frac{1}{2}u^2\Big), \quad \lambda>0.$$

In order to parameterize v and w, we use neural networks v_ϑ and w_θ, respectively, with a single hidden layer. To provide non-negativity, we add $(\cdot)^2$ to the final layer [44].

Every 100 steps, we store the policy value and the average loss. If we have an analytical solution for the the policy value ρ^π and stationary distribution correction $w_{\pi/\mathcal{D}}$, we also store the error $|\hat{\rho}^\pi - \rho^\pi|$ and the MSE $\mathbb{E}_{(s,a)\sim d^{\mathcal{D}}}|w_\vartheta - w_{\pi/\mathcal{D}}|^2$. In addition, the *Bellman residual angle (BRA)* between the gradients $\nabla_\vartheta \mathcal{P}^\pi v_\vartheta$ and $\nabla_\vartheta v_\vartheta$ is stored.

4.2 Tabular stationary DIstribution Correction Estimation

Let the state-action space be finite. Define the matrices

$$\vec{\mathcal{P}}^\pi_{p,*} \doteq \vec{\mathcal{P}}^\pi_*(\vec{D}^{\mathcal{D}})^p \quad \text{and} \quad \vec{A}_p \doteq \vec{A}(\vec{D}^{\mathcal{D}})^p, \quad \text{where} \quad p\in\mathbb{R}.$$

Since $\vec{D}$ is a diagonal matrix, its powers can be defined via component wise application. Using the Law of Large Number approximations for $d_0(s_0)$, $T(s'\mid s,a)$, and $d^{\mathcal{D}}(s,a)$, define

$$\hat{d}^\pi_0 \doteq \left(\pi(a_0\mid s_0)\frac{1}{n}\sum_{i=1}^n \mathbb{1}_{s_0=s_{0,i}}\right)_{(s_0,a_0)\in S\times A}, \tag{4.9}$$
$$\hat{d}^{\mathcal{D}} \doteq \left(\frac{1}{n}\sum_{i=1}^n \mathbb{1}_{s=s_i,\ a=a_i}\right)_{(s,a)\in S\times A}, \quad \hat{D}^{\mathcal{D}} \doteq \operatorname{diag}(\hat{d}^{\mathcal{D}}), \tag{4.10}$$
$$\hat{\mathcal{P}}^\pi_* \doteq \left(\pi(a'\mid s')\frac{\frac{1}{n}\sum_{i=1}^n \mathbb{1}_{s=s_i,\ a=a_i,\ s'=s'_i}}{\frac{1}{n}\sum_{i=1}^n \mathbb{1}_{s=s_i,\ a=a_i}}\right)_{(s',a'),(s,a)\in S\times A}, \quad \hat{A} \doteq I - \gamma\hat{\mathcal{P}}^\pi_*, \tag{4.11}$$
$$\hat{\mathcal{P}}^\pi_{p,*} \doteq \hat{\mathcal{P}}^\pi_*(\hat{D}^{\mathcal{D}})^p. \quad \hat{A}_p \doteq \hat{A}(\hat{D}^{\mathcal{D}})^p, \tag{4.12}$$

Notice that

$$\hat{d}_0^\pi = \frac{1}{n}\bar{d}_0^\pi, \qquad \hat{\mathcal{P}}_*^\pi = \bar{\mathcal{P}}_{1,*}^\pi (\bar{D}^\mathcal{D})^{-1}, \qquad \hat{\mathcal{P}}_{p,*}^\pi = \frac{1}{n^p}\bar{\mathcal{P}}_{p,*}^\pi,$$
$$\hat{D}^\mathcal{D} = \frac{1}{n}\bar{D}^\mathcal{D}, \qquad \hat{A} = \bar{A}_1 (\bar{D}^\mathcal{D})^{-1}, \qquad \hat{A}_p = \frac{1}{n^p}\bar{A}_p,$$

where

$$\bar{d}_0^\pi \doteq \sum_{i=1}^n \left(\pi(a_0 \mid s_{0,i}) \mathbb{1}_{s_0 = s_{0,i}}\right)_{(s_0,a_0)\in S\times A}, \tag{4.13}$$

$$\bar{d}^\mathcal{D} \doteq \sum_{i=1}^n \left(\mathbb{1}_{s=s_i,\ a=a_i}\right)_{(s,a)\in S\times A}, \quad \bar{D}^\mathcal{D} \doteq \operatorname{diag}(\bar{d}^\mathcal{D}), \tag{4.14}$$

$$\bar{\mathcal{P}}_*^\pi \doteq \sum_{i=1}^n \left(\pi(a' \mid s_i') \mathbb{1}_{s=s_i,\ a=a_i,\ s'=s_i'}\right)_{(s',a'),(s,a)\in S\times A}, \quad \bar{A} \doteq \bar{D}^\mathcal{D} - \gamma \bar{\mathcal{P}}_*^\pi, \tag{4.15}$$

$$\bar{\mathcal{P}}_{p,*}^\pi \doteq \bar{\mathcal{P}}_{1,*}^\pi (\bar{D}^\mathcal{D})^{p-1}, \quad \bar{A}_p \doteq \bar{A}_1 (\bar{D}^\mathcal{D})^{p-1}. \tag{4.16}$$

Let $x : S \times A \to \mathbb{R}^K$ be a feature function and consider the feature matrix

$$X \doteq (\vec{x}_1, \dots, \vec{x}_K) \in \mathbb{R}^{|S\times A|\times K}. \tag{4.17}$$

From this, we can derive the linear parameterization function space

$$\mathcal{F} \doteq \left\{ \vec{w}_\theta \doteq X\vec{\theta} \;\middle|\; \vec{\theta} \in \mathbb{R}^K \right\}. \tag{4.18}$$

In order to approximate the stationary distribution correction exactly, we would need to chose the feature matrix such that assumption 4.2.1 is satisfied. Note, that the choice $X \doteq I$ corresponds do a one-hot-encoding. In particular, this choice of feature matrix would satisfy the assumption 4.2.1.

Assumption 4.2.1. The stationary distribution correction $\vec{w}_{\pi/\mathcal{D}}$ is part of the column span of the feature matrix X.

Oftentimes, our feature matrix is not representative enough. An algorithm working with the feature matrix X will, at best, approximate the projection $Pw_{\pi/\mathcal{D}}$ of the stationary distribution correction onto the range of X. Finally, Remark 2.4.4 tells us that this projection can be expressed as

$$P = X(X^\top D^\mathcal{D} X)^{-1} X^\top D^\mathcal{D}.$$

4.3 Dual stationary DIstribution Correction Estimation

Using the Q-LP from Lemma 3.3.1 directly as a way of policy evaluation bears a major problem for applications, where the state-action space is infinite. To circumvent this issue, Nachum et al. [34] introduce the algorithm *DualDICE*.

We merge the contents of Nachum et al. [34] and [31] to create a better understanding of DualDICE. On top of that, we include all the technical details, omitted by Nachum et al. [34].

4.3.1 Objectives

Primal Objective

We consider the dual Q-LP (3.11). Since it is over-constrained, the optimal solution does not change if we substitute the objective by the ϕ-divergence $-D_\phi(d \parallel d^{\mathcal{D}})$. Even though this has no impact on the divergence, we will assume that ϕ is defined for all real numbers. We get

$$\begin{aligned} \max_{d:S\times A\to\mathbb{R}_{\geq 0}} \; & -D_\phi(d \parallel d^{\mathcal{D}}) \\ \text{s.t.} \quad & \forall (s,a) \in S \times A: \; d(s,a) = (1-\gamma) d_0^\pi(s,a) + \gamma \mathcal{P}_*^\pi d(s,a). \end{aligned}$$

Now, this is not an LP anymore, since the objective is not linear any more, but merely convex. However, it is a Fenchel optimization problem

$$\begin{gathered} -\min_d f(d) + g(Ad), \quad \text{where} \\ f(d) \doteq D_\phi(d \parallel d^{\mathcal{D}}), \quad g = \delta_{(1-\gamma)d_0^\pi}, \quad A = I - \gamma \mathcal{P}_*^\pi. \end{gathered}$$

Building the dual, we get

$$\begin{gathered} -\max_v -f_*(-A_* v) - g_*(v), \quad \text{where} \\ f_* = \mathbb{E}_{d^{\mathcal{D}}}[\phi_*(\cdot)], \quad g_* = (1-\gamma)\mathbb{E}_{d_0^\pi}[\cdot], \quad A_* = I - \gamma \mathcal{P}^\pi, \end{gathered}$$

resulting in

$$\min_{v:S\times A\to\mathbb{R}} J(v) \doteq (1-\gamma)\mathbb{E}_{(s_0,a_0)\sim d_0^\pi}[v(s_0,a_0)] + \mathbb{E}_{(s,a)\sim d^{\mathcal{D}}}[\phi_*(\gamma \mathcal{P}^\pi v(s,a) - v(s,a))].$$

Dual Objective

Unrolling the definition of the convex conjugate ϕ_* in the second expectation of $J(v)$, we get

$$\begin{aligned} &\mathbb{E}_{(s,a)\sim d^{\mathcal{D}}}[\phi_*(\gamma \mathcal{P}^\pi v(s,a) - v(s,a))] \\ &\quad = \mathbb{E}_{(s,a)\sim d^{\mathcal{D}}}[\max_{w\in\mathbb{R}} (\gamma \mathcal{P}^\pi v(s,a) - v(s,a)) w - \phi(w)] \\ &\quad = \max_{w:S\times A\to\mathbb{R}} \mathbb{E}_{(s,a)\sim d^{\mathcal{D}}}[(\gamma \mathcal{P}^\pi v(s,a) - v(s,a)) w(s,a) - \phi(w(s,a))] \\ &\quad = \max_{w:S\times A\to\mathbb{R}} \gamma \mathbb{E}_{(s,a)\sim d^{\mathcal{D}},\; (s',a')\sim T^\pi(s,a)}[v(s',a') w(s,a)] \\ &\qquad\quad - \mathbb{E}_{(s,a)\sim d^{\mathcal{D}}}[v(s,a) w(s,a)] \\ &\qquad\quad - \mathbb{E}_{(s,a)\sim d^{\mathcal{D}}}[\phi(w(s,a))]. \end{aligned}$$

This leads to the dual objective

$$\begin{aligned} &\min_{v:S\times A\to\mathbb{R}} \max_{w:S\times A\to\mathbb{R}} J(v,w) \\ &\quad \doteq (1-\gamma)\mathbb{E}_{(s_0,a_0)\sim d_0^\pi}[v(s_0,a_0)] \\ &\qquad + \mathbb{E}_{(s,a)\sim d^{\mathcal{D}},\; (s',a')\sim T^\pi(s,a)}\left[w(s,a)\left(\gamma v(s',a') - v(s,a)\right)\right] \\ &\qquad - \mathbb{E}_{(s,a)\sim d^{\mathcal{D}}}[\phi(w(s,a))]. \end{aligned}$$

4.3.2 The Dual Variable

For any function $v : S \times A \to \mathbb{R}$, consider the change of variables $x = -A_* v$. According to Remark 3.2.7, A_* is bijective, so we can also reverse this substitution. Now, calculate

$$\begin{aligned}
&\mathbb{E}_{(s,a)\sim d^\pi}[x(s,a)] \\
&\quad = \mathbb{E}_{(s,a)\sim d^\pi}\left[\gamma \mathbb{E}_{(s',a')\sim T^\pi(s,a)}[v(s',a')] - v(s,a)\right] \\
&\quad = (1-\gamma)\sum_{t=0}^{\infty} \gamma^t \mathbb{E}_{(s,a)\sim d_t^\pi}\left[\gamma \mathbb{E}_{(s',a')\sim T^\pi(s,a)}[v(s',a')] - v(s,a)\right] \\
&\quad = (1-\gamma)\sum_{t=0}^{\infty} \gamma^{t+1} \mathbb{E}_{(s',a')\sim d_{t+1}^\pi}[v(s',a')] - (1-\gamma)\sum_{t=0}^{\infty} \gamma^t \mathbb{E}_{(s,a)\sim d_t^\pi}[v(s,a)] \\
&\quad = -(1-\gamma)\mathbb{E}_{(s_0,a_0)\sim d_0^\pi}[v(s_0,a_0)].
\end{aligned}$$

This lets us reformulate the primal objective

$$\begin{aligned}
J(v) &= \mathbb{E}_{(s,a)\sim d^{\mathcal{D}}}[\phi_*(x(s,a))] - \mathbb{E}_{(s,a)\sim d^\pi}[x(s,a)] \\
&= \sum_{(s,a)\in S\times A} \phi_*(x(s,a)) d^{\mathcal{D}}(s,a) - x(s,a) d^\pi(s,a).
\end{aligned}$$

By taking the derivative with respect to x and setting it to zero, we gather that for the optimal x^*, we get

$$\phi_*'(-A_* v^*) = \phi_*'(x^*) = d^\pi / d^{\mathcal{D}} = w_{\pi/\mathcal{D}}.$$

Alternatively, we can use the optimal solution v^* of the dual Fenchel optimization problem to derive that for the solution d^* of the primal we have

$$d^\pi = d^* = f_*'(-A_* v^*) = d^{\mathcal{D}} \phi_*'(-A_* v^*).$$

For a fixed $v : S \times A \to \mathbb{R}$, taking the derivative of $J(v, w)$ with respect to w and setting it to zero, we gather that

$$\phi'(w_v^*) = \gamma \mathcal{P}^\pi v - v = -A_* v.$$

By Lemma 2.2.9, we have $(\phi')^{-1} = \phi_*'$, so

$$w^* = \phi_*'(-A_* v^*) = d^\pi / d^{\mathcal{D}} = w_{\pi/\mathcal{D}}. \tag{4.19}$$

4.3.3 Implementation

Tabular

For a finite state-action space, and $\phi \doteq \frac{1}{2}(\cdot)^2$, we can rewrite the primal objective with vector notation as

$$J(v) = (1-\gamma)\langle \vec{v}, \vec{d_0^\pi} \rangle + \frac{1}{2} \|\vec{A}^\top \vec{v}\|^2_{\ell_2(d^{\mathcal{D}})}.$$

Using Remark 2.4.2, we can build the gradient

$$\nabla_v J(v) = (1-\gamma)\vec{d_0^\pi} + \vec{A}\vec{D}^{\mathcal{D}}\vec{A}^\top \vec{v}.$$

Setting the gradient to zero, we get the linear equation system

$$-\vec{A}_{1/2}\vec{A}_{1/2}^{\top}\vec{v}^{*} = (1-\gamma)\vec{d}_0^{\pi}.$$

Using (4.19), and $\phi_* = \frac{1}{2}(\cdot)^2$ from Example 2.2.5, we get

$$\vec{w}_{\pi/\mathcal{D}} = -\vec{A}^{\top}\vec{v}^{*}.$$

Considering the law of large numbers approximations $\hat{d}_0^{\pi}$, $\hat{A}$, and $\hat{D}^{\mathcal{D}}$, $\hat{A}_p$, from (4.9), (4.10), (4.11), and (4.12), respectively, we get the approximate linear equation system,

$$-\hat{A}_{1/2}\hat{A}_{1/2}^{\top}\hat{v} = (1-\gamma)\hat{d}_0^{\pi} \quad \text{and} \quad \hat{w}_{\pi/\mathcal{D}} = -\hat{A}^{\top}\hat{v}.$$

Continuous

We specify $\phi \doteq \frac{1}{p}(\cdot)^p$, for some $p > 1$. For v and w we use the function classes $\mathcal{F}_v$ and $\mathcal{F}_w$, respectively, e.g. by parameterization using neural networks. Now, we have access to samples taken with respect to $p^{\mathcal{D}}$ and even explicit access to π. This lets us perform SGD on v and SGA on w. Since we have a saddle point problem, we use SGDA, i.e., each time we sample a batch of experience to approximate the gradient, we perform a gradient step on both v and w in parallel. Building the gradients, we get

$$\begin{aligned}
\nabla_{\vartheta} J(v_\vartheta, w_\theta) &= (1-\gamma)\mathbb{E}_{(s_0,a_0)\sim d_0^{\pi}}[\nabla_{\vartheta} v_\vartheta(s_0,a_0)] \\
&\quad + \mathbb{E}_{(s,a)\sim d^{\mathcal{D}},\ (s',a')\sim T^{\pi}(s,a)}[w_\theta(s,a)(\gamma\nabla_{\vartheta} v_\vartheta(s',a') - \nabla_{\vartheta} v_\vartheta(s,a))], \\
\nabla_{\theta} J(v_\vartheta, w_\theta) &= \mathbb{E}_{(s,a)\sim d^{\mathcal{D}},\ (s',a')\sim T^{\pi}(s,a)}[(\gamma v_\vartheta(s',a') - v_\vartheta(s,a))\nabla_{\theta} w_\theta(s,a)] \\
&\quad - \mathbb{E}_{(s,a)\sim d^{\mathcal{D}}}[w_\theta(s,a)^{p-1}\nabla_{\theta} w_\theta(s,a)].
\end{aligned}$$

4.3.4 Convergence

Parameterizing v and w in $J(v,w)$, e.g. by neural networks, induces an inherent approximation error $\epsilon_{\text{approx}}(\mathcal{F}_v, \mathcal{F}_w)$. We can apply SGDA to $J(v,w)$, and obtain an approximation $\hat{w}_{\pi/\mathcal{D}}$ of $w_{\pi/\mathcal{D}}$. Let ϵ_{opt} denote the error, we get from SGDA. Finally, we can get the approximation $\hat{\rho}^{\pi}$ of ρ^{π}, as in (3.15). The MSE of $\hat{\rho}^{\pi}$ is discussed in Theorem 4.3.3.

Assumption 4.3.1. The stationary distribution correction is uniformly bounded, i.e.,

$$\exists C_w > 0 : \quad \sup_{(s,a)\in S\times A} |w_{\pi/\mathcal{D}}(s,a)| \leq C_w.$$

Assumption 4.3.2. The observed reward $\hat{r}$ is uniformly bounded, i.e.,

$$\exists C_r > 0 : \quad \sup_{(s,a)\in S\times A} |\hat{r}(s,a)| \leq C_r.$$

Theorem 4.3.3. *Let Assumptions 4.3.1 and 4.3.2 hold. Also, choose $\phi \doteq \frac{1}{2}(\cdot)^2$. Then, the MSE of DualDICE's estimate is bounded by*

$$\mathbb{E}|\hat{\rho}^{\pi} - \rho^{\pi}|^2 = \mathcal{O}_{log}\left(\epsilon_{approx}(\mathcal{F}_v, \mathcal{F}_w) + \epsilon_{opt} + \frac{1}{\sqrt{n}}\right).$$

The expectation is taken with respect to randomness both in the sampling of $\mathcal{D} \sim p^{\mathcal{D}}$ and in the algorithm. $\mathcal{O}_{log}$ simply ignores logarithmic factors.

4.4 Generalized stationary DIstribution Correction Estimation

A major downside of DualDICE is that it only works in the discounted setting. On top of that, pushing the discount factor γ towards 1 has a negative influence on the estimators accuracy [44]. The algorithm *GenDICE* by Zhang et al. [44] aims to also include the undiscounted setting in its policy value approximation and improve stability for higher discount factors.

4.4.1 Objectives

Primal Objective

The idea behind GenDICE is to find the stationary distribution correction $w_{\pi/\mathcal{D}}$, by starting off with the modified backwards Bellman equations (3.17). A naive approach to the solution of these equations in the continuous setting is to use a positive definite discrimination function $D(\cdot \parallel \cdot)$ and consider the optimization problem

$$\min_{w:S\times A\to\mathbb{R}_{\geq 0}} D(D^{\mathcal{D}}w \parallel \mathcal{T}^{\pi}D^{\mathcal{D}}w).$$

Now, $w_{\pi/\mathcal{D}}$ would indeed be a solution, but any scaled version $cw_{\pi/\mathcal{D}}$ by a constant $c \geq 0$ also solves the problem. In particular, the trivial degenerate solution $w \equiv 0$ cannot be ruled out. Therefore, consider the *norm penalization coefficient* $\lambda > 0$ and the optimization problem

$$\min_{w:S\times A\to\mathbb{R}_{\geq 0}} J(w) \doteq D(D^{\mathcal{D}}w \parallel \mathcal{T}^{\pi}D^{\mathcal{D}}w) + \frac{\lambda}{2}\left(\mathbb{E}_{(s,a)\sim d^{\mathcal{D}}}[w(s,a)] - 1\right)^2.$$

According to Remarks 3.5.3 and 3.5.4, the stationary distribution correction $w_{\pi/\mathcal{D}}$ is the only solution to $J(w) = 0$.

Dual Objective

However, since we only have access to samples of $\mathcal{T}^{\pi}D^{\mathcal{D}}$, we cannot evaluate it at arbitrary points. Therefore, the objective becomes intractable. In order to make our objective tractable, we must further specify our discrimination function. Using a ϕ-divergence,

$$D_{\phi}(\mathcal{T}^{\pi}D^{\mathcal{D}}w \parallel d^{\mathcal{D}}w) = \int_{S\times A} d^{\mathcal{D}}(s,a)w(s,a)\phi\left(\frac{\mathcal{T}^{\pi}D^{\mathcal{D}}w(s,a)}{d^{\mathcal{D}}(s,a)w(s,a)}\right)\,\mathrm{d}s\,\mathrm{d}a.$$

We apply the Fenchel–Moreau Theorem 2.2.10 to ϕ and get

$$\begin{aligned}
&\int_{S\times A} d^{\mathcal{D}}(s',a')w(s',a') \max_{v\in\mathbb{R}} \left(\frac{\mathcal{T}^\pi D^{\mathcal{D}} w(s',a')}{d^{\mathcal{D}}(s',a')w(s',a')} v - \phi_*(v) \right) \mathrm{d}s'\ \mathrm{d}a' \\
&\quad = \max_{v:S\times A\to\mathbb{R}} \int_{S\times A} \mathcal{T}^\pi D^{\mathcal{D}} w(s',a')v(s',a') - d^{\mathcal{D}}(s',a')w(s',a')\phi_*(v(s',a'))\ \mathrm{d}s'\ \mathrm{d}a' \\
&\quad = \max_{v:S\times A\to\mathbb{R}} \int_{S\times A} \Bigg((1-\gamma) d_0(s')\pi(a' \mid s') \\
&\qquad\quad + \gamma \int_{S\times A} T^\pi(s',a' \mid s,a) d^{\mathcal{D}}(s,a) w(s,a)\ \mathrm{d}s\ \mathrm{d}a \Bigg) v(s',a') \\
&\qquad\quad - d^{\mathcal{D}}(s',a')w(s',a')\phi_*(v(s',a'))\ \mathrm{d}s'\ \mathrm{d}a' \\
&\quad = \max_{v:S\times A\to\mathbb{R}} (1-\gamma) \int_{S\times A} v(s',a') d_0^\pi(s',a')\ \mathrm{d}s'\ \mathrm{d}a' \\
&\qquad\quad + \gamma \int_{S\times A} \int_{S\times A} w(s,a)v(s',a')T^\pi(s',a' \mid s,a) d^{\mathcal{D}}(s,a)\ \mathrm{d}s\ \mathrm{d}a\ \mathrm{d}s'\ \mathrm{d}a' \\
&\qquad\quad - \int_{S\times A} w(s',a')\phi_*(v(s',a'))d^{\mathcal{D}}(s',a')\ \mathrm{d}s'\ \mathrm{d}a' \\
&\quad = \max_{v:S\times A\to\mathbb{R}} (1-\gamma)\mathbb{E}_{(s,a)\sim d_0^\pi}[v(s,a)] \\
&\qquad\quad + \gamma \mathbb{E}_{(s,a)\sim d^{\mathcal{D}},\ (s',a')\sim T^\pi(s,a)}[w(s,a)v(s',a')] \\
&\qquad\quad - \mathbb{E}_{(s,a)\sim d^{\mathcal{D}}}[w(s,a)\phi_*(v(s,a))].
\end{aligned}$$

Also applying it to $\frac{1}{2}(\cdot)^2$ gives us a dual objective

$$\begin{aligned}
&\min_{w:S\times A\to\mathbb{R}_{\geq 0}} \max_{v:S\times A\to\mathbb{R},\ u\in\mathbb{R}} J(w,v,u) \\
&\qquad \doteq (1-\gamma)\mathbb{E}_{(s_0,a_0)\sim d_0^\pi}[v(s_0,a_0)] \\
&\qquad\qquad + \gamma \mathbb{E}_{(s,a)\sim d^{\mathcal{D}},\ (s',a')\sim T^\pi(s,a)}[w(s,a)v(s',a')] \\
&\qquad\qquad - \mathbb{E}_{(s,a)\sim d^{\mathcal{D}}}[w(s,a)\phi_*(v(s,a))] \\
&\qquad\qquad + \lambda \left(\mathbb{E}_{(s,a)\sim d^{\mathcal{D}}}[uw(s,a) - u] - \frac{1}{2}u^2 \right).
\end{aligned}$$

4.4.2 Implementation

Continuous

The χ^2-divergence was chosen as an f-divergence. By Example 2.2.12, this results in

$$\begin{aligned}
&\min_{w:S\times A\to\mathbb{R}_{\geq 0}} \max_{v:S\times A\to\mathbb{R},\ u\in\mathbb{R}} J(w,v,u) \\
&\quad = (1-\gamma)\mathbb{E}_{(s_0,a_0)\sim d_0^\pi}[v(s_0,a_0)] \\
&\qquad + \mathbb{E}_{(s,a)\sim d^{\mathcal{D}},\ (s',a')\sim T^\pi(s,a)}\left[w(s,a)\left(\gamma v(s',a') - v(s,a)\right)\right] \\
&\qquad + \frac{1}{4}\mathbb{E}_{(s,a)\sim d^{\mathcal{D}}}\left[w(s,a)v(s,a)^2\right] \\
&\qquad + \lambda \left(\mathbb{E}_{(s,a)\sim d^{\mathcal{D}}}[uw(s,a) - u] - \frac{1}{2}u^2 \right).
\end{aligned}$$

The functions ϕ_* and $(\cdot)^2$ are convex. Therefore, the objective $J(w, v, u)$ is convex in w and concave in v and u and we have a *convex-concave saddle-point problem (CCSP)*. Recall the estimation from DualDICE. For w and v, We will use function spaces $\mathcal{F}_w$ and $\mathcal{F}_v$, respectively. The variables w and v are parameterized by neural networks. Call the parameters $\theta, \vartheta \in \mathbb{R}^K$, respectively. To assure that the first network only outputs non-negative values, an extra positive neuron was added to the end, such as

$$\exp(\cdot), \quad \log(1+\exp(\cdot)), \quad \text{or} \quad (\cdot)^2. \tag{4.20}$$

Building the gradients, we get

$$\begin{aligned}
\nabla_\theta J(w_\theta, v_\vartheta, u) &= \mathbb{E}_{(s,a)\sim d^{\mathcal{D}},\ (s',a')\sim T^\pi(s,a)}[(\gamma v_\vartheta(s',a') - v_\vartheta(s,a))\nabla_\theta w_\theta(s,a)] \\
&\quad - \frac{1}{4}\mathbb{E}_{(s,a)\sim d^{\mathcal{D}}}[v_\vartheta(s,a)^2 \nabla_\theta w_\theta(s,a)] \\
&\quad + \lambda u \mathbb{E}_{(s,a)\sim d^{\mathcal{D}}}[w_\theta(s,a)], \\
\nabla_\vartheta J(w_\theta, v_\vartheta, u) &= (1-\gamma)\mathbb{E}_{(s_0,a_0)\sim d_0^\pi}[\nabla_\vartheta v_\vartheta(s_0,a_0)] \\
&\quad + \mathbb{E}_{(s,a)\sim d^{\mathcal{D}},\ (s',a')\sim T^\pi(s,a)}\left[w_\theta(s,a)\left(\gamma \nabla_\vartheta v_\vartheta(s',a') - \nabla_\vartheta v_\vartheta(s,a)\right)\right] \\
&\quad - \frac{1}{2}\mathbb{E}_{(s,a)\sim d^{\mathcal{D}}}[w_\theta(s,a) v_\vartheta(s,a) \nabla_\vartheta v_\vartheta(s,a)], \\
\nabla_u J(w_\theta, v_\vartheta, u) &= \lambda\left(\mathbb{E}_{(s,a)\sim d^{\mathcal{D}}}[w_\theta(s,a) - 1] - u\right).
\end{aligned}$$

4.4.3 Convergence

Since we now also include the undiscounted setting $\gamma = 1$, we have to include the additional Assumption 4.4.1.

Assumption 4.4.1 (Markov chain regularity)**.** The backwards Bellman equations (3.9) have a unique solution, i.e., the stationary distribution d^π exists and is unique.

Theorem 4.4.2. *Let Assumptions 4.3.1, 4.3.2, and 4.4.1 hold. Also, let ϕ_* be Lipschitz-continuous, let the psuedo-dimension of $\mathcal{F}_w$ and $\mathcal{F}_v$ be bounded and*

$$\exists C_{\mathcal{F}_w} > 0 : \ \forall w \in \mathcal{F}_w : \ \|w\|_\infty \leq C_{\mathcal{F}_w}.$$

Then, the error between GenDICE's estimate $\hat{w}_{\pi/\mathcal{D}}$ and $w_{\pi/\mathcal{D}}$ is bounded by

$$\mathbb{E}\left[J(\hat{w}_{\pi/\mathcal{D}}) - J(w_{\pi/\mathcal{D}})\right] = \mathcal{O}_{log}\left(\epsilon_{approx}(\mathcal{F}_w, \mathcal{F}_v) + \epsilon_{opt} + \frac{1}{\sqrt{n}}\right).$$

The expectation is taken with respect to randomness both in the sampling of $\mathcal{D} \sim p^{\mathcal{D}}$ and in the algorithm. $\mathcal{O}_{log}$ simply ignores logarithmic factors.

4.5 Gradient stationary DIstribution Correction Estimation

GenDICE fixes some of the problems that DualDICE has. However, in doing so, it introduces other problems [41].

Firstly, note that f-divergences are originally defined only for probability distributions. Extending the inputs of D_ϕ to generic functions will cause it to lose non-negativity. For example, if

$q > p > 0$, then $D_{\mathrm{KL}}(p \,\|\, q) < 0$. However, as long as $\min\{p, q\} > 0$, we still have $D_{\chi^2}(p \,\|\, q) \geq 0$, which is fortunate, since GenDICE actually uses this f-divergence.

Nevertheless, another problem arises when using the extra non-linear positive neuron (4.20), to ensure $w_\theta \geq 0$. Since objective $J(w, v, u)$ is not necessarily non-decreasing in each w_i, we cannot assure that $J(w_\theta, v_\vartheta, u)$ is convex in θ, even if the extra positive neuron is convex [12, p. 86].

Based on the ideas of GenDICE, Zhang et al. [41] present the algorithm *GradientDICE*. It tries to solve the problems of GenDICE mentioned above, by using the $L_2(d^{\mathcal{D}})$-norm instead of the χ^2-divergence, thereby removing the the necessity of the constraint $w \geq 0$.

Not only do we cover the algorithm as in this section, we also mention important details, omitted by Zhang et al. [41].

4.5.1 Objectives

Primal Objective

The primal objective that GradientDICE uses is similar to GenDICE,

$$\min_{w: S \times A \to \mathbb{R}} J(w) \doteq \frac{1}{2} \| \mathcal{T}_{\mathcal{D}}^{\pi} w - w \|_{L_2(d^{\mathcal{D}})}^2 + \frac{\lambda}{2} \left(\mathbb{E}_{(s,a) \sim d^{\mathcal{D}}} [w(s,a)] - 1 \right)^2 .$$

Dual Objective

We apply the Fenchel–Moreau Theorem 2.2.10 to $\frac{1}{2}(\cdot)^2$ and rewrite the first summand as

$$\begin{aligned}
\frac{1}{2} \| \mathcal{T}_{\mathcal{D}}^{\pi} w - w \|_{L_2(d^{\mathcal{D}})}^2 &= \mathbb{E}_{d^{\mathcal{D}}} \left[\frac{1}{2} \left(\mathcal{T}_{\mathcal{D}}^{\pi} w - w \right)^2 \right] = \mathbb{E}_{d^{\mathcal{D}}} \left[\max_{v \in \mathbb{R}} \left(\mathcal{T}_{\mathcal{D}}^{\pi} w - w \right) v - \frac{1}{2} v^2 \right] \\
&= \max_{v: S \times A \to \mathbb{R}} \mathbb{E}_{d^{\mathcal{D}}} \left[(D^{\mathcal{D}})^{-1} \left((1-\gamma) d_0^{\pi} + \gamma \mathcal{P}_*^{\pi} D^{\mathcal{D}} w \right) v \right] - \mathbb{E}_{d^{\mathcal{D}}} [wv] - \frac{1}{2} \mathbb{E}_{d^{\mathcal{D}}} [v^2]. \\
&= \max_{v: S \times A \to \mathbb{R}} (1-\gamma) \langle d_0^{\pi}, v \rangle + \gamma \langle \mathcal{P}_*^{\pi} D^{\mathcal{D}} w, v \rangle - \langle D^{\mathcal{D}} w, v \rangle - \frac{1}{2} \langle d^{\mathcal{D}}, v^2 \rangle \\
&= \max_{v: S \times A \to \mathbb{R}} (1-\gamma) \langle d_0^{\pi}, v \rangle + \gamma \langle d^{\mathcal{D}}, w \mathcal{P}^{\pi} v \rangle - \langle d^{\mathcal{D}}, wv \rangle - \frac{1}{2} \langle d^{\mathcal{D}}, v^2 \rangle.
\end{aligned}$$

Also doing this for the second summand, like we did for GenDICE, we get an objective

$$\begin{aligned}
&\min_{w: S \times A \to \mathbb{R}_{\geq 0}} \max_{v: S \times A \to \mathbb{R},\ u \in \mathbb{R}} J(w, v, u) \\
&\quad \doteq (1-\gamma) \mathbb{E}_{(s_0, a_0) \sim d_0^{\pi}} [v(s_0, a_0)] \\
&\qquad + \mathbb{E}_{(s,a) \sim d^{\mathcal{D}},\ (s', a') \sim T^{\pi}(s,a)} \left[w(s,a) \left(\gamma v(s', a') - v(s,a) \right) \right] \\
&\qquad - \frac{1}{2} \mathbb{E}_{(s,a) \sim d^{\mathcal{D}}} \left[v(s,a)^2 \right] \\
&\qquad + \lambda \left(\mathbb{E}_{(s,a) \sim d^{\mathcal{D}}} [uw(s,a) - u] - \frac{1}{2} u^2 \right).
\end{aligned}$$

4.5.2 Implementation

Tabular

Using Remark 2.4.2, we can build the derivatives

$$\begin{aligned}
\partial_w \frac{1}{2}\|\vec{\mathcal{T}}^\pi \vec{D}^{\mathcal{D}}\vec{w} - \vec{D}^{\mathcal{D}}\vec{w}\|^2_{(\vec{D}^{\mathcal{D}})^{-1}} &= \partial_w \frac{1}{2}\|(1-\gamma)\vec{d}_0^\pi - \vec{A}_1\vec{w}\|^2_{(\vec{D}^{\mathcal{D}})^{-1}} \\
&= -\big((1-\gamma)\vec{d}_0^\pi - \vec{A}_1\vec{w}\big)^\top (\vec{D}^{\mathcal{D}})^{-1}\vec{A}_1 \\
&= -(1-\gamma)(\vec{d}_0^\pi)^\top (\vec{D}^{\mathcal{D}})^{-1}\vec{A}_1 + \vec{w}^\top \vec{A}_1^\top (\vec{D}^{\mathcal{D}})^{-1}\vec{A}_1, \\
\partial_w \frac{\lambda}{2}((\vec{d}_0^\pi)^\top \vec{w} - 1)^2 &= \lambda((\vec{d}_0^\pi)^\top \vec{w} - 1)(\vec{d}_0^\pi)^\top \\
&= \lambda \vec{w}^\top \vec{d}_0^\pi (\vec{d}_0^\pi)^\top - \lambda(\vec{d}_0^\pi)^\top,
\end{aligned}$$

hence, the gradient is

$$\nabla_w J(w) = \big(\vec{A}_1^\top (\vec{D}^{\mathcal{D}})^{-1}\vec{A}_1 + \lambda \vec{d}_0^\pi (\vec{d}_0^\pi)^\top\big)\vec{w} - \big((1-\gamma)\vec{A}_1^\top (\vec{D}^{\mathcal{D}})^{-1}\vec{d}_0^\pi + \lambda \vec{d}_0^\pi\big).$$

Setting the gradient to zero, and applying Remarks 3.5.3 and 3.5.4, we get the linear equation system

$$\big(\vec{A}_1^\top (\vec{D}^{\mathcal{D}})^{-1}\vec{A}_1 + \lambda \vec{d}_0^\pi (\vec{d}_0^\pi)^\top\big)\vec{w}_{\pi/\mathcal{D}} = (1-\gamma)\vec{A}_1^\top (\vec{D}^{\mathcal{D}})^{-1}\vec{d}_0^\pi + \lambda \vec{d}_0^\pi.$$

Considering the Law of Large Numbers approximations $\hat{d}_0^\pi$, $\hat{A}$, and $\hat{D}^{\mathcal{D}}$, $\hat{A}_p$, from (4.9), (4.10), (4.11), and (4.12), respectively, we get the approximate linear equation system,

$$\big(\hat{A}_1^\top (\hat{D}^{\mathcal{D}})^{-1}\hat{A}_1 + \lambda \hat{d}_0^\pi (\hat{d}_0^\pi)^\top\big)\hat{w}_{\pi/\mathcal{D}} = (1-\gamma)\hat{A}_1^\top (\hat{D}^{\mathcal{D}})^{-1}\hat{d}_0^\pi + \lambda \hat{d}_0^\pi.$$

Continuous

The implementation for the continuous setting is similar to that of DualDICE. In contrast to GradientDICE, we do not require our parameterization for w to ensure non-negativity. Building the gradients, we get

$$\begin{aligned}
\nabla_\theta J(w_\theta, v_\vartheta, u) &= \mathbb{E}_{(s,a)\sim d^{\mathcal{D}},\ (s',a')\sim T^\pi(s,a)}[(\gamma v_\vartheta(s',a') - v_\vartheta(s,a))\nabla_\theta w_\theta(s,a)] \\
&\quad + \lambda u \mathbb{E}_{(s,a)\sim d^{\mathcal{D}}}[\nabla_\theta w_\theta(s,a)], \\
\nabla_\vartheta J(w_\theta, v_\vartheta, u) &= (1-\gamma)\mathbb{E}_{(s_0,a_0)\sim d_0^\pi}[\nabla_\vartheta v_\vartheta(s_0,a_0)] \\
&\quad + \mathbb{E}_{(s,a)\sim d^{\mathcal{D}},\ (s',a')\sim T^\pi(s,a)}\left[w_\theta(s,a)\left(\gamma \nabla_\vartheta v_\vartheta(s',a') - \nabla_\vartheta v_\vartheta(s,a)\right)\right] \\
&\quad - \mathbb{E}_{(s,a)\sim d^{\mathcal{D}}}[v_\vartheta(s,a)\nabla_\vartheta v_\vartheta(s,a)], \\
\nabla_u J(w_\theta, v_\vartheta, u) &= \lambda\left(\mathbb{E}_{(s,a)\sim d^{\mathcal{D}}}[w_\theta(s,a) - 1] - u\right).
\end{aligned}$$

4.5.3 Convergence

We now want to prove some convergence results for GradientDICE, using a linear parameterization. To this end, consider the feature function $x : S \times A \to \mathbb{R}^K$ and feature matrix $X \in \mathbb{R}^{|S\times A|\times K}$ from (4.17) and linear parameterizations

$$w_\theta(s,a) \doteq \langle x(s,a), \theta\rangle \quad \text{and} \quad v_\vartheta(s,a) \doteq \langle x(s,a), \vartheta\rangle \quad \text{for} \quad (s,a) \in S \times A.$$

We can choose a constant $\xi \geq 0$ and perform ridge regularization,

$$\min_{\theta \in \mathbb{R}^K} \max_{\vartheta \in \mathbb{R}^K,\ u \in \mathbb{R}} J_\xi(w_\theta, v_\vartheta, u) \doteq J(w_\theta, v_\vartheta, u) + \frac{\xi}{2}\|\theta\|_2^2.$$

Building the gradients, using Remark 2.4.2, we get

$$\begin{aligned}
\nabla_\theta J_\xi(w_\theta, v_\vartheta, u) = \gamma \mathbb{E}_{(s,a)\sim d^{\mathcal{D}},\ (s',a')\sim T^\pi(s,a)}[\langle x(s',a'), \vartheta\rangle x(s,a)] \\
- \mathbb{E}_{(s,a)\sim d^{\mathcal{D}}}[\langle x(s,a), \vartheta\rangle x(s,a)] \\
- \lambda u \mathbb{E}_{(s,a)\sim d^{\mathcal{D}}}[x(s,a)] \\
+ \xi\theta, \\
\nabla_\vartheta J_\xi(w_\theta, v_\vartheta, u) = (1-\gamma)\mathbb{E}_{(s_0,a_0)\sim d_0^\pi}[x(s_0,a_0)] \\
+ \gamma \mathbb{E}_{(s,a)\sim d^{\mathcal{D}},\ (s',a')\sim T^\pi(s,a)}[\langle x(s,a), \theta\rangle x(s',a')] \\
- \mathbb{E}_{(s,a)\sim d^{\mathcal{D}}}[\langle x(s,a), \theta\rangle x(s,a)] \\
- \mathbb{E}_{(s,a)\sim d^{\mathcal{D}}}[\langle x(s,a), \vartheta\rangle x(s,a)], \\
\nabla_u J_\xi(w_\theta, v_\vartheta, u) = \lambda\left(\langle \mathbb{E}_{(s,a)\sim d^{\mathcal{D}}}[x(s,a)], \theta\rangle - 1 - u\right).
\end{aligned}$$

Replacing the expected values by samples,

$$\begin{aligned}
(s_{0,t}, s_t, a_t, s_t') \sim p^{\mathcal{D}}, \quad a_{0,t} \sim \pi(s_{0,t}), \quad a_t' \sim \pi(s_t'), \\
x_{0,t} \doteq x(s_{0,t}, a_{0,t}), \quad x_t \doteq x(s_t, a_t), \quad x_t' \doteq x(s_t', a_t'),
\end{aligned}$$

and choosing a learning rate sequence $(\alpha_t)_{t\in\mathbb{N}}$, satisfying the Robbins-Monro conditions (2.18), we get the SGDA formulation

$$\begin{aligned}
\theta_{t+1} &\doteq \theta_t - \alpha_t\left(\gamma\langle x_t', \vartheta_t\rangle x_t - \langle x_t, \vartheta_t\rangle x_t + \lambda u_t x_t + \xi\theta_t\right), \\
\vartheta_{t+1} &\doteq \vartheta_t + \alpha_t\left((1-\gamma)x_{0,t} + \gamma\langle x_t, \theta_t\rangle x_t' - \langle x_t, \theta_t\rangle x_t - \langle x_t, \vartheta_t\rangle x_t\right), \\
u_{t+1} &\doteq u_t + \alpha_t\lambda\left(\langle x_t, \theta_t\rangle - 1 - u_t\right).
\end{aligned}$$

We can collect all parameters into a single vector $\kappa_t^\top \doteq (\theta_t^\top, \vartheta_t^\top, u_t)$. Then, we can rewrite a step as $\kappa_{t+1} \doteq \kappa_t + \alpha_t(G_{t+1}\kappa_t + g_{t+1})$, where

$$G_{t+1} \doteq \begin{pmatrix} -\xi I & x_t(x_t - \gamma x_t')^\top & -\lambda x_t \\ -(x_t - \gamma x_t')x_t^\top & -x_t x_t^\top & \vec{0} \\ \lambda x_t^\top & \vec{0} & -\lambda \end{pmatrix} \quad \text{and} \quad g_{t+1} \doteq \begin{pmatrix} \vec{0} \\ (1-\gamma)x_{0,t} \\ -\lambda \end{pmatrix}.$$

We want to calculate the expected values $G \doteq \mathbb{E}[G_{t+1}]$ and $g \doteq \mathbb{E}[g_{t+1}]$. For the individual

parts we have

$$\begin{aligned}
&\mathbb{E}[x_t(x_t - \gamma x_t')^\top] \\
&= \mathbb{E}_{(s,a)\sim d^{\mathcal{D}},\ (s',a')\sim T^\pi(s,a)}[x(s,a)(x(s,a) - \gamma x(s',a'))^\top] \\
&= \sum_{(s,a)\in S\times A} x(s,a) d^{\mathcal{D}}(s,a) \left(x(s,a) - \gamma \sum_{(s',a')\in S\times A} T^\pi(s',a' \mid s,a) x(s',a') \right)^\top \\
&= X^\top \vec{D}^{\mathcal{D}} (I - \gamma \vec{\mathcal{P}}^\pi) X = \left(X^\top (I - \gamma \vec{\mathcal{P}}_*^\pi) \vec{D}^{\mathcal{D}} X \right)^\top,
\end{aligned}$$

$$\begin{aligned}
\mathbb{E}[x_t] &= \mathbb{E}_{(s,a)\sim d^{\mathcal{D}}}[x(s,a)] = \sum_{(s,a)\in S\times A} x(s,a) d^{\mathcal{D}}(s,a) = X^\top d^{\mathcal{D}}, \\
\mathbb{E}[x_{t,0}] &= \mathbb{E}_{(s,a)\sim d_0^\pi}[x(s,a)] = \sum_{(s,a)\in S\times A} x(s,a) d_0^\pi(s,a) = X^\top d_0^\pi, \\
\mathbb{E}[x_t x_t^\top] &= \sum_{(s,a)\in S\times A} x(s,a) d^{\mathcal{D}}(s,a) x(s,a)^\top = X^\top \vec{D}^{\mathcal{D}} X.
\end{aligned}$$

We define the matrices

$$E_\gamma \doteq X^\top (I - \gamma \vec{\mathcal{P}}_*^\pi) \vec{D}^{\mathcal{D}} X \quad \text{and} \quad E_0 \doteq X^\top \vec{D}^{\mathcal{D}} X.$$

Then, we get

$$G = \begin{pmatrix} -\xi I & E_\gamma^\top & -\lambda X^\top \vec{d}^{\mathcal{D}} \\ -E_\gamma & -E_0 & \vec{0} \\ \lambda (X^\top \vec{d}^{\mathcal{D}})^\top & \vec{0} & -\lambda \end{pmatrix} \quad \text{and} \quad g = \begin{pmatrix} \vec{0} \\ (1-\gamma) X^\top \vec{d}_0^\pi \\ -\lambda \end{pmatrix}.$$

Now, we can finally formulate the SGDA convergence result in Theorem 4.5.6.

Assumption 4.5.1. X has full rank, i.e., linearly independent columns.

Remark 4.5.2. If we assume that $D^{\mathcal{D}} > 0$, Assumption 4.5.1 implies that E_0 is positive definite. Furthermore, since $E_0 \in \mathbb{R}^{K\times K}$ is symmetric, $\|\cdot\|_{E_0}$ is a norm on $\mathbb{C}^K$. ■

Assumption 4.5.3. E_γ is non-singular or $\xi > 0$.

Remark 4.5.4. Sutton and Barto show that $(I - \gamma \vec{\mathcal{P}}_*^\pi)\vec{D}^{\mathcal{D}}$ is positive definite for $0 \le \gamma < 1$ [38, pp. 206–207]. By Assumption 4.5.1, E_γ is positive definite. Remark 2.4.3 further implies that Assumption 4.5.3 is satisfied for $0 \le \gamma < 1$, even without ridge regularization, i.e., $\xi = 0$. ■

Assumption 4.5.5. The features $x_{0,t}$, x_t and x_t' have uniformly bounded second moments.

Theorem 4.5.6. *Let Assumptions 4.5.1, 4.5.3, and 4.5.5 hold. Then, we have*

$$\lim_{t\to\infty} \kappa_t = -G^{-1} g \quad \textit{almost surely.}$$

Proof. Our goal is to apply the ODE Lemma 2.5.7. Rewrite the update for κ_t as

$$\begin{aligned}
\kappa_{t+1} \doteq \kappa_t + \alpha_t(G_{t+1}\kappa_t + g_{t+1}) = \kappa_t + \alpha_t(h(\kappa_t) + M_{t+1}), \quad \text{where} \\
h(\kappa) \doteq G\kappa + g, \quad \text{and} \quad M_{t+1} \doteq (G_{t+1} - G)\kappa_t + (g_{t+1} - g).
\end{aligned} \tag{4.21}$$

1. We show that Assumption 2.5.5 is satisfied. For the first part, we verify that the sequence $(M_t)_{t\in\mathbb{N}}$ is a martingale difference sequence with respect to the filtration

$$\mathcal{F}_t \doteq \sigma(\kappa_{i-1}, M_i)_{i=1}^{t} = \sigma(x_{0,i}, x_i, x_i')_{i=1}^{t-1}.$$

Because all the random variables that G_{t+1} and g_{t+1} are constructed from are $x_{0,t}$, x_t and x_t', they are independent to $\mathcal{F}_t$. On the other hand, the way we have rewritten κ_t in (4.21), we can obtain it by using $(\kappa_i)_{i=0}^{t-1}$ and $(M_i)_{i=1}^{t}$, so it is $\mathcal{F}_t$-measurable. This leads to

$$\begin{aligned}
\mathbb{E}[M_{t+1} \mid \mathcal{F}_t] &= \mathbb{E}[(G_{t+1} - G)\kappa_t \mid \mathcal{F}_t] + \mathbb{E}[g_{t+1} - g \mid \mathcal{F}_t] \\
&= \mathbb{E}[G_{t+1} - G \mid \mathcal{F}_t]\kappa_t + \mathbb{E}[g_{t+1} - g \mid \mathcal{F}_t] \\
&= (\mathbb{E}[G_{t+1} \mid \mathcal{F}_t] - G)\kappa_t + (\mathbb{E}[g_{t+1} \mid \mathcal{F}_t] - g) \\
&= (\mathbb{E}[G_{t+1}] - G)\kappa_t + (\mathbb{E}[g_{t+1}] - g) = 0.
\end{aligned}$$

For the last part, we define $\tilde{G}_t \doteq G_t - G$ and $\tilde{g}_t \doteq g_t - g$. Note that for all $x \in \mathbb{R}$,

$$0 \le (x-1)^2 = x^2 - 2x + 1, \quad \text{so} \quad 2x \le x^2 + 1 \le 2x^2 + 1.$$

Now, define $C_t \doteq 2\max\{\|\tilde{G}_{t+1}\|^2 + \|\tilde{G}_{t+1}\| \cdot \|\tilde{g}_{t+1}\|, \|\tilde{g}_{t+1}\|^2\}$ and use the Cauchy-Schwarz inequality, to show

$$\begin{aligned}
\|M_{t+1}\|^2 &= \|\tilde{G}_{t+1}\kappa_t\|^2 + 2\langle \tilde{G}_{t+1}\kappa_t, \tilde{g}_{t+1}\rangle + \|\tilde{g}_{t+1}\|^2 \\
&\le \|\tilde{G}_{t+1}\kappa_t\|^2 + 2\|\tilde{G}_{t+1}\kappa_t\| \cdot \|\tilde{g}_{t+1}\| + \|\tilde{g}_{t+1}\|^2 \\
&\le \|\tilde{G}_{t+1}\|^2\|\kappa_t\|^2 + \|\tilde{G}_{t+1}\|(2\|\kappa_t\|)\|\tilde{g}_{t+1}\| + \|\tilde{g}_{t+1}\|^2 \\
&\le \|\tilde{G}_{t+1}\|^2(2\|\kappa_t\|^2 + 1) + \|\tilde{G}_{t+1}\|(2\|\kappa_t\|^2 + 1)\|\tilde{g}_{t+1}\| + \|\tilde{g}_{t+1}\|^2 \\
&\le (\|\tilde{G}_{t+1}\|^2 + \|\tilde{G}_{t+1}\| \cdot \|\tilde{g}_{t+1}\|)(2\|\kappa_t\|^2 + 1) + \|\tilde{g}_{t+1}\|^2 \\
&\le C_t(\|\kappa_t\|^2 + 1).
\end{aligned}$$

Again, C_t is independent of $\mathcal{F}_t$ and κ_t is $\mathcal{F}_t$-measurable, so

$$\begin{aligned}
\mathbb{E}[\|M_{t+1}\|^2 \mid \mathcal{F}_t] &\le \mathbb{E}[C_t(\|\kappa_t\|^2 + 1) \mid \mathcal{F}_t] \\
&= \mathbb{E}[C_t \mid \mathcal{F}_t](\|\kappa_t\|^2 + 1) \\
&= \mathbb{E}[C_t](\|\kappa_t\|^2 + 1) \le \sup_{i\in\mathbb{N}} \mathbb{E}[C_i](\|\kappa_t\|^2 + 1).
\end{aligned}$$

Since we assumed $x_{0,t}$, x_t and x_t' to have uniformly bounded second moments, we also get $\sup_{i\in\mathbb{N}} \mathbb{E}[C_i] < \infty$.

2. We show that Assumption 2.5.4 is satisfied.

i. The function h is indeed Lipschitz continuous, because

$$\|h(y_1) - h(y_2)\| = \|(Gy_1 - g) - (Gy_2 - g)\| = \|G(y_1 - y_2)\| \le \|G\| \cdot \|y_1 - y_2\|.$$

Defining the function $h_\infty(y) \doteq Gy$, we get

$$\left\| \frac{h(ry)}{r} - h_\infty(y) \right\| = \left\| Gy - \frac{g}{r} - Gy \right\| = \frac{1}{r}\|g\| \xrightarrow{r\to\infty} 0.$$

ii. We need to prove that the origin $\vec{0} \in \mathbb{R}^n$ is an asymptotically stable equilibrium of the ODE $y'(t) = h_\infty(y(t)) = Gy(t)$. This can be done by checking that the real parts $\Re(l) < 0$ for

all eigenvalues l of G. For now, let $l \neq 0$ be an eigenvalue of G with normalized eigenvector $\kappa \neq 0$, i.e., $\overline{\kappa}^\top \kappa = \|\kappa\|^2 = 1$. If we let $\kappa^\top \doteq (\theta^\top, \vartheta^\top, u)$, where $\theta \in \mathbb{C}^K$, $\vartheta \in \mathbb{C}^K$ and $u \in \mathbb{C}$, then

$$
\begin{aligned}
l = l\overline{\kappa}^\top \kappa = \overline{\kappa}^\top G \kappa &= \begin{pmatrix} \overline{\theta} \\ \overline{\vartheta} \\ \overline{u} \end{pmatrix}^\top \begin{pmatrix} -\xi I & E_\gamma^\top & -\lambda X^\top \vec{d}^{\mathcal{D}} \\ -E_\gamma & -E_0 & \vec{0} \\ \lambda (X^\top \vec{d}^{\mathcal{D}})^\top & \vec{0} & -\lambda \end{pmatrix} \begin{pmatrix} \theta \\ \vartheta \\ u \end{pmatrix} \qquad (4.22) \\
&= -\xi \overline{\theta}^\top \theta + \overline{\theta}^\top E_\gamma^\top \vartheta - \lambda u \overline{\theta}^\top X^\top \vec{d}^{\mathcal{D}} - \overline{\vartheta}^\top E_\gamma \theta - \overline{\vartheta}^\top E_0 \vartheta + \lambda \overline{u} (X^\top \vec{d}^{\mathcal{D}})^\top \theta - \lambda \overline{u} u \\
&= -\xi \|\theta\|^2 - \|\vartheta\|_{E_0}^2 - \lambda |u|^2 + \Im\left(\overline{\theta}^\top E_\gamma^\top \vartheta - \overline{\vartheta}^\top E_\gamma \theta\right) - \Im\left(\lambda u \overline{\theta}^\top X^\top \vec{d}^{\mathcal{D}} - \lambda \overline{u} (X^\top \vec{d}^{\mathcal{D}})^\top \theta\right).
\end{aligned}
$$

In order to show that $\Re(l) < 0$, we consider two cases. When $\gamma < 1$, we have $\xi = 0$ and $l \neq 0$ implies that $\vartheta \neq \vec{0}$ or $u \neq 0$. When $\gamma = 1$, we have $\xi > 0$ and $l \neq 0$ implies that $\theta \neq \vec{0}$, $\vartheta \neq \vec{0}$, or $u \neq 0$.

iii. It only remains to show that $\kappa^* \doteq -G^{-1} g$ is the unique globally asymptotically stable equilibrium for the ODE $y'(t) = h(y(t)) = Gy(t) + g$.

Firstly, we want to show that G is non-singular, so we check for $\det(G) \neq 0$. Applying (2.15) twice, we get

$$
\begin{aligned}
\det(G) &= -\lambda \det\left(\begin{pmatrix} -\xi I & E_\gamma^\top \\ -E_\gamma & -E_0 \end{pmatrix} + \lambda^{-1} \begin{pmatrix} -\lambda^2 (X^\top \vec{d}^{\mathcal{D}})(X^\top \vec{d}^{\mathcal{D}})^\top & \vec{0} \\ \vec{0} & \vec{0} \end{pmatrix}\right) \\
&= (-1)^{2K+1} \lambda \det \begin{pmatrix} \xi I + \lambda (X^\top \vec{d}^{\mathcal{D}})(X^\top \vec{d}^{\mathcal{D}})^\top & -E_\gamma^\top \\ E_\gamma & E_0 \end{pmatrix} \\
&= (-1)^{2K+1} \lambda \det(E_0) \det\left(\xi I + \lambda (X^\top \vec{d}^{\mathcal{D}})(X^\top \vec{d}^{\mathcal{D}})^\top + E_\gamma^\top E_0^{-1} E_\gamma\right).
\end{aligned}
$$

Now, because $\lambda > 0$, all of our summands are positive semi-definite. According to Assumption 4.5.3, ξI or $E_\gamma^\top E_0^{-1} E_\gamma$ is strictly positive definite. Because the sum of positive semi-definite matrices is positive semi-definite, and even positive definite if a single summand is positive definite, this ensures $\det(G) \neq 0$. Since G is now verified to be non-singular, the linear equation system $0 = h(\kappa) = G\kappa + g$ only has κ^* as solution, which means that it is the unique equilibrium.

For the global asymptotic stability, we use Ljapunov's method and $L(y) \doteq \frac{1}{2}\|Gy + g\|^2$. Applying the chain rule and Remark 2.4.2, we get $\nabla L(y) = G^\top (Gy + g)$. To verify that L is a strict Ljapunov function for h, we consider $y \in \mathbb{R}^n$, which is not an equilibrium, i.e., $y \neq -G^{-1} g$. This means that $\kappa \doteq Gy + g \neq 0$ is a real vector. We can reuse our calculation (4.22) from earlier, and conclude analogously that

$$
\langle \nabla L(y), h(y) \rangle = \langle G^\top (Gy + g), Gy + g \rangle = \kappa^\top G \kappa = -\xi \|\theta\|^2 - \|\vartheta\|_{E_0}^2 - \lambda |u|^2 < 0.
$$

Thus, L is a strict Ljapunov function. Now, check that κ^* is a strict minimum of L. We notice that $\nabla L(\kappa^*) = 0$ and the hessian matrix $\nabla^2 L(y) = GG^\top$ is positive definite, since G is non-singular, so for all $y \in \mathbb{R}^{2K+1}$,

$$
y^\top \nabla^2 L(y) y = (G^\top y)^\top (G^\top y) = \|G^\top y\|^2 = 0 \iff G^\top y = \vec{0} \iff y = \vec{0}.
$$

□

Remark 4.5.7. A different perspective to solving for the optimal solution of our objective is to set its gradient to zero. To do this as efficiently as possible, we combine our arguments into a

single vector $\kappa^\top \doteq (\theta^\top, \vartheta^\top, u)$ again. Then we can simplify the objective even more as

$$\begin{aligned}
&\min_{\theta\in\mathbb{R}^K} \max_{\vartheta\in\mathbb{R}^K,\ u\in\mathbb{R}} J_\xi(w_\theta, v_\vartheta, u) \\
&\quad= \frac{1}{2}\left(\theta^\top \xi I\theta - \vartheta^\top X^\top D^{\mathcal{D}} X\vartheta - u\lambda u\right) \\
&\qquad - \frac{1}{2}\left(\vartheta^\top X^\top (I-\gamma\vec{\mathcal{P}}_*^\pi) D^{\mathcal{D}} X\theta + \theta^\top X^\top (I-\gamma\vec{\mathcal{P}}_*^\pi) D^{\mathcal{D}} X\vartheta\right) \\
&\qquad - \frac{1}{2}\left(u\lambda (X^\top \vec{d}^{\mathcal{D}})^\top\theta + \theta^\top\lambda (X^\top\vec{d}^{\mathcal{D}})^\top u\right) \\
&\qquad - \left((1-\gamma)X^\top \vec{d}_0^{\pi}\right)^\top \vartheta - \lambda u \\
&\quad= \frac{1}{2}\kappa^\top H\kappa + g^\top\kappa,
\end{aligned}$$

where

$$H \doteq \begin{pmatrix} \xi I & -E_\gamma^\top & \lambda x^\top \vec{d}^{\mathcal{D}} \\ -E_\gamma & -E_0 & \vec{0} \\ \lambda(X^\top\vec{d}^{\mathcal{D}})^\top & \vec{0} & -\lambda \end{pmatrix}.$$

Now, the overall gradient is simply

$$\nabla_\kappa J_\xi(w_\theta, v_\vartheta, u) = H\kappa + g.$$

Define the invertible matrix $\sigma \doteq \operatorname{diag}(-I_K, I_K, 1)$. Since the θ-component of g is $\vec{0}$, we have $\sigma g = g$. Also, by definition, $\sigma H = G$. If we set the gradient to zero and multiply with σ, we get

$$\vec{0} = \sigma(H\kappa + g) = G\kappa + g.$$

■

We now want to prove some consistency results, i.e., that the algorithm actually approximates the stationary distribution correction $\vec{w}_{\pi/\mathcal{D}}$.

To this end, we invert the matrix

$$-G = \left(\begin{array}{cc|c} \xi I & -E_\gamma^\top & \lambda X^\top\vec{d}^{\mathcal{D}} \\ E_\gamma & E_0 & \vec{0} \\ \hline -\lambda(X^\top\vec{d}^{\mathcal{D}})^\top & \vec{0} & \lambda \end{array}\right).$$

Then, we use Theorem 4.5.6, to obtain θ, the part of κ_∞, which relates to $\vec{w}_{\pi/\mathcal{D}}$. We do this, by using Lemma 2.4.5 twice.

Firstly, we define

$$M \doteq \begin{pmatrix} \xi I & -E_\gamma^\top \\ E_\gamma & E_0 \end{pmatrix} \quad\text{and}\quad \Xi \doteq (M/E_0)^{-1} = (\xi I + E_\gamma^\top E_0^{-1} E_\gamma)^{-1}.$$

Note that by definition of E_0, we have that Ξ is symmetric. We apply (2.17) to M and get

$$M^{-1} = \begin{pmatrix} \Xi & \Xi E_\gamma^\top E_0^{-1} \\ -E_0^{-1}E_\gamma\Xi & E_0^{-1} + E_0^{-1}E_\gamma\Xi E_\gamma^\top E_0^{-1} \end{pmatrix}.$$

Secondly, we define

$$z \doteq \Xi X^\top \vec{d}^{\mathcal{D}} \quad \text{and} \quad \beta \doteq \frac{1}{\lambda}(-G/M)$$
$$= \frac{1}{\lambda}\left(\lambda - \begin{pmatrix} -\lambda(X^\top \vec{d}^{\mathcal{D}})^\top & \vec{0} \end{pmatrix} M^{-1} \begin{pmatrix} \lambda X^\top \vec{d}^{\mathcal{D}} \\ \vec{0} \end{pmatrix}\right)$$
$$= 1 + \lambda (X^\top \vec{d}^{\mathcal{D}})^\top \Xi (X^\top \vec{d}^{\mathcal{D}})$$
$$= 1 + \lambda z^\top \Xi^{-1} z.$$

Now, we can calculate the blocks of $-G^{-1}$ as

$$M^{-1} + M^{-1} \begin{pmatrix} \lambda X^\top \vec{d}^{\mathcal{D}} \\ \vec{0} \end{pmatrix} \frac{\beta^{-1}}{\lambda} \begin{pmatrix} -\lambda (X^\top \vec{d}^{\mathcal{D}})^\top & \vec{0} \end{pmatrix} M^{-1}$$
$$= M^{-1} - \lambda \beta^{-1} \begin{pmatrix} \Xi X^\top \vec{d}^{\mathcal{D}} \\ -E_0^{-1} E_\gamma \Xi X^\top \vec{d}^{\mathcal{D}} \end{pmatrix} \begin{pmatrix} \Xi X^\top \vec{d}^{\mathcal{D}} \\ E_0^{-1} E_\gamma \Xi X^\top \vec{d}^{\mathcal{D}} \end{pmatrix}^\top$$
$$= M^{-1} - \lambda \beta^{-1} \begin{pmatrix} \Xi (X^\top \vec{d}^{\mathcal{D}})(X^\top \vec{d}^{\mathcal{D}})^\top \Xi & \Xi (X^\top \vec{d}^{\mathcal{D}})(X^\top \vec{d}^{\mathcal{D}})^\top \Xi E_\gamma^\top E_0^{-1} \\ -E_0^{-1} E_\gamma \Xi (X^\top \vec{d}^{\mathcal{D}})(X^\top \vec{d}^{\mathcal{D}})^\top \Xi & -E_0^{-1} E_\gamma \Xi (X^\top \vec{d}^{\mathcal{D}})(X^\top \vec{d}^{\mathcal{D}})^\top \Xi E_\gamma^\top E_0^{-1} \end{pmatrix}$$
$$= \begin{pmatrix} \Xi + \lambda \beta^{-1} z z^\top & \Xi E_\gamma^\top E_0^{-1} - \lambda \beta^{-1} z z^\top E_\gamma^\top E_0^{-1} \\ -E_0^{-1} E_\gamma \Xi + \lambda \beta^{-1} E_0^{-1} E_\gamma z z^\top & E_0^{-1} + E_0^{-1} E_\gamma \Xi E_\gamma^\top E_0^{-1} - \lambda \beta^{-1} E_0^{-1} E_\gamma z z^\top E_\gamma^\top E_0^{-1} \end{pmatrix},$$

$$-M^{-1} \begin{pmatrix} \lambda X^\top \vec{d}^{\mathcal{D}} \\ \vec{0} \end{pmatrix} \frac{\beta^{-1}}{\lambda} = -\beta^{-1} \begin{pmatrix} \Xi X^\top \vec{d}^{\mathcal{D}} \\ -E_0^{-1} E_\gamma \Xi X^\top \vec{d}^{\mathcal{D}} \end{pmatrix} = \begin{pmatrix} -\beta^{-1} z \\ \beta^{-1} E_0^{-1} E_\gamma z \end{pmatrix},$$

$$-\frac{\beta^{-1}}{\lambda} \begin{pmatrix} -\lambda (X^\top \vec{d}^{\mathcal{D}})^\top & \vec{0} \end{pmatrix} M^{-1} = \beta^{-1} \begin{pmatrix} (X^\top \vec{d}^{\mathcal{D}})^\top \Xi & -(X^\top \vec{d}^{\mathcal{D}})^\top \Xi E_\gamma^\top E_0^{-1} \end{pmatrix}$$
$$= \begin{pmatrix} \beta^{-1} z^\top & -\beta^{-1} z^\top E_\gamma^\top E_0^{-1} \end{pmatrix}.$$

Finally, we use (2.16) to calculate

$$-G^{-1} = \left(\begin{array}{cc|c} \Xi + \lambda \beta^{-1} z z^\top & \Xi E_\gamma^\top E_0^{-1} - \lambda \beta^{-1} z z^\top E_\gamma^\top E_0^{-1} & -\beta^{-1} z \\ -E_0^{-1} E_\gamma \Xi + \lambda \beta^{-1} E_0^{-1} z z^\top & E_0^{-1} + E_0^{-1} E_\gamma \Xi E_\gamma^\top E_0^{-1} - \lambda \beta^{-1} E_0^{-1} E_\gamma z z^\top E_\gamma^\top E_0^{-1} & \beta^{-1} E_0^{-1} E_\gamma z \\ \hline \beta^{-1} z^\top & -\beta^{-1} z^\top E_\gamma^\top E_0^{-1} & \lambda \beta^{-1} \end{array}\right).$$

Multiplying with g, we can derive a formula for

$$\theta_\infty = \left(\Xi E_\gamma^\top E_0^{-1} - \lambda \beta^{-1} z z^\top E_\gamma^\top E_0^{-1}\right)(1-\gamma) X^\top \vec{d}_0^{\vec{\pi}} + \beta^{-1} z \lambda$$
$$= (1-\gamma) \Xi E_\gamma^\top E_0^{-1} X^\top \vec{d}_0^{\vec{\pi}} + \lambda \beta^{-1} z \left(1 - (1-\gamma) z^\top E_\gamma^\top E_0^{-1} X^\top \vec{d}_0^{\vec{\pi}}\right). \tag{4.23}$$

While Zhang et al. [41] have a consistency proof for the undiscounted case with Proposition 4.5.10, they are missing one for the discounted setting. We add such a statement in the form of Proposition 4.5.8.

Proposition 4.5.8. *Let Assumptions 4.5.1, 4.5.5 and 4.2.1 hold. Furthermore, assume that $\xi = 0$ and E_γ is non-singular. Then*

$$X\theta_\infty = \vec{w}_{\pi/\mathcal{D}}.$$

Proof. By the definition of the stationary distribution correction $w_{\pi/\mathcal{D}}$, and the backwards Bellman equations (3.9) respectively,

$$\begin{aligned}(\vec{D}^{\mathcal{D}})^{-1}\vec{d}^{\pi} = \vec{w}_{\pi/\mathcal{D}} &\iff \vec{d}^{\pi} = \vec{D}^{\mathcal{D}}\vec{w}_{\pi/\mathcal{D}},\\ \vec{d}^{\pi} = (1-\gamma)\vec{d}_0^{\pi} + \gamma\vec{\mathcal{P}}_*^{\pi}\vec{d}^{\pi} &\iff (1-\gamma)\vec{d}_0^{\pi} = (I-\gamma\vec{\mathcal{P}}_*^{\pi})\vec{d}^{\pi} = A_1\vec{w}_{\pi/\mathcal{D}}.\end{aligned}$$

The following matrix can be checked to be a projection onto the range of X,

$$P \doteq X(X^\top A_1 X)^{-1}X^\top A_1.$$

Applying both of these identities yields

$$(1-\gamma)\; XE_\gamma^{-1}X^\top \vec{d}_0^{\pi} = P\vec{w}_{\pi/\mathcal{D}}.$$

Because $\xi = 0$, we have

$$\begin{aligned}\Xi^{-1} &= E_\gamma^\top E_0^{-1}E_\gamma, & z &= E_\gamma^{-1}E_0(E_\gamma^{-1})^\top X^\top \vec{d}^{\mathcal{D}},\\ \Xi &= E_\gamma^{-1}E_0(E_\gamma^{-1})^\top, & z^\top &= (\vec{d}^{\mathcal{D}})^\top XE_\gamma^{-1}E_0(E_\gamma^{-1})^\top.\end{aligned}$$

Putting it together with (4.23), we get

$$\begin{aligned}X\theta_\infty &= (1-\gamma)XE_\gamma^{-1}E_0(E_\gamma^{-1})^\top E_\gamma^\top E_0^{-1}X^\top \vec{d}_0^{\pi}\\ &\qquad + \lambda\beta^{-1}z(1-(1-\gamma)(\vec{d}^{\mathcal{D}})^\top XE_\gamma^{-1}E_0(E_\gamma^{-1})^\top E_\gamma^\top E_0^{-1}X^\top \vec{d}_0^{\pi})\\ &= (1-\gamma)XE_\gamma^{-1}X^\top \vec{d}_0^{\pi} + \lambda\beta^{-1}z(1-(1-\gamma)(\vec{d}^{\mathcal{D}})^\top XE_\gamma^{-1}X^\top \vec{d}_0^{\pi})\\ &= P\vec{w}_{\pi/\mathcal{D}} + \frac{1-(\vec{d}^{\mathcal{D}})^\top P\vec{w}_{\pi/\mathcal{D}}}{1/\lambda + z^\top \Xi^{-1}z}z.\end{aligned}$$

Finally, we use assumption 4.2.1 to get

$$P\vec{w}_{\pi/\mathcal{D}} = \vec{w}_{\pi/\mathcal{D}} \quad \text{and} \quad (\vec{d}^{\mathcal{D}})^\top P\vec{w}_{\pi/\mathcal{D}} = (\vec{d}^{\mathcal{D}})^\top \vec{w}_{\pi/\mathcal{D}} = 1.$$

□

Remark 4.5.9. Unfortunately, in practice we are faced with the issue of having to chose the feature matrix X ad hoc. This means that we cannot take assumption 4.2.1 for granted.

Now, we cannot say anything about the expectation of our projected stationary distribution correction $P\vec{w}_{\pi/\mathcal{D}}$ being equal to one. Hence, the fraction as in the proof of Proposition 4.5.8 does not necessarily vanish.

Also, it is unclear, whether P is an orthogonal projection onto the range of X with respect to the scalar product $\langle , \rangle_{A_1}$. In Remark 4.5.4 we already established that A_1 is positive definite, but it is not necessarily symmetric. By the Hilbert space projection theorem, this would have led to

$$P\vec{w}_{\pi/\mathcal{D}} = \arg\min\left\{\|\vec{w} - \vec{w}_{\pi/\mathcal{D}}\|_{A_1}^2 \;\middle|\; \vec{w} \in \operatorname{ran}(X)\right\}.$$

■

Non-singularity of E_γ for $0 \leq \gamma < 1$ is discussed in Remark 4.5.4. In case $\gamma = 1$, we cannot apply Proposition 4.5.8, but we still have Proposition 4.5.10.

Proposition 4.5.10. *Let Assumptions 4.5.1, 4.5.5 and 4.2.1 hold. Furthermore, assume that* $\gamma = 1$ *and* $XE_0^{-1}X^\top$ *is non-singular. Consider the eigendecomposition*

$$E_\gamma^\top E_0^{-1} E_\gamma = V \Lambda V^\top, \quad \text{where}$$
$$V \text{ orthogonal}, \quad \Lambda = \operatorname{diag}(\lambda_1, \dots, \lambda_r, 0, \dots, 0), \quad \lambda_1 \geq \dots \geq \lambda_r > 0.$$

Let $v \doteq V^\top X^\top \vec{d}^{\mathcal{D}}$ *and* $j = r+1, \dots, K$, *such that* $v_j \neq 0$. *Then*

$$X\theta_{\infty,\xi} \xrightarrow{\xi \downarrow 0} \vec{w}_{\pi/\mathcal{D}}.$$

Proof. Since $XE_0^{-1}X^\top$ is symmetric and positive semi-definite and we assumed that it is non-singular, it even is positive definite. Therefore, according to Remarks 3.5.3 and 3.5.4, it suffices to show that

$$\epsilon_1(\xi) \doteq \|\vec{D}^{\mathcal{D}} \vec{\mathcal{T}}_{\mathcal{D}}^{\pi} X\theta_{\infty,\xi} - \vec{D}^{\mathcal{D}} X\theta_{\infty,\xi}\|^2_{XE_0^{-1}X^\top} \xrightarrow{\xi \downarrow 0} 0 \quad \text{and} \quad \epsilon_2(\xi) \doteq (X^\top \vec{d}^{\mathcal{D}})^\top \theta_{\infty,\ell} - 1 \xrightarrow{\xi \downarrow 0} 0.$$

We calculate the auxiliary quantities

$$\begin{aligned}
\Xi &= (\xi I + V\Lambda V^\top)^{-1} = (V(\xi I + \Lambda)V^\top)^{-1} = V(\xi I + \Lambda)^{-1} V^\top, \\
z &= V(\xi I + \Lambda)^{-1} V^\top X^\top \vec{d}^{\mathcal{D}} = V(\xi I + \Lambda)^{-1} v, \\
\beta &= 1 + \lambda (\vec{d}^{\mathcal{D}})^\top X V (\xi I + \Lambda)^{-1} V^\top X^\top \vec{d}^{\mathcal{D}} = 1 + \lambda v^\top (\xi I + \Lambda)^{-1} v.
\end{aligned}$$

Applying this, together with $\gamma = 1$, to (4.23), we get

$$\theta_{\infty,\xi} = \lambda \beta^{-1} z = \lambda (1 + \lambda v^\top (\xi I + \Lambda)^{-1} v)^{-1} V (\xi I + \Lambda)^{-1} v.$$

Firstly, we take care of

$$\begin{aligned}
\epsilon_1(\xi) &= \|(I - \vec{\mathcal{P}}_*^{\pi}) \vec{D}^{\mathcal{D}} X\theta_{\infty,\xi}\|^2_{XE_0^{-1}X^\top} \\
&= \theta_{\infty,\xi}^\top X^\top \vec{D}^{\mathcal{D}} (I - \vec{\mathcal{P}}^{\pi}) X E_0^{-1} X^\top (I - \vec{\mathcal{P}}_*^{\pi}) \vec{D}^{\mathcal{D}} X\theta_{\infty,\xi} \\
&= \theta_{\infty,\xi}^\top E_\gamma^\top E_0^{-1} E_\gamma \theta_{\infty,\xi} \\
&= (V^\top \theta_{\infty,\xi})^\top \Lambda (V^\top \theta_{\infty,\xi}) \\
&= \frac{\lambda^2 v^\top \Lambda (\xi I + \Lambda)^{-2} v}{(1 + \lambda v^\top (\xi I + \Lambda)^{-1} v)^2}.
\end{aligned}$$

Secondly, we treat

$$\begin{aligned}
\epsilon_2(\xi) + 1 &= (\vec{d}^{\mathcal{D}})^\top X\theta_{\infty,\xi} \\
&= \lambda (1 + v^\top (\xi I + \Lambda)^{-1} v)^{-1} (\vec{d}^{\mathcal{D}})^\top X V (\xi I + \Lambda)^{-1} v \\
&= \frac{\lambda v^\top (\xi I + \Lambda)^{-1} v}{1 + \lambda v^\top (\xi I + \Lambda)^{-1} v}, .
\end{aligned}$$

Notice that

$$v^\top \Lambda (\xi I + \Lambda)^{-2} v = \sum_{i=1}^{r} v_i^2 \frac{\lambda_i^2}{(\xi + \lambda_i)^2} \quad \text{and} \quad v^\top (\xi I + \Lambda)^{-1} v = \sum_{i=1}^{r} v_i^2 \frac{1}{\xi + \lambda_i} + \sum_{i=r+1}^{K} v_i^2 \frac{1}{\xi}.$$

Since $v_j \neq 0$ and applying L'Hospital's rule, respectively,

$$\lim_{\xi \downarrow 0} \epsilon_1(\xi) = 0 \quad \text{and} \quad \lim_{\xi \downarrow 0} \epsilon_2(\xi) + 1 = 1.$$

□

4.6 Confidence Interval DIstribution Correction Estimation

In order to obtain confidence intervals instead of point estimates, in the offline behavior agnostic setting, via distribution correction estimation, Dai et al. [13] introduce the algorithm *CoinDICE*.

Similar to DualDICE, it formulates an objective based on the Q-LP from Lemma 3.3.1. Similar to an approach by Duchi et al. [14], Theorem 2.6.8 is applied to obtain an asymptotic confidence interval for the policy value.

Not only do we cover the algorithm as in this section, we also mention important details, omitted by Dai et al. [13].

4.6.1 Embedded Q-LP

Just like DualDICE, CoinDICE takes the dual of the Q-LP from Lemma 3.3.1 as a starting point. However, a feature map $\phi : S \times A \to \mathbb{R}^K$ is chosen and we consider a relaxation that embeds the constraints in a function space $\mathcal{F}_\phi$. For any $\beta \in \mathbb{R}^K$ we define the function $Q_\beta \doteq \beta^\top \phi$. We collect them inside the K-dimensional subspace

$$\mathcal{F}_\phi \doteq \{Q_\beta \mid \beta \in \mathbb{R}^K\} = \operatorname{span}\{\phi_i\}_{i=1}^K.$$

We get back our original formulation by choosing

$$K = |S \times A| \quad \text{and} \quad \phi(s,a) \doteq (\mathbb{1}_{s'=s,\ a'=a})_{(s',a') \in S \times A}.$$

The following Lemmas 4.6.1 and 4.6.2 discuss an $\mathcal{F}_\phi$-embedded version of the Q-LP from Lemma 3.3.1.

Lemma 4.6.1 (embedded Q-LP)**.** *Let* $0 < \gamma < 1$. *Then, the* primal embedded Q-LP

$$\rho_\phi^\pi = \min_{\beta \in \mathbb{R}^K} (1-\gamma)\mathbb{E}_{(s_0,a_0)\sim d_0^\pi}[Q_\beta(s_0,a_0)] \tag{4.24}$$
$$s.t. \quad \forall (s,a) \in S \times A: \ Q_\beta(s,a) \geq r(s,a) + \gamma \mathcal{P}^\pi Q_\beta(s,a),$$

has the dual embedded Q-LP

$$\rho_\phi^\pi = \max_{d: S\times A \to \mathbb{R}_{\geq 0}} \mathbb{E}_{(s,a)\sim d}[r(s,a)] \tag{4.25}$$
$$s.t. \quad \langle \phi, d \rangle = \langle \phi, (1-\gamma) d_0^\pi + \gamma \mathcal{P}_*^\pi d \rangle.$$

Proof. Let $L_{\mathrm{P}}(\beta, d)$ and $L_{\mathrm{D}}(d, \beta)$ be the Lagrangian of the primal and dual embedded Q-LP (4.24) and (4.25), respectively. By Lemma 3.2.3, $\mathcal{P}_*^\pi$ is the adjoint of $\mathcal{P}^\pi$. Thus, the conditions of Lemma 2.2.2 hold,

$$\begin{aligned} L_{\mathrm{P}}(\beta, d) &= (1-\gamma)\mathbb{E}_{(s_0,a_0)\sim d_0^\pi}[Q_\beta(s_0,a_0)] + \langle d, \mathcal{B}^\pi Q_\beta - Q_\beta \rangle \\ &= (1-\gamma)\langle Q_\beta, d_0^\pi \rangle + \langle r, d \rangle + \gamma \langle \mathcal{P}^\pi Q_\beta, d \rangle - \langle Q_\beta, d \rangle \\ &= \mathbb{E}_{(s,a)\sim d}[r(s,a)] + \beta^\top \langle \phi, (1-\gamma) d_0^\pi + \gamma \mathcal{P}_*^\pi d - d \rangle = L_{\mathrm{D}}(d, \beta). \end{aligned}$$

□

Lemma 4.6.2. *Consider the* modified primal embedded Q-LP

$$\min_{\beta \in \mathbb{R}^K} \|Q^\pi - Q_\beta\|_{L_1(d_0^\pi)} \tag{4.26}$$
$$s.t. \quad \forall (s,a) \in S \times A: \ Q_\beta(s,a) \geq r(s,a) + \gamma \mathcal{P}^\pi Q_\beta(s,a).$$

It shares the same optimal solution with the primal embedded Q-LP (4.24).

Proof. Consider an arbitrary $\beta \in \mathbb{R}^K$, where Q_β is feasible for the primal embedded Q-LP (4.24), i.e., $Q_\beta \geq \mathcal{B}^\pi Q_\beta$. Since $\mathcal{B}^\pi$ is a monotonic γ-contraction, we can apply Banach iteration to Q and get

$$Q_\beta \geq \mathcal{B}^\pi Q_\beta \geq (\mathcal{B}^\pi)^2 Q_\beta \geq (\mathcal{B}^\pi)^3 Q_\beta \geq \cdots \geq \lim_{t\to\infty} (\mathcal{B}^\pi)^t Q_\beta = Q^\pi.$$

Therefore, we can omit the absolute value inside $\|Q_\beta - Q^\pi\|_{L_1(d_0^\pi)}$ and get

$$\begin{aligned}
&\mathbb{E}_{d_0^\pi}[Q^\pi] + \|Q_\beta - Q^\pi\|_{L_1(d_0^\pi)} \\
&\quad = \int_{S\times A} Q^\pi(s,a) d_0^\pi(s,a)\ \mathrm{d}s\ \mathrm{d}a + \int_{S\times A} (Q_\beta(s,a) - Q^\pi(s,a)) d_0^\pi(s,a)\ \mathrm{d}s\ \mathrm{d}a \\
&\quad = \mathbb{E}_{d_0^\pi}[Q_\beta].
\end{aligned}$$

Because $\mathbb{E}_{d_0^\pi}[Q^\pi] = \rho^\pi$ is constant in β,

$$\arg\min_{\beta\in\mathbb{R}^K} \mathbb{E}_{d_0^\pi}[Q_\beta] = \arg\min_{\beta\in\mathbb{R}^K} \|Q^\pi - Q_\beta\|_{L_1(d_0^\pi)}.$$

□

We can use Lemma 4.6.1 for Theorem 4.6.3. It states that the error we make, by embedding our constraints, for estimating the policy value ρ^π by ρ_ϕ^π, can be bounded by how well we can approximate Q^π by functions from $\mathcal{F}_\phi$.

Theorem 4.6.3 (CoinDICE approximation error)**.** *Assume that $\mathcal{F}_\phi$ contains the constant one function. Then*

$$0 \leq \rho_\phi^\pi - \rho^\pi \leq 2 \min_{\beta\in\mathbb{R}^K} \|Q^\pi - Q_\beta\|_\infty.$$

Proof. Consider the optimal solution to the embedded Q-LP from Lemma 4.6.1,

$$(d_*, \beta_*) \doteq \arg\min_{\beta\in\mathbb{R}^K} \max_{d: S\times A\to\mathbb{R}_{\geq 0}} L(\beta, d).$$

Furthermore, let

$$\beta_\star \doteq \arg\min_{\beta\in\mathbb{R}^K} \|Q^\pi - Q_\beta\|_\infty \quad \text{and} \quad \epsilon \doteq \min_{\beta\in\mathbb{R}^K} \|Q^\pi - Q_\beta\|_\infty = \|Q^\pi - Q_{\beta_\star}\|_\infty.$$

By the forwards Bellman equations (3.5) and the fact that the forwards Bellman operator is a γ-contraction with respect to the norm $\|\cdot\|_\infty$, we get

$$\mathcal{B}^\pi Q_{\beta_\star} - Q^\pi \leq \sup_{(s,a)\in S\times A} |\mathcal{B}^\pi Q_{\beta_\star}(s,a) - Q^\pi(s,a)| = \|\mathcal{B}^\pi Q_{\beta_\star} - \mathcal{B}^\pi Q^\pi\|_\infty \leq \gamma\epsilon.$$

Now, let

$$(1-\gamma)c + (1+\gamma)\epsilon \stackrel{!}{=} 0 \iff c \doteq -\frac{1+\gamma}{1-\gamma}\epsilon \implies (1-\gamma)(\epsilon + |c|) = 2\epsilon.$$

Because the expected Bellman operator $\mathcal{P}^\pi$ is linear,

$$\mathcal{B}^\pi(Q_{\beta_\star} - c) = r + \gamma\mathcal{P}^\pi(Q_{\beta_\star} - c) = r + \gamma\mathcal{P}^\pi Q_{\beta_\star} - \gamma c = \mathcal{B}^\pi Q_{\beta_\star} - \gamma c.$$

By definition of $\beta_\star$ and ϵ, we have

$$Q^\pi = Q_{\beta_\star} + Q^\pi - Q_{\beta_\star} \leq Q_{\beta_\star} + \epsilon.$$

Putting all of this together yields

$$\begin{aligned}\mathcal{B}^\pi(Q_{\beta_\star} - c) &= \mathcal{B}^\pi Q_{\beta_\star} - \gamma c \\ &\leq Q^\pi + \gamma\epsilon - \gamma c \\ &\leq Q_{\beta_\star} + \epsilon + \gamma\epsilon - \gamma c \\ &= Q_{\beta_\star} - c + (1-\gamma)c + (1+\gamma)\epsilon \\ &= Q_{\beta_\star} - c.\end{aligned}$$

Recall that the solution Q^π to the primal Q-LP (3.10) is better than the solution Q_{β_*} to the embedded primal Q-LP (4.24). Thus,

$$0 \leq \rho_\phi^\pi - \rho^\pi = (1-\gamma)\mathbb{E}_{(s,a)\sim d_0^\pi}[Q_{\beta_*}] - (1-\gamma)\mathbb{E}_{(s,a)\sim d_0^\pi}[Q^\pi].$$

By our assumption, $Q_{\beta_\star} - c \in \mathcal{F}_\phi$. Therefore, there exists a $\bar{\beta} \in \mathbb{R}^K$, such that $Q_{\bar{\beta}} = Q_{\beta_\star} - c$. We have just shown that $Q_{\bar{\beta}} \geq \mathcal{B}^\pi Q_{\bar{\beta}}$. So, $Q_{\bar{\beta}}$ is a feasible solution. By Lemma 4.6.2, Q_{β_*} is the optimal solution to the modified embedded primal Q-LP (4.26). Therefore, we further get

$$\leq (1-\gamma)\|Q_{\beta_*} - Q^\pi\|_{L_1(d_0^\pi)} \leq (1-\gamma)\|Q_{\bar{\beta}} - Q^\pi\|_{L_1(d_0^\pi)}.$$

Because d_0^π is a distribution, $\|\cdot\|_{L_1(d_0^\pi)} \leq \|\cdot\|_\infty$. Applying the triangle inequality, we further calculate

$$\begin{aligned}&\leq (1-\gamma)\|Q_{\bar{\beta}} - Q^\pi\|_\infty \leq (1-\gamma)(\|Q_{\bar{\beta}} - Q_{\beta_\star}\|_\infty + \|Q_{\beta_\star} - Q^\pi\|_\infty) \\ &= (1-\gamma)(|c| + \epsilon) = 2\epsilon.\end{aligned}$$

The claim follows from the definition of ϵ.

□

4.6.2 Generalized Estimating Equations

We now want to rewrite the dual constraints from the embedded Q-LP in Lemma 4.6.1 as *generalized estimating equations* in Lemma 4.6.4. To this end, we introduce a notation to bundle the samples in the spaces

$$X \doteq S \times S \times A \times \mathbb{R} \times S \quad \text{and} \quad Y \doteq S \times A \times S \times A \times \mathbb{R} \times S \times A.$$

For any $w : S \times A \to \mathbb{R}_{\geq 0}$, define the function

$$\iota(\cdot; w) : \begin{cases} Y \to \mathbb{R}^K \\ y \mapsto (1-\gamma)\phi(s_0, a_0) + w(s,a)(\gamma\phi(s', a') - \phi(s,a)). \end{cases}$$

Integrating over a_0 and a' by using π, we define

$$\iota^\pi(\cdot; w) : \begin{cases} X \to \mathbb{R}^K \\ x \mapsto \mathbb{E}_{a_0\sim\pi(s_0),\ a'\sim\pi(s')}[\iota(s_0, a_0, s, a, r, s', a'; w)]. \end{cases}$$

Lemma 4.6.4 (generalized estimating equations)**.** *For any $d \in \Delta_{d^{\mathcal{D}}}(S \times A)$ we let $w \doteq d/d^{\mathcal{D}}$ and for any $w : S \times A \to \mathbb{R}_{\geq 0}$ with $\mathbb{E}_{(s,a)\sim d^{\mathcal{D}}}[w(s,a)] = 1$, we let $d \doteq wd^{\mathcal{D}}$. Then, we have*

$$\langle \phi, d \rangle = \langle \phi, (1-\gamma)d_0^\pi + \gamma \mathcal{P}_*^\pi d \rangle \iff \mathbb{E}_{x\sim p^{\mathcal{D}}}[\iota^\pi(x; w)] = \vec{0}.$$

Proof. Firstly,

$$\mathbb{E}_{(s_0,a_0)\sim d_0^\pi}[(1-\gamma)\phi(s_0,a_0)] = \int_{S\times A} (1-\gamma)\phi(s_0,a_0) d_0^\pi(s_0,a_0)\ \mathrm{d}s_0\ \mathrm{d}a_0 = \langle \phi, (1-\gamma)d_0^\pi \rangle,$$

and secondly,

$$\begin{aligned}
&\mathbb{E}_{(s,a)\sim d^{\mathcal{D}},\ (s',a')\sim T^\pi(s,a)}[w(s,a)(\gamma\phi(s',a') - \phi(s,a))] \\
&= \int_{S\times A}\int_{S\times A} \frac{d(s,a)}{d^{\mathcal{D}}(s,a)}(\gamma\phi(s',a') - \phi(s,a)) T^\pi(s',a' \mid s,a) d^{\mathcal{D}}(s,a)\ \mathrm{d}s\ \mathrm{d}a\ \mathrm{d}s'\ \mathrm{d}a' \\
&= \int_{S\times A} \phi(s',a')\gamma \int_{S\times A} d(s,a) T^\pi(s',a' \mid s,a)\ \mathrm{d}s\ \mathrm{d}a\ \mathrm{d}s'\ \mathrm{d}a' \\
&\qquad - \int_{S\times A} \phi(s,a) d(s,a) \int_{S\times A} T^\pi(s',a' \mid s,a)\ \mathrm{d}s'\ \mathrm{d}a'\ \mathrm{d}s\ \mathrm{d}a \\
&= \int_{S\times A} \phi(s',a')\gamma\mathcal{P}_*^\pi d(s',a')\ \mathrm{d}s'\ \mathrm{d}a' - \int_{S\times A} \phi(s,a) d(s,a)\ \mathrm{d}s\ \mathrm{d}a \\
&= \langle \phi, \gamma\mathcal{P}_*^\pi d \rangle - \langle \phi, d \rangle.
\end{aligned}$$

Therefore,

$$\begin{aligned}
\mathbb{E}_{x\sim p^{\mathcal{D}}}[\iota^\pi(x; w)] &= \mathbb{E}_{(s_0,a_0)\sim d_0^\pi}[(1-\gamma)\phi(s_0,a_0)] \\
&\quad + \mathbb{E}_{(s,a)\sim d^{\mathcal{D}},\ (s',a')\sim T^\pi(s,a)}[w(s,a)(\gamma\phi(s',a') - \phi(s,a))] \\
&= \langle \phi, (1-\gamma)d_0^\pi + \gamma\mathcal{P}_*^\pi d \rangle - \langle \phi, d \rangle.
\end{aligned}$$

□

By substituting $w = d/d^{\mathcal{D}}$ and using the generalized estimating equations, we rewrite the dual embedded Q-LP (4.25) as

$$\begin{aligned}
\rho_\phi^\pi = \max_{w: S\times A \to \mathbb{R}_{\geq 0}}\ & \mathbb{E}_{(s,a)\sim d^{\mathcal{D}}}[w(s,a) r(s,a)] \\
\text{s.t.}\quad & \mathbb{E}_{x\sim p^{\mathcal{D}}}[\iota^\pi(x; w)] = \vec{0}.
\end{aligned} \tag{4.27}$$

Consider the Lagrangian integrand

$$\ell(\cdot; w, \beta) : \begin{cases} Y \to \mathbb{R} \\ y \mapsto w(s,a) r(s,a) + \beta^\top \iota(y; w). \end{cases}$$

We can rewrite it as

$$\begin{aligned}
\ell(y; w, \beta) &= w(s,a) r(s,a) \\
&\quad + (1-\gamma)\beta^\top \phi(s_0,a_0) \\
&\quad + w(s,a)(\gamma\beta^\top\phi(s',a') - \beta^\top\phi(s,a)) \\
&= (1-\gamma) Q_\beta(s_0,a_0) \\
&\quad + w(s,a)(r(s,a) + \gamma Q_\beta(s',a') - Q_\beta(s,a)).
\end{aligned} \tag{4.28}$$

Integrating over a_0 and a' by using π, we define

$$\ell^\pi(\cdot; w, \beta) : \begin{cases} X \to \mathbb{R} \\ x \mapsto \mathbb{E}_{a_0 \sim \pi(s_0),\ a' \sim \pi(s')}[\ell(s_0, a_0, s, a, r, s', a'; w, \beta)]. \end{cases}$$

Now, (4.27) is an LP. By Lagrange duality (2.9), we get

$$\rho^\pi_\phi = \max_{w: S\times A \to \mathbb{R}_{\geq 0}} \min_{\beta \in \mathbb{R}^K} \mathbb{E}_{x\sim p^{\mathcal{D}}}[\ell^\pi(x; w, \beta)] = \min_{\beta \in \mathbb{R}^K} \max_{w: S\times A \to \mathbb{R}_{\geq 0}} \mathbb{E}_{x\sim p^{\mathcal{D}}}[\ell^\pi(x; w, \beta)]. \tag{4.29}$$

4.6.3 Confidence Interval Derivation

We now want to leverage Theorem 2.6.8, to find a confidence interval for our estimator ρ^π_ϕ. Our confidence interval can be formulated by using (4.27) or (4.29), as

$$\begin{aligned} C^f_{n,\xi} &\doteq \left\{ \max_{w: S\times A \to \mathbb{R}_{\geq 0}} \mathbb{E}_{x\sim p}[w(s,a)r(s,a)] \;\middle|\; p \in B^f_{\xi/n}(\hat{p}^{\mathcal{D}}),\ \mathbb{E}_{x\sim p}[\iota^\pi(x; w)] = 0 \right\} \\ &= \left\{ \max_{w: S\times A \to \mathbb{R}_{\geq 0}} \min_{\beta \in \mathbb{R}^K} \mathbb{E}_{x\sim p}[\ell^\pi(x; w, \beta)] \;\middle|\; p \in B^f_{\xi/n}(\hat{p}^{\mathcal{D}}) \right\} \\ &= \left\{ \min_{\beta \in \mathbb{R}^K} \max_{w: S\times A \to \mathbb{R}_{\geq 0}} \mathbb{E}_{x\sim p}[\ell^\pi(x; w, \beta)] \;\middle|\; p \in B^f_{\xi/n}(\hat{p}^{\mathcal{D}}) \right\} \end{aligned}$$

Since the objective of $C^f_{n,\xi} \subset \mathbb{R}$ is convex in p and the constraints are a convex, the set is also convex. The same goes for closedness. Therefore, we are dealing with a closed interval. Theorem 4.6.9 claims, that $C^f_{n,\xi}$ is a confidence interval.

Assumption 4.6.5 (Stationary ratio regularity)**.** Let $\mathcal{H}_w$ be a bounded RKHS, with kernel function k bounded by $K < \infty$. The stationary distribution correction $w_{\pi/\mathcal{D}}$ is part of a compact subset $\mathcal{F}_w \subset \mathcal{H}_w$, where

$$\exists C_w < \infty :\ \forall w \in \mathcal{F}_w :\ \|w\|_\infty \leq C_w.$$

Assumption 4.6.5 together with (2.10) shows that

$$\forall w \in \mathcal{F}_w,\ \forall (s,a) \in S \times A :\ |w(s,a)| \leq K\|w\|_{\mathcal{F}_w}.$$

Assumption 4.6.6 (Embedding feature regularity)**.** Let $\mathcal{F}_\beta \subseteq \mathbb{R}^K$ be a compact set of feature coefficients. The features and their coefficients are universally bounded, i.e.,

$$\exists C_\phi < \infty :\ \|\phi\|_2 \leq C_\phi \quad \text{and} \quad \exists C_\beta < \infty :\ \forall \beta \in \mathcal{F}_\beta :\ \|\beta\|_2 \leq C_\beta.$$

Assumption 4.6.6 and the Cauchy-Schwarz inequality imply

$$\forall \beta \in \mathcal{F}_\beta,\ \forall (s,a) \in S \times A :\ |Q_\beta(s,a)| = |\beta^\top \phi(s,a)| \leq \|\beta\|_2 \|\phi(s,a)\|_2 \leq C_\beta C_\phi.$$

Lemma 4.6.7. *Let Assumptions 4.6.5 and 4.6.6 hold. Then $\ell(y; w, \beta)$ is bounded and C_ℓ-Lipschitz-continuous in (w, β) with some $C_\ell \in \mathbb{R}$.*

Proof. 1. $\ell(y; w, \beta)$ is bounded, because

$$\begin{aligned} |\ell(y; w, \beta)| &= (1-\gamma)|Q_\beta(s_0, a_0)| + |w(s,a)|(|r(s,a)| + \gamma|Q_\beta(s', a')| + |Q_\beta(s,a)|) \\ &\leq (1-\gamma)C_\beta C_\phi + C_w(r_{\max} + (1+\gamma)C_\beta C_\phi). \end{aligned}$$

2. $\ell(y; w, \beta)$ is Lipschitz-continuous in (w, β), because

$$
\begin{aligned}
&|\ell(y; w_1, \beta_1) - \ell(y; w_2, \beta_2)| \\
&\quad = \Big| \Big(w_1(s,a)r(s,a) + \beta_1^\top((1-\gamma)\phi(s_0,a_0) + w_1(s,a)(\gamma\phi(s',a') - \phi(s,a)))\Big) \\
&\qquad - \Big(w_2(s,a)r(s,a) + \beta_2^\top((1-\gamma)\phi(s_0,a_0) + w_2(s,a)(\gamma\phi(s',a') - \phi(s,a)))\Big)\Big| \\
&\quad \le (1-\gamma)|(\beta_1 - \beta_2)^\top\phi(s_0,a_0)| + |(w_1(s,a) - w_2(s,a))r(s,a)| \\
&\qquad + \Big| \Big(w_1(s,a)\beta_1^\top(\gamma\phi(s',a') - \phi(s,a))\Big) + \Big(w_1(s,a)\beta_2^\top(\gamma\phi(s',a') - \phi(s,a))\Big) \\
&\qquad - \Big(w_1(s,a)\beta_2^\top(\gamma\phi(s',a') - \phi(s,a))\Big) - \Big(w_2(s,a)\beta_2^\top(\gamma\phi(s',a') - \phi(s,a))\Big)\Big| \\
&\quad \le (1-\gamma)|(\beta_1 - \beta_2)^\top\phi(s_0,a_0)| + |(w_1(s,a) - w_2(s,a))r(s,a)| \\
&\qquad + \Big|w_1(s,a)(\beta_1 - \beta_2)^\top(\gamma\phi(s',a') - \phi(s,a))\Big| \\
&\qquad + \Big|(w_1(s,a) - w_2(s,a))\beta_2^\top(\gamma\phi(s',a') - \phi(s,a))\Big| \\
&\quad \le (1-\gamma)\|\beta_1 - \beta_2\|_2\|\phi(s_0,a_0)\|_2 + |w_1(s,a) - w_2(s,a)|r_{\max} \\
&\qquad + |w_1(s,a)|\|\beta_1 - \beta_2\|_2(\gamma\|\phi(s',a')\|_2 + \|\phi(s,a)\|_2) \\
&\qquad + |w_1(s,a) - w_2(s,a)|\|\beta_2\|_2(\gamma\|\phi(s',a')\|_2 + \|\phi(s,a)\|_2) \\
&\quad \le (1-\gamma)\|\beta_1 - \beta_2\|_2 C_\phi + K\|w_1 - w_2\|_{\mathcal{F}_w} r_{\max} \\
&\qquad + C_w\|\beta_1 - \beta_2\|_2(1+\gamma)C_\phi + \|w_1 - w_2\|_{\mathcal{F}_w} C_\beta(1+\gamma)C_\phi \\
&\quad \le C_\ell(\|\beta_1 - \beta_2\|_2 + \|w_1 - w_2\|_{\mathcal{F}_w}),
\end{aligned}
$$

where

$$C_\ell \doteq \max\{(1+C_w)(1-\gamma)C_\phi, Kr_{\max} + (1+\gamma)C_\phi C_\beta\}.$$

□

Lemma 4.6.8. *Let Assumptions 4.6.6 and 4.6.5 hold. Consider the class of functions*

$$\mathcal{H} \doteq \{\ell^\pi(\cdot; w, \beta) \mid w \in \mathcal{F}_w,\ \beta \in \mathcal{F}_\beta\}$$

and the functional

$$L^* : \begin{cases} \mathcal{P}(X) \to \mathbb{R} \\ P \mapsto \min_{\beta\in\mathcal{F}_\beta} \max_{w\in\mathcal{F}_w} \mathbb{E}_P[\ell^\pi(\cdot; w, \beta)]. \end{cases}$$

Let

$$P \in \mathcal{P}(X) \quad \text{and} \quad (\beta^*, w^*) \doteq \arg\min_{\beta\in\mathcal{F}_\beta} \max_{w\in\mathcal{F}_w} \mathbb{E}_P[\ell^\pi(\cdot; w, \beta)].$$

Then L^* is Hadamard differentiable at P tangentially to $B(\mathcal{H}, P) \subset L_\infty(\mathcal{H})$ with Hadamard derivative

$$\partial L_P^*(H) = H\ell^\pi(\cdot; w^*, \beta^*), \quad \text{where} \quad H \in B(\mathcal{H}, P).$$

In particular, ∂L_P^* is a bounded linear functional on the space of bounded measures with the canonical gradient as influence function

$$L^{(1)}(\cdot; P) \doteq \ell^\pi(\cdot; w^*, \beta^*) - \mathbb{E}_P[\ell^\pi(\cdot; w^*, \beta^*)].$$

Proof. Chose $(t_n)_{n\in\mathbb{N}} \subset \mathbb{R}$ and $(H_n)_{n\in\mathbb{N}} \subset \mathcal{H}$, such that

$$t_n \xrightarrow{n\to\infty} 0, \quad \|H_n - H\|_{L_\infty(\mathcal{H})} \xrightarrow{n\to\infty} 0, \quad \text{and} \quad P + t_n H_n \in \mathcal{P}(X) \text{ for all } n \in \mathbb{N}.$$

Firstly, we show upper bound convergence. Start with

$$\begin{aligned}
&L^*(P + t_n H_n) - L^*(P) \\
&\quad = \min_{\beta\in\mathcal{F}_\beta} \max_{w\in\mathcal{F}_w} \big(\mathbb{E}_P[\ell^\pi(\cdot; w, \beta)] + t_n H_n \ell^\pi(\cdot; w, \beta)\big) - \min_{\beta\in\mathcal{F}_\beta} \max_{w\in\mathcal{F}_w} \mathbb{E}_P[\ell^\pi(\cdot; w, \beta)] \\
&\quad \leq \max_{w\in\mathcal{F}_w} \big(\mathbb{E}_P[\ell^\pi(\cdot; w, \beta^*)] + t_n H_n \ell^\pi(\cdot; w, \beta^*)\big) - \mathbb{E}_P[\ell^\pi(\cdot; w, \beta^*)] \\
&\quad \leq \max_{w\in\mathcal{F}_w} t_n H_n \ell^\pi(\cdot; w, \beta^*).
\end{aligned}$$

Define

$$w_n^* \doteq \arg\max_{w\in\mathcal{F}_w} H_n \ell^\pi(\cdot; w, \beta^*).$$

Then,

$$\max_{w\in\mathcal{F}_w} H_n \ell^\pi(\cdot; w, \beta^*) - \max_{w\in\mathcal{F}_w} H \ell^\pi(\cdot; w, \beta^*) \leq H_n \ell^\pi(\cdot; w_n^*, \beta^*) - H\ell^\pi(\cdot; w_n^*, \beta^*) \leq \|H_n - H\|_{L_\infty(\mathcal{H})}.$$

Therefore,

$$\limsup_{n\to\infty} \frac{L^*(P + t_n H_n) - L^*(P)}{t_n} \leq H\ell^\pi(\cdot; w^*, \beta^*).$$

Secondly, we show lower bound convergence. Define

$$w_n(\beta) \doteq \arg\max_{w\in\mathcal{F}_w} \big(\mathbb{E}_P[\ell^\pi(\cdot; w, \beta)] + t_n H_n \ell^\pi(\cdot; w, \beta)\big)$$

Then,

$$\begin{aligned}
&L^*(P + t_n H_n) \\
&\quad = \min_{\beta\in\mathcal{F}_\beta} \max_{w\in\mathcal{F}_w} \mathbb{E}_P[\ell^\pi(\cdot; w, \beta)] + t_n H_n \ell^\pi(\cdot; w, \beta) \\
&\quad = \min_{\beta\in\mathcal{F}_\beta} \mathbb{E}_P[\ell^\pi(\cdot; w_n(\beta), \beta)] \\
&\qquad\qquad + t_n\big(H_n \ell^\pi(\cdot; w_n(\beta), \beta) - H\ell^\pi(\cdot; w_n(\beta), \beta)\big) + t_n H \ell^\pi(\cdot; w_n(\beta), \beta) \\
&\quad \leq \min_{\beta\in\mathcal{F}_\beta} \mathbb{E}_P[\ell^\pi(\cdot; w_n(\beta), \beta)] + t_n \|H_n - H\|_{L_\infty(\mathcal{H})} + t_n \|H\|_{L_\infty(\mathcal{H})} \\
&\quad \leq \min_{\beta\in\mathcal{F}_\beta} \mathbb{E}_P[\ell^\pi(\cdot; w_n(\beta), \beta)] + \mathcal{O}(t_n).
\end{aligned}$$

Define the ϵ-ball

$$B_\epsilon(P) \doteq \left\{ \beta' \in \mathcal{F}_\beta \;\middle|\; \max_{w\in\mathcal{F}_w} \mathbb{E}_P[\ell^\pi(\cdot; w, \beta')] \leq \min_{\beta\in\mathcal{F}_\beta} \max_{w\in\mathcal{F}_w} \mathbb{E}_P[\ell^\pi(\cdot; w, \beta)] + \epsilon \right\}.$$

For every $n \in \mathbb{N}$, take a solution $\beta_n^* \in B_0(P + t_n H_n)$. According to the above, this means that there exists a $C > 0$ such that $\beta_n^* \in B_{t_n C}(P)$, for every $n \in \mathbb{N}$. This means that $(\beta_n^*)_{n\in\mathbb{N}}$ is bounded and therefore, has a convergent sub sequence against $\beta^* \in B_0(P)$. W.l.o.g. let

$(\beta_n^*)_{n\in\mathbb{N}}$ be that sub sequence itself. Because $\ell^\pi(\cdot; w, \beta)$ is bounded and Lipschitz-continuous in (w, β), we have that

$$\lim_{n\to\infty} \max_{w\in\mathcal{F}_w} \mathbb{E}_P[\ell^\pi(\cdot; w, \beta_n^*)] = L^*(P).$$

Due to optimality, we also have

$$\inf_{n\in\mathbb{N}} \max_{w\in\mathcal{F}_w} \mathbb{E}_P[\ell^\pi(\cdot; w, \beta_n^*)] \geq L^*(P).$$

Define

$$w_n \doteq \arg\max_{w\in\mathcal{F}_w} \mathbb{E}_P[\ell^\pi(\cdot; w, \beta_n^*)].$$

Then,

$$\begin{aligned}
&L^*(P + t_n H_n) - L^*(P) \\
&\quad\geq \max_{w\in\mathcal{F}_w} \mathbb{E}_P[\ell^\pi(\cdot; w, \beta_n^*)] + t_n H_n \ell^\pi(\cdot; w, \beta_n^*) - \max_{w\in\mathcal{F}_w} \mathbb{E}_P[\ell^\pi(\cdot; w, \beta_n^*)] \\
&\quad\geq \mathbb{E}_P[\ell^\pi(\cdot; w_n, \beta_n^*)] + t_n H_n \ell^\pi(\cdot; w_n, \beta_n^*) - \mathbb{E}_P[\ell^\pi(\cdot; w_n, \beta_n^*)] \\
&\quad= t_n H_n \ell^\pi(\cdot; w_n, \beta_n^*).
\end{aligned}$$

Also, we have $w_n \xrightarrow{n\to\infty} w^*$, so

$$\begin{aligned}
&|H_n \ell^\pi(\cdot; w_n, \beta_n^*) - H\ell^\pi(\cdot; w^*, \beta^*)| \\
&\quad\leq |H_n \ell^\pi(\cdot; w_n, \beta_n^*) - H\ell^\pi(\cdot; w_n, \beta_n^*)| + |H\ell^\pi(\cdot; w_n, \beta_n^*) - H\ell^\pi(\cdot; w^*, \beta^*)| \\
&\quad\leq \|H_n - H\|_{L_\infty(\mathcal{H})} + |H\ell^\pi(\cdot; w_n, \beta_n^*) - H\ell^\pi(\cdot; w^*, \beta^*)| \xrightarrow{n\to\infty} 0.
\end{aligned}$$

Therefore,

$$\liminf_{n\to\infty} \frac{L^*(P + t_n H_n) - L^*(P)}{t_n} \geq H\ell^\pi(\cdot; w^*, \beta^*).$$

□

Theorem 4.6.9 (CoinDICE asymptotic coverage)**.** *Let $\mathcal{D}$ contain i.i.d. samples and the embedded Q-LP from Lemma 4.6.1 have a unique solution. Also, let Assumptions 2.6.6, 4.6.5 and 4.6.6 hold. Then, we have that $C_{n,\chi^{2,1-\alpha}_{(1)}}$ is an asymptotic $(1-\alpha)$-confidence interval of ρ^π, i.e.,*

$$\lim_{n\to\infty} \mathbb{P}\left(\rho^\pi \in C^f_{n,\xi}\right) = \mathbb{P}\left(\chi^2_{(1)} \leq \xi\right).$$

Proof. We will apply Theorem 2.6.8 to prove this claim. Recall the definitions in lemma 4.6.7 4.6.8. By Lemma 4.6.7, $\ell^\pi(\cdot; w, \beta)$ is bounded and Lipschitz continuous in (w, β). By Assumptions 4.6.5 and 4.6.6, the sets $\mathcal{F}_w$ and $\mathcal{F}_\beta$ are both compact. Finally, $\mathcal{H} \subset L_2(P^\mathcal{D})$. By Lemma 2.6.5, $\mathcal{H}$ is $p^\mathcal{D}$-Donsker with L_2-integrable envelope. Lemma 4.6.8 provides the last requirements to apply Theorem 2.6.8.

□

4.6.4 Confidence Interval Calculation

In the following Theorem 4.6.10, we discuss how to calculate the lower and upper bound for our confidence interval $C^f_{n,\xi}$.

Theorem 4.6.10 (CoinDICE upper and lower confidence bounds)**.** *Let l_n and u_n denote the lower and upper confidence bounds of $C^f_{n,\xi}$, respectively. Then*

$$\begin{aligned}
l_n &= \min_{\beta\in\mathbb{R}^K} \max_{w:S\times A\to\mathbb{R}_{\geq 0}} \min_{p\in B^f_{\xi/n}(\hat{p}^{\mathcal{D}})} \mathbb{E}_{x\sim p}[\ell^\pi(x;w,\beta)], \\
&= \min_{\beta\in\mathbb{R}^K} \max_{w:S\times A\to\mathbb{R}_{\geq 0}} \min_{\substack{\lambda\in\mathbb{R}_{\geq 0},\\ \eta\in\mathbb{R}}} \mathbb{E}_{x\sim\hat{p}^{\mathcal{D}}}\left[-\lambda f_*\left(\frac{\eta-\ell^\pi(x;w,\beta)}{\lambda}\right) - \lambda\frac{\xi}{n} + \eta\right], \\
u_n &= \max_{w:S\times A\to\mathbb{R}_{\geq 0}} \min_{\beta\in\mathbb{R}^K} \max_{p\in B^f_{\xi/n}(\hat{p}^{\mathcal{D}})} \mathbb{E}_{x\sim p}[\ell^\pi(x;w,\beta)]. \\
&= \max_{w:S\times A\to\mathbb{R}_{\geq 0}} \min_{\beta\in\mathbb{R}^K} \min_{\substack{\lambda\in\mathbb{R}_{\geq 0},\\ \eta\in\mathbb{R}}} \mathbb{E}_{x\sim\hat{p}^{\mathcal{D}}}\left[\lambda f_*\left(\frac{\ell^\pi(x;w,\beta)-\eta}{\lambda}\right) + \lambda\frac{\xi}{n} + \eta\right].
\end{aligned}$$

The optimal weights for lower and upper confidence bounds, respectively, are

$$p_l = f'_*\left(\frac{\eta-\ell^\pi(x;w,\beta)}{\lambda}\right)\hat{p}^{\mathcal{D}}(x) \quad \textit{and} \quad p_u = f'_*\left(\frac{\ell^\pi(x;w,\beta)-\eta}{\lambda}\right)\hat{p}^{\mathcal{D}}(x). \tag{4.30}$$

Proof. W.l.o.g. we only calculate the upper bound u_n.

By (4.28), one can see that $\ell(x;w,\beta)$ is linear in w and β, respectively. Also, $B^f_{\xi/n}(\hat{p}^{\mathcal{D}})$ is compact and convex and $\mathbb{R}^K$ is convex anyways. Therefore, we can apply Sion's theorem and get

$$\begin{aligned}
u_n &= \max_{p\in B^f_{\xi/n}(\hat{p}^{\mathcal{D}})} \max_{w:S\times A\to\mathbb{R}_{\geq 0}} \min_{\beta\in\mathbb{R}^K} \mathbb{E}_{x\sim p}[\ell^\pi(x;w,\beta)] \\
&= \max_{w:S\times A\to\mathbb{R}_{\geq 0}} \max_{p\in B^f_{\xi/n}(\hat{p}^{\mathcal{D}})} \min_{\beta\in\mathbb{R}^K} \mathbb{E}_{x\sim p}[\ell^\pi(x;w,\beta)] \\
&= \max_{w:S\times A\to\mathbb{R}_{\geq 0}} \min_{\beta\in\mathbb{R}^K} \max_{p\in B^f_{\xi/n}(\hat{p}^{\mathcal{D}})} \mathbb{E}_{x\sim p}[\ell^\pi(x;w,\beta)].
\end{aligned}$$

This proves the first claim.

For the second claim, we rewrite the inner maximization in terms of a Lagrangian. Since $D_f(p \parallel \hat{p}^{\mathcal{D}})$ and $\|p\|_1$ are convex and linear in $p : S\times A\to\mathbb{R}_{\geq 0}$, respectively, we can apply Lagrange duality (2.9) and get

$$\begin{aligned}
&\max_{p\in B^f_{\xi/n}(\hat{p}^{\mathcal{D}})} \mathbb{E}_{x\sim p}[\ell^\pi(x;w,\beta)] \\
&\quad= \max_{0\leq p\ll\hat{p}^{\mathcal{D}}} \min_{\substack{\lambda\in\mathbb{R}_{\geq 0},\\ \eta\in\mathbb{R}}} \mathbb{E}_{x\sim p}[\ell^\pi(x;w,\beta)] - \lambda\left(D_f(p \parallel \hat{p}^{\mathcal{D}}) - \frac{\xi}{n}\right) - \eta\left(\|p\|_1 - 1\right) \\
&\quad= \min_{\substack{\lambda\in\mathbb{R}_{\geq 0},\\ \eta\in\mathbb{R}}} \max_{0\leq p\ll\hat{p}^{\mathcal{D}}} \mathbb{E}_{x\sim p}[\ell^\pi(x;w,\beta)] - \lambda\left(D_f(p \parallel \hat{p}^{\mathcal{D}}) - \frac{\xi}{n}\right) - \eta\left(\|p\|_1 - 1\right).
\end{aligned}$$

Again, we rewrite the inner maximization. This time, we use the substitution

$$q = p/\hat{p}^{\mathcal{D}} \quad \text{for} \quad 0\leq p\ll\hat{p}^{\mathcal{D}} \quad\Longleftrightarrow\quad p = q\hat{p}^{\mathcal{D}} \quad \text{for} \quad q : S\times A\to\mathbb{R}_{\geq 0}.$$

We get

$$\begin{aligned}
&\max_{0\le p\ll\hat{p}^{\mathcal{D}}} \mathbb{E}_{x\sim p}[\ell^\pi(x;w,\beta)] - \lambda\mathbb{E}_{x\sim\hat{p}^{\mathcal{D}}}\left[f\left(\frac{p(x)}{\hat{p}^{\mathcal{D}}(x)}\right)\right] + \lambda\frac{\xi}{n} - \eta\mathbb{E}_{x\sim p}[1] + \eta \\
&\quad= \max_{q:S\times A\to\mathbb{R}_{\ge 0}} \mathbb{E}_{x\sim\hat{p}^{\mathcal{D}}}[\ell^\pi(x;w,\beta)q(x)] - \lambda\mathbb{E}_{x\sim\hat{p}^{\mathcal{D}}}[f(q(x))] - \eta\mathbb{E}_{x\sim\hat{p}^{\mathcal{D}}}[q(x)] + \lambda\frac{\xi}{n} + \eta \\
&\quad= \mathbb{E}_{x\sim\hat{p}^{\mathcal{D}}}\left[\lambda\max_{q\in\mathbb{R}_{\ge 0}}\left(\frac{\ell^\pi(x;w,\beta)-\eta}{\lambda}q - f(q)\right) + \lambda\frac{\xi}{n} + \eta\right] \\
&\quad= \mathbb{E}_{x\sim\hat{p}^{\mathcal{D}}}\left[\lambda f_*\left(\frac{\ell^\pi(x;w,\beta)-\eta}{\lambda}\right) + \lambda\frac{\xi}{n} + \eta\right].
\end{aligned}$$

Now, the optimal $q(x)$ is given by

$$f'_*\left(\frac{\ell^\pi(x;w,\beta)-\eta}{\lambda}\right).$$

By applying our substitution $p(x) = q(x)\hat{p}^{\mathcal{D}}(x)$, we prove the second claim.

□

We now want to leverage Theorem 4.6.10, to come up with an explicit algorithm, to calculate the lower an upper confidence bound.

Remark 4.6.11. Consider a distribution $p \in \Delta_{\hat{p}^{\mathcal{D}}}$. By (2.2) it takes the form

$$\sum_{i=1}^n p_i \mathbb{1}_{s_{0,i}=s_0,\ s_i=s,\ a_i=a,\ s'_i=s'}, \quad \text{where} \quad \vec{p} = (p_1,\dots,p_n)^\top \in \Delta^n.$$

Therefore,

$$\begin{aligned}
\mathbb{E}_{x\sim p}[\ell^\pi(x;w,\beta)] = \sum_{i=1}^n p_i\ell_i(w,\beta) = \langle\vec{p},\vec{\ell}(w,\beta)\rangle, \quad \text{where} \\
\ell_i(w,\beta) \doteq \ell^\pi(s_{0,i},s_i,a_i,r_i,s'_i;w,\beta), \quad i=1,\dots,n.
\end{aligned}$$

We can obtain $\vec{\ell}(w,\beta)$ from the dataset $\mathcal{D}$ and evaluation policy π.

■

For now, we will fix $\beta \in \mathbb{R}^K$ and $w : S\times A \to \mathbb{R}_{\ge 0}$. From the KKT-conditions, we gather that

$$D_f(p \parallel \hat{p}^{\mathcal{D}}) = \frac{1}{n}\sum_{i=1}^n f(np_i) = \frac{\xi}{n}. \tag{4.31}$$

In order to apply (4.30), we need to further specify our f-divergence. We consider the modified KL divergence (2.23). Using Lemma 2.2.9, we get

$$f'(x) = 2\left(\log x + x\frac{1}{x}\right) - 2 = 2\log x \quad \text{and} \quad f'_*(y) = (f')^{-1}(y) = e^{y/2}.$$

Plugging into (4.30) gives us

$$p_i = \exp\left(\pm\frac{\eta-\ell_i}{2\lambda}\right)\frac{1}{n} = \exp\left(\mp\frac{\ell_i}{2\lambda}\right)\Big/ n\exp\left(\pm\frac{\eta}{2\lambda}\right).$$

Now, we use the fact that $\vec{p}$ must be a distributional vector and get

$$1 = \sum_{i=1}^{n} p_i = \sum_{i=1}^{n} \exp\left(\mp \frac{\ell_i}{2\lambda}\right) \Big/ n \exp\left(\pm \frac{\eta}{2\lambda}\right).$$

Finally the optimal weights for lower and upper confidence bounds, respectively, are

$$\vec{p}_\lambda = \operatorname{softmax}\left(\mp \frac{\ell}{2\lambda}\right),$$

where we have to chose the $\lambda \geq 0$ such that $\vec{p}_\lambda$ that satisfies the KKT-condition (4.31).

5 Environments

5.1 General

Since we require our environments to have an infinite horizon, but in practice, this is rarely the case, we apply two custom `Wrapper`[1] objects from OpenAI Gym to still achieve it.

- `AbsorbingWrapper`. In this scenario, once the `terminate` flag is `True`, the environment neglects all actions. It either stays in the same absorbing state [38] or acts on "auto pilot," e.g. in `Cartpole` the pole would swing without the cart moving. In any case, some absorbing reward (usually zero) is handed out.
- `LoopingWrapper`. Instead of neglecting actions, the environment is immediately reset using its initial state distribution, applying `Env.reset()`. Assuming that the state-action-space is finite and it never enters any loops prior to termination, we achieve ergodicity. This makes looping an attractive option for the undiscounted setting.

Remark 5.1.1. For applications where we want to achieve a specific goal similar to reaching the end of a maze, e.g. successfully curing a patient, we apply absorbing with absorbing reward zero. As reward function we use

$$R(s,a,s') \doteq \begin{cases} 1, & \text{if the goal is reached in } s', \text{ but not in } s, \\ 0, & \text{else.} \end{cases}$$

Also, let H be the (possibly infinite) random variable hitting time for the time step at which the goal is first reached, i.e.,

$$H \doteq \inf\{t \in \mathbb{N} \mid \text{goal is first reached at time } t\}.$$

Considering our reward function, the policy value is computed as follows

$$\rho^{\pi,\gamma} = (1-\gamma)\mathbb{E}[\gamma^H].$$

For lower γ, we get a higher penalty for taking longer to reach the goal. The closer γ moves towards 1, the lower this penalty gets. Taking the limit and using dominated convergence, we see that the scaled policy value converges towards the success rate, i.e.,

$$\begin{aligned} \lim_{\gamma\to 1} \mathbb{E}[\gamma^H] &= \lim_{\gamma\to 1} \Big(\underbrace{\mathbb{E}[\gamma^H \mid H<\infty]}_{<1} \ \mathbb{P}(H<\infty) + \underbrace{\mathbb{E}[\gamma^H \mid H=\infty]}_{0} \ \mathbb{P}(H=\infty) \Big) \\ &= \underbrace{\mathbb{E}\left[\lim_{\gamma\to 1} \gamma^H \,\middle|\, H<\infty \right]}_{1} \ \mathbb{P}(H<\infty) = \mathbb{P}(\text{goal is reached}). \end{aligned}$$

■

[1] https://gymnasium.farama.org/api/wrappers/

5.2 Boyan Chain

In order to compare our algorithms in an environment, where we can determine the stationary distribution correction $w_{\pi/\mathcal{D}}$ and policy value ρ^π analytically, we use `BoyanChain` [11, 41]. The initial state distribution d_0^π, transition matrix $\mathcal{P}^\pi$, and reward function r are explicitly stated in Figure 5.1. We choose $d^\mathcal{D}$ to be uniform on $S \times A$. With this information, we can solve the modified backwards Bellman equations (3.17) explicitly.

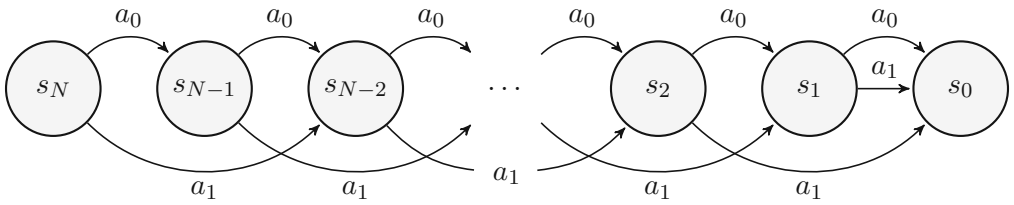


Figure 5.1: We follow the setup by Yao and Liu [11]. The initial distribution is uniform over all states. For all $i \geq 2$, a_0 transitions from s_i to s_{i-1} and a_1 from s_i to s_{i-2}, Both cases yielding a reward of -3. For $\gamma < 1$, at s_1 both a_0 and a_1 lead towards s_0 and we get a reward of -2. We consider s_0 an absorbing state and let both actions steer from s_0 back to s_0, with a reward of 0. For $\gamma = 1$, we want to ensure ergodicity, hence, at s_1 and s_0, both actions reset the environment using the initial state distribution, with rewards -2 and 0, respectively. This boils down to using absorbing and looping.

Considering the reward function, we notice that the goal of this environment is to reach s_0 as quickly as possible. Hence, the optimal policy always chooses a_1 over a_0. However, our policy π has

$$\pi(a_0 \mid s_i) = 0.1 \quad \text{and} \quad \pi(a_1 \mid s_i) = 0.9 \quad \text{for all} \quad i = 0, \ldots, N.$$

We generate a dataset of $n = 100{,}000$ samples for our numerical results and set $N = 12$ following Boyan et al. [11]. Since this environment is tabular, we use one hot encoding, to embed S into $[0, 1]^{N+1}$.

5.3 OpenAI Gym

To further investigate the performance of our estimators, we test them on the OpenAI Gym environments `FrozenLake`[2], `Taxi`[3], and `Cartpole`[4]. Following Dai et al. [13], we use looping for the first two instances and apply an absorbing state with a reward of -1 for the latter.

For `FrozenLake`, we have a *deterministic* and *stochastic* version, depending whether the parameter `is_slippery` is `False` or `True`, respectively. In both cases, we generate $n = 100{,}000$ dataset samples using a uniform dataset distribution $d^\mathcal{D}$ on $S \times A$. The initial state distribution d_0^π, transition matrix $\mathcal{P}^\pi$, and reward function r are obtained analytically. Just like in `BoyanChain`, we can then solve the modified Bellman equations (3.17) explicitly. The evaluation policies are trained with `PPO`[5] [37, 35], without looping in the environment.

In `Taxi`, we use the same environment, behavior policy, and evaluation policy as Dai et al. [13]. We gather 10,000 trajectories with a length of 200.

[2] https://gymnasium.farama.org/environments/toy_text/frozen_lake/
[3] https://gymnasium.farama.org/environments/toy_text/taxi/
[4] https://gymnasium.farama.org/environments/classic_control/cart_pole/
[5] https://stable-baselines3.readthedocs.io/en/master/modules/ppo.html

With `Cartpole`, however, we train a behavior and evaluation policy for 10,000 and 100,000 steps, using `PPO`, also without absorbing in the environment. Here, we gather 500 trajectories with a length of 200. In all of these time steps, the behavior and evaluation policies manage to balance the pole roughly 70% and 100% of the time, respectively.

5.4 Medical

We use a set of prerecorded data of septic patients and their treatment from `AmsterdamUMCdb` [39, 8]. Each state is composed of certain sensory data, such as blood pressure, blood oxygen saturation, etc. Actions are clustered administered dosage of the drug *hydrocortisone*. Once a patient's treatment ends successfully, a reward of 1 is handed out, otherwise, we only get 0. The evaluation policy is obtained in the same way as Bologheanu et al. [8].

In order for the policy value to be easily interpretable, the setup from Remark 5.1.1 serves as a basis for our environment. When a patient is cured or passed away in the state s_H, all subsequent states are chosen to be the same absorbing state, i.e., $s_t = s_H$ for all $t \geq H$. In order to implement this setup, we pad each ending of a trajectory with the respective finite state.

The evaluation was performed using this data directly via `NeuralDualDice`, `NeuralGenDice`, and `NeuralGradientDice`, as well as in a clustered form, as described in Figure 5.2. Clustering the dataset lets us construct a simulator and use algorithms with a more solid convergence theory.

[6] https://scikit-learn.org/1.5/modules/generated/sklearn.cluster.KMeans.html
[7] https://gymnasium.farama.org/api/env/
[8] https://sb3-contrib.readthedocs.io/en/master/modules/ppo_mask.html

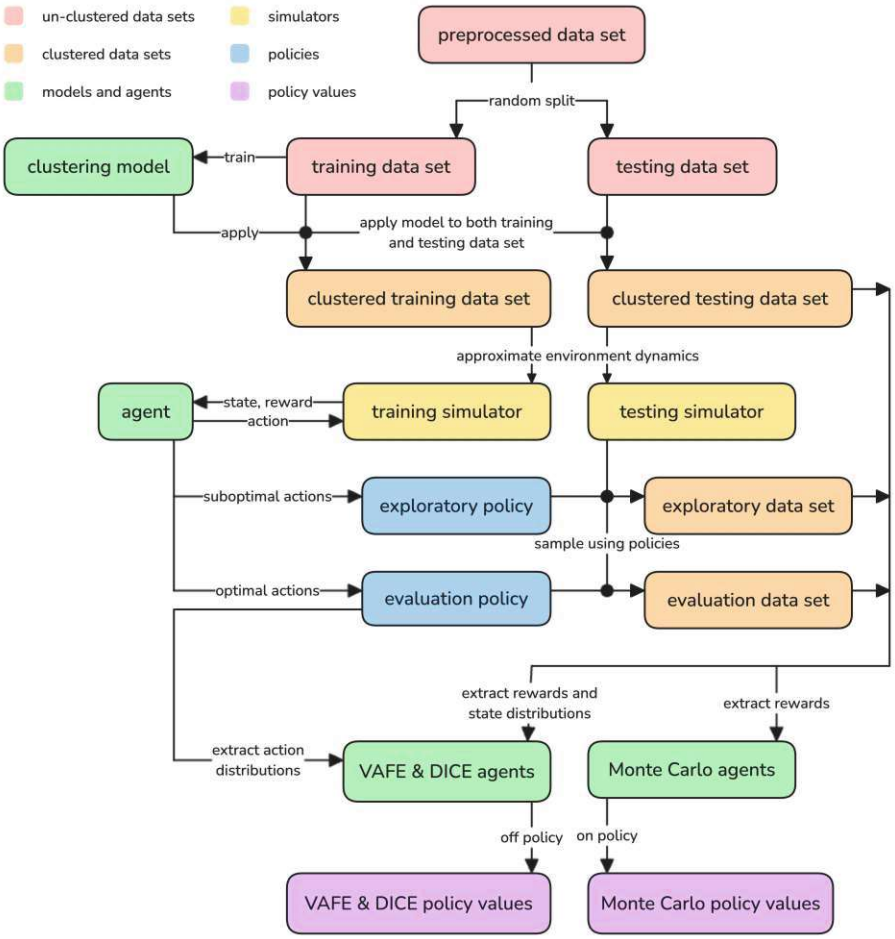


Figure 5.2: We start by splitting the *preprocessed data set* into a *training* and *testing data set*. We select some features from the un-clustered observations in the training data set and train a `KMeans`[6] model from Scikit-learn with 256 clusters, which we use to extract clustered observations from the training and testing data set. Two additional clusters, representing "recovered" and "deceased," are added at the end of each trajectory. The clustered observations are stored in the *clustered training* and *testing data set*, together with the patient IDs, time stamps, actions, and rewards. From these, we construct a *training* and *testing simulator* in the format of an `Env`[7] from OpenAI Gym, by approximating the initial state distribution and transition kernel, respectively. We make sure for each environment to allow an action a in a state s only if (s, a) is part of the associated clustered dataset. On the training simulator we apply `MaskablePPO`[8] [17] from Stable-Baselines3 to train an *exploratory* and *evaluation policy*. We apply these policies to the testing simulator, to produce a *exploratory* and *evaluation data set*. Together with the clustered testing data set, *Monte Carlo agents* use their rewards to approximate their respective policy values via on policy evaluation. Additionally, we use the *VAFE* and *DICE agents* `TabularVafe`, `TabularDice`, `TabularDualDice`, and `TabularGradientDice`, which take the action distributions from the evaluation policy as well as the rewards and state distribution from these datasets to produce an off policy estimate of the evaluation policy value.

6 Numerical Results

We conducted experiments using both tabular and continuous algorithms with various hyperparameters, selecting those that yielded the best performance. We briefly summarize their functionality.

The most important hyperparameter is the discount factor γ, which lets us control how much weight is given to rewards further along a trajectory.

The objectives (4.4) and (4.5) also include the norm penalty coefficient λ. It determines, how high the approximate stationary distribution correction should be penalized, when its expected value under the dataset deviates from one.

The boolean hyperparameter `weighted` lets us choose between a simple (3.15) and weighted Monte Carlo estimator (3.16) for the policy value, provided, we are already given an approximation for the stationary distribution correction.

Tabular algorithms also include `projected`. If active, we first project all vectors and matrices onto the subspace of all indices whose corresponding state-action-pairs actually occur in the dataset, before approximation. Afterwards, we embed back into the original space. In some cases, it is necessary to enforce this assumption in order to satisfy the requirements set forth in Assumption 3.5.1. If already satisfied, it does not make a difference for the estimator.

Finally `TabularDice` has the boolean hyperparameter `modified`, which dictates whether to use the standard backwards (3.9) or modified backwards Bellman equations (3.17).

6.1 Boyan Chain

`BoyanChain Tabular`. This environment was used to investigate the performance of our tabular algorithms, as we increase the number of states. In the episodic variant Subfigures 6.1a and 6.1b, the policy value drops as the chain length N increases. Recall that as the end of the chain is reached, reward zero, instead of -3 and -2, is handed out. Similar behavior can be found in the continuing variant Subfigure 6.1c. The Subfigures 6.1d, 6.1e, and 6.1f, show the policy value error in more detail. We see that our tabular algorithms perform similarly well, if not better than on-policy evaluation. In the episodic setting, the error increases as the number of states goes up. This is not the case with regard to the continuing domain. The underlying cause of this behavior is likely an increase and subsequent decrease in the discrepancy of the rewards, respectively. However, an examination of the stationary distribution corrections, as illustrated in Subfigures 6.1g, 6.1h, and 6.1i, reveals an increase in the MSE as the number of iterations N increases. This means that this experiment was successful and the analytical value was approximated properly.

`BoyanChain Continuous`. Here we provide an environment with a continuous state space and access to an analytical solution to the stationary distribution correction $w_{\pi/\mathcal{D}}$ and policy value ρ^{π}. The Subfigures 6.2a, 6.2b, 6.2c, and 6.2e, 6.2f, 6.2g show that the analytical policy value can be approximated decreasingly well as the discount factor increases. This can be observed in more detail in Subfigures 6.2i, 6.2j, and 6.2k, respectively. Also, the weighted estimator (3.16) performs a more accurate approximation than the simple estimator (3.15), especially for

`NeuralGenDice`. As suggested by Nachum et al. [34] and Zhang et al [41], the performance of `NeuralDualDice`, approximating the stationary distribution correction, is significantly impaired when the discount factor is increased. `NeuralGenDice` and `NeuralGradientDice` also follow this patter, but are more robust in this regard. Mao et al. [25] claim that a small Bellman residual angle of the gradients $\nabla_\vartheta \mathcal{P}^\pi v_\vartheta$ and $\nabla_\vartheta v_\vartheta$ may be responsible for the instability associated with a higher discount factor, as the gradients cancel each other out more and more. In Figure 8.21 we see that the BRA is indeed quite small, at about $\frac{\pi}{8}$. Even so, in Subfigures 6.2d, 6.2h, and 6.2l, it is evident that the undiscounted case, with discount factor $\gamma = 1$, offers an excellent approximation performance, albeit it based in a slightly different environment. We have the same setup as Zhang et al. [41]. However, our results for `NeuralDualDice` and `NeuralGradientDice` show an even smaller MSE on the stationary distribution correction. We also get lower MSEs for the continuing setting. Our analytical solution for the stationary distribution correction was calculated by solving the eigenvalue problem (4.1) and not by iteratively applying the transition matrix to the identity matrix 10,000 times [41]. This may explain the better performance.

6.2 OpenAI Gym

`FrozenLake`. This environment lets us observe the influence of adding randomness to the transition dynamics. Comparing Subfigures 6.3a and 6.3b, we notice that approximating the policy value $\rho^\pi(\gamma)$ gets harder, not only as the discount factor γ increases, but also when we switch from deterministic to stochastic transitions. For those algorithms that allow for an undiscounted evaluation, the undiscounted counterparts converge to their evaluation outcome, with increasing discount factor. This is not the case for episodic on-policy evaluation. The rationale behind this phenomenon can be elucidated with relative ease. We are dealing with a looped environment, but can only gather finite trajectories. With a higher discount factor, rewards further along the episode are given a higher weighting. Not sampling them is equivalent to setting them to zero, which biases the estimator.

`Taxi`. Here we provide a deterministic environment similar to `FrozenLake`, in terms of its objective, but has a significantly higher state space. As a consequence, gathering experience by means of a behavior policy is harder, since we must satisfy Assumption 3.5.1. Numerical evidence for this claim can be found when comparing Figures 8.18 and 8.19. There, we see that the approximations of the initial state distribution $\hat{d}_0^\pi$ and the transition matrix $\hat{\mathcal{P}}^\pi$ are already flawed. Further evidence is given in Subfigure 6.3c, which shows that approximating the policy value does not work as well as for `FrozenLake`. Nevertheless, the approximations are still reasonable, especially for a higher discount factor. This means that our algorithms are to a certain extent robust against flaws in the dataset.

`Cartpole`. This classical environment serves as a more sophisticated continuous state space testing environment than `BoyanChain Continuous`. Although an analytical solution is lacking, we can still undertake a comparison with on-policy evaluation alone. The performance of `NeuralDualDice` in Subfigures 6.3g, 6.3h, and 6.3i, is comparable to Nachum et al. [34]. Except for `NeuralGenDice` with the discount factor $\gamma = 0.1$, the weighted estimators in Subfigures 6.3d, 6.3e, and 6.3f, perform better than their simple counterparts in Subfigures 6.3d, 6.3e, and 6.3f, respectively. This does not agree with the comparison in `BoyanChain Continous`, which shows that there is "no free lunch." `NeuralGenDice` is as accurate as `NeuralDualDice` for $\gamma = 0.1, 0.5$, but for $\gamma = 0.9$ it underestimates the policy value. For `NeuralGenDice`, this

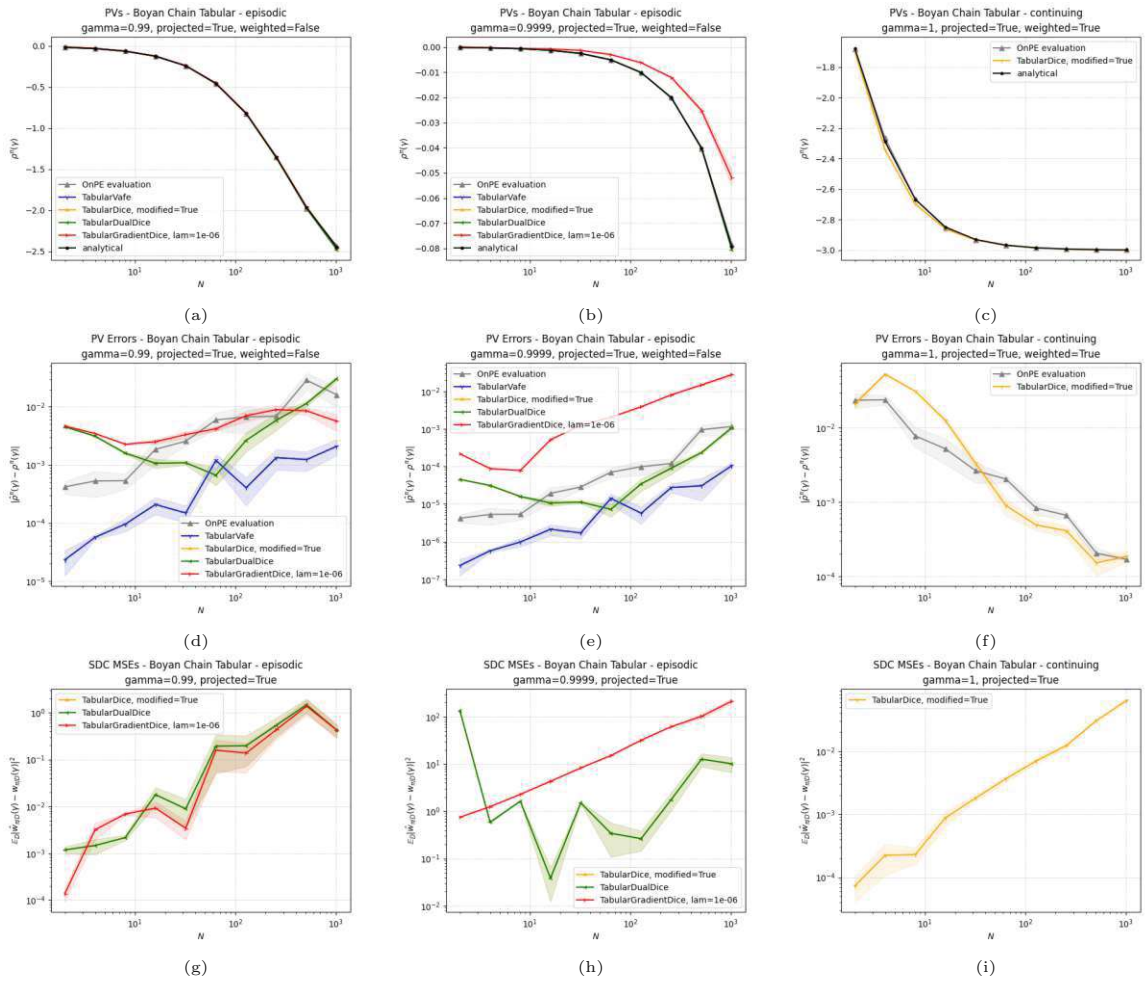
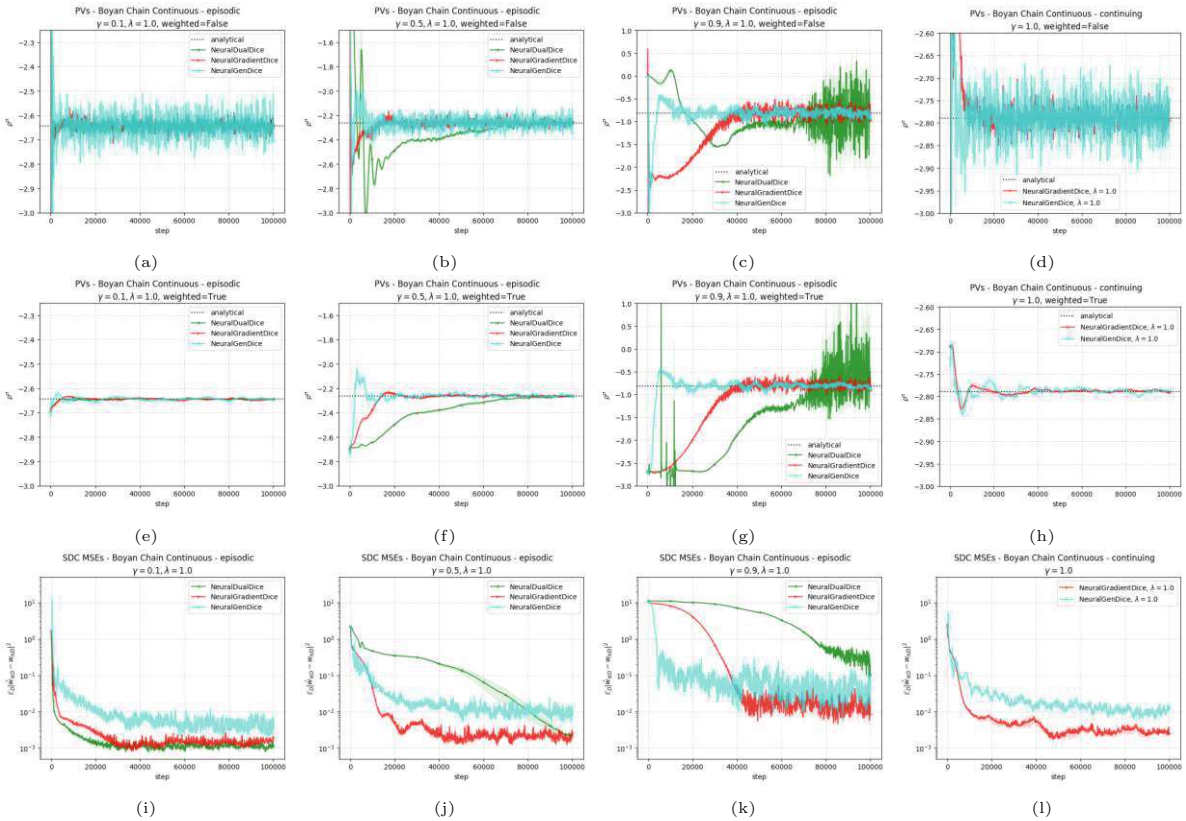


Figure 6.1: `BoyanChain Tabular`. The horizontal axis shows the length of the chain, i.e., $N+1$ is the number of states. The columns of the multi-plot, 6.1a, 6.1d, 6.1g and 6.1b, 6.1e, 6.1h and 6.1c, 6.1f, 6.1i, represent the same runs, displaying the (approximate) policy value ρ^π, policy value error $|\hat{\rho}^\pi - \rho^\pi|$ and stationary distribution correction MSE $\mathbb{E}_\mathcal{D}|w_{\pi/\mathcal{D}} - \hat{w}_{\pi/\mathcal{D}}|^2$. We use various discount factors γ and the same norm penalization coefficients $\lambda = 10^{-6}$. We plot the sample-mean and an area spanning half the standard deviation using runs on four datasets, generated by different seeds.

is already the case for the lower discount factors. Taking a closer look at the loss functions, plotted in Figures 8.10, 8.11, and 8.12, we see that they move increasingly closer to zero, the higher the discount factor gets. Presumably, the Bellman error $\gamma\mathcal{P}^\pi v - v$ in the loss definition is responsible. Figure 8.22 supports this claim, by showing a small Bellman residual angle below $\frac{\pi}{16}$. As already discussed in `BoyanChain Continuous`, this leads to many parts of the gradients $\nabla_\vartheta \mathcal{P}^\pi v_\vartheta$ and $\nabla_\vartheta v_\vartheta$ canceling each other out. Since the BRA in `Cartpole` is half of the BRA in `BoyanChain Continuous`, this may explain why the approximation in this application is even worse for the high discount factor $\gamma = 0.9$.



(a) (b) (c) (d)

(e) (f) (g) (h)

(i) (j) (k) (l)

Figure 6.2: `BoyanChain Continuous`. The horizontal axis shows the training step. The columns, 6.2a, 6.2e, 6.2i, 6.2b, 6.2f, 6.2j, 6.2c, 6.2g, 6.2k, 6.2d, 6.2h, 6.2l, represent the same runs, displaying the analytical and approximate policy value ρ^π, using the simple (3.15) and weighted (3.16) estimator, and the MSE $\mathbb{E}_\mathcal{D}|\hat{w}_{\pi/\mathcal{D}} - w_{\pi/\mathcal{D}}|^2$ to the analytical stationary distribution correction $w_{\pi/\mathcal{D}}$. We use various discount factors γ and norm penalization coefficients λ. We plot the sample-mean and an area spanning half the standard deviation using runs on four datasets, generated by different seeds.

6.3 Medical Application

`Medical Tabular`. In order to be able to perform on-policy evaluation, we cluster our medical dataset, extract the necessary distributions and build a simulator, as described in Figure 5.2. Similar to Nachum et al. [34], we want to compare policy evaluation algorithms, using various datasets. As illustrated in Figure 6.4, the clinician, exploratory, and evaluation policy each demonstrate a distinct level of policy value, with each outperforming the others. Figure 6.6 depicts the convergence of policy values towards respective treatment success rates across datasets. Since the goal of our algorithms is to approximate the evaluation policy value, we notice that resampling the dataset by means of the exploratory policy 6.4b yields the best approximation. The clinician dataset 6.4a is likely too different from the trajectories that the evaluation policy would follow. The dataset resampled with the evaluation policy 6.4c, on the other hand, does not sample enough experience for Assumption 3.5.1 to be satisfied. Similarly to `Taxi`, this can be supported by noticing the already flawed approximations of the initial state distribution $\hat{d}_0^\pi$ and transition matrix $\hat{\mathcal{P}}^\pi$, as illustrated in Figure 8.20. For each algorithm and

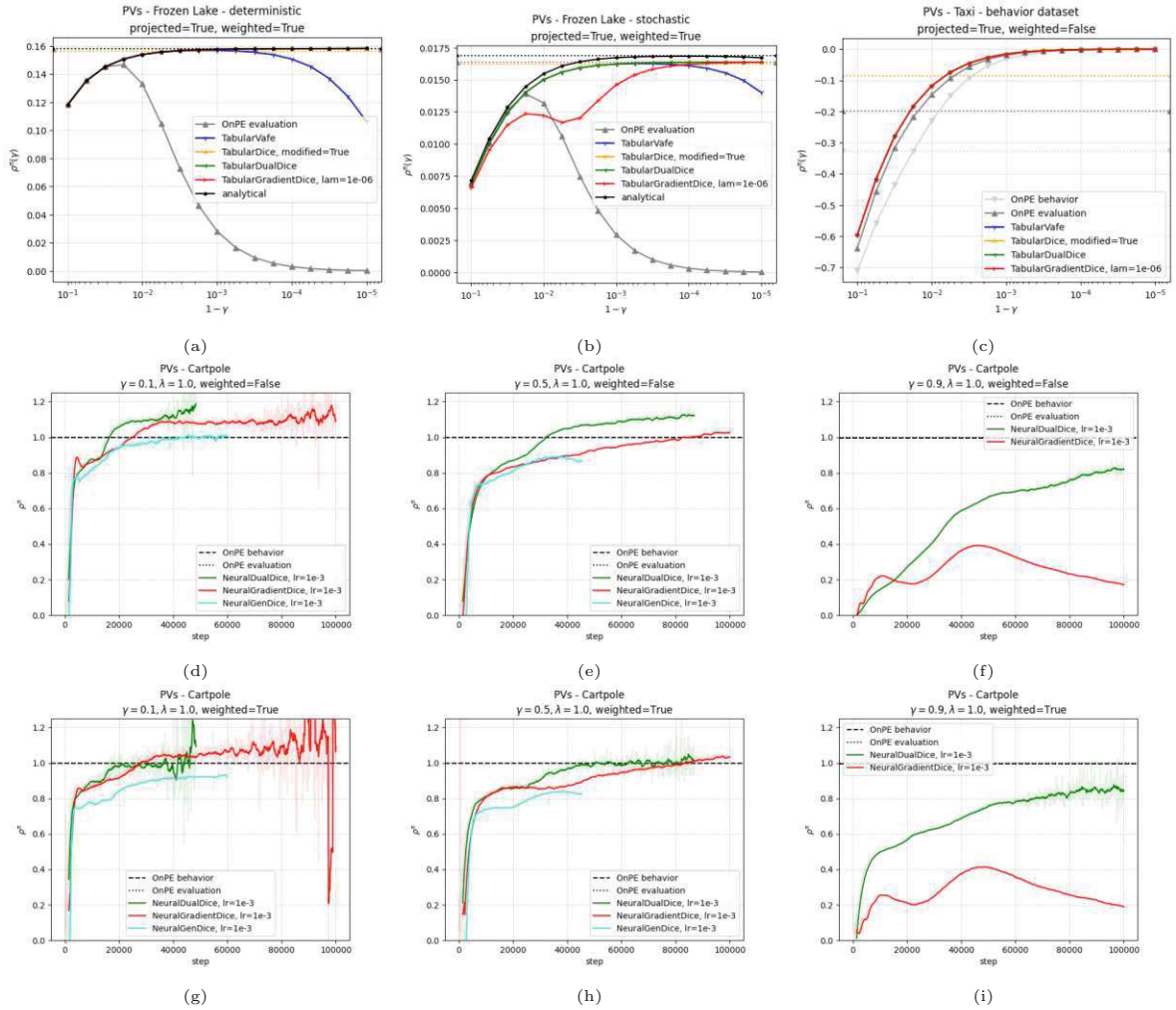
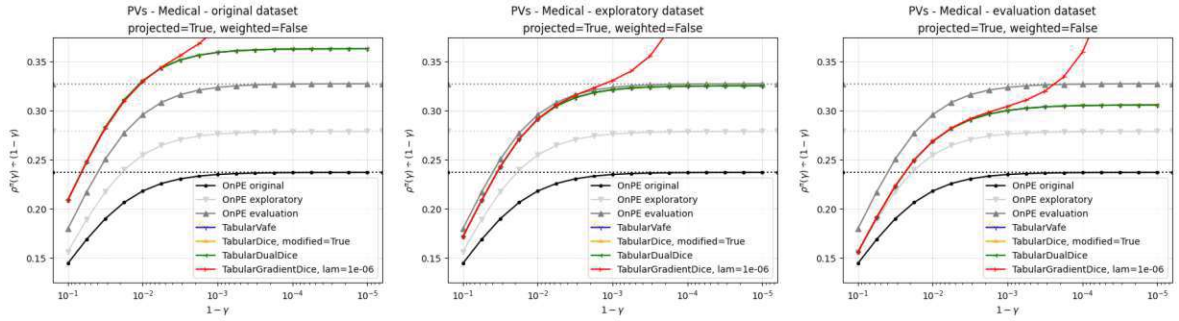


Figure 6.3: `OpenAI Gym`. The top three plots, 6.3a, 6.3b, and 6.3c, show the policy value $\rho^{\pi}(\gamma)$ against the discount factor γ close to one, for `FrozenLake` and `Taxi`. The dotted lines mark the respective undiscounted results for the policy value, i.e., $\gamma = 1$. The bottom three plots, 6.3g, 6.3h, and 6.3i, present the policy value ρ^{π} with discount factor $\gamma = 0.1$, 0.5, and 0.9, respectively, plotted against the training step of the respective algorithm, for `Cartpole`. Only the algorithms that yield sufficient results were chosen. In all of the plots, we present the behavior- and evaluation policy values as a reference point.

the onpolicy evaluation, consider the right most point on their respective curve, i.e., where the discount factor γ is the highest. The differences between the points of an algorithm and on-policy evaluation, are less than 5%, 1%, and 2.5%, respectively. These serve as error margins on the evaluation policy treatment success rate, respectively. Since the difference of the treatment success rates of the clinician and evaluation policy is more than 7.5%, this is already enough to confirm superhuman performance.

`Medical Continuous`. The primary motivation for this work is to evaluate a policy, treating septically ill patients, using offline behavior agnostic policy evaluation algorithms. This means that we operate directly on a dataset, without clustering or inferring the distribution

of the behavior policy. In Figure 6.5, we illustrate the best results available. For all numbers of hidden neurons, `NeuralDualDice` and `NeuralGradientDice` consistently produce a policy value above the clinician's behavior, even when accounting for the standard deviation. Here, we chose the `simple` policy value estimate. In Figures 8.14 and 8.15, we see that the `weighted` estimator is biased upwards [38, p. 105]. On the other hand, `NeuralGenDice` produces unstable learning curves. This is probably due to a high variance of the stationary distribution correction approximate $\hat{w}_{\pi/\mathcal{D}}$, resulting in some of its components being much higher than others. Consequently, in Figure 8.16, the `simple` policy estimate skyrockets or oscillates immensely, while the `weighted` estimator nullifies this effect. The choice of the low discount factor $\gamma = 0.9$ significantly decouples the scaled policy value from the treatment success rate, as shown in the Figure 6.6. However, the policy gets a higher penalty for taking too many steps per episode. Also, the estimators are more stable and less biased, while a discount factor of $\gamma = 0.99$ already leads to implausible results. Our findings in Figure 6.5 are consistent with Bologheanu et al. [8], who used the same method to obtain the evaluation policy. As already discussed for `BoyanChain Continuous` and `Cartpole`, the Bellman residual angle can give an insight into the performance of the estimator. Just like `BoyanChain Continuous` in Figure 8.21, `NeuralGradientDice` in Figure 8.15 shows a BRA of just under $\frac{\pi}{8}$. On the other hand, `NeuralDualDice` in Figure 8.14 and `NeuralGenDice` in Figure 8.16 have a BRA slightly below $\frac{\pi}{16}$, similar to `Cartpole` in Figure 8.22. In the same way that the results for `BoyanChain Continuous` are superior to those of `Cartpole`, `NeuralGradientDice` produces more stable policy value and loss curves than `NeuralDualDice` and `NeuralGenDice`.



(a) (b) (c)

Figure 6.4: `Medical Tabular`. These three plots show the scaled policy value $\rho^\pi(\gamma)/(1-\gamma)$, plotted against the discount factor γ close to one. The dotted lines mark the respective treatment success rate for the clinicians, exploratory and evaluation policy. The data for the VAFE and DICE algorithms is either taken from the clustered test dataset directly 6.4a or resampled using an exploratory policy 6.4b or the evaluation policy 6.4c.

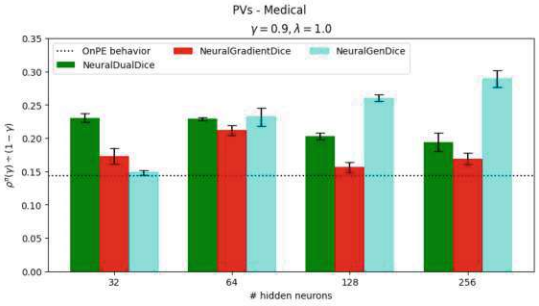


Figure 6.5: `Medical Continuous`. Here, we show the mean and standard deviation of the scaled policy value with a discount factor of $\gamma = 0.9$. The objectives use the same norm penalization coefficient $\lambda = 1.0$, as suggested by Zhang et al. [44, 41]. The primal and dual neural networks have a single hidden layer, with different numbers of neurons, specified on the horizontal axis. The samples were taken from the marked parts of the learning curves in Figures 8.14, 8.15 and 8.16, where the policy value estimate and the loss function settle in an equilibrium and oscillates with a consistent amplitude. For each algorithm, we chose either the simple or the weighted estimator, depending on which produced the more plausible output, respectively. As a reference, we also show the clinician's scaled behavior policy value on the dotted horizontal line.

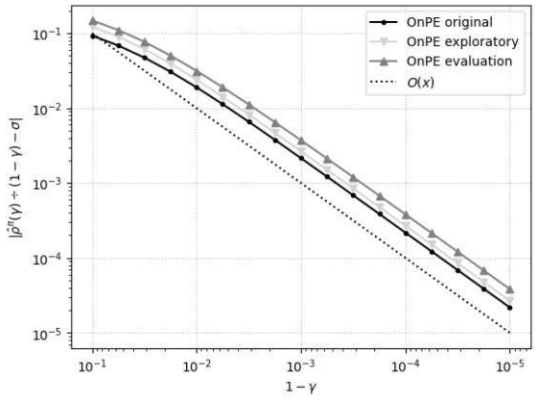


Figure 6.6: We plot the error between the scaled on-policy value $\hat{\rho}^\pi(\gamma)/(1-\gamma)$ and treatment success rate σ, for the clinician as well as an exploratory- and the evaluation policy. As we increase the discount factor γ towards one, the scaled policy value converges linearly against the success rate.

7 Conclusions and Future Work

In this work, we described methods that approximate the policy value for tabular environments, based directly on the Bellman equations (3.5) and (3.9), solving linear equation systems and eigenvalue problems and collected the most prominent DICE methods for continuous environments.

We tested these algorithms on various well established environments retaining certain selected properties, analyzed the results and made connections to the underlying theory. Finally, we executed the algorithms on a carefully constructed environment for medical applications, where the policy values are easily interpretable.

We saw that our estimators yield good results, as long as we provide adequate data. For the clustered medical application, our estimators achieved errors on the treatment success rate ranging from 1% to 5%. This is especially interesting, since practical and theoretical evidence confirms that classical off-policy evaluation methods based on importance sampling suffer from high variance [42]. They also explicitly require the distribution of the clinician's behavior policy, which can only be approximated at best [19, 8].

Theoretical guarantees that the estimator will work in the undiscounted setting are difficult to provide, when dealing with complex continuous environments. However, there are even some issues relating to stability and bias, if one chooses to evaluate in the discounted setting with a high discount factor. It is important to further develop stable and precise algorithms for continuous environments. A clustered environment simulator always deviates from the original. Possible solutions might adapt ideas from Mao et al. [25] from policy optimization to policy evaluation. An alternative to DICE by Mousavi et al. [28] involves the use of reproducing kernel Hilbert spaces and maximum mean discrepancy. Building upon their approach is to find kernels that provide accurate estimates.

Also, the need for safe policy evaluation for medical applications calls for algorithms capable of providing rigorous confidence intervals on the policy value. Algorithms like these, which also only have the requirements of those discussed in this work, i.e., offline and behavior agnostic, should be developed further and tested in the same way that we have done here [13].

It is still an open task, to run these offline behavior agnostic policy evaluation algorithms on more and bigger dataset, including different clustering techniques and feature selections, perhaps also with other treatment objectives. Nevertheless, the medical policy evaluation results are very promising so far, showing that there is a lot of potential for RL and medicine to work together.

Acknowlegement

This work was supported by the project "RELY – Reliable Reinforcement Learning for Sustainable Energy Systems," funded by the Austrian Research Funding Agency (FFG) under the grant number #FO0999899921.

Bibliography

[1] Josh Achiam et al. "GPT-4 Technical Report". In: *arXiv:2303.08774* (2023).

[2] Daniel Alpay. "An Advanced Complex Analysis Problem Book". In: *Topological Vector Spaces, Functional Analysis, and Hilbert Spaces of Analytic Functions. Birkäuser Basel* (2015).

[3] OpenAI: Marcin Andrychowicz et al. "Learning Dexterous In-Hand Manipulation". In: *The International Journal of Robotics Research* 39.1 (2020), pp. 3–20.

[4] Francis Bach. "Breaking the Curse of Dimensionality with Convex Neural Networks". In: *Journal of Machine Learning Research* 18.19 (2017), pp. 1–53.

[5] Francesca Bartolucci et al. "Understanding Neural Networks with Reproducing Lernel Banach Spaces". In: *Applied and Computational Harmonic Analysis* 62 (2023), pp. 194–236.

[6] Alain Berlinet and Christine Thomas-Agnan. *Reproducing Kernel Hilbert Spaces in Probability and Statistics*. Springer Science & Business Media, 2011.

[7] Markus Böck et al. "Superhuman Performance on Sepsis MIMIC-III Data by Distributional Reinforcement Learning". In: *PLoS One* 17.11 (2022), e0275358.

[8] Razvan Bologheanu et al. "Development of a Reinforcement Learning Algorithm to Optimize Corticosteroid Therapy in Critically Ill Patients with Sepsis". In: *Journal of Clinical Medicine* 12.4 (2023), p. 1513.

[9] Vivek S Borkar and Sean P Meyn. "The ODE Method For Convergence of Stochastic Approximation and Reinforcement Learning". In: *SIAM Journal on Control and Optimization* 38.2 (2000), pp. 447–469.

[10] Jonathan Borwein and Adrian Lewis. *Convex Analysis*. Springer, 2006.

[11] Justin A Boyan. "Least-Squares Temporal Difference Learning". In: *ICML*. 1999, pp. 49–56.

[12] Stephen Boyd and Lieven Vandenberghe. *Convex Optimization*. Cambridge University Press, 2004.

[13] Bo Dai et al. "Coindice: Off-policy Confidence Interval Estimation". In: *Advances in Neural Information Processing Systems* 33 (2020), pp. 9398–9411.

[14] John C Duchi, Peter W Glynn, and Hongseok Namkoong. "Statistics of Robust Optimization: A Generalized Empirical Likelihood Approach". In: *Mathematics of Operations Research* 46.3 (2021), pp. 946–969.

[15] Jianfeng Gao, Michel Galley, and Lihong Li. "Neural Approaches to Conversational AI". In: *The 41st International ACM SIGIR Conference on Research & Development in Information Retrieval*. 2018, pp. 1371–1374.

[16] Helmut Horvath. "Deep Reinforcement Learning with Applications to Autonomous Driving". MSc thesis. Technische Universität Wien, 2024.

[17] Shengyi Huang and Santiago Ontañón. "A Closer Look at Invalid Action Masking in Policy Gradient Algorithms". In: *arXiv preprint arXiv:2006.14171* (2020).

[18] Tobias Kietreiber. "Kombination von Maximum Entropy Reinforcement Learning mit Distributional Q-Value Approximation". MSc thesis. Technische Universität Wien, 2024.

[19] Matthieu Komorowski et al. "The Artificial Intelligence Clinician Learns Optimal Treatment Strategies for Sepsis in Intensive Care". In: *Nature Medicine* 24.11 (2018), pp. 1716–1720.

[20] Jongmin Lee et al. "COptiDICE: Offline Constrained Reinforcement Learning via Stationary Distribution Correction Estimation". In: *arXiv preprint arXiv:2204.08957* (2022).

[21] Jongmin Lee et al. "OptiDice: Offline Policy Optimization via Stationary Distribution Correction Estimation". In: *International Conference on Machine Learning*. PMLR. 2021, pp. 6120–6130.

[22] Jiwei Li et al. "Deep Reinforcement Learning for Dialogue Generation". In: *arXiv preprint arXiv:1606.01541* (2016).

[23] Lihong Li et al. "Unbiased Offline Evaluation of Contextual-Bandit-Based News Article Recommendation Algorithms". In: *Proceedings of the Fourth ACM International Conference on Web Search and Data Mining*. 2011, pp. 297–306.

[24] Travis Mandel et al. "Offline Policy Evaluation Across Representations with Applications to Educational Games." In: *AAMAS*. Vol. 1077. 2014.

[25] Liyuan Mao et al. "Odice: Revealing the Mystery of Distribution Correction Estimation via Orthogonal-Gradient Update". In: *arXiv preprint arXiv:2402.00348* (2024).

[26] Sean P Meyn and Richard L Tweedie. *Markov Chains and Stochastic Stability*. Springer Science & Business Media, 2012.

[27] Volodymyr Mnih. "Playing Atari with Deep Reinforcement Learning". In: *arXiv preprint arXiv:1312.5602* (2013).

[28] Ali Mousavi et al. "Black-Box Off-Policy Estimation for Infinite-Horizon Reinforcement Learning". In: *arXiv preprint arXiv:2003.11126* (2020).

[29] Krikamol Muandet et al. "Kernel Mean Embedding of Distributions: A Review and Beyond". In: *Foundations and Trends in Machine Learning* 10.1-2 (2017), pp. 1–141.

[30] Susan A Murphy et al. "Marginal Mean Models for Dynamic Regimes". In: *Journal of the American Statistical Association* 96.456 (2001), pp. 1410–1423.

[31] Ofir Nachum. *Reinforcement Learning Via Convex Duality*. `https://www.bilibili.com/video/BV1Q54y1r7qP/`. Accessed: 20.6.2024. 2020.

[32] Ofir Nachum and Bo Dai. "Reinforcement Learning via Fenchel-Rockafellar Duality". In: *arXiv preprint arXiv:2001.01866* (2020).

[33] Ofir Nachum et al. "Algaedice: Policy Gradient from Arbitrary Experience". In: *arXiv preprint arXiv:1912.02074* (2019).

[34] Ofir Nachum et al. "Dualdice: Behavior-Agnostic Estimation of Discounted Stationary Distribution Corrections". In: *Advances in Neural Information Processing Systems* 32 (2019).

[35] Antonin Raffin et al. "Stable-Baselines3: Reliable Reinforcement Learning Implementations". In: *Journal of Machine Learning Research* 22.268 (2021), pp. 1–8.

[36] Erica Salvato et al. “Crossing the Reality Gap: A Survey on Sim-To-Real Transferability of Robot Controllers in Reinforcement Learning”. In: *IEEE Access* 9 (2021), pp. 153171–153187.

[37] John Schulman et al. “Proximal Policy Optimization Algorithms”. In: *arXiv preprint arXiv:1707.06347* (2017).

[38] Richard S Sutton, Andrew G Barto, et al. *Introduction to Reinforcement Learning. Vol. 135*. MIT press Cambridge, 1998.

[39] Patrick J Thoral et al. “Sharing ICU Patient Data Responsibly under the Society of Critical Care Medicine/European Society of Intensive Care Medicine Joint Data Science Collaboration: The Amsterdam University Medical Centers Database (AmsterdamUMCdb) Example”. In: *Critical Care Medicine* 49.6 (2021), e563–e577.

[40] Jon Wellner et al. *Weak Convergence and Empirical Processes: With Applications to Statistics*. Springer Science & Business Media, 2013.

[41] Shimon Whiteson. “GradientDICE: Rethinking Generalized Offline Estimation of Stationary Values”. In: (2020).

[42] Tengyang Xie, Yifei Ma, and Yu-Xiang Wang. “Towards Optimal Off-Policy Evaluation for Reinforcement Learning with Marginalized Importance Sampling”. In: *Advances in Neural Information Processing Systems* 32 (2019).

[43] Mengjiao Yang et al. “Off-policy Evaluation via the Regularized Lagrangian”. In: *Advances in Neural Information Processing Systems* 33 (2020), pp. 6551–6561.

[44] Ruiyi Zhang et al. “Gendice: Generalized Offline Estimation of Stationary Values”. In: *arXiv preprint arXiv:2002.09072* (2020).

8 Appendix

8.1 Additional Numerical Results

This section presents supplementary graphs and analyses that, while not central to the main contributions of this work, provide valuable insights into the behavior of our algorithms. These results help to contextualize and explain some of the phenomena observed in our main findings. Specifically, we include plots that illustrate the effects of different learning rates and key statistics of our algorithms. These additional visualizations support claims about why performance varies under certain conditions and offer a deeper understanding of how our methods respond to different parameter settings. Collectively, this evidence underscores the robustness and adaptability of our approaches across a range of scenarios.

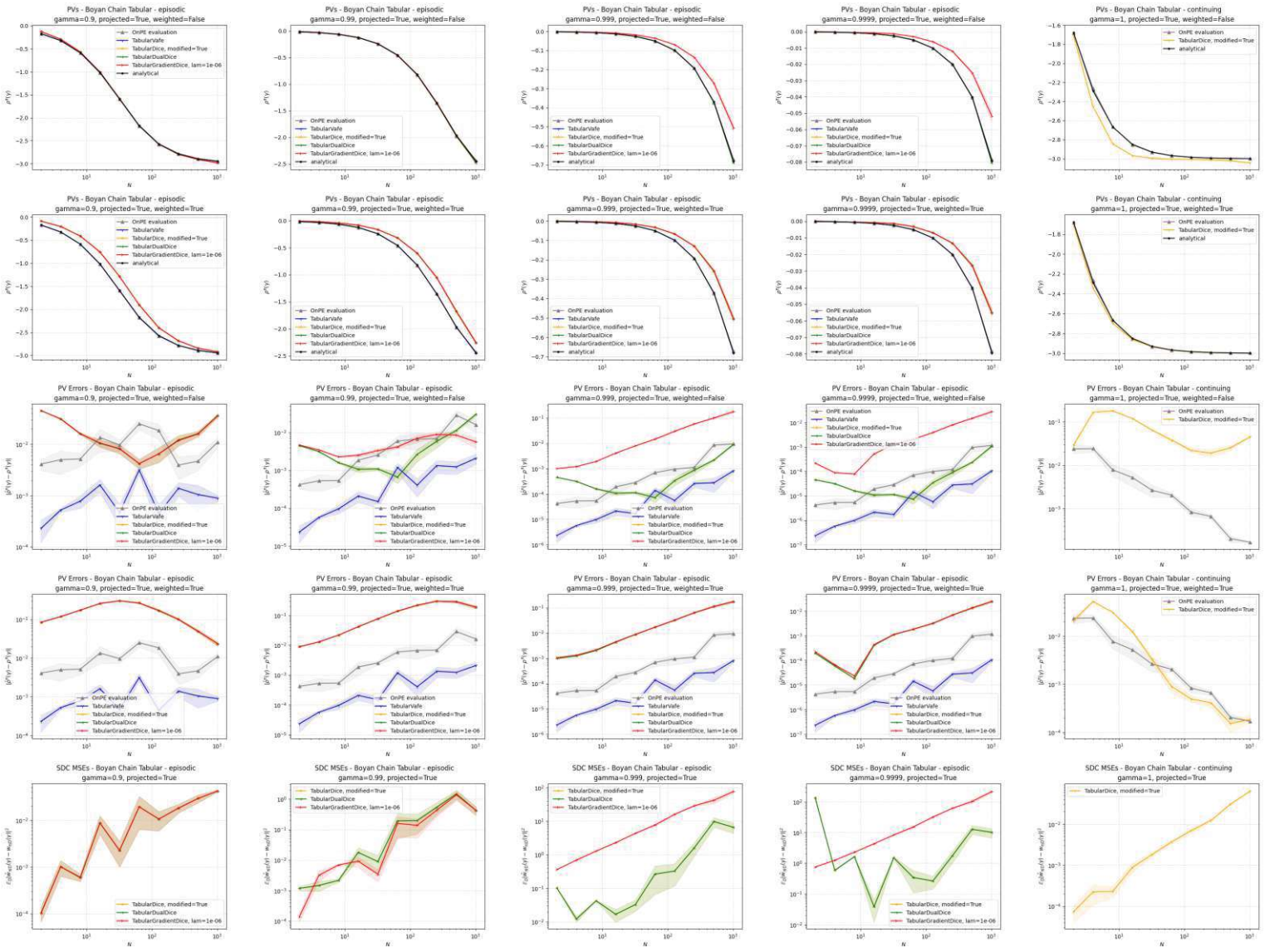


Figure 8.1: `BoyanChain Tabular`

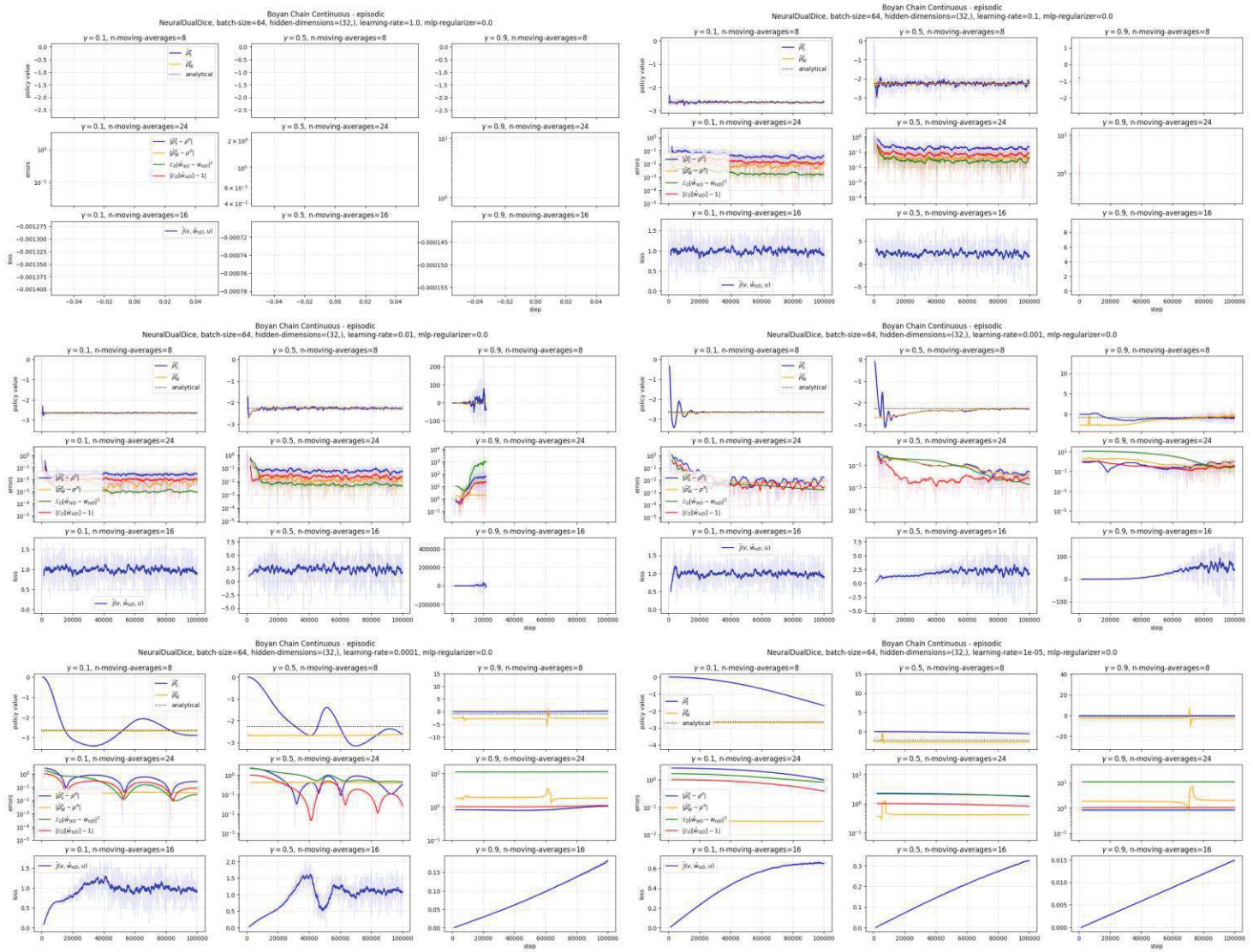


Figure 8.2: `BoyanChain Continuous - episodic - NeuralDualDice`

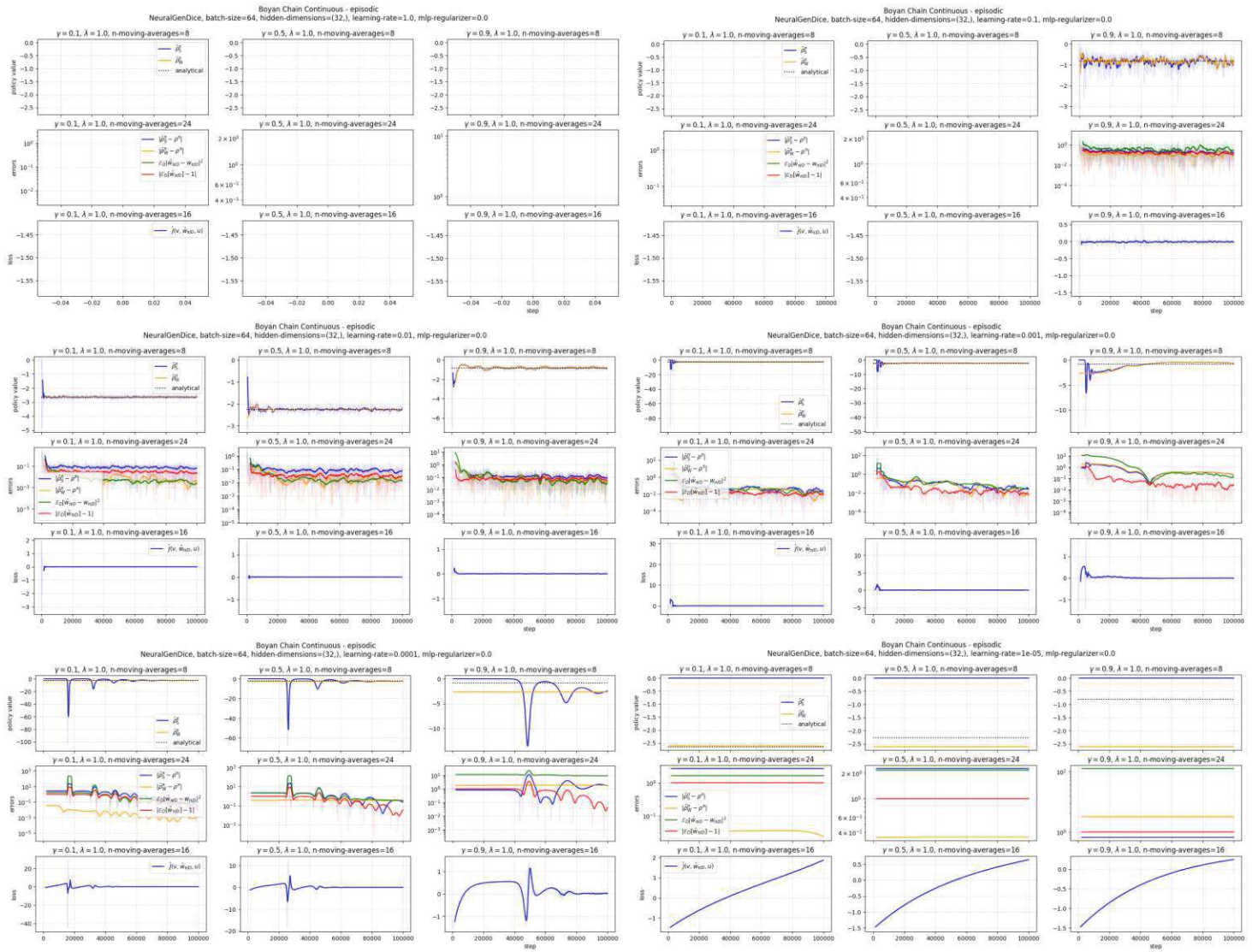


Figure 8.3: `BoyanChain Continuous - episodic - NeuralGenDice`

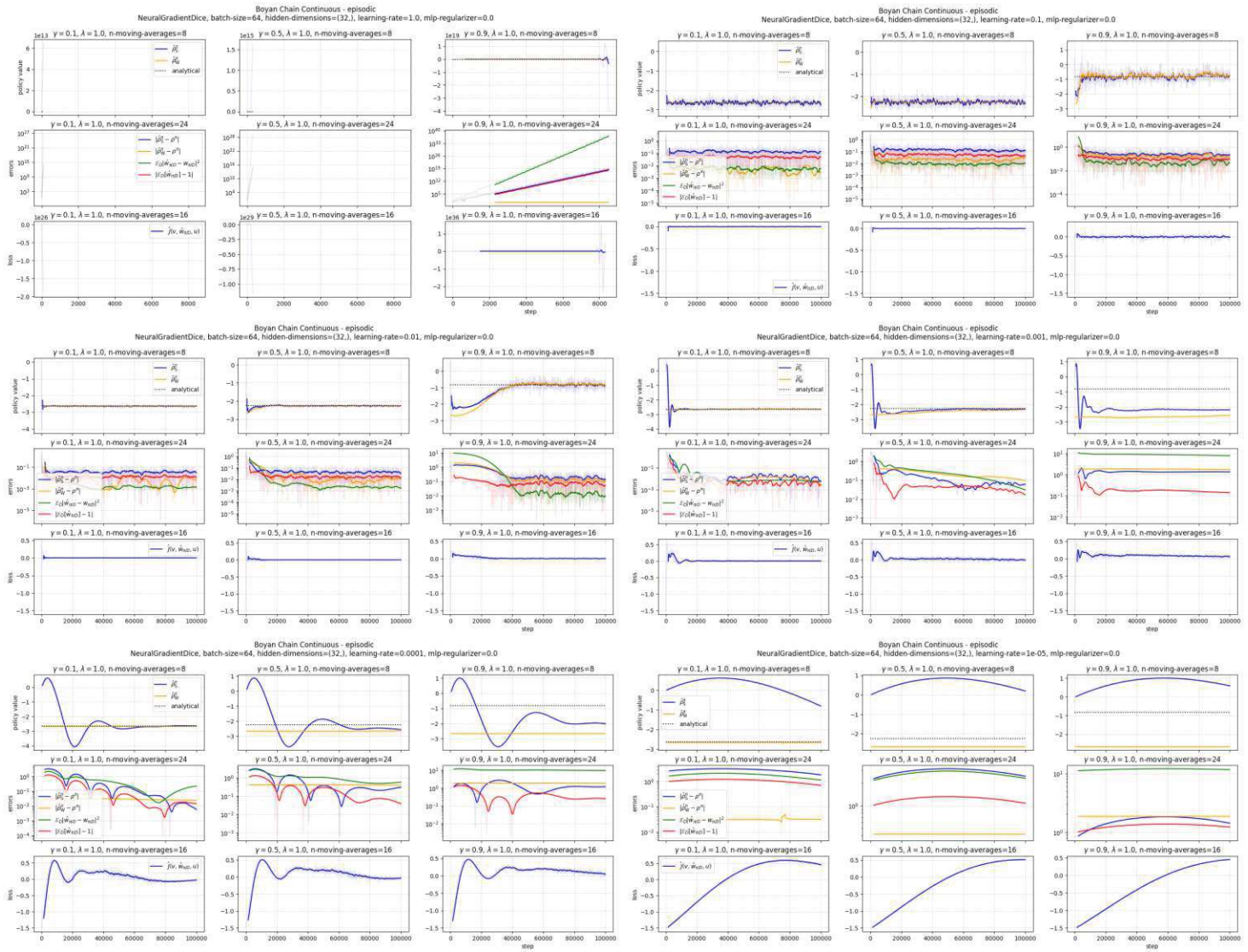


Figure 8.4: `BoyanChain Continuous - episodic - NeuralGradientDice`

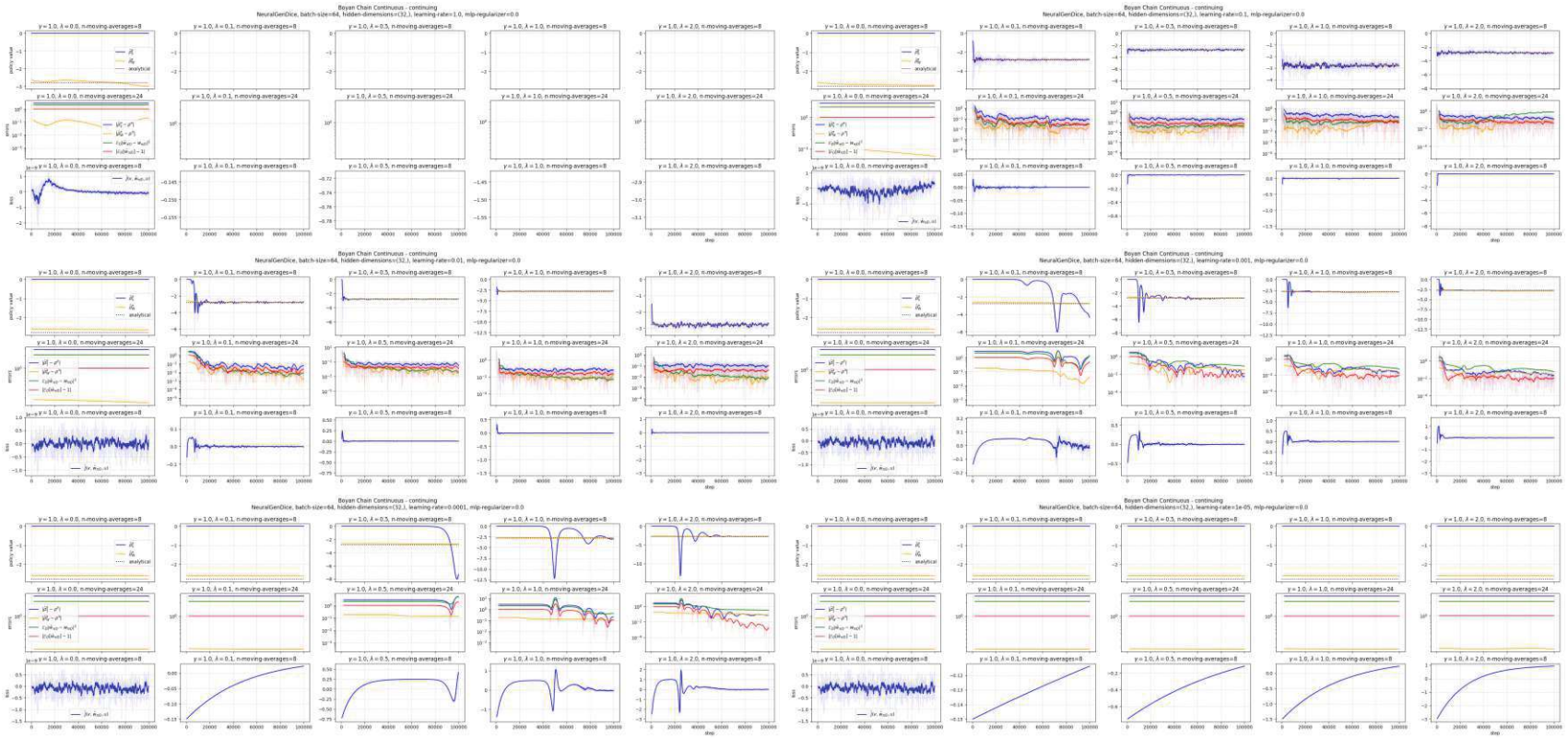


Figure 8.5: `BoyanChain Continuous - continuing - NeuralGenDice`

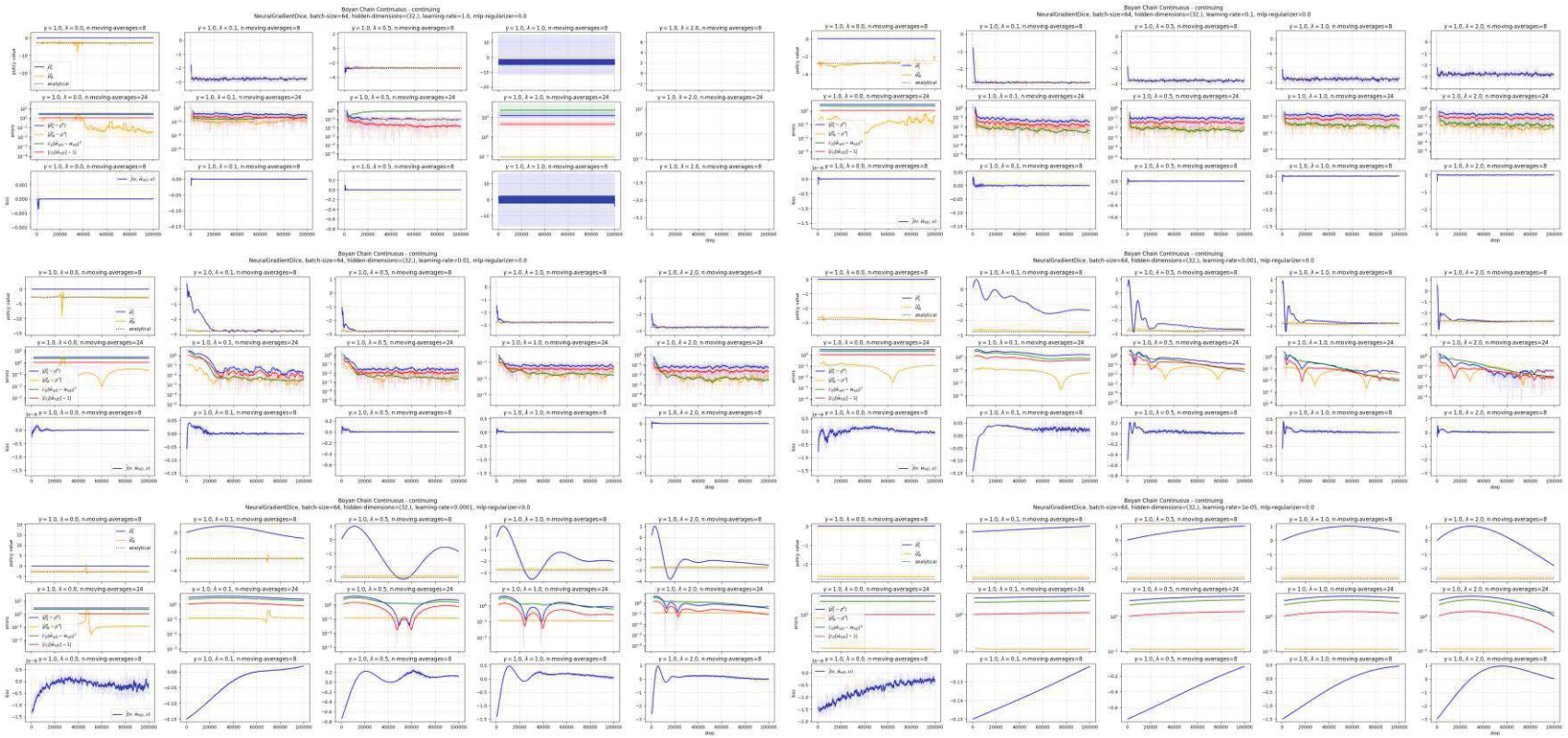


Figure 8.6: `BoyanChain Continuous - continuing - NeuralGradientDice`

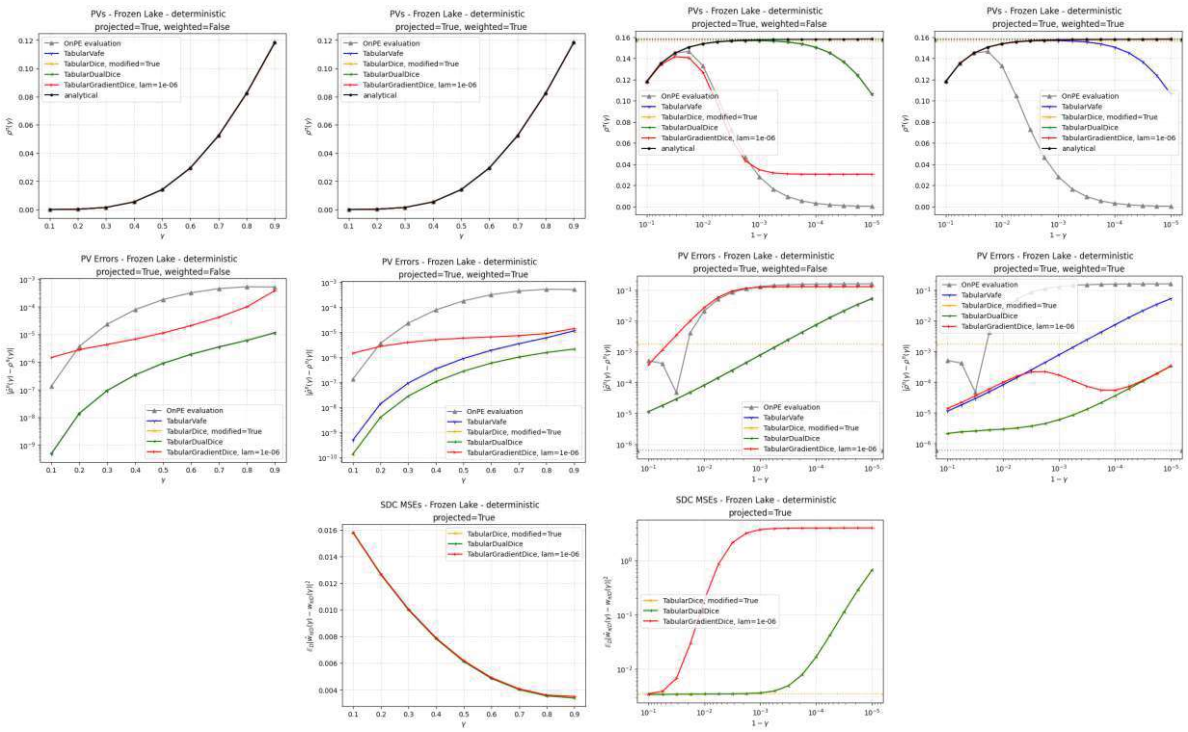


Figure 8.7: `FrozenLake - deterministic`

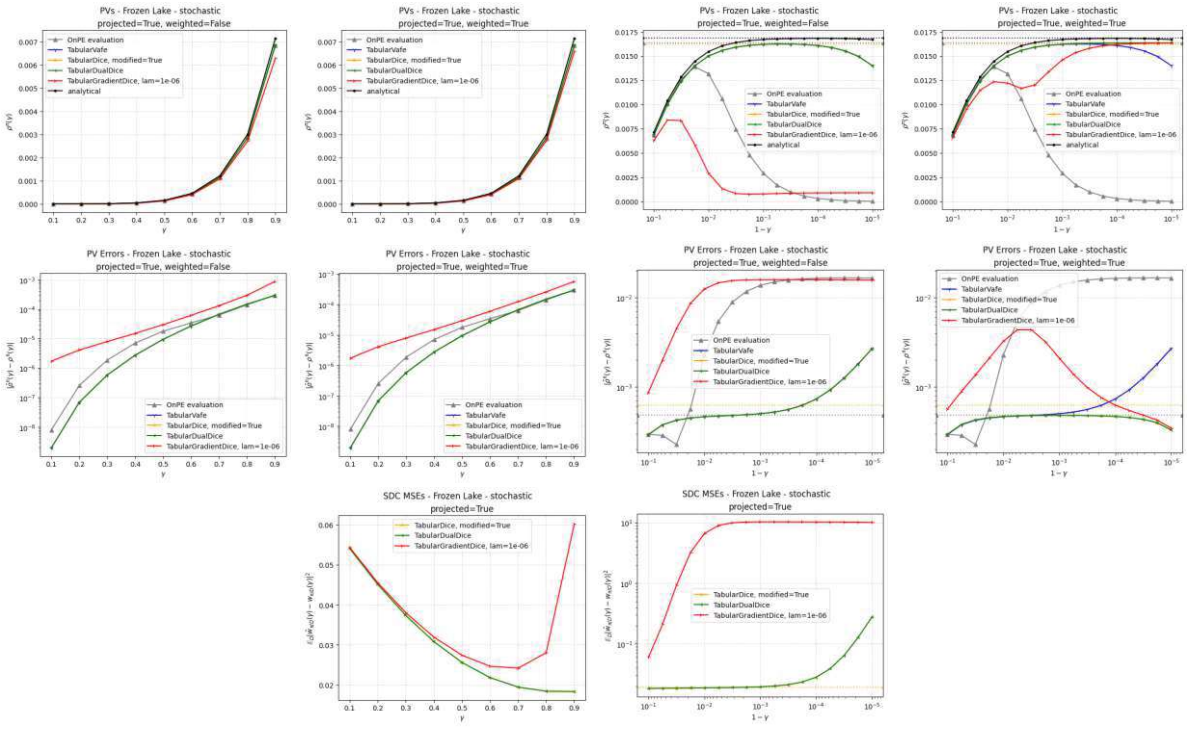


Figure 8.8: `FrozenLake - stochastic`

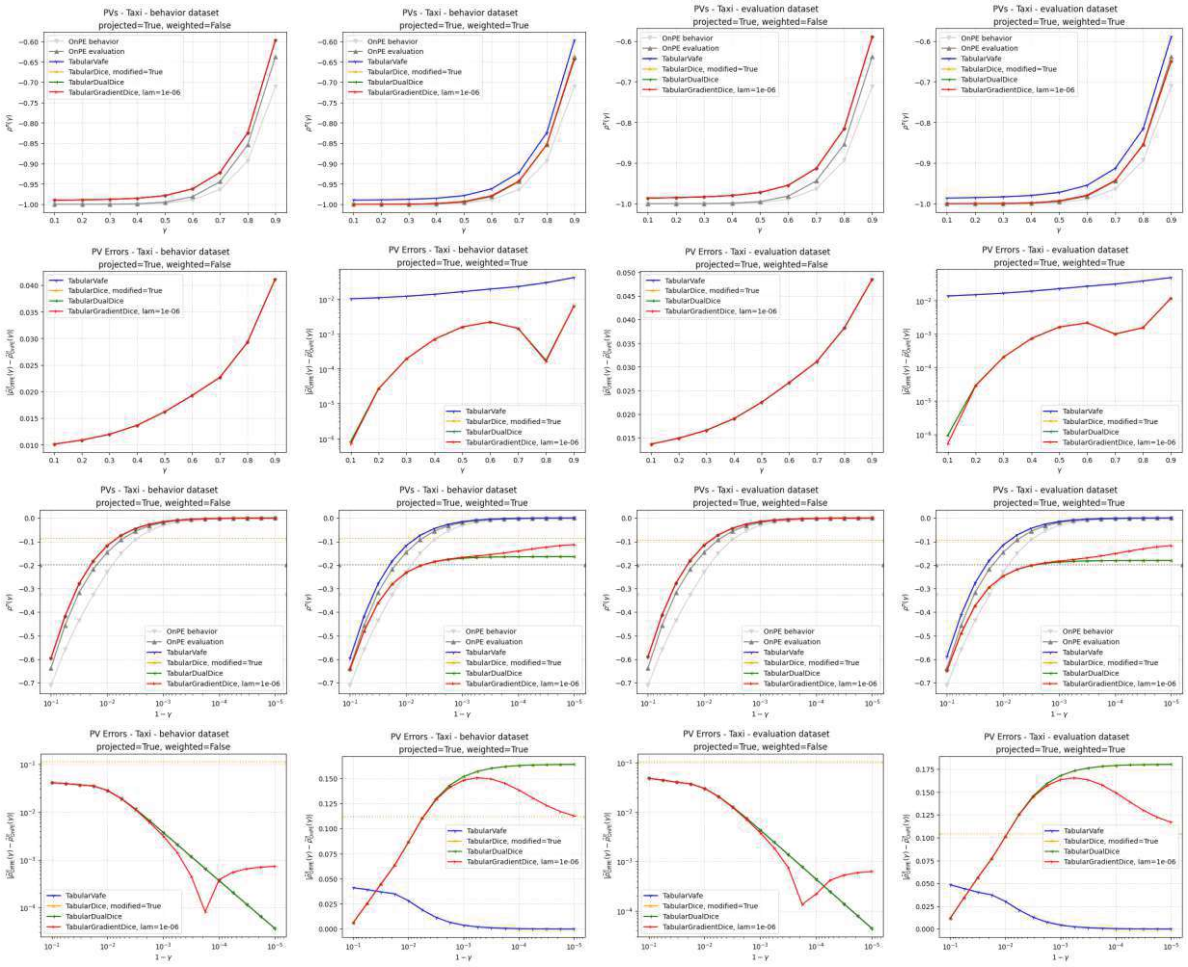


Figure 8.9: `Taxi`

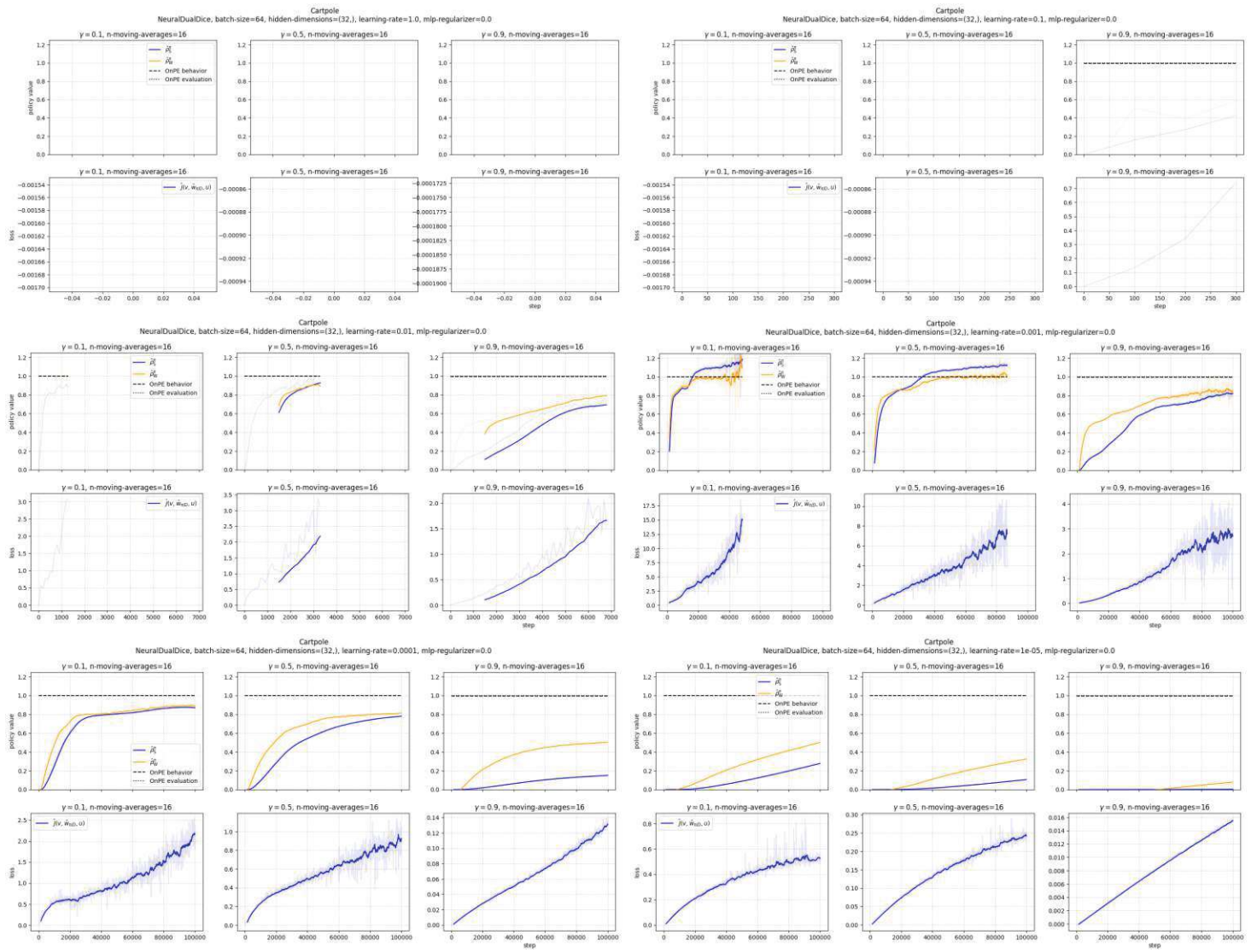


Figure 8.10: `Cartpole - NeuralDualDice`

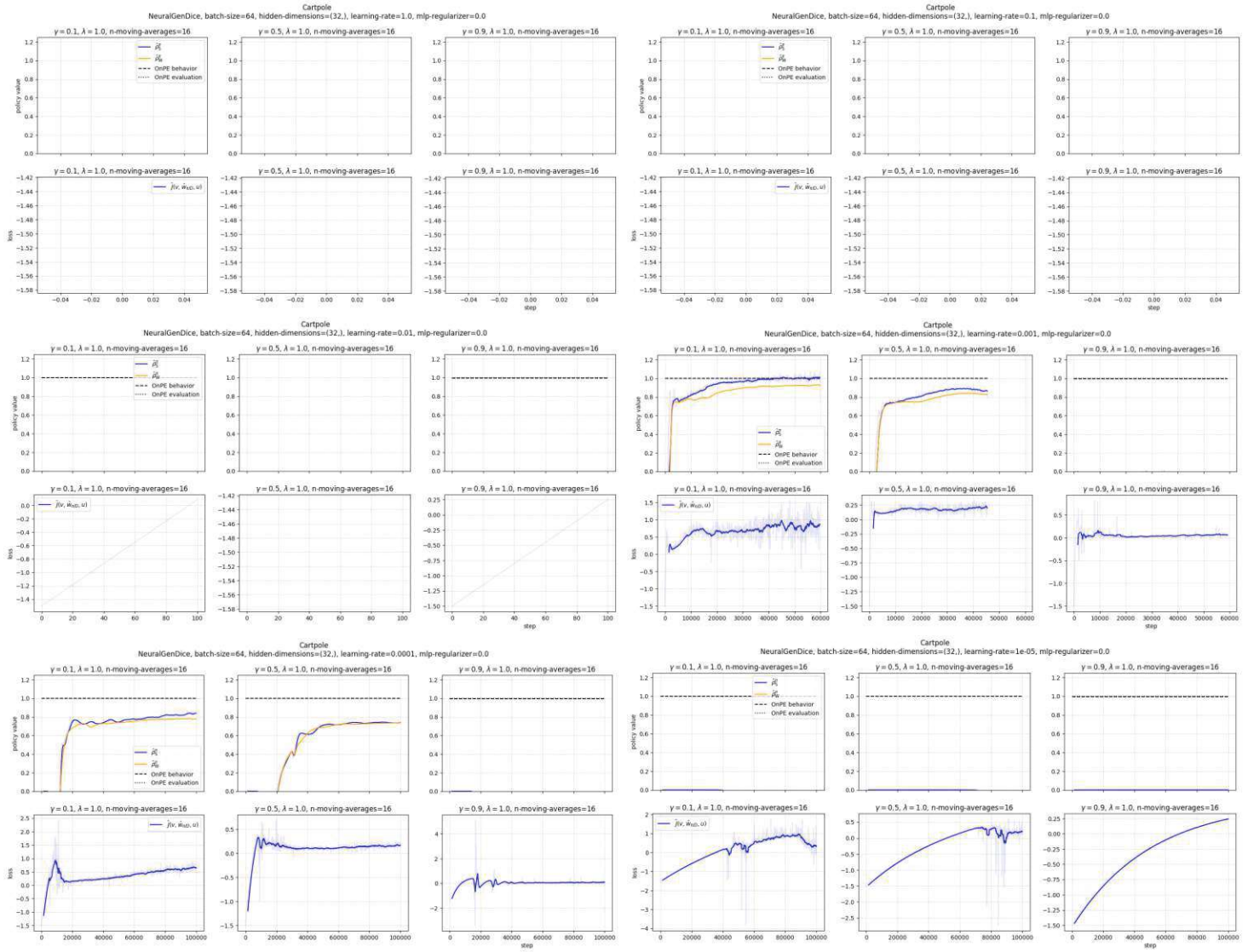


Figure 8.11: `Cartpole - NeuralGenDice`

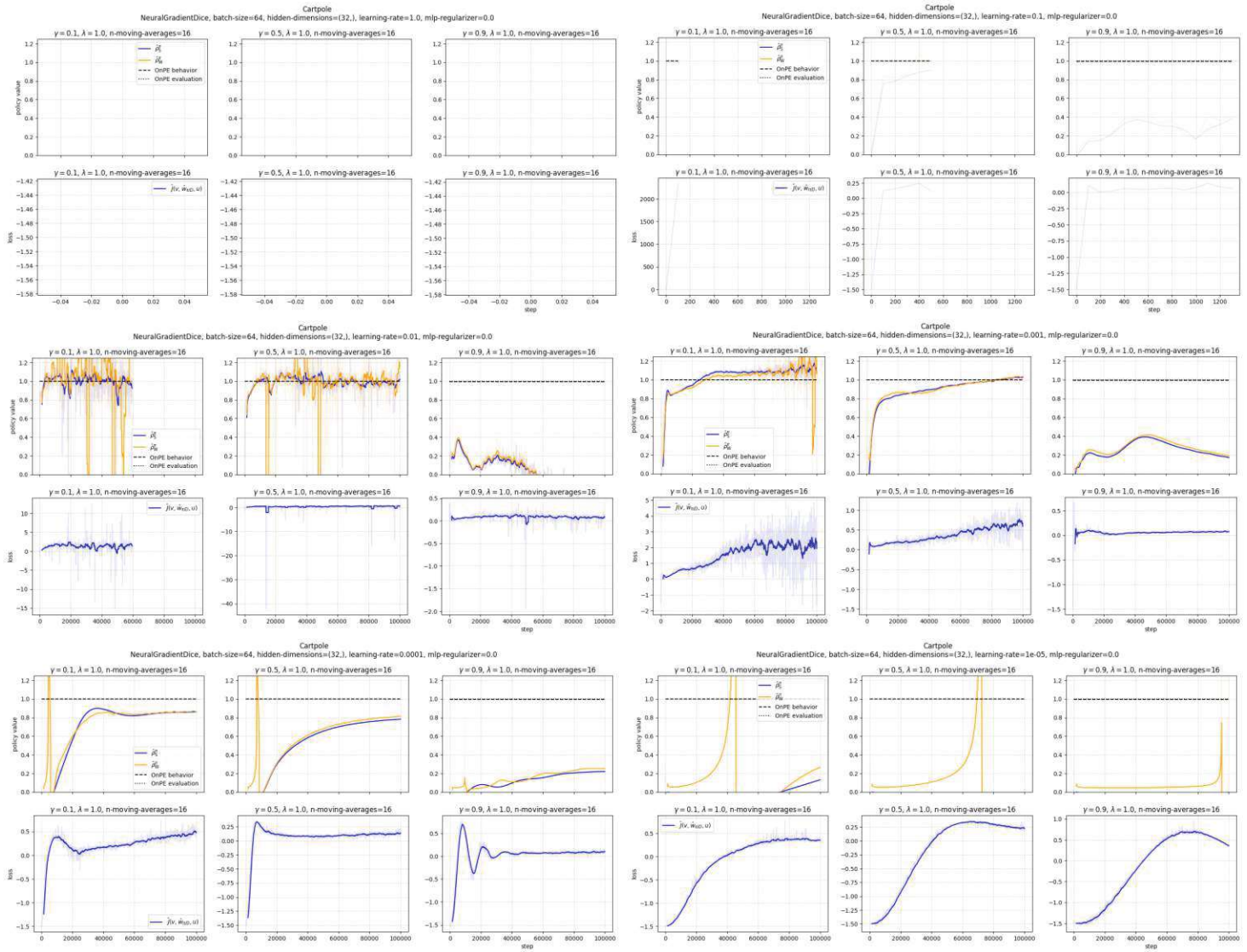


Figure 8.12: `Cartpole - NeuralGradientDice`

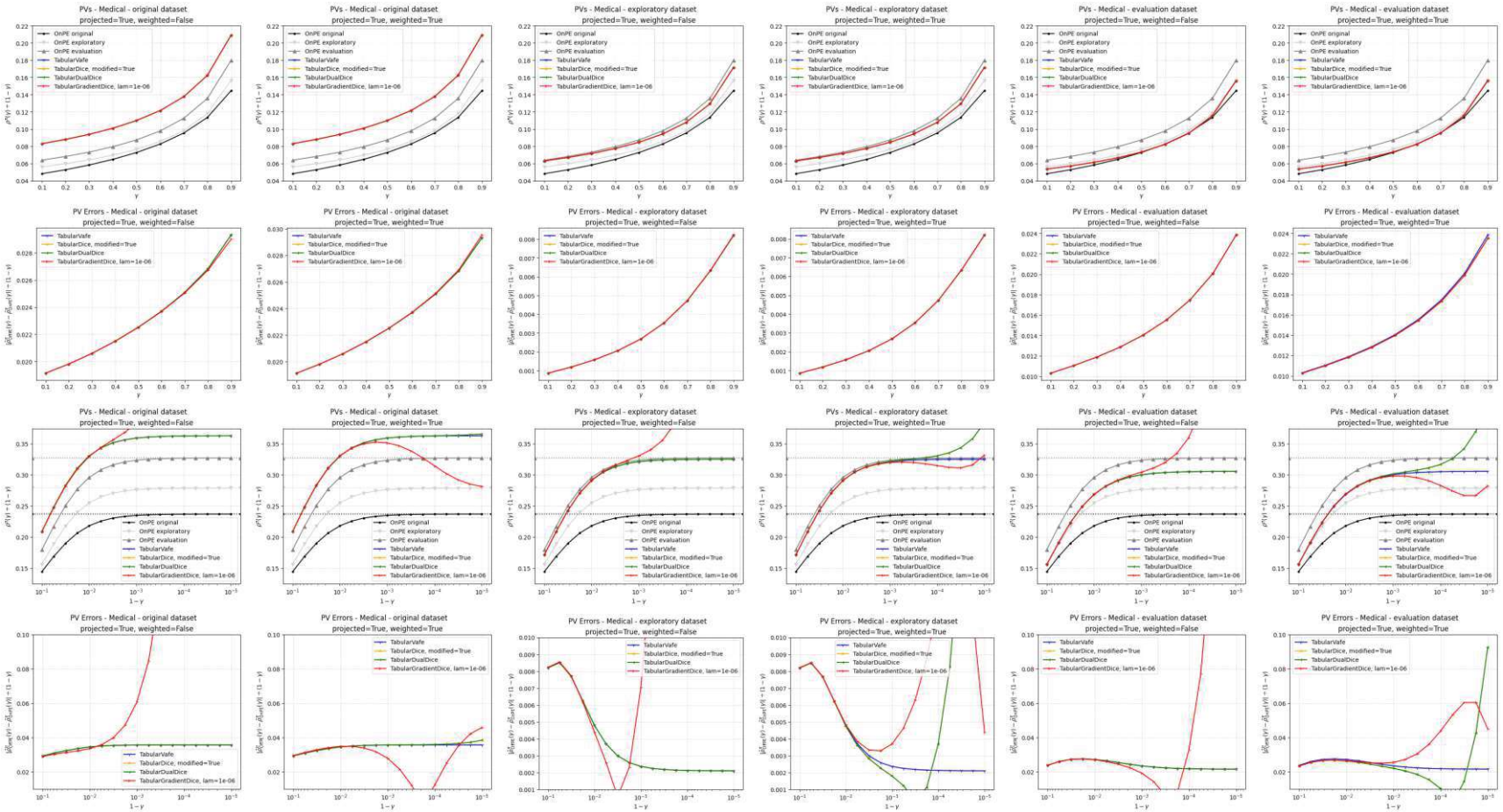


Figure 8.13: `Medical Tabular`

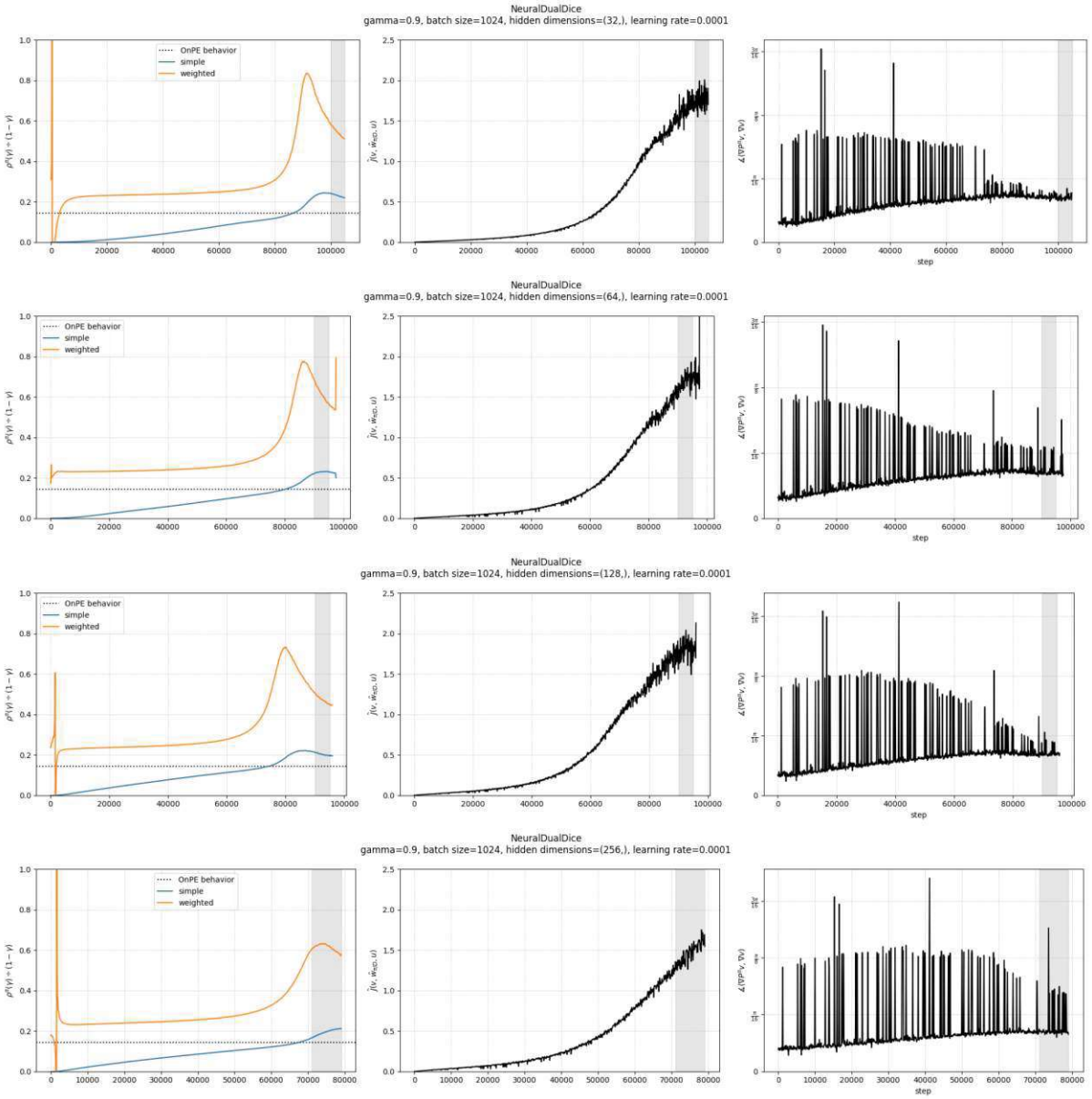


Figure 8.14: `Medical Continuous` - `NeuralDualDice`

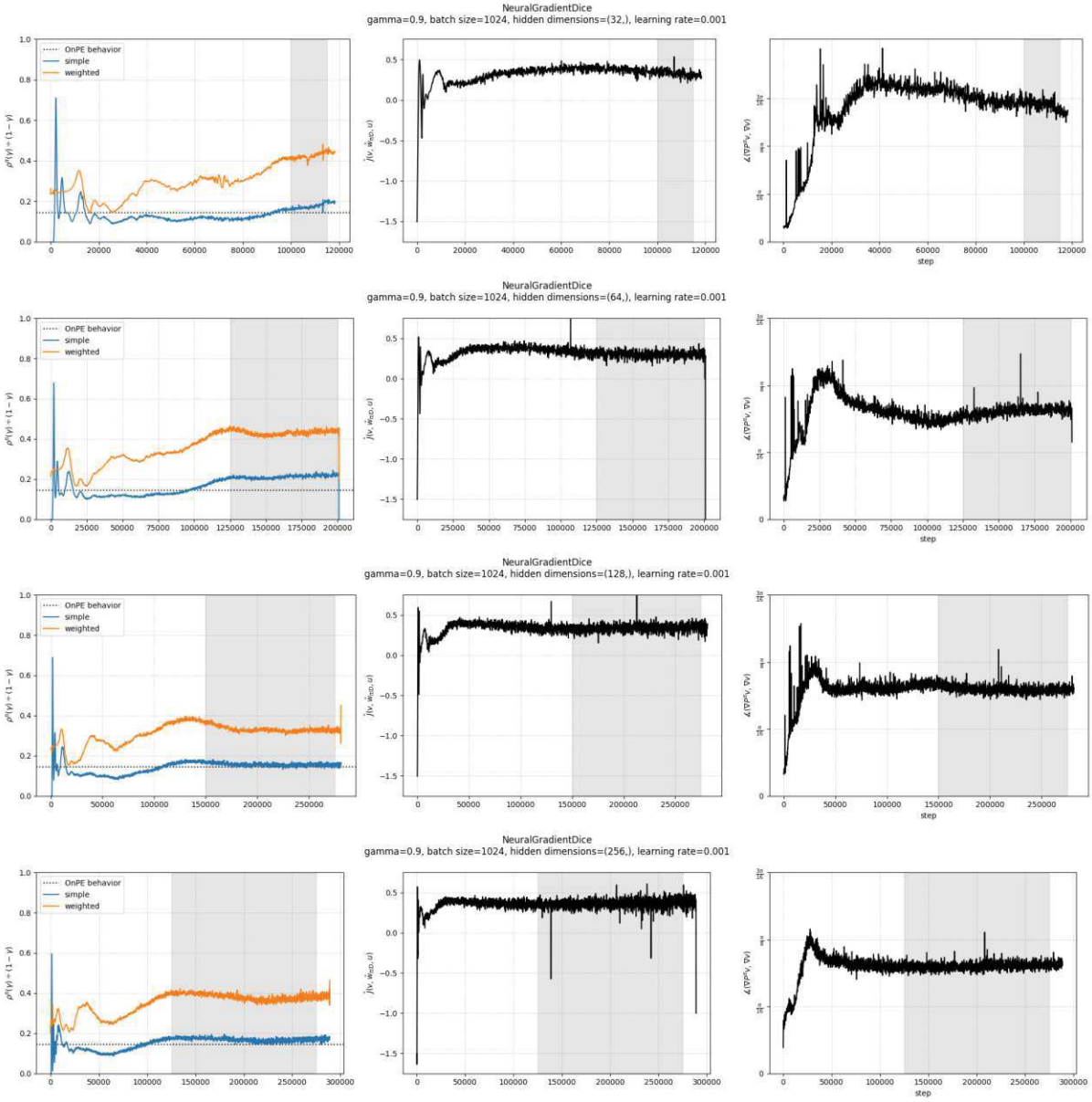


Figure 8.15: `Medical Continuous - NeuralGradientDice`

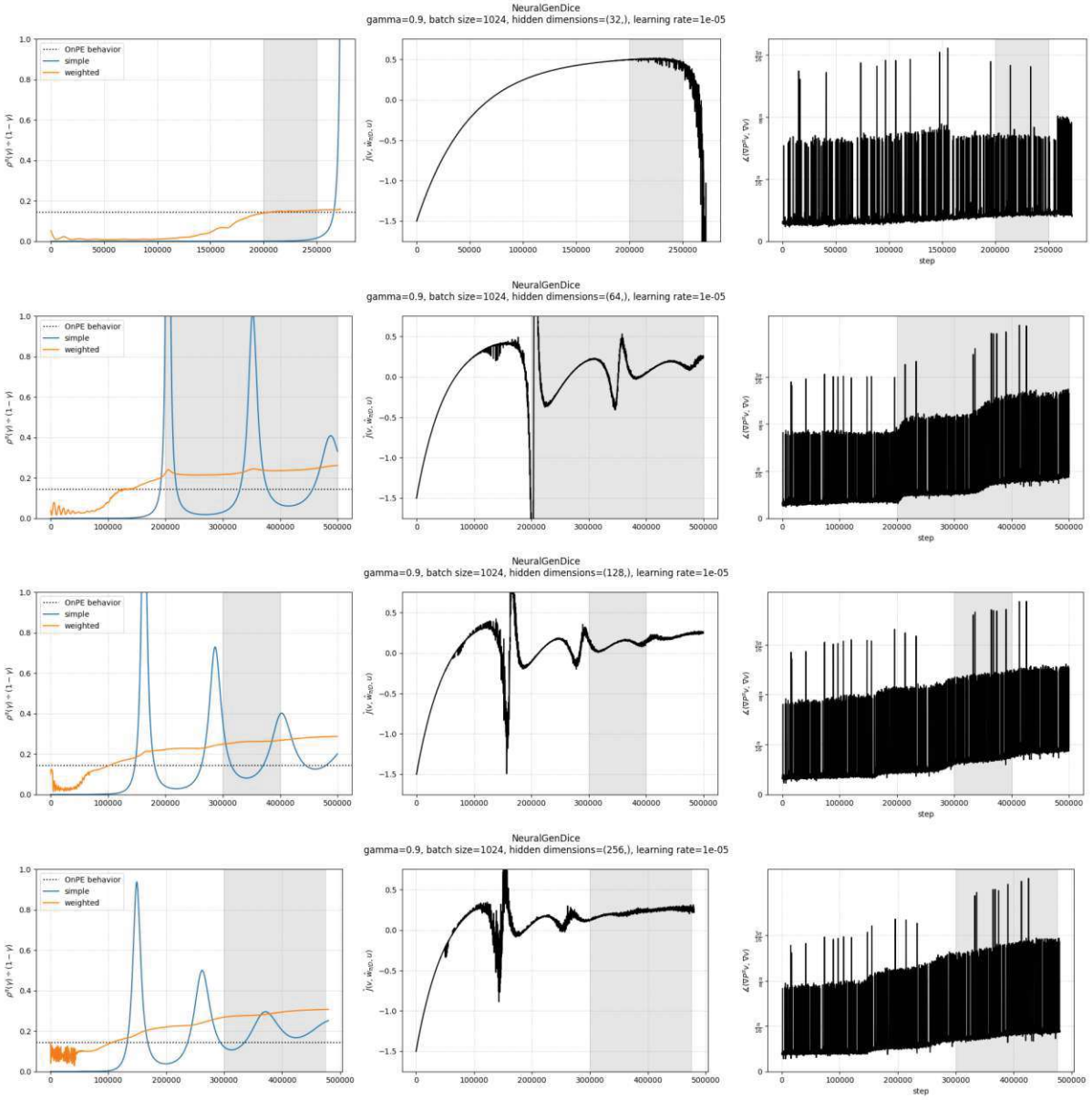


Figure 8.16: `Medical Continuous - NeuralGenDice`

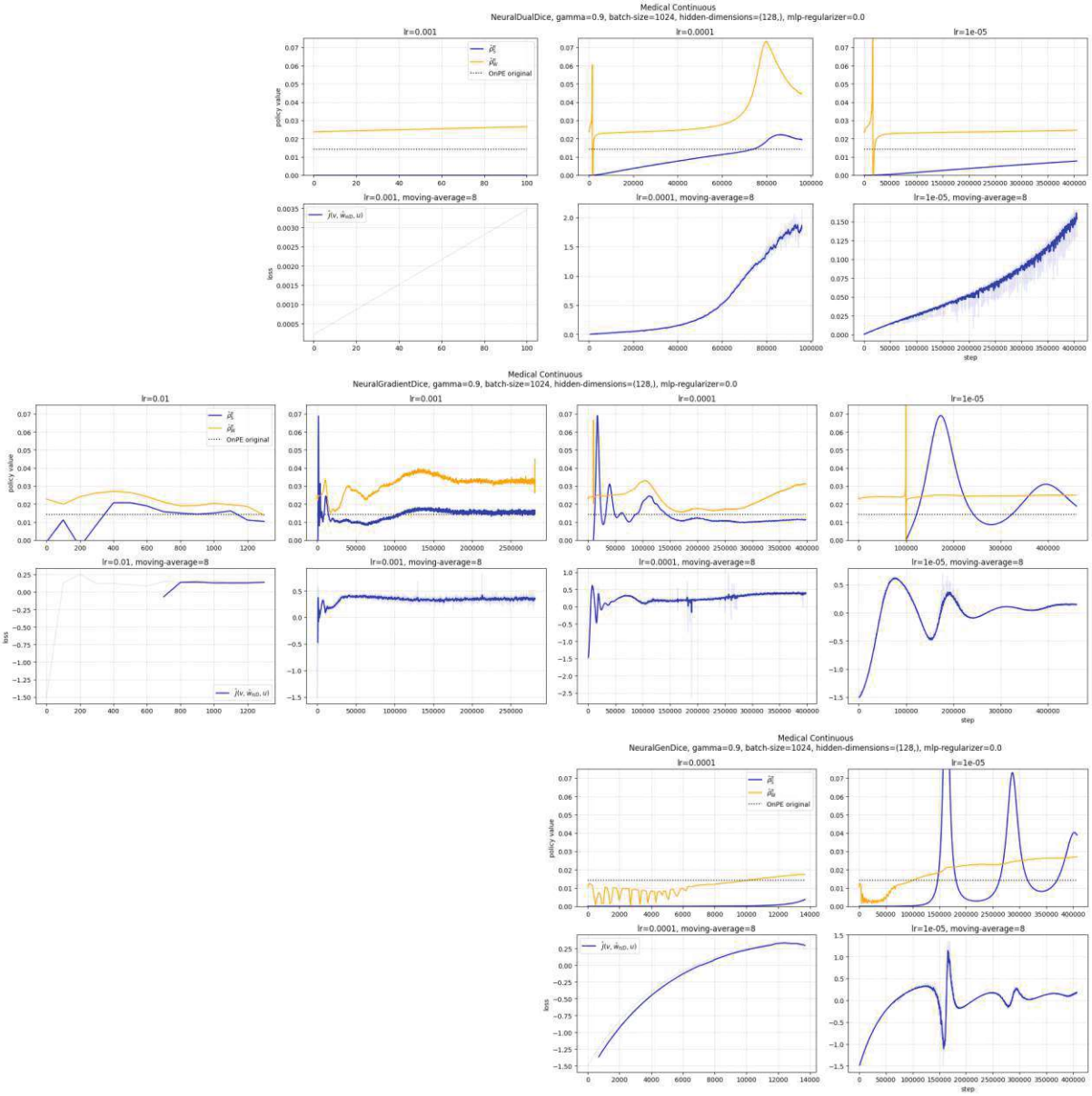


Figure 8.17: `Medical Continuous`

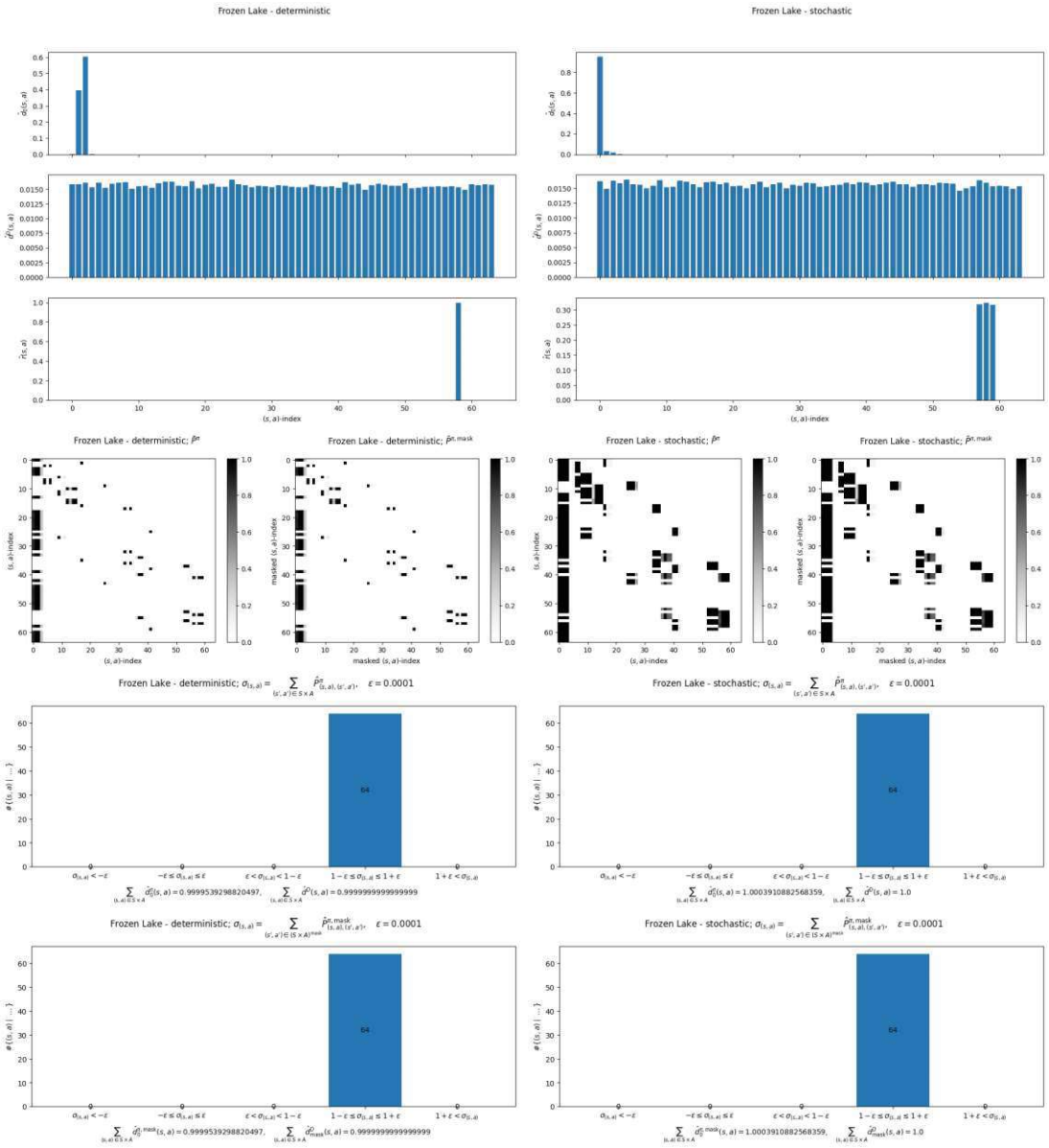


Figure 8.18: `FrozenLake` - Auxiliary Estimates

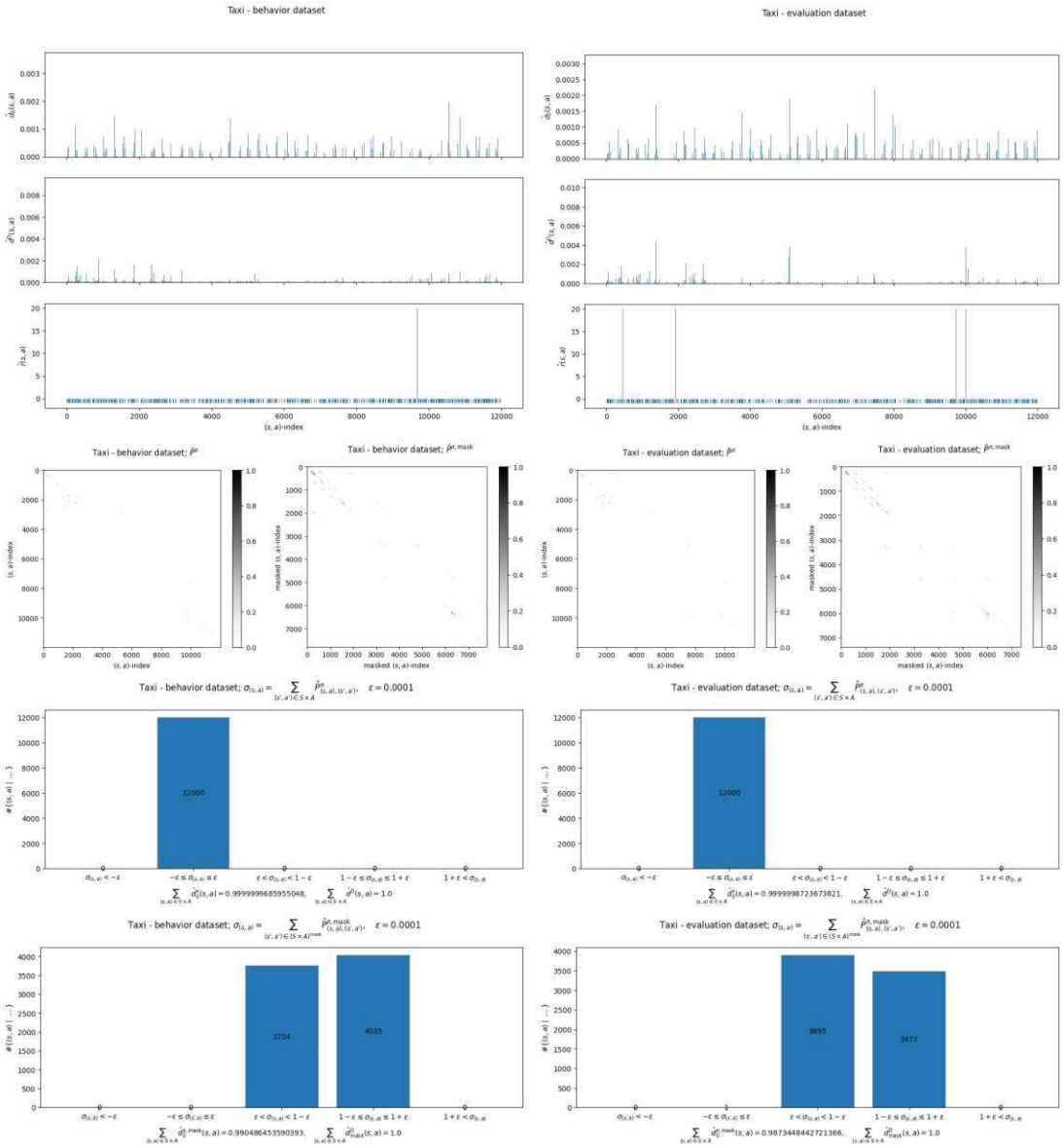


Figure 8.19: `Taxi` - Auxiliary Estimates

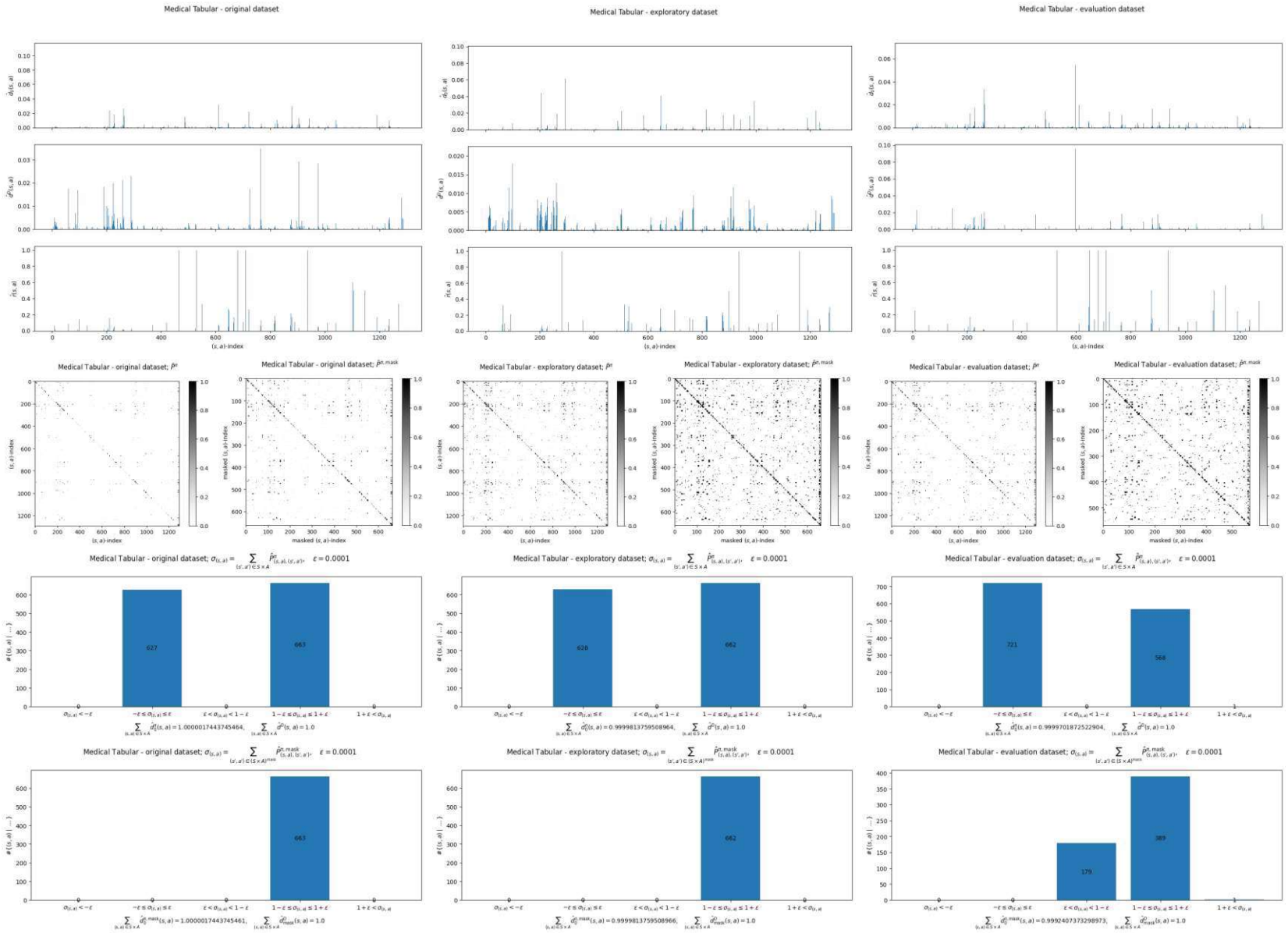


Figure 8.20: `Medical Tabular` - Auxiliary Estimates

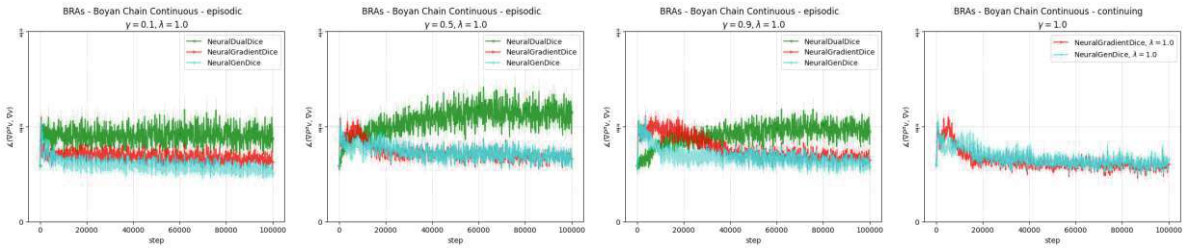


Figure 8.21: `BoyanChain Continuous - BRAs`

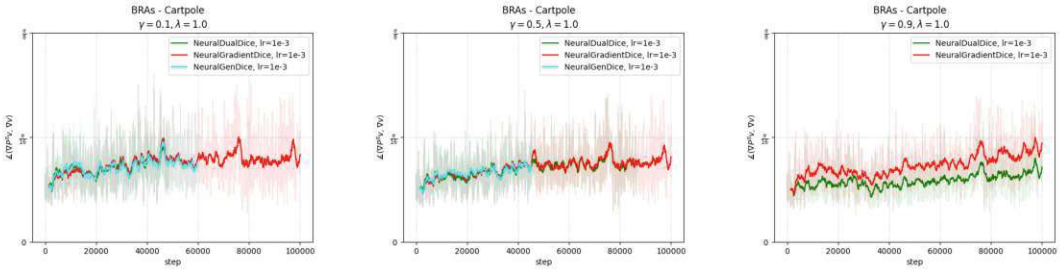


Figure 8.22: `Cartpole - BRAs`