Conditional Importance Sampling for Off-Policy Learning

Mark Rowland **Anna Harutyunyan** **Hado van Hasselt**
Diana Borsa **Tom Schaul** **Rémi Munos** **Will Dabney**
DeepMind

Abstract

The principal contribution of this paper is a conceptual framework for off-policy reinforcement learning, based on conditional expectations of importance sampling ratios. This framework yields new perspectives and understanding of existing off-policy algorithms, and reveals a broad space of unexplored algorithms. We theoretically analyse this space, and concretely investigate several algorithms that arise from this framework.

1 Introduction

Using off-policy data is crucial for many tasks in reinforcement learning (RL), including for acquiring knowledge about diverse aspects of the environment [Sutton et al., 2011], learning from memorised data [Mnih et al., 2015, Schaul et al., 2016], exploration [Watkins and Dayan, 1992], and learning to perform auxiliary tasks [Schaul et al., 2015, Jaderberg et al., 2017, Bellemare et al., 2019]. One of the fundamental techniques for correcting for the difference between the policy that generated the data and the policy that an algorithm aims to learn about is importance sampling (IS) [Metropolis and Ulam, 1949, Kahn and Harris, 1949], which was first introduced in off-policy RL by Precup et al. [2000]. Importance sampling features as a core ingredient of many off-policy algorithms [Maei, 2011, van Hasselt et al., 2014, Munos et al., 2016, Jiang and Li, 2016, Sutton et al., 2016], and is supported by strong theoretical understanding coming from the computational statistics literature [Robert and Casella, 2013, Särkkä, 2013].

Importance sampling often suffers from high variance, especially when multi-step trajectories are considered. This has motivated the study of a wide range of variance reduction techniques in off-policy reinforcement learning. These techniques include importance weight truncation [Munos et al., 2016, Espeholt et al., 2018] weighted importance sampling [Precup et al., 2000, Mahmood et al., 2014], adaptive bootstrapping [Mahmood et al., 2017], variants of emphatic TD [Hallak et al., 2016], saddle-point formulations exploiting low-variance versions of SGD [Du et al., 2017, Johnson and Zhang, 2013, Defazio et al., 2014], empirical proposal estimation [Hanna et al., 2019], doubly-robust approaches [Jiang and Li, 2016, Thomas and Brunskill, 2016], confidence bounds on returns [Thomas et al., 2015b,a, Metelli et al., 2018, Papini et al., 2019] and state distribution estimation [Xie et al., 2018, Liu et al., 2018, Kallus and Uehara, 2019a,b, Uehara and Jiang, 2019, Hallak and Mannor, 2017, Gelada and Bellemare, 2019, Nachum et al., 2019].

In this paper, we propose a new framework for variance reduction in off-policy learning, *conditional importance sampling* (CIS), based on taking conditional expectations of importance weights. This framework is motivated by the observation that when estimating a return off-policy using standard importance sampling, every action along a trajectory contributes to the importance weight, even if the action had no effect on the return observed. Intuitively, it would be preferable for the importance weight to depend only on the return itself; if two policies generate similar distributions of returns, there should be no need to perform importance weighting at all. As just one application of the CIS framework, we make this insight precise, and introduce *return-conditioned importance sampling* (RCIS), a new off-policy evaluation algorithm. Concretely, using notation introduced formally in Section 2, given a random return G, RCIS uses *conditional* importance weights of the form

$$\mathbb{E}\left[\prod_{t=1}^{n-1} \frac{\pi(A_t|X_t)}{\mu(A_t|X_t)} \,\middle|\, G\right],$$

which integrates out noise in the trajectory that is irrelevant in determining the return, leading to a lower-variance importance weight.

Proceedings of the 23rd International Conference on Artificial Intelligence and Statistics (AISTATS) 2020, Palermo, Italy. PMLR: Volume 108.

However, return is just one possible variable to condition on. The central insight of the CIS framework is that there exists a large space of variables that the importance weights can be conditioned on, with each choice leading to a different off-policy algorithm. In the remainder of the paper, we give a mathematical description of the general CIS framework, which then allows us to make several further contributions:

(i) We compare and analyse the statistical properties of CIS algorithms based on properties of the conditioning variables.
(ii) We study several specific instantiations of algorithms from this framework, including RCIS and *state-conditioned importance sampling* (SCIS, given by conditioning on the states visited by a trajectory at each timestep).
(iii) We develop practical versions of these algorithms, based on learning the conditional importance weights in a supervised manner.

We note that concurrently with this work, Liu et al. [2019] also consider conditional importance sampling in off-policy learning, establishing connections with the conditional Monte Carlo literature and undertaking statistical analysis of these estimators.

2 Background

Consider a Markov decision process (MDP) $(\mathcal{X}, \mathcal{A}, \gamma, P, \mathcal{R})$ with finite state space $\mathcal{X}$, finite action space $\mathcal{A}$, discount factor $\gamma \in [0, 1)$, transition kernel $P : \mathcal{X} \times \mathcal{A} \to \mathscr{P}(\mathcal{X})$, reward distribution probability mass function $\mathcal{R} : \mathbb{R} \times \mathcal{X} \times \mathcal{A} \to \mathbb{R}$ (so that $\mathcal{R}(r, x, a)$ encodes the probability of observing reward r after taking action a in state x), and initial state distribution $\nu \in \mathscr{P}(\mathcal{X})$[1].

Given a Markov policy $\pi : \mathcal{X} \to \mathscr{P}(\mathcal{A})$, the distribution of the process $(X_t, A_t, R_t)_{t\geq 0}$ itself is defined by $X_0 \sim \nu$, $A_t|X_{0:t}, A_{0:t-1}, R_{0:t-1} \sim \pi(\cdot|X_t)$, $R_t|X_{0:t}, A_{0:t}, R_{0:t-1} \sim \mathcal{R}(\cdot|X_t, A_t)$, and $X_{t+1}|X_{0:t}, A_{0:t}, R_{0:t} \sim P(\cdot|X_t, A_t)$ for each $t \geq 0$. We denote the full trajectory $(X_t, A_t, R_t)_{t\geq 0}$ by τ, and use the notation $\tau_{s:t}$ to denote the partial trajectory $(X_s, A_s, R_s, X_{s+1}, \ldots, X_t)$. We denote the distribution of τ under the policy π by η^π, and denote the distribution of $\tau_{s:t}$ by $\eta^\pi_{s:t}$ for any $0 \leq s \leq t$. We will also denote conditional versions of these distributions given $(X_0, A_0) = (x, a)$ in the manner $\eta^\pi|_{(x,a)}$.

[1]With some care, it is possible to show through the use of measure theory that versions of many results in this paper hold in much greater generality, such as in classes of MDPs with continuous state and/or action spaces. For the sake of accessibility and clarity of exposition, the main paper focuses on the discrete case, but we discuss how these results generalise in Appendix C.2 for the interested reader.

2.1 Policy evaluation

The *evaluation problem* with *target policy* $\pi : \mathcal{X} \to \mathscr{P}(\mathcal{A})$ is defined as estimation of the Q-function

$$Q^\pi(x, a) := \mathbb{E}_{\eta^\pi|_{(x,a)}}\left[\sum_{t=0}^{\infty} \gamma^t R_t\right], \qquad (1)$$

for all $(x, a) \in \mathcal{X} \times \mathcal{A}$. The fundamental result of value-based RL is that the Q-function in Expression (1) satisfies the *Bellman equation* $T^\pi Q^\pi = Q^\pi$ [Bellman, 1957], where the one-step Bellman evaluation operator $T^\pi : \mathbb{R}^{\mathcal{X}\times\mathcal{A}} \to \mathbb{R}^{\mathcal{X}\times\mathcal{A}}$ is defined by

$$(T^\pi Q)(x, a) = \mathbb{E}_{\eta^\pi|_{(x,a)}}[R_0 + \gamma Q(X_1, A_1)],$$

for all $Q \in \mathbb{R}^{\mathcal{X}\times\mathcal{A}}$ and $(x, a) \in \mathcal{X} \times \mathcal{A}$. As T^π is a contraction in $(\mathbb{R}^{\mathcal{X}\times\mathcal{A}}, \|\cdot\|_\infty)$, Q^π is its *unique* fixed point, and repeated application of T^π to any initial Q-function will converge to Q^π. An evaluation algorithm may therefore seek to (approximately) perform a recursion of the form $Q_{k+1} \leftarrow T^\pi Q_k$ $(k \geq 1)$, with the aim of converging to Q^π. More general classes of contractive operators with fixed point Q^π can also be considered, such as the Retrace operator [Munos et al., 2016], and the n-step Bellman operator, given by

$$((T^\pi)^n Q)(x, a) = \mathbb{E}_{\eta^\pi|_{(x,a)}}\left[\sum_{t=0}^{n-1} \gamma^t R_t + \gamma^n Q(X_n, A_n)\right].$$

2.2 Off-policy policy evaluation

Exact computation of the expectations defining the above operators is often intractable, and so Monte Carlo[2] estimators based on trajectories sampled from the environment are used [Bertsekas and Tsitsiklis, 1996, Szepesvári, 2010, Sutton and Barto, 2018]. Further, it is often desirable, or necessary, to use trajectories sampled from a different distribution η^μ, based on a *behaviour policy* $\mu : \mathcal{X} \to \mathscr{P}(\mathcal{A})$; in such cases, the problem is said to be *off-policy*.

A common estimator for the application of the n-step Bellman operator $(T^\pi)^n$ to a Q-function Q at a specific state-action pair $(x, a) \in \mathcal{X} \times \mathcal{A}$ is given by sampling $\tau_{0:n}$ from $\eta^\mu_{0:n}|_{(x,a)}$, and computing a *bootstrapped return*, defined by

$$\bar{G}^\pi_{0:n} := \sum_{t=0}^{n-1} \gamma^t R_t + \gamma^n V(X_n; \pi), \qquad (2)$$

[2]Throughout, we use the term "Monte Carlo" in its statistical sense, to mean sampled-based approximation of any expectation, including those defining temporal difference algorithms.

where $V(x;\pi) = \mathbb{E}_{A\sim\pi(\cdot|x)}[Q(x,A)]$, and an *importance-weighting correction term*, defined by

$$\rho^{\pi,\mu}_{s:t} := \prod_{i=s}^{t} \frac{\pi(A_i|X_i)}{\mu(A_i|X_i)}\,. \quad (3)$$

for $1 \leq s \leq t$, and finally forming the *ordinary importance sampling* (OIS) estimator

$$\bar{G}^{\mathrm{OIS};\pi,\mu}_{0:n} := \rho^{\pi,\mu}_{1:n-1}\bar{G}^{\pi}_{0:n}\,. \quad (4)$$

Much research in off-policy learning is concerned with constructing such estimators that have desirable statistical properties, such as low variance and consistency. Throughout, we will assume the *support condition*:

$$\mathrm{supp}(\pi(\cdot|x)) \subseteq \mathrm{supp}(\mu(\cdot|x)) \quad \text{for all } x\in\mathcal{X}\,, \quad \text{(SC)}$$

a mild assumption that is sufficient for unbiased importance sampling, which is satisfied by exploratory behaviours such as ε-greedy. This is equivalent to absolute continuity of π with respect to μ at each state; intuitively, this ensures that any trajectory that can arise by following π is also realisable under μ.

3 Preliminary analysis

As a warm-up and motivation for the conceptual framework we present in the next section, we analyse some commonly-used off-policy Monte Carlo estimators.

3.1 Ordinary importance sampling

We begin with a formal proof of the unbiasedness of the OIS estimator, a well-known result in the literature. In this and many results that follow, we will be interested in distributions over trajectories conditioned on some initial state-action pair $(x,a)\in\mathcal{X}\times\mathcal{A}$; this will be present in the notation, but we avoid continuously mentioning it in the text for brevity. We examine the proof of this result in some detail, since it will be informative for the original results that follow. Proofs of other results in the paper are given in Appendix A.

Proposition 3.1. Assume the support condition (SC) holds. For a trajectory drawn from $\eta^{\mu}|_{(x,a)}$, the OIS estimator in Expression (4) is unbiased for the output of the n-step return operator $(T^{\pi})^n$. That is,

$$\mathbb{E}_{\eta^{\mu}|_{(x,a)}}\left[\rho^{\pi,\mu}_{1:n-1}\bar{G}^{\pi}_{0:n}\right] = \mathbb{E}_{\eta^{\pi}|_{(x,a)}}\left[\bar{G}^{\pi}_{0:n}\right]\,. \quad (5)$$

Proof. We first observe that the ratio of policy probabilities that appears within the factor $\rho^{\pi,\mu}_{1:n-1}$ can also be interpreted as the importance ratio for the conditional trajectory distributions $\eta^{\mu}_{0:n}|_{(x,a)}$ and $\eta^{\pi}_{0:n}|_{(x,a)}$, as the following calculation shows:

$$\begin{aligned}
&\frac{\eta^{\pi}_{0:n}|_{(x,a)}(\tau_{0:n})}{\eta^{\mu}_{0:n}|_{(x,a)}(\tau_{0:n})} && (6)\\
=&\frac{P(X_1|x_0,a_0)\mathcal{R}(R_0|x_0,a_0)}{P(X_1|x_0,a_0)\mathcal{R}(R_0|x_0,a_0)}\times\\
&\quad\frac{\prod_{t=1}^{n-1}\pi(A_t|X_t)\mathcal{R}(R_t|X_t,A_t)P(X_{t+1}|X_t,A_t)}{\prod_{t=1}^{n-1}\mu(A_t|X_t)\mathcal{R}(R_t|X_t,A_t)P(X_{t+1}|X_t,A_t)}\\
=&\prod_{t=1}^{n-1}\frac{\pi(A_t|X_t)}{\mu(A_t|X_t)} && (7)\\
=&\rho^{\pi,\mu}_{1:n-1}\,.
\end{aligned}$$

Noting also that the term $\bar{G}^{\pi}_{0:n}$ in Equation (5) is simply a function of the random truncated trajectory $\tau_{0:n}$, we may now appeal to standard importance sampling theory, using the notation $\Psi(\tau_{0:n}) = \bar{G}^{\pi}_{0:n}$, to obtain

$$\begin{aligned}
\mathbb{E}_{\eta^{\mu}|_{(x,a)}}[\rho^{\pi,\mu}_{1:n-1}\bar{G}^{\pi}_{0:n}] =&\mathbb{E}_{\eta^{\mu}|_{(x,a)}}\left[\frac{\eta^{\pi}_{0:n}|_{(x,a)}(\tau_{0:n})}{\eta^{\mu}_{0:n}|_{(x,a)}(\tau_{0:n})}\Psi(\tau_{0:n})\right]\\
=&\mathbb{E}_{\eta^{\pi}|_{(x,a)}}[\Psi(\tau_{0:n})]\\
=&\mathbb{E}_{\eta^{\pi}|_{(x,a)}}\left[\bar{G}^{\pi}_{0:n}\right]\,,
\end{aligned}$$

as required. □

We highlight two points. Firstly, note that the argument above did not depend on any special structure of $\bar{G}^{\pi}_{0:n}$, other than that it was expressible as a function of the truncated trajectory $\tau_{0:n}$; this analysis is therefore readily applicable to many other functions of the trajectory beyond n-step returns, as we will see in Section 4. Secondly, note that within the proof we showed that the familiar product of ratios of *action* probabilities (7) is precisely equal to the ratio of *trajectory* probabilities (6), a fact we will use in the remainder of the paper.

3.2 Per-decision importance sampling

Whilst the OIS target of Expression (4) is straightforwardly understood, it often has very high variance. A popular variant that aims to address this shortcoming is given by the *per-decision importance sampling* (PDIS) [Precup et al., 2000] target:

$$\bar{G}^{\mathrm{PDIS};\pi,\mu}_{0:n} = \sum_{t=0}^{n-1}\rho^{\pi,\mu}_{1:t}\gamma^t R_t + \rho^{\pi,\mu}_{1:n-1}\gamma^n V(X_n;\pi)\,, \quad (8)$$

The intuition behind this estimator is that each individual reward is only weighted by importance ratios for actions that preceded the reward, it being unnecessary to account for the off-policyness of future actions. This estimator is also unbiased, and is described in the literature as often having lower variance than the OIS estimator. We show below that each constituent term of the PDIS estimator *is* lower variance than the counterpart term in the OIS estimator.

Proposition 3.2. Assuming the support condition (SC), each term in the PDIS estimator has variance at most that of the corresponding term in the OIS estimator. That is, for all $0 \leq t \leq n-1$,

$$\mathrm{Var}_{\eta^\mu|_{(x,a)}}\left(\rho^{\pi,\mu}_{1:t}\gamma^t R_t\right) \leq \mathrm{Var}_{\eta^\mu|_{(x,a)}}\left(\rho^{\pi,\mu}_{1:n-1}\gamma^t R_t\right) .$$

The proof technique provides the main insight giving rise to the *conditional importance sampling* framework described in the next section, so we provide a sketch below. The fundamental idea is to show that each term in the estimator $\bar{G}^{\mathrm{PDIS};\pi,\mu}_{0:n}$ can be viewed as a *conditional expectation* of a corresponding term in the estimator $\bar{G}^{\mathrm{OIS};\pi,\mu}_{0:n}$; we can then use the following well known variance decomposition for any two real-valued random variables Z_1 and Z_2 with finite second moments:

$$\begin{aligned}\mathrm{Var}(Z_1) =& \mathrm{Var}(\mathbb{E}\left[Z_1|Z_2\right]) + \mathbb{E}\left[\mathrm{Var}(Z_1|Z_2)\right] \\ \geq& \mathrm{Var}(\mathbb{E}\left[Z_1|Z_2\right])\,,\end{aligned} \tag{9}$$

with the inequality strict whenever Z_1 is not $\sigma(Z_2)$-measurable, or not a function of Z_2, using non-measure-theoretic terminology. This idea is closely related to the notion of Rao-Blackwellisation, a variance reduction technique which is ubiquitous across statistics and signal processing [Casella and Berger, 2002, Särkkä, 2013, Robert and Casella, 2013].

To apply this result to prove Proposition 3.2, consider the term $\rho^{\pi,\mu}_{1:n-1}\gamma^t R_t$ from the OIS estimator, and the term $\rho^{\pi,\mu}_{1:t}\gamma^t R_t$ from the PDIS estimator. A direct calculation yields

$$\begin{aligned}&\mathbb{E}_{\eta^\mu|_{(x,a)}}\left[\rho^{\pi,\mu}_{1:n-1}\gamma^t R_t \middle| X_{0:t}, A_{0:t}, R_t\right] \\ =&\rho^{\pi,\mu}_{1:t}\gamma^t R_t \ \mathbb{E}_{\eta^\mu}\left[\rho^{\pi,\mu}_{t+1:n-1}\middle| X_{0:t}, A_{0:t}, R_t\right] \\ =&\rho^{\pi,\mu}_{1:t}\gamma^t R_t\,.\end{aligned}$$

The final equality follows from the general fact that when the support condition (SC) is satisfied, the expectation of an importance weight with respect to the importance sampling distribution is 1. Thus, the PDIS term really is a conditional expectation of the corresponding term in the OIS estimator. The bootstrap terms in the PDIS and OIS estimators are in fact equal, and hence the result of Proposition 3.2 follows. Note that Liu et al. [2019] also analyse the covariance terms, showing that it is possible for high covariances to outweigh the benefits of smaller per-term variance.

We are now ready to generalise the reasoning presented in this section, and present the main conceptual framework of the paper.

4 Conditional importance sampling: Theory

The proof of Proposition 3.2 highlights an important observation; the PDIS estimator in Expression (8) can be interpreted as taking particular conditional expectations of the OIS estimator in Expression (4) as a means of reducing variance. It will turn out that this process of taking conditional expectations is a productive way of both discovering new off-policy importance sampling methods, and also understanding their statistical properties. For this reason, we take some time to spell out this logic more generally.

Consider the problem of estimating $\mathbb{E}_{\eta^\pi_{0:n}|_{(x,a)}}[\Psi(\tau_{0:n})]$, for some function Ψ of the truncated trajectory $\tau_{0:n}$, via importance sampling. A standard importance estimator, taking $\tau_{0:n} \sim \eta^\mu_{0:n}|_{(x,a)}$, is given by

$$\frac{\eta^\pi_{0:n}|_{(x,a)}(\tau_{0:n})}{\eta^\mu_{0:n}|_{(x,a)}(\tau_{0:n})}\Psi(\tau_{0:n})\,. \tag{10}$$

If Ψ extracts an n-step return from the trajectory, this yields the standard OIS estimator, and if Ψ extracts a single reward R_t, this yields an individual term from the OIS estimator. In Section 3, we saw that in this latter case, a way of reducing the variance of the resulting estimator is to take the conditional expectation given the random variables $(X_{0:t}, A_{0:t}, R_t)$, essentially because $\Psi(\tau_{0:n}) = R_t$ is expressible as a function of $(X_{0:t}, A_{0:t}, R_t)$, and the trajectory importance weight is *not* expressible as a function of $(X_{0:t}, A_{0:t}, R_t)$, allowing some extraneous sources of noise to be integrated out. We now formalise this in greater generality.

Definition 4.1. Given a functional Ψ of a trajectory $\tau_{0:n}$, we say that Ψ *factors through* another functional Φ if there exists a third function h (independent of the MDP) with $\Psi = h \circ \Phi$, or equivalently, if $\Psi(\tau_{0:n})$ can be written as a function of $\Phi(\tau_{0:n})$ for all values of $\tau_{0:n}$. We say that Φ is a *sufficient conditioning functional* (SCF) for Ψ.

This notion of sufficient conditioning functionals suggests the following general framework for constructing off-policy estimators, generalising the perspective of PDIS given in the previous section.

Conditional importance sampling.
Given a target functional $\Psi(\tau_{0:n})$, select an SCF Φ for Ψ and construct the estimator

$$\mathbb{E}_{\eta^\mu|_{(x,a)}}\left[\frac{\eta^\pi_{0:n}|_{(x,a)}(\tau_{0:n})}{\eta^\mu_{0:n}|_{(x,a)}(\tau_{0:n})}\middle|\Phi(\tau_{0:n})\right]\Psi(\tau_{0:n})\,. \tag{11}$$

Through different choices of Ψ and Φ, this yields a wide space of possible off-policy learning algorithms; we refer to this as the *conditional importance sampling*

(CIS) framework. We begin with some basic analysis of the properties of these estimators.

Proposition 4.2. Assume the support condition (SC) holds. Given a trajectory functional Ψ and an associated SCF Φ, the estimator in Expression (11) is unbiased for $\mathbb{E}_{\eta^\pi}[\Psi(\tau_{0:n})]$. Further, its variance is no greater than that of the OIS estimator in Expression (10).

Having established our framework and some basic properties of the associated estimators, we now provide several examples to aid intuition.

Examples:

- By taking $\Psi(\tau_{0:n}) = \bar{G}^\pi_{0:n}$, $\Phi(\tau_{0:n}) = \tau_{0:n}$ we recover the usual OIS estimator.
- By taking $\Psi(\tau_{0:n}) = R_t$, and $\Phi(\tau_{0:n}) = (X_{0:t}, A_{0:t}, R_t)$, we recover the terms of the PDIS estimator, as described in Section 3.2.
- By taking $\Psi(\tau_{0:n}) = R_t$, and $\Phi(\tau_{0:n}) = (X_t, A_t, R_t)$, we recover terms closely related to the marginalised importance sampling estimator of Xie et al. [2018].

4.1 Orderings and optimality

Given the wide space of possible SCFs Φ for a given target Ψ encompassed by the CIS estimators in Expression (11), we now turn our attention to understanding the statistical properties of these estimators.[3]

There is a natural preorder $\precsim$ on SCFs for a given target Ψ, that specifies that for two such conditioners Φ_1 and Φ_2, we have $\Phi_1 \precsim \Phi_2$ if there exists a function h such that $\Phi_1 = h \circ \Phi_2$. The relation $\Phi_1 \precsim \Phi_2$ thus makes rigorous the notion *"all information encoded about the trajectory $\tau_{0:n}$ by $\Phi_1(\tau_{0:n})$ is also encoded by $\Phi_2(\tau_{0:n})$"*.

A second preorder that is particularly relevant to studying the statistical properties of off-policy estimators is that of having lower variance, denoted $\precsim_{\mathbb{V}}$. That is, $\Phi_1 \precsim_{\mathbb{V}} \Phi_2$ if $\mathrm{Var}(\mathbb{E}\left[\rho^{\pi,\mu}_{1:n-1} \middle| \Phi_1(\tau_{0:n})\right]) \leq \mathrm{Var}(\mathbb{E}\left[\rho^{\pi,\mu}_{1:n-1} \middle| \Phi_2(\tau_{0:n})\right])$. Note that whilst the preorder $\precsim$ is invariant to the MDP and policies π and μ in question, the variance preorder $\precsim_{\mathbb{V}}$ is not. This potentially complicates our variance analysis; however, the following proposition establishes a useful relationship between these two preorders.

Proposition 4.3. For any given MDP, and pair of policies π and μ satisfying (SC), and target functional Ψ, the variance preorder *refines* the inclusion preorder. That is, for any two SCFs Φ_1, Φ_2 of Ψ, if $\Phi_1 \precsim \Phi_2$, then we have $\Phi_1 \precsim_{\mathbb{V}} \Phi_2$.

[3]It is possible to get a slightly more streamlined analysis by working with sigma-algebras, rather than functions of the random trajectory. We restrict the exposition in the main paper to the functional perspective for accessibility and simplicity, but provide a measure-theoretic perspective in Appendix C.1.

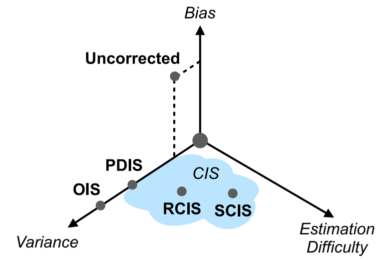

Figure 1: Schematic illustration of three traded-off quantities associated with CIS estimators.

The connection established in Proposition 4.3 will allow us to address the question of optimality: which SCFs for Ψ yield the lowest variance estimator given in Expression (11)?

Proposition 4.4. An SCF for Ψ for which the associated estimator in Expression (11) achieves minimal variance is Ψ itself.

This result gives guidance for choosing a conditioner Φ for a given target Ψ; we study several such algorithms in more detail in Section 5.

4.2 Beyond sufficient conditioning functionals

So far, we have enforced the condition that if Ψ is a target functional, a conditioner Φ used to form the conditional importance-weighted term in Expression (11) should be such that $\Psi(\tau_{0:n})$ is expressible in terms of $\Phi(\tau_{0:n})$. This condition ensures that the resulting estimator is unbiased, as shown in Proposition 4.2. However, relaxing this condition gives an even greater collection of off-policy estimators. Such estimators formed with functionals Φ which are *not* SCFs for Ψ will generally be biased, but in many circumstances may be particularly low-variance, allowing for a bias-variance trade-off to be made.

Example:

- By taking $\Psi(\tau_{0:n}) = \sum_{t=0}^{n-1} \gamma^t R_t + \gamma^n V(X_n;\pi)$, and $\Phi(\tau_{0:n}) = 0$ (i.e., a function independent of the trajectory), we recover n-step uncorrected returns, popularly used in deep reinforcement learning.

4.3 Bias, variance, and estimation difficulty

We now discuss the various trade-offs inherent within the choice of Φ required by the CIS framework. Proposition 4.2 shows that any Φ that is an SCF for Ψ yields an unbiased off-policy estimator. As described in Section 4.2, choosing Φ which is *not* an SCF for Ψ generally results in the introduction of bias, but may also offer a further substantial reduction in variance. In addition, there is the question of whether for a given Φ, the importance weight above is available analytically (as in the case of per-decision importance sampling, for example), or whether the weight itself must be estimated, as is the case for several concrete CIS algorithms, RCIS and SCIS, which we describe in Section 5. Figure 1 schematically illustrates the trade-offs between these three quantities made by several algorithms in the CIS framework.

5 Conditional importance sampling: Algorithms

Having set out the CIS framework, we now investigate several novel algorithms which naturally arise from it.

5.1 Return-conditioned importance sampling

Consider taking the n-step truncated return as our target: $\Psi(\tau_{0:n}) = \sum_{t=0}^{n-1} \gamma^t R_t$, and following the optimality result of Proposition 4.4, taking the conditioner $\Phi = \Psi$ to be this return too. This yields a conditional importance weight of the form

$$\mathbb{E}_{\eta^\mu|_{(x,a)}}\left[\rho^{\pi,\mu}_{1:n-1}\middle|\sum_{t=0}^{n-1}\gamma^t R_t\right] .$$

It is possible to express this conditional importance weight more directly, as the following result shows.

Proposition 5.1. Assume the support condition (SC). For a given policy μ let $p^\mu|_{(x,a)}$ be the probability mass function of $\sum_{t=0}^{n-1} \gamma^t R_t$ under $\eta^\mu|_{(x,a)}$. Then we have

$$\mathbb{E}_{\eta^\mu|_{(x,a)}}\left[\rho^{\pi,\mu}_{1:n-1}\middle|\sum_{t=0}^{n-1}\gamma^t R_t\right] = \frac{p^\pi|_{(x,a)}(\sum_{t=0}^{n-1}\gamma^t R_t)}{p^\mu|_{(x,a)}(\sum_{t=0}^{n-1}\gamma^t R_t)} . \tag{12}$$

That is, the optimal conditional importance weight for the n-step bootstrapped return is the ratio of the probabilities of the returns themselves under the target and behaviour distributions. This is appealing since it shifts the focus from (potentially irrelevant) policy probabilities directly to probabilities of generating a certain return value. Due to this property, we term the corresponding estimator the *return-conditioned importance sampling* (RCIS) estimator, given by:

$$\mathbb{E}_{\eta^\mu|_{(x,a)}}\left[\rho^{\pi,\mu}_{1:n-1}\middle|G\right] G + \rho^{\pi,\mu}_{1:n-1}\gamma^n V(X_n;\pi) ,$$

where $G = \sum_{t=0}^{n-1} \gamma^t R_t$. We note that several further variations of return-conditioned importance sampling are available, such as using an importance weight conditioned on the entire bootstrapped return.

There are strong connections here to distributional reinforcement learning [Morimura et al., 2010, Bellemare et al., 2017, Dabney et al., 2018], in which approximations to return distributions are learnt directly through interaction with the environment.

5.2 Reward-conditioned and state-conditioned importance sampling

The previous section establishes return-conditioned importance sampling as the optimal (with respect to estimator variance) unbiased means of importance weighting an entire return. However, this leaves open the question as to whether improvements can be made by importance weighting the individual terms of a return separately, as in per-decision importance sampling. If we interpret each reward R_t in the return $\bar{G}^\pi_{0:n}$ as a target in its own right, Proposition 4.4 shows that the corresponding optimal unbiased importance weight is

$$\mathbb{E}_{\eta^\mu|_{(x,a)}}\left[\frac{\eta^\pi_{0:n}|_{(x,a)}(\tau_{0:n})}{\eta^\mu_{0:n}|_{(x,a)}(\tau_{0:n})}\middle|R_t\right] .$$

We refer to the use of these weights as *reward-conditioned importance sampling.* Another estimator of interest that we mention due to its connections with existing off-policy evaluation algorithms and model-based reinforcement learning is given by (suboptimally) conditioning on the tuple (X_t, A_t, R_t) instead of R_t itself. In this case, we obtain the importance weight

$$\mathbb{E}_{\eta^\mu|_{(x,a)}}\left[\frac{\eta^\pi_{0:n}|_{(x,a)}(\tau_{0:n})}{\eta^\mu_{0:n}|_{(x,a)}(\tau_{0:n})}\middle|X_t, A_t, R_t\right] , \tag{13}$$

which can be shown (see Appendix A) to be equal to

$$\frac{p^\pi_t|_{(x,a)}(X_t)}{p^\mu_t|_{(x,a)}(X_t)} \times \frac{\pi(A_t|X_t)}{\mu(A_t|X_t)} , \tag{14}$$

where $p^\pi_t|_{(x,a)}$ represents the distribution over the state at time t starting at state-action pair (x, a) and following π thereafter. Thus, learning this conditional importance weight is closely related to learning the difference between the two transition models p^μ_t and p^π_t. For this reason, we refer to the use of the importance weight in Expression (13) as *state-conditioned importance sampling* (SCIS). There are close ties with the state distribution estimation methods mentioned earlier, such as marginalised importance sampling [Xie et al., 2018, 2019], which estimates a similar quantity, but by focusing on learning these transitions distributions separately, rather than their ratio directly, as well as the work of Liu et al. [2018], which learns a ratio of related distributions via a Bellman equation.

5.3 Importance weight regression

A crucial practical question about the conditional importance weights appearing in Equations (12) and (14) (and indeed in the general CIS estimator in Equation (11)), is how these should be estimated when they are not available analytically. A general approach is given by solving the following regression problem:

$$\min_\theta \mathbb{E}_{\eta^\mu|_{(x,a)}}\left[\left(f_\theta(\Phi(\tau_{0:n})) - \frac{\eta^\pi_{0:n}|_{(x,a)}(\tau_{0:n})}{\eta^\mu_{0:n}|_{(x,a)}(\tau_{0:n})}\right)^2\right]. \quad (15)$$

In words, we attempt to predict the trajectory importance weight via the function f_θ parameterised by θ, using solely the information contained in $\Phi(\tau_{0:n})$. In addition, a single regressor could be used across all initial state-action pairs, taking these quantities as additional input (i.e., $f_\theta(x, a, \Phi(\tau_{0:n}))$), and thus allowing for generalisation across actions and states. In practice, global minimisation of this objective will likely not be possible, and it may be desirable to modify the objective to take into account the magnitude of the target term $\Psi(\tau_{0:n})$ (e.g. the n-step return) to reduce the variance of the resulting approximate solution, for example. One such modified objective takes the form

$$\min_\theta \mathbb{E}_{\eta^\mu|_{(x,a)}}\left[\left(\left(f_\theta(\Phi(\tau_{0:n})) - \frac{\eta^\pi_{0:n}|_{(x,a)}(\tau_{0:n})}{\eta^\mu_{0:n}|_{(x,a)}(\tau_{0:n})}\right)\Psi(\tau_{0:n})\right)^2\right]. \quad (16)$$

The following result grounds these objectives.

Proposition 5.2. A global minimum for each of the objectives in Expressions (15) and (16) is given by

$$f_\theta(\Phi(\tau_{0:n})) = \mathbb{E}_{\eta^\mu|_{(x,a)}}\left[\frac{\eta^\pi_{0:n}|_{(x,a)}(\tau_{0:n})}{\eta^\mu_{0:n}|_{(x,a)}(\tau_{0:n})}\middle|\Phi(\tau_{0:n})\right].$$

6 Experiments

To complement the CIS framework and the theoretical analysis conducted in earlier sections, we provide several simple illustrative experiments that demonstrate (i) that CIS algorithms can deliver substantial variance reduction, and (ii) that the regression approach of Section 5.3 can be used to obtain practical implementations of CIS algorithms. We exhibit results on a classic chain environment, with both tabular and linear function approximation methods; full experiment specifications are given in Appendix B.

6.1 Operator estimation

We begin with the task of off-policy estimation of the application of the n-step Bellman operator $(T^\pi)^n$ to a fixed Q-function via trajectories generated by following the behaviour policy μ. This serves as a precursor for off-policy evaluation, and allows us to disentangle the variance reduction achieved by conditional importance sampling from compounding bootstrapping effects.

Results are shown for a chain environment in Figure 2. We plot MSE for both OIS and PDIS, as well as conditional importance sampling versions of these algorithms, RCIS and SCIS, with the conditional importance weights provided by a pre-computed *oracle*. The use of an oracle allows us to separate the variance reduction effects of conditional importance sampling from the potential errors introduced by the regression approach described in Section 5.3. In each of the four sub-plots of Figure 2, we vary one property of the estimation problem, to illustrate how performance of the methods under study changes. In all cases, we plot results for three different settings of the parameter in question, with solid lines corresponding to *low* values of this parameter, and finely-dashed lines corresponding to high values of the parameter; see Table 1. "Noise" refers to the transition noise added to the chain, β controls mismatch between the target π and behaviour μ policies, by replacing the target with a mixture $\beta\pi + (1-\beta)\mu$, and "extra actions" describes how many extra (redundant) copies of each action are added to the environment.

Table 1: Parameter values for Figure 2.

Line type	Noise	n	β	Extra actions
Solid	0%	2	0.1	0
Dashed	10%	4	0.5	1
Finely-dashed	50%	7	1.0	3

In all cases, the CIS methods outperform their existing counterparts, with more pronounced improvements in the presence of larger n, more transition noise, greater off-policyness, and high level of action redundancy.

6.2 Policy evaluation

We now consider the full task of off-policy policy evaluation using n-step returns along trajectories generated by a behaviour policy, with importance weights provided by existing and new CIS algorithms. We report results for the same chain environment as for the operator estimation experiments in Figure 3, with varying levels of transition noise and off-policyness as described in Table 1. We give results for online variants of CIS algorithms by solving the empirical version of Expression (15) (based on the observed trajectories) exactly for each different value of the functional observed; complete results including the oracle versions of the CIS algorithms are given in Appendix B.4. We show results for tabular evaluation, as well as versions using tile-coding linear function approximation [Sutton and Barto, 2018] (full details in Appendix B.3). Generally, we observe that the online versions of the CIS algorithms generally give a noticeable improvement over their non-conditional versions. These results serve as a proof of concept that practical, online versions of the

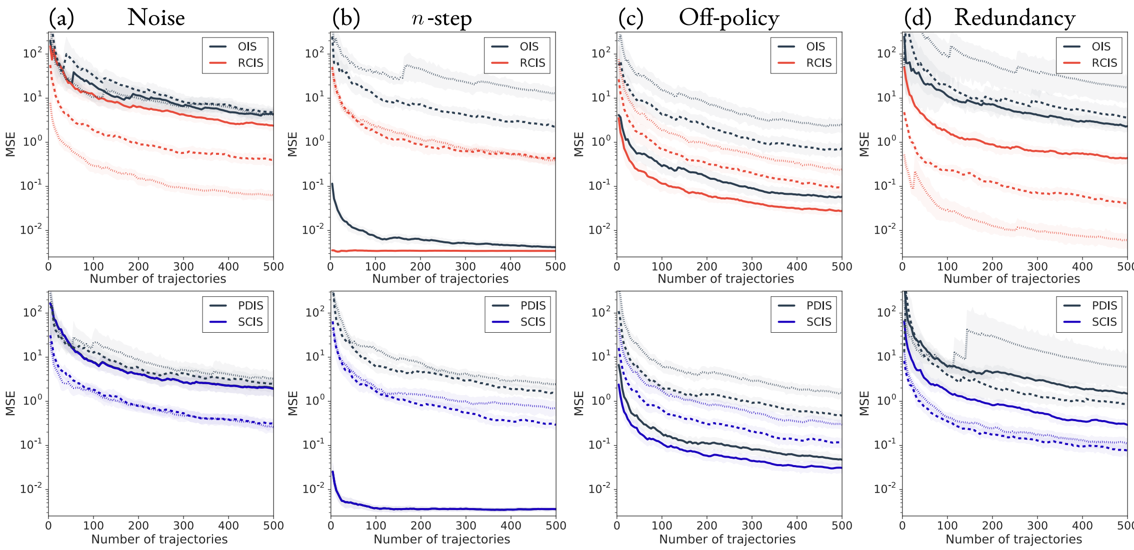

Figure 2: Operator estimation MSE as a function of sample number for OIS, PDIS, RCIS, and SCIS, on a chain MDP with varying (**a**) levels of transition noise, (**b**) n-step updates, (**c**) separation of policies, and (**d**) redundancy in action sets, as outlined in Table 1. Shaded regions indicate bootstrapped 95% confidence intervals.

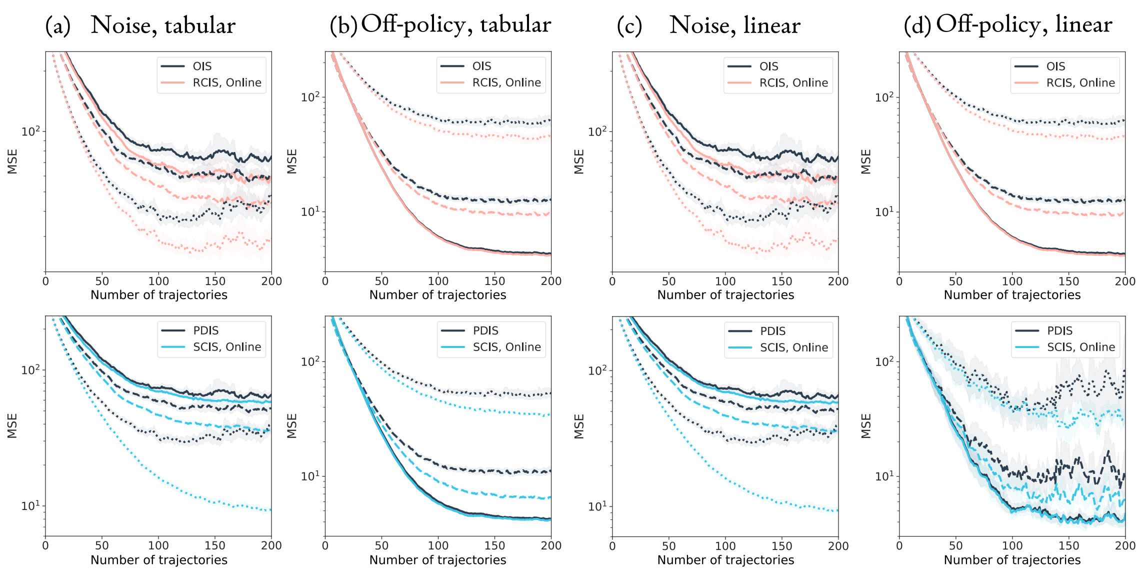

Figure 3: Policy evaluation MSE as a function of number of trajectories for OIS, RCIS, PDIS, and SCIS, with both tabular and function approximation variants. Shaded regions indicate bootstrapped 95% confidence intervals.

CIS algorithms introduced in Section 5 can improve over non-conditional baselines. We expect that with further research into regression methods described in Section 5.3, the gap between oracle and online CIS algorithms can be narrowed.

7 Discussion

We have unified several existing importance sampling algorithms via a new conceptual framework based on conditional expectations of importance weights, allowing for straightforward analysis and comparison, in addition to the development of new algorithms.

There remain many interesting investigations to be carried out towards theoretically and empirically understanding how the CIS framework interacts with complementary approaches for variance reduction, such as weighted importance sampling and importance weight truncation. We expect several further directions to prove fruitful for future work, including further exploration of the space of CIS algorithms, scaling up CIS algorithms to work in combination with deep RL architectures, and further investigation into relationships between particular CIS algorithms with other sub-fields of RL (such as RCIS and distributional RL).

Acknowledgements

We thank Adam White for detailed feedback on an earlier version of this paper, and the anonymous reviewers for helpful comments during the review process.

References

M. G. Bellemare, W. Dabney, and R. Munos. A distributional perspective on reinforcement learning. In *International Conference on Machine Learning (ICML)*, 2017.

M. G. Bellemare, W. Dabney, R. Dadashi, A. A. Taiga, P. S. Castro, N. L. Roux, D. Schuurmans, T. Lattimore, and C. Lyle. A geometric perspective on optimal representations for reinforcement learning. *Neural Information Processing Systems (NeurIPS)*, 2019.

R. Bellman. *Dynamic programming*. Princeton University Press, 1st edition, 1957.

D. P. Bertsekas and S. E. Shreve. *Stochastic Optimal Control: The Discrete-Time Case*. Athena Scientific, 2007.

D. P. Bertsekas and J. N. Tsitsiklis. *Neuro-dynamic programming*, volume 5. Athena Scientific, 1996.

P. Billingsley. *Probability and measure*. Wiley, 3rd edition, 1995.

G. Casella and R. L. Berger. *Statistical inference*. Duxbury, 2002.

W. Dabney, M. Rowland, M. G. Bellemare, and R. Munos. Distributional reinforcement learning with quantile regression. In *AAAI Conference on Artificial Intelligence*, 2018.

A. Defazio, F. Bach, and S. Lacoste-Julien. SAGA: A fast incremental gradient method with support for non-strongly convex composite objectives. In *Neural Information Processing Systems (NIPS)*, 2014.

S. S. Du, J. Chen, L. Li, L. Xiao, and D. Zhou. Stochastic variance reduction methods for policy evaluation. In *International Conference on Machine Learning (ICML)*, 2017.

L. Espeholt, H. Soyer, R. Munos, K. Simonyan, V. Mnih, T. Ward, Y. Doron, V. Firoiu, T. Harley, I. Dunning, S. Legg, and K. Kavukcuoglu. IMPALA: Scalable distributed deep-RL with importance weighted actor-learner architectures. In *International Conference on Machine Learning (ICML)*, 2018.

C. Gelada and M. G. Bellemare. Off-policy deep reinforcement learning by bootstrapping the covariate shift. In *AAAI Conference on Artificial Intelligence*, 2019.

A. Hallak and S. Mannor. Consistent on-line off-policy evaluation. In *International Conference on Machine Learning (ICML)*, 2017.

A. Hallak, A. Tamar, R. Munos, and S. Mannor. Generalized emphatic temporal difference learning: Bias-variance analysis. In *AAAI Conference on Artificial Intelligence*, 2016.

J. P. Hanna, S. Niekum, and P. Stone. Importance sampling policy evaluation with an estimated behavior policy. In *International Conference on Machine Learning (ICML)*, 2019.

M. Jaderberg, V. Mnih, W. M. Czarnecki, T. Schaul, J. Z. Leibo, D. Silver, and K. Kavukcuoglu. Reinforcement learning with unsupervised auxiliary tasks. In *International Conference on Learning Representations (ICLR)*, 2017.

N. Jiang and L. Li. Doubly robust off-policy value evaluation for reinforcement learning. In *International Conference on Machine Learning (ICML)*, 2016.

R. Johnson and T. Zhang. Accelerating stochastic gradient descent using predictive variance reduction. In *Neural Information Processing Systems (NIPS)*, 2013.

H. Kahn and T. E. Harris. Estimation of particle transmission by random sampling. In *Monte Carlo Method*, volume 12 of *Applied Mathematics Series*, pages 27–30. National Bureau of Standards, 1949.

N. Kallus and M. Uehara. Double reinforcement learning for efficient off-policy evaluation in Markov decision processes. *arXiv*, 2019a.

N. Kallus and M. Uehara. Efficiently breaking the curse of horizon: Double reinforcement learning in infinite-horizon processes. *arXiv*, 2019b.

T. P. Lillicrap, J. J. Hunt, A. Pritzel, N. Heess, T. Erez, Y. Tassa, D. Silver, and D. Wierstra. Continuous control with deep reinforcement learning. In *International Conference on Learning Representations (ICLR)*, 2016.

Q. Liu, L. Li, Z. Tang, and D. Zhou. Breaking the curse of horizon: Infinite-horizon off-policy estimation. In *Neural Information Processing Systems (NeurIPS)*, 2018.

Y. Liu, P.-L. Bacon, and E. Brunskill. Understanding the curse of horizon in off-policy evaluation via conditional importance sampling. *arXiv*, 2019.

H. R. Maei. *Gradient temporal-difference learning algorithms*. PhD thesis, University of Alberta, 2011.

A. R. Mahmood, H. van Hasselt, and R. S. Sutton. Weighted importance sampling for off-policy learning with linear function approximation. In *Neural Information Processing Systems (NIPS)*, 2014.

A. R. Mahmood, H. Yu, and R. S. Sutton. Multi-step off-policy learning without importance sampling ratios. *arXiv*, 2017.

A. M. Metelli, M. Papini, F. Faccio, and M. Restelli. Policy optimization via importance sampling. In *Neural Information Processing Systems (NeurIPS)*, 2018.

N. Metropolis and S. Ulam. The Monte Carlo method. *J. Am. Stat. Assoc.*, 44:335, 1949.

V. Mnih, K. Kavukcuoglu, D. Silver, A. A. Rusu, J. Veness, M. G. Bellemare, A. Graves, M. Riedmiller, A. K. Fidjeland, G. Ostrovski, S. Petersen, C. Beattie, A. Sadik, I. Antonoglou, H. King, D. Kumaran, D. Wierstra, S. Legg, and D. Hassabis. Human-level control through deep reinforcement learning. *Nature*, 518(7540):529–533, Feb. 2015.

T. Morimura, M. Sugiyama, H. Kashima, H. Hachiya, and T. Tanaka. Nonparametric return distribution approximation for reinforcement learning. In *International Conference on Machine Learning (ICML)*, 2010.

R. Munos, T. Stepleton, A. Harutyunyan, and M. Bellemare. Safe and efficient off-policy reinforcement learning. In *Neural Information Processing Systems (NIPS)*, 2016.

O. Nachum, Y. Chow, B. Dai, and L. Li. DualDICE: Behavior-agnostic estimation of discounted stationary distribution corrections. In *Neural Information Processing Systems (NeurIPS)*, 2019.

M. Papini, A. M. Metelli, L. Lupo, and M. Restelli. Optimistic policy optimization via multiple importance sampling. In *International Conference on Machine Learning (ICML)*, 2019.

D. Precup, R. S. Sutton, and S. P. Singh. Eligibility traces for off-policy policy evaluation. In *International Conference on Machine Learning (ICML)*, 2000.

C. Robert and G. Casella. *Monte Carlo statistical methods*. Springer, 2013.

S. Särkkä. *Bayesian filtering and smoothing*. Cambridge University Press, 2013.

T. Schaul, D. Horgan, K. Gregor, and D. Silver. Universal value function approximators. In *International Conference on Machine Learning (ICML)*, 2015.

T. Schaul, J. Quan, I. Antonoglou, and D. Silver. Priortized experience replay. In *International Conference on Learning Representations (ICLR)*, 2016.

D. Silver, G. Lever, N. Heess, T. Degris, D. Wierstra, and M. Riedmiller. Deterministic policy gradient algorithms. In *International Conference on Machine Learning (ICML)*, 2014.

R. S. Sutton and A. G. Barto. *Reinforcement Learning: An Introduction*. The MIT Press, 2nd edition, 2018.

R. S. Sutton, J. Modayil, M. Delp, T. Degris, P. M. Pilarski, A. White, and D. Precup. Horde: A scalable real-time architecture for learning knowledge from unsupervised sensorimotor interaction. In *Autonomous Agents and Multiagent Systems (AAMAS)*, 2011.

R. S. Sutton, A. R. Mahmood, and M. White. An emphatic approach to the problem of off-policy temporal-difference learning. *The Journal of Machine Learning Research*, 17(1):2603–2631, 2016.

C. Szepesvári. *Algorithms for reinforcement learning*. Morgan & Claypool Publishers, 2010.

P. Thomas and E. Brunskill. Data-efficient off-policy policy evaluation for reinforcement learning. In *International Conference on Machine Learning (ICML)*, 2016.

P. Thomas, G. Theocharous, and M. Ghavamzadeh. High confidence policy improvement. In *International Conference on Machine Learning (ICML)*, 2015a.

P. S. Thomas, G. Theocharous, and M. Ghavamzadeh. High-confidence off-policy evaluation. In *AAAI Conference on Artificial Intelligence*, 2015b.

M. Uehara and N. Jiang. Minimax weight and Q-function learning for off-policy evaluation. *arXiv*, 2019.

H. van Hasselt, A. R. Mahmood, and R. S. Sutton. Off-policy TD(λ) with a true online equivalence. In *Uncertainty in Artificial Intelligence (UAI)*, 2014.

C. Watkins and P. Dayan. Q-learning. *Machine learning*, 8(3-4):279–292, 1992.

T. Xie, Y.-X. Wang, and Y. Ma. Marginalized off-policy evaluation for reinforcement learning. In *NeurIPS Workshop on Causal Learning*, 2018.

T. Xie, Y. Ma, and Y.-X. Wang. Towards optimal off-policy evaluation for reinforcement learning with marginalized importance sampling. In *Neural Information Processing Systems (NeurIPS)*, 2019.