The Optimal Approximation Factors in Misspecified Off-Policy Value Function Estimation

Philip Amortila [1] **Nan Jiang** [1] **Csaba Szepesvári** [2]

Abstract

Theoretical guarantees in reinforcement learning (RL) are known to suffer multiplicative blow-up factors with respect to the misspecification error of function approximation. Yet, the nature of such *approximation factors*—especially their optimal form in a given learning problem—is poorly understood. In this paper we study this question in linear off-policy value function estimation, where many open questions remain. We study the approximation factor in a broad spectrum of settings, such as presence vs. absence of state aliasing and full vs. partial coverage of the state space. Our core results include instance-dependent upper bounds on the approximation factors with respect to both the weighted L_2-norm (where the weighting is the offline state distribution) and the L_∞ norm. We show that these approximation factors are optimal (in an instance-dependent sense) for a number of these settings. In other cases, we show that the instance-dependent parameters which appear in the upper bounds are necessary, and that the finiteness of either alone cannot guarantee a finite approximation factor even in the limit of infinite data.

1. Introduction

Realizability assumptions are pervasive amongst theoretical guarantees in reinforcement learning (RL) with function approximation. These assumptions posit that the true optimal solution, a value function to be estimated from data, belongs to the function class which is used. In practice, however, the realizability assumption rarely holds, and the degree to which it is violated is largely unknown. Thus, we need algorithms that do not rely on the realizability assumption in the sense that their guarantees *automatically* scale with the degree of misspecification.

When the ground truth solution is not representable by the function class, a natural relaxed objective is to instead recover the *best-in-class function* in the function class, i.e. the function which is closest to the true solution as measured by some norm. The "minimal" error incurred by the best-in-class function is called the *misspecification* error. The ratio between the error of the attained solution and that of the best-in-class solution is called the *approximation factor* or approximation ratio.

Existing error bounds for misspecified RL problems often suffer large approximation factors in addition to other statistical errors (Chen & Jiang, 2019). Unlike the statistical errors, these error terms represent the "bias" of the solution, and thus do not decrease even asymptotically as the sample size goes to infinity. It is rarely the case that attention is brought to whether these blowup factors are necessary, or if the ratios attained are optimal.

In a myriad of settings which are easier than RL (such as in linear regression or empirical risk minimization), it is indeed possible to recover an approximation factor of 1 (or arbitrarily close to 1) (Wainwright, 2019; Shalev-Shwartz & Ben-David, 2014). Whether or not similar guarantees are possible in RL problems, or what the optimal ratios would be, has been largely unstudied. Towards studying this question, we formulate an offline RL problem with linear features, and examine the optimal approximation ratio achieved by any estimator (even *asymptotic* ones). Attainable approximation factors may depend on the number of samples available, but the optimal asymptotic approximation factor is as low as it can be since even "sample-inefficient" estimators are allowed.

Concretely, our learning problem is that of linear off-policy value function estimation in infinite-horizon discounted Markov Reward Processes (MRPs). Despite the apparent simplicity of this setting, even here an understanding of the blowup factors remains open. In this problem, the learner is given access to a feature-map $\varphi : \mathcal{S} \to \mathbb{R}^d$ and an offline dataset of tuples (s_i, r_i, s_i') from the MRP. The states s_i are sampled i.i.d. from an off-policy distribution μ that may be

[1]University of Illinois, Urbana-Champaign [2]University of Alberta. Correspondence to: Philip Amortila <philipa4@illinois.edu>.

Proceedings of the 40^{th} International Conference on Machine Learning, Honolulu, Hawaii, USA. PMLR 202, 2023.

different from the stationary distribution of the MRP. We also study the *aliased* setting where the states can only be observed through their feature mapping. We do not assume anything about the off-policy distribution beyond that it yields a non-degenerate second moment matrix (defined in Section 3).[1] We also do not assume that the value function of the MRP is linear in the given feature mapping, and thus the task of the learner is simply to output the *best possible linear approximation* of the true value function (as measured by some norm). Our question is thus: *"what is the optimal asymptotic approximation factor for linear off-policy value function estimation under misspecification?"*

Recent works (Amortila et al., 2020; Perdomo et al., 2022) have provided some negative results in the realizable setting which demonstrate that the approximation factor may be arbitrarily large in the worst case. In this paper, we provide instance-dependent upper and lower bound results, with the goal of pinning down the optimal approximation ratio for off-policy value function estimation, under both the $L_2(\mu)$ norm and the L_∞ norm. For upper bounds, we analyze the well-known (off-policy) Least Squares Temporal Difference (LSTD) algorithm (Bradtke & Barto, 1996), and provide exact characterizations of its error compared to the optimal linear projection. This leads to an approximation factor for LSTD involving two problem-dependent terms, giving two "failure modes" for this algorithm. Via instance-dependent lower bounds, we show that the approximation factor attained by LSTD is optimal (up to constant factors) in a myriad of settings. In other cases, we show that both factors are necessary: the finiteness of only one of these terms cannot guarantee a finite approximation factor.[2] Our results explain the above unidentifiability results, as well as provides new ones. To our knowledge, the only prior work establishing the optimality of LSTD was in the *on-policy* setting (i.e., when μ is the stationary distribution) and held for sample sizes which were much smaller than the size of the state space (Mou et al., 2020). In particular, no prior work exists on characterizing the necessary blowup of the misspecification error in the off-policy case, even that which is asymptotically achievable. Furthermore, prior LSTD bounds are in the $L_2(\mu)$ norm only, while we also provide additional results in the maximum-norm, L_∞, a norm that allows for distribution-free error guarantees. For ease of reference, a summary of the settings that we study and their associated results can be found in Table 1.

[1] e.g. μ need not cover the entire state space or have good "concentrability" with respect to the stationary distribution

[2] Here and throughout the paper, we use "necessary" in the usual way, i.e. that the problem is intractable without these quantities *unless we make alternate assumptions or introduce other problem-dependent quantities*. See Section 6 for more discussion.

2. Problem setup

This section formalizes linear off-policy value function estimation in discounted Markov Reward Processes.

Notation We write $\mathrm{Dists}(\mathcal{X})$ to denote the set of probability distributions over a set $\mathcal{X}$. We write $I_{n\times n}$ for the $n \times n$ identity matrix, or simply I when the dimension is clear from context. For any matrix X, we let $\lambda_{\min}(X)$ and $\sigma_{\min}(X)$ denote its minimum eigenvalue (if X is square) and minimum singular value, respectively. All vectors are column vectors, and we write $^\top$ for the transpose operator.

Markov Reward Processes Markov Reward Processes arise when a fixed memoryless policy is followed in a a Markov Decision Process (Puterman, 2014; Szepesvári, 2010).

Definition 2.1 (Markov Reward Process)**.** A finite discounted Markov Reward Process (MRP) $\mathcal{M} = \langle \mathcal{S}, \mathcal{R}, \mathcal{P}, \gamma \rangle$ is defined by a finite state space $\mathcal{S} \in \mathbb{N}$, a stochastic reward function $\mathcal{R} : \mathcal{S} \to \mathrm{Dists}([-1, 1])$ with expectation $r(s) = \int x \,\mathrm{d}(\mathcal{R}(s))$, a transition function $\mathcal{P} : \mathcal{S} \to \mathrm{Dists}(\mathcal{S})$, and a discount factor $\gamma \in [0, 1)$.

To simplify the presentation, we consider finite (but arbitrarily large) state spaces. As is standard, we have assumed that the reward distribution at any state is almost-surely bounded. We will write $S \coloneqq |\mathcal{S}|$, and canonically identify $\mathcal{S} = \{1, \cdots, S\}$. We can identify r with a S-dimensional vector and $\mathcal{P}$ with the $S \times S$ row-stochastic matrix. We write $\mathcal{P}(s'|s) = [\mathcal{P}(s)](s') = P_{s,s'}$. The value function of an MRP is the following:

Definition 2.2 (Value function)**.** The value function in an MRP $\mathcal{M}$ is the function $v_{\mathcal{M}} : \mathcal{S} \mapsto [\frac{-1}{1-\gamma}, \frac{1}{1-\gamma}]$ defined by:

$$v_{\mathcal{M}}(s) = \mathbb{E}\left[\sum_{t\geq 0} \gamma^t r(S_t) \,\middle|\, S_0 := s, S_t \sim P(S_{t-1})\right].$$

In vector notation we have

$$v_{\mathcal{M}} = \sum_{t=0}^{\infty} \gamma^t P^t r = (I - \gamma P)^{-1} r,$$

which is an S-dimensional vector.

Policy evaluation with misspecified linear features A feature map $\varphi : \mathcal{S} \to \mathbb{R}^d$ is given, which the learner can use to approximate $v_{\mathcal{M}}$. The task of the learner is to output a function $f : \mathcal{S} \to \mathbb{R}$ that is linear in the features in the sense that for some $\theta \in \mathbb{R}^d$, for every $s \in \mathcal{S}$, $f(s) = \theta^\top \varphi(s)$. We write $\Phi \in \mathbb{R}^{S\times d}$ for the matrix whose s^{th} row $(s \in S)$ is $(\varphi(s))^\top$, and $\mathcal{F}_\Phi = \{f_\theta = \Phi\theta \mid \theta \in \mathbb{R}^d\} \subseteq \mathbb{R}^S$ for the subspace consisting of linear functions. The learner will be evaluated by how far the function f is from $v_{\mathcal{M}}$ in a given

	$L_2(\mu)$ norm	L_∞ norm
$\mu \geq 0$. Aliasing.	$\alpha^\star \approx \sqrt{1 + \left(\gamma \frac{\|\Pi_\mu P\|_\mu}{\sigma_{\min}(\Sigma^{-1/2} A \Sigma^{-1/2})}\right)^2}$	$\alpha^\star \approx 1 + \frac{1+\gamma}{\sigma_{\min}(A)}$
$\mu \geq 0$. No aliasing.	Upper bound: $\alpha^\star \leq \sqrt{1 + \left(\frac{\gamma \|\Pi_\mu P\|_\mu}{\sigma_{\min}(\Sigma^{-1/2} A \Sigma^{-1/2})}\right)^2}$ Lower bounds: $\|\Pi_\mu P\|_\mu = \infty$ or $\sigma_{\min}(\Sigma^{-1/2} A \Sigma^{-1/2}) = 0 \implies \alpha^\star = \infty$.	$\alpha^\star \approx 1 + \frac{1+\gamma}{\sigma_{\min}(A)}$
$\mu > 0$. Aliasing.	$\alpha^\star \approx \sqrt{1 + \left(\gamma \frac{\|\Pi_\mu P\|_\mu}{\sigma_{\min}(\Sigma^{-1/2} A \Sigma^{-1/2})}\right)^2}$	$\alpha^\star = \frac{1}{2(1-\gamma)}$
$\mu > 0$. No aliasing.	$\alpha^\star = 1$.	$\alpha^\star = 1$.

Table 1. The optimal asymptotic approximation factors $\alpha^\star$ for various settings. $\mu \geq 0$: offline distribution is arbitrary. $\mu > 0$: offline distribution has full support. Aliasing: states are only observed through feature mapping (cf. Section 2). The terms Π_μ, Σ, and A are defined in Section 3. $\approx$ indicates matching upper and lower bounds up to constants for certain parameter regimes. $=$ indicates matching upper and lower bounds.

norm, which is also available to the learner. We consider the so-called *misspecified* setting, that is, we do not assume $v_{\mathcal{M}}$ itself is a linear function of the features. Instead, the learner is only asked to produce a function whose error is not much larger than that of the *best linear approximation* of $v_{\mathcal{M}}$ in the given norm (obtained via the projection operators defined in Section 3).

Observation model We study the *offline* setting, meaning that the learner is given a dataset $\mathcal{D}_n$ from the MRP and no interaction is allowed. We will study both the *aliased* and *non-aliased* settings. In the aliased setting (Sutton & Barto, 2018), the states are only seen through the feature mapping. Formally, the observations take the form of n i.i.d. samples, which are generated by the following process

$$\varphi_i = \varphi(s_i) \text{ where } s_i \overset{\text{i.i.d.}}{\sim} \mu, \tag{1}$$
$$R_i \sim \mathcal{R}(s_i), \tag{2}$$
$$\varphi'_i = \varphi(s'_i) \text{ where } s'_i \sim \mathcal{P}(s_i). \tag{3}$$

We refer to the joint distribution over the triplets $(\varphi_i, r_i, \varphi'_i)$ as $\mathbb{Q}_{\mathcal{M},\mu,\varphi}$, and thus the dataset $\mathcal{D}_n = \{(\varphi_i, r_i, \varphi'_i)\}_{i=1}^n$ consists of n i.i.d. samples from $\mathbb{Q}_{\mathcal{M},\mu,\varphi}$.

In the *non-aliased* setting, the learner instead observes $\mathcal{D}_n^\diamond = \{(s_i, \varphi(s_i), r_i, s'_i, \varphi(s'_i))\}_{i=1}^n$, where

$$s_i \overset{\text{i.i.d.}}{\sim} \mu,\ r_i \sim \mathcal{R}(s_i),\ s'_i \sim \mathcal{P}(S_i). \tag{4}$$

We will refer to the joint distribution over non-aliased tuples $(s_i, \varphi(s_i), r_i, s'_i, \varphi(s'_i))$ as $\mathbb{Q}^\diamond_{\mathcal{M},\mu,\varphi}$. We write $\text{supp}(\mu) := \{\mu(s) > 0\} \subseteq \mathcal{S}$ for the support of μ.

We are in the *off-policy* setting, by which we mean that μ is *not* restricted to be a stationary distribution of the transition matrix P. In particular, we do not assume that μ has good "concentrability" or has support over the entire state space.

All of our upper bounds will apply to the aliased setting and thus also for the easier non-aliased setting, so we will only need to distinguish the settings when stating lower bounds. We make some minor "quality of life" assumptions about φ and μ, which are mainly for convenience. Let us write D for the diagonal matrix with the entries of μ along its diagonal (i.e. $D_{s,s} = \mu(s)$, for $s \in S$, and 0 otherwise).

Assumption 2.3 (Feature boundedness & non-degenerate second moment)**.** We have $\max_s \|\varphi(s)\|_2 \leq 1$. Furthermore, we assume that $\Sigma := \Phi^\top D \Phi = \mathbb{E}_\mu[\varphi(s)\varphi(s)^\top]$ is invertible.

Above, the L_2-boundedness of φ just provides a normalization of the features and can be assumed without loss of generality. Furthermore, if Σ is not invertible then the features are redundant; the dimensionality of the feature space can be reduced so that after the reduction Σ is invertible. Hence, this assumption can also be made without loss of generality, and we further know that it is insufficient by itself for the value prediction problem (even under realizability) (Amortila et al., 2020).

Optimal asymptotic approximation factors The quality of a finite-sample estimator is characterized by its *approximation ratio* and its *statistical error*. If, given the dataset $\mathcal{D}_n$ a learner returns the (possibly random) function $\hat{v} = \hat{v}(\mathcal{D}_n) \in \mathcal{F}_\Phi$, one often upper bounds the error of the returned function via an *oracle inequality* of the following form:

$$\|\hat{v} - v_{\mathcal{M}}\| \leq \underbrace{\alpha_n(\mathcal{M}, \mu, \varphi)}_{\text{approximation factor}} \underbrace{\inf_\theta \|\Phi\theta - v_M\|}_{\text{oracle's error}} + \underbrace{\varepsilon_n(\mathcal{M}, \mu, \varphi)}_{\text{statistical error}}, \tag{5}$$

which holds either with high probability or in expectation. The approximation factor measures the magnification of

the oracle approximation error $\inf_\theta \|\Phi\theta - v_\mathcal{M}\|$, and may be due to the imperfection of the learning algorithm or because of a fundamental hurdle that every learner faces (or both). As we will be interested in the fundamental difficulty all learners face in off-policy estimation, regardless of sample-sizes, we will consider the limit of infinite sample sizes, where the statistical error is zero. In particular, we can think of this as the case when the learner is given the distribution $\mathbb{Q}_{\mathcal{M},\mu,\varphi}$ (in the non-aliased case the distribution $\mathbb{Q}^\diamond_{\mathcal{M},\mu,\varphi}$). In the non-aliased case this is equivalent to the learner being given the model $\mathcal{P}(s)$ and $r(s)$ *for all states* $s \in \mathrm{supp}(\mu)$, and its task can be viewed as "completing" this model outside of the data distribution (using the features). A learner is a map from distributions of the above form to linear functions over $\mathcal{F}_\Phi$. The approximation ratio exhibited by a deterministic asymptotic estimator is:

$$\alpha^{\hat{v}}_{\|\cdot\|}(\mathcal{M},\mu,\varphi) = \frac{\|\hat{v}(\mathbb{Q}_{\mathcal{M},\mu,\varphi}) - v_\mathcal{M}\|}{\inf_\theta \|\Phi\theta - v_\mathcal{M}\|}, \tag{6}$$

with the convention that $\frac{0}{0} = 1$ and $\frac{x}{0} = \infty$ whenever $x > 0$. We refer to $\inf_\theta \|\Phi\theta - v_M\|$ as the *misspecification error* of the MRP $\mathcal{M}$. We do not need to consider random asymptotic estimators since, if one measures them by their expected approximation ratio, Jensen's inequality tells us that deterministic estimators are optimal.[3]

We will consider two natural choices for the norms, the weighted $L_2(\mu)$ norm and the L_∞ norm. These are defined by

$$\|v\|_\mu = \left(\sum_s \mu(s)v^2(s)\right)^{1/2} \quad \& \quad \|v\|_\infty = \max_s |v(s)|,$$

where μ is the offline state distribution from Equation (1). For any matrix $X \in \mathbb{R}^{S\times S}$, we will write $\|X\|_\mu$ for it's $L_2(\mu)$-operator norm. The $L_2(\mu)$ norm is a natural choice for function estimation as it only asks to minimize the error on states which have been encountered. In particular, for the simpler problem of linear regression (a special case of our setting for $\gamma = 0$), the least squares estimator attains the optimal approximation ratio of 1 under this norm. Meanwhile, the L_∞ norm is important for obtaining distribution-independent guarantees which we often need for RL, e.g. when value prediction is being used as a subroutine (Lagoudakis & Parr, 2003). We emphasize that our problem setting requires *function estimation* (estimating $v_\mathcal{M}$) rather than simply *return estimation* (estimating $v_\mathcal{M}$ under an initial distribution). Function estimation is a strictly more difficult problem, and there are many applications where one would require a guarantee on the error of off-policy evaluation on the whole space rather than simply at the initial states, e.g. for the aforementioned subroutines or in model selection problems (Huang & Jiang, 2022). We will write α_μ for approximation ratios in the $L_2(\mu)$ norm, and α_∞ for approximation ratios in the L_∞ norm.

[3] Since the averaged estimator $\mathbb{E}[\hat{v}]$ will be deterministic and output functions in $\mathcal{F}_\Phi$, and we have $\|\mathbb{E}[\hat{v}(\mathbb{Q}_{\mathcal{M},\mu,\varphi})] - v_\mathcal{M}\| \le \mathbb{E}[\|\hat{v}(\mathbb{Q}_{\mathcal{M},\mu,\varphi}) - v_\mathcal{M}\|]$.

3. Background

The optimal linear approximations of $v_\mathcal{M}$ are obtained by taking its projection via the projection operators.

Definition 3.1 (Projection operators)**.** We write Π_μ for the linear projection in the $L_2(\mu)$ norm, i.e. $\Pi_\mu v = \mathrm{argmin}_{\hat{v}\in\mathcal{F}_\Phi} \|\hat{v} - v\|_\mu$. This operator has a closed form,

$$\Pi_\mu = \Phi\Sigma^{-1}\Phi^\top D, \tag{7}$$

which is well-defined by Assumption 2.3. We also write Π_∞ for the linear projection in the L_∞ norm, i.e. $\Pi_\infty v \in \mathrm{argmin}_{\hat{v}\in\mathcal{F}_\Phi} \|\hat{v} - v\|_\infty$. The L_∞ projection may not be unique, and we consider that ties can be broken arbitrarily (we will not need to refer to a specific minimizer, only the value of the minimum).

One canonical estimator for the policy evaluation problem is the Least Squares Temporal Difference (LSTD) algorithm (Bradtke & Barto, 1996). In the limit of infinite samples, or *at the population level*, it is defined by the estimator

$$A := \Phi^\top D(I - \gamma P)\Phi = \mathbb{E}_{s,s'}\left[\varphi(s)(\varphi(s) - \gamma\varphi(s'))^\top\right] \tag{8}$$

$$b := \Phi^\top Dr = \mathbb{E}_{s\sim\mu}\left[\varphi(s)r(s)\right] \tag{9}$$

$$\theta_{\texttt{LSTD}} := A^{-1}b\,, \qquad v_{\texttt{LSTD}} = \Phi\theta_{\texttt{LSTD}}\,, \tag{10}$$

whenever A is invertible. In the sequel we will see that we do not need to define $\theta_{\texttt{LSTD}}$ when A is not invertible since in that case no estimator can have a finite approximation ratio. The finite-sample version of LSTD is obtained by replacing A and b by their empirical averages. We note that LSTD is applicable in the aliased setting.

4. Approximation Ratios in the $L_2(\mu)$ Norm

We begin by studying the optimal approximation factor in the $L_2(\mu)$ norm. Section 4.1 provides a general upper bound for the approximation ratio attained by LSTD and then provides a nearly-matching lower bound for the aliased setting. Section 4.2 studies whether this approximation ratio is also optimal in the non-aliased setting. The results of this section are summarized in the left column of Table 1.

4.1. Under aliasing: LSTD attains the optimal approximation factor

Our first result is a tight upper bound for the approximation factor obtained by LSTD.

Theorem 4.1. *Assume that the A matrix from Equation (8) is invertible. Then the population LSTD estimator of Equation (10) has an approximation factor upper bound of*

$$\alpha_\mu^{LSTD} \leq \sqrt{1 + \left(\gamma \|\Phi A^{-1} \Phi^\top D P\|_\mu\right)^2} \tag{11}$$

$$\leq \sqrt{1 + \left(\gamma \frac{\|\Pi_\mu P\|_\mu}{\sigma_{\min}(\Sigma^{-1/2} A \Sigma^{-1/2})}\right)^2} \tag{12}$$

Proof (sketch). This result relies on an *exact error decomposition* of the LSTD solution:

$$\Phi\theta_{\text{LS}} - \Phi\theta_{\text{LSTD}} = \gamma \Phi A^{-1} \Phi^\top D P (\Pi_\mu v_\mathcal{M} - v_\mathcal{M}), \tag{13}$$

where θ_{LS} is the least-squares parameter corresponding to the optimal solution, i.e. satisfying $\Phi\theta_{\text{LS}} = \Pi_\mu v_\mathcal{M}$. See Appendix A.1 for a full proof. □

We note that the vector $\Pi_\mu v_\mathcal{M} - v_\mathcal{M}$ is the component of the value function which is orthogonal to the features, so Equation (13) indicates that the error of LSTD is precisely dictated by action of the linear operator $\Phi A^{-1} \Phi^\top D P$ on this vector. The second upper bound in Theorem 4.1 (Equation (12)) further separates out the two terms $\|\Pi_\mu P\|_\mu$ and $\sigma_{\min}(\Sigma^{-1/2} A \Sigma^{-1/2})$. We identify these two terms as the two instance-dependent factors which control the hardness of the value function estimation problem under the $L_2(\mu)$ norm. We next give an instance-dependent lower bound which shows that for any instance values of the two parameters (within certain domains), there is a nearly-matching lower bound on the achievable asymptotic approximation factor.

Theorem 4.2. *In the aliased setting, $\forall x \in [1, \infty], \forall y \in (0, \frac{1}{2})$, there exists a collection of two instances $\mathbb{M} = \{(\mathcal{M}_1, \mu_1, \varphi_1), (\mathcal{M}_2, \mu_2, \varphi_2)\}$ which both satisfy $\|\Pi_\mu P\|_\mu = x$ and $\sigma_{\min}(\Sigma^{-1/2} A \Sigma^{-1/2}) = y$ and generate the same data distribution $\mathbb{Q}$, yet any estimator $\hat{v}$ will satisfy*

$$\sup_{(\mathcal{M},\mu,\varphi)\in\mathbb{M}} \alpha_\mu^{\hat{v}}(\mathcal{M}, \mu, \varphi) \geq \sqrt{1 + \gamma^2 \frac{\|\Pi_\mu P\|_\mu^2 - 1}{\sigma_{\min}^2(\Sigma^{-1/2} A \Sigma^{-1/2})}} \tag{14}$$

Proof (sketch). We construct two MRPS which, under aliasing, will generate the same data distribution. However, the two MRPs have different value functions and one will be realizable. In particular, the approximation ratio will be infinite if learner doesn't output that particular value function. The lower bound is obtained by calculating the error of this value function as the estimate for the first MRP. See Figure 1 for an illustration of the two MRPs, and Appendix A.2 for a full proof. □

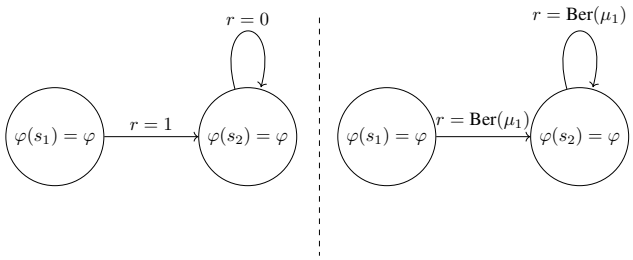


Figure 1. The construction of Theorem 4.2. Left: MRP $\mathcal{M}_1$. Right: MRP $\mathcal{M}_2$. They generate the same aliased distribution $\mathbb{Q}$.

The numerator in the second term of the lower bound is always non-negative due to the restriction on the domain of x. Furthermore, when $x > \sqrt{2}$, then the upper bound (Eq. (12)) and the lower bound (Eq. (14)) differ by at most a multiplicative factor of 2. Thus, in this regime of the instance-dependent parameters, LSTD attains the asymptotically optimal approximation ratio up to constant factors. Our domain restrictions on x and y in the lower bound also do not preclude the interesting regimes of the problem, i.e. the cases where $\|\Pi_\mu P\|_\mu$ is large ($\to \infty$) or $\sigma_{\min}(\Sigma^{-1/2} A \Sigma^{-1/2})$ is small ($\to 0$). Of course, this lower bound heavily relies on the aliased nature of the problem. Our next section examines whether the same lower bound holds in the non-aliased setting, where the learner is less restricted.

4.2. Without aliasing: what is the optimal approximation factor?

In the non-aliased case, the learner can still use the LSTD algorithm, so the upper bound of Theorem 4.1 still holds. For the lower bounds, the class of learners that we are competing against now have more information. We conjecture that the bound in Equation (12) remains optimal, but this remains open. In this work, we instead show the weaker results that both of our instance-dependent factors appearing in Equation (12) are independently necessary, meaning that the finiteness of one alone does not guarantee a finite approximation ratio.

4.2.1. $\|\Pi_\mu P\|_\mu$ IS NECESSARY

The first result of two exhibits a family of instances where $\sigma_{\min}(\Sigma^{-1/2} A \Sigma^{-1/2}) > 0$ yet the approximation ratio of any estimator is infinite. By the upper bound of Theorem 4.1, this must indicate that $\|\Pi_\mu P\|_\mu = \infty$, and indeed this is the case.

Lemma 4.3. *In the non-aliased setting, there exists a family of instances $\mathbb{M} = \{(\mathcal{M}, \mu, \varphi)\}$ which all have an $L_2(\mu)$-misspecification of 0, $\sigma_{\min}(\Sigma^{-1/2} A \Sigma^{-1/2}) > 0$,*

and $\|\Pi P\|_\mu = \infty$*, yet any estimator* $\hat{v}$ *will satisfy*

$$\sup_{(\mathcal{M},\mu,\varphi)\in\mathbb{M}} \alpha_\mu^{\hat{v}}(\mathcal{M},\mu,\varphi) = \infty$$

Proof. We take MRPs which have the same transition dynamics as those in the construction of Theorem 4.2. They have a reward $r(s_1) = 0$ and $r(s_2) = r$, and $\mu(s_1) = 1$, $\mu(s_2) = 0$. The features are arbitrary non-zero vectors. $L_2(\mu)-$realizability is trivially satisfied since only $\text{supp}(\mu) = \{s_1\}$. No estimator can recover the true value function since there is no data on state s_2. See Appendix A.3 for a full proof. □

This example illustrates the interpretation that $\|\Pi_\mu P\|_\mu$ intuitively captures the main source hardness in value function estimation. Namely, it is large (or infinite) when there is a lack of "pushforward" coverage (Xie & Jiang, 2021), meaning that a state $s \in \text{supp}(\mu)$ may transition to a state $s' \notin \text{supp}(\mu)$. Since the value at s depends on the value at s', we may not be able to predict $v_\mathcal{M}(s)$ even under realizability. Our next result shows that, surprisingly, this is not the only source of hardness in the off-policy value estimation problem.

4.2.2. $\sigma_{\min}(\Sigma^{-1/2}A\Sigma^{-1/2})$ IS ALSO NECESSARY

We next examine the case where $\|\Pi_\mu P\|_\mu$ is finite but $\sigma_{\min}(\Sigma^{-1/2}A\Sigma^{-1/2})$ is zero. This case is somewhat restricted, as in the presence of unsupported states the condition $\|\Pi_\mu P\|_\mu < \infty$ implies a strong structure on the features (cf. Lemma 4.5). Our next result demonstrates that even in the presence of this condition, one can find a set of instances where any estimator will have an infinite approximation ratio. The upper bound of Theorem 4.1 implies that $\sigma_{\min}(\Sigma^{-1/2}A\Sigma^{-1/2}) = 0$ must be the case on these instances, and indeed this is the case.

Theorem 4.4. *In the non-aliased setting, there exists a family of instances* $\{(M,\mu,\varphi)\}$ *which all have an* $L_2(\mu)$*-misspecification of* 0, $\|\Pi_\mu P\|_\mu < \infty$*, and* $\sigma_{\min}(\Sigma^{-1/2}A\Sigma^{-1/2}) = 0$*, yet any estimator* $\hat{v}$ *will satisfy*

$$\sup_{(\mathcal{M},\mu,\varphi)\in\mathbb{M}} \alpha_\mu^{\hat{v}}(\mathcal{M},\mu,\varphi) = \infty$$

Proof (sketch). We pick a 5-state MRP with 3 μ-supported states (numbered $1,2,3$) and 2 μ-unsupported states (numbered $4,5$). We set the reward to be zero except for $r(4)$ and $r(5)$ (which will be unknown to the learner). For a fixed transition matrix P, let $\mathbb{d} = (I-\gamma P)^{-1}$ denotes its discounted occupancy matrix, and $\mathbb{d}_4$ and $\mathbb{d}_5$ denote the fourth and fifth columns of this matrix, respectively. Geometrically, the space of possible value functions chosen by varying the reward function corresponds to a 2-dimensional place $\mathcal{V}_\mathcal{M} \coloneqq \{r(4)\cdot\mathbb{d}_4 + r(5)\cdot\mathbb{d}_5\}_{r_4,r_5\in[-1,1]}$. We then pick

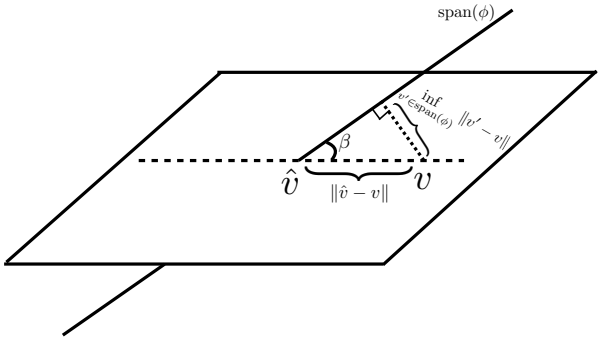


Figure 2. Illustration of proof of Theorem 4.4. The plane is a space of possible value functions $\mathcal{V}_\mathcal{M}$ which the nature can choose from without leaking more information to the learner. The line represents the space of possible linear predictors $\mathcal{F}_\Phi$. The true value function is v, and the best estimator is $\hat{v}$. As shown on the figure, the angle β determines the approximation ratio, and $\beta = 0$ (line in the plane) implies ∞ approximation ratio. In our construction, β is controlled by the magnitude of A.

a 1-dimensional feature map $\Phi = \lambda_1\mathbb{d}_4 + \lambda_2\mathbb{d}_5 \in \mathbb{R}^{5\times 1}$, which is a linear combination of the columns of $\mathbb{d}$ and thus lies in the plane. Thus, there are an infinite number of realizable value functions, and the learner cannot distinguish the correct one without knowing $r(4)$ and $r(5)$ (which occur at unsupported states). This implies that the approximation ratio is infinite. The only thing left to check is that $\|\Pi_\mu P\|_\mu < \infty$. In the presence of unsupported states, this would imply following structural condition.

Lemma 4.5. *Under Assumption 2.3,* $\|\Pi_\mu P\|_\mu < \infty$ *if and only if* $\forall s' \notin \text{supp}(\mu), \mathbb{E}_{s\sim\mu}[\varphi(s)\mathcal{P}(s'|s)] = 0$.

See Appendix A.4 for a proof of Lemma 4.5. This condition (along with the condition that $\text{supp}(\mu) = \{1,2,3\}$) turns out to be a set of bilinear condition in both μ and Φ, and we proceeded by random search to find a problem $(P,\lambda_1,\lambda_2,\mu(1),\mu(2),\mu(3))$ which satisfies this condition. See Appendix A.4 for a full description of the MRP. □

Why is $A = 0$ implied by the construction of the previous proof? While it may appear surprising that the invertibility of some algorithm-specific quantity (the A matrix) can dictate the hardness of value function estimation for *all* estimators, the intuition is that $A = 0$ implies that the linear subspace $\mathcal{F}_\Phi$ can live completely inside of the space of "plausible" value functions which the learner can not distinguish between ($\mathcal{V}_\mathcal{M}$, in the notation of our proof). More formally, when Φ is a linear combination of columns from the discounted occupancy matrix, we have that $D\Phi = D(\gamma P)\Phi$ and thus $A = 0$. In the general case where A is nonzero, its minimum singular value dictates the "angle" between the $\mathcal{F}_\Phi$ and $\mathcal{V}_\mathcal{M}$, and a small angle indicates a large approxima-

tion error (see Figure 2). In conclusion, we have showed that $\|\Pi_\mu P\| < \infty$ and $\sigma_{\min}(\Sigma^{-1/2}A\Sigma^{-1/2}) > 0$ are *both* independently necessary for finite approximation factors in value function estimation under the $L_2(\mu)$ norm.

Cases where the approximation factor is well-behaved Even though the blowup from these two factors is unavoidable in general, along the way we identify several novel conditions under which the optimal value function can be recovered, either by LSTD or by alternative estimators. These are outlined in Appendix B. Particularly interesting conditions are when P maps orthogonal value functions (i.e. functions not lying in the span of Φ) to orthogonal value functions, or when $\|P\|_\mu < \infty$ (noting that this is stronger than just $\|\Pi_\mu P\|_\mu < \infty$, since $\|\Pi_\mu P\|_\mu \le \|\Pi_\mu\|_\mu \|P\|_\mu = \|P\|_\mu$).

5. Approximation Ratios in the L_∞ Norm

In this section we study the optimal asymptotic approximation factor for the L_∞ norm. A summary of the results for this section can be found in the right column of Table 1. Recall that we write α_∞ for approximation factors in this norm.

5.1. LSTD attains the optimal approximation factor

We begin with an upper bound for LSTD. We first note that it is possible (see Appendix D) to convert an approximation ratio bound for the $L_2(\mu)$ norm into an approximation ratio bound for the L_∞ norm by paying an extra factor of $1/\lambda_{\min}(\Sigma)$ (which is finite by Assumption 2.3, but may be arbitrarily large). However, our next result shows that this eigenvalue dependence is not necessary and a more direct approach yields a better result.

Theorem 5.1. *Assume that the A matrix from Equation (8) is invertible. Then the population LSTD estimator has an approximation factor upper bound of*

$$\alpha_\infty^{LSTD} \le 1 + \left\|\Phi A^{-1}\Phi^\top D(I-\gamma P)\right\|_\infty \le 1 + \frac{1+\gamma}{\sigma_{\min}(A)}$$

Proof (sketch). Relies on another exact decomposition of the LSTD error:

$$\Pi_\infty v_\mathcal{M} - \Phi\theta_{\text{LSTD}} = \Phi A^{-1}\Phi^\top D(I-\gamma P)(\Pi_\infty v_\mathcal{M} - v_\mathcal{M}).$$

See Appendix C.1 for a full proof. □

This shows that, in the L_∞ norm, the upper bound obtained by LSTD only depends on the minimum singular value of A (rather than on the same singular value as well as $\|\Pi_\mu P\|_\mu$, as was the case for the $L_2(\mu)$ norm). At first glance it might appear strange that there are *less* problem-dependent factors in the L_∞ bound (which should be a harder norm to minimize), but the resolution to this apparent contradiction is that a guarantee of small misspecification under the L_∞ norm is a substantially stronger assumption. Intuitively, the usefulness of the $L_2(\mu)$ guarantee hinges on our ability to translate the misspecification error on μ-supported states to other parts of the state space, which additionally depends on $\|\Pi_\mu P\|_\mu$.

On the lower bound side, we can combine the ideas of the construction from (Amortila et al., 2020) and our previous lower bound (Theorem 4.4) to establish that LSTD attains the optimal approximation factor for regimes where γ is large enough. Formally, the result is that:

Theorem 5.2. *In the non-aliased setting, for all $\gamma \in [c_1, 1)$ where c_1 is some absolute constant, and for all $y \in [0, 1-\gamma]$, there exists three instances $\{(\mathcal{M}_i, \mu_i, \varphi_i)\}$ which all satisfy $\sigma_{\min}(A) = y$ yet*

$$\inf_{\hat{v}} \sup_{(\mathcal{M}_i,\mu_i,\varphi_i)} \alpha_\infty^{\hat{v}}(\mathcal{M}_i, \mu_i, \varphi_i) \ge \frac{1}{2} + \frac{\gamma}{\sigma_{\min}(A)}.$$

The value of the constant is upper bounded by $c_1 \le 0.7$.

Proof (sketch). The proof uses the MRP construction from (Amortila et al., 2020) (and Lemma 4.3) but perturbs the features by adding a column of the discounted occupancy matrix (similar to the construction of Theorem 4.4). See Appendix C.2 for a full proof. □

We note again that the domain for our problem-dependent parameters ($y \in [0, 1-\gamma]$) do not preclude the interesting regimes, which are when $\sigma_{\min}(A) \to 0$. We also note that, since the lower bound holds for the non-aliased setting, it also holds for the (harder) aliased setting. Towards comparing the upper and lower bound, we can take their ratio, use the bounds on γ and y, and observe that the ratio is always upper bounded by 2. Thus, for this regime of problem parameters, the lower bounds and upper bounds match up to a constant factor.

5.2. Optimal approximation ratio under full support

In this section, we examine a natural additional assumption which enables an alternative model-based estimator that asymptotically achieves a much better approximation ratio of $(1-\gamma)^{-1}$, which is independent of $1/\sigma_{\min}(A)$. On the other hand, this estimator will be much less sample-efficient, as its sample complexity will depend on the cardinality of $|\varphi(\mathcal{S})|$. Formally, the assumption is:

Assumption 5.3 (Full support)**.** The off-policy distribution μ is such that $\text{supp}(\mu) = \mathcal{S}$.

The full support assumption appears somewhat commonly in the literature when L_∞ norms are concerned (Huang & Jiang, 2022; Bertsekas & Tsitsiklis, 1996). We note

that Assumption 5.3 renders the problem trivial in the non-aliased setting as we can asymptotically recover the true value function $v_{\mathcal{M}}$. However, it remains an interesting question whether a similar result is possible under aliasing. Our estimator is based on state abstractions.

State abstractions We call $\varphi(\mathcal{S}) \coloneqq \mathcal{X}$ the *abstract space*, and we denote abstract states by x, x'. Note that $X \coloneqq |\mathcal{X}| \leq |\mathcal{S}|$ and in particular the abstract space is also finite. This estimator ignores the topology on $\mathcal{X}$ and instead learns a pointwise function on the abstract space. Our estimator is defined as the solution to the Bayes model $M_\varphi = (r_\varphi, P_\varphi)$, where: $r_\varphi(x) = \mathbb{E}[r(s) \mid \varphi(s) = x] \in \mathbb{R}^X$ and $P_\varphi(x, x') = \mathbb{P}(x' \mid x) \in \mathbb{R}^{X \times X}$. Note that the Bayes model implicitly depends on the off-policy distribution μ via the condition expectations. The solution to this model is

$$v_\varphi = (I - \gamma P_\varphi)^{-1} r_\varphi \in \mathbb{R}^X, \tag{15}$$

which we call the Bayes value function. The following result shows that v_φ has a well-behaved approximation ratio (see Appendix C.3 for a proof).

Theorem 5.4. *Under Assumption 5.3, the estimator v_φ from Equation (15) has an approximation ratio of $\frac{2}{1-\gamma}$, i.e. we have*

$$\begin{aligned}\|v_\varphi \circ \varphi - v_M\|_\infty &\leq \frac{2}{1-\gamma} \inf_{f : \mathcal{X} \mapsto \mathbb{R}} \|f \circ \varphi - v_{\mathcal{M}}\|_\infty \\ &\leq \frac{2}{1-\gamma} \inf_\theta \|\Phi\theta - v_{\mathcal{M}}\|_\infty\end{aligned}$$

Indeed, one can construct examples where this estimator is infinitely better than LSTD, by taking $\sigma_{\min}(A) \to 0$, which causes the LSTD parameter to diverge. This is illustrated in the counterexample of (Kolter, 2011), where LSTD diverges but our Bayes estimator achieves an approximation ratio of 1. Our next result shows that the approximation ratio $2/(1-\gamma)$ is arbitrarily close to optimal.

Theorem 5.5. *In the aliased setting, under Assumption 5.3, $\forall \varepsilon > 0$, $\forall \gamma \in (0, 1)$, there exists a collection of two instances $\mathbb{M} = \{(\mathcal{M}_1, \mu_1, \varphi_1), (\mathcal{M}_2, \mu_2, \varphi_2)\}$ which generate the same data distribution $\mathbb{Q}$, yet any estimator $\hat{v}$ will satisfy*

$$\sup_{(\mathcal{M}, \mu, \varphi) \in \mathbb{M}} \alpha_\infty^{\hat{v}}(\mathcal{M}, \mu, \varphi) \geq \frac{2}{1-\gamma} - \varepsilon$$

Proof (sketch). We use same construction as Theorem 4.2, but the error remains bounded when we are under the L_∞ norm. See Appendix C.4 for a full proof. □

It is interesting to note that this abstract model-based estimator does not work under $L_2(\mu)$ misspecification. In particular, the construction in the lower bound of Theorem 4.2 satisfies the full-support assumption (Assumption 5.3), yet the minimax $L_2(\mu)$ error can be taken to infinity by taking the "pushforward" parameter $\|\Pi_\mu P\|_\mu \approx \mu(s_1)/\mu(s_2) \to \infty$. We note that the function $v_\varphi \circ \varphi$ may not be a linear function of the features (since it is defined pointwise for each value of φ). We can output a linear function simply by taking the L_∞ projection to the set of linear functions, which results in a final bound of $1 + \frac{2}{1-\gamma}$ (see Corollary C.2 in Appendix C.5).

6. Related works

Existing negative results for off-policy evaluation With finite-horizons MRPs, (Wang et al., 2020) show that the off-policy evaluation problem has an exponential lower bound (either in d or in H, the horizon) even with realizability and good $\lambda_{\min}(\Sigma)$. This was adapted to the infinite-horizon setting by (Amortila et al., 2020) which shows that even with L_∞ realizability and good $\lambda_{\min}(\Sigma)$ the true solution may be asymptotically unidentifiable. (Perdomo et al., 2022) identify that the invertibility of A may be necessary for identifiability in the realizable setting: they show that amongst a class of "linear estimators" (which depend only on certain first-moment quantities), *any* MDP where $A = 0$ can be modified such that "linear estimators" can not recover the true value function. Their lower bound only applies to a restricted class of estimators, whereas ours rules out all estimators. In the misspecified on-policy case, (Mou et al., 2020) show that LSTD has the optimal approximation factor for restricted sample sizes n satisfying $n^2 + d \lesssim S$. In the on-policy case, the asymptotic approximation ratio is 1, so the hardness in their result comes from the sample size restriction, whereas ours comes from the non-stationarity of the off-policy distribution μ. Overall, there was no precise understanding of when this problem is solvable/not solvable (a result which is captured by our instance-dependent bounds), or which blowup is optimal in the off-policy case (even asymptotically).

Existing guarantees for LSTD The LSTD algorithm was originally proposed by (Bradtke & Barto, 1996). There have been several sample complexity analyses, (e.g. Perdomo et al. (2022); Pires & Szepesvári (2012); Tu & Recht (2018); Duan et al. (2021)). In terms of approximation ratios under misspecification, in the on-policy case, (Tsitsiklis & Van Roy, 1997) derive the classical approximation ratio bound of $(1 - \gamma^2)^{-1/2}$, which uses the fact that P (and thus $\Pi_\mu P$) are contractive in the $L_2(\mu)$ norm when μ is the stationary distribution. This bound was sharpened in an instance-dependent fashion by Yu & Bertsekas (2010); Mou et al. (2020), which both consider the more general problem of solving projected fixed point equations. Their approximation bounds are similar to our Theorem 4.1, although our proof relies on a simpler and exact error decomposition. Our proof also enables us to readily derive L_∞ bounds,

whereas only $L_2(\mu)$ bounds are considered the above works. Conversely, the work of (Perdomo et al., 2022) provides approximation bounds only in the L_∞ norm, and these sub-optimally scale with *both* $\lambda_{\min}(\Sigma)$ and $\sigma_{\min}(A)$ (similar to our result in Appendix D). The work of (Mou et al., 2020) studies the optimality of the blowup only in the on-policy setting and with restricted sample sizes. The work of (Scherrer, 2010) studies both LSTD and Bellman Residual Minimization and shows that they are both instances of oblique projections onto certain subspaces, a perspective which yields the approximation factor of $\left\|\Pi_{(I-\gamma P)^\top D\Phi}\right\|_\mu$, where $\Pi_X = \Phi(X^\top\Phi)^{-1}X^\top$ is the oblique projection operator.

OPE at large In the paper we show the quantities $\|\Pi_\mu P\|_\mu$ and $\sigma_{\min}(\Sigma^{-1/2}A\Sigma^{-1/2})$ are both necessary for the $L_2(\mu)$ norm, and that $\sigma_{\min}(A)$ is necessary for the L_∞ norm. This implies that removing the finiteness of any of these quantities leads to unbounded approximation ratio. However, our results do not exclude the possibility that one can come up with alternative assumptions/quantities to replace them. In fact, there are two sets of alternative assumptions that are widely used in the OPE literature: (1) "Bellman-completeness" (Antos et al., 2008; Munos, 2007; Chen & Jiang, 2019; Duan & Wang, 2020), which asserts that the function class is *closed* under the Bellman operator, and (2) the realizability of so-called importance weight functions (Liu et al., 2018; Uehara et al., 2020; Miyaguchi, 2021). However, most of these works focus on the estimation of the expected return at the initial state distribution instead of recovering the full function (with Huang & Jiang (2022) as an exception), and none of them study the optimality of the approximation ratio. Moreover, under these different assumptions, the definition of misspecification error changes (e.g., the violation of Bellman-completeness is sometimes referred to as "inherent Bellman error" (IBE) (Antos et al., 2008)), and so does the behavior of the approximation ratio. Under the $L_2(\mu)$ norm, small misspecification and small IBE do not imply each other, so studying the approximation ratio under these assumptions would be an interesting future direction. We can also compare the $\|\Pi_\mu P\|_\mu$ quantity, a measure of data coverage, with the more classical notion of concentrability (Antos et al., 2008).[4] However, the construction of Theorem 4.4 shows that $\|\Pi_\mu P\|_\mu$ can be small while concentrability could be infinite, so this notion is too loose for our purposes.

[4] In this setting we could for example define concentrability as $\|\rho/\mu\|_\infty$ or $\|d/\mu\|_\infty$, where ρ is the stationary distribution and d is the discounted state occupancy.

7. Conclusion

In this work we have highlighted the importance of understanding the necessary blowups to the approximation factors which occur in misspecified RL problems. We have posed a simple but fundamental learning problem, that of linear off-policy value function estimation, and focused on establishing the optimal approximation ratios for this problem achieved even by asymptotic estimators. We have provided instance-dependent upper and lowers bounds for a variety of settings (the $L_2(\mu)$ and L_∞ norms, aliased and non-aliased observations, partial support and full-support) which established the optimal algorithms and ratios for certain of these settings. In the other settings, it was shown that LSTD is a fundamental algorithm in the sense that whenever it has an infinite error then so does every other estimator.

For future work, it would be fruitful to understand the general lower bound for the non-aliased $L_2(\mu)$ case (Section 4.2). In sections 4.1, 5.1, and 5.2, we have been able to provide matching upper and lower bounds (potentially up to constant factors) for the optimal asymptotic factors. By contrast, our results for the non-aliased $L_2(\mu)$ setting only proved that our instance-dependent quantities were *necessary* (without being able to establish the precise scaling of the bounds).

While we have only been concerned with policy evaluation, it would also be important to consider misspecification in the more complicated policy optimization (or online exploration) problems. Here, there are even more hardness results which preclude a simple answer to this question (Lattimore et al., 2020; Du et al., 2019; Weisz et al., 2021; Foster et al., 2021).

Acknowledgements

PA gratefully acknowledges funding from the Natural Sciences and Engineering Research Council (NSERC) of Canada. NJ acknowledges funding support from NSF IIS-2112471 and NSF CAREER IIS-2141781. CS gratefully acknowledges funding from NSERC and the Canada CIFAR AI Chairs Program through Amii.

References

Amortila, P., Jiang, N., and Xie, T. A variant of the wang-foster-kakade lower bound for the discounted setting. *arXiv preprint arXiv:2011.01075*, 2020.

Antos, A., Szepesvári, C., and Munos, R. Learning near-optimal policies with bellman-residual minimization based fitted policy iteration and a single sample path. *Machine Learning*, 71:89–129, 2008.

Bertsekas, D. and Tsitsiklis, J. N. *Neuro-dynamic program-*

ming. Athena Scientific, 1996.

Bradtke, S. J. and Barto, A. G. Linear least-squares algorithms for temporal difference learning. *Machine learning*, 22(1):33–57, 1996.

Chen, J. and Jiang, N. Information-theoretic considerations in batch reinforcement learning. In *International Conference on Machine Learning*, pp. 1042–1051. PMLR, 2019.

Du, S. S., Kakade, S. M., Wang, R., and Yang, L. F. Is a good representation sufficient for sample efficient reinforcement learning? *arXiv preprint arXiv:1910.03016*, 2019.

Duan, Y. and Wang, M. Minimax-Optimal Off-Policy Evaluation with Linear Function Approximation. *arXiv:2002.09516 [cs, math, stat]*, February 2020. URL http://arxiv.org/abs/2002.09516. arXiv: 2002.09516.

Duan, Y., Wang, M., and Wainwright, M. J. Optimal policy evaluation using kernel-based temporal difference methods. *arXiv:2109.12002 [cs, math, stat]*, September 2021. URL http://arxiv.org/abs/2109.12002. arXiv: 2109.12002.

Foster, D. J., Krishnamurthy, A., Simchi-Levi, D., and Xu, Y. Offline reinforcement learning: Fundamental barriers for value function approximation. *arXiv preprint arXiv:2111.10919*, 2021.

Huang, A. and Jiang, N. Beyond the return: Off-policy function estimation under user-specified error-measuring distributions. In *Advances in Neural Information Processing Systems*, 2022.

Jiang, N. Notes on state abstractions, 2018.

Kolter, J. The fixed points of off-policy td. *Advances in Neural Information Processing Systems*, 24, 2011.

Lagoudakis, M. G. and Parr, R. Least-squares policy iteration. *The Journal of Machine Learning Research*, 4: 1107–1149, 2003.

Lattimore, T., Szepesvari, C., and Weisz, G. Learning with good feature representations in bandits and in rl with a generative model. In *International Conference on Machine Learning*, pp. 5662–5670. PMLR, 2020.

Li, L., Walsh, T. J., and Littman, M. L. Towards a unified theory of state abstraction for mdps. In *AI&M*, 2006.

Liu, Q., Li, L., Tang, Z., and Zhou, D. Breaking the curse of horizon: Infinite-horizon off-policy estimation. *Advances in Neural Information Processing Systems*, 31, 2018.

Miyaguchi, K. Asymptotically exact error characterization of offline policy evaluation with misspecified linear models. *Advances in Neural Information Processing Systems*, 34:28573–28584, 2021.

Mou, W., Pananjady, A., and Wainwright, M. J. Optimal oracle inequalities for solving projected fixed-point equations. *arXiv:2012.05299 [cs, math, stat]*, December 2020. URL http://arxiv.org/abs/2012.05299. arXiv: 2012.05299.

Munos, R. Performance bounds in l_p-norm for approximate value iteration. *SIAM journal on control and optimization*, 46(2):541–561, 2007.

Perdomo, J. C., Krishnamurthy, A., Bartlett, P., and Kakade, S. A Sharp Characterization of Linear Estimators for Offline Policy Evaluation. *arXiv:2203.04236 [cs, stat]*, March 2022. URL http://arxiv.org/abs/2203.04236. arXiv: 2203.04236.

Pires, B. Á. and Szepesvári, C. Statistical linear estimation with penalized estimators: an application to reinforcement learning. In *Proceedings of the 29th International Coference on International Conference on Machine Learning*, pp. 1755–1762, 2012.

Puterman, M. L. *Markov decision processes: discrete stochastic dynamic programming*. John Wiley & Sons, 2014.

Scherrer, B. Should one compute the temporal difference fix point or minimize the bellman residual? the unified oblique projection view. *arXiv preprint arXiv:1011.4362*, 2010.

Shalev-Shwartz, S. and Ben-David, S. *Understanding machine learning: From theory to algorithms*. Cambridge university press, 2014.

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

Szepesvári, C. Algorithms for reinforcement learning. *Synthesis lectures on artificial intelligence and machine learning*, 4(1):1–103, 2010.

Tsitsiklis, J. and Van Roy, B. An analysis of temporal-difference learning with function approximation. *IEEE Transactions on Automatic Control*, 42(5):674–690, May 1997. ISSN 00189286. doi: 10.1109/9.580874. URL http://ieeexplore.ieee.org/document/580874/.

Tu, S. and Recht, B. Least-squares temporal difference learning for the linear quadratic regulator. In *International Conference on Machine Learning*, pp. 5005–5014. PMLR, 2018.

Uehara, M., Huang, J., and Jiang, N. Minimax Weight and Q-Function Learning for Off-Policy Evaluation. *arXiv:1910.12809 [cs, stat]*, October 2020. URL http://arxiv.org/abs/1910.12809. arXiv: 1910.12809.

Wainwright, M. J. *High-dimensional statistics: A non-asymptotic viewpoint*, volume 48. Cambridge university press, 2019.

Wang, R., Foster, D. P., and Kakade, S. M. What are the statistical limits of offline rl with linear function approximation? *arXiv preprint arXiv:2010.11895*, 2020.

Weisz, G., Amortila, P., and Szepesvári, C. Exponential lower bounds for planning in mdps with linearly-realizable optimal action-value functions. In *Algorithmic Learning Theory*, pp. 1237–1264. PMLR, 2021.

Xie, T. and Jiang, N. Batch value-function approximation with only realizability. In *International Conference on Machine Learning*, pp. 11404–11413. PMLR, 2021.

Yu, H. and Bertsekas, D. P. Error Bounds for Approximations from Projected Linear Equations. *Mathematics of Operations Research*, 35(2):306–329, May 2010. ISSN 0364-765X, 1526-5471. doi: 10.1287/moor.1100.0441. URL http://pubsonline.informs.org/doi/abs/10.1287/moor.1100.0441.

A. Proofs for Section 4

A.1. Proof of Theorem 4.1

Theorem 4.1. *Assume that the A matrix from Equation (8) is invertible. Then the population LSTD estimator of Equation (10) has an approximation factor upper bound of*

$$\alpha_\mu^{LSTD} \leq \sqrt{1 + \left(\gamma \|\Phi A^{-1} \Phi^\top D P\|_\mu\right)^2} \tag{11}$$

$$\leq \sqrt{1 + \left(\gamma \frac{\|\Pi_\mu P\|_\mu}{\sigma_{\min}(\Sigma^{-1/2} A \Sigma^{-1/2})}\right)^2} \tag{12}$$

The proof follows from the following exact characterization of LSTD.

Lemma A.1. *If A^{-1} exists, then we have that*

$$\theta_{LS} - \theta_{LSTD} = \gamma A^{-1} \Phi^\top D P v^\perp,$$

where $v^\perp = v_\mathcal{M} - \Pi_\mu v_\mathcal{M}$.

Proof. We have that $v_\mathcal{M} = \Pi_\mu v_\mathcal{M} + v^\perp$, with $\Pi_\mu v \in \text{col}(\Phi)$ and $v^\perp \in (\text{col}(\Phi))^\perp$ (note: the $\perp$ subspace is with respect to the μ-weighted inner product). This equivalently means that $v^\perp \in \ker(\Phi^\top D)$. Then we have:

$$\begin{aligned}
(I - \gamma P) v_M &= r \\
(I - \gamma P)\Phi\theta_{\text{LS}} + (I - \gamma P) v^\perp &= r \\
\Phi^\top D (I - \gamma P)\Phi\theta_{\text{LS}} + \Phi^\top D (I - \gamma P) v^\perp &= \Phi^\top D r && (\Phi^\top D \text{ on both sides}) \\
\left(\Phi^\top D (\Phi - \gamma P \Phi)\right) \theta_{\text{LS}} - \gamma \Phi^\top D P v^\perp &= \Phi^\top D r && (v^\perp \in \ker(\Phi^\top D)) \\
A\theta_{\text{LS}} &= b + \gamma \Phi^\top D P v^\perp && (\text{Defns of } A, b) \\
\theta_{\text{LS}} &= A^{-1} b + \gamma A^{-1} \Phi^\top D P v^\perp && (A^{-1} \text{ exists})
\end{aligned}$$

Meanwhile, the LSTD solution is defined by $\theta_{\text{LSTD}} = A^{-1} b$. Substracting both of these gives:

$$\theta_{\text{LS}} - \theta_{\text{LSTD}} = \gamma A^{-1} \Phi^\top D P v^\perp. \tag{16}$$

Note that we also have:

$$\theta_{\text{LS}} - \theta_{\text{LSTD}} = -A^{-1} \Phi^\top D (I - \gamma P) v^\perp, \tag{17}$$

since $\Phi^\top D v^\perp = 0$.

□

Corollary A.2. *Let $v_{LSTD} = \Phi\theta_{LSTD}$ and $v_{LS} = \Phi\theta_{LS}$. Then we have the two inequalities*

$$\|\Phi\theta_{LSTD} - \Phi\theta_{LS}\|_\mu = \gamma \left\|\Phi A^{-1} \Phi^\top D P v^\perp\right\|_\mu \leq \gamma \left\|\Phi A^{-1} \Phi^\top D P\right\|_\mu \left\|v^\perp\right\|_\mu$$

and

$$\|v_{LSTD} - v_{LS}\|_\mu \leq \frac{\gamma}{\sigma_{\min}(I - \gamma \Sigma^{-1/2} \Sigma_{cr} \Sigma^{-1/2})} \left\|\Pi_\mu P v^\perp\right\|_\mu \leq \frac{\gamma}{\sigma_{\min}(\Sigma^{-1/2} A \Sigma^{-1/2})} \|\Pi_\mu P\|_\mu \left\|v^\perp\right\|_\mu, \tag{18}$$

where $\Sigma = \Phi^\top D \Phi = \mathbb{E}_\mu[\varphi(s)\varphi(s)^\top]$ is the covariance matrix and $\Sigma_{cr} = \Phi^\top D P \Phi = \mathbb{E}_{\mu,P}[\varphi(s)\varphi(s')^\top]$ is the cross-covariance.

Proof. The first equality follows from

$$\|\Phi\theta_{\text{LSTD}} - \Phi\theta_{\text{LS}}\|_\mu = \gamma\left\|\Phi A^{-1}\Phi^\top DPv^\perp\right\|_\mu \le \gamma\left\|\Phi A^{-1}\Phi^\top DP\right\|_\mu \left\|v^\perp\right\|_\mu$$

The second equality follows from:

$$\begin{aligned}
\|\Phi\theta_{\text{LSTD}} - \Phi\theta_{\text{LS}}\|_\mu &= \left\|\Sigma^{1/2}(\theta_{\text{LSTD}} - \theta_{\text{LS}})\right\|_2 \\
&= \gamma\left\|\Sigma^{1/2}A^{-1}\Phi^\top DPv^\perp\right\|_2 \\
&= \gamma\left\|(I - \gamma\Sigma^{-1/2}\Sigma_{\text{cr}}\Sigma^{-1/2})^{-1}\Sigma^{-1/2}\Phi^\top DPv^\perp\right\|_2 \\
&\le \frac{\gamma}{\sigma_{\min}(I - \gamma\Sigma^{-1/2}\Sigma_{\text{cr}}\Sigma^{-1/2})}\left\|\Sigma^{-1/2}\Phi^\top DPv^\perp\right\|_2 \\
&= \frac{\gamma}{\sigma_{\min}(I - \gamma\Sigma^{-1/2}\Sigma_{\text{cr}}\Sigma^{-1/2})}\left\|\Sigma^{1/2}\Sigma^{-1}\Phi^\top DPv^\perp\right\|_2 \\
&= \frac{\gamma}{\sigma_{\min}(I - \gamma\Sigma^{-1/2}\Sigma_{\text{cr}}\Sigma^{-1/2})}\left\|\Sigma^{1/2}(\Phi^\top D\Phi)^{-1}\Phi^\top DPv^\perp\right\|_2 \\
&= \frac{\gamma}{\sigma_{\min}(I - \gamma\Sigma^{-1/2}\Sigma_{\text{cr}}\Sigma^{-1/2})}\left\|\Phi(\Phi^\top D\Phi)^{-1}\Phi^\top DPv^\perp\right\|_\mu \\
&= \frac{\gamma}{\sigma_{\min}(I - \gamma\Sigma^{-1/2}\Sigma_{\text{cr}}\Sigma^{-1/2})}\left\|\Pi_\mu Pv^\perp\right\|_\mu \qquad (19)
\end{aligned}$$

And then note that $I - \gamma\Sigma^{-1/2}\Sigma_{\text{cr}}\Sigma^{-1/2} = \Sigma^{-1/2}(\Sigma - \gamma\Sigma_{\text{cr}})\Sigma^{-1/2} = \Sigma^{-1/2}A\Sigma^{-1/2}$. □

To conclude the proof of Theorem 4.1, we can use the Pythagorean theorem on $v_{\text{LSTD}} - \Pi_\mu v_{\mathcal{M}} \in \text{col}(\Phi)$ and $\Pi_\mu v_{\mathcal{M}} - v_{\mathcal{M}} \in (\text{col}(\Phi))^\perp$:

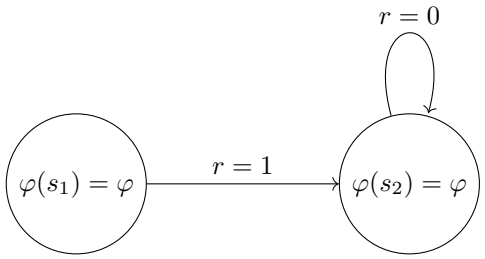
$$\begin{aligned}
\|v_{\text{LSTD}} - v_{\mathcal{M}}\|_\mu &= \sqrt{\|\Pi_\mu v_{\mathcal{M}} - v_{\mathcal{M}}\|_\mu^2 + \|v_{\text{LSTD}} - \Pi_\mu v_{\mathcal{M}}\|_\mu^2} \\
&\le \sqrt{\|\Pi_\mu v_{\mathcal{M}} - v_{\mathcal{M}}\|_\mu^2 + \left(\gamma\|\Phi A^{-1}\Phi^\top DP\|_\mu\right)^2\|\Pi_\mu v_{\mathcal{M}} - v_{\mathcal{M}}\|_\mu^2} \\
&= \sqrt{1 + \left(\gamma\|\Phi A^{-1}\Phi^\top DP\|_\mu\right)^2}\|\Pi_\mu v_{\mathcal{M}} - v_{\mathcal{M}}\|_\mu \\
&\le \sqrt{1 + \left(\gamma\frac{\|\Pi_\mu P\|_\mu}{\sigma_{\min}(\Sigma^{-1/2}A\Sigma^{-1/2})}\right)^2}\|\Pi_\mu v_{\mathcal{M}} - v_{\mathcal{M}}\|_\mu \\
&\le \left(1 + \gamma\frac{\|\Pi_\mu P\|_\mu}{\sigma_{\min}(\Sigma^{-1/2}A\Sigma^{-1/2})}\right)\|\Pi_\mu v_{\mathcal{M}} - v_{\mathcal{M}}\|_\mu \qquad (\sqrt{1+x^2} \le 1 + x \text{ whenever } x \ge 0)
\end{aligned}$$

A.2. Proof of Theorem 4.2

Theorem 4.2. *In the aliased setting, $\forall x \in [1, \infty], \forall y \in (0, \frac{1}{2})$, there exists a collection of two instances $\mathbb{M} = \{(\mathcal{M}_1, \mu_1, \varphi_1), (\mathcal{M}_2, \mu_2, \varphi_2)\}$ which both satisfy $\|\Pi_\mu P\|_\mu = x$ and $\sigma_{\min}(\Sigma^{-1/2}A\Sigma^{-1/2}) = y$ and generate the same data distribution $\mathbb{Q}$, yet any estimator $\hat{v}$ will satisfy*

$$\sup_{(\mathcal{M},\mu,\varphi)\in\mathbb{M}} \alpha_\mu^{\hat{v}}(\mathcal{M}, \mu, \varphi) \ge \sqrt{1 + \gamma^2\frac{\|\Pi_\mu P\|_\mu^2 - 1}{\sigma_{\min}^2(\Sigma^{-1/2}A\Sigma^{-1/2})}} \qquad (14)$$

Proof. Our first instance is $(\mathcal{M}_1, \mu, \varphi)$ defined via



We set $\varphi = 1 \in \mathbb{R}^1$, and define the shorthands $\mu_1 = \mu(s_1)$, $\mu_2 = \mu(s_2)$. The values of γ and μ_1, μ_2 will be picked to ensure that $\sigma_{\min}(\Sigma^{-1/2} A \Sigma^{-1/2}) = y$ and $\|\Pi_\mu P\|_\mu = x$.

By taking the derivative of $\mu_1(\theta - 1)^2 + \mu_2(\theta)^2$ wrt θ, the optimal estimator is $\theta_1 = \theta\varphi = \mu_1$. It has a square error:

$$\|\theta_1 \varphi(s) - v_{\mathcal{M}_1}\|_\mu^2 = \mu_1(\mu_1 - 1)^2 + \mu_2(\mu_1)^2 = \mu_1 - \mu_1^2 = \mu_1\mu_2$$

Note that this instance has $\mathbb{P} = \begin{pmatrix} 0 & 1 \\ 0 & 1 \end{pmatrix}$, $\Sigma^{-1} = 1$, and $\Pi_\mu = \Phi\Phi^\top D = (1,1)(1,1)^\top D = \begin{pmatrix} 1 & 1 \\ 1 & 1 \end{pmatrix} \begin{pmatrix} \mu_1 & 0 \\ 0 & \mu_2 \end{pmatrix} = \begin{pmatrix} \mu_1 & \mu_2 \\ \mu_1 & \mu_2 \end{pmatrix}$. This gives $\Pi_\mu P = \begin{pmatrix} \mu_1 & \mu_2 \\ \mu_1 & \mu_2 \end{pmatrix} \begin{pmatrix} 0 & 1 \\ 0 & 1 \end{pmatrix} = \begin{pmatrix} 0 & 1 \\ 0 & 1 \end{pmatrix}$. The operator norm thus has a value

$$\max_{\|v\|_\mu = 1} \|\Pi_\mu P v\|_\mu = \max_{\|v\|_\mu = 1} \sqrt{\mu_1 v_2^2 + \mu_2 v_2^2},$$

which is maximized by taking $v_1 = 0, v_2 = 1/\sqrt{\mu_2}$, giving a value of

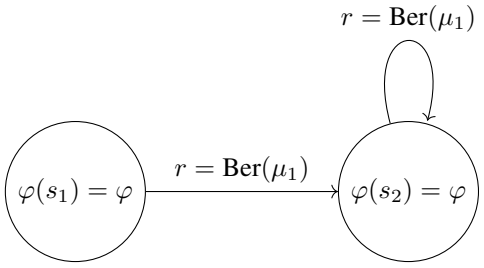
$$\|\Pi_\mu P\|_\mu = \sqrt{\frac{\mu_1}{\mu_2} + 1}.$$

Note that $\|\Pi_\mu P\|_\mu \in [1, \infty]$. We need this to equal x which is easily achieved by solving $1 + \frac{\mu_1}{1-\mu_1} = x^2 \implies \mu_1 = \frac{x^2 - 1}{x^2}$ which lies inside $(0, 1)$ for all $x \in (1, \infty)$. The cases where $x = 1$ or $x = \infty$ are handled by picking $\mu_1 = 1$ or $\mu_1 = 0$, respectively. Meanwhile we also have that

$$A = \varphi^2 - \gamma\varphi^2 = \varphi^2(1 - \gamma) = (1 - \gamma),$$

and $\Sigma^{-1/2} A \Sigma^{-1/2} = A$. We need $A = y$, which is achieved by picking $\gamma = 1 - y$. Note the restriction on the domain of $y \in (0, 1/2)$ means that $1/2 < \gamma < 1$.

The second MRP is the one defined as:



where again $\varphi = 1 \in \mathbb{R}^1$. We take μ_1 and μ_2 to be the same as in the first MRP. This instance also has $A = \varphi^2(1 - \gamma) = (1 - \gamma)$ and $\|\Pi_\mu P\|_\mu = 1 + \frac{\mu_1}{\mu_2}$, which is easily seen since the features and the transition dynamics are the same. Further note that these two MRPs generate the same aliased distribution $\mathbb{Q}_{\mathcal{M},\mu,\varphi}$ since they both generate $(\varphi, 0, \varphi)$ with probability $1 - \mu_1$ and $(\varphi, 1, \varphi)$ with probability μ_1.

The optimal estimator for $\mathcal{M}_2$ is evidently $\theta_2 = \varphi\theta = \mu_1/(1 - \gamma)$, since $v_{\mathcal{M}_2}(s_1) = v_{\mathcal{M}_2}(s_2) = \mu_1/(1 - \gamma)$. In particular, this second MRP is realizable so this forces the estimator to pick $\mu_1/(1 - \gamma)$ when faced against these two examples (μ_1 is

known by looking at the occurrence of the triples $(\varphi, 1, \varphi)$ in $\mathbb{Q}_{\mathcal{M},\mu,\varphi}$). And other choice of estimator will in fact have a worst-case approximation ratio $\sup_{\mathcal{M}\in\{\mathcal{M}_1,\mathcal{M}_2\}} \alpha = \infty$. On the first instance, this estimator will have a squared error

$$\begin{aligned}\|\theta_2\varphi(s) - v_{\mathcal{M}_1}\|_\mu^2 = \mu_1(\frac{\mu_1}{1-\gamma} - 1)^2 + \mu_2(\frac{\mu_1}{1-\gamma})^2 &= \mu_1((\frac{\mu_1}{1-\gamma})^2 - 2\frac{\mu_1}{1-\gamma} + 1) + \mu_2(\frac{\mu_1}{1-\gamma})^2 \\ &= (\frac{\mu_1}{1-\gamma})^2(\mu_1+\mu_2) - 2\frac{\mu_1^2}{1-\gamma} + \mu_1 \\ &= (\frac{\mu_1}{1-\gamma})^2 - 2\frac{\mu_1^2}{1-\gamma} + \mu_1\end{aligned}$$

Taking the ratio of squared errors gives:

$$\begin{aligned}\frac{\|\theta_2\varphi(s) - v_{\mathcal{M}_1}\|_\mu^2}{\|\theta_1\varphi(s) - v_{\mathcal{M}_1}\|_\mu^2} = \frac{(\frac{\mu_1}{1-\gamma})^2 - 2\frac{\mu_1^2}{1-\gamma} + \mu_1}{\mu_1 - \mu_1^2} &= \frac{\frac{\mu_1}{(1-\gamma)^2} - 2\frac{\mu_1}{1-\gamma} + 1}{1-\mu_1} \\ &= \frac{\frac{\mu_1}{(1-\gamma)^2} - 2\frac{\mu_1}{1-\gamma} + 1}{\mu_2} \\ &\geq 1 + \frac{\mu_1}{\mu_2}\left(\frac{2\gamma-1}{(1-\gamma)^2}\right) && (1/\mu_2 \geq 1, \text{ and algebra}) \\ &\geq 1 + \frac{\mu_1}{\mu_2}\left(\frac{1}{(1-\gamma)^2}\right) && (\tfrac{1}{2} \leq \gamma \leq 1) \\ &= 1 + \frac{\|\Pi_\mu P\|_\mu^2 - 1}{\sigma_{\min}(\Sigma^{-1/2}A\Sigma^{-1/2})^2} \\ &\geq 1 + \gamma^2\frac{\|\Pi_\mu P\|_\mu^2 - 1}{\sigma_{\min}(\Sigma^{-1/2}A\Sigma^{-1/2})^2}\end{aligned}$$

The LHS was the ratio of squared errors, so taking square roots gives α_μ and the desired bound.

□

A.3. Proof of Theorem 4.3

Lemma 4.3. *In the non-aliased setting, there exists a family of instances $\mathbb{M} = \{(\mathcal{M}, \mu, \varphi)\}$ which all have an $L_2(\mu)$-misspecification of 0, $\sigma_{\min}(\Sigma^{-1/2}A\Sigma^{-1/2}) > 0$, and $\|\Pi P\|_\mu = \infty$, yet any estimator $\hat{v}$ will satisfy*

$$\sup_{(\mathcal{M},\mu,\varphi)\in\mathbb{M}} \alpha_\mu^{\hat{v}}(\mathcal{M}, \mu, \varphi) = \infty$$

Proof. This example is a slight modification of the two-state example (Amortila et al., 2020). See Figure A.3.

In the construction, we have $\mu(s_A) = 1$ and $\mu(s_B) = 0$. Note that $A = -\gamma^2\varepsilon \neq 0$. Since the reward at state s_B is never observed, any estimator will have constant error $\Omega(1/(1-\gamma))$ error asymptotically. Since $v_\mathcal{M}$ is realizable on s_A (with $\theta = r/(1-\gamma)$), the misspecification is 0. Thus, the approximation ratio is infinite. The last thing to show is that $\|\Pi_\mu P\|_\mu = \infty$. This is because $P = \begin{pmatrix} 0 & 1 \\ 0 & 1 \end{pmatrix}$, and $\Pi_\mu = \begin{pmatrix} 1 & 0 \\ (1+\varepsilon)/\gamma & 0 \end{pmatrix}$ so $\Pi_\mu P = \begin{pmatrix} 0 & 1 \\ 0 & (1+\varepsilon)/\gamma \end{pmatrix}$ thus $\|\Pi_\mu P\|_\mu = \max_{\|v\|_\mu=1} \|\Pi_\mu P v\| = \max_{\|v\|_\mu=1} \left\|(v_2, \frac{1+\varepsilon}{\gamma}v_2)^\top\right\|_\mu = \max_{\|v\|_\mu=1} v_2 = \infty$. In the last step we can take $v_2 \to \infty$ in the maximization since that state is unsupported. □

A.4. Proof of Theorem 4.4

Block matrix notation In this section we will use the following convenient block matrix notation. Noting that $|\operatorname{supp}(\mu)| \leq S$ (with equality iff μ has support on all the states), we will re-arrange the states such that those that are supported

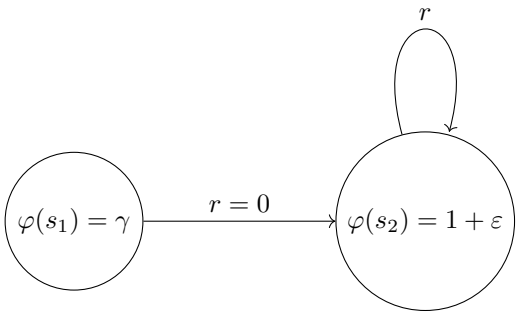


Figure 3. The construction of Lemma 4.3

are numbered $1 \dots |\operatorname{supp}(\mu)|$, and the unsupported ones are numbered $|\operatorname{supp}(\mu)| + 1 \dots S$. Furthermore, for a given vector $v \in \mathbb{R}^S$, we will write $v_\mu = (v_1, \dots, v_{|\operatorname{supp}(\mu)|})$ for the restriction of v to the support states of $\mathcal{S}$, and $v_{\neg\mu} = (v_{|\operatorname{supp}(\mu)|+1}, \dots, v_S)$ for the restriction of v to the unsupported states. Similarly, for a given matrix X, we will write it in block form as

$$X = \left[\begin{array}{c|c} X_{\mu,\mu} & X_{\mu,\neg\mu} \\ \hline X_{\neg\mu,\mu} & X_{\neg\mu,\neg\mu} \end{array}\right]$$

Theorem 4.4. *In the non-aliased setting, there exists a family of instances $\{(M, \mu, \varphi)\}$ which all have an $L_2(\mu)$-misspecification of 0, $\|\Pi_\mu P\|_\mu < \infty$, and $\sigma_{\min}(\Sigma^{-1/2} A \Sigma^{-1/2}) = 0$, yet any estimator $\hat{v}$ will satisfy*

$$\sup_{(\mathcal{M},\mu,\varphi)\in\mathbb{M}} \alpha_\mu^{\hat{v}}(\mathcal{M}, \mu, \varphi) = \infty$$

We start by noting the following property.

Lemma 4.5. *Under Assumption 2.3, $\|\Pi_\mu P\|_\mu < \infty$ if and only if $\forall s' \notin \operatorname{supp}(\mu), \mathbb{E}_{s\sim\mu}[\varphi(s)\mathcal{P}(s'|s)] = 0$.*

Proof. We show that $\|\Pi_\mu P\|_\mu < \infty \iff (\Pi P)_{\mu,\neg\mu} = 0_{\mu,\neg\mu}$, and then that this implies $\forall i \in \mu, k \notin \mu$, $\langle \varphi_i, \Sigma^{-1} \left(\sum_j \mu_j \varphi_j P_{j,k}\right)\rangle = 0$. Lastly we show that If $\lambda_{\min}(\Sigma) > 0$ then $\|\Pi_\mu P\|_\mu < \infty$ if and only if $\sum_j \mu \varphi_j P_{j,k} = 0 \, \forall k \notin \mu$.

The first part is easily observed by noting that $\|\Pi_\mu P\|_\mu < \infty \iff \max_{\|v\|_\mu=1} \|\Pi_\mu P v\|_\mu < \infty \iff \max_{\|v\|_\mu=1} \|((\Pi_\mu P)_{\mu,\mu} v_\mu + (\Pi_\mu P)_{\mu,\neg\mu} v_{\neg\mu}; 0_{\neg\mu})\|_\mu < \infty \iff (\Pi_\mu P)_{\mu,\neg\mu} = 0$, where the last line follows since if it has a non-trivial kernel then we can take $v_{\neg\mu}$ going to infinity while satisfying the constraints $\|v\|_\mu = 1$. The second part is observed by expanding the definition of $(\Pi_\mu P)_{i,k}$ for all $i \in \mu$ and all $k \notin \mu$. For the last part, we note that $\lambda_{\min}(\Sigma) > 0$ implies that the span of $\{\varphi(s_i)\}_{i\in\operatorname{supp}(\mu)} = \mathbb{R}^d$. Thus, for each $k \notin \mu$, the set of equations $\langle \varphi_i, \Sigma^{-1} \left(\sum_j \mu_j \varphi_j P_{j,k}\right)\rangle = 0$ obtained by varying over all $i \in \mu$ must imply that the vector on the RHS must be 0, i.e. $\Sigma^{-1} \left(\sum_j \mu_j \varphi_j P_{j,k}\right) = 0 \implies \sum_j \mu_j \varphi_j P_{j,k} = 0$ for each k. □

Proof (of 4.4). There are $m \coloneqq 3$ states in μ, and $n \coloneqq 2$ states in $\neg\mu$. We number the known states as $1, 2, 3$ and the unknown states as 4 and 5. The states within μ transition amongst each other and to the unknown states. The unknown states simply self-loop. The reward will be

$$\mathbb{R} = (0, 0, 0, r4, r5),$$

where $r4$ and $r5$ are chosen later. We also set

$$\gamma = 9/10.$$

The only things left to choose are $(\mathbb{P}, \varphi, \mu)$. Let us write down the transition matrix.

$$\mathbb{P} = \begin{pmatrix} 0.313 & 0.2322 & 0.2999 & 0.0786 & 0.0763 \\ 0.8483 & 0.0014 & 0.0867 & 0.0484 & 0.0152 \\ 0.1144 & 0.2852 & 0.219 & 0.2437 & 0.1377 \\ 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 1 \end{pmatrix}$$

where the floating point numbers are exact (i.e. can be represented as rationals).

And of course the discounted occupancy matrix is

$$\mathbb{d} \coloneqq (I - \gamma\mathbb{P})^{-1} = \begin{pmatrix} 2.22637 & 0.675069 & 0.814047 & 3.65445 & 2.63005 \\ 1.76839 & 1.56311 & 0.74639 & 3.56891 & 2.35319 \\ 0.85084 & 0.586281 & 1.58849 & 4.3413 & 2.63309 \\ 0 & 0 & 0 & 10 & 0 \\ 0 & 0 & 0 & 0 & 10 \end{pmatrix}$$

The data distribution is not yet chosen but will have the following constraints

$$\mu_1 > 0, \mu_2 > 0, \mu_3 > 0, \mu_4 = \mu_5 = 0$$

The task of the learner is to predict a value function on μ, i.e. on the first 3 states. Let us write $\mathbb{d}_4$ for the 4^{th} column of $\mathbb{d}$, and $\mathbb{d}_5$ for the 5^{th} of $\mathbb{d}$. The space of possible value functions in this MRP is

$$\mathcal{V}_M = \{v = r4 \cdot \mathbb{d}_4 + r5 \cdot \mathbb{d}_5 \mid r4, r5 \in [-1, 1]\} \subseteq \mathbb{R}^5,$$

since we set $r_1 = r_2 = r_3 = 0$. The space of possible value functions restricted to μ is:

$$\mathcal{V}_M^\mu = \left\{(v(s_1), v(s_2), v(s_3))^\top = r4 \cdot \mathbb{d}_{1::3,4} + r5 \cdot \mathbb{d}_{1::3,5} \mid r4, r5 \in [-1, 1]\right\} \subseteq \mathbb{R}^3,$$

where $\mathbb{d}_{1::3,4}$ is the first 3 elements of $\mathbb{d}_4$, and $\mathbb{d}_{1::3,5}$ is the first 3 elements of $\mathbb{d}_5$ (i.e. the column vectors $(3.65445, 3.56891, 4.3413)^\top$ and $(2.63005, 2.35319, 2.63309)^\top$, respectively). This is a $2-$dimensional plane lying in $\mathbb{R}^3$. There is no loss of generality in assuming that the learner will pick a hypothesis whose restriction to μ is in $\mathcal{V}_M^\mu$, as hypothesis lying outside of $\mathcal{V}_\mathcal{M}$ would be incorrect for all choices of reward functions (thus, strictly worse).

We pick a $1-$dimensional feature mapping $\varphi : \mathcal{S} \mapsto \mathbb{R}$ (i.e. $\Phi \in \mathbb{R}^{5\times 1}$). We choose Φ such that it is a linear combination of the last two columns of $\mathbb{d}$, i.e.

$$\Phi = \alpha\mathbb{d}_4 + \beta\mathbb{d}_5,$$

which means that Φ is a vector lying inside $\mathcal{V}_\mathcal{M}$. Our particular choice of α and β give

$$(\varphi_1, \varphi_2, \varphi_3)^\top = -0.5874\mathbb{d}_{1::3,4} + 0.9354\mathbb{d}_{1::3,5} = (0.313528, 0.104797, -0.0870883)^\top \tag{20}$$

The only thing left to pick now is μ. We cannot do this arbitrarily, as we have to ensure that $\|\Pi_\mu P\|_\mu < \infty$. Following the characterization of Lemma 4.5, we need to ensure that $\sum_j \mu_j\varphi_j P_{j,4} = 0$ and $\sum_j \mu_j\varphi_j P_{j,5} = 0$. Since we have chosen φ and P, the above two equations are linear constraints in μ. Together with the constraint that $\mu_1 + \mu_2 + \mu_3 = 1$, we can solve them to find that

$$(\mu_1, \mu_2, \mu_3)^\top = (0.0840949, 0.660425, 0.25548)^\top$$

Note that such a solution—where μ is a valid distribution—is not always possible for different choices of $\mathbb{P}$ and φ, hence the seemingly mysterious choices for $\mathbb{P}$ and φ. This particular instance was found via a random search: we keep generating P and the coefficients in Eq.(20) for defining φ, and stop when we find an instance with $\mu_1, \mu_2, \mu_3 > 0$ (they can be negative).

It remains to show that 1) $\sigma_{\min}(\Sigma^{-1/2}A\Sigma^{-1/2}) = 0$, and 2) the worst-case asymptotic approximation error is ∞. For 1), one can verify that with this choice of (P, φ, μ) we have:

$$\Sigma = 0.0174572 > 0 \,\&\, A = 0 \implies \sigma_{\min}(\Sigma^{-1/2}A\Sigma^{-1/2}) = 0.$$

The last thing to argue is that the error is ∞. The space of possible linear predictors is the line $\{\theta \cdot (\varphi_1, \varphi_2, \varphi_3)\}$. Recall that we picked Φ such that this entire line lies inside $\mathcal{V}_M$. In other words, there are an infinite number of possible realizable value functions that the environment could pick (obtained via $(r4, r5) = \theta(\alpha, \beta)$ for arbitrary θ).

However, from the perspective of the learner, the only information available is the value of the reward inside μ (which is 0), the transitions inside μ (the matrix $\mathbb{P}_{\mu,\mu}$), and the transitions from μ to $\neg\mu$ (the matrix $\mathbb{P}_{\mu,\neg\mu}$). In fact we can assume that the learner knows the whole $\mathbb{P}$ matrix, since $\mathbb{P}_{\neg\mu,\mu} = 0$ and $\mathbb{P}_{\neg\mu,\neg\mu} = \mathtt{Id}_{2\times 2}$ (self-loops). But none of this information is enough to deduce the value of r_4, r_5 (the reward happens at states that are unsupported), and this reward is what determines the true value function. So, for whatever value function the learner picks, we can pick a different realizable value function, thus rendering the approximation factor infinite. (See Figure 4).

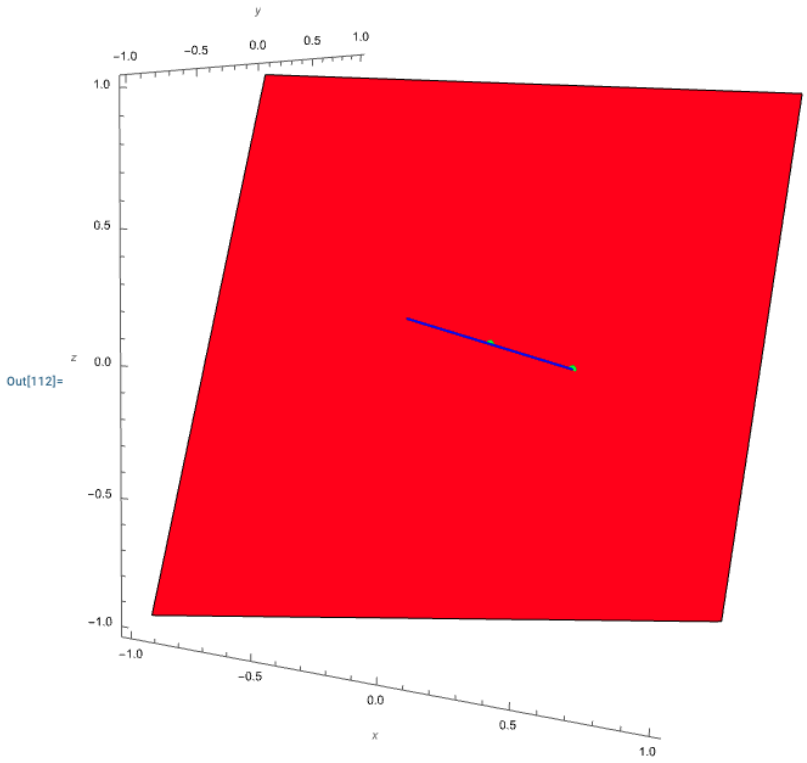


Figure 4. The construction above. Red plane: $\mathcal{V}_M$, space of value functions. Blue line: $\{\theta \cdot (\varphi_1, \varphi_2, \varphi_3) \mid \theta\}$, space of linear predictors, which lies inside $\mathcal{V}_M$. Two green points: a hypothesis value function and the other true value function.

□

B. Cases where $\alpha = 1$ is asymptotically achievable

Thanks to the proof of Equation 12, we identify several scenarios where the true solution can be recovered.

1. $\Phi A^{-1}\Phi^\top DPv^\perp = 0$, in particular $\Phi^\top DPv^\perp = 0$, e.g. when is satisfied under the condition that the orthogonal subspace of $\mathrm{col}(\Phi)$ is closed under P (i.e. P maps orthogonal vectors (of the features) to orthogonal vectors). Then LSTD has an $L_2(\mu)$ approximation factor of 1.
   - Proof: from Equation (13).
2. $\|P\|_\mu < \infty$ implies that $v_\mathcal{M}$ can be learned exactly on the support of μ. Thus with the tabular function class the asymptotic approximation ratio in the $L_2(\mu)$ norm is either 1 if $\|P\|_\mu < \infty$ or ∞ if $\|P\|_\mu$.
   - Proof: If $\|P\|_\mu < \infty$ then we must have the condition $(\mu(s) > 0 \,\&\, P(s'|s) > 0) \implies \mu(s') > 0$, otherwise in the equation $\max_{\|v\|_\mu = 1} \|Pv\|_\mu$ we will have a contribution of $P_{s,s'}v(s')$ for some unsupported state s', and the

value for $v(s')$ can be taken to infinity while satisfying the constraint $\|v\|_\mu = 1$. From this condition it is easy to see that $v_\mathcal{M}$ can be recovered exactly on μ, as in the asymptotic regime we have access to $r(s) \forall s \in \mu$ and $P(s) \forall s \in \mu$, and if a state transitions to s' then we will also have $P(s')$ and $r(s')$.

C. Proofs for Section 5

C.1. Proof of Theorem 5.1

Theorem 5.1. *Assume that the A matrix from Equation (8) is invertible. Then the population LSTD estimator has an approximation factor upper bound of*

$$\alpha_\infty^{LSTD} \leq 1 + \left\|\Phi A^{-1}\Phi^\top D(I-\gamma P)\right\|_\infty \leq 1 + \frac{1+\gamma}{\sigma_{\min}(A)}$$

Proof. We repeat the steps of Lemma A.1. Let us write $v_\mathcal{M} = \Pi_\infty v_\mathcal{M} + \delta := \Phi\theta_\infty + \delta$, so that $\delta = v_\mathcal{M} - \Pi_\infty v_\mathcal{M}$.

Then we have:

$$\begin{aligned}
(I-\gamma P)v_M &= r \\
(I-\gamma P)\Phi\theta_\infty + (I-\gamma P)\delta &= r \\
\Phi^\top D(I-\gamma P)\Phi\theta_\infty + \Phi^\top D(I-\gamma P)\delta &= \Phi^\top Dr && (\Phi^\top D \text{ on both sides}) \\
\left(\Phi^\top D(\Phi - \gamma P\Phi)\right)\theta_\infty &= \Phi^\top Dr - \Phi^\top D(I-\gamma P)\delta \\
A\theta_\infty &= b - \Phi^\top D(I-\gamma P)\delta && (\text{Defns of } A, b) \\
\theta_\infty &= A^{-1}b - A^{-1}\Phi^\top D(I-\gamma P)\delta && (A^{-1} \text{ exists})
\end{aligned}$$

Meanwhile, the LSTD solution is defined by $\theta_{\text{LSTD}} = A^{-1}b$. Substracting both of these gives:

$$\theta_\infty - \theta_{\text{LSTD}} = -A^{-1}\Phi^\top D(I-\gamma P)\delta \tag{21}$$

Now, applying Φ and taking the ∞ norm gives

$$\begin{aligned}
\|\Phi(\theta_\infty - \theta_{\text{LSTD}})\|_\infty &= \left\|\Phi A^{-1}\Phi^\top D(I-\gamma P)\delta\right\|_\infty \\
&\leq \left\|\Phi A^{-1}\Phi^\top D(I-\gamma P)\right\|_\infty \|\delta\|_\infty \\
&\leq \left\|\Phi A^{-1}\Phi^\top D\right\|_\infty \|(I-\gamma P)\|_\infty \|\delta\|_\infty \\
&\leq \left\|\Phi A^{-1}\Phi^\top D\right\|_\infty (1+\gamma)\|\delta\|_\infty
\end{aligned}$$

It remains to relate $\left\|\Phi A^{-1}\Phi^\top D\right\|_\infty$ to $\sigma_{\min}(A)$. Notice that

$$(\Phi A^{-1}\Phi^\top D)_{i,j} = \mu_j \left\langle \varphi_i, A^{-1}\varphi_j \right\rangle,$$

The L_∞ matrix norm is the maximum L_1 norm of a row, thus

$$\begin{aligned}
\left\|\Phi A^{-1}\Phi^\top D\right\|_\infty = \max_i \left(\sum_j \left| \mu_j \left\langle \varphi_i, A^{-1}\varphi_j \right\rangle \right| \right) &\leq \max_i \left(\sum_j \mu_j \|\varphi_i\|_2 \left\|A^{-1}\varphi_j\right\|_2 \right) && (\text{Cauchy-Schwartz}) \\
&= \left\|A^{-1}\right\|_2 \max_i \|\varphi_i\|_2 \left(\sum_j \mu_j \|\varphi_j\|_2 \right) \\
&\leq \left\|A^{-1}\right\|_2 1 && (\|\varphi_i\| \leq 1 \, \forall i) \\
&= \frac{1}{\sigma_{\min}(A)}
\end{aligned}$$

Combining everything and using a triangle inequality gives us:

$$\|v_{\texttt{LSTD}} - v_{\mathcal{M}}\|_\infty \leq (1 + \frac{(1+\gamma)1}{\sigma_{\min}(A)})\|\Phi\theta_\infty - v_{\mathcal{M}}\|_\infty$$

□

C.2. Proof of Theorem 5.2

Theorem 5.2. *In the non-aliased setting, for all $\gamma \in [c_1, 1)$ where c_1 is some absolute constant, and for all $y \in [0, 1-\gamma]$, there exists three instances $\{(\mathcal{M}_i, \mu_i, \varphi_i)\}$ which all satisfy $\sigma_{\min}(A) = y$ yet*

$$\inf_{\hat{v}} \sup_{(\mathcal{M}_i, \mu_i, \varphi_i)} \alpha_\infty^{\hat{v}}(\mathcal{M}_i, \mu_i, \varphi_i) \geq \frac{1}{2} + \frac{\gamma}{\sigma_{\min}(A)}.$$

The value of the constant is upper bounded by $c_1 \leq 0.7$.

Proof. We take $P = \begin{pmatrix} 0 & 1 \\ 0 & 1 \end{pmatrix}$ and $D = \begin{pmatrix} 1 & 0 \\ 0 & 0 \end{pmatrix}$. Note that this is the same MRP as in (Amortila et al., 2020) and Lemma 4.3. This gives a discounted occupancy matrix

$$\mathbb{d} = (I - \gamma P)^{-1} = \begin{pmatrix} 1 & \gamma/(1-\gamma) \\ 0 & 1/(1-\gamma) \end{pmatrix}.$$

Let d_1 denote the first column of $\mathbb{d}$ and d_2 denote the second column. We take $r = (0, r_2)^\top$, i.e. no reward at state 1 and a reward of r_2 at state 2. This gives $v_{\mathcal{M}} = r_2 d_2$. We set one instance to have $r_2 = 1$, one instance to have $r_2 = 0$, and the last instance to have $r_2 = -1$. The three instances are otherwise identical. We take $\Phi = [\alpha d_1 + d_2]\,(1-\gamma) \in \mathbb{R}^{2\times 1}$ (thus $\varphi(s) \in \mathbb{R}$), and we will later impose that $0 \leq \alpha \leq 1$. Assuming that this bound on α holds for now, we can see that $\|\varphi_1\|_2 = [\alpha + \gamma/(1-\gamma)](1-\gamma) \leq (1-\gamma)/(1-\gamma) = 1$ and $\|\varphi_2\|_2 = (1-\gamma)/(1-\gamma)$ and thus $\|\varphi_i\|_2 \leq 1$ for all i. We can directly verify that

$$A = \Phi^\top D(I - \gamma P)\Phi = \left[\alpha^2 + \alpha\gamma/(1-\gamma)\right](1-\gamma)^2 = \alpha^2(1-\gamma)^2 + \alpha\gamma(1-\gamma)$$

We need $\sigma_{\min}(A) = |A| = A = y$, so we can solve the quadratic for α and pick the positive solution to get:

$$\alpha = \frac{-\gamma + \sqrt{\gamma^2 + 4y}}{2(1-\gamma)}$$

Note that α satisfies the bound $0 \leq \alpha \leq 1$ whenever $\gamma < 1$ and $0 \leq y \leq 1 - \gamma$, which holds by the assumption in our theorem statement. The misspecification error is at most:

$$\inf_\theta \|v_{\mathcal{M}} - \Phi\theta\|_\infty = \inf_\theta \|r_2 d_2 - (\alpha d_1 + d_2)\theta(1-\gamma)\|_\infty = \inf_\theta \|(r_2 - (1-\gamma)\theta)d_2 - \alpha(1-\gamma)\theta d_1\|_\infty \leq \alpha\|d_1\|_\infty = \alpha,$$

where the upper bound was obtained by plugging in $\theta = \frac{r_2}{1-\gamma}$. Note that the minimax estimator against these three instances will need to output $\theta = 0$ since the instance with $r_2 = 0$ is realizable with $\theta = 0$. Namely, if the learner does not output

$\theta = 0$ then its worst-case approximation will be ∞. This gives the ratio:

$$\begin{aligned}
\alpha_\infty &\geq \frac{\|v_\mathcal{M} - 0\|_\infty}{\|\Pi_\infty v_\mathcal{M} - v_\mathcal{M}\|_\infty} \\
&\geq \frac{\|v_\mathcal{M}\|_\infty}{\alpha} \\
&= \frac{1}{\alpha(1-\gamma)} \\
&= \frac{1}{y}\left\{\alpha(1-\gamma)+\gamma\right\} && \text{(using that } \alpha(1-\gamma)\left\{\alpha(1-\gamma)+\gamma\right\} = y \text{ by definition of } A\text{)} \\
&= \frac{1}{y}\left\{\frac{-\gamma+\sqrt{\gamma^2+4y}}{2}+\gamma\right\} && \text{(using that } \alpha = \tfrac{-\gamma+\sqrt{\gamma^2+4y}}{2(1-\gamma)}\text{)} \\
&= \frac{\gamma}{2y}\left\{1+\sqrt{1+\frac{4y}{\gamma^2}}\right\} \\
&\geq \frac{\gamma}{2y}\left\{1+1+\frac{2y}{\gamma^2}-\frac{(4y)^2}{8\gamma^4}\right\} && \text{(using that } \sqrt{1+x} \geq 1 + x/2 - x^2/8 \text{ for all } x \geq 0\text{)} \\
&= \frac{\gamma}{y}\left\{1+\frac{y}{\gamma^2}-\frac{y^2}{\gamma^4}\right\} \\
&= \frac{\gamma}{y}+\frac{1}{\gamma}-\frac{y}{\gamma^3} \\
&\geq \frac{\gamma}{y}+\frac{1}{\gamma}-\frac{1-\gamma}{\gamma^3} && \text{(using that } y \leq 1-\gamma\text{)} \\
&\geq \frac{\gamma}{y}+\frac{1}{2}, && \text{(using that } \tfrac{1}{\gamma}-\tfrac{1-\gamma}{\gamma^3} \geq \tfrac{1}{2} \text{ when } \gamma \geq c_1\text{)}
\end{aligned}$$

as desired. The value of c_1 can be taken to be the smallest solution x such that $\frac{1}{x} - \frac{1-x}{x^3} \geq \frac{1}{2}$, which by Mathematica is approximately 0.6889 (but one can verify that 0.7 suffices and that this inequality holds for all $x \geq 0.7$ since the function is increasing). □

C.3. Proof of Theorem 5.4

Theorem 5.4. *Under Assumption 5.3, the estimator v_φ from Equation (15) has an approximation ratio of $\frac{2}{1-\gamma}$, i.e. we have*

$$\begin{aligned}
\|v_\varphi \circ \varphi - v_M\|_\infty &\leq \frac{2}{1-\gamma} \inf_{f:\mathcal{X}\mapsto\mathbb{R}} \|f \circ \varphi - v_\mathcal{M}\|_\infty \\
&\leq \frac{2}{1-\gamma} \inf_{\theta} \|\Phi\theta - v_\mathcal{M}\|_\infty
\end{aligned}$$

Proof. Inspired by the theory of "$q^\star$-irrelevant abstractions" (Li et al., 2006; Jiang, 2018; Xie & Jiang, 2021), we define $v_\mathcal{M}$-irrelevant abstractions as follows:

Definition C.1. A feature map $\varphi : \mathcal{S} \mapsto \mathcal{X}$ is an $\varepsilon-$approximate $v_\mathcal{M}-$irrelevant abstraction for MRP $\mathcal{M}$ if there exists a function $f : \mathcal{X} \mapsto \mathbb{R}$ such that

$$\inf_{f:\mathcal{X}\mapsto\mathbb{R}} \|f \circ \varphi - v_\mathcal{M}\|_\infty = \varepsilon$$

Note that the $\inf$ is taken over all pointwise functions over $\mathcal{X}$, and thus every feature mapping with an L_∞-misspecification error of ε is also a ε-approximate $v_\mathcal{M}$-irrelevant abstraction.

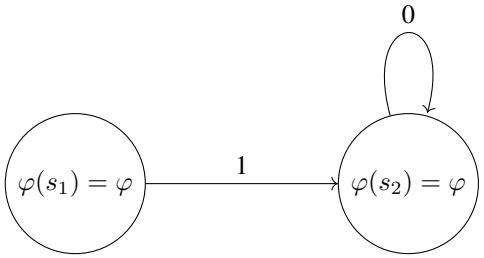
To conclude the proof we use Theorem 5 from (Jiang, 2018), which establishes the analogous claim for the case of $q^\star$-irrelevant abstractions. Indeed, the case of $v_\mathcal{M}$-irrelevant abstractions can be reduced from the more general case of $q^\star$-irrelevant abstractions by considering the case where there is only one action to take in each state. It is easily seen that our Bayes model is then equivalent to the model constructed in Lemma 3 of (Jiang, 2018), which Theorem 5 uses to establish the approximation error bound of $2/(1-\gamma)$. □

C.4. Proof of Theorem 5.5

Theorem 5.5. *In the aliased setting, under Assumption 5.3, $\forall \varepsilon > 0$, $\forall\, \gamma \in (0,1)$, there exists a collection of two instances $\mathbb{M} = \{(\mathcal{M}_1, \mu_1, \varphi_1), (\mathcal{M}_2, \mu_2, \varphi_2)\}$ which generate the same data distribution $\mathbb{Q}$, yet any estimator $\hat{v}$ will satisfy*

$$\sup_{(\mathcal{M},\mu,\varphi)\in\mathbb{M}} \alpha_\infty^{\hat{v}}(\mathcal{M},\mu,\varphi) \geq \frac{2}{1-\gamma} - \varepsilon$$

Proof. The first MRP $\mathcal{M}_1$ is defined as



We set $\varphi = 1$ for simplicity. We place the initial distribution $\mu(s_1) = p$ and $\mu(s_2) = 1 - p$, and should think of $p \to 1$ (we can't actually set $p = 1$ due to the full-support condition, but a limiting argument suffices). Note that $v_{\mathcal{M}}(s_1) = 1$ and $v_{\mathcal{M}}(s_2) = 0$, so the optimal ∞-norm approximation for this MRP is $\theta_1 = \varphi\theta = \frac{1}{2}$.

Our second MRP $\mathcal{M}_2$ is the following:

which generates the same distribution $\mathbb{Q}$. This instance is realizable with value function $\theta_2 = v_{\mathcal{M}} = \frac{p}{1-\gamma}$, which forces our estimator to output θ_2. Let p be large enough such that $\|\theta_2 - v_{\mathcal{M}_1}\|_\infty = \max\{|\frac{p}{1-\gamma} - 1|, |\frac{p}{1-\gamma} - 0|\} = \frac{p}{1-\gamma}$ (i.e. $p > (1-\gamma)/2$). Taking the ratio of approximation errors:

$$\frac{\|\theta_2 - v_{\mathcal{M}_1}\|_\infty}{\|\theta_1 - v_{\mathcal{M}_1}\|_\infty} = \frac{p/(1-\gamma)}{1/2} = \frac{2p}{1-\gamma} \geq \frac{2}{1-\gamma} - \varepsilon,$$

where the last step takes $p \geq 1 - \frac{\varepsilon(1-\gamma)}{2}$. □

C.5. Proof of Corollary C.2

Corollary C.2. *The projected Bayes value function has an approximation factor of*

$$\|\Pi_\infty(v_\varphi \circ \varphi) - v_{\mathcal{M}}\|_\infty \leq \left(1 + \frac{2}{1-\gamma}\right) \inf_\theta \|\Phi\theta - v_{\mathcal{M}}\|_\infty$$

Proof. This amounts to an application of the triangle inequality:

$$\begin{aligned}
\|\Pi_\infty(v_\varphi \circ \varphi) - v_{\mathcal{M}}\|_\infty &\leq \|\Pi_\infty v_\varphi - \Pi_\infty v_{\mathcal{M}}\|_\infty + \|\Pi_\infty v_{\mathcal{M}} - v_{\mathcal{M}}\|_\infty \\
&\leq \|(v_\varphi \circ \varphi) - v_{\mathcal{M}}\|_\infty + \inf_\theta \|\Phi\theta - v_{\mathcal{M}}\|_\infty && (\Pi_\infty \text{ is non-expansive}) \\
&\leq \frac{2}{1-\gamma} \inf_\theta \|\Phi\theta - v_{\mathcal{M}}\|_\infty + \inf_\theta \|\Phi\theta - v_{\mathcal{M}}\|_\infty && (\text{Previous bound}) \\
&= \left(1 + \frac{2}{1-\gamma}\right) \varepsilon_\infty,
\end{aligned}$$

which concludes the proof. □

D. Translating $L_2(\mu)$ oracle inequalities to L_∞ oracle inequalities

This section shows that one can convert an $L_2(\mu)$ oracle inequality to an L_∞ oracle inequality

Lemma D.1. *Assuming we have a bound*

$$\|v_\theta - v_\mathcal{M}\|_\mu \leq \alpha_\mu \|\Pi_\mu v_\mathcal{M} - v_\mathcal{M}\|_\mu.$$

This can be converted to an approximation ratio bound

$$\|v_\theta - v_\mathcal{M}\|_\infty \leq \left(1 + \max_s \left\|\Sigma^{-1/2}\varphi(s)\right\|_2 (1+\alpha_\mu)\right) \|\Pi_\infty v_\mathcal{M} - v_\mathcal{M}\|_\infty$$

Proof. Let $\Phi\theta_\mathcal{M}$ be an L_∞ linear projection, and $\delta(s)$ be such that $v_\mathcal{M}(s) = \delta(s) + \theta_\mathcal{M}^\top \varphi(s)$.

$$\begin{aligned}
\|v_\mathcal{M} - v_\theta\|_\infty &= \max_s |\theta^\top \varphi(s) - v_\mathcal{M}(s)| \\
&= \max_s |\theta^\top \varphi(s) - \theta_\mathcal{M}^\top \varphi(s) - \delta(s)| \\
&\leq \|\delta(s)\|_\infty + \max_s |(\theta^\top \varphi(s) - \theta_\mathcal{M})^\top \varphi(s)| \\
&\leq \|\delta(s)\|_\infty + \max_s \left\|\Sigma^{-1/2}\varphi(s)\right\|_2 \left\|\Sigma^{1/2}(\theta_\mathcal{M} - \theta)\right\|_2 && \text{(Cauchy-Schwartz)} \\
&= \|\delta(s)\|_\infty + \left(\max_s \left\|\Sigma^{-1/2}\varphi(s)\right\|_2\right) \|\Phi(\theta_\mathcal{M} - \theta)\|_\mu \\
&\leq \|\delta(s)\|_\infty + \left(\max_s \left\|\Sigma^{-1/2}\varphi(s)\right\|_2\right) \left(\|\Phi(\theta_\mathcal{M}) - v_\mathcal{M}\|_\mu + \|v_\mathcal{M} - \Phi\theta\|_\mu\right) \\
&\leq \|\delta(s)\|_\infty + \left(\max_s \left\|\Sigma^{-1/2}\varphi(s)\right\|_2\right) \left(\|\Phi(\theta_\mathcal{M}) - v_\mathcal{M}\|_\mu + \alpha_\mu \|v_\mathcal{M} - \Pi_\mu v_\mathcal{M}\|_\mu\right) \\
&\leq \|\delta(s)\|_\infty + \left(\max_s \left\|\Sigma^{-1/2}\varphi(s)\right\|_2\right) \left(\|\Phi(\theta_\mathcal{M}) - v_\mathcal{M}\|_\mu + \alpha_\mu \|v_\mathcal{M} - \Phi\theta_\mathcal{M}\|_\mu\right) \\
& && (\Pi_\mu v_\mathcal{M} = \inf_{\hat{v}\in\mathcal{F}_\Phi} \|v_\mathcal{M} - \hat{v}\|) \\
&\leq \|\delta(s)\|_\infty + \left(\max_s \left\|\Sigma^{-1/2}\varphi(s)\right\|_2\right) \left((1+\alpha_\mu)\|\Phi(\theta_\mathcal{M}) - v_\mathcal{M}\|_\mu\right) \\
&\leq \|\delta(s)\|_\infty + \left(\max_s \left\|\Sigma^{-1/2}\varphi(s)\right\|_2\right) (1+\alpha_\mu)\|\Phi(\theta_\mathcal{M}) - v_\mathcal{M}\|_\infty \\
&= \|\delta(s)\|_\infty + \left(\max_s \left\|\Sigma^{-1/2}\varphi(s)\right\|_2\right) (1+\alpha_\mu)\|\delta(s)\|_\infty \\
&= \left(1 + \max_s \left\|\Sigma^{-1/2}\varphi(s)\right\|_2 (1+\alpha_\mu)\right) \|\delta(s)\|_\infty
\end{aligned}$$

□