An Optimal Tightness Bound for the Simulation Lemma

Sam Lobel
samuel_lobel@brown.edu
Department of Computer Science
Brown University

Ronald Parr
parr@cs.duke.edu
Department of Computer Science
Duke University

Abstract

We present a bound for value-prediction error with respect to model misspecification that is tight, including constant factors. This is a direct improvement of the "simulation lemma," a foundational result in reinforcement learning. We demonstrate that existing bounds are quite loose, becoming vacuous for large discount factors, due to the suboptimal treatment of compounding probability errors. By carefully considering this quantity on its own, instead of as a subcomponent of value error, we derive a bound that is sub-linear with respect to transition function misspecification. We then demonstrate broader applicability of this technique, improving a similar bound in the related subfield of hierarchical abstraction.

1 Introduction

In reinforcement learning, an agent is frequently tasked with making decisions in an environment that it cannot model perfectly. This may occur because the environment is learned about through sampled data, or because the agent's environment model is simplified through some abstraction. In such cases it is natural to ask, how might the quality of this approximation impact an agent's decision making? This is the subject of the "simulation lemma," a foundational result in reinforcement learning that bounds the error in value estimation when the transition and reward function are known only with some specified degree of precision.

The simulation lemma was introduced in the context of exploration and finds use in a variety of domains that utilize imperfect models, such as hierarchical abstraction (Abel et al., 2016) and offline policy evaluation (Yin et al., 2021). Frequently, results of this kind rely on developing a recursive relationship between the value error at subsequent timesteps. We show that this approach implicitly overestimates how probability errors compound over time. By more directly approximating this quantity, we produce a bound on value-estimation error that is demonstrably tight. We then show that existing bounds can be derived as a linearization of our result, and finally apply our result to a hierarchical setting to demonstrate broader applicability.

2 Background and Related Work

We develop our results in the framework of Markov Decision Processes (MDPs): $\mathcal{M} = (\mathcal{S}, \mathcal{A}, R, T, \gamma)$, where $\mathcal{S}$ is the state space, $\mathcal{A}$ is the action space, and $\gamma \in [0, 1]$ is the discount factor. The next-state transition probabilities are given by $T(s'|s, a)$, and the reward function by $R(s, a) \in [0, 1]$. A policy $\pi(a|s)$ gives the probability of taking an action from a given state. The objective in the MDP framework is generally either to construct a policy π that maximizes the expected γ-discounted sum of reward, or to evaluate a given policy on this same measure.

When a model of the environment is given, these quantities can be computed exactly, for example through policy iteration or dynamic programming (Howard, 1960). In reinforcement learning, however, the agent generally is not given this model, and instead must learn about the environment

through interaction. A common approach to this is model-based reinforcement learning (Moerland et al., 2023; Auer & Ortner, 2006), which aims to estimate the environment's transitions and rewards from gathered data. However, when using finite data, the learned model is generally imperfect. This work concerns itself with developing optimal bounds on policy evaluation error in the setting of misspecified models. Here we detail a variety of areas in which such a bound is useful, along with related lines of study.

Exploration The original simulation lemma was introduced in the context of efficient exploration (Kearns & Singh, 2002), to quantify policy evaluation error as a function of state-action visitation counts. Understanding the effect of imperfect modelling is central to efficient exploration (Auer & Ortner, 2006; Auer et al., 2008; Brafman & Tennenholtz, 2002). Methods that use these measures include count-based exploration (Strehl & Littman, 2008) and its pseudocount approximations (Bellemare et al., 2016; Lobel et al., 2023).

Abstraction Model approximation frequently appears in the field of abstraction, where a full model of an MDP is replaced by one that is simpler in some respect. As we show later, our methodology can be used to improve the value error bounds when performing this replacement with state-action abstracted *options* (Sutton et al., 1999). A simple form of state abstraction is *discretization*, where sets of states are grouped by some measure of similarity. A common example of this occurs in the *partially observable* MDP framework (Lee et al., 2007; Grover & Dimitrakakis, 2021), where the continuous belief-state space can be discretized into an approximate, finite MDP.

Offline Policy Evaluation The goal of offline policy evaluation (OPE) is to estimate the value of a policy using a fixed dataset of transitions, often generated by a different policy. Model-based OPE involves fitting an empirical model of transitions and rewards from this dataset, and using this to estimate value (Gottesman et al., 2019). In this setting, the simulation lemma often is a key step in constructing accuracy bounds of the estimated value (Yin & Wang, 2020; Yin et al., 2021).

We also note that a variety of results in the literature bound the value error using different measures of similarity than the original simulation lemma. Perhaps most closely related to our contribution is work that bounds multi-step transition error of imperfectly-modelled Lipschitz transition functions (Asadi et al., 2018). This results in a similar sum of compounding errors to ours, albiet in a different setting. Bisimulation metrics (Ferns et al., 2004) unify transition and reward error into a single quantity that can be used to measure the similarity of MDPs with entirely different state spaces.

3 Main Result

We begin by stating the conditions of the original simulation lemma. We consider two MDPs: $\mathcal{M} = (\mathcal{S}, \mathcal{A}, R, T, \gamma)$, and $\hat{\mathcal{M}} = (\mathcal{S}, \mathcal{A}, \hat{R}, \hat{T}, \gamma)$, which share a state-action space, but have (boundedly) different transition and reward functions. We are interested in the effect of running the same policy π on these two related MDPs. Let P^π be a matrix that contains the policy-conditioned state-state transition probabilities, and R^π be a vector that contains the per-state expected reward:

$$\begin{aligned} P^\pi_{s,s'} &= \mathbb{E}_{a\sim\pi(s)}[T(s'|s,a)] &&= \sum_{a\in\mathcal{A}} T(s'|s,a)\pi(a|s) \\ R^\pi_s &= \mathbb{E}_{a\sim\pi(s)}[R(s,a)] &&= \sum_{a\in\mathcal{A}} R(s,a)\pi(a|s). \end{aligned} \tag{1}$$

We define $\hat{P}^\pi$ and $\hat{R}^\pi$ analogously for MDP $\hat{\mathcal{M}}$. Throughout this work, a single index on a matrix (or vector) extracts the specified row vector (or scalar). Furthermore, P^a and R^a refer to the transition probabilities, and expected reward, of executing action a from each state. Using this notation, we can quantify the difference between two transition or reward functions with the following:

$$\forall s, \pi: \ \|P_s^\pi - \hat{P}_s^\pi\|_1 \ \leq \ \epsilon_T \tag{2}$$
$$\forall \pi: \ \|R^\pi - \hat{R}^\pi\|_\infty \ \leq \ \epsilon_R. \tag{3}$$

We are interested in the value difference between running π on each MDP. The value of a state for a given policy and MDP is defined as the expected discounted sum of rewards:

$$v^\pi(s) = \mathbb{E}_{a_i \sim \pi(s_i)}\big[\sum_{t=0}^{\infty} \gamma^t R(s_t, a_t) \mid s_0 = s, \mathcal{M}\big],$$

where s_t is a random variable representing the state at timestep t. Noting that $\Pr(s_t = s' | s_0 = s, \pi) = (P^\pi)^t_{s,s'}$, we can concisely represent value in vectorized notation as follows:

$$V^\pi = \sum_{t=0}^{\infty} \gamma^t (P^\pi)^t R^\pi \quad , \quad V_s^\pi = \sum_{t=0}^{\infty} \gamma^t \langle (P^\pi)^t_s, R^\pi \rangle,$$

where $\langle \cdot, \cdot \rangle$ denotes the inner product between two vectors. We define $\hat{V}^\pi$ analogously for $\hat{\mathcal{M}}$.

3.1 Original Simulation Lemma

We are interested in quantifying the maximum value difference between running the same policy on two different MDPs. The original simulation lemma bounds this quantity as follows:

$$\forall s, \pi: \ \ |V_s^\pi - \hat{V}_s^\pi| \leq \frac{\epsilon_R}{1-\gamma} + \frac{\gamma \epsilon_T}{2(1-\gamma)^2}. \tag{4}$$

Existing proofs of the simulation lemma frequently take advantage of a recursive representation of value (the Bellman Equation) (Howard, 1960):

$$V^\pi = R^\pi + \gamma P^\pi \sum_{t=0}^{\infty} \gamma^t (P^\pi)^t R^\pi = R^\pi + \gamma P^\pi V^\pi.$$

For a complete proof, please refer to Jiang (2018) or see Appendix A. The key mathematical idea is to establish the following recursive relationship:

$$\forall s, \pi: \ \ |V_s^\pi - \hat{V}_s^\pi| \leq \epsilon_R + \frac{\gamma \epsilon_T}{2(1-\gamma)} + \gamma \|V^\pi - \hat{V}^\pi\|_\infty, \tag{5}$$

which can then be easily transformed into the simulation lemma's bound. Analyzing the recursive relationship above, the first term (ϵ_R) represents a one-step reward-prediction error. The second term ($\frac{\gamma \epsilon_T}{2(1-\gamma)}$) represents the maximum value error that results from misspecifying ϵ_T of the next-state distribution's probability mass. However, by defining the recursive relationship as such, this bound implicitly assumes that the process can continually misspecify ϵ_T of its probability at each timestep. This quickly amounts to misspecifying more than the entire probability mass, leading to a vast overestimate of the value error, in particular when $\epsilon_T > 1 - \gamma$. In contrast, we carefully track the probability drift at each timestep to avoid this issue.

3.2 Bounding Probability Distance

We seek to bound the probability distance tightly at any timestep t. To do so effectively, it is useful to frame distances between probability vectors in terms of their overlap, instead of their L_1 distance.

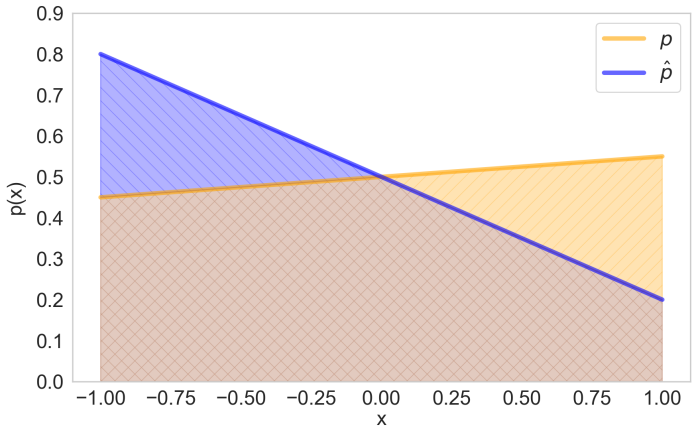

Figure 1: Visualization of relation between L_1 distance and overlap of two probability distributions (Equation 7). The blue and orange shaded regions together comprise the L_1 distance. The brown region represents overlap. Overlap plus *either* the blue or orange sections constitutes a probability distribution, and therefore has total area 1. Thus the blue and orange regions both individually have area $\|p - \hat{p}\|_1/2$, and so $\|\bar{p}\|_1 = 1 - \|p - \hat{p}\|_1/2$.

We note that Jiang et al. (2016) uses similar machinery to bound compounding probability error (Lemma 1), though applies this insight in a different context. For two probability vectors p, $\hat{p}$, we define their overlap as $\bar{p}$, such that for each index i:

$$\bar{p}_i = \min(p_i, \hat{p}_i).$$

Usefully, because each element of $p - \bar{p}$ (and likewise $\hat{p} - \bar{p}$) is non-negative, the L_1 norm of the difference between these two vectors is equal to the difference between the L_1 norms:

$$\|p - \bar{p}\|_1 = \sum_i |p_i - \bar{p}_i| = \sum_i p_i - \sum_i \bar{p}_i = \|p\|_1 - \|\bar{p}\|_1 \tag{6}$$

We use this to derive an equivalence between overlap and L_1 distance, related to the concept of *total variation distance* (Levin & Peres, 2017). Below, we use the notation $[p]^+$ to indicate a thresholded version of p that retains only the non-negative parts, $[p]^+_i = \max(p_i, 0)$:

$$\begin{aligned} \|p - \hat{p}\|_1 &= \|[p - \bar{p}]^+\|_1 + \|[\hat{p} - \bar{p}]^+\|_1 \\ &= \|p - \bar{p}\|_1 + \|\hat{p} - \bar{p}\|_1 \\ &= \|p\|_1 - \|\bar{p}\|_1 + \|\hat{p}\|_1 - \|\bar{p}\|_1 \\ &= 1 + 1 - 2\|\bar{p}\|_1 \\ \implies \|\bar{p}\|_1 &= 1 - \frac{\|p - \hat{p}\|_1}{2}. \end{aligned} \tag{7}$$

See Figure 1 for a demonstration and explanation of this equivalence. This relationship allows for a simple rewriting of the transition-error condition of the simulation lemma (Equation 2):

$$\forall s, \pi : \ \|\bar{P}^\pi_s\|_1 \geq 1 - \frac{\epsilon_T}{2}. \tag{8}$$

Using this framing, we can now lower-bound the overlap of state-distributions at timestep t when starting from s_0, by demonstrating that at every timestep, at least $1 - \epsilon_T/2$ fraction of the prior timestep's distributional overlap is retained. For notational convenience, $P^t_{s_0,s} = (P^\pi)^t_{s_0,s}$, and

$\bar{M}^t_{s_0,s} = \min(P^t_{s_0,s}, \hat{P}^t_{s_0,s})$. Thus,

$$\begin{aligned}
\|\bar{M}^{t+1}_{s_0}\|_1 &= \sum_{s'} \min(P^{t+1}_{s_0,s'}, \hat{P}^{t+1}_{s_0,s'}) \\
&= \sum_{s'} \min(\sum_{s} P^t_{s_0,s} \cdot P^\pi_{s,s'} \ , \ \sum_{s} \hat{P}^t_{s_0,s} \cdot \hat{P}^\pi_{s,s'}) \\
&\geq \sum_{s'} \sum_{s} \min(P^t_{s_0,s} \cdot P^\pi_{s,s'} \ , \ \hat{P}^t_{s_0,s} \cdot \hat{P}^\pi_{s,s'}) \\
&\geq \sum_{s'} \sum_{s} \min\Big(\min(P^t_{s_0,s} \ , \ \hat{P}^t_{s_0,s}) \cdot P^\pi_{s,s'}) \ , \ \min(P^t_{s_0,s} \ , \ \hat{P}^t_{s_0,s}) \cdot \hat{P}^\pi_{s,s'}\Big) \\
&= \sum_{s} \sum_{s'} \min(P^t_{s_0,s} \ , \ \hat{P}^t_{s_0,s}) \min(P^\pi_{s,s'} \ , \ \hat{P}^\pi_{s,s'}) \\
&= \sum_{s} \min(P^t_{s_0,s} \ , \ \hat{P}^t_{s_0,s}) \sum_{s'} \min(P^\pi_{s,s'} \ , \ \hat{P}^\pi_{s,s'}) \\
&\geq \|\bar{M}^t_{s_0}\|_1 \cdot \max_s \|\bar{P}^\pi_s\|_1 \\
\implies \|\bar{M}^{t+1}_{s_0}\|_1 &\geq \|\bar{M}^t_{s_0}\|_1 \cdot (1 - \epsilon_T/2).
\end{aligned}$$

The third line can be understood as providing the minimum operator more options to choose from, in that after bringing the minimum inside of the sum, the two elements in the second line are both still possible choices and so the inequality holds. The fourth line can be understood similarly for multiplication.

With $\bar{M}^0 = I$ as the base case, applying recursion yields

$$\|\bar{M}^t_{s_0}\|_1 \geq (1 - \epsilon_T/2)^t. \tag{9}$$

We contrast this with the equivalent recursive proof of distributional drift using the L_1 formulation of transition misspecification, akin to the recursion employed by the original simulation lemma (Equation 5):

$$\begin{aligned}
\|P^{t+1}_{s_0} - \hat{P}^{t+1}_{s_0}\|_1 &= \|P^t_{s_0} P^\pi - \hat{P}^t_{s_0} \hat{P}^\pi\|_1 \\
&= \frac{1}{2}\|(P^t_{s_0} - \hat{P}^t_{s_0})(P^\pi + \hat{P}^\pi) + (P^t_{s_0} + \hat{P}^t_{s_0})(P^\pi - \hat{P}^\pi)\|_1 \\
&\leq \frac{1}{2}\|P^t_{s_0} - \hat{P}^t_{s_0}\|_1 \|(P^\pi + \hat{P}^\pi)^T\|_1 + \frac{1}{2}\|P^t_{s_0} + \hat{P}^t_{s_0}\|_1 \|(P^\pi - \hat{P}^\pi)^T\|_1 \\
&= \|P^t_{s_0} - \hat{P}^t_{s_0}\|_1 + \|P^\pi - \hat{P}^\pi\|_1 \\
&\leq \|P^t_{s_0} - \hat{P}^t_{s_0}\|_1 + \epsilon_T \\
\implies \|P^{t+1}_{s_0} - \hat{P}^{t+1}_{s_0}\|_1 &\leq (t+1)\,\epsilon_T,
\end{aligned}$$

where $\|\cdot\|_1$ above refers to both the matrix and vector 1-norm, and on the third line we use the identity $\|Ax\|_1 \leq \|A\|_1 \|x\|_1$. This result makes clear the contrast between the two methods for computing distributional drift: Naïvely using the L_1 formulation leads to unbounded accumulation of drift as horizon approaches infinity, while the overlap formulation smoothly decays from 1 to 0. This difference is crucial to generating the tighter bound in the next section.

3.3 A Tight Bound on Value Error

We are now ready to prove our main result, a tight bound on the value error.

Theorem 1 *For two MDPs $\mathcal{M}$ and $\hat{\mathcal{M}}$ related as described in Equations 2 and 3, the following inequality holds:*

$$\forall s, \pi: \ |V_s^\pi - \hat{V}_s^\pi| \ \leq \ \frac{1}{1-\gamma} - \frac{1-\epsilon_R}{1-\gamma(1-\epsilon_T/2)}. \tag{10}$$

Furthermore, this bound is tight.

Proof: Since the conditions of the simulation lemma (Equations 2,3) are symmetric with respect to $\mathcal{M}$ and $\hat{\mathcal{M}}$, without loss of generality we assume $V_{s_0}^\pi \geq \hat{V}_{s_0}^\pi$, and thus $|V_{s_0}^\pi - \hat{V}_{s_0}^\pi| = V_{s_0}^\pi - \hat{V}_{s_0}^\pi$. We now add and subtract the same quantity in a way that allows for discarding a strictly non-positive term:

$$\begin{aligned}
|V_{s_0}^\pi - \hat{V}_{s_0}^\pi| &= V_{s_0}^\pi - \hat{V}_{s_0}^\pi \\
&= \sum_{t=0}^{\infty} \gamma^t \langle P_{s_0}^t, R^\pi \rangle - \gamma^t \langle \hat{P}_{s_0}^t, \hat{R}^\pi \rangle \\
&= \sum_{t=0}^{\infty} \gamma^t \Big(\langle P_{s_0}^t, R^\pi \rangle - \langle \bar{M}_{s_0}^t, R^\pi \rangle + \langle \bar{M}_{s_0}^t, R^\pi \rangle - \langle \bar{M}_{s_0}^t, \hat{R}^\pi \rangle + \langle \bar{M}_{s_0}^t, \hat{R}^\pi \rangle - \langle \hat{P}_{s_0}^t, \hat{R}^\pi \rangle \Big) \\
&= \sum_{t=0}^{\infty} \gamma^t \langle P_{s_0}^t - \bar{M}_{s_0}^t, R^\pi \rangle + \gamma^t \langle \bar{M}_{s_0}^t, R^\pi - \hat{R}^\pi \rangle + \gamma^t \langle \bar{M}_{s_0}^t - \hat{P}_{s_0}^t, \hat{R}^\pi \rangle.
\end{aligned}$$

By construction, $\bar{M}_{s_0}^t$ is the overlap between $P_{s_0}^t$ and $\hat{P}_{s_0}^t$, and thus and entries of $\bar{M}_{s_0}^t - \hat{P}_{s_0}^t$ are strictly non-positive. Since rewards are likewise non-negative, the third inner product in the above sum is always non-positive. Thus, we can drop this term to significantly tighten our bound.

$$\begin{aligned}
V_{s_0}^\pi - \hat{V}_{s_0}^\pi &\leq \sum_{t=0}^{\infty} \gamma^t \langle P_{s_0}^t - \bar{M}_{s_0}^t, R^\pi \rangle + \gamma^t \langle \bar{M}_{s_0}^t, R^\pi - \hat{R}^\pi \rangle \\
&\leq \sum_{t=0}^{\infty} \gamma^t \| P_{s_0}^t - \bar{M}_{s_0}^t \|_1 \cdot \| R^\pi \|_\infty + \gamma^t \| \bar{M}_{s_0}^t \|_1 \cdot \| R^\pi - \hat{R}^\pi \|_\infty \\
&\leq \sum_{t=0}^{\infty} \gamma^t \| P_{s_0}^t - \bar{M}_{s_0}^t \|_1 + \gamma^t \| \bar{M}_{s_0}^t \|_1 \epsilon_R \\
&= \sum_{t=0}^{\infty} \gamma^t \| P_{s_0}^t \|_1 - \gamma^t \| \bar{M}_{s_0}^t \|_1 + \gamma^t \| \bar{M}_{s_0}^t \|_1 \epsilon_R \\
&= \sum_{t=0}^{\infty} \gamma^t + \gamma^t (\epsilon_R - 1) \| \bar{M}_{s_0}^t \|_1 \\
&\leq \sum_{t=0}^{\infty} \gamma^t + \gamma^t (\epsilon_R - 1)(1 - \epsilon_T/2)^t \\
&= \frac{1}{1-\gamma} + (\epsilon_R - 1) \sum_{t=0}^{\infty} (\gamma - \frac{\gamma \epsilon_T}{2})^t \\
\implies |V_{s_0}^\pi - \hat{V}_{s_0}^\pi| &\leq \frac{1}{1-\gamma} - \frac{1-\epsilon_R}{1-\gamma(1-\epsilon_T/2)}.
\end{aligned}$$

■

This proof makes use of Hölder's inequality to bound inner products with L_1 and L_∞ norms, as well as the identity in Equation 6 to split $\| P_{s_0}^t - \bar{M}_{s_0}^t \|_1$ into $\| P_{s_0}^t \|_1 - \| \bar{M}_{s_0}^t \|_1$. We provide a parallel

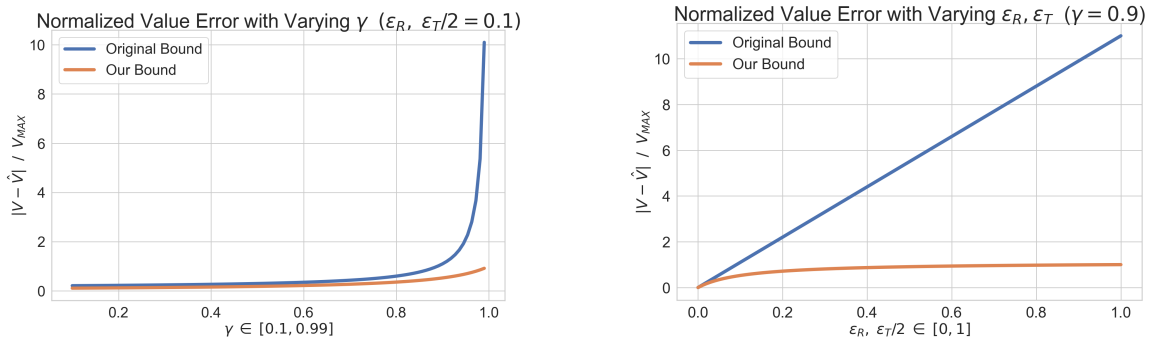

Figure 2: Bounds on value error given by original simulation lemma as well as our tighter bounds, normalized by V_{MAX}. (Left) Bound on value error with increasing gamma shows the original lemma's suboptimality with respect to discount. (Right) Bound on value error with increasing misspecification shows looseness of linear approximation compared to the tight bound.

proof for the finite-horizon undiscounted setting in Appendix B. We briefly remark that this bound matches intuition:

- When $\gamma = 0$, then $|V_s^\pi - \hat{V}_s^\pi| \leq \epsilon_R$ since only the first step contributes to value.
- When $\epsilon_R = 1$, the MDPs can have completely different reward functions and thus $|V_s^\pi - \hat{V}_s^\pi| \leq \frac{1}{1-\gamma} = V_{MAX}$.
- When $\epsilon_R = \epsilon_T = 0$, the MDPs are identical and thus $|V_s^\pi - \hat{V}_s^\pi| = 0$.

Additionally we note the the original simulation lemma can be reproduced as a Taylor expansion of our bound around $\epsilon_R = 0$ and $\epsilon_T = 0$, proving that the original bound is the tightest possible linear approximation to the maximal error as model misspecification approaches 0. Figure 2 presents a comparison of our bound with the original simulation lemma, demonstrating superiority in the large-misspecification and large-discount limits.

3.4 Proof of Tightness

We now demonstrate that this is the tightest possible bound, including constant factors, by constructing a pair of MDPs with exactly this value error. $\mathcal{M}$ consists of two states, both of which transition to themselves, with $R(s_1) = 1$ and $R(s_2) = 0$. We construct $\hat{\mathcal{M}}$ so that $\hat{V}(s_1)$ is as small as possible given ϵ_R, ϵ_T, by setting $\hat{R}(s_1) = 1 - \epsilon_R$, and transitioning from s_1 to s_2 with probability $\epsilon_T/2$ (and thus self-transitions with $\epsilon_T/2$ less probability, so $\|P_{s_1}^\pi - \hat{P}_{s_1}^\pi\|_1 = \epsilon_T$). Hence, $V(s_0) = \frac{1}{1-\gamma}$ and $\hat{V}(s_0) = \frac{1-\epsilon_R}{1-\gamma(1-\epsilon_T/2)}$.

Intuitively, this result makes clear the role of ϵ_T as modifying the discount factor of $\hat{\mathcal{M}}$. A discount can be interpreted as entering an absorbing state with probability $1-\gamma$ at each timestep (Sutton & Barto, 2018). In $\hat{\mathcal{M}}$, this instead occurs more frequently, with probability $1 - \gamma(1 - \epsilon_T/2)$.

3.5 Value Loss of Optimal Policy

The simulation lemma directly applies to bounding the value difference of executing the same policy on two related MDPs. However, in reinforcement learning the task is frequently to learn an *optimal* policy π^*, that has the following property:

$$\forall \pi, s : V_s^{\pi^*} \geq V_s^\pi.$$

It is natural to ask, if one learns the optimal policy $\hat{\pi}^*$ by training on an approximate MDP $\hat{\mathcal{M}}$, how much worse will this policy do than π^* when executed on the actual MDP $\mathcal{M}$? In contrast to the simulation lemma, we are comparing the value loss of *different* policies on the *same* MDP. Noting that $\hat{V}_s^{\hat{\pi}^*} \geq \hat{V}_s^{\pi^*}$:

$$\begin{aligned} V_s^{\pi^*} - V_s^{\hat{\pi}^*} &= V_s^{\pi^*} + (\hat{V}_s^{\pi^*} - \hat{V}_s^{\pi^*}) + (\hat{V}_s^{\hat{\pi}^*} - \hat{V}_s^{\hat{\pi}^*}) - V_s^{\hat{\pi}^*} \\ &= (V_s^{\pi^*} - \hat{V}_s^{\pi^*}) + (\hat{V}_s^{\pi^*} - \hat{V}_s^{\hat{\pi}^*}) + (\hat{V}_s^{\hat{\pi}^*} - V_s^{\hat{\pi}^*}) \\ &\leq (V_s^{\pi^*} - \hat{V}_s^{\pi^*}) + 0 + (\hat{V}_s^{\hat{\pi}^*} - V_s^{\hat{\pi}^*}) \\ &\leq |V_s^{\pi^*} - \hat{V}_s^{\pi^*}| + |\hat{V}_s^{\hat{\pi}^*} - V_s^{\hat{\pi}^*}|. \end{aligned}$$

This is simply twice the value error of executing the *same* policy on *different* MDPs. Thus, by improving the simulation lemma bound, we similarly tighten the estimated value loss when training on an approximate MDP. Similar results are common in inverse RL, e.g., Burchfiel et al. (2016), and have been noted in the context of the simulation lemma as well (Jiang, 2018).

3.6 Application to Hierarchy

Analogs to the simulation lemma exist throughout the reinforcement learning literature; here, we present an extension of our proof to one such instance in the field of hierarchical reinforcement learning. We use the formalism of ϕ-relative options (Abel et al., 2020), a form of approximately value preserving state and action abstractions.

Let $\mathcal{O}_\phi^*$ be a set of options o^* over abstract states $s_\phi \in \mathcal{S}_\phi$, that can be composed to form a policy that is optimal in the base MDP. Let $\hat{\mathcal{O}}_\phi$ be a set of options that approximates $\mathcal{O}_\phi^*$ in that

$$\begin{aligned} &\forall o^* \in \mathcal{O}_\phi^* \;\; \exists \hat{o} \in \hat{\mathcal{O}}_\phi : \\ &\forall s, s' \;\; |P_{s,s'}^{o^*} - P_{s,s'}^{\hat{o}}| \leq \epsilon_T \;\text{ and }\; |R_s^{o^*} - R_s^{\hat{o}}| \leq \epsilon_R, \end{aligned}$$

where R_s^o and $P_{s,s'}^o$ represent the reward and multi-time models of Sutton et al. (1999). We define $V^{\pi_{o^*}}$ as the value of executing the best policy over $\mathcal{O}_\phi^*$, and $V^{\pi_{\hat{o}}}$ as the value of executing an approximately equivalent policy using options from $\hat{\mathcal{O}}_\phi$. By bounding probability distances we arrive at the following relation:

$$|V_s^{\pi_{o^*}} - V_s^{\pi_{\hat{o}}}| \leq \frac{R_{MAX}}{1-\gamma} - \frac{R_{MAX} - \epsilon_R}{1 - \gamma + (|S| - 1)\epsilon_T}.$$

This improves on the existing bound (Abel et al., 2020):

$$|V_s^{\pi_{o^*}} - V_s^{\pi_{\hat{o}}}| \leq \frac{\epsilon_R + |S|\epsilon_T R_{MAX}}{(1-\gamma)^2},$$

in much the same way as our original result improves upon the simulation lemma. A proof, more complete definitions, and an example demonstrating tightness are deferred to Appendix C. The main difference in applying our technique to this domain is careful treatment of the multi-time transition function, where $\sum_{s'} P_{s,s'}^o \neq 1$.

4 Conclusion

The simulation lemma is a widely used result in reinforcement learning that quantifies the effect of model misspecification on value. We demonstrate that the originally provided bound is quite loose,

becoming vacuous when applied to large discount factors frequently used in reinforcement learning. In this work we present a version of this lemma that is optimally tight, along with an example application of this method to hierarchical reinforcement learning. We expect that our bound can be applied to a variety of results throughout the literature, and that the general proof technique can be useful in other domains.

Acknowledgements

We would like to thank George Konidaris for valuable input during the early stages of this work, as well as Tuluhan Akbulut and Ruo Yu Tao for helping inspire the question we ask here. This material is based upon work supported by the National Science Foundation Graduate Research Fellowship under grant #2040433 and ARO grant #W911NF2210251.

References

David Abel, David Hershkowitz, and Michael Littman. Near optimal behavior via approximate state abstraction. In *International Conference on Machine Learning*, pp. 2915–2923. PMLR, 2016.

David Abel, Nate Umbanhowar, Khimya Khetarpal, Dilip Arumugam, Doina Precup, and Michael Littman. Value preserving state-action abstractions. In *International Conference on Artificial Intelligence and Statistics*, pp. 1639–1650. PMLR, 2020.

Kavosh Asadi, Dipendra Misra, and Michael Littman. Lipschitz continuity in model-based reinforcement learning. In *International Conference on Machine Learning*, pp. 264–273. PMLR, 2018.

Peter Auer and Ronald Ortner. Logarithmic online regret bounds for undiscounted reinforcement learning. *Advances in neural information processing systems*, 19, 2006.

Peter Auer, Thomas Jaksch, and Ronald Ortner. Near-optimal regret bounds for reinforcement learning. *Advances in neural information processing systems*, 21, 2008.

Marc Bellemare, Sriram Srinivasan, Georg Ostrovski, Tom Schaul, David Saxton, and Remi Munos. Unifying count-based exploration and intrinsic motivation. *Advances in neural information processing systems*, 29, 2016.

Ronen I Brafman and Moshe Tennenholtz. R-max a general polynomial time algorithm for near-optimal reinforcement learning. *Journal of Machine Learning Research*, 3(Oct):213–231, 2002.

Benjamin Burchfiel, Carlo Tomasi, and Ronald Parr. Distance minimization for reward learning from scored trajectories. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 30, 2016.

Norman Ferns, Prakash Panangaden, and Doina Precup. Metrics for finite Markov decision processes. *Proceedings of the 20th conference on Uncertainty in Artificial Intelligence*, 2004.

Omer Gottesman, Yao Liu, Scott Sussex, Emma Brunskill, and Finale Doshi-Velez. Combining parametric and nonparametric models for off-policy evaluation. In *International Conference on Machine Learning*, pp. 2366–2375. PMLR, 2019.

Divya Grover and Christos Dimitrakakis. Adaptive belief discretization for POMDP planning. *arXiv preprint arXiv:2104.07276*, 2021.

Ronald A Howard. Dynamic programming and Markov processes. 1960.

Nan Jiang. Notes on tabular methods. 2018. URL https://nanjiang.cs.illinois.edu/files/cs542f22/note3.pdf.

Nan Jiang, Satinder Singh, and Ambuj Tewari. On structural properties of mdps that bound loss due to shallow planning. In *IJCAI*, volume 8, pp. 1, 2016.

Sham Kakade, Michael J Kearns, and John Langford. Exploration in metric state spaces. In *International Conference on Machine Learning*, pp. 306–312. PMLR, 2003.

Michael Kearns and Satinder Singh. Near-optimal reinforcement learning in polynomial time. *Machine learning*, 49:209–232, 2002.

Wee Lee, Nan Rong, and David Hsu. What makes some POMDP problems easy to approximate? *Advances in neural information processing systems*, 20, 2007.

David A Levin and Yuval Peres. *Markov chains and mixing times*, volume 107. American Mathematical Soc., 2017.

Sam Lobel, Akhil Bagaria, and George Konidaris. Flipping coins to estimate pseudocounts for exploration in reinforcement learning. In *International Conference on Machine Learning*, pp. 22594–22613. PMLR, 2023.

Thomas M Moerland, Joost Broekens, Aske Plaat, Catholijn M Jonker, et al. Model-based reinforcement learning: A survey. *Foundations and Trends® in Machine Learning*, 16(1):1–118, 2023.

Alexander L Strehl and Michael L Littman. An analysis of model-based interval estimation for Markov decision processes. *Journal of Computer and System Sciences*, 74(8):1309–1331, 2008.

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

Richard S Sutton, Doina Precup, and Satinder Singh. Between MDPs and semi-MDPs: A framework for temporal abstraction in reinforcement learning. *Artificial intelligence*, 112(1-2):181–211, 1999.

Ming Yin and Yu-Xiang Wang. Asymptotically efficient off-policy evaluation for tabular reinforcement learning. In *International Conference on Artificial Intelligence and Statistics*, pp. 3948–3958. PMLR, 2020.

Ming Yin, Yu Bai, and Yu-Xiang Wang. Near-optimal provable uniform convergence in offline policy evaluation for reinforcement learning. In *International Conference on Artificial Intelligence and Statistics*, pp. 1567–1575. PMLR, 2021.

A Full proof of Simulation Lemma

For completeness, we include the proof of the simulation lemma found in Jiang (2018). We adopt notation from Section 3.

$$\begin{aligned}
|V_s^\pi - \hat{V}_s^\pi| &= |R_s^\pi + \gamma\langle P_s^\pi, V^\pi\rangle - \hat{R}_s^\pi - \gamma\langle \hat{P}_s^\pi, \hat{V}^\pi\rangle| \\
&\leq \epsilon_R + \gamma|\langle P_s^\pi, V^\pi\rangle - \langle \hat{P}_s^\pi, V^\pi\rangle + \langle \hat{P}_s^\pi, V^\pi\rangle - \langle \hat{P}_s^\pi, \hat{V}^\pi\rangle| \\
&= \epsilon_R + \gamma|\langle P_s^\pi, V^\pi - \frac{\mathbf{1}}{2(1-\gamma)}\rangle - \langle \hat{P}_s^\pi, V^\pi - \frac{\mathbf{1}}{2(1-\gamma)}\rangle + \langle \hat{P}_s^\pi, V^\pi\rangle - \langle \hat{P}_s^\pi, \hat{V}^\pi\rangle| \\
&\leq \epsilon_R + \gamma\|P_s^\pi - \hat{P}_s^\pi\|_1 \cdot \|V^\pi - \frac{\mathbf{1}}{2(1-\gamma)}\|_\infty + \gamma\|\hat{P}_s^\pi\|_1 \cdot \|V^\pi - \hat{V}^\pi\|_\infty \\
&\leq \epsilon_R + \frac{\gamma\epsilon_T}{2(1-\gamma)} + \gamma\|V^\pi - \hat{V}^\pi\|_\infty \\
\implies |V_s^\pi - \hat{V}_s^\pi| &\leq \frac{\epsilon_R}{1-\gamma} + \frac{\gamma\epsilon_T}{2(1-\gamma)^2}.
\end{aligned}$$

This proof makes use of Hölder's inequality to bound inner products with L_1 and L_∞ norms, as well as centers the value $0 \leq V_s^\pi \leq \frac{1}{1-\gamma}$ through subtracting the midpoint for improved bounds.

B Application to the Finite-Horizon Setting

We now extend our improved bound to the finite-horizon, undiscounted setting, where an agent interacts with an environment for H steps. One difference in this setting is that policies are conditioned on timestep as well as state; hence we define $\pi = [\pi_0, \ldots, \pi_{H-1}]$. Existing bounds in the finite-horizon setting establish a relationship between values at subsequent timesteps. Noting that $0 \leq V^\pi_{h,s} \leq H - h$ (and defining $V^\pi_{H,s} = 0$), Then,

$$
\begin{aligned}
|V^\pi_{h,s} - \hat{V}^\pi_{h,s}| &= |R^{\pi_h}_s + \langle P^{\pi_h}_s, V^\pi_{h+1}\rangle - \hat{R}^{\pi_h}_s - \langle \hat{P}^{\pi_h}_s, \hat{V}^\pi_{h+1}\rangle| \\
&\leq \epsilon_R + |\langle P^{\pi_h}_s, V^\pi_{h+1}\rangle - \langle \hat{P}^{\pi_h}_s, V^\pi_{h+1}\rangle + \langle \hat{P}^{\pi_h}_s, V^\pi_{h+1}\rangle - \langle \hat{P}^{\pi_h}_s, \hat{V}^\pi_{h+1}\rangle| \\
&= \epsilon_R + |\langle P^{\pi_h}_s, V^\pi_{h+1} - \frac{H-h-1}{2}\cdot \mathbf{1}\rangle - \langle \hat{P}^{\pi_h}_s, V^\pi_{h+1} - \frac{H-h-1}{2}\cdot \mathbf{1}\rangle \\
&\quad + \langle \hat{P}^{\pi_h}_s, V^\pi_{h+1}\rangle - \langle \hat{P}^{\pi_h}_s, \hat{V}^\pi_{h+1}\rangle| \\
&\leq \epsilon_R + \|P^{\pi_h}_s - \hat{P}^{\pi_h}_s\|_1 \cdot \|V^\pi_{h+1} - \frac{H-h-1}{2}\cdot \mathbf{1}\|_\infty + \|V^\pi_{h+1} - \hat{V}^\pi_{h+1}\|_\infty \\
&\leq \epsilon_R + \epsilon_T \frac{H-h-1}{2} + \|V^\pi_{h+1} - \hat{V}^\pi_{h+1}\|_\infty \\
\implies |V^\pi_{h,s} - \hat{V}^\pi_{h,s}| &\leq \sum_{i=h}^{H-1} \epsilon_R + \epsilon_T \frac{H-i-1}{2} \\
\implies |V^\pi_{0,s} - \hat{V}^\pi_{0,s}| &\leq \epsilon_R H + \epsilon_T \frac{H(H-1)}{4}
\end{aligned}
$$

For our bound, the only change from the discounted setting is replacing the discounted infinite sums of Section 3.3 with finite undiscounted ones. Redefining $P^t = \prod_{0 \leq i < t} P^{\pi_i}$, and WLOG assuming that $V^\pi_{0,s_0} \geq \hat{V}^\pi_{0,s_0}$ we can show:

$$
\begin{aligned}
|V^\pi_{0,s_0} - \hat{V}^\pi_{0,s_0}| &= V^\pi_{0,s_0} - \hat{V}^\pi_{0,s_0} \\
&= \sum_{t=0}^{H-1} \langle P^t_{s_0}, R^{\pi_t}\rangle - \langle \hat{P}^t_{s_0}, \hat{R}^{\pi_t}\rangle \\
&= \sum_{t=0}^{H-1} \langle P^t_{s_0} - \bar{M}^t_{s_0}, R^{\pi_t}\rangle + \langle \bar{M}^t_{s_0}, R^{\pi_t} - \hat{R}^{\pi_t}\rangle + \langle \bar{M}^t_{s_0} - \hat{P}^t_{s_0}, \hat{R}^{\pi_t}\rangle \\
&\leq \sum_{t=0}^{H-1} \langle P^t_{s_0} - \bar{M}^t_{s_0}, R^{\pi_t}\rangle + \langle \bar{M}^t_{s_0}, R^{\pi_t} - \hat{R}^{\pi_t}\rangle \\
&\leq \sum_{t=0}^{H-1} \|P^t_{s_0} - \bar{M}^t_{s_0}\|_1 \cdot \|R^{\pi_t}\|_\infty + \|\bar{M}^t_{s_0}\|_1 \cdot \|R^{\pi_t} - \hat{R}^{\pi_t}\|_\infty \\
&\leq \sum_{t=0}^{H-1} \|P^t_{s_0} - \bar{M}^t_{s_0}\|_1 + \|\bar{M}^t_{s_0}\|_1 \epsilon_R \\
&= \sum_{t=0}^{H-1} \|P^t_{s_0}\|_1 - \|\bar{M}^t_{s_0}\|_1 + \|\bar{M}^t_{s_0}\|_1 \epsilon_R \\
&\leq \sum_{t=0}^{H-1} 1 + (\epsilon_R - 1)(1 - \epsilon_T/2)^t \\
\implies |V^\pi_{0,s_0} - \hat{V}^\pi_{0,s_0}| &\leq H - (1-\epsilon_R)\frac{2}{\epsilon_T}(1 - (1-\epsilon_T/2)^H)
\end{aligned}
$$

Again, we note that Taylor expanding this relation at $\epsilon_R = 0$ and $\epsilon_T = 0$ recovers the original bound.

C Proof of Hierarchy Bound

This proof exactly mirrors the one in the main body, with additional care taken to handle multi-time models. We first describe the ϕ-relative options framework (definitions largely taken from Abel et al. (2020)), and then provide a tighter bound on value loss.

An *option* $o \in \mathcal{O}$ is an abstract action defined by the tuple $(\mathcal{I}_o, \beta_o, \pi_o)$, where $\mathcal{I}_o \subseteq \mathcal{S}$ is the subset of the state space the option can initiate in, $\beta_0 \subseteq \mathcal{S}$ is the subset the option terminates in, and π_o is a policy. For a given state abstraction $\phi : \mathcal{S} \to \mathcal{S}_\phi$, an option o_ϕ is said to be ϕ-relative if and only if $\exists s_\phi \in \mathcal{S}_\phi$ such that

$$s \in s_\phi \implies s \in \mathcal{I}_{o_\phi} \qquad s \notin s_\phi \implies s \in \beta_{o_\phi} \qquad \forall s \in s_\phi, \;\; \pi_{o_\phi}(s) \to \Delta(\mathcal{A})$$

In words, a ϕ-relative option is one that executes from anywhere in one abstract state, and terminates upon leaving that abstract state. Furthermore, $\mathcal{O}_\phi$ denotes a set of only ϕ-relative options, with at least one option that executes at each abstract state.

Let $\mathcal{O}^*_\phi$ be a set of ϕ-relative options o^* that can be composed to form an optimal policy in the base MDP. Let $\hat{\mathcal{O}}_\phi$ be a set of options that approximates $\mathcal{O}^*_\phi$ in that

$$\begin{aligned} &\forall o^* \in \mathcal{O}^*_\phi \;\; \exists \hat{o} \in \hat{\mathcal{O}}_\phi : \\ &\forall s, s' \;\; |P^{o^*}_{s,s'} - P^{\hat{o}}_{s,s'}| \leq \epsilon_T \;\text{ and }\; |R^{o^*}_s - R^{\hat{o}}_s| \leq \epsilon_R \end{aligned} \tag{11}$$

where R^o_s and $P^o_{s,s'}$ represent the multi-time reward and transition functions described in Sutton et al. (1999):

$$R^o_s = \mathbb{E}_{a\sim o}[\sum_{t=0}^{\infty} \gamma^t r_t] \qquad P^o_{s,s'} = \sum_{t=1}^{\infty} \gamma^t \Pr(s_t = s', t_\beta = t).$$

In words, R^o_s is the expected discounted reward accumulated over the course of an option execution, and $P^o_{s,s'}$ is the total discounted probability that an option terminates in s' when starting from s. Crucially, $\sum_{s'\in\mathcal{S}} P^o_{s,s'} \leq \gamma < 1$. We also note that the ϵ_T bound is per-entry, not per-vector. This was the form of the conditions in the original simulation lemma (Kearns & Singh, 2002), which was replaced with a vectorized version in subsequent work (Kakade et al., 2003).

Since $\|P^o_s\|_1$ may take on different values for different options and starting states, we can no longer directly use a relation similar to Equation 7. However, we can augment the MDP by adding an absorbing state s_x, and modify each option such that

$$R^o_{s_x} = 0 \quad , \quad P^o_{s,s_x} = \gamma - \sum_{s'\neq s_x} P^o_{s,s'}.$$

By doing this, $\|P^o_s\|_1 = \gamma$ without modifying the behavior of the given option in the base MDP. This allows our proof to proceed treating options in roughly the same way as we do actions in the main body. Noting that since $P^o_{s,s} \equiv 0$ by construction, for two options $o^*, \hat{o}^*$ satisfying the relations of Equation 11 we have that:

$$|P^{o^*}_{s,s_x} - P^{\hat{o}}_{s,s_x}| \leq \sum_{s'\neq s_x, s} |P^{o^*}_{s,s'} - P^{\hat{o}}_{s,s'}| \leq (|S| - 1)\epsilon_T.$$

Thus we can recover a condition similar to that of Equation 2:

$$\|P^{o^*}_s - P^{\hat{o}}_s\|_1 = \sum_{s'\in\mathcal{S}+s_x} |P^{o^*}_{s,s'} - P^{\hat{o}}_{s,s'}| \leq 2(|S| - 1)\epsilon_T.$$

Due to the addition of s_x, we can now describe the above bound in terms of overlap. Defining $\bar{P}^{o^*,\hat{o}}_{s,s'} = \min(P^{o^*}_{s,s'}, P^{\hat{o}}_{s,s'})$, we can produce a similar relation to Equation 8:

$$\|\bar{P}^{o^*,\hat{o}}_{s}\|_1 \geq \gamma - (|S| - 1)\epsilon_T.$$

Let $\Pi_{\mathcal{O}_\phi}$ be the set of abstract policies representable by $\mathcal{O}_\phi$. Let π_{o^*} be a policy within $\Pi_{\mathcal{O}^*_\phi}$ that is optimal in the base MDP. Let $\pi_{\hat{o}}$ be a policy in $\Pi_{\hat{\mathcal{O}}_\phi}$ produced by replacing each o^* chosen by π_{o^*} with an option $\hat{o}$ satisfying the relations of Equation 11. Then, we can follow the same algebraic steps as in the main body to produce the following bound:

$$|V^{\pi_{o^*}}_s - V^{\pi_{\hat{o}}}_s| \leq \frac{R_{MAX}}{1-\gamma} - \frac{R_{MAX} - \epsilon_R}{1 - \gamma + (|S| - 1)\epsilon_T}.$$

C.1 Proof of Tightness

We can generate an abstract MDP that achieves this bound using a similar recipe as in Section 3.4. We construct an abstract MDP where each option o^* transitions uniformly to each other state with discounted probability $\frac{\gamma}{|S|-1}$, receiving a reward of R_{MAX}. We then construct a new set of options that uniformly transition with discounted probability $\frac{\gamma}{|S|-1} - \epsilon_T$, receiving reward $R_{MAX} - \epsilon_R$. This exactly reproduces the provided bound.