

Algorithmic Framework for Model-based Reinforcement Learning with Theoretical Guarantees

Huazhe Xu* Yuanzhi Li † Yuandong Tian ‡ Trevor Darrell § Tengyu Ma ¶

Abstract

While model-based reinforcement learning has empirically been shown to significantly reduce the sample complexity that hinders model-free RL, the theoretical understanding of such methods has been rather limited. In this paper, we introduce a novel algorithmic framework for designing and analyzing model-based RL algorithms with theoretical guarantees, and a practical algorithm Optimistic Lower Bounds Optimization (OLBO). In particular, we derive a theoretical guarantee of monotone improvement for model-based RL with our framework. We iteratively build a lower bound of the expected reward based on the estimated dynamical model and sample trajectories, and maximize it jointly over the policy and the model. Assuming the optimization in each iteration succeeds, the expected reward is guaranteed to improve. The framework also incorporates an optimism-driven perspective, and reveals the intrinsic measure for the model prediction error. Preliminary simulations demonstrate that our approach outperforms the standard baselines on continuous control benchmark tasks.

1 Introduction

In recent years reinforcement learning has achieved strong empirical success, including super-human performances on Atari games and Go [26, 35] and learning locomotion and manipulation skills in robotics [22, 33, 24]. Many of these results are achieved by model-free reinforcement learning algorithms that often require a massive number of samples, and therefore their applications are mostly limited to simulated environments. Model-based reinforcement learning, in contrast, exploits the information from state observations explicitly — by planning with a learned dynamical model — and is considered a promising approach to reduce the sample complexity. Indeed, empirical results [8, 9, 22, 28, 20, 30] have shown strong improvements in terms of sample efficiency.

Despite promising empirical findings, many of *theoretical* properties of model-based reinforcement are not well-understood. For example, how does the error of the estimated model affect the estimation of the value function and the planning? Can model-based RL algorithms be guaranteed to improve the policy monotonically and converge to a local maximum of the value function? How do we quantify the uncertainty in the dynamical models? Previous theoretical results [1, 36, 6, 39, 40] mostly focus on linear parametrizations of the value function, policy or dynamics, and thus may not be applicable to complex situations in deep reinforcement learning.

In this paper, we propose a novel algorithmic framework for model-based reinforcement learning with theoretical guarantees. We provide upper bound on how much the error can compound and divert the value of imaginary rollouts from their real value. With this, our algorithm builds a lower bound of

*UC Berkeley. huazhe_xu@eecs.berkeley.edu

†Princeton University. yuanzhi1@cs.princeton.edu

‡Facebook AI Research. yuandong@fb.com

§UC Berkeley. trevor@eecs.berkeley.edu

¶Facebook AI Research. tengyuma@stanford.edu

the true value function from sample trajectories, and maximizes it over both the dynamical model and the policy. The real value function is guaranteed to monotonically increase (assuming the planning succeeds in each iteration.) To the best of our knowledge, this is the first theoretical guarantee of monotone improvement for model-based reinforcement learning. The framework also incorporates an optimism-driven perspective, and reveals the intrinsic measure of the model prediction error.

More concretely, the key idea of the paper is we can use an estimated model and sample trajectories to build a provable lower bound of the real value function V^π :

$$V^\pi \geq \widehat{V}^\pi - D^{\pi, \widehat{M}}. \quad (1.1)$$

Here \widehat{V}^π is the value function of the policy π on the estimated model \widehat{M} , and $D^{\pi, \widehat{M}}$ is a designed discrepancy bound D that captures the intrinsic difference between the estimated model \widehat{M} and the real dynamical model M^* .

Since discrepancy bound $D^{\pi, \widehat{M}}$ is *invariant* to the representation of the state space it may lead to a better design of loss function for learning dynamical models. Under certain assumptions and simplification, our theory can recover the standard model-based RL algorithms but suggests that ℓ_2 or ℓ_1 norm loss is preferable compared the mean-squared error (MSE). As shown below, we indeed observe that ℓ_2 and ℓ_1 losses significantly outperform MSE baseline in continuous tasks in Mujoco [42].

We also justify our framework by showing that jointly optimizing policy and dynamical model yields better results compared to the standard scheme of tuning the model and policy separately. Readers may have realized that optimizing a robust lower bound is reminiscent of robust control and robust optimization. We remark that the vital distinction is that we optimistically and iteratively maximize the RHS of (1.1) jointly over the model \widehat{M} and the policy π , which compensates the conservatism from the lower bound.

Last but not the least, we remark that the most sophisticated theoretical results in Section 4.3 develop and utilize new mathematical tools that measure the difference between policies in χ^2 -divergence (instead of KL or TV). These tools may be of independent interests and used for better analysis of model-free reinforcement learning algorithms such as TRPO [32], PPO [34] and CPO [3].

1.1 Related work

Model-based reinforcement learning are known to require fewer samples than model-free algorithms [9] and have been successfully applied to robotics in both simulation and in the real world [8, 27, 10] using dynamical models ranging from relatively simple models such as Gaussian process [8, 19], time-varying linear models [23, 25, 21, 43], mixture of Gaussians [17], to multi-layer neural networks [14, 28, 20, 41]. In particular, the work of Kurutach et al. [20] uses an ensemble of neural networks to learn the dynamical model significantly reduced the sample complexity compared to model-free approaches. In contrast, we focus on theoretical understanding of model-based RL to design of new algorithms, and our experiments use a single neural network to approximate the dynamical model.

Prior work explores a variety of ways of combining model-free and model-based ideas to achieve the best of the two methods [38, 37]. For example, learned models [23, 13, 16] are used to enrich the replay buffer in the model-free off-policy RL. The work of Pong et al. [41] proposes goal-conditioned value functions trained by model-free algorithms and use it for model-based controls. The work of Feinberg et al. [12] uses dynamical models to improve the estimation of the value functions in model-free algorithms.

Recent work [7, 6] provide the strong finite sample complexity bounds for solving linear quadratic regulator (linear dynamical system with quadratic reward function) using model-based approach. Boczar et al. [4] provide finite-data guarantees for the “coarse-ID control” pipeline, which is composed of a system identification step followed by a robust controller synthesis procedure. Our method, by contrast, applies to non-linear dynamical systems. Our algorithm also estimate the models iteratively based on sampled trajectories from the learned policies.

The work [39, 40] analyzes the behavior of Dyna-like algorithms when both value function and dynamics are linear with TD(0). Abbeel et al.[2] shows such iterative approaches converge to a

local optimum of expected reward, assuming the partial derivative of learned model is in the vicinity of the true models since optimization starts, which could be hard to satisfy. The work of Feinberg et al. [12] bounds the value discrepancy between the real and estimated models by ℓ_2^2 error of the trajectories and use it to improve the value estimation in model-free RL. Our discrepancy bounds, in contrast, depend on the expected error of the model in one step and can be invariant to the state representation. Moreover, our algorithms enjoy convergence guarantees in the sense that the value under the unknown true dynamics is non-decreasing.

2 Notations and Preliminaries

We mostly work with continuous state and action space, although most of the results can be extended to discrete action space as well. We denote the state space by $\mathcal{S} \subset \mathbb{R}^d$, the action space by $\mathcal{A} \subset \mathbb{R}^\ell$. A policy $\pi(\cdot|\cdot)$ specifies the conditional distribution over the action space given a state. A dynamical model $M(\cdot|S, A)$ specifies the conditional distribution of the next state given the current state S and action A . We will use M^* globally to denote the unknown real dynamical model. Let \mathcal{M} denote a (parameterized) family of models that we are interested in. We use Π to denote the parameterized family of policies.

With slight abuse of the notation, we also use π and M the stochastic function defined by π and M : by $\pi(s)$ we mean the random variable with distribution $\pi(\cdot|s)$. Unless otherwise state, we will use capital letters and Greek letters for random variables. For random variable X , we will use p_X to denote its density function.

Let $S_0^{\pi, M}, \dots, S_t^{\pi, M}, \dots$ to denote the random variable of the states steps $0, \dots$, when we execute policy π on dynamic model M . We will omit the subscript when it's clear from the context. We use A_0, \dots, A_t, \dots for actions similarly. We often use τ to denote the random variables for the trajectory $(S_0, A_1, \dots, S_t, A_t, \dots)$. Let γ be the discount factor and $V^{\pi, M}$ be the value function:

$$V^{\pi, M}(s) = \mathbb{E}_{\substack{\forall t \geq 0, A_t \sim \pi(\cdot|S_t) \\ S_{t+1} \sim M(\cdot|S_t, A_t)}} \left[\sum_{t=0}^{\infty} \gamma^t R(S_t, A_t) \mid S_0 = s \right] \quad (2.1)$$

We define $V^{\pi, M} = \mathbb{E}[V^{\pi, M}(S_0)]$ as the expected reward-to-go at step 0 and our goal is to maximize the cumulative rewards V^{π, M^*} . For simplicity, throughout the paper, we set $\kappa = \gamma(1 - \gamma)^{-1}$ since it occurs frequently in our equations. Every policy π induces a distribution of states visited by policy π , as formally defined below.

Definition 2.1. For a policy π , define $\rho^{\pi, M}$ as the discounted distribution of the states visited by π on M . Let ρ^π be a shorthand for ρ^{π, M^*} and we omit the superscript M^* throughout the paper. Concretely, let $p_{S_t^\pi | S_0=s}$ be the distribution of $S_t^\pi \mid S_0^\pi = s$ and let $\rho^\pi = (1 - \gamma) \sum_{t=0}^{\infty} \gamma^t p_{S_t^\pi}$

Given the purpose of the paper is to study model-based RL algorithms with sample efficiency, we assume that given an imaginary model \widehat{M} , we can optimize the imaginary reward (possibly with a large amount of imaginary rollouts.) Indeed, given sufficient samples from the imaginary dynamical models, model-free algorithms such as policy gradient, TRPO [32], DDPG [24], PPO [34], etc., often perform very well.

3 Algorithmic Framework

As alluded in the introduction, towards optimizing V^{π, M^*} , our plan is to build a lower bound for V^{π, M^*} of the following type and optimize it iteratively:

$$V^{\pi, M^*} \geq V^{\pi, \widehat{M}} - D(\widehat{M}, \pi) \quad (3.1)$$

where $D(\widehat{M}, \pi) \in \mathbb{R}_{\geq 0}$ bounds from above the discrepancy between $V^{\pi, \widehat{M}}$ and V^{π, M^*} . Building such an optimizable discrepancy bound globally that holds for all \widehat{M} and π turns out to be rather difficult, if not impossible. Instead, we shoot for establishing such a bound over the neighborhood of a reference policy π_{ref} .

$$V^{\pi, M^*} \geq V^{\pi, \widehat{M}} - D_{\pi_{\text{ref}}}(\widehat{M}, \pi), \quad \forall \pi \text{ s.t. } d(\pi, \pi_{\text{ref}}) \leq \delta \quad (\text{R1})$$

Here $d(\cdot, \cdot)$ is a function that measures the closeness of two policies (that will be chosen later in alignment with the choice of D .) We also note that the bound $D_{\pi_{\text{ref}}}(\widehat{M}, \pi)$ depends on the choice of neighborhood size δ but we omit a subscript for δ for simplicity. We will require our discrepancy bound to vanish when \widehat{M} is an accurate model:

$$\widehat{M} = M^* \implies D_{\pi_{\text{ref}}}(\widehat{M}, \pi) = 0, \quad \forall \pi, \pi_{\text{ref}} \quad (\text{R2})$$

The third requirement for the discrepancy bound D is that it can be estimated and optimized in the sense that

$$D_{\pi_{\text{ref}}}(\widehat{M}, \pi) \text{ is of the form } \mathbb{E}_{\tau \sim \pi_{\text{ref}}, M^*} [f(\widehat{M}, \pi, \tau)] \quad (\text{R3})$$

where f is a known differentiable function. We can estimate such kind of discrepancy bounds for every π in the neighborhood of π_{ref} by sampling empirical trajectories $\tau^{(1)}, \dots, \tau^{(n)}$ once from executing policy π_{ref} on the real environment M^* and compute the average of $f(\widehat{M}, \pi, \tau^{(i)})$'s. Note that we insist the expectation cannot be over the randomness of trajectories from π on M^* , because then we have to re-sample trajectories for every possible π encountered.

For example, one of the valid discrepancy bounds (under some assumptions) that we will prove in Section 4 is the error of the prediction of \widehat{M} on the trajectories from π_{ref} :

$$D_{\pi_{\text{ref}}}(\widehat{M}, \pi) = L \cdot \mathbb{E}_{S_0, \dots, S_t, \tau \sim \pi_{\text{ref}}, M^*} [\|\widehat{M}(S_t) - S_{t+1}\|] \quad (3.2)$$

Suppose we can establish such an discrepancy bound D (and the distance function d) with properties (R1), (R2), and (R3), — which will be the main focus of Section 4 —, then we can devise the following algorithmic framework as shown in Algorithm 1. We iteratively optimize the lower bound over the policy π_{k+1} and the model M_{k+1} , subject to the constraint that the policy is not very far from the reference policy π_k obtained in the previous iteration. For simplicity, we only state the population version with the exact computation of $D_{\pi_{\text{ref}}}(\widehat{M}, \pi)$.

Algorithm 1 General Algorithmic Framework

Inputs: Initial policy π_0 . Discrepancy bound D and distance function d that satisfy equation (R1) and (R2).

For $k = 0$ to T :

$$\pi_{k+1}, M_{k+1} = \underset{\pi \in \Pi, M \in \mathcal{M}}{\text{argmax}} \quad V^{\pi, M} - D_{\pi_k}(M, \pi) \quad (3.3)$$

$$\text{s.t. } d(\pi, \pi_k) \leq \delta \quad (3.4)$$

We remark that the discrepancy bound $D_{\pi_k}(M, \pi)$ in the objective plays the role of learning the dynamical model by ensuring the model to fit to the existing trajectories. For example, using the discrepancy bound is the form of equation (3.2), we recover the standard objective for model learning. Jointly optimizing M and π encourages the algorithm to choose the most optimistic model that can fit to the training data. The optimism compensates the conservatism in the lower bound and allows the algorithm to explore part of the space where the dynamical model are uncertainty with. Moreover, optimism-drive exploration can be more efficient than the random exploration only options that may lead to good rewards under some feasible dynamics are explored. We show formally that the expected reward in the real environment is non-decreasing under the assumption that the real dynamics belongs to our parameterized family \mathcal{M} .⁶

Theorem 3.1. *Suppose that $M^* \in \mathcal{M}$, that D and d satisfy equation (R1) and (R2), and the optimization problem in equation (3.3) is solvable at each iteration. Then, Algorithm 1 produces a sequence of policies π_0, \dots, π_T with monotonically increasing values:*

$$V^{\pi_0, M^*} \leq V^{\pi_1, M^*} \leq \dots \leq V^{\pi_T, M^*} \quad (3.5)$$

Moreover, as $k \rightarrow \infty$, the value V^{π_k, M^} converges to some $V^{\bar{\pi}, M^*}$, where $\bar{\pi}$ is a local maximum of V^{π, M^*} in domain Π .*

⁶We note that such an assumption, though restricted, may not be very far from reality: optimistically speaking, we only need to approximate the dynamical model accurately on the trajectories of the optimal policy. This might be much easier than approximating the dynamical model globally.

Proof. Since D and d satisfy equation (R1), we have that

$$V^{\pi_{k+1}, M^*} \geq V^{\pi_{k+1}, M_{k+1}} - D_{\pi_k}(M_{k+1}, \pi_{k+1})$$

By the definition that π_{k+1} and M_{k+1} are the optimizers of equation (3.3), we have that

$$V^{\pi_{k+1}, M_{k+1}} - D_{\pi_k}(M_{k+1}, \pi_{k+1}) \geq V^{\pi_k, M^*} - D_{\pi_k}(M^*, \pi_k) = V^{\pi_k, M^*} \quad (\text{by equation R2})$$

Combing the two equations above we complete the proof of equation (3.5).

For the second part of the theorem, by compactness, we have that a subsequence of π_k converges to some $\bar{\pi}$. By the monotonicity we have $V^{\pi_k, M^*} \leq V^{\bar{\pi}, M^*}$ for every $k \geq 0$. For the sake of contradiction, we assume $\bar{\pi}$ is not a local maximum, then in the neighborhood of $\bar{\pi}$ there exists π' such that $V^{\pi', M^*} > V^{\bar{\pi}, M^*}$ and $d(\bar{\pi}, \pi') < \delta/2$. Let t be such that π_t is in the $\delta/2$ -neighborhood of $\bar{\pi}$. Then we see that (π', M^*) is a better solution than (π_{t+1}, M_{t+1}) for the optimization in iteration t because $V^{\pi', M^*} > V^{\bar{\pi}, M^*} \geq V^{\pi_{t+1}, M^*} \geq V^{\pi_{t+1}, M_{t+1}} - D_{\pi_t}(M_{t+1}, \pi_t)$. (Here the last inequality uses equation (R1).) This contradicts the assumption that $\bar{\pi}$ is a local maximum. Therefore $\bar{\pi}$ is a local maximum and we complete the proof. \square

4 Discrepancy Bounds Design

In this section, we design discrepancy bounds that can provably satisfy the requirements (R1), (R2), and (R3). We design increasingly stronger discrepancy bounds from Section 4.1 to Section 4.3.

4.1 Norm-based prediction error bounds

In this subsection, we derive the discrepancy bound D of the form $\|\widehat{M}(S, A) - M^*(S, A)\|$ averaged over the observed state-action pair (S, A) under certain conditions on the dynamical model \widehat{M} . This suggests that we should not use the mean-squared error for learning the model, and instead, we should use the norm itself as the metric. In section 6, we will demonstrate that the ℓ_2 norm error consistently outperforms the square of ℓ_2 norm. Through the derivation, we will also introduce a telescoping lemma, which serves as the main building block towards other finer discrepancy bounds.

In this subsection, we make the (strong) assumption that the imaginary value function $V^{\pi, \widehat{M}}$ is L -Lipschitz w.r.t to some norm $\|\cdot\|$ in the sense that

$$\forall s, s' \in \mathcal{S}, |V^{\pi, \widehat{M}}(s) - V^{\pi, \widehat{M}}(s')| \leq L \cdot \|s - s'\| \quad (4.1)$$

In other words, nearby starting points should give similar rewards under the same policy π . We note that not every real environment M^* has this property, let alone imaginary dynamical models. However, once the real dynamical model gives a Lipschitz value function, we can penalize the Lipschitz-ness of the imaginary model during the training.

We start off with a lemma showing that the expected prediction error is an upper bound of the discrepancy between the real and imaginary values.

Lemma 4.1. *Suppose $V^{\pi, \widehat{M}}$ is L -Lipschitz (in the sense of (4.1)). Recall $\kappa = \gamma(1 - \gamma)^{-1}$. Then, we have*

$$|V^{\pi, \widehat{M}} - V^{\pi, M^*}| \leq \kappa L \mathbb{E}_{\substack{S \sim \rho^\pi \\ A \sim \pi(\cdot|S)}} [\|\widehat{M}(S, A) - M^*(S, A)\|] \quad (4.2)$$

However, in RHS in equation 4.2 cannot serve as a discrepancy bound because it doesn't satisfy the requirement (R3) — to optimize it over π we need to collect samples from every possible ρ^π , the state distribution of the policy π on the *real* model M^* . The main proposition of this subsection stated next shows that for every π in the neighborhood of a reference policy π_{ref} , we can replace the distribution ρ^π by a fixed distribution $\rho^{\pi_{\text{ref}}}$ with incurring only a higher order approximation. We use the expected KL divergence between two π and π_{ref} to define the neighborhood:

$$d^{\text{KL}}(\pi, \pi_{\text{ref}}) = \mathbb{E}_{S \sim \rho^{\pi_{\text{ref}}}} [KL(\pi(\cdot|S), \pi_{\text{ref}}(\cdot|S))^{1/2}] \quad (4.3)$$

Proposition 4.2. *In the same setting of Proposition 4.1, assume in addition that π is close to a reference policy π_{ref} in the sense that $d^{\text{KL}}(\pi, \pi_{\text{ref}}) \leq \delta$, and the states in S are uniformly bounded in the sense that $\|s\| \leq B, \forall s \in S$. Then,*

$$|V^{\pi, \widehat{M}} - V^{\pi, M^*}| \leq \kappa L \mathbb{E}_{\substack{S \sim \rho^{\pi_{\text{ref}}} \\ A \sim \pi(\cdot|S)}} \left[\|\widehat{M}(S, A) - M^*(S, A)\| \right] + 2\kappa^2 \delta B \quad (4.4)$$

In a benign scenario, the second term in the RHS of equation (4.4) should be dominated by the first term when the neighborhood size δ is sufficiently small. Moreover, the term B can also be replaced by $\max_{S,A} \|\widehat{M}(S, A) - M^*(S, A)\|$. The dependency on κ may not be tight but we note that most of the analysis of similar nature loses additional κ factor [32, 3]. Towards proving Propositions 4.2, we define the following quantity that captures the discrepancy between a dynamical model \widehat{M} and M^* for a single state-action pair (s, a) .

$$G^{\pi, \widehat{M}}(s, a) = V^{\pi, \widehat{M}}(\widehat{M}(s, a)) - V^{\pi, \widehat{M}}(M^*(s, a)) \quad (4.5)$$

We first give a telescoping lemma that decompose the discrepancy between $V^{\pi, M}$ and V^{π, M^*} into the expected single-step discrepancy G .

Lemma 4.3. *[Telescoping Lemma] Recall that $\kappa := \gamma(1 - \gamma)^{-1}$. For any policy π and dynamical models M, \widehat{M} , we have that*

$$V^{\pi, \widehat{M}} - V^{\pi, M} = \kappa \mathbb{E}_{\substack{S \sim \rho^{\pi, M} \\ A \sim \pi(\cdot|S)}} \left[G^{\pi, \widehat{M}}(S, A) \right] \quad (4.6)$$

The proof is reminiscent of the telescoping expansion in [15] (c.f. [32]) for characterizing the value difference of two policies, but we apply it to deal with the discrepancy between models. The detail is deferred to Supp. Materials. With the telescoping Lemma 4.3, Proposition 4.1 follows straightforwardly from Lipschitzness of the imaginary reward function. Proposition 4.2 follows from that ρ^π and $\rho^{\pi_{\text{ref}}}$ are close. We defer them to Supp. Materials.

4.2 Intrinsic error of the prediction

As alluded in the previous subsection, the prediction error based bounds doesn't apply when the value functions are not Lipschitz. Moreover, there are two limitations:

First, it's not clear, a priori, what is the right norm $\|\cdot\|$ that we should pick to measure the difference between the prediction $\widehat{M}(S, A)$ and the true next state $M^*(S, A)$. The correct one should be problem-dependent but state-representation invariant. Consider a scenario where some coordinates of the states S is redundant fundamentally: suppose the state S is consist of two blocks $[S_1, S_2]$ where S_1 is the intrinsic representation in the sense that the reward only depends on the coordinates in S_1 and the next state S'_1 also only depends S_1 . In this case, a good dynamical model should only care about the prediction of S_2 . However, with the generic norm based loss, the learning algorithm may spend unnecessary efforts in predicting S_2 . Thus, the optimal should metric take into account the intrinsic geometry in the state space induced by the dynamics.

Second, the error metric for the prediction should also depend on the state itself instead of only on the difference $\widehat{M}(S, A) - M^*(S, A)$. It's very possible that when S is at a critical position (e.g., when a robot is about to fall), the prediction error needs to be highly accurate so that the model \widehat{M} can be useful for planning. On the other hand, at other states, the dynamical model is allowed make bigger mistakes because they are not essential for the reward and can be corrected in an open loop system.

We propose the following discrepancy bound towards addressing the limitation above. Recall the definition of $G^{\pi, \widehat{M}}(s, a) = V^{\pi, \widehat{M}}(\widehat{M}(s, a)) - V^{\pi, \widehat{M}}(M^*(s, a))$ which measures the difference between $\widehat{M}(s, a)$ and $M^*(s, a)$ according to their imaginary rewards. We construct a discrepancy bound using the absolute value of G . Let's define ε_1 and ε_{\max} as the average of $|G^{\pi, \widehat{M}}|$ and its maximum.

$$\varepsilon_1 = \mathbb{E}_{S \sim \rho^{\pi_{\text{ref}}}} \left[|G^{\pi, \widehat{M}}(S)| \right] \text{ and } \varepsilon_{\max} = \max_S |G^{\pi, \widehat{M}}(S)| \quad (4.7)$$

We will show that the following discrepancy bound $D_{\pi_{\text{ref}}}^G(\widehat{M}, \pi)$ satisfies the property (R1), (R2).

$$D_{\pi_{\text{ref}}}^G(\widehat{M}, \pi) = \kappa \cdot \varepsilon_1 + \kappa^2 \delta \varepsilon_{\max} \quad (4.8)$$

Proposition 4.4. *Let d^{KL} and D^G be defined as in equation (4.3) and (4.8). Then the choice $d = d^{\text{KL}}$ and $D = D^G$ satisfies the basic requirements (equation (R1) and (R2)).*

The proof follows from the telescoping lemma (Lemma 4.3) and is deferred to Section B. We remark that the first term $\kappa \varepsilon_1$ can in principle be estimated and optimized approximately: the expectation be replaced by empirical samples from $\rho^{\pi_{\text{ref}}}$, and $G^{\pi, \widehat{M}}$ is an analytical function of π and \widehat{M} when they are both deterministic, and therefore can be optimized by back-propagation through time (BPTT). (When π and \widehat{M} are stochastic with a re-parameterizable noise such as Gaussian distribution [18], we can also use back-propagation to estimate the gradient.) The second term in equation (4.8) is difficult to optimize because it involves the maximum. However, it can be in theory considered as a second-order term because δ can be chosen to be a fairly small number. We note that the κ^2 dependency in the second term (4.8) could in theory force us to choose $\delta = 1/\kappa$, which is practically very conservative. To address this, we improve the dependency on both κ and δ in the next subsection with a stronger bound. (Empirically, we also found that the constraint (3.4) can be removed, which suggests that even the improved bound in the next subsection is far from tight.)

Besides BPTT, $G^{\pi, \widehat{M}}$ can also potentially be optimized in a actor-critic framework. Recall that G is precisely the difference between the imaginary critics on $\widehat{M}(S, A)$ and $M^*(S, A)$. Therefore, one could potentially build a critic from temporal difference learning and use the critic to penalize the learning of the dynamical model.

Attentive readers may notice that an adversarial dynamical model \widehat{M} can make $G^{\pi, \widehat{M}}$ equal to zero by constantly predicting some fixed states with constant reward (and therefore the imaginary value function is constant and G is zero.) However, optimism comes into play here: such ‘‘cheating’’ dynamical model admittedly can fool the discrepancy bound, but it won’t give high imaginary reward $V^{\pi, \widehat{M}}$. By optimizing $V - D$ over the model and policy, we discourage the algorithm to find such a cheating solution.

4.3 Refined bounds

The theoretical limitation of the discrepancy bound $D^G(\widehat{M}, \pi)$ is that the second term involving ε_{\max} is not rigorously optimizable by stochastic samples. In the worst case, there seem to exist situations where such infinity norm of $G^{\pi, \widehat{M}}$ is inevitable. In this section we tighten the discrepancy bounds with a different closeness measure d , χ^2 -divergence, in the policy space, and the dependency on the ε_{\max} is smaller (though not entirely removed.) We note that χ^2 -divergence has the same second order approximation as KL-divergence around the local neighborhood the reference policy and thus locally affects the optimization much.

We start by defining a re-weighted version β^π of the distribution ρ^π where examples in later step are slightly weighted up. We can effectively sample from β^π by importance sampling from ρ^π .

Definition 4.5. For a policy π , define β^π as the *re-weighted* version of discounted distribution of the states visited by π on M^* . Recall that $p_{S_t^\pi}$ is the distribution of the state at step t , we define $\beta^\pi = (1 - \gamma)^2 \sum_{t=1}^{\infty} t \gamma^{t-1} p_{S_t^\pi}$.

Then we are ready to state our discrepancy bound. Let

$$d^{\chi^2}(\pi, \pi_{\text{ref}}) = \max\left\{ \mathbb{E}_{S \sim \rho^{\pi_{\text{ref}}}} [\chi^2(\pi(\cdot|S), \pi_{\text{ref}}(\cdot|S))], \mathbb{E}_{S \sim \beta^{\pi_{\text{ref}}}} [\chi^2(\pi(\cdot|S), \pi_{\text{ref}}(\cdot|S))] \right\} \quad (4.9)$$

$$D_{\pi_{\text{ref}}}^{\chi^2}(\widehat{M}, \pi) = \kappa \varepsilon_1 + \kappa \delta \varepsilon_2 + \kappa^{3/2} \delta_1 \delta_2^{1/2} \varepsilon_{\max} \quad (4.10)$$

where $\varepsilon_2 = \mathbb{E}_{S \sim \beta^{\pi_{\text{ref}}}} [f^2]$ and $\varepsilon_1, \varepsilon_{\max}$ are defined in equation (4.7).

Proposition 4.6. *The discrepancy bound D^{χ^2} and closeness measure d^{χ^2} satisfies requirements (R1) and (R2).*

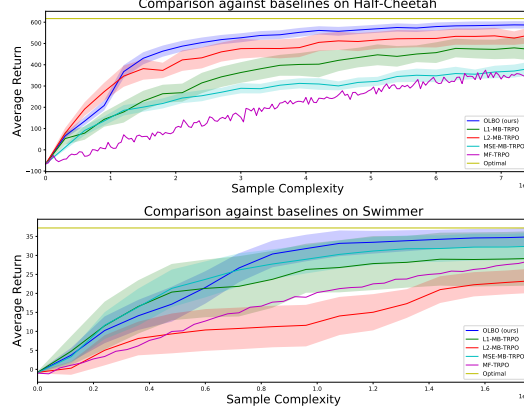


Figure 1: Performance on the benchmark Mujoco tasks. The x-axis shows the sample complexity (in the number of steps). The y-axis is the average return. Solid lines are average values over 5 random seeds for model-based method. Model-free learning curve is shown for reference. The initial 9e4 steps of OLBO uses regular L1-MB-TRPO. The optimal reward is computed by running model-free TRPO until convergence.

5 Implementing Optimism-driven Approach by Reduction to Standard RL

In Algorithm 1, we propose to maximize $V - D$ over the policy and the dynamical model. We have discussed the optimization of discrepancy bounds when define them. In this section we show that maximizing $V^{\pi, M}$ over the policy π and the model M can be reduced to a standard RL problem and therefore we can call any model-free RL algorithms.

The reduction works by designing a new environment \tilde{M} and new policy $\tilde{\pi}$ so that optimizing $\tilde{\pi}$ only with environment \tilde{M} is equivalent to optimizing (π, M) . Concretely, we enlarge the action space to be $\tilde{\mathcal{A}} = \mathcal{A} \times \mathcal{S}$. The policy $\tilde{\pi}$ returns a concatenation of the real action produced by π and the next state, and the dynamical model \tilde{M} just read off the next state from action $\tilde{\pi}(S)$:

$$\forall s \in \mathcal{S}, \tilde{a} \in \mathcal{A} \times \mathcal{S}, \tilde{\pi}(s) \triangleq (\pi(s), M(s, \pi(s))) \text{ and } \tilde{M}(s, \tilde{a}) \triangleq \tilde{a}_S \quad (5.1)$$

Here \tilde{a}_S denotes the restriction of \tilde{a} into the second set of coordinates w.r.t. the space \mathcal{S} .⁷

We have that $(\tilde{M}, \tilde{\pi})$ is equivalent to (M, π) in terms of the distributions of the states, and \tilde{M} is fixed dynamics. Therefore, optimizing $\tilde{\pi}$ is equivalent to optimizing (M, π) . Moreover, we note that such a translation preserves most, if not all, properties of the parameterization of π and M : if π and M are deterministic and differentiable, then so is $\tilde{\pi}$; if stochastic π and M can be efficiently sampled, or are re-parameterizable [18], so is $\tilde{\pi}$; if the density of π and M can be evaluated, then so is $\tilde{\pi}(\tilde{a}|s) = \pi(\tilde{a}_A|s) \cdot M(\tilde{a}_S|s, \tilde{a}_A)$;

6 Experiments

We provide proof-of-concepts evaluations for the theory developed in the previous sections. We first demonstrate that our proposed algorithm, Optimistic Lower Bound Optimization (OLBO), which uses the norm-based discrepancy bound developed in Section 4.1 in the Framework 1, outperforms (in sample complexity) standard model-based RL baselines in swimmer and half-cheetach benchmark in mujoco. Second, we apply the discrepancy bound in Section 4.2 and show that it helps the performance in an artificial noisy half-cheetach environment, which indicates the discrepancy bounds that are more robust to the change of state representations and thus may serve as a better learning objective for the dynamical model. Environment specs are in Appendix E.1.

Baselines The baselines include MSE-MB-TRPO, L1-MB-TRPO and L2-MB-TRPO, where the algorithms iteratively: a) learn the dynamical model with loss function MSE, L1 and L2 norm respectively, using the existing trajectories, b) policy optimization with the estimated dynamical

⁷When π is stochastic, the two occurrences of $\pi(s)$ in equation (5.1) are defined to use the same randomness.

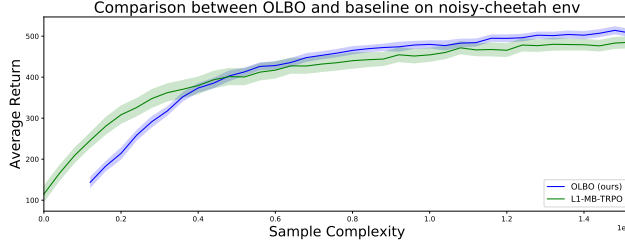


Figure 2: Performance on noisy-cheetah environment. The discrepancy loss helps the agent to focus more on the useful states from noisy redundant states. All the experiments are run on 4 random seeds. It takes 120K data points to build the critic and therefore the curve starts at 120K.

model using TRPO [32], and c) collect more samples from the current policy. We call each outer ‘model-TRPO-data’ iteration a *stage* for simplicity and convenience.

We note that the standard algorithms [28, 20] use mean-squared error (MSE) as the loss function. Our theory in Section 4.1 indicates that the norm (instead of the square of the norm) is a better learning objective and therefore we include L1-MB-TRPO and L2-MB-TRPO as baseline for completeness. Indeed, using the norm (either L1 or L2) as learning objective improves the performance. We perform these three simple but robust baselines with similar hyperparameter settings as detailed in Appendix.

OLBO on the benchmark We implement OLBO following the method in Sec.5. In each stage we optimize lower bound (3.3) (with discrepancy bounds being L1 as justified in Section 4.1) with respect to the model and the policy and collect the trajectories from the current policy. To optimize the lower bound (3.3), we alternate between maximizing the $V^{\pi, \hat{M}}$ term using TRPO w.r.t the model and the policy (sing the reduction method in Section 5) and minimizing the term $D(\hat{M}, \pi)$ with respect to the model stochastically.⁸ Empirically we found the constraint (3.4) is not necessary and therefore drop it in the implementation. Pseudo-codes for OLBO and the baselines can be found in algorithmic boxes 2 and 3.

We warm-start our algorithm after running 3 stages of L1-MB-TRPO baseline (in which $9e4$ steps of real rollouts are used). In Fig. 1, we find that all the model-based methods are more data efficient in the early stage of training compared to model-free TRPO. Only OLBO achieves near optimal performance with comparably small amount of data. We believe that the benefits mostly comes from the robust perspective of the algorithm: the parameter of the estimated model and the policy network are jointly optimized so that the policy have lower chance to overfit a fixed estimated model.⁹

Discrepancy Loss on Noisy Environment We test the proposed discrepancy loss (4.8) on an artificial cheetah environment to test whether our algorithm is more robust to different state representation. We enlarge the state space by 10 coordinates and add noise in them. We perform the L1-MB-TRPO baseline and train a critic network following the setting in DDPG [24] to estimate $V^{\pi, \hat{M}}$ that appear in the discrepancy bound. After 3 stages, we fix the critic network parameters. Then, we train from scratch by using a linear combination of the the discrepancy loss (4.8) and the L1 objective with a coefficient $1e-3$ as the discrepancy bound in (3.3). We find that adding this term will constantly improve the learned model and thus achieve higher rewards as shown in Figure 2. In practice, the critic learning is unstable especially with the estimated model. We slowdown the soft update rate 100 times. See Appendix E for more implementation details.

7 Conclusions

We design a novel algorithmic framework for designing and analyzing model-based RL algorithms with the guarantee to convergence monotonically to a local maximum of the reward. Experimental results show that our proposed algorithm (OLBO) achieve near-optimal reward on the mujoco benchmarks swimmer and half cheetah, with much fewer samples than other model-based baselines and model-free RL algorithms.

⁸Note that the norm discrepancy bound doesn’t depend on the policy.

⁹We empirically found that maximizing the reward $V^{\pi, \hat{M}}$ only over π also works, as long as the two terms in objective (3.3) are optimized alternatively.

A compelling empirical open question is whether pure model-based RL can achieve near-optimal reward on other more complicated mujoco tasks or real-world robotic tasks. We believe that understanding the trade-off between optimism and robustness is essential for designing more sample-efficient algorithms. In our theory, we assume that the parameterized model class contains the true dynamical model. Removing this assumption is also another interesting open question.

Acknowledgments:

We’d like to thank Emma Brunskill, Chelsea Finn, Haoran Tang and Yuping Luo for many helpful comments.

References

- [1] Yasin Abbasi-Yadkori and Csaba Szepesvári. Regret bounds for the adaptive control of linear quadratic systems. In *Proceedings of the 24th Annual Conference on Learning Theory*, pages 1–26, 2011.
- [2] Pieter Abbeel, Morgan Quigley, and Andrew Y Ng. Using inaccurate models in reinforcement learning. In *Proceedings of the 23rd international conference on Machine learning*, pages 1–8. ACM, 2006.
- [3] Joshua Achiam, David Held, Aviv Tamar, and Pieter Abbeel. Constrained policy optimization. *arXiv preprint arXiv:1705.10528*, 2017.
- [4] Ross Boczar, Nikolai Matni, and Benjamin Recht. Finite-data performance guarantees for the output-feedback control of an unknown system. *arXiv preprint arXiv:1803.09186*, 2018.
- [5] Thomas M Cover and Joy A Thomas. *Elements of information theory*. John Wiley & Sons, 2012.
- [6] Sarah Dean, Horia Mania, Nikolai Matni, Benjamin Recht, and Stephen Tu. On the sample complexity of the linear quadratic regulator. *arXiv preprint arXiv:1710.01688*, 2017.
- [7] Sarah Dean, Horia Mania, Nikolai Matni, Benjamin Recht, and Stephen Tu. Regret bounds for robust adaptive control of the linear quadratic regulator. *arXiv preprint arXiv:1805.09388*, 2018.
- [8] Marc Deisenroth and Carl E Rasmussen. Pilco: A model-based and data-efficient approach to policy search. In *Proceedings of the 28th International Conference on machine learning (ICML-11)*, pages 465–472, 2011.
- [9] Marc Peter Deisenroth, Gerhard Neumann, Jan Peters, et al. A survey on policy search for robotics. *Foundations and Trends® in Robotics*, 2(1–2):1–142, 2013.
- [10] Marc Peter Deisenroth, Carl Edward Rasmussen, and Dieter Fox. Learning to control a low-cost manipulator using data-efficient reinforcement learning. 2011.
- [11] Yan Duan, Xi Chen, Rein Houthooft, John Schulman, and Pieter Abbeel. Benchmarking deep reinforcement learning for continuous control. In *International Conference on Machine Learning*, pages 1329–1338, 2016.
- [12] Vladimir Feinberg, Alvin Wan, Ion Stoica, Michael I Jordan, Joseph E Gonzalez, and Sergey Levine. Model-based value estimation for efficient model-free reinforcement learning. *arXiv preprint arXiv:1803.00101*, 2018.
- [13] Shixiang Gu, Timothy Lillicrap, Ilya Sutskever, and Sergey Levine. Continuous deep q-learning with model-based acceleration. In *International Conference on Machine Learning*, pages 2829–2838, 2016.
- [14] K Jetat Hunt, D Sbarbaro, R Żbikowski, and Peter J Gawthrop. Neural networks for control systems—a survey. *Automatica*, 28(6):1083–1112, 1992.
- [15] Sham Kakade and John Langford. Approximately optimal approximate reinforcement learning. In *ICML*, volume 2, pages 267–274, 2002.

- [16] Gabriel Kalweit and Joschka Boedecker. Uncertainty-driven imagination for continuous deep reinforcement learning. In *Conference on Robot Learning*, pages 195–206, 2017.
- [17] S Mohammad Khansari-Zadeh and Aude Billard. Learning stable nonlinear dynamical systems with gaussian mixture models. *IEEE Transactions on Robotics*, 27(5):943–957, 2011.
- [18] Diederik P Kingma and Max Welling. Auto-encoding variational bayes. *arXiv preprint arXiv:1312.6114*, 2013.
- [19] Jonathan Ko and Dieter Fox. Gp-bayesfilters: Bayesian filtering using gaussian process prediction and observation models. *Autonomous Robots*, 27(1):75–90, 2009.
- [20] Thanard Kurutach, Ignasi Clavera, Yan Duan, Aviv Tamar, and Pieter Abbeel. Model-ensemble trust-region policy optimization. *arXiv preprint arXiv:1802.10592*, 2018.
- [21] Sergey Levine and Pieter Abbeel. Learning neural network policies with guided policy search under unknown dynamics. In *Advances in Neural Information Processing Systems*, pages 1071–1079, 2014.
- [22] Sergey Levine, Chelsea Finn, Trevor Darrell, and Pieter Abbeel. End-to-end training of deep visuomotor policies. *The Journal of Machine Learning Research*, 17(1):1334–1373, 2016.
- [23] Sergey Levine and Vladlen Koltun. Guided policy search. In *International Conference on Machine Learning*, pages 1–9, 2013.
- [24] Timothy P Lillicrap, Jonathan J Hunt, Alexander Pritzel, Nicolas Heess, Tom Erez, Yuval Tassa, David Silver, and Daan Wierstra. Continuous control with deep reinforcement learning. *arXiv preprint arXiv:1509.02971*, 2015.
- [25] Rudolf Lioutikov, Alexandros Paraschos, Jan Peters, and Gerhard Neumann. Sample-based informational-theoretic stochastic optimal control. In *Robotics and Automation (ICRA), 2014 IEEE International Conference on*, pages 3896–3902. IEEE, 2014.
- [26] Volodymyr Mnih, Koray Kavukcuoglu, David Silver, Andrei A Rusu, Joel Veness, Marc G Bellemare, Alex Graves, Martin Riedmiller, Andreas K Fidjeland, Georg Ostrovski, et al. Human-level control through deep reinforcement learning. *Nature*, 518(7540):529, 2015.
- [27] Jun Morimoto and Christopher G Atkeson. Minimax differential dynamic programming: An application to robust biped walking. In *Advances in neural information processing systems*, pages 1563–1570, 2003.
- [28] Anusha Nagabandi, Gregory Kahn, Ronald S Fearing, and Sergey Levine. Neural network dynamics for model-based deep reinforcement learning with model-free fine-tuning. *arXiv preprint arXiv:1708.02596*, 2017.
- [29] Frank Nielsen and Richard Nock. On the chi square and higher-order chi distances for approximating f-divergences. *IEEE Signal Processing Letters*, 21(1):10–13, 2014.
- [30] Vitchyr Pong, Shixiang Gu, Murtaza Dalal, and Sergey Levine. Temporal difference models: Model-free deep rl for model-based control. *arXiv preprint arXiv:1802.09081*, 2018.
- [31] Igal Sason and Sergio Verdú. f -divergence inequalities. *IEEE Transactions on Information Theory*, 62(11):5973–6006, 2016.
- [32] John Schulman, Sergey Levine, Pieter Abbeel, Michael Jordan, and Philipp Moritz. Trust region policy optimization. In *International Conference on Machine Learning*, pages 1889–1897, 2015.
- [33] John Schulman, Philipp Moritz, Sergey Levine, Michael Jordan, and Pieter Abbeel. High-dimensional continuous control using generalized advantage estimation. *arXiv preprint arXiv:1506.02438*, 2015.
- [34] John Schulman, Filip Wolski, Prafulla Dhariwal, Alec Radford, and Oleg Klimov. Proximal policy optimization algorithms. *arXiv preprint arXiv:1707.06347*, 2017.

- [35] David Silver, Julian Schrittwieser, Karen Simonyan, Ioannis Antonoglou, Aja Huang, Arthur Guez, Thomas Hubert, Lucas Baker, Matthew Lai, Adrian Bolton, et al. Mastering the game of go without human knowledge. *Nature*, 550(7676):354, 2017.
- [36] Max Simchowitz, Horia Mania, Stephen Tu, Michael I Jordan, and Benjamin Recht. Learning without mixing: Towards a sharp analysis of linear system identification. *arXiv preprint arXiv:1802.08334*, 2018.
- [37] Richard S Sutton. Integrated architectures for learning, planning, and reacting based on approximating dynamic programming. In *Machine Learning Proceedings 1990*, pages 216–224. Elsevier, 1990.
- [38] Richard S Sutton. Dyna, an integrated architecture for learning, planning, and reacting. *ACM SIGART Bulletin*, 2(4):160–163, 1991.
- [39] Richard S Sutton, Csaba Szepesvári, Alborz Geramifard, and Michael P Bowling. Dyna-style planning with linear function approximation and prioritized sweeping. *arXiv preprint arXiv:1206.3285*, 2012.
- [40] Aviv Tamar, Dotan Di Castro, and Ron Meir. Integrating a partial model into model free reinforcement learning. *Journal of Machine Learning Research*, 13(Jun):1927–1966, 2012.
- [41] Voot Tangkaratt, Syogo Mori, Tingting Zhao, Jun Morimoto, and Masashi Sugiyama. Model-based policy gradients with parameter-based exploration by least-squares conditional density estimation. *Neural networks*, 57:128–140, 2014.
- [42] Emanuel Todorov, Tom Erez, and Yuval Tassa. Mujoco: A physics engine for model-based control. In *Intelligent Robots and Systems (IROS), 2012 IEEE/RSJ International Conference on*, pages 5026–5033. IEEE, 2012.
- [43] Michael C Yip and David B Camarillo. Model-less feedback control of continuum manipulators in constrained environments. *IEEE Transactions on Robotics*, 30(4):880–889, 2014.

A Proof of Lemma 4.3

Proof of Lemma 4.3. Let W_j be the cumulative reward when we use dynamical model M for j steps and then \widehat{M} for the rest of the steps, that is,

$$W_j = \mathbb{E}_{\substack{\forall t \geq 0, A_t \sim \pi(\cdot | S_t) \\ \forall j > t \geq 0, S_{t+1} \sim M(\cdot | S_t, A_t) \\ \forall t \geq j, S_{t+1} \sim \widehat{M}(\cdot | S_t, A_t)}} \left[\sum_{t=0}^{\infty} \gamma^t R(S_t, A_t) \mid S_0 = s \right] \quad (\text{A.1})$$

By definition, we have that $W_{\infty} = V^{\pi, M}(s)$ and $W_0 = V^{\pi, \widehat{M}}(s)$. Then, we decompose the target into a telescoping sum,

$$V^{\pi, M}(s) - V^{\pi, \widehat{M}}(s) = \sum_{j=0}^{\infty} (W_{j+1} - W_j) \quad (\text{A.2})$$

Now we re-write each of the summands $W_{j+1} - W_j$. Comparing the trajectory distribution in the definition of W_{j+1} and W_j , we see that they only differ in the dynamical model applied in j -th step. Concretely, W_j and W_{j+1} can be rewritten as $W_j = R + \mathbb{E}_{S_j, A_j \sim \pi, M} \left[\mathbb{E}_{\hat{S}_{j+1} \sim \widehat{M}(\cdot | S_j, A_j)} \left[\gamma^{j+1} V^{\pi, \widehat{M}}(\hat{S}_{j+1}) \right] \right]$ and $W_{j+1} = R + \mathbb{E}_{S_j, A_j \sim \pi, M^*} \left[\mathbb{E}_{S_{j+1} \sim M(\cdot | S_j, A_j)} \left[\gamma^{j+1} V^{\pi, \widehat{M}}(S_{j+1}) \right] \right]$ where R denotes the reward from the first j steps from policy π and model M^* . Canceling the shared term in the two equations above, we get

$$W_{j+1} - W_j = \gamma^{j+1} \mathbb{E}_{S_j, A_j \sim \pi, M} \left[\mathbb{E}_{\substack{\hat{S}_{j+1} \sim \widehat{M}(\cdot | S_j, A_j) \\ S_{j+1} \sim M(\cdot | S_j, A_j)}} \left[V^{\pi, \widehat{M}}(S_{j+1}) - V^{\pi, \widehat{M}}(\hat{S}_{j+1}) \right] \right] \quad (\text{A.3})$$

Combining the equation above with equation (A.2) concludes that

$$V^{\pi, M} - V^{\pi, \widehat{M}} = \frac{\gamma}{1 - \gamma} \mathbb{E}_{S \sim \rho^{\pi}, A \sim \pi(S)} \left[V^{\pi, \widehat{M}}(M(S, A)) - V^{\pi, \widehat{M}}(\widehat{M}(S, A)) \right] \quad (\text{A.4})$$

□

B Missing Proofs in Section 4

Proof of Proposition 4.4. By Lemma 4.3 and triangle inequality, we have that

$$\begin{aligned} \frac{1 - \gamma}{\gamma} |V^{\pi, M} - V^{\pi, \widehat{M}}| &\leq \mathbb{E}_{S \sim \rho^{\pi}} \left[|G^{\pi, \widehat{M}}(S)| \right] && (\text{triangle inequality}) \\ &\leq \mathbb{E}_{S \sim \rho^{\pi_{\text{ref}}}} \left[|G^{\pi, \widehat{M}}(S)| \right] + |\rho^{\pi} - \rho^{\pi_{\text{ref}}}|_1 \cdot \max_S |G^{\pi, \widehat{M}}(S)| && (\text{Holder inequality}) \end{aligned}$$

By Corollary D.7 we have that $|\rho^{\pi} - \rho^{\pi_{\text{ref}}}|_1 \leq \frac{\gamma}{1 - \gamma} \mathbb{E}_{S \sim \rho^{\pi_{\text{ref}}}} [KL(\pi(S), \pi_{\text{ref}}(S))^{1/2} | S] = \frac{\delta \gamma}{1 - \gamma}$. Combining this with the equation above, we complete the proof. □

Proof of Proposition 4.6. Let μ be the distribution of the initial state S_0 , and let P' and P be the state-to-state transition kernel under policy π and π_{ref} . Let $\bar{G} = (1 - \gamma) \sum_{k=0}^{\infty} \gamma^k P^k$ and $\bar{G}' = (1 - \gamma) \sum_{k=0}^{\infty} \gamma^k P'^k$. Under these notations, we can re-write $\rho^{\pi_{\text{ref}}} = \bar{G}\mu$ and $\rho^{\pi} = \bar{G}'\mu$. Moreover, we observe that $\beta^{\pi_{\text{ref}}} = \bar{G}P\bar{G}\mu$.

Let $\delta_1 = (1 - \gamma) \chi_{\bar{G}\mu}^2(P', P)^{1/2}$ and $\delta_2 = (1 - \gamma) \chi_{\bar{G}P\bar{G}\mu}^2(P', P)^{1/2}$ by the χ^2 divergence between P' and P , measured with respect to distributions $\bar{G}\mu = \rho^{\pi_{\text{ref}}}$ and $\bar{G}P\bar{G}\mu = \beta^{\pi_{\text{ref}}}$. By Lemma C.1,

we have that the χ^2 -divergence between the states can be bounded by the χ^2 -divergence between the actions in the sense that:

$$\chi_{\bar{G}\mu}^2(P', P)^{1/2} = \chi_{\rho^{\pi_{\text{ref}}}}^2(P', P)^{1/2} \leq \mathbb{E}_{S \sim \rho^{\pi_{\text{ref}}}} [\chi^2(\pi(\cdot|S), \pi_{\text{ref}}(\cdot|S))] \quad (\text{B.1})$$

$$\chi_{\bar{G}P\bar{G}\mu}^2(P', P)^{1/2} = \chi_{\beta^{\pi_{\text{ref}}}}^2(P', P)^{1/2} \leq \mathbb{E}_{S \sim \beta^{\pi_{\text{ref}}}} [\chi^2(\pi(\cdot|S), \pi_{\text{ref}}(\cdot|S))] \quad (\text{B.2})$$

Therefore we obtain that $\delta_1 \leq (1 - \gamma)\delta$, $\delta_2 \leq (1 - \gamma)\delta$. Let $f(s) = G^{\pi, \widehat{M}}(s)$. We can control the difference between $\langle \rho^{\pi_{\text{ref}}}, f \rangle$ and $\langle \rho^{\pi}, f \rangle$ by

$$\left| \mathbb{E}_{S \sim \rho^{\pi_{\text{ref}}}} \left[G^{\pi, \widehat{M}}(S) \right] - \mathbb{E}_{S \sim \rho^{\pi}} \left[G^{\pi, \widehat{M}}(S) \right] \right| = |\langle \rho^{\pi_{\text{ref}}}, f \rangle - \langle \rho^{\pi}, f \rangle| \quad (\text{B.3})$$

$$\leq (1 - \gamma)^{-1} \left(\delta_1 \langle \bar{G}P\bar{G}\mu, f^2 \rangle^{1/2} + \delta_1 \delta_2^{1/2} \|f\|_{\infty} \right) \quad (\text{B.4})$$

$$= (1 - \gamma)^{-1} \left(\delta_1 \varepsilon_2 + \delta_1 \delta_2^{1/2} \varepsilon_{\max} \right) \quad (\text{B.5})$$

$$\leq \delta \varepsilon_2 + (1 - \gamma)^{-1/2} \delta_1 \delta_2^{1/2} \varepsilon_{\max} \quad (\text{B.6})$$

It follows that

$$\begin{aligned} |V^{\pi, \widehat{M}} - V^{\pi, M}| &\leq \gamma(1 - \gamma)^{-1} \left| \mathbb{E}_{S \sim \rho^{\pi}} \left[G^{\pi, \widehat{M}}(S) \right] \right| \quad (\text{by Lemma 4.3}) \\ &\leq \gamma(1 - \gamma)^{-1} \left(\left| \mathbb{E}_{S \sim \rho^{\pi_{\text{ref}}}} \left[G^{\pi, \widehat{M}}(S) \right] \right| + \left| \mathbb{E}_{S \sim \rho^{\pi_{\text{ref}}}} \left[G^{\pi, \widehat{M}}(S) \right] - \mathbb{E}_{S \sim \rho^{\pi}} \left[G^{\pi, \widehat{M}}(S) \right] \right| \right) \\ &\leq \gamma(1 - \gamma)^{-1} \left| \mathbb{E}_{S \sim \rho^{\pi_{\text{ref}}}} \left[G^{\pi, \widehat{M}}(S) \right] \right| + \gamma(1 - \gamma)^{-1} \delta \varepsilon_2 + \gamma(1 - \gamma)^{-3/2} \delta_1 \delta_2^{1/2} \varepsilon_{\max} \end{aligned}$$

□

Proof of Proposition 4.1 and 4.2 . By definition of G and the Lipschitzness of $V^{\pi, \widehat{M}}$, we have that $|G^{\pi, \widehat{M}}(s, a)| \leq L|\widehat{M}(s, a) - M^*(s, a)|$. Then, by Lemma 4.3 and triangle inequality, we have that

$$\begin{aligned} |V^{\pi, \widehat{M}} - V^{\pi, M^*}| &= \kappa \cdot \left| \mathbb{E}_{\substack{S \sim \rho^{\pi, M} \\ A \sim \pi(\cdot|S)}} \left[G^{\pi, \widehat{M}}(S, A) \right] \right| \leq \kappa \mathbb{E}_{\substack{S \sim \rho^{\pi, M} \\ A \sim \pi(\cdot|S)}} \left[|G^{\pi, \widehat{M}}(S, A)| \right] \\ &\leq \kappa \mathbb{E}_{\substack{S \sim \rho^{\pi} \\ A \sim \pi(\cdot|S)}} \left[\|\widehat{M}(S, A) - M^*(S, A)\| \right]. \end{aligned} \quad (\text{B.7})$$

Thus we proved Proposition 4.1. Note that for any distribution ρ and ρ' and function f , we have $\mathbb{E}_{S \sim \rho} f(S) = \mathbb{E}_{S \sim \rho'} f(S) + \langle \rho - \rho', f \rangle \leq \mathbb{E}_{S \sim \rho'} f(S) + \|\rho - \rho'\|_1 \|f\|_{\infty}$. Thus applying this inequality with $f(S) = \mathbb{E}_{A \sim \pi(\cdot|S)} [\|\widehat{M}(S, A) - M^*(S, A)\|]$, we obtain that

$$\begin{aligned} \mathbb{E}_{\substack{S \sim \rho^{\pi} \\ A \sim \pi(\cdot|S)}} \left[\|\widehat{M}(S, A) - M^*(S, A)\| \right] &\leq \mathbb{E}_{\substack{S \sim \rho^{\pi_{\text{ref}}} \\ A \sim \pi(\cdot|S)}} \left[\|\widehat{M}(S, A) - M^*(S, A)\| \right] \\ &\quad + \|\rho^{\pi_{\text{ref}}} - \rho\|_1 \max_S \mathbb{E}_{A \sim \pi(\cdot|S)} \left[\|\widehat{M}(S, A) - M^*(S, A)\| \right] \\ &\leq \mathbb{E}_{\substack{S \sim \rho^{\pi_{\text{ref}}} \\ A \sim \pi(\cdot|S)}} \left[\|\widehat{M}(S, A) - M^*(S, A)\| \right] + 2\delta \kappa B \end{aligned} \quad (\text{B.8})$$

where the last inequality uses the inequalities (see Corollary D.7) that $\|\rho^{\pi} - \rho^{\pi_{\text{ref}}}\|_1 \leq \frac{\gamma}{1-\gamma} \mathbb{E}_{S \sim \rho^{\pi_{\text{ref}}}} [KL(\pi(S), \pi_{\text{ref}}(S))^{1/2}|S] = \delta \kappa$ and that $\|\widehat{M}(S, A) - M^*(S, A)\| \leq 2B$. Combining (B.8) and (B.7) we complete the proof of Proposition 4.2. □

C χ^2 -Divergence Based Inequalities

Lemma C.1. *Let S be a random variable over the domain \mathcal{S} . Let π and π' be two policies and $A \sim \pi(\cdot | S)$ and $A' \sim \pi'(\cdot | S)$. Let $Y \sim M(\cdot | S, A)$ and $Y' \sim M(\cdot | S, A')$ be the random variables for the next states under two policies. Then,*

$$\mathbb{E} [\chi^2(Y|S, Y'|S)] \leq \mathbb{E} [\chi^2(A|S, A'|S)] \quad (\text{C.1})$$

Proof. By definition, we have that $Y|S = s, A = a$ has the same density as $Y'|S = s, A' = a$ for any a and s . Therefore by Theorem D.4 (setting X, X', Y, Y' in Theorem D.4 by $A|S = s, A'|S = s, Y|S = s, Y'|S = s$ respectively), we have

$$\chi^2(Y|S = s, Y'|S = s) \leq \chi^2(A|S = s, A'|S = s) \quad (\text{C.2})$$

Taking expectation over the randomness of S we complete the proof. \square

C.1 Properties of Markov Processes

In this subsection, we consider bounded the difference of the distributions induced by two markov process starting from the same initial distributions μ . Let P, P' be two transition kernels. Let $G = \sum_{k=0}^{\infty} \gamma^k P^k$ and $\bar{G} = (1 - \gamma)G$. Define G' and \bar{G}' similarly. Therefore we have that $\bar{G}\mu$ is the discounted distribution of states visited by the markov process starting from distribution μ . In other words, if μ is the distribution of S_0 , and P is the transition kernel induced by some policy π , then $\bar{G}\mu = \rho^\pi$.

First of all, let $\Delta = \gamma(P' - P)$ and we note that with simple algebraic manipulation,

$$\bar{G}' - \bar{G} = (1 - \gamma)^{-1} \bar{G}' \Delta \bar{G} \quad (\text{C.3})$$

Let f be some function. We will mostly interested in the difference between $\mathbb{E}_{S \sim \bar{G}\mu} [f]$ and $\mathbb{E}_{S \sim \bar{G}'\mu} [f]$, which can be rewritten as $\langle (\bar{G}' - \bar{G})\mu, f \rangle$. We will bound this quantity from above by some divergence measure between P' and P .

We start off with a simple lemma that controls the form $\langle p - q, f \rangle$ by the χ^2 divergence between p and q . With this lemma we can reduce our problem of bounding $\langle (\bar{G}' - \bar{G})\mu, f \rangle$ to characterizing the χ^2 divergence between $\bar{G}'\mu$ and $\bar{G}\mu$.

Lemma C.2. *Let p and q be probability distributions. Then we have*

$$\langle q - p, f \rangle^2 \leq \chi^2(q, p) \cdot \langle p, f^2 \rangle$$

Proof. By Cauchy-Schwartz inequality, we have

$$\langle q - p, f \rangle^2 \leq \left(\int \frac{(q(x) - p(x))^2}{p(x)} dx \right) \left(\int p(x) f(x)^2 \right) = \chi^2(q, p) \cdot \langle p, f^2 \rangle$$

\square

The following Lemma is a refinement of the lemma above. It deals with the distributions p and q with the special structure $p = WP'\mu$ and $q = WP\mu$.

Lemma C.3. *Let W, P', P be transition kernels and μ be a distribution. Then,*

$$\langle W(P' - P)\mu, f \rangle^2 \leq \chi_\mu^2(P', P) \langle WP\mu, f^2 \rangle \quad (\text{C.4})$$

where $\chi_\mu^2(P', P)$ is a divergence between transitions defined in Definition D.3.

Proof. By Lemma C.2 with $p = WP\mu$ and $q = WP'\mu$, we conclude that

$$\langle W(P' - P)\mu, f \rangle^2 \leq \chi^2(q, p) \cdot \langle p, f^2 \rangle \leq \chi^2(WP'\mu, WP\mu) \langle WP\mu, f^2 \rangle \quad (\text{C.5})$$

By Theorem D.4 and Theorem D.5 we have that $\chi^2(WP'\mu, WP\mu) \leq \chi^2(P'\mu, P\mu) \leq \chi_\mu^2(P', P)$, plugging this into the equation above we complete the proof. \square

Now we are ready to state the main result of this subsection.

Lemma C.4. *Let $\bar{G}, \bar{G}', P', P, f$ as defined in the beginning of this section. Let $\delta_1 = (1 - \gamma)\chi_{\bar{G}\mu}^2(P', P)^{1/2}$ and $\delta_2 = (1 - \gamma)\chi_{\bar{G}P\bar{G}\mu}^2(P', P)^{1/2}$. Then,*

$$(1 - \gamma) |\langle \bar{G}'\mu, f \rangle - \langle \bar{G}\mu, f \rangle| \leq \delta_1 \|f\|_\infty \quad (\text{C.6})$$

$$(1 - \gamma) |\langle \bar{G}'\mu, f \rangle - \langle \bar{G}\mu, f \rangle| \leq \delta_1 \langle \bar{G}P\bar{G}\mu, f^2 \rangle^{1/2} + \delta_1 \delta_2^{1/2} \|f\|_\infty \quad (\text{C.7})$$

Proof. Recall by equation (C.3), we have $(\bar{G}' - \bar{G})\mu = \langle \bar{G}'\Delta\bar{G}\mu, f \rangle$. By Lemma C.3,

$$\langle \bar{G}'\Delta\bar{G}\mu, f \rangle^2 \leq \chi_{\bar{G}\mu}^2(P', P) \langle \bar{G}'P\bar{G}\mu, f^2 \rangle \quad (\text{C.8})$$

Using equation (C.3) again, we have

$$\langle \bar{G}'P\bar{G}\mu, f^2 \rangle = \langle \bar{G}P\bar{G}\mu, f^2 \rangle + \frac{1}{1 - \gamma} \langle \bar{G}'\Delta\bar{G}P\bar{G}\mu, f^2 \rangle \quad (\text{C.9})$$

By Lemma C.3 again, we have that

$$\langle \bar{G}'\Delta\bar{G}P\bar{G}\mu, f^2 \rangle^2 \leq \chi_{\bar{G}P\bar{G}\mu}^2(P', P) \langle \bar{G}'P\bar{G}P\bar{G}\mu, f^4 \rangle \quad (\text{C.10})$$

By Holder inequality and the fact that $\|\bar{G}\|_{1 \rightarrow 1} = 1$, $\|\bar{G}'\|_{1 \rightarrow 1} = 1$ and $\|P\|_{1 \rightarrow 1} = 1$, we have

$$\langle \bar{G}'P\bar{G}P\bar{G}\mu, f^4 \rangle \leq \|\bar{G}'P\bar{G}P\bar{G}\mu\|_1 \|f^4\|_\infty \leq \|f\|_\infty^4 \quad (\text{C.11})$$

Then, combining equation (C.8), (C.9), (C.11), we have

$$\begin{aligned} (1 - \gamma) |\langle \bar{G}'\Delta\bar{G}\mu, f \rangle| &\leq \delta_1 \langle \bar{G}'P\bar{G}\mu, f^2 \rangle^{1/2} && (\text{by equation (C.8)}) \\ &\leq \delta_1 \langle \bar{G}P\bar{G}\mu, f^2 \rangle^{1/2} + \delta_1 \langle \bar{G}'\Delta\bar{G}P\bar{G}\mu, f^2 \rangle^{1/2} && (\text{by equation (C.9) and AM-GM}) \\ &\leq \delta_1 \langle \bar{G}P\bar{G}\mu, f^2 \rangle^{1/2} + \delta_1 \delta_2^{1/2} \|f\|_\infty && (\text{by equation (C.10) and (C.11)}) \end{aligned}$$

□

The following Lemma is a stronger extension of Lemma C.4, which can be used to future improve Proposition 4.6, and may be of other potential independent interests. We state it for completeness.

Lemma C.5. *Let $\bar{G}, \bar{G}', P', P, f$ as defined in the beginning of this section. Let $d_k = (\bar{G}P)^k \bar{G}\mu$ and $\delta_k = (1 - \gamma)\chi_{d_{k-1}}^2(P', P)^{1/2}$, then we have that for any K ,*

$$\begin{aligned} (1 - \gamma) |\langle \bar{G}'\mu, f \rangle - \langle \bar{G}\mu, f \rangle| &\leq \delta_1 \langle d_1, f^2 \rangle^{-1/2} + \delta_1 \delta_2^{-1/2} \langle d_2, f^4 \rangle^{-1/4} + \\ &\quad + \delta_1 \dots \delta_{K-1}^{2^{-K}} \langle d_K, f^{2^K} \rangle^{2^{-K}} + \delta_1 \dots \delta_{K-1}^{2^{-K}} \|f\|_\infty \end{aligned}$$

Proof. We first use induction to prove that:

$$(1 - \gamma) |\langle (\bar{G}' - \bar{G})\mu, f \rangle| \leq \sum_{k=1}^K \left(\prod_{0 \leq s \leq k-1} \delta_{s+1}^{2^{-s}} \right) \langle d_k, f^{2^k} \rangle^{2^{-k}} + \left(\prod_{0 \leq s \leq K-1} \delta_{s+1}^{2^{-s}} \right) \langle \bar{G}'\Delta(\bar{G}P)^K \bar{G}\mu, f^{2^K} \rangle^{2^{-K}} \quad (\text{C.12})$$

By Lemma C.4, we got the case for $K = 1$. Assuming we have prove the case for K , then applying

$$\begin{aligned} \langle \bar{G}'\Delta(\bar{G}P)^K \bar{G}\mu, f^{2^K} \rangle &= \langle \bar{G}\Delta(\bar{G}P)^K \bar{G}\mu, f^{2^K} \rangle + (1 - \gamma)^{-1} \langle \bar{G}'\Delta(\bar{G}P)^{K+1} \bar{G}\mu, f^{2^K} \rangle \\ &\leq \langle \bar{G}\Delta(\bar{G}P)^K \bar{G}\mu, f^{2^K} \rangle + (1 - \gamma)^{-1} \chi_{d_K}^2(P', P)^{1/2} \langle \bar{G}'\Delta(\bar{G}P)^{K+1} \bar{G}\mu, f^{2^{K+1}} \rangle^{1/2} \\ &\leq \langle \bar{G}\Delta(\bar{G}P)^K \bar{G}\mu, f^{2^K} \rangle + \delta_{K+1} \langle \bar{G}'\Delta(\bar{G}P)^{K+1} \bar{G}\mu, f^{2^{K+1}} \rangle^{1/2} \end{aligned} \quad (\text{C.13})$$

By Cauchy-Schwartz inequality, we obtain that

$$\langle \bar{G}'\Delta(\bar{G}P)^K \bar{G}\mu, f^{2^K} \rangle^{2^{-K}} \leq \langle \bar{G}\Delta(\bar{G}P)^K \bar{G}\mu, f^{2^K} \rangle^{2^{-K}} + \delta_{K+1} \langle \bar{G}'\Delta(\bar{G}P)^{K+1} \bar{G}\mu, f^{2^{K+1}} \rangle^{2^{-K-1}}$$

Plugging the equation above into equation (C.12), we provide the induction hypothesis for the case with $K + 1$.

Now applying $\langle \bar{G}'\Delta(\bar{G}P)^K \bar{G}\mu, f^{2^K} \rangle^{2^{-K}} \leq \|f\|_\infty$ with equation (C.12) we complete the proof.

□

D Toolbox

Definition D.1 (χ^2 distance, c.f. [29, 5]). The Neyman χ^2 distance between two distributions p and q is defined as

$$\chi^2(p, q) \triangleq \int \frac{(p(x) - q(x))^2}{q(x)} dx = \int \frac{p(x)^2}{q(x)} dx - 1 \quad (\text{D.1})$$

For notational simplicity, suppose two random variables X and Y has distributions p_X and p_Y , we often write $\chi^2(X, Y)$ as a simplification for $\chi^2(p_X, p_Y)$.

Theorem D.2 ([31]). *The Kullback-Leibler (KL) divergence between two distributions p, q is bounded from above by the χ^2 distance:*

$$KL(p, q) \leq \chi^2(p, q) \quad (\text{D.2})$$

Proof. Since log is a concave function, by Jensen inequality we have

$$\begin{aligned} KL(p, q) &= \int p(x) \log \frac{p(x)}{q(x)} dx \leq \log \int p(x) \cdot \frac{p(x)}{q(x)} dx \\ &= \log(\chi^2(p, q) + 1) \leq \chi^2(p, q) \end{aligned}$$

□

Definition D.3 (χ^2 distance between transitions). Given two transition kernels P, P' . For any distribution μ , we define $\chi_\mu^2(P', P)$ as:

$$\chi_\mu^2(P', P) \triangleq \int \mu(x) \chi^2(P'(\cdot|X=x), P(\cdot|X=x)) dx \quad (\text{D.3})$$

Theorem D.4. *Suppose random variables (X, Y) and (X', Y') satisfy that $p_{Y|X} = p_{Y'|X'}$. Then*

$$\chi^2(Y, Y') \leq \chi^2(X, X') \quad (\text{D.4})$$

Or equivalently, for any transition kernel P and distribution μ, μ' , we have

$$\chi^2(P\mu, P\mu') \leq \chi^2(\mu, \mu') \quad (\text{D.5})$$

Proof. Denote $p_{Y|X}(y|x) = p_{Y'|X'}(y|x)$ by $p(y|x)$, and we rewrite p_X as p and $p_{X'}$ as p' . By Cauchy-Schwarz inequality, we have:

$$\begin{aligned} p_Y(y)^2 &= \left(\int p(y|x)p(x) dx \right)^2 \leq \left(\int p(y|x)p'(x) dx \right) \left(\int p(y|x) \frac{p(x)^2}{p'(x)} dx \right) \\ &= p_{Y'}(y) \left(\int p(y|x) \frac{p(x)^2}{p'(x)} dx \right) \end{aligned} \quad (\text{D.6})$$

It follows that

$$\chi^2(Y, Y') = \int \frac{p_Y(y)^2}{p_{Y'}(y)} dy - 1 \leq \int dy \int p(y|x) \frac{p(x)^2}{p'(x)} dx - 1 = \chi^2(X, X')$$

□

Theorem D.5. *Let X, Y, Y' are three random variables. Then,*

$$\chi^2(Y, Y') \leq \mathbb{E} [\chi^2(Y|X, Y'|X)] \quad (\text{D.7})$$

We note that the expectation on the right hand side is over the randomness of X .¹⁰ As a direct corollary, we have for transition kernel P' and P and distribution μ ,

$$\chi^2(P'\mu, P\mu) \leq \chi_\mu^2(P', P) \quad (\text{D.8})$$

¹⁰Observe $\chi^2(Y|X, Y'|X)$ deterministically depends on X .

Proof. We denote $p_{Y'|X}(y|x)$ by $p'(y|x)$ and $p_{Y|X}(y|x)$ by $p(y|x)$, and let $p(x)$ be a simplification for $p_X(x)$. We have by Cauchy-Schwarz,

$$\frac{p_Y(y)^2}{p_{Y'}(y)} = \frac{(\int p(y|x)p(x)dx)^2}{\int p'(y|x)p(x)dx} \leq \int \frac{p(y|x)^2}{p'(y|x)} p(x)dx \quad (\text{D.9})$$

It follows that

$$\chi^2(Y, Y') = \int \frac{p_Y(y)^2}{p_{Y'}(y)} dy - 1 \leq \int \frac{p(y|x)^2}{p'(y|x)} p(x) dx dy - 1 = \mathbb{E} [\chi^2(Y|X, Y'|X) | X] \quad (\text{D.10})$$

□

Claim D.6. Let μ be a distribution over the state space \mathcal{S} . Let P and P' be two transition kernels. $G = \sum_{k=0}^{\infty} (\gamma P)^k = (\text{Id} - \gamma P)^{-1}$ and $G' = \sum_{k=0}^{\infty} (\gamma P')^k = (\text{Id} - \gamma P')^{-1}$. Let $d = (1 - \gamma)G\mu$ and $d' = (1 - \gamma)G'\mu$ be the discounted distribution starting from μ induced by the transition kernels G and G' . Then,

$$|d - d'|_1 \leq \frac{1}{1 - \gamma} |\Delta d|_1 \quad (\text{D.11})$$

Moreover, let $\gamma(P' - P) = \Delta$. Then, we have

$$G' - G = \sum_{k=1}^{\infty} (G\Delta)^k G \quad (\text{D.12})$$

Proof. With algebraic manipulation, we obtain,

$$\begin{aligned} G' - G &= (\text{Id} - \gamma P')^{-1} ((\text{Id} - \gamma P) - (\text{Id} - \gamma P')(\text{Id} - \gamma P)^{-1}) \\ &= G' \Delta G \end{aligned} \quad (\text{D.13})$$

It follows that

$$\begin{aligned} |d - d'|_1 &= (1 - \gamma) |G' \Delta G \mu|_1 \leq |\Delta G \mu|_1 \quad (\text{since } (1 - \gamma) |G'|_{1 \rightarrow 1} \leq 1) \\ &= \frac{1}{1 - \gamma} |\Delta d|_1 \end{aligned}$$

Replacing G' in the RHS of the equation (D.13) by $G' = G + G' \Delta G$, and doing this recursively gives

$$G' - G = \sum_{k=1}^{\infty} (G\Delta)^k G$$

□

Corollary D.7. Let π and π' be two policies and let ρ^π be defined as in Definition 2.1. Then,

$$|\rho^\pi - \rho'|_1 \leq \frac{\gamma}{1 - \gamma} \mathbb{E}_{S \sim \rho^\pi} \left[KL(\pi(S), \pi'(S))^{1/2} | S \right] \quad (\text{D.14})$$

Proof. Let P and P' be the state-state transition matrix under policy π and π' and $\Delta = \gamma(P' - P)$. By Claim D.6, we have that

$$|\rho^\pi - \rho^{\pi'}|_1 \leq \frac{1}{1 - \gamma} |\Delta \rho^\pi|_1 = \frac{\gamma}{1 - \gamma} \mathbb{E}_{S \sim \rho^\pi} [|p_{M^*(S, \pi(S))|S} - p_{M^*(S, \pi'(S))|S}|_1] \quad (\text{D.15})$$

$$\leq \frac{\gamma}{1 - \gamma} \mathbb{E}_{S \sim \rho^\pi} [|p_{\pi(S)|S} - p_{\pi'(S)|S}|_1] \quad (\text{D.16})$$

$$\leq \frac{\gamma}{1 - \gamma} \mathbb{E}_{S \sim \rho^\pi} \left[KL(\pi(S), \pi'(S))^{1/2} | S \right] \quad (\text{by Pinsker's inequality})$$

□

Table 1: Reward function definition in Swimmer and Half-cheetah environments

	Reward r_t
Swimmer	$s_t^{x_{vel}} - 0.5 \ \frac{a}{50}\ _2^2$
Half-cheetah	$s_t^{x_{vel}} - 0.05 \ \frac{a}{1}\ _2^2$

E Implementation Details

E.1 Environment Specs

We adopt Half-Cheetah and Swimmer environments from Mujoco simulator as our testing environment. The observation space is slightly modified by removing redundant dimensions. The action space is normalized to the space from -1 to 1. We use the standard reward as in rllab [11]. We set 200 as the maximum step number in each rollout. For the noisy-cheetah environment, we concatenate 10 more entries of unit Gaussian variable to the original states. This will make the environment harder to learn in practice.

E.2 Architecture

For all the experiments, we use a two layer fully connected network with 500 hidden units each layer to fit the dynamics of environments. We use another fully connected two-layer network with 32 hidden units per layer as policy network. To approximate the critic function, we use a two-layer network with 100 hidden units per layer. All the activation layers for model learning are using ReLU and for policy network tanh.

E.3 Data Preprocessing and Hyper-parameter Selection

E.3.1 Data Collection

We first define each *stage* means a full pipeline of data collection, model training and policy optimization. We populate the dataset with 60 rollouts in Swimmer and 150 rollouts in cheetah each *stage* that resulted from the execution of parameterized actions a_t from a randomly initialized exploratory policy network. Each rollout started from a selected starting state s_0 with small gaussian noise $\sim \mathcal{N}(0, 0.001)$. During data collection, we use the Ornstein-Uhlenbeck noise with $\sigma = 0.3$ and $\theta = 0.15$ to get an exploratory policy for more diverse trajectories. We set the upper bound scale for the dataset 5 times the amount collected in each *stage* and will drop the oldest ones when the dataset is full.

E.3.2 Imaginary Model Learning

We use Adam optimizer with 1e-3 as learning rate to learn the model. We normalize the input data for model learning with empirical mean and variance computed in the initial *stage*. The model outputs a normalized difference between previous input state and the next state. The batchsize is 512 state and action pairs every step. We train this network 40/min(n,5) epochs where n is the number of *stage* we are at. In each epoch, the network goes over the whole dataset once.

E.3.3 TRPO Hyperparameters

For TRPO we use the version provided by rllab [11] and use a batch size of 4000 every iteration and step size 0.01 for conjugate gradient optimization. Other hyperparameters stay the same as the online version. In each stage we train it for 100 steps. Note that more training steps will make trpo overfit the imaginary model heavily for baselines.

E.3.4 OLBO hyperparameters

We set $K = 80$, $k = 40$ and $m = 10$ in Algorithm 3.

Algorithm 2 L1-model-based-TRPO

θ : parameters for the estimated model; ϕ : parameters for policy network; m : the number of steps for TRPO; \tilde{M} : the estimated model; $\pi^{explore}$ the exploratory policy.

Initialize π_ϕ and \tilde{M}_θ ; A dataset $\mathcal{D} = \emptyset$

for stage $t \in \{1, 2, \dots\}$ **do**

1. Sample N trajectories with $\pi_\phi^{explore}$ and concatenate them in \mathcal{D} ; remove old trajectories if the size of \mathcal{D} exceeds size R .

2. Sample triplets from \mathcal{D} and update the parameters θ of model \tilde{M}_θ by performing k steps of stochastic gradient descents with L1 loss.

3. run TRPO with model \tilde{M}_θ as an environment for m steps.

end for

Algorithm 3 Optimistic Lower Bound Optimization (OLBO)

θ : parameters for the estimated model; ϕ : parameters for policy network; m : the number of steps for TRPO; \tilde{M} : the estimated model; $\pi^{explore}$ the exploratory policy.

Initialize π_ϕ and \tilde{M}_θ ; A dataset $\mathcal{D} = \emptyset$

for stage $t \in \{1, 2, \dots\}$ **do**

Sample N trajectories with $\pi_\phi^{explore}$ and concatenate them in \mathcal{D} ; remove old trajectories if the size of \mathcal{D} exceeds size R .

for $i \in \{1, 2, \dots, K\}$ **do**

Sample triplets from \mathcal{D} , and update the parameters θ of model \tilde{M}_θ by minimizing the discrepancy bounds D with stochastic gradient descent for k steps.

Jointly optimize model \tilde{M}_θ and policy π_ϕ with TRPO as described in 5 for m steps.

end for

end for
