Learning Value Equivalent Models on Policy Improvement Path

Shenao Zhang
Georgia Tech

Zhaoran Wang
Northwestern University

Tuo Zhao
Georgia Tech

1 Introduction

The basic intuition of value-aware model learning [4, 3] is to assign different weights to the transitions according to the *optimal* state value. However, the optimal value function is not known and the VAML loss cannot be readily minimized. Thus, as shown in [6, 5], we may find a model using the value equivalence principle, which is then used for value-based planning. As more value functions are augmented to be considered, the model will less compromise the performance, and eventually collapse to the real environment model. The key is to determine the representative value functions (or the value function of representative policies [5]).

In this work, we focus on model-based RL where the model is used as the simulator to calculate policy gradient. We show that learning value equivalent models through the policy improvement path gives a $T^{-\frac{1}{2}}$ rate of convergence. We first give convergence rates of policy gradient in terms of gradient bias and variance for finite MDPs with direct policy parameterization, and for more general function approximations settings with natural gradient. Then we provide the bounds for model-based likelihood ratio gradient estimator and pathwise gradient estimator, respectively. We show that for both methods, the model learner is equipped with the value equivalence objective in an online learning manner. Specifically, the online learning problem is to select the value functions that are used to obtain the PVE model. Then a loss that depends on how representative the selected value functions are is revealed. If the past policy values are representative enough, we may expect $\hat{f}_t$ to accurately describe the environment, and thus suffering a low loss, with which the bias of policy gradient is small.

For tabular settings and liner function approximations, the model learning process becomes a *convex* problem, where online gradient descent can be adopted to obatin an iterative style of VAML algorithm [3]. However, online GD is not enough for more general nonconvex settings. We bring the family of *follow the leader* algorithms to tackle this model learning problem. That is, vanilla FTL indicates choosing the past policies on the policy improvement path as representative ones, which also shares similarities with the implementations in [2, 5]. We also propose several principled representative policy selection procedure by considering its predictable nature. Then optimistic FTL can be leveraged to obtain tighter regret bound.

There are wide applications of value equivalence principle in modern deep RL agents, *e.g.*, MuZero [12], Predictron [13], VPN [9]. Besides, it is experimentally shown in [5] that MuZero integrating past policies outperforms MuZero. We extend these value-based planning MBRL to model-based policy gradient algorithms, with principled value equivalent model learners.

Work in progress.

2 Policy Gradient Convergence Rates

2.1 Finite MDPs

First consider discounted finite MDPs where projected gradient ascent on the direct policy parameterization is performed.

$$\text{direct policy parameterization:} \quad \pi_\theta(a|s) = \theta_{s,a}$$

$$\text{projected gradient ascent:} \quad \pi_{t+1} = P_{\Delta(\mathcal{A})^{|S|}}(\theta_t + \eta \nabla_\theta \hat{J}(\pi_{\theta_t}))$$

Here, $J(\pi) = \mathbb{E}_{s_0 \sim \zeta(\cdot)}\left[V_\pi(s_0)\right]$ and $\hat{J}(\pi) = \mathbb{E}_{s_0 \sim \zeta(\cdot)}\left[\hat{V}_\pi(s_0)\right]$.

Assumption 1. *V_{π_θ} is L-smooth in θ.*

This assumption holds when the reward r and transition function f are both Lipschitz continuous and smooth (*i.e.*, twice continuously differentiable with Lipschitz continuous first derivative) [1, 15, 11].

Fact 1. *(Bastani, 2021, Lemma D.2). Denote L_h as the Lipschitz constant for function h and $\bar{L}_h = \max\{L_h, L_{\nabla h,1}\}$. Then ∇V_{π_θ} is L-Lipschitz, where $L = 44H^5\bar{L}_r\bar{L}_f^{4H}$. Particularly, $\nabla_\theta V_{\pi_\theta}$ is Lipschitz continuous in θ with Lipschitz constant $24H^5\bar{L}_r\bar{L}_f^{4H}$.*

Define the gradient bias b_t and variance (upper bound) v_t as

$$b_t = \left\| \nabla_\theta J(\pi_{\theta_t}) - \mathbb{E}[\nabla_\theta \hat{J}(\pi_{\theta_t})] \right\|_2$$

$$v_t = \mathbb{E}\left[\left\| \nabla_\theta \hat{J}(\pi_{\theta_t}) - \mathbb{E}[\nabla_\theta \hat{J}(\pi_{\theta_t})] \right\|_2^2 \right]$$

With a similar proof in [15], we have the following bound.

Lemma 1. *Define the gradient mapping $\rho_t = \frac{1}{\eta}\left[P_{\Delta(\mathcal{A})^{|S|}}(\theta_t + \eta \nabla_\theta J(\pi_{\theta_t}))\right]$. For $c = \left(\eta - L\eta^2\right)^{-1}$ and $\eta \le \frac{1}{L}$,*

$$\min_{t\in[T]} \mathbb{E}\left[\|\rho_t\|_2^2\right] \le \frac{4}{T}\left(\sum_{t=1}^{T} c(b_t + \frac{\eta}{2}v_t) + b_t^2 + v_t\right) + \frac{4c}{T}\,\mathbb{E}[J(\pi_{\theta_T}) - J(\pi_{\theta_1})].$$

Theorem 1. *For $\eta \le \frac{1}{L}$,*

$$\min_{t\in[T]} J(\pi^*) - J(\pi_{\theta_t}) \le \left\| \frac{d_\zeta^{\pi^*}}{d_\zeta^{\pi_{\theta_t}}} \right\|_\infty \frac{4}{\sqrt{T}} \left(\left(\sum_{t=1}^{T} c(b_t + \frac{\eta}{2}v_t) + b_t^2 + v_t \right) + c\,\mathbb{E}[J(\pi_{\theta_T}) - J(\pi_{\theta_1})] \right)^{\frac{1}{2}}. \tag{1}$$

Proof. See Appendix 6.2. □

2.2 Function Approximations

Next, we consider discounted MDPs beyond finite settings, which enables function approximations. In particular, we focus on the convergence rates of Natural Policy Gradient.

$$\theta_{t+1} - \theta_t = \eta \tilde{\nabla}_\theta J(\theta_t)$$

$$\tilde{\nabla}_\theta J(\theta) = F(\theta)^{-1}\nabla_\theta J(\theta) = \frac{1}{1-\gamma} F(\theta)^{-1} \mathop{\mathbb{E}}_{s\sim d_{\mu_1}^{\pi_\theta}} \mathop{\mathbb{E}}_{a\sim\pi_\theta} \left[A^{\pi_\theta}(s,a)\nabla_\theta \log \pi_\theta(a|s)\right].$$

Theorem 2. *Assume $\log \pi_\theta(a|s)$ is a ι-smooth function of θ. For $\eta \le \frac{1}{L}$,*

$$\min_{t\in[T]} J(\pi^*) - J(\pi_{\theta_t}) \le \frac{\log|\mathcal{A}|}{T\eta(1-\gamma)} + \frac{\iota}{T(1-\gamma)(1-L\eta)} \left(\mathbb{E}[J(\pi_{\theta_T}) - J(\pi_{\theta_1})] + \eta \sum_{t=1}^{T} (b_t + \frac{v_t}{2}) \right). \tag{2}$$

Proof. See Appendix 6.3. □

What remains is to bound b_t and v_t in the convergence rate.

3 Likelihood Ratio Gradient Estimator

In the following sections, we assume the initial state s_0 is deterministic and known for simplicity, which implies $J(\pi) = V_\pi(s_0)$. We analyze finite-horizon γ-discounted MDPs.

Denote the value by unrolling H steps of model $\hat{f}$ with policy π as $V^{\hat{f}}_{\pi,H}$. One potential way to estimate $V_\pi(s_0)$ is simply calculating $V^{\hat{f}}_{\pi,H}(s_0)$ and setting H as the task horizon. However, such naive estimation will suffer from large bias due to model compounding error, or even cause training failure when adopting pathwise gradient estimator due to the curse of chaos, which we will formalize later. To tackle these challenges, we may adopt another value function $\hat{V}_\phi$, which can be either parametric or non-parametric, and obtain the *H-step model value expansion*:

$$\hat{V}_{\pi,H}(s_0) = V^{\hat{f}}(s_0) + \gamma^H \hat{V}_\phi(\hat{s}_H).$$

Then

$$\nabla_\theta \hat{V}_{\pi,H}(s) = \mathbb{E}[\hat{Q}_{\pi,H}(s,a)\nabla_\theta \log \pi_\theta(a|s)].$$

Proposition 1. *With function approximations, for a stochastic model $\hat{f}$ with form $\hat{s}_{h+1} = \hat{f}(\hat{s}_h, \hat{a}_h) + \xi_h$, where $\xi_h \sim p(\xi)$ and $p(\xi)$ is σ_ξ-subgaussian, we have*

$$\begin{aligned} v_t &\le \left(\frac{L_{\hat{r}}(L_{\hat{f}}|\mathcal{A}| + \sigma_\xi)}{L_{\hat{f}} - 1} \left(\frac{\gamma^H L_{\hat{f}}^{H-2} - \gamma^2}{\gamma L_{\hat{f}} - 1} + \gamma^H L_{\hat{V}_\phi}(L_{\hat{f}}^{H-2} - 1) \right) + \frac{L_{\hat{r}}|\mathcal{A}|}{1-\gamma} \right)^2 |\mathcal{A}|^2 \\ &= O\left((\gamma L_{\hat{f}})^{2H} \right). \end{aligned} \tag{3}$$

Proof. See Appendix 6.4. □

Denote $\mathcal{P}^H_\pi$ as the H-step environment transition operator under policy π: $\mathcal{P}^H_\pi[x](s_0) = \mathbb{E}[x(s_H)|f^*, \pi]$ and $\hat{\mathcal{P}}^H_\pi$ is the counterpart using the model $\hat{f}$ instead of the environment f^*.

Let $\mathcal{T}^H_\pi$ be H applications of the policy's Bellman operator $\mathcal{T}_\pi$, defined as $\mathcal{T}_\pi[v](s) = \mathbb{E}[r(s,a) + \gamma v(s')|f^*, \pi]$. Similarly, $\hat{\mathcal{T}}_\pi$ is the Bellman operator induced by model $\hat{f}$ and policy π.

Proposition 2. *The gradient bias is bounded by*

$$b_t \le \left\| \mathcal{T}^H_\pi \hat{V}^\pi_\phi - \hat{\mathcal{T}}^H_\pi \hat{V}^\pi_\phi \right\| |\mathcal{A}| + \gamma^H \mathcal{L}(\hat{V}^\pi_\phi)|\mathcal{A}|, \tag{4}$$

where loss $\mathcal{L}(\hat{V}^\pi_\phi) = \left\| \mathcal{P}^H_\pi V - \mathcal{P}^H_\pi \hat{V}^\pi_\phi \right\| + \left\| \hat{\mathcal{P}}^H_\pi V - \hat{\mathcal{P}}^H_\pi \hat{V}^\pi_\phi \right\|$.

Proof. See Appendix 6.5. □

The (non)parametric value is optimized to match the true value of π by minimizing $\mathcal{L}(\hat{V}^\pi_\phi)$, and we write $\hat{V}^\pi_\phi$ and $\hat{V}^\pi_\phi$ interchangeably.

4 Pathwise Gradient Estimator

Pathwise gradient estimator is another model-based approach that leverages the first-order gradient information in the function approximations. The pathwise gradient is calculated by backpropagating through the *H-step model value expansion*, but with a parameterized value function $\hat{V}_\phi$. Denote the H-step value expansion as $\hat{V}_{\pi,H}$.

Proposition 3. *Denote $\epsilon = \max \nabla_\theta \pi(a|s)$. With function approximations, we have*

$$v_t \le \left(\frac{2L_r(\epsilon + K^{H-1})}{L_{\hat{f}} - 1} \frac{\gamma^H L_{\hat{f}}^{H+1}}{\gamma L_{\hat{f}} - 1} + \gamma^H \left(2L_{\hat{Q}} K^{H+1} + 2L_{\hat{Q}} L_1 (\epsilon + K^H) L_{\hat{f}} \frac{L_{\hat{f}}^H - 1}{L_{\hat{f}} - 1} \right) \right)^2. \tag{5}$$

Proof. See Appendix 6.6. □

We can see the necessity of introducing $\hat{V}_\phi$. A large truncation step H will lead to the curse of chaos and failure in training [8, 10]. Specifically, v_t depends exponentially on $L_{\hat{f}}$, *i.e.*, small changes in initial conditions can result in diverging states and the gradient variance is also hopelessly large. (One potential way to mitigate it is to constrain the Jacobian of the learned dynamical system [8]. In other words, we can learn a more well-behaved non-chaotic version of physics instead of the rigid body dynamics with sharp changes.)

Proposition 4. *The gradient bias is bounded by*

$$b_t \leq \left\| \nabla_\theta \left(\mathcal{T}_\pi^H \hat{V}_\phi^\pi - \hat{\mathcal{T}}_\pi^H \hat{V}_\phi^\pi \right) \right\|_2 + \gamma^H \left\| \nabla_\theta \mathcal{L}(\hat{V}_\phi^\pi) \right\|_2 . \tag{6}$$

Proof. See Appendix 6.7. □

5 Value Equivalent Model Learning Through Policy Improvement Path

Algorithm 1 Model-Based Policy Gradient

Initialize policy π_0, model $\hat{f}$, parametric value $\hat{V}_\phi$, $t = 1$.

for $\lfloor \frac{T}{k} \rfloor k$ iterations **do**
  Execute policy π_{t-1} in the real environment
  Optimize $\hat{V}_\phi = \operatorname{argmin} \mathcal{L}(\hat{V}_\phi^{\pi_{t-1}})$ and $\hat{f}_t$
  for k gradient updates **do**
    Calculate gradient $\nabla_\theta \hat{J}(\pi_{t-1})$ with LR or Pathwise Gradient Estimator
    Perform policy gradient and obtain updated π_t
    $t \leftarrow t + 1$
  end for
end for

In [5], the Proper Value Equivalence (PVE) objective is

$$\begin{aligned} \left\| V^\pi - \hat{\mathcal{T}}_\pi^H V^\pi \right\| &\leq \left\| V^\pi - \mathcal{T}_\pi^H \hat{V}_\phi \right\| + \left\| \mathcal{T}_\pi^H \hat{V}_\phi - \hat{\mathcal{T}}_\pi^H \hat{V}_\phi \right\| + \left\| \hat{\mathcal{T}}_\pi^H \hat{V}_\phi - \hat{\mathcal{T}}_\pi^H V^\pi \right\| \\ &\leq \gamma^H \left\| \mathcal{P}_\pi^H V^\pi - \mathcal{P}_\pi^H \hat{V}_\phi \right\| + \gamma^H \left\| \hat{\mathcal{P}}_\pi^H V^\pi - \hat{\mathcal{P}}_\pi^H \hat{V}_\phi \right\| + \left\| \mathcal{T}_\pi^H \hat{V}_\phi - \hat{\mathcal{T}}_\pi^H \hat{V}_\phi \right\| . \end{aligned} \tag{7}$$

Thus, the common terms we want to minimize in Proposition 2 and 4 are the upper bound of the PVE loss $\left\| V^\pi - \hat{\mathcal{T}}_\pi^H V^\pi \right\|$. Minimizing this upper bound is equivalent to finding a proper value equivalent model with respect to a single policy.

However, for model-based RL, multiple policy updates are performed with an imagined model to reduce the sample complexity. In other words, the learned model $\hat{f}_t$ continues to have an impact on the following K policy updates. Thus, simply setting $\hat{f}_t = \operatorname{argmin}_{\hat{f}} \left\| \mathcal{T}_{\pi_{t-1}}^H \hat{V}_\phi^{\pi_{t-1}} - \hat{\mathcal{T}}_{\pi_{t-1}}^H \hat{V}_\phi^{\pi_{t-1}} \right\|$ might cause large bias under certain conditions.

Formally, in the convergence rate in Theorem 1, we need the accumulate $\sum_{t=1}^T b_t$ to be small. That is, the true objective for the model learner is

$$\min_{\hat{f}_1, \ldots, \hat{f}_{\lfloor \frac{T}{k} \rfloor k}} \sum_{t=1}^T \mathcal{L}_t(\hat{f}_{\lfloor \frac{T}{k} \rfloor k}) \triangleq \min_{\hat{f}_1, \ldots, \hat{f}_{\lfloor \frac{T}{k} \rfloor k}} \sum_{t=1}^T \left\| \mathcal{T}_{\pi_t}^H \hat{V}_\phi^{\pi_t} - \hat{\mathcal{T}}_{\pi_t}^H \hat{V}_\phi^{\pi_t} \right\| ,$$

where $\hat{\mathcal{T}}_\pi^H$ is the Bellman operator induced by model $\hat{f}_{\lfloor \frac{T}{k} \rfloor k}$.

This model learning process can be cast as an online learning problem. During each iteration t, the learner is tasked with learning a model. Then the model-based policy gradient is returned and a new policy π_t is obtained. The loss function $\mathcal{L}_t$ is revealed to the model learner afterwards. This

procedure is repeated across T rounds. Since $\mathcal{L}_t(f^*) = 0$ for the real model f^*, we have the following equivalence

$$\mathfrak{R} := \sum_{t=1}^{T} \mathcal{L}_t(\hat{f}_{\lfloor \frac{T}{k} \rfloor k}) - \inf_f \sum_{t=1}^{T} \mathcal{L}_t(f) \equiv \sum_{t=1}^{T} \mathcal{L}_t(\hat{f}_{\lfloor \frac{T}{k} \rfloor k}) - \sum_{t=1}^{T} \mathcal{L}_t(f^*) \equiv \sum_{t=1}^{T} \mathcal{L}_t(\hat{f}_{\lfloor \frac{T}{k} \rfloor k})$$

In other words, it is equivalent to solve a regret $\mathfrak{R}$ minimization problem, where many existing online learning approaches can be adopted.

For tabular settings and linear function approximations, the model learning process becomes an online convex learning problem. In this convex case, online gradient descent can be used to obtain $O(T^{\frac{1}{2}})$ regret. The iterative VAML [3] can be viewed as a special case of convex model learning with online gradient descent, but in a value-based planning manner. We may also adopt the abstract state representation when the underlying transition of a high-dimensional MDP can be linearly modeled.

$$\text{Online GD: } \hat{f}_{\lfloor \frac{T}{k} \rfloor k} = \hat{f}_{\lfloor \frac{T}{k} \rfloor k-1} - \eta \nabla \mathcal{L}_{t-1}(\hat{f}_{\lfloor \frac{T}{k} \rfloor k-1})$$

$$\text{FTPL: } \hat{f}_{\lfloor \frac{T}{k} \rfloor k} = \operatorname*{argmin}_{\hat{f} \in \mathcal{F}} \sum_{i=1}^{t-1} \mathcal{L}_i(\hat{f}) + \sigma_t$$

$$\text{OFTPL: } \hat{f}_{\lfloor \frac{T}{k} \rfloor k} = \operatorname*{argmin}_{\hat{f} \in \mathcal{F}} \sum_{i=1}^{t-1} \mathcal{L}_i(\hat{f}) + M_t(\hat{f}) + \sigma_t$$

where σ_t is a random perturbation such that $\sigma_{t,j}$, the j-th coordinate of σ_t, is sampled from $\text{Exp}(\lambda_t)$, the exponential distribution with parameter λ_t.

However, online GD losses its attractiveness in nonconvex setting, where the loss functions encountered by the model learner can potentially be nonconvex. For example, when we use nonlinear function approximation in complex tasks. Or nonlinear fitted value is adopted when value iteration cannot be performed exactly (*e.g.*, due to large state space).

In general non-convex online learning problem, the most straightforward algorithm is *Follow the Leader* (FTL) and its variants, including *Follow the Regularized Leader* (FTRL) and *Follow the Perturbed Leader* (FTPL). We show that FTPL can still achieve $O(T^{\frac{1}{2}})$ regret with access to an offline optimization oracle which takes as input a function and returns an approximate minimizer of this function.

Proposition 5. *Denote the (α, β)-approximate optimization oracle of function g as $\mathcal{O}_{\alpha,\beta}(g)$ in the sense that if $x^* = \mathcal{O}_{\alpha,\beta}(g)$, then $g(x^*) - \langle \sigma, x^* \rangle \leq \inf_x f(x) - \langle \sigma, x \rangle + \alpha + \beta \|\sigma\|_1$. For $\lambda = O(T^{-\frac{1}{2}})$, $\alpha = O(T^{-\frac{1}{2}})$, and $\beta = O(T^{-1})$, we have the regret bound for FTPL and OFTPL that*

$$\mathbb{E}\left[\sum_{t=1}^{T} \mathcal{L}_t(\hat{f}_{\lfloor \frac{T}{k} \rfloor k})\right] \leq O(T^{\frac{1}{2}}).$$

We also give the insights behind using FTL-style algorithms in the model learning process. The basic intuition of value-aware model learning is to assign different weights to the transitions according to the *optimal* state value. However, the optimal value function is not known and the VAML loss cannot be readily minimized. Thus, as shown in [6, 5], we may find models using value equivalence principle. As more value functions are augmented to be considered, the model will less compromise the performance, and eventually collapse to the real environment model. The key is to determine the representative value functions (or the value function of representative policies [5]). The above online learning problem is thus to select the value functions that are used to obtain the PVE model. Then loss $\mathcal{L}_t$ depends on how representative the selected value functions are. With more representative $\hat{V}_\phi^{\pi_{1:t-1}}$, model $\hat{f}_{\lfloor \frac{T}{k} \rfloor k}$ also more perfectly describe the environment, and thus leading to lower online loss $\mathcal{L}_t(\hat{f}_{\lfloor \frac{T}{k} \rfloor k})$ and smaller gradient bias.

The idea that past policies on the policy improvement path can be selected as representative ones shares similarities with recent implementations [2, 5]. For non-oblivious loss which is not adversarially chosen, we may expect the losses predictable from the past. Specifically, our model learning process

is predictable as successive losses depend on the policies obtained by applying policy gradient. Thus, using a predictor M_t might provide useful information and the resulting optimistic algorithms will enjoy tighter bounds. For example, we may simply set $M_t = \mathcal{L}_{t-1}$ to emphasise the policy similarities between consecutive iterations, or set $M_t = \frac{1}{t-1}\sum_{i=1}^{t-1}\mathcal{L}_i$ as the past average, or add importance weights $M_t = \frac{1}{t-1}\sum_{i=1}^{t-1}\alpha_i\mathcal{L}_i$ for fading memory statistics.

Plugging the above regret bound into Proposition 2, 4, and the two theorems, we get a $T^{-\frac{1}{2}}$ rate of convergence for model-based RL with both gradient estimators in both finite MDP settings and function approximation settings.

Since the optimal truncation timestep H depends on the accuracy of the value $\hat{V}_\phi$ and the model $\hat{f}$, it is impractical to determine the best H. Thus, we may jointly minimize the loss w.r.t. multiple possible H, which is also how Predictron, MuZero, VPN, and PVE are implemented.

6 Proofs

6.1 Proof of Lemma 1

Proof. Denote $\beta_t = \frac{\theta_{t+1}-\theta_t}{\eta}$. By Assumption 1, we have

$$\begin{aligned} J(\pi_{\theta_{t+1}}) - J(\pi_{\theta_t}) &\geq \nabla_\theta J(\pi_{\theta_t})^\top(\theta_{t+1}-\theta_t) - \frac{L}{2}\|\theta_{t+1}-\theta_t\|_2^2 \\ &= \eta \nabla_\theta J(\pi_{\theta_t})^\top \beta_t - \frac{L\eta^2}{2}\|\beta_t\|_2^2. \end{aligned} \tag{8}$$

Rewrite the exact gradient $\nabla_\theta J(\pi_{\theta_t})$ as

$$\nabla_\theta J(\pi_{\theta_t}) = \left(\nabla_\theta J(\pi_{\theta_t}) - \mathbb{E}[\nabla_\theta \hat{J}(\pi_{\theta_t})]\right) - \left(\nabla_\theta \hat{J}(\pi_{\theta_t}) - \mathbb{E}[\nabla_\theta \hat{J}(\pi_{\theta_t})]\right) + \nabla_\theta \hat{J}(\pi_{\theta_t}).$$

Then we bound $\nabla_\theta J(\pi_{\theta_t})^\top \beta_t$ in Eq. (8) by bounding the resulting three terms.

$$\left|\left(\nabla_\theta J(\pi_{\theta_t}) - \mathbb{E}[\nabla_\theta \hat{J}(\pi_{\theta_t})]\right)^\top \beta_t\right| \leq \|\beta_t\|_2 \|\nabla_\theta J(\pi_{\theta_t}) - \mathbb{E}[\nabla_\theta \hat{J}(\pi_{\theta_t})]\|_2 = \|\beta_t\|_2 b_t \tag{9}$$

$$\left(\nabla_\theta \hat{J}(\pi_{\theta_t}) - \mathbb{E}[\nabla_\theta \hat{J}(\pi_{\theta_t})]\right)^\top \beta_t \leq \frac{\|\nabla_\theta \hat{J}(\pi_{\theta_t}) - \mathbb{E}[\nabla_\theta \hat{J}(\pi_{\theta_t})]\|_2^2}{2} + \frac{\|\beta_t\|_2^2}{2} \tag{10}$$

$$\nabla_\theta \hat{J}(\pi_{\theta_t})^\top \beta_t \geq \|\beta_t\|_2^2, \tag{11}$$

where Eq. (11) holds due to the fact that $\left(\theta_{t+1} - (\theta_t + \eta \nabla_\theta \hat{J}(\theta_t))\right)^\top (\theta_{t+1} - \theta_t) \leq 0$.

Thus, we can bound Eq. (8) by

$$J(\pi_{\theta_{t+1}}) - J(\pi_{\theta_t}) \geq \eta \left(-\|\beta_t\|_2 b_t - \frac{\|\nabla_\theta \hat{J}(\pi_{\theta_t}) - \mathbb{E}[\nabla_\theta \hat{J}(\pi_{\theta_t})]\|_2^2}{2} + \frac{\|\beta_t\|_2^2}{2}\right) - \frac{L\eta^2}{2}\|\beta_t\|_2^2. \tag{12}$$

By taking expectation in Eq. (12), we have

$$\begin{aligned} \left(\frac{\eta}{2} - \frac{L\eta^2}{2}\right)\mathbb{E}[\|\beta_t\|_2^2] &\leq \mathbb{E}[J(\pi_{\theta_{t+1}}) - J(\pi_{\theta_t})] + \eta\, \mathbb{E}[\|\beta_t\|_2] b_t + \frac{\eta}{2} v_t \\ &\leq \mathbb{E}[J(\pi_{\theta_{t+1}}) - J(\pi_{\theta_t})] + b_t + \frac{\eta}{2} v_t \end{aligned} \tag{13}$$

Besides,

$$\begin{aligned} \|\rho_t - \beta_t\|_2 &= \frac{1}{\eta}\left\| P_{\Delta(\mathcal{A})^{|S|}}(\theta_t + \eta \nabla_\theta J(\pi_{\theta_t})) - P_{\Delta(\mathcal{A})^{|S|}}(\theta_t + \eta \nabla_\theta \hat{J}(\pi_{\theta_t}))\right\|_2 \\ &\leq \left\|\nabla_\theta J(\pi_{\theta_t}) - \nabla_\theta \hat{J}(\pi_{\theta_t})\right\|_2. \end{aligned}$$

Then due to the fact that $\|x+y\|_2^2 \leq 2\|x\|_2^2 + 2\|y\|_2^2$, we have

$$\begin{aligned} \mathbb{E}[\|\rho_t - \beta_t\|_2^2] &\leq \mathbb{E}\left[\left\|\nabla_\theta J(\pi_{\theta_t}) - \nabla_\theta \hat{J}(\pi_{\theta_t})\right\|_2^2\right] \\ &= \mathbb{E}\left[\left\|\nabla_\theta J(\pi_{\theta_t}) - \mathbb{E}[\nabla_\theta \hat{J}(\pi_{\theta_t})] + \mathbb{E}[\nabla_\theta \hat{J}(\pi_{\theta_t})] - \nabla_\theta \hat{J}(\pi_{\theta_t})\right\|_2^2\right] \\ &\leq 2\,\mathbb{E}\left[\left\|\nabla_\theta J(\pi_{\theta_t}) - \mathbb{E}[\nabla_\theta \hat{J}(\pi_{\theta_t})]\right\|_2^2\right] + 2\,\mathbb{E}\left[\left\|\mathbb{E}[\nabla_\theta \hat{J}(\pi_{\theta_t})] - \nabla_\theta \hat{J}(\pi_{\theta_t})\right\|_2^2\right] \\ &= 2\left\|\nabla_\theta J(\pi_{\theta_t}) - \mathbb{E}[\nabla_\theta \hat{J}(\pi_{\theta_t})]\right\|_2^2 + 2\,\mathbb{E}\left[\left\|\mathbb{E}[\nabla_\theta \hat{J}(\pi_{\theta_t})] - \nabla_\theta \hat{J}(\pi_{\theta_t})\right\|_2^2\right] \\ &\leq 2b_t^2 + 2v_t. \end{aligned} \tag{14}$$

For $\eta \leq \frac{1}{L}$, $\frac{\eta}{2} - \frac{L\eta^2}{2} > 0$. From Eq. (13) and Eq. (14), we have for $c = \left(\eta - L\eta^2\right)^{-1}$,

$$\begin{aligned}
\min_{t\in[T]} \mathbb{E}\left[\|\rho_t\|_2^2\right] &\leq \frac{1}{T}\sum_{t=1}^{T} \mathbb{E}\left[\|\rho_t\|_2^2\right] \\
&\leq \frac{2}{T}\sum_{t=1}^{T} \left(\mathbb{E}\left[\|\beta_t\|_2^2\right] + \mathbb{E}\left[\|\rho_t - \beta_t\|_2^2\right]\right) \\
&\leq \frac{4c}{T}\left(\mathbb{E}[J(\pi_{\theta_T}) - J(\pi_{\theta_1})] + \sum_{t=1}^{T}(b_t + \frac{\eta}{2}v_t)\right) + \frac{4}{T}\sum_{t=1}^{T}(b_t^2 + v_t) \\
&= \frac{4}{T}\left(\sum_{t=1}^{T} c(b_t + \frac{\eta}{2}v_t) + b_t^2 + v_t\right) + \frac{4c}{T}\,\mathbb{E}[J(\pi_{\theta_T}) - J(\pi_{\theta_1})].
\end{aligned}$$

□

6.2 Proof of Theorem 1

We first give two lemmas before the proof of Theorem 1.

Lemma 2. *(Agarwal, 2021, Lemma 4.1). Value function V satisfies gradient domination property. For all state distribution $\mu_1, \mu_2 \in \Delta(\mathcal{S})$, we have*

$$\mathop{\mathbb{E}}_{s_0\sim\mu_1}[V_{\pi^*}(s_0)] - \mathop{\mathbb{E}}_{s_0\sim\mu_1}[V_{\pi}(s_0)] \leq \left\|\frac{d_{\mu_1}^{\pi^*}}{d_{\mu_2}^{\pi}}\right\|_\infty \max_{\bar{\pi}} (\bar{\pi} - \pi)^\top \nabla_\pi \mathop{\mathbb{E}}_{s_0\sim\mu_2}[V_\pi(s_0)].$$

Lemma 3. *(Agarwal, 2021, Proposition B.1). Define $G^\eta(\pi) = \frac{1}{\eta}\left[P_{\Delta(\mathcal{A})^{|S|}}(\theta_t + \eta\nabla_\pi \mathbb{E}_{s_0}[V_\pi(s_0)]\right]$. If $\|G^\eta(\pi)\|_2 \leq \epsilon$, then*

$$\max_{\pi + e\in\Delta(\mathcal{A})^{|S|}, \|e\|_2\leq 1} e^\top \nabla_\pi \mathop{\mathbb{E}}_{s_0}[V_\pi(s_0)] \leq \epsilon(\eta L + 1).$$

Proof. From Lemma 2, we know that for initial distribution ζ,

$$J(\pi^*) - J(\pi_{\theta_t}) \leq \left\|\frac{d_\zeta^{\pi^*}}{d_\zeta^{\pi_{\theta_t}}}\right\|_\infty \max_{\bar{\pi}} (\bar{\pi} - \pi_{\theta_t})^\top \nabla_\pi J(\pi_{\theta_t}).$$

Then by Lemma 3, Lemma 1 and $\eta L + 1 \leq 2$, we know

$$\begin{aligned}
&\min_{t\in[T]} J(\pi^*) - J(\pi_{\theta_t}) \\
&\leq \left\|\frac{d_\zeta^{\pi^*}}{d_\zeta^{\pi_{\theta_t}}}\right\|_\infty (\eta L + 1) \min_{t\in[T]} \mathbb{E}\left[\|\rho_t\|_2^2\right] \\
&\leq \left\|\frac{d_\zeta^{\pi^*}}{d_\zeta^{\pi_{\theta_t}}}\right\|_\infty (\eta L + 1) \min_{t\in[T]} \sqrt{\mathbb{E}\left[\|\rho_t\|_2^2\right]} \\
&\leq \left\|\frac{d_\zeta^{\pi^*}}{d_\zeta^{\pi_{\theta_t}}}\right\|_\infty (\eta L + 1)\frac{2}{\sqrt{T}}\left(\left(\sum_{t=1}^{T} c(b_t + \frac{\eta}{2}v_t) + b_t^2 + v_t\right) + c\,\mathbb{E}[J(\pi_{\theta_T}) - J(\pi_{\theta_1})]\right)^{\frac{1}{2}} \\
&\leq \left\|\frac{d_\zeta^{\pi^*}}{d_\zeta^{\pi_{\theta_t}}}\right\|_\infty \frac{4}{\sqrt{T}}\left(\left(\sum_{t=1}^{T} c(b_t + \frac{\eta}{2}v_t) + b_t^2 + v_t\right) + c\,\mathbb{E}[J(\pi_{\theta_T}) - J(\pi_{\theta_1})]\right)^{\frac{1}{2}},
\end{aligned}$$

where the second inequality follows Jensen's inequality with the concave square root function. □

6.3 Proof of Theorem 2

Proof. By smoothness, we have

$$\log \frac{\pi_{t+1}(a|s)}{\pi_t(a|s)} - \nabla_\theta \log \pi_t(a|s)(\theta_{t+1} - \theta_t) \geq -\frac{\iota}{2} \|\theta_{t+1} - \theta_t\|_2^2 . \tag{15}$$

To match the definition in Lemma 1, denote $\beta_t = \frac{\theta_{t+1}-\theta_t}{\eta} = \tilde{\nabla}_\theta \hat{J}(\theta_t)$.

$$\begin{aligned}
\mathop{\mathbb{E}}_{s\sim d^{\pi^*}}[D_{\mathrm{KL}}(\pi^*||\pi_t) - D_{\mathrm{KL}}(\pi^*||\pi_{t+1})] &= \mathop{\mathbb{E}}_{s\sim d^{\pi^*}} \mathop{\mathbb{E}}_{a\sim\pi^*} \left[\log \frac{\pi_{t+1}(a|s)}{\pi_t(a|s)}\right] \\
&\geq \mathop{\mathbb{E}}_{s\sim d^{\pi^*}} \mathop{\mathbb{E}}_{a\sim\pi^*} \left[\nabla_\theta \log \pi_t(a|s)(\theta_{t+1} - \theta_t) - \frac{\iota}{2} \|\theta_{t+1} - \theta_t\|_2^2\right] \\
&= \mathop{\mathbb{E}}_{s\sim d^{\pi^*}} \mathop{\mathbb{E}}_{a\sim\pi^*} \left[\eta A^{\pi_t}(s,a) - \frac{\iota\eta^2}{2} \|\beta_t\|_2^2\right] \\
&= (1-\gamma)\eta \left(J(\pi^*) - J(\pi_t)\right) - \frac{\iota\eta^2}{2} \mathbb{E}[\|\beta_t\|_2^2].
\end{aligned} \tag{16}$$

where the last equality follows the performance difference lemma [7].

Using similar techniques in the proof of Lemma 1, we have for $\eta \leq \frac{1}{L}$ that

$$\mathbb{E}[\|\beta_t\|_2^2] \leq 2(\eta - L\eta^2)^{-1} \left(\mathbb{E}[J(\pi_{\theta_{t+1}}) - J(\pi_{\theta_t})] + \eta b_t + \frac{\eta}{2} v_t\right). \tag{17}$$

Thus,

$$\begin{aligned}
&\min_{t\in[T]} J(\pi^*) - J(\pi_{\theta_t}) \\
&\leq \frac{1}{T}\sum_{t=1}^{T} \left(J(\pi^*) - J(\pi_{\theta_t})\right) \\
&\leq \frac{1}{T\eta(1-\gamma)} \sum_{t=1}^{T} \mathop{\mathbb{E}}_{s\sim d^{\pi^*}}[D_{\mathrm{KL}}(\pi^*||\pi_t) - D_{\mathrm{KL}}(\pi^*||\pi_{t+1})] + \frac{\iota\eta^2}{2} \mathbb{E}[\|\beta_t\|_2^2] \\
&\leq \frac{\mathbb{E}_{s\sim d^{\pi^*}}[D_{\mathrm{KL}}(\pi^*||\pi_1)]}{T\eta(1-\gamma)} + \frac{\iota}{T(1-\gamma)(1-L\eta)} \left(\mathbb{E}[J(\pi_{\theta_T}) - J(\pi_{\theta_1})] + \eta \sum_{t=1}^{T} (b_t + \frac{v_t}{2})\right) \\
&\leq \frac{\log|\mathcal{A}|}{T\eta(1-\gamma)} + \frac{\iota}{T(1-\gamma)(1-L\eta)} \left(\mathbb{E}[J(\pi_{\theta_T}) - J(\pi_{\theta_1})] + \eta \sum_{t=1}^{T} (b_t + \frac{v_t}{2})\right).
\end{aligned}$$

□

6.4 Proof of Proposition 1

Proof. First, we have

$$\begin{aligned}
\|\hat{s}_{h+1} - \mathbb{E}[\hat{s}_{h+1}]\|_2 &= \left\|\hat{f}(\hat{s}_h, \hat{a}_h) + \xi_h - \mathbb{E}[\hat{f}(\hat{s}_h, \hat{a}_h)]\right\|_2 \\
&\leq L_{\hat{f}} \sqrt{\|\hat{s}_h - \mathbb{E}[\hat{s}_h]\|_2^2 + |\mathcal{A}|^2} + \sigma_\xi \\
&\leq L_{\hat{f}} \|\hat{s}_h - \mathbb{E}[\hat{s}_h]\|_2 + L_{\hat{f}}|\mathcal{A}| + \sigma_\xi.
\end{aligned} \tag{18}$$

Since $\|\hat{s}_0 - \mathbb{E}[\hat{s}_0]\|_2 = 0$, we have

$$\|\hat{s}_h - \mathbb{E}[\hat{s}_h]\|_2 \leq \left(L_{\hat{f}}|\mathcal{A}| + \sigma_\xi\right) \frac{L_{\hat{f}}^{h-2} - 1}{L_{\hat{f}} - 1}$$

.

Besides,

$$\begin{aligned}\left\|\hat{V}-\mathbb{E}[\hat{V}]\right\| &= \left\|\sum_{h=0}^{H-1}\gamma^h\left(\hat{r}(\hat{s}_h,\hat{a}_h)-\mathbb{E}[\hat{r}(\hat{s}_h,\hat{a}_h)]\right)+\gamma^H\left(\hat{V}_\phi(\hat{s}_H)-\mathbb{E}[\hat{V}_\phi(\hat{s}_H)]\right)\right\|_2 \\ &\leq \sum_{h=0}^{H-1}\gamma^h L_{\hat{r}}\left(\left\|\hat{s}_h-\mathbb{E}[\hat{s}_h]\right\|_2+|\mathcal{A}|\right)+\gamma^H L_{\hat{V}_\phi}\left\|\hat{s}_H-\mathbb{E}[\hat{s}_H]\right\|_2 \\ &= \frac{L_{\hat{r}}(L_{\hat{f}}|\mathcal{A}|+\sigma_\xi)}{L_{\hat{f}}-1}\left(\frac{\gamma^H L_{\hat{f}}^{H-2}-\gamma^2}{\gamma L_{\hat{f}}-1}+\gamma^H L_{\hat{V}_\phi}(L_{\hat{f}}^{H-2}-1)\right)+\frac{L_{\hat{r}}|\mathcal{A}|}{1-\gamma}.\end{aligned} \tag{19}$$

Since for finite MDP,

$$\nabla_\theta J(\pi_{\theta_t})=\sum_{s\in\mathcal{S}}d^\pi(s)\sum_{a\in\mathcal{A}}Q^\pi(s,a),$$

where $d^\pi(s)=\lim_{h\to\infty}P(s_h=s|s_0,\pi_\theta)$.

Finally,

$$\begin{aligned}v_t &= \mathbb{E}\left[\left\|\nabla_\theta\hat{J}(\pi_{\theta_t})-\mathbb{E}[\nabla_\theta\hat{J}(\pi_{\theta_t})]\right\|_2^2\right] \\ &\leq \left(\frac{L_{\hat{r}}(L_{\hat{f}}|\mathcal{A}|+\sigma_\xi)}{L_{\hat{f}}-1}\left(\frac{\gamma^H L_{\hat{f}}^{H-2}-\gamma^2}{\gamma L_{\hat{f}}-1}+\gamma^H L_{\hat{V}_\phi}(L_{\hat{f}}^{H-2}-1)\right)+\frac{L_{\hat{r}}|\mathcal{A}|}{1-\gamma}\right)^2|\mathcal{A}|^2 \\ &= O\left((\gamma L_{\hat{f}})^{2H}\right).\end{aligned}$$

□

6.5 Proof of Proposition 2

Proof. Let $R_\theta=\mathbb{E}[r(s,a)]$.

For H-step model-based rollout, we have

$$\begin{aligned}\left\|V-\mathbb{E}[\hat{V}]\right\| &= \left\|\sum_{h=0}^{H-1}\gamma^h\left(\mathbb{E}[r(s_h,a_h)]-\mathbb{E}[\hat{r}(\hat{s}_h,\hat{a}_h)]\right)+\gamma^H\left(\mathbb{E}[V(s_H)]-\mathbb{E}[\hat{V}_\phi(\hat{s}_H)]\right)\right\| \\ &= \left\|\sum_{h=0}^{H-1}\gamma^h\left(\mathcal{P}_\pi^h R_\theta-\hat{\mathcal{P}}_\pi^h\hat{R}_\theta\right)+\gamma^H\left(\mathcal{P}_\pi^H V-\hat{\mathcal{P}}_\pi^H\hat{V}_\phi\right)\right\| \\ &= \left\|\sum_{h=0}^{H-1}\gamma^h\left(\mathcal{P}_\pi^h R_\theta-\hat{\mathcal{P}}_\pi^h\hat{R}_\theta\right)+\gamma^H\left(\mathcal{P}_\pi^H\hat{V}_\phi-\hat{\mathcal{P}}_\pi^H\hat{V}_\phi+\mathcal{P}_\pi^H V-\mathcal{P}_\pi^H\hat{V}_\phi\right)\right\| \\ &\leq \left\|\mathcal{T}_\pi^H\hat{V}_\phi-\hat{\mathcal{T}}_\pi^H\hat{V}_\phi\right\|+\gamma^H\left\|\mathcal{P}_\pi^H V-\mathcal{P}_\pi^H\hat{V}_\phi\right\| \\ &\leq \left\|\mathcal{T}_\pi^H\hat{V}_\phi-\hat{\mathcal{T}}_\pi^H\hat{V}_\phi\right\|+\gamma^H\mathcal{L}(\hat{V}_\phi).\end{aligned} \tag{20}$$

The gradient bias is thus

$$\begin{aligned}b_t &= \left\|\nabla_\theta J(\pi_{\theta_t})-\mathbb{E}[\nabla_\theta\hat{J}(\pi_{\theta_t})]\right\|_2 \\ &\leq \left\|\mathcal{T}_\pi^H\hat{V}_\phi-\hat{\mathcal{T}}_\pi^H\hat{V}_\phi\right\||\mathcal{A}|+\gamma^H\mathcal{L}(\hat{V}_\phi)|\mathcal{A}|.\end{aligned} \tag{21}$$

□

6.6 Proof of Proposition 3

Proof. For $\hat{V}_{\pi,H}(s_0) = V^{\hat{f}}(s_0) + \gamma^H \mathbb{E}_a[\hat{Q}_\phi(\hat{s}_H, \hat{a}_H)]$, we have

$$\nabla_\theta \hat{V}_{\pi,H}(s_0) = \nabla_\theta V^{\hat{f}}(s_0) + \gamma^H \nabla_a \hat{Q}_\phi(\hat{s}_H, \hat{a}_H)\nabla_\theta \hat{a}_H + \gamma^H \nabla_s \hat{Q}_\phi(\hat{s}_H, \hat{a}_H)\nabla_\theta \hat{s}_H.$$

Thus, v_t is bounded by

$$\begin{aligned}
&\left\|\nabla_\theta \hat{J}(\pi_{\theta_t}) - \mathbb{E}[\nabla_\theta \hat{J}(\pi_{\theta_t})]\right\|_2 \\
&\leq \left\|\nabla_\theta V_H^{\hat{f}}(s_0) - \mathbb{E}[\nabla_\theta V_H^{\hat{f}}(s_0)]\right\|_2 + \gamma^H \left\|\nabla_a \hat{Q}_\phi(\hat{s}_H, \hat{a}_H)\nabla_\theta \hat{a}_H - \mathop{\mathbb{E}}_{\hat{s}_H, \hat{a}_H}[\nabla_a \hat{Q}_\phi(\hat{s}_H, \hat{a}_H)\nabla_\theta \hat{a}_H]\right\|_2 \\
&+ \gamma^H \left\|\nabla_s \hat{Q}_\phi(\hat{s}_H, \hat{a}_H)\nabla_\theta \hat{s}_H - \mathop{\mathbb{E}}_{\hat{s}_H, \hat{a}_H}[\nabla_s \hat{Q}_\phi(\hat{s}_H, \hat{a}_H)\nabla_\theta \hat{s}_H]\right\|_2 \\
&= \left\|\nabla_\theta V_H^{\hat{f}}(s_0) - \mathbb{E}[\nabla_\theta V_H^{\hat{f}}(s_0)]\right\|_2 + \gamma^H \left\|\nabla \hat{Q}_\phi \nabla_\theta \hat{x}_H - \mathop{\mathbb{E}}_{\hat{x}_H}[\nabla \hat{Q}_\phi \nabla_\theta \hat{x}_H]\right\|_2 \\
&\leq \frac{2L_r(\epsilon + K^{H-1})}{L_{\hat{f}} - 1}\frac{\gamma^H L_{\hat{f}}^{H+1}}{\gamma L_{\hat{f}} - 1} + \gamma^H \left\|\nabla \hat{Q}_\phi \nabla_\theta \hat{x}_H - \mathop{\mathbb{E}}_{\hat{x}_H}[\nabla \hat{Q}_\phi \nabla_\theta \hat{x}_H]\right\|_2,
\end{aligned} \tag{22}$$

where the last inequality follows Lemma 4. The second term is bounded by

$$\begin{aligned}
\left\|\nabla \hat{Q}_\phi \nabla_\theta \hat{x}_H - \mathop{\mathbb{E}}_{\hat{x}_H}[\nabla \hat{Q}_\phi \nabla_\theta \hat{x}_H]\right\|_2 &= \left\|\nabla \hat{Q}_\phi \mathop{\mathbb{E}}_{\hat{x}_H}[\nabla_\theta \hat{x}_H] - \mathop{\mathbb{E}}_{\hat{x}_H}[\nabla \hat{Q}_\phi \nabla_\theta \hat{x}_H] + \nabla \hat{Q}_\phi \nabla_\theta \hat{x}_H - \nabla \hat{Q}_\phi \mathop{\mathbb{E}}_{\hat{x}_H}[\nabla_\theta \hat{x}_H]\right\|_2 \\
&\leq 2L_{\hat{Q}} \left\|\mathop{\mathbb{E}}_{\hat{x}_H}[\nabla_\theta \hat{x}_H]\right\|_2 + L_{\hat{Q}} \left\|\nabla_\theta \hat{x}_H - \mathop{\mathbb{E}}_{\hat{x}_H}[\nabla_\theta \hat{x}_H]\right\|_2 \\
&\leq 2L_{\hat{Q}} K^{H+1} + 2L_{\hat{Q}} L_1(\epsilon + K^H) L_{\hat{f}} \frac{L_{\hat{f}}^H - 1}{L_{\hat{f}} - 1},
\end{aligned} \tag{23}$$

where the last inequality follows Equation (28).

Finally,

$$\begin{aligned}
v_t &= \mathbb{E}\left[\left\|\nabla_\theta \hat{J}(\pi_{\theta_t}) - \mathbb{E}[\nabla_\theta \hat{J}(\pi_{\theta_t})]\right\|_2^2\right] \\
&\leq \left(\frac{2L_r(\epsilon + K^{H-1})}{L_{\hat{f}} - 1}\frac{\gamma^H L_{\hat{f}}^{H+1}}{\gamma L_{\hat{f}} - 1} + \gamma^H \left(2L_{\hat{Q}} K^{H+1} + 2L_{\hat{Q}} L_1(\epsilon + K^H) L_{\hat{f}} \frac{L_{\hat{f}}^H - 1}{L_{\hat{f}} - 1}\right)\right)^2.
\end{aligned} \tag{24}$$

□

Lemma 4. *For H-step model rollout value, we have the following for its pathwise gradient*

$$\left\|\nabla_\theta V_H^{\hat{f}}(s_0) - \mathbb{E}[\nabla_\theta V_H^{\hat{f}}(s_0)]\right\|_2 \leq \frac{2(\epsilon + K^{H-1})L_r}{L_{\hat{f}} - 1}\left(\frac{\gamma^H L_{\hat{f}}^{H+1}}{\gamma L_{\hat{f}} - 1}\right). \tag{25}$$

Proof.

$$\left\|\nabla_\theta V_H^{\hat{f}}(s_0) - \mathbb{E}[\nabla_\theta V_H^{\hat{f}}(s_0)]\right\|_2 \leq \sum_{h=0}^{H-1} \gamma^h \left\|\nabla_\theta r(\hat{s}_h, \hat{a}_h) - \mathbb{E}[\nabla_\theta r(\hat{s}_h, \hat{a}_h)]\right\|_2 \tag{26}$$

Here, $\hat{s}_{h+1} = \hat{f}(\hat{s}_h, \hat{a}_h) + \xi_h$ and $\hat{a}_h \sim \pi(\cdot|\hat{s}_h)$.

Assume $\|\nabla_\theta \hat{x}_h\|_2 \leq K^{h+1}$, where K is related to the Lipschitz constants.

$$\begin{aligned}
\left\|\nabla_\theta \hat{s}_{h+1} - \mathop{\mathbb{E}}_{\hat{s}_{h+1}}[\nabla_\theta \hat{s}_{h+1}]\right\|_2 &= \left\|\nabla_s \hat{f}\nabla_\theta \hat{s}_h + \nabla_a \hat{f}\nabla_\theta \hat{a}_h - \mathop{\mathbb{E}}_{\hat{x}_h}[\nabla_s \hat{f}\nabla_\theta \hat{s}_h + \nabla_a \hat{f}\nabla_\theta \hat{a}_h]\right\|_2 \\
&\leq \left\|\nabla_s \hat{f}\nabla_\theta \hat{s}_h - \mathop{\mathbb{E}}_{\hat{x}_h}[\nabla_s \hat{f}\nabla_\theta \hat{s}_h]\right\|_2 + \left\|\nabla_a \hat{f}\nabla_\theta \hat{a}_h - \mathop{\mathbb{E}}_{\hat{x}_h}[\nabla_a \hat{f}\nabla_\theta \hat{a}_h]\right\|_2 \\
&\leq \left\|\nabla_s \hat{f}\nabla_\theta \hat{s}_h - \mathop{\mathbb{E}}_{\hat{x}_h}[\nabla_s \hat{f}]\nabla_\theta \hat{s}_h\right\|_2 + \left\|\mathop{\mathbb{E}}_{\hat{x}_h}[\nabla_s \hat{f}]\nabla_\theta \hat{s}_h - \mathop{\mathbb{E}}_{\hat{x}_h}[\nabla_s \hat{f}\nabla_\theta \hat{s}_h]\right\|_2 \\
&+ \left\|\nabla_a \hat{f}\nabla_\theta \hat{a}_h - \mathop{\mathbb{E}}_{\hat{x}_h}[\nabla_a \hat{f}]\nabla_\theta \hat{a}_h\right\|_2 + \left\|\mathop{\mathbb{E}}_{\hat{x}_h}[\nabla_a \hat{f}]\nabla_\theta \hat{a}_h - \mathop{\mathbb{E}}_{\hat{x}_h}[\nabla_a \hat{f}\nabla_\theta \hat{a}_h]\right\|_2 \\
&\leq 2L_{\hat{f}} \|\nabla_\theta \hat{x}_h\|_2 + L_{\hat{f}} \left\|\nabla_\theta \hat{s}_h - \mathop{\mathbb{E}}_{\hat{x}_h}[\nabla_\theta \hat{s}_h]\right\|_2 + L_{\hat{f}} \left\|\nabla_\theta \hat{a}_h - \mathop{\mathbb{E}}_{\hat{x}_h}[\nabla_\theta \hat{a}_h]\right\|_2 \\
&\leq 2L_{\hat{f}} K^{h+1} + L_{\hat{f}} \left\|\nabla_\theta \hat{s}_h - \mathop{\mathbb{E}}_{\hat{x}_h}[\nabla_\theta \hat{s}_h]\right\|_2 + 2L_{\hat{f}}\epsilon
\end{aligned} \tag{27}$$

Applying the recursion we have

$$\begin{aligned}
\left\|\nabla_\theta \hat{s}_h - \mathop{\mathbb{E}}_{\hat{s}_h}[\nabla_\theta \hat{s}_h]\right\|_2 &\leq 2L_{\hat{f}}(\epsilon + K^h) \sum_{k=0}^{h-1} L_{\hat{f}}^{h-1-k} \\
&= 2(\epsilon + K^h) \sum_{k=0}^{h-1} L_{\hat{f}}^{h-k} \\
&= 2(\epsilon + K^h) L_{\hat{f}} \frac{L_{\hat{f}}^h - 1}{L_{\hat{f}} - 1}
\end{aligned} \tag{28}$$

Then

$$\|r(\hat{s}_h, \hat{a}_h) - \mathbb{E}[r(\hat{s}_h, \hat{a}_h)]\|_2 \leq L_r \|(\hat{s}_h - \mathbb{E}[\hat{s}_h], \hat{a}_h - \mathbb{E}[\hat{a}_h])\|_2 \tag{29}$$

$$\begin{aligned}
\|\nabla_\theta r(\hat{s}_h, \hat{a}_h) - \mathbb{E}[\nabla_\theta r(\hat{s}_h, \hat{a}_h)]\|_2 &\leq L_r \sqrt{\left(2(\epsilon + K^{h-1}) L_{\hat{f}} \frac{L_{\hat{f}}^h - 1}{L_{\hat{f}} - 1}\right)^2 + (2\epsilon)^2} \\
&\leq 2(\epsilon + K^h) L_r \left(\frac{L_{\hat{f}}^{h+1} - L_{\hat{f}}}{L_{\hat{f}} - 1} + 1\right) \\
&= 2(\epsilon + K^h) L_r \frac{L_{\hat{f}}^{h+1} - 1}{L_{\hat{f}} - 1}
\end{aligned} \tag{30}$$

Thus,

$$\begin{aligned}
\left\|\nabla_\theta V_H^{\hat{f}}(s_0) - \mathbb{E}[\nabla_\theta V_H^{\hat{f}}(s_0)]\right\|_2 &\leq \sum_{h=0}^{H-1} \gamma^h \left\|\nabla_\theta r(\hat{s}_h, \hat{a}_h) - \mathbb{E}[\nabla_\theta r(\hat{s}_h, \hat{a}_h)]\right\|_2 \\
&\leq \frac{2(\epsilon + K^{H-1})L_r}{L_{\hat{f}} - 1} \sum_{h=0}^{H-1} \gamma^h \left(L_{\hat{f}}^{h+1} - 1\right) \\
&\leq \frac{2(\epsilon + K^{H-1})L_r}{L_{\hat{f}} - 1} \left(\frac{\gamma^H L_{\hat{f}}^{H+1} - 1}{\gamma L_{\hat{f}} - 1} - \frac{\gamma^H - 1}{\gamma - 1}\right) \qquad (31) \\
&\leq \frac{2(\epsilon + K^{H-1})L_r}{L_{\hat{f}} - 1} \left(\frac{\gamma^H L_{\hat{f}}^{H+1} - 1}{\gamma L_{\hat{f}} - 1} - \gamma^{H-1}\right) \\
&\leq \frac{2(\epsilon + K^{H-1})L_r}{L_{\hat{f}} - 1} \left(\frac{\gamma^H L_{\hat{f}}^{H+1}}{\gamma L_{\hat{f}} - 1}\right)
\end{aligned}$$

□

6.7 Proof of Proposition 4

Proof. The gradient bias is

$$\begin{aligned}
b_t &= \left\|\nabla_\theta J(\pi_{\theta_t}) - \mathbb{E}[\nabla_\theta \hat{J}(\pi_{\theta_t})]\right\|_2 \\
&= \left\|\sum_{h=0}^{H-1} \gamma^h \left(\mathbb{E}[\nabla_\theta r(s_h, a_h)] - \mathbb{E}[\nabla_\theta \hat{r}(\hat{s}_h, \hat{a}_h)]\right) + \gamma^H \left(\mathbb{E}[\nabla_\theta V(s_H)] - \mathbb{E}[\nabla_\theta \hat{V}_\phi(\hat{s}_H)]\right)\right\|_2 \\
&= \left\|\sum_{h=0}^{H-1} \gamma^h \nabla_\theta (\mathcal{P}_\pi^h R_\theta - \hat{\mathcal{P}}_\pi^h \hat{R}_\theta) + \gamma^H \nabla_\theta (\mathcal{P}_\pi^H V - \hat{\mathcal{P}}_\pi^H \hat{V}_\phi)\right\|_2 \\
&= \left\|\sum_{h=0}^{H-1} \gamma^h \nabla_\theta (\mathcal{P}_\pi^h R_\theta - \hat{\mathcal{P}}_\pi^h \hat{R}_\theta) + \gamma^H \nabla_\theta (\mathcal{P}_\pi^H \hat{V}_\phi - \hat{\mathcal{P}}_\pi^H \hat{V}_\phi) + \gamma^H \nabla_\theta (\mathcal{P}_\pi^H V - \mathcal{P}_\pi^H \hat{V}_\phi)\right\|_2 \\
&\leq \left\|\nabla_\theta \left(\mathcal{T}_\pi^H \hat{V}_\phi - \hat{\mathcal{T}}_\pi^H \hat{V}_\phi\right)\right\|_2 + \gamma^H \left\|\nabla_\theta \left(\mathcal{P}_\pi^H V - \mathcal{P}_\pi^H \hat{V}_\phi\right)\right\|_2 .
\end{aligned} \qquad (32)$$

□

7 Proof of Proposition 5

Proof. For regret $\mathfrak{R}$, we have

$$\begin{aligned}
\mathfrak{R} &= \sum_{t=1}^{T} \mathcal{L}_t(\hat{f}_{\lfloor \frac{T}{k} \rfloor k}) \\
&= \sum_{t=1}^{T} \left[\mathcal{L}_t(\hat{f}_{\lfloor \frac{T}{k} \rfloor k}) - \mathcal{L}_t(\hat{f}_{t+1}) \right] + \sum_{t=1}^{T} \mathcal{L}_t(\hat{f}_{t+1}) \\
&\leq l \sum_{t=1}^{T} \|\hat{f}_{\lfloor \frac{T}{k} \rfloor k} - \hat{f}_{t+1}\|_1 + \sum_{t=1}^{T} \mathcal{L}_t(\hat{f}_{t+1}) \\
&= l \sum_{j=1}^{\lfloor \frac{T}{k} \rfloor k} \sum_{i=1}^{k} \|\hat{f}_{jk} - \hat{f}_{jk+i}\|_1 + \sum_{t=1}^{T} \mathcal{L}_t(\hat{f}_{t+1}) \\
&\leq l \sum_{j=1}^{\lfloor \frac{T}{k} \rfloor k} \sum_{i=1}^{k} \|\hat{f}_{jk} - \hat{f}_{jk+1}\|_1 + \ldots + \|\hat{f}_{jk+i-1} - \hat{f}_{jk+i}\|_1 + \sum_{t=1}^{T} \mathcal{L}_t(\hat{f}_{t+1}) \\
&= l \sum_{j=1}^{\lfloor \frac{T}{k} \rfloor k} \sum_{i=1}^{k} (k+1-i) \|\hat{f}_{jk+i-1} - \hat{f}_{jk+i}\|_1 + \sum_{t=1}^{T} \mathcal{L}_t(\hat{f}_{t+1}).
\end{aligned}$$

Note that we only optimize the model at iteration $\lfloor \frac{T}{k} \rfloor k$, and the virtual $\hat{f}_t$ at other iteration t is not what we have during training. It is introduced for the proof only.

By applying induction, we have for the expected regret

$$\begin{aligned}
\mathbb{E}\left[\sum_{t=1}^{T} \mathcal{L}_t(\hat{f}_{\lfloor \frac{T}{k} \rfloor k}) \right] &\leq l \sum_{t=1}^{T} \frac{k(k+1)}{2} \mathbb{E}[\|\hat{f}_t - \hat{f}_{t+1}\|_1] + \sum_{t=1}^{T} \mathbb{E}[\mathcal{L}_t(\hat{f}_{t+1})] \\
&\leq l \sum_{t=1}^{T} \frac{k(k+1)}{2} \mathbb{E}[\|\hat{f}_t - \hat{f}_{t+1}\|_1] + \frac{d(\beta T + D)}{\lambda} + \alpha T.
\end{aligned} \tag{33}$$

The stability term $\mathbb{E}[\|\hat{f}_t - \hat{f}_{t+1}\|_1]$. From Lemma 5 and Lemma 6 in [14], we have

$$\mathbb{E}[\|\hat{f}_t - \hat{f}_{t+1}\|_1] \leq 125\lambda D l d^2 + \frac{d\beta}{20\lambda l} + 2d\beta + \frac{\alpha}{20l}.$$

Plugging the above bound in Equation (33) gives the bound of expected regret.

$$\mathbb{E}\left[\sum_{t=1}^{T} \mathcal{L}_t(\hat{f}_{\lfloor \frac{T}{k} \rfloor k}) \right] \leq O\left(T\lambda D l^2 d^2 k^2 + dl\beta + \frac{d(\beta T + D)}{\lambda} + \alpha T \right).$$

For $\lambda = O(T^{-\frac{1}{2}})$, $\alpha = O(T^{-\frac{1}{2}})$, and $\beta = O(T^{-1})$, we have the $O(T^{\frac{1}{2}})$ expect regret bound. $\square$

References

[1] Osbert Bastani. Sample complexity of estimating the policy gradient for nearly deterministic dynamical systems. In *International Conference on Artificial Intelligence and Statistics*, pages 3858–3869. PMLR, 2020.

[2] Will Dabney, André Barreto, Mark Rowland, Robert Dadashi, John Quan, Marc G Bellemare, and David Silver. The value-improvement path: Towards better representations for reinforcement learning. *arXiv preprint arXiv:2006.02243*, 2020.

[3] Amir-massoud Farahmand. Iterative value-aware model learning. In *NeurIPS*, pages 9090–9101, 2018.

[4] Amir-massoud Farahmand, Andre Barreto, and Daniel Nikovski. Value-aware loss function for model-based reinforcement learning. In *Artificial Intelligence and Statistics*, pages 1486–1494. PMLR, 2017.

[5] Christopher Grimm, André Barreto, Gregory Farquhar, David Silver, and Satinder Singh. Proper value equivalence. *arXiv preprint arXiv:2106.10316*, 2021.

[6] Christopher Grimm, André Barreto, Satinder Singh, and David Silver. The value equivalence principle for model-based reinforcement learning. *arXiv preprint arXiv:2011.03506*, 2020.

[7] Sham Kakade and John Langford. Approximately optimal approximate reinforcement learning. In *In Proc. 19th International Conference on Machine Learning*. Citeseer, 2002.

[8] Luke Metz, C Daniel Freeman, Samuel S Schoenholz, and Tal Kachman. Gradients are not all you need. *arXiv preprint arXiv:2111.05803*, 2021.

[9] Junhyuk Oh, Satinder Singh, and Honglak Lee. Value prediction network. *arXiv preprint arXiv:1707.03497*, 2017.

[10] Paavo Parmas, Carl Edward Rasmussen, Jan Peters, and Kenji Doya. Pipps: Flexible model-based policy search robust to the curse of chaos. In *International Conference on Machine Learning*, pages 4065–4074. PMLR, 2018.

[11] Matteo Pirotta, Marcello Restelli, and Luca Bascetta. Policy gradient in lipschitz markov decision processes. *Machine Learning*, 100(2):255–283, 2015.

[12] Julian Schrittwieser, Ioannis Antonoglou, Thomas Hubert, Karen Simonyan, Laurent Sifre, Simon Schmitt, Arthur Guez, Edward Lockhart, Demis Hassabis, Thore Graepel, et al. Mastering atari, go, chess and shogi by planning with a learned model. *Nature*, 588(7839):604–609, 2020.

[13] David Silver, Hado Hasselt, Matteo Hessel, Tom Schaul, Arthur Guez, Tim Harley, Gabriel Dulac-Arnold, David Reichert, Neil Rabinowitz, Andre Barreto, et al. The predictron: End-to-end learning and planning. In *International Conference on Machine Learning*, pages 3191–3199. PMLR, 2017.

[14] Arun Sai Suggala and Praneeth Netrapalli. Online non-convex learning: Following the perturbed leader is optimal. In *Algorithmic Learning Theory*, pages 845–861. PMLR, 2020.

[15] Lingxiao Wang, Qi Cai, Zhuoran Yang, and Zhaoran Wang. Neural policy gradient methods: Global optimality and rates of convergence. *arXiv preprint arXiv:1909.01150*, 2019.