Learning Value Equivalent Models on Policy Improvement Path

1 Introduction

The basic intuition of value-aware model learning [7, 6] is to assign different weights to the transitions according to the *optimal* state value. By disregarding aspects of the environment which are not related to decision-making, VAML and its variants can be shown more effective and can outperform their classical counterparts under the same capacity constraints. However, the optimal value function is not known and the VAML loss cannot be readily minimized. Thus, as shown in [9, 8], 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.

Intuitively, the model that is value equivalent (VE) with $\{V\} = \{V^\pi | \pi \in \Pi\}$ accurately describe the real environment when Π contains all possible policies. However, in most cases we cannot enumerate all possible policies and their corresponding values. One potential solution is to train the VE model with *representative* policies (and their values). The key is to determine the principled set Π_t at iteration t in the training process that characterizes an effective VE model. In this work, we propose a new VE model learning formulation based on an algorithm's *policy-improvement path* that allows good asymptotic policies and fast convergence rates.

We focus on model-based RL in two frameworks and build the connections between the value equivalence principle and online model learning. Firstly, we study sample-based policy search with VE models, where the model generated samples are used to perform (model-free) policy optimization. We show that the convergence rate depends on the Rademacher complexity of the model loss, for which we design several online learners. In general, learning VE models through the policy improvement path gives a $T^{-\frac{1}{2}}$ rate of convergence for natural policy gradient. We also analyze the iterative model learning approaches in convex settings such as linear approximations. In nonconvex settings, we provide several extensions inspired by FTPL and optimistic FTPL with predictors for tighter regret bound.

We also analyze the wider usage of value equivalence principle in model-based planning. Modern agents such as MuZero have shown the superiority of VE models in complex tasks. Recent work [8] has observed even better experimental performance when constraining the model update w.r.t. $\Pi_t = \{\pi_i\}_{i=1}^{t-1}$, instead of $\Pi_t = \{\pi_{t-1}\}$ as used in MuZero. We investigate such phenomenon and provide theoretical guidance for designing the policy cover Π_t. Besides, from the local regret perspective, we prove an approximate local optimality with an online model learner.

2 Preliminaries

Value-Aware Model Learning is a framework that addresses the *objective mismatch* issue [12, 5] in transition-predictive Model-Based Reinforcement Learning (MBRL). The conventional approach for one-step ahead predictive model learning is to minimize the KL-divergence between the empirical data and the model, which leads to the Maximum Likelihood Estimator (MLE). Objective mismatch thus arises between training the forward dynamics model and the overall goal of RL that maximizes the average policy value. To solve this misalignment between model learning and decision making,

Work in progress.

Farahmand et al. proposed VAML [7, 6, 20] that minimizes the discrepancy between the Bellman optimality operator induced by model $\hat{f}$ and the real f^* on the *optimal* value function V:

$$l(\hat{f}) = \int_{s,a} |\langle f^*(\cdot|s,a) - \hat{f}(\cdot|s,a), V\rangle|. \tag{1}$$

However, V is unknown during training and different approaches are proposed to approximate objective (1). For example, the optimal value in [7] is replaced with the supremum over a function space to minimize the worst-case discrepancy. IterVAML [6] performs planning in an iterative manner, with the unknown value replaced by the estimation in each iteration. Recently, VaGraM [20] provides several modifications towards a more practical algorithm.

Value Equivalence is another principle that supports decision-aware model learning, which states that two models are value equivalent (VE) w.r.t. a set of functions $\mathbb{V}$ and a set of policies $\mathbb{\Pi} := \{\pi|\pi : \mathcal{S} \mapsto P(\mathcal{A})\}$ if they yield the same updates under corresponding Bellman operators [9, 8]. The class $\mathcal{F}$ of models that are VE with f^* are expressed as

$$\mathcal{F}^1(\mathbb{\Pi}, \mathbb{V}) = \{\hat{f} \in \mathcal{F} : \hat{\mathcal{T}}_\pi v = \mathcal{T}_\pi v, \forall \pi \in \mathbb{\Pi}, v \in \mathbb{V}\}, \tag{2}$$

where Bellman operator $\mathcal{T}_\pi[v](s) = \mathbb{E}[r(s,a) + v(s')|f^*, \pi]$ and $\hat{\mathcal{T}}_\pi[v](s) = \mathbb{E}[\hat{r}(s,a) + v(s')|\hat{f}, \pi]$.

Define the set of *all* possible policies and functions as $\mathbb{\Pi}_{\text{all}} := \{\pi|\pi : \mathcal{S} \mapsto P(\mathcal{A})\}$ and $\mathbb{V}_{\text{all}} := \{v|v : \mathcal{S} \mapsto \mathbb{R}\}$. It is shown in [9] that $\mathcal{F}^1(\mathbb{\Pi}_{\text{all}}, \mathbb{V}_{\text{all}})$ either contains only the environment or is empty. It makes intuitive sense since the increasing number of policies and functions poses more stringent constraints on the VE model, which leads to the shrink of the VE model class.

Grimm et al. [8] further reduced the required functions for characterizing an effective model sufficient for planning. They show that the VE principle can be formulated w.r.t. policy value functions only. Specifically, define an order-k VE model class as

$$\mathcal{F}^k(\mathbb{\Pi}, \mathbb{V}) = \{\hat{f} \in \mathcal{F} : \hat{\mathcal{T}}_\pi^k v = \mathcal{T}_\pi^k v, \forall \pi \in \mathbb{\Pi}, v \in \mathbb{V}\}, \tag{3}$$

where $\mathcal{T}_\pi^k$ denote k applications of $\mathcal{T}_\pi$. When $k \to \infty$, the *Proper Value Equivalence* $\hat{\mathcal{T}}_\pi^\infty v = \mathcal{T}_\pi^\infty v$ becomes $\hat{V}^\pi = V^\pi$. The need for selecting functions $\mathbb{V}$ is removed and $\mathcal{F}^k(\mathbb{\Pi}, \mathbb{V})$ becomes $\mathcal{F}^k(\mathbb{\Pi})$.

We note one property given in [8] regarding the structure of the model class: for any $k \in \mathbb{Z}^+$, $\mathcal{F}^\infty(\mathbb{\Pi}) = \bigcap_{\pi \in \mathbb{\Pi}} \mathcal{F}^k(\{\pi\}, \{V^\pi\})$. This provides an alternative way to describe the order-∞ PVE model as the intersection of $\mathcal{F}^k$ with respect to the singleton policies in $\mathbb{\Pi}$ and their respective value functions. We denote any VE model in $\mathcal{F}^\infty(\mathbb{\Pi})$ as $\hat{f}_{\mathbb{\Pi}}$.

Sequential Rademacher Complexity is an important concept for online learning problems. We adopt the notation from [15] to define the sequential Rademacher complexity $\mathfrak{R}_T(\mathbb{L}; \boldsymbol{x}, \boldsymbol{y})$ of the loss function class $\mathbb{L} = \{(x,y) \mapsto \mathcal{L}((x,y); \hat{f})\}$. Denote a sequence of Rademacher random variables $\epsilon = (\epsilon_1, \cdots, \epsilon_T)$. For any $\mathcal{X}$-valued tree $\boldsymbol{x}$ and any $\mathcal{Y}$-valued tree $\boldsymbol{y}$, $\mathfrak{R}_T(\mathbb{L}) = \sup_{\boldsymbol{x},\boldsymbol{y}} \mathfrak{R}_T(\mathbb{L}; \boldsymbol{x}, \boldsymbol{y})$, where

$$\mathfrak{R}_T(\mathbb{L}; \boldsymbol{x}, \boldsymbol{y}) \triangleq \mathop{\mathbb{E}}_{\epsilon,\xi}\left[\sup_{\mathcal{L}\in\mathbb{L}} \sum_{t=1}^{T} \epsilon_t \mathcal{L}((\boldsymbol{x}(\epsilon), \xi_t), \boldsymbol{y}(\epsilon))\right]. \tag{4}$$

3 Value Equivalence in the Policy-Improvement Path

It is shown in [8] that the optimal policy in $\hat{f}_{\mathbb{\Pi}_{\text{all}}}$ is also an optimal policy in the environment. In practice, however, it is a stringent requirement to enumerate every $\pi \in \mathbb{\Pi}_{\text{all}}$ and learn the corresponding (proper) VE model. We now investigate if and how set $\mathbb{\Pi}_{\text{all}}$ can be effectively reduced to its subset $\mathbb{\Pi} \subset \mathbb{\Pi}_{\text{all}}$ for VE to still succeed.

We show that the application of value-equivalence principle can be tailored as the learning of a model that allows for accurate value predictions of policies in an algorithm's *policy-improvement path*.

Specifically, we define the policy-improvement path $\mathfrak{P}(\cdot) = \{\pi_1, \cdots, \pi_T\}$ as a sequence of policies given by an algorithm, *e.g.*, $\mathfrak{P}(\hat{f}_{\mathbb{\Pi}})$ formed by policy updates with $\hat{f}_{\mathbb{\Pi}}$. We provide two properties regarding the structure of $\mathfrak{P}(\hat{f}_{\mathbb{\Pi}})$.

Firstly, suppose w.l.o.g. that we are interested in obtaining a unique optimal policy π^* at the end of training. When $\pi_T = \pi^*$, all possible policy-improvement paths compose a tree-like structure in which the optimal policy is the root, the first level has all policies π_{T-1} that lead to π^* with one iteration of policy update, and so on. This indicates that an optimal VE model should always give accurate value predictions of V^{π^*}.

Secondly, it suffices to set $\mathbb{\Pi}$ such that $\hat{f}_{\mathbb{\Pi}}$ is value equivalent with all $\pi \in \mathfrak{P}(\hat{f}_{\mathbb{\Pi}})$. The long-term performance of VE-based planning is only affected by the model's ability to approximate policy values in its policy improvement path. We formalize these two properties with the following theorem.

Theorem 1. *For a value equivalent model $\hat{f}_{\mathbb{\Pi}}$ learned on $\mathbb{\Pi}$, the policy path $\mathfrak{P}(\hat{f}_{\mathbb{\Pi}})$ is formed by following the update rule $\pi_{t+1} = \operatorname{argmax}_\pi \hat{\mathcal{T}}_\pi^k V^{\pi_t}$. Suppose the VE error in $\mathfrak{P}(\hat{f}_{\mathbb{\Pi}})$ is bounded by some $\epsilon \geq 0$ for distribution μ, i.e.,*

$$\|V^\pi - \hat{\mathcal{T}}_\pi^k V^\pi\|_\mu \leq \epsilon, \forall \pi \in \mathfrak{P}(\hat{f}_{\mathbb{\Pi}}),$$

where $\hat{\mathcal{T}}$ is induced by $\hat{f}_{\mathbb{\Pi}}$. Then for any reference policy $\overline{\pi}$, the value gap in the real MDP between $\overline{\pi}$ and the asymptotic policy π_∞ in $\mathfrak{P}(\hat{f}_{\mathbb{\Pi}})$ is bounded by

$$\limsup_{t\to\infty} \left\|V^{\overline{\pi}} - V^{\pi_t}\right\|_\mu \leq \epsilon \left(I + \frac{e_{\overline{\pi}}}{\epsilon} + 2\gamma^k (I - \gamma^k \hat{\mathcal{P}}_{\pi_{t+1}}^k)^{-1} \hat{\mathcal{P}}_{\pi_{t+1}}^k\right)(I - \gamma^k \hat{\mathcal{P}}_{\overline{\pi}}^k),$$

where $e_{\overline{\pi}} = \|V^{\overline{\pi}} - \hat{\mathcal{T}}_{\overline{\pi}}^k V^\pi\|_\mu$ is the VE error of $\hat{f}_{\mathbb{\Pi}}$ w.r.t. the reference policy $\overline{\pi}$.

Proof. See Appendix 8.1. □

We can freely set the reference policy $\overline{\pi}$, *e.g.*, to the optimal policy, and obtain the necessary conditions for global asymptotic optimality: the VE error w.r.t. all $\pi \in \{\pi^*, \mathfrak{P}(\hat{f}_{\mathbb{\Pi}})\}$ should be small.

Theorem 1 shows the structure of policy set in VE learning and the promising results of replacing the large $\mathbb{\Pi}_{\text{all}}$ with its subset $\mathbb{\Pi}$. Unfortunately, it does not directly suggest a training objective or support a practical algorithm, as both the optimal policy and the policy-improvement path are not known beforehand. For this reason, we will discuss how the intuition in this section can guide the choices of policy sets $\mathbb{\Pi}$. We will also provide positive theoretical guarantees for different VE algorithmic frameworks in Sec. 4 and 5 based on similar online learning perspectives.

4 Sample-Based Policy Search with VE Models

4.1 Convergence Rate of Natural Policy Gradient

In this section, we study a specific model-based policy search framework that generates samples from a learned model and performs policy optimization with the samples. In particular, we analyze natural policy gradient with value equivalent models and study the property discounted MDPs with function approximations.

Denote the state value and state-action value by unrolling model $\hat{f}$ with policy π as $\hat{V}^\pi$ and $\hat{Q}^\pi$, respectively. That is, $\hat{V}^\pi = \sum_{h=1}^\infty \gamma^h \hat{r}(\hat{s}_h, \hat{a}_h)$, where $\hat{s}_h \sim \hat{f}(\cdot|\hat{s}_{h-1}, \hat{a}_{h-1})$, $\hat{a}_h \sim \pi(\cdot|\hat{s}_h)$, and the objective is $J(\pi) = \mathbb{E}_{s_0}\left[V^\pi(s_0)\right]$. For natural gradient $\tilde{\nabla}_\theta J(\theta_t)$, a Fisher information matrix is adopted for update $\theta_{t+1} - \theta_t = \eta \tilde{\nabla}_\theta J(\theta_t)$, where

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

In practice, $\hat{J}(\pi_{\theta_t})$ is estimated with B samples replacing the expectation.

Theorem 2. *Assume V_{π_θ} is L-smooth in θ and for all $s \in \mathcal{S}$ and $a \in \mathcal{A}$ that $\log \pi_\theta(a|s)$ is a ι-smooth function of θ. For $\eta \leq \frac{1}{L}$,*

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

where the gradient bias b_t and the upper bound v_t of gradient variance are defined by $b_t = \left\|\nabla_\theta J(\pi_{\theta_t}) - \mathbb{E}[\nabla_\theta \hat{J}(\pi_{\theta_t})]\right\|_2$ and $v_t = \mathbb{E}\left[\left\|\nabla_\theta \hat{J}(\pi_{\theta_t}) - \mathbb{E}[\nabla_\theta \hat{J}(\pi_{\theta_t})]\right\|_2^2\right]$.

Proof. See Appendix 8.2. □

We analyze the gradient bias and variance terms below as their sums might still be linear in T.

4.2 Online VE Model Learning

For convenience, we assume a bounded gradient variance. We show in Appendix 8.3 that a wide range of RL problems can be contained with the assumptions held.

Assumption 1. *There exists an absolute constant $c \geq 0$ such that $v_t \leq c^2/B$ for all $t \in [T]$.*

We first focus on the likelihood ratio gradient estimator that leverages the score function to calculate gradient. The gradient bias can be represented by $b_{t+1} = \|V^{\pi_t} - \hat{\mathcal{T}}_{\pi_t}^{\infty} V^{\pi_t}\|$, which corresponds to a proper VE loss [8] w.r.t. a single policy π_t. Different from pure model-free algorithms, in MBPG (Alg. 1) multiple PG updates are performed with model-generated data (for sample complexity reduction). In other words, the learned model $\hat{f}_t$ continues to have an impact on the following n policy updates. Thus, simply setting $\hat{f}_t = \operatorname{argmin}_{\hat{f}} b_t$ only guarantees a low-bias gradient update from π_{t-1} to π_t, leading to potentially large bias for the following $n-1$ policy gradients.

Formally, in the convergence rate Theorem 2, we need the accumulate $\sum_{t=1}^{T} b_t$ to be small. That is, the true objective for the model learner $\mathcal{R}$ is to minimize the following regret REG_T

$$\min_{\hat{f}_1,\ldots,\hat{f}_{\lfloor \frac{T}{n} \rfloor n}} \sum_{t=1}^{T} \mathcal{L}_t \left(\left((s,a), \hat{V}^{\pi} \right) ; \hat{f}_{\lfloor \frac{t}{n} \rfloor n} \right) \triangleq \min_{\hat{f}_1,\ldots,\hat{f}_{\lfloor \frac{T}{n} \rfloor n}} \sum_{t=1}^{T} \left\| V^{\pi_t} - \hat{\mathcal{T}}_{\pi_t}^{\infty} V^{\pi_t} \right\|,$$

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

For model learner gives predictions based on $\mathcal{R}(\{x_i, y_i\}_{i=1}^{t-1})$, where x and y are training inputs (s,a) and predictions $\hat{V}^{\pi}$, standard results in [15] showed the existence of an online model learner $\mathcal{R}$ that gives $\sum_{t=1}^{T} b_t \leq 2n\mathfrak{R}_T(\mathbb{L})$. Typically, we have the Rademacher complexity $\mathfrak{R}_T(\mathbb{L})$ bounded by $\tilde{O}(\sqrt{T})$, which is desired in the convergence rate theorem. Next, we discuss several such designs.

General model learners. Unlike Π that contains predefined policies, we denote Π_t as the policy set that changes during training. We first provide a learner in general settings which maintains $\Pi_t = \{\pi_i\}_{i=1}^{t-1}$. Define the model learning error at iteration t as $\delta_t = \sum_{i=1}^{t-1} \mathbb{E}_{s \sim d^{\pi_i}} \left\| (\mathcal{T}_{\pi_t}^{\infty} - \hat{\mathcal{T}}_{\pi_t}^{\infty}) V^{\pi_i} \right\|$, where $\hat{\mathcal{T}}_{\pi_t}$ is induced by model $\hat{f}_{\lfloor \frac{t}{n} \rfloor n}$.

Proposition 1. *With model learning error $\delta = \max_{t \in [1,T]} \delta_t$,*

$$REG_T \leq (\delta + \frac{1}{(1-\gamma)}) \sqrt{T(1 + \frac{1}{(1-\gamma)^2}) \log(1 + \frac{T}{(1-\gamma)^2})}. \tag{5}$$

Suppose we have access to an offline optimization oracle that gives a finite constant σ, then regret $REG_T \leq O(\sqrt{T \log T})$.

Selecting past policies in the policy-improvement path as Π_t has shown its effectiveness in recent implementation [8], which further improves the state-of-the-art MuZero performance on Atari tasks.

Online convex learning. Despite the bound developed in general settings, for linear function approximations, the model learning process becomes an online *convex* optimization problem, where online gradient descent suffices to obtain $O(\sqrt{T})$ regret. In this case, we can simply leverage *iterative* model learning algorithms that set $\Pi_t = \pi_{t-1}$ and minimize $\delta_t' = \|V^{\pi_{t-1}} - \hat{\mathcal{T}}_{\pi_{t-1}}^{\infty} V^{\pi_{t-1}}\|$. When the underlying transition of a high-dimensional MDP can be linearly modeled, we may also first convert the states to an abstract space for the convexity to be satisfied.

Extensions in nonconvex settings. Several extensions of Proposition 1 can be obtained in nonconvex settings, *e.g.*, with nonlinear function approximation. In nonconvex online learning, a heuristic yet provable algorithm is *Follow the Perturbed Leader* (FTPL), which adds a perturbation noise to the loss: $\delta_t' = \delta_t + \sigma_t$. For non-oblivious loss which is not adversarially chosen, *e.g.*, the VE model loss, we may expect it predictable from the past. Thus, minimizing $\delta'_t = \delta_t + M_t + \sigma_t$ with a predictor M_t leads to optimistic algorithms (OFTPL) and tighter regret bound when M_t encodes useful information. For example, we may simply set $M_t = \mathcal{L}_{t-1}$ to emphasise the policy similarities between consecutive iterations, or importance weighted past losses $M_t = \frac{1}{t-1}\sum_{i=1}^{t-1}\alpha_i \mathcal{L}_i$ for fading memory statistics, or combined with additional VE losses given by random policies.

We show that FTPL and OFTPL achieve $O(\sqrt{T})$ regret with access to an offline optimization oracle.

Proposition 2. *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 $g(x^*) - \langle \sigma, x^* \rangle \leq \inf_x f(x) - \langle \sigma, x \rangle + \alpha + \beta\|\sigma\|_1$. Suppose the losses encountered by the model learner are l-Lipschitz w.r.t l_1 norm and σ_t is an exponential distribution where each coordinate is sampled from Exp(λ_t). For $\lambda = O(T^{-\frac{1}{2}})$, $\alpha = O(T^{-\frac{1}{2}})$, and $\beta = O(T^{-1})$, we have for FTPL and OFTPL that $\mathbb{E}\left[REG_T\right] \leq O(\sqrt{T})$.*

Discussions. All the above designs guarantee an $\tilde{O}(T^{-\frac{1}{2}})$ rate of convergence in Theorem 2. However, value equivalence might not suffice to support broader problems of interest such as pathwise gradient estimator and PG without natural gradients. Instead, policy-aware or gradient-aware model learning [1, 4] should be analyzed since they directly learn a model for accurate gradient estimation, by imposing the score magnitude as an additional weight.

5 VE-Model-Based Planning

Another important usage of models is to perform model-based planning. Much attention has been paid to the planning with value equivalent models, such as Predictron, MuZero, VPN, and PVE.

5.1 Case Study of MuZero

MuZero [18] is a modern architecture that shows the power of value equivalence principle in model-based planning. The agent maintains a predictive model that outputs abstract state transitions, values, and rewards, then the action returned by MCTS (with upper confidence bound applied to trees) is executed in the real MDPs. The two stages are carried out iteratively.

Grimm et al. [8] demonstrate the close connection between the VE loss and the MuZero objective. Specifically, MuZero's per-state model loss can be expressed as:

$$l(s_h) = \sum_{i=0}^{k}(V_{h+i} - v(z_h^i))^2 + (r_{h+i} - \hat{r}(z_h^i))^2,$$

where $V_{h+i} = r_{h+i} + \cdots + r_{h+i+k'-1} + v(z^0_{h+i+k'})$ is the k'-step bootstrapping. z_h^i is the abstract state after i step model rollout from the real state s_h. $s_{h:H}$, $a_{h:H}$, and $r_{h:H}$ are the real trajectory.

It is shown that the MuZero model loss upper bounds the order-k VE loss, *i.e.*, $\|V^{\pi_t} - \hat{\mathcal{T}}^k_{\pi_t}V^{\pi_t}\|^2_{d_\pi} \leq c\,\mathbb{E}_{d_\pi}[l(s)]$. Optimizing the MuZero model is equivalent to finding a VE model w.r.t. $\Pi_t = \pi_{t-1}$.

Since experiments in [8] show that even better results can be obtained by adding VE loss induced by past policies to MuZero, we investigate why $\Pi_t = \{\pi_i\}_{i=1}^{t-1}$ is preferred over $\Pi_t = \{\pi_{t-1}\}$.

UCB1 is an algorithm for multi-armed bandit that achieves logarithm regret with the number of actions taken. When applied to MCTS (UCT [11] or PUCT [16]), convergence to the optimal policy can be shown, if given access to the accurate payoffs. However, learning a unified model for value prediction is exactly the objective of MuZero, which results in iterative executions of tree search and model update. Thus, to derive the convergence of MuZero, we also need the predicted model value $\hat{\mathcal{T}}^k_{\pi_t}V^{\pi_t}$ to converge to the true MDP as iteration t increases, otherwise it can get stuck at local optima. This can be handled by online model learning.

Formally, for an optimal arm (action) with reward μ^* at a given state, define the regret to be the loss caused by the policy not always playing the best arm at the end of each iteration (T in total), *i.e.*, $\text{REG}_T = T\mu^* - \mathbb{E}[\sum_{t=1}^{T}\mu_{it}]$.

Proposition 3. *Denote the number of MCTS runs with model $\hat{f}_t$ in a single iteration t as n. Suppose $n = O(T)$. For MuZero with model learner $\mathcal{R}$, denote the Rademacher complexity for the class L where $l \in L$ as $\mathfrak{R}_T$. Then the regret REG$_T$ is bounded by*

Algorithm 1 Model-Based Policy Gradient

```
Initialize policy $\pi_0$, model $\hat{f}$, $t = 1$
for $\lfloor \frac{T}{n} \rfloor n$ iterations do
    Execute policy $\pi_{t-1}$ in the real MDP
    Update VE model $\hat{f}_t$ (by minimizing $\delta_t$ or $\delta'_t$)
    for $n$ gradient updates do
        Calculate gradient $\nabla_\theta \hat{J}(\pi_{t-1})$
        Perform PG and obtain updated $\pi_t$
        $t \leftarrow t + 1$
    end for
end for
```

Algorithm 2 Model-Based Planning

```
Initialize policy $\pi_0$, model $\hat{f}$, $t = 1$
for $\lfloor \frac{T}{n} \rfloor n$ iterations do
    Execute policy $\pi_{t-1}$ in the real MDP
    Update VE model $\hat{f}_t$
    for $n$ planning iterations do
        Perform model-based planning (e.g., MCTS)
        $t \leftarrow t + 1$
    end for
    Return $\pi_t$
end for
```

$$\mathit{REG}_T \leq O(\mathfrak{R}_T) + O(\sqrt{T \log T}).$$

Proof. See Appendix 8.6. □

It is typical for online model learners that $\mathfrak{R}_T$ is bounded by $\tilde{O}(\sqrt{T})$. For example, define the model loss as $\sum_{\pi \in \Pi_t} \mathbb{E}_{d_\pi}[l(s)]$ and $\Pi_t = \{\pi_i\}_{i=1}^{t-1}$.

5.2 Planning in General Settings

Note that the UCB algorithms used in MCTS and MuZero can only be applied to independent arms. For more general settings such as nonlinear models, recent works [3] have shown pessimistic results towards *global* convergence. Specifically, Dong et al. show that the complexity (*e.g.*, Eluder dimension [17, 13]) of nonlinear models cannot be polynomially bounded even for one-layer neural networks. It is thus suggested to find an (ϵ_g, ϵ_h)-approximate local optimal policy π^* and use local regret instead as the evaluation metric, defined as

$$\text{REG}^{\text{LC}}_{\epsilon_g,\epsilon_h}(T) = \sum_{t=1}^{T} \left(\sup_{\pi^* \in \mathfrak{C}_{\epsilon_g,\epsilon_h}} V^{\pi^*} - V^{\pi_t} \right)$$

Proposition 4. *Define the model loss $l_t = \|V^{\pi_{t-1}} - \hat{V}^{\pi_{t-1}}\| + \|V^{\pi_t} - \hat{V}^{\pi_t}\|$ and $\mathfrak{R}_T(L)$ the Rademacher complexity for class L, where $l \in L$. Under Lipschitz assumptions, we have for the $(\epsilon, 6\sqrt{\zeta\epsilon})$-regret that*

$$\mathit{REG}^{\mathit{LC}}_{\epsilon,6\sqrt{\zeta\epsilon}}(T) \leq O(\sqrt{T\mathfrak{R}_T}).$$

Notably, the loss l_t contains the prediction error w.r.t. both π_{t-1} at the previous iteration and π_t, the *upcoming* policy. The intuition is to optimize the model so that monotonic policy improvement can be obtained to achieve the local optima within the convex hull. The simplest way to meet such condition is to give accurate predictions for successive policy updates. This motivates the usage of VE objectives with online model learners, such as FTL-style algorithms.

6 Experiments

7 Related Work

MuZero, VPN, etc.

8 Proofs

8.1 Proof of Theorem 1

Proof. Define in the real MDP the value gap between policy π_t and $\overline{\pi}$ as $x_t = V^{\overline{\pi}} - V^{\pi_t}$, the policy gain between successive policies as $g_t = V^{\pi_{t+1}} - V^{\pi_t}$.

We decompose and bound x_{t+1} as

$$\begin{aligned} x_{t+1} &= V^{\overline{\pi}} - V^{\pi_{t+1}} \\ &= V^{\overline{\pi}} - \hat{\mathcal{T}}^k_{\overline{\pi}} V^{\overline{\pi}} + \hat{\mathcal{T}}^k_{\overline{\pi}} V^{\overline{\pi}} - \hat{\mathcal{T}}^k_{\overline{\pi}} V^{\pi_t} + \hat{\mathcal{T}}^k_{\overline{\pi}_T} V^{\pi_t} - \hat{\mathcal{T}}^k_{\pi_{t+1}} V^{\pi_t} \\ &+ \hat{\mathcal{T}}^k_{\pi_{t+1}} V^{\pi_t} - \hat{\mathcal{T}}^k_{\pi_{t+1}} V^{\pi_{t+1}} + \hat{\mathcal{T}}^k_{\pi_{t+1}} V^{\pi_{t+1}} - V^{\pi_{t+1}}. \end{aligned} \tag{6}$$

Due to the VE error bound and the optimality of policy π_{t+1}, *i.e.*, $\pi_{t+1} = \mathrm{argmax}_\pi \hat{\mathcal{T}}^k_\pi V^{\pi_t}$, we have

$$\|x_{t+1}\|_\mu \leq e_{\overline{\pi}} + \gamma^k \hat{\mathcal{P}}^k_{\overline{\pi}} \|x_t\|_\mu - \gamma^k \hat{\mathcal{P}}^k_{\pi_{t+1}} \|g_t\|_\mu + \epsilon, \tag{7}$$

For the policy gain g_t, we have

$$\begin{aligned} g_t &= V^{\pi_{t+1}} - V^{\pi_t} \\ &= V^{\pi_{t+1}} - \hat{\mathcal{T}}^k_{\pi_{t+1}} V^{\pi_{t+1}} + \hat{\mathcal{T}}^k_{\pi_{t+1}} V^{\pi_{t+1}} - \hat{\mathcal{T}}^k_{\pi_{t+1}} V^{\pi_t} \\ &+ \hat{\mathcal{T}}^k_{\pi_{t+1}} V^{\pi_t} - \hat{\mathcal{T}}^k_{\pi_t} V^{\pi_t} + \hat{\mathcal{T}}^k_{\pi_t} V^{\pi_t} - V^{\pi_t} \\ &\geq -2\epsilon + \gamma^k \hat{\mathcal{P}}^k_{\pi_{t+1}} g_t, \end{aligned} \tag{8}$$

where the inequality holds since $\hat{\mathcal{T}}^k_{\pi_{t+1}} V^{\pi_t} \geq \hat{\mathcal{T}}^k_{\pi_t} V^{\pi_t}$.

Thus,

$$-g_t \leq 2\epsilon (I - \gamma^k \hat{\mathcal{P}}^k_{\pi_{t+1}})^{-1}. \tag{9}$$

Plugging into Equation (7), we obtain

$$\|x_{t+1}\|_\mu \leq e_{\overline{\pi}} + \epsilon + 2\epsilon\gamma^k (I - \gamma^k \hat{\mathcal{P}}^k_{\pi_{t+1}})^{-1} \hat{\mathcal{P}}^k_{\pi_{t+1}} I + \gamma^k \hat{\mathcal{P}}^k_{\overline{\pi}} \|x_t\|_\mu. \tag{10}$$

By taking the limit superior component-wise, we have

$$\limsup_{t \to \infty} \|x_t\|_\mu \leq \epsilon \left(I + \frac{e_{\overline{\pi}}}{\epsilon} + 2\gamma^k (I - \gamma^k \hat{\mathcal{P}}^k_{\pi_{t+1}})^{-1} \hat{\mathcal{P}}^k_{\pi_{t+1}} \right) (I - \gamma^k \hat{\mathcal{P}}^k_{\overline{\pi}}).$$

□

8.2 Proof of Theorem 2

Proof. Denote $\beta_t = \frac{\theta_{t+1} - \theta_t}{\eta} = \tilde{\nabla}_\theta \hat{J}(\theta_t)$. By Lipschitz assumption we have

$$\begin{aligned} J(\pi_{\theta_{t+1}}) - J(\pi_{\theta_t}) &\geq \nabla_\theta J(\pi_{\theta_t})^\top (\theta_{t+1} - \theta_t) - \frac{L}{2} \|\theta_{t+1} - \theta_t\|_2^2 \\ &= \eta \nabla_\theta J(\pi_{\theta_t})^\top \beta_t - \frac{L\eta^2}{2} \|\beta_t\|_2^2. \end{aligned} \tag{11}$$

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

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

Then we bound $\nabla_\theta J(\pi_{\theta_t})^\top \beta_t$ in Eq. (11) by bounding the resulting three terms.

$$\left| \Big(\nabla_\theta J(\pi_{\theta_t}) - \mathbb{E}[\nabla_\theta \hat{J}(\pi_{\theta_t})] \Big)^\top \beta_t \right| \leq \|\beta_t\|_2 \|\nabla_\theta J(\pi_{\theta_t}) - \mathbb{E}[\nabla_\theta \hat{J}(\pi_{\theta_t})]\|_2 = \|\beta_t\|_2 b_t \tag{12}$$

$$\Big(\nabla_\theta \hat{J}(\pi_{\theta_t}) - \mathbb{E}[\nabla_\theta \hat{J}(\pi_{\theta_t})] \Big)^\top \beta_t \leq \frac{\|\nabla_\theta \hat{J}(\pi_{\theta_t}) - \mathbb{E}[\nabla_\theta \hat{J}(\pi_{\theta_t})]\|_2^2}{2} + \frac{\|\beta_t\|_2^2}{2} \tag{13}$$

$$\nabla_\theta \hat{J}(\pi_{\theta_t})^\top \beta_t \geq \|\beta_t\|_2^2, \tag{14}$$

where Eq. (14) 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. (11) by

$$J(\pi_{\theta_{t+1}}) - J(\pi_{\theta_t}) \geq \eta \left(-\|\beta_t\|_2 b_t - \frac{\|\nabla_\theta \hat{J}(\pi_{\theta_t}) - \mathbb{E}[\nabla_\theta \hat{J}(\pi_{\theta_t})]\|_2^2}{2} + \frac{\|\beta_t\|_2^2}{2} \right) - \frac{L\eta^2}{2} \|\beta_t\|_2^2. \tag{15}$$

By taking expectation in Eq. (15), we have

$$\begin{aligned} \left(\frac{\eta}{2} - \frac{L\eta^2}{2} \right) \mathbb{E}[\|\beta_t\|_2^2] &\leq \mathbb{E}[J(\pi_{\theta_{t+1}}) - J(\pi_{\theta_t})] + \eta\, \mathbb{E}[\|\beta_t\|_2] b_t + \frac{\eta}{2} v_t \\ &\leq \mathbb{E}[J(\pi_{\theta_{t+1}}) - J(\pi_{\theta_t})] + b_t + \frac{\eta}{2} v_t \end{aligned} \tag{16}$$

We have for $\eta \leq \frac{1}{L}$ that

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

By smoothness, we have

$$\log \frac{\pi_{t+1}(a|s)}{\pi_t(a|s)} - \nabla_\theta \log \pi_t(a|s)(\theta_{t+1} - \theta_t) \geq -\frac{\iota}{2} \left\| \theta_{t+1} - \theta_t \right\|_2^2. \tag{18}$$

Then we obtain by the definition of KL divergence that

$$\begin{aligned} \mathbb{E}_{s \sim d^{\pi^*}} [D_{\mathrm{KL}}(\pi^* || \pi_t) - D_{\mathrm{KL}}(\pi^* || \pi_{t+1})] &= \mathbb{E}_{s \sim d^{\pi^*}} \mathbb{E}_{a \sim \pi^*} \left[\log \frac{\pi_{t+1}(a|s)}{\pi_t(a|s)} \right] \\ &\geq \mathbb{E}_{s \sim d^{\pi^*}} \mathbb{E}_{a \sim \pi^*} \left[\nabla_\theta \log \pi_t(a|s)(\theta_{t+1} - \theta_t) - \frac{\iota}{2} \left\| \theta_{t+1} - \theta_t \right\|_2^2 \right] \\ &= \mathbb{E}_{s \sim d^{\pi^*}} \mathbb{E}_{a \sim \pi^*} \left[\eta A^{\pi_t}(s,a) - \frac{\iota \eta^2}{2} \left\| \beta_t \right\|_2^2 \right] \\ &= (1-\gamma)\eta \left(J(\pi^*) - J(\pi_t) \right) - \frac{\iota \eta^2}{2} \mathbb{E}[\|\beta_t\|_2^2]. \end{aligned} \tag{19}$$

where the last equality follows the performance difference lemma [10].

Thus,

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

□

8.3 Examples

In Theorem 2 it is assumed that V_{π_θ} is L-smooth in θ. We first provide a possible condition in Example 1 for this assumption to hold when the reward r and transition function f are both Lipschitz continuous and smooth. Such assumptions are also consider by many previous works [2, 21, 14].

Example 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 $24H^5 \bar{L}_r \bar{L}_f^{4H}$.*

We also give an example setting that satisfies Assumption 1.

Example 2. *With function approximations, for a stochastic model 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. Assume a Lipschitz continuous $\hat{f}$ and $\nabla \hat{r}$. For initial state distribution $\|s_0 - \mathbb{E}[s_0]\|_2 \le \|\mathcal{S}\|_2$ and $\|a\|_2 \le 1$, we have that $v_r \le c_t^2/B$ where*

$$c_t = \left(\frac{L_{\nabla \hat{r}}(L_{\hat{f}} + \sigma_\xi)}{L_{\hat{f}} - 1} + L_{\nabla \hat{r}} \|\mathcal{S}\|_2 \right) \frac{2L_{\hat{f}}(L_{\hat{f}}^k - 1)}{L_{\hat{f}} - 1} + kL_{\nabla \hat{r}}. \tag{20}$$

Setting c as the largest c_t satisfies Assumption 1.

Proof. First, we have

$$\begin{aligned} \|\hat{s}_{h+1} - \mathbb{E}[\hat{s}_{h+1}]\|_2 &= \left\| \hat{f}(\hat{s}_h, \hat{a}_h) + \xi_h - \mathbb{E}[\hat{f}(\hat{s}_h, \hat{a}_h)] \right\|_2 \\ &\le L_{\hat{f}} \sqrt{\|\hat{s}_h - \mathbb{E}[\hat{s}_h]\|_2^2 + 1} + \sigma_\xi \\ &\le L_{\hat{f}} \|\hat{s}_h - \mathbb{E}[\hat{s}_h]\|_2 + L_{\hat{f}} + \sigma_\xi. \end{aligned} \tag{21}$$

Since $\|s_0 - \mathbb{E}[s_0]\|_2 \le \|\mathcal{S}\|_2$, we have for $h \ge 2$ that

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

.

Besides,

$$\begin{aligned} \left\| \nabla_\theta \hat{J}(\pi_{\theta_t}) - \mathbb{E}[\nabla_\theta \hat{J}(\pi_{\theta_t})] \right\|_2 &= \left\| \sum_{h=0}^{k} \nabla \hat{r}(\hat{s}_h, \hat{a}_h) - \mathbb{E}[\nabla \hat{r}(\hat{s}_h, \hat{a}_h)] \right\|_2 \\ &\le \sum_{h=0}^{k} L_{\nabla \hat{r}} \left(\|\hat{s}_h - \mathbb{E}[\hat{s}_h]\|_2 + 1 \right) \\ &\le \left(\frac{L_{\nabla \hat{r}}(L_{\hat{f}} + \sigma_\xi)}{L_{\hat{f}} - 1} + L_{\nabla \hat{r}} \|\mathcal{S}\|_2 \right) \frac{2L_{\hat{f}}(L_{\hat{f}}^k - 1)}{L_{\hat{f}} - 1} + kL_{\nabla \hat{r}}. \end{aligned} \tag{22}$$

□

8.4 Proof of Proposition 1

Proof.

$$\begin{aligned} \left\| V^{\pi_t} - \hat{\mathcal{T}}_{\pi_t}^k V^{\pi_t} \right\|^2 &= \left\| (\mathcal{T}_{\pi_t}^k - \hat{\mathcal{T}}_{\pi_t}^k) V^{\pi_t} \right\|^2 \\ &\le \|V^{\pi_t}\|_{\mathcal{V}_{t-1}^{-1}}^2 \left(\sum_{i=1}^{t-1} \mathop{\mathbb{E}}_{s \sim d^{\pi_i}} \left\| (\mathcal{T}_{\pi_t}^k - \hat{\mathcal{T}}_{\pi_t}^k) V^{\pi_i} \right\|^2 \right) \\ &\le (\delta_t^2 + \frac{1}{(1-\gamma)^2}) \|V^{\pi_t}\|_{\mathcal{V}_{t-1}^{-1}}^2 \end{aligned} \tag{23}$$

where $\mathcal{V}_t = 1 + \sum_{i=1}^{t} \mathbb{E}_{s\sim d^{\pi_i}}[V^{\pi_i}(s)^2]$. By writing it in a recursive form,

$$\begin{aligned}
\mathcal{V}_t &= \mathcal{V}_{t-1} + \mathop{\mathbb{E}}_{s\sim d^{\pi_t}}[V^{\pi_i}(s)^2] \\
&= \mathcal{V}_{t-1}\left(1 + \mathcal{V}_{t-1}^{-1} \mathop{\mathbb{E}}_{s\sim d^{\pi_t}}[V^{\pi_i}(s)^2]\right)
\end{aligned} \tag{24}$$

Then we have

$$\begin{aligned}
\log \mathcal{V}_t &= \log \mathcal{V}_{t-1} + \log\left(1 + \mathcal{V}_{t-1}^{-1} \mathop{\mathbb{E}}_{s\sim d^{\pi_t}}[V^{\pi_i}(s)^2]\right) \\
&\geq \log \mathcal{V}_{t-1} + \frac{\mathcal{V}_{t-1}^{-1}\,\mathbb{E}_{s\sim d^{\pi_t}}[V^{\pi_i}(s)^2]}{1 + \mathcal{V}_{t-1}^{-1}\,\mathbb{E}_{s\sim d^{\pi_t}}[V^{\pi_i}(s)^2]} \\
&\geq \log \mathcal{V}_{t-1} + \frac{\|V^{\pi_t}\|^2_{\mathcal{V}_{t-1}^{-1}}}{1 + \frac{1}{(1-\gamma)^2}}.
\end{aligned} \tag{25}$$

Thus,

$$\begin{aligned}
\sum_{t=1}^{T} \|V^{\pi_t}\|_{\mathcal{V}_{t-1}^{-1}} &\leq \sqrt{T}\sqrt{\sum_{t=1}^{T} \|V^{\pi_t}\|^2_{\mathcal{V}_{t-1}^{-1}}} \\
&\leq \sqrt{T}\sqrt{\sum_{t=1}^{T}(1 + \frac{1}{(1-\gamma)^2})\,(\log \mathcal{V}_t - \log \mathcal{V}_{t-1})} \\
&= \sqrt{T}\sqrt{(1 + \frac{1}{(1-\gamma)^2})\log \mathcal{V}_T} \\
&\leq \sqrt{T(1 + \frac{1}{(1-\gamma)^2})\log(1 + \frac{T}{(1-\gamma)^2})}
\end{aligned} \tag{26}$$

Plugging Eq. (26) into Eq, (23) gives the result.

□

8.5 Proof of Proposition 2

Proof. For regret REG_T, we have

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

Note that we only optimize the model at iteration $\lfloor \frac{T}{n} \rfloor n$, 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

$$\begin{aligned}\mathbb{E}\left[\sum_{t=1}^{T} \mathcal{L}_t(\hat{f}_{\lfloor \frac{T}{n} \rfloor n})\right] &\leq l \sum_{t=1}^{T} \frac{n(n+1)}{2}\, \mathbb{E}[\|\hat{f}_t - \hat{f}_{t+1}\|_1] + \sum_{t=1}^{T} \mathbb{E}[\mathcal{L}_t(\hat{f}_{t+1})] \\ &\leq l \sum_{t=1}^{T} \frac{n(n+1)}{2}\, \mathbb{E}[\|\hat{f}_t - \hat{f}_{t+1}\|_1] + \frac{d(\beta T + D)}{\lambda} + \alpha T.\end{aligned} \tag{27}$$

The stability term $\mathbb{E}[\|\hat{f}_t - \hat{f}_{t+1}\|_1]$ remains to be bounded. From Lemma 5 and Lemma 6 in [19], we have

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

Plugging the above bound in Equation (27) gives the bound of expected regret.

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

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

8.6 Proof of Proposition 3

Proof. Denote $a^*_{\hat{f}}$ as the optimal arm under $\hat{f}$ and $\mu^*_{\hat{f}}$ as the return by playing $a^*_{\hat{f}}$. Then we can decompose the regret by

$$\begin{aligned}\text{REG}_T &= T\mu^* - \mathbb{E}[\sum_{t=1}^{T} \mu_{it}] \\ &= T\mu^* - \sum_{t=1}^{T} \mu^*_{\hat{f}_t} + \sum_{t=1}^{T} \mu^*_{\hat{f}_t} - \mathbb{E}[\sum_{t=1}^{T} \mu_{it}] \\ &\leq T\mu^* - \sum_{t=1}^{T} \mu^*_{\hat{f}_t} + O(T\sqrt{\frac{\log n}{n}}) \\ &\leq O(\mathfrak{R}_T) + O(\sqrt{T \log T}),\end{aligned} \tag{28}$$

where the first inequality follows the results in PUCT [16]. In the last inequality, the first term is due to the property of Rademacher complexity and the second term is because $n = O(T)$ and $O(T\sqrt{\frac{\log n}{n}}) \leq O(T\sqrt{\frac{\log T}{T}}) \leq O(\sqrt{T \log T})$. □

References

[1] Romina Abachi, Mohammad Ghavamzadeh, and Amir-massoud Farahmand. Policy-aware model learning for policy gradient methods. *arXiv preprint arXiv:2003.00030*, 2020.

[2] 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.

[3] Kefan Dong, Jiaqi Yang, and Tengyu Ma. Provable model-based nonlinear bandit and reinforcement learning: Shelve optimism, embrace virtual curvature. *Advances in Neural Information Processing Systems*, 34, 2021.

[4] Pierluca D'Oro, Alberto Maria Metelli, Andrea Tirinzoni, Matteo Papini, and Marcello Restelli. Gradient-aware model-based policy search. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 3801–3808, 2020.

[5] Benjamin Eysenbach, Alexander Khazatsky, Sergey Levine, and Ruslan Salakhutdinov. Mismatched no more: Joint model-policy optimization for model-based rl. *arXiv preprint arXiv:2110.02758*, 2021.

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

[7] 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.

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

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

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

[11] Levente Kocsis and Csaba Szepesvári. Bandit based monte-carlo planning. In *European conference on machine learning*, pages 282–293. Springer, 2006.

[12] Nathan Lambert, Brandon Amos, Omry Yadan, and Roberto Calandra. Objective mismatch in model-based reinforcement learning. *arXiv preprint arXiv:2002.04523*, 2020.

[13] Ian Osband and Benjamin Van Roy. Model-based reinforcement learning and the eluder dimension. *Advances in Neural Information Processing Systems*, 27, 2014.

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

[15] Alexander Rakhlin, Karthik Sridharan, and Ambuj Tewari. Online learning: Stochastic, constrained, and smoothed adversaries. *Advances in neural information processing systems*, 24, 2011.

[16] Christopher D Rosin. Multi-armed bandits with episode context. *Annals of Mathematics and Artificial Intelligence*, 61(3):203–230, 2011.

[17] Daniel Russo and Benjamin Van Roy. Eluder dimension and the sample complexity of optimistic exploration. *Advances in Neural Information Processing Systems*, 26, 2013.

[18] 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.

[19] 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.

[20] Claas A Voelcker, Victor Liao, Animesh Garg, and Amir-massoud Farahmand. Value gradient weighted model-based reinforcement learning. In *International Conference on Learning Representations*, 2021.

[21] 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.