Low-Switching Policy Gradient with Exploration via Online Sensitivity Sampling

Yunfan Li [1] Yiran Wang [1] Yu Cheng [2] Lin Yang [1]

Abstract

Policy optimization methods are powerful algorithms in Reinforcement Learning (RL) for their flexibility to deal with policy parameterization and ability to handle model misspecification. However, these methods usually suffer from slow convergence rates and poor sample complexity. Hence it is important to design provably sample efficient algorithms for policy optimization. Yet, recent advances for this problems have only been successful in tabular and linear setting, whose benign structures cannot be generalized to non-linearly parameterized policies. In this paper, we address this problem by leveraging recent advances in value-based algorithms, including bounded eluder-dimension and online sensitivity sampling, to design a low-switching sample-efficient policy optimization algorithm, *LPO*, with general non-linear function approximation. We show that, our algorithm obtains an ε-optimal policy with only $\widetilde{O}(\frac{\text{poly}(d)}{\varepsilon^3})$ samples, where ε is the suboptimality gap and d is a complexity measure of the function class approximating the policy. This drastically improves previously best-known sample bound for policy optimization algorithms, $\widetilde{O}(\frac{\text{poly}(d)}{\varepsilon^8})$. Moreover, we empirically test our theory with deep neural nets to show the benefits of the theoretical inspiration.

1. Introduction

Reinforcement learning (RL) has achieved great success in many practical areas by adopting policy gradient methods with deep neural networks (Schulman et al., 2015a; 2017; Haarnoja et al., 2018). These policy optimization methods are some of the most classic (Williams, 1992; Konda & Tsitsiklis, 1999) approaches for RL. Although their theoretical convergence properties have been established in (Geist et al., 2019; Abbasi-Yadkori et al.; Agarwal et al., 2020b; Bhandari & Russo, 2019) with assumptions that the state space is already well-explored, it is usually not the case in practice. To resolve this issue, policy-based approaches with active exploration in the environment have been proposed in simple tabular (Shani et al., 2020), linear function approximation (Cai et al., 2020; Agarwal et al., 2020a) and general function approximation (Feng et al., 2021) models.

Among these exploration-based approaches, Agarwal et al. (2020a) and Feng et al. (2021) are specially designed to handle model-misspecification more robustly than existing value-based approaches (Jin et al., 2020; Wang et al., 2020b) by performing policy gradient methods to solve a sequence of optimistic MDPs. However, the robustness of both (Agarwal et al., 2020a) and (Feng et al., 2021) pays a huge price: to obtain an ε-suboptimal policy, Agarwal et al. (2020a) requires $\sim \widetilde{O}(1/\varepsilon^{11})$, and Feng et al. (2021) requires $\sim \widetilde{O}(1/\varepsilon^{8})$ number of samples to obtain an ε-optimal policy. Recently, Zanette et al. (2021) has designed a low switching (i.e. reducing the number of policy changes) policy-based algorithm with linear function approximation, which largely reduces the sample complexity of Agarwal et al. (2020a). However, it is still unknown how to improve sample complexity of policy-based algorithms with good robustness in the non-linear setting.

As for the value-based methods, low-switching techniques (Bai et al., 2019; Gao et al., 2021; Kong et al., 2021) are utilized to reduce the policy changes of the algorithm. Among them, Kong et al. (2021) proposed a novel notion of online sensitivity score, which measures the importance of a data point relative to a given dataset over some *general* function class. By using this sensitivity score, Kong et al. (2021) established an online sub-sampling technique which greatly reduced the average *computing time* of previous work (Wang et al., 2020b). Nevertheless, it is unknown whether such low-switching techniques can be applied to save *samples* in policy-based approaches.

In this paper, we present a low-switching policy-based algorithm **LPO** (***L****ow-Switching* ***P****olicy Gradient and Explo-*

[1] Department of Electical and Computer Engineering, University of California, Los Angeles, Los Angeles, CA, USA [2] Microsoft Research, Redmond, WA, USA. Correspondence to: Yunfan Li <yunfanli@g.ucla.edu>, Lin Yang <linyang@ee.ucla.edu>.

Proceedings of the 40^{th} International Conference on Machine Learning, Honolulu, Hawaii, USA. PMLR 202, 2023. Copyright 2023 by the author(s).

*ration via **O**nline Sensitivity Sampling*), which leverages techniques in policy-based approaches, such as (Feng et al., 2021; Zanette et al., 2021) and value-based approach, such as (Kong et al., 2021) to establish efficient policy gradient on non-linear function class while preserving the low-switching property to save samples and running time. Our algorithm follows an actor-critic framework, where the critic guides the exploration of the policy via exploration bonuses derived from the non-linear function class, and policy-gradient (PG) updates the policy to guarantee robustness and stability. The low-switching technique is applied primarily to derive a slowly updating critic, while preserving the quality of learning. Since one of the major terms in sample complexity originates from the PG policy update, slowly updating critic can drastically save the sample complexity as it requires only a few policy updates. Concretely, our approach only update the policy for $\sim \log T$ times for running T rounds of the algorithm, whereas existing approaches, e.g., (Feng et al., 2021), which also targets on the policy-based exploration with non-linear function approximation, takes at least $\sim T$ policy updates. We also derive new PG approaches aware of the structure of non-linear function class to further save samples in updating the policy.

Our Contribution

- We design a new policy-based exploration algorithm, **LPO**, with non-linear function approximation. The algorithm enjoys the same stability guarantee in terms of model-misspecification as presented in existing approaches (Feng et al., 2021). This algorithm leverages efficient value-based techniques (online sensitivity sampling) to slowly update its policy and thus enjoys a sample complexity of $\widetilde{O}(\text{poly}(d)/\varepsilon^3)$, whereas existing approach takes at least $\widetilde{O}(\text{poly}(d)/\varepsilon^8)$ samples to obtain an ε-optimal policy, where d is related to the eluder-dimension, measuring the complexity of the function class.
- While enjoying a theoretical guarantee at special cases where the function class has a bounded complexity, the algorithm itself can be implemented using neural networks. We further empirically tested the theoretical inspiration of online sensitivity sampling with existing deep RL frameworks. The experimental results demonstrated the efficacy of our approaches.

Related Work With regards to exploration methods in RL, there are many provable results in the tabular case (Kearns & Singh, 2002; Brafman & Tennenholtz, 2002; Kearns, 1989; Jin et al., 2018) and linear (Yang & Wang, 2019; 2020; Jin et al., 2020) settings with value or model-based methods. Recent papers (Shani et al., 2020; Cai et al., 2020; Agarwal et al., 2020a) have developed policy-based methods also in tabular and linear settings and Zanette et al. (2021) greatly reduces the sample complexity of Agarwal et al. (2020a) mainly by using the doubling trick for determinant of empircial cumulative covariance. However, relative few provable results are achieved in non-linear setting.

For general function approximation, complexity measures are essential for non-linear function class, and Russo & Van Roy (2013) proposed the concept of eluder dimension. Recent papers have extended it to more general framework (e.g. Bellman Eluder dimension (Jin et al., 2021), Decision-Estimation Coefficient (Foster et al., 2021), Admissible Bellman Characterization (Chen et al., 2022)). However, the use of eluder dimension allows computational tractable optimization methods. Based on the eluder dimension, the value-based technique of Wang et al. (2020b) describes a UCB-VI style algorithm that can explore the environment driven by a well-designed width function and Kong et al. (2021) devises an online sub-sampling method which largely reduces the average computation time of Wang et al. (2020b).

For policy-based method in the general setting, Feng et al. (2021) proposes a model-free algorithm with abundant exploration to environment using the indicator of width function. Moreover, it has better robustness to model misspecification compared to (misspecified) Linear MDP (Jin et al., 2020). However, Feng et al. (2021) suffers from huge sample complexity. In this paper, instead of directly finding a similar notion in the non-linear setting just like determinant in linear setting (Zanette et al., 2021), we adopt an online sensitivity-sampling method to quantify the sensitivity of new-coming data obtained from the environment. Moreover, the importance of designing a sophisticated and efficient reward bonus is mentioned in (Zanette et al., 2021) and we significantly generalize this approach to the non-linear setting by combining the width function and its indicator and our reward bonuses save samples and computing time compared to (Feng et al., 2021).

Notations. We use $[n]$ to represent index set $\{1, \cdots n\}$. For $x \in \mathbb{R}$, $\lfloor x \rfloor$ represents the largest integer not exceeding x and $\lceil x \rceil$ represents the smallest integer exceeding x. Given $a, b \in \mathbb{R}^d$, we denote by $a^\top b$ the inner product between a and b and $||a||_2$ the Euclidean norm of a. Given a matrix A, we use $||A||_2$ for the spectral norm of A, and for a positive definite matrix Σ and a vector x, we define $||x||_\Sigma = \sqrt{x^\top \Sigma x}$. We abbreviate Kullback-Leibler divergence to **KL** and use O to lead orders in asymptotic upper bounds and $\widetilde{O}$ to hide the polylog factors. For a finite set $\mathcal{A}$, we denote the cardinality of $\mathcal{A}$ by $|\mathcal{A}|$, all distributions over $\mathcal{A}$ by $\Delta(\mathcal{A})$, and especially the uniform distribution over $\mathcal{A}$ by Unif$(\mathcal{A})$.

For a function $f : \mathcal{S} \times \mathcal{A} \to \mathbb{R}$, define

$$\|f\|_\infty = \max_{(s,a)\in\mathcal{S}\times\mathcal{A}} |f(s,a)|.$$

Similarly, for a function $v : \mathcal{S} \to \mathbb{R}$, define

$$\|v\|_\infty = \max_{s\in\mathcal{S}} |v(s)|.$$

For a set of state-action pairs $\mathcal{Z} \subseteq \mathcal{S} \times \mathcal{A}$, for a function $f : \mathcal{S} \times \mathcal{A} \to \mathbb{R}$, we define the $\mathcal{Z}$-norm of f as

$$\|f\|_{\mathcal{Z}} = \left(\sum_{(s,a)\in\mathcal{Z}} (f(s,a))^2\right)^{1/2}.$$

2. Preliminaries

Markov Decision Process In this paper, we consider discounted Markov decision process (MDP) environment, which can be specified by a tuple, $\mathcal{M} = (\mathcal{S}, \mathcal{A}, P, r, \gamma)$, where $\mathcal{S}$ is a possibly infinite state space, $\mathcal{A}$ is a finite action space and we denote $A = |\mathcal{A}|$, $P : \mathcal{S} \times \mathcal{A} \to \Delta(\mathcal{S})$ specifies a transition kernel and is unknown to the learner, $r : \mathcal{S} \times \mathcal{A} \to [0, 1]$ is a reward function, and $\gamma \in (0, 1)$ is a discount factor that discounts the reward received in a future time step.

Suppose an RL agent chooses an action $a \in \mathcal{A}$ at the current state s, the environment brings the agent to a new state s' with the unknown probability $P(s' \mid s, a)$ and the agent receives an instant reward $r(s, a)$. The goal for a leaner is to find a policy[1] $\pi : \mathcal{S} \to \Delta(\mathcal{A})$ such that the expected long-term rewards are maximized. In particular, the quality of a policy can be measured by the the Q-value function $Q^\pi : \mathcal{S} \times \mathcal{A} \to \mathbb{R}$ is defined as:

$$Q^\pi(s,a) := \mathbb{E}^\pi\left[\sum_{t=0}^{\infty} \gamma^t r(s_t, a_t) \mid s_0 = s, a_0 = a\right],$$

where the expectation is taken over the trajectory following π – this measures the expected discounted total returns of playing action a at state s and then playing policy π (for an indefinite amount of time). And after taking expectation over the action space, we get the value function: $V^\pi(s) := \mathbb{E}_{a\sim\pi(\cdot|s)}[Q^\pi(s,a)]$, which measures the total expected discounted returns of playing policy π starting from state s. From V^π and Q^π, we can further define the advantage function of π as $A^\pi(s,a) = Q^\pi(s,a) - V^\pi(s)$, which measures whether the action a can be further improved. Moreover, if a policy π^* is optimal, then the Bellman equation (Puterman, 2014) states that $A^{\pi^*}(s,a) \le 0$ for all s, a and $\mathbb{E}_{a\sim\pi^*(\cdot|s)}[A^{\pi^*}(s,a)] = 0$. In practice, we may also restrict the policy space being considered as Π (which may be parameterized by a function class).

We also define the discounted state-action distribution $d^\pi_{\tilde{s}}(s,a)$ induced by π as:

$$d^\pi_{\tilde{s}}(s,a) = (1-\gamma)\sum_{t=0}^{\infty} \gamma^t \Pr{}^\pi(s_t = s, a_t = a \mid s_0 = \tilde{s}),$$

where $\Pr^\pi(s_t = s, a_t = a \mid s_0 = \tilde{s})$ is the probability of reaching (s,a) at the t_{th} step starting from $\tilde{s}$ following π. Similarly, the definition of $d^\pi_{\tilde{s},\tilde{a}}(s,a)$ can be easily derived as the distribution of state-actions if the agent starts from state $\tilde{s}$ and selects an action $\tilde{a}$. For any initial state-actions distribution $\nu \in \Delta(\mathcal{S} \times \mathcal{A})$, we denote by $d^\pi_\nu(s,a) := \mathbb{E}_{(\tilde{s},\tilde{a})\sim\nu}\left[d^\pi_{(\tilde{s},\tilde{a})}(s,a)\right]$ and $d^\pi_\nu(s) := \sum_a d^\pi_\nu(s,a)$. Given an initial state distribution $\rho \in \Delta(\mathcal{S})$, we define $V^\pi_\rho := \mathbb{E}_{s_0\sim\rho}[V^\pi(s_0)]$. With these notations, the reinforcement learning (RL) problem with respect to the policy class Π is reduced to solving the following optimization problem.

$$\underset{\pi\in\Pi}{\text{maximize}} V^\pi_{\rho_0},$$

for some initial distribution ρ_0. We further, without loss of generality [2], assume ρ_0 is a singleton on some state s_0.

Policy Space and Width Function We now formally define the policy parameterization class, which is compatible with a neural network implementation. For a set of functions $\mathcal{F} \subseteq \{f : \mathcal{S} \times \mathcal{A} \to \mathbb{R}\}$, we consider a policy space as $\Pi_{\mathcal{F}} := \{\pi_f, f \in \mathcal{F}\}$ by applying the softmax transform to functions in $\mathcal{F}$, i.e., for any $f \in \mathcal{F}$,

$$\pi_f(a|s) := \frac{\exp(f(s,a))}{\sum_{a'\in\mathcal{A}} \exp(f(s,a'))}.$$

Given $\mathcal{F}$, we define its function difference class $\Delta\mathcal{F} := \{\Delta f \mid \Delta f = f - f', f, f' \in \mathcal{F}\}$ and width function on $\Delta\mathcal{F}$ for a state-action pair (s,a) as

$$w(\Delta\mathcal{F}, s, a) = \sup_{\Delta f\in\Delta\mathcal{F}} |\Delta f(s,a)|.$$

As we will show shortly, this width function will be used to design exploration bonuses for our algorithm.

3. Algorithms

[1] We here only consider stationary policies as one can always find a stationary optimal-policy in a discounted MDP (Puterman, 2014).

[2] Otherwise, we can modify the MDP and add a dummy state s_0 with ρ_0 as its state transition for all actions played at s_0.

Algorithm 1 LPO

```
Input: Function class $\mathcal{F}$
Hyperparameters: $N, \delta, \beta$
For all $s \in S$, initialize $\pi^0(\cdot|s) = \text{Unif}(\mathcal{A})$, $\widehat{\mathcal{Z}}^1 = \emptyset$
for $n = 1, 2, \cdots, N$ do
    Update policy cover $\pi^n_{cov} = \pi^{0:n-1}$
    $\widehat{\mathcal{Z}}^n \leftarrow$ S-Sampling$(\mathcal{F}, \widehat{\mathcal{Z}}^{n-1}, (s_{n-1}, a_{n-1}), \delta)$
    if $\widehat{\mathcal{Z}}^n \neq \widehat{\mathcal{Z}}^{\underline{n}}$ or $n = 1$ then
        Update the known set and bonus function
        $\mathcal{K}^n = \{(s,a) \mid \omega(\widehat{\mathcal{F}}^n, s, a) < \beta\}$
        $b^n(s,a) = \frac{3}{1-\gamma} \cdot \mathbf{1}\{\omega(\widehat{\mathcal{F}}^n, s, a) \geq \beta\} + \frac{2}{\beta} \cdot \omega(\widehat{\mathcal{F}}^n, s, a) \cdot \mathbf{1}\{\omega(\widehat{\mathcal{F}}^n, s, a) < \beta\}$
        Set $\underline{n} \leftarrow n$
        $\pi^n \leftarrow$ Policy Update$(\pi^n_{cov}, b^n, \mathcal{K}^n)$
    else
        $\pi^n \leftarrow \pi^{\underline{n}}, \mathcal{K}^n \leftarrow \mathcal{K}^{\underline{n}}, b^n \leftarrow b^{\underline{n}}$
    end if
    $(s_n, a_n) \leftarrow$ d-sampler$(\pi^{\underline{n}}, \nu)$
end for
Output: Unif$(\pi^0, \pi^1, \cdots, \pi^{N-1})$
```

In this section, we present our algorithm *Low-Switching Policy Gradient and Exploration via Online Sensitivity Sampling* (**LPO**). The algorithm takes a function class $\mathcal{F}$ as an input and interacts with the environment to produce a near-optimal policy. The complete pseudocode is in Algorithm 1. We first give an overview of our algorithm before describing the details of our improvements.

3.1. Overview of our Algorithm

Our algorithm **LPO** (Algorithm 1) has two loops. The outer loop produces a series of well-designed optimistic MDPs by adding a reward bonus and choosing an initial state distribution which are then solved with regression in the inner loop. These optimistic MDPs will encourage the agent to explore unseen part of the environment. In our **LPO**, we construct the initial state distribution by using the uniform mixture of previous well-trained policies (also called policy cover).

Specifically, at the beginning of n-th iteration, we have already collected sample (s_n, a_n) using the last policy $\pi^{\underline{n}}$. Then at iteration n, we use **S-Sampling** (i.e. **Sensitivity-Sampling**) (Algorithm 3) to measure the change that the new sample brings to the dataset. If the current sample can provide sufficiently new information relative to the formal dataset, then with great probability, we choose to store this data and invoke the inner loop to update the policy. Otherwise, we just abandon this data and continue to collect data under $\pi^{\underline{n}}$. Through this process, a policy cover $\pi^n_{cov} = \text{Unif}(\pi^0, \pi^1, \cdots, \pi^{n-1})$ is constructed to provide an initial distribution for the inner routine. To this end, we define the known state-actions $\mathcal{K}^n$, which can be visited with high probability under π^n_{cov}. Using $\mathcal{K}^n$ and a reward bonus b^n, we create an optimistic MDP to encourage the agent to explore outside $\mathcal{K}^n$ as well as to refine its estimates inside $\mathcal{K}^n$.

In the inner routine, the algorithm **Policy Update** (Algorithm 4) completes the task to find an approximately optimal policy π^n in the optimistic MDP through general function approximation. This policy π^n would produce new samples which will be measured in the next iteration. Under the procedure of our algorithm, the policy cover will gain sufficient coverage over the state-action space and the bonus will shrink. Therefore, the near-optimal policies in the optimistic MDPs eventually behave well in the original MDP. Next, we will describe the details of each part of our algorithm.

3.2. Outer Loop

Now we describe the details of three important parts in the outer loop.

Lazy Updates of Optimistic MDPs via Online Sensitivity-Sampling The lazy or infrequent updates of the optimistic MDPs in **LPO** play a crucial role of improving sample complexity, which reduce the number of **Policy Update** invocations from $O(N)$ to $O(\text{poly}(\log N))$. For the linear case, (Zanette et al., 2021) achieves this result by monitoring the determinant of the empirical cumulative covariance matrix. However, in our general setting, we can not count on the linear features anymore. Instead, we introduce our online sensitivity sampling technique, which is also mentioned in (Wang et al., 2020b; Kong et al., 2021).

Now we describe the procedure for constructing the sensitivity dataset $\widehat{\mathcal{Z}}^n$. At the beginning of iteration n, the algorithm receives the current sensitivity dataset $\widehat{\mathcal{Z}}^{n-1}$ and the new data (s_{n-1}, a_{n-1}) from the last iteration. We first calculate the online sensitivity score in (1) to measure the importance of (s_{n-1}, a_{n-1}) relative to $\widehat{\mathcal{Z}}^{n-1}$.

$$\begin{aligned}&\text{sensitivity}_{\mathcal{Z}^n,\mathcal{F}}(z)\\ &= \sup_{f_1, f_2 \in \mathcal{F}} \frac{(f_1(z) - f_2(z))^2}{\min\{||f_1 - f_2||^2_{\mathcal{Z}^n}, 4NW^2\} + 1}\end{aligned} \qquad (1)$$

Then the algorithm adds (s_{n-1}, a_{n-1}) to $\widehat{\mathcal{Z}}^{n-1}$ with probability decided by its online sensitivity score. Intuitively, the more important the new sample is, the more likely it is to be added. At the same time, the algorithm set the number of copies added to the dataset according to the sampling probability, if added. In addition, due to the technical obstacle in theoretical proof, we need to replace the original data z with the data $\widehat{z} \in \mathcal{C}(\mathcal{S} \times \mathcal{A}, 1/16\sqrt{64N^3/\delta})$ in the ε-cover set (defined in Assumption 4.3), which satisfies:

$$\sup_{f\in\mathcal{F}} |f(z) - f(\hat{z})| \leq 1/16\sqrt{64N^3/\delta} \tag{2}$$

Furthermore, the chance that the new sample is sensitive gets smaller when the dataset gets enough samples, which means that the policy will not change frequently in the later period of training. As will be demonstrated later, the number of switches (i.e. the number of policy changes in the running of the outer loop) achieve the logarithmic result. To this end, the size of sensitivity dataset is bounded and provides good approximation to the original dataset, which serves as a benign property for theoretical proof.

Known and Unknown state-actions According to the value of width function (defined in Section 2) under the sensitivity dataset, the state-action space $\mathcal{S} \times \mathcal{A}$ is divided into two sets, namely the known set $\mathcal{K}^n$ in (3) and its complement the unknown set.

$$\mathcal{K}^n = \{(s,a) \in \mathcal{S} \times \mathcal{A} |\, \omega(\widehat{\mathcal{F}}^n, s, a) < \beta\} \tag{3}$$

where

$$\widehat{\mathcal{F}}^n = \{\Delta f \in \Delta\mathcal{F} |\, ||\Delta f||^2_{\widehat{\mathcal{Z}}^n} \leq \epsilon\}$$

In fact, the width function $\omega(\widehat{\mathcal{F}}^n, s, a)$ serves as a prediction error for a new state-action pair (s,a) where the training data is $\widehat{\mathcal{Z}}^n$, which is the general form of $||\phi(s,a)||_{(\widehat{\Sigma}^n)^{-1}}$ in the linear case. Therefore, the known set $\mathcal{K}^n$ represents the state-action pairs easily visited under the policy cover π^n_{cov}. If all the actions for one state are in the $\mathcal{K}^n$, we say this state is known.

$$\mathcal{K}^n = \{s \in \mathcal{S} |\, \forall a \in \mathcal{A}, \omega(\widehat{\mathcal{F}}^n, s, a) < \beta\} \tag{4}$$

Bonus Function In a more refined form, **LPO** devises bonus function in both the known and unknown sets.

$$\begin{aligned} b^n(s,a) &= 2b^n_w(s,a) + b^n_1(s,a), \text{ where} \\ b^n_w(s,a) &= \frac{1}{\beta}\,\omega(\widehat{\mathcal{F}}^n, s, a)\mathbf{1}\{s \in \mathcal{K}^n\}, \text{ and} \\ b^n_1(s,a) &= \frac{3}{1-\gamma}\mathbf{1}\{s \notin \mathcal{K}^n\} \end{aligned} \tag{5}$$

On the unknown state-actions, the bonus is a constant $\frac{3}{1-\gamma}$, which is the largest value of the original reward over a trajectory. This will force the agent out of the known set and explore the unknown parts of the **MDP**. On the known state-actions, the uncertainty is measured by the width function.

Notice that our algorithm explore the environment in a much more sophisticated and efficient way than (Feng et al., 2021) does. Our algorithm **LPO** not only explores the unknown part using the indicator $b^n_1(s,a)$, but also takes the uncertainty information $b^n_w(s,a)$ in the known set into account. Consequently, **LPO** still explores the state-action pair in the known set until it is sufficiently understood. Moreover, since the size of sensitivity dataset is bounded by $O(d \log N)$, where d is the eluder dimension, the average computing time of our bonus function is largely reduced.

3.3. Inner Loop

In the inner routine, the **Policy Update** initializes the policy to be a uniform distribution and encourages the policy to explore the unknown state-actions. Next, we adopt the online learning algorithm to update the policy, which is an actor-critic pattern. This update rule is equivalent to Natural Policy Gradient (NPG) algorithm for log-linear policies (Kakade, 2001; Agarwal et al., 2020b), where we fit the critic with Monte Carlo samples and update the actor using exponential weights. As mentioned in (Agarwal et al., 2020b), Monte Carlo sampling has an advantage of assuring better robustness to model misspecification, but produces huge source of sample complexity.

Sample efficient policy evaluation oracle via importance sampling. In the **Policy Update** routine, the policy obtained in each iteration needs to be evaluated. In a most direct way, the agent needs to interact with the environment by Monte Carlo sampling and estimate the Q-function for each policy, and this leads to the waste of samples. In order to improve the sample complexity of **Policy Update** while keeping the robustness property, we design a sample efficient policy evaluation oracle by applying trajectory-level importance sampling on past Monte Carlo return estimates (Precup, 2000). To be specific, at iteration $\underline{k}$ in the inner loop, the agent will collect data by routine **Behaviour Policy Sampling** (Algorithm 5), and the dataset obtained in this iteration will be reused for the next κ turns. At iteration k ($k \leq \underline{k} + \kappa$), the **Policy Evaluation Oracle** (Algorithm 6) can estimate the Monte Carlo return for the current policy π_k by reweighting the samples with importance sampling. With the reweighted random return, the oracle fits the critic via least square regression and outputs an estimated $\widehat{Q}_k$ for policy π_k. To this end, we update the policy by following the NPG rule. Although the technique above can largely reduce the number of interactions with environment (from K to $\lceil\frac{K}{\kappa}\rceil$), the selection of κ greatly influences the variance of importance sampling estimator, which ultimately challenges the robustness property. In fact, We need to set $\kappa = \widetilde{O}(\sqrt{K})$ in order to keep a stable variance of the estimator (see Lemma E.4 and Remark E.5 for details).

Remark 3.1. If we have obtained a full coverage dataset over state-action space (e.g. using bonus-driven reward to collect data in (Jin et al., 2020; Wang et al., 2020a)), the policy evaluation oracle can evaluate all the policies by using the above mentioned dataset and only needs $\widetilde{O}(\frac{1}{\varepsilon^2})$ samples.

However, this oracle depends on the (ζ-approximate) linear MDP, which is a stronger assumption than that in our setting.

Pessimistic critic to produce one-sided errors From the line 9 in Algorithm 6, we find that the critic fitting is actually the Monte Carlo return minus the initial bonus. Therefore, an intuitive form for the critic estimate is $\widehat{Q}^k_{b^n}(s,a) = f_k(s,a) + b^n(s,a)$. However, in line 10 of Algorithm 6, we only partially make up for the subtracted term and define the critic estimate as $\widehat{Q}^k_{b^n}(s,a) = f_k(s,a) + \frac{1}{2}b^n(s,a)$. This introduces a negative bias in the estimate. However, in the later proof we can see that $\widehat{Q}^k_{b^n}(s,a)$ is still optimistic relative to $Q^k(s,a)$. This one-sided error property plays an essential role of bounding the gap between $Q^{k,*}_{b^n}$ and $\widehat{Q}^k_{b^n}(s,a)$, which ultimately improves the sample complexity.

4. Theoretical Guarantee

In this section, we will provide our main result of **LPO** under some assumptions. The main theorem (Theorem 4.8) is presented in this section and the complete proof is in the appendix.

First of all, the sample complexity of algorithms with function approximation depends on the complexity of the function class. To measure this complexity, we adopt the notion of eluder dimension which is first mentioned in (Russo & Van Roy, 2013).

Definition 4.1. (Eluder dimension). $\varepsilon \geq 0$ and $\mathcal{Z} = \{(s_i, a_i)\}_{i=1}^n \subseteq \mathcal{S} \times \mathcal{A}$ be a sequence of state-action pairs.

- A state-action pair $(s,a) \in \mathcal{S} \times \mathcal{A}$ is ε-dependent on $\mathcal{Z}$ with respect to $\mathcal{F}$ if any $f, f' \in \mathcal{F}$ satisfying $\|f - f'\|_{\mathcal{Z}} \leq \varepsilon$ also satisfies $|f(s,a) - f'(s,a)| \leq \varepsilon$.
- An (s,a) is ε-independent of $\mathcal{Z}$ with respect to $\mathcal{F}$ if (s,a) is not ε-dependent on $\mathcal{Z}$.
- The eluder dimension $\dim_E(\mathcal{F}, \varepsilon)$ of a function class $\mathcal{F}$ is the length of the longest sequence of elements in $\mathcal{S} \times \mathcal{A}$ such that, for some $\varepsilon' \geq \varepsilon$, every element is ε'-independent of its predecessors.

Remark 4.2. In fact, (Russo & Van Roy, 2013) has shown several bounds for eluder dimension in some special cases. For example, when $\mathcal{F}$ is the class of linear functions, i.e. $f_\theta(s,a) = \theta^\top \phi(s,a)$ with a given feature $\phi : \mathcal{S} \times \mathcal{A} \to \mathbb{R}^d$, or the class of generalized linear functions of the form $f_\theta(s,a) = g(\theta^\top \phi(s,a))$ where g is a differentiable and strictly increasing function, $\dim_E(\mathcal{F}, \varepsilon) = O(d \log(1/\varepsilon))$. Recently, more general complexity measures for non-linear function class have been proposed in (Jin et al., 2021; Foster et al., 2021; Chen et al., 2022). However, the adoption of eluder dimension allows us to use computation-friendly optimization methods (e.g. dynamic programming, policy gradient) whereas theirs do not directly imply computationally tractable and implementable approaches. If only statistical complexities are considered, we believe our techniques could be extended to their complexity measures.

Next, we assume that the function class $\mathcal{F}$ and the state-actions $\mathcal{S} \times \mathcal{A}$ have bounded covering numbers.

Assumption 4.3. (ε-cover). For any $\varepsilon > 0$, the following holds:

1. there exists an ε-cover $\mathcal{C}(\mathcal{F}, \varepsilon) \subseteq \mathcal{F}$ with size $|\mathcal{C}(\mathcal{F}, \varepsilon)| \leq \mathcal{N}(\mathcal{F}, \varepsilon)$, such that for any $f \in \mathcal{F}$, there exists $f' \in \mathcal{C}(\mathcal{F}, \varepsilon)$ with $\|f - f'\|_\infty \leq \varepsilon$;

2. there exists an ε-cover $\mathcal{C}(\mathcal{S} \times \mathcal{A}, \varepsilon)$ with size $|\mathcal{C}(\mathcal{S} \times \mathcal{A}, \varepsilon)| \leq \mathcal{N}(\mathcal{S} \times \mathcal{A}, \varepsilon)$, such that for any $(s,a) \in \mathcal{S} \times \mathcal{A}$, there exists $(s', a') \in \mathcal{C}(\mathcal{S} \times \mathcal{A}, \varepsilon)$ with $\max_{f \in \mathcal{F}} |f(s,a) - f(s', a')| \leq \varepsilon$.

Remark 4.4. Assumption 4.3 is rather standard. Since our algorithm complexity depends only logarithmically on $\mathcal{N}(\mathcal{F}, \cdot)$ and $\mathcal{N}(\mathcal{S} \times \mathcal{A}, \cdot)$, it is even acceptable to have exponential size of these covering numbers.

Next, we enforce a bound on the values of the functions class:

Assumption 4.5. (Regularity). We assume that $\sup_{f \in \mathcal{F}} \|f\|_\infty \leq W$.

Assumption 4.5 is mild as nearly all the function classes used in practice have bounded magnitude in the input domain of interests. In general, one shall not expect an arbitrary function class could provide good guarantees in approximating a policy. In this section, we apply the following completeness condition to characterize whether the function class $\mathcal{F}$ fits to solve RL problems.

Assumption 4.6. ($\mathcal{F}$-closedness).

$$\mathcal{T}^\pi f(s,a) := \mathbb{E}^\pi \left[r(s,a) + \gamma f(s', a') \mid s, a \right].$$

We assume that for all $\pi \in \{\mathcal{S} \to \Delta(\mathcal{A})\}$ and $g : \mathcal{S} \times \mathcal{A} \to [0, W]$, we have $\mathcal{T}^\pi g \in \mathcal{F}$.

Remark 4.7. Assumption 4.6 is a closedness assumption on $\mathcal{F}$, which enhances its representability to deal with critic fitting. For some special cases, like linear f in the linear MDP (Yang & Wang, 2019; Jin et al., 2020), this assumption always holds. If this assumption does not hold, which means Q-function may not realizable in our function class $\mathcal{F}$, then there exists a ϵ_{bias} when we approximate the true Q-function. This model misspecified setting will be discussed in Assumption B.5.

With the above assumptions, we have the following sample complexity result for **LPO**.

Theorem 4.8. *(Main Results: Sample Complexity of LPO). Under Assumption 4.3, 4.5, and 4.6, for any comparator $\widetilde{\pi}$, a fixed failure probability δ, eluder dimension $d = dim(\mathcal{F}, 1/N)$, a suboptimality gap ε and appropriate input hyperparameters: $N \geq \widetilde{O}\left(\frac{d^2}{(1-\gamma)^4\varepsilon^2}\right), K = \widetilde{O}\left(\frac{\ln|\mathcal{A}|W^2}{(1-\gamma)^2\varepsilon^2}\right), M \geq \widetilde{O}\left(\frac{d^2}{(1-\gamma)^4\varepsilon^2}\right), \eta = \widetilde{O}\left(\frac{\sqrt{\ln|\mathcal{A}|}}{\sqrt{KW}}\right), \kappa = \widetilde{O}\left(\frac{1-\gamma}{\eta W}\right)$, our algorithm returns a policy π^{LPO}, satisfying*

$$\left(V^{\widetilde{\pi}} - V^{\pi^{LPO}}\right)(s_0) \leq \varepsilon.$$

with probability at least $1 - \delta$ after taking at most $\widetilde{O}\left(\frac{d^3}{(1-\gamma)^8\varepsilon^3}\right)$ samples.

Remark 4.9. The complete proof of our main theorem is presented in Theorem B.12. For the model misspecified case, which means Assumption 4.6 does not hold, there exists a ϵ_{bias} in our regret bound (see details in Lemma B.11)

5. Practical Implementation and Experiments

In this section, we introduce a practical approach to implementing our proposed theoretical algorithm and show our experiment results.

5.1. Practical Implementation of LPO

The width function in our theoretical framework enables our policy gradient algorithm to explore efficiently. The value of the width function should be large over novel state-action pairs and shrink over not novel. Intuitively, the width function over all state-action pairs should be its maximum at the beginning of learning and decrease along the way. To this end, we leverage the random network distillation technique proposed by (Burda et al., 2018). We randomly initialize two neural networks f and f' that maps from $\mathcal{A} \times \mathcal{S}$ to $\mathbb{R}^d$ with the same architecture (but different parameters). And our bonus $b(s, a)$ is defined as $b(s, a) = \|f(s, a) - f'(s, a)\|^2$. During training, we fix f' and train f to minimize $\|f(s, a) - f'(s, a)\|^2$ with gradient descent over state-action pairs that the agent has seen during the sampling procedure, i.e. distilling the randomly initialized network f' into a trained one f. Over time, this residual-dependent bonus will decrease over state-action pairs that agents commonly visit.

With bonus calculated in the way described above, at each Monte Carlo session, we can calculate an intrinsic advantage estimation $\hat{A}^{(int)}$, which will affect our policy gradient along with the extrinsic advantage estimation $\hat{A}^{(ext)}$. The gradient of policy parametrized by ϕ is given by:

$$\hat{A}_t^{(total)} = \alpha \hat{A}_t^{(ext)} + \beta \hat{A}_t^{(int)} \tag{6}$$

Algorithm 2 LPO (Practical Implementation)

```
Input: Width function $(f, f')$, Policy $\pi_{\phi_0}$, Value networks $(V^{(ext)}, V^{(int)})$
Hyperparameters: $N, K, \gamma, \lambda, \alpha, \beta$
For all $s \in S$, initialize $\pi^0(\cdot|s) = \text{Unif}(\mathcal{A})$
for $k = 0, 1, 2, 3, ..., K$ do
    $T \leftarrow \lceil (1 + \frac{1}{K})^k N \rceil$
    Run policy $\pi_\phi$ for $T$ steps $\rightarrow D_k$
    Calculate $A^{(ext)}$, $A^{(int)}$ using 7 using $\lambda$
    Calculate $A^{(total)}$ using 6 using $\alpha$, $\beta$
    Optimize $\pi_\phi$, $(V^{(ext)}, V^{(int)})$ using PPO with $A^{(total)}$ as advantage estimation
    Optimize $f$ to fit $f'$ w.r.t. $D_k$
end for
Output: Unif($\pi^0, \pi^1, \cdots, \pi^{N-1}$)
```

where α and β are hyperparameters that control how much we want to emphasize the importance of exploration, in our experiment, we use $\alpha = 2$ and $\beta = 1$. And the $\hat{A}_{ext}$, $\hat{A}_{int}$ are calculated with generalized advantage estimation (Schulman et al., 2015b), and the estimation of advantages over a time period of T is given by:

$$\begin{aligned}\hat{A}_t^{(ext)} &= \sum_{i=t}^{T} (\gamma^{(ext)}\lambda)^{i-t}[r_i + \gamma^{(ext)}V(s_{i+1}) - V(s_i)] \\ \hat{A}_t^{(int)} &= \sum_{i=t}^{T} (\gamma^{(int)}\lambda)^{i-t}[b_i + \gamma^{(int)}V^{(int)}(s_{i+1}) \\ &\quad - V^{(int)}(s_i)]\end{aligned} \tag{7}$$

where V and $V^{(int)}$ are two value estimator parametrized that predicts extrinsic and intrinsic value separately. We train value functions to fit the Monte Carlo estimation of values of the current policy.

In our theoretical framework, the sensitivity is computed using (1) and only designed to achieve boundness on the final theoretical guarantee, but not for practical implementation. We overcome this issue by introducing a coarse, yet effective approximation of sensitivity sampling: gradually increasing the samples we collect for Monte Carlo estimation. For each sampling procedure at time t, we collect $max\{N, (1 + \frac{1}{T})^t N\}$ samples (N is the number of samples we collect at the first sampling procedure). This simple mechanism makes the agent collect more and more samples for each training loop, which allow the agent to explore more freely at the early stage of learning, and forces the agent to explore more carefully at the late stage, as using more data for each training loop will shrink the value of width function in a more stable way. The algorithm is shown in Algorithm 2.

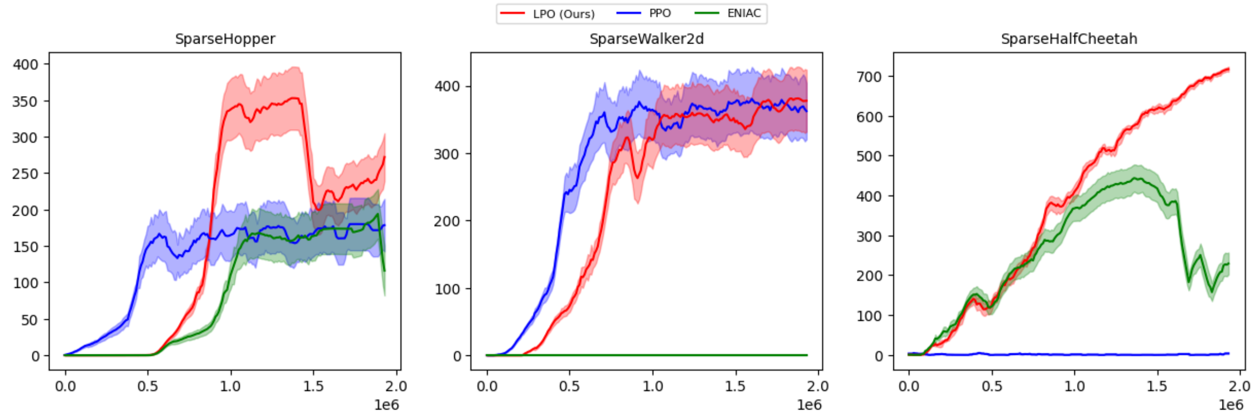

Figure 1. Performance of **PPO**-Based algorithms on sparse-reward MuJoCo localmotion environments. Lines are evaluation results averaged over 5 random seeds every 10k steps, the shaded area represents the standard deviation.

5.2. Experiments

To further illustrate the effectiveness of our width function and our proposed sensitivity sampling, we compare (Schulman et al., 2017; Feng et al., 2021) with our proposed **LPO** in sparse reward MuJoCo environments (Todorov et al., 2012). We re-implement (Feng et al., 2021) with the random network distillation method, as we find the original implementation of width function was not numerically stable. More details are in the discussion section.

The sparse MuJoCo environment is different from the usual localmotion task, where rewards are dense and given according to the speed of the robots, in sparse environments, reward $(+1)$ is only given whenever the robot exceeds a certain speed, and no reward is given otherwise. Such environments are more difficult than their dense reward counterpart in the sense that the agent needs to actively explore the environment strategically and find a good policy. **PPO** manages to find rewards in SparseHopper and SparseWalker2d, but fails in SparseHalfCheetah. Although **ENIAC** (Feng et al., 2021) also uses intrinsic rewards for strategic exploration, it still fails in the SparseWalker2d environment. This might be due to the intrinsic reward of **ENIAC** switching too fast, thus the exploration is not persistent enough for the agent to find the reward. In contrast, our method succeeds in all three environments, the result is shown Figure 1.

5.3. Limitation of Previous Implementations

Note that we do not compare our method directly with implementations in (Agarwal et al., 2020a; Feng et al., 2021), as we discovered some limitations presented in their implementations. We will discuss this in more details in Section F about their limitations in terms of batch normalization and mis-implementations of core components in existing approaches. We illustrate the issue by directly running their published code. As shown in Figure 2, our approaches and the other baseline approaches (Raffin et al., 2021) successfully solve the problem in a few epochs, while their implementations fail to achieve similar performance.

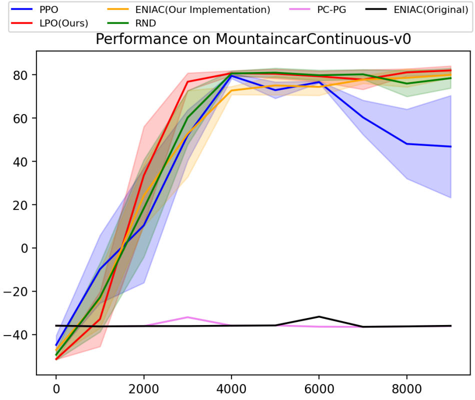

Figure 2. Performance comparison between the original implementations of **ENIAC, PC-PG** and our implementation of **ENIAC, LPO**. Lines are evaluation results averaged over 5 random seeds every 10k steps, the shaded area represents the standard deviation.

6. Conclusion

In this paper, we establish a low-switching sample-efficient policy optimization algorithm with general function approximation using online sensitivity sampling and data reuse techniques. Our algorithm largely improves the sample complexity in (Feng et al., 2021), while still keeping its robustness to model misspecification. Our numerical experiments also empirically prove the efficiency of the low-switching technique in policy gradient algorithms.

7. Acknowledgements

This work was supported in part by DARPA under agreement HR00112190130, NSF grant 2221871, and an Amazon Research Grant. The authors would also like to thank Dingwen Kong for useful discussions.

References

Abbasi-Yadkori, Y., Bartle, P. L., Bhatia, K., Lazić, N., Szepesvári, C., and Weisz, G. P: Regret bounds for policy iteration using expert prediction.

Agarwal, A., Henaff, M., Kakade, S., and Sun, W. Pc-pg: Policy cover directed exploration for provable policy gradient learning. *Advances in neural information processing systems*, 33:13399–13412, 2020a.

Agarwal, A., Kakade, S. M., Lee, J. D., and Mahajan, G. Optimality and approximation with policy gradient methods in markov decision processes. In *Conference on Learning Theory*, pp. 64–66. PMLR, 2020b.

Bai, Y., Xie, T., Jiang, N., and Wang, Y.-X. Provably efficient q-learning with low switching cost. *Advances in Neural Information Processing Systems*, 32, 2019.

Bellemare, M. G., Naddaf, Y., Veness, J., and Bowling, M. The arcade learning environment: An evaluation platform for general agents. *Journal of Artificial Intelligence Research*, 47:253–279, jun 2013.

Beygelzimer, A., Langford, J., Li, L., Reyzin, L., and Schapire, R. Contextual bandit algorithms with supervised learning guarantees. In *Proceedings of the Fourteenth International Conference on Artificial Intelligence and Statistics*, pp. 19–26. JMLR Workshop and Conference Proceedings, 2011.

Bhandari, J. and Russo, D. Global optimality guarantees for policy gradient methods. *arXiv preprint arXiv:1906.01786*, 2019.

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

Burda, Y., Edwards, H., Storkey, A., and Klimov, O. Exploration by random network distillation. *arXiv preprint arXiv:1810.12894*, 2018.

Cai, Q., Yang, Z., Jin, C., and Wang, Z. Provably efficient exploration in policy optimization. In *International Conference on Machine Learning*, pp. 1283–1294. PMLR, 2020.

Chen, Z., Li, C. J., Yuan, A., Gu, Q., and Jordan, M. I. A general framework for sample-efficient function approximation in reinforcement learning. *arXiv preprint arXiv:2209.15634*, 2022.

Feng, F., Yin, W., Agarwal, A., and Yang, L. Provably correct optimization and exploration with non-linear policies. In *International Conference on Machine Learning*, pp. 3263–3273. PMLR, 2021.

Foster, D. J., Kakade, S. M., Qian, J., and Rakhlin, A. The statistical complexity of interactive decision making. *arXiv preprint arXiv:2112.13487*, 2021.

Freedman, D. A. On tail probabilities for martingales. *the Annals of Probability*, pp. 100–118, 1975.

Gao, M., Xie, T., Du, S. S., and Yang, L. F. A provably efficient algorithm for linear markov decision process with low switching cost. *arXiv preprint arXiv:2101.00494*, 2021.

Geist, M., Scherrer, B., and Pietquin, O. A theory of regularized markov decision processes. In *International Conference on Machine Learning*, pp. 2160–2169. PMLR, 2019.

Haarnoja, T., Zhou, A., Abbeel, P., and Levine, S. Soft actor-critic: Off-policy maximum entropy deep reinforcement learning with a stochastic actor. In *International conference on machine learning*, pp. 1861–1870. PMLR, 2018.

Jin, C., Allen-Zhu, Z., Bubeck, S., and Jordan, M. I. Is q-learning provably efficient? *Advances in neural information processing systems*, 31, 2018.

Jin, C., Yang, Z., Wang, Z., and Jordan, M. I. Provably efficient reinforcement learning with linear function approximation. In *Conference on Learning Theory*, pp. 2137–2143. PMLR, 2020.

Jin, C., Liu, Q., and Miryoosefi, S. Bellman eluder dimension: New rich classes of rl problems, and sample-efficient algorithms. *Advances in neural information processing systems*, 34:13406–13418, 2021.

Kakade, S. M. A natural policy gradient. *Advances in neural information processing systems*, 14, 2001.

Kearns, M. and Singh, S. Near-optimal reinforcement learning in polynomial time. *Machine learning*, 49(2):209–232, 2002.

Kearns, M. J. *Computational Complexity of Machine Learning*. PhD thesis, Department of Computer Science, Harvard University, 1989.

Konda, V. and Tsitsiklis, J. Actor-critic algorithms. *Advances in neural information processing systems*, 12, 1999.

Kong, D., Salakhutdinov, R., Wang, R., and Yang, L. F. Online sub-sampling for reinforcement learning with general function approximation. *arXiv preprint arXiv:2106.07203*, 2021.

Langley, P. Crafting papers on machine learning. In Langley, P. (ed.), *Proceedings of the 17th International Conference on Machine Learning (ICML 2000)*, pp. 1207–1216, Stanford, CA, 2000. Morgan Kaufmann.

Li, Q., Zhai, Y., Ma, Y., and Levine, S. Understanding the complexity gains of single-task rl with a curriculum. *arXiv preprint arXiv:2212.12809*, 2022.

Precup, D. Eligibility traces for off-policy policy evaluation. *Computer Science Department Faculty Publication Series*, pp. 80, 2000.

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

Raffin, A., Hill, A., Gleave, A., Kanervisto, A., Ernestus, M., and Dormann, N. Stable-baselines3: Reliable reinforcement learning implementations. *Journal of Machine Learning Research*, 22(268):1–8, 2021. URL http://jmlr.org/papers/v22/20-1364.html.

Russo, D. and Van Roy, B. Eluder dimension and the sample complexity of optimistic exploration. *Advances in Neural Information Processing Systems*, 26, 2013.

Schulman, J., Levine, S., Abbeel, P., Jordan, M., and Moritz, P. Trust region policy optimization. In *International conference on machine learning*, pp. 1889–1897. PMLR, 2015a.

Schulman, J., Moritz, P., Levine, S., Jordan, M., and Abbeel, P. High-dimensional continuous control using generalized advantage estimation. *arXiv preprint arXiv:1506.02438*, 2015b.

Schulman, J., Wolski, F., Dhariwal, P., Radford, A., and Klimov, O. Proximal policy optimization algorithms. *arXiv preprint arXiv:1707.06347*, 2017.

Shani, L., Efroni, Y., Rosenberg, A., and Mannor, S. Optimistic policy optimization with bandit feedback. In *International Conference on Machine Learning*, pp. 8604–8613. PMLR, 2020.

Todorov, E., Erez, T., and Tassa, Y. Mujoco: A physics engine for model-based control. In *2012 IEEE/RSJ International Conference on Intelligent Robots and Systems*, pp. 5026–5033, 2012. doi: 10.1109/IROS.2012.6386109.

Wang, R., Du, S. S., Yang, L., and Salakhutdinov, R. R. On reward-free reinforcement learning with linear function approximation. *Advances in neural information processing systems*, 33:17816–17826, 2020a.

Wang, R., Salakhutdinov, R. R., and Yang, L. Reinforcement learning with general value function approximation: Provably efficient approach via bounded eluder dimension. *Advances in Neural Information Processing Systems*, 33:6123–6135, 2020b.

Williams, R. J. Simple statistical gradient-following algorithms for connectionist reinforcement learning. *Machine learning*, 8(3):229–256, 1992.

Yang, L. and Wang, M. Sample-optimal parametric q-learning using linearly additive features. In *International Conference on Machine Learning*, pp. 6995–7004. PMLR, 2019.

Yang, L. and Wang, M. Reinforcement learning in feature space: Matrix bandit, kernels, and regret bound. In *International Conference on Machine Learning*, pp. 10746–10756. PMLR, 2020.

Zanette, A., Cheng, C.-A., and Agarwal, A. Cautiously optimistic policy optimization and exploration with linear function approximation. In *Conference on Learning Theory*, pp. 4473–4525. PMLR, 2021.

Algorithm 3 **S-Sampling (Sensitivity-Sampling)**

```
Input: Function class $\mathcal{F}$, current sensitivity dataset $\widehat{\mathcal{Z}}$, a new state-action pair $z$, failure probability $\delta$.
Compute the sensitivity factor of $z$ relative to the dataset $\mathcal{Z}$:
    $$s_z = C \cdot \text{sensitivity}_{\widehat{\mathcal{Z}},\mathcal{F}}(z) \cdot \log(N\mathcal{N}(\mathcal{F}, \sqrt{\delta/64N^3})/\delta)$$
Let $\widehat{z}$ be the neighbor of $z$ satisfying the condition (2)
if $s_z \geq 1$ then
    Add 1 copy of $\widehat{z}$ into $\mathcal{Z}$
else
    Let $n_z = \lfloor \frac{1}{s_z} \rfloor$ and add $n_z$ copies of $\widehat{z}$ into $\widehat{\mathcal{Z}}$ with probability $\frac{1}{n_z}$
end if
return $\widehat{\mathcal{Z}}$.
```

Algorithm 4 **Policy Update**

```
Input: $\pi_{cov}, b, \mathcal{K}$
Initialize: $\pi_0(\cdot|s) = \text{Unif}(\mathcal{A})$ if $s \in \mathcal{K}$ $and$
            $\pi_0(\cdot|s) = \text{Unif}(\{a|(s,a) \notin \mathcal{K}\})$ if $s \notin \mathcal{K}$
for $k = 1, 2, \cdots, K-1$ do
    if $k - \underline{k} > \kappa$ or $k = 0$ then
        $\underline{k} \leftarrow k$
        $D \leftarrow$ Behaviour Policy Sampling$(\pi_{\underline{k}}, \pi_{\text{cov}})$
    end if
    $\widehat{Q}_k \leftarrow$ Policy Evaluation Oracle$(\pi_k, D, \pi_{\underline{k}}, b)$
    Update policy: $\forall s \in \mathcal{K}$
    $\pi_{k+1}(\cdot|s) \propto \pi_k(\cdot|s) e^{\eta \widehat{Q}_k(\cdot|s)}$
end for
return: $\pi_{0:K-1} = \text{Unif}\{\pi_0, \cdots, \pi_{K-1}\}$
```

Algorithm 5 **Behaviour Policy Sampling**

```
Input: Behaviour Policy $\pi$, Policy Cover $\pi^{1:n}$
$D = \emptyset$
for $i = 1, \cdots, M$ do
    Sample j uniformly at random in $1 : n$
    $(s, a) \leftarrow$ d-sampler$(\pi_j, \nu)$
    Sample $h \geq 1$ with probability $\gamma^{h-1}(1-\gamma)$
    Continue the rollout from $(s, a)$ by executing $\pi$ for $h - 1$ steps
    Storing the rollout $D[i] \leftarrow \{(s_1, a_1, \cdots, s_h, a_h)\}$ where $(s_1, a_1) = (s, a)$
end for
return: $D$
```

Algorithm 6 Policy Evaluation Oracle

```
**Input**: Evaluate policy $\pi$, Dataset $D$, Behaviour policy $\underline{\pi}$, Bonus function $b$
**for** $i = 1, 2, \cdots, M$ **do**
    $\lambda_i \leftarrow \prod_{\tau=2}^{|D[i]|} \frac{\pi(a_\tau|s_\tau)}{\underline{\pi}(a_\tau|s_\tau)}$
    $(s_i^F, a_i^F) \leftarrow D[i]$ 's first sample
    $(s_i^L, a_i^L) \leftarrow D[i]$ 's last sample
    $G_i \leftarrow \frac{1}{1-\gamma}[r(s_i^L, a_i^L) + b(s_i^L, a_i^L)]$
**end for**
**end for**
$\widehat{f} \leftarrow \underset{f \in \mathcal{F}}{\text{argmin}} \sum_{i=1}^{M} \left(f(s_i^F, a_i^F) - (\lambda_i G_i - b(s_i^F, a_i^F))\right)^2$
**return:** $\widehat{Q}(s,a) = \widehat{f}(s,a) + \frac{1}{2}b(s,a),\ \forall s \in \mathcal{K}^n\ \ and\ \widehat{Q}(s,a) = \widehat{f}(s,a) + b(s,a)$, otherwise
```

Algorithm 7 d-sampler

```
**Input**: $\nu \in \Delta(S \times A), \pi$.
Sample $s_0, a_0 \sim \nu$.
Sample $\tau \geq 1$ with probability $\gamma^{\tau-1}(1-\gamma)$.
Execute $\pi$ for $\tau - 1$ steps, giving state s.
Sample action $a \sim \pi(\cdot|s)$.
**return** $(s, a)$.
```

A. Remaining Algorithm Pseudocodes

We provide the remaining algorithms including Sensitivity-Sampling (Algorithm 3), Policy Update (Algorithm 4), Behaviour Policy Sampling (Algorithm 5), Policy Evaluation Oracle (Algorithm 6), and the visitation distribution sampler (Algorithm 7).

B. Main Analysis

In this section, we provide the main guarantees for **LPO**.

B.1. Proof Setup

Bonus and auxiliary MDP To begin with, we introduce the concept of optimisic (bonus-added) and auxiliary MDP, which is also mentioned in (Agarwal et al., 2020a; Feng et al., 2021). However, we design these MDPs with very different bonus functions.

For each epoch $n \in [N]$, we consider an arbitrary fixed comparator policy $\widetilde{\pi}$ (e.g., an optimal policy) and three MDPs: the original MDP $\mathcal{M} := (\mathcal{S}, \mathcal{A}, P, r, \gamma)$, the bonus-added MDP $\mathcal{M}_{b^n} := (\mathcal{S}, \mathcal{A}, P, r + b^n, \gamma)$, and an auxiliary MDP $\mathcal{M}^n := (\mathcal{S}, \mathcal{A} \cup \{a^\dagger\}, P^n, r^n, \gamma)$, where $a^\dagger$ is an extra action which is only available for $s \notin \mathcal{K}^n$. For all $(s, a) \in \mathcal{S} \times \mathcal{A}$,

$$P^n(\cdot|s, a) = P(\cdot|s, a),\ r^n(s, a) = r(s, a) + b^n(s, a).$$

For $s \notin \mathcal{K}^n$

$$P^n(s|s, a^\dagger) = 1, r^n(s, a^\dagger) = 3$$

The above equation actually exhibits its property to absorb and provide maximum rewards for agent outside the known states. For readers' convenience, we present the definition of bonus function and known states.

The bonus function $b^n : \mathcal{S} \times \mathcal{A} \to \mathbb{R}$ defined as

$$\begin{aligned} b^n_w(s, a) &= \frac{1}{\beta}\, \omega(\widehat{\mathcal{F}}^n, s, a)\mathbf{1}\{s \in \mathcal{K}^n\} \\ b^n_1(s, a) &= \frac{3}{1-\gamma}\mathbf{1}\{s \notin \mathcal{K}^n\} \\ b^n(s, a) &= 2b^n_w(s, a) + b^n_1(s, a) \end{aligned} \tag{8}$$

Known states

$$\begin{aligned} \mathcal{K}^n =& \{s \in \mathcal{S} |\ \forall a \in \mathcal{A}, \omega(\widehat{\mathcal{F}}^n, s, a) < \beta\} \\ \mathcal{K}^n =& \{(s, a) \in \mathcal{S} \times \mathcal{A} |\ \omega(\widehat{\mathcal{F}}^n, s, a) < \beta\} \end{aligned} \tag{9}$$

Given the auxiliary MDP $\mathcal{M}^n$, we define $\widetilde{\pi}^n(\cdot|s) = \widetilde{\pi}(\cdot|s)$ for $s \in \mathcal{K}^n$ and $\widetilde{\pi}^n(a^\dagger|s) = 1$ for $s \notin \mathcal{K}^n$. We denote by $\widetilde{d}_{\mathcal{M}^n}$ the state-action distribution induced by $\widetilde{\pi}^n$ on $\mathcal{M}^n$ and $d^{\widetilde{\pi}}$ the state-action distribution induced by $\widetilde{\pi}$ on $\mathcal{M}$.

Given a policy π, we denote by $V^\pi_{b^n}, Q^\pi_{b^n}$, and $A^\pi_{b^n}$ the state-value, Q-value, and the advantage function of π on $\mathcal{M}_{b^n}$, and $V^\pi_{\mathcal{M}^n}, Q^\pi_{\mathcal{M}^n}$,and $A^\pi_{\mathcal{M}^n}$ for the counterparts on $\mathcal{M}^n$, and we define $Q^\pi(s, \widetilde{\pi}) := \mathbb{E}_{a\sim\widetilde{\pi}(\cdot|s)}[Q^\pi(s, a)]$.

Based on the above definitions and notations, we have the following lemmas.

Lemma B.1. *(Distribution Dominance) (Feng et al., 2021). Consider any state $s \in \mathcal{K}^n$, we have:*

$$\tilde{d}_{\mathcal{M}^n}(s, a) \leq d^{\tilde{\pi}}(s, a), \quad \forall a \in \mathcal{A}.$$

Proof. We prove by induction over the time steps along the horizon h. We denote the state-action distribution at the h_{th} step following $\tilde{\pi}^n$ on $\mathcal{M}^n$ as $\tilde{d}_{\mathcal{M}^n,h}$.

Starting at $h = 0$, if $s_0 \in \mathcal{K}^n$, then $\tilde{\pi}^n(\cdot \mid s_0) = \tilde{\pi}(\cdot \mid s_0)$ and

$$\tilde{d}_{\mathcal{M}^n,0}\left(s_0,a\right)=d_0^{\tilde{\pi}}\left(s_0,a\right), \quad \forall a\in\mathcal{A}.$$

Now we assume that at step h, for all $s\in\mathcal{K}^n$, it holds that

$$\tilde{d}_{\mathcal{M}^n,h}(s,a)\leq d_h^{\tilde{n}}(s,a), \forall a\in\mathcal{A}$$

Then, for step $h+1$, by definition we have that for $s\in\mathcal{K}^n$

$$\begin{aligned}
\tilde{d}_{\mathcal{M}^n,h+1}(s) &= \sum_{s',a'}\tilde{d}_{\mathcal{M}^n,h}\left(s',a'\right)P_{\mathcal{M}^n}\left(s\mid s',a'\right)\\
&= \sum_{s',a'}1\left\{s'\in\mathcal{K}^n\right\}\tilde{d}_{\mathcal{M}^n,h}\left(s',a'\right)P_{\mathcal{M}^n}\left(s\mid s',a'\right)\\
&= \sum_{s',a'}1\left\{s'\in\mathcal{K}^n\right\}\tilde{d}_{\mathcal{M}^n,h}\left(s',a'\right)P\left(s\mid s',a'\right)
\end{aligned}$$

where the second line is due to that if $s'\notin\mathcal{K}^n, \tilde{\pi}$ will deterministically pick $a^\dagger$ and $P_{\mathcal{M}^n}\left(s\mid s',a^\dagger\right)=0$. On the other hand, for $d_{h+1}^{\tilde{\pi}}(s,a)$, it holds that for $s\in\mathcal{K}^n$,

$$\begin{aligned}
d_{h+1}^{\tilde{\pi}}(s) &= \sum_{s',a'}d_h^{\tilde{\pi}}\left(s',a'\right)P\left(s\mid s',a'\right)\\
&= \sum_{s',a'}1\left\{s'\in\mathcal{K}^n\right\}d_h^{\tilde{\pi}}\left(s',a'\right)P\left(s\mid s',a'\right)+\sum_{s',a'}1\left\{s'\notin\mathcal{K}^n\right\}d_h^{\tilde{\pi}}\left(s',a'\right)P\left(s\mid s',a'\right)\\
&\geq \sum_{s',a'}1\left\{s'\in\mathcal{K}^n\right\}d_h^{\tilde{\pi}}\left(s',a'\right)P\left(s\mid s',a'\right)\\
&\geq \sum_{s',a'}1\left\{s'\in\mathcal{K}^n\right\}\tilde{d}_{\mathcal{M}^n,h}\left(s',a'\right)P\left(s\mid s',a'\right)\\
&= \tilde{d}_{\mathcal{M}^n,h+1}(s).
\end{aligned}$$

Using the fact that $\tilde{\pi}^n(\cdot\mid s)=\tilde{\pi}(\cdot\mid s)$ for $s\in\mathcal{K}^n$, we conclude that the inductive hypothesis holds at $h+1$ as well. Using the definition of the average state-action distribution, we conclude the proof.

□

Lemma B.2. *(Partial optimism) (Zanette et al., 2021). Fix a policy $\widetilde{\pi}$ that never takes $a^\dagger$. In any episode n it holds that*

$$V_{\mathcal{M}^n}^{\widetilde{\pi}^n}(\widetilde{s})-V^{\widetilde{\pi}}(\widetilde{s})\geq\frac{1}{1-\gamma}\mathbb{E}_{s\sim d_{\widetilde{s}}^{\widetilde{\pi}}}2b_\omega^n(s,\widetilde{\pi})$$

Proof. Notice that if $s\notin\mathcal{K}^n$, then $V_{\mathcal{M}^n}^{\widetilde{\pi}^n}(s)=\frac{3}{1-\gamma}$ as the policy self-loops in s by taking $a^\dagger$ there. Otherwise,

$$\begin{aligned}
V_{\mathcal{M}^n}^{\widetilde{\pi}^n}(s) &= \mathbb{E}_{a\sim\widetilde{\pi}^n(\cdot|s)}Q_{\mathcal{M}^n}^{\widetilde{\pi}^n}(s,a)\\
&= \frac{1}{1-\gamma}\mathbb{E}_{a\sim\tilde{\pi}^n(\cdot|s)}\mathbb{E}_{(s',a')\sim\tilde{d}_{\mathcal{M}^n}|(s,a)}r^n(s',a')\\
&\leq \frac{3}{1-\gamma}
\end{aligned} \tag{10}$$

Therefore, $V^{\widetilde{\pi}^n}_{\mathcal{M}^n}(s) \leq \frac{3}{1-\gamma}$. Using the performance difference lemma in (Kakade, 2001), we get:

$$\begin{aligned}
&(1-\gamma)(V^{\widetilde{\pi}^n}_{\mathcal{M}^n}(\widetilde{s}) - V^{\widetilde{\pi}}_{\mathcal{M}^n}(\widetilde{s})) = \mathbb{E}_{(s,a)\sim d^{\widetilde{\pi}}_{\widetilde{s}}}[-A^{\widetilde{\pi}^n}_{\mathcal{M}^n}(s,a)] \\
&= \mathbb{E}_{(s,a)\sim d^{\widetilde{\pi}}_{\widetilde{s}}}[V^{\widetilde{\pi}^n}_{\mathcal{M}^n}(s) - Q^{\widetilde{\pi}^n}_{\mathcal{M}^n}(s,a)] \\
&= \mathbb{E}_{s\sim d^{\widetilde{\pi}}_{\widetilde{s}}}[Q^{\widetilde{\pi}^n}_{\mathcal{M}^n}(s,\widetilde{\pi}^n) - Q^{\widetilde{\pi}^n}_{\mathcal{M}^n}(s,\widetilde{\pi})] \\
&= \mathbb{E}_{s\sim d^{\widetilde{\pi}}_{\widetilde{s}}}[\Big(Q^{\widetilde{\pi}^n}_{\mathcal{M}^n}(s,\widetilde{\pi}^n) - Q^{\widetilde{\pi}^n}_{\mathcal{M}^n}(s,\widetilde{\pi})\Big)\mathbf{1}\{s\notin\mathcal{K}^n\}] \\
&= \mathbb{E}_{s\sim d^{\widetilde{\pi}}_{\widetilde{s}}}[\left(\frac{3}{1-\gamma} - r(s,\widetilde{\pi}) - 2b^n_\omega(s,\widetilde{\pi}) - b^n_1(s,\widetilde{\pi}) - \gamma\mathbb{E}_{s'\sim P(s,\widetilde{\pi})}V^{\widetilde{\pi}^n}_{\mathcal{M}^n}(s')\right)\mathbf{1}\{s\notin\mathcal{K}^n\}]
\end{aligned} \tag{11}$$

Since $r(s,\widetilde{\pi}) + 2b^n_\omega(s,\widetilde{\pi}) + \gamma\mathbb{E}_{s'\sim P(s,\widetilde{\pi})}V^{\widetilde{\pi}^n}_{\mathcal{M}^n}(s') \leq 1 + 2 + \frac{3\gamma}{1-\gamma} \leq \frac{3}{1-\gamma}$

Thus,

$$\begin{aligned}
V^{\widetilde{\pi}^n}_{\mathcal{M}^n}(\widetilde{s}) &\geq V^{\widetilde{\pi}}_{\mathcal{M}^n}(\widetilde{s}) - \frac{1}{1-\gamma}\mathbb{E}_{s\sim d^{\widetilde{\pi}}_{\widetilde{s}}}b^n_1(s,\widetilde{\pi}) \\
&= V^{\widetilde{\pi}}(\widetilde{s}) + \frac{1}{1-\gamma}\mathbb{E}_{s\sim d^{\widetilde{\pi}}_{\widetilde{s}}}b^n(s,\widetilde{\pi}) - \frac{1}{1-\gamma}\mathbb{E}_{s\sim d^{\widetilde{\pi}}_{\widetilde{s}}}b^n_1(s,\widetilde{\pi}) \\
&= V^{\widetilde{\pi}}(\widetilde{s}) + \frac{1}{1-\gamma}\mathbb{E}_{s\sim d^{\widetilde{\pi}}_{\widetilde{s}}}2b^n_\omega(s,\widetilde{\pi})
\end{aligned} \tag{12}$$

□

Lemma B.3. *(Negative Advantage) (Zanette et al., 2021).*

$$A^{\pi^n}_{\mathcal{M}^n}(s,\widetilde{\pi}^n)\boldsymbol{1}\{s\notin\mathcal{K}^n\} \leq 0$$

Proof. Assume $s\notin\mathcal{K}^n$, then $Q^{\pi^n}_{\mathcal{M}^n}(s,\widetilde{\pi}^n) = 3 + \gamma V^{\pi^n}_{\mathcal{M}^n}(s)$. Note that if $s\notin\mathcal{K}^n$, π^n takes an action $a\neq a^\dagger$ such that $b^n_1(s,a) = \frac{3}{1-\gamma}$. Therefore, $V^{\pi^n}_{\mathcal{M}^n}(s) \geq \frac{3}{1-\gamma}$.
Combining the two expressions we obtain that, in any state $s\notin\mathcal{K}^n$,

$$A^{\pi^n}_{\mathcal{M}^n}(s,\widetilde{\pi}^n) = Q^{\pi^n}_{\mathcal{M}^n}(s,\widetilde{\pi}^n) - V^{\pi^n}_{\mathcal{M}^n}(s) = 3 + \gamma V^{\pi^n}_{\mathcal{M}^n}(s) - V^{\pi^n}_{\mathcal{M}^n}(s) \leq 0$$

□

B.2. Proof Sketch

In order to bound the gap of values between the output policy π^{LPO} =Unif($\pi^0,\pi^1,\cdots,\pi^{N-1}$) and the comparator $\widetilde{\pi}$, we need to analyze the gap between $V^{\widetilde{\pi}}$ and V^{π^n} for each $n\in[N]$. With the above three lemmas based on the structure of the well-designed MDPs, we are able to obtain the following regret decomposition (see details in Lemma B.11 (Regret decomposition)):

$$\begin{aligned}
\left(V^{\widetilde{\pi}} - V^{\pi^n}\right)(s_0) \leq \frac{1}{1-\gamma}\Bigg[&\underbrace{\sup_{s\in\mathcal{K}^n}\widehat{A}^{\pi^n}_{\mathcal{M}^n}(s,\widetilde{\pi})\mathbf{1}\{s\in\mathcal{K}^n\}}_{\text{term 1}} + \underbrace{\mathbb{E}_{s\sim\widetilde{d}_{\mathcal{M}^n}}[A^{\pi^n}_{\mathcal{M}^n}(s,\widetilde{\pi}) - A^{*}_{\mathcal{M}^n}(s,\widetilde{\pi})]\mathbf{1}\{s\in\mathcal{K}^n\}}_{\text{term 2}} \\
&+ \underbrace{\mathbb{E}_{s\sim\widetilde{d}_{\mathcal{M}^n}}[A^{*}_{\mathcal{M}^n}(s,\widetilde{\pi}) - \widehat{A}^{\pi^n}_{\mathcal{M}^n}(s,\widetilde{\pi})]\mathbf{1}\{s\in\mathcal{K}^n\}}_{\text{term 3}} - \underbrace{\mathbb{E}_{s\sim d^{\widetilde{\pi}}|s_0}2b^n_\omega(s,\widetilde{\pi})}_{\text{term 4}} + \underbrace{\mathbb{E}_{s\sim d^n|s_0}b^n(s,\pi^n)}_{\text{term 5}}\Bigg]
\end{aligned} \tag{13}$$

Now we discuss the details of each term.

- term 1 represents the *Solver Error*, which measures the performance of policy π^n in terms of empirical advantage function on known states. This term can be bounded by Lemma B.10 (NPG Analysis).

- term 2 represents the *Approximation Error*, which exists when our function class $\mathcal{F}$ do not have enough representability to deal with critic fitting, and this term can be controlled with Assumption B.5 (Bounded Transfer Error) and Lemma B.9
- term 3 represents the *Statistical Error*, which is the average error between the empirical and optimal advantage function under known states. This term can be bounded by term 4 (the expectation of width function) according to Lemma B.1 and Lemma E.8.
- term 5 is the expectation of bonus function under policy π^n, and the bound of bonuses is achieved in Lemma D.3

B.3. Analysis of LPO

In this part, we follow the above steps of proof to establish the result of our main theorem.

First, we introduce some notions and assumptions to handle the model misspecification. These notions have been discussed in (Agarwal et al., 2020a; Feng et al., 2021).

Definition B.4. (Transfer error). Given a target function $g : \mathcal{S} \times \mathcal{A} \to \mathbb{R}$, we define the critic loss function $L(f; d; g)$ with $d \in \Delta(\mathcal{S} \times \mathcal{A})$ as:

$$L(f; d; g) := \mathbb{E}_{(s,a)\sim d}\left[(f(s,a) - g(s,a))^2\right]$$

We let $Q_{b^n}^{\pi^n}$, $Q_{b^n}^{\pi_k}$ to be the Q-function in the bonus-added MDP for a given outer iteration n and an inner iteration k. Then the transfer error for a fixed comparator $\widetilde{\pi}$ is defined as

$$\epsilon_k^n := \inf_{f \in \mathcal{F}_k^n} L\left(f; \tilde{d}, Q_{b^n}^{\pi_k} - b^n\right),$$

where $\mathcal{F}_k^n := \operatorname{argmin}_{f\in\mathcal{F}} L\left(f; \rho_{cov}^n, Q_{b^n}^{\pi_k} - b^n\right)$ and $\tilde{d}(s,a) := d_{s_0}^{\tilde{\pi}}(s) \circ \mathrm{Unif}(\mathcal{A})$.

Assumption B.5. (Bounded Transfer Error). For the fixed comparator policy $\tilde{\pi}$, for every epoch $n \in [N]$ and every iteration k inside epoch n, we assume that there exists a constant ϵ_{bias} such that

$$\epsilon_k^n \leq \epsilon_{\text{bias}},$$

Notice that ϵ_{bias} measures both approximation error and distribution shift error. By the definition of transfer error, we can select a function $\tilde{f}_k^n \in \mathcal{F}_k^n$, such that

$$L\left(\tilde{f}_k^n; \tilde{d}; Q_{b^n}^{\pi_k} - b^n\right) \leq 2\epsilon_{\text{bias}}$$

Assumption B.6. For the same loss L in the Definition B.4 and the fitter $\tilde{f}_k^n$ in Assumption B.5, we assume that there exists some $C \geq 1$ and $\epsilon_0 \geq 0$ such that for any $f \in \mathcal{F}$,

$$\mathbb{E}_{(s,a)\sim\rho_{cov}^n}\left[\left(f(s,a) - \tilde{f}_k^n(s,a)\right)^2\right] \leq C \cdot \left(L\left(f; \rho_{cov}^n, Q_{b^n}^{\pi_k} - b^n\right) - L\left(\tilde{f}_k^n; \rho_{cov}^n, Q_{b^n}^{\pi_k} - b^n\right)\right) + \epsilon_0$$

for $n \in [N]$ and $0 \leq k \leq K-1$.

Remark B.7. Under Assumption 4.6, for every $n \in [N]$ and $k \in [K]$, $Q_{b^n}^{\pi_k}(s,a) - b^n(s,a) = \mathbb{E}^{\pi_k^n}\left[r(s,a) + \gamma Q_{b^n}^{\pi_k}(s',a') \,|s,a\right] \in \mathcal{F}$. Thus, ϵ_{bias} can take value 0 and $\tilde{f}_k^n = Q_{b^n}^{\pi_k} - b^n$. Further in Assumption B.6, we have

$$\mathbb{E}_{(s,a)\sim\rho_{cov}^n}\left[\left(f(s,a) - \tilde{f}_k^n(s,a)\right)^2\right] = L\left(f; \rho_{cov}^n, Q_{b^n}^{\pi_k} - b^n\right).$$

Hence, C can take value 1 and $\epsilon_0 = 0$. If $Q_{b^n}^{\pi_k} - b^n$ is not realizable in $\mathcal{F}$, ϵ_{bias} and ϵ_0 could be strictly positive. Hence, the above two assumptions are generalized version of the closedness condition considering model misspecification. Next, we define the optimal regression fit considering the loss function and its corresponding advantage functions.

Definition B.8.

$$\begin{aligned} & f^{n,*} \in \operatorname*{argmin}_{f\in\mathcal{F}} L(f;\rho^n, Q^{\pi^n}_{b^n} - b^n),\ f^*_k \in \operatorname*{argmin}_{f\in\mathcal{F}} L(f;\rho^n, Q^{\pi_k}_{b^n} - b^n) \\ & A^*_{b^n}(s,a) = f^{n,*}(s,a) + b^n(s,a) - \mathbb{E}_{a'\sim\pi^n(\cdot|s)}\left[f^{n,*}(s,a') + b^n(s,a')\right] \\ & A^{k,*}_{b^n}(s,a) = f^*_k(s,a) + b^n(s,a) - \mathbb{E}_{a'\sim\pi_k(\cdot|s)}\left[f^*_k(s,a') + b^n(s,a')\right] \end{aligned} \tag{14}$$

In the later proof, we select $f^{n,*}$, f^*_k not only to be the optimal solution with respect to the above loss function, but also satisfy the inequality in Assumption B.6, just like $\tilde{f}^n_k$.

Lemma B.9. *(Advantage Transfer Error decomposition). We have that*

$$\mathbb{E}_{(s,a)\sim\tilde{d}_{\mathcal{M}^n}}\left(A^k_{b^n} - A^{k,*}_{b^n}\right)\mathbf{1}\left\{s\in\mathcal{K}^n\right\} < 4\sqrt{|\mathcal{A}|\epsilon_{bias}}\,.$$

Proof. We have

$$\begin{aligned} & \mathbb{E}_{(s,a)\sim\tilde{d}_{\mathcal{M}^n}}\left(A^k_{b^n} - A^{k,*}_{b^n}\right)1\left\{s\in\mathcal{K}^n\right\} \\ & = \mathbb{E}_{(s,a)\sim\tilde{d}_{\mathcal{M}^n}}\left(Q^k_{b^n} - f^*_k - b^n\right)\mathbf{1}\left\{s\in\mathcal{K}^n\right\} - \mathbb{E}_{s\sim\tilde{d}_{\mathcal{M}^n},a\sim\pi_k(\cdot|s)}\left(Q^k_{b^n} - f^*_k - b^n\right)1\left\{s\in\mathcal{K}^n\right\} \\ & \leq \sqrt{\mathbb{E}_{(s,a)\sim\tilde{d}_{\mathcal{M}^n}}\left(Q^k_{b^n} - f^*_k - b^n\right)^2\mathbf{1}\left\{s\in\mathcal{K}^n\right\}} + \sqrt{\mathbb{E}_{s\sim\tilde{d}_{\mathcal{M}^n},a\sim\pi_k(|s)}\left(Q^k_{b^n} - f^*_k - b^n\right)^2 1\left\{s\in\mathcal{K}^n\right\}} \\ & \leq \sqrt{\mathbb{E}_{(s,a)\sim d^{\tilde{\pi}}}\left(Q^k_{b^n} - f^*_k - b^n\right)^2\mathbf{1}\left\{s\in\mathcal{K}^n\right\}} + \sqrt{\mathbb{E}_{s\sim d^{\tilde{\pi}},a\sim\pi_k(\cdot|s)}\left(Q^k_{b^n} - f^*_k - b^n\right)^2\mathbf{1}\left\{s\in\mathcal{K}^n\right\}} \\ & = \sqrt{\mathbb{E}_{(s,a)\sim\tilde{d}}|\mathcal{A}|\tilde{\pi}(a\mid s)\cdot\left(Q^k_{b^n} - f^*_k - b^n\right)^2\mathbf{1}\left\{s\in\mathcal{K}^n\right\}} + \sqrt{\mathbb{E}_{(s,a)\sim\tilde{d}}|\mathcal{A}|\pi_k(a\mid s)\cdot\left(Q^k_{b^n} - f^*_k - b^n\right)^2\mathbf{1}\left\{s\in\mathcal{K}^n\right\}} \\ & < 4\sqrt{|\mathcal{A}|\epsilon_{\text{bias}}}, \end{aligned}$$

where the first inequality is by Cauchy-Schwarz, the second inequality is by distribution dominance, and the last two lines follow the bounded transfer error assumption and the definition of f^*_k. □

Next, we provide the NPG Analysis.

Lemma B.10. *(NPG Analysis) (Agarwal et al., 2020a).*

Take stepsize $\eta = \sqrt{\frac{\log(|\mathcal{A}|)}{16W^2K}}$ *in the algorithm, then for any* $n\in[N]$ *we have*

$$\sum_{k=0}^{K-1}\mathbb{E}_{(s,a)\sim\tilde{d}_{\mathcal{M}^n}}\left[\widehat{A}^k_{b^n}(s,a)\mathbf{1}\left\{s\in\mathcal{K}^n\right\}\right] \leq 8W\sqrt{\log(|\mathcal{A}|)K}$$

Proof. Here we omit the index n for simplicity. From the NPG update rule

$$\begin{aligned} \pi_{k+1}(\cdot\mid s) & \propto \pi_k(\cdot\mid s)e^{\eta\widehat{Q}_k(s,\cdot)} \\ & \propto \pi_k(\cdot\mid s)e^{\eta\widehat{Q}_k(s,\cdot)}e^{-\eta\widehat{V}_k(s)} \\ & = \pi_k(\cdot\mid s)e^{\eta\widehat{A}_k(s,\cdot)} \end{aligned}$$

if we define the normalizer

$$z_k(s) = \sum_{a'}\pi_k\left(a'\mid s\right)e^{\eta\widehat{A}_k\left(s,a'\right)}$$

then the update can be written as

$$\pi_{k+1}(\cdot \mid s) = \frac{\pi_k(\cdot \mid s) e^{\eta \widehat{A}_k(s,\cdot)}}{z_k(s)}$$

Then for any $s \in \mathcal{K}^n$,

$$\begin{aligned}
&\mathbf{KL}\left(\tilde{\pi}(\cdot \mid s), \pi_{k+1}(\cdot \mid s)\right) - \mathbf{KL}\left(\widetilde{\pi}(\cdot \mid s), \pi_k(\cdot \mid s)\right) \\
&= \sum_a \tilde{\pi}(a \mid s) \log \frac{\widetilde{\pi}(a \mid s)}{\pi_{k+1}(a \mid s)} - \sum_a \widetilde{\pi}(a \mid s) \log \frac{\widetilde{\pi}(a \mid s)}{\pi_k(a \mid s)} \\
&= \sum_a \tilde{\pi}(a \mid s) \log \frac{\pi_k(a \mid s)}{\pi_{k+1}(a \mid s)} \\
&= \sum_a \tilde{\pi}(a \mid s) \log \left(z_k e^{-\eta \widehat{A}_k(s,a)} \right) \\
&= -\eta \sum_a \widetilde{\pi}(a \mid s) \widehat{A}_k(s,a) + \log z_k(s)
\end{aligned}$$

Since $\left|\widehat{A}_k(s,a)\right| \leq 4W$ and when $T > \log(|\mathcal{A}|), \eta < 1/(4W)$, we have $\eta \widehat{A}_k(s,a) \leq 1$. By the inequality that $\exp(x) \leq 1 + x + x^2$ for $x \leq 1$ and $\log(1+x) \leq x$ for $x > -1$

$$\log\left(z_k(s)\right) \leq \eta \sum_a \pi_k(a \mid s) \widehat{A}_k(s,a) + 16\eta^2 W^2 = 16\eta^2 W^2$$

Thus for $s \in \mathcal{K}^n$ we have

$$\mathbf{KL}\left(\tilde{\pi}(\cdot \mid s), \pi_{k+1}(\cdot \mid s)\right) - \mathbf{KL}\left(\tilde{\pi}(\cdot \mid s), \pi_k(\cdot \mid s)\right) \leq -\eta \mathbb{E}_{a \sim \tilde{\pi}^n(\cdot \mid s)} \left[\widehat{A}_k(s,a)\right] + 16\eta^2 W^2$$

By taking sum over k, we get

$$\begin{aligned}
&\sum_{k=0}^{K-1} \mathbb{E}_{(s,a) \sim \tilde{d}_{\mathcal{M}^n}} \left[\widehat{A}_k(s,a) \mathbf{1}\left\{s \in \mathcal{K}^n\right\}\right] \\
=& \sum_{k=0}^{K-1} \frac{1}{\eta} \mathbb{E}_{s \sim \tilde{d}_{\mathcal{M}^n}} \left[\left(\mathbf{KL}\left(\tilde{\pi}(\cdot \mid s), \pi_0(\cdot \mid s)\right) - \mathbf{KL}\left(\tilde{\pi}(\cdot \mid s), \pi_K(\cdot \mid s)\right)\right) \mathbf{1}\left\{s \in \mathcal{K}^n\right\}\right] + 16\eta K W^2 \\
\leq& \log(|\mathcal{A}|)/\eta + 16\eta K W^2 \\
\leq& 8W\sqrt{\log(|\mathcal{A}|)K}.
\end{aligned}$$

□

Now we are ready to analyze the regret decomposition.

Lemma B.11. *(Regret decomposition). With probability at least* $1 - \delta$ *it holds that*

$$\frac{1}{N} \sum_{n=1}^{N} \left(V^{\widetilde{\pi}} - V^{\pi^n}\right)(s_0) \leq \frac{\mathcal{R}(K)}{(1-\gamma)K} + \frac{2\sqrt{2A\epsilon_{bias}}}{1-\gamma} + \frac{1}{\sqrt{N}} \widetilde{O}\left(\frac{\sqrt{d^2\epsilon}}{(1-\gamma)^2\beta}\right) \tag{15}$$

Proof. Fix a policy $\widetilde{\pi}$ on $\mathcal{M}$. Consider the following decomposition for an outer episode n,

$$\begin{aligned}\left(V^{\widetilde{\pi}} - V^{\pi^n}\right)(s_0) = {} & \underbrace{V^{\widetilde{\pi}}(s_0) + \frac{1}{1-\gamma}\mathbb{E}_{s\sim d^{\widetilde{\pi}}|s_0} 2b^n_\omega(s,\widetilde{\pi})}_{\le V^{\tilde{\pi}^n}_{\mathcal{M}^n}(s_0) \text{ by Lemma B.2}} - \underbrace{V^{\pi^n}(s_0) - \frac{1}{1-\gamma}\mathbb{E}_{s\sim d^n|s_0} b^n(s,\pi^n)}_{=-V^{\pi^n}_{b^n}} \\ & + \frac{1}{1-\gamma}\underbrace{\left[-\mathbb{E}_{s\sim d^{\widetilde{\pi}}|s_0} 2b^n_\omega(s,\widetilde{\pi}) + \mathbb{E}_{s\sim d^n|s_0} b^n(s,\pi^n)\right]}_{\overset{def}{=} B^n}\end{aligned} \tag{16}$$

We put the term B^n aside for a moment and use performance difference lemma to obtain

$$\begin{aligned}V^{\tilde{\pi}^n}_{\mathcal{M}^n}(s_0) - V^{\pi^n}_{b^n}(s_0) &= V^{\tilde{\pi}^n}_{\mathcal{M}^n}(s_0) - V^{\pi^n}_{\mathcal{M}^n}(s_0) \\ &= \frac{1}{1-\gamma}\mathbb{E}_{s\sim\widetilde{d}_{\mathcal{M}^n}}\left[A^{\pi^n}_{\mathcal{M}^n}(s,\widetilde{\pi}^n)\right] \\ &= \frac{1}{1-\gamma}\mathbb{E}_{s\sim\widetilde{d}_{\mathcal{M}^n}}\left[A^{\pi^n}_{\mathcal{M}^n}(s,\widetilde{\pi}^n)\mathbf{1}\{s\in\mathcal{K}^n\} + \underbrace{A^{\pi^n}_{\mathcal{M}^n}(s,\widetilde{\pi}^n)\mathbf{1}\{s\notin\mathcal{K}^n\}}_{\le 0 \text{ by Lemma B.3}}\right] \\ &\le \frac{1}{1-\gamma}\mathbb{E}_{s\sim\widetilde{d}_{\mathcal{M}^n}}\left[A^{\pi^n}_{\mathcal{M}^n}(s,\widetilde{\pi})\mathbf{1}\{s\in\mathcal{K}^n\}\right]\end{aligned} \tag{17}$$

where the last step is because on states $s\in\mathcal{K}^n$ we have $\widetilde{\pi}^n(\cdot|s) = \widetilde{\pi}(\cdot|s)$.
Define

$$\begin{aligned}\widehat{A}^k_{b^n}(s,a) &= \widehat{Q}^k_{b^n}(s,a) - \widehat{V}^k_{b^n}(s) \\ &= f_k(s,a) + b^n_\omega(s,a) - \mathbb{E}_{a'\sim\pi_k(\cdot|s)}\left[f_k(s,a') + b^n_\omega(s,a')\right]\end{aligned} \tag{18}$$

and

$$\widehat{A}^{\pi^n}_{b^n}(s,a) = \frac{1}{K}\sum_{k=0}^{K-1}\widehat{A}^k_{b^n}(s,a) \tag{19}$$

Then we get

$$\begin{aligned}&= \frac{1}{1-\gamma}\left[\mathbb{E}_{s\sim\widetilde{d}_{\mathcal{M}^n}}\widehat{A}^{\pi^n}_{\mathcal{M}^n}(s,\widetilde{\pi})\mathbf{1}\{s\in\mathcal{K}^n\} + \mathbb{E}_{s\sim\widetilde{d}_{\mathcal{M}^n}}\left[A^{\pi^n}_{\mathcal{M}^n}(s,\widetilde{\pi}) - \widehat{A}^{\pi^n}_{\mathcal{M}^n}(s,\widetilde{\pi})\right]\mathbf{1}\{s\in\mathcal{K}^n\}\right] \\ &\le \frac{1}{1-\gamma}\left[\underbrace{\sup_{s\in\mathcal{K}^n}\widehat{A}^{\pi^n}_{\mathcal{M}^n}(s,\widetilde{\pi})\mathbf{1}\{s\in\mathcal{K}^n\}}_{\text{term 1}} + \underbrace{\mathbb{E}_{s\sim\widetilde{d}_{\mathcal{M}^n}}[A^{\pi^n}_{\mathcal{M}^n}(s,\widetilde{\pi}) - A^*_{\mathcal{M}^n}(s,\widetilde{\pi})]\mathbf{1}\{s\in\mathcal{K}^n\}}_{\text{term 2}}\right. \\ &\qquad \left. + \underbrace{\mathbb{E}_{s\sim\widetilde{d}_{\mathcal{M}^n}}[A^*_{\mathcal{M}^n}(s,\widetilde{\pi}) - \widehat{A}^{\pi^n}_{\mathcal{M}^n}(s,\widetilde{\pi})]\mathbf{1}\{s\in\mathcal{K}^n\}}_{\text{term 3}}\right]\end{aligned} \tag{20}$$

The first term can be bounded by Lemma B.10 (NPG lemma):

$$\sup_{s\in\mathcal{K}^n}\widehat{A}^{\pi^n}_{\mathcal{M}^n}(s,\widetilde{\pi})\mathbf{1}\{s\in\mathcal{K}^n\} = \sup_{s\in\mathcal{K}^n}\frac{1}{K}\sum_{k=0}^{K-1}\mathbb{E}_{a\sim\widetilde{\pi}(\cdot|s)}\widehat{A}^k_{b^n}(s,a)\mathbf{1}\{s\in\mathcal{K}^n\} \le \frac{\mathcal{R}(K)}{K} \tag{21}$$

The second term can be bounded by Lemma B.1, Lemma B.9

$$\begin{aligned}&\mathbb{E}_{s\sim\widetilde{d}_{\mathcal{M}^n}}[A^{\pi^n}_{\mathcal{M}^n}(s,\widetilde{\pi}) - A^*_{\mathcal{M}^n}(s,\widetilde{\pi})]\mathbf{1}\{s\in\mathcal{K}^n\} \\ &\le \mathbb{E}_{s\sim d^{\widetilde{\pi}}}[A^{\pi^n}_{\mathcal{M}^n}(s,\widetilde{\pi}) - A^*_{\mathcal{M}^n}(s,\widetilde{\pi})]\mathbf{1}\{s\in\mathcal{K}^n\} \\ &\le 2\sqrt{2A\epsilon_{\text{bias}}}\end{aligned} \tag{22}$$

The third term can be bounded by Lemma E.8, which ensures that with probability at least $1-\frac{\delta}{2}$ it holds that

$$\forall n\in[N],\ \forall k\in\{0,\cdots,K-1\},\ \forall(s,a)\in\mathcal{K}^n:\ 0\le Q^{k,*}_{b^n}(s,a)-\widehat{Q}^k_{b^n}(s,a)\le 2b^n_\omega(s,a) \tag{23}$$

Then we have: $\forall n\in[N],\forall(s,a)\in\mathcal{K}^n$:

$$\begin{aligned}A^*_{\mathcal{M}^n}(s,a)-\widehat{A}^{\pi^n}_{\mathcal{M}^n}(s,a)&=\frac{1}{K}\sum_{k=0}^{K-1}\left[\left(Q^{k,*}_{b^n}(s,a)-\widehat{Q}^k_{b^n}(s,a)\right)-\underbrace{\left(Q^{k,*}_{b^n}(s,\pi^n_k)-\widehat{Q}^k_{b^n}(s,\pi^n_k)\right)}_{\ge 0}\right]\\&\le Q^*_{\mathcal{M}^n}(s,a)-\widehat{Q}^{\pi^n}_{\mathcal{M}^n}(s,a)\end{aligned} \tag{24}$$

Apply the Lemma B.1, we have

$$\begin{aligned}\mathbb{E}_{s\sim\widetilde{d}_{\mathcal{M}^n}}[A^*_{\mathcal{M}^n}(s,\widetilde{\pi})-\widehat{A}^{\pi^n}_{\mathcal{M}^n}(s,\widetilde{\pi})]\mathbf{1}\{s\in\mathcal{K}^n\}&\le\mathbb{E}_{s\sim\widetilde{d}_{\mathcal{M}^n}}[Q^*_{\mathcal{M}^n}(s,\widetilde{\pi})-\widehat{Q}^{\pi^n}_{\mathcal{M}^n}(s,\widetilde{\pi})]\mathbf{1}\{s\in\mathcal{K}^n\}\\&\le\mathbb{E}_{s\sim d^{\widetilde{\pi}}}[Q^*_{\mathcal{M}^n}(s,\widetilde{\pi})-\widehat{Q}^{\pi^n}_{\mathcal{M}^n}(s,\widetilde{\pi})]\mathbf{1}\{s\in\mathcal{K}^n\}\end{aligned} \tag{25}$$

As a result,

$$\begin{aligned}\left(V^{\widetilde{\pi}}-V^{\pi^n}\right)(s_0)&\le\frac{1}{1-\gamma}\left[\frac{\mathcal{R}(K)}{K}+2\sqrt{2A\epsilon_{\text{bias}}}+\mathbb{E}_{s\sim d^{\widetilde{\pi}}}2b^n_\omega(s,\widetilde{\pi})\mathbf{1}\{s\in\mathcal{K}^n\}+B^n\right]\\&=\frac{1}{1-\gamma}\left[\frac{\mathcal{R}(K)}{K}+2\sqrt{2A\epsilon_{\text{bias}}}+\mathbb{E}_{s\sim d^n}b^n(s,\pi^n)\right]\end{aligned} \tag{26}$$

And finally using the concentration of bonuses (Lemma D.3) we get

$$\begin{aligned}\frac{1}{N}\sum_{n=1}^N\left(V^{\widetilde{\pi}}-V^{\pi^n}\right)(s_0)&\le\frac{\mathcal{R}(K)}{(1-\gamma)K}+\frac{2\sqrt{2A\epsilon_{\text{bias}}}}{1-\gamma}+\frac{1}{N(1-\gamma)}\sum_{n=1}^N\mathbb{E}_{s\sim d^n}b^n(s,\pi^n)\\&\le\frac{\mathcal{R}(K)}{(1-\gamma)K}+\frac{2\sqrt{2A\epsilon_{\text{bias}}}}{1-\gamma}+\frac{1}{\sqrt{N}}\widetilde{O}\left(\frac{\sqrt{d^2\epsilon}}{(1-\gamma)^2\beta}\right)\end{aligned} \tag{27}$$

□

Combining all previous lemmas, we have the following theorem about the sample complexity of our LPO.

Theorem B.12. *(Main Results: Sample Complexity of LPO). Under Assumption 4.3, 4.5, and 4.6, for any comparator $\widetilde{\pi}$, a fixed failure probability δ, eluder dimension $d=dim(\mathcal{F},1/N)$, a suboptimality gap ε and appropriate input hyperparameters: $N\ge\widetilde{O}\left(\frac{d^2}{(1-\gamma)^4\varepsilon^2}\right), K=\widetilde{O}\left(\frac{\ln|\mathcal{A}|W^2}{(1-\gamma)^2\varepsilon^2}\right), M\ge\widetilde{O}\left(\frac{d^2}{(1-\gamma)^4\varepsilon^2}\right), \eta=\widetilde{O}\left(\frac{\sqrt{\ln|\mathcal{A}|}}{\sqrt{K}W}\right), \kappa=\widetilde{O}\left(\frac{1-\gamma}{\eta W}\right)$, our algorithm returns a policy π^{LPO}, satisfying*

$$\left(V^{\widetilde{\pi}}-V^{\pi^{LPO}}\right)(s_0)\le\varepsilon.$$

with probability at least $1-\delta$ after taking at most $\widetilde{O}\left(\frac{d^3}{(1-\gamma)^8\varepsilon^3}\right)$ samples.

Proof. First, let's consider Lemma B.11 (Regret decomposition). We need ensure

$$\frac{\mathcal{R}(K)}{(1-\gamma)K}=\frac{8W}{(1-\gamma)}\sqrt{\frac{\ln|\mathcal{A}|}{K}}\le\frac{\varepsilon}{2}\quad\longrightarrow\quad K=\widetilde{O}\left(\frac{\ln|\mathcal{A}|W^2}{(1-\gamma)^2\varepsilon^2}\right)$$

This gives the inner iteration complexity. Next, β can be any constant between 0 and 1, and recall that ϵ is the parameter that controls the width function (3). We set ϵ in the following form (see Lemma E.6 for justification)

$$\epsilon=100\left(\frac{3}{2}C_1N\cdot\epsilon_{\text{stat}}+20NW\epsilon_1+\frac{1}{2}C_2\cdot\ln\left(\frac{N\mathcal{N}(\Delta\mathcal{F},2\epsilon_1)}{\delta'}\right)\right)$$

and

$$\epsilon_{\text{stat}} = \frac{500C \cdot W^4 \cdot \log\left(\frac{\mathcal{N}(\mathcal{F},\epsilon_2)}{\delta_3}\right)}{M} + 13W^2 \cdot \epsilon_2$$

where ϵ_1, ϵ_2 represents the function cover radius. Since our algorithm complexity depends only logarithmically on the covering numbers, we can set the cover radius with any polynomial degree of ε. In fact, $\epsilon_1 = O(\varepsilon^3)$, $\epsilon_2 = O(\varepsilon^3)$, $\epsilon = O(\log N)$,

$$\frac{1}{\sqrt{N}}\widetilde{O}\left(\frac{\sqrt{d^2\epsilon}}{(1-\gamma)^2\beta}\right) \leq \frac{\varepsilon}{2} \longrightarrow M = N \geq \widetilde{O}\left(\frac{d^2}{(1-\gamma)^4\varepsilon^2}\right)$$

gives the outer iteration complexity and the number of samples collected by a single Monte Carlo trajectory.

Under Assumption 4.6, which means Q-function is realizable in our function class $\mathcal{F}$, $\epsilon_{\text{bias}} = 0$ (see Remark B.7 for justification). After setting the hyperparameters above, with probability at least $1-\delta$, we have

$$\frac{1}{N}\sum_{n=1}^{N}\left(V^{\widetilde{\pi}} - V^{\pi^n}\right)(s_0) \leq \varepsilon$$

Remember that our algorithm outputs a uniform mixture of policy cover π^{LPO} =Unif($\pi^0, \pi^1, \cdots, \pi^{N-1}$), so we have

$$\left(V^{\widetilde{\pi}} - V^{\pi^{\text{LPO}}}\right)(s_0) \leq \varepsilon.$$

Next, we consider the total samples we need to collect through steps of the algorithm.

Every time the bonus switches, Algorithm 4 is invoked, and runs for K iterations. From Lemma E.4 we know that once data are collected, they can be reused for the next κ policies. Therefore, we actually run Algorithm 5 for $\lceil \frac{K}{\kappa} \rceil$ times, and whenever invoking Algorithm 5, we need M samples by Monte Carlo sampling. We denote S to be the number of bonus switches given in Proposition C.7 (Number of Switches).

In total, the sample complexity of our algorithm is

$$\begin{aligned}
&\underbrace{S}_{\text{number of inner loops invoked}} \times \underbrace{\left\lceil \frac{K}{\kappa} \right\rceil}_{\text{the inner iteration}} \times \underbrace{M}_{\text{Monte Carlo trajectories}} \\
&= \widetilde{O}\left(d \times \frac{2\ln(1/\delta)\left(\frac{\sqrt{\ln|\mathcal{A}|}}{\sqrt{K}W}\right)(B+W)}{(1-\gamma)\ln 2} \times K \times M\right) \\
&= \widetilde{O}\left(d\left(\frac{B}{W}+1\right)\frac{\sqrt{K}}{1-\gamma}M\right) \\
&= \widetilde{O}\left(\frac{d}{1-\gamma} \times \frac{W}{(1-\gamma)\varepsilon} \times \frac{d^2}{(1-\gamma)^4\varepsilon^2}\right) \\
&= O\left(\frac{d^3}{(1-\gamma)^8\varepsilon^3}\right)
\end{aligned}$$

We complete the proof of our main theorem. □

C. The Number of Switches

In this section, we will give the bound of the number of switching policies in the outer loop.

Recall that the width function is

$$\omega(\widehat{\mathcal{F}}^n, s, a) = \sup_{f_1, f_2 \in \mathcal{F}, ||f_1 - f_2||^2_{\widehat{\mathcal{Z}}^n} \leq \epsilon} |f_1(s,a) - f_2(s,a)|$$

The parameter ϵ will be defined later in (35). In fact, we will show that $\epsilon = O(\log N)$ in Lemma E.6 and Lemma E.7. First, we need to show that for every $n \in [N]$, the sensitivity dataset $\widehat{\mathcal{Z}}^n$ approximates the original dataset $\mathcal{Z}^n$. Our approach is inspired by (Kong et al., 2021).

For all $n \in [N]$ and $\alpha \in [\epsilon, +\infty)$, we define the following quantities

$$\begin{aligned}
\underline{\mathcal{B}}^n(\alpha) &:= \left\{(f_1, f_2) \in \mathcal{F} \times \mathcal{F} \mid \|f_1 - f_2\|_{\mathcal{Z}^n}^2 \leq \alpha/100\right\} \\
\mathcal{B}^n(\alpha) &:= \left\{(f_1, f_2) \in \mathcal{F} \times \mathcal{F} \mid \min\left\{\|f_1 - f_2\|_{\widehat{\mathcal{Z}}^n}^2, 4NW^2\right\} \leq \alpha\right\} \\
\overline{\mathcal{B}}^n(\alpha) &:= \left\{(f_1, f_2) \in \mathcal{F} \times \mathcal{F} \mid \|f_1 - f_2\|_{\mathcal{Z}^n}^2 \leq 100\alpha\right\}
\end{aligned}$$

For each $n \in [N]$, we use $\mathcal{E}^n(\alpha)$ to denote the event that

$$\underline{\mathcal{B}}^n(\alpha) \subseteq \mathcal{B}^n(\alpha) \subseteq \overline{\mathcal{B}}^n(\alpha)$$

Furthermore, we denote that

$$\mathcal{E}^n := \bigcap_{j=0}^{\infty} \mathcal{E}^n\left(100^j \epsilon\right),$$

Our goal is to show that the event $\mathcal{E}^n$ holds with great probability, which means $\widehat{\mathcal{Z}}^n$ is a good approximation to $\mathcal{Z}^n$.

Before presenting the proof, we need the following concentration inequality proved in (Freedman, 1975).

Lemma C.1. *Consider a real-valued martingale $\{Y_k : k = 0, 1, 2, \cdots\}$ with difference sequence $\{X_k : k = 0, 1, 2, \cdots\}$. Assume that the difference sequence is uniformly bounded:*

$$|X_k| \leq R \quad \textit{almost surely for} \quad k = 1, 2, 3, \cdots$$

For a fixed $n \in \mathbb{N}$, assume that

$$\sum_{k=1}^{n} \mathbb{E}_{k-1}\left(X_k^2\right) \leq \sigma^2$$

almost surely. Then for all $t \geq 0$,

$$P\left\{|Y_n - Y_0| \geq t\right\} \leq 2\exp\left\{-\frac{t^2/2}{\sigma^2 + Rt/3}\right\}$$

Furthermore, we need a bound on the number of elements in the sensitivity dataset. This is established in (Kong et al., 2021).

Lemma C.2. *With probability at least $1 - \delta/64N$,*

$$\left|\widehat{\mathcal{Z}}^n\right| \leq 64N^3/\delta \quad \forall n \in [N].$$

The following lemma shows that if $\mathcal{E}^n$ happens, $\widehat{\mathcal{Z}}^n$ is a good approximation to $\mathcal{Z}^n$. And this is proved in (Kong et al., 2021).

Lemma C.3. *If $\mathcal{E}^n$ occurs, then*

$$\frac{1}{10000}\|f_1 - f_2\|_{\mathcal{Z}^n}^2 \leq \min\left\{\|f_1 - f_2\|_{\widehat{\mathcal{Z}}^n}^2, 4NW^2\right\} \leq 10000\|f_1 - f_2\|_{\mathcal{Z}^n}^2, \quad \forall \|f_1 - f_2\|_{\mathcal{Z}^n}^2 > 100\epsilon$$

and

$$\min\left\{\|f_1-f_2\|_{\widehat{\mathcal{Z}}^n}^2, 4NW^2\right\} \leq 10000\epsilon, \quad \forall \|f_1-f_2\|_{\mathcal{Z}^n}^2 \leq 100\epsilon$$

To establish our result, we need the following lemma. The proof follows the approach of (Kong et al., 2021). We present it here for completeness.

Lemma C.4. *For* $\alpha \in \left[\epsilon, 4NW^2\right]$

$$\operatorname{Pr}\left(\mathcal{E}^1\mathcal{E}^2\ldots\mathcal{E}^{n-1}\left(\mathcal{E}^n(\alpha)\right)^c\right) \leq \delta/\left(32N^2\right)$$

Proof. We use $\overline{\mathcal{Z}}^n$ to denote the dataset without rounding, i.e., we replace every element $\hat{z}$ with z. Denote $C_1 = C \cdot \log\left(N \cdot \mathcal{N}\left(\mathcal{F}, \sqrt{\delta/64N^3}\right)/\delta\right)$, where C will be chosen appropriately later. We consider any fixed pair $(f_1, f_2) \in \mathcal{C}\left(\mathcal{F}, \sqrt{\delta/\left(64N^3\right)}\right) \times \mathcal{C}\left(\mathcal{F}, \sqrt{\delta/\left(64N^3\right)}\right)$.

For each $i \geq 2$, define

$$Z_i = \max\left\{\|f_1-f_2\|_{\mathcal{Z}^i}^2, \min\left\{\|f_1-f_2\|_{\widehat{\mathcal{Z}}^{i-1}}^2, 4NW^2\right\}\right\}$$

and

$$Y_i = \begin{cases} \frac{1}{p_{z_{i-1}}}\left(f_1\left(z_{i-1}\right)-f_2\left(z_{i-1}\right)\right)^2 & z_{i-1} \text{ is added into } \overline{\mathcal{Z}}^i \text{ and } Z_i \leq 2000000\alpha \\ 0 & z_{i-1} \text{ is not added into } \overline{\mathcal{Z}}^i \text{ and } Z_i \leq 2000000\alpha \\ \left(f_1\left(z_{i-1}\right)-f_2\left(z_{i-1}\right)\right)^2 & Z_i > 2000000\alpha \end{cases}$$

Note that Z_i is constant under F_{i-1} and Y_i is adapted to the filtration F_i, thus

$$\mathbb{E}_{i-1}\left[Y_i\right] = \left(f_1\left(z_{i-1}\right)-f_2\left(z_{i-1}\right)\right)^2$$

now we bound Y_i and its variance in order to use Freedman's inequality.

If $p_{z_{i-1}} = 1$ or $Z_i > 2000000\alpha$, then $Y_i - \mathbb{E}_{i-1}\left[Y_i\right] = \operatorname{Var}_{i-1}\left[Y_i - \mathbb{E}_{i-1}\left[Y_i\right]\right] = 0$. Otherwise by the definition of p_z we have

$$\begin{aligned}\left|Y_i - \mathbb{E}_{i-1}\left[Y_i\right]\right| &\leq \left(\min\left\{\|f_1-f_2\|_{\widehat{\mathcal{Z}}^{i-1}}^2, 4NW^2\right\}+1\right)/C_1 \\ &\leq 3000000\alpha/C_1\end{aligned}$$

and

$$\begin{aligned}\operatorname{Var}_{i-1}\left[Y_i - \mathbb{E}_{i-1}\left[Y_i\right]\right] &\leq \frac{1}{p_{z_{i-1}}}\left(f_1\left(z_{i-1}\right)-f_2\left(z_{i-1}\right)\right)^4 \\ &\leq \left(f_1\left(z_{i-1}\right)-f_2\left(z_{i-1}\right)\right)^2 \cdot 3000000\alpha/C_1\end{aligned}$$

Taking sum with respect to i yields

$$\sum_{i=2}^{n} \operatorname{Var}_{i-1}\left[Y_i - \mathbb{E}_{i-1}\left[Y_i\right]\right] \leq (3000000\alpha)^2/C_1$$

By Freedman's inequality, we have

$$
\begin{aligned}
&\mathbb{P}\left\{\left|\sum_{i=2}^{n}\left(Y_i-\mathbb{E}_{i-1}\left[Y_i\right]\right)\right| \geq \alpha/100\right\} \\
&\leq 2\exp\left\{-\frac{(\alpha/100)^2/2}{(3000000\alpha)^2/C_1+\alpha\cdot 3000000\alpha/3C_1}\right\} \\
&\leq \left(\delta/64N^2\right)/\left(\mathcal{N}\left(\mathcal{F},\sqrt{\delta/\left(64N^3\right)}\right)\right)^2
\end{aligned}
$$

where the last inequality is guaranteed by taking C appropriately large.

Taking a union bound over all $(f_1, f_2) \in \mathcal{C}\left(\mathcal{F},\sqrt{\delta/\left(64N^3\right)}\right) \times \mathcal{C}\left(\mathcal{F},\sqrt{\delta/\left(64N^3\right)}\right)$, with probability at least $1-\delta/\left(64T^2\right)$, we have

$$
\left|\sum_{i=2}^{n}\left(Y_i-\mathbb{E}_{i-1}\left[Y_i\right]\right)\right| \leq \alpha/100.
$$

In the rest of the proof, we condition on the event above and the event defined in Lemma C.2.

Part 1 $(\underline{\mathcal{B}}^n(\alpha) \subseteq \mathcal{B}^n(\alpha))$: Consider any pair $f_1, f_2 \in \mathcal{F}$ with $\|f_1-f_2\|_{\mathcal{Z}^n}^2 \leq \alpha/100$. From the definition we know that there exist $\left(\hat{f}_1,\hat{f}_2\right) \in \mathcal{C}\left(\mathcal{F},\sqrt{\delta/\left(64N^3\right)}\right) \times \mathcal{C}\left(\mathcal{F},\sqrt{\delta/\left(64N^3\right)}\right)$ such that $\left\|\hat{f}_1-f_1\right\|_\infty, \left\|\hat{f}_2-f_2\right\|_\infty \leq \sqrt{\delta/\left(64N^3\right)}$. Then we have that

$$
\begin{aligned}
\left\|\hat{f}_1-\hat{f}_2\right\|_{\mathcal{Z}^n}^2 &\leq \left(\|f_1-f_2\|_{\mathcal{Z}^n}+\left\|f_1-\hat{f}_1\right\|_{\mathcal{Z}^n}+\left\|\hat{f}_2-f_2\right\|_{\mathcal{Z}^n}\right)^2 \\
&\leq \left(\|f_1-f_2\|_{\mathcal{Z}^n}+2\cdot\sqrt{|\mathcal{Z}^n|}\cdot\sqrt{\delta/\left(64N^3\right)}\right)^2 \\
&\leq \alpha/50
\end{aligned}
$$

We consider the Y_i 's which correspond to $\hat{f}_1$ and $\hat{f}_2$. Because $\left\|\hat{f}_1-\hat{f}_2\right\|_{\mathcal{Z}^n}^2 \leq \alpha/50$, we also have that $\left\|\hat{f}_1-\hat{f}_2\right\|_{\mathcal{Z}^{n-1}}^2 \leq \alpha/50$. From $\mathcal{E}^{n-1}$ we know that $\min\left\{\left\|\hat{f}_1-\hat{f}_2\right\|_{\widehat{\mathcal{Z}}^{n-1}}^2, 4NW^2\right\} \leq 10000\alpha$. Then from the definition of Y_i we have

$$
\left\|\hat{f}_1-\hat{f}_2\right\|_{\overline{\mathcal{Z}}^n}^2 = \sum_{i=2}^{n} Y_i
$$

Then $\left\|\hat{f}_1-\hat{f}_2\right\|_{\overline{\mathcal{Z}}^n}^2$ can be bounded in the following manner:

$$
\begin{aligned}
\left\|\hat{f}_1-\hat{f}_2\right\|_{\mathcal{Z}^n}^2 &= \sum_{i=2}^{n} Y_i \\
&\leq \sum_{i=2}^{n}\mathbb{E}_{i-1}\left[Y_i\right]+\alpha/100 \\
&= \left\|\hat{f}_1-\hat{f}_2\right\|_{\mathcal{Z}^n}^2+\alpha/100 \\
&\leq 3\alpha/100
\end{aligned}
$$

As a result, $\left\|\hat{f}_1-\hat{f}_2\right\|_{\overline{\mathcal{Z}}^n}^2$ can also be bounded:

$$
\begin{aligned}
\|f_1 - f_2\|^2_{\overline{\mathcal{Z}}^n} &\le \left(\left\|\hat{f}_1 - \hat{f}_2\right\|_{\overline{\mathcal{Z}}^n} + \left\|f_1 - \hat{f}_1\right\|_{\overline{\mathcal{Z}}^n} + \left\|f_2 - \hat{f}_2\right\|_{\overline{\mathcal{Z}}^n} \right)^2 \\
&\le \left(\left\|\hat{f}_1 - \hat{f}_2\right\|_{\overline{\mathcal{Z}}^n} + 2 \cdot \sqrt{\left|\overline{\mathcal{Z}}^n\right|} \cdot \sqrt{\delta / \left(64N^3\right)} \right)^2 \\
&\le \alpha/25
\end{aligned}
$$

Finally we could bound $\|f_1 - f_2\|^2_{\widehat{\mathcal{Z}}^n}$:

$$
\begin{aligned}
\|f_1 - f_2\|^2_{\widehat{\mathcal{Z}}^n} &\le \left(\|f_1 - f_2\|_{\overline{\mathcal{Z}}^n} + \sqrt{64N^3/\delta} / \left(8\sqrt{64N^3/\delta} \right) \right)^2 \\
&\le \alpha
\end{aligned}
$$

We conclude that for any pair $f_1, f_2 \in \mathcal{F}$ with $\|f_1 - f_2\|^2_{\mathcal{Z}^n} \le \alpha/100$, it holds that $\|f_1 - f_2\|^2_{\widehat{\mathcal{Z}}^n} \le \alpha$. Thus we must have $\underline{\mathcal{B}}^n(\alpha) \subseteq \mathcal{B}^n(\alpha)$.

Part 2 $\left(\mathcal{B}^n(\alpha) \subseteq \overline{\mathcal{B}}^n(\alpha)\right)$: Consider any pair $f_1, f_2 \in \mathcal{F}$ with $\|f_1 - f_2\|^2_{\mathcal{Z}^n} > 100\alpha$. From the definition we know that there exist $\left(\hat{f}_1, \hat{f}_2\right) \in \mathcal{C}\left(\mathcal{F}, \sqrt{\delta/\left(64N^3\right)}\right) \times \mathcal{C}\left(\mathcal{F}, \sqrt{\delta/\left(64N^3\right)}\right)$ such that $\left\|\hat{f}_1 - f_1\right\|_\infty, \left\|\hat{f}_2 - f_2\right\|_\infty \le \sqrt{\delta/\left(64N^3\right)}$. Then we have that

$$
\begin{aligned}
\left\|\hat{f}_1 - \hat{f}_2\right\|^2_{\mathcal{Z}^n} &\ge \left(\|f_1 - f_2\|_{\mathcal{Z}^n} - \left\|f_1 - \hat{f}_1\right\|_{\mathcal{Z}^n} - \left\|\hat{f}_2 - f_2\right\|_{\mathcal{Z}^n} \right)^2 \\
&\ge \left(\|f_1 - f_2\|_{\mathcal{Z}^n} - 2 \cdot \sqrt{|\mathcal{Z}^n|} \cdot \sqrt{\delta/\left(64N^3\right))} \right)^2 \\
&> 50\alpha
\end{aligned}
$$

Thus we have $\left\|\hat{f}_1 - \hat{f}_2\right\|^2_{\mathcal{Z}^n} > 50\alpha$. We consider the Y_i 's which correspond to $\hat{f}_1$ and $\hat{f}_2$. Here we want to prove that $\left\|\hat{f}_1 - \hat{f}_2\right\|^2_{\widehat{\mathcal{Z}}^n} > 40\alpha$. For the sake of contradicition we assume that $\left\|\hat{f}_1 - \hat{f}_2\right\|^2_{\widehat{\mathcal{Z}}^n} \le 40\alpha$.

Case 1 : $\left\|\hat{f}_1 - \hat{f}_2\right\|^2_{\mathcal{Z}^n} \le 2000000\alpha$. From the definition of Y_i we have

$$
\left\|\hat{f}_1 - \hat{f}_2\right\|^2_{\overline{\mathcal{Z}}^n} = \sum_{i=2}^{n} Y_i
$$

Combined with the former result, we conclude that

$$
\left\|\hat{f}_1 - \hat{f}_2\right\|^2_{\overline{\mathcal{Z}}^n} = \sum_{i=2}^{n} Y_i \ge \sum_{i=2}^{n} \mathbb{E}_{i-1}\left[Y_i\right] - \alpha/100 = \left\|\hat{f}_1 - \hat{f}_2\right\|^2_{\mathcal{Z}^n} - \alpha/100 > 50\alpha - \alpha/100 > 49\alpha
$$

Then we have

$$
\begin{aligned}
&\left\|\hat{f}_1 - \hat{f}_2\right\|^2_{\widehat{\mathcal{Z}}^n} \ge \left(\left\|\hat{f}_1 - \hat{f}_2\right\|_{\overline{\mathcal{Z}}^n} - \sqrt{64N^3/\delta} / \left(8\sqrt{64N^3/\delta} \right) \right)^2 \\
&> 40\alpha
\end{aligned}
$$

which leads to a contradiction.

Case 2 : $\left\|\hat{f}_1 - \hat{f}_2\right\|^2_{\mathcal{Z}^{n-1}} > 1000000\alpha$. From $\mathcal{E}^{n-1}$ we deduce that $\left\|\hat{f}_1 - \hat{f}_2\right\|^2_{\widehat{\mathcal{Z}}^{n-1}} > 100\alpha$ which directly leads to a contradiction.

Case 3 : $\left\|\hat{f}_1 - \hat{f}_2\right\|_{\mathcal{Z}^n}^2 > 2000000\alpha$ and $\left\|\hat{f}_1 - \hat{f}_2\right\|_{\mathcal{Z}^{n-1}}^2 \leq 1000000\alpha$. It is clear that $\left(\hat{f}_1\left(z_{n-1}\right) - \hat{f}_2\left(z_{n-1}\right)\right)^2 > 1000000\alpha$. From the definition of sensitivity we know that z_{n-1} will be added into $\overline{\mathcal{Z}}^n$ almost surely, which leads to a contradiction.

We conclude that $\left\|\hat{f}_1 - \hat{f}_2\right\|_{\widehat{\mathcal{Z}}^n}^2 > 40\alpha$. Finally we could bound $\|f_1 - f_2\|_{\widehat{\mathcal{Z}}^n}^2$:

$$\begin{aligned}
\|f_1 - f_2\|_{\widehat{\mathcal{Z}}^n}^2 &\geq \left(\left\|\hat{f}_1 - \hat{f}_2\right\|_{\widehat{\mathcal{Z}}^n} - \left\|f_1 - \hat{f}_1\right\|_{\widehat{\mathcal{Z}}^n} - \left\|\hat{f}_2 - f_2\right\|_{\widehat{\mathcal{Z}}^n}\right)^2 \\
&\geq \left(\left\|\hat{f}_1 - \hat{f}_2\right\|_{\widehat{\mathcal{Z}}^n} - 2 \cdot \sqrt{\left|\widehat{\mathcal{Z}}^n\right|} \cdot \sqrt{\delta / \left(64N^3\right)}\right)^2 \\
&> \alpha
\end{aligned}$$

We conclude that for any pair $f_1, f_2 \in \mathcal{F}$ with $\|f_1 - f_2\|_{\mathcal{Z}^n}^2 > 10000\alpha$, it holds that $\|f_1 - f_2\|_{\widehat{\mathcal{Z}}^n}^2 > 100\alpha$. This implies that $\mathcal{B}^n(\alpha) \subseteq \overline{\mathcal{B}}^n(\alpha)$. □

Next, we give a bound of the summation of online sensitivity scores.

Lemma C.5. *(Bound of sensitivity scores). We have*

$$\sum_{n=1}^{N-1} \text{sensitivity}_{\mathcal{Z}^n, \mathcal{F}}\left(z^n\right) \leq C \cdot \dim_E(\mathcal{F}, 1/4N) \log_2\left(4NW^2\right) \log N$$

for some absolute constant $C > 0$.

Proof. Note that $\mathcal{Z}^n = \{(s_i, a_i)\}_{i \in [n-1]}$, so $|\mathcal{Z}^n| \leq N$.

Notice that

$$\begin{aligned}
\text{sensitivity}_{\mathcal{Z}^n, \mathcal{F}}\left(z^n\right) &= \sup_{f_1, f_2 \in \mathcal{F}} \frac{\left(f_1(z) - f_2(z)\right)^2}{\min\{\|f_1 - f_2\|_{\mathcal{Z}^n}^2, 4NW^2\} + 1} \\
&= \sup_{f_1, f_2 \in \mathcal{F}} \frac{\left(f_1\left(z_n\right) - f_2\left(z_n\right)\right)^2}{\|f_1 - f_2\|_{\mathcal{Z}^n}^2 + 1}
\end{aligned}$$

For each $n \in [N-1]$, let $f_1, f_2 \in \mathcal{F}$ be an arbitrary pair of functions, such that

$$\frac{\left(f_1\left(z_n\right) - f_2\left(z_n\right)\right)^2}{\|f_1 - f_2\|_{\mathcal{Z}^n}^2 + 1}$$

is maximized, and we define $L\left(z_n\right) = \left(f_1\left(z_n\right) - f_2\left(z_n\right)\right)^2$ for such f_1, f_2.

Note that $0 \leq L\left(z_n\right) \leq 4W^2$. Let $\mathcal{Z}^N = \cup_{\alpha=0}^{\left\lfloor \log_2\left(4W^2N\right)\right\rfloor} \mathcal{Z}_\alpha \cup \mathcal{Z}_\infty$ be a dyadic decomposition with respect to $L(\cdot)$, where for each $0 \leq \alpha \leq \left\lfloor \log_2\left(4W^2N\right)\right\rfloor$, we define

$$\mathcal{Z}_\alpha = \left\{z_n \in \mathcal{Z}^N \mid L\left(z_n\right) \in \left(4W^2 \cdot 2^{-(\alpha+1)}, 4W^2 \cdot 2^{-\alpha}\right]\right\}$$

and

$$\mathcal{Z}_\infty = \left\{z_n \in \mathcal{Z}^N \mid L\left(z_n\right) \leq 4W^2 \cdot 2^{-\left\lfloor \log_2\left(4W^2N\right)\right\rfloor - 1}\right\}$$

Therefore, for any $z_n \in \mathcal{Z}_\infty$, $\text{sensitivity}_{\mathcal{Z}^n, \mathcal{F}}\left(z_n\right) \leq 4W^2 \cdot 2^{-\left\lfloor \log_2\left(4W^2N\right)\right\rfloor - 1} < 1/N$, and thus

$$\sum_{z_n \in \mathcal{Z}_\infty} \text{sensitivity}_{\mathcal{Z}^n,\mathcal{F}}\left(z_n\right) \le N \cdot \frac{1}{N} = 1$$

For each α, let $N_\alpha = \left|\mathcal{Z}_\alpha\right| / \dim_E\left(\mathcal{F}, 4W^2 \cdot 2^{-(\alpha+1)}\right)$, and we decompose $\mathcal{Z}_\alpha$ into $(N_\alpha + 1)$ disjoint subsets, i.e., $\mathcal{Z}_\alpha = \cup_{j=1}^{N_\alpha+1} \mathcal{Z}_\alpha^j$, by using the following procedure:

Initialize $Z_\alpha^j = \{\}$ for all j and consider each $z_n \in \mathcal{Z}_\alpha$ sequentially.

For each $z_n \in \mathcal{Z}_\alpha$, find the smallest $1 \le j \le N_\alpha$, such that z_n is $4W^2 \cdot 2^{-(\alpha+1)}$-independent of $\mathcal{Z}_\alpha^j$ with respect to $\mathcal{F}$.

Set $j = N_\alpha + 1$ if such j does not exist, use $j\left(z_n\right) \in \left[N_\alpha + 1\right]$ to denote the choice of j for z_n, and add z_n into Z_α^j.

Now, for each $z_n \in \mathcal{Z}_\alpha, z_n$ is $4W^2 \cdot 2^{-(\alpha+1)}$-dependent on each of $\mathcal{Z}_\alpha^1, \mathcal{Z}_\alpha^2, \cdots, \mathcal{Z}_\alpha^{j(z_n)-1}$.

Next, we will show that: For each $z_n \in \mathcal{Z}_\alpha$,

$$\text{sensitivity}_{\mathcal{Z}^n,\mathcal{F}}\left(z_n\right) \le \frac{4}{j\left(z_n\right)}$$

For any $z_n \in \mathcal{Z}_\alpha$, let $f_1, f_2 \in \mathcal{F}$ be an arbitrary pair of functions, such that

$$\frac{\left(f_1\left(z_n\right) - f_2\left(z_n\right)\right)^2}{\left\|f_1 - f_2\right\|_{\mathcal{Z}^n}^2 + 1}$$

is maximized. Since $z_n \in \mathcal{Z}_\alpha$, we must have $\left(f_1\left(z_n\right) - f_2\left(z_n\right)\right)^2 > 4W^2 \cdot 2^{-(\alpha+1)}$, since z_n is $4W^2 \cdot 2^{-(\alpha+1)}$ dependent on each of $\mathcal{Z}_\alpha^1, \mathcal{Z}_\alpha^2, \cdots, \mathcal{Z}_\alpha^{j(z_n)-1}$, for each $1 \le t < j\left(z_n\right)$, we have $\left\|f_1 - f_2\right\|_{\mathcal{Z}_\alpha^t}^2 \ge 4W^2.\ 2^{-(\alpha+1)}$, note that $\mathcal{Z}_\alpha^1, \mathcal{Z}_\alpha^2, \cdots, \mathcal{Z}_\alpha^{j(z_n)-1} \subset \mathcal{Z}^n$ due to the design of the partition procedure. Thus,

$$\begin{aligned}\text{sensitivity}_{\mathcal{Z}^n,\mathcal{F}}\left(z_n\right) &\le \frac{\left(f_1\left(z_n\right) - f_2\left(z_n\right)\right)^2}{\left\|f_1 - f_2\right\|_{\mathcal{Z}^n}^2 + 1} \le \frac{4W^2 \cdot 2^{-\alpha}}{\left\|f_1 - f_2\right\|_{\mathcal{Z}^n}^2} \le \frac{4W^2 \cdot 2^{-\alpha}}{\sum_{t=1}^{j(z_n)-1} \left\|f_1 - f_2\right\|_{\mathcal{Z}_\alpha^t}^2}, \\ &\le \frac{4W^2 \cdot 2^{-\alpha}}{\left(j\left(z_n\right) - 1\right) \cdot 4W^2 \cdot 2^{-(\alpha+1)}} \le \frac{2}{j\left(z_n\right) - 1}\end{aligned}$$

Therefore,

$$\text{sensitivity}_{\mathcal{Z}^n,\mathcal{F}}\left(z_n\right) \le \min\left\{\frac{2}{j\left(z_n\right) - 1}, 1\right\} \le \frac{4}{j\left(z_n\right)}$$

In addition, by the definition of $4W^2 \cdot 2^{-(\alpha+1)}$-independent, we have $\left|\mathcal{Z}_\alpha^j\right| \le \dim_E\left(\mathcal{F}, 4W^2 \cdot 2^{-(\alpha+1)}\right)$ for all $1 \le j \le N_\alpha$. Therefore,

$$\begin{aligned}\sum_{z_n \in \mathcal{Z}_\alpha} \text{sensitivity}\ _{\mathcal{Z}^n,\mathcal{F}}\left(z_n\right) &\le \sum_{1 \le j \le N_\alpha} \left|\mathcal{Z}_\alpha^j\right| \cdot \frac{4}{j} + \sum_{z \in \mathcal{Z}_\alpha^{N_\alpha+1}} \frac{4}{N_\alpha} \\ &\le 4 \dim_E\left(\mathcal{F}, 4W^2 \cdot 2^{-(\alpha+1)}\right) \cdot \left(\ln\left(N_\alpha\right) + 1\right) + \left|\mathcal{Z}_\alpha\right| \cdot \frac{4W^2 \cdot 2^{-(\alpha+1)}}{\left|\mathcal{Z}_\alpha\right|} \\ &= 4 \dim_E\left(\mathcal{F}, 4W^2 \cdot 2^{-(\alpha+1)}\right) \cdot \left(\ln\left(N_\alpha\right) + 2\right) \\ &\le 8 \dim_E\left(\mathcal{F}, 4W^2 \cdot 2^{-(\alpha+1)}\right) \cdot \ln N\end{aligned}$$

Now, by the monotonicity of eluder dimension, it follows that:

$$\begin{aligned}\sum_{n=1}^{N-1} \text{sensitivity }_{\mathcal{Z}^n,\mathcal{F}}(z_n) &\leq \sum_{\alpha=0}^{\lfloor \log_2(4W^2N) \rfloor} \sum_{z_n \in \mathcal{Z}_\alpha} \text{sensitivity }_{\mathcal{Z}^n,\mathcal{F}}(z_n) + \sum_{z_n \in \mathcal{Z}^\infty} \text{sensitivity }_{\mathcal{Z}^n,\mathcal{F}}(z_n) \\ &\leq 8\left(\left\lfloor \log_2\left(4W^2N\right)\right\rfloor + 1\right) \dim_E(\mathcal{F}, 1/4N) \ln N + 1 \\ &\leq 9\left(\left\lfloor \log_2\left(4W^2N\right)\right\rfloor + 1\right) \dim_E(\mathcal{F}, 1/4N) \ln N\end{aligned}$$

□

The following proposition verifies that $\bigcap_{n=1}^{\infty} \mathcal{E}^n$ happens with high probability.

Proposition C.6.

$$\mathbb{P}\left(\bigcap_{n=1}^{\infty} \mathcal{E}^n\right) \geq 1 - \delta/32$$

Proof. For all $n \in [N]$ it holds that

$$\begin{aligned}&\mathbb{P}\left(\mathcal{E}^1\mathcal{E}^2\ldots\mathcal{E}^{n-1}\right) - \mathbb{P}\left(\mathcal{E}^1\mathcal{E}^2\ldots\mathcal{E}^{n}\right) \\ =&\mathbb{P}\left(\mathcal{E}^1\mathcal{E}^2\ldots\mathcal{E}^{n-1}\left(\mathcal{E}^n\right)^c\right) \\ =&\mathbb{P}\left(\mathcal{E}^1\mathcal{E}^2\ldots\mathcal{E}^{n-1}\left(\bigcap_{j=0}^{\infty}\mathcal{E}^n\left(100^j\epsilon\right)\right)^c\right) \\ =&\mathbb{P}\left(\mathcal{E}^1\mathcal{E}^2\ldots\mathcal{E}^{n-1}\bigcup_{j=0}^{\infty}\left(\mathcal{E}^n\left(100^j\epsilon\right)\right)^c\right) \\ \leq&\sum_{j=0}^{\infty}\mathbb{P}\left(\mathcal{E}^1\mathcal{E}^2\ldots\mathcal{E}^{n-1}\left(\mathcal{E}^n\left(100^j\epsilon\right)\right)^c\right) \\ =&\sum_{j\geq 0, 100^j\epsilon \leq 4NW^2}\mathbb{P}\left(\mathcal{E}^1\mathcal{E}^2\ldots\mathcal{E}^{n-1}\left(\mathcal{E}^n\left(100^j\epsilon\right)\right)^c\right)\end{aligned}$$

where the last equality holds because $\mathbb{P}\left(\mathcal{E}^n(\alpha)\right) = 1$ while $\alpha > 4NW^2$. Combining this with Lemma C.4 yields

$$\mathbb{P}\left(\mathcal{E}^1\mathcal{E}^2\ldots\mathcal{E}^{n-1}\right) - \mathbb{P}\left(\mathcal{E}^1\mathcal{E}^2\ldots\mathcal{E}^{n}\right) \leq \delta/\left(32N^2\right)\cdot\left(\log\left(4NW^2/\epsilon\right)+2\right) \leq \delta/(32N)$$

thus

$$\begin{aligned}&\mathbb{P}\left(\bigcap_{n=1}^{N}\mathcal{E}^n\right) \\ =&1 - \sum_{n=1}^{N}\left(\mathbb{P}\left(\mathcal{E}^1\mathcal{E}^2\ldots\mathcal{E}^{n-1}\right) - \mathbb{P}\left(\mathcal{E}^1\mathcal{E}^2\ldots\mathcal{E}^{n}\right)\right) \\ \geq&1 - N\cdot(\delta/32N) \\ =&1-\delta/32\end{aligned}$$

□

With Lemma C.5, we are now ready to prove:

Proposition C.7. *With probability at least $1-\delta/8$, the following statements hold:*

(i) The subsampled dataset $\widehat{\mathcal{Z}}^n$ changes for at most

$$S_{\max} = C \cdot \log\left(N\mathcal{N}\left(\mathcal{F}, \sqrt{\delta/64N^3}\right)/\delta\right) \cdot \dim_E(\mathcal{F}, 1/N) \cdot \log^2 N$$

where $C > 0$ is some absolute constant.

(ii) For any $n \in [N]$, $\left|\widehat{\mathcal{Z}}^n\right| \leq 64N^3/\delta$.

Proof. Conditioning on $\mathcal{E}^n$, we have

$$\mathbb{I}\{\mathcal{E}^n\} \cdot \text{sensitivity}_{\widehat{\mathcal{Z}}^n,\mathcal{F}}(z_n) \leq C \cdot \text{sensitivity}_{\mathcal{Z}^n,\mathcal{F}}(z_n)$$

for some constant $C > 0$ according to Lemma C.3. By definition of p_z we have

$$p_z \lesssim \text{sensitivity}_{\widehat{\mathcal{Z}},\mathcal{F}}(z) \cdot \log\left(N\mathcal{N}\left(\mathcal{F}, \sqrt{\delta/64N^3}\right)/\delta\right)$$

thus by Lemma C.5 we have

$$\begin{aligned}
\sum_{n=1}^{N-1} \mathbb{I}\{\mathcal{E}^n\} \cdot p_{z_n} &\lesssim \sum_{n=1}^{N-1} \mathbb{I}\{\mathcal{E}^n\} \cdot \text{sensitivity}_{\widehat{\mathcal{Z}}^n,\mathcal{F}}(z_n) \cdot \log\left(N\mathcal{N}\left(\mathcal{F}, \sqrt{\delta/64N^3}\right)/\delta\right) \\
&\lesssim \sum_{n=1}^{N-1} \text{sensitivity}_{\mathcal{Z}^n,\mathcal{F}}(z_n) \cdot \log\left(N\mathcal{N}\left(\mathcal{F}, \sqrt{\delta/64N^3}\right)/\delta\right) \\
&\lesssim \log\left(N\mathcal{N}\left(\mathcal{F}, \sqrt{\delta/64N^3}\right)/\delta\right) \dim_E(\mathcal{F}, 1/N) \log^2 N
\end{aligned}$$

and by choosing C in the proposition appropriately, we may assume that

$$\sum_{n=1}^{N-1} \mathbb{I}\{\mathcal{E}^n\} \cdot p_{z_n} \leq S_{\max}/3$$

For $2 \leq n \leq N$, define random variables $\{X_n\}$ as

$$X_n = \begin{cases} \mathbb{I}\left\{\mathcal{E}^{n-1}\right\} & \hat{z}_{n-1} \text{ is added into } \widehat{\mathcal{Z}}^n \\ 0 & \text{otherwise} \end{cases}$$

Then X_n is adapted to the filtration F_n. We have that $\mathbb{E}_{n-1}[X_n] = p_{z_{n-1}} \cdot \mathbb{I}\left\{\mathcal{E}^{n-1}\right\}$ and $\mathbb{E}_{n-1}\left[(X_n - \mathbb{E}_{n-1}[X_n])^2\right] = \mathbb{I}\left\{\mathcal{E}^{n-1}\right\} \cdot p_{z_{n-1}}\left(1 - p_{z_{n-1}}\right)$. Note that $X_n - \mathbb{E}_{n-1}[X_n]$ is a martingale difference sequence with respect to F_n and

$$\begin{aligned}
\sum_{n=2}^{N} \mathbb{E}_{n-1}\left[(X_n - \mathbb{E}_{n-1}[X_n])^2\right] &= \sum_{n=2}^{N} \mathbb{I}\{\mathcal{E}^n\} p_{z_{n-1}}\left(1 - p_{z_{n-1}}\right) \leq \sum_{n=2}^{N} \mathbb{I}\{\mathcal{E}_{n-1}\} \cdot p_{z_{n-1}} \leq S_{\max}/3 \\
\sum_{n=2}^{N} \mathbb{E}_{n-1}[X_n] &= \sum_{n=2}^{n} p_{z_{n-1}} \mathbb{I}\{\mathcal{E}_{n-1}\} \leq S_{\max}/3
\end{aligned}$$

thus by applying Freedman's inequality (Lemma C.1), we deduce that

$$\begin{aligned}
&\mathbb{P}\left\{\sum_{n=2}^{N} X_n \geq S_{\max}\right\} \\
\leq &\mathbb{P}\left\{\left|\sum_{n=2}^{N} \left(X_n - \mathbb{E}_{n-1}\left[X_n\right]\right)\right| \geq 2S_{\max}/3\right\} \\
\leq &2\exp\left\{-\frac{\left(2S_{\max}/3\right)^2/2}{S_{\max}/3 + 2S_{\max}/9}\right\} \\
\leq &\delta/(32N)
\end{aligned}$$

With a union bound we know that with probability at least $1-\delta/32$,

$$\sum_{n=2}^{N} X_n < S_{\max}$$

We condition on the event above and $\bigcap_{n=1}^{N} \mathcal{E}^n$. In this case, we add elements into $\widehat{\mathcal{Z}}^n$ for at most $S_{\max}$ times. Combining the result above with Lemma C.2 completes the proof.

□

D. Concentration of Bonuses

Before bounding the bonuses, we need the following concentration inequality proved in (Beygelzimer et al., 2011).

Lemma D.1. *(Bernstein for Martingales).*

Consider a sequence of random variables $X_1, X_2, \cdots, X_T$. Assume that for all t, $X_t \leq R$, and $\mathbb{E}_t[X_t] \overset{def}{=} \mathbb{E}[X_t|X_1, \cdots, X_{t-1}] = 0$. Then for any $\delta > 0$, there exists constant c_1, c_2, such that

$$\mathbf{P}\left(\sum_{t=1}^{T} X_t \leq c_1 \times \sqrt{\sum_{t=1}^{T} \mathbb{E}_t[X_t^2] \ln\frac{1}{\delta}} + c_2 \times \ln\frac{1}{\delta}\right) \geq 1-\delta$$

Lemma D.2. *(Bound of Indicators). For any episode n during the execution of the algorithm, with probability $1-\delta/2$,*

$$\sum_{n=1}^{N} \mathbb{E}_{(s,a)\sim d^n} b_1^n(s,a) \leq \widetilde{O}\left(\frac{\sqrt{Nd^2\epsilon}}{(1-\gamma)\beta}\right) \tag{28}$$

where $d = dim_E(\mathcal{F}, 1/N)$.

Proof.

$$\begin{aligned}
\sum_{n=1}^{N} \mathbb{E}_{(s,a)\sim d^n} b_1^n(s,a) &\leq \frac{3}{1-\gamma}\sum_{n=1}^{N} \mathbb{E}_{(s,a)\sim d^n} \mathbf{1}\{\omega(\widehat{\mathcal{F}}^n, s, a) \geq \beta\} \\
&= \frac{3}{1-\gamma}\sum_{n=1}^{N} \mathbb{E}_{(s,a)\sim d^n} \mathbf{1}\{\frac{1}{\beta}\omega(\widehat{\mathcal{F}}^n, s, a) \geq 1\} \\
&\leq \frac{3}{1-\gamma}\cdot\frac{1}{\beta}\sum_{n=1}^{N} \mathbb{E}_{(s,a)\sim d^n} \omega(\widehat{\mathcal{F}}^n, s, a) \\
&\leq \widetilde{O}\left(\frac{\sqrt{Nd^2\epsilon}}{(1-\gamma)\beta}\right) \text{ (by Lemma D.3)}
\end{aligned} \tag{29}$$

□

Lemma D.3. *(Bound of Bonuses). For any episode n during the execution of the algorithm, with probability $1-\delta/2$*

$$\sum_{n=1}^{N}\mathbb{E}_{(s,a)\sim d^n}\omega(\widehat{\mathcal{F}}^n,s,a)\leq O\left(\sqrt{Nd^2\epsilon}+\ln(\frac{2}{\delta})\right)=\widetilde{O}\left(\sqrt{Nd^2\epsilon}\right) \tag{30}$$

where $d = dim_E(\mathcal{F},1/N)$.

Proof. We define the random dataset $\mathcal{D}_{1:n}$ to represent all the information at the beginning of iteration n of the algorithm. Then we define

$$\xi_n=\mathbb{E}_{(s,a)\sim d^n}[\omega(\widehat{\mathcal{F}}^n,s,a)|\mathcal{D}_{1:n}]-\omega(\widehat{\mathcal{F}}^n,s_n,a_n)$$

and let

$$A=\sum_{n=1}^{N}\mathbb{E}_{(s,a)\sim d^n}[\omega(\widehat{\mathcal{F}}^n,s,a)|\mathcal{D}_{1:n}]=\sum_{n=1}^{N}\omega(\widehat{\mathcal{F}}^n,s_n,a_n)+\sum_{n=1}^{N}\xi_n$$

Now we bound $\sum_{n=1}^{N}\omega(\widehat{\mathcal{F}}^n,s_n,a_n)$:
We condition on the event in the Proposition C.6, we have

$$\omega(\widehat{\mathcal{F}}^n,s,a)\leq\sup_{f_1,f_2\in\mathcal{F},||f_1-f_2||^2_{\mathcal{Z}^n}\leq 100\epsilon}|f_1(s,a)-f_2(s,a)|\overset{def}{=}\bar{b}^n(s,a)$$

For any given $\alpha>0$, let $\mathcal{L}=\{(s_n,a_n)|n\in[N],\bar{b}^n(s_n,a_n)>\alpha\}$, let $|\mathcal{L}|=L$.
Next we show that there exists $z_k:=(s_k,a_k)\in\mathcal{L}$, such that (s_k,a_k) is α-dependent on at least $N=L/\text{dim}_E(\mathcal{F},\alpha)-1$ disjoint subsequences in $\mathcal{Z}^k\cap\mathcal{L}$. We decompose the $\mathcal{L}=\cup_{j=1}^{N+1}\mathcal{L}^j$ by the following procedure. We initialize $\mathcal{L}^j=\{\}$ for all j and consider $z_k\in\mathcal{L}$ sequentially. For each $z_k\in\mathcal{L}$, we find the smallest j $(1\leq j\leq N)$, such that z_k is α-independent on $\mathcal{L}^j$ with respect to $\mathcal{F}$. We set $j=N+1$ if such j does not exist. We add z_k into $\mathcal{L}^j$ afterwards. When the decomposition of $\mathcal{L}$ is finished, $\mathcal{L}^{N+1}\neq\emptyset$, as $\mathcal{L}^j$ contains at most $\text{dim}_E(\mathcal{F},\alpha)$ elements for $j\in[N]$. For any $z_k\in\mathcal{L}^{N+1}$, z_k is α-dependent on at least $N=L/\text{dim}_E(\mathcal{F},\alpha)-1$ disjoint subsequences in $\mathcal{Z}^k\cap\mathcal{L}$.
On the other hand, there exists $f_1,f_2\in\mathcal{F}$ with $||f_1-f_2||^2_{\mathcal{Z}^k}\leq 100\epsilon$, such that $|f_1(s,a)-f_2(s,a)|>\alpha$. So we have:

$$(L/\text{dim}_E(\mathcal{F},\alpha)-1)\cdot\alpha^2\leq||f_1-f_2||^2_{\mathcal{Z}^k}\leq 100\epsilon$$

which implies

$$L\leq(\frac{100\epsilon}{\alpha^2}+1)\text{dim}_E(\mathcal{F},\alpha)$$

Now let $b_1\geq b_2\geq\cdots b_N$ to be a permulation of $\{\bar{b}^n(s_n,a_n)\}_{n=1}^N$. For any $b_n\geq\frac{1}{N}$, we have

$$n\leq(\frac{100\epsilon}{b_n^2}+1)\text{dim}_E(\mathcal{F},b_n)\leq(\frac{100\epsilon}{b_n^2}+1)\text{dim}_E(\mathcal{F},\frac{1}{N})$$

which implies that

$$b_n\leq\left(\frac{n}{\text{dim}_E(\mathcal{F},\frac{1}{N})}-1\right)^{-\frac{1}{2}}\sqrt{100\epsilon},\ \text{when } b_n\geq 1/N$$

Moreover, we have $b_n\leq 2W$, so

$$\begin{aligned}\sum_{n=1}^{N}b_n&\leq N\cdot\frac{1}{N}+2W\cdot\text{dim}_E(\mathcal{F},1/N)+\sum_{\text{dim}_E(\mathcal{F},1/N)<n\leq N}\left(\frac{n}{\text{dim}_E(\mathcal{F},\frac{1}{N})}-1\right)^{-\frac{1}{2}}\sqrt{100\epsilon}\\&\leq 1+2W\cdot\text{dim}_E(\mathcal{F},1/N)+C\cdot\sqrt{\text{dim}_E(\mathcal{F},1/N)\cdot N\cdot\epsilon}\end{aligned} \tag{31}$$

For simplicity, we denote $d:=\text{dim}_E(\mathcal{F},1/N)$, then $\sum_{n=1}^{N}\omega(\widehat{\mathcal{F}}^n,s_n,a_n)\leq O(\sqrt{Nd^2\epsilon})$.

Then we will bound the sum of noise terms:

$$\begin{aligned}\sum_{n=1}^{N} \mathbb{E}_{(s,a)\sim d^n}[\xi_n^2|\mathcal{D}_{1:n}] &= \sum_{n=1}^{N} \mathbb{E}_{(s,a)\sim d^n}[\omega^2(\widehat{\mathcal{F}}^n, s, a)|\mathcal{D}_{1:n}] \\ &\leq 2W \cdot \sum_{n=1}^{N} \mathbb{E}_{(s,a)\sim d^n}[\omega(\widehat{\mathcal{F}}^n, s, a)|\mathcal{D}_{1:n}]\end{aligned} \tag{32}$$

Now using the Lemma D.1 (Bernstein for Martingales) gives with probability at least $1-\frac{\delta}{2}$ for some constant c

$$\begin{aligned}\sum_{n=1}^{N} \xi_n &\leq c \times \left(\sqrt{2\sum_{n=1}^{N} \mathbb{E}_{(s,a)\sim d^n}[\xi_n^2|\mathcal{D}_{1:n}]\ln(2/\delta)} + \frac{\ln(2/\delta)}{3}\right) \\ &= c \times \left(\sqrt{4WA\ln(2/\delta)} + \frac{\ln(2/\delta)}{3}\right)\end{aligned} \tag{33}$$

Solving for A finally gives with high probability

$$A = O\left(\sqrt{Nd^2\epsilon} + \ln(\frac{2}{\delta})\right) \tag{34}$$

□

E. Analysis of Policy Evaluation Oracle

In this section, we provide the theoretical guarantee of our policy evaluation oracle using importance sampling technique.

Definition E.1. (Importance Sampling Estimator). Let t be a positive discrete random variable with probability mass function $\mathbf{P}(t=\tau) = \gamma^{\tau-1}(1-\gamma)$, and let $\{(s_\tau, a_\tau, r_\tau)\}_{\tau=1,\ldots,t}$ be a random trajectory of length t obtained by following a fixed "behavioral" policy $\underline{\pi}$ from (s,a). The importance sampling estimator of the target policy π is:

$$\left(\Pi_{\tau=2}^{t} \frac{\pi(s_\tau, a_\tau)}{\underline{\pi}(s_\tau, a_\tau)}\right) \frac{r_t}{1-\gamma}.$$

Notice that our inner loop solves the bonus-added MDP problem, so r_t is replaced by G_t in the following formula.

$$G_t = \begin{cases} \dfrac{1}{1-\gamma}[r_t + b(s_t, a_t)], & \text{if } t \geq 2 \\ \dfrac{1}{1-\gamma}[r_t], & \text{if } t = 1 \end{cases}$$

Definition E.2. We define $B = \frac{3}{1-\gamma}$, $G_{\max} = \frac{2+B}{(1-\gamma)}$, $\delta_1 = \gamma^\alpha$

Remark E.3. From the definition of bonus function, we know that $0 \leq b(\cdot,\cdot) \leq B$. In addition, the random return from a single Monte Carlo trajectory $\frac{G_t}{1-\gamma}$ has a deterministic upper bound $G_{\max}$. For a concise bound, we can assume $2G_{\max} \leq W$ in the following proof.

Lemma E.4. *(Stability of Importance Sampling Estimator) When*

$$k - \underline{k} \leq \kappa \overset{def}{=} \frac{(1-\gamma)\ln 2}{2\ln(1/\delta_1)\,\eta(B+W)},$$

then with probability $1-\delta_1$,

$$\left(\Pi_{\tau=2}^{t} \frac{\pi(s_\tau, a_\tau)}{\underline{\pi}(s_\tau, a_\tau)}\right) \frac{G_t}{1-\gamma} \leq 2G_{max}$$

Remark E.5. This lemma indicates that if we want to get a stable importance sampling estimator during policy evaluation process, κ should be $O(\sqrt{K})$ (since η has an order of $O(\frac{1}{\sqrt{K}})$ by Lemma B.10).

Proof. This lemma combines the results of Appendix G in (Zanette et al., 2021). First of all, we need to figure out the policy form on the known set. In fact, we have the following conclusion.

$$\forall(s,a), \quad \pi_k(a \mid s) = \pi_{\underline{k}}(a \mid s) \times \frac{e^{c(s,a)}}{\sum_{a'} \pi_{\underline{k}}\left(a' \mid s\right) e^{c(s,a')}}$$

where

$$c(s,a) = \eta \cdot \sum_{i=\underline{k}}^{k-1} [b(s,a) + f_i(s,a)]$$

We assume $k > \underline{k}$, then according to the algorithm,

$$\begin{aligned} \pi_k(\cdot|s) &\propto \pi_{k-1}(\cdot|s) e^{\eta[f_{k-1}(s,\cdot)+b(s,\cdot)]} \\ &\propto \pi_{\underline{k}}(\cdot|s) e^{\eta \sum_{i=\underline{k}}^{k-1}[f_i(s,\cdot)+b(s,\cdot)]} \end{aligned}$$

We define

$$c(s,a) = \eta \cdot \sum_{i=\underline{k}}^{k-1} [b(s,a) + f_i(s,a)]$$

So the desired result is obtained.
To simplify the notation, we use c to denote $\sup_{(s,a)} |c(s,a)|$. Then the following chain of inequalities is true.

$$e^{-2c} \leq \frac{e^{-c}}{\sum_{a'} \underline{\pi}\left(a' \mid s\right) e^{c}} \leq \frac{\pi(a \mid s)}{\underline{\pi}(a \mid s)} \leq \frac{e^{c(s,a)}}{\sum_{a'} \pi\left(a' \mid s\right) e^{c(s,a')}} \leq \frac{e^{c}}{\sum_{a'} \underline{\pi}\left(a' \mid s\right) e^{-c}} = e^{2c}$$

So we can bound the policy ratio.

$$e^{-2c} \leq \sup_{(s,a)} \frac{\pi(a \mid s)}{\pi(a \mid s)} \leq e^{2c}$$

Notice that

$$\eta \cdot \kappa \cdot (B + W) \geq \sup_{(s,a)} |c(s,a)|$$

Then we have

$$c = \sup_{(s,a)} |c(s,a)| \leq \frac{(1-\gamma)\ln 2}{2\ln(1/\delta_1)}$$

Remember that t is small with high probability:

$$\begin{aligned} \mathbf{P}(t > \alpha) &= \sum_{t=\alpha+1}^{\infty} \gamma^{\alpha-1}(1-\gamma) \\ &= \gamma^{\alpha} \sum_{t=0}^{\infty} \gamma^{\alpha}(1-\gamma) \\ &= \gamma^{\alpha} \overset{\text{def}}{=} \delta_1. \end{aligned}$$

This implies

$$\alpha = \frac{\ln \delta_1}{\ln \gamma} = \frac{\ln 1/\delta_1}{\ln 1/\gamma} \leq \frac{\ln 1/\delta_1}{1-\gamma}$$

In the complement of the above event:

$$\left(\sup_{(s,a)} \frac{\pi(a \mid s)}{\pi(a \mid s)}\right)^{t-1} \leq e^{2(\alpha-1)c} \leq e^{\frac{(\alpha-1)(1-\gamma)\ln 2}{\ln(1/\delta_1)}} \leq 2.$$

Then with probability at least $1-\delta_1$ if the importance sampling ratio is upper bounded

$$\Pi_{\tau=2}^{t} \frac{\pi\left(s_\tau, a_\tau\right)}{\underline{\pi}\left(s_\tau, a_\tau\right)} \leq\left(\sup _{(s, a)} \frac{\pi(a \mid s)}{\pi(a \mid s)}\right)^{t-1} \leq 2$$

And $\frac{G_t}{1-\gamma}$ is bounded by $G_{\max}$ in absolute value, which guarantees our result. □

Lemma E.6. *(Concentration of Width Function). If we set*

$$\frac{1}{100} \epsilon=\frac{3}{2} C_1 N \cdot \epsilon_{\text{stat}}+20 N W \epsilon_1+\frac{1}{2} C_2 \cdot \ln \left(\frac{N \mathcal{N}\left(\Delta \mathcal{F}, 2 \epsilon_1\right)}{\delta}\right) \tag{35}$$

where ϵ_1 denotes the function cover radius, C_1, C_2 is some constant defined in the following proof, ϵ_{stat} will be determined in Lemma E.7.

Then with probalility at least $1-\frac{1}{2}\delta$, for all $n \in [N]$

$$||\Delta f_k||_{\widehat{\mathcal{Z}}^n}^2 \leq \epsilon \tag{36}$$

Proof. Conditioned on the Proposition C.6, we only need to prove

$$||\Delta f_k||_{\mathcal{Z}^n}^2 \leq 100\epsilon \tag{37}$$

Let $\mathcal{C}\left(\Delta \mathcal{F}, 2 \epsilon_1\right)$ be a cover set of $\Delta \mathcal{F}$. Then for every $\Delta f \in \Delta \mathcal{F}$, there exists a $\Delta g \in \mathcal{C}\left(\Delta \mathcal{F}, 2 \epsilon_1\right)$ such that $\|\Delta f-\Delta g\|_{\infty} \leq 2 \epsilon_1$. Consider a fixed $\Delta g \in \mathcal{C}\left(\Delta \mathcal{F}, 2 \epsilon_1\right)$, we define n random variables:

$$X_i=\frac{1}{8 W^2}\left(\left(\Delta g\left(s_i, a_i\right)\right)^2-\mathbb{E}_{(s, a) \sim d^{\pi^i}}\left[(\Delta g(s, a))^2\right]\right), i \in[n]$$

Notice that for all $i \in [n]$, $X_i \leq 1$, $\mathbb{E}_i[X_i]=0$, and

$$\mathbb{E}_i[X_i^2] \leq \mathbb{E}_i[|X_i|] \leq \mathbb{E}_i\left[\frac{\left(\Delta g\left(s_i, a_i\right)\right)^2}{4 W^2}\right]=\frac{1}{4 W^2} \mathbb{E}_{(s, a) \sim d^{\pi^i}}(\Delta g(s, a))^2$$

Then by using Lemma D.1 (Bernstein for Martinglaes), we have the following inequality: With probalility at least $1-\delta_2$,

$$\sum_{i=1}^n X_i \leq c_1 \times \sqrt{\frac{\ln \frac{1}{\delta_2}}{4 W^2} \sum_{i=1}^n \mathbb{E}_{(s, a) \sim d^{\pi^i}}(\Delta g(s, a))^2}+c_2 \times \ln \frac{1}{\delta_2}$$

which means

$$\frac{1}{n} \sum_{i=1}^n\left[\left(\Delta g\left(s_i, a_i\right)\right)^2-\mathbb{E}_{(s, a) \sim d^{\pi^i}}(\Delta g(s, a))^2\right] \leq c \times\left(\sqrt{\frac{\ln \frac{1}{\delta_2}}{n^2} \sum_{i=1}^n \mathbb{E}_{(s, a) \sim d^{\pi^i}}(\Delta g(s, a))^2}+\frac{1}{n} \ln \frac{1}{\delta_2}\right)$$

We now proof that if $\lambda=C \cdot \ln \left(\frac{1}{\delta_2}\right)$, which C is a constant appropriate large, then

$$c \times\left(\sqrt{\frac{\ln \frac{1}{\delta_2}}{n^2} \sum_{i=1}^n \mathbb{E}_{(s, a) \sim d^{\pi^i}}(\Delta g(s, a))^2}+\frac{1}{n} \ln \frac{1}{\delta_2}\right) \leq \frac{1}{2}\left(\frac{1}{n}\left(\sum_{i=1}^n \mathbb{E}_{(s, a) \sim d^{\pi^i}}(\Delta g(s, a))^2\right)+\frac{\lambda}{n}\right)$$

To simplify the notation, we define $A=\sum_{i=1}^n \mathbb{E}_{(s, a) \sim d^{\pi^i}}(\Delta g(s, a))^2$.

Case 1: $A \leq \lambda$ According to the selection of λ, there exists constant c', c'' appropriate small, such that

$$\frac{1}{n}\ln\left(\frac{1}{\delta_2}\right) \leq c' \cdot \left(\frac{\lambda}{n}\right)$$

$$\sqrt{\frac{\lambda}{n^2}\ln\left(\frac{1}{\delta_2}\right)} \leq c'' \cdot \left(\frac{\lambda}{n}\right)$$

Then

$$\textbf{LHS} \leq c \times \left(\sqrt{\frac{\lambda}{n^2}\ln\left(\frac{1}{\delta_2}\right)} + \frac{1}{n}\ln\left(\frac{1}{\delta_2}\right)\right) \leq c \times (c' + c'')\left(\frac{\lambda}{n}\right) \leq \frac{1}{2}\left(\frac{\lambda}{n} + \frac{A}{n}\right) = \textbf{RHS}$$

Case 2: $A \geq \lambda$ there also exists constant c', c'' appropriate small, such that

$$\frac{1}{n}\ln\left(\frac{1}{\delta_2}\right) \leq c' \left(\frac{A}{n}\right)$$

$$\sqrt{\frac{A}{n^2}\ln\left(\frac{1}{\delta_2}\right)} \leq c'' \cdot \left(\frac{A}{n}\right)$$

Then

$$\textbf{LHS} \leq c \times \left(\sqrt{\frac{A}{n^2}\ln\left(\frac{1}{\delta_2}\right)} + \frac{1}{n}\ln\left(\frac{1}{\delta_2}\right)\right) \leq c \times (c' + c'')\left(\frac{A}{n}\right) \leq \frac{1}{2}\left(\frac{\lambda}{n} + \frac{A}{n}\right) = \textbf{RHS}$$

After taking the union bound on the function cover $\mathcal{C}(\Delta\mathcal{F}, 2\epsilon_1)$, we have the following result: With probalility at least $1 - N\mathcal{N}(\Delta\mathcal{F}, 2\epsilon_1)\delta_2 \overset{def}{=} 1 - \frac{1}{8}\delta$, by setting $\lambda = C \cdot \ln\left(\frac{N\mathcal{N}(\Delta\mathcal{F}, 2\epsilon_1)}{\delta}\right)$, we have

$$\forall n, \forall \Delta g \in \mathcal{C}(\Delta\mathcal{F}, 2\epsilon_1), \sum_{i=1}^{n}\left[(\Delta g(s_i, a_i))^2 - \mathbb{E}_{(s,a)\sim d^{\pi^i}}(\Delta g(s,a))^2\right] \leq \frac{1}{2}\left(\sum_{i=1}^{n}\mathbb{E}_{(s,a)\sim d^{\pi^i}}(\Delta g(s,a))^2 + \lambda\right)$$

Next, we transform to an arbitrary function $\Delta f \in \Delta\mathcal{F}$. Since there exists a $\Delta g \in \mathcal{C}(\Delta\mathcal{F}, 2\epsilon_1)$ such that $\|\Delta f - \Delta g\|_\infty \leq 2\epsilon_1$, we have that for all $i \in [n]$,

$$\begin{aligned} &\left|(\Delta f(s_i, a_i))^2 - (\Delta g(s_i, a_i))^2\right| \\ &= |\Delta f(s_i, a_i) - \Delta g(s_i, a_i)| \cdot |\ \Delta f(s_i, a_i) + \Delta g(s_i, a_i))\ | \leq 8W\epsilon_1 \end{aligned}$$

and

$$\begin{aligned} &\left|\mathbb{E}_{(s,a)\sim d^{\pi^i}}\left[(\Delta f(s,a))^2\right] - \mathbb{E}_{(s,a)\sim d^{\pi^i}}\left[(\Delta g(s,a))^2\right]\right| \\ &\leq \mathbb{E}_{(s,a)\sim d^{\pi^i}}|\Delta f(s,a) - \Delta g(s,a)| \cdot |\Delta f(s,a) + \Delta g(s,a)| \leq 8W\epsilon_1 \end{aligned}$$

Therefore,

$$
\begin{aligned}
&\sum_{i=1}^{n}\left[(\Delta f(s_i,a_i))^2-\mathbb{E}_{(s,a)\sim d^{\pi^i}}(\Delta f(s,a))^2\right]\\
\leq&\left|\sum_{i=1}^{n}\left[(\Delta f(s_i,a_i))^2-(\Delta g(s_i,a_i))^2\right]\right|+\left|\sum_{i=1}^{n}\left[(\Delta g(s_i,a_i))^2-\mathbb{E}_{(s,a)\sim d^{\pi^i}}(\Delta g(s,a))^2\right]\right|\\
+&\left|\sum_{i=1}^{n}\left[\mathbb{E}_{(s,a)\sim d^{\pi^i}}(\Delta f(s,a))^2-\mathbb{E}_{(s,a)\sim d^{\pi^i}}(\Delta g(s,a))^2\right]\right|\\
\leq&\frac{1}{2}\left(\sum_{i=1}^{n}\mathbb{E}_{(s,a)\sim d^{\pi^i}}(\Delta g(s,a))^2+\lambda\right)+16nW\epsilon_1\\
\leq&\frac{1}{2}\left(\sum_{i=1}^{n}\mathbb{E}_{(s,a)\sim d^{\pi^i}}(\Delta f(s,a))^2+8nW\epsilon_1+\lambda\right)+16nW\epsilon_1
\end{aligned}
$$

Then,

$$
\forall n\in[N],\forall\Delta f\in\Delta\mathcal{F},||\Delta f||_{\mathcal{Z}^n}^2\leq\frac{3}{2}\sum_{i=1}^{n}\mathbb{E}_{(s,a)\sim d^{\pi^i}}(\Delta f(s,a))^2+\frac{1}{2}\lambda+20nW\epsilon_1
$$

Then we have with probability at least $1-\frac{1}{8}\delta$,

$$
\|\Delta f_k\|_{\mathcal{Z}^n}^2\leq\frac{3}{2}n\cdot\mathbb{E}_{\rho_{\text{cov}}^n}\left[(\Delta f_k)^2\right]+20nW\epsilon_1+\frac{1}{2}\lambda,\ \ \forall n\in[N]
$$

By Assumption 6,

$$
\begin{aligned}
\mathbb{E}_{\rho_{\text{cov}}^n}\left[(\Delta f_k)^2\right]&=\mathbb{E}_{(s,a)\sim\rho_{\text{cov}}^n}\left[(f_k^*(s,a)-f_k(s,a))^2\right]\\
&\leq C\cdot\left(L\left(f_k;\rho_{\text{cov}}^n,Q_{b^n}^k-b^n\right)-L\left(f_k^*;\rho_{\text{cov}}^n,Q_{b^n}^k-b^n\right)\right)\\
&\leq C\cdot\epsilon_{\text{stat}}\ \ \text{(by Lemma E.7)}
\end{aligned}
$$

By the choice of ϵ, $\|\Delta f_t\|_{\mathcal{Z}^n}^2\leq 100\epsilon,\ \forall n\in[N]$ with probability at least $1-\frac{1}{4}\delta$. Combining the above result with Proposition C.7, we finish our proof of Lemma E.6.

Next, we give an explicit form of ϵ_{stat} as defined in the next lemma. □

Lemma E.7. *(Concentration of statistical error). Following the same notation as in Lemma E.6, it holds with probability at least $1-\frac{1}{8}\delta$ that*

$$
L\left(f_k;\rho_{cov}^n,Q_{b^n}^k-b^n\right)-L\left(f_k^*;\rho_{cov}^n,Q_{b^n}^k-b^n\right)\leq\frac{500C\cdot W^4\cdot\log\left(\frac{\mathcal{N}(\mathcal{F},\epsilon_2)}{\delta_3}\right)}{M}+13W^2\cdot\epsilon_2,
$$

where C,ϵ_0 are defined in Assumption B.6, and $\epsilon_2>0$ denotes the function cover radius which will be determined later.

Proof. This proof builds on Feng et al. (2021)'s Lemma C.4, but deals with the concentration of importance sampling estimator. First note that in the loss function, the expectation has a nested structure: the outer expectation is taken over $(s,a)\sim\rho_{\text{cov}}^n$ and the inner conditional expectation is $Q_{b^n}^k(s,a)=\mathbb{E}^{\pi_k}\left[\sum_{h=0}^{\infty}\gamma^h\left(r\left(s_h,a_h\right)+b^n\left(s_h,a_h\right)\right)\mid\left(s_0,a_0\right)=(s,a)\right]$ given a sample of $(s,a)\sim\rho_{\text{cov}}^n$. To simplify the notation, we use x to denote $(s,a),y\mid x$ for an unbiased sample of $Q_{b^n}^k(s,a)-b^n(s,a)$ through importance sampling, and ν for ρ_{cov}^n, the marginal distribution over x, then the loss function can be recast as

$$\begin{aligned}
&\mathbb{E}_{x\sim\nu}\left[\left(f_k(x)-\mathbb{E}[y\mid x]\right)^2\right]:=L\left(f_k;\rho_{\text{cov}}^n,Q_{b^n}^k-b^n\right)\\
&\mathbb{E}_{x\sim\nu}\left[\left(f_k^*(x)-\mathbb{E}[y\mid x]\right)^2\right]:=L\left(f_k^*;\rho_{\text{cov}}^n,Q_{b^n}^k-b^n\right)
\end{aligned}$$

In particular, f_k can be rewritten as

$$f_k\in\underset{f\in\mathcal{F}}{\operatorname{argmin}}\sum_{i=1}^{M}\left(f\left(x_i\right)-y_i\right)^2,$$

where (x_i,y_i) are drawn i.i.d.: x_i is generated following the marginal distribution ν and y_i is generated conditioned on x_i.

Note that y_i is collected by importance sampling estimator, which does not necessarily come from Monte Carlo sampling. However, in the latest time when the agent interacts with the environment, the samples are drawn i.i.d., which guaranteed the same property for the importance sampling process.

For any function f, we have:

$$\begin{aligned}
&\mathbb{E}_{x,y}\left[\left(f_k(x)-y\right)^2\right]\\
=&\mathbb{E}_{x,y}\left[\left(f_k(x)-\mathbb{E}[y\mid x]\right)^2\right]+\mathbb{E}_{x,y}\left[\left(\mathbb{E}[y\mid x]-y\right)^2\right]+2\mathbb{E}_{x,y}\left[\left(f_k(x)-\mathbb{E}[y\mid x]\right)\left(\mathbb{E}[y\mid x]-y\right)\right]\\
=&\mathbb{E}_{x,y}\left[\left(f_k(x)-\mathbb{E}[y\mid x]\right)^2\right]+\mathbb{E}_{x,y}\left[\left(\mathbb{E}[y\mid x]-y\right)^2\right],
\end{aligned}$$

where the last step follows from the cross term being zero. Thus we can rewrite the generalization error as

$$\begin{aligned}
&\mathbb{E}_{x}\left[\left(f_k(x)-\mathbb{E}[y\mid x]\right)^2\right]-\mathbb{E}_{x}\left[\left(f_k^*(x)-\mathbb{E}[y\mid x]\right)^2\right]\\
=&\mathbb{E}_{x,y}\left(f_k(x)-y\right)^2-\mathbb{E}_{x,y}\left(f_k^*(x)-y\right)^2.
\end{aligned}$$

Next, we establish a concentration bound on f_k. Since f_k depends on the training set $\{(x_i,y_i)\}_{i=1}^M$, as discussed in Lemma B.9, we use a function cover on $\mathcal{F}$ for a uniform convergence argument. We denote by F_k^n the σ-algebra generated by randomness before epoch n iteration k. Recall that $f_k^*\in\operatorname{argmin}_{f\in\mathcal{F}}L\left(f;\rho_{\text{cov}}^n,Q_{b^n}^k-b^n\right)$. Conditioning on $F_k^n,\rho_{\text{cov}}^n,Q_{b^n}^k-b^n$, and f_k^* are all deterministic. For any $f\in\mathcal{F}$, we define

$$Z_i(f):=\left(f\left(x_i\right)-y_i\right)^2-\left(f_k^*\left(x_i\right)-y_i\right)^2,\quad i\in[M]$$

Then $Z_1(f),\ldots,Z_M(f)$ are i.i.d. random variables and notice that y_i is drawn from importance sampling estimator. From Lemma E.4, we know that with probability at least $1-M\delta_1$, $y_i\leq 2G_{\max}\leq W,\ i\in[M]$.

Conditioned on this event, we have

$$\begin{aligned}
\mathbb{V}\left[Z_k(f)\mid F_k^n\right]&\leq\mathbb{E}\left[Z_i(f)^2\mid F_k^n\right]\\
&=\mathbb{E}\left[\left(\left(f\left(x_i\right)-y_i\right)^2-\left(f_k^*\left(x_i\right)-y_i\right)^2\right)^2\mid F_k^n\right]\\
&=\mathbb{E}\left[\left(f\left(x_i\right)-f_k^*\left(x_i\right)\right)^2\cdot\left(f\left(x_i\right)+f_k^*\left(x_i\right)-2y_i\right)^2\mid F_k^n\right]\\
&\leq 36W^4\cdot\mathbb{E}\left[\left(f\left(x_i\right)-f_k^*\left(x_i\right)\right)^2\mid F_k^n\right]\\
&\leq 36W^4\cdot\left(C\cdot\mathbb{E}\left[Z_i(f)\mid F_k^n\right]\right)
\end{aligned}$$

where the last inequality is by Assumption B.6. Next, we apply Bernstein's inequality on the function cover $\mathcal{C}(\mathcal{F}, \epsilon_2)$ and take the union bound. Specifically, with probability at least $1-\delta_3$, for all $g \in \mathcal{C}(\mathcal{F}, \epsilon_2)$,

$$\begin{aligned}
&\mathbb{E}\left[Z_i(g) \mid F_k^n\right] - \frac{1}{M}\sum_{i=1}^{M} Z_i(g) \\
&\leq \sqrt{\frac{2\mathbb{V}\left[Z_i(g) \mid F_k^n\right] \cdot \log \frac{\mathcal{N}(\mathcal{F},,\epsilon_2)}{\delta_3}}{M}} + \frac{12W^4 \cdot \log \frac{\mathcal{N}(\mathcal{F},,\epsilon_2)}{\delta_3}}{M} \\
&\leq \sqrt{\frac{72W^4\left(C \cdot \mathbb{E}\left[Z_i(g) \mid F_k^n\right]\right) \cdot \log \frac{\mathcal{N}(\mathcal{F},\epsilon_2)}{\delta_3}}{M}} + \frac{12W^4 \cdot \log \frac{\mathcal{N}(\mathcal{F},\epsilon_2)}{\delta_3}}{M}.
\end{aligned}$$

For f_t, there exists $g \in \mathcal{C}(\mathcal{F}, \epsilon_2)$ such that $\|f_k - g\|_\infty \leq \epsilon_2$ and

$$\begin{aligned}
\left|Z_i\left(f_k\right) - Z_i(g)\right| &= \left|\left(f_k\left(x_i\right) - y_i\right)^2 - \left(g\left(x_i\right) - y_i\right)^2\right| \\
&= \left|f_k\left(x_i\right) - g\left(x_i\right)\right| \cdot \left|f_k\left(x_i\right) + g\left(x_i\right) - 2y_i\right| \leq 6W^2\epsilon_2.
\end{aligned}$$

Therefore, with probability at least $1-\delta_3$,

$$\begin{aligned}
&\mathbb{E}\left[Z_i\left(f_k\right) \mid F_k^n\right] - \frac{1}{M}\sum_{i=1}^{M} Z_i\left(f_k\right) \\
&\leq \mathbb{E}\left[Z_i(g) \mid F_k^n\right] - \frac{1}{M}\sum_{i=1}^{M} Z_i(g) + 12W^2\epsilon_2 \\
&\leq \sqrt{\frac{72W^4\left(C \cdot \mathbb{E}\left[Z_i(g) \mid F_k^n\right]\right) \log \frac{\mathcal{N}(\mathcal{F},\epsilon_2)}{\delta_3}}{M}} + \frac{12W^4 \log \frac{\mathcal{N}(\mathcal{F},\epsilon_2)}{\delta_3}}{M} + 12W^2\epsilon_2 \\
&\leq \sqrt{\frac{72W^4\left(C \cdot \mathbb{E}\left[Z_i(f_k) \mid F_k^n\right] + 6CW^2\epsilon_2\right) \log \frac{\mathcal{N}(\mathcal{F},\epsilon_2)}{\delta_3}}{M}} + \frac{12W^4 \log \frac{\mathcal{N}(\mathcal{F},\epsilon_2)}{\delta_3}}{M} + 12W^2\epsilon_2.
\end{aligned}$$

Since f_k is an empirical minimizer, we have $\frac{1}{M}\sum_{i=1}^{M} Z_i\left(f_k\right) \leq 0$. Thus,

$$\mathbb{E}\left[Z_i\left(f_k\right) \mid F_k^n\right] \leq \sqrt{\frac{72W^4\left(C \cdot \mathbb{E}\left[Z_i(f_k) \mid F_k^n\right] + 6CW^2\epsilon_2\right) \log \frac{\mathcal{N}(\mathcal{F},\epsilon_2)}{\delta_3}}{M}} + \frac{12W^4 \log \frac{\mathcal{N}(\mathcal{F},\epsilon_2)}{\delta_3}}{M} + 12W^2\epsilon_2.$$

Solving the above inequality with quadratic formula and using $\sqrt{a+b} \leq \sqrt{a} + \sqrt{b}$, $\sqrt{ab} \leq a/2 + b/2$ for $a > 0, b > 0$, we obtain

$$\mathbb{E}\left[Z_i\left(f_k\right) \mid F_k^n\right] \leq \frac{500C \cdot W^4 \cdot \log \frac{\mathcal{N}(\mathcal{F},\epsilon_2)}{\delta_3}}{M} + 13W^2 \cdot \epsilon_2$$

Since the right-hand side is a constant, through taking another expectation, we have

$$\mathbb{E}\left[Z_i\left(f_k\right)\right] \leq \frac{500C \cdot W^4 \cdot \log \frac{\mathcal{N}(\mathcal{F},\epsilon_2)}{\delta_3}}{M} + 13W^2 \cdot \epsilon_2.$$

Notice that $\mathbb{E}\left[Z_i\left(f_k\right)\right] = L\left(f_k; \rho_{\text{cov}}^n, Q_{b^n}^k - b^n\right) - L\left(f_k^*; \rho_{\text{cov}}^n, Q_{b^n}^k - b^n\right)$.

Finally, we let $(1 - M\delta_1)(1-\delta_3) \geq 1 - \frac{1}{8}\delta$, so the desired result is obtained. □

Lemma E.8. *(One-sided error). With probability at least* $1 - \frac{\delta}{2}$ *it holds that*

$$\forall n \in [N],\ \forall k \in \{0, \cdots, K-1\},\ \forall (s,a) \in \mathcal{K}^n:\ 0 \leq Q_{b^n}^{k,*}(s,a) - \widehat{Q}_{b^n}^k(s,a) \leq 2b_\omega^n(s,a) \tag{38}$$

Proof. When $(s,a) \in \mathcal{K}^n$,

$$\widehat{Q}^k_{b^n}(s,a) = f_k(s,a) + b^n_\omega(s,a)$$

$$Q^{k,*}_{b^n}(s,a) = f^*_k(s,a) + b^n(s,a) = f^*_k(s,a) + 2b^n_\omega(s,a)$$

Then,

$$|Q^{k,*}_{b^n}(s,a) - \widehat{Q}^k_{b^n}(s,a) - b^n_\omega(s,a)| = |f^*_k(s,a) - f_k(s,a)| = |\Delta f_k(s,a)|$$

According to Lemma E.6, with probability at least $1 - \frac{1}{2}\delta$, $||\Delta f_k||^2_{\widehat{\mathcal{Z}}^n} \leq \epsilon$, $\forall n \in [N]$
Using the definition of $b^n_\omega(s,a)$, we have

$$|\Delta f_k(s,a)| \leq \omega(\widehat{\mathcal{F}}^n, s, a) \leq b^n_\omega(s,a) \ (\beta < 1)$$

Finally, Lemma E.8 concludes. □

F. Limitation of Previous Implementations

Note that we do not compare our method directly with implementations in (Agarwal et al., 2020a; Feng et al., 2021), as we discovered some limitations presented in their implementations. We show our insights in this section and provide an empirical evaluation of the quality of implementations of our algorithm and previous ones.

Observation normalization is also very crucial for on-policy algorithms, but it is missing in those implementations. For the MountainCar environment, we find that the difficulty is not from the exploration problem, but from the ill-shaped observation. In their experiments, **PPO**-based exploration algorithms take up to 10k episodes to learn a near-optimal policy in MountainCar environment, however, with a running mean-std observation normalization, it only takes PPO-based algorithms a few episodes to learn the task.

Furthermore, both of their implementations strictly follow the theoretical algorithms and use a “Roll-In” mechanism in order to get the previous distribution ρ. Although a recent study (Li et al., 2022) shows evidence of leveraging the “Roll-In” mechanism in single-task RL problems for the off-policy algorithms, it still remains unknown whether such mechanism benefits on-policy algorithms in single-task RL problems. In our experiment, we find that **PC-PG** or **ENIAC** with “Roll-In” does not provide efficiency compared to its counterpart variant. We hypothesize that it is because the stochasticity of **PPO** and the environment is enough for the policy itself to recover the state distribution, thus the additionally introduced “Roll-In” is not needed.

Additionally, experiments from previous works (Agarwal et al., 2020a; Feng et al., 2021) compared exploration capability with **RND**, the current state-of-the-art algorithm on Montezuma’s Revenge (Bellemare et al., 2013; Burda et al., 2018). However, we find there is some discrepancy between their implementation and the original implementation of **RND**. Most importantly, their implementation does not use next state s' to determine the intrinsic reward of state action pair (s,a). The reason why this is crucial is that using s' to determine the intrinsic reward integrates the novelty of the (s,a) while using s will lose the information of the action.

To demonstrate our point, we tested the original implementation of (Agarwal et al., 2020a; Feng et al., 2021) on MountainCarContinuous with running observation normalization (for all running algorithms). With observation normalization, our implemented algorithms easily learn the task within 10000 steps, significantly better than results reported in (Agarwal et al., 2020a; Feng et al., 2021). Moreover, we also test their implementations along with observation normalization. The performance of their implementations does not improve much over the course of 10000 steps, which demonstrates our point that their “Roll-In” mechanism may not provide efficiency.

Our implementations (Raffin et al., 2021), including **RND** and **PPO**, succeed to find rewards in the environments, while implementations from previous works do not. The result is shown in Figure 2.

G. Hyperparameters

We implemented our method based on the open source package (Raffin et al., 2021), and the performance of **PPO** is obtained by running the built-in implemented **PPO**. Following (Burda et al., 2018), we use smaller batch size (compared to 64 in standard MuJoCo environment (Schulman et al., 2017)), specifically 32 in SparseHopper and 16 in SparseWalker2d and SparseHalfCheetah. The detailed hyperparameters are showed in the table G.

Hyperparameter	**Value** (**LPO**, **ENIAC**)	**Value** (**PPO**)
N	2048	2048
T	2e6	2e6
λ	0.95	0.95
$\gamma^{(int)}$	0.999	-
$\gamma^{(ext)}$	0.99	0.99
α	2	-
β	1	-
Learning rate	1e-4	1e-4
Batch size	32, 16	32, 16
Number of epoch per iteration	10	10