Algorithm 3 Value Estimators

```
Routine: $V^\pi$-ESTIMATOR
    Input: starting state $s$.
    Execute $\pi$ from $s$; at any step $t$ with $(s_t, a_t)$, terminate with probability $1 - \gamma$.
    Return: $\hat{V}^\pi(s) = \sum_{i=0}^{t} r(s_i, a_i)$, where $s_0 = s$.
Routine: $Q^\pi$-ESTIMATOR
    Input: starting state-action $(s, a)$.
    Execute $\pi$ from $(s, a)$; at any step $t$ with $(s_t, a_t)$, terminate with probability $1 - \gamma$.
    Return: $\hat{Q}^\pi(s, a) = \sum_{i=0}^{t} r(s_i, a_i)$, where $(s_0, a_0) = (s, a)$.
```

Algorithm 4 d^π Sampler

```
Routine: $d_\nu^\pi$-SAMPLER
    Input: $\nu \in \Delta(\mathcal{S} \times \mathcal{A}), \pi$.
    Sample $s_0, a_0 \sim \nu$;
    Execute $\pi$ from $s_0, a_0$; at any step $t$ with $(s_t, a_t)$, terminate with probability $1 - \gamma$.
    Return: $(s_t, a_t)$.
```

A. Omitted pseudocodes from main text

We give the pseudocodes for value estimators and visitation distribution sampler in Algorithms 3 and 4 respectively. Combining them, we are able to generate samples for critic fit.

B. Proof Setup

B.1. Definition and Notation

We denote by $\mathcal{M}$ the original MDP and $\tilde{\pi}$ an arbitrary fixed comparator policy (e.g., an optimal policy). Our target is to show that after N epochs, ENIAC is able to output a policy whose value is larger than $V^{\tilde{\pi}}$ minus some problem-dependent constant. First we describe the construction of some auxiliary MDPs, which is conceptually similar to Agarwal et al. (2020a), modulo the difference in the bonus functions.

For each epoch $n \in [N]$, we consider three MDPs: the original MDP $\mathcal{M}$, the bonus-added MDP $\mathcal{M}_{b^n} := (\mathcal{S}, \mathcal{A}, P, r + b^n, \gamma)$, and an auxiliary MDP $\mathcal{M}^n$. $\mathcal{M}^n$ is defined as $(\mathcal{S}, \mathcal{A} \cup \{a^\dagger\}, P^n, r^n, \gamma)$, where $a^\dagger$ is an extra action which is only available for $s \notin \mathcal{K}^n$ (recall that $s \in \mathcal{K}^n$ if and only if $b^n(s, a) \equiv 0$ for all $a \in \mathcal{A}$). For all $(s, a) \in \mathcal{S} \times \mathcal{A}$,

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

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

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

Basically, $a^\dagger$ allows the agent to stay in a state $s \notin \mathcal{K}^n$ while accumulating maximum instant rewards.

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

Additional Notations Given a policy π, we denote by $V_{b^n}^\pi$, $Q_{b^n}^\pi$, and $A_{b^n}^\pi$ the state-value, Q-value, and advantage function of π on $\mathcal{M}_{b^n}$ and $V_{\mathcal{M}^n}^\pi$, $Q_{\mathcal{M}^n}^\pi$, and $A_{\mathcal{M}^n}^\pi$ for the counterparts on $\mathcal{M}^n$. For the policy π_t^n, i.e., the policy at the t_{th} iteration in the n_{th} epoch of ENIAC, we further simplify the notation as $V_{b^n}^t$, $Q_{b^n}^t$, and $A_{b^n}^t$ and also $V_{\mathcal{M}^n}^t$, $Q_{\mathcal{M}^n}^t$, and $A_{\mathcal{M}^n}^t$.

Remark 2. *Note that only $\tilde{\pi}^n$ can take the action $a^\dagger$ for $s \notin \mathcal{K}^n$. All policies $\{\pi_t^n\}$ is not aware of $a^\dagger$ and therefore, $V_{b^n}^t = V_{\mathcal{M}^n}^t$, $Q_{b^n}^t = Q_{\mathcal{M}^n}^t$, and $A_{b^n}^t = A_{\mathcal{M}^n}^t$.*

Based on the above definitions, we directly have the following two lemmas.

Lemma B.1. *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. The proof follows that of Lemma B.1. in (Agarwal et al., 2020a). We present below for the readers' convenience.

We prove by induction over the time steps along the horizon. Recall $\tilde{d}_{\mathcal{M}^n}$ is the state-action distribution of $\tilde{\pi}^n$ over $\mathcal{M}^n$ and $d^{\tilde{\pi}}$ is the state-action distribution of $\tilde{\pi}$ on both $\mathcal{M}_{b^n}$ and $\mathcal{M}$ as they share the same dynamics. We use another subscript h to indicate the step index, e.g., $\tilde{d}_{\mathcal{M}^n,h}$ is the state-action distribution at the h_{th} step following $\tilde{\pi}^n$ on $\mathcal{M}^n$.

Starting at $h = 0$, if $s_0 \in \mathcal{K}^n$, then $\tilde{\pi}^n(\cdot|s_0) = \tilde{\pi}(\cdot|s_0)$ and we can easily get:

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

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

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

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

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

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

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

Using the fact that $\tilde{\pi}^n(\cdot|s) = \tilde{\pi}(\cdot|s)$ 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. *For any epoch $n \in [N]$, we have*

$$V_{\mathcal{M}^n}^{\tilde{\pi}^n} \ge V_{\mathcal{M}}^{\tilde{\pi}}.$$

Proof. The result is straightforward since if following $\tilde{\pi}^n$ we run into some $s \notin \mathcal{K}^n$, then by definition, $\tilde{\pi}^n$ is able to collect maximum instant rewards for all steps later. ■

B.2. Proof Sketch

We intend to compare the values of the output policy $\pi_{\text{ave}}^N := \text{Unif}(\pi^2, \pi^3, \ldots, \pi^{N+1})$ and the comparator $\tilde{\pi}$. To achieve this, we use two intermediate quantities $V_{b^n}^{\pi^{n+1}}$ and $V_{\mathcal{M}^n}^{\tilde{\pi}^n}$ and build the following inequalities as bridges:

$$V^{\pi_{\text{ave}}^N} = \frac{1}{N}\sum_{n=1}^{N} V^{\pi^{n+1}} \ge \frac{1}{N}\sum_{n=1}^{N} V_{b^n}^{\pi^{n+1}} - A, \quad V_{b^n}^{\pi^{n+1}} = V_{\mathcal{M}^n}^{\pi^{n+1}} \ge V_{\mathcal{M}^n}^{\tilde{\pi}^n} - B, \quad V_{\mathcal{M}^n}^{\tilde{\pi}^n} \ge V^{\tilde{\pi}},$$

where A and B are two terms to be specified. If the above relations all hold, the desired result is natually induced. For these inequalities, we observe that

1. The leftmost inequality is about the value differences of a sequence of policies $(\pi^2, \pi^3, \ldots, \pi^{N+1})$ on two different reward functions (with or without the bonus). Thus, it is bounded by the cumulative bonus, or equivalently, the *expected bonus* over the state-action measure induced by these policies, which we use the eluder dimension of the approximation function class to bound. We present this result for SPI-Sample, SPI-Compute, and NPG-Sample in Lemma C.1, C.5, and D.2, respectively.

2. The rightmost inequality is proved in Lemma B.2.
3. To show the middle inequality, we analyze the convergence of actor-critic updates, leveraging properties of the multiplicative weight updates for a regret bound following the analysis of Agarwal et al. (2020c).

In the sequel, we present sample complexity analysis for ENIAC-SPI-SAMPLE, ENIAC-SPI-COMPUTE, and ENIAC-NPG-SAMPLE. ENIAC-NPG-COMPUTE can be easily adapted with minor changes of the assumptions. In particular, we provide general results considering model misspecification and the theorems in the main body fall as special cases under Assumption 4.1 or 4.4.

C. Analysis of ENIAC-SPI

In this section, we provide analysis for ENIAC-SPI-SAMPLE and ENIAC-SPI-COMPUTE. We start with stating the assumptions which quantifies model misspecification.

Assumption C.1 (Bounded 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:*

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

For the fixed comparator policy $\tilde{\pi}$ (defined at the beginning of Section B.1), we define $\tilde{d}(s,a) := d_{s_0}^{\tilde{\pi}}(s) \circ \mathrm{Unif}(\mathcal{A})$. In ENIAC-SPI (both sample and compute versions), for every epoch $n \in [N]$ and every iteration t inside epoch n, we assume that

$$\inf_{f \in \mathcal{F}_t^n} L(f; \tilde{d}, Q_{b^n}^t - b^n) \leq \epsilon_{bias},$$

where $\mathcal{F}_t^n := \mathrm{argmin}_{f\in\mathcal{F}} L(f; \rho_{cov}^n, Q_{b^n}^t - b^n)$ and $\epsilon_{bias} \geq 0$ is some problem-dependent constant.

ϵ_{bias} measures both approximation error and distribution shift error. In later proof, we select a particular function in $\tilde{f}_t^n \in \mathcal{F}_t^n$ such that

$$L(\tilde{f}_t^n; \tilde{d}, Q_{b^n}^t - b^n) \leq 2\epsilon_{\text{bias}}. \tag{16}$$

We establish complexity results by comparing the empirical minimizer f_t^n of (6) with this optimal fitter $\tilde{f}_t^n$.

Assumption C.2. *For the same loss L as defined in Assumption C.1 and the fitter $\tilde{f}_t^n$, we assume that there exists some $C \geq 1$ and $\epsilon_0 \geq 0$ such that for any $f \in \mathcal{F}$,*

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

for $n \in [N]$ and $0 \leq t \leq T-1$.

Remark 3. *Under Assumption 4.1, $Q_{b^n}^t - b^n = \mathbb{E}^{\pi_t^n}[r(s,a) + \gamma Q_{b^n}^t(s',a')] \in \mathcal{F}$. Thus, ϵ_{bias} can take value 0 and $\tilde{f}_t^n = Q_{b^n}^t - b^n$. Further in Assumption C.2, we have*

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

Thus, C can take value 1 and $\epsilon_0 = 0$. If $Q_{b^n}^t - b^n$ 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.

C.1. Sample Complexity of ENIAC-SPI-SAMPLE

We follow the proof steps in Section B.2 and first establish a bonus bound.

Lemma C.1 (SPI-SAMPLE: The Bound of Bonus). *With probability at least $1 - N\delta$, it holds that*

$$\sum_{n=1}^{N}\left(V_{b^n}^{\pi^{n+1}} - V^{\pi^{n+1}}\right) \leq \frac{2\epsilon^2 + 8KW^2 + \beta^2}{(1-\gamma)\beta^2 K} \cdot dim_E(\mathcal{F}, \beta) + \frac{N}{1-\gamma}\sqrt{\frac{\log(2/\delta)}{2K}}.$$

Proof.

$$\begin{aligned}\sum_{n=1}^{N}\left(V_{b^n}^{\pi^{n+1}} - V^{\pi^{n+1}}\right) &\leq \sum_{n=1}^{N} \mathbb{E}_{(s,a)\sim d^{n+1}} \mathbf{1}\{(s,a) \notin \mathcal{K}^n\}/(1-\gamma) \\ &= \sum_{n=1}^{N} \mathbb{E}_{(s,a)\sim d^{n+1}} \mathbf{1}\{w(\tilde{\mathcal{F}}^n, s, a) \geq \beta\}/(1-\gamma),\end{aligned}$$

where d^{n+1} denotes the state-action distribution induced by π^{n+1} on $\mathcal{M}$. We denote by $\mathcal{D}^n$ the sampled dataset $\{(s_i, a_i)\}_{i=1}^K \sim d^n$ at the beginning of epoch n. Then $\mathcal{Z}^n = \mathcal{Z}^{n-1} \cup \mathcal{D}^n$. By Hoeffding's inequality, with probability at least $1 - \delta$,

$$\mathbb{E}_{(s,a)\sim d^{n+1}} \mathbf{1}\{w(\tilde{\mathcal{F}}^n, s, a) \geq \beta\} \leq \frac{1}{K} \sum_{(s,a)\in\mathcal{D}^{n+1}} \mathbf{1}\{w(\tilde{\mathcal{F}}^n, s, a) \geq \beta\} + \sqrt{\frac{\log(2/\delta)}{2K}}.$$

Taking the union bound, with probability at least $1 - N\delta$, we have

$$\sum_{n=1}^N V_{b^n}^{\pi^{n+1}} - V^{\pi^{n+1}} \leq \frac{1}{K(1-\gamma)} \sum_{n=1}^N \sum_{(s,a)\in\mathcal{D}^{n+1}} \mathbf{1}\{w(\tilde{\mathcal{F}}^n, s, a) \geq \beta\} + \frac{N}{1-\gamma}\sqrt{\frac{\log(2/\delta)}{2K}}. \tag{17}$$

Next we bound the first term in Equation (17) following a similar process as in (**?**)Proposition 3]russo2013eluder. We simplify $w(\tilde{\mathcal{F}}^n, \cdot, \cdot)$ as $w^n(\cdot, \cdot)$ and label all samples in $\mathcal{Z}^n$ in lexical order, e.g., (s_i^{n+1}, a_i^{n+1}) denotes the ith sample in $\mathcal{D}^{n+1}$. For every (s_i^{n+1}, a_i^{n+1}), we define a sequence S_{i-1}^{n+1} which contains all samples generated before (s_i^{n+1}, a_i^{n+1}), i.e.,

$$S_{i-1}^{n+1} := \big((s_1^1, a_1^1), \dots, (s_K^1, a_K^1), (s_1^2, a_1^2), \cdots (s_K^n, a_K^n), (s_1^{n+1}, a_1^{n+1}), \dots, (s_{i-1}^{n+1}, a_{i-1}^{n+1})\big) \tag{18}$$

Next we show that,

$$\sum_{n=1}^N \sum_{(s,a)\in\mathcal{D}^{n+1}} \mathbf{1}\{w^n(s,a) \geq \beta\} \leq \Big(2\epsilon^2/\beta^2 + 8W^2K/\beta^2 + 1\Big) \cdot \dim_E(\mathcal{F}, \beta). \tag{19}$$

For $n \leq N$, if $w^n(s_i^{n+1}, a_i^{n+1}) > \beta$ then (s_i^{n+1}, a_i^{n+1}) is β-dependent with respect to $\mathcal{F}$ on fewer than $8(\epsilon)^2/\beta^2 + 32W^2K/\beta^2$ disjoint subsequences of S_{i-1}^{n+1}. To see this, note that if $w^n(s_i^{n+1}, a_i^{n+1}) > \beta$, there exists $\bar{f}, \underline{f} \in \mathcal{F}$ such that $\bar{f} - \underline{f} \in \tilde{\mathcal{F}}^n$ and $\bar{f}(s_i^{n+1}, a_i^{n+1}) - \underline{f}(s_i^{n+1}, a_i^{n+1}) \geq \beta$. By definition, if (s_i^{n+1}, a_i^{n+1}) is β-dependent on a subsequence $\big((s_{t_1}, a_{t_1}), \dots, (s_{t_k}, a_{t_k})\big)$ of S_{i-1}^{n+1}, then $\sum_{j=1}^k (\bar{f}\big(s_{t_j}, a_{t_j}) - \underline{f}(s_{t_j}, a_{t_j})\big)^2 \geq \beta^2$. It follows that, if (s_i^{n+1}, a_i^{n+1}) is β-dependent on L disjoint subsequences of S_{i-1}^{n+1} then $\|\bar{f} - \underline{f}\|_{S_{i-1}^{n+1}}^2 \geq L\beta^2$, where we recall our notation $\|f\|_S = \sqrt{\sum_{x\in S} f(x)^2}$. By the definition of $\tilde{\mathcal{F}}^n$ and $S_{i-1}^{n+1} = \mathcal{Z}^n \cup \{(s_j^{n+1}, a_j^{n+1})\}_{j=1}^{i-1}$, we have

$$\|\bar{f} - \underline{f}\|_{S_{i-1}^{n+1}} \leq \|\bar{f} - \underline{f}\|_{\mathcal{Z}^n} + \|\bar{f} - \underline{f}\|_{\{(s_j^{n+1}, a_j^{n+1})\}_{j=1}^{i-1}} \leq \epsilon + 2W\sqrt{i-1} \leq \epsilon + 2W\sqrt{K},$$

where W is an upper bound of $\|f\|_\infty$. Hence, $L < 2\epsilon^2/\beta^2 + 8W^2K/\beta^2$.

Next, we show that in any state-action sequence $((s_1, a_1), \dots, (s_\tau, a_\tau))$, there is some $j \leq \tau$ such that the element (s_j, a_j) is β-dependent with respect to $\mathcal{F}$ on at least $\tau/d - 1$ disjoint subsequences of the subset $((s_1, a_1), \dots, (s_{j-1}, a_{j-1}))$, where $d := \dim_E(\mathcal{F}, \beta)$. Here we assume that $\tau \geq d$ since otherwise the claim is trivially true. To see this, for an integer L safistying $Ld + 1 \leq \tau \leq (L+1) \cdot d$, we will construct L disjoint subsequences $S_1, \dots, S_L$ *one element at a time*. First, for each $i \in [L]$ add (s_i, a_i) to the subsequence S_i. Now, if (s_{L+1}, a_{L+1}) is β-dependent on all subsequences $S_1, \dots, S_L$, our claim is established. Otherwise, select a subsequence S_i such that (s_{L+1}, a_{L+1}) is β-independent of it and append (s_{L+1}, a_{L+1}) to S_i. Repeat this process for elements with indices $j > L + 1$ until (s_j, a_j) is β-dependent on all subsequences or $j = \tau$. In the latter scenario, since $\tau - 1$ elements have already been put in subsequences, we have that $\sum |S_j| \geq L \cdot d$. However, by the definition of $\dim_E(\mathcal{F}, \beta)$, since each element of a subsequence S_j is β-independent of its predecessors, we must have $|S_j| \leq d, \forall j \in [L]$ and therefore, $\sum |S_j| \leq L \cdot d$. In this case, (s_τ, a_τ) must be β-dependent on all subsequences.

Now consider the subsequence $S_\beta := \big((s_{i_1}^{n_1}, a_{i_1}^{n_1}), \dots, (s_{i_\tau}^{n_\tau}, a_{i_\tau}^{n_\tau})\big)$ of S_K^{N+1} which consists of all elements such that $w_n\big((s_i^{n+1}, a_i^{n+1})\big) \geq \beta$. With that being said, S_β consists of all sample points where large width occurs from epoch 1 to epoch N. The indices in S_β are in lexical order and $(s_{i_j}^{n_j}, a_{i_j}^{n_j})$ denotes the j_{th} element in S_β. As we have established, each $(s_{i_j}^{n_j}, a_{i_j}^{n_j})$ is β-dependent on fewer than $2\epsilon^2/\beta^2 + 8W^2K/\beta^2$ disjoint subsequences of $S_{i_j-1}^{n_j}$ (recall the definition in Equation (18)). It follows that each $(s_{i_j}^{n_j}, a_{i_j}^{n_j})$ is β-dependent on fewer than $2\epsilon^2/\beta^2 + 8W^2K/\beta^2$ disjoint subsequences of $((s_{i_1}^{n_1}, a_i^{n_1}), \dots, (s_{i_{j-1}}^{n_{j-1}}, a_{i_{j-1}}^{n_{j-1}})) \subset S_\beta$, i.e., the elements in S_β before $(s_{i_j}^{n_j}, a_{i_j}^{n_j})$. Combining this with the fact we have established that there exists some $(s_{i_j}^{n_j}, a_{i_j}^{n_j})$ that is β-dependent on at least $\tau/d - 1$ disjoint subsequences of $((s_{i_1}^{n_1}, a_i^{n_1}), \dots, (s_{i_{j-1}}^{n_{j-1}}, a_{i_{j-1}}^{n_{j-1}}))$, we have $\tau/d - 1 \leq 2\epsilon^2/\beta^2 + 8W^2K/\beta^2$. It follows that $\tau \leq \big(2\epsilon^2/\beta^2 + 8W^2K/\beta^2 + 1\big) \cdot d$, which is Equation (19).

Combining all above results, with probability at least $1 - N\delta$,

$$\sum_{n=1}^N \Big(V_{b^n}^{\pi^{n+1}} - V^{\pi^{n+1}}\Big) \leq \frac{2\epsilon^2 + 8KW^2 + \beta^2}{(1-\gamma)\beta^2 K} \cdot \dim_E(\mathcal{F}, \beta) + \frac{N}{1-\gamma}\sqrt{\frac{\log(2/\delta)}{2K}}.$$

■

Next we prove the last step in Section B.2. For notation brevity, we focus on a specific epoch n and drop the dependence on n in the policy and critic functions. We define

$$\widehat{A}^t_{b^n}(s,a) := f_t(s,a) + b^n(s,a) - \mathbb{E}_{a'\sim\pi_t(\cdot|s)}[f_t(s,a') + b^n(s,a')], \tag{20}$$

where f_t is the output of the critic fit step at iteration t in epoch n. It can be easily verified that $\mathbb{E}_{a\sim\pi_t(\cdot|s)}\widehat{A}^t_{b^n}(s,a) = 0$ and the SPI-SAMPLE update in Equation (7) is equivalent to

$$\pi_{t+1}(\cdot|s) \propto \pi_t(\cdot|s)\exp\left(\eta\widehat{A}^t_{b^n}(s,\cdot)\mathbf{1}\{s\in\mathcal{K}^n\}\right), \quad \forall s\in\mathcal{S}. \tag{21}$$

$\widehat{A}^t_{b^n}$ is indeed our approximation to the true advantage function $A^t_{b^n}$. In the sequel, we show that the actor-critic convergence is upper bounded by the approximation error which can further be controlled with sufficient samples under our assumptions.

Lemma C.2 (SPI-SAMPLE: Actor-Critic Convergence)**.** *In ENIAC-SPI-SAMPLE, let $\widehat{A}^t_{b^n}$ be as defined in Equation* (20) *and the stepsize $\eta = \sqrt{\frac{\log(|\mathcal{A}|)}{16W^2T}}$. For any epoch $n\in[N]$, SPI-SAMPLE obtains a sequence of policies $\{\pi_t\}_{t=0}^{T-1}$ such that when comparing to $\tilde{\pi}^n$:*

$$\begin{aligned}\frac{1}{T}\sum_{t=0}^{T-1}(V^{\tilde{\pi}^n}_{\mathcal{M}^n} - V^t_{b^n}) &= \frac{1}{T}\sum_{t=0}^{T-1}(V^{\tilde{\pi}^n}_{\mathcal{M}^n} - V^t_{\mathcal{M}^n}) \\ &\le \frac{1}{1-\gamma}\Big(8W\sqrt{\frac{\log(|\mathcal{A}|)}{T}} + \frac{1}{T}\sum_{t=0}^{T-1}\mathbb{E}_{(s,a)\sim\tilde{d}_{\mathcal{M}^n}}\left[\left(A^t_{b^n}(s,a) - \widehat{A}^t_{b^n}(s,a)\right)\mathbf{1}\{s\in\mathcal{K}^n\}\right]\Big).\end{aligned}$$

Proof. The equality is mentioned in Remark 2. We first show that $A^t_{\mathcal{M}^n}(s,a^\dagger)\le 0$ for any $s\notin\mathcal{K}^n$. Since π_t uniformly randomly selects an unfamiliar action with bonus $1/(1-\gamma)$ for $s\notin\mathcal{K}^n$, we have $V^t_{\mathcal{M}^n}(s)\ge 1/(1-\gamma)$. Thus,

$$A^t_{\mathcal{M}^n}(s,a^\dagger) = Q^t_{\mathcal{M}^n}(s,a^\dagger) - V^t_{\mathcal{M}^n}(s) = 1-(1-\gamma)\cdot V^t_{\mathcal{M}^n}(s)\le 0,\ \ \forall s\notin\mathcal{K}^n,$$

where $Q^t_{\mathcal{M}^n}(s,a^\dagger) = 1+\gamma V^t_{\mathcal{M}^n}(s)$ ($a^\dagger$ leads s to s). Based on the above result, we have

$$\begin{aligned}&V^{\tilde{\pi}^n}_{\mathcal{M}^n} - V^t_{\mathcal{M}^n} = \frac{1}{1-\gamma}\sum_{(s,a)}\tilde{d}_{\mathcal{M}^n}(s,a)A^t_{\mathcal{M}^n}(s,a) \\ &= \frac{1}{1-\gamma}\sum_{(s,a)}\tilde{d}_{\mathcal{M}^n}(s,a)A^t_{\mathcal{M}^n}(s,a)\mathbf{1}\{s\in\mathcal{K}^n\} + \frac{1}{1-\gamma}\sum_{(s,a)}\tilde{d}_{\mathcal{M}^n}(s,a)A^t_{\mathcal{M}^n}(s,a)\mathbf{1}\{s\notin\mathcal{K}^n\} \\ &= \frac{1}{1-\gamma}\sum_{(s,a)}\tilde{d}_{\mathcal{M}^n}(s,a)A^t_{\mathcal{M}^n}(s,a)\mathbf{1}\{s\in\mathcal{K}^n\} + \frac{1}{1-\gamma}\sum_{s}\tilde{d}_{\mathcal{M}^n}(s)A^t_{\mathcal{M}^n}(s,a^\dagger)\mathbf{1}\{s\notin\mathcal{K}^n\} \\ &\le \frac{1}{1-\gamma}\sum_{(s,a)}\tilde{d}_{\mathcal{M}^n}(s,a)A^t_{\mathcal{M}^n}(s,a)\mathbf{1}\{s\in\mathcal{K}^n\} \\ &= \frac{1}{1-\gamma}\sum_{(s,a)}\tilde{d}_{\mathcal{M}^n}(s,a)A^t_{b^n}(s,a)\mathbf{1}\{s\in\mathcal{K}^n\} \\ &= \frac{1}{1-\gamma}\left(\mathbb{E}_{(s,a)\sim\tilde{d}_{\mathcal{M}^n}}\left[\widehat{A}^t_{b^n}(s,a)\mathbf{1}\{s\in\mathcal{K}^n\}\right] + \mathbb{E}_{(s,a)\sim\tilde{d}_{\mathcal{M}^n}}\left[(A^t_{b^n}(s,a) - \widehat{A}^t_{b^n}(s,a))\mathbf{1}\{s\in\mathcal{K}^n\}\right]\right)\end{aligned} \tag{22}$$

where the first line is by the performance difference lemma in Kakade (2003), the third line is due to that $\tilde{\pi}^n$ deterministically picks $a^\dagger$ for $s\notin\mathcal{K}^n$, and the fifth line follows that π_t never picks $a^\dagger$ so for any action $a\in\mathcal{A}$ we have $A^t_{\mathcal{M}^n} = A^t_{b^n}$.

Next we establish an upper bound of the first term in Equation (22). Recall that in SPI-SAMPLE the policy update is equivalent to (21). Thus, for $s\in\mathcal{K}^n$, we have

$$\mathbf{KL}\big(\tilde{\pi}^n(\cdot|s),\pi_{t+1}(\cdot|s)\big) - \mathbf{KL}\big(\tilde{\pi}^n(\cdot|s),\pi_t(\cdot|s)\big) = \mathbb{E}_{a\sim\tilde{\pi}^n(\cdot|s)}[-\eta\widehat{A}^t_{b^n}(s,a) + \log(z^t(s))],$$

where $z^t(s) := \sum_a \pi_t(a|s)\exp(\eta\widehat{A}^t_{b^n}(s,a))$. Since $|\widehat{A}^t_{b^n}(s,a)|\le 4W$ and when $T>\log(|\mathcal{A}|)$, $\eta<1/(4W)$, we have $\eta\widehat{A}^t_{b^n}(s,a)\le 1$. By the inequality that $\exp(x)\le 1+x+x^2$ for $x\le 1$ and $\log(1+x)\le x$ for $x>-1$,

$$\log(z^t(s))\le\eta\sum_a\pi_t(a|s)\widehat{A}^t_{b^n}(s,a) + 16\eta^2W^2 = 16\eta^2W^2.$$

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

$$\mathbf{KL}(\tilde{\pi}^n(\cdot|s), \pi_{t+1}(\cdot|s)) - \mathbf{KL}(\tilde{\pi}^n(\cdot|s), \pi_t(\cdot|s)) \leq -\eta \mathbb{E}_{a\sim\tilde{\pi}^n(\cdot|s)}[\widehat{A}^t_{b^n}(s,a)] + 16\eta^2 W^2.$$

Adding both sides from $t = 0$ to $T-1$ and taking $\eta = \sqrt{\frac{\log(|\mathcal{A}|)}{16W^2T}}$, we get

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

where the inequality follows that $\pi_0(\cdot|s) = \mathrm{Unif}(\mathcal{A})$. Lastly, combining with Equation (22), the regret on $\mathcal{M}^n$ satisfies

$$\sum_{t=0}^{T-1}(V^{\tilde{\pi}^n}_{\mathcal{M}^n} - V^t_{\mathcal{M}^n}) \leq \frac{1}{1-\gamma}\left(8W\sqrt{\log(|\mathcal{A}|)T} + \sum_{t=1}^{T}\mathbb{E}_{(s,a)\sim\tilde{d}_{\mathcal{M}^n}}\left[\Big(A^t_{b^n}(s,a) - \widehat{A}^t_{b^n}(s,a)\Big)\mathbf{1}\{s \in \mathcal{K}^n\}\right]\right).$$

■

Next, we analyze the approximation error and build an upper bound on $A^t_{b^n} - \widehat{A}^t_{b^n}$. Recall that $A^t_{b^n}$ is the true advantage of policy π^n_t in the bonus-added MDP and $\widehat{A}^t_{b^n}$ is an approximation to $A^t_{b^n}$ with the empirical minimizer f_t as defined in (20). We still focus on a specific epoch n and simplify the notation $\tilde{f}^n_t$ as defined in (16) to f^*_t.

Lemma C.3 (SPI-SAMPLE: Approximation Bound)**.** *At epoch n, assume for all $0 \leq t \leq T-1$:*

$$L(f_t; \rho^n_{cov}, Q^t_{b^n} - b^n) \leq L(f^*_t; \rho^n_{cov}, Q^t_{b^n} - b^n) + \epsilon_{stat}, \tag{23}$$

where $\epsilon_{stat} > 0$ is to be determined in the next lemma, and let

$$\epsilon^2 = NK\big(C \cdot \epsilon_{stat} + \epsilon_0 + 16W\epsilon_1\big) + 8W^2\log(\mathcal{N}(\mathcal{F}, \epsilon_1)/\delta) \cdot \sqrt{NK}, \tag{24}$$

where ϵ is used in bonus function (see Section 3.3) and C, ϵ_0 are defined in Assumption C.2, and $\epsilon_1 > 0$ denotes the function cover radius which will be determined later. Under Assumption C.1 and C.2, we have that for every $0 \leq t \leq T-1$, with probability at least $1-\delta$,

$$\mathbb{E}_{(s,a)\sim\tilde{d}_{\mathcal{M}^n}}\Big(A^t_{b^n}(s,a) - \widehat{A}^t_{b^n}(s,a)\Big)\mathbf{1}\{s \in \mathcal{K}^n\} \leq 4\sqrt{|\mathcal{A}|\epsilon_{bias}} + 2\beta.$$

Proof. To analyze the difference between $A^t_{b^n}$ and $\widehat{A}^t_{b^n}$, we introduce an intermediate variable $A^*_t(s,a) := f^*_t + b^n - \mathbb{E}_{a'\sim\pi_t(\cdot|s)}[f^*_t + b^n]$, i.e., the approximated advantage generated by the selected best on-policy fit. Then

$$\mathbb{E}_{(s,a)\sim\tilde{d}_{\mathcal{M}^n}}(A^t_{b^n} - \widehat{A}^t_{b^n})\mathbf{1}\{s \in \mathcal{K}^n\} = \mathbb{E}_{(s,a)\sim\tilde{d}_{\mathcal{M}^n}}\left[(A^t_{b^n} - A^*_t)\mathbf{1}\{s \in \mathcal{K}^n\} + (A^*_t - \widehat{A}^t_{b^n})\mathbf{1}\{s \in \mathcal{K}^n\}\right].$$

For the first difference, we have

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

where the first inequality is by Cauchy-Schwarz, the second inequality is by Lemma B.1, and the last two lines follow Assumption C.1 and the definition of f^*_t.

For the second difference,

$$\begin{aligned}
&\mathbb{E}_{(s,a)\sim\tilde{d}_{\mathcal{M}^n}}(A^*_t - \widehat{A}^t_{b^n})\mathbf{1}\{s \in \mathcal{K}^n\} \\
=&\mathbb{E}_{(s,a)\sim\tilde{d}_{\mathcal{M}^n}}(f^*_t - f_t)\mathbf{1}\{s \in \mathcal{K}^n\} - \mathbb{E}_{s\sim\tilde{d}_{\mathcal{M}^n}, a\sim\pi_t(\cdot|s)}(f^*_t - f_t)\mathbf{1}\{s \in \mathcal{K}^n\}
\end{aligned} \tag{25}$$

Next we show that $\Delta f_t := (f_t^* - f_t) \in \tilde{\mathcal{F}}^n$. Recall that $\tilde{\mathcal{F}}^n := \{\Delta f \in \Delta\mathcal{F} \mid \|\Delta f\|_{\mathcal{Z}^n} \leq \epsilon\}$. We only need to show that $\|\Delta f_t\|_{\mathcal{Z}^n} \leq \epsilon$. To achieve this, we plan to utilize the fact that f_t is trained with samples generated from $\rho_{\text{cov}}^n := \text{Unif}(d_{s_0}^{\pi^1}, d_{s_0}^{\pi^2}, \ldots, d_{s_0}^{\pi^n})$ while $\mathcal{Z}^n$ is sequentially constructed with samples from $d_{s_0}^{\pi^i}, i \in [n]$. However, such a correlation does not guarantee a trivial concentration bound. We need to deal with the subtle randomness dependency therein: 1. π^i depends on $\pi^{[i-1]}$ thus the samples in $\mathcal{Z}^n$ are not independent; 2. $\mathcal{Z}^n$ determines $\tilde{\mathcal{F}}^n$, $\tilde{\mathcal{F}}^n$ defines the bonus b^n, and Δf_t is obtained based on b^n. So Δf_t and $\mathcal{Z}^n$ are not independent. Nevertheless, we carefully leverage function cover on $\Delta\mathcal{F}$ to establish a martingale convergence on every anchor function in the cover set, then transform to a bound on the realization Δf_t.

Let $\mathcal{C}(\Delta\mathcal{F}, 2\epsilon_1)$ be a cover set of $\Delta\mathcal{F}$. Then for every $\Delta f \in \Delta\mathcal{F}$, there exists a $\Delta g \in \mathcal{C}(\Delta\mathcal{F}, 2\epsilon_1)$ such that $\|\Delta f - \Delta g\|_\infty \leq 2\epsilon_1$. We rank the samples in $\mathcal{Z}^n$ in lexical order, i.e., (s_k^i, a_k^i) is the k_{th} sample generated following $d_{s_0}^{\pi^i}$ at the beginning of the i_{th} epoch. There are in total nK samples in $\mathcal{Z}^n$. For every $\Delta g \in \mathcal{C}(\Delta\mathcal{F}, 2\epsilon_1)$, we define nK corresponding random variables:

$$X_{(i,k)}^{\Delta g} := (\Delta g(s_k^i, a_k^i))^2 - \mathbb{E}_{(s,a)\sim d_{s_0}^{\pi^i}}[(\Delta g(s,a))^2], \;\; i \in [n], k \in [K]$$

We rank $\{X_{(i,k)}^{\Delta g}\}$ in lexical order and upon which, we define a martingale:

$$Y_{0,0}^{\Delta g} = 0, \;\; Y_{(i,k)}^{\Delta g} = \sum_{(i',k')=(1,1)}^{(i,k)} X_{(i',k')}^{\Delta g}, \quad i \in [n], k \in [K].$$

Then by single-sided Azuma-Hoeffding's inequality, with probability at least $1-\delta$, for all $\Delta g \in \mathcal{C}(\Delta\mathcal{F}, 2\epsilon_1)$, it holds that

$$Y_{(n,K)}^{\Delta g} \leq \sqrt{32W^4 \cdot nK \cdot \log\Big(\frac{\mathcal{N}(\Delta\mathcal{F}, 2\epsilon_1)}{\delta}\Big)} \leq \sqrt{64W^4 \cdot nK \cdot \log\Big(\frac{\mathcal{N}(\mathcal{F}, \epsilon_1)}{\delta}\Big)}, \tag{26}$$

where the right inequality is by Lemma E.1. Next, we transform to Δf_t. Since there exists a $\Delta g \in \mathcal{C}(\Delta\mathcal{F}, 2\epsilon_1)$ such that $\|\Delta f_t - \Delta g\|_\infty \leq 2\epsilon_1$, we have that for all $i \in [n]$ and $k \in [K]$,

$$\begin{aligned} &\big|(\Delta f_t(s_k^i, a_k^i))^2 - (\Delta g(s_k^i, a_k^i))^2\big| \\ &= |\Delta f_t(s_k^i, a_k^i) - \Delta g(s_k^i, a_k^i)| \cdot |\Delta f_t(s_k^i, a_k^i) + \Delta g(s_k^i, a_k^i))| \leq 8W\epsilon_1 \end{aligned}$$

and

$$\begin{aligned} &\Big|\mathbb{E}_{(s,a)\sim d_{s_0}^{\pi^i}}[(\Delta f_t(s,a))^2] - \mathbb{E}_{(s,a)\sim d_{s_0}^{\pi^i}}[(\Delta g(s,a))^2]\Big| \\ &\leq \mathbb{E}_{(s,a)\sim d_{s_0}^{\pi^i}}|\Delta f_t(s,a) - \Delta g(s,a)| \cdot |\Delta f_t(s,a) + \Delta g(s,a)| \leq 8W\epsilon_1 \end{aligned}$$

Therefore,

$$\begin{aligned} Y_{(n,K)}^{\Delta f_t} &= \sum_{(i,k)=(1,1)}^{(n,K)} (\Delta f_t(s_k^i, a_k^i))^2 - \mathbb{E}_{(s,a)\sim d_{s_0}^{\pi^i}}[(\Delta f_t(s,a))^2] \\ &\leq \sum_{(i,k)=(1,1)}^{(n,K)} (\Delta g(s_k^i, a_k^i))^2 - \mathbb{E}_{(s,a)\sim d_{s_0}^{\pi^i}}[(\Delta g(s,a))^2] + nK \cdot 16W\epsilon_1 \\ &= Y_{(n,K)}^{\Delta g} + nK \cdot 16W\epsilon_1. \end{aligned} \tag{27}$$

Note that

$$Y_{(n,K)}^{\Delta f_t} = \|\Delta f_t\|_{\mathcal{Z}^n}^2 - \sum_{i=1}^n K \cdot \mathbb{E}_{d_{s_0}^{\pi^i}}[(\Delta f_t)^2] = \|\Delta f_t\|_{\mathcal{Z}^n}^2 - nK \cdot \mathbb{E}_{\rho_{\text{cov}}^n}[(\Delta f_t)^2]. \tag{28}$$

Combining (26), (27), and (28), we have that

$$\|\Delta f_t\|_{\mathcal{Z}^n}^2 \leq nK \cdot \mathbb{E}_{\rho_{\text{cov}}^n}[(\Delta f_t)^2] + nK \cdot 16W\epsilon_1 + \sqrt{64W^4 \cdot nK \cdot \log\Big(\frac{\mathcal{N}(\mathcal{F}, \epsilon_1)}{\delta}\Big)}.$$

By Assumption C.2,

$$\begin{aligned} \mathbb{E}_{\rho_{\text{cov}}^n}[(\Delta f_t)^2] &= \mathbb{E}_{(s,a)\sim\rho_{\text{cov}}^n}[(f_t^* - f_t)^2] \leq C \cdot (L(f_t; \rho_{\text{cov}}^n, Q_{b^n}^t - b^n) - L(f_t^*; \rho_{\text{cov}}^n, Q_{b^n}^t - b^n)) + \epsilon_0 \\ &\leq C \cdot \epsilon_{\text{stat}} + \epsilon_0. \end{aligned}$$

By the choice of ϵ, $\|\Delta f_t\|_{\mathcal{Z}^n}^2 \leq \epsilon^2$ with probability at least $1-\delta$. Thus, $\Delta f_t \in \tilde{\mathcal{F}}^n$ and for all $(s,a) \in \mathcal{K}^n$, $|f_t^*(s,a) - f_t(s,a)| \leq \beta$. Plugging into (25), we have (25) $\leq 2\beta$. The desired result is obtained. ■

Next, we give an explicit form of ϵ_{stat} as defined in Equation (23).

Lemma C.4. *Following the same notation as in Lemma C.3, it holds with probability at least $1-\delta$ that*

$$L(f_t;\rho^n_{cov},Q^t_{b^n}-b^n)-L(f^*_t;\rho^n_{cov},Q^t_{b^n}-b^n)\le\frac{500C\cdot W^4\cdot\log\left(\frac{\mathcal{N}(\mathcal{F},\epsilon_2)}{\delta}\right)}{M}+13W^2\cdot\epsilon_2+\epsilon_0,$$

where C, ϵ_0 are defined in Assumption C.2, and $\epsilon_2>0$ denotes the function cover radius which will be determined later.

Proof. First note that in the loss function, the expectation has a nested structure: the outer expectation is taken over $(s,a)\sim\rho^n_{\text{cov}}$ and the inner conditional expectation is $Q^t_{b^n}(s,a)=\mathbb{E}^{\pi_t}[\sum_{h=0}^{\infty}\gamma^h\big(r(s_h,a_h)+b^n(s_h,a_h)\big)|(s_0,a_0)=(s,a)]$ given a sample of $(s,a)\sim\rho^n_{\text{cov}}$. To simplify the notation, we use x to denote (s,a), $y|x$ for an unbiased sample of $Q^t_{b^n}(s,a)-b^n(s,a)$, and ν for ρ^n_{cov}, the marginal distribution over x, then the loss function can be recast as

$$\mathbb{E}_{x\sim\nu}[(f_t(x)-\mathbb{E}[y|x])^2]:=L(f_t;\rho^n_{\text{cov}},Q^t_{b^n}-b^n)$$
$$\mathbb{E}_{x\sim\nu}[(f^*_t(x)-\mathbb{E}[y|x])^2]:=L(f^*_t;\rho^n_{\text{cov}},Q^t_{b^n}-b^n).$$

In particular, f_t can be rewritten as

$$f_t\in\operatorname*{argmin}_{f\in\mathcal{F}}\sum_{i=1}^{M}(f(x_i)-y_i)^2,$$

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. For any function f, we have:

$$\begin{aligned}&\mathbb{E}_{x,y}[(f_t(x)-y)^2]\\ =\;&\mathbb{E}_{x,y}[(f_t(x)-\mathbb{E}[y|x])^2]+\mathbb{E}_{x,y}[(\mathbb{E}[y|x]-y)^2]+2\mathbb{E}_{x,y}[(f_t(x)-\mathbb{E}[y|x])(\mathbb{E}[y|x]-y)]\\ =\;&\mathbb{E}_{x,y}[(f_t(x)-\mathbb{E}[y|x])^2]+\mathbb{E}_{x,y}[(\mathbb{E}[y|x]-y)^2],\end{aligned}$$

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

$$\begin{aligned}&\mathbb{E}_x[(f_t(x)-\mathbb{E}[y|x])^2]-\mathbb{E}_x[(f^*_t(x)-\mathbb{E}[y|x])^2]\\ =\;&\mathbb{E}_{x,y}(f_t(x)-y)^2-\mathbb{E}_{x,y}(f^*_t(x)-y)^2.\end{aligned}\tag{29}$$

Next, we establish a concentration bound on f_t. Since f_t depends on the training set $\{(x_i,y_i)\}_{i=1}^M$, as in Assumption C.3, we use a function cover on $\mathcal{F}$ for a uniform convergence argument. We denote by $\mathscr{F}^n_t$ the σ-algebra generated by randomness before epoch n iteration t. Recall that $f^*_t\in\operatorname{argmin}_{f\in\mathcal{F}}L(f;\rho^n_{\text{cov}},Q^t_{b^n}-b^n)$. Conditioning on $\mathscr{F}^n_t$, ρ^n_{cov}, $Q^t_{b^n}-b^n$, and f^*_t are all deterministic. For any $f\in\mathcal{F}$, we define

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

Then $Z_1(f),\dots,Z_M(f)$ are i.i.d. random variables and

$$\begin{aligned}\mathbb{V}[Z_i(f)\mid\mathscr{F}^n_t]&\le\mathbb{E}[Z_i(f)^2\mid\mathscr{F}^n_t]\\ &=\mathbb{E}\left[\Big((f(x_i)-y_i)^2-(f^*_t(x_i)-y_i)^2\Big)^2\mid\mathscr{F}^n_t\right]\\ &=\mathbb{E}\left[\big(f(x_i)-f^*_t(x_i)\big)^2\cdot\big(f(x_i)+f^*_t(x_i)-2y_i\big)^2\mid\mathscr{F}^n_t\right]\\ &\le 36W^4\cdot\mathbb{E}[\big(f(x_i)-f^*_t(x_i)\big)^2\mid\mathscr{F}^n_t]\\ &\le 36W^4\cdot(C\cdot\mathbb{E}[Z_i(f)\mid\mathscr{F}^n_t]+\epsilon_0),\end{aligned}$$

where the last inequality is by Assumption C.2 and Equation (29). 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 $1-\delta$, for all $g\in\mathcal{C}(\mathcal{F},\epsilon_2)$,

$$\begin{aligned}&\mathbb{E}[Z_i(g)\mid\mathscr{F}^n_t]-\frac{1}{M}\sum_{i=1}^{M}Z_i(g)\\ \le&\sqrt{\frac{2\mathbb{V}[Z_i(g)\mid\mathscr{F}^n_t]\cdot\log\frac{\mathcal{N}(\mathcal{F},\epsilon_2)}{\delta}}{M}}+\frac{12W^4\cdot\log\frac{\mathcal{N}(\mathcal{F},\epsilon_2)}{\delta}}{M}\\ \le&\sqrt{\frac{72W^4(C\cdot\mathbb{E}[Z_i(g)\mid\mathscr{F}^n_t]+\epsilon_0)\cdot\log\frac{\mathcal{N}(\mathcal{F},\epsilon_2)}{\delta}}{M}}+\frac{12W^4\cdot\log\frac{\mathcal{N}(\mathcal{F},\epsilon_2)}{\delta}}{M}.\end{aligned}$$

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

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

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

$$\begin{aligned}&\mathbb{E}[Z_i(f_t) \mid \mathscr{F}_t^n] - \frac{1}{M}\sum_{i=1}^{M} Z_i(f_t) \\ \leq& \mathbb{E}[Z_i(g) \mid \mathscr{F}_t^n] - \frac{1}{M}\sum_{i=1}^{M} Z_i(g) + 12W^2\epsilon_2 \\ \leq& \sqrt{\frac{72W^4(C \cdot \mathbb{E}[Z_i(g) \mid \mathscr{F}_t^n] + \epsilon_0)\log\frac{\mathcal{N}(\mathcal{F},\epsilon_2)}{\delta}}{M}} + \frac{12W^4\log\frac{\mathcal{N}(\mathcal{F},\epsilon_2)}{\delta}}{M} + 12W^2\epsilon_2 \\ \leq& \sqrt{\frac{72W^4(C \cdot \mathbb{E}[Z_i(f_t) \mid \mathscr{F}_t^n] + 6CW^2\epsilon_2 + \epsilon_0)\log\frac{\mathcal{N}(\mathcal{F},\epsilon_2)}{\delta}}{M}} + \frac{12W^4\log\frac{\mathcal{N}(\mathcal{F},\epsilon_2)}{\delta}}{M} + 12W^2\epsilon_2.\end{aligned}$$

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

$$\mathbb{E}[Z_i(f_t) \mid \mathscr{F}_t^n] \leq \sqrt{\frac{72W^4(C \cdot \mathbb{E}[Z_i(f_t) \mid \mathscr{F}_t^n] + 6CW^2\epsilon_2 + \epsilon_0)\log\frac{\mathcal{N}(\mathcal{F},\epsilon_2)}{\delta}}{M}} + \frac{12W^4\log\frac{\mathcal{N}(\mathcal{F},\epsilon_2)}{\delta}}{M} + 12W^2\epsilon_2.$$

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

$$\mathbb{E}[Z_i(f_t) \mid \mathscr{F}_t^n] \leq \frac{500C \cdot W^4 \cdot \log\frac{\mathcal{N}(\mathcal{F},\epsilon_2)}{\delta}}{M} + 13W^2 \cdot \epsilon_2 + \epsilon_0.$$

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

$$\mathbb{E}[Z_i(f_t)] \leq \frac{500C \cdot W^4 \cdot \log\frac{\mathcal{N}(\mathcal{F},\epsilon_2)}{\delta}}{M} + 13W^2 \cdot \epsilon_2 + \epsilon_0.$$

Notice that $\mathbb{E}[Z_i(f_t)] = L(f_t; \rho_{\text{cov}}^n, Q_{b^n}^t - b^n) - L(f_t^*; \rho_{\text{cov}}^n, Q_{b^n}^t - b^n)$. The desired result is obtained. ■

Combining all previous lemmas, we have the following theorem which states the detailed sample complexity of ENIAC-SPI-SAMPLE (a detailed version of Theorem 4.1)

Theorem C.1 (Main Result: Sample Complexity of ENIAC-SPI-SAMPLE). *Let $\delta \in (0, 1)$ and $\varepsilon \in (0, 1/(1-\gamma))$. With Assumptions C.1, C.2, 4.2, and 4.3, we set the hyperparameters as:*

$$\begin{aligned}&\beta = \frac{\varepsilon(1-\gamma)}{2},\;\; T = \frac{64W^2 \cdot \log|\mathcal{A}|}{\varepsilon^2(1-\gamma)^2},\;\; N \geq \frac{32W^2 \cdot \mathit{dim}_E(\mathcal{F},\beta)}{\varepsilon^3(1-\gamma)^3},\;\; \eta = \sqrt{\frac{\log(|\mathcal{A}|)}{16W^2T}} \\ &\epsilon_1 = \frac{(1-\gamma)^3\varepsilon^3}{128W \cdot \mathit{dim}_E(\mathcal{F},\beta)},\;\; K = \frac{128W^2 \cdot \mathit{dim}_E(\mathcal{F},\beta) \cdot \left(\log(\frac{3NT \cdot \mathcal{N}(\mathcal{F},\epsilon_1)}{\delta})\right)^2 \cdot \log(\frac{6NT}{\delta})}{\varepsilon^3(1-\gamma)^3}, \\ &\epsilon_2 = \frac{(1-\gamma)^3\varepsilon^3}{110C \cdot W^2 \cdot \mathit{dim}_E(\mathcal{F},\beta)},\;\; M = \frac{4000C^2W^4 \cdot \mathit{dim}_E(\mathcal{F},\beta) \cdot \log(\frac{3NT \cdot \mathcal{N}(\mathcal{F},\epsilon_2)}{\delta})}{\varepsilon^3(1-\gamma)^3},\end{aligned}$$

and ϵ satisfies Equation (24) correspondingly. Then with probability at least $1 - \delta$, for the average policy $\pi_{ave}^N := \pi_{ave}^N :=$ $\text{Unif}(\pi^2, \ldots, \pi^{N+1})$, we have

$$V^{\pi_{ave}^N} \geq V^{\tilde{\pi}} - \frac{4\sqrt{|\mathcal{A}|\epsilon_{bias}}}{1-\gamma} - \epsilon_0 \cdot \frac{16C\mathit{dim}_E(\mathcal{F},\beta)}{\varepsilon^2(1-\gamma)^3} - 9\varepsilon$$

for any comparator $\tilde{\pi}$ with total number of samples:

$$\tilde{\mathcal{O}}\Big(\frac{C^2W^8 \cdot \left(\mathit{dim}_E(\mathcal{F},\beta)\right)^2 \cdot \left(\log(\mathcal{N}(\mathcal{F},\epsilon'))\right)^2}{\varepsilon^8(1-\gamma)^8}\Big),$$

where $\epsilon' = \min(\epsilon_1, \epsilon_2)$.

Proof. By Lemma C.1, we have that with probability at least $1 - N\delta_1$,

$$V^{\pi_{\text{ave}}^N} \geq \frac{1}{N}\sum_{n=1}^{N} V_{b^n}^{\pi^{n+1}} - \frac{2\epsilon^2 + 8KW^2 + \beta^2}{(1-\gamma)\beta^2 NK} \cdot \dim_E(\mathcal{F}, \beta) + \frac{1}{1-\gamma}\sqrt{\frac{\log(2/\delta_1)}{2K}}. \tag{30}$$

By Lemma C.2, C.3, and B.2, we have that for every $n \in [N]$, with probability at least $1 - 2T\delta_1$,

$$V_{b^n}^{\pi^{n+1}} \geq V^{\tilde{\pi}} - \frac{1}{1-\gamma}\Big(8W\sqrt{\frac{\log(|\mathcal{A}|)}{T}} + 4\sqrt{|\mathcal{A}|\epsilon_{\text{bias}}} + 2\beta\Big). \tag{31}$$

Combining inequalities (30) and (31), we have with probability at least $1 - 3NT\delta_1$,

$$\begin{aligned} V^{\pi_{\text{ave}}^N} \geq V^{\tilde{\pi}} - \frac{1}{1-\gamma}\bigg(&\frac{2\epsilon^2 + 8KW^2 + \beta^2}{\beta^2 NK} \cdot \dim_E(\mathcal{F}, \beta) + \sqrt{\frac{\log(2/\delta_1)}{2K}} \\ &+ 8W\sqrt{\frac{\log(|\mathcal{A}|)}{T}} + 4\sqrt{|\mathcal{A}|\epsilon_{\text{bias}}} + 2\beta\bigg). \end{aligned} \tag{32}$$

We plug in the value of ϵ^2 in Equation (24) with the bound on ϵ_{stat} in Lemma C.4 and choose hyperparameters such that every term in (32) (except for the ones with ϵ_0 or ϵ_{bias}) is bounded by ε. Finally, we set $\delta_1 = \delta/(3NT)$ and $\epsilon' = \min(\epsilon_1, \epsilon_2)$. In total, the sample complexity is

$$N(K + TM) = \tilde{\mathcal{O}}\Big(\frac{C^2 W^8 \cdot \big(\dim_E(\mathcal{F}, \beta)\big)^2 \cdot \big(\log(\mathcal{N}(\mathcal{F}, \epsilon'))\big)^2}{\varepsilon^8(1-\gamma)^8}\Big).$$

■

Corollary 1. *If Assumption 4.1 holds, with proper hyperparameters, the average policy $\pi_{ave}^N := \mathrm{Unif}(\pi^2, \ldots, \pi^{N+1})$ of ENIAC-SPI-SAMPLE achieves $V^{\pi_{ave}^N} \geq V^{\tilde{\pi}} - \varepsilon$ with probability at least $1 - \delta$ and the sample complexity is*

$$\tilde{\mathcal{O}}\Big(\frac{W^8 \cdot \big(dim_E(\mathcal{F}, \beta)\big)^2 \cdot \big(\log(\mathcal{N}(\mathcal{F}, \epsilon'))\big)^2}{\varepsilon^8(1-\gamma)^8}\Big).$$

Proof. The result is straightforward as mentioned in Remark 3 that under Assumption 4.1, $\epsilon_{\text{bias}} = 0$, $C = 1$, and $\epsilon_0 = 0$. ■

C.2. Sample Complexity of ENIAC-SPI-COMPUTE

In this section, we prove the result for ENIAC-SPI-COMPUTE. SPI-COMPUTE only differs from SPI-SAMPLE at two places: the value of the bonus and the actor update rule. These differences cause changes in the bonus bound result and the convergence analysis while Lemma C.3 and C.4 still hold with the same definition of $\widehat{A}_{b^n}^t$ as in (20). In the sequel, we present the bonus bound and the convergence result for SPI-COMPUTE.

Lemma C.5 (SPI-COMPUTE: The Bound of Bonus)**.** *With probability at least $1 - N\delta$,*

$$\sum_{n=1}^{N} V_{b^n}^{\pi^{n+1}} - V^{\pi^{n+1}} \leq \frac{|\mathcal{A}|}{(1-\gamma)\alpha} \cdot \frac{2\epsilon^2 + 8W^2K + \beta^2}{\beta^2 K} \cdot dim_E(\mathcal{F}, \beta) + \frac{N|\mathcal{A}|}{(1-\gamma)\alpha}\sqrt{\frac{\log(2/\delta)}{2K}}.$$

The proof is similar to Lemma C.1. We only need to revise the bonus value from $\frac{1}{1-\gamma}$ to $\frac{|\mathcal{A}|}{(1-\gamma)\alpha}$.

As for the actor-critic convergence, we focus on a specific epoch n and still define

$$\widehat{A}_{b^n}^t(s,a) := f_t(s,a) + b^n(s,a) - \mathbb{E}_{a' \sim \pi_t(\cdot|s)}[f_t(s,a') + b^n(s,a')]. \tag{33}$$

It is easy to verify that $\mathbb{E}_{a\sim\pi_t(\cdot|s)}[\widehat{A}_{b^n}^t] = 0$ and for $s \in \mathcal{K}^n$, the actor update in SPI-COMPUTE is equivalent to

$$\pi'_{t+1}(a|s) \propto \pi'_t(a|s)\exp\big(\eta \widehat{A}_{b^n}^t(s,a)\big),\ \pi_{t+1} = (1-\alpha)\pi'_{t+1} + \alpha\mathrm{Unif}(\mathcal{A})$$

since $b^n(s,\cdot) = 0$ for $s \in \mathcal{K}^n$. As before, we use $\widehat{A}_{b^n}(s,a)$ to approximate the true advantage of π_t^n on $\mathcal{M}_{b^n}$. Then we have the following result.

Lemma C.6 (SPI-COMPUTE: Actor-Critic Convergence)**.** *In ENIAC-SPI-COMPUTE, let $\widehat{A}_{b^n}^t$ be as defined in Equation (33), $\eta = \sqrt{\frac{\log(|\mathcal{A}|)}{16W^2T}}$, and $\alpha = \frac{1}{1+\sqrt{T}}$. For any epoch $n \in [N]$, SPI-COMPUTE obtains a sequence of policies $\{\pi_t\}_{t=0}^{T-1}$ such that when comparing to $\tilde{\pi}^n$:*

$$\begin{aligned}
&\frac{1}{T}\sum_{t=0}^{T-1}(V_{\mathcal{M}^n}^{\tilde{\pi}^n} - V_{b^n}^t) = \frac{1}{T}\sum_{t=0}^{T-1}(V_{\mathcal{M}^n}^{\tilde{\pi}^n} - V_{\mathcal{M}^n}^t)\\
&\leq \frac{1}{1-\gamma}\Big(12W\sqrt{\frac{\log(|\mathcal{A}|)}{T}} + \frac{1}{T}\sum_{t=0}^{T-1}\Big(\mathbb{E}_{(s,a)\sim\tilde{d}_{\mathcal{M}^n}}\big(A_{b^n}^t(s,a) - \widehat{A}_{b^n}^t(s,a)\big)\mathbf{1}\{s\in\mathcal{K}^n\}\Big).
\end{aligned}$$

Proof of Lemma C.6. Similar to the reasoning in Lemma C.2, we first have that $A_{\mathcal{M}^n}^t(s,a^\dagger)\leq 0$ for any $s\notin\mathcal{K}^n$. To see this, note that for $s\notin\mathcal{K}^n$, there exists an action with bonus $b^n = |\mathcal{A}|/\big((1-\gamma)\alpha\big)$ and π_t has probability at least $\alpha/|\mathcal{A}|$ selects that action. Therefore, $V_{\mathcal{M}^n}^t(s)\geq 1/(1-\gamma)$ and

$$A_{\mathcal{M}^n}^t(s,a^\dagger) = Q_{\mathcal{M}^n}^t(s,a^\dagger) - V_{\mathcal{M}^n}^t(s) = 1-(1-\gamma)\cdot V_{\mathcal{M}^n}^t(s)\leq 0,\ \ \forall s\notin\mathcal{K}^n.$$

Recall that $\tilde{\pi}^n$ deterministically picks $a^\dagger$ for $s\notin\mathcal{K}^n$. Based on the above inequality, it holds that

$$\begin{aligned}
V_{\mathcal{M}^n}^{\tilde{\pi}^n} - V_{\mathcal{M}^n}^t &= \frac{1}{1-\gamma}\sum_{(s,a)}\tilde{d}_{\mathcal{M}^n}(s,a)A_{\mathcal{M}^n}^t(s,a) \leq \frac{1}{1-\gamma}\sum_{(s,a)}\tilde{d}_{\mathcal{M}^n}(s,a)A_{\mathcal{M}^n}^t(s,a)\mathbf{1}\{s\in\mathcal{K}^n\}\\
&= \frac{1}{1-\gamma}\sum_{(s,a)}\tilde{d}_{\mathcal{M}^n}(s,a)A_{b^n}^t(s,a)\mathbf{1}\{s\in\mathcal{K}^n\}. \qquad (34)
\end{aligned}$$

Next we restrict on $s\in\mathcal{K}^n$ and establish the consecutive KL difference on $\{\pi_t'(\cdot|s)\}$. Specifically, since for $s\in\mathcal{K}^n$, $\pi_{t+1}'(\cdot|s)\propto\pi_t'(\cdot|s)\exp(\eta\widehat{A}_{b^n}^t(s,a))$,

$$\mathbf{KL}(\tilde{\pi}^n(\cdot|s),\pi_{t+1}'(\cdot|s)) - \mathbf{KL}(\tilde{\pi}^n(\cdot|s),\pi_t'(\cdot|s)) = \mathbb{E}_{a\sim\tilde{\pi}^n(\cdot|s)}[-\eta\widehat{A}_{b^n}^t(s,a)+\log(z^t)],$$

where $z^t := \sum_a \pi_t'(a|s)\exp(\eta\widehat{A}_{b^n}^t(s,a))$. With the assumptions that $|\widehat{A}_{b^n}^t(s,a)|\leq 4W$ and $\eta\leq 1/(4W)$ when $T>\log(|\mathcal{A}|)$, we have that $\eta\widehat{A}_{b^n}^t(s,a)\leq 1$. By the inequality that $\exp(x)\leq 1+x+x^2$ for $x\leq 1$, we have that

$$\begin{aligned}
\log(z^t) &\leq \log(1+\eta\sum_a\pi_t'(a|s)\widehat{A}_{b^n}^t(s,a)+16\eta^2W^2)\\
&= \log\left(1+\eta\sum_a\Big(\frac{\pi_t(a|s)}{1-\alpha} - \frac{\alpha\cdot\mathrm{Unif}(\mathcal{A})}{1-\alpha}\Big)\cdot\widehat{A}_{b^n}^t(s,a)+16\eta^2W^2\right)\\
&= \log\left(1-\frac{\eta\alpha}{(1-\alpha)|\mathcal{A}|}\sum_a\widehat{A}_{b^n}^t(s,a)+16\eta^2W^2\right)\\
&\leq \log(1+\eta\frac{4W\alpha}{1-\alpha}+16\eta^2W^2)\\
&\leq \frac{4W\eta\alpha}{1-\alpha}+16\eta^2W^2,
\end{aligned}$$

where the second line follows from that $\pi_t' = \frac{\pi_t}{1-\alpha} - \frac{\alpha\mathrm{Unif}(\mathcal{A})}{1-\alpha}$ and the last line follows that $\log(1+x)\leq x$ for $x>0$. Hence, for $s\in\mathcal{K}^n$,

$$\mathbf{KL}(\tilde{\pi}^n(\cdot|s),\pi_{t+1}'(\cdot|s)) - \mathbf{KL}(\tilde{\pi}^n(\cdot|s),\pi_t'(\cdot|s)) \leq -\eta\mathbb{E}_{a\sim\tilde{\pi}^n(\cdot|s)}[\widehat{A}_{b^n}^t(s,a)] + \frac{4W\eta\alpha}{1-\alpha}+16\eta^2W^2.$$

Take $\alpha = \frac{1}{1+\sqrt{T}}$. Adding both sides from $t=0$ to $T-1$, we get

$$\begin{aligned}
&\sum_{t=0}^{T-1}\mathbb{E}_{(s,a)\sim\tilde{d}_{\mathcal{M}^n}}[\widehat{A}_{b^n}^t(s,a)\mathbf{1}\{s\in\mathcal{K}^n\}]\\
\leq&\frac{1}{\eta}\mathbb{E}_{s\sim\tilde{d}_{\mathcal{M}^n}}\left[\big(\mathbf{KL}(\tilde{\pi}^n(\cdot|s),\pi_0'(\cdot|s)) - \mathbf{KL}(\tilde{\pi}^n(\cdot|s),\pi_T'(\cdot|s))\big)\mathbf{1}\{s\in\mathcal{K}^n\}\right] + 4W\sqrt{T}+16\eta TW^2\\
\leq&\log(|\mathcal{A}|)/\eta + 4W\sqrt{T}+16\eta TW^2 \leq 12W\sqrt{\log(|\mathcal{A}|)T}.
\end{aligned}$$

Combining with Equation (34), the regret on $\mathcal{M}^n$ satisfies

$$
\begin{aligned}
&\sum_{t=0}^{T-1}(V^{\tilde{\pi}^n}_{\mathcal{M}^n} - V^t_{\mathcal{M}^n}) \\
&\le \frac{1}{1-\gamma}\Bigg(\sum_{t=0}^{T-1}\mathbb{E}_{(s,a)\sim\tilde{d}_{\mathcal{M}^n}}\left[\widehat{A}^t_{b^n}(s,a)\mathbf{1}\{s\in\mathcal{K}^n\}\right] + \sum_{t=0}^{T-1}\mathbb{E}_{(s,a)\sim\tilde{d}_{\mathcal{M}^n}}\left[A^t_{b^n}(s,a) - \widehat{A}^t_{b^n}(s,a))\mathbf{1}\{s\in\mathcal{K}^n\}\right]\Bigg) \\
&\le \frac{1}{1-\gamma}\Bigg(12W\sqrt{\log(|\mathcal{A}|)T} + \sum_{t=0}^{T-1}\mathbb{E}_{(s,a)\sim\tilde{d}_{\mathcal{M}^n}}\left[\big(A^t_{b^n}(s,a) - \widehat{A}^t_{b^n}(s,a)\big)\mathbf{1}\{s\in\mathcal{K}^n\}\right]\Bigg).
\end{aligned}
$$

■

Since the definition of $\widehat{A}^t_{b^n}$ is the same as the one for SPI-SAMPLE, Lemma C.3 and Lemma C.4 are directly applied. In total, we have the following theorem for the sample complexity of ENIAC-SPI-COMPUTE.

Theorem C.2 (Main Result: Sample Complexity of ENIAC-SPI-COMPUTE). *Let $\delta\in(0,1)$ and $\varepsilon\in(0,1/(1-\gamma))$. With Assumptions C.1, C.2, 4.2, and 4.3, we set the hyperparameters as:*

$$
\begin{aligned}
&\beta = \frac{\varepsilon(1-\gamma)}{2},\ T = \frac{144W^2\cdot\log|\mathcal{A}|}{\varepsilon^2(1-\gamma)^2},\ N \ge \frac{384W^3|\mathcal{A}|\log(|\mathcal{A}|)\cdot \mathit{dim}_E(\mathcal{F},\beta)}{\varepsilon^4(1-\gamma)^4},\ \eta = \sqrt{\frac{\log(|\mathcal{A}|)}{16W^2T}}, \\
&\alpha = \frac{1}{1+\sqrt{T}},\ \epsilon_1 = \frac{(1-\gamma)^4\varepsilon^4}{1536W^2|\mathcal{A}|\log(|\mathcal{A}|)\cdot\mathit{dim}_E(\mathcal{F},\beta)},\ \epsilon_2 = \frac{(1-\gamma)^4\varepsilon^4}{1248CW^3|\mathcal{A}|\log(|\mathcal{A}|)\mathit{dim}_E(\mathcal{F},\beta)}, \\
&K = \frac{1536W^3|\mathcal{A}|^2(\log(|\mathcal{A}|))^2\cdot\mathit{dim}_E(\mathcal{F},\beta)\cdot\big(\log(\frac{3NT\cdot\mathcal{N}(\mathcal{F},\epsilon_1)}{\delta})\big)^2\cdot\log(\frac{6NT}{\delta})}{\varepsilon^4(1-\gamma)^4}, \\
&M = \frac{48000C^2W^5|\mathcal{A}|\log(|\mathcal{A}|)\mathit{dim}_E(\mathcal{F},\beta)\log(\frac{3NT\cdot\mathcal{N}(\mathcal{F},\epsilon_2)}{\delta})}{\varepsilon^4(1-\gamma)^4},
\end{aligned}
$$

and ϵ satisfies Equation (24) correspondingly. Then with probability at least $1-\delta$, for the average policy $\pi^N_{ave} := \mathrm{Unif}(\pi^2,\dots,\pi^{N+1})$, we have

$$
V^{\pi^N_{ave}} \ge V^{\tilde{\pi}} - \frac{4\sqrt{|\mathcal{A}|\epsilon_{bias}}}{1-\gamma} - \epsilon_0\cdot\frac{200CW\cdot|\mathcal{A}|\log(|\mathcal{A}|)\cdot\mathit{dim}_E(\mathcal{F},\beta)}{\varepsilon^3(1-\gamma)^4} - 9\varepsilon
$$

for any comparator $\tilde{\pi}$ with total number of samples:

$$
\tilde{\mathcal{O}}\Big(\frac{C^2W^{10}\cdot|\mathcal{A}|^2\cdot\big(\mathit{dim}_E(\mathcal{F},\beta)\big)^2\cdot\big(\log(\mathcal{N}(\mathcal{F},\epsilon'))\big)^2}{\varepsilon^{10}(1-\gamma)^{10}}\Big),
$$

where $\epsilon' = \min(\epsilon_1,\epsilon_2)$.

Corollary 2. *If Assumption 4.1 holds, with proper hyperparameters, the average policy $\pi^N_{ave} := \mathrm{Unif}(\pi^2,\dots,\pi^{N+1})$ of ENIAC-SPI-COMPUTE achieves $V^{\pi^N_{ave}} \ge V^{\tilde{\pi}} - \varepsilon$ with probability at least $1-\delta$ and total number of samples:*

$$
\tilde{\mathcal{O}}\Big(\frac{W^{10}\cdot|\mathcal{A}|^2\cdot\big(\mathit{dim}_E(\mathcal{F},\beta)\big)^2\cdot\big(\log(\mathcal{N}(\mathcal{F},\epsilon'))\big)^2}{\varepsilon^{10}(1-\gamma)^{10}}\Big).
$$

D. Analysis of ENIAC-NPG

In this section, we provide the sample complexity of ENIAC-NPG-SAMPLE. For ENIAC-NPG-COMPUTE, it can be adapted from ENIAC-SPI-COMPUTE and ENIAC-NPG-SAMPLE.

The analysis of ENIAC-NPG-SAMPLE is in parallel to that of ENIAC-SPI-SAMPLE. As before, we provide a general result which considers model misspecification and Theorem 4.2 falls as a special case under the closedness Assumption 4.4.

We simplify the notation as π_θ for $\pi_{f_\theta}(a|s) := \frac{\exp(f_\theta(s,a))}{\sum_{a'}\exp(f_\theta(s,a'))}$. Then for epoch n iteration t in ENIAC-NPG-SAMPLE,

$$
\pi^n_t(\cdot|s) = \begin{cases}\pi_{\theta^n_t}(\cdot|s), & s\in\mathcal{K}^n \\ \mathrm{Unif}(\{a\in\mathcal{A}:(s,a)\notin\mathcal{K}^n\}), & o.w.\end{cases}
$$

We state the following assumptions to quantify the misspecification error.

Assumption D.1 (Bounded Transfer Error)**.** *Given a target function $g : \mathcal{S} \times \mathcal{A} \to \mathbb{R}$, we define the critic loss function $L(u; d, g, \pi_\theta)$ with $d \in \Delta(\mathcal{S} \times \mathcal{A})$ as:*

$$L(u; d, g, \pi_\theta) := \mathbb{E}_{(s,a)\sim d}\left[(u^\top \nabla_\theta \log \pi_\theta - g)^2\right].$$

For the fixed comparator policy $\tilde{\pi}$ as mentioned in Section B.1, we define a state-action distribution $\tilde{d}(s,a) := d_{s_0}^{\tilde{\pi}}(s) \circ \mathrm{Unif}(\mathcal{A})$. In ENIAC-NPG-SAMPLE, for every epoch $n \in [N]$ and every iteration t inside epoch n, we assume that

$$\inf_{u \in \mathcal{U}_t^n} L(u; \tilde{d}, A_{b^n}^t - \bar{b}_t^n, \pi_{\theta_t^n}) \leq \epsilon_{bias},$$

where $\mathcal{U}_t^n := \mathrm{argmin}_{u \in \mathcal{U}} L(u; \rho_{cov}^n, A_{b^n}^t - \bar{b}_t^n, \pi_{\theta_t^n})$ and $\epsilon_{bias} \geq 0$ is a problem-dependent constant.

Recall that $\left(A_{b^n}^t - \bar{b}_t^n\right)(s,a) = Q_{b^n}^t(s,a) - b^n(s,a) - \mathbb{E}_{a\sim\pi_t^n(\cdot|s)}[Q_{b^n}^t(s,a) - b^n(s,a)]$. As before, we denote by $\tilde{u}_t^n$ a particular vector in $\mathcal{U}_t^n$ such that $L(\tilde{u}_t^n; \tilde{d}, A_{b^n}^t - \bar{b}_t^n, \pi_{\theta_t^n}) \leq 2\epsilon_{\text{bias}}$. Note that we use $\nabla_\theta \log \pi_{\theta_t^n}$ as the linear features for critic fit at iteration t epoch n, even though π_t^n is not the same as $\pi_{\theta_t^n}$. Nevertheless, we show later that this choice of features is sufficient for good critic fitting on the known states, where we measure our critic error.

Remark 4. *Under the closedness condition Assumption 4.4,*

$$\begin{aligned} A_{b^n}^t(s,a) - \bar{b}^n(s,a) &= Q_{b^n}^t(s,a) - b^n(s,a) - \mathbb{E}_{a'\sim\pi_t^n}(Q_{b^n}^t - b^n(s,a')) \\ &= \mathbb{E}^{\pi_t^n}[r(s,a) + \gamma Q_{b^n}^t(s',a')] - \mathbb{E}_{a'\sim\pi_t^n}[\mathbb{E}^{\pi_t^n}[r(s,a') + \gamma Q_{b^n}^t(s'',a'')]] \\ &\in \mathcal{G}_{f_{\theta_t^n}}, \end{aligned}$$

where the last step follows, since π_t^n can be described as $\pi_{\theta_t^n, \mathcal{K}^n}$ under the notation of Assumption 4.4, whence the containment of $\mathcal{G}_{f_{\theta_t^n}}$ follows. Thus, there exists a vector $u \in \mathcal{U}$ such that $u^\top \nabla \log \pi_{f_{\theta_t^n}} = A_{b^n}^t - \bar{b}^n$ everywhere. We can then take ϵ_{bias} as 0 and $\tilde{u}_t^n = u$. Assumption D.1 therefore is a generalized version of the closedness condition.

For NPG, the loss function L is convex in the parameters u since the features are fixed for every individual iteration. As a result, we naturally have an inequality as in Assumption C.2 for SPI. We present it in the lemma below, which essentially follows a similar result for the linear case in Agarwal et al. (2020a).

Lemma D.1. *For the same loss function L as defined in Assumption D.1, it holds that*

$$\begin{aligned} &\mathbb{E}_{(s,a)\sim\rho_{cov}^n}\left[\left((u_t^n - \tilde{u}_t^n)^\top \nabla_\theta \log \pi_{\theta_t^n}\right)^2\right] \\ \leq &L(u_t^n; \rho_{cov}^n, A_{b^n}^t - \bar{b}_t^n, \pi_{\theta_t^n}) - L(\tilde{u}_t^n; \rho_{cov}^n, A_{b^n}^t - \bar{b}_t^n, \pi_{\theta_t^n}). \end{aligned}$$

Proof. For the left-hand side, we have that

$$\begin{aligned} &\mathbb{E}_{(s,a)\sim\rho_{\text{cov}}^n}\left[\left((u_t^n)^\top \nabla_\theta \log \pi_{\theta_t^n} - (\tilde{u}_t^n)^\top \nabla_\theta \log \pi_{\theta_t^n}\right)^2\right] \\ =&\mathbb{E}_{(s,a)\sim\rho_{\text{cov}}^n}\left[\left((u_t^n)^\top \nabla_\theta \log \pi_{\theta_t^n} + \bar{b}_t^n - A_{b^n}^t\right)^2\right] - \mathbb{E}_{(s,a)\sim\rho_{\text{cov}}^n}\left[\left((\tilde{u}_t^n)^\top \nabla_\theta \log \pi_{\theta_t^n} + \bar{b}_t^n - A_{b^n}^t\right)^2\right] \\ &- 2\mathbb{E}_{(s,a)\sim\rho_{\text{cov}}^n}\left[\left((u_t^n)^\top \nabla_\theta \log \pi_{\theta_t^n} - (\tilde{u}_t^n)^\top \nabla_\theta \log \pi_{\theta_t^n}\right) \cdot \left((\tilde{u}_t^n)^\top \nabla_\theta \log \pi_{\theta_t^n} + \bar{b}_t^n - A_{b^n}^t\right)\right] \end{aligned}$$

Since $\tilde{u}_t^n$ is a minimizer. By first-order optimality condition, the cross term is greater or equal to 0. The desired result is obtained. ■

D.1. Sample Complexity of ENIAC-NPG-SAMPLE

We follow the same steps as listed in B.2 and start with the bonus bound.

Lemma D.2 (NPG-SAMPLE: The Bound of Bonus)**.** *With probability at least $1 - N\delta$,*

$$\sum_{n=1}^{N} V_{b^n}^{\pi^{n+1}} - V^{\pi^{n+1}} \leq \frac{2\epsilon^2 + 32G^2B^2K + \beta^2}{(1-\gamma)\beta^2 K} \cdot dim_E(\mathcal{G}_\mathcal{F}, \beta) + \frac{N}{1-\gamma}\sqrt{\frac{\log(2/\delta)}{2K}}.$$

The proof is similar to Lemma C.1. The only thing changed is the function approximation space. Thus we have $\dim_E(\mathcal{G}_\mathcal{F}, \beta)$ instead of $\dim_E(\mathcal{F}, \beta)$ and $\|g_\theta^u\|_\infty \leq 2GB, \forall g_\theta^u \in \mathcal{G}_\mathcal{F}$.

Next, we establish the convergence result of NPG update. We focus on a specific episode n and for each iteration t, we define

$$\widehat{A}_{b^n}^t(s,a) := u_t^\top \nabla f_{\theta_t}(s,a) + b^n - \mathbb{E}_{a'\sim\pi_{\theta_t}(\cdot|s)}[u_t^\top \nabla f_{\theta_t}(s,a') + b^n(s,a')]. \tag{35}$$

Since $\pi_t(\cdot|s) = \pi_{\theta_t}(\cdot|s)$ for $s \in \mathcal{K}^n$, $\mathbb{E}_{a' \sim \pi_t(\cdot|s)}[\widehat{A}^t_{b^n}(s,a')] = 0$ for $s \in \mathcal{K}^n$.

From the algorithm we can see that $\widehat{A}^t_{b^n}$ is indeed our approximation to the real advantages $A^t_{b^n}$. In contrary to ENIAC-SPI, the actor update in ENIAC-NPG does not use $\widehat{A}^t_{b^n}$ directly but by modifying the parameter θ. In the next lemma, we show how to link the NPG update to a formula of $\widehat{A}^t_{b^n}$ and eventually are able to bound the policy sub-optimality with function approximation error.

Lemma D.3 (NPG-SAMPLE: Convergence)**.** *In ENIAC-NPG-SAMPLE, let $\widehat{A}^t_{b^n}$ be as defined in Equation (35) and $\eta = \sqrt{\frac{\log(|\mathcal{A}|)}{(16D^2+\Lambda B^2)T}}$. For any epoch $n \in [N]$, NPG-SAMPLE obtains a sequence of policies $\{\pi_t\}_{t=0}^{T-1}$ such that when comparing to $\tilde{\pi}$:*

$$\begin{aligned}
&\frac{1}{T}\sum_{t=0}^{T-1}(V^{\tilde{\pi}^n}_{\mathcal{M}^n} - V^t_{b^n}) = \frac{1}{T}\sum_{t=0}^{T-1}(V^{\tilde{\pi}^n}_{\mathcal{M}^n} - V^t_{\mathcal{M}^n})\\
&\le \frac{1}{1-\gamma}\left(2\sqrt{\frac{\log(|\mathcal{A}|)(16D^2+\Lambda B^2)}{T}} + \frac{1}{T}\sum_{t=0}^{T-1}\mathbb{E}_{(s,a)\sim \tilde{d}_{\mathcal{M}^n}}\left[\left(A^t_{b^n}(s,a) - \widehat{A}^t_{b^n}(s,a)\right)\mathbf{1}\{s\in\mathcal{K}^n\}\right]\right).
\end{aligned}$$

Proof. For the same reason as in Lemma C.2, we have

$$V^{\tilde{\pi}^n}_{\mathcal{M}^n} - V^t_{\mathcal{M}^n} \le \frac{1}{1-\gamma}\sum_{(s,a)}\tilde{d}_{\mathcal{M}^n}(s,a)A^t_{b^n}(s,a)\mathbf{1}\{s\in\mathcal{K}^n\}. \tag{36}$$

We focus on on $s \in \mathcal{K}^n$. Then $\pi_t(\cdot|s) \propto \exp(f_{\theta_t}(s,\cdot))$ and $b(s,\cdot) = 0$. It holds that

$$\begin{aligned}
&\mathbf{KL}(\tilde{\pi}^n(\cdot|s), \pi_{t+1}(\cdot|s)) - \mathbf{KL}(\tilde{\pi}^n(\cdot|s), \pi_t(\cdot|s))\\
&= -\mathbb{E}_{a\sim\tilde{\pi}^n(\cdot|s)}\left[f_{\theta_{t+1}}(s,a) - f_{\theta_t}(s,a)\right] + \log\frac{\sum_a \exp(f_{\theta_{t+1}}(s,a))}{\sum_a \exp(f_{\theta_t}(s,a))}\\
&\le -\mathbb{E}_{a\sim\tilde{\pi}^n(\cdot|s)}[\eta\cdot u_t^\top \nabla_\theta f_{\theta_t} - \eta^2\frac{\Lambda B^2}{2}] + \log\frac{\sum_a \exp(f_{\theta_t}(s,a) + \eta\cdot u_t^\top\nabla_\theta f_{\theta_t} + \eta^2\Lambda B^2/2)}{\sum_a \exp(f_{\theta_t}(s,a))}\\
&= -\eta\cdot\mathbb{E}_{a\sim\tilde{\pi}^n(\cdot|s)}[\widehat{A}^t_{b^n}(s,a)] - \eta\cdot\mathbb{E}_{a'\sim\pi_t(\cdot|s)}u_t^\top\nabla_\theta f_{\theta_t}(s,a')\\
&\quad + \log\left(\sum_a \pi_t(s,a)\exp\left(\eta\cdot\widehat{A}^t_{b^n}(s,a) + \eta\cdot\mathbb{E}_{a'\sim\pi_t(\cdot|s)}u_t^\top\nabla_\theta f_{\theta_t}\right)\right) + \eta^2\Lambda B^2\\
&= -\mathbb{E}_{a\sim\tilde{\pi}^n(\cdot|s)}[\eta\widehat{A}^t_{b^n}(s,a)] + \log\left(\sum_a \pi_t(a|s)\exp\left(\eta\widehat{A}^t_{b^n}(s,a)\right)\right) + \eta^2\Lambda B^2.
\end{aligned}$$

where the inequality is by Taylor expansion and the regularity assumption 4.5:

$$f_{\theta_t} + (\theta_{t+1}-\theta_t)^\top\nabla_\theta f_{\theta_t} - \frac{\Lambda}{2}\|\theta_{t+1}-\theta_t\|_2^2 \le f_{\theta_{t+1}} \le f_{\theta_t} + (\theta_{t+1}-\theta_t)^\top\nabla_\theta f_{\theta_t} + \frac{\Lambda}{2}\|\theta_{t+1}-\theta_t\|_2^2.$$

Since $|\widehat{A}^t_{b^n}(s,a)| \le 4D$ and $\eta \le 1/(4D)$ when $T > \log(|\mathcal{A}|)$, $\eta\widehat{A}^t_{b^n}(s,a) \le 1$. By the inequality that $\exp(x) \le 1 + x + x^2$ for $x \le 1$, we have that

$$\begin{aligned}
&\log\left(\sum_a \pi_t(a|s)\exp\left(\eta\widehat{A}^t_{b^n}(s,a)\right)\right)\\
\le&\log\left(1 + \mathbb{E}_{a\sim\pi_t(\cdot|s)}[\eta\widehat{A}^t_{b^n}(s,a)] + 16\eta^2D^2\right) \le 16\eta^2D^2.
\end{aligned}$$

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

$$\mathbf{KL}(\tilde{\pi}^n(\cdot|s), \pi_{t+1}(\cdot|s)) - \mathbf{KL}(\tilde{\pi}^n(\cdot|s), \pi_t(\cdot|s)) \le -\eta\mathbb{E}_{a\sim\tilde{\pi}^n(\cdot|s)}[\widehat{A}^t_{b^n}(s,a)] + \eta^2(16D^2+\Lambda B^2).$$

Adding both sides from $t = 0$ to $T-1$ and taking $\eta = \sqrt{\frac{\log(|\mathcal{A}|)}{(16D^2+\Lambda B^2)T}}$, we get

$$\begin{aligned}
&\sum_{t=0}^{T-1}\mathbb{E}_{(s,a)\sim\tilde{d}_{\mathcal{M}^n}}\left[\widehat{A}^t_{b^n}(s,a)\mathbf{1}\{s\in\mathcal{K}^n\}\right]\\
\le&\frac{1}{\eta}\mathbb{E}_{s\sim\tilde{d}_{\mathcal{M}^n}}\left[\left(\mathbf{KL}(\tilde{\pi}^n(\cdot|s), \pi_0(\cdot|s)) - \mathbf{KL}(\tilde{\pi}^n(\cdot|s), \pi_T(\cdot|s))\right)\mathbf{1}\{s\in\mathcal{K}^n\}\right] + \eta T(16D^2+\Lambda B^2)\\
\le&\log(|\mathcal{A}|)/\eta + \eta T(16D^2+\Lambda B^2) \le 2\sqrt{\log(|\mathcal{A}|)\cdot(16D^2+\Lambda B^2)\cdot T}.
\end{aligned}$$

Combining with Equation (36), the regret on $\mathcal{M}^n$ satisfies

$$\begin{aligned}
&\sum_{t=0}^{T-1}(V_{\mathcal{M}^n}^{\tilde{\pi}^n} - V_{\mathcal{M}^n}^t) \\
&\leq \frac{1}{1-\gamma}\sum_{t=0}^{T-1}\mathbb{E}_{(s,a)\sim\tilde{d}_{\mathcal{M}^n}}\left[\widehat{A}_{b^n}^t(s,a)\mathbf{1}\{s\in\mathcal{K}^n\}\right] + \frac{1}{1-\gamma}\sum_{t=0}^{T-1}\mathbb{E}_{(s,a)\sim\tilde{d}_{\mathcal{M}^n}}\left[A_{b^n}^t(s,a) - \widehat{A}_{b^n}^t(s,a))\mathbf{1}\{s\in\mathcal{K}^n\right] \\
&\leq \frac{1}{1-\gamma}\left(2\sqrt{\log(|\mathcal{A}|)(16D^2+\Lambda B^2)T} + \sum_{t=0}^{T-1}\mathbb{E}_{(s,a)\sim\tilde{d}_{\mathcal{M}^n}}\left[\big(A_{b^n}^t(s,a) - \widehat{A}_{b^n}^t(s,a)\big)\mathbf{1}\{s\in\mathcal{K}^n\}\right]\right).
\end{aligned}$$

■

Next, we establish two lemmas to bound the difference between the true advantage $A_{b^n}^t(s,a)$ and the approximation $\widehat{A}_{b^n}^t(s,a)$.

Lemma D.4 (Approximation Bound)**.** *At epoch n, assume for all $0 \leq t \leq T-1$,*

$$L(u_t^n;\rho_{cov}^n, A_{b^n}^t - \bar{b}_t^n, \pi_{\theta_t^n}) \leq L(\tilde{u}_t^n;\rho_{cov}^n, A_{b^n}^t - \bar{b}_t^n, \pi_{\theta_t^n}) + \epsilon_{stat},$$

where $\epsilon_{stat} > 0$ is to be determined later, and

$$\epsilon^2 = NK\big(\epsilon_{stat} + 16D\epsilon_1\big) + 8D^2\log(\mathcal{N}(\mathcal{G}_{\mathcal{F}},\epsilon_1)/\delta)\cdot\sqrt{NK}, \tag{37}$$

where ϵ is used in bonus function design (see Section 3.3) and ϵ_1 is to be determined. Under Assumption D.1 and 4.5, we have that for every $0 \leq t \leq T-1$, with probability at least $1-(n+1)\delta$,

$$\mathbb{E}_{(s,a)\sim\tilde{d}_{\mathcal{M}^n}}\Big(A_{b^n}^t(s,a) - \widehat{A}_{b^n}^t(s,a)\Big) \leq 4\sqrt{|\mathcal{A}|\epsilon_{bias}} + 2\beta.$$

Lemma D.5. *Following the same notation as in Lemma D.4, it holds with probability at least $1-\delta$ that*

$$L(u_t^n;\rho_{cov}^n, A_{b^n}^t - \bar{b}_t^n, \pi_{\theta_t^n}) - L(\tilde{u}_t^n;\rho_{cov}^n, A_{b^n}^t - \bar{b}_t^n, \pi_{\theta_t^n}) \leq \frac{500D^4\cdot d\log\left(\frac{6D}{\epsilon_2\delta}\right)}{M} + 13D^2\cdot\epsilon_2,$$

where d is the linear dimension of u.

The proofs of the above lemmas can be easily adapted from Lemma C.3 or Lemma C.4 by replacing f_t with $u_t^\top\nabla f_{\theta_t}$, $\tilde{f}_t^n$ with $(\tilde{u}_t^n)^\top\nabla f_{\theta_t}$, and $\mathcal{F}$ with $\mathcal{G}_{\mathcal{F}}$. In particular, for Lemma D.5, since the linear feature is fixed for critic fit at iteration t epoch n, the function cover is defined on the space $\mathcal{G}_{f_{\theta_t^n}}$. By Lemma E.2, the covering number is therefore represented with the linear dimension of u, d.

In the following, we present the detailed form of the sample complexity of NPG-SAMPLE.

Theorem D.1 (Main Result: Sample Complexity of ENIAC-NPG-SAMPLE)**.** *Let $\delta \in (0,1)$ and $\varepsilon \in (0, 1/(1-\gamma))$. With Assumptions D.1 and 4.5, we set the hyperparameters as:*

$$\begin{aligned}
&\beta = \frac{\varepsilon(1-\gamma)}{2}, T = \frac{64(D^2+\Lambda B^2)\cdot\log|\mathcal{A}|}{\varepsilon^2(1-\gamma)^2}, N \geq \frac{128B^2G^2\cdot dim_E(\mathcal{G}_{\mathcal{F}},\beta)}{\varepsilon^3(1-\gamma)^3}, \eta = \sqrt{\frac{\log(|\mathcal{A}|)}{(16D^2+\Lambda B^2)T}} \\
&\epsilon_1 = \frac{(1-\gamma)^3\varepsilon^3}{128D\cdot dim_E(\mathcal{G}_{\mathcal{F}},\beta)},\;\; K = \frac{32D^2\cdot dim_E(\mathcal{G}_{\mathcal{F}},\beta)\cdot\big(\log(\frac{3NT\cdot\mathcal{N}(\mathcal{G}_{\mathcal{F}},\epsilon_1)}{\delta})\big)^2\cdot\log(\frac{6NT}{\delta})}{\varepsilon^3(1-\gamma)^3}, \\
&\epsilon_2 = \frac{(1-\gamma)^3\varepsilon^3}{110D^2\cdot dim_E(\mathcal{G}_{\mathcal{F}},\beta)},\;\; M = \frac{4000D^4\cdot dim_E(\mathcal{G}_{\mathcal{F}},\beta)\cdot d\log(\frac{18DNT}{\epsilon_2\delta})}{\varepsilon^3(1-\gamma)^3},
\end{aligned}$$

and ϵ satisfies Equation (37) correspondingly. Then with probability at least $1-\delta$, for the average policy $\pi_{ave}^N :=$ Unif$(\pi^2,\ldots,\pi^{N+1})$, we have

$$V^{\pi_{ave}^N} \geq V^{\tilde{\pi}} - \frac{4\sqrt{|\mathcal{A}|\epsilon_{bias}}}{1-\gamma} - 9\varepsilon$$

for any comparator $\tilde{\pi}$ with total number of samples:

$$\tilde{\mathcal{O}}\Big(\frac{D^6(D^2+\Lambda B^2)\cdot\big(dim_E(\mathcal{G}_{\mathcal{F}},\beta)\big)^2\cdot\big(\log(\mathcal{N}(\mathcal{G}_{\mathcal{F}},\epsilon'))\big)^2}{\varepsilon^8(1-\gamma)^8}\Big),$$

where $\epsilon' = \min(\epsilon_1,\epsilon_2)$ such that $\log(\mathcal{N}(\mathcal{G}_{\mathcal{F}},\epsilon')) = \Omega(d)$.

The proof is similar to that of Theorem C.1. We also have the following result when the closedness assumption is satisfied.

Corollary 3. *If Assumption 4.4 holds, with proper hyperparameters, the average policy $\pi_{ave}^N := \mathrm{Unif}(\pi^2, \ldots, \pi^{N+1})$ of ENIAC-NPG-SAMPLE achieves $V^{\pi_{ave}^N} \geq V^{\tilde{\pi}} - \varepsilon$ with probability at least $1-\delta$ and total number of samples:*

$$\tilde{\mathcal{O}}\Big(\frac{D^6(D^2+\Lambda B^2)\cdot \big(dim_E(\mathcal{G}_{\mathcal{F}},\beta)\big)^2 \cdot \big(\log(\mathcal{N}(\mathcal{G}_{\mathcal{F}},\epsilon'))\big)^2}{\varepsilon^8(1-\gamma)^8}\Big)$$

Note that under Assumption 4.4, as mentioned in Remark 4, $\epsilon_{\text{bias}} = 0$.

E. Auxiliary Lemmas

Lemma E.1. *Given a function class $\mathcal{F}$, for its covering number, we have $\mathcal{N}(\Delta\mathcal{F}, \epsilon) \leq \mathcal{N}(\mathcal{F}, \epsilon/2)^2$.*

Proof. Let $\Delta\mathcal{C}(\mathcal{F}, \epsilon/2) := \{f - f' | f, f' \in \mathcal{C}(\mathcal{F}, \epsilon/2)\}$. Then $\Delta\mathcal{C}(\mathcal{F}, \epsilon/2)$ is an ϵ-cover for $\Delta\mathcal{F}$ and $|\Delta\mathcal{C}(\mathcal{F}, \epsilon/2)| \leq |\mathcal{C}(\mathcal{F}, \epsilon/2)|^2 \leq \mathcal{N}(\mathcal{F}, \epsilon/2)^2$. ■

Lemma E.2. *Given $f \in \mathcal{F}$, under the regularity Assumption 4.5, we have that the covering number of the linear class $\mathcal{G}_f := \{u^\top \nabla_\theta \log \pi_f, u \in \mathcal{U} \subset \mathbb{R}^d, f \in \mathcal{F}\}$ achieves $\mathcal{N}(\mathcal{G}_f, \epsilon) \leq \big(\frac{3D}{\epsilon}\big)^d$.*

Proof. In order to construct a cover set of $\mathcal{G}_f$ with radius ϵ_2, we need that for any $u \in \mathcal{U} \subset \mathbb{R}^d$, there exist a $\tilde{u}$, such that

$$\|u^\top \nabla_\theta \log \pi_f(s,a) - \tilde{u}^\top \nabla_\theta \log \pi_f(s,a)\|_\infty \leq \epsilon_2.$$

where the infinity norm is taken over all $(s,a) \in \mathcal{S} \times \mathcal{A}$. By Cauchy-Schwarz inequality, we have

$$\|u^\top \nabla_\theta \log \pi_f - \tilde{u}^\top \nabla_\theta \log \pi_f\|_\infty = \|(u-\tilde{u})^\top \nabla_\theta \log \pi_f\|_\infty \leq 2G\|u - \tilde{u}\|_2.$$

Thus, it is enough to have $\|u - \tilde{u}\|_2 \leq \epsilon_2/(2G)$, which is equivalent to cover a ball in $\mathbb{R}^d$ with radius B (recall that $\|u\| \leq B$) with small balls of radius $\epsilon_2/(2G)$. The latter has a covering number bounded by $\left(\frac{6BG}{\epsilon_2}\right)^d \leq \left(\frac{6D}{\epsilon_2}\right)^d$[7]. ■

F. Algorithm Hyperparameters

In this section, we present more details about the implementation in our experiments. All algorithms were based on the PPO implementation of (Shangtong, 2018). The network structure is described in the main body and the last layer outputs the parameters of a 1D Gaussian for action selection.

The width training process is presented in Algorithm 5. Recall that our training loss is

$$\sum_{(s,a)\in\mathcal{Z}_Q^n} \frac{\lambda\big(f(s,a) - f'(s,a)\big)^2}{|\mathcal{Z}_Q^n|} - \sum_{(s',a')\in\mathcal{Z}^n} \frac{\big(f(s',a') - f'(s',a')\big)^2}{|\mathcal{Z}^n|} - \sum_{(s,a)\in\mathcal{Z}_Q^n} \frac{\lambda_1\big(f(s,a) - f'(s,a)\big)}{|\mathcal{Z}_Q^n|}. \tag{38}$$

To stabilize training, for each iteration we sample a minibatch $\mathcal{D}_Q$ from the query batch, then run several steps of stochastic gradient descent with changing minibatches on $\mathcal{Z}^n$ while fixing $\mathcal{D}_Q$. The hyperparameters for width training are listed in Table 1.

For PC-PG, we follow the same implementation as mentioned in (Agarwal et al., 2020a); for PPO-RND, the RND network has the same architecture as the policy network, except that the last linear layer mapping hidden units to actions is removed. We found that tuning the intrinsic reward coefficient was important for getting good performance for RND. The hyperparameters for optimization are listed in Table 2 and 3.

[7] The covering number of Euclidean balls can be easily found in literature.

Algorithm 5 Width Training in ENIAC

```
Input: Replay buffer $\mathcal{Z}^n$, query batch $\mathcal{Z}_Q^n$.
Initialize $f$ with the same network structure as the critic.
Copy $f'$ as $f$ and fix $f'$ during training.
for $i = 1$ to $I$ do
    Sample a minibatch $\mathcal{D}_Q$ from $\mathcal{Z}_Q^n$
    for $j = 1$ to $J$ do
        Sample a minibatch $\mathcal{D}_j$ from $\mathcal{Z}^n$
        Do one step of gradient descent on $f$ with loss in Equation (38) and $\mathcal{D}_Q$ and $\mathcal{D}_j$.
    end for
end for
Output: $w^n := |f - f'|$
```

Table 1. ENIAC Width Training Hyperparameters

Hyperparameter	2-layer	4-layer	6-layer
λ	0.1	0.1	0.1
λ_1	0.01	0.01	0.01
$\|Z_Q\|$	20000	20000	20000
Learning Rate	0.001	0.001	0.0015
$\|\mathcal{D}_j\|$	160	160	160
$\|\mathcal{D}_Q\|$	20	20	10
Gradient Clippling	5.0	5.0	5.0
I	1000	1000	1000
J	10	10	10

Table 2. ENIAC/PC-PG Optimization Hyperparameters

Hyperparameter	Values Considered	2-layer	4-layer	6-layer
Learning Rate	$e^{-3}, 5e^{-4}, e^{-4}$	$5e^{-4}$	$5e^{-4}$	$5e^{-4}$
τ_{GAE}	0.95	0.95	0.95	0.95
Gradient Clippling	0.5, 1, 2, 5	5.0	5.0	5.0
Entropy Bonus	0.01	0.01	0.01	0.01
PPO Ratio Clip	0.2	0.2	0.2	0.2
PPO Minibatch	160	160	160	160
PPO Optimization Epochs	5	5	5	5
ϵ-greedy sampling	0, 0.01, 0.05	0.05	0.05	0.05

Table 3. PPO-RND Hyperparameters

Hyperparameter	Values Considered	2-layer	4-layer	6-layer
Learning Rate	$e^{-3}, 5e^{-4}, e^{-4}$	e^{-4}	e^{-4}	e^{-4}
τ_{GAE}	0.95	0.95	0.95	0.95
Gradient Clippling	5.0	5.0	5.0	5.0
Entropy Bonus	0.01	0.01	0.01	0.01
PPO Ratio Clip	0.2	0.2	0.2	0.2
PPO Minibatch	160	160	160	160
PPO Optimization Epochs	5	5	5	5
Intrinsic Reward Normalization	true, false	false	false	false
Intrinsic Reward Coefficient	0.5, 1, $e, e^2, e^3, 5e^3, e^4$	$5e^3$	e^3	e^3