Cautiously Optimistic Policy Optimization and Exploration with Linear Function Approximation

Andrea Zanette ZANETTE@STANFORD.EDU
Stanford University

Ching-An Cheng CHINGANC@MICROSOFT.COM
Microsoft Research

Alekh Agarwal ALEKHA@MICROSOFT.COM
Microsoft Research

Editors: Mikhail Belkin and Samory Kpotufe

Abstract

Policy optimization methods are popular reinforcement learning algorithms, because their incremental and on-policy nature makes them more stable than the value-based counterparts. However, the same properties also make them slow to converge and sample inefficient, as the on-policy requirement precludes data reuse and the incremental updates couple large iteration complexity into the sample complexity. These characteristics have been observed in experiments as well as in theory in the recent work of Agarwal et al. (2020a), which provides a policy optimization method PC-PG that can robustly find near optimal polices for approximately linear Markov decision processes but suffers from an extremely poor sample complexity compared with value-based techniques.

In this paper, we propose a new algorithm, COPOE, that overcomes the sample complexity issue of PC-PG while retaining its robustness to model misspecification. Compared with PC-PG, COPOE makes several important algorithmic enhancements, such as enabling data reuse, and uses more refined analysis techniques, which we expect to be more broadly applicable to designing new reinforcement learning algorithms. The result is an improvement in sample complexity from $\widetilde{O}(1/\epsilon^{11})$ for PC-PG to $\widetilde{O}(1/\epsilon^{3})$ for COPOE, nearly bridging the gap with value-based techniques.

Keywords: Exploration, Optimization, Reinforcement Learning, Natural Policy Gradient, Mirror Descent, Importance Sampling, Sample Complexity

1. Introduction

In building real-world learning systems, it is desirable to have algorithms that possess strong sample complexity guarantees under favorable model assumptions, while being robust to model misspecification. This need of robust adaptivity is particularly crucial to reinforcement learning (RL) applications, where we do not have the luxury of tuning our modeling choices through repeated experimentation with a static dataset.

Nonetheless, the intertwined complexity of credit assignment and exploration inherent to RL makes designing such an algorithm challenging. Most provably efficient RL algorithms with function approximations (Yang and Wang, 2020; Jin et al., 2020; Zanette et al., 2020c; Ayoub et al., 2020; Zhou et al., 2020a; Jiang et al., 2017) require certain structural assumptions on the environment's regularity in order to provide sample complexity guarantees. These conditions are in some sense necessary, especially for high dimensional problems; otherwise, the learner in the worst case would require exponentially many samples before discovering any useful information (see e.g. (Kakade et al., 2003; Krishnamurthy et al., 2016; Weisz et al., 2020)). However, these provably efficient

RL algorithms are typically not robust to model misspecification, because their performance guarantees allow for only small ℓ_∞-bounded perturbations from their assumptions. Can we design RL algorithms that offer strong performance guarantees when the model assumption holds and degrade more gracefully with model misspecification, such as according to *average* errors?

In this paper, we study this question in the context of policy optimization methods in the exploration setting of an approximately linearly parametrizable Markov decision process (MDP) model, which includes approximately linear or low-rank MDPs studied for example by Yang and Wang (2020); Jin et al. (2020); Zanette et al. (2020a). Policy optimization methods are some of the most classical (Williams, 1992; Sutton et al., 1999; Konda and Tsitsiklis, 2000; Kakade, 2001) as well as widely used approaches for RL (Schulman et al., 2015, 2017). Their practical success is largely due to the flexibility to work with differentiable policy parameterization and the capability of directly optimizing the objective of interest. The latter aspect, in particular, has been theoretically shown to make these techniques robust to model misspecification to a much greater degree than the value- or model-based counterparts (Agarwal et al., 2020b). However, for the exploration setting which we study here, relatively few results exist for provably efficient policy optimization methods (Agarwal et al., 2020a; Shani et al., 2020; Cai et al., 2020); we include additional related work in Appendix B.

The closest work motivated by similar reasons to ours is the recent paper of Agarwal et al. (2020a), which proposes an algorithm called PC-PG that optimizes policies by performing natural gradient ascent (Kakade, 2001) to solve a sequence of optimistic MDPs. The guarantees of PC-PG exhibits the sort of robustness to misspecification that we desire: the algorithm succeeds whenever the state-action features can linearly approximate the state-action value functions of the learner's policies, with an approximation error defined in an average sense under the visitation distribution of a fixed comparator policy of interest (a notion called the *transfer error* in (Agarwal et al., 2020b)). As shown in Agarwal et al. (2020a), this type of error dependency allows for nicer guarantees in misspecification settings such as approximate state aggregations and in individual examples where value- or model-based techniques fail.

However, the robustness of PC-PG comes at a steep price: to learn an ϵ-suboptimal policy PC-PG requires $\tilde{\Omega}(1/\epsilon^{11})$ number of samples! This sample inefficiency leads us to ask whether such a trade-off is necessary for a nicer notion of model misspecification.

In this work, we present a new algorithm COPOE (*Cautiously Optimistic Policy Optimization and Exploration*), which builds on PC-PG but improves its sample complexity in three crucial ways:

- **Pessimistic evaluation with optimistic bonus:** Like most exploration methods, we use the idea of reward bonuses to realize optimism in the face of uncertainty. However, in the optimistic MDP with bonus, we perform *pessimistic* value function estimation for our policies (hence COPOE is cautiously optimistic). This trick leads to one-sided errors in our value estimates, which in turn yields important savings in sample complexity.

- **Adaptive schedule for MDP update:** We devise an adaptive scheme to construct the optimistic MDPs in order to avoid repeatedly solving similar optimistic MDPs. While PC-PG collects a fixed number of samples with the solution it finds for each optimistic MDP, we use a variable number of samples based on a data-dependent quantity and a doubling schedule. This effectively replaces $O(N)$ rounds of data collection in PC-PG with $O(d \log N)$ rounds in COPOE when we perform N iterations with a d-dimensional feature map. This is the *primary source of our sample complexity improvements*, and is enabled by a new concentration inequality for inverse covariance matrices.

- **Data reuse via importance sampling:** We show that due to the relative stability of natural gradient ascent in policy optimization, data collected by one policy can be reused to perform many policy updates with basic importance weighted Monte Carlo return estimates, without incurring excessive estimation variance as commonly conjectured. This observation allows us to avoid collecting fresh samples for every policy update as in PC-PG, while keeping its robustness property originating from using Monte Carlo estimation.

These algorithmic innovations, along with improvements in the analysis, yield the following informal result for linear MDPs. Please see Theorem 4 for the general results.

Theorem 1 (Informal result for linear MDPs) *For a linear MDP (Jin et al., 2020) with a d-dimensional feature map,* COPOE *finds an ϵ-optimal policy with probability at least $1-\delta$ using at most $\widetilde{O}\left(\frac{d^3 \log(1/\delta)}{(1-\gamma)^{13}\epsilon^3}\right)$ samples from the MDP.*

COPOE further retains the same dependence on transfer error as PC-PG when the linear MDP assumption is violated, thereby yielding an improved sample in complexity *without any sacrifice of the robustness to model misspecification*. In addition to the aforementioned algorithmic improvements, our analysis leverages a new covariance matrix concentration result (Lemma 39), which might be of independent interest.

While our algorithm is motivated by approximately linear MDPs, our new result begets the question of whether our algorithm and analysis can be further improved to match the $\approx \frac{d^2}{\epsilon^2}$ sample complexity of the best value-based methods for linear MDPs (Zanette et al., 2020b) (notice that our $\widetilde{O}$ notation hides a dependence on $\log|\mathcal{A}|$). We believe that this requires an even stronger data reuse as the variance of importance sampling limits how far back we can go in terms of reusing data from past policies. Estimators based on Bellman backups, such as Fitted Q-iteration and Least Square Policy Evaluation (Bertsekas et al., 1995; Sutton and Barto, 2018), can perform a more effective data reuse, but it is unclear if they exhibit a similar robustness to model misspecification. Further investigating these questions is a promising future direction.

2. Preliminaries

We consider a discounted infinite-horizon MDP (Puterman, 1994) $M = (\mathcal{S}, \mathcal{A}, p, r, \gamma)$ defined by a possibly infinite state space $\mathcal{S}$, a finite action space $\mathcal{A}$, a discount factor $\gamma \in [0,1)$, and for every state-action pair (s,a), a reward function $r(s,a)$ and a transition kernel $p(\cdot \mid s,a)$ over the next state. A stationary, stochastic policy π maps a state $s \in \mathcal{S}$ to a probability function $\pi(\cdot \mid s)$ over the actions in $\mathcal{A}$. A policy π then induces a distribution over states and actions $d^\pi_{s_0}(s,a) = (1-\gamma)\sum_{t=0}^\infty \gamma^t \mathbf{P}^\pi(s_t = s, a_t = a | s_0)$, which is the normalized discounted sum of probabilities that the state action (s_t, a_t) at time step t equals (s,a) under the probability function $\mathbf{P}^\pi$ associated to the Markov chain induced by π, with the start state being s_0. Sometimes we also condition on an initial state and an initial action, and we omit any conditioning when it is clear from the context; when the conditioning on the first state and action (s,a) is made explicit, we write $\mathbb{E}_{(s',a')\sim\pi|(s,a)}\cdot$. A policy π also defines a state-action value function Q^π and a state value function V^π, which are

$$Q^\pi(s,a) \stackrel{def}{=} \sum_{t=0}^\infty \gamma^t \, \mathbb{E}_{(s',a')\sim\pi|(s,a)} \, r(s',a') \quad \text{and} \quad V^\pi(s) \stackrel{def}{=} \mathbb{E}_{a\sim\pi(\cdot|s)} \, Q^\pi(s,a).$$

For a function $Q : \mathcal{S} \times \mathcal{A} \to \mathbb{R}$, we also overload the notation $Q(s, \pi)$ to denote $\mathbb{E}_{a\sim\pi} Q(s, a)$ (e.g. we can write $V^\pi(s) = Q^\pi(s, \pi)$). The corresponding advantage function for π is defined as $A^\pi(s, a) \stackrel{def}{=} Q^\pi(s, a) - V^\pi(s)$. Under some regularity assumptions there exists an optimal stationary policy $\pi^\star$ whose state and state-action value functions are $V^\star(s) = \sup_\pi V^\pi(s)$ and $Q^\star(s, a) = \sup_\pi Q^\pi(s, a)$. We also write $Q^\pi(s, a; r)$ or $V^\pi(s; r)$ when emphasizing that the reward function r defines these values.

In this paper, we study linear function approximation under an approximate version of the linear MDP model below; the exact approximation notion is given in Definition 3 *(Transfer Error)*.

Definition 2 (Linear MDP (Jin et al., 2020)) *An MDP $\mathcal{M} = (\mathcal{S}, \mathcal{A}, p, r, \gamma)$ is* linear *if there exists a known mapping $\phi : \mathcal{S} \times \mathcal{A} \to \mathbb{R}^d$ and a positive measure $\mu : \mathcal{S} \to \mathbb{R}^d$ such that for all s, a, s', we have $p(s'|s, a) = \phi(s, a)^\top \mu(s')$.*

Linear MDPs have the attractive property that for any function $f : \mathcal{S} \to \mathbb{R}$, there is $w_f \in \mathbb{R}^d$ such that $\mathbb{E}_{s'\sim p(\cdot|s,a)} f(s') = w_f^\top \phi(s, a)$. We make the normalization assumption that $\|\phi(s, a)\|_2 \le 1$ and for any function $f : \mathcal{S} \to \mathbb{R}$ such that $\|f\|_\infty \le \frac{1}{(1-\gamma)^2}$, we have $\|w_f\|_2 \le W = \widetilde{O}(\frac{1}{(1-\gamma)^2})$. The $O(\cdot)$ notation hides constant values and the $\widetilde{O}(\cdot)$ notation hides constants and $\mathrm{polylog}(d, \frac{1}{1-\gamma}, \frac{1}{\epsilon}, \frac{1}{\delta})$, where δ is the failure probability and ϵ is the suboptimality. For a symmetric positive definite matrix Σ and a vector x, we define $\|x\|_\Sigma = \sqrt{x^\top \Sigma x}$.

3. Algorithm

We present the algorithm *Cautiously Optimistic Policy Optimization and Exploration* (COPOE), which is summarized in Algorithm 1. COPOE builds on the PC-PG algorithm of Agarwal et al. (2020a), with important improvements in design to obtain a better sample complexity. Like PC-PG, COPOE is a two-loop algorithm, where the outer loop sets up a sequence of optimistic policy optimization problems which are then solved in the inner loop.

In the nth outer loop, we define the *policy cover*, π^n_{cov}, as the mixture of all the policies discovered so far, and update its empirical cumulative covariance matrix $\widehat{\Sigma}^n$. We use $\widehat{\Sigma}^n$ to estimate the state-actions that the current policy cover π^n_{cov} can confidently explore.

Algorithm 1 COPOE: Cautiously Optimistic Policy Optimization and Exploration

Parameters: N, λ, β
Initialize $\widehat{\Sigma}^1 = \lambda I, \pi^0(\cdot \mid \cdot) = \mathrm{Unif}(\mathcal{A}), \underline{n} = 1$
for $n = 1, 2, \ldots, N$ **do**
  Update policy cover $\pi^n_{cov} = \pi^{0:n-1}$
  if $\det(\widehat{\Sigma}^n) > 2\det(\widehat{\Sigma}^{\underline{n}})$ or $n = 1$ **then**
    Update known set $\mathcal{K}^n$ in (1) and bonus b^n in (2); Set $\underline{n} \leftarrow n$
    $\pi^n \leftarrow \textsc{Solver}(\pi^n_{cov}, b^n, \mathcal{K}^n)$
  else
    $\pi^n \leftarrow \pi^{\underline{n}}, \mathcal{K}^n \leftarrow \mathcal{K}^{\underline{n}}, b^n \leftarrow b^{\underline{n}}$
  end if
  Sample $\phi^n \leftarrow \textsc{FeatureSampler}(\pi^{\underline{n}})$ and update $\widehat{\Sigma}^{n+1} = \widehat{\Sigma}^n + (\phi^n)(\phi^n)^\top$
end for

Algorithm 2 SOLVER$(\pi_{cov}, b, \mathcal{K})$

Parameters: K, η, κ
$\pi_0(\cdot \mid s) = \text{Unif}(\mathcal{A})$ if $s \in \mathcal{K}$ and
$\pi_0(\cdot \mid s) = \text{Unif}(\{a \mid (s,a) \notin \mathcal{K}\})$ if $s \notin \mathcal{K}$
for $k = 0, 1, \ldots, K-1$ **do**
  if $k - \underline{k} > \kappa$ or $k = 0$ **then**
    $\underline{k} \leftarrow k$
    $\mathcal{D} \leftarrow \text{MONTECARLO}(\pi_{cov}, \pi_k, b)$
  end if
  $\widehat{Q}_k \leftarrow \text{CRITIC}(\mathcal{D}, \pi_{\underline{k}}, \pi_k, b)$
  Update policy: $\forall s \in \mathcal{K}$,
  $\pi_{k+1}(\cdot \mid s) \propto \pi_k(\cdot \mid s) e^{\eta \widehat{Q}_k(\cdot|s)}$
end for
Return: $\pi_{0:K-1} = \{\pi_0, \ldots, \pi_{K-1}\}$

Algorithm 3 CRITIC$(\mathcal{D}, \underline{\pi}, \pi, b)$

Parameters: W
for $i = 1, \ldots, |\mathcal{D}|$ **do**
  $(x_i, \mathcal{P}_i, G_i, b_i) \leftarrow \mathcal{D}[i]$
  $\rho_i \leftarrow \Pi_{\tau=2}^{|\mathcal{P}_i|} \frac{\pi(a_\tau|s_\tau)}{\underline{\pi}(a_\tau|s_\tau)}$
end for
$\widehat{w} = \min_{\|w\|_2 \le W} \sum_{i=1}^{|\mathcal{D}|} \left(x_i^\top w - \rho_i G_i - b_i \right)^2$
Return: $\widehat{Q}(s,a) = \phi(s,a)^\top \widehat{w} + \frac{1}{2} b(s,a)$,
$\forall s \in \mathcal{K}^n$ and $\widehat{Q}(s,a) = b(s,a)$ otherwise

If the current policy cover π_{cov}^n can explore a sufficiently larger space than the old policy cover can (which is measured as the change of the covariance matrices in line 5), we proceed to update the learner's policy π^n. To this end, we first define the *known state-actions*, $\mathcal{K}^n$, based on $\widehat{\Sigma}^n$, which can be thought of as the subset of state-actions that can be reached with enough probability under π_{cov}^n. Using $\mathcal{K}^n$, we create the optimistic MDP for the inner policy optimization by augmenting the original MDP $\mathcal{M}$ with a reward bonus b^n based on $\mathcal{K}^n$, so that solving the optimistic MDP would encourage the learner to explore state-actions outside $\mathcal{K}^n$ as well as to refine its estimates inside $\mathcal{K}^n$.

The policy optimization routine (in line 7 of Algorithm 1) takes these objects and returns an optimistic policy π^n. This policy π^n updates the policy cover to π_{cov}^{n+1}, which will define the next optimistic MDP when a sufficient covariance change is made again. Over the course of learning, the optimistic MDPs gradually converge to the original MDP $\mathcal{M}$.

3.1. COPOE : Outer loop

Here we describe the details of three major components used in the outer loop of COPOE (the policy cover, the known state-actions, and the reward bonus) and our adaptive rule for updating optimistic MDPs in line 5 of Algorithm 1.

Policy Cover At iteration n, we define the policy cover as $\pi_{cov}^n = \pi^{0:n-1}$, which is the uniform mixture of prior policies. When sampling from π_{cov}^n, we first sample j uniformly from $\{0, \ldots, n-1\}$ and then run π^j to generate a trajectory. Note that, in the policy cover, the policies π^j and π^{j+1} differ only if invoke SOLVER in line 7 is invoked at the $(j+1)$th outer iteration, so the cover contains many copies of each policy. As we will discuss at the end of this section, *there are only* $O(d \log n)$ *unique policies in the policy cover* π_{cov}^n.

Known state-actions The state-action space $\mathcal{S} \times \mathcal{A}$ is partitioned into two sets, namely the set $\mathcal{K}^n$ described in (1) of known state-actions and its complement. When the empirical cumulative covariance matrix $\widehat{\Sigma}^n$ is significantly different from the old one (line 5), we update the known state-action set,

$$\mathcal{K}^n = \{(s,a) \mid \sqrt{\beta} \|\phi(s,a)\|_{(\widehat{\Sigma}^n)^{-1}} < 1\}. \tag{1}$$

Intuitively, the set $\mathcal{K}^n$ represents the state-action pairs easily reached under π^n_{cov}, because state-action pairs with a small quadratic form lie in a direction that has a reasonable visitation under the policy cover π^n_{cov}, as noted in many prior works in linear bandits (Dani et al., 2008; Abbasi-Yadkori et al., 2011) and RL (Jin et al., 2020; Agarwal et al., 2020a). If the features for all actions at a state lie in the $\mathcal{K}^n$, we say the state is known; without possibility of confusion, we denote with $\mathcal{K}^n = \{s \mid \sqrt{\beta}\|\phi(s,a)\|_{(\widehat{\Sigma}^n)^{-1}} < 1, \forall a\}$ the set of known states. Unlike PC-PG, our algorithmic choices allow using a much smaller threshold β to define a substantially larger known set, as we will see in the next section.

Reward bonus At a high level, COPOE performs exploration *both* in the known and unknown regions. On unknown states $\mathcal{S} \setminus \mathcal{K}^n$ the algorithm roughly tries to emulate R-MAX (Brafman and Tennenholtz, 2002), which is reasonable when the uncertainty is very high; within the known space $\mathcal{K}^n$, the algorithm has sufficient information to explore in a much more sophisticated and efficient way, which is enabled by the bonus described below:

$$b^n(s,a) = 2b^n_\phi(s,a) + b^n_{\mathbb{1}}(s,a), \quad \text{where} \tag{2}$$
$$b^n_\phi(s,a) = \sqrt{\beta}\|\phi(s,a)\|_{(\widehat{\Sigma}^n)^{-1}}\mathbb{1}\{s \in \mathcal{K}^n\}, \quad \text{and} \quad b^n_{\mathbb{1}}(s,a) = \frac{3}{1-\gamma}\mathbb{1}\{(s,a) \notin \mathcal{K}^n\}$$

In other words, the bonus is assigned differently on the known and unknown spaces. On unknown state-actions, the assigned bonus equals $b^n_{\mathbb{1}}(s,a) = \frac{3}{1-\gamma}$, which is the largest value of the original reward over a trajectory. Consequently, visiting any such state-action pair is strictly preferable to staying within the known subset of the MDP and the known set $\mathcal{K}^n$ is expanded. In the known region, the uncertainty is quantified by the bonus $b^n_\phi(s,a) = \sqrt{\beta}\|\phi(s,a)\|_{(\widehat{\Sigma}^n)^{-1}}$ which is the only one active inside the known space. This form of the bonus b^n_ϕ is standard from the linear bandit literature (e.g., (Dani et al., 2008; Abbasi-Yadkori et al., 2011) and linear MDPs (Jin et al., 2020).

Our definition of bonus differs from that in the related PC-PG algorithm. Unlike COPOE, PC-PG only explores using the bonus $b_{\mathbb{1}}$, ignoring the amount of information (or uncertainty) encoded in the quadratic form $\|\phi(s,a)\|_{(\widehat{\Sigma}^n)^{-1}}$. As a result, PC-PG stops exploring a state-action immediately after it becomes known (i.e. in $\mathcal{K}^n$). We found that such a behavior is undesirable, because doing so would couple the threshold used in defining the known set $\mathcal{K}^n$ with the policy performance suboptimality ϵ, ultimately resulting in a more sample inefficient exploration.

Adaptive updates of optimistic MDPs It remains to explain why infrequent or *lazy updates* of the optimistic MDP (line 5 of Algorithm 1) are beneficial. Recall that in iteration n we seek to find

$$\pi^n \approx \max_\pi \mathbb{E}_{s\sim\pi^n_{cov}} V^\pi(s; r + b^n), \tag{3}$$

and add it to the policy cover. However, finding this policy entails a significant sample complexity because the SOLVER (Algorithm 2) — which relies on Monte Carlo estimations to evaluate its policies — must be invoked.

This suggests to call the SOLVER only when the returned policy π^n is expected to be significantly better than the prior one $\pi^{\underline{n}}$ or to make a significant contribution to the policy cover. Because the optimistic MDP is defined by the bonus b^n, which is a function of $\widehat{\Sigma}^n$ (see (1) and (2)), the optimistic MDP only changes significantly when the updated $\widehat{\Sigma}^n$ is very different as measured by its determinant. Therefore, each time the determinant doubles (line 5), we update the known set and the bonus according to (1) and (2), respectively, based on the latest $\widehat{\Sigma}^n$. Then we invoke the SOLVER to

find a new policy to update the policy cover. As a result, the number of solver invocations is reduced from $O(N)$ to $O(d \log N)$, providing substantial sample complexity gains.

3.2. COPOE: Inner loop

We now turn our attention to the SOLVER, Algorithm 2. At a high-level, we initialize the policy to be a uniform distribution that prefers at a state s to take an unknown action a such that $(s, a) \notin \mathcal{K}^n$, and employ the online learning algorithm (the exponentiated weight update (Freund and Schapire, 1997)) on the *known* states to update the policy. This update rule is equivalent to the Natural Policy Gradient (NPG) algorithm for log-linear policies (Kakade, 2001; Agarwal et al., 2020b).

The update rule is an actor-critic scheme, where we fit the critic by regressing on the observed Monte Carlo returns and update the actor using exponentiated weights. As argued in (Agarwal et al., 2020b), using Monte Carlo as critic is an essential technique to provide better robustness to model misspecification compared with a least squares policy evaluation (LSPE) (Bertsekas and Ioffe, 1996) method, but is also a significant source of sample complexity.

Fitting the critic with nearly on-policy data To improve the sample complexity of Algorithm 2, we devise a way to *reuse past data while keeping the robustness property of Monte Carlo*.

Our estimator reuses data by applying trajectory-level importance sampling on past Monte Carlo return estimates (Precup, 2000). While trajectory-level importance sampling has been typically associated with exponentially high variance, we found that its variance is constant when we properly control how much into the past the data are reused, because the policies produced by the online learning here do not change significantly between successive updates but induce similar trajectories.

At iteration k in Algorithm 2, we have access to a dataset of trajectories previously drawn in Algorithm 4 by first sampling $s, a \sim \pi_{cov}$, and then following the policy $\pi_{\underline{k}}$ for some prior iteration $\underline{k} \geq k - \kappa$ (see Algorithm 4 in the appendix for details). We use this dataset to obtain a Monte Carlo return estimate for the current policy π_k by reweighting the samples with importance sampling (see Algorithm 3 for details). Subsequently, we learn a critic by training a linear function to map $\phi(s, a)$ — the feature vectors for the initial state and action sampled from π_{cov} — to the reweighted random return via least squares linear regression (line 8 in Algorithm 2 and Algorithm 3).

Following prior works (Jin et al., 2020; Agarwal et al., 2020a), we offset the sampled return by the bonus value at the initial state-action in the trajectory in line 6 of Algorithm 3. This offset ensures that the regression target is perfectly realizable using a linear function in $\phi(s, a)$ when the MDP is exactly linear, despite the non-linear bonus function.

Cautious optimism and one-sided errors Since the critic fitting in line 6 of Algorithm 3 is offset by the initial bonus to preserve linearly of the representation, it would be natural to define the critic estimates as $\widehat{Q}(s, a) = \widehat{w}^\top \phi(s, a) + b^n(s, a)$, which would exactly correct for the offset (this is the approach taken in PC-PG). However, in line 7 of Algorithm 3, we only partially correct for the offset and instead define the critic estimate as $\widehat{Q}(s, a) = \widehat{w}^\top \phi(s, a) + \frac{1}{2} b^n(s, a)$. This introduces a *negative bias* in the estimate. However, since our critic is being fit to the bonus augmented returns, we are able to show in our analysis that $\widehat{Q}_k(s, a)$ (in line 8 of Algorithm 2) is still optimistic relative to $Q^{\pi_k}(s, a; r)$, while being an underestimate of $Q^{\pi_k}(s, a; r + b^n)$. This one-sided error property plays a crucial role of improving a factor of $O(\frac{1}{\epsilon})$ in sample complexity.

Actor updates With the critic computed above, line 9 in Algorithm 2 updates the policy on the known states using the exponentiated weight updates, with the critic function as the negative loss. We change the data collection policy every κ iterations to collect a fresh dataset for critic fitting.

4. Main Result

In this section we provide the main guarantees for COPOE. We make the following *transfer error* assumption, originally introduced in (Agarwal et al., 2020b) for policy gradient algorithms.

Definition 3 (Transfer Error) *Define the loss functional*

$$\mathcal{L}(w, d, f) \stackrel{def}{=} \frac{1}{2}\,\mathbb{E}_{(s,a)\sim d}\left[\phi(s,a)^\top w - f\right]^2. \tag{4}$$

For a given outer iteration n (in Algorithm 1) and an inner iteration k (in Algorithm 2) let

$$Q_k^n(s,a) = Q^{\pi_k}(s,a; r+b^n), \quad Q^n(s,a) = Q^{\pi^n}(s,a; r+b^n) \tag{5}$$

be the optimistic action-value functions. Define the 'best' regression parameters

$$w_k^{n,\star} \in \arg\min_{\|w\|_2\le W} \mathcal{L}(w, \rho^n, Q_k^n - b^n), \qquad w^{n,\star} \in \arg\min_{\|w\|_2\le W} \mathcal{L}(w, \rho^n, Q^n - b^n). \tag{6}$$

Then the transfer error with respect to a fixed comparator $\widetilde{\pi}$ is defined as[1]

$$\mathcal{E}_k^n \stackrel{def}{=} \mathcal{L}(w_k^{n,\star}, d^{\widetilde{\pi}} \circ \mathit{Unif}(|\mathcal{A}|), Q_k^n - b^n). \tag{7}$$

For compactness we denote the average approximation error across N and K (the inner and outer iterations of the algorithm) as $\sqrt{\mathcal{E}} \stackrel{def}{=} \frac{1}{NK}\sum_{n=1}^{N}\sum_{k=0}^{K-1}\sqrt{\mathcal{E}_k^n}$.

The transfer error measures the average prediction error of the agent's estimator in the limit of infinite data on unseen samples. For the transfer error to be small, the estimator does not need to be pointwise accurate but only accurate in *expectation* along a *fixed distribution*, namely the state-action distribution induced by the comparator $\widetilde{\pi}$ (typically the optimal policy $\pi^\star$). These are substantially weaker requirements than the typical ℓ_∞ error assumption arising from the use of temporal difference methods. In particular, on the low-rank or linear MDP model (Yang and Wang, 2020; Jin et al., 2020; Zanette et al., 2020a) the transfer error is zero. In this case, we say that the linear model is not misspecified; for more details please see Appendix C.

Theorem 4 (Sample Complexity Analysis of COPOE) *Fix a failure probability δ; for appropriate input parameters,*

$$(N, K, \eta, \lambda, \kappa, W) = \widetilde{O}\left(\frac{d^2}{(1-\gamma)^8\epsilon^2}, \frac{\ln|\mathcal{A}|W^2}{(1-\gamma)^2\epsilon^2}, \frac{\sqrt{\ln|\mathcal{A}|}}{\sqrt{K}W}, d, \frac{1-\gamma}{\eta W}, \frac{1}{(1-\gamma)^2}\right),$$

COPOE *returns with probability at least $1-\delta$ a policy π^{COPOE} such that*

$$\left(V^\star - V^{\pi^{\text{COPOE}}}\right)(s_0) \le \epsilon + \frac{2\sqrt{2|\mathcal{A}|\mathcal{E}}}{1-\gamma},$$

using at most $\widetilde{O}(\frac{d^3}{(1-\gamma)^{13}\epsilon^3})$ samples.

1. Shifting the Q values below by the bonus b^n in regression and adding the bonus afterwards is a standard practice in exploration methods (see, e.g., (Jin et al., 2020)).

We now discuss some aspects of our result and compare it to the most relevant prior works.

Better robustness compared to LSVI-UCB Compared to (Jin et al., 2020) on well-specified linear MDPs, COPOE provides PAC bounds to find an ϵ-optimal policy and inherents the same $O(d^3)$ dependence on the feature dimension as (Jin et al., 2020), while being $1/\epsilon$ worse in sample complexity and in horizon dependence; a $\log|\mathcal{A}|$ factor is also implicitly hidden in our $\widetilde{O}$ notation. However, the transfer error of COPOE in Definition 3 can be a significantly weaker assumption, as discussed in (Agarwal et al., 2020a).

Sample complexity improvement relative to PC-PG COPOE operates under an essentially identical notion of transfer error as PC-PG in (Agarwal et al., 2020a) and shares several PC-PG's algorithmic principles (e.g. the exponentiated weights rule for policy update, Monte Carlo for policy evaluation, and the concept of policy cover). But importantly, because COPOE uses a better bonus structure, adaptive bonus updates, and performs importance sampling to reuse Monte Carlo data, COPOE is able to lower the sample complexity from the slow $\widetilde{O}(\frac{1}{\epsilon^{11}})$ rate of PC-PG to the faster $\widetilde{O}(\frac{1}{\epsilon^3})$ rate. Note that unlike PC-PG, we do not extend our analysis to the infinite dimensional setting, though we expect it to be possible using the covering arguments from Yang et al. (2020a).

Better sample complexity in the optimization setting Finally, COPOE 's analysis is based on the natural policy gradient algorithm (Kakade, 2001), which has recently been analyzed in (Agarwal et al., 2020b) when a good sampling distribution is already given (for example, through a generative model). For solving the policy optimization subproblem, COPOE improves the $\widetilde{O}(\frac{1}{\epsilon^4})$ rate obtained in (Agarwal et al., 2020b) to $\widetilde{O}(\frac{1}{\epsilon^3})$ by the data reuse scheme described in Section 3.2.

5. Technical Analysis

In this section we briefly sketch the analysis of COPOE and prove Theorem 4. We start by giving a regret decomposition analysis of the policies computed by COPOE in Section 5.1. This result will be used as the foundation of the proof of the main result in Section 5.2

Notation We introduce a few more notations to simplify the presentation. The outer policy π^n is a uniform mixture of the policies $\pi_0^n, \ldots, \pi_{K-1}^n$ returned by the SOLVER (see Algorithm 2) in outer iteration n. For the policy π^n in the outer iteration n in Algorithm 1, we denote with $Q^n(s,a) = Q^{\pi^n}(s,a; r+b^n)$ the state-action value function, with $V^n(s) = V^{\pi^n}(s; r+b^n)$ the state value function, and with $A^n(s,a) = Q^n(s,a) - V^n(s)$ the advantage function on the optimistic MDP. Similarly, for the linear approximation (given by the Monte Carlo regression in line 8 of Algorithm 2), we write $\widehat{Q}^n(s,a) = \frac{1}{K}\sum_{k=0}^{K-1} \widehat{Q}^{\pi_k^n}(s,a)$, $\widehat{V}^n(s) = \frac{1}{K}\sum_{k=0}^{K-1} \widehat{Q}^{\pi_k^n}(s,\pi_k^n)$, and $\widehat{A}^n(s,a) = \widehat{Q}^n(s,a) - \widehat{V}^n(s)$. Using the best regressed parameter $w^{n,\star}$ in Definition 3, we also define the best predictor $Q^{n,\star}(s,a) = \phi(s,a)^\top w^{n,\star} + b^n(s,a)$ and its advantage function $A^{n,\star}(s,a) = Q^{n,\star}(s,a) - V^{n,\star}(s)$. In absence of misspecification, we note that $Q^{n,\star} = Q^n$.

5.1. Regret Decomposition

Fix an outer iteration index n. We start the analysis by giving the following performance lemma, which is obtained by combining the performance difference lemma (Kakade and Langford, 2002) with several properties of our algorithm.

Lemma 5 (Performance Analysis; (44) in appendix) *With high probability,* COPOE *ensures*

$$(1-\gamma)(V^\star - V^{\pi^n})(s_0) \leq \underbrace{\sup_{s\in\mathcal{K}^n} \widehat{A}^n(s,\pi^\star)}_{\text{Solver error}} + \mathbb{E}_{(s,a)\sim\pi^\star} \underbrace{\left|A^n(s,a) - A^{n,\star}(s,a)\right| \mathbb{1}\{s\in\mathcal{K}^n\}}_{\text{Approximation error on states in } \mathcal{K}^n}$$
$$+ \mathbb{E}_{(s,a)\sim\pi^\star} \underbrace{\left(Q^{n,\star}(s,a) - \widehat{Q}^n(s,a)\right)\mathbb{1}\{s\in\mathcal{K}^n\}}_{\text{Statistical error along } \pi^\star \text{ on states in } \mathcal{K}^n}$$
$$- \underbrace{\mathbb{E}_{(s,a)\sim\pi^\star} 2b^n_\phi(s,a)\mathbb{1}\{s\in\mathcal{K}^n\}}_{\text{Bonus along } \pi^\star \text{ on states in } \mathcal{K}^n} + \underbrace{\mathbb{E}_{(s,a)\sim\pi^n} b^n(s,a)}_{\text{Bonus along } \pi^n \text{ on the full space}} . \qquad (8)$$

We discuss each of these terms in detail below.

5.1.1. SOLVER ERROR

The first term in (8) measures how well the policy π^n performs in terms of our empirical advantage function on known states; generating such a policy is done using the regret guarantee of our online learning rule in Algorithm 2. We have the following lemma (see also (Agarwal et al., 2020b,a)).

Lemma 6 (Online regret of softmax; Lemma 19 in appendix) *Using an appropriate learning rate η, Algorithm 2 identifies a mixture policy π^n that satisfies* $\sup_{s\in\mathcal{K}^n}\sup_{a\in\mathcal{A}} \widehat{A}^n(s,a) = \widetilde{O}\left(\frac{1}{(1-\gamma)^2}\sqrt{\frac{1}{K}}\right)$.

Thus the solver error in (8) can be reduced arbitrarily, although the number of iterations K directly affects the sample complexity; see Section 5.2.

5.1.2. APPROXIMATION ERROR

The second term in (8) is an approximation error in advantages under $\pi^\star$ and is non-zero only when the linear MDP assumption is not exactly satisfied. The performance bound of Lemma 5 highlights that the *approximation error* is measured *1)* in expectation and *2)* along the distribution induced by $\pi^\star$. For brevity, we neglect the approximation error in this proof sketch; we note that this quantity can be controlled in the general version of the result using the transfer error condition (Definition 3).

5.1.3. STATISTICAL ERROR

The third term in (8) is perhaps the most surprising: it reasons about the statistical error in our critic fitting on the known states, *but only under states and actions chosen according to $\pi^\star$*. In other words, the agent's estimator does not need to be correct for arbitrary distributions; otherwise, an ℓ_∞ guarantee over the known-set is needed (as needed by Agarwal et al. (2020a)). Such result is enabled by the following key lemma, which contributes the underestimation property of $\widehat{Q}$ needed in the proof of Lemma 5. (Recall the regression target is subtracted with $b^n(s,a)$ but the final predictor adds back only $\frac{1}{2}b^n(s,a)$.)

Lemma 7 (One sided errors; Lemma 30 in appendix) *Let $w^{n,\star}$ be defined in Definition 3 and let $\widehat{w}^n$ be the corresponding empirical minimizer. Define the agent's predictor on $(s,a)\in\mathcal{K}^n$ as $\widehat{Q}^{n,\star}(s,a) \stackrel{def}{=} \phi(s,a)^\top \widehat{w}^n + \frac{1}{2}b^n(s,a)$. Then with high probability, jointly $\forall n$ and $\forall(s,a)\in\mathcal{K}^n$,*

$$0 \leq (Q^{n,\star} - \widehat{Q}^n)(s,a) \leq b^n(s,a) = 2b_\phi(s,a). \qquad (9)$$

5.1.4. BONUS DIFFERENCE AND CONCENTRATION

The final two terms in (8) arise as we optimize policies in the optimistic MDP, but the performance difference of interest is defined for the original MDP. The negative bonus term under the comparator $\pi^\star$ helps cancel some of the statistical errors (i.e. the third term), which is crucial for the overall sample complexity results.

For the other bonus term under π^n, we can bound it using the elliptic potential lemma (e.g., (Abbasi-Yadkori et al., 2011)) and a martingale argument.

Lemma 8 (Concentration on Bonus; Lemmas 32 and 33 in appendix) *With high probability, it holds that* $\sum_{n=1}^{N} \mathbb{E}_{(s,a)\sim\pi^n} b^n(s,a) = \widetilde{O}\left(\frac{d\sqrt{N}}{(1-\gamma)^3}\right)$.

5.2. Sample Complexity Analysis (Proof of theorem 4)

In order to bound the sample complexity for obtaining Theorem 4, we need to bound the following quantities:

1. the number of outer iterations N to control the number of samples collected for the matrix $\widehat{\Sigma}^n$,
2. the number of calls to SOLVER across N iterations,
3. the number of data collection rounds in SOLVER for critic fitting, and
4. the number of samples in each dataset that SOLVER collects. We start with bounding the number of inner and outer iterations.

Lemma 9 (Convergence rate of COPOE ; Proposition 16 in appendix) *With high probability* COPOE *, computes policies* $\pi_1, \ldots, \pi_N$ *such that*

$$\frac{1}{N}\sum_{n=1}^{N}\left(V^\star - V^{\pi^n}\right)(s_0) \le \underbrace{\widetilde{O}\left(\frac{1}{(1-\gamma)^3\sqrt{K}}\right)}_{\textit{Solver error}} + \textit{Approx. error} + \underbrace{\widetilde{O}\left(\frac{d}{(1-\gamma)^4\sqrt{N}}\right)}_{\textit{Average statistical uncertainty}}.$$

where Approx. error *denotes the second term in Lemma 5.*

Using the above proposition we can give a proof of Theorem 4.

Proof (of Theorem 4) To ensure the average suboptimality gap is below ϵ we need to ensure:

$$\frac{1}{N}\sum_{n=1}^{N}\left(V^\star - V^{\pi^n}\right)(s_0) \le \epsilon \qquad \longrightarrow \qquad K \approx \frac{1}{(1-\gamma)^6\epsilon^2}, \quad N \approx \frac{d^2}{(1-\gamma)^8\epsilon^2}. \tag{10}$$

Next we bound the number of calls to SOLVER; this is controlled by the lazy update (line 5 in Algorithm 1) and the bonus structure.

Lemma 10 (Number of solver calls; Lemma 38) COPOE *invokes* SOLVER *at most* $O(d\log N)$ *times.*

Every time it is invoked, SOLVER runs for K iterations in (10) and at every iteration it needs to receive an evaluation on the performance of the current policy ($\widehat{Q}_k$ estimator from the critic, Algorithm 3). Using importance sampling we can avoid collecting fresh Monte Carlo data for every policy. We control the number of data collection rounds based on importance sampling variance.

Lemma 11 (Stability of the Importance Sampling Estimator; Lemma 24 and (129) in appendix) *The importance sampling ratio used in Algorithm 3 is bounded by a constant with high probability:*

$$\textit{If} \quad k - \underline{k} = \widetilde{O}\left(\sqrt{K}(1-\gamma)\right), \quad \textit{then} \quad \Pi_{\tau=2}^{t} \frac{\pi_k(s_\tau, a_\tau)}{\pi_{\underline{k}}(s_\tau, a_\tau)} \leq 2, \qquad \forall \{s_1, a_1, \ldots, s_t, a_t\}.$$

In other words, after fresh Monte Carlo trajectories are collected, the importance sampling estimator can be used to make stable predictions of the value roughly for the future $\sqrt{K}(1-\gamma)$ policies. This implies that we need to collect fresh data at most once every $K/((1-\gamma)\sqrt{K}) = \widetilde{O}\left(\frac{\sqrt{K}}{1-\gamma}\right)$ iterations.

It remains to specify the number of samples we collect in each round of data collection. Note that in our statistical analysis, we want the critic error to be bounded by $b^n(s,a)$, which roughly goes down as $O(1/\sqrt{n})$, as the matrix $\widehat{\Sigma}^n$ that defines the bonus grows linearly in n. This vague intuition can be formalized by appealing to standard linear regression analysis to show that we need to collect $O(n)$ Monte Carlo returns to fit the critic in outer iteration n.

Lemma 12 (Number of Monte Carlo Trajectories) *When the Monte Carlo procedure is invoked at the outer iteration $\underline{n}$, at most $\underline{n} \leq N$ trajectories are collected.*

Finally, $\widetilde{O}(\frac{\log(1/\delta)}{1-\gamma})$ is a uniform high probability bound on the length of each Monte Carlo trajectory, which implies the total sample complexity of COPOE is

$$\underbrace{\widetilde{O}(d)}_{\text{\# calls to Algorithm 2}} \times \underbrace{\widetilde{O}\left(\frac{\sqrt{K}}{1-\gamma}\right)}_{\text{\# calls to Algorithm 4}} \times \underbrace{\widetilde{O}\left(N\right)}_{\text{\# Monte Carlo trajectories}} \times \underbrace{\widetilde{O}\left(\frac{1}{1-\gamma}\right)}_{\text{\# samples per trajectory}} = \widetilde{O}\left(\frac{d^3}{(1-\gamma)^{13}\epsilon^3}\right).$$

■

6. Discussion

In this paper, we advance the theoretical understanding of sample-efficient policy optimization methods with strategic exploration and robustness to model misspecification. While we carry out our analysis for a specific algorithm, we expect the insights developed here for sample complexity improvements to be more broadly applicable. For instance, the exponentiated weight updates in our policy optimization subroutine can generally be substituted with other no-regret algorithms from the Follow The Regularized Leader family. As usual, we expect different choices to offer varying trade-offs in their dependence on problem parameters; with reasonable choices, they are still amenable to the importance sampling based data reuse. Similarly, the lazy updates for the bonus are generically applicable. Note that our algorithmic choices strike a particular balance of a very infrequent bonus update and a fairly accurate optimization. Prior works in different, but related problems (Agarwal et al., 2014) have shown that often there is flexibility in these choices, such as more regular updates followed by coarser optimization, which might be empirically preferable.

Perhaps the most important outstanding question not addressed here is how to close the gap between the $\widetilde{O}(1/\epsilon^2)$ sample complexity that is known to be achievable in linear MDPs (see e.g. Jin et al. (2020)) and our worse dependence of $\widetilde{O}(1/\epsilon^3)$. There appears to be a trade-off in terms of the allowable assumption on model misspecification, and approaches based on Least Square Policy

Evaluation for data reuse (such as LSVI-UCB) fail to work under our transfer error assumption and the special cases in Agarwal et al. (2020a). Whether this trade-off is fundamental, or if a single method can be developed to be robust to transfer error, while enjoying an optimal sample complexity guarantee in the absence of misspecification is an interesting direction for future work.

Acknowledgment

Most of the work was completed while Andrea Zanette was interning at Microsoft Research and the remaining part of the work was done while Andrea Zanette was visiting the Simons Institute for the Theory of Computing.

Acknowledgments

The authors are grateful to the reviewers for their helpful comments.

References

Yasin Abbasi-Yadkori, David Pal, and Csaba Szepesvari. Improved algorithms for linear stochastic bandits. In *Advances in Neural Information Processing Systems (NIPS)*, 2011.

Yasin Abbasi-Yadkori, Peter Bartlett, Kush Bhatia, Nevena Lazic, Csaba Szepesvári, and Gellért Weisz. Politex: Regret bounds for policy iteration using expert prediction. In *Proceedings of the 36th International Conference on Machine Learning*, volume 97, pages 3692–3702, 2019.

Alekh Agarwal, Daniel Hsu, Satyen Kale, John Langford, Lihong Li, and Robert Schapire. Taming the monster: A fast and simple algorithm for contextual bandits. In *International Conference on Machine Learning*, pages 1638–1646, 2014.

Alekh Agarwal, Mikael Henaff, Sham Kakade, and Wen Sun. Pc-pg: Policy cover directed exploration for provable policy gradient learning. *arXiv preprint arXiv:2007.08459*, 2020a.

Alekh Agarwal, Sham M Kakade, Jason D Lee, and Gaurav Mahajan. Optimality and approximation with policy gradient methods in markov decision processes. In *Conference on Learning Theory*, pages 64–66, 2020b.

Alex Ayoub, Zeyu Jia, Csaba Szepesvari, Mengdi Wang, and Lin F Yang. Model-based reinforcement learning with value-targeted regression. *arXiv preprint arXiv:2006.01107*, 2020.

Mohammad Gheshlaghi Azar, Bert Kappen, et al. Dynamic policy programming with function approximation. In *International Conference on Artificial Intelligence and Statistics (AISTATS)*, 2011.

Mohammad Gheshlaghi Azar, Ian Osband, and Remi Munos. Minimax regret bounds for reinforcement learning. In *International Conference on Machine Learning (ICML)*, 2017.

Dimitri P Bertsekas and Sergey Ioffe. Temporal differences-based policy iteration and applications in neuro-dynamic programming. *Lab. for Info. and Decision Systems Report LIDS-P-2349, MIT, Cambridge, MA*, 14, 1996.

Dimitri P Bertsekas, Dimitri P Bertsekas, Dimitri P Bertsekas, and Dimitri P Bertsekas. *Dynamic programming and optimal control*, volume 1. Athena scientific Belmont, MA, 1995.

Alina Beygelzimer, John Langford, Lihong Li, Lev Reyzin, and Robert Schapire. Contextual bandit algorithms with supervised learning guarantees. In *Proceedings of the Fourteenth International Conference on Artificial Intelligence and Statistics*, pages 19–26. JMLR Workshop and Conference Proceedings, 2011.

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

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

Qi Cai, Zhuoran Yang, Chi Jin, and Zhaoran Wang. Provably efficient exploration in policy optimization. In *International Conference on Machine Learning*, pages 1283–1294. PMLR, 2020.

Varsha Dani, T. Hayes, and Sham M. Kakade. Stochastic linear optimization under bandit feedback. In *COLT*, 2008.

Christoph Dann, Nan Jiang, Akshay Krishnamurthy, Alekh Agarwal, John Langford, and Robert E Schapire. On oracle-efficient pac rl with rich observations. In *Advances in Neural Information Processing Systems (NIPS)*, pages 1429–1439, 2018.

Christoph Dann, Lihong Li, Wei Wei, and Emma Brunskill. Policy certificates: Towards accountable reinforcement learning. In *International Conference on Machine Learning*, pages 1507–1516, 2019.

Simon S Du, Sham M Kakade, Ruosong Wang, and Lin F Yang. Is a good representation sufficient for sample efficient reinforcement learning? *arXiv preprint arXiv:1910.03016*, 2019.

Simon S. Du, Jason D. Lee, Gaurav Mahajan, and Ruosong Wang. Agnostic q-learning with function approximation in deterministic systems: Tight bounds on approximation error and sample complexity, 2020.

Yonathan Efroni, Nadav Merlis, Mohammad Ghavamzadeh, and Shie Mannor. Tight regret bounds for model-based reinforcement learning with greedy policies. In *Advances in Neural Information Processing Systems*, 2019.

Eyal Even-Dar, Sham M Kakade, and Yishay Mansour. Online markov decision processes. *Mathematics of Operations Research*, 34(3):726–736, 2009.

Maryam Fazel, Rong Ge, Sham Kakade, and Mehran Mesbahi. Global convergence of policy gradient methods for the linear quadratic regulator. In *International Conference on Machine Learning*, pages 1467–1476. PMLR, 2018.

Yoav Freund and Robert E Schapire. A decision-theoretic generalization of on-line learning and an application to boosting. *Journal of computer and system sciences*, 55(1):119–139, 1997.

Matthieu Geist, Bruno Scherrer, and Olivier Pietquin. A theory of regularized markov decision processes. *arXiv preprint arXiv:1901.11275*, 2019.

Nan Jiang, Akshay Krishnamurthy, Alekh Agarwal, John Langford, and Robert E. Schapire. Contextual decision processes with low Bellman rank are PAC-learnable. In Doina Precup and Yee Whye Teh, editors, *International Conference on Machine Learning (ICML)*, volume 70 of *Proceedings of Machine Learning Research*, pages 1704–1713, International Convention Centre, Sydney, Australia, 06–11 Aug 2017. PMLR. URL http://proceedings.mlr.press/v70/jiang17c.html.

Chi Jin, Zeyuan Allen-Zhu, Sebastien Bubeck, and Michael I Jordan. Is q-learning provably efficient? In *Advances in Neural Information Processing Systems*, pages 4863–4873, 2018.

Chi Jin, Zhuoran Yang, Zhaoran Wang, and Michael I Jordan. Provably efficient reinforcement learning with linear function approximation. In *Conference on Learning Theory*, 2020.

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

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

Sham Machandranath Kakade et al. *On the sample complexity of reinforcement learning*. PhD thesis, University of London London, England, 2003.

Vijay R Konda and John N Tsitsiklis. Actor-critic algorithms. In *Advances in neural information processing systems*, pages 1008–1014. Citeseer, 2000.

Akshay Krishnamurthy, Alekh Agarwal, and John Langford. Pac reinforcement learning with rich observations. In *Advances in Neural Information Processing Systems (NIPS)*, pages 1840–1848, 2016.

Tor Lattimore and Csaba Szepesvári. *Bandit Algorithms*. Cambridge University Press, 2020.

Boyi Liu, Qi Cai, Zhuoran Yang, and Zhaoran Wang. Neural proximal/trust region policy optimization attains globally optimal policy. *arXiv preprint arXiv:1906.10306*, 2019.

Nishant Mehta. Fast rates with high probability in exp-concave statistical learning. In *Artificial Intelligence and Statistics*, pages 1085–1093. PMLR, 2017.

Gergely Neu, Anders Jonsson, and Vicenç Gómez. A unified view of entropy-regularized markov decision processes. *arXiv preprint arXiv:1705.07798*, 2017.

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

Martin L. Puterman. *Markov Decision Processes: Discrete Stochastic Dynamic Programming*. John Wiley & Sons, Inc., New York, NY, USA, 1994. ISBN 0471619779.

Daniel Russo. Worst-case regret bounds for exploration via randomized value functions. In *Advances in Neural Information Processing Systems*, 2019.

Bruno Scherrer and Matthieu Geist. Local policy search in a convex space and conservative policy iteration as boosted policy search. In *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*, pages 35–50. Springer, 2014.

John Schulman, Sergey Levine, Pieter Abbeel, Michael Jordan, and Philipp Moritz. Trust region policy optimization. In *International conference on machine learning*, pages 1889–1897. PMLR, 2015.

John Schulman, Filip Wolski, Prafulla Dhariwal, Alec Radford, and Oleg Klimov. Proximal policy optimization algorithms. *arXiv preprint arXiv:1707.06347*, 2017.

Lior Shani, Yonathan Efroni, Aviv Rosenberg, and Shie Mannor. Optimistic policy optimization with bandit feedback. In *International Conference on Machine Learning*, pages 8604–8613. PMLR, 2020.

Richard S Sutton and Andrew G Barto. *Reinforcement learning: An introduction*. MIT Press, 2018.

Richard S Sutton, David A McAllester, Satinder P Singh, Yishay Mansour, et al. Policy gradient methods for reinforcement learning with function approximation. In *NIPs*, volume 99, pages 1057–1063. Citeseer, 1999.

Roman Vershynin. Introduction to the non-asymptotic analysis of random matrices. *arXiv preprint arXiv:1011.3027*, 2010.

Ruosong Wang, Dean P Foster, and Sham M Kakade. What are the statistical limits of offline rl with linear function approximation? *arXiv preprint arXiv:2010.11895*, 2020a.

Ruosong Wang, Ruslan Salakhutdinov, and Lin F. Yang. Provably efficient reinforcement learning with general value function approximation, 2020b.

Yining Wang, Ruosong Wang, Simon S Du, and Akshay Krishnamurthy. Optimism in reinforcement learning with generalized linear function approximation. *arXiv preprint arXiv:1912.04136*, 2019.

Gellert Weisz, Philip Amortila, and Csaba Szepesvári. Exponential lower bounds for planning in mdps with linearly-realizable optimal action-value functions. *arXiv preprint arXiv:2010.01374*, 2020.

Zheng Wen and Benjamin Van Roy. Efficient exploration and value function generalization in deterministic systems. In *Advances in Neural Information Processing Systems (NIPS)*, 2013.

Ronald J Williams. Simple statistical gradient-following algorithms for connectionist reinforcement learning. *Machine learning*, 8(3-4):229–256, 1992.

Lin F Yang and Mengdi Wang. Reinforcement leaning in feature space: Matrix bandit, kernels, and regret bound. In *International Conference on Machine Learning (ICML)*, 2020.

Zhuoran Yang, Chi Jin, Zhaoran Wang, Mengdi Wang, and Michael I Jordan. Bridging exploration and general function approximation in reinforcement learning: Provably efficient kernel and neural value iterations. *arXiv preprint arXiv:2011.04622*, 2020a.

Zhuoran Yang, Chi Jin, Zhaoran Wang, Mengdi Wang, and Michael I. Jordan. On function approximation in reinforcement learning: Optimism in the face of large state spaces, 2020b.

Andrea Zanette. Exponential lower bounds for batch reinforcement learning: Batch rl can be exponentially harder than online rl. *arXiv preprint arXiv:2012.08005*, 2020.

Andrea Zanette and Emma Brunskill. Tighter problem-dependent regret bounds in reinforcement learning without domain knowledge using value function bounds. In *International Conference on Machine Learning (ICML)*, 2019. URL http://proceedings.mlr.press/v97/zanette19a.html.

Andrea Zanette, David Brandfonbrener, Matteo Pirotta, and Alessandro Lazaric. Frequentist regret bounds for randomized least-squares value iteration. In *AISTATS*, 2020a.

Andrea Zanette, Alessandro Lazaric, Mykel Kochenderfer, and Emma Brunskill. Learning near optimal policies with low inherent bellman error. In *International Conference on Machine Learning (ICML)*, 2020b.

Andrea Zanette, Alessandro Lazaric, Mykel J Kochenderfer, and Emma Brunskill. Provably efficient reward-agnostic navigation with linear value iteration. In *Advances in Neural Information Processing Systems*, 2020c.

Zihan Zhang, Xiangyang Ji, and Simon S Du. Is reinforcement learning more difficult than bandits? a near-optimal algorithm escaping the curse of horizon. *arXiv preprint arXiv:2009.13503*, 2020.

Dongruo Zhou, Quanquan Gu, and Csaba Szepesvari. Nearly minimax optimal reinforcement learning for linear mixture markov decision processes. *arXiv preprint arXiv:2012.08507*, 2020a.

Dongruo Zhou, Jiafan He, and Quanquan Gu. Provably efficient reinforcement learning for discounted mdps with feature mapping. *arXiv preprint arXiv:2006.13165*, 2020b.

Appendix A. Remaining Algorithm Components

Algorithm 4 MONTECARLO($\pi_{1:q}, \pi, b$)

Inputs: Policy cover $\pi^{1:q}$, evaluation policy π, additional reward b
$\mathcal{D} = \emptyset$
for iteration $i = 1, \ldots, q$ **do**
  Sample j uniformly at random in $[q]$
  Sample $\tau \geq 1$ with probability $\gamma^{\tau-1}(1-\gamma)$
  Execute π_j for $\tau - 1$ steps from a sampled initial state, giving state s
  Sample action $a \sim \pi_j(\cdot \mid s)$
  Sample $h \geq 1$ with probability $\gamma^{h-1}(1-\gamma)$
  Continue the rollout from (s, a) by executing π for $h - 1$ steps,
  giving the rollout $\mathcal{P} = \{(s_1, a_1, \ldots, s_h, a_h)\}$ where $(s_1, a_1) = (s, a)$
  $G = \frac{1}{1-\gamma}[r(s_h, a_h) + b(s_h, a_h)]$
  $\mathcal{D} \leftarrow \mathcal{D} \cup \{(\phi(s, a), \mathcal{P}, G, b(s, a))\}$
end for
return $\mathcal{D}$

Algorithm 5 FEATURESAMPLER(π)

Sample $\tau \geq 1$ with probability $\gamma^{\tau-1}(1-\gamma)$
Execute π for $\tau - 1$ steps from a sampled initial state, giving state s
Sample action $a \sim \pi(\cdot \mid s)$
return $\phi(s, a)$

Appendix B. Additional Related Literature

Exploration has been widely studied in the tabular setting (Azar et al., 2017; Zanette and Brunskill, 2019; Efroni et al., 2019; Jin et al., 2018; Dann et al., 2019; Zhang et al., 2020; Russo, 2019), but obtaining formal guarantees for exploration with function approximation is a challenge even in the linear case due to recent lower bounds (Du et al., 2019; Weisz et al., 2020; Zanette, 2020; Wang et al., 2020a). When the action-value function is only approximately linear, several ideas from tabular exploration and linear bandits (Lattimore and Szepesvári, 2020) have been combined to obtain provably efficient algorithms in low-rank MDPs (Yang and Wang, 2020; Zanette et al., 2020a; Jin et al., 2020) and their extensions (Wang et al., 2019, 2020b). Minimax regret bounds for under little or zero inherent Bellman error (a superset of low-rank MDPs) is given in (Zanette et al., 2020b) and a computationally tractable algorithm for that setting has recently been proposed (Zanette et al., 2020c). No inherent Bellman error is a subset of a more general framework of MDPs with low Bellman rank (Jiang et al., 2017) where the inherent Bellman error is allowed to have a low rank structure but no computationally tractable algorithm are known for such general setting (Dann et al., 2018).

Extensions of the linear or low-rank MDP models to kernel and neural function approximation have recently been presented in Yang et al. (2020b). Other linear transition models recently considered include those presented by (Ayoub et al., 2020; Zhou et al., 2020b); for the latter, a minimax algorithm has recently been proposed (Zhou et al., 2020a).

If linearity holds only for the optimal action-value function and one is only interested in identifying an optimal policy (as opposed to a near optimal one), then (Du et al., 2020) provide an algorithm for such setting, although a sample complexity proportional to the inverse gap (which can be exponentially small) must be suffered. Deterministic systems with linear value functions are also learnable in finite horizon by just assuming realizability (Wen and Van Roy, 2013).

Finally there is a rich literature on the convergence properties of policy gradient methods (Kakade and Langford, 2002; Azar et al., 2011; Scherrer and Geist, 2014; Neu et al., 2017; Even-Dar et al., 2009; Geist et al., 2019; Liu et al., 2019; Abbasi-Yadkori et al., 2019; Bhandari and Russo, 2019; Fazel et al., 2018; Agarwal et al., 2020b) although these do not address the exploration setting. Notable exceptions include: (Shani et al., 2020) on tabular domains and (Cai et al., 2020) on a linear MDP model different than the one we consider here and the aforementioned work of Agarwal et al. (2020a).

Appendix C. Additional Notation and MDP Construction

In table Table 1 we define some frequently used symbols that will be used in the following analyses.

Table 1: Symbols

B	$\stackrel{def}{=}$	$\frac{3}{1-\gamma}$
G_{max}	$\stackrel{def}{=}$	$\frac{2+B}{(1-\gamma)}$
W	$\stackrel{def}{=}$	$2G_{max}$
κ	$\stackrel{def}{=}$	see Eq. (129)
λ	$\stackrel{def}{=}$	λ_{min}, see Eq. (212)
β	$\stackrel{def}{=}$	see Eq. (136)
$\mathcal{B}$	$\stackrel{def}{=}$	$\{v \in \mathbb{R}^d \mid \|v\|_2 \leq 1\}$
t_{max}	$\stackrel{def}{=}$	$\frac{\ln(16N^2K/\delta)}{1-\gamma}$ (maximum high probability trajectory length Lemma 31 *(Trajectory Boundness)*)

We denote with n the *outer* iterations (see Algorithm 1) and with k the *inner* iterations (see Algorithm 2). We use the outer iteration index n as superscript and the inner iteration index k as subscript to indicate that a certain quantity that is computed in the outer iteration n and the inner iteration k, respectively.

Transfer error on linear MDPs On linear MDPs, the transfer error in Definition 3 is exactly zero, i.e., $\mathcal{E} = 0$. This follows by combining Claim D.1 with Lemma D.1 in (Agarwal et al., 2020a).

Average policy and cover In the analysis we use the concept of average policy or policy mixture.

Definition 13 (Average Policy) *Given policies $\pi^0, \ldots, \pi^{n-1}$ let the average policy $\pi^{0:n-1}$ be defined as follows: sample $i \in \{0, 1, \ldots, n-1\}$ with uniform probability and the follow π^i for the episode.*

Let d^π be the distribution over state-actions induced by policy π, and let $\rho^n_{cov} = \frac{1}{n}\sum_{i=0}^{n-1} d^{\pi_i}$ be that induced by $\pi^{0:n-1}$.

Remark on expressions containing mixture policies We highlight that when the mixture policy π^n appears in an expression, for notational convenience it is intended that the whole expression is averaged. For example, when writing the expected bonus $\mathbb{E}_{s\sim\pi^n} b(s, \pi^n) = \frac{1}{K}\sum_{k=0}^{K-1} \mathbb{E}_{s\sim\pi^k_n} b(s, \pi^n_k)$. This is to be consistent with the way the mixture policies are defined (and the way the algorithm operates), where an index j in $\{0, \ldots, K-1\}$ is sampled uniformly at random and then policy π_j is followed for the full episode; to be consistent, all quantities must then refer to the same policy π_j, for example $\widehat{A}^n(s,a) = \widehat{Q}^n(s,a) - \widehat{Q}^n(s,\pi^n) = \sum_{k=0}^{K-1}\left(\widehat{Q}^n_j(s,a) - \widehat{Q}^n_j(s,\pi^n_j)\right) = \sum_{k=0}^{K-1}\left(\widehat{Q}^n_j(s,a) - \widehat{V}^n_j(s)\right)$.

Known states We define the set of known state-actions in a certain outer episode n (this stays constant for all inner iterations k of Algorithm 2 as n is fixed) the following set (we overload the notation as there is no possibility of confusion)

$$\mathcal{K}^n \stackrel{def}{=} \left\{ s \in \mathcal{S} \mid \forall a \in \mathcal{A}, \sqrt{\beta}\|\phi(s,a)\|_{(\widehat{\Sigma}^n)^{-1}} < 1 \right\} \tag{11}$$

$$\mathcal{K}^n \stackrel{def}{=} \left\{ (s,a) \in \mathcal{S} \times \mathcal{A} \mid \sqrt{\beta}\|\phi(s,a)\|_{(\widehat{\Sigma}^n)^{-1}} < 1 \right\}. \tag{12}$$

Inner policies The inner policies $\pi_0, \pi_1, \ldots$ are those computed by Algorithm 2 and are defined as:

$$\forall s \in \mathcal{K}^n : \qquad \pi_{k+1}(\cdot \mid s) \propto \pi_k(\cdot \mid s) e^{\eta \widehat{Q}_k(\cdot|s)} \tag{13}$$

$$\forall s \notin \mathcal{K}^n : \qquad \pi_{k+1}(\cdot \mid s) = \text{Unif}(\{a \mid (s,a) \notin \mathcal{K}^n\}) \tag{14}$$

The initialization is

$$\forall s \in \mathcal{K}^n : \qquad \pi_0(\cdot \mid s) = \text{Unif}(\mathcal{A}) \tag{15}$$

$$\forall s \notin \mathcal{K}^n : \qquad \pi_0(\cdot \mid s) = \text{Unif}(\{a \mid (s,a) \notin \mathcal{K}^n\}) \tag{16}$$

Outer policies The outer policies $\pi^1, \pi^2, \ldots$ are those maintained by Algorithm 1 and they are a mixture of the inner policies computed by the SOLVER. In particular, when the SOLVER terminates it returns a mixture of policies $\pi_{0:K-1}$, and π^n is set to be equivalent to that mixture.

Bonus and optimistic MDP Consider $\mathcal{M} = (\mathcal{S}, \mathcal{A}, p, r, \gamma)$. In iteration n we construct an optimistic MDP with bonus $b^n : \mathcal{S} \times \mathcal{A} \to \mathbb{R}$ defined as $\mathcal{M}^n = (\mathcal{S}, \mathcal{A} \cup \{a^\dagger\}, p, r + b^n, \gamma)$. The bonus function reads as

$$b^n_\phi(s,a) \stackrel{def}{=} \sqrt{\beta}\|\phi(s,a)\|_{(\widehat{\Sigma}^n)^{-1}} \mathbb{1}\{s \in \mathcal{K}^n\} \tag{17}$$

$$b^n_{\mathbb{1}}(s,a) \stackrel{def}{=} \frac{3}{1-\gamma} \mathbb{1}\{(s,a) \notin \mathcal{K}^n\} \tag{18}$$

$$b^n(s,a) \stackrel{def}{=} 2b^n_\phi(s,a) + b^n_{\mathbb{1}}(s,a). \tag{19}$$

The bonus behaves as follows. In any state, if $s \in \mathcal{K}^n$ then $b^n(s,a) = 2b^n_\phi(s,a)$ and if $(s,a) \notin \mathcal{K}^n$ then $b^n(s,a) = b^n_{\mathbb{1}}(s,a)$. In particular, the bonus are in the range $[0,2)$ if the state-action is known, and otherwise the bonus is deterministically set to $\frac{3}{(1-\gamma)}$. Notice that a state-action (s,a) such that $s \notin \mathcal{K}^n$ but $(s,a) \in \mathcal{K}^n$ has zero bonus (or generally we can set an arbitrarily value here); the specific bonus value at such a state-action is irrelevant as the algorithm's policy π^n by construction (cf. Eq. (13)) always takes an action with the indicator bonus $b^n_{\mathbb{1}}(s,a)$ if the state $s \notin \mathcal{K}^n$.

The optimistic MDP has an extra action $a^\dagger$ that self loops in the current state with probability 1 with a reward $r(s,a^\dagger) = 3$. The bonus function $b^n(s,a^\dagger) = 0$. (The agent is not even aware of the existence of $a^\dagger$; this extra action $a^\dagger$ is introduced purely for analysis.) Denote the state-action value function of a generic policy π on $\mathcal{M}^n$ with $Q^{n,\pi}(s,a) = \frac{1}{1-\gamma}\mathbb{E}_{(s',a')\sim\pi|(s,a)}[r(s',a') + b^n(s',a')]$. The state value function is denoted with $V^{n,\pi}(s) = \mathbb{E}_{a\sim\pi(\cdot|s)} Q^{n,\pi}(s,a)$.

Let π^n be the policy identified by the agent in the outer episode n (the policy returned by Algorithm 2). We define $Q^n = Q^{n,\pi^n}$, $V^n = V^{n,\pi^n}$ for brevity.

Approximators On the known states, in outer iteration n and inner iteration k, we define the best Q-approximator $Q_k^{n,\star}$ and the agent's approximator $\widehat{Q}_k^n$ as

$$\text{if } s \in \mathcal{K}^n : \begin{cases} Q_k^{n,\star}(s,a) & = \phi(s,a)^\top w_k^{n,\star} + 2b_\phi^n(s,a) \\ \widehat{Q}_k^n(s,a) & = \phi(s,a)^\top \widehat{w}_k^n + b_\phi^n(s,a) \end{cases} \tag{20}$$

Otherwise, we set them to be the same as $b^n(s,a)$. We omit either n or k when there is no possibility of confusion.

Appendix D. Main Analysis

We start our analysis by showing some auxiliary lemmas which we will later use to prove Proposition 16. In particular, Lemma 14 and Lemma 15 are variations of the corresponding lemmas in (Agarwal et al., 2020a).

We start by recalling the performance difference lemma (e.g., (Kakade and Langford, 2002)) which states that for any two policies π, π' we can write

$$(V^{\pi'} - V^{\pi})(s_0) = \frac{1}{1-\gamma}\,\mathbb{E}_{(s,a)\sim\pi}\, A^{\pi'}(s,a) \tag{21}$$

where A^{π} is the advantage function associated with π.

The following lemma is similar to lemma B.2 (Agarwal et al., 2020a).

Lemma 14 (Partial optimism) *Fix a policy $\widetilde{\pi}$ that never takes $a^{\dagger}$. Define the policy $\widetilde{\pi}^n$ on $\mathcal{M}^n$ such that $\widetilde{\pi}^n(\cdot \mid s) = \widetilde{\pi}(\cdot \mid s)$ if $s \in \mathcal{K}^n$ and $\widetilde{\pi}^n(a^{\dagger} \mid s) = 1$ if $s \notin \mathcal{K}^n$. In any episode n it holds that*

$$V^{\widetilde{\pi}}(s_0) + \frac{1}{1-\gamma}\,\mathbb{E}_{s\sim\widetilde{\pi}|s_0}\, 2b^n_{\phi}(s,\widetilde{\pi}) \le V^{n,\widetilde{\pi}^n}(s_0). \tag{22}$$

Proof Notice that $\widetilde{\pi}^n$ always takes an action where $b^n_{\mathbb{1}}(s,a) = 0$. A quick computation gives:

$$V^{n,\widetilde{\pi}^n}(s) \le \frac{3}{1-\gamma}. \tag{23}$$

and in particular, if $s \notin \mathcal{K}^n$ then $V^{n,\widetilde{\pi}^n}(s) = \frac{3}{1-\gamma}$ as the policy self-loops in s by taking $a^{\dagger}$ there. Using the performance difference lemma we get:

$$(1-\gamma)\left(V^{n,\widetilde{\pi}^n}(s_0) - V^{n,\widetilde{\pi}}(s_0)\right) = \tag{24}$$

$$= \mathbb{E}_{(s,a)\sim\widetilde{\pi}|s_0}\left[Q^{n,\widetilde{\pi}^n}(s,\widetilde{\pi}^n) - Q^{n,\widetilde{\pi}^n}(s,\widetilde{\pi})\right] \tag{25}$$

$$= \mathbb{E}_{(s,a)\sim\widetilde{\pi}|s_0}\left[\left(Q^{n,\widetilde{\pi}^n}(s,\widetilde{\pi}^n) - Q^{n,\widetilde{\pi}^n}(s,\widetilde{\pi})\right)\mathbb{1}\{s \notin \mathcal{K}^n\}\right] \tag{26}$$

$$= \mathbb{E}_{(s,a)\sim\widetilde{\pi}|s_0}\left[\left(\frac{3}{1-\gamma} - Q^{n,\widetilde{\pi}^n}(s,\widetilde{\pi})\right)\mathbb{1}\{s \notin \mathcal{K}^n\}\right] \tag{27}$$

$$= \mathbb{E}_{(s,a)\sim\widetilde{\pi}|s_0}\left[\left(\frac{3}{1-\gamma} - \underbrace{r(s,\widetilde{\pi})}_{\le 1} - \underbrace{2b^n_{\phi}(s,\widetilde{\pi})}_{\le 2} - b^n_{\mathbb{1}}(s,\widetilde{\pi}) - \underbrace{\gamma\,\mathbb{E}_{s'\sim p(s,\widetilde{\pi})}\, V^{n,\widetilde{\pi}^n}(s')}_{\le \frac{3\gamma}{1-\gamma} \text{ by Eq. (23)}}\right)\mathbb{1}\{s \notin \mathcal{K}^n\}\right]. \tag{28}$$

We have $\underbrace{r(s,\widetilde{\pi})}_{\le 1} + \underbrace{2b^n_{\phi}(s,\widetilde{\pi})}_{\le 2} + \underbrace{\gamma\,\mathbb{E}_{s'\sim p(s,\widetilde{\pi})}\, V^{n,\widetilde{\pi}^n}(s')}_{\le \frac{3\gamma}{1-\gamma} \text{ by Eq. (23)}} \le 3 + \frac{3\gamma}{1-\gamma} \le \frac{3}{1-\gamma}$. Continuing the chain above:

$$\ge \mathbb{E}_{(s,a)\sim\widetilde{\pi}|s_0}\left[-b^n_{\mathbb{1}}(s,\widetilde{\pi})\mathbb{1}\{s \notin \mathcal{K}^n\}\right] \tag{29}$$

$$= \mathbb{E}_{(s,a)\sim\widetilde{\pi}|s_0}\left[-b^n_{\mathbb{1}}(s,\widetilde{\pi})\right]. \tag{30}$$

Thus,

$$V^{n,\widetilde{\pi}^n}(s_0) \geq V^{n,\widetilde{\pi}}(s_0) - \frac{1}{1-\gamma}\,\mathbb{E}_{(s,a)\sim\widetilde{\pi}|s_0}\, b^n_{\mathbb{1}}(s,a) \tag{31}$$

$$= V^{\widetilde{\pi}}(s_0) + \frac{1}{1-\gamma}\,\mathbb{E}_{(s,a)\sim\widetilde{\pi}|s_0}\, b^n(s,a) - \frac{1}{1-\gamma}\,\mathbb{E}_{(s,a)\sim\widetilde{\pi}|s_0}\, b^n_{\mathbb{1}}(s,a) \tag{32}$$

$$= V^{\widetilde{\pi}}(s_0) + \frac{1}{1-\gamma}\,\mathbb{E}_{(s,a)\sim\widetilde{\pi}|s_0}\, 2b^n_{\phi}(s,a) \tag{33}$$

■

The following lemma is similar to lemma A.1 in (Agarwal et al., 2020a).

Lemma 15 (Negative Advantage) *We have*

$$A^n(s,\widetilde{\pi}^n)\mathbb{1}\{s \notin \mathcal{K}^n\} \leq 0.$$

Proof Assume $s \notin \mathcal{K}^n$. In such state, $\widetilde{\pi}^n$ takes action $a^\dagger$ and self-loops in s, where a reward $= 3$ is received for the first timestep. Thus, for $Q^n = Q^{n,\pi^n}$,

$$Q^n(s,\widetilde{\pi}^n) = 3 + \gamma V^n(s).$$

In addition, in $s \notin \mathcal{K}^n$ an action $a \neq a^\dagger$ such that $b^n_{\mathbb{1}}(s,a) = \frac{3}{1-\gamma}$ must exist. In such case, π^n always takes one such action; this is because π^n by definition is a mixture of the policies $\pi_1, \ldots, \pi_{K-1}$ computed by Algorithm 2, and they all choose an action with the indicator bonus if the state $s \notin \mathcal{K}^n$, see line 2 in Algorithm 2. Therefore

$$V^n(s) \geq \frac{3}{1-\gamma}.$$

Combining the two expressions we obtain that, in any state $s \notin \mathcal{K}^n$,

$$A^n(s,\widetilde{\pi}^n) = Q^n(s,\widetilde{\pi}^n) - V^n(s) = \Big[3 + \gamma V^n(s) - V^n(s)\Big] = 3 - (1-\gamma)V^n(s) \leq 0.$$

■

Proposition 16 (Analysis of COPOE) *With probability at least* $1-\delta$ *it holds that*

$$\frac{1}{N}\sum_{n=1}^{N}\left(V^{\widetilde{\pi}} - V^{\pi^n}\right)(s_0) \leq \frac{\mathcal{R}(K)}{(1-\gamma)K} + \frac{2\sqrt{2A\mathcal{E}_n}}{1-\gamma} + \frac{1}{\sqrt{N}} \times \widetilde{O}\left(\frac{\sqrt{\beta d}}{(1-\gamma)^2}\right) \tag{34}$$

Proof Fix a policy $\widetilde{\pi}$ on $\mathcal{M}$ ($\widetilde{\pi}$ does not take $a^\dagger$ since $a^\dagger$ is not available on $\mathcal{M}$). Consider the following decomposition for an outer episode n (recall the policy π^n is the mixture policy of the policies $\pi_0, \ldots, \pi_{K-1}$ computed by the SOLVER, see Appendix C for more details)

$$\left(V^{\widetilde{\pi}} - V^{\pi^n}\right)(s_0) = \underbrace{V^{\widetilde{\pi}}(s_0) + \frac{1}{1-\gamma}\,\mathbb{E}_{s\sim\widetilde{\pi}|s_0}\, 2b^n_{\phi}(s,\widetilde{\pi})}_{\leq V^{n,\widetilde{\pi}^n}(s_0) \text{ by Lemma 14}} \underbrace{- V^{\pi^n}(s_0) - \frac{1}{1-\gamma}\,\mathbb{E}_{s\sim\pi^n|s_0}\, b^n(s,\pi^n)}_{\stackrel{def}{=} -V^n(s_0)} \tag{35}$$

$$+ \frac{1}{1-\gamma}\underbrace{\Big[-\mathbb{E}_{s\sim\widetilde{\pi}|s_0}\, 2b^n_{\phi}(s,\widetilde{\pi}) + \mathbb{E}_{s\sim\pi^n|s_0}\, b^n(s,\pi^n)\Big]}_{\stackrel{def}{=} B^n}$$

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

$$\begin{aligned} V^{n,\widetilde{\pi}^n}(s_0) - V^n(s_0) &= \frac{1}{1-\gamma}\,\mathbb{E}_{s\sim\widetilde{\pi}^n|s_0}\Big[\underbrace{Q^n(s,\widetilde{\pi}^n) - V^n(s)}_{A^n(s,\widetilde{\pi}^n)}\Big] \\ &= \frac{1}{1-\gamma}\,\mathbb{E}_{s\sim\widetilde{\pi}^n|s_0}\Big[A^n(s,\widetilde{\pi}^n)\mathbb{1}\{s\in\mathcal{K}^n\} + \underbrace{A^n(s,\widetilde{\pi}^n)\mathbb{1}\{s\notin\mathcal{K}^n\}}_{\leq 0 \text{ by Lemma 15}}\Big] \\ &= \frac{1}{1-\gamma}\,\mathbb{E}_{s\sim\widetilde{\pi}^n|s_0}\Big[A^n(s,\widetilde{\pi})\mathbb{1}\{s\in\mathcal{K}^n\}\Big] \end{aligned} \tag{36}$$

where the last step is because on states $s \in \mathcal{K}^n$ we have $\widetilde{\pi}^n(\cdot \mid s) = \widetilde{\pi}(\cdot \mid s)$; using this, we can derive

$$\begin{aligned} &= \frac{1}{1-\gamma}\Big[\mathbb{E}_{s\sim\widetilde{\pi}^n|s_0}\,\widehat{A}^n(s,\widetilde{\pi})\mathbb{1}\{s\in\mathcal{K}^n\} + \mathbb{E}_{s\sim\widetilde{\pi}^n|s_0}\Big[A^n(s,\widetilde{\pi}) - \widehat{A}^n(s,\widetilde{\pi})\Big]\mathbb{1}\{s\in\mathcal{K}^n\}\Big] \\ &\leq \frac{1}{1-\gamma}\Bigg[\underbrace{\sup_{s\in\mathcal{K}^n}\widehat{A}^n(s,\widetilde{\pi})}_{\text{term 1}} + \underbrace{\mathbb{E}_{s\sim\widetilde{\pi}^n|s_0}\Big[A^n(s,\widetilde{\pi}) - A^{n,\star}(s,\widetilde{\pi})\Big]\mathbb{1}\{s\in\mathcal{K}^n\}}_{\text{term 2}} \\ &\qquad + \underbrace{\mathbb{E}_{s\sim\widetilde{\pi}^n|s_0}\Big[A^{n,\star}(s,\widetilde{\pi}) - \widehat{A}^n(s,\widetilde{\pi})\Big]\mathbb{1}\{s\in\mathcal{K}^n\}}_{\text{term 3}}\Bigg]. \end{aligned} \tag{37}$$

The second term is the approximation error, and we can bound it as follows by taking absolute values and using Lemma 18 *(Distribution Dominance)*

$$\begin{aligned} \mathbb{E}_{s\sim\widetilde{\pi}^n|s_0}\Big[A^n(s,\widetilde{\pi}) - A^{n,\star}(s,\widetilde{\pi})\Big]\mathbb{1}\{s\in\mathcal{K}^n\} &\leq \mathbb{E}_{s\sim\widetilde{\pi}^n|s_0}\Big|A^n(s,\widetilde{\pi}) - A^{n,\star}(s,\widetilde{\pi})\Big|\mathbb{1}\{s\in\mathcal{K}^n\} \\ &\leq \mathbb{E}_{s\sim\widetilde{\pi}|s_0}\Big|A^n(s,\widetilde{\pi}) - A^{n,\star}(s,\widetilde{\pi})\Big|\mathbb{1}\{s\in\mathcal{K}^n\} \\ &= \mathbb{E}_{s\sim\widetilde{\pi}|s_0}\Big|\frac{1}{K}\sum_{k=0}^{K-1} A_k^n(s,\widetilde{\pi}) - A_k^{n,\star}(s,\widetilde{\pi})\Big|\mathbb{1}\{s\in\mathcal{K}^n\} \end{aligned} \tag{38}$$

Now we focus on the third term in (37); Lemma 30 *(Validity of Confidence Intervals)* ensures that with probability at least $1-\frac{\delta}{2}$ it holds that

$$\forall n\in[N],\ \forall k\in\{0,\dots,K-1\},\ \forall(s,a)\in\mathcal{K}^n:\quad 0 \leq Q_k^{n,\star}(s,a) - \widehat{Q}_k^n(s,a) \leq 2b_\phi^n(s,a). \tag{39}$$

In what follows we omit the subscript k as we need the bound to hold only for the mixture policy π^n. Then $\forall n \in [N],\ \forall (s,a) \in \mathcal{K}^n$:

$$A^{n,\star}(s,a) - \widehat{A}^n(s,a) = \frac{1}{K}\sum_{k=0}^{K-1}\Bigg[\left(Q_k^{n,\star}(s,a) - \widehat{Q}_k^n(s,a)\right) - \underbrace{\left(Q_k^{n,\star}(s,\pi_k^n) - \widehat{Q}_k^n(s,\pi_k^n)\right)}_{\leq 0}\Bigg] \tag{40}$$

$$\leq Q^{n,\star}(s,a) - \widehat{Q}^n(s,a). \tag{41}$$

The right hand side is by definition positive using Eq. (39). Thus Lemma 18 *(Distribution Dominance)* can be applied to obtain

$$\mathbb{E}_{s\sim\widetilde{\pi}^n|s_0}\left[A^{n,\star}(s,\widetilde{\pi}) - \widehat{A}^n(s,\widetilde{\pi})\right]\mathbb{1}\{s\in\mathcal{K}^n\} \leq \mathbb{E}_{s\sim\widetilde{\pi}^n|s_0}\left[Q^{n,\star}(s,\widetilde{\pi}) - \widehat{Q}^n(s,\widetilde{\pi})\right]\mathbb{1}\{s\in\mathcal{K}^n\} \tag{42}$$

$$\leq \mathbb{E}_{s\sim\widetilde{\pi}|s_0}\left[Q^{n,\star}(s,\widetilde{\pi}) - \widehat{Q}^n(s,\widetilde{\pi})\right]\mathbb{1}\{s\in\mathcal{K}^n\}. \tag{43}$$

Together, plugging Eqs. (38) and (42) back into Eq. (37), Eq. (36) and finally Eq. (35) gives

$$\left(V^{\widetilde{\pi}} - V^{\pi^n}\right)(s_0) \leq \frac{1}{1-\gamma}\Bigg[\underbrace{\sup_{s\in\mathcal{K}^n}\widehat{A}^n(s,\widetilde{\pi})}_{\text{term 1}} + \underbrace{\mathbb{E}_{s\sim\widetilde{\pi}|s_0}\left|A^n(s,\widetilde{\pi}) - A^{n,\star}(s,\widetilde{\pi})\right|\mathbb{1}\{s\in\mathcal{K}^n\}}_{\text{term 2}} \tag{44}$$

$$+\underbrace{\mathbb{E}_{s\sim\widetilde{\pi}|s_0}\left[Q^{n,\star}(s,\widetilde{\pi}) - \widehat{Q}^n(s,\widetilde{\pi})\right]\mathbb{1}\{s\in\mathcal{K}^n\}}_{\text{term 3}} + B^n\Bigg]$$

We can bound term 1 using Lemma 19 *(NPG lemma)*. We then obtain (the online regret $\mathcal{R}(K)$ is defined in the lemma)

$$\sup_{s\in\mathcal{S}}\widehat{A}^n(s,\widetilde{\pi})\mathbb{1}\{s\in\mathcal{K}^n\} = \sup_{s\in\mathcal{S}}\frac{1}{K}\sum_{k=0}^{K-1}\mathbb{E}_{a\sim\widetilde{\pi}(\cdot|s)}\widehat{A}_k^n(s,a)\mathbb{1}\{s\in\mathcal{K}^n\} \leq \frac{\mathcal{R}(K)}{K}. \tag{45}$$

We can bound the second term in the prior display by invoking Lemma 17 *(Advantage Transfer Error Decomposition)*. The third term is finally bounded by Eq. (39). As a result, the performance difference has an upper bound:

$$\left(V^{\widetilde{\pi}} - V^{\pi^n}\right)(s_0) \leq \frac{1}{1-\gamma}\left[\frac{\mathcal{R}(K)}{K} + 2\sqrt{2A\mathcal{E}^n} + \mathbb{E}_{(s,a)\sim\widetilde{\pi}|s_0}\, 2b_\phi^n(s,a)\mathbb{1}\{s\in\mathcal{K}^n\} + B^n\right] \tag{46}$$

$$= \frac{1}{1-\gamma}\left[\frac{\mathcal{R}(K)}{K} + 2\sqrt{2A\mathcal{E}^n} + \mathbb{E}_{s\sim\pi^n|s_0}\, b^n(s,\pi^n)\right] \tag{47}$$

Averaging over the outer rounds $n \in [N]$ and defining $\sqrt{\mathcal{E}} \overset{def}{=} \frac{1}{N}\sum_{n=1}^N \sqrt{\mathcal{E}^n}$ (where $\sqrt{\mathcal{E}^n}$ itself is an average of the SOLVER's errors $\sqrt{\mathcal{E}^n} \overset{def}{=} \frac{1}{K}\sum_{k=0}^{K-1}\sqrt{\mathcal{E}_k^n}$) gives

$$\frac{1}{N}\sum_{n=1}^N\left(V^{\widetilde{\pi}} - V^{\pi^n}\right)(s_0) \leq \frac{\mathcal{R}(K)}{(1-\gamma)K} + \frac{2\sqrt{2A\mathcal{E}}}{1-\gamma} + \frac{1}{N(1-\gamma)}\sum_{n=1}^N\mathbb{E}_{s\sim\pi^n|s_0}\, b^n(s,\pi^n) \tag{48}$$

Finally, Lemma 33 *(Bounding the Sum of Bonuses)* and Lemma 32 *(Bounding the Sum of Indicators)* and a union bound conclude. ■

The following lemma is similar to lemma C.1 in (Agarwal et al., 2020a).

Lemma 17 (Advantage Transfer Error Decomposition) *We have*

$$\mathbb{E}_{s\sim\widetilde{\pi}|s_0}\left|A^n(s,\widetilde{\pi})-\widehat{A}^{n,\star}(s,\widetilde{\pi})\right|\mathbb{1}\{s\in\mathcal{K}^n\}\leq 2\sqrt{2A\mathcal{E}^n}. \tag{49}$$

Proof We leave the conditioning on the starting state s_0 implicit. Using Definition 3 *(Transfer Error)*

$$\begin{aligned}
&= \mathbb{E}_{s\sim\widetilde{\pi}}\,\mathbb{E}_{a\sim\widetilde{\pi}(\cdot|s)}\left|\left(A^n(s,a)-A^{n,\star}(s,a)\right)\mathbb{1}\{s\in\mathcal{K}^n\}\right| && (50)\\
&\leq \mathbb{E}_{s\sim\widetilde{\pi}}\,\mathbb{E}_{a\sim\widetilde{\pi}(\cdot|s)}\left|\left(Q^n(s,a)-Q^{n,\star}(s,a)\right)\mathbb{1}\{s\in\mathcal{K}^n\}\right| \\
&\quad+\mathbb{E}_{s\sim\widetilde{\pi}}\,\mathbb{E}_{a\sim\pi^n(\cdot|s)}\left|\left(Q^n(s,a)-Q^{n,\star}(s,a)\right)\mathbb{1}\{s\in\mathcal{K}^n\}\right| && (51)\\
&\overset{\text{Jensen}}{\leq} \sqrt{\mathbb{E}_{s\sim\widetilde{\pi}}\,\mathbb{E}_{a\sim\widetilde{\pi}(\cdot|s)}\left[\left(Q^n(s,a)-Q^{n,\star}(s,a)\right)^2\mathbb{1}\{s\in\mathcal{K}^n\}\right]} && (52)\\
&\qquad+\sqrt{\mathbb{E}_{s\sim\widetilde{\pi}}\,\mathbb{E}_{a\sim\pi^n(\cdot|s)}\left[\left(Q^n(s,a)-Q^{n,\star}(s,a)\right)^2\mathbb{1}\{s\in\mathcal{K}^n\}\right]} && (53)\\
&\leq \sqrt{|\mathcal{A}|\,\mathbb{E}_{s\sim\widetilde{\pi}}\,\mathbb{E}_{a\sim\text{Unif}|\mathcal{A}|}\left[\left(Q^n(s,a)-Q^{n,\star}(s,a)\right)^2\right]}+\sqrt{|\mathcal{A}|\,\mathbb{E}_{s\sim\widetilde{\pi}}\,\mathbb{E}_{a\sim\text{Unif}|\mathcal{A}|}\left[\left(Q^n(s,a)-Q^{n,\star}(s,a)\right)^2\right]} && (54)\\
&\leq 2\sqrt{|\mathcal{A}|}\sqrt{2\mathcal{L}(w^{n,\star},d^{\widetilde{\pi}}\circ\text{Unif}|\mathcal{A}|,Q^n-b^n)} && (55)\\
&\leq 2\sqrt{2|\mathcal{A}|\mathcal{E}^n}. && (56)
\end{aligned}$$

■

The following lemma is similar to B.1 in (Agarwal et al., 2020a).

Lemma 18 (Distribution Dominance) *If $f:\mathcal{S}\to\mathbb{R}$ is a positive function then we have*

$$\mathbb{E}_{s\sim\widetilde{\pi}^n}f(s)\mathbb{1}\{s\in\mathcal{K}^n\}\leq\mathbb{E}_{s\sim\widetilde{\pi}}f(s)\mathbb{1}\{s\in\mathcal{K}^n\}.$$

Proof Consider the MDP $\mathcal{M}^n$ but with $f(s)\mathbb{1}\{s\in\mathcal{K}^n\}$ as the total reward function in s and let $\widetilde{Q},\widetilde{V}$ be the value functions. Recall that the reward is positive and that $\widetilde{\pi}^n$ circles back to $s\notin\mathcal{K}^n$

once such state is reached. Then the performance difference lemma ensures

$$\mathbb{E}_{s\sim\widetilde{\pi}^n} f(s)\mathbb{1}\{s \in \mathcal{K}^n\} - \mathbb{E}_{s\sim\widetilde{\pi}} f(s)\mathbb{1}\{s \in \mathcal{K}^n\} \tag{57}$$

$$= (1-\gamma)(\widetilde{V}^{\widetilde{\pi}^n} - \widetilde{V}^{\widetilde{\pi}}) \tag{58}$$

$$= \mathbb{E}_{s\sim\widetilde{\pi}}\left[\left(\underbrace{\widetilde{Q}^{\widetilde{\pi}^n}(s,\widetilde{\pi}^n)}_{=0} - \widetilde{Q}^{\widetilde{\pi}^n}(s,\widetilde{\pi})\right)\mathbb{1}\{s \notin \mathcal{K}^n\}\right] \tag{59}$$

$$+ \mathbb{E}_{s\sim\widetilde{\pi}}\left[\left(\underbrace{\widetilde{Q}^{\widetilde{\pi}^n}(s,\widetilde{\pi}^n)}_{=\widetilde{Q}^{\widetilde{\pi}^n}(s,\widetilde{\pi})} - \widetilde{Q}^{\widetilde{\pi}^n}(s,\widetilde{\pi})\right)\mathbb{1}\{s \in \mathcal{K}^n\}\right] \tag{60}$$

$$\leq 0. \tag{61}$$

■

Appendix E. NPG Guarantees

Consider a fixed episode n where NPG is invoked (we omit the dependence on n in the notation) and notice that the set $\mathcal{K}^n$ is fixed.

Lemma 19 (NPG lemma) *Fix n. If $K \geq 4 \ln |\mathcal{A}|$ and the learning rate is $\eta = \frac{\sqrt{\ln |\mathcal{A}|}}{\sqrt{KW}}$, then $\eta |\widehat{A}_k(\cdot, \cdot)| \leq 1$ and we have for any fixed state $s \in \mathcal{K}^n$ and distribution $\widetilde{\pi}(\cdot \mid s)$*

$$\sum_{k=0}^{K-1} \mathbb{E}_{a \sim \widetilde{\pi}(\cdot|s)} \widehat{A}_k(s,a) \mathbb{1}\{s \in \mathcal{K}^n\} \leq 2W \sqrt{\ln |\mathcal{A}| K} \stackrel{def}{=} \mathcal{R}(K). \tag{62}$$

Proof The update rule in known states reads

$$\begin{aligned} \pi_{k+1}(\cdot \mid s) &\propto \pi_k(\cdot \mid s) e^{\eta \widehat{Q}_k(s,\cdot)} && (63)\\ &\propto \pi_k(\cdot \mid s) e^{\eta \widehat{Q}_k(s,\cdot)} e^{-\eta \widehat{V}_k(s)} && (64)\\ &= \pi_k(\cdot \mid s) e^{\eta \widehat{A}_k(s,\cdot)}. && (65) \end{aligned}$$

Denote the normalizer $z_k(s) = \sum_{a'} \pi_k(a' \mid s) e^{\eta \widehat{A}_k(s,a')}$. The update rule in known states can be written as

$$\pi_{k+1}(\cdot \mid s) = \frac{\pi_k(\cdot \mid s) e^{\eta \widehat{A}_k(s,\cdot)}}{z_k(s)}. \tag{66}$$

Then we have the following equality for any state $s \in \mathcal{K}$:

$$\begin{aligned} &\mathrm{KL}(\widetilde{\pi}(\cdot \mid s) \,||\, \pi_{k+1}(\cdot \mid s)) - \mathrm{KL}(\widetilde{\pi}(\cdot \mid s) \,||\, \pi_k(\cdot \mid s))\\ &= \sum_a \widetilde{\pi}(a \mid s) \ln \frac{\widetilde{\pi}(a \mid s)}{\pi_{k+1}(a \mid s)} - \sum_a \widetilde{\pi}(a \mid s) \ln \frac{\widetilde{\pi}(a \mid s)}{\pi_k(a \mid s)}\\ &= \sum_a \widetilde{\pi}(a \mid s) \ln \frac{\pi_k(a \mid s)}{\pi_{k+1}(a \mid s)}\\ &= \sum_a \widetilde{\pi}(a \mid s) \ln \left(z_k e^{-\eta \widehat{A}_k(s,a)} \right)\\ &= -\eta \sum_a \widetilde{\pi}(a \mid s) \widehat{A}_k(s,a) + \ln z_k(s). \end{aligned} \tag{67}$$

We show that for any know state s we have $\ln z_k(s) \leq \eta^2 W^2$. To see this, we use the fact that $|\eta \widehat{A}_k(\cdot, \cdot)| \leq 1$ which allows us to use the inequality $e^x \leq 1 + x + x^2$ to claim for any known state

$$\begin{aligned} \ln z_k(s) &= \ln \left(\sum_{a'} \pi_k(a' \mid s) e^{\eta \widehat{A}_k(s,a')} \right) && (68)\\ &\leq \ln \left(\sum_{a'} \pi_k(a' \mid s) (1 + \eta \widehat{A}_k(s,a') + \eta^2 \widehat{A}_k^2(s,a')) \right) && (69)\\ &\leq \ln \left(1 + \eta^2 W^2\right) \leq \eta^2 W^2. && (70) \end{aligned}$$

Plugging the above result into Eq. (67) and summing over k gives

$$\mathrm{KL}(\widetilde{\pi}(\cdot \mid s) \mid\mid \pi_K(\cdot \mid s)) - \mathrm{KL}(\widetilde{\pi}(\cdot \mid s) \mid\mid \pi_0(\cdot \mid s)) \tag{71}$$

$$= \sum_{k=0}^{K-1} \Big[\mathrm{KL}(\widetilde{\pi}(\cdot \mid s) \mid\mid \pi_{k+1}(\cdot \mid s)) - \mathrm{KL}(\widetilde{\pi}(\cdot \mid s) \mid\mid \pi_k(\cdot \mid s)) \Big] \tag{72}$$

$$\leq -\eta \sum_{k=0}^{K-1} \sum_{a} \widetilde{\pi}(a \mid s) \widehat{A}_k(s,a) + \eta^2 W^2 K. \tag{73}$$

Recalling that the KL divergence is positive and that $\mathrm{KL}(\widetilde{\pi}(\cdot \mid s) \mid\mid \pi_0(\cdot \mid s)) \leq \ln |\mathcal{A}|$ for know states gives:

$$\eta \sum_{k=0}^{K-1} \mathbb{E}_{a \sim \widetilde{\pi}(\cdot|s)} \widehat{A}_k(s,a) \mathbb{1}\{s \in \mathcal{K}^n\} \leq \ln |\mathcal{A}| + \eta^2 W^2 K. \tag{74}$$

Choosing $\eta = \frac{\sqrt{\ln |\mathcal{A}|}}{\sqrt{KW}}$ finally gives

$$\sum_{k=0}^{K-1} \mathbb{E}_{a \sim \widetilde{\pi}(\cdot|s)} \widehat{A}_k(s,a) \mathbb{1}\{s \in \mathcal{K}^n\} \leq 2W \sqrt{\ln |\mathcal{A}| K} \stackrel{def}{=} \mathcal{R}(K). \tag{75}$$

■

Appendix F. Iteration and Sample Complexity

In this section we examine the sample and iteration complexity of the algorithm

Lemma 20 (Iteration Complexity) *With probability at least* $1-\delta$ *we have*

$$\frac{1}{N}\sum_{n=1}^{N}\left(V^{\widetilde{\pi}}-V^{\pi^n}\right)(s_0)\leq\epsilon+\frac{2\sqrt{2A\mathcal{E}}}{1-\gamma} \tag{76}$$

with the number of inner iterations K *and the number of outer iterations* N *no larger than*

$$K=\widetilde{O}\left(\frac{\ln|\mathcal{A}|W^2}{(1-\gamma)^2\epsilon^2}\right),\qquad N=\widetilde{O}\left(\frac{d^2}{(1-\gamma)^8\epsilon^2}\right). \tag{77}$$

Proof Consider Proposition 16 *(Analysis of* COPOE*)*. We need ensure

$$\frac{\mathcal{R}(K)}{(1-\gamma)K}=\frac{2W}{(1-\gamma)}\sqrt{\frac{\ln|\mathcal{A}|}{K}}\leq\frac{\epsilon}{2}\quad\longrightarrow\quad K=\widetilde{O}\left(\frac{\ln|\mathcal{A}|W^2}{(1-\gamma)^2\epsilon^2}\right). \tag{78}$$

This gives the inner iteration complexity. Next (β comes from Eq. (136))

$$\begin{aligned}\frac{1}{\sqrt{N}}\times\widetilde{O}\left(\frac{\sqrt{\beta d}}{(1-\gamma)^2}\right)\leq\frac{\epsilon}{2}\quad\longrightarrow\quad N&=\widetilde{O}\left(\frac{d\beta}{(1-\gamma)^4\epsilon^2}\right) &(79)\\ &=\widetilde{O}\left(\frac{d}{(1-\gamma)^4\epsilon^2}\right)\times\widetilde{O}\left(\frac{d}{(1-\gamma)^4}\right) &(80)\\ &=\widetilde{O}\left(\frac{d^2}{(1-\gamma)^8\epsilon^2}\right) &(81)\end{aligned}$$

gives the outer iteration complexity. ■

Lemma 21 (Sample Complexity) *In the same setting as Theorem 20* (Iteration Complexity)*, the total number of sampled trajectories is*

$$\widetilde{O}\left(\frac{d^3}{(1-\gamma)^{12}\epsilon^3}\right) \tag{82}$$

or equivalently

$$\widetilde{O}\left(\frac{d^3}{(1-\gamma)^{13}\epsilon^3}\right) \tag{83}$$

samples.

Proof Every time the bonus switches, Algorithm 2 is invoked, and runs for K iterations. From Lemma 29 *(Union Bound)* we know that once data are collected, they can be reused for the next κ policies (defined in Eq. (129)). Let S be the number of bonus switches given in Lemma 38 *(Number*

of Switches); then fresh data is collected a total of (for the definitions of the symbols, please see Table 1)

$$S \times \lceil \frac{K}{\kappa} \rceil = \widetilde{O}\left(d \times \frac{2\ln(1/\delta)(\frac{\sqrt{\ln|\mathcal{A}|}}{\sqrt{KW}})\,(B+W)}{(1-\gamma)\ln 2} \times K \right) \tag{84}$$

$$= \widetilde{O}\left(d \times \left(\frac{B}{W} + 1 \right) \frac{\sqrt{K}}{1-\gamma} \right) \tag{85}$$

$$= \widetilde{O}\left(d \times \frac{1}{1-\gamma} \times \frac{W}{(1-\gamma)\epsilon} \right) \tag{86}$$

$$= O\left(\frac{d}{(1-\gamma)^4 \epsilon} \right) \tag{87}$$

times (as $W \geq B$). Every time data is collected by the critic at most N rollouts are performed, giving the total number of trajectories:

$$N \times \widetilde{O}\left(\frac{d}{(1-\gamma)^4 \epsilon} \right) = \widetilde{O}\left(\frac{d^3}{(1-\gamma)^{12} \epsilon^3} \right). \tag{88}$$

The sample complexity is then obtained by multiplying the above result by t_{max} in Table 1, which is a uniform bound on the trajectory length in the event we consider. ■

Appendix G. Regression with Monte Carlo and Importance Sampling

In this section we derive high probability confidence intervals for Monte Carlo with importance sampling.

- Appendix G.1 *(Importance Sampling Estimator)* gives generic properties of the importance sampling estimator when the target and behavioral policies are not too different.
- Appendix G.2 *(Small Perturbations to Policies)* examines the effect of small perturbations to policies; this is needed in the union bound in the section described below.
- Appendix G.3 *(Regression Guarantees with Importance Sampling)* gives the actual confidence intervals for the Q-values for the algorithm we examine in this paper

G.1. Importance Sampling Estimator

The importance sampling ratio used in this work starts from the timestep $t = 2$: since we are estimating the Q-values of policies, the first state-action from the cover is always fixed, the two policies in the ratio at $t = 1$ cancel each other out.

Definition 22 (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 $\{(s_\tau, a_\tau, r_\tau)\}_{\tau=1,\dots,t}$ 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:*

$$\left(\Pi_{\tau=2}^{t} \frac{\pi(s_\tau, a_\tau)}{\underline{\pi}(s_\tau, a_\tau)}\right) \frac{r_t}{1-\gamma}. \tag{89}$$

In this section, we will focus on a specific type of behavior and target policies, which are related as

$$\forall (s,a), \qquad \pi(a \mid s) = \underline{\pi}(a \mid s) \times \frac{e^{c(s,a)}}{\sum_{a'} \underline{\pi}(a' \mid s) e^{c(s,a')}}, \qquad \sup_{(s,a)} |c(s,a)| \le \frac{(1-\gamma)\ln(1+\epsilon')}{2\ln(1/\delta')}. \tag{90}$$

For such policies, we have the following results.

Lemma 23 (Policy Ratio) *Assume that $\pi, \underline{\pi}$ are related through Eq. (90) and that $c \overset{def}{=} \sup_{(s,a)} |c(s,a)|$. Then*

$$e^{-2c} \le \sup_{(s,a)} \frac{\pi(a \mid s)}{\underline{\pi}(a \mid s)} \le e^{2c}. \tag{91}$$

Proof The following chain of inequalities is true.

$$e^{-2c} \le \frac{e^{-c}}{\sum_{a'} \underline{\pi}(a' \mid s) e^{c}} \le \frac{\pi(a \mid s)}{\underline{\pi}(a \mid s)} = \frac{e^{c(s,a)}}{\sum_{a'} \underline{\pi}(a' \mid s) e^{c(s,a')}} \le \frac{e^{c}}{\sum_{a'} \underline{\pi}(a' \mid s) e^{-c}} = e^{2c}. \tag{92}$$

■

In addition, for the policies in Eq. (90) we can also examine the bias and variance of the importance sampling estimator.

Lemma 24 (Bias and Variance of Importance Sampling Estimator) *Let $\underline{\pi}$ be a fixed behavioral policy. If π is a fixed target policy with the same support as $\underline{\pi}$ then Eq. (89) is an unbiased estimator of the value of π from (s,a). In addition, assume $\pi, \underline{\pi}$ are related by Eq. (90) where in particular $c(s,a)$ satisfies the constraint in Eq. (90). Let R_{max} be a deterministic upper bound to the maximum absolute value of the reward r_t. Then with probability at least $1-\delta'$ the importance sampling estimator in Eq. (89) is bounded in absolute value by $\frac{1+\epsilon'}{1-\gamma}R_{max}$ and the random timestep t in the importance sampling estimator is bounded by $\frac{\ln 1/\delta'}{1-\gamma}$.*

Proof It is well known that the importance sampling estimator is unbiased (Precup, 2000). For the high probability bound we proceed as follows. Using Lemma 23 *(Policy Ratio)* we claim

$$\left(\sup_{(s,a)} \frac{\pi(a \mid s)}{\underline{\pi}(a \mid s)}\right)^{t-1} \leq e^{2(t-1)c}. \tag{93}$$

We show that t is small with high probability:

$$\begin{aligned}
\mathbf{P}(t > \tau) &= \sum_{t=\tau+1}^{\infty} \gamma^{\tau-1}(1-\gamma) && (94)\\
&= \gamma^{\tau} \sum_{t=0}^{\infty} \gamma^{\tau}(1-\gamma) && (95)\\
&= \gamma^{\tau} \stackrel{def}{=} \delta'. && (96)
\end{aligned}$$

This implies

$$\tau = \frac{\ln \delta'}{\ln \gamma} = \frac{\ln 1/\delta'}{\ln 1/\gamma} \leq \frac{\ln 1/\delta'}{1-\gamma} \tag{97}$$

In the complement of the above event:

$$\left(\sup_{(s,a)} \frac{\pi(a \mid s)}{\underline{\pi}(a \mid s)}\right)^{t-1} \leq e^{2(\tau-1)c}. \tag{98}$$

We require that the exponential above be $\leq 1+\epsilon'$, leading to the condition:

$$2(\tau-1)c \leq \ln\left(1+\epsilon'\right) \Rightarrow c \leq \frac{\ln\left(1+\epsilon'\right)}{2(\tau-1)}. \tag{99}$$

Under the assumption of Eq. (90), the above condition holds because

$$\sup_{(s,a)} |c(s,a)| = c \leq \frac{(1-\gamma)\ln\left(1+\epsilon'\right)}{2\ln(1/\delta')} \leq \frac{\ln\left(1+\epsilon'\right)}{2\tau} \leq \frac{\ln\left(1+\epsilon'\right)}{2(\tau-1)}, \tag{100}$$

Therefore, we can ensure

$$\mathbf{P}\left(\left\{\forall \pi \text{ satisfying } Eq.\ (90), \left(\sup_{(s,a)} \frac{\pi(s,a)}{\underline{\pi}(s,a)}\right)^{t-1} \leq \left(1+\epsilon'\right)\right\} \bigcap \left\{t \leq \frac{\ln 1/\delta'}{1-\gamma}\right\}\right) \geq 1-\delta'. \tag{101}$$

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

$$\Pi_{\tau=2}^{t}\frac{\pi(s_\tau,a_\tau)}{\underline{\pi}(s_\tau,a_\tau)} \leq \left(\sup_{(s,a)}\frac{\pi(a\mid s)}{\underline{\pi}(a\mid s)}\right)^{t-1} \leq 1+\epsilon' \tag{102}$$

the thesis follows. ■

G.2. Small Perturbations to Policies

In this section we examine the effect on the loss of small perturbations to the algorithm policies. This is useful when dealing with a discretization argument in Appendix G.3 *(Regression Guarantees with Importance Sampling)*. We highlight that the ϵ'' in this section concerns the discretization error in the union bound in Appendix G.3 *(Regression Guarantees with Importance Sampling)*, and is not to be confused with the value that ϵ' takes in Appendix G.1 *(Importance Sampling Estimator)* (in particular, ϵ' is implicitly defined in Eq. (129)).

Lemma 25 (Difference and Ratio of Nearby Policies) *Fix $\underline{\pi}$ and assume that π, π', ϵ'' satisfy $\forall(s,a)$ the following conditions for some function $b(s,a)$:*

$$\begin{aligned}
&\|\phi(s,a)\|_2 \leq 1, \quad \|w-w'\|_2 \overset{def}{=} \epsilon'' \leq 1\\
&\pi'(a\mid s) \overset{def}{=} \underline{\pi}(a\mid s)\times\frac{e^{c'(s,a)}}{\sum_{a'}\underline{\pi}(a'\mid s)e^{c'(s,a')}}, \qquad \textit{where} \quad c'(s,a)\overset{def}{=} b(s,a)+\phi(s,a)^\top w'\\
&\pi(a\mid s) \overset{def}{=} \underline{\pi}(a\mid s)\times\frac{e^{c(s,a)}}{\sum_{a'}\underline{\pi}(a'\mid s)e^{c(s,a')}}, \qquad \textit{where} \quad c(s,a)\overset{def}{=} b(s,a)+\phi(s,a)^\top w.
\end{aligned} \tag{103}$$

If $\underline{\pi}(a\mid s)=0$ then $\pi(a\mid s)=\pi'(a\mid s)=0$. Otherwise we have the following inequalities:

$$\frac{\pi'(a\mid s)}{\pi(a\mid s)} \leq 1+4\epsilon'', \qquad \frac{\pi(a\mid s)}{\pi'(a\mid s)} \leq 1+4\epsilon'', \qquad \sum_a |\pi'(a\mid s)-\pi(a\mid s)| \leq 8\epsilon''. \tag{104}$$

Proof Dividing the two expressions gives

$$\frac{\pi'(a\mid s)}{\pi(a\mid s)} = \frac{e^{c'(s,a)}}{e^{c(s,a)}}\times\frac{\sum_{a'}\underline{\pi}(a'\mid s)e^{c(s,a')}}{\sum_{a'}\underline{\pi}(a'\mid s)e^{c'(s,a')}} \tag{105}$$

$$= e^{\phi(s,a)^\top(w'-w)}\times\sum_{a''}\underline{\pi}(a''\mid s)\frac{e^{b(s,a'')+\phi(s,a'')^\top w' e^{\phi(s,a'')^\top(w-w')}}}{\sum_{a'}\underline{\pi}(a'\mid s)e^{c'(s,a')}} \tag{106}$$

$$= e^{\phi(s,a)^\top(w'-w)}\times\sum_{a''}\pi'(a''\mid s)e^{\phi(s,a'')^\top(w-w')} \tag{107}$$

$$\leq e^{2\|w-w'\|_2} \leq 1+4\|w-w'\|_2. \tag{108}$$

The last step follows if $\|w-w'\|_2 \leq 1$ by the inequality $e^x \leq 1+2x$ if $x\in[0,1]$. By symmetry, we obtain the other inequality concerning the ratio of the policies in the statement of the

lemma. Using the expression derived above we can write (the second expression below follows by symmetry)

$$\pi'(a \mid s) - \pi(a \mid s) \leq 4\|w - w'\|_2 \pi(a \mid s) \tag{109}$$
$$\pi(a \mid s) - \pi'(a \mid s) \leq 4\|w - w'\|_2 \pi'(a \mid s) \tag{110}$$

Taking absolute values and summing over the actions leads to the third expression in the lemma's statement. ■

Lemma 26 (Stability of the Q-values) *Let π', π, ϵ'' as in Eq. (103). It holds that*

$$\forall (s,a), \qquad \left| \left(Q_k^{\pi'} - Q_k^{\pi}\right)(s,a) \right| \leq \frac{8\epsilon''}{1-\gamma} \times \sup_{(s'',a''),\pi'' \in \{\pi,\pi'\}} |Q_k^{\pi''}(s'',a'')|. \tag{111}$$

Proof Using the performance difference lemma and Lemma 25 *(Difference and Ratio of Nearby Policies)* we can write

$$\left(Q_k^{\pi'} - Q_k^{\pi}\right)(s,a) = \sum_{t=2}^{\infty} \gamma^{t-1} \mathbb{E}_{s_t \sim \pi|(s,a)} \left[\sum_{a_t} \pi'(a_t \mid s_t) Q_k^{\pi'}(s_t,a_t) - \sum_{a_t} \pi(a_t \mid s_t) Q_k^{\pi'}(s_t,a_t) \right] \tag{112}$$

$$\leq \frac{1}{1-\gamma} \times \sup_s \sum_a |\pi'(a \mid s) - \pi(a \mid s)| \times \sup_{(s'',a'')} |Q_k^{\pi'}(s'',a'')| \tag{113}$$

$$\leq \frac{8\epsilon''}{1-\gamma} \sup_{(s'',a'')} |Q_k^{\pi'}(s'',a'')|. \tag{114}$$

Symmetry concludes. ■

Lemma 27 (Stability of the Empirical Q-values) *Let π', π, ϵ'' as in Eq. (103). For any trajectory $\{s_1, a_1, r_1, \ldots, s_t, a_t, r_t\}$ of length $t \leq t_{max}$ if $4\epsilon'' t_{max} \leq 1$ it holds that*

$$\left| \left(\Pi_{\tau=2}^{t} \frac{\pi'(s_\tau,a_\tau)}{\underline{\pi}(s_\tau,a_\tau)} \right) \frac{r_t}{1-\gamma} - \left(\Pi_{\tau=2}^{t} \frac{\pi(s_\tau,a_\tau)}{\underline{\pi}(s_\tau,a_\tau)} \right) \frac{r_t}{1-\gamma} \right| \leq 8\epsilon'' t_{max} \times \max_{\pi'' \in \{\pi,\pi'\}} \left(\Pi_{\tau=2}^{t} \frac{\pi''(s_\tau,a_\tau)}{\underline{\pi}(s_\tau,a_\tau)} \right) \frac{r_t}{1-\gamma} \tag{115}$$

Proof Using Lemma 25 *(Difference and Ratio of Nearby Policies)* we can write

$$\left(\Pi_{\tau=2}^{t} \frac{\pi'(s_\tau,a_\tau)}{\underline{\pi}(s_\tau,a_\tau)} - \Pi_{\tau=2}^{t} \frac{\pi(s_\tau,a_\tau)}{\underline{\pi}(s_\tau,a_\tau)} \right) \frac{r_t}{1-\gamma} \tag{116}$$

$$= \Pi_{\tau=2}^{t} \frac{\pi(s_\tau,a_\tau)}{\underline{\pi}(s_\tau,a_\tau)} \left(\Pi_{\tau=2}^{t} \frac{\pi'(s_\tau,a_\tau)}{\pi(s_\tau,a_\tau)} - 1 \right) \frac{r_t}{1-\gamma}. \tag{117}$$

Now, apply Lemma 25 *(Difference and Ratio of Nearby Policies)* and the condition $t \leq t_{max}$ to the middle term to derive

$$\Pi_{\tau=2}^{t} \frac{\pi'(s_\tau,a_\tau)}{\pi(s_\tau,a_\tau)} - 1 \leq (1 + 4\epsilon'')^{t_{max}} - 1. \tag{118}$$

Recall $t_{max} \geq 1$; for $x \in \mathbb{R}$, when $xt_{max} \leq 1$ but $x \geq 0$, we have the following inequalities:

$$1 + x \leq e^x \rightarrow (1+x)^{t_{max}} \leq e^{t_{max}x} \leq 1 + 2t_{max}x. \tag{119}$$

Let $x = 4\epsilon''$; if $4\epsilon'' t_{max} \leq 1$ then we have

$$(1 + 4\epsilon'')^{t_{max}} - 1 \leq 8\epsilon'' t_{max}. \tag{120}$$

Symmetry concludes. ■

G.3. Regression Guarantees with Importance Sampling

In this section we examine the rate of convergence of the linear regression that uses the importance sampling estimator in the way it is implemented in the algorithm. The bonus b^n and the 'reference' expected and empirical covariance matrices $\Sigma^n, \widehat{\Sigma}^n$ are fixed throughout this section (as they are fixed in all inner iterations of the algorithm). As the outer iteration index n is constant, we often omit it for brevity.

Remark: for additional notation please see Table 1.

Remark: in Eq. (121), if the trajectory is of length 1 then the bonus is not added to accommodate the linear MDP framework, and instead it is (half) added directly to the predictor $\widehat{Q}$, resulting in the pessimistic biased estimate described in the main text.

Lemma 28 (Statistical Rate for a Fixed Target Policy Regression with Importance Sampling) *Fix a behavioral policy $\underline{\pi}$ and a target policy π satisfying Eq. (90) with $\epsilon' = 1$. Fix a bonus function $0 \leq b(\cdot,\cdot) \leq B$. Consider drawing n samples, as follows. For every sample $i \in [n]$ first draw a timestep $t_i \geq 1$ with probability $\mathbf{P}(t_i = \tau) = \gamma^{\tau-1}(1-\gamma)$ and a starting state-action $(s_{i1}, a_{i1}) \sim \rho$ for some distribution ρ. Second, draw a trajectory $\{s_{i1}, a_{i1}, r_{i1}, \ldots, s_{it_i}, a_{it_i}, r_{it_i}\}$ from (s_{i1}, a_{i1}) by following $\underline{\pi}$ for $t_i - 1$ timesteps. Define the random return G_i as*

$$G_i = \begin{cases} \frac{1}{1-\gamma}\left[r_{it_i} + b(s_{it_i}, a_{it_i})\right] & \text{if } t_i \geq 2 \\ \frac{1}{1-\gamma}\left[r_{it_i}\right] & \text{if } t_i = 1 \end{cases} \tag{121}$$

Let G_{max} be a deterministic upper bound to any realization of G_i above (its value is defined in Table 1). Define the empirical loss and the empirical minimizer

$$\widehat{\mathcal{L}}(w,\pi) = \frac{1}{2n}\sum_{i=1}^{n}\left(\phi(s_{i1},a_{i1})^\top w - \Pi_{\tau=2}^{t_i}\frac{\pi(s_{i\tau},a_{i\tau})}{\underline{\pi}(s_{i\tau},a_{i\tau})}G_i\right)^2, \qquad \widehat{w} = \underset{\|w\|_2 \leq W}{\arg\min}\,\widehat{\mathcal{L}}(w,\pi). \tag{122}$$

Define the true minimizer of the loss in Eq. (4)

$$w^\star \overset{def}{=} \underset{\|w\|_2 \leq W}{\arg\min}\,\mathcal{L}(w, \rho, Q^{b,\pi} - b) \tag{123}$$

where $Q^{b,\pi}$ is the state-action value function of π on $\mathcal{M}(\mathcal{S},\mathcal{A},p,r+b,\gamma)$. Let $\Sigma = n\,\mathbb{E}_{(s,a)\sim\rho}\,\phi(s,a)\phi(s,a)^\top + \lambda I$. With probability at least $1-(n+1)\delta'$

$$\|w^\star - \widehat{w}\|_\Sigma^2 \leq 2(C_1 + C_2 \ln\frac{1}{\delta'}) + 2\lambda W^2 = \widetilde{O}(dW^2). \tag{124}$$

Proof The hypotheses ensure through Lemma 24 *(Bias and Variance of Importance Sampling Estimator)* that the importance sampling estimator in Eq. (121) is unbiased estimate of $Q^{b,\pi}(s_{i_1}, a_{i_1}) - b(s_{i_1}, a_{i_1})$ and bounded by $2G_{max}$ in absolute value with probability at least $1 - n\delta'$ for all n samples. Combining this with Lemma 43 *(Statistical Rates for Linear Regression; Theorem 1 in (Mehta, 2017))* we obtain that for a fixed target policy with probability at least $1 - (n+1)\delta'$ we must have

$$\mathcal{L}(\widehat{w}, \pi) - \min_{\|w\|_2 \leq W} \mathcal{L}(w, \rho, Q^{b,\pi} - b) \leq \frac{C_1 + C_2 \ln \frac{1}{\delta'}}{n}, \tag{125}$$

where

$$C_1 = \widetilde{O}\left((W^2 + G_{max}^2)d\right) \tag{126}$$

$$C_2 = \widetilde{O}\left((W^2 + G_{max}^2)\right). \tag{127}$$

Finally, using Lemma 41 *(Σ-norm to Excess Risk)* we conclude. ■

Lemma 29 (Union Bound) *Assume $0 \leq k - \underline{k} \leq \kappa$ where κ is defined in Eq. (129). Let $w_k^\star$ be as in Definition 3 (Transfer Error), and let $\widehat{w}_k$ be the parameter computed during regression by Algorithm 2 in line 8. For some universal constant c we have*

$$\mathbf{P}\left(\|\widehat{w}_k - w_k^\star\|_{(\Sigma^n)^{-1}} \leq c\sqrt{C_1 + C_2 \ln \frac{1}{\delta'} + \lambda W^2}\right) \geq 1 - \left[\mathrm{poly}(N, K, \frac{1}{1-\gamma}, G_{max}, \ln(1/\delta), W)\right]^d \delta' - \frac{\delta}{4} \tag{128}$$

Proof Assume

$$k - \underline{k} \leq \kappa \stackrel{def}{=} \frac{(1-\gamma)\ln 2}{2\ln(8N^2K/\delta)\eta\,(B+W)}. \tag{129}$$

From we know that the all trajectories are bounded by t_{max} with probability $> 1 - \delta/8$. Define the unit ball $\mathcal{B} \stackrel{def}{=} \{v \in \mathbb{R}^d \mid \|v\|_2 \leq 1\}$. As the target policy π_k is a priori unknown we do a union bound over the possible vectors $v \stackrel{def}{=} \sum_{i=\underline{k}}^{k-1} \widehat{w}_i$ which is a priori unknown and data-dependent. Since $v \in (\kappa W)\mathcal{B}$, consider the discretization $\mathcal{D} \subset (\kappa W)\mathcal{B}$ given in Lemma 44 *(Discretization of Euclidean Ball)* with ϵ'' to be determined in this proof. The lemma ensures that for any $v \in (\kappa W)\mathcal{B}$, $\exists v' \in \mathcal{D}, \|v - v'\|_2 \leq \epsilon''$ and that $|\mathcal{D}| = (1 + \frac{2\kappa W}{\epsilon''})^d$.

Now fix v and let π be the policy induced by v and π' the policy induced by v'. Consider the empirical loss in Eq. (122) and let $\widehat{w}_v$ be the empirical minimizer corresponding to π and $\widehat{w}_{v'}$ that corresponding to π'. In addition let $w_v^\star$ be the minimizer corresponding to π and $w_{v'}^\star$ the minimizer corresponding to π' of the true loss in Eq. (123). We can write (Σ is the expected covariance matrix which is fixed throughout the inner iterations)

$$\|\widehat{w}_v - w_v^\star\|_\Sigma \leq \|\widehat{w}_v - \widehat{w}_{v'}\|_\Sigma + \|\widehat{w}_{v'} - w_{v'}^\star\|_\Sigma + \|w_{v'}^\star - w_v^\star\|_\Sigma. \tag{130}$$

We bound each term above.

In the event defined at the beginning of this proof that all trajectories are bounded in length by t_{max} the importance sampling estimator in Eq. (89) is bounded in absolute value by $2G_{max} =$

$\frac{2}{1-\gamma}(3B)$ for all policies π satisfying Eq. (90) ($3B$ is the maximum absolute value of the reward including the bonus) and for all n samples; In particular, the random timestep t of any trajectory is bounded by t_{max}. Then Lemma 27 *(Stability of the Empirical Q-values)* ensures that the importance sampling estimator for π and π' only differ by $8\epsilon'' t_{max} G_{max}$. Plugging this into Lemma 42 *(Stability of the Loss Minimizer)* ensures

$$\|\widehat{w}_{v'} - \widehat{w}_v\|_\Sigma^2 \leq 2n(8\epsilon'' t_{max} G_{max})W + 2\lambda W^2. \tag{131}$$

Likewise, Lemma 26 *(Stability of the Q-values)* ensures that the true Q values for π and π' on the optimistic MDP differ by at most $\frac{8\epsilon''}{1-\gamma}(2G_{max})$. Then Lemma 42 *(Stability of the Loss Minimizer)* ensures

$$\|\widehat{w}^\star_{v'} - \widehat{w}^\star_v\|_\Sigma^2 \leq 2n\left(\frac{8\epsilon''}{1-\gamma}(2G_{max})\right)W + 2\lambda W^2. \tag{132}$$

Setting $\frac{1}{\epsilon''} = \mathrm{poly}(N, K, \frac{1}{1-\gamma}, G_{max}, \ln(1/\delta))$ ensures that the rhs of Eqs. (131) and (132) is, say, $\leq 4\lambda W^2$ (we will have $\lambda > 1$ and $W > 1$) and also satisfies the requirement $4\epsilon'' t_{max} \leq 1$ of Lemma 27 *(Stability of the Empirical Q-values)* (t_{max} was defined at the beginning of the proof).

The ϵ'' just computed determines the size of the discretization set $|\mathcal{D}|$ which is $\left[\mathrm{poly}(n, \frac{1}{1-\gamma}, G_{max})\right]^d$. A union bound over all $v' \in \mathcal{D}$ coupled with Lemma 28 *(Statistical Rate for a Fixed Target Policy Regression with Importance Sampling)* ensures that with probability at least

$$1 - \left[\mathrm{poly}(N, K, \frac{1}{1-\gamma}, G_{max}, \ln(1/\delta), W)\right]^d \frac{\delta'}{2} - \frac{\delta}{8}$$

for an appropriate universal constant c

$$\forall v' \in \mathcal{D}: \qquad \|\widehat{w}_{v'} - \widehat{w}^\star_{v'}\|_\Sigma^2 \leq c\left(C_1 + C_2 \ln\frac{1}{\delta'} + \lambda W^2\right). \tag{133}$$

Plugging back to Eq. (130) concludes. ■

Lemma 30 (Validity of Confidence Intervals) *With probability at least $1 - \frac{\delta}{2}$ for all inner and outer iterations $n \in [N], k = 0, \ldots, K-1$ of the algorithm it holds that*

$$\forall s \in \mathcal{K}^n, \forall a: \qquad |Q_k^{n,\star}(s,a) - \widehat{Q}_k^n(s,a) - b_\phi^n(s,a)| \leq b_\phi^n(s,a) \stackrel{def}{=} \sqrt{\beta}\|\phi(s,a)\|_{(\widehat{\Sigma}^n)^{-1}} \tag{134}$$

where β is defined in Eq. (136).

Proof Define an appropriate $\delta = \left[\mathrm{poly}(N, K, \frac{1}{1-\gamma}, G_{max}, \ln(1/\delta), W)\right]^d \times \delta'$ and invoke Lemma 29. A union bound over all inner and outer iterations ensures

$$\mathbf{P}\left(\forall n \in [N], \forall k = 0, 1, \ldots, K-1 : \|\widehat{w}_k^n - w_k^{n,\star}\|_{(\Sigma^n)^{-1}} \leq \frac{1}{3}\sqrt{\beta}\right) \geq 1 - \frac{\delta}{4} \tag{135}$$

where

$$\beta = \widetilde{O}\left(C_1 + C_2 d \times \ln[\text{poly}(N, K, \frac{1}{1-\gamma}, G_{max}, \ln(1/\delta), W)] + \lambda W^2\right) \tag{136}$$

$$= \widetilde{O}(dW^2 + dG_{max}^2) \tag{137}$$

Combining the above result with Lemma 39 *(Concentration of Inverse Covariances)* gives with probability $1 - \delta/2$:

$$|\phi(s,a)^\top (w_k^{n,\star} - \widehat{w}_k^n)| \leq \|\phi(s,a)\|_{(\Sigma^n)^{-1}} \|\widehat{w}_k^n - w_k^{n,\star}\|_{\Sigma^n} \tag{138}$$

$$\leq 3\|\phi(s,a)\|_{(\widehat{\Sigma}^n)^{-1}} \|\widehat{w}_k^n - w_k^{n,\star}\|_{\Sigma^n} \tag{139}$$

$$= \sqrt{\beta}\|\phi(s,a)\|_{(\widehat{\Sigma}^n)^{-1}}. \tag{140}$$

In other words, thanks to Lemma 39 *(Concentration of Inverse Covariances)* we can use the empirical covariance in place of the full covariance. This implies that we can write the confidence intervals fully as a function of known quantities, in particular, using the empirical covariance matrix $\widehat{\Sigma}^n$ that Algorithm 1 maintains.

Using the definitions for the Q values (still under the same event in known states):

$$|\widehat{Q}_k^n(s,a) + b_\phi^n(s,a) - Q_k^{n,\star}(s,a)| = |\phi(s,a)^\top \widehat{w}_k^n + b_\phi^n(s,a) + b_\phi^n(s,a) - \phi(s,a)^\top w_k^{n,\star} - 2b_\phi^n(s,a)| \tag{141}$$

$$\leq |\phi(s,a)^\top \widehat{w}_k^n - \phi(s,a)^\top w_k^{n,\star}| \tag{142}$$

$$\leq \sqrt{\beta}\|\phi(s,a)\|_{(\widehat{\Sigma}^n)^{-1}} \stackrel{def}{=} b_\phi^n(s,a). \tag{143}$$

■

Lemma 31 (Trajectory Boundness) *Under the conditions on κ in Lemma 29 (Union Bound), all trajectories sampled by Algorithms 4 and 5 are bounded in length by $t_{max} = \frac{\ln(16N^2K/\delta)}{1-\gamma}$ with probability at least $1 - \delta/8$.*

Proof Then Lemma 40 *(Policy Form on Known Set)* ensures that the policies $\pi_k, \pi_{\underline{k}}$ take the form described in Eq. (90) with $\epsilon' = 1, \delta' = \delta/(8N^2K)$ and in particular, Lemma 24 *(Bias and Variance of Importance Sampling Estimator)* ensures that the trajectory lengths are all bounded by $t_{max} = \frac{\ln(16N^2K/\delta)}{1-\gamma}$ with probability at least $1 - \delta/8$ after a union bound over N trajectories collected possibly collected at each of the K solver's iterations, times at most N calls to the SOLVER, and a final union bound over the trajectories samples by Algorithms 4 and 5. ■

Appendix H. Concentration of Bonuses

The proof proceeds with the empirical covariance matrices since the determinant conditions is checked on the empirical matrices.

Notation: In this section for notational convenience the subscripts refer to the outer episode n; for example, we denote the covariance matrix with Σ_n instead of Σ^n

Lemma 32 (Bounding the Sum of Indicators) *For any outer episode n during the execution of the algorithm, let $\sigma(n)$ be the last episode smaller than n where the bonus was updated. If π^n takes an action where $b_{\mathbb{1}}$ is nonzero in a state $s \notin \mathcal{K}^n$ and $\lambda \geq 1$ then under the event of Lemma 33 (Bounding the Sum of Bonuses) we have*

$$\sum_{n=1}^{N} \mathbb{E}_{(s,a)\sim\pi^n|s_0} b_{\mathbb{1}}^{\sigma(n)}(s,a) \overset{def}{=} \frac{3}{1-\gamma} \sum_{n=1}^{N} \mathbb{E}_{s\sim\pi^n|s_0} \mathbb{1}\{s \notin \mathcal{K}^n\} \leq \widetilde{O}\left(\frac{\sqrt{\beta N d}}{1-\gamma}\right). \tag{144}$$

Proof

$$\sum_{n=1}^{N} \mathbb{E}_{s\sim\pi^n|s_0} \mathbb{1}\{s \notin \mathcal{K}^n\} = \sum_{n=1}^{N} \mathbb{E}_{(s,a)\sim\pi^n|s_0} \mathbb{1}\{\sqrt{\beta}\|\phi(s,a)\|_{\widehat{\Sigma}_{\sigma(n)}^{-1}} \geq 1\} \tag{145}$$

$$\leq \sum_{n=1}^{N} \mathbb{E}_{(s,a)\sim\pi^n|s_0} \sqrt{\beta}\|\phi(s,a)\|_{\widehat{\Sigma}_{\sigma(n)}^{-1}}. \tag{146}$$

Finally Lemma 33 *(Bounding the Sum of Bonuses)* concludes. ■

Lemma 33 (Bounding the Sum of Bonuses) *For any outer episode n during the execution of the algorithm, let $\sigma(n)$ be the last episode smaller than n where the bonus was updated. If $\lambda \geq 1$ then with probability at least $1-\delta'$*

$$\sum_{n=1}^{N} \mathbb{E}_{(s,a)\sim\pi^n|s_0} b_{\phi}^{\sigma(n)}(s,a) = \sqrt{\beta} \sum_{n=1}^{N} \mathbb{E}_{(s,a)\sim\pi^n|s_0} \|\phi(s,a)\|_{\widehat{\Sigma}_{\sigma(n)}^{-1}} \tag{147}$$

$$\leq \sqrt{\beta} O\left(\sqrt{ND} + \ln(1/\delta')\right) = \widetilde{O}(\sqrt{\beta N d}), \tag{148}$$

where D is defined in Eq. (176).

Proof

We can write

$$\sum_{n=1}^{N} \mathbb{E}_{(s,a)\sim\pi^n|s_0} b_{\phi}^{\sigma(n)}(s,a) = \sqrt{\beta} \sum_{n=1}^{N} \mathbb{E}_{(s,a)\sim\pi^n|s_0} \|\phi(s,a)\|_{\widehat{\Sigma}_{\sigma(n)}^{-1}}. \tag{149}$$

We need to bound the summation for any realization of the sequence of $(\pi^n, \widehat{\Sigma}_{\sigma(n)})$. Define the random dataset $\mathcal{D}_{1:n}$ containing all the information (i.e., the realization of the random variables) at the *beginning* of iteration n of the algorithm. Conditioning on $\mathcal{D}_{1:n}$ fixes the policy π^n and the covariance $\widehat{\Sigma}_{\sigma(n)}$ and the distribution over ϕ.

Notice that in each episode n the random variable ϕ and the collected feature ϕ_n are identically distributed when conditioned on $\mathcal{D}_{1:n}$ (since their distribution is uniquely determined by the policy π^n in that episode, which is fixed under the conditioning on $\mathcal{D}_{1:n}$). Therefore we can define the 'noise' in the sampled feature

$$\xi_n = \mathbb{E}_{(s,a)\sim\pi^n|s_0}\left[\|\phi(s,a)\|_{\widehat{\Sigma}^{-1}_{\sigma(n)}} \mid \mathcal{D}_{1:n}\right] - \|\phi_n\|_{\widehat{\Sigma}^{-1}_{\sigma(n)}} \tag{150}$$

and write

$$A \overset{def}{=} \sum_{n=1}^{N} \mathbb{E}_{(s,a)\sim\pi^n|s_0}\left[\|\phi(s,a)\|_{\widehat{\Sigma}^{-1}_{\sigma(n)}} \mid \mathcal{D}_{1:n}\right] = \sum_{n=1}^{N} \|\phi_n\|_{\widehat{\Sigma}^{-1}_{\sigma(n)}} + \sum_{n=1}^{N} \xi_n \tag{151}$$

$$\leq \sqrt{N \sum_{n=1}^{N} \|\phi_n\|^2_{\widehat{\Sigma}^{-1}_{\sigma(n)}}} + \sum_{n=1}^{N} \xi_n \tag{152}$$

The first summation on the rhs is bounded by Lemma 36 *(Potential Argument)* by D; it remains to bound the sum of the noise terms. Conditioned on $\mathcal{D}_{1:n}$, the noise ξ_n is mean-zero. Summing over the conditional second moments gives:

$$\sum_{n=1}^{N} \mathbb{E}_{(s,a)\sim\pi^n|s_0}\left[\xi_n^2 \mid \mathcal{D}_{1:n}\right] = \sum_{n=1}^{N} \mathbb{E}_{(s,a)\sim\pi^n|s_0}\left[\|\phi(s,a)\|^2_{\widehat{\Sigma}^{-1}_{\sigma(n)}} \mid \mathcal{D}_{1:n}\right] \tag{153}$$

$$\leq \sum_{n=1}^{N} \mathbb{E}_{(s,a)\sim\pi^n|s_0}\left[\|\phi(s,a)\|_{\widehat{\Sigma}^{-1}_{\sigma(n)}} \mid \mathcal{D}_{1:n}\right] = A. \tag{154}$$

The last step follows because if $\lambda \geq 1$ and $\|\phi(\cdot,\cdot)\|_2 \leq 1$ we have $\|\phi(\cdot,\cdot)\|_{\widehat{\Sigma}^{-1}_{\sigma(n)}} \leq \|\phi(\cdot,\cdot)\|_2 \leq 1$ giving $\|\phi(\cdot,\cdot)\|^2_{\widehat{\Sigma}^{-1}_{\sigma(n)}} \leq \|\phi(\cdot,\cdot)\|_{\widehat{\Sigma}^{-1}_{\sigma(n)}}$. Now, Lemma 45 *(Bernstein for Martinglaes)* gives with probability at least $1-\delta'$ for some constant c

$$\sum_{n=1}^{N} \xi_n \leq c \times \left(\sqrt{2\sum_{n=1}^{N} \mathbb{E}_{(s,a)\sim\pi^n|s_0}[\xi_n^2 \mid \mathcal{D}_{1:n}] \ln(1/\delta')} + \frac{\ln(1/\delta')}{3}\right) \tag{155}$$

$$= c \times \left(\sqrt{2A\ln(1/\delta')} + \frac{\ln(1/\delta')}{3}\right). \tag{156}$$

Combining with Eq. (151) we have shown that with probability at least $1-\delta'$ we must have the following relation

$$A \leq \sqrt{ND} + c \times \left(\sqrt{2A\ln(1/\delta')} + \frac{\ln(1/\delta')}{3}\right). \tag{157}$$

Solving for A finally gives with high probability

$$A = O\left(\sqrt{ND} + \ln(1/\delta')\right). \tag{158}$$

■

The following lemma is used to claim that whenever the determinant condition is violated (triggering a new call to the SOLVER) then the condition is not violated by much.

Lemma 34 (Maximum Determinant Ratio) *If* $\lambda \geq 1$ *and* $\|\phi(\cdot,\cdot)\|_2 \leq 1$ *then* $\det\left(\widehat{\Sigma}_n\right) \leq 4\det\left(\widehat{\Sigma}_{\underline{n}}\right)$.

Proof If $\det\left(\widehat{\Sigma}_n\right) \leq 2\det\left(\widehat{\Sigma}_{\underline{n}}\right)$ the statement holds; if $\det\left(\widehat{\Sigma}_n\right) > 2\det\left(\widehat{\Sigma}_{\underline{n}}\right)$ then by construction we must have $\det\left(\widehat{\Sigma}_{n-1}\right) \leq 2\det\left(\widehat{\Sigma}_{\underline{n}}\right)$ (as the algorithm switches to a new policy once such condition is violated). Use Lemma 37 *(Determinant Ratio)* and recall $\|\phi_{n-1}\|^2_{\widehat{\Sigma}_{n-1}^{-1}} \leq \|\phi_{n-1}\|_2^2 \leq 1$ for $\lambda \geq 1$ to write

$$\det\left(\widehat{\Sigma}_n\right) = \det\left(\widehat{\Sigma}_{n-1}\right)\left(1 + \|\phi_{n-1}\|^2_{\widehat{\Sigma}_{n-1}^{-1}}\right) \leq 2\det\left(\widehat{\Sigma}_{n-1}\right) \leq 4\det\left(\widehat{\Sigma}_{\underline{n}}\right). \tag{159}$$

■

The following lemma is key. It implicitly quantifies the loss due to the delayed update of the covariance matrix; the effect of such delay are rather mild, as they only affect a numerical constant.

Lemma 35 (Trace to LogDeterminant) *Let* Σ *be a positive define matrix and let* M *be a symmetric positive semidefinite matrix. Let* $\Sigma' = \Sigma + M$. *Then if* $\det(\Sigma') \leq 4\det(\Sigma)$ *we have*

$$\ln\det(\Sigma') \geq \ln\det(\Sigma) + \frac{1}{3}\operatorname{Tr}(M\Sigma^{-1}). \tag{160}$$

Proof We have

$$\begin{aligned}
\det(\Sigma') &= \det(\Sigma + M) && (161)\\
&= \det(\Sigma^{\frac{1}{2}})\det(I + \Sigma^{-\frac{1}{2}}M\Sigma^{-\frac{1}{2}})\det(\Sigma^{\frac{1}{2}}) && (162)\\
&= \det(\Sigma)\det(I + \Sigma^{-\frac{1}{2}}M\Sigma^{-\frac{1}{2}}). && (163)
\end{aligned}$$

Denote with $\lambda_1, \ldots, \lambda_d$ the eigenvalues of $\Sigma^{-\frac{1}{2}}M\Sigma^{-\frac{1}{2}}$. We must have by hypothesis

$$4 \geq \frac{\det(\Sigma')}{\det(\Sigma)} = \det(I + \Sigma^{-\frac{1}{2}}M\Sigma^{-\frac{1}{2}}) = \prod_{j=1}^{d}(1+\lambda_j). \tag{164}$$

Taking $\ln$ gives:

$$\ln 4 \geq \sum_{j=1}^{d}\ln(1+\lambda_j). \tag{165}$$

Since all λ_j's must be positive, the $\ln$ terms in the rhs above are positive, and each must satisfy

$$\ln 4 \geq \ln(1+\lambda_j), \qquad \forall j \in [d] \tag{166}$$

and so in particular (by exponentiating the above display)

$$3 \geq \lambda_j, \qquad \forall j \in [d] \tag{167}$$

which allows us to use the following inequality

$$\ln(1+\lambda_j) \geq \frac{1}{3}\lambda_j. \tag{168}$$

Going back to Eq. (164) (and again taking ln) gives

$$\ln\det(\Sigma') = \ln\det(\Sigma) + \sum_{j=1}^{d}\ln(1+\lambda_j) \tag{169}$$

$$\geq \ln\det(\Sigma) + \frac{1}{3}\sum_{j=1}^{d}\lambda_j \tag{170}$$

$$\geq \ln\det(\Sigma) + \frac{1}{3}\operatorname{Tr}(\Sigma^{-\frac{1}{2}}M\Sigma^{-\frac{1}{2}}) \tag{171}$$

$$= \ln\det(\Sigma) + \frac{1}{3}\operatorname{Tr}(M\Sigma^{-1}). \tag{172}$$

■

Lemma 36 (Potential Argument) *Let $\underline{n}_1, \underline{n}_2, \ldots, \underline{n}_{last}$ be the indexes in the sequence $n = 1, \ldots, N$ where the bonus gets updated and let $\sigma(n)$ be the last episode smaller than n where the bonus was updated, i.e.,*

$$\widehat{\Sigma}_{\underline{n}_{i+1}} = \widehat{\Sigma}_{\underline{n}_i} + \sum_{n=\underline{n}_i}^{\underline{n}_{i+1}-1}\phi_n\phi_n^\top \tag{173}$$

$$\widehat{\Sigma}_n = \widehat{\Sigma}_{\sigma(n)}, \quad \underline{n}_i \leq n < \underline{n}_{i+1} \tag{174}$$

$$\det(\widehat{\Sigma}_{\underline{n}_{i+1}}) \leq 4\det(\widehat{\Sigma}_{\underline{n}_i}). \tag{175}$$

We have

$$\sum_{n=1}^{N}\|\phi_n\|^2_{\widehat{\Sigma}^{-1}_{\sigma(n)}} \leq 3\left(\ln\det(\widehat{\Sigma}_{N+1}) - \ln\det(\widehat{\Sigma}_1)\right) \stackrel{def}{=} D = \widetilde{O}(d). \tag{176}$$

Proof Let $\underline{n}_{last}$ be the index of the last switch. Use Lemma 35 *(Trace to LogDeterminant)* twice with the following inputs (notice that the determinant ratio condition is satisfied in both cases)

$$\Sigma' = \widehat{\Sigma}_{\underline{n}_{i+1}}, \quad \Sigma = \widehat{\Sigma}_{\underline{n}_i}, \quad M = \sum_{n=\underline{n}_i}^{\underline{n}_{i+1}-1}\phi_n\phi_n^\top \tag{177}$$

$$\Sigma' = \widehat{\Sigma}_{N+1}, \quad \Sigma = \widehat{\Sigma}_{\underline{n}_{last}}, \quad M = \sum_{n=\underline{n}_{last}}^{N}\phi_n\phi_n^\top \tag{178}$$

to obtain

$$\ln\det(\widehat{\Sigma}_{\underline{n}_{i+1}}) - \ln\det(\widehat{\Sigma}_{\underline{n}_i}) \geq \frac{1}{3}\operatorname{Tr}(\sum_{n=\underline{n}_i}^{\underline{n}_{i+1}-1} \phi_n\phi_n^\top \widehat{\Sigma}_{\underline{n}_i}^{-1}) \tag{179}$$

$$= \frac{1}{3}\sum_{n=\underline{n}_i}^{\underline{n}_{i+1}-1} \operatorname{Tr}(\phi_n\phi_n^\top \widehat{\Sigma}_{\sigma(n)}^{-1}) \tag{180}$$

and likewise

$$\ln\det(\widehat{\Sigma}_{N+1}) - \ln\det(\widehat{\Sigma}_{\underline{n}_{last}}) \geq \frac{1}{3}\sum_{n=n_{last}}^{N} \operatorname{Tr}(\phi_n\phi_n^\top \widehat{\Sigma}_{\sigma(n)}^{-1}) \tag{181}$$

Summing over the switches, recalling $\underline{n}_1 = 1$ and adding the above display gives (after cancelling the terms in the telescoping sum)

$$\ln\det(\widehat{\Sigma}_{N+1}) - \ln\det(\widehat{\Sigma}_{\underline{n}_1}) \geq \frac{1}{3}\sum_{n=1}^{N} \operatorname{Tr}(\phi_n\phi_n^\top \widehat{\Sigma}_{\sigma(n)}^{-1}). \tag{182}$$

Finally, consider

$$\|\phi_n\|^2_{\Sigma^{-1}_{\sigma(n)}} = \phi_n^\top \widehat{\Sigma}_{\sigma(n)}^{-1}\phi_n \tag{183}$$

$$= \operatorname{Tr}\left(\phi_n^\top \widehat{\Sigma}_{\sigma(n)}^{-1}\phi_n\right) \tag{184}$$

$$= \operatorname{Tr}\left(\phi_n\phi_n^\top \widehat{\Sigma}_{\sigma(n)}^{-1}\right). \tag{185}$$

Combining with the prior display concludes. ■

Lemma 37 (Determinant Ratio) *If $\Sigma^+ = \Sigma + \phi\phi^\top$ and Σ is strictly symmetric positive definite then* $\det(\Sigma^+) = \det(\Sigma)\left(1 + \|\phi\|^2_{\Sigma^{-1}}\right)$.

Proof The inverse of Σ exists because Σ is strictly positive definite. We can write

$$\det(\Sigma^+) = \det(\Sigma + \phi\phi^\top) \tag{186}$$

$$= \det\left[\Sigma^{\frac{1}{2}}\left(I + \Sigma^{-\frac{1}{2}}\phi\phi^\top\Sigma^{-\frac{1}{2}}\right)\Sigma^{\frac{1}{2}}\right] \tag{187}$$

$$= \det(\Sigma)\det\left(I + \Sigma^{-\frac{1}{2}}\phi\phi^\top\Sigma^{-\frac{1}{2}}\right). \tag{188}$$

We use the matrix determinant lemma to continue and write

$$= \det(\Sigma)\left(1 + \phi^\top\Sigma^{-\frac{1}{2}}\Sigma^{-\frac{1}{2}}\phi\right) \tag{189}$$

$$= \det(\Sigma)\left(1 + \|\phi\|^2_{\Sigma^{-1}}\right). \tag{190}$$

■

Lemma 38 (Number of Switches) *Using the same notation as Lemma 36 we have that the number of times the bonus is updated is* $\widetilde{O}(d)$.

Proof Notice that $\det(\widehat{\Sigma}_1) \geq \det(\widehat{\Sigma}_0) = \lambda_{min}^d$ (for the definition of λ_{min}, please see Eq. (212)) and $\det(\widehat{\Sigma}_N) \leq (\lambda_{min} + \frac{(N+1)}{d})^d$ (see proof of lemma 11 in (Abbasi-Yadkori et al., 2011)). Let $\underline{n}_1, \underline{n}_2, \ldots, \underline{n}_{last}$ be the indexes in the sequence $n = 1, \ldots, N$ where the bonus b^n gets updated. Every time the bonus is updated we have

$$\det(\widehat{\Sigma}_{\underline{n}_{i+1}}) \geq 2 \det(\widehat{\Sigma}_{\underline{n}_i}). \tag{191}$$

Let S denote the number of times the bonus is updated. By induction,

$$\left(\lambda_{min} + \frac{(N+1)}{d}\right)^d \geq \det(\Sigma_{\underline{n}_{last}}) \geq 2^S \det(\Sigma_1) \geq 2^S \lambda_{min}^d \tag{192}$$

It follows that

$$S \leq d \ln_2 \left(1 + \frac{(N+1)}{d\lambda_{min}}\right) = \widetilde{O}(d) \tag{193}$$

■

Appendix I. Inverse Covariance Matrix Estimation

Lemma 39 (Concentration of Inverse Covariances) *Let μ_i be the conditional distribution of ϕ given the sampled $\phi_1, \ldots, \phi_{i-1}$. Assume $\|\phi\|_2 \leq 1$ for any realization of the vector. Define $\Sigma = \frac{1}{n}\sum_{i=1}^n \mathbb{E}_{\phi\sim\mu_i}\phi\phi^\top$. If*

$$\lambda \geq \lambda_{min} = \Omega(d\ln(n/\delta'')). \tag{194}$$

where λ_{min} is defined in Eq. (212) then we have

$$\mathbf{P}\left(\forall n \geq 1, \quad \frac{3}{1}\left(n\Sigma + \lambda I\right)^{-1} \succeq \left(\sum_{i=1}^n \phi_i\phi_i^\top + \lambda I\right)^{-1} \succeq \frac{3}{5}\left(n\Sigma + \lambda I\right)^{-1}\right) \geq 1 - \delta''. \tag{195}$$

In the same event as above the following event must hold as well

$$\forall n \geq 1, \quad \frac{1}{3}\left(n\Sigma + \lambda I\right) \preceq \left(\sum_{i=1}^n \phi_i\phi_i^\top + \lambda I\right) \preceq \frac{5}{3}\left(n\Sigma + \lambda I\right). \tag{196}$$

Proof Consider any x such that $\|x\|_2 = 1$. Let $\Sigma_i = \mathbb{E}_{\phi\sim\mu_i}\phi\phi^\top$ and $\Sigma \stackrel{def}{=} \frac{1}{n}\sum_{i=1}^n \Sigma_i$. We have

$$\mathbb{E}_{\phi\sim\mu_i} x^\top\phi\phi^\top x = \mathbb{E}_{\phi\sim\mu_i}\left(x^\top\phi\right)^2 = x^\top\Sigma_i x. \tag{197}$$

The random variable $\left(x^\top\phi\right)^2$, $\phi \sim \mu_i$ is positive with maximum value $\left(x^\top\phi\right)^2 \leq \|x\|_2^2\|\phi\|_2^2 \leq 1$ and mean $x^\top\Sigma_i x$; therefore the conditional variance is at most $x^\top\Sigma_i x$ as well, as we show below

$$\operatorname*{Var}_{\phi\sim\mu_i}(\phi^\top x)^2 \leq \mathbb{E}_{\phi\sim\mu_i}(\phi^\top x)^2 = x^\top\Sigma_i x. \tag{198}$$

Now Lemma 45 *(Bernstein for Martinglaes)* gives with probability at least $1-\delta'$ for some constant c

$$\left|\frac{1}{n}\sum_{i=1}^n\left[\left(x^\top\phi_i\right)^2 - x^\top\Sigma_i x\right]\right| = \left|\frac{1}{n}\sum_{i=1}^n\left(x^\top\phi_i\right)^2 - x^\top\Sigma x\right| \leq c\left(\sqrt{2\frac{x^\top\Sigma x}{n}\ln(2/\delta')} + \frac{\ln(2/\delta')}{3n}\right). \tag{199}$$

We require

$$c\left(\sqrt{2\frac{x^\top\Sigma x}{n}\ln(2/\delta')} + \frac{\ln(2/\delta')}{3n}\right) \leq \frac{1}{2}\left(x^\top\Sigma x + \frac{\lambda}{n}\right). \tag{200}$$

We will show that if

$$\lambda \geq \Omega(\ln(1/\delta')) \tag{201}$$

then Eq. (200) holds for any fixed value of n.

Case $x^\top \Sigma x \leq \frac{\lambda}{n}$. In this case it is sufficient to satisfy for some constants c', c''

$$\frac{\ln(2/\delta')}{3n} \leq c'\left(\frac{\lambda}{n}\right) \quad \longleftrightarrow \quad \Omega(\ln(1/\delta')) \leq \lambda$$
$$\sqrt{2\frac{\lambda}{n^2}\ln(2/\delta')} \leq c''\left(\frac{\lambda}{n}\right) \quad \longleftrightarrow \quad \Omega(\ln(1/\delta')) \leq \lambda.$$

Case $x^\top \Sigma x > \frac{\lambda}{n}$. In this case to satisfy proceed as in the above display (first equation). For the second equation it is sufficient to satisfy for some constant c'''

$$\sqrt{2\frac{x^\top \Sigma x}{n}\ln(2/\delta')} \leq c'''\left(x^\top \Sigma x\right) \quad \longleftrightarrow \quad \frac{\ln(2/\delta')}{(x^\top \Sigma x)n} \leq O(1).$$

Using the condition $x^\top \Sigma x > \frac{\lambda}{n}$ we can conclude that satisfying $\Omega(\ln(1/\delta')) \leq \lambda$ suffices.

This ultimately implies that for any fixed x such that $\|x\|_2 = 1$ we have

$$\left|x^\top\left(\frac{1}{n}\sum_{i=1}^n \phi_i\phi_i^\top - \Sigma\right)x\right| \leq \frac{1}{2}\left(x^\top\Sigma x + \frac{\lambda}{n}\right) = \frac{1}{2}x^\top\left(\Sigma + \frac{\lambda}{n}I\right)x. \tag{202}$$

with probability at least $1-\delta'$. Define $\partial\mathcal{B} = \{\|x\| = 1\}$; using a standard discretization argument, (e.g., lemma 5.2 in (Vershynin, 2010)) we have that

$$\forall \epsilon > 0,\ \exists \mathcal{B}_\epsilon \subseteq \mathcal{B} \quad \text{such that} \quad \forall x \in \mathcal{B}, \exists x' \in \mathcal{B}_\epsilon \subseteq \mathcal{B} \quad \text{such that} \quad \|x - x'\|_2 \leq \epsilon \tag{203}$$

and

$$|\mathcal{B}_{\epsilon'}| \leq \left(\frac{3}{\epsilon'}\right)^d \stackrel{def}{=} \mathcal{N}. \tag{204}$$

Therefore, applying the result of Eq. (202) to any such $x' \in \mathcal{B}_\epsilon$ gives after a union bound over the x' and the number of samples n that with probability at least $1 - n\mathcal{N}\delta' \stackrel{def}{=} 1 - \delta''$ we have that

$$\forall n,\ \forall x' \in \mathcal{B}_{\epsilon'}, \qquad \left|(x')^\top\left(\frac{1}{n}\sum_{i=1}^n \phi_i\phi_i^\top\right)x' - (x')^\top\Sigma x'\right| \leq \frac{1}{2}(x')^\top\left(\Sigma + \frac{\lambda}{n}I\right)x'. \tag{205}$$

Now for any $x \in \partial\mathcal{B}$ consider the closest $x' \in \mathcal{B}_{\epsilon'}$. We have that for an spd matrix A such that $\|A\|_2 \leq 1$

$$\begin{aligned} x^\top Ax - (x')^\top Ax' &= x^\top Ax - (x')^\top Ax + (x')^\top Ax - (x')^\top Ax' && (206)\\ &= (x-x')^\top Ax + (x')^\top A(x-x') && (207)\\ &\leq 2\epsilon'\|A\|_2 \max\{\|x\|_2, \|x'\|_2\} && (208)\\ &\leq 2\epsilon'. && (209)\end{aligned}$$

Apply this to the case $A = \Sigma$ and $A = \frac{1}{n}\sum_{i=1}^n \phi_i\phi_i^\top$ (notice that $\|\Sigma\|_2 \leq 1$ and $\|\frac{1}{n}\sum_{i=1}^n \phi_i\phi_i^\top\|_2 \leq 1$ follow from hypothesis) to obtain

$$\left|x^\top\left(\frac{1}{n}\sum_{i=1}^n \phi_i\phi_i^\top\right)x - (x')^\top\left(\frac{1}{n}\sum_{i=1}^n \phi_i\phi_i^\top\right)x'\right| \leq 2\epsilon' \tag{210}$$

$$\left|x^\top\Sigma x - (x')^\top\Sigma x'\right| \leq 2\epsilon'. \tag{211}$$

This implies that, if

$$\lambda \geq \Omega\left(\ln\left(\frac{2n\mathcal{N}}{\delta''}\right)\right) \stackrel{def}{=} \lambda_{min} \tag{212}$$

then

$$\forall n,\ \forall x \in \mathcal{B}, \qquad \left| x^\top \left[\left(\frac{1}{n}\sum_{i=1}^{n} \phi_i \phi_i^\top + \frac{\lambda}{n} I \right) - \left(\Sigma + \frac{\lambda}{n} I \right) \right] x \right| \leq \frac{1}{2} x^\top \left(\Sigma + \frac{\lambda}{n} I \right) x + 4\epsilon' \tag{213}$$

$$\leq \frac{2}{3} x^\top \left(\Sigma + \frac{\lambda}{n} I \right) x \tag{214}$$

by setting $\epsilon' = \mathcal{O}\left(\frac{1}{n}\right)$ (as we set $\lambda > 1$ and in addition $x \in \partial\mathcal{B}$). This implies

$$\frac{1}{3}\left(\Sigma + \frac{\lambda}{n} I \right) \preceq \frac{1}{n}\sum_{i=1}^{n} \phi_i \phi_i^\top + \frac{\lambda}{n} \preceq \frac{5}{3}\left(\Sigma + \frac{\lambda}{n} I \right) \tag{215}$$

and finally the thesis. ■

Appendix J. Technical Results

Lemma 40 (Policy Form on Known Set) *Fix n (the outer iteration index) and the bonus b in that outer iteration. Let k be an inner episode of the algorithm and let $\underline{k}$ be the last time data were collected. The policy π_k computed by the algorithm in the inner episode k reads in any known state $s \in \mathcal{K}^n$ for some $w_{\underline{k}}, \dots, w_{k-1}$:*

$$\pi_k(a \mid s) = \pi_{\underline{k}}(a \mid s) \times \frac{e^{c(s,a)}}{\sum_{a'} \pi_{\underline{k}}(a' \mid s) e^{c(s,a')}}, \qquad \textit{where} \quad c(s,a) = \eta \sum_{i=\underline{k}}^{k-1} \left[b(s,a) + \phi(s,a)^\top \widehat{w}_i \right]. \tag{216}$$

Proof

Assume $k > \underline{k}$, otherwise the the statement is trivially true. The update rule reads

$$\pi_k(\cdot \mid s) \propto \pi_{k-1}(\cdot \mid s) e^{\eta \widehat{Q}_{k-1}(\cdot|s)} \tag{217}$$

$$= \pi_{k-1}(\cdot \mid s) e^{\eta \left[\phi(s,\cdot)^\top \widehat{w}_{k-1} + b(s,\cdot) \right]} \tag{218}$$

Using induction gives

$$\pi_k(\cdot \mid s) \propto \pi_{\underline{k}}(\cdot \mid s) \prod_{I=\underline{k}}^{k-1} e^{\eta \left[\phi(s,a)^\top \widehat{w}_i + b(s,\cdot) \right]} \tag{219}$$

$$\propto \pi_{\underline{k}}(\cdot \mid s) \times e^{c(s,\cdot)}. \tag{220}$$

Normalization concludes. ■

Lemma 41 (Σ-norm to Excess Risk) *Fix $\lambda > 0$. Define*

$$\mathcal{L}(w) = \frac{1}{2} \mathbb{E}_{(x,y)} \left(x^\top w - y \right)^2 \tag{221}$$

$$w^\star \in \operatorname*{arg\,min}_{\|w\|_2 \le W} \mathcal{L}(w). \tag{222}$$

Then for any scalar $M > 0$

$$\|w - w^\star\|^2_{(M \mathbb{E}_{(x,y)} x x^\top + \lambda I)} \le 2M \left(\mathcal{L}(w) - \mathcal{L}(w^\star) \right) + \lambda \|w - w^\star\|_2^2. \tag{223}$$

Proof We write $\mathbb{E}$ in place of $\mathbb{E}_{(x,y)}$ for short. The optimality condition reads (for any feasible w)

$$\mathbb{E} \left(x^\top w^\star - y \right) x^\top (w - w^\star) \ge 0. \tag{224}$$

Therefore

$$2\Big[\mathcal{L}(w) - \mathcal{L}(w^\star)\Big] = \mathbb{E}\left(x^\top w - y\right)^2 - \mathbb{E}\left(x^\top w^\star - y\right)^2 \tag{225}$$

$$= \mathbb{E}\left[\left(x^\top w - y\right) - \left(x^\top w^\star - y\right)\right]\left[x^\top w - y + x^\top w^\star - y\right] \tag{226}$$

$$= \mathbb{E}\left[x^\top (w - w^\star)\right]\left[x^\top w - y + x^\top w^\star - y\right] \tag{227}$$

$$= \mathbb{E}\left[x^\top (w - w^\star)\right]\left[x^\top (w - w^\star) + x^\top w^\star - y + x^\top w^\star - y\right] \tag{228}$$

$$= (w - w^\star)^\top \mathbb{E}\left(xx^\top\right)(w - w^\star) + 2\,\mathbb{E}\left[x^\top (w - w^\star)\right]\left[x^\top w^\star - y\right] \tag{229}$$

$$\geq (w - w^\star)^\top \mathbb{E}\left(xx^\top\right)(w - w^\star) \tag{230}$$

$$= (w - w^\star)^\top \left[\mathbb{E}\, xx^\top + \frac{\lambda}{M} I\right](w - w^\star) - \frac{\lambda}{M}\|w - w^\star\|_2^2. \tag{231}$$

The inequality follows from the optimality conditions in the prior display. ■

Lemma 42 (Stability of the Loss Minimizer) *Let*

$$\mathcal{L}(w) = \frac{1}{2}\,\mathbb{E}_{(x,y)}\left(x^\top w - y\right)^2, \quad w_\star \in \operatorname*{arg\,min}_{\|w\|_2 \leq W} \mathcal{L}(w) \tag{232}$$

$$\mathcal{L}'(w) = \frac{1}{2}\,\mathbb{E}_{(x,y)}\left(x^\top w - y - f(y)\right)^2, \quad w'_\star \in \operatorname*{arg\,min}_{\|w\|_2 \leq W} \mathcal{L}'(w). \tag{233}$$

If $\lambda > 0$ and the perturbation $|f(y)| \leq \epsilon_f$ for every y and $\Sigma = \mathbb{E}_{(x,y)}\, xx^\top$ and $M > 0$ then

$$\|w'_\star - w_\star\|^2_{M\Sigma + \lambda I} \leq 2M\epsilon_f W + 2\lambda W^2. \tag{234}$$

Proof

Define $y' = y + f(y)$ for short. We write $\mathbb{E}$ instead of $\mathbb{E}_{(x,y)}$ for brevity. The optimality conditions at the minimizers $w_\star$ and $w'_\star$ for the real losses $\mathcal{L}(w)$ and $\mathcal{L}'(w)$ read

$$\frac{\partial}{\partial w}\mathcal{L}(w_\star)(w - w_\star) = \mathbb{E}(x^\top w_\star - y)x^\top (w - w_\star) \geq 0 \tag{235}$$

$$\frac{\partial}{\partial w}\mathcal{L}'(w'_\star)(w - w'_\star) = \mathbb{E}(x^\top w'_\star - y')x^\top (w - w'_\star) \geq 0. \tag{236}$$

Take the first condition and evaluate it at $w = w'_\star$ to write

$$0 \leq \mathbb{E}(x^\top w_\star - y)x^\top (w'_\star - w_\star) \tag{237}$$

$$= \mathbb{E}\left(x^\top w'_\star + x^\top (w_\star - w'_\star) - y' + (y' - y)\right)x^\top (w'_\star - w_\star) \tag{238}$$

$$= \mathbb{E}\left(x^\top w'_\star - y'\right)x^\top (w'_\star - w_\star) \tag{239}$$

$$+ \mathbb{E}\left(x^\top (w_\star - w'_\star) + (y' - y)\right)x^\top (w'_\star - w_\star) \tag{240}$$

The first term in the above rhs must be negative due to the second optimality condition (for $w = w_\star$) in the previous display; therefore, at the very least the second term in the rhs above must be positive

$$0 \leq \mathbb{E}\left(x^\top(w_\star - w'_\star) + (y' - y)\right) x^\top(w'_\star - w_\star) \tag{241}$$

$$= \mathbb{E}(w'_\star - w_\star)^\top x x^\top (w_\star - w'_\star) + \mathbb{E}(y' - y)x^\top(w'_\star - w_\star) \tag{242}$$

and next

$$M(w'_\star - w_\star)^\top \mathbb{E}\, x x^\top (w'_\star - w_\star) \leq M\, \mathbb{E}(y' - y)x^\top(w'_\star - w_\star) \tag{243}$$

and finally

$$(w'_\star - w_\star)^\top \left(M\, \mathbb{E}\, x x^\top + \lambda I\right)(w'_\star - w_\star) \leq M\, \mathbb{E}(y' - y)x^\top(w'_\star - w_\star) + \lambda \|w_\star - w'_\star\|_2^2 \tag{244}$$

$$\leq 2M\epsilon_f W + 2\lambda W^2. \tag{245}$$

■

Lemma 43 (Statistical Rates for Linear Regression; Theorem 1 in (Mehta, 2017)) *With $z = (\phi, y)$ let $l_w(z) \mapsto \frac{1}{2}(\phi^\top w - y)^2$. Assume $Z \sim P$ and $\|\phi\|_2 \leq 1, \|w\|_2 \leq W, |y| \leq y_{max}$. Let $\widehat{w}$ be the empirical risk minimizer with n i.i.d. samples from P and let $w^\star = \arg\min_{\|w\|_2 \leq W} \mathbb{E}_{Z\sim P}\, l_w(Z)$. With probability at least $1 - \delta'$ we have*

$$\mathbb{E}_{Z\sim P}[l_{\widehat{w}}(Z) - l_{w^\star}(Z)] \leq \frac{1}{n}\left[32(W + y_{max})^2 \times \left[d \ln(16(W + y_{max})(2W)n) + \ln\frac{1}{\delta'}\right] + 1\right] \tag{246}$$

Proof The maximum value the loss can take is $L_{max}^2 = (W + y_{max})^2$. The statement then follows as an application of Theorem 1 in (Mehta, 2017) to linear regression, which is $1/(4L_{max}^2)$-exp-concave (end of section 3 in (Mehta, 2017). ■

Lemma 44 (Discretization of Euclidean Ball) *The Euclidean sphere $R\mathcal{B} = \{x \mid \|x\|_2 \leq R, x \in \mathbb{R}^d\}$ with radius R equipped with the Euclidean metric admits a discretization for every $\epsilon > 0$: $\mathcal{D}_\epsilon = \{y_1, \ldots, y_{\mathcal{N}_\epsilon}\} \subseteq R\mathcal{B}$ with*

$$\mathcal{N}_\epsilon \leq \left(1 + \frac{2R}{\epsilon}\right)^d \tag{247}$$

such that

$$\forall x \in R\mathcal{B}, \exists y \in \mathcal{D}_\epsilon \quad \textit{such that} \quad \|x - y\|_2 \leq \epsilon. \tag{248}$$

Proof By scaling the unit ball to have radius R and using lemma 5.2 in (Vershynin, 2010). ■

Lemma 45 (Bernstein for Martinglaes) *Consider the stochastic process $\{X_n\}$ adapted to the filtration $\{\mathcal{F}_n\}$. Assume $\mathbb{E}\, X_n = 0$ and $cX_n \leq 1$ for every n; then for every constant $z \neq 0$ it holds that*

$$\mathbf{P}\left(\sum_{n=1}^{N} X_n \leq z \sum_{n=1}^{N} \mathbb{E}(X_n^2 \mid \mathcal{F}_n) + \frac{1}{z} \ln \frac{1}{\delta}\right) \geq 1 - \delta. \tag{249}$$

This implies

$$\mathbf{P}\left(\sum_{n=1}^{N} X_n \leq c \times \sqrt{\sum_{n=1}^{N} \mathbb{E}(X_n^2 \mid \mathcal{F}_n) \ln \frac{1}{\delta}} + \ln \frac{1}{\delta}\right) \geq 1 - \delta. \tag{250}$$

Proof The first inequality follows from Theorem 1 in (Beygelzimer et al., 2011); the second follows from optimizing the bound as a function of z depending upon which term in the rhs is larger. ■