2024-07-27

Multiple-policy evaluation via density estimation

Multiple-policy Evaluation via Density Estimation

Yilei Chen
Boston University

Aldo Pacchiano*
Boston University
Broad Institute of MIT and Harvard

Ioannis Ch. Paschalidis*
Boston University

Abstract

In this work, we focus on the multiple-policy evaluation problem where we are given a set of K target policies and the goal is to evaluate their performance (the expected total rewards) to an accuracy ϵ with probability at least $1-\delta$. We propose an algorithm named CAESAR to address this problem. Our approach is based on computing an approximate optimal offline sampling distribution and using the data sampled from it to perform the simultaneous estimation of the policy values. CAESAR consists of two phases. In the first one we produce coarse estimates of the vistation distributions of the target policies at a low order sample complexity rate that scales with $\tilde{O}(\frac{1}{\epsilon})$. In the second phase, we approximate the optimal offline sampling distribution and compute the importance weighting ratios for all target policies by minimizing a step-wise quadratic loss function inspired by the objective in DualDICE Nachum et al. [2019]. Up to low order and logarithm terms CAESAR achieves a sample complexity $\tilde{O}\left(\frac{H^4}{\epsilon^2}\sum_{h=1}^{H}\max_{k\in[K]}\sum_{s,a}\frac{(d_h^{\pi^k}(s,a))^2}{\mu_h^*(s,a)}\right)$, where d^π is the visitation distribution of policy π and μ^* is the optimal sampling distribution.

1 Introduction

Policy evaluation is a fundamental problem in Reinforcement Learning (RL) Sutton and Barto [2018] of which the goal is to estimate the expected total rewards of a given policy. This process serves as an integral component in various RL methodologies, such as policy iteration and policy gradient approaches Sutton et al. [1999], wherein the current policy undergoes evaluation followed by potential updates. Policy evaluation is also paramount in scenarios where prior to deploying a trained policy, thorough evaluation is necessary imperative to ensure its safety and efficacy.

Broadly speaking there exist two scenarios where the problem of policy evaluation has been considered, known as online and offline data regimes. In online scenarios a learner is interacting sequentially with the environment and is tasked with using its online deployments to collect helpful data for policy evaluation. The simplest method for online policy evaluation is Monte-Carlo estimation Fonteneau et al. [2013]. One can collect multiple trajectories by following the target policy, and use the empirical mean of the rewards as the estimator. These on-policy methods typically require executing the policy we want to estimate which may be unpractical or dangerous in many cases. For example, in the medical treatment scenario, implementing an untrustworthy policy can cause unfortunate consequences Thapa et al. [2005]. In these cases, offline policy evaluation may be preferable. In the offline case, the learner has access to a batch of data and is tasked to use this in the best way possible to estimate the value of a target policy. There are many works focus on this field based on different techniques such as importance-sampling, model-based estimation and doubly-robust estimators Yin and Wang [2020], Jiang and Li [2016], Yin et al. [2021], Xie et al. [2019], Li et al. [2015].

Motivated by the applications where people often have multiple policies that they would like to evaluate, e.g. multiple policies trained using different hyperparameters, Dann et al. [2023] considered multiple-policy evaluation which aims to estimate the performance of a set of K target policies instead of a single policy. From the simplest perspective, multiple-policy evaluation does not pose challenges beyond single-policy evaluation since one can always use single-policy evaluation methods by K times to solve the multiple-policy evaluation problem. However, this can be extremely

*Joint senior-authorship

sample-inefficient as it neglects potential similarities among the K target policies. Consequently, its sample complexity invariably escalates linearly as a function of K.

Dann et al. [2023] proposed an on-policy algorithm that leverages the possible similarity among target policies based on an idea related to trajectory synthesis Wang et al. [2020]. The basic technique is that if more than one policy take the same action at a certain state, then only one sample is needed at that state which can be reused to synthesize trajectories for these policies. Their algorithm achieves an instance-dependent sample complexity which gives much better results when target policies have many overlaps.

In the context of single policy off-policy evaluation, the theoretical guarantees depend on the overlap between the offline data distribution and the visitations of the evaluated policy Xie et al. [2019], Yin and Wang [2020], Duan et al. [2020]. These coverage conditions which ensure that the data logging distribution Xie et al. [2022] adequately covers the state space are typically captured by the ratio between the densities corresponding to the offline data distribution and the policy to evaluate, also known as importance ratios.

A single offline dataset can be used to evaluate multiple policies simultaneously. The policy evaluation guarantees will be different for each of the policies in the set depending on how much overlap the offline distribution has with the policy visitation distributions. These observations inform an approach to the multiple policy evaluation problem different from Dann et al. [2023] that can also leverage the policy visitation overlap in a meaningful way. Our algorithm is based on the idea of designing a behavior distribution with enough coverage of the target policy set. Once this distribution is computed, i.i.d. samples from the behavior distribution can be used to estimate the value of the target policies using ideas inspired in the offline policy optimization literature. Our algorithms consist of two phases:

1. Build coarse estimators of the policy visitation distributions and use them to compute a mixture policy that achieves a low visitation ratio with respect to all K policies to evaluate.
2. Sample from this approximately optimal mixture policy and use these to construct mean reward estimators for all K policies.

Coarse estimation of the visitation distributions up to constant multiplicative accuracy can be achieved at a cost that scales linearly, instead of quadratically with the inverse of the accuracy parameter (see Section 4.1) and polynomially in parameters such as the size of the state and action spaces, and the logarithm of the policy evaluation set. Estimating the policy visitation distributions up to multiplicative accuracy is enough to find an approximately optimal behavior distribution that minimizes the maximum visitation ratio among all policies to estimate (see Section 4.2). The samples generated from this behavior distribution are used to estimate the target policy values via importance weighting. Since the importance weights are not known to sufficient accuracy, we propose the IDES or Importance Density Estimation Algorithm (see Algorithm 1) for estimating these distribution ratios by minimizing a series of loss functions inspired by the $\mathrm{DualDICE}$ Nachum et al. [2019] method (see Section 4.3). Combining these steps we arrive at our main algorithm (CAESAR) or Coarse and Adaptive Estimation with Approximate Reweighing for Multi-Policy Evaluation (see Algorithm 2) that achieves a high probability finite sample complexity for the problem of multi-policy evaluation.

2 Related Work

There is a rich family of off-policy estimators for policy evaluation Liu et al. [2018], Jiang and Li [2016], Dai et al. [2020], Feng et al. [2021], Jiang and Huang [2020]. But none of them is effective in our setting. Importance-sampling is a simple and popular method for off-policy evaluation but suffers exponential variance in horizon Liu et al. [2018]. Marginalized importance-sampling has been proposed to get rid of the exponential variance. However, existing works all focus on function approximations which only produce approximately correct estimators Dai et al. [2020] or are designed for the infinite-horizon case Feng et al. [2021]. Doubly robust estimator Jiang and Li [2016], Hanna et al. [2017], Farajtabar et al. [2018] also solves the exponential variance problem, but no finite sample result is available. Our algorithm is based on marginalized importance-sampling and addresses the above limitations in the sense that our algorithm provides non-asymptotic sample complexity results and works for finite-horizon Markov Decision Processes.

Another popular estimator is called model-based estimator which evaluates the policy by estimating the transition function of the environment Dann et al. [2019], Zanette and Brunskill [2019]. Yin and Wang [2020] provides a similar sample complexity to our results. However, there are some significant differences between their result and our result. First, our sampling distribution is optimal which is calculated based on the coarse distribution estimator. Second, our sample complexity is non-asymptotic while their result is asymptotic. Third, the true distributions appear in our sample complexity can be replaced by known distribution estimators without inducing additional costs which means we can provide a known sample complexity while their result is always unknown since we do not know the true visitation distributions of target policies.

The work that most aligns with ours is Dann et al. [2023] which proposed an on-policy algorithm based on the idea of trajectory synthesis. The authors propose the first instance-dependent sample complexity analysis of the multiple-policy evaluation problem. Different from their work, our algorithm uses off-policy evaluation based on importance-weighting and achieves a better sample complexity with simpler techniques and analysis.

Our algorithm also uses some techniques modified from other works which we summarize here. DualDICE is a technique for estimating distribution ratios by minimizing some loss functions proposed by Nachum et al. [2019]. We build on this idea and make some modifications to meet the need in our setting. Besides, we utilize stochastic gradient descent algorithms and their convergence rate for strongly-convex and smooth functions in the optimization literature Hazan and Kale [2011]. Finally, we adopt the Median of Means estimator Minsker [2023] to convert in-expectation results to high-probability results.

3 Preliminaries and Problem Setup

Notations We denote the set $\{1, 2, \ldots, N\}$ by $[N]$. $\{X_n\}_{n=1}^N$ represents the set $\{X_1, X_2, \ldots, X_N\}$. $\mathbb{E}_\pi$ denotes the expectation over the trajectories produced by following policy π. $\tilde{O}$ hides constants, logarithmic and lower-order terms. And we use $\mathbb{V}[X]$ to represent the variance of random variable X.

Reinforcement learning framework We consider episodic tabular Markov Decision Processes (MDPs) defined by a tuple $\{\mathcal{S}, \mathcal{A}, H, \{P_h\}_{h=1}^H, \{r_h\}_{h=1}^H, \nu\}$ where $\mathcal{S}$ and $\mathcal{A}$ represents the state and action space respectively with S the cardinality of the state space $\mathcal{S}$ and A the cardinality of the action space $\mathcal{A}$. H is the horizon which defines the number of steps the agent can take before the end of an episode. $P_h(\cdot|s,a) \in \Delta\mathcal{S}$ is the transition function which represents the probability of transitioning to the next state if the agent takes action a at state s. And $r_h(s,a)$ is the reward function denotes the reward the agent can get if the agent takes action a at state s. In this work, we assume that the reward is deterministic and bounded $r_h(s,a) \in [0,1]$ which is consistent with prior work Dann et al. [2023]. We denote the initial state distribution by $\nu \in \Delta\mathcal{S}$.

A policy $\pi = \{\pi_h\}_{h=1}^H$ is a mapping from the state space to the probability distribution space over the action space. $\pi_h(a|s)$ denotes the probability of taking action a at state s and step h. The value function $V_h^\pi(s)$ of a policy π is the expected total rewards the agent can receive by starting from step$h$, state s and following the policy π, i.e., $V_h^\pi(s) = \mathbb{E}_\pi[\sum_{l=h}^H r_l|s]$. The performance $J(\pi)$ of a policy π is defined as the expected total rewards the agent can get. By the definition of the value function, there is the relationship $J(\pi) = V_1^\pi(s|s \sim \nu)$. For simplicity, in the following context, we use V_1^π to denote $V_1^\pi(s|s \sim \nu)$.

The state visitation distribution $d_h^\pi(s)$ of a policy π represents the probability of reaching state s at step h if the agent starts from a state sampled from the initial state distribution ν at step $l = 1$ and following policy π subsequently, i.e. $d_h^\pi(s) = \mathbb{P}[s_h = s|s_1 \sim \nu, \pi]$. Similarly, the state-action visitation distribution $d_h^\pi(s,a)$ is defined as $d_h^\pi(s,a) = d_h^\pi(s)\pi(a|s)$. Based on the definition of the visitation distribution, the performance of policy π can also be expressed as $J(\pi) = V_1^\pi = \sum_{h=1}^H \sum_{s,a} d_h^\pi(s,a) r_h(s,a)$.

Multiple-policy evaluation problem setup In multiple-policy evaluation, we are given a set of known policies $\{\pi^k\}_{k=1}^K$ and a pair of factors $\{\epsilon, \delta\}$. The objective is to evaluate the performance of these given policies such that with probability at least $1-\delta, \forall \pi \in \{\pi^k\}_{k=1}^K, |\hat{V}_1^\pi - V_1^\pi| \leq \epsilon$ where $\hat{V}_1^\pi$ is the performance estimator.

Dann et al. [2023] proposed an algorithm based on the idea of trajectory stitching and achieved sample complexity,

$$\tilde{O}\left(\frac{H^2}{\epsilon^2}\mathbb{E}\left[\sum_{(s,a)\in\mathcal{K}^{1:H}} \frac{1}{d^{max}(s)}\right] + \frac{SH^2K}{\epsilon}\right) \tag{1}$$

where $d^{max} = \max_{k\in[K]} d^{\pi^k}$ and $\mathcal{K}^h(s,a) \subseteq [K]$ keeps track of which target policies visit state-action pair (s,a) at step h in their trajectories. This sample complexity can be extremely bad due to the existence of $\frac{1}{d^{\max}(s)}$. The authors Dann et al. [2023] also mentioned that their sample complexity is suboptimal by giving a concrete case.

Another way to reuse samples for evaluation of different policies is model-based method. Based on the model-based estimator proposed by Yin and Wang [2020], an asymptotic convergence rate can be derived,

$$\sqrt{\frac{H}{n}} \cdot \sqrt{\sum_{h=1}^H \mathbb{E}_{\pi^k}\left[\frac{d^{\pi^k}(s_h,a_h)}{\mu(s_h,a_h)}\right]} + o(\frac{1}{\sqrt{n}}) \tag{2}$$

where μ is the distribution of the behavior policy used to generate the offline dataset. Though, it looks similar to our results, we have claimed in the section of related work that there are significant differences.

3.1 Contributions

In this work, we propose an algorithm named CAESAR for multiple-policy evaluation with two phases. In the first phase, we roughly estimate the visitation distributions of target polices with a lower-order sample complexity $\tilde{O}(\frac{1}{\epsilon})$ such that $|\hat{d}_h^{\pi^k}(s,a) - d_h^{\pi^k}(s,a)| \leq \max\{\epsilon, \frac{d_h^{\pi^k}(s,a)}{4}\}, k \in [K]$ where $d_h^{\pi^k}$ is the true visitation distribution of policy π at step h and $\hat{d}_h^{\pi^k}$ is our coarse distribution estimator. In the second phase, with the coarse distribution estimators, we can solve a convex optimization problem to get our optimal sampling distribution $\tilde{\mu}$ with which we estimate the performance of target policies using marginal importance weighting. CAESAR finally achieves that with number of trajectories $n = \tilde{O}\left(\frac{H^4}{\epsilon^2}\sum_{h=1}^H \max_{k\in[K]}\sum_{s,a}\frac{(d_h^{\pi^k}(s,a))^2}{\mu_h^*(s,a)}\right)$ and probability at least $1-\delta$, we can evaluate the performance of all target policies up to ϵ error. CAESAR is consistently better than the naive uniform sampling strategy over target policies. CAESAR also improves the results (1) in Dann et al. [2023] in some cases where their results have a dependency on K while we do not (see Section 4.5).

In addition to our main result, we also provide some results that may spark interest beyond the specific multi-policy evaluation problem we addressed in this work. To utilize marginal importance weighting, we propose an algorithm named IDES to estimate the marginal importance ratio by minimizing a carefully designed step-wise loss function using stochastic gradient descent which is modified from the idea of DualDICE Nachum et al. [2019]. We also utilize a Median-of-Means estimator Minsker [2023] to convert the in-expectation result to the high-probability result which can be of interest.

4 Main Results and Algorithm

In this section, we introduce our algorithm and present our main results. Different from on-policy evaluation, we try to build a single sampling distribution with which we can estimate the performance of all target policies using importance weighting. We achieve it by the following procedures. We first roughly estimate the visitation distributions of target policies at the cost of a lower-order sample complexity. Based on these coarse distribution estimators, we can calculate an optimal sampling distribution by solving a convex optimization problem. Finally, we utilize the idea of DualDICE Nachum et al. [2019] with some modifications to estimate the importance-weighting ratio. We provide the main algorithm scheme in Algorithm 2. In the following sections, we explain these procedures of our algorithm in details.

4.1 Estimation of visitation distributions of target policies

We first introduce a proposition that shows we can roughly estimate the visitation distributions of target policies with just lower-order sample complexity $\tilde{O}(\frac{1}{\epsilon})$. Though, the estimator is coarse and cannot be used to estimate the performance directly which is our ultimate goal, it has some nice properties which enable us to build the optimal sampling distribution and estimate the importance weighting ratio in the following sections.

The idea behind this estimator is based on the following Lemma that shows estimating the mean value of a Bernoulli random variable up to constant multiplicative accuracy only requires $\tilde{O}(\frac{1}{\epsilon})$ samples.

Lemma 4.1. *Let Z_ℓ be i.i.d. samples $Z_\ell \overset{i.i.d.}{\sim} \mathrm{Ber}(p)$, for some known constant $C > 0$, setting $t \geq \frac{C\log(C/\epsilon\delta)}{\epsilon}$, we have that with probability at least $1-\delta$, the empirical mean estimator $\hat{p}_t = \frac{1}{t}\sum_{\ell=1}^t Z_\ell$ satisfies,*

$$|\hat{p}_t - p| \leq \max\{\epsilon, \frac{p}{4}\}.$$

Proof. See Appendix A.1. □

Lemma 4.1 can be used to derive coarse estimators $\hat{d}^{\pi^k} = \{\hat{d}_h^{\pi^k}\}_{h=1}^H$ with constant multiplicative accuracy with respect to the true visitation probabilities $d^{\pi^k} = \{d_h^{\pi^k}\}_{h=1}^H$.

Proposition 4.2. *With number of trajectories $n \geq \frac{CK\log(CK/\epsilon\delta)}{\epsilon} = \tilde{O}(\frac{1}{\epsilon})$, we can estimate $\hat{d}^{\pi^k} = \{\hat{d}_h^{\pi^k}\}_{h=1}^H$ such that with probability at least $1-\delta$, $|\hat{d}_h^{\pi^k}(s,a) - d_h^{\pi^k}(s,a)| \leq \max\{\epsilon, \frac{d_h^{\pi^k}(s,a)}{4}\}$, $\forall s \in \mathcal{S}, a \in \mathcal{A}, h \in [H], k \in [K]$.*

Proof. Based on Lemma 4.1, the proposition can be directly derived by regarding $d_h^{\pi^k}(s,a)$ as a $\mathrm{Bernoulli}$ variable. $\square$

We next show that based on these coarse visitation estimators, we can ignore those states and actions with low estimated visitation probability without inducing significant errors.

Lemma 4.3. *Suppose we have an estimator $\hat{d}(s,a)$ of $d(s,a)$ such that $|\hat{d}(s,a) - d(s,a)| \leq \max\{\epsilon', \frac{d(s,a)}{4}\}$. If $\hat{d}(s,a) \geq 5\epsilon'$, then $\max\{\epsilon', \frac{d(s,a)}{4}\} = \frac{d(s,a)}{4}$, and if $\hat{d}(s,a) \leq 5\epsilon'$, then $d(s,a) \leq 7\epsilon'$.*

Proof. Proof of the first claim: If $\max\{\epsilon', \frac{d(s,a)}{4}\} = \epsilon'$, then we have $d(s,a) \leq 4\epsilon'$ and

$$\hat{d}(s,a) - d(s,a) \leq \epsilon',$$

hence, $\hat{d}(s,a) \leq d(s,a) + \epsilon' \leq 5\epsilon'$ which implies that $\hat{d}(s,a) \leq 5\epsilon'$ is a necessary condition for $\max\{\epsilon', \frac{d(s,a)}{4}\} = \epsilon'$. Proof of the second claim: We have $\hat{d}(s,a) \leq 5\epsilon'$. If $\max\{\epsilon', \frac{d(s,a)}{4}\} = \epsilon'$, then $d(s,a) \leq 4\epsilon'$. If $\max\{\epsilon', \frac{d(s,a)}{4}\} = \frac{d(s,a)}{4}$, then $d(s,a) \leq \frac{4}{3}\hat{d}(s,a) \leq \frac{20}{3}\epsilon' \leq 7\epsilon'$. $\square$

Based on Lemma 4.3, we can ignore the state-action pairs such that $\hat{d}(s,a) \leq 5\epsilon'$. Since if we replace ϵ' by $\frac{\epsilon}{14SA}$, the error of performance estimation induced by ignoring these state-action pair is at most $\frac{\epsilon}{2}$. For simplicity of presentation, we can set $\hat{d}^{\pi}(s,a) = d^{\pi}(s,a) = 0$ if $\hat{d}^{\pi}(s,a) < \frac{5\epsilon}{14SA}$. Hence, we have that,

$$|\hat{d}_h^{\pi^k}(s,a) - d_h^{\pi^k}(s,a)| \leq \frac{d_h^{\pi^k}(s,a)}{4}, \ \forall s \in \mathcal{S}, a \in \mathcal{A}, h \in [H], k \in [K]. \tag{3}$$

4.2 Optimal sampling distribution

We evaluate the expected total rewards of target policies by importance weighting using samples $\{s_1^i, a_1^i, s_2^i, a_2^i, \ldots, s_H^i, a_H^i\}_{i=1}^n$ sampled from a sampling distribution $\{\mu_h\}_{h=1}^H$.

$$\hat{V}_1^{\pi^k} = \frac{1}{n}\sum_{i=1}^{n}\sum_{h=1}^{H}\frac{d_h^{\pi^k}(s_h^i, a_h^i)}{\mu_h(s_h^i, a_h^i)} r(s_h^i, a_h^i), \ k \in [K].$$

From the perspective of minimizing the variance of our estimator (see Appendix A.2), we can find the optimal sampling distribution by solving the following optimization problem,

$$\mu_h^* = \arg\min_{\mu} \max_{k\in[K]} \sum_{s,a} \frac{(d_h^{\pi^k}(s,a))^2}{\mu(s,a)}, \ h \in [H]. \tag{4}$$

The above problem is a convex optimization problem. However, the optimal μ_h^* can be outside of the convex hull of $\{d_h^{\pi^k}\}_{k=1}^K$. In some cases, the optimal μ^* may not be realized by any policy (see Appendix A.3). Hence, in order to have an easy way of sampling from μ^*, we constrain μ to be inside of the convex hull of $\{d_h^{\pi^k}\}_{k=1}^K$ and solve the constrained optimization problem,

$$\mu_h^* = \arg\min_{\mu\in M_h} \max_{k\in[K]} \sum_{s,a} \frac{(d_h^{\pi^k}(s,a))^2}{\mu(s,a)}, \ h = 1, \ldots, H, \tag{5}$$

where $M_h = \{\mu : \mu = \sum_{k=1}^K \alpha_k \hat{d}_h^{\pi^k}, \alpha_k \geq 0, \sum_{k=1}^K \alpha_k = 1\}$. Denote by $\mu_h^* = \sum_{k=1}^K \alpha_k^* d_h^{\pi^k}$ the optimal solution.

Notice that we do not know d_h^{π}, so we can only solve the approximate optimization problem,

$$\hat{\mu}_h^* = \arg\min_{\mu\in M_h} \max_{k\in[K]} \sum_{s,a} \frac{(\hat{d}_h^{\pi^k}(s,a))^2}{\mu(s,a)}, \ h = 1, \ldots, H, \tag{6}$$

and we denote the optimal solution by $\hat{\mu}_h^* = \sum_{k=1}^K \hat{\alpha}_k^* \hat{d}_h^{\pi^k}$. Correspondingly, our real sampling distribution would be $\tilde{\mu}_h^* = \sum_{k=1}^K \hat{\alpha}_k^* d_h^{\pi^k}$.

Remark 4.1. *We can constrain μ to be the convex combination of the visitation distributions of all deterministic policies instead of the target policies. In that case, we can potentially get a better sampling distribution at the cost of exponential dependency of S and A in the lower order sample complexity since we need to roughly estimate the visitation distributions of all deterministic policies. However, we believe there exists a clever way of coarse estimation to get rid of the exponential dependency in the lower order sample complexity.*

The next lemma tells us that the optimal sampling distribution also has the same property as the coarse distribution estimators.

Lemma 4.4. *If property (3) holds: $|\hat{d}_h^{\pi^k}(s,a) - d_h^{\pi^k}(s,a)| \leq \frac{d_h^{\pi^k}(s,a)}{4}$, $\forall s \in \mathcal{S}, a \in \mathcal{A}, h \in [H], k \in [K]$, then*

1. $|\tilde{\mu}_h^*(s,a) - \hat{\mu}_h^*(s,a)| \leq \frac{\tilde{\mu}_h^*(s,a)}{4}$ *and therefore* $\frac{1}{\tilde{\mu}_h^*(s,a)} \leq \frac{5}{4\hat{\mu}_h^*(s,a)}$.
2. $\frac{d_h^{\pi^k}(s,a)}{\tilde{\mu}_h^*(s,a)} \leq \frac{5\hat{d}_h^{\pi^k}(s,a)}{3\hat{\mu}_h^*(s,a)}$ $\forall k \in [K]$.

Proof. First,$|\hat{\mu}_h^*(s,a) - \tilde{\mu}_h^*(s,a)| = |\sum_{k=1}^K \hat{\alpha}_k^*(\hat{d}_h^{\pi^k}(s,a) - d_h^{\pi^k}(s,a))| \leq \sum_{k=1}^K \hat{\alpha}_k^* |\hat{d}_h^{\pi^k}(s,a) - d_h^{\pi^k}(s,a)| \leq \frac{1}{4}\sum_{k=1}^K \hat{\alpha}_k^* d_h^{\pi^k}(s,a) = \frac{\tilde{\mu}_h^*(s,a)}{4}$. Second, $\frac{d_h^{\pi^k}(s,a)}{\tilde{\mu}_h^*(s,a)} \leq \frac{4\hat{d}_h^{\pi^k}(s,a)}{3\tilde{\mu}_h^*(s,a)} \leq \frac{5\hat{d}_h^{\pi^k}(s,a)}{3\hat{\mu}_h^*(s,a)}$. ☐

4.3 Estimation of the importance weighting ratio

In this section, we introduce our algorithm for estimating the importance weighting ratios which is sketched out in Algorithm 1. Our algorithm is based on the idea of DualDICE Nachum et al. [2019]. In DualDICE, they propose the following loss function

$$\ell^\pi(w) = \frac{1}{2}\mathbb{E}_{s,a\sim\mu}\left[w^2(s,a)\right] - \mathbb{E}_{s,a\sim d^\pi}\left[w(s,a)\right], \tag{7}$$

the optimal minimum is achieved at $w^{\pi,*}(s,a) = \frac{d^\pi(s,a)}{\mu(s,a)}$ which is the distribution ratio. They address the on-policy limitation of the second term in 7 by changing the variable based on Bellman's equation. However, their method only works for infinite horizon MDPs and it becomes unclear how to optimize the loss function after the variable change. We propose a new step-wise loss function which works for finite horizon MDPs. More importantly, the loss function is strongly-convex and smooth, hence, we can give finite sample results based on the convergence rate in the optimization literature.

Specifically, we define a loss function at each step and optimize them from step $h = 1$ to step $h = H$. We define the loss function of policy π at step h as,

$$\begin{aligned}
\ell_h^\pi(w) &= \frac{1}{2}\sum_{s,a}\frac{\tilde{\mu}_h(s,a)}{\hat{\mu}_h(s,a)}w^2(s,a) - \sum_{s,a}\sum_{s',a'}\tilde{\mu}_{h-1}(s',a')P(s|s',a')\pi(a|s)\frac{\hat{w}_{h-1}(s',a')}{\hat{\mu}_{h-1}(s',a')}w(s,a) \\
&= \frac{1}{2}\mathbb{E}_{s,a\sim\tilde{\mu}_h}\left[\frac{w^2(s,a)}{\hat{\mu}_h(s,a)}\right] - \mathbb{E}_{s',a'\sim\tilde{\mu}_{h-1},s\sim P_{h-1}(\cdot|s',a')}\left[\sum_a \frac{\hat{w}_{h-1}(s',a')}{\hat{\mu}_{h-1}(s',a')}w(s,a)\pi(a|s)\right]
\end{aligned}$$

where $\tilde{\mu}_h = \sum_{k=1}^K \hat{\alpha}_k d_h^{\pi^k}$ is the sampling distribution, and $\hat{\mu}_h = \sum_{k=1}^K \hat{\alpha}_k \hat{d}_h^{\pi^k}$ is the optimal solution of the approximate opt problem (refer to Section 4.2) and we set $\tilde{\mu}_0$ as empty, $\hat{w}_0 = \hat{\mu}_0 = 1$ for notation simplicity.

It is clear that the loss function is strongly convex and smooth. An even better property about our loss function is that the smoothness factor ξ and strongly-convexity factor γ are well bounded based on the property of our coarse distribution estimator, i.e. $\frac{4}{5} \leq \frac{\tilde{\mu}_h(s,a)}{\hat{\mu}_h(s,a)} \leq \frac{4}{3}$ which is a trivial conclusion from Lemma 4.4,

$$\xi = \max_{s,a}\frac{\tilde{\mu}_h(s,a)}{\hat{\mu}_h(s,a)},\ \gamma = \min_{s,a}\frac{\tilde{\mu}_h(s,a)}{\hat{\mu}_h(s,a)},\ \frac{\xi}{\gamma} \leq \frac{5}{3}. \tag{8}$$

This property actually plays an important role in the final sample complexity which we will revisit later. In the following lemma, we show that our step-wise loss function has an another nice property on step-to-step error propagation.

Lemma 4.5. *Suppose we have an estimator $\hat{w}_{h-1}$ at step $h-1$ such that,*

$$\sum_{s,a}\left|\tilde{\mu}_{h-1}(s,a)\frac{\hat{w}_{h-1}(s,a)}{\hat{\mu}_{h-1}(s,a)} - d_{h-1}^\pi(s,a)\right| \leq \epsilon,$$

Algorithm 1 **I**mportance **D**ensity **Es**timation (IDES)

Input: Horizon H, feasible sets $\{D_h\}_{h=1}^H$, accuracy ϵ, target policy π, coarse estimator $\{\hat{d}_h^\pi\}_{h=1}^H$ and $\{\hat{\mu}_h\}_{h=1}^H$.
Initialize $w_h^0 = 0,\ h = 1, \ldots, H$ and assume $\mu_0 = Empty$ for simple presentation.
for $h = 1$ **to** H **do**
  Set the iteration number of optimization, $n_h = C_h \left(\frac{H^4}{\epsilon^2} \sum_{s,a} \frac{(\hat{d}_h^\pi(s,a))^2}{\hat{\mu}_h(s,a)} + \frac{(\hat{d}_{h-1}^\pi(s,a))^2}{\hat{\mu}_{h-1}(s,a)} \right)$, where C_h is a known constant.
  for $i = 1$ **to** n_h **do**
    Sample $\{s_h^i, a_h^i\}$ from μ_h and $\{s_{h-1}^i, a_{h-1}^i, s_h^{i'}\}$ from μ_{h-1}.
    Calculate gradient $g(w_h^{i-1})$,

$$g(w_h^{i-1})(s,a) = \frac{w_h^{i-1}(s,a)}{\hat{\mu}_h(s,a)} \mathbb{I}(s_h^i = s, a_h^i = a) - \frac{\hat{w}_{h-1}(s_{h-1}^i, a_{h-1}^i)}{\hat{\mu}_{h-1}(s_{h-1}^i, a_{h-1}^i)} \pi(a|s) \mathbb{I}(s_h^{i'} = s)$$

    Update $w_h^i = Proj_{w \in D_h}\{w_h^{i-1} - \eta_h^i g(w_h^{i-1})\}$.
  end for
  Output the estimator $\hat{w}_h = \frac{1}{\sum_{i=1}^{n_h} i} \sum_{i=1}^{n_h} w_h^i$.
end for

then by minimizing the loss function $\ell_h^\pi(w)$ at step h to $\|\nabla \ell_h^\pi(\hat{w}_h(s,a))\|_1 \le \epsilon$, we have,

$$\sum_{s,a} \left| \tilde{\mu}_h(s,a) \frac{\hat{w}_h(s,a)}{\hat{\mu}_h(s,a)} - d_h^\pi(s,a) \right| \le 2\epsilon.$$

Proof. See Appendix A.4. □

Before introducing the sample complexity of estimating the distribution ratios by Algorithm 1, we first give the following convergence rate of minimizing strongly-convex loss functions with stochastic gradient descent.

Lemma 4.6 (Theorem 25.2 in Notes)**.** *For a $\lambda-$strongly convex loss function $L(w)$ satisfying $\|w^*\| \le D$ for some known D, there exists a stochastic gradient descent algorithm that can output $\hat{w}$ after T iterations such that,*

$$\mathbb{E}[L(\hat{w}) - L(w^*)] \le \frac{2G^2}{\lambda(T+1)},$$

where G^2 is the variance bound of the stochastic gradient.

Based on the above results, implementing Algorithm 1 with iterations n_h specified in the algorithm, we can achieve an accurate estimation of the distribution ratios which we formalize in the following lemma.

Lemma 4.7. *Implement Algorithm 1 with parameters n_h specified in the algorithm, we have,*

$$\mathbb{E}\left[\sum_{s,a} \left| \tilde{\mu}_h(s,a) \frac{\hat{w}_h(s,a)}{\hat{\mu}_h(s,a)} - d_h^{\pi^k}(s,a) \right| \right] \le \frac{\epsilon}{4H}.$$

Proof. See Appendix A.5 for complete proof. Here we provide proof sketch. By Lemma 4.5 and the property of smoothness $\|\nabla \ell_h^\pi(\hat{w}_h)\|_1^2 \le 2\xi(\ell_h^\pi(\hat{w}_h) - \ell_h^\pi(w_h^*))$, we need to minimize the loss function at each step h such that $\ell_h^\pi(\hat{w}_h) - \ell_h^\pi(w_h^*) \le \frac{\epsilon^2}{32\xi H^4}$. Invoking the convergence rate in Lemma 4.6, we need samples $n = O\left(\frac{\xi}{\gamma} \frac{H^4 G^2}{\epsilon^2}\right)$. Remember that we have shown in (8) that $\frac{\xi}{\gamma} \le \frac{5}{3}$, this nice property helps us to get rid of the undesired ratio of the smoothness factor and the strongly-convexity factor, i.e. $\frac{\max_{s,a} \mu(s,a)}{\min_{s,a} \mu(s,a)}$ of the original loss function (7) which can be extremely bad. Finally, replacing G^2 by $O\left(\sum_{s,a} \frac{(\hat{d}_h^\pi(s,a))^2}{\hat{\mu}_h(s,a)} + \frac{(\hat{d}_{h-1}^\pi(s,a))^2}{\hat{\mu}_{h-1}(s,a)} \right)$ which is the variance bound of the stochastic gradient in our case. □

4.4 Main results

In this section, we provide our final sample complexity result of estimating the performance of all target policies based on the results we have in previous sections. We first give the main theorem and then provide the derivations.

Theorem 4.8. *Implement Algorithm 2, then with probability at least* $1-\delta$*, for all target policies, we have that* $|\hat{V}_1^{\pi^k} - V_1^{\pi^k}| \le \epsilon$*. And the total number of trajectories sampled is,*

$$n = \tilde{O}\left(\frac{H^4}{\epsilon^2}\sum_{h=1}^{H}\max_{k\in[K]}\sum_{s,a}\frac{(d_h^{\pi^k}(s,a))^2}{\mu_h^*(s,a)}\right) \tag{9}$$

Besides, the unknown true visitation distributions can be replaced by the coarse estimator to provide a concrete sample complexity,

$$n = \tilde{O}\left(\frac{H^4}{\epsilon^2}\sum_{h=1}^{H}\max_{k\in[K]}\sum_{s,a}\frac{(\hat{d}_h^{\pi^k}(s,a))^2}{\hat{\mu}_h(s,a)}\right)$$

Now, we explain how the above result is derived. We first introduce a Median-of-Means (MoM) estimator and data splitting technique which can conveniently convert Lemma 4.7 to a version holds with high probability.

Lemma 4.9. *For a one-dimension value* μ*, suppose we have a stochastic estimator* $\hat{\mu}$ *such that* $\mathbb{E}[|\hat{\mu}-\mu|] \le \frac{\epsilon}{4}$*, then if we generate* $N = O(\log(1/\delta))$ *i.i.d. estimators* $\{\hat{\mu}_1, \hat{\mu}_2, \dots, \hat{\mu}_N\}$ *and choose* $\hat{\mu}_{MoM} = Median(\hat{\mu}_1, \hat{\mu}_2, \dots, \hat{\mu}_N)$*, we have with probability at least* $1-\delta$*,*

$$|\hat{\mu}_{MoM} - \mu| \le \epsilon$$

Proof. See Appendix A.6. □

For step h, Algorithm 1 can output a solution $\hat{w}_h$ such that $\mathbb{E}[\ell_h^\pi(\hat{w}_h) - \ell_h^\pi(w_h^*)] \le \frac{\epsilon^2}{32\xi H^4}$. We can apply Lemma 4.9 on our algorithm which means that we can run the algorithm for $N = O(\log(1/\delta))$ times. Hence, we will get N solutions $\{\hat{w}_{h,1}, \hat{w}_{h,2}, \dots, \hat{w}_{h,N}\}$. Set $\hat{w}_{h,MoM}$ as the solution such that $\ell_h^\pi(\hat{w}_{h,MoM}) = Median(\ell_h^\pi(\hat{w}_{h,1}), \ell_h^\pi(\hat{w}_{h,2}), \dots, \ell_h^\pi(\hat{w}_{h,N}))$. Based on Lemma 4.9, we have that with probability at least $1-\delta$, it holds $\ell_h^\pi(\hat{w}_{h,MoM}) - \ell_h^\pi(w_h^*) \le \frac{\epsilon^2}{32\xi H^4}$. With a little abuse of notation, we just denote $\hat{w}_{h,MoM}$ by $\hat{w}_h$ in the following content.

Now we are ready to estimate the total expected rewards of target policies, With the importance weighting ratio estimator $\frac{\hat{w}_h(s,a)}{\hat{\mu}_h(s,a)}$ from Algorithm 1, we can estimate the performance of policy π^k,

$$\hat{V}_1^{\pi^k} = \frac{1}{n}\sum_{i=1}^{n}\sum_{h=1}^{H}\frac{\hat{w}_h^{\pi^k}(s_h^i, a_h^i)}{\hat{\mu}_h(s_h^i, a_h^i)} r_h(s_h^i, a_h^i) \tag{10}$$

where $\{s_h^i, a_h^i\}_{i=1}^n$ is sampled from $\tilde{\mu}_h$.

Lemma 4.10. *With samples* $n = \tilde{O}\left(\frac{H^2}{\epsilon^2}\sum_{h=1}^{H}\max_{k\in[K]}\sum_{s,a}\frac{(\hat{d}_h^{\pi^k}(s,a))^2}{\hat{\mu}_h(s,a)}\right)$*, we have with probability at least* $1-\delta$*,* $|\hat{V}_1^{\pi^k} - V_1^{\pi^k}| \le \frac{\epsilon}{2},\ k \in [K]$.

Proof. First, we can decompose the error $|\hat{V}_1^{\pi^k} - V_1^{\pi^k}| = |\hat{V}_1^{\pi^k} - \mathbb{E}[\hat{V}_1^{\pi^k}] + \mathbb{E}[\hat{V}_1^{\pi^k}] - V_1^{\pi^k}| \le |\hat{V}_1^{\pi^k} - \mathbb{E}[\hat{V}_1^{\pi^k}]| + |\mathbb{E}[\hat{V}_1^{\pi^k}] - V_1^{\pi^k}|$. Then, by Bernstein's inequality, with samples $n = \tilde{O}\left(\frac{H^2}{\epsilon^2}\sum_{h=1}^{H}\max_{k\in[K]}\sum_{s,a}\frac{(\hat{d}_h^{\pi^k}(s,a))^2}{\hat{\mu}_h(s,a)}\right)$, we have, $|\hat{V}_1^{\pi^k} - \mathbb{E}[\hat{V}_1^{\pi^k}]| \le \frac{\epsilon}{4}$. Based Lemma 4.7, we have, $|\mathbb{E}[\hat{V}_1^{\pi^k}] - V_1^{\pi^k}| \le \frac{\epsilon}{4}$. □

Remember that in Section 4.1, we ignore those states and actions with low estimated visitation distribution for each target policy which induce at most $\frac{\epsilon}{2}$ error. Combined with Lemma 4.10, our estimator $\hat{V}_1^{\pi^k}$ finally achieves that with probability at least $1-\delta$, $|\hat{V}_1^{\pi^k} - V_1^{\pi^k}| \le \epsilon, k \in [K]$.

And for sample complexity, in our algorithm, we need to sample data in three procedures. First, for the coarse estimation of the visitation distribution, we need $\tilde{O}(\frac{1}{\epsilon})$ samples. Second, to estimate the importance-weighting ratio,

Algorithm 2 **C**oarse and **A**daptive **Es**timation with **A**pproximate **R**eweighing for Multi-Policy Evaluation (CAESAR)

Input: Accuracy ϵ, confidence δ, target policies $\{\pi^k\}_{k=1}^K$
Roughly estimate visitation distributions of all target policies and get $\{\hat{d}^{\pi^k}\}_{k=1}^K$.
Solve the approximate optimization problem (6) and get $\{\hat{\alpha}_k^*\}_{k=1}^K$.
Implement Algorithm 1 and get $\{\hat{w}^{\pi^k}\}_{k=1}^K$.
Build the final performance estimator $\{\hat{V}_1^{\pi^k}\}_{k=1}^K$ by (10).
Output: $\{\hat{V}_1^{\pi^k}\}_{k=1}^K$.

we need samples $\tilde{O}\left(\frac{H^4}{\epsilon^2}\sum_{h=1}^H \max_{k\in[K]}\sum_{s,a}\frac{(d_h^{\pi^k}(s,a))^2}{\mu_h^*(s,a)}\right)$. Last, to build the final performance estimator (10), we need samples $\tilde{O}\left(\frac{H^2}{\epsilon^2}\sum_{h=1}^H \max_{k\in[K]}\sum_{s,a}\frac{(\hat{d}_h^{\pi^k}(s,a))^2}{\hat{\mu}_h(s,a)}\right)$. Therefore, the total trajectories needed,

$$n = \tilde{O}\left(\frac{H^4}{\epsilon^2}\sum_{h=1}^H \max_{k\in[K]}\sum_{s,a}\frac{(d_h^{\pi^k}(s,a))^2}{\mu_h^*(s,a)}\right).$$

Moreover, notice that,

$$\max_{k\in[K]}\sum_{s,a}\frac{(\hat{d}_h^{\pi^k}(s,a))^2}{\hat{\mu}_h(s,a)} \leq \max_{k\in[K]}\sum_{s,a}\frac{(\hat{d}_h^{\pi^k}(s,a))^2}{\mu_h^*(s,a)} \leq \frac{25}{16}\sum_{s,a}\frac{(d_h^{\pi}(s,a))^2}{\mu_h^*(s,a)} \tag{11}$$

where μ_h^* is the optimal solution of the optimization problem (5), the first inequality holds due to $\hat{\mu}_h$ is the minimum of the approximate optimization problem (6) and the second inequality holds due to $\hat{d}_h^{\pi}(s,a) \leq \frac{5}{4}d_h^{\pi}(s,a)$. Based on (11), we can substitute the coarse distribution estimator in the sample complexity bound by the exact one,

$$n = \tilde{O}\left(\frac{H^4}{\epsilon^2}\sum_{h=1}^H \max_{k\in[K]}\sum_{s,a}\frac{(d_h^{\pi^k}(s,a))^2}{\mu_h^*(s,a)}\right).$$

4.5 Discussion

In this section, we analyse our sample complexity and compare it with existing results. For off-policy evaluation, the CR-lower bound proposed by Jiang and Li [2016] (Theorem 3) shows that there exists a MDP such that the variance of any unbiased estimator is lower bounded by

$$\sum_{h=1}^H \mathbb{E}_\mu\left[\left(\frac{d_h^\pi(s_h,a_h)}{\mu_h(s_h,a_h)}\right)^2 \mathbb{V}[V_h^\pi(s_h)]\right],$$

where π is the policy to evaluate and μ is the sampling distribution generated by some behavior policy. If we simply bound the variance of the value function by H^2 due to the boundness of rewards and omit the dependency on H which we will discuss later, our sample complexity matches the lower bound for multiple-policy evaluation since our sampling distribution is optimal.

One interesting proposition from our sample complexity result is that in the case where all target policies are identical, i.e. $d^{\pi^1} = d^{\pi^2} = \cdots = d^{\pi^K} = d$. Then, the optimal sampling distribution is $\mu^* = d$, hence, the sample complexity is $\frac{H^5}{\epsilon^2}$ which has no dependency on S or A.

We can also derive a non-instance dependent sample complexity based on our results. Let the sampling distribution μ_h' be $\frac{1}{SA}\sum_{s,a} d_h^{\pi_{s,a}}$, where $\pi_{s,a} = \arg\max_{k\in[K]} d_h^{\pi^k}(s,a)$. Since μ_h^* is the optimal solution and μ_h' is a feasible solution, we have our sample complexity (9) is bounded,

$$n \leq \tilde{O}\left(\frac{H^5 SA}{\epsilon^2}\right).$$

Now, we compare our result with the one achieved by Dann et al. [2023]. One big problem of the result by Dann et al. [2023] is the existence of unfavorable $\frac{1}{d^{max}(s)}$ which can induce an undesired dependency on K. We next illustrate it by an example. Consider such a MDP with two layers, a single initial state $s_{1,1}$ in the first layer and two terminal states in the second layer $s_{2,1}, s_{2,2}$. The transition function is same for all actions, i.e. $P(s_{2,1}|s_{1,1}, a) = p$ and p is sufficiently small. The agents only receive rewards at state $s_{2,1}$ no matter what actions they take. Hence, to evaluate the performance of some policy under this MDP, it is enough to only consider the second layer.

Now, suppose we have K target policies to evaluate such that they take different actions at state $s_{1,1}$ and take the same action at any state in the second layer. Since the transition function at state $s_{1,1}$ is same for any action, the visitation distribution at state $s_{2,1}$ of all target policies is identical.

Since p is sufficiently small, we have that the probability of $s_{2,1}$ being reached is $\mathbb{P}[s_{2,1} \in \mathcal{K}^2] = 1 - (1-p)^K \approx pK$. Invoke the result (1) of Dann et al. [2023], the sample complexity in this case has a dependency on K. At the same time, our result can show the sample complexity without dependency on K. However, it is unclear whether our result is better in all cases.

Finally, comparing with the naive uniform sampling strategy over target policies, our method is clearly at advantage since our sampling distribution is optimal within all possible combinations of the target policies which include the naive uniform one.

5 Conclusion and Future Work

In this work, we consider the problem of multi-policy evaluation. And we propose an algorithm CAESAR based on computing an approximate optimal offline sampling distribution and using the data sampled from it to perform the simultaneous estimation of the policy values. CAESAR achieves that with number of trajectories $n = \tilde{O}\left(\frac{H^4}{\epsilon^2}\sum_{h=1}^{H}\max_{k\in[K]}\sum_{s,a}\frac{(d_h^{\pi^k}(s,a))^2}{\mu_h^*(s,a)}\right)$ and probability at least $1-\delta$, we can evaluate the performance of all target policies up to ϵ error. The algorithm consists of three techniques. First, we obtain a coarse distribution estimator at the cost of lower-order sample complexity. Second, based on the coarse distribution estimator, we show an achievable optimal sampling distribution by solving an convex optimization problem. Last, we propose a novel step-wise loss function for finite-horizon MDPs. By minimizing the loss function step to step, we are able to get the importance weighting ratio and a non-asymptotic sample complexity is available due to the smoothness and strongly-convexity of the loss function.

Beyond the result of this work, there are still some open questions of interest. First, our sample complexity has a dependency on H^4 which is induced by the error propagation in the estimation of the importance weighting ratios. Specifically, the error of minimizing the loss function at early steps, e.g $h = 1$ will propagate to later steps e.g $h = H$. We conjecture a dependency on H^2 is possible by considering a comprehensive loss function includes the whole horizon instead of step-wise loss functions which require step by step optimization. Besides, considering a reward-dependent sample complexity is also an interesting direction. For example, consider a MDP with sparse rewards where only one state-action has non-zero reward, then a better sample complexity may be possible by just focusing on state-action pairs with non-zero rewards. Another future direction is to apply the coarse distribution estimator on more scenarios. In our work, the coarse distribution estimator plays an important role throughout the algorithm. And we believe this type of estimator has potentiality in other different settings and tasks.

Acknowledgements. This work was supported in part by funding from the Eric and Wendy Schmidt Center at the Broad Institute of MIT and Harvard, the NSF under grants CCF-2200052 and IIS-1914792, the ONR under grants N00014-19-1-2571 and N00014-21-1-2844, the NIH under grant 1UL1TR001430, the DOE under grant DE-AC02-05CH11231, the ARPA-E under grant and DE-AR0001282, and the Boston University Kilachand Fund for Integrated Life Science and Engineering.

References

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. PMLR, 2014.

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.

Bo Dai, Ofir Nachum, Yinlam Chow, Lihong Li, Csaba Szepesvári, and Dale Schuurmans. Coindice: Off-policy confidence interval estimation. *Advances in neural information processing systems*, 33:9398–9411, 2020.

Chris Dann, Mohammad Ghavamzadeh, and Teodor V Marinov. Multiple-policy high-confidence policy evaluation. In *International Conference on Artificial Intelligence and Statistics*, pages 9470–9487. PMLR, 2023.

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

Yaqi Duan, Zeyu Jia, and Mengdi Wang. Minimax-optimal off-policy evaluation with linear function approximation. In *International Conference on Machine Learning*, pages 2701–2709. PMLR, 2020.

Mehrdad Farajtabar, Yinlam Chow, and Mohammad Ghavamzadeh. More robust doubly robust off-policy evaluation. In *International Conference on Machine Learning*, pages 1447–1456. PMLR, 2018.

Yihao Feng, Ziyang Tang, Na Zhang, and Qiang Liu. Non-asymptotic confidence intervals of off-policy evaluation: Primal and dual bounds. *arXiv preprint arXiv:2103.05741*, 2021.

Raphael Fonteneau, Susan A Murphy, Louis Wehenkel, and Damien Ernst. Batch mode reinforcement learning based on the synthesis of artificial trajectories. *Annals of operations research*, 208:383–416, 2013.

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

Josiah P Hanna, Peter Stone, and Scott Niekum. Bootstrapping with models: Confidence intervals for off-policy evaluation. In *Proceedings of the 16th Conference on Autonomous Agents and MultiAgent Systems*, pages 538–546, 2017.

Elad Hazan and Satyen Kale. Beyond the regret minimization barrier: an optimal algorithm for stochastic strongly-convex optimization. In *Proceedings of the 24th Annual Conference on Learning Theory*, pages 421–436. JMLR Workshop and Conference Proceedings, 2011.

Nan Jiang and Jiawei Huang. Minimax value interval for off-policy evaluation and policy optimization. *Advances in Neural Information Processing Systems*, 33:2747–2758, 2020.

Nan Jiang and Lihong Li. Doubly robust off-policy value evaluation for reinforcement learning. In *International conference on machine learning*, pages 652–661. PMLR, 2016.

Lihong Li, Rémi Munos, and Csaba Szepesvári. Toward minimax off-policy value estimation. In *Artificial Intelligence and Statistics*, pages 608–616. PMLR, 2015.

Qiang Liu, Lihong Li, Ziyang Tang, and Dengyong Zhou. Breaking the curse of horizon: Infinite-horizon off-policy estimation. *Advances in neural information processing systems*, 31, 2018.

Stanislav Minsker. Efficient median of means estimator. In *The Thirty Sixth Annual Conference on Learning Theory*, pages 5925–5933. PMLR, 2023.

Ofir Nachum, Yinlam Chow, Bo Dai, and Lihong Li. Dualdice: Behavior-agnostic estimation of discounted stationary distribution corrections. *Advances in neural information processing systems*, 32, 2019.

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

Richard S Sutton, David McAllester, Satinder Singh, and Yishay Mansour. Policy gradient methods for reinforcement learning with function approximation. *Advances in neural information processing systems*, 12, 1999.

Devinder Thapa, In-Sung Jung, and Gi-Nam Wang. Agent based decision support system using reinforcement learning under emergency circumstances. In *Advances in Natural Computation: First International Conference, ICNC 2005, Changsha, China, August 27-29, 2005, Proceedings, Part I 1*, pages 888–892. Springer, 2005.

Ruosong Wang, Simon S Du, Lin F Yang, and Sham M Kakade. Is long horizon reinforcement learning more difficult than short horizon reinforcement learning? *arXiv preprint arXiv:2005.00527*, 2020.

Tengyang Xie, Yifei Ma, and Yu-Xiang Wang. Towards optimal off-policy evaluation for reinforcement learning with marginalized importance sampling. *Advances in neural information processing systems*, 32, 2019.

Tengyang Xie, Dylan J. Foster, Yu Bai, Nan Jiang, and Sham M. Kakade. The role of coverage in online reinforcement learning. *ArXiv*, abs/2210.04157, 2022. URL `https://api.semanticscholar.org/CorpusID:252780137`.

Ming Yin and Yu-Xiang Wang. Asymptotically efficient off-policy evaluation for tabular reinforcement learning. In *International Conference on Artificial Intelligence and Statistics*, pages 3948–3958. PMLR, 2020.

Ming Yin, Yu Bai, and Yu-Xiang Wang. Near-optimal provable uniform convergence in offline policy evaluation for reinforcement learning. In *International Conference on Artificial Intelligence and Statistics*, pages 1567–1575. PMLR, 2021.

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*, pages 7304–7312. PMLR, 2019.

A Proof of theorems and lemmas in Section 4

A.1 Proof of Lemma 4.1

Our results relies on the following variant of Bernstein inequality for martingales, or Freedman's inequality Freedman [1975], as stated in e.g., Agarwal et al. [2014], Beygelzimer et al. [2011].

Lemma A.1 (Simplified Freedman's inequality)**.** *Let $X_1, ..., X_T$ be a bounded martingale difference sequence with $|X_\ell| \leq R$. For any $\delta' \in (0, 1)$, and $\eta \in (0, 1/R)$, with probability at least $1 - \delta'$,*

$$\sum_{\ell=1}^{T} X_\ell \leq \eta \sum_{\ell=1}^{T} \mathbb{E}_\ell[X_\ell^2] + \frac{\log(1/\delta')}{\eta}. \tag{12}$$

where $\mathbb{E}_\ell[\cdot]$ is the conditional expectation[2] induced by conditioning on $X_1, \cdots, X_{\ell-1}$.

Lemma A.2 (Anytime Freedman)**.** *Let $\{X_t\}_{t=1}^{\infty}$ be a bounded martingale difference sequence with $|X_t| \leq R$ for all $t \in \mathbb{N}$. For any $\delta' \in (0, 1)$, and $\eta \in (0, 1/R)$, there exists a universal constant $C > 0$ such that for all $t \in \mathbb{N}$ simultaneously with probability at least $1 - \delta'$,*

$$\sum_{\ell=1}^{t} X_\ell \leq \eta \sum_{\ell=1}^{t} \mathbb{E}_\ell[X_\ell^2] + \frac{C \log(t/\delta')}{\eta}. \tag{13}$$

where $\mathbb{E}_\ell[\cdot]$ is the conditional expectation induced by conditioning on $X_1, \cdots, X_{\ell-1}$.

Proof. This result follows from Lemma A.1. Fix a time-index t and define $\delta_t = \frac{\delta'}{12t^2}$. Lemma A.1 implies that with probability at least $1 - \delta_t$,

$$\sum_{\ell=1}^{t} X_\ell \leq \eta \sum_{\ell=1}^{t} \mathbb{E}_\ell \left[X_\ell^2\right] + \frac{\log(1/\delta_t)}{\eta}.$$

A union bound implies that with probability at least $1 - \sum_{\ell=1}^{t} \delta_t \geq 1 - \delta'$,

$$\begin{aligned}\sum_{\ell=1}^{t} X_\ell &\leq \eta \sum_{\ell=1}^{t} \mathbb{E}_\ell \left[X_\ell^2\right] + \frac{\log(12t^2/\delta')}{\eta} \\ &\overset{(i)}{\leq} \eta \sum_{\ell=1}^{t} \mathbb{E}_\ell \left[X_\ell^2\right] + \frac{C \log(t/\delta')}{\eta}.\end{aligned}$$

holds for all $t \in \mathbb{N}$. Inequality (i) holds because $\log(12t^2/\delta') = \mathcal{O}\left(\log(t\delta')\right)$.

□

Proposition A.3. *Let $\delta' \in (0, 1)$, $\beta \in (0, 1]$ and $Z_1, \cdots, Z_T$ be an adapted sequence satisfying $0 \leq Z_\ell \leq \tilde{B}$ for all $\ell \in \mathbb{N}$. There is a universal constant $C' > 0$ such that,*

$$(1-\beta)\sum_{t=1}^{T} \mathbb{E}_t[Z_t] - \frac{2\tilde{B}C' \log(t/\delta')}{\beta} \leq \sum_{\ell=1}^{T} Z_\ell \leq (1+\beta)\sum_{t=1}^{T} \mathbb{E}_t[Z_t] + \frac{2\tilde{B}C' \log(t/\delta')}{\beta}$$

with probability at least $1 - 2\delta'$ simultaneously for all $T \in \mathbb{N}$.

Proof. Consider the martingale difference sequence $X_t = Z_t - \mathbb{E}_t[Z_t]$. Notice that $|X_t| \leq \tilde{B}$. Using the inequality of Lemma A.2 we obtain for all $\eta \in (0, 1/B^2)$.

$$\begin{aligned}\sum_{\ell=1}^{t} X_\ell &\leq \eta \sum_{\ell=1}^{t} \mathbb{E}_\ell[X_\ell^2] + \frac{C \log(t/\delta')}{\eta} \\ &\overset{(i)}{\leq} 2\eta B^2 \sum_{\ell=1}^{t} \mathbb{E}_\ell[Z_\ell] + \frac{C \log(t/\delta')}{\eta}\end{aligned}$$

[2]We will use this notation to denote conditional expectations throughout this work.

for all $t \in \mathbb{N}$ with probability at least $1 - \delta'$. Inequality (i) holds because $\mathbb{E}_t[X_t^2] \leq B^2\mathbb{E}[|X_t|] \leq 2B^2\mathbb{E}_t[Z_t]$ for all $t \in \mathbb{N}$. Setting $\eta = \frac{\beta}{2B^2}$ and substituting $\sum_{\ell=1}^t X_\ell = \sum_{\ell=1}^t Z_\ell - \mathbb{E}_\ell[Z_\ell]$,

$$\sum_{\ell=1}^t Z_\ell \leq (1+\beta)\sum_{\ell=1}^t \mathbb{E}_\ell[Z_\ell] + \frac{2B^2 C\log(t/\delta')}{\beta} \tag{14}$$

with probability at least $1 - \delta'$. Now consider the martingale difference sequence $X'_t = \mathbb{E}[Z_t] - Z_t$ and notice that $|X'_t| \leq B^2$. Using the inequality of Lemma A.2 we obtain for all $\eta \in (0, 1/B^2)$,

$$\begin{aligned}\sum_{\ell=1}^t X'_\ell &\leq \eta \sum_{\ell=1}^t \mathbb{E}_\ell[(X'_\ell)^2] + \frac{C\log(t/\delta')}{\eta} \\ &\leq 2\eta B^2 \sum_{\ell=1}^t \mathbb{E}_\ell[Z_\ell] + \frac{C\log(t/\delta')}{\eta}\end{aligned}$$

Setting$\eta = \frac{\beta}{2B^2}$ and substituting $\sum_{\ell=1}^t X'_\ell = \sum_{\ell=1}^t \mathbb{E}[Z_\ell] - Z_\ell$ we have,

$$(1-\beta)\sum_{\ell=1}^t \mathbb{E}[Z_\ell] \leq \sum_{\ell=1}^t Z_\ell + \frac{2B^2 C\log(t/\delta')}{\beta} \tag{15}$$

with probability at least $1 - \delta'$. Combining Equations 14 and 15 and using a union bound yields the desired result.

□

Proposition A.3 can be used to show,

Let the Z_ℓ be i.i.d. samples $Z_\ell \overset{i.i.d.}{\sim} \mathrm{Ber}(p)$. The empirical mean estimator, $\widehat{p}_t = \frac{1}{t}\sum_{\ell=1}^t Z_\ell$ satisfies,

$$(1-\beta)p - \frac{2C'\log(t/\delta')}{\beta t} \leq \widehat{p}_t \leq (1+\beta)p + \frac{2C'\log(t/\delta')}{\beta t}$$

with probability at least $1 - 2\delta'$ for all $t \in \mathbb{N}$ where $C' > 0$ is a (known) universal constant. Given $\epsilon > 0$ set $t \geq \frac{8C'\log(t/\delta')}{\beta\epsilon}$ (notice the dependence of t on the RHS - this can be achieved by setting $t \geq \frac{C\log(C/\beta\epsilon\delta')}{\beta\epsilon}$ for some (known) universal constant $C > 0$).

In this case observe that,

$$(1-\beta)p - \epsilon/8 \leq \widehat{p}_t \leq (1+\beta)p + \epsilon/8$$

Setting $\beta = 1/8$,

$$7p/8 - \epsilon/8 \leq \widehat{p}_t \leq 9p/8 + \epsilon/8$$

so that,

$$p - \widehat{p}_t \leq p/8 + \epsilon/8.$$

and

$$\widehat{p}_t - p \leq p/8 + \epsilon/8.$$

and therefore $|\widehat{p}_t - p| \leq p/8 + \epsilon/8 \leq 2\max(p/8, \epsilon/8) = \max(p/4, \epsilon/4)$.

A.2 Proof of the optimal sampling distribution (4)

Our performance estimator is,

$$\hat{V}_1^{\pi^k} = \frac{1}{n}\sum_{i=1}^n \sum_{h=1}^H \frac{d_h^{\pi^k}(s_h^i, a_h^i)}{\mu_h(s_h^i, a_h^i)} r(s_h^i, a_h^i),\ k \in [K].$$

Denote $\sum_{h=1}^{H} \frac{d_h^{\pi^k}(s_h^i,a_h^i)}{\mu_h(s_h^i,a_h^i)} r_h(s_h^i, a_h^i)$ by X_i. And for simplicity, denote $\mathbb{E}_{(s_1,a_1)\sim\mu_1,\ldots,(s_H,a_H)\sim\mu_H}$ by $\mathbb{E}_\mu$, the variance of our estimator is bounded by,

$$\begin{aligned}
\mathbb{E}_\mu[X_i^2] &= \mathbb{E}_\mu\left[\left(\sum_{h=1}^{H} \frac{d_h^{\pi^k}(s_h^i,a_h^i)}{\mu_h(s_h^i,a_h^i)} r_h(s_h^i,a_h^i)\right)^2\right] \\
&\le \mathbb{E}_\mu\left[H\cdot\sum_{h=1}^{H}\left(\frac{d_h^{\pi^k}(s_h^i,a_h^i)}{\mu_h(s_h^i,a_h^i)} r_h(s_h^i,a_h^i)\right)^2\right] \\
&\le \mathbb{E}_\mu\left[H\cdot\sum_{h=1}^{H}\left(\frac{d_h^{\pi^k}(s_h^i,a_h^i)}{\mu_h(s_h^i,a_h^i)}\right)^2\right] \\
&= H\cdot\sum_{h=1}^{H}\mathbb{E}_{d_h^{\pi^k}}\left[\frac{d_h^{\pi^k}(s_h^i,a_h^i)}{\mu_h(s_h^i,a_h^i)}\right].
\end{aligned}$$

The first inequality holds by $\mathrm{Cauchy-Schwarz}$ inequality. The second inequality holds due to the assumption $r_h(s,a)\in[0,1]$.

Denote $\sum_{h=1}^{H}\mathbb{E}_{d_h^{\pi^k}}\left[\frac{d_h^{\pi^k}(s_h^i,a_h^i)}{\mu_h(s_h^i,a_h^i)}\right]$ by $\rho_{\mu,k}$. Applying $\mathrm{Bernstein's}$ inequality, we have that with probability at least $1-\delta$ and n samples, it holds,

$$|\hat{V}_1^{\pi^k} - V_1^{\pi^k}| \le \sqrt{\frac{2H\rho_{\mu,k}\log(1/\delta)}{n}} + \frac{2M_k\log(1/\delta)}{3n}$$

where $M_k = \max_{s_1,a_1,\ldots,s_H,a_H}\sum_{h=1}^{H}\frac{d_h^{\pi^k}(s_h,a_h)}{\mu_h(s_h,a_h)} r_h(s_h,a_h)$.

To achieve an ϵ accuracy of evaluation, we need samples,

$$n_{\mu,k} \le \frac{8H\rho_{\mu,k}\log(1/\delta)}{\epsilon^2} + \frac{4M_k\log(1/\delta)}{3\epsilon}$$

Take the union bound over all target policies,

$$n_\mu \le \frac{8H\max_{k\in[K]}\rho_{\mu,k}\log(K/\delta)}{\epsilon^2} + \frac{4M\log(K/\delta)}{3\epsilon}$$

where $M = \max_{k\in[K]} M_k$.

We define the optimal sampling distribution μ^* as the one minimizing the higher order sample complexity,

$$\begin{aligned}
\mu_h^* &= \arg\min_{\mu_h}\max_{k\in[K]}\mathbb{E}_{d_h^{\pi^k}(s,a)}\left[\frac{d_h^{\pi^k}(s,a)}{\mu_h(s,a)}\right] \\
&= \arg\min_{\mu_h}\max_{k\in[K]}\sum_{s,a}\frac{\left(d_h^{\pi^k}(s,a)\right)^2}{\mu_h(s,a)},\ h=1,\ldots,H.
\end{aligned}$$

A.3 An example of unrealizable optimal sampling distribution

Here, we give an example to illustrate the assertion that in some cases, the optimal sampling distribution cannot be realized by any policy.

Consider such a MDP with two layers, in the first layer, there is a single initial state $s_{1,1}$, in the second layer, there are two states $s_{2,1}, s_{2,2}$. The transition function at state $s_{1,1}$ is identical for any action, $\mathbb{P}(s_{2,1}|s_{1,1},a) = \mathbb{P}(s_{2,2}|s_{1,1},a) = \frac{1}{2}$. Hence, for any policy, the only realizable state visitation distribution at the second layer is $d_2(s_{2,1}) = d_2(s_{2,2}) = \frac{1}{2}$.

Suppose the target policies take $K\ge 2$ different actions at state $s_{2,1}$ while take the same action at state $s_{2,2}$.

By solving the optimization problem (4), we have the optimal sampling distribution at the second layer,

$$\mu_2^*(s_{2,1}) = \frac{K^2}{1+K^2},\ \mu_2^*(s_{2,2}) = \frac{1}{1+K^2},$$

which is clearly not realizable by any policy.

A.4 Proof of Lemma 4.5

Proof. The gradient of $\ell_h^\pi(w)$,

$$\nabla_{w(s,a)}\ell_h^\pi(w) = \frac{\tilde{\mu}_h(s,a)}{\hat{\mu}_h(s,a)}w(s,a) - \sum_{s',a'}\tilde{\mu}_{h-1}(s',a')P(s|s',a')\pi(a|s)\frac{\hat{w}_{h-1}(s',a')}{\hat{\mu}_{h-1}(s',a')}$$

Suppose by some SGD algorithm, we can converge to a point $\hat{w}_h$ such that the gradient of the loss function is less than ϵ,

$$\|\nabla\ell_h^\pi(\hat{w}_h)\|_1 = \sum_{s,a}\left|\frac{\tilde{\mu}_h(s,a)}{\hat{\mu}_h(s,a)}\hat{w}_h(s,a) - \sum_{s',a'}\tilde{\mu}_{h-1}(s',a')P(s|s',a')\pi(a|s)\frac{\hat{w}_{h-1}(s',a')}{\hat{\mu}_{h-1}(s',a')}\right| \le \epsilon.$$

By decomposing,

$$\begin{aligned}
&\left|\frac{\tilde{\mu}_h(s,a)}{\hat{\mu}_h(s,a)}\hat{w}_h(s,a) - \sum_{s',a'}\tilde{\mu}_{h-1}(s',a')P(s|s',a')\pi(a|s)\frac{\hat{w}_{h-1}(s',a')}{\hat{\mu}_{h-1}(s',a')}\right| \\
=&\left|\frac{\tilde{\mu}_h(s,a)}{\hat{\mu}_h(s,a)}\hat{w}_h(s,a) - d_h^\pi(s,a) + d_h^\pi(s,a) - \sum_{s',a'}\tilde{\mu}_{h-1}(s',a')P(s|s',a')\pi(a|s)\frac{\hat{w}_{h-1}(s',a')}{\hat{\mu}_{h-1}(s',a')}\right| \\
\ge&\left|\frac{\tilde{\mu}_h(s,a)}{\hat{\mu}_h(s,a)}\hat{w}_h(s,a) - d_h^\pi(s,a)\right| - \left|d_h^\pi(s,a) - \sum_{s',a'}\tilde{\mu}_{h-1}(s',a')P(s|s',a')\pi(a|s)\frac{\hat{w}_{h-1}(s',a')}{\hat{\mu}_{h-1}(s',a')}\right| \\
=&\left|\tilde{\mu}_h(s,a)\frac{\hat{w}_h(s,a)}{\hat{\mu}_h(s,a)} - d_h^\pi(s,a)\right| - \left|\sum_{s',a'}P(s|s',a')\pi(a|s)\left(d_{h-1}^\pi(s',a') - \tilde{\mu}_{h-1}(s',a')\frac{\hat{w}_{h-1}(s',a')}{\hat{\mu}_{h-1}(s',a')}\right)\right|
\end{aligned}$$

Hence, we have,

$$\begin{aligned}
\sum_{s,a}\left|\tilde{\mu}_h(s,a)\frac{\hat{w}_h(s,a)}{\hat{\mu}_h(s,a)} - d_h^\pi(s,a)\right| &\le \epsilon + \sum_{s,a}\left|\sum_{s',a'}P(s|s',a')\pi(a|s)\left(d_{h-1}^\pi(s',a') - \tilde{\mu}_{h-1}(s',a')\frac{\hat{w}_{h-1}(s',a')}{\hat{\mu}_{h-1}(s',a')}\right)\right| \\
&\le \epsilon + \sum_{s',a'}\left|d_{h-1}^\pi(s',a') - \tilde{\mu}_{h-1}(s',a')\frac{\hat{w}_{h-1}(s',a')}{\hat{\mu}_{h-1}(s',a')}\right| \\
&\le 2\epsilon
\end{aligned}$$

□

A.5 Proof of Lemma 4.7

Proof. The minimum w_h^* of the loss function $\ell_h^\pi(w)$ is $w_h^*(s,a) = \frac{d_h^\pi(s,a)}{\tilde{\mu}_h(s,a)}\hat{\mu}_h(s,a)$ if $\hat{w}_{h-1}$ achieves optimum. By the property of the coarse distribution estimator, we have,

$$w_h^*(s,a) = \frac{d_h^\pi(s,a)}{\tilde{\mu}_h(s,a)}\hat{\mu}_h(s,a) \le \frac{\frac{4}{3}\hat{d}_h^\pi(s,a)}{\frac{4}{5}\hat{\mu}_h(s,a)}\hat{\mu}_h(s,a) = \frac{5}{3}\hat{d}_h^\pi(s,a)$$

We can define a feasible set for the optimization problem, i.e. $w_h(s,a) \in [0, D_h(s,a)]$, $D_h(s,a) = 2\hat{d}_h^\pi(s,a)$.

Next, we analyse the variance of the stochastic gradient. We denote the stochastic gradient as $g_h(w)$, $\{s_1^i, a_1^i, \ldots, s_H^i, a_H^i\}$ a trajectory sampled from $\tilde{\mu}_h$ and $\{s_1^j, a_1^j, \ldots, s_H^j, a_H^j\}$ a trajectory sampled from $\tilde{\mu}_{h-1}$.

$$g_h(w)(s,a) = \frac{w(s,a)}{\hat{\mu}_h(s,a)}\mathbb{I}(s_h^i = s, a_h^i = a) - \frac{\hat{w}_{h-1}(s_{h-1}^j, a_{h-1}^j)}{\hat{\mu}_{h-1}(s_{h-1}^j, a_{h-1}^j)}\pi(a|s)\mathbb{I}(s_h^j = s)$$

the variance bound is,

$$\mathbb{V}[g_h(w)] \le \mathbb{E}[\|g_h(w)\|^2] \le \sum_{s,a} \tilde{\mu}_h(s,a) \left(\frac{w(s,a)}{\hat{\mu}_h(s,a)}\right)^2 + \tilde{\mu}_{h-1}(s,a) \left(\frac{\hat{w}_{h-1}(s,a)}{\hat{\mu}_{h-1}(s,a)}\right)^2$$

$$\le O\left(\sum_{s,a} \frac{(\hat{d}_h^\pi(s,a))^2}{\hat{\mu}_h(s,a)} + \frac{(\hat{d}_{h-1}^\pi(s,a))^2}{\hat{\mu}_{h-1}(s,a)}\right) \tag{16}$$

the last inequality is due to the bounded feasible set for w and the property of coarse distribution estimator $\tilde{\mu}_h(s,a) \le \frac{4}{3}\hat{\mu}_h(s,a)$.

Based on the error propagation lemma 4.5, if we can achieve $\|\nabla \ell_h^\pi(\hat{w}_h)\|_1 \le \frac{\epsilon}{4H^2}$ from step $h = 1$ to step $h = H$, then we have,

$$\sum_{s,a} \left| \tilde{\mu}_h(s,a) \frac{\hat{w}_h(s,a)}{\hat{\mu}_h(s,a)} - d_h^\pi(s,a) \right| \le \frac{\epsilon}{4H}, \forall h = 1, 2, \dots, H$$

which can enable us to build the final estimator of the performance of policy π with at most error ϵ.

By the property of smoothness, to achieve $\|\nabla \ell_h^\pi(\hat{w}_h)\|_1 \le \frac{\epsilon}{4H^2}$, we need to achieve $\ell_h^\pi(\hat{w}_h) - \ell_h^\pi(w_h^*) \le \frac{\epsilon^2}{32\xi H^4}$ where ξ is the smoothness factor, because,

$$\|\nabla \ell_h^\pi(\hat{w}_h)\|_1^2 \le 2\xi(\ell_h^\pi(\hat{w}_h) - \ell_h^\pi(w_h^*)) \le \frac{\epsilon^2}{16H^4}.$$

Invoke the convergence rate for strongly-convex and smooth loss functions, i.e. Lemma 4.6, we have that the number of samples needed to achieve $\ell_h^\pi(\hat{w}_h) - \ell_h^\pi(w_h^*) \le \frac{\epsilon^2}{32\xi H^4}$ is,

$$n = O\left(\frac{\xi}{\gamma} \frac{H^4 G^2}{\epsilon^2}\right)$$

We have shown in (8) that $\frac{\xi}{\gamma} \le \frac{5}{3}$, this nice property helps us to get rid of the undesired ratio of the smoothness factor and the strongly-convexity factor, i.e. $\frac{\max_{s,a} \mu(s,a)}{\min_{s,a} \mu(s,a)}$ of the original loss function (7) which can be extremely bad. Replacing G^2 by our variance bound (16), we have,

$$n_h^\pi = O\left(\frac{H^4}{\epsilon^2} \left(\sum_{s,a} \frac{(\hat{d}_h^\pi(s,a))^2}{\hat{\mu}_h(s,a)} + \frac{(\hat{d}_{h-1}^\pi(s,a))^2}{\hat{\mu}_{h-1}(s,a)}\right)\right)$$

For each step h, we need the above number of trajectories, sum over h, we have the total sample complexity,

$$n^\pi = O\left(\frac{H^4}{\epsilon^2} \sum_{h=1}^{H} \sum_{s,a} \frac{(\hat{d}_h^\pi(s,a))^2}{\hat{\mu}_h(s,a)}\right)$$

To evaluate K policies, we need trajectories,

$$n = O\left(\frac{H^4}{\epsilon^2} \sum_{h=1}^{H} \max_{k \in [K]} \sum_{s,a} \frac{(\hat{d}_h^{\pi^k}(s,a))^2}{\hat{\mu}_h(s,a)}\right).$$

□

A.6 Proof of Lemma 4.9

Proof. By Markov's inequality, we have,

$$\mathbb{P}(|\hat{\mu} - \mu| \ge \epsilon) \le \frac{\mathbb{E}[|\hat{\mu} - \mu|]}{\epsilon} \le \frac{1}{4}.$$

The event that $|\hat{\mu}_{MoM} - \mu| > \epsilon$ belongs to the event where more than half estimators $\hat{\mu}_i$ are outside of the desired range $|\hat{\mu}_i - \mu| > \epsilon$, hence, we have,

$$\mathbb{P}(|\hat{\mu}_{MoM} - \mu| > \epsilon) \le \mathbb{P}(\sum_{i=1}^{N} \mathbb{I}(|\hat{\mu}_i - \mu| > \epsilon) \ge \frac{N}{2})$$

Denote $\mathbb{I}(|\hat{\mu}_i - \mu| > \epsilon)$ by Z_i and $\mathbb{E}[Z_i] = p$,

$$\begin{aligned}
\mathbb{P}(|\hat{\mu}_{MoM} - \mu| > \epsilon) &= \mathbb{P}(\sum_{i=1}^{N} Z_i \geq \frac{N}{2}) \\
&= \mathbb{P}(\frac{1}{N}\sum_{i=1}^{N}(Z_i - p) \geq \frac{1}{2} - p) \\
&\leq e^{-2N(\frac{1}{2}-p)^2} \\
&\leq e^{-\frac{N}{8}}
\end{aligned}$$

the first inequality holds by Hoeffding's inequality and the second inequality holds due to $p \leq \frac{1}{4}$. Set $\delta = e^{-\frac{N}{8}}$, we have, with $N = O(\log(1/\delta))$, with probability at least $1 - \delta$, it holds $|\hat{\mu}_{MoM} - \mu| \leq \epsilon$. □