On the Sample Complexity of Differentially Private Policy Gradient

Yi He [1] **Xingyu Zhou** [1]

Abstract

Policy Optimization (PO) is a cornerstone of modern reinforcement learning (RL), with applications ranging from robotics and healthcare to training large language models. However, the growing use of PO in sensitive domains raises pressing privacy concerns. In this paper, we initiate the study of differentially private policy optimization (PO). We begin by defining a suitable notion of differential privacy tailored to PO, addressing challenges unique to its on-policy learning dynamics and the definition of the privacy unit. Focusing on policy gradient (PG) with the REINFORCE estimator, we propose a differentially private variant and analyze its sample complexity. Our results establish bounds for both first-order stationary point (FOSP) convergence and global optimality, showing that privacy can be achieved with provably lower-order overhead.

1. Introduction

Policy Optimization (PO) is one of the most widely used methods in Reinforcement Learning (RL), with applications spanning games (Silver et al., 2016), robotics (Levine & Koltun, 2013), healthcare (Yu et al., 2021), and, more recently, the training of large language models (Ouyang et al., 2022; Guo et al., 2025). Unlike value-based RL, PO directly optimizes the policy and encompasses a variety of algorithms such as vanilla policy gradient (PG) (Williams, 1992; Sutton et al., 1999), natural policy gradient (NPG) (Kakade, 2001), trust region policy optimization (TRPO) (Schulman et al., 2015), proximal policy optimization (PPO) (Schulman et al., 2017) and more recently, group relative policy optimization (GRPO) (Shao et al., 2024). Due to its popularity, there is a rich literature that provides various theoretical understandings of different PO methods (e.g., computational efficiency or sample complexity) (Agarwal et al., 2021b; Yuan et al., 2022; Shani et al., 2020; Liu et al., 2019).

As PO becomes increasingly prevalent in real-world applications, privacy concerns are emerging as a critical challenge. For instance, in personalized medical care, patient interactions—where the state represents medical history, the action corresponds to prescribed medication, and the reward reflects treatment effectiveness—constitute sensitive data that must be protected. Similarly, in RL-based training of large language models, user prompts may contain private information that requires protection. Addressing these privacy concerns is essential for ensuring the responsible deployment of PO methods in sensitive domains.

In this paper, we initiate the study of differentially private policy optimization, making the following key contributions. First, we formally define a suitable notion of differential privacy (DP) (Dwork et al., 2006) for PO, highlighting its distinctions from the standard DP definitions used in supervised learning. These differences stem from the unique learning dynamics and the notion of the privacy unit in PO. Second, as an initial step, we focus on the most basic yet fundamental PO method—policy gradient with REINFORCE—and develop a differentially private variant. Beyond providing formal privacy guarantees, we also establish sample complexity bounds for this method, analyzing both convergence to a first-order stationary point (FOSP) and global optimality. Notably, all of our sample complexity bounds consist of two components: the leading terms match the standard non-private results in Yuan et al. (2022), while the privacy cost appears as lower-order additive terms.

2. Preliminaries

Policy optimization (PO) in bandit. In this work, instead of considering a general Markov decision process (MDP), we focus on the simpler bandit formulation, which allows us to easily demonstrate the key ideas. We note that generalizing it to MDP is standard, as done in the literature (Yuan et al., 2022). This bandit formulation already captures many interesting real-world applications, such as personalized medical care (Zhou et al., 2023) and alignment/reasoning training in large language models (LLMs) (Ouyang et al., 2022). In particular, given an initial state $s \in \mathcal{S}$ (e.g., a medical status or a prompt in LLMs) sampled from a distribution ρ, an action $a \in \mathcal{A}$ (e.g., a medical prescription or a response in LLMs) is generated according to a policy π and a reward $R(s,a) \in [-R_{\max}, R_{\max}]$ is observed. In policy optimization, we parameterize the policy π by π_θ with $\theta \in \Theta = \mathbb{R}^d$ (e.g., a neural network), and the goal is to

[1]Wayne State University, USA.

leverage interactions (sample trajectories) to find an optimal policy that maximizes the following objective:

$$J(\theta) \stackrel{\text{def}}{=} \mathbb{E}_{s\sim\rho, a\sim\pi_\theta(\cdot|s)}\left[R(s,a)\right].$$

Vanilla policy gradient (PG). One simple and direct approach to solving the above policy optimization problem is via vanilla policy gradient, given by

$$\theta_{t+1} = \theta_t + \eta \nabla J(\theta_t),$$

where $\eta > 0$ is some learning rate, $\nabla J(\theta_t)$ is the gradient at step t, and θ_1 is some initial value. The gradient can be written as follows by the classic policy gradient theorem

$$\nabla J(\theta) = \mathbb{E}_{s\sim\rho, a\sim\pi_\theta(\cdot|s)}\left[R(s,a)\nabla_\theta \log \pi_\theta(a|s)\right].$$

However, in practice, it is often hard to compute the full gradient above due to both statistical (e.g., without knowing ρ) and computational (averaging over all possible trajectories) issues. Instead, we will build an unbiased estimator of it using samples. One classic gradient estimator is the REINFORCE (Williams, 1992), given by

$$\widehat{\nabla}_m J(\theta) := \frac{1}{m}\sum_{i=1}^{m} R(s_i,a_i)\nabla_\theta \log \pi_\theta(a_i|s_i), \quad (1)$$

where $m > 0$ is the batch size for a batch of i.i.d on-policy samples $\{(s_i,a_i)\}_{i=1}^m$, where $s_i \sim \rho$ and $a_i \sim \pi_\theta(\cdot|s_i)$. With this gradient estimator used in every step $t \in [T]$, the overall sample complexity is given by $N = Tm$.

Standard sample complexity. In the standard non-private case, previous work has established various sample complexity bounds for PG with REINFORCE (Yuan et al., 2022; Liu et al., 2020; Zhang et al., 2021). The typical result is that for an accuracy of $\alpha > 0$, for either first-order stationary point of $J(\theta)$ or the global optimum, the required total sample is on the order of $N = O\left(\frac{1}{\alpha^k}\right)$ for some $k \in \mathbb{R}$, depending on the specific scenarios, see Yuan et al. (2022).

In this paper, our goal is to formally introduce differential privacy (DP) into the problem of policy optimization and derive the sample complexity bounds under privacy.

3. Differential Privacy in Policy Optimization

We first recall the standard DP definition with a fixed dataset.

Definition 1 (Dwork et al. (2006))**.** A randomized mechanism $\mathcal{M}$ satisfies (ε,δ)-DP if for adjacent datasets D, D' differing by one record, and $\forall S \subseteq \text{Range}(\mathcal{M})$:

$$\mathbb{P}[\mathcal{M}(D) \in S] \leqslant e^\varepsilon \cdot \mathbb{P}[\mathcal{M}(D') \in S] + \delta.$$

This standard DP notion can be directly used in supervised learning problems with D being a set of i.i.d samples $\{(x_i,y_i)\}_{i=1}^N$ from an unknown distribution and $\mathcal{M}(D)$ being the final policy. This has been utilized in private empirical risk minimization (ERM) (Chaudhuri et al., 2011; Bassily et al., 2014) as well as private stochastic optimization (both convex and non-convex), e.g., Bassily et al. (2019). For example, the optimal excess population loss for stochastic convex optimization is $O_\delta\left(\sqrt{\frac{1}{N}} + \frac{\sqrt{d}}{N\varepsilon}\right)$ for (ε,δ)-DP.

One may attempt to adopt the above notion to PO with the dataset D being $\{(s_i,a_i)\}_{i=1}^N$ and $\mathcal{M}(D)$ being the final policy. However, this does not make too much sense because (i) there is no such a fixed dataset in policy gradient as the actions are generated in the on-policy fashion, i.e., using the most recent policy; (ii) the neighboring relation of differing in one sample (s_i,a_i) (i.e., privacy unit) actually does not hold as changing one sample will lead to difference in all future samples due to different policy onward. Thus, we need a new definition that can address the above two issues.

To this end, we borrow the idea from private online bandit and RL literature (Vietri et al., 2020; Chowdhury & Zhou, 2022), which essentially considers a set of "users" as the dataset. For instance, the dataset could be N unique patients interacting with the learning agent, and each user has an initial state (e.g., medical status), which is distributed according to ρ. We can fix the "users" in advance (or arrive online) and the privacy unit is now for each patient, hence resolving both issues above. Moreover, the set of "users" can also represent N prompts in the training of LLMs, with each "user" contributing one prompt. Note that although we use "users" to align with personalization application, this is still an item-level DP, as each "user" appears only once (as a patient or prompt). The learning agent can interact with each user to observe (s,a) and $R(s,a)$ *dynamically*. With the above notion of dataset, the privacy protection in PO is that changing one "user" in the dataset will not change the final policy too much, leading to the following definition.

Definition 2 (DP in PO)**.** Consider any policy optimization algorithm $\mathcal{M}$ interacting with a set D of N "users" and $\mathcal{M}(D)$ being the final output policy. We say $\mathcal{M}$ is (ε,δ)-DP if for adjacent datasets D, D' differing by one "user", and $\forall S \subseteq \text{Range}(\mathcal{M})$:

$$\mathbb{P}[\mathcal{M}(D) \in S] \leqslant e^\varepsilon \cdot \mathbb{P}[\mathcal{M}(D') \in S] + \delta.$$

Remark 1. We emphasize once again that the above DP notion is defined for PO in RL, analogous to the standard DP in statistical learning (e.g., supervised learning). Vanilla policy gradient is merely one specific method, just as (stochastic) gradient descent is a particular method for stochastic optimization. In this paper, we aim to design a private PG method and analyze its sample complexity. In the future, as the next step, one can also consider designing private versions of other PO methods, such as natural policy gradient (NPG) and proximal policy optimization (PPO).

Algorithm 1 Differentially Private Policy Gradient (DP-PG)

input Number of "users" N, batch size m, privacy parameters ε, δ, learing rate η

1: Initialize policy parameter $\theta_1 \in \Theta$ and set $T = N/m$
2: **for** $t = 1$ to T **do**
3: Compute gradient estimator $\widehat{\nabla}_m J(\theta_t)$ using the t-th fresh batch of "users" via (1)
4: Add noise $\widetilde{\nabla}_m J(\theta_t) = \widehat{\nabla}_m J(\theta_t) + \mathcal{N}(0, \sigma^2 I_d)$
5: Update policy parameter $\theta_{t+1} = \theta_t + \eta \widetilde{\nabla}_m J(\theta_t)$
6: **end for**

4. Differentially Private Policy Gradient

In this section, we present a private version of policy gradient with the REINFORCE estimator. At a high level, it is a one-pass mini-batch stochastic gradient ascent with additional Gaussian noise for privacy protection.

Our algorithm named DP-PG is given in Algorithm 1. It will run T-step update with $T = N/m$ due to the one-pass algorithm. For each step t, we first leverage a *fresh* batch of "users" to construct an *unbiased* estimator $\widehat{\nabla}_m J(\theta)$. Then, a Gaussian noise with variance σ^2 at each dimension is added, where σ^2 depends on the privacy parameters.

Remark 2. The main reason for one-pass here is to ensure that $\widehat{\nabla}_m J(\theta_t)$ is an unbiased estimator of the true gradient, similar to one-pass SGD for stochastic optimization. This also leads to a simpler privacy analysis. Our DP-PG can also be used in the online setting where a stream of "users" arrive sequentially, as in standard online RL/bandits.

Theorem 1 (Privacy guarantee)**.** *Assume for any $s \in \mathcal{S}$ and $\theta \in \Theta$, there exists a constant G such that $\|\nabla_\theta \log \pi_\theta(a \mid s)\| \leqslant G$. Then, setting $\sigma^2 = \frac{8 \ln(1.25/\delta) R_{\max}^2 G^2}{m^2 \varepsilon^2}$ in Algorithm 1 ensures (ε, δ)-DP as in Definition 2.*

The above result directly follows from the privacy guarantee of the Gaussian mechanism and (adaptive) parallel composition due to our one-pass algorithm. The assumption of G is satisfied by softmax policy as well as Gaussian policy. In fact, they satisfy an even stronger condition in Assumption 1, as will be discussed shortly.

Remark 3. By the so-called *billboard lemma* (Hsu et al., 2016), our DP-PG also satisfies the commonly used *joint differential privacy* (JDP) in private online RL/bandits (Vietri et al., 2020; Shariff & Sheffet, 2018; Chowdhury & Zhou, 2022; Zhou, 2022). Roughly speaking, JDP guarantees that changing one "user" (say u) will not change all the actions prescribed to all other "users" except u.

5. Sample Complexity under Privacy

In this section, we aim to establish the sample complexity bounds of our DP-PG for both first-order stationary point (FOSP) and global optimum convergence.

5.1. First-order Stationary Point Convergence

We start with the sample complexity for FOSP convergence. This result is not only of its own importance, but will also be useful for our later result on the global optimum convergence. In particular, we will consider the following general class of policies, which is widely studied in previous non-private work and also includes commonly used policies such as softmax and Gaussian policy (Yuan et al., 2022).

Assumption 1 (Lipschitz Smoothness (LS))**.** There exist constants $G, F > 0$ such that for every state $s \in \mathcal{S}$, the gradient and Hessian of $\log \pi_\theta(\cdot \mid s)$ of any $\theta \in \Theta$ satisfy

$$\|\nabla_\theta \log \pi_\theta(a \mid s)\| \leqslant G \text{ and } \|\nabla_\theta^2 \log \pi_\theta(a \mid s)\| \leqslant F.$$

Remark 4. For simplicity, as in previous work, we will often view G and F as constants $\Theta(1)$, hence omitted in the sample complexity bound.

Theorem 2 (FOSP convergence)**.** *Under the same setting of Theorem 1 and Assumption 1, there exists a proper parameter choices of m and η, such that*

$$\mathbb{E}\left[\|\nabla J(\theta_U)\|^2\right] \leqslant O\left(\frac{1}{\sqrt{N}} + \left(\frac{\sqrt{d}}{N\varepsilon}\right)^{2/3}\right), \quad (2)$$

where θ_U is uniformly sampled from $\{\theta_1, \ldots, \theta_T\}$.

Remark 5. Several remarks are in order. First, we can see that the first term in (2) matches the previous non-private term, i.e., for an accuracy of α, the sample complexity is $O(1/\alpha^4)$ (Yuan et al., 2022); Second, the privacy cost is a lower order additive term (for constant ε and d), i.e., the additional sample complexity due to privacy is $O_\delta\left(\frac{\sqrt{d}}{\alpha^3 \varepsilon}\right)$.

5.2. Global Optimum Convergence

We now turn our focus to the global optimum convergence in the sense of average regret, i.e., $J^* - \frac{1}{T}\sum_{t=1}^T \mathbb{E}\left[J(\theta_t)\right]$. Following the non-private work (Yuan et al., 2022), we will also consider two different scenarios and aim to establish the corresponding sample complexities in the private case.

5.2.1. Fisher-non-degenerate Parameterization

In the first scenario, in addition to Assumption 1, we further assume the following two conditions on the policy class, both of which are commonly used in the non-private case.

The first condition is the so-called *Fisher-non-degenerate policy*, formally defined below.

Assumption 2 (Fisher-non-degenerate, adapted from Assumption 2.1 of Ding et al. (2022))**.** For all $\theta \in \mathbb{R}^d$, there exists $\mu > 0$ s.t. the Fisher information matrix $F_\rho(\theta)$ induced by policy π_θ and initial state distribution ρ satisfies

$$F_\rho(\theta) \stackrel{\text{def}}{=} \mathbb{E}_{s\sim\rho, a\sim\pi_\theta(\cdot|s)}\left[\nabla_\theta \log \pi_\theta(a|s) \nabla_\theta \log \pi_\theta(a|s)^\top\right] \geqslant \mu \mathbf{I}_d.$$

This assumption is commonly used in the literature on non-private PG methods (Yuan et al., 2022; Ding et al., 2022; Agarwal et al., 2021a; Liu et al., 2022). As shown in Sec B.2 in Ding et al. (2022), this assumption is satisfied by the Gassuian policy and even certain neural policies.

The next condition is the so-called *compatible function approximation*, which is also a common assumption in the PG literature to handle function approximation error in the non-tabular case.

Assumption 3 (Compatible, adapted from Assumption 4.6 in Ding et al. (2022))**.** For all $\theta \in \mathbb{R}^d$, there exists $\alpha_{\text{bias}} > 0$ such that the *transferred compatible function approximation error* satisfies

$$\mathbb{E}_{s\sim\rho,a\sim\pi_{\theta^*}(\cdot|s)}\left[(A^{\pi_\theta}(s,a) - u^{*\top}\nabla_\theta \log \pi_\theta(a|s))^2\right] \leqslant \alpha_{\text{bias}},$$

where π_{θ^*} is an optimal policy and $u^* = (F_\rho(\theta))^\dagger \nabla J(\theta)$.

Remark 6. The intuition behind "compatible" here is that we are approximating the advantage function $A^{\pi_\theta}(s,a)$ using the $\nabla_\theta \log \pi_\theta(a|s)$ as the feature vector; The "transfer error" here means that we are shifting to the expectation in terms of an optimal policy (rather than the current policy). The approximation error α_{bias} is zero for a softmax tabular policy, and the error is small when π_θ is a rich neural policy. (Ding et al., 2022; Liu et al., 2022; Wang et al., 2019).

With the above two additional assumptions along with the LS assumption in Assumption 1, we have the following important result, i.e., the objective $J(\theta)$ satisfies the so-called *relaxed weak gradient domination*.

Lemma 1 (Lemma 4.7 in Ding et al. (2022))**.** *If the policy π_θ satisfies Assumptions 1, 2 and 3, then*

$$J^* - J(\theta) \leqslant \frac{G}{\mu}\|\nabla J(\theta)\| + \sqrt{\alpha_{\text{bias}}}.$$

This lemma essentially allows us to easily translate a guarantee in terms of FOSP to a certain global optimum convergence. This leads to our next main result with its proof given in Appendix B.

Theorem 3. *Consider the same setting of Theorem 2 and further let Assumptions 2 and 3 hold. Then, for any $\alpha > 0$, Algorithm 1 enjoys the following average regret guarantee*

$$J^* - \frac{1}{T}\sum_{t=1}^{T}\mathbb{E}\left[J(\theta_t)\right] \leqslant O(\alpha) + O\left(\sqrt{\alpha_{\text{bias}}}\right),$$

when the sample size satisfies $N \geqslant O\left(\frac{1}{\alpha^4\mu^4} + \frac{\sqrt{d}}{\alpha^3\mu^3\varepsilon}\right)$.

Remark 7. In the above bound, we explicitly include the parameter μ to clearly illustrate its impact. The first term $O\left(\frac{1}{\alpha^4\mu^4}\right)$ matches the non-private one in Yuan et al. (2022) while the second term is the privacy cost. As we can see, for both terms, there exists an additional $1/\mu$ factor compared to the sample complexity of FOSP. This indicates that for very small but still positive μ, our bound could be large.

5.2.2. Tabular Softmax with Log-barrier Regularization

In this section, we move to the second scenario for our study of global convergence where we consider the tabular case with the classic softmax policy:

Definition 3 (Tabular softmax policy)**.** Consider a finite state space $\mathcal{S}$ and action space $\mathcal{A}$. For any state-action pair $(s,a) \in \mathcal{S}\times\mathcal{A}$, the softmax policy is given by

$$\pi_\theta(a \mid s) = \frac{\exp(\theta_{s,a})}{\sum_{a'\in\mathcal{A}}\exp(\theta_{s,a'})},$$

where $\theta \in \mathbb{R}^{|\mathcal{S}||\mathcal{A}|}$.

One key motivation here is to leverage the tabular structure and the specific property of softmax policy to establish a sample complexity of global optimum convergence that is independent of the parameter μ. To this end, as in the non-private case (Agarwal et al., 2021a; Yuan et al., 2022), we will consider a regularized problem, whose FOSP turns out to be an approximate global optimal solution of the unregularized (original) objective, for proper choice of regularization. In particular, we consider the following log-barrier regularization objective:

$$\begin{aligned} J_\lambda(\theta) &\overset{\text{def}}{=} J(\theta) - \lambda\mathbb{E}_{s\sim\text{Unif}_\mathcal{S}}\left[\text{KL}(\text{Unif}_\mathcal{A}, \pi_\theta(\cdot|s))\right] \\ &= J(\theta) + \frac{\lambda}{|\mathcal{A}||S|}\sum_{s,a}\log\pi_\theta(a|s) + \lambda\log|\mathcal{A}|, \quad (3) \end{aligned}$$

where the KL divergence is $\text{KL}(p,q) = \mathbb{E}_{x\sim p}\left[\log\frac{p(x)}{q(x)}\right]$, Unif_χ denotes the uniform distribution over a set χ and $\lambda > 0$ is the regularization constant.

We will run our DP-PG over this regularized objective by using the sample-based gradient estimator at each step with proper choices of batch size and learning rate. Then, we have the following main result regarding the global optimum convergence in terms of the unregularized $J(\theta)$. The proof is given in Appendix C.

Theorem 4. *Consider Algorithm 1 applied to $J_\lambda(\theta)$. For any $m > 0$, setting $\sigma^2 = \frac{8\ln(1.25/\delta)\cdot R_{\max}^2 G^2}{m^2\varepsilon^2}$ ensures (ε,δ)-DP. Further, there exist proper choices of parameters for m and η, such that the following holds*

$$J^* - \frac{1}{T}\sum_{t=1}^{T}\mathbb{E}\left[J(\theta_t)\right] \leqslant O(\alpha),$$

when the sample size satisfies $N \geqslant O\left(\frac{1}{\alpha^6} + \frac{\sqrt{d}}{\alpha^{9/2}\varepsilon}\right)$.

Remark 8. The first term in the sample complexity bound matches the non-private one in Yuan et al. (2022), while the second term is the lower-order privacy cost (for constant ε and d). We note that while the dependence on α is worse than the previous one, there is no dependence on μ in the bound, which could offer benefits when μ is quite small.

References

Agarwal, A., Kakade, S. M., Lee, J. D., and Mahajan, G. On the theory of policy gradient methods: Optimality, approximation, and distribution shift. *Journal of Machine Learning Research*, 22(98):1–76, 2021a.

Agarwal, A., Kakade, S. M., Lee, J. D., and Mahajan, G. On the theory of policy gradient methods: Optimality, approximation, and distribution shift. *Journal of Machine Learning Research*, 22(98):1–76, 2021b.

Bassily, R., Smith, A., and Thakurta, A. Private empirical risk minimization: Efficient algorithms and tight error bounds. In *2014 IEEE 55th annual symposium on foundations of computer science*, pp. 464–473. IEEE, 2014.

Bassily, R., Feldman, V., Talwar, K., and Guha Thakurta, A. Private stochastic convex optimization with optimal rates. *Advances in neural information processing systems*, 32, 2019.

Chaudhuri, K., Monteleoni, C., and Sarwate, A. D. Differentially private empirical risk minimization. *Journal of Machine Learning Research*, 12(3), 2011.

Chowdhury, S. R. and Zhou, X. Differentially private regret minimization in episodic markov decision processes. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, pp. 6375–6383, 2022.

Ding, Y., Zhang, J., and Lavaei, J. On the global optimum convergence of momentum-based policy gradient. In Camps-Valls, G., Ruiz, F. J. R., and Valera, I. (eds.), *Proceedings of The 25th International Conference on Artificial Intelligence and Statistics*, volume 151 of *Proceedings of Machine Learning Research*, pp. 1910–1934. PMLR, 28–30 Mar 2022.

Dwork, C., McSherry, F., Nissim, K., and Smith, A. Calibrating noise to sensitivity in private data analysis. In *Theory of Cryptography: Third Theory of Cryptography Conference, TCC 2006, New York, NY, USA, March 4-7, 2006. Proceedings 3*, pp. 265–284. Springer, 2006.

Guo, D., Yang, D., Zhang, H., Song, J., Zhang, R., Xu, R., Zhu, Q., Ma, S., Wang, P., Bi, X., et al. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning. *arXiv preprint arXiv:2501.12948*, 2025.

Hsu, J., Huang, Z., Roth, A., Roughgarden, T., and Wu, Z. Private matchings and allocations. *SIAM Journal on Computing*, 45:1953–1984, 2016.

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

Levine, S. and Koltun, V. Guided policy search. In *International conference on machine learning*, pp. 1–9. PMLR, 2013.

Liu, B., Cai, Q., Yang, Z., and Wang, Z. Neural trust region/proximal policy optimization attains globally optimal policy. *Advances in neural information processing systems*, 32, 2019.

Liu, Y., Zhang, K., Basar, T., and Yin, W. An improved analysis of (variance-reduced) policy gradient and natural policy gradient methods. *Advances in Neural Information Processing Systems*, 33:7624–7636, 2020.

Liu, Y., Zhang, K., Başar, T., and Yin, W. An improved analysis of (variance-reduced) policy gradient and natural policy gradient methods, 2022. URL https://arxiv.org/abs/2211.07937.

Ouyang, L., Wu, J., Jiang, X., Almeida, D., Wainwright, C., Mishkin, P., Zhang, C., Agarwal, S., Slama, K., Ray, A., et al. Training language models to follow instructions with human feedback. *Advances in neural information processing systems*, 35:27730–27744, 2022.

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

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

Shani, L., Efroni, Y., and Mannor, S. Adaptive trust region policy optimization: Global convergence and faster rates for regularized mdps. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pp. 5668–5675, 2020.

Shao, Z., Wang, P., Zhu, Q., Xu, R., Song, J., Bi, X., Zhang, H., Zhang, M., Li, Y., Wu, Y., et al. Deepseekmath: Pushing the limits of mathematical reasoning in open language models. *arXiv preprint arXiv:2402.03300*, 2024.

Shariff, R. and Sheffet, O. Differentially private contextual linear bandits. *Advances in Neural Information Processing Systems*, 31, 2018.

Silver, D., Huang, A., Maddison, C. J., Guez, A., Sifre, L., Van Den Driessche, G., Schrittwieser, J., Antonoglou, I., Panneershelvam, V., Lanctot, M., et al. Mastering the game of go with deep neural networks and tree search. *nature*, 529(7587):484–489, 2016.

Sutton, R. S., McAllester, D., Singh, S., and Mansour, Y. Policy gradient methods for reinforcement learning with function approximation. *Advances in neural information processing systems*, 12, 1999.

Vietri, G., Balle, B., Krishnamurthy, A., and Wu, S. Private reinforcement learning with pac and regret guarantees. In *International Conference on Machine Learning*, pp. 9754–9764. PMLR, 2020.

Wang, L., Cai, Q., Yang, Z., and Wang, Z. Neural policy gradient methods: Global optimality and rates of convergence, 2019. URL https://arxiv.org/abs/1909.01150.

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

Yu, C., Liu, J., Nemati, S., and Yin, G. Reinforcement learning in healthcare: A survey. *ACM Computing Surveys (CSUR)*, 55(1):1–36, 2021.

Yuan, R., Gower, R. M., and Lazaric, A. A general sample complexity analysis of vanilla policy gradient. In *International Conference on Artificial Intelligence and Statistics*, pp. 3332–3380. PMLR, 2022.

Zhang, J., Kim, J., O'Donoghue, B., and Boyd, S. Sample efficient reinforcement learning with reinforce. In *Proceedings of the AAAI conference on artificial intelligence*, volume 35, pp. 10887–10895, 2021.

Zhou, T., Wang, Y., Yan, L., and Tan, Y. Spoiled for choice? personalized recommendation for healthcare decisions: A multiarmed bandit approach. *Information Systems Research*, 34(4):1493–1512, 2023.

Zhou, X. Differentially private reinforcement learning with linear function approximation. *Proceedings of the ACM on Measurement and Analysis of Computing Systems*, 6 (1):1–27, 2022.

A. Proof of Theorem 2

A.1. ABC Assumption and Smoothness

Lemma 2 (ABC). *There exists constants $A, B, C \geqslant 0$ such that the policy gradient estimator satisfies:*

$$\mathbb{E}\left[\left\|\widetilde{\nabla}_m J(\theta)\right\|^2\right] \leqslant 2A(J^* - J(\theta)) + B\left\|\nabla \widehat{J}(\theta)\right\|^2 + C, \tag{4}$$

where $\nabla \widehat{J}(\theta) := \mathbb{E}_{s,a}[R(s,a)\nabla_\theta \log(\pi_\theta(a|s))]$, and $A = 0, B = 1 - 1/m, C = \frac{R_{\max}^2 G^2}{m} + d\sigma^2$

Proof. For notation simplicity, we let $g_\theta(\tau_i) := R(s_i, a_i)\nabla_\theta \log \pi_\theta(a_i|s_i)$. Thus, we have $\widetilde{\nabla}_m J(\theta) = \frac{1}{m}\sum_i g_\theta(\tau_i) + Z$. Notice that $\mathbb{E}[g_\theta(\tau_i)] = \mathbb{E}\left[\widetilde{\nabla}_m J(\theta)\right] = \nabla \widehat{J}(\theta)$, cause Z is the Gaussian bias, which expectation is 0.

Now, we have

$$\begin{aligned}
\mathbb{E}\left[\left\|\widetilde{\nabla}_m J(\theta)\right\|^2\right] &= \mathbb{E}\left[\left\|\frac{1}{m}\sum_i g_\theta(\tau_i) + Z\right\|^2\right] \\
&= \mathbb{E}\left[\left\|\frac{1}{m}\sum_i g_\theta(\tau_i)\right\|^2\right] + \mathbb{E}\left[\|Z\|^2\right] + 2\cdot\mathbb{E}\left[\left\langle \frac{1}{m}\sum_i g_\theta(\tau_i), Z\right\rangle\right] \\
&= \mathbb{E}\left[\left\|\frac{1}{m}\sum_i g_\theta(\tau_i)\right\|^2\right] + d\sigma^2 + 0 \\
&= \mathbb{E}\left[\left\|\frac{1}{m}\sum_i g_\theta(\tau_i) - \nabla \widehat{J}(\theta) + \nabla \widehat{J}(\theta)\right\|^2\right] + d\sigma^2 \\
&= \left\|\nabla \widehat{J}(\theta)\right\|^2 + \mathbb{E}\left[\left\|\frac{1}{m}\sum_i g_\theta(\tau_i) - \nabla \widehat{J}(\theta)\right\|^2\right] + d\sigma^2 \\
&= \left\|\nabla \widehat{J}(\theta)\right\|^2 + \frac{1}{m^2}\sum_i \mathbb{E}\left[\left\|g_\theta(\tau_i) - \nabla \widehat{J}(\theta)\right\|^2\right] + d\sigma^2 \\
&= \left\|\nabla \widehat{J}(\theta)\right\|^2 + \frac{1}{m}\cdot\mathbb{E}\left[\|g_\theta(\tau_1)\|^2 - \left\|\nabla \widehat{J}(\theta)\right\|^2\right] + d\sigma^2.
\end{aligned}$$

To proceed, we need to establish an upper bound on $\mathbb{E}\left[\|g_\theta(\tau_1)\|^2\right]$.

In particular, we have

$$\begin{aligned}
\mathbb{E}\left[\|g_\theta(\tau_1)\|^2\right] &= \mathbb{E}\left[|R(s_1,a_1)|^2 \|\nabla_\theta \log \pi_\theta(a_1 \mid s_1)\|^2\right] \\
&\leqslant R_{\max}^2 G^2,
\end{aligned}$$

which follows from Assumption 1 (LS).

Hence, we conclude that:

$$\mathbb{E}\left[\left\|\widetilde{\nabla}_m J(\theta)\right\|^2\right] \leqslant \left(1 - \frac{1}{m}\right)\left\|\nabla \widehat{J}(\theta)\right\|^2 + \frac{R_{\max}^2 G^2}{m} + d\sigma^2.$$

i.e., ABC condition in (4) is satisfied with $A = 0, B = 1 - 1/m, C = \frac{R_{\max}^2 G^2}{m} + d\sigma^2$ □

Lemma 3 (Smoothness under LS). *Under LS assumption in Assumption 1, $J(\cdot)$ is L-smooth, namely $\left\|\nabla^2 J(\theta)\right\| \leqslant L$ for all θ, with*

$$L = R_{\max}(G^2 + F).$$

Proof. For smoothness, it suffices to bound the operator norm of Hessian, i.e., $\left\|\nabla^2 J(\theta)\right\|$.

By definition, we have

$$\begin{aligned}
\nabla^2 J(\theta) &= \nabla_\theta \mathbb{E}_{s\sim\rho, a\sim\pi_\theta(\cdot|s)} \left[R(s,a)\nabla_\theta \log \pi_\theta(a|s)\right] \\
&\overset{(a)}{=} \nabla_\theta \int p_\theta(s,a) \left(R(s,a)\nabla_\theta \log \pi_\theta(a|s)\right) d(s,a) \\
&\overset{(b)}{=} \int \nabla_\theta p_\theta(s,a) \left(R(s,a)\nabla_\theta \log \pi_\theta(a|s)\right)^\top d(s,a) + \int p_\theta(s,a) \left(R(s,a)\nabla_\theta^2 \log \pi_\theta(a|s)\right) d(s,a) \\
&= \mathbb{E}_{s,a\sim p_\theta} \left[R(s,a)\nabla_\theta \log \pi_\theta(a|s) \log \pi_\theta(a|s)^\top\right] + \mathbb{E}_{s,a\sim p_\theta} \left[R(s,a)\nabla_\theta^2 \log \pi_\theta(a|s)\right]
\end{aligned}$$

where in (a), we let $p_\theta(s,a) := \rho(s)\pi_\theta(a|s)$, and (b) holds by chain rules.

Thus, we have

$$\left\|\nabla_\theta^2 J(\theta)\right\| \leq \underbrace{\mathbb{E}_{s,a}\left[|R(s,a)| \left\|\nabla_\theta \log \pi_\theta(a|s)\right\|^2\right]}_{\mathcal{T}_1} + \underbrace{\mathbb{E}_{s,a}\left[|R(s,a)| \left\|\nabla_\theta^2 \log \pi_\theta(a|s)\right\|\right]}_{\mathcal{T}_2}.$$

For $\mathcal{T}_1$ and $\mathcal{T}_2$, by Assumption 1, we have

$$\mathcal{T}_1 \leq R_{\max} G^2, \quad \mathcal{T}_2 \leq R_{\max} F,$$

which hence completes the proof. □

A.2. FOSP convergence

Proof.

Lemma 4 (Adapted from Theorem 3.4 in Yuan et al. (2022)). *Suppose that J is smoothness and satisfy Assumption 2. Consider the iterates θ_t of the PG method with step size $\eta_t = \eta \in (0, \frac{2}{LB})$, let $\delta_0 = J^* - J(\theta_0)$, if $A = 0$, we have:*

$$E\left[\left\|\nabla J(\theta_U)\right\|^2\right] \leqslant \frac{2\delta_0}{\eta T(2 - LB\eta)} + \frac{LC\eta}{2 - LB\eta}. \tag{5}$$

where θ_U is uniformly sampled from $\{\theta_0, ..., \theta_{T-1}\}$

Followed by Lemma 4, when $\eta < \frac{1}{LB}$, we can imply $\frac{1}{2-LB\eta} < 1$, then we can simply the equation into this:

$$E\left[\left\|\nabla J(\theta_U)\right\|^2\right] \leqslant \frac{2\delta_0}{\eta T} + LC\eta, \tag{6}$$

where $B = 1 - 1/m, \delta_0 = J^* - J(\theta_0), L = R_{max}(G^2 + F), C = \frac{R_{max}^2 G^2}{m} + d\sigma^2, G$ and F are constants.

From Theorem 1, to make sure our algorithm satisfy the (ε, δ)-DP, we set $\sigma^2 = \frac{8\ln(1.25/\delta)\cdot R_{\max}^2 G^2}{m^2\varepsilon^2}$.

Based on Lemma 4 and Equation 6, choose $\eta = \min\{\frac{1}{LB}, \frac{\sqrt{2\delta_0}}{\sqrt{TLC}}\}$, we have:

$$\begin{aligned}\mathbb{E}\left[\|\nabla J(\theta_U)\|^2\right] &\leqslant \frac{2\delta_0 LB}{T} + \frac{2\sqrt{2\delta_0 LC}}{\sqrt{T}} \\ &= O\left(\frac{1}{T} + \frac{\sqrt{C}}{\sqrt{T}}\right) \\ &= O\left(\frac{m}{N} + \frac{1}{\sqrt{N}} + \frac{\sigma\sqrt{md}}{\sqrt{N}}\right) \\ &= O\left(\frac{m}{N} + \frac{1}{\sqrt{N}} + \frac{\sqrt{d}}{\varepsilon\sqrt{Nm}}\right).\end{aligned}$$

Thus, we can determine the value of m.

To balance the terms in the convergence bound:

$$O\left(\frac{m}{N} + \frac{1}{\sqrt{N}} + \frac{\sqrt{d}}{\varepsilon\sqrt{Nm}}\right).$$

Setting $\frac{m}{N} = \frac{\sqrt{d}}{\varepsilon\sqrt{Nm}}$, we solve for:

$$m = \left(\frac{\sqrt{d}}{\varepsilon}\right)^{2/3} N^{1/3} = (1/\varepsilon)^{2/3} (Nd)^{1/3}.$$

Substituting back, the convergence bound simplifies to:

$$O\left(\frac{1}{\sqrt{N}} + \left(\frac{\sqrt{d}}{N\varepsilon}\right)^{2/3}\right).$$

□

B. Proof of Theorem 3

Proof. We know that:

$$E\left[\|\nabla J(\theta_U)\|^2\right] = \frac{1}{T}\sum_{t=1}^{T} \mathbb{E}\left[\|\nabla J(\theta_t)\|^2\right].$$

Besides, followed by Lemma 1, we obtain that:

$$(J^* - J(\theta))^2 \leqslant \left(\frac{G}{\mu}\|\nabla J(\theta_U)\| + \sqrt{\alpha_{bias}}\right)^2 \leqslant 2\frac{G^2}{\mu^2}\|\nabla J(\theta_U)\|^2 + 2\alpha_{bias}$$

which holds by $(p+q)^2 \leqslant 2p^2 + 2q^2$.

Taking expectation over both sides condition on θ_t, yields that

$$\frac{1}{T}\sum_{t=1}^{T} E\left[(J^* - J(\theta))^2\right] \leqslant 2\frac{G^2}{\mu^2}\frac{1}{T}\sum_{t=1}^{T} E\left[\|\nabla J(\theta_t)\|^2\right] + 2\alpha_{bias} \overset{(a)}{=} O\left(\frac{1}{\mu^2}\left(\frac{1}{\sqrt{N}} + \left(\frac{\sqrt{d}}{N\varepsilon}\right)^{2/3}\right)\right) + O(\alpha_{bias})$$

where (a) is hold by Theorem 2.

By applying Jensen inequality twice, we have:

$$\frac{1}{T}\sum_{t=1}^{T} E\left[(J^* - J(\theta))^2\right] \geqslant E\left[\left(J^* - \frac{1}{T}\sum_{t=1}^{T} J(\theta_t)\right)^2\right] \geqslant \left(J^* - \frac{1}{T}\sum_{t=1}^{T} E\left[J(\theta_t)\right]\right)^2$$

So we have:

$$\left(J^* - \frac{1}{T}\sum_{t=1}^{T} E\left[J(\theta_t)\right]\right)^2 \leqslant O\left(\frac{1}{\mu^2}\left(\frac{1}{\sqrt{N}} + \left(\frac{\sqrt{d}}{N\varepsilon}\right)^{2/3}\right)\right) + O(\alpha_{bias})$$

In that case, we finally get the result:

$$J^* - \frac{1}{T}\sum_{t=1}^{T} \mathbb{E}\left[J(\theta_t)\right] = O\left(\frac{1}{\mu}\left(N^{-1/4} + \left(\frac{\sqrt{d}}{N\varepsilon}\right)^{1/3}\right)\right) + O(\sqrt{\alpha_{bias}}).$$

Here, we suppose $J^* - \frac{1}{T}\sum_{t=1}^{T} \mathbb{E}\left[J(\theta_t)\right] \leqslant O(\alpha) + O(\sqrt{\alpha_{bias}})$.

Hence, we have:

$$N \geq O\left(\frac{1}{\alpha^4\mu^4} + \frac{\sqrt{d}}{\alpha^3\mu^3\varepsilon}\right)$$

□

C. Proof of Theorem 4

Based on softmax settings, by simple calculus, we have

$$\begin{aligned} \frac{\partial \log \pi_\theta(a|s)}{\partial \theta_s} &= \mathbf{1}_a - \pi_s(\theta), \\ \frac{\partial^2 \log \pi_\theta(a|s)}{\partial \theta_s^2} &= -\mathbf{H}(\pi_s(\theta)), \end{aligned} \tag{7}$$

where $\mathbf{1}_a \in \mathbb{R}^{|A|}$ is a vector with all zero entries except being 1 for the entry corresponding to action a, and $\mathbf{H}(\pi_s(\theta)) = \text{Diag}(\pi_s(\theta)) - \pi_s(\theta)\pi_s(\theta)^\top$.

In particular, for softmax, we can determine the G and F in Assumption 1

$$\begin{aligned} \|\nabla_\theta \log \pi_\theta(a \mid s)\| &\leqslant G := \sqrt{1 - \frac{1}{|\mathcal{A}|}} \\ \|\nabla_\theta^2 \log \pi_\theta(a \mid s)\| &\leqslant F := 1. \end{aligned}$$

C.1. FOSP of Softmax Policy

Lemma 5. *The regularized gradient estimator $\widetilde{\nabla}_m J_\lambda(\theta)$ satisfies Lemma 2 with parameters:*

$$\begin{aligned} A &= 0, \quad B = 1 - \frac{1}{m} \\ C &= \frac{2}{m}\left(1 - \frac{1}{|\mathcal{A}|}\right)\left(R_{\max}^2 + \frac{\lambda^2}{|S|}\right) + d\sigma^2, \end{aligned}$$

Specifically, we have the variance bound:

$$\mathbb{E}\left[\left\|\widetilde{\nabla}_m J_\lambda(\theta)\right\|^2\right] \leqslant \left(1-\frac{1}{m}\right)\|\nabla J_\lambda(\theta)\|^2 + d\sigma^2 + \frac{2}{m}\left(1-\frac{1}{|\mathcal{A}|}\right)\left(R_{\max}^2+\frac{\lambda^2}{|S|}\right).$$

Proof. Let $g(\tau \mid \theta)$ be a stochastic gradient estimator of one single sampled trajectory τ. Thus $\widetilde{\nabla}_m J(\theta) = \frac{1}{m}\sum_{i=1}^m g(\tau_i \mid \theta) + Z$.

From equation (3) we have the following gradient estimator

$$\widetilde{\nabla}_m J_\lambda(\theta) = \widetilde{\nabla}_m J(\theta) + \frac{\lambda}{|\mathcal{A}||S|}\sum_{s,a}\nabla_\theta \log \pi_{s,a}(\theta).$$

For a state $s \in \mathcal{S}$, we have

$$\begin{aligned}\frac{\lambda}{|\mathcal{A}||S|}\sum_{a\in\mathcal{A}}\frac{\partial \log \pi_{s,a}(\theta)}{\partial \theta_s} &\overset{(7)}{=} \frac{\lambda}{|\mathcal{A}||S|}\sum_{a\in\mathcal{A}}(\mathbf{1}_a - \pi_s(\theta)) \\ &= \frac{\lambda \mathbf{1}_{|A|}}{|\mathcal{A}||S|} - \frac{\lambda}{|S|}\pi_s(\theta) \\ &= \frac{\lambda}{|S|}\left(\frac{\mathbf{1}_{|\mathcal{A}|}}{|\mathcal{A}|} - \pi_s(\theta)\right),\end{aligned}$$

where $\mathbf{1}_{|\mathcal{A}|} \in \mathbb{R}^{|\mathcal{A}|}$ is a vector of all ones.

Thus we have

$$\widetilde{\nabla}_m J_\lambda(\theta) = \widetilde{\nabla}_m J(\theta) + \frac{\lambda}{|S|}\left(\frac{\mathbf{1}_{|\mathcal{A}|}}{|\mathcal{A}|} - [\pi_s(\theta)]_{s\in S}\right), \tag{8}$$

where $\mathbf{1} \in \mathbb{R}^{|S||\mathcal{A}|}$ and $[\pi_s(\theta)]_{s\in S} = [\pi_{s_1}(\theta); ...; \pi_{s_{|S|}}(\theta)] \in \mathbb{R}^{|S||\mathcal{A}|}$ is the stacking of the vectors $\pi_s(\theta)$.

Next, taking expectation on the trajectories, we have

$$\begin{aligned}
\mathbb{E}\left[\left\|\widetilde{\nabla}_m J_\lambda(\theta)\right\|^2\right] &\overset{(8)}{=} \mathbb{E}\left[\left\|\widetilde{\nabla}_m J(\theta) + \frac{\lambda}{|S|}\left(\frac{\mathbf{1}_{|\mathcal{A}|}}{|\mathcal{A}|} - [\pi_s(\theta)]_{s\in S}\right)\right\|^2\right] \\
&= \mathbb{E}\left[\left\|\nabla J(\theta) + \frac{\lambda}{|S|}\left(\frac{\mathbf{1}_{|\mathcal{A}|}}{|\mathcal{A}|} - [\pi_s(\theta)]_{s\in S}\right) + \widetilde{\nabla}_m J(\theta) - \nabla J(\theta)\right\|^2\right] \\
&\overset{(a)}{=} \|\nabla J_\lambda(\theta)\|^2 + \mathbb{E}\left[\left\|\widetilde{\nabla}_m J(\theta) - \nabla J(\theta)\right\|^2\right] \\
&\overset{(b)}{=} \|\nabla J_\lambda(\theta)\|^2 + \mathbb{E}\left[\left\|\widehat{\nabla}_m J(\theta) + \mathbf{Z} - \nabla J(\theta)\right\|^2\right] \\
&= \|\nabla J_\lambda(\theta)\|^2 + \frac{\mathbb{E}\left[\|g(\tau_1|\theta) - \nabla J(\theta)\|^2\right]}{m} + d\sigma^2 \\
&= \|\nabla J_\lambda(\theta)\|^2 + d\sigma^2 \\
&\quad + \frac{\mathbb{E}\left[\left\|g(\tau_1 \mid \theta) + \frac{\lambda}{|S|}\left(\frac{1}{|A|} - [\pi_s(\theta)]_{s\in S}\right) - \nabla J(\theta) - \frac{\lambda}{|S|}\left(\frac{1}{|A|} - [\pi_s(\theta)]_{s\in S}\right)\right\|^2\right]}{m} \\
&\overset{(c)}{=} \left(1 - \frac{1}{m}\right)\|\nabla J_\lambda(\theta)\|^2 + \frac{\mathbb{E}\left[\left\|g(\tau_1 \mid \theta) + \frac{\lambda}{|S|}\left(\frac{1}{|A|} - [\pi_s(\theta)]_{s\in S}\right)\right\|^2\right]}{m} + d\sigma^2 \\
&\overset{(d)}{\leqslant} \left(1 - \frac{1}{m}\right)\|\nabla J_\lambda(\theta)\|^2 + \frac{2\mathbb{E}\left[\|g(\tau_1|\theta)\|^2\right] + 2\left\|\frac{\lambda}{|S|}\left(\frac{\mathbf{1}_{|\mathcal{A}|}}{|\mathcal{A}|} - [\pi_s(\theta)]_{s\in S}\right)\right\|^2}{m} + d\sigma^2,
\end{aligned}$$

where (a) and (c) holds by definition of $\nabla J_\lambda(\theta)$; (b) holds by definition of $\widetilde{\nabla}_m J(\theta)$; (d) holds by $(p+q)^2 \leqslant 2p^2 + 2q^2$.

In particular, we have

$$\left\|\frac{\lambda}{|S|}\left(\frac{\mathbf{1}_{|\mathcal{A}|}}{|\mathcal{A}|} - [\pi_s(\theta)]_{s\in S}\right)\right\|^2 \leqslant \frac{\lambda^2}{|S|^2}\left(\frac{|S||\mathcal{A}|}{|\mathcal{A}|^2} - 2\frac{|S|}{|\mathcal{A}|} + |S|\right) = \frac{\lambda^2}{|S|}\left(1 - \frac{1}{|\mathcal{A}|}\right),$$

where the inequality is obtained by using $\|\pi_s(\theta)\|^2 \leqslant 1$.

As for $\mathbb{E}\left[\|g(\tau_1 \mid \theta)\|^2\right]$, we have

$$\mathbb{E}\left[\|g(\tau_1 \mid \theta)\|^2\right] \leqslant R_{\max}^2 G^2 = R_{\max}^2\left(1 - \frac{1}{|A|}\right),$$

where the equality is obtained by Assumption 1 with $G^2 = \left(1 - \frac{1}{|A|}\right)$.

Combining above, we have that the gradient estimator $\widetilde{\nabla}_m J_\lambda(\theta)$ satisfies ABC assumption with

$$\mathbb{E}\left[\left\|\widetilde{\nabla}_m J_\lambda(\theta)\right\|^2\right] \leqslant \left(1 - \frac{1}{m}\right)\|\nabla J_\lambda(\theta)\|^2 + \frac{2}{m}\left(1 - \frac{1}{|A|}\right)\left(R_{\max}^2 + \frac{\lambda^2}{|S|}\right) + d\sigma^2,$$

where $A = 0, B = 1 - \frac{1}{m}, C = \frac{2}{m}\left(1 - \frac{1}{|\mathcal{A}|}\right)\left(R_{\max}^2 + \frac{\lambda^2}{|S|}\right) + d\sigma^2$. □

Lemma 6 (Regularized FOSP Convergence). *Under the learning rate condition $\eta < \frac{1}{LB}$, the iterates satisfy:*

$$\mathbb{E}\left[\|\nabla J_\lambda(\theta_U)\|^2\right] \leqslant \frac{2\delta_0}{\eta T} + LC\eta, \tag{9}$$

where $B = 1 - 1/m$, $\delta_0 = J^* - J(\theta_0)$, $L = R_{\max}(2 - \frac{1}{|\mathcal{A}|})$, and C as defined in Lemma 5.

Proof. From Lemma E.3 in Yuan et al. (2022), We know that $J_\lambda(\cdot)$ is smooth and Lipschitz.

Then, based on Lemma 4, we know that:

$$\mathbb{E}\left[\|\nabla J_\lambda(\theta_U)\|^2\right] \leqslant \frac{2\delta_0}{\eta T(2-LB\eta)} + \frac{LC\eta}{2-LB\eta}.$$

let $\eta < \frac{1}{LB}$, we can simply the equation into this:

$$\mathbb{E}\left[\|\nabla J_\lambda(\theta_U)\|^2\right] \leqslant \frac{2\delta_0}{\eta T} + LC\eta,$$

where $B = 1 - 1/m$, $\delta_0 = J^* - J(\theta_0)$, $L = R_{max}(2 - \frac{1}{|A|})$, $C = \frac{2}{m}\left(1 - \frac{1}{|A|}\right)\left(R_{\max}^2 + \frac{\lambda^2}{|S|}\right) + d\sigma^2$, $G^2 = 1 - \frac{1}{|\mathcal{A}|}$ and $F = 1$. □

Then we need to choose proper σ to satisfy the (ε, δ)-DP. Note that the sensitivity Δ of the gradient estimator $\widehat{\nabla}_m J_\lambda(\theta)$ is dominated by the data-dependent term. Despite introducing the regularization term λ, this term only depends on the policy parameters θ (independent of data), thus it does not affect the sensitivity. The ℓ_2-sensitivity of the gradient remains $\Delta = \frac{2R_{\max}G}{m}$.

Lemma 7. *let* $\sigma^2 = \frac{8\ln(1.25/\delta)\cdot R_{\max}^2 G^2}{m^2\varepsilon^2}$*, the batch size* m *be set as:* $m = (1/\varepsilon)^{2/3}(Nd)^{1/3}$*, and* $\eta = min(\frac{1}{LB}, \frac{\sqrt{2\delta_0}}{\sqrt{TLC}})$*, we have:*

$$\mathbb{E}\left[\|\nabla J_\lambda(\theta_U)\|^2\right] \leqslant O\left(\frac{1}{\sqrt{N}} + \left(\frac{\sqrt{d}}{N\varepsilon}\right)^{2/3}\right). \tag{10}$$

Proof. for $\eta = min(\frac{1}{LB}, \frac{\sqrt{2\delta_0}}{\sqrt{TLC}})$ we know:

$$\mathbb{E}\left[\|\nabla J_\lambda(\theta_U)\|^2\right] \leqslant \frac{2\delta_0 LB}{T} + \frac{2\sqrt{2\delta_0 LC}}{\sqrt{T}} = O\left(\frac{1}{T} + \frac{\sqrt{C}}{\sqrt{T}}\right) = O\left(\frac{m}{N} + \frac{1}{\sqrt{N}} + \frac{\sigma\sqrt{md}}{\sqrt{N}}\right).$$

Plug in $\sigma^2 = \frac{8\ln(1.25/\delta)\cdot R_{\max}^2 G}{m^2\varepsilon^2}$ and $m = (1/\varepsilon)^{2/3}(Nd)^{1/3}$, we have:

$$\mathbb{E}\left[\|\nabla J_\lambda(\theta_U)\|^2\right] \leqslant O\left(\frac{1}{\sqrt{N}} + (\frac{\sqrt{d}}{N\varepsilon})^{2/3}\right).$$

□

C.2. Global Optimum Convergence

We first introduce an important proposition to bound our global private optimum convergence of softmax with log barrier regularization.

Proposition 1 (Adapted from Theorem 5.2 in Agarwal et al. (2021a))**.** *Suppose* θ *is such that* $\|\nabla J_\lambda(\theta)\| \leqslant \frac{\lambda}{2|S||\mathcal{A}|}$*. Then for every initial distribution* ρ*, we have*

$$J^* - J(\theta) \leqslant 2\lambda. \tag{11}$$

Proof. Firstly, we define the following set of "bad" iterates:

$$I^+ \triangleq \left\{ t \in \{1, \ldots, T\} \;\middle|\; \|\nabla J_\lambda(\theta_t)\| \geqslant \frac{\lambda}{2|\mathcal{S}||\mathcal{A}|} \right\},$$

with$\lambda = \frac{\alpha}{2}$.

From Proposition 1, we know that if $\|\nabla J_\lambda(\theta)\| \leqslant \frac{\lambda}{2|\mathcal{S}||\mathcal{A}|}$, we have $J^* - J(\theta) \leqslant 2\lambda$.

Hence, we have:

$$\begin{aligned} J^* - \frac{1}{T}\sum_{t=1}^{T} J(\theta_t) &= \frac{1}{T}\sum_{t\in I^+} J^* - J(\theta_t) + \frac{1}{T}\sum_{t\notin I^+} J^* - J(\theta_t) \\ &\overset{(a)}{\leqslant} \frac{|I^+|}{T}\cdot 2\mathcal{R}_{\max} + \frac{1}{T}\sum_{t\notin I^+} J^* - J(\theta_t) \\ &\overset{(11)}{\leqslant} \frac{|I^+|}{T}\cdot 2\mathcal{R}_{\max} + \frac{T-|I^+|}{T}\cdot 2\lambda \\ &\leqslant \frac{|I^+|}{T}\cdot 2\mathcal{R}_{\max} + 2\lambda \\ &\leqslant \frac{|I^+|}{T}\cdot 2\mathcal{R}_{\max} + \alpha, \end{aligned} \tag{12}$$

where (a) holds by $J(\cdot) \leqslant R_{\max}$, then $J^* - J(\theta_t) \leqslant J^* + J(\theta_t) \leqslant 2R_{\max}$.

Now we turn to bound $|I^+|$

$$\sum_{t=1}^{T} \|\nabla J_\lambda(\theta_t)\|^2 \geqslant \sum_{t\in I^+} \|\nabla J_\lambda(\theta_t)\|^2 \geqslant \frac{|I^+|\lambda^2}{4|\mathcal{S}|^2|\mathcal{A}|^2}.$$

Through a straightforward mathematical transformation, we get

$$\begin{aligned} \frac{|I^+|}{T} &\leqslant \frac{4|\mathcal{S}|^2|\mathcal{A}|^2}{\lambda^2}\cdot\frac{1}{T}\sum_{t=1}^{T}\|\nabla J_\lambda(\theta_t)\|^2 \\ &= \frac{16}{\alpha^2}\cdot|\mathcal{S}|^2|\mathcal{A}|^2\cdot\frac{1}{T}\sum_{t=1}^{T}\|\nabla J_\lambda(\theta_t)\|^2. \end{aligned}$$

Thus, we have

$$J^* - \frac{1}{T}\sum_{t=1}^{T} J(\theta_t) \overset{(12)}{\leqslant} \frac{32R_{\max}}{\alpha^2}|\mathcal{S}|^2|\mathcal{A}|^2\cdot\frac{1}{T}\sum_{t=1}^{T}\|\nabla J_\lambda(\theta_t)\|^2 + \alpha.$$

Taking expectation over the iterations on both sides, we have

$$J^* - \frac{1}{T}\sum_{t=1}^{T}\mathbb{E}\left[J(\theta_t)\right] \leqslant \frac{32R_{\max}}{\alpha^2}|\mathcal{S}|^2|\mathcal{A}|^2\cdot\frac{1}{T}\sum_{t=1}^{T}\mathbb{E}\left[\|\nabla J_\lambda(\theta_t)\|^2\right] + \alpha.$$

To guarantee that $J^* - \frac{1}{T}\sum_{t=1}^{T}\mathbb{E}\left[J(\theta_t)\right] \leqslant \alpha$, we need to show:

$$\frac{1}{T}\sum_{t=1}^{T}\mathbb{E}\left[\|\nabla J_\lambda(\theta_t)\|^2\right] \leqslant \alpha^3,$$

Hence, based on Equation 10 in Lemma 7, it is obvious to show that:

$$N \geqslant O\left(\frac{1}{\alpha^6} + \frac{\sqrt{d}}{\alpha^{9/2}\varepsilon}\right).$$

□