The Role of Baselines in Policy Gradient Optimization

Anonymous Author(s)
Affiliation
Address
email

Abstract

We study the effect of baselines in on-policy stochastic policy gradient optimization, and close the gap between the theory and practice of policy optimization methods. Our first contribution is to show that the *state value* baseline allows on-policy stochastic *natural* policy gradient (NPG) to converge to a globally optimal policy at an $O(1/t)$ rate, which was not previously known. The analysis relies on two novel findings: the expected progress of the NPG update satisfies a stochastic version of the non-uniform Łojasiewicz (NŁ) inequality, and with probability 1 the state value baseline prevents the optimal action's probability from vanishing, thus ensuring sufficient exploration. Importantly, these results provide a new understanding of the role of baselines in stochastic policy gradient: by showing that the variance of natural policy gradient estimates remains unbounded with or without a baseline, we find that variance reduction *cannot* explain their utility in this setting. Instead, the analysis reveals that the primary effect of the value baseline is to reduce the aggressiveness of the updates rather than their variance. That is, we demonstrate that a finite variance is *not necessary* for almost sure convergence of stochastic NPG, while controlling update aggressiveness is both necessary and sufficient. Additional experimental results verify these theoretical findings.

1 Introduction

The policy gradient (PG) [19] is a key concept in reinforcement learning (RL), lying at the foundation of policy-based and actor-critic methods, and responsible for some of the most prominent practical achievements in RL [17, 18, 5]. However, progress in the theoretical understanding of PG methods is recent, and a number of the techniques used in practice still lack rigorous support, particularly in the online stochastic regime where an action is sampled from the current policy at each iteration. We study stochastic policy optimization in more detail to close this gap between theory and practice.

In stochastic policy optimization, the two most common techniques for improving the basic algorithm are to include on-policy importance sampling (IS) and subtract a baseline. Including on-policy IS provides unbiased gradient estimates, but introduces high variance when an action's sampling probability is close to 0. Meanwhile, subtracting a baseline remains a heuristic [16] that has strong empirical but limited theoretical support. One possible benefit of a baseline is that it provides variance reduction [20], which has motivated work on designing alternative baselines that further reduce variance [20]. However, other work [3] has shown that variance reduction is not necessarily aligned with policy learning quality. To date, it has remained unclear how a baseline impacts the quality of the ultimate solution found by policy gradient optimization. We resolve this question in this work.

Recent progress in the theory of deterministic PG has shown that, given exact gradients, softmax policy gradient is able to converge to a globally optimal policy at a $O(1/t)$ rate [14]. Unfortunately, despite this guarantee, the constants in this rate can be extremely large [10] due to initialization sensitivity and poor performance at escaping sub-optimal plateaus [13]. Therefore, in the exact gradient setting, several techniques have been considered for mitigating the weaknesses of softmax

Submitted to 36th Conference on Neural Information Processing Systems (NeurIPS 2022). Do not distribute.

PG, leading to better constants [1] or even exponentially faster rates of $O(e^{-c \cdot t})$ for $c > 0$. Such improvements include adding entropy regularization [14, 2], normalizing the gradients [12], or applying natural policy gradient (NPG) [2, 7, 11].

However, in the on-policy *stochastic* optimization case, recent studies [11, 3] have shown that naively applying the above techniques, such as normalizing policy gradients or appying NPG, leads to unexpectedly *worse* performance than stochastic PG. That is, techniques that accelerate convergence in the exact policy gradient setting become *unsound* in the stochastic gradient setting, by inducing a non-zero probability of failure (i.e., failing to converge to a globally optimal solution) [11]. Such failures occur even when stochastic PG can still converge to a global optimum with probability 1. Previous work has indicated that one key reason behind the failure of these acceleration strategies arises from their "over-committal behaviour" in the stochastic setting, which occurs independently of the variance of the gradient estimates [3]. That is, baseline techniques with higher variance can still better avoid over-committal behaviour (i.e., premature convergence) and ultimately achieve better policy optimization [3].

To resolve this issue, we develop a deeper understanding of the role of baselines in stochastic policy optimization based on the following contributions. First, we establish a new result that combining on-policy IS with a value function baseline and natural policy gradient (NPG) can achieve almost sure convergence to a globally optimal policy at a $O(1/t)$ rate. This result is based on two novel findings: **(i)** At any iteration t, the conditional expected progress of the algorithm's next iterate obeys a stochastic non-uniform Łojasiewicz (NŁ) inequality. **(ii)** The use of the state value baseline (with appropriate learning rate control) almost surely prevents the probability of the optimal action from vanishing. These findings show that a key role of the value baseline is to automatically ensure "sufficient exploration" during on-policy stochastic optimization. Next, we provide a more detailed understanding of how the baseline modulates the circular interaction between stochastic action sampling and policy updates. Although a baseline has no effect on the exact gradient, it can play a major role in stochastic gradients. In this respect, we first show that the variance of the PG estimator is unbounded with or without a baseline, hence variance reduction cannot be the primary effect. Instead, our analysis reveals that the key role the baseline plays in ensuring global convergence is to reduce the aggressiveness of updates. That is, finite variance of the gradient estimates is not necessary for ensuring global convergence, while properly controlling update aggressiveness is both necessary and sufficient.

The remainder of the paper is organized as follows. Section 2 provides the main results that establish the almost sure convergence of stochastic NPG with on-policy IS and state value baseline to a globally optimal policy at a $O(1/t)$ rate. Section 3 then develops the new understanding of the role of the baseline by going beyond standard variance reduction arguments. To further support these theoretical findings, in Section 4 we provide some simulations to verify the results. Finally, Section 5 concludes the paper with a brief discussion.

2 On-policy Stochastic Natural Policy Gradient

We first consider a one-state Markov Decision Process (MDP) defined by a finite action space $[K] \coloneqq \{1, 2, \ldots, K\}$ where the true mean reward vector is given by $r \in (0, 1]^K$. The policy optimization problem is to maximize the expected reward,

$$\max_{\theta : [K] \to \mathbb{R}} \; \mathop{\mathbb{E}}_{a \sim \pi_\theta(\cdot)} [r(a)], \tag{1}$$

where the policy π_θ is parameterized by θ using the standard softmax parameterization,

$$\pi_\theta(a) = \frac{\exp\{\theta(a)\}}{\sum_{a' \in [K]} \exp\{\theta(a')\}}, \quad \text{for all } a \in [K]. \tag{2}$$

Our focus in this paper is on on-policy optimization, where at each iteration $t \geq 1$ the current policy π_{θ_t} is used to sample one action and perform one update.

For reward observations, rather than assume the true mean reward $r(a_t)$ is observed for the sampled action a_t, we make the more practical assumption that only a noisy observation $x(a) \in \mathbb{R}$ is drawn from an unknown distribution with expected value $r(a_t)$. In particular, the observed reward $x(a)$ is assumed to be sampled from a bounded distribution $x(a) \in [-R_{\max}, R_{\max}]$.

Assumption 1 (Bounded sampled reward)**.** *For each action $a \in [K]$, the true mean reward $r(a)$ is the expectation of a bounded reward distribution, i.e.,*

$$r(a) = \int_{-R_{\max}}^{R_{\max}} x \cdot P_a(x)\mu(dx) \tag{3}$$

where μ is a finite measure over $[-R_{\max}, R_{\max}]$, and $P_a(x) \geq 0$ is the probability density function with respect to μ, and $R_{\max} > 0$ is the reward range. We let R_a denote the reward distribution for action a defined by the density P_a and base measure μ.

Then, given a sampled reward observation $x(a) \sim R_a$, an unbiased estimate of the expected reward vector r can be formed by on-policy importance sampling (IS).

Definition 1 (On-policy importance sampling (IS))**.** *At iteration t, sample one action $a_t \sim \pi_{\theta_t}(\cdot)$ and observe one reward sample $x(a_t) \sim R_{a_t}$. Let $\hat{x}(a) = \mathbb{I}\{a = a_t\}\, x(a_t)$ for all $a \in [K]$. Then the IS reward estimate is constructed as $\hat{r}_t(a) = \frac{\mathbb{I}\{a=a_t\}}{\pi_{\theta_t}(a)} \cdot \hat{x}(a)$ for all $a \in [K]$.*

2.1 Failure Without a Baseline

First, to establish context, we review an existing negative result for the representative algorithm, natural policy gradient (NPG) [6], which for the softmax parameterization is defined as follows.

Update 1 (NPG with on-policy stochastic gradient)**.** $\theta_{t+1} \leftarrow \theta_t + \eta \cdot \hat{r}_t$*, where $\pi_\theta(a)$ is by Eq.* (2)*.*

It is well known that NPG behaves problematically with on-policy IS, even if the true mean reward $r(a_t)$ is observed. In particular, NPG converges to a sub-optimal deterministic policy with a constant positive probability in this case, as shown by [3, 11].

Proposition 1 ([11])**.** *Using Update 1, we have the following:* ***(i)*** *with positive probability, $\sum_{a \neq a^*} \pi_{\theta_t}(a) \to 1$ as $t \to \infty$;* ***(ii)*** *$\forall a \in [K]$, with positive probability, $\pi_{\theta_t}(a) \to 1$, as $t \to \infty$.*

Essentially Proposition 1 asserts that Update 1 is too aggressive: if sub-optimal actions are sampled t times successively, their probabilities will become exponentially close to 1; i.e., $1 - \sum_{a \neq a^*} \pi_{\theta_t}(a) \in O(e^{-c \cdot t})$. It follows that $\prod_{t=1}^{\infty} \sum_{a \neq a^*} \pi_{\theta_t}(a) > 0$; that is, the on-policy sampling process $a_t \sim \pi_{\theta_t}(\cdot)$ has a non-zero probability of sampling sub-optimal actions forever, which implies that there is a positive probability that π_{θ_t} fails to converge to an optimal deterministic policy.

2.2 Global Convergence with a Value Baseline

Despite the above failure, we now prove that subtracting a value baseline rectifies the problem for NPG. Consider the modified update that includes a baseline.

Update 2 (NPG, on-policy stochastic gradient with value baseline)**.** $\theta_{t+1} \leftarrow \theta_t + \eta \cdot \left(\hat{r}_t - \hat{b}_t\right)$*, where $\pi_\theta(a)$ is parameterized by Eq.* (2)*, $\hat{r}_t$ is given by Definition 1, $b_t \coloneqq \pi_{\theta_t}^\top r$, and $\hat{b}_t(a) = \left(\frac{\mathbb{I}\{a_t=a\}}{\pi_{\theta_t}(a)} - 1\right) \cdot b_t$ for all $a \in [K]$.*

Note that, since $\mathrm{softmax}(\theta) = \mathrm{softmax}(\theta + c \cdot \mathbf{1})$ for all $c \in \mathbb{R}$, Update 2 is equivalent to $\theta_{t+1} \leftarrow \theta_t + \eta \cdot \frac{\mathbb{I}\{a_t=a\}}{\pi_{\theta_t}(a)} \cdot \left(\hat{r}(a) - \pi_{\theta_t}^\top r\right)$. That is, for a_t the update reduces to $\theta_{t+1}(a_t) \leftarrow \theta_t(a_t) + \eta \cdot \frac{\hat{r}(a_t) - \pi_{\theta_t}^\top r}{\pi_{\theta_t}(a_t)}$, while for the other actions $a \neq a_t$ not sampled at iteration t, the parameters $\theta_{t+1}(a) \leftarrow \theta_t(a)$ remain unchanged. Unfortunately, the variance of this update is not uniformly bounded whenever $\pi_{\theta_t}(a)$ is close to 0 for at least one action $a \in [K]$, therefore standard stochastic gradient analysis for bounded variance estimators [15, 22, 9, 21] cannot be applied. Instead, we need to develop two new techniques to establish global convergence results, both of which rely heavily on using baselines.

The following lemma provides the first key technique, which we refer to as the stochastic non-uniform Łojasiewciz (NŁ) inequality.

Lemma 1 (Stochastic NŁ)**.** *Suppose Assumption 1 holds. Let $a^* \coloneqq \arg\max_{a \in [K]} r(a)$ denote the optimal action, and $\Delta \coloneqq r(a^*) - \max_{a \neq a^*} r(a)$ denote the reward gap. Using Update 2 with on-policy sampling $a_t \sim \pi_{\theta_t}(\cdot)$ and IS estimator of Definition 1 with learning rate,*

$$\eta = \frac{\pi_{\theta_t}(a_t) \cdot \left|r(a_t) - \pi_{\theta_t}^\top r\right|}{8 \cdot R_{\max}^2}, \tag{4}$$

we have, for all $t \geq 1$,

$$\mathbb{E}_t[\pi_{\theta_{t+1}}^\top r] - \pi_{\theta_t}^\top r \geq \frac{\eta}{2} \cdot \sum_{i=1}^{K} \pi_{\theta_t}(i) \cdot \left(r(i) - \pi_{\theta_t}^\top r\right)^2 \qquad \left(= \frac{\eta}{2} \cdot \mathrm{Var}_{a \sim \pi_\theta(\cdot)}[r(a)]\right) \tag{5}$$

$$\geq \frac{\eta}{2} \cdot \Delta \cdot \pi_{\theta_t}(a^*) \cdot \left(\pi^* - \pi_{\theta_t}\right)^\top r, \tag{6}$$

where $\mathbb{E}_t[\cdot]$ is on randomness from on-policy sampling $a_t \sim \pi_{\theta_t}(\cdot)$ and reward sampling $x \sim R_{a_t}$, and $\pi^* := \arg\max_{\pi \in \Delta(K)} \pi^\top r$ is the globally optimal policy.

Remark 1. *The learning rate Eq. (4) is $\eta \in O(1/t)$ in retrospect later with knowing the convergence rate. If true mean reward $r(a_t)$ is observed, then using Update 2 with learning rate $\eta \in \Theta(1)$ also achieves a $O(1/t)$ global convergence rate. We only present the results and proofs for observing true mean reward/value and using constant learning rates in the appendix, which is of independent interest under scenarios whenever the reward/value can be accurately estimated.*

We refer to $\pi_{\theta_t}(a^*)$ in Lemma 1 the **stochastic NŁ coefficient**. Lemma 1 is a stochastic generalization of the NŁ inequality, which has been widely used in proving global convergence of true gradient softmax PG variants [14, 13, 12, 11, 23]. It is stochastic because Lemma 1 contains an expectation. It is non-uniform because Lemma 1 depends on the current parameter θ_t, which cannot be uniformly lower bounded away from 0 across the entire domain of $\theta \in \mathbb{R}^K$ (that is, one can always find θ such that $\pi_\theta(a^*)$ is arbitrarily close to 0).

Note that, since $\eta \geq 0$ and $\mathrm{Var}_{a \sim \pi_\theta(\cdot)}[r(a)] \geq 0$, we have $\mathbb{E}_t[\pi_{\theta_{t+1}}^\top r] - \pi_{\theta_t}^\top r \geq 0$, which implies that after one update, Update 2 achieves non-negative progress *in expectation*. Combining Lemma 1 with Doob's supermartingale convergence theorem then leads to the following result.

Corollary 1. *The sequence $\{\pi_{\theta_t}^\top r\}_{t \geq 1}$ converges with probability one.*

Corollary 1 asserts that, the random sequence $\pi_{\theta_t}^\top r$ produced by Update 2 asymptotically approaches some finite value (since $\pi_\theta^\top r \in [0, 1]$), ruling out the possibility of divergence (oscillating forever). However, this does not necessarily imply that $\pi_{\theta_t}^\top r$ will approach the globally optimal policy value $r(a^*)$. A subtlety arises in bounding the stochastic NŁ coefficient in Lemma 1 away from 0.

A second key technique is required to show that $\pi_{\theta_t}(a^*)$ is non-vanishing.

Lemma 2 (Non-vanishing stochastic NŁ coefficient / "automatic exploration")**.** *Using Update 2 with the same assumptions as in Lemma 1, for an arbitrary initialization $\theta_1 \in \mathbb{R}^K$, we have,*

$$c := \inf_{t \geq 1} \pi_{\theta_t}(a^*) > 0, \qquad \textit{almost surely (a.s.).} \tag{7}$$

In particular, Lemma 2 further excludes the possibility that $\pi_{\theta_t}^\top r$ approaches values other than $r(a^*)$. In fact, using the special "variance-like" structure of the expected progress (Lemma 1), Corollary 1 implies that π_{θ_t} approaches a "generalized one-hot policy" as $t \to \infty$.

From these results, we can then argue by contradiction that π_{θ_t} cannot approach a sub-optimal "generalized one-hot policy" as $t \to \infty$. That is, we show that the probability of the union of sub-optimal actions cannot approach 1, which will imply that the optimal action's probability must approach 1 and achieve Eq. (7). Further details on the intuitions are presented later in Section 3, which reveal that the above arguments hold because Update 2 achieves a form of "automatic exploration" by using a baseline. Finally, by combining the two key results of Lemmas 1 and 2, we not only establish asymptotic convergence of NPG to a global optimum, but also establish a global convergence rate of $O(1/t)$ in terms of the sub-optimality gap.

Theorem 1 (Almost sure global convergence rate)**.** *Using Update 2 with on-policy sampling $a_t \sim \pi_{\theta_t}(\cdot)$, the IS estimator in Definition 1, and any initialization $\theta_1 \in \mathbb{R}^K$, there exists a $C > 0$ such that for all $t \geq 1$,*

$$\mathbb{E}_t[(\pi^* - \pi_{\theta_t})^\top r] \leq C/t, \qquad \textit{a.s.} \tag{8}$$

where we use $\mathbb{E}_t[\cdot]$ to denote $\mathbb{E}_t[\cdot|\mathcal{F}_t]$ for brevity, and $\mathcal{F}_t$ is the σ-algebra generated by $a_1, \ldots, a_{t-1}$.

2.3 General MDPs

Next, we generalize these results to finite Markov decision processes (MDPs). Given a finite set $\mathcal{X}$, let $\Delta(\mathcal{X})$ denote the set of all probability distributions on $\mathcal{X}$. A finite MDP is defined as a tuple $\mathcal{M} := (\mathcal{S}, \mathcal{A}, r, \mathcal{P}, \gamma)$, where $\mathcal{S}$ and $\mathcal{A}$ are finite state and action spaces, respectively. $r : \mathcal{S} \times \mathcal{A} \to \mathbb{R}$ is the expected reward function, $\mathcal{P} : \mathcal{S} \times \mathcal{A} \to \Delta(\mathcal{S})$ is the probability transition function, and $\gamma \in [0, 1)$ is the discount factor. We also extend Assumption 1 to every $(s, a) \in \mathcal{S} \times \mathcal{A}$ and assume there is a reward distribution $R_{s,a}$ with expectation $r(s, a)$, uniformly bounded within $[-R_{\max}, R_{\max}]$. Given a policy $\pi : \mathcal{S} \to \Delta(\mathcal{A})$, at each time $t \geq 0$, an agent is given a state $s_t \in \mathcal{S}$, takes an action $a_t \sim \pi(\cdot|s_t)$, then receives a scalar reward observation $x(s_t, a_t) \sim R_{s_t,a_t}$ and a next-state $s_{t+1} \sim \mathcal{P}(\cdot|s_t, a_t)$. The value function of π at state s is defined as

$$V^\pi(s) := \mathop{\mathbb{E}}_{\substack{a_t \sim \pi(\cdot|s_t), \\ s_{t+1} \sim \mathcal{P}(\cdot|s_t,a_t)}} \left[\sum_{t=0}^{\infty} \gamma^t r(s_t, a_t) \,\middle|\, s_0 = s \right]. \tag{9}$$

The policy optimization problem for a general MDP is to maximize the expected value of the policy,

$$\max_{\theta:\mathcal{S}\times\mathcal{A}\to\mathbb{R}} V^{\pi_\theta}(\rho) := \max_{\theta:\mathcal{S}\times\mathcal{A}\to\mathbb{R}} \mathop{\mathbb{E}}_{s\sim\rho(\cdot)} [V^{\pi_\theta}(s)], \tag{10}$$

where $\rho \in \Delta(\mathcal{S})$ is an initial state distribution, and $\pi_\theta(\cdot|s) = \mathrm{softmax}(\theta(s, \cdot))$,

$$\pi_\theta(a|s) = \frac{\exp\{\theta(s, a)\}}{\sum_{a' \in \mathcal{A}} \exp\{\theta(s, a')\}}, \text{ for all } (s, a) \in \mathcal{S} \times \mathcal{A}. \tag{11}$$

Given a policy π, its state-action value function is defined as $Q^\pi(s, a) := r(s, a) + \gamma \cdot \sum_{s'} \mathcal{P}(s'|s, a) \cdot V^\pi(s')$, and its advantage function is defined as $A^\pi(s, a) := Q^\pi(s, a) - V^\pi(s)$, for $(s, a) \in \mathcal{S} \times \mathcal{A}$. The state distribution of π is defined as $d^\pi_{s_0}(s) := (1 - \gamma) \cdot \sum_{t=0}^{\infty} \gamma^t \cdot \Pr(s_t = s|s_0, \pi, \mathcal{P})$. We also denote $d^\pi_\rho(s) := \mathbb{E}_{s_0\sim\rho(\cdot)} \left[d^\pi_{s_0}(s)\right]$. Given ρ, there exists an optimal policy π^* such that $V^{\pi^*}(\rho) = \max_{\pi:\mathcal{S}\to\Delta(\mathcal{A})} V^\pi(\rho)$. We denote $V^*(\rho) := V^{\pi^*}(\rho)$ for conciseness.

For a general MDP, we need to assume the initial state distribution μ is "sufficiently exploratory" [1, 14].

Assumption 2 (Sufficient exploration)**.** *The initial state distribution satisfies* $\min_s \mu(s) > 0$.

Algorithm 1 NPG, on-policy stochastic natural gradient

Input: Learning rate $\eta > 0$.
Output: Policies $\pi_{\theta_t} = \mathrm{softmax}(\theta_t)$.
Initialize parameter $\theta_1(s, a)$ for all $(s, a) \in \mathcal{S} \times \mathcal{A}$.
while $t \geq 1$ **do**
  Sample $s_t \sim d^{\pi_{\theta_t}}_\mu(\cdot)$, $a_t \sim \pi_{\theta_t}(\cdot|s)$ and $x_t \sim R_{a_t}$.
  Estimate $\hat{Q}^{\pi_{\theta_t}}(s_t, a_t)$ for $Q^{\pi_{\theta_t}}(s_t, a_t)$.
  $\theta_{t+1}(s_t, a_t) \leftarrow \theta_t(s_t, a_t) + \eta \cdot \frac{\hat{Q}^{\pi_{\theta_t}}(s_t,a_t) - V^{\pi_{\theta_t}}(s_t)}{\pi_{\theta_t}(a_t|s_t)}$.
end while

At iteration t, the NPG method uses the current state distribution to sample one state $s_t \sim d^{\pi_{\theta_t}}_\mu(\cdot)$, then uses on-policy sampling to sample one action $a_t \sim \pi_{\theta_t}(\cdot|s)$ and observe a reward $x(s_t, a_t) \sim R_{s_t,a_t}$. For the sampled state action pair $(s_t, a_t) \in \mathcal{S} \times \mathcal{A}$, an estimated state-action value $\hat{Q}^{\pi_{\theta_t}}(s_t, a_t)$ is then used to perform update, which can be viewed as generated from the value distribution of true state-action value $Q^{\pi_{\theta_t}}(s_t, a_t)$. The current state value function $V^{\pi_{\theta_t}}(s_t)$ is used as the baseline, as shown in Algorithm 1.

According to the performance difference lemma, we have, for all $\rho \in \Delta(\mathcal{S})$,

$$V^{\pi_{\theta_{t+1}}}(\rho) - V^{\pi_{\theta_t}}(\rho) = \frac{1}{1-\gamma} \cdot \sum_s d^{\pi_{\theta_{t+1}}}_\rho(s) \cdot \sum_a \left(\pi_{\theta_{t+1}}(a|s) - \pi_{\theta_t}(a|s)\right) \cdot Q^{\pi_{\theta_t}}(s, a), \tag{12}$$

where the inner summation over actions is similar to $\left(\pi_{\theta_{t+1}} - \pi_{\theta_t}\right)^\top r$ in one-state MDPs. This connection allows us to generalize Lemma 1 to the following result.

Lemma 3 (Stochastic NŁ)**.** *Using Algorithm 1 with learning rate,*

$$\eta_t = \frac{\pi_{\theta_t}(a_t|s_t) \cdot |Q^{\pi_{\theta_t}}(s_t, a_t) - V^{\pi_{\theta_t}}(s_t)|}{8 \cdot Q^2_{\max}}, \tag{13}$$

where $Q_{\max} = \frac{1}{1-\gamma} \cdot R_{\max}$ according to Assumption 1, we have, for all $t \geq 1$,

$$\mathbb{E}_t[V^{\pi_{\theta_{t+1}}}(\rho)] - V^{\pi_{\theta_t}}(\rho) \geq \frac{1}{1-\gamma} \cdot \frac{\eta}{2} \cdot \sum_s d_\mu^{\pi_{\theta_t}}(s) \cdot d_\rho^{\pi_{\theta_{t+1}}}(s) \cdot \sum_a \pi_{\theta_t}(a|s) \cdot (Q^{\pi_{\theta_t}}(s,a) - V^{\pi_{\theta_t}}(s))^2$$

$$\geq \frac{\eta}{2} \cdot (1-\gamma) \cdot \left\| \frac{d_\rho^{\pi^*}}{d_\mu^{\pi_{\theta_t}}} \right\|_\infty^{-1} \cdot \left\| \frac{1}{d_\rho^{\pi^*}} \right\|_\infty^{-1} \cdot \left\| \frac{d_\rho^{\pi^*}}{\rho} \right\|_\infty^{-1} \cdot \tilde{\Delta} \cdot \min_{s \in \mathcal{S}} \pi_{\theta_t}(a^*(s)|s) \cdot [V^*(\rho) - V^{\pi_{\theta_t}}(\rho)],$$

where $\mathbb{E}_t[\cdot]$ is on randomness from state sampling $s_t \sim d_\mu^{\pi_{\theta_t}}(\cdot)$, on-policy sampling $a_t \sim \pi_{\theta_t}(\cdot|s_t)$, and sampling $\hat{Q}^{\pi_{\theta_t}}(s_t, a_t)$ from the value distribution of $Q^{\pi_{\theta_t}}(s_t, a_t)$, and $a^*(s)$ is the action selected by the optimal policy π^* under state s, and $\tilde{\Delta} > 0$ for any "non-trivial" MDPs.

Next, similar to Lemma 2, we can develop a set of contradictions that establish the following result.

Lemma 4 (Non-vanishing stochastic NŁ coefficient / "automatic exploration")**.** *Using Algorithm 1 with the same assumptions as Lemma 3, with arbitrary initialization $\theta_1 \in \mathbb{R}^{\mathcal{S} \times \mathcal{A}}$, we have,*

$$c := \inf_{t \geq 1, s \in \mathcal{S}} \pi_{\theta_t}(a^*(s)|s) > 0, \qquad a.s. \tag{14}$$

By combining Lemmas 3 and 4, we obtain the following result that generalizes Theorem 1.

Theorem 2 (Almost sure global convergence rate)**.** *Using Algorithm 1 with any initialization $\theta_1 \in \mathbb{R}^K$, under the same assumptions as Lemmas 3, there exists a $C > 0$ such that for all $t \geq 1$,*

$$\mathbb{E}_t[V^{\pi^*}(\rho) - V^{\pi_{\theta_t}}(\rho)] \leq C/t, \qquad a.s. \tag{15}$$

where we use $\mathbb{E}_t[\cdot]$ to denote $\mathbb{E}_t[\cdot|\mathcal{F}_t]$ for brevity, and $\mathcal{F}_t$ is the σ-algebra generated by $(s_1, a_1), (s_2, a_2), \ldots, (s_{t-1}, a_{t-1})$.

3 Understanding Baselines in On-policy Stochastic Policy Optimization

Section 2 shows that using a value function baseline in on-policy stochastic NPG can ensure convergence to a globally optimal policy. However, the mechanism behind this finding requires further elucidation. Preliminary studies [3, 11] have observed that subtracting a baseline can reduce the committal behavior of PG-based estimators, suggesting that this effect might be more important than variance reduction. However, the concept of "committal behavior" was intuitive.

In this section, we provide a new, deeper understanding of how a baseline improves the convergence behaviour of a stochastic PG based method. To make the explanations clearer, we will use the deterministic reward setting to show the main ideas, even though the results generalize to stochastic reward settings. That is, we consider the stronger assumption that the true mean reward $r(a_t)$ is observed for sampled actions a_t, which leads to the simplified IS estimator.

Definition 2 (Simplified on-policy importance sampling (IS))**.** *At iteration t, sample one action $a_t \sim \pi_{\theta_t}(\cdot)$. The IS reward estimate is then constructed as $\hat{r}_t(a) = \frac{\mathbb{I}\{a_t = a\}}{\pi_{\theta_t}(a)} \cdot r(a)$ for all $a \in [K]$.*

3.1 Baselines Do Not Control Update Variance in NPG

We begin from the well known result that value baselines have no effect on exact policy gradients.

Proposition 2 (Unbiasedness of NPG)**.** *For NPG with and without a state value baseline, corresponding to Updates 1 and 2 respectively, we have $\mathbb{E}_{a_t \sim \pi_{\theta_t}(\cdot)}[\hat{r}_t] = \mathbb{E}_{a \sim \pi_{\theta_t}(\cdot)}[\hat{r}_t - \hat{b}_t] = r$.*

According to Proposition 2, Updates 1 and 2 become identical if the exact policy gradient is available, hence both enjoy an $O(e^{-c \cdot t})$ convergence rate to a global optimum ($c > 0$) [7, 11]. Therefore, a state value baseline can only have an effect if the policy gradient has to be estimated from a stochastic sample. However, we find that the variance of the NPG updates remains unbounded in the stochastic setting, regardless of whether a state value baseline is used.

Proposition 3 (Unboundedness of NPG)**.** *For NPG without a baseline, Update 1, we have $\mathbb{E}_{a_t \sim \pi_{\theta_t}(\cdot)} \|\hat{r}_t\|_2^2 = \sum_{a \in [K]} \frac{r(a)^2}{\pi_{\theta_t}(a)}$. For NPG with a state value baseline, Update 2, we have $\mathbb{E}_{a_t \sim \pi_{\theta_t}(\cdot)} \|\hat{r}_t - \hat{b}_t\|_2^2 = \sum_{a \in [K]} \frac{(r(a) - \pi_{\theta_t}^\top r)^2}{\pi_{\theta_t}(a)} - K \cdot (\pi_{\theta_t}^\top r)^2 + 2 \cdot (\pi_{\theta_t}^\top r) \cdot (r^\top \mathbf{1})$.*

According to Proposition 3, whenever π_{θ_t} nears a one-hot probability distribution over $[K]$ (which it must converge to), there exists at least one action $a \in [K]$ such that both $\frac{r(a)^2}{\pi_{\theta_t}(a)}$ and $\frac{(r(a)-\pi_{\theta_t}^\top r)^2}{\pi_{\theta_t}(a)}$ become unbounded, implying an unbounded scale for both Updates 1 and 2. Yet we know from Proposition 1 that not using a baseline fails with positive probability, while from Theorem 1 subtracting a state value baseline ensures almost sure convergence to a global optimum. The fact that the variance of both updates is unbounded suggests that it is difficult to draw conclusions on the effect of the baseline from a variance reduction perspective alone. An alternative analysis is required to explain the fundamental difference between Updates 1 and 2.

3.2 Coupled Sampling and Updating

In on-policy stochastic policy optimization, sampling and updating are coupled as shown in Figure 1. At iteration t, the data collected depends on the current policy, since on-policy sampling is used $a_t \sim \pi_{\theta_t}(\cdot)$, while the policy is updated from the observations collected based on a_t. This coupling introduces complexity in the optimization process as well as in the analysis. However, this coupling is also fundamental to understanding the circular interaction created by any on-policy stochastic optimization method. That is, on-policy stochastic optimization faces an exploration-exploitation dilemma: a learning algorithm can improve the policy and increase the probability of choosing actions that yield higher rewards (exploitation), but it must not do so too aggressively lest it fail to identify possibly higher-reward actions (exploration). Striking a proper balance between exploration and exploitation is key to achieving good convergence properties. Different levels of update aggression create different circular effects between sampling and updating, which is central to determining almost sure convergence to a global optimum.

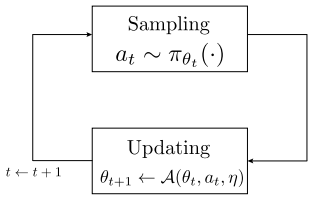


Figure 1: Coupled on-policy sampling and updating.

3.3 The "Vicious Circle" of Being Too Aggressive

First we illustrate a negative effect, the "vicious circle" of being too aggressive.

Lemma 5 (Bad sampling)**.** *Let $\pi_{\theta_t}(a) \in (0,1)$ be the probability of sampling action a using online sampling $a_t \sim \pi_{\theta_t}(\cdot)$, for all $t \geq 1$. If $1 - \pi_{\theta_t}(a) \in o(1/t)$, then $\prod_{t=1}^{\infty} \pi_{\theta_t}(a) > 0$.*

Note that Lemma 5 characterizes sampling behaviour under general conditions that do not otherwise depend on specific updates. However, according to Lemma 5, if an action's probability approaches 1 strictly faster than $O(1/t)$, by whatever means, it becomes possible to not sample any other action forever, which creates a "lack of exploration" phenomenon as it is known in RL. In particular, on-policy stochastic NPG without a baseline can produce such a sequence of $\{\pi_{\theta_t}(a)\}_{t\geq 1}$.

Lemma 6 (NPG aggressiveness)**.** *Fix sampling $a_t = a$ for all $t \geq 1$, using Update 1 with constant learning rate $\eta > 0$, we have $1 - \pi_{\theta_t}(a) \in O(e^{-c \cdot t})$ for all $t \geq 1$, where $c > 0$.*

Note that Lemma 6 only characterizes the aggressiveness of Update 1 with the sampling fixed to be $a_t = a$ for all $t \geq 1$. Lemmas 5 and 6 together describe the "vicious circle" between sampling and updating that can be created by overly aggressive updates. First, in on-policy sampling, there will always be a non-zero probability of "bad luck"; that is, with positive probability a set of sub-optimal actions can be sequentially sampled for multiple steps. Second, an overly aggressive update will only exaggerate the weakness of the sampling procedure by increasing the sampled sub-optimal actions' probabilities rapidly (Lemma 6). Third, this exaggeration can worsen data collection for subsequent updating by further increasing the prevalence of sub-optimal actions. Such a vicious circular interaction between sampling and updating can happen repeatedly, and its self-reinforcing nature can create a non-zero probability that the cycle occurs forever (Lemma 5), resulting in convergence to a sub-optimal deterministic policy (a stationary point for both sampling and updating).

3.4 The "Virtuous Circle" of Not Being Too Aggressive

Next, we demonstrate a positive effect, the "virtuous circle" of not being too aggressive.

Lemma 7 (Good sampling)**.** *Let $\pi_{\theta_t}(a) \in (0,1)$ be the probability of sampling action a using online sampling $a_t \sim \pi_{\theta_t}(\cdot)$, for all $t \geq 1$. If $1 - \pi_{\theta_t}(a) \in \Omega(1/t)$, then $\prod_{t=1}^{\infty} \pi_{\theta_t}(a) = 0$.*

As in Lemma 5, Lemma 7 only characterizes the effect of sampling behaviour under general conditions that do not otherwise depend on specific updates. Here we see that if an action's probability approaches 1 no faster than $O(1/t)$, it is no longer possible to avoid sampling any other action forever; that is, sufficiently slow modification of the sampling probabilities forces persistent exploration such that every action is sampled within some finite time with probability 1. In particular, subtracting a value baseline in on-policy stochastic NPG produces such a sequence $\{\pi_{\theta_t}(a)\}_{t\geq 1}$.

Lemma 8 (Value baselines reduce NPG aggressiveness)**.** *Fix sampling $a_t = a$ for all $t \geq 1$. Then using Update 2 with a constant learning rate $\eta > 0$ obtains $1 - \pi_{\theta_t}(a) \in \Omega(1/t)$ for all $t \geq 1$.*

As in Lemma 6, Lemma 8 only characterizes the conservativeness of Update 2 with fixed sampling of $a_t = a$ for all $t \geq 1$. Lemmas 7 and 8 now describe a "virtuous circle" between sampling and updating that is created by using not too aggressive updates. First, even in a worst case situation (e.g., an adversarial initialization), where a sub-optimal action has a dominant probability $\pi_{\theta_t}(a) \approx 1$, under on-policy sampling all actions will eventually be sampled. Second, conservative updating will mitigate the effect of the extreme sampler by not increasing the sub-optimal action's probability too rapidly (Lemma 8). Third, sustained diversity in sampling will eventually draw a better action than the current dominating sub-optimal action (Lemma 7). Finally, once better actions are sampled, the update will improve subsequent sampling by decreasing the probability of the dominating sub-optimal action. In particular, this is achieved by increasing value baselines to be larger than the dominating sub-optimal action's true mean reward, such that the dominating sub-optimal action will start losing probabilities. This virtuous circular interaction between sampling and updating ensures sufficient exploration, which prevents the iteration from converging to a sub-optimal deterministic policy.

3.5 How a State Value Baseline Reduces Update Aggressiveness

Based on Lemmas 5 and 7, the boundary between "too aggressive" and "not too aggressive" is precisely $\Theta(1/t)$. We now explain how a state value baseline in NPG will control update aggressiveness. First, without a baseline, sampling a sub-optimal action $a \in [K]$ for t times makes its parameter behave as $\theta_t(a) \in \Theta(t)$, since $r(a) \in \Theta(1)$. On the other hand, other action parameters will behave as $\theta_t(a') \in \Theta(1)$ if they are only sampled a constant number of times. Under the softmax parameterization Eq. (2), this will imply that $1 - \pi_{\theta_t}(a) \in O(e^{-c\cdot t})$, which is far too aggressive. Second using a state value baseline, under repeated sampling the parameter increase for a sub-optimal action $a \in [K]$ will be damped. In particular, whenever the policy is close to deterministic, say $\pi_{\theta_t}(a) \approx 1$, we also have $\pi_{\theta_t}^\top r \approx r(a)$. Therefore, since

$$r(a) - \pi_{\theta_t}^\top r = \sum_{a' \neq a} \pi_{\theta_t}(a') \cdot (r(a) - r(a')) \leq 1 - \pi_{\theta_t}(a), \tag{16}$$

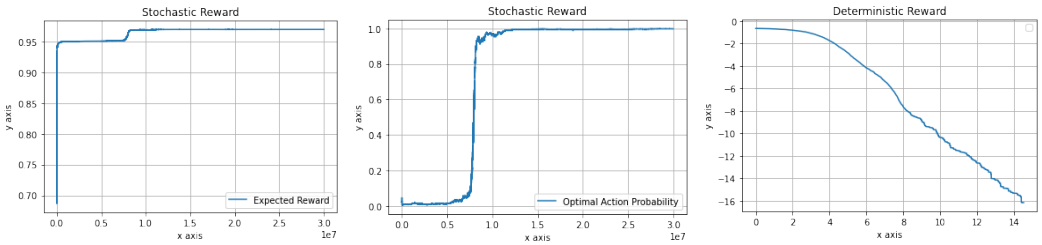
the closer $1 - \pi_{\theta_t}(a)$ is to 0, the smaller $r(a) - \pi_{\theta_t}^\top r$ will be. This means even if a is sampled repeatedly for t times, we obtain $\theta_t(a) \in O(\log t)$ and $1 - \pi_{\theta_t}(a) \in \Omega(1/t)$ (Lemma 8). Thus, the effect of baseline is to modify the sampling to lie exactly on the boundary of being good enough. From this argument the key role of the value baseline is to reduce update aggressiveness to achieve a particular effect on long-term sampling, rather than simply reduce variance. It also shows how using an appropriately un-aggressive update is both necessary (Lemma 5) and sufficient (Lemma 7) to achieve almost sure convergence to a global optimum in on-policy stochastic policy optimization.

4 Simulations

We conducted simulations to verify the two main results above: asymptotic convergence toward globally optimal policy π^* in Lemma 2, and the $O(1/t)$ convergence rate in Theorem 1.

4.1 Asymptotic Convergence

We first considered a one-state MDP with $K = 20$ actions and true mean reward vector $r \in (0,1)^K$, where the optimal action is $a^* = 1$ with true mean reward $r(1) \approx 0.97$ and best sub-optimal



(a) $\pi_{\theta_t}^\top r$. (b) $\pi_{\theta_t}(a^*)$. (c) $\log\left(\pi^* - \pi_{\theta_t}\right)^\top r$.

Figure 2: Adversarial initialization (a) and (b); uniform initialization (c).

action's true mean reward $r(2) \approx 0.95$. The sampled reward is observed with a large noise, e.g., $x \approx -2.03$ and $x \approx 3.97$ with both 0.5 probability for the optimal action, such that $r(1) \approx 0.5 \cdot (-2.03) + 0.5 \cdot 3.97$. Details about r and the reward distributions can be found in the appendix.

To verify asymptotic convergence to a globally optimal policy in Lemma 2, we consider the iteration behaviors of Update 2 under an adversarial initialization, where $\pi_{\theta_1}(2) \approx 0.88$, i.e., a sub-optimal action starts with a dominating probability. This is the worst case scenario for Lemma 2, where the optimal action only has a small chance to be sampled, while the sampled reward noise is very large.

As shown in Figure 2a, the expected reward $\pi_{\theta_t}^\top r$ quickly approaches and remains stuck around $r(2) \approx 0.95$ initially, as expected. However, after about 8×10^6 iterations, the policy π_{θ_t} finally escapes the sub-optimal plateau and approaches the optimal reward $r(1) \approx 0.97$. This simulation result is consistent with Lemma 2, i.e., for an arbitrary initialization, the introduction of a value baseline eventually makes π_{θ_t} approach a globally optimal policy within finite time, while additionally the optimal action's probability never vanishes, $\inf_{t \geq 1} \pi_{\theta_t}(a^*) > 0$, as shown in Figure 2b.

4.2 Convergence Rate

We run Update 2 with a uniform initialization, i.e., $\pi_{\theta_1}(a) = 1/K$ for all $a \in [K]$, and calculate averaged sub-optimality gap $\left(\pi^* - \pi_{\theta_t}\right)^\top r$ across 100 independent runs. As shown in Figure 2c, where both axes are in $\log$ scale, the slope is aproximately -1, indicating that $\log\left(\pi^* - \pi_{\theta_t}\right)^\top r = -\log t + C$, or equivalently $\left(\pi^* - \pi_{\theta_t}\right)^\top r = C'/t$, which is consistent with Theorem 1.

5 Conclusion

This work clarifies some of the longstanding mysteries the have separated the theory and practice of policy gradient optimization. The major finding is a state value baseline reduces the aggressiveness of the on-policy stochastic NPG update, which turns out to be necessary and sufficient for achieving almost sure convergence to a global optimum. The deeper understanding of the circular dependence between on-policy sampling and updating also dispels a common misconception about variance reduction, showing that bounded variance estimators are not necessary for achieving global convergence. The main technical innovation is the stochastic NŁ inequality, and the subsequent arguments that establish global convergence, both of which depend critically on the value baseline.

This work leaves open a number of interesting questions. *First*, the $O(1/t)$ convergence rate contains an initialization dependent constant in Lemma 2, resulting from plateaus as observed in Figure 2a, which does not appear in results that use the direct parameterization [4]. Thus the difficulty appears due to the non-linear softmax transform. Removing or improving this constant would impact practical performance, so investigating other techniques, such as regularization, optimism or momentum might be helpful. *Second*, the results in this paper use the true state values as the baselines. It would be interesting to consider the effect of estimating the value baseline or using alternative baselines in policy optimization. *Finally*, the $O(1/t)$ last iteration convergence rate implies an optimal $O(\log T)$ regret in stochastic bandit problems [8]. The explanation of the circular dependence between sampling and updating is specific to on-policy PG optimization, but it is also consistent with the exploration exploitation dilemma in RL. In other words, this work suggests a completely new approach to the exploration-exploitation trade-off, achieving provable bounds with ever requiring explicit uncertainty estimates, nor any concrete instantiation of the principle of optimism under uncertainty.

References

[1] Alekh Agarwal, Sham M Kakade, Jason D Lee, and Gaurav Mahajan. On the theory of policy gradient methods: Optimality, approximation, and distribution shift. *Journal of Machine Learning Research*, 22(98):1–76, 2021.

[2] Shicong Cen, Chen Cheng, Yuxin Chen, Yuting Wei, and Yuejie Chi. Fast global convergence of natural policy gradient methods with entropy regularization. *Operations Research*, 2021.

[3] Wesley Chung, Valentin Thomas, Marlos C Machado, and Nicolas Le Roux. Beyond variance reduction: Understanding the true impact of baselines on policy optimization. *arXiv preprint arXiv:2008.13773*, 2020.

[4] Denis Denisov and Neil Walton. Regret analysis of a markov policy gradient algorithm for multi-arm bandits. *arXiv preprint arXiv:2007.10229*, 2020.

[5] Tuomas Haarnoja, Aurick Zhou, Pieter Abbeel, and Sergey Levine. Soft actor-critic: Off-policy maximum entropy deep reinforcement learning with a stochastic actor. In *International Conference on Machine Learning*, pages 1861–1870, 2018.

[6] Sham M Kakade. A natural policy gradient. In *Advances in Neural Information Processing Systems*, pages 1531–1538, 2002.

[7] Sajad Khodadadian, Prakirt Raj Jhunjhunwala, Sushil Mahavir Varma, and Siva Theja Maguluri. On the linear convergence of natural policy gradient algorithm. *arXiv preprint arXiv:2105.01424*, 2021.

[8] Tze Leung Lai, Herbert Robbins, et al. Asymptotically efficient adaptive allocation rules. *Advances in applied mathematics*, 6(1):4–22, 1985.

[9] Guanghui Lan. Policy mirror descent for reinforcement learning: Linear convergence, new sampling complexity, and generalized problem classes. *arXiv preprint arXiv:2102.00135*, 2021.

[10] Gen Li, Yuting Wei, Yuejie Chi, Yuantao Gu, and Yuxin Chen. Softmax policy gradient methods can take exponential time to converge. In *Conference on Learning Theory*, pages 3107–3110. PMLR, 2021.

[11] Jincheng Mei, Bo Dai, Chenjun Xiao, Csaba Szepesvari, and Dale Schuurmans. Understanding the effect of stochasticity in policy optimization. *arXiv preprint arXiv:2110.15572*, 2021.

[12] Jincheng Mei, Yue Gao, Bo Dai, Csaba Szepesvari, and Dale Schuurmans. Leveraging non-uniformity in first-order non-convex optimization. In *International Conference on Machine Learning*, pages 7555–7564. PMLR, 2021.

[13] Jincheng Mei, Chenjun Xiao, Bo Dai, Lihong Li, Csaba Szepesvári, and Dale Schuurmans. Escaping the gravitational pull of softmax. *Advances in Neural Information Processing Systems*, 33:21130–21140, 2020.

[14] Jincheng Mei, Chenjun Xiao, Csaba Szepesvari, and Dale Schuurmans. On the global convergence rates of softmax policy gradient methods. In *International Conference on Machine Learning*, pages 6820–6829. PMLR, 2020.

[15] Arkadi Nemirovski, Anatoli Juditsky, Guanghui Lan, and Alexander Shapiro. Robust stochastic approximation approach to stochastic programming. *SIAM Journal on optimization*, 19(4):1574–1609, 2009.

[16] Ben Recht. Updates on policy gradients. *http://www.argmin.net/2018/03/13/pg-saga/*, 2018.

[17] John Schulman, Sergey Levine, Pieter Abbeel, Michael Jordan, and Philipp Moritz. Trust region policy optimization. In *International Conference on Machine Learning*, pages 1889–1897, 2015.

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

[19] Richard S Sutton, David A McAllester, Satinder P Singh, and Yishay Mansour. Policy gradient methods for reinforcement learning with function approximation. In *Advances in Neural Information Processing Systems*, pages 1057–1063, 2000.

[20] George Tucker, Surya Bhupatiraju, Shixiang Gu, Richard Turner, Zoubin Ghahramani, and Sergey Levine. The mirage of action-dependent baselines in reinforcement learning. In *International conference on machine learning*, pages 5015–5024. PMLR, 2018.

[21] Junyu Zhang, Chengzhuo Ni, Zheng Yu, Csaba Szepesvari, and Mengdi Wang. On the convergence and sample efficiency of variance-reduced policy gradient method. *arXiv preprint arXiv:2102.08607*, 2021.

[22] Junzi Zhang, Jongho Kim, Brendan O'Donoghue, and Stephen Boyd. Sample efficient reinforcement learning with reinforce. *arXiv preprint arXiv:2010.11364*, 2020.

[23] Runyu Zhang, Jincheng Mei, Bo Dai, Dale Schuurmans, and Na Li. On the effect of log-barrier regularization in decentralized softmax gradient play in multiagent systems. *arXiv preprint arXiv:2202.00872*, 2022.

Checklist

1. For all authors...
   (a) Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope? [Yes]
   (b) Did you describe the limitations of your work? [Yes]
   (c) Did you discuss any potential negative societal impacts of your work? [N/A]
   (d) Have you read the ethics review guidelines and ensured that your paper conforms to them? [Yes]
2. If you are including theoretical results...
   (a) Did you state the full set of assumptions of all theoretical results? [Yes]
   (b) Did you include complete proofs of all theoretical results? [Yes]
3. If you ran experiments...
   (a) Did you include the code, data, and instructions needed to reproduce the main experimental results (either in the supplemental material or as a URL)? [Yes]
   (b) Did you specify all the training details (e.g., data splits, hyperparameters, how they were chosen)? [Yes]
   (c) Did you report error bars (e.g., with respect to the random seed after running experiments multiple times)? [Yes]
   (d) Did you include the total amount of compute and the type of resources used (e.g., type of GPUs, internal cluster, or cloud provider)? [N/A]
4. If you are using existing assets (e.g., code, data, models) or curating/releasing new assets...
   (a) If your work uses existing assets, did you cite the creators? [N/A]
   (b) Did you mention the license of the assets? [N/A]
   (c) Did you include any new assets either in the supplemental material or as a URL? [N/A]
   (d) Did you discuss whether and how consent was obtained from people whose data you're using/curating? [N/A]
   (e) Did you discuss whether the data you are using/curating contains personally identifiable information or offensive content? [N/A]
5. If you used crowdsourcing or conducted research with human subjects...
   (a) Did you include the full text of instructions given to participants and screenshots, if applicable? [N/A]
   (b) Did you describe any potential participant risks, with links to Institutional Review Board (IRB) approvals, if applicable? [N/A]
   (c) Did you include the estimated hourly wage paid to participants and the total amount spent on participant compensation? [N/A]