Proceedings of Machine Learning Research vol 242:1–13, 2024

A Large Deviations Perspective on Policy Gradient Algorithms

Wouter Jongeneel WOUTER.JONGENEEL@EPFL.CH
EPFL

Daniel Kuhn DANIEL.KUHN@EPFL.CH
EPFL

Mengmeng Li MENGMENG.LI@EPFL.CH
EPFL

Abstract

Motivated by policy gradient methods in the context of reinforcement learning, we identify a large deviation rate function for the iterates generated by stochastic gradient descent for possibly non-convex objectives satisfying a Polyak-Łojasiewicz condition. Leveraging the contraction principle from large deviations theory, we illustrate the potential of this result by showing how convergence properties of policy gradient with a softmax parametrization and an entropy regularized objective can be naturally extended to a wide spectrum of other policy parametrizations.

1. Introduction and related work.

Policy gradient methods are at the core of several reinforcement learning (RL) algorithms. *e.g.*, see Sutton et al. (1999); Degris et al. (2012); Agarwal et al. (2021). As such, a wide adoption of these algorithms is aided by precisely understanding their performance. To that end, in scenarios where policy gradients can only be evaluated stochastically, it is crucial to understand the underlying distribution governing the evolution of policy iterates. Despite the popularity of stochastic policy gradient methods for over 20 years, global convergence has been only understood recently (Fazel et al., 2018; Zhang et al., 2020; Bhandari and Russo, 2019; Agarwal et al., 2021). Although global convergence proofs are actively developed for different policy gradient algorithm templates (Bhandari and Russo, 2019; Agarwal et al., 2021; Cen et al., 2022), two aspects of the current convergence analysis would benefit from further study: 1) global convergence guarantees are often stated in terms of *expected* suboptimality; and 2) the choice of policy parametrizations impacts convergence behavior (Mei et al., 2020a). In light of these observations, this paper aims to derive sharp convergence rates in *probability* and to provide a unifying approach towards understanding the effect of different policy parametrizations. As we build upon a rich history of work, we succinctly comment on: (i) *high probability analysis* in stochastic optimization and RL; and (ii) *policy gradient* methods in RL.

(i) Despite stochastic gradient descent (SGD) being over 70 years old (Robbins and Monro, 1951), there has been a surge of interest in terms of *high probability analysis* for (SGD): for strongly-convex objectives (Harvey et al., 2019); non-convex objectives satisfying a Polyak-Łojasiewicz (PL) condition (Madden et al., 2020); and non-convex objectives (Ghadimi and Lan, 2013; Liu et al., 2023) with Lipschitz gradients. A motivation for studying tight convergence bounds in terms of probabilities over convergence in expectation is due to the fact that the iterates generated by SGD can be brittle and *large deviations* from the expected value can occur, *e.g.*, see Gower et al. (2020) and references therein. Hence, a high-probability bound is attractive since it reassures the practitioner that certain behavior occurs at least with a close-to-1 probability for a single realization of iterates.

(ii) *Policy gradient* methods can be interpreted as gradient descent applied to the policy optimization problem in RL (Sutton and Barto, 2018, Ch. 13), see in particular (Sutton and Barto, 2018, p. 337) for historical remarks. Therefore, when the state space is large (Bottou, 2010) and due to necessary approximation, one oftentimes resorts to stochastic variants, whose convergence guarantees are often stated in terms of expectations (Lan, 2023), with the exception of (Ding et al., 2021), see also (Madden et al., 2020, § 2.2). The high-probability convergence rate provided in Ding et al. (2021) can be seen as a loosened bound of the *large deviation rate* we aim to derive. The pursuit of a tight high-probability concentration bound for the iterations produced by stochastic policy gradient methods holds significant potential for bolstering the practical implementation and interpretability of policy gradient methods, thereby improving the overall applicability of RL. Studying the concentration behavior of SGD iterates through the lens of large deviations is pioneered by Bajovic et al. (2023). It is worth mentioning that, however, their analysis exploits strong convexity of the optimization objective. Then, a first step towards understanding more precisely the concentration behavior of stochastic policy gradient iterates would be to study its entropy-regularized variant, which already faces the difficulties of non-convexity and non-uniformity of the PL constant.

As the theory of *large deviations* is the key tool in this paper, we briefly introduce it already at this stage. Let $X_t \in \mathbb{R}^d$ be the t^{th} iterate of some stochastic algorithm aimed at driving X_t to $x^\star$ for $t \to +\infty$, then, many aforementioned works provide statistical guarantees of the form $\mathbb{P}(\|X_t - x^\star\|_2 \leq \varepsilon) \geq 1 - \beta\ \forall t \geq T$ for an appropriately chosen triple (T, ε, β). Although often impressive pieces of work, there are some remarks to be made. First, it is often not clear if these high-probability bounds are *tight*. Secondly, the discrepancy is often measured by a sufficiently simple function, like the ℓ_2-norm in this case, however, in practice one might be interested in vastly different sublevel sets, *e.g.*, sublevel sets capturing harmful events, the ones that are of interest in safe RL. Third, the guarantees are oftentimes states for particular policy parametrizations, a unifying framework that generalizes performance analysis across different policy parametrizations remains largely unexplored. The theory of *large deviations* is precisely powerful when one aims to address these aforementioned points. Let us clarify terminology and provide intuition, yet, while skipping some details. A sequence of finite (probability) measures $(\mu_t)_t$ on $\mathbb{R}^d$ is said to satisfy a *large deviation principle* (LDP) with *rate function* $I : \mathbb{R}^d \to [0, +\infty]$ when for any Borel set $\Theta \subseteq \mathbb{R}^d$

$$-\mathring{r} := -\inf_{\theta \in \mathring{\Theta}} I(\theta) \leq \liminf_{t \to +\infty} \frac{1}{t} \log \mu_t(\Theta) \leq \limsup_{t \to +\infty} \frac{1}{t} \log \mu_t(\Theta) \leq -\inf_{\theta \in \overline{\Theta}} I(\theta) =: -\overline{r}, \tag{1}$$

where $\mathring{\Theta}$ denotes the interior of Θ, and $\overline{\Theta}$ denotes the closure of Θ. Now, identifying $(\mu_t)_t$ with a sequence of random variables $(X_t)_t$, we have after rearranging (1) that $\exp(-\mathring{r}t + o(t)) \leq \mathbb{P}(X_t \in \Theta) \leq \exp(-\overline{r}t + o(t))$. As such, an LDP captures—possibly tight, *e.g.*, under mild topological assumptions—convergence rates. Indeed, regarding applicability, once the rate function is identified, the corresponding inaccuracy rate can be obtained for *any* new given region of interest. For a detailed exposition of large deviations, we refer to an overview paper by Varadhan (2008) and the book by Dembo and Zeitouni (2009). As alluded to above, we will use the theory of large deviations to study stochastic policy gradient iterates, that is, to study the iterates of a non-convex optimization algorithm. Although large deviation tools appeared in seminal work on non-convex optimization Ghadimi and Lan (2013), that work does not provide an LDP. What is more, large deviations are intimately connected to rare events, whose analysis is of interest to RL (Frank et al., 2008). Also, efficient adaptive sampling techniques become available thanks to careful deployment of large deviation theory (Dupuis and Wang, 2004), an approach that is recently attracting interest in the context of SGD (Lahire, 2023).

Contributions.

(i) We find, with high probability, an lower bound on the rate function for iterates generated by softmax policy gradient with an entropy regularized objective (4), see Theorem 5. We also recover results similar to Madden et al. (2020), yet, via large deviation theory, see Lemma 3.

(ii) We demonstrate the wide applicability of having established a large deviation rate by leveraging the contraction principle to establish large deviation rates for a wide range of tabular policy parametrizations beyond the softmax parametrization.

(iii) Effectively, we establish a LDP upper bound for SGD under a PL condition, which is of independent interest.

In Section 2 we describe the RL setting under consideration whereas in Section 3 we derive a preliminary exponential bound on the convergence of the value function in probability. Then, in Section 4 we provide our main result: a LDP upper bound for the stochastic policy gradient iterates. At last, we show in Section 5 that this LDP upper bound can be lifted from a softmax parametrization to a whole family of parametrizations, which directly leads to exponential convergence rates for existing and new policy parametrizations, with high probability.

2. Problem statement.

Markov decision process (MDP). We consider a finite MDP given by a six-tuple $(\mathcal{S}, \mathcal{A}, P, c, \gamma, \rho)$ consisting of a finite state space $\mathcal{S} = \{1, \dots, S\}$, a finite action space $\mathcal{A} = \{1, \dots, A\}$, a transition kernel $P : \mathcal{S} \times \mathcal{A} \to \mathcal{P}(\mathcal{S})$, a cost-per-stage function $c : \mathcal{S} \times \mathcal{A} \to \mathbb{R}$, assumed to be bounded[1], a discount factor $\gamma \in (0, 1)$, and an initial distribution $\rho \in \mathcal{P}(\mathcal{S})$. Here, we use $\mathcal{P}(\mathcal{S}) := \{p \in \mathbb{R}^{|\mathcal{S}|} : \sum_{i=1}^{|\mathcal{S}|} p_i = 1, p \geq 0\}$ to denote the probability simplex over $\mathcal{S}$. Throughout the rest of the paper we restrict attention to stationary policies, which are described by a stochastic kernel $\pi \in \Pi := \mathcal{P}(\mathcal{A})^{|\mathcal{S}|}$. The value function $V^\pi : \mathcal{P}(\mathcal{S}) \to \mathbb{R}$ associated with π is defined through

$$V^\pi(\rho) := \mathbb{E}\big[\textstyle\sum_{k=0}^{\infty} \gamma^k c\left(s_k, a_k\right) | s_0 \sim \rho, a_k \sim \pi\left(\cdot|s_k\right), s_{k+1} \sim P(\cdot|s_k, a_k)\big], \tag{2}$$

and the main objective is to minimize $V^\pi(\rho)$ across all $\pi \in \Pi$.

Entropy-regularized reinforcement learning (RL). It is sometimes useful to work with the modified objective where a regularizer is added, that is, $V_\tau^\pi(\rho) := V^\pi(\rho) + \tau \cdot \mathbb{H}(\rho, \pi)$ (Neu et al., 2017; Mei et al., 2020b), where $\tau > 0$ and

$$\mathbb{H}(\rho, \pi) := \mathbb{E}\big[\textstyle\sum_{k=0}^{\infty} -\gamma^k \log \pi\left(a_k|s_k\right) | s_0 \sim \rho, a_k \sim \pi\left(\cdot|s_k\right), s_{k+1} \sim P\left(\cdot|s_k, a_k\right)\big].$$

An optimal policy $\pi^\star \in \Pi$ is defined through the condition $V_\tau^{\pi^\star}(\rho) \leq V_\tau^\pi(\rho)$ for all $\pi \in \Pi$.

[1] One can set the cost of a particular stage to be ∞ but that results in an infeasible/unbounded optimization problem.

Softmax policy gradient for entropy-regularized objective. The softmax transform $\pi_\theta(\cdot|s) := \mathrm{softmax}(\theta(s,\cdot))$ of any function $\theta : \mathcal{S} \times \mathcal{A} \to \mathbb{R}$ is defined through

$$\pi_\theta(a|s) = \frac{\exp\left(\theta(s,a)\right)}{\sum_{a' \in \mathcal{A}} \exp\left(\theta\left(s,a'\right)\right)} \quad \forall a \in \mathcal{A}. \tag{3}$$

In the remainder, we frequently think of θ as a policy parameter in $\mathbb{R}^{d=|\mathcal{S}||\mathcal{A}|}$ and we use $V_\tau^\theta(\rho)$ for $V_\tau^{\pi_\theta}(\rho)$. By (Nachum et al., 2017, Thm. 3) there exists $\theta^\star \in \mathbb{R}^d$ such that $\pi_{\theta^\star}$ is the optimal policy of the entropy-regularized MDP. By the policy gradient theorem (Sutton et al., 1999), the policy gradient $g(\theta_t) := \partial V_\tau^\theta(\rho)/\partial\theta|_{\theta=\theta_t}$ is given by $\partial V_\tau^\theta(\rho)/\partial\theta(s,a) = 1/(1-\gamma) \cdot d_\rho^\theta(s) \cdot \pi_\theta(a|s) \cdot A_\tau^\theta(s,a)$, where d_ρ^θ denotes the discounted state-visitation frequency measure, and A_τ^θ is the advantage function associated with the regularized objective, see also (Mei et al., 2020b, Lem. 1). In practice, however, computing the advantage function is computationally expensive. It is therefore convenient to work with estimates of the exact gradient. Throughout the rest of the paper we assume unbiased stochastic sample access to $\partial_\theta V_\tau^{\pi_\theta}(\rho)|_{\theta=\theta'}$, denoted as $\tilde{g}(\theta')$. Stochastic softmax policy gradient with entropy-regularized objective thus suggests to update policy parameters via

$$\theta_{t+1} \leftarrow \theta_t - \eta_t \cdot \tilde{g}(\theta_t). \tag{4}$$

Assumption 1 (Warm start with sufficient exploration) *The initial state distribution satisfies* $\min_{s\in\mathcal{S}} \rho(s) > 0$. *In addition, the initial iterate θ_1 is chosen around a Δ-neighborhood of $\theta^\star$ in the sense that* $\min_{\theta^\star\in\Theta^\star} \|\theta_1 - \theta^\star\|_2 \leq \Delta$ *where $\Theta^\star$ is the set of all optimal solutions.*

The above assumption simplifies the exposition and such an initial θ_1 can be obtained by a stochastic policy gradient method such as (Ding et al., 2021, Alg. 3.2). We additionally define $Z_t := g(\theta_t) - \tilde{g}(\theta_t)$. As in (Bajovic et al., 2023) we make the following assumption.

Assumption 2 (Gradient estimation uncertainty) *The stochastic process $(Z_t)_t$ satisfies: (i) Z_t depends on the past only through θ_t; (ii) $\mathbb{E}[Z_t|\theta_t = \theta] = 0$ for any θ and (iii) the distribution of Z_t is independent of the iterate index $t \in \mathbb{N}$.*

By Assumption 2 (i), the process $(\theta_t)_{t\geq 0}$ generated by (4) is a Markov chain. We use $\Lambda(\lambda;\theta) = \log \mathbb{E}[\exp(\langle\lambda, Z_t\rangle) \mid \theta_t = \theta]$ to denote the conditional *log-moment generating function* (LMGF) of $Z_t \in \mathbb{R}^d$. We also define $M(\nu;\theta) = \mathbb{E}[\exp(\nu\|Z_t\|_2^2) \mid \theta_t = \theta]$ as the *conditional moment-generating function* (MGF) of $\|Z_t\|_2^2$. Now, we impose the following technical assumption to ensure that the tail probabilities of the disturbances $(Z_t)_t$ decay sufficiently fast.

Assumption 3 (Sub-Gaussian process) *All elements of the sequence $(Z_t)_t$ follow a σ-sub-Gaussian distribution for some $\sigma > 0$, i.e.,*

$$\Lambda(\lambda;\theta) \leq \tfrac{1}{2}\sigma^2\|\lambda\|_2^2 \quad \forall \lambda, \theta \in \mathbb{R}^d. \tag{5}$$

A Monte-Carlo gradient estimation method that satisfies the above assumptions is proposed by Ding et al. (2021). The method operates by generating trajectories of the MDP with policy parameter θ and then estimate the advantage function based on these trajectories.

Lemma 1 (L_1-smoothness of the value function (Mei et al., 2020b, Lem. 7 & 14)) *There exists a constant $L_1 > 0$ such that*

$$|V_\tau^{\pi_\theta}(\rho) - V_\tau^{\pi_{\theta'}}(\rho) - \langle g(\theta'), \theta - \theta'\rangle| \leq \tfrac{1}{2}L_1\|\theta - \theta'\|_2^2 \quad \forall \theta, \theta' \in \mathbb{R}^d. \tag{6}$$

For the dependency of L_1 on the four-tuple $(c, \gamma, \tau, |\mathcal{A}|)$ see (Mei et al., 2020b, Lem. 7 & 14).

3. Preliminaries.

Lemma 2 (Non-uniform Polyak-Łojasiewicz condition) *If Assumption 1 holds, then $\|g(\theta)\|_2^2 \geq \mu(\theta)\big(V_\tau^\theta(\rho) - V_\tau^{\theta^\star}(\rho)\big)$, where $\mu(\theta) = 2\tau|\mathcal{S}|^{-1}\min_s \rho(s)\min_{s,a}\pi_\theta(a|s)^2\|d_\rho^{\pi_{\theta^\star}}/\rho\|_\infty^{-1}$. Moreover, if Assumption 2 and 3 hold and $\eta_t \leq \eta/(t+\sqrt{3\sigma^2/(2\epsilon\Delta)})$, for iterates generated by (4), we have $\mu = \inf_{t\geq 1}\mu(\theta_t) > 0$ with probability at least $1-\epsilon/6$.*

Proof By (Mei et al., 2020b, Lem. 15), we have $\|g(\theta)\|_2^2 \geq \mu(\theta)\big(V_\tau^\theta(\rho) - V_\tau^{\theta^\star}(\rho)\big)$, where $\mu(\theta) = 2\tau|\mathcal{S}|^{-1}\min_s \rho(s)\min_{s,a}\pi_\theta(a|s)^2\|d_\rho^{\pi_{\theta^\star}}/\rho\|_\infty^{-1}$. According to (Ding et al., 2021, Lem. 6.4) we further have $\inf_{t\geq 1}\pi_{\theta_t}(a|s) > 0$ with probability $1-\epsilon/6$, which concludes the proof. ■

Lemma 2 states that along a trajectory of (4), with high probability, we have that $\mu(\theta_t) > 0$ for all t. In what follows we will exploit that $\mu = \inf_{t\geq 1}\mu(\theta_t)$ is strictly positive with high probability. Next, we derive elementary recursive inequalities and provide the first main result. If η_t satisfies $0 < \eta_t \leq 1/L_1$ for all $t = 1,\dots,T$, then (4) implies that

$$\begin{aligned}
V_\tau^{\theta_{t+1}}(\rho) &= V_\tau^{\theta_t - \eta_t g(\theta_t) + \eta_t Z_t}(\rho) \\
&\leq V_\tau^{\theta_t}(\rho) - \eta_t\langle g(\theta_t), g(\theta_t) - Z_t\rangle + \tfrac{1}{2}\eta_t^2 L_1\|g(\theta_t) - Z_t\|_2^2 \\
&\leq V_\tau^{\theta_t}(\rho) - \eta_t\|g(\theta_t)\|_2^2 + \eta_t\langle g(\theta_t), Z_t\rangle + \eta_t^2 L_1\big(\|g(\theta_t)\|_2^2 + \|Z_t\|_2^2\big) \\
&\leq V_\tau^{\theta_t}(\rho) + \mu\left(\eta_t^2 L_1 - \eta_t\right)\big(V_\tau^{\theta_t}(\rho) - V_\tau^{\theta^\star}(\rho)\big) + \eta_t\langle g(\theta_t), Z_t\rangle + \eta_t^2 L_1\|Z_t\|_2^2,
\end{aligned}$$

where the first inequality follows from the L_1-smoothness of the value function, the second inequality exploits that $\|x-y\|_2^2 \leq 2\|x\|_2^2 + 2\|y\|_2^2$ for all $x, y \in \mathbb{R}^d$, and the last inequality holds because $\eta_t \leq 1/L_1$ and the PL condition from Lemma 2. Subtracting $V_\tau^{\theta^\star}(\rho)$ from both sides yields

$$V_\tau^{\theta_{t+1}}(\rho) - V_\tau^{\theta^\star}(\rho) \leq (1 - \mu\eta_t + \mu\eta_t^2 L_1)(V_\tau^{\theta_t}(\rho) - V_\tau^{\theta^\star}(\rho)) + \eta_t\langle g(\theta_t), Z_t\rangle + \eta_t^2 L_1\|Z_t\|_2^2. \quad (7)$$

Also, observe that since[2] $\eta_t \leq 1/L_1$, $\mu \leq L_1$ and $\eta_t, L_1 > 0$, we have $\mu\eta_t(\eta_t L_1 - 1) \in (-1, 0]$ and thus $(1 - \mu\eta_t + \mu\eta_t^2 L_1) \in (0, 1]$ for $t = 1,\dots,T$. Next, by substituting the points $\theta' = \theta - 1/L_1 \cdot g(\theta)$ and θ into (6), we obtain $V_\tau^{\theta - 1/L_1\cdot g(\theta)}(\rho) \leq V_\tau^\theta(\rho) - \frac{1}{2}L_1^{-1}\|g(\theta)\|_2^2$. Using the optimality of $\theta^\star$, we may then conclude that $V_\tau^{\theta^\star}(\rho) \leq V_\tau^{\theta - 1/L_1\cdot g(\theta)}(\rho) \leq V_\tau^\theta(\rho) - \frac{1}{2}L_1^{-1}\|g(\theta)\|_2^2$ which implies that

$$\|g(\theta)\|_2^2 \leq 2L_1(V_\tau^\theta(\rho) - V_\tau^{\theta^\star}(\rho)) \quad \forall \theta \in \mathbb{R}^d. \quad (8)$$

Lemma 3 (Exponential upper bound) *Suppose Assumptions 1, 2, and 3 hold, and choose $T > 1$, $\epsilon \in (0,1)$. Then, there is a universal constant $C > 0$ such that the following event holds with probability at least $1-\epsilon/6$. Let $C_{\mathsf{M}} = (\sigma\sqrt{|\mathcal{S}||\mathcal{A}|}C)^2$ where σ is as in (5), and set $\eta > 0$ such that $(\mu\eta - 1) > \sigma^2/C_{\mathsf{M}}$. Let $\eta_t = \eta/(t + t_0 + 1)$ for all $t = 1,\dots,T$ and choose t_0 and K such that*

$$\begin{aligned}
t_0 &\geq \max\left\{\frac{\eta^2 L_1}{(\mu\eta - 1) - B_0 C_0 \eta^2} - 1, L_1\eta - 2, \sqrt{\frac{3\sigma^2}{2\epsilon\Delta}} - 1\right\}, \\
K &\geq \max_{t=1,\dots,T}\left\{B_0^{-1}, (t_0+1)\big(V_\tau^{\theta_1}(\rho) - V_\tau^{\theta^\star}(\rho)\big), \frac{2c_t C_{\mathsf{M}}}{1 - (a_t + B_0 C_0 b_t^2)}\right\},
\end{aligned} \quad (9)$$

[2] To show that $\mu \leq L_1$, one exploits that L_1-smoothness implies Lipschitz continuity of the gradient and recalls that the PL condition implies a quadratic growth condition (Karimi et al., 2016, App. A).

where $B_0 = 1/(2\eta^2 L_1 C_{\mathsf{M}})$*,* $C_0 = 2L_1\sigma^2$*,*

$$a_t = \frac{t+t_0+1}{t+t_0}\left(1 - \mu\eta_t + \mu\eta_t^2 L_1\right), \; b_t = \frac{\eta}{\sqrt{t+t_0}}, \textit{ and } c_t = \frac{\eta^2 L_1}{t+t_0+1}. \tag{10}$$

Then, for any $\delta \geq 0$ *and* $t = 1, \ldots, T$ *we have that* $\mathbb{P}\left(\left(V_\tau^{\theta_{t+1}}(\rho) - V_\tau^{\theta^\star}(\rho)\right) \geq \delta\right) \leq e^{1-(t+t_0+1)\delta/K}$.

The proof is inspired by (Bajovic et al., 2023), with the difference being that they work with strongly convex objective functions. Unfortunately, $V_\tau^\theta(\rho)$ fails to be convex in θ.

Proof Define an auxiliary process $Y_t = (t+t_0)(V_\tau^{\theta_t}(\rho) - V_\tau^{\theta^\star}(\rho))$. Then, the recursion (7) implies that $Y_{t+1} \leq a_t Y_t + b_t\sqrt{t+t_0}\langle g(\theta_t), Z_t\rangle + c_t\|Z_t\|^2$, where a_t, b_t and c_t are as in (10). Defining the MGFs associated with Y_t as $\Phi_t(\nu) = \mathbb{E}\left[\exp\left(\nu Y_t\right)\right]$ and $\Phi_{t+1|t}\left(\nu;\theta_t\right) = \mathbb{E}\left[\exp\left(\nu Y_{t+1}\right) \mid \theta_t\right]$ for $\nu \in \mathbb{R}$, the recursion (7) further implies that

$$\begin{aligned}
\Phi_{t+1|t}\left(\nu;\theta_t\right) &\leq \exp\left(a_t\nu Y_t\right)\mathbb{E}\left[\exp\left(2b_t\nu\sqrt{t+t_0}\langle g(\theta_t), Z_t\rangle\right) \mid \theta_t\right]^{1/2}\mathbb{E}\Big[\exp\left(2c_t\nu\left\|Z_t\right\|_2^2\right) \mid \theta_t\Big]^{1/2} \\
&\leq \exp\left(a_t\nu Y_t\right)\exp\left(\sigma^2 b_t^2\nu^2(t+t_0)\|g(\theta_t)\|_2^2\right)^{1/2}\mathbb{E}\left[\exp\left(2c_t\nu\left\|Z_t\right\|_2^2\right) \mid \theta_t\right]^{1/2} \\
&\leq \exp\left(a_t\nu Y_t\right)\exp\left(2L_1\sigma^2 b_t^2\nu^2 Y_t\right)^{1/2}\mathbb{E}\left[\exp\left(2c_t\nu\left\|Z_t\right\|_2^2\right) \mid \theta_t\right]^{1/2},
\end{aligned}$$

where the first inequality follows from *Hölder's inequality*, the second inequality follows from (5) and the last inequality follows from (8). Recall that $d = \dim(Z_t)$ and define $B_0 = 1/(2\eta^2 L_1(\sigma\sqrt{d}C)^2)$. Then, there is a constant $C > 0$ such that for all $\nu \in [0, B_0]$, we have by monotonicity of c_t, Hölder's inequality for $p = 1/(\nu 2\eta^2 L_1(\sigma\sqrt{d}C)^2) \geq 1$ and (Jin et al., 2019, Lem. 2) that $M(2c_t\nu;\theta) \leq \exp(2c_t\nu(\sigma\sqrt{d}C)^2)$. Recall that $(\sigma\sqrt{d}C)^2 = C_{\mathsf{M}}$ because $d = |\mathcal{S}||\mathcal{A}|$. Moreover, as $\exp(x) \geq \exp(x)^{1/2}$ for $x \geq 0$, we can simplify the above inequality to

$$\begin{aligned}
\Phi_{t+1|t}\left(\nu;\theta_t\right) &\leq \exp\left(a_t\nu Y_t\right)\exp\left(2L_1\sigma^2 b_t^2\nu^2 Y_t\right)\exp(2c_t\nu C_{\mathsf{M}}) \\
&= \exp\left(\nu\left(a_t + \nu C_0 b_t^2\right)Y_t\right)\exp(2c_t\nu C_{\mathsf{M}}) \quad \forall\nu \leq B_0
\end{aligned} \tag{11}$$

for $C_0 = 2L_1\sigma^2$. Taking expectations on both sides of (11) yields

$$\Phi_{t+1}(\nu) \leq \Phi_t\left(\nu(a_t + B_0 C_0 b_t^2)\right)\exp(2c_t\nu C_{\mathsf{M}}) \quad \forall\nu \leq B_0. \tag{12}$$

We aim to show that $\Phi_{t+1}(\nu) \leq \exp(\nu K)$ for $t = 1, \ldots, T$ by studying the magnitude of $(a_t + B_0 C_0 b_t^2)$ and using induction. First, it can be shown that $a_t < 1$ for all $t \geq 1$. To do so, rewrite a_t as

$$a_t = 1 - \frac{\mu\eta - 1}{t+t_0}\left(1 - \frac{\mu\eta^2 L_1}{(\mu\eta - 1)(t+t_0+1)}\right), \tag{13}$$

where we exploit the identity $1 + 1/s = (s+1)/s$ for $s \neq 0$. We have that $(\mu\eta - 1) > 0$. Now since $t_0 \geq L_1\eta - 2$ by assumption, we have $\eta_t \leq 1/L_1$. Recall as well the fact that $\mu \leq L_1$, which leads to $\eta_t \leq 1/\mu$ and thus $\mu\eta_t \leq 1$, that is, $\mu\eta/(t+t_0+1) \leq 1$. Exactly this implies that $(\mu\eta - 1)/(t+t_0) \in (0, 1]$. This covers the first non-trivial fraction of (13). Now for the second part, we have the implication that

$$t_0 + 1 \geq \frac{\mu\eta^2 L_1}{\mu\eta - 1} \implies \frac{\mu\eta^2 L_1}{(\mu\eta - 1)(t+t_0+1)} \in (0, 1],$$

from where we can conclude that $a_t \in [0, 1)$ for any finite $t \geq 1$. However, we will need $(a_t + B_0C_0b_t^2) < 1$. As b_t decays with t_0, this can be achieved by selecting a sufficiently large t_0. Explicitly, let $\mu\eta = 1 + \varepsilon$ for some $\varepsilon > 0$. By the definition of b_t and a_t as in (13) we then obtain

$$a_t + B_0C_0b_t^2 = 1 - \frac{\varepsilon}{t + t_0}\left(1 - \frac{B_0C_0\eta^2}{\varepsilon} - \frac{\mu\eta^2 L_1}{\varepsilon(t + t_0 + 1)}\right). \tag{14}$$

By assumption, we have $\mu\eta - 1 > \sigma^2/C_{\mathsf{M}}$. This implies that $\varepsilon > B_0C_0\eta^2$. It then follows that $a_t + B_0C_0b_t^2 \in [0, 1)$ for any finite $t \geq 1$ and $t_0 + 1 \geq \mu\eta^2L_1/(\varepsilon - B_0C_0\eta^2)$. Recall also that $t_0 \geq L_1\eta - 2$ by assumption. Thus, we have $\eta_t \leq 1/L_1$. Assumption 1 allows then for the following inductive procedure. Let $\nu \in [0, 1/K]$ for K as in (9) then, for $t = 1$, we have $\Phi_1(\nu) \leq \exp(\nu K)$ since $K \geq t_0\left(V_\tau^{\theta_1}(\rho) - V_\tau^{\theta^\star}(\rho)\right)$. Now suppose $\Phi_t(\nu) \leq \exp(\nu K)$ for some arbitrary $t \geq 1$, then, we have by the inductive assumption that $\Phi_{t+1} \leq \exp(2c_tC_{\mathsf{M}}\nu)\exp(\nu(a_t + B_0C_0b_t^2)K)$. However, since for $t = 1, \dots, T$ we also have that $K \geq 2c_tC_{\mathsf{M}}/(1 - (a_t + B_0C_0b_t^2))$, which is well-defined since $(a_t + B_0C_0b_t^2) < 1$, by our selection of t_0 and η, it follows that $\Phi_{t+1}(\nu) \leq \exp(\nu K)$ for $t = 1, \dots, T$. By (Harvey et al., 2019, Claim A.7), if a random variable $X \in \mathbb{R}$ satisfies $\mathbb{E}[\exp(\lambda X)] \leq C_1\exp(\lambda C_2)$ for all $\lambda \leq 1/C_2$ and some universal constants C_1 and C_2, then $\mathbb{P}(X \geq C_2\log(1/\delta)) \leq C_1e\delta$. Hence, after applying (Harvey et al., 2019, Claim A.7) one finds that $\mathbb{P}(Y_t \geq C_3) \leq \exp(1 - C_3/K)$, for any choice of $C_3 \geq 0$ and $t = 1, \dots, T$. The proof is concluded by substituting the expression for Y_t into the previous inequality and a union bound. ■

We note that by the expressions for (a_t, b_t, c_t) and (14) it follows that $\lim_{t\to+\infty} 2c_tC_{\mathsf{M}}/(1 - (a_t + B_0C_0b_t^2))$ is bounded, that is, one can freely select T, which we will exploit later in Theorem 4.

4. Large deviations.

To provide our main result, we need to impose another assumption on the random variables $(Z_t)_t$.

Assumption 4 (Conditional LMGF regularity) *There is a $L_\Lambda \geq 0$ such that*

$$|\Lambda(\lambda; \theta_1) - \Lambda(\lambda; \theta_2)| \leq L_\Lambda\|\lambda\|_2^2\|\theta_1 - \theta_2\|_2 \quad \forall(\lambda, \theta_1, \theta_2) \in \mathbb{R}^d \times \mathbb{R}^d \times \mathbb{R}^d.$$

We refer to (Bajovic et al., 2023, p. 5) for a discussion of this assumption. For instance, when the conditional distribution of Z_t given $\theta_t = \theta$ is Gaussian, and its covariance matrix is Lipschitz continuous in θ, then Assumption 4 holds. In particular, if the covariance is independent of θ, this trivially holds true. Next, we define the MGF of θ_t as $\mathbb{R}^d \ni \lambda \mapsto \Gamma_t(\lambda) := \mathbb{E}\left[\exp(\langle\lambda, \theta_t - \theta^\star\rangle)\right]$ and similarly the LMGF as $\mathbb{R}^d \ni \lambda \mapsto \log\Gamma_t(\lambda)$. Now we are equipped to provide our main technical result, leading directly to a large deviation principle (LDP) upper bound. Note that $\Gamma_t(\lambda)$ depends on the optimal solution $\theta^\star$ which is fixed and unknown.

Theorem 4 (Limiting LMGF) *Suppose that Assumptions 1, 2, 3 and 4 hold and set $\eta_t = \eta/(t + t_0 + 1)$ as in Lemma 3. Then, conditioning on the event that $\mu = \inf_{t\geq1}\mu(\theta_t) > 0$, we have that*

$$\limsup_{t\to+\infty}\frac{1}{t}\log\Gamma_t(t\lambda) \leq r(\lambda) + \int_0^1 \Lambda\left(\eta QD(x)Q^\top\lambda; \theta^\star\right)\mathrm{d}x =: \Psi(\lambda), \tag{15}$$

where $r(\lambda) = O(\|\lambda\|_2^3)$ with its expression presented in the proof below, and Q and $D(x)$ are such that $H(\theta^\star) = QDQ^\top$, $QQ^\top = I_d$, $D = \mathrm{diag}(\rho_1, \dots, \rho_n)$ being the diagonalization of $H(\theta^\star)$, and $D(x) = \mathrm{diag}(x^{\eta\rho_1 - 1}, \dots, x^{\eta\rho_n - 1})$.

To prove Theorem 4, we will largely follow the proof strategy of (Bajovic et al., 2023, Lem. 6), yet, with technical deviations regarding indexing of the sequences. In addition, we cannot appeal to the strong convexity exploited in (Bajovic et al., 2023).

Proof We provide a brief overview of the proof. Step 0 constructs a deterministic sequence as a basis for deriving recursions of MGF. Step 1 defines a ball around the optimal parameter and split the cases depending on whether θ_ℓ resides within the ball. Step 2-4 study individual cases leading to their respective recursive relation. Finally, Step 5 summarizes the previously derived recursions, takes the limit and concludes. *Step 0, sequence.* Fix some vector $\lambda \in \mathbb{R}^d$, integer $t \geq 1$ and define the sequence of vectors $\zeta_\ell = B_{t,\ell}\zeta_t$ via $B_{t,\ell} = (I - \eta_\ell H(\theta^\star)) \cdot \ldots \cdot (I - \eta_{t-1}H(\theta^\star))$ and $\zeta_t = t\lambda$, with $1 \leq \ell < t$. As $\eta_t \leq 1/L_1$ we have that $\|I - \eta_t H(\theta^\star)\|_2 \leq 1 - \eta_t \lambda_{\min}(H(\theta^\star))$. Hence, using (Bajovic et al., 2023, Lem. 2) and exploiting $\eta\lambda_{\min}(H(\theta^\star)) > 1$, *i.e.*, recall $\eta\mu > 1$ from Lemma 3, one can show that

$$\|\zeta_\ell\|_2 \leq t\left(\frac{\ell + t_0 + 1}{t + t_0 + 1}\right)^{\eta\lambda_{\min}(H(\theta^\star))} \|\lambda\|_2 \leq (\ell + t_0 + 1)\|\lambda\|_2.$$

Step 1, conditional MGF. Let μ_ℓ and ν_ℓ be Borel measures induced by θ_ℓ and $\|\theta_\ell - \theta^\star\|_2$, respectively. Then, one readily shows that $\Gamma_{\ell+1}(\zeta_\ell) = \int_{\mathbb{R}^d} \Gamma_{\ell+1|\ell}(\zeta_\ell;\theta)\mu_\ell(\mathrm{d}\theta)$. Next, define the sets $B_\ell(\theta^\star,\delta) = \{\theta_\ell : \|\theta_\ell - \theta^\star\|_2 \leq \delta\}$. Now we can construct the decomposition $\Gamma_{\ell+1}(\zeta_\ell) = \Gamma_{\ell+1|B_\ell(\theta^\star,\delta)}(\zeta_\ell) + \Gamma_{\ell+1|B^c_\ell(\theta^\star,\delta)}(\zeta_\ell)$, which we study separately. *Step 2, $\theta \in B_\ell(\theta^\star,\delta)$.* We have that $\Gamma_{\ell+1|\ell}(\zeta_\ell;\theta) = \exp\left(\Lambda(\eta_\ell\zeta_\ell;\theta) + \langle\zeta_\ell, \theta - \eta_\ell g(\theta) - \theta^\star\rangle\right)$. Let $H(\theta)$ denote the Hessian of $V_\tau^{\pi_\theta}(\rho)$ at θ and define the residual term $h(\theta) = g(\theta) - H(\theta^\star)(\theta - \theta^\star)$. Recall Step 0, Assumption 4 and define the largest residual term $\bar{h}(\delta) = \sup_{\theta \in B_\ell(\theta^\star,\delta)} \|h(\theta)\|_2$, it follows that

$$\begin{aligned}\Gamma_{\ell+1|\ell}(\zeta_\ell;\theta) \leq &\exp\left(\Lambda(\eta_\ell\zeta_\ell;\theta^\star) + L_\Lambda\eta_\ell^2\|\zeta_\ell\|_2^2\delta + \eta_\ell\|\zeta_\ell\|_2\bar{h}(\delta) + \langle\zeta_{\ell-1}, \theta - \theta^\star\rangle\right)\\ \leq &\exp\left(\Lambda(\eta_\ell\zeta_\ell;\theta^\star) + r_0(\lambda,\delta)\right)\exp(\langle\zeta_{\ell-1}, \theta - \theta^\star\rangle)\end{aligned}$$

for $r_0(\lambda,\delta) = 4L_\Lambda\eta^2\|\lambda\|_2^2\delta + 2\eta\|\lambda\|_2\bar{h}(\delta)$. Hence, integrating, we find that

$$\Gamma_{\ell+1|B_\ell(\theta^\star,\delta)}(\zeta_\ell) \leq \exp\left(\Lambda(\eta_\ell\zeta_\ell;\theta^\star) + r_0(\lambda,\delta)\right)\Gamma_\ell(\zeta_{\ell-1}). \tag{16}$$

Step 3, contraction. We construct another recursive formula. Start from

$$\|\theta_{t+1} - \theta^\star\|_2 = \|\theta_t - \eta_t g(\theta_t) + \eta_t Z_t - \theta^\star\|_2 \leq \|\theta_t - \eta_t g(\theta_t) - \theta^\star\|_2 + \eta_t\|Z_t\|_2,$$

and expand $\|\theta_t - \eta_t g(\theta_t) - \theta^\star\|_2^2 = \|\theta_t - \theta^\star\|_2^2 - 2\eta_t\langle\theta_t - \theta^\star, g(\theta_t)\rangle + \eta_t^2\|g(\theta_t)\|_2^2$. Due to optimality of $\theta^\star$ and L_1-smoothness of V_τ^θ we have $\|g(\theta_t)\|_2^2 \leq L_1^2\|\theta_t - \theta^\star\|_2^2$ and

$$\begin{aligned}-\langle\theta_t - \theta^\star, g(\theta_t)\rangle &\leq \tfrac{1}{2}L_1\|\theta_t - \theta^\star\|_2^2 + (V_\tau^{\theta_t}(\rho) - V_\tau^{\theta^\star}(\rho))\\ &\leq \tfrac{1}{2}L_1\|\theta_t - \theta^\star\|_2^2 + \frac{1}{2\mu(\theta_t)}\|g(\theta_t)\|_2^2 \leq \tfrac{1}{2}L_1\|\theta_t - \theta^\star\|_2^2 + \frac{L_1^2}{2\mu(\theta_t)}\|\theta_t - \theta^\star\|_2^2.\end{aligned}$$

As such, one can set $\gamma_t := (1 + \eta_t L_1 + \eta_t L_1^2/M\mu(\theta_t) + \eta_t^2 L_1^2)^{1/2} \leq (3 + L_1/\mu)^{1/2} =: \bar{\gamma}$ such that

$$\|\theta_{t+1} - \theta^\star\|_2 \leq \gamma_t\|\theta_t - \theta^\star\|_2 + \eta_t\|Z_t\|_2 \quad t = 1,\ldots,T. \tag{17}$$

Step 4, $\theta \in B^c_\ell(\theta^\star,\delta)$. Using Step 3 and Assumption 3 one can show that $\Gamma_{\ell+1|\ell}(\zeta_\ell;\theta) \leq \exp\left(\frac{1}{2}\sigma^2\eta^2\|\lambda\|_2^2\right)\exp(\bar{\gamma}(\ell + t_0 + 1)\|\lambda\|_2\|\theta - \theta^\star\|_2)$. Hence, we have

$$\Gamma_{\ell+1|B^c_\ell(\theta^\star,\delta)}(\zeta_\ell) \leq \exp\left(\tfrac{1}{2}\sigma^2\eta^2\|\lambda\|_2^2\right)\int_{\tau \geq \delta} \exp(\bar{\gamma}(\ell + t_0 + 1)\|\lambda\|_2\tau)\nu_\ell(\mathrm{d}\tau).$$

Exploiting Lemma 3, *i.e.* ν_ℓ being induced by $\mathbb{P}\left(\|\theta_\ell - \theta^\star\|_2 \geq \delta\right) \leq e^{1-(\ell+t_0)L_1\delta^2/(2K)}$, we know there is a sufficiently large $\delta(\lambda) = O(\bar{\gamma}\|\lambda\|_2 K/L_1)$, denoted as $\bar{\delta}(\lambda)$ such that

$$\Gamma_{\ell+1|B^c_\ell(\theta^\star,\delta)}(\zeta_\ell) \leq \bar{C}\mathrm{exp}\left(\tfrac{1}{2}\sigma^2\eta^2\|\lambda\|_2^2\right) \tag{18}$$

for some constant $\bar{C}$ *cf.* (Bajovic et al., 2023, Proof of Lem. 6). *Step 5, limiting recursion.* Combining (16) and (18) then yields $\Gamma_{\ell+1}(\zeta_\ell) \leq \bar{C}\mathrm{exp}\left(\frac{1}{2}\sigma^2\eta^2\|\lambda\|_2^2\right) + \exp\left(\Lambda(\eta_\ell\zeta_\ell;\theta^\star) + r(\lambda)\right)\Gamma_\ell(\zeta_{\ell-1})$, where $r(\lambda) = r_0(\lambda, \bar{\delta}(\lambda))$. Continuing with the recursion provides us with

$$\begin{aligned}\Gamma_{t+1}(t\lambda) \leq &\mathrm{exp}\left(\textstyle\sum_{\ell=1}^t \Lambda(\eta_\ell\zeta_\ell;\theta^\star) + r(\lambda)\right)\Gamma_1(\zeta_1)\\ &+ \bar{C}\mathrm{exp}\left(\tfrac{1}{2}\sigma^2\eta^2\|\lambda\|_2^2\right)\textstyle\sum_{\ell=1}^t \exp\left(\sum_{j=\ell}^t \Lambda(\eta_j\zeta_j;\theta^\star) + r(\lambda)\right)\end{aligned}$$

Note that $\Gamma_1(\zeta_1)$ is finite by Assumption 1. Then we immediately obtain that

$$\limsup_{t\to+\infty} t^{-1}\log\Gamma_{t+1}(t\lambda) \leq r(\lambda) + \limsup_{t\to+\infty} t^{-1}\textstyle\sum_{\ell=1}^t \Lambda(\eta_\ell\zeta_\ell;\theta^\star).$$

To complete the proof, we appeal to (Bajovic et al., 2023, Lemma C.3) and find that

$$\lim_{t\to+\infty} t^{-1}\textstyle\sum_{\ell=1}^t \Lambda(\eta_\ell\zeta_\ell;\theta^\star) = \int_0^1 \Lambda\left(\eta Q D(x) Q^\top\lambda;\theta^\star\right)\mathrm{d}x.$$

■

To provide our main result, we recall that the *Legendre-Fenchel transform* of a function $\Psi : \mathbb{R}^d \to \mathbb{R}$ is defined by $I(\theta') = \sup_{\lambda\in\mathbb{R}^d}\langle\theta',\lambda\rangle - \Psi(\lambda)$ for all θ'.

The Theorem below is a strict generalization of Lemma 3. Specifically, Theorem 5 characterizes the concentration rate applicable to any Borel set.

Theorem 5 (LDP upper bound) *Suppose that Assumption 1, 2, 3 and 4 hold and set* $\eta_t = \eta/(t + t_0 + 1)$ *as in Lemma 3. Define* Ψ *as in Theorem 4. Then, with probability at least* $1 - \epsilon/6$ *we have that the sequence* $(\theta_t)_t$ *satisfies a LDP upper bound with a rate function* I*. The function* I *is the Legendre-Fenchel transformation of* Ψ*. That is, for any Borel set* $\Theta \subseteq \mathbb{R}^{|\mathcal{S}||\mathcal{A}|}$*, we have that*

$$\limsup_{t\to+\infty}\frac{1}{t}\log\mathbb{P}(\theta_t \in \Theta) \leq -\inf_{\theta'+\theta^\star\in\overline{\Theta}} I(\theta'). \tag{19}$$

Proof Directly from Theorem 4 and the *Gärtner-Ellis theorem* (Gärtner, 1977; Ellis, 1984). ■

Note, one can also bring the offset term $\theta^\star$ into the left-hand-side of (19), *i.e.*, consider $\mathbb{P}(\theta_t - \theta^\star \in \Theta)$. Indeed, strictly speaking, the LDP upper bound is derived for the sequence $(\theta_t - \theta^\star)_t$.

5. Ramifications

One could argue that the results (bounds) from above contribute towards a better understanding of the softmax transformation, the effect of the regularization parameter τ, the effect of initialization and the analysis of restrictive policy parametrizations, all active topics of research Hennes et al. (2020); Mei et al. (2020a); Li et al. (2021); Mei et al. (2020b). However, more interesting is that Theorem 5 also allows for a simple proof of exponential decay, with high probability, for a whole family of policy parametrizations different from softmax.

The structure of Ψ (non-negativity and finiteness) immediately reveals that its Fenchel conjugate (Legendre-Fenchel transform) I satisfies $I(0) = 0$, better yet, $I(\theta') > 0$ for all $\theta' \neq 0$. Hence, I can be identified as a *good* rate function (Dembo and Zeitouni, 2009, p. 4). Moreover, for *any* Borel set Θ such that $\theta^\star \notin \overline{\Theta}$ we have, with high probability, that the probability $\mathbb{P}(\theta_t \in \Theta)$ decays exponentially fast. Not only that, we can provide a bound on the convergence rate for any of those sets, *i.e.*, let $r := \inf_{\theta' + \theta^\star \in \overline{\Theta}} I(\theta')$, then, $\mathbb{P}(\theta_t \in \Theta) \leq \exp(-rt + o(t))$. To lift this observation to any continuous transformation of the softmax parametrization we need the following principle from the theory of large deviations. To avoid pathological examples, we assume all our sets to be topological Hausdorff spaces.

Theorem 6 (Contraction principle (Dembo and Zeitouni, 2009, § 4.2.1)) *Consider $U \subseteq \mathbb{R}^d$ and $W \subseteq \mathbb{R}^q$, and $f : U \to W$ a continuous map. Define $I' : W \to [0, +\infty]$ through $I'(w) = \inf_{u \in U, w = f(u)} I(u)$ for all $w \in W$, where $I : U \to [0, +\infty]$ is a good rate function. Then, I' is also a good rate function on W.*

Although we studied a softmax policy, the *contraction principle*, originally used by Donsker and Varadhan (1976), allows us to extend our LDP results to any continuous transformation f of $\theta_t - \theta^\star$, *e.g.*, one recovers the *escort transformation* (Mei et al., 2020a) by transforming θ_t componentwise by the map $\mathbb{R} \ni w \mapsto p \log |w|$ for $p \geq 1$. To be explicit, whereas the softmax policy is given by (3), the escort parametrization is given by $\pi_\theta(a|s) = |\theta(s,a)|^p / \sum_{a' \in \mathcal{A}} |\theta(s, a')|^p$. Hence, we directly infer that the exponential decay rate extends from the softmax- to the escort parametrization. Indeed, $w \mapsto p \log |w|$ is not a smooth function, yet, for $p \geq 2$ the escort policy parametrization is differentiable, which is precisely the setting for which an exponential decay rate under entropic regularization is provided by (Mei et al., 2020a, Thm. 4). To formally summarize this observation, by combining Theorem 5 and Theorem 6 we get the following.

Corollary 7 (Going beyond softmax) *Let $(\theta_t - \theta^\star)_t \subset \mathbb{R}^{|\mathcal{S}||\mathcal{A}|}$ be a sequence corresponding to softmax policy gradient and let the sequence $(\omega_t - \omega^\star)_t \subset \mathbb{R}^q$ be defined via a continuous map $f : \mathbb{R}^{|\mathcal{S}||\mathcal{A}|} \to \mathbb{R}^q$ through $\omega_t - \omega^\star := f(\theta_t - \theta^\star)$ $\forall t$, then, with probability at least $1 - \epsilon/6$,*

$$\limsup_{t \to +\infty} \frac{1}{t} \log \mathbb{P}(\omega_t \in \Omega) \leq - \inf_{\omega' + \omega^\star \in \overline{\Omega}} I'(\omega'), \tag{20}$$

for any Borel set $\Omega \subseteq \mathbb{R}^q$ and I' as in Theorem 6.

Similar to the escort transformation, one can study *spherical*- and *Taylor* softmax (de Brébisson and Vincent, 2016). However, Corollary 7 is ought to be most interesting to apply to prospective policy parametrizations leading to (20) being non-trivial. Indeed, we emphasize that one must verify that the solution to the non-linear optimization problem (20) is non-trivial, there is no free lunch. Moreover, the gradient estimation errors needs to remain well-behaved after the transformation, which may not be true for an arbitrary continuous mapping f. We leave it to future work to explore the exact characterizations of the scope of Corollary 7.

To remove the $(1 - \epsilon/6)$-high probability statements throughout, future work aims at studying the multiphase algorithm from (Ding et al., 2021) in the context of large deviations, this requires being able to study a controlled sequence $(Z_t)_t$, which is a non-trivial extension. We also remark that in the context of SGD for an objective satisfying a PL condition, the LDP upper bound always holds, simply since $\mu > 0$, *cf.* Lemma 2.

References

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.

Dragana Bajovic, Dusan Jakovetic, and Soummya Kar. Large deviations rates for stochastic gradient descent with strongly convex functions. In *International Conference on Artificial Intelligence and Statistics*, pages 10095–10111, 2023.

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

Léon Bottou. Large-scale machine learning with stochastic gradient descent. In *International Conference on Computational Statistics*, pages 177–186, 2010.

Shicong Cen, Chen Cheng, Yuxin Chen, Yuting Wei, and Yuejie Chi. Fast global convergence of natural policy gradient methods with entropy regularization. *Operations Research*, 70(4): 2563–2578, 2022.

Alexandre de Brébisson and Pascal Vincent. An exploration of softmax alternatives belonging to the spherical loss family. In *International Conference on Learning Representations*, 2016.

Thomas Degris, Patrick M Pilarski, and Richard S Sutton. Model-free reinforcement learning with continuous action in practice. In *IEEE American Control Conference*, pages 2177–2182, 2012.

Amir Dembo and Ofer Zeitouni. *Large Deviations Techniques and Applications*. Springer, 2009.

Yuhao Ding, Junzi Zhang, and Javad Lavaei. Beyond exact gradients: Convergence of stochastic soft-max policy gradient methods with entropy regularization. *arXiv preprint arXiv:2110.10117*, 2021.

Monroe D Donsker and S.R. Srinivasa Varadhan. Asymptotic evaluation of certain Markov process expectations for large time—III. *Communications on pure and applied Mathematics*, 29(4): 389–461, 1976.

Paul Dupuis and Hui Wang. Importance sampling, large deviations, and differential games. *Stochastics: An International Journal of Probability and Stochastic Processes*, 76(6):481–508, 2004.

Richard S Ellis. Large deviations for a general class of random vectors. *The Annals of Probability*, 12(1):1–12, 1984.

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

Jordan Frank, Shie Mannor, and Doina Precup. Reinforcement learning in the presence of rare events. In *International Conference on Machine Learning*, pages 336–343, 2008.

Jürgen Gärtner. On large deviations from the invariant measure. *Theory of Probability & Its Applications*, 22(1):24–39, 1977.

Saeed Ghadimi and Guanghui Lan. Stochastic first-and zeroth-order methods for nonconvex stochastic programming. *SIAM Journal on Optimization*, 23(4):2341–2368, 2013.

Robert M Gower, Mark Schmidt, Francis Bach, and Peter Richtárik. Variance-reduced methods for machine learning. *Proceedings of the IEEE*, 108(11):1968–1983, 2020.

Nicholas JA Harvey, Christopher Liaw, Yaniv Plan, and Sikander Randhawa. Tight analyses for non-smooth stochastic gradient descent. In *Conference on Learning Theory*, pages 1579–1613, 2019.

Daniel Hennes, Dustin Morrill, Shayegan Omidshafiei, Rémi Munos, Julien Perolat, Marc Lanctot, Audrunas Gruslys, Jean-Baptiste Lespiau, Paavo Parmas, Edgar Duéñez-Guzmán, et al. Neural replicator dynamics: Multiagent learning via hedging policy gradients. In *International Conference on Autonomous Agents and Multiagent Systems*, pages 492–501, 2020.

Chi Jin, Praneeth Netrapalli, Rong Ge, Sham M Kakade, and Michael I Jordan. A short note on concentration inequalities for random vectors with subGaussian norm. *arXiv preprint arXiv:1902.03736*, 2019.

Hamed Karimi, Julie Nutini, and Mark Schmidt. Linear convergence of gradient and proximal-gradient methods under the Polyak-Łojasiewicz condition. In *European Conference on Machine Learning and Knowledge Discovery in Databases*, pages 795–811, 2016.

Thibault Lahire. Importance sampling for stochastic gradient descent in deep neural networks. *arXiv preprint arXiv:2303.16529*, 2023.

Guanghui Lan. Policy mirror descent for reinforcement learning: Linear convergence, new sampling complexity, and generalized problem classes. *Mathematical programming*, 198(1):1059–1106, 2023.

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, 2021.

Zijian Liu, Ta Duy Nguyen, Thien Hang Nguyen, Alina Ene, and Huy Lê Nguyen. High probability convergence of stochastic gradient methods. *arXiv preprint arXiv:2302.14843*, 2023.

Liam Madden, Emiliano Dall'Anese, and Stephen Becker. High-probability convergence bounds for non-convex stochastic gradient descent. *arXiv preprint arXiv:2006.05610*, 2020.

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

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, 2020b.

Ofir Nachum, Mohammad Norouzi, Kelvin Xu, and Dale Schuurmans. Bridging the gap between value and policy based reinforcement learning. *Advances in neural information processing systems*, 30, 2017.

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

Herbert Robbins and Sutton Monro. A Stochastic Approximation Method. *The Annals of Mathematical Statistics*, 22(3):400 – 407, 1951.

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

Richard S Sutton, David McAllester, Satinder Singh, and Yishay Mansour. Policy gradient methods for reinforcement learning with function approximation. In *Advances in Neural Information Processing Systems*, pages 1057–1063, 1999.

S.R. Srinivasa Varadhan. Special invited paper: Large deviations. *The Annals of Probability*, 36(2): 397–419, 2008.

Kaiqing Zhang, Alec Koppel, Hao Zhu, and Tamer Basar. Global convergence of policy gradient methods to (almost) locally optimal policies. *SIAM Journal on Control and Optimization*, 58(6): 3586–3612, 2020.