Landscape Analysis of Stochastic Policy Gradient Methods

Xingtu Liu ✉

Simon Fraser University, Burnaby BC, Canada
xingtu_liu@sfu.ca

Abstract. Policy gradient methods are among the most important techniques in reinforcement learning. Despite the inherent non-concave nature of policy optimization, these methods demonstrate good behavior, both in practice and in theory. Hence, it is important to study the non-concave optimization landscape. This paper aims to provide a comprehensive landscape analysis of the objective function optimized by stochastic policy gradient methods. Using tools borrowed from statistics and topology, we prove a uniform convergence result for the empirical objective function, (and its gradient, Hessian and stationary points) to the corresponding population counterparts. Specifically, we derive $\tilde{O}(\sqrt{|\mathcal{S}||\mathcal{A}|}/(1-\gamma)\sqrt{n})$ rates of convergence, with the sample size n, the state space $\mathcal{S}$, the action space $\mathcal{A}$, and the discount factor γ. Furthermore, we prove the one-to-one correspondence of the non-degenerate stationary points between the population and the empirical objective. In particular, our findings are agnostic to the choice of the algorithm and hold for a wide range of gradient-based methods. Consequently, we are able to recover and improve numerous existing results through the vanilla policy gradient. To the best of our knowledge, this is the first work theoretically characterizing optimization landscapes of stochastic policy gradient methods.

Keywords: Reinforcement Learning · Policy Gradient · Nonconvex Optimization · Sample Complexity.

1 Introduction

Reinforcement learning (RL) is a machine learning paradigm aimed at building learning agents capable of making sequential decisions in a dynamic environment. Recent years have witnessed great successes of RL in many real-world problems, such as strategy games [1,2], robotics [3], and large language models [4,5]. Among RL methods, policy gradient (PG) methods [6,7,8] represent a critical class of algorithms used to search the optimal policy. PG are widely used due to their inherent advantages. They are applicable to tasks with both discrete and continuous action spaces. They can function as both model-free and model-based methods, showcasing their versatility. Moreover, they are capable of handling function approximation and high-dimensional state-action spaces, among other benefits.

The asymptotic convergence of policy gradient has been established decades ago [6,7,9,8]. With the smoothness and gradient dominance condition, softmax policy gradient methods converge to global optimum at sub-linear and linear rates [10,11,12,13]. However, these results require a sufficient exploration assumption and these rates are associated with problem-dependent constants that could be exponentially large [14]. Moreover, in practice, obtaining access to the exact gradient of the objective function is not feasible. Thus, stochastic policy gradient methods are studied [15,16,17,18]. In the stochastic setting, there exists a body of work studying the sample complexity of global convergence [19,20,21,22,23,24]. These results rely on techniques such as importance sampling, the use of second-order information or fisher information, and acceleration. The algorithms and techniques therein each possess their own limitations. Given the non-concave nature of policy optimization, studying the convergence to first-order stationary points (FOSPs) becomes a natural course of inquiry [6,25,26,27,28,20,29]. However, FOSPs do not necessarily correspond to local maxima. The algorithms designed for finding FOSPs may converge to local minima and saddle points that are undesirable for RL tasks. Therefore, this motivates research on second-order stationary points (SOSPs) [15,28,29]. This paper delves into these concerns by analyzing the underlying optimization landscape. We provide a uniform convergence of the empirical objective function towards the population objective function, as well as its gradient and Hessian. For the set of non-degenerate stationary points encompassing maximal, minimal, and saddle points of both the population and empirical objective functions, we prove there is a one-to-one correspondence between them. This provides us with a clearer perspective on the optimization landscape of stochastic policy gradient methods. Our results also match some existing sample complexity results with a simpler algorithm design.

Despite a substantial body of work analyzing the nonconvex optimization landscape in machine learning and deep learning [30,31,32,33,34,35,36,37,38,39], a thorough theoretical examination of the optimization landscape of policy gradient algorithms has remained elusive. We notice two recent work [40,41] that explore the landscape of policy gradient methods, within the contexts of solving combinatorial problems with deep neural networks and static output feedback (SOF) control in discrete-time linear time-invariant (LTI) systems, respectively. Therefore, the settings and results of these work are inherently distinct from ours.

2 Preliminaries

An infinite-horizon discounted Markov decision process (MDP) [42] is defined by $\mathcal{M}(\mathcal{S}, \mathcal{A}, \mathcal{P}, r, \gamma, \rho)$. We use $\Delta(\mathcal{X})$ to denote the set of the probability distribution over the set $\mathcal{X}$. $\mathcal{M}(\mathcal{S}, \mathcal{A}, \mathcal{P}, r, \gamma, \rho)$ is specified by a finite state space $\mathcal{S}$, a finite action space $\mathcal{A}$, transition function $\mathcal{P} : \mathcal{S} \times \mathcal{A} \to \Delta(\mathcal{S})$, reward function $r : \mathcal{S} \times \mathcal{A} \to \mathbb{R}$, and the discount factor $\gamma \in [0, 1)$. We assume the reward is bounded,

i.e. $r(s,a) \in [0,1]$[1], $\forall (s,a)$. Given a policy $\pi_{\boldsymbol{\theta}} : \mathcal{S} \to \Delta(\mathcal{A})$, the expected value function $\boldsymbol{V}(\boldsymbol{\theta})$ is defined as

$$\boldsymbol{V}(\boldsymbol{\theta}) := \mathbb{E}_{\boldsymbol{\tau}}[V(\boldsymbol{\theta}, \boldsymbol{\tau})] := \mathbb{E}_{\boldsymbol{\tau}}\left[\sum_{t=0}^{\infty} \gamma^t r(a_t, s_t)\right] \tag{1}$$

where the expectation is computed over the following distribution

$$\boldsymbol{\tau} \sim \rho(s_0) \prod_{t=0}^{\infty} \pi_{\boldsymbol{\theta}}(a_t|s_t)\mathcal{P}(s_{t+1}|s_t, a_t).$$

In reality, the dynamics of the environment are not fully known or are too complex to model accurately, and it is impossible to compute a sum over an infinite sequence. Therefore, in practice, we sample n trajectories with a finite horizon $H \in \mathbb{N}$, specifically,

$$\{\boldsymbol{\tau}_i\} = \left\{(s_0^{(i)}, a_0^{(i)}, r(s_0^{(i)}, a_0^{(i)}), \ldots, s_{H-1}^{(i)}, a_{H-1}^{(i)}, r(s_{H-1}^{(i)}, a_{H-1}^{(i)}))\right\},$$

and the empirical estimate of the value function is defined as

$$\hat{\boldsymbol{V}}_n(\boldsymbol{\theta}) := \frac{1}{n}\sum_{i=1}^{n} \hat{V}_H(\boldsymbol{\theta}, \boldsymbol{\tau}_i) := \frac{1}{n}\sum_{i=1}^{n}\sum_{t=0}^{H} \gamma^t r(a_t^{(i)}, s_t^{(i)}). \tag{2}$$

By the policy gradient theorem [43], for a differentiable map $\boldsymbol{\theta} \mapsto \pi_{\boldsymbol{\theta}}$, the gradient of the expected value function is computed by

$$\nabla \boldsymbol{V}(\boldsymbol{\theta}) = \mathbb{E}_{\boldsymbol{\tau}}\left[\sum_{t=0}^{\infty} \gamma^t r(s_t, a_t) \sum_{k=0}^{\infty} \nabla_{\boldsymbol{\theta}} \log \pi_{\boldsymbol{\theta}}(a_k|s_k)\right], \tag{3}$$

and the REINFORCE gradient estimator [6] is computed by

$$\nabla \hat{\boldsymbol{V}}_n(\boldsymbol{\theta}) = \frac{1}{n}\sum_{i=1}^{n}\sum_{t=0}^{H-1} \gamma^t r(s_t^{(i)}, a_t^{(i)}) \sum_{k=0}^{H-1} \nabla_{\boldsymbol{\theta}} \log \pi_{\boldsymbol{\theta}}(a_k^{(i)}|s_k^{(i)}). \tag{4}$$

3 Uniform Convergence Results

We first present the following assumptions regarding the norm of the first-order, second-order, and third-order derivatives of the score function $\log \pi_{\boldsymbol{\theta}}(a|s)$. Note that these assumptions are common and realistic in many scenarios. We instantiate the results with softmax parametrization in Lemma 1.

Assumption 1. *There exists $G_g > 0$ such that for every action $a \in \mathcal{A}$ and state $s \in \mathcal{S}$, the gradient of $\log \pi_{\boldsymbol{\theta}}(a|s)$ satisfies*

$$\|\nabla_{\boldsymbol{\theta}} \log \pi_{\boldsymbol{\theta}}(a|s)\|_2 \leq G_g.$$

[1] For rewards in $[R_{\min}, R_{\max}]$ simply rescale these bounds.

Assumption 2. *There exists $G_h > 0$ such that for every action $a \in \mathcal{A}$ and state $s \in \mathcal{S}$, the Hessian of $\log \pi_{\boldsymbol{\theta}}(a|s)$ satisfies*

$$\left\|\nabla^2_{\boldsymbol{\theta}} \log \pi_{\boldsymbol{\theta}}(a|s)\right\|_F \leq G_h.$$

Assumption 3. *There exists $G_t > 0$ such that for every action $a \in \mathcal{A}$ and state $s \in \mathcal{S}$, the third-order partial derivatives of $\log \pi_{\boldsymbol{\theta}}(a|s)$ satisfies*

$$\left\|\nabla^3_{\boldsymbol{\theta}} \log \pi_{\boldsymbol{\theta}}(a|s)\right\|_F \leq G_t.$$

Remark 1. Softmax policy, Gaussian policy with bounded action spaces, and relative entropy policy serve as instances that satisfy the above assumptions. The focus on the score function is due to an analytical simplicity. It is important to note that the main results in this paper remain applicable if we replace the score function with the objective function. In that case, policy with direct parametrization would also satisfy the assumptions, and we left it as a future work.

Given the function $\boldsymbol{\theta} : \mathcal{S} \times \mathcal{A} \to \mathbb{R}$, the tabular policy with softmax parametrization is defined as

$$\pi_{\boldsymbol{\theta}}(a|s) = \frac{\exp(\boldsymbol{\theta}(s,a))}{\sum_{a'} \exp(\boldsymbol{\theta}(s,a'))}.$$

Lemma 1. *The softmax policy satisfies Assumption 1, 2, and 3 with $G_g = \sqrt{2}$, $G_h = 1$, and $G_t = 3\sqrt{|A|}$.*

Proof. In the subsequent proof, subscripts are omitted for simplicity. Recall that $\frac{\partial \pi(a|s)}{\partial \boldsymbol{\theta}(s',\cdot)} = 0$ for $s' \neq s$. Moreover,

$$\frac{\partial \pi(a|s)}{\partial \boldsymbol{\theta}(s,a)} = \pi(a|s) - \pi(a|s)^2$$

and

$$\frac{\partial \pi(a|s)}{\partial \boldsymbol{\theta}(s,a')} = -\pi(a|s)\pi(a'|s).$$

By applying the chain rule, we obtain

$$\frac{\partial \log \pi(a|s)}{\partial \boldsymbol{\theta}(s,\cdot)} = \frac{1}{\pi(a|s)} \frac{\partial \pi(a|s)}{\partial \boldsymbol{\theta}(s,\cdot)}.$$

Now by letting $D(a,\boldsymbol{\theta}) = \frac{\partial \pi(a|s)}{\partial \boldsymbol{\theta}(s,\cdot)}$ and

$$\mathbf{1}_{ij} = \begin{cases} 1, & \text{if } i = j \\ 0, & \text{otherwise,} \end{cases}$$

it follows that

$$\begin{aligned} D_i(a,\boldsymbol{\theta}) &= \mathbf{1}_{ia}(\pi(a|s)-\pi(a|s)^2)+(1-\mathbf{1}_{ia})(-\pi(a|s)\pi(a'|s)) \\ &= \pi(a|s)(\mathbf{1}_{ia}(1-\pi(a|s))+(1-\mathbf{1}_{ia})(-\pi(a'|s))), \end{aligned}$$

and

$$\frac{\partial \log \pi(a|s)}{\partial \boldsymbol{\theta}(s,i)} = \mathbf{1}_{ia}(1-\pi(a|s))+(1-\mathbf{1}_{ia})(-\pi(a'|s)). \tag{5}$$

Therefore,

$$\begin{aligned} \left\| \frac{\partial \log \pi(a|s)}{\partial \boldsymbol{\theta}(s,\cdot)} \right\|_2 &\leq \sqrt{1+\sum_{a'}\pi(a'|s)^2} \\ &\leq \sqrt{2}. \end{aligned}$$

Next, from (5) we observe that

$$\begin{aligned} \frac{\partial^2 \log \pi(a|s)}{\partial \boldsymbol{\theta}(s,\cdot)^2} &= -\frac{\partial \pi(\cdot|s)}{\partial \boldsymbol{\theta}(s,\cdot)} \\ &= -\mathrm{Diag}(\pi(\cdot|s))+\pi(\cdot|s)\pi(\cdot|s)^\top \end{aligned}$$

and

$$\frac{\partial^3 \log \pi(\cdot|s)}{\partial \boldsymbol{\theta}(s,\cdot)^3} = -\frac{\partial^2 \pi(a|s)}{\partial \boldsymbol{\theta}(s,\cdot)^2}.$$

From [20,11], we have $\left\| \frac{\partial^2 \log \pi(a|s)}{\partial \boldsymbol{\theta}(s,\cdot)^2} \right\|_F \leq 1$. Now letting $H(a,) = \frac{\partial^2 \pi(a|s)}{\partial \boldsymbol{\theta}^2(s,\cdot)}$, we obtain

$$\begin{aligned} &H_{i,j}(a,\boldsymbol{\theta}) \\ &= \frac{\partial \left\{\mathbf{1}_{ia}\pi(a|s)-\pi_{\boldsymbol{\theta}}(a|s)\pi(i|s)\right\}}{\partial \boldsymbol{\theta}(s,j)} \\ &= \mathbf{1}_{ia}\left(\mathbf{1}_{ja}\pi(a|s)-\pi(a|s)\pi(j|s)\right)-\pi(a|s)\left(\mathbf{1}_{ij}\pi(j|s)-\pi(i|s)\pi(j|s)\right) \\ &\quad -\pi(i|s)\left(\mathbf{1}_{ja}\pi(a|s)-\pi(a|s)\pi(j|s)\right) \\ &= \pi(a|s)\left(\mathbf{1}_{ia}\left(\mathbf{1}_{ja}-\pi(j|s)\right)-\pi(j|s)\left(\mathbf{1}_{ij}-\pi(i|s)\right)-\pi(i|s)\left(\mathbf{1}_{ja}-\pi(j|s)\right)\right) \end{aligned}$$

and

$$
\begin{aligned}
&\left\| \frac{\partial^3 \log \pi(\cdot|s)}{\partial \boldsymbol{\theta}(s,\cdot)^3} \right\|_F \\
&= \left\| -H_{i,j}(\cdot, \boldsymbol{\theta}) \right\|_F \\
&= \left(\sum_a \sum_j \sum_i H_{i,j}(a, \boldsymbol{\theta})^2 \right)^{1/2} \\
&= \left(\sum_a \pi(a|s)^2 \left(\sum_{j \neq a} \sum_i \left(\frac{H_{i,j}(a, \boldsymbol{\theta})}{\pi(a|s)} \right)^2 + \sum_i \left(\frac{H_{i,a}(a, \boldsymbol{\theta})}{\pi(a|s)} \right)^2 \right) \right)^{1/2} \\
&\leq \left(\sum_{j \neq a} \sum_i \left(\frac{H_{i,j}(a, \boldsymbol{\theta})}{\pi(a|s)} \right)^2 + \sum_i \left(\frac{H_{i,a}(a, \boldsymbol{\theta})}{\pi(a|s)} \right)^2 \right)^{1/2} \\
&\leq \left(\sum_{j \neq a} \sum_i \pi(j|s)^2 (2\pi(i|s) - \mathbf{1}_{ia} - \mathbf{1}_{ij})^2 + \sum_i \left(\frac{H_{i,a}(a, \boldsymbol{\theta})}{\pi(a|s)} \right)^2 \right)^{1/2} \\
&\leq \Bigg(\sum_i 4 \sum_{j \neq a} \pi(j|s) \\
&\qquad + \sum_i \left(\mathbf{1}_{ia} \left(1 - \pi(a|s)\right) - \pi(a|s) \left(\mathbf{1}_{ia} - \pi(i|s) \right) - \pi(i|s) \left(1 - \pi(a|s)\right) \right)^2 \Bigg)^{1/2} \\
&\leq \left(\sum_i 4 \sum_{j \neq a} \pi(j|s) + \sum_{i \neq a} \left(\pi(i|s)(2\pi(a|s) - 1) \right)^2 + 4 \right)^{1/2} \\
&\leq \left(4|A| + \sum_i \pi(i|s) + 4 \right)^{1/2} \\
&= \sqrt{4|A| + 5} \\
&\leq 3\sqrt{|A|}.
\end{aligned}
$$

This concludes the proof. □

3.1 Uniform Convergence of Objective

Theorem 1. *Suppose Assumption 1 holds. Then the truncated estimator defined in (2) uniformly converges to the expected infinite-horizon discounted value function. Specifically, if* $n \geq \max\left\{ \frac{18(\log(4/\mathcal{E}) + |\mathcal{S}||\mathcal{A}| \log(n/\mathcal{E}))}{(1-\gamma)^2}, \frac{18HG_g r}{(1-\gamma)} \right\}$, *then we*

have

$$\sup_{\boldsymbol{\theta}\in\Theta}\left|\hat{\boldsymbol{V}}_n(\boldsymbol{\theta})-\boldsymbol{V}(\boldsymbol{\theta})\right|\leq\sqrt{\frac{36|\mathcal{S}||\mathcal{A}|\log(n/\mathcal{E}))}{n(1-\gamma)^2}}+\frac{18HG_gr}{n(1-\gamma)}$$

with probability at least $1-\mathcal{E}$*, where* $\Theta=B^{|\mathcal{S}||\mathcal{A}|}(r)$.

Remark 2. In our analysis, the parameter space is confined within a sphere with radius r, which is permitted to be significantly large, on the order of $O(|\mathcal{S}||\mathcal{A}|/\epsilon(1-\gamma)G_g)$. Direct parametrization and Gaussian parametrization (with bounded action space) are examples that have a constrained parameter space. However, to search for a deterministic optimal policy using softmax parametrization, the parameter space must be unbounded, marking a limitation of our results. Nevertheless, we can quantify an error term between the reward inferred from the constrained parameter and the optimal parameter. This error diminishes with an increasingly large radius r.

From Theorem 1, one can observe that with increasingly larger sample size n, the difference between empirical objective and population objective decreases monotonically. According to the definition of uniform convergence [44], the uniform convergence rate of the empirical objective to its population objective is $O(1/\sqrt{n})$ up to a log factor. A direct consequence of this result is that for any algorithm that finds global optima using the exact gradient across T iterations, it can be deduced that the stochastic version of the algorithm exhibits a sample complexity of $O(T/\epsilon^2)$. For algorithms demonstrating linear convergence, such as the natural policy gradient and geometry-aware policy gradient [45,12], this allows for achieving the optimal sample complexity.

3.2 Uniform Convergence of Gradient

Theorem 2. *Suppose assumption 1 and 2 hold. Then the truncated gradient estimator defined in (2) uniformly converges to the gradient of the expected infinite-horizon discounted value function in Euclidean norm. Specifically, if* $n\geq\max\left\{\frac{64HG_g|\mathcal{S}||\mathcal{A}|\log(n/\mathcal{E})}{(1-\gamma)^2},\frac{18HG_hr}{(1-\gamma)}\right\}$*, then we have*

$$\sup_{\boldsymbol{\theta}\in\Theta}\left\|\nabla\hat{\boldsymbol{V}}_n(\boldsymbol{\theta})-\nabla\boldsymbol{V}(\boldsymbol{\theta})\right\|_2\leq\sqrt{\frac{64HG_g|\mathcal{S}||\mathcal{A}|\log(n/\mathcal{E})}{n(1-\gamma)^2}}+\frac{18HG_hr}{n(1-\gamma)}.$$

with probability at least $1-\mathcal{E}$*, where* $\Theta=B^{|\mathcal{S}||\mathcal{A}|}(r)$.

According to Theorem 2, if a point $\boldsymbol{\theta}$ is an $\epsilon/2$-stationary point of $\boldsymbol{V}(\boldsymbol{\theta})$, then for $n=(1/\epsilon^2)$, $\boldsymbol{\theta}$ is also an ϵ-stationary point of $\hat{\boldsymbol{V}}_n(\boldsymbol{\theta})$ with high probability. With vanilla policy gradient, Theorem 2 provides a sample complexity of $O(1/\epsilon^4)$ for finding FOSPs. This improves over the $O(1/\epsilon^{4.5})$ rate in [28], and matches the $O(1/\epsilon^4)$ rates in [6,27,20].

3.3 Uniform Convergence of Hessian

Theorem 3. *Suppose assumption 2 and 3 hold. Then the Hessian of the estimator uniformly converges to the Hessian of the expected infinite-horizon discounted value function in operator norm. Specifically, if* $n \geq \max\left\{\frac{128HG_h|\mathcal{S}||\mathcal{A}|\log(n/\mathcal{E})}{(1-\gamma)^2}, \frac{18HG_tr}{(1-\gamma)}\right\}$*, then we have*

$$\sup_{\boldsymbol{\theta}\in\Theta}\left\|\nabla^2\hat{\boldsymbol{V}}_n(\boldsymbol{\theta})-\nabla^2\boldsymbol{V}(\boldsymbol{\theta})\right\|_{op}\leq\sqrt{\frac{128HG_h|\mathcal{S}||\mathcal{A}|\log(n/\mathcal{E})}{n(1-\gamma)^2}}+\frac{18HG_tr}{n(1-\gamma)}.$$

with probability at least $1-\mathcal{E}$*, where* $\Theta = B^{|\mathcal{S}||\mathcal{A}|}(r)$*.*

Recall that an ϵ-stationary point does not necessarily indicate a local maximum, which motivates the study on convergence towards SOSPs. Specifically, the goal is to achieve convergence to first-order stationary points that are also local maxima. For this purpose, the eigenvalues of the Hessian must be non-positive. An ϵ-second-order stationary point and an $(\epsilon, \sqrt{\chi\epsilon})$-second-order stationary point are defined as follows.

Definition 1 (Second-Order Stationary Point [46]). *For the* χ*-Hessian-Lipschitz function* $J(\cdot)$*, we say that* $\boldsymbol{\theta}$ *is an* ϵ*-second-order stationary point if*

$$\|\nabla J(\boldsymbol{\theta})\|_2 \leq \epsilon \quad \text{and} \quad \lambda_{\max}(\nabla^2 J(\boldsymbol{\theta})) \leq 0;$$

we say $\boldsymbol{\theta}$ *is an* $(\epsilon, \sqrt{\chi\epsilon})$*-second-order stationary point if*

$$\|\nabla J(\boldsymbol{\theta})\|_2 \leq \epsilon \quad \text{and} \quad \lambda_{\max}(\nabla^2 J(\boldsymbol{\theta})) \leq \sqrt{\chi\epsilon}.$$

Note that if one matrix qualifies as an ϵ-SOSP and is close to another matrix, this does not necessarily mean that the second matrix is also an ϵ-SOSP. Consider a simple example

$$M_1 = \begin{pmatrix} -1 & 0 \\ 0 & \epsilon/2 \end{pmatrix} \quad \text{and} \quad M_2 = \begin{pmatrix} -1 & 0 \\ 0 & -\epsilon/2 \end{pmatrix}.$$

It follows that $\|M_1 - M_2\|_{op} \leq \|M_1 - M_2\|_F \leq \epsilon$, yet $\lambda_{\max}(M_1) > 0$ and $\lambda_{\max}(M_2) < 0$. Hence, based solely on Theorem 3 only, we cannot ensure that if a point $\boldsymbol{\theta}$ is an $\epsilon/2$-SOSP of $\boldsymbol{V}(\boldsymbol{\theta})$, then $\boldsymbol{\theta}$ is an ϵ-SOSP of $\hat{\boldsymbol{V}}_n(\boldsymbol{\theta})$, regardless of the value of the sample size n. Surprisingly, such results are attainable through further analysis, which we will discuss in the subsequent section. We will analyze non-degenerate stationary points using the results presented in this section. Non-degenerate stationary points are defined to include local maxima, thereby fulfilling the criteria for ϵ-SOSPs as desired.

3.4 Proof Sketch

The proof framework of Theorems 1, 2, and 3 are similar. Thus, we only sketch the proof of Theorem 1. First, we construct an ϵ-covering net of the ball $B^{|\mathcal{S}||\mathcal{A}|}(r)$, $\boldsymbol{\theta}_\epsilon = \{\boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_{N_\epsilon}\}$. Next, we define four events $\boldsymbol{E} = \sup_{\boldsymbol{\theta}} \left|\hat{\boldsymbol{V}}_n(\boldsymbol{\theta}) - \boldsymbol{V}(\boldsymbol{\theta})\right| \geq t$, $\boldsymbol{E}_1 = \{\sup_{\boldsymbol{\theta}} \left|\frac{1}{n}\sum_{i=1}^n (\hat{V}(\boldsymbol{\theta}, \boldsymbol{\tau}_i) - \hat{V}(\boldsymbol{\theta}_{i(\boldsymbol{\theta})}, \boldsymbol{\tau}_i))\right| \geq \frac{t}{3}\}$, $\boldsymbol{E}_2 = \{\sup_{i\in[N_\epsilon]} \left|\frac{1}{n}\sum_{i=1}^n \hat{V}(\boldsymbol{\theta}_{i(\boldsymbol{\theta})}, \boldsymbol{\tau}_i) - \mathbb{E}[V(\boldsymbol{\theta}_{i(\boldsymbol{\theta})}, \boldsymbol{\tau})]\right| \geq \frac{t}{3}\}$, and $\boldsymbol{E}_3 = \{\sup_{\boldsymbol{\theta}} \left|\mathbb{E}[V(\boldsymbol{\theta}_{i(\boldsymbol{\theta})}, \boldsymbol{\tau})] - \mathbb{E}[V(\boldsymbol{\theta}, \boldsymbol{\tau})]\right| \geq \frac{t}{3}\}$. Now we have $\mathbb{P}(\boldsymbol{E}) \leq \mathbb{P}(\boldsymbol{E}_1) + \mathbb{P}(\boldsymbol{E}_2) + \mathbb{P}(\boldsymbol{E}_3)$ and we bound them separately. For $\mathbb{P}(\boldsymbol{E}_1)$ and $\mathbb{P}(\boldsymbol{E}_3)$, we use the Lipschitz constant of the objective function and the properties of ϵ-net to prove $\mathbb{P}(\boldsymbol{E}_1) \leq \mathcal{E}/2$ and $\mathbb{P}(\boldsymbol{E}_3) = 0$. Since bounding $\mathbb{P}(\boldsymbol{E}_2)$ necessitating a further decomposition, the analysis was extracted in lemmas provided in the appendix. They subdivide $\mathbb{P}(\boldsymbol{E}_2)$ into a concentration inequality and a bias term due to truncation. With specific choices of the sample size and truncation length, it satisfies that $\mathbb{P}(\boldsymbol{E}_2) \leq \mathcal{E}/2$. Lastly, by combining the bounds of the three terms, we arrive at the desired results.

4 Characterizing the Landscape of Policy Gradient Methods

Given a set of stationary points $\{\boldsymbol{\theta}'_1, \ldots, \boldsymbol{\theta}'_{m_1}\}$ of the empirical estimate of the value function $\hat{\boldsymbol{V}}_n(\boldsymbol{\theta})$, and another set of stationary points $\{\boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_{m_2}\}$ of the expected value function $\boldsymbol{V}(\boldsymbol{\theta})$, what are the properties of them? Does there exist a clear relationship between the two sets of stationary points? This section is dedicated to addressing these questions.

Definition 2 (Non-degenerate stationary points [47]). *If a stationary point $\boldsymbol{x}$ is said to be a non-degenerate stationary point of $\boldsymbol{F}(\boldsymbol{x})$, then it satisfies*

$$\inf_i \left|\lambda_i\left(\nabla^2 \boldsymbol{F}(\boldsymbol{x})\right)\right| \geq \zeta$$

where $\lambda_i\left(\nabla^2 \boldsymbol{F}(\boldsymbol{x})\right)$ denotes the i-th eigenvalue of the Hessian $\nabla^2 \boldsymbol{F}(\boldsymbol{x})$ and ζ is a positive constant.

Definition 3 (Index of non-degenerate stationary points [48]). *The index of a symmetric non-degenerate matrix is the number of its negative eigenvalues, and the index of a non-degenerate stationary point $\boldsymbol{x}$ of a smooth function $\boldsymbol{F}$ is simply the index of its Hessian $\nabla^2 \boldsymbol{F}(\boldsymbol{x})$.*

We focus on the non-degenerate stationary points of $\hat{\boldsymbol{V}}_n(\boldsymbol{\theta})$ and $\boldsymbol{V}(\boldsymbol{\theta})$. Non-degenerate stationary points are stationary points that are geometrically isolated, meaning they are distinct points rather than flat areas. Non-degenerate stationary points include local minima, local maxima, and non-degenerate saddle points, in contrast to degenerate stationary points, which are degenerate saddle points. Now we present our main result.

Theorem 4. *Suppose assumption 1, 2, and 3 hold. Then if* $n \geq \max\left\{\frac{C^2 H G_h |\mathcal{S}||\mathcal{A}| \log(n/\mathcal{E})}{(1-\gamma)^2}, \frac{CHG_t r}{(1-\gamma)}\right\}$ *where* C *is a constant, for* $k \in \{1, \ldots, m\}$, *there exists a non-degenerate stationary point* $\boldsymbol{\theta}'_k$ *of* $\hat{\boldsymbol{V}}_n(\boldsymbol{\theta})$ *which corresponds to the non-degenerate stationary point* $\boldsymbol{\theta}_k$ *of* $\boldsymbol{V}(\boldsymbol{\theta})$. *In addition,* $\boldsymbol{\theta}'_k$ *and* $\boldsymbol{\theta}_k$ *have the same non-degenerate index and they satisfy*

$$\|\boldsymbol{\theta}'_k - \boldsymbol{\theta}_k\|_2 \leq \sqrt{\frac{C^2 H G_g |\mathcal{S}||\mathcal{A}| \log(n/\mathcal{E})}{n(1-\gamma)^2}} + \frac{CHG_h r}{n(1-\gamma)}$$

with probability at least $1 - \mathcal{E}$.

Theorem 4 establishes a one-to-one correspondence between the non-degenerate stationary points of the empirical objective function $\hat{\boldsymbol{V}}_n(\boldsymbol{\theta})$ and those of the population objective function $\boldsymbol{V}(\boldsymbol{\theta})$. Furthermore, the corresponding pairs have the same non-degenerate index, indicating their corresponding Hessian matrices have the same properties, such as the same number of negative eigenvalues. Therefore, given a sufficiently large sample size, the stationary points of $\hat{\boldsymbol{V}}_n(\boldsymbol{\theta})$ and $\boldsymbol{V}(\boldsymbol{\theta})$ share similar properties in that they have exactly matching local minima, local maxima, and saddle points. Although Theorem 4 focuses on non-degenerate stationary points, Theorem 2 and Theorem 3 ensure that for any degenerate point $\boldsymbol{\theta}$, the gradient and Hessian of $\hat{\boldsymbol{V}}_n(\boldsymbol{\theta})$ are close to those of $\boldsymbol{V}(\boldsymbol{\theta})$.

Before proving the theorem, we first introduce two lemmas proved in [33]. The first lemma follows a classical result in differential topology (Theorem 14.4.4 in [48]). It states that under certain conditions, the behavior of the critical points of two distinct functions aligns within a particular region. Lemma 3 states that the set of non-degenerate critical points can be decomposed into disjoint sets such that each set contains at most one critical point.

Lemma 2. *Let* $D \subseteq \mathbb{R}^d$ *be a compact set with a* C^2 *boundary* ∂D, *and* $f, g : A \to \mathbb{R}$ *be* C^2 *functions defined on an open set* A, *with* $D \subseteq A$. *Assume that for all* $\boldsymbol{\theta} \in \partial D$ *and all* $t \in [0,1]$, $t\nabla f(\boldsymbol{\theta}) + (1-t)\nabla g(\boldsymbol{\theta}) \neq \mathbf{0}$. *Finally, assume that the Hessian* $\nabla^2 f(\boldsymbol{\theta})$ *is non-degenerate and has index equal to* r *for all* $\boldsymbol{\theta} \in D$. *Then the following properties hold:*
(1) If g *has no critical point in* D, *then* f *has no critical point in* D.
(2) If g *has a unique critical point* $\boldsymbol{\theta}$ *in* D *that is non-degenerate with an index of* r, *then* f *also has a unique critical point* $\boldsymbol{\theta}'$ *in* D *with the index equal to* r.

Lemma 3. *Suppose that* $F(\boldsymbol{\theta}) : \boldsymbol{\Theta} \to \mathbb{R}$ *is a* C^2 *function where* $\boldsymbol{\theta} \in \boldsymbol{\Theta}$. *Assume that* $\{\boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_m\}$ *is its non-degenerate critical points and let* $D = \{\boldsymbol{\theta} \in \boldsymbol{\Theta} : \|\nabla F(\boldsymbol{\theta})\|_2 < \epsilon$ *and* $\inf_i \left|\lambda_i\left(\nabla^2 F(\boldsymbol{\theta})\right)\right| \geq \zeta\}$. *Then* D *can be decomposed into (at most) countably components, with each component containing either exactly one critical point, or no critical point. Concretely, there exist disjoint open sets* $\{D_k\}_{k \in \mathbb{N}}$, *with* D_k *possibly empty for* $k \geq m+1$, *such that*

$$D = \cup_{k=1}^{\infty} D_k.$$

Furthermore, $\boldsymbol{\theta}_k \in D_k$ *for* $1 \leq k \leq m$ *and each* $D_i, k \geq m+1$ *contains no stationary points.*

We now present the proof of Theorem 4.

Proof. Consider that the set $\{\boldsymbol{\theta}_1, \boldsymbol{\theta}_2, \cdots, \boldsymbol{\theta}_m\}$ consists of the non-degenerate critical points for $\boldsymbol{V}(\boldsymbol{\theta})$. As defined, for any $\boldsymbol{\theta}_k$, it follows

$$\inf_i \left|\lambda_i^k \left(\nabla^2 \boldsymbol{V}\left(\boldsymbol{\theta}_k\right)\right)\right| \geq \zeta,$$

where $\lambda_i^k \left(\nabla^2 \boldsymbol{V}\left(\boldsymbol{\theta}_k\right)\right)$ denotes the i-th eigenvalue of the Hessian $\nabla^2 \boldsymbol{V}\left(\boldsymbol{\theta}_k\right)$ and ζ is a constant. Let $D = \left\{\boldsymbol{\theta} \in B^{|\mathcal{S}||\mathcal{A}|}(r) \mid \|\nabla \boldsymbol{V}(\boldsymbol{\theta})\|_2 \leq \epsilon \text{ and } \inf_i \left|\lambda_i \left(\nabla^2 \boldsymbol{V}\left(\boldsymbol{\theta}_k\right)\right)\right| \geq \zeta\right\}$. By Lemma 3, $D = \cup_{k=1}^{\infty} D_k$ where each D_k is a disjoint component with $\boldsymbol{\theta}_k \in D_k$ for $k \leq m$ and D_k does not contain any critical point of $\boldsymbol{V}(\boldsymbol{\theta})$ for $k \geq m+1$. Moreover, by the continuity of $\nabla \boldsymbol{V}(\boldsymbol{\theta})$, it yields $\|\nabla \boldsymbol{V}(\boldsymbol{\theta})\|_2 = \epsilon$ for $\boldsymbol{\theta} \in \partial D_k$. Note that the value of ϵ below is a function related to n. According to Theorem 2, when the sample is sufficiently large, we have

$$\sup_{\boldsymbol{\theta} \in B^{|\mathcal{S}||\mathcal{A}|}(r)} \left\|\nabla \hat{\boldsymbol{V}}_n(\boldsymbol{\theta}) - \nabla \boldsymbol{V}(\boldsymbol{\theta})\right\|_2 \leq \sqrt{\frac{64HG_g|\mathcal{S}||\mathcal{A}|\log(n/\mathcal{E})}{n(1-\gamma)^2}} + \frac{18HG_h r}{n(1-\gamma)} \triangleq \frac{\epsilon}{2}$$

with probability at least $1-\mathcal{E}$. This implies that for arbitrary $\boldsymbol{\theta} \in D_k$, we have

$$\begin{aligned}
&\inf_{\boldsymbol{\theta} \in B^{|\mathcal{S}||\mathcal{A}|}(r)} \left\|t\nabla \hat{\boldsymbol{V}}_n(\boldsymbol{\theta}) + (1-t)\nabla \boldsymbol{V}(\boldsymbol{\theta})\right\|_2 \\
&= \inf_{\boldsymbol{\theta} \in B^{|\mathcal{S}||\mathcal{A}|}(r)} \left\|t\left(\nabla \hat{\boldsymbol{V}}_n(\boldsymbol{\theta}) - \nabla \boldsymbol{V}(\boldsymbol{\theta})\right) + \nabla \boldsymbol{V}(\boldsymbol{\theta})\right\|_2 \\
&\geq \inf_{\boldsymbol{\theta} \in B^{|\mathcal{S}||\mathcal{A}|}(r)} \|\nabla \boldsymbol{V}(\boldsymbol{\theta})\|_2 - \sup_{\boldsymbol{\theta} \in B^{|\mathcal{S}||\mathcal{A}|}(r)} t\left\|\nabla \hat{\boldsymbol{V}}_n(\boldsymbol{\theta}) - \nabla \boldsymbol{V}(\boldsymbol{\theta})\right\|_2 \\
&\geq \frac{\epsilon}{2}.
\end{aligned} \tag{6}$$

Similarly, by Theorem 3, when n is sufficiently large we have

$$\begin{aligned}
\sup_{\boldsymbol{\theta} \in B^{|\mathcal{S}||\mathcal{A}|}(r)} \left\|\nabla^2 \hat{\boldsymbol{V}}_n(\boldsymbol{\theta}) - \nabla^2 \boldsymbol{V}(\boldsymbol{\theta})\right\|_{\text{op}} &\leq \sqrt{\frac{128HG_h|\mathcal{S}||\mathcal{A}|\log(n/\mathcal{E})}{n(1-\gamma)^2}} + \frac{18HG_t r}{n(1-\gamma)} \\
&\leq \frac{\zeta}{2}
\end{aligned}$$

with probability at least $1-\mathcal{E}$. Assuming that $\boldsymbol{y} \in \mathbb{R}^d$ is a vector satisfying $\boldsymbol{y}^T\boldsymbol{y} = 1$, we bound $\lambda_i^k \left(\nabla^2 \hat{\boldsymbol{V}}_n(\boldsymbol{\theta})\right)$ for arbitrary $\boldsymbol{\theta} \in D_k$ as follows:

$$
\begin{aligned}
&\inf_{\boldsymbol{\theta}\in B^{|\mathcal{S}||\mathcal{A}|}(r)} \left|\lambda_i^k\left(\nabla^2\hat{\boldsymbol{V}}_n(\boldsymbol{\theta})\right)\right| \\
&= \inf_{\boldsymbol{\theta}\in D_k} \min_{\boldsymbol{y}^T\boldsymbol{y}=1} \left|\boldsymbol{y}^T\nabla^2\hat{\boldsymbol{V}}_n(\boldsymbol{\theta})\boldsymbol{y}\right| \\
&= \inf_{\boldsymbol{\theta}\in B^{|\mathcal{S}||\mathcal{A}|}(r)} \min_{\boldsymbol{y}^T\boldsymbol{y}=1} \left|\boldsymbol{y}^T\left(\nabla^2\hat{\boldsymbol{V}}_n(\boldsymbol{\theta})-\nabla^2\boldsymbol{V}(\boldsymbol{\theta})\right)\boldsymbol{y}+\boldsymbol{y}^T\nabla^2\boldsymbol{V}(\boldsymbol{\theta})\boldsymbol{y}\right| \\
&\geq \inf_{\boldsymbol{\theta}\in B^{|\mathcal{S}||\mathcal{A}|}(r)} \min_{\boldsymbol{y}^T\boldsymbol{y}=1} \left|\boldsymbol{y}^T\nabla^2\boldsymbol{V}(\boldsymbol{\theta})\boldsymbol{y}\right| - \min_{\boldsymbol{y}^T\boldsymbol{y}=1} \left|\boldsymbol{y}^T\left(\nabla^2\hat{\boldsymbol{V}}_n(\boldsymbol{\theta})-\nabla^2\boldsymbol{V}(\boldsymbol{\theta})\right)\boldsymbol{y}\right| \\
&\geq \inf_{\boldsymbol{\theta}\in B^{|\mathcal{S}||\mathcal{A}|}(r)} \min_{\boldsymbol{y}^T\boldsymbol{y}=1} \left|\boldsymbol{y}^T\nabla^2\boldsymbol{V}(\boldsymbol{\theta})\boldsymbol{y}\right| - \max_{\boldsymbol{y}^T\boldsymbol{y}=1} \left|\boldsymbol{y}^T\left(\nabla^2\hat{\boldsymbol{V}}_n(\boldsymbol{\theta})-\nabla^2\boldsymbol{V}(\boldsymbol{\theta})\right)\boldsymbol{y}\right| \\
&= \inf_{\boldsymbol{\theta}\in B^{|\mathcal{S}||\mathcal{A}|}(r)} \inf_i \left|\lambda_i^k\left(\nabla^2 V(\boldsymbol{\theta}_k)\right)\right| - \left\|\nabla^2\hat{\boldsymbol{V}}_n(\boldsymbol{\theta})-\nabla^2\boldsymbol{V}(\boldsymbol{\theta})\right\|_{\mathrm{op}} \\
&\geq \frac{\zeta}{2}. \qquad (7)
\end{aligned}
$$

This implies that in each set $D_k, \nabla^2\hat{\boldsymbol{V}}_n(\boldsymbol{\theta})$ has no zero eigenvalues. Combining (6) and (7), by Lemma 2 we know that if the population risk $\boldsymbol{V}(\boldsymbol{\theta})$ has no critical point in D_k, then the empirical risk $\hat{\boldsymbol{V}}_n(\boldsymbol{\theta})$ has also no critical point in D_k. By Lemma 2, we can also obtain that in D_k, if $\boldsymbol{V}(\boldsymbol{\theta})$ has a unique critical point $\boldsymbol{\theta}_k$ with non-degenerate index r_k, then $\hat{\boldsymbol{V}}_n(\boldsymbol{\theta})$ also has a unique critical point $\boldsymbol{\theta}'_k$ in D_k with the same index r_k. This establishes the first conclusion.

Next, we bound the distance between the corresponding critical points of $\boldsymbol{V}(\boldsymbol{\theta})$ and $\hat{\boldsymbol{V}}_n(\boldsymbol{\theta})$. Assume that in $D_k, \boldsymbol{V}(\boldsymbol{\theta})$ has a unique critical point $\boldsymbol{\theta}_k$ and $\hat{\boldsymbol{V}}_n(\boldsymbol{\theta})$ also has a unique critical point $\boldsymbol{\theta}'_k$. Then, there exists $t\in[0,1]$ such that for any $\boldsymbol{z}\in\partial\mathrm{B}^d(1)$, we have

$$
\begin{aligned}
\epsilon &\geq \left\|\nabla\boldsymbol{V}\left(\boldsymbol{\theta}'_k\right)\right\|_2 \\
&= \max_{\boldsymbol{z}^T\boldsymbol{z}=1} \left\langle\nabla\boldsymbol{V}\left(\boldsymbol{\theta}'_k\right),\boldsymbol{z}\right\rangle \\
&= \max_{\boldsymbol{z}^T\boldsymbol{z}=1} \left\langle\nabla\boldsymbol{V}\left(\boldsymbol{\theta}_k\right),\boldsymbol{z}\right\rangle + \left\langle\nabla^2\boldsymbol{V}\left(\boldsymbol{\theta}_k+t\left(\boldsymbol{\theta}'_k-\boldsymbol{\theta}_k\right)\right)\left(\boldsymbol{\theta}'_k-\boldsymbol{\theta}_k\right),\boldsymbol{z}\right\rangle \\
&\overset{(a)}{\geq} \left\langle\left(\nabla^2\boldsymbol{V}\left(\boldsymbol{\theta}_k\right)\right)^2\left(\boldsymbol{\theta}'_k-\boldsymbol{\theta}_k\right),\left(\boldsymbol{\theta}'_k-\boldsymbol{\theta}_k\right)\right\rangle^{1/2} \\
&\overset{(b)}{\geq} \left\|\boldsymbol{\theta}'_k-\boldsymbol{\theta}_k\right\|_2,
\end{aligned}
$$

where (a) holds since $\nabla\boldsymbol{V}\left(\boldsymbol{\theta}_k\right)=\mathbf{0}$ and (b) holds since $\boldsymbol{\theta}_k+t\left(\boldsymbol{\theta}'_k-\boldsymbol{\theta}_k\right)$ is in D_k and for any $\boldsymbol{\theta}\in D_k$ we have $\inf_i\left|\lambda_i\left(\nabla^2\boldsymbol{V}(\boldsymbol{\theta})\right)\right|\geq\zeta$. Consider the conditions in Theorem 2 and Theorem 3 we obtain that if $n\geq\frac{CH}{(1-\gamma)}\max\left\{\frac{CG_h|\mathcal{S}||\mathcal{A}|\log(n/\mathcal{E})}{(1-\gamma)},G_t r\right\}$, then

$$
\left\|\boldsymbol{\theta}'_k-\boldsymbol{\theta}_k\right\|_2\leq\sqrt{\frac{C^2HG_g|\mathcal{S}||\mathcal{A}|\log(n/\mathcal{E})}{n(1-\gamma)^2}}+\frac{CHG_h r}{n(1-\gamma)}
$$

holds with probability at least $1-\mathcal{E}$, where $C=\frac{18}{\zeta}$ is a constant. □

5 Conclusion

This paper provides a theoretical analysis on the optimization landscape of stochastic policy gradient methods. We establish uniform convergence of the empirical objective function, its gradient, and its Hessian, to their expected counterparts at the rate of $\tilde{O}(\sqrt{|\mathcal{S}||\mathcal{A}|}/(1-\gamma)\sqrt{n})$. Furthermore, we characterize the optimization landscape by establishing a one-to-one correspondence between the non-degenerate stationary points of the empirical and population objectives. Although this paper focuses on the landscape from a statistical and topological perspective, it still matches some existing results even when employing vanilla policy gradient methods. We hope these results will provide more insights and deeper understanding of stochastic policy gradient methods.

References

1. Silver, D., Huang, A., Maddison, C.J., Guez, A., Sifre, L., van den Driessche, G., Schrittwieser, J., Antonoglou, I., Panneershelvam, V., Lanctot, M., Dieleman, S., Grewe, D., Nham, J., Kalchbrenner, N., Sutskever, I., Lillicrap, T., Leach, M., Kavukcuoglu, K., Graepel, T., Hassabis, D.: Mastering the game of go with deep neural networks and tree search. Nature 529(7587), 484-489 (2016)
2. Silver, D., Schrittwieser, J., Simonyan, K., Antonoglou, I., Huang, A., Guez, A., Hubert, T., Baker, L., Lai, M., Bolton, A., Chen, Y., Lillicrap, T., Hui, F., Sifre, L., van den Driessche, G., Graepel, T., Hassabis, D.: Mastering the game of go without human knowledge. Nature 550(7676), 354-359 (2017)
3. Nalpantidis, L., S. Polydoros, A.: Survey of model-based reinforcement learning: Applications on robotics. Journal of Intelligent Robotic Systems 86(2), 153-173 (2017)
4. Christiano, P.F., Leike, J., Brown, T., Martic, M., Legg, S., Amodei, D.: Deep reinforcement learning from human preferences. In: Advances in Neural Information Processing Systems. vol. 30 (2017)
5. OpenAI: Gpt-4 technical report (2024)
6. Williams, R.J.: Simple statistical gradient-following algorithms for connectionist reinforcement learning. Machine Learning p. 8:229-256 (1992)
7. Sutton, R.S., McAllester, D., Singh, S., Mansour, Y.: Policy gradient methods for reinforcement learning with function approximation. In: Advances in Neural Information Processing Systems. vol. 12. MIT Press (1999)
8. Kakade, S.M.: A natural policy gradient. In: Advances in Neural Information Processing Systems. vol. 14 (2001)
9. Baxter, J., Bartlett, P.L.: Infinite-horizon policy-gradient estimation. J. Artif. Int. Res. 15(1), 319-350 (nov 2001)
10. Agarwal, A., Kakade, S.M., Lee, J.D., Mahajan, G.: On the theory of policy gradient methods: optimality, approximation, and distribution shift. J. Mach. Learn. Res. 22(1) (jan 2021)
11. Mei, J., Xiao, C., Szepesvari, C., Schuurmans, D.: On the global convergence rates of softmax policy gradient methods (2022)
12. Mei, J., Gao, Y., Dai, B., Szepesvari, C., Schuurmans, D.: Leveraging non-uniformity in first-order non-convex optimization. In: Proceedings of the 38th International Conference on Machine Learning. Proceedings of Machine Learning Research, vol. 139, pp. 7555-7564. PMLR (18-24 Jul 2021)

13. Xiao, L.: On the convergence rates of policy gradient methods. Journal of Machine Learning Research 23(282), 1-36 (2022)
14. Li, G., Wei, Y., Chi, Y., Gu, Y., Chen, Y.: Softmax policy gradient methods can take exponential time to converge. In: In Proceedings of Thirty Fourth Conference on Learning Theory. vol. 134, p. 3107-3110 (2021)
15. Zhang, K., Koppel, A., Zhu, H., BaÅsar, T.: Global convergence of policy gradient methods to (almost) locally optimal policies. SIAM Journal on Control and Optimization 58(6), 3586-3612 (2020)
16. Mei, J., Dai, B., Xiao, C., Szepesvari, C., Schuurmans, D.: Understanding the effect of stochasticity in policy optimization. In: Advances in Neural Information Processing Systems. vol. 34, pp. 19339-19351 (2021)
17. Mei, J., Zhong, Z., Dai, B., Agarwal, A., Szepesvari, C., Schuurmans, D.:Stochastic gradient succeeds for bandits. In: Proceedings of the 40th International Conference on Machine Learning. Proceedings of Machine Learning Research, vol. 202, pp. 24325-24360. PMLR (23-29 Jul 2023)
18. Lu, M., Aghaei, M., Raj, A., Vaswani, S.: Practical principled policy optimization for finite MDPs. In: OPT 2023: Optimization for Machine Learning (2023)
19. Ding, Y., Zhang, J., Lavaei, J.: On the global optimum convergence of momentum-based policy gradient. In: Proceedings of The 25th International Conference on Artificial Intelligence and Statistics. Proceedings of Machine Learning Research, vol. 151, pp. 1910-1934. PMLR (28-30 Mar 2022)
20. Yuan,R.,Gower,R.M.,Lazaric,A.:A general sample complexity analysis of vanilla policy gradient. In: Proceedings of The 25th International Conference on Artificial Intelligence and Statistics. vol. 151, pp. 3332-3380. PMLR (2022)
21. Masiha, S., Salehkaleybar, S., He, N., Kiyavash, N., Thiran, P.: Stochastic second-order methods improve best-known sample complexity of SGD for gradient-dominated functions. In: Advances in Neural Information Processing Systems (2022)
22. Liu, Y., Zhang, K., Basar, T., Yin, W.: An improved analysis of (variance-reduced) policy gradient and natural policy gradient methods. In: Advances in Neural Information Processing Systems. vol. 33, pp. 7624-7636 (2020)
23. Fatkhullin, I., Barakat, A., Kireeva, A., He, N.: Stochastic policy gradient methods: Improved sample complexity for Fisher-non-degenerate policies. In: Proceedings of the 40th International Conference on Machine Learning. Proceedings of Machine Learning Research, vol. 202, pp. 9827-9869. PMLR (23-29 Jul 2023)
24. Fatkhullin, I., Barakat, A., Kireeva, A., He, N.: Stochastic policy gradient methods: Improved sample complexity for Fisher-non-degenerate policies. In: Proceedings of the 40th International Conference on Machine Learning. Proceedings of Machine Learning Research, vol. 202, pp. 9827-9869. PMLR (23-29 Jul 2023)
25. Shen, Z., Ribeiro, A., Hassani, H., Qian, H., Mi, C.: Hessian aided policy gradient. In: Proceedings of the 36th International Conference on Machine Learning. Proceedings of Machine Learning Research, vol. 97, pp. 5729-5738. PMLR (09-15 Jun 2019)
26. Wachi, A., Wei, Y., Sui, Y.: Safe policy optimization with local generalized linear function approximations. CoRR abs/2111.04894 (2021)
27. Papini, M.: Safe policy optimization (2021)
28. Yang, L., Zheng, Q., Pan, G.: Sample complexity of policy gradient finding second-order stationary points. Proceedings of the AAAI Conference on Artificial Intelligence 35(12), 10630-10638 (May 2021)
29. Maniyar, M.P., Mondal, A., A., P.L., Bhatnagar, S.: A cubic regularized policy newton algorithm for reinforcement learning (2023)

30. Ge, R., Lee, J.D., Ma, T.: Matrix completion has no spurious local minimum. In: Advances in Neural Information Processing Systems. vol. 29 (2016)
31. Kawaguchi, K.: Deep learning without poor local minima. In: Advances in Neural Information Processing Systems. vol. 29 (2016)
32. Sun, J., Qu, Q., Wright, J.: A geometric analysis of phase retrieval. CoRR abs/1602.06664 (2016)
33. Mei, S., Bai, Y., Montanari, A.: The landscape of empirical risk for non-convex losses (2017)
34. Ge, R., Ma, T.: On the optimization landscape of tensor decompositions (2017)
35. Jin, C., Ge, R., Netrapalli, P., Kakade, S.M., Jordan, M.I.: How to escape saddle points efficiently. In: Proceedings of the 34th International Conference on Machine Learning. Proceedings of Machine Learning Research, vol. 70, pp. 1724-1732. PMLR (06-11 Aug 2017)
36. Du, S., Lee, J., Li, H., Wang, L., Zhai, X.: Gradient descent finds global minima of deep neural networks. In: Proceedings of the 36th International Conference on Machine Learning. vol. 97, pp. 1675-1685. PMLR (09-15 Jun 2019)
37. Sun, R.: Optimization for deep learning: theory and algorithms. CoRR abs/1912.08957 (2019)
38. Soltanolkotabi, M., Javanmard, A., Lee, J.D.: Theoretical insights into the optimization landscape of over-parameterized shallow neural networks. IEEE Transactions on Information Theory 65(2), 742-769 (2019)
39. Liu, X.: Neural networks with complex-valued weights have no spurious local minima. CoRR abs/2103.07287 (2021)
40. Caramanis, C., Fotakis, D., Kalavasis, A., Kontonis, V., Tzamos, C.: Optimizing solution-samplers for combinatorial problems: The landscape of policy-gradient method. In: Thirty-seventh Conference on Neural Information Processing Systems (2023)
41. Duan, J., Li, J., Chen, X., Zhao, K., Li, S.E., Zhao, L.: Optimization landscape of policy gradient methods for discrete-time static output feedback. IEEE Transactions on Cybernetics p. 1-14 (2024)
42. Puterman, M.L.: Markov Decision Processes: Discrete Stochastic Dynamic Programming. John Wiley Sons, Inc (1994)
43. Sutton, R.S., McAllester, D., Singh, S., Mansour, Y.: Policy gradient methods for reinforcement learning with function approximation. In Advances in neural information processing systems p. 1057-1063 (2000)
44. Vapnik, V.N., Vapnik, V., et al.: Statistical learning theory (1998)
45. Khodadadian, S., Jhunjhunwala, P.R., Varma, S.M., Maguluri, S.T.: On the linear convergence of natural policy gradient algorithm. CoRR (2021)
46. Nesterov, Y., Polyak, B.T.: Cubic regularization of newton method and its global performance. Mathematical programming 108(1), 177-205 (2006)
47. Dubrovin, B., Fomenko, A., Novikov, S.: On differentiable functions with isolated critical points. Topology 8(4), 361-369 (1969). https://doi.org/https://doi.org/10.1016/0040-9383(69)90022-6
48. Dubrovin, B., Fomenko, A., Novikov, S.: Modern geometry-methods and applications: Part II: The geometry and topology of manifolds. Springer Science Business Media (2012)
49. Vershynin, R.: Introduction to the non-asymptotic analysis of random matrices. Compressed Sensing (Y. C. Eldar and G. Kutyniok, eds.), Cambridge University Press, Cambridge (2012)
50. Hoeffding, W.: Probability inequalities for sums of bounded random variables. Journal of the American Statistical Association, 58(301), 13-30. (1963)

A Technical Lemmas

Lemma 4 ([49]). *Let $\boldsymbol{x} \in \mathbb{R}^d$ and $\boldsymbol{\lambda}_\epsilon = \{\boldsymbol{\lambda}_1, \ldots, \boldsymbol{\lambda}_N\}$ be an ϵ-covering net of $B_{\mathbf{0}}^d(1)$, then*

$$\|\boldsymbol{x}\|_2 \leq \frac{1}{1-\epsilon} \sup_{\boldsymbol{\lambda} \in \boldsymbol{\lambda}_\epsilon} \langle \boldsymbol{\lambda}, \boldsymbol{x} \rangle.$$

Lemma 5 ([49]). *Let $\boldsymbol{X} \in \mathbb{R}^{d \times d}$ be a symmetric matrix and $\boldsymbol{\lambda}_\epsilon = \{\boldsymbol{\lambda}_1, \ldots, \boldsymbol{\lambda}_N\}$ be an ϵ-covering net of $B_{\mathbf{0}}^d(1)$, then*

$$\|\boldsymbol{X}\|_{op} \leq \frac{1}{1-2\epsilon} \sup_{\boldsymbol{\lambda} \in \boldsymbol{\lambda}_\epsilon} |\langle \boldsymbol{\lambda}, \boldsymbol{X}\boldsymbol{\lambda} \rangle| .$$

Lemma 6. *Suppose assumption 1, 2, and 3 hold. For $\hat{V}_H(\boldsymbol{\theta}, \boldsymbol{\tau})$ defined in (2), we have*

$$\left\| \nabla \hat{V}_H(\boldsymbol{\theta}, \boldsymbol{\tau}) \right\|_2 \leq \frac{HG_g}{1-\gamma}, \qquad \left\| \nabla^2_{\boldsymbol{\theta}} \hat{V}_H(\boldsymbol{\theta}, \boldsymbol{\tau}) \right\|_F \leq \frac{HG_h}{1-\gamma},$$

and

$$\left\| \nabla^3_{\boldsymbol{\theta}} \hat{V}_H(\boldsymbol{\theta}, \boldsymbol{\tau}) \right\|_F \leq \frac{HG_t}{1-\gamma}$$

for any trajectory $\boldsymbol{\tau}$ with horizon $H \in \mathbb{N} \cup \{\infty\}$ and $\boldsymbol{\theta} \in B^{|\mathcal{S}||\mathcal{A}|}(r)$.

Proof. Let $j \in \{1, 2, 3\}$. From (4), we have

$$\begin{aligned} \left\| \nabla^j_{\boldsymbol{\theta}} \hat{V}_H(\boldsymbol{\theta}, \boldsymbol{\tau}) \right\| &\leq \sum_{t=0}^{H-1} \gamma^t r(s_t, a_t) \sum_{k=0}^{H-1} \left\| \nabla^j_{\boldsymbol{\theta}} \log \pi_{\boldsymbol{\theta}}(a_k|s_k) \right\| \\ &\leq \frac{H}{1-\gamma} \sup_i \left\| \nabla^j_{\boldsymbol{\theta}} \log \pi_{\boldsymbol{\theta}}(a_k|s_k) \right\| \\ &\leq \frac{HG_j^2}{1-\gamma} \end{aligned}$$

where $G_1 = G_g, G_2 = G_h$, and $G_3 = G_t$. □

Lemma 7. *Under assumption 2 and 3, we have*

$$\left\| \mathbb{E}\left[\nabla_{\boldsymbol{\theta}} \hat{V}_H(\boldsymbol{\theta}, \boldsymbol{\tau}) \right] - \mathbb{E}\left[\nabla_{\boldsymbol{\theta}} V(\boldsymbol{\theta}, \boldsymbol{\tau}) \right] \right\|_2 \leq \frac{\gamma^H G_g}{1-\gamma} \sqrt{\frac{1}{1-\gamma} + H}.$$

and

$$\left\| \mathbb{E}\left[\nabla^2_{\boldsymbol{\theta}} \hat{V}_H(\boldsymbol{\theta}, \boldsymbol{\tau}) \right] - \mathbb{E}\left[\nabla^2_{\boldsymbol{\theta}} V(\boldsymbol{\theta}, \boldsymbol{\tau}) \right] \right\|_F \leq \frac{\gamma^H G_h}{1-\gamma} \sqrt{\frac{1}{1-\gamma} + H}.$$

While this result was known [20], we included a proof for the sake of completeness.

Proof. Let $j \in \{1,2\}$. From (3), we have

$$\begin{aligned}
&\left\|\mathbb{E}\left[\nabla_{\boldsymbol{\theta}}^{j}\hat{V}_H(\boldsymbol{\theta},\boldsymbol{\tau})\right]-\mathbb{E}\left[\nabla_{\boldsymbol{\theta}}^{j}V(\boldsymbol{\theta},\boldsymbol{\tau})\right]\right\|^2\\
&\leq \mathbb{E}_\tau\left[\left\|\sum_{t=H}^{\infty}\gamma^{t/2}r\left(s_t,a_t\right)\gamma^{t/2}\left(\sum_{k=0}^{t}\nabla_\theta^j\log\pi_\theta\left(a_k\mid s_k\right)\right)\right\|^2\right]\\
&\leq \mathbb{E}_\tau\left[\left(\sum_{t=H}^{\infty}\gamma^t r\left(s_t,a_t\right)^2\right)\left(\sum_{k=H}^{\infty}\gamma^k\left\|\sum_{k'=0}^{k}\nabla_\theta^j\log\pi_\theta\left(a_{k'}\mid s_{k'}\right)\right\|^2\right)\right]\\
&\leq \frac{\gamma^H}{1-\gamma}\mathbb{E}_\tau\left[\sum_{k=H}^{\infty}\gamma^k\left\|\sum_{k'=0}^{k}\nabla_\theta^j\log\pi_\theta\left(a_{k'}\mid s_{k'}\right)\right\|^2\right]\\
&= \frac{\gamma^H}{1-\gamma}\sum_{k=H}^{\infty}\gamma^k\sum_{k'=0}^{k}\mathbb{E}_\tau\left[\left\|\nabla_\theta^j\log\pi_\theta\left(a_{k'}\mid s_{k'}\right)\right\|^2\right]\\
&\leq \frac{G_j^2\gamma^H}{1-\gamma}\sum_{k=H}^{\infty}\gamma^k(k+1)\\
&= \frac{G_j^2\gamma^{2H}}{1-\gamma}\sum_{k=0}^{\infty}\gamma^k(k+1+H)\\
&= \left(\frac{1}{1-\gamma}+H\right)\frac{G_j^2\gamma^{2H}}{(1-\gamma)^2}
\end{aligned}$$

where $G_1 = G_g$ and $G_2 = G_h$. □

Lemma 8. *For any $t > 0$, the estimate function $\hat{V}_H(\boldsymbol{\pi},\boldsymbol{\tau})$ with $H \geq \frac{\log(2/t(1-\gamma))}{\log(1/\gamma)}$ obeys*

$$\mathbb{P}\left(\left|\frac{1}{n}\sum_{i=1}^{n}\hat{V}_H(\boldsymbol{\theta},\boldsymbol{\tau}_i)-\mathbb{E}[V(\boldsymbol{\theta},\boldsymbol{\tau})]\right|\geq t\right)\leq 2\exp\left(-\frac{1}{2}nt^2\left(1-\gamma\right)^2\right).$$

Proof. Let $H \geq \frac{\log(2/t(1-\gamma))}{\log(1/\gamma)}$, we have

$$\begin{aligned}
&\left|\frac{1}{n}\sum_{i=1}^{n}\hat{V}_H(\boldsymbol{\theta},\boldsymbol{\tau}_i)-\mathbb{E}[V(\boldsymbol{\theta},\boldsymbol{\tau})]\right|\\
&= \left|\frac{1}{n}\sum_{i=1}^{n}\hat{V}_H(\boldsymbol{\theta},\boldsymbol{\tau}_i)-\mathbb{E}\left[\hat{V}_H(\boldsymbol{\theta},\boldsymbol{\tau})\right]+\mathbb{E}\left[\hat{V}_H(\boldsymbol{\theta},\boldsymbol{\tau})\right]-\mathbb{E}[V(\boldsymbol{\theta},\boldsymbol{\tau})]\right|\\
&\leq \left|\frac{1}{n}\sum_{i=1}^{n}\hat{V}_H(\boldsymbol{\theta},\boldsymbol{\tau}_i)-\mathbb{E}\left[\hat{V}_H(\boldsymbol{\theta},\boldsymbol{\tau})\right]\right|+\left|\mathbb{E}\left[\hat{V}_H(\boldsymbol{\theta},\boldsymbol{\tau})\right]-\mathbb{E}[V(\boldsymbol{\theta},\boldsymbol{\tau})]\right|\\
&\leq \left|\frac{1}{n}\sum_{i=1}^{n}\hat{V}_H(\boldsymbol{\theta},\boldsymbol{\tau}_i)-\mathbb{E}\left[\hat{V}_H(\boldsymbol{\theta},\boldsymbol{\tau})\right]\right|+\frac{\gamma^H}{1-\gamma}
\end{aligned}$$

$$\leq \left| \frac{1}{n} \sum_{i=1}^{n} \hat{V}_H(\boldsymbol{\theta}, \boldsymbol{\tau}_i) - \mathbb{E}\left[\hat{V}_H(\boldsymbol{\theta}, \boldsymbol{\tau})\right] \right| + \frac{t}{2}.$$

By Hoeffding's Inequality [50],

$$\mathbb{P}\left(\left| \frac{1}{n} \sum_{i=1}^{n} \hat{V}_H(\boldsymbol{\theta}, \boldsymbol{\tau}_i) - \mathbb{E}[\hat{V}_H(\boldsymbol{\theta}, \boldsymbol{\tau})] \right| \geq \frac{t}{2} \right) \leq 2\exp\left(-\frac{1}{2} n t^2 (1-\gamma)^2 \right).$$

Combining the above inequalities, we conclude the proof. □

Lemma 9. *For any $t > 0$, the gradient $\nabla \hat{V}_H(\boldsymbol{\theta}, \boldsymbol{\tau})$ with $H \geq \frac{\log(2G_g\sqrt{H+(1/1-\gamma)}/t(1-\gamma))}{\log(1/\gamma)}$ obeys*

$$\begin{aligned} &\mathbb{P}\left(\left\| \frac{1}{n} \sum_{i=1}^{n} \nabla_{\boldsymbol{\theta}} \hat{V}_H(\boldsymbol{\theta}, \boldsymbol{\tau}_i) - \mathbb{E}\left[\nabla_{\boldsymbol{\theta}} V(\boldsymbol{\theta}, \boldsymbol{\tau})\right] \right\|_2 \geq t \right) \\ &\leq 2\exp\left(-\frac{1}{32 H G_g} n t^2 (1-\gamma)^2 + |A| \log 5 \right). \end{aligned}$$

Proof. Let $H \geq \frac{\log(2G_g\sqrt{H+(1/1-\gamma)}/t(1-\gamma))}{\log(1/\gamma)}$, we have

$$\begin{aligned} &\left\| \frac{1}{n} \sum_{i=1}^{n} \nabla_{\boldsymbol{\theta}} \hat{V}_H(\boldsymbol{\theta}, \boldsymbol{\tau}_i) - \mathbb{E}\left[\nabla_{\boldsymbol{\theta}} V(\boldsymbol{\theta}, \boldsymbol{\tau})\right] \right\|_2 \\ &= \left\| \frac{1}{n} \sum_{i=1}^{n} \nabla_{\boldsymbol{\theta}} \hat{V}_H(\boldsymbol{\theta}, \boldsymbol{\tau}_i) - \mathbb{E}\left[\nabla_{\boldsymbol{\theta}} \hat{V}_H(\boldsymbol{\theta}, \boldsymbol{\tau})\right] + \mathbb{E}\left[\nabla_{\boldsymbol{\theta}} \hat{V}_H(\boldsymbol{\theta}, \boldsymbol{\tau})\right] - \mathbb{E}\left[\nabla_{\boldsymbol{\theta}} V(\boldsymbol{\theta}, \boldsymbol{\tau})\right] \right\|_2 \\ &\leq \left\| \frac{1}{n} \sum_{i=1}^{n} \nabla_{\boldsymbol{\theta}} \hat{V}_H(\boldsymbol{\theta}, \boldsymbol{\tau}_i) - \mathbb{E}\left[\nabla_{\boldsymbol{\theta}} \hat{V}_H(\boldsymbol{\theta}, \boldsymbol{\tau})\right] \right\|_2 + \left\| \mathbb{E}\left[\nabla_{\boldsymbol{\theta}} \hat{V}_H(\boldsymbol{\theta}, \boldsymbol{\tau})\right] - \mathbb{E}\left[\nabla_{\boldsymbol{\theta}} V(\boldsymbol{\theta}, \boldsymbol{\tau})\right] \right\|_2 \\ &\leq \left\| \frac{1}{n} \sum_{i=1}^{n} \nabla_{\boldsymbol{\theta}} \hat{V}_H(\boldsymbol{\theta}, \boldsymbol{\tau}_i) - \mathbb{E}\left[\nabla_{\boldsymbol{\theta}} \hat{V}_H(\boldsymbol{\theta}, \boldsymbol{\tau})\right] \right\|_2 + \frac{\gamma^H G_g}{1-\gamma} \sqrt{\frac{1}{1-\gamma} + H} \\ &\leq \left\| \frac{1}{n} \sum_{i=1}^{n} \nabla_{\boldsymbol{\theta}} \hat{V}_H(\boldsymbol{\theta}, \boldsymbol{\tau}_i) - \mathbb{E}\left[\nabla_{\boldsymbol{\theta}} \hat{V}_H(\boldsymbol{\theta}, \boldsymbol{\tau})\right] \right\|_2 + \frac{t}{2} \end{aligned}$$

where the second inequality holds since we utilize Lemma 7. Let $\boldsymbol{\lambda}_{1/2}$ be the $\frac{1}{2}$-covering net of $B^{|A|}(1)$ with $\left|\boldsymbol{\lambda}_{1/2}\right| \leq 5^{|A|}$. From Lemma 4 we know that

$$\begin{aligned} &\left\| \frac{1}{n} \sum_{i=1}^{n} \nabla_{\boldsymbol{\theta}} \hat{V}_H(\boldsymbol{\theta}, \boldsymbol{\tau}_i) - \mathbb{E}\left[\nabla_{\boldsymbol{\theta}} \hat{V}_H(\boldsymbol{\theta}, \boldsymbol{\tau})\right] \right\|_2 \\ &\leq 2 \sup_{\boldsymbol{\lambda} \in \boldsymbol{\lambda}_{1/2}} \left\langle \boldsymbol{\lambda}, \frac{1}{n} \sum_{i=1}^{n} \nabla_{\boldsymbol{\theta}} \hat{V}_H(\boldsymbol{\theta}, \boldsymbol{\tau}_i) - \mathbb{E}\left[\nabla_{\boldsymbol{\theta}} \hat{V}_H(\boldsymbol{\theta}, \boldsymbol{\tau})\right] \right\rangle \end{aligned}$$

$$\leq 2 \sup_{\boldsymbol{\lambda} \in \boldsymbol{\lambda}_{1/2}} \left\{ \left\langle \boldsymbol{\lambda}, \frac{1}{n} \sum_{i=1}^{n} \nabla_{\boldsymbol{\theta}} \hat{V}_H(\boldsymbol{\theta}, \boldsymbol{\tau}_i) \right\rangle - \mathbb{E}\left[\left\langle \boldsymbol{\lambda}, \nabla_{\boldsymbol{\theta}} \hat{V}_H(\boldsymbol{\theta}, \boldsymbol{\tau}) \right\rangle \right] \right\}.$$

Now we have

$$\begin{aligned}
&\mathbb{P}\left(\left\| \frac{1}{n} \sum_{i=1}^{n} \nabla_{\boldsymbol{\theta}} \hat{V}_H(\boldsymbol{\theta}, \boldsymbol{\tau}_i) - \mathbb{E}\left[\nabla_{\boldsymbol{\theta}} V(\boldsymbol{\theta}, \boldsymbol{\tau}) \right] \right\|_2 \geq \frac{t}{2} \right) \\
&\leq \mathbb{P}\left(\sup_{\boldsymbol{\lambda} \in \boldsymbol{\lambda}_{1/2}} \left\{ \left\langle \boldsymbol{\lambda}, \frac{1}{n} \sum_{i=1}^{n} \nabla_{\boldsymbol{\theta}} \hat{V}_H(\boldsymbol{\theta}, \boldsymbol{\tau}_i) \right\rangle - \mathbb{E}\left[\left\langle \boldsymbol{\lambda}, \nabla_{\boldsymbol{\theta}} \hat{V}_H(\boldsymbol{\theta}, \boldsymbol{\tau}) \right\rangle \right] \right\} \geq \frac{t}{4} \right) \\
&\leq 5^{|A|} \sup_{\boldsymbol{\lambda} \in \boldsymbol{\lambda}_{1/2}} \mathbb{P}\left(\left\langle \boldsymbol{\lambda}, \frac{1}{n} \sum_{i=1}^{n} \nabla_{\boldsymbol{\theta}} \hat{V}_H(\boldsymbol{\theta}, \boldsymbol{\tau}_i) \right\rangle - \mathbb{E}\left[\left\langle \boldsymbol{\lambda}, \nabla_{\boldsymbol{\theta}} \hat{V}_H(\boldsymbol{\theta}, \boldsymbol{\tau}) \right\rangle \right] \geq \frac{t}{4} \right)
\end{aligned}$$

It can be verified as in the proof of Lemma 6 that the value of $\left\langle \boldsymbol{\lambda}, \frac{1}{n} \sum_{i=1}^{n} \nabla_{\boldsymbol{\theta}} \hat{V}_H(\boldsymbol{\theta}, \boldsymbol{\tau}_i) \right\rangle$ lie in the range $\left[-\frac{HG_g}{1-\gamma}, \frac{HG_g}{1-\gamma} \right]$ for any $\boldsymbol{\lambda}$. By Hoeffding's Inequality [50],

$$\begin{aligned}
&\mathbb{P}\left(\left\langle \boldsymbol{\lambda}, \frac{1}{n} \sum_{i=1}^{n} \nabla_{\boldsymbol{\theta}} \hat{V}_H(\boldsymbol{\theta}, \boldsymbol{\tau}_i) \right\rangle - \mathbb{E}\left[\left\langle \boldsymbol{\lambda}, \nabla_{\boldsymbol{\theta}} \hat{V}_H(\boldsymbol{\theta}, \boldsymbol{\tau}) \right\rangle \right] \geq \frac{t}{4} \right) \\
&\leq 2 \exp\left(-\frac{1}{32 H G_g} n t^2 (1-\gamma)^2 \right).
\end{aligned}$$

Therefore,

$$\begin{aligned}
&\mathbb{P}\left(\left\| \frac{1}{n} \sum_{i=1}^{n} \nabla_{\boldsymbol{\theta}} \hat{V}_H(\boldsymbol{\theta}, \boldsymbol{\tau}_i) - \mathbb{E}\left[\nabla_{\boldsymbol{\theta}} V(\boldsymbol{\theta}, \boldsymbol{\tau}) \right] \right\|_2 \geq \frac{t}{2} \right) \\
&\leq 2 \exp\left(-\frac{1}{32 H G_g} n t^2 (1-\gamma)^2 + |A| \log 5 \right).
\end{aligned}$$

□

Lemma 10. *For any $t > 0$, the gradient $\nabla \hat{V}_H(\boldsymbol{\theta}, \boldsymbol{\tau})$ with $H \geq \frac{\log(2G_h \sqrt{H + (1/1-\gamma)}/t(1-\gamma))}{\log(1/\gamma)}$ obeys*

$$\begin{aligned}
&\mathbb{P}\left(\left\| \frac{1}{n} \sum_{i=1}^{n} \nabla^2_{\boldsymbol{\theta}} \hat{V}_H(\boldsymbol{\theta}, \boldsymbol{\tau}_i) - \mathbb{E}\left[\nabla^2_{\boldsymbol{\theta}} V(\boldsymbol{\theta}, \boldsymbol{\tau}) \right] \right\|_F \geq t \right) \\
&\leq 2 \exp\left(-\frac{1}{64 H G_h} n t^2 (1-\gamma)^2 + |A| \log 9 \right).
\end{aligned}$$

Proof. Let $H \geq \frac{\log(2G_h \sqrt{H + (1/1-\gamma)}/t(1-\gamma))}{\log(1/\gamma)}$, we have

$$\left\| \frac{1}{n} \sum_{i=1}^{n} \nabla^2_{\boldsymbol{\theta}} \hat{V}_H(\boldsymbol{\theta}, \boldsymbol{\tau}_i) - \mathbb{E}\left[\nabla^2_{\boldsymbol{\theta}} V(\boldsymbol{\theta}, \boldsymbol{\tau}) \right] \right\|_2$$

$$
\begin{aligned}
&= \left\| \frac{1}{n}\sum_{i=1}^{n} \nabla^2_{\boldsymbol{\theta}} \hat{V}_H(\boldsymbol{\theta}, \boldsymbol{\tau}_i) - \mathbb{E}\left[\nabla^2_{\boldsymbol{\theta}} \hat{V}_H(\boldsymbol{\theta}, \boldsymbol{\tau})\right] + \mathbb{E}\left[\nabla^2_{\boldsymbol{\theta}} \hat{V}_H(\boldsymbol{\theta}, \boldsymbol{\tau})\right] - \mathbb{E}\left[\nabla^2_{\boldsymbol{\theta}} V(\boldsymbol{\theta}, \boldsymbol{\tau})\right] \right\|_2 \\
&\leq \left\| \frac{1}{n}\sum_{i=1}^{n} \nabla^2_{\boldsymbol{\theta}} \hat{V}_H(\boldsymbol{\theta}, \boldsymbol{\tau}_i) - \mathbb{E}\left[\nabla^2_{\boldsymbol{\theta}} \hat{V}_H(\boldsymbol{\theta}, \boldsymbol{\tau})\right] \right\|_2 + \left\| \mathbb{E}\left[\nabla^2_{\boldsymbol{\theta}} \hat{V}_H(\boldsymbol{\theta}, \boldsymbol{\tau})\right] - \mathbb{E}\left[\nabla^2_{\boldsymbol{\theta}} V(\boldsymbol{\theta}, \boldsymbol{\tau})\right] \right\|_2 \\
&\leq \left\| \frac{1}{n}\sum_{i=1}^{n} \nabla^2_{\boldsymbol{\theta}} \hat{V}_H(\boldsymbol{\theta}, \boldsymbol{\tau}_i) - \mathbb{E}\left[\nabla^2_{\boldsymbol{\theta}} \hat{V}_H(\boldsymbol{\theta}, \boldsymbol{\tau})\right] \right\|_2 + \frac{\gamma^H G_h}{1-\gamma}\sqrt{\frac{1}{1-\gamma} + H} \\
&\leq \left\| \frac{1}{n}\sum_{i=1}^{n} \nabla^2_{\boldsymbol{\theta}} \hat{V}_H(\boldsymbol{\theta}, \boldsymbol{\tau}_i) - \mathbb{E}\left[\nabla^2_{\boldsymbol{\theta}} \hat{V}_H(\boldsymbol{\theta}, \boldsymbol{\tau})\right] \right\|_2 + \frac{t}{2}
\end{aligned}
$$

where the second inequality holds since we utilize Lemma 7. Let $\boldsymbol{\lambda}_{1/4}$ be the $\frac{1}{4}$-covering net of $B^{|A|}(1)$ with $\left|\boldsymbol{\lambda}_{1/4}\right| \leq 9^{|A|}$. From Lemma 5 we know that

$$
\begin{aligned}
&\left\| \frac{1}{n}\sum_{i=1}^{n} \nabla^2_{\boldsymbol{\theta}} \hat{V}_H(\boldsymbol{\theta}, \boldsymbol{\tau}_i) - \mathbb{E}\left[\nabla^2_{\boldsymbol{\theta}} \hat{V}_H(\boldsymbol{\theta}, \boldsymbol{\tau})\right] \right\|_2 \\
&\leq 2 \sup_{\boldsymbol{\lambda} \in \boldsymbol{\lambda}_{1/4}} \left| \left\langle \boldsymbol{\lambda}, \frac{1}{n}\sum_{i=1}^{n} \nabla^2_{\boldsymbol{\theta}} \hat{V}_H(\boldsymbol{\theta}, \boldsymbol{\tau}_i) - \mathbb{E}\left[\nabla^2_{\boldsymbol{\theta}} \hat{V}_H(\boldsymbol{\theta}, \boldsymbol{\tau})\right] \boldsymbol{\lambda} \right\rangle \right| \\
&\leq 2 \sup_{\boldsymbol{\lambda} \in \boldsymbol{\lambda}_{1/4}} \left| \left\langle \boldsymbol{\lambda}, \frac{1}{n}\sum_{i=1}^{n} \nabla^2_{\boldsymbol{\theta}} \hat{V}_H(\boldsymbol{\theta}, \boldsymbol{\tau}_i)\boldsymbol{\lambda} \right\rangle - \mathbb{E}\left[\left\langle \boldsymbol{\lambda}, \nabla^2_{\boldsymbol{\theta}} \hat{V}_H(\boldsymbol{\theta}, \boldsymbol{\tau})\boldsymbol{\lambda} \right\rangle\right] \right|.
\end{aligned}
$$

Now we have

$$
\begin{aligned}
&\mathbb{P}\left(\left\| \frac{1}{n}\sum_{i=1}^{n} \nabla^2_{\boldsymbol{\theta}} \hat{V}_H(\boldsymbol{\theta}, \boldsymbol{\tau}_i) - \mathbb{E}\left[\nabla^2_{\boldsymbol{\theta}} v(\boldsymbol{\theta}, \boldsymbol{\tau})\right] \right\|_2 \geq \frac{t}{2} \right) \\
&\leq \mathbb{P}\left(\sup_{\boldsymbol{\lambda} \in \boldsymbol{\lambda}_{1/4}} \left| \left\langle \boldsymbol{\lambda}, \frac{1}{n}\sum_{i=1}^{n} \nabla^2_{\boldsymbol{\theta}} \hat{V}_H(\boldsymbol{\theta}, \boldsymbol{\tau}_i)\boldsymbol{\lambda} \right\rangle - \mathbb{E}\left[\left\langle \boldsymbol{\lambda}, \nabla^2_{\boldsymbol{\theta}} \hat{V}_H(\boldsymbol{\theta}, \boldsymbol{\tau})\boldsymbol{\lambda} \right\rangle\right] \right| \geq \frac{t}{4} \right) \\
&\leq 9^{|A|} \sup_{\boldsymbol{\lambda} \in \boldsymbol{\lambda}_{1/4}} \mathbb{P}\left(\left| \left\langle \boldsymbol{\lambda}, \frac{1}{n}\sum_{i=1}^{n} \nabla^2_{\boldsymbol{\theta}} \hat{V}_H(\boldsymbol{\theta}, \boldsymbol{\tau}_i)\boldsymbol{\lambda} \right\rangle - \mathbb{E}\left[\left\langle \boldsymbol{\lambda}, \nabla^2_{\boldsymbol{\theta}} \hat{V}_H(\boldsymbol{\theta}, \boldsymbol{\tau})\boldsymbol{\lambda} \right\rangle\right] \right| \geq \frac{t}{4} \right)
\end{aligned}
$$

By Lemma 6, we know that $\left\langle \boldsymbol{\lambda}, \frac{1}{n}\sum_{i=1}^{n} \nabla^2_{\boldsymbol{\theta}} \hat{V}_H(\boldsymbol{\theta}, \boldsymbol{\tau}_i)\boldsymbol{\lambda} \right\rangle$ is upper bounded by $\frac{HG_h}{1-\gamma}$ for any $\boldsymbol{\lambda}$. By Hoeffding's Inequality [50],

$$
\begin{aligned}
&\mathbb{P}\left(\left| \left\langle \boldsymbol{\lambda}, \frac{1}{n}\sum_{i=1}^{n} \nabla^2_{\boldsymbol{\theta}} \hat{V}_H(\boldsymbol{\theta}, \boldsymbol{\tau}_i)\boldsymbol{\lambda} \right\rangle - \mathbb{E}\left[\left\langle \boldsymbol{\lambda}, \nabla^2_{\boldsymbol{\theta}} \hat{V}_H(\boldsymbol{\theta}, \boldsymbol{\tau})\boldsymbol{\lambda} \right\rangle\right] \right| \geq \frac{t}{4} \right) \\
&\leq 2\exp\left(-\frac{1}{64HG_h} n t^2 (1-\gamma)^2 \right).
\end{aligned}
$$

Therefore,

$$
\mathbb{P}\left(\left\| \frac{1}{n}\sum_{i=1}^{n} \nabla^2_{\boldsymbol{\theta}} \hat{V}_H(\boldsymbol{\theta}, \boldsymbol{\tau}_i) - \mathbb{E}\left[\nabla^2_{\boldsymbol{\theta}} V(\boldsymbol{\theta}, \boldsymbol{\tau})\right] \right\|_2 \geq \frac{t}{2} \right)
$$

$$\leq 2\exp\left(-\frac{1}{64HG_h}nt^2\left(1-\gamma\right)^2+|A|\log 9\right).$$

□

B Proof of Theorem 1

Theorem 5 (Theorem 1 restated). *Suppose Assumption 1 holds. Then the truncated estimator defined in (2) uniformly converges to the expected infinite-horizon discounted value function. Specifically, if* $n \geq \max\left\{\frac{18(\log(4/\mathcal{E})+|\mathcal{S}||\mathcal{A}|\log(n/\mathcal{E}))}{(1-\gamma)^2}, \frac{18HG_g r}{(1-\gamma)}\right\}$*, then we have*

$$\sup_{\boldsymbol{\theta}\in\Theta}\left|\hat{\boldsymbol{V}}_n(\boldsymbol{\theta})-\boldsymbol{V}(\boldsymbol{\theta})\right| \leq \sqrt{\frac{36|\mathcal{S}||\mathcal{A}|\log(n/\mathcal{E}))}{n(1-\gamma)^2}}+\frac{18HG_g r}{n(1-\gamma)}$$

with probability at least $1-\mathcal{E}$*, where* $\Theta = B^{|\mathcal{S}||\mathcal{A}|}(r)$*.*

Proof. According to ϵ-covering theory in [49], it is known that the ϵ-covering number N_ϵ of the ball $B^{|\mathcal{S}||\mathcal{A}|}(r)$ is upper bounded by $\left(\frac{3r}{\epsilon}\right)^{|\mathcal{S}||\mathcal{A}|}$. We assume $\boldsymbol{\theta}_\epsilon = \{\boldsymbol{\theta}_1,\ldots,\boldsymbol{\theta}_{N_\epsilon}\}$ is the ϵ-covering net of $B^{|\mathcal{S}||\mathcal{A}|}(r)$. Let $\boldsymbol{\theta} \in B^{|\mathcal{S}||\mathcal{A}|}(r)$ be an arbitrary vector, we have $\|\boldsymbol{\theta}-\boldsymbol{\theta}_i\| \leq \epsilon$ for some $\boldsymbol{\theta}_i \in \{\boldsymbol{\theta}_1,\ldots,\boldsymbol{\theta}_{N_\epsilon}\}$. We denote $V(\boldsymbol{\theta},\boldsymbol{\tau}) = \lim_{H\to\infty}\hat{V}_H(\boldsymbol{\theta},\boldsymbol{\tau})$. Note that $\hat{\boldsymbol{V}}_n(\boldsymbol{\theta}) = \frac{1}{n}\sum_{i=1}^n \hat{V}_H(\boldsymbol{\theta},\boldsymbol{\tau}_i)$ and $\boldsymbol{V}(\boldsymbol{\theta}) = \mathbb{E}_{\boldsymbol{\tau}}[V(\boldsymbol{\theta},\boldsymbol{\tau})]$. Having clarified the aforementioned context, the subscripts will henceforth be omitted for simplicity. Now for any $\boldsymbol{\theta} \in B^{|\mathcal{S}||\mathcal{A}|}(r)$ and by the decomposition strategy we have

$$\begin{aligned}
&\left|\hat{\boldsymbol{V}}_n(\boldsymbol{\theta})-\boldsymbol{V}(\boldsymbol{\theta})\right| = \left|\frac{1}{n}\sum_{i=1}^n \hat{V}(\boldsymbol{\theta},\boldsymbol{\tau}_i)-\mathbb{E}[V(\boldsymbol{\theta},\boldsymbol{\tau})]\right| \\
&= \Bigg|\frac{1}{n}\sum_{i=1}^n \hat{V}(\boldsymbol{\theta},\boldsymbol{\tau}_i)-\frac{1}{n}\sum_{i=1}^n \hat{V}(\boldsymbol{\theta}_{i(\boldsymbol{\theta})},\boldsymbol{\tau}_i)+\frac{1}{n}\sum_{i=1}^n \hat{V}(\boldsymbol{\theta}_{i(\boldsymbol{\theta})},\boldsymbol{\tau}_i)-\mathbb{E}[V(\boldsymbol{\theta}_{i(\boldsymbol{\theta})},\boldsymbol{\tau})] \\
&\quad +\mathbb{E}[V(\boldsymbol{\theta}_{i(\boldsymbol{\theta})},\boldsymbol{\tau})]-\mathbb{E}[V(\boldsymbol{\theta},\boldsymbol{\tau})]\Bigg| \\
&\leq \left|\frac{1}{n}\sum_{i=1}^n(\hat{V}(\boldsymbol{\theta},\boldsymbol{\tau}_i)-\hat{V}(\boldsymbol{\theta}_{i(\boldsymbol{\theta})},\boldsymbol{\tau}_i))\right|+\left|\frac{1}{n}\sum_{i=1}^n \hat{V}(\boldsymbol{\theta}_{i(\boldsymbol{\theta})},\boldsymbol{\tau}_i)-\mathbb{E}[V(\boldsymbol{\theta}_{i(\boldsymbol{\theta})},\boldsymbol{\tau})]\right| \\
&\quad +\left|\mathbb{E}[V(\boldsymbol{\theta}_{i(\boldsymbol{\theta})},\boldsymbol{\tau})]-\mathbb{E}[V(\boldsymbol{\theta},\boldsymbol{\tau})]\right|
\end{aligned}$$

where $i(\boldsymbol{\theta}) = \arg\min_{i\in[N_\epsilon]}\|\boldsymbol{\theta}-\boldsymbol{\theta}_i\|_2$. Then, we define three events $\boldsymbol{E}_1, \boldsymbol{E}_2, \boldsymbol{E}_3$ as

$$\begin{aligned}
\boldsymbol{E}_1 &= \left\{\sup_{\boldsymbol{\theta}\in B^{|\mathcal{S}||\mathcal{A}|}(r)}\left|\frac{1}{n}\sum_{i=1}^n(\hat{V}(\boldsymbol{\theta},\boldsymbol{\tau}_i)-\hat{V}(\boldsymbol{\theta}_{i(\boldsymbol{\theta})},\boldsymbol{\tau}_i))\right| \geq \frac{t}{3}\right\}, \\
\boldsymbol{E}_2 &= \left\{\sup_{i\in[N_\epsilon]}\left|\frac{1}{n}\sum_{i=1}^n \hat{V}(\boldsymbol{\theta}_{i(\boldsymbol{\theta})},\boldsymbol{\tau}_i)-\mathbb{E}[V(\boldsymbol{\theta}_{i(\boldsymbol{\theta})},\boldsymbol{\tau})]\right| \geq \frac{t}{3}\right\},
\end{aligned}$$

$$\boldsymbol{E}_3 = \left\{ \sup_{\boldsymbol{\theta}\in B^{|\mathcal{S}||\mathcal{A}|}(r)} \left|\mathbb{E}[V(\boldsymbol{\theta}_{i(\boldsymbol{\theta})}, \boldsymbol{\tau})] - \mathbb{E}[V(\boldsymbol{\theta}, \boldsymbol{\tau})]\right| \geq \frac{t}{3} \right\}.$$

Therefore, we have

$$\mathbb{P}\left(\sup_{\boldsymbol{\theta}\in B^{|\mathcal{S}||\mathcal{A}|}(r)} \left|\hat{\boldsymbol{V}}_n(\boldsymbol{\theta}) - \boldsymbol{V}(\boldsymbol{\theta})\right| \geq t \right) \leq \mathbb{P}(\boldsymbol{E}_1) + \mathbb{P}(\boldsymbol{E}_2) + \mathbb{P}(\boldsymbol{E}_3).$$

Next, we first upper bound $\mathbb{P}(\boldsymbol{E}_1)$.

$$\begin{aligned}
\mathbb{P}(\boldsymbol{E}_1) &= \mathbb{P}\left(\sup_{\boldsymbol{\theta}\in B^{|\mathcal{S}||\mathcal{A}|}(r)} \left| \frac{1}{n}\sum_{i=1}^n (\hat{V}(\boldsymbol{\theta}, \boldsymbol{\tau}_i) - \hat{V}(\boldsymbol{\theta}_{i(\boldsymbol{\theta})}, \boldsymbol{\tau}_i)) \right| \geq \frac{t}{3} \right) \\
&\leq \frac{3}{t}\mathbb{E}\left[\sup_{\boldsymbol{\theta}\in B^{|\mathcal{S}||\mathcal{A}|}(r)} \left| \frac{1}{n}\sum_{i=1}^n (\hat{V}(\boldsymbol{\theta}, \boldsymbol{\tau}_i) - \hat{V}(\boldsymbol{\theta}_{i(\boldsymbol{\theta})}, \boldsymbol{\tau}_i)) \right| \right] \\
&\leq \frac{3}{t}\mathbb{E}\left[\sup_{\boldsymbol{\theta}\in B^{|\mathcal{S}||\mathcal{A}|}(r)} \frac{\left| \frac{1}{n}\sum_{i=1}^n (\hat{V}(\boldsymbol{\theta}, \boldsymbol{\tau}_i) - \hat{V}(\boldsymbol{\theta}_{i(\boldsymbol{\theta})}, \boldsymbol{\tau}_i)) \right|}{\|\boldsymbol{\theta} - \boldsymbol{\theta}_{i(\boldsymbol{\theta})}\|_2} \sup_{\boldsymbol{\theta}\in B^{|\mathcal{S}||\mathcal{A}|}(r)} \|\boldsymbol{\theta} - \boldsymbol{\theta}_{i(\boldsymbol{\theta})}\|_2 \right] \\
&\leq \frac{3\epsilon}{t}\mathbb{E}\left[\sup_{\boldsymbol{\theta}\in B^{|\mathcal{S}||\mathcal{A}|}(r)} \|\nabla \hat{\boldsymbol{V}}_n(\boldsymbol{\theta})\|_2 \right],
\end{aligned}$$

where the first inequality holds by Markov inequality. By Lemma 6, we have

$$\begin{aligned}
\mathbb{E}\left[\sup_{\boldsymbol{\theta}\in B^{|\mathcal{S}||\mathcal{A}|}(r)} \|\nabla \hat{\boldsymbol{V}}_n(\boldsymbol{\theta})\|_2 \right] &= \mathbb{E}\left[\sup_{\boldsymbol{\theta}\in B^{|\mathcal{S}||\mathcal{A}|}(r)} \left\| \frac{1}{n}\sum_{i=1}^n \nabla \hat{V}(\boldsymbol{\theta}, \boldsymbol{\tau}_i) \right\|_2 \right] \\
&= \mathbb{E}\left[\sup_{\boldsymbol{\theta}\in B^{|\mathcal{S}||\mathcal{A}|}(r)} \left\| \nabla \hat{V}(\boldsymbol{\theta}, \boldsymbol{\tau}) \right\|_2 \right] \\
&\leq \frac{HG_g}{1-\gamma}.
\end{aligned}$$

Thus,

$$\mathbb{P}(\boldsymbol{E}_1) \leq \frac{3HG_g\epsilon}{t(1-\gamma)}$$

and by letting $t \geq \frac{6HG_g\epsilon}{(1-\gamma)\mathcal{E}}$ we have

$$\mathbb{P}(\boldsymbol{E}_1) \leq \frac{\mathcal{E}}{2}.$$

Secondly, we bound $\mathbb{P}(\boldsymbol{E}_2)$ as follows:

$$\mathbb{P}(\boldsymbol{E}_2) = \mathbb{P}\left(\sup_{i\in[N_\epsilon]} \left| \frac{1}{n}\sum_{i=1}^n \hat{V}(\boldsymbol{\theta}_{i(\boldsymbol{\theta})}, \boldsymbol{\tau}_i) - \mathbb{E}[V(\boldsymbol{\theta}_{i(\boldsymbol{\theta})}, \boldsymbol{\tau})] \right| \geq \frac{t}{3} \right)$$

$$\leq \left(\frac{3r}{\epsilon}\right)^{|\mathcal{S}||\mathcal{A}|} \sup_{i\in[N_\epsilon]} \mathbb{P}\left(\left|\frac{1}{n}\sum_{i=1}^{n}\hat{V}(\boldsymbol{\theta}_{i(\boldsymbol{\theta})},\boldsymbol{\tau}_i) - \mathbb{E}[V(\boldsymbol{\theta}_{i(\boldsymbol{\theta})},\boldsymbol{\tau})]\right| \geq \frac{t}{3}\right)$$
$$\leq \left(\frac{3r}{\epsilon}\right)^{|\mathcal{S}||\mathcal{A}|} 2\exp\left(-\frac{1}{18}nt^2(1-\gamma)^2\right)$$

where the last inequality is due to Lemma 8. Therefore, if we let

$$t \geq \sqrt{\frac{18(\log(4/\mathcal{E}) + |\mathcal{S}||\mathcal{A}|\log(3r/\epsilon))}{n(1-\gamma)^2}},$$

then we have

$$\mathbb{P}(\boldsymbol{E}_2) \leq \frac{\mathcal{E}}{2}.$$

Thirdly, we upper bound $\mathbb{P}(\boldsymbol{E}_3)$. Similarly,

$$\begin{aligned}
\mathbb{P}(\boldsymbol{E}_3) &= \mathbb{P}\left(\sup_{\boldsymbol{\theta}\in\mathbb{S}^{|A|-1}} |\mathbb{E}[V(\boldsymbol{\theta}_{i(\boldsymbol{\theta})},\boldsymbol{\tau})] - \mathbb{E}[V(\boldsymbol{\theta},\boldsymbol{\tau})]| \geq \frac{t}{3}\right) \\
&\leq \mathbb{P}\left(\sup_{\boldsymbol{\theta}\in\mathbb{S}^{|A|-1}} \frac{\left|\mathbb{E}[V(\boldsymbol{\theta}_{i(\boldsymbol{\theta})},\boldsymbol{\tau})] - \mathbb{E}[V(\boldsymbol{\theta},\boldsymbol{\tau})]\right|}{\|\boldsymbol{\theta}-\boldsymbol{\theta}_{i(\boldsymbol{\theta})}\|_2} \sup_{\boldsymbol{\theta}\in\mathbb{S}^{|A|-1}} \|\boldsymbol{\theta}-\boldsymbol{\theta}_{i(\boldsymbol{\theta})}\|_2 \geq \frac{t}{3}\right) \\
&\leq \mathbb{P}\left(\epsilon\mathbb{E}\left[\sup_{\boldsymbol{\theta}\in\mathbb{S}^{|A|-1}} \|\nabla \boldsymbol{V}(\boldsymbol{\theta},\boldsymbol{\tau})\|_2\right] \geq \frac{t}{3}\right) \\
&\leq \mathbb{P}\left(\frac{\epsilon H G_g}{1-\gamma} \geq \frac{t}{3}\right).
\end{aligned}$$

By setting $\epsilon \leq \frac{t(1-\gamma)}{3HG_g}$, we ensure that $\mathbb{P}(\boldsymbol{E}_3) = 0$. Lastly, we let $\epsilon = \frac{3r\mathcal{E}}{n}$ and

$$\begin{aligned}
t &\geq \max\left\{\frac{6HG_g\epsilon}{(1-\gamma)\mathcal{E}}, \sqrt{\frac{18(\log(4/\mathcal{E}) + |\mathcal{S}||\mathcal{A}|\log(3r/\epsilon))}{n(1-\gamma)^2}}\right\} \\
&= \max\left\{\frac{18HG_gr}{n(1-\gamma)}, \sqrt{\frac{18(\log(4/\mathcal{E}) + |\mathcal{S}||\mathcal{A}|\log(n/\mathcal{E}))}{n(1-\gamma)^2}}\right\}
\end{aligned}$$

to ensure $\mathbb{P}\left(\sup_{\boldsymbol{\theta}\in\mathbb{S}^{|A|-1}} \left|\hat{\boldsymbol{V}}_n(\boldsymbol{\theta}) - \boldsymbol{V}(\boldsymbol{\theta})\right| \geq t\right) \leq \mathcal{E}$. Therefore, if $n \geq \max\left\{\frac{18(\log(4/\mathcal{E})+|\mathcal{S}||\mathcal{A}|\log(n/\mathcal{E}))}{(1-\gamma)^2}, \frac{18HG_gr}{(1-\gamma)}\right\}$, then with probability at least $1-\mathcal{E}$

$$\sup_{\boldsymbol{\theta}\in\Theta} \left|\hat{\boldsymbol{V}}_n(\boldsymbol{\theta}) - \boldsymbol{V}(\boldsymbol{\theta})\right| \leq \sqrt{\frac{36|\mathcal{S}||\mathcal{A}|\log(n/\mathcal{E}))}{n(1-\gamma)^2}} + \frac{18HG_gr}{n(1-\gamma)}.$$

□

C Proof of Theorem 2

Theorem 6 (Theorem 2 restated). *Suppose assumption 1 and 2 hold. Then the truncated gradient estimator defined in (2) uniformly converges to the gradient of the expected infinite-horizon discounted value function in Euclidean norm. Specifically, if* $n \geq \frac{64H}{(1-\gamma)} \max\left\{\frac{G_g|\mathcal{S}||\mathcal{A}|\log(n/\mathcal{E})}{(1-\gamma)}, G_h r\right\}$*, then we have*

$$\sup_{\boldsymbol{\theta}\in\Theta}\left\|\nabla\hat{\boldsymbol{V}}_n(\boldsymbol{\theta})-\nabla\boldsymbol{V}(\boldsymbol{\theta})\right\|_2 \leq \sqrt{\frac{64HG_g|\mathcal{S}||\mathcal{A}|\log(n/\mathcal{E})}{n(1-\gamma)^2}}+\frac{18HG_hr}{n(1-\gamma)}.$$

with probability at least $1-\mathcal{E}$*, where* $\Theta = B^{|\mathcal{S}||\mathcal{A}|}(r)$*.*

Proof. We adopt a similar proof strategy as in the proof of Theorem 1. Note that the ϵ-covering number N_ϵ of the ball $B^{|\mathcal{S}||\mathcal{A}|}(r)$ is upper bounded by $\left(\frac{3r}{\epsilon}\right)^{|\mathcal{S}||\mathcal{A}|}$. We assume $\boldsymbol{\theta}_\epsilon = \{\boldsymbol{\theta}_1,\ldots,\boldsymbol{\theta}_{N_\epsilon}\}$ is the ϵ-covering net of $B^{|\mathcal{S}||\mathcal{A}|}(r)$. Let $\boldsymbol{\theta}\in B^{|\mathcal{S}||\mathcal{A}|}(r)$ be an arbitrary vector, we have $\|\boldsymbol{\theta}-\boldsymbol{\theta}_i\|\leq\epsilon$ for some $\boldsymbol{\theta}_i\in\{\boldsymbol{\theta}_1,\ldots,\boldsymbol{\theta}_{N_\epsilon}\}$. We denote $V(\boldsymbol{\theta},\boldsymbol{\tau}) = \lim_{H\to\infty}\hat{V}_H(\boldsymbol{\theta},\boldsymbol{\tau})$. For any $\boldsymbol{\theta}\in B^{|\mathcal{S}||\mathcal{A}|}(r)$, by the decomposition strategy we have

$$\begin{aligned}
&\left\|\nabla\hat{\boldsymbol{V}}_n(\boldsymbol{\theta})-\nabla\boldsymbol{V}(\boldsymbol{\theta})\right\|_2 = \left\|\frac{1}{n}\sum_{i=1}^n\nabla\hat{V}(\boldsymbol{\theta},\boldsymbol{\tau}_i)-\mathbb{E}[\nabla V(\boldsymbol{\theta},\boldsymbol{\tau})]\right\|_2\\
&=\left\|\frac{1}{n}\sum_{i=1}^n\nabla\hat{V}(\boldsymbol{\theta},\boldsymbol{\tau}_i)-\frac{1}{n}\sum_{i=1}^n\nabla\hat{V}(\boldsymbol{\theta}_{i(\boldsymbol{\theta})},\boldsymbol{\tau}_i)+\frac{1}{n}\sum_{i=1}^n\nabla\hat{V}(\boldsymbol{\theta}_{i(\boldsymbol{\theta})},\boldsymbol{\tau}_i)-\mathbb{E}[\nabla V(\boldsymbol{\theta}_{i(\boldsymbol{\theta})},\boldsymbol{\tau})]\right.\\
&\quad\left.+\mathbb{E}[\nabla V(\boldsymbol{\theta}_{i(\boldsymbol{\theta})},\boldsymbol{\tau})]-\mathbb{E}[\nabla V(\boldsymbol{\theta},\boldsymbol{\tau})]\right\|_2\\
&\leq\left\|\frac{1}{n}\sum_{i=1}^n(\nabla\hat{V}(\boldsymbol{\theta},\boldsymbol{\tau}_i)-\nabla\hat{V}(\boldsymbol{\theta}_{i(\boldsymbol{\theta})},\boldsymbol{\tau}_i))\right\|_2+\left\|\frac{1}{n}\sum_{i=1}^n\nabla\hat{V}(\boldsymbol{\theta}_{i(\boldsymbol{\theta})},\boldsymbol{\tau}_i)-\mathbb{E}[\nabla V(\boldsymbol{\theta}_{i(\boldsymbol{\theta})},\boldsymbol{\tau})]\right\|_2\\
&\quad+\left\|\mathbb{E}[\nabla V(\boldsymbol{\theta}_{i(\boldsymbol{\theta})},\boldsymbol{\tau})]-\mathbb{E}[\nabla V(\boldsymbol{\theta},\boldsymbol{\tau})]\right\|_2
\end{aligned}$$

where $i(\boldsymbol{\theta}) = \arg\min_{i\in[N_\epsilon]}\|\boldsymbol{\theta}-\boldsymbol{\theta}_i\|_2$. Then, we define three events $\boldsymbol{E}_1,\boldsymbol{E}_2,\boldsymbol{E}_3$ as

$$\begin{aligned}
\boldsymbol{E}_1&=\left\{\sup_{\boldsymbol{\theta}\in B^{|\mathcal{S}||\mathcal{A}|}(r)}\left\|\frac{1}{n}\sum_{i=1}^n(\nabla\hat{V}(\boldsymbol{\theta},\boldsymbol{\tau}_i)-\nabla\hat{V}(\boldsymbol{\theta}_{i(\boldsymbol{\theta})},\boldsymbol{\tau}_i))\right\|_2\geq\frac{t}{3}\right\},\\
\boldsymbol{E}_2&=\left\{\sup_{i\in[N_\epsilon]}\left\|\frac{1}{n}\sum_{i=1}^n\nabla\hat{V}(\boldsymbol{\theta}_{i(\boldsymbol{\theta})},\boldsymbol{\tau}_i)-\mathbb{E}[\nabla V(\boldsymbol{\theta}_{i(\boldsymbol{\theta})},\boldsymbol{\tau})]\right\|_2\geq\frac{t}{3}\right\},\\
\boldsymbol{E}_3&=\left\{\sup_{\boldsymbol{\theta}\in B^{|\mathcal{S}||\mathcal{A}|}(r)}\left\|\mathbb{E}[\nabla V(\boldsymbol{\theta}_{i(\boldsymbol{\theta})},\boldsymbol{\tau})]-\mathbb{E}[\nabla V(\boldsymbol{\theta},\boldsymbol{\tau})]\right\|_2\geq\frac{t}{3}\right\}.
\end{aligned}$$

Accordingly, we have

$$\mathbb{P}\left(\sup_{\boldsymbol{\theta}\in B^{|\mathcal{S}||\mathcal{A}|}(r)}\left\|\nabla\hat{\boldsymbol{V}}_n(\boldsymbol{\theta})-\nabla\boldsymbol{V}(\boldsymbol{\theta})\right\|_2\geq t\right)\leq\mathbb{P}(\boldsymbol{E}_1)+\mathbb{P}(\boldsymbol{E}_2)+\mathbb{P}(\boldsymbol{E}_3).$$

Next, we first upper bound $\mathbb{P}(\boldsymbol{E}_1)$.

$$\begin{aligned}
\mathbb{P}(\boldsymbol{E}_1) &= \mathbb{P}\left(\sup_{\boldsymbol{\theta}\in B^{|\mathcal{S}||\mathcal{A}|}(r)} \left\|\frac{1}{n}\sum_{i=1}^{n}(\nabla\hat{V}(\boldsymbol{\theta},\boldsymbol{\tau}_i)-\nabla\hat{V}(\boldsymbol{\theta}_{i(\boldsymbol{\theta})},\boldsymbol{\tau}_i))\right\|_2 \geq \frac{t}{3}\right)\\
&\leq \frac{3}{t}\mathbb{E}\left[\sup_{\boldsymbol{\theta}\in B^{|\mathcal{S}||\mathcal{A}|}(r)} \left\|\frac{1}{n}\sum_{i=1}^{n}(\nabla\hat{V}(\boldsymbol{\theta},\boldsymbol{\tau}_i)-\nabla\hat{V}(\boldsymbol{\theta}_{i(\boldsymbol{\theta})},\boldsymbol{\tau}_i))\right\|_2\right]\\
&\leq \frac{3}{t}\mathbb{E}\left[\sup_{\boldsymbol{\theta}\in B^{|\mathcal{S}||\mathcal{A}|}(r)} \frac{\left\|\frac{1}{n}\sum_{i=1}^{n}(\nabla\hat{V}(\boldsymbol{\theta},\boldsymbol{\tau}_i)-\nabla\hat{V}(\boldsymbol{\theta}_{i(\boldsymbol{\theta})},\boldsymbol{\tau}_i))\right\|_2}{\|\boldsymbol{\theta}-\boldsymbol{\theta}_{i(\boldsymbol{\theta})}\|_2} \sup_{\boldsymbol{\theta}\in B^{|\mathcal{S}||\mathcal{A}|}(r)} \|\boldsymbol{\theta}-\boldsymbol{\theta}_{i(\boldsymbol{\theta})}\|_2\right]\\
&\leq \frac{3\epsilon}{t}\mathbb{E}\left[\sup_{\boldsymbol{\theta}\in B^{|\mathcal{S}||\mathcal{A}|}(r)} \left\|\nabla^2\hat{\boldsymbol{V}}_n(\boldsymbol{\theta})\right\|_2\right],
\end{aligned}$$

where the first inequality holds by Markov inequality. By Lemma 6, we have

$$\begin{aligned}
\mathbb{E}\left[\sup_{\boldsymbol{\theta}\in B^{|\mathcal{S}||\mathcal{A}|}(r)} \left\|\nabla^2\hat{\boldsymbol{V}}_n(\boldsymbol{\theta})\right\|_2\right] &= \mathbb{E}\left[\sup_{\boldsymbol{\theta}\in B^{|\mathcal{S}||\mathcal{A}|}(r)} \left\|\frac{1}{n}\sum_{i=1}^{n}\nabla^2\hat{V}(\boldsymbol{\theta},\boldsymbol{\tau}_i)\right\|_2\right]\\
&= \mathbb{E}\left[\sup_{\boldsymbol{\theta}\in B^{|\mathcal{S}||\mathcal{A}|}(r)} \left\|\nabla^2\hat{V}(\boldsymbol{\theta},\boldsymbol{\tau})\right\|_2\right]\\
&\leq \frac{HG_h}{(1-\gamma)}.
\end{aligned}$$

Thus,

$$\mathbb{P}(\boldsymbol{E}_1) \leq \frac{3HG_h\epsilon}{t(1-\gamma)}$$

and by letting $t \geq \frac{6HG_h\epsilon}{(1-\gamma)\mathcal{E}}$ we have

$$\mathbb{P}(\boldsymbol{E}_1) \leq \frac{\mathcal{E}}{2}.$$

Secondly, we bound $\mathbb{P}(\boldsymbol{E}_2)$ as follows:

$$\begin{aligned}
\mathbb{P}(\boldsymbol{E}_2) &= \mathbb{P}\left(\sup_{i\in[N_\epsilon]} \left\|\frac{1}{n}\sum_{i=1}^{n}\nabla\hat{V}(\boldsymbol{\theta}_{i(\boldsymbol{\theta})},\boldsymbol{\tau}_i)-\mathbb{E}[\nabla V(\boldsymbol{\theta}_{i(\boldsymbol{\theta})},\boldsymbol{\tau})]\right\|_2 \geq \frac{t}{3}\right)\\
&\leq \left(\frac{3r}{\epsilon}\right)^{|\mathcal{S}||\mathcal{A}|} \sup_{i\in[N_\epsilon]} \mathbb{P}\left(\left\|\frac{1}{n}\sum_{i=1}^{n}\nabla\hat{V}(\boldsymbol{\theta}_{i(\boldsymbol{\theta})},\boldsymbol{\tau}_i)-\mathbb{E}[\nabla V(\boldsymbol{\theta}_{i(\boldsymbol{\theta})},\boldsymbol{\tau})]\right\|_2 \geq \frac{t}{3}\right)\\
&\leq \left(\frac{3r}{\epsilon}\right)^{|\mathcal{S}||\mathcal{A}|} 2\exp\left(-\frac{1}{32HG_g}nt^2(1-\gamma)^2+|A|\log 5\right)
\end{aligned}$$

where the last inequality is due to Lemma 9. Therefore, if we let

$$t \geq \sqrt{\frac{32HG_g(\log(4/\mathcal{E})+|\mathcal{S}||\mathcal{A}|\log(3r/\epsilon)+|A|\log(5))}{n(1-\gamma)^2}},$$

then we have

$$\mathbb{P}(\boldsymbol{E}_2) \leq \frac{\mathcal{E}}{2}.$$

Thirdly, we upper bound $\mathbb{P}(\boldsymbol{E}_3)$. Similarly,

$$\begin{aligned}
\mathbb{P}(\boldsymbol{E}_3) &= \mathbb{P}\left(\sup_{\boldsymbol{\theta}\in B^{|\mathcal{S}||\mathcal{A}|}(r)} \left\|\mathbb{E}[\nabla V(\boldsymbol{\theta}_{i(\boldsymbol{\theta})}, \boldsymbol{\tau})] - \mathbb{E}[\nabla V(\boldsymbol{\theta}, \boldsymbol{\tau})]\right\|_2 \geq \frac{t}{3}\right) \\
&\leq \mathbb{P}\left(\sup_{\boldsymbol{\theta}\in B^{|\mathcal{S}||\mathcal{A}|}(r)} \frac{\left\|\mathbb{E}[\nabla V(\boldsymbol{\theta}_{i(\boldsymbol{\theta})}, \boldsymbol{\tau})] - \mathbb{E}[\nabla V(\boldsymbol{\theta}, \boldsymbol{\tau})]\right\|_2}{\|\boldsymbol{\theta} - \boldsymbol{\theta}_{i(\boldsymbol{\theta})}\|_2} \sup_{\boldsymbol{\theta}\in B^{|\mathcal{S}||\mathcal{A}|}(r)} \|\boldsymbol{\theta} - \boldsymbol{\theta}_{i(\boldsymbol{\theta})}\|_2 \geq \frac{t}{3}\right) \\
&\leq \mathbb{P}\left(\epsilon\mathbb{E}\left[\sup_{\boldsymbol{\theta}\in B^{|\mathcal{S}||\mathcal{A}|}(r)} \left\|\nabla^2 \boldsymbol{V}(\boldsymbol{\theta}, \boldsymbol{\tau})\right\|_2\right] \geq \frac{t}{3}\right) \\
&\leq \mathbb{P}\left(\frac{\epsilon H G_h}{(1-\gamma)} \geq \frac{t}{3}\right).
\end{aligned}$$

By setting $\epsilon \leq \frac{t(1-\gamma)}{3HG_h}$, we ensure that $\mathbb{P}(\boldsymbol{E}_3) = 0$. Lastly, we let $\epsilon = \frac{3r\mathcal{E}}{n}$ and

$$\begin{aligned}
t &\geq \max\left\{\frac{6HG_h\epsilon}{(1-\gamma)\mathcal{E}}, \sqrt{\frac{32HG_g(\log(4/\mathcal{E}) + |\mathcal{S}||\mathcal{A}|\log(3r/\epsilon) + |A|\log(5))}{n(1-\gamma)^2}}\right\} \\
&= \max\left\{\frac{18HG_h r}{n(1-\gamma)}, \sqrt{\frac{32HG_g(\log(4/\mathcal{E}) + |\mathcal{S}||\mathcal{A}|\log(n/\mathcal{E}) + |A|\log(5))}{n(1-\gamma)^2}}\right\}
\end{aligned}$$

to ensure $\mathbb{P}\left(\sup_{\boldsymbol{\theta}\in B^{|\mathcal{S}||\mathcal{A}|}(r)} \left\|\nabla\hat{\boldsymbol{V}}_n(\boldsymbol{\theta}) - \nabla\boldsymbol{V}(\boldsymbol{\theta})\right\|_2 \geq t\right) \leq \mathcal{E}$. Therefore, if $n \geq \frac{64H}{(1-\gamma)}\max\left\{\frac{G_g|\mathcal{S}||\mathcal{A}|\log(n/\mathcal{E})}{(1-\gamma)}, G_h r\right\}$, then with probability at least $1-\mathcal{E}$

$$\sup_{\boldsymbol{\theta}\in\Theta} \left\|\nabla\hat{\boldsymbol{V}}_n(\boldsymbol{\theta}) - \nabla\boldsymbol{V}(\boldsymbol{\theta})\right\|_2 \leq \sqrt{\frac{64HG_g|\mathcal{S}||\mathcal{A}|\log(n/\mathcal{E})}{n(1-\gamma)^2}} + \frac{18HG_h r}{n(1-\gamma)}.$$

□

D Proof of Theorem 3

Theorem 7 (Theorem 3 restated). *Suppose assumption 2 and 3 hold. Then the Hessian of the estimator uniformly converges to the Hessian of the expected infinite-horizon discounted value function in operator norm. Specifically, if* $n \geq \frac{128H}{(1-\gamma)}\max\left\{\frac{G_h|\mathcal{S}||\mathcal{A}|\log(n/\mathcal{E})}{(1-\gamma)}, G_t r\right\}$, *then we have*

$$\sup_{\boldsymbol{\theta}\in\Theta} \left\|\nabla^2\hat{\boldsymbol{V}}_n(\boldsymbol{\theta}) - \nabla^2\boldsymbol{V}(\boldsymbol{\theta})\right\|_{op} \leq \sqrt{\frac{128HG_h|\mathcal{S}||\mathcal{A}|\log(n/\mathcal{E})}{n(1-\gamma)^2}} + \frac{18HG_t r}{n(1-\gamma)}.$$

with probability at least $1-\mathcal{E}$, *where* $\Theta = B^{|\mathcal{S}||\mathcal{A}|}(r)$.

Proof. We adopt a similar proof strategy as in the proof of Theorem 1. Note that the ϵ-covering number N_ϵ of the ball $B^{|\mathcal{S}||\mathcal{A}|}(r)$ is upper bounded by $\big(\frac{3r}{\epsilon}\big)^{|\mathcal{S}||\mathcal{A}|}$. We assume $\boldsymbol{\theta}_\epsilon = \{\boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_{N_\epsilon}\}$ is the ϵ-covering net of $B^{|\mathcal{S}||\mathcal{A}|}(r)$. Let $\boldsymbol{\theta} \in B^{|\mathcal{S}||\mathcal{A}|}(r)$ be an arbitrary vector, we have $\|\boldsymbol{\theta} - \boldsymbol{\theta}_i\| \leq \epsilon$ for some $\boldsymbol{\theta}_i \in \{\boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_{N_\epsilon}\}$. We denote $V(\boldsymbol{\theta}, \boldsymbol{\tau}) = \lim_{H\to\infty} \hat{V}_H(\boldsymbol{\theta}, \boldsymbol{\tau})$. For any $\boldsymbol{\theta} \in B^{|\mathcal{S}||\mathcal{A}|}(r)$, by the decomposition strategy we have

$$
\begin{aligned}
&\Big\|\nabla^2 \hat{\boldsymbol{V}}_n(\boldsymbol{\theta}) - \nabla^2 \boldsymbol{V}(\boldsymbol{\theta})\Big\|_{op} = \bigg\|\frac{1}{n}\sum_{i=1}^n \nabla^2 \hat{V}(\boldsymbol{\theta}, \boldsymbol{\tau}_i) - \mathbb{E}[\nabla^2 V(\boldsymbol{\theta}, \boldsymbol{\tau})]\bigg\|_{op} \\
&= \bigg\|\frac{1}{n}\sum_{i=1}^n \nabla^2 \hat{V}(\boldsymbol{\theta}, \boldsymbol{\tau}_i) - \frac{1}{n}\sum_{i=1}^n \nabla^2 \hat{V}(\boldsymbol{\theta}_{i(\boldsymbol{\theta})}, \boldsymbol{\tau}_i) + \frac{1}{n}\sum_{i=1}^n \nabla^2 \hat{V}(\boldsymbol{\theta}_{i(\boldsymbol{\theta})}, \boldsymbol{\tau}_i) - \mathbb{E}[\nabla^2 V(\boldsymbol{\theta}_{i(\boldsymbol{\theta})}, \boldsymbol{\tau})] \\
&\quad + \mathbb{E}[\nabla^2 V(\boldsymbol{\theta}_{i(\boldsymbol{\theta})}, \boldsymbol{\tau})] - \mathbb{E}[\nabla^2 V(\boldsymbol{\theta}, \boldsymbol{\tau})]\bigg\|_{op} \\
&\leq \bigg\|\frac{1}{n}\sum_{i=1}^n (\nabla^2 \hat{V}(\boldsymbol{\theta}, \boldsymbol{\tau}_i) - \nabla^2 \hat{V}(\boldsymbol{\theta}_{i(\boldsymbol{\theta})}, \boldsymbol{\tau}_i))\bigg\|_{op} + \bigg\|\frac{1}{n}\sum_{i=1}^n \nabla^2 \hat{V}(\boldsymbol{\theta}_{i(\boldsymbol{\theta})}, \boldsymbol{\tau}_i) - \mathbb{E}[\nabla^2 V(\boldsymbol{\theta}_{i(\boldsymbol{\theta})}, \boldsymbol{\tau})]\bigg\|_{op} \\
&\quad + \big\|\mathbb{E}[\nabla^2 V(\boldsymbol{\theta}_{i(\boldsymbol{\theta})}, \boldsymbol{\tau})] - \mathbb{E}[\nabla^2 V(\boldsymbol{\theta}, \boldsymbol{\tau})]\big\|_{op}
\end{aligned}
$$

where $i(\boldsymbol{\theta}) = \arg\min_{i\in[N_\epsilon]} \|\boldsymbol{\theta} - \boldsymbol{\theta}_i\|_2$. Then, we define three events $\boldsymbol{E}_1, \boldsymbol{E}_2, \boldsymbol{E}_3$ as

$$
\begin{aligned}
\boldsymbol{E}_1 &= \left\{\sup_{\boldsymbol{\theta}\in B^{|\mathcal{S}||\mathcal{A}|}(r)} \bigg\|\frac{1}{n}\sum_{i=1}^n (\nabla^2 \hat{V}(\boldsymbol{\theta}, \boldsymbol{\tau}_i) - \nabla^2 \hat{V}(\boldsymbol{\theta}_{i(\boldsymbol{\theta})}, \boldsymbol{\tau}_i))\bigg\|_{op} \geq \frac{t}{3}\right\}, \\
\boldsymbol{E}_2 &= \left\{\sup_{i\in[N_\epsilon]} \bigg\|\frac{1}{n}\sum_{i=1}^n \nabla^2 \hat{V}(\boldsymbol{\theta}_{i(\boldsymbol{\theta})}, \boldsymbol{\tau}_i) - \mathbb{E}[\nabla^2 V(\boldsymbol{\theta}_{i(\boldsymbol{\theta})}, \boldsymbol{\tau})]\bigg\|_{op} \geq \frac{t}{3}\right\}, \\
\boldsymbol{E}_3 &= \left\{\sup_{\boldsymbol{\theta}\in B^{|\mathcal{S}||\mathcal{A}|}(r)} \big\|\mathbb{E}[\nabla^2 V(\boldsymbol{\theta}_{i(\boldsymbol{\theta})}, \boldsymbol{\tau})] - \mathbb{E}[\nabla^2 V(\boldsymbol{\theta}, \boldsymbol{\tau})]\big\|_{op} \geq \frac{t}{3}\right\}.
\end{aligned}
$$

Accordingly, we have

$$
\mathbb{P}\left(\sup_{\boldsymbol{\theta}\in B^{|\mathcal{S}||\mathcal{A}|}(r)} \Big\|\nabla^2 \hat{\boldsymbol{V}}_n(\boldsymbol{\theta}) - \nabla^2 \boldsymbol{V}(\boldsymbol{\theta})\Big\|_{op} \geq t\right) \leq \mathbb{P}(\boldsymbol{E}_1) + \mathbb{P}(\boldsymbol{E}_2) + \mathbb{P}(\boldsymbol{E}_3).
$$

Next, we first upper bound $\mathbb{P}(\boldsymbol{E}_1)$.

$$
\begin{aligned}
\mathbb{P}(\boldsymbol{E}_1) &= \mathbb{P}\left(\sup_{\boldsymbol{\theta}\in B^{|\mathcal{S}||\mathcal{A}|}(r)} \bigg\|\frac{1}{n}\sum_{i=1}^n (\nabla^2 \hat{V}(\boldsymbol{\theta}, \boldsymbol{\tau}_i) - \nabla^2 \hat{V}(\boldsymbol{\theta}_{i(\boldsymbol{\theta})}, \boldsymbol{\tau}_i))\bigg\|_{op} \geq \frac{t}{3}\right) \\
&\leq \frac{3}{t}\mathbb{E}\left[\sup_{\boldsymbol{\theta}\in B^{|\mathcal{S}||\mathcal{A}|}(r)} \bigg\|\frac{1}{n}\sum_{i=1}^n (\nabla^2 \hat{V}(\boldsymbol{\theta}, \boldsymbol{\tau}_i) - \nabla^2 \hat{V}(\boldsymbol{\theta}_{i(\boldsymbol{\theta})}, \boldsymbol{\tau}_i))\bigg\|_{op}\right]
\end{aligned}
$$

$$\leq \frac{3}{t}\mathbb{E}\left[\sup_{\boldsymbol{\theta}\in B^{|\mathcal{S}||\mathcal{A}|}(r)} \frac{\left\|\frac{1}{n}\sum_{i=1}^{n}(\nabla^2\hat{V}(\boldsymbol{\theta},\boldsymbol{\tau}_i)-\nabla^2\hat{V}(\boldsymbol{\theta}_{i(\boldsymbol{\theta})},\boldsymbol{\tau}_i))\right\|_{op}}{\|\boldsymbol{\theta}-\boldsymbol{\theta}_{i(\boldsymbol{\theta})}\|_2} \sup_{\boldsymbol{\theta}\in B^{|\mathcal{S}||\mathcal{A}|}(r)} \|\boldsymbol{\theta}-\boldsymbol{\theta}_{i(\boldsymbol{\theta})}\|_2\right]$$

$$\leq \frac{3\epsilon}{t}\mathbb{E}\left[\sup_{\boldsymbol{\theta}\in B^{|\mathcal{S}||\mathcal{A}|}(r)} \left\|\nabla^3\hat{\boldsymbol{V}}_n(\boldsymbol{\theta})\right\|_F\right],$$

where the first inequality holds by Markov inequality. By Lemma 6, we have

$$\begin{aligned}\mathbb{E}\left[\sup_{\boldsymbol{\theta}\in B^{|\mathcal{S}||\mathcal{A}|}(r)} \left\|\nabla^3\hat{\boldsymbol{V}}_n(\boldsymbol{\theta})\right\|_F\right] &= \mathbb{E}\left[\sup_{\boldsymbol{\theta}\in B^{|\mathcal{S}||\mathcal{A}|}(r)} \left\|\frac{1}{n}\sum_{i=1}^{n}\nabla^3\hat{V}(\boldsymbol{\theta},\boldsymbol{\tau}_i)\right\|_F\right] \\ &= \mathbb{E}\left[\sup_{\boldsymbol{\theta}\in B^{|\mathcal{S}||\mathcal{A}|}(r)} \left\|\nabla^3\hat{V}(\boldsymbol{\theta},\boldsymbol{\tau})\right\|_F\right] \\ &\leq \frac{HG_t\epsilon}{(1-\gamma)}.\end{aligned}$$

Thus,

$$\mathbb{P}(\boldsymbol{E}_1) \leq \frac{3HG_t\epsilon}{t(1-\gamma)}$$

and by letting $t \geq \frac{6HG_t\epsilon}{(1-\gamma)\mathcal{E}}$ we have

$$\mathbb{P}(\boldsymbol{E}_1) \leq \frac{\mathcal{E}}{2}.$$

Secondly, we bound $\mathbb{P}(\boldsymbol{E}_2)$ as follows:

$$\begin{aligned}\mathbb{P}(\boldsymbol{E}_2) &= \mathbb{P}\left(\sup_{i\in[N_\epsilon]}\left\|\frac{1}{n}\sum_{i=1}^{n}\nabla^2\hat{V}(\boldsymbol{\theta}_{i(\boldsymbol{\theta})},\boldsymbol{\tau}_i)-\mathbb{E}[\nabla^2 V(\boldsymbol{\theta}_{i(\boldsymbol{\theta})},\boldsymbol{\tau})]\right\|_{op} \geq \frac{t}{3}\right) \\ &\leq \left(\frac{3r}{\epsilon}\right)^{|\mathcal{S}||\mathcal{A}|} \sup_{i\in[N_\epsilon]} \mathbb{P}\left(\left\|\frac{1}{n}\sum_{i=1}^{n}\nabla^2\hat{V}(\boldsymbol{\theta}_{i(\boldsymbol{\theta})},\boldsymbol{\tau}_i)-\mathbb{E}[\nabla^2 V(\boldsymbol{\theta}_{i(\boldsymbol{\theta})},\boldsymbol{\tau})]\right\|_{op} \geq \frac{t}{3}\right) \\ &\leq \left(\frac{3r}{\epsilon}\right)^{|\mathcal{S}||\mathcal{A}|} 2\exp\left(-\frac{1}{64HG_h}nt^2\left(1-\gamma\right)^2+|A|\log 9\right)\end{aligned}$$

where the last inequality is due to Lemma 10. Therefore, if we let

$$t \geq \sqrt{\frac{64HG_h(\log(4/\mathcal{E})+|\mathcal{S}||\mathcal{A}|\log(3r/\epsilon)+|A|\log(9))}{n(1-\gamma)^2}},$$

then we have

$$\mathbb{P}(\boldsymbol{E}_2) \leq \frac{\mathcal{E}}{2}.$$

Thirdly, we upper bound $\mathbb{P}(\boldsymbol{E}_3)$. Similarly,

$$\begin{aligned}
\mathbb{P}(\boldsymbol{E}_3) &= \mathbb{P}\left(\sup_{\boldsymbol{\theta}\in\mathbb{S}^{|A|-1}} \left\|\mathbb{E}[\nabla^2 V(\boldsymbol{\theta}_{i(\boldsymbol{\theta})},\boldsymbol{\tau})] - \mathbb{E}[\nabla^2 V(\boldsymbol{\theta},\boldsymbol{\tau})]\right\|_{op} \geq \frac{t}{3}\right) \\
&\leq \mathbb{P}\left(\sup_{\boldsymbol{\theta}\in\mathbb{S}^{|A|-1}} \frac{\left\|\mathbb{E}[\nabla^2 V(\boldsymbol{\theta}_{i(\boldsymbol{\theta})},\boldsymbol{\tau})] - \mathbb{E}[\nabla^2 V(\boldsymbol{\theta},\boldsymbol{\tau})]\right\|_{op}}{\|\boldsymbol{\theta}-\boldsymbol{\theta}_{i(\boldsymbol{\theta})}\|_2} \sup_{\boldsymbol{\theta}\in\mathbb{S}^{|A|-1}} \|\boldsymbol{\theta}-\boldsymbol{\theta}_{i(\boldsymbol{\theta})}\|_2 \geq \frac{t}{3}\right) \\
&\leq \mathbb{P}\left(\epsilon\mathbb{E}\left[\sup_{\boldsymbol{\theta}\in\mathbb{S}^{|A|-1}} \left\|\nabla^3 \boldsymbol{V}(\boldsymbol{\theta},\boldsymbol{\tau})\right\|_F\right] \geq \frac{t}{3}\right) \\
&\leq \mathbb{P}\left(\frac{\epsilon H G_t \epsilon}{(1-\gamma)} \geq \frac{t}{3}\right).
\end{aligned}$$

By setting $\epsilon \leq \frac{t(1-\gamma)}{3HG_t}$, we ensure that $\mathbb{P}(\boldsymbol{E}_3) = 0$. Lastly, we let $\epsilon = \frac{3r\mathcal{E}}{n}$ and

$$\begin{aligned}
t &\geq \max\left\{\frac{6HG_t\epsilon}{(1-\gamma)\mathcal{E}}, \sqrt{\frac{64HG_h(\log(4/\mathcal{E}) + |\mathcal{S}||\mathcal{A}|\log(3r/\epsilon) + |A|\log(9))}{n(1-\gamma)^2}}\right\} \\
&= \max\left\{\frac{18HG_t r}{n(1-\gamma)}, \sqrt{\frac{64HG_h(\log(4/\mathcal{E}) + |\mathcal{S}||\mathcal{A}|\log(n/\mathcal{E}) + |A|\log(9))}{n(1-\gamma)^2}}\right\}
\end{aligned}$$

to ensure $\mathbb{P}\left(\sup_{\boldsymbol{\theta}\in\mathbb{S}^{|A|-1}} \left\|\nabla^2 \hat{\boldsymbol{V}}_n(\boldsymbol{\theta}) - \nabla^2 \boldsymbol{V}(\boldsymbol{\theta})\right\|_{op} \geq t\right) \leq \mathcal{E}$. Therefore, if $n \geq \frac{128H}{(1-\gamma)}\max\left\{\frac{G_h|\mathcal{S}||\mathcal{A}|\log(n/\mathcal{E})}{(1-\gamma)}, G_t r\right\}$, then with probability at least $1-\mathcal{E}$

$$\sup_{\boldsymbol{\theta}\in\Theta} \left\|\nabla^2 \hat{\boldsymbol{V}}_n(\boldsymbol{\theta}) - \nabla^2 \boldsymbol{V}(\boldsymbol{\theta})\right\|_{op} \leq \sqrt{\frac{128HG_h|\mathcal{S}||\mathcal{A}|\log(n/\mathcal{E})}{n(1-\gamma)^2}} + \frac{18HG_t r}{n(1-\gamma)}.$$

□