Stochastic Second-Order Methods Improve Best-Known Sample Complexity of SGD for Gradient-Dominated Functions

Saeed Masiha[*]
College of Management of Technology
EPFL, Lausanne, Switzerland
`mohammadsaeed.masiha@epfl.ch`

Saber Salehkaleybar[*]
School of Computer and Communication Sciences
EPFL, Lausanne, Switzerland
`saber.salehkaleybar@epfl.ch`

Niao He
Department of Computer Science
ETH, Zurich, Switzerland
`niao.he@inf.ethz.ch`

Negar Kiyavash
College of Management of Technology
EPFL, Lausanne, Switzerland
`negar.kiyavash@epfl.ch`

Patrick Thiran
School of Computer and Communication Sciences
EPFL, Lausanne, Switzerland
`patrick.thiran@epfl.ch`

Abstract

We study the performance of Stochastic Cubic Regularized Newton (SCRN) on a class of functions satisfying gradient dominance property with $1 \leq \alpha \leq 2$ which holds in a wide range of applications in machine learning and signal processing. This condition ensures that any first-order stationary point is a global optimum. We prove that the total sample complexity of SCRN in achieving ϵ-global optimum is $\mathcal{O}(\epsilon^{-7/(2\alpha)+1})$ for $1 \leq \alpha < 3/2$ and $\tilde{\mathcal{O}}(\epsilon^{-2/(\alpha)})$ for $3/2 \leq \alpha \leq 2$. SCRN improves the best-known sample complexity of stochastic gradient descent. Even under a weak version of gradient dominance property, which is applicable to policy-based reinforcement learning (RL), SCRN achieves the same improvement over stochastic policy gradient methods. Additionally, we show that the average sample complexity of SCRN can be reduced to $\mathcal{O}(\epsilon^{-2})$ for $\alpha = 1$ using a variance reduction method with time-varying batch sizes. Experimental results in various RL settings showcase the remarkable performance of SCRN compared to first-order methods.

1 Introduction

Consider the following unconstrained stochastic non-convex optimization problem:

$$\min_{\mathbf{x} \in \mathbb{R}^d} F(\mathbf{x}) := \mathbb{E}_\xi[f(\mathbf{x}, \xi)], \tag{1}$$

where the random variable ξ is sampled from an underlying distribution P_ξ. In order to optimize the objective function $F(\mathbf{x})$, we have access to the first and second derivatives of stochastic function $f(\mathbf{x}, \xi)$. The above optimization problem covers a wide range of problems, from the offline setting where the objective function is minimized over a fixed number of samples, to the online setting where the samples are drawn sequentially.

---

[*]equal contribution

36th Conference on Neural Information Processing Systems (NeurIPS 2022).

In the deterministic case (where we have access to the derivatives of $F(\mathbf{x})$), the gradient descent (GD) algorithm in the non-convex setting only guarantees convergence to a first-order stationary point (FOSP) (i.e., a point $\mathbf{x}$ such that $\|\nabla F(\mathbf{x})\| = 0$), which can be a local minimum, a local maximum, or a saddle point. In contrast, second-order methods accessing the Hessian of F (or Hessian of $f(\mathbf{x}, \xi)$ in the stochastic setting) can exploit the curvature information to effectively escape saddle points and converge to a second-order stationary point (SOSP) (i.e., such that $\|\nabla F(\mathbf{x})\| = 0$, $\nabla^2 F(\mathbf{x}) \succeq 0$). In their seminal work, Nesterov and Polyak [25] proposed the so-called cubic-regularized Newton (CRN) algorithm which exploits Hessian information and globally converges to an SOSP at a sub-linear rate of $\mathcal{O}(1/k^{2/3})$, where k is the number of iterations.

In recent years, the performance of stochastic CRN (SCRN) for general non-convex functions has been the focus of several studies (for more details, see the related work in Section 1.1). A variance-reduced version of SCRN [3] can find (ϵ, γ)-SOSP (i.e., a point $\mathbf{x}$ such that $\|\nabla F(\mathbf{x})\| \leq \epsilon$ and $\nabla^2 F(\mathbf{x}) \succeq -\gamma I$) with the sample complexity of $\tilde{\mathcal{O}}(\epsilon^{-3})$. Moreover, this rate is optimal for achieving ϵ-approximate FOSP (i.e., a point $\mathbf{x}$ such that $\|\nabla F(\mathbf{x})\| \leq \epsilon$) and it cannot be improved using any stochastic p-th order methods for $p \geq 2$ [3].

Nesterov and Polyak [25] studied CRN under the gradient dominance property (See Assumption 2). Under Assumption 2 with $\alpha = 2$, they showed that iterates of $F(\mathbf{x}_t) - \min_\mathbf{x} F(\mathbf{x})$ converges to zero super-linearly. The case of $\alpha = 2$ (commonly called Polyak-Łojasiewicz (PL) condition) was originally introduced by Polyak in [28], who showed that GD achieves linear convergence rate. The gradient dominance property or its weak variations are satisfied in a quite wide range of machine learning applications such as neural networks with one hidden layer [18] or ResNet with linear activation [12] (for more details, see Section 1.1). A particularly important application of the weak version of gradient dominance property with $\alpha = 1$ (Assumption 6) is in policy-based reinforcement learning (RL) [41].

Khaled et al. [16] showed that under PL condition, the stochastic GD (SGD) with time-varying step-size returns a point $\hat{\mathbf{x}}$ with a sample complexity of $\mathcal{O}(1/\epsilon)$, to reach $\mathbb{E}[F(\hat{\mathbf{x}})] - \min_\mathbf{x} F(\mathbf{x}) \leq \epsilon$. Furthermore, the dependency of the sample complexity of SGD on ϵ is optimal [26]. Recently, Fontaine et al. [10] obtained a sample complexity of $\mathcal{O}(\epsilon^{-\frac{4}{\alpha}+1})$ for SGD under gradient dominance property with $1 \leq \alpha \leq 2$. This shows that the worst sample complexity occurs for $\alpha = 1$, which is precisely the value of α that finds important applications in policy-based RL. Indeed, under a weak version of gradient dominance property with $\alpha = 1$, it has been shown that stochastic policy gradient (SPG) converges to the optimum point with a sample complexity of $\tilde{\mathcal{O}}(\epsilon^{-3})$ [41, 9].

We know that in the deterministic case, CRN outperforms GD under gradient dominance property for all $\alpha \in [1, 2]$ [25, 50][2]. Therefore, a natural question that arises is whether this holds true in the stochastic setting as well? That is, does SCRN improve upon SGD under the gradient dominance property?

Herein, we address this question. Specifically, our main contributions are as follows:

- We analyze the sample complexity of SCRN under gradient dominance property for $1 \leq \alpha \leq 2$ in order to return an ϵ-global stationary point $\hat{\mathbf{x}}$ satisfying $F(\hat{\mathbf{x}}) - \min_\mathbf{x} F(\mathbf{x}) \leq \epsilon$ (in expectation or with high probability). As stated in Table 1, SCRN improves upon the best-known sample complexity of SGD for all $1 \leq \alpha < 2$. The largest improvement is for $\alpha = 1$, and is $\mathcal{O}(\epsilon^{-0.5})$.

- In the setting of policy-based RL, under the weak version of gradient dominance property with $\alpha = 1$ (Assumption 6), we show that SCRN achieves a sample complexity of $\tilde{\mathcal{O}}(\epsilon^{-2.5})$, improving over the best-known sample complexity of SPG by a factor of $\tilde{\mathcal{O}}(\epsilon^{-0.5})$.

- We show that an adaptation of a variance-reduced SCRN [47] with time-varying batch sizes further improves the sample complexity of SCRN. For $\alpha = 1$, the average sample complexity is reduced to $\mathcal{O}(\epsilon^{-2})$.

---

[2]In particular, for $\alpha \in [1, 3/2)$, the number of iterations of CRN is $\mathcal{O}(1/\epsilon^{3/(2\alpha)-1})$, for $\alpha = 3/2$ is $\mathcal{O}(\log(1/\epsilon))$, and for $\alpha \in (3/2, 2]$, the number of iterations is $\mathcal{O}(\log \log(1/\epsilon))$. For $\alpha \in [1, 2)$, the number of iterations of GD is $\mathcal{O}(1/\epsilon^{2/\alpha-1})$ and for $\alpha = 2$, is $\mathcal{O}(\log(1/\epsilon))$ [50].

2

Table 1: Comparison of sample complexities of SGD on Lipschitz gradient functions and SCRN on Lipschitz Hessian functions to achieve ϵ-global stationary point under gradient dominance property with $\alpha \in [1, 2]$. The last column indicates the improvement of SCRN with respect to SGD.

α	SGD [10]	SCRN (Ours)	Improvement
$[1, \frac{3}{2})$	$\mathcal{O}(\epsilon^{-4/\alpha+1})$	$\mathcal{O}(\epsilon^{-7/(2\alpha)+1})$	$\mathcal{O}(\epsilon^{-1/(2\alpha)})$
$\frac{3}{2}$	$\mathcal{O}(\epsilon^{-4/\alpha+1}) = \mathcal{O}(\epsilon^{-5/3})$	$\tilde{\mathcal{O}}(\epsilon^{-7/(2\alpha)+1}) = \tilde{\mathcal{O}}(\epsilon^{-4/3})$	$\tilde{\mathcal{O}}(\epsilon^{-1/3})$
$(\frac{3}{2}, 2)$	$\mathcal{O}(\epsilon^{-4/\alpha+1})$	$\tilde{\mathcal{O}}(\epsilon^{-2/\alpha})$ w.h.p	$\tilde{\mathcal{O}}(\epsilon^{-2/\alpha+1})$
2	$\mathcal{O}(\epsilon^{-1})$	$\tilde{\mathcal{O}}(\epsilon^{-2/\alpha}) = \tilde{\mathcal{O}}(\epsilon^{-1})$ w.h.p	–

1.1 Related work

Gradient dominance property and its applications: The gradient dominance property with $\alpha = 2$ (commonly called PL condition) was originally introduced by Polyak in [28]. It was shown by Karimi et al. [15] to be weaker than the most recent global optimality conditions that appeared in the literature of machine learning [19, 23, 42]. The gradient dominance property is also satisfied (sometimes locally rather than globally, and also under distributional assumptions) for the population risk in some learning models including neural networks with one hidden layer [18], ResNet with linear activation [12], and generalized linear model and robust regression [11]. Moreover, in policy-based reinforcement learning (RL), a weak version of gradient dominance property with $\alpha = 1$ (see Assumption 6) holds for some classes of policies (such as Gaussian policy and log-linear policy).

Variants of cubic regularized Newton method: For non-convex optimization, Nesterov and Polyak [25] proposed the CRN algorithm, which converges to a SOSP with the convergence rate of $\mathcal{O}(1/k^{2/3})$ (where k is the number of iterations) by solving a cubic sub-problem in each iteration. Cartis et al. [5] presented an adaptive framework for cubic regularization method. In [17, 39], sub-sampled versions of gradient and Hessian were used in CRN to overcome the computational burden of Hessian matrix evaluations in high dimensional settings. In the context of stochastic optimization, Tripuraneni et al. [33] proposed a stochastic cubic regularization algorithm that obtains ϵ-SOSP with sample complexity of $\tilde{\mathcal{O}}(\epsilon^{-3.5})$. Arjavani et al. [3] improved the sample complexity to $\tilde{\mathcal{O}}(\epsilon^{-3})$ using variance reduction. In the convex setting, Song et al. [30] presented a proximal CRN and its accelerated version, and proved a sample complexity of $\tilde{\mathcal{O}}(\epsilon^{-2})$ to reach ϵ-global stationary point as long as the approximated Hessian in each iteration satisfies certain properties. In finite-sum non-convex setting, Zhou et al. [48] proposed an adaptive sub-sampled CRN method that requires $\tilde{\mathcal{O}}(N + N^{4/5}\epsilon^{-3/2})$ to find ϵ-SOSP, where N is the total number of samples. Sample complexity was further reduced to $\tilde{\mathcal{O}}(N + N^{2/3}\epsilon^{-3/2})$ using various variance reduction methods [36, 45, 49]. To the best of our knowledge, no previous work on analyzing SCRN for gradient-dominant functions exists.

1.2 Notations

We adopt the following notation in the sequel. Calligraphic letters (e.g., $\mathcal{S}$) denote spaces. Upper-case bold letters (e.g., $\mathbf{A}$) denote matrices, and the lower-case bold letters (e.g., $\mathbf{x}$) denote vectors. $\| \cdot \|$ denote the ℓ_2-norm for vectors and the operator norm for matrices ($\|\mathbf{A}\| := \lambda_{\max}(\mathbf{A}^T\mathbf{A})$ where $\lambda_{\max}(\mathbf{X})$ is the maximum eigenvalue of matrix $\mathbf{X}$), respectively. $\mathbf{A} \succeq \mathbf{B}$ indicates that $\mathbf{A} - \mathbf{B}$ is positive semi-definite. We use the notation $\mathcal{O}$ to hide constants, and the notation $\tilde{\mathcal{O}}$ to hide both constants and logarithmic factors. $X \leq_{1-\delta} Y$ denotes that random variable X is less than or equal to random variable Y with the probability at least $1 - \delta$.

2 Setup

Recall the stochastic non-convex optimization problem in (1), where the goal is to minimize the objective function $F(\mathbf{x})$ having access to stochastic gradients $\nabla f(\mathbf{x}, \xi)$ and stochastic Hessian matrices $\nabla^2 f(\mathbf{x}, \xi)$. We make the following assumption about the objective function in (1).

Assumption 1. *The Hessian of F is Lipschitz continuous with constant L_2, i.e.,*

$$\|\nabla^2 F(\mathbf{x}) - \nabla^2 F(\mathbf{y})\| \leq L_2 \|\mathbf{x} - \mathbf{y}\|_2, \quad \forall \mathbf{x}, \mathbf{y} \in \mathbb{R}^d. \tag{2}$$

3

Consider the empirical estimators $\mathbf{g}_t := \frac{1}{n_1}\sum_{i=1}^{n_1}\nabla f(\mathbf{x}_t, \xi_i)$, and $\mathbf{H}_t := \frac{1}{n_2}\sum_{i=1}^{n_2}\nabla^2 f(\mathbf{x}_t, \xi_i)$ where n_1 and n_2 are the numbers of samples used for estimating the gradient vector and Hessian matrix, respectively.

Algorithm 1 Stochastic cubic regularized Newton method with stopping criterion

---

Input: Batch sizes n_1, n_2, initial point $\mathbf{x}_0$, accuracy ϵ, cubic penalty parameter M, maximum number of iterations T

1: $t \leftarrow 1$
2: $\|\boldsymbol{\Delta}_0\| = \infty$
3: **while** $\|\boldsymbol{\Delta}_{t-1}\| \geq \sqrt[2\alpha]{\epsilon}$ or $t \leq T$ **do**
4:      $\mathbf{g}_t \leftarrow \frac{1}{n_1}\sum_{i=1}^{n_1}\nabla f(\mathbf{x}_t, \xi_i)$
5:      $\mathbf{H}_t \leftarrow \frac{1}{n_2}\sum_{i=1}^{n_2}\nabla^2 f(\mathbf{x}_t, \xi_i)$
6:      $\boldsymbol{\Delta}_t \leftarrow \arg\min_{\boldsymbol{\Delta}\in\mathbb{R}^d}\langle \mathbf{g}_t, \boldsymbol{\Delta}\rangle + \frac{1}{2}\langle\boldsymbol{\Delta}, \mathbf{H}_t\boldsymbol{\Delta}\rangle + \frac{M}{6}\|\boldsymbol{\Delta}\|^3$
7:      $\mathbf{x}_{t+1} \leftarrow \mathbf{x}_t + \boldsymbol{\Delta}_t$
8:      $t \leftarrow t+1$
9: **end while**
10: **return** $\mathbf{x}_t$

Definition 1 (Total sample complexity). *Given $\epsilon, \delta > 0$, the total sample complexity is the number of calls (queries) of stochastic gradient and stochastic Hessian along the iterations until reaching a point $\mathbf{x}$ that satisfies one of the following: (1) for high probability analysis: $F(\mathbf{x}) - F(\mathbf{x}^*) \leq \epsilon$ with probability at least $1 - \delta$; or (2) for analysis in expectation: $\mathbb{E}[F(\mathbf{x})] - F(\mathbf{x}^*) \leq \epsilon$.*

Our goal is to study the performance of stochastic cubic regularized Newton (SCRN) for objective functions that satisfy the gradient dominance property. Algorithm 1 describes the steps in SCRN. At each iteration t, we take batches of stochastic gradient vectors and Hessian matrices (lines 4 and 5) and then solve the following sub-problem to obtain $\boldsymbol{\Delta}_t$ (line 6):

$$\min_{\boldsymbol{\Delta}\in\mathbb{R}^d} m_t(\boldsymbol{\Delta}) := \langle\mathbf{g}_t, \boldsymbol{\Delta}\rangle + \frac{1}{2}\langle\boldsymbol{\Delta}, \mathbf{H}_t\boldsymbol{\Delta}\rangle + \frac{M}{6}\|\boldsymbol{\Delta}\|^3. \tag{3}$$

We assume that there is an oracle that returns a global solution for this sub-problem (This assumption will be relaxed subsequently, see Remark 6). Finally, we update $\mathbf{x}_t$ in line 7.

3   SCRN under gradient dominance property

We shall study the performance of SCRN for functions satisfying gradient dominance property, defined as follows.

Assumption 2. *Function $F(\mathbf{x})$ satisfies gradient dominance property when for every $\mathbf{x} \in \mathbb{R}^d$,*

$$F(\mathbf{x}) - F(\mathbf{x}^*) \leq \tau_F\|\nabla F(\mathbf{x})\|^\alpha, \tag{4}$$

where $\mathbf{x}^* \in \arg\min_{\mathbf{x}} F(\mathbf{x})$, $\tau_F > 0$, and $\alpha \in [1, 2]$ are two constants.

The case $\alpha = 2$ is often referred to as PL condition [28, 15]. In this paper, we consider all α's in the interval $[1, 2]$. The gradient dominance property holds for a large class of functions including sub-analytic functions, logarithm, and exponential functions, and semi-algebraic functions. These function classes cover some of the most common non-convex objectives used in practice (see related work in Section 1.1).

In the following lemma, we present a recursion inequality that captures the behaviour of the function $F(\mathbf{x}_t) - F(\mathbf{x}^*)$ at each iteration t for SCRN under gradient dominance property.

Lemma 1. *Assume that function F satisfies Assumption 1 (Lipschitz Hessian) and Assumption 2 (gradient dominance property) for $\alpha \geq 1$. Then the resulting update $\mathbf{x}_{t+1}$ in Algorithm 1 (line 7) after plugging in $\boldsymbol{\Delta}_t$, the solution of sub-problem in (3), satisfies the following:*

$$F(\mathbf{x}_{t+1}) - F(\mathbf{x}^*) \leq$$
$$C(F(\mathbf{x}_t) - F(\mathbf{x}_{t+1}))^{2\alpha/3} + C_g\|\nabla F(\mathbf{x}_t) - \mathbf{g}_t\|^\alpha + C_H\|\nabla^2 F(\mathbf{x}_t) - \mathbf{H}_t\|^{2\alpha}, \tag{5}$$

where $C, C_g, C_H > 0$ are constants depending on M, L_2, and τ_F, and defined in (34).

4

Due to space limitations, all proofs are moved to the appendix.

In the following, we first provide an analysis in expectation of SCRN under gradient dominance property with $\alpha \in [1, 3/2]$. Next, we study the same algorithm for $\alpha \in (3/2, 2]$ using a high probability analysis.

3.1 SCRN under gradient dominance property with $\alpha \in [1, 3/2]$

We make the following assumption on the stochastic gradients and Hessians.

Assumption 3. *For a given $\alpha \in [1, 3/2]$ and for each query point $\mathbf{x} \in \mathbb{R}^d$:*

$$\mathbb{E}[\nabla f(\mathbf{x}, \xi)] = \nabla F(\mathbf{x}), \quad \mathbb{E}[\|\nabla f(\mathbf{x}, \xi) - \nabla F(\mathbf{x})\|_2^2] \leq \sigma_1^2, \tag{6}$$

$$\mathbb{E}[\nabla^2 f(\mathbf{x}, \xi)] = \nabla^2 F(\mathbf{x}), \quad \mathbb{E}[\|\nabla^2 f(\mathbf{x}, \xi) - \nabla^2 F(\mathbf{x})\|^{2\alpha}] \leq \sigma_{2,\alpha}^2, \tag{7}$$

where σ_1 and $\sigma_{2,\alpha}$ are two constants.

Remark 1. *The assumption $\mathbb{E}[\|\nabla^2 f(\mathbf{x}, \xi) - \nabla^2 F(\mathbf{x})\|^{2\alpha}] \leq \sigma_{2,\alpha}^2$ for $1 \leq \alpha \leq 3/2$ is slightly stronger than the usual assumption $\mathbb{E}[\|\nabla^2 f(\mathbf{x}, \xi) - \nabla^2 F(\mathbf{x})\|^2] \leq \sigma_2^2$. We need this assumption because of the specific form of the error of Hessian estimator in recursion inequality in (5). As a result of the assumption in (7), using a version of matrix moment inequality (See Lemma 4 in Appendix), we show that the dependency of the Hessian sample complexity in Theorem 1 on dimension d is in the order of $\log d$.*

Theorem 1. *Let $F(\mathbf{x})$ satisfy Assumptions 1 and 2 for a given α and the stochastic gradient and Hessian satisfy Assumption 3 for the same α. Moreover, assume that an exact solver for sub-problem (3) exists. Then Algorithm 1 outputs a point $\mathbf{x}_T$ such that $\mathbb{E}[F(\mathbf{x}_T)] - F(\mathbf{x}^*) \leq \epsilon$ after T iterations, where*

(i) if $\alpha \in [1, 3/2)$, $T = \mathcal{O}(\epsilon^{-\frac{3-2\alpha}{2\alpha}})$, with access to the following numbers of samples of the stochastic gradient and Hessian per iteration:

$$n_1 \geq \frac{C_g^{2/\alpha}}{C^{6/\alpha}} \cdot \frac{4^{2/\alpha} \sigma_1^{2/\alpha}}{\epsilon^{2/\alpha}}, \quad n_2 \geq \frac{C_H'^{1/\alpha}}{C^{3/\alpha}} \cdot \frac{4^{1/\alpha} \sigma_{2,\alpha}^{2/\alpha}}{\epsilon^{1/\alpha}}, \tag{8}$$

where C_H' is defined in (38) and depends on $\log(d)$.

(ii) if $\alpha = 3/2$, $T = \mathcal{O}(\log(1/\epsilon))$ with the same numbers of samples per iteration as in (35).

3.2 SCRN under gradient dominance property with $\alpha \in (3/2, 2]$

Definition 2 (Bernstein's condition for matrices). *A zero-mean symmetric random matrix $\mathbf{X}$ satisfies the Bernstein condition with parameter $b > 0$ if*

$$\mathbb{E}[\mathbf{X}^k] \preccurlyeq \frac{1}{2} k! \, b^{k-2} \, \mathrm{Var}(\mathbf{X}), \quad for \ k = 3, 4, \ldots \tag{9}$$

where $\mathrm{Var}(\mathbf{X}) := \mathbb{E}[\mathbf{X}^2] - (\mathbb{E}[\mathbf{X}])^2$.

Assumption 4. *We assume that the symmetric version of each centered gradient estimator $\mathbf{G}(\mathbf{x}, \xi) := \begin{bmatrix} \mathbf{0}_{1 \times 1} & \mathbf{g}(\mathbf{x}, \xi)^T \\ \mathbf{g}(\mathbf{x}, \xi) & \mathbf{0}_{d \times d} \end{bmatrix}$ where $\mathbf{g}(\mathbf{x}, \xi) := \nabla f(\mathbf{x}, \xi) - \nabla F(\mathbf{x})$ and each centered Hessian estimator $\mathbf{H}(\mathbf{x}, \xi) := \nabla^2 f(\mathbf{x}, \xi) - \nabla^2 F(\mathbf{x})$ satisfy Bernstein's condition (9) with parameters M_1 and M_2, respectively.*

Remark 2. *It is noteworthy that most previous work analyzing SCRN [33, 49, 36] assumed that centered gradient and centered Hessian estimators are bounded, i.e., $\|\nabla f(\mathbf{x}, \xi) - \nabla F(\mathbf{x})\|_2 \overset{a.s.}{\leq} M_1, \|\nabla^2 f(\mathbf{x}, \xi) - \nabla^2 F(\mathbf{x})\| \overset{a.s.}{\leq} M_2$. This is a stronger assumption than Assumption 4 as it implies Bernstein's condition (9) for $\mathbf{g}(\mathbf{x}, \xi)$ and $\mathbf{H}(\mathbf{x}, \xi)$.*

Theorem 2. *Suppose that $F(\mathbf{x})$ satisfies Assumptions 1, 2, the stochastic gradient and Hessian satisfy Assumption 3 (with $\alpha = 1$) and Assumption 4, and there exists an exact solver for sub-problem (3). Then, Algorithm 1, with probability $1 - \delta$, outputs a solution $\mathbf{x}_T$ such that $F(\mathbf{x}_T) - F(\mathbf{x}^*) \leq \epsilon$*

5

after $T = \mathcal{O}\left(\log\left(\log(1/\epsilon)\right)\right)$ iterations with the following numbers of samples for the stochastic gradient and Hessian, respectively per iteration:

$$n_1 \geq \frac{8}{3} \max\left(\frac{\tilde{C}^{1/\alpha} M_1}{\epsilon^{1/\alpha}}, \frac{\tilde{C}^{2/\alpha} \sigma_1^2}{\epsilon^{2/\alpha}}\right) \log\left(\frac{4(T+1)d}{\delta}\right), \tag{10}$$

$$n_2 \geq \frac{8}{3} \max\left(\frac{\tilde{C}^{1/(2\alpha)} M_2}{\epsilon^{1/(2\alpha)}}, \frac{\tilde{C}^{1/\alpha} \sigma_{2,1}^2}{\epsilon^{1/\alpha}}\right) \log\left(\frac{4(T+1)d}{\delta}\right), \tag{11}$$

where $\tilde{C} = 1 + \frac{\tau_F}{2}\left(\frac{M+L_2+4}{2}\right)^\alpha$.

Remark 3. *It is noteworthy that the assumptions in Theorem 2 suffice to obtain the sample complexities in Theorem 1 for $1 \leq \alpha \leq 3/2$ up to a logarithmic factor. However, the assumptions of Theorem 2 are stronger than those in Theorem 1.*

Remark 4. *Theorem 2 implies that the sample complexity of Algorithm 1 for $\alpha = 2$ is $\mathcal{O}\left(\log\log\log(1/\epsilon) \cdot \log(\log(1/\epsilon))/\epsilon\right)$.*

We summarized the sample complexity of SCRN in Theorems 1 and 2 and the best-known sample complexity of SGD under gradient dominance property in Table 1. The optimal sample complexity of SGD under gradient dominance property with $\alpha = 2$ is discussed in detail in [16, Section 5.2]. For the general case of $\alpha \in [1, 2]$, the sample complexity of SGD under gradient dominance property was derived in [10, Theorem 10]. This rate is achieved with a time-varying step-size. In the fourth column of Table 1, we provide the improvement of the sample complexity of SCRN with respect to that of SGD. The improvement is $\mathcal{O}\left(\epsilon^{-1/(2\alpha)}\right)$ for $\alpha \in [1, 3/2)$ and is $\tilde{\mathcal{O}}(\epsilon^{-2/\alpha+1})$ for $\alpha \in [3/2, 2]$, which are decreasing functions of α. The largest improvement is for $\alpha = 1$ and is $\mathcal{O}(\epsilon^{-0.5})$.

Remark 5. *The special case $\alpha = 2$ generalizes strong convexity. For strongly convex objectives, the dependence of sample complexity of SGD on τ_F is $\mathcal{O}(\tau_F^2/\epsilon)$ [16, Corollary 2] while the sample complexity of SCRN is $\mathcal{O}(\tau_F^{7/4}/\epsilon)$. This is an improvement by a factor of $\tau_F^{1/4}$. In Appendix A.2.3, we provide simulation results comparing the performance of SCRN and SGD over synthetic functions by varying α and τ_F.*

Remark 6. *In our analysis, we assumed that we have access to the exact solution of sub-problem in (3). Although no closed-form solution nor exact solver exists for this sub-problem, there are algorithms that approximate the exact solution with high probability [2, 4]. We emphasize that solving the sub-problem in (3) requires extra computation, but extra gradient and Hessian queries are not needed. In particular, Carmon and Duchi [4] proposed a perturbed GD-based algorithm that returns an approximate solution $\tilde{\boldsymbol{\Delta}}_t$ such that $m_t(\tilde{\boldsymbol{\Delta}}_t) \leq_{1-\delta'} m_t(\boldsymbol{\Delta}_t) + \epsilon'$ with $\mathcal{O}(\log(1/\delta')/\epsilon')$ iterations, for any given $\delta', \epsilon' > 0$.*

In Appendix A.2.2, we prove the following lemma:

Lemma 2. *Theorems 1 and 2 are still true if we were to use an inexact sub-solver which returned an approximate solution $\tilde{\boldsymbol{\Delta}}_t$ such that $\|\nabla m_t(\tilde{\boldsymbol{\Delta}}_t)\| \leq \epsilon^{1/\alpha}$ ($\epsilon^{1/\alpha}$-stationary point). Moreover, under some mild assumptions (same as those in [4]), a GD-based algorithm indeed returns such a solution in $\mathcal{O}(\epsilon^{-2/\alpha})$ iterations.*

Remark 7. *In the iterations of GD-based algorithm, instead of directly computing the Hessian matrix (which could be computationally expensive in high dimensions), we could compute Hessian-vector products by running Pearlmutter's algorithm [27]. Thus, the total computational complexity of evaluating gradients and Hessian vector products of SCRN is in the order of $\mathcal{O}(d\epsilon^{-3/\alpha})$ while the one of SGD is $\mathcal{O}(d\epsilon^{-4/\alpha+1})$, where d is the dimension of $\mathbf{x}$. For $\alpha \in [1, 3/2)$, it can be seen that the computational complexity of SGD is less than the one of SCRN by a factor of $\mathcal{O}(\epsilon^{-1/(2\alpha)})$. For $\alpha \in [3/2, 2]$, this factor is $\tilde{\mathcal{O}}(\epsilon^{-(1-1/\alpha)})$.*

3.3 Further improvements

In our analysis of SCRN, we set the batch sizes such that the stochastic error terms $C_g\|\nabla F(\mathbf{x}_t) - \mathbf{g}_t\|^\alpha$ and $C_H\|\nabla^2 F(\mathbf{x}_t) - \mathbf{H}_t\|^{2\alpha}$ in (5) are in the order of ϵ (either in expectation or with high probability). However, for $1 \leq \alpha < 3/2$, it is just needed to make sure that the error terms at iteration t are $\mathcal{O}(t^{-(2\alpha)/(3-2\alpha)})$, which equals the convergence rate of the function values $F(\mathbf{x}_t) - F(\mathbf{x}^*)$ (see Lemma 11 in Appendix A.3). Moreover, as stated in the following theorem, incorporating time-varying batch sizes in conjunction with variance reduction improves sample complexity results.

6

Assumption 5. *We assume that $f(\mathbf{x}, \xi)$ satisfies L_1'-average smoothness and L_2'-average Hessian Lipschitz continuity, i.e., $\mathbb{E}[\|\nabla f(\mathbf{x}, \xi) - \nabla f(\mathbf{y}, \xi)\|^2] \leq L_1'^2 \|\mathbf{x} - \mathbf{y}\|^2$ and $\mathbb{E}[\|\nabla^2 f(\mathbf{x}, \xi) - \nabla^2 f(\mathbf{y}, \xi)\|^2] \leq L_2'^2 \|\mathbf{x} - \mathbf{y}\|^2$ for all $\mathbf{x}, \mathbf{y} \in \mathbb{R}^d$.*

Theorem 3. *Suppose that $F(\mathbf{x})$ satisfies the gradient dominance property with $\alpha = 1$. Assume that Assumptions 1, 3, and 5 for $\alpha = 1$ hold. Then variance-reduced SCRN (See Algorithm 2 in Appendix A.3) achieves ϵ-global stationary point in expectation by making $\mathcal{O}(\epsilon^{-2})$ stochastic gradients and $\mathcal{O}(\epsilon^{-1})$ stochastic Hessian on average, and the iteration complexity is $\mathcal{O}(\epsilon^{-1/2})$.*

We will see in the next section that there are a class of objective functions in RL setting that satisfies a weak version of gradient dominance with $\alpha = 1$ and Lipschitz gradient property and then we can take advantage of variance-reduced SCRN (See Remark 11).

4 SCRN under weak gradient dominance property in RL Setting

In this section, we showcase practical relevance of our result in Section 3.1 by applying SCRN to model-free RL. Specifically, in Theorem 4, we prove that as long as the expected return satisfies the weak version of gradient dominance property (Assumption 6), SCRN improves upon the best-known sample complexity of stochastic policy gradient (SPG) [41] by a factor of $\mathcal{O}(1/\sqrt{\epsilon})$ (see Appendix A.4.1 for the related work).

4.1 RL Setup

Consider a discrete Markov decision process (MDP) $\mathcal{M} = (\mathcal{S}, \mathcal{A}, P, R, \rho, \gamma)$, where $\mathcal{S}$ is the state space and $\mathcal{A}$ is the action space. $P(s'|s, a)$ denotes the probability of state transition from s to s' after taking action a and $R(\cdot, \cdot) : \mathcal{S} \times \mathcal{A} \rightarrow [-R_{\max}, R_{\max}]$ is a bounded reward function, where $R_{\max}$ is a positive scalar. ρ represents the initial distribution on state space $\mathcal{S}$ and $\gamma \in (0, 1)$ is the discount factor.

The parametric policy π_θ is a probability distribution over $\mathcal{S} \times \mathcal{A}$ with parameter $\theta \in \mathbb{R}^d$, and $\pi_\theta(a|s)$ denotes the probability of taking action a at a given state s. Let $\tau = \{s_t, a_t\}_{t \geq 0} \sim p(\tau|\pi_\theta)$ be a trajectory generated by the policy π_θ, where $p(\tau|\pi_\theta) := \rho(s_0) \prod_{t=0}^{\infty} \pi_\theta(a_t|s_t) P(s_{t+1}|s_t, a_t)$. The expected return of π is defined as $J(\pi_\theta) := \mathbb{E}_{\tau \sim p(\cdot|\pi_\theta)} \left[\sum_{t=0}^{\infty} \gamma^t R(s_t, a_t) \right]$. In the sequel, we consider a set of parameterized policies $\{\pi_\theta : \theta \in \mathbb{R}^d\}$, with the assumption that π_θ is differentiable with respect to θ. For ease of presentation, we denote $J(\pi_\theta)$ by $J(\theta)$.

The goal of policy-based RL is to find $\theta^* = \arg\max_\theta J(\theta)$. However, in many cases, $J(\theta)$ is a non-concave function and instead we settle for obtaining an ϵ-FOSP, $\hat{\theta}$, such that $\|\nabla J(\hat{\theta})\| \leq \epsilon$. It can be shown that: $\nabla J(\theta) = \mathbb{E} \left[\sum_{h=0}^{\infty} \left(\sum_{t=h}^{\infty} \gamma^t R(s_t, a_t) \right) \nabla \log \pi_\theta(a_h|s_h) \right]$. In practice, the full gradient cannot be computed due to the infinite horizon length. Instead, it is commonly truncated to a length H horizon as follows. $\nabla J_{\mathsf{H}}(\theta) = \mathbb{E}[\sum_{h=0}^{\mathsf{H}-1} \Psi_h(\tau) \nabla \log \pi_\theta(a_h|s_h)]$, where $\Psi_h(\tau) = \sum_{t=h}^{\mathsf{H}-1} \gamma^t R(s_t, a_t)$.

Assume that we sample m trajectories $\tau^i = \{s_t^i, a_t^i\}_{t \geq 0}$, $1 \leq i \leq m$, and then compute $\hat{\nabla}_m J(\theta) = \frac{1}{m} \sum_{i=1}^{m} \sum_{h=0}^{\mathsf{H}-1} \Psi_h(\tau^i) \nabla \log \pi_\theta(a_h^i|s_h^i)$, which is an unbiased estimator for $\nabla J_{\mathsf{H}}(\theta)$. The vanilla SPG method is based on the following update: $\theta \leftarrow \theta + \eta \hat{\nabla}_m J(\theta)$, where η is the learning rate.

It can be shown that the Hessian matrix of $J_{\mathsf{H}}(\theta)$ can be obtained as follows [29, Appendix 7.2]: $\nabla^2 J_{\mathsf{H}}(\theta) = \mathbb{E}[\nabla \Phi(\theta; \tau) \nabla \log p(\tau|\pi_\theta)^T + \nabla^2 \Phi(\theta; \tau)]$, where $\Phi(\theta; \tau) = \sum_{h=0}^{\mathsf{H}-1} \sum_{t=h}^{\mathsf{H}-1} \gamma^t r(s_t, a_t) \log \pi_\theta(a_h|s_h)$. As a result, for trajectories $\tau^i = \{s_t^i, a_t^i\}_{t \geq 0}$, $1 \leq i \leq m$, $\hat{\nabla}_m^2 J(\theta) = \frac{1}{m} \sum_{i=1}^{m} \nabla \Phi(\theta; \tau^i) \nabla \log p(\tau^i|\pi_\theta)^T + \nabla^2 \Phi(\theta; \tau^i)$ is an unbiased estimator of Hessian matrix $\nabla^2 J_{\mathsf{H}}(\theta)$.

4.2 Sample Complexity of SCRN

In our analysis, we consider the recently introduced relaxed weak gradient dominance property with $\alpha = 1$ [41].

Assumption 6 (Weak gradient dominance property with $\alpha = 1$). *J satisfies the weak gradient dominance property if for all $\theta \in \mathbb{R}^d$, there exist $\tau_J > 0$ and $\epsilon' > 0$ such that*

$$\epsilon' + \tau_J \|\nabla J(\theta)\| \geq J^* - J(\theta), \tag{12}$$

7

where $J^* := \max_\theta J(\theta)$.

Remark 8. *Two commonly made assumptions in the literature: non-degenerate Fisher matrix [20, 9] and transferred compatible function approximation error [35, 1] imply Assumption 6. See Appendix A.4.1 for more details.*

Further, we also make Lipschitz and smooth policy (LS) assumptions which are widely adopted in the analysis of vanilla policy gradient (PG) [46] as well as variance-reduced PG methods, e.g. in [29].

Assumption 7 (LS)**.** *There exist constants $G_1, G_2 > 0$ such that for every state $s \in \mathcal{S}$, the gradient and Hessian of $\log \pi_\theta(\cdot|s)$ satisfy $\|\nabla_\theta \log \pi_\theta(a|s)\| \leq G_1$ and $\|\nabla_\theta^2 \log \pi_\theta(a|s)\| \leq G_2$.*

Lemma 3. *Under Assumption 7, we have $\|\nabla J(\theta) - \nabla J_{\mathsf{H}}(\theta)\| \leq D_g \gamma^{\mathsf{H}}$ and $\|\nabla^2 J(\theta) - \nabla^2 J_{\mathsf{H}}(\theta)\| \leq D_H \gamma^{\mathsf{H}}$, where $D_g = \frac{G_1 R_{\max}}{1-\gamma}\sqrt{\frac{1}{1-\gamma} + \mathsf{H}}$ and $D_H = \frac{R_{\max}(G_2+G_1^2)}{1-\gamma}\left(\mathsf{H} + \frac{1}{1-\gamma}\right).$*

Assumption 8 (Lipschitz Hessian)**.** *There exists a constant $\bar{L}_2$ such that the Hessian of $\log \pi_\theta(a|s)$ satisfies*

$$\|\nabla^2 \log \pi_\theta(a|s) - \nabla^2 \log \pi_{\theta'}(a|s)\| \leq \bar{L}_2 \|\theta - \theta'\|. \tag{13}$$

The Lipschitz Hessian assumption is commonly used to find SOSP in policy gradient algorithms [40]. For the Gaussian policy (136), $\nabla^2 \log \pi_\theta(a|s)$ reduces to the matrix $-\phi(s)\phi(s)^T/\sigma^2$, which is a constant function of θ and thus satisfies condition (13). Soft-max policy also satisfies this assumption (See appendix A.4.4).

Theorem 4. *For a policy π_θ satisfying Assumptions 7, 8, and the corresponding objective function $J(\theta)$ satisfying Assumption 6, SCRN outputs the solution θ_T such that $J^* - \mathbb{E}[J(\theta_T)] \leq \epsilon + \epsilon'$ and the sample complexity (the number of observed state-action pairs) is: $T \times m \times \mathsf{H} = \tilde{\mathcal{O}}(\epsilon^{-2.5})$ for $\epsilon' = 0$ and $T \times m \times \mathsf{H} = \tilde{\mathcal{O}}(\epsilon^{-0.5}\epsilon'^{-2})$ for $\epsilon' > 0$.*

Remark 9. *Under weak gradient dominance property with $\alpha = 1$ (Assumption 6), it has been shown that the sample complexity of SPG is $\tilde{\mathcal{O}}(\epsilon^{-3})$ in case of $\epsilon' = 0$ and $\tilde{\mathcal{O}}(\epsilon^{-1}\epsilon'^{-2})$ in case of $\epsilon' > 0$ [41, Theorem C.1]. Therefore, SCRN improves upon the best-known sample complexity of SPG in both cases $\epsilon' = 0$ and $\epsilon' > 0$ by a factor of $\mathcal{O}(\epsilon^{-0.5})$.*

Remark 10. *Having access to exact gradient and Hessian (the deterministic case), under Assumption 6, PG algorithm achieves global convergence (i.e., $J^* - J(\theta_T) \leq \epsilon$) with $\tilde{\mathcal{O}}(\epsilon^{-1})$ iterations [41] while CRN requires $\tilde{\mathcal{O}}(\epsilon^{-0.5})$ iterations.*

Remark 11. *Under the same assumptions as in Theorem 4 and Assumption 5 for $\alpha = 1$ and bounded variance of importance sampling weights (See Assumption 15), a variance-reduced version of SCRN (See Algorithm 3) achieves global convergence (i.e., $J^* - \mathbb{E}[J(\theta_T)] \leq \epsilon$) with a sample complexity of $\tilde{\mathcal{O}}(\epsilon^{-2})$. See Appendix A.4.6 for a proof.*

5 Experiments

In this section, we evaluate the performance of SCRN in the two following RL settings. First, we consider grid world environments with finite state and action spaces, and next some robotic control tasks with continuous state and action spaces. The details of the implementations for all methods and some additional experiments appear in the appendix and the codes are available in the supplementary material.

Environments with finite state and action spaces: We consider two grid world environments in our experiments: cliff walking [31, Example 6.6], and random mazes [51]. In cliff walking, the agent's aim is to reach a goal state from a start state, avoiding a region of cells called "cliff". The episode is terminated if the agent enters the cliff region, or the number of steps exceeds 100 without reaching the goal. Moreover, we consider a soft-max tabular policy in the experiments of this part. Indeed, it has been shown that a variant of gradient dominance property with $\alpha = 1$ holds for soft-max tabular policy in environments with finite state and action spaces [22].

We compare SCRN with two existing first-order methods: vanilla SPG and REINFORCE [37]. For the first-order methods, we use a time-varying learning rate and tune the parameters. To improve the performance of the first-order methods, we also add an entropy regularization term to the reward function.

In Fig. 1, the average length of paths traversed by the agent and the average episode return are depicted against the number of episodes for each method. The results are averaged over 64 instances

8

(a) Avg. path length   (b) Avg. episode return
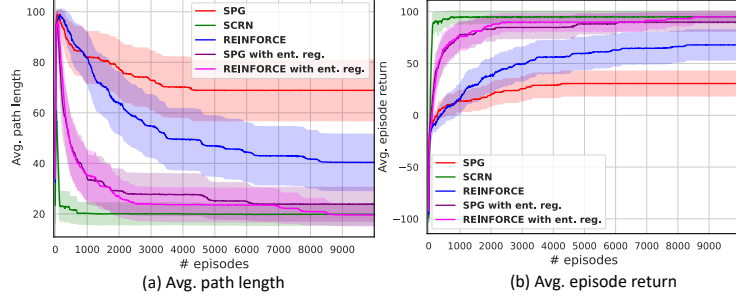
Figure 1: Comparison of SCRN with first-order methods in cliff walking environment. The percentages of successful instances for SPG, SCRN, REINFORCE, SPG with entropy regularization, and REINFORCE with entropy regularization are 32.8%, 100%, 54.7%, 100%, and 92.2%, respectively.
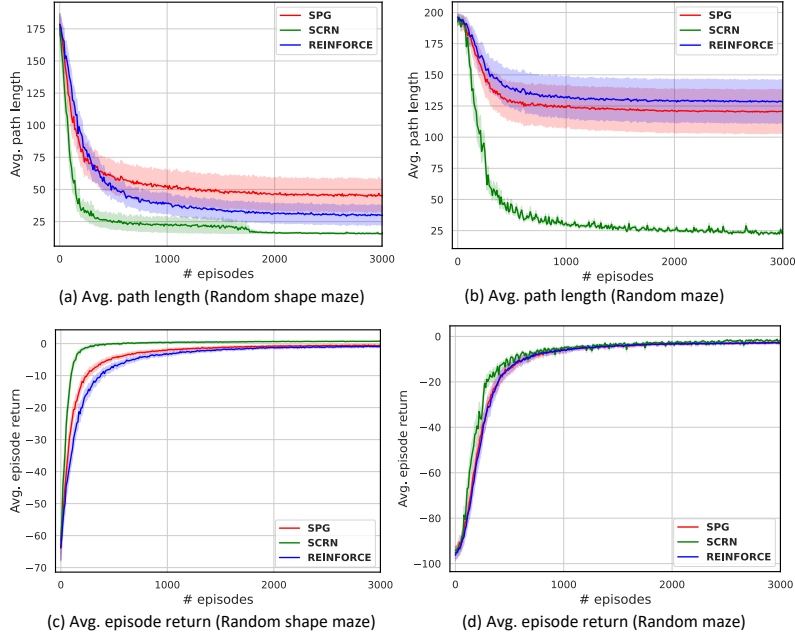
Figure 2: Comparison of SCRN with first-order methods in maze environments. In random shape maze, the percentages of successful instances for SPG, SCRN, and REINFORCE are 86%, 100%, and 95.3%, respectively. In random maze, the percentages of successful instances for SPG, SCRN, and REINFORCE are 45.3%, 97%, and 40.6% respectively.

and the shaded region shows the 90% confidence interval. As can be seen in Fig. 1 (a), all the algorithms have a phase at the beginning during which the agent falls off the cliff in most episodes and the average length of paths are small. Then, the agent learns to avoid the cliff but could still not reach the goal in most cases. Finally, it finds a path to the goal and tries to reduce the path length. Note that SCRN finds the path very quickly and significantly outperforms the other algorithms. In fact, SPG and REINFORCE get stuck in the start state for some period of time, while SCRN easily escapes the flat plateau (for more details, please see the demonstrations in the supplementary file). The performance of SPG and REINFORCE improves with entropy regularization, but SCRN still outperforms them. Moreover, SPG and REINFORCE fail to reach the goal in some instances while SCRN is successful in almost all instances. In the captions of figures, for each algorithm, we also provide the percentage of instances in which the agent reached the goal. Please note that these percentages are obtained based on the parameters from the last update of each algorithm.

Additionally, we studied the performance of the three aforementioned algorithms on a random maze and a random shape maze [51]. In the random shape maze, random shape blocks are placed on a grid and the agent tries to reach the goal state finding the shortest path. As shown in Fig. 2, SCRN again

9
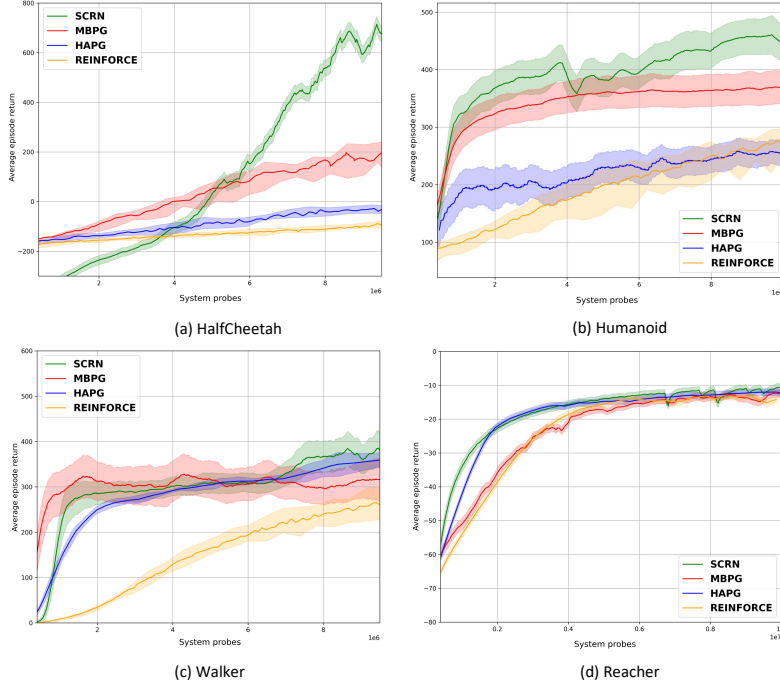
Figure 3: Comparison of SCRN with REINFORCE and variance-reduced SPG methods in MuJoCo environments.

outperforms the first-order methods in both environments. In Appendix A.5, we also provided results for SPG and REINFROCE with entropy regularization which have slightly better performance.

Environments with continuous state and action spaces: We consider the following control tasks in MuJoCo simulator [32]: Walker, Humanoid, Reacher, and HalfCheetah. We compare SCRN with first-order methods such as REINFORCE, and two state-of-the-art representatives of variance-reduced PG methods, HAPG [29] and MBPG [13], both with guaranteed convergence to ϵ-FOSP in general non-convex settings. HAPG uses second order information (Hessian vector products) for variance reduction and MBPG is a recent work based on STORM, a batch-free state-of-the-art variance reduction approach [8].

We report average episode return against system probes as our performance measure. That is, the number of observed state-action pairs (see Fig. 3). At each point, we run the trained policy 10 times and compute the empirical estimate of the mean and the 90% confidence interval of the episode return. As seen in Fig. 3, SCRN outperforms the other methods, especially in more complex environments such as HalfCheetah and Humanoid. In Appendix A.5, we also provided results for variance-reduced SCRN which improves upon SCRN in Humanoid and Reacher environments.

6 Conclusion

We studied the performance of SCRN for objectives satisfying the gradient dominance property for $1 \leq \alpha \leq 2$, which holds in various machine learning applications. We showed that SCRN improves the best-known sample complexity of SGD. The largest improvement is in the case of $\alpha = 1$. Moreover, for $\alpha = 1$, the average sample complexity of SCRN can be reduced to $\mathcal{O}(\epsilon^{-2})$ by utilizing a variance reduction method with time-varying batch sizes. A weak version of gradient dominance for $\alpha = 1$ is satisfied in some policy-based RL settings. In the RL setting, we showed that SCRN achieves the same improvement over SPG under the weak version of gradient dominance property for $\alpha = 1$.

7 Acknowledgement

The authors would like to thank Mohammadsadegh Khorasani for conducting the experiments for continuous state-action environments. This research is partially supported by National Centre of Competence in Research (NCCR), grant agreement no. 51NF40_180545.

10

References

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

[2] Naman Agarwal, Zeyuan Allen-Zhu, Brian Bullins, Elad Hazan, and Tengyu Ma. Finding approximate local minima for nonconvex optimization in linear time. *arXiv preprint arXiv:1611.01146*, 2016.

[3] Yossi Arjevani, Yair Carmon, John C Duchi, Dylan J Foster, Ayush Sekhari, and Karthik Sridharan. Second-order information in non-convex stochastic optimization: Power and limitations. In *Conference on Learning Theory*, pages 242–299. PMLR, 2020.

[4] Yair Carmon and John C Duchi. Gradient descent efficiently finds the cubic-regularized non-convex newton step. *arXiv preprint arXiv:1612.00547*, 2016.

[5] Coralia Cartis, Nicholas IM Gould, and Philippe L Toint. Adaptive cubic regularisation methods for unconstrained optimization. part i: motivation, convergence and numerical results. *Mathematical Programming*, 127(2):245–295, 2011.

[6] Semih Cayci, Niao He, and R Srikant. Linear convergence of entropy-regularized natural policy gradient with linear function approximation. *arXiv preprint arXiv:2106.04096*, 2021.

[7] RY Chen, A Gittens, and JA Tropp. The masked sample covariance estimator: An analysis via matrix concentration inequalities. arxiv e-print. *arXiv preprint arXiv:1109.1637*, 2011.

[8] Ashok Cutkosky and Francesco Orabona. Momentum-based variance reduction in non-convex sgd. *Advances in neural information processing systems*, 32, 2019.

[9] Yuhao Ding, Junzi Zhang, and Javad Lavaei. On the global convergence of momentum-based policy gradient. *arXiv preprint arXiv:2110.10116*, 2021.

[10] Xavier Fontaine, Valentin De Bortoli, and Alain Durmus. Convergence rates and approximation results for sgd and its continuous-time counterpart. In *Conference on Learning Theory*, pages 1965–2058. PMLR, 2021.

[11] Dylan J Foster, Ayush Sekhari, and Karthik Sridharan. Uniform convergence of gradients for non-convex learning and optimization. *Advances in Neural Information Processing Systems*, 31, 2018.

[12] Moritz Hardt and Tengyu Ma. Identity matters in deep learning. *arXiv preprint arXiv:1611.04231*, 2016.

[13] Feihu Huang, Shangqian Gao, Jian Pei, and Heng Huang. Momentum-based policy gradient methods. In *International Conference on Machine Learning*, pages 4422–4433. PMLR, 2020.

[14] Sham Kakade and John Langford. Approximately optimal approximate reinforcement learning. In *In Proc. 19th International Conference on Machine Learning*. Citeseer, 2002.

[15] Hamed Karimi, Julie Nutini, and Mark Schmidt. Linear convergence of gradient and proximal-gradient methods under the polyak-łojasiewicz condition. In *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*, pages 795–811. Springer, 2016.

[16] Ahmed Khaled and Peter Richtárik. Better theory for sgd in the nonconvex world. *arXiv preprint arXiv:2002.03329*, 2020.

[17] Jonas Moritz Kohler and Aurelien Lucchi. Sub-sampled cubic regularization for non-convex optimization. In *International Conference on Machine Learning*, pages 1895–1904. PMLR, 2017.

[18] Yuanzhi Li and Yang Yuan. Convergence analysis of two-layer neural networks with relu activation. *Advances in neural information processing systems*, 30, 2017.

11

[19] Ji Liu, Steve Wright, Christopher Ré, Victor Bittorf, and Srikrishna Sridhar. An asynchronous parallel stochastic coordinate descent algorithm. In *International Conference on Machine Learning*, pages 469–477. PMLR, 2014.

[20] Yanli Liu, Kaiqing Zhang, Tamer Basar, and Wotao Yin. An improved analysis of (variance-reduced) policy gradient and natural policy gradient methods. *Advances in Neural Information Processing Systems*, 33:7624–7636, 2020.

[21] Lester Mackey, Michael I Jordan, Richard Y Chen, Brendan Farrell, and Joel A Tropp. Matrix concentration inequalities via the method of exchangeable pairs. *The Annals of Probability*, 42 (3):906–945, 2014.

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

[23] Ion Necoara, Yu Nesterov, and Francois Glineur. Linear convergence of first order methods for non-strongly convex optimization. *Mathematical Programming*, 175(1):69–107, 2019.

[24] Yurii Nesterov. *Introductory lectures on convex optimization: A basic course*, volume 87. Springer Science & Business Media, 2003.

[25] Yurii Nesterov and Boris T Polyak. Cubic regularization of newton method and its global performance. *Mathematical Programming*, 108(1):177–205, 2006.

[26] Phuong Ha Nguyen, Lam Nguyen, and Marten van Dijk. Tight dimension independent lower bound on the expected convergence rate for diminishing step sizes in sgd. *Advances in Neural Information Processing Systems*, 32, 2019.

[27] Barak A Pearlmutter. Fast exact multiplication by the hessian. *Neural computation*, 6(1): 147–160, 1994.

[28] Boris Teodorovich Polyak. Gradient methods for minimizing functionals. *Zhurnal Vychislitel'noi Matematiki i Matematicheskoi Fiziki*, 3(4):643–653, 1963.

[29] Zebang Shen, Alejandro Ribeiro, Hamed Hassani, Hui Qian, and Chao Mi. Hessian aided policy gradient. In *International conference on machine learning*, pages 5729–5738. PMLR, 2019.

[30] Chaobing Song, Ji Liu, and Yong Jiang. Inexact proximal cubic regularized newton methods for convex optimization. *arXiv preprint arXiv:1902.02388*, 2019.

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

[32] Emanuel Todorov, Tom Erez, and Yuval Tassa. Mujoco: A physics engine for model-based control. In *2012 IEEE/RSJ international conference on intelligent robots and systems*, pages 5026–5033. IEEE, 2012.

[33] Nilesh Tripuraneni, Mitchell Stern, Chi Jin, Jeffrey Regier, and Michael I Jordan. Stochastic cubic regularization for fast nonconvex optimization. *Advances in neural information processing systems*, 31, 2018.

[34] Martin J Wainwright. *High-dimensional statistics: A non-asymptotic viewpoint*, volume 48. Cambridge University Press, 2019.

[35] Lingxiao Wang, Qi Cai, Zhuoran Yang, and Zhaoran Wang. Neural policy gradient methods: Global optimality and rates of convergence. *arXiv preprint arXiv:1909.01150*, 2019.

[36] Zhe Wang, Yi Zhou, Yingbin Liang, and Guanghui Lan. Stochastic variance-reduced cubic regularization for nonconvex optimization. In *The 22nd International Conference on Artificial Intelligence and Statistics*, pages 2731–2740. PMLR, 2019.

[37] Ronald J Williams. Simple statistical gradient-following algorithms for connectionist reinforcement learning. *Machine learning*, 8(3):229–256, 1992.

12

[38] Pan Xu, Felicia Gao, and Quanquan Gu. An improved convergence analysis of stochastic variance-reduced policy gradient. In *Uncertainty in Artificial Intelligence*, pages 541–551. PMLR, 2020.

[39] Peng Xu, Fred Roosta, and Michael W Mahoney. Second-order optimization for non-convex machine learning: An empirical study. In *Proceedings of the 2020 SIAM International Conference on Data Mining*, pages 199–207. SIAM, 2020.

[40] Long Yang, Qian Zheng, and Gang Pan. Sample complexity of policy gradient finding second-order stationary points. *arXiv preprint arXiv:2012.01491*, 2020.

[41] Rui Yuan, Robert M Gower, and Alessandro Lazaric. A general sample complexity analysis of vanilla policy gradient. *arXiv preprint arXiv:2107.11433*, 2021.

[42] Haibin Zhang, Jiaojiao Jiang, and Zhi-Quan Luo. On the linear convergence of a proximal gradient method for a class of nonsmooth convex minimization problems. *Journal of the Operations Research Society of China*, 1(2):163–186, 2013.

[43] Junyu Zhang, Alec Koppel, Amrit Singh Bedi, Csaba Szepesvari, and Mengdi Wang. Variational policy gradient method for reinforcement learning with general utilities. *Advances in Neural Information Processing Systems*, 33:4572–4583, 2020.

[44] Junyu Zhang, Chengzhuo Ni, Csaba Szepesvari, Mengdi Wang, et al. On the convergence and sample efficiency of variance-reduced policy gradient method. *Advances in Neural Information Processing Systems*, 34, 2021.

[45] Junyu Zhang, Lin Xiao, and Shuzhong Zhang. Adaptive stochastic variance reduction for subsampled newton method with cubic regularization. *INFORMS Journal on Optimization*, 4 (1):45–64, 2022.

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

[47] Dongruo Zhou and Quanquan Gu. Stochastic recursive variance-reduced cubic regularization methods. In *International Conference on Artificial Intelligence and Statistics*, pages 3980–3990. PMLR, 2020.

[48] Dongruo Zhou, Pan Xu, and Quanquan Gu. Stochastic variance-reduced cubic regularized newton methods. In *International Conference on Machine Learning*, pages 5990–5999. PMLR, 2018.

[49] Dongruo Zhou, Pan Xu, and Quanquan Gu. Stochastic variance-reduced cubic regularization methods. *J. Mach. Learn. Res.*, 20(134):1–47, 2019.

[50] Yi Zhou, Zhe Wang, and Yingbin Liang. Convergence of cubic regularization for nonconvex optimization under kl property. *Advances in Neural Information Processing Systems*, 31, 2018.

[51] Xingdong Zuo. mazelab: A customizable framework to create maze and gridworld environments. https://github.com/zuoxingdong/mazelab, 2018.

13

Checklist

1. For all authors...

    I Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope? [Yes]

    II Did you describe the limitations of your work? [Yes] In all the statements in the paper (theorems, lemmas,...), we explicitly mentioned what are the assumptions under which the statements holds.

    III Did you discuss any potential negative societal impacts of your work? [No] This is a theoretical work and it does not have any meaningful negative societal impacts.

    IV Have you read the ethics review guidelines and ensured that your paper conforms to them? [Yes]

2. If you are including theoretical results...

    I Did you state the full set of assumptions of all theoretical results? [Yes]

    II Did you include complete proofs of all theoretical results? [Yes]

3. If you ran experiments...

    I Did you include the code, data, and instructions needed to reproduce the main experimental results (either in the supplemental material or as a URL)? [Yes] The codes with the instructions are available as the supplementary material.

    II Did you specify all the training details (e.g., data splits, hyperparameters, how they were chosen)? [Yes] We discuss them in the appendix.

    III Did you report error bars (e.g., with respect to the random seed after running experiments multiple times)? [Yes] We run our experiments on multiple instances and reported confidence intervals.

    IV Did you include the total amount of compute and the type of resources used (e.g., type of GPUs, internal cluster, or cloud provider)? [Yes] We mentioned them in the details of experiments in the appendix.

4. If you are using existing assets (e.g., code, data, models) or curating/releasing new assets...

    I If your work uses existing assets, did you cite the creators? [Yes] We cited them in the experiments section.

    II Did you mention the license of the assets? [N/A] The codes that were used are publicly available.

    III Did you include any new assets either in the supplementary material or as a URL? [Yes] We provided the implementations of our algorithms in a supplemental material.

    IV Did you discuss whether and how consent was obtained from people whose data you're using/curating? [N/A]

    V Did you discuss whether the data you are using/curating contains personally identifiable information or offensive content? [N/A]

5. If you used crowdsourcing or conducted research with human subjects...

    I Did you include the full text of instructions given to participants and screenshots, if applicable? [N/A]

    II Did you describe any potential participant risks, with links to Institutional Review Board (IRB) approvals, if applicable? [N/A]

    III Did you include the estimated hourly wage paid to participants and the total amount spent on participant compensation? [N/A]

14

A Appendix

A.1 Proofs of Section 3.1

We first provide some lemmas and then prove Theorem 1.

Lemma 4. *Let $\mathbf{Y}_1, \ldots, \mathbf{Y}_N$ be centered symmetric random $d \times d$ matrices. Then*

$$\left(\mathbb{E} \left\| \sum_{i=1}^{n} \mathbf{Y}_i \right\|^p \right)^{\frac{1}{p}} \leq 2\sqrt{e \cdot \max\{p, 2\log(d)\}} \left(\mathbb{E} \left\| \sum_{i=1}^{n} \mathbf{Y}_i^2 \right\|^{p/2} \right)^{\frac{1}{p}} \tag{14}$$

The proof of Lemma 4 is given in Appendix A.1.2.

In our setting, we are going to bound $\mathbb{E}\left[\|\mathbf{H}_t - \nabla^2 F(\mathbf{x}_t)\|^{2\alpha}\right]$ for $\alpha \geq 1$ using Lemma 4

$$\mathbb{E}\|\mathbf{H}_t - \nabla^2 F(\mathbf{x})\|^{2\alpha} = \mathbb{E} \left\| \frac{1}{n_2} \sum_{i=1}^{n_2} (\nabla^2 f(\mathbf{x}, \xi_i) - \nabla^2 F(\mathbf{x})) \right\|^{2\alpha}$$

$$\leq 2^{2\alpha} (e \cdot \max\{2\alpha, \log d\})^{\alpha} \mathbb{E} \left\| \frac{1}{n_2^2} \sum_{i=1}^{n_2} (\nabla^2 f(\mathbf{x}, \xi_i) - \nabla^2 F(\mathbf{x}))^2 \right\|^{\alpha} \tag{15}$$

$$= \frac{2^{2\alpha} (e \cdot \max\{2\alpha, \log d\})^{\alpha}}{n_2^{\alpha}} \mathbb{E} \left\| \frac{1}{n_2} \sum_{i=1}^{n_2} (\nabla^2 f(\mathbf{x}, \xi_i) - \nabla^2 F(\mathbf{x}))^2 \right\|^{\alpha}$$

$$\leq \frac{2^{2\alpha} (e \cdot \max\{2\alpha, \log d\})^{\alpha}}{n_2^{\alpha}} \frac{1}{n_2} \sum_{i=1}^{n_2} \mathbb{E} \left\| (\nabla^2 f(\mathbf{x}, \xi_i) - \nabla^2 F(\mathbf{x}))^2 \right\|^{\alpha} \tag{16}$$

$$\leq \frac{2^{2\alpha} (e \cdot \max\{2\alpha, \log d\})^{\alpha}}{n_2^{\alpha}} \frac{1}{n_2} \sum_{i=1}^{n_2} \mathbb{E} \left\| (\nabla^2 f(\mathbf{x}, \xi_i) - \nabla^2 F(\mathbf{x})) \right\|^{2\alpha} \tag{17}$$

$$\leq \frac{2^{2\alpha} (e \cdot \max\{2\alpha, \log d\})^{\alpha}}{n_2^{\alpha}} \cdot \sigma_{2,\alpha}^2 \tag{18}$$

where (15) comes from Lemma 4 and (16) is derived by the convexity of $\|X\|^{\alpha}$. (17) is obtained by $\|AB\| \leq \|A\|\|B\|$ for the square matrices A, B. The last inequality is obtained from Assumption 3.

Lemma 5. *(i). Assume that function F satisfies Assumption 1 (Lipschitz Hessian). Then the solution of sub-problem, $\boldsymbol{\Delta}_t$, in Algorithm 1 (line 7), satisfies the following conditions:*

$$\|\nabla F(\mathbf{x}_t + \boldsymbol{\Delta}_t)\| \leq \frac{M + L_2}{2}\|\boldsymbol{\Delta}_t\|^2 + \|\nabla F(\mathbf{x}_t) - \mathbf{g}_t\| + \frac{\|\nabla^2 F(\mathbf{x}_t) - \mathbf{H}_t\|^2}{2} + \frac{\|\boldsymbol{\Delta}_t\|^2}{2}, \tag{19}$$

$$\frac{3M - 2L_2 - 8}{12}\|\boldsymbol{\Delta}_\mathbf{t}\|^3 \leq F(\mathbf{x}_t) - F(\mathbf{x}_t + \boldsymbol{\Delta}_t) + \frac{2\|\nabla F(\mathbf{x}_t) - \mathbf{g}_t\|^{3/2}}{3} + \frac{\|\nabla^2 F(\mathbf{x}_t) - \mathbf{H}_t\|^3}{6}. \tag{20}$$

(ii). With the same assumption in (i) and using Assumption 2 (gradient dominance property) for $\alpha \in [1, \infty)$, we have

$$F(\mathbf{x}_t + \boldsymbol{\Delta}_t) - F(\mathbf{x}^*) \leq$$
$$C(F(\mathbf{x}_t) - F(\mathbf{x}_t + \boldsymbol{\Delta}_t))^{2\alpha/3} + C_g\|\nabla F(\mathbf{x}_t) - \mathbf{g}_t\|^{\alpha} + C_H\|\nabla^2 F(\mathbf{x}_t) - \mathbf{H}_t\|^{2\alpha}, \tag{21}$$

where $C, C_g, C_H > 0$ are some constants depending on M, L_2, and τ_F.

Proof. Proof of (19): We know from Lipschitzness of Hessian of F that

$$\|\nabla F(\mathbf{x}_t + \boldsymbol{\Delta}_t) - \nabla F(\mathbf{x}_t) - \nabla^2 F(\mathbf{x}_t)\boldsymbol{\Delta}_t\| \leq \frac{L_2}{2}\|\boldsymbol{\Delta}_t\|^2. \tag{22}$$

15

Moreover, $\boldsymbol{\Delta}_t$ is an optimal solution of the problem in Algorithm 1 (line 10) and therefore, it satisfies the optimality condition: $\mathbf{g}_t + \mathbf{H}_t\boldsymbol{\Delta}_t + \frac{M}{2}\|\boldsymbol{\Delta}_t\|\boldsymbol{\Delta}_t = 0$. Therefore, we have:

$$\|\mathbf{g}_t + \mathbf{H}_t\boldsymbol{\Delta}_t\| = \frac{M}{2}\|\boldsymbol{\Delta}_t\|^2. \tag{23}$$

By summing the above two equations and using triangle inequality, we have:

$$\|\nabla F(\mathbf{x}_t + \boldsymbol{\Delta}_t)\| \leq \frac{M + L_2}{2}\|\boldsymbol{\Delta}_t\|^2 + \|\nabla F(\mathbf{x}_t) - \mathbf{g}_t\| + \|(\nabla^2 F(\mathbf{x}_t) - \mathbf{H}_t)\boldsymbol{\Delta}_t\|$$

$$\leq \frac{M + L_2}{2}\|\boldsymbol{\Delta}_t\|^2 + \|\nabla F(\mathbf{x}_t) - \mathbf{g}_t\| + \|\nabla^2 F(\mathbf{x}_t) - \mathbf{H}_t\|\|\boldsymbol{\Delta}_t\| \tag{24}$$

$$\leq \frac{M + L_2}{2}\|\boldsymbol{\Delta}_t\|^2 + \|\nabla F(\mathbf{x}_t) - \mathbf{g}_t\| + \frac{\|\nabla^2 F(\mathbf{x}_t) - \mathbf{H}_t\|^2}{2} + \frac{\|\boldsymbol{\Delta}_t\|^2}{2}, \tag{25}$$

where the last inequality is due to Young's inequality.
Proof of (20): For the sub-problem, we have from [25] that

$$\mathbf{g}_t^T\boldsymbol{\Delta}_t + \boldsymbol{\Delta}_t^T\mathbf{H}_t\boldsymbol{\Delta}_t + \frac{M}{2}\|\boldsymbol{\Delta}_t\|^3 = 0, \quad \mathbf{H}_t + \frac{M\|\boldsymbol{\Delta}_t\|}{2}I_{d\times d} \succeq 0 \tag{26}$$

which yields

$$\mathbf{g}_t^T\boldsymbol{\Delta}_t \leq 0. \tag{27}$$

From Lipschitzness of Hessian, we have the following:

$$F(\mathbf{x}_t + \boldsymbol{\Delta}_t) \leq F(\mathbf{x}_t) + \langle \nabla F(\mathbf{x}_t), \boldsymbol{\Delta}_t \rangle + \frac{1}{2}\langle \boldsymbol{\Delta}_t, \nabla^2 F(\mathbf{x}_t)\boldsymbol{\Delta}_t \rangle + \frac{L_2}{6}\|\boldsymbol{\Delta}_t\|^3 \tag{28}$$

$$= F(\mathbf{x}_t) + \langle \nabla F(\mathbf{x}_t) - \mathbf{g}_t, \boldsymbol{\Delta}_t \rangle + \frac{1}{2}\langle \boldsymbol{\Delta}_t, (\nabla^2 F(\mathbf{x}_t) - \mathbf{H}_t)\boldsymbol{\Delta}_t \rangle + \mathbf{g}_t^T\boldsymbol{\Delta}_t + \frac{1}{2}\boldsymbol{\Delta}_t^T\mathbf{H}_t\boldsymbol{\Delta}_t + \frac{L_2}{6}\|\boldsymbol{\Delta}_t\|^3$$

$$\overset{(a)}{\leq} F(\mathbf{x}_t) + \langle \nabla F(\mathbf{x}_t) - \mathbf{g}_t, \boldsymbol{\Delta}_t \rangle + \frac{1}{2}\langle \boldsymbol{\Delta}_t, (\nabla^2 F(\mathbf{x}_t) - \mathbf{H}_t)\boldsymbol{\Delta}_t \rangle + \frac{2L_2 - 3M}{12}\|\boldsymbol{\Delta}_t\|^3 \tag{29}$$

$$\overset{(b)}{\leq} F(\mathbf{x}_t) + \|\nabla F(\mathbf{x}_t) - \mathbf{g}_t\|_2\|\boldsymbol{\Delta}_t\|_2 + \frac{1}{2}\|\nabla^2 F(\mathbf{x}_t) - \mathbf{H}_t\|\|\boldsymbol{\Delta}_t\|_2^2 + \frac{2L_2 - 3M}{12}\|\boldsymbol{\Delta}_t\|^3 \tag{30}$$

$$\overset{(c)}{\leq} F(\mathbf{x}_t) + \frac{2\|\nabla F(\mathbf{x}_t) - \mathbf{g}_t\|^{3/2}}{3} + \frac{\|\nabla^2 F(\mathbf{x}_t) - \mathbf{H}_t\|^3}{6} + \frac{8 + 2L_2 - 3M}{12}\|\boldsymbol{\Delta}_t\|^3, \tag{31}$$

where (a) comes from (26) and (27), (b) comes from Cauchy-Schwartz inequality and (c) is due to Young's inequality.
Proof of (21): Based on gradient dominance property and (19), we have:

$$(F(\mathbf{x}_t + \boldsymbol{\Delta}_t) - F(\mathbf{x}^*)) \leq \tau_F\|\nabla F(\mathbf{x}_t + \boldsymbol{\Delta}_t)\|^\alpha$$

$$\leq \tau_F\Big(\frac{M + L_2 + 1}{2}\|\boldsymbol{\Delta}_t\|^2 + \|\nabla F(\mathbf{x}_t) - \mathbf{g}_t\| + \frac{\|\nabla^2 F(\mathbf{x}_t) - \mathbf{H}_t\|^2}{2}\Big)^\alpha \tag{32}$$

$$\leq 3^{\alpha-1}\tau_F\Big[\Big(\frac{M + L_2 + 1}{2}\Big)^\alpha\|\boldsymbol{\Delta}_t\|^{2\alpha} + \|\nabla F(\mathbf{x}_t) - \mathbf{g}_t\|^\alpha + \frac{\|\nabla^2 F(\mathbf{x}_t) - \mathbf{H}_t\|^{2\alpha}}{2^\alpha}\Big]$$

where we used $(a + b + c)^\alpha \leq 3^{\alpha-1}(a^\alpha + b^\alpha + c^\alpha), \forall a, b, c > 0$ in the last two above inequalities.
Substituting (20) into (32), we have:

$$F(\mathbf{x}_t + \boldsymbol{\Delta}_t) - F(\mathbf{x}^*) \leq$$

$$3^{\alpha-1}\tau_F\Big[\Big(\frac{M + L_2 + 1}{2}\Big)^\alpha\Big(\frac{12}{3M - 2L_2 - 8}\Big)^{2\alpha/3}\Big(F(\mathbf{x}_t) - F(\mathbf{x}_t + \boldsymbol{\Delta}_t) + \frac{2\|\nabla F(\mathbf{x}_t) - \mathbf{g}_t\|^{3/2}}{3} +$$

$$\frac{\|\nabla^2 F(\mathbf{x}_t) - \mathbf{H}_t\|^3}{6}\Big)^{2\alpha/3} + \|\nabla F(\mathbf{x}_t) - \mathbf{g}_t\|^\alpha + \frac{\|\nabla^2 F(\mathbf{x}_t) - \mathbf{H}_t\|^{2\alpha}}{2^\alpha}\Big] \tag{33}$$

$$\leq C(F(\mathbf{x}_t) - F(\mathbf{x}_t + \boldsymbol{\Delta}_t))^{2\alpha/3} + C_g\|\nabla F(\mathbf{x}_t) - \mathbf{g}_t\|^\alpha + C_H\|\nabla^2 F(\mathbf{x}_t) - \mathbf{H}_t\|^{2\alpha},$$

where in the last inequality we used $(a + b + c)^{2\alpha/3} \leq 3^{2\alpha/3 - 1/3}(a^{2\alpha/3} + b^{2\alpha/3} + c^{2\alpha/3})$ which is derived by the following inequalities:

$$(a + b + c)^{2\alpha} \leq 3^{2\alpha-1}(a^{2\alpha} + b^{2\alpha} + c^{2\alpha}) \leq 3^{2\alpha-1}(a^{2\alpha/3} + b^{2\alpha/3} + c^{2\alpha/3})^3$$

16

for any $a, b, c \in \mathbb{R}^+$ and,

$$C = 3^{(5\alpha-4)/3} \tau_F \left(\frac{M + L_2 + 1}{2}\right)^\alpha \left(\frac{12}{3M - 2L_2 - 8}\right)^{2\alpha/3}$$

$$C_g = 2^{2\alpha/3} \times 3^{\frac{5\alpha-7}{3}} \tau_F \left(\frac{M + L_2 + 1}{2}\right)^\alpha \left(\frac{12}{3M - 2L_2 - 8}\right)^{2\alpha/3} + 3^{\alpha-1} \tau_F \quad (34)$$

$$C_H = 2^{-2\alpha/3} \times 3^{\frac{5\alpha-7}{3}} \tau_F \left(\frac{M + L_2 + 1}{2}\right)^\alpha \left(\frac{12}{3M - 2L_2 - 8}\right)^{2\alpha/3} + 3^{\alpha-1} 2^{-\alpha} \tau_F$$

and the proof is complete. □

Theorem 1. *Let $F(\mathbf{x})$ satisfy Assumptions 1 and 2 for a given α and the stochastic gradient and Hessian satisfy Assumption 3 for the same α. Moreover, assume that an exact solver for sub-problem (3) exists. Then Algorithm 1 outputs a point $\mathbf{x}_T$ such that $\mathbb{E}[F(\mathbf{x}_T)] - F(\mathbf{x}^*) \le \epsilon$ after T iterations, where*

(i) *if $\alpha \in [1, 3/2)$, $T = \mathcal{O}(\epsilon^{-\frac{3-2\alpha}{2\alpha}})$, with access to the following numbers of samples of the stochastic gradient and Hessian per iteration:*

$$n_1 \ge \frac{C_g^{2/\alpha}}{C^{6/\alpha}} \cdot \frac{2^{2/\alpha} \sigma_1^{2/\alpha}}{\epsilon^{2/\alpha}}, \quad n_2 \ge \frac{{C'_H}^{1/\alpha}}{C^{3/\alpha}} \cdot \frac{2^{1/\alpha} \sigma_{2,\alpha}^{2/\alpha}}{\epsilon^{1/\alpha}}, \quad (35)$$

where C'_H is defined in (38) and depends on $\log(d)$.

(ii) *if $\alpha = 3/2$, $T = \mathcal{O}(\log(1/\epsilon))$ with the same numbers of samples per iteration as in (35).*

A.1.1 Proof of Theorem 1

Part (i): By Lemma 5, we have:

$$F(\mathbf{x}_t + \boldsymbol{\Delta}_t) - F(\mathbf{x}^*) \le C(F(\mathbf{x}_t) - F(\mathbf{x}_t + \boldsymbol{\Delta}_t))^{2\alpha/3} + C_g \|\nabla F(\mathbf{x}_t) - \mathbf{g}_t\|^\alpha + C_H \|\nabla^2 F(\mathbf{x}_t) - \mathbf{H}_t\|^{2\alpha}$$

Taking expectation of both sides given $\mathbf{x}_t$

$$\mathbb{E}[F(\mathbf{x}_t + \boldsymbol{\Delta}_t) - F(\mathbf{x}^*)] \le$$
$$C(\mathbb{E}F(\mathbf{x}_t) - \mathbb{E}F(\mathbf{x}_t + \boldsymbol{\Delta}_t))^{2\alpha/3} + C_g \mathbb{E}\|\nabla F(\mathbf{x}_t) - \mathbf{g}_t\|^\alpha + C_H \mathbb{E}\|\nabla^2 F(\mathbf{x}_t) - \mathbf{H}_t\|^{2\alpha}$$

$$\le C(\mathbb{E}F(\mathbf{x}_t) - \mathbb{E}F(\mathbf{x}_t + \boldsymbol{\Delta}_t))^{2\alpha/3} + C_g \frac{\sigma_1^\alpha}{n_1^{\alpha/2}} + C_H \mathbb{E}\|\nabla^2 F(\mathbf{x}_t) - \mathbf{H}_t\|^{2\alpha} \quad (36)$$

$$\le C(\mathbb{E}F(\mathbf{x}_t) - \mathbb{E}F(\mathbf{x}_t + \boldsymbol{\Delta}_t))^{2\alpha/3} + C_g \frac{\sigma_1^\alpha}{n_1^{\alpha/2}} + C_H \frac{2^{2\alpha}(e \cdot \max\{2\alpha, \log d\})^\alpha}{n_2^\alpha} \cdot \sigma_{2,\alpha}^2. \quad (37)$$

where (36) comes from the following facts that,

$$\mathbb{E}[\|\nabla F(\mathbf{x}_t) - \mathbf{g}_t\|^\alpha] \le (\mathbb{E}[\|\nabla F(\mathbf{x}_t) - \mathbf{g}_t\|^2])^{\alpha/2} \le \frac{\sigma_1^\alpha}{n_1^{\alpha/2}},$$

which is comes from Jensen's inequality. $\mathbb{E}[A^{2\alpha/3}] \le (\mathbb{E}[A])^{2\alpha/3}$. Inequality (37) is derived by Equation (18) for bounding the Hessian error. Let define

$$C'_H := C_H 2^{2\alpha}(e \cdot \max\{2\alpha, \log d\})^\alpha \quad (38)$$

and $P(n_1, n_2) := \frac{C_g \sigma_1}{n_1^{\alpha/2}} + \frac{C'_H \sigma_{2,\alpha}^2}{n_2^\alpha}$ and $\delta_t := \frac{\mathbb{E}[F(\mathbf{x}_t)] - F(\mathbf{x}^*)}{C^{3/(3-2\alpha)}} - \frac{P(n_1, n_2)}{C^{3/(3-2\alpha)}}$. Then rewrite Inequality (37) as follows,

$$\delta_{t+1} \le (\delta_t - \delta_{t+1})^{2\alpha/3}. \quad (39)$$

First we show that $\delta_t \to 0$ when $t \to \infty$ and then with good choices of batch sizes, $P(n_1, n_2) \to 0$. Let $h(t) := \frac{2\alpha}{3-2\alpha} t^{1-\frac{3}{2\alpha}}$. To show $\delta_t \to 0$, We have two cases:

17

Case (1): Assume that $\delta_{t+1} \geq 2^{-\frac{2\alpha}{3}} \delta_t$ for some t. Then

$$h(\delta_{t+1}) - h(\delta_t) = \int_{\delta_t}^{\delta_{t+1}} \frac{d}{dt} h(t) dt = \int_{\delta_{t+1}}^{\delta_t} t^{-\frac{3}{2\alpha}} dt \geq (\delta_t - \delta_{t+1}) \delta_t^{-\frac{3}{2\alpha}}$$

$$\geq (\delta_t - \delta_{t+1}) \frac{1}{2} \delta_{t+1}^{-\frac{3}{2\alpha}} \geq \frac{1}{2} \tag{40}$$

Case (2): If we have $\delta_{t+1} \leq 2^{-\frac{2\alpha}{3}} \delta_t$ for some $t \geq 0$, $\delta_{t+1}^{1-\frac{3}{2\alpha}} \geq 2^{\frac{3-2\alpha}{3}} \delta_t^{1-\frac{3}{2\alpha}}$.

$$h(\delta_{t+1}) - h(\delta_t) = \frac{2\alpha}{3-2\alpha} (\delta_{t+1}^{1-\frac{3}{2\alpha}} - \delta_t^{1-\frac{3}{2\alpha}}) \geq \frac{2\alpha}{3-2\alpha} (2^{\frac{3-2\alpha}{3}} - 1) \delta_t^{1-\frac{3}{2\alpha}}$$

$$\geq \frac{2\alpha}{3-2\alpha} (2^{\frac{3-2\alpha}{3}} - 1) \delta_0^{1-\frac{3}{2\alpha}}. \tag{41}$$

Let define $D := \min\{\frac{1}{2}, \frac{2\alpha}{3-2\alpha}(2^{\frac{3-2\alpha}{3}} - 1)\delta_0^{1-\frac{3}{2\alpha}}\}$. Then we have

$$h(\delta_{t+1}) - h(\delta_t) \geq D$$

which implies

$$h(\delta_T) \geq \sum_{t=1}^{T} h(\delta_t) - h(\delta_{t-1}) \geq D \cdot T.$$

Then we get

$$\delta_T \leq \left(\frac{2\alpha}{3-2\alpha}\right)^{\frac{2\alpha}{3-2\alpha}} \frac{1}{(DT)^{\frac{2\alpha}{3-2\alpha}}}. \tag{42}$$

Hence, $F(x_T) - F(x^*)$ converges to a stationary point $P(n_1, n_2)$ at a rate of $\mathcal{O}\left(\frac{1}{T^{\frac{2\alpha}{3-2\alpha}}}\right)$. We can choose $n_1 \geq \frac{4^{2/\alpha} C_g^{2/\alpha} \sigma_1^{2/\alpha}}{C^{\frac{2}{\alpha} \cdot \frac{3}{3-2\alpha}} \epsilon^{2/\alpha}}$, and $n_2 \geq \frac{4^{1/\alpha} C_H'^{1/\alpha} \sigma_{2,\alpha}^{2/\alpha}}{C^{\frac{1}{\alpha} \cdot \frac{3}{3-2\alpha}} \epsilon^{1/\alpha}}$ to have $P(n_1, n_2) \leq C^{3/(3-2\alpha)} \frac{\epsilon}{2}$. With $T \geq \frac{2\alpha}{3-2\alpha} \frac{1}{D(\epsilon/2)^{\frac{3-2\alpha}{2\alpha}}}$, we have $\delta_T \leq \epsilon/2$ and then $\mathbb{E}[F(\mathbf{x}_T)] - F(x^*) \leq C^{\frac{3}{3-2\alpha}} \epsilon$. Finally, the total sample complexity for having $\mathbb{E}[F(\mathbf{x}_T)] - F(x^*) \leq \epsilon$ would be

$$T \cdot (n_1 + n_2) = \left(\frac{2\alpha}{3-2\alpha} D \frac{C^{3/(2\alpha)}}{(\epsilon/2)^{\frac{3-2\alpha}{2\alpha}}}\right) \left(\frac{4^{2/\alpha} C_g^{2/\alpha} \sigma_1^{2/\alpha}}{\epsilon^{2/\alpha}} + \frac{4^{1/\alpha} C_H'^{1/\alpha} \sigma_{2,\alpha}^{2/\alpha}}{\epsilon^{1/\alpha}}\right) = \mathcal{O}\left(\frac{1}{\epsilon^{\frac{7}{2\alpha}-1}}\right).$$

Part (ii): By Lemma 5, we have:

$$F(\mathbf{x}_t + \boldsymbol{\Delta}_t) - F(\mathbf{x}^*) \leq C(F(\mathbf{x}_t) - F(\mathbf{x}_t + \boldsymbol{\Delta}_t)) + C_g \|\nabla F(\mathbf{x}_t) - \mathbf{g}_t\|^{3/2} + C_H \|\nabla^2 F(\mathbf{x}_t) - \mathbf{H}_t\|^3. \tag{43}$$

Taking expectation of both sides given $\mathbf{x}_t$ and using the same arguments as in (36) and (37), we get

$$\mathbb{E}[F(\mathbf{x}_t + \boldsymbol{\Delta}_t) - F(\mathbf{x}^*)] \leq C(\mathbb{E}F(\mathbf{x}_t) - \mathbb{E}F(\mathbf{x}_t + \boldsymbol{\Delta}_t)) + C_g \frac{\sigma_1^\alpha}{n_1^{3/4}} + C_H' \frac{\sigma_{2,1.5}^2}{n_2^{3/2}}. \tag{44}$$

Let define $\delta_t := \mathbb{E}[F(\mathbf{x}_t)] - F(\mathbf{x}^*) - C_g \frac{\sigma_1}{n_1^{3/4}} - C_H' \frac{\sigma_{2,1.5}^2}{n_2^{3/2}}$ and rewrite the above inequality as follows,

$$\delta_{t+1} \leq C(\delta_t - \delta_{t+1}). \tag{45}$$

We get $\delta_T \leq \left(\frac{C}{C+1}\right)^T \delta_0$. Thus, $F(x_T) - F(x^*)$ converges to a stationary point

$$C_g \frac{\sigma_1}{n_1^{3/4}} + C_H' \frac{\sigma_{2,1.5}^2}{n_2^{3/2}}$$

at a rate of $\left(\frac{C}{C+1}\right)^T \cdot \delta_0$. In order to have $\mathbb{E}[F(\mathbf{x}_T)] - F(\mathbf{x}^*) \leq \epsilon$, we can choose $T \geq \frac{\log(2\delta_0/\epsilon)}{\log\left(\frac{C+1}{C}\right)}$, $n_1 \geq \frac{C_g^{4/3}}{C^4} \frac{4^{4/3} \sigma_1^{4/3}}{\epsilon^{4/3}}$, and $n_2 \geq \frac{C_H'^{2/3}}{C^2} \frac{4^{2/3} \sigma_{2,1.5}^{4/3}}{\epsilon^{2/3}}$. Finally, the total sample complexity would be

$$T \cdot (n_1 + n_2) = \frac{\log\left(\frac{2\delta_0}{\epsilon}\right)}{\log\left(\frac{C+1}{C}\right)} \cdot \left(\frac{C_g^{4/3}}{C^4} \frac{4^{4/3} \sigma_1^{4/3}}{\epsilon^{4/3}} + \frac{C_H'^{2/3}}{C^2} \frac{4^{2/3} \sigma_{2,1.5}^{4/3}}{\epsilon^{2/3}}\right).$$

18

A.1.2 Proof of Lemma 4

The proof is mainly adapted from the symmetrization argument [34, 7]. Consider $\mathbf{Y}'_i$ as an independent copy of $\mathbf{Y}_i$ for $i = 1, \ldots, n$. Then

$$\mathbb{E}\left\|\sum_{i=1}^n \mathbf{Y}_i\right\|^p = \mathbb{E}\left\|\sum_{i=1}^n \mathbb{E}_{\mathbf{Y}'_i}(\mathbf{Y}_i - \mathbf{Y}'_i)\right\|^p$$

$$\leq \mathbb{E}_{\mathbf{Y}}\mathbb{E}_{\mathbf{Y}'}\left\|\sum_{i=1}^n (\mathbf{Y}_i - \mathbf{Y}'_i)\right\|^p = \mathbb{E}\left\|\sum_{i=1}^n \epsilon_i(\mathbf{Y}_i - \mathbf{Y}'_i)\right\|^p \tag{46}$$

$$\leq \mathbb{E}\left[2^{p-1}\left(\left\|\sum_{i=1}^n \epsilon_i \mathbf{Y}_i\right\|^p + \left\|\sum_{i=1}^n \epsilon_i \mathbf{Y}'_i\right\|^p\right)\right] = 2^p \mathbb{E}\left\|\sum_{i=1}^n \epsilon_i \mathbf{Y}_i\right\|^p \tag{47}$$

where (46) comes from the fact that $\mathbf{Y}_i - \mathbf{Y}'_i$ has the same distribution as $\mathbf{Y}'_i - \mathbf{Y}_i$ and ϵ_i's are Rademacher random variables (i.e. $\mathbb{P}[\epsilon_i = -1] = \mathbb{P}[\epsilon_i = 1] = \frac{1}{2}$). (47) comes from the inequalities $\|\mathbf{A} - \mathbf{H}\| \leq \|\mathbf{A}\| + \|\mathbf{H}\|$ and $\left(\frac{a+b}{2}\right)^q \leq \frac{a^q + b^q}{2}$. Hence,

$$\left[\mathbb{E}\left\|\sum_{i=1}^n \mathbf{Y}_i\right\|^p\right]^{\frac{1}{p}} \leq 2\left[\mathbb{E}\left\|\sum_{i=1}^n \epsilon_i \mathbf{Y}_i\right\|^p\right]^{\frac{1}{p}}$$

Let define the Schatten p-norm of a matrix $\mathbf{A}$ as $\|\mathbf{A}\|_p := \left(\mathrm{Tr}[(\mathbf{A}^T\mathbf{A})^{p/2}]\right)^{\frac{1}{p}} := \left(\sum_{i \geq 1} s_i^p(A)\right)^{\frac{1}{p}}$ where $s_i(A)$'s are the singular value of $\mathbf{A}$. Schatten ∞-norm of a matrix is its operator norm by the definition ($\|\mathbf{A}\| = \|\mathbf{A}\|_\infty$). Then for $q \geq p$, we have

$$\left[\mathbb{E}\left\|\sum_{i=1}^n \mathbf{Y}_i\right\|^p\right]^{\frac{1}{p}} \leq 2\left[\mathbb{E}\left\|\sum_{i=1}^n \epsilon_i \mathbf{Y}_i\right\|^p\right]^{\frac{1}{p}}$$

$$\leq 2\left[\mathbb{E}\left\|\sum_{i=1}^n \epsilon_i \mathbf{Y}_i\right\|_q^p\right]^{\frac{1}{p}} \leq 2\left[\mathbb{E}_{\mathbf{Y}}\left(\mathbb{E}_{\boldsymbol{\epsilon}}\left\|\sum_{i=1}^n \epsilon_i \mathbf{Y}_i\right\|_q^q\right)^{p/q}\right]^{\frac{1}{p}} \tag{48}$$

where (48) comes from the fact that $(\mathbb{E}\|X\|^p)^{1/p} \leq (\mathbb{E}\|X\|^q)^{1/q}$ for $q \geq p$.

The matrix Khintchine inequality is as follows:

Lemma 6. *[21] Suppose $q > 2$ and consider the deterministic, symmetric matrices $\mathbf{A}_i$, $1 \leq i \leq n$. Then*

$$\left(\mathbb{E}_{\boldsymbol{\epsilon}}\left\|\sum_{i=1}^n \epsilon_i \mathbf{A}_i\right\|_q^q\right)^{1/q} \leq \sqrt{q}\left\|\left[\sum_{i=1}^n \mathbf{A}_i^2\right]^{1/2}\right\|_q$$

Assume that $q = \max\{p, 2\log d\} \geq p$. From (48),

$$\left[\mathbb{E}\left\|\sum_{i=1}^n \mathbf{Y}_i\right\|^p\right]^{\frac{1}{p}} \leq 2\left[\mathbb{E}_{\mathbf{Y}}\left(\mathbb{E}_{\boldsymbol{\epsilon}}\left\|\sum_{i=1}^n \epsilon_i \mathbf{Y}_i\right\|_q^q\right)^{p/q}\right]^{\frac{1}{p}}$$

$$\leq 2\sqrt{q}\left[\mathbb{E}_{\mathbf{Y}}\left\|\left(\sum_{i=1}^n \mathbf{Y}_i^2\right)^{1/2}\right\|_q^p\right]^{\frac{1}{p}} \tag{49}$$

$$\leq 2\sqrt{q}\left[\mathbb{E}_{\mathbf{Y}}\left(d^{\frac{1}{q}}\left\|\left(\sum_{i=1}^n \mathbf{Y}_i^2\right)^{1/2}\right\|\right)^p\right]^{\frac{1}{p}} \tag{50}$$

19

$$\leq 2\sqrt{eq} \left[\mathbb{E}_{\mathbf{Y}} \left(\left\| \left(\sum_{i=1}^{n} \mathbf{Y}_i^2 \right)^{1/2} \right\|^p \right) \right]^{\frac{1}{p}} \tag{51}$$

$$= 2\sqrt{eq} \left[\mathbb{E}_{\mathbf{Y}} \left(\left\| \sum_{i=1}^{n} \mathbf{Y}_i^2 \right\|^{p/2} \right) \right]^{\frac{1}{p}} \tag{52}$$

where (49) is derived by Lemma 6 and (50) comes from $\|\mathbf{A}\|_q \leq d^{\frac{1}{q}} \|A\|$. (51) is due to $d^{1/q} \leq d^{\frac{1}{2\log d}} \leq \sqrt{e}$ and (52) follows from $\|\mathbf{A}\|^2 = \lambda_{\max}(\mathbf{A}^T\mathbf{A}) = \lambda_{\max}(\mathbf{A}^2) = \|\mathbf{A}^2\|$ when $\mathbf{A}$ is positive semi-definite ($\mathbf{A} \succeq 0$).

A.2 Proofs of Section 3.2

We begin by providing some lemmas that will be used in the main proof.

Lemma 7. *Under Assumptions 3 and 4, we can adjust gradient and Hessian mini-batch sizes*

$$n_1 \geq \frac{8}{3} \max \left(\frac{M_1}{\epsilon}, \frac{\sigma_1^2}{\epsilon^2} \right) \log \frac{2d}{\delta},$$

$$n_2 \geq \frac{8}{3} \max \left(\frac{M_2}{\sqrt{\epsilon}}, \frac{\sigma_2^2}{\epsilon} \right) \log \frac{2d}{\delta},$$

such that with probability at least $1 - \delta$,

$$\|\mathbf{g}_t - \nabla F(\mathbf{x}_t)\| \leq \epsilon, \tag{53}$$

$$\|\mathbf{H}_t - \nabla^2 F(\mathbf{x}_t)\| \leq \sqrt{\epsilon}. \tag{54}$$

Proof of Lemma 7. Recall $\mathbf{g}(\mathbf{x}, \xi) = \nabla f(\mathbf{x}, \xi) - \nabla F(\mathbf{x})$, $\mathbf{G}(\mathbf{x}, \xi) = \begin{bmatrix} \mathbf{0}_{1\times1} & \mathbf{g}(\mathbf{x}, \xi)^T \\ \mathbf{g}(\mathbf{x}, \xi) & \mathbf{0}_{d\times d} \end{bmatrix}$, and $\mathbf{H}(\mathbf{x}, \xi) = \nabla^2 f(\mathbf{x}, \xi) - \nabla^2 F(\mathbf{x})$.

We use the matrix Bernstein's inequality from [34, Theorem 6.17] to control the estimation error in the stochastic gradients and stochastic Hessians under Assumptions 3 and 4. For a given $\mathbf{x}_t$,

$$\mathbb{P}\left[\left\| \frac{1}{n_1} \sum_{i=1}^{n_1} \mathbf{G}(\mathbf{x}_t, \xi_i) \right\| \geq t \right] \leq 2d \exp\left(\frac{-n_1 t^2}{2(\sigma_g^2 + M_1 t)} \right) \leq 2d \exp\left(-\frac{n_1}{8} \min\left\{ \frac{t}{M_1}, \frac{t^2}{\sigma_g^2} \right\} \right), \tag{55}$$

where M_1 is the parameter of Bernstein's condition for $\mathbf{G}(\mathbf{x}_t, \xi_i)$ in Assumption 4 and

$$\sigma_g^2 := \max \left\{ \left\| \frac{1}{n_1} \sum_{i=1}^{n_1} \mathbb{E}[\mathbf{g}(\mathbf{x}_t, \xi_i)\mathbf{g}(\mathbf{x}_t, \xi_i)^T] \right\|_2, \left\| \frac{1}{n_1} \sum_{i=1}^{n_1} \mathbb{E}[\mathbf{g}(\mathbf{x}_t, \xi_i)^T\mathbf{g}(\mathbf{x}_t, \xi_i)] \right\|_2 \right\}.$$

Note that $\sigma_g^2 \leq \sigma_1^2$ where σ_1^2 is the variance parameter in Assumption 3.

$$\mathbb{P}\left[\left\| \frac{1}{n_2} \sum_{i=1}^{n_2} \mathbf{H}(\mathbf{x}_t, \xi_i) \right\| \geq t \right] \leq 2d \exp\left(\frac{-n_2 t^2}{2(\sigma_2^2 + M_2 t)} \right) \leq 2d \exp\left(-\frac{n_2}{8} \min\left\{ \frac{t}{M_2}, \frac{t^2}{\sigma_2^2} \right\} \right). \tag{56}$$

where M_2 is the parameter of Bernstein's condition for $\mathbf{H}(\mathbf{x}_t, \xi_i)$ in Assumption 4 and

$$\sigma_H^2 := \left\| \frac{1}{n_1} \sum_{i=1}^{n_1} \mathbb{E}[\mathbf{H}^2(\mathbf{x}_t, \xi_i)] \right\|.$$

Note that $\sigma_H^2 \leq \sigma_2^2$ where σ_2^2 is the parameter for the bounded variance condition 3. First, we claim that for every $\mathbf{v} \in \mathbb{R}^d$ we have

$$\|\mathbf{v}\|_2 = \left\| \begin{bmatrix} \mathbf{0}_{1\times1} & \mathbf{v}^T \\ \mathbf{v} & \mathbf{0}_{d\times d} \end{bmatrix} \right\| \tag{57}$$

20

Using Inequality (57), we can obtain $\mathbf{g}_t - \nabla F(\mathbf{x}_t) \leq \frac{1}{n_1} \sum_{i=1}^{n} \mathbf{G}(\mathbf{x}_t, \xi_i)$. Note that $\mathbf{H}_t - \nabla^2 F(\mathbf{x}_t) = \frac{1}{n_2} \sum_{i=1}^{n} \mathbf{H}(\mathbf{x}_t, \xi_i)$. Hence,

$$\mathbb{P}\left[\|\mathbf{g}_t - \nabla F(\mathbf{x}_t)\| \geq t\right] \leq 2d \exp\left(-\frac{n_1}{8} \min\left\{\frac{t}{M_1}, \frac{t^2}{\sigma_1^2}\right\}\right)$$

$$\mathbb{P}\left[\|\mathbf{H}_t - \nabla^2 F(\mathbf{x}_t)\| \geq t\right] \leq 2d \exp\left(-\frac{n_2}{8} \min\left\{\frac{t}{M_2}, \frac{t^2}{\sigma_2^2}\right\}\right)$$

In other words, for

$$n_1 \geq 8 \max\left(\frac{M_1}{\epsilon}, \frac{\sigma_1^2}{\epsilon^2}\right) \log \frac{2d}{\delta}, \tag{58}$$

$$n_2 \geq 8 \max\left(\frac{M_2}{\sqrt{\epsilon}}, \frac{\sigma_2^2}{\epsilon}\right) \log \frac{2d}{\delta}, \tag{59}$$

we have:

$$\|\mathbf{g}_t - \nabla F(\mathbf{x}_t)\| \leq_{1-\delta} \epsilon, \tag{60}$$

$$\|\mathbf{H}_t - \nabla^2 F(\mathbf{x}_t)\| \leq_{1-\delta} \sqrt{\epsilon}. \tag{61}$$

The claims of Lemma 7 are established.

Proof of (57): For every symmetric matrix $\mathbf{A}$ we have $|\lambda_{\max}(\mathbf{A})| = \sigma_{\max}(\mathbf{A})$. For matrix $\mathbf{A} = \begin{bmatrix} \mathbf{0}_{1\times 1} & \mathbf{v}^T \\ \mathbf{v} & \mathbf{0}_{d\times d} \end{bmatrix}$ we have $\det(\lambda I - \mathbf{A}) = \lambda^d - \lambda^{d-2}\|\mathbf{v}\|^2 = 0$. Thus, $\lambda_{\max} = \|\mathbf{v}\|$.

$\square$

Lemma 8. *Under Assumptions 1, 3, and 4 and for*

$$n_1 \geq 8 \max\left(\frac{M_1}{\epsilon^{1/\alpha}}, \frac{\sigma_1^2}{\epsilon^{2/\alpha}}\right) \log \frac{2d}{\delta}, \tag{62}$$

$$n_2 \geq 8 \max\left(\frac{M_2}{\epsilon^{1/(2\alpha)}}, \frac{\sigma_2^2}{\epsilon^{1/\alpha}}\right) \log \frac{2d}{\delta} \tag{63}$$

where $\alpha \geq 0$, we have for Algorithm 1:

$$\|\nabla F(\mathbf{x}_{t+1})\| \leq_{1-2\delta} \frac{3}{2}\epsilon^{1/\alpha} + \frac{M + L_2 + 1}{2}\|\mathbf{\Delta}_t\|^2. \tag{64}$$

Proof of Lemma 8. We know from Lipschitzness of Hessian of F that

$$\|\nabla F(\mathbf{x}_{t+1}) - \nabla F(\mathbf{x}_t) - \nabla^2 F(\mathbf{x}_t)(\mathbf{x}_{t+1} - \mathbf{x}_t)\| \leq \frac{L_2}{2}\|\mathbf{x}_{t+1} - \mathbf{x}_t\|^2. \tag{65}$$

Moreover, $\mathbf{\Delta}_t$ is an optimal solution of the problem in (3) and therefore, it satisfies the optimality condition: $\mathbf{g}_t + \mathbf{H}_t\mathbf{\Delta}_t + \frac{M}{2}\|\mathbf{\Delta}_t\|\mathbf{\Delta}_t = 0$. Therefore, we have:

$$\|\mathbf{g}_t + \mathbf{H}_t\mathbf{\Delta}_t\| = \frac{M}{2}\|\mathbf{\Delta}_t\|^2. \tag{66}$$

By summing the above two equations and using triangle inequality, we have:

$$\|\nabla F(\mathbf{x}_{t+1})\| \leq \frac{M + L_2}{2}\|\mathbf{\Delta}_t\|^2 + \|\nabla F(\mathbf{x}_t) - \mathbf{g}_t\| + \|(\nabla^2 F(\mathbf{x}_t) - \mathbf{H}_t)(\mathbf{x}_{t+1} - \mathbf{x}_t)\|$$

$$\leq \frac{M + L_2}{2}\|\mathbf{\Delta}_t\|^2 + \|\nabla F(\mathbf{x}_t) - \mathbf{g}_t\| + \|\nabla^2 F(\mathbf{x}_t) - \mathbf{H}_t\|\|\mathbf{\Delta}_t\| \tag{67}$$

$$\overset{(a)}{\leq}_{1-2\delta} \frac{M + L_2}{2}\|\mathbf{\Delta}_t\|^2 + \epsilon^{1/\alpha} + \epsilon^{1/(2\alpha)}\|\mathbf{\Delta}_t\| \tag{68}$$

$$\overset{(b)}{\leq}_{1-2\delta} \frac{M + L_2 + 1}{2}\|\mathbf{\Delta}_t\|^2 + \frac{3}{2}\epsilon^{1/\alpha}. \tag{69}$$

(a) Due to Lemma 7 if we consider the batch sizes considered in (62) and (63).
(b) According to Young's inequality, we have: $ab \leq \frac{a^2 + b^2}{2}$. $\square$

21

Lemma 9. *Under the assumptions of Lemma 8, we have*

$$F(\mathbf{x}_{t+1}) \leq_{1-2\delta} F(\mathbf{x}_t) + \epsilon^{1/\alpha}\|\mathbf{\Delta}_t\|_2 + \frac{1}{2}\epsilon^{1/(2\alpha)}\|\mathbf{\Delta}_t\|_2^2 + \frac{2L_2 - 3M}{12}\|\mathbf{\Delta}_t\|^3. \tag{70}$$

Proof of Lemma 9. For the sub-problem mentioned in (3), we have from [25, Proposition 1] that

$$\mathbf{g}_t^T\mathbf{\Delta}_t + \mathbf{\Delta}_t^T\mathbf{H}_t\mathbf{\Delta}_t + \frac{M}{2}\|\mathbf{\Delta}_t\|^3 = 0, \quad \mathbf{H}_t + \frac{M\|\mathbf{\Delta}_t\|}{2}I_{d\times d} \succeq 0 \tag{71}$$

which yields

$$\mathbf{g}_t^T\mathbf{\Delta}_t \leq 0. \tag{72}$$

From Lipschitzness of Hessian, we have the following:

$$F(\mathbf{x}_{t+1}) \leq F(\mathbf{x}_t) + \langle \nabla F(\mathbf{x}_t), \mathbf{\Delta}_t \rangle + \frac{1}{2}\langle \mathbf{\Delta}_t, \nabla^2 F(\mathbf{x}_t)\mathbf{\Delta}_t \rangle + \frac{L_2}{6}\|\mathbf{\Delta}_t\|^3 \tag{73}$$

$$= F(\mathbf{x}_t) + \langle \nabla F(\mathbf{x}_t) - \mathbf{g}_t, \mathbf{\Delta}_t \rangle + \frac{1}{2}\langle \mathbf{\Delta}_t, (\nabla^2 F(\mathbf{x}_t) - \mathbf{H}_t)\mathbf{\Delta}_t \rangle + \mathbf{g}_t^T\mathbf{\Delta}_t + \frac{1}{2}\mathbf{\Delta}_t^T\mathbf{H}_t\mathbf{\Delta}_t + \frac{L_2}{6}\|\mathbf{\Delta}_t\|^3$$

$$\overset{(a)}{\leq} F(\mathbf{x}_t) + \langle \nabla F(\mathbf{x}_t) - \mathbf{g}_t, \mathbf{\Delta}_t \rangle + \frac{1}{2}\langle \mathbf{\Delta}_t, (\nabla^2 F(\mathbf{x}_t) - \mathbf{H}_t)\mathbf{\Delta}_t \rangle + \frac{2L_2 - 3M}{12}\|\mathbf{\Delta}_t\|^3 \tag{74}$$

$$\overset{(b)}{\leq} F(\mathbf{x}_t) + \|\nabla F(\mathbf{x}_t) - \mathbf{g}_t\|_2\|\mathbf{\Delta}_t\|_2 + \frac{1}{2}\|\nabla^2 F(\mathbf{x}_t) - \mathbf{H}_t\|\|\mathbf{\Delta}_t\|_2^2 + \frac{2L_2 - 3M}{12}\|\mathbf{\Delta}_t\|^3 \tag{75}$$

$$\overset{(c)}{\leq}_{1-2\delta} F(\mathbf{x}_t) + \epsilon^{1/\alpha}\|\mathbf{\Delta}_t\|_2 + \frac{1}{2}\epsilon^{1/(2\alpha)}\|\mathbf{\Delta}_t\|_2^2 + \frac{2L_2 - 3M}{12}\|\mathbf{\Delta}_t\|^3 \tag{76}$$

where (a) comes from (72) and (71). Inequality (b) comes from Cauchy-Schwartz inequality and (c) is derived by Lemma 7. □

Theorem 2. *When $F(\mathbf{x})$ satisfies Assumptions 1, 2, the stochastic gradient and Hessian satisfy Assumption 3 (with $\alpha = 1$) and Assumption 4, and there exists an exact solver for sub-problem (3), Algorithm 1, with probability $1 - \delta$, outputs a solution $\mathbf{x}_T$ such that $F(\mathbf{x}_T) - F(\mathbf{x}^*) \leq \epsilon$ after $T = \mathcal{O}\left(\log\left(\log(1/\epsilon)\right)\right)$ iterations with the following numbers of samples for the stochastic gradient and Hessian, respectively per iteration:*

$$n_1 \geq \frac{8}{3}\max\left(\frac{\tilde{C}^{1/\alpha}M_1}{\epsilon^{1/\alpha}}, \frac{\tilde{C}^{2/\alpha}\sigma_1^2}{\epsilon^{2/\alpha}}\right)\log\left(\frac{4(T+1)d}{\delta}\right), \tag{77}$$

$$n_2 \geq \frac{8}{3}\max\left(\frac{\tilde{C}^{1/(2\alpha)}M_2}{\epsilon^{1/(2\alpha)}}, \frac{\tilde{C}^{1/\alpha}\sigma_{2,1}^2}{\epsilon^{1/\alpha}}\right)\log\left(\frac{4(T+1)d}{\delta}\right), \tag{78}$$

where $\tilde{C} = 1 + \frac{\tau_F}{2}\left(\frac{M+L_2+4}{2}\right)^\alpha$.

A.2.1 Proof of Theorem 2

Now, we are ready to prove the main theorem. Suppose that $t = N_0$ is the first time that $\|\mathbf{\Delta}_t\| \leq \sqrt[2\alpha]{\epsilon}$. In other words, for $t = 0, \ldots, N_0 - 1$, we have $\|\mathbf{\Delta}_t\| > \sqrt[2\alpha]{\epsilon}$. The following analysis is valid for $t = 0, \ldots, N_0 - 1$. From Lemma 8, for

$$n_1 \geq 8\max\left(\frac{M_1}{\epsilon^{1/\alpha}}, \frac{\sigma_1^2}{\epsilon^{2/\alpha}}\right)\log\frac{2d}{\delta} \tag{79}$$

$$n_2 \geq 8\max\left(\frac{M_2}{\epsilon^{1/(2\alpha)}}, \frac{\sigma_2^2}{\epsilon^{1/\alpha}}\right)\log\frac{2d}{\delta}, \tag{80}$$

we have:

$$\|\nabla F(\mathbf{x}_{t+1})\| \leq_{1-2\delta} \frac{3}{2}\sqrt[\alpha]{\epsilon} + \frac{M+L_2+1}{2}\|\mathbf{\Delta}_t\|^2 \tag{81}$$

$$\leq_{1-2\delta} \frac{3}{2}\|\mathbf{\Delta}_t\|^2 + \frac{M+L_2+1}{2}\|\mathbf{\Delta}_t\|^2. \tag{82}$$

22

Let $A := \frac{3}{2} + \frac{M+L_2+1}{2}$. Then, we can rewrite the above inequality in the following form:

$$\frac{1}{A}\|\nabla F(\mathbf{x}_{t+1})\| \leq_{1-2\delta} \|\boldsymbol{\Delta}_t\|^2. \tag{83}$$

Using Lemma 9, we have:

$$F(\mathbf{x}_{t+1}) \leq_{1-2\delta} F(\mathbf{x}_t) + \epsilon^{1/\alpha}\|\boldsymbol{\Delta}_t\|_2 + \frac{1}{2}\epsilon^{1/(2\alpha)}\|\boldsymbol{\Delta}_t\|_2^2 + \frac{2L_2 - 3M}{12}\|\boldsymbol{\Delta}_t\|^3 \tag{84}$$

$$\leq_{1-2\delta} F(\mathbf{x}_t) - \left(\frac{3M - 2L_2}{12} - \frac{3}{2}\right)\|\boldsymbol{\Delta}_t\|^3 \tag{85}$$

$$\leq_{1-4\delta} F(\mathbf{x}_t) - \left(\frac{3M - 2L_2}{12} - \frac{3}{2}\right)\frac{1}{A^{3/2}}\|\nabla F(\mathbf{x}_{t+1})\|^{3/2}, \tag{86}$$

where the last inequality is according to (83). Let $B := \left(\frac{3M-2L_2}{12} - \frac{3}{2}\right)/A^{3/2}$. Using gradient dominance property, we have:

$$(F(\mathbf{x}_{t+1}) - F(\mathbf{x}^*)) \leq_{1-4\delta} \tau_F \frac{(F(\mathbf{x}_t) - F(\mathbf{x}_{t+1}))^{2\alpha/3}}{B^{2\alpha/3}}. \tag{87}$$

Let $h_t := \frac{\tau_F^{\frac{3}{2\alpha-3}}}{B^{\frac{2\alpha}{2\alpha-3}}}[F(\mathbf{x}_t) - F(\mathbf{x}^*)]$. Then,

$$h_{t+1}^{3/2\alpha} \leq_{1-4\delta} h_t - h_{t+1} \tag{88}$$

We define $K_{N_0} := \max_{0 \leq t \leq N_0-1} h_t$. Therefore,

$$1 + \frac{1}{K_{N_0}^{\frac{2\alpha-3}{2\alpha}}} \leq 1 + \frac{1}{h_{t+1}^{\frac{2\alpha-3}{2\alpha}}} \leq_{1-4\delta} \frac{h_t}{h_{t+1}}.$$

Finally, we can find $1 \leq N_1 \leq N_0$ such that

$$h_{N_1} \leq_{1-4N_1\delta} h_0 \cdot \left(1 + \frac{1}{K_{N_0}^{\frac{2\alpha-3}{2\alpha}}}\right)^{-N_1} \leq \frac{1}{2}. \tag{89}$$

From above inequality, we imply

$$N_1 \geq \frac{\log 2h_0}{\log\left(1 + \frac{1}{K_{N_0}^{\frac{2\alpha-3}{2\alpha}}}\right)} \tag{90}$$

For iteration $t \geq N_1$, we use the following recursion: $h_{t+1} \leq_{1-4\delta} h_t^{2\alpha/3}$. Thus

$$h_{N_0} \leq_{1-4(N_0-N_1)\delta} h_{N_1}^{(\frac{2\alpha}{3})^{N_0-N_1}} \leq_{1-4N_0\delta} \left(\frac{1}{2}\right)^{(\frac{2\alpha}{3})^{N_0-N_1}} \tag{91}$$

On the other hand, from (81), gradient dominance property, and stopping criterion $\|\boldsymbol{\Delta}\|_2 \leq \sqrt[2\alpha]{\epsilon}$, we have

$$F(\mathbf{x}_{N_0}) - F(\mathbf{x}^*) \leq \tau_F\|\nabla F(\mathbf{x}_{N_0})\|^\alpha \leq_{1-4\delta} \tau_F\left(\frac{M + L_2 + 4}{2}\right)^\alpha \cdot \epsilon. \tag{92}$$

Summing up (91) and (92), we get

$$F(\mathbf{x}_{N_0}) - F(\mathbf{x}^*) \leq_{1-(4N_0+4)\delta} \frac{B^{\frac{2\alpha}{2\alpha-3}}}{\tau_F^{\frac{3}{2\alpha-3}}}\left(\frac{1}{2}\right)^{(\frac{2\alpha}{3})^{N_0-N_1}} + \frac{\tau_F}{2}\left(\frac{M + L_2 + 4}{2}\right)^\alpha \cdot \epsilon. \tag{93}$$

First denote $\tilde{C} := 1 + \frac{\tau_F}{2}\left(\frac{M+L_2+4}{2}\right)^\alpha$ and $D := \frac{B^{\frac{2\alpha}{2\alpha-3}}}{\tau_F^{\frac{3}{2\alpha-3}}}$. Hence, we have to choose $N_0 \geq$ $N_1 + \frac{\log\left(\frac{\log D + \log\frac{1}{\epsilon}}{\log 2}\right)}{\log(2\alpha/3)}$, in order to guarantee that $F(\mathbf{x}_{N_0}) - F(\mathbf{x}^*) \leq \tilde{C}\epsilon$ with probability at least $1 - (4N_0 + 4)\delta$. By plugging $\delta' = 4(N_0 + 1)\delta$ and $\epsilon' = \tilde{C}\epsilon$ in the sample complexities n_1, n_2 in (79) and (80), we can get the desired result in Theorem 2 and the proof is complete.

23

A.2.2 Inexact cubic sub-solver (Proof of Lemma 2)

In this part, we show that under some mild assumptions, we can use a gradient descant based algorithm to find an approximate solution $\tilde{\boldsymbol{\Delta}}_t$ of sub-problem in (3) such that Theorem 1 and Theorem 2 still hold. To do so, we provide a new version of (21) for $\tilde{\boldsymbol{\Delta}}_t$ and then show that the order of error terms in the recursion is not changed. Recall that the sub-problem (3) has the following form:

$$\min_{\boldsymbol{\Delta} \in \mathbb{R}^d} m_t(\boldsymbol{\Delta}) := \langle \mathbf{g}_t, \boldsymbol{\Delta} \rangle + \frac{1}{2} \langle \boldsymbol{\Delta}, \mathbf{H}_t \boldsymbol{\Delta} \rangle + \frac{M}{6} \|\boldsymbol{\Delta}\|^3. \tag{94}$$

We need the following assumption:

Assumption 9. *Hessian estimators satisfy $\|\mathbf{H}_t\| \leq \beta$ for all $1 \leq t \leq T$ where β is a positive constant.*

Consider any iteration $t \in [1, T]$ of Algorithm 1 and in what follows, the iteration t is fixed. Thus, we drop the subscription t for simplifying the notations. We use the gradient descent algorithm as our sub-solver with the following update and initial point $\tilde{\boldsymbol{\Delta}}^{(0)}$:

$$\tilde{\boldsymbol{\Delta}}^{(k+1)} = \tilde{\boldsymbol{\Delta}}^{(k)} - \eta \nabla m(\tilde{\boldsymbol{\Delta}}^k) = \left(I - \eta \mathbf{H} - \eta \frac{M}{2} \|\tilde{\boldsymbol{\Delta}}^{(k)}\| \right) \tilde{\boldsymbol{\Delta}}^{(k)} - \eta \mathbf{g}.$$

We define the following quantity

$$R_c \triangleq -\frac{\mathbf{g}_t^T \mathbf{H} \mathbf{g}}{2M \|\mathbf{g}_t\|^2} + \sqrt{\frac{\|\mathbf{g}\|}{M} + \left(\frac{\mathbf{g}^T \mathbf{H} \mathbf{g}}{2M \|\mathbf{g}_t\|^2} \right)^2},$$

which is the Cauchy radius [4] and

$$R \triangleq \frac{\|\mathbf{H}\|}{2M} + \sqrt{\left(\frac{\|\mathbf{H}\|}{2M} \right)^2 + \frac{\|\mathbf{g}\|}{M}},$$

which is an upper bound on $\|\boldsymbol{\Delta}\|$ (See [4, Claim 2.1]). We make the following assumptions.

Assumption 10. *The step-size η satisfies $0 \leq \eta \leq \frac{1}{4(\|\mathbf{H}\| + MR)}$.*

Assumption 11. *The initialization $\tilde{\boldsymbol{\Delta}}^{(0)}$ satisfies $\tilde{\boldsymbol{\Delta}}^{(0)} = -r \frac{\mathbf{g}}{\|\mathbf{g}\|}$, with $0 \leq r \leq R_c$.*

In [4, Lemma 2.3], Carmon and Duchi showed that with the Assumptions 10 and 11, $\mathbf{g}^T \tilde{\boldsymbol{\Delta}}^{(k)} \leq 0$ and in [4, Lemma 2.2], they showed that $\nabla m(\tilde{\boldsymbol{\Delta}}^{(k)})^T \tilde{\boldsymbol{\Delta}}^{(k)} \leq 0$ for all $1 \leq k \leq K$. Assumptions 9, 11, and 10 implies

$$\|\nabla^2 m(\tilde{\boldsymbol{\Delta}}^{(k)})\| = \left\| \mathbf{H} + \frac{M \|\tilde{\boldsymbol{\Delta}}^{(k)}\|}{2} I_{d \times d} \right\| \leq \|\mathbf{H}\| + \frac{M \|\tilde{\boldsymbol{\Delta}}^{(k)}\|}{2} \stackrel{(a)}{\leq} \beta + \frac{MR}{2}$$

where (a) comes from Assumption 9 and $\|\tilde{\boldsymbol{\Delta}}^{(k)}\| \leq R$ which is from [4, Lemma 2.2]. Hence, the Lipschitzness of gradient of $m(\cdot)$ is established and the gradient descent algorithm obtains $\|\nabla m(\tilde{\boldsymbol{\Delta}}^{(K)})\| \leq \epsilon'$ with $K = \mathcal{O}(\epsilon'^{-2})$ [24]. Let $\tilde{\boldsymbol{\Delta}}_t = \tilde{\boldsymbol{\Delta}}^{(K)}$ be the inexact solution of above sub-solver.

From Lipschitzness of Hessian, we have the following:

$$F(\mathbf{x}_t + \tilde{\boldsymbol{\Delta}}_t) \leq F(\mathbf{x}_t) + \langle \nabla F(\mathbf{x}_t), \tilde{\boldsymbol{\Delta}}_t \rangle + \frac{1}{2} \langle \tilde{\boldsymbol{\Delta}}_t, \nabla^2 F(\mathbf{x}_t) \tilde{\boldsymbol{\Delta}}_t \rangle + \frac{L_2}{6} \|\boldsymbol{\Delta}_t\|^3$$

$$= F(\mathbf{x}_t) + \langle \nabla F(\mathbf{x}_t) - \mathbf{g}_t, \tilde{\boldsymbol{\Delta}}_t \rangle + \frac{1}{2} \langle \tilde{\boldsymbol{\Delta}}_t, (\nabla^2 F(\mathbf{x}_t) - \mathbf{H}_t) \tilde{\boldsymbol{\Delta}}_t \rangle + \frac{L_2}{6} \|\tilde{\boldsymbol{\Delta}}_t\|^3 + \mathbf{g}_t^T \tilde{\boldsymbol{\Delta}}_t + \frac{1}{2} \tilde{\boldsymbol{\Delta}}_t^T \mathbf{H}_t \tilde{\boldsymbol{\Delta}}_t$$

$$= F(\mathbf{x}_t) + \langle \nabla F(\mathbf{x}_t) - \mathbf{g}_t, \tilde{\boldsymbol{\Delta}}_t \rangle + \frac{1}{2} \langle \tilde{\boldsymbol{\Delta}}_t, (\nabla^2 F(\mathbf{x}_t) - \mathbf{H}_t) \tilde{\boldsymbol{\Delta}}_t \rangle + \frac{2L_2 - 3M}{12} \|\tilde{\boldsymbol{\Delta}}_t\|^3 + \frac{1}{2} \mathbf{g}_t^T \tilde{\boldsymbol{\Delta}}_t$$

$$+ \frac{1}{2} \nabla m(\tilde{\boldsymbol{\Delta}}_t)^T \tilde{\boldsymbol{\Delta}}_t$$

$$\stackrel{(a)}{\leq} F(\mathbf{x}_t) + \langle \nabla F(\mathbf{x}_t) - \mathbf{g}_t, \tilde{\boldsymbol{\Delta}}_t \rangle + \frac{1}{2} \langle \tilde{\boldsymbol{\Delta}}_t, (\nabla^2 F(\mathbf{x}_t) - \mathbf{H}_t) \tilde{\boldsymbol{\Delta}}_t \rangle + \frac{2L_2 - 3M}{12} \|\tilde{\boldsymbol{\Delta}}_t\|^3$$

$$\stackrel{(b)}{\leq} F(\mathbf{x}_t) + \|\nabla F(\mathbf{x}_t) - \mathbf{g}_t\|_2 \|\tilde{\boldsymbol{\Delta}}_t\|_2 + \frac{1}{2} \|\nabla^2 F(\mathbf{x}_t) - \mathbf{H}_t\| \|\tilde{\boldsymbol{\Delta}}_t\|_2^2 + \frac{2L_2 - 3M}{12} \|\tilde{\boldsymbol{\Delta}}_t\|^3$$

$$\stackrel{(c)}{\leq} F(\mathbf{x}_t) + \frac{2 \|\nabla F(\mathbf{x}_t) - \mathbf{g}_t\|^{3/2}}{3} + \frac{\|\nabla^2 F(\mathbf{x}_t) - \mathbf{H}_t\|^3}{6} + \frac{2L_2 - 3M + 8}{12} \|\tilde{\boldsymbol{\Delta}}_t\|^3, \tag{95}$$

24

where (a) comes from the fact that $\mathbf{g}^T\tilde{\boldsymbol{\Delta}}_t \leq 0$ and $\nabla m(\tilde{\boldsymbol{\Delta}}_t)^T\tilde{\boldsymbol{\Delta}}_t \leq 0$ and (b) comes from Cauchy-Schwartz inequality and (c) is due to Young's inequality.

We know from Lipschitzness of Hessian of F that

$$\|\nabla F(\mathbf{x}_t + \tilde{\boldsymbol{\Delta}}_t) - \nabla F(\mathbf{x}_t) - \nabla^2 F(\mathbf{x}_t)\tilde{\boldsymbol{\Delta}}_t\| \leq \frac{L_2}{2}\|\tilde{\boldsymbol{\Delta}}_t\|^2. \tag{96}$$

Moreover, the gradient descent returns $\tilde{\boldsymbol{\Delta}}_t$ which is ϵ'-approximate first-order stationary point: $\|\nabla m(\tilde{\boldsymbol{\Delta}}_t)\| = \|\mathbf{g}_t + \mathbf{H}_t\tilde{\boldsymbol{\Delta}}_t + \frac{M}{2}\|\tilde{\boldsymbol{\Delta}}_t\|\tilde{\boldsymbol{\Delta}}_t\| \leq \epsilon'$. Therefore, we have:

$$\|\mathbf{g}_t + \mathbf{H}_t\tilde{\boldsymbol{\Delta}}_t\| \leq \epsilon' + \frac{M}{2}\|\tilde{\boldsymbol{\Delta}}_t\|^2. \tag{97}$$

By summing the above two equations and using triangle inequality, we have:

$$\begin{aligned}
\|\nabla F(\mathbf{x}_t + \tilde{\boldsymbol{\Delta}}_t)\| &\leq \frac{M + L_2}{2}\|\tilde{\boldsymbol{\Delta}}_t\|^2 + \|\nabla F(\mathbf{x}_t) - \mathbf{g}_t\| + \|(\nabla^2 F(\mathbf{x}_t) - \mathbf{H}_t)\tilde{\boldsymbol{\Delta}}_t\| + \epsilon' \\
&\leq \frac{M + L_2}{2}\|\tilde{\boldsymbol{\Delta}}_t\|^2 + \|\nabla F(\mathbf{x}_t) - \mathbf{g}_t\| + \|\nabla^2 F(\mathbf{x}_t) - \mathbf{H}_t\|\|\tilde{\boldsymbol{\Delta}}_t\| + \epsilon' \qquad (98) \\
&\leq \frac{M + L_2}{2}\|\tilde{\boldsymbol{\Delta}}_t\|^2 + \|\nabla F(\mathbf{x}_t) - \mathbf{g}_t\| + \frac{\|\nabla^2 F(\mathbf{x}_t) - \mathbf{H}_t\|^2}{2} + \frac{\|\tilde{\boldsymbol{\Delta}}_t\|^2}{2} + \epsilon',
\end{aligned}$$
$$\tag{99}$$

Based on gradient dominance property and (99), we have:

$$\begin{aligned}
(F(\mathbf{x}_t + \tilde{\boldsymbol{\Delta}}_t) - F(\mathbf{x}^*)) &\leq \tau_F\|\nabla F(\mathbf{x}_t + \tilde{\boldsymbol{\Delta}}_t)\|^\alpha \\
&\leq \tau_F\Big(\frac{M + L_2 + 1}{2}\|\tilde{\boldsymbol{\Delta}}_t\|^2 + \|\nabla F(\mathbf{x}_t) - \mathbf{g}_t\| + \frac{\|\nabla^2 F(\mathbf{x}_t) - \mathbf{H}_t\|^2}{2} + \epsilon'\Big)^\alpha \qquad (100) \\
&\leq 4^{\alpha-1}\tau_F\Big[\Big(\frac{M + L_2 + 1}{2}\Big)^\alpha\|\tilde{\boldsymbol{\Delta}}_t\|^{2\alpha} + \|\nabla F(\mathbf{x}_t) - \mathbf{g}_t\|^\alpha + \frac{\|\nabla^2 F(\mathbf{x}_t) - \mathbf{H}_t\|^{2\alpha}}{2^\alpha} + {\epsilon'}^\alpha\Big]
\end{aligned}$$

where we used $(a + b + c + d)^\alpha \leq 4^{\alpha-1}(a^\alpha + b^\alpha + c^\alpha + d^\alpha), \forall a, b, c, d > 0$ in the last inequality. Substituting (95) into (100), we have:

$$\begin{aligned}
F(\mathbf{x}_t &+ \tilde{\boldsymbol{\Delta}}_t) - F(\mathbf{x}^*) \leq \\
&4^{\alpha-1}\tau_F\Big[\Big(\frac{M + L_2 + 1}{2}\Big)^\alpha\Big(\frac{12}{3M - 2L_2 - 6}\Big)^{2\alpha/3}\Big(F(\mathbf{x}_t) - F(\mathbf{x}_t + \tilde{\boldsymbol{\Delta}}_t) + \frac{2\|\nabla F(\mathbf{x}_t) - \mathbf{g}_t\|^{3/2}}{3} + \\
&\qquad \frac{\|\nabla^2 F(\mathbf{x}_t) - \mathbf{H}_t\|^3}{6}\Big)^{2\alpha/3} + \|\nabla F(\mathbf{x}_t) - \mathbf{g}_t\|^\alpha + \frac{\|\nabla^2 F(\mathbf{x}_t) - \mathbf{H}_t\|^{2\alpha}}{2^\alpha} + {\epsilon'}^\alpha\Big] \\
&\leq C(F(\mathbf{x}_t) - F(\mathbf{x}_t + \tilde{\boldsymbol{\Delta}}_t))^{2\alpha/3} + C_g\|\nabla F(\mathbf{x}_t) - \mathbf{g}_t\|^\alpha + C_H\|\nabla^2 F(\mathbf{x}_t) - \mathbf{H}_t\|^{2\alpha} + C_{\epsilon'}{\epsilon'}^\alpha,
\end{aligned}$$
$$\tag{101}$$

where in the last inequality we used $(a + b + c)^{2\alpha/3} \leq 3^{2\alpha/3-1/3}(a^{2\alpha/3} + b^{2\alpha/3} + c^{2\alpha/3})$ which is derived by the following inequalities:

$$(a + b + c)^{2\alpha} \leq 3^{2\alpha-1}(a^{2\alpha} + b^{2\alpha} + c^{2\alpha}) \leq 3^{2\alpha-1}(a^{2\alpha/3} + b^{2\alpha/3} + c^{2\alpha/3})^3$$

for any $a, b, c \in \mathbb{R}^+$ and,

$$\begin{aligned}
C &= 4^{\alpha-1}3^{(2\alpha-1)/3}\tau_F\Big(\frac{M + L_2 + 1}{2}\Big)^\alpha\Big(\frac{12}{3M - 2L_2 - 8}\Big)^{2\alpha/3} \\
C_g &= 2^{8\alpha/3-2} \times 3^{-\frac{1}{3}}\tau_F\Big(\frac{M + L_2 + 1}{2}\Big)^\alpha\Big(\frac{12}{3M - 2L_2 - 8}\Big)^{2\alpha/3} + 2^{\alpha-2}\tau_F \\
C_H &= 2^{4\alpha/3-2} \times 3^{-\frac{1}{3}}\tau_F\Big(\frac{M + L_2 + 1}{2}\Big)^\alpha\Big(\frac{12}{3M - 2L_2 - 8}\Big)^{2\alpha/3} + 2^{\alpha-2}\tau_F \\
C_{\epsilon'} &= 4^{\alpha-1}\tau_F
\end{aligned} \tag{102}$$

If $\epsilon' = \epsilon^{1/\alpha}$, we have the following recursion inequality.

$$\delta_{t+1} \leq C(\delta_t - \delta_{t+1})^{2\alpha/3} + C_g\|\nabla F(\mathbf{x}_t) - \mathbf{g}_t\|^\alpha + C_H\|\nabla^2 F(\mathbf{x}_t) - \mathbf{H}_t\|^{2\alpha} + C_{\epsilon'}\epsilon. \tag{103}$$

25
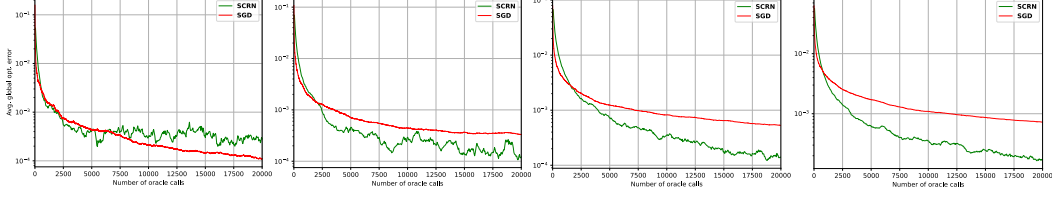
Figure 4: Performance of SCRN versus SGD with adaptive step-size for different values of α. From left to right, average global optimization error of SCRN and SGD versus number of oracle calls for $\alpha = 4/3, 5/4, 6/5, 7/6$.
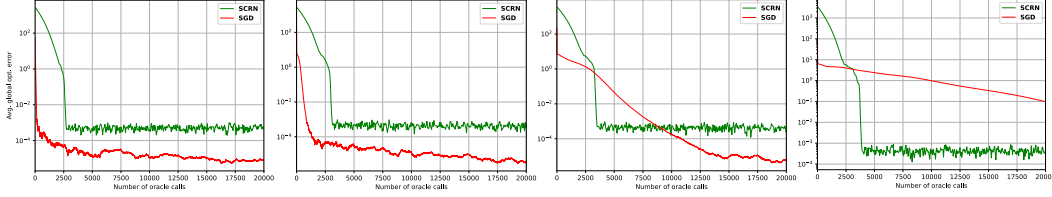
Figure 5: Performance of SCRN versus SGD on a function satisfying gradient dominance property with $\alpha = 2$ for different values of τ_F. From left to right, average global optimization error versus number of oracle calls for $\tau_F \approx 8, 32, 85, 190$.

Recall that in the analysis of Sections 3.1 and 3.2, we keep the error terms $\|\nabla F(\mathbf{x}_t) - \mathbf{g}_t\|^{\alpha}$ and $\|\nabla^2 F(\mathbf{x}_t) - \mathbf{H}_t\|^{2\alpha}$ in the order of ϵ (in expectation or with high probability) and note that the effect of inexactness in sub-solver appears in the term $C_{\epsilon'}\epsilon$ with the same order. Hence, we still have a valid analysis of the global optimum when using an inexact sub-solver based on gradient descent. It is important to note that when using the inexact sub-solver, finding a local minimum solution of sub-problem is sufficient for the global convergence analysis of SCRN and despite second order convergence analysis of SCRN [4, 33, 49], its global convergence analysis does not need to approximate the exact solution of sub-solver, i.e., $\boldsymbol{\Delta}_t$.

A.2.3 Simulations of SCRN and SGD on synthetic functions satisfying gradient dominance property

Example 1. *We consider function $x^{p/q} : [-1, 1] \to [0, 1]$, satisfying gradient dominance property with $\alpha = \frac{p}{p-q} \in (1, \infty)$ where p is an even positive integer and $q < p$ is a positive integer. For this choice of objective function, we compare average global optimization error $(\mathbb{E}[F(x)] - F(x^*))$ of SCRN and SGD for $1 < \alpha < 3/2$ in Figure 4. We observe that the improvement of SCRN upon SGD is increasing when α decreases from $4/3$ to $7/6$.*

Example 2. *As we discussed in Remark 5, SCRN improves SGD in terms of dependency on τ_F. We consider the function $F(x) = x^2 + a\sin^2(x)$ for $0 \le a \le 4.60\bar{3}$. This choice of function satisfy gradient dominance property of $\alpha = 2$ with different values of τ_F. When a goes to $4.60\bar{3}$, τ_F goes to infinity. By tuning the parameter a, we run SCRN and SGD with the best adaptive step-size on function $F(x)$ for four different values of τ_F in Figure 5. As can be seen, SCRN outperforms SGD for large values of τ_F.*

A.3 variance-reduced SCRN under gradient dominance property with $\alpha = 1$

We first provide two lemmas that are used in analyzing variance-reduced version of SCRN. They are stated for more general case $1 \le \alpha < 3/2$ in the following. We will use them in the proof of Theorem 3 for the case $\alpha = 1$.

Lemma 10. *Consider the following recursive inequality for $1 \le \alpha < 3/2$:*

$$\delta_{t+1} \le (\delta_t - \delta_{t+1})^{2\alpha/3}. \tag{104}$$

26

Then,

- *(i) As far as $\delta_t \geq 1$, we have: $\delta_t \leq (1/2)^t \delta_0$.*

- *(ii) For any pair (t, t') such that $t' \leq t$:*

$$\delta_t \leq \beta^\beta \left(\beta \delta_{t'}^{-1/\beta} + D(t - t') \right)^{-\beta}, \tag{105}$$

where $\beta := \frac{2\alpha}{3 - 2\alpha}$ and $D := \min\{\frac{1}{2}, \beta(2^{\frac{3-2\alpha}{3}} - 1)\delta_{t'}^{-1/\beta}\}$.

Proof. For part (i), as long as $\delta_t \geq 1$, we have:

$$\frac{\delta_t}{\delta_{t-1}} \leq \frac{1}{1 + \delta_t^{-1 + 3/(2\alpha)}} \leq \frac{1}{2}, \tag{106}$$

where the first inequality is due to (104) and the second inequality comes from the fact that $\delta_t \geq 1$. Therefore, $\prod_{k=1}^t \delta_k / \delta_{k-1} \leq (1/2)^t$ which we can imply that: $\delta_t \leq (1/2)^t \delta_0$.

For part (ii), let $h(t) := \frac{2\alpha}{3 - 2\alpha} t^{1 - \frac{3}{2\alpha}}$. Consider the following two cases:

Case (1): Suppose that $\delta_{t+1} \geq 2^{-\frac{2\alpha}{3}} \delta_t$ for all $t \geq t'$. Then

$$h(\delta_{t+1}) - h(\delta_t) = \int_{\delta_t}^{\delta_{t+1}} \frac{d}{dt} h(t) dt = \int_{\delta_{t+1}}^{\delta_t} t^{-\frac{3}{2\alpha}} dt \geq (\delta_t - \delta_{t+1}) \delta_t^{-\frac{3}{2\alpha}}$$

$$\geq (\delta_t - \delta_{t+1}) \frac{1}{2} \delta_{t+1}^{-\frac{3}{2\alpha}} \geq \frac{1}{2}. \tag{107}$$

Case (2): Suppose that $\delta_{t+1} \leq 2^{-\frac{2\alpha}{3}} \delta_t$ for all $t \geq t'$. Then, we have: $\delta_{t+1}^{1 - \frac{3}{2\alpha}} \geq 2^{\frac{3-2\alpha}{3}} \delta_t^{1 - \frac{3}{2\alpha}}$. Therefore,

$$h(\delta_{t+1}) - h(\delta_t) = \frac{2\alpha}{3 - 2\alpha} (\delta_{t+1}^{1 - \frac{3}{2\alpha}} - \delta_t^{1 - \frac{3}{2\alpha}}) \geq \frac{2\alpha}{3 - 2\alpha} (2^{\frac{3-2\alpha}{3}} - 1)\delta_t^{1 - \frac{3}{2\alpha}}$$

$$\geq \frac{2\alpha}{3 - 2\alpha} (2^{\frac{3-2\alpha}{3}} - 1)\delta_{t'}^{1 - \frac{3}{2\alpha}}. \tag{108}$$

Let $D := \min\{\frac{1}{2}, \frac{2\alpha}{3-2\alpha}(2^{\frac{3-2\alpha}{3}} - 1)\delta_{t'}^{1 - \frac{3}{2\alpha}}\}$. Then we have

$$h(\delta_{t+1}) - h(\delta_t) \geq D$$

which implies

$$h(\delta_t) - h(\delta_{t'}) = \sum_{k=t'+1}^t h(\delta_k) - h(\delta_{k-1}) \geq D \cdot (t - t').$$

Then we get

$$\delta_t \leq \left(\frac{2\alpha}{3 - 2\alpha} \right)^{\frac{2\alpha}{3-2\alpha}} \frac{1}{(h(\delta_{t'}) + D(t - t'))^{\frac{2\alpha}{3-2\alpha}}}. \tag{109}$$

$\square$

Lemma 11. *Consider the following recursive inequality for $1 \leq \alpha < 3/2$:*

$$\delta_t \leq C(\delta_{t-1} - \delta_t)^{2\alpha/3} + \frac{C'}{(\lceil t/S \rceil S)^\beta}, \qquad \forall t \geq 1, \tag{110}$$

where $\beta := \frac{2\alpha}{3-2\alpha}$, C, and C' are some positive constants and S is a positive integer. Then,

$$\delta_t \leq \frac{C_\delta C^{3/(3-2\alpha)}}{(\lfloor t/S \rfloor S)^\beta} + \frac{C'}{(\lceil t/S \rceil S)^\beta}, \tag{111}$$

for all $t \geq S$ where C_δ satisfies $C_\delta^{-1/\beta} \left(1 + A^{1/(1+\beta)}\right) \leq \frac{1}{2\beta}$ where $A := \beta C'/(C^{3/(3-2\alpha)}C_\delta)$ and $S > \max\{C_\delta^{1/\beta}(1 + A)^{1/\beta}, C_\delta^{1/\beta}(t_0 - 2\beta)/(C_\delta^{1/\beta} - 2\beta)\}$, and $t_0 \leq \lceil \log(\delta_0/C^{3/(3-2\alpha)})/\log(2) \rceil$.

27

Remark 12. *In the analysis of SCRN where $\delta_t = \mathbb{E}[F(\mathbf{x}_t)] - F(\mathbf{x}^*)$, and $C' = C_g + C_H$ which are chosen in (34), if $C_\delta = (2\beta)^{\beta+1}C'$, then the condition $C_\delta^{-1/\beta}\left(1 + \left(\frac{\beta C'}{C^{3/(3-2\alpha)}C_\delta}\right)^{1/(1+\beta)}\right) \leq \frac{1}{2\beta}$ turns to*

$$1 + \frac{1}{C \cdot 2^{\frac{\beta}{1+\beta}}\beta^{\frac{\beta}{1+\beta}}} \leq \beta^{1/\beta}C'^{1/\beta}$$

which is satisfied for large M. Hence, after T iterations, if the errors of gradient and Hessian are

$$\mathbb{E}[\|\nabla F(\mathbf{x}_t) - \mathbf{g}_t\|^\alpha] \leq \frac{1}{(\lceil t/S \rceil S)^\beta}, \quad \mathbb{E}[\|\nabla^2 F(\mathbf{x}_t) - \mathbf{H}_t\|^{2\alpha}] \leq \frac{1}{(\lceil t/S \rceil S)^\beta},$$

then

$$\delta_T \leq \frac{(2\beta)^{\beta+1}(C_g + C_H)C^{3/(3-2\alpha)}}{(\lfloor T/S \rfloor S)^\beta} + \frac{C_g + C_H}{(\lceil T/S \rceil S)^\beta}. \tag{112}$$

Proof. We define $\delta_t^k := (\delta_t - C'/(kS)^\beta)/C^{3/(3-2\alpha)}$ for $(k-1)S \leq t \leq kS$ and $k \geq 1$. Therefore, we have the following recursive inequality: $\delta_t^k \leq (\delta_t^k - \delta_{t-1}^k)^{2\alpha/3}$ for $(k-1)S + 1 \leq t \leq kS$. By induction on k, we will show that $\delta_{kS}^k \leq C_\delta/(kS)^\beta$ for all $k \geq 1$. We first prove the statement for the base case $k = 1$. Suppose t_0 is the first iteration such that $\delta_{t_0}^1 < 1$. From Lemma 10 (part (i)), we know that $t_0 \leq \lceil \log(\delta_0/C^{3/(3-2\alpha)})/\log(2) \rceil$. Moreover, from part (ii) in Lemma 10, by setting $t' = t_0$, we have:

$$\begin{aligned}
\delta_S^1 &\leq \beta^\beta \left(\beta(\delta_{t_0}^1)^{-1/\beta} + D(S - t_0)\right)^{-\beta} \\
&\leq \beta^\beta \left(\beta + (S - t_0)/2\right)^{-\beta} \\
&\leq \frac{C_\delta}{S^\beta},
\end{aligned} \tag{113}$$

where the second inequality is due to fact that $D = 1/2$ as $\delta_{t_0}^1 < 1$ and the third inequality comes from the constraints on C_δ and S in the statement of lemma.

Now, suppose that the statement holds up to some k. We will show that it also holds for $k + 1$. From the induction hypothesis, we know that

$$\begin{aligned}
\delta_{kS} &\leq \frac{C^{3/(3-2\alpha)}C_\delta}{(kS)^\beta} + \frac{C'}{(kS)^\beta} \\
\implies \delta_{kS}^{k+1} &\leq \frac{C_\delta}{(kS)^\beta} + \frac{C'}{C^{3/(3-2\alpha)}}\left(\frac{1}{(kS)^\beta} - \frac{1}{((k+1)S)^\beta}\right) \\
&\leq \frac{C_\delta}{(kS)^\beta}\left(1 + \frac{A}{k^{\beta+1}}\right),
\end{aligned} \tag{114}$$

where $A = \beta C'/(C^{3/(3-2\alpha)}C_\delta)$ and the second inequality is due to definition of δ_{kS}^{k+1}.

From the constraint on C_δ in the statement of lemma, we have:

$$\begin{aligned}
&C_\delta^{-1/\beta}\left(1 + A^{1/(1+\beta)}\right) \leq \frac{1}{2\beta} \\
&\overset{(a)}{\implies} C_\delta^{-1/\beta}\left(1 + \frac{A^{1/\beta}}{k^{1/\beta} + A^{1/\beta}/k}\right) \leq \frac{1}{2\beta} \\
&\implies C_\delta^{-1/\beta}\left(k + 1 - \frac{k}{(1 + A^{1/\beta}/k^{1/\beta+1})}\right) \leq \frac{1}{2\beta} \\
&\overset{(b)}{\implies} C_\delta^{-1/\beta}\left(k + 1 - \frac{k}{(1 + A/k^{\beta+1})^{1/\beta}}\right) \leq \frac{1}{2\beta} \\
&\implies C_\delta^{-1/\beta}(k+1)S \leq \frac{kS}{(C_\delta(1 + A/k^{\beta+1}))^{1/\beta}} + S/(2\beta),
\end{aligned} \tag{115}$$

28

where (a) is due to fact that $A^{1/(\beta(\beta+1))} < (\beta^{1/(1+\beta)} + \beta^{-\beta/(1+\beta)})A^{1/(\beta(1+\beta))} \le k^{1/\beta} + A^{1/\beta}/k$ for $\beta > 1$ and $k > 0$ and (b) comes from $(a+b)^{1/\beta} \le (a^{1/\beta} + b^{1/\beta})$ for $\beta > 1$ and any $a, b > 0$.

Now, from part (ii) of Lemma 10, we have:

$$
\begin{aligned}
\delta_{(k+1)S}^{k+1} &\le \beta^\beta \left(\beta(\delta_{kS}^{k+1})^{-1/\beta} + DS \right)^{-\beta} \\
&\overset{(a)}{\le} \left(\frac{kS}{(C_\delta(1 + A/k^{\beta+1}))^{1/\beta}} + S/(2\beta) \right)^{-\beta} \\
&\overset{(b)}{\le} \frac{C_\delta}{((k+1)S)^\beta},
\end{aligned}
\tag{116}
$$

where (a) is according to (114) and the fact that $D = 1/2$ (as $\delta_{kS}^{k+1} < 1$ due to the constraint on S in the statement of lemma) and (b) comes from (115).

From the definition of δ_{kS}^k, we can imply that for all $k \ge 1$,

$$
\delta_{kS} \le \frac{C^{3/(3-2\alpha)}C_\delta + C'}{(kS)^\beta}.
\tag{117}
$$

Moreover, from $(\delta_t^k)^{3/2\alpha} + \delta_t^k \le \delta_{t-1}^k$ for $(k-1)S+1 \le t \le kS$, we can imply that: $\delta_t^k \le \delta_{(k-1)S}^k$ for $(k-1)S+1 \le t \le kS$. Therefore,

$$
\delta_t \le \frac{C_\delta C^{3/(3-2\alpha)}}{(\lfloor t/S \rfloor S)^\beta} + \frac{C'}{(\lceil t/S \rceil S)^\beta} \quad \mathrm{mod}(t, S) \ne 0.
\tag{118}
$$

$\square$

The above lemma shows that it suffices to bound the expectations of error terms for stochastic gradient and Hessian by $\mathcal{O}(1/t^\beta)$ and $\delta_t := \mathbb{E}[F(\mathbf{x}_t)] - F(\mathbf{x}^*)$ still has the same convergence rate in Theorem 1. In the following, we show that incorporating time-varying batch sizes in conjunction with variance reduction improves sample complexity results.

Algorithm 2 variance-reduced stochastic cubic regularized Newton method

---

Input: Maximum number of iterations T, batch sizes $\{n_g^t\}_{t=1}^T$, $\{n_H^t\}_{t=1}^T$, the period length S, initial point $\mathbf{x}_0$, and cubic penalty parameter M.

1: $t \leftarrow 0$
2: **while** $t \le T$ **do**
3:      Sample index set $\mathcal{J}_t$ with $|\mathcal{J}_t| = n_g^t$; $\mathcal{I}_t$ with $|\mathcal{I}_t| = n_H^t$
4:
$$
\mathbf{v}_t \leftarrow \begin{cases} \nabla f_{\mathcal{J}_t}(\mathbf{x}_t), & \mathrm{mod}(t, S) = 0 \\ \nabla f_{\mathcal{J}_t}(\mathbf{x}_t) - \nabla f_{\mathcal{J}_t}(\mathbf{x}_{t-1}) + \mathbf{v}_{t-1}, & \text{else} \end{cases}
$$
5:
$$
\mathbf{U}_t \leftarrow \begin{cases} \nabla^2 f_{\mathcal{I}_t}(\mathbf{x}_t), & \mathrm{mod}(t, S) = 0 \\ \nabla^2 f_{\mathcal{I}_t}(\mathbf{x}_t) - \nabla^2 f_{\mathcal{I}_t}(\mathbf{x}_{t-1}) + \mathbf{U}_{t-1}, & \text{else} \end{cases}
$$
6:      $\mathbf{\Delta}_t \leftarrow \arg\min_{\mathbf{\Delta} \in \mathbb{R}^d} \langle \mathbf{v}_t, \mathbf{\Delta} \rangle + \frac{1}{2}\langle \mathbf{\Delta}, \mathbf{U}_t\mathbf{\Delta} \rangle + \frac{M}{6}\|\mathbf{\Delta}\|^3$
7:      $\mathbf{x}_{t+1} \leftarrow \mathbf{x}_t + \mathbf{\Delta}_t$
8:      $t \leftarrow t + 1$
9: **end while**
10: **return** $\mathbf{x}_t$

We consider the following assumptions:

Assumption 12. *We assume that $f(\mathbf{x}, \xi)$ satisfies L_1'-average smoothness and (L_2', α)-average Hessian Lipschitz continuity, i.e., $\mathbb{E}[\|\nabla f(\mathbf{x}, \xi) - \nabla f(\mathbf{y}, \xi)\|^2] \le L_1'^2 \|\mathbf{x} - \mathbf{y}\|^2$ and $\mathbb{E}[\|\nabla^2 f(\mathbf{x}, \xi) - \nabla^2 f(\mathbf{y}, \xi)\|^{2\alpha}] \le L_2'^{2\alpha} \|\mathbf{x} - \mathbf{y}\|^{2\alpha}$ for all $\mathbf{x}, \mathbf{y} \in \mathbb{R}^d$ and $1 \le \alpha < 3/2$.*

29

It is noteworthy that the above assumption hold in policy-based RL setting that is considered in this paper.

Let define $\nabla f_{\mathcal{J}_t}(\mathbf{x}_t) := \frac{1}{|\mathcal{J}_t|} \sum_{j \in \mathcal{J}_t} \nabla f(\mathbf{x}_t, \xi_j)$ and $\nabla^2 f_{\mathcal{I}_t}(\mathbf{x}_t) := \frac{1}{|\mathcal{I}_t|} \sum_{i \in \mathcal{I}_t} \nabla^2 f(\mathbf{x}_t, \xi_i)$. By adapting the variance reduction method in [47] (see Algorithm 2), we can have the following bound on the errors of estimated gradient $\mathbf{v}_t$ and Hessian $\mathbf{U}_t$:

Lemma 12. *Let n_g^t and n_H^t be the number of stochastic gradients and Hessian matrices that are taken at time t. Suppose that for a given $\epsilon > 0$, we take the following number of samples at checkpoints:*

$$n_g^t \geq \frac{2\sigma_1^2}{\epsilon^{2/\alpha}}, n_H^t \geq \frac{2^{1/\alpha}(8e \cdot \log d)^2 \cdot \sigma_{2,\alpha}^{2/\alpha} S^{1-1/\alpha}}{\epsilon^{1/\alpha}}. \tag{119}$$

and at the other iterations, we take the following number of samples:

$$n_g^t(\mathbf{x}_t, \mathbf{x}_{t-1}) \geq \frac{4L_1'^2 S \|\boldsymbol{\Delta}_{t-1}\|^2}{\epsilon^{2/\alpha}}, n_H^t(\mathbf{x}_t, \mathbf{x}_{t-1}) \geq \frac{4 \times 2^{1/\alpha}(8e \cdot \log d)^2 \cdot L_2'^2 S \|\boldsymbol{\Delta}_{t-1}\|^2}{\epsilon^{1/\alpha}}, \tag{120}$$

Then, under Assumptions 3 and 12 for a specific amount of $\alpha \in [1, 3/2)$, we have:

$$\mathbb{E}[\|\nabla F(\mathbf{x}_t) - \mathbf{v}_t\|^\alpha] \leq \epsilon, \qquad \mathbb{E}[\|\nabla^2 F(\mathbf{x}_t) - \mathbf{U}_t\|^{2\alpha}] \leq \epsilon. \tag{121}$$

Proof. First for the gradient estimator $\mathbf{v}_t$, we have $\mathbf{v}_t - \nabla F(\mathbf{x}_t) = \sum_{k=\lfloor t/S \rfloor S}^t \mathbf{u}_k$ such that

$$\mathbf{u}_k = \begin{cases} \nabla f_{\mathcal{J}_k}(\mathbf{x}_k) - \nabla F(\mathbf{x}_k), & k = \lfloor t/S \rfloor S, \\ \nabla f_{\mathcal{J}_k}(\mathbf{x}_k) - \nabla f_{\mathcal{J}_k}(\mathbf{x}_{k-1}) - \nabla F(\mathbf{x}_k) + \nabla F(\mathbf{x}_{k-1}), & k > \lfloor t/S \rfloor S. \end{cases}$$

We know that $\mathbb{E}[\mathbf{u}_k|\mathcal{G}_k] = 0$ where $\mathcal{G}_k = \sigma(\mathbf{x}_0, \ldots, \mathbf{x}_k, \mathbf{u}_0, \cdots, \mathbf{u}_{k-1})$. Conditioned on $\mathcal{G}_k$ and for $k > \lfloor t/S \rfloor S$, we have

$$\mathbb{E}[\|\mathbf{u}_k\|^2|\mathcal{G}_k] = \mathbb{E}\left\| \frac{1}{n_g^k} \sum_{i=1}^{n_g^k} \nabla f(\mathbf{x}_k, \xi_i) - \nabla f(\mathbf{x}_{k-1}, \xi_i) - \nabla F(\mathbf{x}_k) + \nabla F(\mathbf{x}_{k-1}) \right\|^2$$

$$\overset{(a)}{=} \frac{1}{n_g^k} \mathbb{E} \|\nabla f(\mathbf{x}_k, \xi_1) - \nabla f(\mathbf{x}_{k-1}, \xi_1) - \nabla F(\mathbf{x}_k) + \nabla F(\mathbf{x}_{k-1})\|^2$$

$$\overset{(b)}{\leq} \frac{2}{n_g^k} \mathbb{E} \|\nabla f(\mathbf{x}_k, \xi_1) - \nabla f(\mathbf{x}_{k-1}, \xi_1)\|^2 + \frac{2}{n_g^k} \mathbb{E} \|\nabla F(\mathbf{x}_k) - \nabla F(\mathbf{x}_{k-1})\|^2$$

$$\leq \frac{4L_1'^2}{n_g^k} \|\mathbf{x}_k - \mathbf{x}_{k-1}\|^2 \tag{122}$$

where (a) comes from $[\nabla f(\mathbf{x}_k, \xi_i) - \nabla f(\mathbf{x}_{k-1}, \xi_i) - \nabla F(\mathbf{x}_k) + \nabla F(\mathbf{x}_{k-1})]$'s are i.i.d conditioned on $\mathcal{G}_k$ for $1 \leq i \leq n_g^k$. Inequality (b) is from $(a + b)^2 \leq 2a^2 + 2b^2$ and (c) is derived by Assumption 12. For $k = \lfloor t/S \rfloor S$, $\mathbb{E}[\|\mathbf{u}_k\|^2] \leq \frac{\sigma_1^2}{n_g^k}$.

$$\mathbb{E}[\|\nabla F(\mathbf{x}_t) - \mathbf{v}_t\|^2] = \mathbb{E}\left\| \sum_{k=\lfloor t/S \rfloor S}^t \mathbf{u}_k \right\|^2 \leq \sum_{k=\lfloor t/S \rfloor S}^t \mathbb{E}[\mathbb{E}[\|\mathbf{u}_k\|^2|\mathcal{G}_k]]$$

$$\leq \frac{\sigma_1^2}{n_g^{\lfloor t/S \rfloor S}} + \sum_{k=\lfloor t/S \rfloor S+1}^t 4L_1'^2 \mathbb{E}\left[\frac{\|\mathbf{x}_k - \mathbf{x}_{k-1}\|^2}{n_g^k} \right] \tag{123}$$

where the first inequality comes from the fact that $\mathbb{E}[\mathbf{u}_k^T \mathbf{u}_{k'}] = \mathbb{E}[\mathbf{u}_{k'}^T \mathbb{E}[\mathbf{u}_k|\mathcal{G}_k]] = 0$ for $k > k'$.

If we take $n_g^k \geq \frac{2\sigma_1^2}{\epsilon^{2/\alpha}}$ samples at checkpoints ($k = \lfloor t/S \rfloor S$) and and at the other iterations, we take $n_g^k \geq \frac{4L_1'^2 S \|\boldsymbol{\Delta}_{k-1}\|^2}{\epsilon^{2/\alpha}}$ samples, we have

$$\mathbb{E}[\|\nabla F(\mathbf{x}_t) - \mathbf{v}_t\|^\alpha] \leq \left(\mathbb{E}[\|\nabla F(\mathbf{x}_t) - \mathbf{v}_t\|^2]\right)^{\alpha/2} \leq \epsilon.$$

Now we give a proof for Hessian estimator which is similar to the gradient one.

30

First for the Hessian estimator $\mathbf{U}_t$, we have $\nabla^2 F(\mathbf{x}_t) - \mathbf{U}_t = \sum_{k=\lfloor t/S \rfloor S}^{t} \mathbf{V}_k$ such that

$$\mathbf{V}_k = \begin{cases} \nabla^2 f_{\mathcal{J}_k}(\mathbf{x}_k) - \nabla^2 F(\mathbf{x}_k), & k = \lfloor t/S \rfloor S, \\ \nabla^2 f_{\mathcal{J}_k}(\mathbf{x}_k) - \nabla^2 f_{\mathcal{J}_k}(\mathbf{x}_{k-1}) - \nabla^2 F(\mathbf{x}_k) + \nabla^2 F(\mathbf{x}_{k-1}), & k > \lfloor t/S \rfloor S. \end{cases}$$

Let $\mathcal{H}_k = \sigma(\mathbf{x}_0, \cdots, \mathbf{x}_k, \mathbf{V}_0, \ldots, \mathbf{V}_{k-1})$. Then, we have $\mathbb{E}[\mathbf{V}_k | \mathcal{H}_k] = 0$. Conditioned on $\mathcal{H}_k$ and for $k > \lfloor t/S \rfloor S$, we have

$$\mathbb{E}[\|\mathbf{V}_k\|^{2\alpha} | \mathcal{H}_k] = \mathbb{E} \left\| \frac{1}{n_H^k} \sum_{i=1}^{n_H^k} \nabla^2 f(\mathbf{x}_k, \xi_i) - \nabla^2 f(\mathbf{x}_{k-1}, \xi_i) - \nabla^2 F(\mathbf{x}_k) + \nabla^2 F(\mathbf{x}_{k-1}) \right\|^{2\alpha}$$

$$\overset{(a)}{\leq} (8e \log d)^\alpha \cdot \mathbb{E} \left\| \frac{1}{(n_H^k)^2} \sum_{i=1}^{n_H^k} [\nabla^2 f(\mathbf{x}_k, \xi_i) - \nabla^2 f(\mathbf{x}_{k-1}, \xi_i) - \nabla^2 F(\mathbf{x}_k) + \nabla^2 F(\mathbf{x}_{k-1})]^2 \right\|^{\alpha}$$

$$\overset{(b)}{\leq} \frac{(8e \log d)^\alpha}{(n_H^k)^\alpha} \cdot \frac{1}{n_H^k} \sum_{i=1}^{n_H^k} \mathbb{E} \left\| [\nabla^2 f(\mathbf{x}_k, \xi_i) - \nabla^2 f(\mathbf{x}_{k-1}, \xi_i) - \nabla^2 F(\mathbf{x}_k) + \nabla^2 F(\mathbf{x}_{k-1})]^2 \right\|^{\alpha}$$

$$\overset{(c)}{\leq} \frac{(8e \log d)^\alpha}{(n_H^k)^\alpha} \mathbb{E} \left\| [\nabla^2 f(\mathbf{x}_k, \xi_1) - \nabla^2 f(\mathbf{x}_{k-1}, \xi_1) - \nabla^2 F(\mathbf{x}_k) + \nabla^2 F(\mathbf{x}_{k-1})] \right\|^{2\alpha}$$

$$\overset{(d)}{\leq} \frac{2^{2\alpha-1}(8e \log d)^\alpha}{(n_H^k)^\alpha} \mathbb{E} \|\nabla^2 f(\mathbf{x}_k, \xi_1) - \nabla^2 f(\mathbf{x}_{k-1}, \xi_1)\|^{2\alpha} + \frac{2^{2\alpha-1}(8e \log d)^\alpha}{(n_H^k)^\alpha} \mathbb{E} \|\nabla^2 F(\mathbf{x}_k) - \nabla^2 F(\mathbf{x}_{k-1})\|^{2\alpha}$$

$$\overset{(e)}{\leq} \frac{2^{2\alpha}(8e \log d)^\alpha (L_1')^{2\alpha}}{(n_H^k)^\alpha} \|\mathbf{x}_k - \mathbf{x}_{k-1}\|^{2\alpha}, \tag{124}$$

where (a) comes from Lemma 4 and (b) comes from Jensen's inequality for operator norm $\|\cdot\|^\alpha$. (c) is derived by $\|AB\| \leq \|A\|\|B\|$ and the fact that $[\nabla^2 f(\mathbf{x}_k, \xi_i) - \nabla^2 f(\mathbf{x}_{k-1}, \xi_i) - \nabla^2 F(\mathbf{x}_k) + \nabla^2 F(\mathbf{x}_{k-1})]$'s are i.i.d conditioned on $\mathcal{H}_k$ for $1 \leq i \leq n_g^k$. Inequality (d) is from inequality $(a+b)^{2\alpha} \leq 2^{2\alpha-1}(a^{2\alpha} + b^{2\alpha})$ and (e) is derived by Assumption 12. For $k = \lfloor t/S \rfloor S$, $\mathbb{E}[\|\mathbf{V}_k\|^{2\alpha}] \leq \frac{(8e \log d)^\alpha \cdot \sigma_{2,\alpha}^2}{(n_H^k)^\alpha}$ from Lemma 4. Then

$$\mathbb{E}[\|\nabla^2 F(\mathbf{x}_t) - \mathbf{U}_t\|^{2\alpha}] = \mathbb{E} \left\| \sum_{k=\lfloor t/S \rfloor S}^{t} \mathbf{V}_k \right\|^{2\alpha} \overset{(a)}{\leq} (8e \log d)^\alpha \mathbb{E} \left\| \frac{S}{S} \cdot \sum_{k=\lfloor t/S \rfloor S}^{t} \mathbf{V}_k^2 \right\|^{\alpha}$$

$$\overset{(b)}{\leq} (8e \log d)^\alpha S^{\alpha-1} \sum_{k=\lfloor t/S \rfloor S}^{t} \mathbb{E}[\|\mathbf{V}_k^2\|^\alpha] \overset{(c)}{\leq} (8e \log d)^\alpha S^{\alpha-1} \sum_{k=\lfloor t/S \rfloor S}^{t} \mathbb{E}[\|\mathbf{V}_k\|^{2\alpha}]$$

$$\leq \frac{(8e \log d)^{2\alpha} \sigma_{2,\alpha}^2 S^{\alpha-1}}{(n_H^{\lfloor t/S \rfloor S})^\alpha} + \sum_{k=\lfloor t/S \rfloor S+1}^{t} 2^{2\alpha}(8e \log d)^{2\alpha} S^{\alpha-1}(L_2')^{2\alpha} \mathbb{E} \left[\frac{\|\mathbf{x}_k - \mathbf{x}_{k-1}\|^{2\alpha}}{(n_H^k)^\alpha} \right], \tag{125}$$

where (a) is derived by Lemma 4. (b) comes from Jensen's inequality for $\|\cdot\|^\alpha$ and (c) is from $\|AB\| \leq \|A\|\|B\|$.

If we take $n_H^k \geq \frac{2^{1/\alpha}(8e \cdot \log d)^2 \cdot \sigma_{2,\alpha}^{2/\alpha} S^{1-1/\alpha}}{\epsilon^{1/\alpha}}$ samples at checkpoints ($k = \lfloor t/S \rfloor S$) and and at the other iterations, we take $n_H^k \geq \frac{4 \times 2^{1/\alpha}(8e \cdot \log d)^2 \cdot L_2'^2 S \|\boldsymbol{\Delta}_{k-1}\|^2}{\epsilon^{1/\alpha}}$ samples, we have

$$\mathbb{E}[\|\nabla^2 F(\mathbf{x}_t) - \mathbf{U}_t\|^{2\alpha}] \leq \epsilon.$$

$\square$

Theorem 3. *Under gradient dominance property with $\alpha = 1$, Assumptions 1, 3 for $\alpha = 1$, and 12 for $\alpha = 1$, Algorithm 2 can achieve ϵ-global stationary point in expectation by querying $\mathcal{O}(\epsilon^{-2})$ stochastic gradients and $\mathcal{O}(\epsilon^{-1})$ stochastic Hessian on average.*

31

Proof. From Lemma 5, using the estimates of gradient and Hessian from $\mathbf{v}_t$ and $\mathbf{U}_t$, we have for $\alpha = 1$:

$$F(\mathbf{x}_t + \mathbf{\Delta}_t) - F(\mathbf{x}^*) \leq C(F(\mathbf{x}_t) - F(\mathbf{x}_t + \mathbf{\Delta}_t))^{2/3} + C_g \|\nabla F(\mathbf{x}_t) - \mathbf{v}_t\| + C_H \|\nabla^2 F(\mathbf{x}_t) - \mathbf{U}_t\|^2. \tag{126}$$

By defining $\delta_t := \mathbb{E}[F(\mathbf{x}_t)] - F(\mathbf{x}^*)$ and using Jensen's inequality, we can rewrite the above inequality as follows:

$$\delta_{t+1} \leq C(\delta_t - \delta_{t+1})^{2/3} + C_g \mathbb{E}[\|\nabla F(\mathbf{x}_t) - \mathbf{v}_t\|] + C_H \mathbb{E}[\|\nabla^2 F(\mathbf{x}_t) - \mathbf{U}_t\|^2]. \tag{127}$$

Now, suppose for $(k-1)S < t \leq kS$, $k \geq 1$, we set ϵ to $(kS)^{-2}$ in n_g^t's and n_H^t's. Then, from Lemma 12, for all $t \geq 0$,

$$\delta_{t+1} \leq C(\delta_t - \delta_{t+1})^{2/3} + \frac{C_g + C_H}{(\lceil t/S \rceil S)^2}, \tag{128}$$

and from Lemma 11,

$$\delta_T \leq \frac{16(C_g + C_H)C^3}{(\lfloor T/S \rfloor S)^2} + \frac{C_g + C_H}{(\lceil T/S \rceil S)^2}. \tag{129}$$

After

$$T \geq \frac{\left[16(C_g + C_H)C^3 + C_g + C_H\right]^{1/2}}{\epsilon^{1/2}} \tag{130}$$

iterations, $\delta_T \leq \epsilon$. Furthermore, according to Lemma 5 (20), we have:

$$\frac{3M - 2L_2 - 8}{12}\|\mathbf{\Delta}_t\|^3 \leq F(\mathbf{x}_t) - F(\mathbf{x}_t + \mathbf{\Delta}_t) + \frac{2\|\nabla F(\mathbf{x}_t) - \mathbf{g}_t\|^{3/2}}{3} + \frac{\|\nabla^2 F(\mathbf{x}_t) - \mathbf{H}_t\|^3}{6} \tag{131}$$

and using inequality $(a + b + c)^{2/3} \leq 3^{2/3-1}(a^{2/3} + b^{2/3} + c^{2/3})$ for $a, b, c > 0$ and then taking expectation, we have:

$$\mathbb{E}[\|\mathbf{\Delta}_t\|^2] \overset{(a)}{\leq}$$
$$C''[(\mathbb{E}[F(\mathbf{x}_t)] - \mathbb{E}[F(\mathbf{x}_t + \mathbf{\Delta}_t)])^{2/3} + c_g \mathbb{E}\|\nabla F(\mathbf{x}_t) - \mathbf{g}_t\| + c_H \mathbb{E}[\|\nabla^2 F(\mathbf{x}_t) - \mathbf{H}_t\|^2]]$$

$$\implies \sum_{t=(k-1)S}^{kS-1} \mathbb{E}[\|\mathbf{\Delta}_t\|^2] \leq C'' \left[\sum_{t=(k-1)S}^{kS-1} (\delta_t - \delta_{t+1})^{2/3} + c_g \mathbb{E}\|\nabla F(\mathbf{x}_t) - \mathbf{g}_t\| + c_H \mathbb{E}[\|\nabla^2 F(\mathbf{x}_t) - \mathbf{H}_t\|^2] \right]$$

$$\overset{(b)}{\leq} C'' \left[S^{1/3}(\delta_{(k-1)S} - \delta_{kS})^{2/3} + \frac{(c_g + c_H)}{(kS)^{\beta/\alpha}} \right]$$

$$\overset{(c)}{\leq} C'' \left[S^{1/3} \left(\frac{16(C_g + C_H)C^3}{((k-1)S)^2} + \frac{C_g + C_H}{((k-1)S)^2} \right)^{2/3} + \frac{(c_g + c_H)}{(kS)^2} \right] = \mathcal{O}(S^{1/3}(kS)^{-4/3}), \tag{132}$$

where $C'' = \frac{12^{2/3}}{3^{1/3}(3M - 2L_2 - 8)^{2/3}}$, $c_g = \frac{2}{3}$, and $c_H = \frac{1}{6}$ in (a), and in (b), we used Jensen's inequality for the first term in the sum:

$$\frac{1}{S} \sum_{t=(k-1)S}^{kS-1} (\delta_t - \delta_{t+1})^{2/3} \leq \left(\frac{1}{S} \sum_{t=(k-1)S}^{kS-1} \delta_t - \delta_{t+1} \right)^{2/3} = \left(\frac{1}{S}(\delta_{(k-1)S} - \delta_{kS}) \right)^{2/3}$$

and we set n_g^t's and n_H^t's such that

$$\mathbb{E}[\|\nabla F(\mathbf{x}_t) - \mathbf{g}_t\|] \leq \frac{1}{(kS)^2}$$

$$\mathbb{E}[\|\nabla^2 F(\mathbf{x}_t) - \mathbf{H}_t\|^2] \leq \frac{1}{(kS)^2}.$$

32

Moreover, (c) is according to Lemma 11.

Hence, according to Lemma 12 for the case $\alpha = 1$, in order to obtain the following errors:

$$\mathbb{E}[\|\nabla F(\mathbf{x}_t) - \mathbf{g}_t\|]) \leq \frac{1}{(kS)^2}, \quad \mathbb{E}[\|\nabla^2 F(\mathbf{x}_t) - \mathbf{H}_t\|^2] \leq \frac{1}{(kS)^2},$$

the average sample complexity of gradient $\mathbb{E}\left[\sum_{t=1}^{T} n_g^t\right]$ must be in order of

$$\sum_{\mathrm{mod}\ (t,S)=0} \frac{2\sigma_1^2}{t^{-4}} + \sum_{\mathrm{mod}\ (t,S)\neq 0} \frac{4 \cdot L_1'^2 S \mathbb{E}[\|\mathbf{\Delta}_{t-1}\|^2]}{(\lceil t/S \rceil S)^{-4}}. \tag{133}$$

It is enough to have $\mathbb{E}\left[\sum_{t=1}^{T} n_g^t\right]$ be greater than the following upper bound for (133):

$$\sum_{\mathrm{mod}\ (t,S)=0} \frac{2\sigma_1^2}{t^{-4}} + \sum_{\mathrm{mod}\ (t,S)\neq 0} \frac{4 \cdot L_1'^2 S \mathbb{E}[\|\mathbf{\Delta}_{t-1}\|^2]}{(\lceil t/S \rceil S)^4}$$

$$= \sum_{k=1}^{\lceil T/S \rceil} 2\sigma_1^2 (kS)^4 + \sum_{k=1}^{\lceil T/S \rceil} \sum_{t=(k-1)S+1}^{kS} 4 \cdot L_1'^2 S (kS)^4 \mathbb{E}[\|\mathbf{\Delta}_{t-1}\|^2]$$

$$\overset{(a)}{\leq} 2\sigma_1^2 \sum_{k=1}^{\lceil T/S \rceil} (kS)^4 + \sum_{k=1}^{\lceil T/S \rceil} 4 \cdot L_1'^2 S (kS)^4 C'' \left[S^{1/3} \left(\frac{16(C_g + C_H)C^3}{(kS)^2} + \frac{C_g + C_H}{(kS)^2} \right)^{2/3} + \frac{(c_g + c_H)}{(kS)^2} \right]$$

$$\leq 2\sigma_1^2 \frac{T^5}{S} + 4c \cdot L_1'^2 C'' \left[\left[(16(C_g + C_H)C^3) + (C_g + C_H) \right]^{2/3} + \frac{c_g + c_H}{(kS)^{2(1-2/3)}} \right] S^{1/3} T^{5-4/3} \tag{134}$$

where in (a), we used (132). If $S = \lfloor \frac{T}{q} \rfloor$ where q is a constant integer, the average sample complexity of gradient, would be in the order of $\mathcal{O}(T^4)$. As $T \geq \frac{\left[16(C_g+C_H)C^3+C_g+C_H\right]^{1/2}}{\epsilon^{1/2}}$, the average sample complexity of gradient is at least

$$\mathcal{O}\left(\frac{2\sigma_1^2 \left[16(C_g + C_H)C^3 + C_g + C_H\right]^2}{\epsilon^2} \right)$$

in order to get $\delta_T \leq \epsilon$.

Moreover, for the average sample complexity of Hessian $\mathbb{E}\left[\sum_{t=1}^{T} n_H^t\right]$, we have:

$$\sum_{\mathrm{mod}\ (t,S)=0} \frac{(8e \cdot \log d)^2 \cdot \sigma_{2,1}^2}{(t)^{-2}} + \sum_{\mathrm{mod}\ (t,S)\neq 0} \frac{2(8e \cdot \log d)^2 \cdot L_2'^2 S \mathbb{E}[\|\mathbf{\Delta}_{t-1}\|^2]}{(\lceil t/S \rceil S)^{-2}}$$

$$= \sum_{k=1}^{\lceil T/S \rceil} (8e \cdot \log d)^2 \cdot \sigma_{2,1}^2 (kS)^2 + \sum_{k=1}^{\lceil T/S \rceil} \sum_{t=(k-1)S+1}^{kS} 2(8e \cdot \log d)^2 \cdot L_2'^2 S (kS)^2 \mathbb{E}[\|\mathbf{\Delta}_{t-1}\|^2]$$

$$\leq (8e \cdot \log d)^2 \cdot \sigma_{2,1}^2 S^2 \sum_{k=1}^{\lceil T/S \rceil} k^2 +$$

$$\sum_{k=1}^{\lceil T/S \rceil} 2(8e \cdot \log d)^2 \cdot L_2'^2 S (kS)^2 C'' \left[S^{1/3} \left(\frac{16(C_g + C_H)C^3}{(kS)^2} + \frac{C_g + C_H}{(kS)^2} \right)^{2/3} + \frac{(c_g + c_H)}{(kS)^2} \right]$$

$$\leq (8e \cdot \log d)^2 \cdot \sigma_{2,1}^2 T^3 +$$

$$2(8e \cdot \log d)^2 \cdot L_2'^2 S^{1/3} (T)^{3-4/3} C'' \left([16(C_g + C_H)C^3 + (C_g + C_H)]^{2/3} + \frac{c_g + c_H}{(kS)^{2(1-2/3)}} \right) \tag{135}$$

By setting $S = \lfloor \frac{T}{q} \rfloor$ where q is a positive integer constant, the average sample complexity of Hessian, would be in the order of $\mathcal{O}(T^2)$. As $T = \mathcal{O}(\epsilon^{-\frac{1}{2}})$, the average sample complexity of Hessian is in the order of $\mathcal{O}\left(\frac{1}{\epsilon}\right)$.

33

$\square$

A.4   Proof of RL results

Two important classes of policies: 1) Scalar-action, fixed-variance Gaussian policy:

$$\pi_\theta(a|s) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left\{-\frac{1}{2}\left(\frac{a - \theta^T\phi(s)}{\sigma}\right)^2\right\}. \tag{136}$$

2) Softmax tabular policy:

$$\pi_\theta(a|s) = \frac{\exp(\theta_{s,a})}{\sum_{a'\in\mathcal{A}}\exp(\theta_{s,a'})} \tag{137}$$

where $\theta = \{\theta_{s,a}\}_{(s,a)\in\mathcal{S}\times\mathcal{A}}$ are parameters of the policy.

A.4.1   Discussion on gradient dominance property with $\alpha = 1$ for Policies

Assumption 13 (Fisher-non-degenerate.). *For all $\theta \in \mathbb{R}^d$, there exists $\mu_F > 0$ such that the Fisher information matrix $F_\rho(\theta)$ induced by policy π_θ and initial distribution ρ satisfies*

$$F_\rho(\theta) := \mathbb{E}_{(s,a)\sim v_\rho^{\pi_\theta}}[\nabla_\theta\log\pi_\theta(a|s)\nabla_\theta\log\pi_\theta(a|s)^T] \succeq \mu_F I_{d\times d}, \tag{138}$$

where $v_\rho^{\pi_\theta}(s,a) := (1-\gamma)\mathbb{E}_{s_0\sim\rho}\sum_{t=0}^\infty \gamma^t\mathbb{P}(s_t = s, a_t = a|s_0, \pi_\theta)$ is the state-action visitation measure.

This assumption is commonly used in the literature [20, 9]. The Fisher-non-degenerate setting implicitly guarantees that the agent is able to explore the state-action space under the considered policy class. Similar conditions of the Fisher-non-degeneracy is also required in other global optimum convergence framework (Assumption 6.5 in [1] on the relative condition number and Assumption 3 in [6] on the regularity of the parametric model). Assumption 13 is satisfied by a wide families of policies, including the Gaussian policy (136) and certain neural policy. Without the non-degenerate Fisher information matrix condition, the global optimum convergence of more general parameterizations would be hard to analyze without introducing the additional exploration procedures in the non-tabular setting [9, Sec. 8].

Assumption 14. *For all $\theta \in \mathbb{R}^d$, there exists $\epsilon_{bias} > 0$ such that the transferred compatible function approximation error with $(s,a) \sim v_\rho^{\pi_{\theta^*}}$ satisfies*

$$\mathbb{E}[(A^{\pi_\theta}(s,a) - (1-\gamma)u^{*T}\nabla_\theta\log\pi_\theta(a|s))^2] \leq \epsilon_{bias}, \tag{139}$$

where $v_\rho^{\pi_{\theta^*}}$ is the state-action distribution induced by an optimal policy, and $u^* = (F_\rho(\theta))^\dagger\nabla J(\theta)$.

This is also a common assumption [35, 1, 41, 20, 9]. In particular, when π_θ is a soft-max tabular policy (137), ϵ_{bias} is 0 [9]; Moreover, it can be shown that when π_θ is a rich neural policy, ϵ_{bias} is small [35].

By Lemma 4.7 in [9], Assumptions 13 and 14 yield relaxed gradient dominance property with $\alpha = 1$ (See Assumption 6). We provide the proof in the sequel:

From performance difference lemma [14], we have

$$\mathbb{E}_{(s,a)\sim v_\rho^{\pi_{\theta^*}}}[A^{\pi_{\theta_t}}(s,a)] = (1-\gamma)(J_\rho(\theta^*) - J_\rho(\theta_t)), \tag{140}$$

and from Assumption 14 and Jensen's inequality

$$\mathbb{E}[A^{\pi_\theta}(s,a) - (1-\gamma)u^{*T}\nabla_\theta\log\pi_\theta(a|s)] \leq \sqrt{\epsilon_{bias}}. \tag{141}$$

Substituting (140) into (141), we get

$$\begin{aligned}
J_\rho(\theta^*) - J_\rho(\theta_t) &\leq \frac{1}{1-\gamma}\sqrt{\epsilon_{bias}} + \mathbb{E}[u^{*T}\nabla_\theta\log\pi_\theta(a|s)] \\
&\leq \frac{1}{1-\gamma}\sqrt{\epsilon_{bias}} + \frac{G_1}{\mu_F}\|\nabla J_\rho(\theta_t)\|,
\end{aligned} \tag{142}$$

34

where the last inequality comes from following Cauchy-Schwartz inequality

$$\mathbb{E}[u^{*T}\nabla_\theta \log \pi_\theta(a|s)] = \mathbb{E}[\nabla J(\theta)^T (F_\rho(\theta))^\dagger \nabla_\theta \log \pi_\theta(a|s)]$$

$$\leq \mathbb{E}\|\nabla J(\theta)\|_2 \mathbb{E}\|(F_\rho(\theta))^\dagger \nabla_\theta \log \pi_\theta(a|s)\|_2 \leq \frac{1}{\mu_F}\mathbb{E}\|\nabla_\theta \log \pi_\theta(a|s)\|_2 \mathbb{E}\|\nabla J(\theta)\|_2 \leq \frac{G_1}{\mu_F}\mathbb{E}\|\nabla J(\theta)\|_2$$

where two last inequalities are from Assumptions 13 and 7. Thus, by setting $\epsilon' := \frac{1}{1-\gamma}\sqrt{\epsilon_{bias}}$ and $\tau_J := \frac{G}{\mu_F}$,

$$J^* - J(\theta_t) \leq \tau_J\|\nabla J(\theta_t)\|_2 + \epsilon'. \tag{143}$$

In the past few years, several work have attempted to establish that PG and SPG converge to a global optimal point for various classes of policies [43, 44]. For instance, for Fisher non-degenerate policies, Liu et al. [20] proposed a variance-reduced SPG method that converges to a global optimal point with sample complexity of $\tilde{\mathcal{O}}(\epsilon^{-3})$. By considering a momentum term, Ding et al. [9] showed that the sample complexities of SPG for the soft-max policy and Fisher-non-degenerate policies are in the order of $\tilde{\mathcal{O}}(\epsilon^{-4.5})$ and $\tilde{\mathcal{O}}(\epsilon^{-3})$, respectively. In the case of weak gradient dominant functions with $\alpha = 1$ (See Assumption 6), Yuan et al. [41] showed that the sample complexity of SPG is bounded by $\tilde{\mathcal{O}}(\epsilon^{-3})$.

Theorem 4. *For a policy π_θ satisfying Assumptions 7, 8, and the corresponding objective function $J(\theta)$ satisfying Assumption 6, SCRN outputs the solution θ_T such that $J^* - \mathbb{E}[J(\theta_T)] \leq \epsilon + \epsilon'$ and the sample complexity (the number of observed state action pairs) is: $T \times m \times \mathsf{H} = \tilde{\mathcal{O}}(\epsilon^{-2.5})$ for $\epsilon' = 0$ and $T \times m \times \mathsf{H} = \tilde{\mathcal{O}}(\epsilon^{-0.5}\epsilon'^{-2})$ for $\epsilon' > 0$.*

A.4.2   Proof of Theorem 4

In order to obtain Lipschitz Hessian for $J(\theta)$, we use the following lemma in [46].

Lemma 13. *Under Assumptions 7, and 8, we have*

$$\|\nabla^2 J(\theta) - \nabla^2 J(\theta')\| \leq \tilde{L}\|\theta - \theta'\|_2, \tag{144}$$

where

$$\tilde{L} = \frac{R_{max}G_1G_2}{(1-\gamma)^2} + \frac{R_{max}G_1^3(1+\gamma)}{(1-\gamma)^3} + \frac{R_{max}G_1}{1-\gamma}\max\left\{G_2, \frac{\gamma G_1^2}{1-\gamma}, \frac{\bar{L}_2}{G_1}, \frac{G_2\gamma}{1-\gamma}, \frac{G_1(1+\gamma)+G_2\gamma(1-\gamma)}{1-\gamma^2}\right\}.$$

It is good to mention that $\tilde{L} = \mathcal{O}(1/(1-\gamma)^3)$. Using (21) for $\alpha = 1$ in Lemma 5 for $F(\theta) := -J(\theta)$, we get

$$J^* - J(\theta_{t+1}) - \epsilon' \leq \tau_J\|\nabla J(\theta_{t+1})\|$$

$$\leq C(J(\theta_{t+1}) - J(\theta_t))^{2/3} + C_g\|\nabla J(\theta_t) - \hat{\nabla}_m J(\theta_t)\| + C_H\|\nabla^2 J(\theta_t) - \hat{\nabla}_m^2 J(\theta_t)\|^2, \tag{145}$$

where

$$C = 3^{1/3}\tau_F\left(\frac{M + \tilde{L} + 1}{2}\right)\left(\frac{12}{3M - 2\tilde{L} - 8}\right)^{2/3}$$

$$C_g = 2^{2/3} \times 3^{\frac{-2}{3}}\tau_J\left(\frac{M + \tilde{L} + 1}{2}\right)\left(\frac{12}{3M - 2\tilde{L} - 8}\right)^{2/3} + \tau_J \tag{146}$$

$$C_H = 2^{-2/3} \times 3^{\frac{-2}{3}}\tau_J\left(\frac{M + \tilde{L} + 1}{2}\right)\left(\frac{12}{3M - 2\tilde{L} - 8}\right)^{2/3} + 2^{-1}\tau_J.$$

We know that $\hat{\nabla}_m J(\theta_t)$ and $\hat{\nabla}_m^2 J(\theta_t)$ are in fact unbiased estimators of $\nabla J_\mathsf{H}(\theta_t)$ and $\nabla^2 J_\mathsf{H}(\theta_t)$. Thus,

$$J^* - J(\theta_{t+1}) \leq C(J(\theta_{t+1}) - J(\theta_t))^{2/3} + C_g\|\nabla J_\mathsf{H}(\theta_t) - \hat{\nabla}_m J(\theta_t)\| + 2C_H\|\nabla^2 J_\mathsf{H}(\theta_t) - \hat{\nabla}_m^2 J(\theta_t)\|^2 + \epsilon'$$

$$+ C_g\|\nabla J(\theta_t) - \nabla J_\mathsf{H}(\theta_t)\| + 2C_H\|\nabla^2 J(\theta_t) - \nabla^2 J_\mathsf{H}(\theta_t)\|^2$$

$$\leq C(J(\theta_{t+1}) - J(\theta_t))^{2/3} + C_g\|\nabla J_\mathsf{H}(\theta_t) - \hat{\nabla}_m J(\theta_t)\| + 2C_H\|\nabla^2 J_\mathsf{H}(\theta_t) - \hat{\nabla}_m^2 J(\theta_t)\|^2 + \epsilon'$$

$$+ C_g D_g\gamma^\mathsf{H} + 2C_H D_H\gamma^{2\mathsf{H}}.$$

35

where C_g and C_H is defined in Lemma 3. We take a expectation from both sides given θ_t and use Jensen's inequality as follows:

$$J^* - \mathbb{E}J(\theta_{t+1}) \le C(\mathbb{E}J(\theta_{t+1}) - \mathbb{E}J(\theta_t))^{2/3} + C_g\mathbb{E}[\|\nabla J_{\mathsf{H}}(\theta_t) - \hat{\nabla}_m J(\theta_t)\|]$$
$$+ 2C_H\mathbb{E}\left[\|\nabla^2 J_{\mathsf{H}}(\theta_t) - \hat{\nabla}_m^2 J(\theta_t)\|^2\right] + C_g D_g \gamma^{\mathsf{H}} + 2C_H D_H \gamma^{2\mathsf{H}} + \epsilon'. \tag{147}$$

We use the following lemma to bound the error terms of the gradient estimator and the Hessian estimator.

Lemma 14. *Under Assumption 7, we have*

$$\mathbb{E}\|\hat{\nabla}_m J(\theta) - \nabla J_{\mathsf{H}}(\theta)\|^2 \le \frac{\mathsf{H}G_1^2 R_{\max}^2}{m(1-\gamma)^2},$$
$$\mathbb{E}\|\nabla^2 J_{\mathsf{H}}(\theta) - \hat{\nabla}_m^2 J(\theta)\|^2 \le \frac{4e \cdot \max\{2, \log d\} R_{\max}^2(\mathsf{H}^2 G_1^4 + G_2^2)}{m(1-\gamma)^4}. \tag{148}$$

The proof of Lemma 14 is given in Appendix A.4.3.

Now we have from Jensen's inequality

$$\mathbb{E}\|\nabla J_{\mathsf{H}}(\theta_t) - \hat{\nabla}_m J(\theta_t)\| \le \left(\mathbb{E}\|\nabla J_{\mathsf{H}}(\theta_t) - \hat{\nabla}_m J(\theta_t)\|^2\right)^{1/2} \le \frac{\mathsf{H}G R_{\max}}{(1-\gamma)\sqrt{m}}.$$

Then we get

$$J^* - \mathbb{E}J(\theta_{t+1}) \le C(\mathbb{E}J(\theta_{t+1}) - \mathbb{E}J(\theta_t))^{2/3} + C_g\frac{\mathsf{H}G R_{\max}}{(1-\gamma)\sqrt{m}} + 2C_H\frac{4e \cdot \max\{2, \log d\} R_{\max}^2(F^2 + G^4)}{(1-\gamma)^4 m}$$
$$+ \epsilon' + C_g D_g \gamma^{\mathsf{H}} + 2C_H D_H \gamma^{2\mathsf{H}}. \tag{149}$$

Let define a stationary value for $J^* - \mathbb{E}J(\theta_t)$ as $P(m, \epsilon')$ which is as follows:

$$P(m, \epsilon') := C_g\frac{G R_{\max}}{(1-\gamma)^{3/2}\sqrt{m}} + 2C_H\frac{4e \cdot \max\{2, \log d\} R_{\max}^2(F^2 + G^4)}{(1-\gamma)^4 m} + \epsilon' + C_g D_g \gamma^{\mathsf{H}} + 2C_H D_H \gamma^{2\mathsf{H}}. \tag{150}$$

We define:

$$\delta_t := \frac{J^* - \mathbb{E}J(\theta_t) - P(m, \epsilon')}{C^3}.$$

Then we get following recursion from (149):

$$\delta_{t+1} \le (\delta_t - \delta_{t+1})^{2/3}.$$

We know from (42) in the proof of Theorem 1 that

$$\delta_T \le \frac{4}{(DT)^2} \tag{151}$$

where $D = \min\{1/2, 2(2^{1/3} - 1)\delta_0^{-1/2}\}$. Hence,

$$J^* - \mathbb{E}J(\theta_T) - P(m, \epsilon) \le \frac{4C^3}{(DT)^2}.$$

Therefore, $J^* - \mathbb{E}J(\theta_t)$ converges to $P(m, \epsilon)$ with the convergence rate of $\mathcal{O}(1/T^2)$. Let

$$m \ge \max\left\{\frac{C_g^2 \mathsf{H}^2 G_1^2 R_{\max}^2}{(1-\gamma)^2 \epsilon'^2}, \frac{C_H 4e \cdot \max\{2, \log d\} R_{\max}^2(G_2^2 + G_1^4)}{(1-\gamma)^4 \epsilon'}\right\},$$

$$\mathsf{H} \ge \max\left\{\frac{\log\left(\frac{C_g D_g}{\epsilon'}\right)}{\log(1/\gamma)}, \frac{\log\left(\frac{2C_H D_H}{\epsilon'}\right)}{2\log(1/\gamma)}\right\}.$$

Then, θ_T satisfies:

$$J^* - \mathbb{E}J(\theta_T) \le 5\epsilon' + \epsilon \tag{152}$$

36

where $T \geq \frac{2C^{3/2}D}{\sqrt{\epsilon}}$. Finally, the number of observed state-action pairs are required to get (152) is as follows:

$$T \times \mathsf{H} \times m \geq \mathcal{O}\left(\frac{C^{3/2}DC_g^2 G_1^2 R_{\max}^2 \log^2\left(\frac{C_g D_g}{\epsilon'}\right)}{(1-\gamma)^2 \log^2(1/\gamma)\epsilon'^2 \epsilon^{0.5}}\right) \overset{(a)}{=} \mathcal{O}\left(\frac{G_1^{2+\frac{7}{6}} G_2^{\frac{7}{6}} R_{\max}^{2+\frac{7}{6}}}{(1-\gamma)^{4+\frac{7}{3}} \epsilon'^2 \epsilon^{0.5}}\right) \quad (153)$$

where (a) comes from $C = \mathcal{O}(M^{1/3})$ and $C_g = \mathcal{O}(M^{1/3})$ and the fact that $M = \mathcal{O}(\tilde{L})$ and from Lemma 13, $\tilde{L} = \frac{R_{\max} G_1 G_2}{(1-\gamma)^2}$.

A.4.3 Proof of Lemma 14

In [41, Lemma 4.2], it has been shown that

$$\mathbb{E}\|\hat{\nabla}_m J(\theta) - \nabla J_{\mathsf{H}}(\theta)\|^2 \leq \frac{\mathsf{H}G_1^2 R_{\max}^2}{m(1-\gamma)^2}. \quad (154)$$

We know from [29, Lemma 4.1] that

$$\|\hat{\nabla}^2(\theta, \tau)\|^2 \leq \frac{R_{\max}^2(\mathsf{H}^2 G_1^4 + G_2^2)}{(1-\gamma)^4}, \quad (155)$$

where $\hat{\nabla}^2(\theta, \tau) := \nabla\Phi(\theta; \tau)\nabla \log p(\tau|\pi_\theta)^T + \nabla^2\Phi(\theta; \tau)$. We have an upper bound on the variance of Hessian of value function for each trajectory as follows:

$$\mathbb{E}[\|\nabla^2 J_{\mathsf{H}}(\theta) - \hat{\nabla}^2(\theta, \tau)\|^2] \leq \mathbb{E}[\|\hat{\nabla}^2(\theta, \tau)\|^2] \leq \frac{R_{\max}^2(\mathsf{H}^2 G_1^4 + G_2^2)}{(1-\gamma)^4}.$$

Denote $\sigma_{2,1}^2 := \frac{R_{\max}^2(\mathsf{H}^2 G_1^4 + G_2^2)}{(1-\gamma)^4}$. Then from Equation (18), we can get

$$\mathbb{E}[\|\nabla^2 J_{\mathsf{H}}(\theta) - \hat{\nabla}_m^2 J(\theta)\|^2] \leq \frac{4e \cdot \max\{2, \log d\}R_{\max}^2(\mathsf{H}^2 G_1^4 + G_2^2)}{m(1-\gamma)^4} \quad (156)$$

A.4.4 Soft-max policies satisfy Lipschitz Hessian

Recall that soft-max tabular policy as follows:

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

For any $(s, a, a') \in \mathcal{S} \times \mathcal{A} \times \mathcal{A}$ with $a' \neq a$, we get the following partial derivatives for the soft-max tabular policy

$$\frac{\partial \pi_\theta(a|s)}{\partial \theta_{s,a}} = \pi_\theta(a|s)(1 - \pi_\theta(a|s)), \quad \frac{\partial \pi_\theta(a|s)}{\partial \theta_{s,a'}} = -\pi_\theta(a|s)\pi_\theta(a'|s). \quad (157)$$

Note that for $s' \in \mathcal{S}$ with $s' \neq s$, we have $\frac{\partial \pi_\theta(a|s)}{\partial \theta_{s',a}} = 0$. We denote $\theta_s = [\theta_{s,a}]_{a\in\mathcal{A}} \in \mathbb{R}^{|\mathcal{A}|}$. We obtain the gradient and Hessian of $\log \pi_\theta(a|s)$ w.r.t. θ_s as follows:

$$\frac{\partial \log \pi_\theta(a|s)}{\partial \theta_s} = \mathbf{1}_a - \pi_\theta(\cdot|s), \quad \frac{\partial^2 \log \pi_\theta(a|s)}{\partial \theta_s^2} = \pi_\theta(\cdot|s)\pi_\theta(\cdot|s)^T - \text{Diag}(\pi_\theta(\cdot|s)) \quad (158)$$

where $\mathbf{1}_a \in \mathbb{R}^{|\mathcal{A}|}$ is a vector with zero entries except one non-zero entry 1 corresponding to the action a and $\pi_\theta(\cdot|s) = [\pi_\theta(a|s)]_{a\in\mathcal{A}} \in \mathbb{R}^{|\mathcal{A}|}$ is a vector consists of all policies with the same state s. Now we have

$$\|\nabla^2 \log \pi_{\theta'}(a|s) - \nabla^2 \log \pi_\theta(a|s)\| = \left\|\frac{\partial^2}{\partial \theta_s^2} \log \pi_{\theta'}(a|s) - \frac{\partial^2}{\partial \theta_s^2} \log \pi_\theta(a|s)\right\|$$

$$\leq \underbrace{\left\|\pi_{\theta'}(\cdot|s)\pi_{\theta'}(\cdot|s)^T - \pi_\theta(\cdot|s)\pi_\theta(\cdot|s)^T\right\|}_{(1)} + \underbrace{\left\|\text{Diag}(\pi_{\theta'}(\cdot|s)) - \text{Diag}(\pi_\theta(\cdot|s))\right\|}_{(2)}. \quad (159)$$

37

The first term (1) can be bounded as follows:

$$
\begin{aligned}
(1) &= \left\| \pi_\theta(\cdot|s)\pi_\theta(\cdot|s)^T - \pi_{\theta'}(\cdot|s)\pi_\theta(\cdot|s)^T + \pi_{\theta'}(\cdot|s)\pi_\theta(\cdot|s)^T - \pi_{\theta'}(\cdot|s)\pi_{\theta'}(\cdot|s)^T \right\| \\
&\leq (\|\pi_\theta(\cdot|s)\|_2 + \|\pi_{\theta'}(\cdot|s)\|_2)\|\pi_\theta(\cdot|s) - \pi_{\theta'}(\cdot|s)\|_2 \\
&\overset{(a)}{\leq} 2\|\pi_\theta(\cdot|s) - \pi_{\theta'}(\cdot|s)\|_2 \overset{(b)}{\leq} 4\|\theta - \theta'\|_2.
\end{aligned}
\tag{160}
$$

where (a) comes from the fact that as $\sum_a \pi(a|s) = 1$ then the $\|\pi_\theta(\cdot|s)\|_2 \leq 1$ is less than one. Inequality (b) is derived by $\max_\theta \|\nabla_\theta \pi_\theta(a|s)\| \leq 2$ which yields $\pi_\theta(a,s)$ is 2-Lipschitz.

The second term (2) is bounded by $2\|\theta - \theta'\|_2$. Hence, $\log \pi_\theta(a|s)$ for soft-max policy is 6-Lipschitz Hessian.

Lemma 2. *Under Assumption 7, we have $\|\nabla J(\theta) - \nabla J_{\mathsf{H}}(\theta)\| \leq D_g \gamma^{\mathsf{H}}$ and $\|\nabla^2 J(\theta) - \nabla^2 J_{\mathsf{H}}(\theta)\| \leq D_H \gamma^{\mathsf{H}}$, where $D_g = \frac{G_1 R_{\max}}{1-\gamma}\sqrt{\frac{1}{1-\gamma} + \mathsf{H}}$ and $D_H = \frac{R_{\max}(G_2 + G_1^2)}{1-\gamma}\left(\mathsf{H} + \frac{1}{1-\gamma}\right)$.*

A.4.5 Proof of Lemma 3

Under Assumption 7, it has been shown that $D_g = \frac{G_1 R_{\max}}{1-\gamma}\sqrt{\frac{1}{1-\gamma} + \mathsf{H}}$ [41, Lemma 4.5]. We will show that under Assumption 7, $D_H = \frac{R_{\max}(G_2 + G_1^2)}{1-\gamma}\left(\mathsf{H} + \frac{1}{1-\gamma}\right)$.

It can be shown that [41, Proof of Lemma 4.4]

$$
\begin{aligned}
\nabla^2 J(\theta) &= \mathbb{E}_\tau \left[\sum_{t'=0}^{\infty} \nabla_\theta \log \pi_\theta(a_{t'}|s_{t'}) \left(\sum_{t=0}^{\infty} \gamma^t R(s_t, a_t) \left(\sum_{k=0}^{t} \nabla_\theta \log \pi_\theta(a_k|s_k) \right) \right)^T \right] \\
&+ \mathbb{E}_\tau \left[\sum_{t=0}^{\infty} \gamma^t R(s_t, a_t) \left(\sum_{k=0}^{t} \nabla_\theta^2 \log \pi_\theta(a_k|s_k) \right) \right]
\end{aligned}
$$

and then

$$
\begin{aligned}
\nabla^2 J_{\mathsf{H}}(\theta) &= \mathbb{E}_\tau \left[\sum_{t'=0}^{\mathsf{H}-1} \nabla_\theta \log \pi_\theta(a_{t'}|s_{t'}) \left(\sum_{t=0}^{\mathsf{H}-1} \gamma^t R(s_t, a_t) \left(\sum_{k=0}^{t} \nabla_\theta \log \pi_\theta(a_k|s_k) \right) \right)^T \right] \\
&+ \mathbb{E}_\tau \left[\sum_{t=0}^{\mathsf{H}-1} \gamma^t R(s_t, a_t) \left(\sum_{k=0}^{t} \nabla_\theta^2 \log \pi_\theta(a_k|s_k) \right) \right].
\end{aligned}
$$

Therefore,

$$
\begin{aligned}
\nabla^2 J(\theta) - \nabla^2 J_{\mathsf{H}}(\theta) &= \underbrace{\mathbb{E}_\tau \left[\sum_{t=\mathsf{H}}^{\infty} \gamma^t R(s_t, a_t) \left(\sum_{k=0}^{t} \nabla_\theta^2 \log \pi_\theta(a_k|s_k) \right) \right]}_{(1)} \\
&+ \underbrace{\mathbb{E}_\tau \left[\sum_{t'=\mathsf{H}}^{\infty} \nabla_\theta \log \pi_\theta(a_{t'}|s_{t'}) \left(\sum_{t=0}^{\infty} \gamma^t R(s_t, a_t) \left(\sum_{k=0}^{t} \nabla_\theta \log \pi_\theta(a_k|s_k) \right) \right)^T \right]}_{(2)} \\
&+ \underbrace{\mathbb{E}_\tau \left[\sum_{t'=0}^{\mathsf{H}-1} \nabla_\theta \log \pi_\theta(a_{t'}|s_{t'}) \left(\sum_{t=\mathsf{H}}^{\infty} \gamma^t R(s_t, a_t) \left(\sum_{k=0}^{t} \nabla_\theta \log \pi_\theta(a_k|s_k) \right) \right)^T \right]}_{(3)}
\end{aligned}
$$

38

First we bound $\|(1)\|_2$ as follows:

$$\|(1)\|_2 \leq \mathbb{E}_\tau \left[\sum_{t=\mathsf{H}}^\infty \gamma^t |R(s_t, a_t)| \left(\sum_{k=0}^t \|\nabla_\theta \log \pi_\theta(a_k|s_k)\| \right) \right]$$

$$\leq R_{\max} \left[\sum_{t=\mathsf{H}}^\infty \gamma^t \left(\sum_{k=0}^t \mathbb{E}_\tau \|\nabla_\theta \log \pi_\theta(a_k|s_k)\| \right) \right]$$

$$\leq G_2 R_{\max} \sum_{t=\mathsf{H}}^\infty \gamma^t (1+t) = \gamma^\mathsf{H} G_2 R_{\max} \sum_{t=0}^\infty \gamma^t (1+t+\mathsf{H}) = \gamma^H G_2 R_{\max} \left(\frac{1}{(1-\gamma)^2} + \frac{\mathsf{H}}{1-\gamma} \right).$$

To bound the second term (2), it is good to note that for $0 \leq t < t'$

$$\mathbb{E}_\tau \left[\nabla_\theta \log \pi_\theta(a_{t'}|s_{t'}) \left(\gamma^t R(s_t, a_t) \left(\sum_{k=0}^t \nabla_\theta \log \pi_\theta(a_k|s_k) \right) \right)^T \right] = 0.$$

Because

$$\mathbb{E}_\tau \left[\nabla_\theta \log \pi_\theta(a_{t'}|s_{t'}) \left(\gamma^t R(s_t, a_t) \left(\sum_{k=0}^t \nabla_\theta \log \pi_\theta(a_k|s_k) \right) \right)^T \right]$$

$$= \mathbb{E}_{s_{0:t'}, a_{0:(t'-1)}} \left[\underbrace{\mathbb{E}_{a_{t'}} \left[\nabla_{\theta'} \log \pi_\theta(a_{t'}|s_{t'})|s_{t'} \right]}_{=0} \cdot \left(\gamma^t R(s_t, a_t) \left(\sum_{k=0}^t \nabla_\theta \log \pi_\theta(a_k|s_k) \right) \right)^T \right] = 0$$

$$\tag{161}$$

where

$$\mathbb{E}_{a_{t'}} \left[\nabla_{\theta'} \log \pi_\theta(a_{t'}|s_{t'})|s_{t'} \right] = \int \nabla_\theta \pi_\theta(a_{t'}|s_{t'}) da_{t'} = \nabla_\theta \underbrace{\int \pi_\theta(a_{t'}|s_{t'}) da_{t'}}_{=1} = 0.$$

Hence, we get

$$\|(2)\| = \left\| \mathbb{E}_\tau \left[\sum_{t'=\mathsf{H}}^t \nabla_\theta \log \pi_\theta(a_{t'}|s_{t'}) \left(\sum_{t=\mathsf{H}}^\infty \gamma^t R(s_t, a_t) \left(\sum_{k=0}^t \nabla_\theta \log \pi_\theta(a_k|s_k) \right) \right)^T \right] \right\|$$

$$\overset{(a)}{=} \left\| \mathbb{E}_\tau \left[\sum_{t=\mathsf{H}}^\infty \gamma^t R(s_t, a_t) \left(\sum_{t'=\mathsf{H}}^t \nabla_\theta \log \pi_\theta(a_{t'}|s_{t'}) \right) \left(\sum_{k=\mathsf{H}}^t \nabla_\theta \log \pi_\theta(a_k|s_k) \right)^T \right] \right\|$$

$$\leq R_{\max} \sum_{t=\mathsf{H}}^\infty \gamma^t \mathbb{E}_\tau \left[\left\| \sum_{k=\mathsf{H}}^t \nabla_\theta \log \pi_\theta(a_k|s_k) \right\|^2 \right]$$

$$\overset{(b)}{=} R_{\max} \sum_{t=\mathsf{H}}^\infty \gamma^t \sum_{k=\mathsf{H}}^t \mathbb{E}_\tau \left[\|\nabla_\theta \log \pi_\theta(a_k|s_k)\|^2 \right]$$

$$\leq R_{\max} G_1^2 \sum_{t=\mathsf{H}}^\infty \gamma^t (t - \mathsf{H} + 1) = R_{\max} G_1^2 \gamma^\mathsf{H} \sum_{t=0}^\infty \gamma^t (t+1) \leq \frac{G_1^2 R_{\max} \gamma^\mathsf{H}}{(1-\gamma)^2}$$

where (a) comes from the similar argument to (161) and (b) comes from the fact that for any $k \neq k'$

$$\mathbb{E}_\tau \left[\nabla_\theta \log \pi_\theta(a_k|s_k)^T \nabla_\theta \log \pi_\theta(a_{k'}|s_{k'}) \right] = 0. \tag{162}$$

39

To bound the third term (3), first, with the similar argument to (161), we can rewrite (3) as follows:

$$\|(3)\| = \left\| \mathbb{E}_\tau \left[\sum_{t'=0}^{\mathsf{H}-1} \nabla_\theta \log \pi_\theta(a_{t'}|s_{t'}) \left(\sum_{t=\mathsf{H}}^{\infty} \gamma^t R(s_t, a_t) \left(\sum_{k=0}^{t} \nabla_\theta \log \pi_\theta(a_k|s_k) \right) \right)^T \right] \right\|$$

$$= \left\| \mathbb{E}_\tau \left[\sum_{t=\mathsf{H}}^{\infty} \gamma^t R(s_t, a_t) \cdot \left(\sum_{t'=0}^{\mathsf{H}-1} \nabla_\theta \log \pi_\theta(a_{t'}|s_{t'}) \right) \cdot \left(\sum_{k=0}^{t} \nabla_\theta \log \pi_\theta(a_k|s_k) \right)^T \right] \right\|$$

$$\overset{(a)}{=} \left\| \mathbb{E}_\tau \left[\sum_{t=\mathsf{H}}^{\infty} \gamma^t R(s_t, a_t) \cdot \left(\sum_{t'=0}^{\mathsf{H}-1} \nabla_\theta \log \pi_\theta(a_{t'}|s_{t'}) \right) \cdot \left(\sum_{k=0}^{\mathsf{H}-1} \nabla_\theta \log \pi_\theta(a_k|s_k) \right)^T \right] \right\|$$

$$\leq R_{\max} \sum_{t=\mathsf{H}}^{\infty} \gamma^t \cdot \mathbb{E}_\tau \left[\left\| \sum_{k=0}^{\mathsf{H}-1} \nabla_\theta \log \pi_\theta(a_k|s_k) \right\|^2 \right]$$

$$\leq R_{\max} \sum_{t=\mathsf{H}}^{\infty} \gamma^t \cdot \sum_{k=0}^{\mathsf{H}-1} \mathbb{E}_\tau \left[\| \nabla_\theta \log \pi_\theta(a_k|s_k) \|^2 \right] \leq R_{\max} G_1^2 \sum_{t=\mathsf{H}}^{\infty} \gamma^t \mathsf{H} = \frac{\gamma^{\mathsf{H}} R_{\max} G_1^2 \mathsf{H}}{1-\gamma}$$

where (a) comes from (162).

Using triangle inequality and combining all bounds for (1), (2), and (3), we have

$$\|\nabla^2 J(\theta) - \nabla^2 J_{\mathsf{H}}(\theta)\| \leq \|(1)\| + \|(2)\| + \|(3)\|$$

$$\leq \gamma^{\mathsf{H}} \left(G_2 R_{\max} \left(\frac{1}{(1-\gamma)^2} + \frac{\mathsf{H}}{1-\gamma} \right) + \frac{G_1^2 R_{\max}}{(1-\gamma)^2} + \frac{R_{\max} G_1^2 \mathsf{H}}{1-\gamma} \right). \tag{163}$$

A.4.6 Analysis of the variance-reduced SCRN under weak gradient dominance property with $\alpha = 1$

Definition 3 (Importance Sampling). *For any given trajectory τ and two parameters $\theta, \theta' \in \mathbb{R}^d$, we define the importance sampling weight as*

$$w(\tau|\theta', \theta) = \frac{p(\tau|\pi_{\theta'})}{p(\tau|\pi_\theta)} = \prod_{i=1}^{\mathsf{H}} \frac{\pi_{\theta'}(a_h|s_h)}{\pi_\theta(a_h|s_h)}, \tag{164}$$

where H is the horizon length.

Assumption 15. *There exists a constant $W \in (0, \infty)$, such that, for any pair of policies generated by Algorithm 3, the following holds*

$$\mathrm{Var}(w(\tau|\theta', \theta)) \leq W, \quad \theta, \theta' \in \mathbb{R}^d, \ \tau \sim p(\cdot|\pi_{\theta'}). \tag{165}$$

In following, we present a version of Algorithm 2 adapted for RL setting by utilizing importance sampling. In what follows, $\hat{\nabla} J(\theta'; \tau)$ and $\hat{\nabla}^2 J(\theta'; \tau)$ are gradient and Hessian estimators of the value function, respectively which are sampled from distribution $p(\tau|\pi_\theta)$ for some θ.

Assumptions 3 with $\alpha = 1$ and Assumption 12 are satisfied in the considered RL setting. To see this, $\|\hat{\nabla}^2 J(\theta, \tau)\|^2 \leq \frac{\mathsf{H}^2 G_1^4 R_{\max}^2 + G_2^2 R_{\max}^2}{(1-\gamma)^4}$ from [29, Lemma 4.1]. Then, $\hat{\nabla} J(\theta, \tau)$ is Lipschitz. Moreover, similar to the arguments in Lemma 13, one can show that $\|\hat{\nabla}^2 J(\theta, \tau) - \hat{\nabla}^2 J(\theta', \tau)\| \leq \tilde{L}\|\theta - \theta'\|_2$. Hence, we can imply that Lipschitzness of $\hat{\nabla} J(\theta, \tau)$ and $\hat{\nabla}^2 J(\theta, \tau)$ yield Assumption 12. Furthermore, from proof of Lemma 3 (Section A.4.3), the variances of $\hat{\nabla} J(\theta, \tau)$ and $\hat{\nabla}^2 J(\theta, \tau)$ are bounded.

Now, with

$$\mathsf{H} \geq \max \left\{ \frac{\log\left(\frac{C_g D_g}{\epsilon} \right)}{\log(1/\gamma)}, \frac{\log\left(\frac{2 C_H D_H}{\epsilon} \right)}{2 \log(1/\gamma)} \right\},$$

Equation (147) can be rewritten as follows:

$$J^* - \mathbb{E}J(\theta_{t+1}) \leq C(\mathbb{E}J(\theta_{t+1}) - \mathbb{E}J(\theta_t))^{2/3} + C_g \mathbb{E}[\|\nabla J_{\mathsf{H}}(\theta_t) - \mathbf{v}_t\|]$$
$$+ 2 C_H \mathbb{E}\left[\|\nabla^2 J_{\mathsf{H}}(\theta_t) - \mathbf{U}_t\|^2 \right] + 2\epsilon. \tag{166}$$

40

Algorithm 3 IS-VR stochastic cubic regularized Newton method

Input: Maximum number of iterations T, batch sizes $\{n_g^t\}_{t=1}^T$, $\{n_H^t\}_{t=1}^T$, the period length S, initial point θ_0, and cubic penalty parameter M.

1: $t \leftarrow 0$
2: **while** $t \leq T$ **do**
3:      Sample trajectory set $\mathcal{J}_t = \{\tau_j^t\}_{j=1}^{n_g^t}$ and trajectory set $\mathcal{I}_t = \{\tau_i^t\}_{i=1}^{n_H^t}$ according to policy $p(\tau|\pi_{\theta_t})$
4:
$$
\mathbf{v}_t \leftarrow \begin{cases} \frac{1}{n_g^t} \sum_{j=1}^{n_g^t} \hat{\nabla} J(\theta_t; \tau_j^t), & \mod(t,S) = 0 \\ \frac{1}{n_g^t} \sum_{j=1}^{n_g^t} [\hat{\nabla} J(\theta_t; \tau_j^t) - w(\tau_j^t|\theta_{t-1}, \theta_t) \hat{\nabla} J(\theta_{t-1}; \tau_j^t)] + \mathbf{v}_{t-1}, & \text{else} \end{cases}
$$
5:
$$
\mathbf{U}_t \leftarrow \begin{cases} \frac{1}{n_H^t} \sum_{i=1}^{n_H^t} \hat{\nabla}^2 J(\theta_t; \tau_i^t), & \mod(t,S) = 0 \\ \frac{1}{n_H^t} \sum_{i=1}^{n_H^t} [\hat{\nabla}^2 J(\theta_t; \tau_i^t) - w(\tau_i^t|\theta_{t-1}, \theta_t) \hat{\nabla}^2 J(\theta_{t-1}; \tau_i^t)] + \mathbf{U}_{t-1}, & \text{else} \end{cases}
$$
6:      $\mathbf{\Delta}_t \leftarrow \arg\min_{\mathbf{\Delta} \in \mathbb{R}^d} \langle \mathbf{v}_t, \mathbf{\Delta} \rangle + \frac{1}{2} \langle \mathbf{\Delta}, \mathbf{U}_t \mathbf{\Delta} \rangle + \frac{M}{6} \|\mathbf{\Delta}\|^3$
7:      $\theta_{t+1} \leftarrow \theta_t + \mathbf{\Delta}_t$
8:      $t \leftarrow t + 1$
9: **end while**
10: **return** θ_t

Lemma 15 (Lemma 6.1 in [38]). *For a policy π_θ satisfying Assumptions 7 and 15, and for any $\theta, \theta' \in \mathbb{R}^d$, we have*

$$
\mathrm{Var}(w(\tau|\theta, \theta')) \leq C_w \|\theta - \theta'\|^2, \tag{167}
$$

where $C_w = \mathsf{H}(2\mathsf{H}G_1^2 + G_2)(W + 1)$.

In order to prove Theorem 3, we must show that with the following number of samples at checkpoints:

$$
n_g^t = \Theta\left(\frac{1}{\epsilon^2}\right), n_H^t = \Theta\left(\frac{(\log d)^2}{\epsilon}\right), \tag{168}
$$

and the following number of samples at the other iterations:

$$
n_g^t(\theta_t, \theta_{t-1}) = \Theta\left(\frac{S\|\mathbf{\Delta}_{t-1}\|^2}{\epsilon^2}\right), n_H^t(\theta_t, \theta_{t-1}) = \Theta\left(\frac{(\log d)^2 \cdot S\|\mathbf{\Delta}_{t-1}\|^2}{\epsilon}\right), \tag{169}
$$

we have:
$$
\mathbb{E}[\|\nabla J_{\mathsf{H}}(\theta_t) - \mathbf{v}_t\|] \leq \epsilon, \qquad \mathbb{E}[\|\nabla^2 J_{\mathsf{H}}(\theta_t) - \mathbf{U}_t\|^2] \leq \epsilon. \tag{170}
$$
We use the same argument in the proof of Lemma 12 and we borrow the notations in Lemma 12. Let us define $\mathbf{u}_k$ and $\mathbf{V}_k$. For the gradient estimator $\mathbf{v}_t$, we have $\mathbf{v}_t - \nabla J_{\mathsf{H}}(\theta_t) = \sum_{k=\lfloor t/S \rfloor S}^{t} \mathbf{u}_k$ such that:

$$
\mathbf{u}_k = \begin{cases} \frac{1}{n_g^k} \sum_{j=1}^{n_g^k} \hat{\nabla} J(\theta_k; \tau_j^k) - \nabla J_{\mathsf{H}}(\theta_k), & k = \lfloor t/S \rfloor S, \\ \frac{1}{n_g^k} \sum_{j=1}^{n_g^k} [\hat{\nabla} J(\theta_k; \tau_j^k) - w(\tau_j^k|\theta_{k-1}, \theta_k) \hat{\nabla} J(\theta_{k-1}; \tau_j^k)] - \nabla J_{\mathsf{H}}(\theta_k) + \nabla J_{\mathsf{H}}(\theta_{k-1}), & k > \lfloor t/S \rfloor S, \end{cases}
$$

and for the Hessian estimator $\mathbf{U}_t$, we have $\mathbf{U}_t - \nabla^2 J_{\mathsf{H}}(\theta_t) = \sum_{k=\lfloor t/S \rfloor S}^{t} \mathbf{V}_k$ such that

$$
\mathbf{V}_k = \begin{cases} \frac{1}{n_H^k} \sum_{i=1}^{n_H^k} \hat{\nabla}^2 J(\theta_k; \tau_i^k) - \nabla^2 J_{\mathsf{H}}(\theta_k), & k = \lfloor t/S \rfloor S, \\ \frac{1}{n_H^k} \sum_{i=1}^{n_H^k} [\hat{\nabla}^2 J(\theta_k; \tau_i^k) - w(\tau_i^k|\theta_{k-1}, \theta_k) \hat{\nabla}^2 J(\theta_{k-1}; \tau_i^k)] - \nabla^2 J_{\mathsf{H}}(\theta_k) + \nabla^2 J_{\mathsf{H}}(\theta_{k-1}), & k > \lfloor t/S \rfloor S. \end{cases}
$$

To prove that (170), as the importance sampling terms appears in $\mathbf{u}_k$ and $\mathbf{V}_k$ for $k > \lfloor t/S \rfloor S$, we must only modify the proof of $\mathbb{E}[\|\mathbf{u}_k\|^2|\mathcal{G}_k] \leq \frac{C_1}{n_g^k} \|\theta_k - \theta_{k-1}\|^2$ and $\mathbb{E}[\|\mathbf{V}_k\|^2|\mathcal{H}_k] \leq \frac{C_2}{n_g^k} \|\theta_k - \theta_{k-1}\|^2$

41

for $k > \lfloor t/S \rfloor S$. We prove them as follows:

$$\mathbb{E}[\|\mathbf{u}_k\|^2|\mathcal{G}_k] = \mathbb{E}\left\|\frac{1}{n_g^k}\sum_{j=1}^{n_g^k}[\hat{\nabla}J(\theta_k;\tau_j^k) - w(\tau_j^k|\theta_{k-1},\theta_k)\hat{\nabla}J(\theta_{k-1};\tau_j^k)] - \nabla J_{\mathsf{H}}(\theta_k) + \nabla J_{\mathsf{H}}(\theta_{k-1})\right\|^2$$

$$\stackrel{(a)}{=} \frac{1}{n_g^k}\mathbb{E}\left\|\hat{\nabla}J(\theta_k;\tau_1^k) - w(\tau_1^k|\theta_{k-1},\theta_k)\hat{\nabla}J(\theta_{k-1};\tau_1^k) - \nabla J_{\mathsf{H}}(\theta_k) + \nabla J_{\mathsf{H}}(\theta_{k-1})\right\|^2$$

$$\stackrel{(b)}{\leq} \frac{2}{n_g^k}\mathbb{E}\|\hat{\nabla}J(\theta_k;\tau_1^k) - w(\tau_1^k|\theta_{k-1},\theta_k)\hat{\nabla}J(\theta_{k-1};\tau_1^k)\|^2 + \frac{2}{n_g^k}\mathbb{E}\|\nabla J_{\mathsf{H}}(\theta_k) - \nabla J_{\mathsf{H}}(\theta_{k-1})\|^2$$

$$\stackrel{(c)}{\leq} \frac{4}{n_g^k}\mathbb{E}\|\hat{\nabla}J(\theta_k;\tau_1^k) - \hat{\nabla}J(\theta_{k-1};\tau_1^k)\|^2 + \frac{4}{n_g^k}\mathbb{E}[(w(\tau_1^k|\theta_{k-1},\theta_k) - 1)^2\|\hat{\nabla}J(\theta_{k-1};\tau_1^k)\|^2]$$
$$+ \frac{2}{n_g^k}\mathbb{E}\|\nabla J_{\mathsf{H}}(\theta_k) - \nabla J_{\mathsf{H}}(\theta_{k-1})\|^2$$

$$\stackrel{(d)}{\leq} \frac{4}{n_g^k}\mathbb{E}\|\hat{\nabla}J(\theta_k;\tau_1^k) - \hat{\nabla}J(\theta_{k-1};\tau_1^k)\|^2 + \frac{4G_1^2}{n_g^k}\mathbb{E}[(w(\tau_1^k|\theta_{k-1},\theta_k) - 1)^2] + \frac{2}{n_g^k}\mathbb{E}\|\nabla J_{\mathsf{H}}(\theta_k) - \nabla J_{\mathsf{H}}(\theta_{k-1})\|^2$$

$$\stackrel{(e)}{\leq} \frac{6L_1'^2 + 4G_1^2 C_w}{n_g^k}\|\theta_k - \theta_{k-1}\|^2 \tag{171}$$

where (a) comes from $\hat{\nabla}J(\theta_k;\tau_j^k) - w(\tau_j^k|\theta_{k-1},\theta_k)\hat{\nabla}J(\theta_{k-1};\tau_j^k) - \nabla J_{\mathsf{H}}(\theta_k) + \nabla J_{\mathsf{H}}(\theta_{k-1})$'s are i.i.d conditioned on $\mathcal{G}_k$ for $1 \leq i \leq n_g^k$. Inequalities (b) and (c) are from $(a+b)^2 \leq 2a^2 + 2b^2$. (d) is due to Assumption 7. (e) is derived by Assumption 12 and Lemma 15.

Similarly, for Hessian estimator we have

$$\mathbb{E}[\|\mathbf{V}_k\|^2|\mathcal{H}_k] = \mathbb{E}\left\|\frac{1}{n_H^k}\sum_{i=1}^{n_H^k}[\hat{\nabla}^2J(\theta_k,\tau_i) - w(\tau_i^k|\theta_{k-1},\theta_k)\hat{\nabla}^2J(\theta_{k-1},\tau_i)] - \nabla^2 J_{\mathsf{H}}(\theta_k) + \nabla^2 J_{\mathsf{H}}(\theta_{k-1})\right\|^2$$

$$\stackrel{(a)}{\leq} 8e\log d \cdot \mathbb{E}\left\|\frac{1}{(n_H^k)^2}\sum_{i=1}^{n_H^k}\left[[\hat{\nabla}^2J(\theta_k,\tau_i) - w(\tau_i^k|\theta_{k-1},\theta_k)\hat{\nabla}^2J(\theta_{k-1},\tau_i)] - \nabla^2 J_{\mathsf{H}}(\theta_k) + \nabla^2 J_{\mathsf{H}}(\theta_{k-1})\right]^2\right\|$$

$$\stackrel{(b)}{\leq} \frac{8e\log d}{n_H^k} \cdot \frac{1}{n_H^k}\sum_{i=1}^{n_H^k}\mathbb{E}\left\|[\hat{\nabla}^2J(\theta_k,\tau_i) - w(\tau_i^k|\theta_{k-1},\theta_k)\hat{\nabla}^2J(\theta_{k-1},\tau_i) - \nabla^2 J_{\mathsf{H}}(\theta_k) + \nabla^2 J_{\mathsf{H}}(\theta_{k-1})]^2\right\|$$

$$\stackrel{(c)}{\leq} \frac{8e\log d}{n_H^k}\mathbb{E}\left\|\hat{\nabla}^2J(\theta_k,\tau_1) - w(\tau_1^k|\theta_{k-1},\theta_k)\hat{\nabla}^2J(\theta_{k-1},\tau_1) - \nabla^2 J_{\mathsf{H}}(\theta_k) + \nabla^2 J_{\mathsf{H}}(\theta_{k-1})\right\|^2$$

$$\stackrel{(d)}{\leq} \frac{16e\log d}{n_H^k}\mathbb{E}\|\hat{\nabla}^2J(\theta_k,\tau_1) - w(\tau_1^k|\theta_{k-1},\theta_k)\hat{\nabla}^2J(\theta_{k-1},\tau_1)\|^2 + \frac{16e\log d}{n_H^k}\|\nabla^2 J_{\mathsf{H}}(\theta_k) - \nabla^2 J_{\mathsf{H}}(\theta_{k-1})\|^2$$

$$\stackrel{(e)}{\leq} \frac{32e\log d}{n_H^k}\mathbb{E}\|\hat{\nabla}^2J(\theta_k,\tau_1) - \hat{\nabla}^2J(\theta_{k-1},\tau_1)\|^2 + \frac{32e\log d}{n_H^k}\mathbb{E}\left[(1 - w(\tau_1^k|\theta_{k-1},\theta_k))^2\|\hat{\nabla}^2J(\theta_{k-1},\tau_1)\|^2\right]$$
$$+ \frac{16e\log d}{n_H^k}\|\nabla^2 J_{\mathsf{H}}(\theta_k) - \nabla^2 J_{\mathsf{H}}(\theta_{k-1})\|^2$$

$$\stackrel{(f)}{\leq} \frac{32e\log d}{n_H^k}\mathbb{E}\|\hat{\nabla}^2J(\theta_k,\tau_1) - \hat{\nabla}^2J(\theta_{k-1},\tau_1)\|^2 + \frac{32e\cdot G_2^2\log d}{n_H^k}\mathbb{E}\left[(1 - w(\tau_1^k|\theta_{k-1},\theta_k))^2\right]$$
$$+ \frac{16e\log d}{n_H^k}\|\nabla^2 J_{\mathsf{H}}(\theta_k) - \nabla^2 J_{\mathsf{H}}(\theta_{k-1})\|^2$$

$$\stackrel{(g)}{\leq} \frac{32e\log d(2L_2'^2 + C_w G_2^2)}{n_H^k}\|\theta_k - \theta_{k-1}\|^2, \tag{172}$$

where (a) comes from Lemma 4 and the fact that $[\hat{\nabla}^2J(\theta_k,\tau_i) - w(\tau_i^k|\theta_{k-1},\theta_k)\hat{\nabla}^2J(\theta_{k-1},\tau_i) - \nabla^2 J_{\mathsf{H}}(\theta_k) + \nabla^2 J_{\mathsf{H}}(\theta_{k-1})]$'s are i.i.d conditioned on $\mathcal{H}_k$ for $1 \leq i \leq n_H^k$. (b) comes from Jensen's

42

(a) Avg. path length (random shape maze)

(b) Avg. episode return (random shape maze)

(c) Avg. path length (random maze)

(d) Avg. episode return (random maze)
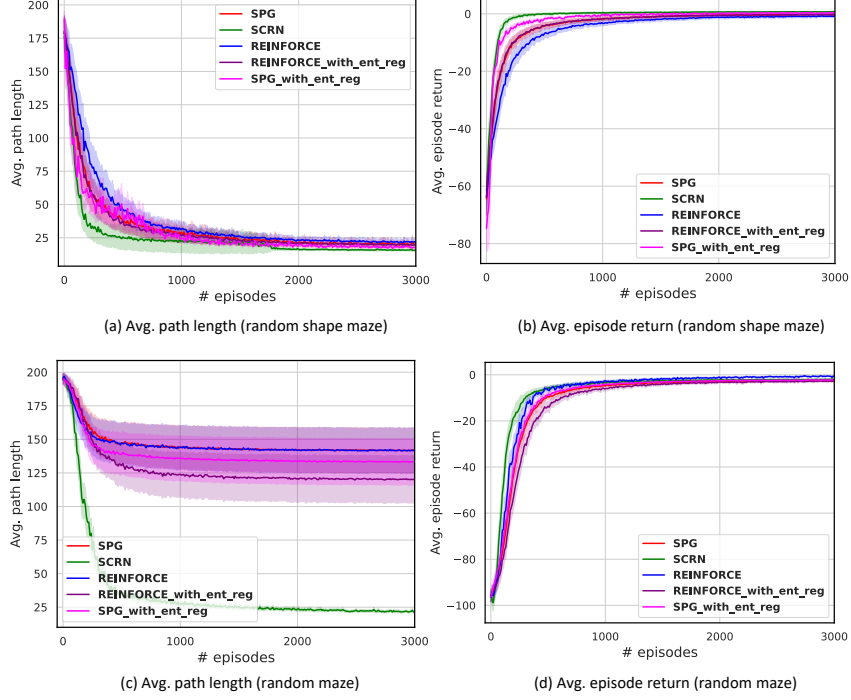
Figure 6: Comparison of SCRN with first-order methods in maze environments. In random shape maze, the percentages of successful instances for SPG, SCRN, REINFORCE, SPG with entropy regularization, and REINFORCE with entropy regularization are 86%, 100%, 95.3%, 92.2%, and 95.3%, respectively. In random maze, the percentages of successful instances for SPG, SCRN, REINFORCE, SPG with entropy regularization, and REINFORCE with entropy regularization are 45.3%, 97%, 40.6%, 54.7%, and 54.7%, respectively.

inequality for operator norm $\|\cdot\|$. (c) is derived by $\|AB\| \leq \|A\|\|B\|$ and the fact that $[\hat{\nabla}^2 J(\theta_k, \tau_i) - w(\tau_i^k|\theta_{k-1}, \theta_k)\hat{\nabla}^2 J(\theta_{k-1}, \tau_i) - \nabla^2 J_{\mathsf{H}}(\theta_k) + \nabla^2 J_{\mathsf{H}}(\theta_{k-1})]$'s are i.i.d conditioned on $\mathcal{H}_k$ for $1 \leq i \leq n_H^k$. Inequalities (d) and (e) are from inequality $(a+b)^2 \leq 2a^2 + 2b^2$ and (f) is derived by Assumption 7. (g) comes from Assumption 12 and Lemma 15. Similar to the proof of Lemma 12, we can argue that for Algorithm 3, (170) holds.

Now, based on what we proved above, with the same arguments in Theorem 3, Algorithm 3 achieves ϵ-global stationary point by querying $\mathcal{O}(\epsilon^{-2})$ stochastic gradients and $\mathcal{O}(\epsilon^{-1})$ stochastic Hessian on average. More precisely, the following average number of queries of stochastic gradient and Hessian are required in order to achieve ϵ-global stationary point:

$$2\mathsf{H} \times \sum_{i=1}^{T} n_g^t = \mathcal{O}\left(\frac{\sigma_1^2 \mathsf{H} \left[4^2(C_g + C_H)C^3 + C_g + C_H \right]^2}{\epsilon^2} \right) = \mathcal{O}\left(\frac{(G_1^{2+\frac{8}{3}} G_2^{\frac{8}{3}} R_{\max}^{2+\frac{8}{3}}}{(1-\gamma)^{4+\frac{16}{3}} \epsilon^2} \right).$$

where (a) comes from $\sigma_1^2 = \frac{\mathsf{H}G_1^2 R_{\max}^2}{(1-\gamma)^2}$ Lemma 14 and $C = \mathcal{O}(M^{1/3})$, $C_g = \mathcal{O}(M^{1/3})$, and $C_H = \mathcal{O}(M^{1/3})$ and the fact that $M = \mathcal{O}(\tilde{L})$ and from Lemma 13, $\tilde{L} = \mathcal{O}\left(\frac{R_{\max}G_1 G_2}{(1-\gamma)^2} \right)$.

A.5 Further experimental results

Discussion on results: In Fig. 6 (a), in the random shape maze environment, SCRN finds paths with average length of 25 after about 600 episodes while SPG and REINFORCE can only achieve the average path lengths of greater than 25 after 3000 episodes. In the random maze environment, SCRN again outperforms the other two algorithms and achieves the average path length of 25 after about 2000 episodes. However, for SPG and REINFROCE, the average path lengths are more than 100 which also indicates the low percentages of successful instances for these two algorithms in this

43

a) HalfCheetah

b) Humanoid

c) Walker

d) Reacher
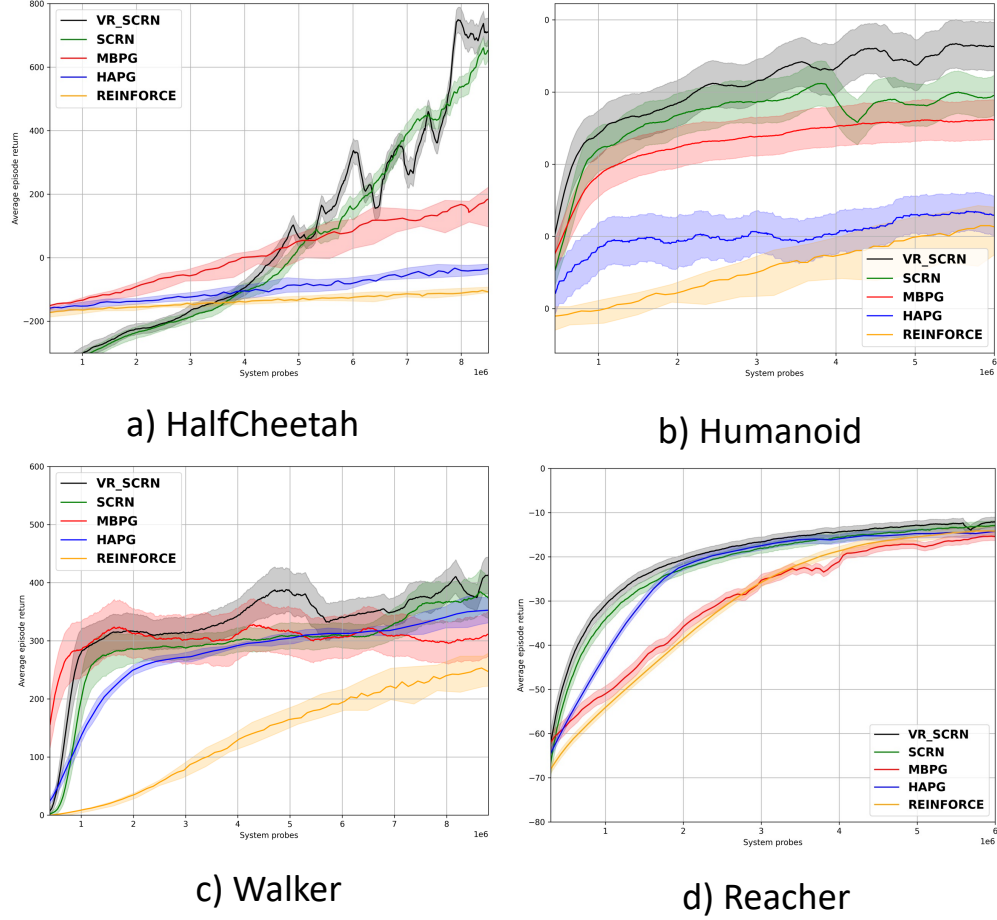
Figure 7: Comparison of SCRN and its variance-reduced version with REINFORCE and variance-reduced SPG methods in MuJoCo environments.

environment. Furthermore, as shown in Fig. 6 (c,d), in both random shape maze and random maze environments, SCRN has better performance than SPG and REINFORCE in terms of average episode return. We also provide the results for the random shape maze and random maze environments where we add entropy regularization term to the objective functions of SPG and REINFORCE. As can be seen, the results with entropy regularization term are slightly better both in the average path length and average episode return.

In Fig. 7, for control tasks in MuJoCo simulator, we compared SCRN and its variance-reduced version with first-order methods such as REINFORCE, and two state-of-the-art representatives of variance-reduced PG methods, HAPG [29] and MBPG [13], both with guaranteed convergence to ϵ-FOSP in general non-convex settings. In HalfCheetah environment, SCRN and variance-reduced SCRN outperform the other methods by achieving a score of around 600. In the Humanoid environment, variance-reduced SCRN achieves the best performance among other algorithms. In Walker environment, the performances of SCRN, its variance-reduced version, and HAPG perform better than MBPG and REINFORCE. In Reacher environment, the variance-reduced SCRN has the best performance among the considered algorithms.

A.6 Implementation details

Environments:

- Environments with finite state and action spaces: We considered two grid world environments in our experiments: [31, Example 6.6], and random mazes [51]. In cliff walking,

44
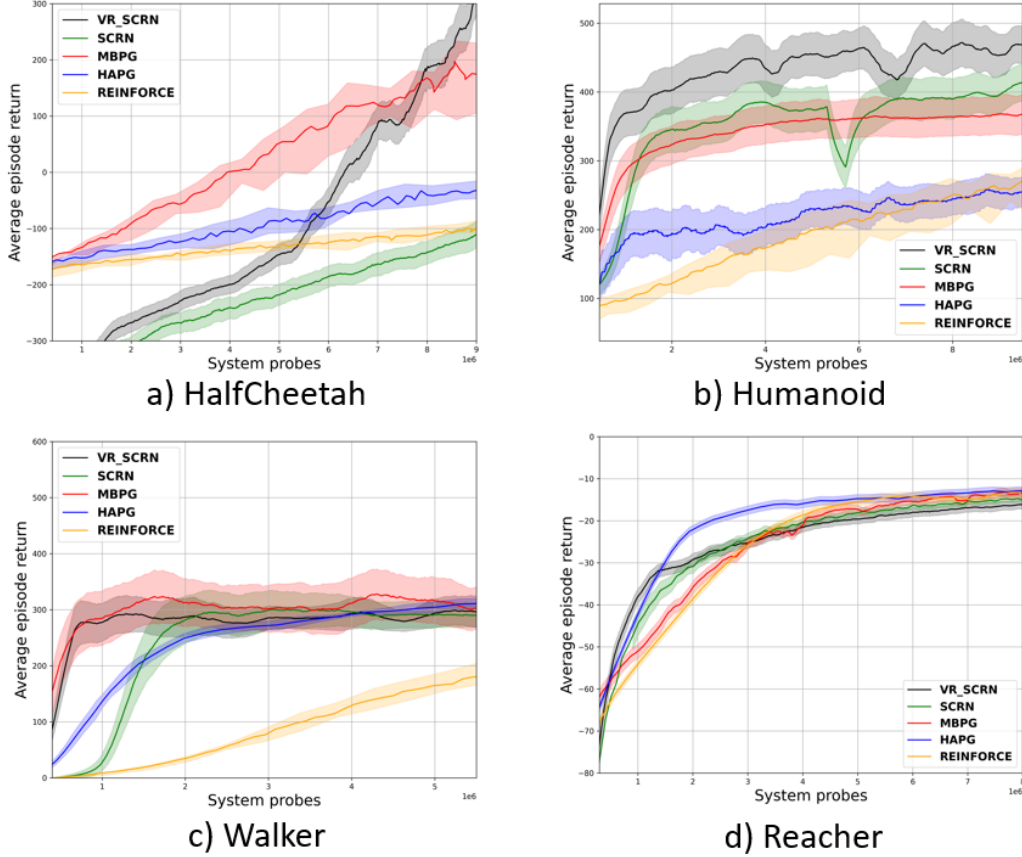
Figure 8: Comparison of SCRN and its variance-reduced version (without using the trick in updates) with REINFORCE and variance-reduced SPG methods in MuJoCo environments.

the agent's aim is to reach a goal state from a start state, avoiding a region of cells called "cliff". The episode is terminated if the agent enters the cliff region or the number of steps exceeds 100 without reaching the goal. The reward is -0.1 in all transitions except those into the cliff where the reward is -100. The reward of reaching the goal is 100. Moreover, we considered a softmax tabular policy for all the experiments of this part.

In random mazes, the size of each maze is 10×10. In the random shape maze, random shape blocks are placed on a grid and the agent tries to reach the goal state finding the shortest path, avoiding blocks. The reward is -0.1 in all transitions except if the agent tries to go to a cell which belongs to a block where the reward is -1. Moreover, the reward of reaching the goal is 1. An episode is terminated if the agent could not reach the goal after 200 steps.

- Environments with continuous state and action spaces: We considered the following control tasks in MuJoCo simulator [32]: Walker, Humanoid, Reacher, and HalfCheeta. We compared SCRN with first-order methods such as REINFORCE, and two state-of-the-art representatives of variance-reduced PG methods, HAPG [29] and MBPG [13]. For each task, we utilized a Gaussian multi-layer perceptron (MLP) policy whose mean and variance are parameterized by an MLP with two hidden layers of 64 neurons. For a fair comparison, we considered the same network architecture for all the methods.

Algorithms:

For the environments with finite state and action spaces, we provided an implementation of SCRN, SPG, and REINFORCE with NumPy where the gradient and Hessian (just for SCRN) of a given trajectory are computed based on their closed forms. We also implemented SCRN with PyTorch in Garage library in order to execute it on environments with continuous state and action spaces. In

45

our experiments, we used a Linux server with Intel Xeon Gold 6240 CPU (36 cores) operating at 2.60GHz with 377 GB DDR4 of memory, and Nvidia Titan X GPU.

Regarding SPG and REINFORCE in grid world environments, we adapted a time-varying learning rate by checking various forms and the form of $a/(\lfloor t/P \rfloor + b)$ provided the best performance for the first-order methods where a, b, and P are some constants that are needed to be tuned. The parameter P shows the number of episodes that the learning rate remains unchanged.

A trick in updates of SCRN and its variance-reduced version: Experiments in Figures 3 and 7 are conducted using the following trick to prevent updating based on a large $\|\mathbf{\Delta}_t\|$. If $\|\mathbf{\Delta}_t\|$ precedes a threshold, we neglect that whole iteration and do not update $\mathbf{x}_t$. Please note that we take the number of observed state-actions of that neglected iteration into account in computing system probes.

For sake of completeness, in Figure 8, we also provided the performances of SCRN and its variance-reduced version without using the trick described above, and compared with other variance-reduced methods.

Demonstrations:

We studied how the parameters of the softmax tabular policy are evolving over time in cliff walking environment. For each cell, we considered an arrow with four directions: up, down, left, and right. At any time, the direction of each arrow shows which one of the four directions has the highest probability of being taken by the agent in the corresponding cell. The color of the arrow becomes darker as this probability increases. For each algorithm, we run 100 episodes and each episode, we demonstrated how the parameters are being updated. For SPG and REINFORCE, for a long period of time, at the start state, the agent takes the "up" action in order to avoid falling off the cliff. In fact, it takes many episodes until the parameters of the start state are updated such that the agent tries to find a path to the goal. In contrast, SCRN finds a path to the goal after few episodes and then tries to improve the path length.

46