Faster stochastic cubic regularized Newton methods with momentum

Yiming Yang [*]   Chuan He [†]   Xiao Wang [‡]   Zheng Peng [*]

July 17, 2025

Abstract

Cubic regularized Newton (CRN) methods have attracted significant research interest because they offer stronger solution guarantees and lower iteration complexity. With the rise of the big-data era, there is growing interest in developing stochastic cubic regularized Newton (SCRN) methods that do not require exact gradient and Hessian evaluations. In this paper, we propose faster SCRN methods that incorporate gradient estimation with small, controlled errors and Hessian estimation with momentum-based variance reduction. These methods are particularly effective for problems where the gradient can be estimated accurately and at low cost, whereas accurate estimation of the Hessian is expensive. Under mild assumptions, we establish the iteration complexity of our SCRN methods by analyzing the descent of a novel potential sequence. Finally, numerical experiments show that our SCRN methods can achieve comparable performance to deterministic CRN methods and vastly outperform first-order methods in terms of both iteration counts and solution quality.

Keywords: Stochastic cubic regularized Newton methods, variance reduction, momentum, iteration complexity

Mathematics Subject Classification: 49M15, 90C25, 90C30

1 Introduction

In this paper, we consider the smooth unconstrained optimization problem

$$\min_{x \in \mathbb{R}^n} f(x), \tag{1}$$

where $f : \mathbb{R}^n \to \mathbb{R}$ is twice continuously differentiable. We assume that problem (1) has at least one optimal solution. Over the past few years, second-order methods have gained popularity for handling problem (1) due to their ability to converge in fewer iterations than first-order methods and to deliver higher solution quality. However, the computational overhead incurred per evaluation of the Hessian matrix hinders the scalability of second-order methods in modern large-scale settings. To better leverage second-order information in these settings, this paper aims to propose practical second-order methods—particularly variants of the cubic regularized Newton (CRN) method—to solve problem (1) and analyze their iteration complexity for finding an approximate second-order stationary point (SOSP) of (1).

Second-order methods have recently received considerable attention for their strong solution guarantees and rapid convergence, with substantial progress made in designing new second-order methods with complexity guarantees for solving problem (1). In particular, CRN methods [1, 6, 8, 28], trust-region methods [11, 12, 27], second-order line-search method [31], inexact regularized Newton method [13],

---

[*]Department of Mathematics, Xiangtan University, China (email: `yiming9780@outlook.com,pzheng@xtu.edu.cn`).

[†]Department of Mathematics, Linköping University, Sweden (email: `chuan.he@liu.se`).

[‡]School of Computer Science and Engineering, Sun Yat-sen University, China (email: `wangx936@mail.sysu.edu.cn`).

quadratic regularization method [5], and Newton-CG methods [20, 21, 30] were developed for finding an $(\epsilon, \sqrt{\epsilon})$-SOSP x of problem (1) satisfying

$$\|\nabla f(x)\| \leq \epsilon, \qquad \lambda_{\min}(\nabla^2 f(x)) \geq -\sqrt{\epsilon},$$

where $\epsilon \in (0,1)$ is a tolerance parameter and $\lambda_{\min}(\cdot)$ denotes the minimum eigenvalue of the associated matrix. Under suitable assumptions, it was shown that these methods achieve an iteration complexity of $\mathcal{O}(\epsilon^{-3/2})$ for finding an $(\epsilon, \sqrt{\epsilon})$-SOSP, which has been proven to be optimal in [7, 9]. Besides, several gradient-based methods with random perturbations (e.g., [2, 24, 37]) have also been developed to find an $(\epsilon, \sqrt{\epsilon})$-SOSP with high probability.

Despite the significant advantages of second-order methods, their high per-iteration cost limits their use in large-scale problems. To address this limitation, many research works have focused on developing inexact and stochastic variants of second-order methods. In particular, inexact and stochastic versions of CRN methods [3, 8, 16, 19, 25, 32, 33, 36], trust-region method [36], Newton-CG method [38], and subsampling line-search method [4], have been developed to seek an approximate SOSP of problem (1). These methods achieve a similar order of complexity bounds as their exact variants. Yet, in each iteration, these methods require a fairly accurate approximation of the gradient and Hessian with small errors depending on the desired tolerance ϵ (potentially after a certain number of iterations), which remains quite restrictive in practice.

Furthermore, several recent works [10, 23, 34, 39, 40] have developed stochastic second-order methods leveraging variance reduction techniques. These methods apply variance reduction to iteratively correct estimation errors using stochastic derivative information from previous iterations, thereby avoiding the need to construct accurate derivative estimates at each step. Such desirable features enable these methods to maintain low per-iteration costs, making them more practical for large-scale problems. Among these works, [10] is the only one that do not assuming a finite-sum structure for the objective function. This work proposed two methods that achieve iteration complexity bounds of $\mathcal{O}(\epsilon^{-7/2})$ and $\mathcal{O}(\epsilon^{-10/3})$, respectively, for finding an approximate stochastic stationary point x satisfying $\mathbb{E}[\|\nabla f(x)\|] \leq \epsilon$. However, these complexity bounds leave a significant gap compared to the bound of $\mathcal{O}(\epsilon^{-3/2})$ achieved by the deterministic CRN method; moreover, they are worse than the iteration complexity of $\mathcal{O}(\epsilon^{-3})$ achieved by stochastic first-order methods (e.g., [15, 17, 26]).

Motivated by the aforementioned discussions, we aim to rethink the design of stochastic second-order methods and identify the types of stochastic optimization problems for which they are preferable. In this paper, we focus on a class of problems where the gradient can be estimated relatively easily with small errors (see Assumption 1(c)), but estimating the Hessian is more expensive; therefore, only unbiased stochastic Hessian estimators with bounded moments (see Assumption 1(d)) are available at each step. Specifically, we propose two new variants of SCRN methods for solving such problems. Under mild assumptions, we establish their iteration complexity for finding an (ϵ_g, ϵ_H)-stochastic second-order stationary point (SSOSP) of problem (1) (see Definition 1), based on an analysis of the descent of a novel potential sequence (see (17)). For ease of comparison, we summarize the iteration complexity, the number of samples per iteration, smoothness conditions, and stationary measures for vanilla gradient descent, stochastic first-order methods, CRN, variants of SCRN, and our methods for nonconvex optimization in Table 1.

The main contributions of this paper are highlighted below.

- We propose two new SCRN methods (Algorithms 1 and 2), which adopt stochastic gradients with small errors and incorporate momentum-based variance reduction for estimating Hessian. Under mild assumptions, we establish their iteration complexity based on an analysis of the descent of a

2

Table 1: Comparison of vanilla gradient descent (GD), stochastic gradient methods with momentum (SGD-M), CRN, SCRN, and SCRN with momentm (SCRN-M) in terms of iteration complexity, the number of samples per iteration, smoothness conditions, and stationary measures.

	First-order methods			
	iteration complexity	gradient samples per iteration	smoothness condition	stationary measure
GD	$\mathcal{O}(\epsilon^{-2})$	—	∇f Lipschitz	$\|\nabla f(x)\| \leq \epsilon$
SGD [18]	$\mathcal{O}(\epsilon^{-4})$	1	∇f Lipschitz	$\mathbb{E}[\|\nabla f(x)\|^2] \leq \epsilon^2$
SGD-M [14, 22]	$\widetilde{\mathcal{O}}(\epsilon^{-(3p+1)/p})$	1	$\mathcal{D}^p f$ Lipschitz[1]	$\mathbb{E}[\|\nabla f(x)\|] \leq \epsilon$
SGD-M [15]	$\mathcal{O}(\epsilon^{-3})$	1	G average Lipschitz	$\mathbb{E}[\|\nabla f(x)\|^2] \leq \epsilon^2$
	Second-order methods			
	iteration complexity	Hessian samples per iteration	smoothness condition	stationary measure
CRN [28]	$\mathcal{O}(\max\{\epsilon_g^{-3/2}, \epsilon_H^{-3}\})$	—	$\nabla^2 f$ Lipschitz	$\|\nabla f(x)\| \leq \epsilon_g, \lambda_{\min}(\nabla^2 f(x)) \geq -\epsilon_H$
SCRN [32]	$\mathcal{O}(\epsilon^{-3/2})$	$\widetilde{\mathcal{O}}(\epsilon^{-1})$	$\nabla f, \nabla^2 f$ Lipschitz	$\|\nabla f(x)\| \leq \epsilon, \lambda_{\min}(\nabla^2 f(x)) \geq -\sqrt{\epsilon}$ w.h.p.
SCRN-M [10]	$\mathcal{O}(\epsilon^{-7/2})$	1	$\nabla^2 f$ Lipschitz	$\mathbb{E}[\|\nabla f(x)\|^{3/2}] \leq \epsilon^{3/2}$
Algorithm 1 (ours)	$\mathcal{O}(\max\{\epsilon_g^{-7/4}, \epsilon_H^{-7}\})$	1	$\nabla^2 f$ Lipschitz	$\mathbb{E}[\|\nabla f(x)\|^{3/2}] \leq \epsilon_g^{3/2}, \mathbb{E}[\lambda_{\min}(\nabla^2 f(x))^3] \geq -\epsilon_H^3$
Algorithm 2 (ours)	$\mathcal{O}(\max\{\epsilon_g^{-5/3}, \epsilon_H^{-5}\})$	1	H average Lipschitz	$\mathbb{E}[\|\nabla f(x)\|^{3/2}] \leq \epsilon_g^{3/2}, \mathbb{E}[\lambda_{\min}(\nabla^2 f(x))^3] \geq -\epsilon_H^3$

  novel potential sequence. To the best of our knowledge, the obtained complexity bounds are new to the literature.

- We conduct numerical experiments (Section 5) to compare our SCRN methods with deterministic CRN, other SCRN variants, and first-order methods. The numerical results show that our methods achieve performance comparable to deterministic CRN methods and significantly outperform other SCRN variants and first-order methods.

The remainder of this paper is organized as follows. In Section 2, we introduce the notation and assumptions used throughout the paper. Sections 3 and 4 present two new variants of SCRN and establish their iteration complexity. Section 5 reports preliminary numerical results. Finally, Section 6 provides the proofs of the main results.

2   Notation and assumptions

Throughout this paper, we use $\mathbb{R}^n$ to denote the n-dimensional Euclidean space endowed with the standard inner product $\langle \cdot, \cdot \rangle$. We let $\| \cdot \|$ denote the Euclidean norm for vectors and the spectral norm for matrices, and use $\| \cdot \|_F$ to denote the Frobenius norm for matrices. For a given matrix $H \in \mathbb{R}^{n \times n}$, we use $\lambda_{\min}(H)$ to denote its minimum eigenvalue, and use $\text{Tr}(H)$ to denote the trace of H. We let I be the $n \times n$ identity matrix. In addition, we use $\widetilde{\mathcal{O}}(\cdot)$ to represent $\mathcal{O}(\cdot)$ with polylogarithmic terms omitted.

  We now make the following assumptions throughout this paper.

Assumption 1. (a) *There exists a finite f_{low} such that $f(x) \geq f_{\text{low}}$ for all $x \in \mathbb{R}^n$.*

(b) *There exist $L > 0$ and $L_F > 0$ such that $\|\nabla^2 f(y) - \nabla^2 f(x)\| \leq L\|y - x\|$ and $\|\nabla^2 f(y) - \nabla^2 f(x)\|_F \leq L_F\|y - x\|$ hold for all $x, y \in \mathbb{R}^n$.*

---

[1]G and $\mathcal{D}^p f$ represent the stochastic gradient and the pth-order derivative of f, respectively.

(c) *For any $\delta \in (0, 1)$, we have access to a stochastic gradient estimator $G_\delta : \mathbb{R}^n \times \mathcal{Z} \to \mathbb{R}^n$ satisfying*

$$\mathbb{E}_\zeta[\|G_\delta(x;\zeta) - \nabla f(x)\|^{3/2}] \leq \delta^{3/2} \qquad \forall x \in \mathbb{R}^n. \tag{2}$$

(d) *We have access to a stochastic Hessian estimator $H : \mathbb{R}^n \times \Xi \to \mathbb{R}^{n \times n}$ satisfying*

$$\mathbb{E}_\xi[H(x;\xi)] = \nabla^2 f(x), \quad \mathbb{E}_\xi[\|H(x;\xi) - \nabla^2 f(x)\|_F^3] \leq \sigma^3 \qquad \forall x \in \mathbb{R}^n \tag{3}$$

for some $\sigma > 0$.

Remark 1. *(i) Assumptions 1(a) and 1(b) are common in the literature on CRN methods (e.g., see [28, 39]). It follows from Assumption 1(b) that*

$$\|\nabla f(y) - \nabla f(x) - \nabla^2 f(x)(y-x)\| \leq \frac{L}{2}\|y-x\|^2 \quad \forall x, y \in \mathbb{R}^n, \tag{4}$$

$$f(y) \leq f(x) + \nabla f(x)^T(y-x) + \frac{1}{2}(y-x)^T \nabla^2 f(x)(y-x) + \frac{L}{6}\|y-x\|^3 \quad \forall x, y \in \mathbb{R}^n. \tag{5}$$

In addition, Assumption 1(c) states that $G_\delta(\cdot;\xi)$ approximates the true gradient $\nabla f(\cdot)$ to any desired accuracy in expectation, while Assumption 1(d) states that the stochastic Hessian $H(\cdot;\xi)$ is an unbiased estimator of $\nabla^2 f(\cdot)$ and has a bounded third-order central moment.

(ii) We made two Lipschitz continuity assumptions on $\nabla^2 f$ in Assumption 1(b). In particaulr, the Lipschitz continuity with respect to the spectral norm is used to estimate the reduction of function values at each iteration of our SCRN methods (see Lemma 3 below and the classical analysis in [28]). In comparison, the Lipschitz continuity with respect to the Frobenius norm is used to derive a recursive relation for the decreasing estimation error given by the momentum update (see Lemmas 4 and 6). In our complexity analysis, we found that within the Schatten family of matrix norms, only the Frobenius norm seems effective for analyzing Hessian estimation error with momentum updates, while other norms, such as the spectral and nuclear norms, do not appear to be useful. Our explanation is that for any Schatten-p norm $\|\cdot\|_{S_p}$, our analysis requires the norm $\|\cdot\|_{S_p}$ to be continuously differentiable. However, this condition is satisfied only when $p = 2$, which corresponds to the Frobenius norm.

We next introduce the definition of an approximate SSOSP, which our methods aim to achieve.

Definition 1. For any $\epsilon_g, \epsilon_H \in (0, 1)$, we say that $x \in \mathbb{R}^n$ is an (ϵ_g, ϵ_H)-stochastic second-order stationary point (SSOSP) of problem (1) if it satisfies $\mathbb{E}[\|\nabla f(x)\|^{3/2}] \leq \epsilon_g^{3/2}$ and $\mathbb{E}[\lambda_{\min}(\nabla^2 f(x))^3] \geq -\epsilon_H^3$.

3 SCRN with Polyak momentum

In this section, we propose an SCRN method with Polyak momentum, and then establish its iteration complexity under Assumption 1.

Specifically, our SCRN method with Polyak momentum generate three sequences, $\{g^k\}$, $\{M_k\}$, and $\{x^k\}$. At the kth iteration, this method first computes g^k as a stochastic gradient of f at x^k with error δ_k, and then computes M_k as a weighted average of the stochastic Hessians evaluated at $x^0, \ldots, x^k$. The next iterate x^{k+1} is obtained by solving a cubic regularized Newton subproblem. Details of this method are described in Algorithm 1, with a specific choice of input parameters given in Theorem 1.

The following theorem establishes the iteration complexity of Algorithm 1 for computing an (ϵ_g, ϵ_H)-SSOSP of problem (1). Its proof is provided in Section 6.2.

Algorithm 1 SCRN with Polyak momentum

Input: starting point $x^0 \in \mathbb{R}^n$, regularization parameters $\{\eta_k\} \subset (0, \infty)$, error parameters $\{\delta_k\} \subset (0, 1)$, momentum parameters $\{\theta_k\} \subset (0, 1)$.
Initialize: $M_{-1} = 0$ and $\theta_{-1} = 1$.
for $k = 0, 1, 2, \ldots$ **do**
    Construct the gradient and Hessian estimators:

$$g^k = G_{\delta_k}(x^k; \zeta^k), \quad M_k = (1 - \theta_{k-1})M_{k-1} + \theta_{k-1}H(x^k; \xi^k). \tag{6}$$

    Update the next iterate:

$$x^{k+1} \in \underset{x \in \mathbb{R}^n}{\text{Arg min}}\Big\{(g^k)^T(x - x^k) + \frac{1}{2}(x - x^k)^T M_k(x - x^k) + \frac{1}{6\eta_k}\|x - x^k\|^3\Big\}.$$

end for

Theorem 1. *Suppose that Assumption 1 holds. Define*

$$M_{\text{pm}} := 54\Big(f(x^0) - f_{\text{low}} + \sigma^3 L_F^{-2} + L_F^{3/2}\sigma^3 + 1\Big), \tag{7}$$

where f_{low}, L_F, and σ be given in Assumption 1. Let $\{x^k\}$ be all iterates generated by Algorithm 1 with input parameters $\{(\eta_k, \theta_k, \delta_k)\}$ given by

$$\eta_k = \frac{1}{9K^{2/7}}, \quad \theta_k = \frac{7L_F}{3K^{2/7}}, \quad \delta_k = \frac{1}{9K^{4/7}}, \quad \forall k \geq 0. \tag{8}$$

Then, for any $\epsilon_g, \epsilon_H \in (0, 1)$, x^{ι_K} is an (ϵ_g, ϵ_H)-SSOSP of problem (1) for all K satisfying

$$K \geq \max\left\{\Big(\frac{(3M_{\text{pm}})^{2/3}}{\epsilon_g}\Big)^{7/4}, \Big(\frac{(108M_{\text{pm}})^{1/3}}{\epsilon_H}\Big)^7, \Big(\frac{2L}{9}\Big)^{7/2}, \Big(\frac{7L_F}{3}\Big)^{7/2}, 1\right\}, \tag{9}$$

where ι_K is uniformly drawn from $\{1, \ldots, K\}$.

Remark 2. *From Theorem 1, we see that Algorithm 1 with input parameters given by (8) achieves an iteration complexity of $\mathcal{O}(\max\{\epsilon_g^{-7/4}, \epsilon_H^{-7}\})$ for finding an (ϵ_g, ϵ_H)-SSOSP of problem (1).*

4   SCRN with recursive momentum

In this section, we propose an SCRN method with recursive momentum, and then establish its iteration complexity.

Specifically, our SCRN method with recursive momentum generate three sequences, $\{g^k\}$, $\{M_k\}$, and $\{x^k\}$. At the kth iteration, this method first compute g^k as a stochastic gradient of f at x^k with error δ_k, and compute M_k as a weighted average of the stochastic Hessian evaluated at $x^0, \ldots, x^k$ using the recursive momentum scheme proposed in [15]. The next iterate x^{k+1} is obtained by solving a cubic regularized Newton subproblem. Details of this method are provided in Algorithm 2, with a specific choice of input parameters given in Theorem 2.

Before analyzing Algorithm 2, we make the following assumption regarding the *mean-cubed smoothness* of the stochastic Hessian estimator $H(\cdot; \xi)$.

Assumption 2. *There exists $L_H > 0$ such that $\mathbb{E}_\xi[\|H(y; \xi) - H(x; \xi)\|_F^3] \leq L_H^3\|y - x\|_F^3$ holds for all $x, y \in \mathbb{R}^n$.*

Algorithm 2 SCRN with recursive momentum

Input: starting point $x^0 \in \mathbb{R}^n$, regularization parameters $\{\eta_k\} \subset (0, \infty)$, error control parameters $\{\delta_k\} \subset (0, 1)$, momentum parameters $\{\theta_k\} \subset (0, 1)$.

Initialize: $M_{-1} = 0$ and $\theta_{-1} = 1$.

for $k = 0, 1, 2, \ldots$ **do**

    Construct the gradient and Hessian estimators:

$$g^k = G_{\delta_k}(x^k; \zeta^k), \quad M_k = (1 - \theta_{k-1})M_{k-1} + H(x^k; \xi^k) - (1 - \theta_{k-1})H(x^{k-1}; \xi^k). \tag{10}$$

    Update the next iterate:

$$x^{k+1} \in \underset{x \in \mathbb{R}^n}{\text{Arg min}}\left\{(g^k)^T(x - x^k) + \frac{1}{2}(x - x^k)^T M_k (x - x^k) + \frac{1}{6\eta_k}\|x - x^k\|^3\right\}.$$

end for

The following theorem establishes an iteration complexity bound of Algorithm 2 for computing an (ϵ_g, ϵ_H)-SSOSP of problem (1). Its proof is relegated to Section 6.3.

Theorem 2. *Suppose that Assumptions 1 and 2 hold. Define*

$$M_{\text{rm}} := 75(f(x^0) - f_{\text{low}} + \sigma^3(L_F^3 + L_H^3)^{-2/3} + (L_F^3 + L_H^3)\sigma^3 + 1), \tag{11}$$

where f_{low}, L_F, and σ are given in Assumption 1, and L_H is given in Assumption 2. Let $\{x^k\}$ be all iterates generated by Algorithm 2 with input parameters $\{(\eta_k, \theta_k, \delta_k)\}$ given by

$$\eta_k = \frac{1}{17K^{1/5}}, \quad \theta_k = \frac{625(L_F^3 + L_H^3)^{2/3}}{289K^{2/5}}, \quad \delta_k = \frac{1}{17K^{3/5}} \quad \forall k \geq 0. \tag{12}$$

Then, for any $\epsilon_g, \epsilon_H \in (0, 1)$, x^{ι_K} is an (ϵ_g, ϵ_H)-SSOSP of problem (1) for all K satisfying

$$K \geq \max\left\{\left(\frac{(3M_{\text{rm}})^{2/3}}{\epsilon_g}\right)^{5/3}, \left(\frac{(281M_{\text{rm}})^{1/3}}{\epsilon_H}\right)^5, \left(\frac{2L}{17}\right)^5, 7(L_F^3 + L_H^3)^{5/3}, 1\right\}, \tag{13}$$

where ι_K is uniformly drawn from $\{1, \ldots, K\}$.

Remark 3. *From Theorem 2, we observe that Algorithm 2 with input parameters given by (12) achieves an iteration complexity of $\mathcal{O}(\max\{\epsilon_g^{-5/3}, \epsilon_H^{-5}\})$ for finding an (ϵ_g, ϵ_H)-SSOSP of problem (1). This bound improves upon the one for Algorithm 1 established in Theorem 1.*

5   Numerical experiments

In this section, we conduct numerical experiments to evaluate the performance of Algorithms 1 and 2, abbreviated as SCRN-PM and SCRN-RM, respectively. We compare these methods with the adaptive cubic regularized Newton method [8] (A-CRN), stochastic cubic regularized Newton method with momentum [10] (SCRN-M), and SpaRSA [35]. The experiments are conducted on three nonconvex statistical learning problems using datasets from LIBSVM[2]. All the algorithms are coded in Python, and all computations are performed on a laptop with an Intel Core i7 processor and 10 GB of RAM.

---

[2]https://www.csie.ntu.edu.tw/~cjlin/libsvmtools/datasets/

5.1 Regularized logistic regression problems

In this subsection, we consider the regularized logistic regression problem:

$$\min_{x\in\mathbb{R}^n} \ \sum_{i=1}^m \left(b_i \ln(\phi(x^T a_i)) + (1-b_i)\ln(1-\phi(x^T a_i)) + \lambda\sum_{j=1}^n \frac{(\gamma x_j)^2}{1+(\gamma x_j)^2}\right), \tag{14}$$

where $\phi(t) = e^t/(1+e^t)$ denotes the sigmoid function, $\{(a_i,b_i)\}_{1\le i\le m} \subset \mathbb{R}^n \times \mathbb{R}$ is the given data, and $(\lambda,\gamma) = (0.001, 10)$. We consider three datasets 'a9a', 'phishing', and 'w8a' from LIBSVM.

We apply SCRN-PM, SCRN-RM, A-CRN, SCRN-M, and SpaRSA to solve problem (14). All methods are initialized at $[0.5,\dots,0.5]^T$. For SCRN-M, we choose 50% of the elements from the gradient and Hessian, respectively, to construct unbiased estimators of ∇f and $\nabla^2 f$. For SCRN-PM and SCRN-RM, we choose set g^k as full gradients $\nabla f(x^k)$ for all $k \ge 0$, and choose 50% of the elements from the Hessian to construct unbiased Hessian estimators. For CRN and all SCRN methods, we adopt the Lanczos method used in [8] to solve the cubic regularized subproblems. We compare these methods in terms of the function value gap defined by $f(x^k) - f^*$, where f^* is the minimum objective value found during the first 2000 iterations across all methods. The algorithmic parameters are selected to suit each method well in terms of computational performance.

For each dataset, we plot the function value gap in Figure 1 to illustrate the convergence behavior of all competing methods. From Figure 1, we observe that SCRN-PM and SCRN-RM vastly outperform SCRN-M and SpaRSA. In addition, SCRN-PM and SCRN-RM achieve a comparable performance to CRN in terms the number of iterations, while outperforming CRN in terms of CPU time. These observations indicate that full gradients significantly improve the performance of the SCRN algorithm, bringing it close to that of the deterministic CRN while reducing computation time. However, when SCRN uses stochastic gradients, its convergence becomes much slower and may offer little to no advantage over first-order methods. Furthermore, we observe that SCRN-RM slightly outperforms SCRN-PM, which corroborates our theoretical results.

5.2 Regularized nonlinear least-squares problems

In this subsection, we consider the regularized nonlinear least-squares problem:

$$\min_{x\in\mathbb{R}^n} \ \frac{1}{m}\sum_{i=1}^m (b_i - \phi(a_i^\top x))^2 + \lambda\sum_{j=1}^n \frac{(\gamma x_j)^2}{1+(\gamma x_j)^2}, \tag{15}$$

where $\phi(t) = e^t/(1+e^t)$ denotes the sigmoid function, $\{(a_i,b_i)\}_{1\le i\le m} \subset \mathbb{R}^n \times \mathbb{R}$ is the given data, and $(\lambda,\gamma) = (0.001, 1)$. We consider three datasets 'a9a', 'phishing', and 'w8a' from LIBSVM.

We apply SCRN-PM, SCRN-RM, A-CRN, SCRN-M, and SpaRSA to solve problem (14). All methods are initialized at $[0.5,\dots,0.5]^T$. For SCRN-M, SCRN-PM, and SCRN-RM, we construct gradient and Hessian estimators using the same strategy as described in Section 5.1. For CRN and all SCRN methods, we adopt the Lanczos method used in [8] to solve the cubic regularized subproblems. We compare these methods in terms of the function value gap defined by $f(x^k) - f^*$, where f^* is the minimum objective value found during the first 2000 iterations across all methods. The algorithmic parameters are selected to suit each method well in terms of computational performance.

For each dataset, we plot the function value gap in Figure 2 to illustrate the convergence behavior of all competing methods. As shown in Figure 2, SCRN-PM and SCRN-RM significantly outperform both SCRN-M and SpaRSA. In addition, SCRN-PM and SCRN-RM achieve performance comparable to that of CRN in terms of iteration counts, while outperforming CRN in CPU time. This suggests that using
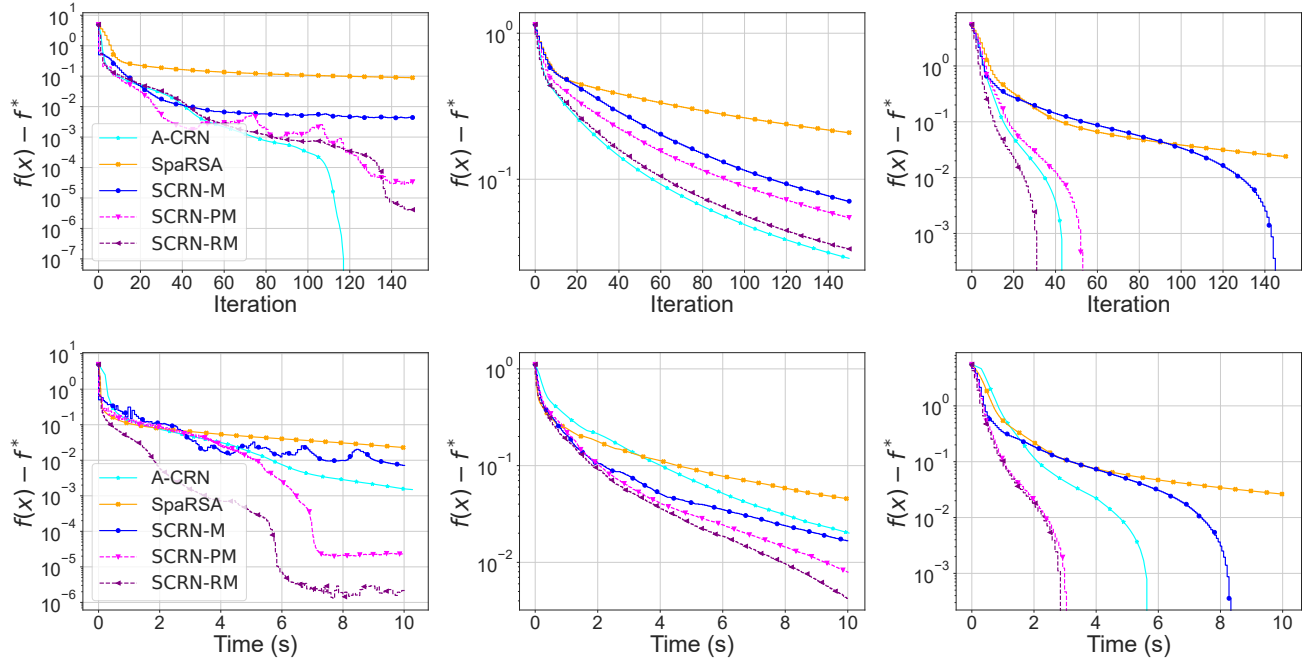
Figure 1: Convergence behavior of objective value gap for problem (14). Correspond to the results on the 'a9a'(left), 'phishing'(middle), and 'w8a'(right) datasets, respectively.
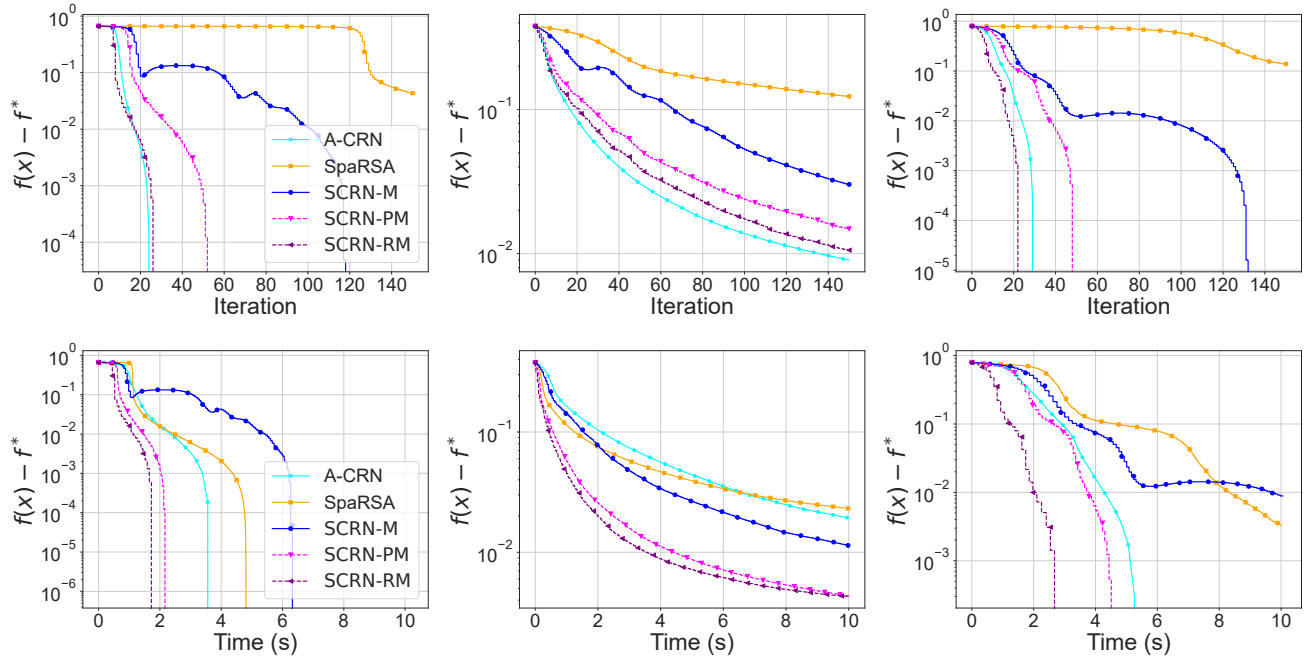

Figure 2: Convergence behavior of objective value gap for problem (15). Correspond to the results on the 'a9a'(left), 'phishing'(middle), and 'w8a'(right) datasets, respectively.

full gradients greatly improves the convergence speed of SCRN, bringing it close to the deterministic CRN while reducing the computational time per iteration. In contrast, when stochastic gradients are used, SCRN converges much more slowly and may offer little to no advantage over first-order methods. In addition, SCRN-RM slightly outperforms SCRN-PM, which is consistent with our theoretical results.

5.3 Robust linear regression

In this subsection, we consider the robust linear regression problem:

$$\min_{x \in \mathbb{R}^n} \frac{1}{m} \sum_{i=1}^{m} \phi(b_i - a_i^\top x), \tag{16}$$

where $\phi(t) = \ln\left(t^2/2 + 1\right)$ is a nonconvex loss function, $\{(a_i, b_i)\}_{1 \le i \le n} \subset \mathbb{R}^m \times \mathbb{R}$ is the given data, and $(\lambda, \gamma) = (0.001, 1)$. We consider three datasets 'ijcnn1', 'phishing', and 'w8a' from LIBSVM.

We apply SCRN-PM, SCRN-RM, A-CRN, SCRN-M, and SpaRSA to solve problem (16). All methods are initialized at $[0.5, \ldots, 0.5]^T$. For SCRN-M, SCRN-PM, and SCRN-RM, we construct gradient and Hessian estimators using the same strategy as described in Section 5.1. For CRN and all SCRN methods, we adopt the Lanczos method used in [8] to solve the cubic regularized subproblems. We compare these methods in terms of the function value gap defined by $f(x^k) - f^*$, where f^* is the minimum objective value found during the first 2000 iterations across all methods. The algorithmic parameters are selected to suit each method well in terms of computational performance.

For each dataset, we plot the function value gap in Figure 3 to demonstrate the convergence behavior of all competing methods. As illustrated, SCRN-PM and SCRN-RM substantially outperform SCRN-M and SpaRSA. In addition, our SCRN-PM and SCRN-RM achieve a comparable performance to CRN in terms of the number of iterations, while outperforming CRN in CPU time. This indicates that incorporating full gradients significantly accelerates the convergence of SCRN, bringing its performance close to that of the deterministic CRN while reducing the computational time per iteration. Conversely, when stochastic gradients are employed, SCRN converges much more slowly and may provide little to no advantage over first-order methods. In addition, SCRN-RM slightly outperforms SCRN-PM, aligning with our theoretical results.

6 Proof of the main results

In this section, we provide proofs of Theorems 1 and 2.

For notational convenience, we define a sequence of potentials for Algorithms 1 and 2 as

$$\mathcal{P}_k := f(x^k) + p_k \|M_k - \nabla^2 f(x^k)\|_F^3 \qquad \forall k \ge 0, \tag{17}$$

where the sequence $\{(x^k, M_k)\}$ is generated by each respective algorithm, and $\{p_k\}$ is a sequence of positive scalars that will be specified separately for each case. We also define the following quantity for measuring the approximate first- and second-order stationarity of problem (1):

$$\mu_\eta(x) := \max\left\{\frac{1}{3}\|\nabla f(x)\|^{3/2}, -\frac{\eta^{3/2}}{4}\lambda_{\min}(\nabla^2 f(x))^3\right\} \tag{18}$$

for some $\eta > 0$.

The following lemma provides expansions for the cubed Frobenius norm, generalizing the well-known identity $\|U + V\|_F^2 = \|U\|_F^2 + 2\text{Tr}(U^T V) + \|V\|_F^2$ and inequality $\|U + V\|_F^2 \le (1+c)\|U\|_F^2 + (1+1/c)\|V\|_F^2$ for all $U, V \in \mathbb{R}^{n \times n}$ and $c > 0$.
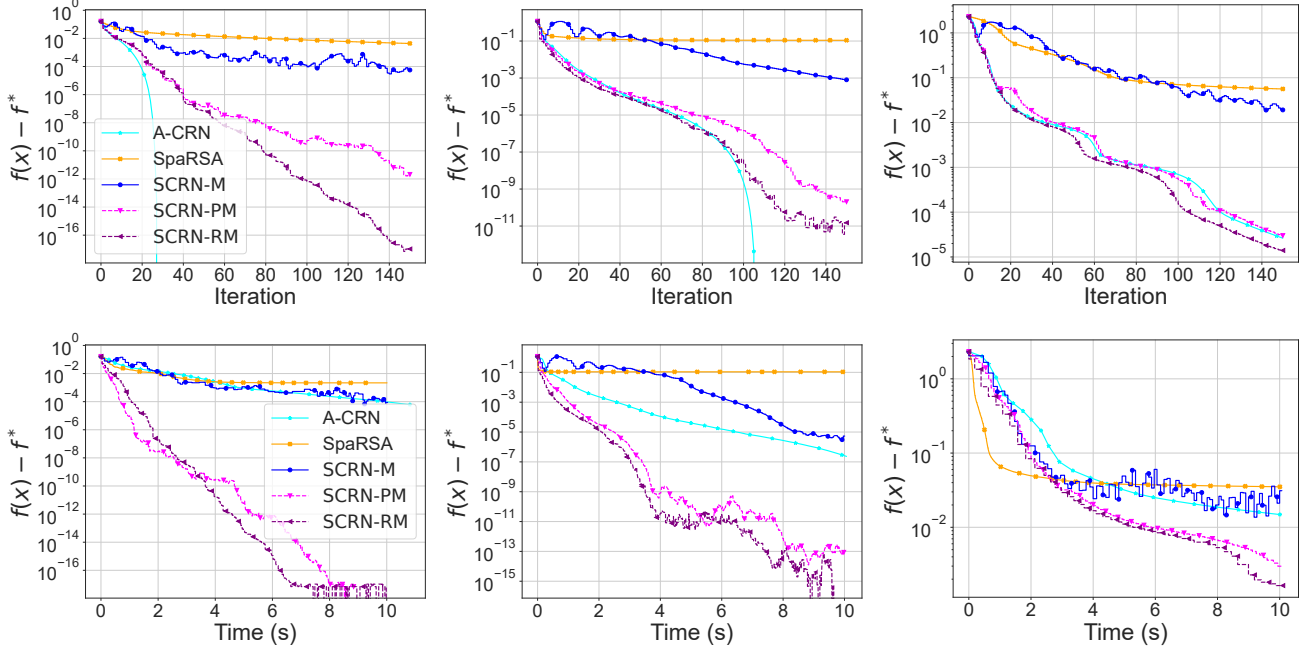
Figure 3: Convergence behavior of objective value gap for problem (16). Correspond to the results on the 'ijcnn1'(left), 'phishing'(middle), and 'w8a'(right) datasets, respectively.

Lemma 1. *For any $U, V \in \mathbb{R}^{n \times n}$, it holds that*

$$\|U + V\|_F^3 \le (1 + c)\|U\|_F^3 + 3\|U\|_F \mathrm{Tr}(U^T V) + 2(1 + c^{-1/2})\|V\|_F^3 \quad \forall c > 0, \tag{19}$$

$$\|U + V\|_F^3 \le (1 + 2c)\|U\|_F^3 + 2(1 + c^{-1/2} + 2c^{-2})\|V\|_F^3 \quad \forall c > 0. \tag{20}$$

Proof. Fix any $U, V \in \mathbb{R}^{n \times n}$. For convenience, we vectorize U and V by letting $u = \mathrm{vec}(U) \in \mathbb{R}^{n^2}$ and $v = \mathrm{vec}(V) \in \mathbb{R}^{n^2}$. Let $\phi(w) := \|w\|^3$ for all $w \in \mathbb{R}^{n^2}$. It follows from [29, Theorem 6.3] that

$$\|\nabla^2 \phi(w) - \nabla^2 \phi(w')\| \le 9\|w - w'\| \quad \forall w, w' \in \mathbb{R}^{n^2}.$$

By this and (5), one has that

$$\phi(u + v) \le \phi(u) + \nabla\phi(u)^T v + \frac{1}{2} v^T \nabla^2 \phi(u) v + \frac{3}{2}\|v\|^3.$$

This together with $\phi(u) = \|u\|^3$, $\nabla\phi(u) = 3\|u\|u$, and $\nabla^2\phi(u) = 3(uu^T/\|u\| + \|u\|I)$ implies that

$$\|u + v\|^3 \le \|u\|^3 + 3\|u\|u^T v + 3v^T(uu^T/\|u\| + \|u\|I)v/2 + 2\|v\|^3$$
$$\le \|u\|^3 + 3\|u\|u^T v + 3\|u\|\|v\|^2 + 2\|v\|^3 \le (1 + c)\|u\|^3 + 3\|u\|u^T v + 2(1 + c^{-1/2})\|v\|^3 \quad \forall c > 0, \tag{21}$$

where the last inequality is due to the Young's inequality. Using again the Young's inequality and (21), we obtain that

$$\|u + v\|^3 \le (1 + 2c)\|u\|^3 + 2(1 + c^{-1/2} + 2c^{-2})\|v\|^3 \quad \forall c > 0. \tag{22}$$

In view of (21), (22), $u = \mathrm{vec}(U)$, and $v = \mathrm{vec}(V)$, we see that (19) and (20) hold as desired. $\qquad \square$

6.1 Some properties of cubic subproblems

In this subsection, we present some properties of the cubic regularized subproblem:

$$x^+ \in \operatorname*{Arg\,min}_{x' \in \mathbb{R}^n} \left\{ g^T(x' - x) + \frac{1}{2}(x' - x)^T M(x' - x) + \frac{1}{6\eta}\|x' - x\|^3 \right\} \tag{23}$$

for given $x \in \mathbb{R}^n$, $M \in \mathbb{R}^{n \times n}$, and $\eta > 0$. The first- and second-order optimality condition of (23) yield

$$g + M(x^+ - x) + \frac{1}{2\eta}\|x^+ - x\|(x^+ - x) = 0, \quad M + \frac{1}{2\eta}\|x^+ - x\|I \succeq 0. \tag{24}$$

The next lemma provides an upper bound for the first- and second-order stationary measure at x^+, which can be seen as an inexact variant of [28, Lemma 5].

Lemma 2. *Suppose that Assumption 1 holds. Assume that $\eta \in (0, (2L)^{-1})$ holds, where L is given in Assumption 1(b). Let $x \in \mathbb{R}^n$ and $M \in \mathbb{R}^{n \times n}$ be given, and let x^+ be a solution to (23). Then,*

$$\|\nabla f(x^+)\|^{3/2} \le \frac{3}{\eta^{3/2}}\|x^+ - x\|^3 + \frac{3\eta^{3/2}}{4}\|M - \nabla^2 f(x)\|_F^3 + 3\|g - \nabla f(x)\|^{3/2}, \tag{25}$$

$$-\eta^{3/2}\lambda_{\min}(\nabla^2 f(x^+))^3 \le \frac{4}{\eta^{3/2}}\|x^+ - x\|^3 + 4\eta^{3/2}\|\nabla^2 f(x) - M\|_F^3. \tag{26}$$

Consequently, one has

$$\mu_\eta(x^+) \le \eta^{-3/2}\|x^+ - x\|^3 + \eta^{3/2}\|M - \nabla^2 f(x)\|_F^3 + \|g - \nabla f(x)\|^{3/2}, \tag{27}$$

where μ_η is defined in (18).

Proof. Using (4) with $y = x^+$, we obtain that

$$\|\nabla f(x^+) - g - M(x^+ - x) + g - \nabla f(x) + (M - \nabla^2 f(x))(x^+ - x)\| \le \frac{L}{2}\|x^+ - x\|^2.$$

This along with the first relation in (24) implies that

$$\begin{aligned}
\|\nabla f(x^+)\| &\le \frac{L}{2}\|x^+ - x\|^2 + \|g + M(x^+ - x)\| + \|g - \nabla f(x)\| + \|(M - \nabla^2 f(x))(x^+ - x)\| \\
&\overset{(24)}{=} \left(\frac{L}{2} + \frac{1}{2\eta}\right)\|x^+ - x\|^2 + \|g - \nabla f(x)\| + \|(M - \nabla^2 f(x))(x^+ - x)\| \\
&\le \frac{3}{4\eta}\|x^+ - x\|^2 + \|g - \nabla f(x)\| + \|M - \nabla^2 f(x)\|\|x^+ - x\| \\
&\le \frac{5}{4\eta}\|x^+ - x\|^2 + \frac{\eta}{2}\|M - \nabla^2 f(x)\|^2 + \|g - \nabla f(x)\|
\end{aligned}$$

where the second inequality is due to the spectral norm inequality and $L \le 1/(2\eta)$, and the third inequality follows from the Young's inequality. This inequality further implies that

$$\begin{aligned}
\|\nabla f(x^+)\|^{3/2} &\le \left(\frac{5}{4\eta}\|x^+ - x\|^2 + \frac{\eta}{2}\|M - \nabla^2 f(x)\|^2 + \|g - \nabla f(x)\|\right)^{3/2} \\
&\le \sqrt{3}\left(\left(\frac{5}{4\eta}\right)^{3/2}\|x^+ - x\|^3 + \left(\frac{\eta}{2}\right)^{3/2}\|M - \nabla^2 f(x)\|^3 + \|g - \nabla f(x)\|^{3/2}\right) \\
&\le \frac{3}{\eta^{3/2}}\|x^+ - x\|^3 + \frac{3\eta^{3/2}}{4}\|M - \nabla^2 f(x)\|^3 + 3\|g - \nabla f(x)\|^{3/2},
\end{aligned}$$

11

where the second inequality is due to $(a + b + c)^{3/2} \le \sqrt{3}(a^{3/2} + b^{3/2} + c^{3/2})$ for all $a, b, c \ge 0$. This together with the fact that the spectral norm of a matrix is bounded above by the Frobenius norm proves (25) as desired.

We next prove (26). Using the Lipschitz continuity of $\nabla^2 f$ and (24), we obtain that

$$\nabla^2 f(x^+) \succeq \nabla^2 f(x) - L\|x^+ - x\|I \succeq M - \|M - \nabla^2 f(x)\|I - L\|x^+ - x\|I$$

$$\overset{(24)}{\succeq} -((2\eta)^{-1} + L)\|x^+ - x\|I - \|M - \nabla^2 f(x)\|I \succeq \eta^{-1}\|x^+ - x\|I - \|M - \nabla^2 f(x)\|I,$$

where the last relation is due to $L \le (2\eta)^{-1}$. It then follows that

$$-\lambda_{\min}(\nabla^2 f(x^+))^3 \le (\eta^{-1}\|x^+ - x\| + \|M - \nabla^2 f(x)\|)^3 \le 4\eta^{-3}\|x^+ - x\|^3 + 4\|M - \nabla^2 f(x)\|^3,$$

where the second inequality is due to $(a + b)^3 \le 4a^3 + 4b^3$ for all $a, b \ge 0$. In view of the above inequality and the fact that the spectral norm of a matrix is bounded above by the Frobenius norm, we obtain that (26) holds.

Combining (25) and (26) with the definition of μ_η in (18), we obtain that (27) holds, which completes the proof of this lemma. $\qquad\square$

We next show that solving a cubic regularized subproblem yields a descent property of f.

Lemma 3. *Suppose that Assumption 1 holds. Assume that $\eta \in (0, (2L)^{-1})$ holds, where L is given in Assumption 1(b). Let $x \in \mathbb{R}^n$ and $M \in \mathbb{R}^{n \times n}$ be given, and let x^+ be a solution to (23). Then,*

$$f(x^+) \le f(x) - \frac{1}{9\eta}\|x^+ - x\|^3 + 24\eta^2\|\nabla^2 f(x) - M\|_F^3 + 3\eta^{1/2}\|\nabla f(x) - g\|^{3/2}. \tag{28}$$

Proof. It follows from (24) that

$$g^T(x^+ - x) = -(x^+ - x)^T M(x^+ - x) - \frac{1}{2\eta}\|x^+ - x\|^3, \tag{29}$$

$$-(x^+ - x)^T M(x^+ - x) \le \frac{1}{2\eta}\|x^+ - x\|^3. \tag{30}$$

Using these and (5) with $y = x^+$, we obtain that

$$f(x^+) \overset{(5)}{\le} f(x) + \nabla f(x)^T(x^+ - x) + \frac{1}{2}(x^+ - x)^T \nabla^2 f(x)(x^+ - x) + \frac{L}{6}\|x^+ - x\|^3$$

$$= f(x) + g^T(x^+ - x) + \frac{1}{2}(x^+ - x)^T M(x^+ - x) + \frac{L}{6}\|x^+ - x\|^3$$

$$\quad + (\nabla f(x) - g)^T(x^+ - x) + \frac{1}{2}(x^+ - x)^T(\nabla^2 f(x) - M)(x^+ - x)$$

$$\overset{(29)}{=} f(x) - \frac{1}{2}(x^+ - x)^T M(x^+ - x) - \left(\frac{1}{2\eta} - \frac{L}{6}\right)\|x^+ - x\|^3$$

$$\quad + (\nabla f(x) - g)^T(x^+ - x) + \frac{1}{2}(x^+ - x)^T(\nabla^2 f(x) - M)(x^+ - x)$$

$$\overset{(30)}{\le} f(x) - \left(\frac{1}{4\eta} - \frac{L}{6}\right)\|x^+ - x\|^3 + (\nabla f(x) - g)^T(x^+ - x) + \frac{1}{2}(x^+ - x)^T(\nabla^2 f(x) - M)(x^+ - x)$$

$$\le f(x) - \frac{1}{6\eta}\|x^+ - x\|^3 + \|x^+ - x\|\|\nabla f(x) - g\| + \frac{1}{2}\|x^+ - x\|^2\|\nabla^2 f(x) - M\|$$

$$\le f(x) - \frac{1}{9\eta}\|x^+ - x\|^3 + 24\eta^2\|\nabla^2 f(x) - M\|^3 + 3\eta^{1/2}\|\nabla f(x) - g\|^{3/2},$$

where the third inequality is due to $L \leq (2\eta)^{-1}$ and the spectral norm inequality, and the last inequality is due to Young's inequality in two forms: $ab \leq a^3/(36\eta) + 2\sqrt{12}\eta^{1/2}b^{3/2}/3$ and $ab \leq a^{3/2}/(18\eta) + 48\eta^2 b^3$ for all $a, b > 0$. In view of the above inequality and the fact that the spectral norm of a matrix is bounded above by the Frobenius norm, we obtain that this lemma holds as desired. $\quad\square$

6.2 Proof of the main results in Section 3

In this subsection, we present some technical lemmas and then use them to prove Theorem 1. The following lemma gives the recurrence for the estimation error of the Hessian estimators $\{M_k\}$ generated by Algorithm 1.

Lemma 4. *Suppose that Assumption 1 holds. Let $\{(x^k, M_k)\}$ be the sequence generated by Algorithm 1 with momentum parameters $\{\theta_k\}$. Then, it holds that for all $k \geq 0$,*

$$\mathbb{E}_{\xi^{k+1}}[\|M_{k+1} - \nabla^2 f(x^{k+1})\|_F^3] \leq (1-\theta_k)\|M_k - \nabla^2 f(x^k)\|_F^3 + 21L_F^3\theta_k^{-2}\|x^{k+1} - x^k\|^3 + 5\sigma^3\theta_k^{5/2}, \quad (31)$$

where L_F and σ are given in Assumption 1.

Proof. Fix any $k \geq 0$. It follows from (6) that

$$M_{k+1} - \nabla^2 f(x^{k+1}) \overset{(6)}{=} (1-\theta_k)M_k + \theta_k H(x^{k+1}; \xi^{k+1}) - \nabla^2 f(x^{k+1})$$
$$= (1-\theta_k)(M_k - \nabla^2 f(x^k)) + (1-\theta_k)(\nabla^2 f(x^k) - \nabla^2 f(x^{k+1})) + \theta_k(H(x^{k+1}; \xi^{k+1}) - \nabla^2 f(x^{k+1})). \quad (32)$$

Observe from Assumption 1 that $\|\nabla^2 f(x^{k+1}) - \nabla^2 f(x^k)\|_F \leq L_F\|x^{k+1} - x^k\|$, $\mathbb{E}_{\xi^{k+1}}[H(x^{k+1}; \xi^{k+1})] = \nabla^2 f(x^{k+1})$ and $\mathbb{E}_{\xi^{k+1}}[\|H(x^{k+1}; \xi^{k+1}) - \nabla^2 f(x^{k+1})\|_F^3] \leq \sigma^3$. Using these, (19), (20), and (32), we obtain that for all $c > 0$,

$$\mathbb{E}_{\xi^{k+1}}[\|M_{k+1} - \nabla^2 f(x^{k+1})\|_F^3]$$
$$\overset{(32)}{=} \mathbb{E}_{\xi^{k+1}}\left[\|(1-\theta_k)(M_k - \nabla^2 f(x^k)) + (1-\theta_k)(\nabla^2 f(x^k) - \nabla^2 f(x^{k+1})) + \theta_k(H(x^{k+1}; \xi^{k+1}) - \nabla^2 f(x^{k+1}))\|_F^3\right]$$
$$\overset{(19)}{\leq} (1+c)\|(1-\theta_k)(M_k - \nabla^2 f(x^k)) + (1-\theta_k)(\nabla^2 f(x^k) - \nabla^2 f(x^{k+1}))\|_F^3$$
$$\quad + 2(1+c^{-1/2})\mathbb{E}_{\xi^{k+1}}\left[\|\theta_k(H(x^{k+1}; \xi^{k+1}) - \nabla^2 f(x^{k+1}))\|_F^3\right]$$
$$\overset{(20)}{\leq} (1+c)(1+2c)(1-\theta_k)^3\|M_k - \nabla^2 f(x^k)\|_F^3$$
$$\quad + 2(1+c)(1+c^{-1/2} + 2c^{-2})(1-\theta_k)^3\|\nabla^2 f(x^{k+1}) - \nabla^2 f(x^k)\|_F^3$$
$$\quad + 2(1+c^{-1/2})\theta_k^3\mathbb{E}_{\xi^{k+1}}\left[\|H(x^{k+1}; \xi^{k+1}) - \nabla^2 f(x^{k+1})\|_F^3\right]$$
$$\leq (1+c)(1+2c)(1-\theta_k)^3\|M_k - \nabla^2 f(x^k)\|_F^3 + 2(1+c)(1+c^{-1/2} + 2c^{-2})(1-\theta_k)^3 L_F^3\|x^{k+1} - x^k\|^3$$
$$\quad + 2(1+c^{-1/2})\sigma^3\theta_k^3, \quad (33)$$

where the first inequality is due to (19) and $\mathbb{E}_{\xi^{k+1}}[H(x^{k+1}; \xi^{k+1})] = \nabla^2 f(x^{k+1})$, the second inequality follows from (20), and the last inequality follows from $\|\nabla^2 f(x^{k+1}) - \nabla^2 f(x^k)\|_F \leq L_F\|x^{k+1} - x^k\|$ and $\mathbb{E}_{\xi^{k+1}}[\|H(x^{k+1}; \xi^{k+1}) - \nabla^2 f(x^{k+1})\|_F^3] \leq \sigma^3$.

Letting $c = \theta_k/(2(1-\theta_k))$ in (33) and using $\theta_k \in (0, 1)$, we obtain that $c^{-1/2} = (2(1-\theta_k)/\theta_k)^{1/2} \leq \sqrt{2}\theta_k^{-1/2}$ and $c^{-2} = 4(1-\theta_k)^2/\theta_k^2 \leq 4\theta_k^{-2}$. Combining these with (33), we obtain that

$$\mathbb{E}_{\xi^{k+1}}\left[\|M_{k+1} - \nabla^2 f(x^{k+1})\|_F^3\right] \leq (1-\theta_k/2)(1-\theta_k)\|M_k - \nabla^2 f(x^k)\|_F^3$$
$$\quad + 2(1-\theta_k/2)(1-\theta_k)^2(1 + \sqrt{2}\theta_k^{-1/2} + 8\theta_k^{-2})L_F^3\|x^{k+1} - x^k\|^3 + 2(1 + \sqrt{2}\theta_k^{-1/2})\sigma^3\theta_k^3$$
$$\leq (1-\theta_k)\|M_k - \nabla^2 f(x^k)\|_F^3 + 21L_F^3\theta_k^{-2}\|x^{k+1} - x^k\|^3 + 5\sigma^3\theta_k^{5/2},$$

where the second inequality is due to $\theta_k \in (0, 1)$. Hence, the conclusion of this lemma holds as desired. $\quad\square$

The following lemma establishes a descent property for the potential sequence $\{\mathcal{P}_k\}$ defined below.

Lemma 5. *Suppose that Assumption 1 holds. Let $\{(x^k, M_k)\}$ be the sequence generated by Algorithm 1 with input parameters $\{(\eta_k, \theta_k)\}$. Assume that $\{\eta_k\} \subset (0, (2L)^{-1})$ and $\{\theta_k\} \subset (0,1)$, where L is given in Assumption 1(b). Let $\{\mathcal{P}_k\}$ be defined in (17) for $\{(x^k, M_k)\}$ and any positive sequence $\{p_k\}$ satisfying*

$$p_{k+1} = \frac{\theta_k^2}{378 L_F^3 \eta_k}, \quad \frac{433\eta_k^2}{18} + (1 - \theta_k)p_{k+1} \leq p_k \quad \forall k \geq 0, \tag{34}$$

where L_F is given in Assumption 1(b). Then, it holds that

$$\mathbb{E}_{\xi^{k+1}}[\mathcal{P}_{k+1}] \leq \mathcal{P}_k - \eta_k^{1/2} \mu_{\eta_k}(x^{k+1})/18 + 55\eta_k^{1/2}\|g^k - \nabla f(x^k)\|^{3/2}/18 + 5\sigma^3\theta_k^{5/2}p_{k+1} \quad \forall k \geq 0, \tag{35}$$

where σ is given in Assumption 1(c), and μ_η is defined in (18).

Proof. Fix any $k \geq 0$. Notice that $\eta_k \in (0, (2L)^{-1})$. It follows from (27) and (28) with $(x^+, x, M, \eta) = (x^{k+1}, x^k, M_k, \eta_k)$ that

$$\mu_{\eta_k}(x^{k+1}) \leq \eta_k^{-3/2}\|x^{k+1} - x^k\|^3 + \eta_k^{3/2}\|M_k - \nabla^2 f(x^k)\|_F^3 + \|g^k - \nabla f(x^k)\|^{3/2}, \tag{36}$$

$$f(x^{k+1}) \leq f(x^k) - (9\eta_k)^{-1}\|x^{k+1} - x^k\|^3 + 24\eta_k^2\|\nabla^2 f(x^k) - M_k\|_F^3 + 3\eta_k^{1/2}\|\nabla f(x^k) - g^k\|^{3/2}. \tag{37}$$

Combining these with (17) and (31) , we obtain that

$$
\begin{aligned}
\mathbb{E}_{\xi^{k+1}}[\mathcal{P}_{k+1}] &\overset{(17)}{=} \mathbb{E}_{\xi^{k+1}}[f(x^{k+1}) + p_{k+1}\|M_{k+1} - \nabla^2 f(x^{k+1})\|_F^3] \\
&\overset{(31)(37)}{\leq} f(x^k) - ((9\eta_k)^{-1} - 21L_F^3\theta_k^{-2}p_{k+1})\|x^{k+1} - x^k\|^3 \\
&\quad + (24\eta_k^2 + (1 - \theta_k)p_{k+1})\|M_k - \nabla^2 f(x^k)\|_F^3 + 3\eta_k^{1/2}\|g^k - \nabla f(x^k)\|^{3/2} + 5\sigma^3\theta_k^{5/2}p_{k+1} \\
&\overset{(36)}{\leq} f(x^k) - \eta_k^{3/2}((9\eta_k)^{-1} - 21L_F^3\theta_k^{-2}p_{k+1})\mu_{\eta_k}(x^{k+1}) \\
&\quad + (\eta_k^3((9\eta_k)^{-1} - 21L_F^3\theta_k^{-2}p_{k+1}) + 24\eta_k^2 + (1 - \theta_k)p_{k+1})\|M_k - \nabla^2 f(x^k)\|_F^3 \\
&\quad + (3\eta_k^{1/2} + \eta_k^{3/2}((9\eta_k)^{-1} - 21L_F^3\theta_k^{-2}p_{k+1}))\|g^k - \nabla f(x^k)\|^{3/2} + 5\sigma^3\theta_k^{5/2}p_{k+1} \\
&= f(x^k) - \eta_k^{1/2}\mu_{\eta_k}(x^{k+1})/18 + (433\eta_k^2/18 + (1 - \theta_k)p_{k+1})\|M_k - \nabla^2 f(x^k)\|_F^3 \\
&\quad + 55\eta_k^{1/2}\|g^k - \nabla f(x^k)\|^{3/2}/18 + 5\sigma^3\theta_k^{5/2}p_{k+1} \\
&\overset{(17)}{\leq} \mathcal{P}_k - \eta_k^{1/2}\mu_{\eta_k}(x^{k+1})/18 + 55\eta_k^{1/2}\|g^k - \nabla f(x^k)\|^{3/2}/18 + 5\sigma^3\theta_k^{5/2}p_{k+1},
\end{aligned}
$$

where the second equality is due to $p_{k+1} = \theta_k^2/(378L_F^3\eta_k)$, and the last inequality follows from (17) and $433\eta_k^2/18 + (1 - \theta_k)p_{k+1} \leq p_k$. The conclusion (35) then follows from the above inequality. $\square$

We are now ready to prove Theorem 1.

Proof of Theorem 1. For convenience, let $\eta = 1/(9K^{2/7})$. Then, we have $\eta_k = \eta$, $\theta_k = 21L_F\eta$, and $\delta_k = 9\eta^2$ for all $k \geq 0$. Also, we define $p_k = 7\eta/(6L_F)$ for all $k \geq 0$. Then, one can verify that (34) holds for $\{(\eta_k, \theta_k, \delta_k)\}$ defined in (8) and $\{p_k\}$ defined above. In addition, by (8), one has that $\{\eta_k\} \subset (0, (2L)^{-1})$ and $\{\theta_k\} \subset (0,1)$ holds for all $K \geq \max\{(2L/9)^{7/2}, (7L_F/3)^{7/2}, 1\}$, Thus, Lemma 5 holds for $\{(\eta_k, \theta_k, \delta_k)\}$ defined in (8) and $\{p_k\}$ defined above. By the definition of $\{p_k\}$, $M_0 = H(x^0; \xi^0)$, (2) and (3), one has

$$\mathbb{E}[\mathcal{P}_0] = f(x^0) + p_0\mathbb{E}[\|M_0 - \nabla^2 f(x^0)\|_F^3] \leq f(x^0) + p_0\sigma^3 = f(x^0) + 7\eta\sigma^3/(6L_F), \tag{38}$$

$$\mathbb{E}[\mathcal{P}_K] = \mathbb{E}[f(x^K) + p_K\|M_K - \nabla^2 f(x^K)\|_F^3] \geq f_{\text{low}}. \tag{39}$$

Notice that $\mathbb{E}_{\zeta^k}[\|g^k - \nabla f(x^k)\|^{3/2}] \leq \delta_k^{3/2}$. Taking expectation of both sides of (35) with respect to $\{\xi^i\}_{i=0}^{k+1}$ and $\{\zeta^i\}_{i=0}^{k}$, and substituting $\eta_k = \eta$, $\theta_k = 21L_F\eta$, $\delta_k = 9\eta^2$ and $p_k = 7\eta/(6L_F)$, we obtain that for all $k \geq 0$,

$$\mathbb{E}[\mathcal{P}_{k+1}] \leq \mathbb{E}[\mathcal{P}_k] - \eta^{1/2}\mathbb{E}[\mu_\eta(x^{k+1})]/18 + (11789\sigma^3 L_F^{3/2} + 83)\eta^{7/2}.$$

Summing up this inequality over $k = 0, \ldots, K-1$, and using (38) and (39), we can see that for all $K \geq \max\{(2L/9)^{7/2}, (7L_F/3)^{7/2}, 1\}$,

$$f_{\text{low}} \overset{(38)}{\leq} \mathbb{E}[\mathcal{P}_K] \leq \mathbb{E}[\mathcal{P}_0] - (\eta^{1/2}/18)\sum_{k=0}^{K-1}\mathbb{E}[\mu_\eta(x^{k+1})] + (11789\sigma^3 L_F^{3/2} + 83)K\eta^{7/2}$$

$$\overset{(39)}{\leq} f(x^0) + 7\eta\sigma^3/(6L_F) - (\eta^{1/2}/18)\sum_{k=0}^{K-1}\mathbb{E}[\mu_\eta(x^{k+1})] + (11789\sigma^3 L_F^{3/2} + 83)K\eta^{7/2}.$$

Rearranging the terms of this inequality and using $\eta = 1/(9K^{2/7})$, we obtain the following holds for all $K \geq \max\{(2L/9)^{7/2}, (7L_F/3)^{7/2}, 1\}$,

$$\frac{1}{K}\sum_{k=0}^{K-1}\mathbb{E}[\mu_\eta(x^{k+1})] \leq 18\Big(\frac{f(x^0) - f_{\text{low}} + 7\eta\sigma^3/(6L_F)}{K\eta^{1/2}} + (11789\sigma^3 L_F^{3/2} + 83)\eta^3\Big)$$

$$\leq 54(f(x^0) - f_{\text{low}} + \sigma^3/(L_F^2) + L_F^{3/2}\sigma^3 + 1)K^{-6/7} \overset{(7)}{=} M_{\text{pm}}K^{-6/7}.$$

Recall that ι_K is uniformly drawn from $\{1, \ldots, K\}$. This along with the above inequality implies that for all $K \geq \max\{(2L/9)^{7/2}, (7L_F/3)^{7/2}, 1\}$,

$$\mathbb{E}[\mu_\eta(x^{\iota_K})] = \frac{1}{K}\sum_{k=0}^{K-1}\mathbb{E}[\mu_\eta(x^{k+1})] \leq M_{\text{pm}}K^{-6/7},$$

which along with the definition of μ_η in (18) and the fact that $\eta = 1/(9K^{2/7})$ implies the following holds for all $K \geq \max\{(2L/9)^{7/2}, (7L_F/3)^{7/2}, 1\}$,

$$\mathbb{E}[\|\nabla f(x^{\iota_K})\|^{3/2}] \leq 3M_{\text{pm}}K^{-6/7}, \quad \mathbb{E}[\lambda_{\min}(\nabla f(x^{\iota_K}))^3] \geq -4M_{\text{pm}}K^{-6/7}\eta^{-3/2} = -108M_{\text{pm}}K^{-3/7}.$$

In view of this, we can see that x^{ι_K} is an (ϵ_g, ϵ_H)-SSOSP of (1) for all K satisfying (9). Hence, the conclusion of this theorem holds as desired. $\qquad\square$

6.3 Proof of the main results in Section 4

In this subsection, we present some technical lemmas and then use them to prove Theorem 2. The following lemma gives the recurrence for the estimation error of the Hessian estimators $\{M_k\}$ generated by Algorithm 2.

Lemma 6. *Suppose that Assumptions 1 and 2 hold. Let $\{(x^k, M_k)\}$ be the sequence generated by Algorithm 2 with momentum parameters $\{\theta_k\}$. Then, it holds that for all $k \geq 0$,*

$$\mathbb{E}_{\xi^{k+1}}[\|M^{k+1} - \nabla^2 f(x^{k+1})\|_F^3] \leq (1-\theta_k)\|M_k - \nabla^2 f(x^k)\|_F^3 + 36(L_F^3 + L_H^3)\theta_k^{-1/2}\|x^{k+1} - x^k\|^3 + 36\theta_k^{5/2}\sigma^3, \tag{40}$$

where L_F and σ are given in Assumption 1, and L_H is given in Assumption 2.

Proof. Fix any $k \geq 0$. It follows from (10) that

$$M_{k+1} - \nabla^2 f(x^{k+1}) \overset{(10)}{=} (1 - \theta_k)(M_k - \nabla^2 f(x^k)) + H(x^{k+1}; \xi^{k+1}) - \nabla^2 f(x^{k+1})$$
$$+ (1 - \theta_k)(\nabla^2 f(x^k) - H(x^k; \xi^{k+1})) \tag{41}$$

Observe from Assumptions 1 and 2 that $\|\nabla^2 f(x^{k+1}) - \nabla^2 f(x^k)\|_F \leq L_F \|x^{k+1} - x^k\|$, $\mathbb{E}_{\xi^{k+1}}[H(x^{k+1}; \xi^{k+1})] = \nabla^2 f(x^{k+1})$, $\mathbb{E}_{\xi^{k+1}}[\|H(x^{k+1}; \xi^{k+1}) - \nabla^2 f(x^{k+1})\|_F^3] \leq \sigma^3$, and $\mathbb{E}_{\xi^{k+1}}[\|H(x^{k+1}; \xi^{k+1}) - H(x^k; \xi^{k+1})\|_F^3] \leq L_H^3 \|x^{k+1} - x^k\|_F^3$. Using these, (19), and (41), we obtain that

$$\mathbb{E}_{\xi^{k+1}}[\|M_{k+1} - \nabla^2 f(x^{k+1})\|_F^3]$$

$$\overset{(41)}{=} \mathbb{E}_{\xi^{k+1}}[\|(1 - \theta_k)(M_k - \nabla^2 f(x^k)) + H(x^{k+1}; \xi^{k+1}) - \nabla^2 f(x^{k+1}) + (1 - \theta_k)(\nabla^2 f(x^k) - H(x^k; \xi^{k+1}))\|_F^3]$$

$$\overset{(19)}{\leq} (1 + c)\|(1 - \theta_k)(M_k - \nabla^2 f(x^k))\|_F^3$$
$$+ 2(1 + c^{-1/2})\mathbb{E}_{\xi^{k+1}}\|H(x^{k+1}; \xi^{k+1}) - \nabla^2 f(x^{k+1}) + (1 - \theta_k)(\nabla^2 f(x^k) - H(x^k; \xi^{k+1}))\|_F^3$$

$$= (1 + c)(1 - \theta_k)^3\|M_k - \nabla^2 f(x^k)\|_F^3 + 2(1 + c^{-1/2})\mathbb{E}_{\xi^{k+1}}\|H(x^{k+1}; \xi^{k+1}) - H(x^k; \xi^{k+1})$$
$$+ \nabla^2 f(x^k) - \nabla^2 f(x^{k+1}) - \theta_k(\nabla^2 f(x^k) - H(x^k; \xi^{k+1}))\|_F^3$$

$$\leq (1 + c)(1 - \theta_k)^3\|M_k - \nabla^2 f(x^k)\|_F^3 + 18(1 + c^{-1/2})\mathbb{E}_{\xi^{k+1}}[\|H(x^{k+1}; \xi^{k+1}) - H(x^k; \xi^{k+1})\|_F^3]$$
$$+ 18(1 + c^{-1/2})\|\nabla^2 f(x^k) - \nabla^2 f(x^{k+1})\|_F^3 + 18(1 + c^{-1/2})\theta_k^3 \mathbb{E}_{\xi^{k+1}}[\|\nabla^2 f(x^k) - H(x^k; \xi^{k+1})\|_F^3]$$

$$\leq (1 + c)(1 - \theta_k)^3\|M_k - \nabla^2 f(x^k)\|_F^3 + 18(1 + c^{-1/2})(L_F^3 + L_H^3)\|x^k - x^{k+1}\|^3 + 18\sigma^3(1 + c^{-1/2})\theta_k^3, \tag{42}$$

where the first inequality follows from (19) and $\mathbb{E}_{\xi^{k+1}}[H(x^{k+1}; \xi^{k+1})] = \nabla^2 f(x^{k+1})$, the second inequality is due to $\|A + B + C\|_F^3 \leq 9(\|A\|_F^3 + \|B\|_F^3 + \|C\|_F^3)$ for all $A, B, C \in \mathbb{R}^{n \times n}$, and the last inequality follows from $\|\nabla^2 f(x^{k+1}) - \nabla^2 f(x^k)\|_F \leq L_F \|x^{k+1} - x^k\|$, $\mathbb{E}_{\xi^{k+1}}[\|H(x^{k+1}; \xi^{k+1}) - \nabla^2 f(x^{k+1})\|_F^3] \leq \sigma^3$, and $\mathbb{E}_{\xi^{k+1}}[\|H(x^{k+1}; \xi^{k+1}) - H(x^k; \xi^{k+1})\|_F^3] \leq L_H^3 \|x^{k+1} - x^k\|_F^3$.

Letting $c = \theta_k/(1 - \theta_k)$ in (42), and using $\theta_k \in (0, 1)$, we obtain $c^{-1/2} = (1 - \theta_k)^{1/2}\theta_k^{-1/2} \leq \theta_k^{-1/2}$. Combining this with (42), we obtain that

$$\mathbb{E}_{\xi^{k+1}}[\|M_{k+1} - \nabla^2 f(x^{k+1})\|_F^3] \leq (1 - \theta_k)^2\|M_k - \nabla^2 f(x^k)\|_F^3$$
$$+ 18(L_F^3 + L_H^3)(1 + \theta_k^{-1/2})\|x^{k+1} - x^k\|^3 + 18\sigma^3(1 + \theta_k^{-1/2})\theta_k^3,$$

which along with $\theta_k \in (0, 1)$ implies that (40) holds as desired. $\square$

The following lemma establishes a descent property for the potential sequence $\{\mathcal{P}_k\}$ defined below.

Lemma 7. *Suppose that Assumptions 1 and 2 hold. Let $\{(x^k, M_k)\}$ be the sequence generated by Algorithm 2 with input parameters $\{(\eta_k, \theta_k)\}$. Assume that $\{\eta_k\} \subset (0, (2L)^{-1})$ and $\{\theta_k\} \subset (0, 1)$, where L is given in Assumption 1(b). Let $\{\mathcal{P}_k\}$ be defined in (17) for $\{(x^k, M_k)\}$ and any positive sequence $\{p_k\}$ satisfying*

$$p_{k+1} = \frac{\theta_k^{1/2}}{648(L_F^3 + L_H^3)\eta_k}, \quad \frac{433\eta_k^2}{18} + (1 - \theta_k)p_{k+1} \leq p_k \quad \forall k \geq 0, \tag{43}$$

where L_F is given in Assumption 1(b) and L_H is given in Assumption 2. Then, it holds that

$$\mathbb{E}_{\xi^{k+1}}[\mathcal{P}_{k+1}] \leq \mathcal{P}_k - \eta_k^{1/2}\mu_{\eta_k}(x^{k+1})/18 + 55\eta_k^{1/2}\|g^k - \nabla f(x^k)\|^{3/2}/18 + 36\theta_k^{5/2}p_{k+1}\sigma^3 \quad \forall k \geq 0, \tag{44}$$

where σ is given in Assumption 1, and $\mu_\eta(x)$ is defined in (18).

16

Proof. Fix any $k \geq 0$. Notice that $\eta_k \in (0, (2L)^{-1})$. It follows from (27) and (28) with $(x^+, x, M, \eta) = (x^{k+1}, x^k, M_k, \eta_k)$ that

$$\mu_{\eta_k}(x^{k+1}) \leq \eta_k^{-3/2}\|x^{k+1} - x^k\|^3 + \eta_k^{3/2}\|M_k - \nabla^2 f(x^k)\|_F^3 + \|g^k - \nabla f(x^k)\|^{3/2}, \tag{45}$$

$$f(x^{k+1}) \leq f(x^k) - (9\eta_k)^{-1}\|x^{k+1} - x^k\|^3 + 24\eta_k^2\|\nabla^2 f(x^k) - M_k\|_F^3 + 3\eta_k^{1/2}\|\nabla f(x^k) - g^k\|^{3/2}. \tag{46}$$

Combining these with (17) and (40), we obtain that

$$\mathbb{E}_{\xi^{k+1}}[\mathcal{P}_{k+1}] \overset{(17)}{=} \mathbb{E}_{\xi^{k+1}}[f(x^{k+1}) + p_{k+1}\|M_{k+1} - \nabla^2 f(x^{k+1})\|_F^3]$$

$$\overset{(40)(46)}{\leq} f(x^k) - ((9\eta_k)^{-1} - 36(L_F^3 + L_H^3)\theta_k^{-1/2}p_{k+1})\|x^{k+1} - x^k\|^3$$
$$+ (24\eta_k^2 + (1 - \theta_k)p_{k+1})\|M_k - \nabla^2 f(x^k)\|_F^3 + 3\eta_k^{1/2}\|g^k - \nabla f(x^k)\|^{3/2} + 36\sigma^3\theta_k^{5/2}p_{k+1}$$

$$\overset{(45)}{\leq} f(x^k) - \eta_k^{3/2}((9\eta_k)^{-1} - 36(L_F^3 + L_H^3)\theta_k^{-1/2}p_{k+1})\mu_{\eta_k}(x^{k+1})$$
$$+ (\eta_k^3((9\eta_k)^{-1} - 36(L_F^3 + L_H^3)\theta_k^{-1/2}p_{k+1}) + 24\eta_k^2 + (1 - \theta_k)p_{k+1})\|M_k - \nabla^2 f(x^k)\|_F^3$$
$$+ (3\eta_k^{1/2} + \eta_k^{3/2}((9\eta_k)^{-1} - 36(L_F^3 + L_H^3)\theta_k^{-1/2}p_{k+1}))\|g^k - \nabla f(x^k)\|^{3/2} + 36\sigma^3\theta_k^{5/2}p_{k+1}$$

$$= f(x^k) - \eta_k^{1/2}\mu_{\eta_k}(x^{k+1})/18 + (433\eta_k^2/18 + (1 - \theta_k)p_{k+1})\|M_k - \nabla^2 f(x^k)\|_F^3$$
$$+ 55\eta_k^{1/2}\|g^k - \nabla f(x^k)\|^{3/2}/18 + 36\sigma^3\theta_k^{5/2}p_{k+1}$$

$$\overset{(17)}{\leq} \mathcal{P}_k - \eta_k^{1/2}\mu_{\eta_k}(x^{k+1})/18 + 55\eta_k^{1/2}\|g^k - \nabla f(x^k)\|^{3/2}/18 + 36\sigma^3\theta_k^{5/2}p_{k+1},$$

where the second equality is due to $p_{k+1} = \theta_k^{1/2}/(648(L_F^3 + L_H^3)\eta_k)$, and the last inequality follows from (17) and $433\eta_k^2/18 + (1 - \theta_k)p_{k+1} \leq p_k$. The conclusion (44) then follows from the above inequality. $\square$

We now provide a proof of Theorem 2.

Proof of Theorem 2. For convenience, let $\eta = 1/(17K^{1/5})$. Then, we have $\eta_k = \eta$, $\theta_k = 625(L_F^3 + L_H^3)^{2/3}\eta^2$ and $\delta_k = 289\eta^3$ for all $k \geq 0$. In addition, we define $p_k = 625^{1/2}/(648(L_F^3 + L_H^3)^{2/3})$ for all $k \geq 0$. Then, one can verify that (43) holds for $\{(\eta_k, \theta_k, \delta_k)\}$ defined in (12) and $\{p_k\}$ defined above. In addition, by (12), one has that $\{\eta_k\} \subset (0, (2L)^{-1})$ and $\{\theta_k\} \subset (0, 1)$ holds for all $K \geq \max\{(2L/17)^5, 7(L_F^3 + L_H^3)^{5/3}, 1\}$, Thus, Lemma 7 holds for $\{(\eta_k, \theta_k, \delta_k)\}$ defined in (12) and $\{p_k\}$ defined above. By the definition of $\{p_k\}$, $M_0 = H(x^0; \xi^0)$, and (3), one has

$$\mathbb{E}[\mathcal{P}_0] = f(x^0) + p_0\mathbb{E}[\|M_0 - \nabla^2 f(x^0)\|_F^3] \leq f(x^0) + p_0\sigma^3 \leq f(x^0) + \sigma^3/(L_F^3 + L_H^3)^{2/3}, \tag{47}$$

$$\mathbb{E}[\mathcal{P}_K] = \mathbb{E}[f(x^K) + p_K\|M_K - \nabla^2 f(x^K)\|_F^3] \geq f_{\text{low}}. \tag{48}$$

Notice that $\mathbb{E}_{\zeta^k}[\|g^k - \nabla f(x^k)\|^{3/2}] \leq \delta_k^{3/2}$. Taking expectation of both sides of (44) with respect to $\{\xi^i\}_{i=0}^{k+1}$ and $\{\zeta^i\}_{i=0}^k$, and substituting $\eta_k = \eta$, $\theta_k = 625(L_F^3 + L_H^3)^{2/3}\eta^2$, and $p_k = 625^{1/2}/(648(L_F^3 + L_H^3)^{2/3})$, we obtain that for all $k \geq 0$,

$$\mathbb{E}[\mathcal{P}_{k+1}] \leq \mathbb{E}[\mathcal{P}_k] - \eta^{1/2}\mathbb{E}[\mu_\eta(x^{k+1})]/18 + \left(\frac{55}{18} \cdot 17^3 + \frac{36 \cdot 625^3}{648}\sigma^3(L_F^3 + L_H^3)\right)\eta^5.$$

Summing this inequality over $k = 0, \ldots, K - 1$, and using (47) and (48), it follows that for all $K \geq \max\{(2L/17)^5, 7(L_F^3 + L_H^3)^{5/3}, 1\}$,

$$f_{\text{low}} \overset{(48)}{\leq} \mathbb{E}[\mathcal{P}_K] \leq \mathbb{E}[\mathcal{P}_0] - (\eta^{1/2}/18)\sum_{k=0}^{K-1}\mathbb{E}[\mu_\eta(x^{k+1})] + \left(\frac{55}{18} \cdot 17^3 + \frac{36 \cdot 625^3}{648}\sigma^3(L_F^3 + L_H^3)\right)K\eta^5$$

$$\overset{(47)}{\leq} f(x^0) + \sigma^3/(L_F^3 + L_H^3)^{2/3} - (\eta^{1/2}/18)\sum_{k=0}^{K-1}\mathbb{E}[\mu_\eta(x^{k+1})] + \left(\frac{55}{18} \cdot 17^3 + \frac{36 \cdot 625^3}{648}\sigma^3(L_F^3 + L_H^3)\right)K\eta^5.$$

17

Rearranging the terms of this inequality and using the definition of $\eta = 1/(17K^{1/5})$, we obtain that for all $K \geq \max\{(2L/17)^5, 7(L_F^3 + L_H^3)^{5/3}, 1\}$,

$$\frac{1}{K} \sum_{k=0}^{K-1} \mathbb{E}[\mu_\eta(x^{k+1})] \leq 18 \left(\frac{f(x^0) - f_{\text{low}} + \sigma^3/(L_F^3 + L_H^3)^{2/3}}{K\eta^{1/2}} + \left(\frac{55}{18} \cdot 17^3 + \frac{36 \cdot 625^3}{648} \sigma^3(L_F^3 + L_H^3) \right) \eta^{9/2} \right)$$

$$\leq 75(f(x^0) - f_{\text{low}} + \sigma^3/(L_F^3 + L_H^3)^{2/3} + (L_F^3 + L_H^3)\sigma^3 + 1)K^{-9/10} \overset{(11)}{=} M_{\text{rm}} K^{-9/10}.$$

Recall that ι_K is uniformly drawn from $\{1, \ldots, K\}$. This along with the above inequality implies that for all $K \geq \max\{(2L/17)^5, 7(L_F^3 + L_H^3)^{5/3}, 1\}$,

$$\mathbb{E}[\mu_\eta(x^{\iota_K})] = \frac{1}{K} \sum_{k=0}^{K-1} \mathbb{E}[\mu_\eta(x^{k+1})] \leq M_{\text{rm}} K^{-9/10},$$

which along with the definition of μ_η in (18) and $\eta = 1/(17K^{1/5})$ implies that for all $K \geq \max\{(2L/17)^5, 7(L_F^3 + L_H^3)^{5/3}, 1\}$,

$$\mathbb{E}[\|\nabla f(x^{\iota_K})\|^{3/2}] \leq 3M_{\text{rm}} K^{-9/10}, \quad \mathbb{E}[\lambda_{\min}(\nabla f(x^{\iota_K}))^3] \geq -4M_{\text{rm}} K^{-9/10} \eta^{-3/2} = -281 M_{\text{rm}} K^{-3/5}.$$

In view of these, we can see that x^{ι_K} is an (ϵ_g, ϵ_H)-SSOSP of (1) for all K satisfying (13). Hence, the conclusion of this theorem holds as desired. $\qquad \square$

Acknowledgment

References

[1] N. Agarwal, Z. Allen-Zhu, B. Bullins, E. Hazan, and T. Ma. Finding approximate local minima faster than gradient descent. In *Proceedings of the 49th Annual ACM SIGACT Symposium on Theory of Computing*, pages 1195–1199, 2017.

[2] Z. Allen-Zhu and Y. Li. Neon2: Finding local minima via first-order oracles. In *Advances in Neural Information Processing Systems*, volume 31, 2018.

[3] S. Bellavia, G. Gurioli, and B. Morini. Adaptive cubic regularization methods with dynamic inexact Hessian information and applications to finite-sum minimization. *IMA Journal of Numerical Analysis*, 41(1):764–799, 2021.

[4] E. H. Bergou, Y. Diouane, V. Kunc, V. Kungurtsev, and C. W. Royer. A subsampling line-search method with second-order results. *INFORMS Journal on Optimization*, 4(4):403–425, 2022.

[5] E. G. Birgin and J. M. Martínez. The use of quadratic regularization with a cubic descent condition for unconstrained optimization. *SIAM Journal on Optimization*, 27(2):1049–1074, 2017.

[6] Y. Carmon and J. Duchi. Gradient descent finds the cubic-regularized nonconvex Newton step. *SIAM Journal on Optimization*, 29(3):2146–2178, 2019.

[7] Y. Carmon, J. C. Duchi, O. Hinder, and A. Sidford. Lower bounds for finding stationary points I. *Mathematical Programming*, 184(1):71–120, 2020.

[8] C. Cartis, N. I. Gould, and P. L. Toint. Adaptive cubic regularisation methods for unconstrained optimization. Part I: Motivation, convergence and numerical results. *Mathematical Programming*, 127(2):245–295, 2011.

[9] C. Cartis, N. I. Gould, and P. L. Toint. Worst-case evaluation complexity and optimality of second-order methods for nonconvex smooth optimization. In *Proceedings of the International Congress of Mathematicians: Rio de Janeiro 2018*, pages 3711–3750, 2018.

[10] E. M. Chayti, N. Doikov, and M. Jaggi. Improving stochastic cubic Newton with momentum. In *International Conference on Artificial Intelligence and Statistics*, volume 258, pages 1441–1449, 2025.

[11] F. E. Curtis, D. P. Robinson, C. W. Royer, and S. J. Wright. Trust-region Newton-CG with strong second-order complexity guarantees for nonconvex optimization. *SIAM Journal on Optimization*, 31(1):518–544, 2021.

[12] F. E. Curtis, D. P. Robinson, and M. Samadi. A trust region algorithm with a worst-case iteration complexity of $\mathcal{O}(\epsilon^{-3/2})$ for nonconvex optimization. *Mathematical Programming*, 162:1–32, 2017.

[13] F. E. Curtis, D. P. Robinson, and M. Samadi. An inexact regularized Newton framework with a worst-case iteration complexity of for nonconvex optimization. *IMA Journal of Numerical Analysis*, 39(3):1296–1327, 2019.

[14] A. Cutkosky and H. Mehta. Momentum improves normalized SGD. In *International Conference on Machine Learning*, pages 2260–2268, 2020.

[15] A. Cutkosky and F. Orabona. Momentum-based variance reduction in non-convex SGD. *Advances in Neural Information Processing Systems*, 32, 2019.

[16] N. Doikov and G. N. Grapiglia. First and zeroth-order implementations of the regularized Newton method with lazy approximated Hessians. *Journal of Scientific Computing*, 103(1):32, 2025.

[17] C. Fang, C. J. Li, Z. Lin, and T. Zhang. Spider: Near-optimal non-convex optimization via stochastic path-integrated differential estimator. In *Advances in neural information processing systems*, volume 31, 2018.

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

[19] G. N. Grapiglia, M. L. Gonçalves, and G. Silva. A cubic regularization of Newton's method with finite difference Hessian approximations. *Numerical Algorithms*, pages 1–24, 2022.

[20] C. He, H. Huang, and Z. Lu. Newton-CG methods for nonconvex unconstrained optimization with Hölder continuous Hessian. *Mathematics of Operations Research*, 2025.

[21] C. He, Z. Lu, and T. K. Pong. A Newton-CG based augmented Lagrangian method for finding a second-order stationary point of nonconvex equality constrained optimization with complexity guarantees. *SIAM Journal on Optimization*, 33(3):1734–1766, 2023.

[22] C. He, Z. Lu, D. Sun, and Z. Deng. Complexity of normalized stochastic first-order methods with momentum under heavy-tailed noise. *arXiv preprint arXiv:2506.11214*, 2025.

[23] M. Jaggi, N. Doikov, et al. Unified convergence theory of stochastic and variance-reduced cubic Newton methods. *Transactions on Machine Learning Research*, 2024.

[24] C. Jin, P. Netrapalli, and M. I. Jordan. Accelerated gradient descent escapes saddle points faster than gradient descent. pages 1042–1085, 2018.

[25] J. M. Kohler and A. Lucchi. Sub-sampled cubic regularization for non-convex optimization. pages 1895–1904, 2017.

[26] Z. Li, H. Bao, X. Zhang, and P. Richtárik. PAGE: A simple and optimal probabilistic gradient estimator for nonconvex optimization. In *International Conference on Machine Learning*, pages 6286–6295, 2021.

[27] J. M. Martínez and M. Raydan. Cubic-regularization counterpart of a variable-norm trust-region method for unconstrained minimization. *Journal of Global Optimization*, 68:367–385, 2017.

[28] Y. Nesterov and B. T. Polyak. Cubic regularization of Newton method and its global performance. *Mathematical Programming*, 108(1):177–205, 2006.

[29] A. Rodomanov and Y. Nesterov. Smoothness parameter of power of Euclidean norm. *Journal of Optimization Theory and Applications*, 185:303–326, 2020.

[30] C. W. Royer, M. O'Neill, and S. J. Wright. A Newton-CG algorithm with complexity guarantees for smooth unconstrained optimization. *Mathematical Programming*, 180:451–488, 2020.

[31] C. W. Royer and S. J. Wright. Complexity analysis of second-order line-search algorithms for smooth nonconvex optimization. *SIAM Journal on Optimization*, 28(2):1448–1477, 2018.

[32] N. Tripuraneni, M. Stern, C. Jin, J. Regier, and M. I. Jordan. Stochastic cubic regularization for fast nonconvex optimization. *Advances in Neural Information Processing Systems*, 31, 2018.

[33] Z. Wang, Y. Zhou, Y. Liang, and G. Lan. A note on inexact gradient and Hessian conditions for cubic regularized Newton's method. *Operations Research Letters*, 47(2):146–149, 2019.

[34] Z. Wang, Y. Zhou, Y. Liang, and G. Lan. Stochastic variance-reduced cubic regularization for nonconvex optimization. pages 2731–2740, 2019.

[35] S. J. Wright, R. D. Nowak, and M. A. Figueiredo. Sparse reconstruction by separable approximation. *IEEE Transactions on Signal Processing*, 57(7):2479–2493, 2009.

[36] P. Xu, F. Roosta, and M. W. Mahoney. Newton-type methods for non-convex optimization under inexact Hessian information. *Mathematical Programming*, 184(1):35–70, 2020.

[37] Y. Xu, R. Jin, and T. Yang. Neon+: Accelerated gradient methods for extracting negative curvature for non-convex optimization. *arXiv preprint arXiv:1712.01033*, 2017.

[38] Z. Yao, P. Xu, F. Roosta, S. J. Wright, and M. W. Mahoney. Inexact Newton-CG algorithms with complexity guarantees. *IMA Journal of Numerical Analysis*, 43(3):1855–1897, 2023.

[39] J. Zhang, L. Xiao, and S. Zhang. Adaptive stochastic variance reduction for subsampled Newton method with cubic regularization. *INFORMS Journal on Optimization*, 4(1):45–64, 2022.

[40] D. Zhou, P. Xu, and Q. Gu. Stochastic variance-reduced cubic regularization methods. *Journal of Machine Learning Research*, 20(134):1–47, 2019.