Stochastic Variance-Reduced Cubic Regularization for Nonconvex Optimization

Zhe Wang
Ohio State University
wang.10982@osu.edu

Yi Zhou
Duke University
yi.zhou610@duke.edu

Yingbin Liang
Ohio State University
liang.889@osu.edu

Guanghui Lan
Georgia Institute of Technology
george.lan@isye.gatech.edu

Abstract

Cubic regularization (CR) is an optimization method with emerging popularity due to its capability to escape saddle points and converge to second-order stationary solutions for nonconvex optimization. However, CR encounters a high sample complexity issue for finite-sum problems with a large data size. In this paper, we propose a stochastic variance-reduced cubic-regularization (SVRC) method under random sampling, and study its convergence guarantee as well as sample complexity. We show that the iteration complexity of SVRC for achieving a second-order stationary solution within ϵ accuracy is $\mathcal{O}(\epsilon^{-3/2})$, which matches the state-of-art result on CR types of methods. Moreover, our proposed variance reduction scheme significantly reduces the per-iteration sample complexity. The resulting total Hessian sample complexity of our SVRC is $\tilde{\mathcal{O}}(N^{2/3}\epsilon^{-3/2})$, which outperforms the state-of-art result by a factor of $\tilde{\mathcal{O}}(N^{2/15})$. We also study our SVRC under random sampling without replacement scheme, which yields a lower per-iteration sample complexity, and hence justifies its practical applicability.

1 Introduction

Many machine learning problems are formulated as finite-sum nonconvex optimization problems that take the form

$$\min_{\mathbf{x}\in\mathbb{R}^d} F(\mathbf{x}) \triangleq \frac{1}{N}\sum_{i=1}^{N} f_i(\mathbf{x}), \tag{1}$$

where each component function f_i corresponds to the loss on the i-th data sample. While finding global optimal solutions of generic nonconvex optimization problems are challenging, various nonconvex problems in the form of eq. (1) have been shown to possess good landscape properties that facilitate convergence. For example, the square loss of a shallow linear neural network is shown to have only strict saddle points other than local minimum (Baldi and Hornik, 1989; Zhou and Liang, 2018). The same property also holds for some other nonconvex problems such as phase retrieval (Sun et al., 2017) and matrix factorization (Ge et al., 2016; Bhojanapalli et al., 2016). Such a remarkable property has motivated a growing research interest in designing algorithms that can escape strict saddle points and have guaranteed convergence to local minimum, and even to global minimum for problems without spurious local minimum.

Various algorithms have been designed to have the capability to escape strict saddle points in nonconvex optimization. Such a desired property requires that the obtained solution $\mathbf{x}^\star$ satisfies the second-order stationary conditions within an ϵ accuracy, i.e.,

$$\|\nabla F(\mathbf{x}^\star)\| \leqslant \epsilon, \qquad \nabla^2 F(\mathbf{x}^\star) \succcurlyeq -\sqrt{\epsilon}\boldsymbol{I}. \tag{2}$$

Therefore, upon convergence, the gradient is guaranteed to be close to zero and the Hessian is guaranteed to be almost positive semidefinite, which thresh-out the possibility to converge to strict saddle points. Among these algorithms (which are reviewed in related work), the cubic-regularized Newton's method (also called cubic regularization or CR) (Nesterov and Polyak, 2006) is a popular method that provides the second-order stationary guarantee for the obtained solution. At each iteration k, CR solves a sub-problem that approximates the objective function in eq. (1) with a cubic-regularized second-order Taylor's expansion at the current iterate $\mathbf{x}_k$. In specific, the update rule of CR can be written as

$$\mathbf{s}_{k+1} = \operatorname*{argmin}_{\mathbf{s}\in\mathbb{R}^d} \nabla F(\mathbf{x}_k)^\top \mathbf{s} + \frac{1}{2}\mathbf{s}^\top \nabla^2 F(\mathbf{x}_k)\mathbf{s} + \frac{M}{6}\|\mathbf{s}\|^3,$$

Proceedings of the 22nd International Conference on Artificial Intelligence and Statistics (AISTATS) 2019, Naha, Okinawa, Japan. PMLR: Volume 89.

$$\mathbf{x}_{k+1} = \mathbf{x}_k + \mathbf{s}_{k+1}. \tag{3}$$

It has been shown that CR converges to a point satisfying the second-order stationary condition (eq. (2)) within $\mathcal{O}(\epsilon^{-3/2})$ number of iterations. However, fully solving the exact cubic sub-problem in eq. (3) requires a high computation complexity, especially due to the computation of the Hessian matrices for loss functions on all the data samples. To evaluate the complexity of CR type algorithms, we define the stochastic Hessian oracle (SHO) as follows. Given a point $\mathbf{x}$ and the component number i, the oracle returns the corresponding Hessian $\nabla^2 f_i(\mathbf{x})$. Moreover, we define the subproblem oracle (SO) as a subroutine, which for a given a point $\mathbf{x}$, returns the minimizer of eq. (3). In Cartis et al. (2011), the authors proposed an inexact cubic-regularized (inexact-CR) Newton's method, which formulates the cubic sub-problem in eq. (3) with an inexact Hessian $\mathbf{H}_k$ that satisfies

$$\left\|(\mathbf{H}_k - \nabla^2 F(\mathbf{x}_k))\mathbf{s}_{k+1}\right\| \leqslant C \left\|\mathbf{s}_{k+1}\right\|^2, \tag{4}$$

where $C \geqslant 0$ is a certain numerical constant. In particular, Cartis et al. (2011) showed that such an inexact method achieves the same order of theoretical guarantee as the original CR. This inexact condition has been explored in various situations (Kohler and Lucchi, 2017; Cartis et al., 2012a,b; Zhou et al., 2018). Especially, in order to satisfy the inexact Hessian condition in eq. (4), Kohler and Lucchi (2017) proposed a practical sub-sampling scheme (referred to SCR) to implement the inexact-CR. Specifically, at each iteration k, SCR collects two index sets $\xi_g(k), \xi_H(k)$ whose elements are sampled uniformly from $\{1, \ldots, N\}$ at random, and then evaluates respectively the gradients and Hessians of the corresponding component functions, i.e., $\mathbf{g}_k \triangleq \frac{1}{|\xi_g(k)|}\sum_{i\in\xi_g(k)} \nabla f_i(\mathbf{x}_k)$ and $\mathbf{H}_k \triangleq \frac{1}{|\xi_H(k)|}\sum_{i\in\xi_H(k)} \nabla^2 f_i(\mathbf{x}_k)$. Then, SCR solves the following cubic sub-problem at the k-th iteration.

$$\mathbf{s}_{k+1} = \operatorname*{argmin}_{\mathbf{s}\in\mathbb{R}^d} \mathbf{g}_k^\top \mathbf{s} + \tfrac{1}{2}\mathbf{s}^\top \mathbf{H}_k \mathbf{s} + \tfrac{M}{6}\left\|\mathbf{s}\right\|^3.$$

Kohler and Lucchi (2017) showed that if the mini-batch sizes to satisfy

$$|\xi_g(k)| \geqslant \mathcal{O}\left(\frac{1}{\left\|\mathbf{s}_{k+1}\right\|^4}\right), |\xi_H(k)| \geqslant \mathcal{O}\left(\frac{1}{\left\|\mathbf{s}_{k+1}\right\|^2}\right), \tag{5}$$

then the sub-sampled mini-batch of Hessians $\mathbf{H}_k$ satisfies eq. (4) and the sub-sampled mini-batch of gradients $\mathbf{g}_k$ satisfies

$$\left\|\mathbf{g}_k - \nabla F(\mathbf{x}_k)\right\| \leqslant C_1 \left\|\mathbf{s}_{k+1}\right\|^2, \tag{6}$$

where $C_1 \geqslant 0$ is a certain numerical constant, which further guarantee the same convergence rate for SCR as that the original exact CR.

Three important issues here motivate our design of a new sub-sampling CR algorithm.

- It can be seen from eq. (5) that as the algorithm converges, i.e., $\mathbf{s}_{k+1} \to \mathbf{0}$, the required sample size of SCR in Kohler and Lucchi (2017) grows polynomially fast, resulting significant increase in computational complexity. Thus, an important open issue here is to design an improved sub-sampling CR algorithm that reduces the sample complexity (and correspondingly computational complexity) particularly when the algorithm approaches to convergence.
- Another reason for the above pessimistic bound is because that Kohler and Lucchi (2017) analyzed the sample complexity for sampling *with* replacement, whereas in practice sampling *without* replacement can potentially have much lower sample complexity. As a clear evidence, the sample complexity for sampling *with* replacement to achieve a certain accuracy can be *unbounded*, whereas this for sampling *without* replacement can only be as large as the total sample size. Thus, the second open issue is to develop bounds for sampling *without* replacement in order to provide more precise guidance for sub-sampled CR methods.
- We also observe that eqs. (4) and (6) involve $\|\mathbf{s}_{k+1}\|$ (and hence $\mathbf{x}_{k+1}$), which is not available at iteration k. Kohler and Lucchi (2017) used s_k to replace s_{k+1} in experiments but not theory. A more recent study Wang et al. (2019) theoretically justified such a replacement with the convergence analysis, but not for stochastic sub-sampling scheme, for which the convergence analysis requires considerable efforts.

In this paper, we address the aforementioned open issues, and our contributions are summarized as follows.

Our Contributions

We propose a stochastic variance reduced cubic-regularized (SVRC) Newton's algorithm, which combines the variance reduced technique with concentration inequality under sub-sampling scheme. We show that the computation of the full Hessian and gradient can facilitate many steps of efficient inner-loop iteration as well as accurate approximation of Hessian and gradient under high probability perspective. SVRC can be associated with two sampling schemes, respectively with and without replacement.

We establish the convergence guarantee of SVRC *with high probability* under the implementable inexact condition similar with $\left\|\mathbf{H}_k - \nabla^2 F(\mathbf{x}_k)\right\| \leqslant C\left\|\mathbf{s}_k\right\|$. We show that the convergence of SVRC is at the same rate

[1] We note that SVRC(ZSG) does not need the objective function and its gradient to be Lipschitz but we adopt such assumptions.

Algorithms		Total SHO	Total SO
CR	(Nesterov and Polyak, 2006)	$\mathcal{O}(N\epsilon^{-3/2})$	$\mathcal{O}(\epsilon^{-3/2})$
SCR	(Kohler and Lucchi, 2017)	$\mathcal{O}(\epsilon^{-5/2})$	$\mathcal{O}(\epsilon^{-3/2})$
Inexact CR	(Xu et al., 2017)	$\mathcal{O}(\epsilon^{-5/2})$	$\mathcal{O}(\epsilon^{-3/2})$
SVRC(ZXG)	(Zhou et al., 2018)	$\mathcal{O}(N^{4/5}\epsilon^{-3/2})$	$\mathcal{O}(\epsilon^{-3/2})$
SVRC	(This Work)	$\tilde{\mathcal{O}}(N^{2/3}\epsilon^{-3/2})$	$\mathcal{O}(\epsilon^{-3/2})$

Table 1: Comparison of total Hessian sample complexity

($O(\epsilon^{-3/2})$) as the original CR (Nesterov and Polyak, 2006) or the other type of inexact-CR in Cartis et al. (2011, 011b); Kohler and Lucchi (2017).

We then develop the bounds on the *total* Hessian sample complexity of SVRC. We show that SVRC achieves $\tilde{\mathcal{O}}(N^{2/3}\epsilon^{-3/2})$ Hessian sample complexity (where we use $\tilde{\mathcal{O}}$ to hide the dependence on log factors), which outperforms CR (Nesterov and Polyak, 2006) by an order of $\tilde{\mathcal{O}}(N^{1/3})$ and outperform SCR (Kohler and Lucchi, 2017) in the regime of high accuracy requirement. Furthermore, our proposed SVRC order-wise outperforms the algorithm SVRC(ZSG) (Zhou et al., 2018) by an order of $\tilde{\mathcal{O}}(N^{2/15})$, which is also a variance reduced cubic regularized method *concurrently proposed*. A detailed comparison among these algorithms are summarized in Table 1.

We further provide an analysis for the case under sampling without replacement by developing a new concentration bound for sampling without replacement for random *matrices* by generalizing that for *scalar* random variables in Bardenet and Maillard (2015). Our result shows that sample replacement has lower sample complexity than that of with replacement in each iteration.

Related Works

Escaping saddle points: Various algorithms have been developed to escape strict saddle points and converge to local minimum for nonconvex optimization. The first-order such algorithms include the gradient descent algorithm with random initialization (Lee et al., 2016) and with injection of random noise (Rong et al., 2015; Chi et al., 2017). Various second-order algorithms were also proposed. In particular, Xu et al. (2017); Liu and Yang (2017); Carmon et al. (2016) proposed algorithms that exploit the negative curvature of Hessian to escape saddle points. The CR method as we describe below is another type of second-order algorithm that has been shown to escape strict saddle points.

CR type of algorithms: The CR method was shown in Nesterov and Polyak (2006) that converges to a point that satisfies the first- and second-order optimality condition for nonconvex optimization. Its accelerated version was proposed in Nesterov (2008) and the convergence rate was characterized for convex optimization. Several methods have been proposed to solve the cubic sub-problem in CR more efficiently. Cartis et al. (2011) proposed to approximately solve the cubic sub-problem in Krylov space. Agarwal et al. (2017) proposed an alternative fast way to solve the sub-problem. Carmon and Duchi (2016) proposed a method based on gradient descent. Zhou et al. (2018) studied asymptotic convergence rate of CR under the nonconvex KŁ condition, and Wang et al. (2018) established convergence guarantee for CR with momentum in nonconvex optimization.

Inexact CR algorithms: Various inexact approaches were proposed to approximate Hessian and gradient in order to reduce the computational complexity for CR type of algorithms. In particular, Ghadimi et al. (2017) studied the inexact CR and accelerated CR for convex optimization, where the inexactness is fixed throughout the iterations. Tripuraneni et al. (2017) studied a similar inexact CR for nonconvex optimization. Alternatively, Cartis et al. (2011, 011b) studied the inexact CR for nonconvex optimization, where the inexact condition is adaptive during the iterations. Wang et al. (2019) established the convergence result of CR under a more reasonable inexact condition. Jiang et al. (2017) studied the adaptive inexact accelerated CR for convex optimization. In practice, sub-sampling is a very common approach to implement inexact algorithms. Kohler and Lucchi (2017) proposed a sub-sampling scheme that adaptively changes the sample complexity to guarantee the inexactness condition in Cartis et al. (2011, 011b). Xu et al. (2017) proposed uniform and nonuniform sub-sampling algorithms with fixed inexactness condition for nonconvex optimization.

Stochastic variance reduced algorithms: Stochastic variance reduced algorithms have been applied to various first-order algorithms (known as SVRG algorithms), and the convergence rate has been studied for convex functions in, e.g., Johnson and Zhang (2013); Xiao and Zhang (2014) and for nonconvex functions in, e.g., Reddi et al. (2016); Li et al. (2017); Fang et al. (2018); Wang et al. (2018). Zhou et al. (2018) proposed a variance reduction version of CR. In this paper, we proposed another type of stochastic variance reduction to the second-order CR method to improve the state-of-art sample complexity result of approximating Hessian and gradient in probability perspective, and analyzed it in with and without replacement schemes.

Sampling without replacement: The sampling without replacement scheme for first-order methods has been studied by various papers. Recht and Re (2012) and Shamir (2016) studied stochastic gradient descent under sampling without replacement for least square problems. Gürbüzbalaban et al. (2015) provided convergence rate of the random reshuffling method. As for the sampling without replacement bounds, Hoeffding (1963) showed that the bound for sampling with replacement also holds for sampling without replacement. Friedlander and Schmidt (2012) provided deterministic bounds for without replacement sampling schemes for gradient approximations under certain assumptions. Bardenet and Maillard (2015) provided tight concentration bounds for sampling without replacement for scalar random variables, while bounds for random matrices remain unclear. We fill this gap, and provide a tight bound for random matrices under sampling without replacement in this paper.

2 Stochastic Variance Reduction Scheme for Cubic Regularization

In this paper, we are interested in solving the finite-sum problem given in eq. (1), which is rewritten below.

$$\min_{\mathbf{x}\in\mathbb{R}^d} F(\mathbf{x}) \triangleq \frac{1}{N}\sum_{i=1}^{N} f_i(\mathbf{x}), \tag{7}$$

where the component functions $f_i, i = 1, \ldots, N$ correspond to the loss of the i-th data samples, respectively, and is nonconvex. More specifically, we adopt the following standard assumptions on the objective function in eq. (7) throughout the paper

Assumption 1. *The objective function in eq. (7) satisfies*

1. *Function F is bounded below, i.e., $\inf_{\mathbf{x}\in\mathbb{R}^d} F(\mathbf{x}) > -\infty$;*
2. *For all component functions $f_i, i = 1, \ldots, N$, the function value f_i, the gradient ∇f_i, and the Hessian $\nabla^2 f_i$ are L_0, L_1 and L_2-Lipschitz, respectively.*

Classical first-order stochastic optimization methods such as stochastic gradient descent has a low sample complexity per-iteration (Nemirovski et al., 2009). However, due to the variance of the stochastic gradients, the convergence rate is slow even with the incorporation of momentum (Lan, 2012; Ghadimi and Lan, 2016). A popular approach to maintain the sample complexity yet achieve a faster convergence rate that is comparable to that of the full batch first-order methods is the stochastic variance reduction scheme (Johnson and Zhang, 2013; Xiao and Zhang, 2014).

Motivated by the success of the variance reduction scheme in improving the sample complexity of first-order methods, we propose a *stochastic variance reduced cubic*-regularized Newton's method, and refer to it as SVRC. The detailed steps of SVRC are presented in Algorithm 1. To briefly elaborate the notation in Algorithm 1, we sequentially index the iterate variable $\mathbf{x}$ across all inner loops by k for $k = 0, 1, \ldots$, so that for each $\mathbf{x}_k$, the initial variable of its inner loop is indexed as $\mathbf{x}_{\lfloor k/m \rfloor \cdot m}$ (where m is the number of iterations in each inner loop). For notational simplicity, we denote such an initial variable of each inner loop as $\tilde{\mathbf{x}}$ and denote its corresponding full gradient and Hessian as $\tilde{\mathbf{g}}$ and $\widetilde{\mathbf{H}}$, whenever there is no confusion.

Algorithm 1 SVRC

Input: $\mathbf{x}_0 \in \mathbb{R}^d$, and $\epsilon_1, m, M \in \mathbb{R}^+$.
while k **do**
  if $k \bmod m = 0$ **then**
    Set $\mathbf{g}_k = \nabla F(\mathbf{x}_k)$, $\mathbf{H}_k = \nabla^2 F(\mathbf{x}_k)$, $\widetilde{\mathbf{g}} = \mathbf{g}_k, \tilde{\mathbf{x}} = \mathbf{x}_k$ and $\widetilde{\mathbf{H}} = \mathbf{H}_k$.
  else
    Sample index sets $\xi_g(k)$ and $\xi_H(k)$ from $\{1, ..., n\}$ uniformly at random.
    Compute

$$\mathbf{g}_k = \tfrac{1}{|\xi_g(k)|}\left[\textstyle\sum_{i\in\xi_g(k)} \left(\nabla f_i(\mathbf{x}_k) - \nabla f_i(\tilde{\mathbf{x}})\right)\right] + \tilde{\mathbf{g}},$$
$$\mathbf{H}_k = \tfrac{1}{|\xi_H(k)|}\left[\textstyle\sum_{i\in\xi_H(k)} (\nabla^2 f_i(\mathbf{x}_k) - \nabla^2 f_i(\tilde{\mathbf{x}}))\right] + \widetilde{\mathbf{H}}.$$

  end if
  $\mathbf{s}_{k+1} = \mathrm{argmin}_{\mathbf{s}\in\mathbb{R}^d}\, \mathbf{g}_k^\top \mathbf{s} + \frac{1}{2}\mathbf{s}^\top \mathbf{H}_k \mathbf{s} + \frac{M}{6}\|\mathbf{s}\|^3$.
  $\mathbf{x}_{k+1} = \mathbf{x}_k + \mathbf{s}_{k+1}$.
  if $\max\{\|s_{k+1}\|, \|s_k\|\} \leqslant \epsilon_1$ **then**
    return x_{k+1}
  end if
end while

To elaborate the algorithm, SVRC calculates a full gradient $\widetilde{\mathbf{g}}$ and a full Hessian $\widetilde{\mathbf{H}}$ in every outer loop (i.e., for every m iterations), which are further used to construct the stochastic variance reduced gradients $\mathbf{g}_k$ and Hessians $\mathbf{H}_k$ in the inner loops. Note that the index sets $\xi_g(k), \xi_H(k)$ for the sampled gradients and Hessians are generated by a random sampling scheme. More specifically, we consider the following two types of sampling schemes in this paper.

Sampling with replacement: For $k = 0, 1, \dots$, each element of the index sets $\xi_g(k)$ and $\xi_H(k)$ is sampled uniformly at random from $\{1, \dots, N\}$.

Sampling without replacement: For $k = 0, 1, \dots$, the index sets $\xi_g(k)$ and $\xi_H(k)$ are sampled uniformly at random from all subsets of $\{1, \dots, N\}$ with cardinality $|\xi_g(k)|$ and $|\xi_H(k)|$, respectively.

To elaborate, the sampling with replacement scheme may sample the same index multiple times within each mini-batch, whereas the sampling without replacement scheme samples each index at most once within each mini-batch. Therefore, the sampling without replacement scheme has a smaller variance compared to that of the sampling with replacement scheme. Consequently, these sampling schemes lead to inexact gradients and inexact Hessians with different guarantees to meet the inexactness criterion.

3 Sample Complexity of SVRC

In this section, we study the sample complexity of SVRC for achieving a second-order stationary point via three technical steps, each corresponding to one subsection below.

3.1 Iteration Complexity under Modified Inexact Condition

In order to analyze the sample complexity of SVRC for achieving a second-order stationary point, it turns out that the inexact condition (Wang et al., 2019) on the estimated gradients and Hessians is not sufficient. Thus, we propose a modified inexact condition below, and then analyze the convergence to a second-order stationary point if SVRC satisfies such a condition.

Assumption 2. *The approximate Hessian $\mathbf{H}_k$ and approximate gradient $\mathbf{g}_k$ satisfy, for all $k = 0, \cdots$,*

$$\left\|\mathbf{H}_k - \nabla^2 F(\mathbf{x}_k)\right\| \leqslant \alpha \max\left\{\|\mathbf{s}_k\|, \epsilon_1\right\} \tag{8}$$

$$\|\mathbf{g}_k - \nabla F(\mathbf{x}_k)\| \leqslant \beta \max\left\{\|\mathbf{s}_k\|^2, \epsilon_1^2\right\} \tag{9}$$

where ϵ_1, α and β are universal positive constants.

The inexact conditions in eqs. (8) and (9) introduce a slack variable ϵ_1 to avoid full batch sampling when $\|s_k\|$ is very close to zero upon convergence. It turns out introduction of such a variable is essential for characterizing the total sample complexity of our proposed variance reduction scheme in Algorithm 1. Furthermore, since eqs. (8) and (9) are different from that in (Wang et al., 2019), and hence require the convergence analysis if SVRC satisfies such conditions. The following theorem presents the iteration complexity analysis under the modified conditions. The technical proof in fact requires considerable extra effort than that in Wang et al. (2019).

Theorem 1. *Suppose Assumption 1 holds, and SVRC satisfies Assumption 2. Let*

$$\tau \triangleq \min\left\{ \left(\frac{L+M}{2} + 2\beta + 2\alpha\right)^{-\frac{1}{2}}, \left(\frac{M+2L}{2} + 2\alpha\right)^{-1} \right\},$$

set

$$\epsilon_1 = \tau\sqrt{\epsilon}, \tag{10}$$

and properly choose M, α and $\beta \in \mathbb{R}$ such that

$$\gamma \triangleq \left(\frac{3M - 2L_2}{24} - \frac{5}{2}\beta - \frac{5}{4}\alpha\right) > 0. \tag{11}$$

Then, the SVRC algorithm outputs an ϵ-approximate second-order stationary point, i.e.,

$$\|\nabla f(\mathbf{x}_{k+1})\| \leqslant \epsilon \quad \text{and} \quad \nabla^2 f(\mathbf{x}_{k+1}) \succcurlyeq -\sqrt{\epsilon}\mathbf{I} \tag{12}$$

within at most $k = O\left(\epsilon^{-3/2}\right)$ number of iterations. Moreover, the following inequality holds

$$\sum_{i=1}^{k+1} \|\mathbf{s}_i\|^3 \leqslant C, \tag{13}$$

where $C \triangleq \left(f(\mathbf{x}_0) - f^* + (2\beta + \alpha + 2\gamma)\,\epsilon_1^3\right)/\gamma$.

As stated in Theorem 1, SVRC outputs an ϵ-approximate second-order stationary point with $k = O\left(\epsilon^{-3/2}\right)$. Such an iteration complexity matches the state-of-art result and is the best result that one can expect on nonconvex optimization.

3.2 Per-iteration Sample Complexity

In this subsection, we bound the per-iteration sample complexity in order for SVRC (under sampling with replacement) to satisfy the inexact conditions in eqs. (8) and (9). We apply Bernstein's inequality and obtain the following theorem.

Theorem 2. *Let Assumption 1 hold. Consider SVRC under the sampling with replacement scheme. Then, the sub-sampled mini-batch of gradients $\mathbf{g}_k, k = 0, 1, \dots$*

satisfies Assumption 2 with probability at least $1-\zeta$ *provided that*

$$|\xi_g(k)| \geqslant \bigg(\frac{8L_1^2}{\beta^2 \max\{\|\mathbf{s}_k\|^4, \epsilon_1^4\}} \|\mathbf{x}_k - \tilde{\mathbf{x}}\|^2 + \frac{4L_1}{3\beta \max\{\|\mathbf{s}_k\|^2, \epsilon_1^2\}} \|\mathbf{x}_k - \tilde{\mathbf{x}}\|\bigg) \log\left(\frac{2(d+1)}{\zeta}\right), \tag{14}$$

Furthermore, the sub-sampled mini-batch of Hessians $\mathbf{H}_k, k = 0, 1, \ldots$ *of SVRC satisfies Assumption 2 with probability at least* $1-\zeta$ *provided that*

$$|\xi_H(k)| \geqslant \bigg(\frac{8L_2^2}{\alpha^2 \max\{\|\mathbf{s}_k\|^2, \epsilon_1^2\}} \|\mathbf{x}_k - \tilde{\mathbf{x}}\|^2 + \frac{4L_2}{3\alpha \max\{\|\mathbf{s}_k\|, \epsilon_1\}} \|\mathbf{x}_k - \tilde{\mathbf{x}}\|\bigg) \log\left(\frac{4d}{\zeta}\right). \tag{15}$$

We next compare the per-iteration Hessian sample complexity of SVRC under the sampling with replacement scheme (eq. (15)) with that of SCR under the same sampling scheme developed in Kohler and Lucchi (2017), which is rewritten below

$$|\xi_H(k)| \geqslant \mathcal{O}\left(\tfrac{1}{\|\mathbf{s}_{k+1}\|^2}\right). \tag{16}$$

To compare, our Theorem 2 requires a Hessian sample complexity of roughly the order

$$|\xi_H(k)| \geqslant \mathcal{O}\left(\tfrac{\|\mathbf{x}_k - \tilde{\mathbf{x}}\|^2}{\|\mathbf{s}_k\|^2}\right). \tag{17}$$

It can be seen that the sample complexity bounds for SVRC in eq. (17) have an additional term $\|\mathbf{x}_k - \tilde{\mathbf{x}}\|^2$ in the numerators comparing to their corresponding bound for SCR in eq. (16). Intuitively, $\|\mathbf{x}_k - \tilde{\mathbf{x}}\| \to 0$ as the algorithm converges, and thus our variance reduction scheme requires a lower sample complexity than the stochastic sampling in SCR.

3.3 Total Sample Complexity of SVRC

Theorem 2 provides the sample complexity per iteration (each iteration in SVRC inner loop). We next provide our result on the sample complexity over the running process of SVRC, which is a key factor that impacts the computational complexity of SVRC.

Theorem 3. *Let Assumptions 1 hold. For a given* ϵ *and* δ, *set* $m = N^{1/3}$, *then SVRC under the sampling with replacement scheme outputs an point* $\mathbf{x}_{k+1}$ *such that satisfies* $\|\nabla F(\mathbf{x}_{k+1})\| \leqslant \epsilon$ *and* $\nabla^2 F(\mathbf{x}_{k+1}) \succcurlyeq -\epsilon \boldsymbol{I}$ *with probability at least* $1-\delta$, *and the total Hessian sample complexity of SVRC is bounded by*

$$\sum_{i=1}^{K} |\xi_H(i)| \leqslant \frac{CN^{2/3}}{\epsilon^{3/2}} \log\left(\frac{8d}{\epsilon\delta}\right).$$

We next compare the total Hessian sample complexity of SVRC with that of other CR-type algorithms, which are given below.

$$\text{SVRC:} \quad \sum_{i=1}^{K} |\xi_H(i)| = \tilde{\mathcal{O}}\left(\frac{N^{2/3}}{\epsilon^{3/2}}\right), \tag{18}$$

$$\text{SVRC (ZXG):} \quad \sum_{i=1}^{K} |\xi_H(i)| = \mathcal{O}\left(\frac{N^{4/5}}{\epsilon^{3/2}}\right), \tag{19}$$

$$\text{CR:} \quad \sum_{i=1}^{K} |\xi_H(i)| \leqslant \mathcal{O}\left(\frac{N}{\epsilon^{3/2}}\right), \tag{20}$$

$$\text{SCR:} \quad \sum_{i=1}^{K} |\xi_H(i)| \leqslant \mathcal{O}\left(\frac{1}{\epsilon^{5/2}}\right). \tag{21}$$

Comparing eqs. (18) to (20). Clearly, our SVRC has lower total sample complexity than CR and SVRC(ZXG) by an order of $\tilde{\mathcal{O}}(N^{1/3})$ and $\tilde{\mathcal{O}}(N^{2/15})$, respectively. Therefore, our stochastic variance reduction scheme is sample efficient when applied to CR type of methods. Also, comparing the sample complexity of the two subsampled algorithms in eqs. (18) and (21), we observe that SVRC enjoys a lower-order complexity bound than SCR if $\epsilon = o(N^{-2/3})$, and hence performs better in the high accuracy regime.

4 SVRC under Sampling without Replacement Scheme

In this section, we explore the sample complexity of SVRC under the sampling without replacement scheme, which is commonly used in practice.

To this end, we first develop some technical concentration inequalities in the next subsection.

4.1 Concentration Inequality under Sampling without Replacement

The statistics of sampling without replacement is very different and more stable than that of sampling with replacement. However, theoretical analysis of sampling *without* replacement turns out to be very difficult. A common approach is to apply the concentration bound for sampling with replacement, which also holds for sampling without replacement (Tropp, 2012). However, such analysis can be too loose to capture the essence of the scheme of sampling without replacement. For example, the sample complexity for sampling *with* replacement to achieve a certain accuracy can be *unbounded*, whereas sampling *without* replacement can at most sample the total sample size.

Thus, in order to develop a tight sample complexity bound for SVRC under sampling without replacement, we first leverage a recently developed Hoeffding-type of concentration inequality for sampling *without* replacement (Bardenet and Maillard, 2015). There, the result is applicable only for scalar random variables, whereas our analysis here needs to deal with sub-sampled gradients and Hessians, which are vectors and matrices. This motivates us to first establish the matrix version of the Hoeffding-Serfling inequality. Such a concentration bound can be of independent interest in various other domains. The proof turns out to be very involved and is provided in the supplementary materials.

Theorem 4. *Let $\mathcal{X} := \{\mathbf{A}_1, \cdots, \mathbf{A}_N\}$ be a collection of real-valued matrices in $\mathbb{R}^{d_1 \times d_2}$ with bounded spectral norm, i.e., $\|\mathbf{A}_i\| \leqslant \sigma$ for all $i = 1, \ldots, N$ and some $\sigma > 0$. Let $\mathbf{X}_1, \cdots, \mathbf{X}_n$ be $n < N$ samples from $\mathcal{X}$ under the sampling without replacement. Denote $\mu := \frac{1}{N}\sum_{i=1}^N \mathbf{A}_i$. Then, for any $\epsilon > 0$, the following bound holds.*

$$P\left(\left\|\frac{1}{n}\sum_{i=1}^n \mathbf{X}_i - \mu\right\| \geqslant \epsilon\right) \leqslant 2(d_1 + d_2)\exp\left(-\frac{n\epsilon^2}{8\sigma^2(1+1/n)(1-n/N)}\right).$$

To further understand the above theorem, consider symmetric random matrix $\mathbf{X}_i \in \mathbb{R}^{d \times d}$. Suppose we want $\left\|\frac{1}{n}\sum_{i=1}^n \mathbf{X}_i - \mu\right\| \leqslant \epsilon$ to hold with probability $1-\zeta$. Then the above theorem requires the sample size to satisfy

$$n_w \geqslant \left(\frac{1}{N} + \frac{\epsilon^2}{16\sigma^2 \log(4d)/\zeta)}\right)^{-1}. \tag{22}$$

We consider two regimes to understand the bound in eq. (22). (a) Low accuracy regime: Suppose ϵ is large enough so that the second term in eq. (22) dominates. In this case, we roughly have $n_w \geqslant \frac{16\sigma^2 \log(4d/\zeta)}{\epsilon^2}$, which has the same order as the suggested sample size by the matrix version of the Hoeffding inequality for sampling *with* replacement given below

$$n_b \geqslant \frac{8\sigma^2 \log(2d/\zeta)}{\epsilon^2}. \tag{23}$$

Thus, the sample size is approximately the same for sampling with and without replacement to achieve a low accuracy concentration. (b) High accurary regime: Suppose ϵ is small enough so that the first term in eq. (22) dominates. Hence, eq. (22) roughly reduces to $n_w \geqslant N$, whereas the matrix version of the Hoeffding bound in eq. (23) for sampling with replacement requires infinite samples as $\epsilon \to 0$. Thus, the sample size is highly different for sampling with and without replacement to achieve a high accuracy concentration.

4.2 Per-iteration Sample Complexity

We apply Theorem 4 to analyze the sample complexity of SVRC under sampling without replacement. Our next theorem characterizes the sample size needed for SVRC in order to satisfy the inexact condition in Assumption 2.

Theorem 5. *Let Assumption 1 hold. Consider SVRC under sampling without replacement. The sub-sampled mini-batches of gradients $\mathbf{g}_k, k = 0, 1, \ldots$ satisfy eq. (6) with probability at least $1-\zeta$ provided that*

$$|\xi_g(k)| \geqslant \left(\frac{1}{N} + \frac{\beta^2 \max\{\|\mathbf{s}_k\|^4, \epsilon_1^4\}}{64L_1^2 \|\mathbf{x}_k - \tilde{\mathbf{x}}\|^2 \log(2(d+1)/\zeta)}\right)^{-1}, \tag{24}$$

Furthermore, the sub-sampled mini-batches of Hessians $\mathbf{H}_k, k = 0, 1, \ldots$ satisfy eq. (4) with probability at least $1-\zeta$ provided that

$$|\xi_H(k)| \geqslant \left(\frac{1}{N} + \frac{\alpha^2 \max\{\|\mathbf{s}_k\|^2, \epsilon_1^2\}}{64L_2^2 \|\mathbf{x}_k - \tilde{\mathbf{x}}\|^2 \log(4d/\zeta)}\right)^{-1}. \tag{25}$$

In order to further understand the sample complexity in Theorem 5 and what improvement that SVRC makes in terms of sample complexity compared to the SCR algorithm in Kohler and Lucchi (2017), we next characterize the corresponding sample complexity for SCR under sampling without replacement below. (We note that the sample complexity for SCR under sampling with replacement was provided in Kohler and Lucchi (2017).)

Proposition 6. *Let Assumptions 1 hold. Consider the SCR algorithm in Kohler and Lucchi (2017) under sampling without replacement. The sub-sampled mini-batch of gradients $\mathbf{g}_k, k = 0, 1, \ldots$ satisfies eq. (6) with probability at least $1-\zeta$ provided that for all k*

$$|\xi_g(k)| \geqslant \left(\frac{1}{N} + \frac{C_1^2 \|\mathbf{x}_{k+1} - \mathbf{x}_k\|^4}{64L_0^2 \log(2(d+1)/\zeta)}\right)^{-1}. \tag{26}$$

Furthermore, the sub-sampled mini-batch of Hessians $\mathbf{H}_k, k = 0, 1, \ldots$ satisfies eq. (4) with probability at least $1-\zeta$ provided that for all k

$$|\xi_H(k)| \geqslant \left(\frac{1}{N} + \frac{C_2^2 \|\mathbf{x}_{k+1} - \mathbf{x}_k\|^2}{64L_1^2 \log(4d/\zeta)}\right)^{-1}. \tag{27}$$

To compare the sample complexity for SVRC in Theorem 5 and SCR in Proposition 6, we take the sample complexity for mini-batch of gradients as an example. Comparing eq. (24) and eq. (26), the second term in the denominator in eq. (24) is additionally divided by $\|\mathbf{x}_k - \tilde{\mathbf{x}}\|^2$, which converges to zero as the algorithms converge. Thus, $\|\mathbf{x}_{k+1} - \mathbf{x}_k\|^2$ in eq. (26) converges to zero much faster than $\frac{\|\mathbf{x}_{k+1}-\mathbf{x}_k\|^4}{\|\mathbf{x}_k-\tilde{\mathbf{x}}\|^2}$ in eq. (24), so that the term $1/N$ dominates the denominator and results in the sample size close to the number of total samples much earlier in the iteration of SCR than SVRC.

We also note that Proposition 6 shows that as SCR approaches the convergence, the sample size goes to the total number of samples with technical rigor, whereas such a fact was only intuitively discussed in Kohler and Lucchi (2017).

4.3 Total Sample Complexity

We next characterize the total Hessian sample complexity of SVRC under sampling without replacement.

Theorem 7. *Let Assumptions 1 hold. For a given ϵ and δ, set $m = N^{1/3}$, then SVRC under sampling without replacement outputs an point $\mathbf{x}_{k+1}$ such that satisfies $\|\nabla F(\mathbf{x}_{k+1})\| \leqslant \epsilon$ and $\nabla^2 F(\mathbf{x}_{k+1}) \succcurlyeq -\epsilon \boldsymbol{I}$ with probability at least $1-\delta$. Then the total sample complexity for Hessian used in SVRC is bounded by*

$$\sum_{i=0}^{k} |\xi_H(k)| \leqslant \frac{CN^{2/3}}{\epsilon^{3/2}} \log\left(\frac{8d}{\epsilon\delta}\right). \tag{28}$$

In this theorem, we show that total sample complexity of SVRC under sampling without replacement is at least as good as SVRC under sampling with replacement. And the comparison of this bound with other bound follows similarly as we discuss in Section 3.3.

5 Discussion

Storage Issue: The proposed algorithm involves the storage of a Hessian, which requires $\mathcal{O}(d^2)$ space for storage. In this perspective, the proposed algorithm can be directly applied for solving small or medium scale machine learning problems. As for large scale problems, using PCA to store the main component of Hessian can be a possible solution.

With and Without replacement: We show that the total sample complexity of SVRC under sampling without replacement is at least as good as SVRC under sampling with replacement. Actually, if we compare the per iteration complexity of the two, i.e., we compare Theorem 5 with Theorem 2, the without replacement scheme has a better complexity than that with replacement in each iteration since there is a $1/N$ term in the denominator on the bound for the scheme without replacement. This does suggest the same total sample complexity for the two schemes is likely due to the technicality issue.

6 Conclusion

In this paper, we proposed a stochastic variance-reduced cubic regularization method. We characterized the per iteration sample complexity for Hessian and gradient that guarantees convergence of SVRC to a second-order optimality condition, under both sampling with and without replacement. We also developed the total sample size for Hessian. Our theoretic results imply that SVRC outperforms the state-of-art result by an factor of $O(N^{2/15})$. Moreover, Our study demonstrates that variance reduction can bring substantial advantage in sample size as well as computational complexity for second-order algorithms, along which direction we plan to explore further in the future.

Acknowledgement

The work of Z. Wang, Y. Zhou and Y. Liang was supported in part by the U.S. National Science Foundation under the grant CCF-1761506, and the work of G. Lan was supported in part by Army Research Office under the grant W911NF-18-1-0223 and U.S. National Science Foundation under the grant CMMI-1254446.

References

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

Baldi, P. and Hornik, K. (1989). Neural networks and principal component analysis: Learning from examples without local minima. *Neural Networks*, 2(1):53 – 58.

Bardenet, R. and Maillard, O. (2015). Concentration inequalities for sampling without replacement. *Bernoulli*, 21(3):1361–1385.

Bhatia, R. (2007). *Positive Definite Matrices*. Princeton University Press.

Bhojanapalli, S., Neyshabur, B., and Srebro, N. (2016). Global optimality of local search for low rank matrix recovery. In *Proc. Advances in Neural Information Processing Systems (NIPS)*.

Carmon, Y. and Duchi, J. C. (2016). Gradient descent efficiently finds the cubic-regularized non-convex Newton step. *ArXiv: 1612.00547*.

Carmon, Y., Duchi, J. C., Hinder, O., and Sidford, A. (2016). Accelerated methods for non-convex optimization. *ArXiv:1611.00756*.

Cartis, C., Gould, N. I. M., and Toint, P. (2011b). Adaptive cubic regularization methods for unconstrained optimization. part ii: worst-case function-and derivative-evaluation complexity. *Mathematical Programming*, 130(2):295–319.

Cartis, C., Gould, N. I. M., and Toint, P. L. (2011). Adaptive cubic regularization methods for unconstrained optimization. part i : motivation, convergence and numerical results. *Mathematical Programming*.

Cartis, C., Gould, N. I. M., and Toint, P. L. (2012a). An adaptive cubic regularization algorithm for nonconvex optimization with convex constraints and its function-evaluation complexity. *IMA Journal of Numerical Analysis*, 32(4):1662 – 1695.

Cartis, C., Gould, N. I. M., and Toint, P. L. (2012b). Complexity bounds for second-order optimality in unconstrained optimization. *Journal of Complexity*, 28(1):93 – 108.

Chi, J., Rong, G., Praneeth, N., K., S. M., and Michael, I. J. (2017). How to escape saddle points efficiently. In *Proc. 34th International Conference on Machine Learning (ICML)*, volume 70, pages 1724–1732.

Fang, C., Li, C., Lin, Z., and Zhang, T. (2018). SPIDER: Near-optimal non-convex optimization via stochastic path-integrated differential estimator. In *Proc. Advances in Neural Information Processing Systems (NIPS)*, pages 687–697.

Friedlander, M. and Schmidt, M. (2012). Hybrid deterministic-stochastic methods for data fitting. *SIAM Journal on Scientific Computing*, 34(3):A1380–A1405.

Ge, R., Lee, J., and Ma, T. (2016). Matrix completion has no spurious local minimum. In *Proc. Advances in Neural Information Processing Systems (NIPS)*, pages 2973–2981.

Ghadimi, S. and Lan, G. (2016). Accelerated gradient methods for nonconvex nonlinear and stochastic programming. *Mathematical Programming*, 156:59–99.

Ghadimi, S., Liu, H., and Zhang, T. (2017). Second-order methods with cubic regularization under inexact information. *ArXiv: 1710.05782*.

Gürbüzbalaban, M., Ozdaglar, A., and Parrilo, P. (2015). Why Random Reshuffling Beats Stochastic Gradient Descent. *ArXiv:1510.08560*.

Hoeffding, W. (1963). Probability inequalities for sums of bounded random variables. *Journal of the American Statistical Association*, 58(301):13–30.

Jiang, B., Lin, T., and Zhang, S. (2017). A unified scheme to accelerate adaptive cubic regularization and gradient methods for convex optimization. *ArXiv:1710.04788*.

Johnson, R. and Zhang, T. (2013). Accelerating stochastic gradient descent using predictive variance reduction. In *Proc. 26th Advances in Neural Information Processing Systems (NIPS)*, pages 315–323.

Kohler, J. M. and Lucchi, A. (2017). Sub-sampled cubic regularization for non-convex optimization. In *Proc. 34th International Conference on Machine Learning (ICML)*, volume 70, pages 1895–1904.

Lan, G. (2012). An optimal method for stochastic composite optimization. *Mathematical Programming*, 133(1):365–397.

Lee, J. D., Simchowitz, M., Jordan, M. I., and Recht, B. (2016). Gradient descent only converges to minimizers. In *Proc. 29th Annual Conference on Learning Theory (COLT)*, volume 49, pages 1246–1257.

Li, Q., Zhou, Y., Liang, Y., and Varshney, P. K. (2017). Convergence analysis of proximal gradient with momentum for nonconvex optimization. In *Proc. 34th International Conference on Machine Learning (ICML*, volume 70, pages 2111–2119.

Liu, M. and Yang, T. (2017). On noisy negative curvature descent: competing with gradient descent for faster non-convex optimization. *ArXiv: 1709.08571*.

Nemirovski, A. S., Juditsky, A., Lan, G., and Shapiro, A. (2009). Robust stochastic approximation approach to stochastic programming. *SIAM Journal on Optimization*, 19:1574–1609.

Nesterov, Y. (2008). Accelerating the cubic regularization of newton's method on convex problems. *Mathematical Programming*, 112(1):159–181.

Nesterov, Y. and Polyak, B. (2006). Cubic regularization of newton's method and its global performance. *Mathematical Programming*.

Recht, B. and Re, C. (2012). Toward a noncommutative arithmetic-geometric mean inequality: conjectures, case-studies, and consequences. *Conference on Learning Theory*.

Reddi, S., Hefny, A., Sra, S., B., P., and A., S. (2016). Stochastic variance reduction for nonconvex optimization. In *Proc. 33rd International Conference on Machine Learning (ICML)*, volume 48, pages 314–323.

Rong, G.and Furong, H., Chi, J., and Yang, Y. (2015). Escaping from saddle points — online stochastic gradient for tensor decomposition. In *Proc. 28th Conference on Learning Theory (COLT)*, volume 40, pages 797–842.

Shamir, O. (2016). Without-replacement sampling for stochastic gradient methods. In *Proc. 29th Advances in Neural Information Processing Systems (NIPS)*, pages 46–54.

Sun, J., Qu, Q., and Wright, J. (2017). A geometrical analysis of phase retrieval. *Foundations of Computational Mathematics*, pages 1–68.

Tripuraneni, N., Stern, M., Jin, C., Regier, J., and Jordan, M. I. (2017). Stochastic Cubic Regularization for Fast Nonconvex Optimization. *ArXiv: 711.02838*.

Tropp, J. A. (2012). User-friendly tail bounds for sums of random matrices. *Foundations of Computational Mathematics*, 12(4):389–434.

Wang, Z., Ji, K., Zhou, Y., Liang, Y., and Tarokh, V. (2018). SpiderBoost: A class of faster variance-reduced algorithms for nonconvex optimization. *ArXiv:1810.10690*.

Wang, Z., Zhou, Y., Liang, Y., and Lan, G. (2018). Cubic Regularization with Momentum for Nonconvex Optimization. *arXiv:1810.03763*.

Wang, Z., Zhou, Y., Liang, Y., and Lan, G. (2019). A note on inexact gradient and Hessian conditions for cubic regularized Newtons method. *Operations Research Letters*.

Xiao, L. and Zhang, T. (2014). A proximal stochastic gradient method with progressive variance reduction. *SIAM Journal on Optimization*, 24(4):2057–2075.

Xu, P., Roosta-Khorasani, F., and Mahoney, M. W. (2017). Newton-type methods for non-convex optimization under inexact hessian information. *ArXiv: 1708.07164*.

Xu, Y., Jin, R., and Yang, T. (2017). NEON+: Accelerated gradient methods for extracting negative curvature for non-convex optimization. *ArXiv: 1712.01033*.

Zhou, D., Xu, P., and Gu, Q. (2018). Stochastic Variance-Reduced Cubic Regularized Newton Method. In *Proc. 35th International Conference on Machine Learning (ICML)*.

Zhou, Y. and Liang, Y. (2018). Critical points of linear neural networks: Analytical forms and landscape properties. In *Proc. International Conference on Learning Representations (ICLR)*.

Zhou, Y., Wang, Z., and Liang, Y. (2018). Convergence of cubic regularization for nonconvex optimization under KL property. In *Proc. 32nd Advances in Neural Information Processing Systems (NIPS)*.