Cubic Regularization with Momentum for Nonconvex Optimization

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

Yi Zhou
EECS Dept.
Duke University
yi.zhou610@duke.edu

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

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

Abstract

Momentum is a popular technique to accelerate the convergence in practical training, and its impact on convergence guarantee has been well-studied for first-order algorithms. However, such a successful acceleration technique has not yet been proposed for second-order algorithms in nonconvex optimization. In this paper, we apply the momentum scheme to cubic regularized (CR) Newton's method and explore the potential for acceleration. Our numerical experiments on various nonconvex optimization problems demonstrate that the momentum scheme can substantially facilitate the convergence of cubic regularization, and perform even better than the Nesterov's acceleration scheme for CR. Theoretically, we prove that CR under momentum achieves the best possible convergence rate to a second-order stationary point for nonconvex optimization. Moreover, we study the proposed algorithm for solving problems satisfying an error bound condition and establish a local quadratic convergence rate. Then, particularly for finite-sum problems, we show that the proposed algorithm can allow computational inexactness that reduces the overall sample complexity without degrading the convergence rate.

1 INTRODUCTION

In the era of machine learning, deep models such as neural networks have achieved great success in solving a variety of challenging tasks. However, training deep models is in general a difficult task and traditional first-order algorithms can easily get stuck at sub-optimal points such as saddle points, which have been shown to bottleneck the performance of practical training (Dauphin et al., 2014). Motivated by this, there is a rising interest in designing algorithms that can escape saddle points in general nonconvex optimization, and the cubic regularization (CR) Newton's method is such a type of popular optimization algorithm.

More specifically, consider the following generic nonconvex optimization problem.

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

where $f:\mathbb{R}^d\to\mathbb{R}$ is a twice-differentiable and nonconvex function. The CR algorithm (Nesterov and Polyak, 2006) takes an initialization $\mathbf{x}_0\in\mathbb{R}^d$, a proper parameter $M>0$, and generates a sequence $\{\mathbf{x}_k\}_k$ for solving eq. (1) via the following update rule.

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

Intuitively, the main step of CR solves a cubic minimization subproblem that is formulated by the second-order Taylor expansion at the current iterate with a cubic regularizer. Such a cubic subproblem can be efficiently solved by many dedicated solvers (Cartis et al., 2011a; Carmon and Duchi, 2016; Agarwal et al., 2017) that induce a low overall computation complexity (see Section 4.3 for further elaboration). By exploiting second order information (i.e., gradient and Hessian) of the objective function, the CR algorithm has been shown to produce a solution $\mathbf{x}$ that satisfies the ϵ-second-order stationary condition, i.e.,

$$\|\nabla f(\mathbf{x})\| \leqslant \epsilon \quad \text{and} \quad \lambda_{\min}\big(\nabla^2 f(\mathbf{x})\big) \geqslant -\sqrt{\epsilon}, \tag{2}$$

where $\lambda_{\min}\big(\nabla^2 f(\mathbf{x})\big)$ denotes the minimum eigenvalue of the Hessian $\nabla^2 f(\mathbf{x})$. Unlike the first-order stationary condition (i.e., $\|\nabla f(\mathbf{x})\|\leqslant\epsilon$) which does not rule out the possibility of converging to a saddle point, the second-order stationary condition requires the corresponding Hessian to be almost positive semidefinite and hence can avoid convergence to strict saddle points (i.e., at which Hessian has negative eigenvalue). In particular, a variety of nonconvex machine learning problems such as phase retrieval (Sun et al., 2017), dictionary learning (Sun et al., 2015) and tensor decomposition (Ge et al., 2015) have been shown to have only strict saddle points. Therefore, CR is guaranteed to escape all the saddle points and converge to a local minimum in solving these problems.

While most existing studies on the CR algorithm focus on reducing the computation complexity by various sampling schemes, e.g., mini-batch sampling (Xu et al., 2017), subsampling (Kohler and Lucchi, 2017), variance-reduced sampling (Wang et al., 2019b; Zhou et al., 2018), less attention has been paid to the design of new schemes for accelerating CR. The only exception is Nesterov (2008), where an acceleration scheme was proposed for CR, but has been shown to achieve a faster convergence rate than CR only for *convex* problems. Such an accelerated scheme consists of hyperparameters that are fine-tuned in the context of convex optimization, and hence may not guarantee to produce a second-order stationary solution in *nonconvex* optimization. There does not exist any accelerated CR algorithm that has provable convergence for nonconvex optimization. Therefore, the aim of this paper is to design a momentum-based scheme for CR with provable second-order stationary convergence guarantee for nonconvex optimization as well as yielding faster convergence in practical scenarios.

1.1 OUR CONTRIBUTIONS

Our major contribution lies in proposing the first CR algorithm that incorporates momentum technique, which has provable convergence guarantee to a second-order stationary point in nonconvex optimization. We also performed a comprehensive study of this algorithm from various aspects both in theory and experiments to demonstrate the appealing attributes of the proposed algorithm. Our specific contribution are listed as follows.

- We propose a CR type algorithm with momentum acceleration (referred to as CRm), which includes a cubic regularization step, a momentum step for acceleration and a monotone step. The momentum step introduces negligible computation complexity compared to that of the cubic regularization step in original CR, but can provide substantial advantage of acceleration.
- We establish the global convergence of CRm to a second-order stationary point in nonconvex optimization. The corresponding convergence rate is as fast as that of CR in the order-level, which is the best one can expect for nonconvex optimization. Our experiments demonstrate that CRm substantially outperforms CR as well as Nesterov's accelerated CR (which does not have guaranteed performance for nonconvex optimization).
- We also show that CRm enjoys the local quadratic convergence property under a local error bound condition, which establishes the advantage of the second-order algorithms than the first-order algorithms in nonconvex optimization.
- We further show that the inexact variant of CRm significantly improves the computational complexity without losing the convergence rate. We also study the finite-sum problem, where we implement the inexact CRm via a subsampling approach, and established the total Hessian sample complexity to guarantee the convergence with high probability.

On the core of our proof technique, we rely on the delicate design of the adaptive momentum parameter in eq. (4), and the monotone step in the algorithm, which makes it possible to establish the convergence result under nonconvex optimization but with momentum acceleration. To the best of our knowledge, there is no result on accelerated CR type algorithms that have such good convergence property, or even the convergence property under nonconvex optimization.

1.2 RELATED WORKS

Escaping saddle points: A number of algorithms have been proposed to escape saddle points in order to find local minima. In general, There are three lines of research. It has been shown that with random perturbation, gradient descent algorithm (Jin et al., 2017), the stochastic gradient descent (Ge et al., 2015), the zero-th order method (Jin et al., 2018), and the accelerated gradient descent (Jin et al., 2017) can escape saddle points. The gradient descent has also been incorporated with the negative curvature descent in Carmon et al. (2016); Liu and Yang (2017); Xu et al. (2017) in order to converge to the second-order stationary points. Furthermore, the cubic regularized (CR) algorithm, which first appeared in Griewank (1981), has been shown by Nesterov and Polyak (2006) to converge to the second-order stationary points. Cartis et al. (2011a,b) then proposed an adaptive CR method with an approximate sub-problem solver. Agarwal et al. (2017) established an efficient sub-problem solver for CR by using the Hessian-vector product technique, and Carmon and Duchi (2016) showed that gradient descent can efficiently solve the sub-problem in CR. Zhou et al. (2018) studied the CR algorithm in nonconvex optimization. This paper further accelerates the CR algorithm with momentum and establishes its convergence rate to a second-order stationary point.

Algorithms with momentum for nonconvex optimization: Ghadimi and Lan (2016); Li and Lin (2015) proposed accelerated gradient descent type of algorithms for nonconvex optimization, which are guaranteed to converge as fast as gradient descent for nonconvex problems. Yao et al. (2017) proposed an efficient accelerated proximal gradient descent algorithm for nonconvex problems, which requires only one proximal step in each iteration as compared to the requirement of two proximal steps in each iteration in the algorithm proposed in Li and Lin (2015). Then Li et al. (2017) analyzed the algorithm in Yao et al. (2017) under the KL condition. While the existing studies analyzed only convergence to first-order stationary points, this paper proposes the CR algorithms with momentum that converge to a second-order stationary point.

Inexact CR algorithms: To reduce the computational complexity for the CR type of algorithms, various inexact Hessian and gradient approaches were proposed. In particular, Ghadimi et al. (2017) studied the inexact Hessian CR and accelerated CR for convex optimization, where the inexact level is fixed during iterations. Tripuraneni et al. (2017) studied an inexact CR for nonconvex optimization, which allows both the gradient and Hessian to be inexact. Alternatively, Cartis et al. (2011a,b) studied the inexact Hessian CR for nonconvex optimization, where the inexact condition is adaptive during iterations. Jiang et al. (2017) studied a unified scheme of inexact accelerated adaptive CR and gradient descent for convex optimization. Furthermore, Kohler and Lucchi (2017) proposed a subsampling CR (SCR) that adaptively changes the sample batch size to guarantee the inexactness condition in Cartis et al. (2011a,b), Wang et al. (2019a) relaxed the inexact condition in Kohler and Lucchi (2017); Cartis et al. (2011a,b), and Xu et al. (2017) proposed uniform and non-uniform sampling algorithms with fixed inexactness for nonconvex optimization. Wang et al. (2019b); Zhou et al. (2018) proposed stochastic variance reduced subsampling CR algorithms. This paper establishes the convergence rate for the inexact scenarios of the proposed CR algorithm with momentum.

Local quadratic convergence: The Newton's method and cubic regularized algorithm have been shown to converge quadratically to the global minimum under the strongly convex condition in Nesterov and Polyak (2006); Nesterov (2008), respectively. Furthermore, various Newton-type algorithms, i.e., the Levenberg-Marquardt method (Yamashita and Fukushima, 2001; Fan and Yuan, 2005), the regularized Newton method (Li et al., 2004), the regularized proximal Newton's method (Yue et al., 2016), and the CR algorithm (Yue et al., 2018), have been shown to have the local quadratic convergence under the more relaxed local error bound condition. This paper further establishes such a property for the proposed CR with momentum algorithm.

2 CRm: CUBIC REGULARIZATION WITH MOMENTUM

In this section, we propose a CR-type algorithm that adopts a momentum scheme (referred to as CRm). The algorithm steps of CRm are summarized in Algorithm 1.

At each iteration, the proposed CRm conducts a cubic step (eq. (3)), a momentum step (eqs. (4) and (5)), and a monotone step (eq. (6)). In particular, the cubic step solves a subproblem of the second-order Taylor expansion with a cubic regularizer at the current iterate $\mathbf{x}_k$. The cubic step can be implemented efficiently by adopting the solver based on the Hessian-vector product approach (see Section 4.3 for details). The momentum step is an extrapolation step that aims to accelerate the algorithm. We note that the momentum step requires very little additional computation compared to the cubic step, but offers substantial advantage for accelerating the algorithm. The monotone step chooses the next iteration point between the cubic step and the momentum step to achieve the minimum function value. This guarantees that the algorithm outputs a desirable monotonically decreasing function value sequence, and helps to establish the convergence guarantee under nonconvex optimization.

We further highlight the ideas in the design of CRm. First, we choose the momentum in the direction of $\mathbf{y}_{k+1} - \mathbf{y}_k$, which has been used for the first-order methods with momentum for nonconvex problems (Li et al., 2017; Yao et al., 2017). Second, the momentum parameter β_{k+1} in eq. (4) is set to be adaptive (in fact proportional) to the norm of the progress made in the cubic regularization step and the norm of gradient, i.e., $\|\mathbf{y}_{k+1} - \mathbf{x}_k\|$ and $\|\nabla f(\mathbf{y}_{k+1})\|$. In this way, if the iterate is far away from a second-order stationary point, $\|\mathbf{y}_{k+1} - \mathbf{x}_k\|$ and $\|\nabla f(\mathbf{y}_{k+1})\|$ are large so that the momentum takes a large stepsize to make good progress. On the other hand, as the iterate is close to the stationary point, $\|\mathbf{y}_{k+1} - \mathbf{x}_k\|$ and $\|\nabla f(\mathbf{y}_{k+1})\|$ are small so that the momentum takes a small momentum stepsize in order not to miss the stationary point. It turns out that such a choice of the momentum parameter is critical to guarantee the convergence of CRm (as can be seen in the proof) as well as achieving acceleration. Our experiments (see Section 5) show that such a momentum scheme can substantially accelerate the convergence of CR in various nonconvex optimization problems. Therefore, the requirement of the adaptive step

Algorithm 1 CRm

1: **Input:** Initialization $\mathbf{x}_0 = \mathbf{y}_0 \in \mathbb{R}^d, \rho < 1, M > L_2$

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

3: **Cubic step:**

$$\mathbf{s}_{k+1} = \underset{\mathbf{s}}{\operatorname{argmin}} \nabla f(\mathbf{x}_k)^\top \mathbf{s} + \frac{1}{2}\mathbf{s}^\top \nabla^2 f(\mathbf{x}_k)\mathbf{s} + \frac{M}{6}\|\mathbf{s}\|^3$$

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

4: **Momentum step:**

$$\beta_{k+1} = \min\{\rho, \|\nabla f(\mathbf{y}_{k+1})\|, \|\mathbf{y}_{k+1} - \mathbf{x}_k\|\} \tag{4}$$

$$\mathbf{v}_{k+1} = \mathbf{y}_{k+1} + \beta_{k+1}(\mathbf{y}_{k+1} - \mathbf{y}_k) \tag{5}$$

5: **Monotone Step:**

$$\mathbf{x}_{k+1} = \underset{\mathbf{x} \in \{\mathbf{y}_{k+1}, \mathbf{v}_{k+1}\}}{\operatorname{argmin}} f(\mathbf{x}) \tag{6}$$

6: **end for**

size β_k in eq. (4) is not only intuitively reasonable but also theoretically sound.

In the monotone step, the algorithm compares the function values between the cubic regularization step and the momentum step, and choose the better one to perform the next step. In this way, the proposed accelerated CR algorithm is guaranteed to be monotone, i.e., the generated function value sequences are monotonically decreasing. This monotone step is not required in convex optimization, but it seems crucial in nonconvex optimization due to the landscape of nonconvex function does not have strong structure as convex function. We further note that although the momentum step may not play a role in every iteration due to the monotone step, our experiments show that the momentum step does participate for most iterations during the course of convergence, validating its importance to accelerate the algorithm.

3 CONVERGENCE ANALYSIS OF CRm

In this section, we establish both the global and the local convergence rates of CRm to a second-order stationary point.

3.1 GLOBAL CONVERGENCE OF CRm

First recall that our goal is to minimize a twice-differentiable nonconvex function $f(\mathbf{x})$ (c.f. eq. (1)). We adopt the following standard assumptions on the objective function.

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

1. *f is twice-continuously differentiable and bounded below, i.e., $f^\star \triangleq \inf_{\mathbf{x} \in \mathbb{R}^d} f(\mathbf{x}) > -\infty$;*
2. *For all $\alpha \in \mathbb{R}$, the sublevel set $\{\mathbf{x} : f(\mathbf{x}) \leqslant \alpha\}$ of f is bounded;*
3. *The gradient $\nabla f(\cdot)$ and Hessian $\nabla^2 f(\cdot)$ are L_1 and L_2-Lipschitz continuous, respectively.*

Assumption 1 imposes standard conditions on the nonconvex objective function f. In particular, the bounded sublevel set condition in item 2 is satisfied whenever f is coercive, i.e., $f(\mathbf{x}) \to +\infty$ as $\|\mathbf{x}\| \to +\infty$. This is true for many non-negative loss functions under mild conditions.

Based on Assumption 1, we characterize the global convergence rate of CRm to a second-order stationary point in the following result. We refer the readers to the supplementary materials for the proof.

Theorem 1 (Global convergence rate)**.** *Let Assumption 1 hold and fix any $\epsilon \leqslant 1$. Then, the sequence $\{\mathbf{x}_k\}_{k \geqslant 0}$ generated by CRm contains an ϵ-second-order stationary point provided that the total number of iterations k satisfies that*

$$k \geqslant \frac{C}{\epsilon^{3/2}}, \tag{7}$$

where C is a universal positive constant and is specified in the proof.

Theorem 1 establishes the global convergence rate to an ϵ-second-order stationary point for CRm. Although the obtained convergence rate of CRm achieves the same order as that of the original CR algorithm in Nesterov and Polyak (2006), which is in fact the best that one can expect for general nonconvex optimization, the technical proof critically exploits the design of the momentum scheme, and requires substantial machinery to handle the momentum step. Further in Section 5, we demonstrate via various experiments that CRm do enjoy the momentum acceleration and converge much faster than the original CR algorithm.

3.2 LOCAL CONVERGENCE OF CRm

It is well known that Newton-type second-order algorithms enjoy a local quadratic convergence rate for minimizing strongly convex functions. While strong convexity is a restrictive condition in nonconvex optimization, many nonconvex problems such as phase retrieval Zhang et al. (2017) and low-rank matrix recovery Tu et al. (2016) have been shown to satisfy the following more relaxed local error bound condition (Yue et al., 2018).

Assumption 2 (Local error bound)**.** *Denote $\mathcal{X}$ as the set of second-order stationary points of f. There exists $\kappa, r > 0$ such that for all $\mathbf{x} \in \{\mathbf{x} : \mathrm{dist}(\mathbf{x}, \mathcal{X}) \leqslant r\}$, it holds that*

$$\mathit{dist}(\mathbf{x}, \mathcal{X}) \leqslant \kappa \|\nabla f(\mathbf{x})\|, \tag{8}$$

where dist$(\mathbf{x}, \mathcal{X})$ denotes the point-to-set distance between $\mathbf{x}$ and $\mathcal{X}$.

One can easily check that all strongly convex functions satisfy the above local error bound condition. Therefore, the local error bound condition is a more general geometry than strong convexity.

Next, we explore the local convergence property for CRm under the local error bound condition. Typically, such a property is due to the usage of the Hessian information in the algorithm. In CRm, the momentum step does not directly exploit the Hessian information. Hence, it is not clear *a priori* by including the momentum step whether CRm still enjoys the local quadratic convergence property. The following theorem provides an affirmative answer.

Theorem 2. *Let Assumptions 1 and 2 hold. Then, the sequence $\{\mathbf{x}_k\}_{k\geqslant 0}$ generated by CRm with $M > L_2$ converges quadratically to a point $\mathbf{x}^\star \in \mathcal{X}$, where $\mathcal{X}$ is the set of second-order stationary points of f. That is, there exists an integer k_1 such that for all $k \geqslant k_1$,*

$$\|\mathbf{x}_{k+1} - \mathbf{x}^\star\| \leqslant C\|\mathbf{x}_k - \mathbf{x}^\star\|^2, \tag{9}$$

where C is a universal positive constant and is specified in the proof.

Under the local error bound condition, Theorem 2 shows that CRm enjoys a quadratic convergence rate as shown in eq. (9). To elaborate, note that Theorem 1 guarantees the convergence of CRm to a second-order stationary point, i.e., $\|\mathbf{x}_k - \mathbf{x}^\star\| \to 0$ as $k \to \infty$. Thus, the recursion in eq. (9) implies that $C\|\mathbf{x}_k - \mathbf{x}^\star\| \leqslant (C\|\mathbf{x}_{k_1} - \mathbf{x}^\star\|)^{2^{k-k_1}}$, which is at a quadratic converge rate. In particular, the region of quadratic convergence is defined by $\|\mathbf{x}_k - \mathbf{x}^\star\| \leqslant 1/C$. Such quadratic convergence achieves an ϵ-accuracy second-order stationary point within $k = O(\log\log(1/\epsilon))$ number of iterations, which is much faster than the linear converge rate of fisrt-order methods in local region.

Local quadratic convergence has also been established for the original CR algorithm under the local error bound condition Yue et al. (2018). As a comparison, our proof of Theorem 2 for CRm exploits the proposed momentum scheme, which results in additional terms that requires extra effort to handle.

4 INEXACT VARIANTS OF CRm

The major computational load of CRm lies in the cubic step, which requires to solve a computationally costly optimization problem. In this section, we explore three implementation schemes that can efficiently perform the cubic step without sacrificing the acceleration performance.

4.1 CUBIC STEP WITH INEXACT HESSIAN

The cubic step requires the full Hessian information, which can be too costly in practice. Instead, we consider performing the following cubic step with an inexact approximation of the Hessian.

$$\hat{\mathbf{x}}_{k+1} = \operatorname*{argmin}_{\mathbf{s} \triangleq \mathbf{x} - \mathbf{x}_k} \nabla f(\mathbf{x}_k)^\top \mathbf{s} + \frac{1}{2}\mathbf{s}^\top \mathbf{H}_k \mathbf{s} + \frac{M}{6}\|\mathbf{s}\|^3, \tag{10}$$

where $\mathbf{H}_k$ denotes the inexact estimation of the full Hessian $\nabla^2 f(\mathbf{x}_k)$, and their difference is assumed to satisfy the following criterion. Section 4.2 proposes a subsampling scheme to achieve Assumption 3 for the finite-sum problem.

Assumption 3. *The inexact Hessian $\mathbf{H}_k$ in eq. (10) satisfies, for all $k \geqslant 0$,*

$$\|\mathbf{H}_k - \nabla^2 f(\mathbf{x}_k)\| \leqslant \epsilon_1.$$

Assumption 3 assumes that the inexact Hessian is close to the exact one in terms of a small operator norm gap. Such inexact criterion has been considered in Tripuraneni et al. (2017); Xu et al. (2017) to study the convergence property of the inexact CR algorithm.

Next, we study the convergence of the inexact variant of CRm by replacing the cubic step in eq. (3) with the inexact cubic step in eq. (10). Our main result is summarized as follows, and the proof is provided in the supplemental materials.

Theorem 3. *Let Assumptions 1 and 3 hold and fix any $\epsilon \leqslant 1$. Then, the sequence $\{\mathbf{x}_k\}_{k\geqslant 0}$ generated by the inexact CRm with $M > 2L_2/3 + 2$ and $\epsilon_1 = \theta\sqrt{\epsilon}$ contains an ϵ-second-order stationary point provided that the total number of iterations k satisfies that*

$$k \geqslant \frac{C}{\epsilon^{3/2}}, \tag{11}$$

where C, θ are universal constants, and are specified in the proof.

Theorem 3 shows that, under a proper inexact criterion, the iteration complexity of inexact CRm is on the same order as that of exact CRm for achieving an ϵ-second-order stationary point. Since the inexact Hessian saves the computation in each iteration comparing to the full Hessian, it is clear that the overall computation complexity of the inexact CRm is less than that of the exact cases.

In Appendix A.2, we verify through experiments that the inexact algorithm do perform much better than the corresponding exact version.

We note that the proof of Theorem 3 suggests that the condition that $\|\mathbf{y}_{k+1} - \mathbf{x}_k\| \leqslant \epsilon_1$ implies the point $\mathbf{x}_{k+1}$ is an ϵ-second-order stationary point, where $\epsilon_1 = \theta\sqrt{\epsilon}$. Thus, the implementation of the inexact CRm can terminate by checking the satisfaction of the condition $\|\mathbf{y}_{k+1} - \mathbf{x}_k\| \leqslant \epsilon_1$.

4.2 INEXACT CRm VIA SUBSAMPLING

In this subsection, we consider a general finite-sum optimization problem, where inexact CRm can be implemented via subsampling. More specifically, consider to solve the following optimization problem:

$$f(x) \triangleq \sum_{i=1}^{n} f_i(x), \tag{12}$$

where $f_i(\cdot)$ is possibly nonconvex. Furthermore, we assume that Assumption 1 holds for each $f_i(\cdot)$. For finite-sum problems, the full Hessian can be approximated by the Hessian of a mini-batch of data samples each uniformly randomly drawn from the dataset, i.e.,

$$\mathbf{H}_k = \frac{1}{|S_1|} \sum_{i \in S_1} \nabla^2 f_i(\mathbf{x}_k). \tag{13}$$

We use the subsampling technique introduced in Kohler and Lucchi (2017) to satisfy the inexact condition in Assumption 3. The following theorem provides our characterization of the overall Hessian sample complexity in order to guarantee the convergence of the subsampling algorithm with high probability over the entire iteration process.

Theorem 4 (Total Hessian sample complexity)**.** *Assuming that Assumption 1 holds for each $f_i(\cdot)$, and let the sub-sampled mini-batch of Hessians $\mathbf{H}_k, k = 0, 1, \ldots$ satisfies*

$$|S_1| = \left(\frac{8L_1^2}{\theta^2\epsilon} + \frac{4L_1}{3\theta\sqrt{\epsilon}}\right) \log\left(\frac{4d}{\epsilon^{3/2}\delta}\right),$$

then the sequence $\{\mathbf{x}_k\}_{k \geqslant 0}$ generated by the inexact CRm with $M > L_2 + 2$ outputs an ϵ-second-order stationary point with probability at least $1 - \delta$ by taking at most the following number of Hessian samples in total:

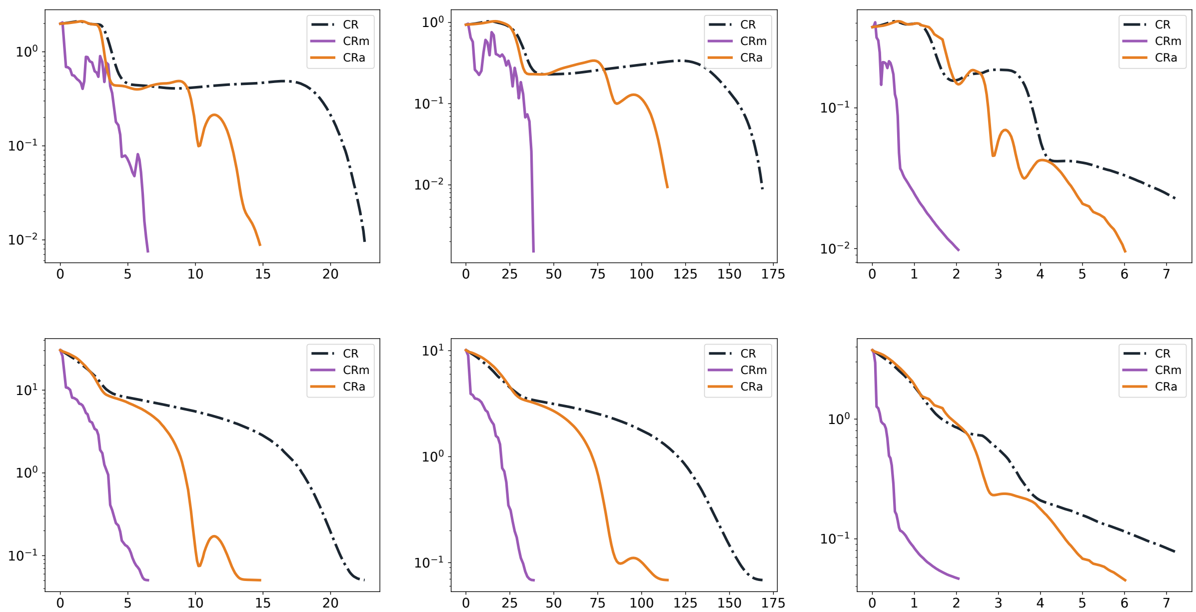
$$S \leqslant C \left(\frac{8L_1^2}{\theta^2\epsilon^{5/2}} + \frac{4L_1}{3\theta\epsilon^2}\right) \log\left(\frac{4d}{\epsilon\delta}\right).$$

Theorem 4 characterizes the total Hessian sample complexity to guarantee the convergence of CRm with high probability. This is the first such a type result for stochastic CR algorithms. Note that previous studies Kohler and Lucchi (2017); Xu et al. (2017) on subsampling CR provide only the Hessian sample complexity per iteration to guarantee inexactness condition with high probability. Our result indicates that even over the entire iteration process, the convergence is still guaranteed with high probability. In fact, if we let N denote the total sample complexity, Theorem 4 implies that the failure probability δ decays exponentially fast as the total sample complexity N becomes asymptotically large. Such a result by nature is stronger than those that characterize the convergence only in expectation, not in (high) probability, in existing literature.

4.3 EFFICIENT SOLVERS FOR CUBIC STEP

Since the cubic step does not have a closed form solution, an inexact solver is typically used for solving the cubic step. Various solvers have been proposed to approximately solve such a subproblem. The first type of solver is based on the Lanczos method (Cartis et al., 2011a,b), which solves the cubic subproblem in a Krylov subspace $\mathcal{K} = \operatorname{span}\{\nabla f(\mathbf{x}_k), \nabla^2 f(\mathbf{x}_k)\nabla f(\mathbf{x}_k), \cdots\}$ instead of in the entire space. Each step of the solver can be implemented efficiently with a computation cost of $O(d)$ (Kohler and Lucchi, 2017). Moreover, building the subspace requires a Hessian-vector product, which introduces a cost of $O(nd)$ per additional subspace dimension for finite-sum problem with n data samples. The second type of solver is proposed by Agarwal et al. (2017), which is based on the techniques of Hessian-vector product and binary search. The proposed solver can find an approximate solution of the cubic subproblem with a total cost of $O(nd/\epsilon^{1/4})$ for finite-sum problems, where ϵ is the desired accuracy. Carmon and Duchi (2016) proposed another solver based on the gradient descent method. The solver finds an approximate solution of the cubic subproblem within $O(\epsilon^{-1}\log(1/\epsilon))$ iterations for large ϵ and $O(\log(1/\epsilon))$ iterations for small ϵ.

All of these solvers can be applied to solve the cubic step in CRm. Note that the momentum and monotone steps in CRm introduce order-level less computation complexity compared to these solvers for solving the cubic step. Thus, CRm have the same per-iteration computational complexity as CR when implementing the same solver, and have at least the same overall computational complexity as CR (in fact, much less overall computational complexity in practice as demonstrated by our experiments). As the solvers solve the cubic subproblem up to certain accuracy in practice, the total computation complexity of CRm to achieve a second-order stationary point can still be established.



(a) a9a (b) covtype (c) ijcnn1

Figure 1: Nonconvex logistic regression. Top: gradient norm v.s. time. Bottom: function value gap v.s. time.

5 EXPERIMENTS

5.1 SETUP

We compare the performance among the following six algorithms: cubic regularization algorithm (**CR**) in Nesterov and Polyak (2006), accelerated cubic regularization algorithm (**CRa**) in Nesterov (2008) (whose convergence guarantee has not been established), cubic regularization algorithm (**CRm**) with momentum, cubic regularization algorithm with inexact Hessian (**CR_I**), accelerated cubic regularization with inexact Hessian (**CRa_I**), CRm with inexact Hessian (**CRm_I**). In this section, we present the comparison among the three exact algorithms. The comparisons among the inexact variants are presented in Appendix A due to space limitation. The details of the experiment settings can also be found in Appendix A.

We conduct two experiments. The first experiment solves the following logistic regression problem with a nonconvex regularizer

$$\min_{\mathbf{w}\in\mathbb{R}^d} -\left(\frac{1}{n}\sum_{i=1}^{n} y_i \log\left(\frac{1}{1+e^{-\mathbf{w}^T\mathbf{x}}}\right)\right.$$
$$\left. +(1-y_i)\log\left(\frac{e^{\mathbf{w}^T\mathbf{x}}}{1+e^{-\mathbf{w}^T\mathbf{x}}}\right)\right) + \alpha\sum_{i=1}^{d}\frac{w_i^2}{1+w_i^2},$$

where we set $\alpha = 0.1$ in our experiment. The second experiment solves the following nonconvex robust linear regression problem

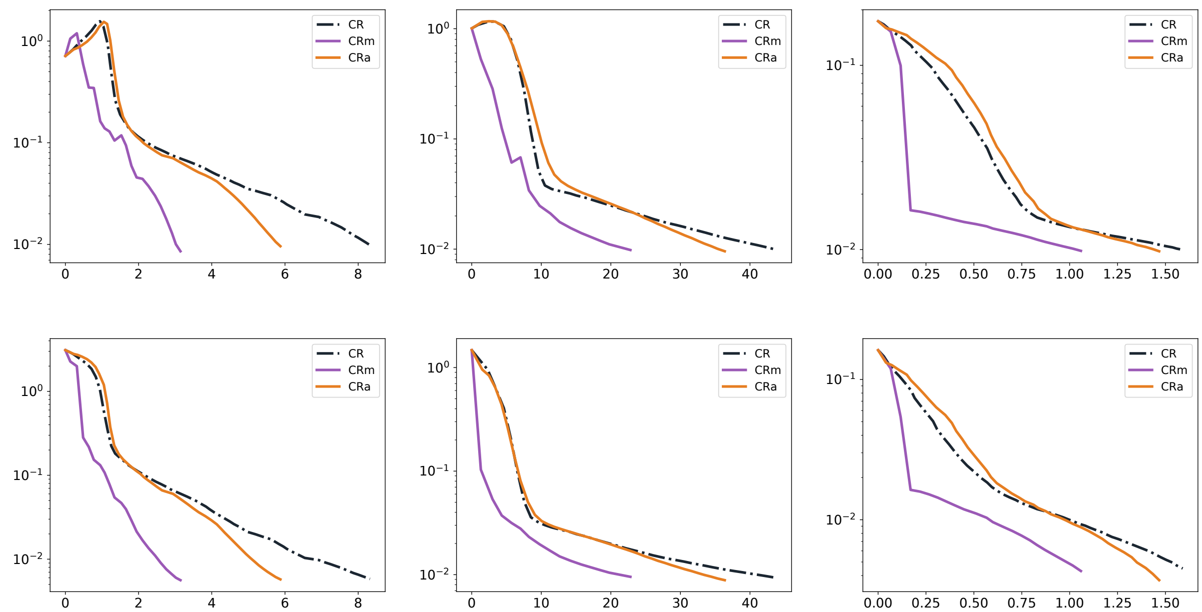
$$\min_{\mathbf{w}\in\mathbb{R}^d} \frac{1}{n}\sum_{i=1}^{n}\eta(y_i - \mathbf{w}^T\mathbf{x}_i), \tag{14}$$

where $\eta(x) = \log(\frac{x^2}{2}+1)$. Each experiment is performed over three datasets, i.e., a9a, covtype, and ijcnn (Chang and Lin, 2011).

5.2 RESULTS

Figures 1 and 2 show the results of the two experiments for comparing the three exact algorithms, respectively. From both figures, it can be seen that CRm outperforms the vanilla CR, which demonstrates that the momentum step in CRm significantly accelerates the CR algorithm for nonconvex problems. Also, CRm outperforms CRa in the experiments with datasets a9a and covtype, while its performance is comparable to CRa in the experiments with dataset ijcnn1. Thus, our proposed momentum step achieves a faster convergence than the Nesterov's acceleration scheme for CR.

We note that similar comparisons are observed in the comparison of the corresponding three inexact variants of the algorithms (see Figures 3 and 4 in Appendix A), i.e., our momentum scheme with inexact Hessian outperforms other inexact CR algorithms. We also note that all inexact variants of the algorithms outperform their exact counterparts (see Figures 5 and 6 in Appendix A). Hence,



(a) a9a (b) covtype (c) ijcnn1

Figure 2: Robust linear regression. Top: gradient norm v.s. time. Bottom: function value gap v.s. time.

the inexact implementation plays an important role in reducing the computation complexity of these CR type of algorithms in practice.

6 CONCLUSION

In this paper, we proposed a momentum scheme to accelerate the cubic regularization algorithm. We showed that the order of the global convergence rate of the proposed algorithm CRm is at least as fast as its vanilla version. We also established the local quadratic convergence property for the proposed algorithm, and extended our analysis for the proposed algorithm to the inexact Hessian case and established the total Hessian sample complexity to guarantee the convergence with high probability. We further conducted various experiments to demonstrate the advantage of applying momentum for accelerating the cubic regularized algorithm.

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.

Carmon, Y. and Duchi, J. C. (2016). Gradient descent efficiently finds the cubic-regularized non-convex Newton step. *arXiv:1803.09357*.

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. L. (2011a). Adaptive cubic regularization methods for unconstrained optimization. Part I : Motivation, convergence and numerical results. *Mathematical Programming*, 127(2):245–295.

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

Chang, C. and Lin, C. (2011). LIBSVM: A library for support vector machines. *ACM Transactions on Intelligent Systems and Technology*, 2(3):1–27.

Dauphin, Y. N., Pascanu, R., Gulcehre, C., Cho, K., Ganguli, S., and Bengio, Y. (2014). Identifying and attacking the saddle point problem in high-dimensional nonconvex optimization. In *Proc. International Conference on Neural Information Processing Systems (NIPS)*, pages 2933–2941.

Fan, J. and Yuan, Y. (2005). On the quadratic convergence of the levenberg-marquardt method without nonsingularity assumption. *Computing*, 74(1):23–39.

Ge, R., Huang, F., Jin, C., and Yuan, Y. (2015). Escaping from saddle points — online stochastic gradient for tensor decomposition. In *Proc. Conference on Learning Theory (COLT)*, volume 40, pages 797–842.

Ghadimi, S. and Lan, G. (2016). Gradient methods for nonconvex nonlinear and stochastic programming. *Mathematical Programming*, 156(1):59–99.

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

Griewank, A. (1981). The modification of newtons method for unconstrained optimization by bounding cubic terms. *Technical report, Department of Applied Mathematics and Theoretical Physics, University of Cambridge*.

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*.

Jin, C., Ge, R., Netrapalli, P., Kakade, S. M., and Jordan, M. I. (2017). How to escape saddle points efficiently. In *Proc. 34th International Conference on Machine Learning (ICML)*, volume 70, pages 1724–1732.

Jin, C., Liu, L., Ge, R., and Jordan, M. I. (2018). Minimizing nonconvex population risk from rough empirical risk. *arXiv:1803.09357*.

Jin, C., Netrapalli, P., and Jordan, M. I. (2017). Accelerated gradient descent escapes saddle points faster than gradient descent. *arXiv:1711.10456*.

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.

Li, D.-H., Fukushima, M., Qi, L., and Yamashita, N. (2004). Regularized newton methods for convex minimization problems with singular solutions. *Computational Optimization and Applications*, 28(2):131–147.

Li, H. and Lin, Z. (2015). Accelerated proximal gradient methods for nonconvex programming. In *Proc. 28th Advances in Neural Information Processing Systems (NIPS)*, pages 379–387.

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*.

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. T. (2006). Cubic regularization of newton method and its global performance. *Mathematical Programming*, 108(1):177–205.

Sun, J., Qu, Q., and Wright, J. (2015). Complete dictionary recovery using nonconvex optimization. In *Proc. 32nd International Conference on Machine Learning (ICML)*, volume 37, pages 2351–2360.

Sun, J., Qu, Q., and Wright, J. (2017). A geometric analysis of phase retrieval. *Foundations of Computational Mathematics*, pages 2379–2383.

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.

Tu, S., Boczar, R., Simchowitz, M., Soltanolkotabi, M., and Recht, B. (2016). Low-rank solutions of linear matrix equations via Procrustes flow. In *Proc. 33rd International Conference on Machine Learning (ICML)*, pages 964–973.

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

Wang, Z., Zhou, Y., Liang, Y., and Lan, G. (2019b). Sample complexity of stochastic variance-reduced cubic regularization for nonconvex optimization. *In Proc. 22nd International Conference on Artificial Intelligence and Statistics (AISTATS)*.

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*.

Yamashita, N. and Fukushima, M. (2001). On the rate of convergence of the levenberg-marquardt method. *Topics in Numerical Analysis*, pages 239–249.

Yao, Q., Kwok, J. T., Gao, F., Chen, W., and Liu, T.-Y. (2017). Efficient inexact proximal gradient algorithm for nonconvex problems. In *Proc. 26th International Joint Conference on Artificial Intelligence (IJCAI-17)*, pages 3308–3314.

Yue, M., Zhou, Z., and So, A. (2018). On the quadratic convergence of the cubic regularization method under a local error bound condition. *arXiv:1801.09387*.

Yue, M.-C., Zhou, Z., and Man-Cho So, A. (2016). A Family of inexact SQA methods for non-smooth convex minimization with provable convergence guarantees based on the Luo-Tseng error bound property. *arXiv:1605.07522*.

Zhang, H., Liang, Y., and Chi, Y. (2017). A nonconvex approach for phase retrieval: Reshaped Wirtinger flow and incremental algorithms. *Journal of Machine Learning Research*, 18(141):1–35.

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