Curvature-oriented variance reduction methods for nonconvex stochastic optimization*

Qiankun Shi† Shiji Zuo‡ Yongxiang Liu§ Xiao Wang¶

February 14, 2026

Abstract

When pursuing an approximate second-order stationary point in nonconvex constrained stochastic optimization, is it possible to design a stochastic second-order method that achieves the same sample complexity order as in the unconstrained setting? To address this question in this paper, we first introduce Carme, a curvature-oriented variance reduction method designed for unconstrained nonconvex stochastic optimization. Under the smoothness assumption of the stochastic objective function, the sample complexity of Carme, regarding evaluations of first- and second-order oracles, improves the best-known results in the literature. We then propose Carme-ALM, an augmented Lagrangian-based variant of Carme tailored to nonconvex stochastic optimization with deterministic constraints. Under suitable conditions, we prove that Carme-ALM achieves a sample complexity for finding an approximate second-order stationary point that is comparable to that of the unconstrained case. This provides a positive, yet conditional, answer to the question posed above.

Keywords: Nonconvex constrained optimization, cubic regularization, second-order stationarity, sample complexity

MSC Classification: 90C30, 90C26, 65K05

1 Introduction

In this paper, we consider the following nonconvex stochastic optimization problem:

$$\min_{x\in\mathbb{R}^n} \quad f(x) := \mathbb{E}_\xi[F(x,\xi)] \quad \text{s.t.} \quad x \in X. \tag{1}$$

Here, $\mathbb{E}_\xi$ represents the expectation taken with respect to the random variable ξ, that is independent of x and defined on a probability space Ξ. For any $\xi \in \Xi$, the mapping $F(\cdot,\xi) : \mathbb{R}^n \to \mathbb{R}$ is continuously differentiable and possibly nonconvex. Nonconvex stochastic optimization is prevalent in modern machine learning, data science, and operations research, with applications in deep neural network training [22], robust control [45], and constrained matrix factorization [15]. These problems feature complex objective landscapes with multiple local minima and saddle points, further complicated by nonlinear constraints. Large-scale datasets have necessitated the use of stochastic methods, which rely on noisy gradient and Hessian estimates to achieve computational efficiency. However, the interplay of nonconvexity, stochasticity and (potentially nonconvex) constraints poses significant theoretical and practical challenges, including

*This work is partially supported by National Science Foundation of China (No. 12271278).

†Sun Yat-sen University, Guangzhou, China, and Pengcheng Laboratory, Shenzhen, China. Email: shiqk@mail2.sysu.edu.cn

‡Rensselaer Polytechnic Institute, New York, United States. Email: zuos@rpi.edu

§Pengcheng Laboratory, Shenzhen, China. Email: liuyx@pcl.ac.cn

¶Corresponding author. Sun Yat-sen University, Guangzhou, China. Email: wangx936@mail.sysu.edu.cn

escaping saddle points and ensuring constraint satisfaction. In this paper, we are interested in two types of nonconvex stochastic optimization problems. The first type is the unconstrained stochastic optimization, where the feasible set X is the whole n-dimensional space, i.e.,

$$X = \mathbb{R}^n.$$

The second one is the general equality-constrained stochastic optimization, where the feasible set is determined by possibly nonlinear constraints, i.e.,

$$X = \{x \in \mathbb{R}^n : c_i(x) = 0, i \in \mathcal{E}\}. \tag{2}$$

Here, $\mathcal{E}$, with $|\mathcal{E}| = m$, denotes the index set of equality constraints, and $c_i : \mathbb{R}^n \to \mathbb{R}, i \in \mathcal{E}$, are twice continuously differentiable and possibly nonconvex.

Traditional first-order methods for unconstrained optimization, such as gradient descent methods, are known to converge to first-order stationary points with small gradient norms. However, in nonconvex settings, such points may correspond to saddle points or even local maxima, leading to suboptimal solutions. To address these issues, second-order methods that exploit curvature information have been widely studied and shown to converge to second-order stationary points. Among second-order optimization techniques, cubic regularization, first introduced by Nesterov and Polyak [35], has emerged as a powerful framework for nonconvex optimization. By augmenting the quadratic Newton step with a cubic regularization term, this approach ensures convergence to second-order stationary points and enables efficient escape from saddle points. It also achieves superior iteration complexity compared with first-order methods. Stochastic variants of cubic regularization, which employ stochastic approximations of gradients and Hessians, have been developed to handle large-scale problems [40, 48]. These advances have reduced the per-iteration computational cost while preserving strong convergence guarantees, making stochastic cubic methods particularly appealing for large scale unconstrained optimization problems. Despite these successes, the application of stochastic cubic regularization to nonconvex constrained optimization remains underexplored. Constrained problems add further complexity, as they require maintaining feasibility while optimizing a nonconvex objective. Classical approaches, such as interior-point methods and penalty methods, have been adapted to handle constraints, but their integration with stochastic cubic regularization remains limited. Early efforts, such as [2, 9, 26], demonstrate robustness of cubic methods for nonconvex constrained problems, yet they rely on deterministic gradients, making them computationally prohibitive for large-scale applications. In a recent work [42], Wang considers to incorporate second-order information to solve constrained stochastic optimization problems. However, the treatment of stochasticity in [42] is suboptimal, resulting in relatively high computational complexity and reduced ability to efficiently escape saddle points.

These observations raise a key question:

> *Is it possible to design a stochastic second-order method for nonconvex constrained stochastic optimization to obtain an approximate second-order stationary point, while achieving the same sample complexity order as in unconstrained setting?*

To address this question, we will first propose a curvature-oriented variance reduction method for unconstrained stochastic optimization that owns improved sample complexity than existing methods. Second, we will develop an augmented Lagrangian–based curvature-oriented algorithm for nonconvex constrained optimization, which, under certain conditions, attains the same complexity order as in the unconstrained setting.

1.1 Related work

Second-order optimization

Second-order optimization methods are pivotal for identifying second-order stationary points, effectively avoiding saddle points and local maxima in nonconvex optimization landscapes. Newton's method, a cornerstone of second-order approaches, has been extensively studied in its classical form [21], extensions to nonconvex settings [36], and stochastic variants [37, 43]. The cubic-regularized Newton method has garnered significant attention due to its enhanced global and local convergence properties [27, 35]. However, the computational complexity of exactly solving cubic subproblems has prompted research into inexact cubic regularization techniques and efficient solvers [7, 10, 11].

Despite these progress, traditional cubic regularization methods rely on deterministic oracles, requiring exact gradient and Hessian information, which becomes prohibitively expensive in large-scale data regimes. To address this, cubic regularization variants leveraging stochastic oracles have been proposed to approximate gradient and Hessian information [12, 23, 25, 30, 40]. Nevertheless, these methods often necessitate large batch sizes to ensure convergence, resulting in high sample complexity [13]. Chayti et al. [14] introduced a momentum-based stochastic cubic algorithm that samples gradients and Hessians only once per iteration, reducing computational overhead. However, this approach overlooks the potential mismatch in optimal complexities for gradient and Hessian computations, leading to inefficient Hessian sampling. An alternative strategy employs slightly larger batch sizes for stochastic gradients and Hessians, with distinct batch sizes combined with variance reduction techniques to achieve lower sample complexity [48, 50]. However, these methods typically require periodic checkpoints to maintain convergence guarantees. While most studies focus on finding $(\epsilon, \sqrt{\epsilon})$-stationary points, recent efforts have explored the more general notion of (ϵ, γ)-stationary points by incorporating negative curvature search into cubic regularization frameworks [1]. In this work, we consider the same (ϵ, γ)-stationarity criterion as our optimality notion.

Constrained stochastic optimization

Constrained optimization methods face additional challenges to ensure the solution's feasibility. For linear constraints, projective set constraints, and even convex functional constraints, the theoretical framework is relatively well-established [4, 36]. However, when constraint functions exhibit nonconvexity and the objective functions are stochastic, ensuring feasibility becomes significantly more challenging. Algorithms for solving such nonconvex constrained stochastic optimization problems primarily include proximal point methods [5, 6, 28], sequential quadratic programming methods [3, 18, 19, 33, 34], directional decomposition methods [38], penalty methods [16, 17, 29, 31, 32, 39]. The main concept behind the first three classes of methods is to transform complex nonconvex constrained optimization problems into a sequence of simpler constrained subproblems. In contrast, penalty methods, including augmented Lagrangian methods, reformulate the problem as an unconstrained optimization task for solution. Among those methods, the state-of-the-art sample complexity for finding an ϵ-KKT point of (1)-(2) is $O(\epsilon^{-3})$ [28, 32, 39], under mild conditions. However, these methods purely focus on identifying first-order stationary points and cannot guarantee second-order stationarity, even when some methods incorporate Hessian information [33, 34].

To obtain an approximate second-order stationary point, Cartis et al. [9] study high-order optimality conditions for smooth nonlinear constrained optimization and propose a two-phase framework capable of achieving approximate first-, second-, and third-order criticality. Their analysis establishes worst-case evaluation complexity bounds and reveals intrinsic difficulties of attaining high-order constrained critical points using standard penalty-based approaches. Xie and Wright [46] analyze the worst-case complexity of a proximal augmented Lagrangian (AL) framework for nonconvex optimization with nonlinear equality constraints. They show that approximate first- and second-order KKT points can be obtained under suitable choices of the penalty and proximal parameters, with iteration complexity depending explicitly on

the penalty growth. Goyens et al. [26] revisit Fletcher's augmented Lagrangian for equality-constrained optimization and propose a Gradient–Eigenstep method grounded in a Riemannian optimality framework. Under standard regularity assumptions, their method achieves (ϵ, γ)-SSPs in $O(\epsilon^{-2} + \gamma^{-3})$ iterations for deterministic problems. Building on these deterministic advances, Wang [42] extends cubic regularization techniques to stochastic equality-constrained optimization by embedding an inexact cubic solver into a linearized augmented Lagrangian framework. However, achieving second-order stationarity under stochastic oracles incurs relatively high complexity. Overall, these results highlight the intrinsic difficulty of attaining second-order stationarity for stochastic nonconvex constrained optimization with sample complexity comparable to that of unconstrained problems.

1.2 Contributions

This paper studies stochastic optimization from unconstrained to equality-constrained settings, with a focus on sample complexity of algorithms for finding approximate second-order stationary points. We first propose Carme, a curvature-oriented variance reduction method, to find an (ϵ, γ)-SSP of unconstrained stochastic optimization. Compared with existing works (See Table 1), it matches the best-known iteration complexity and gradient-related sample complexity, and improved Hessian-related sample complexity under smoothness of stochastic functions. We then extend Carme to an augmented Lagrangian-based method to solve stochastic optimizaiton with nonconvex equality constraints. The corresponding sample complexity to find an (ϵ, γ)-SSP can match the unconstrained counterpart, under certain conditions with suitable choices of γ (see Table 2).

Algorithm	Setting	Stationary Point	Complexity (Iteration)	Complexity (Gradient)	Complexity (Hessian)
CR [35] ARC [10, 11]	Deterministic	$(\epsilon, \sqrt{\epsilon})$-SSP	$O(\epsilon^{-3/2})$	$O(\epsilon^{-3/2})$	
SCR [30, 47]	Finite-sum	$(\epsilon, \sqrt{\epsilon})$-SSP	$O(\epsilon^{-3/2})$	$\tilde{O}(n\epsilon^{-3/2} + \epsilon^{-7/2})$	$\tilde{O}(n\epsilon^{-3/2} + \epsilon^{-5/2})$
Lite-SVRC [44, 49, 50]	Finite-sum Assumption 3	$(\epsilon, \sqrt{\epsilon})$-SSP	$O(\epsilon^{-3/2})$	$\tilde{O}(n\epsilon^{-3/2})$	$\tilde{O}(n^{2/3}\epsilon^{-3/2})$
SRVRC [48]	Finite-sum Assumption 3	$(\epsilon, \sqrt{\epsilon})$-SSP	$O(\epsilon^{-3/2})$	$\tilde{O}(n\epsilon^{-3/2} \wedge n^{1/2}\epsilon^{-2} \wedge \epsilon^{-3})^\star$	$\tilde{O}(n^{1/2}\epsilon^{-3/2} \wedge \epsilon^{-2})$
SCRTR [1]	Expect. Obj.	(ϵ, γ)-SSP	$O(\epsilon^{-3/2} + \gamma^{-3})$	$\tilde{O}(\epsilon^{-3} + \gamma^{-2}\epsilon^{-2} + \gamma^{-3})$	$\tilde{O}(\epsilon^{-3} + \gamma^{-2}\epsilon^{-2} + \gamma^{-5})$
Carme (ours)	Expect. Obj. Assumption 3	(ϵ, γ)-SSP	$O(\epsilon^{-3/2} + \gamma^{-3})$	$\tilde{O}(\epsilon^{-3} + \gamma^{-2}\epsilon^{-2} + \gamma^{-3})$	$\tilde{O}(\epsilon^{-2} + \gamma^{-4})$

$\star$ $a \wedge b$ means $\min\{a, b\}$.

Table 1: Complexity of second-order algorithms for nonconvex *unconstrained* optimization.

1.3 Notation and preliminaries

We denote the σ-algebra generated by a set of random variables $\{v_1, \ldots, v_m\}$ as $\mathcal{F}(v_1, \ldots, v_m)$. For simplicity, denote $\sum_{t=1}^{0} \rho_t := 0$. Unless otherwise specified, $\|\cdot\|$ denotes the standard Euclidean norm on $\mathbb{R}^n$ and its induced matrix norm. The notation $\mathbf{1}(A)$ refers to the indicator function and it equals 1 if A holds, and 0 otherwise. The notation Pr represents the probability a random event occurs. Big-O notations $\mathcal{O}$ (resp. $\tilde{\mathcal{O}}$) and Ω (resp. $\tilde{\Omega}$) hide constants (resp. logarithmic factors) regarding the upper and lower bound, respectively. In the remainder of this paper, we impose the following assumptions.

Assumption 1 *Let $\mathcal{X}$ be an open convex set that contains $\{x_k\}$ generated by the associated algorithm, and f is lower bounded and $c_i, i = 1, \ldots, m$ are bounded over $\mathcal{X}$, namely there exists $C > 0$ such that $f(x) \geq -C$ and $\|c(x)\| \leq C$ for any $x \in \mathcal{X}$.*

Algorithm	Setting	Stationary Point	Complexity (Iteration)	Complexity (Gradient)	Complexity (Hessian)
OUTER [9]	Deterministic	ϵ-SCP†	$O(\epsilon^{-5})$	$O(\epsilon^{-5})$	
Proximal AL [46]	Deterministic	(ϵ,γ)-SSP	$O(\epsilon^{-11/2}+\gamma^{-7})$	$O(\epsilon^{-11/2}+\gamma^{-7})$	
Gradient-Eigenstep [26]	Deterministic	(ϵ,γ)-SSP	$O(\epsilon^{-2}+\gamma^{-3})$	$O(\epsilon^{-2}+\gamma^{-3})$	
SCPD [42]	Expect. Obj.	(ϵ,γ)-SSP	$O(\epsilon^{-3}\gamma^{-3})$	$\tilde{O}(\epsilon^{-5}\gamma^{-7})$	$\tilde{O}(\epsilon^{-3}\gamma^{-5})$
Carme-ALM (ours)	Expect. Obj. Assumption 3	(ϵ,γ)-SSP	$O(\epsilon^{-3/2}+\gamma^{-3})$	$\tilde{O}(\epsilon^{-3}+\gamma^{-2}\epsilon^{-2}+\gamma^{-4})$	$\tilde{O}(\epsilon^{-2}+\gamma^{-4})$

† ϵ-SCP denotes an ϵ-approximate second-order critical point of the constrained problem or of the feasibility problem $\min \|c(x)\|^2$.

Table 2: Complexity of second-order algorithms for nonconvex *constrained* optimization.

Assumption 2 *Function f is twice continuously differentiable with bounded gradient, L_g^f-Lipschitz continuous gradient, and L_H^f-Lipschitz continuous Hessian. Functions $c_i, i\in\mathcal{E}$ are twice continuously differentiable with bounded gradient, L_g^c-Lipschitz continuous gradient, and L_H^c-Lipschitz continuous Hessian. That is, for any $x,y\in\mathcal{X}$,*

$$\|\nabla f(x)\|\le L_f,\quad \|\nabla f(x)-\nabla f(y)\|\le L_g^f\|x-y\|,\quad \|\nabla^2 f(x)-\nabla^2 f(y)\|\le L_H^f\|x-y\|,$$
$$\|\nabla c_i(x)\|\le L_c,\quad \|\nabla c_i(x)-\nabla c_i(y)\|\le L_g^c\|x-y\|,\quad \|\nabla^2 c_i(x)-\nabla^2 c_i(y)\|\le L_H^c\|x-y\|,\ i\in\mathcal{E}.$$

Assumption 3 *For almost any ξ, $F(\cdot,\xi)$ is twice continuously differentiable with L_g^f-Lipschitz continuous gradients and L_H^f-Lipschitz continuous Hessians.*

Assumption 4 *For any x, $\mathbb{E}_\xi[\nabla F(x,\xi)]=\nabla f(x)$ and $\mathbb{E}_\xi[\nabla^2 F(x,\xi)]=\nabla^2 f(x)$, and for any ξ, $\|\nabla F(x,\xi)-\nabla f(x)\|\le\sigma_g$ and $\|\nabla^2 F(x,\xi)-\nabla^2 f(x)\|\le\sigma_h$.*

Remark 1 *In the study of stochastic cubic regularization methods for unconstrained stochastic optimization [41] and equality-constrained stochastic optimization [42], an assumption analogous to Assumption 4 is adopted. This assumption facilitates the use of vector and matrix concentration inequalities, which are essential to compute the sample complexity of stochastic cubic regularization methods [30, 44, 50].*

In this paper, we study algorithms for finding approximate second-order stationary points (SSPs) of problem (1) which are defined as below.

Definition 1 ((ϵ,γ)-SSP) *Given $\epsilon,\gamma>0$, we call x an (ϵ,γ)-SSP of (1) with $X=\mathbb{R}^n$, if*

$$\|\nabla f(x)\|\le\epsilon\quad\text{and}\quad d^\top\nabla^2 f(x)d\ge-\gamma\|d\|^2\ \text{for any } d\in\mathbb{R}^n. \tag{3}$$

We call x an (ϵ,γ)-SSP of (1)-(2), if there exists $\lambda\in\mathbb{R}^m$ such that

$$\|\nabla f(x)+\nabla c(x)\lambda\|\le\epsilon,\quad \|c(x)\|\le\epsilon \tag{4}$$

and

$$d^\top(\nabla^2 f(x)+\sum_{i=1}^m\lambda_i\nabla^2 c_i(x))d\ge-\gamma\|d\|^2\ \text{for any } d\in\mathrm{Null}(\nabla c(x)^\top). \tag{5}$$

Treating γ as an independent parameter allows for a more flexible notion of approximate second-order stationarity and has become standard in recent studies of stochastic second-order methods [1, 26].

2 Hybrid stochastic estimator

To address the challenges posed by stochasticity in nonconvex optimization, particularly in the context of cubic regularization for problem (1), we propose a hybrid stochastic estimator for variance reduction. This technique extends the STORM (Stochastic Recursive Momentum) method [20], originally developed for gradient variance reduction, to reduce the variance of both stochastic gradient and Hessian estimates simultaneously. Unlike prior stochastic cubic regularization methods [48], which rely on periodic checkpoints such as large-batch or full gradient and Hessian evaluations, our approach eliminates the use of such checkpoints by maintaining recursively updated stochastic gradient and Hessian estimates. This makes it well-suited for large-scale nonconvex optimization problems.

2.1 Motivations

Stochastic cubic regularization methods rely on approximate gradient and Hessian estimates to solve the subproblem:

$$\min_s \quad m_k(s) = \langle g_k, s\rangle + \frac{1}{2}\langle s, H_k s\rangle + \frac{M}{6}\|s\|^3,$$

where $g_k \approx \nabla f(x_k)$ and $H_k \approx \nabla^2 f(x_k)$ are stochastic approximations. The variance in these estimates, as bounded by Assumption 4 ($\|\nabla F(x,\xi) - \nabla f(x)\| \le \sigma_g$, $\|\nabla^2 F(x,\xi) - \nabla^2 f(x)\| \le \sigma_h$), can destabilize convergence and necessitate large batch sizes or checkpoints to ensure accuracy. Checkpoints, as used in [48], involve computing large (even full) batch gradients and Hessians periodically, incurring significant computational overhead, especially in high-dimensional settings.

The STORM method [20] addresses gradient variance by maintaining a recursive momentum-based estimator, achieving near-optimal convergence rates without checkpoints. Inspired by this, we generalize STORM to construct a hybrid estimator that reduces variance in both gradient and Hessian estimates, leveraging the Lipschitz continuity properties of ∇f and $\nabla^2 f$ (Assumptions 2, 3) and the stochasticity bounds (Assumption 4). While the idea of extending STORM to stochastic cubic regularization may appear straightforward, a direct application is nontrivial due to the interaction between variance reduction and cubic regularization, which requires careful control of higher-order error terms.

2.2 Hybrid stochastic estimator

Our hybrid stochastic estimator operates iteratively, updating gradient and Hessian estimates using a momentum-based recursion that exploits the continuity of the stochastic functions. At kth iteration with $k \ge 1$, batches of samples $\mathcal{B}_k^g$ and $\mathcal{B}_k^h$, we compute

$$g_k = \frac{1}{B_k^g}\sum_{i\in\mathcal{B}_k^g}\left[\nabla F(x_k,\xi_i) + (1-\alpha_g)(g_{k-1} - \nabla F(x_{k-1},\xi_i))\right], \tag{6}$$

$$H_k = \frac{1}{B_k^h}\sum_{i\in\mathcal{B}_k^h}\left[\nabla^2 F(x_k,\xi_i) + (1-\alpha_h)\left(H_{k-1} - \nabla^2 F(x_{k-1},\xi_i)\right)\right], \tag{7}$$

and at initial point x_0 we compute

$$g_0 = \frac{1}{B_0^g}\sum_{i\in\mathcal{B}_0^g}\nabla F(x_0,\xi_i), \quad H_0 = \frac{1}{B_0^h}\sum_{i\in\mathcal{B}_0^h}\nabla^2 F(x_0,\xi_i), \tag{8}$$

where $\alpha_g, \alpha_h \in [0,1]$ are the gradient and Hessian momentum parameters, $\mathcal{B}_k^g$ and $\mathcal{B}_k^h$ are index sets with $|\mathcal{B}_k^g| = B_k^g$ and $|\mathcal{B}_k^h| = B_k^h$. We allow $\mathcal{B}_k^g$ and $\mathcal{B}_k^h$ to be sampled independently, although sharing samples

is also possible without affecting the analysis. For notation simplicity, in the following we introduce the deviations of approximate gradients and Hessians by defining

$$\epsilon_k^g = g_k - \nabla f(x_k), \quad \epsilon_k^h = H_k - \nabla^2 f(x_k), \quad k \geq 0. \tag{9}$$

A key feature of the estimator is the recursive correction term, which reuses the same batch $\mathcal{B}_k^g$ (resp. $\mathcal{B}_k^h$) to evaluate gradients (resp. Hessians) at both x_k and x_{k-1}. This recursive update eliminates the need for periodic checkpoints, which typically require substantially larger batch sizes to control the estimator variance. In the classical STORM analysis, variance reduction is achieved by bounding the squared estimation errors of gradients (and Hessians) in terms of $\|x_{k+1} - x_k\|^2$. However, in cubic regularization algorithms, the step size typically scales as $\|x_{k+1} - x_k\|^3$, and this mismatch in polynomial order gives rise to a fundamental technical difficulty in controlling the accumulation of estimation errors. Directly adapting the standard STORM arguments may therefore lead to suboptimal complexity bounds. To overcome this issue, we develop an analytical framework that carefully bounds the estimation errors and employs an adaptive batch size strategy to decouple the algorithmic iterations from the error dynamics.

The following two lemmas provide uniform high-probability bounds on the gradient and Hessian estimation errors, which will be used in the subsequent complexity analysis.

Lemma 1 *Given $K \geq 1$ and $\delta \leq \frac{1}{2K}$, suppose that $\{B_k^g\}$ satisfies*

$$B_0^g = \frac{6480\sigma_g^2 \log^2(1/\delta)}{\bar{\epsilon}_{g,1}^2} \tag{10}$$

and

$$B_k^g = \max\left(\frac{25920(L_g^f)^2\|x_k - x_{k-1}\|^2 \log^2(1/\delta)}{\alpha_g \bar{\epsilon}_{g,1}^2}, \frac{6480\sigma_g^2 \log^2(1/\delta)}{\bar{\epsilon}_{g,2}}\right), \quad k \geq 1, \tag{11}$$

where $\bar{\epsilon}_g > 0$. Then with probability at least $1 - 2K\delta$ we have that

$$\|\epsilon_k^g\|^2 \leq \frac{3\bar{\epsilon}_{g,1}^2 + 2\alpha_g \bar{\epsilon}_{g,2}}{80}, \quad k = 0, \ldots, K-1.$$

Proof. See Appendix A. □

Remark 2 *Lemma 1 introduces three tunable parameters α_g, $\bar{\epsilon}_{g,1}$, and $\bar{\epsilon}_{g,2}$, which jointly control the accuracy of the gradient estimator and the resulting batch size B_k^g. The guiding principle for selecting these parameters is to minimize the batch size B_k^g while ensuring a prescribed estimation accuracy.*

To illustrate this principle, suppose that we aim to guarantee $\|\epsilon_k^g\| \leq \varepsilon/4$. Ignoring absolute constants, the error bound in Lemma 1 suggests that the two terms on the right-hand side should be of order ε^2. This motivates us to set $\bar{\epsilon}_{g,1} = \varepsilon$ and $\alpha_g \bar{\epsilon}_{g,2} = \varepsilon^2$. Then the dominant terms in the batch size B_k^g scales as

$$\frac{\|x_k - x_{k-1}\|^2}{\alpha_g \varepsilon^2} + \frac{1}{\bar{\epsilon}_{g,2}},$$

leading to the auxiliary optimization problem

$$\min_{\alpha_g, \bar{\epsilon}_{g,2}} \quad \sum_{k=1}^{K} \left(\frac{\|x_k - x_{k-1}\|^2}{\alpha_g \varepsilon^2} + \frac{1}{\bar{\epsilon}_{g,2}}\right) \quad \text{s.t.} \quad \alpha_g \bar{\epsilon}_{g,2} = \varepsilon^2.$$

This reveals that the optimal choice of α_g and $\bar{\epsilon}_{g,2}$ depends explicitly on the step size $\|x_k - x_{k-1}\|$. In the subsequent analysis, this dependence is handled by controlling $\|x_k - x_{k-1}\|$ through the trust-region mechanism. Accordingly, we directly specify parameter choices that satisfy the above principle and yield near-minimal batch sizes under the imposed step-size bounds.

Similarly, the following lemma characterizes the error bound of the stochastic Hessian estimator under appropriate batch-size choices.

Lemma 2 *Given $K \geq 1$ and $\delta \leq \frac{1}{2K}$, suppose that $\{B_k^h\}$ satisfies*

$$B_0^h = \frac{3240\sigma_h^2 \log^2(n/\delta)}{\bar{\epsilon}_{h,1}^2} \tag{12}$$

and

$$B_k^h = \max\left(\frac{12960(L_H^f)^2\|x_k - x_{k-1}\|^2 \log^2(n/\delta)}{\alpha_h \bar{\epsilon}_{h,1}^2}, \frac{3240\sigma_h^2 \log^2(n/\delta)}{\bar{\epsilon}_{h,2}}\right), \tag{13}$$

where $\bar{\epsilon}_h > 0$. Then with probability at least $1 - 2K\delta$ it holds that

$$\|\epsilon_k^h\|^2 \leq \frac{3\bar{\epsilon}_{h,1}^2 + 2\alpha_h \bar{\epsilon}_{h,2}}{40}, \quad k = 0, \ldots, K-1.$$

Proof. See Appendix B. □

Lemmas 1 and 2 provide high-probability uniform bounds on the gradient and Hessian estimation errors generated by the hybrid stochastic estimator. The batch sizes are chosen adaptively to balance two sources of error: the stochastic noise inherent in the oracle and the recursive bias introduced by the momentum correction, which depends on the displacement $\|x_k - x_{k-1}\|$. Such bounds are essential for ensuring that the cubic model constructed at each iteration remains sufficiently accurate to yield descent and negative curvature detection guarantees.

3 Carme for unconstrained optimization

In this section, we introduce the Curvature-oriented variance reduction method (Carme) for unconstrained stochastic optimization (1) with $X = \mathbb{R}^n$, i.e.,

$$\min_{x \in \mathbb{R}^n} f(x) = \mathbb{E}_\xi[F(x, \xi)], \tag{14}$$

with the goal of finding an (ϵ, γ)-SSP satisfying (3). Carme combines the hybrid stochastic gradient and Hessian estimator developed in Section 2 within a cubic-regularized framework. In addition, an explicit negative-curvature step is incorporated to enhance the ability to escape saddle points under stochastic oracles. In Subsection 3.1 we will present the algorithm framework of Carme with detailed complexity analysis provided in Subsection 3.2.

3.1 Algorithm framework

Carme proceeds iteratively and, at each iteration, primarily performs a cubic-regularized Newton step, while occasionally exploiting negative curvature through a randomized mechanism. At iteration k, to perform a cubic-regularized Newton step, a cubic-regularized subproblem with a trust-region constraint is solved:

$$\min_{\|s\| \leq \eta} m_k(s) = \langle g_k, s\rangle + \frac{1}{2}\langle s, H_k s\rangle + \frac{M}{6}\|s\|^3. \tag{15}$$

Here, g_k and H_k are variance-reduced estimates of the gradient $\nabla f(x_k)$ and Hessian $\nabla^2 f(x_k)$, defined in (6) and (7) respectively, $M > 0$ is a cubic regularization parameter, and $\eta > 0$ is the trust-region radius.

The integration of the trust-region scheme with cubic regularization is well-established as a globalization mechanism [1]. In this work, the trust-region radius η is also used to control the adaptive batch sizes in (6)–(7), since $\|x_k - x_{k-1}\| = \|s_{k-1}\| \leq \eta$ enters their definitions. This design allows us to control the variance of the stochastic estimators without inflating the per-iteration sampling cost, which is crucial for the stability and complexity guarantees of the proposed algorithm.

Solving the cubic-regularized subproblem (15) to global optimality can be computationally expensive, especially when the Hessian approximation H_k is indefinite. Rather than requiring an exact solution, we allow for inexact solutions that satisfy mild descent and approximate stationarity conditions. Such inexactness is standard in the analysis of cubic regularization methods and is sufficient to ensure the desired theoretical guarantees.

Specifically, we impose the following conditions on the inexact solution of (15):

Condition A *An inexact solution s_k of subproblem (15) satisfies the following conditions:*

$$m_k(s_k) \leq 0, \tag{16a}$$

$$\|\nabla m_k(s_k)\| \leq M\omega^{2/3}, \text{ when } \|s_k\| < \eta, \tag{16b}$$

where the parameter $\omega > 0$ controls the solution accuracy.

Such an inexact solution can be efficiently obtained, for instance, by a truncated gradient descent scheme applied to (15); see Appendix C for details. Importantly, the procedure for solving the cubic subproblem only accesses the stochastic estimates g_k and H_k that are already computed, and therefore does not incur any additional stochastic calls.

To further enhance Carme's ability to achieve (ϵ, γ)-SSPs with arbitrary γ, as defined in Definition 1, we also incorporate the negative curvature search technique, as it enables efficient escape from saddle points by exploiting negative curvature directions. While cubic regularization implicitly incorporates curvature information, explicitly incorporating a negative-curvature step is essential in the stochastic and inexact setting to reliably certify (ϵ, γ)-second-order stationarity without incurring excessive computational cost. The complete algorithmic framework of Carme is presented in Algorithm 1.

3.2 Complexity analysis

In this subsection, we analyze the iteration and sample complexity of the proposed method for computing an (ϵ, γ)-SSP. We begin by establishing the descent properties of the cubic-regularized Newton step.

Lemma 3 *Let s_k be an inexact solution of (15) satisfying Condition A. Suppose Assumptions 1 and 2 hold, then for any $M \geq 4L_H^f$ and $\eta > 0$, the point $x_k + s_k$ satisfies*

$$f(x_k) - f(x_k + s_k) \geq \frac{M}{12}\|s_k\|^3 - \frac{8}{\sqrt{M}}\|\epsilon_k^g\|^{3/2} - \frac{4\eta^{3/2}}{\sqrt{M}}\|\epsilon_k^h\|^{3/2}. \tag{17}$$

Proof. Since $\nabla^2 f$ is L_H^f-Lipschitz continuous, Taylor's theorem gives

$$\begin{aligned} f(x_k + s_k) - f(x_k) &\leq \langle \nabla f(x_k), s_k \rangle + \frac{1}{2}\langle s_k, \nabla^2 f(x_k) s_k \rangle + \frac{L_H^f}{6}\|s_k\|^3 \\ &= m_k(s_k) + \frac{L_H^f - M}{6}\|s_k\|^3 + \langle \nabla f(x_k) - g_k, s_k \rangle + \frac{1}{2}\langle s_k, (\nabla^2 f(x_k) - H_k) s_k \rangle. \end{aligned}$$

Using $m_k(s_k) \leq 0$ and $M \geq 4L_H^f$ (so $(L_H^f - M)/6 \leq -M/8$), we obtain

$$f(x_k + s_k) - f(x_k) \leq -\frac{M}{8}\|s_k\|^3 + \|\nabla f(x_k) - g_k\|\|s_k\| + \frac{1}{2}\|(\nabla^2 f(x_k) - H_k)s_k\|\|s_k\|. \tag{18}$$

Algorithm 1: Carme

Input: Initial point $x_0 \in \mathbb{R}^n$, parameters $\epsilon, \gamma, \omega, \alpha_g, \alpha_h, \bar{\epsilon}_{g,1}, \bar{\epsilon}_{g,2}, \bar{\epsilon}_{h,1}, \bar{\epsilon}_{h,1}, \eta, M, p > 0$ and $K > 0$.

for $k = 0, \ldots, K$ **do**

Compute g_k and H_k through (6) and (7).

Sample $Q_k \sim$ Bernoulli(p) with bias p.

if $Q_k = 1$; `// cubic-regularized Newton step`

then

Solve subproblem

$$\min_{\|s\| \le \eta} f(x_k) + \langle g_k, s \rangle + \frac{1}{2}\langle s, H_k s \rangle + \frac{M}{6}\|s\|^3$$

obtaining an inexact solution s_k satisfying Condition A.

Set $x_{k+1} = x_k + s_k$.

else

if $\lambda_{\min}(H_k) \le -4\gamma$; `// negative-curvature step`

then

Find a unit vector u_k such that $u_k^\top H_k u_k \le -2\gamma$.

Set $x_{k+1} = x_k + \frac{\gamma}{L_H^f} \cdot r_k \cdot u_k$, where $r_k \sim \text{Uniform}(\{-1, 1\})$.

else

Set $x_{k+1} = x_k$;

Additionally, applying the scaled Young's inequality $ab \le \frac{8}{\sqrt{M}}a^{3/2} + \frac{M}{64}b^3$ yields

$$\|\nabla f(x_k) - g_k\|\|s_k\| \le \frac{8}{\sqrt{M}}\|\nabla f(x_k) - g_k\|^{3/2} + \frac{M}{64}\|s_k\|^3,$$
$$\|(\nabla^2 f(x_k) - H_k)s_k\|\|s_k\| \le \frac{8}{\sqrt{M}}\|(\nabla^2 f(x_k) - H_k)s_k\|^{3/2} + \frac{M}{64}\|s_k\|^3.$$

Then plugging these bounds into (18), we obtain

$$\begin{aligned} f(x_k + s_k) - f(x_k) &\le -\frac{M}{12}\|s_k\|^3 + \frac{8}{\sqrt{M}}\|\nabla f(x_k) - g_k\|^{3/2} + \frac{4}{\sqrt{M}}\|(\nabla^2 f(x_k) - H_k)s_k\|^{3/2} \\ &\le -\frac{M}{12}\|s_k\|^3 + \frac{8}{\sqrt{M}}\|\nabla f(x_k) - g_k\|^{3/2} + \frac{4}{\sqrt{M}}\|\nabla^2 f(x_k) - H_k\|^{3/2} \cdot \eta^{3/2}. \end{aligned}$$

Rearranging the terms yields the conclusion. □

Lemma 3 shows that the objective decrease along the direction s_k is cubic in its length, up to additive terms induced by the gradient and Hessian estimation errors. Next, the following lemma links the post-step gradient norm $\|\nabla f(x_k + s_k)\|$ to the step length $\|s_k\|$ and the estimator errors, which will be used to translate step-based descent into first-order stationarity measure.

Lemma 4 *Under the same conditions as Lemma 3, it holds that*

$$\mathbf{1}\left\{\|\nabla f(x_k + s_k)\| \ge \frac{M\eta^2}{2}\right\} \le \frac{2}{\eta^2}\|s_k\|^2 + \frac{2}{M\eta^2}\left(\|\epsilon_k^g\| + \eta\|\epsilon_k^h\| + M\omega^{2/3}\right). \tag{19}$$

Proof. First, one of the following two cases must occur: either $\|s_k\| = \eta$, or $\|s_k\| < \eta$.

Case 1: $\|s_k\| = \eta$**.** In this case, we have $\frac{2}{\eta^2}\|s_k\|^2 = 2$. Since the indicator function on the left-hand side of (19) is at most one, the inequality holds trivially.

Case 2: $\|s_k\| < \eta$. Using the L_H^f-Lipschitz continuity of $\nabla^2 f$, we obtain

$$\begin{aligned}\|\nabla f(x_k + s_k)\| &\le \|\nabla f(x_k + s_k) - \nabla f(x_k) - \nabla^2 f(x_k)s_k\| + \|\nabla f(x_k) + \nabla^2 f(x_k)s_k\| \\ &\le \frac{L_H^f}{2}\|s_k\|^2 + \|\nabla f(x_k) - g_k\| + \|(\nabla^2 f(x_k) - H_k)s_k\| + \|g_k + H_k s_k\|.\end{aligned}$$

Since $\|s_k\| < \eta$ and s_k satisfies Condition A, we further have

$$\|(\nabla^2 f(x_k) - H_k)s_k\| \le \eta\|\nabla^2 f(x_k) - H_k\|, \qquad \|g_k + H_k s_k\| \le \frac{M}{2}\|s_k\|^2 + M\omega^{2/3}.$$

Combining the above inequalities yields

$$\|\nabla f(x_k + s_k)\| \le \frac{L_H^f + M}{2}\|s_k\|^2 + \|\nabla f(x_k) - g_k\| + \eta\|\nabla^2 f(x_k) - H_k\| + M\omega^{2/3}. \tag{20}$$

Suppose that $\|\nabla f(x_k + s_k)\| \ge \frac{M\eta^2}{2}$. Since $M \ge L_H^f$, inequality (20) implies

$$\frac{M\eta^2}{2} \le M\|s_k\|^2 + \|\nabla f(x_k) - g_k\| + \eta\|\nabla^2 f(x_k) - H_k\| + M\omega^{2/3}.$$

Dividing both sides by $\frac{M\eta^2}{2}$ gives

$$\mathbf{1}\left\{\|\nabla f(x_k + s_k)\| \ge \frac{M\eta^2}{2}\right\} \le \frac{2}{\eta^2}\|s_k\|^2 + \frac{2}{M\eta^2}\left(\|\nabla f(x_k) - g_k\| + \eta\|\nabla^2 f(x_k) - H_k\| + M\omega^{2/3}\right).$$

Finally, recalling that $\|\nabla f(x_k) - g_k\| = \|\epsilon_k^g\|$ and $\|\nabla^2 f(x_k) - H_k\| = \|\epsilon_k^h\|$ by (9), we obtain (19). □

Combining Lemmas 3 and 4, we can relate the expected decrease of the objective function to the first-order stationarity, which forms the basis of the iteration and sample complexity analysis.

Lemma 5 *Under the same conditions as Lemma 3, then for k with $Q_k = 1$, it holds that*

$$\mathbb{E}\left[f(x_k) - f(x_{k+1})\right] \ge \frac{M\eta^3}{72}\Pr\left\{\|\nabla f(x_{k+1})\| \ge \frac{M\eta^2}{2}\right\} - \frac{9}{\sqrt{M}}\mathbb{E}\left[\|\epsilon_k^g\|^{3/2}\right] - \frac{5\eta^2}{\sqrt{M}}\mathbb{E}\left[\|\epsilon_k^h\|^{3/2}\right] - \frac{M}{12}\omega,$$

where $\Pr(\cdot)$ *and* $\mathbb{E}[\cdot]$ *are taken w.r.t. the randomness over* g_k *and* H_k.

Proof. From Lemma 3, taking expectations on both sides of (17) yields

$$\mathbb{E}[f(x_k) - f(x_{k+1})] \ge \frac{M}{12}\mathbb{E}\left[\|s_k\|^3\right] - \frac{8}{\sqrt{M}}\mathbb{E}\left[\|\epsilon_k^g\|^{3/2}\right] - \frac{4}{\sqrt{M}}\mathbb{E}\left[(\eta\|\epsilon_k^h\|)^{3/2}\right], \tag{21}$$

where the second inequality follows from Jensen's inequality. Next, from Lemma 4, since the indicator function takes values in $\{0, 1\}$, raising both sides of (19) to the power $3/2$ preserves the inequality. Using Jensen's inequality $(\sum_{i=1}^n a_i)^q \le n^{q-1}\sum_{i=1}^n a_i^q$ for $q = 3/2$, we obtain

$$\begin{aligned}\mathbf{1}\left\{\|\nabla f(x_{k+1})\| \ge \frac{M\eta^2}{2}\right\} &\le \left(\frac{2}{\eta^2}\|s_k\|^2 + \frac{2}{M\eta^2}\left(\|\epsilon_k^g\| + \eta\|\epsilon_k^h\| + M\omega^{2/3}\right)\right)^{3/2} \\ &\le \frac{6}{\eta^3}\|s_k\|^3 + \frac{6}{M^{3/2}\eta^3}\left(\|\epsilon_k^g\|^{3/2} + (\eta\|\epsilon_k^h\|)^{3/2} + M^{3/2}\omega\right).\end{aligned}$$

Taking expectations and rearranging terms yield

$$\mathbb{E}\left[\|s_k\|^3\right] \ge \frac{\eta^3}{6}\Pr\left\{\|\nabla f(x_{k+1})\| \ge \frac{M\eta^2}{2}\right\} - M^{-3/2}\mathbb{E}\left[\|\epsilon_k^g\|^{3/2} + (\eta\|\epsilon_k^h\|)^{3/2}\right] - \omega. \tag{22}$$

Substituting (22) into (21) yields the desired result, where numerical constants are obtained by collecting and simplifying the bounds. □

Lemma 5 establishes a fundamental link between the expected objective decrease of a cubic-regularized Newton step and the probability of violating first-order stationarity. Specifically, it shows that whenever the gradient norm at the next iterate remains above a threshold of order $\frac{M\eta^2}{2}$, the algorithm guarantees an expected decrease in the objective value, up to additive terms induced by the gradient and Hessian estimation errors and the inexactness parameter ω.

Next, we characterize the expected function decrease produced by the negative-curvature step. In particular, we show that whenever the estimated Hessian detects sufficient negative curvature, this step yields a decrease of order γ^3, up to a bias term related to the Hessian estimation error.

Lemma 6 *Suppose Assumptions 1 and 2 hold, then for k with $Q_k = 0$, it holds that*

$$\mathbb{E}\left[f(x_k) - f(x_{k+1})\right] \geq \frac{5\gamma^3}{6(L_H^f)^2}\Pr\left\{\lambda_{\min}(H_k) \leq -4\gamma\right\} - \frac{\gamma^2}{2(L_H^f)^2}\mathbb{E}\left[\|\epsilon_k^h\|\right], \tag{23}$$

where $\Pr(\cdot)$ and $\mathbb{E}[\cdot]$ are taken w.r.t. the randomness in H_k and r_k.

Proof. If $\lambda_{\min}(H_k) > -4\gamma$, then $x_{k+1} = x_k$ and hence $f(x_{k+1}) = f(x_k)$. If $\lambda_{\min}(H_k) \leq -4\gamma$, we choose a unit vector u_k satisfying $u_k^\top H_k u_k \leq -2\gamma$, and set $\tilde{s}_k := \frac{\gamma}{L_H^f} r_k u_k$. Using the L_H^f-Lipschitz continuity of $\nabla^2 f$, we have

$$f(x_{k+1}) \leq f(x_k) + \langle \nabla f(x_k), \tilde{s}_k\rangle + \frac{1}{2}\tilde{s}_k^\top \nabla^2 f(x_k)\tilde{s}_k + \frac{L_H^f}{6}\|\tilde{s}_k\|^3.$$

Conditioning on H_k (and hence on u_k), $\mathbb{E}[r_k] = 0$ and $\mathbb{E}[r_k^2] = 1$ imply

$$\mathbb{E}_{r_k}\left[\langle \nabla f(x_k), \tilde{s}_k\rangle + \tfrac{1}{2}\tilde{s}_k^\top \nabla^2 f(x_k)\tilde{s}_k \mid H_k\right] = \frac{\gamma^2}{2(L_H^f)^2} u_k^\top \nabla^2 f(x_k) u_k$$

$$= \frac{\gamma^2}{2(L_H^f)^2}\Big(u_k^\top H_k u_k + u_k^\top(\nabla^2 f(x_k) - H_k)u_k\Big) \leq -\frac{\gamma^3}{(L_H^f)^2} + \frac{\gamma^2}{2(L_H^f)^2}\|\nabla^2 f(x_k) - H_k\|.$$

Since $\|\tilde{s}_k\| = \frac{\gamma}{L_H^f}$, we obtain

$$\mathbb{E}_{r_k}[f(x_{k+1}) \mid H_k] \leq f(x_k) - \frac{5\gamma^3}{6(L_H^f)^2} + \frac{\gamma^2}{2(L_H^f)^2}\|\nabla^2 f(x_k) - H_k\|$$

on the event $\{\lambda_{\min}(H_k) \leq -4\gamma\}$. Therefore, conditioning on H_k we have

$$\mathbb{E}_{r_k}[f(x_{k+1}) \mid H_k] \leq f(x_k) - \frac{5\gamma^3}{6(L_H^f)^2}\,\mathbf{1}\{\lambda_{\min}(H_k) \leq -4\gamma\} + \frac{\gamma^2}{2(L_H^f)^2}\|\nabla^2 f(x_k) - H_k\|.$$

Taking expectation over H_k and rearranging yield (23). □

Lemma 6 shows that, whenever the negative-curvature branch is selected, the randomized step yields an expected decrease of order $\mathcal{O}(\gamma^3)$ if $\lambda_{\min} \leq -4\gamma$, with the only degradation relative to the deterministic setting arises from the Hessian estimation error $\mathbb{E}[\|\epsilon_k^h\|]$.

We are now ready to state the main complexity result of Carme for unconstrained stochastic optimization. The following theorem shows that, under appropriate parameter choices, Carme finds an (ϵ, γ)-SSP with reasonably high probability, while achieving improved sample complexity compared with existing stochastic cubic regularization methods. First, given the target accuracies $\epsilon > 0$ and $\gamma > 0$, and define the auxiliary tolerance $\varepsilon := \frac{\epsilon}{450}$, we specify the parameter settings as follows.

Algorithmic parameters.

$$\Delta = f(x_0) + C, \qquad M = 4L_H^f, \qquad \eta = 30M^{-1/2}\varepsilon^{1/2},$$
$$\omega = \min\Big\{M^{-1}, \frac{3\varepsilon^{3/2}}{4M^{3/2}}\Big\}, \qquad p = \frac{4\sqrt{M}\,\gamma^3}{4\sqrt{M}\,\gamma^3 + 135M^2\varepsilon^{3/2}}, \qquad K = \left\lceil \frac{90\Delta M^2}{\gamma^3} + \frac{16\Delta\sqrt{M}}{5\varepsilon^{3/2}} \right\rceil.$$

Estimator parameters.

$$\begin{aligned} &\delta = \frac{1}{200K}, \quad \bar{\epsilon}_{g,1} = \varepsilon, \quad \bar{\epsilon}_{g,2} = \left(\frac{\varepsilon^{-3/2}+\gamma^{-3}}{\varepsilon^{-1/2}+\gamma^{-1}}\right)^{1/2}\bar{\epsilon}_{g,1}^2, \quad \alpha_g = \frac{\bar{\epsilon}_{g,1}^2}{\bar{\epsilon}_{g,2}}, \\ &\bar{\epsilon}_{h,1} = \frac{1}{10}\min\Big\{\frac{M^{1/2}\varepsilon^{1/2}}{4}, \frac{\gamma}{10}\Big\}, \quad \bar{\epsilon}_{h,2} = \left(\frac{\varepsilon^{-3/2}+\gamma^{-3}}{\varepsilon^{-1/2}+\gamma^{-1}}\right)^{1/2}\bar{\epsilon}_{h,1}^2, \quad \alpha_h = \frac{\bar{\epsilon}_{h,1}^2}{\bar{\epsilon}_{h,2}}. \end{aligned} \tag{24}$$

And B_k^g and B_k^h can be chosen as in Lemmas 1 and 2. Note that the above constants are chosen conservatively to streamline the subsequent probability bookkeeping and to keep the final stationarity conditions in a clean form. This choice does not affect the oracle complexity, but only the absolute numerical factors. Under the above parameters setting, we have the results below.

Theorem 1 *Under Assumptions 1-4, given $\epsilon \in (0,1)$ and $\gamma \in \big(0, (15\Delta M^2)^{1/3}\big)$, the following statements hold true with probability at least 0.96:*

(i) Carme returns a point x_R such that

$$\|\nabla f(x_R)\| \le \epsilon, \text{ and } \lambda_{\min}(\nabla^2 f(x_R)) \ge -5\gamma, \tag{25}$$

within $K = O(\Delta(L_H^f)^2\gamma^{-3} + \Delta(L_H^f)^{1/2}\epsilon^{-3/2})$ iterations.

(ii) To reach an (ϵ,γ)-SSP of (14), Carme requires at most $\tilde{O}(\epsilon^{-3} + \epsilon^{-2}\gamma^{-2} + \gamma^{-3})$ stochastic gradient queries in expectation and $\tilde{O}(\epsilon^{-2} + \gamma^{-4})$ stochastic hessian queries.

Proof. We divide the proof into two parts. We first establish the high-probability second-order stationarity guarantee, and then bound the total sample complexity.

Part I: Second-order stationarity. We begin by controlling the stochastic estimation errors. By Lemma 1 with $\alpha_g = \bar{\epsilon}_g^{1/2} = \varepsilon^{1/2}$, it holds with probability at least $1 - 2K\delta = 0.99$ that

$$\|\nabla f(x_k) - g_k\| \le \frac{\varepsilon}{4}, \qquad k = 0, \dots, K-1. \tag{26}$$

Similarly, by Lemma 2 with $\alpha_h = \bar{\epsilon}_h^{1/2} = \frac{1}{10}\min\{\frac{M^{1/2}\varepsilon^{1/2}}{4}, \frac{\gamma}{10}\}$, with probability at least $1 - 2K\delta = 0.99$,

$$\|\nabla^2 f(x_k) - H_k\|^2 \le \frac{1}{800}\min\left\{\frac{M\varepsilon}{16}, \frac{\gamma^2}{100}\right\}, \qquad k = 0, \dots, K-1. \tag{27}$$

By a union bound, the above two bounds hold simultaneously for all $k = 0, \dots, K-1$with probability at least $1 - 4K\delta = 0.98$. In the sequel, we work on this event. At each iteration, Carme either performs a cubic-regularized step ($Q_k = 1$) or a negative-curvature step ($Q_k = 0$). We analyze the two cases separately.

Case 1: $Q_k = 1$**.** Applying Lemma 5 together with (26), (27) and the choice $\eta = 30M^{-1/2}\varepsilon^{1/2}$ yields

$$\begin{aligned} \mathbb{E}\left[f(x_k) - f(x_{k+1}) \mid Q_k = 1\right] &\ge \frac{375\varepsilon^{3/2}}{\sqrt{M}}\Pr\left\{\|\nabla f(x_{k+1})\| \ge 450\varepsilon\right\} - \frac{9\varepsilon^{3/2}}{8\sqrt{M}} - \frac{11\varepsilon^{3/2}}{16\sqrt{M}} - \frac{\varepsilon^{3/2}}{16\sqrt{M}} \\ &= \frac{375\varepsilon^{3/2}}{\sqrt{M}}\left(\Pr\left\{\|\nabla f(x_{k+1})\| \ge 450\varepsilon\right\} - \frac{1}{200}\right). \end{aligned} \tag{28}$$

Case 2: $Q_k = 0$. By Lemma 6 and (27), we have $\|\epsilon_k^h\|^2 \le \gamma^2/(80000)$, and hence

$$\begin{aligned}\mathbb{E}\left[f(x_k) - f(x_{k+1}) \mid Q_k = 0\right] &\ge \frac{5\gamma^3}{6(L_H^f)^2}\Pr\left\{\lambda_{\min}(H_k) \le -4\gamma\right\} - \frac{\gamma^3}{400\sqrt{2}(L_H^f)^2} \\ &\ge \frac{5\gamma^3}{6(L_H^f)^2}\left(\Pr\left\{\lambda_{\min}(H_k) \le -4\gamma\right\} - \frac{1}{400}\right).\end{aligned} \tag{29}$$

Combining (28) and (29), taking expectation over Q_k, we obtain

$$\begin{aligned}&\mathbb{E}\left[f(x_k) - f(x_{k+1})\right] = \sum_{q\in\{0,1\}} \Pr(Q_t = q)\mathbb{E}\left[f(x_k) - f(x_{k+1}) \mid Q_t = q\right] \\ &\ge (1-p)\cdot\frac{5\gamma^3}{6(L_H^f)^2}\left(\Pr\left\{\lambda_{\min}(H_k) \le -4\gamma\right\} - \frac{1}{400}\right) + p\cdot\frac{375\varepsilon^{3/2}}{\sqrt{M}}\left(\Pr\left\{\|\nabla f(x_{k+1})\| \ge 450\varepsilon\right\} - \frac{1}{200}\right).\end{aligned}$$

Let x_R be generated randomly from $\{x_k\}_{k=0}^{K-1}$. Telescoping the inequality above for k from 0 to $K-1$, and using the bound $\mathbb{E}\left[f(x_0) - f(x_k)\right] \le \Delta$, we obtain

$$\begin{aligned}\Delta &\ge \mathbb{E}\left[f(x_0) - f(x_K)\right] \\ &\ge \frac{5(1-p)\gamma^3}{6(L_H^f)^2}\sum_{k=0}^{K-1}\left(\Pr\{\lambda_{\min}(H_k) \le -4\gamma\} - \frac{1}{400}\right) + \frac{375p\varepsilon^{3/2}}{\sqrt{M}}\sum_{k=1}^{K}\left(\Pr\left\{\|\nabla f(x_k)\| \ge 450\varepsilon\right\} - \frac{1}{200}\right) \\ &\ge 1200\Delta\left(\frac{1}{K}\sum_{k=0}^{K-1}\Pr\{\lambda_{\min}(H_k) \le -4\gamma\} + \frac{1}{K}\sum_{k=1}^{K}\Pr\left\{\|\nabla f(x_k)\| \ge 450\varepsilon\right\} - \frac{3}{400}\right) \\ &\ge 1200\Delta\left(\frac{5}{6(K-1)}\sum_{k=1}^{K-1}\left(\Pr\{\lambda_{\min}(H_k) \le -4\gamma\} + \Pr\left\{\|\nabla f(x_k)\| \ge 450\varepsilon\right\}\right) - \frac{3}{400}\right) \\ &\ge 1200\Delta\left(\frac{5}{6}\left(\Pr\{\lambda_{\min}(H_R) \le -4\gamma\} + \Pr\left\{\|\nabla f(x_R)\| \ge 450\varepsilon\right\}\right) - \frac{3}{400}\right),\end{aligned}$$

where the third inequality follows from Lemma 17, the fourth inequality given by ignoring some (non-negative) terms on the right-hand side and using the fact that $K \ge 6$. Since $\varepsilon = \frac{\epsilon}{450}$, then rearranging the terms yields $\Pr\{\lambda_{\min}(H_R) \le -4\gamma\} + \Pr\left\{\|\nabla f(x_R)\| \ge \epsilon\right\} \le 0.01$, which further implies that the returned point x_R satisfies

$$\Pr\left\{(\lambda_{\min}(H_R) > -4\gamma) \wedge (\|\nabla f(x_R)\| < \epsilon)\right\} \ge 0.99. \tag{30}$$

Besides, we know that $\|\nabla^2 f(x_R) - H_R^0\| \le \frac{\gamma}{200\sqrt{2}}$ from (27). Hence, from $\lambda_{\min}(H_R) \ge -4\gamma$, one can derive

$$\lambda_{\min}(\nabla^2 f(x_R)) \ge -5\gamma.$$

Combining with the event of (26) and (27) holding (probability ≥ 0.98) yields overall success probability at least $0.99 \times 0.98 \ge 0.97$.

Iteration complexity. By construction, the total number of iterations is

$$K = \left\lceil \frac{1440\,\Delta(L_H^f)^2}{\gamma^3} + \frac{16\Delta\sqrt{M}}{5\,\varepsilon^{3/2}} \right\rceil = O\Big(\Delta(L_H^f)^2\gamma^{-3} + \Delta(L_H^f)^{1/2}\epsilon^{-3/2}\Big), \tag{31}$$

where we recall that $\varepsilon = \epsilon/450$ and $M = 4L_H^f$.

Part II: Sample complexity. We now bound the total number of stochastic oracle queries required to reach a point satisfying (25). We first estimate the sample complexity per iteration and then aggregate over all iterations. At each iteration $k \geq 1$, Carme either performs a cubic-regularized step ($Q_{k-1} = 1$) or a negative-curvature step ($Q_{k-1} = 0$). We analyze these two cases separately.

Case 1: $Q_{k-1} = 1$**.** In this case, the trust-region radius is $\eta = O(M^{-1/2}\varepsilon^{1/2})$, which implies $\|x_k - x_{k-1}\|^2 \leq \eta^2 = O\left(\frac{\varepsilon}{L_H^f}\right)$. By Lemma 1 with $\alpha_g \bar{\epsilon}_{g,2} = \bar{\epsilon}_{g,1}^2$ and $\delta = 1/(200K)$, the batch size of stochastic gradients satisfies

$$B_k^g = \max\left(\frac{25920(L_g^f)^2\|x_k - x_{k-1}\|_2^2 \log^2(1/\delta)}{\alpha_g \bar{\epsilon}_{g,1}^2}, \frac{6480\sigma_g^2 \log^2(1/\delta)}{\bar{\epsilon}_{g,2}}\right) = \tilde{O}\left(\frac{\bar{\epsilon}_{g,2}\varepsilon}{\bar{\epsilon}_{g,1}^4} + \frac{1}{\bar{\epsilon}_{g,2}}\right).$$

Similarly, by Lemma 2 with $\alpha_h \bar{\epsilon}_{h,2} = \bar{\epsilon}_{h,1}^2$ and $\delta = 1/(200K)$, the batch size of stochastic Hessians satisfies

$$B_k^h = \max\left(\frac{12960(L_H^f)^2\|x_k - x_{k-1}\|_2^2 \log^2(n/\delta)}{\alpha_h \bar{\epsilon}_{h,1}^2}, \frac{3240\sigma_h^2 \log^2(n/\delta)}{\bar{\epsilon}_{h,2}}\right) = \tilde{O}\left(\frac{\bar{\epsilon}_{h,2}\varepsilon}{\bar{\epsilon}_{h,1}^4} + \frac{1}{\bar{\epsilon}_{h,2}}\right).$$

Case 2: $Q_{k-1} = 0$**.** From the update rule of negative-curvature step in Algorithm 1, we have $\|x_k - x_{k-1}\|^2 = O\left(\frac{\gamma^2}{(L_H^f)^2}\right)$. Consequently, the gradient and Hessian batch sizes satisfy

$$B_k^g = \tilde{O}\left(\frac{\bar{\epsilon}_{g,2}\gamma^2}{\bar{\epsilon}_{g,1}^4} + \frac{1}{\bar{\epsilon}_{g,2}}\right), \quad B_k^h = \tilde{O}\left(\frac{\bar{\epsilon}_{h,2}\gamma^2}{\bar{\epsilon}_{h,1}^4} + \frac{1}{\bar{\epsilon}_{h,2}}\right).$$

Total sample complexity. Let $\mathcal{K}_1 = \{k : Q_{k-1} = 1\}$ and $\mathcal{K}_0 = \{k : Q_{k-1} = 0\}$. By Lemma 17, we have $\mathbb{E}[|\mathcal{K}_1|] = O(Kp) = O(\varepsilon^{-3/2})$ and $\mathbb{E}[|\mathcal{K}_0|] = O(K(1-p)) = O(\gamma^{-3})$. Therefore, the in-expectation total number of stochastic gradient evaluations is

$$\begin{aligned}
&\mathbb{E}\left[\sum_{k=0}^{K} B_k^g\right] = B_0^g + \mathbb{E}\left[\sum_{k\in\mathcal{K}_1} B_k^g\right] + \mathbb{E}\left[\sum_{k\in\mathcal{K}_0} B_k^g\right] \\
&= \tilde{O}\left(\varepsilon^{-2} + \left(\varepsilon^{-3/2} + \gamma^{-3}\right)^{1/2}\left(\varepsilon^{-1/2} + \gamma^{-1}\right)^{1/2}\varepsilon^{-2} + \gamma^{-3}\right) = \tilde{O}\left(\epsilon^{-3} + \epsilon^{-2}\gamma^{-2} + \gamma^{-3}\right),
\end{aligned}$$

where γ^{-3} comes from the fact that $B_k^g \geq 1$ and the last inequality uses Young's inequality $ab \leq \frac{a^p}{p} + \frac{b^q}{q}$, where $\frac{1}{p} + \frac{1}{q} = 1$. For the stochastic Hessian oracles, the in-expectation total number is

$$\begin{aligned}
&\mathbb{E}\left[\sum_{k=0}^{K} B_k^h\right] = B_0^h + \mathbb{E}\left[\sum_{k\in\mathcal{K}_1} B_k^h\right] + \mathbb{E}\left[\sum_{k\in\mathcal{K}_0} B_k^h\right] \\
&= \tilde{O}\left(\left(1 + \left(\varepsilon^{-3/2} + \gamma^{-3}\right)^{1/2}\left(\varepsilon^{-1/2} + \gamma^{-1}\right)^{1/2}\right)\left(\varepsilon^{-1} + \gamma^{-2}\right) + \gamma^{-3}\right) = \tilde{O}\left(\epsilon^{-2} + \gamma^{-4}\right),
\end{aligned}$$

where $\varepsilon^{-1} + \gamma^{-2}$ comes from the fact that $\bar{\epsilon}_{h,1}^{-2} = (\frac{1}{10}\min\{\frac{M^{1/2}\varepsilon^{1/2}}{4}, \frac{\gamma}{10}\})^{-2} = O(\varepsilon^{-1} + \gamma^{-2})$. Then, let $T_g \triangleq \sum_{k=0}^{K} B_k^g$, $T_h \triangleq \sum_{k=0}^{K} B_k^h$. By Markov's inequality, for any $\delta' \in (0,1)$, we have $\Pr\big(T_g \geq \frac{1}{\delta'}\mathbb{E}[T_g]\big) \leq \delta'$, $\Pr\big(T_h \geq \frac{1}{\delta'}\mathbb{E}[T_h]\big) \leq \delta'$. Hence, if we set $\delta' = 0.01$, then it holds with probability at least 0.99 that

$$T_g \leq 100 \cdot \mathbb{E}[T_g] = \tilde{O}\big(\epsilon^{-3} + \epsilon^{-2}\gamma^{-2} + \gamma^{-3}\big), \qquad T_h \leq 100 \cdot \mathbb{E}[T_h] = \tilde{O}\big(\epsilon^{-2} + \gamma^{-4}\big).$$

This completes the proof with the union probability bound $0.99 \times 0.97 \geq 0.96$. □

Remark 3 *We note that the two terms in (31) correspond respectively to the lower bounds on iteration complexity for deterministic second-order optimization algorithms: $\Delta(L_H^f)^{1/2}\epsilon^{-3/2}$ for finding first-order stationary points (see [8, Theorem 2]) and $\Delta(L_H^f)^2\gamma^{-3}$ for second-order stationary points (see [1, Theorem 6]). For the stochastic case, the sample complexity's lower bound for second-order methods to find second-order stationary points, without Assumption 3, is $\Omega(\epsilon^{-3}+\gamma^{-5})$ in terms of stochastic second-order oracle queries (see [1, (14)]). However, under Assumption 3, no such lower bound has been established yet. We compared the complexity results of several cubic-regularized algorithms in Table 1. As can be seen, the development of cubic regularized algorithms has shifted from deterministic problems to stochastic problems in the finite-sum form and further to those in the expectation form. This work complements the existing literature by providing an oracle-complexity analysis for expectation-form stochastic cubic algorithms under the sample-wise smooth condition (Assumption 3). In addition, we study stationary points with separated first- and second-order accuracies to help identify clearer impact of the accuracies on complexity orders.*

To summarize, in this section we propose a stochastic second-order method, referred to as the Curvature-oriented variance reduction method (Carme), which computes an approximate second-order stationary point for unconstrained optimization. The key idea is to integrate variance reduction directly into the curvature exploitation mechanism, so that the stochastic cubic subproblems admit sufficiently accurate descent directions without requiring giant stochastic gradient or Hessian samples. As a result, under smoothness assumptions of stochastic functions, Carme achieves a strictly improved sample complexity over the state-of-the-art stochastic second-order algorithms.

4 Carme-ALM for equality-constrained optimization

In this section, we extend the curvature-oriented variance reduction method (Carme) to equality-constrained stochastic optimization (1)-(2), i.e.

$$\min_{x\in\mathbb{R}^n} \; f(x)=\mathbb{E}_\xi[F(x,\xi)] \quad \text{subject to} \quad c(x)=0, \tag{32}$$

where $c(x)=(c_1(x),\dots,c_m(x))^\top$. Our goal is to compute an (ϵ,γ)-SSP of (32), in the sense of Definition 1, which simultaneously achieves approximate first-order stationarity, second-order curvature conditions along feasible directions, and approximate feasibility.

To solve (32), a direct way is to apply Carme to the penalized problem via quadratic penalty function:

$$\min_{x\in\mathbb{R}^n} \phi(x;\rho) \;:=\; f(x)+\frac{\rho}{2}\|c(x)\|^2, \tag{33}$$

where $\rho>0$ is a penalty parameter. However, such a direct quadratic-penalty-based extension inevitably suffers from a deterioration in complexity relative to the unconstrained case. The main reason lies in the necessity of using a large penalty parameter $\rho=\Theta(\epsilon^{-1})$ to guarantee approximate feasibility, which in turn leads to a significant increase in the Hessian Lipschitz constant L_H^ϕ of the penalized objective. This coupling between feasibility enforcement and curvature growth substantially deteriorates the iteration complexity. Indeed, the iteration complexity of Carme scales polynomially with L_H^ϕ, namely as

$$O\big(\Delta(L_H^\phi)^{1/2}\epsilon^{-3/2}+\Delta(L_H^\phi)^2\gamma^{-3}\big)=O(\epsilon^{-2}\gamma^{-3}).$$

This phenomenon reveals an intrinsic limitation of quadratic-penalty-based extensions in stochastic cubic regularization. Even though the underlying algorithmic structure of Carme remains unchanged, the geometry of the optimization landscape is fundamentally altered by the penalty term once high-accuracy

feasibility is required. As a result, quadratic penalty methods are unable to preserve the favorable second-order behavior of Carme without incurring a substantial additional complexity cost. We cannot help to ask whether there exists an alternative penalty function that ensures feasibility using a bounded penalty parameter, thereby achieving complexity comparable to that of Carme in the unconstrained setting.

Fortunately, the answer is affirmative. And the penalty function is the classic Fletcher's augmented Lagrangian function [24, 26]. Fletcher's augmented Lagrangian is a modified augmented Lagrangian designed for equality-constrained optimization, which takes the form of

$$\mathcal{L}_\rho(x) = f(x) + \langle \lambda(x), c(x)\rangle + \frac{\rho}{2}\|c(x)\|^2,$$

where the multiplier estimate $\lambda(x) = -\nabla c(x)^\dagger \nabla f(x)$ is defined as the least-squares multiplier that approximately satisfies the KKT stationarity condition $\nabla f(x) + \nabla c(x)\lambda = 0$. In contrast to classical augmented Lagrangian methods, the multiplier is not treated as an independent dual variable, but is computed directly from first-order information of f and c. The pseudo-inverse $\nabla c(x)^\dagger$ ensures that this definition remains valid even when the Jacobian is rank-deficient.

Next, to apply Carme to $\min_{x\in\mathbb{R}^n} \mathcal{L}_\rho(x)$, we first characterize the gradient and Hessian of $\mathcal{L}_\rho(x)$. By direct differentiation, the gradient of $\mathcal{L}_\rho(x)$ is given by

$$\nabla \mathcal{L}_\rho(x) = \nabla f(x) + \nabla c(x)\lambda(x) + \nabla \lambda(x) c(x) + \rho \nabla c(x) c(x),$$

where $\lambda(x) = -\nabla c(x)^\dagger \nabla f(x)$, and the Hessian of $\mathcal{L}_\rho(x)$ can be expressed as

$$\begin{aligned}\nabla^2 \mathcal{L}_\rho(x) = &\nabla^2 f(x) + \sum_{i=1}^m \lambda_i(x)\nabla^2 c_i(x) + \rho \nabla c(x)\nabla c(x)^\top + \rho \sum_{i=1}^m c_i(x)\nabla^2 c_i(x) \\ &+ \nabla c(x)\nabla \lambda(x)^\top + \nabla \lambda(x)\nabla c(x)^\top + \rho \sum_{i=1}^m c_i(x)\nabla^2 \lambda_i(x).\end{aligned}$$

One can see that if $c(x) = 0$, $\nabla \mathcal{L}_\rho(x) = 0$ recovers the first-order stationarity, which further together with the condition of $d^\top \nabla^2 \mathcal{L}_\rho(x) d \geq 0$ for $\forall d \in \mathrm{Null}(\nabla c(x)^\top)$ yields the second-order stationarity.

However, the above expressions reveal two major challenges for constrained stochastic optimization. First, the multiplier $\lambda(x)$ depends on both $\nabla f(x)$ and $\nabla c(x)$, and consequently $\nabla^2 \mathcal{L}_\rho(x)$ involves third-order derivatives of f and c through $\nabla^2 \lambda(x)$. Computing these quantities is highly undesirable in practice. Second, in the stochastic setting, both $\nabla f(x)$ and $\nabla^2 f(x)$ are subject to sampling noise, which also propagates through $\lambda(x)$ to the gradient and Hessian of the augmented Lagrangian. These observations motivate the construction of stochastic estimators that (i) avoid explicit third-order derivatives, and (ii) control the variance introduced by multiplier estimation. To this end, we adopt the hybrid stochastic estimators g_k and H_k developed in Section 2. Specifically, we define the following approximations of the gradient and Hessian of $\mathcal{L}_\rho(x)$:

$$\tilde{\nabla} \mathcal{L}_\rho(x_k) = g_k + \nabla c(x_k)\tilde{\lambda}_k + \tilde{\nabla}\lambda_k\, c(x_k) + \rho \nabla c(x_k) c(x_k), \tag{34}$$

$$\tilde{\nabla}^2 \mathcal{L}_\rho(x_k) = H_k + \sum_{i=1}^m (\tilde{\lambda}_{k,i} + \rho c_i(x_k))\nabla^2 c_i(x_k) + \nabla c(x_k)(\tilde{\nabla}\lambda_k + \rho \nabla c(x_k))^\top + \tilde{\nabla}\lambda_k \nabla c(x_k)^\top, \tag{35}$$

where $\tilde{\lambda}_k = \nabla c(x_k)^\dagger g_k$ and $\tilde{\nabla}\lambda_k = \nabla c(x_k)^\dagger H_k + \nabla(\nabla c(x_k)^\dagger g_k)$. We emphasize that $\tilde{\nabla}^2 \mathcal{L}_\rho(x_k)$ is not an unbiased estimator of the exact Hessian $\nabla^2 \mathcal{L}_\rho(x_k)$. In particular, higher-order terms involving $\nabla^2 \lambda_i(x)$ are deliberately omitted to avoid third-order derivative information. The resulting approximation error is structured and proportional to the constraint violation $c(x_k)$. To explicitly control this error, the

cubic subproblem incorporates an additional quadratic correction term of the form $\frac{\rho L_H^\lambda \|c_k\|_1}{2}\|s\|^2$, which accounts for the worst-case curvature contribution of the omitted higher-order terms through an explicit, feasibility-dependent regularization. This results in an adaptive regularization mechanism whose strength scales with the current constraint violation $\|c_k\|$, allowing the model to safely compensate for the omitted higher-order information. In the negative-curvature step, the algorithm further restricts curvature tests to approximately feasible points, so that the neglected higher-order terms become negligible. Together, these mechanisms guarantee that $\tilde{\nabla}^2\mathcal{L}_\rho(x_k)$ provides reliable curvature information for both cubic-regularized and negative-curvature updates without requiring third-order derivatives. This design allows Carme to be applied to Fletcher's augmented Lagrangian without increasing the curvature constants or requiring higher-order derivatives, which will be key to achieving improved complexity bounds in the sequel. We now present the resulting algorithmic framework in Algorithm 2. The algorithm preserves the overall structure of Carme, while incorporating Fletcher's augmented Lagrangian, hybrid stochastic estimators, and feasibility-adaptive curvature regularization.

Algorithm 2: Carme-ALM

Input: Initial point $x_0 \in \mathbb{R}^n$, parameters $\epsilon, \gamma, \omega, \alpha_g, \alpha_h, \bar{\epsilon}_{g,1}, \bar{\epsilon}_{g,2}, \bar{\epsilon}_{h,1}, \bar{\epsilon}_{h,1}, \rho, \eta, M, p > 0$ and $K > 0$.

for $k = 0, \ldots, K$ **do**

Compute g_k and H_k through (34) and (35).

Sample $Q_k \sim$ Bernoulli(p) with bias p.

if $Q_k = 1$; `// cubic-regularized Newton step`

then

Solve subproblem

$$\min_{\|s\| \le \eta} m_k(s) = \langle \tilde{\nabla}\mathcal{L}_\rho(x_k), s\rangle + \frac{1}{2}\langle s, \tilde{\nabla}^2\mathcal{L}_\rho(x_k)s\rangle + \frac{\rho L_H^\lambda \|c_k\|_1}{2}\|s\|^2 + \frac{M}{6}\|s\|^3 \tag{36}$$

obtaining an inexact solution s_k satisfying Condition A.

Set $x_{k+1} = x_k + s_k$.

else

if $\|c(x_k)\| \le \frac{\gamma}{\sqrt{m}\rho L_H^\lambda}$ and $u^\top \tilde{\nabla}^2\mathcal{L}_\rho(x_k)u \le -4\gamma$ *for any* $u \in \mathrm{Null}(\nabla c(x_k)^\top)$ with $\|u\| = 1$;

`// negative-curvature step`

then

Find a unit vector u_k such that $u_k^\top \tilde{\nabla}^2\mathcal{L}_\rho(x_k)u_k \le -2\gamma$, $\forall u_k \in \mathrm{Null}(\nabla c(x_k)^\top)$.

Set $x_{k+1} = x_k + \frac{\gamma}{L_H^f} \cdot r_k \cdot u_k$, where $r_k \sim$ Uniform$(\{-1, 1\})$.

else

Set $x_{k+1} = x_k$;

4.1 Complexity analysis

In this subsection, we analyze the iteration and sample complexity of Carme-ALM. Compared with the unconstrained setting, the presence of equality constraints introduces additional technical challenges in the complexity analysis. In particular, the multiplier $\lambda(x) = \nabla c(x)^\dagger \nabla f(x)$ and its derivatives play a central role in both the gradient and Hessian approximations, and their stability is crucial for controlling the curvature of the augmented Lagrangian. To ensure that the multiplier mapping $\lambda(x)$ is well-defined and Lipschitz continuous along the iterates, we impose a uniform constraint qualification on the Jacobian of

the constraints. Specifically, we adopt a global version of the Linear Independence Constraint Qualification (LICQ), which guarantees that the Jacobian remains nondegenerate throughout the algorithm and allows us to uniformly bound the sensitivity of the pseudo-inverse $\nabla c(x)^\dagger$.

Assumption 5 (Strong LICQ) *The constraint Jacobian $\nabla c(x) \in \mathbb{R}^{n\times m}$ has full column rank at all iterates. Moreover, its smallest singular value is uniformly bounded away from zero; that is,*

$$\sigma_{\min}(\nabla c(x_k)) \geq \sqrt{\nu} \text{ for some } \nu > 0, \quad \forall k \geq 0.$$

Under Assumption 5, the pseudo-inverse $\nabla c(x)^\dagger$ is well-defined and uniformly bounded along the iterates. Consequently, the multiplier admits the explicit representation

$$\lambda(x) = -(\nabla c(x)^\top \nabla c(x))^{-1} \nabla c(x)^\top \nabla f(x),$$

and depends smoothly on both $\nabla f(x)$ and $\nabla c(x)$. Since $\lambda(x)$ enters both the gradient and Hessian of the Fletcher's augmented Lagrangian, controlling its sensitivity is essential for the curvature-based analysis of Carme-ALM. To simplify notation and avoid explicitly tracking third-order derivatives of f and c arising from $\nabla^2\lambda(x)$, we impose the following regularity condition on the multiplier mapping.

Assumption 6 *There exist positive constants L_g^λ and L_H^λ such that*

$$\|\nabla\lambda(x_k)\| \leq L_g^\lambda, \qquad \|\nabla^2\lambda_i(x_k)\| \leq L_H^\lambda, \quad \forall k \geq 0. \tag{37}$$

Besides, for any $\rho > 0$ there exist a constant $L_H^\rho > 0$ such that

$$\|\nabla^2\mathcal{L}_\rho(x_{k+1}) - \nabla^2\mathcal{L}_\rho(x_k)\| \leq L_H^\rho \|x_{k+1} - x_k\|, \quad \forall k \geq 0. \tag{38}$$

The above assumption is mild and holds under sufficient smoothness and regularity conditions. In particular, under Assumption 5, if the objective f and the constraint functions c_i admit bounded first-, second-, and third-order derivatives along the iterates, then the bounds in (37) follow directly. These regularity conditions ensure that the gradient and Hessian of the Fletcher's augmented Lagrangian $\mathcal{L}_\rho$ remain uniformly bounded along the iterates. Specifically, for any fixed $\rho < +\infty$, there exist constants $L_\rho > 0$ and $L_g^\rho > 0$ such that

$$\|\nabla\mathcal{L}_\rho(x_k)\| \leq L_\rho, \qquad \|\nabla^2\mathcal{L}_\rho(x_k)\| \leq L_g^\rho, \quad \forall k \geq 0. \tag{39}$$

Moreover, a sufficient condition for (38) to hold is that the fourth-order derivatives of f and c_i are bounded.

In what follows, we establish a key lemma connecting constraint $c(x)$ and gradient $\nabla\mathcal{L}_\rho(x)$. The lemma further shows that first-order stationarity of $\mathcal{L}_\rho(x)$ implies the KKT conditions of the constrained problem, confirming the exactness of the Fletcher's augmented Lagrangian. For brevity, denote $\mathcal{L}_k^\rho := \mathcal{L}_\rho(x_k)$, $f_k := f(x_k)$, $c_k := c(x_k)$ and $\lambda_k := \lambda(x_k)$.

Lemma 7 *Suppose Assumptions 1, 2, 5 and 6 hold, and let $\rho > \nu^{-1}L_cL_g^\lambda$. Then $\nabla\mathcal{L}_\rho(x_k) = 0$ implies $\nabla f_k + \nabla c_k\lambda_k = 0$ and $c_k = 0$. Moreover, if $\|\nabla\mathcal{L}_\rho(x_k)\| \leq \varepsilon$ and $\rho \geq \nu^{-1}L_c(L_g^\lambda + 1)$, we have $\|c_k\| \leq \varepsilon$ and $\|\nabla f_k + \nabla c_k\lambda_k\| \leq (1 + L_g^\lambda + \rho L_c)\varepsilon$.*

Proof. Left-multiplying $\nabla\mathcal{L}_k^\rho$ by $(\nabla c_k^\top\nabla c_k)^{-1}\nabla c_k^\top$ and using the definition $\lambda_k = -(\nabla c_k^\top\nabla c_k)^{-1}\nabla c_k^\top\nabla f_k$, we obtain

$$(\nabla c_k^\top\nabla c_k)^{-1}\nabla c_k^\top\nabla\mathcal{L}_k^\rho = \left(\rho I_m + (\nabla c_k^\top\nabla c_k)^{-1}\nabla c_k^\top\nabla\lambda_k\right)c_k.$$

Since $\rho > \|(\nabla c_k^\top\nabla c_k)^{-1}\nabla c_k^\top\nabla\lambda_k\|$, it follows that $\rho I_m + (\nabla c_k^\top\nabla c_k)^{-1}\nabla c_k^\top\nabla\lambda_k \succ 0$, with minimum eigenvalue at least $\rho - \nu^{-1}L_cL_g^\lambda > 0$. It together with $\nabla\mathcal{L}_k^\rho = 0$ yields $c_k = 0$, and then $\nabla f_k + \nabla c_k\lambda_k = \nabla\mathcal{L}_k^\rho - \nabla\lambda_k c_k - \rho\nabla c_k c_k = 0$.

Besides, if $\|\nabla\mathcal{L}_\rho(x_k)\| \le \varepsilon$, we have

$$\begin{aligned}\|c_k\| &\le \left\| \left(\rho I_m + (\nabla c_k^\top \nabla c_k)^{-1}\nabla c_k^\top \nabla \lambda_k\right)^{-1} (\nabla c_k^\top \nabla c_k)^{-1}\nabla c_k^\top \nabla \mathcal{L}_k^\rho \right\| \\ &\le \frac{\|(\nabla c_k^\top \nabla c_k)^{-1}\nabla c_k^\top\|\|\nabla \mathcal{L}_k^\rho\|}{\rho - \|(\nabla c_k^\top \nabla c_k)^{-1}\nabla c_k^\top \nabla \lambda_k\|} \le \varepsilon,\end{aligned}$$

where the second inequality uses the property of Neumann series $\|\,(\rho I_m + A_k)^{-1}\,\| \le \frac{1}{\rho - \|A_k\|}$ with $A_k := (\nabla c_k^\top \nabla c_k)^{-1}\nabla c_k^\top \nabla \lambda_k$, and the last inequality is due to $\rho \ge \nu^{-1}L_c(L_g^\lambda + 1)$. Then for the stationarity condition, it holds that

$$\|\nabla f_k + \nabla c_k \lambda_k\| = \|\nabla \mathcal{L}_k^\rho - \nabla\lambda_k c_k - \rho \nabla c_k c_k\| \le \|\nabla \mathcal{L}_k^\rho\| + (\|\nabla\lambda_k\| + \|\rho\nabla c_k\|)\|c_k\| \le (1 + L_g^\lambda + \rho L_c)\varepsilon.$$

The proof is completed. □

The following lemma bounds the gap between the stochastic approximations and their true counterparts.

Lemma 8 *Suppose that Assumptions 1, 2 and 5 hold. Then for each iteration k, the stochastic gradient and Hessian estimators $\tilde{\nabla}\mathcal{L}_\rho(x_k)$ and $\tilde{\nabla}^2\mathcal{L}_\rho(x_k)$ satisfy*

$$\|\tilde{\nabla}\mathcal{L}_\rho(x_k) - \nabla\mathcal{L}_\rho(x_k)\| \le \kappa_1\|g_k - \nabla f_k\| + \kappa_2\|c_k\|\|H_k - \nabla^2 f_k\|$$

and

$$\|\tilde{\nabla}^2\mathcal{L}_\rho(x_k) - \nabla^2\mathcal{L}_\rho(x_k) - \rho\sum_{i=1}^m c_i(x)\nabla^2\lambda_i(x)\| \le \kappa_3\|g_k - \nabla f_k\| + \kappa_4\|H_k - \nabla^2 f_k\|,$$

where $\kappa_1 = 1 + \nu^{-1}L_c^2 + \sqrt{m}L_g^c C(\nu^{-1} + 2\nu^{-2}L_c^2)$, $\kappa_2 = \nu^{-1}L_c$, $\kappa_3 = \nu^{-1}L_g^c L_c\left(m + 2\sqrt{m}(1 + 2\nu^{-1}L_c^2)\right)$ and $\kappa_4 = 1 + 2\nu^{-1}L_c^2$.

Proof. From the definitions of $\tilde{\nabla}\mathcal{L}_\rho(x_k)$ and $\nabla\mathcal{L}_\rho(x_k)$, we have

$$\begin{aligned}\|\tilde{\nabla}\mathcal{L}_\rho(x_k) - \nabla\mathcal{L}_\rho(x_k)\| &= \|g_k + \nabla c(x_k)\tilde{\lambda}_k + \tilde{\nabla}\lambda_k c(x_k) - \nabla f_k - \nabla c(x_k)\lambda_k - \nabla\lambda_k c(x_k)\| \\ &\le \|g_k - \nabla f_k\| + \|\nabla c_k(\nabla c_k^\top \nabla c_k)^{-1}\nabla c_k^\top (g_k - \nabla f_k)\| + \|(\tilde{\nabla}\lambda_k - \nabla\lambda_k)c(x_k)\|.\end{aligned}$$

For the last term $\|(\tilde{\nabla}\lambda_k - \nabla\lambda_k)c(x_k)\|$, we have

$$\begin{aligned}\|\tilde{\nabla}\lambda_k - \nabla\lambda_k\| &= \|(\nabla c_k^\top \nabla c_k)^{-1}\nabla c_k^\top (H_k - \nabla^2 f_k) + \nabla((\nabla c_k^\top \nabla c_k)^{-1}\nabla c_k^\top (g_k - \nabla f_k))\| \\ &\le \nu^{-1}L_c\|H_k - \nabla^2 f_k\| + \sqrt{m}L_g^c(\nu^{-1} + 2\nu^{-2}L_c^2)\|g_k - \nabla f_k\|,\end{aligned} \tag{40}$$

where the inequality uses Lemma 18. Hence, it holds that

$$\|\tilde{\nabla}\mathcal{L}_\rho(x_k) - \nabla\mathcal{L}_\rho(x_k)\| \le \left(1 + \nu^{-1}L_c^2 + \sqrt{m}L_g^c C(\nu^{-1} + 2\nu^{-2}L_c^2)\right)\|g_k - \nabla f_k\| + \nu^{-1}L_c\|c_k\|\|H_k - \nabla^2 f_k\|.$$

For the gap of Hessian, we have

$$\begin{aligned}&\|\tilde{\nabla}^2\mathcal{L}_\rho(x_k) - \nabla^2\mathcal{L}_\rho(x_k) - \rho\sum_{i=1}^m c_i(x)\nabla^2\lambda_i(x)\| \\ &\le \|H_k - \nabla^2 f_k\| + mL_g^c\|(\nabla c_k^\top \nabla c_k)^{-1}\nabla c_k^\top (g_k - \nabla f_k)\| + 2L_c\|\tilde{\nabla}\lambda_k - \nabla\lambda_k\| \\ &\le (1 + 2\nu^{-1}L_c^2)\|H_k - \nabla^2 f_k\| + \nu^{-1}L_g^c L_c\left(m + 2\sqrt{m}(1 + 2\nu^{-1}L_c^2)\right)\|g_k - \nabla f_k\|,\end{aligned}$$

where the last inequality comes from (40). The proof is completed. □

Next, we analyze the relationship between the constraint residual c_k and the regularized Newton step s_k. This is crucial because, for computational efficiency, we omit the Hessian term $\rho\sum_{i=1}^m c_i(x)\nabla^2\lambda_i(x)$ in our algorithm, which depends explicitly on $c(x)$. By quantifying how c_k evolves along s_k, we can control the error introduced by ignoring this term and ensure that the resulting Hessian approximation remains accurate enough for our complexity analysis. The following lemma describes this relationship.

Lemma 9 *Suppose that Assumptions 1, 2, 5 and 6 hold, and let* $\|g_k - \nabla f_k\| \le \epsilon_g$, $\|H_k - \nabla^2 f_k\| \le \epsilon_h$ *and* $\nu^{-1}L_c(L_g^\lambda + \nu^{-1}L_c\epsilon_h + \sqrt{m}L_g^c(\nu^{-1} + 2\nu^{-2}L_c^2)\epsilon_g + 1) \le \rho < +\infty$ *for all* $k \ge 0$. *Then for the cubic-regularized Newton step,* $\|s_k\| < \eta$ *implies*

$$\|c_k\| \le M\omega^{2/3} + \kappa_5\|s_k\|,$$

where κ_5 *is a constant satisfying* $\kappa_5 \ge L_g^\rho + 2\sqrt{m}\rho L_H^\lambda C + \frac{M\eta}{2} + \kappa_3\epsilon_g + \kappa_4\epsilon_h$ *and* L_g^ρ *is introduced in* (39).

Proof. From (16b) in Condition A, if $\|s_k\| < \eta$, we have

$$\left\|\tilde{\nabla}\mathcal{L}_k^\rho + \tilde{\nabla}^2\mathcal{L}_k^\rho s_k + \rho L_H^\lambda\|c_k\|s_k + \frac{M}{2}\|s_k\|s_k\right\| \le M\omega^{2/3},$$

then there exists a vector $z_k \in \mathbb{R}^n$ with $\|z_k\| \le M\omega^{2/3}$ such that

$$\tilde{\nabla}\mathcal{L}_k^\rho + \tilde{\nabla}^2\mathcal{L}_k^\rho s_k + \rho L_H^\lambda\|c_k\|_1 s_k + \frac{M}{2}\|s_k\|s_k = z_k.$$

Hence, left-multiplying both sides by $(\nabla c_k^\top \nabla c_k)^{-1}\nabla c_k^\top$ and using the definition of $\tilde{\nabla}\mathcal{L}_k^\rho$, it holds that

$$\left(\rho I_m + (\nabla c_k^\top \nabla c_k)^{-1}\nabla c_k^\top \tilde{\nabla}\lambda_k\right) c_k = (\nabla c_k^\top \nabla c_k)^{-1}\nabla c_k^\top \left(z_k - \tilde{\nabla}^2\mathcal{L}_k^\rho s_k - \rho L_H^\lambda\|c_k\|_1 s_k - \frac{M}{2}\|s_k\|s_k\right),$$

where we use $(\nabla c_k^\top \nabla c_k)^{-1}\nabla c_k^\top(\nabla f_k + \nabla c_k\tilde{\lambda}_k) = 0$. It then together with Cauchy–Schwarz inequality yields

$$\begin{aligned}
\|c_k\| &\le \left\|\left(\rho I_m + (\nabla c_k^\top \nabla c_k)^{-1}\nabla c_k^\top \tilde{\nabla}\lambda_k\right)^{-1}\right\| \left\|(\nabla c_k^\top \nabla c_k)^{-1}\nabla c_k^\top\right\| \left\|z_k - \tilde{\nabla}^2\mathcal{L}_k^\rho s_k - \rho L_H^\lambda\|c_k\|_1 s_k - \frac{M}{2}\|s_k\|s_k\right\| \\
&\le \frac{\nu^{-1}L_c}{\rho - \left\|(\nabla c_k^\top \nabla c_k)^{-1}\nabla c_k^\top \tilde{\nabla}\lambda_k\right\|}\left(M\omega^{2/3} + (L_g^\rho + 2\sqrt{m}\rho L_H^\lambda C + \frac{M\eta}{2} + \kappa_3\epsilon_g + \kappa_4\epsilon_h)\|s_k\|\right) \\
&\le M\omega^{2/3} + \kappa_5\|s_k\|,
\end{aligned}$$

where the second uses the property of Neumann series and $\|\tilde{\nabla}^2\mathcal{L}_k^\rho\| \le L_g^\rho + \sqrt{m}\rho L_H^\lambda C + \kappa_3\epsilon_g + \kappa_4\epsilon_h$ according to Lemma 8 and (39), and the last line is due to

$$\rho - \left\|(\nabla c_k^\top \nabla c_k)^{-1}\nabla c_k^\top \tilde{\nabla}\lambda_k\right\| \ge \nu^{-1}L_c(L_g^\lambda + \nu^{-1}L_c\epsilon_h + \sqrt{m}L_g^c(\nu^{-1} + 2\nu^{-2}L_c^2)\epsilon_g + 1) - \nu^{-1}L_c\|\tilde{\nabla}\lambda_k\| \ge \nu^{-1}L_c$$

thanks to (40). □

Remark 4 *In Lemma 9,* κ_5 *is a constant relying on* ϵ_g, ϵ_h *and* η. *We will prove the boundedness of* ϵ_g *and* ϵ_h *later and give the specific parameter settings for* η, *providing the setting of* κ_5 *in* (48).

We now establish a series of lemmas to relate the decrease of the Fletcher's augmented Lagrangian along the regularized Newton step to the gradient norm. We first bound the function decrease in terms of the step s_k generated by the regularized Newton update. We then derive a lower bound on $\|\nabla\mathcal{L}_\rho(x_{k+1})\|$ in terms of $\|s_k\|$, where the key ingredient is Lemma 9 describing the relationship between c_k and s_k. Combining these two results yields the desired decrease estimate directly in terms of the gradient of the Fletcher's augmented Lagrangian.

Lemma 10 *Under the same conditions as in Lemma 9, let* s_k *be an inexact solution of* (36) *satisfying Condition A. Then, for any* $M \ge 4L_H^\rho$ *and* $\eta \in (0, +\infty)$, *the point* $x_k + s_k$ *satisfies*

$$\mathcal{L}_\rho(x_k) - \mathcal{L}_\rho(x_k + s_k) \ge \frac{M}{15}\|s_k\|^3 - \frac{8\kappa_6^{3/2}}{\sqrt{M}}\|\nabla f_k - g_k\|^{3/2} - \frac{8(\kappa_7^{3/2}\eta^{3/2} + \kappa_2^{3/2}M^{3/2}\omega)}{\sqrt{M}}\|\nabla^2 f_k - H_k\|^{3/2}. \quad (41)$$

where $\kappa_6 \ge \kappa_1 + \kappa_3\eta$ *and* $\kappa_7 = \kappa_2\kappa_5 + \kappa_4$ *are constants.*

Proof. Using the Lipschitz continuity of $\nabla^2\mathcal{L}_\rho$ and the quadratic correction term that compensates for the omitted term $\rho\sum_i c_i\nabla^2\lambda_i$, we obtain

$$\begin{aligned}
&\mathcal{L}_\rho(x_k+s_k)-\mathcal{L}_k^\rho \le \langle\nabla\mathcal{L}_k^\rho, s_k\rangle + \frac{1}{2}\langle s_k, (\nabla^2\mathcal{L}_k^\rho - \rho\sum_{i=1}^m c_i(x_k)\nabla^2\lambda_i(x_k))s_k\rangle + \frac{\rho L_H^\lambda\|c_k\|_1}{2}\|s_k\|^2 + \frac{L_H^\rho}{6}\|s_k\|^3\\
&= m_k(s_k) + \frac{L_H^\rho - M}{6}\|s_k\|^3 + \langle\nabla\mathcal{L}_k^\rho - \tilde\nabla\mathcal{L}_k^\rho, s_k\rangle + \frac{1}{2}\langle s_k, (\nabla^2\mathcal{L}_k^\rho - \tilde\nabla^2\mathcal{L}_k^\rho - \rho\sum_{i=1}^m c_i(x_k)\nabla^2\lambda_i(x_k))s_k\rangle\\
&\le -\frac{M}{8}\|s_k\|^3 + \|\nabla\mathcal{L}_k^\rho - \tilde\nabla\mathcal{L}_k^\rho\|\|s_k\| + \frac{1}{2}\|\nabla^2\mathcal{L}_k^\rho - \tilde\nabla^2\mathcal{L}_k^\rho - \rho\sum_{i=1}^m c_i(x_k)\nabla^2\lambda_i(x_k)\|\|s_k\|^2\\
&\le -\frac{M}{8}\|s_k\|^3 + (\kappa_1+\kappa_3\eta)\|\nabla f_k - g_k\|\|s_k\| + (\kappa_2\kappa_5+\kappa_4)\|\nabla^2 f_k - H_k\|\|s_k\|^2 + \kappa_2 M\omega^{2/3}\|\nabla^2 f_k - H_k\|\|s_k\|,
\end{aligned}$$

where the second inequality uses $M \ge 4L_H^\rho$ and the last inequality uses Lemma 8. Then using the scaled Young's inequality $ab \le \frac{8}{\sqrt{M}}a^{3/2} + \frac{M}{64}b^3$ yields

$$\begin{aligned}
(\kappa_1+\kappa_3\eta)\|\nabla f_k - g_k\|\|s_k\| &\le \frac{8(\kappa_1+\kappa_3\eta)^{3/2}}{\sqrt{M}}\|\nabla f_k - g_k\|^{3/2} + \frac{M}{64}\|s_k\|^3,\\
(\kappa_2\kappa_5+\kappa_4)\|\nabla^2 f_k - H_k\|\|s_k\|^2 &\le \frac{8(\kappa_2\kappa_5+\kappa_4)^{3/2}}{\sqrt{M}}\|\nabla^2 f_k - H_k\|^{3/2}\eta^{3/2} + \frac{M}{64}\|s_k\|^3,\\
\kappa_2 M\omega^{2/3}\|\nabla^2 f_k - H_k\|\|s_k\| &\le \frac{8\kappa_2^{3/2}M^{3/2}\omega}{\sqrt{M}}\|\nabla^2 f_k - H_k\|^{3/2} + \frac{M}{64}\|s_k\|^3.
\end{aligned}$$

Then plugging these bounds into (18), we obtain

$$\mathcal{L}_\rho(x_k+s_k)-\mathcal{L}_k^\rho \le -\frac{M}{15}\|s_k\|^3 + \frac{8\kappa_6^{3/2}}{\sqrt{M}}\|\nabla f_k - g_k\|^{3/2} + \frac{8(\kappa_7^{3/2}\eta^{3/2} + \kappa_2^{3/2}M^{3/2}\omega)}{\sqrt{M}}\|\nabla^2 f_k - H_k\|^{3/2},$$

where $\kappa_6 \ge \kappa_1+\kappa_3\eta$ and $\kappa_7 = \kappa_2\kappa_5+\kappa_4$. Rearranging the terms yields the conclusion. □

The following lemma establishes the relationship between the regularized Newton step s_k and the gradient $\nabla\mathcal{L}_\rho(x_k+s_k)$.

Lemma 11 *Under the same conditions as Lemma 10, it holds that for any $M \ge L_H^\rho$,*

$$\begin{aligned}
&\mathbf{1}\left\{\|\nabla\mathcal{L}_\rho(x_k+s_k)\| \ge \frac{(M+\sqrt{m}\rho L_H^\lambda\kappa_5)\eta^2}{2}\right\}\\
&\le \frac{2}{\eta^2}\|s_k\|^2 + \frac{2\left(\kappa_6\|\epsilon_k^g\| + (\kappa_2 M\omega^{2/3}+\kappa_7\eta)\|\epsilon_k^h\| + (1+2\sqrt{m}\rho L_H^\lambda\eta)M\omega^{2/3}\right)}{(M+\sqrt{m}\rho L_H^\lambda\kappa_5)\eta^2},
\end{aligned} \tag{42}$$

where $\kappa_6 \ge \kappa_1+\kappa_3\eta$ and $\kappa_7 = \kappa_2\kappa_5+\kappa_4$, defined in Lemma 10.

Proof. In the regularized Newton step, one of two cases must occur: either $\|s_k\| = \eta$, or $\|s_k\| < \eta$. In

the second case, together with $\epsilon_k^g = \nabla f_k - g_k$ and $\epsilon_k^h = \nabla^2 f(x) - H_k$ we have

$$\begin{aligned}
&\|\nabla \mathcal{L}_\rho(x_k + s_k)\| \le \left\|\nabla \mathcal{L}_\rho(x_k + s_k) - \nabla \mathcal{L}_k^\rho - \nabla^2 \mathcal{L}_k^\rho s_k\right\| + \left\|\nabla \mathcal{L}_k^\rho + \nabla^2 \mathcal{L}_k^\rho s_k\right\| \\
\le\ & \frac{L_H^\rho}{2}\|s_k\|^2 + \left\|\nabla \mathcal{L}_k^\rho - \tilde{\nabla} \mathcal{L}_k^\rho\right\| + \left\|(\nabla^2 \mathcal{L}_k^\rho - \tilde{\nabla}^2 \mathcal{L}_k^\rho - \rho \sum_{i=1}^m c_i(x)\nabla^2 \lambda_i(x))s_k\right\| \\
&+ \left\|\tilde{\nabla} \mathcal{L}_k^\rho + \tilde{\nabla}^2 \mathcal{L}_k^\rho s_k + \rho L_H^\lambda \|c_k\|_1 s_k + \frac{M}{2}\|s_k\| s_k\right\| + \frac{M}{2}\|s_k\|^2 + 2\rho L_H^\lambda \|c_k\|_1 \|s_k\| \qquad (43) \\
\le\ & \frac{L_H^\rho + M}{2}\|s_k\|^2 + (\kappa_1 + \kappa_3\eta)\|\epsilon_k^g\| + (\kappa_2\|c_k\| + \kappa_4\eta)\|\epsilon_k^h\| + M\omega^{2/3} + 2\rho L_H^\lambda \|c_k\|_1 \|s_k\| \\
\le\ & \frac{L_H^\rho + M + 2\sqrt{m}\rho L_H^\lambda \kappa_5}{2}\|s_k\|^2 + \kappa_6\|\epsilon_k^g\| + (\kappa_2 M\omega^{2/3} + \kappa_7\eta)\|\epsilon_k^h\| + (1 + 2\sqrt{m}\rho L_H^\lambda \eta)M\omega^{2/3},
\end{aligned}$$

where the first and second inequalities use triangle inequality, the second inequality also follows from the L_H^ρ-Lipschitz continuity of $\nabla^2 \mathcal{L}_\rho$, the third inequality is due to (16b) in Condition A, Lemma 8 and $\|s_k\| < \eta$, and the last one comes from Lemma 9. Rearranging the terms in (43) yields

$$\|s_k\|^2 \ge \frac{2\left(\|\nabla \mathcal{L}_\rho(x_k + s_k)\| - \kappa_6\|\epsilon_k^g\| - (\kappa_2 M\omega^{2/3} + \kappa_7\eta)\|\epsilon_k^h\| - (1 + 2\sqrt{m}\rho L_H^\lambda \eta)M\omega^{2/3}\right)}{L_H^\rho + M + 2\sqrt{m}\rho L_H^\lambda \kappa_5}.$$

Since one of the two cases ($\|s_k\| < \eta$ or $\|s_k\| = \eta$) must hold, we have

$$\begin{aligned}
&\|s_k\|^2 + \frac{2\left(\kappa_6\|\epsilon_k^g\| + (\kappa_2 M\omega^{2/3} + \kappa_7\eta)\|\epsilon_k^h\| + (1 + 2\sqrt{m}\rho L_H^\lambda \eta)M\omega^{2/3}\right)}{L_H^\rho + M + 2\sqrt{m}\rho L_H^\lambda \kappa_5} \\
&\qquad \ge \min\left\{\eta^2, \frac{2}{L_H^\rho + M + 2\sqrt{m}\rho L_H^\lambda \kappa_5}\|\nabla \mathcal{L}_\rho(x_k + s_k)\|\right\}.
\end{aligned}$$

Rearranging the terms and applying the bound $M \ge L_H^\rho$ yields

$$\begin{aligned}
&(M + \sqrt{m}\rho L_H^\lambda \kappa_5)\|s_k\|^2 + \kappa_6\|\epsilon_k^g\| + (\kappa_2 M\omega^{2/3} + \kappa_7\eta)\|\epsilon_k^h\| + (1 + 2\sqrt{m}\rho L_H^\lambda \eta)M\omega^{2/3} \\
\ge\ & \min\left\{\frac{(M + \sqrt{m}\rho L_H^\lambda \kappa_5)\eta^2}{2}, \|\nabla \mathcal{L}_\rho(x_k + s_k)\|\right\} \\
\ge\ & \frac{(M + \sqrt{m}\rho L_H^\lambda \kappa_5)\eta^2}{2} \cdot \mathbf{1}\left\{\|\nabla \mathcal{L}_\rho(x_k + s_k)\| \ge \frac{(M + \sqrt{m}\rho L_H^\lambda \kappa_5)\eta^2}{2}\right\},
\end{aligned}$$

where the last inequality uses the fact that for any $a, b \ge 0$, $\min\{a, b\} \ge a \cdot \mathbf{1}\{b \ge a\}$. Hence, the conclusion is derived. □

The following lemma relates the expected decrease of the Fletcher's augmented Lagrangian to the probability of observing a large gradient at the next iterate.

Lemma 12 *Under the same conditions as in Lemma 10, for any iteration k with $Q_k = 1$, it holds that*

$$\begin{aligned}
&\mathbb{E}\left[\mathcal{L}_\rho(x_k) - \mathcal{L}_\rho(x_{k+1})\right] \ge \frac{M\eta^3}{90}\Pr\left\{\|\nabla \mathcal{L}_\rho(x_k + s_k)\| \ge \frac{(M + \sqrt{m}\rho L_H^\lambda \kappa_5)\eta^2}{2}\right\} \\
&- \frac{9\kappa_6^{3/2}}{\sqrt{M}}\mathbb{E}\left[\|\epsilon_k^g\|^{3/2}\right] - \frac{9(\kappa_7^{3/2}\eta^{3/2} + \kappa_2^{3/2}M^{3/2}\omega)}{\sqrt{M}}\mathbb{E}\left[\|\epsilon_k^h\|^{3/2}\right] - \frac{(1 + 2\sqrt{m}\rho L_H^\lambda \eta)^{3/2}M}{15}\omega,
\end{aligned}$$

where $\Pr(\cdot)$ and $\mathbb{E}[\cdot]$ are taken w.r.t. the randomness over g_k and H_k.

Proof. Using Lemma 10 and taking expectations in (41), we obtain

$$\mathbb{E}[\mathcal{L}_\rho(x_k) - \mathcal{L}_\rho(x_{k+1})] \geq \frac{M}{15}\mathbb{E}\big[\|s_k\|^3\big] - \frac{8\kappa_6^{3/2}}{\sqrt{M}}\mathbb{E}\Big[\|\epsilon_k^g\|^{3/2}\Big] - \frac{8(\kappa_7^{3/2}\eta^{3/2} + \kappa_2^{3/2}M^{3/2}\omega)}{\sqrt{M}}\mathbb{E}\Big[\|\epsilon_k^h\|^{3/2}\Big]. \tag{44}$$

Next, to relate the update s_k to the gradient norm, we invoke Lemma 11. To match the cubic term $\|s_k\|^3$, we raise (42) to power $3/2$ and apply the inequality $(\sum_{i=1}^n a_i)^q \leq n^{q-1}\sum_{i=1}^n a_i^q$ for $q = 3/2$, which gives

$$\begin{aligned}
&\mathbf{1}\left\{\|\nabla\mathcal{L}_\rho(x_k + s_k)\| \geq \frac{(M + \sqrt{m}\rho L_H^\lambda\kappa_5)\eta^2}{2}\right\} \\
&\leq \left(\frac{2}{\eta^2}\|s_k\|^2 + \frac{2\left(\kappa_6\|\epsilon_k^g\| + (\kappa_2 M\omega^{2/3} + \kappa_7\eta)\|\epsilon_k^h\| + (1 + 2\sqrt{m}\rho L_H^\lambda\eta)M\omega^{2/3}\right)}{(M + \sqrt{m}\rho L_H^\lambda\kappa_5)\eta^2}\right)^{3/2} \\
&\leq \frac{6}{\eta^3}\|s_k\|^3 + \frac{6\left(\kappa_6^{3/2}\|\epsilon_k^g\|^{3/2} + (\kappa_2 M\omega^{2/3} + \kappa_7\eta)^{3/2}\|\epsilon_k^h\|^{3/2} + (1 + 2\sqrt{m}\rho L_H^\lambda\eta)^{3/2}M^{3/2}\omega\right)}{(M + \sqrt{m}\rho L_H^\lambda\kappa_5)^{3/2}\eta^3}.
\end{aligned}$$

Taking expectations and applying Jensen's inequality again yields

$$\begin{aligned}
\mathbb{E}\big[\|s_k\|^3\big] \geq \frac{\eta^3}{6}\Pr&\left\{\|\nabla\mathcal{L}_\rho(x_k + s_k)\| \geq \frac{(M + \sqrt{m}\rho L_H^\lambda\kappa_5)\eta^2}{2}\right\} \\
&- \frac{\kappa_6^{3/2}\mathbb{E}\left[\|\epsilon_k^g\|^{3/2}\right] + (\kappa_2 M\omega^{2/3} + \kappa_7\eta)^{3/2}\mathbb{E}\left[\|\epsilon_k^h\|^{3/2}\right] + (1 + 2\sqrt{m}\rho L_H^\lambda\eta)^{3/2}M^{3/2}\omega}{(M + \sqrt{m}\rho L_H^\lambda\kappa_5)^{3/2}}.
\end{aligned} \tag{45}$$

Combining (44) and (45), and using $M + \sqrt{m}\rho L_H^\lambda\kappa_5 \geq M$, the desired inequality follows. □

Lemma 12 serves the same purpose as the corresponding descent lemma (Lemma 5) in the unconstrained Carme analysis. The difference stems from the deliberate omission of third-order derivative terms involving the constraints in the Hessian approximation. This omission introduces additional error terms, which are explicitly controlled via feasibility-aware regularization, while the underlying descent mechanism remains unchanged. The next lemma is the counterpart of the negative-curvature descent result in unconstrained Carme. The proof follows the same argument, with the difference also introduced by the omitted term.

Lemma 13 *Under the same conditions as in Lemma 10, for any iteration k with $Q_k = 0$, it holds that*

$$\mathbb{E}\left[\mathcal{L}_\rho(x_k) - \mathcal{L}_\rho(x_{k+1})\right] \geq \frac{\gamma^3}{3(L_H^\rho)^2}\cdot\Pr\left(\mathcal{E}_k\right) - \frac{\gamma^2}{2(L_H^\rho)^2}\left(\kappa_3\|\epsilon_k^g\| + \kappa_4\|\epsilon_k^h\|\right), \tag{46}$$

where $\Pr(\cdot)$ *and* $\mathbb{E}[\cdot]$ *are taken w.r.t. the randomness in* H_k *and* r_k*, the event* $\mathcal{E}_k$ *is defined by*

$$\mathcal{E}_k = \left\{\|c(x_k)\| \leq \frac{\gamma}{\sqrt{m}\rho L_H^\lambda} \text{ and } u^\top\tilde{\nabla}^2\mathcal{L}_\rho(x_k)u \leq -4\gamma,\ \forall u \in \mathrm{Null}(\nabla c(x_k)^\top) \text{ with } \|u\| = 1\right\}. \tag{47}$$

Proof. If the event $\mathcal{E}_k$ does not occur, then the algorithm sets $x_{k+1} = x_k$, and hence $\mathcal{L}_\rho(x_{k+1}) = \mathcal{L}_\rho(x_k)$. Otherwise, when $\mathcal{E}_k$ occurs, we choose a unit vector $u_k \in \mathrm{Null}(\nabla c(x_k)^\top)$ satisfying $u_k^\top\tilde{\nabla}^2\mathcal{L}_\rho(x_k)u_k \leq -2\gamma$, and define

$$\tilde{s}_k := \frac{\gamma}{L_H^\rho}\, r_k\, u_k,$$

where r_k is a Rademacher random variable. Using the L_H^ρ-Lipschitz continuity of $\nabla^2\mathcal{L}_\rho$, we have

$$\mathcal{L}_\rho(x_{k+1}) \leq \mathcal{L}_\rho(x_k) + \langle\nabla\mathcal{L}_\rho(x_k), \tilde{s}_k\rangle + \frac{1}{2}\tilde{s}_k^\top\nabla^2\mathcal{L}_\rho(x_k)\tilde{s}_k + \frac{L_H^\rho}{6}\|\tilde{s}_k\|^3.$$

Conditioning on H_k (and hence on u_k), and using $\mathbb{E}[r_k] = 0$ and $\mathbb{E}[r_k^2] = 1$, we obtain

$$\begin{aligned}
&\mathbb{E}_{r_k}\Big[\langle \nabla \mathcal{L}_\rho(x_k), \tilde{s}_k\rangle + \tfrac{1}{2}\tilde{s}_k^\top \nabla^2 \mathcal{L}_\rho(x_k)\tilde{s}_k \mid H_k\Big] \\
&= \frac{\gamma^2}{2(L_H^\rho)^2}\left[u_k^\top \tilde{\nabla}^2 \mathcal{L}_\rho(x_k) u_k + u_k^\top \left(\nabla^2 \mathcal{L}_\rho(x_k) - \tilde{\nabla}^2 \mathcal{L}_\rho(x_k)\right) u_k\right] \\
&\le -\frac{\gamma^3}{(L_H^\rho)^2} + \frac{\gamma^2}{2(L_H^\rho)^2}\left(\kappa_3\|\nabla f(x_k) - g_k\| + \kappa_4\|\nabla^2 f(x_k) - H_k\| + \sqrt{m}\rho L_H^\lambda \|c(x_k)\|\right),
\end{aligned}$$

where we used the definition of u_k and Lemma 8. Since the event $\mathcal{E}_k$ ensures $\|c(x_k)\| \le \frac{\gamma}{\sqrt{m}\rho L_H^\lambda}$, the last term can be absorbed, yielding

$$\mathbb{E}_{r_k}\Big[\langle \nabla \mathcal{L}_\rho(x_k), \tilde{s}_k\rangle + \tfrac{1}{2}\tilde{s}_k^\top \nabla^2 \mathcal{L}_\rho(x_k)\tilde{s}_k \mid H_k\Big] \le -\frac{\gamma^3}{2(L_H^\rho)^2} + \frac{\gamma^2}{2(L_H^\rho)^2}\left(\kappa_3\|\epsilon_k^g\| + \kappa_4\|\epsilon_k^h\|\right).$$

Since $\|\tilde{s}_k\| = \frac{\gamma}{L_H^\rho}$, we further obtain

$$\mathbb{E}_{r_k}[\mathcal{L}_\rho(x_{k+1}) \mid H_k] \le \mathcal{L}_\rho(x_k) - \frac{\gamma^3}{3(L_H^\rho)^2} + \frac{\gamma^2}{2(L_H^\rho)^2}\left(\kappa_3\|\epsilon_k^g\| + \kappa_4\|\epsilon_k^h\|\right)$$

on the event E_k. Combining the two cases, conditioning on H_k, we have

$$\mathbb{E}_{r_k}[\mathcal{L}_\rho(x_{k+1}) \mid H_k] \le \mathcal{L}_\rho(x_k) - \frac{\gamma^3}{3(L_H^\rho)^2}\mathbf{1}\{\mathcal{E}_k\} + \frac{\gamma^2}{2(L_H^\rho)^2}\left(\kappa_3\|\epsilon_k^g\| + \kappa_4\|\epsilon_k^h\|\right).$$

Taking expectation over H_k and rearranging yields (46). □

Lemma 13 characterizes the expected decrease of the Fletcher's augmented Lagrangian along a negative-curvature step. Compared with the analysis on Carme in unconstrained setting, the main difference lies in the presence of feasibility-induced error terms arising from the omission of constraint-dependent higher-order derivatives. The restriction imposed by the event $\mathcal{E}_k$ guarantees that the additional error terms introduced by omitting constraint-dependent higher-order derivatives remain negligible. This allows the analysis on negative-curvature descent to follow the same logical structure as Carme, up to controllable estimation errors.

Combining Lemma 12 and Lemma 13, together with the bounds on the stochastic gradient and Hessian errors, we obtain a uniform lower bound on the per-iteration expected decrease of the Fletcher's augmented Lagrangian, namely on $\mathbb{E}[\mathcal{L}_\rho(x_k) - \mathcal{L}_\rho(x_{k+1})]$ under the randomized step selection in Carme-ALM. Summing this bound over iterations yields the following complexity result. For notational convenience, we define

$$\kappa_8 := \sqrt{m}\,\rho\, L_H^\lambda \kappa_5,\ \ \kappa_9 := 30M^{1/2}\kappa_7 + (M+\kappa_8)^{1/2}\kappa_2,$$

where

$$\begin{aligned}
\rho = \nu^{-1}L_c(L_g^\lambda + \nu^{-1}L_c + \sqrt{m}L_g^c(\nu^{-1} + 2\nu^{-2}L_c^2) + 1),\ \ M = 4L_H^\rho \\
\kappa_5 = L_g^\rho + 2\sqrt{m}\rho L_H^\lambda C + M^{1/2} + \kappa_3 + \kappa_4.
\end{aligned} \tag{48}$$

Next, given $\epsilon \in (0,1)$ and $\gamma \ge 0$, define the auxiliary tolerance

$$\varepsilon = \min\left\{\frac{\epsilon}{450\max\{1, L_g^\lambda + \rho L_c\}}, \frac{\gamma}{450\sqrt{m}\rho L_H^\lambda}, \frac{4\kappa_6(M+\kappa_8)}{M}\right\}, \tag{49}$$

where $\kappa_6 = \kappa_1 + \frac{2\kappa_3}{M^{1/2}}$. Then we specify the remaining parameters used in our analysis as follows:

Algorithmic parameters.

$$\Delta = \mathcal{L}_\rho(x_0) + (\nu^{-1}L_c L_f + 1)C, \quad \eta = \frac{30\varepsilon^{1/2}}{(M+\kappa_8)^{1/2}} \le \frac{2}{M^{1/2}}, \quad p = \frac{(M+\kappa_8)^{3/2}\gamma^3}{(M+\kappa_8)^{3/2}\gamma^3 + 900M(L_H^\rho)^2\varepsilon^{3/2}},$$
$$\omega = \min\left\{M^{-1}, M^{-9/4}\varepsilon^{3/4}, \frac{\varepsilon^{3/2}}{(M+\kappa_8)^{3/2}(1+2\sqrt{m}\rho L_H^\lambda \eta)}\right\}, \quad K = \left\lceil \frac{3600\,\Delta(L_H^\rho)^2}{\gamma^3} + \frac{4\Delta(M+\kappa_8)^{3/2}}{M\varepsilon^{3/2}}\right\rceil.$$

Estimator parameters.

$$\begin{aligned}
&\delta = \frac{1}{200K}, \quad \bar\epsilon_{g,1} = \min\left\{\frac{M\varepsilon}{\kappa_6(M+\kappa_8)}, \frac{\gamma}{250\kappa_3}\right\}, \quad \bar\epsilon_{g,2} = \left(\frac{\varepsilon^{-3/2}+\gamma^{-3}}{\varepsilon^{-1/2}+\gamma^{-1}}\right)^{1/2}\bar\epsilon_{g,1}^2, \quad \alpha_g = \frac{\bar\epsilon_{g,1}^2}{\bar\epsilon_{g,2}}, \\
&\bar\epsilon_{h,1} = \frac{1}{10}\min\left\{\left(\frac{M^3\varepsilon}{(M+\kappa_8)\kappa_9^2}\right)^{1/2}, \frac{\gamma}{100\kappa_4}\right\}, \quad \bar\epsilon_{h,2} = \left(\frac{\varepsilon^{-3/2}+\gamma^{-3}}{\varepsilon^{-1/2}+\gamma^{-1}}\right)^{1/2}\bar\epsilon_{h,1}^2, \quad \alpha_h = \frac{\bar\epsilon_{h,1}^2}{\bar\epsilon_{h,2}}.
\end{aligned} \tag{50}$$

And B_k^g and B_k^h can be chosen as in Lemmas 1 and 2. These choices follow the same design as in the unconstrained Carme analysis, with additional constants accounting for feasibility control. Under the above parameters setting, we have the convergence result below.

Theorem 2 *Under Assumptions 1–6, let $\epsilon \in (0,1)$ and $\gamma \in (0, (600\Delta(L_H^\rho)^2)^{1/3})$. Consider Carme-ALM with the parameter choices specified in (48)–(50) (in particular, ρ, M, ε, η, p, ω, K, and $\delta = 1/(200K)$), and with the gradient/Hessian estimators chosen as in Lemmas 1 and 2. Then the following statements hold with probability at least 0.96:*

(i) Carme-ALM returns an iterate x_R for which there exists a multiplier λ_R such that

$$\|\nabla f(x_R) + \nabla c(x_R)\lambda_R\| \le 2\epsilon, \qquad \|c(x_R)\| \le \epsilon,$$

and

$$u^\top \nabla^2 \mathcal{L}_\rho(x_R)u \ge -6\gamma, \quad \forall\, u \in \mathrm{Null}(\nabla c(x_R)^\top),\ \|u\| = 1,$$

within $K = O\big(\gamma^{-3} + \epsilon^{-3/2}\big)$ iterations.

(ii) To reach an (ϵ,γ)-SSP of (14), Carme-ALM requires at most $\tilde O(\epsilon^{-3} + \epsilon^{-2}\gamma^{-2} + \gamma^{-4})$ stochastic gradient queries in expectation and $\tilde O(\epsilon^{-2} + \gamma^{-4})$ stochastic Hessian queries.

Proof. We divide the proof into two parts. We first establish the high-probability (ϵ,γ)-SSP guarantee, and then bound the total sample complexity.

Part I: Approximate second-order stationarity and iteration complexity. We begin by controlling the stochastic estimation errors uniformly over $k = 0, \dots, K-1$. By Lemma 1, with probability at least $1 - 2K\delta = 0.99$, we have

$$\|\nabla f(x_k) - g_k\| \le \min\Big\{\frac{M\varepsilon}{4\kappa_6(M+\kappa_8)}, \frac{\gamma}{1000\kappa_3}\Big\} \le 1, \qquad k = 0, \dots, K-1. \tag{51}$$

Similarly, by Lemma 2, with probability at least $1 - 2K\delta = 0.99$,

$$\|\nabla^2 f(x_k) - H_k\|^2 \le \frac{1}{800}\min\Big\{\frac{M^3\varepsilon}{(M+\kappa_8)\kappa_9^2}, \frac{\gamma^2}{10000\kappa_4^2}\Big\} \le 1, \qquad k = 0, \dots, K-1. \tag{52}$$

By a union bound, (51)–(52) hold simultaneously for all $k = 0, \dots, K-1$ with probability at least $1 - 4K\delta = 0.98$. In the sequel, we work on this event. Thus, the settings of ρ and κ_5 in (48) satisfy the requirements in Lemma 9.

At each iteration k, Carme-ALM selects either a cubic-regularized step ($Q_k = 1$) or a negative-curvature step ($Q_k = 0$). We analyze the two cases separately.

Case 1: $Q_k = 1$. Invoking Lemma 12 together with (51)–(52) (and the definitions of $\kappa_8, \kappa_9, \eta, \omega$) yields

$$\begin{aligned}
&\mathbb{E}\left[\mathcal{L}_\rho(x_k) - \mathcal{L}_\rho(x_{k+1}) \mid Q_k = 1\right] \\
&\geq \frac{300M\varepsilon^{3/2}}{(M+\kappa_8)^{3/2}}\Pr\left\{\|\nabla\mathcal{L}_\rho(x_{k+1})\| \geq 450\varepsilon\right\} - \frac{9M\varepsilon^{3/2}}{8(M+\kappa_8)^{3/2}} - \frac{3M\varepsilon^{3/2}}{50(M+\kappa_8)^{3/2}} - \frac{M\varepsilon^{3/2}}{15(M+\kappa_8)^{3/2}} \qquad (53)\\
&\geq \frac{300M\varepsilon^{3/2}}{(M+\kappa_8)^{3/2}}\left(\Pr\left\{\|\nabla\mathcal{L}_\rho(x_{k+1})\| \geq 450\varepsilon\right\} - \frac{1}{200}\right),
\end{aligned}$$

where we used $a + b \leq (a^{2/3} + b^{2/3})^{3/2}$ to combine the residual terms.

Case 2: $Q_k = 0$. By Lemma 13 and (52), we have

$$\mathbb{E}\left[\mathcal{L}_\rho(x_k) - \mathcal{L}_\rho(x_{k+1}) \mid Q_k = 0\right] \geq \frac{\gamma^3}{3(L_H^\rho)^2}\Pr\left\{E_k\right\} - \frac{3\gamma^3}{4000(L_H^\rho)^2} \geq \frac{\gamma^3}{3(L_H^\rho)^2}\left(\Pr\left\{E_k\right\} - \frac{1}{400}\right). \qquad (54)$$

Taking expectation over Q_k and combining (53)–(54), we obtain

$$\mathbb{E}[\mathcal{L}_\rho(x_k) - \mathcal{L}_\rho(x_{k+1})] \geq \frac{\gamma^3(1-p)}{3(L_H^\rho)^2}\Big(\Pr\{E_k\} - \frac{1}{400}\Big) + \frac{300M\varepsilon^{3/2}p}{(M+\kappa_8)^{3/2}}\Big(\Pr\{\|\nabla\mathcal{L}_\rho(x_{k+1})\| \geq 450\varepsilon\} - \frac{1}{200}\Big). \qquad (55)$$

Let x_R be sampled uniformly from $\{x_k\}_{k=0}^{K-1}$. Telescoping (55) from $k = 0$ to $K-1$ and using $\mathbb{E}[\mathcal{L}_\rho(x_0) - \mathcal{L}_\rho(x_K)] \leq \Delta$, we obtain (by the same averaging argument as in the unconstrained proof and Lemma 17)

$$\begin{aligned}
\Delta &\geq \mathbb{E}\left[\mathcal{L}_\rho(x_0) - \mathcal{L}_\rho(x_K)\right] \\
&\geq \frac{\gamma^3(1-p)}{3(L_H^\rho)^2}\sum_{k=0}^{K-1}\left(\Pr\left\{E_k\right\} - \frac{1}{400}\right) + \frac{300M\varepsilon^{3/2}p}{(M+\kappa_8)^{3/2}}\sum_{k=1}^{K}\left(\Pr\left\{\|\nabla\mathcal{L}_\rho(x_k)\| \geq 450\varepsilon\right\} - \frac{1}{200}\right) \\
&\geq 1200\Delta\left(\frac{1}{K}\sum_{k=0}^{K-1}\Pr\{E_k\} + \frac{1}{K}\sum_{k=1}^{K}\Pr\left\{\|\nabla\mathcal{L}_\rho(x_k)\| \geq 450\varepsilon\right\} - \frac{3}{400}\right) \\
&\geq 1200\Delta\left(\frac{5}{6(K-1)}\sum_{k=1}^{K-1}\left(\Pr\{E_k\} + \Pr\left\{\|\nabla\mathcal{L}_\rho(x_k)\| \geq 450\varepsilon\right\}\right) - \frac{3}{400}\right) \\
&\geq 1200\Delta\left(\frac{5}{6}\left(\Pr\{E_R\} + \Pr\left\{\|\nabla\mathcal{L}_\rho(x_R)\| \geq 450\varepsilon\right\}\right) - \frac{3}{400}\right),
\end{aligned}$$

where the third inequality follows from Lemma 17, the fourth inequality given by ignoring some (non-negative) terms on the right-hand side and using the fact that $K \geq 6$. Rearranging the terms yields $\Pr\{\mathcal{E}_R\} + \Pr\{\|\nabla\mathcal{L}_\rho(x_R)\| \geq 450\varepsilon\} \leq 0.01$, where $\mathcal{E}_k$ is defined in (47). Hence, it holds that

$$\Pr(\mathcal{E}_R^c \ \wedge \ \{\|\nabla\mathcal{L}_\rho(x_R)\| < 450\varepsilon\}) \geq 0.99, \qquad (56)$$

where $\mathcal{E}_R^c$ is the complement of the event $\mathcal{E}_R$. We now translate (56) into an (ϵ, γ)-SSP guarantee. First, since $\rho \geq \nu^{-1}L_c(L_g^\lambda + 1)$ and by the choice of ε, Lemma 7 implies that on $\{\|\nabla\mathcal{L}_\rho(x_R)\| < 450\varepsilon \leq \epsilon\}$,

$$\|c(x_R)\| \leq \epsilon, \qquad \|\nabla f(x_R) + \nabla c(x_R)\lambda_R\| \leq \|\nabla\mathcal{L}_\rho(x_R)\| + (L_g^\lambda + \rho L_c)\|c(x_R)\| \leq 2\epsilon,$$

where the last inequality uses $\|c(x_R)\| \leq 450\varepsilon \leq \frac{\epsilon}{L_g^\lambda + \rho L_c}$. Second, using the choice of ε again, we have $\|c(x_R)\| \leq \gamma/(\sqrt{m}\,\rho L_H^\lambda)$ and thus

$$\|\nabla^2\mathcal{L}_\rho(x_R) - \tilde{\nabla}^2\mathcal{L}_\rho(x_R)\| \leq \rho\Big\|\sum_{i=1}^{m} c_i(x_R)\nabla^2\lambda_i(x_R)\Big\| + \kappa_3\|\nabla f(x_R) - g_R\| + \kappa_4\|\nabla^2 f(x_R) - H_R\| \leq 2\gamma.$$

And the event $\mathcal{E}_R^c$ implies $u^\top \tilde{\nabla}^2 \mathcal{L}_\rho(x_R)u > -4\gamma$ for all unit $u \in \mathrm{Null}(\nabla c(x_R)^\top)$, so $u^\top \nabla^2 \mathcal{L}_\rho(x_R)u \geq -6\gamma$ on the same null space. This completes the proof of (i) with the overall success probability at least $0.99 \times 0.98 \geq 0.97$ via a union bound.

Iteration complexity. By the parameter choice, the iteration complexity is in order

$$K = \left\lceil \frac{3600\,\Delta(L_H^\rho)^2}{\gamma^3} + \frac{4\Delta(M+\kappa_8)^{3/2}}{M\varepsilon^{3/2}} \right\rceil = O(\gamma^{-3} + \epsilon^{-3/2}), \tag{57}$$

where the last equality is due to the setting of ε in (49).

Part II: Sample complexity. We now bound the total number of stochastic oracle calls, and again separate the two update regimes.

Case 1: $Q_{k-1} = 1$. Since $\|x_k - x_{k-1}\| \leq \eta = \frac{30\varepsilon^{1/2}}{(M+\kappa_8)^{1/2}}$, we have $\|x_k - x_{k-1}\|^2 = O(\varepsilon)$. With $\delta = 1/(200K)$ and the estimator choices in Lemmas 1–2, the batch sizes satisfy

$$B_k^g = \tilde{O}\left(\frac{\bar{\epsilon}_{g,2}\varepsilon}{\bar{\epsilon}_{g,1}^4} + \frac{1}{\bar{\epsilon}_{g,2}}\right), \qquad B_k^h = \tilde{O}\left(\frac{\bar{\epsilon}_{h,2}\varepsilon}{\bar{\epsilon}_{h,1}^4} + \frac{1}{\bar{\epsilon}_{h,2}}\right).$$

Case 2: $Q_{k-1} = 0$. From the negative-curvature update, $\|x_k - x_{k-1}\|^2 = O(\gamma^2)$. Consequently,

$$B_k^g = \tilde{O}\left(\frac{\bar{\epsilon}_{g,2}\gamma^2}{\bar{\epsilon}_{g,1}^4} + \frac{1}{\bar{\epsilon}_{g,2}}\right), \quad B_k^h = \tilde{O}\left(\frac{\bar{\epsilon}_{h,2}\gamma^2}{\bar{\epsilon}_{h,1}^4} + \frac{1}{\bar{\epsilon}_{h,2}}\right).$$

Aggregating over $k = 0, \ldots, K$ and using the expected counts of those steps, we obtain

$$\mathbb{E}\Big[\sum_{k=0}^{K} B_k^g\Big] = \tilde{O}\left(\left(1 + \left(\varepsilon^{-3/2} + \gamma^{-3}\right)^{1/2}\left(\varepsilon^{-1/2} + \gamma^{-1}\right)^{1/2}\right)\left(\varepsilon^{-2} + \gamma^{-2}\right) + \gamma^{-3}\right) = \tilde{O}\left(\epsilon^{-3} + \epsilon^{-2}\gamma^{-2} + \gamma^{-4}\right),$$

$$\mathbb{E}\Big[\sum_{k=0}^{K} B_k^h\Big] = \tilde{O}\left(\left(1 + \left(\varepsilon^{-3/2} + \gamma^{-3}\right)^{1/2}\left(\varepsilon^{-1/2} + \gamma^{-1}\right)^{1/2}\right)\left(\varepsilon^{-1} + \gamma^{-2}\right) + \gamma^{-3}\right) = \tilde{O}\left(\epsilon^{-2} + \gamma^{-4}\right),$$

where mixed terms are absorbed by Young's inequality. This proves (ii) thanks to Markov's inequality (with the same proof as Theorem 1) and completes the proof. □

To address the well-known drawback of quadratic penalty methods—namely, the requirement of an excessively large penalty parameter—we adopt Fletcher's augmented Lagrangian and develop Carme-ALM. This modification leads to a clear improvement: at the level of iteration complexity, Carme-ALM matches that of Carme in the unconstrained setting, while requiring only a bounded penalty parameter to enforce feasibility. However, a gap remains at the level of sample complexity. Importantly, this gap does not merely stem from the omission of higher-order derivative terms. Rather, the fundamental reason lies in the structure of the Fletcher's augmented Lagrangian itself. Specifically, the multiplier $\lambda(x)$ depends explicitly on the (stochastic) gradient of the objective function. As a consequence, stochastic errors in the gradient estimation propagate into the Hessian of the augmented Lagrangian, as quantified in Lemma 8. This coupling effect, the Hessian estimation accuracy must be tightened when a smaller second-order tolerance γ is required. This, in turn, increases the required sample size of stochastic gradients, leading to an extra γ^{-4} term compared with the unconstrained Carme.

Notably, when the second-order tolerance satisfies $\gamma = \Omega(\epsilon)$, this coupling effect becomes inactive, and the sample complexity of Carme-ALM coincides with that of the unconstrained method. The discrepancy arises only when higher second-order accuracy is demanded. These observations suggest that while

Fletcher's augmented Lagrangian successfully enforces feasibility with a bounded penalty parameter and preserves the iteration complexity of Carme, achieving fully identical sample complexity remains challenging due to the intrinsic interaction between stochastic gradients and second-order information in the augmented Lagrangian.

5 Conclusion

In this work, we proposed Carme as a unified framework for stochastic nonconvex optimization with second-order stationarity guarantees. We first considered the unconstrained setting and developed a momentum-based stochastic estimator that simultaneously controls gradient and Hessian noise without periodic checkpoints, leading to improved sample complexity for finding (ϵ, γ)-second-order stationary points. We then studied the extension of Carme to constrained problems. We developed Carme-ALM based on Fletcher's augmented Lagrangian. Our analysis establishes high-probability complexity of Carme-ALM to find an (ϵ, γ)-SSP of the constrained problem. While the iteration complexity matches that of the unconstrained method, the sample complexity exhibits a mild dependence on the second-order tolerance. This effect stems from the interaction between stochastic gradient noise and curvature estimation introduced by the augmented Lagrangian. Under certain conditions the sample complexity order matches the one in the unconstrained settings.

References

[1] Y. Arjevani, Y. Carmon, J. C. Duchi, D. J. Foster, A. Sekhari, and K. Sridharan. Second-order information in non-convex stochastic optimization: Power and limitations. In *CoLT*, pages 242–299. PMLR, 2020.

[2] H. Y. Benson and D. F. Shanno. Interior-point methods for nonconvex nonlinear programming: cubic regularization. *Comput. Optim. Appl.*, 58:323–346, 2014.

[3] A. S. Berahas, F. E. Curtis, D. Robinson, and B. Zhou. Sequential quadratic optimization for nonlinear equality constrained stochastic optimization. *SIAM J. Optim.*, 31(2):1352–1379, 2021.

[4] D. P. Bertsekas. *Constrained optimization and Lagrange multiplier methods*. Academic press, 2014.

[5] D. Boob, Q. Deng, and G. Lan. Stochastic first-order methods for convex and nonconvex functional constrained optimization. *Math. Program.*, 197(1):215–279, 2023.

[6] D. Boob, Q. Deng, and G. Lan. Level constrained first order methods for function constrained optimization. *Math. Program.*, 209(1):1–61, 2025.

[7] Y. Carmon and J. Duchi. Gradient descent finds the cubic-regularized nonconvex newton step. *SIAM J. Optim.*, 29(3):2146–2178, 2019.

[8] Y. Carmon, J. Duchi, O. Hinder, and A. Sidford. Lower bounds for finding stationary points i. *Math. Program.*, 184(1):71–120, 2020.

[9] C. Cartis, N. I. Gould, and P. L. Toint. Optimality of orders one to three and beyond: characterization and evaluation complexity in constrained nonconvex optimization. *J. Complex.*, 53:68–94, 2019.

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

[11] C. Cartis, N. I. M. Gould, and P. L. Toint. Adaptive cubic regularisation methods for unconstrained optimization. Part II: Worst-case function- and derivative-evaluation complexity. *Math. Program.*, 130(2):295–319, 2011.

[12] C. Cartis and K. Scheinberg. Global convergence rate analysis of unconstrained optimization methods based on probabilistic models. *Math. Program.*, 169(2):337–375, 2018.

[13] E. M. Chayti, N. Doikov, and M. Jaggi. Unified convergence theory of stochastic and variance-reduced cubic newton methods. *Transact. Mach. Learn. Res.*, 2024.

[14] E. M. Chayti, N. Doikov, and M. Jaggi. Improving stochastic cubic newton with momentum. In *AISTATS*, pages 1441–1449. PMLR, 2025.

[15] Y. Chi, Y. M. Lu, and Y. Chen. Nonconvex optimization meets low-rank matrix factorization: An overview. *IEEE Trans. Signal Process.*, 67(20):5239–5269, 2019.

[16] Y. Cui, Q. Shi, X. Wang, and X. Xiao. An exact penalty method for stochastic equality-constrained optimization. *Optimization Online*, 2025.

[17] Y. Cui, X. Wang, and X. Xiao. A two-phase stochastic momentum-based algorithm for nonconvex expectation-constrained optimization. *J. Sci. Comput.*, 2024.

[18] F. E. Curtis, M. J. O'Neill, and D. P. Robinson. Worst-case complexity of an sqp method for nonlinear equality constrained stochastic optimization. *Math. Program.*, 205(1):431–483, 2024.

[19] F. E. Curtis, D. P. Robinson, and B. Zhou. Sequential quadratic optimization for stochastic optimization with deterministic nonlinear inequality and equality constraints. *SIAM J. Optim.*, 34(4):3592–3622, 2024.

[20] A. Cutkosky and F. Orabona. Momentum-based variance reduction in non-convex sgd. *NeurIPS*, 32, 2019.

[21] J. E. Dennis Jr and R. B. Schnabel. *Numerical methods for unconstrained optimization and nonlinear equations*. SIAM, 1996.

[22] K. Ding, J. Li, and K. Toh. Nonconvex stochastic bregman proximal gradient method with application to deep learning. *J. Mach. Learn. Res.*, 26(39):1–44, 2025.

[23] N. Doikov, S. U. Stich, and M. Jaggi. Spectral preconditioning for gradient methods on graded non-convex functions. *arXiv:2402.04843*, 2024.

[24] R. Fletcher. An exact penalty function for nonlinear programming with inequalities. *Math. Program.*, 5(1):129–150, 1973.

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

[26] F. Goyens, A. Eftekhari, and N. Boumal. Computing second-order points under equality constraints: revisiting fletcher's augmented lagrangian. *J. Optim. Theory Appl.*, 201(3):1198–1228, 2024.

[27] A. Griewank. The modification of Newton's method for unconstrained optimization by bounding cubic terms. *Technical report*, 1981.

[28] B. M. Idrees, L. Arora, and K. Rajawat. Constrained stochastic recursive momentum successive convex approximation. *arXiv:2404.11790*, 2024.

[29] L. Jin and X. Wang. A stochastic primal-dual method for a class of nonconvex constrained optimization. *Comput. Optim. Appl.*, 83:143–180, 2022.

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

[31] W. Liu and Y. Xu. A single-loop spider-type stochastic subgradient method for expectation-constrained nonconvex nonsmooth optimization. *arXiv:2501.19214*, 2025.

[32] Z. Lu, S. Mei, and Y. Xiao. Variance-reduced first-order methods for deterministically constrained stochastic nonconvex optimization with strong convergence guarantees. *arXiv:2409.09906*, 2024.

[33] Sen Na, Mihai Anitescu, and Mladen Kolar. An adaptive stochastic sequential quadratic programming with differentiable exact augmented lagrangians. *Math. Program.*, 199:721–791, 2023.

[34] Sen Na, Mihai Anitescu, and Mladen Kolar. Inequality constrained stochastic nonlinear optimization via active-set sequential quadratic programming. *Math. Program.*, 202:279–353, 2023.

[35] Y. Nesterov and B. T. Polyak. Cubic regularization of newton method and its global performance. *Math. Program.*, 2006.

[36] J. Nocedal and S. Wright. *Numerical Optimization*. Springer, 2006.

[37] F. Roosta-Khorasani and M. W. Mahoney. Sub-sampled newton methods. *Math. Program.*, 174:293–326, 2019.

[38] Q. Shi and H. Wang. Adaptive directional decomposition methods for nonconvex constrained optimization. *arXiv:2511.03210*, 2025.

[39] Q. Shi, X. Wang, and H. Wang. A momentum-based linearized augmented Lagrangian method for nonconvex constrained stochastic optimization. *Math. Oper. Res.*, 2025.

[40] N. Tripuraneni, M. Stern, C. Jin, J. Regier, and M. I. Jordan. Stochastic cubic regularization for fast nonconvex optimization. *NeurIPS*, 31, 2018.

[41] N. Tripuraneni, M. Stern, C. Jin, J. Regier, and M. I. Jordan. Stochastic cubic regularization for fast nonconvex optimization. *NeurIPS*, pages 2899–2908, 2018.

[42] X. Wang. Complexity analysis of inexact cubic-regularized primal-dual methods for finding second-order stationary points. *Math. Comput.*, 94:2961–3008, 2025.

[43] X. Wang, S. Ma, D. Goldfarb, and W. Liu. Stochastic quasi-newton methods for nonconvex stochastic optimization. *SIAM J. Optim.*, 27(2):927–956, 2017.

[44] Z. Wang, Y. Zhou, Y. Liang, and G. Lan. Stochastic variance-reduced cubic regularization for non-convex optimization. In *AISTATS*, pages 2731–2740. PMLR, 2019.

[45] Y. Watanabe, F. Liao, and Y. Zheng. Policy optimization in robust control: Weak convexity and subgradient methods. *arXiv:2509.25633*, 2025.

[46] Y. Xie and S. J. Wright. Complexity of proximal augmented lagrangian for nonconvex optimization with nonlinear equality constraints. *J. Sci. Comput.*, 86:1–30, 2021.

[47] P. Xu, J. Yang, F. Roosta, Christopher Ré, and Michael W Mahoney. Sub-sampled newton methods with non-uniform sampling. *NeurIPS*, 29, 2016.

[48] D. Zhou and Q. Gu. Stochastic recursive variance-reduced cubic regularization methods. In *AISTATS*, pages 3980–3990. PMLR, 2020.

[49] D. Zhou, P. Xu, and Q. Gu. Sample efficient stochastic variance-reduced cubic regularization method. *arXiv:1811.11989*, 2018.

[50] D. Zhou, P. Xu, and Q. Gu. Stochastic variance-reduced cubic regularization methods. *J. Mach. Learn. Res.*, 20(134):1–47, 2019.

A Proof of Lemma 1

Proof. For each $i \in \mathcal{B}_k^g$, we denote for brevity that $F_i(x) := F(x, \xi_i)$ and let $a_i = \nabla F_i(x_k) - \nabla F_i(x_{k-1}) - \nabla f(x_k) + \nabla f(x_{k-1})$, then we have $\mathbb{E}_i a_i = 0$, a_i i.i.d., and

$$\|a_i\| \le \|\nabla F_i(x_k) - \nabla F_i(x_{k-1})\| + \|\nabla f(x_k) - \nabla f(x_{k-1})\| \le 2L_g^f \|x_k - x_{k-1}\|,$$

where the second inequality holds due to the L_g^f-smoothness of f and F_i. Thus, by vector Azuma-Hoeffding inequality in Lemma 15, we have that with probability at least $1-\delta$,

$$\begin{aligned}
&\left\|\nabla F_{\mathcal{B}_k^g}(x_k) - \nabla F_{\mathcal{B}_k^g}(x_{k-1}) - \nabla f(x_k) + \nabla f(x_{k-1})\right\| \\
&= \frac{1}{B_k^g}\left\| \sum_{i\in\mathcal{B}_k^g} \left[\nabla F_i(x_k) - \nabla F_i(x_{k-1}) - \nabla f(x_k) + \nabla f(x_{k-1})\right]\right\| \le 6L_g^f \sqrt{\frac{\log(1/\delta)}{B_k^g}} \|x_k - x_{k-1}\|,
\end{aligned}$$

where $F_{\mathcal{B}_k^g}(x) := \frac{1}{B_k^g}\sum_{i\in\mathcal{B}_k^g} F_i(x)$. For each $i \in \mathcal{B}_k^g$, we have $\mathbb{E}[\nabla F_i(x_k) - \nabla f(x_k)] = 0$, and $\|\nabla F_i(x_k) - \nabla f(x_k)\| \le \sigma_g$. Thus, using Lemma 15 again, we have that with probability at least $1-\delta$,

$$\|\nabla F_{\mathcal{B}_k^g}(x_k) - \nabla f(x_k)\| = \frac{1}{B_k^g}\left\| \sum_{i\in\mathcal{B}_k^g} \left[\nabla F_i(x_k) - \nabla f(x_k)\right]\right\| \le 3\sigma_g \sqrt{\frac{\log(1/\delta)}{B_k^g}}.$$

It is easy to note that $g_t^0 - \nabla f(x_t) = \sum_{k=0}^t (1-\alpha_g)^k u_{t-k}$, where

$$u_k = \begin{cases} \nabla F_{\mathcal{B}_k^g}(x_k) - \nabla f(x_k) + (1-\alpha_g)\left(\nabla f(x_{k-1}) - \nabla F_{\mathcal{B}_k^g}(x_{k-1})\right), & k > 0, \\ \nabla F_{\mathcal{B}_k^g}(x_k) - \nabla f(x_k), & k = 0. \end{cases}$$

Clearly, we have $\mathbb{E}[u_k|\mathcal{F}_{k-1}] = 0$. For $k = 0$, with probability at least $1-\delta$ it holds that

$$\|u_0\| \le 3\sigma_g \sqrt{\frac{\log(1/\delta)}{B_0^g}} \le \frac{\bar{\epsilon}_{g,1}}{\sqrt{720\log(1/\delta)}}, \tag{58}$$

where the second inequality holds due to the setting of B_k^g in (10). For $k > 0$, conditioned on $\mathcal{F}_{k-1}$ the following inequality holds with probability at least $1-\delta$:

$$\|u_k\| \le 3\alpha_g\sigma_g \sqrt{\frac{\log(1/\delta)}{B_k^g}} + 6(1-\alpha_g)L_g^f \sqrt{\frac{\log(1/\delta)}{B_k^g}} \|x_k - x_{k-1}\| \le \frac{(1-\alpha_g)\alpha_g^{1/2}\bar{\epsilon}_{g,1} + \alpha_g \bar{\epsilon}_{g,2}^{1/2}}{\sqrt{720\log(1/\delta)}}, \tag{59}$$

where the second inequality holds due to the setting of B_k^g in (11). By the union bound, with probability at least $1-K\delta$, (59) and (58) hold for all $0 \le k \le K-1$. Then for given k, by Lemma 15, conditioned on $\mathcal{F}_k$ and the event of (59) and (58) occur, with probability at least $1-\delta$ we have

$$\begin{aligned}\|g_k^0 - \nabla f(x_k)\|^2 &= \left\|\sum_{t=0}^{k}(1-\alpha_g)^t u_{k-t}\right\|^2 \\ \le 9\log(1/\delta)\left(\frac{\bar{\epsilon}_{g,1}^2 + 2\alpha_g^{1/2}\bar{\epsilon}_{g,1}\bar{\epsilon}_{g,2}^{1/2} + \alpha_g\bar{\epsilon}_{g,2}}{720\log(1/\delta)} + \frac{\bar{\epsilon}_{g,1}^2}{720\log(1/\delta)}\right) &\le \frac{3\bar{\epsilon}_{g,1}^2 + 2\alpha_g\bar{\epsilon}_{g,2}}{80},\end{aligned} \tag{60}$$

where the first inequality uses $\sum_{k=0}^{t}(1-\alpha_g)^{2k} \le 1/\alpha_g$. Finally, by the union bound, we have that with probability at least $1-2K\delta$, (60) holds for all $0 \le k \le K-1$. □

B Proof of Lemma 2.

Proof. For each $i \in \mathcal{B}_k^h$, let $A_i = \nabla^2 F_i(x_k) - \nabla^2 f(x_k) + \nabla^2 f(x_{k-1}) - \nabla^2 F_i(x_{k-1})$, then we have $\mathbb{E}_i A_i = 0, A_i^\top = A_i$, A_i i.i.d. and

$$\|A_i\| \le \left\|\nabla^2 F_i(x_k) - \nabla^2 F_i(x_{k-1})\right\| + \left\|\nabla^2 f(x_k) - \nabla^2 f(x_{k-1})\right\| \le 2L_H^f\|x_k - x_{k-1}\|,$$

where the second inequality holds due to L_H^f-Lipschitz continuity of $\nabla^2 f_i$ and $\nabla^2 F$. Then by the matrix Azuma inequality in Lemma 16, we have that with probability at least $1-\delta$,

$$\begin{aligned}&\left\|\nabla^2 F_{\mathcal{B}_k^h}(x_k) - \nabla^2 f(x_k) + \nabla^2 f(x_{k-1}) - \nabla^2 F_{\mathcal{B}_k^h}(x_{k-1})\right\| \\ &= \frac{1}{B_k^h}\left\|\sum_{i\in\mathcal{B}_k^h}\left[\nabla^2 F_i(x_k) - \nabla^2 f(x_k) + \nabla^2 f(x_{k-1}) - \nabla^2 F_i(x_{k-1})\right]\right\| \le 6L_H^f\sqrt{\frac{\log(n/\delta)}{B_k^h}}\|x_k - x_{k-1}\|.\end{aligned}$$

Note that for each $i \in \mathcal{B}_k^h$, $\mathbb{E}[\nabla^2 F_i(x) - \nabla^2 f_i(x)] = 0$ and $\|\nabla^2 F_i(x_k) - \nabla^2 f(x_k)\| \le \sigma_h$. Then using Lemma 16 again, we have that with probability at least $1-\delta$,

$$\|\nabla^2 F_{\mathcal{B}_k^h}(x_k) - \nabla^2 f(x_k)\| = \frac{1}{B_k^h}\left\|\sum_{i\in\mathcal{B}_k^h}\left[\nabla^2 F_i(x_k) - \nabla^2 f(x_k)\right]\right\| \le 3\sigma_h\sqrt{\frac{\log(n/\delta)}{B_k^h}}.$$

Then, we have $H_t - \nabla^2 f(x_t) = \sum_{k=0}^{t}(1-\alpha_h)^k V_{t-k}$, where

$$V_k = \begin{cases}\nabla^2 F_{\mathcal{B}_k^h}(x_k) - \nabla^2 f(x_k) + (1-\alpha_h)\left(\nabla^2 f(x_{k-1}) - \nabla^2 F_{\mathcal{B}_k^h}(x_{k-1})\right), & k > 0, \\ \nabla^2 F_{\mathcal{B}_k^h}(x_k) - \nabla^2 f(x_k), & k = 0.\end{cases}$$

Meanwhile, we have $\mathbb{E}[V_k|\mathcal{F}(V_{k-1}, ..., V_0)] = 0$. Conditioned on $\mathcal{F}_{k-1}$, for $k=0$ with probability at least $1-\delta$, we obtain from (12) that

$$\|V_k\| \le 3\sigma_h\sqrt{\frac{\log(n/\delta)}{B_k^h}} \le \frac{\bar{\epsilon}_{h,1}}{\sqrt{360\log(n/\delta)}}. \tag{61}$$

For $k > 0$, with probability at least $1 - \delta$, the following inequality holds:

$$\begin{aligned}\|V_k\| &\leq 6(1-\alpha_h)L_H^f\sqrt{\frac{\log(n/\delta)}{B_k^h}}\|x_k - x_{k-1}\| + 3\alpha_h\sigma_h\sqrt{\frac{\log(n/\delta)}{B_k^h}} \\ &\leq (1-\alpha_h)\frac{\alpha_h^{1/2}\bar{\epsilon}_h}{\sqrt{360\log(n/\delta)}} + \frac{\alpha_h\bar{\epsilon}_h^{1/2}}{\sqrt{360\log(n/\delta)}},\end{aligned} \tag{62}$$

where the second inequality holds due to (13). By union bound, with probability at least $1 - K\delta$, (61) and (62) hold for all $0 \leq k \leq K-1$. Then for given k, by Lemma 16, conditioned on $\mathcal{F}_k$ and the event of (62) and (61) that occur, with probability at least $1 - \delta$ it holds that

$$\|H_k^0 - \nabla^2 f(x_k)\|^2 = \left\|\sum_{t=0}^{k}(1-\alpha_h)^t V_{k-t}\right\|^2 \leq 9\log(n/\delta)\cdot\frac{2\bar{\epsilon}_{h,1}^2 + 2\alpha_h^{1/2}\bar{\epsilon}_{h,1}\bar{\epsilon}_{h,2}^{1/2} + \alpha_h\bar{\epsilon}_h}{360\log(n/\delta)} = \frac{3\bar{\epsilon}_h + 2\alpha_h\bar{\epsilon}_h}{40}. \tag{63}$$

Finally by union bound, with probability at least $1 - 2K\delta$, (63) holds for all $0 \leq k \leq K-1$. □

C Truncated gradient descent for the cubic subproblem

In this section, we will show that Condition A is easy to achieve when applying a truncated gradient descent algorithm to (15). To be specific, let $s_{k,0} = 0$ we iterate with $s_{k,t+\frac{1}{2}} = s_{k,t} - \tau\nabla m_k(s_{k,t})$, where $\tau > 0$. If $\|s_{k,t+\frac{1}{2}}\| \geq \eta$, terminate with $s_k = s_{k,t+1} = s_{k,t} - \vartheta\nabla m_k(s_{k,t})$ such that $\|s_k\| = \eta$ for a certain $\vartheta \in (0, \tau]$. Else if (16b) holds for $s_{k,t+\frac{1}{2}}$, terminate with $s_k = s_{k,t+1} = s_{k,t+\frac{1}{2}}$. Otherwise, the iteration continues with $s_{k,t+1} = s_{k,t+\frac{1}{2}}$. The full algorithm framework is stated in Algorithm 3 and we provide the convergence result of the proposed truncated gradient descent algorithm in the following lemma.

Algorithm 3: Truncated Gradient Descent for the Cubic Subproblem

Input: Model $m_k(\cdot)$, trust-region radius $\eta > 0$, stepsize $\tau > 0$, tolerance $\omega > 0$
Output: An inexact solution s_k satisfying Condition A
Initialize $s_{k,0} = 0$, $t = 0$;
while *true* **do**
  Compute the gradient $\nabla m_k(s_{k,t})$;
  Set $s_{k,t+\frac{1}{2}} = s_{k,t} - \tau\nabla m_k(s_{k,t})$;
  if $\|s_{k,t+\frac{1}{2}}\| \geq \eta$ **then**
    Find $\vartheta \in (0, \tau]$ such that $\|s_{k,t} - \vartheta\nabla m_k(s_{k,t})\| = \eta$;
    Set $s_k = s_{k,t} - \vartheta\nabla m_k(s_{k,t})$;
    break;
  if $\|\nabla m_k(s_{k,t+\frac{1}{2}})\| \leq M\omega^{2/3}$ **then**
    Set $s_k = s_{k,t+\frac{1}{2}}$;
    break;
  Set $s_{k,t+1} = s_{k,t+\frac{1}{2}}$, $t \leftarrow t+1$;
return s_k

Lemma 14 *Suppose that* $\tau = (L_H^f + \min\{\epsilon^{1/2}M^{1/2}, \gamma\}/20\sqrt{2} + M\eta)^{-1}$. *Then with probability at least* $1 - 2K\delta$, *the truncated gradient descent algorithm stop and output* s_k *satisfying* (16a) *and* (16b) *before*

$$T = \left\lceil \frac{(L_H^f + \min\{\epsilon^{1/2}M^{1/2}, \gamma\}/20\sqrt{2} + M\eta)(2(L_f + \epsilon/4)\eta + (L_H^f + \min\left\{\epsilon^{1/2}M^{1/2}, \gamma\right\}/20\sqrt{2})\eta^2)}{M^2\omega^{4/3}} \right\rceil.$$

Proof. It is worthy to note that for any $k \geq 1$, by Lemma 1 with $\alpha_g = \bar{\epsilon}_g^{1/2} = \epsilon^{1/2}$, it holds with probability at least $1 - 2K\delta$ that $\|\nabla f(x_k) - g_k\|^2 \leq \frac{\epsilon^2}{16}$, $k = 0, \ldots, K-1$. Further, by Lemma 2 with $\alpha_h = \bar{\epsilon}_h^{1/2} = \frac{1}{10}\min\{\epsilon^{1/2}M^{1/2}, \gamma\}$, with probability at least $1 - 2K\delta$ it holds that $\|\nabla^2 f(x_k) - H_k\|^2 \leq \frac{1}{800}\min\left\{\epsilon M, \gamma^2\right\}$, $k = 0, \ldots, K-1$. When $s \leq \eta$, the gradient of m_k is Lipschitz continuous with constant $L_H^f + M\eta$ from the formulation:

$$\|\nabla^2 m_k(s)\| = \left\| H_k + \frac{M}{2}\left(\frac{ss^T}{\|s\|} + \|s\|I\right)\right\| \leq L_H^f + \frac{\min\left\{\epsilon^{1/2}M^{1/2}, \gamma\right\}}{20\sqrt{2}} + M\eta,$$

where $\frac{ss^T}{\|s\|}$ is defined as $\mathbf{0}$ when $s = 0$. Then it follows from the taylor expansion of m_k that

$$m_k(s_{k,t+1}) - m_k(s_{k,t}) \leq \langle \nabla m_k(s_k), s_{k,t+1} - s_{k,t}\rangle + \frac{1}{2}\left(L_H^f + \frac{\min\left\{\epsilon^{1/2}M^{1/2}, \gamma\right\}}{20\sqrt{2}} + M\eta\right)\|s_{k,t+1} - s_{k,t}\|^2. \tag{64}$$

From $\vartheta \in (0, \tau]$ and $\tau = \frac{1}{L_H^f + \min\left\{\epsilon^{1/2}M^{1/2}, \gamma\right\}/20\sqrt{2} + M\eta}$, it follows that the R.H.S. of (64) is non-positive, which together with $m_k(s_{k,0}) = 0$ implies $m_k(s_k) \leq 0$. Thus (16a) holds. From the termination condition of the truncated gradient descent algorithm, it follows that (16b) holds. Then we use the contradiction to show the truncated gradient descent algorithm stops in at most T steps. First we assume that the iteration number of this algorithm is at least $T + 1$. From (64), it indicates that for $0 \leq t \leq T - 1$,

$$\frac{\|\nabla m_k(s_{k,t})\|^2}{2(L_H^f + \min\left\{\epsilon^{1/2}M^{1/2}, \gamma\right\}/20\sqrt{2} + M\eta)} \leq m_k(s_{k,t}) - m_k(s_{k,t+1}).$$

Summing up above inequality over $t = 0, \ldots, T-1$ leads to

$$\begin{aligned}\frac{\sum_{t=0}^{T-1}\|\nabla m_k(s_{k,t})\|^2}{2(L_H^f + \min\left\{\epsilon^{1/2}M^{1/2}, \gamma\right\}/20\sqrt{2} + M\eta)} &\leq m_k(0) - m_k(s_{k,t}) = -m_k(s_{k,t}) \\ &\leq (L_f + \epsilon/4)\eta + \frac{1}{2}(L_H^f + \min\left\{\epsilon^{1/2}M^{1/2}, \gamma\right\}/20\sqrt{2})\eta^2.\end{aligned}$$

Dividing both sides of the above inequality by T and using the definition of T, we obtain

$$\frac{1}{T}\sum_{t=0}^{T-1}\|\nabla m_k(s_{k,t})\|^2 \leq M^2\omega^{4/3}.$$

This implies that there must exist some $t \in [0, T-1]$ such that $\|\nabla m_k(s_{k,t})\|^2 \leq M^2\omega^{4/3}$, which gives a contradiction. Hence, the truncated gradient descent algorithm stops in at most T steps. □

D Supported lemmas

Lemma 15 (Vector Azuma-Hoeffding inequality) *Let $\{v_k\}$ be a vector-valued martingale difference, where $\mathbb{E}[v_k|\mathcal{F}(v_1,...,v_{k-1})] = 0$ and $\|v_k\| \le a_k$. Then with probability at least $1-\delta$ it holds that*

$$\left\| \sum_k v_k \right\| \le 3\sqrt{\log(1/\delta) \sum_k a_k^2}.$$

Lemma 16 (Matrix Azuma inequality) *Let $\{X_k\}$ be a finite adapted sequence of self-adjoint matrices with dimension n, and $\{A_k\}$ be a fixed sequence of self-adjoint matrices that satisfy $\mathbb{E}[X_k|\mathcal{F}(X_{k-1},...,X_1)] = \mathbf{0}$ and $X_k^2 \preceq A_k^2$ almost surely. Then with probability at least $1-\delta$ it holds that*

$$\left\| \sum_k X_k \right\| \le 3\sqrt{\log(n/\delta) \sum_k \|A_k\|^2}.$$

Lemma 17 *Given $\Delta, L_H^f, M, \varepsilon > 0$, $\gamma \in (0, \Delta^{1/3}(L_H^f)^{2/3}]$. The following statements hold true.*

(i) If

$$K = \left\lceil \frac{1440\Delta(L_H^f)^2}{\gamma^3} + \frac{16\Delta\sqrt{M}}{5\varepsilon^{3/2}} \right\rceil \quad \text{and} \quad p = \frac{\sqrt{M}\gamma^3}{\sqrt{M}\gamma^3 + 450(L_H^f)^2\varepsilon^{3/2}},$$

then

$$K(1-p) \ge \frac{1440\Delta(L_H^f)^2}{\gamma^3} \quad \text{and} \quad Kp \ge \frac{16\Delta\sqrt{M}}{5\varepsilon^{3/2}}.$$

(ii) If

$$K = \left\lceil \frac{3600\Delta(L_H^\rho)^2}{\gamma^3} + \frac{4\Delta(M+\kappa_8)^{3/2}}{M\varepsilon^{3/2}} \right\rceil \quad \text{and} \quad p = \frac{(M+\kappa_8)^{3/2}\gamma^3}{(M+\kappa_8)^{3/2}\gamma^3 + 900M(L_H^\rho)^2\varepsilon^{3/2}},$$

then

$$K(1-p) \ge \frac{3600\Delta(L_H^\rho)^2}{\gamma^3} \quad \text{and} \quad Kp \ge \frac{4\Delta(M+\kappa_8)^{3/2}}{M\varepsilon^{3/2}}.$$

Proof. (i) Using the fact that $\lceil x \rceil \ge x$ for any $x \ge 1$ and by the value of K and p we obtain

$$\begin{aligned} K(1-p) &= \left\lceil \frac{1440\Delta(L_H^f)^2}{\gamma^3} + \frac{16\Delta\sqrt{M}}{5\varepsilon^{3/2}} \right\rceil \cdot \left(1 - \frac{\sqrt{M}\gamma^3}{\sqrt{M}\gamma^3 + 450(L_H^f)^2\varepsilon^{3/2}}\right) \\ &\ge \left(\frac{1440\Delta(L_H^f)^2}{\gamma^3} + \frac{16\Delta\sqrt{M}}{5\varepsilon^{3/2}}\right) \cdot \frac{450(L_H^f)^2\varepsilon^{3/2}}{\sqrt{M}\gamma^3 + 450(L_H^f)^2\varepsilon^{3/2}} = \frac{1440\Delta(L_H^f)^2}{\gamma^3}. \end{aligned}$$

The bound on $K \cdot p$ follows similarly. (ii) can also be proved in analogy to (i). □

Lemma 18 *Let $c : \mathbb{R}^n \to \mathbb{R}^m$ be twice continuously differentiable and define $J(x) := \nabla c(x) \in \mathbb{R}^{n\times m}$ and $A(x) := J(x)^\top J(x) \in \mathbb{R}^{m\times m}$. Assume that for all x,*

(i) $\|J(x)\| \le L_c$;

(ii) each component Hessian satisfies $\|\nabla^2 c_j(x)\| \le L_g^c$ for $j = 1,\dots,m$;

(iii) $A(x)$ is nonsingular and $\|A(x)^{-1}\| \le \nu^{-1}$.

Let $v \in \mathbb{R}^n$ be any fixed vector, and define $h(x) := A(x)^{-1}J(x)^\top v$. Then h is Lipschitz differentiable and its Jacobian satisfies $\|\nabla h(x)\| \le \sqrt{m}L_g^c\big(2\nu^{-2}L_c^2 + \nu^{-1}\big)\|v\|$ for any x.

Proof. We use directional derivatives. For any $s \in \mathbb{R}^n$, the directional derivative of h at x along s is

$$Dh(x)[s] = D\big(A(x)^{-1}J(x)^\top v\big)[s] = DP(x)[s]\, v,$$

where we have set $P(x) := A(x)^{-1}J(x)^\top \in \mathbb{R}^{m\times n}$.

By the product rule and the matrix inverse derivative identity $D(A^{-1})(x)[s] = -A(x)^{-1}\big(DA(x)[s]\big)A(x)^{-1}$, we obtain

$$\begin{aligned} DP(x)[s] &= D(A^{-1})(x)[s]\, J(x)^\top + A(x)^{-1}D(J(x)^\top)[s] \\ &= -A(x)^{-1}\big(DA(x)[s]\big)A(x)^{-1}J(x)^\top + A(x)^{-1}D(J(x)^\top)[s]. \end{aligned}$$

Since $A(x) = J(x)^\top J(x)$, another application of the product rule gives $DA(x)[s] = D(J(x)^\top)[s]\, J(x) + J(x)^\top DJ(x)[s]$. Substituting into the expression for $DP(x)[s]$ yields

$$DP(x)[s] = -A(x)^{-1}\Big(D(J(x)^\top)[s]\, J(x) + J(x)^\top DJ(x)[s]\Big)A(x)^{-1}J(x)^\top + A(x)^{-1}D(J(x)^\top)[s].$$

We now bound the operator norm of $DP(x)[s]$. By assumption, $\|A(x)^{-1}\| \le \nu^{-1}$ and $\|J(x)\| \le L_c$. Furthermore, the componentwise Hessian bounds $\|\nabla^2 c_j(x)\| \le L_g^c$ imply that for any $s \in \mathbb{R}^n$, $\big\|\big(DJ(x)[s]\big)_{(:,j)}\big\| \le L_g^c\|s\|$. Let $w \in \mathbb{R}^m$ be arbitrary. Then we have $DJ(x)[s]\, w = \sum_{j=1}^m w_j\big(DJ(x)[s]\big)_{(:,j)}$, and hence

$$\|DJ(x)[s]\, w\| \le \sum_{j=1}^m |w_j|\, \big\|\big(DJ(x)[s]\big)_{(:,j)}\big\| \le \sum_{j=1}^m |w_j|\, L_g^c\, \|s\| = L_g^c\, \|s\|\, \|w\|_1 \le L_g^c\, \|s\|\, \sqrt{m}\, \|w\|,$$

where we used $\|w\|_1 \le \sqrt{m}\, \|w\|_2$ in the last inequality. Taking the supremum over all w with $\|w\| = 1$ yields

$$\|DJ(x)[s]\| = \sup_{\|w\|=1} \|DJ(x)[s]\, w\| \le \sqrt{m}\, L_g^c\, \|s\|.$$

Moreover, since the spectral norm is invariant under transposition, we obtain

$$\|D(J(x)^\top)[s]\| = \|DJ(x)[s]^\top\| = \|DJ(x)[s]\| \le \sqrt{m}\, L_g^c\, \|s\|,$$

which indicates

$$\|D(J(x)^\top)[s]\, J(x)\| \le \|D(J(x)^\top)[s]\|\, \|J(x)\| \le \sqrt{m}L_g^c\|s\|\, L_c,$$

and similarly

$$\|J(x)^\top DJ(x)[s]\| \le \|J(x)^\top\|\, \|DJ(x)[s]\| \le \sqrt{m}L_c\, L_g^c\|s\|.$$

Therefore, we derive

$$\|D(J(x)^\top)[s]\, J(x) + J(x)^\top DJ(x)[s]\| \le 2\sqrt{m}L_cL_g^c\|s\|.$$

Using the submultiplicativity of the operator norm yields

$$\begin{aligned} &\big\| - A(x)^{-1}\big(D(J(x)^\top)[s]\, J(x) + J(x)^\top DJ(x)[s]\big)A(x)^{-1}J(x)^\top\big\| \\ &\le \|A(x)^{-1}\|^2\, \|D(J(x)^\top)[s]\, J(x) + J(x)^\top DJ(x)[s]\|\, \|J(x)^\top\| \le 2\sqrt{m}\nu^{-2}L_c^2L_g^c\|s\|. \end{aligned}$$

Combining the two estimates gives $\|DP(x)[s]\| \le \sqrt{m}L_g^c\big(2\nu^{-2}L_c^2 + \nu^{-1}\big)\, \|s\|$. Finally, since $Dh(x)[s] = DP(x)[s]\, v$, we obtain $\|Dh(x)[s]\| \le \|DP(x)[s]\|\, \|v\|$. Taking the supremum over all s with $\|s\| = 1$ yields

$$\|\nabla h(x)\| = \sup_{\|s\|=1} \|Dh(x)[s]\| \le \sqrt{m}L_g^c\big(2\nu^{-2}L_c^2 + \nu^{-1}\big)\|v\|.$$

The proof is completed. □