

Normalized-step stochastic SQP methods for nonconvex constrained optimization under heavy-tailed noise

Chunhao Han^{*} Xiao Wang[†] Pengxiang Xu[‡] Jin Zhang[§]

October 26, 2026

Abstract

In this paper, we study nonconvex constrained optimization with expectation equality constraints and non-negative bounds under heavy-tailed noise. The objective gradients, constraint values, and constraint Jacobian are accessed through stochastic oracles whose errors have bounded p th central moments for some $p \in (1, 2)$. These conditions allow infinite-variance errors for the three types of stochastic oracles, rendering traditional variance-bounded frameworks inapplicable. To address this challenge, we propose a normalized-step stochastic sequential quadratic programming method, NS-SQP. Normalization recasts the descent analysis in terms of first-moment error bounds while retaining a controlled remainder. To handle possible loss of sampled constraint regularity, we combine local sensitivity estimates on sampling-accuracy events with uniform direction bounds and probability estimates on their complements. This yields first-moment direction perturbation bounds under geometric regularity conditions imposed solely on deterministic reference problems. Combining these estimates with sampled normal-step contraction, we establish the expected ϵ -KKT guarantee conditional on the event that all prescribed iterations are completed, which occurs with high probability. We then derive separate sample-complexity bounds for the variants of NS-SQP with mini-batch estimation and with SPIDER-type recursive estimation under an additional mean p th-moment smoothness condition. We also apply the NS-SQP to semi-stochastic case and expectation equality-only case. Numerical results on a flood insurance risk-control problem, an importance-weighted constrained classification problem and CUTest problems illustrate the promising performance of the proposed methods.

Keywords Nonconvex constrained optimization, sequential quadratic programming, stochastic oracle, heavy-tailed noise, sample complexity

1 Introduction

In this paper, we study nonconvex constrained optimization in the form

$$\begin{aligned} \min_{x \in \mathbb{R}^n} \quad & f(x) := \mathbb{E}_\xi[\mathbf{F}(x; \xi)], \\ \text{s.t.} \quad & c(x) := \mathbb{E}_\zeta[\mathbf{C}(x; \zeta)] = \mathbf{0}, \quad x \geq \mathbf{0}, \end{aligned} \tag{1.1}$$

where ξ and ζ are independent random variables following distributions Ξ and $\mathcal{Z}$, respectively. For almost any $\xi \in \Xi$ and $\zeta \in \mathcal{Z}$, the sampled objective function $\mathbf{F}(\cdot; \xi) : \mathbb{R}^n \rightarrow \mathbb{R}$ and the sampled constraint mapping $\mathbf{C}(\cdot; \zeta) : \mathbb{R}^n \rightarrow \mathbb{R}^m$ are continuously differentiable. For (1.1), at an inquiry point x , we can only get access to stochastic oracles $\nabla \mathbf{F}(x; \xi)$, $\mathbf{C}(x; \zeta)$ and $\nabla \mathbf{C}(x; \zeta)$ with columns being $\nabla \mathbf{C}_i(x; \zeta)$, $i = 1, \dots, m$ by sampling i.i.d. samples ξ and ζ . Most existing algorithms designed to solve the nonconvex stochastic constrained optimization problems typically rely on the bounded-variance assumptions on stochastic oracle errors (or

^{*}Southern University of Science and Technology, and Pengcheng Laboratory, Shenzhen, China (hanchh2024@163.com).

[†]Sun Yat-sen University, Guangzhou, China (wangx936@mail1.sysu.edu.cn).

[‡]Pengcheng Laboratory, Shenzhen, China (xupx@pcl.ac.cn).

[§]Southern University of Science and Technology, Shenzhen, China (zhangj9@sustech.edu.cn).

stronger assumptions, such as almost-surely bounded errors or light-tailed noise). Instead, in our considered setting, the stochastic oracles only possess bounded p th central moments for some $p \in (1, 2)$:

$$\mathbb{E}_\xi [\|\nabla \mathbf{F}(x, \xi) - \nabla f(x)\|^p] \leq V_f^p, \quad \mathbb{E}_\zeta [\|\mathbf{C}(x, \zeta) - c(x)\|^p] \leq V_c^p, \quad \mathbb{E}_\zeta [\|\nabla \mathbf{C}(x, \zeta) - \nabla c(x)\|_F^p] \leq V_J^p, \quad (1.2)$$

where $V_f, V_c, V_J \geq 0$. Then only p th central moments are bounded, allowing oracle errors with infinite variance.

Heavy-tailed uncertainty naturally arises across diverse optimization applications. For example, heavy-tailed models have been used to analyze aggregated claims from the national flood insurance program and assess extreme-loss risk [1]. In off-policy evaluation for recommendation systems, importance weighting can produce heavy-tailed weights when the logging policy lacks sufficient coverage of the target policy, inducing highly variable constraint estimates [2, 3]. Power-law risk factors have also been used in portfolio optimization and insurance risk management to model extreme financial losses [4]. In distributed estimation, impulsive communication noise can introduce large perturbations into the information exchanged between agents, often yielding infinite-variance observations [5]. When such random inputs enter constraint functions and their derivatives, the corresponding stochastic evaluations may have finite means but infinite variances. These practical scenarios motivate the finite p th-moment setting adopted here, which permits the analysis of objective gradient, constraint value, and constraint Jacobian errors without requiring bounded second moments.

Nonconvex stochastic optimization with inequality constraints has received considerable attention in recent years. By introducing non-negative slacks and splitting free variables, general inequality-constrained problems can be reformulated as (1.1) under certain conditions. Boob et al. [6] developed an inexact proximal-point framework for nonconvex inequality-constrained optimization, using the ConEx primal-dual method to approximately solve convex subproblems formed by regularizing the objective and constraints. They further developed a level-constrained stochastic proximal-gradient method in a subsequent work [7], and established $\mathcal{O}(\epsilon^{-4})$ stochastic-gradient sample complexity under strictly feasible initialization and bounded subproblem multipliers. For problems with deterministic equality and inequality constraints, Na et al. [8] developed an active-set SQP method without requiring strict complementarity. Shi et al. [9] developed a momentum-based linearized ALM for deterministic functional constraints and established an $\mathcal{O}(\epsilon^{-4})$ sample complexity bound under extended Mangasarian–Fromovitz Constraint Qualification (MFCQ) and mean-squared smoothness assumption. Recently, Zuo et al. [10] proposed a single-loop stochastic penalty method with adaptive penalty updates. Under a local Linear Independence Constraint Qualification (LICQ) condition imposed on the stationary set of the constraint-violation function, they established a $\tilde{\mathcal{O}}(\epsilon^{-4})$ sample complexity bound for an ϵ -KKT point with high probability. For problems with deterministic equality and non-negativity constraints, Curtis et al. [11] developed a stochastic SQP method based on a two-subproblem construction and established conditional convergence guarantees in expectation. Their analysis uses a uniform contraction condition on linearized infeasibility to establish merit-parameter stability.

For stochastic optimization with expectation constraints, stochastic proximal multiplier methods [12] and adaptive primal-dual methods [13] have provided objective-error and feasibility guarantees in the convex setting. For nonconvex problems, the inexact proximal-point framework with ConEx also provides near-KKT guarantees in expectation, with a total sample complexity bound of $\mathcal{O}(\epsilon^{-6})$ under bounded variance and a strong feasibility condition [6]. Here, ϵ measures stationarity and proximity in norm. Liu and Xu [14] avoided nested constrained subproblem solves by combining a smoothed exact penalty with SPIDER-type recursive estimation of constraint values. Under quantitative uniform Slater-type conditions, unbiased bounded-variance estimates, and mean-square Lipschitz continuity of the sample constraint functions, they established $\mathcal{O}(\epsilon^{-4})$ sample complexity bounds for objective and constraint subgradient evaluations and $\mathcal{O}(\epsilon^{-6})$ for constraint value evaluations, yielding an (ϵ, ϵ) -KKT point in expectation. Moreover, the normal-tangential approach of Shi and Wang [15] attains high-probability sample complexity of $\tilde{\mathcal{O}}(\epsilon^{-4})$, $\tilde{\mathcal{O}}(\epsilon^{-4})$, and $\tilde{\mathcal{O}}(\epsilon^{-6})$ for objective gradients, constraint Jacobians and constraint values, respectively. With recursive estimation, these improve to $\tilde{\mathcal{O}}(\epsilon^{-3})$, $\tilde{\mathcal{O}}(\epsilon^{-3})$ and $\tilde{\mathcal{O}}(\epsilon^{-5})$. Their analysis, however, relies on almost-surely bounded oracle errors to establish sampled constraint regularity, while the recursive variant further assumes sample-wise smoothness.

Much of the theory of stochastic optimization under heavy-tailed noise has been developed for unconstrained problems. Zhang et al. [16] proposed a clipping-based stochastic gradient descent (SGD) method, achieving an $\mathcal{O}(\epsilon^{\frac{3p-2}{1-p}})$ sample complexity under bounded p th central moment assumptions. Combining

clipping with momentum and normalization yields high-probability guarantees [17], while variance reduction improves the sample bound to $\tilde{\mathcal{O}}(\epsilon^{\frac{2p-1}{1-p}})$ under uniform Lipschitz continuity of the sample gradients [18]. Recent work has further clarified the role of gradient normalization under heavy-tailed noise. Hübler et al. [19] analyze normalized SGD with either mini-batch estimation or momentum-based gradient estimation, and show that both variants attain $\mathcal{O}(\epsilon^{\frac{3p-2}{1-p}})$ sample complexity. Batched normalized momentum also admits guarantees under generalized smoothness, including a separate rate when the tail index p is unknown [20]. Sun et al. [21] further study normalization alone and in combination with clipping, and develop variance-reduced and accelerated variants under sample-wise and higher-order smoothness assumptions. Their variance-reduced normalized SGD achieves an expected sample complexity of $\mathcal{O}(\epsilon^{\frac{2p-1}{1-p}})$. Polyak, multi-extrapolated, and recursive momentum mechanisms are further analyzed in [22] under parameter settings independent of Lipschitz constants and noise bounds. Specifically, their analysis of recursive momentum leverages weak average smoothness to establish the sample complexity of $\mathcal{O}(\epsilon^{\frac{2p-1}{1-p}})$. Moreover, Fradin et al. [23] develop lower bounds under relaxed average-smoothness, identify matching expectation guarantees for normalized momentum variance reduction, and introduce a double-clipped variant for high-probability analysis. Without either clipping or normalization, mini-batch SGD achieves a suboptimal sample complexity of $\mathcal{O}(\epsilon^{\frac{2p}{1-p}})$ under standard smoothness [24]. The composite framework of Han et al. [25] can be applied to fixed-parameter penalty formulations under heavy-tailed noise. However, its stationarity guarantees do not directly ensure feasibility or approximate KKT conditions for the original constrained problem.

To the best of our knowledge, heavy-tailed extensions of nonconvex stochastic constrained optimization remain limited. Two notable exceptions are the trust-region SQP methods for deterministic equality-constrained optimization in [26, 27]. Specifically, Fang et al. [26] established high-probability first- and second-order iteration-complexity and sample complexity bounds, subject to noise-dependent accuracy thresholds. Their analysis allows irreducible oracle noise and relaxes the sub-exponential tail assumption on objective-value estimation errors to a bounded $(1 + \delta)$ th-moment condition for some $\delta > 0$. Wang et al. [27] developed an objective-value-free alternative by combining normal-tangential decomposition, normalization, and Polyak momentum. Under unbiased gradient errors with bounded p th central moments, they established almost-sure first-order convergence, but did not provide explicit sample complexity bounds. Both methods use exact constraint values and Jacobians. Our work instead allows simultaneous heavy-tailed errors in objective gradients, constraint values, and Jacobians for expectation equality constraints, with additional non-negativity bounds.

1.1 Main Challenges

The main challenges of solving (1.1) under heavy-tailed noise are twofold.

- (1) In problem (1.1), objective gradients, constraint values, and constraint Jacobians are accessed through stochastic oracle with non-negativity bounds imposed. Constraint-value errors affect the estimated violation and the feasibility-restoration target, while Jacobian errors perturb the tangent spaces and the linearized constraints defining the search directions. With heavy-tailed noise, the original oracle errors may have infinite second moments, rendering traditional variance-bounded frameworks inapplicable. Challenge arises when controlling both step perturbations and feasibility restoration under bounded p th central moments.
- (2) The stochastic SQP frameworks for expectation constraints and non-negativity bounds impose regularity conditions on the sampled subproblems, as in [15]. This, however, can be too restrictive under heavy-tailed noise. Under finite p th moments alone, large Jacobian errors may produce rank-deficient sampled Jacobians or invalidate the inward-pointing directions required by MFCQ. Consequently, multiplier bounds and sensitivity estimates based on uniform sampled regularity are not directly applicable. The analysis therefore needs to relate the geometric conditions on deterministic reference problems to the regularity of sampled subproblems on suitable accuracy events, while accounting for the complementary events in the expected residual bounds.

1.2 Contributions

We develop normalized-step stochastic SQP (NS-SQP) methods for nonconvex constrained optimization problem (1.1) under heavy-tailed noise. These methods allow stochastic estimates of objective gradient, constraint Jacobian and constraint values and adopt normalized-step iterate updates to control the step length, thereby suppressing anomalous sampled total steps induced by heavy-tailed errors. Crucially, this normalization enables descent analysis via first-moment error bounds with controlled remainders, thereby accommodating oracle errors with bounded p th central moments. Combining this mechanism with an analysis of total step perturbations and feasibility restoration, we establish the expected ϵ -KKT guarantees under bounded p th central moments (see Figure 1). Although an early termination may be triggered, we can ensure the expected ϵ -KKT guarantee conditioned on a high-probability event. We further derive separate sample complexity for NS-SQPs with mini-batch estimation and with SPIDER-type recursive estimation (see Table 1). We also specialize our methods to the semi-stochastic variant of (1.1) where the equality constraints are deterministic and to problems with only expectation equality constraints, reporting the corresponding sample complexity bounds for finding an expected ϵ -KKT point. We finally report numerical experiment results on three test examples.

Compared with [26, 27], we allow stochastic evaluations of equality-constraint values and Jacobians and additional non-negative constraints. Beyond the bounded-variance setting in [11, 15, 28, 29, 30, 31, 32, 33], our analysis assumes bounded p th central moments, allowing for potential infinite-variance errors. As $p \uparrow 2$, the sample-complexity bounds in semi-stochastic case match those in the literature such as [9, 10]. Unlike the stochastic framework of [15], which uses uniform sampled-Jacobian regularity and sampled MFCQ, we construct subproblems with unique solutions even for rank-deficient sampled Jacobians and impose a projected feasibility error bound on the deterministic feasibility problem and near-active LICQ on the deterministic total-step subproblem. Overcoming the challenges aforementioned in theoretical analysis, we establish sampled normal-step contraction under sufficient sampling accuracy and derive a total-step perturbation bound in expectation by combining local sensitivity estimates with uniform total-step bounds. We further bound the non-progressive probabilities and explicitly account for the contributions of inaccurate samples to the expected KKT residual bounds. More comparison of sample complexity bounds with existing works are provided in Discussions 1 to 3.

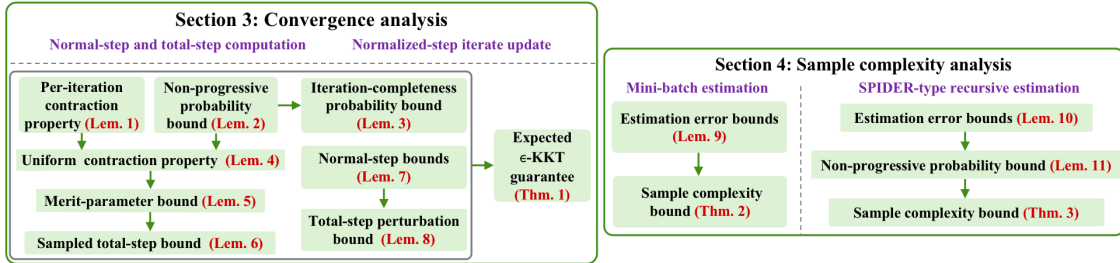


Figure 1: Workflow of the main theoretical analysis.

Estimation Type	Fully-stochastic (1.1) ($\tilde{\nabla}f; \tilde{\nabla}c; \tilde{c}$) ($1 < p < 2$)	Semi-stochastic (4.12) ($\tilde{\nabla}f$) ($1 < p < 2$)	Equality-only (4.16) ($\tilde{\nabla}f; \tilde{\nabla}c; \tilde{c}$) ($1 < p < 2$)
Mini-batch	$\mathcal{O}(\epsilon^{\frac{3p-2}{1-p}}); \mathcal{O}(\epsilon^{\frac{3p+1}{1-p}}); \mathcal{O}(\epsilon^{\frac{3p+1}{1-p}})$	$\mathcal{O}(\epsilon^{\frac{3p-2}{1-p}})$	$\mathcal{O}(\epsilon^{\frac{3p-2}{1-p}}); \mathcal{O}(\epsilon^{\frac{3p-2}{1-p}}); \mathcal{O}(\epsilon^{\frac{4p-2}{1-p}})$
SPIDER-type Recursive	$\mathcal{O}(\epsilon^{\frac{2p-1}{1-p}}); \mathcal{O}(\epsilon^{\frac{2p+1}{1-p}}); \mathcal{O}(\epsilon^{\frac{2p+1}{1-p}})$	$\mathcal{O}(\epsilon^{\frac{2p-1}{1-p}})$	$\mathcal{O}(\epsilon^{\frac{2p-1}{1-p}}); \mathcal{O}(\epsilon^{\frac{2p-1}{1-p}}); \mathcal{O}(\epsilon^{\frac{3p-1}{1-p}})$

¹ The sample complexity for the fully-stochastic case is evaluated conditioned on the event that all prescribed iterations are completed, which occurs with probability at least $1 - \epsilon$.

Table 1: Sample complexity of the NS-SQP methods.

1.3 Notations, preliminaries and organization

For a positive integer T , define $[T] := \{1, \dots, T\}$. For vectors $x, y \in \mathbb{R}^d$, we use $\langle x, y \rangle := x^\top y$ and $\|x\| := \sqrt{\langle x, x \rangle}$ to denote the Euclidean inner product and norm, respectively. For a matrix $A \in \mathbb{R}^{q \times d}$, $\|A\| := \max_{\|x\|=1} \|Ax\|$ denotes the spectral norm, and $\|A\|_F := \sqrt{\text{tr}(A^\top A)}$ denotes the Frobenius norm. Let $\mathbf{0}$ and I denote the zero vector and the identity matrix of appropriate dimensions, respectively. For an event $\mathcal{A}$, $\mathbb{I}_{\mathcal{A}}$ denotes its indicator function, equal to 1 when $\mathcal{A}$ occurs and 0 otherwise, and $\mathcal{A}^c$ denotes the complementary event of $\mathcal{A}$. Let $\mathcal{F}_t$ denote the history prior to the sampling at iteration t , and write $\mathbb{E}_t[\cdot] := \mathbb{E}[\cdot \mid \mathcal{F}_t]$ and $\mathbb{P}_t(\cdot) := \mathbb{P}(\cdot \mid \mathcal{F}_t)$. For $r > 0$, we use $\mathbb{B}(r)$ to denote the closed ball of radius r centered at the origin in either $\mathbb{R}^d$ or $\mathbb{R}^{n \times m}$: $\mathbb{B}(r) := \{u \in \mathbb{R}^d : \|u\| \leq r\}$ or $\mathbb{B}(r) := \{A \in \mathbb{R}^{n \times m} : \|A\|_F \leq r\}$, with the ambient space understood from context. We denote $\Pi_{\mathbb{B}}(\cdot)$ as the projection onto a set $\mathbb{B}$ w.r.t. the Euclidean norm for vectors and Frobenius norm for matrices. Let $\mathcal{O}(\cdot)$ denote an asymptotic upper bound up to constant factors, while $\tilde{\mathcal{O}}(\cdot)$ additionally suppresses logarithmic factors.

Throughout our analysis, we make the following common assumptions on problem (1.1). Let $\mathcal{X}$ be an open convex set containing all iterates generated by the proposed algorithm.

Assumption 1 *The functions $f(x)$ and $c(x)$ are continuously differentiable. The objective function $f(x)$ is lower bounded by $f^* > -\infty$. Moreover, there exist positive constants M_f and M_c such that $\|\nabla f(x)\| \leq M_f$ and $\|c(x)\| \leq M_c$ for any $x \in \mathcal{X}$.*

Assumption 2 *The functions ∇f , c and ∇c are respectively L_f -, L_c - and L_J -Lipschitz continuous over $\mathcal{X}$ w.r.t. $\|\cdot\|$, where $L_f, L_c, L_J > 0$.*

Without loss of generality, we assume that the constant L_c is chosen sufficiently large enough to ensure $\|\nabla c(x)\|_F \leq L_c$ for any $x \in \mathcal{X}$. Assumptions 1 and 2 impose boundedness and the smoothness conditions commonly used in stochastic SQP analyses [11, 26, 29, 34, 35, 36].

Assumption 3 *For any $x \in \mathcal{X}$, stochastic oracles satisfy $\mathbb{E}_\xi[\nabla \mathbf{F}(x, \xi)] = \nabla f(x)$, $\mathbb{E}_\zeta[\mathbf{C}(x, \zeta)] = c(x)$, $\mathbb{E}_\zeta[\nabla C(x, \zeta)] = \nabla c(x)$, and there exist constants $V_f, V_c, V_J > 0$ such that (1.2) holds.*

Compared with the commonly used variance-bound condition (i.e., $p = 2$ in (1.2)), the p th-moment boundedness with $p \in (1, 2)$ is strictly weaker. It allows oracle errors with finite p th moments and infinite second moments. This extension changes the integrability regime rather than merely the magnitude of the noise variance. In the present setting, stochastic errors also perturb the linearized constraints and the resulting restoration and total steps. The analysis must therefore combine p th-moment error control with the deterministic geometric conditions to establish sampled constraint reduction and stability of the computed directions.

Our primary goal of this paper is to propose algorithms to pursue an approximate KKT point of (1.1).

Definition 1 *For a prescribed tolerance $\epsilon > 0$, a point $x \geq \mathbf{0}$ is called an ϵ -KKT point of (1.1), if there exist associated multipliers $\lambda \in \mathbb{R}^m$ and $\mu \in \mathbb{R}_{\geq \mathbf{0}}^n$ such that*

$$\|\nabla f(x) + \nabla c(x)\lambda - \mu\| \leq \epsilon, \quad \|c(x)\| \leq \epsilon, \quad |\mu^\top x| \leq \epsilon. \quad (1.3)$$

If (1.3) holds in expectation, we call x an expected ϵ -KKT point of (1.1).

In the presence of non-negativity constraints $x \geq \mathbf{0}$, the three conditions in (1.3) quantify approximate stationarity, equality feasibility, and complementary slackness, respectively. However, non-negativity constraints $x \geq \mathbf{0}$ may prevent further reduction of the equality constraint violation. In such cases, the feasibility problem $\min_{x \in \mathbb{R}_{\geq \mathbf{0}}^n} \frac{1}{2} \|c(x)\|^2$ may become approximately stationary while $\|c(x)\|$ remains bounded away from zero. For a prescribed tolerance $\epsilon \geq 0$, a point $x \in \mathbb{R}_{\geq \mathbf{0}}^n$ is called an ϵ -IS point of (1.1) if there exists $z \in \mathbb{R}_{\geq \mathbf{0}}^n$ such that

$$\|c(x)\| > \epsilon, \quad \|\nabla c(x)c(x) - z\| \leq \epsilon, \quad |x^\top z| \leq \epsilon.$$

Let $\{x_t\}$ be an iterate sequence generated by an algorithm for solving (1.1). For iterate x_t , use abbreviations $\nabla f_t := \nabla f(x_t)$, $c_t := c(x_t)$ and $\nabla c_t := \nabla c(x_t)$ with corresponding stochastic estimators denoted by $\tilde{\nabla} f_t$, $\tilde{c}_t$ and $\tilde{\nabla} c_t$. We impose the following Conditions 1 and 2 on these stochastic estimators. We use these two conditions throughout the analysis in Section 3. In Section 4 we will propose robust estimator constructions to guarantee those estimator-error bounds required in Conditions 1 and 2.

Condition 1 *Assumptions 1 and 2 hold. And the stochastic estimators satisfy $\|\tilde{\nabla} f_t\| \leq M_f$, $\|\tilde{c}_t\| \leq M_c$ and $\|\tilde{\nabla} c_t\|_F \leq L_c$, for all $t \geq 0$.*

Condition 2 *The expected estimation errors are upper bounded:*

$$\mathbb{E}[\|\tilde{\nabla} f_t - \nabla f_t\|] \leq E_{t,f}, \quad \mathbb{E}[\|\tilde{c}_t - c_t\|] \leq E_{t,c}, \quad \mathbb{E}[\|\tilde{\nabla} c_t - \nabla c_t\|] \leq E_{t,J}. \quad (1.4)$$

The rest of this paper is organized as follows. Section 2 presents the normalized-step stochastic SQP methods for problem (1.1). We analyze their convergence properties and establish the expected ϵ -KKT guarantees in Section 3. Section 4 establishes the sample complexity of the algorithm variants with mini-batch estimation and SPIDER-type recursive estimation, as well as the results when adapted to the semi-stochastic and expectation equality-only cases. Section 5 presents three numerical experiments demonstrating the effectiveness and comparative performances of the proposed methods under heavy-tailed noise. Section 6 concludes the paper.

2 Algorithm framework

In this section, we will propose an algorithm framework for normalized-step stochastic SQP methods for the problem (1.1) under heavy-tailed noise. The presence of the non-negativity constraints $x \geq \mathbf{0}$ complicates both the construction of search directions and the optimality conditions. Motivated by the bound-constrained SQP structure of [11, 15], to balance objective minimization and equality-constraint restoration in the presence of heavy-tailed noise and the non-negativity constraints, we adopt a two-step directional construction. The normal step determines a linearized feasibility target, while the total step minimizes a quadratic objective model subject to this target and the non-negativity bounds. However, unlike [11, 15], we consider problems whose objective gradient, constraint Jacobian and constraint values are estimated under heavy-tailed noise. Such potentially larger-scale errors may distort both objective reduction and constraint restoration.

At current iterate $x_t \geq \mathbf{0}$, after drawing independent stochastic samples to obtain estimators $\tilde{\nabla} f_t$, $\tilde{c}_t$, and $\tilde{\nabla} c_t$, we first compute the normal direction $\tilde{v}_t \in \mathbb{R}^n$. This direction aims to minimize the linearized constraint violation and is obtained by solving the following regularized, strongly convex normal-step subproblem:

$$\min_{v \in \mathbb{R}^n} \quad \frac{1}{2} \|\tilde{c}_t + \tilde{\nabla} c_t^\top v\|^2 + \frac{\gamma_t}{2} \|v\|^2 \quad \text{s.t.} \quad x_t + v \geq \mathbf{0}, \quad (2.1)$$

where γ_t is a regularization parameter satisfying $0 < \gamma_{\min} \leq \gamma_t \leq \gamma_{\max}$. This formulation directly regularizes the full step, avoiding the auxiliary decomposition variables used in [11, 15] and ensuring a unique solution even when the sampled Jacobian is rank deficient. The normal step balances the reduction of the sampled linearized constraint violation and the control of the step norm, thereby providing a feasible restoration target $\tilde{\nabla} c_t^\top \tilde{v}_t$ for the subsequent total-step subproblem. However, the non-negativity constraints may prevent further reduction of the sampled equality-constraint violation. Given a preset small tolerance $\epsilon' > 0$, when $\|\tilde{c}_t\| > \epsilon'$, we test for the sampled infeasible stationarity using the projected gradient residual of the normal-step subproblem (2.1) at $v = \mathbf{0}$. Specifically, let

$$\tilde{v}_t^c := \Pi_{\mathbb{R}_{\geq \mathbf{0}}^n}(x_t - \tilde{\nabla} c_t \tilde{c}_t) - x_t. \quad (2.2)$$

At iterations with $\|\tilde{c}_t\| > \epsilon'$ and if $\|\tilde{v}_t^c\| \leq \epsilon' \beta_{\text{IS}}$, we terminate the algorithm and return x_t as a sampled ϵ' -IS point, where β_{IS} is a stopping threshold. Subsequently, the algorithm determines the total search direction $\tilde{d}_t \in \mathbb{R}^n$ by minimizing a local quadratic approximation of the objective function. We restrict this search to directions that strictly preserve the constraint improvement identified by the normal step. We solve the following total-step subproblem to obtain $\tilde{d}_t$:

$$\min_{d \in \mathbb{R}^n} \quad \tilde{\nabla} f_t^\top d + \frac{1}{2} d^\top H_t d \quad \text{s.t.} \quad \tilde{\nabla} c_t^\top d = \tilde{\nabla} c_t^\top \tilde{v}_t, \quad x_t + d \geq \mathbf{0}, \quad (2.3)$$

where H_t is a symmetric and positive definite matrix satisfying $\sigma_l I \leq H_t \leq \sigma_u I$ for any $t \geq 0$ with two positive constants $\sigma_u \geq \sigma_l$. This subproblem is inherently well-posed and feasible, as the previously computed $d = \tilde{v}_t$ is always a trivial feasible solution.

From the computation of $\tilde{d}_t$, we can see that the constraint restoration is promoted through the normal-step subproblem (2.1). Due to $\tilde{\nabla} c_t^\top \tilde{d}_t = \tilde{\nabla} c_t^\top \tilde{v}_t$, the predicted reduction achieved by $\tilde{d}_t$ through $\|\tilde{c}_t\| - \|\tilde{c}_t + \tilde{\nabla} c_t^\top \tilde{d}_t\|$ is equal to $\|\tilde{c}_t\| - \|\tilde{c}_t + \tilde{\nabla} c_t^\top \tilde{v}_t\|$.

Proposition 1 Suppose Condition 1 holds. At iterate $x_t \geq \mathbf{0}$, the normal step $\tilde{v}_t$ obtained from subproblem (2.1) guarantees a non-negative predicted constraint reduction, satisfying

$$\Delta\tilde{c}_t := \|\tilde{c}_t\| - \|\tilde{c}_t + \tilde{\nabla}c_t^\top \tilde{v}_t\| \geq \frac{\gamma_t}{2M_c} \|\tilde{v}_t\|^2. \quad (2.4)$$

Moreover, $\Delta\tilde{c}_t = 0$ implies $\tilde{v}_t = \mathbf{0}$ and $\tilde{\nabla}f_t^\top \tilde{d}_t + \frac{1}{2}\tilde{d}_t^\top H_t \tilde{d}_t \leq 0$.

Proof. By the design of the normal-step subproblem (2.1), the zero vector is always feasible since $x_t \geq \mathbf{0}$. The optimality of $\tilde{v}_t$ gives $\frac{1}{2}\|\tilde{c}_t + \tilde{\nabla}c_t^\top \tilde{v}_t\|^2 + \frac{\gamma_t}{2}\|\tilde{v}_t\|^2 \leq \frac{1}{2}\|\tilde{c}_t\|^2$. Rearranging the terms yields

$$\gamma_t \|\tilde{v}_t\|^2 \leq \|\tilde{c}_t\|^2 - \|\tilde{c}_t + \tilde{\nabla}c_t^\top \tilde{v}_t\|^2 = (\|\tilde{c}_t\| - \|\tilde{c}_t + \tilde{\nabla}c_t^\top \tilde{v}_t\|) (\|\tilde{c}_t\| + \|\tilde{c}_t + \tilde{\nabla}c_t^\top \tilde{v}_t\|). \quad (2.5)$$

Since this optimality also implies $\|\tilde{c}_t + \tilde{\nabla}c_t^\top \tilde{v}_t\| \leq \|\tilde{c}_t\|$, the second factor on the right-hand side of (2.5) is bounded by $2\|\tilde{c}_t\| \leq 2M_c$. Substituting the definition of $\Delta\tilde{c}_t$ yields $\gamma_t \|\tilde{v}_t\|^2 \leq 2M_c \Delta\tilde{c}_t$, which leads to (2.4). Moreover, if $\Delta\tilde{c}_t = 0$, inequality (2.4) together with $\gamma_t > 0$ implies $\tilde{v}_t = \mathbf{0}$. Since $\tilde{d}_t$ is an optimal solution of (2.3), comparison with the feasible point $d = \mathbf{0}$ gives $\tilde{\nabla}f_t^\top \tilde{d}_t + \frac{1}{2}\tilde{d}_t^\top H_t \tilde{d}_t \leq \tilde{\nabla}f_t^\top \mathbf{0} + \frac{1}{2}\mathbf{0}^\top H_t \mathbf{0} = 0$. $\square$

To measure the algorithm's progress during the iteration process, we use the merit function

$$\Phi_\rho(x) := f(x) + \rho\|c(x)\|, \quad \rho > 0. \quad (2.6)$$

The non-negativity bounds are enforced directly by the subproblems and the iterate update. Proposition 1 motivates the following selection strategy of the merit parameter ρ . If $\tilde{\nabla}f_t^\top \tilde{d}_t + \frac{1}{2}\tilde{d}_t^\top H_t \tilde{d}_t \leq 0$, then the local quadratic objective model is already nonincreasing along $\tilde{d}_t$. Thus, no increase of the merit parameter is required. On the other hand, if $\tilde{\nabla}f_t^\top \tilde{d}_t + \frac{1}{2}\tilde{d}_t^\top H_t \tilde{d}_t > 0$, Proposition 1 implies $\Delta\tilde{c}_t > 0$ by contrapositive. Hence, the merit parameter can be chosen sufficiently large so that the predicted constraint reduction compensates for the increase in the local quadratic objective model. Specifically, we require

$$\tilde{\nabla}f_t^\top \tilde{d}_t + \frac{1}{2}\tilde{d}_t^\top H_t \tilde{d}_t \leq (1 - \sigma_\rho)\tilde{\rho}_t \Delta\tilde{c}_t, \quad \text{if } \tilde{\nabla}f_t^\top \tilde{d}_t + \frac{1}{2}\tilde{d}_t^\top H_t \tilde{d}_t > 0,$$

where $\sigma_\rho \in (0, 1)$ is a predefined parameter. The factor $(1 - \sigma_\rho)$ ensures that only a portion of the weighted constraint reduction is used to compensate for the objective-model increase, leaving $\sigma_\rho \tilde{\rho}_t \Delta\tilde{c}_t$ as a decrease margin in the merit-function analysis. This motivates the following trial merit parameter:

$$\tilde{\rho}_t^{\text{trial}} := \begin{cases} \frac{\tilde{\nabla}f_t^\top \tilde{d}_t + \frac{1}{2}\tilde{d}_t^\top H_t \tilde{d}_t}{(1 - \sigma_\rho)\Delta\tilde{c}_t}, & \text{if } \tilde{\nabla}f_t^\top \tilde{d}_t + \frac{1}{2}\tilde{d}_t^\top H_t \tilde{d}_t > 0, \\ 0, & \text{otherwise.} \end{cases} \quad (2.7)$$

Note that the first branch is well defined because $\Delta\tilde{c}_t > 0$ whenever its numerator is positive. The merit parameter is then updated according to

$$\tilde{\rho}_t := \begin{cases} \tilde{\rho}_{t-1}, & \text{if } \tilde{\rho}_{t-1} \geq \tilde{\rho}_t^{\text{trial}}, \\ \max \left\{ \frac{\tilde{\rho}_{t-1}}{1 - \epsilon_\rho}, \tilde{\rho}_t^{\text{trial}} \right\}, & \text{if } \tilde{\rho}_{t-1} < \tilde{\rho}_t^{\text{trial}}, \end{cases} \quad (2.8)$$

where $\epsilon_\rho \in (0, 1)$ is a predefined parameter and $\tilde{\rho}_0 > 0$ is the initial merit parameter. The trial value $\tilde{\rho}_t^{\text{trial}}$ provides the threshold for sufficient decrease in the sampled linearized merit model, whereas the multiplicative safeguard ensures that strict increase of the merit parameter $\tilde{\rho}_t$ has a fixed relative size. The latter property will be used to prove that $\tilde{\rho}_t$ can be increased only finitely many times once a uniform upper bound on $\tilde{\rho}_t^{\text{trial}}$ has been established. Note that we adopt a similar update structure as [11], but with sampled constraint information, particularly with the reduction $\Delta\tilde{c}_t$ produced by the sampled normal step.

To mitigate extreme direction fluctuations caused by heavy-tailed noise, we employ a normalized-step iterate update:

$$x_{t+1} = x_t + \eta_t \tilde{d}_t, \quad \text{with } \eta_t = \min \left\{ \frac{\kappa_t}{(L_f + \tilde{\rho}_t L_J)\|\tilde{d}_t\|}, 1 \right\}, \quad (2.9)$$

where $\kappa_t > 0$ is a predefined parameter, with the convention that $1/\|\tilde{d}_t\| = +\infty$ when $\tilde{d}_t = \mathbf{0}$. The condition $x_t + \tilde{d}_t \geq \mathbf{0}$ combined with $x_t \geq \mathbf{0}$ ensures $x_{t+1} \geq \mathbf{0}$ under the stepsize restriction $\eta_t \leq 1$. The normalization by $\|\tilde{d}_t\|$ ensures that $\|x_{t+1} - x_t\| = \eta_t \|\tilde{d}_t\| \leq \kappa_t / (L_f + \tilde{\rho}_t L_J)$, thereby limiting the actual displacement even when the stochastic direction has a large norm. This bound also controls the quadratic Taylor remainders by deterministic multiples of κ_t^2 and the error-step products by κ_t times the corresponding error norms. These estimates allow the subsequent descent analysis to bound the expected step norm using first-moment estimation errors derived from the finite p th-moment oracle assumption, without requiring bounded variance of the oracles. We now present the NS-SQP method for (1.1) under heavy-tailed noise in Algorithm 1.

Algorithm 1 NS-SQP method for problem (1.1) under heavy-tailed noise

Input: Initial variable $x_1 \in \mathbb{R}_{\geq \mathbf{0}}^n$; sequence $\{\kappa_t\} \subset \mathbb{R}_{>0}$, merit-parameter $\tilde{\rho}_0 > 0$ and parameters $\epsilon' > 0$; stopping threshold $0 < \beta_{\text{IS}} \leq 1/\max\{1, L_c M_c\}$; symmetric positive-definite matrices $\{H_t\}$; and iteration horizon T .

Output: The sampled ϵ' -IS point or x_R with R drawn independently from $[T]$ and $\mathbb{P}(R=t) = \kappa_t/K_T$ with $K_T := \sum_{k=1}^T \kappa_k$.

```

1: for  $t = 1, 2, \dots, T$  do
2:   Construct the stochastic estimators  $\tilde{\nabla} f_t, \tilde{c}_t$  and  $\tilde{\nabla} c_t$ . // Stochastic Estimation
3:   if  $\|\tilde{c}_t\| \leq \epsilon'$  then
4:     Set  $\tilde{v}_t = \mathbf{0}$  and  $\Delta \tilde{c}_t = 0$ .
5:   else
6:     Compute the projected gradient residual  $\tilde{v}_t^c := \Pi_{\mathbb{R}_{\geq \mathbf{0}}^n}(x_t - \tilde{\nabla} c_t \tilde{c}_t) - x_t$ .
7:     if  $\|\tilde{v}_t^c\| \leq \epsilon' \beta_{\text{IS}}$  then
8:       Terminate and return  $x_t$  as a sampled  $\epsilon'$ -IS point. // Early Termination Test
9:     else
10:      Compute the normal step  $\tilde{v}_t$  by solving (2.1). // Constraint Restoration
11:    end if
12:  end if
13:  Compute the total search direction  $\tilde{d}_t$  by solving (2.3). // Total Step Computation
14:  Update the merit parameter  $\tilde{\rho}_t$  by (2.8). // Merit Parameter Update
15:  Update the iterate  $x_{t+1}$  through (2.9). // Normalized-step Update
16: end for

```

Notice that Algorithm 1 may encounter early termination. Suppose the early termination test is triggered at iteration $t \in [T]$. Then $\|\tilde{c}_t\| > \epsilon'$ and $\|\tilde{v}_t^c\| \leq \epsilon' \beta_{\text{IS}} \leq \frac{\epsilon'}{\max\{1, L_c M_c\}}$. Define the projection $p_t := \Pi_{\mathbb{R}_{\geq \mathbf{0}}^n}(x_t - \tilde{\nabla} c_t \tilde{c}_t)$ and the dual multiplier $z_t := p_t - x_t + \tilde{\nabla} c_t \tilde{c}_t$. By the optimality conditions of the Euclidean projection onto the non-negative orthant, it holds that $z_t \geq \mathbf{0}$, $p_t \geq \mathbf{0}$ and $p_t^\top z_t = 0$. Consequently, we have

$$\|\tilde{\nabla} c_t \tilde{c}_t - z_t\| = \|x_t - p_t\| = \|\tilde{v}_t^c\| \leq \frac{\epsilon'}{\max\{1, L_c M_c\}} \leq \epsilon'.$$

For projection onto the non-negative orthant, z_t can be expressed component-wise as $z_{t,i} = \max\{0, (\tilde{\nabla} c_t \tilde{c}_t)_i - x_{t,i}\}$. Since $x_t \geq \mathbf{0}$, we have $0 \leq z_{t,i} \leq \max\{0, (\tilde{\nabla} c_t \tilde{c}_t)_i\} \leq |(\tilde{\nabla} c_t \tilde{c}_t)_i|$. Therefore, using the uniform estimator bounds yields $\|z_t\| \leq \|\tilde{\nabla} c_t \tilde{c}_t\| \leq L_c M_c$. To bound the complementarity error $x_t^\top z_t$, we substitute $x_t = p_t + (\tilde{\nabla} c_t \tilde{c}_t - z_t)$ and apply the exact complementarity condition $p_t^\top z_t = 0$ to obtain

$$0 \leq x_t^\top z_t = p_t^\top z_t + (x_t - p_t)^\top z_t = (x_t - p_t)^\top z_t \leq \|x_t - p_t\| \|z_t\| = \|\tilde{v}_t^c\| \|z_t\| \leq \epsilon' \beta_{\text{IS}} L_c M_c \leq \epsilon'.$$

Let $\mathcal{S}_t$ denote the event that Algorithm 1 does not trigger the early termination test at current iterate t and let $\bar{\mathcal{S}}$ denote the intersection of all $\mathcal{S}_t$, $t = 1, \dots, T$, respectively defined by

$$\mathcal{S}_t := \{\|\tilde{c}_t\| \leq \epsilon'\} \cup \{\|\tilde{v}_t^c\| > \epsilon' \beta_{\text{IS}}\}, \quad t = 1, \dots, T, \quad \text{and} \quad \bar{\mathcal{S}} := \bigcap_{t=1}^T \mathcal{S}_t. \quad (2.10)$$

Then $\bar{\mathcal{S}}$ represents the event that no early termination is triggered, thus all T iterations are completed.

3 Convergence analysis

In this section, we will explore the convergence properties of Algorithm 1, and establish the expected ϵ -KKT guarantee at the output point. Figure 1 presents the workflow of the theoretical analysis in this section.

We first consider the boundedness of the merit parameters $\{\tilde{\rho}_t\}$. It is worth noting that updating $\tilde{\rho}_t$ requires no further analysis at iterations satisfying $\|\tilde{c}_t\| \leq \epsilon'$. Under this circumstance, indeed, Algorithm 1 sets $\tilde{v}_t = \mathbf{0}$ and $\Delta\tilde{c}_t = 0$. By applying Proposition 1, it holds that $\tilde{\rho}_t^{\text{trial}} = 0$ and $\tilde{\rho}_t = \tilde{\rho}_{t-1}$ according to their definitions (2.7) and (2.8). Hence, it remains to consider the restoration iterations for which $\|\tilde{c}_t\| > \epsilon'$ and $\|\tilde{v}_t^c\| > \epsilon'\beta_{\text{IS}}$. At a restoration iteration where the early termination test is not triggered, Algorithm 1 proceeds to compute a normal step. However, this non-terminating branch does not by itself provide a quantitative guarantee on the amount of constraint restoration. To this end, we require a uniform relation between the first-order stationary measure of the deterministic feasibility problem and the corresponding constraint violation. This motivates the following error-bound condition on the deterministic feasibility problem $\min_{x \in \mathbb{R}_{\geq 0}^n} \frac{1}{2}\|c(x)\|^2$.

Assumption 4 *There exists a constant $\alpha_v > 0$ such that $\|v_t^c\| \geq \alpha_v\|c_t\|$ for all iterations of Algorithm 1, where $v_t^c := \Pi_{\mathbb{R}_{\geq 0}^n}(x_t - \nabla c_t c_t) - x_t$ denotes the true projected gradient residual.*

Note that this assumption pertains solely to the geometry of the deterministic feasibility problem without imposing analogous bounds on the stochastic estimators. Indeed, projection ensures bounded estimates, but finite p th-moment conditions alone do not guarantee sufficiently small estimation errors on every sample path. In the subsequent analysis, this deterministic error bound is combined with explicit estimates of the errors in $\tilde{c}_t$ and $\tilde{\nabla}c_t$ to derive a lower bound for the sampled projected residual, thereby distinguishing degeneracy of the underlying feasibility geometry from perturbations caused by sampling.

Remark 1 *Assumption 4 is motivated by the related condition in [37, AS.10]. Indeed, $v_t^c = \mathbf{0}$ with $c_t \neq \mathbf{0}$ characterizes an exact IS point of the deterministic feasibility problem, whereas Assumption 4 also excludes sequences satisfying $v_t^c \neq \mathbf{0}$ at every finite iteration but $\|v_t^c\|/\|c_t\| \rightarrow 0$. Such sequences are not exactly stationary at any finite iteration, but their constraint-restoration efficiency degenerates asymptotically. Thus, Assumption 4 also rules out this asymptotic degeneracy by imposing a uniform lower bound on the projected feasibility residual relative to the constraint violation. Curtis et al. [11, Lemma 4.22] assume uniform contraction of the linearized constraint residual, whereas we derive sampled normal-step contraction from Assumption 4 under sufficiently accurate constraint estimates.*

To proceed, for notational simplicity we define

$$\begin{aligned}\mathcal{E}_t &:= (\alpha_v + L_c)\|\tilde{c}_t - c_t\| + M_c\|\tilde{\nabla}c_t - \nabla c_t\|_F, \\ \mathcal{E}'_t &:= (\alpha_v + L_c)(\mathbb{E}_t[\|\tilde{c}_t - c_t\|^p])^{1/p} + M_c(\mathbb{E}_t[\|\tilde{\nabla}c_t - \nabla c_t\|_F^p])^{1/p}, \quad t \geq 1.\end{aligned}\tag{3.1}$$

Although $\tilde{v}_t^c$ is introduced primarily for checking a sampled ϵ' -IS point, it also serves as a Cauchy benchmark for the normal-step subproblem (2.1). The following lemma uses Assumption 4 together with the estimation-error bounds to establish a per-iteration contraction property for the sampled linearized constraint residual.

Lemma 1 *Suppose that Assumption 4 and Condition 1 hold. If $\|\tilde{c}_t\| > \epsilon'$ with a given tolerance $\epsilon' > 0$, the normal step $\tilde{v}_t$ obtained from (2.1) satisfies $\|\tilde{c}_t + \tilde{\nabla}c_t^\top \tilde{v}_t\| \leq \vartheta_t\|\tilde{c}_t\|$, with $\vartheta_t = (1 - \frac{[\alpha_v - \mathcal{E}_t/\|\tilde{c}_t\|_+^2]}{1 + L_c^2 + \gamma_{\max}})^{1/2}$.*

Proof. By the variational characterization of the Euclidean projection onto the closed convex set $\mathbb{R}_{\geq 0}^n$, it holds that for any point $y \in \mathbb{R}^n$ and $z \in \mathbb{R}_{\geq 0}^n$, $(y - \Pi_{\mathbb{R}_{\geq 0}^n}(y))^\top(z - \Pi_{\mathbb{R}_{\geq 0}^n}(y)) \leq 0$. Let $y = x_t - \tilde{\nabla}c_t \tilde{c}_t$. By the definition of $\tilde{v}_t^c$ (2.2), it follows that $\Pi_{\mathbb{R}_{\geq 0}^n}(y) = x_t + \tilde{v}_t^c$. Since $x_t \in \mathbb{R}_{\geq 0}^n$, we can set $z = x_t$. Substituting these terms into the variational inequality yields $((x_t - \tilde{\nabla}c_t \tilde{c}_t) - (x_t + \tilde{v}_t^c))^\top(x_t - (x_t + \tilde{v}_t^c)) \leq 0$. It can be further simplified to $(\tilde{\nabla}c_t \tilde{c}_t)^\top \tilde{v}_t^c \leq -\|\tilde{v}_t^c\|^2$. Moreover, the vector $\tilde{v}_t^c/(1 + L_c^2 + \gamma_{\max})$ is also feasible for normal-step

subproblem (2.1) due to $x_t \geq 0$ and $x_t + \tilde{v}_t^c \geq 0$. It then yields

$$\begin{aligned} \frac{1}{2} \left\| \tilde{c}_t + \frac{\tilde{\nabla} c_t^\top \tilde{v}_t^c}{1 + L_c^2 + \gamma_{\max}} \right\|^2 + \frac{\gamma_t}{2} \left\| \frac{\tilde{v}_t^c}{1 + L_c^2 + \gamma_{\max}} \right\|^2 &= \frac{1}{2} \|\tilde{c}_t\|^2 + \frac{(\tilde{\nabla} c_t \tilde{c}_t)^\top \tilde{v}_t^c}{1 + L_c^2 + \gamma_{\max}} + \frac{\|\tilde{\nabla} c_t^\top \tilde{v}_t^c\|^2 + \gamma_t \|\tilde{v}_t^c\|^2}{2(1 + L_c^2 + \gamma_{\max})^2} \\ &\leq \frac{1}{2} \|\tilde{c}_t\|^2 - \frac{\|\tilde{v}_t^c\|^2}{1 + L_c^2 + \gamma_{\max}} + \frac{(L_c^2 + \gamma_{\max}) \|\tilde{v}_t^c\|^2}{2(1 + L_c^2 + \gamma_{\max})^2} \\ &\leq \frac{1}{2} \|\tilde{c}_t\|^2 - \frac{\|\tilde{v}_t^c\|^2}{2(1 + L_c^2 + \gamma_{\max})}. \end{aligned}$$

Since $\tilde{v}_t$ minimizes the normal-step objective, its objective value is no larger than that of the feasible vector $\tilde{v}_t^c/(1 + L_c^2 + \gamma_{\max})$. Dropping the non-negative term $\frac{\gamma_t}{2} \|\tilde{v}_t\|^2$ then gives

$$\|\tilde{c}_t + \tilde{\nabla} c_t^\top \tilde{v}_t\|^2 \leq \|\tilde{c}_t\|^2 - \frac{\|\tilde{v}_t^c\|^2}{1 + L_c^2 + \gamma_{\max}}. \quad (3.2)$$

Recall that $v_t^c = \Pi_{\mathbb{R}_{\geq 0}^n}(x_t - \nabla c_t c_t) - x_t$ is the true counterpart of $\tilde{v}_t^c$. By the non-expansiveness of the Euclidean projection, we obtain $\|\tilde{v}_t^c - v_t^c\| \leq \|\tilde{\nabla} c_t \tilde{c}_t - \nabla c_t c_t\| \leq L_c \|\tilde{c}_t - c_t\| + M_c \|\tilde{\nabla} c_t - \nabla c_t\|_F$. Therefore, using Assumption 4 and the inequality $\|c_t\| \geq \|\tilde{c}_t\| - \|\tilde{c}_t - c_t\|$, it follows that

$$\begin{aligned} \|\tilde{v}_t^c\| &\geq \|v_t^c\| - \|\tilde{v}_t^c - v_t^c\| \geq \alpha_v \|c_t\| - L_c \|\tilde{c}_t - c_t\| - M_c \|\tilde{\nabla} c_t - \nabla c_t\|_F \\ &\geq \alpha_v \|\tilde{c}_t\| - (\alpha_v + L_c) \|\tilde{c}_t - c_t\| - M_c \|\tilde{\nabla} c_t - \nabla c_t\|_F \\ &= \alpha_v \|\tilde{c}_t\| - \mathcal{E}_t = \|\tilde{c}_t\|(\alpha_v - \mathcal{E}_t/\|\tilde{c}_t\|). \end{aligned} \quad (3.3)$$

Since $\|\tilde{v}_t^c\| \geq 0$, (3.3) implies $\|\tilde{v}_t^c\| \geq \|\tilde{c}_t\|[\alpha_v - \mathcal{E}_t/\|\tilde{c}_t\|]_+$. Combining this with the bound in (3.2) gives

$$\|\tilde{c}_t + \tilde{\nabla} c_t^\top \tilde{v}_t\|^2 \leq \|\tilde{c}_t\|^2 - \frac{\|\tilde{v}_t^c\|^2}{1 + L_c^2 + \gamma_{\max}} \leq \|\tilde{c}_t\|^2 \left(1 - \frac{[\alpha_v - \mathcal{E}_t/\|\tilde{c}_t\|]_+^2}{1 + L_c^2 + \gamma_{\max}} \right). \quad (3.4)$$

Thus ϑ_t is well defined and conclusion of the lemma is derived. $\square$

For each generated iteration t of Algorithm 1 and a constant $\eta_c \in (0, 1)$, we define event

$$\mathcal{B}_t := \{\|\tilde{c}_t\| \leq \epsilon'\} \cup \{\mathcal{E}_t \leq \eta_c \alpha_v \epsilon'\}, \quad t = 1, \dots, T, \quad \text{and} \quad \bar{\mathcal{B}} := \bigcap_{t=1}^T \mathcal{B}_t. \quad (3.5)$$

The first event in $\mathcal{B}_t$ corresponds to the branch that no normal-step contraction is required, whereas the second one guarantees sufficient sampling accuracy when a normal step is computed. The complementary event $\bar{\mathcal{B}}^c$ consists of sample paths on which the prescribed sampling-accuracy condition fails at some generated iteration with $\|\tilde{c}_t\| > \epsilon'$. We regard the occurrence of $\mathcal{B}_t^c$ and $\bar{\mathcal{B}}^c$ as non-progressive, on which the desired normal-step contraction is not guaranteed. The following lemma gives a sufficient conditional moment condition ensuring the prescribed non-progressive probability bounds for $\mathcal{B}_t^c$ and $\bar{\mathcal{B}}^c$.

Lemma 2 *For each $t \geq 1$, let $\delta_t \in (0, 1)$ be a prescribed probability budget and $\sum_{t=1}^T \delta_t \in (0, 1)$. If*

$$\mathcal{E}_t' \leq \eta_c \alpha_v \epsilon' \delta_t^{1/p} \quad \text{almost surely}, \quad (3.6)$$

then $\mathbb{P}_t(\mathcal{B}_t^c) \leq \delta_t$ and hence $\mathbb{P}(\bar{\mathcal{B}}^c) \leq \sum_{t=1}^T \delta_t$.

Proof. By the definition of $\mathcal{E}_t$ in (3.1) and the conditional Minkowski's inequality, we have $\mathbb{E}_t[\mathcal{E}_t^p]^{1/p} = \mathbb{E}_t[\left((\alpha_v + L_c)\|\tilde{c}_t - c_t\| + M_c\|\tilde{\nabla} c_t - \nabla c_t\|_F\right)^p]^{1/p} \leq \mathcal{E}_t'$. It follows from (3.6) that $\mathbb{E}_t[\mathcal{E}_t^p] \leq (\eta_c \alpha_v \epsilon')^p \delta_t$. Moreover, taking the complement of (3.5) gives $\mathcal{B}_t^c = \{\|\tilde{c}_t\| > \epsilon'\} \cap \{\mathcal{E}_t > \eta_c \alpha_v \epsilon'\}$. Therefore, we obtain from Markov's inequality that $\mathbb{P}_t(\mathcal{B}_t^c) \leq \mathbb{P}_t(\mathcal{E}_t > \eta_c \alpha_v \epsilon') \leq \frac{\mathbb{E}_t[\mathcal{E}_t^p]}{(\eta_c \alpha_v \epsilon')^p} \leq \delta_t$. Finally, applying the union bound and the tower property yields $\mathbb{P}(\bar{\mathcal{B}}^c) = \mathbb{P}(\bigcup_{t=1}^T \mathcal{B}_t^c) \leq \sum_{t=1}^T \mathbb{P}(\mathcal{B}_t^c) = \sum_{t=1}^T \mathbb{E}[\mathbb{P}_t(\mathcal{B}_t^c)] \leq \sum_{t=1}^T \delta_t$. $\square$

Note that condition (3.6) is a sampling-accuracy requirement rather than an additional structural assumption. By employing sufficiently large batches to estimate c_t and ∇c_t , the conditional failure probability

$\mathbb{P}_t(\mathcal{B}_t^c)$ can be bounded by any prescribed tolerance $\delta_t \in (0, 1)$ that is arbitrarily small. Section 4 specifies the corresponding sample-size requirements. For the subsequent analysis of Algorithm 1, we further assume

$$\beta_{\text{IS}} \leq \min \left\{ \frac{1}{\max\{1, L_c M_c\}}, \frac{(1 - \eta_c)\alpha_v}{2} \right\}. \quad (3.7)$$

The following lemma provides a probability guarantee for Algorithm 1 completing all prescribed iterations. The key is to prove $\bar{\mathcal{B}} \subseteq \bar{\mathcal{S}}$, thus $\mathbb{P}(\bar{\mathcal{S}}) \geq 1 - \mathbb{P}(\bar{\mathcal{B}}^c)$.

Lemma 3 *Suppose Assumption 4, Condition 1 and (3.6) holds. Then with probability at least $1 - \sum_{t=1}^T \delta_t$, Algorithm 1 completes all T prescribed iterations, i.e. $\mathbb{P}(\bar{\mathcal{S}}) \geq 1 - \sum_{t=1}^T \delta_t$.*

Proof. We first show that early termination cannot occur on $\bar{\mathcal{B}}$. At a generated iteration t , if $\|\tilde{c}_t\| \leq \epsilon'$, the early termination test is not performed. Otherwise, $\mathcal{B}_t$ implies $\mathcal{E}_t \leq \eta_c \alpha_v \epsilon'$. The projected-residual bound established in (3.3) and the choice (3.7) give

$$\|\tilde{v}_t^c\| \geq \alpha_v \|\tilde{c}_t\| - \mathcal{E}_t \geq (1 - \eta_c) \alpha_v \epsilon' \geq 2\epsilon' \beta_{\text{IS}} > \epsilon' \beta_{\text{IS}} \quad \text{on } \mathcal{B}_t, \quad \text{if } \|\tilde{c}_t\| > \epsilon'.$$

Consequently, Algorithm 1 completes all prescribed T iterations on $\bar{\mathcal{B}}$, which implies $\bar{\mathcal{B}} \subseteq \bar{\mathcal{S}}$. That is to say, $\bar{\mathcal{S}}^c \subseteq \bar{\mathcal{B}}^c$. Then, Lemma 2 yields $\mathbb{P}(\bar{\mathcal{S}}^c) \leq \mathbb{P}(\bar{\mathcal{B}}^c) \leq \sum_{t=1}^T \delta_t$. Therefore, the probability of completing all prescribed iterations satisfies $\mathbb{P}(\bar{\mathcal{S}}) = 1 - \mathbb{P}(\bar{\mathcal{S}}^c) \geq 1 - \mathbb{P}(\bar{\mathcal{B}}^c) \geq 1 - \sum_{t=1}^T \delta_t$, which completes the proof. $\square$

To establish expected KKT residual bound conditioned on $\bar{\mathcal{S}}$, we first derive the following uniform normal-step contraction bound on $\bar{\mathcal{B}} \subseteq \bar{\mathcal{S}}$.

Lemma 4 *Suppose the conditions of Lemma 1 hold. On the event $\bar{\mathcal{B}}$, for any generated iteration t satisfying $\|\tilde{c}_t\| > \epsilon'$, it holds that $\|\tilde{c}_t + \tilde{\nabla} c_t^\top \tilde{v}_t\| \leq \vartheta \|\tilde{c}_t\|$ with $\vartheta := (1 - \frac{(1 - \eta_c)^2 \alpha_v^2}{1 + L_c^2 + \gamma_{\max}})^{1/2} \in (0, 1)$.*

Proof. Fix a generated iteration t satisfying $\|\tilde{c}_t\| > \epsilon'$ and consider the event $\bar{\mathcal{B}}$. Since $\bar{\mathcal{B}} \subseteq \mathcal{B}_t$, the definition of $\mathcal{B}_t$ gives $\mathcal{E}_t \leq \eta_c \alpha_v \epsilon' < \eta_c \alpha_v \|\tilde{c}_t\|$. By the inequality (3.3), we further obtain $(1 - \eta_c) \alpha_v \|\tilde{c}_t\| \leq \|\tilde{v}_t^c\|$. Note that $x_t \in \mathbb{R}_{\geq 0}^n$ implies $\Pi_{\mathbb{R}_{\geq 0}^n}(x_t) = x_t$. By Assumption 4 and the non-expansiveness of the projection, it holds that $\|\tilde{v}_t^c\| = \|\Pi_{\mathbb{R}_{\geq 0}^n}(x_t - \tilde{\nabla} c_t \tilde{c}_t) - \Pi_{\mathbb{R}_{\geq 0}^n}(x_t)\| \leq \|\tilde{\nabla} c_t \tilde{c}_t\| \leq L_c \|\tilde{c}_t\|$, and hence $(1 - \eta_c) \alpha_v \leq L_c$ due to $\|\tilde{c}_t\| > 0$ and $0 < \eta_c < 1$. Then, it follows that $0 < \frac{(1 - \eta_c)^2 \alpha_v^2}{1 + L_c^2 + \gamma_{\max}} \leq \frac{L_c^2}{1 + L_c^2 + \gamma_{\max}} < 1$. Thus, $\vartheta \in (0, 1)$ is well defined. Moreover, the preceding bound on $\mathcal{E}_t$ implies $[\alpha_v - \mathcal{E}_t / \|\tilde{c}_t\|]_+ \geq (1 - \eta_c) \alpha_v$. It follows from Lemma 1 that $\|\tilde{c}_t + \tilde{\nabla} c_t^\top \tilde{v}_t\| \leq \vartheta \|\tilde{c}_t\|$ holds on $\bar{\mathcal{B}}$. $\square$

Using Lemma 4, the following lemma establishes uniform upper bounds on both the trial and actual merit parameters on $\bar{\mathcal{B}}$.

Lemma 5 *Suppose that Assumption 4 and Condition 1 hold. On the event $\bar{\mathcal{B}}$, for any generated iteration $t \in [T]$, it then holds that $0 \leq \tilde{\rho}_t^{\text{trial}} \leq \rho_{\max}^{\text{trial}}$ and $0 < \tilde{\rho}_0 \leq \tilde{\rho}_t \leq \rho_{\max}$ with $\rho_{\max}^{\text{trial}} := \frac{M_f / \sqrt{\gamma_{\min}} + \sigma_u M_c / (2\gamma_{\min})}{(1 - \sigma_\rho)(1 - \vartheta)}$ and $\rho_{\max} := \max\{\tilde{\rho}_0, \frac{\rho_{\max}^{\text{trial}}}{1 - \epsilon_\rho}\}$, where $\vartheta \in (0, 1)$ is defined in Lemma 4.*

Proof. We first establish the upper bound on $\tilde{\rho}_t^{\text{trial}}$ by considering the two branches in (2.7). If $\tilde{\nabla} f_t^\top \tilde{d}_t + \frac{1}{2} \tilde{d}_t^\top H_t \tilde{d}_t \leq 0$, the second branch of (2.7) yields $\tilde{\rho}_t^{\text{trial}} = 0$, and hence $0 \leq \tilde{\rho}_t^{\text{trial}} \leq \rho_{\max}^{\text{trial}}$, $t \in [T]$ holds trivially. Consider now the case $\tilde{\nabla} f_t^\top \tilde{d}_t + \frac{1}{2} \tilde{d}_t^\top H_t \tilde{d}_t > 0$, which only occurs when $\|\tilde{c}_t\| > \epsilon'$. We next bound the numerator in the first branch of (2.7) (i.e., $\tilde{\nabla} f_t^\top \tilde{d}_t + \frac{1}{2} \tilde{d}_t^\top H_t \tilde{d}_t$). Since $v = \mathbf{0}$ is also feasible for the normal-step subproblem (2.1), the optimality of $\tilde{v}_t$ yields $\|\tilde{c}_t + \tilde{\nabla} c_t^\top \tilde{v}_t\|^2 + \gamma_t \|\tilde{v}_t\|^2 \leq \|\tilde{c}_t\|^2$. Thus, we obtain $\|\tilde{v}_t\| \leq \|\tilde{c}_t\| / \sqrt{\gamma_t} \leq \|\tilde{c}_t\| / \sqrt{\gamma_{\min}}$. Note that $\tilde{v}_t$ satisfies both constraints of the total-step subproblem (2.3), then $\tilde{\nabla} f_t^\top \tilde{d}_t + \frac{1}{2} \tilde{d}_t^\top H_t \tilde{d}_t \leq \tilde{\nabla} f_t^\top \tilde{v}_t + \frac{1}{2} \tilde{v}_t^\top H_t \tilde{v}_t$. By the positive definiteness of H_t and $\|\tilde{c}_t\| \leq M_c$, we further obtain

$$\tilde{\nabla} f_t^\top \tilde{d}_t + \frac{1}{2} \tilde{d}_t^\top H_t \tilde{d}_t \leq M_f \|\tilde{v}_t\| + \frac{\sigma_u}{2} \|\tilde{v}_t\|^2 \leq \frac{M_f}{\sqrt{\gamma_{\min}}} \|\tilde{c}_t\| + \frac{\sigma_u}{2\gamma_{\min}} \|\tilde{c}_t\|^2 \leq \left(\frac{M_f}{\sqrt{\gamma_{\min}}} + \frac{\sigma_u M_c}{2\gamma_{\min}} \right) \|\tilde{c}_t\|.$$

By Lemma 4 with $\|\tilde{c}_t\| > \epsilon'$, we have $\Delta \tilde{c}_t \geq (1 - \vartheta) \|\tilde{c}_t\|$. The first branch of (2.7) gives $\tilde{\rho}_t^{\text{trial}} \leq \rho_{\max}^{\text{trial}}$.

We next prove the bound on the merit parameter $\tilde{\rho}_t$. Suppose first that $\tilde{\rho}_{t-1} \geq \tilde{\rho}_t^{\text{trial}}$. By the first branch of (2.8), we then obtain $\tilde{\rho}_t = \tilde{\rho}_{t-1}$. Otherwise, it follows that $\tilde{\rho}_{t-1} < \tilde{\rho}_t^{\text{trial}} \leq \rho_{\max}^{\text{trial}}$, and hence the second branch of (2.8) gives $\tilde{\rho}_t = \max\{\frac{\tilde{\rho}_{t-1}}{1-\epsilon_\rho}, \tilde{\rho}_t^{\text{trial}}\} \leq \frac{\rho_{\max}^{\text{trial}}}{1-\epsilon_\rho}$. Starting from $\tilde{\rho}_0 > 0$, we can obtain $0 < \tilde{\rho}_0 \leq \tilde{\rho}_t \leq \rho_{\max}$ by induction and $\epsilon_\rho \in (0, 1)$. $\square$

Upon the boundedness of the merit parameter established in Lemma 5, the following lemma bounds the contribution of $\tilde{\mathcal{B}}$ to the expected norm of the sampled total step $\tilde{d}_R$.

Lemma 6 *Under Assumption 4, Conditions 1 and 2, it holds that for any $t \in [T]$,*

$$\begin{aligned} \mathbb{E}[\|\tilde{d}_R\| \cdot \mathbb{I}_{\tilde{\mathcal{B}}}] &\leq \frac{2\ell_{\max}(f_1 - f^* + \rho_{\max}M_c)}{\sigma_l K_T} + \frac{(4\ell_{\max} + \sigma_l) \sum_{t=1}^T \kappa_t^2}{4\sigma_l \ell_{\min} K_T} \\ &\quad + \frac{1}{\sigma_l K_T} \sum_{t=1}^T \left(\frac{2\ell_{\max} \kappa_t}{\ell_{\min}} E_{t,f} + 4\ell_{\max} \rho_{\max} E_{t,c} + \frac{2\ell_{\max} \rho_{\max} \kappa_t}{\ell_{\min}} E_{t,J} \right) =: \mathcal{L}_T, \end{aligned} \quad (3.8)$$

with notations $\ell_t := L_f + \tilde{\rho}_t L_J$, $\ell_{\min} := L_f + \tilde{\rho}_0 L_J$ and $\ell_{\max} := L_f + \rho_{\max} L_J$.

Proof. By the L_f -smoothness of f and the L_J -smoothness of c , we have

$$\begin{aligned} f_{t+1} + \tilde{\rho}_t \|c_{t+1}\| &\leq f_t + \eta_t \nabla f_t^\top \tilde{d}_t + \tilde{\rho}_t \|c_t\| + \eta_t \nabla c_t^\top \tilde{d}_t + \frac{\eta_t^2 L_f + \eta_t^2 \tilde{\rho}_t L_J}{2} \|\tilde{d}_t\|^2 \\ &\leq f_t + \eta_t \tilde{\nabla} f_t^\top \tilde{d}_t + \eta_t \|\nabla f_t - \tilde{\nabla} f_t\| \|\tilde{d}_t\| + \frac{\ell_t}{2} \eta_t^2 \|\tilde{d}_t\|^2 \\ &\quad + \tilde{\rho}_t \|\tilde{c}_t + \eta_t \tilde{\nabla} c_t^\top \tilde{d}_t\| + \tilde{\rho}_t \|\tilde{c}_t - c_t\| + \tilde{\rho}_t \eta_t \|\tilde{\nabla} c_t - \nabla c_t\| \|\tilde{d}_t\|. \end{aligned} \quad (3.9)$$

Crucially, the total step satisfies $\tilde{\nabla} c_t^\top \tilde{d}_t = \tilde{\nabla} c_t^\top \tilde{v}_t$. Using the convexity of the norm and $\eta_t \in (0, 1]$, we bound the linearized model as

$$\|\tilde{c}_t + \eta_t \tilde{\nabla} c_t^\top \tilde{d}_t\| = \|\tilde{c}_t + \eta_t \tilde{\nabla} c_t^\top \tilde{v}_t\| \leq (1 - \eta_t) \|\tilde{c}_t\| + \eta_t \|\tilde{c}_t\| + \tilde{\nabla} c_t^\top \tilde{v}_t = \|\tilde{c}_t\| - \eta_t \Delta \tilde{c}_t.$$

By the selection of $\tilde{\rho}_t$ in (2.8), we further obtain

$$\tilde{\nabla} f_t^\top \tilde{d}_t - \tilde{\rho}_t \Delta \tilde{c}_t \leq -\frac{1}{2} \tilde{d}_t^\top H_t \tilde{d}_t - \sigma_\rho \tilde{\rho}_t \Delta \tilde{c}_t \leq -\frac{\sigma_l}{2} \|\tilde{d}_t\|^2,$$

where the second inequality follows from the positive definiteness of H_t and $\sigma_\rho \tilde{\rho}_t \Delta \tilde{c}_t \geq 0$. With the setting of $\eta_t = \min\{\kappa_t/(\ell_t \|\tilde{d}_t\|), 1\}$ from (2.9), we apply $0 < \ell_{\min} \leq \ell_t \leq \ell_{\max}$ on $\tilde{\mathcal{B}}$ to obtain

$$\eta_t \|\tilde{d}_t\|^2 \geq \min \left\{ \frac{\kappa_t \|\tilde{d}_t\|}{\ell_{\max}}, \|\tilde{d}_t\|^2 \right\} = \frac{\kappa_t^2}{\ell_{\max}^2} \min \left\{ \frac{\ell_{\max} \|\tilde{d}_t\|}{\kappa_t}, \frac{(\ell_{\max} \|\tilde{d}_t\|)^2}{\kappa_t^2} \right\} \geq \frac{\kappa_t}{\ell_{\max}} \|\tilde{d}_t\| - \frac{\kappa_t^2}{4\ell_{\max}^2},$$

where we invoke Lemma 12 in the second inequality, and $\eta_t \|\tilde{d}_t\| \leq \kappa_t/\ell_t \leq \kappa_t/\ell_{\min}$. Assembling the above bounds and substituting them into (3.9) gives

$$\begin{aligned} f_{t+1} + \tilde{\rho}_t \|c_{t+1}\| &\leq f_t + \tilde{\rho}_t \|c_t\| + \left(\frac{1}{2\ell_{\min}} + \frac{\sigma_l}{8\ell_{\max}\ell_{\min}} \right) \kappa_t^2 - \frac{\sigma_l \kappa_t \|\tilde{d}_t\|}{2\ell_{\max}} \\ &\quad + \frac{\kappa_t}{\ell_{\min}} \|\tilde{\nabla} f_t - \nabla f_t\| + 2\tilde{\rho}_t \|\tilde{c}_t - c_t\| + \frac{\tilde{\rho}_t \kappa_t}{\ell_{\min}} \|\tilde{\nabla} c_t - \nabla c_t\|, \end{aligned}$$

where we use $\sigma_l/\ell_{\max}^2 \leq \sigma_l/(\ell_{\max}\ell_{\min})$ in this inequality. Lemma 5 also gives $0 < \tilde{\rho}_0 \leq \tilde{\rho}_t \leq \rho_{\max}$ on $\tilde{\mathcal{B}}$. Together with the monotonicity of $\tilde{\rho}_t$, this yields

$$\begin{aligned} \sum_{t=1}^T [f_t - f_{t+1} + \tilde{\rho}_t (\|c_t\| - \|c_{t+1}\|)] &= f_1 - f_{T+1} + \tilde{\rho}_1 \|c_1\| + \sum_{t=2}^T (\tilde{\rho}_t - \tilde{\rho}_{t-1}) \|c_t\| - \tilde{\rho}_T \|c_{T+1}\| \\ &\leq f_1 - f^* + \rho_{\max} M_c, \end{aligned}$$

where we used $f_{T+1} \geq f^*$ and $\|c_t\| \leq M_c$. We now multiply the preceding one-step inequality by $\mathbb{I}_{\tilde{\mathcal{B}}}$, sum over $t \in [T]$, and take expectations. Using the bound of $\tilde{\rho}_t$ based on Lemma 5 and Condition 2, with $0 \leq \mathbb{I}_{\tilde{\mathcal{B}}} \leq 1$ for the nonnegative error terms, we obtain

$$\begin{aligned} \sum_{t=1}^T \kappa_t \mathbb{E}[\|\tilde{d}_t\| \cdot \mathbb{I}_{\tilde{\mathcal{B}}}] &\leq \frac{2\ell_{\max}}{\sigma_l} (f_1 - f^* + \rho_{\max} M_c) + \frac{(4\ell_{\max} + \sigma_l) \sum_{t=1}^T \kappa_t^2}{4\sigma_l \ell_{\min}} \\ &\quad + \frac{1}{\sigma_l} \sum_{t=1}^T \left(\frac{2\ell_{\max} \kappa_t}{\ell_{\min}} E_{t,f} + 4\ell_{\max} \rho_{\max} E_{t,c} + \frac{2\ell_{\max} \rho_{\max} \kappa_t}{\ell_{\min}} E_{t,J} \right). \end{aligned}$$

Applying the output rule yields the desired result, completing this proof. $\square$

Lemma 6 controls the total step $\tilde{d}_R$ computed from stochastic information, whereas the KKT analysis requires to address its true counterpart d_R . We thus let d_t denote the true total step, defined by the deterministic reference problem corresponding to (2.3), that is,

$$\min_{d \in \mathbb{R}^n} \nabla f_t^\top d + \frac{1}{2} d^\top H_t d \quad \text{s.t.} \quad \nabla c_t^\top d = \nabla c_t^\top v_t, \quad x_t + d \geq \mathbf{0}. \quad (3.10)$$

where the true normal step v_t is defined as the solution of

$$\min_{v \in \mathbb{R}^n} \frac{1}{2} \|c_t + \nabla c_t^\top v\|^2 + \frac{\gamma_t}{2} \|v\|^2 \quad \text{s.t.} \quad x_t + v \geq \mathbf{0}. \quad (3.11)$$

The following lemma characterizes the uniform bounds on v_t and $\tilde{v}_t$ as well as $\|\tilde{v}_t - v_t\|$. Note that in Algorithm 1, when a total step is computed, $\tilde{v}_t$ might be set as zero (see Step 4) or computed through (2.1) (see Step 10). The case $v_t = 0$ automatically satisfies the bound.

Lemma 7 *Under Condition 1, the true normal step v_t and sampled normal step $\tilde{v}_t$ are uniformly bounded by $M_v := M_c/\sqrt{\gamma_{\min}}$. Moreover, for every iteration at which a total step is computed, it holds that*

$$\|\tilde{v}_t - v_t\| \leq \frac{1}{\gamma_{\min}} \left((2M_v L_c + M_c) \|\tilde{\nabla} c_t - \nabla c_t\| + L_c \|\tilde{c}_t - c_t\| + \sqrt{\gamma_{\min}} \epsilon' \right). \quad (3.12)$$

Proof. We will establish the uniform boundedness of the true normal step v_t , and then the upper bound for $\tilde{v}_t$ is obtained via a similar argument. Since $x_t \geq \mathbf{0}$, the trivial step $v = \mathbf{0}$ is always feasible. By the optimality of v_t for problem (3.11), we have $\frac{1}{2} \|c_t + \nabla c_t^\top v_t\|^2 + \frac{\gamma_t}{2} \|v_t\|^2 \leq \frac{1}{2} \|c_t\|^2$. Dropping the first non-negative term on the left-hand side and using the lower bound on γ_t yields $\|v_t\| \leq \|c_t\|/\sqrt{\gamma_{\min}}$. Under the boundedness of $\|c_t\|$ (i.e., $\|c_t\| \leq M_c$) by Assumption 1, the normal step $\|v_t\|$ is uniformly bounded by $M_c/\sqrt{\gamma_{\min}} =: M_v$. The same bound holds for the normal step $\tilde{v}_t$ executed by Algorithm 1. Specifically, if $\|\tilde{c}_t\| \leq \epsilon'$, then $\tilde{v}_t = \mathbf{0}$. Otherwise, $\tilde{v}_t$ solves the sampled normal-step subproblem (2.1), and an identical comparison with the zero step holds. Then $\|\tilde{v}_t\| \leq M_v$ can be derived by Condition 1.

We next bound the deviation $\|\tilde{v}_t - v_t\|$. Recall that in Algorithm 1, $\tilde{v}_t$ is either $\mathbf{0}$ or determined by (2.1) depending on the value of $\|\tilde{c}_t\|$. For simplicity of analysis, we introduce the auxiliary sampled normal step

$$\hat{v}_t := \arg \min_{v \geq -x_t} \left\{ \frac{1}{2} \|\tilde{c}_t + \tilde{\nabla} c_t^\top v\|^2 + \frac{\gamma_t}{2} \|v\|^2 \right\}.$$

By the same argument as above, it holds that $\|\hat{v}_t\| \leq \|\tilde{c}_t\|/\sqrt{\gamma_{\min}} \leq M_v$. The first-order optimality condition for (3.11) is characterized by the variational inequality $(v - v_t)^\top (\nabla c_t (c_t + \nabla c_t^\top v_t) + \gamma_t v_t) \geq 0$ for any $v \geq -x_t$. Similarly, the stochastic step $\hat{v}_t$ satisfies $(v - \hat{v}_t)^\top (\tilde{\nabla} c_t (\tilde{c}_t + \tilde{\nabla} c_t^\top \hat{v}_t) + \gamma_t \hat{v}_t) \geq 0$. Substituting $v = \hat{v}_t$ into the first variational inequality and $v = v_t$ into the second, and summing them yields

$$(\hat{v}_t - v_t)^\top \left[(\nabla c_t \nabla c_t^\top + \gamma_t I) v_t - (\tilde{\nabla} c_t \tilde{\nabla} c_t^\top + \gamma_t I) \hat{v}_t + \nabla c_t c_t - \tilde{\nabla} c_t \tilde{c}_t \right] \geq 0.$$

Rearranging terms and using the γ_t -strong monotonicity (i.e., $(\hat{v}_t - v_t)^\top \gamma_t (\hat{v}_t - v_t) = \gamma_t \|\hat{v}_t - v_t\|^2$) yields

$$\gamma_t \|\hat{v}_t - v_t\|^2 \leq (\hat{v}_t - v_t)^\top \left[(\nabla c_t \nabla c_t^\top - \tilde{\nabla} c_t \tilde{\nabla} c_t^\top) v_t + \nabla c_t c_t - \tilde{\nabla} c_t \tilde{c}_t \right].$$

With the Cauchy-Schwarz inequality applied to the right-hand side, the bound (3.12) trivially holds if $\hat{v}_t = v_t$. For $\|\hat{v}_t - v_t\| > 0$, dividing both sides by $\|\hat{v}_t - v_t\|$ and applying the triangle inequality yields

$$\gamma_t \|\hat{v}_t - v_t\| \leq \|(\nabla c_t \nabla c_t^\top - \tilde{\nabla} c_t \tilde{\nabla} c_t^\top) v_t\| + \|\nabla c_t c_t - \tilde{\nabla} c_t \tilde{c}_t\|. \quad (3.13)$$

Based on the uniform boundedness of $v_t, \nabla c_t, \tilde{\nabla} c_t, \tilde{c}_t$, we have

$$\begin{aligned} \|(\nabla c_t \nabla c_t^\top - \tilde{\nabla} c_t \tilde{\nabla} c_t^\top) v_t\| &\leq M_v (\|\nabla c_t\| + \|\tilde{\nabla} c_t\|) \|\tilde{\nabla} c_t - \nabla c_t\| \leq 2M_v L_c \|\tilde{\nabla} c_t - \nabla c_t\|, \\ \|\nabla c_t c_t - \tilde{\nabla} c_t \tilde{c}_t\| &\leq \|\nabla c_t\| \|\tilde{c}_t - c_t\| + \|\tilde{c}_t\| \|\tilde{\nabla} c_t - \nabla c_t\| \leq L_c \|\tilde{c}_t - c_t\| + M_c \|\tilde{\nabla} c_t - \nabla c_t\|. \end{aligned}$$

Substituting these decomposed bounds back into (3.13) gives

$$\|\hat{v}_t - v_t\| \leq \frac{1}{\gamma_{\min}} ((2M_v L_c + M_c) \|\tilde{\nabla} c_t - \nabla c_t\| + L_c \|\tilde{c}_t - c_t\|). \quad (3.14)$$

It remains to relate $\hat{v}_t$ to the normal step $\tilde{v}_t$ actually used by Algorithm 1. If $\|\tilde{c}_t\| > \epsilon'$ and Algorithm 1 proceeds to the total-step computation, then $\tilde{v}_t = \hat{v}_t$. Otherwise, when $\|\tilde{c}_t\| \leq \epsilon'$, Algorithm 1 sets $\tilde{v}_t = 0$, whereas $\|\hat{v}_t\| \leq \|\tilde{c}_t\|/\sqrt{\gamma_{\min}} \leq \epsilon'/\sqrt{\gamma_{\min}}$. Hence, in both cases, it holds that $\|\tilde{v}_t - \hat{v}_t\| \leq \epsilon'/\sqrt{\gamma_{\min}}$. Combining this inequality with (3.14) and the triangle inequality proves the desired result. $\square$

To control the sensitivity of (3.10), we impose the following constraint qualification. It requires uniform linear independence of the constraint gradients and the coordinate vectors associated with near-active bounds.

Assumption 5 *Let d_t be the unique solution of the subproblem (3.10). There exist constants $\varrho, \nu > 0$ such that for all $t \geq 0$,*

$$\|[\nabla c_t \quad I_{\mathcal{I}_t^a}] z\| \geq \nu \|z\|, \quad \forall z \in \mathbb{R}^{m+|\mathcal{I}_t^a|}, \quad (3.15)$$

where $\mathcal{I}_t^a := \{i \in [n] : [x_t + d_t]_i \leq \varrho\}$, and $I_{\mathcal{I}_t^a}$ is the matrix whose columns are standard coordinate vectors $e_i, i \in \mathcal{I}_t^a$. By convention, if $\mathcal{I}_t^a$ is empty, (3.15) reduces to $\|\nabla c_t z\| \geq \nu \|z\|$ for any $z \in \mathbb{R}^m$.

Remark 2 *Assumption 5 imposes uniform linear independence of the equality-constraint Jacobians and the coordinate vectors associated with near-active bounds. Na et al. [8] impose LICQ at a KKT point and require quantitative active-set regularity for their global analysis. Here, the fixed threshold ϱ includes bounds that may become active under small perturbations, while ν provides a uniform lower bound for the corresponding reference matrices. The stochastic framework of [15] requires a uniformly nondegenerate sampled equality Jacobian and a unit direction orthogonal to the sampled equality gradients, with uniformly positive components at active bounds. Our condition is instead imposed on the exact total-step subproblem, and the effects of inaccurate sampled constraints are controlled through perturbation and probability bounds (see [15, Assumption 6]). Moreover, Shi et al. [9] used an extended MFCQ condition requiring uniformly bounded directions that reduce violated constraints by a fixed amount, supporting multiplier bounds and feasibility analysis. Zuo et al. [10] used local LICQ on the stationary set of the constraint-violation function to derive a feasibility error bound and exclude infeasible stationary points. In our analysis, feasibility restoration is controlled separately by Assumption 4, whereas Assumption 5 imposes uniform joint independence of the equality gradients and near-active bound gradients for the total-step perturbation analysis.*

The following lemma establishes a total-step perturbation bound in expectation, relating the sampled total step to its true counterpart.

Lemma 8 *Under Assumption 5 and Conditions 1, 2, the true total step d_t and the sampled total step $\tilde{d}_t$ are uniformly bounded by $M_d = (M_f + \sqrt{M_f^2 + 2\sigma_l M_f M_v + \sigma_l \sigma_u M_v^2})/\sigma_l$. Moreover, there exist constants $\delta_R > 0$ and $L_d > 0$ such that, for every $t \in [T]$,*

$$\mathbb{E}[\|\tilde{d}_t - d_t\|] \leq \left(L_d + \frac{2M_d}{\delta_R} \right) (E_{t,f} + E_{t,J} + \mathbb{E}[\|\tilde{v}_t - v_t\|]) =: \mathcal{R}_t. \quad (3.16)$$

Proof. We first prove the uniform boundedness of the true total step d_t and the sampled total step $\tilde{d}_t$. Substituting $d = v_t$ trivially satisfies both the equality and inequality constraints of (3.10). Then by the optimality conditions for (3.10), we obtain from Lemma 7 that

$$\nabla f_t^\top d_t + \frac{1}{2} d_t^\top H_t d_t \leq \nabla f_t^\top v_t + \frac{1}{2} v_t^\top H_t v_t \leq M_f M_v + \frac{\sigma_u}{2} M_v^2 := Q_{\max}. \quad (3.17)$$

On the other hand, it follows from $\sigma_l I \leq H_t$ that

$$\nabla f_t^\top d_t + \frac{1}{2} d_t^\top H_t d_t \geq -\|\nabla f_t\| \|d_t\| + \frac{\sigma_l}{2} \|d_t\|^2 \geq -M_f \|d_t\| + \frac{\sigma_l}{2} \|d_t\|^2. \quad (3.18)$$

Combining the lower bound (3.18) with (3.17) yields $\frac{\sigma_l}{2} \|d_t\|^2 - M_f \|d_t\| - Q_{\max} \leq 0$. Solving this inequality directly yields the true absolute bound, i.e.,

$$\|d_t\| \leq \frac{M_f + \sqrt{M_f^2 + 2\sigma_l Q_{\max}}}{\sigma_l} = \frac{M_f + \sqrt{M_f^2 + 2\sigma_l M_f M_v + \sigma_l \sigma_u M_v^2}}{\sigma_l} = M_d. \quad (3.19)$$

By a strictly analogous geometric argument, the sampled normal step $\tilde{v}_t$ serves as a feasible point for the perturbed stochastic set $\{d \in \mathbb{R}^n \mid \tilde{\nabla} c_t^\top d = \tilde{\nabla} c_t^\top \tilde{v}_t, d \geq -x_t\}$. Evaluating the objective value of (2.3) at $\tilde{v}_t$ and leveraging the positive definiteness of H_t yields an identical argument. Consequently, the sampled total step is also uniformly bounded by M_d .

To establish the perturbation bound between $\tilde{d}_t$ and d_t , we first derive a uniform bound on the KKT multipliers of the deterministic total-step subproblem (3.10). Let (λ_t, μ_t) denote these multipliers, with the following sign convention. Then the KKT system of (3.10) is

$$H_t d_t + \nabla f_t - \nabla c_t \lambda_t - \mu_t = \mathbf{0}, \quad \nabla c_t^\top (d_t - v_t) = \mathbf{0}, \quad 0 \leq x_t + d_t \perp \mu_t \geq \mathbf{0}.$$

By complementarity, μ_t is supported on the exact active set, which is contained in $\mathcal{I}_t^a$. Hence, it holds that

$$\begin{bmatrix} \nabla c_t & I_{\mathcal{I}_t^a} \end{bmatrix} \begin{bmatrix} \lambda_t \\ [\mu_t]_{\mathcal{I}_t^a} \end{bmatrix} = H_t d_t + \nabla f_t,$$

where the entries of $[\mu_t]_{\mathcal{I}_t^a}$ corresponding to inactive constraints are zero. Assumption 5, together with the uniform bounds on ∇f_t and d_t , therefore yield

$$\left\| \begin{bmatrix} \lambda_t \\ [\mu_t]_{\mathcal{I}_t^a} \end{bmatrix} \right\| \leq \frac{\sigma_u M_d + M_f}{\nu} =: M_{\lambda\mu}. \quad (3.20)$$

In particular, we obtain $\|\lambda_t\| \leq M_{\lambda\mu}$ and $\|\mu_t\| \leq M_{\lambda\mu}$.

We next establish a uniform perturbation bound. Fix an iteration at which a total step is computed and write $\mathcal{I} = \mathcal{I}_t^a$. Retaining only the bounds indexed by $\mathcal{I}$, we define

$$y = (d, \lambda, \mu_{\mathcal{I}}), \quad K = \mathbb{R}^n \times \mathbb{R}^m \times \mathbb{R}_+^{|\mathcal{I}|}, \quad F_t(y; \omega) := \begin{bmatrix} H_t d + \nabla f - \nabla c \lambda - I_{\mathcal{I}} \mu_{\mathcal{I}} \\ \nabla c^\top (d - v) \\ I_{\mathcal{I}}^\top (x_t + d) \end{bmatrix}, \quad \omega = (\nabla f, \nabla c, v).$$

Set $\omega_t = (\nabla f_t, \nabla c_t, v_t)$, $\tilde{\omega}_t = (\tilde{\nabla} f_t, \tilde{\nabla} c_t, \tilde{v}_t)$, and $y_t = (d_t, \lambda_t, [\mu_t]_{\mathcal{I}})$. All omitted bounds are inactive at d_t and have zero multipliers, then it holds that $0 \in F_t(y_t; \omega_t) + \mathcal{N}_K(y_t)$. For any $r = (r_d, r_e, r_{\mathcal{I}})$, the generalized equation $r \in F_t(y; \omega_t) + \mathcal{N}_K(y)$ is the KKT system of

$$\min_d \quad \frac{1}{2} d^\top H_t d + (\nabla f_t - r_d)^\top d \quad \text{s.t.} \quad \nabla c_t^\top (d - v_t) = r_e, \quad I_{\mathcal{I}}^\top (x_t + d) \geq r_{\mathcal{I}}.$$

By Assumption 5, $[\nabla c_t \ I_{\mathcal{I}}]^\top$ is surjective. Hence, these constraints are feasible for every r , even if all retained bounds are imposed as equalities. The objective is strongly convex, and every active subset satisfies LICQ. Thus, the primal-dual solution is unique for every r , which is denoted by $S_t(r)$.

For each subset $\mathcal{J} \subseteq \mathcal{I}_t^a$ of the bounds retained in the auxiliary quadratic program, define $A_{t,\mathcal{J}} := [\nabla c_t \ I_{\mathcal{J}}]$. The corresponding active-set KKT matrix is

$$\mathcal{K}_{t,\mathcal{J}} := \begin{bmatrix} H_t & -A_{t,\mathcal{J}} \\ A_{t,\mathcal{J}}^\top & 0 \end{bmatrix}.$$

For every $\mathcal{J} \subseteq \mathcal{I}_t^a$, extending the coefficient vector by zeros and applying Assumption 5 gives $\|A_{t,\mathcal{J}} z\| \geq \nu \|z\|$ for all $z \in \mathbb{R}^{m+|\mathcal{J}|}$. Consequently, it holds that $A_{t,\mathcal{J}}^\top H_t^{-1} A_{t,\mathcal{J}} \geq \frac{\nu^2}{\sigma_u} I$. Moreover, it holds that $\|H_t^{-1}\| \leq 1/\sigma_l$

and $\|A_{t,\mathcal{J}}\| \leq \sqrt{L_c^2 + 1}$. The Schur-complement formula therefore shows that there exists a constant $C_K > 0$, independent of t and $\mathcal{J}$, such that $\|\mathcal{K}_{t,\mathcal{J}}^{-1}\| \leq C_K$ for every $\mathcal{J} \subseteq \mathcal{I}_t^a$. For a fixed active subset, the KKT solution is affine in r . Its validity region is a closed polyhedron defined by primal feasibility and nonnegative multipliers. These finitely many regions cover the perturbation space, and the affine expressions agree on overlaps by uniqueness. Each affine piece has linear part of norm at most C_K , since selecting active right-hand-side components and inserting zero inactive multipliers have norm at most one. Partitioning a line segment between r and r' across these regions gives $\|S_t(r) - S_t(r')\| \leq L_R \|r - r'\|$ for all r and r' , where $L_R := C_K$. In particular, $S_t(0) = y_t$. Set $C_0 := 1 + M_{\lambda\mu} + M_d + M_v + L_c$ and choose

$$\delta_R := \min \left\{ \frac{1}{2L_R}, \frac{\varrho}{4L_R C_0} \right\}. \quad (3.21)$$

Both L_R and δ_R are independent of t , T , and the prescribed tolerances.

Furthermore, we define the good event

$$\mathcal{C}_t := \left\{ \max\{\|\tilde{\nabla}c_t - \nabla c_t\|, \|\tilde{\nabla}f_t - \nabla f_t\|, \|\tilde{v}_t - v_t\|\} \leq \delta_R \right\} \quad (3.22)$$

and denote its complement by $\mathcal{C}_t^c$. Unlike $\bar{\mathcal{B}}$, which guarantees normal-step contraction throughout the restoration iterations, $\mathcal{C}_t$ localizes the KKT perturbation analysis at iteration t . On $\mathcal{C}_t$, we use the local sensitivity estimate, while on its complement, we rely on uniform total-step bounds. By the law of total expectation, we decompose the expected deviation over the probability space into two complementary components, i.e.,

$$\mathbb{E}[\|\tilde{d}_t - d_t\|] = \mathbb{E}[\|\tilde{d}_t - d_t\| \cdot \mathbb{I}_{\mathcal{C}_t}] + \mathbb{E}[\|\tilde{d}_t - d_t\| \cdot \mathbb{I}_{\mathcal{C}_t^c}]. \quad (3.23)$$

Bound on the event $\mathcal{C}_t$. Define the perturbation mapping $h_t(y) := F_t(y; \tilde{\omega}_t) - F_t(y; \omega_t)$. For any $y = (d, \lambda, \mu_{\mathcal{I}})$ and $z = (d', \lambda', \mu'_{\mathcal{I}})$, we have

$$h_t(y) - h_t(z) = \begin{bmatrix} -(\tilde{\nabla}c_t - \nabla c_t)(\lambda - \lambda') \\ (\tilde{\nabla}c_t - \nabla c_t)^\top (d - d') \\ \mathbf{0} \end{bmatrix}.$$

Consequently, it holds that $\text{Lip}(h_t) \leq \|\tilde{\nabla}c_t - \nabla c_t\|$, where $\text{Lip}(\cdot)$ denotes the global Lipschitz modulus of h_t as defined in (A.1) (see Lemma 14). On the event $\mathcal{C}_t$, the definition of δ_R gives

$$L_R \text{Lip}(h_t) \leq L_R \|\tilde{\nabla}c_t - \nabla c_t\| \leq L_R \delta_R \leq \frac{1}{2}. \quad (3.24)$$

Applying Lemma 14 with the inverse S_t established above, the perturbed auxiliary KKT system admits a unique solution $\hat{y}_t = (\hat{d}_t, \hat{\lambda}_t, \hat{\mu}_{t,\mathcal{I}})$ satisfying

$$\|\hat{y}_t - y_t\| \leq \frac{L_R}{1 - L_R \text{Lip}(h_t)} \|h_t(y_t)\| \leq 2L_R \|h_t(y_t)\|. \quad (3.25)$$

At y_t , it holds that

$$h_t(y_t) = \begin{bmatrix} (\tilde{\nabla}f_t - \nabla f_t) - (\tilde{\nabla}c_t - \nabla c_t)\lambda_t \\ (\tilde{\nabla}c_t - \nabla c_t)^\top (d_t - v_t) - \tilde{\nabla}c_t^\top (\tilde{v}_t - v_t) \\ \mathbf{0} \end{bmatrix}.$$

Using the boundedness of $\|\lambda_t\|$, $\|d_t\|$, $\|v_t\|$ and $\|\tilde{\nabla}c_t\|$, we obtain

$$\|h_t(y_t)\| \leq \|\tilde{\nabla}f_t - \nabla f_t\| + (M_{\lambda\mu} + M_d + M_v) \|\tilde{\nabla}c_t - \nabla c_t\| + L_c \|\tilde{v}_t - v_t\|. \quad (3.26)$$

On event $\mathcal{C}_t$, (3.22) and (3.26) imply $\|h_t(y_t)\| \leq C_0 \delta_R$. Further combining with (3.25) gives

$$\|\hat{d}_t - d_t\| \leq \|\hat{y}_t - y_t\| \leq 2L_R C_0 \delta_R \leq \frac{\varrho}{2}. \quad (3.27)$$

By the definition of $\mathcal{I}_t^a$ in Assumption 5, we have $[x_t + d_t]_i > \varrho$ for every $i \notin \mathcal{I}_t^a$. Hence, (3.27) yields $[x_t + \hat{d}_t]_i \geq [x_t + d_t]_i - \|\hat{d}_t - d_t\| > \frac{\varrho}{2}, i \notin \mathcal{I}_t^a$, thus the non-negative constraints $[x_t + d]_i \geq 0$ with $i \notin \mathcal{I}_t^a$

are strictly satisfied at $d = \hat{d}_t$. Extending $\hat{\mu}_{t,\mathcal{I}}$ by zeros outside $\mathcal{I}_t^a$ therefore yields a KKT point of the full sampled subproblem (2.3). Since this subproblem is strongly convex, its primal minimizer is unique, and hence $\hat{d}_t = \tilde{d}_t$ on $\mathcal{C}_t$. Combining this identity with (3.25) and (3.26), we obtain

$$\|\tilde{d}_t - d_t\| \cdot \mathbb{I}_{\mathcal{C}_t} \leq L_d \left(\|\tilde{\nabla} f_t - \nabla f_t\| + \|\tilde{\nabla} c_t - \nabla c_t\| + \|\tilde{v}_t - v_t\| \right) \mathbb{I}_{\mathcal{C}_t}, \quad (3.28)$$

where $L_d := 2L_R \max\{1, M_{\lambda\mu} + M_d + M_v, L_c\}$. Taking expectations in (3.28), using $0 \leq \mathbb{I}_{\mathcal{C}_t} \leq 1$, and applying the definitions of $E_{t,f}$ and $E_{t,J}$ gives

$$\mathbb{E}[\|\tilde{d}_t - d_t\| \mathbb{I}_{\mathcal{C}_t}] \leq L_d (E_{t,f} + E_{t,J} + \mathbb{E}[\|\tilde{v}_t - v_t\|]). \quad (3.29)$$

Bound on the event $\mathcal{C}_t^c$. The total step d_t and $\tilde{d}_t$ are bounded by M_d , implying $\|\tilde{d}_t - d_t\| \leq \|\tilde{d}_t\| + \|d_t\| \leq 2M_d$. This yields a worst-case error bounded by $\mathbb{E}[\|\tilde{d}_t - d_t\| \cdot \mathbb{I}_{\mathcal{C}_t^c}] \leq 2M_d \cdot \mathbb{P}(\mathcal{C}_t^c)$. By the definition of $\mathcal{C}_t$, the union bound and Markov's inequality give

$$\mathbb{P}(\mathcal{C}_t^c) \leq \mathbb{P}(\|\tilde{\nabla} c_t - \nabla c_t\| > \delta_R) + \mathbb{P}(\|\tilde{\nabla} f_t - \nabla f_t\| > \delta_R) + \mathbb{P}(\|\tilde{v}_t - v_t\| > \delta_R) \leq \frac{E_{t,J} + E_{t,f} + \mathbb{E}[\|\tilde{v}_t - v_t\|]}{\delta_R}.$$

By combining the above inequalities, we further obtain

$$\mathbb{E}[\|\tilde{d}_t - d_t\| \cdot \mathbb{I}_{\mathcal{C}_t^c}] \leq \frac{2M_d}{\delta_R} (E_{t,J} + E_{t,f} + \mathbb{E}[\|\tilde{v}_t - v_t\|]). \quad (3.30)$$

Finally, plugging (3.29) and (3.30) back into the law of total expectation (3.23) completes the proof. $\square$

We now combine the preceding estimates to control the true total step norm and the equality-constraint violation. The true total step bound combines the good-event descent estimate, the total-step perturbation bound, and the contribution of $\tilde{\mathcal{B}}^c$. The feasibility bound instead uses normal-step contraction and the constraint value estimation error. For notational convenience, we collect these contributions into $\Delta_{T,1}$ and $\Delta_{T,2}$, respectively, with ϵ' added separately to the feasibility bound. Combining the estimates in preceding Lemmas, we obtain the following explicit bounds:

$$\begin{aligned} \Delta_{T,1} := & \frac{2\ell_{\max}(f_1 - f^* + \rho_{\max}M_c)}{\sigma_l K_T} + \frac{(4\ell_{\max} + \sigma_l) \sum_{t=1}^T \kappa_t^2}{4\sigma_l \ell_{\min} K_T} + \frac{4\ell_{\max} \rho_{\max} \sum_{t=1}^T E_{t,c}}{\sigma_l K_T} + \frac{(L_d + 2M_d/\delta_R)\epsilon'}{\sqrt{\gamma_{\min}}} \\ & + \frac{1}{K_T} \sum_{t=1}^T \kappa_t \left(\left(\frac{2\ell_{\max}}{\ell_{\min} \sigma_l} + L_d + \frac{2M_d}{\delta_R} \right) E_{t,f} + \left(L_d + \frac{2M_d}{\delta_R} \right) \frac{L_c E_{t,c}}{\gamma_{\min}} \right. \\ & \left. + \left(\frac{2\ell_{\max} \rho_{\max}}{\ell_{\min} \sigma_l} + \left(L_d + \frac{2M_d}{\delta_R} \right) \left(1 + \frac{2M_v L_c + M_c}{\gamma_{\min}} \right) \right) E_{t,J} \right) + M_d \sum_{t=1}^T \delta_t, \end{aligned} \quad (3.31)$$

$$\begin{aligned} \Delta_{T,2} := & \frac{2\ell_{\max} L_c (f_1 - f^* + \rho_{\max}M_c)}{(1 - \vartheta) \sigma_l K_T} + \frac{(4\ell_{\max} + \sigma_l) L_c \sum_{t=1}^T \kappa_t^2}{4(1 - \vartheta) \sigma_l \ell_{\min} K_T} + \frac{\sum_{t=1}^T \kappa_t E_{t,c}}{K_T} \\ & + \frac{2L_c \sum_{t=1}^T (\ell_{\max} \kappa_t E_{t,f}/\ell_{\min} + 2\ell_{\max} \rho_{\max} E_{t,c} + \ell_{\max} \rho_{\max} \kappa_t E_{t,J}/\ell_{\min})}{(1 - \vartheta) \sigma_l K_T} + M_c \sum_{t=1}^T \delta_t. \end{aligned} \quad (3.32)$$

These quantities satisfy

$$\mathcal{L}_T + \frac{\sum_{t=1}^T \kappa_t \mathcal{R}_t}{K_T} + M_d \mathbb{P}(\tilde{\mathcal{B}}^c) \leq \Delta_{T,1}, \quad \frac{\sum_{t=1}^T \kappa_t E_{t,c}}{K_T} + \frac{L_c}{1 - \vartheta} \mathcal{L}_T + M_c \mathbb{P}(\tilde{\mathcal{B}}^c) \leq \Delta_{T,2}. \quad (3.33)$$

The following theorem presents the expected KKT residual bounds for the randomized output of Algorithm 1, conditioned on $\bar{\mathcal{S}}$, the event that all prescribed iterations are completed.

Theorem 1 *Suppose Assumptions 4, 5, and Conditions 1, 2 hold, and let $\epsilon' > 0$ and $\eta_c \in (0, 1)$ be given. Let $\{\delta_t\}_{t=1}^T \subset (0, 1)$ be a prescribed deterministic sequence satisfying $\sum_{t=1}^T \delta_t < 1$, such that (3.6) holds almost*

surely for every generated iteration t . Then Algorithm 1 completes all T prescribed iterations with probability $\mathbb{P}(\bar{\mathcal{S}}) \geq 1 - \sum_{t=1}^T \delta_t$. On $\bar{\mathcal{S}}$, let x_R denote its randomized output. Then, it holds that

$$\mathbb{E}[\|d_R\| \mid \bar{\mathcal{S}}] \leq \Delta_{T,1}, \quad \mathbb{E}[\|c_R\| \mid \bar{\mathcal{S}}] \leq \epsilon' + \Delta_{T,2}. \quad (3.34)$$

Moreover, given a tolerance $\epsilon \in (0, 1)$, if

$$\Delta_{T,1} \leq \min \left\{ \frac{\epsilon}{\sigma_u}, \frac{\nu\epsilon}{\sigma_u M_d + M_f} \right\}, \quad \epsilon' + \Delta_{T,2} \leq \epsilon, \quad (3.35)$$

there exist $\lambda_R \in \mathbb{R}^m$ and $\mu_R \in \mathbb{R}_{\geq 0}^n$ such that

$$\mathbb{E}[\|\nabla f_R + \nabla c_R \lambda_R - \mu_R\| \mid \bar{\mathcal{S}}] \leq \epsilon, \quad \mathbb{E}[|\mu_R^\top x_R| \mid \bar{\mathcal{S}}] \leq \epsilon, \quad \mathbb{E}[\|c_R\| \mid \bar{\mathcal{S}}] \leq \epsilon. \quad (3.36)$$

Proof. By Lemma 3, it holds that $\bar{\mathcal{B}} \subseteq \bar{\mathcal{S}}$ and $\mathbb{P}(\bar{\mathcal{S}}) \geq 1 - \sum_{t=1}^T \delta_t > 0$. Since $\bar{\mathcal{S}} = \bar{\mathcal{B}} \cup (\bar{\mathcal{S}} \setminus \bar{\mathcal{B}})$, we have

$$\mathbb{P}(\bar{\mathcal{S}} \setminus \bar{\mathcal{B}}) = \mathbb{P}(\bar{\mathcal{B}}^c) - \mathbb{P}(\bar{\mathcal{S}}^c) \leq \sum_{t=1}^T \delta_t - \mathbb{P}(\bar{\mathcal{S}}^c).$$

We first bound the true total step d_R on the event $\bar{\mathcal{S}}$. By the triangle inequality and the randomized output rule of Algorithm 1, we apply Lemmas 6, 7 and 8 to obtain

$$\mathbb{E}[\|d_R\| \cdot \mathbb{I}_{\bar{\mathcal{B}}} \mid \bar{\mathcal{S}}] \leq \mathbb{E}[\|\tilde{d}_R\| \cdot \mathbb{I}_{\bar{\mathcal{B}}} \mid \bar{\mathcal{S}}] + \mathbb{E}[\|d_R - \tilde{d}_R\| \cdot \mathbb{I}_{\bar{\mathcal{B}}} \mid \bar{\mathcal{S}}] \leq \mathcal{L}_T + \frac{\sum_{t=1}^T \kappa_t \mathcal{R}_t}{K_T} \leq \Delta_{T,1} - M_d \sum_{t=1}^T \delta_t,$$

where we use $\mathbb{E}[\|\tilde{d}_t - d_t\| \cdot \mathbb{I}_{\bar{\mathcal{B}}} \mid \bar{\mathcal{S}}] \leq \mathbb{E}[\|\tilde{d}_t - d_t\| \mid \bar{\mathcal{S}}]$. On event $\bar{\mathcal{S}} \setminus \bar{\mathcal{B}}$, the uniform bound $\|d_R\| \leq M_d$ gives

$$\mathbb{E}[\|d_R\| \cdot \mathbb{I}_{\bar{\mathcal{S}} \setminus \bar{\mathcal{B}}} \mid \bar{\mathcal{S}}] \leq M_d \mathbb{P}(\bar{\mathcal{S}} \setminus \bar{\mathcal{B}}) \leq M_d \left(\sum_{t=1}^T \delta_t - \mathbb{P}(\bar{\mathcal{S}}^c) \right).$$

Combining these estimates yields $\mathbb{E}[\|d_R\| \cdot \mathbb{I}_{\bar{\mathcal{S}}} \mid \bar{\mathcal{S}}] \leq \Delta_{T,1} - M_d \mathbb{P}(\bar{\mathcal{S}}^c)$. If $\Delta_{T,1} \leq M_d$, then $\Delta_{T,1} - M_d \mathbb{P}(\bar{\mathcal{S}}^c) \leq \Delta_{T,1}(1 - \mathbb{P}(\bar{\mathcal{S}}^c)) = \Delta_{T,1} \mathbb{P}(\bar{\mathcal{S}})$. Dividing by $\mathbb{P}(\bar{\mathcal{S}})$ gives $\mathbb{E}[\|d_R\| \mid \bar{\mathcal{S}}] \leq \Delta_{T,1}$. If $\Delta_{T,1} > M_d$, the same conclusion follows directly from $\|d_R\| \leq M_d$ on $\bar{\mathcal{S}}$. Thus, we have $\mathbb{E}[\|d_R\| \mid \bar{\mathcal{S}}] \leq \Delta_{T,1}$.

We next bound the true constraint violation on the event $\bar{\mathcal{S}}$. On $\bar{\mathcal{B}}$, if $\|\tilde{c}_t\| \leq \epsilon'$, it follows that $\|c_t\| \leq \epsilon' + \|\tilde{c}_t - c_t\|$. Otherwise, Lemma 4 and the total-step equality constraint yield $(1 - \vartheta)\|\tilde{c}_t\| \leq \|\tilde{\nabla} c_t^\top \tilde{v}_t\| = \|\tilde{\nabla} c_t^\top \tilde{d}_t\| \leq L_c \|\tilde{d}_t\|$. Thus, in both cases, we obtain $\|c_t\| \leq \epsilon' + \|\tilde{c}_t - c_t\| + \frac{L_c}{1 - \vartheta} \|\tilde{d}_t\|$ on $\bar{\mathcal{B}}$. Using the randomized output rule on $\bar{\mathcal{B}}$ and Lemma 6 gives

$$\mathbb{E}[\|c_R\| \cdot \mathbb{I}_{\bar{\mathcal{B}}} \mid \bar{\mathcal{S}}] \leq \epsilon' + \frac{\sum_{t=1}^T \kappa_t E_{t,c}}{K_T} + \frac{L_c}{1 - \vartheta} L_T \leq \epsilon' + \frac{\sum_{t=1}^T \kappa_t E_{t,c}}{K_T} + \frac{L_c}{1 - \vartheta} \mathcal{L}_T \leq \epsilon' + \Delta_{T,2} - M_c \sum_{t=1}^T \delta_t.$$

On the event $\bar{\mathcal{S}} \setminus \bar{\mathcal{B}}$, the bound $\|c_R\| \leq M_c$ implies

$$\mathbb{E}[\|c_R\| \cdot \mathbb{I}_{\bar{\mathcal{S}} \setminus \bar{\mathcal{B}}} \mid \bar{\mathcal{S}}] \leq M_c \mathbb{P}(\bar{\mathcal{S}} \setminus \bar{\mathcal{B}}) \leq M_c \left(\sum_{t=1}^T \delta_t - \mathbb{P}(\bar{\mathcal{S}}^c) \right).$$

Hence, combining these two inequalities yields $\mathbb{E}[\|c_R\| \cdot \mathbb{I}_{\bar{\mathcal{S}}} \mid \bar{\mathcal{S}}] \leq \epsilon' + \Delta_{T,2} - M_c \mathbb{P}(\bar{\mathcal{S}}^c)$. If $\epsilon' + \Delta_{T,2} \leq M_c$, the last expression is bounded above by $(\epsilon' + \Delta_{T,2})(1 - \mathbb{P}(\bar{\mathcal{S}}^c)) = (\epsilon' + \Delta_{T,2}) \mathbb{P}(\bar{\mathcal{S}})$. Dividing by $\mathbb{P}(\bar{\mathcal{S}})$ gives $\mathbb{E}[\|c_R\| \mid \bar{\mathcal{S}}] \leq \epsilon' + \Delta_{T,2}$. If $\epsilon' + \Delta_{T,2} > M_c$, the same conclusion follows from $\|c_R\| \leq M_c$ on $\bar{\mathcal{S}}$. This proves (3.34).

Finally, we derive the KKT residual bounds on the event $\bar{\mathcal{S}}$. At each iteration on $\bar{\mathcal{S}}$, choose the multipliers (λ_t, μ_t) of the true total-step subproblem with the sign convention $\nabla f_t + \nabla c_t \lambda_t - \mu_t = -H_t d_t$ and $\mu_t^\top (x_t + d_t) = 0$. Since $\|H_t\| \leq \sigma_u$, it follows that $\|\nabla f_t + \nabla c_t \lambda_t - \mu_t\| \leq \sigma_u \|d_t\|$. Taking expectations conditioned on $\bar{\mathcal{S}}$ under the randomized output rule gives

$$\mathbb{E}[\|\nabla f_R + \nabla c_R \lambda_R - \mu_R\| \mid \bar{\mathcal{S}}] \leq \sigma_u \mathbb{E}[\|d_R\| \mid \bar{\mathcal{S}}] \leq \sigma_u \Delta_{T,1}.$$

As presented in (3.20), $\|\mu_t\| \leq M_{\lambda\mu} = (\sigma_u M_d + M_f)/\nu$. Since $\mu_t^\top(x_t + d_t) = 0$ and $x_t, \mu_t \geq \mathbf{0}$, we obtain $0 \leq \mu_t^\top x_t = -\mu_t^\top d_t \leq (\sigma_u M_d + M_f)\|d_t\|/\nu$. Taking expectations conditioned on $\bar{\mathcal{S}}$ under the randomized output rule yields

$$\mathbb{E}[\|\mu_R^\top x_R\| \mid \bar{\mathcal{S}}] \leq M_{\lambda\mu} \mathbb{E}[\|d_R\| \mid \bar{\mathcal{S}}] \leq \frac{(\sigma_u M_d + M_f)\Delta_{T,1}}{\nu}.$$

If $\Delta_{T,1} \leq \min\{\epsilon/\sigma_u, \nu\epsilon/(\sigma_u M_d + M_f)\}$ and $\epsilon' + \Delta_{T,2} \leq \epsilon$, then we can verify the stationarity, feasibility, and complementarity requirements (3.36). $\square$

Therefore, it remains to choose the iteration horizon T , sequence κ_t , batch sizes, and probability budgets δ_t so that (3.6) and (3.35) are satisfied, in order to derive the sample complexity bounds for Algorithm 1.

4 Sample Complexity Analysis

In this section, we will establish the sample complexity of Algorithm 1 to return an expected ϵ -KKT point conditioned on $\bar{\mathcal{S}}$, the event that all prescribed iterations are completed. The probability budgets are chosen so that $\sum_{t=1}^T \delta_t \leq \epsilon$, ensuring that $\bar{\mathcal{S}}$ occurs with probability at least $1 - \epsilon$. Throughout this section, we omit constant factors depending on the problem parameters and algorithmic parameters, while explicitly tracking the dependence on tolerance ϵ , maximum iteration number T , tail index p , sequence κ_t , batch sizes and probability budgets δ_t . Recall that Theorem 1 requires the error terms bounded by $\mathcal{O}(\epsilon)$. It then suffices to show that

$$\max \left\{ \Delta_{T,1}, \epsilon' + \Delta_{T,2}, \max_{1 \leq t \leq T} \mathcal{E}'_t / \delta_t^{1/p} \right\} = \mathcal{O}(\epsilon).$$

For each fixed $p \in (1, 2)$, the proportionality constants in the parameter choices below are independent of ϵ and are chosen to satisfy the accuracy requirements of the preceding convergence theorem. To this end, we construct estimators satisfying Conditions 1 and 2 and quantify their estimation errors.

4.1 Complexity bounds with mini-batch estimation

In this subsection, we will address the sample complexity of Algorithm 1 using mini-batch estimators. To ensure the uniform boundedness required in Condition 1 under heavy-tailed noise, we project the mini-batch averages onto prescribed bounded sets. Specifically, the stochastic estimators are computed by

$$\tilde{\nabla} f_t = \Pi_{\mathbb{B}(M_f)} \left(\frac{\sum_{i=1}^{B_{t,f}} \nabla \mathbf{F}(x_t; \xi_t^{(i)})}{B_{t,f}} \right), \quad \tilde{c}_t = \Pi_{\mathbb{B}(M_c)} \left(\frac{\sum_{i=1}^{B_{t,c}} \mathbf{C}(x_t; \zeta_t^{(i)})}{B_{t,c}} \right), \quad \tilde{\nabla} c_t = \Pi_{\mathbb{B}(L_c)} \left(\frac{\sum_{i=1}^{B_{t,J}} \nabla \mathbf{C}(x_t; \bar{\zeta}_t^{(i)})}{B_{t,J}} \right), \quad (4.1)$$

where $B_{t,f}$, $B_{t,c}$, and $B_{t,J}$ are the corresponding batch sizes, and $\mathbb{B}(\cdot)$ denotes the norm ball. Conditionally on the history $\mathcal{F}_t$ prior to iteration t , the three batches are mutually independent, and the samples within each batch are independent and identically distributed. By construction, it follows that $\|\tilde{\nabla} f_t\| \leq M_f$, $\|\tilde{c}_t\| \leq M_c$ and $\|\tilde{\nabla} c_t\| \leq \|\tilde{\nabla} c_t\|_F \leq L_c$, which ensures that the uniform boundedness in Condition 1 holds.

While such a truncation mechanism inherently introduces an estimation bias, its magnitude can be effectively controlled through appropriately chosen batch sizes. The following lemma bounds the expected estimation errors of the constructed stochastic estimators, thereby ensuring that Condition 2 holds.

Lemma 9 *Suppose that Assumptions 1, 2, and 3 hold. Then, for any $t \in [T]$, the expected estimation errors of the truncated estimators constructed in (4.1) satisfy, almost surely,*

$$\begin{aligned} \mathbb{E}_t[\|\tilde{\nabla} f_t - \nabla f_t\|^p] &\leq 2^{2-p} V_f^p B_{t,f}^{1-p}, \quad \mathbb{E}_t[\|\tilde{c}_t - c_t\|^p] \leq 2^{2-p} V_c^p B_{t,c}^{1-p}, \quad \mathbb{E}_t[\|\tilde{\nabla} c_t - \nabla c_t\|_F^p] \leq 2^{2-p} V_J^p B_{t,J}^{1-p}, \\ \mathbb{E}_t[\|\tilde{\nabla} f_t - \nabla f_t\|] &\leq E_{t,f}, \quad \mathbb{E}_t[\|\tilde{c}_t - c_t\|] \leq E_{t,c}, \quad \mathbb{E}_t[\|\tilde{\nabla} c_t - \nabla c_t\|] \leq E_{t,J}, \end{aligned}$$

where

$$E_{t,i} := 2^{\frac{2-p}{p}} V_i B_{t,i}^{\frac{1-p}{p}} \quad \text{for } i \in \{f, c, J\}. \quad (4.2)$$

Proof. For brevity, we only provide the detailed derivation for the upper bound of $\mathbb{E}[\|\tilde{\nabla} f_t - \nabla f_t\|^p]$ and $\mathbb{E}[\|\tilde{\nabla} f_t - \nabla f_t\|]$. The upper bounds for the remaining terms follow from analogous arguments, with the

only exception that bounding $\mathbb{E}[\|\tilde{\nabla}c_t - \nabla c_t\|]$ additionally employs the matrix norm inequality $\|A\| \leq \|A\|_F$. Note that $\Pi_{B(M_f)}(\nabla f(x_t)) = \nabla f(x_t)$ holds under Assumption 1. By the non-expansiveness of the Euclidean projection operator $\Pi_{B(M_f)}(\cdot)$, we can bound the approximation error as

$$\mathbb{E}_t[\|\tilde{\nabla}f_t - \nabla f_t\|^p] \leq \mathbb{E}_t\left[\left\|\frac{1}{B_{t,f}} \sum_{i=1}^{B_{t,f}} \left(\nabla \mathbf{F}(x_t; \xi_t^{(i)}) - \nabla f(x_t)\right)\right\|^p\right] \leq \frac{1}{B_{t,f}^p} \mathbb{E}_t\left[\left\|\sum_{i=1}^{B_{t,f}} \left(\nabla \mathbf{F}(x_t; \xi_t^{(i)}) - \nabla f(x_t)\right)\right\|^p\right].$$

Denote the sample-wise filtration

$$\mathcal{G}_{t,0}^f := \mathcal{F}_t, \quad \mathcal{G}_{t,i}^f := \mathcal{F}_t \vee \sigma(\xi_t^{(1)}, \dots, \xi_t^{(i)}), \quad i = 1, \dots, B_{t,f}.$$

Here, $\vee$ denotes the smallest σ -algebra containing both constituent σ -algebras. Let $X_t^i := \nabla \mathbf{F}(x_t; \xi_t^{(i)}) - \nabla f(x_t)$ for $i = 1, \dots, B_{t,f}$ denote the gradient estimation errors. The current batch is sampled independently of $\mathcal{F}_t$, and the samples within the batch are independently and identically distributed. Since x_t is $\mathcal{F}_t$ -measurable and the stochastic gradient oracle is unbiased, it holds that $\mathbb{E}[X_t^i \mid \mathcal{G}_{t,i-1}^f] = \mathbb{E}[\nabla \mathbf{F}(x_t; \xi_t^{(i)}) \mid \mathcal{G}_{t,i-1}^f] - \nabla f(x_t) = \mathbf{0}$. Moreover, according to Assumption 3, the p th moment of each individual error satisfies $\mathbb{E}[\|X_t^i\|^p \mid \mathcal{G}_{t,i-1}^f] \leq V_f^p < +\infty$. Applying Lemma 13 conditionally on $\mathcal{F}_t$ yields

$$\mathbb{E}_t\left[\left\|\sum_{i=1}^{B_{t,f}} \left(\nabla \mathbf{F}(x_t; \xi_t^{(i)}) - \nabla f(x_t)\right)\right\|^p\right] \leq 2^{2-p} \sum_{i=1}^{B_{t,f}} \mathbb{E}_t\left[\|\nabla \mathbf{F}(x_t; \xi_t^{(i)}) - \nabla f(x_t)\|^p\right] \leq 2^{2-p} V_f^p B_{t,f}.$$

Hence, combining the preceding two inequalities gives $\mathbb{E}_t[\|\tilde{\nabla}f_t - \nabla f_t\|^p] \leq 2^{2-p} V_f^p B_{t,f}^{1-p}$. By invoking Jensen's inequality, we further obtain $\mathbb{E}_t[\|\tilde{\nabla}f_t - \nabla f_t\|] \leq \mathbb{E}_t[\|\tilde{\nabla}f_t - \nabla f_t\|^p]^{\frac{1}{p}} \leq 2^{\frac{2-p}{p}} V_f^{\frac{1-p}{p}} B_{t,f}^{\frac{1-p}{p}} = E_{t,f}$. $\square$

The sample complexity of Algorithm 1 with mini-batch estimation is provided in the following theorem.

Theorem 2 *Suppose Assumptions 1, 2, 3, 4 and 5 hold. Given a tolerance $\epsilon \in (0, 1)$ and $t \in [T]$, set $\epsilon' = c_\epsilon \epsilon$, where $c_\epsilon > 0$ is a sufficiently small fixed constant independent of ϵ , and choose the parameters*

$$\kappa_t \asymp \epsilon, \quad B_{t,f} \asymp \left\lceil \epsilon^{\frac{p}{1-p}} \right\rceil, \quad B_{t,c} \asymp \left\lceil \epsilon^{\frac{p+3}{1-p}} \right\rceil, \quad B_{t,J} \asymp \left\lceil \epsilon^{\frac{p+3}{1-p}} \right\rceil, \quad \delta_t \asymp \frac{\epsilon}{T} \text{ with } \sum_{t=1}^T \delta_t \leq \epsilon.$$

Then Algorithm 1 with mini-batch estimation completes all prescribed iterations with probability at least $1 - \epsilon$, and conditioned on this, it returns an expected ϵ -KKT point after $T \asymp \lceil \epsilon^{-2} \rceil$ iterations, with sample complexity of order $\mathcal{O}(\epsilon^{\frac{3p-2}{1-p}})$, $\mathcal{O}(\epsilon^{\frac{3p+1}{1-p}})$ and $\mathcal{O}(\epsilon^{\frac{3p+1}{1-p}})$ for computing $\tilde{\nabla}f_t$, $\tilde{\nabla}c_t$ and $\tilde{c}_t$, respectively.

Proof. The definitions of $\Delta_{T,1}$ in (3.31), $\Delta_{T,2}$ in (3.32), and $\mathcal{E}'_t$ in (3.6) give

$$\begin{aligned} \Delta_{T,1} &= \mathcal{O}\left(\frac{1 + \sum_{t=1}^T \kappa_t^2 + \sum_{t=1}^T E_{t,c}}{K_T} + \frac{\sum_{t=1}^T \kappa_t (E_{t,f} + E_{t,c} + E_{t,J})}{K_T} + \sum_{t=1}^T \delta_t + \epsilon'\right), \\ \Delta_{T,2} &= \mathcal{O}\left(\frac{1 + \sum_{t=1}^T \kappa_t^2 + \sum_{t=1}^T E_{t,c}}{K_T} + \frac{\sum_{t=1}^T \kappa_t (E_{t,f} + E_{t,c} + E_{t,J})}{K_T} + \sum_{t=1}^T \delta_t\right), \\ \frac{\mathcal{E}'_t}{\delta_t^{1/p}} &= \mathcal{O}\left(\left(\frac{\mathbb{E}_t[\|\tilde{c}_t - c_t\|^p]}{\delta_t}\right)^{\frac{1}{p}} + \left(\frac{\mathbb{E}_t[\|\tilde{\nabla}c_t - \nabla c_t\|_F^p]}{\delta_t}\right)^{\frac{1}{p}}\right). \end{aligned}$$

With the choices of parameters, we obtain

$$E_{t,f} = \mathcal{O}(\epsilon), \quad E_{t,c} = \mathcal{O}(\epsilon^{\frac{p+3}{p}}), \quad E_{t,J} = \mathcal{O}(\epsilon^{\frac{p+3}{p}}), \quad \mathbb{E}_t[\|\tilde{c}_t - c_t\|^p] = \mathcal{O}(\epsilon^{p+3}), \quad \mathbb{E}_t[\|\tilde{\nabla}c_t - \nabla c_t\|_F^p] = \mathcal{O}(\epsilon^{p+3}), \quad K_T \asymp \lceil \epsilon T \rceil.$$

Hence, it holds that

$$\Delta_{T,1} = \mathcal{O}\left(\frac{1}{\epsilon T} + \epsilon^{\frac{p+3}{p}} + \epsilon^{\frac{3}{p}} + \epsilon\right), \quad \epsilon' + \Delta_{T,2} = \mathcal{O}\left(\frac{1}{\epsilon T} + \epsilon^{\frac{p+3}{p}} + \epsilon^{\frac{3}{p}} + \epsilon\right), \quad \mathcal{E}'_t / \delta_t^{1/p} = \mathcal{O}\left(\epsilon^{\frac{p+2}{p}} T^{\frac{1}{p}}\right).$$

With $T \asymp \lceil \epsilon^{-2} \rceil$, these estimates give $\Delta_{T,1} = \mathcal{O}(\epsilon)$, $\epsilon' + \Delta_{T,2} = \mathcal{O}(\epsilon)$ and $\mathcal{E}'_t/\delta_t^{1/p} = \mathcal{O}(\epsilon)$. Lemma 3 therefore yields $\mathbb{P}(\bar{\mathcal{S}}) \geq 1 - \sum_{t=1}^T \delta_t \geq 1 - \epsilon$. By applying Theorem 1, the randomized output conditioned on $\bar{\mathcal{S}}$ satisfies the ϵ -KKT conditions in expectation. Moreover, the total number of samples required for finding an expected ϵ -KKT point are

$$\sum_{t=1}^T B_{t,f} = \mathcal{O}\left(\epsilon^{\frac{p}{1-p}} T\right) = \mathcal{O}\left(\epsilon^{\frac{3p-2}{1-p}}\right), \quad \sum_{t=1}^T B_{t,J} = \mathcal{O}\left(\epsilon^{\frac{p+3}{1-p}} T\right) = \mathcal{O}\left(\epsilon^{\frac{3p+1}{1-p}}\right), \quad \sum_{t=1}^T B_{t,c} = \mathcal{O}\left(\epsilon^{\frac{p+3}{1-p}} T\right) = \mathcal{O}\left(\epsilon^{\frac{3p+1}{1-p}}\right),$$

respectively, which completes the proof. $\square$

Discussion 1 Compared with [11], which uses exact constraint information, Theorem 2 provides sample-complexity bounds for expected ϵ -KKT guarantees conditioned on event $\bar{\mathcal{S}}$, allowing simultaneous stochastic errors in the three oracles. As $p \uparrow 2$, the sample-complexity bound for objective-gradient evaluations has the limiting form $\mathcal{O}(\epsilon^{-4})$, matching the corresponding exponent in [15] up to logarithmic factors, whereas the bounds for constraint-value and Jacobian evaluations both have the limiting form $\mathcal{O}(\epsilon^{-7})$. These higher constraint-sampling costs arise in the analysis from controlling contraction failures through p th-moment error bounds, whereas [15] used sharper concentration bounds under almost-surely bounded original estimation errors.

4.2 Complexity bounds with SPIDER-type recursive estimation

In this subsection, we will establish sample complexity for Algorithm 1 using SPIDER-type recursive estimators. Specifically, Algorithm 1 proceeds in epochs, where each epoch $t \in [T]$ consists of τ_t inner recursive steps ($j \in \{0, \dots, \tau_t - 1\}$). At the beginning of each epoch ($j = 0$), we construct the estimators by identically applying the truncation (4.1) over large batches $B_{t,\cdot}$, i.e.,

$$\tilde{\nabla} f_{t,0} = \Pi_{\mathbb{B}(M_f)} \left(\frac{\sum_{i=1}^{B_{t,f}} \nabla \mathbf{F}(x_{t,0}; \xi_{t,0}^{(i)})}{B_{t,f}} \right), \quad \tilde{c}_{t,0} = \Pi_{\mathbb{B}(M_c)} \left(\frac{\sum_{i=1}^{B_{t,c}} \mathbf{C}(x_{t,0}; \zeta_{t,0}^{(i)})}{B_{t,c}} \right), \quad \tilde{\nabla} c_{t,0} = \Pi_{\mathbb{B}(L_c)}^F \left(\frac{\sum_{i=1}^{B_{t,J}} \nabla \mathbf{C}(x_{t,0}; \bar{\zeta}_{t,0}^{(i)})}{B_{t,J}} \right). \quad (4.3)$$

For the subsequent inner iterations $j \in \{1, \dots, \tau_t - 1\}$, the stochastic estimators are updated using relatively smaller sample sizes $S_{t,\cdot}$, i.e.,

$$\begin{aligned} \tilde{\nabla} f_{t,j} &= \Pi_{\mathbb{B}(M_f)} \left(\tilde{\nabla} f_{t,j-1} + \frac{1}{S_{t,f}} \sum_{i=1}^{S_{t,f}} \left(\nabla \mathbf{F}(x_{t,j}; \xi_{t,j}^{(i)}) - \nabla \mathbf{F}(x_{t,j-1}; \xi_{t,j}^{(i)}) \right) \right), \\ \tilde{\nabla} c_{t,j} &= \Pi_{\mathbb{B}(L_c)} \left(\tilde{\nabla} c_{t,j-1} + \frac{1}{S_{t,J}} \sum_{i=1}^{S_{t,J}} \left(\nabla \mathbf{C}(x_{t,j}; \bar{\zeta}_{t,j}^{(i)}) - \nabla \mathbf{C}(x_{t,j-1}; \bar{\zeta}_{t,j}^{(i)}) \right) \right), \\ \tilde{c}_{t,j} &= \Pi_{\mathbb{B}(M_c)} \left(\tilde{c}_{t,j-1} + \frac{1}{S_{t,c}} \sum_{i=1}^{S_{t,c}} \left(\mathbf{C}(x_{t,j}; \zeta_{t,j}^{(i)}) - \mathbf{C}(x_{t,j-1}; \zeta_{t,j}^{(i)}) \right) \right). \end{aligned} \quad (4.4)$$

In each SPIDER difference, the two oracle evaluations use the same sample, drawn independently of the history preceding the batch. Within the inner loop for $j = 0, \dots, \tau_t - 1$, the iterate is updated by taking a normalized-step along the stochastic direction $\tilde{d}_{t,j}$, i.e.,

$$x_{t,j+1} = x_{t,j} + \eta_{t,j} \tilde{d}_{t,j}, \quad \text{with} \quad \eta_{t,j} = \min \left\{ \frac{\kappa_t}{(L_f + \tilde{\rho}_{t,j} L_J) \|\tilde{d}_{t,j}\|}, 1 \right\}. \quad (4.5)$$

After completing the τ_t inner steps, we set the last inner iterate as the initial point for the next epoch, i.e., $x_{t+1,0} = x_{t,\tau_t}$. The merit parameter $\tilde{\rho}_{t,j}$ is updated after computing $\tilde{d}_{t,j}$ and before updating the iterate. The merit parameter is carried over between epochs and is not reinitialized when the estimators are refreshed. Algorithm 1 with SPIDER-type estimation strategy returns a random point x_R , which is selected from $\{x_{t,j} : 1 \leq t \leq T, 0 \leq j \leq \tau_t - 1\}$ according to the probability distribution $\mathbb{P}(R = (t, j)) = \kappa_t / K_T$ with $K_T := \sum_{k=1}^T \tau_k \kappa_k$ throughout this subsection.

Moreover, to analyze the estimation errors under SPIDER-type recursive estimation, we impose the following mean p th-moment smoothness condition on the stochastic estimates.

Assumption 6 *There exist constants $L_f > 0$, $L_c > 0$, and $L_J > 0$ such that for any $x, y \in \mathbb{R}^n$,*

$$\begin{aligned}\mathbb{E}[\|\nabla \mathbf{F}(x; \xi) - \nabla \mathbf{F}(y; \xi)\|^p] &\leq L_f^p \|x - y\|^p, & \mathbb{E}[\|\mathbf{C}(x; \zeta) - \mathbf{C}(y; \zeta)\|^p] &\leq L_c^p \|x - y\|^p, \\ \mathbb{E}[\|\nabla \mathbf{C}(x; \zeta) - \nabla \mathbf{C}(y; \zeta)\|_F^p] &\leq L_J^p \|x - y\|^p.\end{aligned}$$

Remark 3 *For notational simplicity, we use the same constants L_f , L_c , and L_J as in Assumption 2. Indeed, the unbiasedness of the stochastic oracles together with Jensen's inequality yields*

$$\|\nabla f(x) - \nabla f(y)\| = \|\mathbb{E}[\nabla \mathbf{F}(x; \xi) - \nabla \mathbf{F}(y; \xi)]\| \leq (\mathbb{E}[\|\nabla \mathbf{F}(x; \xi) - \nabla \mathbf{F}(y; \xi)\|^p])^{1/p} \leq L_f \|x - y\|.$$

Analogous arguments establish $\|c(x) - c(y)\| \leq L_c \|x - y\|$ and $\|\nabla c(x) - \nabla c(y)\| \leq L_J \|x - y\|$. Thus, Assumption 6 is a slightly stronger requirement that implies Assumption 2. In particular, if $p = 2$, Assumption 6 reduces to the widely adopted mean-squared smoothness condition.

Let $\mathcal{F}_{t,j}$ denote the history prior to the batch sampling of epoch (t, j) , and write $\mathbb{E}_{t,j}[\cdot] := \mathbb{E}[\cdot \mid \mathcal{F}_{t,j}]$ in this subsection. The following lemma establishes the estimation error bounds for the SPIDER-type estimators.

Lemma 10 *Suppose that Assumptions 1, 3 and 6 hold. For any $t \in [T]$ and $j \in \{0, \dots, \tau_t - 1\}$, the expected estimation errors of the stochastic estimators constructed in (4.3) and (4.4) satisfy, almost surely,*

$$\begin{aligned}\mathbb{E}_{t,0}[\|\tilde{\nabla} f_{t,j} - \nabla f_{t,j}\|^p] &\leq \frac{2^{4-p} \tau_t \kappa_t^p L_f^p S_{t,f}^{1-p}}{\ell_{\min}^p} + 2^{2-p} V_f B_{t,f}^{1-p}, & \mathbb{E}_{t,0}[\|\tilde{\nabla} f_{t,j} - \nabla f_{t,j}\|] &\leq \frac{2^{\frac{4-p}{p}} \tau_t^{\frac{1}{p}} \kappa_t L_f S_{t,f}^{\frac{1-p}{p}}}{\ell_{\min}} + 2^{\frac{2-p}{p}} V_f B_{t,f}^{\frac{1-p}{p}}, \\ \mathbb{E}_{t,0}[\|\tilde{c}_{t,j} - c_{t,j}\|^p] &\leq \frac{2^{4-p} \tau_t \kappa_t^p L_c^p S_{t,c}^{1-p}}{\ell_{\min}^p} + 2^{2-p} V_c B_{t,c}^{1-p}, & \mathbb{E}_{t,0}[\|\tilde{c}_{t,j} - c_{t,j}\|] &\leq \frac{2^{\frac{4-p}{p}} \tau_t^{\frac{1}{p}} \kappa_t L_c S_{t,c}^{\frac{1-p}{p}}}{\ell_{\min}} + 2^{\frac{2-p}{p}} V_c B_{t,c}^{\frac{1-p}{p}}, \\ \mathbb{E}_{t,0}[\|\tilde{\nabla} c_{t,j} - \nabla c_{t,j}\|_F^p] &\leq \frac{2^{4-p} \tau_t \kappa_t^p L_J^p S_{t,J}^{1-p}}{\ell_{\min}^p} + 2^{2-p} V_J B_{t,J}^{1-p}, & \mathbb{E}_{t,0}[\|\tilde{\nabla} c_{t,j} - \nabla c_{t,j}\|] &\leq \frac{2^{\frac{4-p}{p}} \tau_t^{\frac{1}{p}} \kappa_t L_J S_{t,J}^{\frac{1-p}{p}}}{\ell_{\min}} + 2^{\frac{2-p}{p}} V_J B_{t,J}^{\frac{1-p}{p}}.\end{aligned}$$

Proof. Given the structural symmetry among the three SPIDER-type estimators, we only detail the derivation for the objective gradient error $\mathbb{E}_{t,0}[\|\tilde{\nabla} f_{t,j} - \nabla f_{t,j}\|]$ and $\mathbb{E}_{t,0}[\|\tilde{\nabla} f_{t,j} - \nabla f_{t,j}\|^p]$, and the theoretical guarantees for the remaining terms follow.

At the beginning of each epoch (i.e., $j = 0$), the SPIDER-type estimators are constructed identically to the standard mini-batch strategy. Thus, it readily follows from Lemma 9 that the initial estimation error is bounded by $\mathbb{E}_{t,0}[\|\tilde{\nabla} f_{t,0} - \nabla f_{t,0}\|^p] \leq 2^{2-p} V_f B_{t,f}^{1-p}$, where we implicitly invoke the non-expansiveness of the Euclidean projection. For each inner iteration $j \geq 1$, $x_{t,j-1}$, $x_{t,j}$, $\tilde{\nabla} f_{t,j-1}$, $\tilde{c}_{t,j-1}$ and $\tilde{\nabla} c_{t,j-1}$ are $\mathcal{F}_{t,j-1}$ -measurable. The samples drawn at iteration (t, j) are independent of $\mathcal{F}_{t,j-1}$, and the samples within each batch are conditionally independent and identically distributed. For $\tilde{\nabla} f_{t,j}$, we define the sample-wise filtration

$$\mathcal{G}_{t,j,0}^f := \mathcal{F}_{t,j}, \quad \mathcal{G}_{t,j,i}^f := \mathcal{F}_{t,j} \vee \sigma\left(\xi_{t,j}^{(1)}, \dots, \xi_{t,j}^{(i)}\right), \quad i = 1, \dots, S_{t,f}.$$

Moreover, let $\tilde{g}_{t,j} = \tilde{\nabla} f_{t,j-1} + \frac{1}{S_{t,f}} \sum_{i=1}^{S_{t,f}} (\nabla \mathbf{F}(x_{t,j}; \xi_{t,j}^{(i)}) - \nabla \mathbf{F}(x_{t,j-1}; \xi_{t,j}^{(i)}))$ for simplicity, and then $\tilde{\nabla} f_{t,j} = \Pi_{\mathbb{B}(M_f)}(\tilde{g}_{t,j})$. By the non-expansiveness of the Euclidean projection, it holds that

$$\|\tilde{\nabla} f_{t,j} - \nabla f_{t,j}\| \leq \|\tilde{g}_{t,j} - \nabla f_{t,j}\|. \quad (4.6)$$

We now proceed to bound the term $\mathbb{E}_{t,j}[\|\tilde{g}_{t,j} - \nabla f_{t,j}\|^p]$. By the Hölder continuity of the gradient of $\|\cdot\|^p$ (i.e., $\|a + b\|^p \leq 2^{2-p} \|a\|^p + p \|b\|^{p-2} \langle a, b \rangle + \|b\|^p$ for any $a, b \in \mathbb{R}^n$ and $b \neq 0$) [38], we obtain

$$\begin{aligned}\mathbb{E}_{t,j}[\|\tilde{g}_{t,j} - \nabla f_{t,j}\|^p] &\leq 2^{2-p} \mathbb{E}_{t,j}[\|\tilde{g}_{t,j} - \mathbb{E}_{t,j}[\tilde{g}_{t,j}]\|^p] + \mathbb{E}_{t,j}[\|\mathbb{E}_{t,j}[\tilde{g}_{t,j}] - \nabla f_{t,j}\|^p] \\ &\quad + p \mathbb{E}_{t,j} \left[\frac{\langle \tilde{g}_{t,j} - \mathbb{E}_{t,j}[\tilde{g}_{t,j}], \mathbb{E}_{t,j}[\tilde{g}_{t,j}] - \nabla f_{t,j} \rangle}{\|\mathbb{E}_{t,j}[\tilde{g}_{t,j}] - \nabla f_{t,j}\|^{2-p}} \right].\end{aligned}$$

Observe that $\nabla f_{t,j}$ and $\mathbb{E}_{t,j}[\tilde{g}_{t,j}]$ are $\mathcal{F}_{t,j}$ -measurable, which ensures that the numerator of the cross-product term satisfies $\mathbb{E}_{t,j}[\langle \tilde{g}_{t,j} - \mathbb{E}_{t,j}[\tilde{g}_{t,j}], \mathbb{E}_{t,j}[\tilde{g}_{t,j}] - \nabla f_{t,j} \rangle] = 0$. When $\mathbb{E}_{t,j}[\tilde{g}_{t,j}] = \nabla f_{t,j}$, the fraction in the preceding display is defined to be zero, and the inequality holds directly since $2^{2-p} > 1$. Hence, we obtain

$$\mathbb{E}_{t,j}[\|\tilde{g}_{t,j} - \nabla f_{t,j}\|^p] \leq 2^{2-p} \mathbb{E}_{t,j}[\|\tilde{g}_{t,j} - \mathbb{E}_{t,j}[\tilde{g}_{t,j}]\|^p] + \mathbb{E}_{t,j}[\|\mathbb{E}_{t,j}[\tilde{g}_{t,j}] - \nabla f_{t,j}\|^p] \quad (4.7)$$

According to the $\mathcal{F}_{t,j}$ -measurability of $\tilde{\nabla} f_{t,j-1}$ and the unbiasedness of the stochastic oracle (Assumption 3), it holds that $\mathbb{E}_{t,j}[\tilde{g}_{t,j}] = \tilde{\nabla} f_{t,j-1} + \nabla f_{t,j} - \nabla f_{t,j-1}$. Substituting this term back into the decomposed bound (4.7) allows us to decouple the estimation error:

$$\begin{aligned} \mathbb{E}_{t,j}[\|\tilde{g}_{t,j} - \nabla f_{t,j}\|^p] &\leq \frac{2^{2-p}}{S_{t,f}^p} \mathbb{E}_{t,j} \left[\left\| \sum_{i=1}^{S_{t,f}} \left(\nabla \mathbf{F}(x_{t,j}; \xi_{t,j}^{(i)}) - \nabla \mathbf{F}(x_{t,j-1}; \xi_{t,j}^{(i)}) - \nabla f_{t,j} + \nabla f_{t,j-1} \right) \right\|^p \right] \\ &\quad + \mathbb{E}_{t,j}[\|\tilde{\nabla} f_{t,j-1} - \nabla f_{t,j-1}\|^p]. \end{aligned}$$

Let $X_{t,j}^{f,i} := \nabla F(x_{t,j}; \xi_{t,j}^{(i)}) - \nabla F(x_{t,j-1}; \xi_{t,j}^{(i)}) - \nabla f_{t,j} + \nabla f_{t,j-1}$ for $i = 1, \dots, S_{t,f}$. Since the current sample is independent of $\mathcal{G}_{t,j,i-1}^f$ and the stochastic oracle is unbiased, $\{X_{t,j}^{f,i}\}_{i=1}^{S_{t,f}}$ is a martingale difference sequence with respect to $\{\mathcal{G}_{t,j,i}^f\}_{i=0}^{S_{t,f}}$. Further applying the inequality $\|a + b\|^p \leq 2^{p-1}(\|a\|^p + \|b\|^p)$ and Assumption 3 leads to

$$\mathbb{E}[\|X_{t,j}^{f,i}\|^p] \leq 2^{p-1} \left(\mathbb{E}[\|\nabla \mathbf{F}(x_{t,j}; \xi_{t,j}^{(i)}) - \nabla f_{t,j}\|^p] + \mathbb{E}[\|\nabla \mathbf{F}(x_{t,j-1}; \xi_{t,j}^{(i)}) - \nabla f_{t,j-1}\|^p] \right) \leq 2^p V_f^p < +\infty.$$

With this moment condition satisfied, by invoking Burkholder-type inequality (Lemma 13) and the inequality $\|a + b\|^p \leq 2^{p-1}(\|a\|^p + \|b\|^p)$ again, the bound on $\mathbb{E}[\|\tilde{g}_{t,j} - \nabla f_{t,j}\|^p]$ becomes

$$\begin{aligned} \mathbb{E}_{t,j}[\|\tilde{g}_{t,j} - \nabla f_{t,j}\|^p] &\leq \frac{2^{3-p}}{S_{t,f}^p} \sum_{i=1}^{S_{t,f}} \left(\mathbb{E}_{t,j} \left[\|\nabla \mathbf{F}(x_{t,j}; \xi_{t,j}^{(i)}) - \nabla \mathbf{F}(x_{t,j-1}; \xi_{t,j}^{(i)})\|^p \right] + \|\nabla f_{t,j} - \nabla f_{t,j-1}\|^p \right) \\ &\quad + \|\tilde{\nabla} f_{t,j-1} - \nabla f_{t,j-1}\|^p \\ &\leq 2^{4-p} L_f^p S_{t,f}^{1-p} \|x_{t,j} - x_{t,j-1}\|^p + \|\tilde{\nabla} f_{t,j-1} - \nabla f_{t,j-1}\|^p \\ &\leq \frac{2^{4-p} \kappa_t^p L_f^p S_{t,f}^{1-p}}{\ell_{\min}^p} + \|\tilde{\nabla} f_{t,j-1} - \nabla f_{t,j-1}\|^p, \end{aligned}$$

where the second inequality follows from Assumption 6, and the last inequality is due to the fact that $\|x_{t,j} - x_{t,j-1}\| \leq \kappa_t / (L_f + \tilde{\rho}_{t,j-1} L_J) \leq \kappa_t / \ell_{\min}$. Further substituting into (4.6) yields

$$\mathbb{E}_{t,j}[\|\tilde{\nabla} f_{t,j} - \nabla f_{t,j}\|^p] \leq \frac{2^{4-p} \kappa_t^p L_f^p S_{t,f}^{1-p}}{\ell_{\min}^p} + \|\tilde{\nabla} f_{t,j-1} - \nabla f_{t,j-1}\|^p. \quad (4.8)$$

Since $\mathcal{F}_{t,0} \subseteq \mathcal{F}_{t,j}$, iterating this recursion together with the tower property gives

$$\mathbb{E}_{t,0}[\|\tilde{\nabla} f_{t,j} - \nabla f_{t,j}\|^p] \leq \frac{2^{4-p} j \kappa_t^p L_f^p S_{t,f}^{1-p}}{\ell_{\min}^p} + \mathbb{E}_{t,0}[\|\tilde{\nabla} f_{t,0} - \nabla f_{t,0}\|^p] \leq \frac{2^{4-p} \tau_t \kappa_t^p L_f^p S_{t,f}^{1-p}}{\ell_{\min}^p} + 2^{2-p} V_f^p B_{t,f}^{1-p},$$

where we apply $j \leq \tau_t$ in the last inequality. By the conditional Jensen's inequality, we obtain

$$\mathbb{E}_{t,0}[\|\tilde{\nabla} f_{t,j} - \nabla f_{t,j}\|] \leq \frac{2^{\frac{4-p}{p}} \tau_t^{\frac{1}{p}} \kappa_t L_f S_{t,f}^{\frac{1-p}{p}}}{\ell_{\min}} + 2^{\frac{2-p}{p}} V_f B_{t,f}^{\frac{1-p}{p}}.$$

The constraint value bounds follow by the same argument. For the Jacobian estimator, the argument is carried out w.r.t. $\|\cdot\|_F$. If $\hat{J}_{t,j}$ denotes the corresponding unprojected estimator, we then have $\|\tilde{\nabla} c_{t,j} - \nabla c_{t,j}\| \leq \|\tilde{\nabla} c_{t,j} - \nabla c_{t,j}\|_F \leq \|\hat{J}_{t,j} - \nabla c_{t,j}\|_F$, where the first inequality uses $\|A\| \leq \|A\|_F$ for any matrix A , while the second follows from the non-expansiveness of $\Pi_{\mathbb{B}(L_c)}^F$ in the Frobenius norm. The Frobenius-norm moment conditions in Assumptions 3 and 6 then yield the desired results via an analogous argument. $\square$

Under the SPIDER-type recursive estimation, we retain the notation of $E_{t,i}$ for $i \in \{f, c, J\}$, but reinterpret them as total estimation-error bounds over all inner iterations of epoch t . Specifically, we choose these quantities

$$E_{t,i} = \frac{2^{\frac{4-p}{p}} \tau_t^{\frac{p+1}{p}} \kappa_t L_i S_{t,i}^{\frac{1-p}{p}}}{\ell_{\min}} + 2^{\frac{2-p}{p}} \tau_t V_i B_{t,i}^{\frac{1-p}{p}}, \quad i \in \{f, c, J\}. \quad (4.9)$$

Then summing the conditional bounds in Lemma 10 over $j = 0, \dots, \tau_t - 1$ yields

$$\sum_{j=0}^{\tau_t-1} \mathbb{E}_{t,0}[\|\tilde{\nabla} f_{t,j} - \nabla f_{t,j}\|] \leq E_{t,f}, \quad \sum_{j=0}^{\tau_t-1} \mathbb{E}_{t,0}[\|\tilde{c}_{t,j} - c_{t,j}\|] \leq E_{t,c}, \quad \sum_{j=0}^{\tau_t-1} \mathbb{E}_{t,0}[\|\tilde{\nabla} c_{t,j} - \nabla c_{t,j}\|] \leq E_{t,J}.$$

We next analyze Algorithm 1 with SPIDER-type recursive estimation. The additional task is to control the constraint-restoration events throughout each epoch. Following the argument leading to the inequality (4.8) in the proof of Lemma 10, we obtain that for every $j \geq 1$,

$$\begin{aligned} \mathbb{E}_{t,j}[\|\tilde{c}_{t,j} - c_{t,j}\|^p] &\leq \frac{2^{4-p} \kappa_t^p L_c^p S_{t,c}^{1-p}}{\ell_{\min}^p} + \|\tilde{c}_{t,j-1} - c_{t,j-1}\|^p, \\ \mathbb{E}_{t,j}[\|\tilde{\nabla} c_{t,j} - \nabla c_{t,j}\|_F^p] &\leq \frac{2^{4-p} \kappa_t^p L_J^p S_{t,J}^{1-p}}{\ell_{\min}^p} + \|\tilde{\nabla} c_{t,j-1} - \nabla c_{t,j-1}\|_F^p, \quad \text{almost surely.} \end{aligned} \quad (4.10)$$

Extending (3.1) to the inner iterations (t, j) , we define

$$\mathcal{E}_{t,j} := (\alpha_v + L_c) \|\tilde{c}_{t,j} - c_{t,j}\| + M_c \|\tilde{\nabla} c_{t,j} - \nabla c_{t,j}\|_F,$$

and the corresponding contraction event, the epoch-wise and global contraction events

$$\mathcal{B}_{t,j} := \{\|\tilde{c}_{t,j}\| \leq \epsilon'\} \cup \{\mathcal{E}_{t,j} \leq \eta_c \alpha_v \epsilon'\}, \quad \mathcal{B}_t^{\text{sp}} := \bigcap_{j=0}^{\tau_t-1} \mathcal{B}_{t,j}, \quad \bar{\mathcal{B}}^{\text{sp}} := \bigcap_{t=1}^T \mathcal{B}_t^{\text{sp}},$$

respectively. For Algorithm 1, the completion event in this subsection is

$$\bar{\mathcal{S}} := \bigcap_{t=1}^T \bigcap_{j=0}^{\tau_t-1} (\{\|\tilde{c}_{t,j}\| \leq \epsilon'\} \cup \{\|\tilde{v}_{t,j}^c\| > \beta_{\text{IS}} \epsilon'\}).$$

Thus, $\bar{\mathcal{S}}$ denotes completion of all prescribed inner iterations. To control sampling-accuracy failures throughout each epoch, define

$$a_t := \frac{8\kappa_t^p}{\ell_{\min}^p} \left((\alpha_v + L_c)^p L_c^p S_{t,c}^{1-p} + M_c^p L_J^p S_{t,J}^{1-p} \right), \quad b_t := 2 \left((\alpha_v + L_c)^p V_c^p B_{t,c}^{1-p} + M_c^p V_J^p B_{t,J}^{1-p} \right).$$

Here, b_t bounds the initial weighted error moment, while a_t bounds its one-step conditional increase. We set

$$\mathcal{E}'_{t,\text{sp}} := (b_t + \tau_t a_t)^{1/p}. \quad (4.11)$$

The following lemma provides the epoch-wise non-progressive probability bounds for Algorithm 1 with SPIDER-type recursive estimation.

Lemma 11 *Suppose the assumptions of Lemma 10 hold. For each epoch t , let $\delta_t \in (0, 1)$ be a prescribed deterministic probability budget. Suppose that $\mathcal{E}'_{t,\text{sp}} \leq \eta_c \alpha_v \epsilon' \delta_t^{1/p}$ almost surely. Then, it holds that $\mathbb{P}_{t,0}((\mathcal{B}_t^{\text{sp}})^c) \leq \delta_t$ almost surely, where $\mathbb{P}_{t,0}(\cdot) := \mathbb{P}(\cdot \mid \mathcal{F}_{t,0})$. Consequently, it holds that $\mathbb{P}((\bar{\mathcal{B}}^{\text{sp}})^c) \leq \sum_{t=1}^T \delta_t$.*

Proof. For a fixed t , let $\mathcal{H}_{t,j}$ denote the history immediately after the batches at inner iteration (t, j) have been sampled. Thus, it holds that $\mathcal{F}_{t,0} \subseteq \mathcal{H}_{t,0} \subseteq \mathcal{F}_{t,1} \subseteq \mathcal{H}_{t,1} \subseteq \dots$. Since $(a + b)^p \leq 2^{p-1}(a^p + b^p)$, we have

$$\mathcal{E}_{t,j}^p \leq 2^{p-1} \left((\alpha_v + L_c)^p \|\tilde{c}_{t,j} - c_{t,j}\|^p + M_c^p \|\tilde{\nabla} c_{t,j} - \nabla c_{t,j}\|_F^p \right) =: U_{t,j}.$$

Moreover, the stochastic estimate bounds at $(t, 0)$ and (4.10) imply

$$\mathbb{E}_{t,0}[U_{t,0}] \leq b_t, \quad \mathbb{E}[U_{t,j} \mid \mathcal{H}_{t,j-1}] \leq U_{t,j-1} + a_t, \quad j \geq 1,$$

where the second inequality also uses the tower property and $\mathcal{H}_{t,j-1} \subseteq \mathcal{F}_{t,j}$. For $j = 0, \dots, \tau_t - 1$, define $Z_{t,j} := U_{t,j} + (\tau_t - 1 - j)a_t$. Then, we have $Z_{t,j} \geq 0$ and $\mathbb{E}[Z_{t,j} \mid \mathcal{H}_{t,j-1}] \leq U_{t,j-1} + (\tau_t - j)a_t = Z_{t,j-1}$

for $j = 1, \dots, \tau_t - 1$. Thus, $\{Z_{t,j}\}$ forms an integrable non-negative super-martingale. Since $\mathcal{E}_{t,j}^p \leq U_{t,j} \leq Z_{t,j}$, applying Ville's maximal inequality [39, Lemma 1] conditionally on $\mathcal{H}_{t,0}$ and then taking conditional expectations given $\mathcal{F}_{t,0}$ via the tower property yields

$$\mathbb{P}_{t,0}((\mathcal{B}_t^{\text{sp}})^c) \leq \mathbb{P}_{t,0} \left(\max_{0 \leq j < \tau_t} Z_{t,j} \geq (\eta_c \alpha_v \epsilon')^p \right) \leq \frac{\mathbb{E}_{t,0}[Z_{t,0}]}{(\eta_c \alpha_v \epsilon')^p} \leq \frac{b_t + (\tau_t - 1)a_t}{(\eta_c \alpha_v \epsilon')^p}.$$

Noting that the notation in (4.9) and the definition of $\mathcal{E}'_{t,\text{sp}}$ in (4.11) imply $(\mathcal{E}'_{t,\text{sp}})^p = \tau_t a_t + b_t$. Combining $a_t \geq 0$, it follows that $b_t + (\tau_t - 1)a_t \leq (\mathcal{E}'_{t,\text{sp}})^p$. Further invoking the condition $\mathcal{E}'_{t,\text{sp}} \leq \eta_c \alpha_v \epsilon' \delta_t^{1/p}$ almost surely gives $\mathbb{P}_{t,0}((\mathcal{B}_t^{\text{sp}})^c) \leq \delta_t$. Finally, the union bound and the tower property yield

$$\mathbb{P}((\bar{\mathcal{B}}^{\text{sp}})^c) \leq \sum_{t=1}^T \mathbb{P}((\mathcal{B}_t^{\text{sp}})^c) = \sum_{t=1}^T \mathbb{E}[\mathbb{P}_{t,0}((\mathcal{B}_t^{\text{sp}})^c)] \leq \sum_{t=1}^T \delta_t,$$

which completes the proof. $\square$

The following theorem establishes the sample complexity bounds for Algorithm 1 equipped with SPIDER-type recursive estimation.

Theorem 3 *Suppose Assumptions 1, 3, 4, 5 and 6 hold. Given a tolerance $\epsilon \in (0, 1)$, set $\epsilon' = c_\epsilon \epsilon$, where $c_\epsilon > 0$ is a sufficiently small fixed constant independent of ϵ , and choose*

$$\begin{aligned} \kappa_t &\asymp \epsilon, & B_{t,f} &\asymp \left\lceil \epsilon^{\frac{p}{1-p}} \right\rceil, & B_{t,c} &\asymp \left\lceil \epsilon^{\frac{p+2}{1-p}} \right\rceil, & B_{t,J} &\asymp \left\lceil \epsilon^{\frac{p+2}{1-p}} \right\rceil, \\ \tau_t &\asymp \lceil \epsilon^{-1} \rceil, & S_{t,f} &\asymp \left\lceil \epsilon^{\frac{1}{1-p}} \right\rceil, & S_{t,c} &\asymp \left\lceil \epsilon^{\frac{3}{1-p}} \right\rceil, & S_{t,J} &\asymp \left\lceil \epsilon^{\frac{3}{1-p}} \right\rceil, & \delta_t &\asymp \frac{\epsilon}{T} \text{ with } \sum_{t=1}^T \delta_t \leq \epsilon, \end{aligned}$$

Then Algorithm 1 with SPIDER-type recursive estimation completes all prescribed iterations with probability at least $1 - \epsilon$, and conditioned on this, it returns an expected ϵ -KKT point after $T \asymp \lceil \epsilon^{-1} \rceil$ iterations, corresponding to $\mathcal{O}(\epsilon^{-2})$ inner iterations. The sample complexity for computing $\tilde{\nabla} f_t$, $\tilde{\nabla} c_t$ and $\tilde{c}_t$ are of order $\mathcal{O}(\epsilon^{-\frac{2p-1}{p-1}})$, $\mathcal{O}(\epsilon^{-\frac{2p+1}{p-1}})$ and $\mathcal{O}(\epsilon^{-\frac{2p+1}{p-1}})$, respectively.

Proof. The definitions of $\Delta_{T,1}$ in (3.31), $\Delta_{T,2}$ in (3.32), and $\mathcal{E}'_{t,\text{sp}}$ in (4.11) give

$$\begin{aligned} \Delta_{T,1} &= \mathcal{O} \left(\frac{1 + \sum_{t=1}^T \tau_t \kappa_t^2 + \sum_{t=1}^T E_{t,c}}{K_T} + \frac{\sum_{t=1}^T \kappa_t (E_{t,f} + E_{t,c} + E_{t,J})}{K_T} + \sum_{t=1}^T \delta_t + \epsilon' \right), \\ \Delta_{T,2} &= \mathcal{O} \left(\frac{1 + \sum_{t=1}^T \tau_t \kappa_t^2 + \sum_{t=1}^T E_{t,c}}{K_T} + \frac{\sum_{t=1}^T \kappa_t (E_{t,f} + E_{t,c} + E_{t,J})}{K_T} + \sum_{t=1}^T \delta_t \right), \\ \frac{\mathcal{E}'_{t,\text{sp}}}{\delta_t^{1/p}} &= \left(\frac{b_t + \tau_t a_t}{\delta_t} \right)^{1/p} = \mathcal{O} \left(\left(\frac{B_{t,c}^{1-p} + B_{t,J}^{1-p} + \tau_t \kappa_t^p (S_{t,c}^{1-p} + S_{t,J}^{1-p})}{\delta_t} \right)^{1/p} \right). \end{aligned}$$

It follows from the parameter choices that

$$\begin{aligned} E_{t,f} &= \mathcal{O}(1), & E_{t,c} &= \mathcal{O}(\epsilon^{\frac{2}{p}}), & E_{t,J} &= \mathcal{O}(\epsilon^{\frac{2}{p}}), & B_{t,c}^{1-p} &= \mathcal{O}(\epsilon^{p+2}), \\ B_{t,J}^{1-p} &= \mathcal{O}(\epsilon^{p+2}), & S_{t,c}^{1-p} &= \mathcal{O}(\epsilon^3), & S_{t,J}^{1-p} &= \mathcal{O}(\epsilon^3), & K_T &\asymp T. \end{aligned}$$

Further combining the setting of ϵ' , κ_t , τ_t , and δ_t gives

$$\Delta_{T,1} = \mathcal{O} \left(\frac{1}{T} + \epsilon^{\frac{p+2}{p}} + \epsilon^{\frac{2}{p}} + \epsilon \right), \quad \epsilon' + \Delta_{T,2} = \mathcal{O} \left(\frac{1}{T} + \epsilon^{\frac{p+2}{p}} + \epsilon^{\frac{2}{p}} + \epsilon \right), \quad \frac{\mathcal{E}'_{t,\text{sp}}}{\delta_t^{1/p}} = \mathcal{O} \left(\epsilon^{\frac{p+1}{p}} T^{\frac{1}{p}} \right).$$

Further combining $T \asymp \lceil \epsilon^{-1} \rceil$ gives $\Delta_{T,1} = \mathcal{O}(\epsilon)$, $\epsilon' + \Delta_{T,2} = \mathcal{O}(\epsilon)$ and $\mathcal{E}'_{t,\text{sp}}/\delta_t^{1/p} = \mathcal{O}(\epsilon)$. Choosing the proportionality constants sufficiently small or large as specified above ensures $\mathcal{E}'_{t,\text{sp}} \leq \eta_c \alpha_v \epsilon' \delta_t^{1/p}$ and the

accuracy requirements of Theorem 1, with errors aggregated over epochs. Lemma 11 gives $\mathbb{P}((\bar{\mathcal{B}}^{\text{sp}})^c) \leq \sum_{t=1}^T \delta_t$. The same early termination test argument as in Lemma 3 shows that $\bar{\mathcal{B}}^{\text{sp}} \subseteq \bar{\mathcal{S}}$. Hence, it holds that $\mathbb{P}(\bar{\mathcal{S}}) \geq 1 - \sum_{t=1}^T \delta_t \geq 1 - \epsilon$. Applying the argument of Theorem 1 with epoch-wise error bounds, the randomized output conditioned on $\bar{\mathcal{S}}$ satisfies the ϵ -KKT conditions in expectation. Moreover, the total number of samples required for finding an expected ϵ -KKT point are

$$\begin{aligned} \sum_{t=1}^T (B_{t,f} + 2(\tau_t - 1)S_{t,f}) &= \mathcal{O}\left(\left(\epsilon^{\frac{p}{1-p}} + \epsilon^{\frac{p}{1-p}}\right)T\right) = \mathcal{O}\left(\epsilon^{\frac{2p-1}{1-p}}\right), \\ \sum_{t=1}^T (B_{t,J} + 2(\tau_t - 1)S_{t,J}) &= \mathcal{O}\left(\left(\epsilon^{\frac{p+2}{1-p}} + \epsilon^{\frac{p+2}{1-p}}\right)T\right) = \mathcal{O}\left(\epsilon^{\frac{2p+1}{1-p}}\right), \\ \sum_{t=1}^T (B_{t,c} + 2(\tau_t - 1)S_{t,c}) &= \mathcal{O}\left(\left(\epsilon^{\frac{p+2}{1-p}} + \epsilon^{\frac{p+2}{1-p}}\right)T\right) = \mathcal{O}\left(\epsilon^{\frac{2p+1}{1-p}}\right), \end{aligned}$$

which completes the proof. $\square$

Discussion 2 For Algorithm 1, the recursive variance reduction and epoch-wise error control improve the sample complexity bounds in Theorem 2. As $p \uparrow 2$, the sample-complexity bounds have the limiting forms $\mathcal{O}(\epsilon^{-3})$ for objective-gradient evaluations and $\mathcal{O}(\epsilon^{-5})$ for both constraint-value and Jacobian evaluations. The objective-gradient and constraint-value exponents match those in [15], up to logarithmic factors, although our Jacobian bound is higher. Their stochastic framework requires a uniformly nondegenerate sampled equality Jacobian and a quantitative MFCQ for the sampled constraints [15, Assumption 6]. Their complexity analysis establishes these properties with high probability under almost-surely bounded oracle errors, with sample-wise smoothness additionally required for recursive estimation. In our analysis, the geometric regularity conditions are imposed on the true feasibility problem and the deterministic total-step subproblem. Specifically, the projected feasibility error bound supports normal-step contraction, while near-active LICQ controls the multipliers and total-step sensitivity. We establish the required sampled contraction and perturbation bounds on estimation-accuracy events and explicitly control the contributions of their complements to the expected residual bounds. This analysis accommodates bounded p th central moments and mean p th-moment smoothness condition, without requiring almost-surely bounded original oracle errors or uniform regularity of the sampled constraints on every realization.

4.3 Application to the semi-stochastic Case

In this subsection, we specialize the methods to the semi-stochastic setting, with a stochastic objective and deterministic constraints, that is

$$\min_{x \in \mathbb{R}^n} f(x) := \mathbb{E}_{\xi \sim \Xi}[\mathbf{F}(x; \xi)] \quad \text{s.t.} \quad c(x) = \mathbf{0}, \quad x \geq \mathbf{0}. \quad (4.12)$$

In this setting, the constraint values and Jacobians are evaluated exactly at query points. Only the objective gradient requires stochastic estimation. Thus, Algorithm 1 uses

$$\tilde{c}_t = c_t, \quad \tilde{\nabla} c_t = \nabla c_t, \quad (4.13)$$

and the same convention applies at every SPIDER inner iteration. Consequently, the constraint estimation bounds in Condition 2 can be taken as $E_{t,c} = E_{t,J} = 0$. Under the respective CQ assumptions, the constraint-related good events hold almost surely, so their complementary-event probability contributions vanish in the convergence analysis of Algorithm 1.

We retain the projected objective gradient estimators from Section 4. The objective gradient part of Condition 1 is therefore satisfied, while its constraint bounds follow from exact evaluations and the deterministic assumptions. Only the objective gradient part of Assumption 3 is needed for the estimation-error analysis. For SPIDER-type recursive estimation, we additionally require the objective gradient part of Assumption 6. The remaining deterministic smoothness and geometric assumptions are unchanged.

We apply Algorithm 1 to (4.12), using exact constraint information and retaining the stopping coefficient β_{IS} (3.7) chosen in Section 3. The projected residuals satisfy $\tilde{v}_t^c = v_t^c$ because $\tilde{c}_t = c_t$ and $\tilde{\nabla} c_t = \nabla c_t$.

Thus, under Assumption 4, the early termination test cannot be triggered, and all prescribed iterations are completed almost surely. Consequently, it holds that $\mathcal{E}_t = \mathcal{E}'_t = 0$ and $\mathbb{P}(\bar{\mathcal{B}}) = \mathbb{P}(\bar{\mathcal{S}}) = 1$, and the following expected ϵ -KKT guarantees are unconditional. Moreover, the contraction and merit-parameter bounds in Lemma 4 and Lemma 5 hold almost surely without a probability budget.

The identity $\tilde{v}_t^c = v_t^c$ does not imply $\tilde{v}_t = v_t$. Indeed, the normal step v_t retains its definition as the minimizer of the deterministic normal-step subproblem, whereas Algorithm 1 sets $\tilde{v}_t = 0$ when $\|c_t\| \leq \epsilon'$. Using the bound $\|v_t\| \leq \frac{\|c_t\|}{\sqrt{\gamma_{\min}}}$ from Lemma 7, we obtain $\mathbb{E}[\|\tilde{v}_t - v_t\|] \leq \frac{\epsilon'}{\sqrt{\gamma_{\min}}}$. The uniform step bounds in Lemmas 7 and 8 remain valid. Further combining the Lemmas 6 and 8 with $E_{t,c} = E_{t,J} = 0$, the quantities defined in (3.31) and (3.32) in this setting reduce to

$$\begin{aligned} \bar{\Delta}_{T,1} &:= \frac{2\ell_{\max}(f_1 - f^* + \rho_{\max}M_c)}{\sigma_l K_T} + \frac{(4\ell_{\max} + \sigma_l) \sum_{t=1}^T \kappa_t^2}{4\sigma_l \ell_{\min} K_T} + \frac{(L_d + 2M_d/\delta_R)\epsilon'}{\sqrt{\gamma_{\min}}} \\ &\quad + \frac{(2\ell_{\max}/(\sigma_l \ell_{\min}) + L_d + 2M_d/\delta_R) \sum_{t=1}^T \kappa_t E_{t,f}}{K_T}, \end{aligned} \quad (4.14)$$

$$\bar{\Delta}_{T,2} := \frac{2\ell_{\max}L_c(f_1 - f^* + \rho_{\max}M_c)}{(1 - \vartheta)\sigma_l K_T} + \frac{(4\ell_{\max} + \sigma_l)L_c \sum_{t=1}^T \kappa_t^2}{4(1 - \vartheta)\sigma_l \ell_{\min} K_T} + \frac{2\ell_{\max}L_c \sum_{t=1}^T \kappa_t E_{t,f}}{(1 - \vartheta)\ell_{\min}\sigma_l K_T}, \quad (4.15)$$

where $\vartheta \in (0, 1)$ is the contraction factor in Lemma 4. The following corollary specializes the convergence guarantees to the semi-stochastic setting.

Corollary 1 *Suppose Assumptions 4, 5 and Conditions 1, 2 hold. Fix $\epsilon' > 0$ and let x_R be the randomized output generated by Algorithm 1 for (4.12) under (4.13), where R is drawn independently with $\mathbb{P}(R = t) = \kappa_t/K_T$ for $t \in [T]$. Then, it holds that $\mathbb{E}[\|d_R\|] \leq \bar{\Delta}_{T,1}$ and $\mathbb{E}[\|c_R\|] \leq \epsilon' + \bar{\Delta}_{T,2}$. Moreover, given a tolerance $\epsilon \in (0, 1)$, if $\bar{\Delta}_{T,1} \leq \min\{\frac{\epsilon}{\sigma_u}, \frac{\nu\epsilon}{\sigma_u M_d + M_f}\}$ and $\epsilon' + \bar{\Delta}_{T,2} \leq \epsilon$, there exist $\lambda_R \in \mathbb{R}^m$ and $\mu_R \in \mathbb{R}_{\geq 0}^n$ such that (3.36) holds without conditioning.*

Proof. With exact constraint information, $\bar{\mathcal{B}}$ and $\bar{\mathcal{S}}$ hold almost surely. Applying the proof of Theorem 1 with $E_{t,c} = E_{t,J} = 0$ and zero complementary-event contributions yields the stated unconditional bounds. $\square$

The following corollary provides the corresponding sample complexity for Algorithm 1 applied to (4.12).

Corollary 2 *Suppose Assumptions 1, 2, 4 and 5 hold, together with the objective gradient part of Assumption 3. Let $\epsilon \in (0, 1)$ and set $\epsilon' = c_\epsilon \epsilon$, where $c_\epsilon > 0$ is a sufficiently small fixed constant independent of ϵ , and apply Algorithm 1 to (4.12) using (4.13).*

- (1) *Mini-batch estimation. Choose $\kappa_t \asymp \epsilon$, $B_{t,f} \asymp \lceil \epsilon^{\frac{p}{1-p}} \rceil$. Then, the algorithm outputs an expected ϵ -KKT point after $T \asymp \lceil \epsilon^{-2} \rceil$ iterations, with the sample complexity for computing $\tilde{\nabla} f_t$ of order $\mathcal{O}(\epsilon^{\frac{3p-2}{1-p}})$.*
- (2) *SPIDER-type recursive estimation. Additionally, suppose that the objective gradient part of Assumption 6 holds. Choose $\kappa_t \asymp \epsilon$, $B_{t,f} \asymp \lceil \epsilon^{\frac{p}{1-p}} \rceil$, $S_{t,f} \asymp \lceil \epsilon^{\frac{1}{1-p}} \rceil$, $\tau_t \asymp \lceil \epsilon^{-1} \rceil$. Then, the algorithm outputs an expected ϵ -KKT point after $T \asymp \lceil \epsilon^{-1} \rceil$ iterations, corresponding to $\mathcal{O}(\epsilon^{-2})$ inner iterations. The sample complexity for computing $\tilde{\nabla} f_t$ is of order $\mathcal{O}(\epsilon^{\frac{2p-1}{1-p}})$.*

Proof. The result follows from the arguments in Theorems 2 and 3 specialized to the case $E_{t,c} = E_{t,J} = 0$, and hence the proof is omitted. $\square$

Discussion 3 *As $p \uparrow 2$, the objective-gradient sample-complexity bounds in Corollary 2 have the limiting forms $\mathcal{O}(\epsilon^{-4})$ and $\mathcal{O}(\epsilon^{-3})$ under mini-batch and SPIDER-type recursive estimation, respectively. These recover the corresponding ϵ -dependence of the momentum-based linearized ALM in [9], whose $\mathcal{O}(\epsilon^{-3})$ result additionally requires an $\mathcal{O}(\sqrt{\epsilon})$ -feasible initialization. Our SPIDER-type result attains this order under the stated geometric and mean p th-moment smoothness conditions, without prescribing an accuracy-dependent initial feasibility tolerance. Curtis et al. [11] and Na et al. [8] established convergence in expectation and almost-sure convergence, respectively, for stochastic SQP methods with deterministic constraints. Corollary 2 further quantifies the sample complexity under bounded p th central moments for $p \in (1, 2)$, including potentially infinite-variance gradient errors, while controlling stationarity, feasibility, and complementarity.*

4.4 Specialization to equality-only problems

We now consider problem (1.1) without non-negativity bounds, i.e.,

$$\min_{x \in \mathbb{R}^n} f(x) := \mathbb{E}_\xi[\mathbf{F}(x; \xi)], \quad \text{s.t.} \quad c(x) := \mathbb{E}_\zeta[\mathbf{C}(x; \zeta)] = \mathbf{0}. \quad (4.16)$$

This problem has received considerable attention (see, e.g., [15, 31, 32, 33, 40, 41]), with existing analyses relying on bounded-variance oracle errors or stronger almost-sure boundedness assumptions. For notational simplicity, we reuse notation where no ambiguity arises, with any redefinitions applying only to the equality-constrained setting.

We retain the normal-total-step structure of the mixed-constrained case in Section 2, but the absence of bound constraints permits an explicit normal-tangential decomposition. For the normal-step subproblem (2.1) without non-negative constraint, we use the negative gradient of the normal-step objective at $v = 0$ rather than its regularized least-squares minimizer. This choice gives

$$\tilde{v}_t = - \nabla_v \left(\frac{1}{2} \|\tilde{c}_t + \tilde{\nabla} c_t^\top v\|^2 + \frac{\gamma_t}{2} \|v\|^2 \right) \Big|_{v=0} = -\tilde{\nabla} c_t \tilde{c}_t.$$

This first-order correction lies in $\text{Range}(\tilde{\nabla} c_t)$, allowing an orthogonal normal-tangential decomposition. Define $\tilde{U}_t := \text{Null}(\tilde{\nabla} c_t^\top)$ and $P_{\tilde{U}_t} := I - \tilde{\nabla} c_t (\tilde{\nabla} c_t)^\dagger$, where $\dagger$ denotes the Moore-Penrose pseudoinverse. With $H_t = I$ and no bound constraints, the total-step subproblem (2.3) reduces to

$$\min_{d \in \mathbb{R}^n} \tilde{\nabla} f_t^\top d + \frac{1}{2} \|d\|^2 \quad \text{s.t.} \quad \tilde{\nabla} c_t^\top d = \tilde{\nabla} c_t^\top \tilde{v}_t.$$

The strong convexity gives a unique primal solution even for a rank-deficient sampled Jacobian. We omit the early termination test and control the contribution of ill-conditioned sampling events directly in expectation. Since $\tilde{v}_t = -\tilde{\nabla} c_t \tilde{c}_t$ belongs to $\text{Range}(\tilde{\nabla} c_t)$, the unique solution is

$$\tilde{d}_t = -P_{\tilde{U}_t} \tilde{\nabla} f_t + \tilde{v}_t = -P_{\tilde{U}_t} \tilde{\nabla} f_t - \tilde{\nabla} c_t \tilde{c}_t. \quad (4.17)$$

Without bound constraints, the near-active bound set in Assumption 5 is empty (i.e., $\mathcal{I}_t^a = \emptyset$). Assumption 5 therefore reduces to strong LICQ for the true constraint Jacobian. Specifically, there exists a constant $\nu > 0$ such that $\nabla c(x)^\top \nabla c(x) \geq \nu^2 I_m$ at every generated iterate. Moreover, the true projected gradient residual becomes $v_t^c = \Pi_{\mathbb{R}^n}(x_t - \nabla c_t c_t) - x_t = -\nabla c_t c_t$. Thus, Assumption 4 reduces to the error bound $\|\nabla c_t c_t\| \geq \alpha_v \|c_t\|$. This bound follows automatically with $\alpha_v = \nu$, since $\|\nabla c_t c_t\|^2 = c_t^\top \nabla c_t^\top \nabla c_t c_t \geq \nu^2 \|c_t\|^2$. Hence, no separate feasibility error-bound assumption is needed. Unlike the event $\mathcal{B}_t$ used in the mixed-constrained analysis, we consider the event $\mathcal{A}_t := \{\sigma_{\min}(\tilde{\nabla} c_t) \geq \nu/2\}$ and its complement.

Using the same merit function $\Phi_\rho(x) = f(x) + \rho \|c(x)\|$, we next choose the merit parameter to ensure sufficient decrease in its sampled linearized model. On event $\mathcal{A}_t$, the stepsize restriction $\eta_t \leq L_c^{-2}$ gives $\|\tilde{c}_t + \eta_t \tilde{\nabla} c_t^\top \tilde{d}_t\| \leq (1 - \eta_t (\sigma_{\min}(\tilde{\nabla} c_t))^2) \|\tilde{c}_t\|$. Hence, to obtain a predicted decrease of at least $\frac{\eta_t}{2} \|\tilde{d}_t\|^2$ on $\mathcal{A}_t$, it suffices to require $\tilde{\nabla} f_t^\top \tilde{d}_t - \tilde{\rho}_t (\sigma_{\min}(\tilde{\nabla} c_t))^2 \|\tilde{c}_t\| \leq -\frac{1}{2} \|\tilde{d}_t\|^2$. This motivates the merit-parameter update

$$\tilde{\rho}_t = \begin{cases} \max \left\{ \tilde{\rho}_{t-1}, \frac{\tilde{\nabla} f_t^\top \tilde{d}_t + \frac{1}{2} \|\tilde{d}_t\|^2}{(\sigma_{\min}(\tilde{\nabla} c_t))^2 \|\tilde{c}_t\|} \right\}, & \text{on } \mathcal{A}_t \cap \{\tilde{c}_t \neq \mathbf{0}\}, \\ \tilde{\rho}_{t-1}, & \text{otherwise,} \end{cases} \quad (4.18)$$

with $\tilde{\rho}_0 > 0$. We then incorporate the normalization strategy into the iterate update:

$$x_{t+1} = x_t + \eta_t \tilde{d}_t \quad \text{with} \quad \eta_t := \min \left\{ \frac{\kappa_t}{(L_f + \tilde{\rho}_t L_J) \|\tilde{d}_t\|}, \frac{1}{L_c^2} \right\}, \quad (4.19)$$

with $1/\|\tilde{d}_t\| = +\infty$ when $\tilde{d}_t = \mathbf{0}$. The complete algorithm is presented in Algorithm 2.

Remark 4 Detailed statements, parameter choices, and proofs are provided in Appendix B. The main results for Algorithm 2 are summarized below.

Algorithm 2 NS-SQP for equality-constrained problem (4.16)

Input: Initial point $x_1 \in \mathbb{R}^n$, sequence $\{\kappa_t\} \subset \mathbb{R}_{>0}$ and iteration horizon T .

Output: x_R , with R drawn independently from $[T]$ and $\mathbb{P}(R = t) = \kappa_t/K_T$, where $K_T = \sum_{t=1}^T \kappa_t$.

```

1: for  $t = 1, \dots, T$  do
2:   Construct the stochastic estimators  $\tilde{\nabla} f_t$ ,  $\tilde{c}_t$  and  $\tilde{\nabla} c_t$ .           // Stochastic Estimation
3:   Compute the search direction  $\tilde{d}_t$  through (4.17).                     // Total step Computation
4:   Compute the merit parameter  $\tilde{\rho}_t$  through (4.18).                     // Merit Parameter Update
5:   Update the variable  $x_{t+1}$  through (4.19).                           // Normalized-step Update
6: end for
```

(1) **Control of inaccurate sampled Jacobians.** Under Assumption 5 with $\mathcal{I}_t^a = \emptyset$ and Condition 2, it holds that $\mathbb{P}(\mathcal{A}_t^c) \leq 2E_{t,J}/\nu$ for any $t \in [T]$. Algorithm 2 continues on $\mathcal{A}_t^c$, and the contribution of this event is controlled in expectation rather than through an early termination test.

(2) **Unconditional convergence guarantees.** Under Assumption 5 with $\mathcal{I}_t^a = \emptyset$, Conditions 1 and 2, for the randomized output x_R generated by Algorithm 2, there exists $\lambda_R \in \mathbb{R}^m$ such that

$$\max \{ \mathbb{E}[\|\nabla f_R + \nabla c_R \lambda_R\|], \nu \mathbb{E}[\|c_R\|] \} \leq \frac{C}{K_T} \left(1 + \sum_{t=1}^T [\kappa_t^2 + E_{t,c} + \kappa_t(E_{t,f} + E_{t,c} + E_{t,J})] \right).$$

Here, $C > 0$ depends only on fixed problem and algorithmic parameters. An expected ϵ -KKT point is obtained when the right-hand side is at most $\min\{\epsilon, \nu\epsilon\}$. Unlike the mixed-constrained result, this guarantee does not require conditioning on the event that all iterations are completed.

(3) **Sample complexity bounds.** Under Assumptions 1, 2, 3, and 5 with $\mathcal{I}_t^a = \emptyset$, suitable parameter choices yield an expected ϵ -KKT point with sample complexity of $\mathcal{O}(\epsilon^{\frac{3p-2}{1-p}})$ for both the objective gradient and constraint Jacobian oracles, and $\mathcal{O}(\epsilon^{\frac{4p-2}{1-p}})$ for constraint values, under mini-batch estimation. Under additional Assumption 6, SPIDER-type recursive estimation improves these bounds to $\mathcal{O}(\epsilon^{\frac{2p-1}{1-p}})$ and $\mathcal{O}(\epsilon^{\frac{3p-1}{1-p}})$, respectively. The constraint-value and Jacobian sample-complexity bounds are lower than their mixed-constrained counterparts. As $p \uparrow 2$, the bounds for objective gradient, constraint Jacobian, and constraint value evaluations have the limiting forms $(\mathcal{O}(\epsilon^{-4}), \mathcal{O}(\epsilon^{-4}), \mathcal{O}(\epsilon^{-6}))$ with mini-batch estimation and $(\mathcal{O}(\epsilon^{-3}), \mathcal{O}(\epsilon^{-3}), \mathcal{O}(\epsilon^{-5}))$ with SPIDER-type estimation. These agree with the corresponding orders in [15] and [15, 31], respectively, up to logarithmic factors. The recursive analysis uses mean p th-moment smoothness instead of the sample-wise smoothness in [15].

5 Numerical Experiments

In this section, we evaluate Algorithm 1 on a flood insurance risk-control problem and an importance-weighted constrained classification problem, and test Algorithm 2 on CUTest experiments with injected heavy-tailed noise. The variants using mini-batch and SPIDER-type recursive estimation are denoted by NS-SQP-MB and NS-SQP-SP, respectively. In the experiments, we replace $\kappa_t/(L_f + \tilde{\rho}_t L_J)$ by a tuned effective stepsize $h > 0$, while retaining direction normalization and the corresponding stepsize caps. We evaluate residual trajectories and terminal iterates under fixed sampling budgets.

5.1 Flood insurance risk-control problem

We consider an expected-excess risk-control problem using flood insurance claim records from the Open-FEMA FIMA NFIP Redacted Claims v2 dataset.¹ The cleaned dataset contains 1,357,948 observations from 2000-2024, with building-and-contents payments normalized by 10^5 US dollars and grouped into $G = 101$ geographical categories. These records define the uniform empirical distribution $\hat{P}_N$. Let r_j be the retention

¹Federal Emergency Management Agency, *FIMA NFIP Redacted Claims*, version 2. <https://www.fema.gov/openfema-data-page/fima-nfip-redacted-claims-v2>.

proportion for category j , K the risk capital, and $x = (r_1, \dots, r_G, K)$. Introducing non-negative slack variables $s \in \mathbb{R}_+^{G+1}$ and $t \in \mathbb{R}_+$, we formulate the problem with expectation equality constraints and non-negative variables:

$$\min_{z=(x,s,t) \geq 0} f(z) := \sum_{j=1}^G p_j(1-r_j) + 0.2K, \quad \text{s.t.} \quad \mathbb{E}_{\xi=(L,C) \sim \hat{P}_N} \left[\left(\phi_\varepsilon(r_C L - K) - b + t \right) \right] = 0,$$

where L is the normalized payment, C its category, $U = (1, \dots, 1, 10)$, $p_j = 1.2 \mathbb{E}_{\hat{P}_N} [L \mathbb{I}_{\{C=j\}}]$, and $\phi_\varepsilon(u) = \varepsilon \log(1 + \exp(\frac{u}{\varepsilon}))$, with $\varepsilon = 0.005$ and $b = 0.05$. The problem has 205 nonnegative variables and 103 equality constraints.

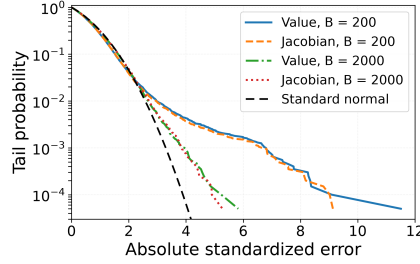


Figure 2: Empirical tails of mini-batch oracle errors.

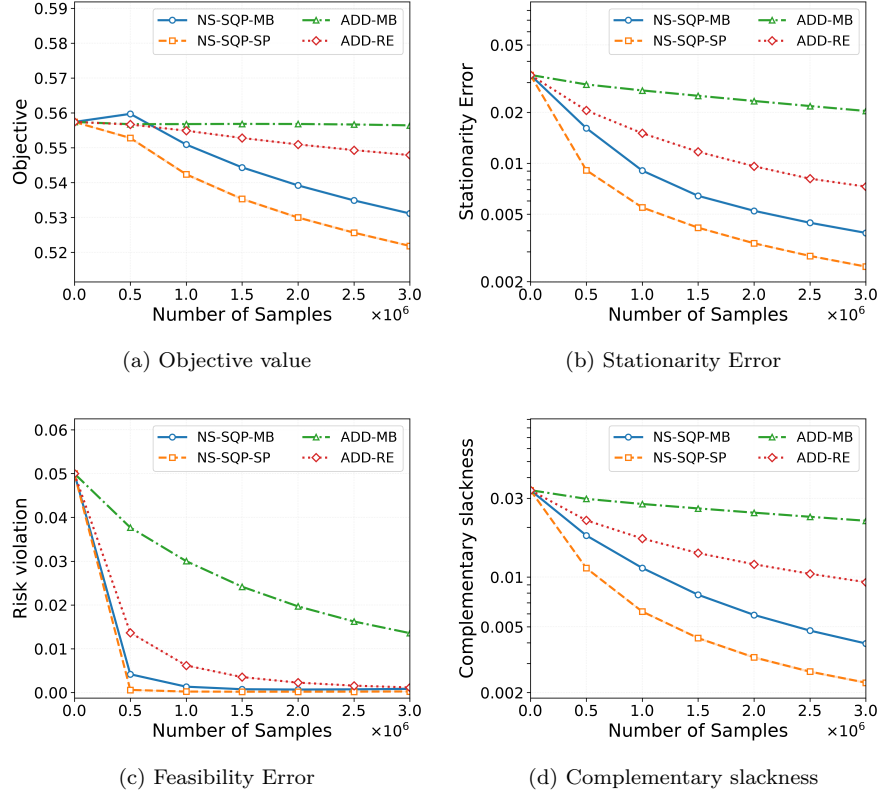


Figure 3: Performance comparison on the flood insurance risk-control problem.

We compare NS-SQP-MB and NS-SQP-SP with ADD-MB and ADD-RE [15], reporting results over 50 random seeds at 3×10^6 sample oracle calls. All methods use batch size 2000, with inner batch size 200 for NS-SQP-SP and ADD-RE. The effective stepsize parameters are tuned over 0.0001, 0.0005, 0.001, 0.003, 0.005, 0.01 for the two NS-SQP variants and 0.01, 0.05, 0.1, 0.2, 0.3, 0.4 for the two ADD variants. Figure 2 shows the empirical tail probabilities of absolute standardized constraint-value and directional Jacobian estimation

r_0	δ_0	Algorithm	Stationarity Error	Feasibility Error	Complementary slackness
0.3	0.075	NS-SQP-MB	1.905×10^{-3}	2.162×10^{-3}	2.844×10^{-3}
		NS-SQP-SP	1.388×10^{-3}	0.893×10^{-3}	2.065×10^{-3}
		ADD-MB	1.729×10^{-2}	2.229×10^{-2}	2.756×10^{-2}
		ADD-RE	7.129×10^{-3}	3.304×10^{-3}	1.387×10^{-2}
0.4	0.1	NS-SQP-MB	3.285×10^{-3}	4.457×10^{-3}	6.979×10^{-3}
		NS-SQP-SP	1.283×10^{-3}	1.539×10^{-3}	2.361×10^{-3}
		ADD-MB	1.503×10^{-2}	3.370×10^{-2}	2.679×10^{-2}
		ADD-RE	7.204×10^{-3}	6.135×10^{-3}	1.465×10^{-2}
0.5	0.15	NS-SQP-MB	1.571×10^{-3}	5.132×10^{-3}	3.951×10^{-3}
		NS-SQP-SP	0.421×10^{-3}	1.795×10^{-3}	1.108×10^{-3}
		ADD-MB	1.301×10^{-2}	5.379×10^{-2}	2.68×10^{-2}
		ADD-RE	7.016×10^{-3}	1.081×10^{-2}	1.534×10^{-2}

Table 2: Results for different initializations on the flood insurance risk-control problem.

errors, based on 20000 independent batches for each batch size in $\{200, 2000\}$. Both errors exhibit substantially heavier empirical tails than the Gaussian benchmark. Define $g(x) = \mathbb{E}_{\hat{P}_N}[\phi_\varepsilon(r_C L - K)] - b$ and $V(x) = \max\{g(x), 0\}$. For each prescribed retention ratio r_0 and initial risk violation $\delta_0 > 0$, all methods share the initialization $r_j^0 = r_0$, $s^0 = U - x^0$, and $t^0 = 0$, with K^0 determined by $\mathbb{E}_{\hat{P}_N}[\phi_\varepsilon(r_0 L - K^0)] - b = \delta_0$. Let $c(z)$ denote the empirical equality-constraint map and $\nabla c(z)$ denote its Jacobian matrix. Common diagnostic multipliers $q \in \mathbb{R}^m$ and $v \in \mathbb{R}_+^n$ are obtained by minimizing $\|\nabla f(z) + \nabla c(z)q - v\|_2^2 + \|z \odot v\|_2^2$, where $\odot$ denotes component-wise multiplication. We define the stationarity error as $\|\nabla f(z) + \nabla c(z)q - v\|_\infty$, the feasibility error as $\max\{\|c(z)\|_\infty, \|[-z]_+\|_\infty\}$, and the complementarity error as $\|z \odot v\|_\infty$, where $[\cdot]_+$ denotes the component-wise positive part.

Figure 3 compares the four methods with initialization $(r_0, \delta_0) = (0.2, 0.05)$ under the same sampling budget. Both NS-SQP variants reduce the stationarity error, risk constraint violation, and complementarity error faster than ADD-MB and ADD-RE, while attaining lower final objective values. In particular, NS-SQP-SP brings the risk violation close to zero within approximately 5×10^5 sample evaluations and achieves the lowest final objective value, stationarity error, and complementarity error. In contrast, ADD-MB retains a noticeable risk violation at the final budget of 3×10^6 sample evaluations. Moreover, we examine the results for different initializations of $(r_0, \delta_0) \in \{(0.3, 0.075), (0.4, 0.1), (0.5, 0.15)\}$. Table 2 reports the mean stationarity error, feasibility error and complementarity slackness across 50 runs. Across all three initializations, both NS-SQP variants attain lower values of all three measures than ADD-MB and ADD-RE, with NS-SQP-SP consistently achieving the lowest means. As r_0 and δ_0 increase together, the feasibility error increases for every method, but NS-SQP-SP exhibits the smallest absolute increase.

5.2 Importance-weighted constrained classification

To examine the robustness of our methods when heavy-tailed noise affects both the objective and constraint estimates, we evaluate our methods on an importance-weighted constrained classification problem constructed from the Open Bandit Dataset² [2], comprising 4,077,727 item–position exposures. The objective penalizes missed clicks, while six expectation constraints control groupwise false-positive surrogate risks. Using the smooth nonconvex surrogate $s(t) = 2/(1 + e^{-t})$, we formulate

$$\begin{aligned}
\min_{x \in [-5, 5]^{518}} \quad & f(x) = \mathbb{E}_\xi[\omega_\xi s(-\psi_\xi^\top x)] + \frac{10^{-3}}{2} \|x\|^2, \\
\text{s.t.} \quad & g_j(x) = \mathbb{E}_\zeta[\omega_{\zeta,j} s(\psi_{\zeta,j}^\top x)] - r \leq 0, \quad j = 1, \dots, 6,
\end{aligned}$$

where ξ and ζ_j are sampled uniformly from the clicked pools and the unclicked pools in group j , respectively. Each feature vector contains 512 encoded shared coordinates and six group indicators. The groups are defined by the three display positions and whether an item’s numerical feature exceeds the median across the 34 catalogue items. Importance weights are inverse logged propensities, normalized once to have mean one

²<https://huggingface.co/datasets/zozonext/open-bandit>

within each sampling pool. Both the objective and constraints are nonconvex. All expectations are taken over the fixed empirical pools, with randomness arising from oracle sampling.

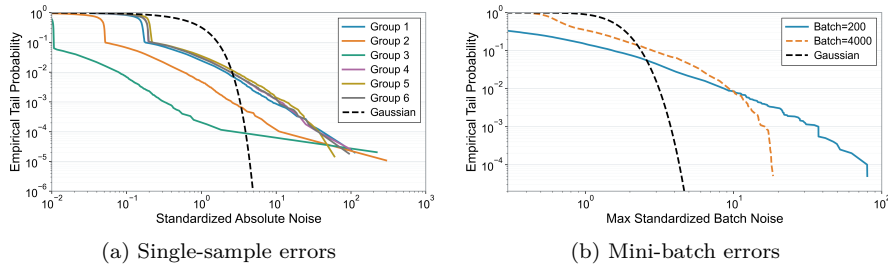


Figure 4: Empirical tails of single-sample and mini-batch oracle errors.

As shown in Figure 4(a), the single-sample constraint errors across all six groups exhibit substantially heavier empirical tails than the Gaussian reference, indicating rare yet extreme oracle fluctuations. These tail deviations persist after mini-batch averaging, as shown in Figure 4(b), where the batch errors remain markedly heavier-tailed than the corresponding Gaussian reference.

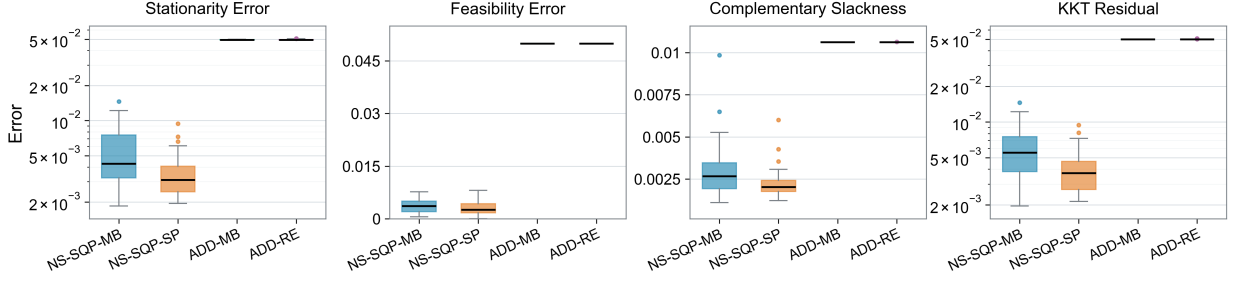
We compare NS-SQP-MB and NS-SQP-SP with ADD-MB and ADD-RE [15] for $r \in \{0.02, 0.05, 0.1, 0.2\}$. All methods start from the same initial point and are run with a budget of 6×10^6 sample evaluations. In the main experiment, the large and inner batch sizes are 4000 and 400, respectively, and the estimators in NS-SQP-SP are refreshed every 5 iterations. We select the effective stepsize parameters from $\{0.001, 0.005, 0.03, 0.1, 0.2, 0.3\}$ and evaluate each selected configuration over 50 independent runs. We report the KKT residual together with its stationarity, feasibility, and complementarity components as defined in subsection 5.1. The boxplots in Figure 5 summarize the terminal residuals over 50 runs per setting. Both NS-SQP variants achieve lower median stationarity, feasibility, complementarity, and KKT errors than the two ADD baselines at all four risk limits. NS-SQP-SP attains smaller KKT residuals at $r = 0.02$ and 0.05 , whereas NS-SQP-MB performs better at $r = 0.1$ and 0.2 . Moreover, Figure 6 shows that both NS-SQP variants maintain substantially smaller maximum group proxy risks than the ADD. At $r = 0.05$, both methods satisfy the risk constraints in all 50 formal runs. Although residual constraint violations remain at $r = 0.2$, both NS-SQP variants still achieve substantially lower mean maximum group proxy risks than the ADD methods. These comparisons across r involve changes in both the constraint limit and the initial infeasibility, which is prescribed as $\max\{0.05, 2r\}$. Thus, the NS methods provide improved constraint control within the available budget.

The effective stepsize sensitivity results in Figure 7 show that small effective stepsizes lead to larger terminal KKT residuals within the prescribed budget. Among the tested candidates, the lowest tuning means are attained at 0.3 for NS-SQP-MB and 0.2 for NS-SQP-SP when $r \in \{0.02, 0.05, 0.2\}$, while both methods prefer 0.1 at $r = 0.1$. These results indicate that moderate to relatively large step sizes perform best within the tested range. Moreover, Figure 8 examines sensitivity to batch configurations $\{500/50, 1000/100, 2000/200, 4000/400\}$. Increasing the batch size improves both methods at $r = 0.02$, but produces larger final KKT residuals at $r = 0.2$. Under the fixed sampling budget, increasing the batch sizes from 500/50 to 4000/400 reduces the number of iterations by approximately a factor of eight. The potential improvement in estimation accuracy may therefore be offset by the reduced number of iterations.

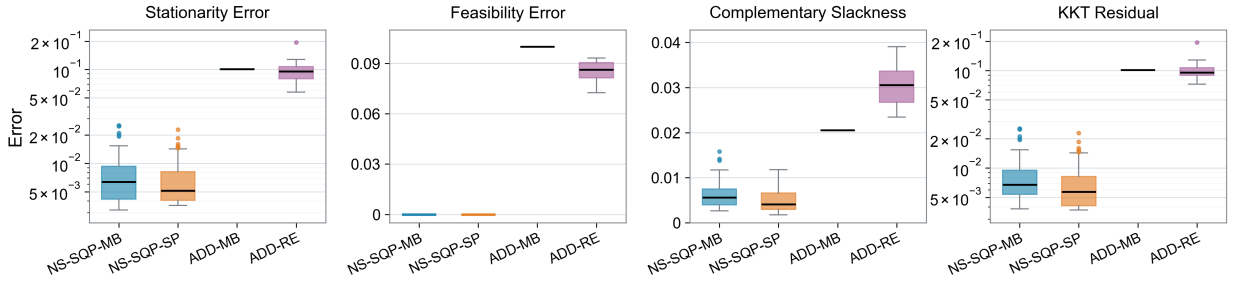
The effective stepsize sensitivity results in Figure 7 show that small effective stepsizes lead to larger terminal KKT residuals within the prescribed budget. Among the tested candidates, the lowest tuning means are attained at 0.3 for NS-SQP-MB and 0.2 for NS-SQP-SP when $r \in \{0.02, 0.05, 0.2\}$, while both methods prefer 0.1 at $r = 0.1$. These results indicate that moderate to relatively large step sizes perform best within the tested range. Moreover, Figure 8 examines sensitivity to batch configurations $\{500/50, 1000/100, 2000/200, 4000/400\}$. Increasing the batch size improves both methods at $r = 0.02$, but produces larger final KKT residuals at $r = 0.2$. Under the fixed sampling budget, increasing the batch sizes from 500/50 to 4000/400 reduces the number of iterations by approximately a factor of eight. The potential improvement in estimation accuracy may therefore be offset by the reduced number of iterations.

5.3 CUTEst experiments under heavy-tailed noise

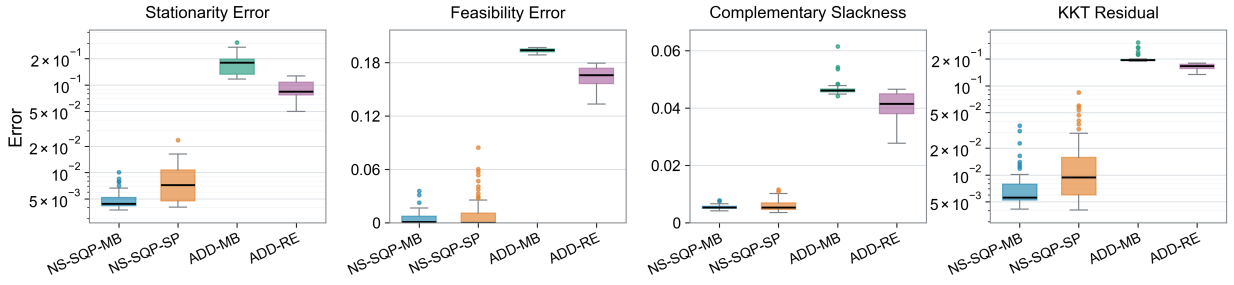
To evaluate our methods under heavy-tailed noise with varying tail indices, we consider 100 equality-constrained CUTEst instances with injected heavy-tailed noise. We compare NS-SQP-MB and NS-SQP-SP with SSQP [32], Stoc-iALM [41], and Prox-EPM [31]. The stochastic objective gradient, constraint value, and constraint Jacobian oracles are perturbed using either symmetric α -stable (S α S) noise or centered Pareto noise. For each distribution, we set $\alpha \in \{1.5, 1.3, 1.1\}$ and the noise-scale parameter to 0.1. For each problem and noise setting, we set a sample budget of 5×10^5 per run. All methods use six batch sizes $\{5, 10, 20, 50, 100, 200\}$. For NS-SQP-SP, we use the (outer batch size, refresh period, inner batch size) triples $\{(5, 2, 2), (10, 2, 4), (20, 4, 4)\}$.



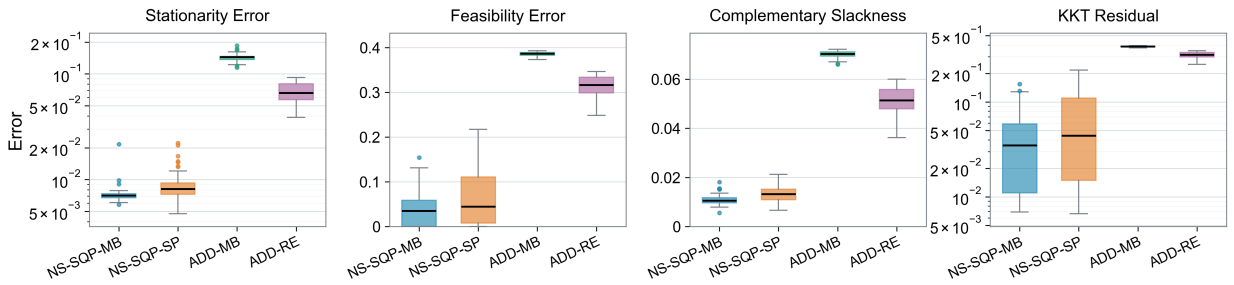
(a) $r = 0.02$



(b) $r = 0.05$



(c) $r = 0.1$



(d) $r = 0.2$

Figure 5: Boxplots of the KKT residual and its components for different risk thresholds r .

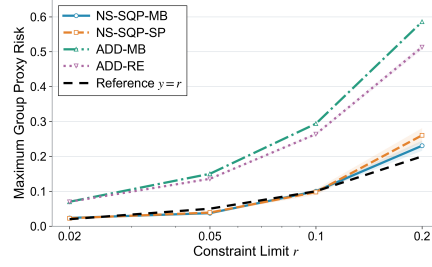


Figure 6: Maximum group proxy risk across constraint limits.

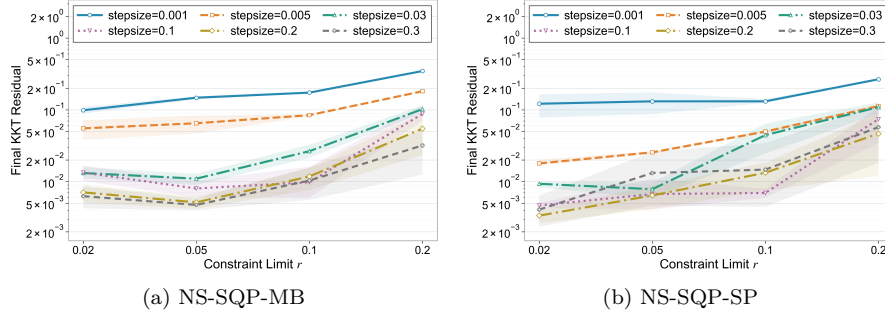


Figure 7: KKT residual with different effective stepsizes.

$(50, 5, 10), (100, 5, 16), (200, 10, 16)\}$. Each batch setting is combined with six method-specific parameter configurations, yielding 36 profiles per method. Specifically, these configurations vary effective stepsize for NS-SQP-MB and NS-SQP-SP, the acceptance parameter for SSQP, the stepsize scale for Stoc-iALM, and the prox-linear stepsize used in the inner penalty subproblems of Prox-EPM. Moreover, the profile with the smallest final KKT residual is selected, with matched random seeds across methods. To evaluate the iterates generated by the algorithms, we define the stationarity error as $\|\nabla f(x) + \nabla c(x)\lambda(x)\|_\infty$, feasibility error as $\|c(x)\|_\infty$ and KKT residual as $\max\{\|\nabla f(x) + \nabla c(x)\lambda(x)\|_\infty, \|c(x)\|_\infty\}$, where $\lambda(x)$ is the solution to $\min_{\lambda \in \mathbb{R}^m} \{\|\nabla f(x) + \nabla c(x)\lambda\|_2^2 + 10^{-8}\|\lambda\|_2^2\}$. The stationarity-error summaries exclude feasibility-only and constant-objective models, whereas the feasibility-error and KKT-residual summaries include all test problems.

Across the six noise settings in Table 3, NS-SQP-MB and NS-SQP-SP jointly obtain the majority of wins and maintain a zero numerical failure rate, whereas each comparison method fails on at least some problems. In every noise setting, the lowest median for each of the three residual measures is attained by an NS-SQP variant. NS-SQP-MB achieves the most wins in all six settings and the lowest median KKT residual in five, while NS-SQP-SP attains the lowest median KKT residual under Pareto noise with $\alpha = 1.1$. As the tail index decreases, the median KKT residual increases for every method under both noise distributions. For each comparison method, the failure rate at $\alpha = 1.1$ exceeds that at $\alpha = 1.5$, while the NS-SQP variants remain free of numerical failures despite the increase in their residuals.

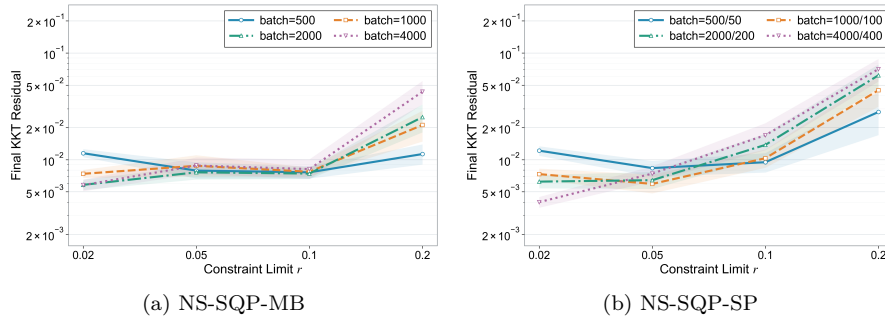


Figure 8: KKT residual with different batch sizes.

Noise	Algorithm	Wins	Failure rate	Median KKT	Median stationarity	Median feasibility
S α S $\alpha = 1.5$	NS-SQP-MB	45	0%	1.620×10^{-2}	5.656×10^{-3}	1.334×10^{-2}
	NS-SQP-SP	17	0%	2.802×10^{-2}	1.245×10^{-2}	1.788×10^{-2}
	SSQP	15	4%	4.801×10^{-2}	4.002×10^{-2}	3.118×10^{-2}
	Stoc-iALM	8	20%	1.379×10^{-1}	7.370×10^{-2}	4.918×10^{-2}
	Prox-EPM	15	15%	4.792×10^{-2}	1.935×10^{-2}	2.633×10^{-2}
S α S $\alpha = 1.3$	NS-SQP-MB	41	0%	2.605×10^{-2}	9.294×10^{-3}	2.124×10^{-2}
	NS-SQP-SP	33	0%	2.860×10^{-2}	1.101×10^{-2}	2.146×10^{-2}
	SSQP	11	4%	8.817×10^{-2}	6.612×10^{-2}	5.605×10^{-2}
	Stoc-iALM	6	18%	2.702×10^{-1}	2.545×10^{-1}	8.164×10^{-2}
	Prox-EPM	9	17%	7.050×10^{-2}	5.719×10^{-2}	5.065×10^{-2}
S α S $\alpha = 1.1$	NS-SQP-MB	43	0%	4.029×10^{-2}	1.516×10^{-2}	3.338×10^{-2}
	NS-SQP-SP	38	0%	4.259×10^{-2}	1.270×10^{-2}	3.971×10^{-2}
	SSQP	13	11%	2.946×10^{-1}	1.680×10^{-1}	1.819×10^{-1}
	Stoc-iALM	2	22%	1.205×10^0	1.360×10^0	2.675×10^{-1}
	Prox-EPM	4	24%	4.035×10^{-1}	2.045×10^{-1}	2.080×10^{-1}
Pareto $\alpha = 1.5$	NS-SQP-MB	41	0%	5.105×10^{-2}	2.165×10^{-2}	4.983×10^{-2}
	NS-SQP-SP	22	0%	5.839×10^{-2}	2.879×10^{-2}	4.720×10^{-2}
	SSQP	13	3%	1.059×10^{-1}	6.680×10^{-2}	7.623×10^{-2}
	Stoc-iALM	15	19%	2.945×10^{-1}	2.819×10^{-1}	5.783×10^{-2}
	Prox-EPM	9	17%	1.320×10^{-1}	9.076×10^{-2}	7.104×10^{-2}
Pareto $\alpha = 1.3$	NS-SQP-MB	37	0%	8.599×10^{-2}	4.820×10^{-2}	7.210×10^{-2}
	NS-SQP-SP	23	0%	1.010×10^{-1}	4.464×10^{-2}	8.825×10^{-2}
	SSQP	11	11%	2.653×10^{-1}	1.294×10^{-1}	1.922×10^{-1}
	Stoc-iALM	11	22%	6.445×10^{-1}	4.685×10^{-1}	2.316×10^{-1}
	Prox-EPM	18	15%	1.744×10^{-1}	1.319×10^{-1}	1.086×10^{-1}
Pareto $\alpha = 1.1$	NS-SQP-MB	43	0%	3.527×10^{-1}	1.853×10^{-1}	3.246×10^{-1}
	NS-SQP-SP	19	0%	3.458×10^{-1}	1.764×10^{-1}	3.384×10^{-1}
	SSQP	17	12%	1.066×10^0	3.824×10^{-1}	6.542×10^{-1}
	Stoc-iALM	11	22%	1.683×10^0	1.641×10^0	4.733×10^{-1}
	Prox-EPM	10	21%	1.538×10^0	5.160×10^{-1}	8.421×10^{-1}

Table 3: Comparison under heavy-tailed noise on 100 CUTEst problems.

6 Conclusion

In this paper, we developed NS-SQP methods for nonconvex optimization with expectation equality constraints and non-negativity bounds under heavy-tailed noise. A regularized normal-step subproblem computes a bound-feasible restoration step, while the total-step subproblem minimizes a sampled quadratic objective model that retains the linearized constraint reduction achieved by the normal step and satisfies the bounds. Normalized updates limit large steps caused by heavy-tailed oracle errors and enable descent analysis based on first-moment error bounds. We established expected ϵ -KKT guarantees, conditioned on completing all prescribed iterations which occurs with probability at least $1 - \epsilon$. We derived separate sample-complexity bounds for objective-gradient, constraint-value, and constraint-Jacobian evaluations under mini-batch estimation, and obtained improved bounds with SPIDER-type recursive estimation. We further specialized the methods to the semi-stochastic setting with exact constraint information and to equality-only problems, establishing unconditional expected ϵ -KKT guarantees and corresponding sample-complexity bounds. We evaluated our methods on a flood insurance risk control problem and an importance-weighted constrained classification problem, and 100 equality-constrained CUTEst problems injected with heavy-tailed noise. These experiments demonstrate the effectiveness of the proposed methods and their advantages in sample efficiency and numerical robustness on real-data models and under injected heavy-tailed noise with varying tail indices.

For problem (1.1), as $p \uparrow 2$, Algorithm 1 with SPIDER-type estimator attains $O(\epsilon^{-5})$ complexity for c and ∇c by using recursive variance reduction and epoch-wise error control (i.e., (4.11) and Lemma 11), whereas mini-batch estimation re-estimates c and ∇c independently at every iteration, only achieving a $O(\epsilon^{-7})$ guarantee under the current analysis. A natural direction for future work is to investigate whether sharper non-progressive probability analysis or improved mini-batch estimators can yield an intuitive $O(\epsilon^{-6})$ sample complexity for estimating c and ∇c .

Declaration of AI use

The authors used ChatGPT to assist with proofreading and language editing, and Codex to assist with developing and debugging the codes for the numerical experiments. The authors reviewed all AI-assisted content and take full responsibility for the final manuscript.

A Technical Lemmas

Lemma 12 *For any $a \in \mathbb{R}$ and $b > 0$, it holds that $\min\{|a|, \frac{a^2}{b^2}\} \geq |a| - \frac{b^2}{4}$.*

Proof. For any $a \in \mathbb{R}$, it holds that $(\frac{|a|}{b^2})^2 \geq (\frac{|a|}{b^2}) - \frac{1}{4}$, implying $\frac{a^2}{b^2} \geq |a| - \frac{b^2}{4}$. Moreover, $|a| \geq |a| - \frac{b^2}{4}$ always holds. $\square$

Lemma 13 (Burkholder-type inequality) *Let $p \in (1, 2)$ and $\{X_i\}_{i=1}^B$ be a martingale difference sequence taking values in $\mathbb{R}^n$ w.r.t. a filtration $\{\mathcal{G}_i\}_{i=0}^B$. That is, $\mathbb{E}[X_i | \mathcal{G}_{i-1}] = 0$ and $\mathbb{E}[\|X_i\|^p] < \infty$ for all $i \in \{1, \dots, B\}$. Then, it holds that $\mathbb{E}[\|\sum_{i=1}^B X_i\|^p] \leq 2^{2-p} \sum_{i=1}^B \mathbb{E}[\|X_i\|^p]$ and $\mathbb{E}[\|\sum_{i=1}^B X_i\|^p | \mathcal{G}_0] \leq 2^{2-p} \sum_{i=1}^B \mathbb{E}[\|X_i\|^p | \mathcal{G}_0]$ almost surely. Furthermore, this result extends directly to matrix-valued $M_j \in \mathbb{R}^{q \times d}$ upon replacing $\|\cdot\|$ with $\|\cdot\|_F$.*

The following contraction estimate is used for the affine KKT mapping in the proof of Lemma 8.

Lemma 14 (Perturbation of an affine generalized equation [42]) *Let $K \subseteq \mathbb{R}^N$ be a nonempty closed convex set and let $F_0 : \mathbb{R}^N \rightarrow \mathbb{R}^N$ be affine. Let $\mathcal{N}_K$ denote the normal cone mapping of K . For every $r \in \mathbb{R}^N$, suppose that the generalized equation $r \in F_0(y) + \mathcal{N}_K(y)$ has a unique solution $S(r)$, and that there exists $L > 0$ for which $\|S(r) - S(r')\| \leq L\|r - r'\|$ for all $r, r' \in \mathbb{R}^N$. Let $h : \mathbb{R}^N \rightarrow \mathbb{R}^N$ satisfy*

$$\text{Lip}(h) := \sup_{y \neq z} \frac{\|h(y) - h(z)\|}{\|y - z\|} < \frac{1}{L}. \quad (\text{A.1})$$

Then $0 \in F_0(y) + h(y) + \mathcal{N}_K(y)$ has a unique solution $\hat{y}$. With $y^* = S(0)$, it satisfies $\|\hat{y} - y^*\| \leq \frac{L}{1 - L \text{Lip}(h)} \|h(y^*)\|$.

Proof. The mapping $y \mapsto S(-h(y))$ is a contraction on $\mathbb{R}^N$, with contraction factor at most $L \text{Lip}(h) < 1$. Its unique fixed point is precisely the solution of the perturbed generalized equation. Moreover, it holds that $\|\hat{y} - y^*\| = \|S(-h(\hat{y})) - S(0)\| \leq L\|h(y^*)\| + L \text{Lip}(h) \|\hat{y} - y^*\|$. Rearranging proves the estimate. $\square$

Lemma 15 (Weyl's Inequality) *Let matrices $A, B \in \mathbb{R}^{m \times n}$ with $q = \min\{m, n\}$, and let $\sigma_1(\cdot) \geq \dots \geq \sigma_q(\cdot) \geq 0$ denote their singular values in descending order. Then, for any $i \in \{1, \dots, q\}$, it holds that $|\sigma_i(A) - \sigma_i(B)| \leq \|A - B\|_2$.*

Lemma 16 (Perturbation of orthogonal projections) *Let $A, B \in \mathbb{R}^{n \times m}$ have the same rank, and set $P = AA^\dagger$ and $Q = BB^\dagger$. Then, it holds that $\|P - Q\| \leq \min\{\|A^\dagger\|, \|B^\dagger\|\} \|A - B\|$.*

Proof. The rank-zero case is immediate. Otherwise, let $U, V \in \mathbb{R}^{n \times r}$ have orthonormal columns spanning the ranges of A and B , where $r = \text{rank}(A) = \text{rank}(B)$. Then $P = UU^\top$ and $Q = VV^\top$. Define $D = U^\top V$, we have $\|(I - Q)P\|^2 = \lambda_{\max}(I_r - DD^\top)$ and $\|(I - P)Q\|^2 = \lambda_{\max}(I_r - D^\top D)$. Then the equality $\|(I - Q)P\| = \|(I - P)Q\|$ holds because the square matrices $DD^\top$ and $D^\top D$ have the same eigenvalues, including multiplicities. Moreover, it holds that $(P - Q)^2 = P(I - Q)P + (I - P)Q(I - P)$. Since $(P - Q)^2$ is block diagonal w.r.t.

$\text{Range}(P) \oplus \text{Null}(P)$, and $P - Q$ is symmetric, we obtain $\|P - Q\|^2 = \max\{\|P(I - Q)P\|, \|(I - P)Q(I - P)\|\} = \max\{\|(I - Q)P\|^2, \|(I - P)Q\|^2\}$, where the second equality uses $\|A^\top A\| = \|AA^\top\| = \|A\|^2$. Taking square roots and using the established equality $\|(I - Q)P\| = \|(I - P)Q\|$ yields $\|P - Q\| = \|(I - Q)P\|$. Further using $(I - Q)B = 0$ gives $\|P - Q\| = \|(I - Q)(A - B)A^\dagger\| \leq \|I - Q\| \|A - B\| \|A^\dagger\| \leq \|A - B\| \|A^\dagger\|$, where the last inequality uses $\|I - Q\| \leq 1$ for orthogonal projections. Interchanging A and B gives the other bound. $\square$

B Proofs for Section 4.4

We first bound the probability that the sampled Jacobian is nearly singular.

Lemma 17 *Under Assumption 5 with $\mathcal{I}_t^a = \emptyset$ and Condition 2, $\mathbb{P}(\mathcal{A}_t^c) \leq 2E_{t,J}/\nu$ for every $t \in [T]$.*

Proof. Under Assumption 5 with $\mathcal{I}_t^a = \emptyset$, the exact constraint Jacobian satisfies $\sigma_{\min}(\nabla c_t) \geq \nu$. By invoking Weyl's inequality for singular values (Lemma 15), we can bound the minimum singular value of the stochastic estimator from below, i.e., $\sigma_{\min}(\tilde{\nabla} c_t) \geq \sigma_{\min}(\nabla c_t) - \|\tilde{\nabla} c_t - \nabla c_t\| \geq \nu - \|\tilde{\nabla} c_t - \nabla c_t\|$. If $\mathcal{A}_t^c = \{\sigma_{\min}(\tilde{\nabla} c_t) < \nu/2\}$ occurs, it follows that $\|\tilde{\nabla} c_t - \nabla c_t\| > \nu/2$. This indicates that $\mathcal{A}_t^c$ can only occur when the estimation error is exceptionally large. Moreover, we have the set inclusion $\mathcal{A}_t^c \subseteq \{\|\tilde{\nabla} c_t - \nabla c_t\| > \nu/2\} \subseteq \{\|\tilde{\nabla} c_t - \nabla c_t\| \geq \nu/2\}$. Consequently, applying Markov's inequality and Condition 2 yields $\mathbb{P}(\mathcal{A}_t^c) \leq \mathbb{P}(\|\tilde{\nabla} c_t - \nabla c_t\| \geq \nu/2) \leq 2E_{t,J}/\nu$, which completes the proof. $\square$

We next bound the merit parameter uniformly over all sample paths, allowing both sampling events to be handled in the expected descent analysis.

Lemma 18 *Suppose Assumption 5 with $\mathcal{I}_t^a = \emptyset$ and Condition 1 hold. For any $t \in [T]$, the merit parameter $\tilde{\rho}_t$ is uniformly upper bounded by $\rho_{\max} := \max\{\tilde{\rho}_0, (4M_f L_c + 2M_c L_c^2)/\nu^2\}$.*

Proof. By applying the orthogonality of $\tilde{u}_t$ and $\tilde{v}_t$, we decouple the numerator of the first branch of $\tilde{\rho}_t$ into tangential and normal components, i.e.,

$$\tilde{\nabla} f_t^\top \tilde{d}_t + \frac{1}{2} \|\tilde{d}_t\|^2 = \tilde{\nabla} f_t^\top \tilde{u}_t + \frac{1}{2} \|\tilde{u}_t\|^2 + \tilde{\nabla} f_t^\top \tilde{v}_t + \frac{1}{2} \|\tilde{v}_t\|^2. \quad (\text{B.1})$$

For the tangential component $\tilde{u}_t = -P_{\tilde{U}_t} \tilde{\nabla} f_t$, leveraging the symmetry and idempotence of the orthogonal projection matrix $P_{\tilde{U}_t}$ yields $\tilde{\nabla} f_t^\top \tilde{u}_t = -\|P_{\tilde{U}_t} \tilde{\nabla} f_t\|^2$. The tangential component becomes $\tilde{\nabla} f_t^\top \tilde{u}_t + \frac{1}{2} \|\tilde{u}_t\|^2 = -\frac{1}{2} \|P_{\tilde{U}_t} \tilde{\nabla} f_t\|^2 \leq 0$. For the normal component $\tilde{v}_t = -\tilde{\nabla} c_t \tilde{c}_t$, it satisfies $\|\tilde{v}_t\| \leq \|\tilde{\nabla} c_t\| \|\tilde{c}_t\|$. Thus, the normal component is linearly bounded by $\|\tilde{c}_t\|$, i.e.,

$$\tilde{\nabla} f_t^\top \tilde{v}_t + \frac{1}{2} \|\tilde{v}_t\|^2 \leq (\|\tilde{\nabla} f_t\| + \frac{1}{2} \|\tilde{\nabla} c_t\| \|\tilde{c}_t\|) \|\tilde{\nabla} c_t\| \|\tilde{c}_t\| \leq (M_f L_c + \frac{1}{2} L_c^2 M_c) \|\tilde{c}_t\|,$$

where the last inequality follows from Condition 1. Hence, (B.1) implies $\tilde{\nabla} f_t^\top \tilde{d}_t + \frac{1}{2} \|\tilde{d}_t\|^2 \leq (M_f L_c + \frac{1}{2} L_c^2 M_c) \|\tilde{c}_t\|$. On $\mathcal{A}_t \cap \{\tilde{c}_t \neq \mathbf{0}\}$, the preceding bound and $\sigma_{\min}(\tilde{\nabla} c_t) \geq \nu/2$ yield

$$\frac{\tilde{\nabla} f_t^\top \tilde{d}_t + \frac{1}{2} \|\tilde{d}_t\|^2}{(\sigma_{\min}(\tilde{\nabla} c_t))^2 \|\tilde{c}_t\|} \leq \frac{M_f L_c + \frac{1}{2} M_c L_c^2}{(\sigma_{\min}(\tilde{\nabla} c_t))^2} \leq \frac{4M_f L_c + 2M_c L_c^2}{\nu^2} \leq \rho_{\max}.$$

On $\mathcal{A}_t^c$ or when $\tilde{c}_t = \mathbf{0}$, the merit parameter remains unchanged. Since $0 < \tilde{\rho}_0 \leq \rho_{\max}$, an induction using the definition of $\tilde{\rho}_t$ gives $0 < \tilde{\rho}_t \leq \rho_{\max}$ for every $t \in [T]$. $\square$

The following lemma bounds the expected norm of the sampled total step at the randomized output of Algorithm 2.

Lemma 19 *Under Assumption 5 with $\mathcal{I}_t^a = \emptyset$, Conditions 1 and 2, the output x_R of Algorithm 2 satisfies*

$$\begin{aligned} \mathbb{E}[\|\tilde{d}_R\|] &\leq \frac{2\ell_{\max}(f_1 - f^* + \rho_{\max} M_c)}{K_T} + \frac{(4\ell_{\max} + L_c^2) \sum_{t=1}^T \kappa_t^2}{4\ell_{\min} K_T} \\ &\quad + \frac{1}{K_T} \sum_{t=1}^T \left(\frac{2\ell_{\max} \kappa_t}{\ell_{\min}} E_{t,f} + 4\ell_{\max} \rho_{\max} E_{t,c} + \frac{\ell_{\max}}{\ell_{\min}} \left(2\rho_{\max} \kappa_t + \frac{(4M_f + 2L_c M_c) \kappa_t}{\nu} \right) E_{t,J} \right), \end{aligned}$$

with notations $\ell_t := L_f + \tilde{\rho}_t L_J$, $\ell_{\min} := L_f + \tilde{\rho}_0 L_J$ and $\ell_{\max} := L_f + \rho_{\max} L_J$.

Proof. The smoothness of f and c implies

$$\begin{aligned} f_{t+1} + \tilde{\rho}_t \|c_{t+1}\| &\leq f_t + \eta_t \nabla f_t^\top \tilde{d}_t + \tilde{\rho}_t \|c_t\| + \eta_t \nabla c_t^\top \tilde{d}_t + \frac{\eta_t^2 L_f + \eta_t^2 \tilde{\rho}_t L_J}{2} \|\tilde{d}_t\|^2 \\ &\leq f_t + \eta_t \tilde{\nabla} f_t^\top \tilde{d}_t + \tilde{\rho}_t \|\tilde{c}_t\| - \eta_t \tilde{\nabla} c_t^\top \tilde{\nabla} c_t \tilde{c}_t + \frac{\ell_t}{2} \eta_t^2 \|\tilde{d}_t\|^2 \\ &\quad + \eta_t \|\nabla f_t - \tilde{\nabla} f_t\| \|\tilde{d}_t\| + \tilde{\rho}_t \|\tilde{c}_t - c_t\| + \tilde{\rho}_t \eta_t \|\tilde{\nabla} c_t - \nabla c_t\| \|\tilde{d}_t\|, \end{aligned} \quad (\text{B.2})$$

where we use the orthogonality of the update direction in last inequality. By conditioning on the event $\mathcal{A}_t$ and its complement $\mathcal{A}_t^c$, the expectation of $\tilde{\rho}_t \|\tilde{c}_t - \eta_t \tilde{\nabla} c_t^\top \tilde{\nabla} c_t \tilde{c}_t\|$ can be expanded as

$$\mathbb{E}[\tilde{\rho}_t \|\tilde{c}_t - \eta_t \tilde{\nabla} c_t^\top \tilde{\nabla} c_t \tilde{c}_t\|] = \mathbb{E}[\tilde{\rho}_t \|\tilde{c}_t - \eta_t \tilde{\nabla} c_t^\top \tilde{\nabla} c_t \tilde{c}_t\| \cdot \mathbb{I}_{\mathcal{A}_t}] + \mathbb{E}[\tilde{\rho}_t \|\tilde{c}_t - \eta_t \tilde{\nabla} c_t^\top \tilde{\nabla} c_t \tilde{c}_t\| \cdot \mathbb{I}_{\mathcal{A}_t^c}]. \quad (\text{B.3})$$

On $\mathcal{A}_t$, the step-size condition $\eta_t \leq 1/L_c^2$ ensures that inequality $\|\tilde{c}_t - \eta_t \tilde{\nabla} c_t^\top \tilde{\nabla} c_t \tilde{c}_t\| \leq (1 - \eta_t (\sigma_{\min}(\tilde{\nabla} c_t))^2) \|\tilde{c}_t\|$ holds. This guarantees that

$$\begin{aligned} \mathbb{E}[\tilde{\rho}_t \|\tilde{c}_t - \eta_t \tilde{\nabla} c_t^\top \tilde{\nabla} c_t \tilde{c}_t\| \cdot \mathbb{I}_{\mathcal{A}_t}] &\leq \mathbb{E}[\tilde{\rho}_t (1 - \eta_t (\sigma_{\min}(\tilde{\nabla} c_t))^2) \|\tilde{c}_t\| \cdot \mathbb{I}_{\mathcal{A}_t}] \\ &\leq \mathbb{E}\left[\left(\tilde{\rho}_t \|\tilde{c}_t\| - \eta_t \tilde{\nabla} f_t^\top \tilde{d}_t - \frac{\eta_t}{2} \|\tilde{d}_t\|^2\right) \cdot \mathbb{I}_{\mathcal{A}_t}\right]. \end{aligned} \quad (\text{B.4})$$

The last inequality follows from the merit-parameter update rule when $\tilde{c}_t \neq \mathbf{0}$, and it follows from $\tilde{\nabla} f_t^\top \tilde{d}_t = -\|\tilde{d}_t\|^2$ when $\tilde{c}_t = \mathbf{0}$. Turning to the second term on the right-hand side of (B.3), the bounds $\eta_t \leq 1/L_c^2$ and $\|\tilde{\nabla} c_t\| \leq L_c$ imply $0 \leq I - \eta_t \tilde{\nabla} c_t^\top \tilde{\nabla} c_t \leq I$. Consequently, it holds that $\mathbb{E}[\tilde{\rho}_t \|\tilde{c}_t - \eta_t \tilde{\nabla} c_t^\top \tilde{\nabla} c_t \tilde{c}_t\| \cdot \mathbb{I}_{\mathcal{A}_t^c}] \leq \mathbb{E}[\tilde{\rho}_t \|\tilde{c}_t\| \cdot \mathbb{I}_{\mathcal{A}_t^c}]$. Moreover, the tangential-step identity and the orthogonality of $\tilde{u}_t$ and $\tilde{v}_t$ give

$$\eta_t \left(\tilde{\nabla} f_t^\top \tilde{d}_t + \frac{1}{2} \|\tilde{d}_t\|^2 \right) = \eta_t \left(-\frac{1}{2} \|\tilde{u}_t\|^2 + \tilde{\nabla} f_t^\top \tilde{v}_t + \frac{1}{2} \|\tilde{v}_t\|^2 \right) \leq \eta_t \left(M_f + \frac{1}{2} L_c M_c \right) \|\tilde{v}_t\| \leq \left(M_f + \frac{1}{2} L_c M_c \right) \frac{\kappa_t}{\ell_t},$$

where we invoke the bounds $\|\tilde{v}_t\| \leq L_c M_c$, $\|\tilde{v}_t\| \leq \|\tilde{d}_t\|$, and $\eta_t \|\tilde{d}_t\| \leq \kappa_t / \ell_t$. It follows that

$$\begin{aligned} \mathbb{E}[\tilde{\rho}_t \|\tilde{c}_t - \eta_t \tilde{\nabla} c_t^\top \tilde{\nabla} c_t \tilde{c}_t\| \cdot \mathbb{I}_{\mathcal{A}_t^c}] &\leq \mathbb{E}[\tilde{\rho}_t \|\tilde{c}_t\| \cdot \mathbb{I}_{\mathcal{A}_t^c}] \\ &= \mathbb{E}\left[\left(\tilde{\rho}_t \|\tilde{c}_t\| - \eta_t \tilde{\nabla} f_t^\top \tilde{d}_t - \frac{\eta_t}{2} \|\tilde{d}_t\|^2\right) \cdot \mathbb{I}_{\mathcal{A}_t^c}\right] + \mathbb{E}\left[\left(\eta_t \tilde{\nabla} f_t^\top \tilde{d}_t + \frac{\eta_t}{2} \|\tilde{d}_t\|^2\right) \cdot \mathbb{I}_{\mathcal{A}_t^c}\right] \\ &\leq \mathbb{E}\left[\left(\tilde{\rho}_t \|\tilde{c}_t\| - \eta_t \tilde{\nabla} f_t^\top \tilde{d}_t - \frac{\eta_t}{2} \|\tilde{d}_t\|^2\right) \cdot \mathbb{I}_{\mathcal{A}_t^c}\right] + \left(M_f + \frac{1}{2} L_c M_c\right) \frac{\kappa_t}{\ell_{\min}} \mathbb{P}(\mathcal{A}_t^c), \end{aligned}$$

where the last inequality follows from $0 < \ell_{\min} \leq \ell_t$. Plugging it as well as (B.4) into (B.3) and invoking Lemma 17 (i.e., $\mathbb{P}(\mathcal{A}_t^c) \leq 2E_{t,J}/\nu$) yields

$$\mathbb{E}[\tilde{\rho}_t \|\tilde{c}_t - \eta_t \tilde{\nabla} c_t^\top \tilde{\nabla} c_t \tilde{c}_t\|] \leq \mathbb{E}[\tilde{\rho}_t \|\tilde{c}_t\|] - \mathbb{E}[\eta_t \tilde{\nabla} f_t^\top \tilde{d}_t] - \frac{1}{2} \mathbb{E}[\eta_t \|\tilde{d}_t\|^2] + \frac{(2M_f + L_c M_c) \kappa_t}{\nu \ell_{\min}} E_{t,J}.$$

Taking the expectation on both sides of (B.2) and substituting the preceding inequality back into it, we obtain

$$\begin{aligned} \mathbb{E}[f_{t+1} + \tilde{\rho}_t \|c_{t+1}\|] &\leq \mathbb{E}[f_t + \tilde{\rho}_t \|c_t\|] - \frac{\mathbb{E}[\eta_t \|\tilde{d}_t\|^2]}{2} + \frac{\mathbb{E}[\ell_t \eta_t^2 \|\tilde{d}_t\|^2]}{2} + \mathbb{E}[\eta_t \|\nabla f_t - \tilde{\nabla} f_t\| \|\tilde{d}_t\|] \\ &\quad + 2\mathbb{E}[\tilde{\rho}_t \|\tilde{c}_t - c_t\|] + \mathbb{E}[\tilde{\rho}_t \eta_t \|\tilde{\nabla} c_t - \nabla c_t\| \|\tilde{d}_t\|] + \frac{(2M_f + L_c M_c) \kappa_t}{\nu \ell_{\min}} E_{t,J}. \end{aligned} \quad (\text{B.5})$$

Recalling the definition $\eta_t = \min\{\kappa_t/(\ell_t \|\tilde{d}_t\|), 1/L_c^2\}$, it follows from $0 < \ell_{\min} \leq \ell_t \leq \ell_{\max}$ that

$$\eta_t \|\tilde{d}_t\|^2 \geq \min\left\{\frac{\kappa_t \|\tilde{d}_t\|}{\ell_{\max}}, \frac{\|\tilde{d}_t\|^2}{L_c^2}\right\} = \frac{\kappa_t^2}{\ell_{\max}^2} \min\left\{\frac{\ell_{\max} \|\tilde{d}_t\|}{\kappa_t}, \frac{(\ell_{\max} \|\tilde{d}_t\|)^2}{\kappa_t^2 L_c^2}\right\} \geq \frac{\kappa_t}{\ell_{\max}} \|\tilde{d}_t\| - \frac{\kappa_t^2 L_c^2}{4\ell_{\max}^2},$$

where we invoke Lemma 12 in the second inequality, and $\eta_t \|\tilde{d}_t\| \leq \kappa_t / \ell_t \leq \kappa_t / \ell_{\min}$. Plugging these inequalities into (B.5) yields

$$\begin{aligned} \mathbb{E}[f_{t+1} + \tilde{\rho}_t \|c_{t+1}\|] &\leq \mathbb{E}[f_t + \tilde{\rho}_t \|c_t\|] - \frac{\kappa_t \mathbb{E}[\|\tilde{d}_t\|]}{2\ell_{\max}} + \left(\frac{1}{2\ell_{\min}} + \frac{L_c^2}{8\ell_{\max}\ell_{\min}}\right) \kappa_t^2 + \frac{\kappa_t}{\ell_{\min}} \mathbb{E}[\|\nabla f_t - \tilde{\nabla} f_t\|] \\ &\quad + 2\mathbb{E}[\tilde{\rho}_t \|\tilde{c}_t - c_t\|] + \frac{\kappa_t}{\ell_{\min}} \mathbb{E}[\tilde{\rho}_t \|\tilde{\nabla} c_t - \nabla c_t\|] + \frac{(2M_f + L_c M_c) \kappa_t}{\nu \ell_{\min}} E_{t,J}, \end{aligned}$$

where we use $1/\ell_{\max}^2 \leq 1/(\ell_{\max}\ell_{\min})$ in this inequality. Then, arranging the terms and summing from $t = 1$ to T leads to

$$\begin{aligned}
& \sum_{t=1}^T \kappa_t \mathbb{E}[\|\tilde{d}_t\|] \\
& \leq 2\ell_{\max} \mathbb{E} \left[\sum_{t=1}^T (f_t - f_{t+1} + \tilde{\rho}_t(\|c_t\| - \|c_{t+1}\|)) \right] + \frac{(4\ell_{\max} + L_c^2) \sum_{t=1}^T \kappa_t^2}{4\ell_{\min}} \\
& + \sum_{t=1}^T \left(\frac{2\ell_{\max}\kappa_t}{\ell_{\min}} \mathbb{E}[\|\nabla f_t - \tilde{\nabla} f_t\|] + 4\ell_{\max} \mathbb{E}[\tilde{\rho}_t\|c_t - \tilde{c}_t\|] + \frac{2\ell_{\max}\kappa_t}{\ell_{\min}} \mathbb{E}[\tilde{\rho}_t\|\nabla c_t - \tilde{\nabla} c_t\|] + \frac{(4M_f + 2L_c M_c)\ell_{\max}\kappa_t}{\nu\ell_{\min}} E_{t,J} \right) \\
& \leq 2\ell_{\max}(f_1 - f^* + \rho_{\max} M_c) + \frac{(4\ell_{\max} + L_c^2) \sum_{t=1}^T \kappa_t^2}{4\ell_{\min}} \\
& + \sum_{t=1}^T \left(\frac{2\ell_{\max}\kappa_t}{\ell_{\min}} E_{t,f} + 4\ell_{\max}\rho_{\max} E_{t,c} + \left(\frac{2\ell_{\max}\rho_{\max}\kappa_t}{\ell_{\min}} + \frac{(4M_f + 2L_c M_c)\ell_{\max}\kappa_t}{\nu\ell_{\min}} \right) E_{t,J} \right),
\end{aligned}$$

where the last inequality follows from the fact that the merit parameter $\tilde{\rho}_t$ is nondecreasing and uniformly bounded, and $\|c_t\| \leq M_c$. Consequently, the conclusion follows from $\mathbb{E}[\|\tilde{d}_R\|] = \sum_{t=1}^T \kappa_t \mathbb{E}[\|\tilde{d}_t\|]/K_T$. $\square$

To obtain the stationarity and feasibility guarantee at the output, we still need a total-step perturbation bound linking the sampled total step $\tilde{d}_t$ to its deterministic counterpart, i.e. the true total step d_t defined by

$$d_t = u_t + v_t = -P_{U_t} \nabla f_t - \nabla c_t c_t, \quad \text{with } U_t := \{u \in \mathbb{R}^n : \nabla c_t^\top u = \mathbf{0}\}.$$

Lemma 20 Suppose that Assumption 5 with $\mathcal{I}_t^a = \emptyset$, Conditions 1 and 2 hold. For any $t \in [T]$, the expected deviation between the sampled direction $\tilde{d}_t$ and the true direction d_t satisfies $\mathbb{E}[\|\tilde{d}_t - d_t\|] \leq E_{t,f} + L_c E_{t,c} + \frac{(M_f + \nu M_c) E_{t,J}}{\nu}$.

Proof. By applying the triangle inequality, we decouple the total direction error into its tangential and normal components, i.e., $\mathbb{E}[\|\tilde{d}_t - d_t\|] \leq \mathbb{E}[\|\tilde{u}_t - u_t\|] + \mathbb{E}[\|\tilde{v}_t - v_t\|]$. For tangential components, it holds that

$$\|\tilde{u}_t - u_t\| \leq \|P_{\tilde{U}_t}(\tilde{\nabla} f_t - \nabla f_t)\| + \|(P_{\tilde{U}_t} - P_{U_t})\nabla f_t\| \leq \|\tilde{\nabla} f_t - \nabla f_t\| + M_f \|P_{\tilde{U}_t} - P_{U_t}\|, \quad (\text{B.6})$$

where we use the fact that $\|P_{\tilde{U}_t}\| \leq 1$ and the boundedness of $\|\nabla f_t\|$. Next, we focus on bounding the term $\|P_{\tilde{U}_t} - P_{U_t}\|$. By Assumption 5 with $\mathcal{I}_t^a = \emptyset$, ∇c_t has full column rank and $\|(\nabla c_t)^\dagger\| \leq 1/\nu$. If $\tilde{\nabla} c_t$ also has full column rank, invoking Lemma 16 on the associated column-space orthogonal projectors yields

$$\|P_{\tilde{U}_t} - P_{U_t}\| \leq \min \{ \|(\nabla c_t)^\dagger\|, \|(\tilde{\nabla} c_t)^\dagger\| \} \|\tilde{\nabla} c_t - \nabla c_t\| \leq \frac{1}{\nu} \|\tilde{\nabla} c_t - \nabla c_t\|.$$

Otherwise, choose $z \in \mathbb{R}^m$ with $\|z\| = 1$ and $\tilde{\nabla} c_t z = \mathbf{0}$. Then, it follows that $\|\tilde{\nabla} c_t - \nabla c_t\| \geq \|(\tilde{\nabla} c_t - \nabla c_t)z\| = \|\nabla c_t z\| \geq \nu$. Since the difference of two orthogonal projectors has spectral norm at most one, we obtain $\|P_{\tilde{U}_t} - P_{U_t}\| \leq 1 \leq \frac{1}{\nu} \|\tilde{\nabla} c_t - \nabla c_t\|$. Substituting it into (B.6) and taking expectations on both sides yields $\mathbb{E}[\|\tilde{u}_t - u_t\|] \leq E_{t,f} + M_f E_{t,J}/\nu$. For normal components, the boundedness of c_t and $\tilde{\nabla} c_t$ implies

$$\mathbb{E}[\|\tilde{v}_t - v_t\|] = \mathbb{E}[\|\tilde{\nabla} c_t \tilde{c}_t - \nabla c_t c_t\|] \leq \mathbb{E}[\|\tilde{\nabla} c_t\| \|\tilde{c}_t - c_t\|] + \mathbb{E}[\|\tilde{\nabla} c_t - \nabla c_t\| \|c_t\|] \leq L_c E_{t,c} + M_c E_{t,J}.$$

Combining the bounds for the tangential and normal error derives the result. $\square$

Combining Lemmas 19 and 20 yields a bound on $\mathbb{E}[\|d_R\|]$ and hence on the expected KKT residuals. Define

$$\begin{aligned}
\Delta_{T,3} := & \left(1 + \frac{2\ell_{\max}}{\ell_{\min}} \right) \frac{\sum_{t=1}^T \kappa_t E_{t,f}}{K_T} + \left(\frac{M_f + \nu M_c}{\nu} + \frac{\ell_{\max}}{\ell_{\min}} \left(2\rho_{\max} + \frac{4M_f + 2L_c M_c}{\nu} \right) \right) \frac{\sum_{t=1}^T \kappa_t E_{t,J}}{K_T} \\
& + \frac{L_c \sum_{t=1}^T \kappa_t E_{t,c}}{K_T} + \frac{4\ell_{\max}\rho_{\max} \sum_{t=1}^T E_{t,c}}{K_T} + \frac{2\ell_{\max}(f_1 - f^* + \rho_{\max} M_c)}{K_T} + \frac{(4\ell_{\max} + L_c^2) \sum_{t=1}^T \kappa_t^2}{4\ell_{\min} K_T}.
\end{aligned}$$

Theorem 4 Suppose Assumption 5 with $\mathcal{I}_t^a = \emptyset$, Conditions 1 and 2 hold. Let x_R be the randomized output of Algorithm 2 with probability $\mathbb{P}(R = t) = \kappa_t/K_T$. Then, it holds that $\mathbb{E}[\|d_R\|] \leq \Delta_{T,3}$. Moreover, given $\epsilon \in (0, 1)$, if $\Delta_{T,3} \leq \min\{\epsilon, \nu\epsilon\}$, there exists $\lambda_R \in \mathbb{R}^m$ such that $\mathbb{E}[\|\nabla f_R + \nabla c_R \lambda_R\|] \leq \epsilon, \mathbb{E}[\|c_R\|] \leq \epsilon$.

Proof. By applying the triangle inequality, it holds that

$$\mathbb{E}[\|d_R\|] \leq \mathbb{E}[\|\tilde{d}_R\|] + \mathbb{E}[\|d_R - \tilde{d}_R\|] = \mathbb{E}[\|\tilde{d}_R\|] + \frac{\sum_{t=1}^T \kappa_t \mathbb{E}[\|d_t - \tilde{d}_t\|]}{K_T}.$$

Combining Lemmas 19 and 20 with bounded estimation errors yields $\mathbb{E}[\|d_R\|] \leq \Delta_{T,3}$. Defining $\lambda_R = -(\nabla c_R^\top \nabla c_R)^{-1} \nabla c_R^\top \nabla f_R \in \mathbb{R}^m$, we have $(\nabla f_R + \nabla c_R \lambda_R) \perp \nabla c_R c_R$. Then it follows from the definition of true total step d_R that $\|\nabla f_R + \nabla c_R \lambda_R\| \leq \|d_R\|$ and $\|\nabla c_R c_R\| \leq \|d_R\|$. Taking the expectation on both sides and using the condition $\Delta_{T,3} \leq \min\{\epsilon, \nu\epsilon\}$, we obtain

$$\mathbb{E}[\|\nabla f_R + \nabla c_R \lambda_R\|] \leq \mathbb{E}[\|d_R\|] \leq \Delta_{T,3} \leq \epsilon, \quad \text{and} \quad \mathbb{E}[\|\nabla c_R c_R\|] \leq \mathbb{E}[\|d_R\|] \leq \Delta_{T,3} \leq \nu\epsilon.$$

Finally, leveraging Assumption 5 with $\mathcal{I}_t^a = \emptyset$, the expected constraint violation is bounded by the normal component, which yields $\mathbb{E}[\|c_R\|] \leq \mathbb{E}[\|\nabla c_R c_R\|]/\nu \leq \epsilon$. The proof is thus completed. $\square$

To establish the sample complexity for finding the expected KKT point using Algorithm 2, it suffices to ensure that $\Delta_{T,3} = \mathcal{O}(\epsilon)$ with a sufficiently small implied constant. With the mini-batch estimation and SPIDER-type recursive estimation, we provide the corresponding sample complexity in the following theorem.

Theorem 5 Suppose Assumptions 1, 2, 3 and 5 with $\mathcal{I}_t^a = \emptyset$ hold. Given a tolerance $\epsilon \in (0, 1)$ and $t \in [T]$, either of the following sampling strategies returns an expected ϵ -KKT point.

- (1) Mini-batch estimation. Choose $\kappa_t \asymp \epsilon$, $B_{t,f} \asymp \lceil \epsilon^{\frac{p}{1-p}} \rceil$, $B_{t,c} \asymp \lceil \epsilon^{\frac{2p}{1-p}} \rceil$, $B_{t,J} \asymp \lceil \epsilon^{\frac{p}{1-p}} \rceil$. Then, Algorithm 2 returns an expected ϵ -KKT point after $T \asymp \lceil \epsilon^{-2} \rceil$ iterations, with the sample complexity of order $\mathcal{O}(\epsilon^{\frac{3p-2}{1-p}})$, $\mathcal{O}(\epsilon^{\frac{3p-2}{1-p}})$ and $\mathcal{O}(\epsilon^{\frac{4p-2}{1-p}})$ for computing $\tilde{\nabla} f_t$, $\tilde{\nabla} c_t$ and $\tilde{c}_t$, respectively.
- (2) SPIDER-type recursive estimation. Additionally, suppose that Assumption 6 holds. Choose $\kappa_t \asymp \epsilon$, $B_{t,f} \asymp \lceil \epsilon^{\frac{p}{1-p}} \rceil$, $B_{t,c} \asymp \lceil \epsilon^{\frac{2p}{1-p}} \rceil$, $B_{t,J} \asymp \lceil \epsilon^{\frac{p}{1-p}} \rceil$, $S_{t,f} \asymp \lceil \epsilon^{\frac{1}{1-p}} \rceil$, $S_{t,c} \asymp \lceil \epsilon^{\frac{p+1}{1-p}} \rceil$, $S_{t,J} \asymp \lceil \epsilon^{\frac{1}{1-p}} \rceil$, $\tau_t \asymp \lceil \epsilon^{-1} \rceil$. Then, Algorithm 2 returns an expected ϵ -KKT point after $T \asymp \lceil \epsilon^{-1} \rceil$ iterations, corresponding to $\mathcal{O}(\epsilon^{-2})$ inner iterations. The sample complexity for computing $\tilde{\nabla} f_t$, $\tilde{\nabla} c_t$ and $\tilde{c}_t$ are of order $\mathcal{O}(\epsilon^{\frac{2p-1}{1-p}})$, $\mathcal{O}(\epsilon^{\frac{2p-1}{1-p}})$ and $\mathcal{O}(\epsilon^{\frac{3p-1}{1-p}})$, respectively.

Proof. Under the parameter setting of (1), it follows from the mini-batch estimation that $E_{t,f} = \mathcal{O}(\epsilon)$, $E_{t,c} = \mathcal{O}(\epsilon^2)$, $E_{t,J} = \mathcal{O}(\epsilon)$ and $K_T \asymp \lceil \epsilon T \rceil$. Based on the definition $\Delta_{T,3}$ in Theorem 4, we further obtain

$$\Delta_{T,3} = \mathcal{O} \left(\frac{1 + \sum_{t=1}^T \kappa_t^2 + \sum_{t=1}^T E_{t,c}}{K_T} + \frac{\sum_{t=1}^T \kappa_t (E_{t,f} + E_{t,c} + E_{t,J})}{K_T} \right) = \mathcal{O} \left(\frac{1}{\epsilon T} + \epsilon^2 + \epsilon \right).$$

Combining with $T \asymp \lceil \epsilon^{-2} \rceil$ yields the desired $\Delta_{T,3} = \mathcal{O}(\epsilon)$. Accordingly, the total number of samples over all T iterations are given by

$$\sum_{t=1}^T B_{t,f} = \mathcal{O} \left(\epsilon^{\frac{p}{1-p}} T \right) = \mathcal{O} \left(\epsilon^{\frac{3p-2}{1-p}} \right), \quad \sum_{t=1}^T B_{t,c} = \mathcal{O} \left(\epsilon^{\frac{2p}{1-p}} T \right) = \mathcal{O} \left(\epsilon^{\frac{4p-2}{1-p}} \right), \quad \sum_{t=1}^T B_{t,J} = \mathcal{O} \left(\epsilon^{\frac{p}{1-p}} T \right) = \mathcal{O} \left(\epsilon^{\frac{3p-2}{1-p}} \right),$$

Combining the parameter setting of (2) with the results in SPIDER-type recursive estimation gives $E_{t,f} = \mathcal{O}(1)$, $E_{t,c} = \mathcal{O}(\epsilon)$, $E_{t,J} = \mathcal{O}(1)$ and $K_T \asymp T$. Recalling the definition of $\Delta_{T,3}$ in Theorem 4 with inner recursive iteration, it holds that

$$\Delta_{T,3} = \mathcal{O} \left(\frac{1 + \sum_{t=1}^T \tau_t \kappa_t^2 + \sum_{t=1}^T E_{t,c}}{K_T} + \frac{\sum_{t=1}^T \kappa_t (E_{t,f} + E_{t,c} + E_{t,J})}{K_T} \right) = \mathcal{O} \left(\frac{1}{T} + \epsilon^2 + \epsilon \right).$$

Together with $T \asymp \lceil \epsilon^{-1} \rceil$, this yields the desired bound $\Delta_{T,3} = \mathcal{O}(\epsilon)$. Consequently, the sample complexity across all T iterations are given by

$$\begin{aligned} \sum_{t=1}^T (B_{t,f} + 2(\tau_t - 1)S_{t,f}) &= \mathcal{O} \left(\left(\epsilon^{\frac{p}{1-p}} + \epsilon^{\frac{p}{1-p}} \right) T \right) = \mathcal{O} \left(\epsilon^{\frac{2p-1}{1-p}} \right), \\ \sum_{t=1}^T (B_{t,c} + 2(\tau_t - 1)S_{t,c}) &= \mathcal{O} \left(\left(\epsilon^{\frac{2p}{1-p}} + \epsilon^{\frac{2p}{1-p}} \right) T \right) = \mathcal{O} \left(\epsilon^{\frac{3p-1}{1-p}} \right), \\ \sum_{t=1}^T (B_{t,J} + 2(\tau_t - 1)S_{t,J}) &= \mathcal{O} \left(\left(\epsilon^{\frac{p}{1-p}} + \epsilon^{\frac{p}{1-p}} \right) T \right) = \mathcal{O} \left(\epsilon^{\frac{2p-1}{1-p}} \right), \end{aligned}$$

which completes the proof. $\square$

References

- [1] Li, X., Joe, H., Genest, C.: A factor-copula latent-vine time series model for extreme flood insurance losses. *Journal of the American Statistical Association* **121**(553), 149–162 (2026)
- [2] Saito, Y., Aihara, S., Matsutani, M., Narita, Y.: Open Bandit Dataset and Pipeline: Towards realistic and reproducible off-policy evaluation. In: *Proceedings of the Neural Information Processing Systems Track on Datasets and Benchmarks*, vol. 1 (2021)
- [3] Vehtari, A., Simpson, D., Gelman, A., Yao, Y., Gabry, J.: Pareto smoothed importance sampling. *Journal of Machine Learning Research* **25**(72), 1–58 (2024)
- [4] Blanchet, J., Zhang, F., Zwart, B.: Efficient scenario generation for heavy-tailed chance constrained optimization. *Stochastic Systems* **14**(1), 22–46 (2024)
- [5] Jakovetic, D., Vukovic, M., Bajovic, D., Sahu, A.K., Kar, S.: Distributed recursive estimation under heavy-tail communication noise. *SIAM Journal on Control and Optimization* **61**(3), 1582–1609 (2023)
- [6] Boob, D., Deng, Q., Lan, G.: Stochastic first-order methods for convex and nonconvex functional constrained optimization. *Mathematical Programming* **197**(1), 215–279 (2023)
- [7] Boob, D., Deng, Q., Lan, G.: Level constrained first order methods for function constrained optimization. *Mathematical Programming* **209**(1), 1–61 (2025)
- [8] Na, S., Anitescu, M., Kolar, M.: Inequality constrained stochastic nonlinear optimization via active-set sequential quadratic programming. *Mathematical Programming* **202**(1), 279–353 (2023)
- [9] Shi, Q., Wang, X., Wang, H.: A momentum-based linearized augmented lagrangian method for nonconvex constrained stochastic optimization. *Mathematics of Operations Research* **51**(1), 92–133 (2026)
- [10] Zuo, S., Wang, X., Wang, H.: An adaptive single-loop stochastic penalty method for nonconvex constrained stochastic optimization. *SIAM Journal on Optimization*, to appear (2026)
- [11] Curtis, F.E., Robinson, D.P., Zhou, B.: Sequential quadratic optimization for stochastic optimization with deterministic nonlinear inequality and equality constraints. *SIAM Journal on Optimization* **34**(4), 3592–3622 (2024)
- [12] Zhang, L., Zhang, Y., Xiao, X., Wu, J.: Stochastic approximation proximal method of multipliers for convex stochastic programming. *Mathematics of Operations Research* **48**(1), 177–193 (2023)
- [13] Yan, Y., Xu, Y.: Adaptive primal-dual stochastic gradient method for expectation-constrained convex stochastic programs. *Mathematical Programming Computation* **14**(2), 319–363 (2022)
- [14] Liu, W., Xu, Y.: A SPIDER-type stochastic subgradient method for expectation-constrained nonconvex nonsmooth optimization. *SIAM Journal on Optimization* **36**(2), 1125–1153 (2026)

- [15] Shi, Q., Wang, X.: Adaptive directional decomposition methods for nonconvex constrained optimization. arXiv preprint arXiv:2511.03210 (2025)
- [16] Zhang, J., Karimireddy, S.P., Veit, A., Kim, S., Reddi, S., Kumar, S., Sra, S.: Why are adaptive methods good for attention models? In: Advances in Neural Information Processing Systems, vol. 33, pp. 15383–15393 (2020)
- [17] Cutkosky, A., Mehta, H.: High-probability bounds for non-convex stochastic optimization with heavy tails. In: Advances in Neural Information Processing Systems, vol. 34 (2021)
- [18] Liu, Z., Zhang, J., Zhou, Z.: Breaking the lower bound with (little) structure: Acceleration in non-convex stochastic optimization with heavy-tailed noise. In: Annual Conference on Learning Theory, pp. 2266–2290 (2023)
- [19] Hübler, F., Fatkhullin, I., He, N.: From gradient clipping to normalization for heavy tailed SGD. In: Proceedings of the 28th International Conference on Artificial Intelligence and Statistics. Proceedings of Machine Learning Research, vol. 258, pp. 2413–2421. PMLR, Mai Khao, Thailand (2025)
- [20] Liu, Z., Zhou, Z.: Nonconvex stochastic optimization under heavy-tailed noises: Optimal convergence without gradient clipping. In: International Conference on Learning Representations, vol. 2025, pp. 92529–92554 (2025)
- [21] Sun, T., Liu, X., Yuan, K.: Revisiting gradient normalization and clipping for nonconvex SGD under heavy-tailed noise: Necessity, sufficiency, and acceleration. Journal of Machine Learning Research **26**(237), 1–42 (2025)
- [22] He, C., Lu, Z., Sun, D., Deng, Z.: Complexity of normalized stochastic first-order methods with momentum under heavy-tailed noise. arXiv preprint arXiv:2506.11214 (2025)
- [23] Fradin, A., Sadiev, A., Condat, L., Richtárik, P.: Tight lower bounds and optimal algorithms for stochastic nonconvex optimization with heavy-tailed noise. arXiv preprint arXiv:2512.18713 (2025)
- [24] Fatkhullin, I., Hübler, F., Lan, G.: Can SGD handle heavy-tailed noise? arXiv preprint arXiv:2508.04860 (2025)
- [25] Han, C., Wang, X., Xu, P., Zhang, J.: Non-convex stochastic compositional optimization under heavy-tailed noise. <https://optimization-online.org/?p=33308> (2026)
- [26] Fang, Y., Lavaei, J., Na, S.: High probability complexity bounds of trust-region stochastic sequential quadratic programming with heavy-tailed noise. Mathematical Programming, 1–73 (2026)
- [27] Wang, H., Fang, Y., Na, S.: TR-SSQP: A trust-region method for constrained stochastic optimization under heavy-tailed noise. arXiv preprint arXiv:2609.30732 (2026)
- [28] Berahas, A.S., Curtis, F.E., O’Neill, M.J., Robinson, D.P.: A stochastic sequential quadratic optimization algorithm for nonlinear-equality-constrained optimization with rank-deficient Jacobians. Mathematics of Operations Research **49**(4), 2212–2248 (2024)
- [29] Wang, Q., Piermarini, C., Zhu, Y., Curtis, F.E.: Projected stochastic momentum methods for nonlinear equality-constrained optimization for machine learning. arXiv preprint arXiv:2601.11795 (2026)
- [30] Curtis, F.E., O’Neill, M.J., Robinson, D.P.: Worst-case complexity of an SQP method for nonlinear equality constrained stochastic optimization. Mathematical Programming **205**(1), 431–483 (2024)
- [31] Cui, Y., Shi, Q., Wang, X., Xiao, X.: An exact penalty method for stochastic equality-constrained optimization. <https://optimization-online.org/?p=32429> (2025)
- [32] Shen, H., Zeng, Y., Zhou, B.: Sequential quadratic optimization for solving expectation equality constrained stochastic optimization problems. arXiv preprint arXiv:2503.09490v4 (2026)

- [33] Cui, Y., Shi, Q., Wang, X., Xiao, X.: A fletcher’s augmented lagrangian-based stochastic first-order method for nonconvex equality-constrained optimization. arXiv preprint arXiv:2606.28230 (2026)
- [34] Berahas, A.S., Curtis, F.E., Robinson, D.P., Zhou, B.: Sequential quadratic optimization for nonlinear equality constrained stochastic optimization. *SIAM Journal on Optimization* **31**(2), 1352–1379 (2021)
- [35] O’Neill, M.J., Tang, A.: Complexity of an inexact stochastic SQP algorithm for equality constrained optimization. arXiv preprint arXiv:2604.14351 (2026)
- [36] Fang, Y., Na, S., Mahoney, M.W., Kolar, M.: Fully stochastic trust-region sequential quadratic programming for equality-constrained optimization problems. *SIAM Journal on Optimization* **34**(2), 2007–2037 (2024)
- [37] Gratton, S., Toint, P.L.: An objective-function-free algorithm for nonconvex stochastic optimization with deterministic equality and inequality constraints. arXiv preprint arXiv:2603.29685 (2026)
- [38] Rodomanov, A., Nesterov, Y.: Smoothness parameter of power of Euclidean norm. *Journal of Optimization Theory and Applications* **185**(2), 303–326 (2020)
- [39] Howard, S.R., Ramdas, A., McAuliffe, J., Sekhon, J.: Time-uniform Chernoff bounds via nonnegative supermartingales. *Probability Surveys* **17**, 257–317 (2020)
- [40] Alacaoglu, A., Wright, S.J.: Complexity of single loop algorithms for nonlinear programming with stochastic objective and constraints. In: *International Conference on Artificial Intelligence and Statistics*, pp. 4627–4635 (2024). PMLR
- [41] Li, Z., Chen, P.-Y., Liu, S., Lu, S., Xu, Y.: Stochastic inexact augmented lagrangian method for nonconvex expectation constrained optimization. *Computational Optimization and Applications* **87**(1), 117–147 (2024)
- [42] Robinson, S.M.: Strongly regular generalized equations. *Mathematics of Operations Research* **5**(1), 43–62 (1980)