A two-phase stochastic momentum-based algorithm for nonconvex expectation-constrained optimization*

Yawen Cui† Xiao Wang‡ Xiantao Xiao§

May 16, 2024

Abstract

In this paper we focus on nonconvex optimization problems with expectation constraints. To address the challenges posed by possibly nonconvex constraints and the stochastic nature of the problem, we propose a two-phase stochastic momentum-based algorithm TStoM. The first phase of TStoM aims to minimize the infeasibility measure searching for a nearly feasible point in the expectation sense. This point is used to initialize the second phase. In each iteration of the second phase, we perform a proximal stochastic gradient step to update the primal variable, while the dual update relies on stochastic constraint function values calculated in a moving average way. Under certain conditions, TStoM can find a stochastic ϵ-stationary point with a sample complexity in order $\mathcal{O}(\epsilon^{-6})$. Furthermore, under a nonsingularity condition we show that the sample complexity is in order $\mathcal{O}(\epsilon^{-5})$ to reach a stochastic ϵ-KKT point. At this point the expected error of approximate constraint values is bounded by $\mathcal{O}(I^{-1/5})$ with I being the number of samples generated during the algorithmic process. Numerical experiments are conducted to demonstrate the efficiency and effectiveness of TStoM.

Keywords: Nonconvex optimization, expectation constraints, stochastic approximation, momentum, sample complexity

MSCcodes: 90C26, 90C15, 90C30, 62L20

1 Introduction

In this paper, we consider the nonconvex constrained optimization problem

$$\begin{aligned} \min_{x\in X} \quad & \{f_0(x) := f(x) + h(x) \ \text{ s.t. } \ \mathbf{c}(x) = \mathbf{0}\}, \\ \text{with} \quad & f(x) = \mathbb{E}_\xi[F(x;\xi)],\ \mathbf{c}(x) = \mathbb{E}_\xi[\mathbf{C}(x;\xi)]\,, \end{aligned} \tag{1.1}$$

where $X \subseteq \mathbb{R}^n$ is a closed convex set, ξ is a random variable in the probability space Ξ and independent of x, and $\mathbb{E}_\xi$ refers to the expectation taken with respect to ξ. Here $F : \mathbb{R}^n \times \Xi \to \mathbb{R}$ and $\mathbf{C} : \mathbb{R}^n \times \Xi \to \mathbb{R}^m$ are continuously differentiable with respect to x but possibly nonconvex, and $h : \mathbb{R}^n \to \mathbb{R}$ is a proper lower-semicontinuous and convex function. We assume that the feasible set of (1.1) is nonempty. Although only equality constraints appear in formulation (1.1), it actually covers more general problems. For problems with inequality constraints, we can simply introduce auxiliary

*This work was funded by the National Natural Science Foundation of China (Nos. 12271076 and 12271278) and the Major Key Project of PCL (No. PCL2022A05).

†School of Mathematical Sciences, Dalian University of Technology, Dalian, China. (ywcui@mail.dlut.edu.cn)

‡Pengcheng Laboratory, Shenzhen, China. (wangx07@pcl.ac.cn)

§School of Mathematical Sciences, Dalian University of Technology, Dalian, China. (xtxiao@dlut.edu.cn)

variables and reformulate the problem into form (1.1). Problem (1.1) arises in many application fields, including risk averse machine learning [19, 27], Neyman-Pearson classification [25, 26], the fairness constrained problems [12, 22], physics-informed neural networks [18] and churn rate constrained problems [14].

Extensive research efforts have been dedicated to investigating numerical algorithms for stochastic optimization with functional constraints. In the context of general convex stochastic optimization with deterministic constraints, the exact information of constraints is accessible while only stochastic oracles of the objective function are available. Xu et al. [32] introduce a single-loop primal-dual stochastic gradient method by leveraging the linearized augmented Lagrangian (AL) function and investigate the convergence rate of the algorithm. Bollapragada et al. [4] consider problems with linear deterministic constraints and propose a double-loop augmented Lagrangian method (ALM) with an adaptive sampling strategy incorporated to control the accuracy of stochastic gradients. For nonconvex stochastic optimization with deterministic constraints, Berahas et al. [3] propose a line-search stochastic sequential quadratic programming (SQP) algorithm with adaptive Lipschitz constant estimates and analyze the algorithm's global convergence properties. Na et al. [23] present an active-set stochastic SQP algorithm for problems with both equality and inequality constraints. They utilize a differentiable exact AL function as the merit function and establish almost sure global convergence. For the same class of problems, Na et al. [24] develop a fully online stochastic SQP method and analyze the iteration complexity required to achieve ϵ-stationarity. Curtis et al. [9] propose a stochastic SQP algorithm that incorporates an adaptive strategy to update merit parameters. Under a strong linear independence constraint qualification (LICQ) condition, they prove the worst-case iteration complexity bound.

Penalty methods have also been studied for nonconvex stochastic optimization with deterministic constraints. Wang et al. [31] introduce a penalty method that minimizes an exact penalty function at each iteration using only stochastic first-order or zeroth-order information. The worst-case complexity in terms of calls to first- and zeroth-order oracles to find an ϵ-stochastic critical point is investigated. In the work by Jin and Wang [17], a class of constrained optimization problems with objective functions composed of two expectation functions is studied. Under a nonsingularity condition, the proposed stochastic nested primal-dual algorithm can find an ϵ-stationary point after $\mathcal{O}(\epsilon^{-4})$ iterations with a sample complexity bounded by $\mathcal{O}(\epsilon^{-6})$. If the algorithm starts from a feasible point, the iteration and sample complexities are reduced to $\mathcal{O}(\epsilon^{-3})$ and $\mathcal{O}(\epsilon^{-5})$, respectively. Shi et al. [30] investigate linearized ALM based on momentum [10] and analyze its global convergence properties. Besides, the sample complexity of the algorithm, which generates an ϵ-stationary point and an ϵ-KKT point under a mean-squared smoothness condition and a constraint qualification condition, is bounded by $\mathcal{O}(\epsilon^{-5})$. When the initial point is nearly feasible, the complexities can also be reduced, to the order of $\mathcal{O}(\epsilon^{-4})$. Another two related work focuses on feasible methods for inequality constrained problems. In [6] the proposed proximal point method transforms the original problem into a sequence of convex subproblems by introducing quadratic terms to objective function and to constraints, with each subproblem solved by a constraint extrapolation (ConEx) approach. Then an inexact proximal point method is presented for inequality constrained stochastic optimization. Under a strong feasibility condition, the iteration complexity of ConEx to achieve an approximate KKT point is established. The subsequent work [7] introduces a proximal gradient method, which incorporates increasing constraint level parameters for each subproblem and is extended to solve programs in stochastic settings.

Addressing problems involving stochasticity in functional constraints like (1.1) is particularly challenging. For convex programs, Yan et al. [34] provide a primal-dual stochastic gradient method and adopts an adaptive scheme to update the primal and dual variables. Convergence rates in terms of objective error and constraint violation are investigated. Zhang et al. [35] present a stochastic augmented Lagrangian-type algorithm that achieves $\mathcal{O}(K^{-1/2})$ expected convergence rates for both objective reduction and constraint violation, with K representing the number of iterations. A similar result is obtained using the proximal method of multipliers proposed in [36]. Regarding nonconvex programs, inexact quadratically regularized constrained methods, proposed by Ma et al. [22], aim for inequality

constrained optimization whose objective and constraint functions are weakly convex. These methods are feasible in the way that they transform the initial problem into a sequence of strongly convex subproblems by adding quadratic terms to objective and to constraints as well. Under a uniform Slater's condition the complexities for finding a nearly ϵ-stationary point are investigated. The aforementioned work [6] also analyzes the complexity of finding an approximate KKT point for expectation constrained optimization merely with inequality constraints in fully-stochastic case. Jin and Wang [16] introduce a stochastic primal-dual (SPD) method for a class of nonconvex optimization with a large number of inequality constraints. To ease the computational burden caused by simultaneously evaluating all the constraints at a single point, SPD randomly selects a small number of constraints at each iteration to construct a stochastic approximation to the linearized AL function. When the initial point is nearly feasible, the iteration and sample complexities can be further reduced. Another two works that are closely related to ours are [20] and [1]. Li et al. [20] employ the standard inexact ALM framework and propose stochastic inexact ALMs (Stoc-iALM), which incorporate subroutines using momentum-based variance-reduced proximal stochastic gradient approaches. They establish an sample complexity in order $\mathcal{O}(\epsilon^{-5})$ to find an ϵ-KKT point. It is noteworthy that their method involves a double-loop structure with intricate subproblems. Alacaoglu and Wright [1] consider single-loop momentum-based algorithms for smooth equality constrained optimization. In particular, under a nonsingularity condition (refer to Assumption 4.1 below), [1] establishes the $\tilde{\mathcal{O}}(\epsilon^{-5})$-sample complexity to find a point satisfying ϵ-approximate first-order conditions (refer to the stochastic ϵ-KKT point defined below).

1.1 Contributions

In this paper, we propose a two-phase stochastic momentum-based algorithm, TStoM, for nonconvex expectation-constrained optimization. The first phase involves applying a stochastic momentum-based approach to solve the infeasibility minimization problem, pursuing a nearly feasible point (in expectation sense) which initializes the next phase. In the second phase we adopt a single-loop approach trying to reduce the criticality measure. To update primal variables we compute stochastic gradients of the linearized augmented Lagrangian function based on momentum. This enables us to construct much simpler subproblems, facilitating the optimization process. For the dual update, we utilize a moving average strategy to obtain stochastic approximations of constraint function values. This strategy helps us effectively estimate the constraints and incorporate them into the optimization process. Under certain conditions, the sample complexity to find a stochastic ϵ-stationary point is bounded by $\mathcal{O}(\epsilon^{-6})$. By assuming a nonsingularity condition, we show that the sample complexity of the algorithm to reach a stochastic ϵ-KKT point is in order $\mathcal{O}(\epsilon^{-5})$. At this point the expected error of the approximate constraint value is in order $O(I^{-1/5})$ with I representing the number of samples. We also conduct numerical experiments on solving three instances: MIMO transmit signal design with imperfect channel state information, the multi-class Neyman-Pearson classification problem, and a chance constrained program. Numerical results demonstrate promising performances of the proposed algorithm.

1.2 Notation and preliminaries

Without any specification, $\|\cdot\|$ denotes the Euclidean norm. The distance between $X, Y \subseteq \mathbb{R}^n$ is defined as $\mathbf{d}(X,Y) = \inf_{x\in X, y\in Y} \|x-y\|$. We define $\nabla \mathbf{c}(x) = (\nabla c_1(x), \nabla c_2(x), \ldots, \nabla c_m(x))$, $\nabla \mathbf{C}(x;\xi) = (\nabla C_1(x;\xi), \nabla C_2(x;\xi), \ldots, \nabla C_m(x;\xi))$ and $[k] = \{1,\ldots,k\}$ for a positive integer k. $\mathbb{E}[\cdot]$ refers to the full expectation taken with respect to all random variables generated during an algorithmic process. The subgradient set of h at x is defined as $\partial h(x) = \{v \in \mathbb{R}^n \mid h(y) \geq h(x) + \langle v, y-x\rangle, \forall y \in \mathbb{R}^n\}$. The normal cone to X at $\bar{x} \in X$ is given by $\mathcal{N}_X(\bar{x}) = \{v \mid \langle v, x-\bar{x}\rangle \leq 0,\ \forall x \in X\}$.

In general, finding a global or even a local minimizer of nonconvex constrained optimization is NP-hard. Therefore, our primary focus is to pursue a more trackable point, a KKT point. A point

$x^* \in X$ is called a KKT point of (1.1), if there exists $\lambda^* \in \mathbb{R}^m$, such that

$$\mathbf{d}(\nabla f(x^*) + \partial h(x^*) - \nabla \mathbf{c}(x^*)\lambda^*, -\mathcal{N}_X(x^*)) = 0 \quad \text{and} \quad \mathbf{c}(x^*) = \mathbf{0}.$$

However, in the course of an algorithmic process, it is inevitable that the iteration might be stuck around an infeasible stationary point which is a solution to the problem:

$$\min_{x \in X} \quad \frac{1}{2}\|\mathbf{c}(x)\|^2. \tag{1.2}$$

We next give definitions of approximate solutions of problem (1.1).

DEFINITION **1.1.** *Given $\epsilon > 0$, we call $x \in X$ an ϵ-KKT point of (1.1), if there exists $\lambda \in \mathbb{R}^m$ such that*

$$\mathbf{d}(\nabla f(x) + \partial h(x) - \nabla \mathbf{c}(x)\lambda, -\mathcal{N}_X(x)) \le \epsilon \ \text{ and } \ \|\mathbf{c}(x)\| \le \epsilon. \tag{1.3}$$

A point $x \in X$ is called an ϵ-stationary point of (1.1), if there exists $\lambda \in \mathbb{R}^m$, such that

$$\mathbf{d}(\nabla f(x) + \partial h(x) - \nabla \mathbf{c}(x)\lambda, -\mathcal{N}_X(x)) \le \epsilon \ \text{ and } \ \mathbf{d}(\nabla \mathbf{c}(x)\mathbf{c}(x), -\mathcal{N}_X(x)) \le \epsilon. \tag{1.4}$$

We call $x \in X$ a stochastic ϵ-KKT point (resp. stochastic ϵ-stationary point) of (1.1), if (1.3) (resp. (1.4)) holds in expectation.

Next, we lay out assumptions that are utilized throughout the rest of this paper.

Assumption 1.1. *The set X is closed and convex. The objective function value of (1.1) over X is lower bounded by C^*. And there exist $M, G > 0$ such that $\|\mathbf{C}(x;\xi)\| \le M$ for all $\xi \in \Xi$,*

$$\|\nabla f(x)\| \le G \ \text{ and } \ \|v\| \le G, \ \forall v \in \partial h(x), x \in X. \tag{1.5}$$

Assumption 1.2. *$F(\cdot;\xi)$ and $\mathbf{C}(\cdot;\xi)$ are differentiable over X almost surely for any $\xi \in \Xi$, and there exist σ_f, $\sigma_c > 0$ such that for any $x \in X$,*

$$\begin{aligned}
&\mathbb{E}_\xi[\nabla F(x;\xi)] = \nabla f(x), \quad &&\mathbb{E}_\xi[\|\nabla F(x;\xi) - \nabla f(x)\|^2] \le \sigma_f^2,\\
&\mathbb{E}_\xi[\nabla \mathbf{C}(x;\xi)] = \nabla \mathbf{c}(x), \quad &&\mathbb{E}_\xi[\|\nabla \mathbf{C}(x;\xi) - \nabla \mathbf{c}(x)\|^2] \le \sigma_c^2,\\
&\mathbb{E}_\xi[\mathbf{C}(x;\xi)] = \mathbf{c}(x), \quad &&\mathbb{E}_\xi[\|\mathbf{C}(x;\xi) - \mathbf{c}(x)\|^2] \le \sigma_c^2.
\end{aligned}$$

Assumption 1.3. *For any $x, y \in X$, there exists $L > 0$ such that*

$$\mathbb{E}_\xi[\|\nabla F(x;\xi) - \nabla F(y;\xi)\|^2] \le L^2\|x-y\|^2, \quad \mathbb{E}_\xi[\|\nabla \mathbf{C}(x;\xi) - \nabla \mathbf{C}(y;\xi)\|^2] \le L^2\|x-y\|^2,$$

and $\mathbb{E}_\xi[\|\mathbf{C}(x;\xi) - \mathbf{C}(y;\xi)\|^2] \le G^2\|x-y\|^2$.

It is worthy to note that Assumption 1.3 also refers to the mean-squared smoothness assumption, which is widely used in works on stochastic approximation methods, such as [2, 20, 30]. By Jensen's inequality, Assumption 1.3 implies that

$$\|\nabla f(x) - \nabla f(y)\| \le L\|x-y\|, \ \|\nabla \mathbf{c}(x) - \nabla \mathbf{c}(y)\| \le L\|x-y\| \text{ and } \|\mathbf{c}(x) - \mathbf{c}(y)\| \le G\|x-y\|, \tag{1.6}$$

which indicates that $\|\nabla \mathbf{c}(x)\| \le G$. And it follows from Assumption 1.1 and Assumption 1.2 that

$$\|\mathbf{c}(x)\| \le M, \quad \mathbb{E}_\xi[\|\nabla \mathbf{C}(x;\xi)\|^2] \le G_c^2 := 2G^2 + 2\sigma_c^2. \tag{1.7}$$

In addition, with ξ_1 and ξ_2 being independent, Assumptions 1.1-1.3 imply that for any $x, y \in X$,

$$\begin{aligned}
&\mathbb{E}_{\xi_1,\xi_2}\left[\|\nabla \mathbf{C}(x;\xi_1)\mathbf{C}(x;\xi_2) - \nabla \mathbf{C}(y;\xi_1)\mathbf{C}(y;\xi_2)\|^2\right]\\
&= \mathbb{E}_{\xi_1,\xi_2}\left[\|(\nabla \mathbf{C}(x;\xi_1) - \nabla \mathbf{C}(y;\xi_1))\mathbf{C}(x;\xi_2) + \nabla \mathbf{C}(y;\xi_1)(\mathbf{C}(x;\xi_2) - \mathbf{C}(y;\xi_2))\|^2\right]
\end{aligned}$$

$$\begin{aligned}
&\leq 2\mathbb{E}_{\xi_1}\left[\|\nabla \mathbf{C}(x;\xi_1)-\nabla \mathbf{C}(y;\xi_1)\|^2\right]\mathbb{E}_{\xi_2}\left[\|\mathbf{C}(x;\xi_2)\|^2\right]\\
&\quad+2\mathbb{E}_{\xi_1}\left[\|\nabla \mathbf{C}(y;\xi_1)\|^2\right]\mathbb{E}_{\xi_2}\left[\|\mathbf{C}(x;\xi_2)-\mathbf{C}(y;\xi_2)\|^2\right]\\
&\leq L_0^2\|x-y\|^2,\ \text{where } L_0:=\sqrt{2(L^2M^2+G_c^2G^2)}. \qquad (1.8)
\end{aligned}$$

Similarly, we can derive

$$\mathbb{E}_{\xi_1,\xi_2}[\|\nabla \mathbf{C}(x;\xi_1)\mathbf{C}(x;\xi_2)-\nabla \mathbf{c}(x)\mathbf{c}(x)\|^2]\leq \sigma_J^2:=2\sigma_c^2(M^2+G_c^2). \qquad (1.9)$$

Unlike (1.7), the uniform boundedness of $\nabla \mathbf{C}(x;\xi)$ for all $\xi\in\Xi$ is assumed in [1]. Besides, [1] requires the objective function be upper bounded, which is not assumed in this paper.

1.3 Outline

The remainder of this paper is outlined as follows. In section 2, we present details of a two-phase stochastic momentum-based algorithm for (1.1). In section 3, we give the auxiliary lemmas and in section 4 we conduct a complexity analysis of the proposed algorithm towards approximate solutions. In section 5, we report numerical experimental results on solving three test examples. Finally, we draw conclusions.

2 A two-phase stochastic momentum-based algorithm for (1.1)

As studied in previous works [16, 17, 30], the (near) feasibility of the initial point can induce lower sample complexities of algorithms for nonconvex constrained optimization under certain conditions. Motivated by this, we introduce a two-phase stochastic approximation algorithm TStoM for solving (1.1), presented in Algorithm 2.1. The first phase is a feasibility pursuing phase, which aims for a nearly feasible point. By using this point as an initial guess, we delve into the second phase for an approximate solution of (1.1).

Our main strategy in Phase I of TStoM is to apply a stochastic gradient approach to minimize the infeasibility measure, i.e. solving (1.2). More specifically, given $z^1\in X$, we randomly generate i.i.d. samples $\varsigma_1^1,\varsigma_2^1$ from Ξ and compute $W^1=\nabla \mathbf{C}(z^1;\varsigma_1^1)\mathbf{C}(z^1;\varsigma_2^1)$. Then for any $t\geq 1$, we update z^{t+1} and compute stochastic gradient W^{t+1} based on momentum technique through

$$z^{t+1}=\arg\min_{z\in X}\{\langle W^t,z-z^t\rangle+\frac{V}{2}\|z-z^t\|^2\}, \qquad (2.1)$$

$$W^{t+1}=v^{t+1}+(1-\gamma_t)(W^t-u^{t+1}) \qquad (2.2)$$

with $V>0$, $\gamma_t\in(0,1)$, $v^{t+1}=\nabla \mathbf{C}(z^{t+1};\varsigma_1^{t+1})\mathbf{C}(z^{t+1};\varsigma_2^{t+1})$ and $u^{t+1}=\nabla \mathbf{C}(z^t;\varsigma_1^{t+1})\mathbf{C}(z^t;\varsigma_2^{t+1})$, where $\varsigma_1^{t+1},\varsigma_2^{t+1}$ are i.i.d. samples from Ξ. After T iterations, we randomly choose an iterate from $z^{t+1},t\in[T]$ as the output of Phase I and set it as the initial point for Phase II.

The aim of Phase II is to pursue an approximate solution of (1.1). To proceed, let us first recall the augmented Lagrangian function associated with (1.1). It can be expressed as $\mathcal{L}_\beta(x,\lambda)=f(x)-\lambda^{\mathsf{T}}\mathbf{c}(x)+\frac{\beta}{2}\|\mathbf{c}(x)\|^2+h(x)$, where $\beta>0$ is a penalty parameter and $\lambda\in\mathbb{R}^m$. We define

$$\psi_\beta(x,\lambda):=-\lambda^{\mathsf{T}}\mathbf{c}(x)+\frac{\beta}{2}\|\mathbf{c}(x)\|^2\ \text{ and }\ \mathscr{D}_\beta(x,\lambda)=f(x)+\psi_\beta(x,\lambda)$$

for simplicity. Due to the stochastic nature of problem (1.1), the exact function information of f and $\mathbf{c}$ are difficult to obtain. We thus generate i.i.d. samples ξ, ζ_1, ζ_2 from Ξ and define

$$\mathcal{G}_\beta(x,\lambda;\xi,\zeta):=\nabla F(x;\xi)+\bar{\mathcal{G}}_\beta(x,\lambda;\zeta), \qquad (2.3)$$

where $\bar{\mathcal{G}}_\beta(x,\lambda;\zeta) = -\nabla \mathbf{C}(x;\zeta_1)\lambda + \beta \nabla \mathbf{C}(x;\zeta_1)\mathbf{C}(x;\zeta_2)$ with $\zeta := (\zeta_1,\zeta_2)$. Obviously, $\bar{\mathcal{G}}_\beta(x,\lambda;\xi,\zeta)$ and $\mathcal{G}_\beta(x,\lambda;\xi,\zeta)$ are unbiased estimates of $\nabla_x \psi_\beta(x,\lambda)$ and $\nabla_x \mathscr{D}_\beta(x,\lambda)$, respectively. We then compute an approximate gradient d^k through

$$d^k = \begin{cases} \frac{1}{|\mathcal{M}|}\sum\limits_{j\in\mathcal{M}} \mathcal{G}_{\beta_k}(x^1,\lambda^1;\xi_j^1,\zeta_j^1), & k=1, \\ \mathcal{G}_{\beta_k}(x^k,\lambda^k;\xi^k,\zeta^k) + (1-\alpha_{k-1})(d^{k-1} - \mathcal{G}_{\beta_{k-1}}(x^{k-1},\lambda^{k-1};\xi^k,\zeta^k)), & k\geq 2, \end{cases} \tag{2.4}$$

where $\alpha_{k-1}\in(0,1)$, $\zeta_j^1 := (\zeta_{j,1}^1,\zeta_{j,2}^1)$, $\zeta^k := (\zeta_1^k,\zeta_2^k)$ and $\xi_j^1,\zeta_{j,1}^1,\zeta_{j,2}^1, j\in\mathcal{M}, \xi^k,\zeta_1^k,\zeta_2^k, k\geq 2$, are i.i.d samples generated from Ξ. For ease of notation, we denote $\xi^1 := \{\xi_j^1, j\in\mathcal{M}\}$, $\zeta^1 := \{(\zeta_{j,1}^1,\zeta_{j,2}^1), j\in\mathcal{M}\}$. Then the primal iterate is updated through

$$x^{k+1} = \arg\min_{x\in X}\{\langle d^k, x\rangle + \frac{1}{2\eta_k}\|x-x^k\|^2 + h(x)\}, \tag{2.5}$$

where $\eta_k > 0$. Note that in the work for deterministic constrained stochastic optimization like [16, 17, 30], the dual variable is updated by $\lambda^{k+1} = \lambda^k - \rho_k \mathbf{c}(x^{k+1})$, where $\rho_k \in (0,\beta_k]$, while the work for optimization with expectation constraints such as [1] applies $\lambda^{k+1} = \lambda^k - \rho_k \mathbf{C}(x^{k+1};\zeta)$ based on a randomly generated sample ζ. Different from previous work, we adopt a moving average way to approximate $\mathbf{c}(x^{k+1})$ based on which we compute λ^{k+1} through

$$\lambda^{k+1} = \lambda^k - \rho_k y^{k+1}, \text{ where } y^{k+1} = (1-\tau_k)y^k + \tau_k \mathbf{C}(x^{k+1};\theta^k), \quad k\geq 1, \tag{2.6}$$

with $y^1 = \mathbf{C}(x^1;\theta^0)$, and $\theta^k, k\geq 0$ being randomly and independently chosen from Ξ. Since y^{k+1} is a weighted average of the previous constraint value estimate y^k and newly updated stochastic constraint function value, it is expected to have a lower variance compared to $\mathbf{C}(x^{k+1};\theta^k)$. In particular, we will show that $\mathbb{E}[\|y^{R+1} - \mathbf{c}(x^{R+1})\|^2] = \mathcal{O}(K^{-2/5})$, as demonstrated in Proposition 4.1.

Algorithm 2.1 Two-phase Stochastic Momentum-based algorithm (TStoM)

Input: Initial point z^1 and dual point $\lambda^1 = \mathbf{0}$, positive integers T and K, parameters $\gamma_t \in (0,1)$ for $t\in[T]$, $V>0$, and $\beta_k>0$, $\eta_k>0$, $\rho_k\in(0,\beta_k]$, $\alpha_k\in(0,1)$, $\tau_k\in(0,1)$ for $k\in[K]$.

Output: x^{R+1} where $R\in[K]$ is uniformly chosen at random.

Phase I

Generate i.i.d. samples $\varsigma_1^1,\varsigma_2^1$ from Ξ and compute $W^1 = \nabla \mathbf{C}(z^1;\varsigma_1^1)\mathbf{C}(z^1;\varsigma_2^1)$.
for $t=1,\ldots,T$ **do**
  Compute z^{t+1} through (2.1).
  Compute W^{t+1} through (2.2).
end for
Set $x^1 = z^{R_0+1}$ with R_0 selected from $[T]$ uniformly at random.

Phase II

for $k=1,\ldots,K$ **do**
  Compute d^k through (2.4).
  Compute x^{k+1} through (2.5).
  Compute λ^{k+1} through (2.6).
end for

3 Auxiliary lemmas

In this section, we will present auxiliary lemmas which are useful for the forthcoming sample complexity analysis of TStoM.

LEMMA **3.1.** *Under Assumption 1.1, it holds that $\|y^k\|\leq M$ for any $k\geq 1$.*

Proof. We show the result by induction. The result holds obviously for $k = 1$. Assume $\|y^k\| \leq M$, then it derives from Assumption 1.1 and (2.6) that $\|y^{k+1}\| = \|(1-\tau_k)y^k + \tau_k \mathbf{C}(x^{k+1};\theta^k)\| \leq M$. □

LEMMA **3.2.** *Under Assumption 1.1, it holds that for any $k \in [K]$, $\|\lambda^{k+1} - \lambda^k\| \leq \rho_k M$ and $\|\lambda^k\| \leq M\sum_{t=1}^{k-1}\rho_t$.*

Proof. Firstly, it follows from Lemma 3.1 that, $\|\lambda^{k+1} - \lambda^k\| = \rho_k\|y^{k+1}\| \leq \rho_k M$. It further yields $\|\lambda^k\| \leq M\sum_{t=1}^{k-1}\rho_t$ by $\|\lambda^k\| = \|\lambda^k - \lambda^1\| \leq \sum_{t=1}^{k-1}\|\lambda^{t+1} - \lambda^t\|$. □

LEMMA **3.3.** *Suppose that Assumptions 1.1-1.3 hold. Then for any $k \geq 1$, we have*

$$\mathbb{E}[\|\nabla_x \mathscr{D}_{\beta_k}(x^{k+1},\lambda^k) - \nabla_x \mathscr{D}_{\beta_k}(x^k,\lambda^k)\|^2] \leq L_{\beta_k}^2 \mathbb{E}[\|x^{k+1} - x^k\|^2], \tag{3.1}$$

$$\mathbb{E}[\|\mathcal{G}_{\beta_k}(x^{k+1},\lambda^k;\xi^{k+1},\zeta^{k+1}) - \mathcal{G}_{\beta_k}(x^k,\lambda^k;\xi^{k+1},\zeta^{k+1})\|^2] \leq L_{\beta_k}^2 \mathbb{E}[\|x^{k+1} - x^k\|^2], \tag{3.2}$$

$$\mathbb{E}[\|\mathcal{G}_{\beta_k}(x^k,\lambda^k;\xi^k,\zeta^k) - \nabla_x \mathscr{D}_{\beta_k}(x^k,\lambda^k)\|^2] \leq \sigma_{\beta_k}^2, \tag{3.3}$$

where $L_{\beta_k} := \beta_k\tilde{L}$ with $\tilde{L} := \max\{\sqrt{2}(G^2 + ML), \sqrt{2}L_0\} + \beta_k^{-1}(L + \sqrt{2}ML\sum_{t=1}^{k-1}\rho_t)$, and $\sigma_{\beta_k} := \sigma_f + \sqrt{2}(\sigma_c M\sum_{t=1}^{k-1}\rho_t + \beta_k\sigma_J)$.

Proof. It follows from the definition of $\psi_\beta(x,\lambda)$, (1.6)-(1.7) and Lemma 3.2 that

$$\begin{aligned}
&\mathbb{E}[\|\nabla_x\psi_{\beta_k}(x^{k+1},\lambda^k) - \nabla_x\psi_{\beta_k}(x^k,\lambda^k)\|] \\
&= \mathbb{E}[\|\nabla\mathbf{c}(x^{k+1})[(\beta_k\mathbf{c}(x^{k+1}) - \lambda^k) - (\beta_k\mathbf{c}(x^k) - \lambda^k)] + (\nabla\mathbf{c}(x^{k+1}) - \nabla\mathbf{c}(x^k))(\beta_k\mathbf{c}(x^k) - \lambda^k)\|] \\
&\leq \mathbb{E}[\beta_k\|\mathbf{c}(x^{k+1}) - \mathbf{c}(x^k)\|\|\nabla\mathbf{c}(x^{k+1})\| + L\|x^{k+1} - x^k\|\|\beta_k\mathbf{c}(x^k) - \lambda^k\|] \\
&\leq (\beta_k G^2 + \beta_k ML + ML\sum_{t=1}^{k-1}\rho_t)\mathbb{E}[\|x^{k+1} - x^k\|],
\end{aligned}$$

which further indicates

$$\begin{aligned}
&\mathbb{E}[\|\nabla_x \mathscr{D}_{\beta_k}(x^{k+1},\lambda^k) - \nabla_x \mathscr{D}_{\beta_k}(x^k,\lambda^k)\|^2] \\
&\leq \mathbb{E}[\|\nabla f(x^{k+1}) - \nabla f(x^k)\|^2] + 2\mathbb{E}[\|\nabla f(x^{k+1}) - \nabla f(x^k)\|\|\nabla_x\psi_{\beta_k}(x^{k+1},\lambda^k) - \nabla_x\psi_{\beta_k}(x^k,\lambda^k)\|] \\
&\quad + \mathbb{E}[\|\nabla_x\psi_{\beta_k}(x^{k+1},\lambda^k) - \nabla_x\psi_{\beta_k}(x^k,\lambda^k)\|^2] \\
&\leq (L + \sqrt{2}(\beta_k G^2 + \beta_k ML + ML\sum_{t=1}^{k-1}\rho_t))^2\mathbb{E}[\|x^{k+1} - x^k\|^2] \leq L_{\beta_k}^2\mathbb{E}[\|x^{k+1} - x^k\|^2].
\end{aligned}$$

Hence, (3.1) is derived. From (1.8), Assumption 1.3 and Lemma 3.2, we can obtain

$$\begin{aligned}
&\mathbb{E}_{\zeta^{k+1}}[\|\bar{\mathcal{G}}_{\beta_k}(x^{k+1},\lambda^k;\zeta^{k+1}) - \bar{\mathcal{G}}_{\beta_k}(x^k,\lambda^k;\zeta^{k+1})\|] \\
&\leq \mathbb{E}_{\zeta^{k+1}}[\beta_k\|\nabla\mathbf{C}(x^{k+1};\zeta_1^{k+1})\mathbf{C}(x^{k+1};\zeta_2^{k+1}) - \nabla\mathbf{C}(x^k;\zeta_1^{k+1})\mathbf{C}(x^k;\zeta_2^{k+1})\| + \|\lambda^k\|\|\nabla\mathbf{C}(x^k;\zeta_1^{k+1}) \\
&\quad - \nabla\mathbf{C}(x^{k+1};\zeta_1^{k+1})\|] \leq (\beta_k L_0 + ML\sum_{t=1}^{k-1}\rho_t)\|x^{k+1} - x^k\|,
\end{aligned}$$

which implies from Assumption 1.3 and Jensen's inequality that

$$\begin{aligned}
&\mathbb{E}_{\xi^{k+1},\zeta^{k+1}}[\|\mathcal{G}_{\beta_k}(x^{k+1},\lambda^k;\xi^{k+1},\zeta^{k+1}) - \mathcal{G}_{\beta_k}(x^k,\lambda^k;\xi^{k+1},\zeta^{k+1})\|^2] \\
&\leq \mathbb{E}_{\xi^{k+1}}[\|\nabla F(x^{k+1};\xi^{k+1}) - \nabla F(x^k;\xi^{k+1})\|^2] + \mathbb{E}_{\zeta^{k+1}}[\|\bar{\mathcal{G}}_{\beta_k}(x^{k+1},\lambda^k;\zeta^{k+1}) - \bar{\mathcal{G}}_{\beta_k}(x^k,\lambda^k;\zeta^{k+1})\|^2] \\
&\quad + 2\mathbb{E}_{\xi^{k+1},\zeta^{k+1}}[\|\nabla F(x^{k+1};\xi^{k+1}) - \nabla F(x^k;\xi^{k+1})\|\|\bar{\mathcal{G}}_{\beta_k}(x^{k+1},\lambda^k;\zeta^{k+1}) - \bar{\mathcal{G}}_{\beta_k}(x^k,\lambda^k;\zeta^{k+1})\|] \\
&\leq (L + \sqrt{2}(\beta_k L_0 + ML\sum_{t=1}^{k-1}\rho_t))^2\|x^{k+1} - x^k\|^2.
\end{aligned}$$

It then yields (3.2) by taking full expectation on above inequality. Note that due to (1.9), Assumption 1.2 and Lemma 3.2,

$$\begin{aligned}
&\mathbb{E}_{\zeta^k}[\|\bar{\mathcal{G}}_{\beta_k}(x^k,\lambda^k;\zeta^k)-\nabla_x\psi_{\beta_k}(x^k,\lambda^k)\|^2]\\
&=\mathbb{E}_{\zeta^k}[\|-\nabla\mathbf{C}(x^k;\zeta_1^k)\lambda^k+\nabla\mathbf{c}(x^k)\lambda^k+\beta_k\nabla\mathbf{C}(x^k;\zeta_1^k)\mathbf{C}(x^k;\zeta_2^k)-\beta_k\nabla\mathbf{c}(x^k)\mathbf{c}(x^k)\|^2]\\
&\le 2\mathbb{E}_{\zeta^k}[\|\lambda^k\|^2\|\nabla\mathbf{C}(x^k;\zeta_1^k)-\nabla\mathbf{c}(x^k)\|^2]+2\beta_k^2\mathbb{E}_{\zeta^k}[\|\nabla\mathbf{C}(x^k;\zeta_1^k)\mathbf{C}(x^k;\zeta_2^k)-\nabla\mathbf{c}(x^k)\mathbf{c}(x^k)\|^2]\\
&\le 2(\sigma_c M\sum_{t=1}^{k-1}\rho_t)^2+2\beta_k^2\sigma_J^2,\quad k\ge 2.
\end{aligned}$$

Then it derives from the independence of ξ^k and ζ^k that

$$\begin{aligned}
&\mathbb{E}_{\xi^k,\zeta^k}[\|\mathcal{G}_{\beta_k}(x^k,\lambda^k;\xi^k,\zeta^k)-\nabla_x\mathscr{D}_{\beta_k}(x^k,\lambda^k)\|^2]\\
&=\mathbb{E}_{\xi^k}[\|\nabla F(x^k;\xi^k)-\nabla f(x^k)\|^2]+\mathbb{E}_{\zeta^k}[\|\bar{\mathcal{G}}_{\beta_k}(x^k,\lambda^k;\zeta^k)-\nabla_x\psi_{\beta_k}(x^k,\lambda^k)\|^2]\\
&\quad+2\mathbb{E}_{\xi^k,\zeta^k}[\langle\nabla F(x^k;\xi^k)-\nabla f(x^k),\bar{\mathcal{G}}_{\beta_k}(x^k,\lambda^k;\zeta^k)-\nabla_x\psi_{\beta_k}(x^k,\lambda^k)\rangle]\\
&\le\sigma_f^2+2(\sigma_c M\sum_{t=1}^{k-1}\rho_t)^2+2\beta_k^2\sigma_J^2.
\end{aligned}$$

By taking full expectation on above inequality, we establish (3.3) for $k\ge 2$. It is straightforward to derive (3.3) for $k=1$, since

$$\mathbb{E}[\|\mathcal{G}_{\beta_1}(x^1,\lambda^1;\xi^1,\zeta^1)-\nabla_x\mathscr{D}_{\beta_1}(x^1,\lambda^1)\|^2]\le\frac{\sigma_{\beta_1}^2}{|\mathcal{M}|}. \tag{3.4}$$

The proof is completed. □

In the following, we denote the error of d^k by $\varepsilon^k:=d^k-\nabla_x\mathscr{D}_{\beta_k}(x^k,\lambda^k)$.

LEMMA **3.4.** *Let Assumptions 1.1-1.3 be satisfied. Then it holds that for any $k\in[K]$,*

$$\begin{aligned}
&\mathbb{E}[\mathbf{d}^2(\nabla_x\mathscr{D}_{\beta_k}(x^{k+1},\lambda^{k+1})+\partial h(x^{k+1}),-\mathcal{N}_X(x^{k+1}))]\\
&\le 4(\rho_k MG)^2+4(L_{\beta_k}^2+\frac{1}{\eta_k^2})\mathbb{E}[\|x^{k+1}-x^k\|^2]+4\mathbb{E}[\|\varepsilon^k\|]^2.
\end{aligned}$$

Proof. By the definition of $\mathscr{D}_\beta(x,\lambda)$ and Lemma 3.2, we have

$$\mathbb{E}[\|\nabla_x\mathscr{D}_{\beta_k}(x^{k+1},\lambda^{k+1})-\nabla_x\mathscr{D}_{\beta_k}(x^{k+1},\lambda^k)\|^2]\le\mathbb{E}[\|\lambda^{k+1}-\lambda^k\|^2\|\nabla\mathbf{c}(x^{k+1})\|^2]\le(\rho_k MG)^2.$$

Then by optimality conditions for (2.5), i.e. $\mathbf{d}(d^k+\partial h(x^{k+1})+\frac{1}{\eta_k}(x^{k+1}-x^k),-\mathcal{N}_X(x^{k+1}))=0$, it indicates from Jensen's inequality and (3.1) that

$$\begin{aligned}
&\mathbb{E}[\mathbf{d}^2(\nabla_x\mathscr{D}_{\beta_k}(x^{k+1},\lambda^{k+1})+\partial h(x^{k+1}),-\mathcal{N}_X(x^{k+1}))]\\
&\le\mathbb{E}[\|\nabla_x\mathscr{D}_{\beta_k}(x^{k+1},\lambda^{k+1})-d^k-\frac{1}{\eta_k}(x^{k+1}-x^k)\|^2]\\
&=\mathbb{E}[\|\nabla_x\mathscr{D}_{\beta_k}(x^{k+1},\lambda^{k+1})-\nabla_x\mathscr{D}_{\beta_k}(x^{k+1},\lambda^k)+\nabla_x\mathscr{D}_{\beta_k}(x^{k+1},\lambda^k)-\nabla_x\mathscr{D}_{\beta_k}(x^k,\lambda^k)\\
&\quad+\nabla_x\mathscr{D}_{\beta_k}(x^k,\lambda^k)-d^k-\frac{1}{\eta_k}(x^{k+1}-x^k)\|^2]\\
&\le 4(\rho_k MG)^2+4L_{\beta_k}^2\mathbb{E}[\|x^{k+1}-x^k\|^2]+4\mathbb{E}[\|\varepsilon^k\|^2]+\frac{4}{\eta_k^2}\mathbb{E}[\|x^{k+1}-x^k\|^2],
\end{aligned}$$

which completes the proof. □

LEMMA **3.5.** *Assume that Assumptions 1.1-1.3 be satisfied, then for any $k \in [K]$,*

$$\begin{aligned}&(\frac{1}{2\eta_k} - \frac{L_{\beta_k}}{2})\mathbb{E}[\|x^{k+1} - x^k\|^2] \\ &\leq \mathbb{E}[\mathcal{L}_{\beta_k}(x^k, \lambda^k) - \mathcal{L}_{\beta_{k+1}}(x^{k+1}, \lambda^{k+1}) + \frac{\beta_{k+1} - \beta_k}{2}M^2 + \frac{\eta_k}{2}\|\varepsilon^k\|^2 + \rho_k M^2].\end{aligned} \tag{3.5}$$

Proof. Lemma 3.2 together with (1.7) implies that

$$\mathbb{E}[\mathcal{L}_{\beta_k}(x^{k+1}, \lambda^k)] = \mathbb{E}[\mathcal{L}_{\beta_k}(x^{k+1}, \lambda^{k+1}) - (\lambda^k - \lambda^{k+1})^{\mathsf{T}}\mathbf{c}(x^{k+1})] \geq \mathbb{E}[\mathcal{L}_{\beta_k}(x^{k+1}, \lambda^{k+1}) - \rho_k M^2]. \tag{3.6}$$

According to optimality conditions for (2.5), there exists a vector $s \in \partial h(x^{k+1})$ such that

$$\langle d^k + s + \frac{1}{\eta_k}(x^{k+1} - x^k), x - x^{k+1}\rangle \geq 0, \quad \forall x \in X.$$

Then by the convexity of h and setting $x = x^k$, we have

$$\mathbb{E}[h(x^{k+1}) - h(x^k)] \leq \mathbb{E}[\langle s, x^{k+1} - x^k\rangle] \leq -\mathbb{E}[\langle d^k + \frac{1}{\eta_k}(x^{k+1} - x^k), x^{k+1} - x^k\rangle].$$

It thus together with

$$\mathbb{E}[\mathscr{D}_{\beta_k}(x^{k+1}, \lambda^k)] \leq \mathbb{E}[\mathscr{D}_{\beta_k}(x^k, \lambda^k) + \langle \nabla_x \mathscr{D}_{\beta_k}(x^k, \lambda^k), x^{k+1} - x^k\rangle + \frac{L_{\beta_k}}{2}\|x^{k+1} - x^k\|^2]$$

indicates from Young's inequality and $\varepsilon^k = d^k - \nabla_x \mathscr{D}_{\beta_k}(x^k, \lambda^k)$ that

$$\begin{aligned}\mathbb{E}[\mathcal{L}_{\beta_k}(x^{k+1}, \lambda^k)] &\leq \mathbb{E}[\mathcal{L}_{\beta_k}(x^k, \lambda^k) + \langle \nabla_x \mathscr{D}_{\beta_k}(x^k, \lambda^k) - d^k, x^{k+1} - x^k\rangle + (\frac{L_{\beta_k}}{2} - \frac{1}{\eta_k})\|x^{k+1} - x^k\|^2] \\ &\leq \mathbb{E}[\mathcal{L}_{\beta_k}(x^k, \lambda^k) + \frac{\eta_k}{2}\|\varepsilon^k\|^2 + (\frac{L_{\beta_k}}{2} - \frac{1}{2\eta_k})\|x^{k+1} - x^k\|^2].\end{aligned}$$

Together with (3.6), we obtain

$$\mathbb{E}[\mathcal{L}_{\beta_k}(x^{k+1}, \lambda^{k+1}) - \mathcal{L}_{\beta_k}(x^k, \lambda^k)] \leq \frac{\eta_k}{2}\mathbb{E}[\|\varepsilon^k\|^2] + (\frac{L_{\beta_k}}{2} - \frac{1}{2\eta_k})\mathbb{E}[\|x^{k+1} - x^k\|^2] + \rho_k M^2. \tag{3.7}$$

Due to $\mathcal{L}_{\beta_{k+1}}(x, \lambda) = \mathcal{L}_{\beta_k}(x, \lambda) + \frac{\beta_{k+1} - \beta_k}{2}\|\mathbf{c}(x)\|^2$, it holds that

$$\mathbb{E}[\mathcal{L}_{\beta_k}(x^{k+1}, \lambda^{k+1}) - \mathcal{L}_{\beta_k}(x^k, \lambda^k)] = \mathbb{E}[\mathcal{L}_{\beta_{k+1}}(x^{k+1}, \lambda^{k+1}) - \mathcal{L}_{\beta_k}(x^k, \lambda^k) - \frac{\beta_{k+1} - \beta_k}{2}\|\mathbf{c}(x^{k+1})\|^2].$$

Substituting the above equality into (3.7) and rearranging the terms, we derive the conclusion. □

The lemma below provides a recursive bound on ε^k.

LEMMA **3.6.** *Under Assumptions 1.1-1.3, the following relation holds for any $k \in [K]$,*

$$\begin{aligned}\mathbb{E}[\|\varepsilon^{k+1}\|^2] \leq{}& (1 - \alpha_k)^2\mathbb{E}[\|\varepsilon^k\|^2] + 2\alpha_k^2\sigma_{\beta_{k+1}}^2 + 4(1 - \alpha_k)^2 L_{\beta_k}^2\mathbb{E}[\|x^{k+1} - x^k\|^2] \\ &+ 8(1 - \alpha_k)^2(\beta_{k+1} - \beta_k)^2 M^2 G_c^2 + 8(1 - \alpha_k)^2\rho_k^2 M^2 G_c^2.\end{aligned} \tag{3.8}$$

Proof. By the definition of ε^k we know that

$$\begin{aligned}\varepsilon^{k+1} &= (1 - \alpha_k)(d^k - \mathcal{G}_{\beta_k}(x^k, \lambda^k; \xi^{k+1}, \zeta^{k+1})) - \nabla_x \mathscr{D}_{\beta_{k+1}}(x^{k+1}, \lambda^{k+1}) + \mathcal{G}_{\beta_{k+1}}(x^{k+1}, \lambda^{k+1}; \xi^{k+1}, \zeta^{k+1}) \\ &= (1 - \alpha_k)(\nabla_x \mathscr{D}_{\beta_k}(x^k, \lambda^k) - \mathcal{G}_{\beta_k}(x^k, \lambda^k; \xi^{k+1}, \zeta^{k+1})) + (1 - \alpha_k)\varepsilon^k\end{aligned}$$

$$+ \mathcal{G}_{\beta_{k+1}}(x^{k+1}, \lambda^{k+1}; \xi^{k+1}, \zeta^{k+1}) - \nabla_x \mathscr{D}_{\beta_{k+1}}(x^{k+1}, \lambda^{k+1}).$$

By squaring both sides of above equality and taking expectation with respect to ξ^{k+1} and ζ^{k+1}, we obtain

$$\begin{aligned}\mathbb{E}_{\xi^{k+1},\zeta^{k+1}}[\|\varepsilon^{k+1}\|^2] &= (1-\alpha_k)^2\|\varepsilon^k\|^2 + \mathbb{E}_{\xi^{k+1},\zeta^{k+1}}[\|\mathcal{G}_{\beta_{k+1}}(x^{k+1}, \lambda^{k+1}; \xi^{k+1}, \zeta^{k+1}) - \nabla_x \mathscr{D}_{\beta_{k+1}}(x^{k+1}, \lambda^{k+1}) \\ &\quad + (1-\alpha_k)(\nabla_x \mathscr{D}_{\beta_k}(x^k, \lambda^k) - \mathcal{G}_{\beta_k}(x^k, \lambda^k; \xi^{k+1}, \zeta^{k+1}))\|^2] \\ &= (1-\alpha_k)^2\|\varepsilon^k\|^2 + \mathbb{E}_{\xi^{k+1},\zeta^{k+1}}[\|\alpha_k \mathcal{A}_{k+1} + (1-\alpha_k)(\mathcal{B}_{k+1} - \mathcal{C}_{k+1})\|^2],\end{aligned}$$

where the first equality is derived from the relations

$$\mathbb{E}_{\xi^{k+1},\zeta^{k+1}}[\langle \mathcal{G}_{\beta_{k+1}}(x^{k+1}, \lambda^{k+1}; \xi^{k+1}, \zeta^{k+1}) - \nabla_x \mathscr{D}_{\beta_{k+1}}(x^{k+1}, \lambda^{k+1}), \varepsilon^k\rangle] = 0,$$

$$\mathbb{E}_{\xi^{k+1},\zeta^{k+1}}[\langle \mathcal{G}_{\beta_k}(x^k, \lambda^k; \xi^{k+1}, \zeta^{k+1}) - \nabla_x \mathscr{D}_{\beta_k}(x^k, \lambda^k), \varepsilon^k\rangle] = 0,$$

and the second equality uses the notations

$$\mathcal{A}_{k+1} := \mathcal{G}_{\beta_{k+1}}(x^{k+1}, \lambda^{k+1}; \xi^{k+1}, \zeta^{k+1}) - \nabla_x \mathscr{D}_{\beta_{k+1}}(x^{k+1}, \lambda^{k+1}),$$

$$\mathcal{B}_{k+1} := \mathcal{G}_{\beta_{k+1}}(x^{k+1}, \lambda^{k+1}; \xi^{k+1}, \zeta^{k+1}) - \mathcal{G}_{\beta_k}(x^k, \lambda^k; \xi^{k+1}, \zeta^{k+1}),$$

$$\mathcal{C}_{k+1} := \nabla_x \mathscr{D}_{\beta_{k+1}}(x^{k+1}, \lambda^{k+1}) - \nabla_x \mathscr{D}_{\beta_k}(x^k, \lambda^k).$$

It is easy to check from (3.2)-(3.3) and $\mathbb{E}_{\xi^{k+1},\zeta^{k+1}}[\mathcal{B}_{k+1}] = \mathcal{C}_{k+1}$ that

$$\begin{aligned}&\mathbb{E}_{\xi^{k+1},\zeta^{k+1}}[\|\varepsilon^{k+1}\|^2] \le (1-\alpha_k)^2\|\varepsilon^k\|^2 + \mathbb{E}_{\xi^{k+1},\zeta^{k+1}}[2\alpha_k^2\|\mathcal{A}_{k+1}\|^2 + 2(1-\alpha_k)^2\|\mathcal{B}_{k+1} - \mathcal{C}_{k+1}\|^2] \\ &= (1-\alpha_k)^2\|\varepsilon^k\|^2 + \mathbb{E}_{\xi^{k+1},\zeta^{k+1}}[2\alpha_k^2\|\mathcal{A}_{k+1}\|^2 + 2(1-\alpha_k)^2\|\mathcal{B}_{k+1}\|^2 - 2(1-\alpha_k)^2\|\mathcal{C}_{k+1}\|^2] \\ &\le (1-\alpha_k)^2\|\varepsilon^k\|^2 + \mathbb{E}_{\xi^{k+1},\zeta^{k+1}}[2\alpha_k^2\|\mathcal{A}_{k+1}\|^2 + 2(1-\alpha_k)^2\|\mathcal{B}_{k+1}\|^2] \\ &\le (1-\alpha_k)^2\|\varepsilon^k\|^2 + 2\alpha_k^2\sigma_{\beta_{k+1}}^2 \\ &\quad + 4(1-\alpha_k)^2\mathbb{E}_{\xi^{k+1},\zeta^{k+1}}[\|\mathcal{G}_{\beta_k}(x^{k+1}, \lambda^k; \xi^{k+1}, \zeta^{k+1}) - \mathcal{G}_{\beta_k}(x^k, \lambda^k; \xi^{k+1}, \zeta^{k+1})\|^2] \\ &\quad + 4(1-\alpha_k)^2\mathbb{E}_{\xi^{k+1},\zeta^{k+1}}[\|\mathcal{G}_{\beta_{k+1}}(x^{k+1}, \lambda^{k+1}; \xi^{k+1}, \zeta^{k+1}) - \mathcal{G}_{\beta_k}(x^{k+1}, \lambda^k; \xi^{k+1}, \zeta^{k+1})\|^2] \\ &\le (1-\alpha_k)^2\|\varepsilon^k\|^2 + 2\alpha_k^2\sigma_{\beta_{k+1}}^2 + 4(1-\alpha_k)^2L_{\beta_k}^2\|x^{k+1} - x^k\|^2 \\ &\quad + 4(1-\alpha_k)^2\mathbb{E}_{\zeta^{k+1}}[\|\bar{\mathcal{G}}_{\beta_{k+1}}(x^{k+1}, \lambda^{k+1}; \zeta^{k+1}) - \bar{\mathcal{G}}_{\beta_k}(x^{k+1}, \lambda^k; \zeta^{k+1})\|^2].\end{aligned}$$

For the last term of the above inequality, it can be derived from (1.7), Assumption 1.1, Lemma 3.2 and the independence of ζ_1^{k+1} and ζ_2^{k+1} that

$$\begin{aligned}&\mathbb{E}_{\zeta^{k+1}}[\|\bar{\mathcal{G}}_{\beta_{k+1}}(x^{k+1}, \lambda^{k+1}; \zeta^{k+1}) - \bar{\mathcal{G}}_{\beta_k}(x^{k+1}, \lambda^k; \zeta^{k+1})\|^2] \\ &\le 2\mathbb{E}_{\zeta^{k+1}}[\|\bar{\mathcal{G}}_{\beta_{k+1}}(x^{k+1}, \lambda^{k+1}; \zeta^{k+1}) - \bar{\mathcal{G}}_{\beta_k}(x^{k+1}, \lambda^{k+1}; \zeta^{k+1})\|^2] \\ &\quad + 2\mathbb{E}_{\zeta^{k+1}}[\|\bar{\mathcal{G}}_{\beta_k}(x^{k+1}, \lambda^{k+1}; \zeta^{k+1}) - \bar{\mathcal{G}}_{\beta_k}(x^{k+1}, \lambda^k; \zeta^{k+1})\|^2] \\ &\le 2(\beta_{k+1} - \beta_k)^2\mathbb{E}_{\zeta^{k+1}}[\|\nabla \mathbf{C}(x^{k+1}; \zeta_1^{k+1})\|^2\|\mathbf{C}(x^{k+1}; \zeta_2^{k+1})\|^2] \\ &\quad + 2\mathbb{E}_{\zeta^{k+1}}[\|\lambda^{k+1} - \lambda^k\|^2\|\nabla \mathbf{C}(x^{k+1}; \zeta_1^{k+1})\|^2] \le 2(\beta_{k+1} - \beta_k)^2M^2G_c^2 + 2\rho_k^2M^2G_c^2.\end{aligned}$$

Then the conclusion can be derived by taking full expectation. □

4 Sample complexity analysis

In this section, we analyze sample complexities of TStoM to find approximate solutions. We assume that parameters used in TStoM satisfy

$$\eta_k L_{\beta_k} < \frac{1}{2}, \quad \frac{4\eta_k^2 L_{\beta_k}^2}{1-\eta_k L_{\beta_k}} \le \alpha_k < 1, \quad \frac{4\eta_k^2 L_{\beta_k}^2}{1-\eta_k L_{\beta_k}} \le \tau_k^2 < 1. \tag{4.1}$$

LEMMA **4.1.** *Suppose that Assumptions 1.1-1.3 hold, then it gives that for any $k \in [K]$,*

$$\begin{aligned}\frac{1}{K}\sum_{k=1}^{K}\alpha_k\mathbb{E}[\|\varepsilon^k\|^2] \le \frac{\sigma_{\beta_1}^2}{K|\mathcal{M}|} + \frac{2\sum_{k=1}^{K}\alpha_k^2\sigma_{\beta_{k+1}}^2}{K} + \frac{8M^2G_c^2}{K}\sum_{k=1}^{K}(\beta_{k+1}-\beta_k)^2 + \frac{8M^2G_c^2}{K}\sum_{k=1}^{K}\rho_k^2 \\ + \frac{2}{K}\sum_{k=1}^{K}\frac{\alpha_k(1-\alpha_k)^2}{\eta_k}\mathbb{E}[\mathcal{L}_{\beta_k}(x^k,\lambda^k) - \mathcal{L}_{\beta_{k+1}}(x^{k+1},\lambda^{k+1}) + \frac{\beta_{k+1}-\beta_k}{2}M^2 + \rho_k M^2].\end{aligned}$$

Proof. Substituting (3.5) into (3.8) leads to

$$\begin{aligned}\mathbb{E}[\|\varepsilon^{k+1}\|^2] &\le (1-\alpha_k)^2\mathbb{E}[\|\varepsilon^k\|^2] + 2\alpha_k^2\sigma_{\beta_{k+1}}^2 + 8(\beta_{k+1}-\beta_k)^2M^2G_c^2 + 8\rho_k^2M^2G_c^2 \\ &\quad + 4(1-\alpha_k)^2L_{\beta_k}^2\mathbb{E}[\|x^{k+1}-x^k\|^2] \\ &\le (1-\alpha_k)^2\mathbb{E}[\|\varepsilon^k\|^2] + 2\alpha_k^2\sigma_{\beta_{k+1}}^2 + 8(\beta_{k+1}-\beta_k)^2M^2G_c^2 + 8\rho_k^2M^2G_c^2 \\ &\quad + \frac{8(1-\alpha_k)^2\eta_k L_{\beta_k}^2}{1-\eta_k L_{\beta_k}}\mathbb{E}[\mathcal{L}_{\beta_k}(x^k,\lambda^k) - \mathcal{L}_{\beta_{k+1}}(x^{k+1},\lambda^{k+1}) + \frac{\beta_{k+1}-\beta_k}{2}M^2 + \frac{\eta_k}{2}\|\varepsilon^k\|^2 + \rho_k M^2] \\ &\le (1-\alpha_k)^2\mathbb{E}[\|\varepsilon^k\|^2] + 2\alpha_k^2\sigma_{\beta_{k+1}}^2 + 8(\beta_{k+1}-\beta_k)^2M^2G_c^2 + 8\rho_k^2M^2G_c^2 \\ &\quad + \frac{2\alpha_k(1-\alpha_k)^2}{\eta_k}\mathbb{E}[\mathcal{L}_{\beta_k}(x^k,\lambda^k) - \mathcal{L}_{\beta_{k+1}}(x^{k+1},\lambda^{k+1}) + \frac{\beta_{k+1}-\beta_k}{2}M^2 + \frac{\eta_k}{2}\|\varepsilon^k\|^2 + \rho_k M^2] \\ &= (1-\alpha_k)^2(1+\alpha_k)\mathbb{E}[\|\varepsilon^k\|^2] + 2\alpha_k^2\sigma_{\beta_{k+1}}^2 + 8(\beta_{k+1}-\beta_k)^2M^2G_c^2 + 8\rho_k^2M^2G_c^2 \\ &\quad + \frac{2\alpha_k(1-\alpha_k)^2}{\eta_k}\mathbb{E}[\mathcal{L}_{\beta_k}(x^k,\lambda^k) - \mathcal{L}_{\beta_{k+1}}(x^{k+1},\lambda^{k+1}) + \frac{\beta_{k+1}-\beta_k}{2}M^2 + \rho_k M^2] \\ &\le (1-\alpha_k)\mathbb{E}[\|\varepsilon^k\|^2] + 2\alpha_k^2\sigma_{\beta_{k+1}}^2 + 8(\beta_{k+1}-\beta_k)^2M^2G_c^2 + 8\rho_k^2M^2G_c^2 \\ &\quad + \frac{2\alpha_k(1-\alpha_k)^2}{\eta_k}\mathbb{E}[\mathcal{L}_{\beta_k}(x^k,\lambda^k) - \mathcal{L}_{\beta_{k+1}}(x^{k+1},\lambda^{k+1}) + \frac{\beta_{k+1}-\beta_k}{2}M^2 + \rho_k M^2],\end{aligned}$$

where the third inequality follows from (4.1) and the last inequality comes from $\alpha_k \in (0,1)$. Summing the above inequality over $k = 1,\dots,K$ and using $\mathbb{E}[\|\varepsilon^1\|^2] \le \frac{\sigma_{\beta_1}^2}{|\mathcal{M}|}$ by (3.4), we derive

$$\begin{aligned}\sum_{k=1}^{K}\alpha_k\mathbb{E}[\|\varepsilon^k\|^2] \le \frac{\sigma_{\beta_1}^2}{|\mathcal{M}|} + 2\sum_{k=1}^{K}\alpha_k^2\sigma_{\beta_{k+1}}^2 + 8M^2G_c^2\sum_{k=1}^{K}(\beta_{k+1}-\beta_k)^2 + 8M^2G_c^2\sum_{k=1}^{K}\rho_k^2 \\ + 2\sum_{k=1}^{K}\frac{\alpha_k(1-\alpha_k)^2}{\eta_k}\mathbb{E}[\mathcal{L}_{\beta_k}(x^k,\lambda^k) - \mathcal{L}_{\beta_{k+1}}(x^{k+1},\lambda^{k+1}) + \frac{\beta_{k+1}-\beta_k}{2}M^2 + \rho_k M^2].\end{aligned} \tag{4.2}$$

The desired result is then obtained by dividing the inequality in (4.2) by K. □

LEMMA **4.2.** *Under Assumptions 1.1-1.3 and (4.1), set $\rho_k \equiv \frac{\rho}{K}$ and positive parameters $\beta_k \equiv \beta_1, \eta_k \equiv \eta_1, \alpha_k \equiv \alpha_1, k \ge 1$, then it holds that with $\bar{\lambda}^{k+1} := \lambda^{k+1} - \beta_k\mathbf{c}(x^{k+1})$, $k \in [K]$,*

$$\mathbb{E}[\mathbf{d}^2(\nabla f(x^{R+1}) + \partial h(x^{R+1}) - \nabla\mathbf{c}(x^{R+1})\bar{\lambda}^{R+1}, -\mathcal{N}_X(x^{R+1}))] \le \frac{4(\rho MG)^2}{K^2} + \frac{14\sigma_{\beta_1}^2}{\alpha_1 K|\mathcal{M}|}$$

$$+\frac{28\sum_{k=1}^{K}\alpha_1\sigma_{\beta_1}^2}{K}+\frac{112M^2G_c^2\rho^2}{\alpha_1K^2}+\frac{20+28(1-\alpha_1)^2}{\eta_1K}(\mathcal{L}_{\beta_1}(x^1,\lambda^1)-C^*+2M^2\rho). \quad (4.3)$$

Proof. To begin with, we provide an upper bound for the stationarity measure as represented in (1.4). It follows from Lemma 3.4 that

$$\begin{aligned}&\frac{1}{K}\sum_{k=1}^{K}\mathbb{E}[\mathbf{d}^2(\nabla_x\mathscr{D}_{\beta_k}(x^{k+1},\lambda^{k+1})+\partial h(x^{k+1}),-\mathcal{N}_X(x^{k+1}))]\\&\leq\frac{4(\rho MG)^2}{K^2}+\frac{4}{K}\sum_{k=1}^{K}(L_{\beta_k}^2+\frac{1}{\eta_k^2})\mathbb{E}[\|x^{k+1}-x^k\|^2]+\frac{4}{K}\sum_{k=1}^{K}\mathbb{E}[\|\varepsilon^k\|^2]. \quad (4.4)\end{aligned}$$

For the second item in R.H.S. of (4.4), it is easy to attain from Lemma 3.5 and $\beta_k\equiv\beta_1$ that

$$\begin{aligned}&\frac{1}{K}\sum_{k=1}^{K}(L_{\beta_k}^2+\frac{1}{\eta_k^2})\mathbb{E}[\|x^{k+1}-x^k\|^2]\\&\leq\frac{1}{K}\sum_{k=1}^{K}\frac{2(1+\eta_k^2L_{\beta_k}^2)}{\eta_k(1-\eta_kL_{\beta_k})}\mathbb{E}[\mathcal{L}_{\beta_k}(x^k,\lambda^k)-\mathcal{L}_{\beta_{k+1}}(x^{k+1},\lambda^{k+1})+\frac{\eta_k}{2}\|\varepsilon^k\|^2+\rho_kM^2]\\&\leq\frac{5}{\eta_1K}\sum_{k=1}^{K}\mathbb{E}[\mathcal{L}_{\beta_k}(x^k,\lambda^k)-\mathcal{L}_{\beta_{k+1}}(x^{k+1},\lambda^{k+1})+\rho_kM^2]+\frac{5}{2K}\sum_{k=1}^{K}\mathbb{E}[\|\varepsilon^k\|]^2, \quad (4.5)\end{aligned}$$

where the second inequality comes from $\eta_k\equiv\eta_1$ for all $k\in[K]$, together with $\frac{1+v^2}{1-v}<\frac{5}{2}$ for $v\in(0,\frac{1}{2})$ and (4.1). Note that

$$\begin{aligned}&\sum_{k=1}^{K}\mathbb{E}[\mathcal{L}_{\beta_k}(x^k,\lambda^k)-\mathcal{L}_{\beta_{k+1}}(x^{k+1},\lambda^{k+1})]=\mathcal{L}_{\beta_1}(x^1,\lambda^1)-\mathbb{E}[\mathcal{L}_{\beta_{K+1}}(x^{K+1},\lambda^{K+1})]\\&=\mathcal{L}_{\beta_1}(x^1,\lambda^1)-\mathbb{E}[f(x^{K+1})+h(x^{K+1})-(\lambda^{K+1})^{\mathsf{T}}\mathbf{c}(x^{K+1})+\frac{\beta_{K+1}}{2}\|\mathbf{c}(x^{K+1})\|^2]\\&\leq\mathcal{L}_{\beta_1}(x^1,\lambda^1)-C^*+M^2\sum_{k=1}^{K}\rho_k=\mathcal{L}_{\beta_1}(x^1,\lambda^1)-C^*+M^2\rho. \quad (4.6)\end{aligned}$$

Substituting (4.6) into (4.5) we can obtain

$$\frac{1}{K}\sum_{k=1}^{K}(L_{\beta_k}^2+\frac{1}{\eta_k^2})\mathbb{E}[\|x^{k+1}-x^k\|^2]\leq\frac{5}{\eta_1K}(\mathcal{L}_{\beta_1}(x^1,\lambda^1)-C^*+2M^2\rho)+\frac{5}{2K}\sum_{k=1}^{K}\mathbb{E}[\|\varepsilon^k\|^2]. \quad (4.7)$$

For the last term of (4.4), by Lemma 4.1, (4.6), $\alpha_k\equiv\alpha_1$, $\beta_k\equiv\beta_1$ and $\rho_k\equiv\frac{\rho}{K}$, we have

$$\frac{1}{K}\sum_{k=1}^{K}\mathbb{E}[\|\varepsilon^k\|^2]\leq\frac{\sigma_{\beta_1}^2}{\alpha_1K|\mathcal{M}|}+\frac{2\sum_{k=1}^{K}\alpha_1\sigma_{\beta_1}^2}{K}+\frac{8M^2G_c^2\rho^2}{\alpha_1K^2}+\frac{2(1-\alpha_1)^2}{\eta_1K}(\mathcal{L}_{\beta_1}(x^1,\lambda^1)-C^*+2M^2\rho). \quad (4.8)$$

Then, plugging (4.7) into (4.4), we obtain

$$\begin{aligned}&\mathbb{E}[\mathbf{d}^2(\nabla f(x^{R+1})+\partial h(x^{R+1})-\nabla\mathbf{c}(x^{R+1})\bar{\lambda}^{R+1},-\mathcal{N}_X(x^{R+1}))]\\&=\frac{1}{K}\sum_{k=1}^{K}\mathbb{E}[\mathbf{d}^2(\nabla_x\mathscr{D}_{\beta_k}(x^{k+1},\lambda^{k+1})+\partial h(x^{k+1}),-\mathcal{N}_X(x^{k+1}))]\end{aligned}$$

$$\leq \frac{4(\rho MG)^2}{K^2} + \frac{20}{\eta_1 K}(\mathcal{L}_{\beta_1}(x^1,\lambda^1) - C^* + 2M^2\rho) + \frac{14}{K}\sum_{k=1}^{K}\mathbb{E}[\|\varepsilon^k\|^2],$$

which yields the conclusion by (4.8). □

LEMMA **4.3.** *Suppose that the conditions in Lemma 4.2 are satisfied, then it holds that*

$$\mathbb{E}[\mathbf{d}^2(\nabla\mathbf{c}(x^{R+1})\mathbf{c}(x^{R+1}), -\mathcal{N}_X(x^{R+1}))] \leq \frac{4}{\beta_1^2}\Big(\frac{4(\rho MG)^2}{K^2} + \frac{14\sigma_{\beta_1}^2}{\alpha_1 K|\mathcal{M}|} + \frac{28\sum_{k=1}^{K}\alpha_1\sigma_{\beta_1}^2}{K}$$
$$+ \frac{112M^2G_c^2\rho^2}{\alpha_1 K^2} + \frac{20+28(1-\alpha_1)^2}{\eta_1 K}(\mathcal{L}_{\beta_1}(x^1,\lambda^1) - C^* + 2M^2\rho) + (2+M^2\rho^2)G^2\Big).$$

Proof. It is apparent that there exists $s^{k+1} \in \partial h(x^{k+1})$ such that

$$\begin{aligned}&\mathbf{d}(\nabla f(x^{k+1}) + s^{k+1} - \nabla\mathbf{c}(x^{k+1})\bar{\lambda}^{k+1}, -\mathcal{N}_X(x^{k+1}))\\ &= \mathbf{d}(\nabla f(x^{k+1}) + \partial h(x^{k+1}) - \nabla\mathbf{c}(x^{k+1})\bar{\lambda}^{k+1}, -\mathcal{N}_X(x^{k+1})).\end{aligned} \tag{4.9}$$

For any $k \geq 1$, it follows from (1.5)-(1.6), Lemma 3.2 and $\bar{\lambda}^{k+1} := \lambda^{k+1} - \beta_k\mathbf{c}(x^{k+1})$ that

$$\begin{aligned}&\mathbb{E}[\mathbf{d}^2(\nabla\mathbf{c}(x^{k+1})\mathbf{c}(x^{k+1}), -\mathcal{N}_X(x^{k+1}))] = \frac{1}{\beta_k^2}\mathbb{E}[\mathbf{d}^2(\beta_k\nabla\mathbf{c}(x^{k+1})\mathbf{c}(x^{k+1}), -\mathcal{N}_X(x^{k+1}))]\\
&\leq \frac{4}{\beta_k^2}\mathbb{E}[\mathbf{d}^2(\nabla f(x^{k+1}) + s^{k+1} - \nabla\mathbf{c}(x^{k+1})\bar{\lambda}^{k+1}, -\mathcal{N}_X(x^{k+1})) + \|\nabla f(x^{k+1})\|^2 + \|s^{k+1}\|^2\\
&\quad + \|\lambda^{k+1}\|^2\|\nabla\mathbf{c}(x^{k+1})\|^2]\\
&\leq \frac{4}{\beta_1^2}\mathbb{E}[\mathbf{d}^2(\nabla f(x^{k+1}) + \partial h(x^{k+1}) - \nabla\mathbf{c}(x^{k+1})\bar{\lambda}^{k+1}, -\mathcal{N}_X(x^{k+1})) + (2 + (M\sum_{t=1}^{k}\rho_t)^2)G^2].\end{aligned}$$

Hence, from Lemma 4.2 and

$$\mathbb{E}[\mathbf{d}^2(\nabla\mathbf{c}(x^{R+1})\mathbf{c}(x^{R+1}), -\mathcal{N}_X(x^{R+1}))] = \frac{1}{K}\sum_{k=1}^{K}\mathbb{E}[\mathbf{d}^2(\nabla\mathbf{c}(x^{k+1})\mathbf{c}(x^{k+1}), -\mathcal{N}_X(x^{k+1}))],$$

we derive the conclusion. □

4.1 Towards a stochastic ϵ-stationary point

We will establish the sample complexity of TStoM to locate a stochastic ϵ-stationary point of (1.1). Following (4.1) we set

$$\beta_k = \beta_0 K^{\upsilon}, \eta_k = \frac{\eta}{L_{\beta_k}K^{2\iota}}, \alpha_k = \frac{4\alpha\eta^2}{K^{2\iota}(K^{2\iota}-\eta)}, \quad \forall k \in [K], \tag{4.10}$$

where $\beta_0, \upsilon, \iota > 0, \eta \in (0, \frac{\sqrt{17}-1}{8})$ and $\alpha \in [1, \frac{1-\eta}{4\eta^2})$ are constants independent of K. With $|\mathcal{M}| = K^{1/3}$, $\sigma_{\beta_k} = \mathcal{O}(\beta_k)$ and $L_{\beta_k} = \mathcal{O}(\beta_k)$, the upper bounds given in Lemmas 4.2 and 4.3 are in order

$$\mathcal{O}(K^{-2} + K^{4\iota+2\upsilon-\frac{4}{3}} + K^{2\upsilon-4\iota} + K^{4\iota-2} + K^{2\iota+2\upsilon-1}),$$
$$\mathcal{O}(K^{-2\upsilon}, K^{-2\upsilon}(K^{-2} + K^{4\iota+2\upsilon-\frac{4}{3}} + K^{2\upsilon-4\iota} + K^{4\iota-2} + K^{2\iota+2\upsilon-1})),$$

respectively. Then to derive the lowest possible complexity order, we can determine υ, ι by solving

$$\min_{\upsilon,\iota>0}\max\{4\iota + 2\upsilon - \frac{4}{3}, 2\upsilon - 4\iota, 2\upsilon + 2\iota - 1, -2\upsilon\}. \tag{4.11}$$

It is easy to verify that its optimal value is reached at $\upsilon = \iota = \frac{1}{6}$. We thus obtain the complexity result to find a stochastic ϵ-stationary point.

THEOREM **4.1.** *Suppose that Assumptions 1.1-1.3 hold, and* $\beta_k = K^{1/6}$, $\rho_k \equiv \frac{\rho}{K}$, $\eta_k \equiv \frac{\eta}{L_{\beta_k} K^{1/3}}$, $|\mathcal{M}| = K^{1/3}$, $\alpha_k \equiv \frac{4\alpha\eta^2}{K^{1/3}(K^{1/3}-\eta)}, k \in [K]$ *with* $\rho \in (0, K^{7/6}], \eta \in (0, \frac{\sqrt{17}-1}{8}), \alpha \in [1, \frac{1-\eta}{4\eta^2})$ *being constants independent of* K*, then*

$$\mathbb{E}[\mathbf{d}^2(\nabla f(x^{R+1}) + \partial h(x^{R+1}) - \nabla \mathbf{c}(x^{R+1})\bar{\lambda}^{R+1}, -\mathcal{N}_X(x^{R+1}))] = \mathcal{O}(K^{-1/3}), \tag{4.12}$$

$$\mathbb{E}[\mathbf{d}^2(\nabla \mathbf{c}(x^{R+1})\mathbf{c}(x^{R+1}), -\mathcal{N}_X(x^{R+1}))] = \mathcal{O}(K^{-1/3}). \tag{4.13}$$

Consequently, to find a stochastic ϵ*-stationary point of* (1.1) *the sample complexity of TStoM with* $T = 0$ *is in order* $\mathcal{O}(\epsilon^{-6})$.

Proof. Under the assumed parameter settings and letting $\upsilon = \iota = \frac{1}{6}$, it is straightforward to obtain from Lemmas 4.2 and 4.3 that (4.12) and (4.13) hold. To attain a stochastic ϵ-stationary point of (1.1), K should be in order $\mathcal{O}(\epsilon^{-6})$. Given that four samples are generated per iteration, the total number of samples used in TStoM with $T = 0$ is bounded by $\mathcal{O}(\epsilon^{-6})$. □

4.2 Towards a stochastic ϵ-KKT point

In this subsection, we will establish the sample complexity of TStoM to find a stochastic ϵ-KKT point of (1.1). As can be seen from Lemmas 4.2 and 4.3, the term $\mathcal{L}_{\beta_1}(x^1, \lambda^1)$, or more specifically $\beta_1\|\mathbf{c}(x^1)\|^2$ has a direct impact on the upper bound of criticality measure. In order to further reduce the sample complexity characterized in (4.11), we must take into account mitigating the impact of possibly large values of β_1. The ideal case is to reduce $\beta_1\|\mathbf{c}(x^1)\|^2$ to the order $\mathcal{O}(1)$, which can be realized when the initial point is sufficiently close to the feasible region. This inspires us to search in Phase I of TStoM for an approximate feasible point to initialize Phase II. To proceed, we impose the following assumption.

Assumption 4.1. *There exists* $\nu > 0$ *such that* $\nu\|\mathbf{c}(x)\| \le \mathbf{d}(\nabla \mathbf{c}(x)\mathbf{c}(x), -\mathcal{N}_X(x))$ *for any* $x \in X$.

Assumption 4.1 requires a nonsingularity condition on the Jacobian of constraint functions. The necessity of such conditions has been revealed in [20, 28] to pursue an approximate KKT point of nonconvex constrained optimization problems. The definition of uniform regularity presented in [5] bears a resemblance to Assumption 4.1. Unlike our work, [6] and [3] require the MFCQ condition and strong LICQ condition, respectively. According to [21], the relationship between the MFCQ condition and Assumption 4.1 is not clearly defined in terms of strength or weakness. Besides, Assumption 4.1 can be implied by assuming a strong LICQ condition. More specifically, suppose that $X := \{x \mid r_j(x) \le 0, j = 1, \ldots, p\}$, where $r_j : \mathbb{R}^n \to \mathbb{R}$ are convex and continuously differentiable. For every $x \in X$, let $\mathcal{I}(x) := \{j \mid r_j(x) = 0\}$, then the normal cone of X at x is given by $\mathcal{N}_X(x) = \{\sum_{j\in\mathcal{I}(x)} \delta_j \nabla r_j(x) : \delta_j \ge 0\}$. Without loss of generality, we define $J(x) := [\nabla c_1(x), \ldots, \nabla c_m(x), \nabla r_1(x), \ldots, \nabla r_q(x)]$ with $\mathcal{I}(x) = \{1, \ldots, q\}$, $q \le p$. By assuming that singular values of $J(x)$ over X are uniformly lower bounded by $\nu > 0$, we can infer that

$$\mathbf{d}(\nabla \mathbf{c}(x)\mathbf{c}(x), -\mathcal{N}_X(x)) = \left\|\sum\nolimits_{i=1}^{m} c_i(x)\nabla c_i(x) + \sum\nolimits_{j=1}^{q} \delta_j \nabla r_j(x)\right\| = \|J(x)\mathbf{c}_\delta(x)\| \ge \nu\|\mathbf{c}(x)\|,$$

where $\mathbf{c}_\delta(x)^\mathsf{T} := [\mathbf{c}(x)^\mathsf{T}, \delta^\mathsf{T}]$ and $\delta \in \mathbb{R}^q$.

Let us first look at Phase I of TStoM. Under Assumptions 1.1-1.3, we know that $\|\nabla \mathbf{c}(u)\mathbf{c}(u) - \nabla \mathbf{c}(v)\mathbf{c}(v)\| \le (G^2 + ML)\|u - v\|$, $\forall u, v \in X$. Therefore, the gradient function of (1.2), denoted by $g(x) := \nabla \mathbf{c}(x)\mathbf{c}(x)$, is $(G^2 + ML)$-Lipschitz continuous over X. Define

$$\mathrm{P}_V(z, d) = V(z - z^+), \text{ where } z^+ = \arg\min_{x\in X}\{\langle d, x\rangle + \frac{V}{2}\|x - z\|^2\}.$$

Let $\{W^t, t \in [T+1]\}$ be generated by Phase I of TStoM. As the approach in Phase I is a momentum-based variance-reduced stochastic gradient method targeted at a least square problem over a convex set, inspired by [33] we obtain the following lemma.

LEMMA **4.4.** *Suppose that Assumptions 1.1-1.3 hold,* $V = 4(G^2 + ML)T^{1/3}$, $\gamma_t = 3[(t+1)^{1/3} - (t+2)^{1/3}]$, $t \in [T]$, *and* $T = \lceil K^{3/10} \rceil$, *then*

$$\mathbb{E}[\|\mathrm{P}_V(z^{R_0}, g^{R_0})\|^2] = \mathcal{O}(K^{-1/5}), \mathbb{E}[\|W^{R_0} - g^{R_0}\|^2] = \mathcal{O}(K^{-1/5}), \mathbb{E}[\|\mathrm{P}_V(z^{R_0}, W^{R_0})\|^2] = \mathcal{O}(K^{-1/5}),$$

where g^{R_0} *is an abbreviation for* $g(z^{R_0})$.

Proof. Under the parameter settings of this lemma, by using Theorem 2.5 in [33] and definition of $\mathrm{P}_V(\cdot,\cdot)$ we obtain $\mathbb{E}[\|\mathrm{P}_V(z^{R_0}, g^{R_0})\|^2] = \mathcal{O}(T^{-2/3}) = \mathcal{O}(K^{-1/5})$. Besides, by (2.34) in [33], we can derive $\mathbb{E}[\|W^{R_0} - g^{R_0}\|^2] = \mathcal{O}(V/T) = \mathcal{O}(K^{-1/5})$. Then due to $\|\mathrm{P}_V(z^{R_0}, W^{R_0}) - \mathrm{P}(z^{R_0}, g^{R_0})\| \leq \|W^{R_0} - g^{R_0}\|$ indicated by Proposition 1 in [13], we attain $\mathbb{E}[\|\mathrm{P}_V(z^{R_0}, W^{R_0})\|^2] = \mathcal{O}(K^{-1/5})$. □

The lemma below ensures that Phase I can find an approximately feasible point of (1.1).

LEMMA **4.5.** *Under Assumptions 1.1-4.1 and same parameter settings as Lemma 4.4, Phase I of TStoM returns a point* x^1 *satisfying* $\mathbb{E}[\|\mathbf{c}(x^1)\|^2] = \mathcal{O}(K^{-1/5})$.

Proof. Due to the independence of ς_1 and ς_2, we can infer that the stochastic gradient of (1.2) satisfies $\mathbb{E}[\nabla \mathbf{C}(x;\varsigma_1)\mathbf{C}(x;\varsigma_2)] = \nabla \mathbf{c}(x)\mathbf{c}(x)$. By Lemma 4.4, we can derive

$$\mathbb{E}[\|g^{R_0+1} - g^{R_0}\|^2] \leq (G^2 + ML)^2 \mathbb{E}[\|z^{R_0+1} - z^{R_0}\|^2] = \frac{(G^2+ML)^2}{V^2} \mathbb{E}[\|\mathrm{P}_V(z^{R_0}, W^{R_0})\|^2],$$

which is bounded by $\mathcal{O}(K^{-2/5})$. Then from the optimality condition for (2.1), i.e. $\mathbf{d}(W^t + V(z^{t+1} - z^t), -\mathcal{N}_X(z^{t+1})) = 0$, together with $V(z^{t+1} - z^t) = -\mathrm{P}_V(z^t, W^t)$ and Lemma 4.4 it follows that

$$\begin{aligned}
&\mathbb{E}[\mathbf{d}^2(\nabla \mathbf{c}(z^{R_0+1})\mathbf{c}(z^{R_0+1}), -\mathcal{N}_X(z^{R_0+1}))] = \mathbb{E}[\mathbf{d}^2(g^{R_0+1}, -\mathcal{N}_X(z^{R_0+1}))] \\
&\leq \mathbb{E}[\|g^{R_0+1} - g^{R_0} + g^{R_0} - (W^{R_0} + V(z^{R_0+1} - z^{R_0}))\|^2] \\
&\leq 3\mathbb{E}[\|g^{R_0+1} - g^{R_0}\|^2] + 3\mathbb{E}[\|W^{R_0} - g^{R_0}\|^2] + 3\mathbb{E}[\|V(z^{R_0+1} - z^{R_0})\|^2] = \mathcal{O}(K^{-1/5}).
\end{aligned}$$

Hence, under Assumption 4.1 and letting $x^1 := z^{R_0+1}$ we obtain $\mathbb{E}[\|\mathbf{c}(x^1)\|^2] = \mathcal{O}(K^{-1/5})$. □

Based on above analysis, we arrive at the following theorem, which characterizes the sample complexity of TStoM to reach a stochastic ϵ-KKT point of (1.1).

THEOREM **4.2.** *Under Assumptions 1.1-4.1 and conditions of Lemma 4.4, suppose that* $|\mathcal{M}| = K^{3/5}$, $\beta_k = K^{1/5}$, $\rho_k \equiv \frac{\rho}{K}$, $\eta_k \equiv \frac{\eta}{L_{\beta_k} K^{2/5}}$, $\alpha_k \equiv \frac{4\alpha\eta^2}{K^{2/5}(K^{2/5}-\eta)}$, $\forall k \in [K]$ *with* $\rho \in (0, K^{6/5}], \eta \in (0, \frac{\sqrt{17}-1}{8})$ *and* $\alpha \in [1, \frac{1-\eta}{4\eta^2})$ *being constants independent of* K. *Then it holds that*

$$\mathbb{E}[\mathbf{d}^2(\nabla f(x^{R+1}) + \partial h(x^{R+1}) - \nabla \mathbf{c}(x^{R+1})\bar{\lambda}^{R+1}, -\mathcal{N}_X(x^{R+1}))] = \mathcal{O}(K^{-2/5}), \tag{4.14}$$

$$\mathbb{E}[\|\mathbf{c}(x^{R+1})\|^2] = \mathcal{O}(K^{-2/5}). \tag{4.15}$$

Consequently, the sample complexity of TStoM to reach a stochastic ϵ*-KKT point of* (1.1) *is in order* $\mathcal{O}(\epsilon^{-5})$.

Proof. As demonstrated in Lemma 4.5, Phase I returns a point x^1 such that $\mathbb{E}[\|\mathbf{c}(x^1)\|^2] = \mathcal{O}(K^{-1/5})$. By the setting of β_k, we obtain $\beta_1\|\mathbf{c}(x^1)\|^2 = \mathcal{O}(1)$. Then it is straightforward to derive (4.14) by substituting the above parameter settings into (4.3). To prove (4.15), recall that there exists $s^{k+1} \in \partial h(x^{k+1})$ such that (4.9) holds, then it follows from Assumption 4.1, (1.5)-(1.6), and Lemma 3.2 that

$$\mathbb{E}[\|\mathbf{c}(x^{R+1})\|^2] = \frac{1}{K}\sum_{k=1}^{K} \mathbb{E}[\|\mathbf{c}(x^{k+1})\|^2] \leq \frac{1}{K}\sum_{k=1}^{K} \Big(\frac{1}{\nu^2\beta_k^2} \mathbb{E}[\mathbf{d}^2(\beta_k \nabla \mathbf{c}(x^{k+1})\mathbf{c}(x^{k+1}), -\mathcal{N}_X(x^{k+1}))] \Big)$$

$$
\begin{aligned}
&\le \frac{4}{\nu^2\beta_1^2 K}\sum_{k=1}^{K}\mathbb{E}[\mathbf{d}^2(\nabla f(x^{k+1})+s^{k+1}+\nabla\mathbf{c}(x^{k+1})(\beta_k\mathbf{c}(x^{k+1})-\lambda^{k+1}),-\mathcal{N}_X(x^{k+1}))\\
&\quad+\|\nabla f(x^{k+1})\|^2+\|s^{k+1}\|^2+\|\lambda^{k+1}\|^2\|\nabla\mathbf{c}(x^{k+1})\|^2]\\
&\le \frac{4}{\nu^2\beta_1^2}\Big(\mathbb{E}[\mathbf{d}^2(\nabla f(x^{R+1})+\partial h(x^{R+1})-\nabla\mathbf{c}(x^{R+1})\bar{\lambda}^{R+1},-\mathcal{N}_X(x^{R+1}))]+(2+M^2\rho^2)G^2\Big),
\end{aligned}
$$

which is in order $\mathcal{O}(K^{-2/5})$.

To attain a stochastic ϵ-KKT point of (1.1), K should be in order $\mathcal{O}(\epsilon^{-5})$. Then in analogy to Theorem 4.1, the number of samples in Phase II is in order $\mathcal{O}(\epsilon^{-3}+\epsilon^{-5})=\mathcal{O}(\epsilon^{-5})$. Meanwhile, since Phase I requires at most two samples per iteration and due to $T=\lceil K^{3/10}\rceil$, we can conclude that the total number of samples is in order $\mathcal{O}(\epsilon^{-5})$. □

As a byproduct, the moving average way to approximate $\mathbf{c}(x)$ as in (2.6) ensures an explicit bound on the expected error of y^{R+1}, which however is not provided in [1].

Proposition 4.1. *Suppose same conditions as Theorem 4.2 hold and set* $\tau_k\equiv\frac{2\tau\eta}{K^{1/5}(K^{2/5}-\eta)^{1/2}}$, $k\in[K]$ *with* $\eta\in(0,\frac{\sqrt{17}-1}{8}),\tau\in[1,\frac{\sqrt{1-\eta}}{2\eta})$. *Then we have* $\mathbb{E}\left[\|y^{R+1}-\mathbf{c}(x^{R+1})\|\right]=\mathcal{O}(K^{-1/5})$, *which is in order* $\mathcal{O}(I^{-1/5})$ *with* I *being the number of samples generated during iteration of TStoM.*

Proof. Denote $\mu^{k+1}:=y^{k+1}-\mathbf{c}(x^{k+1})$. By (2.6), we obtain

$$
\begin{aligned}
\mu^{k+1}&=(1-\tau_k)(y^k-\mathbf{c}(x^k))+\tau_k(\mathbf{C}(x^{k+1};\theta^k)-\mathbf{c}(x^{k+1}))+(1-\tau_k)(\mathbf{c}(x^k)-\mathbf{c}(x^{k+1}))\\
&=(1-\tau_k)\mu^k+\tau_k(p_k+e_k),
\end{aligned}\tag{4.16}
$$

where $p_k:=\mathbf{C}(x^{k+1};\theta^k)-\mathbf{c}(x^{k+1})$, $e_k:=\frac{1-\tau_k}{\tau_k}(\mathbf{c}(x^k)-\mathbf{c}(x^{k+1}))$. Then squaring both sides of (4.16) and taking expectation with respect to θ^k leads to

$$
\begin{aligned}
&\mathbb{E}_{\theta^k}[\|\mu^{k+1}\|^2]=\mathbb{E}_{\theta^k}[\|(1-\tau_k)\mu^k+\tau_k(e_k+p_k)\|^2]\\
&=\mathbb{E}_{\theta^k}[\|(1-\tau_k)\mu^k+\tau_k e_k\|^2]+\tau_k^2\mathbb{E}_{\theta^k}[\|p_k\|^2]+2\tau_k\mathbb{E}_{\theta^k}[\langle(1-\tau_k)\mu^k+\tau_k e_k,p_k\rangle]\\
&\le(1-\tau_k)\|\mu^k\|^2+\tau_k\mathbb{E}_{\theta^k}[\|e_k\|^2]+\tau_k^2\mathbb{E}_{\theta^k}[\|p_k\|^2]\\
&\le(1-\tau_k)\|\mu^k\|^2+\frac{(1-\tau_k)^2G^2}{\tau_k}\mathbb{E}_{\theta^k}[\|x^{k+1}-x^k\|^2]+\tau_k^2\sigma_c^2,
\end{aligned}
$$

where the first inequality follows from the convexity of $\|\cdot\|^2$ and $\mathbb{E}_{\theta^k}[p_k]=0$ and the second inequality is derived from (1.6) and Assumption 1.2. Then by taking full expectation on both sides of the above inequality, summing it over $k=1,\ldots,K$ and applying $\mathbb{E}\left[\|\mu^1\|^2\right]\le\sigma_c^2$, we obtain

$$
\sum_{k=1}^{K}\tau_{k+1}\mathbb{E}\left[\|\mu^{k+1}\|^2\right]\le\sum_{k=1}^{K}\tau_k\mathbb{E}\left[\|\mu^k\|^2\right]+\mathbb{E}\left[\|\mu^{K+1}\|^2\right]\le\sum_{k=1}^{K}\frac{G^2(1-\tau_k)^2}{\tau_k}\mathbb{E}[\|x^{k+1}-x^k\|^2]+(1+\sum_{k=1}^{K}\tau_k^2)\sigma_c^2.
$$

It further implies from (3.5), (4.1) and $\beta_{k+1}\equiv\beta_1$ that

$$
\begin{aligned}
&\sum_{k=1}^{K}\tau_{k+1}\mathbb{E}\left[\|\mu^{k+1}\|^2\right]\\
&\le G^2\sum_{k=1}^{K}\frac{2\eta_k(1-\tau_k)^2}{\tau_k(1-\eta_k L_{\beta_k})}\mathbb{E}[\mathcal{L}_{\beta_k}(x^k,\lambda^k)-\mathcal{L}_{\beta_{k+1}}(x^{k+1},\lambda^{k+1})+\frac{\eta_k}{2}\|\varepsilon^k\|^2+\rho_k M^2]+(1+\sum_{k=1}^{K}\tau_k^2)\sigma_c^2\\
&\le\sum_{k=1}^{K}\frac{G^2\tau_k}{2\eta_k L_{\beta_k}^2}\mathbb{E}[\mathcal{L}_{\beta_k}(x^k,\lambda^k)-\mathcal{L}_{\beta_{k+1}}(x^{k+1},\lambda^{k+1})+\frac{\eta_k}{2}\|\varepsilon^k\|^2+\rho_k M^2]+(1+\sum_{k=1}^{K}\tau_k^2)\sigma_c^2,
\end{aligned}
$$

where the second inequality uses the nonnegativeness of the first term by (3.5). Due to $\tau_{k+1} \equiv \tau_1$ together with (4.6) and (4.8) we have

$$\begin{aligned}
\mathbb{E}\left[\|y^{R+1} - \mathbf{c}(x^{R+1})\|^2\right] &= \frac{1}{K}\sum_{k=1}^{K}\mathbb{E}\left[\|\mu^{k+1}\|^2\right] \le \frac{G^2}{4L_{\beta_1}^2 K}\sum_{k=1}^{K}\mathbb{E}[\|\varepsilon^k\|^2] + \frac{(1+\sum_{k=1}^{K}\tau_k^2)\sigma_c^2}{\tau_1 K} \\
&+ \frac{G^2}{2\eta_1 L_{\beta_1}^2 K}\sum_{k=1}^{K}\mathbb{E}[\mathcal{L}_{\beta_k}(x^k,\lambda^k) - \mathcal{L}_{\beta_{k+1}}(x^{k+1},\lambda^{k+1}) + \rho_k M^2] \\
&\le \frac{G^2\sigma_{\beta_1}^2}{4\alpha_1 L_{\beta_1}^2 K|\mathcal{M}|} + \frac{G^2\sum_{k=1}^{K}\alpha_1\sigma_{\beta_1}^2}{2L_{\beta_1}^2 K} + \frac{2G^2M^2G_c^2\rho^2}{\alpha_1 L_{\beta_1}^2 K^2} + \frac{(1+\sum_{k=1}^{K}\tau_k^2)\sigma_c^2}{\tau_1 K} \\
&+ \frac{(1+(1-\alpha_1)^2)G^2}{2\eta_1 L_{\beta_1}^2 K}(\mathcal{L}_{\beta_1}(x^1,\lambda^1) - C^* + 2M^2\rho) = \mathcal{O}(K^{-\frac{2}{5}}),
\end{aligned}$$

which is in order $\mathcal{O}(I^{-2/5})$. The proof is completed. □

5 Experimental results

5.1 MIMO Transmit Signal Design with Imperfect CSI

For the MIMO transmit signal design problem with imperfect channel state information (CSI), as discussed in [11], the base station is equipped with n antennas and its MIMO signal, based on estimated CSI $\hat{h}_i \in \mathbb{C}^n$, $i = 1, 2, \ldots, k$, simultaneously transmits k data streams to k users. For each i, there is an error between the true CSI h_i and the estimated CSI, which is defined as e_i. To enhance the average MIMO transmission performance, this problem aims for minimizing the total power while enabling each user's expected rate not below a preset threshold. Mathematically, it is formalized as

$$\min_{0\preceq P_i\in\mathbb{R}^{n\times n}} f(P) := \sum_{i=1}^{k}\mathrm{Tr}(P_i) \quad \text{s.t.} \quad \mathbb{E}[G_i(P;E)] \ge r_i, \quad i = 1, 2, \ldots, k,$$

where $G_i(P;E) := \log(1 + \frac{h_i^H P_i h_i}{\sum_{j\neq i} h_i^H P_j h_i + \sigma_i^2})$, $i \in [k]$, $P = \{P_i, i \in [k]\}$, $E = \{e_i, i \in [k]\}$. For user i, P_i denotes the covariance matrix of the transmit signal, σ_i^2 is the variance of the thermal noise, while r_i represents the expected rate. In numerical tests, we set $k = 4$, $n = 8$, $r_i = 0.1$ and $\sigma_i = 0.1$.

We first compare the performance of our proposed TStoM and SPD [16] on this problem. We set the maximum number of samples to 1.2×10^4 and $\alpha_k = 0.6, \tau_k = 0.3$ for TStoM, then adopt identical settings for the remaining parameters in both algorithms to compare the numerical effects of these two algorithms starting from a nearly feasible initial point. From Figure 1 we can see that TStoM by introducing momentum, not only reduces the objective function value more rapidly than SPD but also arrives at a lower constraint violation level. This indicates that the momentum indeed brings benefit to the algorithm's performances. Next, we modify the parameters α_k and τ_k to 0.85 and 0.5, respectively, while ensuring that the remaining parameters of both algorithms maintain their consistency. We then compare the two algorithms starting from a randomly infeasible point within 2×10^4 sample passes, as shown in Figure 2. In addition, under the same parameter settings we also demonstrate the numerical performance of TStoM-P2, which merely implements Phase II of TStoM, in Figure 2. It can be observed that TStoM outperforms the other two in terms of both objective value reduction and constraint violation. This further proves the necessity and effectiveness of finding an initial feasible point.

5.2 Multi-class Neyman-Pearson Classification Problems

We now focus on the multi-class Neyman-Pearson classification problem [21], whose goal is to learn K models $x_k, k \in [K]$ given a set of training data with K classes, and predict the class of a data point

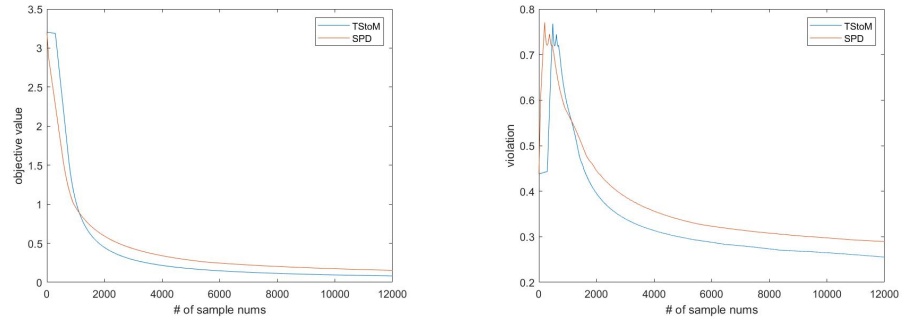

Figure 1: Comparison of TStoM and SPD starting from an feasible initial point

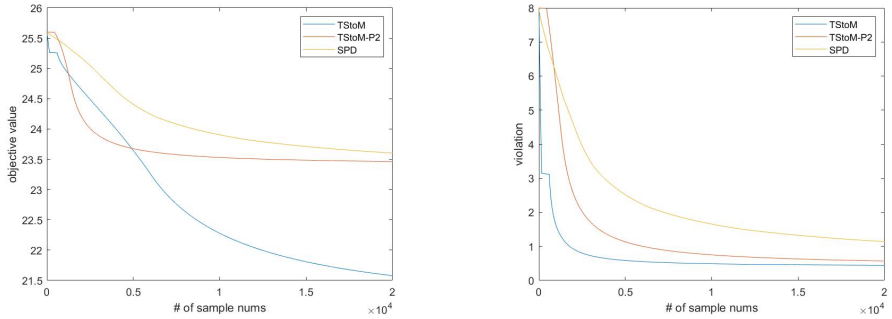

Figure 2: Comparison of TStoM, TStoM-P2 and SPD starting from a randomly infeasible initial point

ξ by choosing the model that maximizes $x_k^\mathsf{T}\xi$. It aims to minimize the loss related to a specific class while controlling the loss values of the other classes, taking the form

$$\begin{aligned} \min_{x_k \in X} \quad & f_1(x) = \frac{1}{|\mathcal{J}_1|} \sum_{p>1} \sum_{\xi \in \mathcal{J}_1} l(x_1^\mathsf{T}\xi - x_p^\mathsf{T}\xi) \\ \text{s.t.} \quad & f_k(x) = \frac{1}{|\mathcal{J}_k|} \sum_{p \neq k} \sum_{\xi \in \mathcal{J}_k} l(x_k^\mathsf{T}\xi - x_p^\mathsf{T}\xi) \leq \gamma_k, \quad k = 2, \ldots, K, \end{aligned}$$

where $X = \{x_k \in \mathbb{R}^n : \|x_k\| \leq \lambda,\ k \in [K]\}$ and $l(z) = 1/(1+e^z)$ is the loss function, $\mathcal{J}_k \subseteq \mathbb{R}^n$ for $k = 1, 2, \ldots, K$ are sets of training data characterized by K classes. We utilize two datasets from LibSVM [8]: *covtype* ($K = 7$) and *mnist* ($K = 10$). We set $\gamma_k = K - 1$, $k = 2, \ldots, K$, and $\lambda = 0.3$.

We evaluate the performances of TStoM compared with SLQPM [1], ICPPC [6] and Stoc-iALM [20]. As can be observed from Figure 3, TStoM shows superior performance on the dataset *covtype*, demonstrating a more rapid decrease in the objective function value and a lower level of constraint violation within the same sample numbers. For the dataset *mnist*, TStoM and Stoc-iALM perform similarly in reducing the objective function, but TStoM stands out in reducing constraint violation, as shown in Figure 4.

5.3 Chance Constrained Program

The chance constrained program is generally given by

$$\min_{x \in \mathcal{C} \subseteq \mathbb{R}^n} \quad \mathbb{E}[F(x;\xi)] \quad \text{s.t.} \quad \mathbb{P}\{G_i(x;\xi) \leq 0, i = 1, \ldots, m\} \geq 1 - \gamma,$$

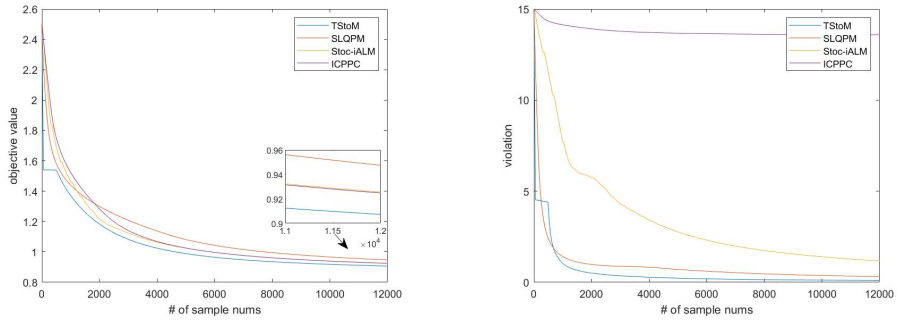

Figure 3: Comparison of TStoM, SLQPM, Stoc-iALM and ICPPC on the covtype dataset

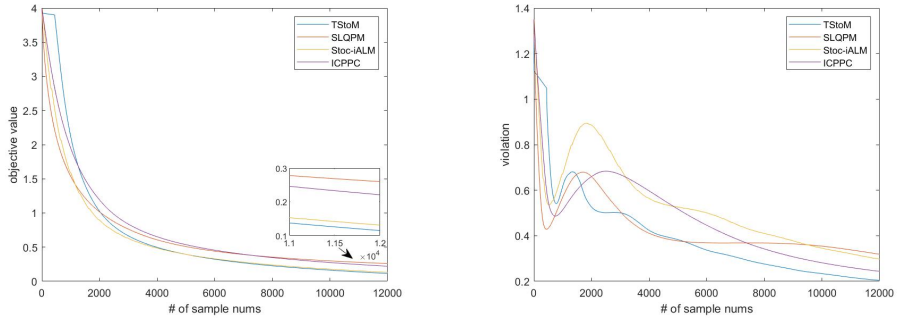

Figure 4: Comparison of TStoM, SLQPM, Stoc-iALM and ICPPC on the mnist dataset

where $\gamma > 0$ is a probability bound. With $G(x;\xi) := \max_{1\le i\le m}\{G_i(x;\xi)\}$, the chance constraint is reformulated as $\mathbb{E}[\mathbb{1}_{[0,\infty)}(G(x;\xi))] \le \gamma$. Nevertheless, since the characteristic function $\mathbb{1}_{[0,\infty)}(\cdot)$ is discontinuous, we introduce a smoothing function $\phi : \mathbb{R} \to \mathbb{R}$ and transform the constraint into $\mathbb{E}[\phi(G(x;\xi))] - \gamma \le 0$. We test the norm optimization problem [15] with slight modifications:

$$\min_{x\in\mathbb{R}^n_+} -\mathbb{E}[\eta^\mathsf{T} x] + \lambda\|x\|^2 \quad \text{s.t.} \quad \mathbb{P}\Big\{\sum_{j=1}^{n} \xi_{ij}^2 x_j^2 \le u^2,\ i \in [m]\Big\} \ge 1-\gamma,$$

where components of η are i.i.d. random variables with both mean value and variance equal to 1, while $\xi_{ij}, i \in [m]$ and $j \in [n]$ are i.i.d. standard normal random variables, $\lambda > 0$ is a regularization parameter. We employ the smoothing function $\phi(y) := (1+\exp(-y/s))^{-1}$ with parameter $s > 0$ [29].

In Figure 5 we report the numerical comparison results between TStoM, ICPPC, Stoc-iALM, SLQPM and SPD. For all five algorithms, the maximum number of samples are set as 1.5×10^4, and $m = 8, n = 3$. It can be observed from Figure 5 that TStoM prevails over the other four in terms of reducing the objective function. With regard to the constraint violation, TStoM exhibits slightly superior performance compared to SPD and SLQPM, outperforming the other two algorithms. Furthermore, while SLQPM excels in KKT residuals, its constraint violation falls short compared to TStoM, and it also performs weakly in objective function reduction. In summary, TStoM demonstrates a more balanced and comprehensivly comparative performance in solving the test problem. Moreover, the comparison between TStoM and SPD further highlights the crucial role that momentum plays.

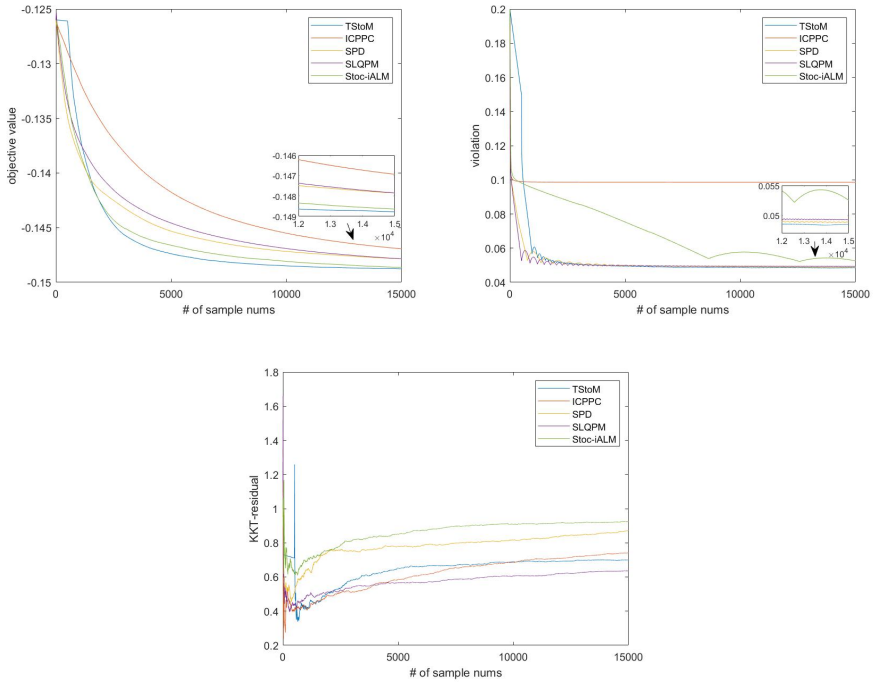

Figure 5: Comparison of TStoM, ICPPC, SPD, SLQPM and Stoc-iALM

6 Conclusion

We study in this paper a two-phase stochastic momentum-based algorithm for nonconvex constrained optimization problems whose objective and constraint functions are in expectation forms. The first phase of algorithm plays as a feasibility search phase, aiming to find an approximately feasible point to initialize the second phase. In the second phase of the algorithm, we incorporate a momentum step to compute the stochastic gradient and construct a stochastic approximation to the linearized augmented Lagrangian function to update the primal variable. The dual update relies on stochastic constraint function values computed through a moving-average scheme. Under certain conditions, we analyze the sample complexities of the proposed algorithm to find a stochastic ϵ-stationary point and a stochastic ϵ-KKT point. We verify the effectiveness of our proposed approach through evaluating its performance in three numerical experiments.

Acknowledgements

We would like to thank Dr. Yangyang Xu and Dr. Zichong Li for kindly sharing their codes regarding the algorithm Stoc-iALM.

References

[1] A. Alacaoglu and S. J. Wright. Complexity of single loop algorithms for nonlinear programming with stochastic objective and constraints. *arXiv:2311.00678*, 2023.

[2] Y. Arjevani, Y. Carmon, J. C. Duchi, D. J. Foster, N. Srebro, and B. Woodworth. Lower bounds for non-convex stochastic optimization. *Mathematical Programming*, 199:165–214, 2023.

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

[4] R. Bollapragada, C. Karamanli, B. Keith, B. Lazarov, S. Petrides, and J. Wang. An adaptive sampling augmented lagrangian method for stochastic optimization with deterministic constraints. *Computers & Mathematics with Applications*, 149:239–258, 2023.

[5] J. Bolte, S. Sabach, and M. Teboulle. Nonconvex lagrangian-based optimization: monitoring schemes and global convergence. *Mathematics of Operations Research*, 43(4):1210–1232, 2018.

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

[7] D. Boob, Q. Deng, and G. Lan. Level constrained first order methods for function constrained optimization. *Mathematical Programming*, 2024.

[8] C. C. Chang and C. J. Lin. Libsvm: A library for support vector machines. *ACM Transactions on Intelligent Systems and Technology*, 2(3, article 27), 2007.

[9] F. E. Curtis, M. J. O'Neill, and D. P. Robinson. Worst-case complexity of an sqp method for nonlinear equality constrained stochastic optimization. *Mathematical Programming*, pages 1–53, 2023.

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

[11] M. Ding and S. D. Blostein. Mimo minimum total mse transceiver design with imperfect csi at both ends. *IEEE Transactions on Signal Processing*, 57(3):1141–1150, 2009.

[12] M. Donini, L. Oneto, S. Ben-David, J. S. Shawe-Taylor, and M. Pontil. Empirical risk minimization under fairness constraints. *Advances in Neural Information Processing Systems*, 31, 2018.

[13] S. Ghadimi, L. G., and H. Zhang. Mini-batch stochastic approximation methods for nonconvex stochastic composite optimization. *Mathematical Programming*, 155:267–305, 2016.

[14] G. Goh, A. Cotter, M. Gupta, and M. P. Friedlander. Satisfying real-world goals with dataset constraints. *Advances in Neural Information Processing Systems*, 29, 2016.

[15] L. J. Hong, Y. Yang, and L. Zhang. Sequential convex approximations to joint chance constrained programs: A monte carlo approach. *Operations Research*, 59(3):617–630, 2011.

[16] L. Jin and X. Wang. A stochastic primal-dual method for a class of nonconvex constrained optimization. *Computational Optimization and Applications*, 83(1):143–180, 2022.

[17] L. Jin and X. Wang. Stochastic nested primal-dual method for nonconvex constrained composition optimization. *to appear in Mathematics of Computation*, 2024.

[18] A. Krishnapriyan, A. Gholami, S. Zhe, R. Kirby, and M. W. Mahoney. Characterizing possible failure modes in physics-informed neural networks. *Advances in Neural Information Processing Systems*, 34:26548–26560, 2021.

[19] P. Krokhmal, J. Palmquist, and S. Uryasev. Portfolio optimization with conditional value-at-risk objective and constraints. *Journal of risk*, 4:43–68, 2002.

[20] Z. Li, P.-Y. Chen, S. Liu, S. Lu, and Y. Xu. Stochastic inexact augmented lagrangian method for nonconvex expectation constrained optimization. *Computational Optimization and Applications*, pages 1–31, 2023.

[21] Q. Lin, R. Ma, and Y. Xu. Complexity of an inexact proximal-point penalty method for constrained smooth non-convex optimization. *Computational Optimization and Applications*, 82(1):175–224, 2022.

[22] R. Ma, Q. Lin, and T. Yang. Quadratically regularized subgradient methods for weakly convex optimization with weakly convex constraints. In *International Conference on Machine Learning*, pages 6554–6564, 2020.

[23] S. Na, M. Anitescu, and M. Kolar. Inequality constrained stochastic nonlinear optimization via active-set sequential quadratic programming. *Mathematical Programming*, pages 1–75, 2023.

[24] S. Na and M. W. Mahoney. Statistical inference of constrained stochastic optimization via sketched sequential quadratic programming. *arXiv:2205.13687*, 2022.

[25] J. Neyman and E. S. Pearson. Ix. on the problem of the most efficient tests of statistical hypotheses. *Philosophical Transactions of the Royal Society of London. Series A, Containing Papers of a Mathematical or Physical Character*, 231(694-706):289–337, 1933.

[26] P. Rigollet and X. Tong. Neyman-pearson classification, convexity and stochastic constraints. *Journal of Machine Learning Research*, pages 2831–2855, 2011.

[27] R. T. Rockafellar, S. Uryasev, et al. Optimization of conditional value-at-risk. *Journal of Risk*, 2:21–42, 2000.

[28] M. F. Sahin, A. Alacaoglu, F. Latorre, V. Cevher, et al. An inexact augmented lagrangian framework for nonconvex optimization with nonlinear constraints. *Advances in Neural Information Processing Systems*, 32, 2019.

[29] F. Shan, L. Zhang, and X. Xiao. A smoothing function approach to joint chance-constrained programs. *Journal of Optimization Theory and Applications*, 163:181–199, 2014.

[30] Q. Shi, X. Wang, and H. Wang. A momentum-based linearized augmented lagrangian method for nonconvex constrained stochastic optimization. *Optimization Online*, 2022.

[31] X. Wang, S. Ma, and Y.-x. Yuan. Penalty methods with stochastic approximation for stochastic nonlinear programming. *Mathematics of computation*, 86(306):1793–1820, 2017.

[32] Y. Xu. Primal-dual stochastic gradient method for convex programs with many functional constraints. *SIAM Journal on Optimization*, 30(2):1664–1692, 2020.

[33] Y. Xu and Y. Xu. Momentum-based variance-reduced proximal stochastic gradient method for composite nonconvex stochastic optimization. *Journal of Optimization Theory and Applications*, 196:266–297, 2023.

[34] Y. Yan and Y. Xu. Adaptive primal-dual stochastic gradient method for expectation-constrained convex stochastic programs. *Mathematical Programming Computation*, 14(2):319–363, 2022.

[35] L. Zhang, Y. Zhang, J. Wu, and X. Xiao. Solving stochastic optimization with expectation constraints efficiently by a stochastic augmented lagrangian-type algorithm. *INFORMS Journal on Computing*, 34(6):2989–3006, 2022.

[36] L. Zhang, Y. Zhang, X. Xiao, and J. Wu. Stochastic approximation proximal method of multipliers for convex stochastic programming. *Mathematics of Operations Research*, 48(1):177–193, 2023.