An adaptive single-loop stochastic penalty method for nonconvex constrained stochastic optimization

Shiji Zuo **Xiao Wang** **Hao Wang**

May 1, 2025

Abstract

Adaptive update schemes for penalty parameters are crucial to enhancing robustness and practical applicability of penalty methods for constrained optimization. However, in the context of general constrained stochastic optimization, additional challenges arise due to the randomness introduced by adaptive penalty parameters. To address these challenges, we propose an **Ada**ptive **S**ingle-loop **S**tochastic **P**enalty method (AdaSSP) in this paper. AdaSSP employs a single-loop algorithmic framework with dynamically updated penalty parameters based on the behavior of iterates. It combines a recursive momentum technique along with clipped stochastic gradient computations to potentially reduce the random variance caused by stochasticity. We present a high-probability oracle complexity analysis for AdaSSP to reach an ϵ-KKT point. We also investigate the in-expectation global convergence regarding the KKT residual at iterates when the penalty parameter sequence is unbounded and bounded, respectively. Finally, preliminary numerical results are reported, revealing the promising performance of the proposed method.

Keywords: Nonconvex constrained optimization · Stochastic approximation · High probability · Oracle complexity · Global convergence

MSCcodes: 90C26 · 90C30 · 65K05 · 62L20

1 Introduction

In this paper we focus on the nonconvex stochastic optimization with deterministic constraints:

$$\begin{aligned} \min_{\boldsymbol{x}\in\mathbb{R}^n} \quad & f(\boldsymbol{x}) \equiv \mathbb{E}_\xi[\mathbf{F}(\boldsymbol{x};\xi)] \\ \text{s. t.} \quad & c_i(\boldsymbol{x}) = \mathbf{0}, \quad i \in \mathcal{E}, \\ & c_i(\boldsymbol{x}) \geq \mathbf{0}, \quad i \in \mathcal{I}, \end{aligned} \tag{1}$$

Shiji Zuo
ShanghaiTech University, Shanghai, China
E-mail: zuoshj@gmail.com

Xiao Wang
Sun Yat-sen University, Guangzhou, China
E-mail: wangx936@mail.sysu.edu.cn

Hao Wang
ShanghaiTech University, Shanghai, China
E-mail: wanghao1@shanghaitech.edu.cn

where ξ is a random variable in the probability space Ξ, and independent of $\boldsymbol{x}$, $\mathcal{E}\cup\mathcal{I}=[m]$ and $\mathcal{E}\cap\mathcal{I}=\emptyset$. For any fixed $\xi\in\Xi$, $\mathbf{F}(\cdot;\xi):\mathbb{R}^n\to\mathbb{R}$ and $c_i:\mathbb{R}^n\to\mathbb{R}, i\in\mathcal{E}\cup\mathcal{I}$, are continuously differentiable and possibly nonconvex. Such problems are prevalent in various fields, such as optimal control [5], PDE-constrained optimization [35], constrained maximum likelihood estimation [13,19], resource allocation [4,37], and constrained deep neural network training [32,34].

For general convex stochastic optimization with deterministic constraints, Xu [44] proposes a single-loop primal-dual stochastic gradient algorithm leveraging the linearized augmented Lagrangian and studies the convergence rate of the algorithm. Bollapragada et al. [8] consider problems with linear deterministic constraints and propose a double-loop augmented Lagrangian method (ALM) with an adaptive sampling strategy to control the accuracy of the stochastic gradient. For nonconvex stochastic optimization problems with deterministic constraints, there are also several classes of algorithms. Na et al. [31] propose a fully online stochastic sequential quadratic programming (SQP) method and establish the iteration complexity required to achieve ϵ-stationarity. Shi et al. [38] present a linearized ALM based on momentum [17] with an improved oracle complexity in order $\mathcal{O}(\epsilon^{-4})$ to find an ϵ-stationary point or an ϵ-KKT point under an extended variant of Mangasarian-Fromovitz Constraint Qualification (MFCQ), where the oracle complexity refers to the total number of stochastic gradient evaluations. Recently, Lu et al. [28] propose an algorithm based on a truncated recursive momentum scheme and establish the oracle complexity in order $\tilde{\mathcal{O}}(\epsilon^{-3})$. Within the framework of a proximal point method, ICPPC [9] and LCSPG [10] aim at inequality-constrained optimization, with $\mathcal{O}(\epsilon^{-4})$ oracle complexity to find an $\sqrt{\epsilon}$ type-I KKT point and $(\sqrt{\epsilon},\sqrt{\epsilon})$-KKT point, respectively. The SPD method [23] adopts a single-loop primal-dual algorithm framework for stochastic optimization with many deterministic constraints and owns the oracle complexity in order $\mathcal{O}(\epsilon^{-5})$ to find an ϵ-stationary point. The STEP algorithm [24] is targeted at a class of constrained optimization whose objective is a composition of two expected-value functions. The cubic-regularized primal-dual algorithms are studied in [40] for finding second-order stationary points in nonconvex equality-constrained optimization, with oracle complexity bounds established for problems with deterministic and stochastic objectives, respectively. It is worth noting that all above algorithms use fixed (deterministic) parameters during their iteration process.

Notably, the adaptive updating of merit/penalty parameters in deterministic constrained optimization is widely used in practical computations to alleviate the need for tedious manual tuning [39] and balance primal and dual residuals to ensure more stable convergence behaviors [45]. However, when solving stochastic optimization problems, the introduction of randomness and the uncontrollable magnitude of adaptive parameters can make the theoretical analysis intractable, e.g., considered in [21] for unconstrained optimization. Curtis et al. [2,14,15] investigate the convergence properties of a stochastic SQP (SSQP) algorithm based on the ℓ_1 merit function associated with an adaptive merit parameter, highlighting the importance of a merit parameter sequence that ultimately becomes sufficiently small while remaining bounded away from zero. Na et al. [29, 30] explore stochastic SQP algorithms with differentiable exact augmented Lagrangians, which incorporates second-order information from both objective and constraint functions. Assuming a strong linear independence constraint qualification (LICQ) condition, they prove the worst-case iteration complexity bound and almost sure convergence, respectively. Berahas et al. [3] tailor noisy function and gradient estimates to propose a stochastic step-search SQP (SS-SQP) to achieve a first order ϵ-stationary point with high probability. O'Neill et al. [33] propose a two stepsize stochastic SQP method with Adagrad-Norm [42] based adaptive stepsizes and provide an oracle complexity of $\tilde{\mathcal{O}}(\epsilon^{-4})$ for stationarity and $\tilde{\mathcal{O}}(\epsilon^{-1})$ for feasibility, respectively. Wang et al. [41] investigate a stochastic penalty method based on ℓ_2 penalty function with an adaptive penalty strategy for equality-constrained optimization and establish the oracle complexity to reach an ϵ-approximate stochastic critical point. As far as we know, however, there has not been any type of ALM that employs an adaptive parameter update strategy to solve the nonconvex constrained stochastic optimization problem (1). The design and analysis of such type of methods will cause additional difficulties since the stochasticity hidden in the adaptive penalty parameters may lead to uncontrolled behavior. The primary goal of this paper is to propose a type

of ALM that adopts the adaptive penalty parameter and enjoys a desirable high-probability oracle complexity to achieve an ϵ-KKT point of (1).

1.1 Contributions

In this paper, we study an adaptive single-loop stochastic penalty method (AdaSSP). AdaSSP adopts an adaptive mechanism to dynamically adjust the penalty parameter based on the level of constraint satisfaction, aiming to balance faster objective value reduction and stabler constraint violation reduction while eliminating the need for manual parameter tuning. Additionally, we employ clipped stochastic gradients with momentum to tackle the potentially large stochastic variance. Under the local LICQ, we establish the high-probability oracle complexity of AdaSSP to reach an ϵ-KKT point. Finally, numerical experiments on a quadratically constrained program and a multiclass Neyman-Pearson classification problem demonstrate the promising performances of the proposed method.

Detailed comparison with several closely related adaptive algorithms for nonconvex constrained stochastic optimization regarding oracle complexity is provided in Table 1.

Table 1 Comparison between algorithms with adaptive merit/penalty parameters. Here, we have $f(\boldsymbol{x}) = \mathbb{E}_\xi[\mathbf{F}(\boldsymbol{x};\xi)]$, and $\lambda_i \geq 0$ for any $i \in \mathcal{I}$. In SSQP, $\tau_{\min}$ is the merit parameter threshold in [14]. $\mathcal{SZO}$ and $\mathcal{SFO}$ denote stochastic zeroth- and first-order oracle, respectively. The "strong LICQ" represents that the Jacobian of constraint functions have singular values that are lower bounded away from zero over a set containing all iterates for all realizations of the random variable.

Algorithm	Problem type	Stationarity measure	Assumptions	Complexity
Algorithm 3.1 [41]	$\min_{\boldsymbol{x}\in\mathbb{R}^n} f(\boldsymbol{x})$ s.t. $c_{\mathcal{E}}(\boldsymbol{x}) = 0, i \in \mathcal{E}$	$\mathbb{E}[\|\nabla f(\boldsymbol{x}) + \nabla \boldsymbol{c}_{\mathcal{E}}(\boldsymbol{x})\boldsymbol{\lambda}_{\mathcal{E}}\|^2] \leq \epsilon^2$, $\mathbb{E}[\|c_{\mathcal{E}}(\boldsymbol{x})\| - \min_{\|\boldsymbol{s}\|\leq 1} \|c_{\mathcal{E}}(\boldsymbol{x}) + \nabla \boldsymbol{c}_{\mathcal{E}}(\boldsymbol{x})\boldsymbol{s}\|] \leq \epsilon$	-	$\mathcal{O}(\epsilon^{-7})$ $\mathcal{SFO}$
SSQP [14]	$\min_{\boldsymbol{x}\in\mathbb{R}^n} f(\boldsymbol{x})$ s.t. $c_i(\boldsymbol{x}) = 0, i \in \mathcal{E}$	$\mathbb{E}[\|\nabla f(\boldsymbol{x}) + \nabla \boldsymbol{c}_{\mathcal{E}}(\boldsymbol{x})\boldsymbol{\lambda}_{\mathcal{E}}\|] \leq \epsilon$, $\mathbb{E}[\sqrt{\|c_{\mathcal{E}}(\boldsymbol{x})\|_1}] \leq \epsilon$	strong LICQ	$\tilde{\mathcal{O}}(\epsilon^{-4})$ $\mathcal{SFO}$ $\mathcal{O}(\epsilon^{-4})$ $\mathcal{SFO}$ (if $\tau_{\min}$ known)
Algorithm 2.1 [33]	$\min_{\boldsymbol{x}\in\mathbb{R}^n} f(\boldsymbol{x})$ s.t. $c_{\mathcal{E}}(\boldsymbol{x}) = 0, i \in \mathcal{E}$	With high probability $\|\nabla f(\boldsymbol{x}) + \nabla \boldsymbol{c}_{\mathcal{E}}(\boldsymbol{x})\boldsymbol{\lambda}_{\mathcal{E}}\| \leq \epsilon,\ \|c_{\mathcal{E}}(\boldsymbol{x})\|_1 \leq \epsilon^4$	strong LICQ	$\tilde{\mathcal{O}}(\epsilon^{-4})$ $\mathcal{SFO}$
SS-SQP [3]	$\min_{\boldsymbol{x}\in\mathbb{R}^n} f(\boldsymbol{x})$ s.t. $c_i(\boldsymbol{x}) = 0, i \in \mathcal{E}$	With high probability $\|\nabla f(\boldsymbol{x}) + \nabla \boldsymbol{c}_{\mathcal{E}}(\boldsymbol{x})\boldsymbol{\lambda}_{\mathcal{E}}\| \leq \epsilon,\ \|c_{\mathcal{E}}(\boldsymbol{x})\| \leq \epsilon^2$	sub-exponential $\mathcal{SZO}$ noise, strong LICQ	$\mathcal{O}(\epsilon^{-6})$ $\mathcal{SZO}$ + $\mathcal{O}(\epsilon^{-4})$ $\mathcal{SFO}$
AdaSSP (this paper)	$\min_{\boldsymbol{x}\in\mathbb{R}^n} f(\boldsymbol{x})$ s.t. $c_i(\boldsymbol{x}) = 0, i \in \mathcal{E}$, $c_i(\boldsymbol{x}) \leq 0, i \in \mathcal{I}$	With high probability $\left(\begin{smallmatrix}\|\nabla \boldsymbol{c}_{\mathcal{E}}(\boldsymbol{x})c_{\mathcal{E}}(\boldsymbol{x}) + \nabla \boldsymbol{c}_{\mathcal{I}}(\boldsymbol{x})[c_{\mathcal{I}}(\boldsymbol{x})]_+\| \leq \epsilon,\\ \|c_{\mathcal{E}}(\boldsymbol{x})\| + \|[c_{\mathcal{I}}(\boldsymbol{x})]_+\| > \epsilon\end{smallmatrix}\right)$ or $\left(\begin{smallmatrix}\|\nabla f(\boldsymbol{x}) + \nabla \boldsymbol{c}(\boldsymbol{x})\boldsymbol{\lambda}\| \leq \epsilon, \|\boldsymbol{c}_{\mathcal{E}}(\boldsymbol{x})\| + \|[\boldsymbol{c}_{\mathcal{I}}(\boldsymbol{x})]_+\| \leq \epsilon,\\ \lambda_i = 0 \text{ if } c_i(\boldsymbol{x}) < -\epsilon \text{ for all } i \in \mathcal{I}\end{smallmatrix}\right)$	mean-squared smoothness	$\tilde{\mathcal{O}}(\epsilon^{-4})$ $\mathcal{SFO}$ (Theorem 1)
		With high probability $\|\nabla f(\boldsymbol{x}) + \nabla \boldsymbol{c}(\boldsymbol{x})\boldsymbol{\lambda}\| \leq \epsilon, \|\boldsymbol{c}_{\mathcal{E}}(\boldsymbol{x})\| + \|[\boldsymbol{c}_{\mathcal{I}}(\boldsymbol{x})]_+\| \leq \epsilon,$ $\lambda_i = 0$ if $c_i(\boldsymbol{x}) < -\epsilon$ for all $i \in \mathcal{I}$	mean-squared smoothness, local LICQ	$\tilde{\mathcal{O}}(\epsilon^{-4})$ $\mathcal{SFO}$ (Theorem 2)

1.2 Notation and organization

We use e to represent the base of the natural logarithm function $\log(\cdot)$ and $\|\cdot\|$ to denote the Euclidean norm of a vector without any specification. For brevity, we introduce $[k] := \{1, \ldots, k\}$ and $\xi^{[k]} := \{\xi^1, \ldots, \xi^k\}$ for any positive integer k. For any $u \in \mathbb{R}$, we define its positive part as $[u]_+ := \max\{0, u\}$. For any $\boldsymbol{u} \in \mathbb{R}^n$, $[\boldsymbol{u}]_+$ is referred to as a component-wise application of the operator $[\cdot]_+$. For $\boldsymbol{a}, \boldsymbol{b} \in \mathbb{R}^n$, $\boldsymbol{a} \succeq \boldsymbol{b}$ denotes that $a_i \geq b_i, i = 1, \ldots, n$, and for $A \in \mathbb{R}^{n\times n}$, $A \succeq 0$ denotes that A is positive semidefinite. The gradient of f at $\boldsymbol{x}$ is denoted by $\nabla f(\boldsymbol{x})$. Given a closed set $\mathcal{C} \subseteq \mathbb{R}^n$, the distance between $\boldsymbol{x}$ and $\mathcal{C}$ is referred to $\mathbf{d}(\boldsymbol{x}, \mathcal{C})$. Given random variables ξ and ζ, $\mathbb{E}_\xi[\cdot]$ represents the expectation w.r.t. ξ and $\mathbb{E}_\xi[\cdot|\zeta]$ represents the expectation w.r.t. ξ conditioned on ζ. The inner product of $\boldsymbol{x}, \boldsymbol{y} \in \mathbb{R}^n$ is denoted by $\langle \boldsymbol{x}, \boldsymbol{y}\rangle$. We define $\boldsymbol{c}_\mathcal{E} : \mathbb{R}^n \to \mathbb{R}^{|\mathcal{E}|}$ with components being $c_i(\cdot)$, $i \in \mathcal{E}$, and $\nabla \boldsymbol{c}_\mathcal{E} : \mathbb{R}^n \to \mathbb{R}^{n\times|\mathcal{E}|}$ with columns being $\nabla c_i(\cdot), i \in \mathcal{E}$, and so forth for $\boldsymbol{c}_\mathcal{I}$, $\nabla \boldsymbol{c}_\mathcal{I}$, $\boldsymbol{\lambda}_\mathcal{I}$ and $\boldsymbol{\lambda}_\mathcal{E}$. For convenience, we abbreviate $\boldsymbol{c}(\boldsymbol{x}^k)$ to $\boldsymbol{c}^k$, and similar abbreviations $\nabla \boldsymbol{c}^k, \nabla \boldsymbol{c}^k_\mathcal{E}$ and $\nabla \boldsymbol{c}^k_\mathcal{I}$ are used. For a positive integer k, we define $\sum_{t=k}^{k-1} \cdot = 0$ and $\prod_{t=k}^{k-1} \cdot = 1$.

For a general nonconvex constrained optimization problem, it is normally impossible to locate its global or even a local minimizer. Instead, main research stream focuses on a more trackable solution, KKT point. A feasible point $\boldsymbol{x}^*$ is called a *KKT point* of (1), if there exists a vector $\boldsymbol{\lambda}^* \in \mathbb{R}^m$ with $\boldsymbol{\lambda}^*_\mathcal{I} \succeq \mathbf{0}$ such that $\nabla f(\boldsymbol{x^*}) + \nabla \boldsymbol{c}(\boldsymbol{x}^*)\boldsymbol{\lambda}^* = 0$ and the complementary slackness condition holds: $\lambda_i c_i(\boldsymbol{x}^*) = 0, \forall i \in \mathcal{I}$. Under certain constraint qualification conditions [43], a local minimizer of the original problem is a KKT point. Throughout this paper, we assume that a KKT point of (1) exists and we aim to find an ϵ-KKT point defined as follows.

Definition 1 (ϵ-KKT point) Given a constant $\epsilon > 0$, a point $\boldsymbol{x} \in \mathbb{R}^n$ is called an ϵ-KKT point of (1), if there exists $\boldsymbol{\lambda} \in \mathbb{R}^m$ with $\boldsymbol{\lambda}_\mathcal{I} \succeq \mathbf{0}$ such that

$$\|\nabla f(\boldsymbol{x}) + \nabla \boldsymbol{c}(\boldsymbol{x})\boldsymbol{\lambda}\| \leq \epsilon, \quad \|\boldsymbol{c}(\boldsymbol{x})\| + \|[\boldsymbol{c}(\boldsymbol{x})]_+\| \leq \epsilon, \quad \lambda_i = 0 \text{ if } c_i(\boldsymbol{x}) \leq -\epsilon, \forall i \in \mathcal{I}.$$

The rest of this paper is organized as follows. In Section 2 we introduce an adaptive single-loop stochastic penalty method for solving problem (1). In Section 3 we present auxiliary lemmas that are essential to the subsequent theoretical analysis. In Sections 4 and 5 we investigate the high-probability oracle complexity of the proposed method to reach an ϵ-KKT point and the global convergence properties of the proposed method, when the penalty parameter sequence is unbounded and bounded, respectively. In Section 6, we present preliminary numerical results on two test problems. Finally, we give conclusions.

2 The AdaSSP method

The augmented Lagrangian (AL) function of problem (1) takes the standard form as described in [36]:

$$\mathcal{L}_\beta(\boldsymbol{x}, \boldsymbol{\lambda}) = f(\boldsymbol{x}) + \psi_\beta(\boldsymbol{c}(\boldsymbol{x}), \boldsymbol{\lambda}), \tag{2}$$

where $\boldsymbol{\lambda} \in \mathbb{R}^m$ is normally referred to as a dual variable, $\beta > 0$ is a penalty parameter, and

$$\psi_\beta(\boldsymbol{z}, \boldsymbol{\lambda}) := \sum_{i\in\mathcal{E}} [\lambda_i z_i + \frac{\beta}{2} z_i^2] + \frac{1}{2\beta} \sum_{i\in\mathcal{I}} [[\lambda_i + \beta z_i]_+^2 - \lambda_i^2].$$

It is easy to obtain that the gradient of $\mathcal{L}_\beta$ with respect to $\boldsymbol{x}$ can be expressed as

$$\nabla_{\boldsymbol{x}} \mathcal{L}_\beta(\boldsymbol{x}, \boldsymbol{\lambda}) = \nabla f(\boldsymbol{x}) + \sum_{i\in\mathcal{E}} (\lambda_i + \beta c_i(\boldsymbol{x}))\nabla c_i(\boldsymbol{x}) + \sum_{i\in\mathcal{I}} [\lambda_i + \beta c_i(\boldsymbol{x})]_+ \nabla c_i(\boldsymbol{x}). \tag{3}$$

Since the computation of exact gradient information of f is costly and sometimes even prohibitive, we assume that a stochastic gradient at any inquiry point can be accessed. More specifically, at current iterate $\boldsymbol{x}^k$, for $k \geq 1$, we compute the stochastic gradient of f through

$$\boldsymbol{g}^k = \alpha_k \boldsymbol{g}^{k-1} + (1-\alpha_k)\boldsymbol{G}^k + \alpha_k \boldsymbol{v}^k, \tag{4}$$

where

$$\begin{cases} \boldsymbol{G}^k = \min\left\{1, \frac{B_k}{\|\nabla \mathbf{F}(\boldsymbol{x}^k;\xi^k)\|}\right\} \nabla \mathbf{F}(\boldsymbol{x}^k;\xi^k), \\ \boldsymbol{v}^k = \mathbb{1}_{k\geq 2} \min\left\{1, \frac{D_k}{\|\nabla \mathbf{F}(\boldsymbol{x}^k;\xi^k) - \nabla \mathbf{F}(\boldsymbol{x}^{k-1};\xi^k)\|}\right\} (\nabla \mathbf{F}(\boldsymbol{x}^k;\xi^k) - \nabla \mathbf{F}(\boldsymbol{x}^{k-1};\xi^k)), \end{cases}$$

$B_k, D_k > 0$, $\alpha_k \in (0,1)$ for $k \geq 1$, ξ^k is a random sample generated from Ξ, $\boldsymbol{G}^k$ is the clipped gradient of $\mathbf{F}(\cdot;\xi^k)$ at $\boldsymbol{x}^k$ and $\boldsymbol{v}^k$ denotes the clipped momentum term. Here, we leverage the momentum technique to efficiently reduce variance with only one sample ξ per iteration, following the methodology of [17,46,38]. Meanwhile, inspired by [27,16], we employ a clipped stochastic gradient and a clipped momentum term to mitigate extreme events, thereby enabling us to establish high-probability theoretical results. Then we (approximately) solve the following subproblem to obtain $\boldsymbol{x}^{k+1}$:

$$\begin{aligned} \min_{\boldsymbol{x}\in\mathbb{R}^n} \quad & Q_k(\boldsymbol{x}) := \langle \boldsymbol{g}^k, \boldsymbol{x}-\boldsymbol{x}^k\rangle + \psi_{\beta_k}(\boldsymbol{c}^k + (\nabla \boldsymbol{c}^k)^{\mathrm{T}}(\boldsymbol{x}-\boldsymbol{x}^k), \boldsymbol{\lambda}^k) + \frac{h_k}{2}\|\boldsymbol{x}-\boldsymbol{x}^k\|^2 \\ \text{s. t.} \quad & \|\boldsymbol{x}-\boldsymbol{x}^k\| \leq \eta_k, \end{aligned} \tag{5}$$

where β_k is a positive penalty parameter and $h_k \in (0, H]$, for some $H > 0$, is a regularization parameter. Note that the objective of (5), $Q_k(\boldsymbol{x})$, is a stochastic approximation to $\mathcal{L}_{\beta_k}(\boldsymbol{x}, \boldsymbol{\lambda}^k)$ and $\boldsymbol{d}^k := \nabla Q_k(\boldsymbol{x}^k)$ is a stochastic approximation to $\nabla_{\boldsymbol{x}} \mathcal{L}_{\beta_k}(\boldsymbol{x}^k, \boldsymbol{\lambda}^k)$, where

$$\boldsymbol{d}^k = \boldsymbol{g}^k + \sum_{i\in\mathcal{E}} (\lambda_i + \beta c_i(\boldsymbol{x}^k))\nabla c_i(\boldsymbol{x}^k) + \sum_{i\in\mathcal{I}} [\lambda_i + \beta c_i(\boldsymbol{x}^k)]_+ \nabla c_i(\boldsymbol{x}^k).$$

For theoretical analysis purposes, it is sufficient to solve (5) until the following condition is satisfied:

$$-\frac{1}{2}\min\{\eta_k, \frac{\|\boldsymbol{d}^k\|}{mG_k^2\beta_k + H}\}\|\boldsymbol{d}^k\| \geq Q_k(\boldsymbol{x}^{k+1}) - Q_k(\boldsymbol{x}^k), \tag{6}$$

where $G_k = \max_{i\in[m]}\{\|\nabla c_i(\boldsymbol{x}^k)\|\}$. This condition (6) can be efficiently satisfied using the approach detailed in Section 6. For the update of $\boldsymbol{\lambda}^k$, we employ the following scheme [38,23,24] by computing

$$\boldsymbol{\lambda}^{k+1} = \boldsymbol{\lambda}^k + \rho_k \begin{bmatrix} \boldsymbol{c}_{\mathcal{E}}(\boldsymbol{x}^{k+1}) \\ \max\{\boldsymbol{c}_{\mathcal{I}}(\boldsymbol{x}^{k+1}), -\boldsymbol{\lambda}^k_{\mathcal{I}}/\beta_k\} \end{bmatrix}, \tag{7}$$

with $\rho_k > 0$.

In Algorithm 1, we present main steps of an adaptive single-loop stochastic penalty method, AdaSSP, to solve problem (1). In this method, the penalty parameters $\{\beta_k\}$ are updated adaptively based on the iterative progress. Due to the stochastic nature of $\{\boldsymbol{g}^k\}_{k\geq 1}$, the penalty parameters are inherently random. This character significantly differs from many stochastic approximation methods for constrained optimization, which use predetermined penalty parameters. Motivated by [6,7] we define $\boldsymbol{V}^{k+1} := \max\{\boldsymbol{c}_{\mathcal{I}}(\boldsymbol{x}^{k+1}), -\boldsymbol{\lambda}^k_{\mathcal{I}}/\beta_k\}$ for $k \geq 2$, with $\boldsymbol{V}^1 := \max\{\boldsymbol{c}_{\mathcal{I}}(\boldsymbol{x}^1), \boldsymbol{0}\}$, to measure the violation and complementary slackness of inequality

constraints at $\boldsymbol{x}^{k+1}$. The penalty parameter stays the same when an *average* of constraint violations at a group of history iterates is improved, i.e.,

$$k=1 \quad \text{or} \quad \frac{1}{\lceil k/2 \rceil} \sum_{t=\lfloor k/2 \rfloor+1}^{k} \Big(\|\boldsymbol{c}_{\mathcal{E}}(\boldsymbol{x}^{t+1})\| + \|\boldsymbol{V}^{t+1}\| \Big) \\ \leq \frac{\tau}{\lceil (k-1)/2 \rceil} \sum_{t=\lfloor (k-1)/2 \rfloor+1}^{k-1} \Big(\|\boldsymbol{c}_{\mathcal{E}}(\boldsymbol{x}^{t+1})\| + \|\boldsymbol{V}^{t+1}\| \Big), \tag{8}$$

where $\tau \in (0,1)$.

Algorithm 1 AdaSSP

Input: $\boldsymbol{x}^1 \in \mathbb{R}^n$, $\beta_1, \Gamma, \vartheta, H > 0$ $\tau \in (0,1)$, and parameters $\{B_k, \eta_k > 0\}, \{\rho_k \in (0, \beta_1]\}, \{\alpha_k \in (0,1),\ h_k \in (0, H]\}$.
Let $\boldsymbol{\lambda}^1 = \boldsymbol{0}$, $\boldsymbol{g}^0 = \boldsymbol{0}$, and $\boldsymbol{V}^1 = \max\{\boldsymbol{c}_{\mathcal{I}}(\boldsymbol{x}^1), \boldsymbol{0}\}$.
For $k = 1, 2, \ldots$ **do**
  Calculate $\boldsymbol{g}^k$ through (4).
  Calculate $\boldsymbol{x}^{k+1}$ by solving (5) such that (6) is fulfilled.
  Calculate $\boldsymbol{\lambda}^{k+1}$ through (7).
  Calculate $\boldsymbol{V}^{k+1} = \max\{\boldsymbol{c}_{\mathcal{I}}(\boldsymbol{x}^{k+1}), -\boldsymbol{\lambda}^k_{\mathcal{I}}/\beta_k\}$.
  If (8) is satisfied, set $\beta_{k+1} = \beta_k$; otherwise, set $\beta_{k+1} = (\beta_k^{1/\vartheta} + \Gamma^{1/\vartheta})^{\vartheta}$.
End For

To simplify notation, we introduce $\tilde{\boldsymbol{\lambda}}^1 = \boldsymbol{\lambda}^1$ and $\tilde{\boldsymbol{\lambda}}^k \in \mathbb{R}^m, k \geq 2$, defined by

$$\tilde{\boldsymbol{\lambda}}^k = \begin{bmatrix} \beta_{k-1}\boldsymbol{c}_{\mathcal{E}}(\boldsymbol{x}^k) + \boldsymbol{\lambda}^{k-1}_{\mathcal{E}} \\ [\beta_{k-1}\boldsymbol{c}_{\mathcal{I}}(\boldsymbol{x}^k) + \boldsymbol{\lambda}^{k-1}_{\mathcal{I}}]_+ \end{bmatrix}. \tag{9}$$

In theory, $\tilde{\boldsymbol{\lambda}}^k$ will play a role as the vector of multipliers at kth iteration. The lemma below provides some insights on how $\boldsymbol{V}^k$ affects the feasibility and complementary slackness error at $\boldsymbol{x}^k$.

Lemma 1 *For any $k \geq 2$ and $\epsilon \geq 0$ such that $\|\boldsymbol{c}_{\mathcal{E}}(\boldsymbol{x}^k)\| + \|\boldsymbol{V}^k\| \leq \epsilon$, it holds that $\|\boldsymbol{c}_{\mathcal{E}}(\boldsymbol{x}^k)\| + \|[\boldsymbol{c}_{\mathcal{I}}(\boldsymbol{x}^k)]_+\| \leq \epsilon$, and $\tilde{\lambda}^k_i = 0$ if $c_i(\boldsymbol{x}^k) \leq -\epsilon$ for all $i \in \mathcal{I}$.*

Proof From $\boldsymbol{\lambda}^1 = \boldsymbol{0}, \rho_k \in (0, \beta_k)$ and (7), we obtain that $\boldsymbol{\lambda}^k_{\mathcal{I}} \succeq \boldsymbol{0}$ for $k \geq 1$ by induction. Then we have

$$|\boldsymbol{V}^k| = |\max\{\boldsymbol{c}_{\mathcal{I}}(\boldsymbol{x}^k), -\boldsymbol{\lambda}^{k-1}_{\mathcal{I}}/\beta_{k-1}\}| \succeq [\boldsymbol{c}_{\mathcal{I}}(\boldsymbol{x}^k)]_+,$$

which implies

$$\|\boldsymbol{c}_{\mathcal{E}}(\boldsymbol{x}^k)\| + \|[\boldsymbol{c}_{\mathcal{I}}(\boldsymbol{x}^k)]_+\| \leq \|\boldsymbol{c}_{\mathcal{E}}(\boldsymbol{x}^k)\| + \|\boldsymbol{V}^k\| \leq \epsilon.$$

Moreover, for any $i \in \mathcal{I}$ it holds that $\max\{c_i(\boldsymbol{x}^k), -\lambda^{k-1}_i/\beta_{k-1}\} \geq -\epsilon$. If $c_i(\boldsymbol{x}^k) \leq -\epsilon$, $-\lambda^{k-1}_i/\beta_{k-1} \geq -\epsilon \geq c_i(\boldsymbol{x}^k)$ and thus $\beta_{k-1}c_i(\boldsymbol{x}^k) + \lambda^{k-1}_i \leq 0$. Hence, it yields that $\tilde{\lambda}^k_i = [\beta_{k-1}c_i(\boldsymbol{x}^k) + \lambda^{k-1}_i]_+ = 0$, which completes the proof. □

Let $(\Omega, \mathcal{F}, \mathbb{P})$ be a probability space defined on the stochastic process $\{\xi_k\}_{k\geq 1}$. Then, with fixed initial parameters, each realization of a run of the algorithm generating $\{\boldsymbol{x}^{k+1}, \beta^{k+1}\}_{k\geq 1}$ is associated with some $\omega \in \Omega$. For simplicity of notation, we omit this dependence in subsequent discussions. We introduce the following set of random variables and the corresponding σ-algebra as

$$\xi^{[k]} := \{\xi^1, \xi^2, \ldots, \xi^k\} \quad \text{and} \quad \mathcal{F}_k = \sigma\big(\xi^{[k]}\big), \quad k \geq 1.$$

From the framework of Algorithm 1, it follows that for any $k \geq 1$, all $\boldsymbol{x}^{k+1}$ and β^{k+1} are $\mathcal{F}_k$-measurable.

For simplicity, we introduce the notations

$$\begin{aligned}
\boldsymbol{e}^k &:= \boldsymbol{g}^k - \nabla f(\boldsymbol{x}^k), \\
\boldsymbol{Z}_k &:= \boldsymbol{v}^k - \mathbb{1}_{k\geq 2}\left(\nabla f(\boldsymbol{x}^k) - \nabla f(\boldsymbol{x}^{k-1})\right) = \boldsymbol{Z}_k^u + \boldsymbol{Z}_k^b \text{ with} \\
&\quad \boldsymbol{Z}_k^u := \boldsymbol{v}^k - \mathbb{E}_{\xi^{[k]}}[\boldsymbol{v}^k] \text{ and } \boldsymbol{Z}_k^b := \mathbb{E}_{\xi^{[k]}}[\boldsymbol{v}^k] - \mathbb{1}_{k\geq 2}\left(\nabla f(\boldsymbol{x}^k) - \nabla f(\boldsymbol{x}^{k-1})\right), \\
\boldsymbol{\theta}_k &:= \boldsymbol{G}^k - \nabla f(\boldsymbol{x}^k) = \boldsymbol{\theta}_k^u + \boldsymbol{\theta}_k^b \text{ with} \\
&\quad \boldsymbol{\theta}_k^u := \boldsymbol{G}^k - \mathbb{E}_{\xi^{[k]}}[\boldsymbol{G}^k] \text{ and } \boldsymbol{\theta}_k^b := \mathbb{E}_{\xi^{[k]}}[\boldsymbol{G}^k] - \nabla f(\boldsymbol{x}^k),
\end{aligned}$$

where $\boldsymbol{e}^k$ denotes error of $\boldsymbol{g}^k$, $\boldsymbol{Z}_k$ denotes error of the clipped momentum, and $\boldsymbol{\theta}_k$ refers to error of the clipped gradient. In the lemma below, we represent the error $\boldsymbol{e}^k$ in terms of $\nabla f(\boldsymbol{x}^1)$, $\{\boldsymbol{Z}_s\}_{s\leq k}$, and $\{\boldsymbol{\theta}_s\}_{s\leq k}$. This form of representation is widely used in the related literature [20,27] and serves as a useful tool for analyzing the rate at which $\boldsymbol{e}^k$ diminishes.

Lemma 2 *For any $k \geq 1$, it holds that*

$$\boldsymbol{e}^k = -\prod_{i=1}^{k}\alpha_i \nabla f(\boldsymbol{x}^1) + \sum_{s=1}^{k}\left(\prod_{i=s}^{k}\alpha_i \boldsymbol{Z}_s\right) + \sum_{s=1}^{k}\left((1-\alpha_s)\prod_{i=s+1}^{k}\alpha_i \boldsymbol{\theta}_s\right), \tag{10}$$

where $\prod_{i=k+1}^{k}\alpha_i = 1$ in convention.

Proof Using the definition of $\boldsymbol{\theta}_k, \boldsymbol{Z}_k$ and $\boldsymbol{e}^k$ we obtain

$$\boldsymbol{e}^1 = \alpha_1 \boldsymbol{g}^0 + (1-\alpha_1)\boldsymbol{G}^1 - \nabla f(\boldsymbol{x}^1) = -\alpha_1 \nabla f(\boldsymbol{x}^1) + (1-\alpha_1)\boldsymbol{\theta}_1.$$

Suppose (10) holds for $k-1 \geq 1$, then

$$\begin{aligned}
\boldsymbol{e}^k &= \alpha_k \boldsymbol{g}^{k-1} + (1-\alpha_k)\boldsymbol{G}^k + \alpha_k \boldsymbol{v}^k - \nabla f(\boldsymbol{x}^k) \\
&= \alpha_k \boldsymbol{e}^{k-1} + \alpha_k \boldsymbol{Z}_k + (1-\alpha_k)\boldsymbol{\theta}_k \\
&= -\prod_{i=1}^{k}\alpha_i \nabla f(\boldsymbol{x}^1) + \sum_{s=1}^{k-1}\left(\prod_{i=s}^{k}\alpha_i \boldsymbol{Z}_s\right) + \sum_{s=1}^{k-1}\left((1-\alpha_s)\prod_{i=s+1}^{k}\alpha_i \boldsymbol{\theta}_s\right) \\
&\qquad + \alpha_k \boldsymbol{Z}_k + (1-\alpha_k)\boldsymbol{\theta}_k,
\end{aligned}$$

which completes the proof by induction. □

Moreover, it follows form [27, Lemma 10] that

$$\left\|\sum_{s=1}^{k}\prod_{i=s}^{k}\alpha_i \boldsymbol{Z}_s\right\| \leq \left|\sum_{s=1}^{k}\bar{\boldsymbol{U}}_s^k\right| + \sqrt{2\left|\sum_{s=1}^{k}\bar{\boldsymbol{R}}_s^k\right|} + \sqrt{2\sum_{s=1}^{k}\mathbb{E}_{\xi^s}\left[\left\|\prod_{i=s}^{k}\alpha_i \boldsymbol{Z}_s^u\right\|^2\right]} + \left\|\sum_{s=1}^{k}\prod_{i=s}^{k}\alpha_i \boldsymbol{Z}_s^b\right\| \tag{11}$$

and

$$\left\|\sum_{s=1}^{k}(1-\alpha_s)\prod_{i=s+1}^{k}\alpha_i \boldsymbol{\theta}_s\right\| \leq \left|\sum_{s=1}^{k}\tilde{\boldsymbol{U}}_s^k\right| + \sqrt{2\left|\sum_{s=1}^{k}\tilde{\boldsymbol{R}}_s^k\right|}$$

$$+\sqrt{2\sum_{s=1}^{k}\mathbb{E}_{\xi^s}\left[\left\|(1-\alpha_s)\prod_{i=s+1}^{k}\alpha_i\boldsymbol{\theta}_s^u\right\|^2\right]}+\left\|\sum_{s=1}^{k}(1-\alpha_s)\prod_{i=s+1}^{k}\alpha_i\boldsymbol{\theta}_s^b\right\|, \tag{12}$$

where

$$\bar{\boldsymbol{U}}_s^k=\begin{cases}\mathbf{0}, & \text{if } s=0 \text{ or } \sum_{j=1}^{s-1}\prod_{i=j}^{k}\alpha_i\boldsymbol{Z}_j^u=\mathbf{0},\\ \operatorname{sign}\left(\sum_{j=1}^{s-1}\bar{\boldsymbol{U}}_j^k\right)\frac{\langle\sum_{j=1}^{s-1}\prod_{i=j}^{k}\alpha_i\boldsymbol{Z}_j^u,\prod_{i=s}^{k}\alpha_i\boldsymbol{Z}_s^u\rangle}{\left\|\sum_{j=1}^{s-1}\prod_{i=j}^{k}\alpha_i\boldsymbol{Z}_j^u\right\|}, & \text{otherwise;}\end{cases}$$

$$\tilde{\boldsymbol{U}}_s^k=\begin{cases}\mathbf{0}, & \text{if } s=0 \text{ or } \sum_{j=1}^{s-1}(1-\alpha_j)\prod_{i=j+1}^{k}\alpha_i\boldsymbol{\theta}_j^u=\mathbf{0},\\ \operatorname{sign}\left(\sum_{j=1}^{s-1}\tilde{\boldsymbol{U}}_j^k\right)\frac{\langle\sum_{j=1}^{s-1}(1-\alpha_j)\prod_{i=j+1}^{k}\alpha_i\boldsymbol{\theta}_j^u,(1-\alpha_s)\prod_{i=s+1}^{k}\alpha_i\boldsymbol{\theta}_s^u\rangle}{\left\|\sum_{j=1}^{s-1}(1-\alpha_j)\prod_{i=j+1}^{k}\alpha_i\boldsymbol{\theta}_j^u\right\|}, & \text{otherwise,}\end{cases}$$

and

$$\begin{aligned}\bar{\boldsymbol{R}}_s^k&=\left\|\prod_{i=s}^{k}\alpha_i\boldsymbol{Z}_s^u\right\|^2-\mathbb{E}_{\xi^s}\left[\left\|\prod_{i=s}^{k}\alpha_i\boldsymbol{Z}_s^u\right\|^2\right],\\ \tilde{\boldsymbol{R}}_s^k&=\left\|(1-\alpha_s)\prod_{i=s+1}^{k}\alpha_i\boldsymbol{\theta}_s^u\right\|^2-\mathbb{E}_{\xi^s}\left[\left\|(1-\alpha_s)\prod_{i=s+1}^{k}\alpha_i\boldsymbol{\theta}_s^u\right\|^2\right].\end{aligned} \tag{13}$$

By [16, Lemma 10], $\{\bar{\boldsymbol{U}}_s^k\}$ and $\{\tilde{\boldsymbol{R}}_s^k\}$ are martingale difference sequences and

$$\left|\bar{\boldsymbol{U}}_s^k\right|\le\|\prod_{i=s}^{k}\alpha_i\boldsymbol{Z}_s^u\|,\quad \left|\tilde{\boldsymbol{U}}_s^k\right|\le\|(1-\alpha_s)\prod_{i=s+1}^{k}\alpha_i\boldsymbol{\theta}_s^u\|. \tag{14}$$

3 Preliminaries

In this section, we provide auxiliary lemmas to prepare the oracle complexity and global convergence analysis in subsequent sections. We first lay out assumptions that are used throughout the remainder of this paper.

Assumption 1 *Let $\mathcal{X}\subseteq\mathbb{R}^n$ be a compact convex set containing all the iterates $\{\boldsymbol{x}^k\}$. Functions f and $c_i, i\in[m]$ are continuously differentiable with L-Lipschitz continuous gradients over $\mathcal{X}$. The objective function value of (1) is lower bounded by C^*. Moreover, there exist $C, G>0$ such that for any $\boldsymbol{x}\in\mathcal{X}$, $|c_i(\boldsymbol{x})|\le C, \forall i\in\mathcal{E}$; $c_i(\boldsymbol{x})\le C, \forall i\in\mathcal{I}$; $\|\nabla f(\boldsymbol{x})\|\le G$, and $\|\nabla c_i(\boldsymbol{x})\|\le G, \forall i\in[m]$.*

Assumption 2 *For any $\boldsymbol{x}\in\mathcal{X}$, $\mathbb{E}_\xi[\nabla\mathbf{F}(\boldsymbol{x};\xi)]=\nabla f(\boldsymbol{x})$ and there exists $\sigma>0$ such that $\mathbb{E}_\xi[\|\nabla\mathbf{F}(\boldsymbol{x};\xi)-\nabla f(\boldsymbol{x})\|^2]\le\sigma^2$.*

Assumption 3 (Mean-squared smoothness) *$\mathbf{F}(\cdot;\xi)$ is continuously differentiable for each $\xi\in\Xi$ and there exists $L>0$ such that for any $\boldsymbol{u},\boldsymbol{v}\in\mathcal{X}$, $\mathbb{E}_\xi[\|\nabla\mathbf{F}(\boldsymbol{u};\xi)-\nabla\mathbf{F}(\boldsymbol{v};\xi)\|^2]\le L^2\|\boldsymbol{u}-\boldsymbol{v}\|^2$.*

Remark 1 In the literature, the assumption that $\mathbf{F}(\cdot;\xi)$ has L-Lipschitz continuous gradients almost surely is usually used, such as in [17,25,27]. Sub-exponential error assumptions, known as light tail distribution, are also often posed to induce the high probability result [3,11,22]. However, thanks to the incorporation of clipping technique in our method, we can realize this goal when only assuming the heavy tail variance boundedness (see Assumption 2) and mean-squared smoothness (see Assumption 3).

The lemma below straightly follows from [38, Lemmas 1, 2, and 4], which demonstrates upper bounds of $\boldsymbol{\lambda}^k$ and the smoothness of $\mathcal{L}_\beta\left(\boldsymbol{x},\boldsymbol{\lambda}^k\right)$ in $\boldsymbol{x}$.

Lemma 3 *Under Assumption 1, it holds that for any $k \geq 1$,*

$$|\lambda_i^k| \leq C\sum_{j=1}^{k-1}\rho_j,\ \forall i\in\mathcal{E};\ 0\leq\lambda_i^k\leq C\sum_{j=1}^{k-1}\rho_j,\ \forall i\in\mathcal{I};\ |\lambda_i^{k+1}-\lambda_i^k|\leq\rho_k\tilde{C},\ \forall i\in\mathcal{E}\cup\mathcal{I},$$

where $\tilde{C}:=\max(\frac{C\sum_{k=1}^{\infty}\rho_k}{\beta_1},C)$ and $\sum_{j=1}^{0}\rho_j:=0$. Moreover, it holds that, for any $\beta\geq\beta_1$ and $\boldsymbol{u},\boldsymbol{v}\in\mathcal{X}$,

$$\|\nabla_{\boldsymbol{x}}\mathcal{L}_\beta(\boldsymbol{u},\boldsymbol{\lambda}^k)-\nabla_{\boldsymbol{x}}\mathcal{L}_\beta(\boldsymbol{v},\boldsymbol{\lambda}^k)\|\leq L_\beta\|\boldsymbol{u}-\boldsymbol{v}\|,\quad k\geq 1, \tag{15}$$

where $L_\beta:=\beta\tilde{L}$ with $\tilde{L}:=\frac{L+mCL\sum_{k=1}^{\infty}\rho_k}{\beta_1}+mG^2+mCL$.

The following lemma provides a bound for $\boldsymbol{d}^k$, which is a stochastic approximation to $\nabla_{\boldsymbol{x}}\mathcal{L}_{\beta_k}(\boldsymbol{x}^k,\boldsymbol{\lambda}^k)$.

Lemma 4 *Under Assumption 1, it holds that for any $k\geq 1$,*

$$\min\{\eta_k,\frac{\|\boldsymbol{d}^k\|}{mG^2\beta_k+H}\}\|\boldsymbol{d}^k\|\leq 2(\mathcal{L}_{\beta_k}(\boldsymbol{x}^k,\boldsymbol{\lambda}^k)-\mathcal{L}_{\beta_k}(\boldsymbol{x}^{k+1},\boldsymbol{\lambda}^k))+\eta_k^2L_{\beta_k}+2\eta_k\|\boldsymbol{e}^k\|. \tag{16}$$

Proof From Lemma 3, (6) and $\|\boldsymbol{x}^{k+1}-\boldsymbol{x}^k\|\leq\eta_k$, we obtain

$$\begin{aligned}
\mathcal{L}_{\beta_k}(\boldsymbol{x}^{k+1},\boldsymbol{\lambda}^k)-\mathcal{L}_{\beta_k}(\boldsymbol{x}^k,\boldsymbol{\lambda}^k)&\leq\langle\nabla_{\boldsymbol{x}}\mathcal{L}_{\beta_k}(\boldsymbol{x}^k,\boldsymbol{\lambda}^k),\boldsymbol{x}^{k+1}-\boldsymbol{x}^k\rangle+\frac{L_{\beta_k}}{2}\|\boldsymbol{x}^{k+1}-\boldsymbol{x}^k\|^2\\
&=\langle\boldsymbol{d}^k-\boldsymbol{e}^k,\boldsymbol{x}^{k+1}-\boldsymbol{x}^k\rangle+\frac{L_{\beta_k}}{2}\|\boldsymbol{x}^{k+1}-\boldsymbol{x}^k\|^2\\
&\leq\langle\boldsymbol{d}^k,\boldsymbol{x}^{k+1}-\boldsymbol{x}^k\rangle+\eta_k\|\boldsymbol{e}^k\|+\frac{L_{\beta_k}\eta_k^2}{2}\\
&\leq Q_k(\boldsymbol{x}^{k+1})-Q_k(\boldsymbol{x}^k)+\eta_k\|\boldsymbol{e}^k\|+\frac{L_{\beta_k}\eta_k^2}{2}\\
&\leq-\frac{1}{2}\min\{\eta_k,\frac{\|\boldsymbol{d}^k\|}{mG_k^2\beta_k+H}\}\|\boldsymbol{d}^k\|+\eta_k\|\boldsymbol{e}^k\|+\frac{L_{\beta_k}\eta_k^2}{2}\\
&\leq-\frac{1}{2}\min\{\eta_k,\frac{\|\boldsymbol{d}^k\|}{mG^2\beta_k+H}\}\|\boldsymbol{d}^k\|+\eta_k\|\boldsymbol{e}^k\|+\frac{L_{\beta_k}\eta_k^2}{2},
\end{aligned}$$

where the third inequality is due to the convexity of $Q_k(\cdot)$ with $\nabla Q_k(\boldsymbol{x}^k)=\boldsymbol{d}^k$ and last inequality is from $G_k=\max_{i\in[m]}\{\|\nabla c_i(\boldsymbol{x}^k)\|\}\leq G$ by Assumption 1. Thus the conclusion is yielded after arranging the terms. □

The next lemma shows an upper bound on the accumulation of differences between the augmented Lagrangian function values at two consecutive iterates.

Lemma 5 *Suppose that Assumption 1 holds, then for any $K \geq 1$,*

$$\sum_{k=1}^{K}(\mathcal{L}_{\beta_k}(\boldsymbol{x}^k,\boldsymbol{\lambda}^k)-\mathcal{L}_{\beta_k}(\boldsymbol{x}^{k+1},\boldsymbol{\lambda}^k)) \leq M_1 + \frac{mC^2\beta_{K+1}}{2}, \tag{17}$$

where $M_1 := \mathcal{L}_{\beta_1}(\boldsymbol{x}^1,\boldsymbol{\lambda}^1) - C^* - \frac{mC^2\beta_1}{2} + 2m\tilde{C}^2\sum_{k=1}^{\infty}\rho_k$.

Proof Firstly, it is easy to obtain from the definitions of ψ_β and $\mathcal{L}_\beta$ that

$$\psi_{\beta_{k+1}}(\boldsymbol{c}^{k+1},\boldsymbol{\lambda}^{k+1}) - \psi_{\beta_k}(\boldsymbol{c}^{k+1},\boldsymbol{\lambda}^{k+1}) \leq \frac{\beta_{k+1}-\beta_k}{2}mC^2, \tag{18}$$

$$\mathcal{L}_{\beta_k}(\boldsymbol{x}^{k+1},\boldsymbol{\lambda}^{k+1}) - \mathcal{L}_{\beta_k}(\boldsymbol{x}^{k+1},\boldsymbol{\lambda}^{k}) \leq m\rho_k\tilde{C}^2. \tag{19}$$

Then combining (18) and (19) gives

$$\begin{aligned}
&\sum_{k=1}^{K}[\mathcal{L}_{\beta_k}(\boldsymbol{x}^k,\boldsymbol{\lambda}^k)-\mathcal{L}_{\beta_k}(\boldsymbol{x}^{k+1},\boldsymbol{\lambda}^k)]\\
&=\sum_{k=1}^{K}[\mathcal{L}_{\beta_k}(\boldsymbol{x}^{k+1},\boldsymbol{\lambda}^{k+1})-\mathcal{L}_{\beta_k}(\boldsymbol{x}^{k+1},\boldsymbol{\lambda}^k)\\
&\quad+\psi_{\beta_{k+1}}(\boldsymbol{c}^{k+1},\boldsymbol{\lambda}^{k+1})-\psi_{\beta_k}(\boldsymbol{c}^{k+1},\boldsymbol{\lambda}^{k+1})+\mathcal{L}_{\beta_k}(\boldsymbol{x}^{k},\boldsymbol{\lambda}^{k})-\mathcal{L}_{\beta_{k+1}}(\boldsymbol{x}^{k+1},\boldsymbol{\lambda}^{k+1})]\\
&\leq\sum_{k=1}^{K}[\mathcal{L}_{\beta_k}(\boldsymbol{x}^k,\boldsymbol{\lambda}^k)-\mathcal{L}_{\beta_{k+1}}(\boldsymbol{x}^{k+1},\boldsymbol{\lambda}^{k+1})+\frac{\beta_{k+1}-\beta_k}{2}mC^2+m\rho_k\tilde{C}^2]\\
&\leq\mathcal{L}_{\beta_1}(\boldsymbol{x}^1,\boldsymbol{\lambda}^1)-\mathcal{L}_{\beta_{K+1}}(\boldsymbol{x}^{K+1},\boldsymbol{\lambda}^{K+1})+\frac{\beta_{K+1}-\beta_1}{2}mC^2+m\tilde{C}^2\sum_{k=1}^{\infty}\rho_k.
\end{aligned} \tag{20}$$

Moreover, we can upper bound $\mathcal{L}_{\beta_1}(\boldsymbol{x}^1,\boldsymbol{\lambda}^1)-\mathcal{L}_{\beta_{K+1}}(\boldsymbol{x}^{k+1},\boldsymbol{\lambda}^{K+1})$ by

$$\begin{aligned}
&\mathcal{L}_{\beta_1}(\boldsymbol{x}^1,\boldsymbol{\lambda}^1)-\mathcal{L}_{\beta_{K+1}}(\boldsymbol{x}^{k+1},\boldsymbol{\lambda}^{K+1})\\
&\leq\mathcal{L}_{\beta_1}(\boldsymbol{x}^1,\boldsymbol{\lambda}^1)-C^*-\Bigg[\sum_{i\in\mathcal{E}}[\lambda_i^{K+1}c_i(\boldsymbol{x}^{k+1})+\frac{\beta_{K+1}}{2}c_i^2(\boldsymbol{x}^{k+1})]\\
&\qquad+\frac{1}{2\beta_{K+1}}\sum_{i\in\mathcal{I}}\left([\lambda_i^{K+1}+\beta_{K+1}c_i(\boldsymbol{x}^{k+1})]_+^2-(\lambda_i^{K+1})^2\right)\Bigg]\\
&\leq\mathcal{L}_{\beta_1}(\boldsymbol{x}^1,\boldsymbol{\lambda}^1)-C^*+\sum_{i\in\mathcal{E}}|\lambda_i^{K+1}c_i(\boldsymbol{x}^{k+1})|+\sum_{i\in\mathcal{I}}\frac{(\lambda_i^{K+1})^2}{2\beta_{K+1}}\\
&\leq\mathcal{L}_{\beta_1}(\boldsymbol{x}^1,\boldsymbol{\lambda}^1)-C^*+\sum_{i\in\mathcal{E}}C^2\sum_{j=1}^{K}\rho_j+\frac{1}{2\beta_1}\sum_{i\in\mathcal{I}}(C\sum_{j=1}^{K}\rho_j)^2\\
&\leq\mathcal{L}_{\beta_1}(\boldsymbol{x}^1,\boldsymbol{\lambda}^1)-C^*+m\tilde{C}^2\sum_{j=1}^{\infty}\rho_j,
\end{aligned} \tag{21}$$

where the third inequality holds by Lemma 3. Thus the result follows after plugging (21) into (20). □

Motivated by [38, Lemma 7], we give the following lemma that links gradients of the augmented Lagrangian function in $\boldsymbol{x}$ with those of the minimization problem

$$\min_{\boldsymbol{x}} \frac{1}{2}\|\boldsymbol{c}_{\mathcal{E}}(\boldsymbol{x})\|^2 + \frac{1}{2}\|[\boldsymbol{c}_{\mathcal{I}}(\boldsymbol{x})]_+\|^2.$$

Lemma 6 *Suppose that Assumption 1 holds, then for any $k \geq 1$, we have*

$$\|\nabla \boldsymbol{c}_{\mathcal{E}}(\boldsymbol{x}^{k+1})\boldsymbol{c}_{\mathcal{E}}(\boldsymbol{x}^{k+1}) + \nabla \boldsymbol{c}_{\mathcal{I}}(\boldsymbol{x}^{k+1})[\boldsymbol{c}_{\mathcal{I}}(\boldsymbol{x}^{k+1})]_+\| \leq \frac{\|\nabla_{\boldsymbol{x}}\mathcal{L}_{\beta_k}(\boldsymbol{x}^{k+1}, \boldsymbol{\lambda}^k)\|}{\beta_k} + \frac{M_2}{\beta_k},$$

where $M_2 := G + mCG\sum_{k=1}^{\infty}\rho_k$.

To proceed we need an analyzing tool provided in the following lemma, which can be derived following the scratch of [18, Lemma 15].

Lemma 7 *Given constants $a \in [0, \infty)$ and $b \in [0, 1)$ with $b < a$, it holds that $\sum_{j=1}^{k} j^{-a}\prod_{i=j+1}^{k}(1 - i^{-b}) \leq \kappa(a, b)k^{b-a}$, where $\kappa(a, b) := 2\exp\left(\frac{1}{1-b}\right) + \left(\left\lceil \max\{a^{\frac{1}{1-b}}, \left(\frac{a-b}{2}\right)^{\frac{1}{1-b}}\}\right\rceil - 1\right)\exp\left(\frac{(a-1)^{1-b}+1+b-a}{1-b}\right)(a-b)^{\frac{a-b}{1-b}}$.*

Proof See Appendix B. □

From Assumptions 2 and 3 as well as $\|\boldsymbol{x}^k - \boldsymbol{x}^{k-1}\| \leq \eta_{k-1}$, it yields that for any $k \geq 2$, the variance of $\nabla\mathbf{F}(\boldsymbol{x}^k; \xi^k) - \nabla\mathbf{F}(\boldsymbol{x}^{k-1}; \xi^k)$ is bounded by $\eta_{k-1}^2 L^2$:

$$\begin{aligned}
&\mathbb{E}_{\xi^k}[\|\nabla\mathbf{F}(\boldsymbol{x}^k; \xi^k) - \nabla\mathbf{F}(\boldsymbol{x}^{k-1}; \xi^k) - (\nabla f(\boldsymbol{x}^k) - \nabla f(\boldsymbol{x}^{k-1}))\|^2] \\
&= \mathbb{E}_{\xi^k}[\|\nabla\mathbf{F}(\boldsymbol{x}^k; \xi^k) - \nabla\mathbf{F}(\boldsymbol{x}^{k-1}; \xi^k)\|^2] - \|\nabla f(\boldsymbol{x}^k) - \nabla f(\boldsymbol{x}^{k-1})\|^2 \\
&\leq \mathbb{E}_{\xi^k}[\|\nabla\mathbf{F}(\boldsymbol{x}^k; \xi^k) - \nabla\mathbf{F}(\boldsymbol{x}^{k-1}; \xi^k)\|^2] \leq \eta_{k-1}^2 L^2.
\end{aligned}$$

Then according to definitions of $\boldsymbol{\theta}_k$ and $\boldsymbol{Z}_k$, it is easy to obtain the upper bounds for $\|\boldsymbol{\theta}_k^b\|$, $\|\boldsymbol{Z}_k^b\|$, $\mathbb{E}_{\xi^k}[\|\boldsymbol{\theta}_k^u\|^2]$, and $\mathbb{E}_{\xi_k}[\|\boldsymbol{Z}_k^u\|^2]$ following [27, Lemma 5].

Lemma 8 *Under Assumptions 2 and 3, it holds that for any $k \geq 1$,*

$$\|\boldsymbol{\theta}_k^u\| \leq 2B_k \text{ and } \|\boldsymbol{Z}_k^u\| \leq 2D_k.$$

Besides, if $\left\|\nabla f(\boldsymbol{x}^k)\right\| \leq B_k/2$ and $\|\nabla f(x_k) - \nabla f(x_{k-1})\| \leq D_k/2$, then

$$\|\boldsymbol{\theta}_k^b\| \leq 2\sigma^2 B_k^{-1}, \quad \|\boldsymbol{Z}_k^b\| \leq 2L^2\eta_{k-1}^2 D_k^{-1}, \tag{22}$$

$$\text{and } \mathbb{E}_{\xi^k}[\|\boldsymbol{\theta}_k^u\|^2] \leq 10\sigma^2, \quad \mathbb{E}_{\xi^k}[\|\boldsymbol{Z}_k^u\|^2] \leq 10L^2\eta_{k-1}^2, \tag{23}$$

where $\eta_0 = 2$ in convention.

In the following, we assume that the parameters in Algorithm 1 satisfy

$$\rho_k = \frac{\rho}{k^{\varpi}},\ \eta_k = k^{-\frac{1}{2}},\ B_k = k^{\frac{1}{4}},\ D_k = \eta_{k-1}^{\frac{1}{2}},\ \alpha_k = 1 - k^{-\frac{1}{2}},\ \forall k \geq 1, \text{ and } \vartheta = \frac{1}{4}, \tag{24}$$

where $\rho > 0$, $\varpi > 1$. It is worthy to note that the parameter setting is problem-free in the convergence analysis, which is different from existing works for penalty methods [1,28,38]. The parameter setting (24) enables more specific upper bounds on the four estimates in (22) and (23).

Lemma 9 *Under Assumptions 1-3, suppose that the parameters in Algorithm 1 satisfy (24). Then for any $k \geq 1$, we have*

$$\|\boldsymbol{\theta}_k^b\| \leq \max\{6G^2, 2\sigma^2\}k^{-\frac{1}{4}}, \quad \|\boldsymbol{Z}_k^b\| \leq 6L^2(k-1)^{-\frac{3}{4}}, \tag{25}$$

$$\mathbb{E}_{\xi^k}[\|\boldsymbol{\theta}_k^u\|^2] \leq \max\{4G^2, 10\sigma^2\}, \quad \textit{and} \quad \mathbb{E}_{\xi^k}[\|\boldsymbol{Z}_k^u\|^2] \leq 10L^2(k-1)^{-1}. \tag{26}$$

Proof See Appendix C. □

Under the parameter setting (24), we can further ensure that the event defined in Lemma 10 occurs with high probability.

Lemma 10 *Under Assumptions 1-3, suppose that the parameters in Algorithm 1 satisfy (24). Then we have* $\mathrm{P}(\zeta_k) \geq 1 - \frac{2\delta}{5(k+1)\log^2(k+1)}$, *where*

$$\zeta_k := \left\{ \left|\sum_{s=1}^k \bar{\boldsymbol{U}}_s^k\right| \leq M_{\bar{U}}k^{-\frac{1}{4}}\log\left(\frac{4(k+1)}{\delta}\right), \left|\sum_{s=1}^k \bar{\boldsymbol{R}}_s^k\right| \leq M_{\bar{R}}k^{-\frac{1}{2}}\log\left(\frac{4(k+1)}{\delta}\right), \right.$$
$$\left. \left|\sum_{s=1}^k \tilde{\boldsymbol{U}}_s^k\right| \leq M_{\tilde{U}}k^{-\frac{1}{4}}\log\left(\frac{4(k+1)}{\delta}\right), \textit{ and } \left|\sum_{s=1}^k \tilde{\boldsymbol{R}}_s^k\right| \leq M_{\tilde{R}}k^{-\frac{1}{2}}\log\left(\frac{4(k+1)}{\delta}\right) \right\}$$

with $M_{\bar{U}} := 18 + 9L\sqrt{10\kappa(1, \frac{1}{2})}$, $M_{\bar{R}} := 72 + 18L\sqrt{10\kappa(\frac{3}{2}, \frac{1}{2})}$, *and*

$$M_{\tilde{U}} := 18 + 9\sqrt{\max\{4G^2, 10\sigma^2\}\kappa(1, \frac{1}{2})}, \ M_{\tilde{R}} := 72 + 18\sqrt{\max\{4G^2, 10\sigma^2\}\kappa(\frac{3}{2}, \frac{1}{2})}.$$

Proof See Appendix D. □

In the lemma below, we provide an upper bound and an iteration complexity on averaged gradients of the augmented Lagrangian function in $\boldsymbol{x}$ at history iterates with high probability. This will serve as a preparation for the oracle complexity analysis in Section 4.

Lemma 11 *Under the condition of Lemma 10 and given a positive constant $\delta \in (0, 1)$, it holds with probability at least $1 - \delta$ that*

$$\frac{1}{K}\sum_{k=1}^K \|\nabla_{\boldsymbol{x}}\mathcal{L}_{\beta_k}(\boldsymbol{x}^{k+1}, \boldsymbol{\lambda}^k)\| \leq M_3 K^{-\frac{1}{4}}\log\left(\frac{4(K+1)}{\delta}\right) \quad \forall K \geq 1, \tag{27}$$

and with probability at least $1 - \delta$ that

$$\frac{1}{K}\sum_{k=1}^K \|\nabla_{\boldsymbol{x}}\mathcal{L}_{\beta_k}(\boldsymbol{x}^{k+1}, \boldsymbol{\lambda}^k)\| \leq \epsilon \tag{28}$$

for any K satisfying

$$K \geq N(\epsilon, \delta) := \left\lceil \max\left\{ e^4, 8^4 M_3{}^4 \epsilon^{-4}\log^4\left(\frac{8\sqrt[4]{8}M_3}{e\delta^{\frac{1}{4}}\epsilon}\right)\right\}\right\rceil,$$

where $M_3 := 2M_1 + 4H + 2mC^2\tilde{\Gamma} + \frac{16}{3}\tilde{\Gamma}\tilde{L} + 4mG^2\tilde{\Gamma} + 3M_e$ *with* $M_e := G\kappa(\frac{1}{2}, \frac{1}{2}) + 4(M_Z + M_\theta)$, $M_Z := M_{\bar{U}} + \sqrt{2M_{\bar{R}}} + 2L\sqrt{5\kappa(1, \frac{1}{2})} + 6L^2\kappa(\frac{3}{4}, \frac{1}{2})$, $M_\theta := M_{\tilde{U}} + \sqrt{2M_{\tilde{R}}} + 2\sigma\sqrt{5\kappa(1, \frac{1}{2})} + \max\{6G^2, 2\sigma^2\}\kappa(\frac{3}{4}, \frac{1}{2})$, $\tilde{\Gamma} := \max\{\beta_1, \Gamma\}$, $\tilde{L}$ *defined in Lemma 3, and M_1 is defined in Lemma 5.*

Proof See Appendix E. □

4 Complexity analysis

In this part, we present an oracle complexity analysis for AdaSSP. Note that only one sample is called at each iteration of AdaSSP. Thus the oracle complexity of AdaSSP is in the same order as its iteration complexity.

The following theorem establishes that, before a certain number of iterations, with high probability we can reach a point $\boldsymbol{x}^k$ that either satisfies

$$\left\|\nabla \boldsymbol{c}_{\mathcal{E}}(\boldsymbol{x}^k)\boldsymbol{c}_{\mathcal{E}}(\boldsymbol{x}^k)+\nabla \boldsymbol{c}_{\mathcal{I}}(\boldsymbol{x}^k)[\boldsymbol{c}_{\mathcal{I}}(\boldsymbol{x}^k)]_+\right\| \leq \epsilon \text{ and } \|\boldsymbol{c}_{\mathcal{E}}(\boldsymbol{x}^k)\|+\|[\boldsymbol{c}_{\mathcal{I}}(\boldsymbol{x}^k)]_+\| > \epsilon, \tag{29}$$

or is an ϵ-*KKT point* satisfying

$$\left\|\nabla f(\boldsymbol{x}^k)+\nabla \boldsymbol{c}(\boldsymbol{x}^k)\tilde{\boldsymbol{\lambda}}^k\right\| \leq \epsilon, \tag{30}$$

$$\|\boldsymbol{c}_{\mathcal{E}}(\boldsymbol{x}^k)\|+\|[\boldsymbol{c}_{\mathcal{I}}(\boldsymbol{x}^k)]_+\| \leq \epsilon, \tag{31}$$

$$\tilde{\lambda}_i^k = 0 \text{ if } c_i(\boldsymbol{x}^k) \leq -\epsilon \text{ for all } i \in \mathcal{I}. \tag{32}$$

Theorem 1 *Under Assumptions 1-3 and given $\epsilon > 0$, $\delta \in (0,1)$, suppose that the parameters in Algorithm 1 satisfy (24). Then it holds with probability at least $1-\delta$ that there exists k satisfying $k \leq \bar{N}+1$, where*

$$\bar{N} := 2N(\frac{\epsilon}{4},\delta) + 2\left\lceil \frac{\log(2m\tilde{C}\epsilon^{-1})}{\log(\tau^{-1})}\right\rceil \times \left\lceil \frac{\bar{\beta}_\epsilon^4-\beta_1^4}{\Gamma^4}+1\right\rceil \tag{33}$$

with $\bar{\beta}_\epsilon = \max\{1, \frac{2M_2}{\epsilon}, \frac{C\sum_{j=1}^{\infty}\rho_j}{\epsilon}\}$, such that either (29) holds or (30)-(32) are satisfied.

Proof Note that

$$\frac{1}{\lceil k/2\rceil}\sum_{t=\lfloor k/2\rfloor+1}^{k}\|\nabla_{\boldsymbol{x}}\mathcal{L}_{\beta_t}(\boldsymbol{x}^{t+1},\boldsymbol{\lambda}^t)\| \leq \frac{2}{k}\sum_{t=1}^{k}\|\nabla_{\boldsymbol{x}}\mathcal{L}_{\beta_t}(\boldsymbol{x}^{t+1},\boldsymbol{\lambda}^t)\|.$$

It then follows from Lemma 11 that the event

$$A(\epsilon,\delta) = \left\{\frac{1}{\lceil k/2\rceil}\sum_{t=\lfloor k/2\rfloor+1}^{k}\|\nabla_{\boldsymbol{x}}\mathcal{L}_{\beta_t}(\boldsymbol{x}^{t+1},\boldsymbol{\lambda}^t)\| \leq \frac{\epsilon}{2},\ \forall k \geq N(\frac{\epsilon}{4},\delta)\right\} \tag{34}$$

occurs with probability at least $1-\delta$. In the remaining part, the analysis is presented under the circumstance that $A(\epsilon,\delta)$ occurs. Define

$$\bar{\mathcal{K}} := \left\{k \in [N(\frac{\epsilon}{4},\delta),\bar{N}] : \frac{1}{\lceil k/2\rceil}\sum_{t=\lfloor k/2\rfloor+1}^{k}\left(\|\boldsymbol{c}_{\mathcal{E}}(\boldsymbol{x}^{t+1})\|+\|\boldsymbol{V}^{t+1}\|\right) \leq \frac{\epsilon}{2}\right\}$$

as the index set including k such that

$$N(\frac{\epsilon}{4},\delta) \leq k \leq \bar{N} \tag{35}$$

and the average of $\|\boldsymbol{c}_{\mathcal{E}}(\boldsymbol{x}^{t+1})\|+\|\boldsymbol{V}^{t+1}\|$ over $t = \lfloor k/2\rfloor+1,\ldots,k$ is sufficiently small. We proceed our analysis by considering two mutually exclusive cases.

Case (i): $\bar{\mathcal{K}} \neq \emptyset$. Then for any $\tilde{k} \in \bar{\mathcal{K}}$, we have that

$$\frac{1}{\lceil \tilde{k}/2 \rceil} \sum_{t=\lfloor \tilde{k}/2 \rfloor + 1}^{\tilde{k}} \left(\|\boldsymbol{c}_{\mathcal{E}}(\boldsymbol{x}^{t+1})\| + \|\boldsymbol{V}^{t+1}\| \right) \leq \frac{\epsilon}{2}, \tag{36}$$

which implies that

$$\frac{1}{\lceil \tilde{k}/2 \rceil} \sum_{t=\lfloor \tilde{k}/2 \rfloor + 1}^{\tilde{k}} \left(\|\boldsymbol{c}_{\mathcal{E}}(\boldsymbol{x}^{t+1})\| + \|\boldsymbol{V}^{t+1}\| + \|\nabla_{\boldsymbol{x}} \mathcal{L}_{\beta_t}(\boldsymbol{x}^{t+1}, \boldsymbol{\lambda}^t)\| \right) \leq \epsilon.$$

Thus there exists $k \in [\lceil \tilde{k}/2 \rceil + 1, \tilde{k}]$ such that

$$\|\boldsymbol{c}_{\mathcal{E}}(\boldsymbol{x}^{k+1})\| + \|\boldsymbol{V}^{k+1}\| + \|\nabla_{\boldsymbol{x}} \mathcal{L}_{\beta_k}(\boldsymbol{x}^{k+1}, \boldsymbol{\lambda}^k)\| \leq \epsilon,$$

which indicates that (30)-(32) hold at $k+1$ by Lemma 1.

Case (ii): $\bar{\mathcal{K}} = \emptyset$. We will show that the first inequality of (29) and (32) hold for some $k \in [\lceil \bar{N}/2 \rceil + 1, \bar{N}]$. First, we need to prove

$$\beta_{\bar{N}/2} \geq \bar{\beta}_\epsilon. \tag{37}$$

Assume (37) were false, that said, $\beta_{\bar{N}/2} < \bar{\beta}_\epsilon$. Then to reach that the penalty parameter increases from $\beta_{N(\frac{\epsilon}{4},\delta)}$ to $\beta_{\bar{N}/2}$, the number of increments will be less than $\lceil \frac{\bar{\beta}_\epsilon^4 - \beta_{N(\frac{\epsilon}{4},\delta)}^4}{\Gamma^4} + 1 \rceil$. Therefore, it means that in $[N(\frac{\epsilon}{4},\delta), \frac{\bar{N}}{2}]$ the longest period of iterations where the penalty parameter stay the same value is larger than

$$\frac{\frac{\bar{N}}{2} - N(\frac{\epsilon}{4},\delta)}{\left\lceil \frac{\bar{\beta}_\epsilon^4 - \beta_{N(\frac{\epsilon}{4},\delta)}^4}{\Gamma^4} + 1 \right\rceil} \geq \frac{\frac{\bar{N}}{2} - N(\frac{\epsilon}{4},\delta)}{\left\lceil \frac{\bar{\beta}_\epsilon^4 - \beta_1^4}{\Gamma^4} + 1 \right\rceil} = \left\lceil \frac{\log(2m\tilde{C}\epsilon^{-1})}{\log(\tau^{-1})} \right\rceil.$$

Suppose that the starting index and ending index of this period is k_1 and k_2, respectively, satisfying $k_1, k_2 \in [N(\frac{\epsilon}{4},\delta), \frac{\bar{N}}{2}]$ and $k_2 - k_1 \geq \left\lceil \frac{\log(2m\tilde{C}\epsilon^{-1})}{\log(\tau^{-1})} \right\rceil$. From Assumption 1 and Lemma 3 we have

$$\begin{aligned} \|\boldsymbol{c}_{\mathcal{E}}(\boldsymbol{x}^{t+1})\| + \|\boldsymbol{V}^{t+1}\| &\leq \|\boldsymbol{c}_{\mathcal{E}}(\boldsymbol{x}^{t+1})\|_1 + \|\max\{[\boldsymbol{c}_{\mathcal{I}}(\boldsymbol{x}^{t+1})]_+, \frac{\boldsymbol{\lambda}_{\mathcal{I}}^t}{\beta_t}\}\|_1 \\ &\leq m \max\{C, \frac{C\sum_{j=1}^{\infty} \rho_j}{\beta_1}\} = m\tilde{C}. \end{aligned} \tag{38}$$

It indicates from the update scheme (8) and the upper bound argument (38) that

$$\begin{aligned} &\frac{1}{\lceil k_2/2 \rceil} \sum_{t=\lfloor k_2/2 \rfloor + 1}^{k_2} \left(\|\boldsymbol{c}_{\mathcal{E}}(\boldsymbol{x}^{t+1})\| + \|\boldsymbol{V}^{t+1}\| \right) \\ &\leq \frac{\tau^{k_2-k_1}}{\lceil k_1/2 \rceil} \sum_{t=\lfloor k_1/2 \rfloor + 1}^{k_1} \left(\|\boldsymbol{c}_{\mathcal{E}}(\boldsymbol{x}^{t+1})\| + \|\boldsymbol{V}^{t+1}\| \right) \\ &\leq \tau^{k_2-k_1} m\tilde{C} \leq \frac{\epsilon}{2}. \end{aligned}$$

This gives a contradiction to $\bar{\mathcal{K}} = \emptyset$, which prove (37).

If $c_i(\boldsymbol{x}^k) < -\epsilon$ for any $k \geq \bar{N}/2 + 1$ and for any $i \in \mathcal{I}$, by Lemma 3 and (37), we have that

$$\lambda_i^{k-1} + \beta_{k-1} c_i(\boldsymbol{x}^k) \leq C \sum_{j=1}^{\infty} \rho_j - \epsilon \bar{\beta}_\epsilon \leq 0 \text{ yielding } \tilde{\lambda}_i^k = \left[\beta_{k-1}\, c_i(\boldsymbol{x}^k) + \lambda_i^{k-1}\right]_+ = 0. \tag{39}$$

Thus (32) holds. Hence, if (34) holds, there exists some $\bar{k} \in [\lceil \frac{\bar{N}+1}{2} \rceil = \frac{\bar{N}}{2} + 1, \bar{N} + 1]$ such that $\|\nabla f(\boldsymbol{x}^{\bar{k}}) + \nabla \boldsymbol{c}(\boldsymbol{x}^{\bar{k}}) \tilde{\boldsymbol{\lambda}}^{\bar{k}}\| \leq \frac{\epsilon}{2}$, which together with Lemma 6 and (37) indicates that

$$\begin{aligned} \left\| \nabla \boldsymbol{c}_{\mathcal{E}}(\boldsymbol{x}^{\bar{k}}) \boldsymbol{c}_{\mathcal{E}}(\boldsymbol{x}^{\bar{k}}) + \nabla \boldsymbol{c}_{\mathcal{I}}(\boldsymbol{x}^{\bar{k}}) [\boldsymbol{c}_{\mathcal{I}}(\boldsymbol{x}^{\bar{k}})]_+ \right\| &\leq \frac{M_2 + \left\| \nabla f(\boldsymbol{x}^{\bar{k}}) + \nabla \boldsymbol{c}(\boldsymbol{x}^{\bar{k}}) \tilde{\boldsymbol{\lambda}}^{\bar{k}} \right\|}{\beta_{\bar{k}}} \\ &\leq \frac{M_2}{\beta_{\bar{k}}} + \frac{\epsilon}{2\beta_{\bar{k}}} \leq \frac{M_2}{\bar{\beta}_\epsilon} + \frac{\epsilon}{2\beta_{\bar{k}}} \leq \epsilon. \end{aligned} \tag{40}$$

In this situation, either (29) or (31) holds at $k = \bar{k}$. The proof is completed. □

To make sure that Algorithm 1 can find an ϵ-KKT point with high-probability, we need to assume the following constraint qualification condition, which is also imposed in [38, Assumption D.2].

Assumption 4 (Local LICQ) *There exists $\nu > 0$ such that singular values of $\{\nabla \boldsymbol{c}(\boldsymbol{x}),\ \boldsymbol{x} \in \mathcal{Y}^*\}$ are uniformly lower bounded away from ν, where $\mathcal{Y}^* := \{\boldsymbol{x} \in \mathbb{R}^n : \|\nabla \boldsymbol{c}_{\mathcal{E}}(\boldsymbol{x}) \boldsymbol{c}_{\mathcal{E}}(\boldsymbol{x}) + \nabla \boldsymbol{c}_{\mathcal{I}}(\boldsymbol{x}) [\boldsymbol{c}_{\mathcal{I}}(\boldsymbol{x})]_+\| = 0\}$.*

Note that $\mathcal{Y}^\star$ is closed due to the continuity of $\nabla \boldsymbol{c}$ and $\boldsymbol{c}$. The reason we call it "local" is that compared with strong LICQ assumed in [3,14,33] which requires LICQ hold in a "global" continuous region containing all realizations of random iterates, we only requires LICQ on the set of all stationary points. Through Assumption 4 we can ensure the feasibility of all points in $\mathcal{Y}^\star$. Moreover, there exists a positive constant $\tilde{\epsilon}$ such that singular values of $\nabla \boldsymbol{c}(\boldsymbol{x})$ are uniformly lower bounded away from $\nu/2$ for any $\boldsymbol{x}$ satisfying $\mathbf{d}(\boldsymbol{x}, \mathcal{Y}^\star) < \tilde{\epsilon}$. Then we have

$$\|\boldsymbol{c}_{\mathcal{E}}(\boldsymbol{x})\| + \|[\boldsymbol{c}_{\mathcal{I}}(\boldsymbol{x})]_+\| \leq \frac{2}{\nu} \|\nabla \boldsymbol{c}_{\mathcal{E}}(\boldsymbol{x}) \boldsymbol{c}_{\mathcal{E}}(\boldsymbol{x}) + \nabla \boldsymbol{c}_{\mathcal{I}}(\boldsymbol{x}) [\boldsymbol{c}_{\mathcal{I}}(\boldsymbol{x})]_+\|,\ \forall \mathbf{d}(\boldsymbol{x}, \mathcal{Y}^\star) < \tilde{\epsilon}. \tag{41}$$

By the compactness of $\mathcal{X}$ (Assumption 1), the set $\{\boldsymbol{x} \in \mathcal{X} : \mathbf{d}(\boldsymbol{x}, \mathcal{Y}^\star) \geq \tilde{\epsilon}\}$ is also compact. Therefore, there exists a constant $\hat{\epsilon}$ such that

$$0 < \hat{\epsilon} := \min_{\{\boldsymbol{x} \in \mathcal{X} : \mathbf{d}(\boldsymbol{x}, \mathcal{Y}^\star) \geq \tilde{\epsilon}\}} \|\nabla \boldsymbol{c}_{\mathcal{E}}(\boldsymbol{x}) \boldsymbol{c}_{\mathcal{E}}(\boldsymbol{x}) + \nabla \boldsymbol{c}_{\mathcal{I}}(\boldsymbol{x}) [\boldsymbol{c}_{\mathcal{I}}(\boldsymbol{x})]_+\|. \tag{42}$$

Then we derive the following theorem, locating an ϵ-KKT point $\boldsymbol{x}^k$ and establishing the boundedness of $\tilde{\boldsymbol{\lambda}}^k$ accordingly.

Theorem 2 *Under Assumptions 1-4 and given $\epsilon \in (0, \frac{2\hat{\epsilon}}{\nu})$, $\delta \in (0,1)$, suppose that the parameters in Algorithm 1 satisfy (24). Then it holds with probability at least $(1-\delta)$ that there exists a certain k satisfying $k \leq \hat{N} + 1$, where*

$$\hat{N} := 2N(\frac{\epsilon}{4}, \delta) + 2 \left\lceil \frac{\log(2m\tilde{C}\epsilon^{-1})}{\log(\tau^{-1})} \right\rceil \times \left\lceil \frac{\hat{\beta}_\epsilon^4 - \beta_1^4}{\Gamma^4} + 1 \right\rceil \tag{43}$$

with $\hat{\beta}_\epsilon = \max\{\frac{1}{\nu}, \frac{2M_2}{\nu\epsilon}, \frac{C\sum_{j=1}^{\infty} \rho_j}{\epsilon}\}$, such that (30)-(32) hold and

$$\|\tilde{\boldsymbol{\lambda}}^k\| \leq \max\left\{ mC \sum_{j=1}^{\infty} \rho_j + \max\left\{ \frac{2\hat{\epsilon}}{\nu^2}, \frac{2M_2}{\nu}, C\sum_{j=1}^{\infty} \rho_j \right\}, \frac{2G}{\nu} + \frac{2\hat{\epsilon}}{\nu^2} \right\}. \tag{44}$$

Proof It follows from Lemma 11 that $A(\epsilon,\delta)$, as defined in (34), occurs with probability at least $1-\delta$. We proceed our analysis under this occurrence. Similar to Theorem 1, we define

$$\hat{\mathcal{K}} := \left\{ k \in [N(\frac{\epsilon}{4},\delta), \hat{N}] : \frac{1}{\lfloor k/2 \rfloor} \sum_{t=\lceil k/2 \rceil+1}^{k} \|\boldsymbol{c}_{\mathcal{E}}(\boldsymbol{x}^{t+1})\| + \|\boldsymbol{V}^{t+1}\| \leq \frac{\epsilon}{2} \right\}.$$

Recalling the proof of Theorem 1, we can obtain the following results:

- When $\hat{\mathcal{K}} \neq \emptyset$, for any given $k \in \hat{\mathcal{K}}$ there exists some $\tilde{k} \in [\lceil k/2 \rceil + 2, k+1]$ such that (30)-(32) hold.
- When $\mathcal{K} = \emptyset$, it occurs that

$$\beta_{\lceil \hat{N}/2 \rceil+1} \geq \hat{\beta}_{\epsilon}. \tag{45}$$

For the case where $\mathcal{K} = \emptyset$, under Assumption 4 and by (40), (43), and (45), there exists a certain $\hat{k} \in [\lceil \hat{N}/2 \rceil + 1, \hat{N}]$ such that

$$\begin{aligned} \left\| \nabla \boldsymbol{c}_{\mathcal{E}}(\boldsymbol{x}^{\hat{k}}) \boldsymbol{c}_{\mathcal{E}}(\boldsymbol{x}^{\hat{k}}) + \nabla \boldsymbol{c}_{\mathcal{I}}(\boldsymbol{x}^{\hat{k}}) [\boldsymbol{c}_{\mathcal{I}}(\boldsymbol{x}^{\hat{k}})]_+ \right\| &\leq \frac{M_2 + \left\| \nabla f(\boldsymbol{x}^{\hat{k}}) + \nabla \boldsymbol{c}(\boldsymbol{x}^{\hat{k}}) \tilde{\boldsymbol{\lambda}}^{\hat{k}} \right\|}{\beta_k} \\ &\leq \frac{M_2}{\beta_k} + \frac{\epsilon}{2\beta_k} \leq \frac{M_2}{\hat{\beta}_{\epsilon}} + \frac{\epsilon}{2\hat{\beta}_{\epsilon}} \leq \frac{\nu\epsilon}{2} \leq \hat{\epsilon}. \end{aligned} \tag{46}$$

It derives from (41) and (42) that

$$\|\boldsymbol{c}_{\mathcal{E}}(\boldsymbol{x}^{\hat{k}})\| + \|[\boldsymbol{c}_{\mathcal{I}}(\boldsymbol{x}^{\hat{k}})]_+\| \leq \frac{2}{\nu} \|\nabla \boldsymbol{c}_{\mathcal{E}}(\boldsymbol{x}^{\hat{k}}) \boldsymbol{c}_{\mathcal{E}}(\boldsymbol{x}^{\hat{k}}) + \nabla \boldsymbol{c}_{\mathcal{I}}(\boldsymbol{x}^{\hat{k}}) [\boldsymbol{c}_{\mathcal{I}}(\boldsymbol{x}^{\hat{k}})]_+\| \leq \epsilon.$$

Now if $c_i(\boldsymbol{x}^{\hat{k}}) < -\epsilon$ for $i \in \mathcal{I}$, similar to (39), we have that $\tilde{\lambda}_i^{\hat{k}} = 0$, which indicates that (30)-(32) is still valid in this case.

When (30)-(32) hold at k, two cases occur.

(i) $\beta_k \leq \beta_{\epsilon}$. Recalling the definition of $\tilde{\boldsymbol{\lambda}}^k$, i.e. (9), gives

$$\begin{aligned} \|\tilde{\boldsymbol{\lambda}}^k\| &\leq \|\boldsymbol{\lambda}^{k-1}\| + \beta_k (\|\boldsymbol{c}_{\mathcal{E}}(\boldsymbol{x}^k)\| + \|[\boldsymbol{c}_{\mathcal{I}}(\boldsymbol{x}^k)]_+\|) \\ &\leq mC \sum_{j=1}^{k-1} \rho_j + \max \left\{ \frac{\epsilon}{\nu}, \frac{2M_2}{\nu}, C \sum_{j=1}^{\infty} \rho_j \right\}, \end{aligned} \tag{47}$$

where the first inequality is due to $\beta_{k-1} \leq \beta_k$ and the second one is from Lemma 3 as well as (31).

(ii) $\beta_k > \beta_{\epsilon}$. Then from the skeleton of (46) we can obtain that $\|\nabla \boldsymbol{c}_{\mathcal{E}}(\boldsymbol{x}^k) \boldsymbol{c}_{\mathcal{E}}(\boldsymbol{x}^k) + \nabla \boldsymbol{c}_{\mathcal{I}}(\boldsymbol{x}^k) [\boldsymbol{c}_{\mathcal{I}}(\boldsymbol{x}^k)]_+\| \leq \hat{\epsilon}$, which implies that $\nabla \boldsymbol{c}(\boldsymbol{x}^k)$ is non-singular with singular values uniformly lower bounded away from $\frac{\nu}{2}$ from (42). This, together with

$$\left\| \nabla \boldsymbol{c}(\boldsymbol{x}^k) \tilde{\boldsymbol{\lambda}}^k \right\| \leq \left\| \nabla f(\boldsymbol{x}^k) + \nabla \boldsymbol{c}(\boldsymbol{x}^k) \tilde{\boldsymbol{\lambda}}^k \right\| + \left\| \nabla f(\boldsymbol{x}^k) \right\| \leq \epsilon + G,$$

yields the boundedness of $\|\tilde{\boldsymbol{\lambda}}^k\|$: $\|\tilde{\boldsymbol{\lambda}}^k\| \leq \frac{2(\epsilon+G)}{\nu}$.

Therefore, combining (i) and (ii) together with $\epsilon \leq \frac{2\hat{\epsilon}}{\nu}$ yields (44). □

Remark 2 By Theorem 2, before a number of $\tilde{\mathcal{O}}(\epsilon^{-4})$ calls to $\mathcal{SFO}$, with a high probability one is supposed to observe an ϵ-KKT point. The work by [1] and [28] establishes the oracle complexity of $\tilde{\mathcal{O}}(\epsilon^{-4})$ and $\tilde{\mathcal{O}}(\epsilon^{-3})$, respectively, with predetermined varying penalty parameters being used. However, different from them, our work adopts an adaptive penalty parameter and employs only local constraint qualification condition (see Assumption 4).

Remark 3 We note that there exits a subsequence of $\{\boldsymbol{x}^k\}_{k\geq 1}$ such that (30)-(32) hold almost surely for any $\epsilon > 0$. This is a direct consequence of the Borel-Cantelli lemma, since it follows from Theorem 2 that the probability of failure events (for fixed ϵ decaying exponentially with K) is summable for ϵ, i.e.,

$$\sum_{K=1}^{\infty} \left[1 - \mathbb{P}\left((30)\text{-}(32) \text{ hold for } \epsilon \text{ and any } k \in [K]\right)\right] < \infty.$$

5 Global convergence analysis

In this section we will investigate the in-expectation global convergence of the Algorithm 1.

The following in-expectation results will help promote the sequential global convergence analysis.

Lemma 12 *Under Assumptions 1-3, suppose that the parameters in Algorithm 1 satisfy (24). Then for any $K \geq 1$ we have that*

$$\frac{1}{K}\sum_{k=1}^{K} \mathbb{E}_{\xi^{[k]}}\left[\|\nabla_{\boldsymbol{x}} \mathcal{L}_{\beta_k}(\boldsymbol{x}^{k+1}, \boldsymbol{\lambda}^k)\|\right] \leq M_4 K^{-\frac{1}{4}}, \tag{48}$$

$$\frac{1}{K}\sum_{k=1}^{K} \mathbb{E}_{\xi^{[k]}}\left[\|\nabla \boldsymbol{c}_{\mathcal{E}}(\boldsymbol{x}^{k+1})\boldsymbol{c}_{\mathcal{E}}(\boldsymbol{x}^{k+1}) + \nabla \boldsymbol{c}_{\mathcal{I}}(\boldsymbol{x}^{k+1})[\boldsymbol{c}_{\mathcal{I}}(\boldsymbol{x}^{k+1})]_+\|^{\frac{1}{2}}\right]$$
$$\leq \left(\mathbb{E}_{\xi^{[K]},\infty}\left[\frac{1}{K}\sum_{k=1}^{K} \beta_k^{-1}\right]\right)^{\frac{1}{2}} \left(M_2{}^{\frac{1}{2}} + M_4^{\frac{1}{2}}\right), \tag{49}$$

where $M_4 := 2M_1 + 4H + 2mC^2\tilde{\Gamma} + 8\tilde{\Gamma}\tilde{L} + 4mG^2\tilde{\Gamma} + 72(L^2+G^2+\sigma^2)\kappa(\frac{3}{4},\frac{1}{2}) + 12(L+G+\sigma)\sqrt{10\kappa(1,\frac{1}{2})} + 3G\kappa(\frac{1}{2},\frac{1}{2})$ *and M_2 is defined in Lemma 6.*

Proof See Appendix F. □

Due to the adaptive update scheme of penalty parameters, it may occur that the penalty parameter sequence $\{\beta_k\}$, generated by a single run of algorithm, increases to infinity or keeps upper bounded. Thus the following analysis is separated into two parts addressing above two cases, respectively.

5.1 Unbounded penalty parameters

When penalty parameters are unbounded, we consider the following event to occur with positive probability:

$$\mathbb{P}(\Omega_\infty) > 0, \text{ where } \Omega_\infty := \{\omega \in \Omega : \beta_k(\omega) \to \infty, \text{ as } k \to \infty\}.$$

In this subsection we will provide a global convergence analysis of AdaSSP conditioned on Ω_∞. To proceed, we formalize the properties of stochastic gradients conditioned on Ω_∞ in the assumption below, so that lemmas in the previous section still hold. The expectation conditioned on Ω_∞ is denoted by $\mathbb{E}_{\xi,\infty}[\cdot] := \mathbb{E}_{\xi}[\cdot|\Omega_\infty]$.

Assumption 5 *The function $\mathbf{F}(\cdot;\xi)$ is continuously differentiable for each $\xi \in \Xi$ and there exists $\sigma > 0$ such that $\mathbb{E}_{\xi,\infty}[\nabla \mathbf{F}(\boldsymbol{x}^k;\xi)] = \nabla f(\boldsymbol{x}^k)$ and $\mathbb{E}_{\xi,\infty}[\|\nabla \mathbf{F}(\boldsymbol{x}^k;\xi) - \nabla f(\boldsymbol{x}^k)\|^2] \le \sigma^2$ for any $\boldsymbol{x}^k$. Moreover, there exists $L > 0$ such that $\mathbb{E}_{\xi,\infty}[\|\nabla \mathbf{F}(\boldsymbol{u};\xi) - \nabla \mathbf{F}(\boldsymbol{v};\xi)\|^2] \le L^2\|\boldsymbol{u}-\boldsymbol{v}\|^2$ for all $\boldsymbol{u},\boldsymbol{v} \in \mathcal{X}$.*

The theorem below shows global convergence properties of the averaged stationarity measure at iterates in expectation.

Theorem 3 *Suppose that Assumptions 1 and 5 hold, and the parameters used in Algorithm 1 satisfy (24). Then*

$$\lim_{K\to\infty} \frac{1}{K}\sum_{k=1}^{K} \mathbb{E}_{\xi^{[k]},\infty}\left[\|\nabla f(\boldsymbol{x}^{k+1}) + \nabla \boldsymbol{c}(\boldsymbol{x}^{k+1})\tilde{\boldsymbol{\lambda}}^{k+1}\|\right] = 0, \tag{50}$$

$$\lim_{K\to\infty} \frac{1}{K}\sum_{k=1}^{K} \mathbb{E}_{\xi^{[k]},\infty}\left[\|\nabla \boldsymbol{c}_{\mathcal{E}}(\boldsymbol{x}^{k+1})\boldsymbol{c}_{\mathcal{E}}(\boldsymbol{x}^{k+1}) + \nabla \boldsymbol{c}_{\mathcal{I}}(\boldsymbol{x}^{k+1})[\boldsymbol{c}_{\mathcal{I}}(\boldsymbol{x}^{k+1})]_+\|^{\frac{1}{2}}\right] = 0. \tag{51}$$

Moreover, if $\eta_k = k^{-q}$ with $q > \frac{3}{4}$ for any $k \ge 1$ and there exists $C_f > 0$ such that $f(\boldsymbol{x}^k) \le C_f$ for any $k \ge 1$, then

$$\lim_{K\to\infty} \mathbb{E}_{\xi^{[K]},\infty}\left[\|\boldsymbol{c}_{\mathcal{E}}^{K+1}\|^2 + \|[\boldsymbol{c}_{\mathcal{I}}^{K+1}]_+\|^2\right] \text{ exists and is finite.} \tag{52}$$

Proof First, it implies (50) and (51) from Lemma 12 and $\beta_k(\omega) \to \infty, \forall \omega \in \Omega_\infty$.

We next demonstrate that (52) holds true. It follows from (16) that

$$0 \le 2(\mathcal{L}_{\beta_k}(\boldsymbol{x}^k,\boldsymbol{\lambda}^k) - \mathcal{L}_{\beta_k}(\boldsymbol{x}^{k+1},\boldsymbol{\lambda}^k)) + \eta_k^2 L_{\beta_k} + 2\eta_k\|\boldsymbol{e}^k\|.$$

Dividing both sides of the above inequality by $2\beta_k$ and taking the expectation yield

$$\begin{aligned} J_k := \mathbb{E}_{\xi^{[k]},\infty}\left[\frac{f(\boldsymbol{x}^{k+1}) - f(\boldsymbol{x}^k)}{\beta_k} + \frac{\psi_{\beta_k}(\boldsymbol{c}^{k+1},\boldsymbol{\lambda}^k) - \psi_{\beta_k}(\boldsymbol{c}^k,\boldsymbol{\lambda}^k)}{\beta_k}\right] \\ - \mathbb{E}_{\xi^{[k]},\infty}\left[\frac{\eta_k^2 L_{\beta_k}}{2\beta_k} + \frac{\eta_k\|\boldsymbol{e}^k\|}{\beta_k}\right] \le 0. \end{aligned}$$

Therefore, $\{\sum_{k=1}^K J_k\}_{K\ge1}$ is a non-increasing sequence as K increases to infinity. Recall the definition of $\mathcal{L}_\beta$. It gives

$$\begin{aligned} \sum_{k=1}^{K} J_k = &\sum_{k=1}^{K} \mathbb{E}_{\xi^{[k]},\infty}\left[\frac{f(\boldsymbol{x}^{k+1}) - f(\boldsymbol{x}^k)}{\beta_k}\right] \\ &+ \mathbb{E}_{\xi^{[K]},\infty}\left[\sum_{k=1}^{K} \frac{\psi_{\beta_k}(\boldsymbol{c}^{k+1},\boldsymbol{\lambda}^k) - \psi_{\beta_k}(\boldsymbol{c}^k,\boldsymbol{\lambda}^k)}{\beta_k} - \frac{1}{2}(\|\boldsymbol{c}_{\mathcal{E}}^{K+1}\|^2 + \|[\boldsymbol{c}_{\mathcal{I}}^{K+1}]_+\|^2)\right] \\ &+ \frac{1}{2}\mathbb{E}_{\xi^{[K]},\infty}\left[\|\boldsymbol{c}_{\mathcal{E}}^{K+1}\|^2 + \|[\boldsymbol{c}_{\mathcal{I}}^{K+1}]_+\|^2\right] - \sum_{k=1}^{K} \mathbb{E}_{\xi^{[k]},\infty}\left[\frac{\eta_k^2 L_{\beta_k}}{2\beta_k} + \frac{\eta_k\|\boldsymbol{e}^k\|}{\beta_k}\right]. \end{aligned} \tag{53}$$

Since $\sum_{k=1}^K [f(\boldsymbol{x}^{k+1}) - f(\boldsymbol{x}^k)]$ is bounded and $\{\frac{1}{\beta_k}\}$ is monotonically decreasing to 0, the first term on the R.H.S. of (53) converges by Dirichlet's Test. Additionally, by Lemma [38, Lemma A.1] we are able to prove that

$$\lim_{K\to\infty} \mathbb{E}_{\xi^{[K]},\infty}\left[\sum_{k=1}^{K} \frac{\psi_{\beta_k}(\boldsymbol{c}^{k+1},\boldsymbol{\lambda}^k) - \psi_{\beta_k}(\boldsymbol{c}^k,\boldsymbol{\lambda}^k)}{\beta_k} - \frac{1}{2}\|\boldsymbol{c}^{K+1}\|^2\right] \quad \text{exists.}$$

Moreover, $\sum_{k=1}^{K} \frac{\eta_k^2 L_{\beta_k}}{2\beta_k}$ converges because $\frac{\eta_k^2 L_{\beta_k}}{2\beta_k} \leq \frac{\eta_k^2 \tilde{L}}{2} \leq \frac{k^{-\frac{3}{2}} \tilde{L}}{2}$ with $q > \frac{3}{4}$. Meanwhile, from (87) and $\eta_k = k^{-q} \leq k^{-\frac{1}{2}}$, we have that

$$\sum_{k=1}^{K} \mathbb{E}_{\xi^{[k]},\infty} \left[\frac{\eta_k \|\boldsymbol{e}^k\|}{\beta_k} \right] \leq \frac{1}{\beta_1} \sum_{k=1}^{K} \eta_k \mathbb{E}_{\xi^{[k]},\infty} \left[\|\boldsymbol{e}^k\| \right] \leq \frac{1}{\beta_1} \left(G\kappa(\frac{1}{2}, \frac{1}{2}) + \sum_{k=1}^{K} k^{-q} \left(6(L^2 + G^2 + \sigma^2) k^{-\frac{1}{4}} \kappa(\frac{3}{4}, \frac{1}{2}) + (L + G + \sigma) k^{-\frac{1}{4}} \sqrt{10\kappa(1, \frac{1}{2})} \right) \right),$$

which converges because $-\frac{1}{4} - q < -1$. Then we obtain that the last term on the R.H.S. of (53) also converges. Therefore, due to the lower boundedness of the fourth term on the R.H.S. of (53), the sequence $\{\sum_{k=1}^{K} J_k\}$ is uniformly lower bounded by a finite value. In addition, by the monotonically decreasing property, $\{\sum_{k=1}^{K} J_k\}$ must converge as K increases to infinity. Thus the fourth term on the R.H.S. of (53) converges to a finite value as K approaches infinity, which implies that (52) holds true. □

To further analyze feasibility and the complementary slackness at iterates, we assume the extended variant of MFCQ considered by [38].

Assumption 6 *There exist positive constants ς and P such that for any $k \geq 1$ the linear system*

$$\begin{aligned} \varsigma \cdot \operatorname{sgn}(c_i(\boldsymbol{x}^k)) + \nabla c_i(\boldsymbol{x}^k)^{\mathrm{T}} \boldsymbol{p} = 0, \qquad & i \in \mathcal{E} : c_i(\boldsymbol{x}^k) \neq 0\,; \\ \varsigma + \nabla c_i(\boldsymbol{x}^k)^{\mathrm{T}} \boldsymbol{p} \leq 0, \qquad & i \in \mathcal{I} : c_i(\boldsymbol{x}^k) > 0 \end{aligned}$$

has a solution $\boldsymbol{p}^k \in \mathbb{R}^n$ with $\|\boldsymbol{p}^k\| \leq P$.

Under Assumption 6, we are able to prove the average of vectors $\{\tilde{\boldsymbol{\lambda}}_k\}_{k \geq 1}$, as defined in (9), is upper bounded in the lemma below. The proof idea is motivated by [38], we thus present the details in Appendix G for completeness.

Lemma 13 *Suppose that Assumptions 1, 5, and 6 hold, and the parameters used in Algorithm 1 satisfy (24), then $\{a_K := \mathbb{E}_{\xi^{[K]},\infty}[\frac{1}{K} \sum_{k=1}^{K} \|\tilde{\boldsymbol{\lambda}}^{k+1}\|],\ K \geq 1\}$ is uniformly upper bounded.*

Proof See Appendix G. □

Following Lemma 13 and [38, Theorem 2], we can provide a characterization of global convergence, in which the expected constraint violation and complementary slackness converge to zero, respectively.

Theorem 4 *Suppose that Assumptions 1, 5, and 6 hold, and the parameters used in Algorithm 1 satisfy (24), then we have*

$$\lim_{K \to \infty} \frac{1}{K} \sum_{k=1}^{K} \mathbb{E}_{\xi^{[k]},\infty} \left[\|\boldsymbol{c}_{\mathcal{E}}(\boldsymbol{x}^{k+1})\|^{\frac{1}{2}} + \|[\boldsymbol{c}_{\mathcal{I}}(\boldsymbol{x}^{k+1})]_+\|^{\frac{1}{2}} \right] = 0, \tag{54}$$

$$\lim_{K \to \infty} \frac{1}{K} \sum_{k=1}^{K} \mathbb{E}_{\xi^{[k]},\infty} \left[\sum_{i \in \mathcal{I}} (\tilde{\lambda}_i^{k+1} |c_i(\boldsymbol{x}^{k+1})|)^{\frac{1}{4}} \right] = 0. \tag{55}$$

Proof See Appendix H.

5.2 Bounded penalty parameters

We now investigate the theoretical properties of AdaSSP when all the penalty parameters are bounded by a certain positive value $\beta_{\max}$. In this case, we consider

$$\mathbb{P}(\Omega_0) > 0, \text{ where } \Omega_0 := \{\omega \in \Omega : \beta_k(\omega) \leq \beta_{\max} \text{ for any } k \geq 1\}.$$

We first formalize the properties of stochastic oracles in the assumption below so that the lemmas in Section 3 still hold conditioned on Ω_0. The expectation conditioned on Ω_0 is denoted by $\mathbb{E}_{\xi,0}[\cdot] := \mathbb{E}_{\xi}[\cdot|\Omega_0]$.

Assumption 7 *The function $\mathbf{F}(\cdot;\xi)$ is continuously differentiable for each $\xi \in \Xi$ and there exists $\sigma > 0$ such that $\mathbb{E}_{\xi,0}[\nabla \mathbf{F}(\boldsymbol{x}^k;\xi)] = \nabla f(\boldsymbol{x}^k)$ and $\mathbb{E}_{\xi,0}[\|\nabla \mathbf{F}(\boldsymbol{x}^k;\xi) - \nabla f(\boldsymbol{x}^k)\|^2] \leq \sigma^2$ for any $\boldsymbol{x}^k$. Moreover, there exists $L > 0$ such that $\mathbb{E}_{\xi,0}[\|\nabla \mathbf{F}(\boldsymbol{u};\xi) - \nabla \mathbf{F}(\boldsymbol{v};\xi)\|^2] \leq L^2\|\boldsymbol{u} - \boldsymbol{v}\|^2$ for all $\boldsymbol{u}, \boldsymbol{v} \in \mathcal{X}$.*

We now summarize the global convergence properties of Algorithm 1 conditioned on Ω_0.

Theorem 5 *Suppose that Assumptions 1 and 7 hold, and the parameters used in Algorithm 1 satisfy (24), then*

$$\lim_{K\to\infty} \frac{1}{K} \sum_{k=1}^{K} \mathbb{E}_{\xi^{[k]},0}[\|\nabla f(\boldsymbol{x}^{k+1}) + \nabla \boldsymbol{c}(\boldsymbol{x}^{k+1})\tilde{\boldsymbol{\lambda}}^{k+1}\|] = 0, \tag{56}$$

$$\lim_{K\to\infty} \frac{1}{K} \sum_{k=1}^{K} \mathbb{E}_{\xi^{[k]},0}[\|\boldsymbol{c}_{\mathcal{E}}(\boldsymbol{x}^{k+1})\| + \|[\boldsymbol{c}_{\mathcal{I}}(\boldsymbol{x}^{k+1})]_+\|] = 0, \tag{57}$$

$$\lim_{k\to\infty} \frac{1}{K} \sum_{k=1}^{K} \mathbb{E}_{\xi^{[k]},0}[\sum_{i\in\mathcal{I}} \tilde{\lambda}_i^{k+1}|c_i(\boldsymbol{x}^{k+1})|] = 0. \tag{58}$$

Proof Firstly, (56) can be derived similarly as (50). Since $\{\beta_k\}$ is upper bounded by $\beta_{\max}$, it holds that there exists index k_b such that $\beta_k = \beta_{\max}$ for all $k \geq k_b$. Then following the update rule of β_k and (38), we obtain that

$$\begin{aligned}
&\sum_{t=\lfloor k/2\rfloor+1}^{k} \|\boldsymbol{c}_{\mathcal{E}}(\boldsymbol{x}^{t+1})\| + \|[\boldsymbol{c}_{\mathcal{I}}(\boldsymbol{x}^{t+1})]_+\| \leq \sum_{t=\lfloor k/2\rfloor+1}^{k} \|\boldsymbol{c}_{\mathcal{E}}(\boldsymbol{x}^{t+1})\| + \|\boldsymbol{V}^{t+1}\| \\
&\leq \frac{\lceil k/2\rceil}{\lceil k_b/2\rceil}\tau^{k-k_b} \sum_{t=\lfloor k_b/2\rfloor+1}^{k_b} (\|\boldsymbol{c}_{\mathcal{E}}(\boldsymbol{x}^{t+1})\| + \|\boldsymbol{V}^{t+1}\|) \leq k\tau^{k-k_b} m\tilde{C}, \quad \forall k > k_b,
\end{aligned}$$

which yields (57). Moreover, it implies from Lemma 1 that for all $k > \lfloor k_b/2\rfloor + 1$,

$$c_i(\boldsymbol{x}^k) < k\tau^{k-k_b} m\tilde{C} \text{ and } \tilde{\lambda}_i^k = 0 \text{ if } c_i(\boldsymbol{x}^k) < -k\tau^{k-k_b} m\tilde{C} \quad \text{for all } i \in \mathcal{I}.$$

Then we obtain that for any $i \in \mathcal{I}$,

$$\begin{aligned}
\tilde{\lambda}_i^k|c_i(\boldsymbol{x}^k)| &= [\lambda_i^k + \beta_k c_i(\boldsymbol{x}^k)]_+|c_i(\boldsymbol{x}^k)| \\
&= \left((\lambda_i^k + \beta_k c_i(\boldsymbol{x}^k))|c_i(\boldsymbol{x}^k)|\right) \mathbb{1}_{|c_i(\boldsymbol{x}^k)|\leq k\tau^{k-k_b}m\tilde{C}} \\
&\leq \left(\lambda_i^k|c_i(\boldsymbol{x}^k)| + \beta_k|c_i(\boldsymbol{x}^k)|^2\right) \mathbb{1}_{|c_i(\boldsymbol{x}^k)|\leq k\tau^{k-k_b}m\tilde{C}} \\
&\leq \max_{0\leq u\leq k\tau^{k-k_b}m\tilde{C}} \lambda_i^k u + \beta_k u^2 \leq C\sum_{t=1}^{\infty} \rho_k\tau^{k-k_b}m\tilde{C} + \beta_k^2 k^2\tau^{2(k-k_b)}m^2\tilde{C}^2,
\end{aligned}$$

which converges to 0 as k approaches ∞. Thus (58) holds true. □

6 Subproblem solver

We now focus on solving subproblem (5), which is equivalent to computing $\boldsymbol{x}^{k+1} = \boldsymbol{x}^k + \boldsymbol{r}^k$, where $\boldsymbol{r}^k$ (approximately) solves

$$\min_{\|\boldsymbol{r}\| \le \eta_k} \varphi(\boldsymbol{r}) := \boldsymbol{r}^{\mathrm{T}} \boldsymbol{g}^k + \frac{\beta_k}{2} \|[(\nabla \boldsymbol{c}_{\mathcal{I}}^k)^{\mathrm{T}} \boldsymbol{r} + \boldsymbol{c}_{\mathcal{I}}^k + \frac{\boldsymbol{\lambda}_k}{\beta_k}]_+\|^2 + \frac{\beta_k}{2} \|(\nabla \boldsymbol{c}_{\mathcal{E}}^k)^{\mathrm{T}} \boldsymbol{r} + \boldsymbol{c}_{\mathcal{E}}^k + \frac{\boldsymbol{\lambda}_k}{\beta_k}\|^2 + \frac{h_k}{2} \|\boldsymbol{r}\|^2.$$

From (3), we know $\nabla \varphi(\boldsymbol{0}) = \boldsymbol{d}^k$. Consider $\bar{\boldsymbol{r}}^k = -\varrho_k \eta_k \frac{\boldsymbol{d}^k}{\|\boldsymbol{d}^k\|}$, where

$$\varrho_k = \arg \min_{\varrho \in [0,1]} \iota \varrho + \frac{1}{2} \|[\varrho \boldsymbol{a}_{\mathcal{I}} + \boldsymbol{b}_{\mathcal{I}}]_+\|^2 + \frac{1}{2} \|\varrho \boldsymbol{a}_{\mathcal{E}} + \boldsymbol{b}_{\mathcal{E}}\|^2 + \frac{h_k \eta_k^2 \rho^2}{2} \tag{59}$$

with $\iota = -\frac{\eta_k (\boldsymbol{d}^k)^{\mathrm{T}} \boldsymbol{g}^k}{\|\boldsymbol{d}^k\|}$, $\boldsymbol{a}_{\mathcal{I}} = \frac{-\sqrt{\beta_k} \eta_k \nabla (\boldsymbol{c}_{\mathcal{I}}^k)^{\mathrm{T}} \boldsymbol{d}^k}{\|\boldsymbol{d}^k\|}$, $\boldsymbol{a}_{\mathcal{E}} = \frac{-\sqrt{\beta_k} \eta_k \nabla (\boldsymbol{c}_{\mathcal{E}}^k)^{\mathrm{T}} \boldsymbol{d}^k}{\|\boldsymbol{d}^k\|}$,

$$\boldsymbol{b}_{\mathcal{I}} = \sqrt{\beta_k} \boldsymbol{c}_{\mathcal{I}}^k + \frac{\boldsymbol{\lambda}_k}{\sqrt{\beta_k}}, \text{ and } \boldsymbol{b}_{\mathcal{E}} = \sqrt{\beta_k} \boldsymbol{c}_{\mathcal{E}}^k + \frac{\boldsymbol{\lambda}_k}{\sqrt{\beta_k}}.$$

The following theorem characterizes the sufficient reduction of φ achieved at $\bar{\boldsymbol{r}}^k$, which is also the reduction $Q_k(\boldsymbol{x}^k) - Q_k(\boldsymbol{x}^k + \bar{\boldsymbol{r}}^k)$.

Proposition 1 *Let $\bar{\boldsymbol{r}}^k = -\varrho_k \eta_k \frac{\boldsymbol{d}^k}{\|\boldsymbol{d}^k\|}$, where ϱ_k is the solution of (59), then*

$$\varphi(\boldsymbol{0}) - \varphi(\bar{\boldsymbol{r}}^k) \ge -\frac{1}{2} \min\{\eta_k, \frac{\|\boldsymbol{d}^k\|}{m G_k^2 \beta_k + H}\} \|\boldsymbol{d}^k\|. \tag{60}$$

Proof Consider the function $\Phi(\boldsymbol{x}) = \langle \boldsymbol{d}^k, \boldsymbol{x} - \boldsymbol{x}^k \rangle + \frac{m G_k^2 \beta_k + H}{2} \|\boldsymbol{x} - \boldsymbol{x}^k\|^2 + \varphi(\boldsymbol{0})$. Recall the model function $Q_k(\boldsymbol{x})$ defined in subproblem (5). It follows from 1 that for any $\boldsymbol{u}, \boldsymbol{v} \in \mathbb{R}^n$,

$$\begin{aligned}
&\|\nabla Q_k(\boldsymbol{u}) - \nabla Q_k(\boldsymbol{v})\| \\
&= \big\|\beta_k \nabla \boldsymbol{c}_{\mathcal{E}}^k (\nabla \boldsymbol{c}_{\mathcal{E}}^k)^{\mathrm{T}} (\boldsymbol{u} - \boldsymbol{v}) + \nabla \boldsymbol{c}_{\mathcal{I}}^k \big([\boldsymbol{\lambda}_{\mathcal{I}}^k + \beta_k \boldsymbol{c}_{\mathcal{I}}^k + \beta_k (\nabla \boldsymbol{c}_{\mathcal{I}}^k)^{\mathrm{T}} \boldsymbol{u}]_+ \\
&\quad - [\boldsymbol{\lambda}_{\mathcal{I}}^k + \beta_k \boldsymbol{c}_{\mathcal{I}}^k + \beta_k (\nabla \boldsymbol{c}_{\mathcal{I}}^k)^{\mathrm{T}} \boldsymbol{v}]_+\big) + h_k (\boldsymbol{u} - \boldsymbol{v})\big\| \\
&\le \beta_k (\|\nabla c_{\mathcal{E}}^k (\nabla \boldsymbol{c}_{\mathcal{E}}^k)^{\mathrm{T}}\| + \|\nabla c_{\mathcal{I}}^k (\nabla \boldsymbol{c}_{\mathcal{I}}^k)^{\mathrm{T}}\|) \|\boldsymbol{u} - \boldsymbol{v}\| + h_k \|\boldsymbol{u} - \boldsymbol{v}\| \\
&\le \beta_k \|\boldsymbol{u} - \boldsymbol{v}\| \sum_{i=1}^m \|\nabla \boldsymbol{c}_i^k\|^2 + H \|\boldsymbol{u} - \boldsymbol{v}\| \le (m G_k^2 \beta_k + H) \|\boldsymbol{u} - \boldsymbol{v}\|,
\end{aligned} \tag{61}$$

which indicates the Lipschitz continuity of $\nabla Q_k(\cdot)$ with Lipschitz constant bounded by $m G_k^2 \beta_k + H$. Therefore, $\Phi(\boldsymbol{x})$ is a quadratic function with the gradient $\nabla \Phi(\boldsymbol{x}^k) = \boldsymbol{d}^k$ and $\Phi(\boldsymbol{r} + \boldsymbol{x}^k) \ge \varphi(\boldsymbol{r})$ by (61). Thus it follows from (59) that

$$\begin{aligned}
\varphi(\boldsymbol{0}) - \varphi(\bar{\boldsymbol{r}}_k) &= \varphi(\boldsymbol{0}) - \min_{\varrho \in [0,1]} \{\varphi(-\varrho \eta_k \frac{\boldsymbol{d}^k}{\|\boldsymbol{d}^k\|})\} \\
&\ge \varphi(\boldsymbol{0}) - \varphi(\arg \min_{\varrho \in [0,1]} \{\Phi(\boldsymbol{x}^k - \varrho \eta_k \frac{\boldsymbol{d}^k}{\|\boldsymbol{d}^k\|})\} - \boldsymbol{x}^k) \\
&\ge \Phi(\boldsymbol{0}) - \min_{\varrho \in [0,1]} \{\Phi(\boldsymbol{x}^k - \varrho \eta_k \frac{\boldsymbol{d}^k}{\|\boldsymbol{d}^k\|})\} \ge \frac{1}{2} \min\{\eta_k, \frac{\|\boldsymbol{d}^k\|}{m G_k^2 \beta_k + H}\} \|\boldsymbol{d}^k\|,
\end{aligned}$$

where the last inequality is due to reduction by the Cauchy point from [43, Lemma 4.3]. Thus the result is yielded. □

There are variants of approaches to solve the subproblem (59). For instance, we can simply reformulate it as a convex quadratic programming (QP) and call a state-of-art QP solver. Instead of doing so, we present a more direct way to solve (59). We can exploit the convexity and piecewise quadratic structure of (59) to solve it through traveling the set $\mathcal{A}(\varrho) = \{i : (\boldsymbol{a}_\mathcal{I})_i\varrho + (\boldsymbol{b}_\mathcal{I})_i > 0\}$. After sorting the critical points $q_i = -\frac{(\boldsymbol{b}_\mathcal{I})_i}{(\boldsymbol{a}_\mathcal{I})_i}$, we can split the interval $[0, 1]$ into at most $|\mathcal{I}| + 1$ partition. Then we will travel all the mini interval from left to right, and for each interval we can examine if there is a minimizer easily from the unidemensional quadratic function determined by quadratic terms for $i \in \mathcal{A}(\varrho)$. The complete procedure is summarized in Algorithm 2. Note that the total computational complexity is in order $\mathcal{O}(m \log m)$, primarily due to the sorting operation in Line 2. Actually, this is negligible compared to the computational cost of matrix multiplication in Algorithm 1.

Algorithm 2 Active set method for solving (59)

1: Initialize $q \in \mathbb{R}^{|\mathcal{I}|}$ with $q_i = -\frac{(\boldsymbol{b}_\mathcal{I})_i}{(\boldsymbol{a}_\mathcal{I})_i}$ if $(\boldsymbol{a}_\mathcal{I})_i \neq 0$ and $q_i = -1$ otherwise.
2: Sort $\boldsymbol{q}$ in ascending order to obtain the sorted indices $\mathcal{H} = \{j_1, j_2, \cdots, j_{|\mathcal{I}|}\}$.
3: Set $\mathcal{A} = \{i : (\boldsymbol{b}_\mathcal{I})_i > 0 \text{ or } [(\boldsymbol{b}_\mathcal{I})_i = 0, (\boldsymbol{a}_\mathcal{I})_i < 0]\}$ and $\bar{i}$ as the smallest index in $\mathcal{H}$ such that $q_{j_{\bar{i}}} \geq 0$.
4: Let $\bar{A} = h_k\eta_k^2 + \|\boldsymbol{a}_\mathcal{E}\|^2 + \sum_{i\in\mathcal{A}}(\boldsymbol{a}_\mathcal{I})_i^2$ and $\bar{B} = \iota + \boldsymbol{a}_\mathcal{E}^\mathrm{T}\boldsymbol{b}_\mathcal{E} + \sum_{i\in\mathcal{A}}(\boldsymbol{a}_\mathcal{I})_i(\boldsymbol{b}_\mathcal{I})_i$.
5: **If** $\bar{A} \neq 0$ and $-\frac{\bar{B}}{\bar{A}} \leq q_{j_{\bar{i}}}$, **return** $\varrho = \max\{0, -\frac{\bar{B}}{\bar{A}}\}$.
6: **For** $i = \bar{i}, \cdots, |\mathcal{H}|$ **do**
7: **If** $q_{j_i} \geq 1$, **return** $\varrho = 1$.
8: **If** $(a_\mathcal{I})_{j_i} < 0$ **then**
9: Update $\bar{A} = \bar{A} - (\boldsymbol{a}_\mathcal{I})_i^2$, $\bar{B} = \bar{B} - (\boldsymbol{a}_\mathcal{I})_i(\boldsymbol{b}_\mathcal{I})_i$.
10: **Else**
11: Update $\bar{A} = \bar{A} + (\boldsymbol{a}_\mathcal{I})_i^2$, $\bar{B} = \bar{B} + (\boldsymbol{a}_\mathcal{I})_i(\boldsymbol{b}_\mathcal{I})_i$.
12: **End If**
13: **If** $\bar{A} \neq 0$ and ($i = |\mathcal{H}|$ or $-\frac{\bar{B}}{\bar{A}} \leq q_{j_{i+1}}$), **return** $\varrho = \min\{1, -\frac{\bar{B}}{\bar{A}}\}$.
14: **End For**
15: **Return** $\varrho = 1$.

7 Numerical simulations

In this section, we will present some numerical simulations to demonstrate the practical performance of the proposed method AdaSSP, i.e. Algorithm 1 compared with MLALM [38], ICPPC [9] and SPD [23] for two nonconvex constrained problem with additional convex set constraints. In each iteration we will project the approximate solution of (5) obtained by 2 onto convex set constraints to get the next iterate. All the experiments were implemented in Matlab 2022a running on a 64-bit Ubuntu machine with a 2.00 Ghz Intel(R) Xeon(R) Gold 6330 CPU.

7.1 Quadratically constrained nonconvex program

The quadratically constrained nonconvex programs [23] takes the form

$$\begin{aligned} \min_{\boldsymbol{x}\in X} \quad & f(\boldsymbol{x}) = \frac{1}{N}\sum_{i=1}^{N}\log(1 + \frac{1}{2}\|\boldsymbol{P}_i\boldsymbol{x} - \boldsymbol{c}_i\|^2) \\ \text{s.t.} \quad & f_j(\boldsymbol{x}) = \frac{1}{2}\boldsymbol{x}^T\boldsymbol{Q}_j\boldsymbol{x} + \boldsymbol{a}_j^T\boldsymbol{x} \leq b_j, \quad j = 1, \ldots, M, \end{aligned} \tag{62}$$

where $X = [-10, 10]^n$. For each $i \in [N]$, we generate $\boldsymbol{P}_i \in \mathbb{R}^{p\times n}$ randomly with elements independently following the standard Gaussian distribution. For each $j \in [M]$, we generate diagonal matrices $\boldsymbol{Q}_j \in \mathbb{R}^{n\times n}$ with elements uniformly and randomly chosen from the interval $[0.5, 1]$, i.e., following $\mathrm{U}[0.5, 1]$ and vectors $\boldsymbol{a}_j \sim \mathrm{U}[0.1, 1.1]^n$. Then we generate a random point $\boldsymbol{x}_* \sim \mathrm{U}(0, 1)^n$ and set $\boldsymbol{c}_i = \boldsymbol{P}_i\boldsymbol{x}_*$, $i \in [N]$, $b_j = \frac{1}{2}\boldsymbol{x}_*^T\boldsymbol{Q}_j\boldsymbol{x}_* + \boldsymbol{a}_j^T\boldsymbol{x}_*$, $j \in [M]$. Note that $\boldsymbol{x}_*$ is feasible to (62) and $f(\boldsymbol{x}_*) = 0$, so $\boldsymbol{x}_*$ is the optimal solution of (62).

We first assess the impact of the adaptive penalty parameter β_k on the numerical performance of the proposed algorithm. We adaptively update β_k in Line 8 of Algorithm 1 by selecting $\beta_1 = 1, \Gamma = 100$, comparing to the case with fixed penalty parameter $\beta_k \in \{100, 400, 700\}$. In all experiments, we set $n = M = 50$, $p = 5$, $N = 1000$, and initialize the algorithm with $\boldsymbol{x}^1 = \boldsymbol{0}$. The maximum number of iterations is set to $K = 2000$, the parameters $\eta_k,\ \alpha_k,\ B_k,\ D_k$ and ϑ are chosen as (24) and $\rho_k = 1,\ \tau = 0.99,\ h_k = 1$. In Figure 1, the left subplot presents the trend of averaged objective function values at all previous iterates, while the right one illustrates the averaged constraint violation $\sum_{j=1}^{M}[f_j(\boldsymbol{x}) - b_j]_+$ over past iterates. We observe that the adaptive setting of β_k shows better performance in the reduction of objective function value compared with the fixed β_k in either small or large magnitude, while the constraint violation is almost in the same level for the two cases. This highlights the importance of adaptive penalty parameter to achieve better performance in terms of the objective function value.

We now compare AdaSSP with several closely related algorithms MLALM [38], ICPPC [9] and SPD [23] for solving (62). For all these four algorithms, we set the maximum number of stochastic gradient computations as 4000. We set $n \in \{50, 100\}$, $p = 5$, $N = 1000$, $M \in \{50, 100\}$. For all algorithms, we choose initial point $\boldsymbol{x}^1 = \boldsymbol{0}$, maximum iteration number $K = 2000$. For ICPPC, we set $k_0 = 2$, $\mathcal{M} = 0.1M$ and $(\mu_0, n) = (50, 100), (100, 200)$. All other parameters are set as required by [9]. It is worthy to mention that the maximum inner-iteration number of ICPPC is chosen as 1 according to its good performance in numerical tests. For MLALM, we set $\eta_k = 0.05/K^{1/4}$, $\alpha_k = 0.7$, $\beta_k = K^{1/4}$, $\rho_k = 10$ and for SPD, we set $\eta_k = 0.03/K^{1/4}$, $\beta_k = K^{1/4}$, $\rho_k = K^{1/4}$. As for AdaSSP, the parameters are the same as above. Figure 2 shows the performances of these algorithms regarding objective function values and constraint violations on QCNP problems under different scenarios. All the results are reported with average over 10 runs of each algorithm. The observations from the figures indicate that, within the same number of stochastic gradient evaluations, AdaSSP reduces the objective function value much faster throughout the algorithm's progress, while the speed to reduce the constraint violations is comparable with MLALM and SPD.

7.2 Multi-class Neyman-Pearson classification

In this subsection, we consider multi-class Neyman-Pearson classification (mNPC) problems [26]. The mNPC problem focuses on the task of learning K models $\boldsymbol{x}_k$, where $k \in [K]$, in order to predict the class of a potential data point $\boldsymbol{\xi}$ by selecting the model that maximizes the inner product $\boldsymbol{x}_k^T\boldsymbol{\xi}$. More precisely, the optimization problem aims to minimize the loss associated with one specific class while controlling the loss values for the remaining classes. The problem formulation can be expressed as follows:

$$\begin{aligned} \min_{\|\boldsymbol{x}_k\|\le\lambda,\, k\in[K]} \quad & \frac{1}{|\mathcal{D}_1|}\sum_{l>1}\sum_{\boldsymbol{\xi}\in\mathcal{D}_1} h(\boldsymbol{x}_1^T\boldsymbol{\xi} - \boldsymbol{x}_l^T\boldsymbol{\xi}) \\ \text{s.t.} \quad & \frac{1}{|\mathcal{D}_k|}\sum_{l\neq k}\sum_{\boldsymbol{\xi}\in\mathcal{D}_k} h(\boldsymbol{x}_k^T\boldsymbol{\xi} - \boldsymbol{x}_l^T\boldsymbol{\xi}) \le \gamma_k, \quad k = 2, \dots, K, \end{aligned} \tag{63}$$

where $h(z) = 1/(1 + e^z)$ is the loss function and $\mathcal{D}_k$ represents the training data of the k-th class. We use two datasets from LibSVM [12]: *covtype* ($K = 7$) and *mnist* ($K = 10$). Besides, we set $\gamma_k = 0.5(K - 1)$ and $\lambda = 0.3$.

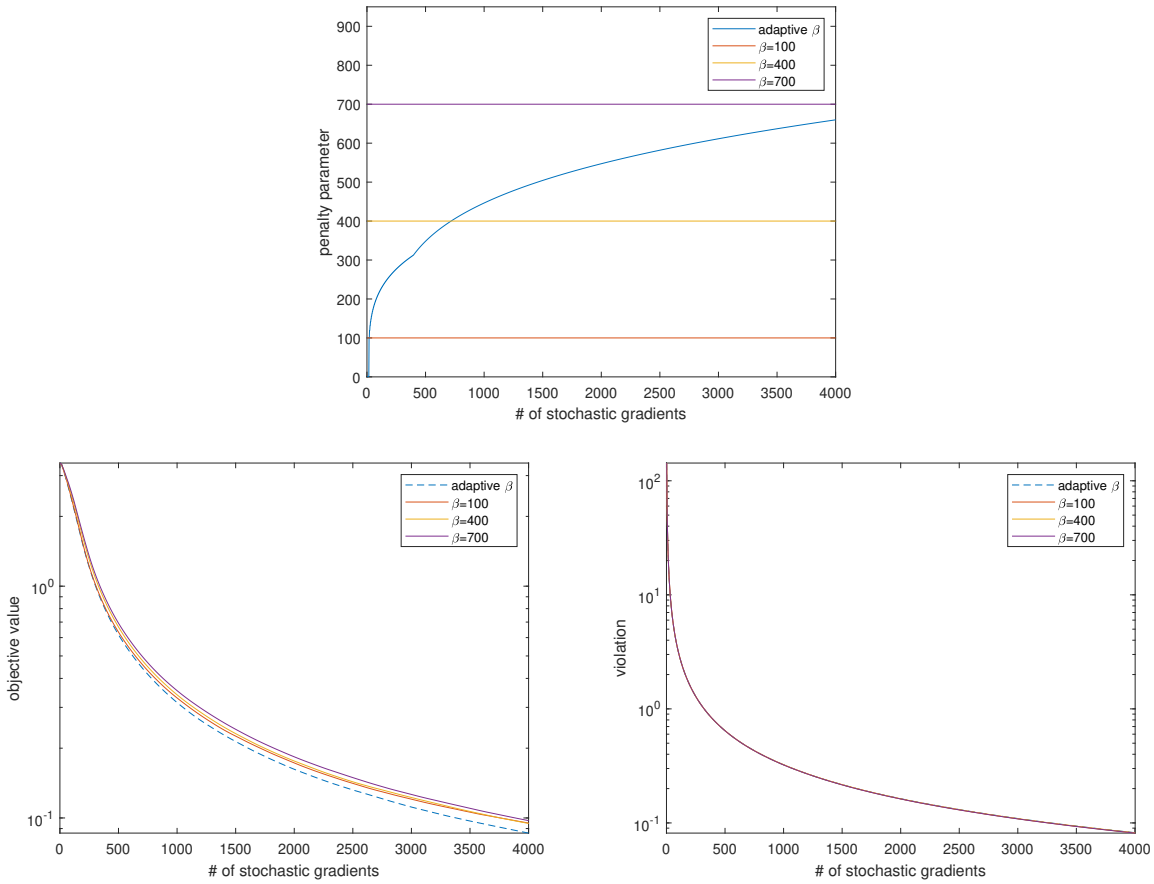

Fig. 1 The impact of β_k on AdaSSP

We now compare AdaSSP with SPD [23], MLALM [38], and ICPPC [9]. For these four algorithms, we set the maximum number of stochastic gradient computations is 4000. For AdaSSP, we set parameter $\alpha_k = 1 - k^{-1/2}$, $\eta_k = k^{-1/2}$ and $\tau = 0.99$ for both datasets. For dataset *covtype*, we set $\beta_1 = 1$, $\Gamma = 15$, $B_k = D_k = k^{-1/4}$, $h_k = 0.001k$, $\rho_k = 0.67$, while for dataset *mnist*, we set $\beta_1 = 10$, $\Gamma = 2$, $B_k = D_k = 100k^{-1}$, $h_k = k$, $\rho_k = 0.0067$. Meanwhile, the parameters are set same for MLALM, SPD and ICPPC as described as [38]. Figures 3 and 4 present the performances of these algorithms on the respective datasets for solving problem (63). All reported results are average from 10 independent runs of each algorithm. From the figures we observe that for the *covtype* dataset, AdaSSP exhibits superior performance in both the objective function value and the constraint violations. Regarding the dataset *mnist*, AdaSSP demonstrates a faster reduction in the objective function value compared to the other algorithms, while ICPPC maintains a lower level of constraint violations. Furthermore, it is worth mentioning that in every scenarios AdaSSP outperforms other algorithms significantly in reduction of objective function values while maintains guaranteed constraint violation convergence, indicating that the incorporation of adaptive penalty parameters bring notable benefits to the numerical performance.

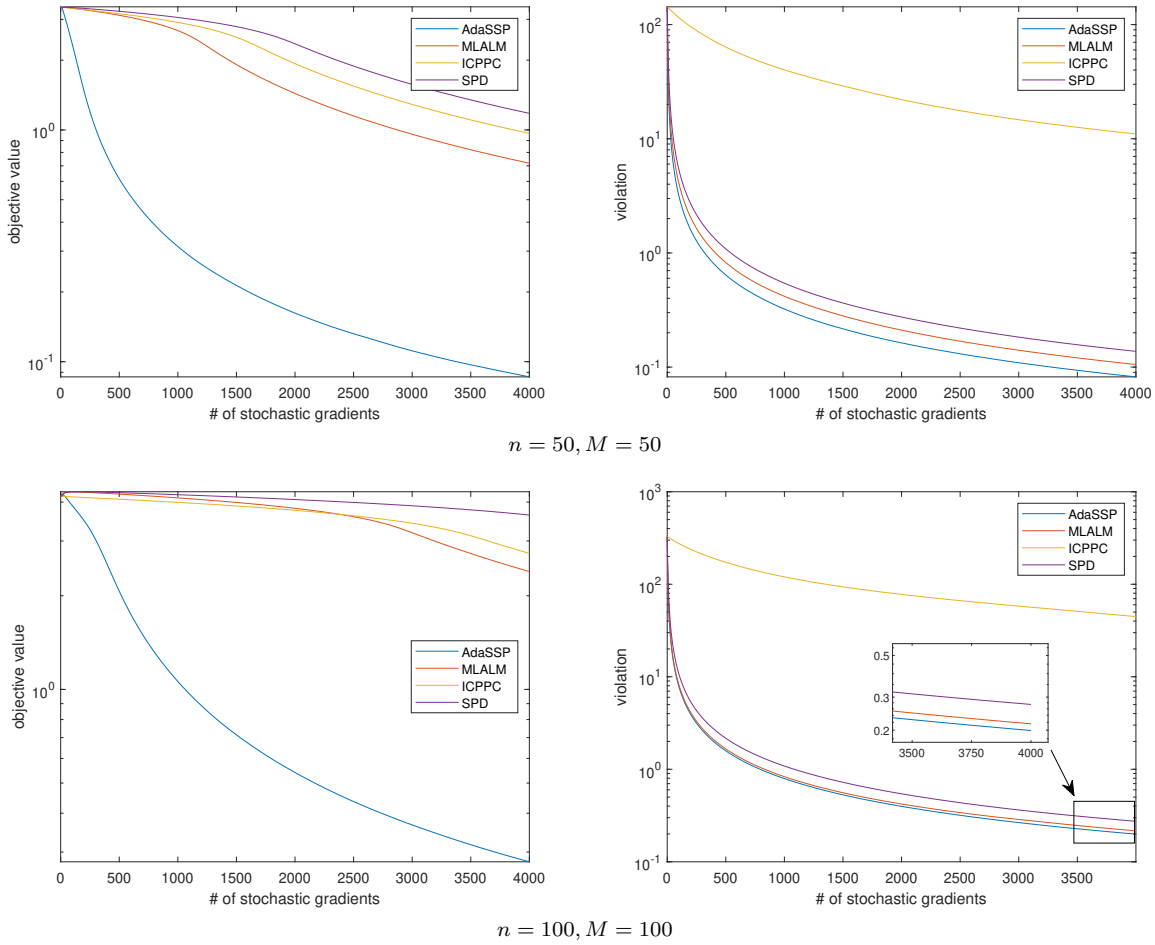

Fig. 2 Comparison between AdaSSP, MLALM, ICPPC and SPD for solving QCNP problems

8 Conclusions

In this paper we propose an adaptive single-loop stochastic penalty method, shorted as AdaSSP. This algorithm employs a single-loop momentum-based framework and adaptively updates penalty parameters along with the iterative progress, which differs from majority of existing methods that use a fixed sequence of penalty parameters. We establish that AdaSSP achieves an $\tilde{\mathcal{O}}(\epsilon^{-4})$ oracle complexity in terms of the total number of stochastic gradient evaluations to find an ϵ-KKT point with high probability, when local LICQ is assumed. We also provide a global convergence analysis for AdaSSP when the penalty parameter sequence is unbounded and bounded, respectively. Specifically, the sequence of average KKT measure at iterates convergence in expectation, either under the assumption of an extended variant of MFCQ or in the case when penalty parameters are bounded. Finally, numerical results on two test problems are reported.

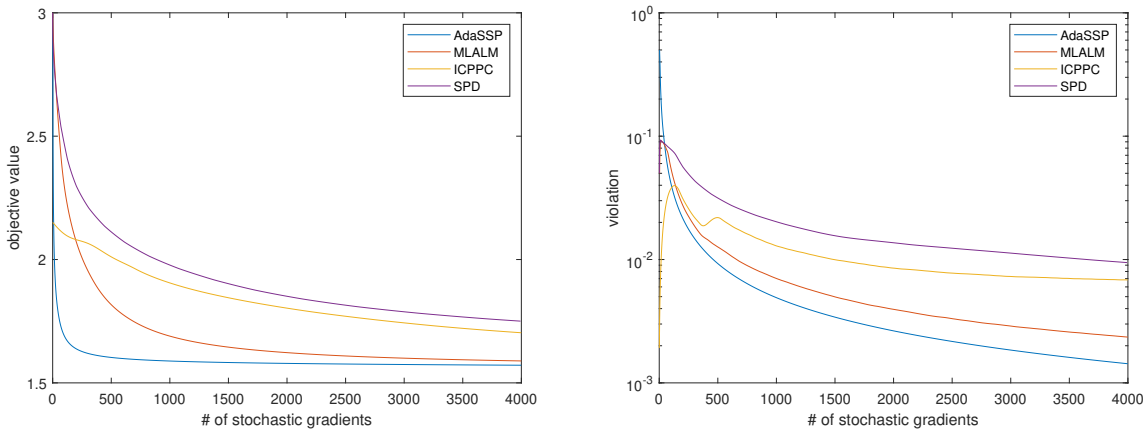

Fig. 3 Comparison between AdaSSP, MLALM, ICPPC and SPD on dataset *covtype*

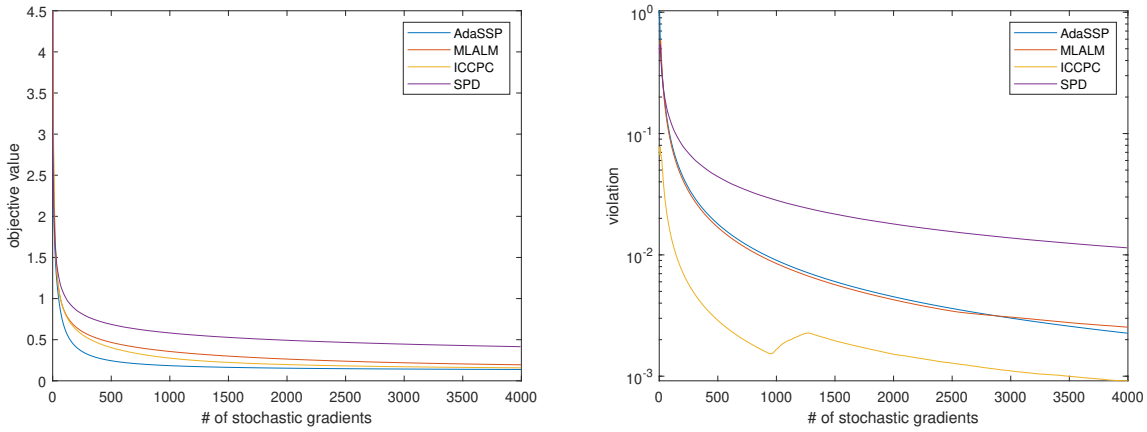

Fig. 4 Comparison between AdaSSP, MLALM, ICPPC and SPD on dataset *mnist*

Acknowledgements This work was partially supported by National Natural Science Foundation of China (No. 11271278) and Guangdong Province Key Laboratory of Computational Science at the Sun Yat-sen University.

References

1. 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. PMLR (2024)
2. Berahas, A.S., Curtis, F.E., Robinson, D., Zhou, B.: Sequential quadratic optimization for nonlinear equality constrained stochastic optimization. SIAM Journal on Optimization **31**(2), 1352–1379 (2021)
3. Berahas, A.S., Xie, M., Zhou, B.: A sequential quadratic programming method eith high-probability complexity bounds for nonlinear equality-constrained stochastic optimization. SIAM Journal on Optimization **35**(1), 240–269 (2025)

4. Bertsimas, D., Gupta, V., Kallus, N.: Robust sample average approximation. Mathematical Programming **171**, 217–282 (2018)
5. Betts, J.T.: Practical methods for optimal control and estimation using nonlinear programming. SIAM (2010)
6. Birgin, E.G., Martínez, J.M.: Practical augmented Lagrangian methods for constrained optimization. SIAM (2014)
7. Birgin, E.G., Martínez, J.M.: Complexity and performance of an augmented lagrangian algorithm. Optimization Methods and Software **35**(5), 885–920 (2020)
8. Bollapragada, R., Karamanli, C., Keith, B., Lazarov, B., Petrides, S., Wang, J.: An adaptive sampling augmented lagrangian method for stochastic optimization with deterministic constraints. Computers & Mathematics with Applications **149**, 239–258 (2023)
9. Boob, D., Deng, Q., Lan, G.: Stochastic first-order methods for convex and nonconvex functional constrained optimization. Mathematical Programming pp. 1436–4646 (2022)
10. Boob, D., Deng, Q., Lan, G.: Level constrained first order methods for function constrained optimization. Mathematical Programming pp. 1–61 (2024)
11. Cao, L., Berahas, A.S., Scheinberg, K.: First- and second-order high probability complexity bounds for trust-region methods with noisy oracles. Mathematical Programming **207**(1), 55–106 (2024)
12. Chang, C.C., Lin, C.J.: LIBSVM: A library for support vector machines. ACM Transactions on Intelligent Systems and Technology **2**, 27:1–27:27 (2011)
13. Chatterjee, N., Chen, Y.H., Maas, P., Carroll, R.J.: Constrained maximum likelihood estimation for model calibration using summary-level information from external big data sources. Journal of the American Statistical Association **111**(513), 107–117 (2016)
14. 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 p. 431–483 (2023)
15. 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)
16. Cutkosky, A., Mehta, H.: High-probability bounds for non-convex stochastic optimization with heavy tails. Advances in Neural Information Processing Systems **34**, 4883–4895 (2021)
17. Cutkosky, A., Orabona, F.: Momentum-based variance reduction in non-convex sgd. Advances in Neural Information Processing Systems **32** (2019)
18. Fatkhullin, I., Barakat, A., Kireeva, A., He, N.: Stochastic policy gradient methods: Improved sample complexity for fisher-non-degenerate policies. In: International Conference on Machine Learning, pp. 9827–9869. PMLR (2023)
19. Geyer, C.J.: Constrained maximum likelihood exemplified by isotonic convex logistic regression. Journal of the American Statistical Association **86**(415), 717–724 (1991)
20. Gorbunov, E., Danilova, M., Gasnikov, A.: Stochastic optimization with heavy-tailed noise via accelerated gradient clipping. Advances in Neural Information Processing Systems **33**, 15042–15053 (2020)
21. Iusem, A.N., Jofré, A., Oliveira, R.I., Thompson, P.: Variance-based extragradient methods with line search for stochastic variational inequalities. SIAM Journal on Optimization **29**(1), 175–206 (2019)
22. Jin, B., Scheinberg, K., Xie, M.: High probability complexity bounds for adaptive step search based on stochastic oracles. SIAM Journal on Optimization **34**(3), 2411–2439 (2024)
23. Jin, L., Wang, X.: A stochastic primal-dual method for a class of nonconvex constrained optimization. Computational Optimization and Applications **83**(1), 143–180 (2022)
24. Jin, L., Wang, X.: Stochastic nested primal-dual method for nonconvex constrained composition optimization. Mathematics of Computation **94**(351), 305–358 (2025)
25. Levy, K., Kavis, A., Cevher, V.: Storm+: Fully adaptive sgd with recursive momentum for nonconvex optimization. Advances in Neural Information Processing Systems **34**, 20571–20582 (2021)
26. Lin, Q., Ma, R., Xu, Y.: Complexity of an inexact proximal-point penalty method for constrained smooth non-convex optimization. Computational Optimization and Applications **82**(1), 175–224 (2022)
27. Liu, Z., Zhang, J., Zhou, Z.: Breaking the lower bound with (little) structure: Acceleration in non-convex stochastic optimization with heavy-tailed noise. In: The Thirty Sixth Annual Conference on Learning Theory, pp. 2266–2290. PMLR (2023)
28. Lu, Z., Mei, S., Xiao, Y.: Variance-reduced first-order methods for deterministically constrained stochastic nonconvex optimization with strong convergence guarantees. arXiv preprint arXiv:2409.09906 (2024)
29. Na, S., Anitescu, M., Kolar, M.: An adaptive stochastic sequential quadratic programming with differentiable exact augmented lagrangians. Mathematical Programming **199**(1), 721–791 (2023)
30. Na, S., Anitescu, M., Kolar, M.: Inequality constrained stochastic nonlinear optimization via active-set sequential quadratic programming. Mathematical Programming (2023)
31. Na, S., Mahoney, M.W.: Statistical inference of constrained stochastic optimization via sketched sequential quadratic programming. arXiv preprint arXiv:2205.13687 (2022)
32. Nandwani, Y., Pathak, A., Singla, P.: A primal dual formulation for deep learning with constraints. Advances in Neural Information Processing Systems **32** (2019)

33. O'Neill, M.J.: A two stepsize SQP method for nonlinear equality constrained stochastic optimization. arXiv preprint arXiv:2408.16656 (2024)
34. Ravi, S.N., Dinh, T., Lokhande, V.S., Singh, V.: Explicitly imposing constraints in deep networks via conditional gradients gives improved generalization and faster convergence. In: Proceedings of the AAAI conference on artificial intelligence, vol. 33, pp. 4772–4779 (2019)
35. Rees, T., Dollar, H.S., Wathen, A.J.: Optimal solvers for pde-constrained optimization. SIAM Journal on Scientific Computing **32**(1), 271–298 (2010)
36. Rockafellar, R.T.: The multiplier method of hestenes and powell applied to convex programming. Journal of Optimization Theory and Applications **12**(6), 555–562 (1973)
37. Shapiro, A., Dentcheva, D., Ruszczynski, A.: Lectures on stochastic programming: modeling and theory. SIAM (2021)
38. Shi, Q., Wang, X., Wang, H.: A momentum-based linearized augmented lagrangian method for nonconvex constrained stochastic optimization. Mathematics of Operations Research (2025)
39. Wang, B.C., Guo, J.J., Huang, P.Q., Meng, X.B.: A two-stage adaptive penalty method based on co-evolution for constrained evolutionary optimization. Complex & Intelligent Systems **9**(4), 4615–4627 (2023)
40. Wang, X.: Complexity analysis of inexact cubic-regularized primal-dual methods for finding second-order stationary points. Mathematics of Computation (2024)
41. Wang, X., Ma, S., Yuan, Y.: Penalty methods with stochastic approximation for stochastic nonlinear programming. Mathematics of Computation **86**(306), 1793–1820 (2017)
42. Ward, R., Wu, X., Bottou, L.: Adagrad stepsizes: Sharp convergence over nonconvex landscapes. Journal of Machine Learning Research **21**(219), 1–30 (2020)
43. Wright, S., Nocedal, J.: Numerical optimization. Springer (2006)
44. Xu, Y.: Primal-dual stochastic gradient method for convex programs with many functional constraints. SIAM Journal on Optimization **30**(2), 1664–1692 (2020)
45. Xu, Y., Liu, M., Lin, Q., Yang, T.: Admm without a fixed penalty parameter: Faster convergence with new adaptive penalization. Advances in Neural Information Processing Systems **30** (2017)
46. Xu, Y., Xu, Y.: Momentum-based variance-reduced proximal stochastic gradient method for composite nonconvex stochastic optimization. Journal of Optimization Theory and Applications **196**(1), 266–297 (2023)

Appendix A: An auxiliary lemma

To give high probability bounds for the estimation error, we will use the following lemma modified slightly from [27, Lemma 14] as a technical tool.

Lemma 14 *Suppose $X_{k\geq 1}$ is a martingale difference sequence adapted to the filtration $\mathcal{F}_{k\geq 1}$ in a Hilbert Space satisfying $\|X_k\| \leq R$ almost surely for some constant R and $\mathbb{E}\left[\|X_k\|^2 \mid \mathcal{F}_{k-1}\right] \leq \sigma_k^2$ almost surely for some constant σ_k^2 . Then with probability at least $1-\delta$, for any $k \geq 1$ it holds that $\|\sum_{s=1}^k X_s\| \leq 3(\sqrt{\sum_{s=1}^k \sigma_s^2} + R)\log\frac{3}{\delta}$.*

Appendix B: Proof of Lemma 7

Proof First, from $1-u \leq e^{-u}$ and the monotonicity of i^{-b}, we obtain

$$\begin{aligned}\sum_{j=1}^k j^{-a}\prod_{i=j+1}^k\left(1-i^{-b}\right) &\leq \sum_{j=1}^k j^{-a}\exp\left(-\sum_{i=j+1}^k i^{-b}\right)\\ &= \exp\left(-\sum_{i=1}^k i^{-b}\right)\sum_{j=1}^k j^{-a}\exp\left(\sum_{i=1}^j i^{-b}\right)\\ &\leq \exp\left(\frac{1-(k+1)^{1-b}}{1-b}\right)\sum_{j=1}^k j^{-a}\exp\left(\frac{j^{1-b}}{1-b}\right).\end{aligned}$$

Because the increasing monotonicity of $j^{-a}\exp\left(\frac{j^{1-b}}{1-b}\right)$ for $j \geq l_j := \left\lceil \max\{a^{\frac{1}{1-b}}, \left(\frac{a-b}{2}\right)^{\frac{1}{1-b}}\}\right\rceil$, one has

$$\sum_{j=l_j}^{k} j^{-a}\exp\left(\frac{j^{1-b}}{1-b}\right) \leq \int_{l_j}^{k+1} j^{-a}\exp\left(\frac{j^{1-b}}{1-b}\right)dj =: I.$$

It follows from partial integration that

$$\begin{aligned} I &= \int_{l_j}^{k+1} j^{b-a}j^{-b}\exp\left(\frac{j^{1-b}}{1-b}\right)dj \\ &= \left[j^{b-a}\exp\left(\frac{j^{1-b}}{1-b}\right)\right]_{j=a}^{j=k+1} - (b-a)\int_{l_j}^{k+1} j^{b-a-1}\exp\left(\frac{j^{1-b}}{1-b}\right)dj \\ &\leq (k+1)^{b-a}\exp\left(\frac{(k+1)^{1-b}}{1-b}\right) - {l_j}^{b-a}\exp\left(\frac{{l_j}^{1-b}}{1-b}\right) + (a-b){l_j}^{b-1}I, \end{aligned}$$

where the last inequality is due to $b \leq 1$. By the setting on l_j, we obtain that $(a-b){l_j}^{b-1} \leq \frac{1}{2}$ and hence

$$I \leq 2(k+1)^{b-a}\exp\left(\frac{(k+1)^{1-b}}{1-b}\right) - 2{l_j}^{b-a}\exp\left(\frac{{l_j}^{1-b}}{1-b}\right) \leq 2(k+1)^{b-a}\exp\left(\frac{(k+1)^{1-b}}{1-b}\right).$$

Combining the above components results in

$$\begin{aligned} \sum_{j=1}^{k} j^{-a}\prod_{i=j+1}^{k}\left(1-i^{-b}\right) &\leq \exp\left(\frac{1-(k+1)^{1-b}}{1-b}\right)\left(2(k+1)^{b-a}\exp\left(\frac{(k+1)^{1-b}}{1-b}\right) + \sum_{j=1}^{l_j-1} j^{-a}\exp\left(\frac{j^{1-b}}{1-b}\right)\right) \\ &\leq 2\exp\left(\frac{1}{1-b}\right)(k+1)^{b-a} + (l_j-1)\exp\left(\frac{(a-1)^{1-b}}{1-b}\right)\exp\left(\frac{1-(k+1)^{1-b}}{1-b}\right) \\ &\leq \left(2\exp\left(\frac{1}{1-b}\right) + (l_j-1)\exp\left(\frac{(a-1)^{1-b}}{1-b}\right)\exp\left(\frac{1+b-a}{1-b}\right)(a-b)^{\frac{a-b}{1-b}}\right)(k+1)^{b-a} \\ &\leq \left(2\exp\left(\frac{1}{1-b}\right) + (l_j-1)\exp\left(\frac{(a-1)^{1-b}+1+b-a}{1-b}\right)(a-b)^{\frac{a-b}{1-b}}\right)k^{b-a}, \end{aligned}$$

where the second inequality is due to the monotonicity of j^{-a} and $\exp\left(\frac{j^{1-b}}{1-b}\right)$, the third is from $\exp\left(\frac{1-(k+1)^{1-b}}{1-b}\right)(k+1)^{a-b} \leq \max_{u>0}\{\exp\left(\frac{1-u^{1-b}}{1-b}\right)u^{a-b}\} = \exp\left(\frac{1+b-a}{1-b}\right)(a-b)^{\frac{a-b}{1-b}}$, and the last is thanks to $b \leq a$. Thus the result is yielded.

Appendix C: Proof of Lemma 9

Proof We first prove (25). It yields from $\|\nabla f(\boldsymbol{x}^k)\| \leq G$ and (22) that

$$\begin{aligned} \|\boldsymbol{\theta}_k^b\| &= \|\boldsymbol{\theta}_k^b\|\mathbb{1}_{\|\nabla f(\boldsymbol{x}^k)\| > \frac{B_k}{2}} + \|\boldsymbol{\theta}_k^b\|\mathbb{1}_{\|\nabla f(\boldsymbol{x}^k)\| \leq \frac{B_k}{2}} \\ &\leq 3\|\nabla f(\boldsymbol{x}^k)\|\mathbb{1}_{\|\nabla f(\boldsymbol{x}^k)\| > \frac{B_k}{2}} + 2\sigma^2 B_k^{-1}\mathbb{1}_{\|\nabla f(\boldsymbol{x}^k)\| \leq \frac{B_k}{2}} \\ &= \left(3\|\nabla f(\boldsymbol{x}^k)\|B_k\mathbb{1}_{k^{\frac{1}{4}} < 2\|\nabla f(\boldsymbol{x}^k)\|} + 2\sigma^2\mathbb{1}_{k^{\frac{1}{4}} \geq 2\|\nabla f(\boldsymbol{x}^k)\|}\right)B_k^{-1} \\ &\leq \max\left\{3\|\nabla f(\boldsymbol{x}^k)\|B_k\mathbb{1}_{B_k < 2\|\nabla f(\boldsymbol{x}^k)\|}, 2\sigma^2\mathbb{1}_{B_k \geq 2\|\nabla f(\boldsymbol{x}^k)\|}\right\}B_k^{-1} \\ &\leq \max\left\{6\|\nabla f(\boldsymbol{x}^k)\|^2, 2\sigma^2\right\}B_k^{-1} \leq \max\left\{6G^2, 2\sigma^2\right\}k^{-\frac{1}{4}}, \end{aligned}$$

where the first inequality is by $\boldsymbol{\theta}_k^b = \mathbb{E}_{\xi^{[k]}}[\boldsymbol{G}^k] - \nabla f(x^k)$, $\|\boldsymbol{G}^k\| \leq B_k$, and

$$\begin{aligned}\|\boldsymbol{\theta}_k^b\|\mathbb{1}_{\|\nabla f(\boldsymbol{x}^k)\|>\frac{B_k}{2}} &\leq (\mathbb{E}_{\xi^k}[\|\boldsymbol{G}^k\|] + \|\nabla f(\boldsymbol{x}^k)\|)\mathbb{1}_{\|\nabla f(\boldsymbol{x}^k)\|>\frac{B_k}{2}} \\ &\leq (B_k + \|\nabla f(\boldsymbol{x}^k)\|)\mathbb{1}_{\|\nabla f(\boldsymbol{x}^k)\|>\frac{B_k}{2}} \\ &\leq 3\|\nabla f(\boldsymbol{x}^k)\|\mathbb{1}_{\|\nabla f(\boldsymbol{x}^k)\|>\frac{B_k}{2}}. \end{aligned} \tag{64}$$

Similarly, with (22) and

$$\|\nabla f(\boldsymbol{x}^k) - \nabla f(\boldsymbol{x}^{k-1})\| \leq L\|\boldsymbol{x}^k - \boldsymbol{x}^{k-1}\| \leq L\eta_{k-1}, \tag{65}$$

we have

$$\|\boldsymbol{Z}_k^b\| \leq \max\{6L^2\eta_{k-1}^2, 2L^2\eta_{k-1}^2\}D_k^{-1} = 6L^2\eta_{k-1}^{\frac{3}{2}}.$$

We now prove (26). Lemma 8 implies that

$$\begin{aligned}\mathbb{E}_{\xi^k}[\|\boldsymbol{\theta}_k^u\|^2] &= \mathbb{E}_{\xi^k}[\|\boldsymbol{\theta}_k^u\|^2]\mathbb{1}_{\|\nabla f(\boldsymbol{x}^k)\|>\frac{B_k}{2}} + \mathbb{E}_{\xi^k}[\|\boldsymbol{\theta}_k^u\|^2]\mathbb{1}_{\|\nabla f(\boldsymbol{x}^k)\|\leq\frac{B_k}{2}} \\ &\leq \mathbb{E}_{\xi^k}[\|\boldsymbol{G}^k\|^2]\mathbb{1}_{\|\nabla f(\boldsymbol{x}^k)\|>\frac{B_k}{2}} + \mathbb{E}_{\xi^k}[\|\boldsymbol{\theta}_k^u\|^2]\mathbb{1}_{\|\nabla f(\boldsymbol{x}^k)\|\leq\frac{B_k}{2}} \\ &\leq B_k^2\mathbb{1}_{\|\nabla f(\boldsymbol{x}^k)\|>\frac{B_k}{2}} + 10\sigma^2\mathbb{1}_{\|\nabla f(\boldsymbol{x}^k)\|\leq\frac{B_k}{2}} \\ &\leq 4\|\nabla f(\boldsymbol{x}^k)\|^2\mathbb{1}_{\|\nabla f(\boldsymbol{x}^k)\|>\frac{B_k}{2}} + 10\sigma^2\mathbb{1}_{\|\nabla f(\boldsymbol{x}^k)\|\leq\frac{B_k}{2}} \\ &\leq \max\{4\|\nabla f(\boldsymbol{x}^k)\|^2, 10\sigma^2\} \leq \max\{4G^2, 10\sigma^2\},\end{aligned}$$

where the first inequality is from

$$\mathbb{E}_{\xi^k}[\|\boldsymbol{\theta}_k^u\|^2] = \mathbb{E}_{\xi^k}[\|\boldsymbol{G}^k - \mathbb{E}_{\xi^k}[\boldsymbol{G}^k]\|^2] = \mathbb{E}_{\xi^k}[\|\boldsymbol{G}^k\|^2] - \mathbb{E}_{\xi^k}[\boldsymbol{G}^k]^2 \leq \mathbb{E}_{\xi^k}[\|\boldsymbol{G}^k\|^2],$$

the second inequality is from $\|\boldsymbol{G}^k\| \leq B_k$, and the last inequality is by $\|\nabla f(\boldsymbol{x}^k)\| \leq G$. Similarly, from (23), (65), and Lemma 8 we can obtain that

$$\mathbb{E}_{\xi^k}[\|\boldsymbol{Z}_k^u\|^2] \leq \max\{4L^2\eta_{k-1}^2, 10L^2\eta_{k-1}^2\} \leq 10L^2\eta_{k-1}^2.$$

Therefore, the proof is completed. □

Appendix D: Proof of Lemma 10

Proof First, from the parameter setting we will prove that for any $k \geq 1$,

$$\eta_{s-1}^2\alpha_s \leq s^{-1} = \eta_s^2. \tag{66}$$

For $s = 1$ and $s = 2$, the inequality holds naturally. For $s \geq 3$ there is

$$s \leq (s-1)^2 \Rightarrow s^{-\frac{1}{2}} \geq (s-1)^{-1} \Rightarrow 1 - s^{-\frac{1}{2}} \leq 1 - (s-1)^{-1} \leq s^{-1}/(s-1)^{-1},$$

where the last inequality is by $(s-2)/(s-1) \leq (s-1)/s$. Thus we have that $\eta_{s-1}^2\alpha_s \leq (s-1)^{-1}(1-s^{-\frac{1}{2}}) \leq s^{-1}$ for $s \geq 3$, which means (66) holds for all cases. It follows from (66) that

$$2\alpha_s D_s = 2\alpha_s\eta_{s-1}^{\frac{1}{2}} = 2(\alpha_s^4\eta_{s-1}^2)^{\frac{1}{4}} \leq 2(\alpha_s\eta_{s-1}^2)^{\frac{1}{4}} \leq 2(\eta_s)^{\frac{1}{2}}, \tag{67}$$

which together with Lemmas 7 and 9 and (14), (65), and (66) implies that

$$\begin{aligned}\sum_{s=1}^{k}\mathbb{E}_{\xi^s}\left[\left(\bar{\boldsymbol{U}}_s^k\right)^2\right] &\le \sum_{s=1}^{k}\mathbb{E}_{\xi^s}\left[\left\|\prod_{i=s}^{k}\alpha_i\boldsymbol{Z}_s^u\right\|^2\right] \le \sum_{s=1}^{k}\prod_{i=s}^{k}\alpha_i^2\mathbb{E}_{\xi^s}\left[\|\boldsymbol{Z}_s^u\|^2\right]\\ &\le \sum_{s=1}^{k}\prod_{i=s}^{k}\alpha_i\cdot 10\eta_{s-1}^2L^2 \le \sum_{s=1}^{k}\prod_{i=s+1}^{k}\alpha_i\cdot 10\eta_s^2L^2 \qquad (68)\\ &= 10L^2\sum_{s=1}^{k}s^{-1}\prod_{i=s+1}^{k}(1-i^{-\frac{1}{2}}) \le 10L^2\kappa(1,\frac{1}{2})k^{-\frac{1}{2}},\end{aligned}$$

and

$$\begin{aligned}\sum_{s=1}^{k}\mathbb{E}_{\xi^s}\left[\left(\bar{\boldsymbol{R}}_s^k\right)^2\right] &\le \sum_{s=1}^{k}\mathbb{E}_{\xi^s}\left[\left\|\prod_{i=s}^{k}\alpha_i\boldsymbol{Z}_s^u\right\|^4\right] \le \sum_{s=1}^{k}\prod_{i=s}^{k}\alpha_i^4 4D_s^2\mathbb{E}_{\xi^s}\left[\|\boldsymbol{Z}_s^u\|^2\right]\\ &\le \sum_{s=1}^{k}\prod_{i=s}^{k}\alpha_i^4\cdot 4D_s^2\eta_{s-1}^2 10L^2 = 40L^2\sum_{s=1}^{k}\prod_{i=s}^{k}\alpha_i^4\eta_{s-1}^3 \qquad (69)\\ &\le 40L^2\sum_{s=1}^{k}\prod_{i=s+1}^{k}\alpha_i^4\eta_s^3 \le 40L^2\sum_{s=1}^{k}\prod_{i=s+1}^{k}\alpha_i\eta_s^3\\ &= 40L^2\sum_{s=1}^{k}s^{-\frac{3}{2}}\prod_{i=s+1}^{k}(1-i^{-\frac{1}{2}}) \le 40L^2\kappa(\frac{3}{2},\frac{1}{2})k^{-\frac{3}{2}}.\end{aligned}$$

Next we will prove the following bound for $k \ge s-1$,

$$\prod_{i=s}^{k}\alpha_i\|\boldsymbol{Z}_s^u\| \le 2\sqrt{\eta_k}. \qquad (70)$$

First, this holds for all $s=1$ because $\boldsymbol{Z}_1^u = 0$. Now we consider the case when $s \ge 2$. For $k = s-1$, we have $\|\boldsymbol{Z}_s^u\| \le 2D_s = \sqrt{\eta_{k-1}}$. Suppose that $\prod_{i=s}^{t}\alpha_i\|\boldsymbol{Z}_s^u\| \le 2\sqrt{\eta_t}$ holds for $k = t \ge s-1$, then for $t+1$ we have $\prod_{i=s}^{t+1}\alpha_i\|\boldsymbol{Z}_s^u\| \le 2\alpha_{t+1}\sqrt{\eta_t} \le 2\sqrt{\eta_{t+1}}$ from (67). Thus (70) follows by induction.

From Lemma 14, (68), and $|\bar{\boldsymbol{U}}_s^k| \le \prod_{i=s}^{k}\alpha_i\|\boldsymbol{Z}_s^u\| \le 2\sqrt{\eta_k} \le 2k^{-\frac{1}{4}}$, we know that with probability at least $1-\frac{\delta}{10(k+1)\log^2(k+1)}$,

$$\begin{aligned}\left|\sum_{s=1}^{k}\bar{\boldsymbol{U}}_s^k\right| &\le 3(2k^{-\frac{1}{4}}+\sqrt{10L^2\kappa(1,\frac{1}{2})k^{-\frac{1}{2}}})\log\frac{30(k+1)\log^2(k+1)}{\delta}\\ &\le \left(18+9L\sqrt{10\kappa(1,\frac{1}{2})}\right)k^{-\frac{1}{4}}\log\left(\frac{4(k+1)}{\delta}\right), \qquad (71)\end{aligned}$$

where the last inequality is from $\log(k+1) \le k+1$ and $30^{1/3} \le 4$. From Lemma 14, (69), and

$$\begin{aligned}|\bar{\boldsymbol{R}}_s^k| &= |\,\|\prod_{i=s}^{k}\alpha_i\boldsymbol{Z}_s^u\|^2 - \mathbb{E}_{\xi^s}[\|\prod_{i=s}^{k}\alpha_i\boldsymbol{Z}_s^u\|^2]\,| \le \|\prod_{i=s}^{k}\alpha_i\boldsymbol{Z}_s^u\|^2 + \mathbb{E}_{\xi^s}[\|\prod_{i=s}^{k}\alpha_i\boldsymbol{Z}_s^u\|^2]\\ &\le \left(\prod_{i=s}^{k}\alpha_i\|\boldsymbol{Z}_s^u\|\right)^2 + \mathbb{E}_{\xi^s}\left[\left(\prod_{i=s}^{k}\alpha_i\|\boldsymbol{Z}_s^u\|\right)^2\right] \le 8\eta_k = 8k^{-\frac{1}{2}}\end{aligned}$$

by (13) and (70), it follows that with probability at least $1-\frac{\delta}{10(k+1)\log^2(k+1)}$,

$$\left|\sum_{s=1}^{k}\bar{\boldsymbol{R}}_s^k\right| \le 3(8k^{-\frac{1}{2}}+\sqrt{40L^2\kappa(\frac{3}{2},\frac{1}{2})k^{-1}})\log\frac{30(k+1)\log^2(k+1)}{\delta}$$

$$\leq \left(72 + 18L\sqrt{10\kappa(\frac{3}{2}, \frac{1}{2})}\right) k^{-\frac{1}{2}} \log\left(\frac{4(k+1)}{\delta}\right). \tag{72}$$

From (14) and Lemmas 7 and 9 we have

$$\begin{aligned}
&\sum_{s=1}^{k} \mathbb{E}_{\xi^s}\left[\left(\tilde{\boldsymbol{U}}_s^k\right)^2\right] \\
&\leq \sum_{s=1}^{k} \mathbb{E}_{\xi^s}\left[\left\|(1-\alpha_s)\prod_{i=s+1}^{k}\alpha_i\boldsymbol{\theta}_s^u\right\|^2\right] \leq \sum_{s=1}^{k}(1-\alpha_s)^2\prod_{i=s+1}^{k}\alpha_i^2\mathbb{E}_{\xi^s}\left[\|\boldsymbol{\theta}_s^u\|^2\right] \\
&\leq \sum_{s=1}^{k}(1-\alpha_s)^2\prod_{i=s+1}^{k}\alpha_i \cdot \max\{4G^2, 10\sigma^2\} \\
&= \max\{4G^2, 10\sigma^2\}\sum_{s=1}^{k} s^{-1}\prod_{i=s+1}^{k}(1-i^{-\frac{1}{2}}) \leq \max\{4G^2, 10\sigma^2\}\kappa(1, \frac{1}{2})k^{-\frac{1}{2}}
\end{aligned} \tag{73}$$

and

$$\begin{aligned}
&\sum_{s=1}^{k} \mathbb{E}_{\xi^s}\left[\left(\tilde{\boldsymbol{R}}_s^k\right)^2\right] \\
&\leq \sum_{s=1}^{k} \mathbb{E}_{\xi^s}\left[\left\|(1-\alpha_s)\prod_{i=s+1}^{k}\alpha_i\boldsymbol{\theta}_s^u\right\|^4\right] \leq \sum_{s=1}^{k}(1-\alpha_s)^4\prod_{i=s+1}^{k}\alpha_i^4 4B_s^2\mathbb{E}_{\xi^s}\left[\|\boldsymbol{\theta}_s^u\|^2\right] \\
&\leq \sum_{s=1}^{k}(1-\alpha_s)^4\prod_{i=s+1}^{k}\alpha_i \cdot 4B_s^2\max\{4G^2, 10\sigma^2\} \\
&\leq 4\max\{4G^2, 10\sigma^2\}\sum_{s=1}^{k} s^{-\frac{3}{2}}\prod_{i=s+1}^{k}(1-i^{-\frac{1}{2}}) \leq 4\max\{4G^2, 10\sigma^2\}\kappa(\frac{3}{2}, \frac{1}{2})k^{-1}.
\end{aligned} \tag{74}$$

Similar to (70), the following bound can be provided for $\boldsymbol{\theta}_s^u$ by induction:

$$(1-\alpha_s)\prod_{i=s+1}^{k}\alpha_i\|\boldsymbol{\theta}_s^u\| \leq 2k^{-\frac{1}{2}}B_k \leq 2k^{-\frac{1}{4}}. \tag{75}$$

Then due to (14), it yields

$$|\tilde{\boldsymbol{U}}_s^k| \leq \|(1-\alpha_s)\prod_{i=s+1}^{k}\alpha_i\boldsymbol{\theta}_s^u\| \leq (1-\alpha_s)\prod_{i=s+1}^{k}\alpha_i\|\boldsymbol{\theta}_s^u\| \leq 2k^{-\frac{1}{4}},$$

which together with Lemma 14 and (73) indicates that

$$\begin{aligned}
\left|\sum_{s=1}^{k}\tilde{\boldsymbol{U}}_s^k\right| &\leq 3(2k^{-\frac{1}{4}} + \sqrt{\max\{4G^2, 10\sigma^2\}\kappa(1, \frac{1}{2})k^{-\frac{1}{2}}})\log\frac{30(k+1)\log^2(k+1)}{\delta} \\
&\leq \left(18 + 9\sqrt{\max\{4G^2, 10\sigma^2\}\kappa(1, \frac{1}{2})}\right)k^{-\frac{1}{4}}\log\left(\frac{4(k+1)}{\delta}\right)
\end{aligned} \tag{76}$$

holds with probability at least $1 - \frac{\delta}{10(k+1)\log^2(k+1)}$. From Lemma 14, (74), and

$$|\tilde{\boldsymbol{R}}_s^k| \leq \left\|(1-\alpha_s)\prod_{i=s+1}^{k}\alpha_i\boldsymbol{\theta}_s^u\right\|^2 + \mathbb{E}_{\xi^s}\left[\left\|(1-\alpha_s)\prod_{i=s+1}^{k}\alpha_i\boldsymbol{\theta}_s^u\right\|^2\right] \leq 8k^{-\frac{1}{2}}$$

by (75), we know that with probability at least $1 - \frac{\delta}{10(k+1)\log^2(k+1)}$,

$$\left|\sum_{s=1}^{k} \tilde{\boldsymbol{R}}_s^k\right| \leq 3(8k^{-\frac{1}{2}} + \sqrt{4\max\{4G^2, 10\sigma^2\}\kappa(\frac{3}{2}, \frac{1}{2})k^{-1}}) \log \frac{30(k+1)\log^2(k+1)}{\delta}$$
$$\leq \left(72 + 18\sqrt{\max\{4G^2, 10\sigma^2\}\kappa(\frac{3}{2}, \frac{1}{2})}\right) k^{-\frac{1}{2}} \log\left(\frac{4(k+1)}{\delta}\right). \tag{77}$$

Thus the conclusion follows after combining (71), (72), (76), and (77). □

Appendix E: Proof of Lemma 11

Proof It follows from Lemma 10 that

$$\mathrm{P}\left(\bigcap_{i=1}^{K} \zeta_i\right) \geq 1 - \sum_{k=1}^{\infty} \frac{2\delta}{5(k+1)\log^2(k+1)} \geq 1 - \delta,$$

where the last inequality is from the fact that

$$\sum_{k=1}^{\infty} \frac{2}{(k+1)\log^2(k+1)} \leq \frac{1}{\log^2 2} + \int_1^{\infty} \frac{2}{(t+1)\log^2(t+1)} \mathrm{d}t = \frac{1}{\log^2 2} + \frac{1}{\log 2} \leq 5.$$

Hence, to derive the conclusion of this lemma it suffices to show that the occurrence of $\cap_{i=1}^{K}\zeta_i$ implies (27). From Lemmas 7 and 9 we obtain

$$\left\|\sum_{s=1}^{k}\prod_{i=s}^{k}\alpha_i \boldsymbol{Z}_s^b\right\| \leq 6L^2\sum_{s=1}^{k}(s-1)^{-\frac{3}{4}}\prod_{i=s}^{k}\alpha_i \leq 6L^2\sum_{s=1}^{k}s^{-\frac{3}{4}}\prod_{i=s+1}^{k}\alpha_i \leq 6L^2\kappa(\frac{3}{4}, \frac{1}{2})k^{-\frac{1}{4}}, \tag{78}$$

and

$$\left\|\sum_{s=1}^{k}(1-\alpha_s)\prod_{i=s+1}^{k}\alpha_i \boldsymbol{\theta}_s^b\right\| \leq \sum_{s=1}^{k}\max\{6G^2, 2\sigma^2\}s^{-\frac{1}{4}}(1-\alpha_s)\prod_{i=s+1}^{k}\alpha_i$$
$$\leq \max\{6G^2, 2\sigma^2\}\sum_{s=1}^{k}s^{-\frac{3}{4}}\prod_{i=s+1}^{k}(1-i^{-\frac{1}{2}}) \leq \max\{6G^2, 2\sigma^2\}\kappa(\frac{3}{4}, \frac{1}{2})k^{-\frac{1}{4}}. \tag{79}$$

From (11), (68), and (78), it implies that under the occurrence of $\cap_{i=1}^{K}\zeta_i$, for all $k \leq K$,

$$\left\|\sum_{s=1}^{k}\prod_{i=s}^{k}\alpha_i \boldsymbol{Z}_s\right\| \leq M_{\bar{U}}k^{-\frac{1}{4}}\log\left(\frac{4(k+1)}{\delta}\right) + \sqrt{2M_{\bar{R}}k^{-\frac{1}{2}}\log\left(\frac{4(k+1)}{\delta}\right)} + 2Lk^{-\frac{1}{4}}\sqrt{5\kappa(1, \frac{1}{2})} + 6L^2k^{-\frac{1}{4}}\kappa(\frac{3}{4}, \frac{1}{2}) \tag{80}$$
$$\leq \left(M_{\bar{U}} + \sqrt{2M_{\bar{R}}} + 2L\sqrt{5\kappa(1, \frac{1}{2})} + 6L^2\kappa(\frac{3}{4}, \frac{1}{2})\right)k^{-\frac{1}{4}}\log\left(\frac{4(K+1)}{\delta}\right)$$
$$= M_Z k^{-\frac{1}{4}}\log\left(\frac{4(K+1)}{\delta}\right).$$

From (12), (73), and (79), it implies that under the occurrence of $\cap_{i=1}^{K}\zeta_i$, for all $k \leq K$ we have:

$$\left\|\sum_{s=1}^{k}(1-\alpha_s)\prod_{i=s+1}^{k}\alpha_i \boldsymbol{\theta}_s\right\|$$

$$\leq M_{\tilde{U}} k^{-\frac{1}{4}} \log\left(\frac{4(k+1)}{\delta}\right) + \sqrt{2M_{\tilde{R}} k^{-\frac{1}{2}} \log\left(\frac{4(k+1)}{\delta}\right)} + 2\sigma k^{-\frac{1}{4}} \sqrt{5\kappa(1,\frac{1}{2})} + \max\{6G^2, 2\sigma^2\} k^{-\frac{1}{4}} \kappa(\frac{3}{4},\frac{1}{2}) \tag{81}$$

$$\leq \left(M_{\tilde{U}} + \sqrt{2M_{\tilde{R}}} + 2\sigma\sqrt{5\kappa(1,\frac{1}{2})} + \max\{6G^2, 2\sigma^2\}\kappa(\frac{3}{4},\frac{1}{2})\right) k^{-\frac{1}{4}} \log\left(\frac{4(K+1)}{\delta}\right)$$

$$= M_\theta k^{-\frac{1}{4}} \log\left(\frac{4(K+1)}{\delta}\right).$$

Then Lemma 2, together with (81) and (80), implies that under the occurrence of $\cap_{i=1}^K \zeta_i$,

$$\begin{aligned}
\sum_{k=1}^K \eta_k \|\boldsymbol{e}^k\| &\leq \sum_{k=1}^K \eta_k \left(\prod_{i=1}^k \alpha_i \left\|\nabla f(\boldsymbol{x}^1)\right\| + \left\|\sum_{s=1}^k \prod_{i=s}^k \alpha_i \boldsymbol{Z}_s\right\| + \left\|\sum_{s=1}^k (1-\alpha_s) \prod_{i=s+1}^k \alpha_i \boldsymbol{\theta}_s\right\|\right) \\
&\leq \sum_{k=1}^K k^{-\frac{1}{2}} \prod_{i=1}^k (1-i^{-\frac{1}{2}}) \left\|\nabla f(\boldsymbol{x}^1)\right\| + \sum_{k=1}^K (M_Z + M_\theta) k^{-\frac{3}{4}} \log\left(\frac{4(K+1)}{\delta}\right) \\
&\leq G\kappa(\frac{1}{2},\frac{1}{2}) + 4(M_Z + M_\theta) K^{\frac{1}{4}} \log\left(\frac{4(K+1)}{\delta}\right) = M_e K^{\frac{1}{4}} \log\left(\frac{4(K+1)}{\delta}\right),
\end{aligned} \tag{82}$$

where the last inequality is due to $\sum_{k=1}^K k^{-\frac{3}{4}} < 1 + \int_1^K k^{-\frac{3}{4}} \mathrm{d}k \leq 4K^{\frac{1}{4}}$. It indicates from Lemmas 4 and 5 that

$$\begin{aligned}
&\sum_{k=1}^K \min\{\eta_k, \frac{\|\boldsymbol{d}^k\|}{mG^2\beta_k + H}\} \|\boldsymbol{d}^k\| \\
&\leq 2\sum_{k=1}^K \left(\mathcal{L}_{\beta_k}(\boldsymbol{x}^k, \boldsymbol{\lambda}^k) - \mathcal{L}_{\beta_k}(\boldsymbol{x}^{k+1}, \boldsymbol{\lambda}^k)\right) + \sum_{k=1}^K \eta_k^2 L_{\beta_k} + 2\sum_{k=1}^K \eta_k \|\boldsymbol{e}^k\| \\
&\leq 2M_1 + mC^2\beta_{K+1} + \sum_{k=1}^K k^{-1}\tilde{L}\beta_k + 2\sum_{k=1}^K \eta_k \|\boldsymbol{e}^k\| \\
&\leq 2M_1 + (2mC^2\tilde{\Gamma} + 4\tilde{\Gamma}\tilde{L}) K^{\frac{1}{4}} + 2\sum_{k=1}^K \eta_k \|\boldsymbol{e}^k\|,
\end{aligned} \tag{83}$$

where the last inequality is from $\beta_k \leq \tilde{\Gamma} k^{1/4}$ with $\tilde{\Gamma} := \max\{\beta_1, \Gamma\}$ that leads to

$$\sum_{k=1}^K k^{-1}\tilde{L}\beta_k \leq \tilde{L}\tilde{\Gamma}\sum_{k=1}^K k^{-\frac{3}{4}} \leq 4\tilde{L}\tilde{\Gamma}K^{\frac{1}{4}} \text{ and } (k+1)^{\frac{1}{4}} \leq 2k^{\frac{1}{4}}, \ \forall k \geq 1.$$

Therefore, we have that

$$\begin{aligned}
&\frac{1}{K}\sum_{k=1}^K \|\boldsymbol{d}^k\| \\
&\leq \frac{1}{K}\sum_{k=1}^K \left(\frac{1}{\eta_k}\min\{\eta_k, \frac{\|\boldsymbol{d}^k\|}{mG^2\beta_k + H}\}\|\boldsymbol{d}^k\| \mathbb{1}_{\eta_k \leq \frac{\|\boldsymbol{d}^k\|}{mG^2\beta_k + H}} + (mG^2\beta_k + H)\eta_k \mathbb{1}_{\eta_k > \frac{\|\boldsymbol{d}^k\|}{mG^2\beta_k + H}}\right) \\
&\leq \frac{1}{K\eta_K}\sum_{k=1}^K \min\{\eta_k, \frac{\|\boldsymbol{d}^k\|}{mG^2\beta_k + H}\}\|\boldsymbol{d}^k\| + \frac{1}{K}\sum_{k=1}^K (mG^2\beta_k + H)\eta_k \\
&\leq \frac{2M_1 + (2mC^2\tilde{\Gamma} + 4\tilde{\Gamma}\tilde{L})K^{\frac{1}{4}} + 2\sum_{k=1}^K \eta_k\|\boldsymbol{e}^k\|}{K\eta_K} + \frac{mG^2\tilde{\Gamma}K^{\frac{1}{2}} + HK^{\frac{1}{4}}}{K}\sum_{k=1}^K k^{-\frac{3}{4}} \\
&\leq \frac{2M_1 + 4H + (2mC^2\tilde{\Gamma} + 4\tilde{\Gamma}\tilde{L} + 4mG^2\tilde{\Gamma})K^{\frac{1}{4}} + 2\sum_{k=1}^K \eta_k\|\boldsymbol{e}^k\|}{K^{1/2}},
\end{aligned} \tag{84}$$

which then yields that

$$\begin{aligned}
&\frac{1}{K}\sum_{k=1}^{K}\|\nabla_{\boldsymbol{x}}\mathcal{L}_{\beta_k}(\boldsymbol{x}^{k+1},\boldsymbol{\lambda}^k)\| \\
&\le \frac{1}{K}\sum_{k=1}^{K}\left(\|\nabla_{\boldsymbol{x}}\mathcal{L}_{\beta_k}(\boldsymbol{x}^{k},\boldsymbol{\lambda}^k)\| + \|\nabla_{\boldsymbol{x}}\mathcal{L}_{\beta_k}(\boldsymbol{x}^{k+1},\boldsymbol{\lambda}^k) - \nabla_{\boldsymbol{x}}\mathcal{L}_{\beta_k}(\boldsymbol{x}^{k},\boldsymbol{\lambda}^k)\|\right) \\
&\le \frac{1}{K}\sum_{k=1}^{K}\left(\|\boldsymbol{d}^k\| + \|\boldsymbol{e}^k\| + L_{\beta_k}\|\boldsymbol{x}^{k+1}-\boldsymbol{x}^k\|\right) \\
&\le \frac{1}{K}\sum_{k=1}^{K}\|\boldsymbol{d}^k\| + \frac{1}{K\eta_K}\sum_{k=1}^{K}\eta_k\|\boldsymbol{e}^k\| + \frac{1}{K}\sum_{k=1}^{K}L_{\beta_k}\eta_k \\
&\le \frac{2M_1 + 4H + (2mC^2\tilde{\Gamma} + 8\tilde{\Gamma}\tilde{L} + 4mG^2\tilde{\Gamma})K^{\frac{1}{4}} + 3\sum_{k=1}^{K}\eta_k\|\boldsymbol{e}^k\|}{K^{1/2}} \\
&\le (2M_1 + 4H + 2mC^2\tilde{\Gamma} + 8\tilde{\Gamma}\tilde{L} + 4mG^2\tilde{\Gamma} + 3M_e)K^{-\frac{1}{4}}\log\left(\frac{4(K+1)}{\delta}\right),
\end{aligned} \tag{85}$$

where the second last inequality is from (83) and

$$\sum_{k=1}^{K}L_{\beta_k}\eta_k \le \tilde{L}\tilde{\Gamma}K^{\frac{1}{2}}\sum_{k=1}^{K}k^{-\frac{3}{4}} \le 4\tilde{L}\tilde{\Gamma}K^{\frac{3}{4}},$$

and the last inequality is from (82) and (84) as well as $1 \le \log(\frac{4(K+1)}{\delta})$. Hence, we obtain from Lemma 10 that (27) holds with probability at least $1-\delta$.

Next we will show that

$$k^{-\frac{1}{4}}\log\left(\frac{k}{\delta}\right) \le \epsilon, \quad \forall k \ge \phi(\epsilon,\delta) := \max\left\{e^4, 4096\epsilon^{-4}\log^4\left(\frac{8}{e\delta^{\frac{1}{4}}\epsilon}\right)\right\}. \tag{86}$$

Let $\phi_0(\hat{\epsilon}) = \max\{1, e^4\hat{\epsilon}^{-4}\log^4\left(\frac{1}{\hat{\epsilon}}\right)\}$ for $\hat{\epsilon} > 0$, then we consider two mutually exclusive cases to prove that $(\phi_0(\hat{\epsilon}))^{-\frac{1}{4}}\log\left(\phi_0(\hat{\epsilon})\right) \le 8e^{-1}\hat{\epsilon}$.

(i) $e^4\hat{\epsilon}^{-4}\log^4\left(\frac{1}{\hat{\epsilon}}\right) \in (-\infty, 1)$. In this case we have $\phi_0(\hat{\epsilon}) < e^4$, together with the increasing monotonicity of $u^{-\frac{1}{4}}\log(u)$ for $u \in (0, e^4)$, we have $(\phi_0(\hat{\epsilon}))^{-\frac{1}{4}}\log\left(\phi_0(\hat{\epsilon})\right) < 0 < 8e^{-1}\hat{\epsilon}$.

(ii) $e^4\hat{\epsilon}^{-4}\log^4\left(\frac{1}{\hat{\epsilon}}\right) \in [1, +\infty)$. In this case we will have $\phi_0(\hat{\epsilon}) = e^4\hat{\epsilon}^{-4}\log^4\left(\frac{1}{\hat{\epsilon}}\right)$, which implies

$$\begin{aligned}
(\phi_0(\hat{\epsilon}))^{-\frac{1}{4}}\log\left(\phi_0(\hat{\epsilon})\right) &= \left(e^4\hat{\epsilon}^{-4}\log^4\left(\frac{1}{\hat{\epsilon}}\right)\right)^{-\frac{1}{4}}\log\left(e^4\hat{\epsilon}^{-4}\log^4\left(\frac{1}{\hat{\epsilon}}\right)\right) \\
&= 4e^{-1}\hat{\epsilon}\left(1 + \frac{1+\log(v)}{v}\right) \le 8e^{-1}\hat{\epsilon},
\end{aligned}$$

where the second equality is from the substitution: $v = \log\left(\frac{1}{\hat{\epsilon}}\right) > 0$ and the inequality is from $1 + \log(v) \le v$.

Let $\hat{\epsilon} = \frac{e\delta^{\frac{1}{4}}\epsilon}{8}$, then it holds that $(\phi_0(\frac{e\epsilon}{8}))^{-\frac{1}{4}}\log\left(\phi_0(\frac{e\epsilon}{8})\right) \le \delta^{\frac{1}{4}}\epsilon$. Because $u^{-\frac{1}{4}}\log(u)$ is monotonically decreasing for $u \ge e^4$, it holds that for any $t \ge \max\{e^4, \phi_0(\frac{e\delta^{\frac{1}{4}}\epsilon}{8})\}$ and $k \ge \delta t \ge e^4$,

$$k^{-\frac{1}{4}}\log\left(\frac{k}{\delta}\right) \le \delta^{-\frac{1}{4}}t^{-\frac{1}{4}}\log(t) \le (\phi_0(\frac{e\delta^{\frac{1}{4}}\epsilon}{8}))^{-\frac{1}{4}}\log(\phi_0(\frac{e\delta^{\frac{1}{4}}\epsilon}{8})) \le \delta^{-\frac{1}{4}}(\delta^{\frac{1}{4}}\epsilon) = \epsilon.$$

Thus (86) holds true with $\delta \in (0,1)$. From Lemma 11 and (86) we obtain that with probability at least $1-\delta$ such that for all $K \ge N(\epsilon,\delta) = \lceil\phi(\frac{\epsilon}{M_3}, \frac{\delta}{8})\rceil$,

$$\frac{1}{K}\sum_{k=1}^{K}\|\nabla_{\boldsymbol{x}}\mathcal{L}_{\beta_k}(\boldsymbol{x}^{k+1},\boldsymbol{\lambda}^k)\| \le M_3K^{-\frac{1}{4}}\log\left(\frac{4(K+1)}{\delta}\right)$$

$$\leq M_3 K^{-\frac{1}{4}} \log\left(\frac{8K}{\delta}\right) \leq M_3 \cdot \frac{\epsilon}{M_3} = \epsilon,$$

which indicates that (28) holds with probability at least $1-\delta$. The proof is completed. □

Appendix F: Proof of Lemma 12

Proof Recalling that $\{\boldsymbol{Z}_k^u\}_{k\geq 0}$ and $\{\boldsymbol{\theta}_k^u\}_{k\geq 0}$ are martingale difference sequences, we obtain from (26) and Lemma 7 that

$$\begin{aligned}
&\mathbb{E}_{\xi^{[k]}}\left[\left\|\sum_{s=1}^{k}\prod_{i=s}^{k}\alpha_i \boldsymbol{Z}_s^u\right\|\right] \leq \sqrt{\mathbb{E}_{\xi^{[k]}}\left[\left\|\sum_{s=1}^{k}\prod_{i=s}^{k}\alpha_i \boldsymbol{Z}_s^u\right\|^2\right]} = \sqrt{\sum_{s=1}^{k}\prod_{i=s}^{k}\alpha_i^2 \mathbb{E}_{\xi^{[k]}}\left[\|\boldsymbol{Z}_s^u\|^2\right]} \\
&\leq \sqrt{\sum_{s=1}^{k}\prod_{i=s}^{k}\alpha_i^2 10\eta_{s-1}^2 L^2} \leq \sqrt{\sum_{s=1}^{k}\prod_{i=s+1}^{k}\alpha_i 10\eta_s^2 L^2} = L\sqrt{10\sum_{s=1}^{k} s^{-1}\prod_{i=s+1}^{k}(1-i^{-\frac{1}{2}})} \\
&\leq Lk^{-\frac{1}{4}}\sqrt{10\kappa(1,\frac{1}{2})},
\end{aligned}$$

and

$$\begin{aligned}
&\mathbb{E}_{\xi^{[k]}}\left[\left\|\sum_{s=1}^{k}(1-\alpha_s)\prod_{i=s+1}^{k}\alpha_i\boldsymbol{\theta}_s^u\right\|\right] \leq \sqrt{\mathbb{E}_{\xi^{[k]}}\left[\left\|\sum_{s=1}^{k}(1-\alpha_s)\prod_{i=s+1}^{k}\alpha_i\boldsymbol{\theta}_s^u\right\|^2\right]} \\
&= \sqrt{\sum_{s=1}^{k}(1-\alpha_s)^2\prod_{i=s}^{k}\alpha_i^2\mathbb{E}_{\xi^{[k]}}\left[\|\boldsymbol{\theta}_s^u\|^2\right]} \leq \sqrt{\sum_{s=1}^{k}(1-\alpha_s)^2\prod_{i=s+1}^{k}\alpha_i \max\{4G^2,10\sigma^2\}} \\
&= \max\{G,\sigma\}\sqrt{10\sum_{s=1}^{k}s^{-1}\prod_{i=s+1}^{k}(1-i^{-\frac{1}{2}})} \leq (G+\sigma)k^{-\frac{1}{4}}\sqrt{10\kappa(1,\frac{1}{2})}.
\end{aligned}$$

Then it follows from (25), (78), and (79) as well as Lemma 2 that

$$\begin{aligned}
&\sum_{k=1}^{K}\eta_k\mathbb{E}_{\xi^{[k]}}\left[\|\boldsymbol{e}^k\|\right] \\
&\leq \sum_{k=1}^{K}\eta_k\mathbb{E}_{\xi^{[k]}}\left[\left(\prod_{i=1}^{k}\alpha_i\|\nabla f(\boldsymbol{x}^1)\| + \left\|\sum_{s=1}^{k}\prod_{i=s}^{k}\alpha_i\boldsymbol{Z}_s\right\| + \left\|\sum_{s=1}^{k}(1-\alpha_s)\prod_{i=s+1}^{k}\alpha_i\boldsymbol{\theta}_s\right\|\right)\right] \\
&\leq G\kappa(\frac{1}{2},\frac{1}{2}) + \sum_{k=1}^{K}\eta_k\mathbb{E}_{\xi^{[k]}}\left[\left(\left\|\sum_{s=1}^{k}\prod_{i=s}^{k}\alpha_i\boldsymbol{Z}_s^b\right\| + \left\|\sum_{s=1}^{k}\prod_{i=s}^{k}\alpha_i\boldsymbol{Z}_s^u\right\| + \left\|\sum_{s=1}^{k}(1-\alpha_s)\prod_{i=s+1}^{k}\alpha_i\boldsymbol{\theta}_s^b\right\| + \left\|\sum_{s=1}^{k}(1-\alpha_s)\prod_{i=s+1}^{k}\alpha_i\boldsymbol{\theta}_s^u\right\|\right)\right] \\
&\leq G\kappa(\frac{1}{2},\frac{1}{2}) + \sum_{k=1}^{K}\eta_k\left(6(L^2+G^2+\sigma^2)k^{-\frac{1}{4}}\kappa(\frac{3}{4},\frac{1}{2}) + (L+G+\sigma)k^{-\frac{1}{4}}\sqrt{10\kappa(1,\frac{1}{2})}\right) \qquad (87) \\
&\leq G\kappa(\frac{1}{2},\frac{1}{2}) + \sum_{k=1}^{K}6(L^2+G^2+\sigma^2)k^{-\frac{3}{4}}\kappa(\frac{3}{4},\frac{1}{2}) + (L+G+\sigma)k^{-\frac{3}{4}}\sqrt{10\kappa(1,\frac{1}{2})} \\
&\leq G\kappa(\frac{1}{2},\frac{1}{2}) + \left(24(L^2+G^2+\sigma^2)\kappa(\frac{3}{4},\frac{1}{2}) + 4(L+G+\sigma)\sqrt{10\kappa(1,\frac{1}{2})}\right)K^{\frac{1}{4}},
\end{aligned}$$

which together with (85) yields

$$\begin{aligned}&\frac{1}{K}\sum_{k=1}^{K}\mathbb{E}_{\xi^{[k]}}\left[\|\nabla_{\boldsymbol{x}}\mathcal{L}_{\beta_k}(\boldsymbol{x}^{k+1},\boldsymbol{\lambda}^k)\|\right]\\&\leq\frac{2M_1+4H+(2mC^2\tilde{\Gamma}+8\tilde{\Gamma}\tilde{L}+4mG^2\tilde{\Gamma})K^{\frac{1}{4}}+3\sum_{k=1}^{K}\eta_k\|\boldsymbol{e}^k\|}{K^{\frac{1}{2}}}\leq\frac{M_4}{K^{\frac{1}{4}}}.\end{aligned}$$

Thus (48) is derived. From Lemma 6 and (48) we have

$$\begin{aligned}&\frac{1}{K}\sum_{k=1}^{K}\mathbb{E}_{\xi^{[k]}}\left[\|\nabla\boldsymbol{c}_{\mathcal{E}}(\boldsymbol{x}^{k+1})\boldsymbol{c}_{\mathcal{E}}(\boldsymbol{x}^{k+1})+\nabla\boldsymbol{c}_{\mathcal{I}}(\boldsymbol{x}^{k+1})[\boldsymbol{c}_{\mathcal{I}}(\boldsymbol{x}^{k+1})]_+\|^{\frac{1}{2}}\right]\\&\leq\frac{1}{K}\sum_{k=1}^{K}\mathbb{E}_{\xi^{[k]}}\left[\beta_k^{-\frac{1}{2}}\|\nabla_{\boldsymbol{x}}\mathcal{L}_{\beta_k}(\boldsymbol{x}^{k+1},\boldsymbol{\lambda}^k)\|^{\frac{1}{2}}+\beta_k^{-\frac{1}{2}}M_2{}^{\frac{1}{2}}\right]\\&\leq\left(\mathbb{E}_{\xi^{[K]}}\left[\frac{1}{K}\sum_{k=1}^{K}\beta_k^{-1}\right]\right)^{\frac{1}{2}}\left\{\left(\frac{1}{K}\sum_{k=1}^{K}\mathbb{E}_{\xi^{[k]}}\left[\|\nabla_{\boldsymbol{x}}\mathcal{L}_{\beta_k}(\boldsymbol{x}^{k+1},\boldsymbol{\lambda}^k)\|\right]\right)^{\frac{1}{2}}+M_2{}^{\frac{1}{2}}\right\}\\&\leq\left(\mathbb{E}_{\xi^{[K]}}\left[\frac{1}{K}\sum_{k=1}^{K}\beta_k^{-1}\right]\right)^{\frac{1}{2}}\left\{\left(M_4K^{-\frac{1}{4}}\right)^{\frac{1}{2}}+M_2{}^{\frac{1}{2}}\right\},\end{aligned}$$

which indicates (49) thanks to Cauchy-Schwarz inequality and Jensen's inequality. The result is yielded. □

Appendix G: Proof of Lemma 13

Proof We assume that $\limsup_{k\to\infty}a_k=+\infty$, that is, for any $M>0$, there exists an infinite subsequence $\{a_k\}_{k\in\mathcal{K}}$ such that $a_k\geq M$, for all $k\in\mathcal{K}$. Following Assumption 6, for any $k\geq1$, there exists $\boldsymbol{p}^{k+1}\in\mathbb{R}^n$ with $\|\boldsymbol{p}^{k+1}\|\leq P$ such that

$$\begin{aligned}&\nabla c_i(\boldsymbol{x}^{k+1})^{\mathrm{T}}\boldsymbol{p}^{k+1}=-\varsigma\cdot\mathrm{sgn}(c_i(\boldsymbol{x}^{k+1})),&&i\in\mathcal{E}:c_i(\boldsymbol{x}^{k+1})\neq0;\\&\nabla c_i(\boldsymbol{x}^{k+1})^{\mathrm{T}}\boldsymbol{p}^{k+1}\leq-\varsigma,&&i\in\mathcal{I}:c_i(\boldsymbol{x}^{k+1})>0.\end{aligned}\tag{88}$$

From Assumption 5, (48), and the definition of $\nabla_{\boldsymbol{x}}\mathcal{L}$, we derive

$$\begin{aligned}&\left|\frac{1}{a_KK}\sum_{k=1}^{K}\mathbb{E}_{\xi^{[k]},\infty}[(\boldsymbol{p}^{k+1})^{\mathrm{T}}\nabla\boldsymbol{c}(\boldsymbol{x}^{k+1})\tilde{\boldsymbol{\lambda}}^{k+1}]\right|\\&\leq\frac{1}{a_KK}\sum_{k=1}^{K}\mathbb{E}_{\xi^{[k]},\infty}[\|\boldsymbol{p}^{k+1}\|\|\nabla f(\boldsymbol{x}^{k+1})-\nabla f(\boldsymbol{x}^{k+1})+\nabla\boldsymbol{c}(\boldsymbol{x}^{k+1})\tilde{\boldsymbol{\lambda}}^{k+1}\|]\\&\leq\frac{P}{a_KK}\sum_{k=1}^{K}\mathbb{E}_{\xi^{[k]},\infty}[\|\nabla f(\boldsymbol{x}^{k+1})\|+\|\nabla f(\boldsymbol{x}^{k+1})+\nabla\boldsymbol{c}(\boldsymbol{x}^{k+1})\tilde{\boldsymbol{\lambda}}^{k+1}\|]\\&\leq\frac{PG}{a_K}+\frac{P}{a_KK}\sum_{k=1}^{K}\mathbb{E}_{\xi^{[k]},\infty}[\|\nabla_{\boldsymbol{x}}\mathcal{L}_{\beta_k}(\boldsymbol{x}^{k+1},\boldsymbol{\lambda}^k)\|]\\&\leq\frac{PG+PM_4K^{-\frac{1}{4}}}{a_K}\leq\frac{PG+PM_4}{M},\end{aligned}$$

which implies that

$$\frac{1}{a_KK}\sum_{k=1}^{K}\mathbb{E}_{\xi^{[k]},\infty}[(\boldsymbol{p}^{k+1})^{\mathrm{T}}\nabla\boldsymbol{c}(\boldsymbol{x}^{k+1})\tilde{\boldsymbol{\lambda}}^{k+1}]\geq-\frac{PG+PM_4}{M}.$$

We next introduce notations:

$$\begin{aligned}&\mathcal{E}_1 := \{i \in \mathcal{E} : c_i(\boldsymbol{x}^{k+1}) \neq 0\}, \quad \mathcal{E}_2 := \{i \in \mathcal{E} : c_i(\boldsymbol{x}^{k+1}) = 0\},\\ &\mathcal{I}_1 := \{i \in \mathcal{I} : c_i(\boldsymbol{x}^{k+1}) > 0\}, \quad \mathcal{I}_2 := \{i \in \mathcal{I} : c_i(\boldsymbol{x}^{k+1}) \leq 0\}.\end{aligned} \tag{89}$$

For the sake of simplicity, and without causing any potential confusion, we omit the index k in the notations as defined in (89). For any $i \in \mathcal{E}_1$, it follows from $|\lambda_i^k| \geq |\tilde{\lambda}_i^{k+1}| - \beta_k |c_i(\boldsymbol{x}^{k+1})|$ that

$$\begin{aligned}(\boldsymbol{p}^{k+1})^{\mathrm{T}} \nabla c_i(\boldsymbol{x}^{k+1}) \tilde{\lambda}_i^{k+1} &= -\varsigma \cdot \mathrm{sgn}(c_i(\boldsymbol{x}^{k+1})) \tilde{\lambda}_i^{k+1} = -\varsigma \cdot \mathrm{sgn}(c_i(\boldsymbol{x}^{k+1}))(\lambda_i^k + \beta_k c_i(\boldsymbol{x}^{k+1}))\\ &\leq \varsigma |\lambda_i^k| - \delta \beta_k |c_i(\boldsymbol{x}^{k+1})| \leq 2\varsigma |\lambda_i^k| - \varsigma |\lambda_i^k| - \delta \beta_k |c_i(\boldsymbol{x}^{k+1})|\\ &\leq -\varsigma |\tilde{\lambda}_i^{k+1}| + 2\varsigma |\lambda_i^k|.\end{aligned}$$

In addition, for any $i \in \mathcal{I}_1$, it follows from (88) and $\tilde{\lambda}_i^{k+1} \geq 0$ that $(\boldsymbol{p}^{k+1})^T \nabla c_i(\boldsymbol{x}^{k+1}) \tilde{\lambda}_i^{k+1} \leq -\varsigma \tilde{\lambda}_i^{k+1}$. We then obtain from (88) that

$$\begin{aligned}&\mathbb{E}_{\xi^{[K]},\infty}[\frac{1}{a_K K} \sum_{k=1}^{K} (\boldsymbol{p}^{k+1})^{\mathrm{T}} \nabla c(\boldsymbol{x}^{k+1}) \tilde{\lambda}^{k+1}] \leq \frac{1}{a_K K} \mathbb{E}_{\xi^{[K]},\infty}[\sum_{k=1}^{K} (\boldsymbol{p}^{k+1})^{\mathrm{T}} \nabla c(\boldsymbol{x}^{k+1}) \tilde{\lambda}^{k+1}]\\
&\leq \frac{1}{a_K K} \mathbb{E}_{\xi^{[K]},\infty}[\sum_{k=1}^{K} (-\varsigma \|\tilde{\lambda}_{\mathcal{E}_1}^{k+1}\|_1 + 2\varsigma \|\lambda_{\mathcal{E}_1}^k\|_1 - \varsigma \|\tilde{\lambda}_{\mathcal{I}_1}^{k+1}\|_1 + (\boldsymbol{p}^{k+1})^{\mathrm{T}} \nabla c_{\mathcal{E}_2 \cup \mathcal{I}_2}(\boldsymbol{x}^{k+1}) \tilde{\lambda}_{\mathcal{E}_2 \cup \mathcal{I}_2}^{k+1})]\\
&= \frac{1}{a_K K} \mathbb{E}_{\xi^{[K]},\infty}[\sum_{k=1}^{K} -\varsigma \|\tilde{\lambda}^{k+1}\|_1 + 2\varsigma \|\lambda_{\mathcal{E}_1}^k\|_1 + \varsigma \|\tilde{\lambda}_{\mathcal{E}_2 \cup \mathcal{I}_2}^{k+1}\|_1 + (\boldsymbol{p}^{k+1})^{\mathrm{T}} \nabla c_{\mathcal{E}_2 \cup \mathcal{I}_2}(\boldsymbol{x}^{k+1}) \tilde{\lambda}_{\mathcal{E}_2 \cup \mathcal{I}_2}^{k+1}]\\
&\leq -\varsigma + \frac{1}{a_K K} \mathbb{E}_{\xi^{[K]},\infty}[\sum_{k=1}^{K} 2\varsigma \|\lambda_{\mathcal{E}_1}^k\|_1 + (\varsigma + PG) \|\tilde{\lambda}_{\mathcal{E}_2 \cup \mathcal{I}_2}^{k+1}\|_1]\\
&\leq -\varsigma + \frac{1}{a_K K} \mathbb{E}_{\xi^{[K]},\infty}[\sum_{k=1}^{K} 2\varsigma \|\lambda_{\mathcal{E}_1}^k\|_1 + (\varsigma + PG)(\|\lambda_{\mathcal{E}_2}^k + \beta_k c_{\mathcal{E}_2}(\boldsymbol{x}^{k+1})\|_1 + \|[\lambda_{\mathcal{I}_2}^k + \beta_k c_{\mathcal{I}_2}(\boldsymbol{x}^{k+1})]_+\|_1)]\\
&\leq -\varsigma + \frac{1}{a_K K} \mathbb{E}_{\xi^{[K]},\infty}[\sum_{k=1}^{K} 2\varsigma \|\lambda_{\mathcal{E}_1}^k\|_1 + (\varsigma + PG) \|\lambda_{\mathcal{E}_2 \cup \mathcal{I}_2}^k\|_1]\\
&\leq -\varsigma + \frac{m\Lambda(2\varsigma + PG)}{a_K} \leq -\varsigma + \frac{m\Lambda(2\varsigma + PG)}{M},\end{aligned}$$

where $\Lambda > 0$ is a constant such that $\|\lambda^k\| \leq \Lambda$ for any $k \geq 1$ by Lemma 3 and the setting of ρ_k, and the fourth inequality is due to $c_{\mathcal{E}_2}(\boldsymbol{x}^{k+1}) = 0$ and $c_{\mathcal{I}_2}(\boldsymbol{x}^{k+1}) \leq 0$. Summarizing above analysis we obtain $M \leq \varsigma^{-1}(PG + PM_4 + m\Lambda(2\varsigma + PG))$. However, this contradicts the arbitrariness of M. Thus we derive that $\limsup_{K \to \infty} a_K < +\infty$. The proof is completed. □

Appendix H: Proof of Theorem 4

Proof As shown in Lemma 13, there exists a constant $\tilde{\Lambda} > 0$ such that

$$\mathbb{E}_{\xi^{[K]},\infty}\Big[\frac{1}{K} \sum_{k=1}^{K} \|\tilde{\lambda}^{k+1}\|\Big] \leq \tilde{\Lambda} \quad \forall\, K \geq 1.$$

Recalling the proof in Lemma 3, there exists $\Lambda > 0$ such that $\|\lambda^k\| \leq \Lambda,\ k \geq 1$. By the definition of $\tilde{\lambda}^k$, we have

$$\mathbb{E}_{\xi^{[K]},\infty}\Big[\frac{1}{K} \sum_{k=1}^{K} \beta_k \Big(\|c_{\mathcal{E}}(\boldsymbol{x}^{k+1})\| + \|[c_{\mathcal{I}}(\boldsymbol{x}^{k+1})]_+\|\Big)\Big] \leq \tilde{\Lambda} + \Lambda \quad \forall\, K \geq 1. \tag{90}$$

Therefore, the following relation holds:

$$\begin{aligned}
&\frac{1}{K}\sum_{k=1}^{K}\mathbb{E}_{\xi^{[k]},\infty}\left[\left(\|c_{\mathcal{E}}(\boldsymbol{x}^{k+1})\|+\|[c_{\mathcal{I}}(\boldsymbol{x}^{k+1})]_+\|\right)^{\frac{1}{2}}\right]\\
&=\frac{1}{K}\sum_{k=1}^{K}\mathbb{E}_{\xi^{[k]},\infty}\left[\beta_k^{-\frac{1}{2}}\beta_k^{\frac{1}{2}}\left(\|c_{\mathcal{E}}(\boldsymbol{x}^{k+1})\|+\|[c_{\mathcal{I}}(\boldsymbol{x}^{k+1})]_+\|\right)^{\frac{1}{2}}\right]\\
&\le\frac{1}{K}\sum_{k=1}^{K}\left(\mathbb{E}_{\xi^{[k]},\infty}\left[\beta_k^{-1}\right]\right)^{\frac{1}{2}}\left(\mathbb{E}_{\xi^{[k]},\infty}\left[\beta_k\left(\|c_{\mathcal{E}}(\boldsymbol{x}^{k+1})\|+\|[c_{\mathcal{I}}(\boldsymbol{x}^{k+1})]_+\|\right)\right]\right)^{\frac{1}{2}}\\
&\le\left(\frac{1}{K}\sum_{k=1}^{K}\mathbb{E}_{\xi^{[k]},\infty}\left[\beta_k^{-1}\right]\right)^{\frac{1}{2}}\left(\frac{1}{K}\sum_{k=1}^{K}\mathbb{E}_{\xi^{[k]},\infty}\left[\beta_k\left(\|c_{\mathcal{E}}(\boldsymbol{x}^{k+1})\|+\|[c_{\mathcal{I}}(\boldsymbol{x}^{k+1})]_+\|\right)\right]\right)^{\frac{1}{2}}\\
&\le\left(\mathbb{E}_{\xi^{[K]},\infty}\left[\frac{1}{K}\sum_{k=1}^{K}\beta_k^{-1}\right]\right)^{\frac{1}{2}}\left(\tilde{\Lambda}+\Lambda\right)^{\frac{1}{2}},
\end{aligned}\tag{91}$$

where the first inequality uses the Cauchy-Schwarz inequality for the expectation of the product of two random variables, the second inequality uses the Cauchy-Schwarz inequality for the average of pairwise products of two sequences, and the last inequality uses (90). From the above inequalities, together with the inequality $a^{\frac{1}{2}}+b^{\frac{1}{2}}\le\sqrt{2}(a+b)^{\frac{1}{2}}$ and the condition on Ω_∞, we can derive (54).

Next, we prove (55). Note that when $c_i(\boldsymbol{x}^{k+1})\le-\frac{\lambda_i^k}{\beta_k}$, $\tilde{\lambda}_i^{k+1}=0$. Letting

$$\mathcal{I}_1:=\{i\in\mathcal{I}:c_i(\boldsymbol{x}^{k+1})>0\}\quad\text{and}\quad\mathcal{I}_2:=\{i\in\mathcal{I}:-\frac{\lambda_i^k}{\beta_k}\le c_i(\boldsymbol{x}^{k+1})\le 0\},\tag{92}$$

we have

$$\begin{aligned}
&\sum_{i\in\mathcal{I}}\left(\tilde{\lambda}_i^{k+1}|c_i(\boldsymbol{x}^{k+1})|\right)^{\frac{1}{4}}\\
&=\sum_{i\in\mathcal{I}_1}\left(\beta_k c_i^2(\boldsymbol{x}^{k+1})+\lambda_i^k c_i(\boldsymbol{x}^{k+1})\right)^{\frac{1}{4}}+\sum_{i\in\mathcal{I}_2}\left(-\lambda_i^k c_i(\boldsymbol{x}^{k+1})-\beta_k(c_i(\boldsymbol{x}^{k+1}))^2\right)^{\frac{1}{4}}\\
&\le\sum_{i\in\mathcal{I}_1}\left(\beta_k^{\frac{1}{4}}c_i^{\frac{1}{2}}(\boldsymbol{x}^{k+1})+(\lambda_i^k c_i(\boldsymbol{x}^{k+1}))^{\frac{1}{4}}\right)+\sum_{i\in\mathcal{I}_2}\left(-\lambda_i^k c_i(\boldsymbol{x}^{k+1})\right)^{\frac{1}{4}}\\
&\le\sum_{i\in\mathcal{I}_1}\beta_k^{\frac{1}{4}}c_i^{\frac{1}{2}}(\boldsymbol{x}^{k+1})+\sum_{i\in\mathcal{I}_1\cup\mathcal{I}_2}(\lambda_i^k|c_i(\boldsymbol{x}^{k+1})|)^{\frac{1}{4}}\\
&\le\beta_k^{\frac{1}{4}}\sum_{i\in\mathcal{I}_1}c_i^{\frac{1}{2}}(\boldsymbol{x}^{k+1})+\Lambda^{\frac{1}{4}}\sum_{i\in\mathcal{I}_1}c_i^{\frac{1}{4}}(\boldsymbol{x}^{k+1})+\Lambda^{\frac{1}{2}}|\mathcal{I}_2|\,\beta_k^{-\frac{1}{4}}\\
&\le|\mathcal{I}_1|^{\frac{3}{4}}\beta_k^{\frac{1}{4}}\|\boldsymbol{c}_{\mathcal{I}_1}(\boldsymbol{x}^{k+1})\|^{\frac{1}{2}}+|\mathcal{I}_1|^{\frac{7}{8}}\Lambda^{\frac{1}{4}}\|\boldsymbol{c}_{\mathcal{I}_1}(\boldsymbol{x}^{k+1})\|^{\frac{1}{4}}+\Lambda^{\frac{1}{2}}|\mathcal{I}_2|\,\beta_k^{-\frac{1}{4}}\\
&\le|\mathcal{I}|^{\frac{3}{4}}\beta_k^{\frac{1}{4}}\|[\boldsymbol{c}_{\mathcal{I}}(\boldsymbol{x}^{k+1})]_+\|^{\frac{1}{2}}+|\mathcal{I}|^{\frac{7}{8}}\Lambda^{\frac{1}{4}}\|[\boldsymbol{c}_{\mathcal{I}}(\boldsymbol{x}^{k+1})]_+\|^{\frac{1}{4}}+\Lambda^{\frac{1}{2}}|\mathcal{I}|\,\beta_k^{-\frac{1}{4}},
\end{aligned}\tag{93}$$

where the first inequality uses the non-negativity of the square function, the second uses the fact that the absolute-value function dominates the original variable, and the fourth inequality uses Jensen's inequality.

Taking full expectation on both sides of (93) and then averaging over the first K iterations yield

$$\begin{aligned}
&\frac{1}{K}\sum_{k=1}^{K}\mathbb{E}_{\xi^{[k]},\infty}\left[\sum_{i\in\mathcal{I}}\left(\tilde{\lambda}_i^{k+1}|c_i(\boldsymbol{x}^{k+1})|\right)^{\frac{1}{4}}\right]\\
&\le\frac{1}{K}\sum_{k=1}^{K}\left(\mathbb{E}_{\xi^{[k]},\infty}\left[|\mathcal{I}|^{\frac{3}{4}}\beta_k^{\frac{1}{4}}\|[c_{\mathcal{I}}(\boldsymbol{x}^{k+1})]_+\|^{\frac{1}{2}}+|\mathcal{I}|^{\frac{7}{8}}\Lambda^{\frac{1}{4}}\|[c_{\mathcal{I}}(\boldsymbol{x}^{k+1})]_+\|^{\frac{1}{4}}+\Lambda^{\frac{1}{2}}|\mathcal{I}|\,\beta_k^{-\frac{1}{4}}\right]\right)
\end{aligned}$$

$$\leq |\mathcal{I}|^{\frac{3}{4}} \Big(\frac{1}{K} \sum_{k=1}^{K} \mathbb{E}_{\xi^{[k]},\infty} \big[\beta_k \| [c_{\mathcal{I}}(\boldsymbol{x}^{k+1})]_+ \| \big] \Big)^{\frac{1}{2}} \Big(\mathbb{E}_{\xi^{[K]},\infty} \big[\frac{1}{K} \sum_{k=1}^{K} \beta_k^{-\frac{1}{2}} \big] \Big)^{\frac{1}{2}}$$
$$+ \ |\mathcal{I}|^{\frac{7}{8}} \Lambda^{\frac{1}{4}} \Big(\frac{1}{K} \sum_{k=1}^{K} \mathbb{E}_{\xi^{[k]},\infty} \big[\| [c_{\mathcal{I}}(\boldsymbol{x}^{k+1})]_+ \|^{\frac{1}{2}} \big] \Big)^{\frac{1}{2}} + \mathbb{E}_{\xi^{[K]},\infty} \big[\frac{|\mathcal{I}| \Lambda^{\frac{1}{2}}}{K} \sum_{k=1}^{K} \beta_k^{-\frac{1}{4}} \big],$$

where the second inequality uses the same Cauchy-Schwarz argument as in (91) and the fact that $(\mathbb{E}[u])^2 \leq \mathbb{E}[u^2]$ for a positive random variable $u > 0$. Finally, combining with (90), (91) and the property that $\beta_k \to \infty$, we obtain the desired conclusion. □