Journal of Scientific Computing (2025) 104:46
https://doi.org/10.1007/s10915-025-02934-w

A Smoothed Augmented Lagrangian Framework for Convex Optimization with Nonsmooth Constraints

Peixuan Zhang[1] · Uday V. Shanbhag[2] · Ethan X. Fang[3]

Received: 26 November 2023 / Revised: 25 December 2024 / Accepted: 16 January 2025 /
Published online: 21 June 2025

Abstract
Augmented Lagrangian (AL) methods have proven remarkably useful in solving optimization problems with complicated constraints. The last decade has seen the development of overall complexity guarantees for inexact AL variants. Yet, a crucial gap persists in addressing nonsmooth convex constraints. To this end, we present a smoothed augmented Lagrangian (AL) framework where nonsmooth terms are progressively smoothed with a smoothing parameter η_k. The resulting AL subproblems are η_k-smooth, allowing for leveraging accelerated schemes. By a careful selection of the inexactness level ϵ_k (for inexact subproblem resolution), the penalty parameter ρ_k, and smoothing parameter η_k at epoch k, we derive rate and complexity guarantees of $\tilde{\mathcal{O}}(1/\boldsymbol{\varepsilon}^{3/2})$ and $\tilde{\mathcal{O}}(1/\boldsymbol{\varepsilon})$ in convex and strongly convex regimes for computing an $\boldsymbol{\varepsilon}$-optimal solution, when ρ_k increases at a geometric rate, a significant improvement over the best available guarantees for AL schemes for convex programs with nonsmooth constraints. Analogous guarantees are developed for settings with $\rho_k = \rho$ as well as $\eta_k = \eta$. Preliminary numerics on a fused Lasso problem display promise.

Keywords Augmented Lagrangian · Convex Optimization · Smoothing

Uday V. Shanbhag would like to dedicate this paper to Prof. Michael A. Saunders for his help, mentorship, and guidance as well as his immense and enduring contributions to the theoretical development and large-scale implementation of algorithms for nonlinear programming.

✉ Uday V. Shanbhag
udaybag@umich.edu

Peixuan Zhang
pqz5090@psu.edu

Ethan X. Fang
ethan.fang@duke.edu

[1] Pennsylvania State University, University Park, Pennsylvania, PA, USA

[2] University of Michigan, Ann Arbor, MI, USA

[3] Duke University, Durham, NC, USA

1 Introduction

We consider the nonsmooth convex program, defined as

$$\min_{\mathbf{x}\in\mathcal{X}} amp;\ \{f(\mathbf{x}) \mid g(\mathbf{x}) \leq 0\}, \tag{NSCopt}$$

where $f : \mathcal{X} \to \mathbb{R}$ is a real-valued convex function and is possibly nonsmooth (but smoothable), $\mathcal{X} \subset \mathbb{R}^n$ is a closed and convex set, and $g(\mathbf{x}) = (g_1(\mathbf{x}), g_2(\mathbf{x}), ..., g_m(\mathbf{x}))^\top$ that each $g_i : \mathcal{X} \to \mathbb{R}, i = 1, 2, \cdots, m$ is a possibly complicated nonsmooth (but smoothable) convex function. Generally, the presence of such constraints precludes usage of projection-based methods to ensure feasibility of iterates. In deterministic regimes, a host of approaches have been employed for contending with complicated constraints, a subset of which include sequential quadratic programming [18, 43], interior point methods [8], and augmented Lagrangian (AL) schemes [38, 39]. Of these, AL schemes have proven to be enormously influential in the context of scientific computing [1, 9, 13], and more specifically in nonlinear programming in the form of solvers such as `minos` [16, 28] and `lancelot` [10] as well as more refined techniques [15, 17]. There has been a significant interest in deriving overall complexity bounds [24, 44] in convex regimes when the Lagrangian subproblem is solved via a first-order method. However, such bounds tend to be poor when constraints are possibly nonsmooth; e.g. standard AL schemes display complexity guarantees of $\mathcal{O}(\boldsymbol{\varepsilon}^{-5})$ for computing an $\boldsymbol{\varepsilon}$-optimal solution in such settings (see Table 1).

Gap and Relevance: Existing ALM schemes for **nonlinear** and **nonsmooth** convex constraints display poor overall complexity in inner (subgradient) steps. Such models are relevant when addressing compositional and risk constraints.

1.1. Related work. Before proceeding, we discuss related prior research. (a) *Augmented Lagrangian Methods.* The augmented Lagrangian method (ALM) was proposed by Hestenes [19] and Powell [37] with a comprehensive rate analysis subsequently provided by Rockafellar [38]. The ALM framework relies on solving a sequence of unconstrained (or relaxed) problems, requiring the minimization of a suitably defined augmented Lagrangian function $\mathcal{L}_\rho(\mathbf{x}, \lambda)$ in $\mathbf{x}$, where ρ and λ denote the penalty parameter and the Lagrange multiplier associated with g, respectively. In high-dimensional settings, the Lagrangian subproblems cannot be solved exactly, leading to the development of variants that allow for inexact resolution of the Lagrangian subproblem. Kang et al. [21] presented an inexact accelerated ALM for strongly convex optimization with linear constraints at a rate of $\mathcal{O}(1/k^2)$, where k is the iteration counter. Non-ergodic convergence guarantees were provided in [24, 25], where either smoothness of f [24] or a composite structure [25] is assumed. Overall complexity guarantees were first provided by Lan and Monteiro [24], Aybat and Iyengar [4], Necoara et al. [29] and most recently Lu and Zhou [26], where the latter three references allowed for conic settings. In fact, Lu and Zhou [26] showed that in conic convex settings with smooth nonlinear constraints, by introducing a regularization, the overall complexity is improved to $\mathcal{O}\left(\boldsymbol{\varepsilon}^{-1}\ln(\boldsymbol{\varepsilon}^{-1})\right)$ with a geometrically increasing penalty parameter. Nedelcu et al. [30] considered convex and strongly convex regimes. Notably, Necoara et al. [29] derived an overall complexity of $\mathcal{O}(\boldsymbol{\varepsilon}^{-\frac{3}{2}})$ and $\mathcal{O}(\boldsymbol{\varepsilon}^{-1})$ for smooth settings objective, respectively. More recently, Xu [44] considered nonlinear but **smooth** regimes in proposing an inexact ALM (under a suitable boundedness requirement) with complexity guarantees of $\mathcal{O}(\boldsymbol{\varepsilon}^{-1})$ (under convex f) and $\mathcal{O}(\boldsymbol{\varepsilon}^{-\frac{1}{2}}\log(\boldsymbol{\varepsilon}^{-1}))$ (under strongly convex f), respectively. Table 1 compares existing complexity guarantees for AL schemes with both our schemes in convex

Table 1 ALM for deterministic convex optimization

Ref.	f	g	Metrics	ρ_k	Rate	Complex.	Comment
[38]	S	NL+S	$\|\mathbf{x}_k - \mathbf{x}^*\|, \|\lambda_k - \lambda^*\|$	ρ_0	–	–	nonlinear
[4]	NS‡	L	$\|f(\mathbf{x}_k) - f^*\|, d\_(g(\mathbf{x}_k))$	$\rho_0\zeta^k$	$\mathcal{O}\left(\frac{1}{\rho_k}\right)$	$\tilde{\mathcal{O}}\left(\boldsymbol{\varepsilon}^{-1}\right)$	Composite conic
[24]	S	L	$(\boldsymbol{\varepsilon}_p, \boldsymbol{\varepsilon}_d)-$optimal	ρ_0	$\mathcal{O}\left(\frac{1}{\sqrt{k}}\right)$	$\mathcal{O}\left(\boldsymbol{\varepsilon}^{-7/4}\right)$	smooth linear
[35]	S	L	$\|f(\mathbf{x}_k) - f^*\|, d\_(g(\mathbf{x}_k))$	ρ_0	$\mathcal{O}\left(\frac{1}{k^2}\right)$	$\mathcal{O}\left(\boldsymbol{\varepsilon}^{-5/4}\right)$	smooth linear
				$\rho_0\zeta^k$	$\mathcal{O}\left(\frac{1}{k^2}\right)$	$\mathcal{O}\left(\boldsymbol{\varepsilon}^{-1}\right)$	smooth linear
[26]	NS‡	NL + S	$\|f(\mathbf{x}_k) - f^*\|, d\_(g(\mathbf{x}_k))$	poly.	-	$\mathcal{O}\left(\boldsymbol{\varepsilon}^{-7/4}\right)$	smooth nonlinear
				$\rho_0\zeta^k$	-	$\tilde{\mathcal{O}}\left(\boldsymbol{\varepsilon}^{-1}\right)^\diamond$	smooth nonlinear
[44]*	NS‡	NL + S	$\|f(\mathbf{x}_k) - f^*\|, d\_(g(\mathbf{x}_k))$	ρ_0	$\mathcal{O}\left(\frac{1}{k}\right)$	$\mathcal{O}\left(\boldsymbol{\varepsilon}^{-3/2}\right)$	smooth nonlinear
				$\rho_0\zeta^k$	$\mathcal{O}\left(\frac{1}{\rho_k}\right)$	$\mathcal{O}\left(\boldsymbol{\varepsilon}^{-1}\right)$	smooth nonlinear
Sm-AL	NS	NL+NS	$\|f^* - \mathcal{D}_\rho(\bar{\lambda}_K)\|$	ρ_0	$\mathcal{O}\left(\frac{1}{\sqrt{k}}\right)$	$\mathcal{O}\left(\boldsymbol{\varepsilon}^{-3}\right)$	nonsmth nonlinear
			$\|f(\mathbf{x}_k) - f^*\|, d\_(g(\mathbf{x}_k))$	$\rho_0\zeta^k$	$\mathcal{O}\left(\frac{1}{\rho_k}\right)$	$\tilde{\mathcal{O}}\left(\boldsymbol{\varepsilon}^{-3/2}\right)$	nonsmth nonlinear
Sm-AL(S†)	NS	NL+NS	$\|f^* - \mathcal{D}_\rho(\bar{\lambda}_K)\|$	ρ_0	$\mathcal{O}\left(\frac{1}{\sqrt{k}}\right)$	$\tilde{\mathcal{O}}\left(\boldsymbol{\varepsilon}^{-2}\right)$	nonsmth nonlinear
			$\|f(\mathbf{x}_k) - f^*\|, d\_(g(\mathbf{x}_k))$	$\rho_0\zeta^k$	$\mathcal{O}\left(\frac{1}{\rho_k}\right)$	$\tilde{\mathcal{O}}\left(\boldsymbol{\varepsilon}^{-1}\right)$	nonsmth nonlinear
N-AL	NS	NL+NS	$\|f^* - \mathcal{D}_\rho(\bar{\lambda}_K)\|$	ρ_0	$\mathcal{O}\left(\frac{1}{\sqrt{k}}\right)$	$\mathcal{O}\left(\boldsymbol{\varepsilon}^{-5}\right)$	nonsmth nonlinear
			$\|f(\mathbf{x}_k) - f^*\|, d\_(g(\mathbf{x}_k))$	$\rho_0\zeta^k$	$\mathcal{O}\left(\frac{1}{\rho_k}\right)$	$\tilde{\mathcal{O}}\left(\boldsymbol{\varepsilon}^{-4}\right)$	nonsmth nonlinear

S: smooth; NS: nonsmooth; L: linear; NL: nonlinear; †Strongly convex; ‡ Composite function: $f(\mathbf{x}) = p(\mathbf{x}) + \gamma(\mathbf{x})$ where $p(\cdot)$ is smooth and $\gamma(\cdot)$ is nonsmooth, proximal; ⋆ Additional boundedness condition required; ⋄ I-AL with regularization terms

(**Sm-AL**) and strongly convex settings (**Sm-AL**(S)) and standard ALM (**N-AL**), where $\tilde{\mathcal{O}}$ suppresses logarithmic terms.

(b) *Smoothing techniques.* While subgradient methods have proven effective in addressing nonsmooth convex objectives [36], smoothing techniques [6] represent an efficient avenue for a subclass of nonsmooth problems. Moreau [27] introduced the (Moreau)-smoothing f_η of a convex function f, with parameter η, defined as

$$f_\eta(\mathbf{x}) \triangleq \inf_{\mathbf{u}} \left\{ f(\mathbf{u}) + \tfrac{1}{\eta}\|\mathbf{u}-\mathbf{x}\|^2 \right\}.$$

Nesterov [33] employed a fixed smoothing parameter in developing a smoothing framework for nonsmooth convex optimization problems with a rate of $\mathcal{O}(\boldsymbol{\varepsilon}^{-1})$, an improvement over $\mathcal{O}(\boldsymbol{\varepsilon}^{-2})$ attainable by subgradient methods. In related work, Aybat and Iyengar [3] designed a smoothed penalty method for obtain $\boldsymbol{\varepsilon}$-optimal solutions for l_1-minimization problems with linear equality constraints in $\tilde{\mathcal{O}}\left(\boldsymbol{\varepsilon}^{-3/2}\right)$ steps. Subsequently, Beck and Teboulle [7] defined an (α, β)-smoothing for a nonsmooth convex f satisfying the following two conditions (i) $f_\eta(\mathbf{x}) \leq f(\mathbf{x}) \leq f_\eta(\mathbf{x}) + \eta\beta$ for all $\mathbf{x}$ and (ii) f_η is (α/η)-smooth. For instance, $f(\mathbf{x}) \triangleq \max\{0, \mathbf{x}\}$ has a smoothing f_η, defined as $f_\eta(\mathbf{x}) \triangleq \eta \log(1+\exp(\frac{\mathbf{x}}{\eta})) - \eta \log 2$. Analogous approaches have been employed for addressing deterministic [12] and stochastic [20] convex optimization problems.

1.2. Applications. We present three applications where nonsmooth convex constraints emerge.

(a) *Regression.* Lasso regression [40] is a model widely used in variable selection in statistical learning. Assuming that the dataset consists of $\{y_i, X_i\}_{i=1}^N$, where (y_i, X_i) denotes the outcome and feature vector for ith instance. Then an *elastic-net* model [46] can be articulated as follows where $C_1 > 0$.

$$\min_{\beta} \left\{ \|y - X\beta\|_2^2 \mid (1-\alpha)\|\beta\|_1 + \alpha\|\beta\|_2 \leq C_1 \right\}. \tag{1}$$

This reduces to standard *Lasso* [40] when $\alpha = 0$ and is generalizable to *fused Lasso* [41] by adding an additional nonsmooth constraint $\sum_{j=2}^p |\beta_j - \beta_{j-1}| \leq C_2$, where $C_2 > 0$.

(b) *Classification.* In statistical learning, the Neyman-Pearson (NP) classification [42] is designed to minimize the type II error while maintaining type I error below a user-specified level α. Consider a labeled training dataset $\{a_i\}_{i=1}^N$ where the positive and negative set are represented by $\{a_i^{(1)}\}_{i=1}^{N_{(1)}}$ and $\{a_i^{(-1)}\}_{i=1}^{N_{(-1)}}$, respectively. The empirical NP classification problem is given by [45] as follows

$$\min_{\mathbf{x}} \left\{ \frac{\sum_{i=i}^{N_{(-1)}} \boldsymbol{\ell}\left(1, \mathbf{x}^\top a_i^{(-1)}\right)}{N_{(-1)}} \,\middle|\, \frac{\sum_{i=i}^{N_{(1)}} \boldsymbol{\ell}\left(-1, \mathbf{x}^\top a_i^{(1)}\right)}{N_{(1)}} - \alpha \leq 0 \right\},$$

where $\boldsymbol{\ell}(\bullet)$ denotes the loss function. Choices of the loss function include nonsmooth variants such as mean absolute error (MAE) and hinge loss.

(c) *Multiple Kernel learning.* Multiple kernel learning (MKL) employs a predefined set of kernels to learn an optimal linear or nonlinear combination of these kernels, defined as follows [22].

$$\begin{aligned} \min_{w,b,(\theta,\xi)\geq 0} \quad & \tfrac{1}{2}\sum_{m=1}^M \frac{\|w_m\|_2^2}{\theta_m} + C\|\xi\|_1 \\ \text{subject to} \quad & y_i\left(\sum_{m=1}^M w_m' \psi_m(\mathbf{x}_i) + b\right) \geq 1 - \xi_i, \quad i = 1, \cdots, m \end{aligned}$$

$$\|\theta\|_p^p \leq 1,$$

where $\psi_i(\bullet), i = 1, \ldots, m$ are predefined kernels, θ is a vector of coefficients for each kernel, w is a weight vector for the primal model for learning with multiple kernels.

1.3. Contributions. We present a smoothed AL framework (**Sm-AL**) where the nonsmooth (but smoothable) objective/constraints are smoothed with a diminishing smoothing parameter η_k. Consequently, the AL subproblem (with penalty parameter ρ_k) is proven to be $\mathcal{O}(\rho_k/\eta_k)$-smooth, allowing for (accelerated) computation of an ϵ_k-exact solution in finite time. By a careful selection of the sequences $\{\epsilon_k, \eta_k, \rho_k\}$, we derive rate and complexity guarantees. Our contributions are formalized next.

(i) In Section 2, we derive an ex-ante bound on the optimal multiplier set of the η-smoothed problem. This result, which is of independent interest, allows for claiming that a saddle-point of the η-smoothed problem is an $\mathcal{O}(\eta)$-saddle point of the original problem, allowing for deriving fixed smoothing schemes. (ii) In Section 3, we establish a dual suboptimality rate of $\mathcal{O}(k^{-1})$ and primal infeasibility rate of $\mathcal{O}(k^{-1/2})$ (constant penalty) while geometric rates of $\mathcal{O}(1/\rho_k)$ on primal infeasibility and suboptimality are derived under geometrically increasing penalty parameters. In Section 4, by employing an accelerated gradient framework for resolving the η_k-smoothed AL subproblem, the overall complexities of (**Sm-AL**) in terms of inner projection steps for obtaining an $\boldsymbol{\varepsilon}$-optimal solution are proven to be $\mathcal{O}(\boldsymbol{\varepsilon}^{-(3+\delta)})$ (constant penalty) and $\tilde{\mathcal{O}}(\boldsymbol{\varepsilon}^{-3/2})$ (geometrically increasing penalty). Analogous bounds in strongly convex settings are given by $\tilde{\mathcal{O}}(\boldsymbol{\varepsilon}^{-(2+\delta)})$ and $\tilde{\mathcal{O}}(\boldsymbol{\varepsilon}^{-1})$ for constant and geometrically increasing penalty parameters, respectively. Similar complexity guarantees are available with a fixed smoothing parameter, akin to those developed in [7, 33] for convex programs with nonsmooth objectives.

(iii) We also develop practical termination criteria in Section 2, which when overlaid with our proposed scheme lead to significantly improved empirical complexity in our numerical experiments with little impact on accuracy.

(iv) Preliminary numerical results are provided in Section 5 before concluding in Section 6.

Organization The remainder of the paper is organized as follows. In Section 2, we introduce the smoothed augmented Lagrangian framework, providing the requisite background and the assumptions. Sections 3 and 4 provide the rate and complexity analysis while Section 5 presents a description of our numerical experiments. The paper concludes in Section 6.

Notation. Let $\|\cdot\|$ denote the Euclidean norm. Given a closed convex set $X \subseteq \mathbb{R}^n$ and $y \in \mathbb{R}^n$, $d_{\mathcal{X}}(y) \triangleq \min_{s\in\mathcal{X}} \|y - s\|$, $d^2_{\mathcal{X}}(y) \triangleq (d_{\mathcal{X}}(y))^2$, and $\Pi_{\mathcal{X}}(y) \triangleq \operatorname{argmin}_{s\in\mathcal{X}} \|y - s\|$; hence, $d_{\mathcal{X}}(y) = \|y - \Pi_{\mathcal{X}}(y)\|$. Moreover, $d^2_{\mathcal{K}}(\cdot)$ is differentiable and its gradient $\nabla d^2_{\mathcal{X}}(y) = 2(y - \Pi_{\mathcal{X}}(y))$. $d_-(u)$ denotes the distance of u to the nonpositive orthant $\mathbb{R}^n_-$, where $d_-(u)$ is defined as $d_-(u) \triangleq \|u - \Pi_{\mathbb{R}^n_-}[u]\|_2$. Finally, $\tilde{\mathcal{O}}(f(n))$ is $\mathcal{O}(f(n))$ up to a $\log(n)$ factor. Finally, $\mathbf{1}$ denotes the column of ones in $\mathbb{R}^n$.

2 A Smoothed Augmented Lagrangian Framework

In this section, we first provide some background and then analyze the smoothed problem, ending with a relation between a saddle-point of the η-smoothed problem and an η-approximate saddle-point of the original problem.

2.1 Background and Assumptions

Corresponding to problem (**NSCopt**), we may define the Lagrangian function $\mathcal{L}_0$ as follows.

$$\mathcal{L}_0(\mathbf{x}, \lambda) \triangleq \begin{cases} f(\mathbf{x}) + \lambda^\top g(\mathbf{x}), & \lambda \geq 0 \\ -\infty. & \text{otherwise} \end{cases}$$

This allows for denoting the set of minimizers of $\mathcal{L}_0(\bullet, \lambda)$ over the set $\mathcal{X}$ by $\mathcal{X}^*(\lambda)$, the dual function by $\mathcal{D}_0(\lambda)$, and the dual solution set by Λ^*, each of which is defined next.

$$\mathcal{X}^*(\lambda) \triangleq \arg\min_{\mathbf{x}\in\mathcal{X}} \mathcal{L}_0(\mathbf{x}, \lambda),\ \mathcal{D}_0(\lambda) \triangleq \inf_{\mathbf{x}\in\mathcal{X}} \mathcal{L}_0(\mathbf{x}, \lambda),\ \text{and}\ \Lambda^* \triangleq \arg\max_{\lambda\geq 0} \mathcal{D}_0(\lambda).$$

By adding a slack variable $\mathbf{v} \in \mathbb{R}^{\mathbf{m}}$, we may recast (**NSCopt**) as follows.

$$\begin{aligned} \min_{\mathbf{x}\in\mathcal{X}, \mathbf{v}\geq \mathbf{0}} \quad & f(\mathbf{x}) \\ \text{subject to} \quad & g(\mathbf{x}) + \mathbf{v} = \mathbf{0}, \qquad (\lambda) \end{aligned}$$

where $\lambda \in \mathbb{R}^m$ denotes the Lagrange multiplier associated with the constraint $g(\mathbf{x}) + \mathbf{v} = \mathbf{0}$. Then the augmented Lagrangian function, denoted by $\mathcal{L}_\rho$, where ρ denotes the penalty parameter, is defined as follows (cf. [38]).

$$\mathcal{L}_\rho(\mathbf{x}, \lambda) \triangleq \min_{\mathbf{v}\geq 0} \left[f(\mathbf{x}) + \lambda^\top (g(\mathbf{x}) + \mathbf{v}) + \tfrac{\rho}{2} \| g(\mathbf{x}) + \mathbf{v}\|^2 \right]. \tag{2}$$

It has been shown that $(\bar{\mathbf{x}}, \bar{\lambda})$ is a saddle-point of the augmented Lagrangian $\mathcal{L}_\rho$ for any $\rho \geq 0$ if and only if $(\bar{\mathbf{x}}, \bar{\lambda})$ is a saddle-point of $\mathcal{L}_0$. Further, if $\bar{\lambda}$ is an optimal dual solution, then $\bar{\mathbf{x}}$ is an optimal solution of (NSCopt) if and only if $\bar{\mathbf{x}}$ minimizes $\mathcal{L}(\bullet, \bar{\lambda})$ over $\mathcal{X}$ [38, Th. 3.5].

If $d_-(u) \triangleq \inf_{v\in\mathbb{R}^n_-} \|u - v\|$ and $\Pi_+[u]$ denotes the Euclidean projection of u onto $\mathbb{R}^m_+$, then the AL function $\mathcal{L}_\rho$ and its gradient can be expressed as follows [38, Sec. 2].

Lemma 1 *Consider the function $\mathcal{L}_\rho$ for $\rho > 0$, $\mathbf{x} \in \mathcal{X}$ and $\lambda \geq 0$. Then*

$$\begin{aligned} \mathcal{L}_\rho(\boldsymbol{x}, \lambda) &= \left(f(\boldsymbol{x}) + \tfrac{\rho}{2} \left(d_- \left(\tfrac{\lambda}{\rho} + g(\boldsymbol{x}) \right) \right)^2 - \tfrac{1}{2\rho} \|\lambda\|^2 \right), \\ \nabla_\lambda \mathcal{L}_\rho(\boldsymbol{x}, \lambda) &= \left(-\tfrac{\lambda}{\rho} + \Pi_+ \left(\tfrac{\lambda}{\rho} + g(\boldsymbol{x}) \right) \right), \\ \text{and } \nabla_{\boldsymbol{x}} \mathcal{L}_\rho(\boldsymbol{x}, \lambda) &= \nabla_{\boldsymbol{x}} f(\boldsymbol{x}) + \rho J_g(\boldsymbol{x}) \left(\tfrac{\lambda}{\rho} + g(\boldsymbol{x}) - \Pi_- \left(\tfrac{\lambda}{\rho} + g(\boldsymbol{x}) \right) \right), \end{aligned}$$

where $J_g(\mathbf{x})$ is Jacobian matrix of g. □

Similarly, the augmented dual function $\mathcal{D}_\rho$, defined as

$$\mathcal{D}_\rho(\lambda) \triangleq \inf_{\mathbf{x}\in\mathcal{X}} \mathcal{L}_\rho(\mathbf{x}, \lambda), \tag{3}$$

can be shown to be differentiable [38, Th. 3.2].

Lemma 2 *Consider the function $\mathcal{D}_\rho$ defined as (3). Then $\mathcal{D}_\rho$ is a C^1 and concave function over $\mathbb{R}^m$ and is the Moreau envelope of $\mathcal{D}_0$, defined as*

$$\mathcal{D}_\rho(\lambda) = \max_{u\in\mathbb{R}^m} \left[\mathcal{D}_0(u) - \tfrac{1}{2\rho} \|u - \lambda\|^2 \right] \text{ and } \nabla_\lambda \mathcal{D}_\rho(\lambda) \triangleq \tfrac{1}{\rho} \left(q_\rho(\lambda) - \lambda \right),$$

where $q_\rho(\lambda) \triangleq \arg\max_u \left[\mathcal{D}_0(u) - \frac{1}{2\rho} \|u - \lambda\|^2 \right]$. □

Since $\mathcal{D}_\rho$ is the Moreau envelope of $\mathcal{D}_0$, $\mathcal{D}_\rho$ has the same set of maximizers as $\mathcal{D}_0$ for any $\rho \geq 0$ [38, Th. 3.2]. Our interest lies in nonsmooth, albeit smoothable, convex functions, defined next [7].

Definition 1 A closed, proper, and convex function $h : \mathbb{R}^n \to \mathbb{R}$ is (α, β) smoothable if for any $\eta > 0$, there exists a convex differentiable function h_η such that

$$\begin{aligned} \left\| \nabla_{\mathbf{x}} h_\eta(\mathbf{x}_1) - \nabla_{\mathbf{x}} h_\eta(\mathbf{x}_2) \right\| &\leq \tfrac{\alpha}{\eta} \|\mathbf{x}_1 - \mathbf{x}_2\|, \quad \forall \mathbf{x}_1, \mathbf{x}_2 \in \mathbb{R}^n, \\ h_\eta(\mathbf{x}) \leq h(\mathbf{x}) &\leq h_\eta(\mathbf{x}) + \eta\beta, \qquad \forall \mathbf{x} \in \mathbb{R}^n. \end{aligned}$$

□

In fact, one may be faced by compositional convex constraints in which the layers may be nonsmooth. In such instances, under suitable conditions, smoothability of the layers implies smoothability of the compositional function but we postpone such avenues for future work. We leverage smoothability assumptions in [7] to state our basic assumptions on the objective and constraint functions. In addition, we impose both compactness requirements on $\mathcal{X}$ as well as a Slater regularity condition. Before stating the required assumptions, we need to define the ϵ-KKT conditions of (NSCopt), which is inspired by KKT conditions.

Definition 2 (ϵ-optimal solution) Let f^* be the optimal value of (NSCOPT). Given $\epsilon \geq 0$, a point $\tilde{\mathbf{x}} \in \mathcal{X}$ is called an ϵ-optimal and ϵ-feasible solution to (NSCOPT) if

$$f(\tilde{\mathbf{x}}) - f^* \leq \epsilon \text{ and } d_-(g(\tilde{\mathbf{x}})) \leq \epsilon, \quad \text{respectively.} \tag{4}$$

□

Then the partial KKT conditions corresponding to relaxing the constraint $g(\mathbf{x}) \leq 0$ are defined as follows, where $\mathcal{L}(\bullet, \bullet)$ denotes the Lagrangian function and $\mathcal{N}_{\mathcal{X}}(x)$ denotes the normal cone of $\mathcal{X}$ at x.

$$0 \in \nabla_{\mathbf{x}} \mathcal{L}(\mathbf{x}, \lambda) + \mathcal{N}_{\mathcal{X}}(\mathbf{x}) \tag{5}$$

$$0 \leq \lambda \perp g(\mathbf{x}) \leq 0. \tag{6}$$

Recall that given an optimization problem, defined as

$$\begin{aligned} &\min_x \ f(\mathbf{x}) \\ &\text{subject to } \ g(\mathbf{x}) \leq 0, \end{aligned} \tag{C-Opt}$$

where f, g_i are smooth functions mapping from $\mathbb{R}^n$ to $\mathbb{R}$ for $i = 1, \cdots, m$. Then under a suitable regularity condition, if x^* is a local minimizer of (C-Opt), then there exists $\lambda \in \mathbb{R}^m_+$ such that

$$\nabla f(\mathbf{x}) + \sum_{i=1}^{m} \lambda_i \nabla g_i(\mathbf{x}) = 0 \tag{7}$$

$$\lambda_i g_i(\mathbf{x}) = 0, \quad i = 1, \cdots, m \tag{8}$$

$$g(\mathbf{x}) \leq 0. \tag{9}$$

In fact, (8)–(9), together with $\lambda \geq 0$, can be compactly stated as

$$\lambda \geq 0, \quad \lambda_i g_i(\mathbf{x}) = 0, \forall i, \quad g(\mathbf{x}) \leq 0.$$

By leveraging the "perp" notation, we have that $\lambda \perp g(x)$ or $\lambda_i g_i(x) = 0$ for all i. Therefore, we may compactly represent the KKT conditions as

$$\nabla f(\mathbf{x}) + \sum_{i=1}^{m} \lambda_i \nabla g_i(\mathbf{x}) = 0 \tag{10}$$

$$0 \leq \lambda \perp g(\mathbf{x}) \leq 0. \tag{11}$$

Note that such a notation is common in complementarity theory (see Cottle, Pang, and Stone [11] or Facchinei and Pang [14]). This allows us to define a (partial) ϵ-KKT point.

Definition 3 (Partial ϵ-KKT condition) Consider the problem (NSCOPT). Then $(\mathbf{x}_\epsilon, \lambda_\epsilon)$ is a partial ϵ-KKT point if $\mathbf{x}_\epsilon \in \mathcal{X}$,

$$\mathcal{L}(\mathbf{x}_\epsilon, \lambda_\epsilon) \leq \mathcal{L}(\mathbf{x}^*, \lambda_\epsilon) + \epsilon, \tag{12}$$

$$0 \leq \lambda_\epsilon, \quad g(\mathbf{x}_\epsilon) \leq \epsilon \mathbf{1}, \text{ and } \lambda_\epsilon^\top g(\mathbf{x}_\epsilon) \geq -\epsilon, \tag{13}$$

where $(\mathbf{x}^*, \lambda^*)$ denotes a KKT point of (NSCOPT) satisfying (5)–(6). □

This allows us to build a simple relation whereby an ϵ-KKT point satisfies ϵ-optimality and ϵ-feasibility.

Lemma 3 *Consider a tuple $(\mathbf{x}_\epsilon, \lambda_\epsilon)$ satisfying the ϵ- KKT conditions given by (12)–(13). Then $(\mathbf{x}_\epsilon, \lambda_\epsilon)$ satisfies 2ϵ-suboptimality and $m\epsilon$-infeasibility, collectively captured by (4).*

Proof We observe that ϵ-primal suboptimality in (4) holds by the following sequence of relations.

$$\begin{aligned} f(\mathbf{x}_\epsilon) - \epsilon &\overset{(13)}{\leq} f(\mathbf{x}_\epsilon) + \lambda_\epsilon^\top g(\mathbf{x}_\epsilon) = \mathcal{L}(\mathbf{x}_\epsilon, \lambda_\epsilon) \\ &\overset{(12)}{\leq} \mathcal{L}(\mathbf{x}^*, \lambda_\epsilon) + \epsilon = f(\mathbf{x}^*) + \underbrace{\lambda_\epsilon^\top g(\mathbf{x}^*)}_{\leq 0} + \epsilon \\ &\leq f(\mathbf{x}^*) + \epsilon \\ \implies f(\mathbf{x}_\epsilon) &\leq f(\mathbf{x}^*) + 2\epsilon. \end{aligned}$$

To show ϵ-feasibility of $\mathbf{x}_\epsilon$ as prescribed in (4), we observe that

$$d_-(g(\mathbf{x}_\epsilon)) \leq \sum_{i=1}^{m} \max\{g_i(\mathbf{x}_\epsilon), 0\} \leq m\epsilon,$$

which completes the proof. □

We now present our ground assumption on the problem of interest and is assumed to hold throughout the paper, unless explicitly mentioned otherwise.

Assumption 1. (a) The function f and the constraint functions $g_1, g_2, \cdots, g_m$ are convex and (α, β)-smoothable real-valued functions.
(b) There exists a point $(\mathbf{x}^*, \lambda^*)$ satisfying the KKT conditions.
(c) The set $\mathcal{X} \subset \mathbb{R}^n$ is a convex and compact set.
(d) (Slater) There exists a vector $\bar{\mathbf{x}} \in \mathcal{X}$ such that $g_i(\bar{\mathbf{x}}) < 0$ for $i = 1, 2, \ldots, m$. □

Condition (d) allows for bounding the set of optimal dual variables (cf. [23]). We now consider the smoothed counterpart of (NSCopt), defined as

$$\begin{aligned} \min_{\mathbf{x} \in \mathcal{X}} \quad & f_\eta(\mathbf{x}) \\ \text{subject to} \quad & g_\eta(\mathbf{x}) \leq 0. \end{aligned} \tag{NSCopt$_\eta$}$$

We note that the solution and multiplier set of (NSCopt$_\eta$) are denoted by X_η^* and Λ_η^*, respectively. Naturally, associated with this problem is the Lagrangian function $\mathcal{L}_{\eta,0}$ of the smoothed problem (referred to as the smoothed Lagrangian) as well as the corresponding dual function $\mathcal{D}_{\eta,0}$; these objects and their augmented counterparts are defined and analyzed in the next subsection.

2.2 Analysis of Smoothed Lagrangians

We now analyze the smoothed Lagrangian framework where f and g are approximated by smoothings f_η and g_η, where the latter is a vector function with components $g_{1,\eta}, \cdots, g_{m,\eta}$. The resulting smoothed Lagrangian function $\mathcal{L}_{\eta,0}$ and the smoothed dual function $\mathcal{D}_{\eta,0}(\lambda)$ are defined as

$$\mathcal{L}_{\eta,0}(\mathbf{x}, \lambda) \triangleq \begin{Bmatrix} f_\eta(\mathbf{x}) + \lambda^\top g_\eta(\mathbf{x}), & \lambda \geq 0 \\ -\infty, & \text{otherwise} \end{Bmatrix} \text{ and } \mathcal{D}_{\eta,0}(\lambda) \triangleq \inf_{\mathbf{x} \in \mathcal{X}} \mathcal{L}_{\eta,0}(\mathbf{x}, \lambda).$$

Then the smoothed augmented Lagrangian function $\mathcal{L}_{\eta,\rho}$ is defined as

$$\begin{aligned} \mathcal{L}_{\eta,\rho}(\mathbf{x}, \lambda) &\triangleq \min_{\mathbf{v} \geq 0} \left[f_\eta(\mathbf{x}) + \lambda^\top (g_\eta(\mathbf{x}) + \mathbf{v}) + \tfrac{\rho}{2} \|g_\eta(\mathbf{x}) + \mathbf{v}\|^2 \right] \\ &= f_\eta(\mathbf{x}) + \tfrac{\rho}{2} \left(d_- \left(\tfrac{\lambda}{\rho} + g_\eta(\mathbf{x}) \right) \right)^2 - \tfrac{1}{2\rho} \|\lambda\|^2. \end{aligned}$$

We may now define $\mathcal{D}_{\eta,\rho}$ and $q_{\eta,\rho}$ as $\mathcal{D}_{\eta,\rho}(\lambda) = \max_u [\, \mathcal{D}_{\eta,0}(u) - \frac{1}{2\rho} \|u - \lambda\|^2 \,]$ and $\nabla_\lambda \mathcal{D}_{\eta,\rho}(\lambda) = \frac{1}{\rho} \left(q_{\eta,\rho}(\lambda) - \lambda \right)$, where $q_{\eta,\rho}(\lambda) \triangleq \operatorname{argmax}_u [\, \mathcal{D}_{\eta,0}(u) - \frac{1}{2\rho} \|u - \lambda\|^2 \,]$. We now relate $\mathcal{D}_\rho$ to $\mathcal{D}_{\eta,\rho}$ and q_ρ to $q_{\eta,\rho}$ in the next lemma.

Lemma 4 *For any $\lambda \in \mathbb{R}_+^m$, the following hold:*
(i) $\left| \mathcal{L}_0(\mathbf{x}, \lambda) - \mathcal{L}_{\eta,0}(\mathbf{x}, \lambda) \right| \leq \eta(\|\lambda\| m + 1)\beta$;
(ii) $|\mathcal{D}_{\eta,0}(\lambda) - \mathcal{D}_0(\lambda)| \leq \eta(\|\lambda\| m + 1)\beta$;
(iii) $|\mathcal{D}_{\eta,\rho}(\lambda) - \mathcal{D}_\rho(\lambda)| \leq \eta(\|\lambda\| m + 1)\beta$. □

Under a Slater regularity condition, the set of optimal multipliers is bounded (cf. [23]). Similar bounds are derived for the η-smoothed problem.

Proposition 1. (a) Let $\bar{\mathbf{x}}$ be as given in Assumption 1(d). Then for any $\eta > 0$, $g_\eta(\bar{\mathbf{x}}) < 0$.
(b) The set of optimal multipliers Λ^* for (NSCopt) is bounded as per

$$\Lambda^* \subseteq \left\{ \lambda \geq 0 \;\middle|\; \sum_{i=1}^{m} \lambda_i \leq b_\lambda \right\} \text{where } \frac{f(\bar{\mathbf{x}}) - f^*}{\min_j\{-g_j(\bar{\mathbf{x}})\}} \leq b_\lambda.$$

(c) For any $\eta > 0$, the set of optimal multipliers Λ_η^* for (NSCopt$_\eta$) is bounded as per with $\tilde{C}^* \triangleq (mb_\lambda + 1)\beta$

$$\Lambda_\eta^* \subseteq B_{\lambda,\eta} = \left\{ \lambda \geq 0 \;\middle|\; \sum_{i=1}^{m} \lambda_i \leq b_{\lambda,\eta} \right\} \text{ where } \frac{f(\bar{\mathbf{x}}) - f^* + \eta(mb_\lambda+1)\beta}{\min_j\{-g_j(\bar{\mathbf{x}})\}} \leq b_{\lambda,\eta}.$$

Proof (a) By Assumption 1(d), there exists a vector $\bar{\mathbf{x}} \in \mathcal{X}$ such that $g(\bar{\mathbf{x}}) < 0$, implying that $g_\eta(\bar{\mathbf{x}}) < 0$ by the property of smoothability (Def. 1).
(b) By the Slater regularity condition, we directly conclude from [23] that

$$\Lambda^* \subseteq \left\{ \lambda \geq 0 \;\middle|\; \sum_{i=1}^{m} \lambda_i \leq \frac{f(\bar{\mathbf{x}}) - \mathcal{D}_0^*}{\min_j\{-g_j(\bar{\mathbf{x}})\}} \right\}, \text{ where } \mathcal{D}_0^* = f^*.$$

(c) Similarly, Λ_η^*, the dual optimal solution set, is bounded as follows.

$$\Lambda_\eta^* \subseteq \left\{ \lambda \geq 0 \mid \sum_{i=1}^{m} \lambda_i \leq \frac{f_\eta(\bar{\mathbf{x}}) - \mathcal{D}_{0,\eta}^*}{\min_j\{-g_{j,\eta}(\bar{\mathbf{x}})\}} \right\} \subseteq \left\{ \lambda \geq 0 \mid \sum_{i=1}^{m} \lambda_i \leq \frac{f(\bar{\mathbf{x}}) - \mathcal{D}_{0,\eta}^*}{\min_j\{-g_{j,\eta}(\bar{\mathbf{x}})\}} \right\}.$$

Recall that $-g_{j,\eta}(\bar{\mathbf{x}}) \geq -g_j(\bar{\mathbf{x}})$ for $j = 1, \cdots, m$. Furthermore, $\min_j\{-g_{j,\eta}(\bar{\mathbf{x}})\} \geq \min_j\{-g_j(\bar{\mathbf{x}})\}$. It follows from (b) that

$$-\mathcal{D}_{0,\eta}(\lambda_\eta^*) \overset{\text{(Optimality of } \lambda_\eta^*)}{\leq} -\mathcal{D}_{0,\eta}(\lambda^*) \overset{\text{(Lemma 4(ii))}}{\leq} -\mathcal{D}_0(\lambda^*) + \eta(mb_\lambda + 1)\beta.$$

Consequently, if $\mathcal{D}_{0,\eta}^* \triangleq \mathcal{D}_{0,\eta}(\lambda_\eta^*)$, $\mathcal{D}_0^* \triangleq \mathcal{D}_0(\lambda^*)$, then

$$\begin{aligned}
\Lambda_\eta^* &\subseteq \left\{ \lambda \geq 0 \mid \sum_{i=1}^{m} \lambda_i \leq \frac{f(\bar{\mathbf{x}}) - \mathcal{D}_{0,\eta}^*}{\min_j\{-g_{j,\eta}(\bar{\mathbf{x}})\}} \right\} \\
&\subseteq \left\{ \lambda \geq 0 \mid \sum_{i=1}^{m} \lambda_i \leq \frac{f(\bar{\mathbf{x}}) - \mathcal{D}_{0,\eta}^*}{\min_j\{-g_j(\bar{\mathbf{x}})\}} \right\} \\
&\subseteq \left\{ \lambda \geq 0 \mid \sum_{i=1}^{m} \lambda_i \leq \frac{f(\bar{\mathbf{x}}) - \mathcal{D}_0^* + \eta(mb_\lambda+1)\beta}{\min_j\{-g_j(\bar{\mathbf{x}})\}} \right\} \\
&\subseteq B_{\lambda,\eta} \triangleq \left\{ \lambda \geq 0 \mid \sum_{i=1}^{m} \lambda_i \leq b_{\lambda,\eta} \right\}.
\end{aligned}$$

□

Both Lemma 4 and Proposition 1 play crucial roles in the convergence analysis presented in Section 3. We now relate a saddle-point $(\mathbf{x}_\eta^*, \lambda_\eta^*)$ of (NSCopt$_\eta$) to an η-saddle-point $(\mathbf{x}^*, \lambda^*)$ of (NSCopt), where the bound on the multipliers for (NScopt) and (NSCopt$_\eta$) are denoted by b_λ and $b_{\lambda,\eta}$, respectively. Next, we relate a saddle-point of (NSCopt$_\eta$) to an η-saddle-point

of (NSCopt), where an η-saddle point satisfies the saddle-point requirements with an $\mathcal{O}(\eta)$ error.

Theorem 1. Let $(\mathbf{x}_\eta^*, \lambda_\eta^*)$ represent a saddle point of (NSCopt$_\eta$).
(a) Suppose $\mathbf{x}_\eta^* \in \mathcal{X}$ is a feasible solution of (NSCopt$_\eta$). Then $\mathbf{x}_\eta^*$ is an $\sqrt{m}\eta\beta$-feasible of (NSCopt), i.e. $d_-(g(\mathbf{x}_\eta^*)) \leq \sqrt{m}\eta\beta$.
(b) $(\mathbf{x}_\eta^*, \lambda_\eta^*)$ is a $2\eta\beta(1+m\max\left\{b_{\lambda,\eta}, \|\lambda\|\right\})$-saddle-point of (NSCopt), i.e. for all $(\mathbf{x}, \lambda) \in \mathcal{X} \times \mathbb{R}_+^m$,

$$\begin{aligned}\mathcal{L}_0(\mathbf{x}_\eta^*, \lambda) - \eta\beta(1+m\max\left\{b_{\lambda,\eta}, \|\lambda\|\right\}) &\leq \mathcal{L}_0(\mathbf{x}_\eta^*, \lambda_\eta^*) \\ &\leq \mathcal{L}_0(\mathbf{x}, \lambda_\eta^*) + \eta\beta(1+mb_{\lambda,\eta}).\end{aligned}$$

Proof (a) Suppose $\mathbf{x}_\eta^* \in \mathcal{X}$ is a feasible solution of (NSCopt$_\eta$). Then $g_\eta(\mathbf{x}_\eta^*) \leq 0$. Furthermore, $g(\mathbf{x}_\eta^*) \leq g_\eta(\mathbf{x}_\eta^*) + \eta\beta\mathbf{1} \leq \eta\beta\mathbf{1}$, implying that $d_-(g(\mathbf{x}_\eta^*)) \leq \eta\beta\|\mathbf{1}\|$.
(b) The dual optimal set Λ_η^* is nonempty and bounded as per Lemma 1. Let $(\mathbf{x}_\eta^*, \lambda_\eta^*)$ be the saddle point of $L_{\eta,0}(\cdot,\cdot)$. We now proceed to show that $(\mathbf{x}_\eta^*, \lambda_\eta^*)$ is an approximate saddle-point of $\mathcal{L}_0$.

$$\begin{aligned}\mathcal{L}_0(\mathbf{x}_\eta^*, \lambda_\eta^*) &= f(\mathbf{x}_\eta^*) + (\lambda_\eta^*)^\top g(\mathbf{x}_\eta^*) \leq f_\eta(\mathbf{x}_\eta^*) + \eta\beta + (\lambda_\eta^*)^\top g_\eta(\mathbf{x}_\eta^*) + \eta b_{\lambda,\eta}\beta\|\mathbf{1}\| \\ &= \mathcal{L}_{0,\eta}(\mathbf{x}_\eta^*, \lambda_\eta^*) + \eta\beta(1+b_{\lambda,\eta}m) \leq \mathcal{L}_{0,\eta}(\mathbf{x}, \lambda_\eta^*) + \eta\beta(1+b_{\lambda,\eta}m) \text{ for all } \mathbf{x} \in \mathcal{X} \\ &\overset{-(\lambda_\eta^*)^\top g(\mathbf{x})\leq 0}{=} \mathcal{L}_0(\mathbf{x}, \lambda_\eta^*) + f_\eta(\mathbf{x}) - f(\mathbf{x}) + (\lambda_\eta^*)^\top(g_\eta(\mathbf{x}) - g(\mathbf{x})) + \eta\beta(1+b_{\lambda,\eta}m) \\ &\leq \mathcal{L}_0(\mathbf{x}, \lambda_\eta^*) + \eta\beta(1+b_{\lambda,\eta}m) \text{ for all } \mathbf{x} \in \mathcal{X}.\end{aligned}$$

The final result follows through the following sequence of inequalities as provided next

$$\begin{aligned}\mathcal{L}_0(\mathbf{x}_\eta^*, \lambda_\eta^*) &= f(\mathbf{x}_\eta^*) + (\lambda_\eta^*)^\top g(\mathbf{x}_\eta^*) \geq f_\eta(\mathbf{x}_\eta^*) + (\lambda_\eta^*)^\top\left(g_\eta(\mathbf{x}_\eta^*)\right) \\ &= \mathcal{L}_{0,\eta}(\mathbf{x}_\eta^*, \lambda_\eta^*) \geq \mathcal{L}_{0,\eta}(\mathbf{x}_\eta^*, \lambda) \quad \text{for all } \lambda \in \mathbb{R}_+^m \\ &= \mathcal{L}_0(\mathbf{x}_\eta^*, \lambda) + f_\eta(\mathbf{x}_\eta^*) - f(\mathbf{x}_\eta^*) + \lambda^\top\left(g_\eta(\mathbf{x}_\eta^*) - g(\mathbf{x}_\eta^*)\right) \\ &\geq \mathcal{L}_0(\mathbf{x}_\eta^*, \lambda) - \eta\beta(1+m\|\lambda\|) \\ &\geq \mathcal{L}_0(\mathbf{x}_\eta^*, \lambda) - \eta\beta\big(1+m\max\{b_{\lambda,\eta}, \|\lambda\|\}\big) \quad \forall\lambda \in \mathbb{R}_+^m.\end{aligned}$$

□

The following Lemma 5 shows the relation between $q_{\eta,\rho}(\bullet)$ and $q_\rho(\bullet)$.

Lemma 5 *For any $\lambda \in \mathbb{R}_+^m$, the following hold:*
(i) $\|q_{\eta,\rho}(\lambda) - q_\rho(\lambda)\| \leq \sqrt{4\rho\eta(\|\lambda\|m + C_m)\beta}$;
(ii) $\|\nabla_\lambda\mathcal{D}_{\eta,\rho}(\lambda) - \nabla_\lambda\mathcal{D}_\rho(\lambda)\| = \frac{1}{\rho}\|q_{\eta,\rho}(\lambda) - q_\rho(\lambda)\| \leq \sqrt{\frac{4\eta(\|\lambda\|m+C_m)\beta}{\rho}}$. □

We now formally state the smoothed AL scheme. The traditional ALM is reliant on solving the subproblem exactly or ϵ_k-inexactly at epoch k. However, in regimes with nonsmooth constraints, the AL subproblem is nonsmooth, precluding the usage of accelerated gradient methods, leading to far poorer performance. Our proposed scheme solves a sequence of η_k-smoothed problems solved within an error tolerance of $\epsilon_k\eta_k^b$ where $b \geq 0$. A formal statement of the scheme is provided next.

Smoothed augmented Lagrangian scheme (Sm-AL).
Given $\mathbf{x}_0$, λ_0, $K > 0$, and sequences $\{\rho_k, \epsilon_k, \eta_k\}$.
For $k = 0, \cdots, K-1$, do

[0] Given η_k, ρ_k, ϵ_k, and λ_k;
[1] $\mathbf{x}_{k+1}$ satisfies $\left\{ \mathcal{L}_{\eta_k,\rho_k}(\mathbf{x}_{k+1}, \lambda_k) - \mathcal{D}_{\eta_k,\rho_k}(\lambda_k) \leq \epsilon_k \eta_k^b \right\}$;
[2] $\lambda_{k+1} = \lambda_k + \rho_k \nabla_\lambda \mathcal{L}_{\eta_k,\rho_k}(\mathbf{x}_{k+1}, \lambda_k)$; $k := k+1$;

Output $\{(\bar{\mathbf{x}}_K, \bar{\lambda}_K)\}$ or $\{(\mathbf{x}_K, \lambda_K)\}$.

Observe that step [1] requires that $\mathbf{x}_{k+1}$ is an $\epsilon_k \eta_k^b$-minimizer of the AL subproblem, given by

$$\min_{x \in X} \ \mathcal{L}_{\eta_k,\rho_k}(\mathbf{x}, \lambda_k),$$

where $\mathcal{D}_{\eta_k,\rho_k}(\lambda_k) = \min_{x \in X} \mathcal{L}_{\eta_k,\rho_k}(\mathbf{x}, \lambda_k)$. Since we have rate guarantees for the accelerated scheme applied to the subproblem, we can determine the minimum number of gradient steps that ensure that $\epsilon_k \eta_k^b$-suboptimality holds. The Lagrange multiplier update can be expressed as follows (cf. [2]).

Lemma 6 *Consider the smoothed augmented Lagrangian scheme (***Sm-AL***). Then for any $k > 0$, step [2] is equivalent to the following equation.*

$$\lambda_{k+1} = \Pi_+ \left[\lambda_k + \rho_k g_{\eta_k}(\mathbf{x}_{k+1}) \right]. \tag{14}$$

The next assumption holds for parameter sequences employed in (**Sm-AL**). Unless mentioned otherwise, Assumptions 1 and 2 hold throughout.

Assumption 2. The positive sequences $\{\epsilon_k, \eta_k, \rho_k\}_{k=1}^K$ satisfy
(i) $\sum_{k=1}^\infty \sqrt{\rho_k \epsilon_k \eta_k^b} < \infty$; (ii) $\sum_{k=1}^\infty \sqrt{\rho_k \eta_k} < \infty$, where $b \geq 0$.

While our rate guarantees for the schemes responsible for resolving the subproblem as well as the outer (dual) problem allow for defining precise lower bounds on the number of steps required, this computational requirement is reliant on a worst-case analysis. In addition, we may attempt to check if the sub-optimality requirement is met at some intermediate step. However, it is not obvious how to check the sub-optimality in the current setting since the optimal value corresponding to either the subproblem or the outer level problem are unavailable. Instead, we appeal to a residual function and consider such an approach next. We emphasize that such a potential early termination of either the subproblem solver or the outer scheme may have computational benefits.

2.3 Termination Criteria

Our inexact augmented Lagrangian framework relies on utilizing inexact solutions to the Lagrangian subproblem, obtained by taking finite but increasing number of gradient-based steps and then leveraging the rate guarantees for accelerated gradient methods. However, we may well meet the required accuracy prior to taking the prescribed number of gradient steps by checking a suitable condition. Such a condition is by no means immediate since a naive assessment of accuracy requires knowing the optimal value to the subproblem; instead, we present a new analysis by leveraging a residual function and present such an analysis next for both the inner and outer loops.

(I). Termination criterion for Inner loop. The inner loop at iteration k terminates when x_{k+1} satisfies the following $\epsilon_k \eta_k^b$-optimality requirement, where ϵ_k is a positive accuracy thresholdat iteration k, η_k is the smoothing parameter at iteration k, and b is a nonnegative scalar that is defined subsequently in the complexity analysis.

$$\mathcal{L}_{\eta_k,\rho_k}(\mathbf{x}_{k+1}, \lambda_k) - \mathcal{D}_{\eta_k,\rho_k}(\lambda_k) \leq \epsilon_k \eta_k^b. \tag{15}$$

In effect, if we view the minimization of the augmented Lagrangian function by the following convex problem, defined as

$$\min_{\mathbf{x}\in\mathcal{X}} h(\mathbf{x}) \triangleq \mathcal{L}_{\eta_k,\rho_k}(\mathbf{x}, \lambda_k), \tag{Opt}$$

where h is a convex and smooth function on $\mathcal{X}$, a closed and convex set. We proceed to show that (15) is equivalent to x_{k+1} approximately satisfying the variational inequality problem.

$$\nabla_x \mathcal{L}_{\eta_k,\rho_k}(\mathbf{x}_{k+1}, \lambda_k)^\top (y - \mathbf{x}_{k+1}) \geq -\epsilon_k \eta_k^b \qquad \forall y \in X. \tag{16}$$

In fact, we now develop a verifiable condition whose satisfaction implies (16).

Lemma 7 *Consider the problem (Opt). Suppose* $\|\mathbf{y}\|^2 \leq C$ *and* $\|\nabla h(\mathbf{y})\|^2 \leq D$ *for any* $\mathbf{y} \in X$ *and* γ *is any positive scalar. Consider the following statements.*

(a) $\mathbf{x}_\epsilon^*$ *is an* ϵ*-optimal solution of (Opt).*
(b) $\nabla h(\mathbf{x}_\epsilon^*)^\top (\mathbf{y} - \mathbf{x}_\epsilon^*) \geq -\epsilon, \quad \forall \mathbf{y} \in \mathcal{X}.$
(c) Suppose there exists $\mathbf{u} \in \mathcal{X}$ *and* $\mathbf{x}_\epsilon^* \in \mathcal{X}$ *such that* $F_{\mathcal{X}}^{\text{nat},\tilde{\epsilon}}(\mathbf{x}_\epsilon^*, \mathbf{u}) = 0$*, where* $F_{\mathcal{X}}^{\text{nat},\tilde{\epsilon}}(\bullet, \bullet)$ *represents the perturbed natural map with a chosen parameter* γ*, defined as*

$$F_{\mathcal{X}}^{\text{nat},\tilde{\epsilon}}(\mathbf{x}, \mathbf{u}) \triangleq \tilde{\epsilon}\left(\mathbf{u} - \Pi_X\left[\mathbf{x} - \gamma \nabla h(\mathbf{x})\right]\right) - \mathbf{x} + \Pi_{\mathcal{X}}\left[\mathbf{x} - \gamma \nabla h(\mathbf{x})\right].$$

Then the following hold.

(i) $(a) \iff (b)$*;*
(ii) $(c) \implies (b)$*, where* $\tilde{\epsilon} = \frac{\gamma\epsilon}{7C+\gamma(C+D)}$ *and* $\epsilon < \frac{7C+\gamma(C+D)}{\gamma}$. □

Observe that the perturbed natural map is rooted in the natural map, a residual function for variational inequality problems [14]. When specialized to the setting of the the smooth convex optimization problem

$$\min_{\mathbf{x}\in\mathcal{X}} f(\mathbf{x}), \tag{COpt}$$

we have that

$$\left[\mathbf{x}^* \text{ solves (COpt)}\right] \iff \left[F_{\mathcal{X}}^{\text{nat}}(\mathbf{x}^*) \triangleq \mathbf{x}^* - \Pi_X\left[\mathbf{x}^* - \gamma \nabla f(\mathbf{x}^*)\right] = 0\right].$$

The lemma above develops a suitably defined $\tilde{\epsilon}$-perturbed counterpart of $F_{\mathcal{X}}^{\text{nat}}$, denoted by $F_{\mathcal{X}}^{\text{nat},\tilde{\epsilon}}$. We observe that $F_{\mathcal{X}}^{\text{nat},\tilde{\epsilon}}(\mathbf{x}, \mathbf{x})$ reduces to

$$F_X^{\text{nat},\tilde{\epsilon}}(\mathbf{x}, \mathbf{x}) \triangleq (1 - \tilde{\epsilon})\left(\Pi_{\mathcal{X}}\left[\mathbf{x} - \gamma \nabla h(\mathbf{x})\right] - \mathbf{x}\right). \tag{17}$$

In other words, for any $\tilde{\epsilon} < 1$,

$$F_{\mathcal{X}}^{\text{nat},\tilde{\epsilon}}(\mathbf{x}, \mathbf{x}) = 0 \iff F_{\mathcal{X}}^{\text{nat}}(\mathbf{x}) = 0, \tag{18}$$

where $F_{\mathcal{X}}^{\text{nat}}(\mathbf{x}) \triangleq -\mathbf{x} + \Pi_{\mathcal{X}}\left[\mathbf{x} - \gamma \nabla h(\mathbf{x})\right]$. Based on the aforementioned result, in the kth iteration, this termination criterion reduces to

$$\textbf{(T1)} \qquad \left\| \tilde{\epsilon}_k \mathbf{u} + (1 - \tilde{\epsilon}_k)\, \Pi_{\mathcal{X}}\left[\mathbf{x}_{k+1} - \gamma \nabla_{\mathbf{x}} \mathcal{L}_{\eta_k,\rho_k}(\mathbf{x}_{k+1}, \lambda_k)\right] - \mathbf{x}_{k+1} \right\| = 0, \tag{19}$$

where $\tilde{\epsilon}_k = \frac{\gamma \epsilon_k \eta_k^b}{7C+\gamma(C+D)}$ and $\mathbf{u} \in \mathcal{X}$.

(II) Termination criterion for outer loop. Here we consider two settings.

(a) Constant penalty parameter. In setting (a), the outer scheme terminates when

$$\left| f(\bar{\mathbf{x}}_K) - f^* \right| \leq \frac{C_1}{\sqrt{K}} + \eta_K \beta \text{ and } d_-\left(g(\bar{\mathbf{x}}_K)\right) \leq \frac{C_2}{\sqrt{K}} + m\eta_K \beta,$$

where $C_1 \triangleq B_5, C_2 \triangleq B_4, B_3, B_4, B_5, B_6$ are defined in Table 3. Since we have access to $g(\bullet)$, it is easy to check $d_-\left(g(\mathbf{x}_K)\right) \leq \sqrt{\epsilon}$. However, evaluating $f(\bar{x}_K) - f^*$ is not directly possible, since f^* is unavailable. Since f is nonsmooth, we apply Lemma 7 to the optimality gap of the smoothed problem $|f_{\eta_K}(\bar{\mathbf{x}}_K) - f^*_{\eta_K}|$ since it is related to the true optimality gap, i.e. by leveraging the property of smoothability of f,

$$\begin{aligned} f(\bar{\mathbf{x}}_K) - f(\mathbf{x}^*) &\leq f_{\eta_K}(\bar{\mathbf{x}}_K) - f_{\eta_K}(\mathbf{x}^*) + \eta_K B \leq \left| f_{\eta_K}(\bar{\mathbf{x}}_K) - f_{\eta_K}(\mathbf{x}^*) \right| + \eta_K B \\ f(\mathbf{x}^*) - f(\bar{\mathbf{x}}_K) &\leq f_{\eta_K}(\mathbf{x}^*) - f_{\eta_K}(\bar{\mathbf{x}}_K) + \eta_K B \leq \left| f_{\eta_K}(\bar{\mathbf{x}}_K) - f_{\eta_K}(\mathbf{x}^*) \right| + \eta_K B \\ &\implies \left| f(\bar{\mathbf{x}}_K) - f(\mathbf{x}^*) \right| \leq \left| f_{\eta_K}(\bar{\mathbf{x}}_K) - f_{\eta_K}(\mathbf{x}^*) \right| + \eta_K B. \end{aligned}$$

Consequently, it suffices to get a bound on each term on the right. To get a bound on $\left| f_{\eta_K}(\bar{\mathbf{x}}_K) - f_{\eta_K}(\mathbf{x}^*) \right|$ given $\hat{\mathbf{x}} \in \mathcal{X}$, we leverage the following residual function that

$$G_{\mathcal{X}}^{\text{nat},\tilde{\epsilon}_K}(\mathbf{x}_{K+1}, \hat{\mathbf{x}}) \triangleq \tilde{\epsilon}_K \hat{\mathbf{x}} + (1 - \tilde{\epsilon}_K) \Pi_X \left[\mathbf{x}_{K+1} - \gamma \nabla \mathcal{L}_{\eta_K}(\mathbf{x}_{K+1}) \right] - \mathbf{x}_{K+1},$$

where $\tilde{\epsilon}_K \triangleq \frac{\gamma C_1}{(7C+\gamma(C+D))\sqrt{K}}$, and C, D are as defined in Lemma 7. (We can set the values for η_K such that the overall optimality gap ($|f - f^*|$) remains controlled below a tighter error tolerance ϵ^2 to ensure the consistency with our complexity analysis.) Therefore, we may employ the following termination criterion (**T2**) at the Kth iterate.

$$\textbf{(T2)} \quad \left\| G_{\mathcal{X}}^{\text{nat},\tilde{\epsilon}_K}(\mathbf{x}_{K+1}, \hat{\mathbf{x}}) \right\| = 0 \text{ and } d_-\left(g(\mathbf{x}_K)\right) \leq \tilde{\epsilon}_K. \tag{20}$$

(b) Increasing penalty parameter. In setting (b), the outer scheme terminates when

$$\left| f(\mathbf{x}_K) - f^* \right| \leq \frac{C_1}{\rho_K} + \eta_K \beta \text{ and } d_-\left(g(\mathbf{x}_K)\right) \leq \frac{C_2}{\rho_K} + m\eta_K \beta,$$

where $C_1 \triangleq B_7$ and $C_2 \triangleq B_8$ as defined in Table 3. While it is easy to check $d_-\left(g(\mathbf{x}_K)\right) \leq \epsilon$, since f^* is unavailable and f is nonsmooth, we apply Lemma 7 to the optimality gap of the smoothed problem $|f_{\eta_K}(\mathbf{x}_K) - f^*_{\eta_K}|$ since it is related to the true optimality gap, i.e. by leveraging the property of smoothability of f, similar to the previous analysis, $|f(\mathbf{x}_K) - f(\mathbf{x}^*)| \leq \left| f_{\eta_K}(\mathbf{x}_K) - f_{\eta_K}(\mathbf{x}^*) \right| + \eta_K B$. Consequently, it suffices to get a bound on both terms on the right. To get a bound on $\left| f_{\eta_K}(\mathbf{x}_K) - f_{\eta_K}(\mathbf{x}^*) \right|$ and given $\hat{x} \in X$, we leverage the following residual function that

$$G_{\mathcal{X}}^{\text{nat},\tilde{\epsilon}_K}(\mathbf{x}_{K+1}, \hat{\mathbf{x}}) \triangleq \tilde{\epsilon}_K \hat{\mathbf{x}} + (1 - \tilde{\epsilon}_K) \Pi_X \left[\mathbf{x}_{K+1} - \gamma \nabla \mathcal{L}_{\eta_K}(\mathbf{x}_{K+1}) \right] - \mathbf{x}_{K+1},$$

where $\tilde{\epsilon}_K \triangleq \frac{\gamma C_1}{(7C+\gamma(C+D))\rho_K}$, and C, D are as defined in Lemma 7. Akin to earlier, we may set the value of η_K such that the overall optimality gap ($|f - f^*|$) remains controlled below ϵ. Therefore, we may employ the following termination criterion (**T2**) at the Kth iterate.

$$\textbf{(T2)} \quad \left\| G_{\mathcal{X}}^{\text{nat},\tilde{\epsilon}_K}(\mathbf{x}_{K+1}, \hat{x}) \right\| = 0 \text{ and } d_-\left(g(\mathbf{x}_K)\right) \leq \tilde{\epsilon}_K. \tag{21}$$

The modified algorithm statement should read as follows.

Smoothed augmented Lagrangian scheme (Sm-AL).
Given $\mathbf{x}_0, \lambda_0, k = 0, K > 0$, and sequences $\{\rho_k, \epsilon_k, \eta_k\}$.
While $k \leq K$ and (**T2**) fails, do;

[0] Given η_k, ρ_k, ϵ_k, and λ_k;
[1] Run subproblem solver for ℓ steps while $\ell \leq M_k$ steps and (**T1**) fails; Output $\mathbf{x}_{k+1}$;
[2] $\lambda_{k+1} = \lambda_k + \rho_k \nabla_\lambda \mathcal{L}_{\eta_k,\rho_k}(\mathbf{x}_{k+1}, \lambda_k)$; $k := k + 1$;

Output $\{(\bar{\mathbf{x}}_K, \bar{\lambda}_K)\}$ or $\{(\mathbf{x}_K, \lambda_K)\}$.

Note that the subproblem solver is essentially an accelerated gradient scheme introduced in Section 4, the minimum number of steps as prescribed by the rate guarantees is denoted by M_k and derived in Section 4.

3 Rate Analysis

In this section, we analyze the rate of convergence for (**Sm-AL**). In 3.1, we provide some preliminaries and then derive rate statements for constant and increasing penalties in Subsections 3.2 and 3.3, respectively.

3.1 Preliminary results

We begin by recalling the following bound, an extension of the result proved in [38, Lemma 4.3].

Lemma 8 *Let $\{\mathbf{x}_k, \lambda_k\}$ be generated by (**Sm-AL**). For any $k \geq 0$, suppose $\mathbf{x}_{k+1}$ satisfies $\mathcal{L}_{\eta_k,\rho_k}(\mathbf{x}_{k+1}, \lambda_k) - \mathcal{D}_{\rho_k,\eta_k}(\lambda_k) \leq \epsilon_k \eta_k^b$ where $b \geq 0$. Then for $k \geq 0$,*

$$\left\| \nabla_\lambda \mathcal{L}_{\eta_k,\rho_k}(x_{k+1}, \lambda_k) - \nabla_\lambda \mathcal{D}_{\eta_k,\rho_k}(\lambda_k) \right\|^2 \leq \frac{2\epsilon_k \eta_k^b}{\rho_k}. \tag{22}$$

By choosing appropriate sequences $\{\epsilon_k, \eta_k, \rho_k\}$, $\{(2\epsilon_k \eta_k^b)/\rho_k\}$ is diminishing (see Lemma 8). We now derive a uniform bound on the sequence $\{\lambda_k\}$.

Lemma 9 *(Bound on λ_k) Consider $\{\lambda_k\}$ generated by (**Sm-AL**).*
(a) $\{\lambda_k\}$ is a convergent sequence. (b) For any K, we have

$$\|\lambda_K - \lambda^*\| \leq \sum_{k=0}^{\infty} \left(\sqrt{2\rho_k \epsilon_k \eta_k^b} + 2\sqrt{\eta_k \rho_k (\|\lambda^*\| m + C_m)\beta} \right) + \|\lambda_0 - \lambda^*\| \triangleq B_\lambda.$$

3.2 Rate analysis under constant ρ_k

Next, we derive rate statements for the dual sub-optimality and primal infeasibility when $\rho_k = \rho$ for all k. Our first result relies on the observation that the augmented dual function $\mathcal{D}_\rho$ has the same set of optimal solutions (and supremum) as the original dual function $\mathcal{D}_0$ (see [38, Th. 3.2]).

Proposition 2. (Dual sub-optimality) Consider the sequence $\{\lambda_k\}$ generated by (**Sm-AL**), where $\rho_k = \rho$ for every $k \geq 0$. If B_λ, B_2 are constants, then the following holds for $\bar{\lambda}_K \triangleq \frac{\sum_{i=1}^K \lambda_i}{K}$ and for any $K > 0$,

$$f^* - \mathcal{D}_\rho(\bar{\lambda}_K) \leq \frac{1}{2\rho K}\|\lambda_0 - \lambda^*\|^2 + \frac{B_\lambda}{K}\sum_{k=0}^{K-1}\frac{\sqrt{2\epsilon_k \eta_k^b}}{\sqrt{\rho}} + \frac{B_2}{K}\sum_{k=0}^{K-1}\eta_k = \mathcal{O}\left(\frac{1}{K}\right).$$

Proof Recall that $\mathcal{D}_{\eta_k,\rho}$ is the Moreau envelope of $\mathcal{D}_{\eta_k,0}$. Consequently, $\nabla_\lambda \mathcal{D}_{\eta_k,\rho}$ is $\frac{1}{\rho}$-Lipschitz. We then have

$$\begin{aligned}
-\mathcal{D}_{\eta_k,\rho}(\lambda_{k+1}) &\leq -\mathcal{D}_{\eta_k,\rho}(\lambda_k) - \nabla_\lambda \mathcal{D}_{\eta_k,\rho}(\lambda_k)^\top(\lambda_{k+1} - \lambda_k) + \tfrac{1}{2\rho}\|\lambda_{k+1} - \lambda_k\|^2 \\
&= -\mathcal{D}_{\eta_k,\rho}(\lambda_k) - \nabla_\lambda \mathcal{D}_{\eta_k,\rho}(\lambda_k)^\top(\lambda_{k+1} - \lambda^*) - \nabla_\lambda \mathcal{D}_{\eta_k,\rho}(\lambda_k)^\top(\lambda^* - \lambda_k) \\
&\quad + \tfrac{1}{2\rho}\|\lambda_{k+1} - \lambda_k\|^2 \\
&\leq -\mathcal{D}_{\eta_k,\rho}(\lambda_k) - \nabla_\lambda \mathcal{D}_{\eta_k,\rho}(\lambda_k)^\top(\lambda_{k+1} - \lambda^*) + (\mathcal{D}_{\eta_k,\rho}(\lambda_k) - \mathcal{D}_{\eta_k,\rho}(\lambda^*)) \\
&\quad + \tfrac{1}{2\rho}\|\lambda_{k+1} - \lambda_k\|^2 \\
&= -\mathcal{D}_{\eta_k,\rho}(\lambda^*) - \nabla_\lambda \mathcal{D}_{\eta_k,\rho}(\lambda_k)^\top(\lambda_{k+1} - \lambda^*) + \tfrac{1}{2\rho}\|\lambda_{k+1} - \lambda_k\|^2,
\end{aligned}$$

where $-\mathcal{D}_{\eta_k,\rho}(\lambda^*) \geq -\mathcal{D}_{\eta_k,\rho}(\lambda_k) - \nabla_\lambda \mathcal{D}_{\eta_k,\rho}(\lambda_k)^\top(\lambda^* - \lambda_k)$. By adding and subtracting $\nabla_\lambda \mathcal{L}_{\eta_k,\rho}(\mathbf{x}_{k+1}, \lambda_k)^\top(\lambda_{k+1} - \lambda^*)$, it follows that

$$\begin{aligned}
-\mathcal{D}_{\eta_k,\rho}(\lambda_{k+1}) &\leq -\mathcal{D}_{\eta_k,\rho}(\lambda^*) - \nabla_\lambda \mathcal{L}_{\eta_k,\rho}(\mathbf{x}_{k+1}, \lambda_k)^\top(\lambda_{k+1} - \lambda^*) + \tfrac{1}{2\rho}\|\lambda_{k+1} - \lambda_k\|^2 \\
&\quad - \left(\nabla_\lambda \mathcal{D}_{\eta_k,\rho}(\lambda_k) - \nabla_\lambda \mathcal{L}_{\eta_k,\rho}(\mathbf{x}_{k+1}, \lambda_k)\right)^\top(\lambda_{k+1} - \lambda^*) \\
&= -\mathcal{D}_{\eta_k,\rho}(\lambda^*) - \tfrac{1}{\rho}(\lambda_{k+1} - \lambda_k)^\top(\lambda_{k+1} - \lambda^*) + \tfrac{1}{2\rho}\|\lambda_{k+1} - \lambda_k\|^2 \\
&\quad - \left(\nabla_\lambda \mathcal{D}_{\eta_k,\rho}(\lambda_k) - \nabla_\lambda \mathcal{L}_{\eta_k,\rho}(\mathbf{x}_{k+1}, \lambda_k)\right)^\top(\lambda_{k+1} - \lambda^*) \\
&\leq -\mathcal{D}_{\eta_k,\rho}(\lambda^*) - \tfrac{1}{\rho}(\lambda_{k+1} - \lambda_k)^\top(\lambda_{k+1} - \lambda^*) + \tfrac{1}{2\rho}\|\lambda_{k+1} - \lambda_k\|^2 \\
&\quad + \|\nabla_\lambda \mathcal{D}_{\eta_k,\rho}(\lambda_k) - \nabla_\lambda \mathcal{L}_{\eta_k,\rho}(\mathbf{x}_{k+1}, \lambda_k)\|\|\lambda_{k+1} - \lambda^*\| \\
&= -\mathcal{D}_{\eta_k,\rho}(\lambda^*) + \tfrac{1}{2\rho}(\|\lambda_k - \lambda^*\|^2 - \|\lambda_{k+1} - \lambda^*\|^2) \\
&\quad + \|\nabla_\lambda \mathcal{D}_{\eta_k,\rho}(\lambda_k) - \nabla_\lambda \mathcal{L}_{\eta_k,\rho}(\mathbf{x}_{k+1}, \lambda_k)\|\|\lambda_{k+1} - \lambda^*\| \\
&\leq -\mathcal{D}_\rho(\lambda^*) + \eta_k(\|\lambda^*\|m + 1)\beta + \tfrac{1}{2\rho}(\|\lambda_k - \lambda^*\|^2 - \|\lambda_{k+1} - \lambda^*\|^2) \\
&\quad + \|\nabla_\lambda \mathcal{D}_{\eta_k,\rho}(\lambda_k) - \nabla_\lambda \mathcal{L}_{\eta_k,\rho}(\mathbf{x}_{k+1}, \lambda_k)\|\|\lambda_{k+1} - \lambda^*\|,
\end{aligned}$$

where the last inequality follows from Lemma 4(iii). By invoking Lemma 9, and $\|\lambda_k\| + \|\lambda^*\| \leq \|\lambda_k - \lambda^*\| + \|\lambda^*\| + \|\lambda^*\| \leq B_\lambda + 2b_\lambda \triangleq \tilde{B}_\lambda$, we obtain

$$\begin{aligned}
-\mathcal{D}_\rho(\lambda_{k+1}) &\leq -\mathcal{D}_\rho(\lambda^*) + \eta_k(\|\lambda_{k+1}\|m + 1)\beta + \eta_k(\|\lambda^*\|m + 1)\beta \\
&\quad + \tfrac{1}{2\rho}(\|\lambda_k - \lambda^*\|^2 - \|\lambda_{k+1} - \lambda^*\|^2) \\
&\quad + \|\nabla_\lambda \mathcal{D}_{\eta_k,\rho}(\lambda_k) - \nabla_\lambda \mathcal{L}_{\eta_k,\rho}(\mathbf{x}_{k+1}, \lambda_k)\|\|\lambda_{k+1} - \lambda^*\| \\
&\leq -\mathcal{D}_\rho(\lambda^*) + \eta_k(\tilde{B}_\lambda m + 1)\beta + \tfrac{1}{2\rho}(\|\lambda_k - \lambda^*\|^2 - \|\lambda_{k+1} - \lambda^*\|^2) \\
&\quad + \|\nabla_\lambda D_{\eta_k,\rho}(\lambda_k) - \nabla_\lambda \mathcal{L}_{\eta_k,\rho}(\mathbf{x}_{k+1}, \lambda_k)\|\|\lambda_{k+1} - \lambda^*\|.
\end{aligned}$$

By summing from $k = 0, \cdots, K-1$ and dividing by K, we obtain

$$\begin{aligned}
&-\left(\frac{1}{K}\sum_{i=0}^{K-1}\mathcal{D}_\rho(\lambda_{i+1}) - \mathcal{D}_\rho(\lambda^*)\right) \\
&\leq \frac{1}{2\rho K}(\|\lambda_0 - \lambda^*\|^2 - \|\lambda_K - \lambda^*\|^2) + \frac{1}{K}\sum_{k=0}^{K-1}\eta_k(\tilde{B}_\lambda m + 1)\beta \\
&\quad + \frac{1}{K}\sum_{k=0}^{K-1}\left\|\nabla_\lambda D_{\eta_k,\rho}(\lambda_k) - \nabla_\lambda \mathcal{L}_{\eta_k,\rho}(\mathbf{x}_{k+1},\lambda_k)\right\|\|\lambda_{k+1} - \lambda^*\| \\
&\leq \frac{1}{2\rho K}\|\lambda_0 - \lambda^*\|^2 + \frac{B_\lambda}{K}\sum_{k=0}^{K-1}\frac{\sqrt{2\epsilon_k\eta_k^b}}{\sqrt{\rho}} + \frac{B_2}{K}\sum_{k=0}^{K-1}\eta_k,
\end{aligned} \tag{23}$$

where boundedness of λ_k follows from Lemma 4 and $\tilde{B}_\lambda$, B_λ, B_2 are constants. Consequently, by invoking the concavity of $\mathcal{D}_\rho$, we may bound the term on the left to obtain the required inequality, where $\bar{\lambda}_K = \frac{1}{K}\sum_{i=1}^{K}\lambda_i$.

$$\begin{aligned}
-\left(\mathcal{D}_\rho(\bar{\lambda}_K) - \mathcal{D}_\rho(\lambda^*)\right) &\leq \frac{1}{2\rho K}(\|\lambda_0 - \lambda^*\|^2 - \|\lambda_K - \lambda^*\|^2) + \frac{1}{K}\sum_{k=0}^{K-1}\eta_k(\tilde{B}_\lambda m + 1)\beta \\
&\quad + \frac{1}{K}\sum_{k=0}^{K-1}\left\|\nabla_\lambda D_{\eta_k,\rho}(\lambda_k) - \nabla_\lambda \mathcal{L}_{\eta_k,\rho}(\mathbf{x}_{k+1},\lambda_k)\right\|\|\lambda_{k+1} - \lambda^*\| \\
&\leq \frac{1}{2\rho K}\|\lambda_0 - \lambda^*\|^2 + \frac{B_\lambda}{K}\sum_{k=0}^{K-1}\frac{\sqrt{2\epsilon_k\eta_k^b}}{\sqrt{\rho}} + \frac{B_2}{K}\sum_{k=0}^{K-1}\eta_k.
\end{aligned} \tag{24}$$

The final result follows by noting that $\mathcal{D}_\rho$ is the Moreau envelope of $\mathcal{D}_0$ and strong duality holds, implying that $\mathcal{D}_\rho(\lambda^*) = \mathcal{D}_0(\lambda^*) = f(\mathbf{x}^*)$. □

Next, we derive a rate statement on the infeasibility.

Proposition 3. (Rate on primal infeasibility) Let $\{(\mathbf{x}_k, \lambda_k)\}$ be a sequence generated by (**SmAL**). Then the following holds for $\bar{\mathbf{x}}_K = \frac{1}{K}\sum_{i=1}^{K}\mathbf{x}_i$ for any $K > 0$, where $C, B_3, B_4 \geq 0$.

$$\begin{aligned}
d_-(g(\bar{\mathbf{x}}_K)) &\leq \frac{\sqrt{m}}{K}\sum_{i=0}^{K-1}\left(\sqrt{\frac{2\epsilon_i\eta_i^b}{\rho}} + \sqrt{m}\eta_i\beta + \sqrt{\frac{2\eta_i\beta\tilde{B}_\lambda}{\rho}}\right) \\
&\quad + \sqrt{\frac{2mC}{\rho K}} \leq \frac{B_3}{K} + \frac{B_4}{\sqrt{K}} = \mathcal{O}\left(\frac{1}{\sqrt{K}}\right).
\end{aligned}$$

Proof We have that $g_{\eta_k}(\mathbf{x}_{k+1})$ can be expressed as

$$g_{\eta_k}(\mathbf{x}_{k+1}) = \nabla_\lambda \mathcal{L}_{\eta_k,\rho}(\mathbf{x}_{k+1},\lambda_k) + \left(\Pi_-\left(\tfrac{\lambda_k}{\rho} + g_{\eta_k}(\mathbf{x}_{k+1})\right)\right).$$

Recall that $d_-(u+v) \leq d_-(u) + \|v\|$ for any $u, v \in \mathbb{R}^m$. Consequently,

$$\begin{aligned}
d_-(g_{\eta_k}(\mathbf{x}_{k+1})) &\leq \|\nabla_\lambda \mathcal{L}_{\eta_k,\rho}(\mathbf{x}_{k+1},\lambda_k)\| + \underbrace{d_-\left(\Pi_-\left(\tfrac{\lambda_k}{\rho} + g_{\eta_k}(\mathbf{x}_{k+1})\right)\right)}_{=0} \\
&= \|\nabla_\lambda \mathcal{L}_{\eta_k,\rho}(\mathbf{x}_{k+1},\lambda_k)\|.
\end{aligned} \tag{25}$$

By definition of $d_-(\bullet)$, convexity of $\max\{g_j(\bullet), 0\}$, and $\|u\|_2 \le \|u\|_1 \le \sqrt{m}\|u\|_2$,

$$
\begin{aligned}
d_-(g(\bar{\mathbf{x}}_K)) &= \inf_{u\in\mathbb{R}^m_-} \|g(\bar{\mathbf{x}}_K) - u\|_2 \le \inf_{u\in\mathbb{R}^m_-} \|g(\bar{\mathbf{x}}_K) - u\|_1 = \sum_{j=1}^m \inf_{u_j\le 0} \left|g_j(\bar{\mathbf{x}}_K) - u_j\right|_1 \\
&= \sum_{j=1}^m \max\{g_j(\bar{\mathbf{x}}_K), 0\} \le \tfrac{1}{K}\sum_{i=0}^{K-1}\sum_{j=1}^m \max\{g_j(\mathbf{x}_{i+1}), 0\} \\
&\le \tfrac{1}{K}\sum_{i=0}^{K-1}\sum_{j=1}^m \max\{g_{j,\eta_i}(\mathbf{x}_{i+1}) + \eta_i\beta, 0\} = \tfrac{1}{K}\sum_{i=0}^{K-1} \inf_{u\in\mathbb{R}^m_-} \|g_{\eta_i}(\mathbf{x}_{i+1}) + \eta_i\beta\mathbf{1} - u\|_1 \\
&\le \tfrac{1}{K}\sum_{i=0}^{K-1} \inf_{u\in\mathbb{R}^m_-} \sqrt{m}\|g_{\eta_i}(\mathbf{x}_{i+1}) + \eta_i\beta\mathbf{1} - u\|_2 = \tfrac{\sqrt{m}}{K}\sum_{k=1}^{K-1} d_-(g_{\eta_i}(\mathbf{x}_{i+1}) + \eta_i\beta\mathbf{1}) \\
&\le \tfrac{\sqrt{m}}{K}\sum_{i=0}^{K-1} \left(d_-(g_{\eta_i}(\mathbf{x}_{i+1})) + \eta_i\beta\|\mathbf{1}\|_2\right) \overset{(25)}{\le} \tfrac{\sqrt{m}}{K}\sum_{i=0}^{K-1} \left(\|\nabla_\lambda \mathcal{L}_{\eta_i,\rho}(\mathbf{x}_{i+1}, \lambda_i)\| + \sqrt{m}\eta_i\beta\right) \\
&\le \tfrac{\sqrt{m}}{K}\sum_{i=0}^{K-1} \left(\|\nabla_\lambda \mathcal{L}_{\eta_i,\rho}(\mathbf{x}_{i+1}, \lambda_i) - \nabla_\lambda \mathcal{D}_{\eta_i,\rho}(\lambda_i)\|\right. \\
&\left. + \|\nabla_\lambda \mathcal{D}_{\eta_i,\rho}(\lambda_i)\| + \sqrt{m}\eta_i\beta\right). \qquad (26)
\end{aligned}
$$

Recall that

$$
\|\nabla_\lambda \mathcal{D}_{\eta_k,\rho}(\lambda_1) - \nabla_\lambda \mathcal{D}_{\eta_k,\rho}(\lambda_2)\| \le \tfrac{1}{\rho}\left\|q_{\eta,\rho}(\lambda_1) - q_{\eta,\rho}(\lambda_2)\right\| + \tfrac{1}{\rho}\|\lambda_1 - \lambda_2\| \le \tfrac{2}{\rho}\|\lambda_1 - \lambda_2\|,
$$

allowing us to claim that $\mathcal{D}_{\eta_k,\rho}$ is a $(2/\rho)$-smooth concave function. Then by leveraging [32] for any $\lambda \ge 0$,

$$
\begin{aligned}
\|\nabla_\lambda \mathcal{D}_{\eta_k,\rho}(\lambda)\| &\le \sqrt{\tfrac{2}{\rho}\left(\mathcal{D}_{\eta_k,\rho}(\lambda^*_{\eta_k}) - \mathcal{D}_{\eta_k,\rho}(\lambda)\right)} \le \sqrt{\tfrac{2}{\rho}\left(\mathcal{D}_{\rho}(\lambda^*_{\eta_k}) - \mathcal{D}_{\rho}(\lambda) + 2\eta_k\beta\tilde{B}_\lambda\right)} \\
&\le \sqrt{\tfrac{2}{\rho}\left(\mathcal{D}_{\rho}(\lambda^*) - \mathcal{D}_{\rho}(\lambda) + 2\eta_k\beta\tilde{B}_\lambda\right)} \le \sqrt{\tfrac{2}{\rho}\left(\mathcal{D}_{\rho}(\lambda^*) - \mathcal{D}_{\rho}(\lambda)\right)} + 2\sqrt{\tfrac{\eta_k\beta\tilde{B}_\lambda}{\rho}},
\end{aligned}
$$

where λ^*_η is a maximizer of $\mathcal{D}_{\eta,\rho}$. By leveraging the concavity of the square-root function, the prior dual sub-optimality bounds, $\sqrt{u+v} \le \sqrt{u} + \sqrt{v}$ for $u, v \ge 0$, the subaddivity of concave functions, we have from (26),

$$
\begin{aligned}
d_-(g(\bar{\mathbf{x}}_K)) &\le \tfrac{\sqrt{m}}{K}\sum_{i=0}^{K-1}\left(\sqrt{\tfrac{2\epsilon_i\eta_i^b}{\rho}} + \sqrt{m}\eta_i\beta\right) + \tfrac{\sqrt{m}}{K}\sum_{i=0}^{K-1}\sqrt{\tfrac{2}{\rho}\left(\mathcal{D}_\rho(\lambda^*) - \mathcal{D}_\rho(\lambda_i)\right)} \\
&+ \tfrac{2\sqrt{m}}{K}\sum_{i=0}^{K-1}\sqrt{\tfrac{\eta_i\beta\tilde{B}_\lambda}{\rho}} \\
&\overset{\text{(Concavity of }\sqrt{\cdot}\text{)}}{\le} \tfrac{\sqrt{m}}{K}\sum_{i=0}^{K-1}\left(\sqrt{\tfrac{2\epsilon_i\eta_i^b}{\rho}} + \sqrt{m}\eta_i\beta\right) + \sqrt{\tfrac{2m}{\rho}\left(\mathcal{D}_\rho(\lambda^*) - \tfrac{1}{K}\sum_{i=0}^{K-1}\mathcal{D}_\rho(\lambda_i)\right)} \\
&+ \tfrac{\sqrt{m}}{K}\sum_{i=0}^{K-1}\sqrt{\tfrac{2\eta_i\beta\tilde{B}_\lambda}{\rho}}.
\end{aligned}
$$

Recall from (24), it follows that

$$\frac{1}{K}\sum_{i=0}^{K-1}\left(\mathcal{D}_\rho(\lambda^*)-\mathcal{D}_\rho(\lambda_i)\right)\leq \frac{1}{2\rho K}\|\lambda_0-\lambda^*\|^2+\frac{B_\lambda}{K}\sum_{k=0}^{K-1}\frac{\sqrt{2\epsilon_k\eta_k^b}}{\sqrt{\rho}}+\frac{B_2}{K}\sum_{k=0}^{K-1}\eta_k=\frac{C}{K},$$

which implies that

$$d_-(g(\bar{\mathbf{x}}_K))\leq \frac{\sqrt{m}}{K}\sum_{i=0}^{K-1}\left(\sqrt{\frac{2\epsilon_i\eta_i^b}{\rho}}+\sqrt{m}\eta_i\beta+\sqrt{\frac{2\eta_i\beta\tilde{B}_\lambda}{\rho}}\right)+\sqrt{\frac{2mC}{\rho K}}$$

where $C\triangleq \frac{\|\lambda_0-\lambda^*\|^2}{2\rho}+\left(B_\lambda\sum_{k=0}^{K-1}\frac{2\epsilon_k\eta_k^b}{\sqrt{\rho}}+B_2\sum_{k=0}^{K-1}\eta_k\right)$. □

We now derive a rate statement for the primal sub-optimality.

Theorem 2. (Rate on primal sub-opt) Consider the sequence $\{(\mathbf{x}_k,\lambda_k)\}$ generated by (**Sm-AL**). Then (27) holds for $\bar{\mathbf{x}}_K=\frac{1}{K}\sum_{i=1}^{K}\mathbf{x}_i$ and for any $K>0$, where $B_5, B_6\geq 0$.

$$-\left(\frac{B_5^2}{K}+\frac{B_5}{\sqrt{K}}+(1+mb_\lambda)\eta_K\beta\right)\leq f(\bar{\mathbf{x}}_K)-f^*\leq\frac{B_6}{K}. \tag{27}$$

Proof Recall that since $\mathbf{x}_k$ may not be feasible with respect to the constraints, we derive upper and lower bounds on the sub-optimality.

(i) *Lower bound.* A rate statement for the lower bound is first constructed. Since $\max_\lambda \mathcal{D}_\rho(\lambda)=\min_{\mathbf{x}\in\mathcal{X}}\mathcal{L}_\rho(\mathbf{x},\lambda^*)=f^*$, the following sequence of inequalities hold where $\bar{\mathbf{x}}_K=\frac{1}{K}\sum_{k=0}^{K-1}\mathbf{x}_k$, $f^*_{\eta_K}=\min_{\mathbf{x}\in\mathcal{X}}\mathcal{L}_{\eta_K,\rho}\left(\mathbf{x},\lambda^*_{\eta_K}\right)$, and $(\mathbf{x}^*_{\eta_K},\lambda^*_{\eta_K})$ is the saddle point of $\mathcal{L}_{\eta_K,0}(\mathbf{x},\lambda)$.

$$\begin{aligned}
f^*_{\eta_K}&=\mathcal{L}_{\eta_K,\rho}(\mathbf{x}^*_{\eta_K},\lambda^*_{\eta_K})\leq\mathcal{L}_{\eta_K,\rho}(\bar{\mathbf{x}}_K,\lambda^*_{\eta_K})\\
&=f_{\eta_K}(\bar{\mathbf{x}}_K)+\frac{\rho}{2}\left(d_-\left(\frac{\lambda^*_{\eta_K}}{\rho}+g_{\eta_K}(\bar{\mathbf{x}}_K)\right)\right)^2-\frac{1}{2\rho}\|\lambda^*_{\eta_K}\|^2\\
&\leq f_{\eta_K}(\bar{\mathbf{x}}_K)+\frac{\rho}{2}\left(d_-\left(g_{\eta_K}(\bar{\mathbf{x}}_K)\right)+\left\|\frac{\lambda^*_{\eta_K}}{\rho}\right\|\right)^2-\frac{1}{2\rho}\|\lambda^*_{\eta_K}\|^2\\
&=f_{\eta_K}(\bar{\mathbf{x}}_K)+\frac{\rho}{2}\left(d_-\left(g_{\eta_K}(\bar{\mathbf{x}}_K)\right)\right)^2+\left\|\lambda^*_{\eta_K}\right\|d_-\left(g_{\eta_K}(\bar{\mathbf{x}}_K)\right)\\
&\overset{\text{Lem. 1}}{\leq} f_{\eta_K}(\bar{\mathbf{x}}_K)+\frac{\rho}{2}\left(d_-\left(g_{\eta_K}(\bar{\mathbf{x}}_K)\right)\right)^2+b_{\lambda,\eta}d_-\left(g_{\eta_K}(\bar{\mathbf{x}}_K)\right).
\end{aligned}$$

By invoking Proposition 3, we obtain the following inequality.

$$f^*_{\eta_K}-f_{\eta_K}(\bar{\mathbf{x}}_K)\leq\frac{B_5^2}{K}+\frac{B_5}{\sqrt{K}}. \tag{28}$$

Let $\mathbf{x}^*\in\mathcal{X}^*$ and $\mathbf{x}^*_{\eta_K}$ is a minimizer of $L_{\eta_K,\rho}(\cdot,\lambda^*_{\eta_K})$. By Lemma 4, we have that

$$\begin{aligned}
f(\mathbf{x}^*)=\mathcal{L}(\mathbf{x}^*,\lambda^*)\leq\mathcal{L}(\mathbf{x}^*_{\eta_K},\lambda^*)&=f(\mathbf{x}^*_{\eta_K})+\sum_{i=1}^{m}\lambda_i^*g_i(x^*_{\eta_K})\\
&\leq f(\mathbf{x}^*_{\eta_K})+\sum_{i=1}^{m}\lambda_i^*(g_i(x^*_{\eta_K})-g_{i,\eta_K}(\mathbf{x}^*_{\eta_K}))\\
&\leq f(\mathbf{x}^*_{\eta_K})+mb_\lambda\eta_K\beta,
\end{aligned} \tag{29}$$

implying that $f(\mathbf{x}^*) \leq f(\mathbf{x}^*_{\eta_K}) + mb_\lambda \beta \eta_K$. By definition of the smoothing, $f(\mathbf{x}^*_{\eta_K}) - f_{\eta_K}(\mathbf{x}^*_{\eta_K}) \leq \beta\eta_K$ and $f_{\eta_K}(\bar{\mathbf{x}}_K) - f(\bar{\mathbf{x}}_K) \leq 0$.

$$f(\mathbf{x}^*) - f(\bar{\mathbf{x}}_K) = \underbrace{f(\mathbf{x}^*) - f(\mathbf{x}^*_{\eta_K})}_{\leq mb_\lambda\beta\eta_K} + \underbrace{f(\mathbf{x}^*_{\eta_K}) - f_{\eta_K}(\mathbf{x}^*_{\eta_K})}_{\leq \beta\eta_K} + \underbrace{f_{\eta_K}(\mathbf{x}^*_{\eta_K}) - f_{\eta_K}(\bar{\mathbf{x}}_K)}_{(28)}$$
$$+ \underbrace{f_{\eta_K}(\bar{\mathbf{x}}_K) - f(\bar{\mathbf{x}}_K)}_{\leq 0} \leq (1 + mb_\lambda)\eta_K\beta + \frac{B_5^2}{K} + \frac{B_5}{\sqrt{K}}.$$

(ii) *Upper bound.* Let $\mathbf{x}^*_{\eta_k,\lambda_k} \in \arg\min_{\mathbf{x}\in\mathcal{X}} \mathcal{L}_{\eta_k,\rho}(\mathbf{x}, \lambda_k)$ and $(\mathbf{x}^*_{\eta_k}, \lambda^*_{\eta_k})$ be the saddle point of $\mathcal{L}_{\eta_k,0}(\mathbf{x}, \lambda)$. Based on the definition of $\mathbf{x}^*_{\eta_k,\lambda_k}$ and $\mathbf{x}^*_{\eta_k}$, the following two inequalities hold.

$$\mathcal{L}_{\eta_k,\rho}(\mathbf{x}_{k+1}, \lambda_k) - \mathcal{L}_{\eta_k,\rho}(\mathbf{x}^*_{\eta_k,\lambda_k}, \lambda_k) \leq \epsilon_k \eta_k^b$$
$$\mathcal{L}_{\eta_k,\rho}(\mathbf{x}^*_{\eta_k,\lambda_k}, \lambda_k) \leq \mathcal{L}_{\eta_k,\rho}(\mathbf{x}^*_{\eta_k}, \lambda_k)$$

By adding the two inequalities, we obtain

$$\begin{aligned}
&\mathcal{L}_{\eta_k,\rho}(\mathbf{x}_{k+1}, \lambda_k) - \mathcal{L}_{\eta_k,\rho}(\mathbf{x}^*_{\eta_k}, \lambda_k) \\
&= \mathcal{L}_{\eta_k,\rho}(\mathbf{x}_{k+1}, \lambda_k) - \mathcal{L}_{\eta_k,\rho}(\mathbf{x}^*_{\eta_k,\rho_k}, \lambda_k) + \mathcal{L}_{\eta_k,\rho}(\mathbf{x}^*_{\eta_k,\rho_k}, \lambda_k) - \mathcal{L}_{\eta_k,\rho}(\mathbf{x}^*_{\eta_k}, \lambda_k) \\
&\leq \epsilon_k \eta_k^b. \qquad (30)
\end{aligned}$$

Consequently, by leveraging (30) and invoking the definition of $\mathcal{L}_{\eta_k,\rho}(\cdot, \lambda_k)$, we have that

$$f_{\eta_k}(\mathbf{x}_{k+1}) - f_{\eta_k}(\mathbf{x}^*_{\eta_k}) \leq \frac{\rho}{2}\left(d_-\left(\frac{\lambda_k}{\rho} + g_{\eta_k}(\mathbf{x}^*_{\eta_k})\right)\right)^2 - \frac{\rho}{2}\left(d_-\left(\frac{\lambda_k}{\rho} + g_{\eta_k}(\mathbf{x}_{k+1})\right)\right)^2 + \epsilon_k\eta_k^b.$$

We observe that

$$d_-(u) = \|\Pi_-(u) - u\| = \|\Pi_-(u) - (\Pi_-(u) + \Pi_+(u))\| = \|-\Pi_+(u)\| = \|\Pi_+(u)\|.$$

By choosing $u = g_{\eta_k}(\mathbf{x}_{k+1}) + \frac{\lambda_k}{\rho}$, it follows from Lemma 6 that

$$d_-\left(g_{\eta_k}(\mathbf{x}_{k+1}) + \tfrac{\lambda_k}{\rho}\right) = \left\|\Pi_+\left(g_{\eta_k}(\mathbf{x}_{k+1}) + \tfrac{\lambda_k}{\rho}\right)\right\| = \left\|\tfrac{\lambda_{k+1}}{\rho}\right\|.$$

Furthermore, we have that $g_{\eta_k}(\mathbf{x}^*_{\eta_k}) \leq 0$ since $\mathbf{x}^*_{\eta_k}$ is feasible with respect to η_k-smoothed objective, implying

$$d_-\left(\tfrac{\lambda_k}{\rho} + g_{\eta_k}(\mathbf{x}^*_{\eta_k})\right) \leq \underbrace{d_-\left(g_{\eta_k}(\mathbf{x}^*_{\eta_k})\right)}_{=0,\ \text{since}\ g_{\eta_k}(\mathbf{x}^*_{\eta_k})\leq 0} + \left\|\tfrac{\lambda_k}{\rho}\right\|$$

which implies

$$f_{\eta_k}(\mathbf{x}_{k+1}) - f_{\eta_k}(\mathbf{x}^*_{\eta_k}) \leq \frac{\rho}{2}\left(\left\|\tfrac{\lambda_k}{\rho}\right\|^2 - \left\|\tfrac{\lambda_{k+1}}{\rho}\right\|^2\right) + \epsilon_k\eta_k^b. \qquad (31)$$

We observe that that $g_{\eta_k}(\mathbf{x}^*) \leq g(\mathbf{x}^*) \leq 0$, implying that $\mathbf{x}^*$ is feasible for the η_k-smoothed problem and consequently,

$$f_{\eta_k}(\mathbf{x}^*_{\eta_k}) - f_{\eta_k}(\mathbf{x}^*) \leq 0. \qquad (32)$$

Summing from $k = 0$ to $K - 1$ and leveraging convexity of f_{η_k} and , we obtain that

$$f(\bar{\mathbf{x}}_K) - f^* \leq \frac{1}{K}\sum_{k=0}^{K-1}\left(f(\mathbf{x}_{k+1}) - f^*\right)$$

$$= \frac{1}{K}\sum_{k=0}^{K-1}\Bigg(\underbrace{f(\mathbf{x}_{k+1}) - f_{\eta_k}(\mathbf{x}_{k+1})}_{\leq \eta_k\beta} + \underbrace{f_{\eta_k}(\mathbf{x}_{k+1}) - f_{\eta_k}(\mathbf{x}^*_{\eta_k})}_{(31)}$$

$$+ \underbrace{f_{\eta_k}(\mathbf{x}^*_{\eta_k}) - f_{\eta_k}(\mathbf{x}^*)}_{\leq 0 \text{ from } (32)} + \underbrace{f_{\eta_k}(\mathbf{x}^*) - f^*}_{\leq 0 \text{ (smoothing)}}\Bigg)$$

$$\leq \frac{1}{K}\left(\frac{\rho}{2}\left(d_-\left(\frac{\lambda_0}{\rho}\right)^2 - d_-\left(\frac{\lambda_K}{\rho}\right)^2\right)\right) + \frac{1}{K}\sum_{k=0}^{K-1}\left(\epsilon_k\eta_k^b + \eta_k\beta\right)$$

$$\leq \frac{\rho}{2K}\|\tfrac{\lambda_0}{\rho}\|^2 + \frac{1}{K}\sum_{k=0}^{K-1}\left(\epsilon_k\eta_k^b + \eta_k\beta\right) \leq \frac{B_6}{K}$$

where $B_6 > 0$ is a constant. □

3.3 Rate analysis under increasing ρ_k

We now consider the setting where $\{\rho_k\}$ is an increasing sequence.

Lemma 10 *(Rate on primal infeasibility) Suppose* $\{(\mathbf{x}_k, \lambda_k)\}$ *is generated by (***Sm-AL***). Then for any* $k \geq 0$, $d_-\left(g(\mathbf{x}_{k+1})\right) \leq \left\|\frac{\lambda_{k+1}-\lambda_k}{\rho_k}\right\| + m\eta_k\beta$.

Proof By the update rule, we have that

$$\lambda_{k+1} := \lambda_k + \rho_k\nabla_\lambda\mathcal{L}_{\eta_k,\rho_k}(\mathbf{x}_{k+1},\lambda_k) = \lambda_k + \rho_k g_{\eta_k}(\mathbf{x}_{k+1}) - \rho_k\Pi_-\left(\frac{\lambda_k}{\rho_k} + g_{\eta_k}(\mathbf{x}_{k+1})\right).$$

It follows that $g_{\eta_k}(\mathbf{x}_{k+1}) = \frac{\lambda_{k+1}-\lambda_k}{\rho_k} + \Pi_-\left(\frac{\lambda_k}{\rho_k} + g_{\eta_k}(\mathbf{x}_{k+1})\right)$, implying

$$d_-\left(g_{\eta_k}(\mathbf{x}_{k+1})\right) \leq d_-\left(\Pi_-\left(\frac{\lambda_k}{\rho_k} + g_{\eta_k}(\mathbf{x}_{k+1})\right)\right) + \left\|\frac{\lambda_{k+1}-\lambda_k}{\rho_k}\right\| = \left\|\frac{\lambda_{k+1}-\lambda_k}{\rho_k}\right\|.$$

Akin to the proof in Proposition 3, we have

$$d_-\left(g(\mathbf{x}_{k+1})\right) \leq d_-\left(g_{\eta_k}(\mathbf{x}_{k+1})\right) + m\eta_k\beta \leq \left\|\frac{\lambda_{k+1}-\lambda_k}{\rho_k}\right\| + m\eta_k\beta.$$

□

Proposition 4. (Rate on primal suboptimality) Suppose $\{(\mathbf{x}_k, \lambda_k)\}$ is generated by **Sm-AL** scheme. Then we have that

$$-\eta_k(1 + b_\lambda m)\beta - \left(\frac{\|\lambda_{k+1}\|^2}{\rho_k} + \frac{\|\lambda^*_{\eta_k} - \lambda_k\|^2}{\rho_k}\right) \leq f(\mathbf{x}_{k+1}) - f^* \leq \eta_k\beta + \frac{\|\lambda_k\|^2}{2\rho_k} + \epsilon_k\eta_k^b.$$

Proof (i) Let $f^*_{\eta_k} \triangleq f_{\eta_k}(\mathbf{x}^*_{\eta_k})$ and $(\mathbf{x}^*_{\eta_k}, \lambda^*_{\eta_k})$ be the saddle point of $\mathcal{L}_{\eta_k,0}(\mathbf{x},\lambda)$. We have that

$$f^*_{\eta_k} \leq \mathcal{L}_{\eta_k,\rho_k}(\mathbf{x}_{k+1},\lambda^*_{\eta_k}) = f_{\eta_k}(\mathbf{x}_{k+1}) + \frac{\rho_k}{2}\left(d_-\left(\frac{\lambda^*_{\eta_k}}{\rho_k} + g_{\eta_k}(\mathbf{x}_{k+1})\right)\right)^2 - \frac{1}{2\rho_k}\|\lambda^*_{\eta_k}\|^2$$

$$= f_{\eta_k}(\mathbf{x}_{k+1}) + \frac{\rho_k}{2}\left(d_-\left(\frac{\lambda_k}{\rho_k} - \frac{\lambda_k}{\rho_k} + \frac{\lambda^*_{\eta_k}}{\rho_k} + g_{\eta_k}(\mathbf{x}_{k+1})\right)\right)^2 - \frac{1}{2\rho_k}\|\lambda^*_{\eta_k}\|^2$$

$$\leq f_{\eta_k}(\mathbf{x}_{k+1}) + \tfrac{\rho_k}{2}\left(d_-\left(\tfrac{\lambda_k}{\rho_k} + g_{\eta_k}(\mathbf{x}_{k+1})\right) + \left\|\tfrac{\lambda_k}{\rho_k} - \tfrac{\lambda^*_{\eta_k}}{\rho_k}\right\|\right)^2 - \tfrac{1}{2\rho_k}\|\lambda^*_{\eta_k}\|^2$$

$$\leq f_{\eta_k}(\mathbf{x}_{k+1}) + \tfrac{\rho_k}{2}\left(\tfrac{\|\lambda_{k+1}\|}{\rho_k} + \left\|\tfrac{\lambda_k}{\rho_k} - \tfrac{\lambda^*_{\eta_k}}{\rho_k}\right\|\right)^2 - \tfrac{1}{2\rho_k}\|\lambda^*_{\eta_k}\|^2$$

$$\leq f_{\eta_k}(\mathbf{x}_{k+1}) + \tfrac{1}{\rho_k}\left(\|\lambda_{k+1}\|^2 + \left\|\lambda_k - \lambda^*_{\eta_k}\right\|^2\right). \tag{33}$$

By adding and subtracting $f(\mathbf{x}^*_{\eta_k})$, $f^*_{\eta_k}$ and $f_{\eta_k}(\mathbf{x}_{k+1})$, it follows that

$$f^* - f(\mathbf{x}_{k+1}) = \underbrace{f^* - f(\mathbf{x}^*_{\eta_k})}_{\leq b_\lambda m\beta\eta_k \text{ from (29)}} + \underbrace{f(\mathbf{x}^*_{\eta_k}) - f^*_{\eta_k}}_{\leq \eta_k\beta} + \underbrace{f^*_{\eta_k} - f_{\eta_k}(\mathbf{x}_{k+1})}_{(33)} + \underbrace{f_{\eta_k}(\mathbf{x}_{k+1}) - f(\mathbf{x}_{k+1})}_{\leq 0}.$$

Consequently, we have that $f(\mathbf{x}_{k+1}) - f(\mathbf{x}^*) \geq -(1 + b_\lambda m)\eta_k\beta - \left(\frac{\|\lambda_{k+1}\|^2}{\rho_k} + \frac{\|\lambda^*_{\eta_k} - \lambda_k\|^2}{\rho_k}\right)$.

(ii) Similar to the previous analysis in Theorem 2, we have

$$\mathcal{L}_{\eta_k,\rho_k}(\mathbf{x}_{k+1}, \lambda_k) - \mathcal{L}_{\eta_k,\rho_k}(\mathbf{x}^*_{\eta_k}, \lambda_k) \leq \epsilon_k\eta_k^b.$$

which implies

$$f_{\eta_k}(\mathbf{x}_{k+1}) - f^*_{\eta_k}$$

$$\leq \tfrac{\rho_k}{2}\left(\left(d_-\left(\tfrac{\lambda_k}{\rho_k} + g_{\eta_k}(\mathbf{x}^*_{\eta_k})\right)\right)^2 - \left(d_-\left(\tfrac{\lambda_k}{\rho_k} + g_{\eta_k}(\mathbf{x}_{k+1})\right)\right)^2\right) + \epsilon_k\eta_k^b$$

$$\leq \tfrac{\rho_k}{2}\left(\left(d_-\left(\tfrac{\lambda_k}{\rho_k} + g_{\eta_k}(\mathbf{x}^*_{\eta_k})\right)\right)^2\right) + \epsilon_k\eta_k^b \leq \tfrac{\rho_k}{2}\left(\left(d_-\left(g_{\eta_k}(\mathbf{x}^*_{\eta_k})\right) + \tfrac{\|\lambda_k\|}{\rho_k}\right)^2\right) + \epsilon_k\eta_k^b$$

$$= \left(\tfrac{\|\lambda_k\|^2}{2\rho_k}\right) + \epsilon_k\eta_k^b$$

$$\implies f(\mathbf{x}_{k+1}) - f^* = \underbrace{f(\mathbf{x}_{k+1}) - f_{\eta_k}(\mathbf{x}_{k+1})}_{\leq \eta_k\beta} + f_{\eta_k}(\mathbf{x}_{k+1}) - f^*_{\eta_k} + \underbrace{f_{\eta_k}(\mathbf{x}^*_{\eta_k}) - f_{\eta_k}(\mathbf{x}^*)}_{\leq 0 \text{ from (32)}}$$

$$+ \underbrace{f_{\eta_k}(\mathbf{x}^*) - f^*}_{\leq 0} \leq \eta_k\beta + \tfrac{\|\lambda_k\|^2}{2\rho_k} + \epsilon_k\eta_k^b.$$

□

We conclude with an overall rate for sub-optimality and infeasibility.

Theorem 3. Suppose $\{(\mathbf{x}_k, \lambda_k)\}$ is generated by (**Sm-AL**). Let $\eta_k = \frac{1}{\rho_k}$. Then the following holds, where $B_7, B_8 \geq 0$ are constants.

$$|f(\mathbf{x}_{k+1}) - f^*| \leq \eta_k\beta(1 + b_\lambda m) + \frac{B_7}{\rho_k} \text{ and } d_-(g(\mathbf{x}_{k+1})) \leq \eta_k\beta m + \frac{B_8}{\rho_k}.$$

Proof Suppose $\rho_k = \rho_0\zeta^k$ where $\zeta > 1$. By choosing $\epsilon_k\eta_k^b = \frac{1}{k^{2+\delta}\rho_k}$, we have that

$$\begin{aligned}
|f(\mathbf{x}_{k+1}) - f^*| &\leq \max\left\{\eta_k\beta(1 + b_\lambda m) + \frac{\|\lambda_{k+1}\|^2}{\rho_k} + \frac{\|\lambda^*_{\eta_k} - \lambda_k\|^2}{\rho_k}, \eta_k\beta + \frac{\|\lambda_k\|^2}{2\rho_k} + \epsilon_k\eta_k^b\right\} \\
&\leq \eta_k\beta(1 + b_\lambda m) + \frac{2\|\lambda_{k+1}\|^2 + 5\|\lambda_k\|^2 + 4\|\lambda^*_{\eta_k}\|^2}{2\rho_k} + \epsilon_k\eta_k^b \leq \eta_k\beta(1 + b_\lambda m) + \frac{\tilde{C}_1}{\rho_k} + \frac{1}{k^{2+\delta}\rho_k} \\
&\leq \eta_k\beta(1 + b_\lambda m) + \frac{B_7}{\rho_k}.
\end{aligned}$$

Next, we derive a rate on the expected infeasibility. Recall from Lemma 4, $g(\mathbf{x}_{k+1}) \leq g_{\eta_k}(\mathbf{x}_{k+1}) + \eta_k\beta\mathbf{1}$, implying that $d_-(g(\mathbf{x}_{k+1}) \leq d_-(g_{\eta_k}(\mathbf{x}_{k+1}) + \eta_k\beta\mathbf{1})$. Therefore,

$$\begin{aligned}
d_-(g(\mathbf{x}_{k+1})) &\leq d_-(g_{\eta_k}(\mathbf{x}_{k+1}) + \eta_k\beta\mathbf{1}) \leq \left\|\tfrac{\lambda_{k+1} - \lambda_k}{\rho_k}\right\| + \eta_k\beta\|\mathbf{1}\| \\
&\leq \eta_k\beta m + \tfrac{2B_\lambda}{\rho_k} = \eta_k\beta m + \tfrac{B_8}{\rho_k}.
\end{aligned}$$

4 Overall Complexity Guarantees

In 4.1, we begin with some preliminaries, including the derivation of Lipschitzian properties for the smoothed AL function. This allows for employing an accelerated gradient framework for inexact resolution of the subproblem, leading to suitable complexity guarantees in 4.2 for convex and strongly convex regimes. In 4.3, overall complexity guarantees for (**Sm-AL**) with a fixed smoothing parameter are presented.

4.1 Preliminaries

We first derive L-smoothness of $\mathcal{L}_{\eta,\rho}(\bullet, \lambda)$ uniformly in λ. Our bound necessitates utilizing an upper bound on η, which we denoted by η^u.

Lemma 11 *Suppose $0 < \eta \leq \eta^u$ and $\rho \geq 1$. Then the following hold.*
(a) For any $\lambda \geq 0$, there exists $\tilde{C}$ such that $\mathcal{L}_{\eta,\rho}(\bullet, \lambda)$ is $\frac{\tilde{C}\rho}{\eta}$-smooth.
(b) $\mathcal{L}_{\eta,\rho}(\mathbf{x}, \lambda)$ is convex in $\mathbf{x} \in \mathcal{X}$ and concave in $\lambda \geq 0$. □

Next, we formally state an accelerated gradient method for resolving the augmented Lagrangian subproblem (ALSub$_{\eta_k,\rho_k}(\lambda_k)$), defined as

$$\min_{\mathbf{x}\in\mathcal{X}} \mathcal{L}_{\eta_k,\rho_k}(\mathbf{x}, \lambda_k). \qquad (\text{ALSub}_{\eta_k,\rho_k}(\lambda_k))$$

Suppose $\mathbf{x}_k^*$ denotes an optimal solution of (ALSub$_{\eta_k,\rho_k}(\lambda_k)$). Since $\mathcal{L}_{\eta_k,\rho_k}(\bullet, \lambda_k)$ is a convex and $\frac{\tilde{C}\rho_k}{\eta_k}$-smooth function, we employ an accelerated gradient method that constructs a sequence $\{\mathbf{y}_j, \mathbf{z}_j\}_{j=0}^{M_k}$ as follows, where $\mathbf{z}_0 = \mathbf{y}_0 = \mathbf{x}_k$.

$$\left\{\begin{aligned}
\mathbf{y}_{j+1} &= \Pi_X\left[\mathbf{z}_j - \beta_j\nabla_{\mathbf{x}}\mathcal{L}_{\eta_k,\rho_k}(\mathbf{z}_j, \lambda_k)\right] \\
\mathbf{z}_{j+1} &= \mathbf{y}_{j+1} + \gamma_j\left(\mathbf{y}_{j+1} - \mathbf{y}_j\right)
\end{aligned}\right\}, \quad j > 0. \qquad (\text{AG})$$

We now restate the convergence guarantees [6, 31, 32, 34] associated with (**AG**).

Theorem 4. (Thm 2.2.1, Lem 2.2.4 [34]) Suppose $\mathcal{X}$ is a convex and compact set where $\|\mathbf{x}-\mathbf{y}\| \le C_1$ for any $\mathbf{x}, \mathbf{y} \in \mathcal{X}$. Suppose $\eta_k \le \eta^u$ and $\rho_k \ge 1$ for any k. Consider a sequence $\{\mathbf{y}_j, \mathbf{z}_j\}$ generated by (**AG**) when applied to $(\text{ALSub}_{\eta_k,\rho_k}(\lambda_k))$.
(i) Suppose $\beta_j = 1/L_k$, $\alpha_j = (1+(1+\alpha_{j-1}^2)^{1/2})/2$, and $\gamma_j = \frac{\alpha_j - 1}{\alpha_{j+1}}$ for $j \ge 0$, where $\alpha_{-1} = 0$. Then $\mathcal{L}_{\eta_k,\rho_k}(\mathbf{y}_{j+1}, \lambda_k) - \mathcal{L}_{\eta_k,\rho_k}(\mathbf{x}_k^*, \lambda_k) \le \frac{C_1 L_k}{j^2}$ for any $j \ge 0$.
(ii) Suppose $\mathcal{L}_{\eta_k,\rho_k}(\bullet, \lambda_k)$ is a μ-strongly convex and L_k-smooth function. Suppose $\beta_j = 1/L_{\eta_k,\rho_k}$ and $\gamma_j = \frac{\sqrt{\kappa_k}-1}{\sqrt{\kappa_k}+1}$ for $j \ge 0$, where $\kappa_k = L_k/\mu$ for $k \ge 0$. Then $\mathcal{L}_{\eta_k,\rho_k}(\mathbf{y}_{j+1}, \lambda_k) - \mathcal{L}_{\eta_k,\rho_k}(\mathbf{x}_k^*, \lambda_k) \le \tilde{C}\left(\frac{\rho_k}{\eta_k}\right)(1-\frac{1}{\sqrt{\kappa_k}})^j$ for $j \ge 0$, where $(\mathcal{L}_{\eta_k,\rho_k}(\mathbf{x}_k, \lambda_k) - \mathcal{L}_{\eta_k,\rho_k}(\mathbf{x}_k^*, \lambda_k) + \mu C_1{}^2/2) \le \tilde{C}\left(\frac{\rho_k}{\eta_k}\right)$ for any k.

4.2 Complexity guarantees for convex and strongly convex *f*

We begin by leveraging Theorem 4 to develop complexity guarantees in convex settings for an $\boldsymbol{\varepsilon}$-optimal solution by leveraging the rate statement for dual suboptimality (in constant penalty settings) and primal sub-optimality (in increasing penalty settings). Throughout, we recall that AL subproblem objective is L_k-smooth, where $L_k = \frac{\tilde{C}\rho_k}{\eta_k}$ and $\|x-y\| \le C_1$ for any $x, y \in X$. Additionally, complexity guarantees are derived by utilizing the rate guarantees presented in Theorem 2 (Constant ρ_0) or Theorem 3 (increasing ρ_k) to determine the number of outer iterations K; specifically, by these results, to ensure $\boldsymbol{\varepsilon}$-suboptimal solutions, we require that $K = \lceil \frac{C}{\boldsymbol{\varepsilon}} \rceil$ (constant ρ) or $K = \lceil \frac{\ln(C/\boldsymbol{\varepsilon})}{\ln(\zeta)} \rceil$ (increasing ρ_k) for a suitable constant C.

Theorem 5. (Overall complexity of **Sm-AL** for convex f) Consider $\{(\mathbf{x}_k, \lambda_k)\}$ generated by (**Sm-AL**). Suppose $\rho_0 \ge 1$, $\boldsymbol{\varepsilon}, \delta > 0$, $b \ge 0$, $\|x-y\| \le C_1$ for any $x, y \in X$ and $\tilde{C}$ is suitably defined based on the smoothing property.
(a) (Constant ρ). Let $\rho_k = \rho_0$, $\eta_k = k^{-(2+\delta)}$, $\epsilon_k = \eta_k^{-b} k^{-(2+\delta)}$, and $M_k = \left\lceil (C_1\tilde{C}\rho_0)^{1/2} k^{2(1+\delta)} \right\rceil$ for $k > 0$. Suppose K is chosen such that $\left(\bar{\mathbf{x}}_K, \bar{\lambda}_K\right)$ satisfies $f^* - \mathcal{D}(\bar{\lambda}_K) \le \boldsymbol{\varepsilon}$ where $\bar{\mathbf{x}}_K = \sum_{i=1}^K \mathbf{x}_i/K$ and $\bar{\lambda}_K = \sum_{i=1}^K \lambda_i/K$. If $K(\boldsymbol{\varepsilon}) = \lceil \frac{C}{\boldsymbol{\varepsilon}} \rceil$, then the overall iteration complexity of computing such an $\bar{\mathbf{x}}_K$ satisfies $\sum_{k=1}^{K(\boldsymbol{\varepsilon})} M_k \le \mathcal{O}\left(\boldsymbol{\varepsilon}^{-(3+\delta)}\right)$.
(b) (Geometrically increasing ρ_k). Let $\rho_k = \rho_0 \zeta^k$, $\eta_k = \frac{1}{\rho_k} k^{-(2+\delta)}$, $\epsilon_k = \frac{1}{\rho_k \eta_k^b} k^{-(2+\delta)}$ and $M_k = \left\lceil \sqrt{C_1\tilde{C}}\rho_k^{3/2} k^{2+\delta} \right\rceil$ for all $k > 0$, where $\zeta > 1$. Suppose K is chosen such that $(\mathbf{x}_K, \lambda_K)$ satisfies $|f^* - f(\mathbf{x}_K)| \le \boldsymbol{\varepsilon}$. If $K(\boldsymbol{\varepsilon}) = \lceil \frac{\ln(C/\boldsymbol{\varepsilon})}{\ln(\zeta)} \rceil$, then the overall iteration complexity of computing such an $\mathbf{x}_K$ satisfies $\sum_{k=1}^{K(\boldsymbol{\varepsilon})} M_k \le \tilde{\mathcal{O}}(\boldsymbol{\varepsilon}^{-\frac{3}{2}})$.

Proof (a) By Theorem 4, M_k is the smallest integer satisfying

$$\mathcal{L}_{\rho_k,\eta_k}(\mathbf{x}_{k+1}, \lambda_k) - \mathcal{D}_{\rho_k,\eta_k}(\lambda_k) \le \left(\frac{C_1 L_k}{M_k^2}\right) = \left(\frac{C_1\tilde{C}\rho_0}{\eta_k M_k^2}\right) \le \epsilon_k \eta_k^b$$
$$\implies M_k = \left\lceil \sqrt{\frac{C_1\tilde{C}\rho_0}{\epsilon_k \eta_k^{b+1}}} \right\rceil = \left\lceil \left(\sqrt{C_1\tilde{C}\rho_0}\right) k^{2(1+\delta)} \right\rceil.$$

Then the iteration complexity of computing a $(\bar{\mathbf{x}}_K, \bar{\lambda}_K)$ where $f^* - \mathcal{D}(\bar{\lambda}_K) \leq \boldsymbol{\varepsilon}$ requires

$$\sum_{k=1}^{K(\boldsymbol{\varepsilon})} M_k = \sum_{k=1}^{\lceil C/\boldsymbol{\varepsilon} \rceil} \left\lceil \left(\sqrt{C_1 \tilde{C} \rho_0} \right) k^{2(1+\delta)} \right\rceil = \mathcal{O}\left(\boldsymbol{\varepsilon}^{-(3+2\delta)} \right).$$

(b) Proceeding similarly, by Theorem 4, M_k is defined as follows.

$$M_k = \left\lceil \sqrt{\frac{C_1 \tilde{C} \rho_k}{\epsilon_k \eta_k^{b+1}}} \right\rceil = \left\lceil \sqrt{\frac{C_1 \tilde{C} \rho_k^2 \eta_k^b k^{(2+\delta)}}{\eta_k^{b+1}}} \right\rceil = \left\lceil \left(\sqrt{C_1 \tilde{C}} \right) \rho_k^{3/2} k^{2+\delta} \right\rceil.$$

Then the iteration complexity of producing an $\mathbf{x}_K$ satisfying $|f^* - f(\mathbf{x}_K)| \leq \boldsymbol{\varepsilon}$ requires

$$\begin{aligned} \sum_{k=1}^{K(\boldsymbol{\varepsilon})} M_k &= \sum_{k=1}^{\lceil \ln \frac{C}{\boldsymbol{\varepsilon}} / \ln \zeta \rceil} \left\lceil \left(\sqrt{C_1 \tilde{C}} \right) \rho_k^{\frac{3}{2}} k^{(2+\delta)} \right\rceil \leq 2 \left(\sqrt{C_1 \tilde{C}} \right) \rho_0^{\frac{3}{2}} \sum_{k=1}^{\log_\zeta \left(\frac{C}{\boldsymbol{\varepsilon}} \right)+1} \zeta^{\frac{3}{2}k} k^{(2+\delta)} \\ &\leq 2 \left(\sqrt{C_1 \tilde{C}} \right) \rho_0^{3/2} \left(\lceil \ln \left(\tfrac{C}{\boldsymbol{\varepsilon}} \right) + 1 \rceil \right)^{3(1+\delta)} \int_1^{\ln_\zeta \left(\frac{C}{\boldsymbol{\varepsilon}} \right)+2} \zeta^{\frac{3}{2}u} du \leq \tilde{\mathcal{O}}\left(\boldsymbol{\varepsilon}^{-\frac{3}{2}} \right). \end{aligned}$$

Remark 1 **(Constant ρ.)** Suppose $\boldsymbol{\varepsilon}$ is a positive scalar. Let $K \triangleq \lceil C/\boldsymbol{\varepsilon} \rceil$ where C is defined in Proposition 2. Suppose **Sm-AL** scheme runs for K iterations and produces $\bar{\mathbf{x}}_K$ and $\bar{\lambda}_K$. Then we have that

$$f^* - \mathcal{D}(\bar{\lambda}_K) \leq \boldsymbol{\varepsilon}, |f^* - f(\bar{\mathbf{x}}_K)| \leq \mathcal{O}\left(\sqrt{\boldsymbol{\varepsilon}} \right), \text{ and } d_-\left(g(\bar{\mathbf{x}}_K) \right) \leq \mathcal{O}\left(\sqrt{\boldsymbol{\varepsilon}} \right).$$

(Increasing ρ_k). Suppose $\boldsymbol{\varepsilon}$ is a positive scalar. Let $K \triangleq \lceil \ln \left(\frac{C}{\boldsymbol{\varepsilon}} \right) / \ln(\zeta) \rceil$ where C is defined in Theorem 3 and $\rho_k = \rho_0 \zeta^k$ with $\zeta > 1$. Suppose **Sm-AL** scheme runs for K iterations and produces $\bar{\mathbf{x}}_K$ and $\bar{\lambda}_K$, where

$$\left| f^* - f\mathbf{x}_K) \right| \leq \boldsymbol{\varepsilon} \text{ and } d_-\left(g(\mathbf{x}_K) \right) \leq \mathcal{O}(\boldsymbol{\varepsilon}).$$

We now produce an extension of the results for strongly convex settings.

Theorem 6. (Overall complexity of **Sm-AL** for strongly convex f) Suppose f is μ-strongly convex on $\mathcal{X}$. Consider a sequence $\{(\mathbf{x}_k, \lambda_k)\}$ generated by (**Sm-AL**). Suppose $\rho_0, \boldsymbol{\varepsilon}, \delta > 0$, $b \geq 0$, $\|x - y\| \leq C_1$ for any $x, y \in X$ and $\tilde{C}$ is suitably defined based on the smoothing property.

(a) (Constant ρ). Let $M_k = \left\lceil \left(\frac{\ln \left(\frac{\tilde{C}\rho_0}{\epsilon_k \eta_k^{b+1}} \right)}{\ln \left(\frac{\sqrt{L_k}}{\sqrt{L_k} - \sqrt{\mu}} \right)} \right) \right\rceil$, $\rho_k = \rho_0$, $\eta_k = k^{-(2+\delta)}$, and $\epsilon_k = \eta_k^{-b} k^{-(2+\delta)}$ for all $k > 0$, where $\delta > 0$. Suppose K is chosen such that $\left(\bar{\mathbf{x}}_K, \bar{\lambda}_K \right)$ satisfies $f^* - \mathcal{D}(\bar{\lambda}_K) \leq \boldsymbol{\varepsilon}$ where $\bar{\mathbf{x}}_K = (\sum_{i=1}^K \mathbf{x}_i)/K$ and $\bar{\lambda}_K = (\sum_{i=1}^K \lambda_i)/K$. If $K(\boldsymbol{\varepsilon}) = \lceil \frac{C}{\boldsymbol{\varepsilon}} \rceil$, then the overall iteration complexity of computing an $\bar{\mathbf{x}}_K$ satisfies $\sum_{k=1}^{K(\boldsymbol{\varepsilon})} M_k \leq \tilde{\mathcal{O}}\left(\frac{1}{\boldsymbol{\varepsilon}^2} \right)$.

(b) (Geometrically increasing ρ_k). Let $M_k = \left\lceil \left(\frac{\ln \left(\frac{\tilde{C}\rho_k}{\epsilon_k \eta_k^{b+1}} \right)}{\ln \left(\frac{\sqrt{L_k}}{\sqrt{L_k} - \sqrt{\mu}} \right)} \right) \right\rceil$, $\rho_k = \rho_0 \zeta^k$, $\eta_k = \rho_k^{-1} k^{-(2+\delta)}$, and $\epsilon_k = \rho_k^{-1} \eta_k^{-b} k^{-(2+\delta)}$ for $k > 0$, where $\delta, \rho > 0, \zeta > 1$. Suppose K is chosen such that $(\mathbf{x}_K, \lambda_K)$ satisfies $|f^* - f(\mathbf{x}_K)| \leq \boldsymbol{\varepsilon}$. If $K(\boldsymbol{\varepsilon}) = \lceil \frac{\ln(C/\boldsymbol{\varepsilon})}{\ln(\zeta)} \rceil$, then the overall iteration complexity of computing an $\mathbf{x}_K$ satisfies $\sum_{k=1}^{K(\boldsymbol{\varepsilon})} M_k \leq \tilde{\mathcal{O}}\left(\frac{1}{\boldsymbol{\varepsilon}} \right)$.

Proof (a) Suppose $\rho_k = \rho_0$ for all k. Suppose M_k represents the least number of steps taken at step k to achieve $(\epsilon_k \eta_k^b)$-optimality of the subproblem. By Theorem 4 and $\ln(x) \geq \frac{x-1}{x}$ for $x > 0$,

$$\mathcal{L}_{\rho_k,\eta_k}(\mathbf{x}_{k+1}, \lambda_k) - \mathcal{D}_{\rho_k,\eta_k}(\lambda_k) \leq \tilde{C}\frac{\rho_0}{\eta_k}\left(1 - \frac{\sqrt{\mu}}{\sqrt{L_k}}\right)^{M_k} \leq \epsilon_k \eta_k^b.$$

$$\begin{aligned}
\implies M_k &= \left\lceil \left(\frac{\ln\left(\frac{\tilde{C}\rho_0}{\epsilon_k \eta_k^{b+1}}\right)}{\ln\left(\frac{\sqrt{L_k}}{\sqrt{L_k}-\sqrt{\mu}}\right)} \right) \right\rceil \leq \left\lceil \left(\frac{\ln\left(\tilde{C}\rho_0 k^{4+2\delta}\right)}{\left(1-\left(\frac{\sqrt{L_k}-\sqrt{\mu}}{\sqrt{L_k}}\right)\right)} \right) \right\rceil \\
&= \left\lceil \left(\frac{\ln\left(\tilde{C}\rho_0 k^{4+2\delta}\right)}{\left(\frac{\sqrt{\mu}}{\sqrt{L_k}}\right)} \right) \right\rceil = \left\lceil \left(\frac{\sqrt{\tilde{C}\rho_0}\ln\left(\tilde{C}\rho_0 k^{4+2\delta}\right)}{(\sqrt{\mu\eta_k})} \right) \right\rceil \\
&= \left\lceil \left(\frac{\sqrt{\tilde{C}\rho_0}\ln\left((\hat{C}k)^{4+2\delta}\right)}{(\sqrt{\mu\eta_k})} \right) \right\rceil, \text{ where } \hat{C} = (\tilde{C}\rho_0)^{1/(4+2\delta)}.
\end{aligned}$$

Consequently, since $K(\boldsymbol{\varepsilon}) = \lceil C/\boldsymbol{\varepsilon} \rceil$ outer steps are required, the overall complexity is

$$\begin{aligned}
\sum_{k=1}^{K(\boldsymbol{\varepsilon})} M_k &= \sum_{k=1}^{\lceil C/\boldsymbol{\varepsilon} \rceil} \left\lceil \left(\frac{\sqrt{\tilde{C}\rho_0}\ln\left((\hat{C}k)^{4+2\delta}\right)}{(\sqrt{\mu\eta_k})} \right) \right\rceil = \sum_{k=1}^{\lceil C/\boldsymbol{\varepsilon} \rceil} \left\lceil \left(\frac{(4+2\delta)k^{1+\delta}\sqrt{\tilde{C}\rho_0}\ln\left((\hat{C}k)\right)}{(\sqrt{\mu})} \right) \right\rceil \\
&\leq \mathcal{O}\left(\frac{1}{\boldsymbol{\varepsilon}^{2+2\delta}} \ln\left(\frac{1}{\boldsymbol{\varepsilon}}\right)\right).
\end{aligned}$$

(b) Consider $\rho_k = \rho_0 \zeta^k$ where $k \geq 0$ and $\zeta > 1$. Proceeding as in (a) and by Theorem 4 and $\ln(x) \geq \frac{x-1}{x}$ for $x > 0$,

$$\begin{aligned}
&\mathcal{L}_{\rho_k,\eta_k}(\mathbf{x}_{k+1}, \lambda_k) - \mathcal{D}_{\rho_k,\eta_k}(\lambda_k) \leq \tilde{C}\left(\frac{\rho_k}{\eta_k}\right)\left(1 - \frac{\sqrt{\mu}}{\sqrt{L_k}}\right)^{M_k} \leq \epsilon_k \eta_k^b \\
\implies M_k &= \left\lceil \left(\frac{\ln\left(\frac{\tilde{C}\rho_k}{\epsilon_k \eta_k^{b+1}}\right)}{\ln\left(\frac{\sqrt{L_k}}{\sqrt{L_k}-\sqrt{\mu}}\right)} \right) \right\rceil \leq \left\lceil \left(\frac{\ln\left(\tilde{C}k^{(4+2\delta)}\rho_k^3\right)}{\left(1-\frac{\sqrt{L_k}-\sqrt{\mu}}{\sqrt{L_k}}\right)} \right) \right\rceil \leq \frac{2\sqrt{\rho_k}\ln\left(\rho_k^3 \tilde{C} k^{(4+2\delta)}\right)}{\sqrt{\mu\eta_k}}.
\end{aligned}$$

Consequently, if $K(\boldsymbol{\varepsilon}) = \lceil \ln(C/\boldsymbol{\varepsilon})/\ln(\zeta) \rceil = \lceil \log_\zeta(C/\boldsymbol{\varepsilon}) \rceil$ outer steps are employed, then the overall complexity can be bounded as follows.

$$\begin{aligned}
\sum_{k=1}^{K(\boldsymbol{\varepsilon})} M_k &= \sum_{k=1}^{\lceil \log_\zeta(C/\boldsymbol{\varepsilon}) \rceil} 2\left\lceil \frac{\sqrt{\rho_k}}{\sqrt{\mu\eta_k}} \ln\left(\rho_k^3 \tilde{C} k^{(4+2\delta)}\right) \right\rceil \\
&\leq \sum_{k=1}^{\lceil \log_\zeta(C/\boldsymbol{\varepsilon}) \rceil} \tilde{C}_1 \left\lceil \rho_k k^{(1+\delta)} \ln\left(\rho_k^3 \tilde{C} k^{(4+2\delta)}\right) \right\rceil \\
&\leq \rho_0 \zeta^{(\lceil \log_\zeta(C/\boldsymbol{\varepsilon}) \rceil)} \left(\lceil \log_\zeta(C/\boldsymbol{\varepsilon}) \rceil\right)^{(1+\delta)} \\
&\times \ln\left(\rho_0^3 \zeta^{3(\lceil \log_\zeta(C/\boldsymbol{\varepsilon}) \rceil)} \tilde{C} \left(\lceil \log_\zeta(C/\boldsymbol{\varepsilon}) \rceil\right)^{(4+2\delta)}\right) \\
&\leq \tilde{\mathcal{O}}\left(\frac{1}{\boldsymbol{\varepsilon}}\right).
\end{aligned}$$

□

Remark 2 **Sm-AL** is designed for convex problems with **nonsmooth nonlinear** convex constraints, achieving an overall complexity of $\tilde{\mathcal{O}}\left(\boldsymbol{\varepsilon}^{-3/2}\right)$ under geometric growth of ρ_k, slightly worse than the best known complexities for contending with **smooth** nonlinear constraints (cf. [26, 44]), i.e. $\mathcal{O}(\boldsymbol{\varepsilon}^{-1})$ (up to log. terms).

4.3 Complexity Analysis for (Sm-AL) with fixed η

Next, we apply (**Sm-AL**) to (NSCopt$_\eta$) with a fixed and appropriately chosen η with the overall goal of finding an $(\bar{\mathbf{x}}_K, \bar{\lambda}_K)$ such that either dual suboptimality is sufficiently small, i.e. $f_\eta^* - \mathcal{D}_{\eta,0}(\bar{\lambda}_K) \leq \boldsymbol{\varepsilon}$ (constant $\rho_k = \rho_0$) or primal suboptimality is sufficiently small $|f_\eta(\mathbf{x}_K) - f_\eta^*| < \boldsymbol{\varepsilon}$ (geometrically increasing ρ_k).
(a) (Constant ρ) Suppose $\eta \leq \tilde{c}\boldsymbol{\varepsilon}$, where $\tilde{c}$ needs specification. After K steps in (**Sm-AL**), $f_\eta^* - \mathcal{D}_{\eta,0}(\bar{\lambda}_K) \leq \frac{\boldsymbol{\varepsilon}}{2}$, where $K = \left\lceil \frac{C}{\boldsymbol{\varepsilon}} \right\rceil$ for a suitable C. By Lemma 4,

$$\begin{aligned} f(\mathbf{x}^*) - \mathcal{D}_0(\bar{\lambda}_K) &\leq f_\eta(\mathbf{x}^*) + \eta\beta - \mathcal{D}_{\eta,0}(\bar{\lambda}_K) + \eta(\|\bar{\lambda}_K\|m + 1)\beta \\ &\leq \underbrace{f_\eta(\mathbf{x}_\eta^*) - \mathcal{D}_{\eta,0}(\bar{\lambda}_K)}_{\leq \frac{\boldsymbol{\varepsilon}}{2}} + \underbrace{\eta\left(\beta(\tilde{B}_\lambda m + 2)\right)}_{\leq \frac{\boldsymbol{\varepsilon}}{2}} \leq \boldsymbol{\varepsilon}. \end{aligned}$$

To ensure that the second term is less than $\boldsymbol{\varepsilon}/2$, we select $\eta \leq \frac{\boldsymbol{\varepsilon}}{2\left(\beta(2+\tilde{B}_\lambda m)\right)}$.
(b) (Geometrically increasing ρ_k). Proceeding similarly, suppose $\eta \leq \tilde{c}\boldsymbol{\varepsilon}$, then by taking K steps in (**Sm-AL**), $|f_\eta(\mathbf{x}_K) - f_\eta^*| \leq \frac{\boldsymbol{\varepsilon}}{2}$, where $K = \lceil \frac{C}{\boldsymbol{\varepsilon}} \rceil$ for a suitable C. Consequently, we have that if $\eta \leq \frac{\boldsymbol{\varepsilon}}{2\beta}$, we have that $f(\mathbf{x}_K) - f^* \leq \boldsymbol{\varepsilon}$.

$$f(\mathbf{x}_K) - f^* \leq f_\eta(\mathbf{x}_K) - f_\eta(\mathbf{x}^*) + \eta\beta \leq \underbrace{f_\eta(\mathbf{x}_K) - f_\eta(\mathbf{x}_\eta^*)}_{\leq \frac{\boldsymbol{\varepsilon}}{2}} + \underbrace{\eta\beta}_{\leq \frac{\boldsymbol{\varepsilon}}{2}} \leq \boldsymbol{\varepsilon}.$$

Similarly, if $\eta \leq \frac{\boldsymbol{\varepsilon}}{2\beta}$, $f^* - f(\mathbf{x}_K) \leq \boldsymbol{\varepsilon}$, implying that if $\eta \leq \frac{\boldsymbol{\varepsilon}}{2\beta}$, $|f(\mathbf{x}_K) - f^*| \leq \boldsymbol{\varepsilon}$.

Proposition 5. (Complexity analysis of **AL** for η-smoothed convex problems) Consider a sequence $\{(\mathbf{x}_k, \lambda_k)\}$ generated by (**Sm-AL**). Suppose $\rho_0, \boldsymbol{\varepsilon} > 0$, $\|x - y\| \leq C_1$ for any $x, y \in X$ and $\tilde{C}$ is suitably defined based on the smoothing property.
(a.) (Constant ρ). Let $\rho_k = \rho_0, \epsilon_k = k^{-(2+\delta)}$, $\eta = \frac{\boldsymbol{\varepsilon}}{2\left(\beta(2+\tilde{B}_\lambda m)\right)}$, and $M_k = \left\lceil \sqrt{\frac{C_1\tilde{C}\rho_0}{\eta\epsilon_k}} \right\rceil$ for $k > 0$, where $\delta > 0$. Suppose K is chosen such that $(\bar{\mathbf{x}}_K, \bar{\lambda}_K)$ satisfies $f^* - \mathcal{D}(\bar{\lambda}_K) \leq \boldsymbol{\varepsilon}$ where $\bar{\mathbf{x}}_K = \sum_{i=1}^K \mathbf{x}_i/K$ and $\bar{\lambda}_K = \sum_{i=1}^K \lambda_i/K$. Let $K(\boldsymbol{\varepsilon}) = \left\lceil \frac{C}{\boldsymbol{\varepsilon}} \right\rceil$ where C is a constant. Then the overall iteration complexity of computing such $\bar{\mathbf{x}}_K$ satisfies $\sum_{k=1}^{K(\boldsymbol{\varepsilon})} M_k \leq \mathcal{O}(\boldsymbol{\varepsilon}^{-\left(\frac{5}{2}+\delta\right)})$.
(b.) (Geometrically increasing ρ_k.) Let $\rho_k = \rho_0\zeta^k, \epsilon_k = \rho_k^{-1}k^{-(2+\delta)}$, $\eta = \frac{\boldsymbol{\varepsilon}}{2\beta}$ and $M_k = \left\lceil \sqrt{\frac{C_1\tilde{C}\rho_k}{\eta\epsilon_k}}k^{1+\delta} \right\rceil$ for all $k > 0$ where $\delta, \rho_0 > 0, \eta > 1$. Suppose K is chosen such that $(\bar{\mathbf{x}}_K, \lambda_K)$ satisfies $|f^* - f(\mathbf{x}_K)| \leq \boldsymbol{\varepsilon}$. Let $K(\boldsymbol{\varepsilon}) = \ln(C/\boldsymbol{\varepsilon})/\ln(\zeta)$ where C is a constant. Then the overall iteration complexity of computing such $\mathbf{x}_K$ satisfies $\sum_{k=1}^{K(\boldsymbol{\varepsilon})} M_k \leq \tilde{O}(\boldsymbol{\varepsilon}^{-\frac{3}{2}})$.

Proof (a.) By Theorem 4, M_k is the smallest integer satisfying

$$\mathcal{L}_{\rho_k,\eta}(\mathbf{x}_{k+1},\lambda_k)-\mathcal{D}_{\rho_k,\eta_k}(\lambda_k)\le\left(\frac{C_1L_k}{M_k^2}\right)\le\left(\frac{C_1\tilde{C}\rho_0}{\eta M_k^2}\right)\le\epsilon_k$$

$$\Longrightarrow M_k=\left\lceil\sqrt{\frac{C_1\tilde{C}\rho_0}{\epsilon_k\eta}}\right\rceil=\left\lceil\left(\sqrt{\frac{2C_1\tilde{C}(\beta(2+B_\lambda m))\rho_0}{\boldsymbol{\varepsilon}}}\right)k^{1+\delta}\right\rceil=\left\lceil\left(\sqrt{\frac{D\rho_0}{\boldsymbol{\varepsilon}}}\right)k^{1+\delta}\right\rceil$$

where $C_1,\tilde{C},\beta,B_\lambda$ are constants and $D\triangleq 2C_1\tilde{C}\,(\beta(2+B_\lambda m))$. Then the complexity of computing a $(\bar{\mathbf{x}}_K,\bar{\lambda}_K)$ where $f^*-\mathcal{D}_0(\bar{\lambda}_K)\le\boldsymbol{\varepsilon}$ requires

$$\sum_{k=1}^{K(\boldsymbol{\varepsilon})}M_k=\sum_{k=1}^{\lceil C/\boldsymbol{\varepsilon}\rceil}\left\lceil\left(\sqrt{\frac{D\rho_0}{\boldsymbol{\varepsilon}}}\right)k^{1+\delta}\right\rceil=\sqrt{D\rho_0}\boldsymbol{\varepsilon}^{-\frac{1}{2}}\sum_{k=1}^{\lceil C/\boldsymbol{\varepsilon}\rceil}\left\lceil k^{1+\delta}\right\rceil\le\mathcal{O}\left(\boldsymbol{\varepsilon}^{-\left(\frac{5}{2}+\delta\right)}\right).$$

(b) Consider $\rho_k=\rho_0\zeta^k$ where $k\ge 0$ and $\zeta>1$. Proceeding as in (a) and by invoking Theorem 4,

$$M_k=\left\lceil\sqrt{\frac{C_1\tilde{C}\rho_k}{\epsilon_k\eta}}\right\rceil=\left\lceil\sqrt{\frac{2C_1\tilde{C}\beta}{\boldsymbol{\varepsilon}}}\rho_k k^{1+\delta}\right\rceil=\left\lceil\sqrt{\frac{D}{\boldsymbol{\varepsilon}}}\rho_k k^{1+\delta}\right\rceil$$

where $C_1,\tilde{C},\beta$ are constants and $D\triangleq 2C_1\tilde{C}\beta$. Then the iteration complexity of producing an $\mathbf{x}_K$ satisfying $|f-f(\mathbf{x}_K)|\le\boldsymbol{\varepsilon}$ leads to the following bound, where $C,D>0$.

$$\sum_{k=1}^{K(\boldsymbol{\varepsilon})}M_k=\sum_{k=1}^{\lceil\ln\frac{C}{\boldsymbol{\varepsilon}}/\ln\zeta\rceil}\left\lceil\left(\sqrt{\frac{D}{\boldsymbol{\varepsilon}}}\right)\rho_k k^{(1+\delta)}\right\rceil\le\left(\sqrt{\frac{D}{\boldsymbol{\varepsilon}}}\right)\rho_0^2\sum_{k=1}^{\log_\zeta\left(\frac{C}{\boldsymbol{\varepsilon}}\right)+1}\zeta^k k^{(1+\delta)}$$

$$\le\sqrt{D}\rho_0^2\boldsymbol{\varepsilon}^{-\frac{1}{2}}\left(\lceil\log_\zeta\left(\tfrac{C}{\boldsymbol{\varepsilon}}\right)+1\rceil\right)^{2(1+\delta)}\int_1^{\log_\zeta\left(\frac{C}{\boldsymbol{\varepsilon}}\right)+2}\zeta^u du\le\tilde{\mathcal{O}}\left(\boldsymbol{\varepsilon}^{-\frac{3}{2}}\right).$$

□

Remark 3 We observe that the complexity guarantees are close to those for diminishing η_k with a slight improvement in the constant ρ_0 regime. We recall that Nesterov [33] and Beck and Teboulle [7] adopted different smoothing techniques with fixed η to get an $\boldsymbol{\varepsilon}$-optimal solution within $\mathcal{O}(1/\boldsymbol{\varepsilon})$. When compared to these smoothing schemes in [7, 33], **Sm-AL** targets problems with nonsmooth constraint functions. Moreover, **Sm-AL** accommodates both fixed and varying η, with an effective complexity rate $\tilde{\mathcal{O}}(\epsilon^{-3/2})$, matching the complexity of a smoothed penalized scheme [3].

Table 2 summarizes rate and complexities for **S-AL**, **S-AL**(η), **S-AL**(S), and **N-AL** where (a). **Sm-AL** is smoothed ALM for convex problems; (b). **Sm-AL**(η) is η-smoothed ALM; (c). **Sm-AL**(S) is **Sm-AL** for strongly convex problems; (d). **N-AL** is original ALM for nonsmooth problems. Additionally, Table 3 captures all of the constants utilized in the results from Sections 3 and 4 in a single table.

5 Numerical Experiments

5.1 Fused Lasso Problems

In this section, we apply (**Sm-AL**) on a fused lasso problem with datasets $\{X_i,y_i\}_{i=1}^N$ where X_i is the d-dimensional feature vector for ith instance and y_i is the corresponding response.

Table 2 Rates & Complexity

	$\rho_k = \rho_0$			$\rho_k = \rho_0 \zeta^k$		
	$f(\bar{\mathbf{x}}_k) - f(\mathbf{x}^*)$	$d_-(g(\bar{\mathbf{x}}_k))$	Complexity†	$f(\mathbf{x}_k) - f(\mathbf{x}^*)$	$d_-(g(\bar{\mathbf{x}}_k))$	Complexity$^\bigstar$
Sm-AL	$\mathcal{O}\left(\frac{1}{\sqrt{K}}\right)$	$\mathcal{O}\left(\frac{1}{\sqrt{K}}\right)$	$\mathcal{O}\left(\boldsymbol{\varepsilon}^{-(3+\delta)}\right)$	$\mathcal{O}\left(\frac{1}{\rho_K}\right)$	$\mathcal{O}\left(\frac{1}{\rho_K}\right)$	$\tilde{\mathcal{O}}\left(\boldsymbol{\varepsilon}^{-3/2}\right)$
Sm-AL(S)	$\mathcal{O}\left(\frac{1}{\sqrt{K}}\right)$	$\mathcal{O}\left(\frac{1}{\sqrt{K}}\right)$	$\tilde{\mathcal{O}}\left(\boldsymbol{\varepsilon}^{-(2+\delta)}\right)$	$\mathcal{O}\left(\frac{1}{\rho_K}\right)$	$\mathcal{O}\left(\frac{1}{\rho_K}\right)$	$\tilde{\mathcal{O}}\left(\boldsymbol{\varepsilon}^{-1}\right)$
N-AL	$\mathcal{O}\left(\frac{1}{\sqrt{K}}\right)$	$\mathcal{O}\left(\frac{1}{\sqrt{K}}\right)$	$\mathcal{O}\left(\boldsymbol{\varepsilon}^{-(5+\delta)}\right)$	$\mathcal{O}\left(\frac{1}{\rho_K}\right)$	$\mathcal{O}\left(\frac{1}{\rho_K}\right)$	$\tilde{\mathcal{O}}\left(\boldsymbol{\varepsilon}^{-4}\right)$
Sm-AL(η)	$\mathcal{O}\left(\frac{1}{\sqrt{K}}\right)$	$\mathcal{O}\left(\frac{1}{\sqrt{K}}\right)$	$\mathcal{O}\left(\boldsymbol{\varepsilon}^{-(5/2+\delta)}\right)$	$\mathcal{O}\left(\frac{1}{\rho_K}\right)$	$\mathcal{O}\left(\frac{1}{\rho_K}\right)$	$\tilde{\mathcal{O}}\left(\boldsymbol{\varepsilon}^{-3/2}\right)$

†: Dual suboptimality

★: Primal suboptimality or Primal infeasibility

Table 3 Constants in Theorems/Propositions

	Const.	Description	Value
Prop.2	B_λ	Bounds on $\|\lambda_K - \lambda^*\|$	$\sum_{k=0}^{\infty}\left(\sqrt{2\rho_k\epsilon_k\eta_k^b} + 2\sqrt{\eta_k\rho_k(\|\lambda^*\|m + C_m)\beta}\right) + \|\lambda_0 - \lambda^*\|$
	B_2	Related to η_k	$(2(B_\lambda + 2b_\lambda)m + 1)\beta$
Prop.3	B_3	Related to η_k	$\sqrt{m}\sum_{i=0}^{\infty}\left(\sqrt{\frac{2\epsilon_i\eta_i^b}{\rho}} + \sqrt{m}\eta_i\beta + \sqrt{\frac{2\eta_i\beta\tilde{B}_\lambda}{\rho}}\right)$
	B_4	Bound for $(\mathcal{D}_\rho(\lambda^*) - \mathcal{D}_\rho(\bar{\lambda}_K))$	$\sqrt{\frac{2mC}{\rho}}$
Thm.2	B_5	Bound for $d_-\left(g_{\eta_K}(\bar{\mathbf{x}}_K)\right)$	$\frac{B_3}{\sqrt{K}} + B_4$
	B_6	Bound for $f(\bar{\mathbf{x}}_K) - f^*$	$\frac{\|\lambda_0\|^2}{2\rho} + \sum_{k=0}^{\infty}(\epsilon_k\eta_k^b + \eta_k\beta)$
Thm.3	B_7	Bound for $\|f_{\eta_k}(\mathbf{x}_{k+1}) - f_{\eta_k}^*\|$	$7B_\lambda^2 + 7b_\lambda^2 + 2b_{\lambda,\eta}^2 + k^{-(2+\delta)}$
	B_8	Bound for $\|\lambda_{k+1} - \lambda_k\|$	$2B_\lambda$
Thm.5	C_1	Constant in AG (Thm.4)	The diameter of $\mathcal{X}$ (Thm 4)
	C_2	Constant in Lipschitz const (Lem.9)	$\mathcal{L}_{\eta,\rho}$ is $\frac{C_2\rho}{\eta}$-smooth
Thm.6	C_3	Constant in AG (Thm.4)	$\mathcal{L}_{\eta_k,\rho_k}(\mathbf{x}_{k+1}, \lambda_k) - \mathcal{L}_{\eta_k,\rho_k}(\mathbf{x}^*, \lambda_k) + \mu B^2/2 \le C_3$ (Thm.4)

Consider the η-smoothing of (1).

$$\begin{aligned} &\min_{\beta\in\mathcal{X}} \quad \|Y - X^\top\beta\|^2 \\ &\text{subject to} \quad \sum_j \left(\sqrt{\beta_j^2+\eta^2} - \eta\right) \le C_1, \sum_j \left(\sqrt{(\beta_j-\beta_{j-1})^2+\eta^2} - \eta\right) \le C_2. \end{aligned}$$

We conducted the experiments on simulated datasets with dimensions of β ranging from 5 to 1000. The results are shown in Table 4. The optimal solutions for each experiment are obtained by using *fmincon* in Matlab. In Table 4, we compare the results from **Sm-AL** with those from **N-AL**. Both **Sm-AL** and **N-AL** terminated at 50 outer iterations except that $n = 1000$ case for **Sm-AL** was stopped at the 30th outer iteration to save time. **N-AL** was terminated when the overall runtime exceeded two hours for higher dimensional problems. In all cases, **Sm-AL** outperforms **N-AL** with respect to primal suboptimality and overall runtime.

Next, we compare the results from **Sm-AL** with **AL** on an η-smoothed problem for a single instance ($n = 5$). We observe that such fixed-smoothing avenues provide relatively coarse approximations compared to their iteratively smoothed counterparts. Finally, we compare empirical rates of **Sm-AL** in two settings of ρ_k for a smaller problem ($n = 5$) in terms of primal suboptimality in Figure 1 and observe alignment with the theoretical rates, represented by blue lines with triangular markers.

The following insights were derived from the analysis of primal suboptimality, as shown in Figure 1.

(i) First, employing a constant η leads to a sequence that converges to an approximate solution while diminishing η_k allows for asymptotic guarantees to a true solution.

(ii) Second, choosing a very small η may impede early progress of the scheme since this leads to a large Lipschitz constant L, constraining the steplength and limiting the progress. On the other hand, selecting a larger η allows for better early progress but the sequence will converge to a solution that may differ significantly from the true solution. A diminishing η_k sequence starts with a larger η (allowing for larger steps and greater progress) but comes with a guarantee that the sequence will converge to a true solution. This is reflected in Figure 1.

(iii) We observe that the complexity guarantees for constant η are close to those for diminishing η_k with a slight improvement in the constant ρ_0 regime (see Theorem X.). We recall that Nesterov [33] and Beck and Teboulle [7] adopted different smoothing techniques with fixed η to get an $\boldsymbol{\varepsilon}$-optimal solution within $\mathcal{O}(1/\boldsymbol{\varepsilon})$. When compared to these smoothing schemes in [7, 33], **Sm-AL** targets problems with *nonsmooth constraint functions*. Moreover, **Sm-AL** accommodates both fixed and varying η, with an effective complexity rate $\tilde{\mathcal{O}}(\boldsymbol{\varepsilon}^{-3/2})$, matching the complexity of a smoothed penalized scheme [3]. When compared to the results in Proposition 2 with constant ρ, **SM-AL** with constant η improves overall complexity by $\mathcal{O}\left(\boldsymbol{\varepsilon}^{-1/2}\right)$. The diminishing nature of η_k slows down the convergence process due to the additional summable requirement for varying η_k.

5.2 Incorporation of termination criteria

Next, we consider the introduction of termination criteria **T1** and **T2** and examine the impact of potentially early termination, measured by $\sum_k N_k$. Table 5 provides a comparison between

Table 4 Numerical results

n	parameters			Sm-AL			N-AL		
	ρ_k	η_k	$\tilde{C}^{\dagger}$	$\bar{f}-f^*$	$d_-(\bar{g})$	Time(s)	$\bar{f}-f^*$	$d_-(\bar{g})$	Time(s)
5	0.1	$k^{-2.01}$	1e+0	4.35e−5	3.84e−4	8.00e−1	3.05e−4	0.00e+0	1.02e+3
	1.01^k	$\frac{1}{\rho_k k^{2.01}}$	5e+2	3.37e−4	0.00e+0	1.68e+0	1.36e−4	0.00e+0	3.52e+3
10	0.1	$k^{-2.01}$	1e+0	2.99e−5	8.12e−4	1.03e+0	2.92e−5	1.40e−3	3.70e+3
	1.01^k	$\frac{1}{\rho_k k^{2.01}}$	5e+2	3.13e−5	2.46e−4	1.79e+0	3.10e−5	0.00e+0	1.05e+4
20	0.1	$k^{-2.01}$	1e+1	3.50e−5	0.00e+0	4.59e+0	3.49e−5	0.00e+0	1.70e+4
	1.01^k	$\frac{1}{\rho_k k^{2.01}}$	8e+2	3.49e−5	0.00e+0	7.05e+0	3.49e−5	0.00e+0	6.36e+4
100	0.1	$k^{-2.01}$	6e+1	6.10e−6	0.00e+0	3.60e+1	5.82e+2	0.00e+0	> 7.2e+3
	1.01^k	$\frac{1}{\rho_k k^{2.01}}$	1e+3	6.21e−6	1.90e−4	7.00e+1	3.40e+3	0.00e+0	> 7.2e+3
200	0.1	$k^{-2.01}$	1e+2	3.71e−5	0.00e+0	8.40e+1	2.44e+3	0.00e+0	> 7.2e+3
	1.01^k	$\frac{1}{\rho_k k^{2.01}}$	1e+3	3.56e−5	0.00e+0	2.19e+2	2.32e+4	0.00e+0	> 7.2e+3
1000	0.1	$k^{-2.01}$	1e+5	-4.34e−5	0.00e+0	8.48e+2	9.41e+3	0.00e+0	> 7.2e+3
	1.01^k	$\frac{1}{\rho_k k^{2.01}}$	1e+4	-4.93e−5	0.00e+0	1.22e+3	4.75e+3	0.00e+0	> 7.2e+3

†: the subproblem $\mathcal{L}_{\rho_k,\eta_k}(\mathbf{x},\lambda)$ is $\left(\tilde{C}\rho_k/\eta_k\right)$-smooth

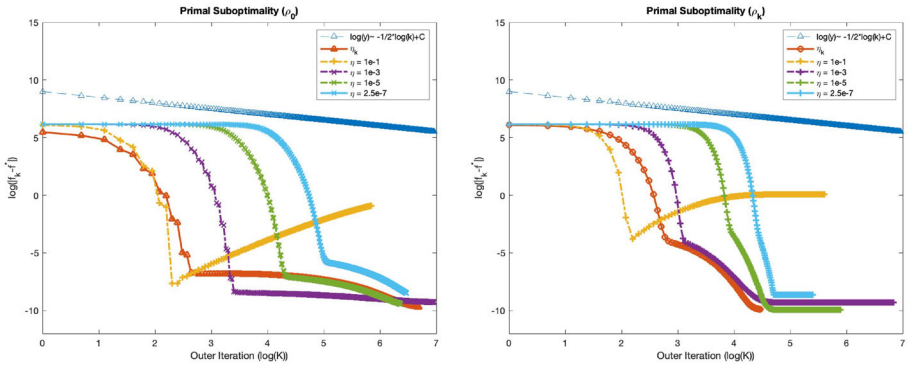

Fig. 1 Primal subopt. for fused lasso problems for constant (L) and increasing ρ_k (R)

the **Sm-AL** scheme with and without termination criteria. It can be observed that the incorporation of these termination criteria leads to significant computational benefits with little (if any) impact on accuracy. A natural question lies in the choice γ in the definition of the residual function $F_{\mathcal{X}}^{\text{nat},\tilde{\epsilon}}$. We observe that when $\mathbf{x} \in \mathcal{X}$,

$$\begin{aligned}\left\| F_{\mathcal{X}}^{\text{nat},\tilde{\epsilon}}(\mathbf{x}, u) \right\| &= \left\| \tilde{\epsilon}\left(\mathbf{u} - \Pi_X\left[\mathbf{x} - \gamma\nabla h(\mathbf{x})\right]\right) - \mathbf{x} + \Pi_{\mathcal{X}}\left[\mathbf{x} - \gamma\nabla h(\mathbf{x})\right] \right\| \\ &\leq \left\| \tilde{\epsilon}\left(\Pi_{\mathcal{X}}[\mathbf{u}] - \Pi_X\left[\mathbf{x} - \gamma\nabla h(\mathbf{x})\right]\right) \right\| + \left\| \Pi_{\mathcal{X}}[\mathbf{x}] - \Pi_{\mathcal{X}}\left[\mathbf{x} - \gamma\nabla h(\mathbf{x})\right] \right\| \\ &\leq \|\tilde{\epsilon}\left(\mathbf{u} - \mathbf{x} + \gamma\nabla h(\mathbf{x})\right)\| + \|\gamma\nabla h(\mathbf{x})\| \\ &\leq 2\tilde{\epsilon}C + \gamma(1+\tilde{\epsilon})D,\end{aligned}$$

where $\|\mathbf{x}\| \leq C$ and $\|\nabla h(\mathbf{x})\| \leq D$ for any $\mathbf{x}, u \in \mathcal{X}$. From the above bound, it may be observed that small choices of γ may lead to early satisfaction of conditions **T1** or **T2** while larger choices of γ may require significantly more iterations. Ideally, since we have already developed convergence guarantees, it would be helpful to relate γ to η_k. Some preliminary numerics are provided where the choice of γ is varied in condition **T2**, leading to some variability in performance. It can be surmised from this table that constant γ leads to poorer performance while diminishing choices for γ lead to far superior behavior. This is less surprising in that for larger values of K, γ is smaller and imposes a more modest threshold for satisfying the condition and thereby allowing for earlier termination.

6 Conclusion

In this paper, we develop a smoothed AL scheme for resolving convex programs with possibly nonsmooth constraints and provide rate and complexity guarantees for convex and strongly convex settings under constant and increasing penalty parameter sequences. The complexity guarantees represent significant improvements over the best available guarantees for AL schemes applied to convex programs with nonsmooth objectives and constraints. A by-product of our analysis develops a relationship between saddle-points of η-smoothed problems and η-saddle points of our original problem. Moreover, to improve the practical behavior of the proposed **Sm-AL** scheme, we have developed termination criteria that allow for premature termination. Our preliminary numerics suggest that such criteria lead to significant improvements in the complexity of our scheme with modest impacts on accuracy

Table 5 Numerical results with termination criteria

n	parameters				Sm-AL			
	Term. Cri.	ρ_k	η_k	$\tilde{C}^\dagger$	$\bar{f} - f^*$	$d_-(\bar{g})$	K	$\sum_k N_k$
5	T_1 & T_2	0.1	$k^{-2.01}$	3e+3	6.16e−5	0.00e+0	807	3.74e+5
		1.01^k	$\frac{1}{\rho_k k^{2.01}}$	3e+3	5.06e−5	0.00e+0	87	2.33e+5
	K & N_k	0.1	$k^{-2.01}$	3e+3	1.82e−4	0.00e+0	400	2.40e+7
		1.01^k	$\frac{1}{\rho_k k^{2.01}}$	3e+3	4.82e−5	0.00e+0	100	1.18e+6
10	T_1 & T_2	0.1	$k^{-2.01}$	3e+3	4.12e−5	0.00e+0	211	2.27e+5
		1.01^k	$\frac{1}{\rho_k k^{2.01}}$	1e+4	6.95e−5	0.00e+0	44	3.33e+4
	K & N_k	0.1	$k^{-2.01}$	3e+3	3.24e−5	1.68e−4	300	1.01e+7
		1.01^k	$\frac{1}{\rho_k k^{2.01}}$	3e+3	3.23e−5	1.01e−4	100	1.18e+6
100	T_1 & T_2	0.1	$k^{-2.01}$	5e+4	6.14e−6	2.49e−4	145	1.18e+4
		1.01^k	$\frac{1}{\rho_k k^{2.01}}$	1e+4	6.14e−6	2.49e−4	69	1.87e+4
	K & N_k	0.1	$k^{-2.01}$	5e+4	6.10e−6	2.97e−4	100	3.69e+5
		1.01^k	$\frac{1}{\rho_k k^{2.01}}$	1e+4	7.11e−4	0.00e+0	50	8.21e+4
200	T_1 & T_2	0.1	$k^{-2.01}$	5e+4	3.60e−5	0.00e+0	36	7.88e+3
		1.01^k	$\frac{1}{\rho_k k^{2.01}}$	1e+4	5.11e−5	0.00e+0	31	3.75e+3
	K & N_k	0.1	$k^{-2.01}$	5e+4	3.56e−5	0.00e+0	100	3.67e+5
		1.01^k	$\frac{1}{\rho_k k^{2.01}}$	1e+4	3.56e−5	0.00e+0	50	4.61e+4
500	T_1 & T_2	0.1	$k^{-2.01}$	5e+4	4.12e−5	0.00e+0	64	2.62e+4
		1.01^k	$\frac{1}{\rho_k k^{2.01}}$	1e+5	5.84e−5	0.00e+0	58	3.84e+4
	K & N_k	0.1	$k^{-2.01}$	5e+4	4.70e−3	0.00e+0	100	3.69e+5
		1.01^k	$\frac{1}{\rho_k k^{2.01}}$	1e+5	2.31e−5	0.00e+0	80	1.89e+5
5	T_1 & T_2	0.1	1e−3	3e+3	9.47e−5	0.00e+0	1057	1.36e+5
		1.01^k	1e−3	3e+3	9.03e−5	0.00e+0	932	2.88e+6
		0.1	1e−5	3e+3	8.15e−5	0.00e+0	580	3.66e+5
		1.01^k	1e−5	3e+3	4.79e−5	0.00e+0	361	5.05e+5
		0.1	2.5e−7	3e+3	2.06e−4	0.00e+0	674	7.35e+5
		1.01^k	2.5e−7	3e+3	1.81e−4	0.00e+0	222	1.14e+6

†: the subproblem $\mathcal{L}_{\rho_k,\eta_k}(\mathbf{x},\lambda)$ is $\left(\tilde{C}\rho_k/\eta_k\right)$-smooth

of the resulting solutions.. We believe that our findings represent a foundation for considering extensions to compositional regimes with expectation-valued and possibly nonsmooth constraints.

Table 6 Performance vs choice of γ

γ	Primal suboptimality	Primal Infeasibility	K
0.1	5.12e-5	0	10000
0.01	5.12e-5	0	10000
1	5.12e-5	0	10000
$\frac{\eta_k}{\rho_k}$	6.16e-5	0	807
η_k	1.72e-4	0	416
$0.1\eta_k$	5.41e-4	0	183

Appendix

Proof of Lemma 1

Proof

$$
\begin{aligned}
\mathcal{L}_\rho(\mathbf{x},\lambda) &\triangleq \min_{\mathbf{v}\geq 0}\left\{f(\mathbf{x})+\lambda^T(g(\mathbf{x})+\mathbf{v})+\tfrac{\rho}{2}\|g(\mathbf{x})+\mathbf{v}\|^2\right\}\\
&= \min_{\mathbf{v}\geq 0}\left\{f(\mathbf{x})+\tfrac{1}{2\rho}\|\lambda\|^2+\lambda^T(g(\mathbf{x})+\mathbf{v})+\tfrac{\rho}{2}\|g(\mathbf{x})+\mathbf{v}\|^2\right\}-\tfrac{1}{2\rho}\|\lambda\|^2\\
&= \min_{\mathbf{v}\geq 0}\left\{f(\mathbf{x})+\tfrac{1}{2}\|\tfrac{1}{\sqrt{\rho}}\lambda+\sqrt{\rho}(g(\mathbf{x})+\mathbf{v})\|^2\right\}-\tfrac{1}{2\rho}\|\lambda\|^2\\
&= \min_{\mathbf{v}\geq 0}\left\{f(\mathbf{x})+\tfrac{1}{2}\|\tfrac{1}{\sqrt{\rho}}\lambda+\sqrt{\rho}g(\mathbf{x})+\sqrt{\rho}\mathbf{v}\|^2\right\}-\tfrac{1}{2\rho}\|\lambda\|^2\\
&= \left\{f(\mathbf{x})+\min_{\mathbf{v}\geq 0}\tfrac{1}{2\rho}\|\lambda+\rho g(\mathbf{x})+\rho\mathbf{v}\|^2\right\}-\tfrac{1}{2\rho}\|\lambda\|^2\\
&= \left\{f(\mathbf{x})+\min_{\mathbf{v}\geq 0}\tfrac{\rho}{2}\|\tfrac{\lambda}{\rho}+g(\mathbf{x})+\mathbf{v}\|^2\right\}-\tfrac{1}{2\rho}\|\lambda\|^2\\
&= \left\{f(\mathbf{x})+\tfrac{\rho}{2}\left(d_+\left(-(\tfrac{\lambda}{\rho}+g(\mathbf{x}))\right)\right)^2-\tfrac{1}{2\rho}\|\lambda\|^2\right\}\\
&= \left\{f(\mathbf{x})+\tfrac{\rho}{2}\left(d_-\left((\tfrac{\lambda}{\rho}+g(\mathbf{x}))\right)\right)^2-\tfrac{1}{2\rho}\|\lambda\|^2\right\},
\end{aligned}
$$

where the last equality follows from $d_+(-v)=d_-(v)$ and $d_{\mathcal{K}}(u)\triangleq\min_{v\in\mathcal{K}}\|v-u\|$. We now derive $\nabla_\lambda\mathcal{L}_\rho(\mathbf{x},\lambda)$ as follows.

$$
\begin{aligned}
\nabla_\lambda\mathcal{L}_\rho(\mathbf{x},\lambda) &= \nabla_\lambda\left[\tfrac{\rho}{2}\left(d_-\left((\tfrac{\lambda}{\rho}+g(\mathbf{x})\right)\right)^2-\tfrac{1}{2\rho}\|\lambda\|^2\right]\\
&= \left(\tfrac{\lambda}{\rho}+g(\mathbf{x})-\Pi_-\left(\tfrac{\lambda}{\rho}+g(\mathbf{x})\right)\right)-\tfrac{\lambda}{\rho}\\
&= \left(g(\mathbf{x})-\Pi_-\left(\tfrac{\lambda}{\rho}+g(\mathbf{x})\right)\right)\\
&= \left(-\tfrac{\lambda}{\rho}+\Pi_+\left(\tfrac{\lambda}{\rho}+g(\mathbf{x})\right)\right),
\end{aligned}
$$

where the second equality is a result of $\nabla_u d^2_{\mathcal{K}}(u)=2(u-\Pi_{\mathcal{K}}[u])$ for any cone $\mathcal{K}$, the last equality is a consequence of $u=\Pi_{-\mathcal{K}}(u)+\Pi_{\mathcal{K}^*}(u)$ and $\mathcal{K}\triangleq\{u:u\geq 0\}$. Similarly, we

derive $\nabla_{\mathbf{x}}\mathcal{L}_\rho(\mathbf{x},\lambda)$ as follows.

$$\begin{aligned}\nabla_{\mathbf{x}}\mathcal{L}_\rho(\mathbf{x},\lambda) &= \nabla_{\mathbf{x}} f(\mathbf{x}) + \nabla_{\mathbf{x}}\left[\frac{\rho}{2}\left(d_-\left(\left(\frac{\lambda}{\rho}+g(\mathbf{x})\right)\right)^2 - \frac{1}{2\rho}\|\lambda\|^2\right]\right.\\ &= \nabla_{\mathbf{x}} f(\mathbf{x}) + \rho J_g(\mathbf{x})\left(\frac{\lambda}{\rho}+g(\mathbf{x}) - \Pi_-\left(\frac{\lambda}{\rho}+g(\mathbf{x})\right)\right),\end{aligned}$$

where $J_g(\mathbf{x})$ is Jacobian matrix of g. □

Proof of Lemma 2

Proof For completeness, we provide this proof which is based on that provided in [38]. Let $u = g(\mathbf{x}) + v$ and $p_\rho(u) \triangleq \inf_{\mathbf{x}\in\mathcal{X}} f(\mathbf{x}) + \frac{\rho}{2}\|u\|^2$, where p_ρ can be regarded as a "permutation" function. Then we have

$$p_\rho = p_0 + \rho q, \quad \text{where } q(u) = \tfrac{1}{2}\|u\|^2.$$

The augmented dual function can be expressed as

$$\begin{aligned}\mathcal{D}_\rho(\lambda) &= \inf_{u\in\mathbb{R}^m}\left\{p_\rho(u) + u^\top\lambda\right\} = -p_\rho^*(-\lambda) = -(p_0+\rho q)^*(-\lambda) = -(p_0^*\square q^*\rho)(-\lambda)\\ &= -\min_{u\in\mathbb{R}^m}\left\{p_0^*(-\lambda) + \rho q^*\left(\tfrac{\lambda-u}{\rho}\right)\right\}\\ &= \max_{u\in\mathbb{R}^m}\left\{-p_0^*(-\lambda) - \rho\left(\tfrac{\lambda-u}{\rho}\right)^2\right\}\\ &= \max_{u\in\mathbb{R}^m}\{\mathcal{D}_0(u) - \tfrac{1}{2\rho}\|u-\lambda\|^2\}\end{aligned}$$

where the infimal convolution of two functions is defined as

$$f\square g(x) = \inf\{f(x-y) + g(y) | y\in\mathbb{R}^n\}.$$

$$\mathcal{D}_\rho(\lambda) = \max_{u\in\mathbb{R}^m}\{\mathcal{D}_0(\lambda) - \tfrac{1}{2\rho}\|u-\lambda\|^2\}$$

Consequently, by Danskin's theorem,

$$\begin{aligned}\nabla_\lambda\mathcal{D}_\rho(\lambda) &= \frac{1}{\rho}\left(\left(\operatorname*{argmax}_u(\mathcal{D}_0(u) - \tfrac{1}{\rho}\|u-\lambda\|^2\right) - \lambda\right)\\ &= \tfrac{1}{\rho}(q_\rho(\lambda) - \lambda).\end{aligned}$$

□

Proof of Lemma 6

Proof We observe that (14) can be expressed as

$$\begin{aligned}\lambda_{k+1} &= \lambda_k + \rho_k\nabla_\lambda\mathcal{L}_{\eta_k,\rho_k}(\mathbf{x}_{k+1},\lambda_k)\\ &\overset{\text{Lemma 1}}{=} \lambda_k + \rho_k\left(-\tfrac{\lambda_k}{\rho_k} + \Pi_+\left[\tfrac{\lambda_k}{\rho_k} + g_{\eta_k}(\mathbf{x}_{k+1})\right]\right)\\ &= \Pi_+\left[\lambda_k + \rho_k g_{\eta_k}(\mathbf{x}_{k+1})\right].\end{aligned} \tag{34}$$

□

Proof of Lemma 4

Proof (i) Since for any $\mathbf{x} \in \mathcal{X}$, we have that

$$|f(\mathbf{x}) - f_\eta(\mathbf{x})| \le \eta\beta \tag{35}$$

$$|g_i(\mathbf{x}) - g_{i,\eta}(\mathbf{x})| \le \eta\beta, \quad i = 1, 2, \ldots, m. \tag{36}$$

Consequently, for any $\lambda \ge 0$, by adding (35) to $\lambda_i \times$ (36) for $i = 1, \cdots, m$,

$$\left|\mathcal{L}_0(\mathbf{x}, \lambda) - \mathcal{L}_{\eta,0}(\mathbf{x}, \lambda)\right| \le \eta(\|\lambda\|m + 1)\beta.$$

(ii) Suppose $\bar{\mathbf{x}} \in \arg\min_{\mathbf{x}\in\mathcal{X}} \mathcal{L}_0(\mathbf{x}, \lambda)$ and $\bar{\mathbf{x}}_\eta \in \arg\min_{\mathbf{x}\in\mathcal{X}} \mathcal{L}_{\eta,0}(\mathbf{x}, \lambda)$. It follows that $\mathcal{D}_0(\lambda) = \mathcal{L}_0(\bar{\mathbf{x}}, \lambda)$ and $\mathcal{D}_{\eta,0}(\lambda) = \mathcal{L}_{\eta,0}(\bar{\mathbf{x}}_\eta, \lambda)$. Let $C = (\|\lambda\|m + 1)\beta$.

$$\mathcal{D}_0(\lambda) = \mathcal{L}_0(\bar{\mathbf{x}}, \lambda) \le \mathcal{L}_0(\bar{\mathbf{x}}_\eta, \lambda) \le \mathcal{L}_{\eta,0}(\bar{\mathbf{x}}_\eta, \lambda) + \eta C = \mathcal{D}_{\eta,0}(\lambda) + \eta C.$$

Similarly, we have that

$$\mathcal{D}_{\eta,0}(\lambda) = \mathcal{L}_{\eta,0}(\bar{\mathbf{x}}_\eta, \lambda) \le \mathcal{L}_{\eta,0}(\bar{\mathbf{x}}, \lambda) \le \mathcal{L}_0(\bar{\mathbf{x}}, \lambda) + \eta C = \mathcal{D}_0(\lambda) + \eta C.$$

This implies that for any $\lambda \in \mathbb{R}^m_+$, $|\mathcal{D}_{\eta,0}(\lambda) - \mathcal{D}_0(\lambda)| \le \eta C$.
(iii) By the prior definitions,

$$\mathcal{D}_{\eta,\rho}(\lambda) = \max_{u\in\mathbb{R}^m} \left[\mathcal{D}_{\eta,0}(u) - \frac{1}{2\rho}\|u - \lambda\|^2 \right] \text{and}$$

$$\mathcal{D}_\rho(\lambda) = \max_{u\in\mathbb{R}^m} \left[\mathcal{D}_0(u) - \frac{1}{2\rho}\|u - \lambda\|^2 \right].$$

For any $\lambda \ge 0$, let $u_1 \in \arg\max_u \mathcal{D}_{\eta,\rho}(\lambda)$ and $u_2 \in \arg\max_u \mathcal{D}_\rho(\lambda)$. Then

$$\begin{aligned}
\mathcal{D}_{\eta,\rho}(\lambda) - \mathcal{D}_\rho(\lambda) &= \max_{u\in\mathbb{R}^m} \left[\mathcal{D}_{\eta,0}(u) - \frac{1}{2\rho}\|u - \lambda\|^2 \right] - \max_{u\in\mathbb{R}^m} \left[\mathcal{D}_0(u) - \frac{1}{2\rho}\|u - \lambda\|^2 \right] \\
&= \max_{u\in\mathbb{R}^m} \left[\mathcal{D}_{\eta,0}(u) - \frac{1}{2\rho}\|u - \lambda\|^2 \right] - \left[\mathcal{D}_0(u_2) - \frac{1}{2\rho}\|u_2 - \lambda\|^2 \right] \\
&\le \max_{u\in\mathbb{R}^m} \left[\mathcal{D}_{\eta,0}(u) - \frac{1}{2\rho}\|u - \lambda\|^2 \right] - \left[\mathcal{D}_0(u_1) - \frac{1}{2\rho}\|u_1 - \lambda\|^2 \right] \\
&\le \left| \mathcal{D}_{\eta,0}(u_1) - \mathcal{D}_0(u_1) \right| \overset{\text{Lemma 4}(ii)}{\le} \eta C.
\end{aligned}$$

Similarly, $\mathcal{D}_\rho(\lambda) - \mathcal{D}_{\eta,\rho}(\lambda) \le \eta C$, implying the result. □

Proof of Lemma 5

Proof (i) By definition, we have that

$$\begin{aligned}
q_\rho(\lambda) &= \arg\max_{u\in\mathbb{R}^m} \left(\mathcal{D}_0(u) - \tfrac{1}{2\rho}\|u - \lambda\|^2 \right) \\
&= \arg\min_{u\in\mathbb{R}^m} \left(-\mathcal{D}_0(u) + \tfrac{1}{2\rho}\|u - \lambda\|^2 \right) = \text{prox}_{-\mathcal{D}_0,\rho}(\lambda). \qquad (37)
\end{aligned}$$

Similarly, $q_{\eta,\rho}(\lambda) = \text{prox}_{-D_{\eta,0},\rho}(\lambda)$. (38)

By strong convexity of $-\mathcal{D}_0(\bullet) + \frac{1}{2\rho}\|\bullet - \lambda\|^2$ and $-\mathcal{D}_{\eta,0}(\bullet) + \frac{1}{2\rho}\|\bullet - \lambda\|^2$ and by noting that $q_\rho(\lambda)$ and $q_{\eta,\rho}(\lambda)$ uniquely minimize (37) and (38), respectively, we obtain that

$$\begin{aligned} -\mathcal{D}_0(q_{\eta,\rho}(\lambda)) + \tfrac{1}{2\rho}\|q_{\eta,\rho}(\lambda) - \lambda\|^2 &\geq -\mathcal{D}_0(q_\rho(\lambda)) + \tfrac{1}{2\rho}\|q_\rho(\lambda) - \lambda\|^2 \\ &\quad + \tfrac{1}{4\rho}\|q_{\eta,\rho}(\lambda) - q_\rho(\lambda)\|^2, \\ -\mathcal{D}_{\eta,0}(q_\rho(\lambda)) + \tfrac{1}{2\rho}\|q_\rho(\lambda) - \lambda\|^2 &\geq -\mathcal{D}_{\eta,0}(q_{\eta,\rho}(\lambda)) + \tfrac{1}{2\rho}\|q_{\eta,\rho}(\lambda) - \lambda\|^2 \\ &\quad + \tfrac{1}{4\rho}\|q_{\eta,\rho}(\lambda) - q_\rho(\lambda)\|^2. \end{aligned}$$

Consequently, by summing the two inequalities above, we have that

$$\begin{aligned} \tfrac{1}{2\rho}\|q_{\eta,\rho}(\lambda) - q_\rho(\lambda)\|^2 &\leq \mathcal{D}_{\eta,0}(q_{\eta,\rho}(\lambda)) - \mathcal{D}_0(q_{\eta,\rho}(\lambda)) + \mathcal{D}_0(q_\rho(\lambda)) - \mathcal{D}_{\eta,0}(q_\rho(\lambda)) \\ &\leq \eta\left(\|q_{\eta,\rho}(\lambda)\|m + 1\right)\beta + \eta\left(\|q_\rho(\lambda)\|m + 1\right)\beta. \end{aligned}$$

By definitions of λ_η^* and λ^*, we have $q_{\eta,\rho}(\lambda_\eta^*) = \lambda_\eta^*$ and $q_\rho(\lambda^*) = \lambda^*$. Therefore, we have the following bounds on $\|q_{\eta,\rho}(\lambda)\|$ and $\|q_\rho(\lambda)\|$.

$$\begin{aligned} \|q_{\eta,\rho}(\lambda)\| = \|q_{\eta,\rho}(\lambda) - q_{\eta,\rho}(\lambda_\eta^*) + \lambda_\eta^*\| &\leq \underbrace{\|q_{\eta,\rho}(\lambda) - q_{\eta,\rho}(\lambda_\eta^*)\|}_{q_{\eta,\rho}(\bullet)\text{ is non-expansive}} + \|\lambda_\eta^*\| \\ &\leq \|\lambda - \lambda_\eta^*\| + \|\lambda_\eta^*\| \leq \|\lambda\| + 2\|\lambda_\eta^*\|. \end{aligned}$$

Similarly, $\|q_\rho(\lambda)\| = \left\|q_\rho(\lambda) - q_\rho(\lambda^*) + \lambda^*\right\| \leq \|\lambda\| + 2\|\lambda^*\|$. Therefore, It follows that for any $\lambda \geq 0$,

$$\begin{aligned} \tfrac{1}{2\rho}\|q_{\eta,\rho}(\lambda) - q_\rho(\lambda)\|^2 &\leq \eta\beta\left(2 + m\left(\|q_{\eta,\rho}(\lambda)\| + \|q_\rho(\lambda)\|\right)\right) \\ &\leq \eta\beta\left(2 + m\left(2\|\lambda\| + 2(b_{\lambda,\eta} + b_\lambda)\right)\right) \\ &= 2\eta\beta\left(C_m + m(\|\lambda\|)\right), \end{aligned}$$

where $C_m \triangleq 1 + m(b_{\lambda,\eta} + b_\lambda)$ is a constant.
(ii) By recalling the definitions of $\nabla_\lambda \mathcal{D}_\rho(\lambda)$ and $\nabla_\lambda \mathcal{D}_{\eta,\rho}(\lambda)$ from Lemma 2,

$$\|\nabla_\lambda \mathcal{D}_{\eta,\rho}(\lambda) - \nabla_\lambda \mathcal{D}_\rho(\lambda)\| = \tfrac{1}{\rho}\|q_{\eta,\rho}(\lambda) - q_\rho(\lambda)\| \leq \sqrt{\tfrac{4\eta(\|\lambda\|m + C_m)\beta}{\rho}}.$$

□

Proof of Lemma 7

Proof (a) $\implies$ (b). Suppose $\mathbf{x}_\epsilon^*$ is an ϵ-optimal solution of (Opt). Suppose (b) does not hold and there exists $\mathbf{y} \in \mathcal{X}$ such that

$$\nabla h(\mathbf{x}_\epsilon^*)^\top\left(\mathbf{y} - \mathbf{x}_\epsilon^*\right) < -\epsilon.$$

Consequently, $h'(\mathbf{x}_\epsilon^*; d) = \nabla h(\mathbf{x}_\epsilon^*)^\top d$ where $d = \mathbf{y} - \mathbf{x}_\epsilon^*$. Since d is a descent direction, by Lemma 4.2 [5], we have that for some $\delta \leq 1$, $h(\mathbf{x}_\epsilon^* + td) - h(\mathbf{x}_\epsilon^*) < -\epsilon$ for any $t \in (0, \delta)$. Note that $\mathbf{x}_\epsilon^* + td \in \mathcal{X}$ since $\mathcal{X}$ is a convex set. It follows that there exists a feasible point $\mathbf{x}_\epsilon^* + td \in \mathcal{X}$ such that $h(\mathbf{x}_\epsilon^* + td) - h(\mathbf{x}_\epsilon^*) < -\epsilon$, violating ϵ-optimality of $\mathbf{x}_\epsilon^*$.
(b) $\implies$ (a). By convexity of h, we have that

$$\begin{aligned} h(\mathbf{y}) &\geq h(\mathbf{x}_\epsilon^*) + \nabla h(\mathbf{x}_\epsilon^*)^\top\left(\mathbf{y} - \mathbf{x}_\epsilon^*\right), \qquad \forall \mathbf{y} \in \mathcal{X} \\ &\geq h(\mathbf{x}_\epsilon^*) - \epsilon, \qquad\qquad\qquad\quad \forall \mathbf{y} \in \mathcal{X}. \end{aligned} \tag{39}$$

Consequently, $h(\mathbf{x}_\epsilon^*) \leq h(\mathbf{x}^*) + \epsilon \leq h(\mathbf{x}) + \epsilon$ for any $x \in \mathcal{X}$, implying that $\mathbf{x}_\epsilon^*$ is an ϵ-optimal solution.

(c) $\implies$ (b). Given $\mathbf{u} \in \mathcal{X}$ and $\mathbf{x}_\epsilon^* \in \mathcal{X}$, we have that $F_{\mathcal{X}}^{\text{nat},\tilde{\epsilon}}(\mathbf{x}_\epsilon^*, \mathbf{u}) = 0$. Consequently, we have that $\mathbf{x}_\epsilon^* = \tilde{\epsilon} v + \Pi_{\mathcal{X}}\left[\mathbf{x}_\epsilon^* - \gamma \nabla h(\mathbf{x}_\epsilon^*)\right]$, where $\tilde{\epsilon} v + \Pi_{\mathcal{X}}\left[\mathbf{x}_\epsilon^* - \gamma \nabla h(\mathbf{x}_\epsilon^*)\right] \in \mathcal{X}$ and $v = \mathbf{u} - \Pi_{\mathcal{X}}\left[\mathbf{x}_\epsilon^* - \gamma \nabla h(\mathbf{x}_\epsilon^*)\right]$. It is easily seen that the former of these assertions holds as observed next.

$$\tilde{\epsilon} v + \Pi_{\mathcal{X}}\left[\mathbf{x}_\epsilon^* - \gamma \nabla h(\mathbf{x}_\epsilon^*)\right] = \tilde{\epsilon}\mathbf{u} + (1-\tilde{\epsilon})\Pi_{\mathcal{X}}\left[\mathbf{x}_\epsilon^* - \gamma \nabla h(\mathbf{x}_\epsilon^*)\right] \in \mathcal{X},$$

since $\mathbf{u} \in \mathcal{X}$, $\tilde{\epsilon} \in (0,1)$, and $\mathcal{X}$ is a convex set. For ease of exposition, we denote $\Pi_{\mathcal{X}}\left[\mathbf{x}_\epsilon^* - \gamma \nabla h(\mathbf{x}_\epsilon^*)\right]$ as $\tilde{\mathbf{x}}$. For any $\mathbf{y} \in \mathcal{X}$,

$$\begin{aligned}
\left(\mathbf{y} - \mathbf{x}_\epsilon^*\right)^\top \nabla h(\mathbf{x}_\epsilon^*) &= \tfrac{1}{\gamma}\left(\mathbf{y} - \mathbf{x}_\epsilon^*\right)^\top \left(\mathbf{x}_\epsilon^* - (\mathbf{x}_\epsilon^* - \gamma \nabla h(\mathbf{x}_\epsilon^*))\right) \\
&= \underbrace{\tfrac{1}{\gamma}\left(\mathbf{y} - \Pi_{\mathcal{X}}\left[\mathbf{x}_\epsilon^* - \gamma \nabla h(\mathbf{x}_\epsilon^*)\right]\right)^\top \left(\tilde{\epsilon} v + \Pi_{\mathcal{X}}\left[\mathbf{x}_\epsilon^* - \gamma \nabla h(\mathbf{x}_\epsilon^*)\right] - (\mathbf{x}_\epsilon^* - \gamma \nabla h(\mathbf{x}_\epsilon^*))\right)}_{\text{Term 1}} \\
&\quad \underbrace{- \tfrac{1}{\gamma}\tilde{\epsilon} v^\top \left(\tilde{\epsilon} v + \Pi_{\mathcal{X}}\left[\mathbf{x}_\epsilon^* - \gamma \nabla h(\mathbf{x}_\epsilon^*)\right] - (\mathbf{x}_\epsilon^* - \gamma \nabla h(\mathbf{x}_\epsilon^*))\right)}_{\text{Term 2}}.
\end{aligned}$$

We first derive a bound on Term 1.

$$\begin{aligned}
\text{Term 1} &= \tfrac{1}{\gamma}\left(\mathbf{y} - \Pi_{\mathcal{X}}\left[\mathbf{x}_\epsilon^* - \gamma \nabla h(\mathbf{x}_\epsilon^*)\right]\right)^\top \left(\tilde{\epsilon} v + \Pi_{\mathcal{X}}\left[\mathbf{x}_\epsilon^* - \gamma \nabla h(\mathbf{x}_\epsilon^*)\right] - (\mathbf{x}_\epsilon^* - \gamma \nabla h(\mathbf{x}_\epsilon^*))\right) \\
&= \underbrace{\tfrac{1}{\gamma}\left(\mathbf{y} - \Pi_{\mathcal{X}}\left[\mathbf{x}_\epsilon^* - \gamma \nabla h(\mathbf{x}_\epsilon^*)\right]\right)^\top \left(\Pi_{\mathcal{X}}\left[\mathbf{x}_\epsilon^* - \gamma \nabla h(\mathbf{x}_\epsilon^*)\right] - (\mathbf{x}_\epsilon^* - \gamma \nabla h(\mathbf{x}_\epsilon^*))\right)}_{\geq 0 \quad \text{(Projection identity)}} \\
&\quad + \tfrac{1}{\gamma}\left(\mathbf{y} - \tilde{\mathbf{x}}\right)^\top \left(\tilde{\epsilon}(\mathbf{u} - \tilde{x})\right) \geq \tfrac{1}{\gamma}\left(\mathbf{y} - \mathbf{x}_\epsilon^*\right)^\top \left(\tilde{\epsilon}\left(\mathbf{u} - \tilde{\mathbf{x}}\right)\right) \\
&\geq -\tfrac{\tilde{\epsilon}}{2\gamma}\left(2\|\mathbf{y}\|^2 + 2\|\mathbf{x}_\epsilon^*\|^2 + 2\|\mathbf{u}\|^2 + 2\|\tilde{\mathbf{x}}\|^2\right) \geq -\tfrac{4}{\gamma}\tilde{\epsilon} C = -\epsilon_1,
\end{aligned} \tag{40}$$

where $\|\mathbf{y}\|^2 \leq C$ for any $\mathbf{y} \in \mathcal{X}$ and $\tilde{\mathbf{x}} \triangleq \Pi_{\mathcal{X}}\left[\mathbf{x}_\epsilon^* - \gamma \nabla h(\mathbf{x}_\epsilon^*)\right]$. Consider Term 2.

$$\begin{aligned}
\text{Term 2} &= -\tfrac{1}{\gamma}\tilde{\epsilon} v^\top \left(\tilde{\epsilon} v + \Pi_{\mathcal{X}}\left[\mathbf{x}_\epsilon^* - \gamma \nabla h(\mathbf{x}_\epsilon^*)\right] - (\mathbf{x}_\epsilon^* - \gamma \nabla h(\mathbf{x}_\epsilon^*))\right) \\
&= -\tfrac{1}{\gamma}\tilde{\epsilon}(\mathbf{u} - \tilde{\mathbf{x}})^\top \left(\tilde{\epsilon}(\mathbf{u} - \tilde{\mathbf{x}}) + \tilde{\mathbf{x}} - (\mathbf{x}_\epsilon^* - \gamma \nabla h(\mathbf{x}_\epsilon^*))\right) \\
&= -\tfrac{1}{\gamma}\tilde{\epsilon}(\mathbf{u} - \tilde{\mathbf{x}})^\top \left(\tilde{\epsilon}\mathbf{u} + (1-\tilde{\epsilon})\tilde{\mathbf{x}} - (\mathbf{x}_\epsilon^* - \gamma \nabla h(\mathbf{x}_\epsilon^*))\right) \\
&= -\tfrac{1}{\gamma}\tilde{\epsilon}\left(\tilde{\epsilon}\|\mathbf{u}\|^2 + (1-2\tilde{\epsilon})\mathbf{u}^\top \tilde{\mathbf{x}} - (1-\tilde{\epsilon})\|\tilde{\mathbf{x}}\|^2 - \mathbf{u}^\top \mathbf{x}_\epsilon^* + \tilde{\mathbf{x}}^\top \mathbf{x}_\epsilon^* + \gamma \mathbf{u}^\top \nabla h(\mathbf{x}_\epsilon^*) - \gamma \tilde{\mathbf{x}}^\top \nabla h(\mathbf{x}_\epsilon^*)\right) \\
&\geq -\tfrac{1}{\gamma}\tilde{\epsilon}\left(\tilde{\epsilon}\|\mathbf{u}\|^2 + \tfrac{(1-2\tilde{\epsilon})}{2}(\|\mathbf{u}\|^2 + \|\tilde{\mathbf{x}}\|^2) + \tfrac{1}{2}(\|\mathbf{u}\|^2 + \|\mathbf{x}_\epsilon^*\|^2) + \tfrac{1}{2}(\|\tilde{\mathbf{x}}\|^2 + \|\mathbf{x}_\epsilon^*\|^2)\right. \\
&\quad \left. + \tfrac{\gamma}{2}(\|\mathbf{u}\|^2 + \|\tilde{\mathbf{x}}\|^2 + 2\|\nabla h(\mathbf{x}_\epsilon^*)\|^2)\right) \\
&\geq -\tfrac{1}{\gamma}\tilde{\epsilon}\left((1 + \tfrac{\gamma}{2})\|\mathbf{u}\|^2 + ((1 - \tilde{\epsilon} + \tfrac{\gamma}{2})\|\tilde{\mathbf{x}}\|^2 + \|\mathbf{x}_\epsilon^*\|^2 + \gamma\|\nabla h(\mathbf{x}_\epsilon^*)\|^2)\right) \\
&\geq -\tfrac{1}{\gamma}\tilde{\epsilon}((3+\gamma)C + \gamma D) \triangleq -\epsilon_2.
\end{aligned}$$

Consequently, we have that

$$\begin{aligned}
\left(\mathbf{y} - \mathbf{x}_\epsilon^*\right)^\top \nabla h(\mathbf{x}_\epsilon^*) \geq -\epsilon, \text{ where } \epsilon = \epsilon_1 + \epsilon_2 = \tfrac{4}{\gamma}\tilde{\epsilon} C + \tfrac{1}{\gamma}\tilde{\epsilon}((3+\gamma)C + \gamma D) \\
\implies \tilde{\epsilon} = \tfrac{\gamma \epsilon}{7C + \gamma(C+D)}.
\end{aligned}$$

□

Proof of Lemma 9

Proof (a) By adding and subtracting $q_{\eta_k,\rho_k}(\lambda_k)$, $q_{\eta_k,\rho_k}(\lambda^*)$, $q_{\rho_k}(\lambda^*)$, it follows that

$$\begin{aligned}\left\|\lambda_{k+1}-\lambda^*\right\| \leq & \|\lambda_{k+1}-q_{\eta_k,\rho_k}(\lambda_k)\|+\|q_{\eta_k,\rho_k}(\lambda_k)-q_{\eta_k,\rho_k}(\lambda^*)\| \\ & +\|q_{\eta_k,\rho_k}(\lambda^*)-q_{\rho_k}(\lambda^*)\|+\underbrace{\|q_{\rho_k}(\lambda^*)-\lambda^*\|}_{=0}.\end{aligned}$$

Next, we derive a bound on $\|\lambda_{k+1}-q_{\eta_k,\rho_k}(\lambda_k)\|$ that

$$\begin{aligned}\|\lambda_{k+1}-q_{\eta_k,\rho_k}(\lambda_k)\| &= \left\|\lambda_k+\rho_k\left(\nabla_\lambda \mathcal{L}_{\eta_k,\rho_k}(\mathbf{x}_{k+1},\lambda_k)\right)-q_{\eta_k,\rho_k}(\lambda_k)\right\| \\ &= \left\|\lambda_k+\rho_k\left(\nabla_\lambda \mathcal{L}_{\eta_k,\rho_k}(\mathbf{x}_{k+1},\lambda_k)\right)-\rho_k\nabla_\lambda \mathcal{D}_{\eta_k,\rho_k}(\lambda_k)-\lambda_k\right\| \\ &\leq \rho_k\left\|\nabla_\lambda \mathcal{L}_{\eta_k,\rho_k}(\mathbf{x}_{k+1},\lambda_k))-\nabla_\lambda \mathcal{D}_{\eta_k,\rho_k}(\lambda_k)\right\| \overset{\text{Lem. 8}}{\leq} \sqrt{2\rho_k\epsilon_k\eta_k^b}.\end{aligned}$$

From Lemma 4, $\|q_{\eta_k,\rho_k}(\lambda^*)-q_{\rho_k}(\lambda^*)\| \leq 2\sqrt{\rho_k\eta_k(\|\lambda^*\|m+C_m)\beta}$, implying that

$$\|\lambda_{k+1}-\lambda^*\| \leq \sqrt{2\rho_k\epsilon_k\eta_k^b}+2\sqrt{\rho_k\eta_k(\|\lambda^*\|m+C_m)\beta}+\|\lambda_k-\lambda^*\|. \tag{41}$$

By leveraging the deterministic form of the Robbins-Siegmund Lemma [36], if $\sqrt{2\rho_k\epsilon_k\eta_k^b}+2\sqrt{\rho_k\eta_k(\|\lambda^*\|m+C_m)\beta}$ is summable, then $\{\|\lambda_k-\lambda^*\|\}$ converges to a non-negative value. It follows that $\{\lambda_k\}$ is convergent.
(b) Summing (41) from $k=0,\cdots,K-1$, we obtain that

$$\begin{aligned}\|\lambda_K-\lambda^*\| &\leq \sum_{k=0}^{K-1}\left(\sqrt{2\rho_k\epsilon_k\eta_k^b}+2\sqrt{\eta_k\rho_k(\|\lambda^*\|m+C_m)\beta}\right)+\|\lambda_0-\lambda^*\| \\ &\leq \sum_{k=0}^{\infty}\left(\sqrt{2\rho_k\epsilon_k\eta_k^b}+2\sqrt{\eta_k\rho_k(\|\lambda^*\|m+C_m)\beta}\right)+\|\lambda_0-\lambda^*\| \triangleq B_\lambda.\end{aligned}$$

□

Proof of Lemma 11

Proof Recall that $\mathcal{L}_{\eta,\rho}(\mathbf{x},\lambda)$ and its gradient $\nabla_{\mathbf{x}}\mathcal{L}_{\eta,\rho}(\mathbf{x},\lambda)$ are defined as

$$\begin{aligned}\mathcal{L}_{\eta,\rho}(\mathbf{x},\lambda) &= f_\eta(\mathbf{x})+\tfrac{\rho}{2}\left(d_-\left(\tfrac{\lambda}{\rho}+g_\eta(\mathbf{x})\right)\right)^2-\tfrac{1}{2\rho}\|\lambda\|^2 \\ \nabla_{\mathbf{x}}\mathcal{L}_{\eta,\rho}(\mathbf{x},\lambda) &= \nabla_{\mathbf{x}} f_\eta(\mathbf{x})+\rho\mathbf{J}_g(\mathbf{x})^\top\left(\tfrac{\lambda}{\rho}+g_\eta(\mathbf{x})-\Pi_-\left[\tfrac{\lambda}{\rho}+g_\eta(\mathbf{x})\right]\right),\end{aligned}$$

where $\left(\mathbf{J}_g(\mathbf{x})\right)^\top \triangleq \left[\nabla_{\mathbf{x}} g_{\eta,1}(\mathbf{x})\ \nabla_{\mathbf{x}} g_{\eta,2}(\mathbf{x})\ \ldots\ \nabla_{\mathbf{x}} g_{\eta,m}(\mathbf{x})\right]$ and $\mathbf{J}_g(\mathbf{x})$ denotes the Jacobian matrix of $g_\eta(\mathbf{x})$. By Assumption 1 and Definition 1, g_η and $\mathbf{J}_g$ are bounded on $\mathcal{X}$ by M_g and M_G, respectively. Since $\mathbf{J}_g$ is bounded, g_η is Lipschitz continuous on $\mathcal{X}$ with constant L_g. By Lemma 9, for all $\mathbf{x}_1,\mathbf{x}_2\in\mathcal{X}$, it follows that

$$\begin{aligned}\left\|\nabla_{\mathbf{x}}\mathcal{L}_{\eta,\rho}(\mathbf{x}_1,\lambda)-\nabla_{\mathbf{x}}\mathcal{L}_{\eta,\rho}(\mathbf{x}_2,\lambda)\right\| \leq & \left\|\nabla_{\mathbf{x}} f_\eta(\mathbf{x}_1)-\nabla_{\mathbf{x}} f_\eta(\mathbf{x}_2)\right\| \\ & +\rho\left\|\mathbf{J}_g(\mathbf{x}_1)^\top\left(\tfrac{\lambda}{\rho}+g_\eta(\mathbf{x}_1)-\Pi_-\left[\tfrac{\lambda}{\rho}+g_\eta(\mathbf{x}_1)\right]\right)\right. \\ & \left.-\mathbf{J}_g(\mathbf{x}_2)^\top\left(\tfrac{\lambda}{\rho}+g_\eta(\mathbf{x}_2)-\Pi_-\left[\tfrac{\lambda}{\rho}+g_\eta(\mathbf{x}_2)\right]\right)\right\|.\end{aligned}$$

Next we show that the second term is Lipschitz continuous in $\mathbf{x}$. By adding and subtracting $-\mathbf{J}_g(\mathbf{x}_2)^\top \left(\frac{\lambda}{\rho} + g_\eta(\mathbf{x}_1) - \Pi_- \left[\frac{\lambda}{\rho} + g_\eta(\mathbf{x}_1)\right]\right)$, we have

$$
\begin{aligned}
&\left\|\mathbf{J}_g(\mathbf{x}_1)^\top \left(\tfrac{\lambda}{\rho} + g_\eta(\mathbf{x}_1) - \Pi_- \left[\tfrac{\lambda}{\rho} + g_\eta(\mathbf{x}_1)\right]\right) - \mathbf{J}_g(\mathbf{x}_2)^\top \left(\tfrac{\lambda}{\rho} + g_\eta(\mathbf{x}_2) - \Pi_- \left[\tfrac{\lambda}{\rho} + g_\eta(\mathbf{x}_2)\right]\right)\right\| \\
&\le \left\|\mathbf{J}_g(\mathbf{x}_1)^\top \left(\tfrac{\lambda}{\rho} + g_\eta(\mathbf{x}_1) - \Pi_- \left[\tfrac{\lambda}{\rho} + g_\eta(\mathbf{x}_1)\right]\right) - \mathbf{J}_g(\mathbf{x}_2)^\top \left(\tfrac{\lambda}{\rho} + g_\eta(\mathbf{x}_1) - \Pi_- \left[\tfrac{\lambda}{\rho} + g_\eta(\mathbf{x}_1)\right]\right)\right\| \\
&+ \left\|\mathbf{J}_g(\mathbf{x}_2)^\top \left(\tfrac{\lambda}{\rho} + g_\eta(\mathbf{x}_1) - \Pi_- \left[\tfrac{\lambda}{\rho} + g_\eta(\mathbf{x}_1)\right]\right) - \mathbf{J}_g(\mathbf{x}_2)^\top \left(\tfrac{\lambda}{\rho} + g_\eta(\mathbf{x}_2) - \Pi_- \left[\tfrac{\lambda}{\rho} + g_\eta(\mathbf{x}_2)\right]\right)\right\| \\
&\le \left\|\mathbf{J}_g(\mathbf{x}_1) - \mathbf{J}_g(\mathbf{x}_2)\right\| \underbrace{\left\|\tfrac{\lambda}{\rho} + g_\eta(\mathbf{x}_1) - \Pi_- \left[\tfrac{\lambda}{\rho} + g_\eta(\mathbf{x}_1)\right]\right\|}_{= \left\|\Pi_+\left[\frac{\lambda}{\rho} + g_\eta(\mathbf{x}_1)\right]\right\|} \\
&+ \|\mathbf{J}_g(\mathbf{x}_2)\| \Bigg(\left\|g_\eta(\mathbf{x}_1) - g_\eta(\mathbf{x}_2)\right\| + \underbrace{\left\|\Pi_- \left[\tfrac{\lambda}{\rho} + g_\eta(\mathbf{x}_1)\right] - \Pi_- \left[\tfrac{\lambda}{\rho} + g_\eta(\mathbf{x}_2)\right]\right\|}_{\text{non-expansive}} \Bigg) \\
&\le \tfrac{m\alpha_g}{\eta} \|\mathbf{x}_1 - \mathbf{x}_2\| \left(\tfrac{b_\lambda}{\rho} + M_g\right) + M_G \left(2L_g \|\mathbf{x}_1 - \mathbf{x}_2\|\right).
\end{aligned}
$$

Consequently, $\mathcal{L}_{\eta,\rho}(\mathbf{x}, \lambda)$ is $(\frac{\tilde{C}\rho}{\eta})$-smooth by observing that

$$
\begin{aligned}
&\left\|\nabla_{\mathbf{x}} \mathcal{L}_{\eta,\rho}(\mathbf{x}_1, \lambda) - \nabla_{\mathbf{x}} \mathcal{L}_{\eta,\rho}(\mathbf{x}_2, \lambda)\right\| \le \tfrac{\alpha_f}{\eta} \|\mathbf{x}_1 - \mathbf{x}_2\| + \rho \left(\tfrac{m\alpha_g}{\eta} \left(\tfrac{b_\lambda}{\rho} + M_g\right) + 2M_G L_g\right) \\
&\times \|\mathbf{x}_1 - \mathbf{x}_2\| \le \tfrac{\tilde{C}\rho}{\eta} \|\mathbf{x}_1 - \mathbf{x}_2\|,
\end{aligned}
$$

where $\rho \ge 1$, $\eta \le \eta^u$, and

$$
\begin{aligned}
\tfrac{\alpha_f}{\eta} + \rho \left(\tfrac{m\alpha_g}{\eta} \left(\tfrac{b_\lambda}{\rho} + M_g\right) + 2M_G L_g\right) &= \tfrac{\alpha_f + m\alpha_g b_\lambda}{\eta} + \rho \left(\tfrac{m\alpha_g M_g}{\eta} + 2M_G L_g\right) \\
&\le \tfrac{(\alpha_f + m\alpha_g b_\lambda)\rho}{\eta} + \rho \left(\tfrac{m\alpha_g M_g}{\eta} + \tfrac{2M_G L_g \eta}{\eta}\right) \\
&\le \tfrac{\rho}{\eta} \left(\alpha_f + m\alpha_g (b_\lambda + M_g) + 2M_G L_g \eta^u\right) \\
&= \tfrac{\tilde{C}\rho}{\eta}.
\end{aligned}
$$

(b) This has been shown in [38, Th. 3.1]. □

Acknowledgements We extend our sincere appreciation to Dr. Qi Wang (University of Michigan at Ann Arbor) for her invaluable suggestions and careful reading of a recent draft of this paper. In addition, the second author would like to acknowledge his early collaboration with Dr. N. Serhat Aybat (Pennsylvania State University) that provided some of the seeds for this study.

Funding P. Zhang and Uday V. Shanbhag are partially supported by ONR Grant N00014-22-1-2589, AFOSR Grant FA9550-24-1-0259, and DOE Grant DE-SC0023303. Ethan X. Fang would like to acknowledge support from NSF Grants DMS-2346292 and DMS-2434666.

References

1. Alger, N., Villa, U., Bui-Thanh, T., Ghattas, O.: A data scalable augmented Lagrangian KKT preconditioner for large-scale inverse problems. SIAM J. Sci. Comput. **39**(5), A2365–A2393 (2017)
2. Aybat, N.S., Ahmadi, H., Shanbhag, U.V.: On the analysis of inexact augmented lagrangian schemes for misspecified conic convex programs. IEEE Transactions on Automatic Control **67**(8), 3981–3996 (2021)
3. Aybat, N.S., Iyengar, G.: A first-order smoothed penalty method for compressed sensing. SIAM Journal on Optimization **21**(1), 287–313 (2011)
4. Aybat, N.S., Iyengar, G.: An augmented Lagrangian method for conic convex programming. arXiv preprint arXiv:1302.6322 (2013)
5. Beck, A.: Introduction to nonlinear optimization: Theory, algorithms, and applications with MATLAB. SIAM (2014)
6. Beck, A.: First-order methods in optimization. SIAM (2017)
7. Beck, A., Teboulle, M.: Smoothing and first order methods: A unified framework. SIAM Journal on Optimization **22**(2), 557–580 (2012)
8. Byrd, R.H., Hribar, M.E., Nocedal, J.: An interior point algorithm for large-scale nonlinear programming. SIAM Journal on Optimization **9**(4), 877–900 (1999)
9. Chang, H., Lou, Y., Ng, M.K., Zeng, T.: Phase retrieval from incomplete magnitude information via total variation regularization. SIAM J. Sci. Comput. **38**(6), A3672–A3695 (2016)
10. Conn, A.R., Gould, G., Toint, P.L.: LANCELOT: a Fortran package for large-scale nonlinear optimization (Release A), vol. 17. Springer Science & Business Media (2013)
11. Cottle, R.W., Pang, J.S., Stone, R.E.: The Linear Complementarity Problem. Academic Press Inc, Boston, MA (1992)
12. Devolder, O., Glineur, F., Nesterov, Y.: Double smoothing technique for large-scale linearly constrained convex optimization. SIAM Journal on Optimization **22**(2), 702–727 (2012)
13. Dong, B., Zhang, Y.: An efficient algorithm for ℓ_0 minimization in wavelet frame based image restoration. J. Sci. Comput. **54**(2–3), 350–368 (2013)
14. Facchinei, F., Pang, J.S.: Finite-dimensional variational inequalities and complementarity problems, vol. I. Springer Series in Operations Research. Springer-Verlag, New York (2003)
15. Friedlander, M.P., Leyffer, S.: Global and finite termination of a two-phase augmented Lagrangian filter method for general quadratic programs. SIAM J. Sci. Comput. **30**(4), 1706–1729 (2008)
16. Friedlander, M.P., Saunders, M.A.: A globally convergent linearly constrained Lagrangian method for nonlinear optimization. SIAM J. Optim. **15**(3), 863–897 (2005)
17. Gao, B., Liu, X., Yuan, Yx.: Parallelizable algorithms for optimization problems with orthogonality constraints. SIAM J. Sci. Comput. **41**(3), A1949–A1983 (2019)
18. Gill, P.E., Murray, W., Saunders, M.A.: SNOPT: an SQP algorithm for large-scale constrained optimization. SIAM Rev. **47**(1), 99–131 (2005). ((**electronic**))
19. Hestenes, M.R.: Multiplier and gradient methods. Journal of Optimization Theory and Applications **4**(5), 303–320 (1969)
20. Jalilzadeh, A., Shanbhag, U.V., Blanchet, J., Glynn, P.W.: Smoothed variable sample-size accelerated proximal methods for nonsmooth stochastic convex programs. Stochastic Systems **12**(4), 373–410 (2022)
21. Kang, M., Kang, M., Jung, M.: Inexact accelerated augmented Lagrangian methods. Computational Optimization and Applications **62**(2), 373–404 (2015)
22. Kloft, M., Brefeld, U., Laskov, P., Müller, K.R., Zien, A., Sonnenburg, S.: Efficient and accurate L_p-norm multiple kernel learning. Advances in Neural Information Processing Systems **22** (2009)
23. Koshal, J., Nedić, A., Shanbhag, U.V.: Multiuser optimization: Distributed algorithms and error analysis. SIAM Journal on Optimization **21**(3), 1046–1081 (2011)
24. Lan, G., Monteiro, R.D.: Iteration-complexity of first-order augmented Lagrangian methods for convex programming. Mathematical Programming **155**(1–2), 511–547 (2016)
25. Liu, Y.F., Liu, X., Ma, S.: On the nonergodic convergence rate of an inexact augmented Lagrangian framework for composite convex programming. Mathematics of Operations Research **44**(2), 632–650 (2019)
26. Lu, Z., Zhou, Z.: Iteration-complexity of first-order augmented Lagrangian methods for convex conic programming. SIAM J. Optim. **33**(2), 1159–1190 (2023)
27. Moreau, J.J.: Proximité et dualité dans un espace hilbertien. Bulletin de la Société mathématique de France **93**, 273–299 (1965)
28. Murtagh, B.A., Saunders, M.A.: A projected Lagrangian algorithm and its implementation for sparse nonlinear constraints. Springer (1982)
29. Necoara, I., Patrascu, A., Glineur, F.: Complexity of first-order inexact Lagrangian and penalty methods for conic convex programming. Optimization Methods and Software **34**(2), 305–335 (2019)

30. Nedelcu, V., Necoara, I., Tran-Dinh, Q.: Computational complexity of inexact gradient augmented Lagrangian methods: application to constrained mpc. SIAM Journal on Control and Optimization **52**(5), 3109–3134 (2014)
31. Nesterov, Y.: A method for unconstrained convex minimization problem with the rate of convergence $\mathcal{O}(1/k^2)$. Doklady an ussr **269**, 543–547 (1983)
32. Nesterov, Y.: Introductory lectures on convex optimization: A basic course, vol. 87. Springer Science & Business Media (2003)
33. Nesterov, Y.: Smooth minimization of non-smooth functions. Mathematical programming **103**(1), 127–152 (2005)
34. Nesterov, Y., et al.: Lectures on convex optimization, vol. 137. Springer (2018)
35. Patrascu, A., Necoara, I., Tran-Dinh, Q.: Adaptive inexact fast augmented Lagrangian methods for constrained convex optimization. Optimization Letters **11**(3), 609–626 (2017)
36. Polyak, B.T.: Introduction to optimization (1987)
37. Powell, M.J.: A method for nonlinear constraints in minimization problems. Optimization pp. 283–298 (1969)
38. Rockafellar, R.T.: A dual approach to solving nonlinear programming problems by unconstrained optimization. Mathematical Programming **5**(1), 354–373 (1973)
39. Rockafellar, R.T.: Augmented Lagrangians and applications of the proximal point algorithm in convex programming. Mathematics of Operations Research **1**(2), 97–116 (1976)
40. Tibshirani, R.: Regression shrinkage and selection via the lasso. Journal of the Royal Statistical Society: Series B (Methodological) **58**(1), 267–288 (1996)
41. Tibshirani, R., Saunders, M., Rosset, S., Zhu, J., Knight, K.: Sparsity and smoothness via the fused Lasso. Journal of the Royal Statistical Society: Series B (Stat. Method.) **67**(1), 91–108 (2005)
42. Tong, X., Xia, L., Wang, J., Feng, Y.: Neyman-Pearson classification: parametrics and sample size requirement. The Jrnl. of Machine Learning Research **21**(1), 380–427 (2020)
43. Wilson, R.B.: A simplicial algorithm for concave programming. Ph. D. Dissertation, Graduate School of Bussiness Administration (1963)
44. Xu, Y.: Iteration complexity of inexact augmented Lagrangian methods for constrained convex programming. Mathematical Programming **185**(1), 199–244 (2021)
45. Zhang, L., Zhang, Y., Wu, J., Xiao, X.: Solving stochastic optimization with expectation constraints efficiently by a stochastic augmented Lagrangian-type algorithm. INFORMS Journal on Computing **34**(6), 2989–3006 (2022)
46. Zou, H., Hastie, T.: Regularization and variable selection via the elastic net. Journal of the Royal Statistical Society: series B (Statistical Methodology) **67**(2), 301–320 (2005)