Noname manuscript No.
(will be inserted by the editor)

A smoothed augmented Lagrangian framework for convex optimization with nonsmooth constraints

Peixuan Zhang · Uday V. Shanbhag* · Ethan X. Fang

Received: date / Accepted: date

The 2nd author 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 Lagrangian schemes.

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 subproblem resolution), the penalty parameter ρ_k, and smoothing parameter η_k at epoch k, we derive rate and complexity guarantees of $\tilde{\mathcal{O}}(1/\varepsilon^{3/2})$ and $\tilde{\mathcal{O}}(1/\varepsilon)$ in convex and strongly convex regimes for computing an ε-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

Mathematics Subject Classification (2020) 90C25 · 90C30

* Corresponding author; P. Zhang and U. V. Shanbhag are at Pennsylvania State University, PA (E-mail: pqz5090,udaybag@psu.edu) · E. X. Fang is at Duke University, NC (E-mail: ethan.fang@duke.edu).

1 Introduction

We consider the nonsmooth convex program, defined as

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

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 [15,39], interior point methods [7], and augmented Lagrangian (AL) schemes [34,35]. Of these, AL schemes have proven to be enormously influential in the context of scientific computing [1,8, 11], and more specifically in nonlinear programming in the form of solvers such as `minos` [25,13] and `lancelot` [9] as well as more refined techniques [14,12]. There has been a significant interest in deriving overall complexity bounds [21, 40] 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 $\mathcal{O}(\varepsilon^{-5})$ for computing an ε-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 [16] and Powell [33] with a comprehensive rate analysis provided by Rockafellar [34]. 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. [18] 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 [21,22], where either smoothness of f [21] or a composite structure [22] is assumed. Overall complexity guarantees were first provided by Lan and Monteiro [21], Aybat and Iyengar [4], Necoara et al. [26] and most recently Lu and Zhou [23], where the latter three references allowed for conic settings. In fact, Lu and Zhou [23] showed that

Ref.	f	g	Metrics	ρ_k	Rate	Complex.	Comment
[34]	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
[21]	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
[31]	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
[23]	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
[40]*	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; ‡ Composite function: $f(\mathbf{x}) = p(\mathbf{x}) + \gamma(\mathbf{x})$ where $p(\cdot)$ is smooth and $\gamma(\cdot)$ is nonsmooth, proximal; $\star$ Additional boundedness condition required; $\diamond$ I-AL with regularization terms.

Table 1 ALM for deterministic convex optimization

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. [27] considered convex and strongly convex regimes. Notably, Necoara et al. [26] derived an overall complexity of $\mathcal{O}(\boldsymbol{\varepsilon}^{-\frac{3}{2}})$ and $\mathcal{O}(\boldsymbol{\varepsilon}^{-1})$ for smooth settings under convex and strongly convex objective f, respectively. More recently, Xu [40] 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 (**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 [32], smoothing techniques [5] represent an efficient avenue for a subclass of nonsmooth problems. Moreau [24] 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 [30] 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 [6] 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 [10] and stochastic [17] convex optimization problems.

1.2. Applications. We present three applications where nonsmooth convex constraints emerge. (a) *Regression.* Lasso regression [36] 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 [43] can be articulated as follows where $C_1 > 0$.

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

This reduces to standard *Lasso* [36] when $\alpha = 0$ and is generalizable to *fused Lasso* [37] 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 [38] 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 [42] 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 [19].

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

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 ε-optimal solution are proven to be $\mathcal{O}(\varepsilon^{-(3+\delta)})$ (constant penalty) and $\tilde{\mathcal{O}}(\varepsilon^{-3/2})$ (geometrically increasing penalty). Analogous bounds in strongly convex settings are given by $\tilde{\mathcal{O}}(\varepsilon^{-(2+\delta)})$ for constant and $\tilde{\mathcal{O}}(\varepsilon^{-1})$ for geometrically increasing penalty parameters. Similar complexity guarantees are available with a fixed smoothing parameter, akin to those developed in [30,6] for convex programs with nonsmooth objectives. (iii) Preliminary numerical results are provided in Section 5 before concluding in Section 6.

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 $\mathcal{L}_0(\bullet, \lambda)$ 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.

$$\begin{array}{ll} \min\limits_{\mathbf{x}\in\mathcal{X}, \mathbf{v}\geq \mathbf{0}} & f(\mathbf{x}) \\ \text{subject to} & g(\mathbf{x}) + \mathbf{v} = \mathbf{0}, \qquad (\lambda) \end{array}$$

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

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

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 [34].

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

$$\mathcal{L}_\rho(\mathbf{x}, \lambda) = \left(f(\mathbf{x}) + \tfrac{\rho}{2} \left(d_- \left(\tfrac{\lambda}{\rho} + g(\mathbf{x}) \right) \right)^2 - \tfrac{1}{2\rho} \|\lambda\|^2 \right)$$
$$\text{and } \nabla_\lambda \mathcal{L}_\rho(\mathbf{x}, \lambda) = \left(-\tfrac{\lambda}{\rho} + \Pi_+ \left(\tfrac{\lambda}{\rho} + g(\mathbf{x}) \right) \right).$$

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

$$\mathcal{D}_\rho(\lambda) \triangleq \inf_{\mathbf{x} \in \mathcal{X}} \mathcal{L}_\rho(\mathbf{x}, \lambda), \tag{2}$$

can be shown to be differentiable [34].

Lemma 2 Consider the function $\mathcal{D}_\rho$ defined as (2). 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 $q_\rho(\lambda) \triangleq \arg\max_u \left[\mathcal{D}_0(u) - \tfrac{1}{2\rho} \|u - \lambda\|^2 \right]$.

Our interest lies in nonsmooth, albeit smoothable, convex functions, defined next.

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*

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

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 [6] 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.

Assumption 21.
(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. [20]). Throughout the rest of this paper, we assume that Assumption 21 holds.

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

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

$$|f(\mathbf{x}) - f_\eta(\mathbf{x})| \leq \eta\beta \tag{3}$$
$$|g_i(\mathbf{x}) - g_{i,\eta}(\mathbf{x})| \leq \eta\beta, \quad i = 1, 2, \ldots, m. \tag{4}$$

Consequently, for any $\lambda \geq 0$, by adding (3) to $\lambda_i \times$ (4) for $i = 1, \cdots, m$,

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

(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 $C = (\|\lambda\|m+1)\beta$.

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

Similarly, we have that

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

This implies that for any $\lambda \in \mathbb{R}^m_+$, $|\mathcal{D}_{\eta,0}(\lambda) - \mathcal{D}_0(\lambda)| \leq \eta C$.
(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 $\lambda \geq 0$, let $u_1 \in \arg\max_u \mathcal{D}_{\eta,\rho}(\lambda)$ and $u_2 \in \arg\max_u \mathcal{D}_\rho(\lambda)$. Then

$$\begin{aligned}
\mathcal{D}_{\eta,\rho}(\lambda) - \mathcal{D}_\rho(\lambda) &= \max_{u\in\mathbb{R}^m} \left[\mathcal{D}_{\eta,0}(u) - \frac{1}{2\rho}\|u-\lambda\|^2 \right] - \max_{u\in\mathbb{R}^m} \left[\mathcal{D}_0(u) - \frac{1}{2\rho}\|u-\lambda\|^2 \right] \\
&= \max_{u\in\mathbb{R}^m} \left[\mathcal{D}_{\eta,0}(u) - \frac{1}{2\rho}\|u-\lambda\|^2 \right] - \left[\mathcal{D}_0(u_2) - \frac{1}{2\rho}\|u_2-\lambda\|^2 \right] \\
&\leq \max_{u\in\mathbb{R}^m} \left[\mathcal{D}_{\eta,0}(u) - \frac{1}{2\rho}\|u-\lambda\|^2 \right] - \left[\mathcal{D}_0(u_1) - \frac{1}{2\rho}\|u_1-\lambda\|^2 \right] \\
&\leq |\,\mathcal{D}_{\eta,0}(u_1) - \mathcal{D}_0(u_1)\,| \overset{\text{Lemma 3(ii)}}{\leq} \eta C.
\end{aligned}$$

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

□

We now consider the smoothed counterpart of (NSCopt), defined as

$$\min_{\mathbf{x}\in\mathcal{X}} \{\, f_\eta(\mathbf{x}) \mid g_\eta(\mathbf{x}) \leq 0 \,\}. \tag{NSCopt$_\eta$}$$

Under a Slater regularity condition, the set of optimal multipliers is bounded (cf. [20]). Similar bounds are derived for the η-smoothed problem.

Proposition 21. (a) For any $\eta > 0$, there exists $\bar{\mathbf{x}} \in \mathcal{X}$ such that $g_\eta(\bar{\mathbf{x}}) < 0$.
(b) The set of optimal multipliers Λ^* for (NSCopt) is bounded as per

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

(c) For any $\eta > 0$, the set of optimal multipliers Λ_η^* for (NSCopt$_\eta$) is bounded as per

$$\Lambda_\eta^* \subseteq B_{\lambda,\eta} = \left\{ \lambda \geq 0 \,\middle|\, \sum_{i=1}^m \lambda_i \leq b_{\lambda,\eta} \right\} \text{ where } b_{\lambda,\eta} \geq \tfrac{f(\bar{\mathbf{x}}) - \mathcal{D}_0^* + \eta(\beta + \tilde{C}^*)}{\min_j\{-g_j(\bar{\mathbf{x}})\}}.$$

Proof. (a) By Assumption 21(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.
(b) By the Slater regularity condition, we directly conclude from [20] that

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

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

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

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

$$-\mathcal{D}_{0,\eta}(\lambda_\eta^*) \overset{\text{(Optimality of } \lambda_\eta^*)}{\leq} -\mathcal{D}_{0,\eta}(\lambda^*) \overset{\text{(Lemma 3(ii))}}{\leq} -\mathcal{D}_0(\lambda^*) + \eta(mb_\lambda + 1)\beta.$$

Consequently, if $\mathcal{D}_{0,\eta}^* \triangleq \mathcal{D}_{0,\eta}(\lambda_\eta^*)$, $\mathcal{D}_0^* \triangleq \mathcal{D}_0(\lambda^*)$ and $\tilde{C}^* \triangleq (mb_\lambda + 1)\beta$, then

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

□

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.

Theorem 21. Let $(\mathbf{x}^*, \lambda^*)$ and $(\mathbf{x}_\eta^*, \lambda_\eta^*)$ represent saddle points of (NSCopt) and (NSCopt$_\eta$), respectively.
(a) Suppose $\mathbf{x}_\eta^* \in \mathcal{X}$ is a feasible solution of (NSCopt$_\eta$). Then $\mathbf{x}_\eta^*$ is an $\eta\beta\|\mathbf{1}\|$-feasible of (NSCopt), i.e. $d_-(g(\mathbf{x}_\eta^*)) \leq \eta\beta\|\mathbf{1}\|$.
(b) Suppose $(\mathbf{x}_\eta^*, \lambda_\eta^*)$ is a saddle-point of (NSCopt$_\eta$). Then $(\mathbf{x}_\eta^*, \lambda_\eta^*)$ is an $2\eta\beta(1 + mb_{\lambda,\eta})$-saddle-point of (NSCopt), i.e. for all $(\mathbf{x}, \lambda) \in \mathcal{X} \times \mathbb{R}_+^m$,

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

Proof. (a) Suppose $\mathbf{x}_\eta^* \in \mathcal{X}$ is a feasible solution of (NSCopt_η). 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 21. Let $(\mathbf{x}_\eta^*, \lambda_\eta^*)$ be the saddle point of $L_\eta(\cdot,\cdot)$. We have

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

The final result follows through the following sequence of inequalities that

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

□

The following Lemma 4 shows the relation between $q_{\eta,\rho}(\bullet)$ and $q_\rho(\bullet)$.

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

Proof. (i) By definition, we have that

$$\begin{aligned}
q_\rho(\lambda) &= \arg\max_{u \in \mathbb{R}^m} \left(\mathcal{D}_0(u) - \tfrac{1}{2\rho}\|u - \lambda\|^2\right) \qquad (5) \\
&= \arg\min_{u \in \mathbb{R}^m} \left(-\mathcal{D}_0(u) + \tfrac{1}{2\rho}\|u - \lambda\|^2\right) = \text{prox}_{-\mathcal{D}_0,\rho}(\lambda).
\end{aligned}$$

$$\text{Similarly, } q_{\eta,\rho}(\lambda) = \text{prox}_{-D_{\eta,0},\rho}(\lambda). \qquad (6)$$

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 (5) and (6), respectively,

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

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, $\|q_\rho(\lambda)\| = \|q_\rho(\lambda) - q_\rho(\lambda^*) + \lambda^*\| \leq \|\lambda\| + 2\|\lambda^*\|$. 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,

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

□

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 $\mathbf{x}_0, \lambda_0$, $K > 0$, and sequences $\{\rho_k, \epsilon_k, \eta_k\}$. For $k = 1, \cdots, K$, we have

[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] $\lambda_{k+1} = \lambda_k + \rho_k \nabla_\lambda \mathcal{L}_{\eta_k,\rho_k}(\mathbf{x}_{k+1}, \lambda_k)$.

The Lagrange multiplier update can be expressed as follows [2].

Lemma 5 Consider the smoothed augmented Lagrangian scheme (**Sm-AL**). Then for any $k > 0$, step [2] is equivalent to the following equation.

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

The next assumption holds for parameter sequences employed in (**Sm-AL**). Unless mentioned otherwise, Assumptions 21 and 22 hold throughout.

Assumption 22. 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 $b \geq 0$.

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 [34] Lemma 4.3.

Lemma 6 Let $\{\mathbf{x}_k, \lambda_k\}$ be generated by (**Sm-AL**). For any $k \geq 0$, suppose $\mathbf{x}_{k+1}$ satisfies $\mathcal{L}_{\eta_k,\rho_k}(\mathbf{x}_{k+1}, \lambda_k) - \mathcal{D}_{\rho_k}(\lambda_k) \leq \epsilon_k \eta_k^b$ where $b \geq 0$. Then for $k \geq 0$

$$\|\nabla_\lambda \mathcal{L}_{\eta_k,\rho_k}(x_{k+1}, \lambda_k) - \nabla_\lambda \mathcal{D}_{\eta_k,\rho_k}(\lambda_k)\|^2 \leq \tfrac{2\epsilon_k \eta_k^b}{\rho_k}. \tag{7}$$

By choosing appropriate sequences $\{\epsilon_k, \eta_k, \rho_k\}$, $\{(2\epsilon_k \eta_k^b)/\rho_k\}$ is diminishing (see Lemma 6). We now derive a uniform bound on the sequence $\{\lambda_k\}$.

Lemma 7 (Bound on λ_k) Consider $\{\lambda_k\}$ generated by (**Sm-AL**).
(a) $\{\lambda_k\}$ is a convergent sequence. (b) For any K, we have

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

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

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

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

$$\begin{aligned}\|\lambda_{k+1} - q_{\eta_k,\rho_k}(\lambda_k)\| &= \|\lambda_k + \rho_k \left(\nabla_\lambda \mathcal{L}_{\eta_k,\rho_k}(\mathbf{x}_{k+1}, \lambda_k)\right) - q_{\eta_k,\rho_k}(\lambda_k)\| \\ &= \|\lambda_k + \rho_k \left(\nabla_\lambda \mathcal{L}_{\eta_k,\rho_k}(\mathbf{x}_{k+1}, \lambda_k)\right) - \rho_k \nabla_\lambda \mathcal{D}_{\eta_k,\rho_k}(\lambda_k) - \lambda_k\| \\ &\leq \rho_k \left\|\nabla_\lambda \mathcal{L}_{\eta_k,\rho_k}(\mathbf{x}_{k+1}, \lambda_k)) - \nabla_\lambda \mathcal{D}_{\eta_k,\rho_k}(\lambda_k)\right\| \overset{\text{Lem. 6}}{\leq} \sqrt{2\rho_k \epsilon_k \eta_k^b}.\end{aligned}$$

From Lemma 3, $\|q_{\eta_k,\rho_k}(\lambda^*) - q_{\rho_k}(\lambda^*)\| \leq 2\sqrt{\rho_k \eta_k (\|\lambda^*\| m + C_m)\beta}$, implying that

$$\|\lambda_{k+1} - \lambda^*\| \leq \sqrt{2\rho_k \epsilon_k \eta_k^b} + 2\sqrt{\rho_k \eta_k (\|\lambda^*\| m + C_m)\beta} + \|\lambda_k - \lambda^*\|. \tag{8}$$

By leveraging the deterministic form of the Robbins-Siegmund Lemma [32], 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 nonnegative value. It follows that $\{\lambda_k\}$ is convergent.

(b) Summing (8) from $k=0,\cdots,K-1$, we obtain that

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

□

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 [34, Th. 3.2]).

Proposition 31 [Dual sub-optimality]. Consider the sequence $\{\lambda_k\}$ generated by (**Sm-AL**), where $\rho_k=\rho$ for every $k\geq 0$. If $\tilde{B}_1,\tilde{B}_2$ are constants, then the following holds for $\bar{\lambda}_K \triangleq \frac{\sum_{i=0}^{K-1}\lambda_i}{KK-1}$ and for any $K_{K-1}\geq 0$,

$$f^*-\mathcal{D}_\rho(\bar{\lambda}_K)\leq \tfrac{1}{2\rho K}\|\lambda_0-\lambda^*\|^2+\tfrac{\tilde{B}_1}{K}\sum_{k=0}\frac{\sqrt{2\epsilon_k\eta_k^b}}{\sqrt{\rho}}+\tfrac{\tilde{B}_2}{K}\sum_{k=0}\eta_k\leq\mathcal{O}\left(\tfrac{1}{K}\right).$$

Proof. Recall that $\mathcal{D}_{\eta_k,\rho}(\lambda)$ 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

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

where $-\mathcal{D}_{\eta_k,\rho}(\lambda^*) \geq -\mathcal{D}_{\eta_k,\rho}(\lambda_k) - \nabla_\lambda \mathcal{D}_{\eta_k,\rho}(\lambda_k)^\top(\lambda^* - \lambda_k)$. It follows that

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

where the last inequality follows from Lemma 21. By invoking Lemma 3(iii), 7, and $\|\lambda_k\| + \|\lambda^*\| \leq \tilde{B}_\lambda = B_\lambda + 2b_\lambda$, we obtain

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

By summing from $k=0$ to $K-1$, dividing by K, and invoking the concavity of D_ρ,

$$\begin{aligned}
-\left(D_\rho(\bar{\lambda}_K) - f^*\right) \leq & \tfrac{1}{2\rho K}(\|\lambda_0-\lambda^*\|^2 - \|\lambda_K-\lambda^*\|^2) + \tfrac{1}{K}\sum_{k=0}^{K-1}\eta_k(2\tilde{B}_\lambda m+1)\beta \\
& + \tfrac{1}{K}\sum_{k=0}^{K-1}\|\nabla_\lambda D_{\eta_k,\rho}(\lambda_k) - \nabla_\lambda \mathcal{L}_{\eta_k,\rho}(\mathbf{x}_{k+1},\lambda_k)\|\,\|\lambda_{k+1}-\lambda^*\| \\
\leq & \tfrac{1}{2\rho K}\|\lambda_0-\lambda^*\|^2 + \tfrac{\tilde{B}_1}{K}\sum_{k=0}^{K-1}\tfrac{\sqrt{2\epsilon_k\eta_k^b}}{\sqrt{\rho}} + \tfrac{\tilde{B}_2}{K}\sum_{k=0}^{K-1}\eta_k,
\end{aligned}$$

where boundedness of λ_k follows from Lemma 3 and $\tilde{B}_\lambda, \tilde{B}_1, \tilde{B}_2$ are constants. □

Next, we derive a rate statement on the infeasibility.

Proposition 32 [Rate on primal infeasibility]. Let$\{\mathbf{x}_k, \lambda_k\}$ be sequence generated by (**Sm-AL**). Then the following holds for any $K > 0$, where

$\tilde{B}, \tilde{C} \geq 0$.

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

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

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

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

$$\begin{aligned} d_-(g_{\eta_k}(\mathbf{x}_{k+1})) &\leq \|\nabla_\lambda \mathcal{L}_{\eta_k,\rho}(\mathbf{x}_{k+1}, \lambda_k)\| + \underbrace{d_-\left(\Pi_-\left(\frac{\lambda_k}{\rho} + g_{\eta_k}(\mathbf{x}_{k+1})\right)\right)}_{=0} \\ &= \|\nabla_\lambda \mathcal{L}_{\eta_k,\rho}(\mathbf{x}_{k+1}, \lambda_k)\|. \end{aligned} \tag{9}$$

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

$$\begin{aligned} d_-(g(\bar{\mathbf{x}}_K)) &= \inf_{u\in\mathbb{R}^m_-}\|g(\bar{\mathbf{x}}_K) - u\|_2 \leq \inf_{u\in\mathbb{R}^m_-}\|g(\bar{\mathbf{x}}_K) - u\|_1 = \sum_{j=1}^m \inf_{u_j\leq 0}|g_j(\bar{\mathbf{x}}_K) - u_j|_1 \\ &= \sum_{j=1}^m \max\{g_j(\bar{\mathbf{x}}_K), 0\} \leq \frac{1}{K}\sum_{i=0}^{K-1}\sum_{j=1}^m \max\{g_j(\mathbf{x}_{i+1}), 0\} \\ &\leq \frac{1}{K}\sum_{i=0}^{K-1}\sum_{j=1}^m \max\{g_{j,\eta_i}(\mathbf{x}_{i+1}) + \eta_i\beta, 0\} = \frac{1}{K}\sum_{i=0}^{K-1}\inf_{u\in\mathbb{R}^m_-}\|g_{\eta_i}(\mathbf{x}_{i+1}) + \eta_i\beta\mathbf{1} - u\|_1 \\ &\leq \frac{1}{K}\sum_{i=0}^{K-1}\inf_{u\in\mathbb{R}^m_-}\sqrt{m}\|g_{\eta_i}(\mathbf{x}_{i+1}) + \eta_i\beta\mathbf{1} - u\|_2 = \frac{\sqrt{m}}{K}\sum_{k=1}^{K} d_-(g_{\eta_i}(\mathbf{x}_{i+1}) + \eta_i\beta\mathbf{1}) \\ &\leq \frac{\sqrt{m}}{K}\sum_{i=0}^{K-1}(d_-(g_{\eta_i}(\mathbf{x}_{i+1})) + \eta_i\beta\|\mathbf{1}\|_2) \overset{(9)}{\leq} \frac{\sqrt{m}}{K}\sum_{i=0}^{K-1}\left(\|\nabla_\lambda \mathcal{L}_{\eta_i,\rho}(\mathbf{x}_{i+1}, \lambda_i)\| + \sqrt{m}\eta_i\beta\right) \\ &\leq \frac{\sqrt{m}}{K}\sum_{i=0}^{K-1}\left(\|\nabla_\lambda \mathcal{L}_{\eta_i,\rho}(\mathbf{x}_{i+1}, \lambda_i) - \nabla_\lambda \mathcal{D}_{\eta_i,\rho}(\lambda_i)\| + \|\nabla_\lambda \mathcal{D}_{\eta_i,\rho}(\lambda_i)\| + \sqrt{m}\eta_i\beta\right). \end{aligned} \tag{10}$$

Recall that

$$\|\nabla_\lambda \mathcal{D}_{\eta_k,\rho}(\lambda_1) - \nabla_\lambda \mathcal{D}_{\eta_k,\rho}(\lambda_2)\| \leq \frac{1}{\rho}\|q_{\eta,\rho}(\lambda_1) - q_{\eta,\rho}(\lambda_2)\| + \frac{1}{\rho}\|\lambda_1 - \lambda_2\| \leq \frac{2}{\rho}\|\lambda_1 - \lambda_2\|.$$

Since $\mathcal{D}_{\eta_k,\rho}$ is a $(2/\rho)$-smooth concave function, then by leveraging [29] for any $\lambda \geq 0$,

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

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, $\sqrt{u+v} \leq \sqrt{u} + \sqrt{v}$ for $u, v \geq 0$, the subaddivity of concave functions, we have from (10),

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

□

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

Theorem 31 [Rate on primal sub-opt]. Consider the sequence $\{\mathbf{x}_k, \lambda_k\}$ generated by (**Sm-AL**). Then (11) holds for any $K > 0$, where $\tilde{B}_1, \tilde{B}_2, \tilde{C}_1 \geq 0$.

$$
-\left(\tfrac{\tilde{B}_1}{K} + \tfrac{\tilde{B}_2}{\sqrt{K}} + \eta_K\beta \right) \leq f(\bar{\mathbf{x}}_K) - f^* \leq \tfrac{\tilde{C}_1}{K} + \tfrac{1}{K} \sum_{k=0}^{K-1} \left(\epsilon_k \eta_k^b + \eta_k \beta \right). \tag{11}
$$

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

$$
\mathbf{x}_{\eta_K}^* \in \arg\min_{\mathbf{x}\in\mathcal{X}} \mathcal{L}_{\eta_K,\rho}\left(\mathbf{x}, \lambda_{\eta_K}^*\right).
$$

$$
\begin{aligned}
f^*_{\eta_K} &\leq \mathcal{L}_{\eta_K,\rho}(\bar{\mathbf{x}}_K, \lambda^*_{\eta_K}) = f_{\eta_K}(\bar{\mathbf{x}}_K) + \tfrac{\rho}{2}\left(d_-\left(\tfrac{\lambda^*_{\eta_K}}{\rho} + g_{\eta_K}(\bar{\mathbf{x}}_K)\right)\right)^2 - \tfrac{1}{2\rho}\|\lambda^*_{\eta_K}\|^2 \\
&\leq f_{\eta_K}(\bar{\mathbf{x}}_K) + \tfrac{\rho}{2}\left(d_-\left(g_{\eta_K}(\bar{\mathbf{x}}_K)\right) + \left\|\tfrac{\lambda^*_{\eta_K}}{\rho}\right\|\right)^2 - \tfrac{1}{2\rho}\|\lambda^*_{\eta_K}\|^2 \\
&= f_{\eta_K}(\bar{\mathbf{x}}_K) + \tfrac{\rho}{2}\left(d_-\left(g_{\eta_K}(\bar{\mathbf{x}}_K)\right)\right)^2 + \left\|\lambda^*_{\eta_K}\right\| d_-\left(g_{\eta_K}(\bar{\mathbf{x}}_K)\right) \\
&\overset{\text{Lem. 21}}{\leq} f_{\eta_K}(\bar{\mathbf{x}}_K) + \tfrac{\rho}{2}\left(d_-\left(g_{\eta_K}(\bar{\mathbf{x}}_K)\right)\right)^2 + b_{\lambda,\eta} d_-\left(g_{\eta_K}(\bar{\mathbf{x}}_K)\right).
\end{aligned}
$$

By invoking Proposition 32, we obtain the following inequality.

$$
f^*_{\eta_K} - f_{\eta_K}(\bar{\mathbf{x}}_K) \leq \tfrac{\tilde{B}_1}{K} + \tfrac{\tilde{B}_2}{\sqrt{K}}. \tag{12}
$$

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 3, it follows that

$$
\begin{aligned}
f(\mathbf{x}^*) - f(\bar{\mathbf{x}}_K) = &\underbrace{f(\mathbf{x}^*) - f(\mathbf{x}^*_{\eta_K})}_{\leq 0} + \underbrace{f(\mathbf{x}^*_{\eta_K}) - f_{\eta_K}(\mathbf{x}^*_{\eta_K})}_{\leq \eta_K \beta} + \underbrace{f_{\eta_K}(\mathbf{x}^*_{\eta_K}) - f_{\eta_K}(\bar{\mathbf{x}}_K)}_{(12)} \\
&+ \underbrace{f_{\eta_K}(\bar{\mathbf{x}}_K) - f(\bar{\mathbf{x}}_K)}_{\leq 0} \leq \eta_K \beta + \tfrac{\tilde{B}_1}{K} + \tfrac{\tilde{B}_2}{\sqrt{K}}.
\end{aligned}
$$

(ii) *Upper bound.* We begin by recalling that $\mathbf{x}_{k+1}$ satisfies the following.

$$
\begin{aligned}
&\mathcal{L}_{\eta_k,\rho}(\mathbf{x}_{k+1}, \lambda_k) - \min_{\mathbf{x}\in\mathcal{X}} \mathcal{L}_{\eta_k,\rho}(\mathbf{x}, \lambda_k) \leq \epsilon_k \eta_k^b \\
\implies &\mathcal{L}_{\eta_k,\rho}(\mathbf{x}_{k+1}, \lambda_k) - \mathcal{L}_{\eta_k,\rho}(\mathbf{x}^*_{\eta_k}, \lambda^*_{\eta_k}) \leq \epsilon_k \eta_k^b.
\end{aligned}
$$

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

$$
\begin{aligned}
f_{\eta_k}(\mathbf{x}_{k+1}) - f^*_{\eta_k} &\leq \tfrac{\rho}{2}\left(d_-\left(\tfrac{\lambda^*_{\eta_k}}{\rho} + g_{\eta_k}(\mathbf{x}^*_{\eta_k})\right)\right)^2 - \tfrac{\rho}{2}\left(d_-\left(\tfrac{\lambda_k}{\rho} + g_{\eta_k}(\mathbf{x}_{k+1})\right)\right)^2 \\
&\quad + \epsilon_k \eta_k^b + \tfrac{1}{2\rho}\left(\|\lambda^*_{\eta_k}\|^2 - \|\lambda_k\|^2\right) \\
&= \tfrac{\rho}{2}\left(\left(d_-\left(\tfrac{\lambda^*_{\eta_k}}{\rho} - \tfrac{\lambda_k}{\rho} + \tfrac{\lambda_k}{\rho} + g_{\eta_k}(\mathbf{x}^*_{\eta_k})\right)\right)^2 - \left(d_-\left(\tfrac{\lambda_k}{\rho} + g_{\eta_k}(\mathbf{x}_{k+1})\right)\right)^2\right) \\
&\quad + \epsilon_k \eta_k^b + \tfrac{1}{2\rho}\left(\|\lambda^*_{\eta_k}\|^2 - \|\lambda_k\|^2\right) \\
&\leq \tfrac{\rho}{2}\left(\left(d_-\left(\tfrac{\lambda_k}{\rho} + g_{\eta_k}(\mathbf{x}^*_{\eta_k})\right)\right)^2 - \left(d_-\left(\tfrac{\lambda_k}{\rho} + g_{\eta_k}(\mathbf{x}_{k+1})\right)\right)^2\right) + \tfrac{1}{2\rho}\left\|\lambda^*_{\eta_k} - \lambda_k\right\|^2 \\
&\quad + \|\lambda^*_{\eta_k} - \lambda_k\|\left(d\left(\tfrac{\lambda_k}{\rho} + g_{\eta_k}(\mathbf{x}^*_{\eta_k})\right)\right) + \epsilon_k \eta_k^b + \tfrac{1}{2\rho}\left(\|\lambda^*_{\eta_k}\|^2 - \|\lambda_k\|^2\right).
\end{aligned}
$$

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 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

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

which implies

$$\begin{aligned}
&f_{\eta_k}(\mathbf{x}_{k+1})-f^*_{\eta_k}\\
&\leq \tfrac{\rho}{2}\left(\left(d_{-}\left(\tfrac{\lambda_k}{\rho}+g_{\eta_k}(\mathbf{x}^*_{\eta_k})\right)\right)^2-\left(d_{-}\left(\tfrac{\lambda_k}{\rho}+g_{\eta_k}(\mathbf{x}_{k+1})\right)\right)^2\right)+\tfrac{1}{2\rho}\left\|\lambda^*_{\eta_k}-\lambda_k\right\|^2\\
&+\|\lambda^*_{\eta_k}-\lambda_k\|\left(d\left(\tfrac{\lambda_k}{\rho}+g_{\eta_k}(\mathbf{x}^*_{\eta_k})\right)\right)+\epsilon_k\eta^b_k+\tfrac{1}{2\rho}\left(\|\lambda^*_{\eta_k}\|^2-\|\lambda_k\|^2\right)\\
&\leq \tfrac{\rho}{2}\left(\left\|\tfrac{\lambda_k}{\rho}\right\|^2-\left\|\tfrac{\lambda_{k+1}}{\rho}\right\|^2\right)+\tfrac{\|\lambda^*_{\eta_k}-\lambda_k\|^2}{2\rho}+\tfrac{\|\lambda^*_{\eta_k}-\lambda_k\|\|\lambda_k\|}{\rho}+\epsilon_k\eta^b_k\\
&+\tfrac{1}{2\rho}\left(2\|\lambda^*_{\eta_k}-\lambda_k\|^2+\|\lambda_k\|^2\right)\\
&\leq \tfrac{\rho}{2}\left(\left\|\tfrac{\lambda_k}{\rho}\right\|^2-\left\|\tfrac{\lambda_{k+1}}{\rho}\right\|^2\right)+\tfrac{\|\lambda^*_{\eta_k}-\lambda_k\|^2}{2\rho}+\tfrac{\|\lambda^*_{\eta_k}-\lambda_k\|^2+\|\lambda^*_{\eta_k}-\lambda_k\|\|\lambda^*_{\eta_k}\|}{\rho}\\
&+\tfrac{4\|\lambda^*_{\eta_k}-\lambda_k\|^2+2\|\lambda^*_{\eta_k}\|^2}{2\rho}+\epsilon_k\eta^b_k\\
&\leq \tfrac{\rho}{2}\left(\left\|\tfrac{\lambda_k}{\rho}\right\|^2-\left\|\tfrac{\lambda_{k+1}}{\rho}\right\|^2\right)+\tfrac{7\|\lambda^*_{\eta_k}-\lambda_k\|^2+2b_{\lambda,\eta}\|\lambda^*_{\eta_k}-\lambda_k\|+2b^2_{\lambda,\eta}}{2\rho}+\epsilon_k\eta^b_k\\
&\leq \tfrac{\rho}{2}\left(\left\|\tfrac{\lambda_k}{\rho}\right\|^2-\left\|\tfrac{\lambda_{k+1}}{\rho}\right\|^2\right)+\tfrac{7\tilde{B}_{\lambda,1}+2\tilde{B}_{\lambda,2}b_{\lambda,\eta}+2b^2_{\lambda,\eta}}{2\rho}+\epsilon_k\eta^b_k \qquad (13)
\end{aligned}$$

where the last inequality holds since

$$\begin{aligned}
&\|\lambda^*_{\eta_k}-\lambda_k\|^2=\|\lambda^*_{\eta_k}-\lambda^*+\lambda^*-\lambda_k\|^2\leq 2\|\lambda_{\eta_k}-\lambda^*\|^2+2\|\lambda_k-\lambda^*\|\leq 4(b^2_\lambda+b^2_{\lambda,\eta})+2B^2_\lambda\\
&\|\lambda^*_{\eta_k}-\lambda_k\|\leq\|\lambda^*_{\eta_k}-\lambda^*\|+\|\lambda_k-\lambda^*\|\leq 2(b_\lambda+b_{\lambda,\eta})+B_\lambda
\end{aligned}$$

and let $\tilde{B}_{\lambda,1}\triangleq 4(b^2_\lambda+b^2_{\lambda,\eta})+2B^2_\lambda$ and $\tilde{B}_{\lambda,2}\triangleq 2(b_\lambda+b_{\lambda,\eta})+B_\lambda$.

Summing from $k=0$ to $K-1$ and leveraging convexity of f_{η_k} and letting $C_{\lambda,\rho}=\frac{7\tilde{B}_{\lambda,1}+2\tilde{B}_{\lambda_2}b_{\lambda,\eta}+2b^2_{\lambda,\eta}}{2\rho}$, we obtain that

$$\begin{aligned}
&f(\bar{\mathbf{x}}_K)-f^*\leq\sum_{k=0}^{K-1}\left(f(\mathbf{x}_{k+1})-f^*\right)\\
&\leq\tfrac{1}{K}\sum_{k=0}^{K-1}\left(\underbrace{f(\mathbf{x}_{k+1})-f_{\eta_k}(\mathbf{x}_{k+1})}_{\leq\eta_k B}+\underbrace{f_{\eta_k}(\mathbf{x}_{k+1})-f^*_{\eta_k}}_{(13)}+\underbrace{f^*_{\eta_k}-f_{\eta_k}(\mathbf{x}^*)}_{\leq 0}+\underbrace{f_{\eta_k}(\mathbf{x}^*)-f^*}_{\leq 0\ \text{(smoothing)}}\right)\\
&\leq\tfrac{1}{K}\left(\tfrac{\rho}{2}\left(d_{-}\left(\tfrac{\lambda_0}{\rho}\right)^2-d_{-}\left(\tfrac{\lambda_K}{\rho}\right)^2\right)\right)+\tfrac{1}{K}\sum_{k=0}^{K-1}\left(\epsilon_k\eta^b_k+C_{\lambda,\rho}+\eta_k\beta\right)\\
&\leq\tfrac{\rho}{2K}\|\tfrac{\lambda_0}{\rho}\|^2+\tfrac{1}{K}\sum_{k=0}^{K-1}\left(\epsilon_k\eta^b_k+C_{\lambda,\rho}+\eta_k\beta\right)\leq\tfrac{\tilde{C}_1}{K}+\tfrac{1}{K}\sum_{k=0}^{K-1}\left(\epsilon_k\eta^b_k+C_{\lambda,\rho}+\eta_k\beta\right)
\end{aligned}$$

where $\tilde{C}_1 > 0$. □

3.3 Rate analysis under increasing ρ_k

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

Lemma 8 (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 32, 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 33. (Rate on primal suboptimality) Suppose $\{\mathbf{x}_k, \lambda_k\}$ is generated by **Sm-AL** scheme. Then we have that

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

Proof. (i) Let $f^*_{\eta_k} \triangleq f_{\eta_k}(\mathbf{x}^*_{\eta_k})$. We have that

$$\begin{aligned}
f^*_{\eta_k} &\leq \mathcal{L}_{\eta_k,\rho_k}(\mathbf{x}_{k+1}, \lambda^*_{\eta_k}) = f_{\eta_k}(\mathbf{x}_{k+1}) + \tfrac{\rho_k}{2}\left(d_-\left(\tfrac{\lambda^*_{\eta_k}}{\rho_k} + g_{\eta_k}(\mathbf{x}_{k+1})\right)\right)^2 - \tfrac{1}{2\rho_k}\|\lambda^*_{\eta_k}\|^2 \\
&\leq f_{\eta_k}(\mathbf{x}_{k+1}) + \tfrac{\rho_k}{2}\left(d_-\left(\tfrac{\lambda_k}{\rho_k} - \tfrac{\lambda_k}{\rho_k} + \tfrac{\lambda^*_{\eta_k}}{\rho_k} + g_{\eta_k}(\mathbf{x}_{k+1})\right)\right)^2 - \tfrac{1}{2\rho_k}\|\lambda^*_{\eta_k}\|^2 \\
&\leq f_{\eta_k}(\mathbf{x}_{k+1}) + \tfrac{\rho_k}{2}\left(d_-\left(\tfrac{\lambda_k}{\rho_k} + g_{\eta_k}(\mathbf{x}_{k+1})\right) + \left\|\tfrac{\lambda_k}{\rho_k} - \tfrac{\lambda^*_{\eta_k}}{\rho_k}\right\|\right)^2 - \tfrac{1}{2\rho_k}\|\lambda^*_{\eta_k}\|^2 \\
&\leq f_{\eta_k}(\mathbf{x}_{k+1}) + \tfrac{\rho_k}{2}\left(\tfrac{\|\lambda_{k+1}\|}{\rho_k} + \left\|\tfrac{\lambda_k}{\rho_k} - \tfrac{\lambda^*_{\eta_k}}{\rho_k}\right\|\right)^2 - \tfrac{1}{2\rho_k}\|\lambda^*_{\eta_k}\|^2 \\
&\leq f_{\eta_k}(\mathbf{x}_{k+1}) + \tfrac{1}{\rho_k}\left(\|\lambda_{k+1}\|^2 + \|\lambda_k - \lambda^*_{\eta_k}\|^2\right). \qquad (14)
\end{aligned}$$

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

$$\begin{aligned}
f^* - f(\mathbf{x}_{k+1}) = &\underbrace{f^* - f(\mathbf{x}^*_{\eta_k})}_{\leq 0} + \underbrace{f(\mathbf{x}^*_{\eta_k}) - f^*_{\eta_k}}_{\leq \eta_k\beta} \\
&+ \underbrace{f^*_{\eta_k} - f_{\eta_k}(\mathbf{x}_{k+1})}_{(14)} + \underbrace{f_{\eta_k}(\mathbf{x}_{k+1}) - f(\mathbf{x}_{k+1})}_{\leq 0}.
\end{aligned}$$

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

(ii) Recall that $\mathbf{x}_{k+1}$ satisfies the following that

$$\begin{aligned}&\mathcal{L}_{\eta_k,\rho_k}(\mathbf{x}_{k+1},\lambda_k)-\min_{\mathbf{x}\in\mathcal{X}}\mathcal{L}_{\eta_k,\rho_k}(\mathbf{x},\lambda_k)\leq\epsilon_k\eta_k^b\\ \Longrightarrow\ &\mathcal{L}_{\eta_k,\rho}(\mathbf{x}_{k+1},\lambda_k)-\mathcal{L}_{\eta_k,\rho}(\mathbf{x}^*_{\eta_k},\lambda^*_{\eta_k})\leq\epsilon_k\eta_k^b.\end{aligned}$$

Moreover, since $g_{\eta_k}(\mathbf{x}^*)\leq g(\mathbf{x}^*)\leq 0$, we have that

$$\begin{aligned}&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)+\tfrac{\left\|\lambda^*_{\eta_k}-\lambda_k\right\|^2}{2\rho_k}\\ &+\|\lambda^*_{\eta_k}-\lambda_k\|\left(d\left(\tfrac{\lambda_k}{\rho_k}+g_{\eta_k}(\mathbf{x}^*_{\eta_k})\right)\right)+\epsilon_k\eta_k^b+\tfrac{1}{2\rho_k}\left(\|\lambda^*_{\eta_k}\|^2-\|\lambda_k\|^2\right)\\ &\leq\tfrac{\rho_k}{2}\left(\left\|\tfrac{\lambda_k}{\rho_k}\right\|^2-\left\|\tfrac{\lambda_{k+1}}{\rho_k}\right\|^2\right)+\tfrac{\|\lambda^*_{\eta_k}-\lambda_k\|^2}{2\rho_k}+\tfrac{\|\lambda^*_{\eta_k}-\lambda_k\|\|\lambda_k\|}{\rho_k}+\epsilon_k\eta_k^b\\ &+\tfrac{1}{2\rho_k}\left(2\|\lambda^*_{\eta_k}-\lambda_k\|^2+\|\lambda_k\|^2\right)\\ &\leq\tfrac{\rho_k}{2}\left(\left\|\tfrac{\lambda_k}{\rho_k}\right\|^2-\left\|\tfrac{\lambda_{k+1}}{\rho}\right\|^2\right)+\tfrac{\|\lambda^*_{\eta_k}-\lambda_k\|^2}{2\rho_k}+\tfrac{\|\lambda^*_{\eta_k}-\lambda_k\|^2+\|\lambda^*_{\eta_k}-\lambda_k\|\|\lambda^*_{\eta_k}\|}{\rho_k}\\ &+\tfrac{4\|\lambda^*_{\eta_k}-\lambda_k\|^2+2\|\lambda^*_{\eta_k}\|^2}{2\rho_k}+\epsilon_k\eta_k^b\\ &\leq\tfrac{\rho_k}{2}\left(\left\|\tfrac{\lambda_k}{\rho_k}\right\|^2-\left\|\tfrac{\lambda_{k+1}}{\rho_k}\right\|^2\right)+\tfrac{7\|\lambda^*_{\eta_k}-\lambda_k\|^2+2b_{\lambda,\eta}\|\lambda^*_{\eta_k}-\lambda_k\|+2b^2_{\lambda,\eta}}{2\rho_k}\\ &\leq\tfrac{\rho_k}{2}\left(\left\|\tfrac{\lambda_k}{\rho_k}\right\|^2-\left\|\tfrac{\lambda_{k+1}}{\rho_k}\right\|^2\right)+\tfrac{7\tilde{B}_{\lambda,1}+2\tilde{B}_{\lambda,2}b_{\lambda,\eta}+2b^2_{\lambda,\eta}}{2\rho_k}\\ \Longrightarrow\ &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}-f_{\eta_k}(\mathbf{x}^*)}_{\leq 0}\\ &+\underbrace{f_{\eta_k}(\mathbf{x}^*)-f^*}_{\leq 0}\leq\eta_k\beta+\tfrac{\|\lambda_k\|^2+\tilde{C}_\lambda}{2\rho_k}+\epsilon_k\eta_k^b\end{aligned}$$

where $\tilde{B}_{\lambda,1}\triangleq 4(b_\lambda^2+b_{\lambda,\eta}^2)+2B_\lambda^2$ and $\tilde{B}_{\lambda,2}\triangleq 2(b_\lambda+b_{\lambda,\eta})+B_\lambda$ and $\tilde{C}_\lambda\triangleq\left(7\tilde{B}_{\lambda,1}+2\tilde{B}_{\lambda_2}b_{\lambda,\eta}+2b^2_{\lambda,\eta}\right)$. □

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

Theorem 32. Suppose $\{\mathbf{x}_k,\lambda_k\}$ is generated by (**Sm-AL**). Let $\eta_k=\frac{1}{\rho_k}$. Then the following holds, where $\tilde{C}_1,\tilde{D}$ are constants.

$$|f(\mathbf{x}_{k+1})-f^*|\leq\eta_k\beta+\tfrac{\tilde{D}}{\rho_k}\text{ and }d_-\left(g(\mathbf{x}_{k+1})\right)\leq\eta_k\beta m+\tfrac{2\tilde{C}_1}{\rho_k}.$$

Proof. Suppose $\rho_k = \rho_0 \zeta^k$ where $\zeta > 1$. Then we have that

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

$$\leq \eta_k\beta + \frac{2\|\lambda_{k+1}\|^2 + 5\|\lambda_k\|^2 + 4\|\lambda^*_{\eta_k}\|^2 + \tilde{C}_\lambda}{2\rho_k} + \epsilon_k\eta_k^b \leq \eta_k\beta + \frac{\tilde{C}_1}{\rho_k} + \frac{1}{k^{2+\delta}\rho_k} \leq \eta_k\beta + \frac{\tilde{D}}{\rho_k}.$$

Next, we derive a rate on the expected infeasibility. Recall from Lemma 3, $g(\mathbf{x}_{k+1}) \leq g_{\eta_k}(\mathbf{x}_{k+1}) + \eta_k\beta\mathbf{1}$, implying that $d_-(g(\mathbf{x}_{k+1}) \leq d_-(g_{\eta_k}(\mathbf{x}_{k+1}) + \eta_k\beta\mathbf{1})$. Therefore,

$$d_-\left(g(\mathbf{x}_{k+1})\right) \leq d_-(g_{\eta_k}(\mathbf{x}_{k+1}) + \eta_k\beta\mathbf{1}) \leq \left\|\tfrac{\lambda_{k+1}-\lambda_k}{\rho_k}\right\| + \eta_k\beta\|\mathbf{1}\| \leq \eta_k\beta m + \frac{2\tilde{C}_1}{\rho_k}.$$

□

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.

Lemma 9 For any $\eta, \rho > 0$, $\lambda \geq 0$, there exists $\tilde{C}$ such that $\mathcal{L}_{\eta,\rho}(\bullet, \lambda)$ is $\frac{\tilde{C}\rho}{\eta}$-smooth.

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

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

where $(\mathbf{J}_g(\mathbf{x}))^\top \triangleq \left[\nabla_{\mathbf{x}} g_{\eta,1}(\mathbf{x})\ \nabla_{\mathbf{x}} g_{\eta,2}(\mathbf{x})\ \ldots\ \nabla_{\mathbf{x}} g_{\eta,m}(\mathbf{x})\right]$ and $\mathbf{J}_g(\mathbf{x})$ denotes the Jacobian matrix of $g_\eta(\mathbf{x})$. By Assumption 21 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 7, for all $\mathbf{x}_1, \mathbf{x}_2 \in \mathcal{X}$, it follows that

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

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 PZ comment: not sure how to make the following inequalities prettier

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

Consequently, $\mathcal{L}_{\eta,\rho}(\mathbf{x}, \lambda)$ is $(\frac{\alpha_4 \rho}{\eta})$-smooth by observing that

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

and the last inequality holds if $\eta \leq 1$ and $\rho \geq 1$. □

The convexity and L_k-smoothness of $\mathcal{L}_{\eta_k,\rho_k}(\bullet, \lambda_k)$ for any non-negative vector λ_k allows for proposing an accelerated scheme for inexactly resolving the augmented Lagrangian subproblem. We formally state an accelerated gradient method for resolving the augmented Lagrangian subproblem $(\text{ALSub}_{\eta_k,\rho_k}(\lambda_k))$. In particular, we have

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

Suppose $\mathbf{x}_k^*$ denotes an optimal solution of $(\text{ALSub}_{\eta_k,\rho_k}(\lambda_k))$. 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 $\mathbf{z}_0 = \mathbf{y}_0 = \mathbf{x}_k$.

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

We now restate the convergence guarantees [5,28,29] associated with (**AG**).

Theorem 41. Suppose $\mathcal{X}$ is a convex and compact set where $\|\mathbf{x} - \mathbf{y}\| \leq B$ for any $\mathbf{x}, \mathbf{y} \in \mathcal{X}$. Further, suppose $\mathcal{L}_{\eta_k,\rho_k}(\bullet, \lambda_k)$ is a convex and L_k-smooth function. 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 $j \geq 0$, where $\alpha_{-1} = 0$. Then $\mathcal{L}_{\eta_k,\rho_k}(\mathbf{y}_{j+1}, \lambda_k) - \mathcal{L}_{\eta_k,\rho_k}(\mathbf{x}_k^*, \lambda_k) \leq \frac{BL_k}{j^2}$ for any $j \geq 0$.
(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 $j \geq 0$, where $\kappa_k = L_k/\mu$ for $k \geq 0$. Then $\mathcal{L}_{\eta_k,\rho_k}(\mathbf{y}_{j+1}, \lambda_k) - \mathcal{L}_{\eta_k,\rho_k}(\mathbf{x}_k^*, \lambda_k) \leq \tilde{C}(1 - \frac{1}{\sqrt{\kappa_k}})^j$ for $j \geq 0$, where $(\mathcal{L}_{\eta_k,\rho_k}(\mathbf{x}_k, \lambda^k) - \mathcal{L}_{\eta_k,\rho_k}(\mathbf{x}_k^*, \lambda^k) + \mu B^2/2) \leq \tilde{C}$ for any k.

4.2 Complexity guarantees for convex and strongly convex f

We begin by leveraging Theorem 41 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 $\|x - y\| \leq B$ for any $x, y \in X$. Additionally, complexity guarantees are derived by utilizing the rate guarantees presented in Theorem 31 (Constant ρ_0) or Theorem 32 (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 42 [Overall complexity of Sm-AL]. Consider $\{(\mathbf{x}_k, \lambda_k)\}$ generated by (**Sm-AL**). Suppose $\rho_0, \boldsymbol{\varepsilon}, \delta > 0$, and $b \geq 0$.
(a) (Constant ρ). Let $\rho_k = \rho_0, \eta_k = k^{-(2+\delta)}, \epsilon_k = \eta_k^{-b} k^{-(2+\delta)}$, and $M_k = \left\lceil (B\tilde{C}\rho_0)^{1/2} k^{2(1+\delta)} \right\rceil$ for $k > 0$. Suppose $\left(\bar{\mathbf{x}}_K, \bar{\lambda}_K\right)$ 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 such an $\bar{\mathbf{x}}_K$ satisfies $\sum_{k=1}^{K(\boldsymbol{\varepsilon})} M_k \leq \mathcal{O}\left(\boldsymbol{\varepsilon}^{-(3+\delta)}\right)$.
(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 $M_k = \left\lceil \sqrt{B\tilde{C}} \rho_k^{3/2} k^{2+\delta} \right\rceil$ for all $k > 0$, where $\zeta > 1$. Suppose $(\mathbf{x}_K, \lambda_K)$ 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 such an $\bar{\mathbf{x}}_K$ satisfies $\sum_{k=1}^{K(\boldsymbol{\varepsilon})} M_k \leq \tilde{\mathcal{O}}(\boldsymbol{\varepsilon}^{-\frac{3}{2}})$.

Proof. (a) By Theorem 41, M_k is the smallest integer satisfying

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

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

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

(b) Proceeding similarly, by Theorem 41, M_k is defined as follows.

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

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

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

□

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

Theorem 43 [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 $\rho_0, \boldsymbol{\varepsilon}, \delta > 0$, and $b \geq 0$.

(a) (Constant ρ). Let $M_k = \left\lceil \left(\frac{\ln\left(\frac{\tilde{C}}{\epsilon_k \eta_k^b} \right)}{\ln\left(\frac{\sqrt{L_k}}{\sqrt{L_k} - \sqrt{\mu}} \right)} \right) \right\rceil$, $\rho_k = \rho_0$, $\eta_k = k^{-(2+\delta)}$, and $\epsilon_k = \eta_k^{-b} k^{-(2+\delta)}$ for all $k > 0$, where $\delta > 0$. Suppose $\left(\bar{\mathbf{x}}_K, \bar{\lambda}_K \right)$ 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 $M_k = \left\lceil \left(\frac{\ln\left(\frac{\tilde{C}}{\epsilon_k \eta_k^b} \right)}{\ln\left(\frac{\sqrt{L_k}}{\sqrt{L_k} - \sqrt{\mu}} \right)} \right) \right\rceil$, $\rho_k = \rho_0 \zeta^k$, $\eta_k = \rho_k^{-1} k^{-(2+\delta)}$, and $\epsilon_k = \rho_k^{-1} \eta_k^{-b} k^{-(2+\delta)}$ for $k > 0$, where $\delta, \rho > 0, \zeta > 1$. Suppose $(\mathbf{x}_K, \lambda_K)$ 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 $\bar{\mathbf{x}}_K$ 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 41 and $\ln(x) \geq \frac{x-1}{x}$ for $x > 0$,

$$\mathcal{L}_{\rho_k,\eta_k}(\mathbf{x}_k, \lambda_k) - \mathcal{L}_{\rho_k,\eta_k}(\mathbf{x}_k{}^*, \lambda_k) \leq \tilde{C}\left(1 - \tfrac{\sqrt{\mu}}{\sqrt{L_k}}\right)^{M_k} \leq \epsilon_k \eta_k^b$$
$$\implies M_k = \left\lceil \left(\frac{\ln\left(\tilde{C}k^{(2+\delta)}\right)}{\left(1 - \frac{\sqrt{L_k}-\sqrt{\mu}}{\sqrt{L_k}}\right)} \right) \right\rceil \leq 2\left\lceil \tfrac{1}{\sqrt{\mu\eta_k}} \ln\left((\hat{C}k)^{(2+\delta)}\right) \right\rceil, \text{ where } \hat{C} = \tilde{C}^{1/(2+\delta)}.$$

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

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

(b) Consider $\rho_k = \rho_0 \zeta^k$ where $k \geq 0$ and $\zeta > 1$. Proceeding as in (a) and by Theorem 41 and $\ln(x) \geq \frac{x-1}{x}$ for $x > 0$,

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

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

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

□

Remark 1 **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. [23,40]), i.e. $\mathcal{O}(\boldsymbol{\varepsilon}^{-1})$ (upto 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 $K = \lceil \frac{C}{\boldsymbol{\varepsilon}} \rceil$ for a suitable C. However, by Lemma 1,

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

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 41 [Complexity analysis of AL for η-smoothed convex problems]. Consider a sequence $\{(\mathbf{x}_k, \lambda_k)\}$ generated by (**Sm-AL**). Suppose $\rho_0, \boldsymbol{\varepsilon} > 0$.
(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 $M_k = \left\lceil \sqrt{\frac{B\tilde{C}\rho_0}{\eta\boldsymbol{\varepsilon}}} k^{1+\delta} \right\rceil$ for $k > 0$, where $\delta > 0$. Suppose $(\bar{\mathbf{x}}_K, \bar{\lambda}_K)$ 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 $K(\boldsymbol{\varepsilon}) = \lceil \frac{C}{\boldsymbol{\varepsilon}} \rceil$ 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 $M_k = \left\lceil \sqrt{\frac{B\tilde{C}\rho_k}{\eta\boldsymbol{\varepsilon}}} k^{1+\delta} \right\rceil$ for all $k > 0$ where $\delta, \rho_0 > 0, \eta > 1$. Suppose $(\bar{\mathbf{x}}_K, \lambda_K)$ 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 $\bar{\mathbf{x}}_K$ satisfies $\sum_{k=1}^{K(\boldsymbol{\varepsilon})} M_k \leq \tilde{O}(\boldsymbol{\varepsilon}^{-\frac{3}{2}})$.

Proof. (a.) By Theorem 41, M_k is the smallest integer satisfying

$$\mathcal{L}_{\rho_k,\eta}(\mathbf{x}_k,\lambda_k) - \mathcal{L}_{\rho_k,\eta}(\mathbf{x}_k^*,\lambda_k) \leq \left(\frac{BL_k}{M_k^2}\right) \leq \left(\frac{B\tilde{C}\rho_0}{\eta M_k^2}\right) \leq \epsilon_k$$
$$\implies M_k = \left\lceil \sqrt{\frac{B\tilde{C}\rho_0}{\epsilon_k \eta}} \right\rceil = \left\lceil \left(\sqrt{2B\tilde{C}\left(\beta(2+B_\lambda m)\right)\rho_0/\boldsymbol{\varepsilon}} \right) k^{1+\delta} \right\rceil = \left\lceil (\sqrt{D/\boldsymbol{\varepsilon}}) k^{1+\delta} \right\rceil.$$

Then the complexity of computing a $(\bar{\mathbf{x}}_K, \bar{\lambda}_K)$ where $f^* - \mathcal{D}_0(\bar{\lambda}_K) \leq \boldsymbol{\varepsilon}$ requires

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

(b) Proceeding as in (a) and by invoking Theorem 41,

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

Then the iteration complexity of producing an $\mathbf{x}_K$ satisfying $|f - f(\mathbf{x}_k)| \leq \boldsymbol{\varepsilon}$ leads to the following bound, where $C, D > 0$.

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

□

Remark 2 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 [30] and Beck and Teboulle [6] 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 [30, **?**], **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].

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.

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	$\mathcal{O}\left(\frac{1}{\sqrt{K}}\right)$	$\mathcal{O}\left(\frac{1}{\sqrt{K}}\right)$	$\mathcal{O}\left(\epsilon^{-(3+\delta)}\right)$	$\mathcal{O}\left(\frac{1}{\rho_K}\right)$	$\mathcal{O}\left(\frac{1}{\rho_K}\right)$	$\tilde{\mathcal{O}}\left(\epsilon^{-3/2}\right)$
Sm-AL(S)	$\mathcal{O}\left(\frac{1}{\sqrt{K}}\right)$	$\mathcal{O}\left(\frac{1}{\sqrt{K}}\right)$	$\tilde{\mathcal{O}}\left(\epsilon^{-(2+\delta)}\right)$	$\mathcal{O}\left(\frac{1}{\rho_K}\right)$	$\mathcal{O}\left(\frac{1}{\rho_K}\right)$	$\tilde{\mathcal{O}}\left(\epsilon^{-1}\right)$
N-AL	$\mathcal{O}\left(\frac{1}{\sqrt{K}}\right)$	$\mathcal{O}\left(\frac{1}{\sqrt{K}}\right)$	$\mathcal{O}\left(\epsilon^{-(5+\delta)}\right)$	$\mathcal{O}\left(\frac{1}{\rho_K}\right)$	$\mathcal{O}\left(\frac{1}{\rho_K}\right)$	$\tilde{\mathcal{O}}\left(\epsilon^{-4}\right)$
Sm-AL(η)	$\mathcal{O}\left(\frac{1}{\sqrt{K}}\right)$	$\mathcal{O}\left(\frac{1}{\sqrt{K}}\right)$	$\mathcal{O}\left(\epsilon^{-(5/2+\delta)}\right)$	$\mathcal{O}\left(\frac{1}{\rho_K}\right)$	$\mathcal{O}\left(\frac{1}{\rho_K}\right)$	$\tilde{\mathcal{O}}\left(\epsilon^{-3/2}\right)$

†: Dual suboptimality ★ : Primal suboptimality or Primal infeasibility

5 Numerical Experiments

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. Consider the η-smoothing of (1).

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

We conducted the experiments on simulated datasets with dimensions of β ranging from 5 to 1000. The results are shown in the 5. The optimal solutions for each experiment are obtained by using *fmincon* in Matlab. In 5, 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 1 and observe alignment with the theoretical rates represented blue lines.

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

Table 3 Numerical results

	parameters			Sm-AL			N-AL		
n	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
50	0.1	$k^{-2.01}$	4e+1	- 4.98e−6	0.00e+0	1.10e+1	2.01e−1	0.00e+0	> 7.2e+3
	1.01^k	$\frac{1}{\rho_k k^{2.01}}$	1e+3	7.00e−5	0.00e+0	1.69e+1	4.42e−2	0.00e+0	> 7.2e+3
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
5	0.1	0.1	1e+0	6.72e+2	0.00e+0				
	1.01^k	0.1	5e+2	8.70e−1	0.00e+0				
5	0.1	0.01	1e+0	1.90e−3	0.00e+0				
	1.01^k	0.01	5e+2	9.10e−3	0.00e+0				
5	0.1	0.001	1e+0	1.00e−3	0.00e+0				
	1.01^k	0.001	5e+2	6.00e−4	0.00e+0				

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

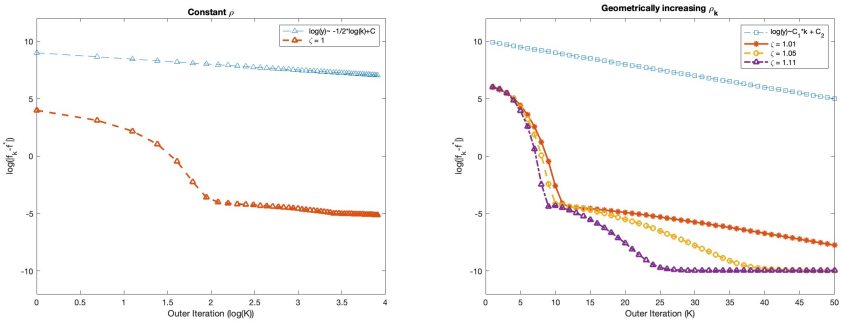

Fig. 1 Primal subopt. for fused lasso problems for constant (L) and increasing ρ_k (R)

problems and η-saddle points of our original problem. We believe that our findings represent a foundation for considering extensions to compositional regimes with expectation-valued and possibly nonsmooth constraints.

7 Declarations

P. Zhang and Uday V. Shanbhag are partially supported by ONR Grant N00014-22-1-2589 and DOE Grant DE-SC0023303

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.: First-order methods in optimization. SIAM (2017)
6. Beck, A., Teboulle, M.: Smoothing and first order methods: A unified framework. SIAM Journal on Optimization **22**(2), 557–580 (2012)
7. 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)
8. 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)
9. 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)
10. 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)
11. 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)
12. 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)
13. Friedlander, M.P., Saunders, M.A.: A globally convergent linearly constrained Lagrangian method for nonlinear optimization. SIAM J. Optim. **15**(3), 863–897 (2005)
14. Gao, B., Liu, X., Yuan, Y.x.: Parallelizable algorithms for optimization problems with orthogonality constraints. SIAM J. Sci. Comput. **41**(3), A1949–A1983 (2019)
15. Gill, P.E., Murray, W., Saunders, M.A.: SNOPT: an SQP algorithm for large-scale constrained optimization. SIAM Rev. **47**(1), 99–131 (electronic) (2005)
16. Hestenes, M.R.: Multiplier and gradient methods. Journal of Optimization Theory and Applications **4**(5), 303–320 (1969)
17. 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)
18. Kang, M., Kang, M., Jung, M.: Inexact accelerated augmented Lagrangian methods. Computational Optimization and Applications **62**(2), 373–404 (2015)
19. 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)
20. Koshal, J., Nedić, A., Shanbhag, U.V.: Multiuser optimization: Distributed algorithms and error analysis. SIAM Journal on Optimization **21**(3), 1046–1081 (2011)
21. Lan, G., Monteiro, R.D.: Iteration-complexity of first-order augmented Lagrangian methods for convex programming. Mathematical Programming **155**(1-2), 511–547 (2016)
22. 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)
23. Lu, Z., Zhou, Z.: Iteration-complexity of first-order augmented Lagrangian methods for convex conic programming. SIAM J. Optim. **33**(2), 1159–1190 (2023)
24. Moreau, J.J.: Proximité et dualité dans un espace hilbertien. Bulletin de la Société mathématique de France **93**, 273–299 (1965)
25. Murtagh, B.A., Saunders, M.A.: A projected Lagrangian algorithm and its implementation for sparse nonlinear constraints. Springer (1982)
26. 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)

27. 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)
28. Nesterov, Y.: A method for unconstrained convex minimization problem with the rate of convergence $\mathcal{O}(1/k^2)$. In: Doklady an ussr, vol. 269, pp. 543–547 (1983)
29. Nesterov, Y.: Introductory lectures on convex optimization: A basic course, vol. 87. Springer Science & Business Media (2003)
30. Nesterov, Y.: Smooth minimization of non-smooth functions. Mathematical programming **103**(1), 127–152 (2005)
31. Patrascu, A., Necoara, I., Tran-Dinh, Q.: Adaptive inexact fast augmented Lagrangian methods for constrained convex optimization. Optimization Letters **11**(3), 609–626 (2017)
32. Polyak, B.T.: Introduction to optimization (1987)
33. Powell, M.J.: A method for nonlinear constraints in minimization problems. Optimization pp. 283–298 (1969)
34. Rockafellar, R.T.: A dual approach to solving nonlinear programming problems by unconstrained optimization. Mathematical Programming **5**(1), 354–373 (1973)
35. Rockafellar, R.T.: Augmented Lagrangians and applications of the proximal point algorithm in convex programming. Mathematics of Operations Research **1**(2), 97–116 (1976)
36. Tibshirani, R.: Regression shrinkage and selection via the lasso. Journal of the Royal Statistical Society: Series B (Methodological) **58**(1), 267–288 (1996)
37. 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)
38. 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)
39. Wilson, R.B.: A simplicial algorithm for concave programming. Ph. D. Dissertation, Graduate School of Bussiness Administration (1963)
40. Xu, Y.: Iteration complexity of inexact augmented Lagrangian methods for constrained convex programming. Mathematical Programming **185**(1), 199–244 (2021)
41. F. Yousefian, A. Nedić, and U. V. Shanbhag. On stochastic gradient and subgradient methods with adaptive steplength sequences. *Automatica*, 48(1):56–67, 2012.
42. 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)
43. 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)