Math Programming Comp. manuscript No.
(will be inserted by the editor)

Linearized augmented Lagrangian framework for solving non-convex problems

Ahmet/Fatih · Armin · Volkan · who else?

Received: / Accepted:

Abstract To be written...

Keywords Primal-dual · Non-linear constraints · Non-convex

This project has received funding from...

Address(es) of author(s) should be given

1 Introduction

We study the nonconvex optimization program

$$\begin{cases} \min\limits_{x} \ f(x) + g(x) \\ A(x) = 0, \end{cases} \tag{1}$$

where (possibly nonconvex) $f : \mathbb{R}^d \to \mathbb{R}$ and (possibly nonlinear) $A : \mathbb{R}^d \to \mathbb{R}^m$ satisfy

$$\|\nabla f(x) - \nabla f(x')\| \le \lambda_f \|x - x'\|, \quad \|\,\mathrm{D}\, A(x) - \mathrm{D}\, A(x')\| \le \lambda_A \|x - x'\|, \tag{2}$$

for every $x, x' \in \mathbb{R}^d$. Above, $\nabla f(x) \in \mathbb{R}^d$ is the gradient of f at x and $\mathrm{D}\, A(x) \in \mathbb{R}^{m\times d}$ is the Jacobian of A at x. Moreover, we assume that $g : \mathbb{R}^d \to \mathbb{R}$ is a proximal-friendly (but possibly nonsmooth) convex function.

A host of problems in computer science [**?**, **?**], machine learning [**?**, **?**], and signal processing [**?**, **?**] naturally fall under the template of (1), including max-cut, clustering, generalized eigenvalue, as well as community detection.

To address these applications, this paper builds up on the classical ideas in linearized augmented Lagrangian framework and proposes a simple, intuitive, and easy-to-implement algorithm to solve 1 with provable convergence rate and under an interpretable geometric condition. In this context, we also develop and analyze the Alternating Direction Method of Multipliers (ADMM). Before we elaborate on the results, let us first motivate (1) with an important application to Semi-Definite Programming (SDP):

Vignette: Burer-Monteiro splitting. A powerful convex relaxation for max-cut, clustering, and several other problems mentioned above is provided by the SDP

$$\begin{cases} \min\limits_{X\in\mathbb{S}^{d\times d}} \langle C, X\rangle \\ B(X) = b, \ X \succeq 0, \end{cases} \tag{3}$$

where $C \in \mathbb{R}^{d\times d}$ and X is a positive semidefinite and symmetric $d \times d$ matrix, and $B : \mathbb{S}^{d\times d} \to \mathbb{R}^m$ is a linear operator. If the unique-games conjecture is true, SDPs achieve the best approximation for the underlying discrete problem [**?**].

Since d is often large, many first- and second-order methods for solving such SDPs are immediately ruled out, not only due to their high computational complexity, but also due to their storage requirements, which are $\mathcal{O}(d^2)$.

A contemporary challenge in optimization therefore is to solve SDPs in small space and in a scalable fashion. A recent algorithm, namely, homotopy conditional gradient method based on Linear Minimization Oracles (LMO), can address this template in small space via sketching [**?**]; however, such LMO-based methods are extremely slow in obtaining accurate solutions.

A key approach for solving (1), dating back to [**?**, **?**], is the so-called Burer-Monteiro (BR) factorization $X = UU^\top$, where $U \in \mathbb{R}^{d\times r}$ and r is selected according to the guidelines in [**?**, **?**]. **AE: maybe we should call this factorization vs splitting following the standard references like Global**

Optimality in Tensor Factorization, Deep Learning, and Beyond. This factorization results in the following nonconvex problem

$$\begin{cases} \min\limits_{U\in\mathbb{R}^{d\times r}} \langle C, UU^\top\rangle \\ B(UU^\top) = b, \end{cases} \tag{4}$$

which can be written in the form of (1). When r is sufficiently large and under some additional assumptions, (3) provably does not have any spurious local minima [**?**, **?**].

The augmented Lagrangian method [**?**] provides a powerful framework to solve (1), reviewed carefully in Section 7. Indeed, for positive β, it is easy to verify that (1) is equivalent to

$$\min_x \max_y \ \mathcal{L}_\beta(x, y) + g(x), \tag{5}$$

where

$$\mathcal{L}_\beta(x, y) := f(x) + \langle A(x), y\rangle + \frac{\beta}{2}\|A(x) - b\|^2, \tag{6}$$

is the augmented Lagrangian corresponding to (1). The equivalent formulation in (5) naturally suggests the following iterative algorithm to solve (1):

$$x_{k+1} \in \operatorname*{argmin}_x \mathcal{L}_\beta(x, y_k) + g(x), \tag{7}$$

$$y_{k+1} = y_k + \sigma_k A(x_{k+1}). \tag{8}$$

Updating x above requires solving the nonconvex problem (7) to global optimality, which is often intractable. The key contribution of this paper is to provably and efficiently address this challenge by proposing and analyzing a linearized augmented Lagrangian algorithm, as well as its ADMM variant.

Contributions. In order to solve (1), this paper proposes to replace the (intractable) problem (7) with the simple update

$$x_{k+1} = P_g(x_k - \gamma_k \nabla \mathcal{L}_{\beta_k}(x_k, y_k)), \tag{9}$$

for carefully selected sequences $\{\beta_k, \gamma_k\}_k$. Here, P_g is the proximal operator corresponding to g, which is often computationally inexpensive.

Put differently, instead of fully solving (7), this paper proposes to apply one iteration of the proximal gradient algorithm for every primal update, which is then followed by a dual update in (8) and an increase in the penalty weight β to gradually enforce the (nonlinear) constraints in (1).

We prove that this fast and scalable Homotopy Linearized Augmented Lagrangian (HoLAL) achieves first-order stationarity for (1) at the rate of $1/\sqrt{k}$. Under standard additional conditions, we also establish local optimality, namely, HoLAL achieves second-order stationarity for (1). We also provide an ADMM variant of HoLAL, with the same convergence rate, which is better suited for a variety of problems that require splitting. **AE: for example?**

AE: How do the rates compare with competitors? Any high level advantage we might have over competitors? Like easy implementation guidelines?

As with several other nonconvex solvers, the success of of HoLAL relies on (a variant of) the *uniform regularity* **AE: what to cite?**, a geometric condition akin to the well-established Slater's condition in convex optimization. In fact, we establish that uniform regularity, when limited to convex problems, is equivalent to the Slater's condition. We also verify the uniform regularity in several important examples.

2 Preliminaries

Notation. We use the notations $\langle \cdot, \cdot \rangle$ and $\| \cdot \|$ for the standard inner product and norm on $\mathbb{R}^d$, respectively. Gradient of differentiable $f : \mathbb{R}^d \to \mathbb{R}$ at x is denoted by $\nabla f(x)$. For a differentiable map $A : \mathbb{R}^d \to \mathbb{R}^m$, $\mathrm{D}\, A(x$ denote its Jacobian at x. For a convex function $g : \mathbb{R}^d \to \mathbb{R}$, the subdifferential at x is denoted by $\partial g(x)$ and the proximal operator $P_g : \mathbb{R}^d \to \mathbb{R}^d$ takes x to

$$P_g(x) = \operatorname*{argmin}_y g(y) + \frac{1}{2}\|x - y\|^2. \tag{10}$$

In addition, if $g = 1_C$ is the indicator function of a convex set or cone, we use the simpler notation P_C, instead of P_{1_C}, to denote the orthogonal projection onto C. Throughout, g^* and ∂g^* will denote the Fenchel conjugate of g and its subdifferential, respectively. For a cone C, we denote its polar by C^*, namely,

$$C^* = \{x : \langle x, x' \rangle \leq 0, \ \forall x' \in C\}. \tag{11}$$

An integer interval is denoted by $[k_0 : k_1] = \{k_0, \cdots, k_1\}$ for integers $k_0 \leq k_1$. For matrices, $\| \cdot \|$ and $\| \cdot \|_F$ denote the spectral and Frobenius norms, respectively.

Necessary optimality conditions. Necessary optimality conditions for (1) are well-studied [**?**]. Indeed, x is a first-order stationary point of (1) if there exists $y \in \mathbb{R}^m$ for which

$$\begin{cases} -\nabla f(x) - \mathrm{D}\, A(x)^\top y \in \partial g(x) \\ A(x) = 0. \end{cases} \tag{12}$$

Recalling (6), we observe that (12) is equivalent to

$$\begin{cases} -\nabla_x \mathcal{L}_\beta(x, y) \in \partial g(x) \\ A(x) = 0, \end{cases} \tag{13}$$

which is in turn the first-order optimality condition for (5). For second-order optimality conditions, we set $g = 0$ in (1) and assume that both f, A are twice-differentiable. In this setting and after recalling (6), x is a local minimum of (1) if there exists $y \in \mathbb{R}^m$ such that

$$\nabla^2_{xx}\mathcal{L}_0(x, y) = \nabla^2 f(x) + \sum_{i=1}^{m} y_i \nabla^2 A_i(x), \tag{14}$$

is positive semidefinite. Above, y_i and A_i are the i^{th} components of y and A, respectively.

Technical lemmas. The following standard results and notions are frequently used throughout this paper and proved in the appendix for completeness. The first result below shows that the augmented Lagrangian is smooth, see Appendix C for the proof.

Lemma 1 (Smoothness) *For fixed $y \in \mathbb{R}^m$ and $\beta, \rho, \rho' \geq 0$, it holds that*

$$\|\nabla_x \mathcal{L}_\beta(x, y) - \nabla_x \mathcal{L}_\beta(x', y)\| \leq \lambda_\beta \|x - x'\|, \quad \forall x, x' \in X_{\rho,\rho'}, \tag{15}$$

where

$$X_{\rho,\rho'} := \{x'' : \|A(x'')\| \leq \rho, \|x''\| \leq \rho'\} \subset \mathbb{R}^d, \tag{16}$$

$$\lambda_\beta := \lambda_f + \sqrt{m}\lambda_A \left(\|y\| + \beta\rho\right) + \beta d \lambda_A'^2, \tag{17}$$

$$\lambda'_A := \max_{\|x\| \leq \rho'} \|\,\mathrm{D}\,A(x)\|, \tag{18}$$

and λ_f, λ_A were defined in (2).

Gradient mapping **AE: what to cite?**, defined below, plays an important role in our convergence analysis.

Definition 1 (Gradient mapping) Given $y \in \mathbb{R}^d$ and $\gamma > 0$, the gradient mapping $G_{\beta,\gamma}(\cdot; y) : \mathbb{R}^d \to \mathbb{R}^d$ takes $x \in \mathbb{R}^d$ to

$$G_{\beta,\gamma}(x, y) = \frac{x - x^+}{\gamma}, \tag{19}$$

where $x^+ = P_g(x - \gamma \nabla_x \mathcal{L}_\beta(x, y))$.

As the name suggests, if in particular we set $g \equiv 0$ in (1), the gradient mapping reduces to $G_{\beta,\gamma}(x, y) = \nabla f(x)$. Note also that $G_{\beta,\gamma}(x, y) = 0$ implies that $-\nabla_x \mathcal{L}_\beta(x, y) \in \partial g(x)$. Therefore, in light of (13), a linear combination of $\|G_{\beta,\gamma}(x, y)\|^2$ and the feasibility gap $\|A(x)\|^2$ is a natural metric to measure the (first-order) stationarity of a pair (x, y) in problem (1).

For a sufficiently small step size γ, the gradient mapping controls the descent in the objective function of (5). The following result is standard [**?**, Lemma 3.2, Remark 3.2(i)], but the proof is given in Appendix D for completeness.

Lemma 2 (Descent lemma) *For $x \in \mathbb{R}^d$ and $y \in \mathbb{R}^m$, let $x^+ = P_g(x - \gamma \nabla_x \mathcal{L}_\beta(x, y))$, where $\gamma < 1/\lambda_\beta$. For $\rho, \rho' \geq 0$, suppose that*

$$x, x^+ \in X_{\rho,\rho'} = \{x' : \|A(x')\| \leq \rho, \|x'\| \leq \rho'\}. \tag{20}$$

Then it holds that

$$\|G_{\beta,\gamma}(x, y)\|^2 \leq \frac{2}{\gamma}(\mathcal{L}_\beta(x, y) + g(x) - \mathcal{L}_\beta(x^+, y) - g(x^+)). \tag{21}$$

In practice, determining the step size γ by computing the right-hand side of (17) is infeasible, since $\lambda_f, \lambda_A, \lambda'_A$ are often unknown. Instead, we can resort to the line search technique, reviewed below and proved in Appendix E.

Lemma 3 (Line search) *Fix $\theta \in (0, 1)$ and $\gamma_0 > 0$. For $\gamma' > 0$, let*

$$x^+_{\gamma'} = P_g(x - \gamma' \nabla_x \mathcal{L}_\beta(x, y)), \tag{22}$$

and define

$$\gamma := \max \Big\{ \gamma' = \gamma_0 \theta^i : \mathcal{L}_\beta(x^+_{\gamma'}, y) \\ \leq \mathcal{L}_\beta(x, y) + \Big\langle x^+_{\gamma'} - x, \nabla_x \mathcal{L}_\beta(x, y) \Big\rangle + \frac{1}{2\gamma'} \|x^+_{\gamma'} - x\|^2 \Big\}. \tag{23}$$

Then, (6) holds and, moreover, we have that

$$\gamma \geq \frac{\theta}{\lambda_\beta}. \tag{24}$$

3 Uniform Regularity

The Slater's condition plays a key role in convex optimization as a sufficient condition for strong duality. As a result, SC guarantees the success of a variety of primal-dual algorithms for constrained convex programming. As a visual example, in problem (1), when $f = 0$, $g = 1_C$ is the indicator function of a bounded convex set $C \subset \mathbb{R}^d$, and A is an affine operator, the Slater's condition removes any pathological cases, such as Figure 1, by ensuring that the affine subspace is not tangent to C.

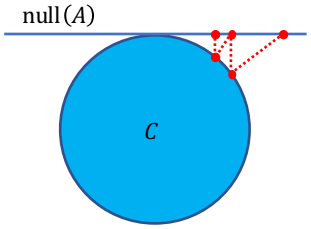

Fig. 1 Solving (1) can be particularly difficult, even when it is a convex program. As an example, this figure shows a pathological case, where the Slater's condition does not apply. See Section 3 for more details.

Likewise, to successfully solve problem (1) in the presence of nonlinear constraints, we require the following condition which, loosely speaking, extends the Slater's condition to the nonconvex setting, as clarified shortly afterwards. This condition, in a sense, also extends the uniform regularity, introduced in [**?**, Definition 2.3], to the more general problem 1.

Definition 2 (Uniform regularity) In problem (1), for $\rho, \rho' > 0$ and subspace $S \subset \mathbb{R}^d$, let us define

$$\nu(g, A, S, \rho, \rho') := \begin{cases} \min\limits_{v,x} \frac{\left\| P_S P_{\mathrm{cone}(\partial g(x))^*} (\mathrm{D}\, A(x)^\top v) \right\|}{\|v\|} \\ \|v\| \le \rho \\ \|x\| \le \rho', \end{cases} \tag{25}$$

where $\mathrm{cone}(\partial g(x))$ is the cone formed by the subdifferential $\partial g(x)$, $P_{\mathrm{cone}(\partial g(x))^*}$ projects onto the polar of this cone, and $\mathrm{D}\, A(x)$ is the Jacobian of A. We say that (1) satisfies the uniform regularity if $\nu(g, A, S, \rho, \rho') > 0$.

Throughout, we will occasionally suppress the dependence of ν on some of its parameters to unburden the notation. A few remarks about uniform regularity are in order.

Jacobian $\mathrm{D}\, A$**.** Let $\mathrm{D}\, A(x)^\top \overset{\mathrm{QR}}{=} Q(x)R(x)$ be the QR decomposition of $\mathrm{D}\, A(x)^\top$. As we will see shortly, $\mathrm{D}\, A(x)^\top$ in (25) might be replaced with its orthonormal basis, namely, $Q(x)$, to broaden the applicability of uniform regularity. For simplicity, we will avoid this minor change and instead, whenever needed, assume that $\mathrm{D}\, A(x)$ is nonsingular; otherwise a simple QR decomposition can remove any redundancy from $A(x) = 0$ in (1).

Subspace S**.** The introduction of a subspace S in (25) broadens the applicability of uniform regularity, as we will see shortly. In particular, when $S = \mathbb{R}^d$, the Moreau decomposition allows us to rewrite (25) as

$$\nu(g, A, S, \rho, \rho') := \begin{cases} \min\limits_{v,x} \frac{\mathrm{dist}\left(\mathrm{D}\, A(x)^\top v, \mathrm{cone}(\partial g(x))\right)}{\|v\|} \\ \|v\| \le \rho \\ \|x\| \le \rho', \end{cases} \tag{26}$$

where $\mathrm{dist}(\cdot, \mathrm{cone}(\partial g(x)))$ returns the Euclidean distance to $\mathrm{cone}(\partial g(x))$.

Convex case. To better parse Definition 2, let us consider the specific example where $f : \mathbb{R}^d \to \mathbb{R}$ is convex, $g = 1_C$ is the indicator function for a bounded convex set $C \subset \mathbb{R}^d$, and A is a nonsingular linear operator represented with the full-rank matrix $A \in \mathbb{R}^{m \times d}$. We also let $T_C(x)$ denote the tangent cone to C at x, and reserve $P_{T_C(x)} : \mathbb{R}^d \to \mathbb{R}^d$ for the orthogonal projection onto this cone.

We can now study the geometric interpretation of uniform regularity in this setting. Using the Moreau decomposition, it is not difficult to rewrite (25) as

$$\nu(g, A, S, \rho, \rho') := \begin{cases} \min\limits_{v,x} \frac{\left\| P_S P_{T_C(x)} A^\top v \right\|}{\|v\|} \\ \|v\| \leq \rho \\ \|x\| \leq \rho' \end{cases}$$
$$= \begin{cases} \min\limits_{x} \eta_{\min}\left(P_S P_{T_C(x)} A^\top \right) \\ \|x\| \leq \rho', \end{cases} \tag{27}$$

where $\eta_{\min}(\cdot)$ returns the smallest singular value of its input matrix. Intuitively then, the uniform regularity ensures that the row span of A is not tangent to C, similar to the Slater's condition, see Figure 1. This close relationship between the uniform regularity and the Slater's condition is formalized next and proved in Appendix F.

Proposition 1 *In (1), suppose that*

- $f : \mathbb{R}^d \to \mathbb{R}$ *is convex,*
- $g = 1_C$ *is the indicator on a convex set* $C \subset \mathbb{R}^d$*,*
- $A : \mathbb{R}^d \to \mathbb{R}^m$ *is a nonsingular linear operator, represented with the full-rank matrix* $A \in \mathbb{R}^{m \times d}$*,*[1]
- *and the problem is feasible, namely, there exists* $x \in C$ *such that* $Ax = 0$*.*

Then,

- *problem (1) satisfies the Slater's condition if there exists a subspace* $S \subseteq \mathbb{R}^d$ *such that* $\nu(g, A, S, \infty, \max_{x \in C} \|x\|) > 0$*.*
- *Moreover, suppose that* S *is the affine hull of* C*. Then, (1) satisfies the Slater's condition if and only if* $\nu(g, A, S, \infty, \max_{x \in C} \|x\|) > 0$*.*

Beyond the Slater's condition. Unlike the Slater's condition, ν also offers information about the convergence speed. For example, suppose that $m = 1$, so that A is a $1 \times d$ row-vector. For a small perturbation vector $\epsilon \in \mathbb{R}^d$, let $C = \{x \in \mathbb{R}^d : (A + \epsilon)x \geq 0\}$ be a half-space. Then the Slater's condition holds regardless of $\|\epsilon\|$. However, even though positive, $\nu(g, A, \mathbb{R}^d)$ can be made arbitrarily small by making $\|\epsilon\|$ small, which can lead to arbitrarily slow convergence.

In this work, we will focus on instances of problem (1) that satisfy the uniform regularity condition. To solve such problems, next section introduces and studies the HoLAL algorithm. **AE: maybe we should say that this has precedent in the literature and that we're not limiting ourselves too much.**

[1] As mentioned earlier, it is easy to remove the full-rank assumption by replacing $\mathrm{D}\, A(x)$ in (25) with its orthonormal basis. We assume A to be full-rank for the sake of clarity at the cost of a simple QR decomposition to remove any "redundant measurements" from problem (1).

4 Linearized AL Algorithm

To solve the equivalent formulation of problem (1) presented in (5), we propose a Homotopy Linearized Augmented Lagrangian algorithm (HoLAL), detailed in Algorithm 4. At every iteration, Algorithm 4 takes a primal descent small followed by a dual ascent step. The increasing sequence of penalty weights $\{\beta_k\}_k$ and the dual updates (Steps 6 and 7) are responsible for continuously enforcing the constraints in (1).

As we will see in the convergence analysis, the particular choice of β_k in Algorithm 4 strikes a balance between reducing the objective of (1) and enforcing its constraints. Moreover, the choice of dual step size σ_k in Algorithm 4 ensures that the dual variable y_k remains bounded; see [**?**] for a precedent in the literature of augmented Lagrangian method with a similar choice for the dual step size.

AE: we should at least say something about the adaptive version, even if we don't include the proof. maybe we should include the algorithm, claim we have proved its convergence (which we did) and then use it in simulations alongside the nonadaptive version.

Algorithm 1 HoLAL algorithm for solving problem (1)

Input: Parameters $\beta_1, \sigma_1, \rho, \tau > 0$, primal initialization $x_1 \in \mathbb{R}^d$ with $\|A(x_1)\| \leq \rho$, dual initialization $y_1 \in \mathbb{R}^m$.

For $k = 1, 2, \ldots$ and until convergence, execute

1. **(Update penalty weight)** $\beta_k \leftarrow \beta_1 \sqrt{k} \log(k+1)/\log 2$.
2. **(Line search)** Use the line search in (23) with $x = x_k, y = y_k, \beta = \beta_k$ and let $\gamma_k \leftarrow \gamma$.
3. **(Primal descent step)** $x_{k+1} \leftarrow P_g(x_k - \gamma_k \nabla \mathcal{L}_{\beta_k}(x_k, y_k))$, where $\mathcal{L}_{\beta_k}$ is the augmented Lagrangian and P_g denotes the proximal operator, defined in (6,10), respectively.
4. **(Stopping criterion)** If $\gamma_k \|G_{\beta_k,\gamma_k}(x_k, y_k)\|^2 + \sigma_k \|A(x_k)\|^2 \leq \tau$, then quit and return x_{k+1}. See (19) for the definition of G_{β_k,γ_k}.
5. **(Update dual step size)** $\sigma_{k+1} \leftarrow \sigma_1 \min\left(\frac{1}{\sqrt{k+1}}, \frac{\|A(x_1)\|}{\|A(x_{k+1})\|} \cdot \frac{\log^2 2}{(k+1)\log^2(k+2)}\right)$.
6. **(Dual ascent step)** $y_{k+1} \leftarrow y_k + \sigma_{k+1} A(x_{k+1})$.

At iteration k, if the primal step size γ_k is sufficiently small, Step 3 of Algorithm 4 reduces the objective of (5). Uniform regularity ensures that this update also reduces the feasibility gap of (1). This intuition is formalized below and proved in Appendix G.

Lemma 4 *For integers $k_0 < k_1$, consider the integer interval $K = [k_0 : k_1]$. Suppose that problem (1) satisfies uniform regularity and, more specifically,*

$$\nu(g, A, S, \rho, \rho') \geq 2\lambda_A \max_{k \in K} \gamma_k \|G_{\beta_k,\gamma_k}(x_k, y_k)\|, \tag{28}$$

where λ_A was defined in (2) and

- $\rho \geq \max_{k\in K} \|A(x_k)\|$,
- $\rho' \geq \max_{k\in K} \|x_k\|$,
- $S \supseteq \bigcup_{k\in K} P_{\mathrm{cone}(\partial g(x_{k+1}))^*} \left(\mathrm{D}\, A(x_{k+1})^\top A(x_{k+1})\right)$.

Then, for every $k \in K$, it holds that

$$\|A(x_k)\| \leq \frac{2}{\nu(g, A, \rho, \rho')\beta_k} \left(\|G_{\beta_k,\gamma_k}(x_k, y_k)\| + \lambda'_f + \lambda'_A \|y_k\|\right), \tag{29}$$

where

$$\lambda'_f := \max_{\|x\|\leq\rho'} \|\nabla f(x)\|, \qquad \lambda'_A := \max_{\|x\|\leq\rho'} \|\,\mathrm{D}\, A(x)\|. \tag{30}$$

Loosely speaking, as the penalty weight β_k increases, the feasibility gap in (1) reduces, as indicated in (29). Note that the larger ν, the more regular problem (1) is, and the smaller feasibility gap becomes. With the aid of Lemma 4, we can derive the convergence rate of Algorithm 4 to a first-order stationary point, with the proof deferred to Appendix A. For the convergence metric, we will use linear combination of the gradient mapping and the feasibility gap of problem (1), as motivated after Definition 1.

Theorem 1 (Convergence rate of HoLAL) *For sufficiently large integers $k_0 < k_1$, consider the interval $K = [k_0 : k_1]$, and consider the output sequence $\{x_k, y_k\}_{k\in K}$ of Algorithm 4. Suppose that*

$$\mu := -\min(0, \inf_k f(x_k) + g(x_k) + \langle A(x_k), y_{k_0}\rangle) < \infty.$$

AE: note that mu and some other quantities are slightly changed from the proof to simplify the presentation which is inconsequential for the proof I think.*For $\rho' \gtrsim \sqrt{\mu}$, in addition to the strong smoothness of f and A quantified in (2), let us define*

$$\lambda'_f = \max_{\|x\|\leq\rho'} \|\nabla f(x)\|, \qquad \lambda'_A = \max_{\|x\|\leq\rho'} \|\,\mathrm{D}\, A(x)\|, \tag{31}$$

to be the (restricted) Lipschitz constants of f and A, respectively. Suppose also that problem (1) satisfies uniform regularity and, more specifically,

$$\nu(g, A, S, \rho, \rho') \gtrsim \max\left(\lambda_A \max_{k\in K} \sqrt{\gamma_k \mu}, \frac{\lambda'_f + \lambda'_A}{\sqrt{\mu}}\right), \tag{32}$$

with

- $\rho' \geq \max_{k\in K} \|x_k\|$,
- $S \supseteq \bigcup_{k\in K} P_{\mathrm{cone}(\partial g(x_{k+1}))^*} \left(\mathrm{D}\, A(x_{k+1})^\top A(x_{k+1})\right)$.

Then the output of Algorithm 4 satisfies

$$\min_{k\in K} \frac{\|G_{\beta_K,\gamma_k}(x_k,y_k)\|^2}{\lambda_A\rho+\lambda_A'^2}\sqrt{\frac{k_0\log^2(k_0+1)}{k_1\log^2(k_1+1)}}+\|A(x_k)\|^2$$
$$\lesssim \frac{1}{k_1-k_0}\left(\frac{\lambda_f'^2+\lambda_A'^2}{\nu(g,A,S,\rho,\rho')^2}+\mu\right), \tag{33}$$

where $\lesssim,\gtrsim$ above suppress the dependence on less important parameters, for the sake of clarity. The exact expressions are found in (76,79,82).

A few remarks about Theorem 1 are in order.

Convergence rates. Loosely speaking, Theorem 1 states that Algorithm 4 achieves first-order stationarity for (1) by reducing the gradient map and the feasibility gap at the rates

$$\|G_{\beta_k,\gamma_k}(x_k,y_k)\|^2=\frac{1}{\widetilde{O}(\sqrt{k})},\qquad \|A(x_k)\|=\frac{1}{\widetilde{O}(\sqrt{k})}. \tag{34}$$

AE: how does this rate compare with others?

Uniform regularity. As confirmed by (48), the larger $\nu(g,A,S,\rho,\rho')$, the more regular (1), and the faster convergence rate of Algorithm 4. In fact, for Algorithm 4 to succeed, Theorem 1 requires ν to be sufficiently large (rather than just positive). We do not know if this is an artifact of the proof of technique or a fundamental problem but it is naturally expected for the convergence rate to at least slow down when ν decreases.

The right-hand side of (47) also depends on the largest primal step size $\max_{k\in K}\gamma_k$. Since γ_k is found by line search in Algorithm 4, we are unable to upper bound this quantity unless we make further assumptions on problem (1), or slightly modify the algorithm to cap primal step sizes. However, recall that the augmented Lagrangian $\mathcal{L}_{\beta_k}(\cdot,y_k)$ is λ_{β_k} Lipschitz gradient and thus typically $\gamma_k\approx 1/\lambda_{\beta_k}$, namely, $\gamma_k\approx 1/\sqrt{k}$ by (73,74).

Lastly note that smoother f,A also improve the convergence rate, see (47,48). Indeed, as f,A becomes smoother, problem (1) more and more resembles a convex program, at least locally.

Subspace S. The freedom over the choice of subspace S specified in Theorem 1 is meant to further strengthen the result in the same spirit of the second result in Proposition 1.

Faster rates. Linear convergence to a global minimizer of problem (1) can be established for Algorithm 4 under restricted strong convexity and smoothness for f in (1) and certain geometric regularities for A therein. **AE: cite paper with Fabian.**

AE: what else should we talk about? what would the reviewers ask? what would better clarify the result for average reader?

5 Local Optimality

Theorem 1 establishes that HoLAL, being a first-order algorithm that does not use any second-order information, achieves first-order stationarity for problem (1) but remains silent about local optimality. As shown in [**?**], finding approximate second-order stationary points of convex-constrained problems is in general NP-hard. For this reason, we focus in this section on the special case of problem (1) with $g = 0$.

Special case. As an important special case of problem (1), if f is strongly convex and the manifold $\{x : A(x) = 0\}$ is smooth enough, then any first-order stationary point of problem (1) is also a local minimum. Intuitively, this happens because the second-order terms of the Lagrangian are locally dominated by those of f. A concrete example is the factorized SDP in (4), when C is positive definite. More formally, suppose that f is strongly convex and both f, A are twice differentiable. For a feasible pair (x, y) in (1), recall from (14) that

$$\begin{aligned}
\nabla^2_{xx}\mathcal{L}_0(x,y) &= \nabla^2 f(x) + \sum_{i=1}^m y_i \nabla^2 A_i(x) \\
&\succcurlyeq \nabla^2 f(x) - \|\sum_{i=1}^m y_i \nabla^2 A_i(x)\| \qquad \text{(Weyl's inequality)} \\
&\succcurlyeq \nabla^2 f(x) - \|y\|_1 \cdot \max_i \|\nabla^2 A_i(x)\| \\
&\succcurlyeq \nabla^2 f(x) - \sqrt{m}\|y\|_2 \cdot \max_i \|\nabla^2 A_i(x)\|. \qquad (35)
\end{aligned}$$

Therefore, if the last line above is positive definite, then x is a local minimum of problem (1). In particular, the proof of Theorem 1 establishes that the output sequence (x_k, y_k) of Algorithm 4 satisfies $\|y_k\| \le y_{\max}$, where $y_{\max}$ is specified in (71). As such, we conclude that the output sequence of Algorithm 4 reaches second-order optimality if

$$\min_{\|x\| \le \rho'} \eta_{\min}(\nabla^2 f(x)) > \sqrt{m} y_{\max} \cdot \max_i \max_{\|x\| \le \rho'} \|\nabla^2 A_i(x)\|, \qquad (36)$$

namely, if f is sufficiently strongly convex and $\{A_i\}_i$ are sufficiently smooth.

General case. More generally, if every saddle point of (1) is strict, then one can extend the analysis of [**?**] to show that Algorithm 4 almost surely does not converge to a saddle point. **AE: haven't actually verified this. should we go down this route?**

AE: beyond this, what else can we say? maybe we can talk about how in a lot of nonconvex problems (2nd order) local optimality implies global optimality. but that's not the focus of this paper and we should mention it in passing.

6 Linearized ADMM

In convex optimization, whenever applicable, Alternating Direction Method of Multipliers (ADMM) [**?**, **?**, **?**] often outperforms the augmented Lagrangian method. Additionally, ADMM often more efficiently incorporates any proximal operators. **AE: this is very vague and hand-wavy. what to say?** Inspired by the success of ADMM in convex optimization, in this section we develop and study a (linearized) ADMM variant of Algorithm 4. More specifically, consider the program

$$\begin{cases} \min\limits_{x,z} f(x) + g(x) + h(z) + l(z) \\ A(x) + B(z) = 0, \end{cases} \tag{37}$$

where $f, h : \mathbb{R}^d \to \mathbb{R}$ and $A, B : \mathbb{R}^d \to \mathbb{R}^d$ are smooth in the sense that

$$\|\nabla f(x) - \nabla f(x')\| \le \lambda_f \|x - x'\|, \ \|DA(x) - DA(x')\| \le \lambda_A \|x - x'\|,$$

$$\|\nabla h(z) - \nabla h(z')\| \le \lambda_h \|z - z'\|, \ \|DB(z) - DB(z')\| \le \lambda_B \|z - z'\|, \tag{38}$$

for every $x, x', z, z' \in \mathbb{R}^d$. Above, $g, l : \mathbb{R}^d \to \mathbb{R}$ are proximal-friendly convex functions. **AE: should we give a "vignette" here too? what would it be?** For penalty weight $\beta \ge 0$, the augmented Lagrangian corresponding to problem (37) is

$$\mathcal{L}_\beta(x, z, y) = f(x) + h(z) + \langle A(x) + B(z), y \rangle + \frac{\beta}{2} \|A(x) + B(z)\|^2, \tag{39}$$

and problem (37) is therefore equivalent to the minimax program

$$\min_{x,z} \max_{y} \mathcal{L}_\beta(x, z, y). \tag{40}$$

To solve the equivalent formulation in (40), we propose the linearized ADMM, detailed in Algorithm 6. Most remarks about Algorithm 4 apply to Algorithm 6 as well and, in particular, note that Algorithm 6 performs two consecutive primal updates, one on x and then one on z.

To parse the details of Algorithm 6, we need to slightly change the gradient map in Definition 1 and the line search procedure in Lemma 3 to match $\mathcal{L}_\beta$ in (39). More specifically, the corresponding gradient maps are defined as

$$G_{\beta,\gamma}(x, z, y) = \frac{x - x^+}{\gamma}, \qquad H_{\beta,\iota}(x, z, y) = \frac{z - z^+}{\iota}, \tag{41}$$

where

$$x^+ = P_g(x - \gamma \nabla_x \mathcal{L}_\beta(x, z, y)), \qquad z^+ = P_l(z - \iota \nabla_z \mathcal{L}_\beta(x, z, y)), \tag{42}$$

and $\gamma, \iota > 0$ are the primal step sizes. The line search procedure too is similar to Lemma 3 and we set

$$x^+_{\gamma'} = P_g(x - \gamma' \nabla_x \mathcal{L}_\beta(x, z, y)),$$

$$z_{\iota'}^{+} = P_l(z - \iota' \nabla_z \mathcal{L}_\beta(x, z, y)), \tag{43}$$

$$\gamma := \max \Big\{ \gamma' = \gamma_0 \theta^i : \mathcal{L}_\beta(x_{\gamma'}^{+}, z, y) \leq \mathcal{L}_\beta(x, z, y) + \Big\langle x_{\gamma'}^{+} - x, \nabla_x \mathcal{L}_\beta(x, z, y) \Big\rangle + \frac{1}{2\gamma'} \|x_{\gamma'}^{+} - x\|^2 \Big\}, \tag{44}$$

$$\iota := \max \Big\{ \iota' = \iota_0 \theta^i : \mathcal{L}_\beta(x, z_{\iota'}^{+}, y) \leq \mathcal{L}_\beta(x, z, y) + \big\langle z_{\iota'}^{+} - z, \nabla_z \mathcal{L}_\beta(x, z, y) \big\rangle + \frac{1}{2\iota'} \|z_{\iota'}^{+} - z\|^2 \Big\}. \tag{45}$$

The analysis of Algorithm 6 is similar to that of Algorithm 4, involving also a similar version of Lemma 4. The convergence rate of Algorithm 6 is detailed below and proved in Appendix B.

Algorithm 2 Linearized ADMM for solving problem (40)

Input: Parameters $\beta_1, \sigma_1, \rho, \tau > 0$, primal initialization $x_1, z_1 \in \mathbb{R}^d$ with $\|A(x_1) + B(z_1)\| \leq \rho$, dual initialization $y_1 \in \mathbb{R}^m$.

For $k = 1, 2, \ldots$, execute

1. **(Update penalty weight)** $\beta_k \leftarrow \beta_1 \sqrt{k} \log(k+1) / \log 2$.
2. **(Line search in x)** Use the line search procedure in (44) by replacing $x = x_k, z = z_k, y = y_k, \beta = \beta_k$ and let $\gamma_k \leftarrow \gamma$ be the output.
3. **(Descent step in x)** $x_{k+1} \leftarrow P_g(x_k - \gamma_k \nabla_x \mathcal{L}_{\beta_k}(x_k, z_k, y_k))$, where $\mathcal{L}_{\beta_k}$ is the augmented Lagrangian and P_g denotes the proximal operator, defined in (6,10), respectively.
4. **(Line search in z)** Use the line search procedure in (23) by replacing $x = x_{k+1}, z = z_k, y = y_k, \beta = \beta_k$ and let $\iota_k \leftarrow \iota$ be the output.
5. **(Descent step in z)** $z_{k+1} \leftarrow P_l(z_k - \iota_k \nabla_z \mathcal{L}_{\beta_k}(x_{k+1}, z_k, y_k))$, where P_l denotes the proximal operator for l.
6. **(Stopping criterion)** If $\gamma_k \|G_{\beta_k, \gamma_k}(x_k, z_k, y_k)\|^2 + \iota_k \|G_{\beta_k, \iota_k}(x_{k+1}, z_k, y_k)\|^2 + \sigma_k \|A(x_k) + B(z_k)\|^2 \leq \tau$, then quit and return x_{k+1}, z_{k+1}. See (19) for the definition of the gradient mapping.
7. **(Update dual step size)** $\sigma_{k+1} \leftarrow \sigma_1 \min\Big(\frac{1}{\sqrt{k+1}}, \frac{\|A(x_1)+B(z_1)\|}{|A(x_{k+1})+B(z_{k+1})\|} \cdot \frac{\log^2 2}{(k+1)\log^2(k+2)}\Big)$.
8. **(Dual ascent step)** $y_{k+1} \leftarrow y_k + \sigma_{k+1}(A(x_{k+1}) + B(z_{k+1}))$.

Theorem 2 (Convergence rate of linearized ADMM) *For sufficiently large integers $k_0 < k_1$, consider the interval $K = [k_0 : k_1]$, and consider the output sequence $\{x_k, y_k\}_{k \in K}$ of Algorithm 4. Suppose that*

$$\mu := -\min(0, \inf_k f(x_k) + g(x_k) + \langle A(x_k), y_{k_0} \rangle) - \min(0, \inf_k h(z_k) + l(z_k) + \langle B(z_k), y_{k_0} \rangle) < \infty.$$

AE: note that mu and some other quantities are slightly changed from the proof to simplify the presentation which is inconsequential for the proof I think. *For $\rho' \gtrsim \sqrt{\mu}$, in addition to the strong smoothness of f and A quantified in (2), let us define*

$$\begin{aligned} \lambda'_f &= \max_{\|x\|\le\rho'} \|\nabla f(x)\|, \qquad & \lambda'_A &= \max_{\|x\|\le\rho'} \|\operatorname{D} A(x)\|, \\ \lambda'_h &= \max_{\|z\|\le\rho'} \|\nabla h(x)\|, \qquad & \lambda'_B &= \max_{\|z\|\le\rho'} \|\operatorname{D} B(x)\|, \end{aligned} \tag{46}$$

to be the (restricted) Lipschitz constants of f, A, h, B. Suppose also that problem (1) satisfies uniform regularity and, more specifically,

$$\nu(g, A, l, B, S, \rho, \rho') \gtrsim \max\left((\lambda_A + \lambda_B) \max_{k\in K} \sqrt{(\gamma_k + \iota_k)\mu}, \frac{\lambda'_f + \lambda'_A}{\sqrt{\mu}} \right), \tag{47}$$

with

- $\rho' \ge \max_{k\in K} \|x_k\|$, $\rho' \ge \max_{k\in K} \|z_k\|$,
- $S \supseteq \bigcup_{k\in K} P_{\operatorname{cone}(\partial g(x_{k+1}))^*} \left(\operatorname{D} A(x_{k+1})^\top (A(x_{k+1}) + B(z_{k+1})) \right)$,
- $S \supseteq \bigcup_{k\in K} P_{\operatorname{cone}(\partial l(z_{k+1}))^*} \left(\operatorname{D} B(z_{k+1})^\top (A(x_{k+1}) + B(z_{k+1})) \right)$.

Then the output of Algorithm 4 satisfies

$$\begin{aligned} \min_{k\in K} \Bigg(& \frac{\|G_{\beta_K,\gamma_k}(x_k, z_k, y_k)\|^2 + \|H_{\beta_K,\iota_k}(x_{k+1}, z_k, y_k)\|^2}{\min(\lambda_A, \lambda_B)\rho + \min(\lambda'^2_A, \lambda'^2_B)} \sqrt{\frac{k_0 \log^2(k_0+1)}{k_1 \log^2(k_1+1)}} \\ & + \|A(x_k) + B(z_k)\|^2 \Bigg) \lesssim \frac{1}{k_1 - k_0} \left(\frac{\lambda'^2_f + \lambda'^2_A + \lambda'^2_h + \lambda'^2_B}{\nu(g, A, l, B, S, \rho, \rho')^2} + \mu \right), \end{aligned} \tag{48}$$

where $\lesssim, \gtrsim$ above suppress the dependence on less important parameters, for the sake of clarity.

Most of the remarks after Theorem 1 apply to Theorem 2 too.

7 Related Works

AE: looks like major revision is required here... Augmented Lagrangian based methods are first proposed in [**?**, **?**]. In the convex setting, standard, inexact and linearized versions of ALM are studied extensively [**?**, **?**, **?**, **?**]. Some works also considered the application of ALM/ADMM to nonconvex problems [**?**,**?**]. These works assume that the operator in (1) is linear, therefore, they do not apply to our setting.

Series of influential papers from Burer and Monteiro [**?**, **?**] proposed using the splitting $X = UU^*$ and suggested solving the problem using ALM. First, they did not have any inexact analysis, their analysis requires primal subproblems to be solved exactly which is not practical. Secondly, they have to put an artificial bound to the primal domain which will be ineffective in practice;

which is impossible to do without knowing the norm of the solution. Lastly, their results are for convergence only, without any rate guarantees.

The authors focused on the special case of SDPs without linear constraints in [**?**] and [**?**]. They prove the convergence of gradient descent on Burer-Monteiro factorized formulation. Their results are not able to extend to linear constraints and general convex functions.

Another line of work focused on solving a specific kind of SDPs by applying gradient descent or trust regions methods on manifolds [**?**, **?**]. The authors show that they can apply gradient descent on manifolds to satisfy the first order stationarity conditions in $\mathcal{O}(1/\epsilon^2)$ iterations. In addition, they apply trust regions methods on manifolds to satisfy the second order stationarity conditions in $\mathcal{O}(1/\epsilon^3)$ iterations. Firstly, these methods have to assume that the problem will be on a smooth manifold, which holds for Maximum Cut and generalized eigenvalue problems, but is not satisfied for other important SDPs such as quadratic programming (QAP) and optimal power flow. Secondly, as noted in [**?**], per iteration cost of their method for Max-Cut problem is $\mathcal{O}(n^6)$ for solving (**??**) which is astronomically larger than our cost of $\mathcal{O}(n^2 r)$ where $r \ll n$.

Another recent line of work [**?**] focused on solving the nonlinear constrained nonconvex problem template (**??**) by adapting the primal-dual method of Chambolle and Pock [**?**]. The authors proved the convergence of the method with rate guarantees by assuming error bound conditions on the objective function, which is not necessarily satisfied for general SDPs.

[**?**] focused on the penalty formulation of (**??**) and studied the optimality of second order stationary points of the formulation. However, their results are for connecting the stationary points of the penalty formulation of non-convex problem to the penalty formulation of convex problem and not to the constrained problem itself.

[**?**] can handle the same problem but their algorithm is much more complicated than ours.

8 Experiments

AE: we need a plan for this.

A Proof of Theorem 1

For the reader's convenience, let us recall the updates of the algorithm in iteration k:

$$\begin{aligned} x_{k+1} &= P_g(x_k - \gamma_k \nabla_x \mathcal{L}_{\beta_k}(x_k, y_k)) \\ &= P_g\Big(x_k - \gamma_k \nabla f(x_k) \\ &\qquad - \gamma_k DA(x_k)^\top \left(y_k + \beta_k A(x_k)\right)\Big), \qquad \text{(see (6))} \end{aligned} \tag{49}$$

$$y_{k+1} = y_k + \sigma_{k+1} A(x_{k+1}). \tag{50}$$

Moreover, we will use the shorthand

$$G_k = G_{\beta_k,\gamma_k}(x_k, y_k) = \frac{x_k - x_{k+1}}{\gamma_k}, \qquad \text{(see (19))} \tag{51}$$

throughout the proof. For integers $k_0 \leq k_1$, consider the interval

$$K = [k_0 : k_1] = \{k_0, \cdots, k_1\}. \tag{52}$$

Since the primal step size γ_k is determined by the line search subroutine in Lemma 3, we may now apply Lemma 2 for every iteration in the interval K to find that

$$\begin{aligned}
\frac{\gamma_k \|G_k\|^2}{2} &\leq \mathcal{L}_{\beta_k}(x_k, y_k) + g(x_k) - \mathcal{L}_{\beta_k}(x_{k+1}, y_k) - g(x_{k+1}) \qquad \text{(see Lemma 2)} \\
&= f(x_k) + g(x_k) - f(x_{k+1}) - g(x_{k+1}) + \langle A(x_k) - A(x_{k+1}), y_k \rangle \\
&\quad + \frac{\beta_k}{2}(\|A(x_k)\|^2 - \|A(x_{k+1})\|^2), \qquad \text{(see (6))}
\end{aligned} \tag{53}$$

for every $k \in K$. On the other hand, note that

$$y_k = y_{k_0-1} + \sum_{i=k_0}^{k} \sigma_i A(x_i), \qquad \text{(see (50))} \tag{54}$$

which, after substituting in (53), yields that

$$\begin{aligned}
\frac{\gamma_k \|G_k\|^2}{2} &\leq f(x_k) + g(x_k) - f(x_{k+1}) - g(x_{k+1}) \\
&\quad + \left\langle A(x_k) - A(x_{k+1}), y_{k_0} + \sum_{i=k_0+1}^{k} \sigma_i A(x_i) \right\rangle \\
&\quad + \frac{\beta_k}{2}(\|A(x_k)\|^2 - \|A(x_{k+1})\|^2).
\end{aligned} \tag{55}$$

By summing up (55) over k from k_0 to k_1, we argue that

$$\begin{aligned}
&\sum_{k=k_0}^{k_1} \frac{\gamma_k \|G_k\|^2}{2} \\
&\leq f(x_{k_0}) + g(x_{k_0}) - f(x_{k_1+1}) - g(x_{k_1+1}) + \langle A(x_{k_0}) - A(x_{k_1+1}), y_{k_0} \rangle \\
&\quad + \sum_{k=k_0}^{k_1} \sum_{i=k_0+1}^{k} \sigma_i \langle A(x_k) - A(x_{k+1}), A(x_i) \rangle \\
&\quad + \sum_{k=k_0}^{k_1} \frac{\beta_k}{2}\|A(x_k)\|^2 - \sum_{k=k_0}^{k_1} \frac{\beta_k}{2}\|A(x_{k+1})\|^2 \qquad \text{(see (55))} \\
&= f(x_{k_0}) + g(x_{k_0}) - f(x_{k_1+1}) - g(x_{k_1+1}) + \langle A(x_{k_0}) - A(x_{k_1+1}), y_{k_0} \rangle \\
&\quad + \sum_{k=k_0}^{k_1} \sum_{i=k_0+1}^{k} \sigma_i \langle A(x_k) - A(x_{k+1}), A(x_i) \rangle \\
&\quad + \sum_{k=k_0}^{k_1} \frac{\beta_k}{2}\|A(x_k)\|^2 - \sum_{k=k_0+1}^{k_1+1} \frac{\beta_{k-1}}{2}\|A(x_k)\|^2.
\end{aligned} \tag{56}$$

By manipulating the last line above, we find that

$$
\begin{aligned}
&\sum_{k=k_0}^{k_1} \frac{\gamma_k \|G_k\|^2}{2} \\
&\leq f(x_{k_0}) + g(x_{k_0}) - f(x_{k_1+1}) - g(x_{k_1+1}) + \langle A(x_{k_0}) - A(x_{k_1+1}), y_{k_0} \rangle \\
&\quad + \frac{\beta_{k_0}}{2} \|A(x_{k_0})\|^2 + \sum_{i=k_0+1}^{k_1} \sum_{k=i}^{k_1} \sigma_i \langle A(x_k) - A(x_{k+1}), A(x_i) \rangle \\
&\quad + \sum_{k=k_0+1}^{k_1} \frac{\beta_k - \beta_{k-1}}{2} \|A(x_k)\|^2 - \frac{\beta_{k_1}}{2} \|A(x_{k_1+1})\|^2 \\
&\leq \mu + \sum_{i=k_0+1}^{k_1} \sigma_i \left\langle A(x_i) - A(x_{k_1+1}), A(x_i) \right\rangle \\
&\quad + \sum_{k=k_0+1}^{k_1} \frac{\beta_k - \beta_{k-1}}{2} \|A(x_k)\|^2 - \frac{\beta_{k_1}}{2} \|A(x_{k_1+1})\|^2 \qquad \text{(see (58))} \\
&= \mu + \sum_{k=k_0+1}^{k_1} \left(\sigma_k + \frac{\beta_k - \beta_{k-1}}{2} \right) \|A(x_k)\|^2 \\
&\quad - \sum_{k=k_0+1}^{k_1} \sigma_k \left\langle A(x_{k_1+1}), A(x_k) \right\rangle - \frac{\beta_{k_1}}{2} \|A(x_{k_1+1})\|^2,
\end{aligned} \tag{57}
$$

where we assumed that

$$
\begin{aligned}
\mu := \max \Bigg(\sup_k \Big(& f(x_{k_0}) + g(x_{k_0}) - f(x_k) - g(x_k) + \langle A(x_{k_0}) - A(x_k), y_{k_0} \rangle \\
& + \frac{\beta_{k_0}}{2} \|A(x_{k_0})\|^2 \Big), 0 \Bigg) < \infty,
\end{aligned} \tag{58}
$$

Given initial step sizes $\beta_{k_0}, \sigma_{k_0} > 0$, recall that the penalty weights and the dual step sizes of Algorithm 4 are set to

$$
\beta_k = \beta_{k_0} \sqrt{\frac{k \log^2(k+1)}{k_0 \log^2(k_0+1)}},
$$

$$
\sigma_k = \sigma_{k_0} \min \left(\sqrt{\frac{k_0}{k}}, \frac{\|A(x_{k_0})\| k_0 \log^2(k_0+1)}{\|A(x_k)\| k \log^2(k+1)} \right), \qquad \forall k \in K. \tag{59}
$$

For future reference, (59) implies that

$$\begin{aligned}
\beta_k - \beta_{k-1} &= \beta_{k-1}\left(\sqrt{\frac{k\log^2(k+1)}{(k-1)\log^2 k}} - 1\right) \\
&\le \beta_{k-1}\cdot\frac{k\log^2(k+1)-(k-1)\log^2 k}{(k-1)\log^2 k} \\
&\le \beta_{k-1}\left(\frac{k\log^2(1+\frac{1}{k})}{(k-1)\log^2 k} + \frac{1}{k-1}\right) \\
&\le \frac{2\beta_{k-1}}{k-1} \qquad (k_0 \gg 1) \\
&\le \frac{2\beta_{k_0}}{k-1}\sqrt{\frac{(k-1)\log^2 k}{k_0\log^2(k_0+1)}} \\
&= \frac{2\beta_{k_0}\log k}{\sqrt{(k-1)k_0}\log(k_0+1)}, \qquad \forall k \in K, \qquad (60)
\end{aligned}$$

when k_0 is sufficiently large. We can therefore further simplify the last line of (57) as

$$\begin{aligned}
&\sum_{k=k_0}^{k_1}\frac{\gamma_k\|G_k\|^2}{2} \\
&\le \mu + \sum_{k=k_0}^{k_1}\left(\sigma_k + \frac{\beta_k-\beta_{k-1}}{2}\right)\|A(x_k)\|^2 \\
&\quad + \sum_{k=k_0}^{k_1}\sigma_k\|A(x_{k_1+1})\|\|A(x_k)\| - \frac{\beta_{k_1}}{2}\|A(x_{k_1+1})\|^2 \qquad \text{(see (57))} \\
&\le \mu + \sum_{k=k_0}^{k_1}\left(\sigma_k + \frac{\beta_k-\beta_{k-1}+1}{2}\right)\|A(x_k)\|^2 \\
&\quad + \frac{1}{2}\left(\sum_{k=k_0}^{k_1}\sigma_k^2 - \beta_{k_1}\right)\|A(x_{k_1+1})\|^2 \qquad (2ab \le a^2+b^2) \\
&\le \mu + 2\sum_{k=k_0}^{k_1}\|A(x_k)\|^2, \qquad \text{(see (59,60))} \qquad (61)
\end{aligned}$$

for sufficiently large k_0. Indeed, the coefficient of $\|A(x_{k_1+1})\|$ in the second-to-last line of (61) is negative because

$$\begin{aligned}
&\sum_{k=k_0}^{k_1}\sigma_k^2 - \beta_{k_1} \\
&\le \sum_{k=k_0}^{k_1}\frac{\sigma_{k_0}^2 k_0}{k} - \beta_{k_0}\sqrt{\frac{k_1\log^2(k_1+1)}{k_0\log^2(k_0+1)}} \\
&\le 2\sigma_{k_0}^2 k_0\int_{k_0}^{k_1}\frac{da}{a} - \beta_{k_0}\sqrt{\frac{k_1\log^2(k_1+1)}{k_0\log^2(k_0+1)}} \\
&\le 0, \qquad (62)
\end{aligned}$$

when k_0 is sufficiently large. Note that (61) bounds the gradient mapping with the feasibility gap. A converse is given by Lemma 4. In order to apply this result, let us assume that the

assumptions in Lemma 4 are met. Lemma 4 is then in force and we may now substitute (29) back into (61) to find that

$$\begin{aligned}
&\sum_{k=k_0}^{k_1} \gamma_k \|G_k\|^2 \\
&\leq 2 \sum_{k=k_0}^{k_1} \|A(x_k)\|^2 + 2\mu \qquad \text{(see (61))} \\
&\leq 2 \sum_{k=k_0}^{k_1} \left(\frac{2}{\nu \beta_k} \left(\|G_k\| + \lambda'_f + \lambda'_A \|y_k\| \right) \right)^2 + 2\mu \qquad \text{(see (29))} \\
&\leq \sum_{k=k_0}^{k_1} \frac{32 \|G_k\|^2}{\nu^2 \beta_k^2} + \sum_{k=k_0}^{k_1} \frac{8 \lambda'^2_f}{\nu^2 \beta_k^2} + \sum_{k=k_0}^{k_1} \frac{8 \lambda'^2_A \|y_k\|^2}{\nu^2 \beta_k^2} + 2\mu, \qquad (63)
\end{aligned}$$

where we used the shorthand $\nu = \nu(g, A, S, \rho, \rho')$ and the last line above uses the inequality

$$\left(\sum_{i=1}^{p} a_i \right)^2 \leq p \sum_{i=1}^{p} a_i^2, \qquad (64)$$

for integer p and scalars $\{a_i\}_i$. If we set

$$B_K = \sum_{k=k_0}^{k_1} \frac{\|y_k\|^2}{k \log^2(k+1)}, \qquad c \geq \sum_{k=1}^{\infty} \frac{1}{k \log^2(k+1)}, \qquad (65)$$

and, after recalling the choice of $\{\beta_k\}_k$ in (59), the last line of (63) can be simplified as

$$\begin{aligned}
\sum_{k=k_0}^{k_1} \gamma_k \|G_k\|^2 &\leq 2 \sum_{k=k_0}^{k_1} \|A(x_k)\|^2 + 2\mu \\
&\leq \sum_{k=k_0}^{k_1} \frac{32 \|G_k\|^2 k_0 \log^2(k_0+1)}{\nu^2 \beta_{k_0}^2 k \log^2(k+1)} + \sum_{k=k_0}^{k_1} \frac{8 \lambda'^2_f k_0 \log^2(k_0+1)}{\nu^2 \beta_{k_0}^2 k \log^2(k+1)} + 2\mu \\
&\quad + \sum_{k=k_0}^{k_1} \frac{8 \lambda'^2_A k_0 \log^2(k_0+1) \|y_k\|^2}{\nu^2 k \log^2(k+1)} + 2\mu \qquad \text{(see (59))} \\
&\leq \sum_{k=k_0}^{k_1} \frac{32 \|G_k\|^2 k_0 \log^2(k_0+1)}{\nu^2 \beta_{k_0}^2 k \log^2(k+1)} + \frac{8 k_0 \log^2(k_0+1)}{\nu^2 \beta_{k_0}^2} \left(c \lambda'^2_f + \lambda'^2_A B_K \right) \\
&\quad + 2\mu. \qquad \text{(see (65))} \qquad (66)
\end{aligned}$$

To simplify the above bound, let us assume that

$$\frac{32 k_0 \log^2(k_0+1)}{\nu^2 \beta_{k_0}^2 k \log^2(k+1)} \leq \frac{\gamma_k}{2}, \qquad \forall k \in K. \qquad (67)$$

After rearranging (66) and applying (67), we arrive at

$$\begin{aligned}
&\sum_{k=k_0}^{k_1} \frac{\gamma_k}{2} \|G_k\|^2 \\
&\leq \sum_{k=k_0}^{k_1} \left(\gamma_k - \frac{32 k_0 \log^2(k_0+1)}{\nu^2 \beta_{k_0}^2 k \log^2(k+1)} \right) \|G_k\|^2 \qquad \text{(see (67))} \\
&\leq \frac{8 k_0 \log^2(k_0+1)}{\nu^2 \beta_{k_0}^2} \left(c \lambda'^2_f + \lambda'^2_A B_K \right) + 2\mu. \qquad \text{(see (68))} \qquad (68)
\end{aligned}$$

In turn, the bound above on the gradient mapping controls the feasibility gap, namely,

$$\begin{aligned}\sum_{k=k_0}^{k_1}\|A(x_k)\|^2 &\leq \sum_{k=k_0}^{k_1}\frac{\gamma_k\|G_k\|^2}{4}+\frac{4k_0\log^2(k_0+1)}{\nu^2\beta_{k_0}^2}\left(c\lambda_f'^2+\lambda_A'^2B_K\right) \qquad \text{(see (66,67))}\\ &\leq \frac{8k_0\log^2(k_0+1)}{\nu^2\beta_{k_0}^2}\left(c\lambda_f'^2+\lambda_A'^2B_K\right)+\mu. \qquad \text{(see (68))} \end{aligned} \tag{69}$$

By adding (68,69), we find that

$$\sum_{k=k_0}^{k_1}\gamma_k\|G_k\|^2+\|A(x_k)\|^2 \leq \frac{24k_0\log^2(k_0+1)}{\nu^2\beta_{k_0}^2}\left(c\lambda_f'^2+\lambda_A'^2B_K\right)+5\mu. \tag{70}$$

In order to interpret (70), we next estimate B_K, defined in (65). To that end, let us first control the growth of the dual sequence $\{y_k\}_k$. Recalling (50) and for every $k\in K$, we write that

$$\begin{aligned}\|y_k\| &\leq \|y_{k_0}\|+\sum_{i=k_0+1}^{k}\sigma_i\|A(x_i)\| \qquad \text{(see (50))}\\ &\leq \|y_{k_0}\|+\sum_{i=k_0+1}^{k}\frac{\rho\sigma_{k_0}k_0\log^2(k_0+1)}{k\log^2(k+1)} \qquad \text{(see (??,59))}\\ &\leq \|y_{k_0}\|+c\rho\sigma_{k_0}k_0\log^2(k_0+1)\\ &=: y_{\max}. \end{aligned} \tag{71}$$

With the growth of the dual sequence discovered above, we evaluate B_K as

$$\begin{aligned}B_K &= \sum_{k=k_0}^{k_1}\frac{\|y_k\|^2}{k\log^2(k+1)} \qquad \text{(see (65))}\\ &\leq \sum_{k=k_0}^{k_1}\frac{y_{\max}^2}{k\log^2(k+1)} \qquad \text{(see (71))}\\ &\leq cy_{\max}^2. \qquad \text{(see (65))} \end{aligned} \tag{72}$$

In order to interpret (70), it still remains to estimate the primal step sizes $\{\gamma_k\}_k$. To invoke (24), we in turn need to gauge how smooth the augmented Lagrangian $\mathcal{L}_{\beta_k}(\cdot,y_k)$ is. For every $k\in K$, note that

$$\begin{aligned}\lambda_{\beta_k} &\leq \lambda_f+\sqrt{m}\lambda_A\left(\|y_k\|+\beta_k\rho\right)+\beta_k d\lambda_A'^2 \qquad \text{(see (17))}\\ &\leq \left(\lambda_f+\sqrt{m}\lambda_A y_{\max}\right)+\beta_k\left(\sqrt{m}\lambda_A\rho+d\lambda_A'^2\right). \qquad \text{(see (71))} \end{aligned} \tag{73}$$

We are now in position to invoke (24) by writing that

$$\begin{aligned}\gamma_k &\geq \frac{\theta}{\lambda_{\beta_k}} \qquad \text{(see (24))}\\ &\geq \frac{\theta}{\left(\lambda_h+\sqrt{m}\lambda_A y_{\max}\right)+\beta_k\left(\sqrt{m}\lambda_A\rho+d\lambda_A'^2\right)} \qquad \text{(see (73))}\\ &\geq \frac{\theta}{2\beta_k\left(\sqrt{m}\lambda_A\rho+d\lambda_A'^2\right)} \qquad \text{((59) and } k_0\gg 1)\\ &\geq \frac{\theta}{2\beta_{k_0}\left(\sqrt{m}\lambda_A\rho+d\lambda_A'^2\right)}\sqrt{\frac{k_0\log^2(k_0+1)}{k\log^2(k+1)}} \qquad \text{(see (59))}\\ &=: \overline{\gamma}\sqrt{\frac{k_0\log^2(k_0+1)}{k\log^2(k+1)}}, \end{aligned} \tag{74}$$

for every $k \in K$. The first consequence of (74) is that (67) holds automatically when k_0 is sufficiently large. Having estimated B_K and $\{\gamma_k\}_k$, we can also rewrite (70). Indeed, (70,72,74) together imply that

$$\begin{aligned}&\sum_{k=k_0}^{k_1} \overline{\gamma}\|G_k\|^2 \sqrt{\frac{k_0 \log^2(k_0+1)}{k \log^2(k+1)}} + \|A(x_k)\|^2 \\ &\leq \frac{24ck_0 \log^2(k_0+1)}{\nu^2 \beta_{k_0}^2} \left(\lambda_f'^2 + \lambda_A'^2 y_{\max}^2\right) + 5\mu, \end{aligned} \tag{75}$$

and, consequently,

$$\begin{aligned}&\min_{k\in K} \overline{\gamma}\|G_k\|^2 \sqrt{\frac{k_0 \log^2(k_0+1)}{k_1 \log^2(k_1+1)}} + \|A(x_k)\|^2 \\ &\leq \frac{24ck_0 \log^2(k_0+1)}{\nu^2 \beta_{k_0}^2 (k_1-k_0)} \left(\lambda_f'^2 + \lambda_A'^2 y_{\max}^2\right) + \frac{5\mu}{k_1-k_0}. \end{aligned} \tag{76}$$

When we applied Lemma 4 earlier, we did not check whether the assumptions on ρ therein hold. Let us revisit this assumption. We first derive a weaker but uniform bound on the feasibility gap. For every $k \in K$, it holds that

$$\begin{aligned}\|A(x_k)\|^2 &\leq \sum_{i=k_0}^{k_1} \|A(x_i)\|^2 \\ &\leq \frac{8k_0 \log^2(k_0+1)}{\nu^2 \beta_{k_0}^2} \left(c\lambda_f'^2 + \lambda_A'^2 B_K\right) + \mu \qquad \text{(see (69))} \\ &\leq \frac{8ck_0 \log^2(k_0+1)}{\nu^2 \beta_{k_0}^2} \left(\lambda_f'^2 + \lambda_A'^2 y_{\max}^2\right) + \mu. \qquad \text{(see (72))} \end{aligned} \tag{77}$$

Therefore, we may replace the assumption on ρ in Lemma 4 with the stronger assumption that

$$\rho^2 \geq \frac{8ck_0 \log^2(k_0+1)}{\nu^2 \beta_{k_0}^2} \left(\lambda_f'^2 + \lambda_A'^2 y_{\max}^2\right) + \mu, \tag{78}$$

which, after rearranging, can be presented as

$$\nu^2 \geq \frac{8ck_0 \log^2(k_0+1)}{\beta_{k_0}^2 (\rho^2-\mu)} \left(\lambda_f'^2 + \lambda_A'^2 y_{\max}^2\right), \qquad \rho > \sqrt{\mu}. \tag{79}$$

Note that, for (79) to hold, it is in particular necessary that $\|A(x_{k_0})\| \leq \rho\sqrt{2/\beta_{k_0}}$, as seen in (58). That is, for Algorithm 4 to success, it must be initialized close enough to the feasible set. Lastly, let us revisit the lower bound on ν in Lemma 4, namely, (28). First we derive a weaker but uniform bound on the gradient mapping. For every $k \in K$, it holds that

$$\begin{aligned}&\max_{k\in K} \gamma_k \|G_k\| \\ &\leq \max_{k\in K} \sqrt{\gamma_k} \cdot \sqrt{\max_{k\in K} \gamma_k \|G_k\|^2} \\ &\leq \max_{k\in K} \sqrt{\gamma_k} \cdot \sqrt{\sum_{k=k_0}^{k_1} \gamma_k \|G_k\|^2} \\ &\leq \max_{k\in K} \sqrt{\gamma_k} \cdot \left(\frac{16ck_0 \log^2(k_0+1)}{\nu^2 \beta_{k_0}^2} \left(\lambda_f'^2 + \lambda_A'^2 y_{\max}^2\right) + 4\mu\right)^{\frac{1}{2}}. \qquad \text{(see (68,72))} \end{aligned} \tag{80}$$

Instead of (28), it therefore suffices to make the stronger assumption that

$$\nu \geq 2\lambda_A \max_{k\in K} \sqrt{\gamma_k} \cdot \left(\frac{16ck_0 \log^2(k_0+1)}{\nu^2 \beta_{k_0}^2} \left(\lambda_f'^2 + \lambda_A'^2 y_{\max}^2 \right) + 4\mu \right)^{\frac{1}{2}}, \tag{81}$$

which can in turn be replaced with the stronger assumptions

$$\nu \geq \max \left(4\sqrt{2}\lambda_A \max_{k\in K} \sqrt{\gamma_k \mu}, \frac{2\sqrt{ck_0}\log(k_0+1)}{\beta_{k_0}\sqrt{\mu}} (\lambda_f' + \lambda_A' y_{\max}) \right) \tag{82}$$

This completes the proof of Theorem 1.

B Proof of Theorem 2

For completeness, let us repeat the technical lemmas and definitions of Section 2, slightly adjusted here for the augmented Lagrangian of problem (40), defined in (39). These standard results are stated below without proof.

Lemma 5 (Smoothness) *Given $\rho, \rho' \geq 0$, it holds that*

$$\|\nabla_x \mathcal{L}_\beta(x,z,y) - \nabla_x \mathcal{L}_\beta(x',z,y)\| \leq \lambda_{\beta,z}\|x-x'\|,$$

$$\|\nabla_z \mathcal{L}_\beta(x,z,y) - \nabla_z \mathcal{L}_\beta(x,z',y)\| \leq \lambda_{\beta,x}\|z-z'\|, \tag{83}$$

for every $(x,z),(x',z),(x,z') \in X_{\rho,\rho'}$ and $y \in \mathbb{R}^m$, where

$$X_{\rho,\rho'} := \{(x'',z'') : \|A(x'')+B(z'')\| \leq \rho, \|x''\| \leq \rho', \|z''\| \leq \rho'\}, \tag{84}$$

$$\lambda_{\beta,x} \leq \lambda_f + \sqrt{m}\lambda_A \left(\|y\| + \beta\rho\right) + \beta d \lambda_A'^2.$$

$$\lambda_{\beta,z} \leq \lambda_h + \sqrt{m}\lambda_B \left(\|y\| + \beta\rho\right) + \beta d \lambda_B'^2, \tag{85}$$

$$\lambda_A' := \max_{\|x\|\leq\rho'} \|\,\mathrm{D}\,A(x)\|, \qquad \lambda_B' := \max_{\|z\|\leq\rho'} \|\,\mathrm{D}\,B(z)\|, \tag{86}$$

and $\lambda_f, \lambda_A, \lambda_h, \lambda_B$ were defined in (38).

Definition 3 (Gradient Mapping) Given $x,z \in \mathbb{R}^d$ and $\gamma,\iota > 0$, the gradient mappings $G_{\beta,\gamma}(\cdot,z,y), G_{\beta,\iota}(x,\cdot,y) : \mathbb{R}^d \to \mathbb{R}^d$ take, respectively, $x,z \in \mathbb{R}^d$ to

$$G_{\beta,\gamma}(x,z,y) = \frac{x-x^+}{\gamma}, \qquad H_{\beta,\iota}(x,z,y) = \frac{z-z^+}{\iota}, \tag{87}$$

where $x^+ = P_g(x - \gamma\nabla_x \mathcal{L}_\beta(x,z,y))$ and $z^+ = P_l(z - \iota\nabla_z \mathcal{L}_\beta(x,z,y))$.

Lemma 6 (Descent lemma) *For $x,z \in \mathbb{R}^d$ and $y \in \mathbb{R}^m$, let x^+, z^+ be as in Definition 3 with $\gamma < 1/\lambda_{\beta,x}$ and $\iota < 1/\lambda_{\beta,z}$. For $\rho,\rho' \geq 0$, suppose that*

$$(x,z),(x^+,z),(x,z^+) \in X_{\rho,\rho'}. \tag{88}$$

Then it holds that

$$\|G_{\beta,\gamma}(x,z,y)\|^2 \leq \frac{2}{\gamma}(\mathcal{L}_\beta(x,z,y) + g(x) - \mathcal{L}_\beta(x^+,z,y) - g(x^+)),$$

$$\|H_{\beta,\iota}(x,z,y)\|^2 \leq \frac{2}{\iota}(\mathcal{L}_\beta(x,z,y) + l(x) - \mathcal{L}_\beta(x,z^+,y) - l(x^+)). \tag{89}$$

Lemma 7 (Line search) *Fix $\theta \in (0,1)$ and $\gamma_0, \iota_0 > 0$. For $\gamma', \iota' > 0$, let*

$$x_{\gamma'}^+ = P_g(x - \gamma' \nabla_x \mathcal{L}_\beta(x,z,y)), \qquad z_{\iota'}^+ = P_l(z - \iota' \nabla_z \mathcal{L}_\beta(x,z,y)), \tag{90}$$

and define

$$\begin{aligned}\gamma := \max \Big\{ \gamma' = \gamma_0 \theta^i : \mathcal{L}_\beta(x_{\gamma'}^+, z, y) \\ \leq \mathcal{L}_\beta(x,z,y) + \Big\langle x_{\gamma'}^+ - x, \nabla_x \mathcal{L}_\beta(x,z,y) \Big\rangle + \frac{1}{2\gamma'} \|x_{\gamma'}^+ - x\|^2 \Big\},\end{aligned}$$

$$\begin{aligned}\iota := \max \Big\{ \iota' = \iota_0 \theta^i : \mathcal{L}_\beta(x, z_{\iota'}^+, y) \\ \leq \mathcal{L}_\beta(x,z,y) + \Big\langle z_{\iota'}^+ - z, \nabla_z \mathcal{L}_\beta(x,z,y) \Big\rangle + \frac{1}{2\iota'} \|z_{\iota'}^+ - z\|^2 \Big\}.\end{aligned} \tag{91}$$

Then, (89) holds and, moreover, we have that

$$\gamma \geq \frac{\theta}{\lambda_{\beta,x}}, \qquad \iota \geq \frac{\theta}{\lambda_{\beta,z}}. \tag{92}$$

For the reader's convenience, let us also recall the updates of Algorithm 6 in iteration k as

$$\begin{aligned} x_{k+1} &= P_g(x_k - \gamma_k \nabla_x \mathcal{L}_{\beta_k}(x_k, z_k, y_k)), \\ z_{k+1} &= P_l(z_k - \iota_k \nabla_z \mathcal{L}_{\beta_k}(x_{k+1}, z_k, y_k)), \\ y_{k+1} &= y_k + \sigma_{k+1}(A(x_{k+1}) + B(z_{k+1})). \end{aligned} \tag{93}$$

For every $k \in K = [k_0 : k_1]$, recall that the primal step sizes γ_k, ι_k are determined by line search in Lemma 7. Moreover, the penalty weights and dual step sizes are set as

$$\begin{aligned} \beta_k &= \beta_{k_0} \sqrt{\frac{k \log^2(k+1)}{k_0 \log^2(k_0+1)}}, \\ \sigma_k &= \sigma_{k_0} \min\left(\sqrt{\frac{k_0}{k}}, \frac{\|A(x_{k_0}) + B(z_{k_0})\|}{\|A(x_k) + B(z_k)\|} \cdot \frac{k_0 \log^2(k_0+1)}{k \log^2(k+1)} \right). \end{aligned} \tag{94}$$

For every $k \in K$, let us set

$$G_k = G_{\beta_k, \gamma_k}(x_k, z_k, y_k) = \frac{x_k - x_{k+1}}{\gamma_k},$$

$$H_k = H_{\beta_k, \iota_k}(x_{k+1}, z_k, y_k) = \frac{z_k - z_{k+1}}{\iota_k}, \tag{95}$$

for short. The convergence analysis of Algorithm 6 only slightly differs from the one in the proof of Theorem 1 and we therefore only present the proof sketch, somewhat informally. Similar to the proof of Theorem 1, two applications of Lemma 6 yields that

$$\begin{aligned} \frac{\gamma_k \|G_k\|^2}{2} &\leq \mathcal{L}_{\beta_k}(x_k, z_k, y_k) + g(x_k) - \mathcal{L}_{\beta_k}(x_{k+1}, z_k, y_k) - g(x_{k+1}), \\ \frac{\iota_k \|H_k\|^2}{2} &\leq \mathcal{L}_{\beta_k}(x_{k+1}, z_k, y_k) + l(z_k) - \mathcal{L}_{\beta_k}(x_{k+1}, z_{k+1}, y_k) - l(z_{k+1}), \end{aligned} \tag{96}$$

for every k. By setting

$$u_k = [\, x_k^\top \;\; z_k^\top \,]^\top \in \mathbb{R}^{2d}, \qquad Q_k = [\, G_k^\top \;\; H_k^\top \,]^\top \in \mathbb{R}^{2d}, \qquad q(u) = f(x) + h(z),$$

$$q'(u) = g(x) + l(z), \qquad D(u) = A(x) + B(z), \qquad \kappa_k = \min(\gamma_k, \iota_k), \tag{97}$$

for every $k \in K$ and after summing up the two inequalities in (96), we reach

$$\frac{\kappa_k \|Q_k\|^2}{2} \leq \mathcal{L}_{\beta_k}(u_k, y_k) + q'(u_k) - \mathcal{L}_{\beta_k}(u_{k+1}, y_k) - q'(u_{k+1}), \qquad \forall k \in K. \tag{98}$$

By following the same steps as in the proof of Theorem 1, we find that

$$\sum_{k=k_0}^{k_1} \frac{\kappa_k \|Q_k\|^2}{2} \leq \mu + 2 \sum_{k=k_0}^{k_1} \|A(x_k) + B(z_k)\|^2, \tag{99}$$

where

$$\mu := \max \Bigg(\sup_k \Big(q(u_{k_0}) + q'(u_{k_0}) - q(u_k) - q'(u_k) + \langle A(x_{k_0}) + B(z_{k_0}) - A(x_k) - B(z_k), y_{k_0} \rangle + \frac{\beta_{k_0}}{2} \|A(x_{k_0}) + B(z_{k_0})\|^2 \Big), 0 \Bigg) < \infty. \tag{100}$$

On the other hand, the x and z updates in (93) imply that

$$G_k - \nabla f(x_k) - \mathrm{D}\, A(x_k)^\top y_k - \beta_k \,\mathrm{D}\, A(x_k)^\top (A(x_k) + B(z_k)) \in \partial g(x_{k+1}), \tag{101}$$

$$H_k - \nabla h(z_k) - \mathrm{D}\, B(z_k)^\top y_k - \beta_k \,\mathrm{D}\, B(z_k)^\top (A(x_{k+1}) + B(z_k)) \in \partial l(z_{k+1}), \tag{102}$$

which can be more compactly written as

$$Q_k - \nabla q(u_k) - \mathrm{D}\, D(u_k)^\top y_k - \beta_k \,\mathrm{D}\, D(u_k)^\top D(u_k) + \beta_k \begin{bmatrix} 0 \\ \mathrm{D}\, B(z_k)^\top (A(x_k) - A(x_{k+1})) \end{bmatrix} \in \partial q'(u_{k+1}). \tag{103}$$

Note that (103) is similar to (121), except for its last term, which satisfies

$$\begin{aligned} &\beta_k \left\| \begin{bmatrix} 0 \\ \mathrm{D}\, B(z_k)^\top (A(x_k) - A(x_{k+1})) \end{bmatrix} \right\| \\ &= \beta_k \| \mathrm{D}\, B(z_k)^\top (A(x_k) - A(x_{k+1})\| \\ &\leq \beta_k \lambda'_A \lambda'_B \|x_k - x_{k+1}\| \qquad \text{(see (86))} \\ &= \beta_k \gamma_k \lambda'_A \lambda'_B G_k, \qquad \text{(see (87))} \end{aligned} \tag{104}$$

and it is not difficult to verify that $\max_{k \geq k_0} \beta_k \gamma_k = O(1)$. From this point, the rest of the proof steps of Lemma 4 and Theorem 1 can be applied directly to complete the proof of Theorem 2.

C Proof of Lemma 1

AE: We assume Hessian exists. We shouldn't assume that for a strictly correct proof! Do you know how to correct this? Note that

$$\mathcal{L}_\beta(x, y) = f(x) + \sum_{i=1}^{m} y_i A_i(x) + \frac{\beta}{2} \sum_{i=1}^{m} (A_i(x))^2, \tag{105}$$

which implies that

$$\begin{aligned}\nabla_x \mathcal{L}_\beta(x,y) &= \nabla f(x) + \sum_{i=1}^m y_i \nabla A_i(x) + \frac{\beta}{2}\sum_{i=1}^m A_i(x)\nabla A_i(x) \\ &= \nabla f(x) + DA(x)^\top y + \beta DA(x)^\top A(x), \end{aligned} \tag{106}$$

where $DA(x)$ is the Jacobian of A at x. By taking another derivative with respect to x, we reach

$$\nabla_x^2 \mathcal{L}_\beta(x,y) = \nabla^2 f(x) + \sum_{i=1}^m \left(y_i + \beta A_i(x)\right)\nabla^2 A_i(x) + \beta \sum_{i=1}^m \nabla A_i(x)\nabla A_i(x)^\top. \tag{107}$$

It follows that

$$\begin{aligned}\|\nabla_x^2 \mathcal{L}_\beta(x,y)\| &\le \|\nabla^2 f(x)\| + \max_i \|\nabla^2 A_i(x)\| \left(\|y\|_1 + \beta\|A(x)\|_1\right) + \beta \sum_{i=1}^m \|\nabla A_i(x)\|^2 \\ &\le \lambda_h + \sqrt{m}\lambda_A \left(\|y\| + \beta\|A(x)\|\right) + \beta\|DA(x)\|_F^2. \end{aligned} \tag{108}$$

For every x such that $\|A(x)\| \le \rho$ and $\|x\| \le \rho'$, we conclude that

$$\|\nabla_x^2 \mathcal{L}_\beta(x,y)\| \le \lambda_f + \sqrt{m}\lambda_A \left(\|y\| + \beta\rho\right) + \beta \max_{\|x\|\le\rho'} \|DA(x)\|_F^2, \tag{109}$$

which completes the proof of Lemma 1.

D Proof of Lemma 2

Throughout, let

$$G = G_{\beta,\gamma}(x,y) = \frac{x - x^+}{\gamma}, \tag{110}$$

for short. Suppose that $\|A(x)\| \le \rho$, $\|x\| \le \rho$, and similarly $\|A(x^+)\| \le \rho$, $\|x^+\| \le \rho'$. An application of Lemma 1 yields that

$$\begin{aligned}\mathcal{L}_\beta(x^+,y) + g(x^+) &\le \mathcal{L}_\beta(x,y) + \langle x^+ - x, \nabla_x \mathcal{L}_\beta(x,y)\rangle + \frac{\lambda_\beta}{2}\|x^+ - x\|^2 + g(x^+) \\ &= \mathcal{L}_\beta(x,y) - \gamma\langle G, \nabla_x \mathcal{L}_\beta(x,y)\rangle + \frac{\gamma^2\lambda_\beta}{2}\|G\|^2 + g(x^+) \end{aligned} \tag{111}$$

Since $x^+ = P_g(x - \gamma\nabla_x \mathcal{L}_\beta(x,y))$, we also have that

$$G - \nabla_x \mathcal{L}_\beta(x,y) = \xi \in \partial g(x^+). \tag{112}$$

By combining (111,112), we find that

$$\begin{aligned}\mathcal{L}_\beta(x^+,y) + g(x^+) &\le \mathcal{L}_\beta(x,y) - \gamma\|G\|^2 + \gamma\langle G, \xi\rangle + \frac{\gamma^2\lambda_\beta}{2}\|G\|^2 + g(x^+) \\ &= \mathcal{L}_\beta(x,y) - \gamma\|G\|^2 + \langle x - x^+, \xi\rangle + \frac{\gamma^2\lambda_\beta}{2}\|G\|^2 + g(x^+) \\ &\le \mathcal{L}_\beta(x,y) + g(x) - \gamma\left(1 - \frac{\gamma\lambda_\beta}{2}\right)\|G\|^2, \end{aligned} \tag{113}$$

where the last line above uses the convexity of g. Recalling that $\gamma \le 1/\lambda_\beta$ completes the proof of Lemma 2.

E Proof of Lemma 3

By optimality of x_γ^+ in (22), we note that

$$x_\gamma^+ - x + \gamma \nabla_x \mathcal{L}_\beta(x,y) = -\gamma\xi \in -\gamma \partial g(x_\gamma^+). \tag{114}$$

By definition in (23), γ also satisfies

$$\begin{aligned}
&\mathcal{L}_\beta(x_\gamma^+, y) + g(x_\gamma^+) \\
&\leq \mathcal{L}_\beta(x,y) + \left\langle x_\gamma^+ - x, \nabla_x \mathcal{L}_\beta(x,y)\right\rangle + \frac{1}{2\gamma}\|x_\gamma^+ - x\|^2 + g(x_\gamma^+) \\
&= \mathcal{L}_\beta(x,y) + \left\langle x - x_\gamma^+, \xi\right\rangle - \frac{1}{2\gamma}\|x_\gamma^+ - x\|^2 + g(x_\gamma^+) \\
&\leq \mathcal{L}_\beta(x,y) - \frac{1}{2\gamma}\|x_\gamma^+ - x\|^2 + g(x) - g(x_\gamma^+) \qquad \text{(convexity of } g\text{)} \\
&= \mathcal{L}_\beta(x,y) - \frac{\gamma}{2}\|G_{\beta,\gamma}(x,y)\|^2 + g(x) - g(x_\gamma^+), \qquad \text{(see Definition 1)}
\end{aligned} \tag{115}$$

which completes the proof of Lemma 3 since (24) follows directly from (23).

F Proof of Proposition 1

To prove the first claim of the proposition, suppose that the Slater's condition does not hold, namely, suppose that

$$\operatorname{null}(A) \cap \operatorname{relint}(C) = \emptyset, \tag{116}$$

where $\operatorname{null}(A)$ and $\operatorname{relint}(C)$ denote the null space of the matrix A and the relative interior of C, respectively. We have assumed that (1) is feasible, namely, there exists $x \in C$ such that $Ax = 0$. It follows from (116) that $x \in \text{boundary}(C)$ and that $\operatorname{null}(A)$ supports C at x, namely, $Ax \geq 0$, for every $x \in C$. (The inequality applies to each entry of the vector Ax.) Consequently, $\operatorname{null}(A) \cap T_C(x) \neq \{0\}$, where $T_C(x)$ is the tangent cone of the set C at x_0. Equivalently, it holds that $\operatorname{row}(A) \cap N_C(x) \neq \{0\}$, where $\operatorname{row}(A)$ is the row space of the matrix A and $N_C(x)$ is the normal cone to C at x. That is, there exists a unit-norm vector v such that $P_{T_C(x)} A^\top v = 0$ and, consequently, $P_S P_{T_C(x)} A^\top v = 0$. Let us take $\rho' = \|x\|$ in (27). We then conclude that

$$\nu(g, A, S, 1, \|x\|) = \nu(g, A, S, \infty, \|x\|) = 0,$$

namely, the uniform regularity also does not hold for any $\rho \geq 0$ and $\rho' = \|x\|$. The above identity follows from the homogenity of the right-hand side of (27). Because increasing ρ' cannot increase the right-hand side of (25), we find that $\nu(g, A, S, \infty, \max_{x \in C} \|x\|) = 0$, which proves the first claim in Proposition 1.

For the converse, we can verify that it suffices to take $\operatorname{row}(A) \subseteq S$. Next, suppose that uniform regularity does not hold, namely, there exists $x \in \mathbb{R}^d$ such that

$$\eta_{\min}(P_S P_{T_C(x)} A^\top) = 0. \tag{117}$$

Throughout, we assume without loss of generality that $x \in C$. Indeed, otherwise x would not be feasible for problem (1) with $g = 1_C$ and cannot be used to study the Slater's condition in (1). Note that (117) can be rewritten as

$$\eta_{\min}(P_S P_{T_C(x)} A^\top) = \eta_{\min}(P_{T_C(x)} A^\top) = 0, \qquad (S = \operatorname{aff}(C)) \tag{118}$$

where $\operatorname{aff}(C)$ is the affine hull of C. Then, since $\|x\| \leq \rho' < \infty$ in (27), we can assume throughout that $\text{boundary}(C) \cap B_{\rho'} \neq \emptyset$ and, moreover, $x \in \text{boundary}(C)$. Here, $B_{\rho'} = \{z :$

$\|z\| \leq \rho'\}$ is the ball of radius ρ' at the origin. Indeed, otherwise if $x \in \operatorname{relint}(C)$, we have that $T_C(x) = S$ and thus

$$\begin{aligned}\eta_{\min}(P_S P_{T_C(x)} A^\top) &= \eta_{\min}(P_{T_C(x)} A^\top) \qquad (S = \operatorname{aff}(C)) \\ &= \eta_{\min}(A^\top) \qquad (\operatorname{row}(A) \subseteq S) \\ &> 0,\end{aligned}$$

which contradicts (117). The last line above holds because, by assumption, A is full-rank. Therefore, by (118), there exists a unit-norm $u \in \operatorname{row}(A)$ such that $u \in N_C(x)$. In turn, this implies that $\operatorname{null}(A) \cap \operatorname{int}(C) = \emptyset$. Indeed, otherwise, any vector $v \in \operatorname{null}(A) \cap \operatorname{int}(C)$ satisfies $\langle u, v \rangle < 0$, which is impossible because $u \in \operatorname{row}(A)$ and $v \in \operatorname{null}(A)$ are orthogonal vectors. That is, the Slater's condition does not hold, which proves the second claim in Proposition 1.

G Proof of Lemma 4

By assumption, we have that

$$\max_{k \in K} \|A(x_k)\| \leq \rho, \qquad \max_{k \in K} \|x_k\| \leq \rho'. \tag{119}$$

From the x update in (49), it follows that

$$x_{k+1} - x_k + \gamma_k \nabla f(x_k) + \gamma_k \operatorname{D} A(x_k)^\top (y_k + \beta_k A(x_k)) \in -\partial g(x_{k+1}), \tag{120}$$

which, after recalling (51), can be written as

$$-\frac{G_k}{\beta_k} + \frac{\nabla f(x_k)}{\beta_k} + \frac{\operatorname{D} A(x_k)^\top y_k}{\beta_k} + \operatorname{D} A(x_k)^\top A(x_k) \in -\frac{\partial g(x_{k+1})}{\beta_k \gamma_k}. \tag{121}$$

Let $\operatorname{cone}(\partial g(x))^*$ denote the polar of

$$\operatorname{cone}(\partial g(x))) = \bigcup_{\alpha \geq 0} \alpha \cdot \partial g(x) \subseteq \mathbb{R}^d. \tag{122}$$

By projecting both sides (121) onto $\operatorname{cone}(\partial g(x_{k+1}))^*$, we find that

$$\begin{aligned}&P_{\operatorname{cone}(\partial g(x_{k+1}))^*} \left(-\frac{G_k}{\beta_k} + \frac{\nabla f(x_k)}{\beta_k} + \frac{\operatorname{D} A(x_k)^\top y_k}{\beta_k} + \operatorname{D} A(x_k)^\top A(x_k) \right) \\ &\in P_{\operatorname{cone}(\partial g(x_{k+1}))^*} \left(-\frac{\partial g(x_{k+1})}{\beta_k \gamma_k} \right) = \{0\},\end{aligned} \tag{123}$$

where the equality above follows from the duality of $\operatorname{cone}(\partial g(x_{k+1}))^*$ and $\operatorname{cone}(\partial g(x_{k+1}))$. Recall also that the subspace $S \subseteq \mathbb{R}^d$ satisfies

$$S \supseteq \bigcup_{k \in K} P_{\operatorname{cone}(\partial g(x_{k+1}))^*} \left(\operatorname{D} A(x_{k+1})^\top A(x_{k+1}) \right), \tag{124}$$

and project both sides of (123) onto S to reach

$$P_S P_{\operatorname{cone}(\partial g(x_{k+1}))^*} \left(-\frac{G_k}{\beta_k} + \frac{\nabla f(x_k)}{\beta_k} + \frac{\operatorname{D} A(x_k)^\top y_k}{\beta_k} + \operatorname{D} A(x_k)^\top A(x_k) \right) = 0. \tag{125}$$

By taking the norm and then applying the triangle inequality above, we argue that

$$\begin{aligned}&\left\| P_S P_{\operatorname{cone}(\partial g(x_{k+1}))^*} (\operatorname{D} A(x_k)^\top A(x_k)) \right\| \\ &\leq \left\| P_S P_{\operatorname{cone}(\partial g(x_{k+1}))^*} \left(-\frac{G_k}{\beta_k} + \frac{\nabla f(x_k)}{\beta_k} + \frac{\operatorname{D} A(x_k)^\top y_k}{\beta_k} \right) \right\| \qquad \text{(see (125))}.\end{aligned} \tag{126}$$

Because proximal map is non-expansive and $P_S P_{\mathrm{cone}(\partial g(x_{k+1}))^*}(0) = 0$, we may upper bound the last line above as

$$\begin{aligned}
&\left\| P_S P_{\mathrm{cone}(\partial g(x_{k+1}))^*}(\mathrm{D}\, A(x_k)^\top A(x_k)) \right\| \\
&\leq \left\| -\frac{G_k}{\beta_k} + \frac{\nabla f(x_k)}{\beta_k} + \frac{\mathrm{D}\, A(x_k)^\top y_k}{\beta_k} \right\| \\
&\leq \frac{1}{\beta_k} \left(\|G_k\| + \|\nabla f(x_k)\| + \|\mathrm{D}\, A(x_k)^\top y_k\| \right). \qquad \text{(triangle inequality)} \\
&\leq \frac{1}{\beta_k} \left(\|G_k\| + \lambda'_f + \lambda'_A \|y_k\| \right), \qquad (127)
\end{aligned}$$

where

$$\lambda'_f := \max_{\|x\| \leq \rho'} \|\nabla f(x)\|, \qquad \lambda'_A := \max_{\|x\| \leq \rho'} \|\mathrm{D}\, A(x)\|. \qquad (128)$$

To lower bound the first line of (127), we invoke the restricted injectivity in Section 3. Indeed, recalling (25) and the first bound in (119), for every $k \in K$, we write that

$$\begin{aligned}
&\left\| P_S P_{\mathrm{cone}(\partial g(x_{k+1}))^*}(\mathrm{D}\, A(x_k)^\top A(x_k)) \right\| \\
&\geq \left\| P_S P_{\mathrm{cone}(\partial g(x_{k+1}))^*}(\mathrm{D}\, A(x_{k+1})^\top A(x_k)) \right\| - \left\| (\mathrm{D}\, A(x_{k+1}) - \mathrm{D}\, A(x_k))^\top A(x_k) \right\| \\
&\geq \nu(g, A, S, \rho, \rho') \|A(x_k)\| - \|\mathrm{D}\, A(x_{k+1}) - \mathrm{D}\, A(x_k)\| \|A(x_k)\|, \qquad \text{(see (25))} \qquad (129)
\end{aligned}$$

where the second line above again uses the non-expansiveness of P_S and $P_{\mathrm{cone}(\partial g(x_{k+1}))^*}$. The remaining term in (129) is bounded as

$$\|\mathrm{D}\, A(x_{k+1}) - \mathrm{D}\, A(x_k)\| \leq \lambda_A \|x_{k+1} - x_k\| = \lambda_A \gamma_k \|G_k\|. \qquad \text{(see (2,119))} \qquad (130)$$

Assuming that

$$\nu(g, A, S, \rho, \rho') \geq 2\lambda_A \max_{k \in K} \gamma_k \|G_k\|, \qquad (131)$$

allows us to simplify the last line of (129) as

$$\left\| P_S P_{\mathrm{cone}(\partial g(x_{k+1}))^*}(\mathrm{D}\, A(x_k)^\top A(x_k)) \right\| \geq \frac{\nu(g, A, S, \rho, \rho')}{2} \|A(x_k)\|, \qquad (132)$$

which, after substituting in (127), yields that

$$\|A(x_k)\| \leq \frac{2}{\beta_k \nu(g, A, S, \rho, \rho')} \left(\|G_k\| + \lambda'_f + \lambda'_A \|y_k\| \right), \qquad (133)$$

and completes the proof of Lemma 4.