Inexact Augmented Lagrangian Framework for Non-Convex Optimization

Anonymous Author(s)
Affiliation
Address
email

Abstract

We propose a practical inexact augmented Lagrangian method (iALM) for nonconvex problems with nonlinear constrains. We characterize the total computational complexity of our method subject to a verifiable geometric condition, which is closely related to the Polyak-Lojsiewicz and Mangasarian-Fromowitz conditions.

In particular, when a first-order solver is used for the inner iterates, we prove that iALM finds a first-order stationary point with $\tilde{\mathcal{O}}(1/\epsilon^3)$ calls to the first-order oracle. If, in addition, the problem is smooth and a second-order solver is used for the inner iterates, iALM finds a second-order stationary point with $\tilde{\mathcal{O}}(1/\epsilon^5)$ calls to the second-order oracle. These complexity results match the known theoretical results in the literature with a simple, implementable and versatile algorithm.

We provide numerical evidence on large-scale machine learning problems, including the Burer-Monteiro factorization of semidefinite programs, and a novel nonconvex relaxation of the standard basis pursuit template. We verify our geometric condition in all these examples.

1 Introduction

We study the nonconvex optimization problem

$$\begin{cases} \min_{x\in\mathbb{R}^d} f(x) + g(x) \\ A(x) = 0, \end{cases} \tag{1}$$

where $f : \mathbb{R}^d \to \mathbb{R}$ is a continuously-differentiable nonconvex function and $A : \mathbb{R}^d \to \mathbb{R}^m$ is a nonlinear operator. We assume that $g : \mathbb{R}^d \to \mathbb{R}$ is a possibly nonsmooth but proximal-friendly convex function [49].

A host of problems in computer science [36, 39, 69], machine learning [42, 58], and signal processing [56, 57] naturally fall under the template (1), including max-cut, clustering, generalized eigenvalue decomposition, as well as the quadratic assignment problem (QAP) [69].

To solve (1), this paper proposes an intuitive and easy-to-implement augmented Lagrangian algorithm, and provides its total iteration complexity under an interpretable geometric condition. Before we elaborate on the results, let us first motivate (1) with an application to semidefinite programming (SDP):

Vignette: Burer-Monteiro splitting. A powerful convex relaxation for max-cut, clustering, and many others is provided by the SDP

$$\begin{cases} \min_{X\in\mathbb{S}^{d\times d}} \langle C, X\rangle \\ B(X) = b, \ X \succeq 0, \end{cases} \tag{2}$$

where $C \in \mathbb{R}^{d\times d}$, X is a positive semidefinite $d \times d$ 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 [54].

Since d is often large, many first- and second-order methods for solving such SDP's 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 is therefore to solve SDPs using little space and in a scalable fashion. The recent homotopy conditional gradient method, which is based on linear minimization oracles (LMOs), can solve (2) in a small space via sketching [68]. However, such LMO-based methods are extremely slow in obtaining accurate solutions.

A different approach for solving (1), dating back to [16, 17], is the so-called Burer-Monteiro (BM) factorization $X = UU^\top$, where $U \in \mathbb{R}^{d\times r}$ and r is selected according to the guidelines in [51, 2], which are shown to be optimal [62]. This factorization does not introduce any extraneous local minima [17] and, moreover, [15] established the connection between the local minimizers of the factorized problem (3) and the global minimizers for (2).

This factorization leads to the nonconvex problem

$$\begin{cases} \min\limits_{U\in\mathbb{R}^{d\times r}} \langle C, UU^\top\rangle \\ B(UU^\top) = b, \end{cases} \tag{3}$$

which can be easily written in the form of (1). To solve (3), the inexact Augmented Lagrangian method (iALM) is widely used [16, 17, 37], due to its cheap per iteration cost and its empirical success. Every (outer) iteration of iALM calls a solver to solve an intermediate augmented Lagrangian subproblem to near stationarity. The user is free in the choice of this solver, which could use first-order, such as the proximal gradient descent [49], or second-order information, such as the trust region method and BFGS [46].[1]

Unlike its convex counterpart [43, 38, 64], the convergence rate and the complexity of iALM for (3) are not well-understood, see Section 5 for a review of the related literature. Indeed, addressing this important theoretical gap is one of the contributions of our work.

Summary of contributions:

∘ We derive the convergence rate of iALM for solving (1) to first- or second-order optimality, and find the total iteration complexity of iALM using different solvers for the augmented Lagrangian subproblems. Our complexity bounds match the best theoretical results in optimization, see Section 5.

∘ Our results are future-proof in the sense that they are independent of the choice of solver called by iALM.

∘ We propose a geometric condition that simplifies the algorithmic analysis for iALM, and clarify its connection to well-known Polyak-Lojasiewicz [35] and Mangasarian-Fromovitz [5] conditions. We also verify this condition for key problems in Section 6.

Roadmap. Section 2 collects the main tools and our notation. We present the iALM in Section 3 and obtain its convergence rate to first- and second-order stationary points in Section 4, alongside their iteration complexities. We provide a comprehensive review of the literature and highlight our key differences in Section 5. Section 6 presents the numerical evidence and comparisons with the state-of-the-art techniques.

2 Preliminaries

Notation. We use the notation $\langle\cdot,\cdot\rangle$ and $\|\cdot\|$ for the standard inner product and the norm on $\mathbb{R}^d$. For matrices, $\|\cdot\|$ and $\|\cdot\|_F$ denote the spectral and the Frobenius norms, respectively. For the convex function $g : \mathbb{R}^d \to \mathbb{R}$, the subdifferential set at $x \in \mathbb{R}^d$ is denoted by $\partial g(x)$ and we will occasionally use the notation $\partial g(x)/\beta = \{z/\beta : z \in \partial g(x)\}$. When presenting iteration complexity results, we often use $\widetilde{O}(\cdot)$ which suppresses the logarithmic dependencies.

[1] Strictly speaking, BFGS is in fact a quasi-Newton method that emulates second-order information.

We use the indicator function $\delta_{\mathcal{X}} : \mathbb{R}^d \to \mathbb{R}$ of a set $\mathcal{X} \subset \mathbb{R}^d$, which takes x to

$$\delta_{\mathcal{X}}(x) = \begin{cases} 0 & x \in \mathcal{X} \\ \infty & x \notin \mathcal{X}. \end{cases} \tag{4}$$

The distance function from a point x to $\mathcal{X}$ is denoted by $\operatorname{dist}(x, \mathcal{X}) = \min_{z \in \mathcal{X}} \|x - z\|$. For integers $k_0 \le k_1$, we denote $[k_0 : k_1] = \{k_0, \dots, k_1\}$.

For an operator $A : \mathbb{R}^d \to \mathbb{R}^m$ with components $\{A_i\}_{i=1}^m$, we let $DA(x) \in \mathbb{R}^{m \times d}$ denote the Jacobian of A, where the ith row of $DA(x)$ is the gradient vector $\nabla A_i(x) \in \mathbb{R}^d$.

Smoothness. We require $f : \mathbb{R}^d \to \mathbb{R}$ and $A : \mathbb{R}^d \to \mathbb{R}^m$ to be smooth, namely, there exist $\lambda_f, \lambda_A \ge 0$ such that

$$\begin{aligned} \|\nabla f(x) - \nabla f(x')\| &\le \lambda_f \|x - x'\|, \\ \|DA(x) - DA(x')\| &\le \lambda_A \|x - x'\|, \end{aligned} \tag{5}$$

for every $x, x' \in \mathbb{R}^d$.

Augmented Lagrangian method (ALM). ALM is a classical algorithm, which first appeared in [32, 53] and extensively studied afterwards in [5, 10]. For solving (1), ALM suggests solving the problem

$$\min_x \max_y \ \mathcal{L}_\beta(x, y) + g(x), \tag{6}$$

where, for penalty weight $\beta > 0$, $\mathcal{L}_\beta$ is the corresponding augmented Lagrangian, defined as

$$\mathcal{L}_\beta(x, y) := f(x) + \langle A(x), y \rangle + \frac{\beta}{2}\|A(x)\|^2. \tag{7}$$

The minimax formulation in (6) naturally suggests the following algorithm for solving (1). For dual step sizes $\{\sigma_k\}_k$, consider the iterations

$$x_{k+1} \in \operatorname*{argmin}_x \ \mathcal{L}_\beta(x, y_k) + g(x), \tag{8}$$

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

However, computing x_{k+1} above requires solving the nonconvex problem (8) to optimality, which is typically intractable. Instead, it is often easier to find an approximate first- or second-order stationary point of (8).

Hence, we argue that by gradually improving the stationarity precision and increasing the penalty weight β above, we can reach a stationary point of the main problem in (1), as detailed in Section 3.

Optimality conditions. First-order necessary optimality conditions for (1) are well-studied. Indeed, $x \in \mathbb{R}^d$ is a first-order stationary point of (1) if there exists $y \in \mathbb{R}^m$ such that

$$\begin{cases} -\nabla f(x) - DA(x)^\top y \in \partial g(x) \\ A(x) = 0, \end{cases} \tag{9}$$

where $DA(x)$ is the Jacobian of A at x. Recalling (7), we observe that (9) is equivalent to

$$\begin{cases} -\nabla_x \mathcal{L}_\beta(x, y) \in \partial g(x) \\ A(x) = 0, \end{cases} \tag{10}$$

which is in turn the necessary optimality condition for (6). Inspired by this, we say that x is an (ϵ_f, β) first-order stationary point of (6) if there exists a $y \in \mathbb{R}^m$ such that

$$\begin{cases} \operatorname{dist}(-\nabla_x \mathcal{L}_\beta(x, y), \partial g(x)) \le \epsilon_f \\ \|A(x)\| \le \epsilon_f, \end{cases} \tag{11}$$

for $\epsilon_f \ge 0$. In light of (11), a suitable metric for evaluating the stationarity of a pair $(x, y) \in \mathbb{R}^d \times \mathbb{R}^m$ is

$$\operatorname{dist}\left(-\nabla_x \mathcal{L}_\beta(x, y), \partial g(x)\right) + \|A(x)\|, \tag{12}$$

which we use as the first-order stopping criterion. As an example, for a convex set $\mathcal{X} \subset \mathbb{R}^d$, suppose that $g = \delta_\mathcal{X}$ is the indicator function on $\mathcal{X}$. Let also $T_\mathcal{X}(x) \subseteq \mathbb{R}^d$ denote the tangent cone to $\mathcal{X}$ at x, and with $P_{T_\mathcal{X}(x)} : \mathbb{R}^d \to \mathbb{R}^d$ we denote the orthogonal projection onto this tangent cone. Then, for $u \in \mathbb{R}^d$, it is not difficult to verify that

$$\text{dist}\,(u, \partial g(x)) = \|P_{T_\mathcal{X}(x)}(u)\|. \tag{13}$$

When $g = 0$, a first-order stationary point $x \in \mathbb{R}^d$ of (1) is also second-order stationary if

$$\lambda_{\text{min}}(\nabla_{xx}\mathcal{L}_\beta(x, y)) \geq 0, \tag{14}$$

where $\nabla_{xx}\mathcal{L}_\beta$ is the Hessian of $\mathcal{L}_\beta$ with respect to x, and $\lambda_{\text{min}}(\cdot)$ returns the smallest eigenvalue of its argument. Analogously, x is an $(\epsilon_f, \epsilon_s, \beta)$ second-order stationary point if, in addition to (11), it holds that

$$\lambda_{\text{min}}(\nabla_{xx}\mathcal{L}_\beta(x, y)) \geq -\epsilon_s, \tag{15}$$

for $\epsilon_s \geq 0$. Naturally, for second-order stationarity, we use $\lambda_{\text{min}}(\nabla_{xx}\mathcal{L}_\beta(x, y))$ as the stopping criterion.

Smoothness lemma. This next result controls the smoothness of $\mathcal{L}_\beta(\cdot, y)$ for a fixed y. The proof is standard but nevertheless is included in Appendix C for completeness.

Lemma 2.1 (smoothness) *For fixed $y \in \mathbb{R}^m$ and $\rho, \rho' \geq 0$, it holds that*

$$\|\nabla_x \mathcal{L}_\beta(x, y) - \nabla_x \mathcal{L}_\beta(x', y)\| \leq \lambda_\beta \|x - x'\|, \tag{16}$$

for every $x, x' \in \{x'' : \|x''\| \leq \rho, \|A(x'')\| \leq \rho'\}$, where

$$\begin{aligned} \lambda_\beta &\leq \lambda_f + \sqrt{m}\lambda_A \|y\| + (\sqrt{m}\lambda_A \rho' + d\lambda_A'^2)\beta \\ &=: \lambda_f + \sqrt{m}\lambda_A \|y\| + \lambda''(A, \rho, \rho')\beta. \end{aligned} \tag{17}$$

Above, λ_f, λ_A were defined in (5) and

$$\lambda'_A := \max_{\|x\| \leq \rho} \|DA(x)\|. \tag{18}$$

3 Algorithm

To solve the equivalent formulation of (1) presented in (6), we propose the inexact ALM (iALM), detailed in Algorithm 1.

At the k^{th} iteration, Step 2 of Algorithm 1 calls a solver that finds an approximate stationary point of the augmented Lagrangian $\mathcal{L}_{\beta_k}(\cdot, y_k)$ with the accuracy of ϵ_{k+1}, and this accuracy gradually increases in a controlled fashion.

The increasing sequence of penalty weights $\{\beta_k\}_k$ and the dual update (Steps 4 and 5) are responsible for continuously enforcing the constraints in (1). The appropriate choice for $\{\beta_k\}_k$ will be specified in Sections 4.1 and 4.2.

The particular choice of the dual step sizes $\{\sigma_k\}_k$ in Algorithm 1 ensures that the dual variable y_k remains bounded, see [4] for a precedent in the ALM literature where a similar dual step size is considered.

4 Convergence Rate

In this section, we derive the total iteration complexity of Algorithm 1 for finding first-order and second-order stationary points of problem (1). All the proofs are deferred to Appendix A. Theorem 4.1 below characterizes the convergence rate of Algorithm 1 for finding stationary points in terms of the number of outer iterations.

Algorithm 1 Inexact ALM for solving (1)

```
Input: Non-decreasing, positive, unbounded sequence ${\beta_k}_{k\geq 1}$, stopping thresholds $\tau_f > 0$ and $\tau_s > 0$.
Initialization: Initial primal variable $x_1 \in \mathbb{R}^d$, initial dual variable $y_0 \in \mathbb{R}^m$, initial dual step size $\sigma_1 > 0$.
for $k = 1, 2, \dots$ do
    1. (Update tolerance) $\epsilon_{k+1} = 1/\beta_k$.
    2. (Inexact primal solution) Obtain $x_{k+1} \in \mathbb{R}^d$ such that
           $\text{dist}(-\nabla_x \mathcal{L}_{\beta_k}(x_{k+1}, y_k), \partial g(x_{k+1})) \leq \epsilon_{k+1}$
       for first-order stationarity
           $\lambda_{\text{min}}(\nabla_{xx} \mathcal{L}_{\beta_k}(x_{k+1}, y_k)) \geq -\epsilon_{k+1}$
       for second-order-stationarity, if $g = 0$ in (1).
    3. (Update dual step size)
           $\sigma_{k+1} = \sigma_1 \min\Big( \frac{\|A(x_1)\| \log^2 2}{\|A(x_{k+1})\| (k+1) \log^2(k+2)}, 1 \Big).$
    4. (Dual ascent) $y_{k+1} = y_k + \sigma_{k+1} A(x_{k+1})$.
    5. (Stopping criterion) If
           $\text{dist}(-\nabla_x \mathcal{L}_{\beta_k}(x_{k+1}), \partial g(x_{k+1})) + \|A(x_{k+1})\| \leq \tau_f,$
       for first-order stationarity and if also $\lambda_{\text{min}}(\nabla_{xx} \mathcal{L}_{\beta_k}(x_{k+1}, y_k)) \geq -\tau_s$ for second-order stationarity, then quit and return $x_{k+1}$ as an (approximate) stationary point of (1).
end for
```

Theorem 4.1 (convergence rate) *For integers $2 \leq k_0 \leq k_1$, consider the interval $K = [k_0 : k_1]$, and let $\{x_k\}_{k\in K}$ be the output sequence of Algorithm 1 on the interval K.*[2] *Let also $\rho := \sup_{k\in[K]} \|x_k\|$.*[3] *Suppose that f and A satisfy (5) and let*

$$\lambda'_f = \max_{\|x\|\leq\rho} \|\nabla f(x)\|, \qquad \lambda'_A = \max_{\|x\|\leq\rho} \|DA(x)\|, \tag{19}$$

be the (restricted) Lipschitz constants of f and A, respectively. With $\nu > 0$, assume that

$$\nu \|A(x_k)\| \leq \text{dist}\left(-DA(x_k)^\top A(x_k), \frac{\partial g(x_k)}{\beta_{k-1}} \right), \tag{20}$$

for every $k \in K$. We consider two cases:

- *If a first-order solver is used in Step 2, then x_k is an $(\epsilon_{k,f}, \beta_k)$ first-order stationary point of (1) with*

$$\begin{aligned} \epsilon_{k,f} &= \frac{1}{\beta_{k-1}} \left(\frac{2(\lambda'_f + \lambda'_A y_{\max})(1 + \lambda'_A \sigma_k)}{\nu} + 1 \right) \\ &=: \frac{Q(f, g, A, \sigma_1)}{\beta_{k-1}}, \end{aligned} \tag{21}$$

for every $k \in K$, where $y_{\max}(x_1, y_0, \sigma_1)$ is specified in (41) due to the limited space.

[2] The choice of $k_1 = \infty$ is valid here too.

[3] If necessary, to ensure that $\rho < \infty$, one can add a small factor of $\|x\|^2$ to $\mathcal{L}_\beta$ in (7). Then it is easy to verify that the iterates of Algorithm 1 remain bounded, provided that the penalty weight β is large enough, $\sup_x \|\nabla f(x)\|/\|x\| < \infty$, $\sup_x \|A(x)\| < \infty$, and $\sup_x \|DA(x)\| < \infty$.

- *If a second-order solver is used in Step 2, then x_k is an $(\epsilon_{k,f}, \epsilon_{k,s}, \beta_k)$ second-order stationary point of (1) with $\epsilon_{k,s}$ specified above and with*

$$\begin{aligned}\epsilon_{k,s} &= \epsilon_{k-1} + \sigma_k \sqrt{m}\lambda_A \frac{2\lambda'_f + 2\lambda'_A y_{\max}}{\nu \beta_{k-1}} \\ &= \frac{\nu + \sigma_k \sqrt{m}\lambda_A 2\lambda'_f + 2\lambda'_A y_{\max}}{\nu \beta_{k-1}} =: \frac{Q'(f,g,A,\sigma_1)}{\beta_{k-1}}.\end{aligned} \tag{22}$$

Loosely speaking, Theorem 4.1 states that Algorithm 1 converges to a (first- or second-) order stationary point of (1) at the rate of $1/\beta_k$, further specified in Sections 4.1 and 4.2. A few remarks are in order about Theorem 4.1.

Regularity. The key geometric condition in Theorem 4.1 is (20) which, broadly speaking, ensures that the primal updates of Algorithm 1 reduce the feasibility gap as the penalty weight β_k grows. We will verify this condition for several examples in Section 6.

This condition in (20) is closely related to those in the existing literature. In the special case where $g = 0$ in (1), it is easy to verify that (20) reduces to the Polyak-Lojasiewicz (PL) condition for minimizing $\|A(x)\|^2$ [35]. PL condition itself is a special case of Kurdyka-Lojasiewicz with $\theta = 1/2$, see [65, Definition 1.1]. When $g = 0$, it is also easy to see that (20) is weaker than the Mangasarian-Fromovitz (MF) condition in nonlinear optimization [12, Assumption 1]. Moreover, when g is the indicator on a convex set, (20) is a consequence of the *basic constraint qualification* in [55], which itself generalizes the MF condition to the case when g is an indicator function of a convex set. **Note:AE: I'm not sure if the claim about basic constraint qualification is true because our condition should locally hold rather globally. Could you add the exact equation number in [55] and double check this? Also consider double checking other claims in this paragraph.**

Loosely speaking, we may think of (20) as a local condition, which should hold within a neighborhood of the constraint set $\{x : A(x) = 0\}$ rather than everywhere in $\mathbb{R}^d$. There is a constant complexity algorithm in [12] to reach this so-called "information zone", which supplements Theorem 4.1. Lastly, in contrast to most conditions in the nonconvex optimization literature, such as [28], the condition in (20) appears to be easier to verify, as we see in Section 6.

AE: the spars review talks about the "Pong-Li" work. Fatih, do you know what is that?

(mfs:) I do not think this work is relevant to our condition but the template seems similar. We can mention in the related works instead. **Note:AE: Ok. Yes, consider adding it to the related work.**

Penalty method. A classical algorithm to solve (1) is the penalty method, which is characterized by the absence of the dual variable ($y = 0$) in (7). Indeed, ALM can be interpreted as an adaptive penalty or smoothing method with a variable center determined by the dual variable. It is worth noting that, with the same proof technique, one can establish the same convergence rate of Theorem 4.1 for the penalty method. However, while both methods have the same convergence rate in theory, iALM outperforms the penalty method in practice by virtue of its variable center and has been excluded from this presentation for this reason.

Computational complexity. Theorem 4.1 specifies the number of (outer) iterations that Algorithm 1 requires to reach a near-stationary point of problem (1) with a prescribed precision and, in particular, specifies the number of calls made to the solver in Step 2. In this sense, Theorem 4.1 does not fully capture the computational complexity of Algorithm 1, as it does not take into account the computational cost of the solver in Step 2.

To better understand the total iteration complexity of Algorithm 1, we consider two scenarios in the following. In the first scenario, we take the solver in Step 2 to be the Accelerated Proximal Gradient Method (APGM), a well-known first-order algorithm [30]. In the second scenario, we will use the second-order trust region method developed in [20].

4.1 First-Order Optimality

Let us first consider the case where the solver in Step 2 is is the first-order algorithm APGM, described in detail in [30]. At a high level, APGM makes use of $\nabla_x \mathcal{L}_\beta(x, y)$ in (7), the proximal operator prox_g, and the classical Nesterov acceleration [45] to reach first-order stationarity for the subproblem in (8). Suppose that $g = \delta_{\mathcal{X}}$ is the indicator function on a bounded convex set $\mathcal{X} \subset \mathbb{R}^d$ and let

$$\rho = \max_{x \in \mathcal{X}} \|x\|, \tag{23}$$

be the radius of a ball centered at the origin that includes $\mathcal{X}$. Then, adapting the results in [30] to our setup, APGM reaches x_k in Step 2 of Algorithm 1 after

$$\mathcal{O}\left(\frac{\lambda_{\beta_k}^2 \rho^2}{\epsilon_{k+1}}\right) \tag{24}$$

(inner) iterations, where λ_{β_k} denotes the Lipschitz constant of $\nabla_x \mathcal{L}_{\beta_k}(x, y)$, bounded in (17). For the clarity of the presentation, we have used a looser bound in (24) compared to [30]. Using (24), we derive the following corollary, describing the total iteration complexity of Algorithm 1 in terms of the number calls made to the first-order oracle in APGM.

Corollary 4.2 *For $b > 1$, let $\beta_k = b^k$ for every k. If we use APGM from [30] for Step 2 of Algorithm 1, the algorithm finds an (ϵ_f, β_k) first-order stationary point, after T calls to the first-order oracle, where*

$$T = \mathcal{O}\left(\frac{Q^3 \rho^2}{\epsilon^3} \log_b \left(\frac{Q}{\epsilon}\right)\right) = \tilde{\mathcal{O}}\left(\frac{Q^3 \rho^2}{\epsilon^3}\right). \tag{25}$$

For Algorithm 1 to reach a near-stationary point with an accuracy of ϵ_f in the sense of (11) and with the lowest computational cost, we therefore need to perform only one iteration of Algorithm 1, with β_1 specified as a function of ϵ_f by (21) in Theorem 4.1. In general, however, the constants in (21) are unknown and this approach is thus not feasible. Instead, the homotopy approach taken by Algorithm 1 ensures achieving the desired accuracy by gradually increasing the penalty weight.[4] This homotopy approach increases the computational cost of Algorithm 1 only by a factor logarithmic in the ϵ_f, as detailed in the proof of Corollary 4.2.

4.2 Second-Order Optimality

Let us now consider the second-order optimality case where the solver in Step 2 is the the trust region method developed in [20]. Trust region method minimizes a quadratic approximation of the function within a dynamically updated trust-region radius. Second-order trust region method that we consider in this section makes use of Hessian (or an approximation of Hessian) of the augmented Lagrangian in addition to first order oracles.

As shown in [47], 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 (1) with $g = 0$.

Let us compute the total computational complexity of Algorithm 1 with the trust region method in Step 2, in terms of the number of calls made to the second-order oracle. By adapting the result in [20] to our setup, we find that the number of (inner) iterations required in Step 2 of Algorithm 1 to produce x_{k+1} is

$$\mathcal{O}\left(\frac{\lambda_{\beta_k,H}^2 (\mathcal{L}_{\beta_k}(x_1, y) - \min_x \mathcal{L}_{\beta_k}(x, y))}{\epsilon_k^3}\right), \tag{26}$$

where $\lambda_{\beta,H}$ is the Lipschitz constant of the Hessian of the augmented Lagrangian, which is of the order of β, as can be proven similar to Lemma 2.1 and x_1 is the initial iterate of the given outer loop. In [20], the term $\mathcal{L}_\beta(x_1, y) - \min_x \mathcal{L}_\beta(x, y)$ is bounded by a constant independent of ϵ. We assume a uniform bound for this quantity for every β_k, instead of for one value of β_k as in [20]. Using (26) and Theorem 4.1, we arrive at the following:

Corollary 4.3 *For $b > 1$, let $\beta_k = b^k$ for every k. We assume that*

$$\mathcal{L}_\beta(x_1, y) - \min_x \mathcal{L}_\beta(x, y) \le L_u, \qquad \forall \beta. \tag{27}$$

If we use the trust region method from [20] for Step 2 of Algorithm 1, the algorithm finds an ϵ-second-order stationary point of (1) in T calls to the second-order oracle where

$$T = \mathcal{O}\left(\frac{L_u Q'^5}{\epsilon^5} \log_b \left(\frac{Q'}{\epsilon}\right)\right) = \widetilde{\mathcal{O}}\left(\frac{L_u Q'^5}{\epsilon^5}\right). \tag{28}$$

[4] In this context, homotopy loosely corresponds to the gradual enforcement of the constraints by increasing the penalty weight.

Note:AE: I can't remember: what is Q' and why isn't it defined? Before closing this section, we note that the remark after Corollary 4.2 applies here as well.

5 Related Work

ALM has a long history in the optimization literature, dating back to [32, 53]. In the special case of (1) with a convex function f and a linear operator A, standard, inexact, and linearized versions of ALM have been extensively studied [38, 43, 60, 64].

Classical works on ALM focused on the general template of (1) with nonconvex f and nonlinear A, with arguably stronger assumptions and required exact solutions to the subproblems of the form (8), which appear in Step 2 of Algorithm 1, see for instance [6].

A similar analysis was conducted in [26] for the general template of (1). The authors considered inexact ALM and proved convergence rates for the outer iterates, under specific assumptions on the initialization of the dual variable. However, unlike our results, the authors did not analyze how to solve the subproblems inexactly and they did not provide total complexity results and verifiable conditions.

Problem (1) with similar assumptions to us is also studied in [9] and [21] for first-order and second-order stationarity, respectively, with explicit iteration complexity analysis. As we have mentioned in Section 4, our iteration complexity results matches these theoretical algorithms with a simpler algorithm and a simpler analysis. In addition, these algorithms require setting final accuracies since they utilize this information in the algorithm. In contrast to [9, 21], Algorithm 1 does not set accuracies a priori.

[19] also considers the same template (1) for first-order stationarity with a penalty-type method instead of ALM. Even though the authors show $\mathcal{O}(1/\epsilon^2)$ complexity, this result is obtained by assuming that the penalty parameter remains bounded. We note that such an assumption can also be used to match our complexity results.

[12] studies the general template (1) with specific assumptions involving local error bound conditions for the (1). These conditions are studied in detail in [11], but their validity for general SDPs (2) has never been established. This work also lacks the total iteration complexity analysis presented here.

Another work [24] focused on solving (1) by adapting the primal-dual method of Chambolle and Pock [22]. The authors proved the convergence of the method and provided convergence rate by imposing error bound conditions on the objective function that do not hold for standard SDPs.

[16, 17] is the first work that proposes the splitting $X = UU^\top$ for solving SDPs of the form (2). Following these works, the literature on Burer-Monteiro (BM) splitting for the large part focused on using ALM for solving the reformulated problem (3).

However, this approach has a few drawbacks: First, it requires exact solutions in Step 2 of Algorithm 1 in theory, which in practice is replaced with inexact solutions. Second, their results only establish convergence without providing the rates. In this sense, our work provides a theoretical understanding of the BM splitting with inexact solutions to Step 2 of Algorithm 1 and complete iteration complexities.

[8, 50] are among the earliest efforts to show convergence rates for BM splitting, focusing on the special case of SDPs without any linear constraints. For these specific problems, they prove the convergence of gradient descent to global optima with convergence rates, assuming favorable initialization. These results, however, do not apply to general SDPs of the form (2) where the difficulty arises due to the linear constraints.

[7] focused on the quadratic penalty formulation of (1), namely,

$$\min_{X \succeq 0} \langle C, X \rangle + \frac{\mu}{2}\|B(x) - b\|^2, \tag{29}$$

which after BM splitting becomes

$$\min_{U \in \mathbb{R}^{d\times r}} \langle C, UU^\top \rangle + \frac{\mu}{2}\|B(UU^\top) - b\|^2, \tag{30}$$

for which they study the optimality of the second-order stationary points. These results are for establishing a connection between the stationary points of (30) and global optima of (29). In contrast, we focus on the relation of the stationary points of (6) to the constrained problem (1).

Another popular method for solving SDPs are due to [14, 13, 15], focusing on the case where the constraints in (1) can be written as a Riemannian manifold after BM splitting. In this case, the authors apply the Riemannian gradient descent and Riemannian trust region methods for obtaining first- and second-order stationary points, respectively. They obtain $\mathcal{O}(1/\epsilon^2)$ complexity for finding first-order stationary points and $\mathcal{O}(1/\epsilon^3)$ complexity for finding second-order stationary points.

While these complexities appear better than ours, the smooth manifold requirement in these works is indeed restrictive. In particular, this requirement holds for max-cut and generalized eigenvalue problems, but it is not satisfied for other important SDPs such as quadratic programming (QAP), optimal power flow and clustering with general affine constraints. In addition, as noted in [13], per iteration cost of their method for max-cut problem is an astronomical $\mathcal{O}(d^6)$.

Lastly, there also exists a line of work for solving SDPs in their original convex formulation, in a storage efficient way [44, 67, 68]. These works have global optimality guarantees by their virtue of directly solving the convex formulation. On the downside, these works require the use of eigenvalue routines and exhibit significantly slower convergence as compared to nonconvex approaches [34].

6 Numerical Evidence

We first begin with a caveat: It is known that quasi-Newton methods, such as BFGS and lBFGS, might not converge for nonconvex problems [25, 40]. For this reason, we have used the trust region method as the second-order solver in our analysis in Section 4, which is well-studied for nonconvex problems [20]. Empirically, however, BFGS and lBGFS are extremely successful and we have therefore opted for those solvers in this section since the subroutine does not affect Theorem 4.1 as long as the subsolver performs well in practice.

6.1 Clustering

Given data points $\{z_i\}_{i=1}^n$, the entries of the corresponding Euclidean distance matrix $D \in \mathbb{R}^{n\times n}$ are $D_{i,j} = \|z_i - z_j\|^2$. Clustering is then the problem of finding a co-association matrix $Y \in \mathbb{R}^{n\times n}$ such that $Y_{ij} = 1$ if points z_i and z_j are within the same cluster and $Y_{ij} = 0$ otherwise. In [52], the authors provide a SDP relaxation of the clustering problem, specified as

$$\begin{cases} \min\limits_{Y \in \mathbb{R}^{nxn}} \operatorname{tr}(DY) \\ Y\mathbf{1} = \mathbf{1},\ \operatorname{tr}(Y) = s,\ Y \succeq 0,\ Y \geq 0, \end{cases} \tag{31}$$

where s is the number of clusters and Y is both positive semidefinite and has nonnegative entries. Standard SDP solvers do not scale well with the number of data points n, since they often require projection onto the semidefinite cone with the complexity of $\mathcal{O}(n^3)$. We instead use the BM factorization to solve (31), sacrificing convexity to reduce the computational complexity. More specifically, we solve the program

$$\begin{cases} \min\limits_{V \in \mathbb{R}^{n\times r}} \operatorname{tr}(DVV^\top) \\ VV^\top\mathbf{1} = \mathbf{1},\ \ \|V\|_F^2 \leq s,\ \ V \geq 0, \end{cases} \tag{32}$$

where $\mathbf{1} \in \mathbb{R}^n$ is the vector of all ones. Note that $Y \geq 0$ in (31) is replaced above by the much stronger but easier-to-enforce constraint $V \geq 0$ in (32), see [37] for the reasoning behind this relaxation. Now, we can cast (32) as an instance of (1). Indeed, for every $i \leq n$, let $x_i \in \mathbb{R}^r$ denote the ith row of V. We next form $x \in \mathbb{R}^d$ with $d = nr$ by expanding the factorized variable V, namely,

$$x = [x_1^\top, \cdots, x_n^\top]^\top \in \mathbb{R}^d,$$

and then set

$$f(x) = \sum_{i,j=1}^{n} D_{i,j} \langle x_i, x_j \rangle, \qquad g = \delta_C,$$

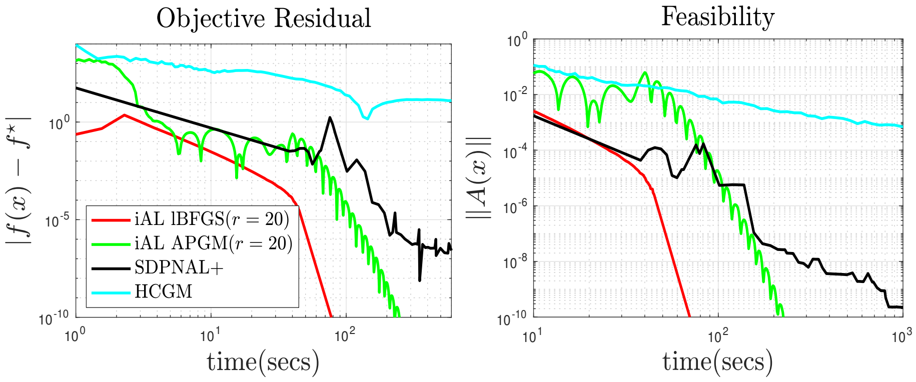

Figure 1: Convergence of different algorithms for k-Means clustering with fashion MNIST dataset. The solution rank for the template (31) is known and it is equal to number of clusters k (Theorem 1. [37]). As discussed in [59], setting rank $r > k$ leads more accurate reconstruction in expense of speed. Therefore, we set the rank to 20.

$$A(x) = [x_1^\top \sum_{j=1}^{n} x_j - 1, \cdots, x_n^\top \sum_{j=1}^{n} x_j - 1]^\top, \tag{33}$$

where C is the intersection of the positive orthant in $\mathbb{R}^d$ with the Euclidean ball of radius $\sqrt{s}$. In Appendix D, we somewhat informally verify that Theorem 4.1 applies to (1) with f, g, A specified above.

In our simulations, we use two different solvers for Step 2 of Algorithm 1, namely, APGM and lBFGS. APGM is a solver for nonconvex problems of the form (8) with convergence guarantees to first-order stationarity, as discussed in Section 4. lBFGS is a limited-memory version of BFGS algorithm in [27] that approximately leverages the second-order information of the problem. We compare our approach against the following convex methods:

- HCGM: Homotopy-based Conditional Gradient Method in [68] which directly solves (31).
- SDPNAL+: A second-order augmented Lagrangian method for solving SDP's with nonnegativity constraints [66].

As for the dataset, our experimental setup is similar to that described by [41]. We use the publicly-available fashion-MNIST data in [63], which is released as a possible replacement for the MNIST handwritten digits. Each data point is a 28×28 gray-scale image, associated with a label from ten classes, labeled from 0 to 9. First, we extract the meaningful features from this dataset using a simple two-layer neural network with a sigmoid activation function. Then, we apply this neural network to 1000 test samples from the same dataset, which gives us a vector of length 10 for each data point, where each entry represents the posterior probability for each class. Then, we form the ℓ_2 distance matrix D from these probability vectors. The results are depicted in Figure 1. We implemented 3 algorithms on MATLAB and used the software package for SDPNAL+ which contains mex files. It is predictable that the performance of our nonconvex approach would even improve by using mex files.

6.2 Basis Pursuit

Basis Pursuit (BP) finds sparsest solutions of an under-determined system of linear equations by solving

$$\begin{cases} \min_z \|z\|_1 \\ Bz = b, \end{cases} \tag{34}$$

where $B \in \mathbb{R}^{n \times d}$ and $b \in \mathbb{R}^n$. BP has found many applications in machine learning, statistics and signal processing [23, 18, 1]. Various primal-dual convex optimization algorithms are available in the literature to solve BP, including [60, 22]. We compare our algorithm against state-of-the-art primal-dual convex methods for solving (34), namely, Chambole-Pock [22], ASGARD [61] and ASGARD-DL [60].

Here, we take a different approach and cast (34) as an instance of (1). Note that any $z \in \mathbb{R}^d$ can be decomposed as $z = z^+ - z^-$, where $z^+, z^- \in \mathbb{R}^d$ are the positive and negative parts of z, respectively. Then consider the change of variables $z^+ = u_1^{\circ 2}$ and $z^- = u_2^{\circ 2} \in \mathbb{R}^d$, where $\circ$ denotes element-wise power. Next, we concatenate u_1 and u_2 as $x := [u_1^\top, u_2^\top]^\top \in \mathbb{R}^{2d}$ and define $\overline{B} := [B, -B] \in \mathbb{R}^{n \times 2d}$. Then, (34) is equivalent to (1) with

$$\begin{aligned} f(x) =& \|x\|^2, \quad g(x) = 0, \\ A(x) =& \overline{B} x^{\circ 2} - b. \end{aligned} \tag{35}$$

In Appendix E, we verify with minimal details that Theorem 4.1 indeed applies to (1) with the above f, A.

We draw the entries of B independently from a zero-mean and unit-variance Gaussian distribution. For a fixed sparsity level k, the support of $z_* \in \mathbb{R}^d$ and its nonzero amplitudes are also drawn from the standard Gaussian distribution. Then the measurement vector is created as $b = Bz + \epsilon$, where ϵ is the noise vector with entries drawn independently from the zero-mean Gaussian distribution with variance $\sigma^2 = 10^{-6}$.

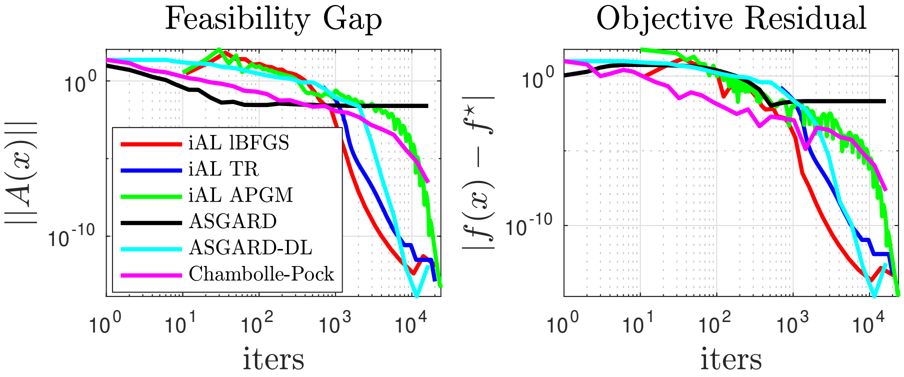

Figure 2: Convergence with different subsolvers for the aforementioned nonconvex relaxation.

The results are compiled in Figure 2. Clearly, the performance of Algorithm 1 with a second-order solver for BP is comparable to the rest. It is, indeed, interesting to see that these type of nonconvex relaxations gives the solution of convex one and first order methods succeed.

Discussion: The true potential of our reformulation is in dealing with more structured norms rather than ℓ_1, where computing the proximal operator is often intractable. One such case is the latent group lasso norm [48], defined as

$$\|z\|_\Omega = \sum_{i=1}^{I} \|z_{\Omega_i}\|,$$

where $\{\Omega_i\}_{i=1}^I$ are (not necessarily disjoint) index sets of $\{1, \cdots, d\}$. Although not studied here, we believe that the nonconvex framework presented in this paper can serve to solve more complicated problems, such as the latent group lasso. We leave this research direction for future work.

References

[1] S. Arora, M. Khodak, N. Saunshi, and K. Vodrahalli. A compressed sensing view of unsupervised text embeddings, bag-of-n-grams, and lstms. 2018.

[2] A. I. Barvinok. Problems of distance geometry and convex properties of quadratic maps. *Discrete & Computational Geometry*, 13(2):189–202, 1995.

[3] A. Beck and M. Teboulle. A fast iterative shrinkage-thresholding algorithm for linear inverse problems. *SIAM journal on imaging sciences*, 2(1):183–202, 2009.

[4] D. P. Bertsekas. On penalty and multiplier methods for constrained minimization. *SIAM Journal on Control and Optimization*, 14(2):216–235, 1976.

[5] D. P. Bertsekas. Constrained optimization and lagrange multiplier methods. *Computer Science and Applied Mathematics, Boston: Academic Press, 1982*, 1982.

[6] D. P. Bertsekas. *Constrained optimization and Lagrange multiplier methods*. Academic press, 2014.

[7] S. Bhojanapalli, N. Boumal, P. Jain, and P. Netrapalli. Smoothed analysis for low-rank solutions to semidefinite programs in quadratic penalty form. *arXiv preprint arXiv:1803.00186*, 2018.

[8] S. Bhojanapalli, A. Kyrillidis, and S. Sanghavi. Dropping convexity for faster semi-definite optimization. In *Conference on Learning Theory*, pages 530–582, 2016.

[9] E. G. Birgin, J. Gardenghi, J. M. Martinez, S. Santos, and P. L. Toint. Evaluation complexity for nonlinear constrained optimization using unscaled kkt conditions and high-order models. *SIAM Journal on Optimization*, 26(2):951–967, 2016.

[10] E. G. Birgin and J. M. Mart_nez. *Practical augmented Lagrangian methods for constrained optimization*, volume 10. SIAM, 2014.

[11] J. Bolte, T. P. Nguyen, J. Peypouquet, and B. W. Suter. From error bounds to the complexity of first-order descent methods for convex functions. *Mathematical Programming*, 165(2):471–507, 2017.

[12] J. Bolte, S. Sabach, and M. Teboulle. Nonconvex lagrangian-based optimization: monitoring schemes and global convergence. *Mathematics of Operations Research*, 2018.

[13] N. Boumal, P.-A. Absil, and C. Cartis. Global rates of convergence for nonconvex optimization on manifolds. *arXiv preprint arXiv:1605.08101*, 2016.

[14] N. Boumal, B. Mishra, P.-A. Absil, and R. Sepulchre. Manopt, a matlab toolbox for optimization on manifolds. *The Journal of Machine Learning Research*, 15(1):1455–1459, 2014.

[15] N. Boumal, V. Voroninski, and A. Bandeira. The non-convex burer-monteiro approach works on smooth semidefinite programs. In *Advances in Neural Information Processing Systems*, pages 2757–2765, 2016.

[16] S. Burer and R. D. Monteiro. A nonlinear programming algorithm for solving semidefinite programs via low-rank factorization. *Mathematical Programming*, 95(2):329–357, 2003.

[17] S. Burer and R. D. Monteiro. Local minima and convergence in low-rank semidefinite programming. *Mathematical Programming*, 103(3):427–444, 2005.

[18] E. J. Candès and M. B. Wakin. An introduction to compressive sampling. *IEEE signal processing magazine*, 25(2):21–30, 2008.

[19] C. Cartis, N. I. Gould, and P. L. Toint. On the evaluation complexity of composite function minimization with applications to nonconvex nonlinear programming. *SIAM Journal on Optimization*, 21(4):1721–1739, 2011.

[20] C. Cartis, N. I. Gould, and P. L. Toint. Complexity bounds for second-order optimality in unconstrained optimization. *Journal of Complexity*, 28(1):93–108, 2012.

[21] C. Cartis, N. I. Gould, and P. L. Toint. Optimality of orders one to three and beyond: characterization and evaluation complexity in constrained nonconvex optimization. *Journal of Complexity*, 2018.

[22] A. Chambolle and T. Pock. A first-order primal-dual algorithm for convex problems with applications to imaging. *Journal of mathematical imaging and vision*, 40(1):120–145, 2011.

[23] S. S. Chen, D. L. Donoho, and M. A. Saunders. Atomic decomposition by basis pursuit. *SIAM review*, 43(1):129–159, 2001.

[24] C. Clason, S. Mazurenko, and T. Valkonen. Acceleration and global convergence of a first-order primal–dual method for nonconvex problems. *arXiv preprint arXiv:1802.03347*, 2018.

[25] Y.-H. Dai. Convergence properties of the bfgs algoritm. *SIAM Journal on Optimization*, 13(3):693–701, 2002.

[26] D. Fernandez and M. V. Solodov. Local convergence of exact and inexact augmented lagrangian methods under the second-order sufficient optimality condition. *SIAM Journal on Optimization*, 22(2):384–407, 2012.

[27] R. Fletcher. *Practical methods of optimization*. John Wiley & Sons, 2013.

[28] F. Flores-Bazán, F. Flores-Bazán, and C. Vera. A complete characterization of strong duality in nonconvex optimization with a single constraint. *Journal of Global Optimization*, 53(2):185–201, 2012.

[29] R. Ge, C. Jin, P. Netrapalli, A. Sidford, et al. Efficient algorithms for large-scale generalized eigenvector computation and canonical correlation analysis. In *International Conference on Machine Learning*, pages 2741–2750, 2016.

[30] S. Ghadimi and G. Lan. Accelerated gradient methods for nonconvex nonlinear and stochastic programming. *Mathematical Programming*, 156(1-2):59–99, 2016.

[31] I. J. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio. Generative Adversarial Networks. *ArXiv e-prints*, June 2014.

[32] M. R. Hestenes. Multiplier and gradient methods. *Journal of optimization theory and applications*, 4(5):303–320, 1969.

[33] A. Ilyas, A. Jalal, E. Asteri, C. Daskalakis, and A. G. Dimakis. The Robust Manifold Defense: Adversarial Training using Generative Models. *arXiv e-prints*, page arXiv:1712.09196, Dec. 2017.

[34] M. Jaggi. Revisiting frank-wolfe: Projection-free sparse convex optimization. In *ICML (1)*, pages 427–435, 2013.

[35] H. Karimi, J. Nutini, and M. Schmidt. Linear convergence of gradient and proximal-gradient methods under the polyak-łojasiewicz condition. In *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*, pages 795–811. Springer, 2016.

[36] S. Khot and A. Naor. Grothendieck-type inequalities in combinatorial optimization. *arXiv preprint arXiv:1108.2464*, 2011.

[37] B. Kulis, A. C. Surendran, and J. C. Platt. Fast low-rank semidefinite programming for embedding and clustering. In *Artificial Intelligence and Statistics*, pages 235–242, 2007.

[38] G. Lan and R. D. Monteiro. Iteration-complexity of first-order augmented lagrangian methods for convex programming. *Mathematical Programming*, 155(1-2):511–547, 2016.

[39] L. Lovász. Semidefinite programs and combinatorial optimization. In *Recent advances in algorithms and combinatorics*, pages 137–194. Springer, 2003.

[40] W. F. Mascarenhas. The bfgs method with exact line searches fails for non-convex objective functions. *Mathematical Programming*, 99(1):49–61, 2004.

[41] D. G. Mixon, S. Villar, and R. Ward. Clustering subgaussian mixtures by semidefinite programming. *arXiv preprint arXiv:1602.06612*, 2016.

[42] E. Mossel, J. Neeman, and A. Sly. Consistency thresholds for the planted bisection model. In *Proceedings of the forty-seventh annual ACM symposium on Theory of computing*, pages 69–75. ACM, 2015.

[43] V. Nedelcu, I. Necoara, and Q. Tran-Dinh. Computational complexity of inexact gradient augmented lagrangian methods: application to constrained mpc. *SIAM Journal on Control and Optimization*, 52(5):3109–3134, 2014.

[44] Y. Nesterov. Primal-dual subgradient methods for convex problems. *Mathematical programming*, 120(1):221–259, 2009.

[45] Y. E. Nesterov. A method for solving the convex programming problem with convergence rate o (1/kˆ 2). In *Dokl. Akad. Nauk SSSR*, volume 269, pages 543–547, 1983.

[46] J. Nocedal and S. Wright. *Numerical Optimization*. Springer Series in Operations Research and Financial Engineering. Springer New York, 2006.

[47] M. Nouiehed, J. D. Lee, and M. Razaviyayn. Convergence to second-order stationarity for constrained non-convex optimization. *arXiv preprint arXiv:1810.02024*, 2018.

[48] G. Obozinski, L. Jacob, and J.-P. Vert. Group lasso with overlaps: the latent group lasso approach. *arXiv preprint arXiv:1110.0413*, 2011.

[49] N. Parikh, S. Boyd, et al. Proximal algorithms. *Foundations and Trends in Optimization*, 1(3):127–239, 2014.

[50] D. Park, A. Kyrillidis, S. Bhojanapalli, C. Caramanis, and S. Sanghavi. Provable burer-monteiro factorization for a class of norm-constrained matrix problems. *arXiv preprint arXiv:1606.01316*, 2016.

[51] G. Pataki. On the rank of extreme matrices in semidefinite programs and the multiplicity of optimal eigenvalues. *Mathematics of operations research*, 23(2):339–358, 1998.

[52] J. Peng and Y. Wei. Approximating K–means–type clustering via semidefinite programming. *SIAM J. Optim.*, 18(1):186–205, 2007.

[53] M. J. Powell. A method for nonlinear constraints in minimization problems. *Optimization*, pages 283–298, 1969.

[54] P. Raghavendra. Optimal algorithms and inapproximability results for every csp? In *Proceedings of the fortieth annual ACM symposium on Theory of computing*, pages 245–254. ACM, 2008.

[55] R. T. Rockafellar. Lagrange multipliers and optimality. *SIAM review*, 35(2):183–238, 1993.

[56] A. Singer. Angular synchronization by eigenvectors and semidefinite programming. *Applied and computational harmonic analysis*, 30(1):20, 2011.

[57] A. Singer and Y. Shkolnisky. Three-dimensional structure determination from common lines in cryo-em by eigenvectors and semidefinite programming. *SIAM journal on imaging sciences*, 4(2):543–572, 2011.

[58] L. Song, A. Smola, A. Gretton, and K. M. Borgwardt. A dependence maximization view of clustering. In *Proceedings of the 24th international conference on Machine learning*, pages 815–822. ACM, 2007.

[59] M. Tepper, A. M. Sengupta, and D. Chklovskii. Clustering is semidefinitely not that hard: Nonnegative sdp for manifold disentangling. *Journal of Machine Learning Research*, 19(82), 2018.

[60] Q. Tran-Dinh, A. Alacaoglu, O. Fercoq, and V. Cevher. An adaptive primal-dual framework for nonsmooth convex minimization. *arXiv preprint arXiv:1808.04648*, 2018.

[61] Q. Tran-Dinh, O. Fercoq, and V. Cevher. A smooth primal-dual optimization framework for nonsmooth composite convex minimization. *SIAM Journal on Optimization*, 28(1):96–134, 2018.

[62] I. Waldspurger and A. Waters. Rank optimality for the burer-monteiro factorization. *arXiv preprint arXiv:1812.03046*, 2018.

[63] H. Xiao, K. Rasul, and R. Vollgraf. Fashion-mnist: a novel image dataset for benchmarking machine learning algorithms, 2017.

[64] Y. Xu. Iteration complexity of inexact augmented lagrangian methods for constrained convex programming. *arXiv preprint arXiv:1711.05812v2*, 2017.

[65] Y. Xu and W. Yin. A globally convergent algorithm for nonconvex optimization based on block coordinate update. *Journal of Scientific Computing*, 72(2):700–734, 2017.

[66] L. Yang, D. Sun, and K.-C. Toh. Sdpnal+: a majorized semismooth newton-cg augmented lagrangian method for semidefinite programming with nonnegative constraints. *Mathematical Programming Computation*, 7(3):331–366, 2015.

[67] A. Yurtsever, Q. T. Dinh, and V. Cevher. A universal primal-dual convex optimization framework. In *Advances in Neural Information Processing Systems*, pages 3150–3158, 2015.

[68] A. Yurtsever, O. Fercoq, F. Locatello, and V. Cevher. A conditional gradient framework for composite convex minimization with applications to semidefinite programming. *arXiv preprint arXiv:1804.08544*, 2018.

[69] Q. Zhao, S. E. Karisch, F. Rendl, and H. Wolkowicz. Semidefinite programming relaxations for the quadratic assignment problem. *Journal of Combinatorial Optimization*, 2(1):71–109, 1998.

A Proof of Theorem 4.1

For every $k \geq 2$, recall from (7) and Step 2 of Algorithm 1 that x_k satisfies

$$\begin{aligned}
&\operatorname{dist}(-\nabla f(x_k) - DA(x_k)^\top y_{k-1} \\
&\qquad - \beta_{k-1} DA(x_k)^\top A(x_k), \partial g(x_k)) \\
&= \operatorname{dist}(-\nabla_x \mathcal{L}_{\beta_{k-1}}(x_k, y_{k-1}), \partial g(x_k)) \leq \epsilon_k.
\end{aligned} \tag{36}$$

With an application of the triangle inequality, it follows that

$$\begin{aligned}
&\operatorname{dist}(-\beta_{k-1} DA(x_k)^\top A(x_k), \partial g(x_k)) \\
&\qquad \leq \|\nabla f(x_k)\| + \|DA(x_k)^\top y_{k-1}\| + \epsilon_k,
\end{aligned} \tag{37}$$

which in turn implies that

$$\begin{aligned}
&\operatorname{dist}(-DA(x_k)^\top A(x_k), \partial g(x_k)/\beta_{k-1}) \\
&\leq \frac{\|\nabla f(x_k)\|}{\beta_{k-1}} + \frac{\|DA(x_k)^\top y_{k-1}\|}{\beta_{k-1}} + \frac{\epsilon_k}{\beta_{k-1}} \\
&\leq \frac{\lambda'_f + \lambda'_A \|y_{k-1}\| + \epsilon_k}{\beta_{k-1}},
\end{aligned} \tag{38}$$

where λ'_f, λ'_A were defined in (19). We next translate (38) into a bound on the feasibility gap $\|A(x_k)\|$. Using the regularity condition (20), the left-hand side of (38) can be bounded below as

$$\operatorname{dist}(-DA(x_k)^\top A(x_k), \partial g(x_k)/\beta_{k-1}) \geq \nu \|A(x_k)\|. \qquad \text{(see (20))} \tag{39}$$

By substituting (39) back into (38), we find that

$$\|A(x_k)\| \leq \frac{\lambda'_f + \lambda'_A \|y_{k-1}\| + \epsilon_k}{\nu \beta_{k-1}}. \tag{40}$$

In words, the feasibility gap is directly controlled by the dual sequence $\{y_k\}_k$. We next establish that the dual sequence is bounded. Indeed, for every $k \in K$, note that

$$\begin{aligned}
\|y_k\| &= \|y_0 + \sum_{i=1}^{k} \sigma_i A(x_i)\| \quad \text{(Step 5 of Algorithm 1)} \\
&\leq \|y_0\| + \sum_{i=1}^{k} \sigma_i \|A(x_i)\| \qquad \text{(triangle inequality)} \\
&\leq \|y_0\| + \sum_{i=1}^{k} \frac{\|A(x_1)\| \log^2 2}{k \log^2(k+1)} \quad \text{(Step 4)} \\
&\leq \|y_0\| + c\|A(x_1)\| \log^2 2 =: y_{\max},
\end{aligned} \tag{41}$$

where

$$c \geq \sum_{i=1}^{\infty} \frac{1}{k \log^2(k+1)}. \tag{42}$$

Substituting (41) back into (40), we reach

$$\begin{aligned}
\|A(x_k)\| &\leq \frac{\lambda'_f + \lambda'_A y_{\max} + \epsilon_k}{\nu \beta_{k-1}} \\
&\leq \frac{2\lambda'_f + 2\lambda'_A y_{\max}}{\nu \beta_{k-1}},
\end{aligned} \tag{43}$$

where the second line above holds if k_0 is large enough, which would in turn guarantees that $\epsilon_k = 1/\beta_{k-1}$ is sufficiently small since $\{\beta_k\}_k$ is increasing and unbounded. It remains to control the first term in (12). To that end, after recalling Step 2 of Algorithm 1 and applying the triangle inequality, we can write that

$$\begin{aligned} &\mathrm{dist}(-\nabla_x \mathcal{L}_{\beta_{k-1}}(x_k, y_k), \partial g(x_k)) \\ &\le \mathrm{dist}(-\nabla_x \mathcal{L}_{\beta_{k-1}}(x_k, y_{k-1}), \partial g(x_k)) \\ &+ \|\nabla_x \mathcal{L}_{\beta_{k-1}}(x_k, y_k) - \nabla_x \mathcal{L}_{\beta_{k-1}}(x_k, y_{k-1})\|. \end{aligned} \tag{44}$$

The first term on the right-hand side above is bounded by ϵ_k, by Step 5 of Algorithm 1. For the second term on the right-hand side of (44), we write that

$$\begin{aligned} &\|\nabla_x \mathcal{L}_{\beta_{k-1}}(x_k, y_k) - \nabla_x \mathcal{L}_{\beta_{k-1}}(x_k, y_{k-1})\| \\ &= \|DA(x_k)^\top (y_k - y_{k-1})\| \qquad \text{(see (7))} \\ &\le \lambda'_A \|y_k - y_{k-1}\| \qquad \text{(see (19))} \\ &= \lambda'_A \sigma_k \|A(x_k)\| \qquad \text{(see Step 5 of Algorithm 1)} \\ &\le \frac{2\lambda'_A \sigma_k}{\nu \beta_{k-1}} (\lambda'_f + \lambda'_A y_{\max}). \qquad \text{(see (43))} \end{aligned} \tag{45}$$

By combining (44,45), we find that

$$\begin{aligned} &\mathrm{dist}(\nabla_x \mathcal{L}_{\beta_{k-1}}(x_k, y_k), \partial g(x_k)) \\ &\le \frac{2\lambda'_A \sigma_k}{\nu \beta_{k-1}} (\lambda'_f + \lambda'_A y_{\max}) + \epsilon_k. \end{aligned} \tag{46}$$

By combining (43,46), we find that

$$\begin{aligned} &\mathrm{dist}(-\nabla_x \mathcal{L}_{\beta_{k-1}}(x_k, y_k), \partial g(x_k)) + \|A(x_k)\| \\ &\le \left(\frac{2\lambda'_A \sigma_k}{\nu \beta_{k-1}} (\lambda'_f + \lambda'_A y_{\max}) + \epsilon_k \right) \\ &\qquad + 2 \left(\frac{\lambda'_f + \lambda'_A y_{\max}}{\nu \beta_{k-1}} \right). \end{aligned} \tag{47}$$

Applying $\sigma_k \le \sigma_1$, we find that

$$\begin{aligned} &\mathrm{dist}(-\nabla_x \mathcal{L}_{\beta_{k-1}}(x_k, y_k), \partial g(x_k)) + \|A(x_k)\| \\ &\le \frac{2\lambda'_A \sigma_1 + 2}{\nu \beta_{k-1}} (\lambda'_f + \lambda'_A y_{\max}) + \epsilon_k. \end{aligned} \tag{48}$$

For the second part of the theorem, we use the Weyl's inequality and Step 5 of Algorithm 1 to write

$$\begin{aligned} \lambda_{\text{min}}(\nabla_{xx} \mathcal{L}_{\beta_{k-1}}(x_k, y_{k-1})) &\ge \lambda_{\text{min}}(\nabla_{xx} \mathcal{L}_{\beta_{k-1}}(x_k, y_k)) \\ &- \sigma_k \| \sum_{i=1}^m A_i(x_k) \nabla^2 A_i(x_k) \|. \end{aligned} \tag{49}$$

The first term on the right-hand side is lower bounded by $-\epsilon_{k-1}$ by Step 2 of Algorithm 1. We next bound the second term on the right-hand side above as

$$\begin{aligned} &\sigma_k \| \sum_{i=1}^m A_i(x_k) \nabla^2 A_i(x_k) \| \\ &\le \sigma_k \sqrt{m} \max_i \|A_i(x_k)\| \|\nabla^2 A_i(x_k)\| \\ &\le \sigma_k \sqrt{m} \lambda_A \frac{2\lambda'_f + 2\lambda'_A y_{\max}}{\nu \beta_{k-1}}, \end{aligned}$$

where the last inequality is due to (5,43). Plugging into (49) gives

$$\begin{aligned} &\lambda_{\text{min}}(\nabla_{xx} \mathcal{L}_{\beta_{k-1}}(x_k, y_{k-1})) \\ &\ge -\epsilon_{k-1} - \sigma_k \sqrt{m} \lambda_A \frac{2\lambda'_f + 2\lambda'_A y_{\max}}{\nu \beta_{k-1}}, \end{aligned}$$

which completes the proof of Theorem 4.1.

B Proof of Corollary 4.2

Let K denote the number of (outer) iterations of Algorithm 1 and let ϵ_f denote the desired accuracy of Algorithm 1, see (11). Recalling Theorem 4.1, we can then write that

$$\epsilon_f = \frac{Q}{\beta_K}, \tag{50}$$

or, equivalently, $\beta_K = Q/\epsilon_f$. We now count the number of total (inner) iterations T of Algorithm 1 to reach the accuracy ϵ_f. From (17) and for sufficiently large k, recall that $\lambda_{\beta_k} \leq \lambda'' \beta_k$ is the smoothness parameter of the augmented Lagrangian. Then, from (24) ad by summing over the outer iterations, we bound the total number of (inner) iterations of Algorithm 1 as

$$\begin{aligned} T &= \sum_{k=1}^{K} \mathcal{O}\left(\frac{\lambda^2_{\beta_{k-1}} \rho^2}{\epsilon_k}\right) \\ &= \sum_{k=1}^{K} \mathcal{O}\left(\beta^3_{k-1}\rho^2\right) \qquad \text{(Step 1 of Algorithm 1)} \\ &\leq \mathcal{O}\left(K\beta^3_{K-1}\rho^2\right) \qquad (\{\beta_k\}_k \text{ is increasing}) \\ &\leq \mathcal{O}\left(\frac{KQ^3\rho^2}{\epsilon_f^3}\right). \qquad \text{(see (50))} \end{aligned} \tag{51}$$

In addition, if we specify $\beta_k = b^k$ for all k, we can further refine T. Indeed,

$$\beta_K = b^K \implies K = \log_b\left(\frac{Q}{\epsilon_f}\right), \tag{52}$$

which, after substituting into (51) gives the final bound in Corollary 4.2.

C Proof of Lemma 2.1

Note that

$$\mathcal{L}_\beta(x,y) = f(x) + \sum_{i=1}^{m} y_i A_i(x) + \frac{\beta}{2}\sum_{i=1}^{m}(A_i(x))^2, \tag{53}$$

which implies that

$$\begin{aligned} &\nabla_x \mathcal{L}_\beta(x,y) \\ &= \nabla f(x) + \sum_{i=1}^{m} y_i \nabla A_i(x) + \frac{\beta}{2}\sum_{i=1}^{m} A_i(x)\nabla A_i(x) \\ &= \nabla f(x) + DA(x)^\top y + \beta DA(x)^\top A(x), \end{aligned} \tag{54}$$

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

$$\begin{aligned} \nabla_x^2 \mathcal{L}_\beta(x,y) = \nabla^2 f(x) &+ \sum_{i=1}^{m}\left(y_i + \beta A_i(x)\right)\nabla^2 A_i(x) \\ &+ \beta \sum_{i=1}^{m} \nabla A_i(x) \nabla A_i(x)^\top. \end{aligned} \tag{55}$$

It follows that

$$\begin{aligned} &\|\nabla_x^2 \mathcal{L}_\beta(x,y)\| \\ &\leq \|\nabla^2 f(x)\| + \max_i \|\nabla^2 A_i(x)\| \left(\|y\|_1 + \beta\|A(x)\|_1\right) \\ &\qquad + \beta\sum_{i=1}^{m}\|\nabla A_i(x)\|^2 \\ &\leq \lambda_h + \sqrt{m}\lambda_A\left(\|y\| + \beta\|A(x)\|\right) + \beta\|DA(x)\|_F^2. \end{aligned} \tag{56}$$

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

$$\|\nabla_x^2 \mathcal{L}_\beta(x,y)\| \le \lambda_f + \sqrt{m}\lambda_A \left(\|y\| + \beta\rho'\right) + \beta \max_{\|x\|\le\rho} \|DA(x)\|_F^2, \tag{57}$$

which completes the proof of Lemma 2.1.

D Clustering

We only verify the condition in (20) here. Note that

$$A(x) = VV^\top \mathbf{1} - \mathbf{1}, \tag{58}$$

$$\begin{aligned} DA(x) &= \begin{bmatrix} w_{1,1}x_1^\top & \cdots & w_{1,n}x_1^\top \\ \vdots & & \\ w_{n,1}x_n^\top & \cdots & w_{n,n}1x_n^\top \end{bmatrix} \\ &= \left[\begin{array}{ccc} V & \cdots & V \end{array}\right] + \begin{bmatrix} x_1^\top & & \\ & \ddots & \\ & & x_n^\top \end{bmatrix}, \end{aligned} \tag{59}$$

where $w_{i.i} = 2$ and $w_{i,j} = 1$ for $i \neq j$. In the last line above, n copies of V appear and the last matrix above is block-diagonal. For x_k, define V_k accordingly and let $x_{k,i}$ be the ith row of V_k. Consequently,

$$\begin{aligned} DA(x_k)^\top A(x_k) &= \begin{bmatrix} (V_k^\top V_k - I_n)V_k^\top \mathbf{1} \\ \vdots \\ (V_k^\top V_k - I_n)V_k^\top \mathbf{1} \end{bmatrix} \\ &\quad + \begin{bmatrix} x_{k,1}(V_kV_k^\top \mathbf{1} - \mathbf{1})_1 \\ \vdots \\ x_{k,n}(V_kV_k^\top \mathbf{1} - \mathbf{1})_n \end{bmatrix}, \end{aligned} \tag{60}$$

where $I_n \in \mathbb{R}^{n\times n}$ is the identity matrix. Let us make a number of simplifying assumptions. First, we assume that $\|x_k\| < \sqrt{s}$ (which can be enforced in the iterates by replacing C with $(1-\epsilon)C$ for a small positive ϵ in the subproblems). Under this assumption, it follows that

$$(\partial g(x_k))_i = \begin{cases} 0 & (x_k)_i > 0 \\ \{a : a \le 0\} & (x_k)_i = 0, \end{cases} \qquad i \le d. \tag{61}$$

Second, we assume that V_k has nearly orthonormal columns, namely, $V_k^\top V_k \approx I_n$. This can also be enforced in each iterate of Algorithm 1 and naturally corresponds to well-separated clusters. While a more fine-tuned argument can remove these assumptions, they will help us simplify the presentation here. Under these assumptions, the (squared) right-hand side of (20) becomes

$$\begin{aligned} &\operatorname{dist}\left(-DA(x_k)^\top A(x_k), \frac{\partial g(x_k)}{\beta_{k-1}}\right)^2 \\ &= \left\| \left(-DA(x_k)^\top A(x_k)\right)_+ \right\|^2 \qquad (a_+ = \max(a,0)) \\ &= \left\| \begin{bmatrix} x_{k,1}(V_kV_k^\top \mathbf{1} - \mathbf{1})_1 \\ \vdots \\ x_{k,n}(V_kV_k^\top \mathbf{1} - \mathbf{1})_n \end{bmatrix} \right\|^2 \qquad (x_k \in C \Rightarrow x_k \ge 0) \\ &= \sum_{i=1}^n \|x_{k,i}\|^2 (V_kV_k^\top \mathbf{1} - \mathbf{1})_i^2 \\ &\ge \min_i \|x_{k,i}\|^2 \cdot \sum_{i=1}^n (V_kV_k^\top \mathbf{1} - \mathbf{1})_i^2 \\ &= \min_i \|x_{k,i}\|^2 \cdot \|V_kV_k^\top \mathbf{1} - \mathbf{1}\|^2. \end{aligned} \tag{62}$$

Therefore, given a prescribed ν, ensuring $\min_i \|x_{k,i}\| \geq \nu$ guarantees (20). When the algorithm is initialized close enough to the constraint set, there is indeed no need to separately enforce (62). In practice, often n exceeds the number of true clusters and a more intricate analysis is required to establish (20) by restricting the argument to a particular subspace of $\mathbb{R}^n$.

E Basis Pursuit

We only verify the regularity condition in (20) for (1) with f, A, g specified in (35). Note that

$$DA(x) = 2\overline{B}\text{diag}(x), \tag{63}$$

where diag$(x) \in \mathbb{R}^{2d \times 2d}$ is the diagonal matrix formed by x. The left-hand side of (20) then reads as

$$\begin{aligned}
&\text{dist}\left(-DA(x_k)^\top A(x_k), \frac{\partial g(x_k)}{\beta_{k-1}}\right) \\
&= \text{dist}\left(-DA(x_k)^\top A(x_k), \{0\}\right) \qquad (g \equiv 0) \\
&= \|DA(x_k)^\top A(x_k)\| \\
&= 2\|\text{diag}(x_k)\overline{B}^\top(\overline{B}x_k^{\circ 2} - b)\|. \qquad \text{(see (63))}
\end{aligned} \tag{64}$$

To bound the last line above, let x_* be a solution of (1) and note that $\overline{B}x_*^{\circ 2} = b$ by definition. Let also $z_k, z_* \in \mathbb{R}^d$ denote the vectors corresponding to x_k, x_*. Corresponding to x_k, also define $u_{k,1}, u_{k,2}$ naturally and let $|z_k| = u_{k,1}^{\circ 2} + u_{k,2}^{\circ 2} \in \mathbb{R}^d$ be the vector of amplitudes of z_k. To simplify matters, let us assume also that B is full-rank. We then rewrite the norm in the last line of (64) as

$$\begin{aligned}
&\|\text{diag}(x_k)\overline{B}^\top(\overline{B}x_k^{\circ 2} - b)\|^2 \\
&= \|\text{diag}(x_k)\overline{B}^\top\overline{B}(x_k^{\circ 2} - x_*^{\circ 2})\|^2 \qquad (\overline{B}x_*^{\circ 2} = b) \\
&= \|\text{diag}(x_k)\overline{B}^\top B(x_k - x_*)\|^2 \\
&= \|\text{diag}(u_{k,1})B^\top B(z_k - z_*)\|^2 \\
&\qquad + \|\text{diag}(u_{k,2})B^\top B(z_k - z_*)\|^2 \\
&= \|\text{diag}(u_{k,1}^{\circ 2} + u_{k,2}^{\circ 2})B^\top B(z_k - z_*)\|^2 \\
&= \|\text{diag}(|z_k|)B^\top B(z_k - z_*)\|^2 \\
&\geq \eta_n(B\text{diag}(|z_k|))^2\|B(z_k - z_*)\|^2 \\
&= \eta_n(B\text{diag}(|z_k|))^2\|Bz_k - b\|^2 \qquad (Bz_* = \overline{B}x_*^{\circ 2} = b) \\
&\geq \min_{|T|=n} \eta_n(B_T) \cdot |z_{k,(n)}|^2\|Bz_k - b\|^2,
\end{aligned} \tag{65}$$

where $\eta_n(\cdot)$ returns the nth largest singular value of its argument. In the last line above, B_T is the restriction of B to the columns indexed by T of size n. Moreover, $z_{k,(n)}$ is the nth largest entry of z in magnitude. Given a prescribed ν, (20) therefore holds if

$$|z_{k,(n)}| \geq \frac{\nu}{2\sqrt{\min_{|T|=n} \eta_n(B_T)}}, \tag{66}$$

for every iteration k. If Algorithm 1 is initialized close enough to the solution z^* and the entries of z^* are sufficiently large in magnitude, there will be no need to directly enforce (66).

E.1 ℓ_∞ Denoising with a Generative Prior

(mfs): We need Fabian's input here.

The authors of [33] have proposed to project onto the range of a Generative Adversarial network (GAN) [31], as a way to defend against adversarial examples. For a given noisy observation $x^* + \eta$, they consider a projection in the ℓ_2 norm. We instead propose to use our augmented Lagrangian method to denoise in the ℓ_∞ norm, a much harder task:

$$\begin{aligned}
\min_{x,z} \quad & \|x^* + \eta - x\|_\infty \\
\text{s.t.} \quad & x = G(z).
\end{aligned} \tag{67}$$

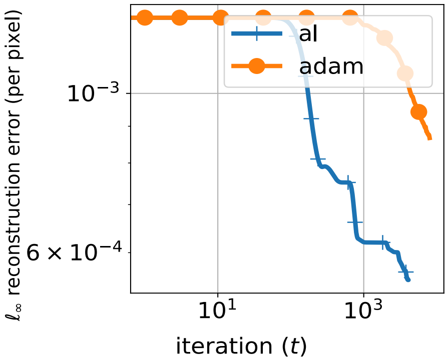

Figure 3: Augmented Lagrangian vs Adam for ℓ_∞ denoising (left). ℓ_2 vs ℓ_∞ denoising as defense against adversarial examples

We use a pretrained generator for the MNIST dataset, given by a standard deconvolutional neural network architecture. We compare the succesful optimizer Adam against our method. Our algorithm involves two forward/backward passes through the network, as oposed to Adam that requires only one. For this reason we let our algorithm run for 4000 iterations, and Adam for 8000 iterations. For a particular example, we plot the objective value vs iteration count in figure E.1. Our method successfully minimizes the objective value, while Adam does not succeed.

E.2 Generalized Eigenvalue Problem

Generalized eigenvalue problem has extensive applications in machine learning, statistics and data analysis [29]. The well-known nonconvex formulation of the problem is [15] given by

$$\begin{cases} \min\limits_{x\in\mathbb{R}^n} x^\top Cx \\ x^\top Bx = 1, \end{cases} \tag{68}$$

where $B, C \in \mathbb{R}^{n\times n}$ are symmetric matrices and B is positive definite, namely, $B \succ 0$. The generalized eigenvector computation is equivalent to performing principal component analysis (PCA) of C in the norm B. It is also equivalent to computing the top eigenvector of symmetric matrix $S = B^{-1/2}CB^{1/2}$ and multiplying the resulting vector by $B^{-1/2}$. However, for large values of n, computing $B^{-1/2}$ is extremely expensive. The natural convex SDP relaxation for (68) involves lifting $Y = xx^\top$ and removing the nonconvex $\text{rank}(Y) = 1$ constraint, namely,

$$\begin{cases} \min\limits_{Y\in\mathbb{R}^{n\times n}} \text{tr}(CY) \\ \text{tr}(BY) = 1, \quad X \succeq 0. \end{cases} \tag{69}$$

Here, however, we opt to directly solve (68) because it fits into our template with

$$\begin{aligned} f(x) =& x^\top Cx, \quad g(x) = 0, \\ A(x) =& x^\top Bx - 1. \end{aligned} \tag{70}$$

We compare our approach against three different methods: manifold based Riemannian gradient descent and Riemannian trust region methods in [13] and the linear system solver in [29], abbrevated as GenELin. We have used Manopt software package in [**?**] for the manifold based methods. For GenELin, we have utilized Matlab's backslash operator as the linear solver. The results are compiled in Figure 4.

Here, we verify the regularity condition in (20) for problem (68). Note that

$$DA(x) = (2Bx)^\top. \tag{71}$$

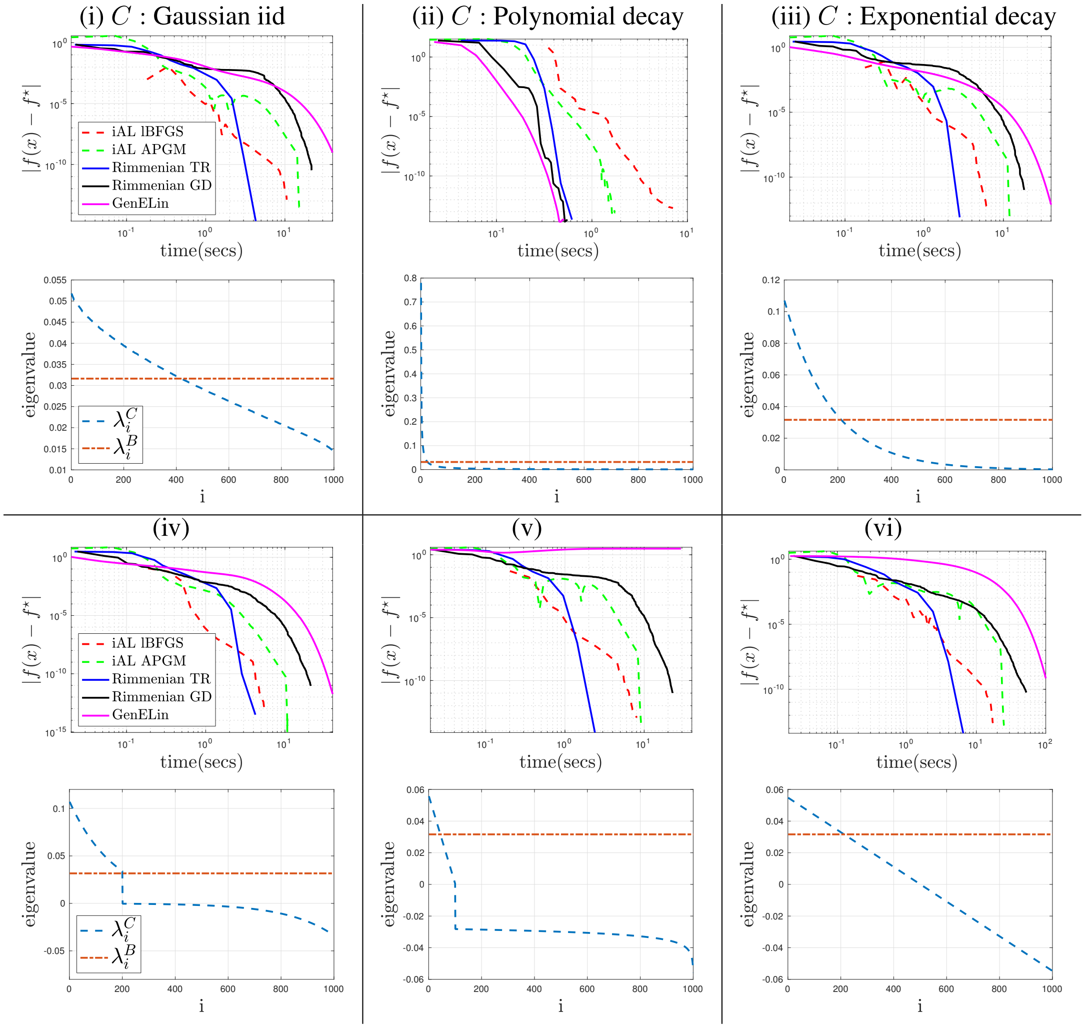

Figure 4: *(Top)* Objective convergence for calculating top generalized eigenvalue and eigenvector of B and C. *(Bottom)* Eigenvalue structure of the matrices. For (i),(ii) and (iii), C is positive semidefinite; for (iv), (v) and (vi), C contains negative eigenvalues. [(i): Generated by taking symmetric part of iid Gaussian matrix. (ii): Generated by randomly rotating diag$(1^{-p}, 2^{-p}, \cdots, 1000^{-p})(p = 1)$. (iii): Generated by randomly rotating diag$(10^{-p}, 10^{-2p}, \cdots, 10^{-1000p})(p = 0.0025)$.]

Therefore,

$$
\begin{aligned}
\operatorname{dist}\left(-DA(x_k)^\top A(x_k), \frac{\partial g(x_k)}{\beta_{k-1}}\right)^2 &= \operatorname{dist}\left(-DA(x_k)^\top A(x_k), \{0\}\right)^2 \qquad (g \equiv 0) \\
&= \|DA(x_k)^\top A(x_k)\|^2 \\
&= \|2Bx_k(x_k^\top Bx_k - 1)\|^2 \qquad \text{(see (71))} \\
&= 4(x_k^\top Bx_k - 1)^2 \|Bx_k\|^2 \\
&= 4\|Bx_k\|^2 \|A(x_k)\|^2 \qquad \text{(see (70))} \\
&\geq \eta_{\min}(B)^2 \|x_k\|^2 \|A(x_k)\|^2, \qquad (72)
\end{aligned}
$$

where $\eta_{\min}(B)$ is the smallest eigenvalue of the positive definite matrix B. Therefore, for a prescribed ν, the regularity condition in (20) holds with $\|x_k\| \geq \nu/\eta_{min}$ for every k. If the algorithm is initialized close enough to the constraint set, there will be again no need to directly enforce this latter condition.