Faster Projection-Free Augmented Lagrangian Methods via Weak Proximal Oracle

Dan Garber
Technion

Tsur Livney
Technion

Shoham Sabach
Technion

Abstract

This paper considers a convex composite optimization problem with affine constraints, which includes problems that take the form of minimizing a smooth convex objective function over the intersection of (simple) convex sets, or regularized with multiple (simple) functions. Motivated by high-dimensional applications in which exact projection/proximal computations are not tractable, we propose a *projection-free* augmented Lagrangian-based method, in which primal updates are carried out using a *weak proximal oracle* (WPO). In an earlier work, WPO was shown to be more powerful than the standard *linear minimization oracle* (LMO) that underlies conditional gradient-based methods (aka Frank-Wolfe methods). Moreover, WPO is computationally tractable for many high-dimensional problems of interest, including those motivated by recovery of low-rank matrices and tensors, and optimization over polytopes which admit efficient LMOs. The main result of this paper shows that under a certain curvature assumption (which is weaker than strong convexity), our WPO-based algorithm achieves an ergodic rate of convergence of $O(1/T)$ for both the objective residual and feasibility gap. This result, to the best of our knowledge, improves upon the $O(1/\sqrt{T})$ rate for existing LMO-based projection-free methods for this class of problems. Empirical experiments on a low-rank and sparse covariance matrix estimation task and the Max Cut semidefinite relaxation demonstrate that of our method can outperform state-of-the-art LMO-based Lagrangian-based methods.

Proceedings of the 26th International Conference on Artificial Intelligence and Statistics (AISTATS) 2023, Valencia, Spain. PMLR: Volume 206.

1 Introduction

Throughout the paper, we consider the following minimization problem

$$\min_{\mathbf{x}\in\mathbb{E}_1,\mathbf{y}\in\mathbb{E}_2} f(\mathbf{x})+\mathcal{R}_{\mathcal{X}}(\mathbf{x})+\mathcal{R}_{\mathcal{Y}}(\mathbf{y}) \text{ s.t. } \mathcal{A}\mathbf{x}=\mathbf{y}, \quad \text{(OP)}$$

where $\mathbb{E}_1$ and $\mathbb{E}_2$ are finite Euclidean spaces, $f:\mathbb{E}_1\to\mathbb{R}$ is a convex and β-smooth function, $\mathcal{R}_{\mathcal{X}}:\mathbb{E}_1\to(-\infty,\infty]$ and $\mathcal{R}_{\mathcal{Y}}:\mathbb{E}_2\to(-\infty,\infty]$ are proper, lower semi-continuous and convex functions, and $\mathcal{A}:\mathbb{E}_1\to\mathbb{E}_2$ is a linear mapping.

Problems that fall into the model (OP) appear in many interesting and important active research areas such as machine learning, signal processing, statistics, and more. For example, the recovery of a matrix or a tensor which is both sparse and of low rank is useful in problems such as covariance matrix estimation Andrews (1991); Driscoll and Kraay (1998); Richard et al. (2012), graph and image denoising Buades et al. (2005a,b); Zhang et al. (2017b, 2020) and link prediction Liben-Nowell and Kleinberg (2003); Lü and Zhou (2011); Zhang et al. (2017a). More applications of Problem (OP) can be found in brain mapping Ogawa et al. (1992); Lancaster et al. (2000); Gramfort et al. (2013) and multiple sequence alignment Corpet (1988); Chenna et al. (2003); Katoh and Toh (2008); Yen et al. (2016).

An important family of efficient methods for solving problems in the form of model (OP) are Lagrangian-based methods, and most notably augmented Lagrangian methods Mizoguchi et al. (1960); Hestenes (1969); Powell (1969). Starting with the classical proximal method of multipliers Rockafellar (1976), and until more recently Sabach and Teboulle (2019); De Marchi (2022); Dhingra et al. (2018); Chambolle and Pock (2016), such methods, which are based on proximal/projection computations (due to the nonsmooth functions $\mathcal{R}_{\mathcal{X}}$ and $\mathcal{R}_{\mathcal{Y}}$ in Problem (OP)) have been successfully developed and corresponding provable convergence rates have been established. However, in many cases of interest such proximal/projection computations are not tractable in high-dimensional problems, for instance when either $\mathcal{R}_{\mathcal{X}}$ or $\mathcal{R}_{\mathcal{Y}}$ is a nuclear norm regularizer for matrices, which underlies many recovery problems of low-rank matrices and tensors (see, for in-

stance, Candès and Recht (2009); Candès et al. (2011); Gandy et al. (2011)), or an indicator function for a polytope. Thus, with the growing interest in recent years in so-called *projection-free* methods, which are mostly based on the use of linear minimization oracles (LMO) instead of proximal/projection oracles through the Frank-Wolfe method (aka conditional gradient, see for instance Jaggi (2013)), and are often much more efficient to implement for high-dimensional problems (e.g., in case $\mathcal{R}_{\mathcal{X}}$ or $\mathcal{R}_{\mathcal{Y}}$ is a nuclear norm regularizer or an indicator function for a polytope which captures some well-studied combinatorial structure, see for instance Jaggi (2013); Hazan and Kale (2012)), have been studied Liu et al. (2019); Yurtsever et al. (2019); Silveti-Falls et al. (2020). However, these methods suffer from slow convergence rates compared to their proximal/projection-based counterparts, with the worst-case guaranteed convergence rate being at best $O(1/\sqrt{T})$, where T is the number of iterations executed. This rate is not known to be improvable even under additional standard curvature assumptions such as strong convexity of the function $f(\cdot)$ in Problem (OP)[1]. A recent attempt to obtain faster projection-free methods under relatively mild assumptions has been made in Gidel et al. (2018), however as we discuss in detail in the appendix (see Section A), there is a major problem with their proof which does not seem easily fixable. We also refer the interested reader to the excellent discussions in Gidel et al. (2018) on major issues with other previous attempts to prove faster rates for projection-free methods.

For the simpler problem of minimizing a smooth convex objective function over a convex and compact set, and in particular in case the feasible set is either a polytope or a nuclear norm ball of matrices or a spectrahedron (set of positive semidefinite matrices with unit trace), several recent works showed how simple modifications of the Frank-Wolfe method can lead to provably faster convergence rates, under standard curvature assumptions, see for instance Garber and Hazan (2016); Lacoste-Julien and Jaggi (2015); Beck and Shtern (2017); Garber (2016); Allen-Zhu et al. (2017); Garber and Kaplan (2019). Thus, in the context of the significantly more complex Problem (OP), our work considers the following natural question:

Can we design a projection-free augmented Lagrangian-based method that, at least under standard curvature assumptions, improves upon the current $O(1/\sqrt{T})$ convergence rate?

We answer this question on the affirmative side by providing a projection-free method with a rate of $O(1/T)$, both in terms of the objective function residual and the feasibility gap of the affine constraint in Problem (OP).

Our approach departs from previous projection-free methods which guarantee only a rate of $O(1/\sqrt{T})$ in two aspects. First, as already suggested, we make a curvature assumption on Problem (OP): we introduce a curvature condition we call *primal quadratic gap* (see definition in the sequel). In particular, this condition holds whenever the smooth function $f(\cdot)$ in Problem (OP) is strongly convex, but also holds in case $f(\cdot)$ is a composition of a strongly convex function with a linear transformation (e.g., a least squares objective, which need not be strongly convex) and $\mathcal{R}_{\mathcal{X}}, \mathcal{R}_{\mathcal{Y}}$ are indicators for polytopes. Second, while previous projection-free methods rely on the availability of a linear minimization oracle (LMO), in this work we consider a slightly stronger oracle which was already considered in recent works Allen-Zhu et al. (2017); Garber et al. (2021); Garber and Kaplan (2019) (these however do not apply to problems such as Problem (OP), which includes affine constraints), namely the *weak proximal oracle* (WPO)[2]. In a nutshell, this oracle solves a certain relaxed version of the proximal/projection problem, which can still be much more efficient to solve than the standard proximal/projection problem, but can provide more informative directions than that of the LMO. Two prime examples for the efficiency of implementing the WPO are when (i) $\mathcal{R}_{\mathcal{X}}$ or $\mathcal{R}_{\mathcal{Y}}$ is an indicator function for a polytope which admits and efficient LMO, then the WPO could be implemented based on a single call to the LMO of the polytope, and (ii) $\mathcal{R}_{\mathcal{X}}$ or $\mathcal{R}_{\mathcal{Y}}$ is an indicator function/regularizer corresponding to the matrix nuclear norm and a unique low-rank optimal solution exists, then implementing the WPO corresponds to a low-rank SVD computation with rank matching that of the low-rank optimal solution, which is much more efficient than proximal/projection computation, which generally requires a full-rank SVD (see a detailed discussion in Allen-Zhu et al. (2017); Garber and Kaplan (2019); Garber et al. (2021)).

The combination of the two ingredients: a curvature condition and the weak proximal oracle, to obtain faster convergence rates for projection-free methods should not come as a surprise since it was already instrumental in achieving similar improvements for projection-free methods in settings that do not include affine constraints as in model (OP), see for instance Garber and Hazan (2016); Lacoste-Julien and Jaggi (2015); Garber (2016); Allen-Zhu et al. (2017); Garber and Kaplan (2019)[3]. To the best of our knowledge,

[1] This is not surprising since it is well-known that in general, and as opposed to projection/proximal-based methods, the Frank-Wolfe method does not benefit from strong convexity, see for instance discussions in Garber and Hazan (2016); Garber (2016); Allen-Zhu et al. (2017).

[2] The term "weak proximal oracle" was originally coined in Garber and Kaplan (2019).

[3] While technically Garber and Hazan (2016); Lacoste-Julien and Jaggi (2015) rely on the use of a standard LMO for the feasible set (which they assume to be a polytope), as we show in the sequel, the way they use the output of the LMO to construct the new descent direction is very similar to the implementation of a WPO. In particular, we rely on observations from Garber and Hazan (2016) to construct an efficient WPO for polytopes.

this is the first time such an approach is used for a problem of the form of model (OP).

1.1 Paper organization

In Section 2, we discuss the augmented Lagrangian approach for solving the Problem (OP), present the Primal Quadratic Gap property (PQG) needed for our algorithm's analysis. We also recall the notion of the Weak Proximal Oracle that will be used in our algorithm and discuss its implementation in several important scenarios. In the Appendix (see Section D), we give examples of problems of interest, for which our algorithm might be appealing to use. In Section 3, we develop our algorithm and prove our main rate of convergence result. In Section 4, we demonstrate the empirical performance of our algorithm.

2 Preliminaries

2.1 Notation

Throughout the paper, we will use the following notation for simplifying the presentation and developments. We use the following compact notations $\mathbf{q} := \begin{bmatrix}\mathbf{x}\\ \mathbf{y}\end{bmatrix} \equiv (\mathbf{x},\mathbf{y}) \in \mathbb{E}_1 \times \mathbb{E}_2 =: \mathbb{E}$, and $\mathcal{R}_{\mathcal{Q}}(\mathbf{q}) := \mathcal{R}_{\mathcal{X}}(\mathbf{x}) + \mathcal{R}_{\mathcal{Y}}(\mathbf{y})$. In addition, we define the linear mapping $\mathcal{K} : \mathbb{E} \to \mathbb{E}_2 := [\mathcal{A}, -\mathcal{I}]$, where $\mathcal{I}$ is an identity linear mapping. This way we can compactly write the constraint $\mathcal{A}\mathbf{x} = \mathbf{y}$ as $\mathcal{K}\mathbf{q} = \mathbf{0}$. For any finite Euclidean space $\mathbb{V}$ of dimension n, we denote the standard Euclidean inner product of any two points $\mathbf{x} := [x_1, \ldots, x_n]^\top$ and $\mathbf{y} := [y_1, \ldots, y_n]^\top \in \mathbb{V}$, by $\langle \mathbf{x}, \mathbf{y}\rangle := \sum_{i=1}^n x_i \cdot y_i$, and we let $\|\mathbf{x}\| := \sqrt{\langle \mathbf{x}, \mathbf{x}\rangle}$ and $\|\mathbf{x}\|_1 := \sum_{i=1}^n |x_i|$ denote the standard Euclidean norm and the ℓ_1 norm, respectively. For any two finite Euclidean spaces $\mathbb{V}_1$ and $\mathbb{V}_2$, the *spectral norm* of the linear mapping $\mathcal{T} : \mathbb{V}_1 \to \mathbb{V}_2$, is denoted by $\|\mathcal{T}\| := \max_{\mathbf{x}\in\mathbb{V}_1} \{\|\mathcal{T}\mathbf{x}\| : \|\mathbf{x}\| = 1\}$. While both norms are denoted the same, throughout the paper it will be clear from the context which of the norms is used in each appearance. In addition, $\|\mathbf{X}\|_F$ denotes the Frobenius norm of a matrix $\mathbf{X}$, $\|\mathbf{X}\|_{\text{nuc}}$ denotes its nuclear norm and $\text{tr}(\mathbf{X})$ denotes its trace. We denote by $\mathbb{S}_+^d$ the set of all positive semidefinite matrices of size $d \times d$ and by $\mathcal{S}_\tau$ the spectrahedron of all matrices in $\mathbb{S}_+^d$ with trace τ. We use the notation $\delta_C(\cdot)$ to denote the indicator function of the set C.

2.2 The Augmented Lagrangian

The *augmented Lagrangian* (AL) of Problem (OP) with a multiplier (dual variable) $\mathbf{w} \in \mathbb{E}_2$ is defined as

$$\begin{aligned}\mathcal{L}_\rho(\mathbf{q},\mathbf{w}) &\equiv \mathcal{L}_\rho(\mathbf{x},\mathbf{y},\mathbf{w})\\ &:= f(\mathbf{x}) + \mathcal{R}_{\mathcal{Q}}(\mathbf{q}) + \langle \mathbf{w}, \mathcal{K}\mathbf{q}\rangle + \frac{\rho}{2}\|\mathcal{K}\mathbf{q}\|^2, \quad (1)\end{aligned}$$

where $\rho > 0$ is a penalty parameter associated to the linear equality constraint. Note that the (standard) Lagrangian of Problem (OP) is recovered when $\rho = 0$.

We also require the two following standard assumptions which we assume to hold true throughout the paper.

Assumption 1. *The augmented Lagrangian has a saddle point, i.e., there exists a point $(\boldsymbol{q}^*, \boldsymbol{w}^*) \in \mathbb{E} \times \mathbb{E}_2$ satisfying*

$$\mathcal{L}_\rho(\boldsymbol{q}^*, \boldsymbol{w}) \le \mathcal{L}_\rho(\boldsymbol{q}^*, \boldsymbol{w}^*) \le \mathcal{L}_\rho(\boldsymbol{q}, \boldsymbol{w}^*), \quad (2)$$

for all $\boldsymbol{q} \in \mathbb{E}$ and $\boldsymbol{w} \in \mathbb{E}_2$.

Assumption 2. *(Slater's Condition) There exist $\boldsymbol{x} \in ri(dom(f) \cap dom(\mathcal{R}_{\mathcal{X}}))$ and $\boldsymbol{y} \in ri(dom(\mathcal{R}_{\mathcal{Y}}))$ such that $\mathcal{A}\boldsymbol{x} = \boldsymbol{y}$, where ri denotes the relative interior of a set.*

We denote by $\mathcal{P}^*$ the set of optimal solutions of the primal problem (OP), and by $\mathcal{D}^*$ the set of optimal solutions of the associated dual problem.

Under the above two assumptions, and thanks to the convexity of Problem (OP), strong duality holds. As a result, and thanks to Proposition 2, the set of saddle points of $\mathcal{L}_\rho$ is non-empty and corresponds to the set of *all* pairs $(\mathbf{q}^*, \mathbf{w}^*)$, where $\mathbf{q}^* \in \mathcal{P}^*$ and $\mathbf{w} \in \mathcal{D}^*$ (see also Appendix B.3).

In order to find solutions of Problem (OP), we will solve the following equivalent saddle point problem

$$\min_{\mathbf{q}\in\mathbb{E}} \max_{\mathbf{w}\in\mathbb{E}_2} \mathcal{L}_\rho(\mathbf{q},\mathbf{w}), \quad (3)$$

whose optimal solutions are the saddle points of $\mathcal{L}_\rho$.

From now on, we will denote the optimal objective function value by $\mathcal{L}_\rho(\mathbf{q}^*, \mathbf{w}^*)$, for all saddle points $(\mathbf{q}^*, \mathbf{w}^*)$, and for short we will write $\mathcal{L}_\rho^*$.

The *smooth* part of the *augmented Lagrangian* (SAL) of Problem (OP) is defined, for any $\mathbf{q} \in \mathbb{E}$ and $\mathbf{w} \in \mathbb{E}_2$, by

$$S(\mathbf{q},\mathbf{w}) \equiv S(\mathbf{x},\mathbf{y},\mathbf{w}) := f(\mathbf{x}) + \langle \mathbf{w}, \mathcal{K}\mathbf{q}\rangle + \frac{\rho}{2}\|\mathcal{K}\mathbf{q}\|^2. \quad (4)$$

The following result will be essential to our developments in the sequel. Its simple proof is deferred to the appendix.

Lemma 2.1. *The function $\boldsymbol{q} \to S(\boldsymbol{q},\boldsymbol{w})$, for any fixed $\boldsymbol{w} \in \mathbb{E}_2$, is smooth with parameter $\beta_S = \beta + \rho(\|\mathcal{A}\| + 1)^2$.*

2.2.1 Primal Quadratic Gap

We now study a new property of the smooth function $\mathbf{q} \to S(\cdot,\mathbf{w})$, for any fixed $\mathbf{w} \in \mathbb{E}_2$, which we refer to as the *Primal Quadratic Gap* (PQG).

Definition 1. (Primal Quadratic Gap) We say that Problem (OP) satisfies the *Primal Quadratic Gap* property with a parameter $\alpha_S > 0$, if for any $\mathbf{q} \in \text{dom}(\mathcal{R}_{\mathcal{Q}})$, the point $\mathbf{q}^* := \arg\min_{\mathbf{q}^*\in\mathcal{P}^*} \|\mathbf{q} - \mathbf{q}^*\|^2 \in \mathcal{P}^*$ satisfies, for all $\mathbf{w} \in \mathbb{E}_2$, the following inequality

$$\langle \mathbf{q}^* - \mathbf{q}, \nabla_{\mathbf{q}} S(\mathbf{q},\mathbf{w})\rangle \le S(\mathbf{q}^*,\mathbf{w}) - S(\mathbf{q},\mathbf{w}) - \frac{\alpha_S}{2}\|\mathbf{q}^* - \mathbf{q}\|^2. \quad (5)$$

Algorithm	Oracle	Rate	Oracle implementation in polytope setup	Oracle implementation in nuclear norm setup	Assumptions
Proximal Method of Multipliers Rockafellar (1976); Sabach and Teboulle (2019)	proximal	$O(1/T)$	projection	full-rank SVD	
Accelerated Primal Dual Chambolle and Pock (2016)	proximal	$O(1/T^2)$	projection	full-rank SVD	strong convexity
Conditional Gradient Augmented Lagrangian Yurtsever et al. (2019)	LMO	$O(1/\sqrt{T})$	LMO	rank-one SVD	
This work	weak proximal	$O(1/T)$	LMO plus convex quadratic opt. over simplex	rank($\mathbf{x}^*$)-SVD	primal quadratic gap (weaker than strong convexity)

Table 1: Comparison of augmented Lagrangian-based methods with different optimization oracles for Problem (OP). The "Rate" column specifies the convergence rate only in terms of the number of iterations T and suppresses all other quantities. "Polytope setup" in the forth column refers to a setting in which $\mathcal{R}_\mathcal{X}$ is an indicator function of some compact and convex polytope and $\mathcal{R}_\mathcal{Y}$ is some proximal friendly function. "Nuclear norm setup" in the fifth column refers to a setting in which $\mathcal{R}_\mathcal{X}$ is a matrix nuclear norm regularizer and $\mathcal{R}_\mathcal{Y}$ is some proximal friendly function. $\mathbf{x}^*$ denotes the optimal solution, which is assumed to be unique. Both columns specify the dominating cost of implementing the appropriate optimization oracle for the $\mathbf{x}$ variable.

This property is a weaker version of the strong convexity property. Here, instead of assuming the inequality (5) holds for any two points, we only require that it holds for any point $\mathbf{q} \in dom(\mathcal{R}_\mathcal{Q})$ and the corresponding closest optimal solution of Problem (OP).

Example 1. If the function $f(\mathbf{x})$ is strongly convex, then the function $\mathbf{q} \to S(\mathbf{q}, \mathbf{w})$, for a fixed $\mathbf{w} \in \mathbb{E}_2$, is strongly convex with a certain parameter α_S. This implies that in this case the SAL $S(\mathbf{q}, \mathbf{w})$ satisfies the primal quadratic gap property, with the same parameter α_S.

Theorem 1. *Suppose that $f : \mathbb{E}_1 \to \mathbb{R}$ is α-strongly convex. Then, Problem* (OP) *admits a unique primal optimal solution $\boldsymbol{q}^*$ and it satisfies the PQG property with the parameter $\alpha_S = \min\{\frac{\alpha}{2}, \frac{\alpha\rho}{\alpha+2\rho\|\mathcal{A}\|^2}\} > 0$. In particular, the function $\boldsymbol{q} \to S(\boldsymbol{q}, \boldsymbol{w})$, for any fixed $\boldsymbol{w} \in \mathbb{E}_2$, is strongly convex.*

The proof is deferred to the Appendix (see Section B).

Example 2. If $\mathcal{R}_\mathcal{Q}$ is an indicator function for a polytope, we can show that the PQG property holds true even when $f(\cdot)$ need not be strongly convex. Alternatively, we will make the following assumption.

Assumption 3. *(i) $f \equiv g \circ \mathcal{B}$, where $\mathcal{B} : \mathbb{E}_1 \to \mathbb{E}_3$ is a linear mapping, and $g : \mathbb{E}_3 \to \mathbb{R}$ is α_g-strongly convex.*

(ii) $\mathcal{R}_\mathcal{Q}(\boldsymbol{q}) := \mathcal{R}_\mathcal{X}(\boldsymbol{x}) + \mathcal{R}_\mathcal{Y}(\boldsymbol{y})$ is an indicator of a compact and convex polytope $\mathcal{F} \equiv \{\boldsymbol{q} \in \mathbb{E} : \mathcal{C}\boldsymbol{q} \leq \boldsymbol{b}\}$, where $\mathcal{C} : \mathbb{E} \to \mathbb{R}^p$ is a linear mapping, and $\boldsymbol{b} \in \mathbb{R}^p$.

Theorem 2. *Suppose that Assumption 3 holds true. Then, there exists a constant $\sigma > 0$ such that if $\rho \geq \alpha_g$, then Problem* (OP) *satisfies the PQG property with the parameter $\alpha_S = \alpha_g\sigma^{-1}$.*

The proof is deferred to the Appendix (see Section B).

2.3 Weak Proximal Oracle

In this section, we present the main ingredient of our algorithm, which is used to update the primal variable $\mathbf{q}$ of Problem (3) — the weak proximal oracle, a concept which we adapt from Garber and Kaplan (2019) to our augmented Lagrangian framework.

To this end, we will need to define the following function. Given two points $\mathbf{q} \in \mathbb{E}$ and $\mathbf{w} \in \mathbb{E}_2$ together with two scalars $\eta \in (0, 1]$ and $\mu > 0$, we define the function

$$\Phi_\lambda(\mathbf{v}) := \mathcal{R}_\mathcal{Q}(\mathbf{v}) + \langle \mathbf{v}, \nabla_\mathbf{q} S(\mathbf{q}, \mathbf{w}) + 2\mu\mathcal{K}^\top\mathcal{K}\mathbf{q}\rangle + \frac{\lambda}{2}\left(\eta(\beta_S + 2\mu\|\mathcal{K}\|^2)\right)\|\mathbf{v} - \mathbf{q}\|^2,$$

where β_S is the smoothness parameter of $S(\cdot, \mathbf{w})$ (independent of $\mathbf{w}$), and $\lambda \geq 1$ is a parameter.

Before we formally define the notion of weak proximal oracle, we would like to define the notion of Strong Proximal Oracle. We say that a procedure is a *(Strong) Proximal Oracle* applied to the augmented Lagrangian $\mathcal{L}_\rho(\mathbf{q}, \mathbf{w}) \equiv S(\mathbf{q}, \mathbf{w}) + \mathcal{R}_\mathcal{Q}(\mathbf{q})$, which is associated with Problem (OP), if it computes the exact minimizer of Φ_λ for $\lambda = 1$. That is, solves the problem $\min_{\mathbf{v}\in\mathbb{E}} \Phi_1(\mathbf{v})$.

Definition 2. (Weak Proximal Oracle) We say that a procedure, which is denoted by $\text{WPO}_\lambda(\mathbf{q}, \mathbf{w}, \eta, \mu)$, is a *$\lambda$-Weak Proximal Oracle* applied to $\mathcal{L}_\rho(\mathbf{q}, \mathbf{w})$, if it returns a point

$\mathbf{v} \in \mathbb{E}$ which satisfies that

$$\forall \mathbf{q}^* \in \mathcal{P}^* : \quad \Phi_1(\mathbf{v}) \leq \Phi_\lambda(\mathbf{q}^*), \tag{6}$$

where we recall that $\mathcal{P}^*$ is the set of optimal solutions of Problem (OP).

Recalling that $\mathbf{q}$ is simply a convenient notation for the concatenation of the original two vector variables $\mathbf{x}$ and $\mathbf{y}$, the implementation of an oracle whose output satisfies (6) is naturally achieved by decoupling the condition (6) into two parts, one w.r.t. the variable $\mathbf{x}$ and the other w.r.t. the variable $\mathbf{y}$. That is, we consider two separate computations of two points, $\mathbf{v}_\mathbf{x} \in \mathbb{E}_1$ and $\mathbf{v}_\mathbf{y} \in \mathbb{E}_2$, satisfying the following inequalities with some $\lambda_\mathbf{x}, \lambda_\mathbf{y} \geq 1$:

$$\forall \mathbf{x}^* \in \mathcal{X}^* : \quad \Phi_1^\mathbf{x}(\mathbf{v}_\mathbf{x}) \leq \Phi_{\lambda_\mathbf{x}}^\mathbf{x}(\mathbf{x}^*), \tag{7}$$

where $\mathcal{X}^* := \{\mathbf{x}^* \in \mathbb{E}_1 : (\mathbf{x}^*, \mathcal{A}\mathbf{x}^*) \in \mathcal{P}^*\}$, and

$$\begin{aligned}\Phi_{\lambda_\mathbf{x}}^\mathbf{x} := \mathcal{R}_\mathcal{X}(\mathbf{v}_\mathbf{x}) + \langle \mathbf{v}_\mathbf{x}, \nabla_\mathbf{x} S(\mathbf{q}, \mathbf{w}) + 2\mu \mathcal{A}^\top \mathcal{K}\mathbf{q}\rangle \\ + \lambda_\mathbf{x} \frac{\eta(\beta_S + 2\mu\|\mathcal{K}\|^2)}{2} \|\mathbf{v}_\mathbf{x} - \mathbf{x}\|^2.\end{aligned} \tag{8}$$

Similarly,

$$\forall \mathbf{y}^* \in \mathcal{Y}^* : \quad \Phi_1^\mathbf{y}(\mathbf{v}_\mathbf{y}) \leq \Phi_{\lambda_\mathbf{y}}^\mathbf{y}(\mathbf{y}^*), \tag{9}$$

where $\mathcal{Y}^* := \{\mathcal{A}\mathbf{x}^* \in \mathbb{E}_2 : \mathbf{x}^* \in \mathcal{X}^*\}$, and

$$\begin{aligned}\Phi_{\lambda_\mathbf{y}}^\mathbf{y}(\mathbf{v}_\mathbf{y}) := \mathcal{R}_\mathcal{Y}(\mathbf{v}_\mathbf{y}) + \langle \mathbf{v}_\mathbf{y}, \nabla_\mathbf{y} S(\mathbf{q}, \mathbf{w}) - 2\mu \mathcal{K}\mathbf{q}\rangle \\ + \lambda_\mathbf{y} \frac{\eta(\beta_S + 2\mu\|\mathcal{K}\|^2)}{2} \|\mathbf{v}_\mathbf{y} - \mathbf{y}\|^2.\end{aligned} \tag{10}$$

The following proposition is a simple observation.

Proposition 1. *Assume that $\boldsymbol{v_x} \in \mathbb{E}_1$ satisfies* (7) *with some parameter $\lambda_\boldsymbol{x} \geq 1$, and $\boldsymbol{v_y} \in \mathbb{E}_2$ satisfies* (9) *with some parameter $\lambda_\boldsymbol{y} \geq 1$. Then, $\boldsymbol{v} = (\boldsymbol{v_x}, \boldsymbol{v_y})$ satisfies* (6) *with $\lambda = \max\{\lambda_\boldsymbol{x}, \lambda_\boldsymbol{y}\} \geq 1$.*

The main difficulty in satisfying (7) and (9) is mainly because of the nonsmooth functions $\mathcal{R}_\mathcal{X}$ and $\mathcal{R}_\mathcal{Y}$, respectively. This motivates the following definition.

Definition 3. (Weak Proximal Friendly) We say that a convex function $\mathcal{R}_\mathcal{X}$ ($\mathcal{R}_\mathcal{Y}$) is *weak proximal friendly* for Problem (OP), if a point $\mathbf{v}_\mathbf{x} \in \mathbb{E}_1$ ($\mathbf{v}_\mathbf{y} \in \mathbb{E}_2$) satisfying (7) ((9)) (for some finite $\lambda_\mathbf{x}$ ($\lambda_\mathbf{y}$)) can be computed efficiently. We say that Problem (OP) is *weak proximal friendly* if $\mathcal{R}_\mathcal{X}$ and $\mathcal{R}_\mathcal{Y}$ are both weak proximal friendly.

Let us now discuss some of the most important and interesting examples of weak proximal friendly functions.

2.3.1 Proximal friendly functions

When $\mathcal{R}_\mathcal{X}$ ($\mathcal{R}_\mathcal{Y}$) is a (strong) proximal friendly function (i.e., an exact minimizer of $\Phi_1^\mathbf{x}$ ($\Phi_1^\mathbf{y}$), as defined in (8) ((10)) could be computed efficiently), it follows immediately that it is also a weak proximal friendly function with parameter $\lambda_\mathbf{x} = 1$ ($\lambda_\mathbf{y} = 1$).

2.3.2 Matrix nuclear norm regularization/constraint

In a typical low-rank matrix recovery setup, in which the nuclear norm is used as a convex surrogate for low-rank (see, for instance, the seminal works Candes and Recht (2012); Candès et al. (2011)), we have that $\mathbb{E}_1 = \mathbb{R}^{m\times n}$ and $\mathcal{R}_\mathcal{X}$ is an indicator function of a nuclear norm ball or a nuclear norm regularizer, or an indicator function of the spectrahedron (the set of all positive semidefinite matrices with trace equals some fixed positive parameter), in case the solution is also required to be positive semidefinite. Assuming there exists a unique optimal low-rank solution, i.e., $\mathcal{X}^* = \{\mathbf{X}^*\}$ with $\text{rank}(\mathbf{X}^*) = k << \min\{m, n\}$, then an oracle for (7) amounts to computing a single rank-k SVD of an $m \times n$ matrix, plus additional computationally-cheaper operations, which in high dimension is far more efficient than a proximal/projection computation, which in general requires a full-rank SVD, see detailed discussions in Allen-Zhu et al. (2017); Garber et al. (2021); Garber and Kaplan (2019). Concretely, to satisfy (7) in this case we solve:

$$\arg\min_{\mathbf{V}\in\mathbb{R}^{m\times n}:\ \text{rank}(\mathbf{V})\leq k} \Phi_1^\mathbf{x}(\mathbf{V}), \tag{11}$$

and the corresponding WPO parameter is $\lambda_\mathbf{x} = 1$.

Note that Problem (11) follows the same structure of the standard proximal computation w.r.t. the function $\mathcal{R}_\mathcal{X}$, only that it is further constrained over the set of bounded rank matrices (which makes the problem more efficient to solve). Formally, we have the following theorem (extracted from the relevant discussions in Allen-Zhu et al. (2017); Garber et al. (2021); Garber and Kaplan (2019)).

Theorem 3. *Let $(\boldsymbol{Q}, \boldsymbol{W}) \in \mathbb{E} \times \mathbb{E}_2$ be a pair of primal and dual points, where $\boldsymbol{Q} = (\boldsymbol{X}, \boldsymbol{Y}) \in \mathbb{R}^{m\times n} \times \mathbb{E}_2$. Let $\hat{\beta} := \beta_S + 2\mu\|\mathcal{K}\|^2$. Denote $\boldsymbol{M} := \boldsymbol{X} - \frac{1}{\eta\hat{\beta}}\left(\nabla_\boldsymbol{X} S(\boldsymbol{Q}, \boldsymbol{W}) + 2\mu\mathcal{A}^\top\mathcal{K}\boldsymbol{Q}\right)$. Let $\boldsymbol{U}, \boldsymbol{\Sigma}, \boldsymbol{V}$ be the SVD matrices of $\boldsymbol{M}$, i.e., $\boldsymbol{M} = \boldsymbol{U}\boldsymbol{\Sigma}\boldsymbol{V}^\top$.*

- *If $\mathcal{R}_\mathcal{X}(\cdot) = \nu\|\cdot\|_{nuc}$, for some $\nu > 0$, then a solution to* (11) *is the matrix $\tilde{\boldsymbol{M}} = \boldsymbol{U}\tilde{\boldsymbol{\Sigma}}\boldsymbol{V}^\top$, for*

$$\begin{aligned}\tilde{\boldsymbol{\Sigma}} = \mathit{diag}(\max\{\sigma_1(\boldsymbol{M}) - \zeta, 0\}, \ldots, \\ \max\{\sigma_k(\boldsymbol{M}) - \zeta, 0\}, 0, \ldots, 0),\end{aligned}$$

  where $\zeta := \frac{\nu}{\eta\hat{\beta}}$.

- *If $\mathcal{R}_\mathcal{X}(\cdot) = \delta_{NB(\tau)}(\cdot)$ is the indicator function for the nuclear norm ball of radius τ, then a solution to* (11) *is given by $\tilde{\boldsymbol{M}} = \boldsymbol{U}\tilde{\boldsymbol{\Sigma}}\boldsymbol{V}^\top$, where here $\tilde{\boldsymbol{\Sigma}}$ is the diagonal matrix whose diagonal is the projection of the vector $(\sigma_1(\boldsymbol{M}), \ldots, \sigma_k(\boldsymbol{M}), 0, \ldots, 0)$ onto the ℓ_1 norm ball of radius τ.*

- *If $\mathcal{R}_\mathcal{X}(\cdot) = \delta_{\mathcal{S}_\tau}(\cdot)$ is the indicator function for the spectahedron $\{\boldsymbol{X} \mid \boldsymbol{X} \succeq 0,\ tr(\boldsymbol{X}) = \tau\}$, for some given $\tau > 0$, then letting $\boldsymbol{U}\Lambda\boldsymbol{U}^\top$ be the eigen-decomposition*

of $\boldsymbol{M}$, a solution to (11) *is the matrix $\tilde{\boldsymbol{M}} = \boldsymbol{U}\tilde{\Lambda}\boldsymbol{U}^\top$, where $\tilde{\Lambda}$ is the diagonal matrix whose diagonal is the projection of the vector $(\lambda_1(\boldsymbol{M}), \ldots, \lambda_k(\boldsymbol{M}), 0, \ldots, 0)$ onto the simplex of radius τ.*

In all these three cases, the runtime to compute $\tilde{\boldsymbol{M}}$ is dominated by the computation of the top k components in the SVD of $\boldsymbol{M}$.

2.3.3 Polytope constraint

In case $\mathcal{R}_\mathcal{X}$ ($\mathcal{R}_\mathcal{Y}$) is an indicator function for a convex and compact polytope for which a linear minimization oracle can be implemented efficiently, then $\mathcal{R}_\mathcal{X}$ ($\mathcal{R}_\mathcal{Y}$) is also weak proximal friendly. Concretely, it is possible to construct an oracle for (7) ((9)) using a single call to the linear minimization oracle, plus some additional computations (that do not require any oracle access).

Theorem 4. *Let $\mathcal{R}_\mathcal{X}$ ($\mathcal{R}_\mathcal{Y}$) be an indicator function for a convex and compact polytope $\mathcal{F}$. Suppose a point $\boldsymbol{x} \in \mathbb{E}_1 = \mathbb{R}^d$ ($\boldsymbol{y} \in \mathbb{E}_2 := \mathbb{R}^d$) is given explicitly as a convex combination of t vertices of the polytope $\{z_1, \ldots, z_t\}$. Let us denote the new output of the LMO of $\mathcal{F}$ w.r.t. the linear objective function determined by the vector $\boldsymbol{p_x} := \nabla_{\boldsymbol{x}} S(\boldsymbol{q}, \boldsymbol{w}) + 2\mu\mathcal{A}^\top\mathcal{K}\boldsymbol{q}$ ($\boldsymbol{p_y} := \nabla_{\boldsymbol{y}} S(\boldsymbol{q}, \boldsymbol{w}) - 2\mu\mathcal{K}\boldsymbol{q}$) by z_{t+1} and let $\boldsymbol{M} = [z_1, \ldots, z_{t+1}] \in \mathbb{R}^{d\times(t+1)}$. Let $\hat{\beta} := \beta_S + 2\mu\|\mathcal{K}\|^2$. Then, we can compute a point satisfying* (7) *((9)) with a parameter $\lambda(\mathcal{F}) \geq 1$ by returning the point $\boldsymbol{v_x} = \boldsymbol{M}\boldsymbol{\gamma}_{\boldsymbol{x}}^*$ ($\boldsymbol{v_y} = \boldsymbol{M}\boldsymbol{\gamma}_{\boldsymbol{y}}^*$), where $\boldsymbol{\gamma}_{\boldsymbol{x}}^*$ ($\boldsymbol{\gamma}_{\boldsymbol{y}}^*$) is an optimal solution to the following convex quadratic problem over the simplex of size $t+1$:*

$$\min_{\boldsymbol{\gamma}\geq 0, \langle \boldsymbol{1}, \boldsymbol{\gamma}\rangle = 1} \langle \boldsymbol{M}\boldsymbol{\gamma}, \boldsymbol{p_x}\rangle + \frac{\eta\hat{\beta}}{2}\|\boldsymbol{M}\boldsymbol{\gamma} - \boldsymbol{x}\|^2 \tag{12}$$

$$\left(\min_{\boldsymbol{\gamma}\geq 0, \langle \boldsymbol{1}, \boldsymbol{\gamma}\rangle = 1} \langle \boldsymbol{M}\boldsymbol{\gamma}, \boldsymbol{p_y}\rangle + \frac{\eta\hat{\beta}}{2}\|\boldsymbol{M}\boldsymbol{\gamma} - \boldsymbol{y}\|^2\right), \tag{13}$$

where $\boldsymbol{1} \in \mathbb{R}^{t+1}$ is the vector of ones.

Note that the returned solution $\boldsymbol{v_x} = \boldsymbol{M}\boldsymbol{\gamma}_{\boldsymbol{x}}^*$ ($\boldsymbol{v_y} = \boldsymbol{M}\boldsymbol{\gamma}_{\boldsymbol{y}}^*$) is now given explicitly in the form of a convex combination of at most $t+1$ vertices of the polytope.

The proof is based on several observations from Garber and Hazan (2016) and is given in the Appendix (see Section C).

The fact that the implementation of the WPO described in Theorem 4 only increases the number of vertices in the support of the computed point $\mathbf{v_x}$ ($\mathbf{v_y}$) by at most one is important, since our algorithm for solving the saddle point problem (3) using this WPO makes a single call to this oracle per iteration. Hence, when the overall number of iterations is not very large, we will have that on each iteration, the input point $\mathbf{x}$ ($\mathbf{y}$) to this oracle will be supported on only a few vertices of the polytope, which means that Problem (12) ((13)) could be solved very efficiently.

The WPO parameter $\lambda(\mathcal{F})$ depends on the geometry of the polytope $\mathcal{F}$ and may depend in worst case on the dimension, more details can be found in the Appendix (see Section C).

2.4 Some Illustrative Examples

In Appendix D we discuss in detail several notable applications of interest which satisfy both of our main assumptions: they satisfy the primal quadratic gap property and they admit efficient implementations of the WPO (though they do not necessarily admit an efficient standard strong proximal oracle). These include two main examples: (i) recovering low-rank matrices and tensors via Frobenius norm minimization subject to nuclear norm constraints/regularizations plus an optional additional simple constraint/regularization (e.g., ℓ_1 regularization to promote sparsity in addition to low-rank), assuming that the optimal solution (matrix or tensor) is indeed low-rank, and (ii) minimizing a least squares objective (not necessarily a strongly convex one) over the intersection of polytopes, each admitting an efficient LMO.

3 Algorithm and Convergence Analysis

Our weak proximal oracle-based algorithm for solving the saddle point problem (3) is given as Algorithm 1. At each iteration, the primal variable is updated using a Frank-Wolfe-style update, in the sense that the updated primal variable is given as a convex combination of the previous primal iterate and the output of some oracle. As opposed to the classical Frank-Wolfe method, which relies on the output of a linear minimization oracle, here we rely on the output of a weak proximal oracle described in the preceding sections. The update of the dual variable is done via a standard gradient ascent step.

Algorithm 1 Weak Proximal Method of Multipliers

1. **Input:** Primal and dual step sizes $\eta \in [0, 1], \mu > 0$, a WPO oracle with parameter $\lambda \geq 1$, and initialization $(\mathbf{q}_0, \mathbf{w}_0) \in \mathrm{dom}(\mathcal{R}_\mathcal{Q}) \times \mathbb{E}_2$.
2. **Main step:** For $t = 0, 1, \ldots$ generate the sequence $\{(\mathbf{q}_t, \mathbf{w}_t)\}_{t\in\mathbb{N}}$ as follows

$$\mathbf{v}_t = \mathrm{WPO}_\lambda(\mathbf{q}_t, \mathbf{w}_t, \eta, \mu), \tag{14}$$
$$\mathbf{q}_{t+1} = (1-\eta)\mathbf{q}_t + \eta\mathbf{v}_t, \tag{15}$$
$$\mathbf{w}_{t+1} = \mathbf{w}_t + \mu\mathcal{K}\mathbf{q}_{t+1}. \tag{16}$$

Before stating our main result, the convergence guarantees of Algorithm 1, let us introduce some helpful notation.

- For any $\mathbf{q} = (\mathbf{x}, \mathbf{y}) \in \mathbb{E}$, we denote the value of the

objective function of Problem (OP) by $h(\mathbf{q}) := f(\mathbf{x}) + \mathcal{R}_{\mathcal{X}}(\mathbf{x}) + \mathcal{R}_{\mathcal{Y}}(\mathbf{y})$.

- Given a sequence $\{\mathbf{q}_t\}_{t\in\mathbb{N}}$, we define the corresponding ergodic sequence $\{\bar{\mathbf{q}}_t\}_{t\in\mathbb{N}}$ by $\bar{\mathbf{q}}_t := \frac{1}{T}\sum_{j=0}^{t-1}\mathbf{q}_{j+1}$ for all $t \in \mathbb{N}$.
- For any $t \in \mathbb{N}$, we denote by $d_t := \mathcal{L}_\rho(\mathbf{q}_t, \mathbf{w}_t) - \mathcal{L}_\rho^*$, where $\mathcal{L}_\rho^*$ denotes the AL value of a saddle point.

To formally present our main result, we also recall that α_S and β_S denote the primal quadratic gap parameter and smoothness parameter of $S(\mathbf{q}, \mathbf{w})$, respectively.

Theorem 5. *(An O(1/T) ergodic convergence rate) Let $(\boldsymbol{q}^*, \boldsymbol{w}^*)$ be a saddle point of $\mathcal{L}_\rho$ and let $\{(\boldsymbol{q}_t, \boldsymbol{w}_t)\}_{t\in\mathbb{N}}$ be a sequence generated by Algorithm 1 with a primal step size $\eta = \frac{\alpha_S}{2\lambda(\beta_S + 2\mu(\|\mathcal{A}\|+1)^2)} \in (0, 1]$, and μ satisfying $0 < \mu \leq \frac{\sqrt{\lambda\alpha_S^2 + \lambda^2\beta_S^2} - \lambda\beta_S}{4\lambda(\|\mathcal{A}\|+1)^2}$. Then, for any $c \geq 2\|\boldsymbol{w}^*\|$ and any integer $T \geq 1$ we have*

$$h(\bar{\boldsymbol{q}}_T) - h(\boldsymbol{q}^*) \leq \frac{B(\rho,\mu)}{T} \quad \text{and} \quad \|\mathcal{K}\bar{\boldsymbol{q}}_T\| \leq \frac{2B(\rho,\mu)}{cT},$$

where

$$B(\rho,\mu) = \frac{(c + \|\boldsymbol{w}_0\|)^2}{2\mu} + \max\left\{0, \frac{2d_1(\beta + (\rho+2\mu)(\|\mathcal{A}\|+1)^2)}{\alpha_S}\right\},$$

$d_1 = \mathcal{L}_\rho(\boldsymbol{q}_1, \boldsymbol{w}_1) - \mathcal{L}_\rho(\boldsymbol{q}^*, \boldsymbol{w}^*)$ and β is the smoothness parameter of f.

The following lemma accounts for the main step in the proof of our fast convergence rates given in Theorem 5 and also accounts for the main technical novelty of our paper. It leverages both the primal quadratic gap property and the weak proximal oracle, to obtain a linear convergence rate for the value of the augmented Lagrangian.

Lemma 3.1. *Let $\{(\boldsymbol{q}_t, \boldsymbol{w}_t)\}_{t\in\mathbb{N}}$ be a sequence generated by Algorithm 1 with η and μ as given in Theorem 5. Then, for all $t \in \mathbb{N}$ we have that $d_{t+1} \leq (1-\eta)^t d_1$.*

4 Numerical Experiments

We compare the empirical performance of our algorithm with the state-of-the-art projection-free conditional gradient-based algorithm from Yurtsever et al. (2019). For each of the algorithms we tested two variants, see Table 4.1 for details. Due to lack of space, some of the implementation details and most of the results are deferred to the Appendix (see Section F).

We consider two tasks which involve optimization with low-rank matrices and are cast as optimization over the spectrahedron: estimation of a low-rank and sparse covariance matrix from noisy observations, and the semidefinite relaxation for Max Cut. For both tasks, implementing the WPO in our algorithms involves a low-rank SVD with rank that is at least that of the optimal solution (see, for instance, Section D.1 of the Appendix). Thus, throughout this section we denote the rank of the optimal solution by r^* and our algorithm's estimate of it by $\hat{r}^*$.

4.1 Low-rank and sparse covariance estimation

We consider the following convex relaxation for recovering a low-rank and sparse positive semidefinite matrix from a noisy matrix observation $\widehat{\mathbf{\Sigma}}$:

$$\min_{\mathbf{S}} \frac{1}{2}\|\mathbf{S} - \widehat{\mathbf{\Sigma}}\|_F^2 \text{ s.t. } \mathbf{S} \succeq 0, \text{tr}(\mathbf{S}) = \tau, \|\mathbf{S}\|_1 \leq s. \quad \text{(CME)}$$

As discussed in Section D.1, this problem satisfies all the assumptions required for our theoretical guarantees to hold.

Data generation. Our experiment is inspired by previous experiments conducted in Richard et al. (2012) and Gidel et al. (2018). We first generate a block diagonal, sparse and low-rank covariance matrix $\mathbf{\Sigma} \in \mathbb{R}^{d\times d}$, where we set $d = 400$. Then, we draw d vectors $\mathbf{z}_i \sim \mathcal{N}(0, \mathbf{\Sigma})$, add a Gaussian noise $\mathcal{N}(0, \sigma = 0.6)$ to each entry of $\mathbf{z}_i$, and create the noisy matrix $\widehat{\mathbf{\Sigma}} := \frac{1}{d}\sum_{i=1}^d \mathbf{z}_i\mathbf{z}_i^\top$. To create the blocks of $\mathbf{\Sigma}$, we use r blocks of the form $\mathbf{u}\mathbf{u}^\top$ where $\mathbf{u} \sim \mathcal{U}([-1, 1])$. This way we ensure that $\mathbf{\Sigma}$ is of rank at most r. In order to enforce sparsity, while ensuring the low rank of $\mathbf{\Sigma}$, *before* computing $\mathbf{u}\mathbf{u}^\top$, we only keep the entries u_i for which $|u_i| > 0.9$ (the rest of the entries become zero). Our choices of τ and s, the radius of the nuclear norm ball and the ℓ_1 norm ball, respectively, are chosen as the nuclear norm and ℓ_1 norm of $\mathbf{\Sigma}$. We use rank values $r \in \{5, 10, 20\}$.

Results. We ran both algorithms (two variants for each) for a fixed number of iterations $T = 2000$. The results are the averages over 20 i.i.d. runs of the experiment (each experiment randomly selects $\mathbf{\Sigma}$ and $\widehat{\mathbf{\Sigma}}$). In the graphs, we plot the (normalized) objective $\frac{\|\mathbf{S}-\widehat{\mathbf{\Sigma}}\|_F^2}{2\|\widehat{\mathbf{\Sigma}}\|_F^2}$, the distance from feasibility $\|\mathbf{S} - P_{\ell_1(s)}(\mathbf{S})\|_F$, where $P_{\ell_1(s)}(\mathbf{S})$ is the projection of $\mathbf{S}$ onto the ℓ_1-norm ball of radius s, and the recovery error of $\mathbf{\Sigma}$ measured by $\frac{\|\mathbf{S}-\mathbf{\Sigma}\|_F^2}{2\|\mathbf{\Sigma}\|_F^2}$. All measures are plotted both w.r.t. the number of iterations and the runtime (in seconds). Initially we set the SVD rank parameter for our algorithm to be exactly r, i.e., we set $\hat{r}^* = r$. To simulate a more challenging and realistic setting, we redid the experiments for $r = 10$ and $r = 20$, but this time we run our algorithm with a 1.5x overestimate of the rank, i.e., we use for it SVD computations of rank $\hat{r}^* = 15$ and $\hat{r}^* = 30$, respectively. All other parameters remain unchanged (note this does not affect the baseline which regardless of the rank

Algorithm	Description	SVD rank
Mean (ours)	Returns the mean (ergodic) of the primal sequence produced by Algorithm 1	$\hat{r}^*$
Last (ours)	Returns the last primal point produced by Algorithm 1	$\hat{r}^*$
CGAL-const	Baseline algorithm from Yurtsever et al. (2019) using "const" option for dual updates	1
CGAL-decr	Baseline algorithm from Yurtsever et al. (2019) using "decr" option for dual updates	1

Table 2: Description of algorithms compared in the empirical study. The last column specifies the rank of SVD needed to update primal variable at each iteration.

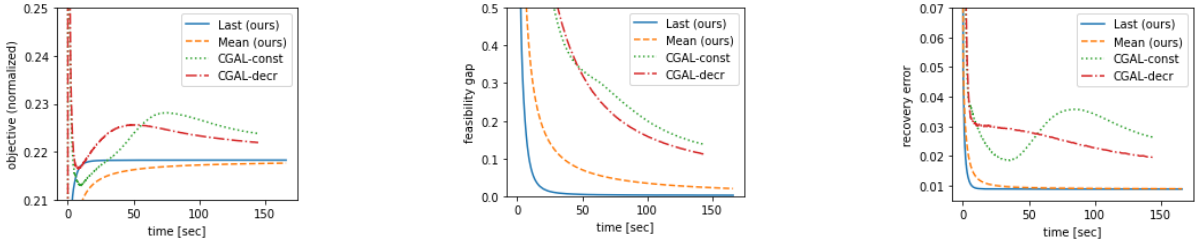

Figure 1: Results for low-rank and sparse covariance estimation with $r = 20$ and $\hat{r}^* = 30$.

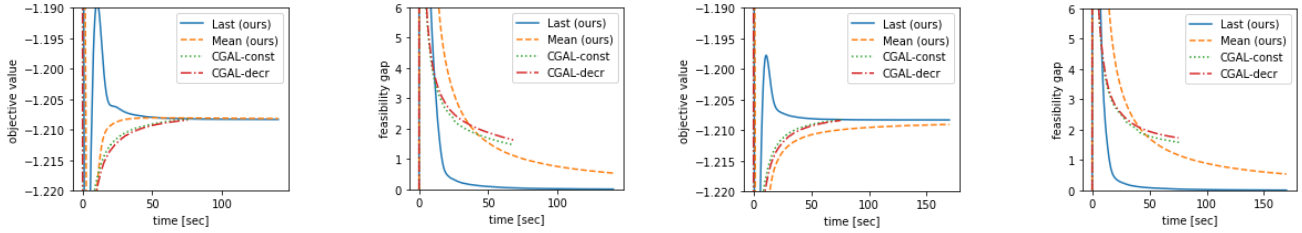

Figure 2: Results for Max Cut with graph G1 when $\hat{r}^* = r^* = 13$ (two leftmost panels) and when $\hat{r}^* = 20$ (two rightmost panels).

performs only rank-one SVD computations). In Figure 1, we show the convergence w.r.t. all measures as a function of runtime when $r = 20$ and we overestimate the rank for our algorithm to $\hat{r}^* = 30$. We can see that w.r.t. all measures, both of our variants clearly outperform the baseline. The rest of the results, which exhibit a very similar trend, are provided in the Appendix (see Section F).

4.2 Max Cut

We now consider the following well-known semidefinite relaxation for the Max Cut problem:

$$\min_{\mathbf{S}} -\text{tr}(\mathbf{CS}) \;\text{ s.t. }\; \mathbf{S} \succeq 0, \text{tr}(\mathbf{S}) = d, \text{diag}(\mathbf{S}) = \mathbf{1}, \quad \text{(MC)}$$

where $\mathbf{C}$ is the Laplacian matrix of a combinatorial graph, $\text{diag}(\mathbf{S})$ is the vector of elements on the main diagonal of the matrix $\mathbf{S} \in \mathbb{S}^d_+$, and $\mathbf{1}$ is a vector of ones of length d.

Note that since the objective function is linear, this problem does not necessarily satisfy our primal quadratic gap condition (Definition 1). Still this does not revoke the applicability of our algorithm for this problem.

Datasets. We used the G1, G2 and G3 graphs from the Gset dataset[4]. These are graphs of size 800×800. The ranks of the optimal solutions for the max cut problem in these graphs are conveniently given in Ding and Udell (2021), Table 1 (the ranks of the optimal solutions are $r^* = 13$ for G1 and G2, and $r^* = 14$ for G3).

Results. We ran all algorithms for 2000 iterations. For each of the datasets we repeated the experiment 10 times and averaged the runtimes for more reliable measurements. Here also, we initially set the SVD rank parameter for our algorithm to be exactly r, i.e., we set $\hat{r}^* = r$, but we also redid the experiment for G1 when overestimating the rank of the optimal solution for our algorithm, taking $\hat{r}^* = 20$ (while $r^* = 13$). We plot the objective value and the feasibility gap measured by $\|\text{diag}(\mathbf{S}) - \mathbf{1}\|$ both w.r.t. number of iterations and runtime. Figure 2 shows the convergence w.r.t. runtime for the graph G1 in case r^* is used exactly or overestimated when setting the SVD rank parameter of our algorithms, $\hat{r}^*$. We can see that when $\hat{r}^* = r$, both of our variants are faster than the baselines. When $\hat{r}^* > r$, while our "Mean" variant is slightly surpassed by the baseline w.r.t. the objective, it converges faster to a feasible solution. Our "Last" variant performs better than the baseline w.r.t. both measurements. The rest of the results are reported in the Appendix (see Section F).

[4] Y. Ye. Gset random graphs, found in: https://www.cise.ufl.edu/research/sparse/matrices/Gset/index.html

Acknowledgements

We thank Atara Kaplan for important discussions regarding problems with the convergence proofs in Gidel et al. (2018) (discussed in Section A).

Dan Garber is supported by the ISRAEL SCIENCE FOUNDATION (grant No. 2267/22).

References

Zeyuan Allen-Zhu, Elad Hazan, Wei Hu, and Yuanzhi Li. Linear convergence of a Frank-Wolfe type algorithm over trace-norm balls. *Advances in Neural Information Processing Systems*, 30, 2017.

Donald WK Andrews. Heteroskedasticity and autocorrelation consistent covariance matrix estimation. *Econometrica: Journal of the Econometric Society*, pages 817–858, 1991.

Amir Beck and Shimrit Shtern. Linearly convergent away-step conditional gradient for non-strongly convex functions. *Mathematical Programming*, 164(1):1–27, 2017.

Antoni Buades, Bartomeu Coll, and J-M Morel. A non-local algorithm for image denoising. In *2005 IEEE computer society conference on computer vision and pattern recognition (CVPR'05)*, volume 2, pages 60–65. Ieee, 2005a.

Antoni Buades, Bartomeu Coll, and Jean-Michel Morel. A review of image denoising algorithms, with a new one. *Multiscale modeling & simulation*, 4(2):490–530, 2005b.

Emmanuel J Candès and Benjamin Recht. Exact matrix completion via convex optimization. *Foundations of Computational mathematics*, 9(6):717–772, 2009.

Emmanuel J Candes and Benjamin Recht. Exact matrix completion via convex optimization. *Communications of the ACM*, 55(6):111–119, 2012.

Emmanuel J Candès, Xiaodong Li, Yi Ma, and John Wright. Robust principal component analysis? *Journal of the ACM (JACM)*, 58(3):1–37, 2011.

Antonin Chambolle and Thomas Pock. On the ergodic convergence rates of a first-order primal–dual algorithm. *Mathematical Programming*, 159(1):253–287, 2016.

Ramu Chenna, Hideaki Sugawara, Tadashi Koike, Rodrigo Lopez, Toby J Gibson, Desmond G Higgins, and Julie D Thompson. Multiple sequence alignment with the clustal series of programs. *Nucleic acids research*, 31(13): 3497–3500, 2003.

Florence Corpet. Multiple sequence alignment with hierarchical clustering. *Nucleic acids research*, 16(22):10881–10890, 1988.

Alberto De Marchi. Constrained structured optimization and augmented Lagrangian proximal methods. *arXiv preprint arXiv:2203.05276*, 2022.

Neil K Dhingra, Sei Zhen Khong, and Mihailo R Jovanović. The proximal augmented Lagrangian method for nonsmooth composite optimization. *IEEE Transactions on Automatic Control*, 64(7):2861–2868, 2018.

Lijun Ding and Madeleine Udell. On the simplicity and conditioning of low rank semidefinite programs. *SIAM Journal on Optimization*, 31(4):2614–2637, 2021.

John C Driscoll and Aart C Kraay. Consistent covariance matrix estimation with spatially dependent panel data. *Review of economics and statistics*, 80(4):549–560, 1998.

Silvia Gandy, Benjamin Recht, and Isao Yamada. Tensor completion and low-n-rank tensor recovery via convex optimization. *Inverse problems*, 27(2):025010, 2011.

Dan Garber. Faster projection-free convex optimization over the spectrahedron. *Advances in Neural Information Processing Systems*, 29, 2016.

Dan Garber. Logarithmic regret for online gradient descent beyond strong convexity. In *The 22nd International Conference on Artificial Intelligence and Statistics*, pages 295–303. PMLR, 2019.

Dan Garber and Elad Hazan. A linearly convergent variant of the conditional gradient algorithm under strong convexity, with applications to online and stochastic optimization. *SIAM Journal on Optimization*, 26(3):1493–1528, 2016.

Dan Garber and Atara Kaplan. Fast stochastic algorithms for low-rank and nonsmooth matrix problems. In *The 22nd International Conference on Artificial Intelligence and Statistics*, pages 286–294. PMLR, 2019.

Dan Garber, Atara Kaplan, and Shoham Sabach. Improved complexities of conditional gradient-type methods with applications to robust matrix recovery problems. *Mathematical Programming*, 186(1):185–208, 2021.

Gauthier Gidel, Fabian Pedregosa, and Simon Lacoste-Julien. Frank-Wolfe splitting via augmented Lagrangian method. In *International Conference on Artificial Intelligence and Statistics*, pages 1456–1465. PMLR, 2018.

Alexandre Gramfort, Bertrand Thirion, and Gaël Varoquaux. Identifying predictive regions from fmri with tv-l1 prior. In *2013 International Workshop on Pattern Recognition in Neuroimaging*, pages 17–20. IEEE, 2013.

Elad Hazan and Satyen Kale. Projection-free online learning. In *Proceedings of the 29th International Conference on Machine Learning, ICML 2012, Edinburgh, Scotland, UK, June 26 - July 1, 2012*. icml.cc / Omnipress, 2012.

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

Martin Jaggi. Revisiting Frank-Wolfe: Projection-free sparse convex optimization. In *International Conference on Machine Learning*, pages 427–435. PMLR, 2013.

Hastad Johan. Tensor rank is np-complete. *Journal of Algorithms*, 4(11):644–654, 1990.

Kazutaka Katoh and Hiroyuki Toh. Recent developments in the mafft multiple sequence alignment program. *Briefings in bioinformatics*, 9(4):286–298, 2008.

Tamara G Kolda and Brett W Bader. Tensor decompositions and applications. *SIAM review*, 51(3):455–500, 2009.

Simon Lacoste-Julien and Martin Jaggi. On the global linear convergence of Frank-Wolfe optimization variants. *Advances in neural information processing systems*, 28, 2015.

Jack L Lancaster, Marty G Woldorff, Lawrence M Parsons, Mario Liotti, Catarina S Freitas, Lacy Rainey, Peter V Kochunov, Dan Nickerson, Shawn A Mikiten, and Peter T Fox. Automated talairach atlas labels for functional brain mapping. *Human brain mapping*, 10(3):120–131, 2000.

David Liben-Nowell and Jon Kleinberg. The link prediction problem for social networks. In *Proceedings of the twelfth international conference on Information and knowledge management*, pages 556–559, 2003.

Ya-Feng Liu, Xin Liu, and Shiqian Ma. On the nonergodic convergence rate of an inexact augmented Lagrangian framework for composite convex programming. *Mathematics of Operations Research*, 44(2):632–650, 2019.

Linyuan Lü and Tao Zhou. Link prediction in complex networks: A survey. *Physica A: statistical mechanics and its applications*, 390(6):1150–1170, 2011.

Toshiyuki Mizoguchi et al. Kj arrow, l. hurwicz and h. uzawa, studies in linear and non-linear programming. *Economic Review*, 11(3):349–351, 1960.

Seiji Ogawa, David W Tank, Ravi Menon, Jutta M Ellermann, Seong G Kim, Helmut Merkle, and Kamil Ugurbil. Intrinsic signal changes accompanying sensory stimulation: functional brain mapping with magnetic resonance imaging. *Proceedings of the National Academy of Sciences*, 89(13):5951–5955, 1992.

Michael JD Powell. A method for nonlinear constraints in minimization problems. *Optimization*, pages 283–298, 1969.

Emile Richard, Pierre-André Savalle, and Nicolas Vayatis. Estimation of simultaneously sparse and low rank matrices. *arXiv preprint arXiv:1206.6474*, 2012.

R Tyrrell Rockafellar. Augmented Lagrangians and applications of the proximal point algorithm in convex programming. *Mathematics of operations research*, 1(2): 97–116, 1976.

Shoham Sabach and Marc Teboulle. Lagrangian methods for composite optimization. *Handbook of Numerical Analysis*, 20:401–436, 2019.

Antonio Silveti-Falls, Cesare Molinari, and Jalal Fadili. Generalized conditional gradient with augmented lagrangian for composite minimization. *SIAM Journal on Optimization*, 30(4):2687–2725, 2020.

Ian En-Hsu Yen, Xin Lin, Jiong Zhang, Pradeep Ravikumar, and Inderjit Dhillon. A convex atomic-norm approach to multiple sequence alignment and motif discovery. In *International Conference on Machine Learning*, pages 2272–2280. PMLR, 2016.

Alp Yurtsever, Olivier Fercoq, and Volkan Cevher. A conditional-gradient-based augmented Lagrangian framework. In *International Conference on Machine Learning*, pages 7272–7281. PMLR, 2019.

Jiawei Zhang, Jianhui Chen, Shi Zhi, Yi Chang, S Yu Philip, and Jiawei Han. Link prediction across aligned networks with sparse and low rank matrix estimation. In *2017 IEEE 33rd International Conference on Data Engineering (ICDE)*, pages 971–982. IEEE, 2017a.

Kai Zhang, Wangmeng Zuo, Yunjin Chen, Deyu Meng, and Lei Zhang. Beyond a Gaussian denoiser: Residual learning of deep cnn for image denoising. *IEEE transactions on image processing*, 26(7):3142–3155, 2017b.

Yongqin Zhang, Ruiwen Kang, Xianlin Peng, Jun Wang, Jihua Zhu, Jinye Peng, and Hangfan Liu. Image denoising via structure-constrained low-rank approximation. *Neural Computing and Applications*, 32(16):12575–12590, 2020.

A Issue with the Convergence Proofs in Gidel et al. (2018)

In Section C of the paper (see pages 13-14), given a compact set $\mathcal{X}$ (formally defined in Subsection 2.2 on page 3), the authors define the following functions (equation numbers are as in the original paper):

$$\mathcal{L} := f(\mathbf{x}) + \mathbf{1}_{\mathcal{X}}(\mathbf{x}) + \langle \mathbf{y}, \mathbf{M}\mathbf{x}\rangle + \frac{\lambda}{2}\|\mathbf{M}\mathbf{x}\|^2, \tag{45}$$

$$d(\mathbf{y}) := \min_{\mathbf{x}\in\mathcal{X}} \mathcal{L}(\mathbf{x},\mathbf{y}), \tag{46}$$

$$d^* = \max_{\mathbf{y}} d(\mathbf{y}).$$

In addition, the authors denote the smoothness parameter of $f(\mathbf{x}) + \frac{\lambda}{2}\|\mathbf{M}\mathbf{x}\|^2$ by L_λ (assuming f is smooth), the set of optimal dual solutions is denoted by $\mathcal{Y}^*$, the minimal distance between $\mathbf{y}$ and a point in $\mathcal{Y}^*$is denoted by $\text{dist}(\mathbf{y},\mathcal{Y}^*)$, and the diameter of $\mathcal{X}$ is denoted by D (defined as the maximum norm between two points in $\mathcal{X}$).

Through these notations, the authors state the following theorem, which is crucial for their convergence analysis.

Theorem 6 (Theorem 1 in Gidel et al. (2018))**.** *Under Slater's condition, there exists $\alpha > 0$, such that for any dual variable $\mathbf{y}$, the following holds:*

$$d^* - d(\mathbf{y}) \geq \frac{1}{2L_\lambda D^2}\min\{\alpha^2 \text{dist}^2(\mathbf{y},\mathcal{Y}^*), \alpha L_\lambda D^2 \text{dist}(\mathbf{y},\mathcal{Y}^*)\}. \tag{92}$$

In order to prove this result, they define the function

$$f_\lambda(\mathbf{x}) := f(\mathbf{x}) + \mathbf{1}_{\mathcal{X}}(\mathbf{x}) + \frac{\lambda}{2}\|\mathbf{M}\mathbf{x}\|^2 \tag{48}$$

and at the beginning of the proof of Lemma 1 (on page 15), for any $\mathbf{x}\in\mathcal{X}$ and any given $\mathbf{u}\in\mathbb{R}^m$, the authors also define the function:

$$g_{\mathbf{x}}(\mathbf{u}) := f_\lambda(\mathbf{u}+\mathbf{x}) - f_\lambda(\mathbf{x}) - \langle \mathbf{u}, \mathbf{n}\rangle, \quad \mathbf{n}\in\partial f_\lambda(\mathbf{x}). \tag{58}$$

From the definition of f_λ we can equivalently write it as follows

$$(f + \frac{1}{2}\|\mathbf{M}\cdot\|^2)(\mathbf{u}+\mathbf{x}) - (f+\frac{1}{2}\|\mathbf{M}\cdot\|^2)(\mathbf{x}) - \langle \mathbf{u}, \nabla(f+\frac{1}{2}\|\mathbf{M}\cdot\|^2)(\mathbf{x})\rangle + \mathbf{1}_{\mathcal{X}}(\mathbf{u}+\mathbf{x}) - \mathbf{1}_{\mathcal{X}}(\mathbf{x}) - \langle \mathbf{u}, \mathbf{n}'\rangle,$$

where $n'\in\partial\mathbf{1}_{\mathcal{X}}(\mathbf{x}) = N_c^{\mathcal{X}}(\mathbf{x}) := \{\mathbf{n}'\in\mathbb{R}^m | \langle \mathbf{n}', \mathbf{x}-\mathbf{x}'\rangle \geq 0, \forall \mathbf{x}'\in\mathcal{X}\}$ is the normal cone of $\mathcal{X}$.

Now, right after the definition of $g_{\mathbf{x}}(\mathbf{u})$ in Eq. (58), the authors define the following function

$$h_{\mathbf{x}}(\mathbf{u}) := \frac{L_\lambda}{2}\|\mathbf{u}\|^2 + \mathbf{1}_{\mathcal{X}}(\mathbf{u}+\mathbf{x}),$$

and use the Descent Lemma[5] to claim that

$$\forall \mathbf{u}\in\mathbb{R}^m: \quad g_{\mathbf{x}}(\mathbf{u}) \leq h_{\mathbf{x}}(\mathbf{u}). \tag{17}$$

However, we observe that using the Descent Lemma and the fact that $\mathbf{x}\in\mathcal{X}$, we obtain that

$$g_{\mathbf{x}}(\mathbf{u}) \leq \frac{L_\lambda}{2}\|\mathbf{u}\|^2 + \mathbf{1}_{\mathcal{X}}(\mathbf{u}+\mathbf{x}) - \langle \mathbf{u}, \mathbf{n}'\rangle = h_{\mathbf{x}}(\mathbf{u}) - \langle \mathbf{u}, \mathbf{n}'\rangle.$$

Now, if $\mathbf{u}+\mathbf{x}\in\mathcal{X}$, then the rightmost term highlighted in red (including the minus sign) is non-negative by the definition of the normal cone (taking $\mathbf{x}' = \mathbf{u}+\mathbf{x}$), and thus can't be omitted, and so the inequality (17) cannot be deduced.

Adding the missing term throughout the rest of the proof, we eventually get a more complicated inequality instead of (75) in the original paper. Thus, the value of the constant α in the theorem above, for which the positivity is proven (see (76) in Gidel et al. (2018)), is irrelevant, and the fixed expression for α for which we can obtain the error bound property (see (92)), becomes even more complicated. It is then unclear that α can be proven to be strictly positive, which is required for the convergence results in Gidel et al. (2018).

[5] Recall that in this case, the Descent Lemma reads as follows
$(f+\frac{1}{2}\|\mathbf{M}\cdot\|^2)(\mathbf{u}+\mathbf{x}) - (f+\frac{1}{2}\|\mathbf{M}\cdot\|^2)(\mathbf{x}) - \langle \mathbf{u}, \nabla(f+\frac{1}{2}\|\mathbf{M}\cdot\|^2)(\mathbf{x})\rangle \leq \frac{L_\lambda}{2}\|\mathbf{u}\|^2$.

B Smoothness and Primal Quadratic Gap Bounds for the Augmented Lagrangian

B.1 Proof of Lemma 2.1(smoothness of S w.r.t. the primal variable)

Proof. Since $f(\mathbf{x})$ is β-smooth we obviously have that $S(\cdot, \mathbf{w})$, for any fixed $\mathbf{w} \in \mathbb{E}_2$, is $(\beta + \rho\|\mathcal{K}\|^2)$-smooth. Moreover, recalling the definition of the spectral norm, we have

$$\|\mathcal{K}\| := \max_{\|\mathbf{q}\|^2=1} \|\mathcal{K}\mathbf{q}\| = \max_{\|\mathbf{x}\|^2+\|\mathbf{y}\|^2=1} \|\mathcal{A}\mathbf{x} - \mathbf{y}\| \leq \max_{\|\mathbf{x}\|^2+\|\mathbf{y}\|^2=1} \|\mathcal{A}\mathbf{x}\| + \|\mathbf{y}\| \leq \max_{\|\mathbf{x}\|^2=1} \|\mathcal{A}\mathbf{x}\| + \max_{\|\mathbf{y}\|^2=1} \|\mathbf{y}\|^2 = \|\mathcal{A}\| + 1.$$

Hence, the desired result follows. □

B.2 Proof of Theorem 1 (primal quadratic gap when f is strongly convex)

Proof. Let us define $\phi_{\mathbf{w}}(\mathbf{q}) := S(\mathbf{q}, \mathbf{w})$ for any $\mathbf{q} \in \mathbb{E}$ and any $\mathbf{w} \in \mathbb{E}_2$. We start the proof by presenting two properties of $\phi_{\mathbf{w}}$, that hold regardless of the strong convexity of f.

First, from the differentiability of f, for any $\mathbf{q} = \begin{bmatrix}\mathbf{x}\\ \mathbf{y}\end{bmatrix} \in \mathbb{E}$, we have that

$$\nabla\phi_{\mathbf{w}}(\mathbf{q}) = \begin{bmatrix}\nabla f(\mathbf{x})\\ \mathbf{0}\end{bmatrix} + \mathcal{K}^\top\mathbf{w} + \rho\mathcal{K}^\top\mathcal{K}\mathbf{q}. \tag{18}$$

Second, from the definition of $\phi_{\mathbf{w}}$, we get, for any $\mathbf{q}_1 = \begin{bmatrix}\mathbf{x}_1\\ \mathbf{y}_1\end{bmatrix}, \mathbf{q}_2 = \begin{bmatrix}\mathbf{x}_2\\ \mathbf{y}_2\end{bmatrix} \in \mathbb{E}$, that

$$\begin{aligned}\phi_{\mathbf{w}}(\mathbf{q}_2) - \phi_{\mathbf{w}}(\mathbf{q}_1) =& f(\mathbf{x}_2) - f(\mathbf{x}_1) + \langle \mathbf{w}, \mathcal{K}(\mathbf{q}_2 - \mathbf{q}_1)\rangle + \frac{\rho}{2}\|\mathcal{K}\mathbf{q}_2\|^2 - \frac{\rho}{2}\|\mathcal{K}\mathbf{q}_1\|^2\\ =& f(\mathbf{x}_2) - f(\mathbf{x}_1) + \langle \mathcal{K}^\top\mathbf{w}, \mathbf{q}_2 - \mathbf{q}_1\rangle + \rho\langle \mathcal{K}\mathbf{q}_1, \mathcal{K}(\mathbf{q}_2 - \mathbf{q}_1)\rangle + \frac{\rho}{2}\|\mathcal{K}\mathbf{q}_2 - \mathcal{K}\mathbf{q}_1\|^2\\ =& f(\mathbf{x}_2) - f(\mathbf{x}_1) + \langle \mathcal{K}^\top\mathbf{w} + \rho\mathcal{K}^\top\mathcal{K}\mathbf{q}_1, \mathbf{q}_2 - \mathbf{q}_1\rangle + \frac{\rho}{2}\|\mathcal{K}\mathbf{q}_2 - \mathcal{K}\mathbf{q}_1\|^2.\end{aligned} \tag{19}$$

Now, from the α-strong convexity of $f(\mathbf{x})$, we have, for any $\mathbf{x}_1, \mathbf{x}_2 \in \mathbb{E}_1$, that

$$f(\mathbf{x}_2) - f(\mathbf{x}_1) \geq \langle \nabla f(\mathbf{x}_1), \mathbf{x}_2 - \mathbf{x}_1\rangle + \frac{\alpha}{2}\|\mathbf{x}_2 - \mathbf{x}_1\|^2. \tag{20}$$

From (18), (19) and (20) we further have

$$\begin{aligned}\phi_{\mathbf{w}}(\mathbf{q}_2) - \phi_{\mathbf{w}}(\mathbf{q}_1) &\geq \langle \nabla f(\mathbf{x}_1), \mathbf{x}_2 - \mathbf{x}_1\rangle + \frac{\alpha}{2}\|\mathbf{x}_2 - \mathbf{x}_1\|^2 + \frac{\rho}{2}\|\mathcal{K}\mathbf{q}_2 - \mathcal{K}\mathbf{q}_1\|^2 + \langle \mathcal{K}^\top\mathbf{w} + \rho\mathcal{K}^\top\mathcal{K}\mathbf{q}_1, \mathbf{q}_2 - \mathbf{q}_1\rangle.\\ &= \langle \nabla\phi_{\mathbf{w}}(\mathbf{q}_1), \mathbf{q}_2 - \mathbf{q}_1\rangle + \frac{\alpha}{2}\|\mathbf{x}_2 - \mathbf{x}_1\|^2 + \frac{\rho}{2}\|\mathcal{A}\mathbf{x}_2 - \mathcal{A}\mathbf{x}_1 + \mathbf{y}_2 - \mathbf{y}_1\|^2,\end{aligned} \tag{21}$$

where the last equality also follows from the definition of $\mathcal{K}$.

Now, since for every $a, b \in \mathbb{R}$, and $s > 0$ it holds that

$$(a+b)^2 \geq (1-s)a^2 + \frac{s-1}{s}b^2,$$

we get that for every $s \geq 1$

$$\begin{aligned}\|\mathcal{A}\mathbf{x}_2 - \mathcal{A}\mathbf{x}_1 + \mathbf{y}_2 - \mathbf{y}_1\|^2 &\geq (\|\mathcal{A}\mathbf{x}_2 - \mathcal{A}\mathbf{x}_1\| - \|\mathbf{y}_2 - \mathbf{y}_1\|)^2\\ &\geq (1-s)\|\mathcal{A}\mathbf{x}_2 - \mathcal{A}\mathbf{x}_1\|^2 + \frac{s-1}{s}\|\mathbf{y}_2 - \mathbf{y}_1\|^2\\ &\geq \|\mathcal{A}\|^2(1-s)\|\mathbf{x}_2 - \mathbf{x}_1\|^2 + \frac{s-1}{s}\|\mathbf{y}_2 - \mathbf{y}_1\|^2,\end{aligned} \tag{22}$$

where the first inequality follows from the well-known Cauchy-Schwarz inequality.

Combining (21) and (22), and denoting $\delta(s) := \min\{\alpha + \rho\|\mathcal{A}\|^2(1-s), \rho(s-1)/s\}$, we get for every $s \geq 1$, that

$$\begin{aligned}\phi_{\mathbf{w}}(\mathbf{q}_2) - \phi_{\mathbf{w}}(\mathbf{q}_1) &\geq \langle \nabla\phi_{\mathbf{w}}(\mathbf{q}_1), \mathbf{q}_2 - \mathbf{q}_1\rangle + \frac{1}{2}\left(\alpha + \rho\|\mathcal{A}\|^2(1-s)\right)\|\mathbf{x}_2 - \mathbf{x}_1\|^2 + \frac{\rho(s-1)}{2s}\|\mathbf{y}_2 - \mathbf{y}_1\|^2 \\ &\geq \langle \nabla\phi_{\mathbf{w}}(\mathbf{q}_1), \mathbf{q}_2 - \mathbf{q}_1\rangle + \frac{\delta(s)}{2}(\|\mathbf{x}_2 - \mathbf{x}_1\|^2 + \|\mathbf{y}_2 - \mathbf{y}_1\|^2) \\ &= \langle \nabla\phi_{\mathbf{w}}(\mathbf{q}_1), \mathbf{q}_2 - \mathbf{q}_1\rangle + \frac{\delta(s)}{2}\|\mathbf{q}_2 - \mathbf{q}_1\|^2.\end{aligned}$$

The result now follows, since taking $\tilde{s} = 1 + \frac{\alpha}{2\rho\|\mathcal{A}\|^2} > 1$, we have

$$\delta(\tilde{s}) = \alpha_S = \min\left\{\frac{\alpha}{2}, \frac{\alpha\rho}{\alpha + 2\rho\|\mathcal{A}\|^2}\right\} > 0,$$

as required. □

B.3 Proof of Theorem 2 (primal quadratic gap when $\mathcal{R}_{\mathcal{Q}}$ is an indicator for a polytope)

The strong duality of a primal dual saddle point problem ensures that for every primal optimal solution there exists a dual optimal solution (and vice versa) such that the pair forms a saddle point. In the case of Problem (3), we can observe the following crucial property, which proves that the set of saddle points is the entire set $\mathcal{P}^* \times \mathcal{D}^*$, meaning that if strong duality holds, every couple consisting of an optimal solution of Problem (OP) and an optimal solution of the dual problem associated with (OP), is a saddle point of $\mathcal{L}_\rho$.

Proposition 2. *Let $(\boldsymbol{q}_1^*, \boldsymbol{w}_1^*)$ and $(\boldsymbol{q}_2^*, \boldsymbol{w}_2^*)$ be two saddle points of $\mathcal{L}_\rho$. Then, $(\boldsymbol{q}_1^*, \boldsymbol{w}_2^*)$ is also a saddle point of $\mathcal{L}_\rho$.*

Proof. Since $\mathbf{q}_1^*$ and $\mathbf{q}_2^*$ are optimal solutions of Problem (OP), they are in particular feasible solutions of Problem (OP) and thus satisfy $\mathcal{K}\mathbf{q}_1^* = \mathcal{K}\mathbf{q}_2^* = \mathbf{0}$. Hence, recalling the definition of $\mathcal{L}_\rho$ we have that $\mathcal{L}_\rho(\mathbf{q}_1^*, \mathbf{w}_1^*) = \mathcal{L}_\rho(\mathbf{q}_1^*, \mathbf{w}_2^*)$ and that $\mathcal{L}_\rho(\mathbf{q}_2^*, \mathbf{w}_1^*) = \mathcal{L}_\rho(\mathbf{q}_2^*, \mathbf{w}_2^*)$.

Therefore, applying the saddle point inequality (2) with $(\mathbf{q}_1^*, \mathbf{w}_1^*)$, we have for all $\mathbf{w} \in \mathbb{E}_2$, that

$$\mathcal{L}_\rho(\mathbf{q}_1^*, \mathbf{w}) \leq \mathcal{L}_\rho(\mathbf{q}_1^*, \mathbf{w}_1^*) = \mathcal{L}_\rho(\mathbf{q}_1^*, \mathbf{w}_2^*). \tag{23}$$

In addition, applying (2) with $(\mathbf{q}_2^*, \mathbf{w}_2^*)$, we have for all $\mathbf{q} \in \mathbb{E}$, that

$$\mathcal{L}_\rho(\mathbf{q}, \mathbf{w}_2^*) \geq \mathcal{L}_\rho(\mathbf{q}_2^*, \mathbf{w}_2^*) = \mathcal{L}_\rho(\mathbf{q}_2^*, \mathbf{w}_1^*) \geq \mathcal{L}_\rho(\mathbf{q}_1^*, \mathbf{w}_1^*) = \mathcal{L}_\rho(\mathbf{q}_1^*, \mathbf{w}_2^*). \tag{24}$$

Note that the second inequality is another application of (2) with the saddle point $(\mathbf{q}_1^*, \mathbf{w}_1^*)$.

By (23) and (24) we get that $(\mathbf{q}_1^*, \mathbf{w}_2^*)$ is a saddle point. □

Lemma B.1. *Let $(\boldsymbol{x}_1^*, \boldsymbol{y}_1^*) \equiv \boldsymbol{q}_1^*$ and $(\boldsymbol{x}_2^*, \boldsymbol{y}_2^*) \equiv \boldsymbol{q}_2^*$ be two optimal solutions of Problem* (OP)*. Then, $\mathcal{B}\boldsymbol{x}_1^* = \mathcal{B}\boldsymbol{x}_2^*$.*

Proof. Let $\mathbf{w}^* \in \mathcal{D}^*$. From the optimality of $\mathbf{q}_1^*, \mathbf{q}_2^*, \mathbf{w}^*$, and by Assumptions 1 and 2, as well as Proposition 2, we have that both $(\mathbf{q}_1^*, \mathbf{w}^*)$ and $(\mathbf{q}_2^*, \mathbf{w}^*)$ are saddle points. By the saddle point property given in (2), we have

$$\mathcal{L}_\rho(\mathbf{q}_1^*, \mathbf{w}^*) = \mathcal{L}_\rho(\mathbf{q}_2^*, \mathbf{w}^*) = \min_{\mathbf{q}\in\mathbb{E}} \mathcal{L}_\rho(\mathbf{q}, \mathbf{w}^*). \tag{25}$$

Assume on the contrary that $\mathcal{B}\mathbf{x}_1^* \neq \mathcal{B}\mathbf{x}_2^*$. Then, by the α_g-strong convexity of g, we have

$$g\left(\mathcal{B}\left(\frac{\mathbf{x}_1^* + \mathbf{x}_2^*}{2}\right)\right) \leq \frac{1}{2}(g(\mathcal{B}\mathbf{x}_1^*) + g(\mathcal{B}\mathbf{x}_2^*)) - \frac{\alpha_g}{8}\|\mathcal{B}\mathbf{x}_1^* - \mathcal{B}\mathbf{x}_2^*\|^2 < \frac{1}{2}(g(\mathcal{B}\mathbf{x}_1^*) + g(\mathcal{B}\mathbf{x}_2^*)). \tag{26}$$

From the convexity of $\mathcal{R}_{\mathcal{X}}$ and $\mathcal{R}_{\mathcal{Y}}$, we have that $\mathcal{R}_{\mathcal{Q}}(\mathbf{q})$ is convex. Hence, the function

$$\psi(\mathbf{q}) := \mathcal{R}_{\mathcal{Q}}(\mathbf{q}) + \langle \mathbf{w}^*, \mathcal{K}\mathbf{q}\rangle + \frac{\rho}{2}\|\mathcal{K}\mathbf{q}\|^2,$$

is convex in $\mathbf{q}$. Thus, from (25) and (26), we have

$$\begin{aligned}\mathcal{L}_\rho\left(\frac{\mathbf{q}_1^*+\mathbf{q}_2^*}{2},\mathbf{w}^*\right) &= g\left(\mathcal{B}\left(\frac{\mathbf{x}_1^*+\mathbf{x}_2^*}{2}\right)\right)+\psi\left(\frac{\mathbf{q}_1^*+\mathbf{q}_2^*}{2}\right)\\ &< \frac{1}{2}g(\mathcal{B}\mathbf{x}_1^*)+\frac{1}{2}g(\mathcal{B}\mathbf{x}_2^*)+\frac{1}{2}\psi(\mathbf{q}_1^*)+\frac{1}{2}\psi(\mathbf{q}_2^*)\\ &= \frac{1}{2}\mathcal{L}_\rho(\mathbf{q}_1^*,\mathbf{w}^*)+\frac{1}{2}\mathcal{L}_\rho(\mathbf{q}_2^*,\mathbf{w}^*)\\ &= \min_{\mathbf{q}\in\mathbb{E}}\mathcal{L}_\rho(\mathbf{q},\mathbf{w}^*),\end{aligned}$$

which is a contradiction. □

The following Lemma is a known property of polytopes, which is stated without a proof. For a proof we refer the reader to Lemma 4 of Garber (2019).

Lemma B.2. *(Hoffman's Lemma) Let $\mathcal{F}:=\{\boldsymbol{q}\in\mathbb{E}|\mathcal{C}\boldsymbol{q}\leq\boldsymbol{b}\}$ be a compact and convex polytope and let $\mathcal{T}:\mathbb{V}_1\to\mathbb{V}_2$ be a linear mapping. Given some $\boldsymbol{c}\in\mathbb{V}_2$, we define the set $\mathcal{F}(\mathcal{T},\boldsymbol{c}):=\{\boldsymbol{q}\in\mathcal{F}|\mathcal{T}\boldsymbol{q}=\boldsymbol{c}\}$. If $\mathcal{F}(\mathcal{T},\boldsymbol{c})\neq\emptyset$, then there exists $\sigma>0$ such that for any $\boldsymbol{q}\in\mathcal{P}$ we have*

$$\text{dist}(\boldsymbol{q},\mathcal{F}(\mathcal{T},\boldsymbol{c}))\leq\sigma\|\mathcal{T}\boldsymbol{q}-\boldsymbol{c}\|^2,$$

where we define $\text{dist}(\boldsymbol{q},\mathcal{F}(\mathcal{T},\boldsymbol{c})):=\min_{\boldsymbol{z}\in\mathcal{F}(\mathcal{T},\boldsymbol{c})}\|\boldsymbol{z}-\boldsymbol{q}\|^2$.

Remark 1. The value of σ depends on $\mathcal{C}$ and $\mathcal{T}$.

Lemma B.3. *There exists a constant $\sigma>0$, such that for any $\boldsymbol{q}\equiv(\boldsymbol{x},\boldsymbol{y})\in\mathcal{F}$, we have*

$$f(\boldsymbol{x}^*)-f(\boldsymbol{x})\geq\langle\nabla f(\boldsymbol{x}),\boldsymbol{x}^*-\boldsymbol{x}\rangle+\frac{\alpha_g\sigma^{-1}}{2}\|\boldsymbol{q}^*-\boldsymbol{q}\|^2-\frac{\alpha_g}{2}\|\mathcal{K}\boldsymbol{q}\|^2,\tag{27}$$

where $\boldsymbol{q}^*\equiv(\boldsymbol{x}^*,\boldsymbol{y}^*)\in\mathcal{P}^*$ is the projection of $\boldsymbol{q}$ onto $\mathcal{P}^*$.

Proof. From Lemma B.1, it follows that there exists some $\mathbf{b}^*\in\mathbb{E}_3$, such that for every optimal solution $\mathbf{x}^*$ of Problem (OP), we have that $\mathcal{B}\mathbf{x}^*=\mathbf{b}^*$.

Now, let us denote $\mathcal{F}_{\mathbf{b}^*}:=\{\mathbf{q}\in\mathcal{F}:\mathcal{B}\mathbf{x}=\mathbf{b}^*,\mathcal{K}\mathbf{q}=\mathbf{0}\}$. We will now show that $\mathcal{F}_{\mathbf{b}^*}=\mathcal{P}^*$.

Let $\mathbf{q}\in\mathcal{P}^*$, then obviously $\mathcal{K}\mathbf{q}=\mathbf{0}$ and $\mathbf{q}\in\mathcal{F}$ (a feasible solution) and from Lemma B.1 it follows that $\mathcal{B}\mathbf{x}=\mathbf{b}^*$. Therefore, $\mathcal{P}^*\subseteq\mathcal{F}_{\mathbf{b}^*}$.

On the other hand, any point $\mathbf{q}\in\mathcal{F}_{\mathbf{b}^*}$ is feasible, as it satisfies $\mathcal{K}\mathbf{q}=\mathbf{0}$. Moreover, by the choice of $\mathbf{b}^*$, any $\mathbf{q}\equiv(\mathbf{x},\mathbf{y})\in\mathcal{F}_{\mathbf{b}^*}$ satisfies $\mathcal{B}\mathbf{x}=\mathbf{b}^*=\mathcal{B}\mathbf{x}^*$, for any optimal primal solution $\mathbf{q}^*\equiv(\mathbf{x}^*,\mathcal{A}\mathbf{x}^*)\in\mathcal{P}^*$. Now, since we also have that $\mathbf{q}\in\mathcal{F}$, the value of the objective of $\mathbf{q}$ satisfies, for any such point $\mathbf{q}^*$

$$g(\mathcal{B}\mathbf{x})+\mathcal{R}_\mathcal{Q}(\mathbf{q})=g(\mathbf{b}^*)=g(\mathcal{B}\mathbf{x}^*)+\mathcal{R}_\mathcal{Q}(\mathbf{q}^*),$$

which means that the objective value of any $\mathbf{q}\in\mathcal{F}_{\mathbf{b}^*}$ is optimal. Therefore, $\mathbf{q}\in\mathcal{P}^*$ and we have that $\mathcal{F}_{\mathbf{b}^*}\subseteq\mathcal{P}^*$, and thus $\mathcal{P}^*=\mathcal{F}_{\mathbf{b}^*}$.

Let $0_{\mathbb{E}_2}:\mathbb{E}_2\to\mathbb{E}_2$ be the zero linear operator $0_{\mathbb{E}_2}(\mathbf{y})=\mathbf{0}$. By applying Hoffman's Lemma with

$$\tilde{\mathcal{B}}:=[\mathcal{B},0_{\mathbb{E}_2}],\quad \mathcal{T}:=\begin{bmatrix}\tilde{\mathcal{B}}\\ \mathcal{K}\end{bmatrix},\quad \mathbf{c}:=\begin{bmatrix}\mathbf{b}^*\\ \mathbf{0}\end{bmatrix},$$

there exists a constant $\sigma>0$, such that for any $\mathbf{q}\in\mathcal{F}$, we have (notice that $\tilde{\mathcal{B}}\mathbf{q}=\mathcal{B}\mathbf{x}$)

$$\text{dist}(\mathbf{q},\mathcal{F}_{\mathbf{b}^*})=\min_{\mathbf{z}\in\mathcal{F}_{\mathbf{b}^*}}\|\mathbf{z}-\mathbf{q}\|^2\leq\sigma\|\mathcal{T}\mathbf{q}-\mathbf{c}\|^2=\sigma\left\|\begin{bmatrix}\tilde{\mathcal{B}}\mathbf{q}-\mathbf{b}^*\\ \mathcal{K}\mathbf{q}\end{bmatrix}\right\|^2=\sigma(\|\tilde{\mathcal{B}}\mathbf{q}-\mathbf{b}^*\|^2+\|\mathcal{K}\mathbf{q}\|^2)=\sigma(\|\mathcal{B}\mathbf{x}-\mathbf{b}^*\|^2+\|\mathcal{K}\mathbf{q}\|^2).$$

Thus, by denoting $\mathbf{q}^*:=\arg\min_{\mathbf{z}\in\mathcal{F}_{\mathbf{b}^*}}\|\mathbf{z}-\mathbf{q}\|^2$, we obtain that

$$\|\mathcal{B}\mathbf{x}-\mathbf{b}^*\|^2\geq\sigma^{-1}\|\mathbf{q}^*-\mathbf{q}\|^2-\|\mathcal{K}\mathbf{q}\|^2.\tag{28}$$

Now, from (28), the α_g-strong convexity of g, and the optimality of $\mathbf{q}^* = \begin{bmatrix}\mathbf{x}^* \\ \mathbf{y}^*\end{bmatrix}$, we have

$$\begin{aligned} f(\mathbf{x}^*) - f(\mathbf{x}) &= g(\mathcal{B}\mathbf{x}^*) - g(\mathcal{B}\mathbf{x}) \\ &\geq \langle \mathbf{x}^* - \mathbf{x}, \mathcal{B}^\top \nabla g(\mathcal{B}\mathbf{x})\rangle + \frac{\alpha_g}{2}\|\mathcal{B}\mathbf{x}^* - \mathcal{B}\mathbf{x}\|^2 \\ &= \langle \mathbf{x}^* - \mathbf{x}, \mathcal{B}^\top \nabla g(\mathcal{B}\mathbf{x})\rangle + \frac{\alpha_g}{2}\|\mathbf{b}^* - \mathcal{B}\mathbf{x}\|^2 \\ &\geq \langle \mathbf{x}^* - \mathbf{x}, \nabla f(\mathbf{x})\rangle + \frac{\alpha_g \sigma^{-1}}{2}\|\mathbf{q}^* - \mathbf{q}\|^2 - \frac{\alpha_g}{2}\|\mathcal{K}\mathbf{q}\|^2. \end{aligned}$$

This completes the desired result since $\mathcal{P}^* = \mathcal{F}_{\mathbf{b}^*}$. □

Applying Lemma B.3, as well as the two properties of $\phi_{\mathbf{w}}(\mathbf{q})$ presented in (18) and (19), we can now prove Theorem 2.

Proof of Theorem 2. For any $\mathbf{q} \in \mathcal{F}$, we denote $\mathbf{q}^* = \arg\min_{\mathbf{z}\in\mathcal{P}^*} \|\mathbf{z} - \mathbf{q}\|^2$. Recalling the beginning of the proof of Theorem 1, by plugging $\mathbf{q}_1 = \mathbf{q} \equiv (\mathbf{x}, \mathbf{y})$ and $\mathbf{q}_2 = \mathbf{q}^* \equiv (\mathbf{x}^*, \mathbf{y}^*)$ in (19), we have (recall that $\mathcal{K}\mathbf{q}^* = \mathbf{0}$)

$$\phi_{\mathbf{w}}(\mathbf{q}^*) - \phi_{\mathbf{w}}(\mathbf{q}) = f(\mathbf{x}^*) - f(\mathbf{x}) + \langle \mathcal{K}^\top \mathbf{w} + \rho \mathcal{K}^\top \mathcal{K}\mathbf{q}, \mathbf{q}^* - \mathbf{q}\rangle + \frac{\rho}{2}\|\mathcal{K}\mathbf{q}\|^2.$$

Combining it with Lemma B.3, and recalling (18), we have

$$\begin{aligned} \phi_{\mathbf{w}}(\mathbf{q}^*) - \phi_{\mathbf{w}}(\mathbf{q}) &\geq \langle \nabla f(\mathbf{x}), \mathbf{x}^* - \mathbf{x}\rangle + \frac{\alpha_g \sigma^{-1}}{2}\|\mathbf{q}^* - \mathbf{q}\|^2 - \frac{\alpha_g}{2}\|\mathcal{K}\mathbf{q}\|^2 + \frac{\rho}{2}\|\mathcal{K}\mathbf{q}^* - \mathcal{K}\mathbf{q}\|^2 + \langle \mathcal{K}^\top \mathbf{w} + \rho \mathcal{K}^\top \mathcal{K}\mathbf{q}, \mathbf{q}^* - \mathbf{q}\rangle \\ &= \langle \nabla \phi_{\mathbf{w}}(\mathbf{q}), \mathbf{q}^* - \mathbf{q}\rangle + \frac{\alpha_g \sigma^{-1}}{2}\|\mathbf{q}^* - \mathbf{q}\|^2 + \frac{\rho - \alpha_g}{2}\|\mathcal{K}\mathbf{q}\|^2 \\ &\geq \langle \nabla \phi_{\mathbf{w}}(\mathbf{q}), \mathbf{q}^* - \mathbf{q}\rangle + \frac{\alpha_g \sigma^{-1}}{2}\|\mathbf{q}^* - \mathbf{q}\|^2, \end{aligned}$$

where the last inequality is true since $\rho \geq \alpha_g$, which proves the desired result with $\alpha_S = \alpha_g \sigma^{-1}$ (recalling $\phi_{\mathbf{w}}(\mathbf{q}) = S(\mathbf{q}, \mathbf{w})$). □

C Proof of Theorem 4 (weak proximal oracle for polytopes)

Proof of Theorem 4. We present a proof that the point $\mathbf{v}_{\mathbf{x}}$ as defined in the theorem indeed satisfies (7). The proof that $\mathbf{v}_{\mathbf{y}}$ satisfies (9) follows the exact same arguments with the obvious modifications. We build on an observation from Garber and Hazan (2016), that given a point $\mathbf{x} \in \mathcal{F}$, that is formed by a convex combination of t vertices of $\mathcal{F}$, a vector $\mathbf{p} \in \mathbb{R}^d$, and a radius $r \in \mathbb{R}_+$, there exists a point $\tilde{\mathbf{x}} \in \mathcal{F}$ such that:

1. $\langle \tilde{\mathbf{x}}, \mathbf{p}\rangle \leq \langle \mathbf{z}, \mathbf{p}\rangle \ \forall \mathbf{z} \in \mathcal{F} \cap B(\mathbf{x}, r)$,
2. $\|\mathbf{x} - \tilde{\mathbf{x}}\| \leq \omega \cdot r$,
3. $\tilde{\mathbf{x}}$ is in the convex hull of the t vertices needed to represent $\mathbf{x}$, and the vertex which is the output of the LMO of $\mathcal{F}$ w.r.t. $\mathbf{p}$,

where $B(\mathbf{x}, r)$ is a ball of radius r centered at $\mathbf{x}$ and $\omega \geq 1$ is some constant that depends on the geometry of $\mathcal{F}$ (see Garber and Hazan (2016) Section 2).

For any $\mathbf{u} \in \mathcal{F}$, let $\tilde{\mathbf{x}}_{\mathbf{u}}$ be a point satisfying the above three properties for $r = \|\mathbf{u} - \mathbf{x}\|$ and $\mathbf{p} = \mathbf{p}_{\mathbf{x}} := \nabla_{\mathbf{x}} S(\mathbf{q}, \mathbf{w}) + 2\mu \mathcal{A}^\top \mathcal{K}\mathbf{q}$. Due to Property 3 above, $\tilde{\mathbf{x}}_{\mathbf{u}}$ can be written as a linear combination of the vertices $\mathbf{z}_1, \ldots, \mathbf{z}_{t+1}$, i.e., $\tilde{\mathbf{x}}_{\mathbf{u}} = \mathbf{M}\boldsymbol{\gamma}$, for some $\boldsymbol{\gamma} \in \mathbb{R}^{t+1}$ in the simplex (recall that the columns of $\mathbf{M}$ are $\mathbf{z}_1, \ldots, \mathbf{z}_{t+1}$). Therefore, by the choice $\mathbf{v}_{\mathbf{x}} = \mathbf{M}\boldsymbol{\gamma}^*_{\mathbf{x}}$, where $\gamma^*_{\mathbf{x}}$ is a minimizer of (12), we have for any $\mathbf{u} \in \mathcal{F}$, that

$$\langle \mathbf{v}_{\mathbf{x}}, \mathbf{p}\rangle + \frac{\eta(\beta_S + 2\mu\|\mathcal{K}\|^2)}{2}\|\mathbf{v}_{\mathbf{x}} - \mathbf{x}\|^2 \leq \langle \tilde{\mathbf{x}}_{\mathbf{u}}, \mathbf{p}\rangle + \frac{\eta(\beta_S + 2\mu\|\mathcal{K}\|^2)}{2}\|\tilde{\mathbf{x}}_{\mathbf{u}} - \mathbf{x}\|^2. \tag{29}$$

By combining Properties 1 and 2 above, we have for any $\mathbf{u} \in \mathcal{F}$, that

$$\langle \tilde{\mathbf{x}}_{\mathbf{u}}, \mathbf{p} \rangle + \frac{\eta(\beta_S + 2\mu\|\mathcal{K}\|^2)}{2}\|\tilde{\mathbf{x}}_{\mathbf{u}} - \mathbf{x}\|^2 \leq \langle \mathbf{u}, \mathbf{p} \rangle + \frac{\omega^2\eta(\beta_S + 2\mu\|\mathcal{K}\|^2)}{2}\|\mathbf{u} - \mathbf{x}\|^2. \tag{30}$$

Combining the last two inequalities, we have that (7) is indeed satisfied with $\mathcal{R}_{\mathcal{X}}(\cdot) = \delta_{\mathcal{F}}(\cdot)$ and $\lambda_{\mathbf{x}} = \omega^2$ for any $\mathbf{u} \in \mathcal{F}$, and in particular for any $\mathbf{x}^* \in \mathcal{X}^* \subseteq \mathcal{F}$. □

D Some Illustrative Examples

In this section, we describe several families of problems of interest for which our two main assumptions — the primal quadratic gap property (Definition 1) and the availability of an efficient weak proximal oracle (Definition 2), hold true.

D.1 Structured Low-Rank Matrix Recovery

Consider the following optimization problem in which the goal is to recover a structured low-rank matrix from some noisy matrix observation $\mathbf{\Sigma}$:

$$\min_{\mathbf{S} \in \mathbb{R}^{m \times n}} \frac{1}{2}\|\mathbf{S} - \mathbf{\Sigma}\|_F^2 + \nu\|\mathbf{S}\|_{\text{nuc}} + \mathcal{R}_{\mathcal{Y}}(\mathcal{A}(\mathbf{S})), \tag{31}$$

where $\nu > 0$, $\mathcal{R}_{\mathcal{Y}}$ is assumed to be a proximal friendly function (or even a weak proximal friendly function, recall Definition 3), and the *unique* optimal solution (guaranteed from strong convexity of the squared Frobenius norm term) is *low rank*.

For instance, when $\mathcal{A} := \mathcal{I}$ is the identity mapping and $\mathcal{R}_{\mathcal{Y}}$ is an ℓ_1 regularizer, Problem (31) is a natural convex relaxation for recovering a matrix that is both low-rank and sparse, see for instance Richard et al. (2012).

Setting $f(\mathbf{S}) := \frac{1}{2}\|\mathbf{S} - \mathbf{\Sigma}\|_F^2$, and $\mathcal{R}_{\mathcal{X}} := \nu\|\mathbf{S}\|_{\text{nuc}}$ we get a problem of the form of model (OP). Primal quadratic gap holds since f is strongly convex (see Theorem 1). By the assumption of low-rank of the optimal solution, as well as its uniqueness, according to the discussion in Section 2.3.2, $\mathcal{R}_{\mathcal{X}}$ is indeed a weak proximal friendly function and in general, Problem (31) is weak proximal friendly.

D.2 Low-Rank Tensor Recovery

The previous example could be extended to the more general and important problem of recovering low-rank tesnors. Analogously to matrices, the rank of a N-way real tensor $\mathbf{X}$, $N > 2$, could be defined as the minimum number of rank one tensors (of the same dimensions) whose sum equals the original tensor, and here we denote it by $\text{rank}(\mathbf{X})$. However, in general, even determining the rank of a given tensor is NP-Hard Johan (1990). An alternative is to use standard matrix rank in an appropriate manner. Let $\mathbf{X} \in \mathbb{R}^{n_1 \times \cdots \times n_N}$ be a N-way tensor. For any $i \in \{1, \ldots, N\}$, let $\mathcal{A}_i : \mathbb{R}^{n_1 \times \cdots \times n_N} \to \mathbb{R}^{n_i \times I_i}$, $I_i := \frac{1}{n_i}\prod_{j=1}^N n_j$, be a linear mapping that flattens the tensor into a $\mathbb{R}^{n_i \times I_i}$ matrix, by flattening all dimensions except for the ith dimension, see exact definition in Kolda and Bader (2009); Gandy et al. (2011). This leads to the definition of the n-rank of a N-way tensor $\mathbf{X}$, which is given by $\text{rank}_n(\mathbf{X}) := (\text{rank}(\mathcal{A}_1\mathbf{X}), \ldots, \text{rank}(\mathcal{A}_N\mathbf{X})) \in \mathbb{N}^N$, where $\text{rank}(\mathcal{A}_i\mathbf{X})$ is the standard matrix rank of $\mathcal{A}_i\mathbf{X}$. It is a fairly simple observation that $\max_{i \in \{1,\ldots,N\}} \text{rank}_n(\mathbf{X})(i) \leq \text{rank}(\mathbf{X})$. This leads to the following natural convex relaxation Gandy et al. (2011), which is analogous to nuclear norm-based relaxations for the matrix case, for the problem of recovering a low-rank tensor from a given noisy tensor measurement $\mathbf{T}$:

$$\begin{aligned} \min_{\mathbf{X} \in \mathbb{R}^{n_1 \times \cdots \times n_N}} & \frac{1}{2}\|\mathbf{X} - \mathbf{T}\|_2^2 + \mathcal{R}_{\mathcal{X}}(\mathbf{X}) + \nu\sum_{i=1}^N \|\mathbf{Y}_i\|_{\text{nuc}} \\ \text{s.t. } & \mathcal{A}_i\mathbf{X} = \mathbf{Y}_i \ \ \forall i \in [N], \end{aligned} \tag{32}$$

where $\|\cdot\|_2$ denotes the ℓ_2 norm for the appropriate tensor space, $\nu > 0$, and $\mathcal{R}_{\mathcal{X}}$ is some (weak) proximal friendly function that may encode additional structure of the tensor to be recovered (e.g., sparsity if we take it to be an ℓ_1 regularizer).

Using the notations of Problem (OP), we will denote the following $f(\mathbf{X}) = \frac{1}{2}\|\mathbf{X} - \mathbf{T}\|_2^2$, $\mathcal{A}\mathbf{X} := [\mathcal{A}_1\mathbf{X}^\top, \ldots, \mathcal{A}_N\mathbf{X}^\top]^\top$, $\mathbf{Y} = \mathbf{Y}_1 \times \cdots \times \mathbf{Y}_N \in \mathbb{R}^{n_1 \times I_1} \times \cdots \times \mathbb{R}^{n_N \times I_N}$, and $\mathcal{R}_{\mathcal{Y}}(\mathbf{Y}) := \nu\sum_{i=1}^N \|\mathbf{Y}_i\|_{\text{nuc}}$.

As in the previous example, the primal quadratic gap property holds since f is strongly convex and the optimal solution $(\mathbf{X}^*, \mathbf{Y}^*)$ is unique. Now, let us assume that $\mathbf{X}^*$ has a low n-rank, meaning $\text{rank}(\mathbf{Y}_i^*) = \text{rank}(\mathcal{A}_i\mathbf{X}^*) \leq k$, for a fairly small

k. In this case, since the variable $\mathbf{Y}$ is given as a cartesian product, the weak proximal computation w.r.t. the variable $\mathbf{Y}$ naturally decouples into N separate weak proximal computation w.r.t. each of the blocks $\mathbf{Y}_1, \ldots, \mathbf{Y}_N$. It should be noted that each block enjoys the same structure as in the low-rank matrix case[6], i.e., amounts to a k-SVD computation of a real matrix (as described in Section 2.3.1). Since additionally $\mathcal{R}_{\mathcal{X}}$ is assumed to be a (weak) proximal friendly function, we have that in general, Problem (32) is weak proximal friendly.

D.3 Least Squares Over Intersection of Polytopes

Let $\mathcal{F}_1, \ldots, \mathcal{F}_n$ be convex and compact polytopes in $\mathbb{R}^d$, for which a linear minimization oracle can be implemented efficiently. Given $\mathbf{M} \in \mathbb{R}^{p \times d}$ and $\mathbf{b} \in \mathbb{R}^p$, we consider the following constrained least squares optimization problem:

$$\min_{\mathbf{x} \in \mathbb{R}^d} \frac{1}{2}\|\mathbf{M}\mathbf{x} - \mathbf{b}\|^2 \quad \text{s.t.} \quad \mathbf{x} \in \bigcap_{i=1}^{n} \mathcal{F}_i. \tag{33}$$

We can see that this problem is of the form of (OP) by setting $f(\mathbf{x}) := \frac{1}{2}\|\mathbf{M}\mathbf{x} - \mathbf{b}\|^2$, $\mathcal{R}_{\mathcal{X}}(\mathbf{x}) := \delta_{\mathcal{F}_1}(\mathbf{x})$, $\mathcal{A} := [\mathcal{I}, \ldots, \mathcal{I}]^\top$ ($n-1$ times), where $\mathcal{I}$ is the identity mapping, and $\mathcal{R}_{\mathcal{Y}}(\mathbf{y}) = \mathcal{R}_{\mathcal{Y}}([\mathbf{y}_1^\top, \ldots, \mathbf{y}_{n-1}^\top]^\top) := \sum_{i=1}^{n-1} \delta_{\mathcal{F}_{i+1}}(\mathbf{y}_i)$, where $\delta_{\mathcal{F}_i}(\cdot)$ is the indicator function of the polytope $\mathcal{F}_i$, $i = 1, \ldots, n$.

Notice we can write f as $g \circ \mathcal{B}$, where $\mathcal{B} = \mathbf{M}$, and $g(\cdot) = \frac{1}{2}\|(\cdot) - \mathbf{b}\|^2$, which is a 1-strongly convex function. In addition, $\mathcal{R}_{\mathcal{Q}}$ here is an indicator function of the product of the polytopes $\mathcal{F}_1 \times \cdots \times \mathcal{F}_n$. Hence, according to Theorem 2, by setting the augmented Lagrangian parameter to $\rho \geq 1$, this problem satisfies the primal quadratic gap property.

Since we assumed each of the polytopes $\mathcal{F}_i$ admits an efficient linear minimization oracle, as discussed in Section 2.3.3, each of the indicator functions $\delta_{\mathcal{F}_i}$ is weak proximal friendly. Further more, since the variable $\mathbf{y}$ admits a simple cartesian product structure, the weak proximal oracle w.r.t. $\mathbf{y}$ naturally decouples into $n-1$ separate weak proximal oracle computations, each w.r.t. to one of the polytopes $\mathcal{F}_i$, $i = 2, \ldots, n$, and so, Problem (33) is also weak proximal friendly.

E Proof of Theorem 5

In order to prove Theorem 5, we first need to establish two intermediate results. The first result was established in Sabach and Teboulle (2019). For the sake of completeness we state it here and reprove it. The other is the proof of Lemma 3.1 stated before.

Lemma E.1. *(Objective and feasibility approximation) Let $(\boldsymbol{q}^*, \boldsymbol{w}^*)$ be a saddle point of $\mathcal{L}_\rho$. Let $\boldsymbol{q} \in \mathbb{E}$, and suppose that $c \geq 2\|\boldsymbol{w}^*\|$, for some $c > 0$. If*

$$h(\boldsymbol{q}) - h(\boldsymbol{q}^*) + c\|\mathcal{K}\boldsymbol{q}\| + \frac{\rho}{2}\|\mathcal{K}\boldsymbol{q}\|^2 \leq \delta, \tag{34}$$

holds for some $\delta \geq 0$, then the following assertions hold

1. $h(\boldsymbol{q}) - h(\boldsymbol{q}^*) \leq \delta$.
2. $\|\mathcal{K}\boldsymbol{q}\| \leq \frac{2\delta}{c}$.

Proof of Lemma E.1. The first assertion holds since $c\|\mathcal{K}\mathbf{q}\| + \frac{\rho}{2}\|\mathcal{K}\mathbf{q}\|^2 \geq 0$. Moreover, since $(\mathbf{q}^*, \mathbf{w}^*)$ is a saddle point of $\mathcal{L}_\rho$, we have

$$h(\mathbf{q}^*) = \mathcal{L}_\rho(\mathbf{q}^*, \mathbf{w}^*) \leq \mathcal{L}_\rho(\mathbf{q}, \mathbf{w}^*) = h(\mathbf{q}) + \langle \mathbf{w}^*, \mathcal{K}\mathbf{q} \rangle + \frac{\rho}{2}\|\mathcal{K}\mathbf{q}\|^2. \tag{35}$$

Combining (34) and (35) we have

$$c\|\mathcal{K}\mathbf{q}\| \leq \delta + \langle \mathbf{w}^*, \mathcal{K}\mathbf{q} \rangle \leq \delta + \|\mathbf{w}^*\| \cdot \|\mathcal{K}\mathbf{q}\| \leq \delta + \frac{c}{2}\|\mathcal{K}\mathbf{q}\|.$$

Rearranging the last inequality, yields the second assertion. □

We now prove Lemma 3.1.

[6] Gandy et al. (2011) formally shows this decoupling for strong proximal computations. It then becomes trivial to apply this for weak proximal computations.

Proof of Lemma 3.1. For any $t \in \mathbb{N}$, let $(\mathbf{q}_t^*, \mathbf{w}^*)$ be a saddle point of $\mathcal{L}_\rho$, such that the pair $\{\mathbf{q}_t, \mathbf{q}_t^*\}$ satisfies the primal quadratic gap (see (5)), i.e., $\mathbf{q}_t^* = \arg\min_{\mathbf{q}^* \in \mathcal{P}^*} \|\mathbf{q}_t - \mathbf{q}^*\|^2$ and $\mathbf{w}^* \in \mathcal{D}^*$. Since $\mathcal{L}_\rho(\mathbf{q}_t^*, \mathbf{w}^*) = \mathcal{L}_\rho^*$, we have

$$\begin{aligned} d_{t+1} &= \mathcal{L}_\rho(\mathbf{q}_{t+1}, \mathbf{w}_{t+1}) - \mathcal{L}_\rho(\mathbf{q}_t^*, \mathbf{w}^*) \\ &= \mathcal{L}_\rho(\mathbf{q}_{t+1}, \mathbf{w}_{t+1}) - \mathcal{L}_\rho(\mathbf{q}_{t+1}, \mathbf{w}_t) + \mathcal{L}_\rho(\mathbf{q}_{t+1}, \mathbf{w}_t) - \mathcal{L}_\rho(\mathbf{q}_t, \mathbf{w}_t) + \mathcal{L}_\rho(\mathbf{q}_t, \mathbf{w}_t) - \mathcal{L}_\rho(\mathbf{q}_t^*, \mathbf{w}^*) \\ &= \langle \mathbf{w}_{t+1} - \mathbf{w}_t, \mathcal{K}\mathbf{q}_{t+1} \rangle + \mathcal{L}_\rho(\mathbf{q}_{t+1}, \mathbf{w}_t) - \mathcal{L}_\rho(\mathbf{q}_t, \mathbf{w}_t) + d_t \\ &= d_t + \mu \|\mathcal{K}\mathbf{q}_{t+1}\|^2 + \mathcal{L}_\rho(\mathbf{q}_{t+1}, \mathbf{w}_t) - \mathcal{L}_\rho(\mathbf{q}_t, \mathbf{w}_t), \end{aligned} \tag{36}$$

where the last equality follows from (16). Now, using the feasibility of $\mathbf{q}_t^*$ ($\mathcal{K}\mathbf{q}_t^* = 0$), we have

$$\begin{aligned} \|\mathcal{K}\mathbf{q}_{t+1}\|^2 &= \|\mathcal{K}\mathbf{q}_{t+1} - \mathcal{K}\mathbf{q}_t + \mathcal{K}\mathbf{q}_t - \mathcal{K}\mathbf{q}_t^*\|^2 \\ &= \|\mathcal{K}\mathbf{q}_{t+1} - \mathcal{K}\mathbf{q}_t\|^2 + 2\langle \mathcal{K}\mathbf{q}_{t+1} - \mathcal{K}\mathbf{q}_t, \mathcal{K}\mathbf{q}_t - \mathcal{K}\mathbf{q}_t^* \rangle + \|\mathcal{K}\mathbf{q}_t - \mathcal{K}\mathbf{q}_t^*\|^2 \\ &= \eta^2 \|\mathcal{K}\mathbf{v}_t - \mathcal{K}\mathbf{q}_t\|^2 + 2\eta \langle \mathcal{K}\mathbf{v}_t - \mathcal{K}\mathbf{q}_t, \mathcal{K}\mathbf{q}_t \rangle + \|\mathcal{K}\mathbf{q}_t - \mathcal{K}\mathbf{q}_t^*\|^2 \\ &\leq \eta^2 \|\mathcal{K}\|^2 \cdot \|\mathbf{v}_t - \mathbf{q}_t\|^2 + \eta \langle \mathbf{v}_t - \mathbf{q}_t, 2\mathcal{K}^\top \mathcal{K}\mathbf{q}_t \rangle + \|\mathcal{K}\|^2 \cdot \|\mathbf{q}_t - \mathbf{q}_t^*\|^2, \end{aligned} \tag{37}$$

where the last equality follows from (15). By the β_S-smoothness of $S(\cdot, \mathbf{w})$, the convexity of $\mathcal{R}_\mathcal{Q}(\mathbf{q})$ and (15), we have

$$\begin{aligned} \mathcal{L}_\rho(\mathbf{q}_{t+1}, \mathbf{w}_t) - \mathcal{L}_\rho(\mathbf{q}_t, \mathbf{w}_t) &= S(\mathbf{q}_{t+1}, \mathbf{w}_t) - S(\mathbf{q}_t, \mathbf{w}_t) + \mathcal{R}_\mathcal{Q}(\mathbf{q}_{t+1}) - \mathcal{R}_\mathcal{Q}(\mathbf{q}_t) \\ &\leq \frac{\eta^2 \beta_S}{2} \|\mathbf{v}_t - \mathbf{q}_t\|^2 + \eta \langle \mathbf{v}_t - \mathbf{q}_t, \nabla_\mathbf{q} S(\mathbf{q}_t, \mathbf{w}_t) \rangle + \eta(\mathcal{R}_\mathcal{Q}(\mathbf{v}_t) - \mathcal{R}_\mathcal{Q}(\mathbf{q}_t)). \end{aligned} \tag{38}$$

Now, from (14), recalling (6), since $\mathbf{q}_t^* \in \mathcal{P}^*$, we have that

$$\begin{aligned} &\mathcal{R}_\mathcal{Q}(\mathbf{v}_t) + \langle \mathbf{v}_t, \nabla_\mathbf{q} S(\mathbf{q}_t, \mathbf{w}_t) + 2\mu \mathcal{K}^\top \mathcal{K}\mathbf{q}_t \rangle + \frac{\eta(\beta_S + 2\mu\|\mathcal{K}\|^2)}{2} \|\mathbf{v}_t - \mathbf{q}_t\|^2 \leq \\ &\mathcal{R}_\mathcal{Q}(\mathbf{q}_t^*) + \langle \mathbf{q}_t^*, \nabla_\mathbf{q} S(\mathbf{q}_t, \mathbf{w}_t) + 2\mu \mathcal{K}^\top \mathcal{K}\mathbf{q}_t \rangle + \frac{\lambda\eta(\beta_S + 2\mu\|\mathcal{K}\|^2)}{2} \|\mathbf{q}_t^* - \mathbf{q}_t\|^2. \end{aligned} \tag{39}$$

By combining (36), (37), (38) and (39) we obtain the following

$$\begin{aligned} d_{t+1} - d_t &= \mu \|\mathcal{K}\mathbf{q}_{t+1}\|^2 + \mathcal{L}_\rho(\mathbf{q}_{t+1}, \mathbf{w}_t) - \mathcal{L}_\rho(\mathbf{q}_t, \mathbf{w}_t) \\ &\leq \frac{2\mu\|\mathcal{K}\|^2}{2} \|\mathbf{q}_t - \mathbf{q}_t^*\|^2 + \frac{\eta^2(\beta_S + 2\mu\|\mathcal{K}\|^2)}{2} \|\mathbf{v}_t - \mathbf{q}_t\|^2 + \eta(\mathcal{R}_\mathcal{Q}(\mathbf{v}_t) - \mathcal{R}_\mathcal{Q}(\mathbf{q}_t)) \\ &+ \eta \langle \mathbf{v}_t - \mathbf{q}_t, \nabla_\mathbf{q} S(\mathbf{q}_t, \mathbf{w}_t) + 2\mu \mathcal{K}^\top \mathcal{K}\mathbf{q}_t \rangle \\ &\leq \frac{1}{2} \left(2\mu\|\mathcal{K}\|^2 + \lambda\eta^2(\beta_S + 2\mu\|\mathcal{K}\|^2)\right) \|\mathbf{q}_t - \mathbf{q}_t^*\|^2 + \eta(\mathcal{R}_\mathcal{Q}(\mathbf{q}_t^*) - \mathcal{R}_\mathcal{Q}(\mathbf{q}_t)) \\ &+ \eta \langle \mathbf{q}_t^* - \mathbf{q}_t, \nabla_\mathbf{q} S(\mathbf{q}_t, \mathbf{w}_t) \rangle + 2\mu\eta \langle \mathcal{K}\mathbf{q}_t^* - \mathcal{K}\mathbf{q}_t, \mathcal{K}\mathbf{q}_t \rangle. \end{aligned} \tag{40}$$

Let us, for convenience, denote $r = 2\mu(\|\mathcal{A}\| + 1)^2 \geq 2\mu\|\mathcal{K}\|^2$. Now, recall that $\mathbf{q}_t^*$ was chosen to satisfy (5). Therefore, from (40) and (5), as well as the feasibility of $\mathbf{q}_t^*$, we obtain that

$$\begin{aligned} d_{t+1} - d_t \leq& \frac{1}{2} \left(2\mu\|\mathcal{K}\|^2 - \eta\alpha_S + \lambda\eta^2(\beta_S + 2\mu\|\mathcal{K}\|^2)\right) \|\mathbf{q}_t - \mathbf{q}_t^*\|^2 - 2\mu\eta\|\mathcal{K}\mathbf{q}_t\|^2 \qquad (41) \\ &+ \eta[S(\mathbf{q}_t^*, \mathbf{w}_t) + \mathcal{R}_\mathcal{Q}(\mathbf{q}_t^*) - S(\mathbf{q}_t, \mathbf{w}_t) - \mathcal{R}_\mathcal{Q}(\mathbf{q}_t)] \\ \leq& \frac{1}{2} \left(r - \eta\alpha_S + \lambda\eta^2(\beta_S + r)\right) \|\mathbf{q}_t - \mathbf{q}_t^*\|^2 + \eta[\mathcal{L}_\rho(\mathbf{q}_t^*, \mathbf{w}_t) - \mathcal{L}_\rho(\mathbf{q}_t, \mathbf{w}_t)]. \qquad (42) \end{aligned}$$

Since $\mathbf{q}_t^*$ is a feasible solution of the Problem (OP), we have that $\mathcal{L}_\rho(\mathbf{q}_t^*, \mathbf{w}_t) = \mathcal{L}_\rho(\mathbf{q}_t^*, \mathbf{w}^*) = \mathcal{L}_\rho^*$. Thus, from (42), after substituting $\eta = \frac{\alpha_S}{2\lambda(\beta_S + r)} \in [0, 1]$, we have for all $t \in \mathbb{N}$ (recall that $\lambda \geq 1$)

$$d_{t+1} \leq (1 - \eta) d_t + \frac{1}{2}(r - \eta\alpha_S + \lambda\eta^2(\beta_S + r)) \|\mathbf{q}_t - \mathbf{q}_t^*\|^2 = (1 - \eta) d_t + \frac{1}{2} \left(r - \frac{\alpha_S^2}{4\lambda(\beta_S + r)}\right) \|\mathbf{q}_t - \mathbf{q}_t^*\|^2.$$

Taking any $\mu > 0$ such that $2\mu(\|\mathcal{A}\| + 1)^2 \equiv r \leq \frac{\sqrt{\lambda\alpha_S^2 + \lambda^2\beta_S^2} - \lambda\beta_S}{2\lambda}$, we get that $r - \frac{\alpha_S^2}{4\lambda(\beta_S + r)} \leq 0$, and therefore we immediately obtain that

$$d_{t+1} \leq (1 - \eta) d_t,$$

which proves the desired result. □

We will now proceed to prove our convergence rate result.

Proof of Theorem 5. First, let $\{\mathbf{w}_t\}_{t\in\mathbb{N}}$ be the sequence of dual points generated by Algorithm 1. For any $t \in \mathbb{N}$ and any $\mathbf{w} \in \mathbb{E}_2$, we define

$$\Delta_t(\mathbf{w}) := \frac{1}{2}\|\mathbf{w} - \mathbf{w}_t\|^2 - \frac{1}{2}\|\mathbf{w} - \mathbf{w}_{t+1}\|^2.$$

We notice that for any $t \in \mathbb{N}$ and $\mathbf{w} \in \mathbb{E}_2$ we have

$$\begin{aligned}
\mathcal{L}_\rho(\mathbf{q}_{t+1}, \mathbf{w}) - \mathcal{L}_\rho(\mathbf{q}_{t+1}, \mathbf{w}_{t+1}) &= \langle \mathbf{w} - \mathbf{w}_{t+1}, \mathcal{K}\mathbf{q}_{t+1}\rangle \\
&= \frac{1}{\mu}\langle \mathbf{w} - \mathbf{w}_{t+1}, \mathbf{w}_{t+1} - \mathbf{w}_t\rangle \\
&= \frac{1}{\mu}\Delta_t(\mathbf{w}) - \frac{1}{2\mu}\|\mathbf{w}_{t+1} - \mathbf{w}_t\|^2 \\
&\leq \frac{1}{\mu}\Delta_t(\mathbf{w}).
\end{aligned}$$

Thus, summing the above inequality for all $t = 0, 1, \ldots, T-1$, we have

$$\begin{aligned}
\sum_{t=0}^{T-1}(\mathcal{L}_\rho(\mathbf{q}_{t+1}, \mathbf{w}) - \mathcal{L}_\rho(\mathbf{q}_{t+1}, \mathbf{w}_{t+1})) &\leq \frac{1}{\mu}\sum_{t=0}^{T-1}\Delta_t(\mathbf{w}) \\
&= \frac{1}{2\mu}\sum_{t=0}^{T-1}(\|\mathbf{w} - \mathbf{w}_t\|^2 - \|\mathbf{w} - \mathbf{w}_{t+1}\|^2) \\
&= \frac{1}{2\mu}(\|\mathbf{w} - \mathbf{w}_0\|^2 - \|\mathbf{w} - \mathbf{w}_T\|^2) \qquad (43) \\
&\leq \frac{1}{2\mu}\|\mathbf{w} - \mathbf{w}_0\|^2 \\
&\leq \frac{1}{2\mu}(\|\mathbf{w}\| + \|\mathbf{w}_0\|)^2. \qquad (44)
\end{aligned}$$

In addition, from Lemma 3.1, we have that $d_{t+1} \leq d_1(1-\eta)^t$. Thus, if $d_1 > 0$, we have that

$$\sum_{t=0}^{T-1}(\mathcal{L}_\rho(\mathbf{q}_{t+1}, \mathbf{w}_{t+1}) - \mathcal{L}_\rho(\mathbf{q}^*, \mathbf{w}^*)) = \sum_{t=0}^{T-1} d_{t+1} \leq d_1\sum_{t=0}^{T-1}(1-\eta)^t \leq d_1\sum_{t=0}^{\infty}(1-\eta)^t = \frac{d_1}{\eta},$$

where the second equality follows from the classical result on the geometric series. Hence, using Lemma 2.1, we obtain that

$$\sum_{t=0}^{T-1}(\mathcal{L}_\rho(\mathbf{q}_{t+1}, \mathbf{w}_{t+1}) - \mathcal{L}_\rho(\mathbf{q}^*, \mathbf{w}^*)) \leq \frac{2d_1}{\alpha_S}\left(\beta + (2\mu + \rho)(\|\mathcal{A}\| + 1)^2\right). \qquad (45)$$

Otherwise, if $d_1 \leq 0$, we have

$$\sum_{t=0}^{T-1}(\mathcal{L}_\rho(\mathbf{q}_{t+1}, \mathbf{w}_{t+1}) - \mathcal{L}_\rho(\mathbf{q}^*, \mathbf{w}^*)) = \sum_{t=0}^{T-1} d_{t+1} \leq \sum_{t=0}^{T-1} d_1(1-\eta)^t \leq 0. \qquad (46)$$

Combining (45) and (46) yields

$$\sum_{t=0}^{T-1}(\mathcal{L}_\rho(\mathbf{q}_{t+1}, \mathbf{w}_{t+1}) - \mathcal{L}_\rho(\mathbf{q}^*, \mathbf{w}^*)) \leq \max\left\{0, \frac{2d_1}{\alpha_S}(\beta + (\rho + 2\mu)(\|\mathcal{A}\| + 1)^2)\right\}. \qquad (47)$$

Now, since $\mathcal{L}_\rho(\cdot, \mathbf{w})$ is convex, and since $\mathcal{L}_\rho(\mathbf{q}^*, \mathbf{w}^*) = \mathcal{L}_\rho(\mathbf{q}^*, \mathbf{w})$ for any $\mathbf{w} \in \mathbb{E}_2$, which follows from the feasibility of

$\mathbf{q}^*$, we get, for any $\mathbf{w} \in \mathbb{E}_2$, that

$$\begin{aligned}
h(\bar{\mathbf{q}}_T) - h(\mathbf{q}^*) + \langle \mathbf{w}, \mathcal{K}\bar{\mathbf{q}}_T \rangle + \frac{\rho}{2}\|\mathcal{K}\bar{\mathbf{q}}_T\|^2 &= \mathcal{L}_\rho(\bar{\mathbf{q}}_T, \mathbf{w}) - \mathcal{L}_\rho(\mathbf{q}^*, \mathbf{w}) \\
&\leq \frac{1}{T}\sum_{t=0}^{T-1}(\mathcal{L}_\rho(\mathbf{q}_{t+1}, \mathbf{w}) - \mathcal{L}_\rho(\mathbf{q}^*, \mathbf{w})) \\
&= \frac{1}{T}\sum_{t=0}^{T-1}(\mathcal{L}_\rho(\mathbf{q}_{t+1}, \mathbf{w}) - \mathcal{L}_\rho(\mathbf{q}^*, \mathbf{w}^*)) \\
&= \frac{1}{T}\left[\sum_{t=0}^{T-1}(\mathcal{L}_\rho(\mathbf{q}_{t+1}, \mathbf{w}) - \mathcal{L}_\rho(\mathbf{q}_{t+1}, \mathbf{w}_{t+1})) \right.\\
&\quad \left. + \sum_{t=0}^{T-1}(\mathcal{L}_\rho(\mathbf{q}_{t+1}, \mathbf{w}_{t+1}) - \mathcal{L}_\rho(\mathbf{q}^*, \mathbf{w}^*))\right] \\
&\leq \frac{1}{T}\left[\frac{(\|\mathbf{w}\| + \|\mathbf{w}_0\|)^2}{2\mu} + \max\left\{0, \frac{2d_1}{\alpha_S}(\beta + (\rho + 2\mu)(\|\mathcal{A}\| + 1)^2)\right\}\right], \qquad (48)
\end{aligned}$$

where the last inequality follows from (44) and (47). Maximizing both sides over $\|\mathbf{w}\| \leq c$, we get

$$h(\bar{\mathbf{q}}_T) - h(\mathbf{q}^*) + c\|\mathcal{K}\bar{\mathbf{q}}_T\| + \frac{\rho}{2}\|\mathcal{K}\bar{\mathbf{q}}_T\|^2 \leq \frac{1}{T}\left[\frac{(c + \|\mathbf{w}_0\|)^2}{2\mu} + \max\left\{0, \frac{2d_1(\beta + (\rho + 2\mu)(\|\mathcal{A}\| + 1)^2)}{\alpha_S}\right\}\right] = \frac{B(\rho, \mu)}{T}.$$

The result now follows from Lemma E.1 by substituting $\delta = B(\rho, \mu)/T$. □

F Implementation Details of Experiments and Additional Results

F.1 Low-rank and sparse covariance estimation

F.1.1 Implementation Details

Our Algorithm:

- Implementing the algorithm, we set $\mathcal{R}_\mathcal{X}$ and $\mathcal{R}_\mathcal{Y}$ to be indicators of the spectrahedron $\mathcal{S}_\tau$ and the ℓ_1-norm ball of radius s, respectively, and we set $\mathcal{A}$ to be the identity operator. Hence in each iteration the main computations of our algorithm are one low-rank SVD and one ℓ_1-norm ball projection. We also set $f(\mathbf{x}) \equiv f(\mathbf{S}) := \frac{1}{2}\|\mathbf{S} - \widehat{\mathbf{\Sigma}}\|_F^2$. In particular, we set the rank estimate for our algorithm, which determines the rank of the SVD computations, as $\hat{r}^* = r$.

- Given the values of ρ and μ, we set the value of the primal step size η to be sent to the WPO according to the formula given in Theorem 5, with respect to the value of μ, and the values of α_S and β_S induced from the value of ρ (notice here the WPO parameter is $\lambda = 1$). Note that for any $\mu > 0$, $\eta \in [0, 1]$.

- For the actual primal step, we used line search. That is, we set:

$$\eta_t = \underset{\eta \in [0,1]}{\arg\min}\{\mu\|\mathcal{K}\mathbf{q}_{t+1}(\eta)\|^2 + \mathcal{L}_\rho(\mathbf{q}_{t+1}(\eta), \mathbf{w}_t)\},$$

  where $\mathbf{q}_{t+1}(\eta) := \eta\mathbf{v}_t + (1-\eta)\mathbf{q}_t$, and $\mathbf{v}_t$ is the output of the WPO. This problem has a closed form solution. It could be easily verified that using this line search does not change the gurantees of Theorem 5.

- We implemented and examined two variants of our algorithm. One which returns the mean of the sequence of primal points generated by Algorithm 1, which we dub "Mean". This is the variant for which we have a theoretically guaranteed convergence rate. The second variant, which we dub "Last", returns the last primal point of the sequence generated by Algorithm 1. This variant seems to be a natural improvement for the Mean variant in practice, in terms of convergence rate.

The Baseline Algorithm: The baseline algorithm for our experiments is the CGAL algorithm introduced in Yurtsever et al. (2019) (Algorithm 1 in the paper). This algorithm solves Problem (OP) if $\mathcal{R}_{\mathcal{X}}$ and $\mathcal{R}_{\mathcal{Y}}$ are indicators of convex sets $\mathcal{X} \subset \mathbb{E}_1$, which is compact, and $\mathcal{Y} \subseteq \mathbb{E}_2$. Yurtsever et al. (2019) suggests two methods for computing the dual step size at every time step — decreasing step sizes (which we dub "decr"), and constant step sizes (which we dub "const"). We present results for both choices. These dual updates also depend on a sequence of dual bounds $\{D_k\}_{k\in\mathbb{N}}$. In accordance with the recommendation in Yurtsever et al. (2019) for best practical performance, we set $D_k = D_{\mathcal{X}}\|\mathcal{A}\|\rho_0, \forall k \in \mathbb{N}$, where $D_{\mathcal{X}}$ is the diameter of $\mathcal{X}$ and ρ_0 is an initial penalty parameter for the augmented Lagrangian (this algorithm uses increasing values of ρ, while we use a constant value). In the implementation of this algorithm we also set $\mathcal{A}$ to be the identity operator, $\mathcal{X}$ to be the spectrahedron, and $\mathcal{Y}$ to be an ℓ_1-norm ball. One iteration of CGAL, thus, requires a single rank-one SVD operation and two projections onto the ℓ_1-norm ball.

SVD implementation and rank overestimation: Both algorithms require low-rank SVD computations. The baseline requires a single rank-one SVD on each iteration, and our algorithm performs a single rank-$\hat{r}^*$ SVD every iteration, where we recall $\hat{r}^*$ is an estimation of r^*, the rank of the optimal solution, which is assumed to be small. We perform these computations for both algorithms using the *scipy.sparse.linalg.eigsh()* built-in function in Python, which gets as an input the rank of the SVD required. For our algorithm, we examined two cases. In the first one, we set $\hat{r}^* = r$. In the second case, we purposely overestimated r^* by a factor of 1.5, meaning we set $\hat{r}^* = 1.5r$. Another parameter of this function is a tolerance parameter, which serves as a stopping condition, which we set to 0.01 for both algorithms, to avoid long running times reaching irrelevant levels of precision.

Manual Tuning of hyper parameters: In our algorithm, given the value of the quadratic penalty parameter ρ, our theory suggests taking values of the dual step size μ which are often highly pessimistic in practice (less than 10^{-3} for this experiment). We thus increase the value of μ beyond its theoretical bound for better practical results and simply set $\mu = 0.2$, which seems to work well. For the tuning of ρ, given the choice $\mu = 0.2$, we started with 1, and kept multiplying/dividing by a factor of 5, until a parameter outperformed its neighbors (the neighbors of 1 are 0.2 and 5, to be clear), over the average of 20 i.i.d. runs. The tuning was done separately for the two variants of our algorithm ("Mean" and "Last"). The tuning of the parameter ρ_0 for the baseline algorithm was done similarly, and was done separately for its two different variants ("decr" and "const"). The values of ρ for our algorithm and ρ_0 for the baseline are presentes in Table 3.

Algorithm \ r	5	10	20
Last (ours)	25	5	1
Mean (ours)	5	5	1
CGAL-const	1	0.2	0.2
CGAL-decr	1	1	1

Table 3: Values of ρ (our algorithm, top two rows) and ρ_0 (baseline algorithm, bottom two rows).

Initialization: For both algorithms we initialized $\mathbf{S}$ to be the projection of $\widehat{\mathbf{\Sigma}}$ onto the spectrahedron of trace τ, and the dual variable to be zero. In addition, for our algorithm, we initialized the additional variable $\mathbf{y}$ to be the projection of $\widehat{\mathbf{\Sigma}}$ onto the ℓ_1 norm ball of radius s.

F.1.2 Results

The complete set of results is given in Figures 3, 4, 5, 6, 7.

F.2 Max Cut

F.2.1 Implementation Details

Our Algorithm We set $\mathcal{R}_{\mathcal{X}}$ to be an indicator of the spectrahedron $\mathcal{S}_\tau$ with $\tau = d$. We set $\mathcal{A}$ to be the identity operator, and set $\mathcal{R}_{\mathcal{Y}}$ to be the indicator function for the set of $d \times d$ matrices with a diagonal of ones. Hence on each iteration, the main computation of our algorithm is a single low-rank SVD. We also set $f(\mathbf{x}) \equiv f(\mathbf{S}) := -\text{tr}(\mathbf{CS})$, for simplicity of computation. As in the CME experiments, we use line search for computing the primal step size on each iteration for this problem as well. In particular, we set the rank estimate for our algorithm, which determines the rank of the SVD computations, as $\hat{r}^* = r^*$.

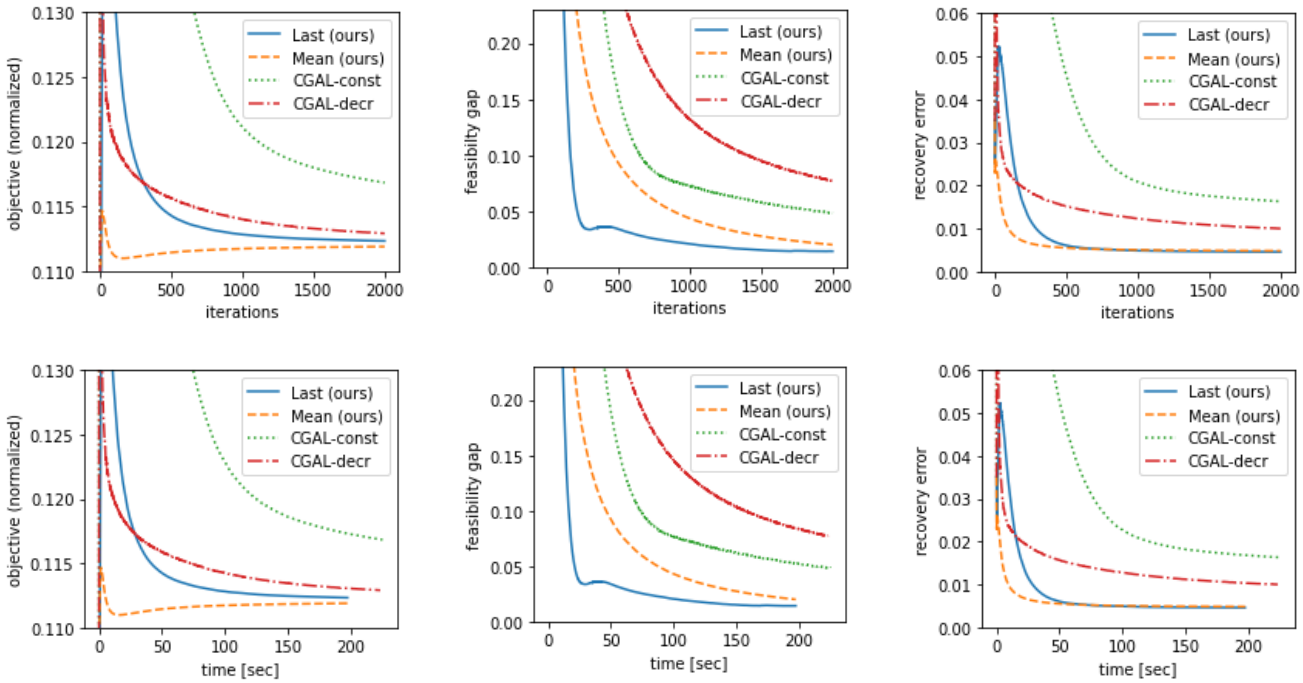

Figure 3: Results for low-rank and sparse covariance estimation for $r = \hat{r}^* = 5$.

The Baseline Algorithm As in the implementation of our algorithm, we set $\mathcal{A}$ to be the identity operator, $\mathcal{X}$ to be the spectrahedron $\mathcal{S}_\tau$ with $\tau = d$, and $\mathcal{Y}$ to be the set of $d \times d$ matrices with diagonal of ones. We used the same objective function as in our algorithm. Hence, for the baseline algorithm, the only expensive computation is a rank-one SVD computation on each iteration.

SVD computations and rank overestimation We used the same Python built-in function for thin SVD computations as in the previous experiment. Here we also set the tolerance parameter of the function to 0.01 for all algorithms. For the G1 graph we also considered overestimating the rank of the optimal solution for our algorithm, taking the SVD rank to be $\hat{r}^* = 20$, where we know $r^* = 13$ is the rank of the optimal solution.

Manual Tuning Since this problem does not satisfy the PQG property, the choice of η becomes heuristic, which adds another free parameter to our algorithm. Here we chose $\mu = 0.2$ and $\eta = 0.2$ arbitrarily, while the choice of $\rho = 1$ for both variants of our algorithm, and $\rho_0 = 1$ for both variants of the baseline was done the same way as in the previous experiment, based on the performance over the graph G1 , where $\hat{r}^* = r^* = 13$ (Figure 8). Here, we used the same parameters for all of four datasets (Figures 8,9,10,11).

Initialization For both algorithms we initialized $\mathbf{S}$ to be the identity matrix, and the dual variable to be zero. In addition, for our algorithm, we initialized the additional variable $\mathbf{y}$ to be the identity matrix as well.

F.2.2 Results

The complete set of results is given in Figures 8, 9, 10, 11.

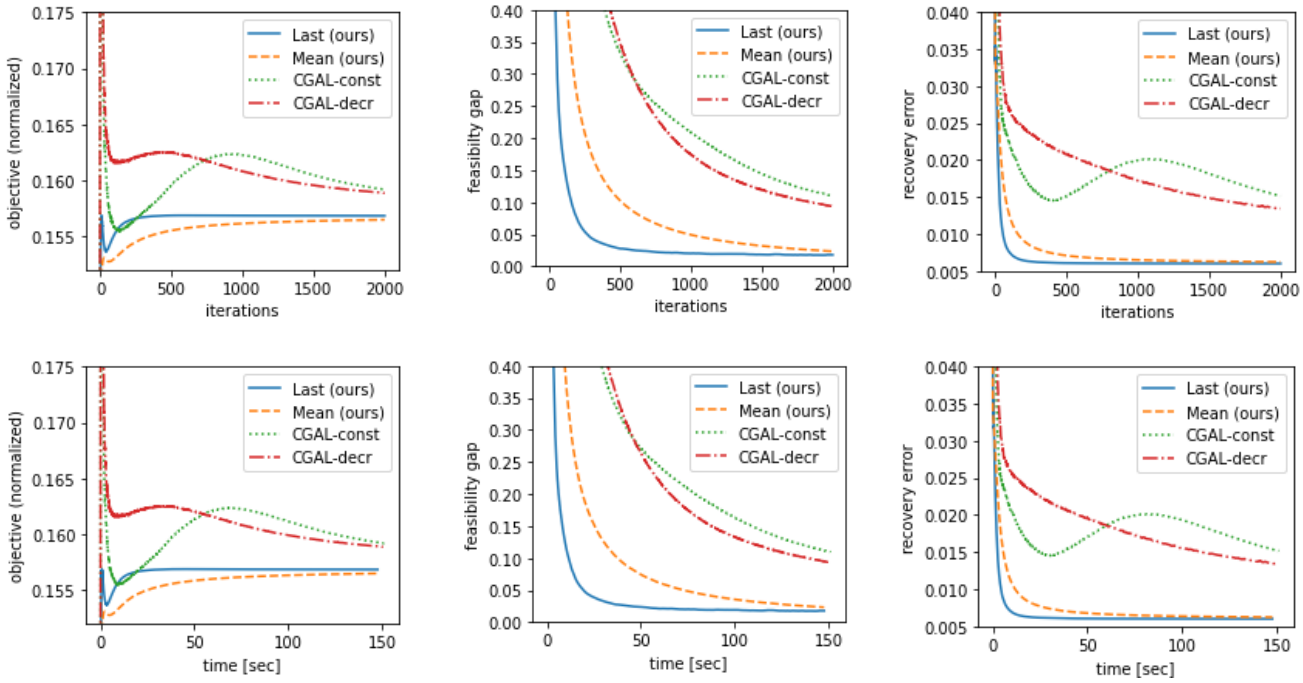

Figure 4: Results for low-rank and sparse covariance estimation for $r = \hat{r}^* = 10$.

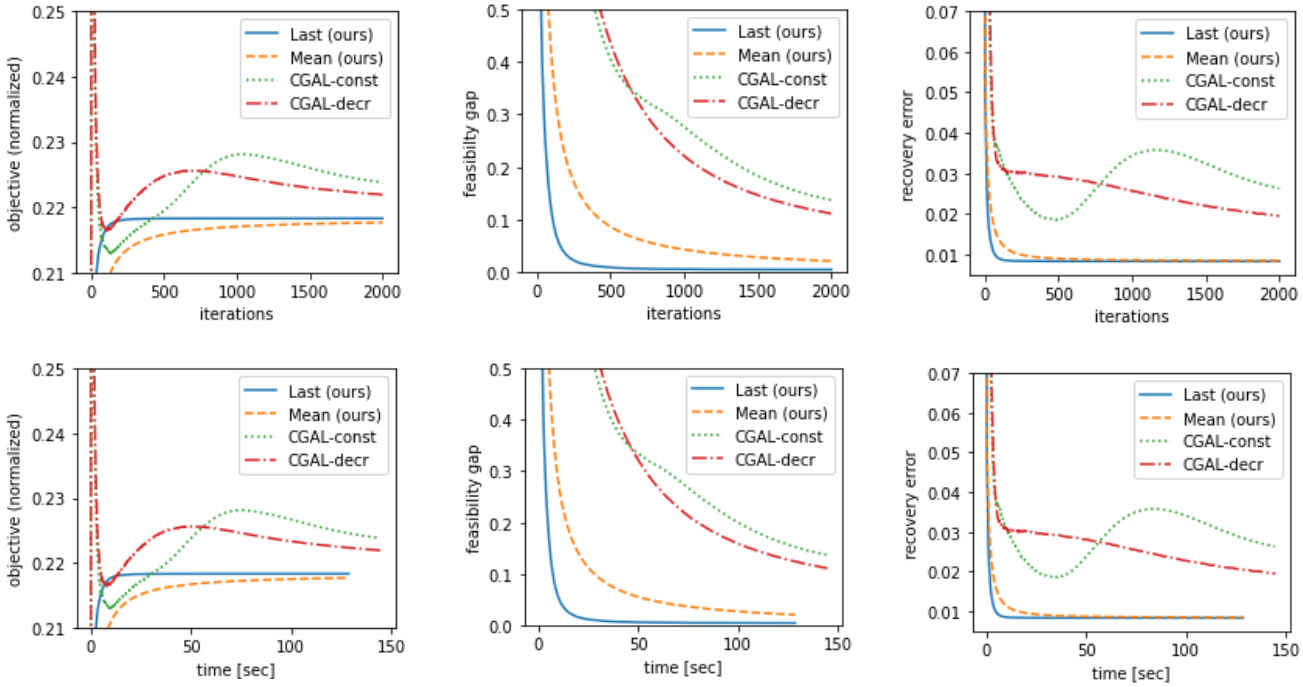

Figure 5: Results for low-rank and sparse covariance estimation for $r = \hat{r}^* = 20$.

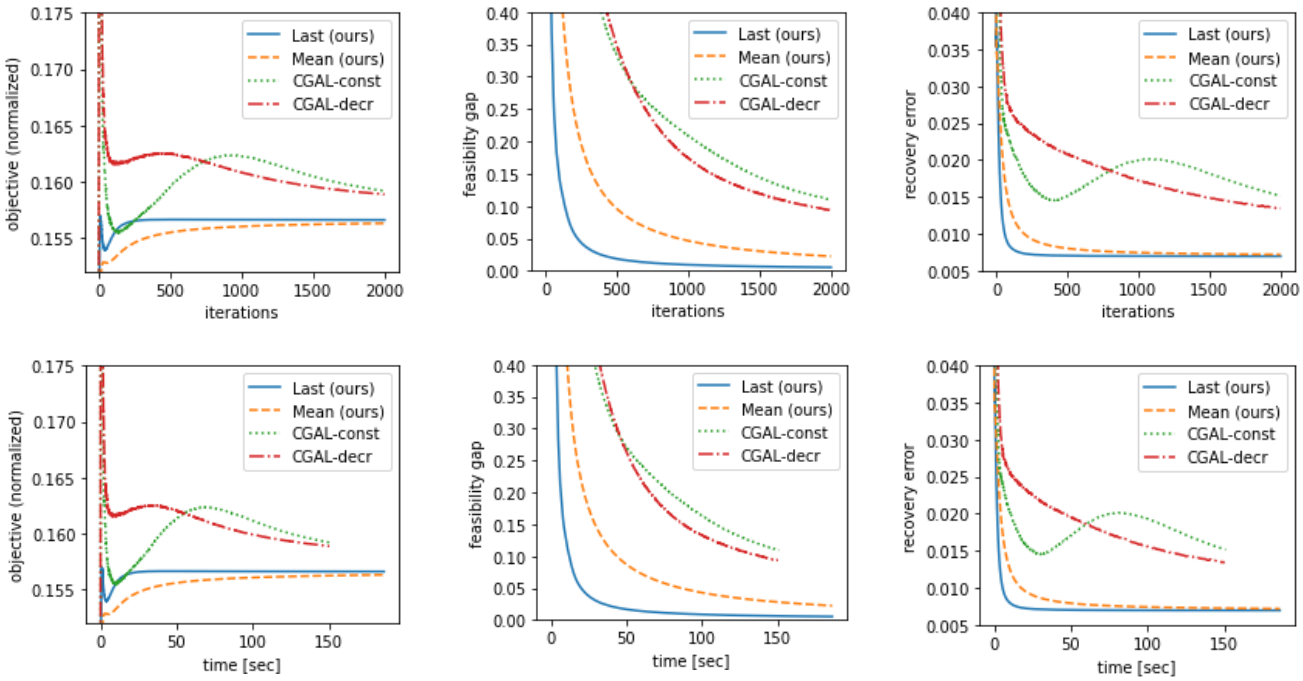

Figure 6: Results for low-rank and sparse covariance estimation for $r = 10$ and $\hat{r}^* = 15$.

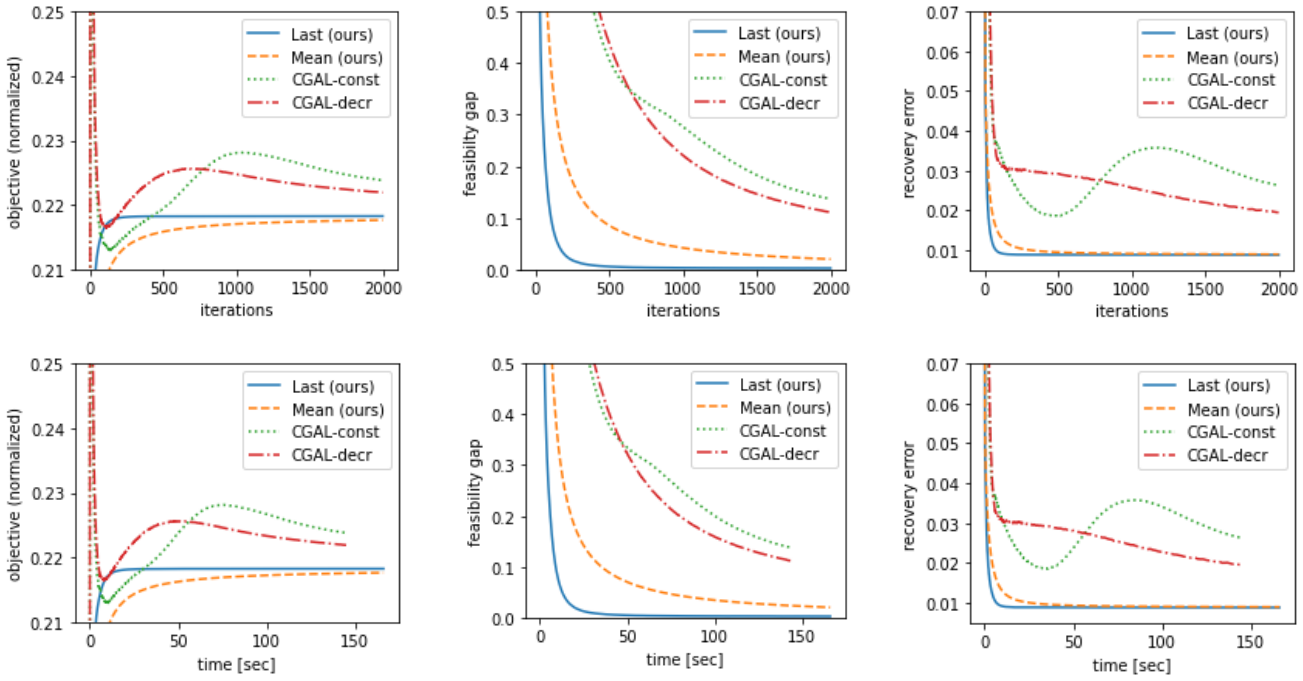

Figure 7: Results for low-rank and sparse covariance estimation for $r = 20$ and $\hat{r}^* = 30$.

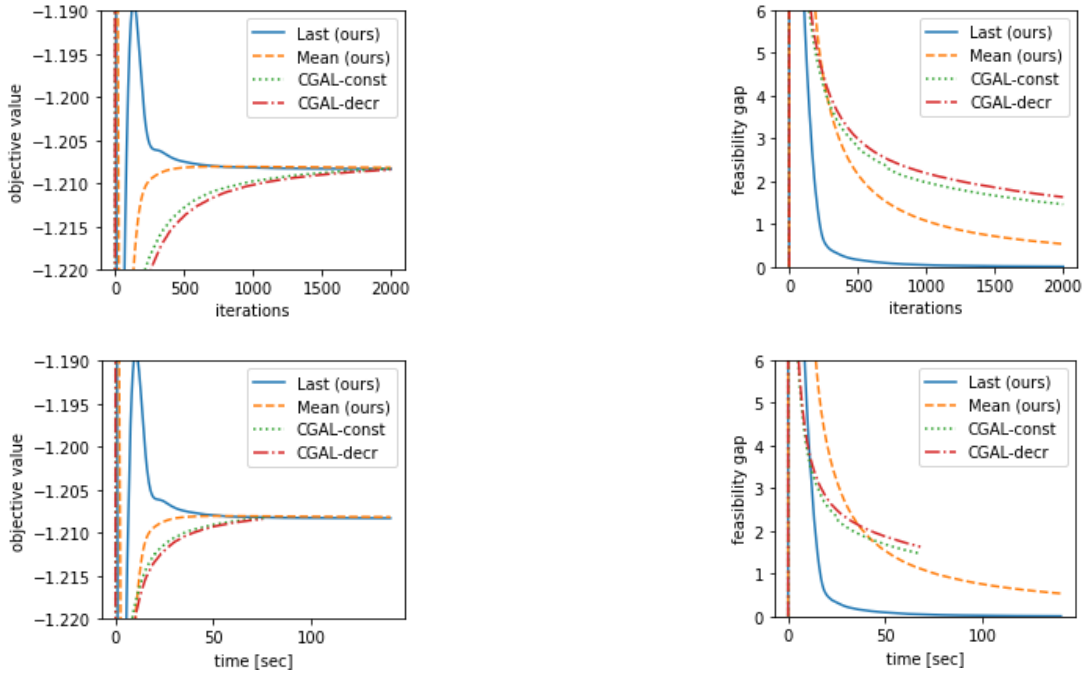

Figure 8: Results for Max Cut with graph G1 ($r^* = \hat{r}^* = 13$).

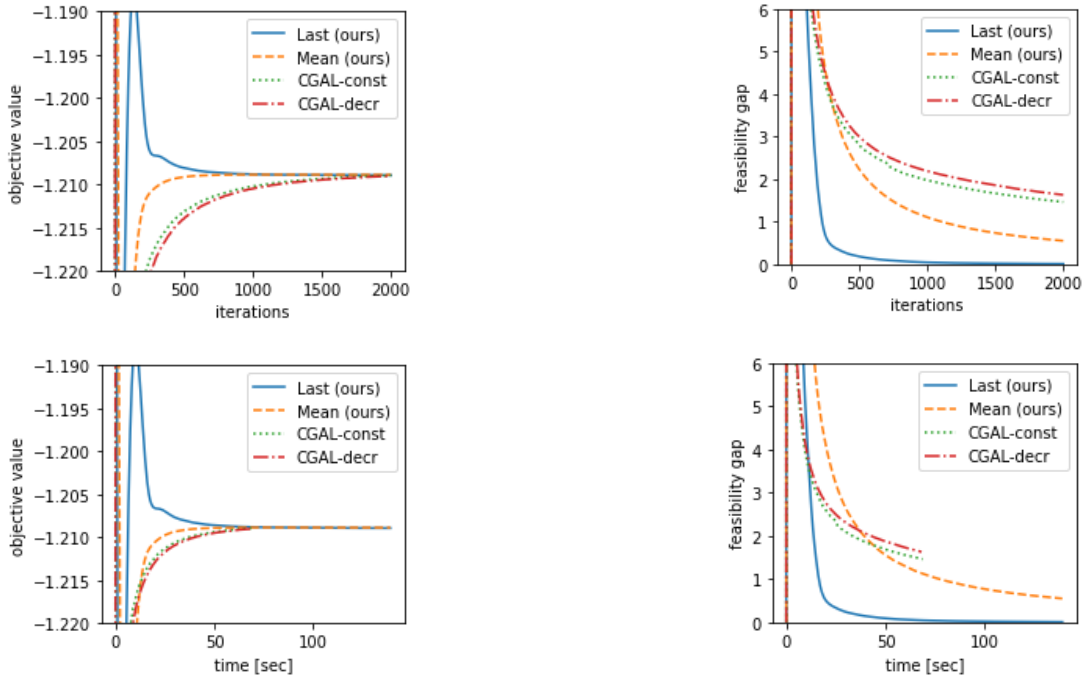

Figure 9: Results for Max Cut with graph G2 ($r^* = \hat{r}^* = 13$).

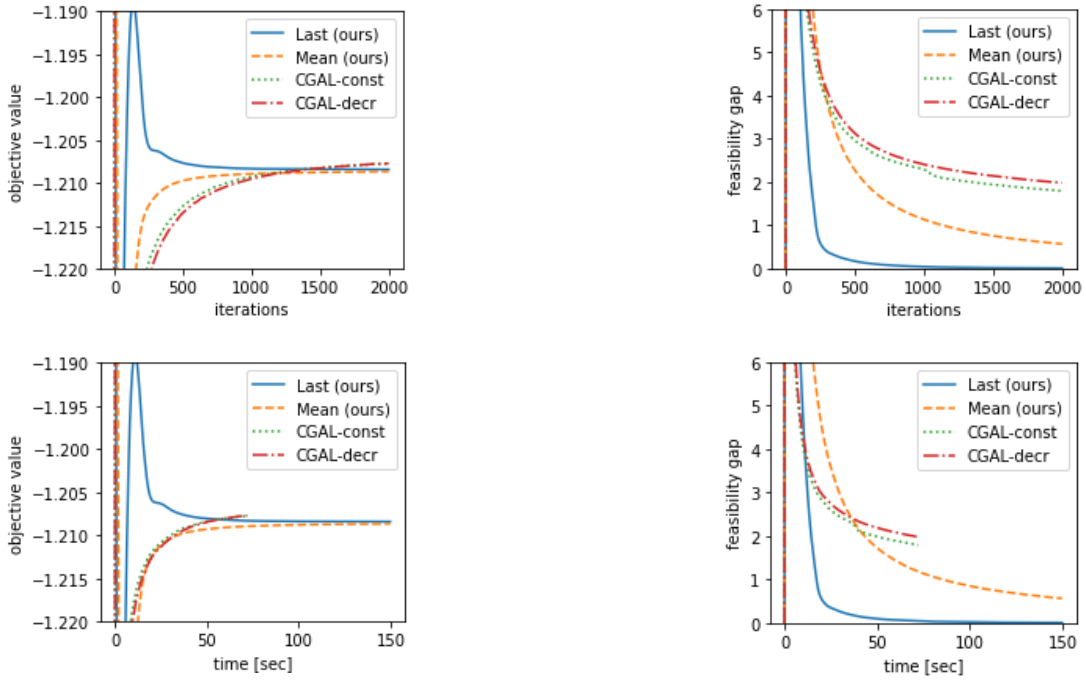

Figure 10: Results for Max Cut with graph G3 ($r^* = \hat{r}^* = 14$).

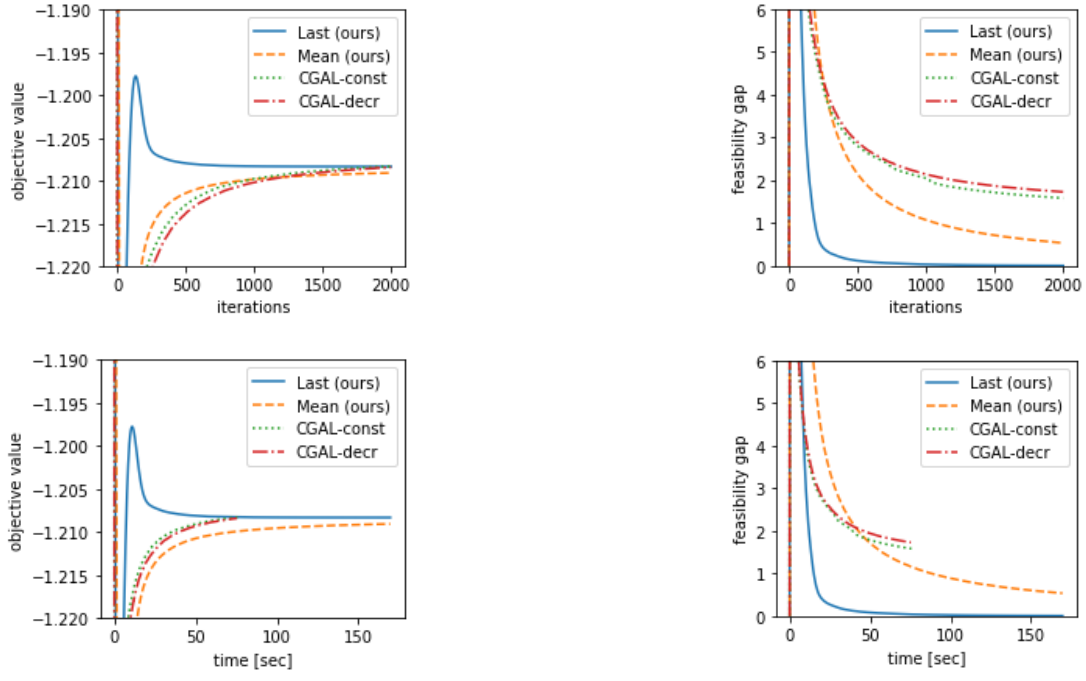

Figure 11: Results for Max Cut with graph G1 ($r^* = 13$) for $\hat{r}^* = 20$.