Spectral Frank-Wolfe Algorithm: Strict Complementarity and Linear Convergence

Lijun Ding [1] **Yingjie Fei** [1] **Qiantong Xu** [2] **Chengrun Yang** [3]

Abstract

We develop a novel variant of the classical Frank-Wolfe algorithm, which we call spectral Frank-Wolfe, for convex optimization over a spectrahedron. The spectral Frank-Wolfe algorithm has a novel ingredient: it computes a few eigenvectors of the gradient and solves a small-scale SDP in each iteration. Such procedure overcomes slow convergence of the classical Frank-Wolfe algorithm due to ignoring eigenvalue coalescence. We demonstrate that strict complementarity of the optimization problem is key to proving linear convergence of various algorithms, such as the spectral Frank-Wolfe algorithm as well as the projected gradient method and its accelerated version.

1. Introduction

We consider solving the following optimization problem with the decision variable $X \in \mathbb{R}^{n\times n}$:

$$\begin{aligned} \text{minimize} \quad & f(X) := g(\mathcal{A}X) + \langle C, X\rangle \\ \text{subject to} \quad & \mathbf{tr}(X) = 1 \quad X \succeq 0. \end{aligned} \tag{1}$$

Problem setup. The setup of Problem (1) is as follows. We assume $C \in \mathbb{R}^{n\times n}$ is a symmetric matrix. The constraint $X \succeq 0$ means that X is symmetric and positive semidefinite. We assume that $\mathcal{A} : \mathbb{S}^n \to \mathbb{R}^m$ is a linear map from the set of symmetric matrices $\mathbb{S}^n$ to the m-dimensional Euclidean space. We also assume that the function $g : \mathbb{R}^m \to \mathbb{R}$ is differentiable and its gradient ∇g is L_g-Lipschitz continuous. We use $\mathbf{tr}(\cdot)$ to denote the standard trace operation, the sum of diagonal entries of the input matrix. We denote by $\mathcal{S}_n$ the feasible region of Problem (1).

[1]School of ORIE, Cornell University, Ithaca, NY 14850, USA [2]Facebook AI Research, Menlo Park, CA 94025, USA [3]School of Electrical and Computer Engineering, Cornell University, Ithaca, NY 14850, USA. Correspondence to: Lijun Ding <ld446@cornell.edu>.

Proceedings of the 37^{th} International Conference on Machine Learning, Online, PMLR 119, 2020.

The set $\mathcal{S}_n$ is called the spectrahedron, which is nonempty and compact. Hence Problem (1) always has an optimal solution. In this paper, we assume Problem (1) admits a unique optimal solution $X_\star$ with rank $r_\star$ for the sake of simplicity. The main results, Theorem 3 and 6 below, can be adapted to the setting where multiple optimal solutions exist; see Section A in the Appendix for a further discussion. It is worth noting that for almost all matrix C, the solution of Problem (1) is indeed unique (Drusvyatskiy & Lewis, 2011, Corollary 3.5).

Applications. The optimization problem covers many low rank matrix recovery problems including matrix sensing (Recht et al., 2010), matrix completion (Candès & Recht, 2009; Jaggi & Sulovskỳ, 2010), phase retrieval (Candes et al., 2015; Yurtsever et al., 2017), and blind deconvolution (Ahmed et al., 2013). The constraints $X \succeq 0$ and $\mathbf{tr}(X) = 1$ impose low-rankness on the solution. The rank $r_\star$ of optimal solutions in these applications is expected to be small comparing to the problem dimension n. We note that the following problem:

$$\text{minimize}_{\|X\|_*\le\alpha} \quad f(X), \tag{2}$$

is sometimes a more direct optimization formulation for aforementioned low rank matrix recovery problems. Since Problem (2) can be re-formulated as Problem (1) (Jaggi & Sulovskỳ, 2010), we consider Problem (1) as our main focus of study in this paper.

Background and related works. A natural but costly algorithm for solving (1) is using the projected gradient descent method (PGD) or its accelerated version (APGD) (Nesterov, 2013). Although the iteration complexity of PGD or APGD is considerably low,[1] each of their iteration requires computing a full eigenvalue decomposition of an $n \times n$ matrix, which scales as $\mathcal{O}(n^3)$ (Trefethen & Bau III, 1997). The high per-iteration cost prevents their large-scale deployment. Hence, projection-free methods are sought, such as the Frank-Wolfe method (FW) (Frank & Wolfe, 1956; Jaggi, 2013) presented in Algorithm 1. In the spectrahedron setting,

[1]PGD or APGD achieves an ϵ-approximate solution in $\mathcal{O}(\log(\frac{1}{\epsilon}))$ iterations for strongly convex f. APGD achieves an ϵ-approximate solution $\mathcal{O}(\frac{1}{\sqrt{\epsilon}})$ for general smooth f.

Algorithm 1 Frank-Wolfe with line search

Input: initialization $X_0 \in \mathcal{S}_n$
for $t = 1, 2, \ldots,$ **do**
  Eigenvalue computation: compute an eigenvector v of $\nabla f(X_t)$ associated with smallest eigenvalue.
  Line search: solve $\hat{\eta} = \arg\min_{\eta \in [0,1]} f(\eta X_t + (1-\eta)vv^\top)$ and set $X_{t+1} = \hat{\eta} X_t + (1-\hat{\eta})vv^\top$.
end for

Algorithm 2 Generalized BlockFW (G-BlockFW)

Input: initialization $X_0 \in \mathcal{S}_n$, a step size $\eta \in [0, 1]$, a smooth parameter β, and an integer $k > 0$
for $t = 1, 2, \ldots,$ **do**
  Eigenvalue computation: compute top k eigenvalues $(\lambda_1, \ldots, \lambda_r)$ and their eigenvectors $V = [v_1, \ldots, v_k]$ of $X_t - \frac{1}{\eta\beta}\nabla f(X_t)$.
  Eigenvalue projection: project $(\lambda_1, \ldots, \lambda_k)$ to the k-dimensional probability simplex $\{x \in \mathbb{R}^k \mid \sum_{i=1}^k x_i = 1, x_i \geq 0\}$, and get the projected point Λ.
  Forming a new iterates: set $X_{t+1} = (1-\eta)X_t + \eta V \mathrm{diag}(\Lambda) V^\top$.
end for

each step only requires computing *one* eigenvector of the gradient of f, which can be efficiently done using the Lanczos method (Kuczyński & Woźniakowski, 1992) by taking advantage of the structure of $\nabla f(X) = \mathcal{A}^*(\nabla g)(\mathcal{A}X) + C$ as well as the sparsity of $\mathcal{A}$ and C. FW converges to an ϵ-approximate solution[2] within $\mathcal{O}(\frac{1}{\epsilon})$ many iterations. However, the iteration complexity $\mathcal{O}(\frac{1}{\epsilon})$ is tight as shown in (Garber, 2016) even if f is strongly convex and no structural assumption is posed on the solution of (1). Considerable recent research effort (Garber, 2016; Freund et al., 2017; Allen-Zhu et al., 2017; Garber, 2019b) has focused on incorporating the low-rankness of solution $X_\star$. Of particular relevance to our work are Garber (2019b) and Allen-Zhu et al. (2017):

- Garber (2019b) shows that Algorithm 1 converges linearly given that the solution is rank *one*, and an eigengap assumption on the gradient $\nabla f(X_\star)$ at the optimal solution is satisfied. We note that the rank-one assumption is crucial for the linear convergence of Algorithm 1 to hold. As we will demonstrate in Section 5, if the solution is not rank one, Algorithm 1 gets stagnant and behaves in the worst case as $\mathcal{O}(\frac{1}{\epsilon})$.

- Allen-Zhu et al. (2017) proposes an algorithm called BlockFW, which is re-formulated as Algorithm 2 for our setting and renamed as generalized BlockFW(G-BlockFW) [3]. It computes only k eigenvectors in each step, and converges linearly so long as $k \geq r_\star = \mathrm{rank}(X_\star)$ and f is strongly convex. However, the method relies critically on the assumption $k \geq r_\star$: no convergence guarantees can be made if this assumption fails. Indeed, we will demonstrate in Section 5 that if $k < r_\star$, G-BlockFW gets stuck at moderate accuracy and cannot make further progress.[4] Moreover, the method needs to store iterates explicitly to compute the eigenvectors. This not only incurs an extra $\mathcal{O}(n^2)$ space complexity, but also increases the burden of computing eigenvectors as the iterates themselves have no structure to be exploited for fast eigenvector computation.[5]

In summary, previous methods converge linearly only when the optimal solution is rank one, or the number of eigenvectors computed in each iteration is no smaller than the rank of the optimal solution.

Our contributions. The contribution of this work is twofold. On the problem structure side:

- We show that the eigengap assumption in (Garber, 2019b) is equivalent to the strict complementarity condition, a well-known regularity condition of semidefinite programming (Alizadeh et al., 1997); see Section 2 for more detail.
- Based on the eigengap condition, or the equivalent strict complementarity condition, we show that Problem (1) satisfies the quadratic growth property (Definition 2 below) when the outer function g is strongly convex over the feasible region $\mathcal{S}_n$ of Problem (1), which is true for all the application being considered. This governs the linear convergence of many first order methods such as PGD, APGD, and our method, Spectral Frank Wolfe.

On the algorithm side, we propose a new algorithm called Spectral Frank-Wolfe (SpecFW) in Section 3, which has the following properties:

- In each of its iteration, it computes k eigenvectors using *only* the current gradient information.
- In each of its iteration, it solves a small-scale subproblem efficiently by APGD for small k.

[2] A matrix X is ϵ-approximate solution to Problem (1) if X is feasible and $f(X) - f(X_\star) \leq \epsilon$.

[3] We note that BlockFW is *not* designed for (1), but rather for (2). Since (2) covers (1), we renamed the algorithm as G-BlockFW.

[4] Allen-Zhu et al. (2017) gives an adaptive k selection procedure which works well in their experiments, but there is no theoretical guarantee for the procedure.

[5] Actually Allen-Zhu et al. (2017) provides a method to avoid the extra space and time costs. However, the method requires knowledge of the strong convexity parameter, which is unavailable in all experiments they perform.

Algorithm	Convergence Rate		
	Worst	Linear	Condition
FW (Alg. 1)	$\frac{8L_f}{t}$	$(1-\frac{\delta}{12L_f})^t$	$r_\star = 1$ and strict comp.
G-BlockFW (Alg. 2)	✗	$(1-\frac{\gamma}{2L_f})^t$	$k \geq r_\star$ and QG
SpecFW (Alg. 3)	$\frac{8L_f}{t}$	$(1-\frac{\min(\delta,\gamma)}{12L_f})^t$	$k \geq r_\star$, QG, and strict comp.

Table 1. Comparision of FW, G-BlockFW and SpecFW. Here, we assume f has gradients ∇f that are L_f-Lipschitz. The optimal solution rank is $r_\star = \text{rank}(X_\star)$. We let t be the number of iterations. Convergence rates are measured by $f(X_t) - f(X_\star)$. We set δ to be the difference between the smallest eigenvalue and the $(r_\star+1)$th-smallest eigenvalue of $\nabla f(X_\star)$, that is, $\delta = \lambda_{n-r_\star}(\nabla f(X_\star)) - \lambda_n(\nabla f(X_\star))$. "Strict comp." means strict complementarity (Definition 1). "QG" means quadratic growth with parameter γ (Definition 2). Both FW and SpecFW have burn-in phases which are bounded by $\frac{72L_f^3}{(\min\{\gamma,\delta\})^3}$. Here, the burn-in phase is the number of iterations in which the method converges with standard rate L_f/t, before shifting to the faster rate (if linear convergence condition is satisfied). The convergence rate of G-BlockFW can be found in Lemma 8 in Section F of the Appendix.

- It always converges at the rate $\mathcal{O}(\frac{1}{\epsilon})$ no matter what choice of k is.
- It converges linearly when $k \geq r_\star$, and the strict complementarity and quadratic growth condition are satisfied. In particular, we do *not* require f to be strongly convex or the rank $r_\star$ to be 1.
- It can easily incorporate the matrix sketching idea from Tropp et al. (2017) and achieves the so-called storage optimality discussed in Yurtsever et al. (2017). The sketching procedure obviates the need for storing the full decision matrix X throughout iterations, thereby saving $\mathcal{O}(n^2)$ space.[6]

Organization. The rest of the paper is organized as follows. In Section 2, we explain the concept of strict complementarity and the classical Frank-Wolfe algorithm, and how they motivate our Spectral Frank-Wolfe. In Section 3, we present the Spectral Frank-Wolfe and its convergence guarantees. In Section 4, we show that the strict complementarity enforces the quadratic growth condition whenever g is strongly convex on $\mathcal{S}_n$. Finally, we demonstrate numerically the effectiveness of the Spectral Frank-Wolfe in Section 5.

Notation. For a symmetric matrix $A \in \mathbb{S}^n$, we denote its i-th largest eigenvalue as $\lambda_i(A)$. The operator two norm, nuclear norm, and Frobenius norm are denoted as $\|A\|_{\text{op}}$, $\|A\|_*$, and $\|A\|_{\text{F}}$, respectively. The inner product $\langle\cdot,\cdot\rangle$ on symmetric matrices is the standard trace inner product. We also equip $\mathbb{R}^m$ with the dot product. For a linear map $\mathcal{B}: \mathbb{S}^d \to \mathbb{R}^l$, the adjoint map of $\mathcal{B}$ is denoted as $\mathcal{B}^*$. We also define its largest and smallest singular values as $\|\mathcal{B}\|_{\text{op}} = \sigma_{\max}(\mathcal{B}) = \max_{\|A\|_{\text{F}}=1} \|\mathcal{B}(A)\|_2$ and $\sigma_{\min}(\mathcal{B}) = \min_{\|A\|_{\text{F}}=1} \|\mathcal{B}(A)\|_2$. Given a matrix $V \in \mathbb{R}^{d\times r}$, we denote the restriction of $\mathcal{B}$ to V as $\mathcal{B}_V : \mathbb{S}^r \to \mathbb{R}^l$ by $\mathcal{B}_V(S) = \mathcal{B}(VSV^\top)$ for any $S \in \mathbb{S}^r$.

[6] Interested readers can find the procedure in Section D in the Appendix. We note the matrix sketching idea cannot be combined with G-BlockFW easily to avoid storing X, as G-BlockFW uses a sum of the current iterate and current gradient to compute the eigenvectors, which destroys the fast matrix-vector product property of the gradient.

2. Motivating SpecFW from complementarity and Frank-Wolfe

In this section, we explain the motivations of the spectral Frank-Wolfe from strict complementarity and its relationship with the classical Frank-Wolfe.

2.1. Observation from complementarity

Let first introduce the KKT condition to see what complementarity means.

KKT condition. By Slater's condition for (1) and the fact that the feasible region $\mathcal{S}_n$ is compact, the following KKT condition of (1) always holds: there is some dual optimal solution $Z_\star \succeq 0$ and $s_\star \in \mathbb{R}$ such that[7]

$$\begin{aligned} \nabla f(X_\star) - Z_\star - s_\star I = 0, &\quad \text{(First Order Condition)} \\ \langle Z_\star, X_\star\rangle = 0, &\quad \text{(Complementarity)} \\ \mathbf{tr}(X_\star) = 1, &\quad \text{(Linear Constraint Feasibility)} \\ Z_\star, X_\star \succeq 0. &\quad \text{(PSD Feasibility)} \end{aligned} \tag{3}$$

Here I is the identity matrix in $\mathbb{S}^n$. We prove in Lemma 1 in the Appendix that the dual solution $(Z_\star, s_\star)$ is actually unique.

Complementarity: extract $X_\star$ from $Z_\star$. we first note that using $Z_\star, X_\star \succeq 0$ and complementarity $\langle Z_\star, X_\star\rangle = 0$, we have $Z_\star X_\star = 0$. This equality implies that

$$\text{range}(X_\star) \subset \text{nullspace}(Z_\star), \tag{4}$$

[7] If there are multiple primal optimal solutions, then the KKT condition holds for any one of them.

and

$$r_\star = \text{rank}(X_\star) \le \dim(\text{nullspace}(Z_\star)) =: k_\star. \quad (5)$$

Hence, if we can compute a matrix $V_\star \in \mathbb{R}^{n\times k_\star}$ with orthonormal columns that span the null space of $Z_\star$, and solve for

$$S_\star = \arg\min_{S\in\mathcal{S}_{k_\star}} f(V_\star S V_\star^\top), \quad (6)$$

then we get the primal optimal solution $X_\star = V_\star S_\star V_\star^\top$.

We note that it is necessary to optimize over the $k_\star$-spectrahedron $\mathcal{S}_{k_\star}$ instead of just a $k_\star$-dimensional probability simplex, as $V_\star$ may not be the eigenvectors of $X_\star$ for $k_\star > 1$. Problem (6) can be solved by APGD rapidly so long as $k_\star$, the *size* of S, is small.

This naturally leads to the following questions:

1. Problem (6) is easy to solve only if $k_\star$ is small; yet for now we only have $k_\star \ge r_\star$. With $r_\star$ expected to be small, can we hope for $k_\star = r_\star$ to hold, so that $k_\star$ is small as well?
2. Suppose we have $k_\star = r_\star$, can we compute $V_\star$ exactly or approximate it well enough?

We answer the first question in the next section by defining strict complementarity and establishing its equivalence to an eigengap condition on $\nabla f(X_\star)$. To answer the second question, we draw relationship between the first order condition in (3) and the classical Frank-Wolfe algorithm in Section 2.3.

2.2. Strict complementarity

We answer why we expect $r_\star = k_\star$ in this section. Using the rank-nullity theorem, we see that the equation $r_\star = \text{rank}(X_\star) \le \dim(\text{nullspace}(Z_\star))$ is equivalent to

$$\text{rank}(X_\star) + \text{rank}(Z_\star) \le n.$$

Strict complementarity (Alizadeh et al., 1997) assumes that we have equality instead of inequality.

Definition 1. *(Strict Complementarity) Let $X_\star, Z_\star$ and $s_\star$ satisfy the KKT condition (3). We say that Problem (1) (or the pair $(X_\star, Z_\star)$) satisfies strict complementarity if*

$$\textit{rank}(X_\star) + \textit{rank}(Z_\star) = n.$$

It is immediately clear that using the rank-nullity theorem again, we see that strict complementarity is equivalent to

$$r_\star = k_\star,$$

which is what we desire. By (4) and given that the solution rank is $r_\star$, strict complementarity is equivalent to

$$\lambda_{n-r_\star}(Z_\star) > 0. \quad (7)$$

Equation (4) also implies that we always have for all $i = 1, \dots, r_\star$,

$$\lambda_{n-r_\star+i}(Z_\star) = 0. \quad (8)$$

Relation with the eigengap assumption. In Garber (2019a;b), the author proposed an eigengap condition:

$$\lambda_{n-r_\star}(\nabla f(X_\star)) - \lambda_n(\nabla f(X_\star)) > 0.$$

This is in fact equivalent to strict complementarity: since $\nabla f(X_\star) = Z_\star + s_\star I$, we have

$$\begin{aligned}&\lambda_{n-r_\star}(\nabla f(X_\star)) - \lambda_n(\nabla f(X_\star))\\ =&\lambda_{n-r_\star}(Z_\star + s_\star I) - \lambda_n(Z_\star + s_\star I)\\ =&\lambda_{n-r_\star}(Z_\star) + s_\star - \lambda_n(Z_\star) - s_\star\\ =&\lambda_{n-r_\star}(Z_\star),\end{aligned}$$

where the last step is due to (8). Using (7), we deduce the equivalence.

Why strict complementarity should hold. Strict complementarity as shown in Drusvyatskiy & Lewis (2011) holds for almost all C (see Lemma 2 for a more detailed derivation). We will also verify this assumption numerically in our experiments in Section 5. Moreover, as demonstrated in Garber (2019b, Lemmas 2 and 10), such assumption should hold if we expect the solution rank $r_\star$ to be stable under small perturbations.

2.3. FW and approximation of nullspace$(Z_\star)$

We have just argued why we expect $r_\star = k_\star$ should hold for Problem (1). In this section, we draw relation of FW and approximation of nullspace$(Z_\star)$.

Denote by $\mathbf{EV}_r(A)$ the eigenspace of the smallest r eigenvalues of a matrix $A \in \mathbb{S}^n$. In view of the first order condition (3), we have

$$\mathbf{EV}_{k_\star}(\nabla f(X_\star)) = \text{nullspace}(Z_\star). \quad (9)$$

Hence nullspace$(Z_\star)$ can be identified using the gradient of f at $X_\star$.

Note that FW indeed uses the eigenvector corresponding to the smallest eigenvalue of $\nabla f(X_t)$ in each of its iteration, and therefore it tries to approximate $\mathbf{EV}_{k_\star}(\nabla f(X_\star))$. This is the main intuition that linear convergence of FW can be established when $r_\star = 1$ as in Garber (2019b). It also reveals that FW fails to converge in a linear rate for $k_\star > 1$, as approximation using one eigenvector is not enough for a $k_\star$-dimensional space. Also, from (8) and the first order condition in the KKT condition, we see the smallest $k_\star$ eigenvalues of the gradient coalesce, and hence it is important to compute the $k_\star$-dimensional space to attain

Algorithm 3 Spectral Frank-Wolfe

Input: initialization $X_0 \in \mathcal{S}_n$, an integer $k > 0$
for $t = 1, 2, \ldots,$ **do**
  Eigenvalue computation: compute the k eigenvectors, $v_1, \ldots, v_k$ of $\nabla f(X_t)$ associated with the k smallest eigenvalues, and form the matrix $V = [v_1, \ldots, v_k] \in \mathbb{R}^{n\times k}$.
  Solving a small-scale SDP: solve $\min_{\eta+\mathbf{tr}(S)=1, S\succeq 0, \eta\geq 0} f(\eta X_t + VSV^\top)$ and get an optimal solution $(\hat{S}, \hat{\eta})$.
  Forming a new iterate: set $X_{t+1} = \hat{\eta} X_t + V\hat{S}V^\top$.
end for

better numerical stability and accuracy. Hence, to overcome this issue, we need to compute at least $k_\star$ eigenvectors and solve a sub-problem like (6) in each iteration.

The above discussion motivates our algorithm, the Spectral Frank-Wolfe (Algorithm 3), described in the next section.

3. Spectral Frank-Wolfe and its Convergence guarantees

In this section, we describe the Spectral Frank-Wolfe algorithm and its theoretical guarantees.

3.1. The Spectral Frank-Wolfe algorithm

The Spectral Frank-Wolfe algorithm is presented in Algorithm 3. We highlight its key mechanism as follows.

Solving a small-scale SDP. The small-scale semidefinite programming (SDP)

$$\min_{\eta+\mathbf{tr}(S)=1, S\succeq 0, \eta\geq 0} f(\eta X_t + VSV^\top). \tag{10}$$

can be solved easily using APGD since projection to the set $\{(\eta, S) \mid \eta + \mathbf{tr}(S) = 1, S \succeq 0, \eta \geq 0\}$ only requires an eigenvalue decomposition of a symmetric matrix of size k and a projection to the $(k+1)$-dimensional probability simplex. The correctness of the procedure for projection can be verified using arguments in Allen-Zhu et al. (2017, Lemma 3.1), and Garber (2019a, Lemma 6). We note that when evaluating gradient is very expensive, instead of minimizing $f(\eta X_t + VSV^\top)$, one can also minimize an upper bound of it (and the guarantees in the next section continue to hold). This is discussed in Section C in the Appendix.

Averaging with current X_t. In addition to the eigenvectors from the current gradient, we also utilize the information of previous iterates when solving the small-scale SDP (10). This follows the same spirit as the classical Frank-Wolfe, which performs a line search over the current iterate and the new atom $vv^\top$. This averaging scheme stabilizes the algorithm and facilitates the $\mathcal{O}(\frac{1}{\epsilon})$ convergence rate.

The choice of k. From the proof of the convergence in the next section, it can be observed that so long as $k \geq k_\star$, Algorithm 3 converges linearly. Of course, one may not know $k_\star$ in advance. In this case, k may be taken as the largest value subject to the user's computational budget or the largest rank of the solution the user can afford in terms of storage. An adaptive strategy may also be employed based on the progress of objetive value decay as in Allen-Zhu et al. (2017, Section 6.2). We do not further the discussion of this issue due to the space limit.

3.2. Theoretical guarantees

To state our result, we first define the notion of quadratic growth.

Definition 2 (Quadratic Growth (QG))**.** *We say that the optimization problem (1) satisfies quadratic growth with parameter $\gamma > 0$, if for every feasible $X \in \mathcal{S}_n$ there holds*

$$f(X) - f(X_\star) \geq \gamma \|X - X_\star\|_F^2.$$

The quadratic growth condition is necessary for linear convergence of gradient descent type methods as shown in Necoara et al. (2019, Theorem 13). Hence we should expect it to hold if we are to show linear convergence of Frank-Wolfe methods. The condition automatically holds for strongly convex f, and more broadly, it is satisfied for almost all C so long as g is semi-algebraic, as shown in Drusvyatskiy et al. (2016, Corollary 4.8). In Section 4, we show that strict complementarity and strong convexity of the outer function g (but not f) implies quadratic growth, as well as an explicit formula of γ in terms of the solution $X_\star$, the map $\mathcal{A}$, and smoothness and strong convexity parameters of g.

We now state the theoretical guarantees for our Algorithm 3.

Theorem 3. *Suppose strict complementarity holds for Problem (1), the optimal solution $X_\star$ is unique with rank $r_\star$, the function g has L_g-Lipschitz continuous gradients, Problem (1) satisfies quadratic growth with parameter γ, and the choice of k satisfies $k \geq r_\star = k_\star$. Define $h_t = f(X_t) - f(X_\star)$ for each t, and $\beta = \|\mathcal{A}\|_{op}^2 L_g$. Then for all t, we have*

$$f(X_t) - f(X_\star) \leq \frac{8\beta}{t}. \tag{11}$$

For all $t \geq T_0 = \frac{72\beta^3}{\gamma\lambda_{n-r_\star}^2(Z_\star)}$, we have

$$h_{t+1} \leq \left(1 - \min\left\{\frac{\gamma}{4\beta}, \frac{\lambda_{n-r_\star}(Z_\star)}{12\beta}\right\}\right) h_t. \tag{12}$$

Discussion on the assumptions. As discussed before, these assumptions are expected to be necessary for linear convergence and robustness of the rank under small perturbations. The assumption of the unique optimal solution is only for the purpose of clear presentation.

Preparation of the proof. Let us first give the definition of the r-th spectral set.

Definition 4. *For each $X \in \mathbb{S}^n$, let $V_X \in \mathbb{R}^{n\times k}$ having orthonormal eigenvectors as columns corresponding to the smallest k eigenvalues of X. Define the spectral k-th set $\mathcal{C}_k(X)$ of X as*

$$\mathcal{C}_k(X) := \left\{ V_X S V_X^\top \in \mathbb{S}^n \mid S \in \mathcal{S}_k \right\}.$$

We next present the following important lemma which is proved in Section E in the Appendix.

Lemma 5. *Given $Y \in \mathbb{S}^n$ which satisfies $\lambda_{n-r}(Y) - \lambda_{n-r+1}(Y) \geq \delta$ for some $\delta > 0$, then for any $X \in \mathbb{S}^n$, $X \succeq 0$, and $\mathbf{tr}(X) = 1$, there is some $W \in \mathcal{C}_r(Y)$ such that*

$$\langle X - W, Y\rangle \geq \frac{\delta}{2}\|X - W\|_F^2.$$

We are now ready to start the proof.

Proof of Theorem 3. Using the Lipschitz smoothness of f, we have for any $t \geq 1$, $\eta \in [0,1]$, and any $W \in \mathcal{C}_{r_\star}(\nabla f(X_t))$:

$$\begin{aligned} f(X_{t+1}) \leq & f(X_t) + (1-\eta)\langle W - X_t, \nabla f(X_t)\rangle \\ & + \frac{(1-\eta)^2\beta}{2}\|W - X_t\|_{\mathrm{F}}^2. \end{aligned} \tag{13}$$

Now choose $W = v_n v_n^\top$ where v_n is the eigenvector of $\nabla f(X_t)$ with the smallest eigenvalue, we can then perform the analysis as normal Frank-Wolfe as is done in (Jaggi, 2013) to reach the first part of the theorem, the inequality (11).

For the second part, we first note that by the discussion after the Definition 1 of strict complementarity, we have $\lambda_{n-r_\star}(\nabla f(X_\star)) - \lambda_{n-r_\star+1}(\nabla f(X_\star)) = \lambda_{n-r_\star}(Z_\star)$, and $\lambda_{n-r_\star+1}(\nabla f(X_\star)) = \cdots = \lambda_n(\nabla f(X_\star))$.

Using Lipschitz continuous gradient of f in step (a), the quadratic growth of f in step (b), and the choice of T_0 in step (c), we find that for all $t \geq T_0$,

$$\begin{aligned} \|\nabla f(X_t) - \nabla f(X_\star)\|_{\mathrm{F}} &\overset{(a)}{\leq} \beta\|X_t - X_\star\|_{\mathrm{F}} \\ &\overset{(b)}{\leq} \beta\left(\frac{f(X_t) - f(X_\star)}{\gamma}\right)^{\frac{1}{2}} \\ &\overset{(c)}{\leq} \frac{1}{3}\lambda_{n-r_\star}(Z_\star). \end{aligned} \tag{14}$$

Using the inequality (14) and Weyl's inequality, we find that

$$\begin{aligned} &\lambda_{n-r_\star}(\nabla f(X_t)) - \lambda_{n-r_\star+1}(\nabla f(X_t)) \\ =& \underbrace{\lambda_{n-r_\star}(\nabla f(X_\star)) - \lambda_{n-r_\star+1}(\nabla f(X_\star))}_{=\lambda_{n-r_\star}(Z_\star)} \\ &+ \underbrace{(\lambda_{n-r_\star}(\nabla f(X_t)) - \lambda_{n-r_\star}(\nabla f(X_\star)))}_{\geq -\frac{1}{3}\lambda_{n-r_\star}(Z_\star)} \\ &+ \underbrace{(\lambda_{n-r_\star+1}(\nabla f(X_\star)) - \lambda_{n-r_\star+1}(\nabla f(X_t)))}_{\geq -\frac{1}{3}\lambda_{n-r_\star}(Z_\star)} \\ \geq& \frac{1}{3}\lambda_{n-r_\star}(Z_\star). \end{aligned}$$

Now we subtract the inequality (13) both sides by $f(X_\star)$, and denote $h_t = f(X_t) - f(X_\star)$ for each t, we reach

$$\begin{aligned} h_{t+1} \leq & h_t + (1-\eta)\underbrace{\langle W - X_t, \nabla f(X_t)\rangle}_{R_1} \\ & + \frac{(1-\eta)^2\beta}{2}\underbrace{\|W - X_t\|_{\mathrm{F}}}_{R_2}. \end{aligned} \tag{15}$$

Using Lemma 5 and the inequality (15), we can choose $W \in \mathcal{C}_{r_\star}(\nabla f(X_t))$ such that

$$\langle W - X_\star, \nabla f(X_t)\rangle \leq -\frac{\lambda_{n-r_\star}(Z_\star)}{6}\|X_\star - W\|_{\mathrm{F}}^2. \tag{16}$$

Let us now analyze the term $R_1 = \langle W - X_t, \nabla f(X_t)\rangle$ using (16) and convexity of f:

$$\begin{aligned} R_1 =& \langle W - X_t, \nabla f(X_t)\rangle \\ =& \langle W - X_\star, \nabla f(X_t)\rangle + \langle X_\star - X_t, \nabla f(X_t)\rangle \\ \leq& -\frac{\lambda_{n-r_\star}(Z_\star)}{6}\|X_\star - W\|_{\mathrm{F}}^2 - h_t. \end{aligned}$$

The term $R_2 = \|X_t - W\|_{\mathrm{F}}^2$ can be bounded by

$$\begin{aligned} R_2 = \|X_t - W\|_{\mathrm{F}}^2 &\overset{(a)}{\leq} 2\left(\|X_t - X_\star\|_{\mathrm{F}}^2 + \|X_\star - W\|_{\mathrm{F}}^2\right) \\ &\overset{(b)}{\leq} \frac{2}{\gamma}h_t + 2\|X_\star - W\|_{\mathrm{F}}^2, \end{aligned}$$

where we use triangle inequality and the basic inequality $(a+b)^2 \leq 2a^2 + 2b^2$ in step (a), and the quadratic growth condition in step (b).

Now combining (15), and the bounds of R_1 and R_2, we reach that there is a $W \in \mathcal{C}_{r_\star}(\nabla f(X_t))$ such that for any

$\xi = 1 - \eta \in [0, 1]$, we have

$$\begin{aligned} h_{t+1} \leq & h_t + \xi \left(-\frac{\lambda_{n-r_\star}(Z_\star)}{6} \|X_\star - W\|_{\mathrm{F}}^2 - h_t \right) \\ & + \frac{\xi^2 \beta}{2} \left(\frac{2}{\gamma} h_t + 2\|X_\star - W\|_{\mathrm{F}}^2 \right) \\ = & \left(1 - \xi + \frac{\xi^2 \beta}{\gamma} \right) h_t \\ & + \left(\xi^2 \beta - \frac{\xi \lambda_{n-r_\star}(Z_\star)}{6} \right) \|X_\star - W\|_{\mathrm{F}}^2. \end{aligned}$$

A detailed calculation and choice of ξ in Section E in Appendix reveals that we can reach the second part of the theorem, the inequality (12). □

4. Quadratic Growth and Linear Convergence of Algorithms

In this section, we show that when g is α-strongly convex (Nesterov, 2013) and strict complementarity of (1) holds, then we have quadratic growth of Problem (1). We also demonstrate when the dual matrix $Z_\star$ has rank $n-1$ then we do not require g to be α-strongly convex. An immediate consequence is the linear convergence of PGD and APGD (Karimi et al., 2016), the generalized blockFW[8] (Algorithm 2, and the spectral Frank-Wolfe (Algorithm 3) as shown in Theorem 3.

Theorem 6. *Suppose strict complementarity of (1) and one of the following conditions hold:*

(i) g is α-strongly convex, and the solution $X_\star$ is unique, or

(ii) the dual matrix $Z_\star$ in the KKT condition (3) has rank $n-1$,

then Problem (1) satisfies quadratic growth. The constant γ takes the form of

(i) $\gamma = \min\left\{ \frac{\lambda_{n-r_\star}(Z_\star)}{4+8\frac{\sigma_{\max}^2(\tilde{\mathcal{A}})}{\sigma_{\min}^2(\tilde{\mathcal{A}}_V)}}, \frac{\alpha \sigma_{\min}^2(\tilde{\mathcal{A}}_V)}{8} \right\}$ in the first case, where $\tilde{\mathcal{A}}(X) = \begin{bmatrix} \mathbf{tr}(X) \\ \mathcal{A}(X) \end{bmatrix}$, and

(ii) $\gamma = \frac{\lambda_{n-r_\star}(Z_\star)}{2}$ in the second case. In addition, the uniqueness of $X_\star$ is implied in the second case.

Proof. The second case has been verified in Garber (2019a, Lemmas 1 and 2). We provide a self-contained and different proof in Section F in Appendix.

Now consider the first case. For any feasible X and the optimal solution $X_\star$, we have

$$\begin{aligned} & f(X) - f(X_\star) \\ = & g(\mathcal{A}X) - g(\mathcal{A}X_\star) + \langle C, X - X_\star \rangle \\ \overset{(a)}{\geq} & \langle (\nabla g)(\mathcal{A}X_\star), \mathcal{A}(X - X_\star) \rangle \\ & + \langle C, X - X_\star \rangle + \frac{\alpha}{2} \|\mathcal{A}X - \mathcal{A}X_\star\|_2^2 \\ \overset{(b)}{=} & \langle \mathcal{A}^*(\nabla g)(\mathcal{A}X_\star) + C, X - X_\star \rangle \\ & + \frac{\alpha}{2} \|\mathcal{A}X - \mathcal{A}X_\star\|_2^2 \\ \overset{(c)}{=} & \langle Z_\star + s_\star I, X - X_\star \rangle + \frac{\alpha}{2} \|\mathcal{A}(X - X_\star)\|_2^2 \\ \overset{(d)}{=} & \langle Z_\star, X \rangle + \frac{\alpha}{2} \|\mathcal{A}(X - X_\star)\|_2^2 \overset{(e)}{\geq} 0 \end{aligned} \tag{17}$$

Here step (a) is due to the strong convexity of g. Step (b) is because of the definition of $\mathcal{A}^*$. For step (c), we uses the first order condition of KKT condition (3) in terms of g and $\mathcal{A}$: $\mathcal{A}^*(\nabla g)(\mathcal{A}X_\star) + C - Z_\star - s_\star I = 0$. The step (d) is due to the complementarity in KKT condition (3) and feasibility of X and $X_\star$. The last inequality (e) is beacause $Z, X \succeq 0$.

We claim that a feasible matrix $X \in \mathbb{S}^n$ is optimal if and only if X satisfies

$$\begin{aligned} \langle Z_\star, X \rangle = 0, \quad & \mathcal{A}X - \mathcal{A}X_\star = 0, \\ \mathbf{tr}(X) = 1, \quad \text{and} \quad & X \succeq 0. \end{aligned} \tag{18}$$

Indeed, if X is optimal, then (17) and feasibility of X implies (18). Conversely, if X satisfies (18), then it satisfies the KKT condition (3) and hence it is optimal because the problem (1) is convex. Since the optimal solution is unique by assumption, we know the system (18) admits a unique solution. Using Lemma 6 in Section F in the Appendix, we have the relationship between $(\langle Z_\star, X \rangle, \|\mathcal{A}(X - X_\star)\|_2)$ and the distance to the solution $\|X - X_\star\|_{\mathrm{F}}$:

$$\begin{aligned} \|X - X_\star\|_{\mathrm{F}}^2 \leq & \left(4 + 8\frac{\sigma_{\max}^2(\tilde{\mathcal{A}})}{\sigma_{\min}^2(\tilde{\mathcal{A}}_V)} \right) \frac{\langle Z_\star, X \rangle}{\lambda_{n-r_\star}(Z_\star)} \\ & + \frac{4}{\sigma_{\min}^2(\tilde{\mathcal{A}}_V)} \|\mathcal{A}(X) - b\|_2^2. \end{aligned} \tag{19}$$

Combining (17) and (19), we see that

$$f(X) - f(X_\star) \geq \gamma \|X - X_\star\|_{\mathrm{F}}^2$$

for $\gamma = \min\left\{ \frac{\lambda_{n-r_\star}(Z_\star)}{4+8\frac{\sigma_{\max}^2(\tilde{\mathcal{A}})}{\sigma_{\min}^2(\tilde{\mathcal{A}}_V)}}, \frac{\alpha \sigma_{\min}^2(\tilde{\mathcal{A}}_V)}{8} \right\}$. □

5. Numerics

In this section, we verify numerically a few of our claims in the paper, and show the advantages of the Spectral Frank-Wolfe algorithm when strict complementarity is satisfied

[8] We show its convergence under quadratic growth in Lemma 8 in Section F in the Appendix.

Dimension n	Avg. gap	Avg. recovery error
100	288.06	0.0013
200	505.16	0.00064
400	961.09	0.00031
600	1358.62	0.00021

Table 2. Verification of low rankness and strict complementarity. The recovery error is measured by $\frac{\|\frac{X_\star}{\tau} - U_\natural U_\natural^\top\|_{\mathrm{F}}}{\|U_\natural U_\natural^\top\|_{\mathrm{F}}}$. The gap is measured by $\lambda_{n-3}(\nabla f(X_\star)) - \lambda_n(\nabla f(X_\star))$. All the results is averaged over 20 iid trials.

and the solution rank is larger than 1. We focus on the quadratic sensing problem (Chen et al., 2015). Given a random matrix $U_\natural \in \mathbb{R}^{n\times r_\natural}$ with $r_\natural = 3$ and Frobenius norm $\|U_\natural\|_{\mathrm{F}}^2 = 1$, we generate Gaussian vectors $a_i \in \mathbb{R}^{n\times 1}, i = 1,\dots,m$ and construct quadratic measurement vectors $y_0(i) = \|U_\natural^\top a_i\|_{\mathrm{F}}^2, i = 1,\dots,m$. We then add noise $\mathtt{n} = c\|y_0\|_2 v$, where c is the inverse signal-to-noise ratio and v is a random unit vector. Our observation is given by $y = y_0 + \mathtt{n}$ and we aim to recover $U_\natural U_\natural^\top$ from y. To this end, we solve the following optimization problem:

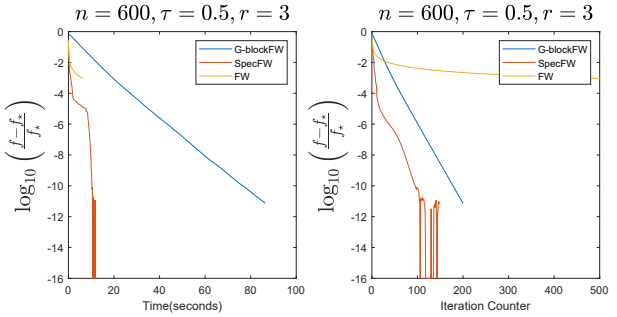
$$\begin{aligned} \text{minimize} \quad & f(X) := \frac{1}{2}\sum_{i=1}^m \left(a_i^\top X a_i - y_i\right)^2 \\ \text{subject to} \quad & \mathbf{tr}(X) = \tau, \quad X \succeq 0. \end{aligned} \tag{20}$$

We set $m = 15nr_\natural$ in all our experiments.

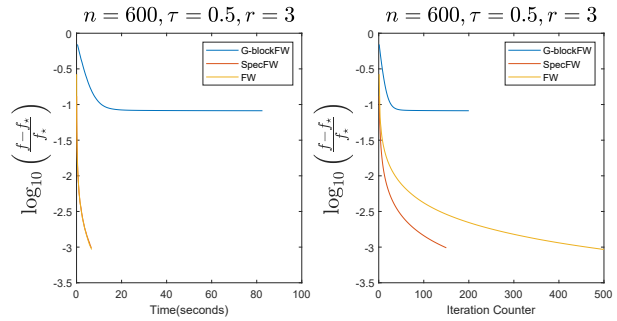
Low rankness and strict complementarity. We verify the low rankness and strict complementarity for $n = 100, 200, 400$ and 600. We set $c = 0.5$ for the noise Level. We also set $\tau = 0.5$, since otherwise, the optimal solution will fit the noise and results in a higher rank matrix. Problem (20) is solved via FASTA (Goldstein et al., 2014; 2015). We found that every optimal solution rank in this case is $r_\star = 3$, and there is indeed a significant gap between $\lambda_{n-3}(\nabla f(X_\star))$ and $\lambda_n(\nabla f(X_\star))$, which verifies strict complementarity. More details can be found in Table 5.

Comparison of algorithms. We now compare the performance of FW, G-BlockFW, and SpecFW. We follow the setting as the previous paragraph for $n = 100, 200, 400$, and 600. We set $k = 4$ for both SpecFW and G-BlockFW, which is larger than $r_\star = 3$. We also set $\eta = 0.4$ and $\beta = 2.5n^2$.[9] The small-scale SDP (10) is solved via FASTA. We plot the relative objective value against both the time and iteration

[9]This choice might appear conservative. But we note that a_i has length around $\sqrt{n}$. Hence the operator norm of $\mathcal{A}$ is around $\sqrt{m}n$ ($\langle a_i a_i^\top, X\rangle \le \|a_i\|_{\mathrm{F}}^2\|X\|_{\mathrm{F}} \approx n\|X\|_{\mathrm{F}}$), which suggests $L_f = \|\mathcal{A}\|_{\mathrm{op}}^2 = n^2 m$ as a safe choice. We have already omitted one m factor here for better algorithmic performance.



(a) $k > r_\star$



(b) $k < r_\star$

Figure 1. Comparison of algorithms under different setting. $f^\star$ is obtained from the best value of the three methods and FASTA.

counter in Figure 1. We only present the plot for the case of $n = 600$ here and those for the other cases can be found in Section G in the Appendix. As can be seen from Figure 1(a), SpecFW converges faster in terms of both the iteration counter and the time. The oscillation in the end may be attributed to the sub-problem solver.

Misspecification of k. We adopt the same setting as before. In this experiment, we set $k = 2$ for both SpecFW and G-BlockFW, which is less than $r_\star = 3$. As can be seen from the Figure 1(b), SpecFW still converges as fast as FW (the two line coincide). G-BlockFW gets stuck around 10^{-1} and stop converging to the optimal solution.

6. Discussion

In this paper, we propose the Spectral Frank-Wolfe algorithm, a novel variant of the classical Frank-Wolfe algorithm, which converges sublinearly for convex smooth optimization problems and converges linearly when strict complementary is satisfied for structural convex optimization problems. We also show that the quadratic growth condition, which is essential for linear convergence of first order methods, holds under strict complementarity.

Here we discuss two potentially interesting extensions of

the current paper:

- **Total computational complexity:** The complexity of subproblem (10) is not discussed and hence leave the total complexity unresolved. Simply using the known $\mathcal{O}(\frac{1}{\sqrt{\epsilon}})$ result for the subproblem complexity seems to be too pessimistic. Is it possible to improve this complexity to $\mathcal{O}(\log(\frac{1}{\epsilon}))$?
- **Solving Subproblem (10) by sub-sampling?** In many applications, f is of a finite sum structure with m terms, e.g., matrix completion, and quadratic sensing. The number m is usually on the order $nr_\star$. In the subproblem (10), the decision variable has size $\mathcal{O}(k^2)$, which is much smaller than m. It might be unwise to use all the m terms. Can we sub-sample the m terms to reduce the burden of computing gradient?

Acknowledgements

L. Ding and Y. Fei were supported by the National Science Foundation CRII award 1657420 and grant 1704828. C. Yang was supported in part by DARPA Award FA8750-17-2- 0101. We would like to thank Yudong Chen, Madeleine Udell, James Renegar, and Adrian Lewis for helpful discussions.

References

Ahmed, A., Recht, B., and Romberg, J. Blind deconvolution using convex programming. *IEEE Transactions on Information Theory*, 60(3):1711–1732, 2013.

Alizadeh, F., Haeberly, J.-P. A., and Overton, M. L. Complementarity and nondegeneracy in semidefinite programming. *Mathematical programming*, 77(1):111–128, 1997.

Allen-Zhu, Z., Hazan, E., Hu, W., and Li, Y. Linear convergence of a frank-wolfe type algorithm over trace-norm balls. In *Advances in Neural Information Processing Systems*, pp. 6191–6200, 2017.

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

Candes, E. J., Eldar, Y. C., Strohmer, T., and Voroninski, V. Phase retrieval via matrix completion. *SIAM review*, 57 (2):225–251, 2015.

Chen, Y., Chi, Y., and Goldsmith, A. J. Exact and stable covariance estimation from quadratic sampling via convex programming. *IEEE Transactions on Information Theory*, 61(7):4034–4059, 2015.

Drusvyatskiy, D. and Lewis, A. S. Generic nondegeneracy in convex optimization. *Proceedings of the American Mathematical Society*, pp. 2519–2527, 2011.

Drusvyatskiy, D., Ioffe, A. D., and Lewis, A. S. Generic minimizing behavior in semialgebraic optimization. *SIAM Journal on Optimization*, 26(1):513–534, 2016.

Frank, M. and Wolfe, P. An algorithm for quadratic programming. *Naval research logistics quarterly*, 3(1-2): 95–110, 1956.

Freund, R. M., Grigas, P., and Mazumder, R. An extended frank-wolfe method with "in-face" directions, and its application to low-rank matrix completion. *SIAM Journal on optimization*, 27(1):319–346, 2017.

Garber, D. Faster projection-free convex optimization over the spectrahedron. In *Advances in Neural Information Processing Systems*, pp. 874–882, 2016.

Garber, D. On the convergence of projected-gradient methods with low-rank projections for smooth convex minimization over trace-norm balls and related problems. *arXiv preprint arXiv:1902.01644*, 2019a.

Garber, D. Linear convergence of frank-wolfe for rank-one matrix recovery without strong convexity. *arXiv preprint arXiv:1912.01467*, 2019b.

Goldstein, T., Studer, C., and Baraniuk, R. A field guide to forward-backward splitting with a FASTA implementation. *arXiv eprint*, abs/1411.3406, 2014. URL http://arxiv.org/abs/1411.3406.

Goldstein, T., Studer, C., and Baraniuk, R. FASTA: A generalized implementation of forward-backward splitting, January 2015. http://arxiv.org/abs/1501.04979.

Jaggi, M. Revisiting frank-wolfe: Projection-free sparse convex optimization. In *Proceedings of the 30th international conference on machine learning*, pp. 427–435, 2013.

Jaggi, M. and Sulovskỳ, M. A simple algorithm for nuclear norm regularized problems. In *Proceedings of the 27th International Conference on International Conference on Machine Learning*, pp. 471–478, 2010.

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

Kuczyński, J. and Woźniakowski, H. Estimating the largest eigenvalue by the power and lanczos algorithms with a random start. *SIAM journal on matrix analysis and applications*, 13(4):1094–1122, 1992.

Necoara, I., Nesterov, Y., and Glineur, F. Linear convergence of first order methods for non-strongly convex optimization. *Mathematical Programming*, 175(1-2):69–107, 2019.

Nesterov, Y. *Introductory lectures on convex optimization: A basic course*, volume 87. Springer Science & Business Media, 2013.

Recht, B., Fazel, M., and Parrilo, P. A. Guaranteed minimum-rank solutions of linear matrix equations via nuclear norm minimization. *SIAM review*, 52(3):471–501, 2010.

Trefethen, L. N. and Bau III, D. *Numerical linear algebra*, volume 50. Siam, 1997.

Tropp, J. A., Yurtsever, A., Udell, M., and Cevher, V. Practical sketching algorithms for low-rank matrix approximation. *SIAM Journal on Matrix Analysis and Applications*, 38(4):1454–1485, 2017.

Yurtsever, A., Udell, M., Tropp, J., and Cevher, V. Sketchy decisions: Convex low-rank matrix optimization with optimal storage. In *Artificial Intelligence and Statistics*, pp. 1188–1196, 2017.