Appendices to “Spectral Frank-Wolfe Algorithm: Strict Complementarity and Linear Convergence”

A. Uniqueness assumption

Here we discuss how to adapt our results to multiple solution setting. First of all, if there are multiple solution, the strict complementarity condition means that there is a primal optimal solution $X_\star$ such that

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

Thus we should set $r_\star$ to be the maximal rank among all primal solutions. Denote the set of primal optimal solution of Problem (1) as $\mathcal{X}_\star$. Quadratic growth in this situation is understood as

$$f(X) - f(X_\star) \geq \gamma \inf_{X_\star \in \mathcal{X}_\star} \|X - X_\star\|_{\text{F}} =: \text{dist}(X, \mathcal{X}_\star),$$

for any $X \succeq 0$ and $\mathbf{tr}(X) = 1$. Now due to strict complementarity, we still have $r_\star = k_\star$ (dual solution $Z_\star$ is unique as shown in the next section). Theorem 3 can be now be proved in the exactly same way by considering the nearest $X_\star \in \mathcal{X}_\star$ to X_t without the uniqueness assumption. To prove Theorem 6, the argument follows exactly as the main proof by considering the nearest $X_\star \in \mathcal{X}_\star$ to X, and replacing Lemma 6 by Lemma 7. In this case, the parameter γ of quadratic growth is $\gamma = \min\left\{\frac{\lambda_{n-r_\star}(Z_\star)}{4+8\frac{\sigma^2_{\max}(\mathcal{A})}{\mu^2}}, \frac{\alpha\mu^2}{8}\right\}$ where $\mu := \sup\{a \geq 0 \mid a \cdot \text{dist}(X, \mathcal{X}_\star) \leq \|\tilde{\mathcal{A}}(X) - b\|_2 \text{ for all } X \in \mathcal{C}_{r_\star}(Z_\star)\}$ and is indeed positive using Lemma 7.

B. Lemmas for Section 2

Lemma 1. *The dual solution $(Z_\star, s_\star)$ of Problem (1) is unique even if the primal solution is not unique.*

Proof. We first show that for any primal solution $X_\star$, its gradient $\nabla f(X_\star)$ is the same. Using β-smoothness of f (the constant β can be taken to be $\|\mathcal{A}\|_{\text{op}}^2 L_g$), we have for any optimal $X_\star$ and $X'_\star$

$$\begin{aligned} &\langle X_\star - X'_\star, \nabla f(X_\star) - \nabla f(X'_\star)\rangle \\ \geq &\frac{1}{\beta}\|\nabla f(X_\star) - \nabla f(X'_\star)\|_{\text{F}}^2. \end{aligned} \tag{1}$$

Since $X_\star$ and $X'_\star$ are optimal solution, we have the following two inequalities using the optimality

$$\langle X_\star - X'_\star, \nabla f(X_\star)\rangle \leq 0, \tag{2}$$
$$\langle X'_\star - X_\star, \nabla f(X'_\star)\rangle \leq 0. \tag{3}$$

Combining the inequalities (1), (2), and (3), we have

$$\|\nabla f(X_\star) - \nabla f(X'_\star)\|_{\text{F}} \leq 0 \implies f(X_\star) = f(X'_\star). \tag{4}$$

This shows that $\nabla f(X_\star)$ is unique. Now for any $Z_\star, s_\star$ and $Z'_\star, s'_\star$ satisfying the KKT condition, we have

$$\begin{aligned} \nabla f(X_\star) + C &= Z_\star + s_\star I \\ &= Z'_\star + s'_\star I \\ \implies Z_\star - Z'_\star &= (s'_\star - s_\star)I. \end{aligned} \tag{5}$$

Now using complementarity in step (a) and feasibility of $X_\star$ in step (b):

$$\begin{aligned}0 &\overset{(a)}{=} \langle Z_\star - Z'_\star, X_\star\rangle = (s'_\star - s_\star)\langle I, X_\star\rangle \\ &\overset{(b)}{=} (s'_\star - s_\star) \\ \implies s_\star &= s'_\star, \quad \text{and} \quad Z_\star = Z'_\star.\end{aligned} \tag{6}$$

Hence the dual solution $Z_\star$ and $s_\star$ is unique. □

Lemma 2. *For almost all C, the strict complementarity condition holds for (1).*

Proof. Let us first define indicator function: for any given $D \subset \mathbb{R}^n$, we define

$$\chi_C(x) = \begin{cases} 0, & x \in D \\ +\infty, & x \notin D. \end{cases}$$

Also denote the relative interior of a set D as relint(D). We utilize the result in Drusvyatskiy & Lewis (2011, Corollary 3.5), that for almost all C, we have

$$\begin{aligned}-C \in &\operatorname{relint}(\partial(g(\mathcal{A}X) \\ &\quad + \chi_{\{\mathbf{tr}(X)=1\}}(X) + \chi_{\{X\succeq 0\}}(X))(X_\star)) \\ &\overset{(a)}{=}\operatorname{relint}(\mathcal{A}^*(\nabla g)(\mathcal{A}X_\star) + \{sI \mid s \in \mathbb{R}\} \\ &\quad + \{-Z \mid Z \succeq 0, \operatorname{range}(Z) \subset \operatorname{nullspace}(X_\star)\}) \\ &\overset{(b)}{=}\mathcal{A}^*(\nabla g)(\mathcal{A}X_\star) + C + \{sI \mid s \in \mathbb{R}\} \\ &\quad + \{-Z \mid Z \succeq 0, \operatorname{range}(Z) = \operatorname{nullspace}(X_\star)\}.\end{aligned} \tag{7}$$

Here we use the sum rule in step (a) as $\frac{1}{n}I$ is in $\{X \mid \mathbf{tr}(X) = 1\}$ and the interior of $\{X \mid X \succeq 0\}$. In step (b), we use the sum rule of relative interior. Hence, there is some $s_\star$ and $Z_\star$ such that

$$\begin{aligned}&\operatorname{range}(Z_\star) = \operatorname{nullspace}(X_\star) \\ \implies &\langle Z_\star, X_\star\rangle = 0, \quad \text{and} \\ &\operatorname{rank}(Z_\star) + \operatorname{rank}(X_\star) = n.\end{aligned} \tag{8}$$

and

$$\mathcal{A}^*(\nabla g)(\mathcal{A}X_\star) + C = Z_\star + s_\star I.$$

We thus conclude $(Z_\star, s_\star)$ satisfies the KKT condition (3), and strict complementarity holds. □

C. SpecFW: minimizing an upper bound of $f(\eta X_t + VSV^\top)$.

When the function f is not fully known or gradient might be hard to query, we may consider the following subproblem instead: solve

$$\begin{aligned}\text{minimize} \quad & g(\mathcal{A}X_t) \\ & + \langle \mathcal{A}(\eta X_t + VSV^\top) - \mathcal{A}X_t, (\nabla g)(\mathcal{A}X_t)\rangle \\ & + \frac{L_g}{2}\|\mathcal{A}(\eta X_t + VSV^\top) - \mathcal{A}X_t\|_2^2 \\ & + \langle C, \eta X_t + VSV^\top\rangle \\ \text{subject to} \quad & \eta + \mathbf{tr}(S) = 1,\ S \succeq 0,\ \text{and } \eta \geq 0.\end{aligned} \tag{9}$$

with decision variable S and η. Then set $X_{t+1} = \eta X_t + VSV^\top$ for the optimal η and S.

The above formulation enjoys the advantage of efficient computation in terms of time when m is small and the linear map $\mathcal{A}$ and $\langle C, \cdot\rangle$ are easy to apply to low rank matrices. One may also save $\mathcal{A}X_t$ during the process to avoid forming X_t and sketching X_t using idea from Tropp et al. (2017) for storage purpose.

One could also consider solving

$$\begin{array}{ll} \text{minimize} & f(X_t) \\ & + \langle \eta X_t + VSV^\top - X_t, \nabla f(X_t)\rangle \\ & + \frac{L_f}{2}\|X_t - (\eta X_t + VSV^\top)\|_{\mathrm{F}} \\ \text{subject to} & \eta + \mathbf{tr}(S) = 1,\ S \succeq 0,\ \text{and } \eta \geq 0. \end{array} \tag{10}$$

Then set $X_{t+1} = \eta X_t + VSV^\top$ for the optimal η and S. Here L_f is the Lipschitz constant of ∇f. This method requires to store X_t in each iteration though.

D. Combination with matrix sketching idea in Tropp et al. (2017)

When m is on the order n, we can employ the matrix sketching idea developed in Tropp et al. (2017) and Yurtsever et al. (2017) to achieve storage reduction. We note that if we store $\mathcal{A}(X_t) = z_t$ and $c_t = \langle C, X_t\rangle$ at each iteration, then we have no problem in doing the small-scale SDP (10), as $f(\eta X_t + VSV^\top) = g(\eta(\mathcal{A}X_t) + \mathcal{A}(VSV^\top)) + \eta\langle C, X_t\rangle + \langle C, VSV^\top\rangle$. If $\mathcal{A}$ and inner product with C can be applied to low rank matrices efficiently, then updating z_t and c_t is not hard due to linearity of our updating scheme $X_{t+1} = \eta X_t + VSV^\top$.

Now we explain how to omit storing the iterate X_t. First, we draw two matrices with independent standard normal entries

$$\begin{aligned} \Psi \in \mathbb{R}^{n\times k} \quad &\text{with} \quad k = 2r+1; \\ \Phi \in \mathbb{R}^{l\times n} \quad &\text{with} \quad l = 4r+3; \end{aligned}$$

Here r is chosen by the user. It either represents the estimate of the true rank of the primal solution or the user's computational budget in dealing with larges matrices.

We use Y_t^C and Y_t^R to capture the column space and the row space of X_t:

$$Y_t^C = X_t\Psi \in \mathbb{R}^{n\times k}, \qquad Y_t^R = \Phi X_t \in \mathbb{R}^{l\times n}. \tag{11}$$

Hence we initially have $Y_0^C = 0$ and $Y_0^R = 0$. Notice that SpecFW does not observe matrix X_t directly. Rather, it observes a stream of rank k updates

$$X_{t+1} = VSV^\top + \eta X_t,$$

where $V \in \mathbb{R}^n \times k$ and $S \in \mathbb{S}^k$.

In this setting, Y_{t+1}^C and Y_{t+1}^R can be directly computed as

$$Y_{t+1}^C = VS(V^\top\Psi) + \eta Y_t^C \in \mathbb{R}^{n\times k}, \tag{12}$$

$$Y_{t+1}^R = (\Psi V)SV^\top + \eta Y_t^R \in \mathbb{R}^{l\times n}. \tag{13}$$

This observation allows us to form the sketch Y_t^C and Y_t^R from the stream of updates.

We then reconstruct X_t and get the reconstructed matrix $\hat{X}_t$ by

$$Y_t^C = Q_tR_t, \quad B_t = (\Phi Q_t)^\dagger Y_t^R, \quad \hat{X}_t = Q_t[B_t]_r, \tag{14}$$

where Q_tR_t is the QR factorization of Y_t^C and $[\cdot]_r$ returns the best rank r approximation in Frobenius norm. Specifically, the best rank r approximation of a matrix Z is $U\Sigma V^*$, where U and V are right and left singular vectors corresponding to the r largest singular values of Z and Σ is a diagonal matrix with r largest singular values of Z. In actual implementation, we may only produce the factors (QU, Σ, V) defining $\hat{X}_T$ in the end instead of reconstructing $\hat{X}_t$ in every iteration. We refer the reader to Tropp et al. (2017, Theorem 5.1) for the theoretical guarantees on the reconstruction matrix $\hat{X}_t$.

Hence we can avoid the *forming a new iteratre* procedure in SpecFW. We remark that the reconstructed matrix $\hat{X}_t$ is not necessarily positive semidefinite. However, this suffices for the purpose of finding a matrices close to X_t. More sophisticated procedure is available for producing a positive semidefinite approximation of X_t (Tropp et al., 2017, Section 7.3).

E. Proofs for Section 3

We first give the detailed calculation of the derivation for (12).

Continuation of proof of Theorem 3. We need to choose $\xi \in [0,1]$ so that $1-\xi+\frac{\xi^2\beta}{\gamma}$ is minimized while keeping $\xi^2\beta - \frac{\xi\lambda_{n-r_\star}(Z_\star)}{6} \leq 0$. For $\xi^2\beta - \frac{\xi\lambda_{n-r_\star}(Z_\star)}{6} \leq 0$, we need $\xi \leq \frac{\lambda_{n-r_\star}(Z_\star)}{6\beta}$. The function $q(\xi) = 1-\xi+\frac{\xi^2\beta}{\gamma}$ is decreasing for $\xi \leq \frac{\gamma}{2\beta}$ and increasing for $\xi \geq \frac{\gamma}{2\beta}$. If $\frac{\gamma}{2\beta} \leq \frac{\lambda_{n-r_\star}(Z_\star)}{6\beta}$, then we can pick $\xi = \frac{\gamma}{2\beta}$, and $q(\xi) = 1-\frac{\gamma}{4\beta}$. If $\frac{\gamma}{2\beta} \geq \frac{\lambda_{n-r_\star}(Z_\star)}{6\beta} \implies \frac{\lambda_{n-r_\star}(Z_\star)}{\gamma} \leq 3$, then we can pick $\xi = \frac{\lambda_{n-r_\star}(Z_\star)}{6\beta}$, and $q(\xi) = 1 - \frac{\lambda_{n-r_\star}(Z_\star)}{6\beta} + \frac{\lambda^2_{n-r_\star}(Z_\star)}{36\gamma\beta} = 1 + \frac{\lambda_{n-r_\star}(Z_\star)}{6\beta}\left(\frac{\lambda_{n-r_\star}(Z_\star)}{6\gamma} - 1\right) \leq 1 - \frac{\lambda_{n-r_\star}(Z_\star)}{12\beta}$. $\square$

We shall prove Lemma 5 in this section. We restate Lemma 5 in a self-contained way.

Lemma 3. *Suppose $Y \in \mathbb{S}^n$ with eigenvalues $\lambda_1(Y) \geq \cdots \geq \lambda_n(Y)$, and $\lambda_{n-r}(Y) - \lambda_{n-r+1}(Y) \geq \delta$. Here $\lambda_i(\cdot)$ denote the operator of taking the i-th largest eigenvalue. Also let $v_1, \ldots, v_n$ be the corresponding orthornomal eigenvectors. Denote the eigenspace corresponding to the last $r$eigenvalus of Y as $\mathcal{V}_{Y,r}$ and the corresponding orthorgonal projection $P_{Y,r} : \mathbb{R}^n \to \mathbb{R}^n$ which is also a matrix in $\mathbb{R}^{n\times n}$. Let $V_{Y,r} \in \mathbb{R}^{n\times r}$ formed by the last r many eigenvectors $v_{n-r+1}, \ldots v_n$ which represents the eigensapce $\mathcal{V}_{Y,r}$. Define $\mathcal{C}_r(Y) = \left\{V_{Y,r}SV_{Y,r}^\top \mid S \succeq 0, \mathbf{tr}(S) = 1\right\}$. Then for any $X \in \mathbb{S}^n$ with $\mathbf{tr}(X) = 1, X \succeq 0$, there is some $W \in \mathcal{C}_r(Y)$ such that*

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

Remark 4. *We note that as long as range(V) = range$(V_{Y,r})$ for some matrix $V \in \mathbb{R}^{n\times r}$ with orthonormal columns, the set $\mathcal{C}_r(Y)$ is the same as $\left\{VSV^\top \mid S \succeq 0, \mathbf{tr}(S) = 1\right\}$.*

Proof of Lemma 5. We first decompose X by

$$X = \underbrace{(X - P_{Y,r}XP_{Y,r})}_{X_1} + \underbrace{P_{Y,r}XP_{Y,r}}_{=:X_2}.$$

Note that $P_{Y,r} = P_{Y,r}^\top$, so $X_2 = P_{Y,r}XP_{Y,r}$ is still symmetric. Let $1-\epsilon = \mathbf{tr}(P_{Y,r}XP_{Y,r})$. Since $\mathbf{tr}(X) = 1$, we have $\epsilon = \mathbf{tr}(X - P_{Y,r}XP_{Y,r})$. We have $\epsilon \in [0,1]$ as $\mathbf{tr}(P_{Y,r}XP_{Y,r}) = \langle X, P_{Y,r}P_{Y,r}\rangle \overset{(a)}{\leq} \|P_{Y,r}\|_{\text{op}}\mathbf{tr}(X) \leq 1$ where step (a) is due to Hölder's inequality.

Consider the eigenvalue decomposition of $X_2 = V_2\Lambda_2V_2^\top$, where $V_2 \in \mathbb{R}^{n\times r}$ and $\Lambda_2 \in \mathbb{S}^r$ with all diagonal nonnegative. Here the column space of V_2 satisfies range$(V_2) = \mathcal{V}_{Y,r}$.

Because $P_{Y,r}XP_{Y,r} = X_2$ is a member in $\mathcal{C}_r(Y)$, we know there is an $W \in \mathcal{C}_r(Y)$ such that $W = V_2\Lambda_WV_2^\top$ where $\Lambda_W \in \mathbb{S}^r$ has nonegative diagonal with $\mathbf{tr}(\Lambda_W) = 1$ and the difference matrix $\Delta = \Lambda_W - \Lambda_2$ has nonnegative entries. We also have $\mathbf{tr}(\Delta) = \epsilon$, as the trace of both Λ_W and X are one.

With such choice of W, let us now analyze $\langle X - W, Y\rangle$:

$$\begin{aligned}\langle X - W, Y\rangle &= \langle X_1, Y\rangle + \langle X_2 - W, Y\rangle \\ &= \underbrace{\langle X - P_{Y,r}XP_{Y,r}, \sum_{i=1}^n \lambda_i(Y)v_iv_i^\top\rangle}_{R_1} \\ &\quad - \underbrace{\langle V_2\Delta V_2^\top, \sum_{i=1}^n \lambda_i(Y)v_iv_i^\top\rangle}_{R_2}.\end{aligned} \tag{15}$$

The first term $R_1 = \langle X - P_{Y,r} X P_{Y,r}, \sum_{i=1}^n \lambda_i(Y) v_i v_i^\top \rangle$ satifies

$$\begin{aligned}
&\langle X - P_{Y,r} X P_{Y,r}, \sum_{i=1}^n \lambda_i(Y) v_i v_i^\top \rangle \\
&\overset{(a)}{=} \sum_{i=1}^n \lambda_i(Y) v_i^\top X v_i - \sum_{i=n-r+1}^n \lambda_i(Y) v_i^\top X v_i \\
&= \sum_{i=1}^{n-r} \lambda_i(Y) v_i^\top X v_i \\
&\overset{(b)}{\geq} (\lambda_{n-r+1}(Y) + \delta) \sum_{i=1}^{n-r} v_i^\top X v_i.
\end{aligned}$$

Here in step (a) we uses the fact that $P_{Y,r} v_i = v_i$ for $i = n-r+1, \dots n$ and is zero for other v_i. In step (b), we use the assumption that $\lambda_{n-r} - \lambda_{n-r+1} \geq \delta$ and each $v_i^\top X v_i \geq 0$ as $X \succeq 0$. We note that $\sum_{i=1}^{n-r} v_i^\top X v_i$ satifies

$$\begin{aligned}
\sum_{i=1}^{n-r} v_i^\top X v_i = \mathbf{tr}\left(X \left(\sum_{i=1}^{n-r} v_i v_i^\top \right) \right) &\overset{(a)}{=} \mathbf{tr}(X(I - P_{Y,r})) \\
&\overset{(b)}{=} \mathbf{tr}(X) - \mathbf{tr}(P_{Y,r} X P_{Y,r}) = \epsilon.
\end{aligned}$$

Here step (a) uses the $P_{Y,r} = V_{Y,r} V_{Y,r}^\top$ and we use $P_{Y,r}^2 = P_{Y,r}$ and cyclic property of trace in step (b).

Now let us analyze the second term R_2:

$$\begin{aligned}
R_2 &= \langle V_2 \Delta V_2^\top, \sum_{i=1}^n \lambda_i(Y) v_i v_i^\top \rangle \\
&\overset{(a)}{=} \langle V_2 \Delta V_2^\top, \sum_{i=n-r+1}^n \lambda_i(Y) v_i v_i^\top \rangle.
\end{aligned}$$

Here we use the fact that $V_2^\top v_i = 0$ for all $v_i, i = 1, \dots n-r$. Since $V_{Y,r}$ and V_2 are both orthonormal representation of $\mathcal{V}_{Y,r}$, we know there is an orthonormal matrix $O \in \mathbb{R}^{r \times r}$ such that $V_{Y,r} = V_2 O$. Define the linear operator diag : $\mathbb{S}^n \to \mathbb{R}^n$, which takes the diagonal of a matrix. Let $\Lambda_{Y,r} = \text{diag}^* \left(\lambda_{n-r+1}(Y), \dots, \lambda_n(Y) \right)$, we see R_2 further equals to

$$\begin{aligned}
R_2 &= \mathbf{tr}\left(V_2 \Delta V_2^\top V_2 O \Lambda_{Y,r} O^\top V_2^\top \right) \\
&\overset{(a)}{=} \mathbf{tr}\left(\Delta O \Lambda_{Y,r} O^\top \right) \\
&\overset{(b)}{\leq} \epsilon \lambda_{n-r+1}(Y).
\end{aligned}$$

Here we use the cyclic property in step (a) and the step (b) is an easy consequence of Δ has nonnegative diagonal and Von Neumann's trace inequality: for symmetric matrices $A, B \in \mathbb{S}^r$, $\langle A, B \rangle \leq \sum_{i=1}^r \lambda_i(A) \lambda_i(B)$. Combining pieces, we find that

$$\langle X - W, Y \rangle \geq (\lambda_{n-r+1}(Y) + \delta)\epsilon - \epsilon \lambda_{n-r+1}(Y) = \delta \epsilon.$$

Now we turn to analyzing the term $\|X - W\|_{\mathrm{F}}^2$. Using $\langle X_1, X_2 \rangle = 0, \langle X_1, W \rangle = 0$, we find that

$$\|X - W\|_{\mathrm{F}}^2 = \|X_1\|_{\mathrm{F}}^2 + \|X_2 - W\|_{\mathrm{F}}^2.$$

The second term $\|X_2 - W\|_{\mathrm{F}}^2$ satisfies

$$\|X_2 - W\|_{\mathrm{F}} = \|V_2 \Delta V_2^\top\|_{\mathrm{F}}^2 = \sum_{i=1}^r \Delta_{ii}^2 \leq \left(\sum_{i=1}^r \Delta_{ii} \right)^2 = \epsilon^2.$$

If we write X in terms of the coordinates given by V_2 and its orthogonal compliment say V_1, then in this new coordinate $V = [V_1, V_2]$:

$$V^\top X V = \begin{bmatrix} A & B \\ B & V_2^\top X_2 V_2 \end{bmatrix}, \quad \text{and} \quad V^\top X_1 V = \begin{bmatrix} A & B \\ B & 0 \end{bmatrix}.$$

Then $\mathbf{tr}(X_1) = \mathbf{tr}(A)$. Lemma 5 implies that

$$\|B\|_{\mathrm{F}}^2 \le \mathbf{tr}(X_2)\mathbf{tr}(A) = \epsilon(1-\epsilon) = \epsilon - \epsilon^2.$$

Hence $\|X_1\|_{\mathrm{F}}^2 = \|A\|_{\mathrm{F}}^2 + 2\|B\|_{\mathrm{F}}^2 \le (\mathbf{tr}(A))^2 + 2\epsilon - 2\epsilon^2 = -\epsilon^2 + 2\epsilon$. Combining pieces and $\epsilon \in [0,1]$, we find that

$$\begin{aligned} \|X - W\|_{\mathrm{F}}^2 &\le 2\epsilon = \frac{2}{\delta}\delta\epsilon \le \frac{2}{\delta}\langle X - W, Y\rangle \\ \Longrightarrow \langle X - W, Y\rangle &\ge \frac{\delta}{2}\|X - W\|_{\mathrm{F}}^2. \end{aligned}$$

□

Lemma 5. *Suppose* $Y = \begin{bmatrix} A & B \\ B^\top & D \end{bmatrix} \succeq 0$. *Then* $\|A\|_{op}\mathbf{tr}(D) \ge \|BB^\top\|_* = \mathbf{tr}(BB^\top) = \|B\|_F^2$.

Proof. For any $\epsilon > 0$, denote $A_\epsilon = A + \varepsilon I$ and $Y_\epsilon = \begin{bmatrix} A_\epsilon & B \\ B^* & D \end{bmatrix}$. We know Y_ϵ is psd, as is its Schur complement $D - B^\top A_\epsilon^{-1} B \succeq 0$ with trace $\mathbf{tr}(D) - \mathbf{tr}(A_\epsilon^{-1} BB^\top) \ge 0$.

Von Neumann's lemma for A_ϵ, $BB^\top \succeq 0$ shows $\mathbf{tr}(A_\epsilon^{-1} BB^*) \ge \frac{1}{\|A_\epsilon\|_{\mathrm{op}}}\|BB^\top\|_*$. Use this with the previous inequality to see $\mathbf{tr}(D) \ge \frac{1}{\|A_\epsilon\|_{\mathrm{op}}}\|BB^\top\|_*$. Multiply by $\|A_\epsilon\|_{\mathrm{op}}$ and let $\varepsilon \to 0$ to complete the proof. □

F. Lemmas for Section 4

We first give a self-contained proof for the second case of Theorem 6.

Proof of second case of Theorem 6. For any feasible X and the optimal solution $X_\star$, we have

$$\begin{aligned} f(X) - f(X_\star) &\overset{(a)}{\ge} \langle \nabla f(X_\star), X - X_\star\rangle \\ &\overset{(b)}{=} \langle Z_\star + s_\star I, X - X_\star\rangle \\ &\overset{(c)}{=} \langle Z_\star, X - X_\star\rangle. \end{aligned}$$

Here step (a) is due to the convexity of f. For step (b), we uses the first order condition of KKT condition (3). The step (c) is due to feasibility of X and $X_\star$.

Since $Z_\star$ has rank $n-1$, using strict complementarity, we reach that any optimal solution $X_\star$ has rank 1 with range$(X_\star) =$ nullspace$(Z_\star)$. Thus any optimal solution $X_\star$ is of the form $X_\star = \xi vv^\top$, v is the non-zero unit vector in the null space of $Z_\star$, and ξ is a nonnegative scaler. Since $X_\star$ has to be feasible, the constraint $\mathbf{tr}(X_\star) = 1$ implies that $\xi = 1$ and hence the solution $X_\star$ is unique. The same argument implies that the set $\mathcal{C}_1(Z_\star) = \{X_\star\}$. Hence using Lemma 5 and $\lambda_n(Z_\star) = 0$, we see that

$$f(X) - f(X_\star) \ge \langle Z_\star, X - X_\star\rangle \ge \frac{\lambda_{n-1}(Z_\star)}{2}\|X - X_\star\|_{\mathrm{F}}^2.$$

□

Next, we establish the lemma that is core to the proof of Theorem 6 under the assumption of uniqueness.

Lemma 6. *Suppose the following system admits a unique solution $X_\star$ with rank $r_\star$:*

$$\langle Z_\star, X_\star \rangle = 0, \mathcal{A}X = b, \quad \text{and} \quad X \succeq 0, \tag{16}$$

for a $Z_\star \succeq 0$ such that rank$(Z_\star)$ + rank$(X_\star) = n$, a linear map $\mathcal{A} : \mathbb{S}^n \to \mathbb{R}^m$, and a vector $b \in \mathbb{R}^m$. Furthur suppose that $\mathcal{A}X = b \implies \mathbf{tr}(X) = 1$. Then for any $X \succeq 0$ with $\mathbf{tr}(X) = 1$, we have

$$\begin{aligned}\|X - X_\star\|_F^2 \leq & \left(4 + 8\frac{\sigma_{\max}(\mathcal{A})}{\sigma_{\min}(\mathcal{A}_V)}\right)\frac{\langle Z_\star, X\rangle}{\lambda_{n-r_\star}(Z_\star)} \\ & + \frac{4}{\sigma_{\min}^2(\mathcal{A}_V)}\|\mathcal{A}(X) - b\|_2^2.\end{aligned} \tag{17}$$

Proof. Let $V \in \mathbb{R}^{n \times r_\star}$ be a matrix with orthonormal columns correpsonding to the eigenspace $\mathcal{V}$ of $X_\star$ of positive eigenvalues. Then $X_\star$ can be written as $X_\star = VS_\star V^\top$ for some $S_\star \in \mathbb{S}^{r_\star}$ such that $S_\star \succ 0$. We claim that the linear map $\mathcal{A}_V$ defined as follows is injective:

$$\begin{aligned}\mathcal{A}_V : \mathbb{S}^{r_\star} &\to \mathbb{R}^m \\ S &\mapsto \mathcal{A}(VSV^\top).\end{aligned}$$

Suppose not, then there is some nonzero $S_0 \in \mathbb{S}^{r_\star}$ such that $\mathcal{A}_V(S_0) = 0$. Then $V(\alpha S_0 + S_\star)V^\top$also satisfies the system (16) for all small enough α. Hence we see that for any $S \in \mathbb{S}^r$

$$\begin{aligned}\|VSV^\top - X_\star\|_{\mathrm{F}} \leq & \frac{1}{\sigma_{\min}(\mathcal{A}_V)}\|\mathcal{A}(VSV^\top) - \mathcal{A}(X_\star)\|_2 \\ = & \frac{1}{\sigma_{\min}(\mathcal{A}_V)}\|\mathcal{A}(VSV^\top) - b\|_2.\end{aligned} \tag{18}$$

Here $\sigma_{\min}(\mathcal{A}_V) = \min_{\|S\|_{\mathrm{F}}=1}\|\mathcal{A}_V(S)\|_2 > 0$.

Using strict complementarity on $Z_\star$ and $X_\star$, we know V is also a representation of the null space of the $Z_\star$. Using Lemma 5, we know there is some $W = VSV^\top \in \mathcal{C}_{r_\star}(Z_\star)$ such that

$$\langle X, Z_\star\rangle \overset{(a)}{=} \langle X - W, Z_\star\rangle \geq \frac{\lambda_{n-r_\star}(Z_\star)}{2}\|X - W\|_{\mathrm{F}}^2, \tag{19}$$

where step (a) is because $\lambda_{n-r_\star+1}(Z_\star) = \cdots = \lambda_n(Z_\star) = 0$. We note if $r_\star = 1$, then $\mathcal{C}_r(Z_\star)$ has $X_\star$ as its only element, as $\mathbf{tr}(X) = 1$ and we are done.

We can bound $\|X - X_\star\|_{\mathrm{F}}^2$ by

$$\begin{aligned}\|X - X_\star\|_{\mathrm{F}}^2 &\overset{(a)}{\leq} 2\|X - W\|_{\mathrm{F}}^2 + 2\|W - X_\star\|_{\mathrm{F}}^2 \\ &\overset{(b)}{\leq} 2\|X - W\|_{\mathrm{F}}^2 + \frac{2}{\sigma_{\min}^2(\mathcal{A}_V)}\|\mathcal{A}(W) - b\|_2^2.\end{aligned} \tag{20}$$

Here we use triangle inequality and basic inequality $(a + c)^2 \leq 2a^2 + 2c^2$ for any real a, c in step (a). In step (b), we use (18).

We can further bound the term $\|\mathcal{A}(W) - b\|_2$ by

$$\begin{aligned}\|\mathcal{A}(W) - b\|_2 &= \|\mathcal{A}(W - X) + \mathcal{A}(X) - b\|_2 \\ &\leq \|\mathcal{A}(W - X)\|_2 + \|\mathcal{A}(X) - b\|_2.\end{aligned} \tag{21}$$

Now combining (20), (21) and $(a+c)^2 \leq 2a^2 + 2c^2$ for any $a, c \in \mathbb{R}$ in the following step (a), we see

$$\begin{aligned}\|X - X_\star\|_{\mathrm{F}}^2 &\overset{(a)}{\leq} 2\|X - W\|_{\mathrm{F}}^2 + \frac{4\|\mathcal{A}(W - X)\|_2^2}{\sigma_{\min}^2(\mathcal{A}_V)} \\ &\quad + \frac{4}{\sigma_{\min}^2(\mathcal{A}_V)}\|\mathcal{A}(X) - b\|_2^2 \\ &\leq \left(2 + 4\frac{\sigma_{\max}^2(\mathcal{A})}{\sigma_{\min}^2(\mathcal{A}_V)}\right)\|X - W\|_{\mathrm{F}}^2 \\ &\quad + \frac{4}{\sigma_{\min}^2(\mathcal{A}_V)}\|\mathcal{A}(X) - b\|_2^2.\end{aligned}$$

Finally using (19) to bound $\|X - W\|_{\mathrm{F}}$, we reached the inequality we want to prove:

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

□

We now establish a lemma to handle the general case that the solution might not be unique. For a convex closed set $\mathcal{X}_\star$, we define the distance to for an arbitrary $X \in \mathbb{S}^n$ to it as

$$\mathrm{dist}(X, \mathcal{X}_\star) := \inf_{X_\star \in \mathcal{X}_\star} \|X - X_\star\|_{\mathrm{F}}.$$

Lemma 7. *Denote the solution set of the following system as $\mathcal{X}_\star$:*

$$\langle Z_\star, X_\star\rangle = 0, \mathcal{A}X = b, \quad \textit{and} \quad X \succeq 0, \tag{22}$$

for a $Z_\star \succeq 0$, a linear map $\mathcal{A} : \mathbb{S}^n \to \mathbb{R}^m$, and a vector $b \in \mathbb{R}^m$. Suppose the system (16) admits a solution $X_\star^0$ with rank $r_\star^0 \geq 1$ such that $\mathrm{rank}(Z_\star) + \mathrm{rank}(X_\star^0) = n$. Further suppose that $\mathcal{A}X = b \implies \mathbf{tr}(X) = 1$. Then the constant $\mu := \sup\{a \geq 0 \mid a \cdot \mathrm{dist}(X, \mathcal{X}_\star) \leq \|\mathcal{A}(X) - b\|_2 \text{ for all } X \in \mathcal{C}_{r_\star}(Z_\star)\}$ is positive, and for any $X \succeq 0$ with $\mathbf{tr}(X) = 1$, we have

$$\begin{aligned}\mathrm{dist}(X, \mathcal{X}_\star)^2 \;\leq\; &\left(4 + 8\frac{\sigma_{\max}(\mathcal{A})}{\mu}\right)\frac{\langle Z_\star, X\rangle}{\lambda_{n-r_\star}(Z_\star)} \\ &+ \frac{4}{\mu^2}\|\mathcal{A}(X) - b\|_2^2.\end{aligned} \tag{23}$$

Proof. Let $V \in \mathbb{R}^{n \times r_\star}$ be a matrix with orthonormal columns corresponding to the eigenspace $\mathcal{V}$ of $r_\star$ zero eigenvalues. Consider the linear map $\mathcal{A}_V$:

$$\begin{aligned}\mathcal{A}_V : \mathbb{S}^{r_\star} &\to \mathbb{R}^m \\ S &\mapsto \mathcal{A}(VSV^\top).\end{aligned}$$

The key replacement of multiple solution setting is to establish an inequality similar to (18), which depicts the injectivity of$\mathcal{A}_V$ for unique solution setting.

Define the solution set $\mathcal{S} \subset \mathbb{S}^{r_\star}$ of the following system:

$$\mathcal{A}_V(S) = b, \quad S \succeq 0. \tag{24}$$

Note that any $S \in \mathcal{S}$ satisfies that $VSV^\top \in \mathcal{X}_\star$. Conversely, for any $X_\star \in \mathcal{X}_\star$, it can be written as $X_\star = VS_\star V^\top$ for some $S_\star \in \mathbb{S}^{r_\star}$ such that $S_\star \succeq 0$ and $\mathcal{A}_V(S_\star) = b$. Hence we have $\mathcal{X}_\star = \{X \mid X = VSV^\top,\ S \in \mathcal{S}\}$.

Now if we take the $X_\star^0 \in \mathcal{X}_\star$ such that $\mathrm{rank}(Z_\star) + \mathrm{rank}(X_\star^0) = n$, then $X_\star^0 = VS_\star^0V^\top$ for some $S_\star^0 \in \mathbb{S}^{r_\star}$ such that $S_\star^0 \succ 0$. This means the system (24) satisfies the condition in Corollary 3 in (Bauschke et al., 1999). By applying this corollary to (24), we know there is a $\mu > 0$ such that for all $S \succeq 0$ and $\mathbf{tr}(S) = 1$,

$$\mathrm{dist}(S, \mathcal{S}) \leq \frac{1}{\mu}\|\mathcal{A}_V S - b\|_2. \tag{25}$$

Translating the inequality to the space $\mathcal{L} = \{X \in \mathbb{S} \mid X = VSV^\top \text{ for some } S \in \mathbb{S}^{r_\star}\}$, we have for all $X \succeq 0$, $\mathbf{tr}(X) = 1$, and $X \in \mathcal{L}$, i.e., $X \in \mathcal{C}_{r_\star}(Z_\star)$:

$$\mathrm{dist}(X, \mathcal{X}_\star) \leq \frac{1}{\mu}\|\mathcal{A}(X) - b\|_2. \tag{26}$$

This is our replacement of (18) in Lemma 6.

Following the proof of Lemma 6, we know there is some $W = VSV^\top \in \mathcal{C}_{r_\star}(Z_\star)$ such that

$$\langle X, Z_\star\rangle = \langle X - W, Z_\star\rangle \geq \frac{\lambda_{n-r_\star}(Z_\star)}{2}\|X - W\|_{\mathrm{F}}^2. \tag{27}$$

To bound $\mathrm{dist}(X, \mathcal{X}_\star)$, we pick an $X_\star \in \mathcal{X}_\star$ such that it is nearest to W (mote $\mathcal{X}_\star$ is compact as $\mathcal{A}(X) = b$ implies $\mathbf{tr}(X) = 1$). Then we have

$$\begin{aligned}
\mathrm{dist}(X, \mathcal{X}_\star)^2 &\leq \|X - X_\star\|_{\mathrm{F}}^2 && (28)\\
&\overset{(a)}{\leq} 2\|X - W\|_{\mathrm{F}}^2 + 2\|W - X_\star\|_{\mathrm{F}}^2 && (29)\\
&\overset{(b)}{\leq} 2\|X - W\|_{\mathrm{F}}^2 + \frac{2}{\mu^2}\|\mathcal{A}(W) - b\|_2^2.
\end{aligned}$$

Here we use triangle inequality and basic inequality $(a + c)^2 \leq 2a^2 + 2c^2$ for any real a, c in step (a). In step (b), we use (18). The rest of the proof is exactly the same as those in Lemma 6. $\square$

The following Lemma establishes the linear convergence of G-BlockFW under quadratic growth condition.

Lemma 8. *Suppose f of Problem (1) is β smooth and Problem (1) satisfies quadratic growth with parameter γ. If $\eta = \frac{\gamma}{\beta}$ and $k \geq r_\star = rank(X_\star)$, where $X_\star$ is an optimal solution of Problem (1), then the generalized Block FW 2 converges linearly:*

$$h_{t+1} \leq (1 - \frac{\gamma}{2\beta})h_t,$$

where $h_t = f(X_t) - f(X_\star)$ for each t.

Proof. Denote $\hat{Y} = V\mathrm{diag}(\Lambda)V^\top$. The Lipschitz smoothness of f shows that

$$f(X_{t+1}) \leq f(X_t) + \eta\langle \hat{Y} - X_t, \nabla f(X_t)\rangle + \frac{\eta^2\beta}{2}\|\hat{Y} - X_t\|_{\mathrm{F}}^2. \tag{30}$$

Using a similar argument as Allen-Zhu et al. (2017, Lemma 3.1), we have

$$\hat{Y} = \arg\min_{Y \in \mathcal{S}_n, \mathrm{rank}(Y) \leq r_\star} \eta\langle \hat{Y} - X_t, \nabla f(X_t)\rangle + \frac{\eta^2\beta}{2}\|\hat{Y} - X_t\|_{\mathrm{F}}^2.$$

Hence, we can replace $\hat{Y}$ in (30) by $X_\star$ in the following step (a),

$$\begin{aligned}
f(X_{t+1}) &\overset{(a)}{\leq} f(X_t) + \eta\langle X_\star - X_t, \nabla f(X_t)\rangle + \frac{\eta^2\beta}{2}\|X_\star - X_t\|_{\mathrm{F}}^2\\
&\overset{(b)}{\leq} f(X_t) - \eta(f(X_t) - f(X_\star)) + \frac{\eta^2\beta}{2\gamma}(f(X_t) - f(X_\star)),
\end{aligned} \tag{31}$$

where step (b) is due to the qudratic growth of Problem (1). Now subtract both sides by $f(X_\star)$, and let $h_t = f(X_t) - f(X_\star)$ for each t, we find that

$$h_{t+1} \leq (1 - \eta + \frac{\eta^2\beta}{2\gamma})h_t.$$

Our choice $\eta = \frac{\gamma}{\beta}$ set $(1 - \eta + \frac{\eta^2\beta}{2\gamma}) = 1 - \frac{\gamma}{2\beta}$ which is what we desired. $\square$

G. Additional Numerics

We include extra numerics for $n = 100, 200, 400$ in Figure 1, 2. As can be seen, SpecFW in these cases are a bit slower than G-BlockFW when $\tau = 0.5$ and $c = 0.5$. SpecFW is as good as FW when k is miss specified.

What if $\nabla f(X_\star) = 0$? Here we also discuss an interesting situation that $c = 0$, and $\tau = 1$, then we see $X_\star = U_\natural U_\natural^\top$ is an optimal solution and gradient in this case is 0. Such situation means strict complementarity fails and the small perturbation to τ will result in a higher-rank solution, meaning the convex relaxation (20) is ill-posed for the purpose of low-rank matrix recovery [Lemma 2](Garber, 2019). Indeed, this is where SpecFW is not advantageous comparing to G-BlockFW as shown in Figure 3. $\tau = 1$ and $c = 0$.

References

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.

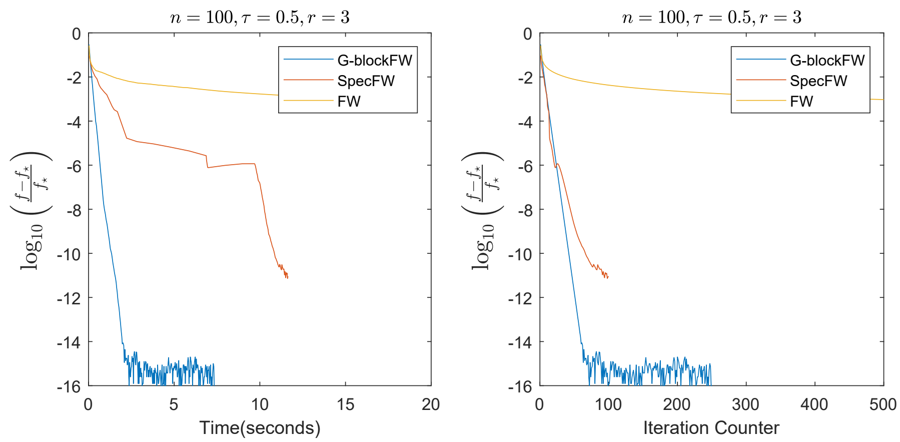
Bauschke, H. H., Borwein, J. M., and Li, W. Strong conical hull intersection property, bounded linear regularity, jameson’s property (g), and error bounds in convex optimization. *Mathematical Programming*, 86(1):135–160, 1999.

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

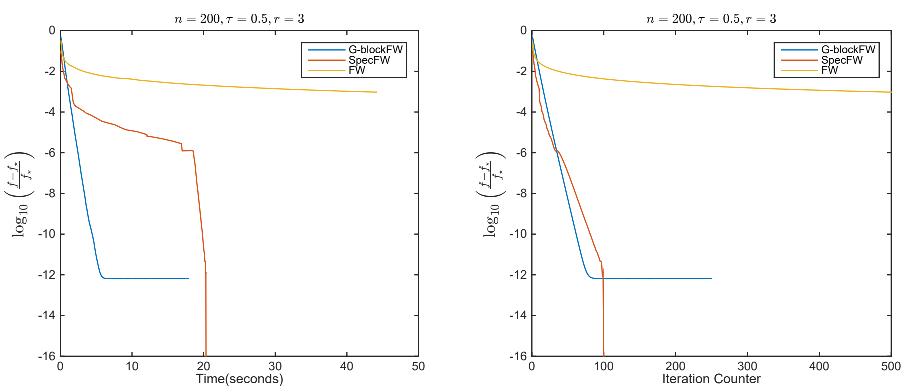
Garber, D. Linear convergence of frank-wolfe for rank-one matrix recovery without strong convexity. *arXiv preprint arXiv:1912.01467*, 2019.

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.

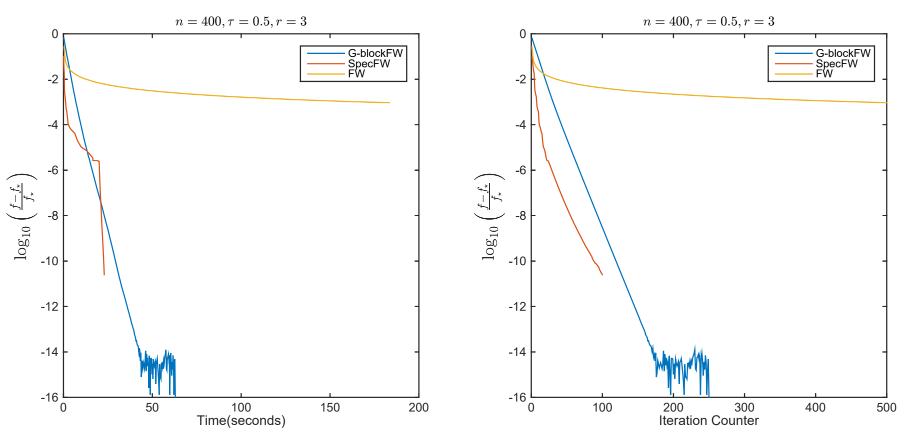
Yurtsever, A., Udell, M., Tropp, J., and Cevher, V. Sketchy decisions: Convex low-rank matrix optimization with optimal storage. In *Artificial Intelligence and Statistics*, pp. 1188–1196, 2017.



(a) $n = 100$

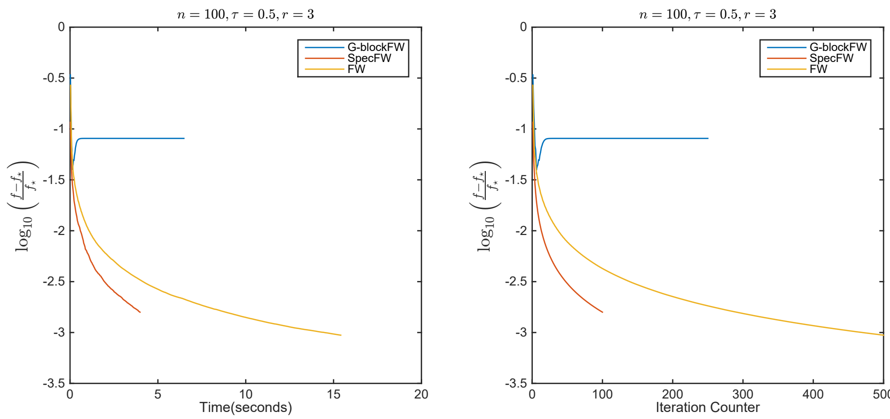


(b) $n = 200$

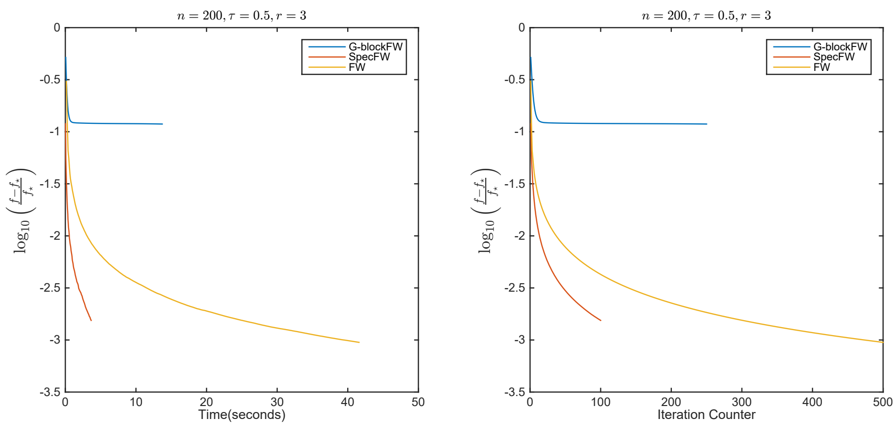


(c) $n = 400$

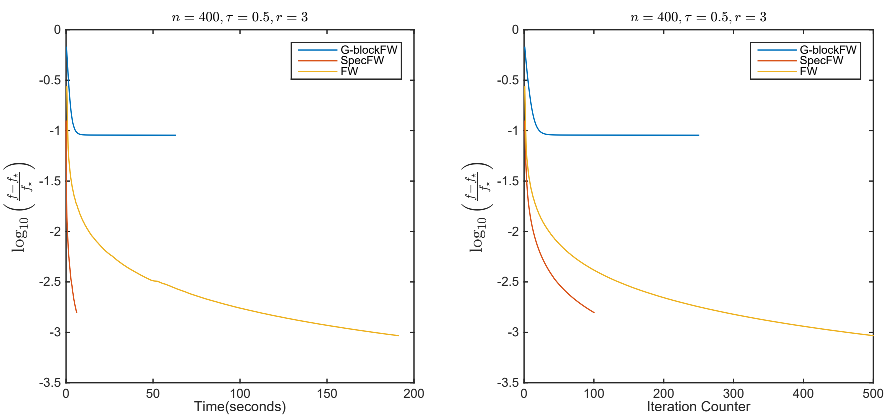
Figure 1. Comparison of algorithms under $\tau = \frac{1}{2}$ and noise level $c = 0.5$.



(a) $n = 100$

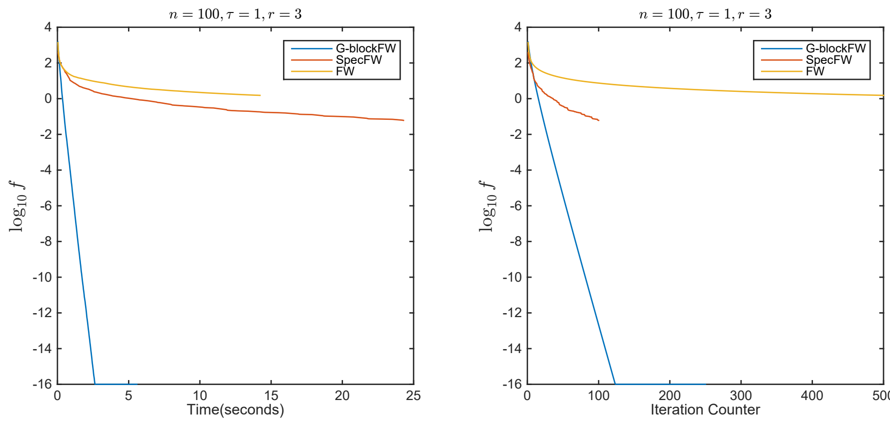


(b) $n = 200$

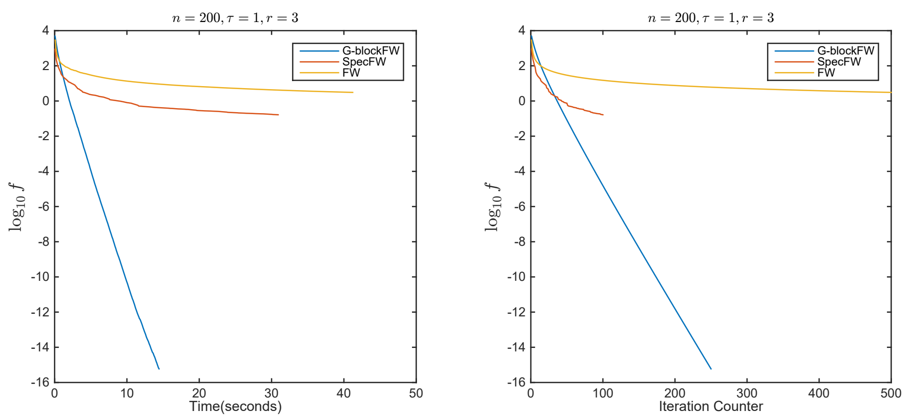


(c) $n = 400$

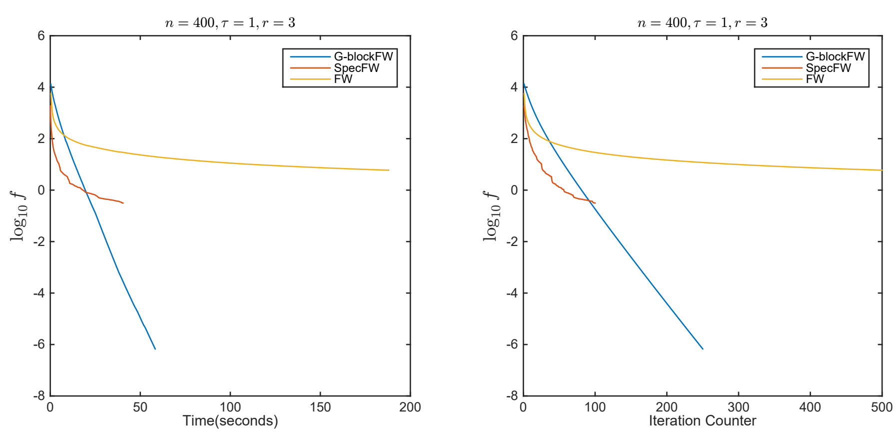
Figure 2. Comparison of algorithms under $\tau = \frac{1}{2}$, noise level $c = 0.5$, and $k = 2 < r_\star$.



(a) $n = 100$



(b) $n = 200$



(c) $n = 400$

Figure 3. Comparison of algorithms under $\tau = 1$, noise level $c = 0$, and $k = 4 > r_\star$.