Supplementary Materials for "Riemannian Pursuit for Big Matrix Recovery"

Mingkui Tan, School of Computer Science, The University of Adelaide, Australia TANMINGKUI@GMAIL.COM
Ivor W. Tsang, QCIS, University of Technology Sydney, Australia IVOR.TSANG@GMAIL.COM
Li Wang, Department of Mathematics, University of California San Diego, USA LIW022@UCSD.EDU
Bart Vandereycken, Department of Mathematics, Princeton University, USA BARTV@PRINCETON.EDU
Sinno Jialin Pan, Institute for Infocomm Research, Singapore JSPAN@I2R.A-STAR.EDU.SG

Abstract

In this supplementary file, we first present the parameter setting for ρ, and then present the proof of the lemmas and theorems appeared in the main paper.

1. Parameter Setting for ρ

In the main paper, we present (14) as a simple and effective method to choose ρ, which is motivated by the thresholding strategy in StOMP for sparse signal recovery (Donoho et al., 2012). Specifically, let $\boldsymbol{\sigma}$ be the singular vector of $\mathcal{A}^*(\mathbf{b})$, where σ_i is arranged in descending order, we choose ρ such that

$$\sigma_i \geq \eta\sigma_1, \ \ \forall i \leq \rho, \tag{1}$$

where $\eta \geq 0.60$ is usually a good choice.

However, it is not trivial to predict the number of singular values that satisfy (1) for big matrices if we do not want to compute a full SVD. Since ρ in general is small, we propose to compute σ_i sequentially until condition (1) is violated. Let $B \geq 1$ be a small integer. We propose to compute B singular values per iteration. Basically, if $\sigma_i > \eta\sigma_1$ (where $i \geq 2$), we can compute the singular values $\sigma_{i+1}, ..., \sigma_{i+B}$ by performing a rank B truncated SVD on $\mathbf{A}_i = \mathcal{A}^*(\mathbf{b}) - \sum_{j=1}^{i} \sigma_j \mathbf{u}_j \mathbf{v}_j^{\mathsf{T}}$ using PROPACK. In practice, we suggest setting $B \geq 2$. The schematic of the *Sequential Truncated SVD for Setting* ρ is presented in Algorithm 1. Notice that, PROPACK involves only matrix-vector product with $\mathbf{A}_i$ and $\mathbf{A}_i^{\mathsf{T}}$ which can be calculated as $\mathbf{U}\text{diag}(\boldsymbol{\sigma})\mathbf{V}^{\mathsf{T}}\mathbf{r}$ by $\mathbf{U}\text{diag}(\boldsymbol{\sigma})(\mathbf{V}^{\mathsf{T}}\mathbf{r})$ for the low-rank term in $\mathbf{A}_i$. We remark that instead of Algorithm 1, a more efficient technique may involve restarting the Krylov-based method, like PROPACK, with an increasingly larger subspace until (1) is satisfied.

Algorithm 1 Sequential Truncated SVD for Setting ρ.

1: Given η and $\mathcal{A}^*(\mathbf{b})$, initialize $\rho = 2$ and $B > 1$.
2: Do the rank-2 truncated SVD on $\mathcal{A}^*(\mathbf{b})$, obtaining $\boldsymbol{\sigma} \in \mathbb{R}^2$, $\mathbf{U} \in \mathbb{R}^{m\times 2}$ and $\mathbf{V} \in \mathbb{R}^{n\times 2}$.
3: If $\sigma_\rho \geq \eta\sigma_1$, stop and return $\rho = 2$.
4: **while** $\sigma_\rho < \eta\sigma_1$ **do**
5:    Let $\rho = \rho + B$.
6:    Do the rank-B truncated SVD on $\mathcal{A}^*(\mathbf{b}) - \mathbf{U}\text{diag}(\boldsymbol{\sigma})\mathbf{V}^{\mathsf{T}}$, obtaining $\boldsymbol{\sigma}^B \in \mathbb{R}^B$, $\mathbf{U}_B \in \mathbb{R}^{m\times B}$ and $\mathbf{V}_B \in \mathbb{R}^{n\times B}$.
7:    Let $\mathbf{U} = [\mathbf{U}\ \mathbf{U}_B]$, $\mathbf{V} = [\mathbf{V}\ \mathbf{V}_B]$ and $\boldsymbol{\sigma} = [\boldsymbol{\sigma}\ \boldsymbol{\sigma}^B]$.
8: **end while**
9: Return ρ.

Proceedings of the 31^{st} International Conference on Machine Learning, Beijing, China, 2014. JMLR: W&CP volume 32.

2. Main Theoretical Results in the Paper

We first repeat the main results in the paper before we prove Lemma 1 and Theorem 1 (the other results were already proven in the paper).

Proposition 1. *In MC, suppose the observed entry set* Ξ *is sampled according to the* Bernoulli model *with each entry* $(i,j) \in \Xi$ *being independently drawn from a probability* p*. There exists a constant* $C > 0$*, for all* $\gamma_r \in (0,1)$*,* $\mu_B \geq 1$*,* $n \geq m \geq 3$*, if* $p \geq C\mu_B^2 r^2 \log(n)/(\gamma_r^2 m)$*, the following RIP condition holds*

$$(1-\gamma_r)p||\mathbf{X}||_F^2 \leq ||\mathcal{P}_\Xi(\mathbf{X})||_F^2 \leq (1+\gamma_r)p||\mathbf{X}||_F^2, \tag{2}$$

for any μ_B*-incoherent matrix* $\mathbf{X} \in \mathbb{R}^{m\times n}$ *of rank at most* r *with probability at least* $1 - \exp(-n\log n)$*.*

Lemma 1. *Let* $\{\mathbf{X}^t\}$ *be the sequence generated by RP, then*

$$f(\mathbf{X}^t) \leq f(\mathbf{X}^{t-1}) - \frac{\tau_t}{2}||\mathbf{H}_2^t||_2^2. \tag{3}$$

where τ_t *satisfies condition in (10) of the paper.*

Theorem 1. *Let* $\{\mathbf{X}^t\}$ *be the sequence generated by RP and* $\zeta = \min\{\tau_1, \cdots, \tau_\iota\}$*. As long as* $f(\mathbf{X}^t) \geq \frac{C}{2}||\mathbf{e}||^2$ *(where* $C > 1$*) and if there exists an integer* $\iota > 0$ *such that* $\gamma_{(\widehat{r}+2\iota\rho)} < \frac{1}{2}$*, then* RP *decreases linearly in objective values when* $t < \iota$*, namely* $f(\mathbf{X}^{t+1}) \leq \nu f(\mathbf{X}^t)$*, where*

$$\nu = 1 - \frac{\rho\zeta}{2\widehat{r}}\left(\frac{C(1-2\gamma_{(\widehat{r}+2\iota\rho)})^2}{(\sqrt{C}+1)^2(1-\gamma_{(\widehat{r}+2\iota\rho)})}\right)\left(1 - \frac{1}{\sqrt{C}}\right)^2.$$

Proposition 2 (Sato & Iwai (2013))**.** *Given the retraction (8) and vector transport (20) on* $\mathcal{M}_r$ *in the paper, there exists a step size* θ_k *that satisfies the strong Wolfe conditions (17) and (18) of the paper.*

Lemma 2. *If* $c_2 < \frac{1}{2}$*, then the search direction* $\boldsymbol{\zeta}_k$ *generated by* NRCG *with Fletcher-Reeves rule and strong Wolfe step size control satisfies*

$$-\frac{1}{1-c_2} \leq \frac{\langle \mathrm{grad}f(\mathbf{X}_k), \boldsymbol{\zeta}_k\rangle}{\langle \mathrm{grad}f(\mathbf{X}_{k-1}), \mathrm{grad}f(\mathbf{X}_{k-1})\rangle} \leq \frac{2c_2-1}{1-c_2}. \tag{4}$$

Theorem 2. *Let* $\{\mathbf{X}_k\}$ *be the sequence generated by NRCG with the strong Wolfe line search, where* $0 < c_1 < c_2 < 1/2$*, we have* $\lim_{k\to\infty} \inf \mathrm{grad}f(\mathbf{X}_k) = \mathbf{0}$.

3. Proof of Lemma 1 in the Paper

The step size τ_t is determined such that

$$f(R_{\mathbf{X}}(-\tau_t\mathbf{H}^{t-1})) \leq f(\mathbf{X}^{t-1}) - \frac{\tau_t}{2}\langle \mathbf{H}^{t-1}, \mathbf{H}^{t-1}\rangle.$$

Since $\langle \mathbf{H}_1^{t-1}, \mathbf{H}_2^{t-1}\rangle = 0$, it follows that

$$\begin{aligned} f(\mathbf{X}^t) &\leq f(\mathbf{X}^{t-1}) - \frac{\tau_t}{2}||\mathbf{H}_1^{t-1}||_2^2 - \frac{\tau_t}{2}||\mathbf{H}_2^{t-1}||_2^2 \\ &\leq f(\mathbf{X}^{t-1}) - \frac{\tau_t}{2}||\mathbf{H}_2^{t-1}||_2^2, \end{aligned}$$

where the equality holds when $\mathbf{H}_1^{t-1} = 0$, which happens if we solve the master problem exactly.

4. Proof of Theorem 1 in the Paper

4.1. Key Lemmas

To complete the proof of Theorem 1, we first recall a property of the orthogonal projection $P_{T_{\mathbf{X}}\mathcal{M}_r}(\mathbf{Z})$.

Lemma 3. *Given the orthogonal projection* $P_{T_{\mathbf{X}}\mathcal{M}_r}(\mathbf{Z}) = P_U\mathbf{Z}P_V + P_U^\perp\mathbf{Z}P_V + P_U\mathbf{Z}P_V^\perp$*, we have rank*$(P_{T_{\mathbf{X}}\mathcal{M}_r}(\mathbf{Z})) \leq 2\min($*rank*$(\mathbf{Z}),$ *rank*$(P_U))$ *for any* $\mathbf{X}$*. In addition, we have* $||P_{T_{\mathbf{X}}\mathcal{M}_r}(\mathbf{Z})||_F \leq ||\mathbf{Z}||_F$ *for any* $\mathbf{Z} \in \mathbb{R}^{m\times n}$*.*

Proof. According to (Shalit et al., 2012), the three terms in $P_{T_{\mathbf{X}}\mathcal{M}_r}(\mathbf{Z})$ are orthogonal to each other. Since $\text{rank}(P_U) = \text{rank}(P_V)$, we have

$$\begin{aligned}\text{rank}(P_{T_{\mathbf{X}}\mathcal{M}_r}(\mathbf{Z})) &= \text{rank}(\mathbf{Z}P_V + P_U\mathbf{Z}P_V^{\perp}) \\ &\leq \min(\text{rank}(\mathbf{Z}), \text{rank}(P_V)) + \min(\text{rank}(\mathbf{Z}), \text{rank}(P_U)) \\ &= 2\min(\text{rank}(\mathbf{Z}), \text{rank}(P_U)).\end{aligned}$$

The relation $||P_{T_{\mathbf{X}}\mathcal{M}_r}(\mathbf{Z})||_F \leq ||\mathbf{Z}||_F$ follows immediately form the fact that $P_{T_{\mathbf{X}}\mathcal{M}_r}$ is an orthogonal projection for the Frobenius norm. □

4.2. Notation

We first introduce some notation.

First, let $\widehat{\mathbf{X}}$ and $\mathbf{e}$ be the ground-truth low-rank matrix and additive noise, respectively. Moreover, let $\{\mathbf{X}^t\}$ be the sequence generated by RP, $\boldsymbol{\xi}^t = \mathcal{A}(\mathbf{X}^t) - \mathbf{b}$ and $\mathbf{G}^t = \mathcal{A}^*(\boldsymbol{\xi}^t)$. In RP, we solve the fixed-rank subproblem approximately by the NRCG method. Recall the definition of the orthogonal projection onto the tangent space of $\mathbf{X} = \mathbf{USV}^{\mathsf{T}}$

$$P_{T_{\mathbf{X}}\mathcal{M}_r}(\mathbf{Z}) := \mathcal{P}_{T_{\mathbf{X}}}(\mathbf{Z}) = P_U\mathbf{Z}P_V + P_U^{\perp}\mathbf{Z}P_V + P_U\mathbf{Z}P_V^{\perp} = P_U\mathbf{Z} + \mathbf{Z}P_V - P_U\mathbf{Z}P_V,$$

where $P_U = \mathbf{UU}^{\mathsf{T}}$ and $P_V = \mathbf{VV}^{\mathsf{T}}$. In addition, denote the projection $\mathcal{P}_{T_{\mathbf{X}}}^{\perp}$ as the complement of $\mathcal{P}_{T_{\mathbf{X}}}$ as

$$\mathcal{P}_{T_{\mathbf{X}}}^{\perp} = (\mathbf{I} - P_U)\mathbf{Z}(\mathbf{I} - P_V).$$

Now recalling

$$\mathbf{E}_t = P_{T_{\mathbf{X}^t}\mathcal{M}_{t\rho}}(\mathbf{G}^t) = P_{T_{\mathbf{X}^t}\mathcal{M}_{t\rho}}(\mathcal{A}^*(\boldsymbol{\xi}^t)),$$

we have

$$\langle \mathbf{X}^t, \mathcal{A}^*(\boldsymbol{\xi}^t)\rangle = \langle \mathbf{X}^t, \mathbf{E}_t\rangle. \tag{5}$$

At the tth iteration, $\text{rank}(\mathbf{X}^t) = t\rho$, thus $\mathbf{X}^t - \widehat{\mathbf{X}}$ is at most of rank $(\widehat{r} + t\rho)$ where $\widehat{r} = \text{rank}(\widehat{\mathbf{X}})$. By the orthogonal projection $\mathcal{P}_{T_{\mathbf{X}}}$, we decompose $\widehat{\mathbf{X}}$ into two mutually orthogonal matrices

$$\widehat{\mathbf{X}} = \widehat{\mathbf{X}}_{C_t} + \widehat{\mathbf{X}}_{Q_t}, \qquad \text{where } \widehat{\mathbf{X}}_{C_t} = \mathcal{P}_{T_{\mathbf{X}^t}}(\widehat{\mathbf{X}}) \text{ and } \widehat{\mathbf{X}}_{Q_t} = \mathcal{P}_{T_{\mathbf{X}^t}}^{\perp}(\widehat{\mathbf{X}}). \tag{6}$$

Based on the above decomposition, it follows that

$$\langle \mathbf{X}^t - \widehat{\mathbf{X}}_{C_t}, \widehat{\mathbf{X}}_{Q_t}\rangle = 0.$$

Without loss of generality, we assume that $\widehat{r} \geq t\rho$. According to Lemma 3, we have $\text{rank}(\widehat{\mathbf{X}}_{C_t}) \leq 2\min(\text{rank}(\widehat{\mathbf{X}}), \text{rank}(P_U)) = 2t\rho$ and $\text{rank}(\widehat{\mathbf{X}}_{Q_t}) \leq \widehat{r}$. Moreover, since the column and row space of $\mathbf{X}^t$ is contained in $\widehat{\mathbf{X}}_{C_t} = \mathcal{P}_{T_{\mathbf{X}^t}}(\widehat{\mathbf{X}})$, we have $\text{rank}(\mathbf{X}^t - \mathcal{P}_{T_{\mathbf{X}}}(\widehat{\mathbf{X}})) \leq 2t\rho$.

Note that, at the $(t+1)$th iteration of RP, we increase the rank of $\mathbf{X}^t$ by ρ by performing a line search using

$$\mathbf{H}^t = \mathbf{H}_1^t + \mathbf{H}_2^t, \qquad \text{where } \mathbf{H}_1^t = \mathcal{P}_{T_{\mathbf{X}^t}}(\mathbf{G}^t) \text{ and } \mathbf{H}_2^t = \mathbf{U}_\rho \text{diag}(\boldsymbol{\sigma}_\rho)\mathbf{V}_\rho^{\mathsf{T}} \in \text{Ran}\,\mathcal{P}_{T_{\mathbf{X}^t}}^{\perp}.$$

For convenience in description, we define an internal variable $\mathbf{Z}^{t+1} \in \mathbb{R}^{m\times n}$ as:

$$\mathbf{Z}^{t+1} = \mathbf{X}^t - \tau_t\mathbf{G}^t, \tag{7}$$

where τ_t is the step size used in (10) in the paper. Based on the decomposition of $\mathbf{H}^t$, we decompose $\mathbf{Z}^{t+1}$ into

$$\mathbf{Z}^{t+1} = \mathbf{Z}_1^{t+1} + \mathbf{Z}_Q^{t+1} + \mathbf{Z}_R^{t+1}, \tag{8}$$

where

$$\begin{aligned}\mathbf{Z}_1^{t+1} &= \mathcal{P}_{T_{\mathbf{X}^t}}(\mathbf{Z}^{t+1}) = \mathbf{X}^t - \tau_t \mathcal{P}_{T_{\mathbf{X}^t}}(\mathbf{G}^t),\\ \mathbf{Z}_Q^{t+1} &= \mathcal{P}_{T_{\widehat{\mathbf{X}}_{Q_t}}}(\mathbf{Z}^{t+1}) = -\tau_t \mathcal{P}_{T_{\widehat{\mathbf{X}}_{Q_t}}}(\mathbf{G}^t),\\ \mathbf{Z}_R^{t+1} &= (I - \mathcal{P}_{T_{\mathbf{X}^t}} - \mathcal{P}_{T_{\widehat{\mathbf{X}}_{Q_t}}})(\mathbf{Z}^{t+1}).\end{aligned} \tag{9}$$

Observe that from (6), we have $\widehat{\mathbf{X}}_{Q_t}^\mathsf{T}\mathbf{X}^t = (\mathbf{X}^t)^\mathsf{T}\widehat{\mathbf{X}}_{Q_t} = 0$. Hence,

$$\mathrm{Ran}\,\mathcal{P}_{T_{\mathbf{X}^t}} \perp \mathrm{Ran}\,\mathcal{P}_{T_{\widehat{\mathbf{X}}_{Q_t}}} \perp \mathrm{Ran}(I - \mathcal{P}_{T_{\mathbf{X}^t}} - \mathcal{P}_{T_{\widehat{\mathbf{X}}_{Q_t}}}), \tag{10}$$

which implies that the three matrices from above are mutually orthogonal. Similarly, $\mathbf{G}^t$ is decomposed into three mutually orthogonal parts

$$\mathbf{G}^t = \mathbf{G}_1^t + \mathbf{G}_Q^t + \mathbf{G}_R^t, \quad \text{where } \mathbf{G}_1^t = \mathcal{P}_{T_{\mathbf{X}^t}}(\mathbf{G}^t),\ \mathbf{G}_Q^t = \mathcal{P}_{T_{\widehat{\mathbf{X}}_{Q_t}}}(\mathbf{G}^t), \text{ and } \mathbf{G}_R^t = (I - \mathcal{P}_{T_{\mathbf{X}^t}} - \mathcal{P}_{T_{\widehat{\mathbf{X}}_{Q_t}}})(\mathbf{G}^t). \tag{11}$$

4.3. Proof of Theorem 1

The proof of Theorem 1 involves three bounds for $f(\mathbf{X}^t)$ in terms of $\widehat{\mathbf{X}}_{Q_t}$, $\mathbf{Z}_Q^{t+1}$ and $||\mathbf{H}_2^t||_F$, respectively. For convenience, we first list these bounds in order to complete the proof of Theorem 1, and we will leave the detailed proof of the three bounds in Section 4.4.

First, the following Lemma gives the bound of $f(\mathbf{X}^t)$ in terms of $\widehat{\mathbf{X}}_{Q_t}$.

Lemma 4. *At the t-th iteration, if $\gamma_{(\widehat{r}+2t\rho)} < 1/2$, then*

$$f(\mathbf{X}^t) \geq \frac{1}{2}\frac{C(1-2\gamma_{(\widehat{r}+2t\rho)})^2}{(\sqrt{C}+1)^2(1-\gamma_{(\widehat{r}+2t\rho)})}||\widehat{\mathbf{X}}_{Q_t}||_F^2.$$

The following lemma bounds $f(\mathbf{X}^t)$ w.r.t. $\mathbf{Z}_Q^{t+1}$.

Lemma 5. *Suppose $||\mathbf{E}_t||_F$ is sufficiently small with $\mathbf{E}_t = \mathcal{P}_{T_{\mathbf{X}^t}}(\mathbf{G}^t)$. For $\gamma_{(\widehat{r}+2t\rho)} < 1/2$ and $C > 1$, we have*

$$\|\mathbf{Z}_Q^{t+1}\|_F^2 \geq \left(\frac{2C\tau_t^2(1-2\gamma_{(\widehat{r}+2t\rho)})^2}{(\sqrt{C}+1)^2(1-\gamma_{(\widehat{r}+2t\rho)})}\right)\left(1-\frac{1}{\sqrt{C}}\right)^2 f(\mathbf{X}^t),$$

By combining Lemma 4 and 5 from above, we shall show the following bound for $f(\mathbf{X}^t)$ w.r.t. $\mathbf{H}_2^t$.

Lemma 6. *If $\gamma_{(\widehat{r}+2t\rho)} < \frac{1}{2}$, at the t-th iteration, we have*

$$\|\mathbf{H}_2^t\|_F^2 > \frac{\rho}{\widehat{r}}\left(\frac{C(1-2\gamma_{(\widehat{r}+2t\rho)})^2}{(\sqrt{C}+1)^2(1-\gamma_{(\widehat{r}+2t\rho)})}\right)(1-\frac{1}{\sqrt{C}})^2 f(\mathbf{X}^t).$$

Proof of Theorem 1. By combining Lemma 1 and Lemma 6, we have

$$\begin{aligned} f(\mathbf{X}^{t+1}) &\leq f(\mathbf{X}^t) - \frac{\tau_t}{2}\|\mathbf{H}_2^t\|^2\\ &\leq \left(1 - \frac{\rho\tau_t}{2\widehat{r}}\left(\frac{C(1-2\gamma_{(\widehat{r}+2t\rho)})^2}{(\sqrt{C}+1)^2(1-\gamma_{(\widehat{r}+2t\rho)})}\right)\left(1-\frac{1}{\sqrt{C}}\right)^2\right) f(\mathbf{X}^t).\end{aligned}$$

The variable τ_t is a step size obtained by the line search. There should exist a ζ and $\zeta = \min\{\tau_1, \cdots, \tau_\iota\}$ such that the above relation holds for each $t < \iota$, where $\gamma_{(\widehat{r}+2\iota\rho)} < 1/2$. Note that $\frac{(1-2\gamma_{(\widehat{r}+2t\rho)})^2}{(1-\gamma_{(\widehat{r}+2t\rho)})}$ is decreasing w.r.t. $\gamma_{(\widehat{r}+2t\rho)}$ in $(0, 1/2)$. In addition, since $\gamma_{(\widehat{r}+2t\rho)} \leq \gamma_{(\widehat{r}+2\iota\rho)}$ holds for all $t \leq \iota$, the following relation holds if $\gamma_{(\widehat{r}+2\iota\rho)} < 1/2$ and $t < \iota$,

$$f(\mathbf{X}^{t+1}) \leq \left(1 - \frac{\rho\zeta}{2\widehat{r}}\left(\frac{C(1-2\gamma_{(\widehat{r}+2\iota\rho)})^2}{(\sqrt{C}+1)^2(1-\gamma_{(\widehat{r}+2\iota\rho)})}\right)\left(1-\frac{1}{\sqrt{C}}\right)^2\right) f(\mathbf{X}^t).$$

This completes the proof of Theorem 1. □

4.4. Detailed Proof of the Three Bounds

4.4.1. KEY LEMMAS

To proceed, we need to recall a property of the constant γ_r in RIP.

Lemma 7. *(Lemma 3.3 in (Candès & Plan, 2010)) For all $\mathbf{X}_P$, $\mathbf{X}_Q \in \mathbb{R}^{m\times n}$ satisfying $\langle \mathbf{X}_P, \mathbf{X}_Q\rangle = 0$, where rank$(\mathbf{X}_P) \leq r_p$, rank$(\mathbf{X}_Q) \leq r_q$,*

$$|\langle \mathcal{A}(\mathbf{X}_P), \mathcal{A}(\mathbf{X}_Q)\rangle| \leq \gamma_{r_p+r_q}\|\mathbf{X}_P\|_F\|\mathbf{X}_Q\|_F. \tag{12}$$

In addition, for any two integers $r \leq s$, then $\gamma_r \leq \gamma_s$ (Dai & Milenkovic, 2009).

Suppose $\mathbf{b}_q = \mathcal{A}(\mathbf{X}_Q)$, for some $\mathbf{X}_Q$ with rank$(\mathbf{X}_Q) = r_q$. Define $\mathbf{b}_p = \mathcal{A}(\mathbf{X}_P)$ where $\mathbf{X}_P$ is the optimal solution of the following problem

$$\min_{\mathbf{X}} \frac{1}{2}||\mathcal{A}(\mathbf{X}) - \mathcal{A}(\mathbf{X}_Q)||_F^2, \qquad \text{s.t.} \quad \text{rank}(\mathbf{X}) = r_p, \;\; \langle \mathbf{X}, \mathbf{X}_Q\rangle = 0. \tag{13}$$

Let $\mathbf{b}_r = \mathbf{b}_p - \mathbf{b}_q = \mathcal{A}(\mathbf{X}_P) - \mathcal{A}(\mathbf{X}_Q)$, then the following relation holds.

Lemma 8. *With $\mathbf{b}_q$, $\mathbf{b}_p$ and $\mathbf{b}_r$ defined above, if $\gamma_{\max(r_p,r_q)} + \gamma_{r_p+r_q} \leq 1$, then*

$$\|\mathbf{b}_p\| \leq \frac{\gamma_{r_p+r_q}}{1-\gamma_{\max(r_p,r_q)}}\|\mathbf{b}_q\|, \quad \text{and} \tag{14}$$

$$\left(1 - \frac{\gamma_{r_p+r_q}}{1-\gamma_{\max(r_p,r_q)}}\right)\|\mathbf{b}_q\| \leq \|\mathbf{b}_r\| \leq \|\mathbf{b}_q\|. \tag{15}$$

Proof. Since $\langle \mathbf{X}_Q, \mathbf{X}_P\rangle = \mathbf{0}$, with Lemma 7, we have

$$\begin{aligned} |\mathbf{b}_p^\mathsf{T}\mathbf{b}_q| &= |\langle \mathcal{A}(\mathbf{X}_P), \mathcal{A}(\mathbf{X}_Q)\rangle| \\ &\leq \gamma_{r_p+r_q}||\mathbf{X}_P||_F||\mathbf{X}_Q||_F \\ &\leq \gamma_{r_p+r_q}\frac{||\mathbf{b}_p||}{\sqrt{1-\gamma_{r_p}}}\frac{||\mathbf{b}_q||}{\sqrt{1-\gamma_{r_q}}} \\ &\leq \frac{\gamma_{r_p+r_q}}{1-\gamma_{\max(r_p,r_q)}}||\mathbf{b}_p||\,||\mathbf{b}_q||. \end{aligned}$$

Now we show that $\mathbf{b}_p^\mathsf{T}\mathbf{b}_r = 0$. Let $\mathbf{X}_P = \mathbf{U}\text{diag}(\boldsymbol{\sigma})\mathbf{V}^\mathsf{T}$. Since $\mathbf{X}_P$ is the minimizer of (13), $\boldsymbol{\sigma}$ is also the minimizer to the following problem:

$$\min_{\boldsymbol{\sigma}} \|\mathbf{b}_q - \mathbf{D}\boldsymbol{\sigma}\|_2^2, \tag{16}$$

where $\mathbf{D} = [\mathcal{A}(\mathbf{u}_1\mathbf{v}_1^\mathsf{T}), ..., \mathcal{A}(\mathbf{u}_{r_p}\mathbf{v}_{r_p}^\mathsf{T})]$. The Galerkin condition for this linear least-square system states that $\mathbf{b}^\mathsf{T}(\mathbf{D}\boldsymbol{\sigma} - \mathbf{b}_q) = 0$ for any $\mathbf{b}$ in the column span of $\mathbf{D}$. Since $\mathbf{b}_p = \mathcal{A}(\mathbf{X}_P)$ is included in the span of $\mathbf{D}$, we obtain $\mathbf{b}_p^\mathsf{T}\mathbf{b}_r = 0$.

Recall now $\mathbf{b}_r = \mathbf{b}_p - \mathbf{b}_q$. Then it follows that $|\mathbf{b}_p^\mathsf{T}\mathbf{b}_q| = |\mathbf{b}_p^\mathsf{T}(\mathbf{b}_p - \mathbf{b}_r)| = ||\mathbf{b}_p||^2$. Accordingly, we have

$$||\mathbf{b}_p|| \leq \frac{\gamma_{r_p+r_q}}{1-\gamma_{\max(r_p,r_q)}}||\mathbf{b}_q||.$$

Using the reverse triangular inequality, $||\mathbf{b}_r|| = ||\mathbf{b}_q - \mathbf{b}_p|| \geq |\; ||\mathbf{b}_q|| - ||\mathbf{b}_p||\; |$, we obtain

$$||\mathbf{b}_r|| \geq \left(1 - \frac{\gamma_{r_p+r_q}}{1-\gamma_{\max(r_p,r_q)}}\right)||\mathbf{b}_q||.$$

By the Galerkin condition of (16), we have $||\mathbf{b}_q||^2 = ||\mathbf{b}_r||^2 + ||\mathbf{b}_p||^2$, we obtain

$$\left(1 - \frac{\gamma_{r_p+r_q}}{1-\gamma_{\max(r_p,r_q)}}\right)||\mathbf{b}_q|| \leq ||\mathbf{b}_r|| \leq ||\mathbf{b}_q||.$$

Finally, the condition $\gamma_{\max(r_p,r_q)} + \gamma_{r_p+r_q} \leq 1$ is for the positiveness of $(1 - \frac{\gamma_{r_p+r_q}}{1-\gamma_{\max(r_p,r_q)}})$. This completes the proof. □

4.4.2. Proof of Lemma 4

Recall (6). Since $\mathbf{b} = \mathcal{A}(\widehat{\mathbf{X}}) + \mathbf{e}$, we have

$$\begin{aligned}
\sqrt{f(\mathbf{X}^t)} &= \sqrt{\frac{1}{2}\|\mathcal{A}(\mathbf{X}^t) - \mathbf{b}\|^2} \\
&= \frac{1}{\sqrt{2}}\|\mathcal{A}(\mathbf{X}^t - \widehat{\mathbf{X}}) - \mathbf{e}\| \\
&\geq \frac{1}{\sqrt{2}}\left(\|\mathcal{A}(\mathbf{X}^t - \widehat{\mathbf{X}})\| - \|\mathbf{e}\|\right) \\
&= \frac{1}{\sqrt{2}}\left(\|\mathcal{A}(\mathbf{X}^t - \widehat{\mathbf{X}}_{C_t}) - \mathcal{A}(\widehat{\mathbf{X}}_{Q_t})\| - ||\mathbf{e}||\right)
\end{aligned}$$

Note that $\langle \mathbf{X}^t - \widehat{\mathbf{X}}_{C_t}, \widehat{\mathbf{X}}_{Q_t}\rangle = 0$, and $\mathrm{rank}(\mathbf{X}^t - \widehat{\mathbf{X}}_{C_t}) \leq 2t\rho$. By applying Lemma 8, where we let $\mathbf{b}_q = \mathcal{A}(\widehat{\mathbf{X}}_{Q_t})$ and $\mathbf{b}_r = \mathcal{A}(\widehat{\mathbf{X}}_{Q_t}) - \mathcal{A}(\mathbf{X}_P)$ with $\mathbf{X}_P$ specified below, it follows that

$$\begin{aligned}
\sqrt{f(\mathbf{X}^t)} &\geq \frac{1}{\sqrt{2}}\left(\min_{\mathrm{rank}(\mathbf{X})=2t\rho, \langle \mathbf{X}, \widehat{\mathbf{X}}_{Q_t}\rangle=0} \|\mathcal{A}(\mathbf{X}) - \mathcal{A}(\widehat{\mathbf{X}}_{Q_t})\| - \|\mathbf{e}\|\right) \\
&\geq \frac{1}{\sqrt{2}}\left(\left(1 - \frac{\gamma_{\widehat{r}+2t\rho}}{1-\gamma_{\max(2t\rho,\widehat{r})}}\right)\|\mathcal{A}(\widehat{\mathbf{X}}_{Q_t})\| - \|\mathbf{e}\|\right) \text{ (by Lemma 8)} \\
&\geq \frac{1}{\sqrt{2}}\left(\left(1 - \frac{\gamma_{\widehat{r}+2t\rho}}{1-\gamma_{\max(2t\rho,\widehat{r})}}\right)\sqrt{1-\gamma_{\widehat{r}}}||\widehat{\mathbf{X}}_{Q_t}||_F - ||\mathbf{e}||\right) \text{ (by RIP condition)} \\
&\geq \frac{1}{\sqrt{2}}\left(\left(1 - \frac{\gamma_{\widehat{r}+2t\rho}}{1-\gamma_{\widehat{r}+2t\rho}}\right)\sqrt{1-\gamma_{\widehat{r}+2t\rho}}||\widehat{\mathbf{X}}_{Q_t}||_F - ||\mathbf{e}||\right) \text{ (by } \gamma_{(\widehat{r}+2t\rho)} \geq \gamma_{\max(2t\rho,\widehat{r})} \geq \gamma_{\widehat{r}}) \\
&\geq \frac{1}{\sqrt{2}}\left(\frac{1-2\gamma_{(\widehat{r}+2t\rho)}}{\sqrt{(1-\gamma_{(\widehat{r}+2t\rho)})}}||\widehat{\mathbf{X}}_{Q_t}||_F - ||\mathbf{e}||\right).
\end{aligned}$$

Recall that $f(\mathbf{X}^t) \geq Cf(\widehat{\mathbf{X}}) \geq \frac{C}{2}||\mathbf{e}||^2$. By rearranging the above inequality, we have

$$f(\mathbf{X}^t) \geq \frac{1}{2}\frac{C(1-2\gamma_{(\widehat{r}+2t\rho)})^2}{(\sqrt{C}+1)^2(1-\gamma_{(\widehat{r}+2t\rho)})}||\widehat{\mathbf{X}}_{Q_t}||_F^2. \tag{17}$$

This completes the proof. □

4.4.3. Proof of Lemma 5

First, we have the following bound of $f(\mathbf{X}^t)$ in terms of $\langle \widehat{\mathbf{X}}_{Q_t}, \mathbf{Z}_Q^{t+1}\rangle$ if $\|\mathbf{E}_t\|_F$ is sufficiently small.

Lemma 9. *Suppose that $\mathbf{X}^t$ is an approximate solution with $\mathbf{E}_t = \mathcal{P}_{T_{\mathbf{X}^t}}(\mathbf{G}^t)$. For $\gamma_{(\widehat{r}+2t\rho)} < 1/2$ and $\|\mathbf{E}_t\|_F$ sufficiently small, the following inequality holds at the t-th iteration:*

$$\frac{1}{2\tau_t}\langle \widehat{\mathbf{X}}_{Q_t}, \mathbf{Z}_Q^{t+1}\rangle \geq \frac{1}{2}(1-\frac{1}{C})f(\mathbf{X}^t). \tag{18}$$

Proof. Recall the decomposition (11) and let $\boldsymbol{\xi}^t = \mathcal{A}(\mathbf{X}^t) - \mathbf{b}$. Then it follows that

$$\begin{aligned}
-\frac{1}{2}\mathcal{A}(\widehat{\mathbf{X}})^{\mathsf{T}}\boldsymbol{\xi}^t &= -\frac{1}{2}\langle \widehat{\mathbf{X}}, \mathcal{A}^*(\boldsymbol{\xi}^t)\rangle \\
&= -\frac{1}{2}\left\langle \begin{bmatrix}\widehat{\mathbf{X}}_{C_t} & \widehat{\mathbf{X}}_{Q_t}\end{bmatrix}, \begin{bmatrix}\mathbf{G}_1^t \\ \mathbf{G}_Q^t\end{bmatrix}\right\rangle \qquad \text{(by (10))} \\
&= -\frac{1}{2}\langle \widehat{\mathbf{X}}_{C_t}, \mathbf{E}_t\rangle - \frac{1}{2}\langle \widehat{\mathbf{X}}_{Q_t}, \mathbf{G}_Q^t\rangle \qquad \text{(by (11))} \\
&= -\frac{1}{2}\langle \widehat{\mathbf{X}}_{C_t}, \mathbf{E}_t\rangle + \frac{1}{2\tau_t}\langle \widehat{\mathbf{X}}_{Q_t}, \mathbf{Z}_Q^{t+1}\rangle. \qquad \text{(by (9))}
\end{aligned} \tag{19}$$

Therefore, we have

$$
\begin{aligned}
f(\mathbf{X}^t) &= \frac{1}{2}\|\mathcal{A}(\mathbf{X}^t)-\mathbf{b}\|^2\\
&= \frac{1}{2}\mathcal{A}(\mathbf{X}^t)^{\mathsf{T}}\boldsymbol{\xi}^t - \frac{1}{2}\mathbf{b}^{\mathsf{T}}\boldsymbol{\xi}^t\\
&= \frac{1}{2}\langle \mathbf{X}^t, \mathcal{A}^*(\boldsymbol{\xi}^t)\rangle - \frac{1}{2}\mathbf{b}^{\mathsf{T}}\boldsymbol{\xi}^t\\
&= \frac{1}{2}\langle \mathbf{E}_t, \mathbf{X}^t\rangle - \frac{1}{2}(\mathcal{A}(\widehat{\mathbf{X}})+\mathbf{e})^{\mathsf{T}}\boldsymbol{\xi}^t \qquad \text{(by (5))}\\
&= \frac{1}{2}\langle \mathbf{E}_t, \mathbf{X}^t\rangle - \frac{1}{2}\mathcal{A}(\widehat{\mathbf{X}})^{\mathsf{T}}\boldsymbol{\xi}^t - \frac{1}{2}\mathbf{e}^{\mathsf{T}}\boldsymbol{\xi}^t\\
&= \frac{1}{2}\langle \mathbf{E}_t, \mathbf{X}^t\rangle + \frac{1}{2\tau_t}\langle \widehat{\mathbf{X}}_{Q_t}, \mathbf{Z}_Q^{t+1}\rangle - \frac{1}{2}\langle \widehat{\mathbf{X}}_{C_t}, \mathbf{E}_t\rangle - \frac{1}{2}\mathbf{e}^{\mathsf{T}}\boldsymbol{\xi}^t. \quad \text{(by (19))}
\end{aligned}
$$

Based on the assumption $f(\mathbf{X}^t) \geq Cf(\widehat{\mathbf{X}}) = \frac{C}{2}||\mathbf{e}||^2$, where $C > 1$, we then have

$$
\frac{1}{2}\left|\mathbf{e}^{\mathsf{T}}\boldsymbol{\xi}^t\right| \leq \frac{1}{2}\|\mathbf{e}\| \times \|\boldsymbol{\xi}^t\| \leq \frac{1}{2}\sqrt{\frac{2}{C}f(\mathbf{X}^t)}\sqrt{2f(\mathbf{X}^t)} = \frac{1}{\sqrt{C}}f(\mathbf{X}^t).
$$

It follows that

$$
\begin{aligned}
\frac{1}{2\tau_t}\langle \widehat{\mathbf{X}}_{Q_t}, \mathbf{Z}_Q^{t+1}\rangle &= f(\mathbf{X}^t) + \frac{1}{2}\mathbf{e}^{\mathsf{T}}\boldsymbol{\xi} - \frac{1}{2}\langle \mathbf{E}_t, \mathbf{X}_t\rangle + \frac{1}{2}\langle \widehat{\mathbf{X}}_{C_t}, \mathbf{E}_t\rangle\\
&\geq (1-\frac{1}{\sqrt{C}})f(\mathbf{X}^t) - \frac{1}{2}\langle \mathbf{E}_t, \mathbf{X}_t\rangle + \frac{1}{2}\langle \widehat{\mathbf{X}}_{C_t}, \mathbf{E}_t\rangle. \qquad (20)
\end{aligned}
$$

Suppose $|\langle \mathbf{E}_t, \widehat{\mathbf{X}}_{C_t} - \mathbf{X}_t\rangle| \leq \vartheta f(\mathbf{X}^t)$ for $\vartheta > 0$, then it follows that

$$
\begin{aligned}
\frac{1}{2\tau_t}\langle \widehat{\mathbf{X}}_{Q_t}, \mathbf{Z}_Q^{t+1}\rangle &\geq (1-\frac{1}{\sqrt{C}})f(\mathbf{X}^t) - \vartheta f(\mathbf{X}^t)\\
&\geq (1-\frac{1}{\sqrt{C}}-\vartheta)f(\mathbf{X}^t). \qquad (21)
\end{aligned}
$$

Suppose $\vartheta \ll \frac{1}{\sqrt{C}}$, we can simplify the formulation by absorbing ϑ into C as

$$
\frac{1}{2\tau_t}\langle \widehat{\mathbf{X}}_{Q_t}, \mathbf{Z}_Q^{t+1}\rangle \geq \left(1-\frac{1}{\sqrt{\widehat{C}}}\right)f(\mathbf{X}^t). \qquad (22)
$$

Let $C := \widehat{C}$, where $C > 1$, and we complete the proof. □

Proof of Lemma 5. Based on Lemma 9, we have

$$
\frac{1}{2\tau_t}\|\widehat{\mathbf{X}}_{Q_t}\|_F\|\mathbf{Z}_Q^{t+1}\|_F \geq \frac{1}{2\tau_t}\langle \widehat{\mathbf{X}}_{Q_t}, \mathbf{Z}_Q^{t+1}\rangle \geq (1-\frac{1}{\sqrt{C}})f(\mathbf{X}^t).
$$

Furthermore, with Lemma 4, it follows that

$$
\begin{aligned}
\|\mathbf{Z}_Q^{t+1}\|_F^2 &\geq \frac{4\tau_t^2}{\|\widehat{\mathbf{X}}_{Q_t}\|_F^2}\left(1-\frac{1}{\sqrt{C}}\right)^2 f(\mathbf{X}^t)^2\\
&\geq \left(\frac{2C\tau_t^2(1-2\gamma_{(\widehat{r}+2t\rho)})^2}{(\sqrt{C}+1)^2(1-\gamma_{(\widehat{r}+2t\rho)})}\right)\left(1-\frac{1}{\sqrt{C}}\right)^2 f(\mathbf{X}^t). \qquad (23)
\end{aligned}
$$

The proof is completed. □

4.4.4. Proof of Lemma 6

Recall that $\mathbf{H}_2^t$ is obtained by the truncated SVD of rank ρ on $\widehat{\mathbf{G}^t} = P^{\perp}_{T_{\mathbf{X}^t}}(\mathbf{G}^t)$, and $r_q = \text{rank}(\mathbf{Z}_Q^{t+1}) \leq 2\widehat{r}$, where $\mathbf{Z}_Q^{t+1} = -\tau_t \mathbf{G}_Q^t = -\tau_t \mathcal{P}_{T_{\widehat{\mathbf{X}}_{Q_t}}}(\widehat{\mathbf{G}})$. Let $\mathbf{H}_Q$ be the truncated SVD of rank ρ on $\mathbf{G}_Q^t = \mathcal{P}_{T_{\widehat{\mathbf{X}}_{Q_t}}}(\mathbf{G}^t)$. Since $\text{Ran}\,\mathcal{P}_{T_{\widehat{\mathbf{X}}_{Q_t}}} \subseteq \text{Ran}\,P^{\perp}_{T_{\mathbf{X}^t}}$, we have

$$||\mathbf{H}_Q||_F^2 \leq ||\mathbf{H}_2^t||_F^2.$$

Accordingly, if $\rho \leq r_q$, we have $\frac{||\mathbf{H}_2^t||_F^2}{\rho} \geq \frac{||\mathbf{H}_Q||_F^2}{\rho} \geq \frac{||\mathbf{Z}_Q^{t+1}||_F^2}{\tau_t^2 r_q}$. It follows from Lemma 5 that

$$||\mathbf{H}_2^t||_F^2 \geq \frac{2\rho}{r_q}\left(\frac{C(1-2\gamma_{(\widehat{r}+2t\rho)})^2}{(\sqrt{C}+1)^2(1-\gamma_{(\widehat{r}+2t\rho)})}\right)\left(1-\frac{1}{\sqrt{C}}\right)^2 f(\mathbf{X}^t).$$

Otherwise, if $\rho > r_q$, we have $\|\mathbf{H}_2^t\|_F \geq \frac{\|\mathbf{Z}_Q^{t+1}\|_F}{\tau_t}$, and the following inequality holds

$$||\mathbf{H}_2^t||_F^2 \geq \left(\frac{2C(1-2\gamma_{(\widehat{r}+2t\rho)})^2}{(\sqrt{C}+1)^2(1-\gamma_{(\widehat{r}+2t\rho)})}\right)\left(1-\frac{1}{\sqrt{C}}\right)^2 f(\mathbf{X}^t).$$

In summary, since $r_q \leq 2\widehat{r}$, we have $||\mathbf{H}_2^t||^2 \geq \frac{\rho}{\widehat{r}}\left(\frac{C(1-2\gamma_{(\widehat{r}+2t\rho)})^2}{(\sqrt{C}+1)^2(1-\gamma_{(\widehat{r}+2t\rho)})}\right)\left(1-\frac{1}{\sqrt{C}}\right)^2 f(\mathbf{X}^t)$. The proof is completed. □

References

Candès, E. J. and Plan, Y. Tight oracle bounds for low-rank matrix recovery from a minimal number of random measurements. *IEEE Trans. on Inform. Theory*, 57(4):2342–2359, 2010.

Dai, W. and Milenkovic, O. Subspace pursuit for compressive sensing signal reconstruction. *IEEE Trans. Inform. Theory*, 55(5):2230–2249, 2009.

Donoho, D. L., Tsaig, Y., Drori, I., and Starck, J. L. Sparse solution of underdetermined systems of linear equations by stagewise orthogonal matching pursuit. *IEEE Trans. Info. Theory*, 58(2):1094–1121, 2012.

Ring, W. and Wirth, B. Optimization methods on Riemannian manifolds and their application to shape space. *SIAM J. Optim.*, 22(2):596–627, 2012.

Sato, H. and Iwai, T. A new, globally convergent Riemannian conjugate gradient method. *Optimization: A Journal of Mathematical Programming and Operations Research*, (ahead-of-print):1–21, 2013.

Shalit, U., Weinshall, D., and Chechik, G. Online learning in the embedded manifold of low-rank matrices. *JMLR*, 13: 429–458, 2012.

Vandereycken, B. Low-rank matrix completion by Riemannian optimization. *SIAM J. Optim.*, 23(2):1214–1236, 2013.