Supplementary Materials for "Proximal Riemannian Pursuit"

A. More detailed preliminaries

In this section, we first present more details about the rank-s matrix submanifold $\mathcal{M}_s = \{\mathbf{X} \in \mathbb{R}^{m\times n} : \text{rank}(\mathbf{X}) = s\}$, and based on $\mathcal{M}_s$ we present the geometries on $\mathcal{M}_{\leq r} = \{\mathbf{X} \in \mathbb{R}^{m\times n} : \text{rank}(\mathbf{X}) \leq r\}$, where $s \leq r$.

A.1. Geometries of fixed-rank matrices $\mathcal{M}_s$

The fixed rank-s matrices lie on a smooth submanifold defined below $\mathcal{M}_s = \{\mathbf{X} \in \mathbb{R}^{m\times n} : \text{rank}(\mathbf{X}) = s\} = \{\mathbf{U}\text{diag}(\boldsymbol{\sigma})\mathbf{V}^\mathsf{T} : \mathbf{U} \in \text{St}_s^m, \mathbf{V} \in \text{St}_s^n, ||\boldsymbol{\sigma}||_0 = s\}$, where $\text{St}_s^m = \{\mathbf{U} \in \mathbb{R}^{m\times s} : \mathbf{U}^\mathsf{T}\mathbf{U} = \mathbf{I}\}$ denotes the Stiefel manifold of $m \times s$ real and orthonormal matrices, and the entries in $\boldsymbol{\sigma}$ are in descending order [51]. Moreover, the tangent space $T_\mathbf{X}\mathcal{M}_s$ at $\mathbf{X}$ is given by

$$T_\mathbf{X}\mathcal{M}_s=\{\mathbf{U}\mathbf{M}\mathbf{V}^\mathsf{T}+\mathbf{U}_p\mathbf{V}^\mathsf{T}+\mathbf{U}\mathbf{V}_p^\mathsf{T} : \mathbf{M} \in \mathbb{R}^{s\times s}, \mathbf{U}_p \in \mathbb{R}^{m\times s}, \mathbf{U}_p^\mathsf{T}\mathbf{U} = \mathbf{0}, \mathbf{V}_p \in \mathbb{R}^{n\times s}, \mathbf{V}_p^\mathsf{T}\mathbf{V} = \mathbf{0}\}. \quad (23)$$

Given $\mathbf{X} \in \mathcal{M}_s$ and $\mathbf{A}, \mathbf{B} \in T_\mathbf{X}\mathcal{M}_s$, by defining a metric $g_\mathbf{X}(\mathbf{A}, \mathbf{B}) = \langle \mathbf{A}, \mathbf{B}\rangle$, $\mathcal{M}_s$ is a **Riemannian manifold** by restricting $\langle \mathbf{A}, \mathbf{B}\rangle$ to the *tangent bundle* [2], which is defined as the disjoint union of all tangent spaces $T\mathcal{M}_s = \bigcup_{\mathbf{X}\in\mathcal{M}_s}\{\mathbf{X}\} \times T_\mathbf{X}\mathcal{M}_s$. The norm of a tangent vector $\boldsymbol{\zeta}_\mathbf{X} \in T_\mathbf{X}\mathcal{M}_s$ evaluated at $\mathbf{X}$ is defined as $||\boldsymbol{\zeta}_\mathbf{X}|| = \sqrt{\langle \boldsymbol{\zeta}_\mathbf{X}, \boldsymbol{\zeta}_\mathbf{X}\rangle}$.

Once the metric is fixed, the notion of the gradient of an objective function can be introduced. For a Riemannian manifold, the **Riemannian gradient** of a smooth function $f : \mathcal{M}_s \to \mathbb{R}$ at $\mathbf{X} \in \mathcal{M}_s$ is defined as the unique tangent vector $\text{grad}f(\mathbf{X})$ in $T_\mathbf{X}\mathcal{M}_s$, such that $\langle \text{grad}f(\mathbf{X}), \boldsymbol{\xi}\rangle = \text{D}f(\mathbf{X})[\boldsymbol{\xi}], \ \forall \boldsymbol{\xi} \in T_\mathbf{X}\mathcal{M}_s$. As $\mathcal{M}_s$ is embedded in $\mathbb{R}^{m\times n}$, the Riemannian gradient of f is given as the **orthogonal projection** of the gradient of f onto the tangent space. Here, the orthogonal projection of any $\mathbf{Z} \in \mathbb{R}^{m\times n}$ onto the tangent space $T_\mathbf{X}\mathcal{M}_s$ at $\mathbf{X} = \mathbf{U}\text{diag}(\boldsymbol{\sigma})\mathbf{V}^\mathsf{T}$ is defined as

$$P_{T_\mathbf{X}\mathcal{M}_s}(\mathbf{Z}) : \mathbf{Z} \mapsto P_U\mathbf{Z}P_V + P_U^\perp\mathbf{Z}P_V + P_U\mathbf{Z}P_V^\perp. \quad (24)$$

where $P_U = \mathbf{U}\mathbf{U}^\mathsf{T}$ and $P_U^\perp = \mathbf{I} - \mathbf{U}\mathbf{U}^\mathsf{T}$. Letting $\mathbf{G} = \nabla f(\mathbf{X})$ be the gradient of $f(\mathbf{X})$ on vector space, it follows that

$$\text{grad}f(\mathbf{X}) = P_{T_\mathbf{X}\mathcal{M}_s}(\mathbf{G}). \quad (25)$$

A *Retraction* mapping on $\mathcal{M}_s$ relates an element in the tangent space to a corresponding point on the manifold. An retraction mapping is actually an approximated Riemannina exp mapping at the first order. In this paper, for a given tangent vector $\boldsymbol{\xi}$ at $\mathbf{X}$, we will make use of the following projection operator as the retraction mapping [2]. One of the issues associated with such retraction mappings is to find the best rank-s approximation to $\mathbf{X} + \boldsymbol{\xi}$ in terms of the Frobenius norm

$$\begin{aligned} R_\mathbf{X}(\boldsymbol{\xi}) =& P_{\mathcal{M}_s}(\mathbf{X} + \boldsymbol{\xi}) \\ =& \arg\min_{\mathbf{Y}\in\mathcal{M}_s} ||\mathbf{Y} - (\mathbf{X} + \boldsymbol{\xi})||_F. \end{aligned} \quad (26)$$

where $\mathbf{X} + \boldsymbol{\xi}$ is defined on the vector space $\mathbb{R}^{m\times n}$. $R_\mathbf{X}(\boldsymbol{\xi})$ can be efficiently computer according to Algorithm 1 in the main paper.

A.2. Variety of low-rank matrices $\mathcal{M}_{\leq r}$

Given an integer $r \geq s \geq 0$, it would be more convenient to consider the closure of $\mathcal{M}_r$:

$$\mathcal{M}_{\leq r} = \{\mathbf{X} \in \mathbb{R}^{m\times n} : \text{rank}(\mathbf{X}) \leq r\}, \tag{27}$$

which is a real-algebraic variety [44]. Let $\text{ran}(\mathbf{X})$ be the column space of $\mathbf{X}$. In the singular points where rank$(\mathbf{X}) = s < r$, we will construct search directions in the tangent cone [44] (instead of the tangent space)

$$T_{\mathbf{X}}\mathcal{M}_{\leq r} = T_{\mathbf{X}}\mathcal{M}_s \oplus \{\Xi_{r-s} \in \mathcal{U}^{\perp} \otimes \mathcal{V}^{\perp}\}, \tag{28}$$

where $\mathcal{U} = \text{ran}(\mathbf{X})$ and $\mathcal{V} = \text{ran}(\mathbf{X}^{\mathsf{T}})$. Essentially, Ξ_{r-s} is a best rank-$(r-s)$ approximation of $\mathbf{G} - P_{T_{\mathbf{X}}\mathcal{M}_s}(\mathbf{G})$, which can be cheaply computed with truncated SVD of rank $(r-s)$. Let grad$f(\mathbf{X}) \in T_{\mathbf{X}}\mathcal{M}_{\leq r}$ be the projection of $\mathbf{G}$ on $T_{\mathbf{X}}\mathcal{M}_{\leq r}$. It can be computed by

$$\text{grad} f(\mathbf{X}) = P_{T_{\mathbf{X}}\mathcal{M}_s}(\mathbf{G}) + \Xi_{r-s}. \tag{29}$$

Given a search direction $\boldsymbol{\xi} \in T_{\mathbf{X}}\mathcal{M}_{\leq r}$, we need perform retraction which finds the best approximation by a matrix of rank at most r as measured in terms of the Frobenius norm, *i.e.*,

$$R^{\leq r}_{\mathbf{X}}(\boldsymbol{\xi}) = \arg\min\nolimits_{\mathbf{Y}\in\mathcal{M}_{\leq r}} ||\mathbf{Y} - (\mathbf{X} + \boldsymbol{\xi})||_F. \tag{30}$$

Since $\Xi_{r-s} \in \mathcal{U}^{\perp} \otimes \mathcal{V}^{\perp}$, $R^{\leq r}_{\mathbf{X}}(\boldsymbol{\xi})$ w.r.t. $\mathcal{M}_{\leq r}$ can be efficiently computed with the same complexity as on $\mathcal{M}_r$. In general, problem (30) can be addressed by performing SVD on $\mathbf{X} + \boldsymbol{\xi}$, which may be computationally expensive.

A.3. Computation of $R^{\leq r}_{\mathbf{X}}(\boldsymbol{\xi})$ on $\mathcal{M}_{\leq r}$

Essentially, Ξ_{r-s} is the best rank-$(r-s)$ approximation of $\mathbf{G} - P_{T_{\mathbf{X}}\mathcal{M}_s}(\mathbf{G})$ (which can be cheaply computed using truncated SVD of rank $r-s$). In other words, Ξ_{r-s} is orthogonal to $\mathbf{G} - P_{T_{\mathbf{X}}\mathcal{M}_s}(\mathbf{G})$. Let $\Xi_s = P_{T_{\mathbf{X}}\mathcal{M}_s}(\mathbf{G}) = \mathbf{U}\mathbf{M}\mathbf{V}^{\mathsf{T}} + \mathbf{U}_p\mathbf{V}^{\mathsf{T}} + \mathbf{U}\mathbf{V}_p^{\mathsf{T}}$, $\mathbf{X} = \mathbf{U}\text{diag}(\boldsymbol{\sigma})\mathbf{V}^{\mathsf{T}} \in \mathcal{M}_s$ and $\boldsymbol{\xi} = \Xi_s + \Xi_{r-s} \in T_{\mathbf{X}}\mathcal{M}_{\leq r}$, where $\Xi_s \in T_{\mathbf{X}}\mathcal{M}_s$ and $\Xi_{r-s} = \mathbf{U}_s\text{diag}(\boldsymbol{\sigma}_s)\mathbf{V}_s^{\mathsf{T}}$. $\mathbf{X} + \boldsymbol{\xi}$ can be written as $[\mathbf{U} \;\; \mathbf{U}_p]\begin{pmatrix}\text{diag}(\boldsymbol{\sigma})+\mathbf{M} & \mathbf{I}_s \\ \mathbf{I}_s & \mathbf{0}\end{pmatrix}[\mathbf{V} \;\; \mathbf{V}_p]^{\mathsf{T}} + \Xi_{r-s}$, where Ξ_{r-s} is orthogonal to first term. With these relations, $R^{\leq r}_{\mathbf{X}}(\boldsymbol{\xi})$ can be calculated via Algorithm 1.

B. Proof of Remark 2

Proof. When updating $\mathbf{X}$ with fixed $\mathbf{E} = \mathbf{E}^{t-1}$, the step size L_t is determined such that

$$\Psi(T_{L_t}(\mathbf{X}^t), \mathbf{E}) \leq \Psi(\mathbf{X}^{t-1}, \mathbf{E}^{t-1}) + \beta\langle \text{grad}(\mathbf{X}^{t-1}, \mathbf{E}^{t-1}), \boldsymbol{\zeta}_{t-1}\rangle / L_t.$$

In PRP, we choose $\boldsymbol{\zeta}_{t-1} = -\text{grad}(\mathbf{X}^{t-1}, \mathbf{E}^{t-1})$. Thus we have

$$\Psi(T_{L_t}(\mathbf{X}^t), \mathbf{E}^{t-1}) \leq \Psi(\mathbf{X}^{t-1}, \mathbf{E}^{t-1}) - \beta\langle \text{grad}(\mathbf{X}^{t-1}, \mathbf{E}^{t-1}), \text{grad}(\mathbf{X}^{t-1}, \mathbf{E}^{t-1})\rangle / L_t.$$

Note that grad$f(\mathbf{X}^{t-1}, \mathbf{E}^{t-1}) = P_{T_{\mathbf{X}^{t-1}}\mathcal{M}_s}(\mathbf{G}) + \Xi_{\kappa}^{t-1}$ (see Step 6), and $\langle P_{T_{\mathbf{X}^{t-1}}\mathcal{M}_s}(\mathbf{G}), \Xi_{\kappa}^{t-1}\rangle = 0$. It follows that

$$\Psi(T_{L_t}(\mathbf{X}^t), \mathbf{E}^{t-1}) \leq \Psi(\mathbf{X}^{t-1}, \mathbf{E}^{t-1}) - \beta||\Xi_{\kappa}^{t-1}||_F^2 / L_t. \tag{31}$$

According to Algorithm 2, $\Psi(T_{L_t}(\mathbf{X}^t), \mathbf{E}^{t-1}) = \Psi(\mathbf{X}_0^t, \mathbf{E}^{t-1})$. Due to the thresholding on $\mathbf{E}$, we have $\Psi(\mathbf{X}_0^t, \mathbf{E}_0^t) \leq \Psi(\mathbf{X}_0^t, \mathbf{E}^{t-1})$. Note that $(\mathbf{X}_0^t, \mathbf{E}_0^t)$ is the starting point of PRG(R). It follows that $\Psi(\mathbf{X}^t, \mathbf{E}^t) \leq \Psi(\mathbf{X}_0^t, \mathbf{E}_0^t) \leq \Psi(\mathbf{X}_0^t, \mathbf{E}^{t-1}) = \Psi(T_{L_t}(\mathbf{X}^t), \mathbf{E}^{t-1}) \leq \Psi(\mathbf{X}^{t-1}, \mathbf{E}^{t-1}) - \beta||\Xi_{\kappa}^{t-1}||_F^2 / L_t$. This completes the proof. $\square$

C. Proof of Lemma 1

Proof. Recall that $\mathbf{X} = \mathbf{Y} + \boldsymbol{\xi}$, where $\mathbf{X}$ lies on the tangent cone $T_{\mathbf{Y}}\mathcal{M}_{\leq r}$ at $\mathbf{Y}$, as illustrated in Figure 3.

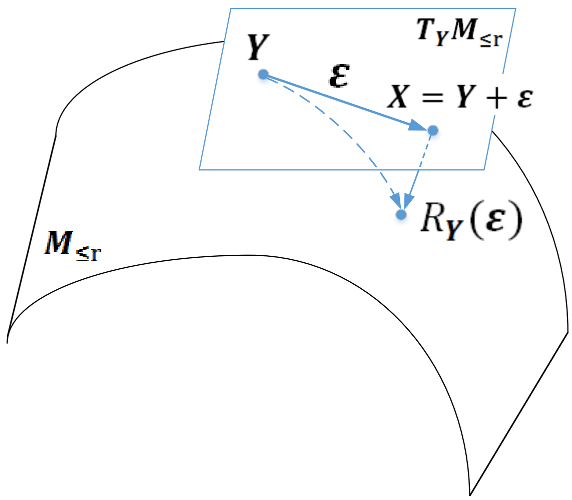

Figure 3. Illustration of Retraction $R_{\mathbf{Y}}(\boldsymbol{\xi})$ on $\mathcal{M}_{\leq r}$.

On the other hand, it is not difficult to verify that

$$
\begin{aligned}
T_L(\mathbf{Y}) \quad &= \arg \min_{\mathbf{X} \in \mathcal{M}_{\leq r}} ||\mathbf{X}||_* + f(\mathbf{Y}) + \langle \mathrm{grad} f(\mathbf{Y}), \boldsymbol{\xi} \rangle + \frac{L}{2} \langle \boldsymbol{\xi}, \boldsymbol{\xi} \rangle \\
&= \arg \min_{\mathbf{X} \in \mathcal{M}_{\leq r}} ||\mathbf{X}||_* + \frac{L}{2} ||\mathbf{X} - \mathbf{Y} + \frac{1}{L} \mathrm{grad} f(\mathbf{Y})||^2, \qquad (32)
\end{aligned}
$$

where we use the fact that $\mathbf{X} = \mathbf{Y} + \boldsymbol{\xi}$ which is restricted on the tangent cone $T_{\mathbf{Y}}\mathcal{M}_{\leq r}$. Let $\mathbf{Z} = \mathbf{Y} - 1/L \mathrm{grad} f(\mathbf{Y})$ and

$$
Q(\mathbf{X}) = f(\mathbf{Y}) + \langle \mathrm{grad} f(\mathbf{Y}), \boldsymbol{\xi} \rangle + L/2 \langle \boldsymbol{\xi}, \boldsymbol{\xi} \rangle,
$$

which is a smooth function. Clearly, $\mathbf{Z}$ is a minimizer of $Q(\mathbf{X})$ when $\boldsymbol{\xi}$ is restricted to $T_{\mathbf{Y}}\mathcal{M}_{\leq r}$, thus $R_{\mathbf{Y}}(\boldsymbol{\xi})$ is a minimizer of $Q(\mathbf{X})$ when $\mathbf{X}$ restricted on $\mathcal{M}_{\leq r}$. This implies that $\mathrm{grad}\Phi(R_{\mathbf{Y}}(\boldsymbol{\xi})) = \mathbf{0}$. In fact, $R_{\mathbf{Y}}(\boldsymbol{\xi})$ is the basic update rule in [51, 49], where the objective function is smooth.

For the non-smooth objective function in (32), following [6], we can show that, there exists $\boldsymbol{\zeta} \in \partial ||\mathbf{X}||_*$ such that $\mathrm{grad}\Phi(T_L(\mathbf{Y})) + \boldsymbol{\zeta} = \mathbf{0}$, *i.e.*, $T_L(\mathbf{Y})$ satisfies the local optimality condition of (17).

On the other hand, from the computation of $T_L(\mathbf{Y})$, we immediately have $\mathrm{rank}(T_L(\mathbf{Y})) \leq \mathrm{rank}(R_{\mathbf{Y}}(\boldsymbol{\xi})) \leq r$. In other words, it is a feasible solution. This completes the proof. $\square$

D. Proof of Lemma 2

Proof. Since $\boldsymbol{\zeta}_k$ is a descent direction, it follows that $\mathbf{0} \notin \mathrm{grad} f(\mathbf{X}_k) + \partial ||\mathbf{X}||_*$ and $\langle \mathrm{grad} f(\mathbf{X}_k), \boldsymbol{\zeta}_k \rangle < 0$. Note that $\Psi(\mathbf{X})$ is bounded below. Since $T_L(\mathbf{X}_k)$ is continuous in L, there must exist an $\widehat{L}$ such that $\Psi(T_L(\mathbf{X}_k)) \leq \Psi(\mathbf{X}_k) + \beta \langle \mathrm{grad} f(\mathbf{X}_k), \boldsymbol{\zeta}_k \rangle / L, \forall L \in [\widehat{L}, +\infty)$. $\square$

Table 3. Computation of $S_\lambda(\mathbf{B})$.

$\Upsilon(\mathbf{E})$	MR: $\|\|\mathbf{E}\|\|_1$	LRR: $\|\|\mathbf{E}\|\|_{2,1}$
$S_\lambda(\mathbf{B})$	$\text{sgn}(\mathbf{B}) \odot \max(\|\mathbf{B}\| - \frac{\lambda}{\gamma}, \mathbf{0})$	$[S_\lambda(\mathbf{B})]_i = \frac{\max(\|\mathbf{b}_i\| - \frac{\lambda}{\gamma}, 0)}{\|\mathbf{b}_i\|}\mathbf{b}_i, \forall i$

E. Proof of Proposition 1

Proof. A point $\mathbf{X}^* \in \mathcal{M}_{\leq r}$ is a local minimizer of (16) if and only if there exists $\varsigma \in \partial||\mathbf{X}||_*$ such that $\text{grad} f(\mathbf{X}) + \varsigma = \mathbf{0}$ [38]. Note that $\Psi(\mathbf{X})$ is bounded below. The proof can be completed by adapting the proof of Theorem 3.9 in [44]. $\square$

F. Proof of Proposition 2

Proof. Note that λ_k is non-increasing, $\mathcal{M}_{\leq r}$ is closed and $\Psi(\mathbf{X}, \mathbf{E})$ is bounded below. $\Psi(\mathbf{X}_{k+1}, \mathbf{E}_{k+1}) \leq \Psi(\mathbf{X}_{k+1}, \mathbf{E}_k) \leq \Psi(\mathbf{X}_k, \mathbf{E}_k)$ holds due to the line search w.r.t. $\mathbf{X}$ and thresholding property on $\mathbf{E}$. The convergence of Algorithm 4 can be established by adapting the proof of Theorem 3.9 in [44]. $\square$

G. Computation of $S_\lambda(\mathbf{B})$

Computation of $S_\lambda(\mathbf{B})$ is shown in Table 3.

H. Complexity comparison on LRR and RPCA

At the tth iteration of PRP, the complexity of PRG or PRG(R) is $O(mnr)$ for a large n. To compute Ξ_κ^t, we need to compute truncated SVD on a $n \times n$ matrix, which takes $O(n^2\kappa)$ time; while the truncated SVD in existing proximal gradient based methods takes $O(n^2 r)$. In contrast, for the LRR solver in [30], the time complexity per iteration is $O(nmr_D + nr_D^2 + r_D^3)$, where r_D denotes the rank of $\mathbf{D}$. Moreover, for the LRR solver in [29], the time complexity per iteration is $O(n^2 r_Z)$, where r_Z denotes the rank of $\mathbf{Z}$ in that iteration.

For RPCA, suppose the data $\mathbf{X}$ is of size $m \times n$. Since $\mathbf{X}$ is not sparse, the cmplexity of RPCA is $O(mnr)$ in general. However, unlike existing methods, the truncated SVDs in the proposed method are warm-started. As a result, the constant term in $O(mnr)$ is much reduced.