Supplementary Material to "Robust Subspace Tracking with Missing Data and Outliers: Novel Algorithm with Convergence Guarantee"

Le Trung Thanh, Nguyen Viet Dung, *Member, IEEE*, Nguyen Linh Trung, *Senior Member, IEEE* and Karim Abed-Meraim, *Fellow, IEEE*

I. Proof of Lemma 1

Follow the line as in previous convergence analysis of ADMM algorithms [1], [2], we can derive the proof of Lemma 1 as follows:

(P-1) The minimizer $\mathbf{u}^{k+1}$ defined in (15) in the main manuscript satisfies

$$\mathcal{L}(\mathbf{s}^k, \mathbf{u}^{k+1}, \mathbf{r}^k, \mathbf{w}^k, \mathbf{e}^k) \leq \mathcal{L}(\mathbf{s}^k, \mathbf{u}^k, \mathbf{r}^k, \mathbf{w}^k, \mathbf{e}^k) - \frac{1+\rho_1}{2}\|\mathbf{u}^k - \mathbf{u}^{k+1}\|_2^2. \tag{1}$$

In particular, the $\mathbf{u}$-update in fact minimizes the following objective function at the k-th iteration, as

$$\mathbf{u}^{k+1} \triangleq \underset{\mathbf{u}}{\operatorname{argmin}}\, \mathcal{L}_{\mathbf{u},k}(\mathbf{u}, .) = \frac{1+\rho_1}{2}\,\|\mathbf{u}\|_2^2 - [\mathbf{P}_t(\mathbf{x}_t - \mathbf{U}_{t-1}\mathbf{w}) - \rho_1(\mathbf{s}^k - \mathbf{r}^k)]^\top \mathbf{u}. \tag{2}$$

The function $\mathcal{L}_{\mathbf{u},k}(\mathbf{u}, .)$ w.r.t the variable $\mathbf{u}$ in (2) is strongly convex with a positive constant $(1+\rho_1)$, i.e., the Hessian of $\mathcal{L}_{\mathbf{u},k}(\mathbf{u}, .)$ is given by

$$\nabla^2 \mathcal{L}_{\mathbf{u},k}(\mathbf{u}, .) = (1+\rho_1)\mathbf{I}.$$

Since $\mathbf{u}^{k+1} = \underset{\mathbf{u}}{\operatorname{argmin}}\, \mathcal{L}_{\mathbf{u},k}(\mathbf{u}, .)$, we have the fact $\mathcal{L}_{\mathbf{u},k}(\mathbf{u}^{k+1}, .) \leq \mathcal{L}_{\mathbf{u},k}(\mathbf{u}^k, .)$. Therefore, we obtain the following inequality

$$\mathcal{L}_{\mathbf{u},k}(\mathbf{u}^{k+1}, .) \leq \mathcal{L}_{\mathbf{u},k}(\mathbf{u}^k, .) - \frac{1+\rho_1}{2}\|\mathbf{u}^{k+1} - \mathbf{u}^k\|_2^2. \tag{3}$$

(P-2) The minimizer $\mathbf{s}^{k+1}$ defined in (16) in the main manuscript satisfies

$$\mathcal{L}(\mathbf{s}^{k+1}, \mathbf{u}^{k+1}, \mathbf{r}^k, \mathbf{w}^k, \mathbf{e}^k) \leq \mathcal{L}(\mathbf{s}^k, \mathbf{u}^{k+1}, \mathbf{r}^k, \mathbf{w}^k, \mathbf{e}^k) - c_s\|\mathbf{s}^k - \mathbf{s}^{k+1}\|_2^2, \tag{4}$$

with a positive constant c_s.

In particular, at the k-th iteration, the variable $\mathbf{s}$ is updated by minimizing the objective function $\mathcal{L}_{\mathbf{s},k}(\mathbf{s}, .)$ as

$$\mathbf{s}^{k+1} \triangleq \underset{\mathbf{s}}{\operatorname{argmin}}\, \mathcal{L}_{\mathbf{s},k}(\mathbf{s}, .) = \rho\,\|\mathbf{s}\|_1 + \frac{\rho_1}{2}\|\mathbf{s} - (\mathbf{u}^{k+1} + \mathbf{r}^k)\|_2^2. \tag{5}$$

Le Trung Thanh is with the University of Engineering and Technology, Vietnam National University, Hanoi, Vietnam, and the PRISME Laboratory, University of Orléans, Orléans, France. Email: thanhletrung@vnu.edu.vn, trung-thanh.le@univ-orleans.fr.

Nguyen Viet Dung is with the University of Engineering and Technology, Vietnam National University, Hanoi and the National Institute of Advanced Technologies of Brittany (ENSTA Bretagne), Brest, France. Email: dungnv@vnu.edu.vn, viet.nguyen@ensta-bretagne.fr.

Nguyen Linh Trung (corresponding author) is with the University of Engineering and Technology, Vietnam National University, Hanoi. E-mail: linhtrung@vnu.edu.vn.

Karim Abed-Meraim is with the PRISME Laboratory, University of Orléans, France. Email: karim.abed-meraim@univ-orleans.fr.

This work was funded by the National Foundation for Science and Technology Development (NAFOSTED) of Vietnam under grant number 102.04-2019.14.

Because the two functions of the ℓ_1-norm $\|\mathbf{s}\|_1$ and ℓ_2-norm $\left\|\mathbf{s}-(\mathbf{u}^{k+1}+\mathbf{r}^k)\right\|_2^2$ are convex, so the $\mathcal{L}_{\mathbf{s},k}(\mathbf{s},.)$ in (5) w.r.t. $\mathbf{s}$ is also convex. It is therefore that for any $\mathbf{s}^k, \mathbf{s}^{k+1} \in \mathbf{R}^n$, we always have

$$\mathcal{L}_{\mathbf{s},k}(\mathbf{s}^k,.) \geq \mathcal{L}_{\mathbf{s},k}(\mathbf{s}^{k+1},.) + \langle \mathbf{s}^k - \mathbf{s}^{k+1}, \nabla\mathcal{L}_{\mathbf{s},k}(\mathbf{s}^{k+1},.)\rangle + \frac{1}{2}\|\mathbf{s}^{k+1}-\mathbf{s}^k\|_2^2. \tag{6}$$

Since $\mathbf{s}^{k+1} = \underset{\mathbf{s}}{\operatorname{argmin}}\, \mathcal{L}_{\mathbf{s},k}(\mathbf{s},.)$, the first derivative $\nabla\mathcal{L}_{\mathbf{s},k}(\mathbf{s}^{k+1},.) = \mathbf{0}$, and hence the inequality

$$\mathcal{L}_{\mathbf{s},k}(\mathbf{s}^{k+1},.) \leq \mathcal{L}_{\mathbf{s},k}(\mathbf{s}^k,.) - \frac{1}{2}\|\mathbf{s}^{k+1}-\mathbf{s}^k\|_2^2.$$

As a result, we have

$$\sum_{k=1}^{K}\frac{1}{2}\|\mathbf{s}^{k+1}-\mathbf{s}^k\|_2^2 \leq \sum_{i=1}^{K}\mathcal{L}_{\mathbf{s},k}(\mathbf{s}^k,.) - \mathcal{L}_{\mathbf{s},k}(\mathbf{s}^{k+1},.) = \mathcal{L}_{\mathbf{s},k}(\mathbf{s}^1,.) - \mathcal{L}_{\mathbf{s},k}(\mathbf{s}^{K+1},.). \tag{7}$$

Let $K \to \infty$, we then have

$$\sum_{k=1}^{\infty}\|\mathbf{s}^{k+1}-\mathbf{s}^k\|_2^2 < \infty. \tag{8}$$

It ends the proof of (P-2).

(P-3) The minimizer $\mathbf{r}^{k+1}$ defined in (14) satisfies

$$\mathcal{L}(\mathbf{s}^{k+1},\mathbf{u}^{k+1},\mathbf{r}^{k+1},\mathbf{w}^k,\mathbf{e}^k) \leq \mathcal{L}(\mathbf{s}^{k+1},\mathbf{u}^{k+1},\mathbf{r}^k,\mathbf{w}^k,\mathbf{e}^k) - c_r\|\mathbf{r}^k-\mathbf{r}^{k+1}\|_2^2. \tag{9}$$

Follow the $\mathbf{r}$-update in (14), it is easy to verify that

$$\begin{aligned}\mathcal{L}(\mathbf{s}^{k+1},\mathbf{u}^{k+1},\mathbf{r}^{k+1},\mathbf{w}^k,\mathbf{e}^k) &= \rho_1(\mathbf{r}^k-\mathbf{s}^{k+1}+\mathbf{u}^{k+1})^\top(\mathbf{u}^{k+1}-\mathbf{s}^{k+1}) + A\\ &= \rho_1(\mathbf{r}^k)^\top(\mathbf{u}^{k+1}-\mathbf{s}^{k+1}) - \rho_1\|\mathbf{u}^{k+1}-\mathbf{s}^{k+1}\|_2^2 + A\\ &= \mathcal{L}(\mathbf{s}^{k+1},\mathbf{u}^{k+1},\mathbf{r}^k,\mathbf{w}^k,\mathbf{e}^k) - \rho_1\|\mathbf{r}^{k+1}-\mathbf{r}^k\|_2^2,\end{aligned}$$

where $A = g(\mathbf{s}^{k+1}) + h(\mathbf{u}^{k+1}) + \frac{\rho_1}{2}\|\mathbf{u}^{k+1}-\mathbf{s}^{k+1}\|$. It results in (P-3).

(P-4) The minimizer $\mathbf{w}^{k+1}$ defined in (20) satisfies

$$\mathcal{L}(\mathbf{s}^{k+1},\mathbf{u}^{k+1},\mathbf{r}^{k+1},\mathbf{w}^{k+1},\mathbf{e}^k) \leq \mathcal{L}(\mathbf{s}^{k+1},\mathbf{u}^{k+1},\mathbf{r}^{k+1},\mathbf{w}^k,\mathbf{e}^k) - c_w\|\mathbf{w}^k-\mathbf{w}^{k+1}\|_2^2. \tag{10}$$

The $\mathbf{w}$-update minimizes the following objective function

$$\mathbf{w}^{k+1} \triangleq \mathcal{L}_{\mathbf{w},k}(\mathbf{w},.) = \frac{\rho_2}{2}\|\mathbf{P}_t(\mathbf{U}_t\mathbf{w}+\mathbf{s}^{k+1}-\mathbf{x}_t)-\mathbf{e}^k\|_2^2$$

Since $\mathcal{L}_{\mathbf{z},k}(\mathbf{w},.)$ is strongly convex, it implies (P-4) that

$$\mathcal{L}(\mathbf{s}^{k+1},\mathbf{u}^{k+1},\mathbf{r}^{k+1},\mathbf{w}^{k+1},\mathbf{e}^k) \leq \mathcal{L}(\mathbf{s}^{k+1},\mathbf{u}^{k+1},\mathbf{r}^{k+1},\mathbf{w}^k,\mathbf{e}^k) - c_w\|\mathbf{w}^k-\mathbf{w}^{k+1}\|_2^2.$$

with a positive number c_w.

(P-5) The minimizer $\mathbf{e}^{k+1}$ defined in (22) satisfies

$$\mathcal{L}(\mathbf{s}^{k+1},\mathbf{u}^{k+1},\mathbf{r}^{k+1},\mathbf{w}^{k+1},\mathbf{e}^{k+1}) \leq \mathcal{L}(\mathbf{s}^{k+1},\mathbf{u}^{k+1},\mathbf{r}^{k+1},\mathbf{w}^{k+1},\mathbf{e}^k) - c_e\|\mathbf{e}^k-\mathbf{e}^{k+1}\|_2^2. \tag{11}$$

Similarly, the Hessian matrix of $\mathcal{L}_{\mathbf{e},k}(\mathbf{e},.)$ is a positive-define matrix, as

$$\nabla^2\mathcal{L}_{\mathbf{e},k}(\mathbf{e},.) = \operatorname{diag}\left(\left[(\mathbf{e}(1)^2+1)^{-3/2},\ldots,(\mathbf{e}(n))^2+1)^{-3/2}\right]\right) + \frac{\rho_2}{2}\mathbf{I}. \tag{12}$$

From the Proposition 2, we have

$$\mathcal{L}_{\mathbf{e},k}(\mathbf{e}^{k+1},.) \leq \mathcal{L}_{\mathbf{e},k}(\mathbf{e}^k,.) - \frac{\rho_2}{2}\left\|\mathbf{e}^{k+1}-\mathbf{e}^k\right\|_2^2. \tag{13}$$

It ends the proof.

II. Proof of Proposition 2

To prove that $g_t(\mathbf{U})$ is strongly convex, we state the following facts: $g_t(\mathbf{U})$ is continuous and differentiable; its second derivative is a positive semi-definite matrix (i.e., $\nabla^2_{\mathbf{U}} g_t(\mathbf{U}) \geq m\mathbf{I}$); and the domain of $g_t(\mathbf{U})$ is convex. In order to satisfy the Lipschitz condition, we show that the first derivative of $g_t(\mathbf{U})$ is bounded.

Stage I: Prove that g_t is a strong convex function.

We show that there exists a positive number m such that

$$|g_t(\mathbf{U}_{t+1}) - g_t(\mathbf{U}_t)| \geq m_1 \left\|\mathbf{U}_{t+1} - \mathbf{U}_t\right\|_F^2. \tag{14}$$

In particular, we state the two claims as follows:

(C-1) $g_t(\mathbf{U})$ is continuous and differentiable.

Proof. Given two variables $\mathbf{A}, \mathbf{B} \in \mathcal{U}$ such that $\|\mathbf{A} - \mathbf{B}\|_F^2 < \gamma$ for some positive constant γ. It is easy to verify that there exists a positive number θ such that $|g_t(\mathbf{A}) - g_t(\mathbf{B})| < \theta$.
Under the given assumptions, we have the following inequality:

$$\begin{aligned}
|g_t(\mathbf{A}) - g_t(\mathbf{B})| =& \frac{1}{t}\left| \sum_{i=1}^{t} \lambda_i^{t-i} \left\|\mathbf{P}_i(\mathbf{A}\mathbf{w}_i + \mathbf{s}_i - \mathbf{x}_i)\right\|_2^2 - \sum_{i=1}^{t} \lambda_i^{t-i} \left\|\mathbf{P}_i(\mathbf{B}\mathbf{w}_i + \mathbf{s}_i - \mathbf{x}_i)\right\|_2^2 \right| \\
&\leq \frac{1}{t} \sum_{i=1}^{t} \lambda_i^{t-i} \left\|\mathbf{P}_i(\mathbf{A} - \mathbf{B})\mathbf{w}_i\right\|_2 \left\|\mathbf{P}_i(\mathbf{A} + \mathbf{B})\mathbf{w}_i + 2(\mathbf{s}_i - \mathbf{x}_i)\right\|_2 \\
&\leq \frac{1}{t} \sum_{i=1}^{t} 2\lambda_i^{t-i} \left\|\mathbf{w}_i\right\|_2^2 \left\|(\mathbf{A} - \mathbf{B})\right\|_F \left\|(\mathbf{A} + \mathbf{B})\right\|_F \left\|\mathbf{s}_i - \mathbf{x}_i\right\|_2 = \theta,
\end{aligned}$$

where $\lambda_i = \lambda(\mathrm{tr}(\mathbf{P}_i)/n)^{1/t-i}$, thanks to the triangle inequality. It is therefore that the set of functions $\{g_t(\mathbf{U})\}_{t=1}^{\infty}$ is continuous on $\mathcal{U}$.
Furthermore, for any $\mathbf{U}^*, \boldsymbol{\Delta} \in \mathcal{U}$, we show that the following limit exists:

$$\begin{aligned}
\lim_{\|\boldsymbol{\Delta}\|\to 0} \frac{|g_t(\mathbf{U}^* + \boldsymbol{\Delta}) - g_t(\mathbf{U}^*)|}{\|\boldsymbol{\Delta}\|} = \lim_{\|\boldsymbol{\Delta}\|\to 0} \frac{1}{t\|\boldsymbol{\Delta}\|} \sum_{i=1}^{t} \lambda_i^{t-i} \Bigg(& \left\|\mathbf{P}_i((\mathbf{U}^* + \boldsymbol{\Delta})\mathbf{w}_i + \mathbf{s}_i - \mathbf{x}_i)\right\|_2^2 \\
& - \left\|\mathbf{P}_i(\mathbf{U}^*\mathbf{w}_i + \mathbf{s}_i - \mathbf{x}_i)\right\|_2^2 \Bigg).
\end{aligned} \tag{15}$$

Specifically, let us denote $\mathbf{y}_i = \mathbf{P}_i(\mathbf{U}^*\mathbf{w}_i + \mathbf{s}_i - \mathbf{x}_i)$, the limit can be written as follows:

$$\begin{aligned}
\lim_{\|\boldsymbol{\Delta}\|\to 0} \frac{|g_t(\mathbf{U}^* + \boldsymbol{\Delta}) - g_t(\mathbf{U}^*)|}{\|\boldsymbol{\Delta}\|} &= \lim_{\|\boldsymbol{\Delta}\|\to 0} \frac{1}{t\|\boldsymbol{\Delta}\|} \sum_{i=1}^{t} \lambda_i^{t-i} \big(\left\|\mathbf{y}_i - \mathbf{P}_i\boldsymbol{\Delta}\mathbf{w}_i\right\|_2^2 - \left\|\mathbf{y}_i\right\|_2^2 \big) \\
&= \lim_{\|\boldsymbol{\Delta}\|\to 0} \frac{1}{t\|\boldsymbol{\Delta}\|} \sum_{i=1}^{t} \lambda_i^{t-i} \big(\left\|\mathbf{P}_i\boldsymbol{\Delta}\mathbf{w}_i\right\|_2^2 - 2\langle \mathbf{y}_i, \mathbf{P}_i\boldsymbol{\Delta}\mathbf{w}_i \rangle \big) \\
&= -\frac{2}{t} \sum_{i=1}^{t} \lambda_i^{t-i} \|\mathbf{y}_i\|_2 \cos(\mathbf{y}_i, \mathbf{P}_i\boldsymbol{\Delta}\mathbf{w}_i) < \infty.
\end{aligned} \tag{16}$$

As a result, the function $g_t(\mathbf{U})$ is differentiable and its first derivative $\nabla g_t(\mathbf{U})$ can be given by

$$\nabla g_t(\mathbf{U}) = \frac{2}{t} \sum_{i=1}^{t} \lambda_i^{t-i} \mathbf{P}_i(\mathbf{U}\mathbf{w}_i + \mathbf{s}_i - \mathbf{x}_i)\mathbf{w}_i^{\top}. \tag{17}$$

In the similar way, it is easy to verify that $\nabla g_t(\mathbf{U})$ is also continuous and the second derivative $\nabla^2 g_t(\mathbf{U})$ is given by

$$\nabla^2 g_t(\mathbf{U}) = \frac{2}{t}\sum_{i=1}^{t} \lambda_i^{t-i}\mathbf{P}_i\mathbf{w}_i\mathbf{w}_i^\top. \tag{18}$$

□

(C-2) The second derivative $\nabla^2 g_t(\mathbf{U})$ is a positive-define matrix. For all $\mathbf{x} \in \mathbb{R}^{p\times 1}$, we have

$$\mathbf{x}^\top\nabla^2 g_t(\mathbf{U})\mathbf{x} = \frac{2}{t}\sum_{i=1}^{t} \lambda_i^{t-i}\mathbf{P}_i(\mathbf{w}_i^\top\mathbf{x})^\top(\mathbf{w}_i^\top\mathbf{x}) = \frac{2}{t}\sum_{i=1}^{t} \lambda_i^{t-i}\mathbf{P}_i(\mathbf{w}_i^\top\mathbf{x})^2 > 0, \quad \forall\lambda, t > 0. \tag{19}$$

It implies that there always exist a positive constant m such that $\nabla^2 g_t(\mathbf{U}) \geq m\mathbf{I}$.

It follows to the claims (C-1), (C-2) and the assumptions showing that the domain of $g_t(\mathbf{U})$ is a convex set that $g_t(\mathbf{U}_t)$ is strongly convex [3, Section 3.1.4].

Stage II: Prove that $g_t(\mathbf{U})$ is also a Lipschitz function:

$$|g_t(\mathbf{U}_{t+1}) - g_t(\mathbf{U}_t)| \leq m_2 \left\|\mathbf{U}_{t+1} - \mathbf{U}_t\right\|_F. \tag{20}$$

Let us denote $d_t(\mathbf{U}) = g_t(\mathbf{U}) - g_{t+1}(\mathbf{U})$. Since $\mathbf{U}_t = \underset{\mathbf{U}}{\operatorname{argmin}}\, g_t(\mathbf{U})$, we exploit that $g_{t+1}(\mathbf{U}_{t+1}) \leq g_{t+1}(\mathbf{U}_t)$ and hence

$$\begin{aligned} g_t(\mathbf{U}_{t+1}) - g_t(\mathbf{U}_t) &= g_t(\mathbf{U}_{t+1}) - g_{t+1}(\mathbf{U}_t) + g_{t+1}(\mathbf{U}_t) - g_t(\mathbf{U}_t) \\ &\leq \big(\underbrace{g_t(\mathbf{U}_{t+1}) - g_{t+1}(\mathbf{U}_{t+1})}_{d_t(\mathbf{U}_{t+1})}\big) - \big(\underbrace{g_t(\mathbf{U}_t) - g_{t+1}(\mathbf{U}_t)}_{d_t(\mathbf{U}_t)}\big). \end{aligned} \tag{21}$$

The first derivative of $d_t(\mathbf{U}) = g_t(\mathbf{U}) - g_{t+1}(\mathbf{U})$ is given by

$$\begin{aligned} \nabla d_t(\mathbf{U}) &= \nabla g_t(\mathbf{U}) - \nabla g_{t+1}(\mathbf{U}) \\ &= \frac{1}{t}\sum_{i=1}^{t} \lambda_i^{t-i}\mathbf{P}_i(\mathbf{U}\mathbf{w}_i + \mathbf{s}_i - \mathbf{x}_i)\mathbf{w}_i^\top - \frac{1}{t+1}\sum_{i=1}^{t+1} \lambda_i^{t+1-i}\mathbf{P}_i(\mathbf{U}\mathbf{w}_i + \mathbf{s}_i - \mathbf{x}_i)\mathbf{w}_i^\top. \end{aligned} \tag{22}$$

Let $\mathbf{A}_t = \sum_{i=1}^{t} \lambda_i^{t-i}\mathbf{P}_i\mathbf{U}\mathbf{w}_i\mathbf{w}_i^\top$ and $\mathbf{B}_t = \sum_{i=1}^{t} \lambda_i^{t-i}\mathbf{P}_i(\mathbf{s}_i - \mathbf{x}_i)$, we can rewrite $\nabla d_t(\mathbf{U})$ as

$$\nabla d_t(\mathbf{U}) = \left(\frac{\mathbf{A}_t}{t} - \frac{\mathbf{A}_{t+1}}{t+1}\right) + \left(\frac{\mathbf{B}_t}{t} - \frac{\mathbf{B}_{t+1}}{t+1}\right). \tag{23}$$

Under the given assumptions, the subspace $\mathbf{U}$, outlier $\{\mathbf{s}_t\}$, signal $\{\mathbf{x}_t\}$ and subspace coefficients $\{\mathbf{w}_t\}$ are bounded, then both $\mathbf{A}_t$ and $\mathbf{B}_t$ are bounded. It is therefore that

$$\left\|\nabla d_t(\mathbf{U})\right\|_F \leq \left\|\frac{\mathbf{A}_t}{t} - \frac{\mathbf{A}_{t+1}}{t+1}\right\|_F + \left\|\frac{\mathbf{B}_t}{t} - \frac{\mathbf{B}_{t+1}}{t+1}\right\|_F \leq m_2 = \mathcal{O}(1/t). \tag{24}$$

Therefore $d_t(\mathbf{U})$ is Lipschiz with the constant m_2,

$$\frac{|d_t(\mathbf{U}_{t+1}) - d_t(\mathbf{U}_t)|}{\left\|\mathbf{U}_{t+1} - \mathbf{U}_t\right\|_F} \leq m_2, \text{ hence } \frac{|g_t(\mathbf{U}_{t+1}) - g_t(\mathbf{U}_t)|}{\left\|\mathbf{U}_{t+1} - \mathbf{U}_t\right\|_F} \leq m_2. \tag{25}$$

This ends the proof.

III. Proof of the Lemma 3

Inspired of the result of convergence analysis for online sparse coding framework in [4, Proposition 2], we derive the convergence of $g_t(\mathbf{U}_t)$ in the similar way. In particular, we first denote the nonnegative stochastic process $\{u_t\}, u_t \triangleq g_t(\mathbf{U}_t) \geq 0$, and then prove that it is a quasi-martingale, i.e., we have to prove the sum of the positive difference of $\{u_t\}_{t=1}^{\infty}$ is bounded,

$$\sum_{t=1}^{\infty} \left|\mathbb{E}[u_{t+1} - u_t|\mathcal{F}_t]\right| < +\infty \quad a.s. \tag{26}$$

We can express $g_{t+1}(\mathbf{U}_t)$ with respect to $g_t(\mathbf{U}_t)$ as follows

$$\begin{aligned} g_{t+1}(\mathbf{U}_t) &= \frac{1}{t+1}\sum_{i=1}^{t+1} \lambda_i^{t+1-i} \left\|\mathbf{P}_i(\mathbf{U}_t\mathbf{w}_i + \mathbf{s}_i - \mathbf{x}_i)\right\|_2^2 + \rho\left\|\mathbf{s}_i\right\|_1 \\ &= \left(\frac{\lambda}{t+1}\sum_{i=1}^{t} \lambda_i^{t-i} \left\|\mathbf{P}_i(\mathbf{U}_t\mathbf{w}_i + \mathbf{s}_i - \mathbf{x}_i)\right\|_2^2 + \rho\left\|\mathbf{s}_i\right\|_1\right) + \left(\frac{1}{t+1}\big(\left\|\mathbf{P}_{t+1}\mathbf{U}_t + \mathbf{s}_{t+1} - \mathbf{x}_{t+1}\right\|_2^2 + \rho\left\|\mathbf{s}_{t+1}\right\|_1\big)\right) \\ &= \frac{\lambda_i t}{t+1} g_t(\mathbf{U}_t) + \frac{1}{t+1}\ell(\mathbf{U}_t, \mathbf{P}_{t+1}, \mathbf{x}_{t+1}). \end{aligned}$$

Since $\mathbf{U}_{t+1} = \underset{\mathbf{U}}{\operatorname{argmin}}\, g_{t+1}(\mathbf{U})$, we have the fact $g_{t+1}(\mathbf{U}_{t+1}) - g_{t+1}(\mathbf{U}_t) \leq 0$, $f_t(\mathbf{U}_t) \leq g_t(\mathbf{U}_t)$, and hence

$$\begin{aligned} u_{t+1} - u_t = g_{t+1}(\mathbf{U}_{t+1}) - g_t(\mathbf{U}_t) &= \underbrace{g_{t+1}(\mathbf{U}_{t+1}) - g_{t+1}(\mathbf{U}_t)}_{\leq 0} + g_{t+1}(\mathbf{U}_t) - g_t(\mathbf{U}_t) \\ &\leq g_{t+1}(\mathbf{U}_t) - g_t(\mathbf{U}_t) = \frac{1}{t+1}\ell(\mathbf{U}_t, \mathbf{P}_{t+1}, \mathbf{x}_{t+1}) - \frac{t(1-\lambda_i)+1}{t+1} g_t(\mathbf{U}_t). \end{aligned} \tag{27}$$

It is therefore that

$$\begin{aligned} \mathbb{E}[u_{t+1} - u_t|\mathcal{F}_t] &\leq \frac{\mathbb{E}[\ell(\mathbf{U}_t, \mathbf{P}_{t+1}, \mathbf{x}_{t+1}) - (t(1-\lambda_i)+1) g_t(\mathbf{U}_t)|\mathcal{F}_t]}{t+1} \\ &\leq \frac{\mathbb{E}[\ell(\mathbf{U}_t, \mathbf{P}_{t+1}, \mathbf{x}_{t+1}) - g_t(\mathbf{U}_t)|\mathcal{F}_t]}{t+1} \leq \frac{\mathbb{E}[\ell(\mathbf{U}_t, \mathbf{P}_{t+1}, \mathbf{x}_{t+1})] - f_t(\mathbf{U}_t)}{t+1} = \frac{f(\mathbf{U}_t) - f_t(\mathbf{U}_t)}{t+1} \end{aligned}$$

because of $f_t(\mathbf{U}_t) \leq g_t(\mathbf{U}_t)$ and $\mathbb{E}[\ell(\mathbf{U}_t, \mathbf{P}_{t+1}, \mathbf{x}_t)] = f(\mathbf{U}_t)$.

Let us define the indicator function δ_t as follows

$$\delta_t \triangleq \begin{cases} 1 & \text{if} \quad \mathbb{E}[g_{t+1}(\mathbf{U}_{t+1}) - g_t(\mathbf{U}_t)|\mathcal{F}_t] > 0 \\ 0 & \text{otherwise,} \end{cases}$$

we then have

$$\mathbb{E}\big[\delta\mathbb{E}[g_{t+1}(\mathbf{U}_{t+1}) - g_t(\mathbf{U}_t)|\mathcal{F}_t]\big] \leq \mathbb{E}[\sqrt{t}(f(\mathbf{U}_t) - f_t(\mathbf{U}_t))]\frac{1}{\sqrt{t}(t+1)}. \tag{28}$$

Under the given assumptions that $\{\mathbf{U}, \mathbf{w}, \mathbf{s}, \mathbf{x}\}$ are bounded, we exploit that the set of measurable functions $\{\ell(\mathbf{U}_i, \mathbf{P}, \mathbf{x})\}_{i\geq 1}$, which is composed of a quadratic norm term and ℓ_1-norm term, is $\mathbb{P}$-Donsker. Therefore, the centered and scaled version of the empirical function $f_t(\mathbf{U}_t)$ satisfies the following proposition:

$$\mathbb{E}[\sqrt{t}(f(\mathbf{U}_t) - f_t(\mathbf{U}_t))] = \mathcal{O}(1), \tag{29}$$

thanks to Proposition 9.

Furthermore, let us consider the convergence of the sum $\sum_{t=1}^{\infty} \frac{\alpha}{\sqrt{t}(t+1)}$. We use the Cauchy-MacLaurin integral test [5] for convergence, as

$$\begin{aligned}\int_{t=1}^{+\infty} \frac{\alpha}{\sqrt{t}(t+1)} dt &= \alpha \int_{x=1}^{+\infty} \frac{1}{x^2+1} dx = \alpha \arctan(x)\big|_1^{+\infty} \\ &= \alpha\big(\arctan(\infty) - \arctan(1)\big) = \alpha\frac{\pi}{4} < \infty.\end{aligned}$$

It is therefore that $\left\{\frac{1}{\sqrt{t}(t+1)}\right\}_{t>0}$ converges and hence

$$\sum_{t=1}^{\infty} \mathbb{E}\big[\delta \mathbb{E}[g_{t+1}(\mathbf{U}_{t+1}) - g_t(\mathbf{U}_t)|\mathcal{F}_t]\big] < \infty. \tag{30}$$

According to quasi-martingale theorem as shown in Proposition 10, we can conclude that $\{g_t(\mathbf{U}_t)\}_{t=1}^{\infty}$ converges almost surely

$$\sum_{t=1}^{\infty} \mathbb{E}[g_{t+1}(\mathbf{U}_{t+1}) - g_t(\mathbf{U}_t)|\mathcal{F}_t] < \infty. \tag{31}$$

We complete the proof.

IV. Proof of Lemma 4

We investigate the convergence of a surrogate sequence $\left\{\big(g_t(\mathbf{U}_t) - f_t(\mathbf{U}_t)\big)\frac{1}{t+1}\right\}$ as follows

$$\begin{aligned}\frac{g_t(\mathbf{U}_t) - f_t(\mathbf{U}_t)}{t+1} &= u_t - u_{t+1} + \underbrace{g_{t+1}(\mathbf{U}_{t+1}) - g_{t+1}(\mathbf{U}_t)}_{\leq 0} + \underbrace{\frac{t(\lambda-1)}{t+1} g_t(\mathbf{U}_t)}_{\leq 0} + \frac{\ell(\mathbf{U}_t, \mathbf{P}_{t+1}, \mathbf{x}_{t+1}) - f_t(\mathbf{U}_t)}{t+1} \\ &\leq \underbrace{u_t - u_{t+1}}_{\text{(S-1)}} + \underbrace{\frac{\ell(\mathbf{U}_t, \mathbf{P}_{t+1}, \mathbf{x}_{t+1}) - f_t(\mathbf{U}_t)}{t+1}}_{\text{(S-2)}}\end{aligned} \tag{32}$$

because of $u_t = g_t(\mathbf{U}_t)$ and $\lambda \leq 1$. Note that, (S-1) $-$ (S-2) converge almost surely:

- The sequence $\mathbb{E}[u_t - u_{t+1}]$ converges almost surely as proved in Lemma 3.
- The sequence (S-2) also converges, thanks to the fact $\mathbb{E}[\ell(\mathbf{U}_t, \mathbf{P}_{t+1}, \mathbf{x}_{t+1})] = f(\mathbf{U}_t)$ and the convergence of $\frac{\mathbb{E}[f(\mathbf{U}_t) - f_t(\mathbf{U}_t)]}{t+1}$ as mentioned in the appendix III.

It is therefore that the sequence $\left\{\big(g_t(\mathbf{U}_t) - f_t(\mathbf{U}_t)\big)\frac{1}{t+1}\right\}$ converges almost surely, i.e.,

$$\sum_{t=0}^{\infty} \big(g_t(\mathbf{U}_t) - f_t(\mathbf{U}_t)\big)\frac{1}{t+1} < \infty. \tag{33}$$

On the other hand, the real sequence $\left\{\frac{1}{t+1}\right\}_{t\geq 0}$ diverges, $\sum_{t=0}^{\infty} \frac{1}{t+1} = \infty$. It implies that $g_t(\mathbf{U}_t) - f_t(\mathbf{U}_t)$ convergences, thanks to the Proposition 7.

V. Proof of Corollary 4

Let $\mathbf{U}_t = \underset{\mathbf{U}}{\operatorname{argmin}}\, g_t(\mathbf{U})$ when $t \to \infty$, we have

$$f_t(\mathbf{U}_t) \leq f_t(\mathbf{U}) + \frac{L}{2}\|\mathbf{U} - \mathbf{U}_t\|_F^2, \forall\ \mathbf{U} \in \mathcal{U}, \tag{34}$$

where L is a positive constant. In other words, $\mathbf{U}_t$ is the minimum point of $f(\mathbf{U})$.

Proof. Let us denote the error function $e_t(\mathbf{U}) = g_t(\mathbf{U}) - f_t(\mathbf{U})$. Then it is easy to have $\nabla e_t(\mathbf{U}) = \nabla g_t(\mathbf{U}) - \nabla f_t(\mathbf{U})$ because the function $f_t(\mathbf{U})$ and its surrogate $g_t(\mathbf{U})$ are continuous and differentiable. We first have the following facts

$$\begin{aligned}\|\nabla e_t(\mathbf{U}) - \nabla e_t(\mathbf{U}_t)\| &= \big\|\big(\nabla g_t(\mathbf{U}) - \nabla g_t(\mathbf{U}_t)\big) - \big(\nabla f_t(\mathbf{U}) - \nabla f_t(\mathbf{U}_t)\big)\big\| \\ &= \big\|\big(\nabla g_t(\mathbf{U})\big) - \big(\nabla f_t(\mathbf{U}) - \nabla f_t(\mathbf{U}_t)\big)\big\| \\ &\leq \|\nabla g_t(\mathbf{U}) - g_t(\mathbf{U}_t)\| + \|\nabla f_t(\mathbf{U}) - \nabla f_t(\mathbf{U}_t)\|, \end{aligned} \tag{35}$$

thanks to the triangle theorem.

As proved in the Proposition 2, the surrogate function $g_t(\mathbf{U})$ is strongly convex, but also Lipschitz and its second derivative are given by

$$\nabla^2 g_t(\mathbf{U}) = \frac{2}{t}\sum_{i=1}^{t} \lambda_i^{t-i}\mathbf{P}_i\mathbf{w}_i\mathbf{w}_i^\top. \tag{36}$$

Under the given assumption that the subspace coefficient vectors $\{\mathbf{v}_i\}_{i\geq 1}$ are bounded, the $\nabla^2_{\mathbf{U}} g_t(\mathbf{U})$ is then bounded. It is therefore that the first derivative $\nabla g_t(\mathbf{U})$ is also a Lipschitz function, that means,

$$\|\nabla g_t(\mathbf{U}) - \nabla g_t(\mathbf{U}_t)\| \leq L_g\|\mathbf{U} - \mathbf{U}_t\|_F. \tag{37}$$

In parallel, we will show that the first derivative of the cost function $f_t(\mathbf{U})$ is Lipschitz too, as

$$\|\nabla f_t(\mathbf{U}) - \nabla f_t(\mathbf{U}_t)\| \leq L_f\|\mathbf{U} - \mathbf{U}_t\|_F. \tag{38}$$

For any two subspace variables $\mathbf{U}_1, \mathbf{U}_2 \in \mathcal{U}$, we have

$$\|\nabla f_t(\mathbf{U}_1) - \nabla f_t(\mathbf{U}_2)\| \leq \frac{1}{t}\sum_{i=1}^{t}\lambda_i^{t-i}\|\nabla \ell(\mathbf{U}_1, \mathbf{P}_i, \mathbf{x}_i) - \nabla \ell(\mathbf{U}_2, \mathbf{P}_i, \mathbf{x}_i)\|. \tag{39}$$

For any signal $\mathbf{x} \in \mathcal{S}$ at time instant t, we also have

$$\begin{aligned}&\|\nabla \ell(\mathbf{U}_1, \mathbf{P}_t, \mathbf{x}) - \nabla \ell(\mathbf{U}_2, \mathbf{P}_t, \mathbf{x})\| \\ &\leq\ \|\mathbf{P}_t\mathbf{U}_1\mathbf{w}^*(\mathbf{U}_1, \mathbf{P}_t, \mathbf{x})\mathbf{w}^*(\mathbf{U}_1, \mathbf{P}_t, \mathbf{x})^\top - \mathbf{P}_t\mathbf{U}_2\mathbf{w}^*(\mathbf{U}_1, \mathbf{P}_t, \mathbf{x})\mathbf{w}^*(\mathbf{U}_1, \mathbf{P}_t, \mathbf{x})^\top\| \\ &+ \|\mathbf{s}^*(\mathbf{U}_1, \mathbf{P}_t, \mathbf{x})\mathbf{w}^*(\mathbf{U}_1, \mathbf{P}_t, \mathbf{x})^\top - \mathbf{s}^*(\mathbf{U}_2, \mathbf{P}_t, \mathbf{x})\mathbf{w}^*(\mathbf{U}_2, \mathbf{P}_t, \mathbf{x})^\top\| + \|\mathbf{x}\mathbf{w}^*(\mathbf{U}_1, \mathbf{P}_t, \mathbf{x})^\top - \mathbf{x}\mathbf{w}^*(\mathbf{U}_2, \mathbf{P}_t, \mathbf{x})^\top\|.\end{aligned}$$

where $(\mathbf{w}^*(\mathbf{U}, \mathbf{P}, \mathbf{x}), \mathbf{s}^*(\mathbf{U}, \mathbf{P}, \mathbf{x})) \triangleq \operatorname{argmin}_{\mathbf{w},\mathbf{s}} \ell(\mathbf{U}, \mathbf{P}, \mathbf{x}, \mathbf{w}, \mathbf{s})$. Note that, $\big(\mathbf{w}^*(\mathbf{U}, \mathbf{P}, \mathbf{x}), \mathbf{s}^*(\mathbf{U}, \mathbf{P}, \mathbf{x})\big)$ can be seen as a continuous function of the two variables. As mentioned in the proof of Lemma 1, the function is not only strongly convex, but also Lipschitz in terms of each variable $\mathbf{s}$ or $\mathbf{w}$. Therefore, we have the following facts:

$$\begin{aligned}\|\mathbf{w}^*(\mathbf{U}_1, \mathbf{P}, \mathbf{x}) - \mathbf{w}^*(\mathbf{U}_2, \mathbf{P}, \mathbf{x})\| &\leq c_1\|\mathbf{P}(\mathbf{U}_1 - \mathbf{U}_2)\|, \\ \|\mathbf{s}^*(\mathbf{U}_1, \mathbf{P}, \mathbf{x}) - \mathbf{s}^*(\mathbf{U}_2, \mathbf{P}, \mathbf{x})\| &\leq c_2\|\mathbf{P}(\mathbf{U}_1 - \mathbf{U}_2)\|,\end{aligned}$$

where c_1 and c_2 are the Lipschitz number of $\mathbf{w}^*(\mathbf{U}_1, \mathbf{P}, \mathbf{x})$ and $\mathbf{w}^*(\mathbf{U}_1, \mathbf{P}, \mathbf{x})$ respectively.

Denote the upper bound for $\mathbf{x}$, $\mathbf{s}$, $\mathbf{w}$ and $\mathbf{U}$ are $\alpha_1, \alpha_2, \alpha_3$ and α_4 respectively. The first part of (E-5) can be bounded as follows:

$$\begin{aligned}
&\|\mathbf{P}_t\mathbf{U}_1\mathbf{w}^*(\mathbf{U}_1,\mathbf{P}_t,\mathbf{x})\mathbf{w}^*(\mathbf{U}_1,\mathbf{P}_t,\mathbf{x})^\top - \mathbf{P}_t\mathbf{U}_2\mathbf{w}^*(\mathbf{U}_1,\mathbf{P}_t,\mathbf{x})\mathbf{w}^*(\mathbf{U}_1,\mathbf{P}_t,\mathbf{x})^\top\| \\
&\le \|\mathbf{P}_t\mathbf{U}_1\|\|\mathbf{w}^*(\mathbf{U}_1,\mathbf{P}_t,\mathbf{x})\|\|\mathbf{w}^*(\mathbf{U}_1,\mathbf{P}_t,\mathbf{x}) - \mathbf{w}^*(\mathbf{U}_2,\mathbf{x})\| \\
&\qquad + \|\mathbf{P}_t\mathbf{U}_1\mathbf{w}^*(\mathbf{U}_1,\mathbf{P}_t,\mathbf{x}) - \mathbf{P}_t\mathbf{U}_2\mathbf{w}^*(\mathbf{U}_1,\mathbf{P}_t,\mathbf{x})\|\|\mathbf{w}^*(\mathbf{U}_1,\mathbf{P}_t,\mathbf{x})\| \\
&\le c_1\alpha_3\alpha_4\|\mathbf{P}_t(\mathbf{U}_1-\mathbf{U}_2)\| + \|\mathbf{w}^*(\mathbf{U}_1,\mathbf{P}_t,\mathbf{x})\|\big(\|\mathbf{P}_t\mathbf{U}_1\|\|\mathbf{w}^*(\mathbf{U}_1,\mathbf{P}_t,\mathbf{x}) - \mathbf{w}^*(\mathbf{U}_2,\mathbf{P}_t,\mathbf{x})\| \\
&\qquad + \|\mathbf{P}_t(\mathbf{U}_1-\mathbf{U}_2)\|\|\mathbf{w}^*(\mathbf{U}_1,\mathbf{P}_t,\mathbf{x})\|\big) \\
&\le \alpha_3\alpha_4\|\mathbf{P}_t(\mathbf{U}_1-\mathbf{U}_2)\| + \alpha_3\big(c_1\alpha_4\|\mathbf{P}_t(\mathbf{U}_1-\mathbf{U}_2)\| + \alpha_3\|\mathbf{P}_t(\mathbf{U}_1-\mathbf{U}_2)\|\big) \\
&= (2c_1\alpha_3\alpha_4 + \alpha_3^2)\|\mathbf{P}_t(\mathbf{U}_1-\mathbf{U}_2)\| \le (2c_1\alpha_3\alpha_4 + \alpha_3^2)\|(\mathbf{U}_1-\mathbf{U}_2)\|, \qquad (40)
\end{aligned}$$

the bounds for the two latter terms are

$$\begin{aligned}
&\|\mathbf{s}^*(\mathbf{U}_1,\mathbf{P}_t,\mathbf{x})\mathbf{w}^*(\mathbf{U}_1,\mathbf{P}_t,\mathbf{x})^\top - \mathbf{s}^*(\mathbf{U}_2,\mathbf{P}_t,\mathbf{x})\mathbf{w}^*(\mathbf{U}_2,\mathbf{P}_t,\mathbf{x})^\top\| \\
&\le \|\mathbf{s}^*(\mathbf{U}_1,\mathbf{P}_t,\mathbf{x})\|\|\mathbf{w}^*(\mathbf{U}_1,\mathbf{P}_t,\mathbf{x})^\top - \mathbf{w}^*(\mathbf{U}_2,\mathbf{P}_t,\mathbf{x})^\top\| + \|\mathbf{s}^*(\mathbf{U}_1,\mathbf{P}_t,\mathbf{x}) - \mathbf{s}^*(\mathbf{U}_2,\mathbf{P}_t,\mathbf{x})\|\|\mathbf{w}^*(\mathbf{U}_2,\mathbf{P}_t,\mathbf{x})\| \\
&\le c_1\alpha_2\|\mathbf{P}_t(\mathbf{U}_1-\mathbf{U}_2)\| + c_2\alpha_3\|\mathbf{P}_t(\mathbf{U}_1-\mathbf{U}_2)\| \\
&\le (c_1\alpha_2 + c_2\alpha_3)\|\mathbf{P}_t(\mathbf{U}_1-\mathbf{U}_2)\| + \alpha_3\|(\mathbf{U}_1-\mathbf{U}_2)\|, \qquad (41)
\end{aligned}$$

and

$$\|\mathbf{x}\mathbf{w}^*(\mathbf{U}_1,\mathbf{P}_t,\mathbf{x})^\top - \mathbf{x}\mathbf{w}^*(\mathbf{U}_2,\mathbf{P}_t,\mathbf{x})^\top\| \le c_1\alpha_1\|(\mathbf{U}_1-\mathbf{U}_2)\|. \qquad (42)$$

From (40), (41) and (42), we can conclude the inequality (38).

From the three facts above above and $g_t(\mathbf{U}_t) \overset{a.s.}{\to} f_t(\mathbf{U}_t)$ when $t \to \infty$, we have $\nabla e_t(\mathbf{U}_t) = \mathbf{0}$ and hence the following inequality

$$|\nabla e_t(\mathbf{U})| \le \frac{L}{2}\|\mathbf{U}-\mathbf{U}_t\|_F.$$

It is therefore that

$$\frac{|e_t(\mathbf{U}) - e_t(\mathbf{U}_t)|}{\|\mathbf{U}-\mathbf{U}_t\|_F} \le \frac{L}{2}\|\mathbf{U}-\mathbf{U}_t\|_F, \qquad (43)$$

thanks to the mean value theorem. In other word, we have $|e_t(\mathbf{U})| \le \frac{L}{2}\|\mathbf{U}-\mathbf{U}_t\|_F^2$ because of $e_t(\mathbf{U}_t) \overset{a.s.}{\to} 0$. In addition, for all $\mathbf{U} \in \mathbf{R}^{n\times r}$, we always have $f_t(\mathbf{U}_t) \le g_t(\mathbf{U})$. Therefore, we can conclude the corollary as follows

$$f_t(\mathbf{U}_t) \le g_t(\mathbf{U}_t) = f_t(\mathbf{U}) + e_t(\mathbf{U}) \le f_t(\mathbf{U}) + \frac{L}{2}\|\mathbf{U}-\mathbf{U}_t\|_F^2. \qquad (44)$$

It ends the proof. □

VI. Technical Propositions

In this section, we would provide the following propositions which help us to derive several important results in the proofs. Their details are provided in well-known materials. [3], [6]–[10].

Proposition 1. *(Strongly Convex): The function f is strongly convex if and only if for all $\mathbf{u}, \mathbf{v} \in \boldsymbol{dom}(f)$ we always have*

$$f(\mathbf{v}) - f(\mathbf{u}) - \frac{1}{2}\|\mathbf{v}-\mathbf{u}\|_2^2 \ge \langle \mathbf{v}-\mathbf{u}, \boldsymbol{\theta}\rangle, \quad \forall \boldsymbol{\theta} \in \partial f(\mathbf{u}).$$

Proposition 2. *The function f is m-strongly convex, with a constant m if and only if for all $\mathbf{u}, \mathbf{v} \in \textbf{\textit{dom}}(f)$, we always have*

$$|f(\mathbf{v}) - f(\mathbf{u})| \geq \frac{m}{2} \|\mathbf{v} - \mathbf{u}\|_2^2 .$$

Proposition 3. *Every norm on $\mathbb{R}^n$ is convex and the sum of convex functions is convex.*

Proposition 4. *(Lipschitz Function) A function $f : \mathcal{V} \to \mathbf{R}$ is called Lipschitz function if there exist a a positive number $L > 0$ such that for all $\mathbf{A}, \mathbf{B} \in \mathcal{V}$, we always have*

$$|f(\mathbf{A}) - f(\mathbf{B})| \leq L\|\mathbf{A} - \mathbf{B}\|.$$

Proposition 5. *(Huber Function): The Huber penalty function replaces the ℓ_1-norm $\|\mathbf{x}\|_1, \mathbf{x} \in \mathbb{R}^n$ is given by the sum $\sum_{i=1}^n f_\mu^{Hub}(x(i))$, where*

$$f_\mu^{Hub}(x(i)) = \begin{cases} \frac{x(i)^2}{2\mu}, & |x(i)| \leq \mu, \\ |x(i)| - \mu/2, & |x| > \mu. \end{cases}$$

There exists a smooth version of the Huber function f_μ^{Hub}, which has derivatives of all degrees, i.e.,

$$\psi_\mu(\mathbf{x}) = \sum_{i=1}^{n} \big((x(i)^2 + \mu^2)^{1/2} - \mu\big).$$

and the first derivative of the pseudo-Huber function ψ_μ is defined by

$$\nabla\psi_\mu(\mathbf{x}) = \big[x(1)(x(1)^2 + \mu^2)^{-1/2}, \ldots, x(n)(x(n)^2 + \mu^2)^{-1/2}\big]^\top.$$

Proposition 6. *Let $\mathcal{V}$ and $\mathcal{W}$ are two vector spaces, and $\mathcal{U} \subset \mathcal{V}$. A function $f : \mathcal{U} \to \mathcal{W}$ is called (Frechet) differentiable at $\mathbf{x} \in \mathcal{U}$ if there exists a bounded linear map $\mathbf{A} : \mathcal{V} \to \mathcal{W}$ such that*

$$\lim_{\|\mathbf{h}\| \to 0} \frac{\|f(\mathbf{x} + \mathbf{h}) - f(\mathbf{x}) - \mathbf{A}\mathbf{x}\|_{\mathcal{W}}}{\|\mathbf{h}\|_{\mathcal{V}}} = 0.$$

Proposition 7. *(Convergence): Let $\{a_t\}_{t=1}^\infty$ and $\{b_t\}_{t=1}^\infty$ be two nonnegative sequences such that $\sum_{i=1}^\infty a_i = \infty$ and $\sum_{i=1}^\infty a_i b_i < \infty$, $|b_{t+1} - b_t| < K a_t$ with some constant K, then $\lim_{t\to\infty} b_t = 0$ or $\sum_{i=1}^\infty b_i < \infty$.*

Proposition 8. *(Convergence): If $\{f_t\}_{t\geq 1}$ and $\{g_t\}_{t\geq 1}$ are sequences of bounded functions which converge uniformly on a set $\mathcal{E}$, then $\{f_t + g_t\}_{t\geq 1}$ and $\{f_t g_t\}_{t\geq 1}$ converge uniformly on $\mathcal{E}$.*

Proposition 9. *($\mathbb{P}$-Donsker classes, Donsker theorem [6, Section 19.2]): Let $F = \{\ell_\theta : \mathcal{X} \to \mathbb{R}\}$ be a set of measurable functions defined on a bounded subset of $\mathbb{R}^n$. For every θ_1, θ_2 and x, if there exists a constant c such that*

$$|\ell_{\theta_1}(x) - \ell_{\theta_2}(x)| < c \|\theta_1 - \theta_2\|_2 ,$$

then F is $\mathbb{P}$-Donsker. For any function ℓ in F, let us define the following functions

$$f_t = \frac{1}{t}\sum_{i=1}^{t} \ell(\mathbf{U}_i), \quad \textit{and} \quad f = \mathbb{E}[f_t(\mathbf{U})].$$

Assume that for all ℓ, $\|\ell\|_\infty < M$ and random variables $\{\mathbf{U}_i\}_{i\geq 1}$ are Borel-measurable, we then have

$$\mathbb{E}[\sqrt{t}\|f_t - f\|_\infty] = \mathcal{O}(1),$$

where $\|\ell\|_\infty \triangleq \inf\{C \geq 0, |f(x)| < C \ \forall\ x\}$.

Proposition 10. *(Quasi Martingales [11, Section 4.4]): Let $(\Omega, \mathcal{F}, \mathbb{P})$ be a probability space, $\{u_t\}_{t>0}$ be a stochastic process on the probability space and $\{\mathcal{F}_t\}_{t>0}$ be a filtration by the past information at time instant t. Let us define the indicator function δ_t as follows*

$$\delta_t \triangleq \begin{cases} 1 & \text{if} \quad \mathbb{E}[u_{t+1} - u_t | \mathcal{F}_t] > 0, \\ 0 & \text{otherwise.} \end{cases}$$

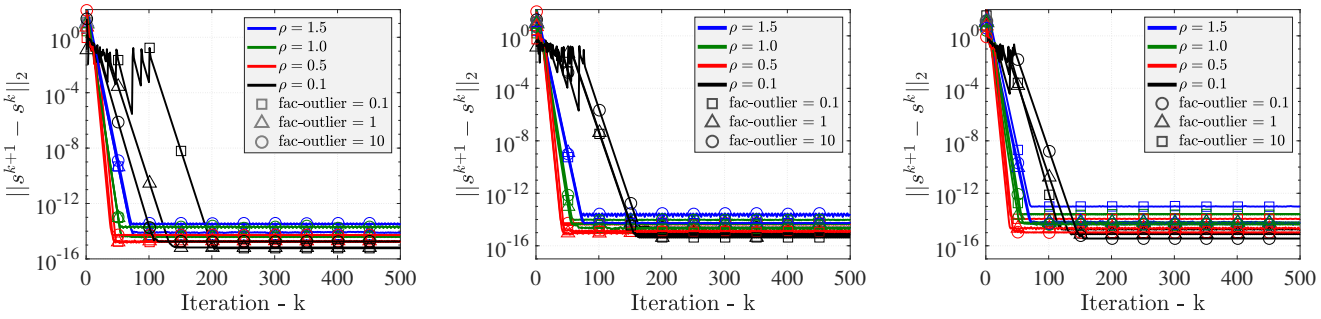
For all t, if $u_t \geq 0$ and $\sum_{i=1}^{\infty} \mathbb{E}[\delta_i (u_{i+1} - u_i) | \mathcal{F}_i] < \infty$, then u_t is a quasi-martingale and converges almost surely, i.e.,

$$\sum_{t=1}^{\infty} \mathbb{E}[u_{t+1} - u_t | \mathcal{F}_t] < \infty.$$

VII. Additional Experimental Results

A. Convergence of PETRELS-ADMM

Fig. 1 shows the typical convergence behavior of PETRELS-ADMM at three noise levels (i.e. $\mathrm{SNR} = \{0, 10, 20\}$ dB) w.r.t the two variables: $\mathrm{fac\text{-}outlier}$ and the weight ρ. The experimental results are practical evidences of Lemma 1 in the main manuscript. Particularly, the variation of $\{\mathbf{s}^k\}_{k\geq 1}$ always converges in all testing cases (i.e., approximate 10^{-14} on average). When the regularization weight $\rho \geq 0.5$, the convergence rate is fast which the variation $\left\|\mathbf{s}^{k+1} - \mathbf{s}^k\right\|_2$ can converge in 50 iterations in both low- and high-noise cases. Similarly, the variations of the sequence $\{\mathbf{U}_t\}_{t\geq 0}$ generated by PETRELS-ADMM also have asymptotic converged behavior as shown in Fig. 2.



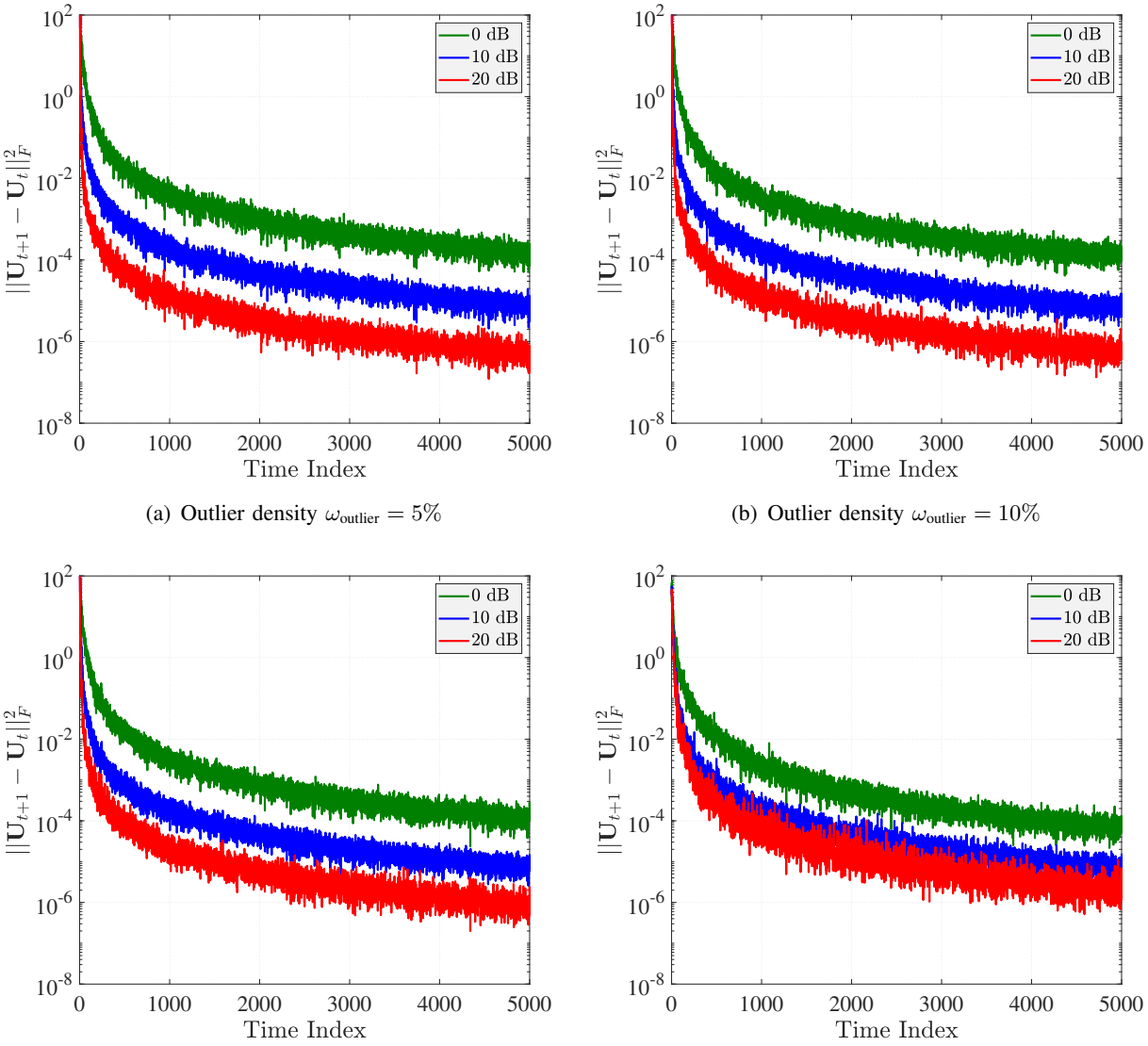
(a) SNR = 0 dB (b) SNR = 10 dB (c) SNR = 20 dB

Fig. 1: Convergence of PETRELS-ADMM in terms of the variation $\left\|\mathbf{s}^{k+1} - \mathbf{s}^k\right\|_2$: $n = 50, r = 2$, 90% entries observed and outlier density of 5%.

B. Outlier Detection

To demonstrate the effectiveness of PETRELS-ADMM, we assess the outlier detection performance of the proposed method in comparison with the well-known GRASTA algorithm [12]. We use a synthetic data whose number of row $n = 50$, rank $r = 2$ and 5000 observations. Outlier density and intensity are varied in the range $[5\% - 40\%]$ and $[0.1, 1, 10]$ respectively, while the value of SNR is set at high (20 dB), moderate (10 dB) and low (5 dB) level.

The results are shown as in Fig. 3-6. Particularly, at low outlier density (e.g. 20%) and high SNR (20 dB), both algorithms can detect outliers effectively. Their detection performance may be degraded when



(a) Outlier density $\omega_{\text{outlier}} = 5\%$

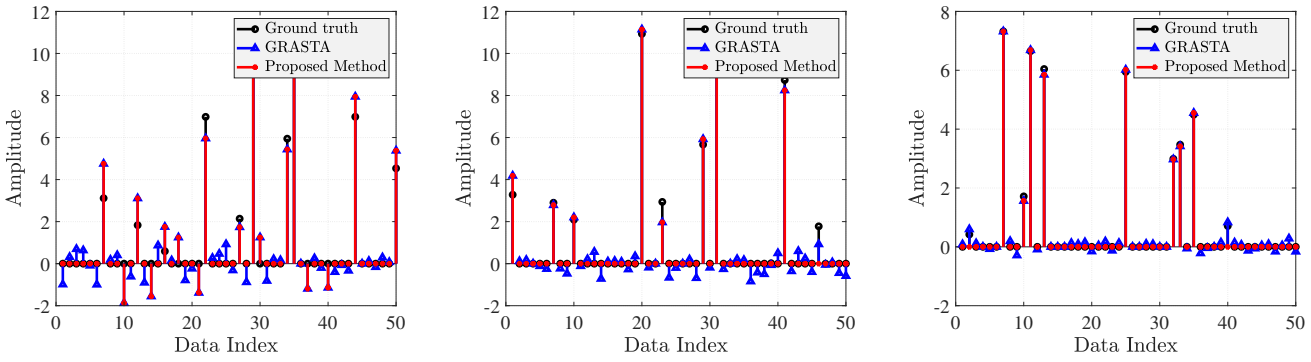
(b) Outlier density $\omega_{\text{outlier}} = 10\%$

(c) Outlier density $\omega_{\text{outlier}} = 20\%$

(d) Outlier density $\omega_{\text{outlier}} = 40\%$

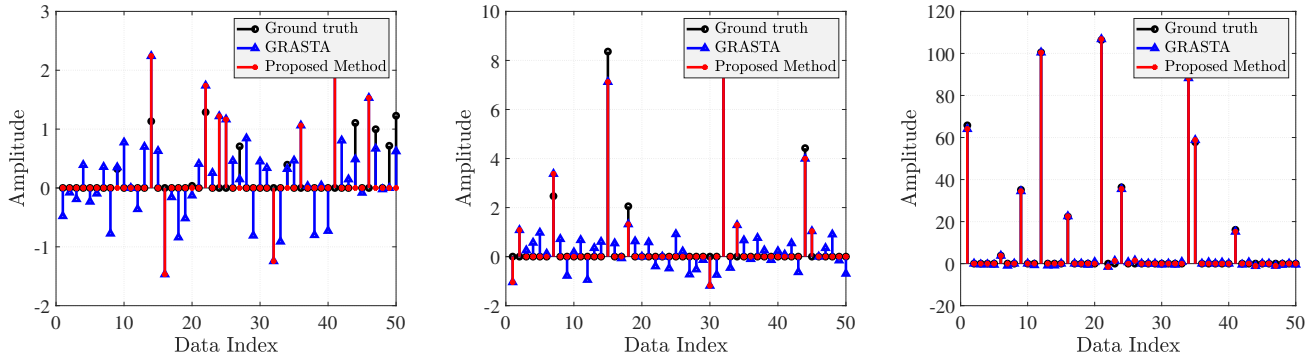
Fig. 2: Convergence of PETRELS-ADMM in terms of the variation $\|\mathbf{U}_{t+1} - \mathbf{U}_t\|_F$: $n = 50, r = 2$, 90% entries observed and outlier intensity $\text{fac-outlier} = 10$.

the effect of the noise is increased (i.e. low SNR). Although the location of outliers can be identified correctly, PERTRELS-ADMM provides better results than GRASTA in terms of sparsity, see Fig. 3(b)-(c). The effect of outlier intensity and density on their outlier detection performance are illustrated in Fig. 4 and Fig. 5 respectively. Our method outperforms GRASTA again. We can see that, when the data is corrupted by "strong" outliers, both methods are able to detect them efficiently. At low SNR, outliers are effectively localized by PETRELS-ADMM even in the presence of high corruptions, while GRASTA labels many locations as outliers, see Fig. 4(a) and Fig. 5(b) for examples. Besides, GRASTA fails to detect outliers in cases of low outlier intensity (e.g. $\text{fac-outlier} = 0.1$), as shown in Fig. 4(a). The overall detection performance of PETRELS-ADMM and GRASTA is reported in Fig. 6.



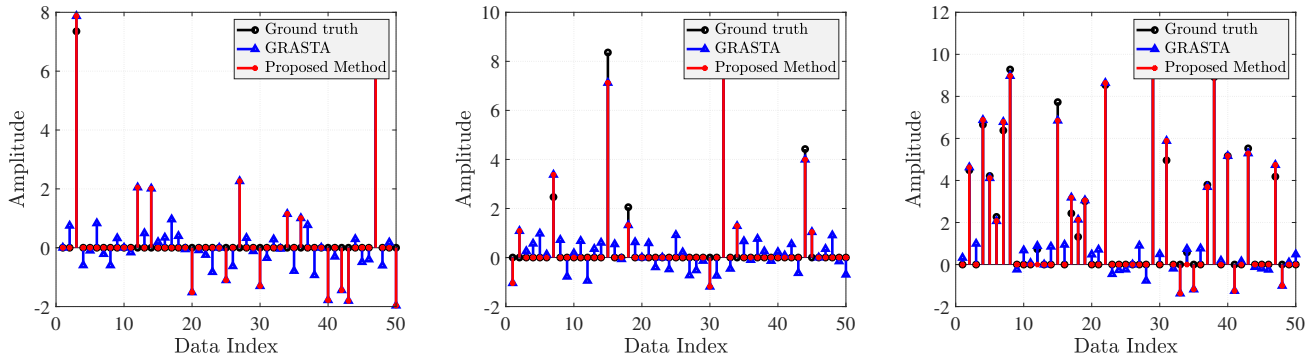
(a) SNR = 5dB (b) SNR = 10dB (c) SNR = 20dB

Fig. 3: Effect of the noise on the outlier detection performance: $n = 50, r = 2$, outlier density of 20% and outlier intensity $\text{fac-outlier} = 1$.



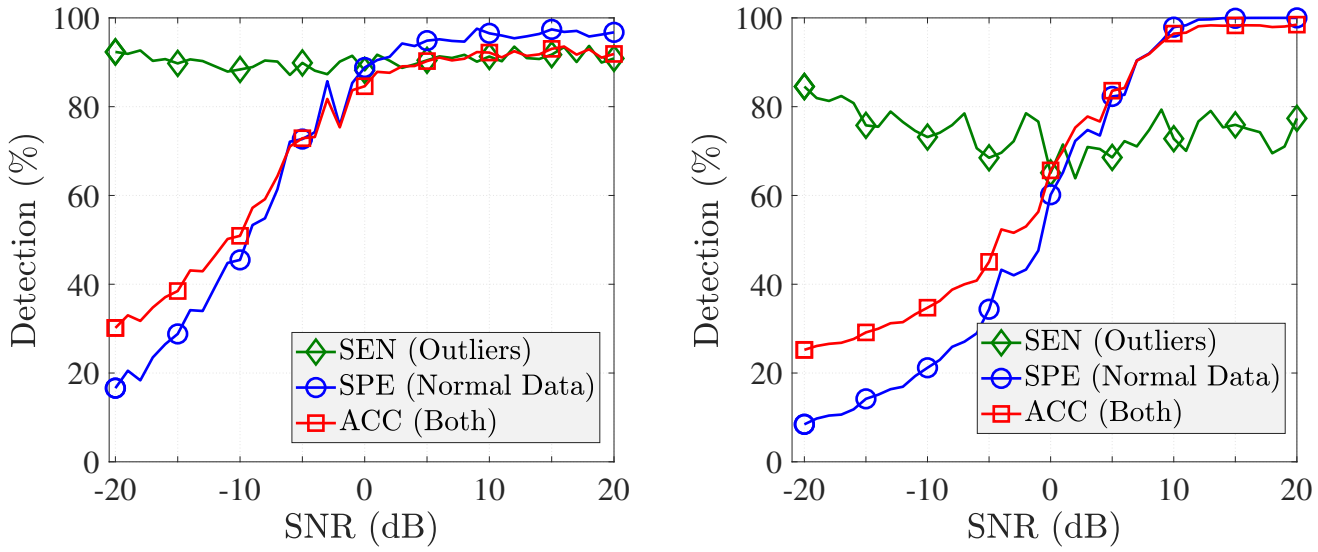
(a) Outlier intensity, fac-outlier = 0.1 (b) Outlier intensity, fac-outlier = 1 (c) Outlier intensity, fac-outlier = 10

Fig. 4: Effect of outlier intensity on the outlier detection performance: $n = 50, r = 2, \text{SNR} = 5$ dB and outlier density of 20%.



(a) Outlier density = 5% (b) Outlier density = 10% (c) Outlier density = 40%

Fig. 5: Effect of outlier density on the outlier detection performance: $n = 50, r = 2, \text{SNR} = 5$ dB and $\text{fac-outlier} = 1$.



(a) PETRELS-ADMM. (b) GRASTA.

Fig. 6: Outlier detection accuracy versus the noise level: $n = 50$, $r = 2$, 80% entries observed and 20% outliers, $\text{fac-outlier} = 1$.

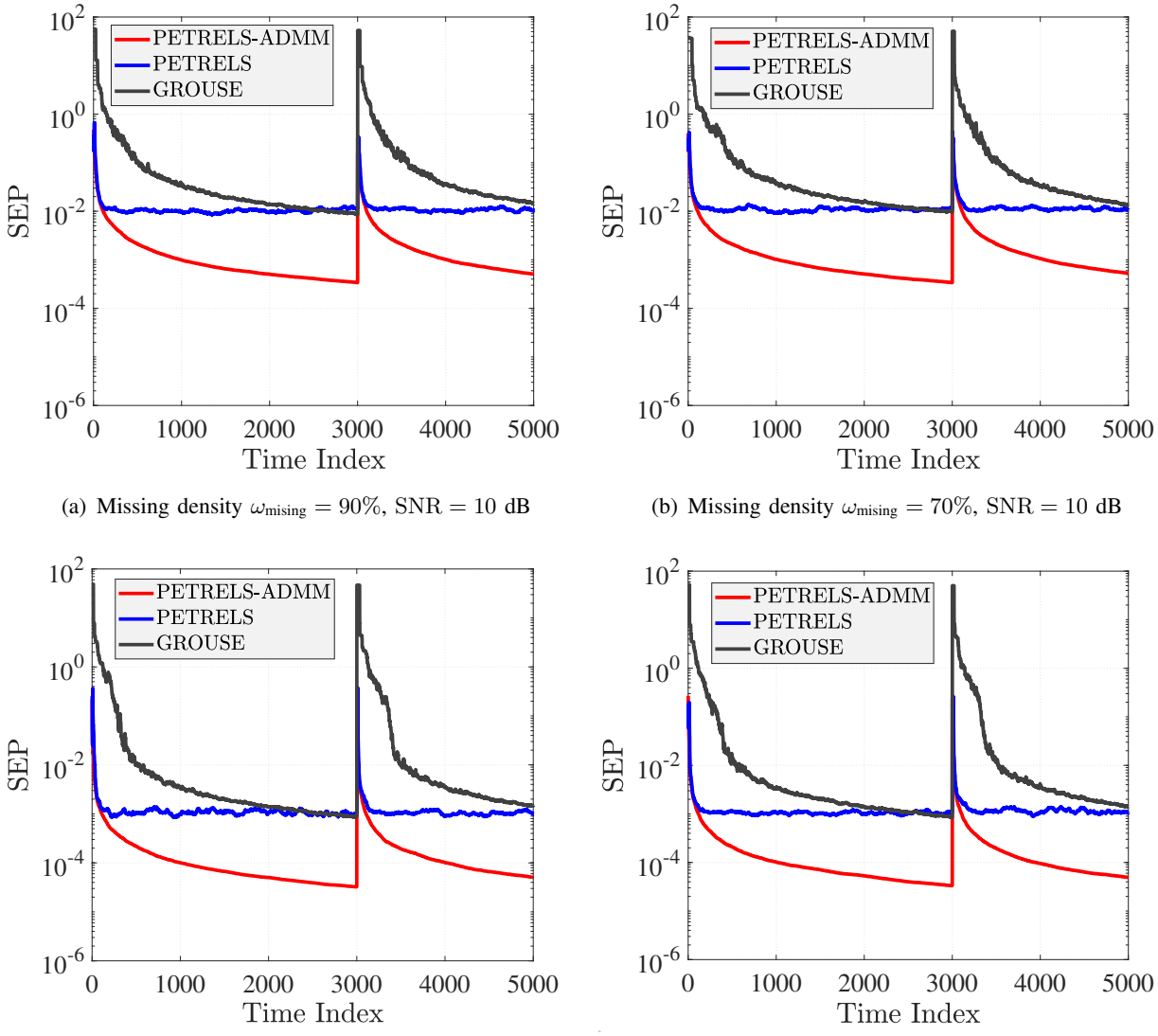
C. Highly Incomplete Observations

In order to illustrate the efficiency and effectiveness of the proposed algorithm for subspace tracking from (very) highly incomplete observations, a performance comparison of PETRELS-ADMM against the original PETRELS [13] and a well-known GROUSE algorithm [14] is conducted. For a fair comparison, effect of outliers is ignored in this task. Following the above experiments, we consider the same data model of $n = 500$, rank $r = 10$ and 5000 observations. The underlying subspace is corrupted abruptly at the time index 3000. The noise level SNR is set at 10 dB and 20 dB.

The results are shown as in Fig. 7. All three algorithms can successfully track the underlying subspace, but PETRELS-ADMM provides better subspace estimation performance than the original PETRELS and GROUSE. Particularly, PETRELS-based algorithms converge faster than GROUSE even with a small number of entries observed at each time. PETRELS-ADMM yields a much better subspace estimation accuracy than the original PETRELS in terms of SEP metric, see Fig. 7.

D. Robustness of PETRELS-ADMM at low Signal-to-Noise Ratio (SNR*)*

Following the same experiment setup in the main manuscript, we demonstrate the effectiveness of PETRELS-ADMM against the-state-of-the-art algorithms at low SNR levels (e.g. $\text{SNR} \in \{0, 5, 10\}$ dB). In particular, the performance of PETRELS-ADMM is investigated with respect to three main aspects: (i) impact of outlier intensity, (ii) impact of outlier density, and (iii) impact of missing density on the subspace estimation accuracy. The results are illustrated in Fig. 8, 9, and 10. In the same manner as in cases of high SNR (20 dB), PETRELS-ADMM outperforms the state-of-the-art subspace tracking algorithms again.



(a) Missing density $\omega_{\text{mising}} = 90\%$, SNR $= 10$ dB

(b) Missing density $\omega_{\text{mising}} = 70\%$, SNR $= 10$ dB

(c) Missing density $\omega_{\text{mising}} = 90\%$, SNR $= 20$ dB

(d) Missing density $\omega_{\text{mising}} = 70\%$, SNR $= 20$ dB

Fig. 7: Performance comparison between the subspace tracking algorithms from highly incomplete observation: $n = 500$, $r = 10$.

REFERENCES

[1] G. Li and T. K. Pong, "Global convergence of splitting methods for nonconvex composite optimization," *SIAM J. Optim.*, vol. 25, no. 4, pp. 2434–2460, 2015.

[2] Y. Wang, W. Yin, and J. Zeng, "Global convergence of ADMM in nonconvex nonsmooth optimization," *J. Sci. Comput.*, vol. 78, no. 1, pp. 29–63, 2019.

[3] S. Boyd and L. Vandenberghe, *Convex optimization*. Cambridge University Press, 2004.

[4] J. Mairal, F. Bach, J. Ponce, and G. Sapiro, "Online learning for matrix factorization and sparse coding," *J. Mach. Learn. Res.*, vol. 11, no. Jan, pp. 19–60, 2010.

[5] K. Knopp, *Theory and application of infinite series*. Courier Corporation, 2013.

[6] A. W. Van der Vaart, *Asymptotic statistics*. Cambridge University Press, 2000.

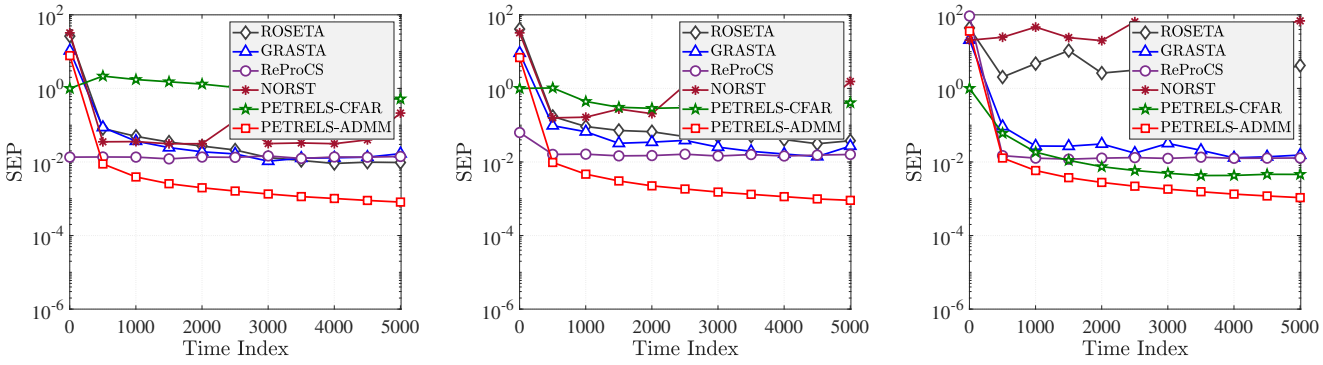
[7] S. Shalev-Shwartz and Y. Singer, "Online learning: Theory, algorithms, and applications," 2007.

[8] K. Fountoulakis and J. Gondzio, "A second-order method for strongly convex ℓ_1-regularization problems," *Math. Program.*, vol. 156, no. 1-2, pp. 189–219, 2016.

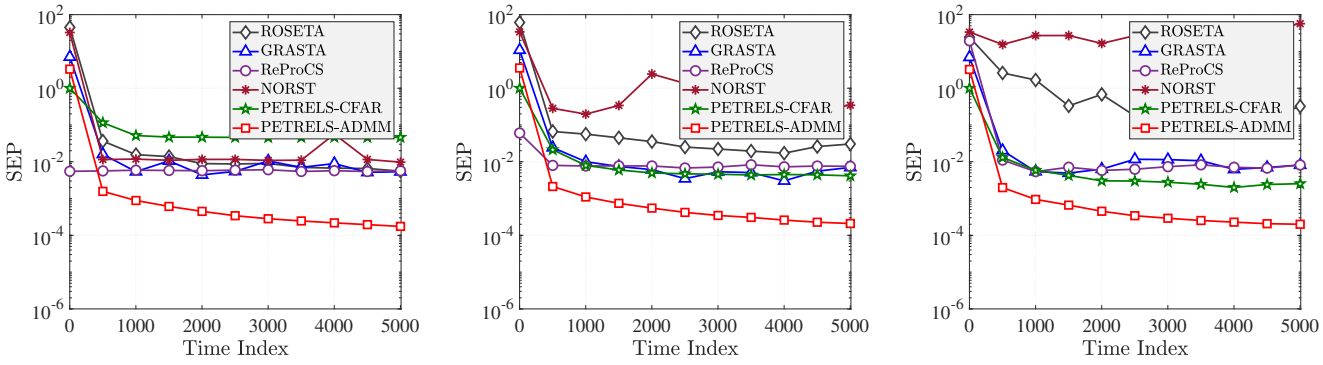
[9] D. P. Bertsekas, "Nonlinear programming," *J Oper. Res. Soc.*, vol. 48, no. 3, pp. 334–334, 1997.

[10] R. Coleman, *Calculus on normed vector spaces*. Springer Science & Business Media, 2012.

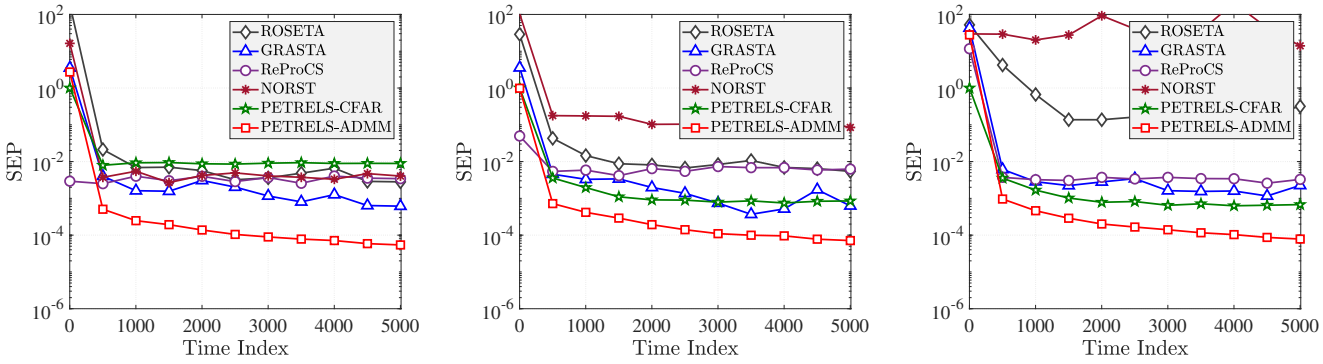
[11] L. Bottou, *On-Line Learning and Stochastic Approximations*. USA: Cambridge University Press, 1999, p. 9–42.



(a) SNR $= 0$ dB and fac-outlier $= 0.1$. (b) SNR $= 0$ dB and fac-outlier $= 1$. (c) SNR $= 0$ dB and fac-outlier $= 10$.



(d) SNR $= 5$ dB and fac-outlier $= 0.1$. (e) SNR $= 5$ dB and fac-outlier $= 1$. (f) SNR $= 5$ dB and fac-outlier $= 10$.



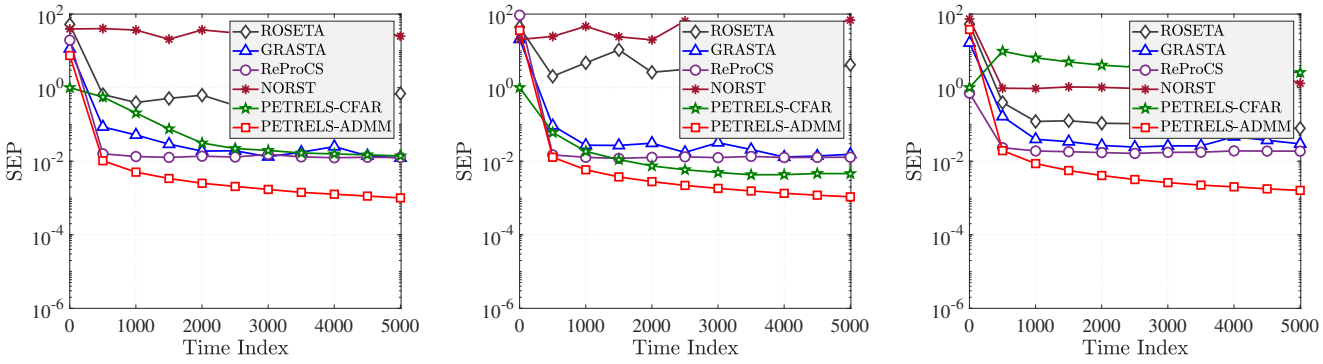
(g) SNR $= 10$ dB and fac-outlier $= 0.1$. (h) SNR $= 10$ dB and fac-outlier $= 1$. (i) SNR $= 10$ dB and fac-outlier $= 10$.

Fig. 8: Impact of outlier intensity on algorithm performance at different (low) noise levels (SNR is chosen among $\{0, 5, 10\}$ dB): $n = 50$, $r = 2$, 90% entries observed, outlier density $\omega_{\text{outlier}} = 0.1$.

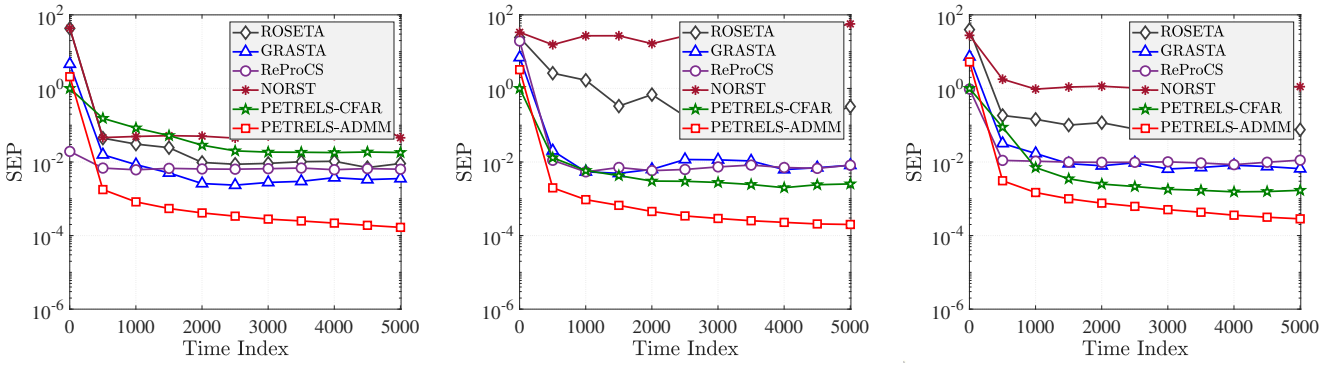
[12] J. He, L. Balzano, and A. Szlam, "Incremental gradient on the Grassmannian for online foreground and background separation in subsampled video," in *IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit.* IEEE, 2012, pp. 1568–1575.

[13] Y. Chi, Y. C. Eldar, and R. Calderbank, "PETRELS: Parallel subspace estimation and tracking by recursive least squares from partial observations," *IEEE Trans. Signal Process.*, vol. 61, no. 23, pp. 5947–5959, 2013.

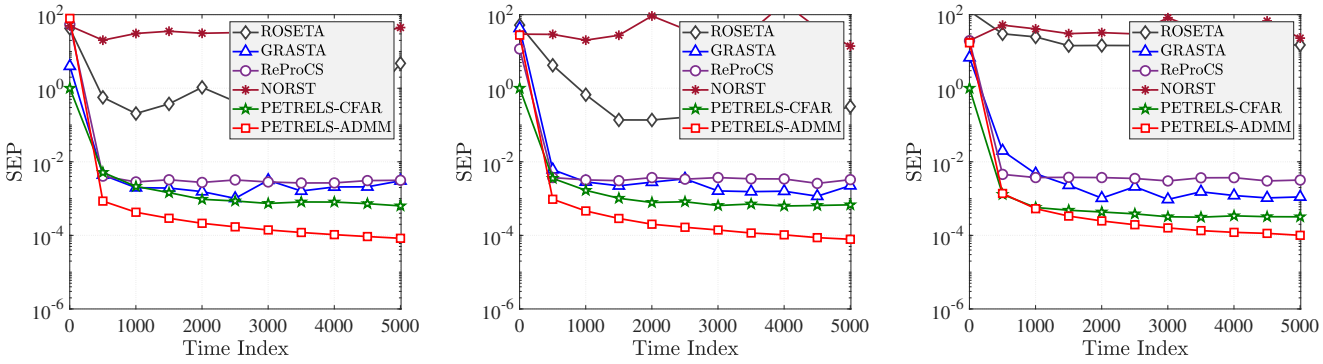
[14] L. Balzano, R. Nowak, and B. Recht, "Online identification and tracking of subspaces from highly incomplete information," in *Ann. Allerton Conf. Commun., Cont. Comput.* IEEE, 2010, pp. 704–711.



(a) SNR $= 0$ dB and $\omega_{\text{outlier}} = 0.05$ (b) SNR $= 0$ dB and $\omega_{\text{outlier}} = 0.1$ (c) SNR $= 0$ dB and $\omega_{\text{outlier}} = 0.2$

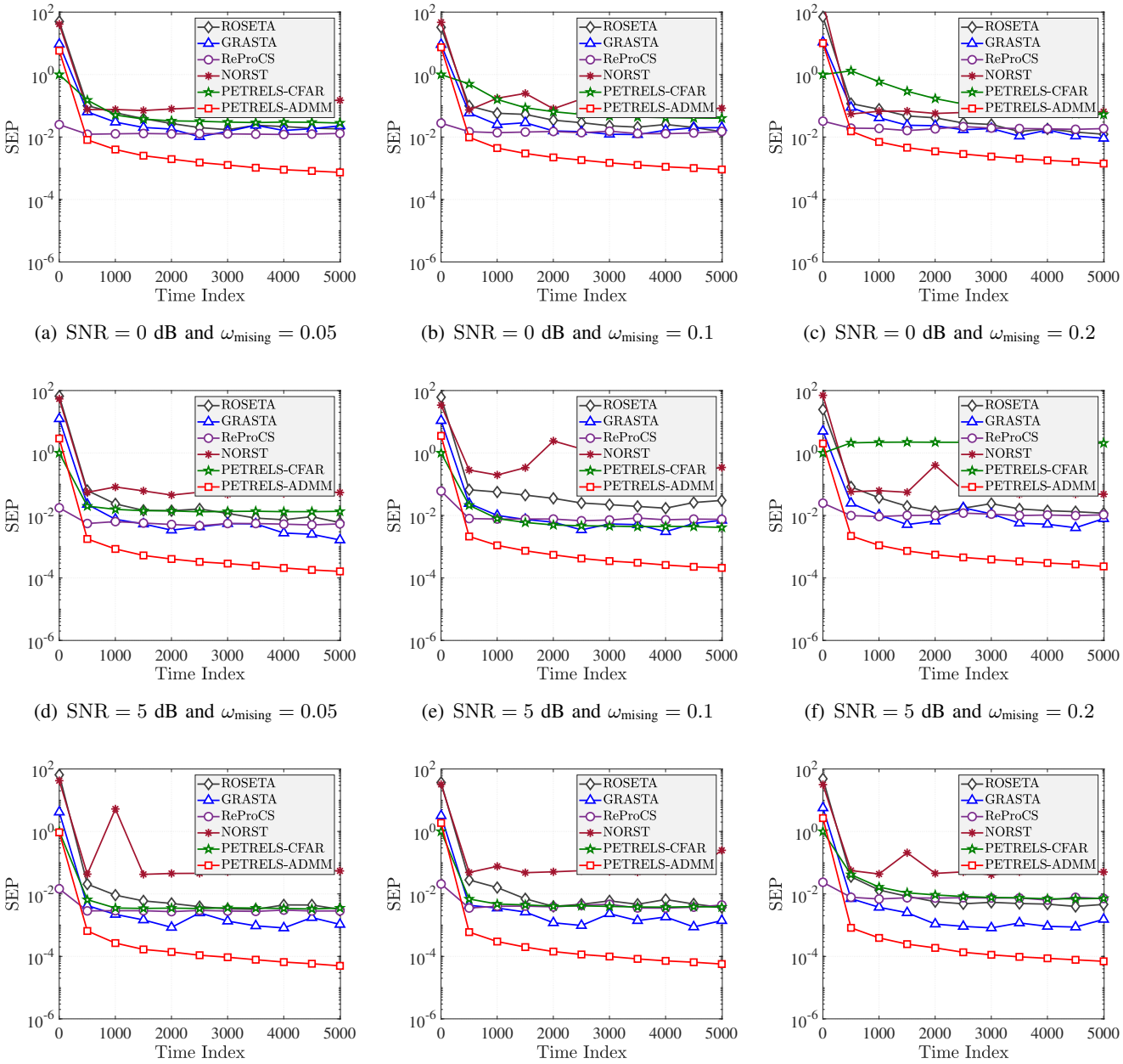


(d) SNR $= 5$ dB and $\omega_{\text{outlier}} = 0.05$ (e) SNR $= 5$ dB and $\omega_{\text{outlier}} = 0.1$ (f) SNR $= 5$ dB and $\omega_{\text{outlier}} = 0.2$



(g) SNR $= 10$ dB and $\omega_{\text{outlier}} = 0.05$ (h) SNR $= 10$ dB and $\omega_{\text{outlier}} = 0.1$ (i) SNR $= 10$ dB and $\omega_{\text{outlier}} = 0.2$

Fig. 9: Impact of outlier density on algorithm performance at different (low) noise levels (SNR is chosen among $\{0, 5, 10\}$ dB): $n = 50$, $r = 2$, 90% entries observed, outlier intensity fac-outlier $= 10$.



(a) SNR $= 0$ dB and $\omega_{\text{mising}} = 0.05$ (b) SNR $= 0$ dB and $\omega_{\text{mising}} = 0.1$ (c) SNR $= 0$ dB and $\omega_{\text{mising}} = 0.2$

(d) SNR $= 5$ dB and $\omega_{\text{mising}} = 0.05$ (e) SNR $= 5$ dB and $\omega_{\text{mising}} = 0.1$ (f) SNR $= 5$ dB and $\omega_{\text{mising}} = 0.2$

(g) SNR $= 10$ dB and $\omega_{\text{mising}} = 0.05$ (h) SNR $= 10$ dB and $\omega_{\text{mising}} = 0.1$ (i) SNR $= 10$ dB and $\omega_{\text{mising}} = 0.2$

Fig. 10: Impact of the density of missing entries on algorithm performance at different (low) noise levels (SNR is chosen among $\{0, 5, 10\}$ dB): $n = 50, r = 2$, outlier density $\omega_{\text{outlier}} = 0.05$, outlier intensity fac-outlier $= 1$.