A. Example Nonconvex Regularizers

Common nonconvex regularizers include capped-ℓ_1 norm (Zhang, 2010b), log-sum-penalty (LSP) (Candès et al., 2008), truncated nuclear norm (TNN) (Hu et al., 2013), smoothed-capped-absolute-deviation (SCAD) (Fan & Li, 2001) and minimax concave penalty (MCP) (Zhang, 2010a). Their definitions are in Table 7 below.

Table 7. Common examples of $\kappa(\sigma_i(\mathbf{X}))$. Here, θ is a constant. For capped-ℓ_1, LSP and MCP, $\theta > 0$; for SCAD, $\theta > 2$; and for TNN, θ is a positive integer.

	$\kappa(\sigma_i(\mathbf{X}))$
capped-ℓ_1 (Zhang, 2010b)	$\min(\sigma_i(\mathbf{X}), \theta)$
LSP (Candès et al., 2008)	$\log(\frac{1}{\theta}\sigma_i(\mathbf{X}) + 1)$
TNN (Hu et al., 2013)	$\begin{cases} \sigma_i(\mathbf{X}) & \text{if } i > \theta \\ 0 & \text{otherwise} \end{cases}$
SCAD (Fan & Li, 2001)	$\begin{cases} \sigma_i(\mathbf{X}) & \text{if } \sigma_i(\mathbf{X}) \le 1 \\ \frac{2\theta\sigma_i(\mathbf{X}) - \sigma_i(\mathbf{X})^2 - 1}{2(\theta-1)} & \text{if } 1 < \sigma_i(\mathbf{X}) \le \theta \\ \frac{(\theta+1)^2}{2} & \text{otherwise} \end{cases}$
MCP (Zhang, 2010a)	$\begin{cases} \sigma_i(\mathbf{X}) - \frac{\alpha^2}{2\theta} & \text{if } \sigma_i(\mathbf{X}) \le \theta \\ \frac{\theta^2}{2} & \text{otherwise} \end{cases}$

B. Proofs

B.1. Proposition 3.1

Proof. Since ϕ is imposed on the unfolding matrix, (13) can be expressed as

$$\mathbf{Z}_t = [\mathcal{Z}_t]_{\langle i \rangle},$$
$$\mathbf{Y}^i_{t+1} = \text{prox}_{\frac{\lambda_i}{\tau}\phi}(\mathcal{Z}_t), \quad i = 1, \dots, K,$$
$$\mathcal{X}_{t+1} = \frac{1}{D}\sum_{d=1}^{D}[\mathbf{Y}^d_{t+1}]^{\langle d \rangle}.$$

(9) can be equivalently rewritten as $\mathcal{Y}^i_{t+1} = \left[\text{prox}_{\frac{\lambda_i}{\tau}\phi}([\mathcal{Z}_t]_{\langle i \rangle})\right]^{\langle i \rangle}$. □

B.2. Proposition 3.2

Proof. For simplicity of exposition, take $\{\pi_1, \pi_2, \pi_3\} = \{1, 2, 3\}$, and consider the case where $\mathbf{U}$ (resp. $\mathbf{V}$) has only one single column $\mathbf{u}$ (resp. $\mathbf{v}$).

We need to fold $\mathbf{u}\mathbf{v}^\top$ along with the first mode and then unfold it along its second mode. In order to avoid the folding and unfolding operations, we consider the structure of $\mathcal{X} = (\mathbf{u}\mathbf{v}^\top)_{\langle 1 \rangle}$. Let $\mathbf{v} = [\mathbf{v}_1; \dots; \mathbf{v}_{I_3}]$, where each $\mathbf{v}_i \in \mathbb{R}^{I_2}$. As $\mathbf{u} \in \mathbb{R}^{I_1}$ and $\mathbf{v} \in \mathbb{R}^{I_2 I_3}$, we have

$$\mathcal{X}_{:,:,i} = \mathbf{v}_i\mathbf{u}^\top.$$

When unfolding $\mathcal{X}$ with the second mode, the unfolding matrix is

$$\left[\mathbf{v}_1\mathbf{u}^\top, \dots, \mathbf{v}_{I_3}\mathbf{u}^\top\right] \in \mathbb{R}^{I_2 \times I_1 I_3}. \tag{21}$$

Thus,

$$\begin{aligned}\mathbf{a}^\top\left[\mathbf{v}_1\mathbf{u}^\top, \dots, \mathbf{v}_{I_3}\mathbf{u}^\top\right] &= \left[(\mathbf{a}^\top\mathbf{v}_1)\mathbf{u}^\top, \cdots, (\mathbf{a}^\top\mathbf{v}_{I_3})\mathbf{u}^\top\right] \\ &= (\mathbf{a}^\top\text{mat}(\mathbf{v})) \otimes \mathbf{u}_p^\top. \end{aligned} \tag{22}$$

Let $\mathbf{b} = [\mathbf{b}_1; \dots; \mathbf{b}_{I_3}]$, where each $\mathbf{b}_i \in \mathbb{R}^{I_1}$. From (21), we have

$$\begin{aligned} & [\mathbf{v}_1\mathbf{u}^\top, \dots, \mathbf{v}_{I_3}\mathbf{u}^\top]\mathbf{b} \\ = & \sum_{i=1}^{I_3}\mathbf{v}_i(\mathbf{u}^\top\mathbf{b}_i) \\ = & [\mathbf{v}_1; \dots; \mathbf{v}_{I_3}]\begin{bmatrix}\mathbf{u}^\top\mathbf{b}_1 \\ \vdots \\ \mathbf{u}^\top\mathbf{b}_{I_3}\end{bmatrix} \\ = & [\mathbf{v}_1; \dots; \mathbf{v}_{I_3}]\left[\mathbf{b}_1; \dots; \mathbf{b}_{I_3}\right]^\top\mathbf{u} \\ = & \text{mat}(\mathbf{v})\text{mat}(\mathbf{b})^\top\mathbf{u}. \end{aligned} \tag{23}$$

When $\mathbf{U}$ (resp. $\mathbf{V}$) has p columns, combining with the fact that $\mathbf{U}\mathbf{V}^\top = \sum_{p=1}^{k}\mathbf{u}_p\mathbf{v}_p^\top$ with (22) and (23), we obtain

$$\mathbf{a}^\top((\mathbf{U}\mathbf{V}^\top)^{\langle 1 \rangle})_{\langle 2 \rangle} = \sum_{p=1}^{k}\mathbf{u}_p^\top \otimes (\mathbf{a}^\top\text{mat}(\mathbf{v}_p)),$$
$$((\mathbf{U}\mathbf{V}^\top)^{\langle 1 \rangle})_{\langle 2 \rangle}\mathbf{b} = \sum_{p=1}^{k}\text{mat}(\mathbf{v}_p)\text{mat}(\mathbf{b})^\top\mathbf{u}_p.$$

The proof does not rely on any specific order of $\{I_1, I_2, I_3\}$. Thus, we can take a permutation of them. □

B.3. Proposition 3.3

Proof. Let $\bar{\lambda}_i = \lambda_i/\tau$. Then,

$$\begin{aligned} & \frac{1}{D}\sum_{d=1}^{D}\text{prox}_{\bar{\lambda}_d\phi}(\mathcal{Z}_{\langle d \rangle}) \\ = & \min_{\{\mathbf{X}_d\}}\frac{1}{D}\sum_{d=1}^{D}\left[\frac{1}{2}\left\|\mathbf{X}_d - \mathcal{Z}_{\langle d \rangle}\right\|_F^2 + \bar{\lambda}_d\phi(\mathbf{X}_d)\right] \\ = & \min_{\{\mathbf{X}_d\}}\frac{1}{2}\|\mathcal{Z}\|_F^2 - \left\langle \mathcal{Z}, \frac{1}{D}\sum_{d=1}^{D}\mathbf{X}_d^{\langle d \rangle}\right\rangle + \frac{1}{2D}\sum_{d=1}^{D}\|\mathbf{X}_d\|_F^2 \\ & + \frac{1}{D}\sum_{d=1}^{D}\bar{\lambda}_d\phi(\mathbf{X}_d) \\ = & \min_{\{\mathbf{X}_d\}}\frac{1}{2}\left\|\mathcal{Z} - \frac{1}{D}\sum_{d=1}^{D}\mathbf{X}_d^{\langle d \rangle}\right\|_F^2 - \frac{1}{2}\left\|\sum_{d=1}^{D}\frac{1}{D}\mathbf{X}_d^{\langle d \rangle}\right\|_F^2 \\ & + \sum_{d=1}^{D}\frac{1}{D}\left[\frac{1}{2}\left\|\mathbf{X}_d^{\langle d \rangle}\right\|_F^2 + \bar{\lambda}_d\phi(\mathbf{X}_d)\right]. \end{aligned} \tag{24}$$

Let $\mathcal{X} = \frac{1}{D}\sum_{d=1}^{D} \mathbf{X}_d^{\langle d \rangle}$. We transform (24) as

$$\min_{\mathcal{X}} \frac{1}{2}\|\mathcal{Z} - \mathcal{X}\|_F^2 + \frac{1}{\tau}\bar{g}(\mathcal{X}) = \text{prox}_{\frac{1}{\tau}\bar{g}}(\mathcal{X}),$$

where $\bar{g}(\mathcal{X})$ is implicitly defined as

$$\bar{g}(\mathcal{X}) = \min_{\{\mathbf{X}_d\}} \frac{\tau}{D}\sum_{d=1}^{D}\left[\frac{1}{2}\left\|\mathbf{X}_d^{\langle d \rangle}\right\|_F^2 + \bar{\lambda}_d\phi(\mathbf{X}_d)\right] - \frac{\tau}{2}\|\mathcal{X}\|_F^2$$
$$\text{s.t. } \frac{1}{D}\sum_{d=1}^{D}\mathbf{X}_d^{\langle d \rangle} = \mathcal{X}. \quad (25)$$

Thus, $\text{prox}_{\frac{1}{\tau}\bar{g}}(\mathcal{Z}) = \frac{1}{D}\sum_{i=1}^{D}\left[\text{prox}_{\bar{\lambda}_d\phi}([\mathcal{Z}]_{\langle i \rangle})\right]^{\langle i \rangle}$. □

B.4. Proposition 3.4

Let $g(\mathcal{X}) = \sum_{d=1}^{D} \frac{\lambda_d}{D}\phi(\mathcal{X}_{\langle d \rangle})$. Before proving Proposition 3.4, we first extend Proposition 2 in (Zhong & Kwok, 2014) in the following Lemma.

Lemma B.1. $0 \leq g(\mathcal{X}) - \bar{g}(\mathcal{X}) \leq \frac{L^2}{2\tau D}\sum_{d=1}^{D}\lambda_d^2$.

Proof. From the definition of $\bar{g}$ in (25), if $\mathcal{X} = \mathbf{X}_1^{\langle 1 \rangle} = \cdots = \mathbf{X}_D^{\langle D \rangle}$, we have

$$\bar{g}(\mathcal{X}) \leq \frac{\tau}{D}\sum_{d=1}^{D}\left[\frac{1}{2}\left\|\mathbf{X}_d^{\langle d \rangle}\right\|_F^2 + \bar{\lambda}_d\phi(\mathbf{X}_d)\right] - \frac{\tau}{2}\|\mathcal{X}\|_F^2$$
$$= \frac{1}{D}\sum_{d=1}^{D}\lambda_d\phi(\mathbf{X}_d) = \frac{1}{D}\sum_{d=1}^{D}\lambda_d\phi(\mathcal{X}_{\langle d \rangle}) = g(\mathcal{X}).$$

Thus, $g(\mathcal{X}) - \bar{g}(\mathcal{X}) \geq 0$.

Next, we prove the "$\leq$" part in the Lemma. Note that

$$\sup_{\mathbf{X}_d} \lambda_d\phi(\mathbf{X}_d) - \tau\text{prox}_{\bar{\lambda}_d\phi}(\mathbf{X}_d^{\langle d \rangle})$$
$$= \sup_{\mathbf{X}_d} \lambda_d\phi(\mathbf{X}_d) - \tau\min_{\mathbf{X}}\left[\frac{1}{2}\left\|\mathbf{X} - \mathbf{X}_d^{\langle d \rangle}\right\|_F^2 + \bar{\lambda}_d\phi(\mathbf{X})\right]$$
$$= \sup_{\mathbf{X}_d, \mathbf{X}} \lambda_d\phi(\mathbf{X}_d) - \frac{\tau}{2}\left\|\mathbf{X} - \mathbf{X}_d^{\langle d \rangle}\right\|_F^2 - \lambda_d\phi(\mathbf{X}). \quad (26)$$

Since ϕ is L-Lipschitz continuous, let $\alpha = \left\|\mathbf{X} - \mathbf{X}_d^{\langle d \rangle}\right\|_F$, we have

$$(26) = \sup_{\mathbf{X}_d, \mathbf{X}} \lambda_d\left[\phi(\mathbf{X}_d) - \phi(\mathbf{X})\right] - \frac{\tau}{2}\left\|\mathbf{X} - \mathbf{X}_d^{\langle d \rangle}\right\|_F^2$$
$$\leq \sup_{\mathbf{X}_d, \mathbf{X}} \lambda_d L\left\|\mathbf{X} - \mathbf{X}_d^{\langle d \rangle}\right\|_F - \frac{\tau}{2}\left\|\mathbf{X} - \mathbf{X}_d^{\langle d \rangle}\right\|_F^2$$
$$= \sup_{\alpha}\left[\lambda_d L\alpha - \frac{\tau}{2}\alpha^2\right]$$
$$= \sup_{\alpha} -\frac{1}{2}\left[\alpha - \frac{\lambda_d L}{\tau}\right]^2 + \frac{\lambda_d^2 L^2}{2} \leq \frac{\lambda_d^2 L^2}{2\tau}. \quad (27)$$

Next, we have

$$g(\mathcal{X}) - \bar{g}(\mathcal{X}) \leq g(\mathcal{X}) - \tau\text{prox}_{\frac{1}{\tau}\bar{g}}(\mathcal{X})$$
$$= \frac{1}{D}\sum_{d=1}^{D}\lambda_d\phi(\mathcal{X}_{\langle d \rangle}) - \frac{\tau}{D}\sum_{d=1}^{D}\text{prox}_{\bar{\lambda}_d\phi}(\mathcal{X}_{\langle d \rangle})$$
$$\leq \frac{1}{D}\sum_{d=1}^{D}\sup_{\mathbf{X}_d}\left[\lambda_d\phi(\mathbf{X}_d) - \tau\text{prox}_{\bar{\lambda}_d\phi}(\mathbf{X}_d)\right]$$
$$\leq \frac{1}{D}\sum_{d=1}^{D}\frac{\lambda_d^2 L^2}{2\tau},$$

where the second inequality comes from (27). Thus, we get the second inequality in the Lemma. □

Now, we prove Proposition 3.4.

Proof. First, we have

$$\min_{\mathcal{X}} F(\mathcal{X}) - \min_{\mathcal{X}} F_\tau(\mathcal{X}) \geq \min_{\mathcal{X}} F(\mathcal{X}) - F_\tau(\mathcal{X})$$
$$= g(\mathcal{X}) - \bar{g}(\mathcal{X}) \geq 0.$$

Let $\mathcal{X}_1 = \arg\min_{\mathcal{X}} F(\mathcal{X})$ and $\mathcal{X}_\tau = \arg\min_{\mathcal{X}} F_\tau(\mathcal{X})$. Then, we have

$$\min_{\mathcal{X}} F(\mathcal{X}) - \min_{\mathcal{X}} F_\tau(\mathcal{X}) = F(\mathcal{X}_1) - F_\tau(\mathcal{X}_\tau)$$
$$\leq F(\mathcal{X}_\tau) - F_\tau(\mathcal{X}_\tau)$$
$$= g(\mathcal{X}_\tau) - \bar{g}(\mathcal{X}_\tau)$$
$$\leq \frac{L^2}{2\tau D}\sum_{d=1}^{D}\lambda_d^2.$$

Thus, $0 \leq \min F - \min F_\tau \leq \frac{L^2}{2\tau D}\sum_{d=1}^{D}\lambda_d^2$. □

B.5. Theorem 3.5

First, we introduce the following Lemmas, which are basic properties for the proximal step.

Lemma B.2 ((Parikh & Boyd, 2013))**.** *Let $\tau > \rho + DL$ and $\eta = \tau - \rho + DL$. Then,*

$$F_\tau(\text{prox}_{\frac{1}{\tau}\bar{g}}(\mathcal{X})) \leq F_\tau(\mathcal{X}) - \frac{\eta}{2}\left\|\mathcal{X} - \text{prox}_{\frac{1}{\tau}\bar{g}}(\mathcal{X})\right\|_F^2.$$

Lemma B.3 ((Parikh & Boyd, 2013))**.** *If $\mathcal{X} = \text{prox}_{\frac{1}{\tau}\bar{g}}(\mathcal{X} - \frac{1}{\tau}\nabla f(\mathcal{X}))$, then $\mathcal{X}$ is a critical point of F_τ.*

Lemma B.4 ((Hare & Sagastizábal, 2009))**.** *The proximal map $\text{prox}_{\frac{1}{\tau}\bar{g}}(\mathcal{X})$ is continuous.*

Now, we prove Theorem 3.5.

Proof. Recall that $\text{prox}_{\frac{1}{\tau}\bar{g}}(\mathcal{X}) = \frac{1}{D}\sum_{i=1}^{D}\text{prox}_{\frac{\lambda_i}{\tau}\phi}(\mathcal{X}_{\langle i \rangle})$. From Lemma B.2,

- If step 8 is performed, we have

$$F_\tau(\mathcal{X}_{t+1}) \leq F_\tau(\mathcal{V}_t) - \frac{\eta}{2}\left\|\mathcal{X}_{t+1} - \mathcal{V}_t\right\|_F^2 \leq F_\tau(\mathcal{X}_t) - \frac{\eta}{2}\left\|\mathcal{X}_{t+1} - \mathcal{X}_t\right\|_F^2. \quad (28)$$

- If step 6 is performed,

$$F_\tau(\mathcal{X}_{t+1}) \leq F_\tau(\mathcal{V}_t) - \frac{\eta}{2}\left\|\mathcal{X}_{t+1} - \mathcal{V}_t\right\|_F^2 \leq F_\tau(\bar{\mathcal{X}}_t) - \frac{\eta}{2}\left\|\mathcal{X}_{t+1} - \bar{\mathcal{X}}_t\right\|_F^2 \leq F_\tau(\mathcal{X}_t) - \frac{\eta}{2}\left\|\mathcal{X}_{t+1} - \bar{\mathcal{X}}_t\right\|_F^2. \quad (29)$$

Combining (28) and (29), we have

$$\frac{2}{\eta}(F_\tau(\mathcal{X}_1) - F_\tau(\mathcal{X}_{T+1})) \geq \sum_{j\in\Omega_1(T)}\left\|\mathcal{X}_{t+1} - \bar{\mathcal{X}}_t\right\|_F^2 + \sum_{j\in\Omega_2(T)}\left\|\mathcal{X}_{t+1} - \mathcal{X}_t\right\|_F^2, \quad (30)$$

where $\Omega_1(T)$ and $\Omega_2(T)$ are a partition of $\{1, \cdots, T\}$ such that when $j \in \Omega_1(T)$ step 6 is performed, and when $j \in \Omega_2(T)$ step 8 is performed.

As F_τ is bounded from below and $\lim_{\|\mathcal{X}\|_F\to\infty} F_\tau(\mathcal{X}) = \infty$, taking $T = \infty$ in (30), we have

$$\sum_{j\in\Omega_1(\infty)}\left\|\mathcal{X}_{t+1} - \mathcal{Y}_t\right\|_F^2 + \sum_{j\in\Omega_2(\infty)}\left\|\mathcal{X}_{t+1} - \mathcal{X}_t\right\|_F^2 = c,$$

where

$$c \leq \frac{2}{\eta}\left[F_\tau(\mathcal{X}_1) - F_\tau^{\min}\right]$$

is a positive constant. Thus, the sequence $\{\mathcal{X}_t\}$ is bounded, and it must have limit points. Besides, one of the following three cases must hold.

1. $\Omega_1(\infty)$ is finite, $\Omega_2(\infty)$ is infinite. Let $\mathcal{X}_*$ be a limit point of $\{\mathcal{X}_t\}$, and $\{\mathcal{X}_{j_t}\}$ be a subsequence that converges to $\mathcal{X}_*$. In this case, on using Lemma B.4, we have

$$\lim_{j_t\to\infty}\left\|\mathcal{X}_{j_t+1} - \mathcal{X}_{j_t}\right\|_F^2 = \lim_{j_t\to\infty}\left\|\text{prox}_{\frac{1}{\tau}\bar{g}}(\mathcal{X}_{j_t} - \frac{1}{\tau}\nabla f(\mathcal{X}_{j_t})) - \mathcal{X}_{j_t}\right\|_F^2 = \left\|\text{prox}_{\frac{1}{\tau}\bar{g}}(\mathcal{X}_* - \frac{1}{\tau}\nabla f(\mathcal{X}_*)) - \mathcal{X}_*\right\|_F^2 = 0.$$

   Thus, $\mathcal{X}_* = \text{prox}_{\frac{1}{\tau}\bar{g}}(\mathcal{X}_* - \frac{1}{\tau}\nabla f(\mathcal{X}_*))$, and $\mathcal{X}_*$ is a critical point of F_τ from Lemma B.3.

2. $\Omega_1(\infty)$ is infinite, $\Omega_2(\infty)$ is finite. Let $\mathcal{X}_*$ be a limit point of $\{\mathcal{X}_t\}$, and $\{\mathcal{X}_{j_t}\}$ be a subsequence that converges to $\mathcal{X}_*$. In this case, we have

$$\lim_{j_t\to\infty}\left\|\mathcal{X}_{j_t+1} - \mathcal{Y}_{j_t}\right\|_F^2 = \lim_{j_t\to\infty}\left\|\text{prox}_{\frac{1}{\tau}\bar{g}}(\mathcal{X}_{j_t} - \frac{1}{\tau}\nabla f(\mathcal{X}_{j_t})) - \mathcal{Y}_{j_t}\right\|_F^2 = \left\|\text{prox}_{\frac{1}{\tau}\bar{g}}(\mathcal{X}_* - \frac{1}{\tau}\nabla f(\mathcal{X}_*)) - \mathcal{X}_*\right\|_F^2 = 0.$$

   Thus, $\mathcal{X}_* = \text{prox}_{\frac{1}{\tau}\bar{g}}(\mathcal{X}_* - \frac{1}{\tau}\nabla f(\mathcal{X}_*))$, and $\mathcal{X}_*$ is a critical point of F_τ from Lemma B.3.

3. Both $\Omega_1(\infty)$ and $\Omega_2(\infty)$ are infinite. From the above cases, we can see that either $\Omega_1(\infty)$ or $\Omega_2(\infty)$ is infinite, and limit points are also the critical points of F_τ.

Thus, all limit points of $\{\mathcal{X}_t\}$ are critical points of F_τ. $\square$

B.6. Corollary 3.6

Proof. Since $\mathcal{X}_{t+1} = \text{prox}_{\frac{1}{\tau}\bar{g}}(\mathcal{V}_t - \frac{1}{\tau}\nabla f(\mathcal{V}_t))$, conclusion (i) directly follows from Lemma B.3.

From (30), we have

$$\min_{1,\ldots,T}\left\|\mathcal{X}_{t+1} - \mathcal{V}_t\right\|_F^2 \leq \frac{1}{T}\sum_{t=1\ldots T}\left\|\mathcal{X}_{t+1} - \mathcal{V}_t\right\|_F^2 \leq \frac{2}{\eta T}(F_\tau(\mathcal{X}_1) - F_\tau(\mathcal{X}_{T+1})) \leq \frac{2}{\eta T}(F_\tau(\mathcal{X}_1) - F_\tau^{\min}).$$

Thus, we obtain Conclusion (ii). $\square$

B.7. Theorem 3.7

We first bound ∂F_τ in Lemma B.5.

Lemma B.5. *For iterations in Algorithm 1, we have* $\min_{\mathcal{U}_t\in\partial F_\tau(\mathcal{X}_t)}\|\mathcal{U}_t\|_F \leq (\tau+\rho)\left\|\mathcal{X}_{t+1} - \mathcal{V}_t\right\|_F$.

Proof. Since $\mathcal{X}_{t+1}$ is generated from the proximal step, i.e., $\mathcal{X}_{t+1} = \text{prox}_{\frac{1}{\tau}\bar{g}}(\mathcal{V}_t - \frac{1}{\tau}\nabla f(\mathcal{V}))$, from its optimality condition, we have

$$\mathcal{X}_{t+1} - \left(\mathcal{V}_t - \frac{1}{\tau}\nabla f(\mathcal{V}_t)\right) + \frac{1}{\tau}\partial\bar{g}(\mathcal{X}_{t+1}) \ni \mathbf{0}.$$

Let

$$\mathcal{U}_t = \tau\left[\mathcal{X}_{t+1} - \mathcal{V}_t\right] - \left[\nabla f(\mathcal{V}_t) - \nabla f(\mathcal{X}_{t+1})\right].$$

We have

$$\partial F_\tau(\mathcal{X}_{t+1}) = \left[\nabla f(\mathcal{X}_{t+1}) + \partial\bar{g}(\mathcal{X}_{t+1})\right] \in \mathcal{U}_t.$$

Table 8. Testing RMSE, CPU time and space required for the synthetic data, when I_3 is small.

		$\bar{c} = 50$, sparsity:5.64%			$\bar{c} = 100$, sparsity: 3.09%		
		RMSE	space (MB)	time (sec)	RMSE	space (MB)	time (sec)
convex	PA-APG	0.0141±0.0012	75.5±0.4	257.2±34.8	0.0149±0.0011	302.4±0.1	2131.7±419.9
(nonconvex) capped-ℓ_1	GDPAN	**0.0103±0.0001**	42.4±2.5	64.4±29.5	**0.0103±0.0001**	171.5±2.2	665.4±99.8
	sNORT	**0.0103±0.0001**	**2.3±0.1**	7.1±4.5	**0.0103±0.0001**	**14.0±0.8**	27.9±5.1
	NORT	**0.0103±0.0001**	3.1±0.1	**2.1±1.4**	**0.0103±0.0001**	14.9±0.9	**5.9±1.6**
(nonconvex) LSP	GDPAN	**0.0103±0.0001**	41.8±2.4	59.1±26.4	0.0104±0.0001	172.2±1.5	654.1±214.7
	sNORT	**0.0103±0.0001**	**2.3±0.1**	4.5±1.5	0.0104±0.0001	14.4±0.1	27.9±5.7
	NORT	**0.0103±0.0001**	**2.3±0.1**	**1.6±1.1**	0.0104±0.0001	15.1±0.1	**5.8±2.8**
(nonconvex) TNN	GDPAN	0.0104±0.0001	41.9±1.6	69.3±26.4	0.0104±0.0001	172.1±1.6	615.0±140.9
	sNORT	0.0104±0.0001	2.5±0.1	6.6±3.8	0.0104±0.0001	14.4±0.1	26.2±4.0
	NORT	0.0104±0.0001	2.5±0.1	**1.4±0.3**	**0.0103±0.0001**	15.1±0.1	**5.3±1.5**

Table 9. Testing RMSE, CPU time and space required for the synthetic data, when I_3 is large.

case (ii)		$\hat{c} = 20$, sparsity:4.77%			$\hat{c} = 40$, sparsity:2.70%		
		RMSE	space (MB)	time (sec)	RMSE	space (MB)	time (sec)
convex	PA-APG	0.0110±0.0007	600.8±70.4	250.1±59.6	0.0098±0.0001	4804.5±598.2	6196.4±2033.4
nonconvex (capped-ℓ_1)	GDPAN	**0.0010±0.0001**	423.1±11.4	179.9±21.5	**0.0006±0.0001**	3243.3±489.6	3670.4±225.8
	sNORT	**0.0010±0.0001**	**10.1±0.1**	22.9±1.1	**0.0006±0.0001**	**44.6±0.3**	575.9±70.9
	NORT	**0.0009±0.0001**	14.4±0.1	5.1±0.3	**0.0006±0.0001**	66.3±0.6	89.4±13.4
nonconvex (LSP)	GDPAN	**0.0010±0.0001**	426.9±9.7	177.8±16.4	**0.0006±0.0001**	3009.3±376.2	3794.0±419.5
	sNORT	**0.0010±0.0001**	10.8±0.1	21.8±0.8	**0.0006±0.0001**	**44.6±0.2**	544.2±75.5
	NORT	**0.0010±0.0001**	14.0±0.1	**4.6±0.7**	**0.0006±0.0001**	62.1±0.5	**81.3±24.9**
nonconvex (TNN)	GDPAN	**0.0010±0.0001**	427.3±10.1	184.1±17.7	**0.0006±0.0001**	3009.2±412.2	3922.9±280.1
	sNORT	**0.0010±0.0001**	**10.2±0.1**	21.8±0.9	**0.0006±0.0001**	**44.7±0.2**	554.7±44.1
	NORT	**0.0010±0.0001**	14.4±0.2	**4.8±0.4**	**0.0006±0.0001**	63.1±0.6	**78.0±9.4**

Thus,

$$\begin{aligned}\|\mathcal{U}_t\|_F &\le \tau \|\mathcal{X}_{t+1} - \mathcal{V}_t\|_F + \|\nabla f(\mathcal{V}_t) - \nabla f(\mathcal{X}_{t+1})\|_F \\ &\le (\tau + \rho) \|\mathcal{X}_{t+1} - \mathcal{V}_t\|_F .\end{aligned}$$

□

Now, we prove Theorem 3.7.

Proof. From Theorem 3.5, we have

$$\lim_{T\to\infty} F_\tau(\mathcal{X}_t) = F_\tau^{\min}.$$

Then, from Lemma B.5, we have

$$\lim_{t\to\infty} \min_{\mathcal{U}_t \in \partial F_\tau(\mathcal{X}_t)} \|\mathcal{U}_t\|_F \le \lim_{t\to\infty} (\tau+\rho) \|\mathcal{X}_{t+1} - \mathcal{V}_t\|_F = 0.$$

Thus, for any ϵ, $c > 0$ and $t > t_0$ where t_0 is a sufficiently large positive integer, we have

$$\begin{aligned}\mathcal{X}_t \in \{\mathcal{X} \,|\, &\min_{\mathcal{U} \in \partial F_\tau(\mathcal{X})} \|\mathcal{U}\|_F \le \epsilon, \\ &F_\tau^{\min} < F_\tau(\mathcal{X}) < F_\tau^{\min} + c\}.\end{aligned}$$

Then, the uniformized KL property implies for all $t \ge t_0$,

$$\begin{aligned}1 &\le \psi'\left(F_\tau(\mathcal{X}_{t+1}) - F_\tau^{\min}\right) \min_{\mathcal{U}_t \in \partial F_\tau(\mathcal{X}_t)} \|\mathcal{U}_t\|_F \\ &= \psi'\left(F_\tau(\mathcal{X}_{t+1}) - F_\tau^{\min}\right) (\tau + \rho) \|\mathcal{X}_{t+1} - \mathcal{V}_t\|_F .\end{aligned} \tag{31}$$

Moreover, from Lemma B.2, we have

$$\|\mathcal{X}_{t+1} - \mathcal{V}_t\|_F^2 \le \frac{2}{\eta} \left[F_\tau(\mathcal{V}_t) - F_\tau(\mathcal{X}_{t+1})\right]. \tag{32}$$

Let $r_t = F_\tau(\mathcal{X}_t) - F_\tau^{\min}$, we have

$$\begin{aligned} r_t - r_{t+1} &= F_\tau(\mathcal{X}_t) - F_\tau^{\min} - \left[F_\tau(\mathcal{X}_{t+1}) - F_\tau^{\min}\right] \\ &\ge F_\tau(\mathcal{V}_t) - F_\tau^{\min} - \left[F_\tau(\mathcal{X}_{t+1}) - F_\tau^{\min}\right] \\ &= F_\tau(\mathcal{V}_t) - F_\tau(\mathcal{X}_{t+1}). \end{aligned} \tag{33}$$

Combine (31), (32) and (33), we have

$$\begin{aligned} 1 &\le \left[\psi'(r_t)\right]^2 (\tau+\rho)^2 \|\mathcal{X}_{t+1} - \mathcal{V}_t\|_F^2 \\ &\le \left[\psi'(r_t)\right]^2 \frac{2(\tau+\rho)^2}{\eta} \left[F_\tau(\mathcal{V}_t) - F_\tau(\mathcal{X}_{t+1})\right] \\ &\le \frac{2(\tau+\rho)^2}{\eta} \left[\psi'(r_{t+1})\right]^2 (r_t - r_{t+1}). \end{aligned} \tag{34}$$

Since $\phi(\alpha) = \frac{C}{\beta}\alpha^\beta$, then $\phi'(\alpha) = C\alpha^{\beta-1}$, (34) becomes

$$1 \le d_1 C^2 r_{t+1}^{2\beta-2} (r_t - r_{t+1}),$$

where $d_1 = \frac{2(\tau+\rho)^2}{\eta}$. Finally, it is shown in (Bolte et al., 2014; Li & Lin, 2015; Li et al., 2017) that the sequence $\{r_t\}$ satisfying the above inequality, convergence to zero with different rates stated in the Theorem. □

Table 10. Algorithms compared on the real-world data sets.

	algorithm	model	basic solver
convex	ADMM (Boyd et al., 2011)	overlapped nuclear norm	ADMM
	FaLRTC (Liu et al., 2013)		Accelerated proximal algorithm for the dual problem
	PA-APG (Yu, 2013)		Accelerated PA algorithm
	FFW (Guo et al., 2017)	latent nuclear norm	efficient Frank-Wolfe algorithm
	TR-MM (Nimishakavi et al., 2018)	squared latent nuclear norm	solved in dual with Riemannian optimization
	TenNN (Zhang & Aeron, 2017)	tensor-SVD	ADMM
factorization	RP (Kasai & Mishra, 2016)	Turker decomposition	Riemannian preconditioning
	TMac (Xu et al., 2013)	multiple matrices factorization	alternative minimization
	CP-WOPT (Acar et al., 2011)	CP decomposition	gradient descent
	TMac-TT (Bengua et al., 2017)	tensor-train decomposition	alternative minimization
nonconvex	GDPAN (Zhong & Kwok, 2014)	nonconvex overlapped regularization	nonconvex PA algorithm
	NORT (Algorithm 1)		proposed algorithm

C. Experimental Details

C.1. Computation of $P_{\Omega}\left(\mathcal{X} - \mathcal{O}\right)$

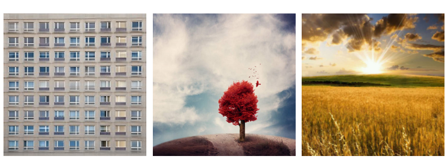
Using (17), each observed element in $P_{\Omega}\left(\mathcal{X}_t - \mathcal{O}\right)$ can be obtained by using Algorithm 2.

Algorithm 2 Computing the pth element in $P_{\Omega}\left(\mathcal{X}_t - \mathcal{O}\right)$.

Require: index $\{i_p^1, i_p^2, i_p^3\}$, factorizations of $\mathbf{Y}_t^1$, $\mathbf{Y}_t^2$, $\mathbf{Y}_t^3$;
$\mathbf{u}_1 \leftarrow$ the i_p^1th row of $\mathbf{U}_t^1$;
$\mathbf{v}_1 \leftarrow$ the $(i_p^2 I_2 + i_p^3)$th row of $\mathbf{V}_t^1$;
$\mathbf{u}_2 \leftarrow$ the i_p^2th row of $\mathbf{U}_t^2$;
$\mathbf{v}_2 \leftarrow$ the $(i_p^3 I_3 + i_p^1)$th row of $\mathbf{V}_t^2$;
if $D = 3$ **then**
  $\mathbf{u}_3 \leftarrow$ the i_p^3th row of $\mathbf{U}_t^3$;
  $\mathbf{v}_3 \leftarrow$ the $(i_p^1 I_1 + i_p^2)$th row of $\mathbf{V}_t^3$;
end if
$o_p \leftarrow p$th element in $P_{\Omega}\left(\mathcal{O}\right)$;
output $v_p = \sum_{i=1}^{D} \alpha_i \mathbf{u}_i^{\top} \mathbf{v}_i - o_p$.

C.2. Color Images

The color images used in Section 4.2.1 are shown in Figure 6.

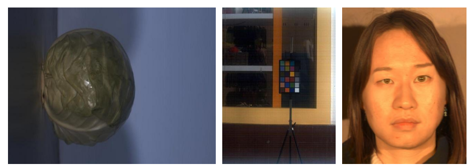


(a) *windows*. (b) *tree*. (c) *rice*.

Figure 6. Color images used in the experiments. All are of size $1000 \times 1000 \times 3$.

C.3. Remote Sensing Data

The hyper-spectral images used in Section 4.2 are shown in Figure 7. The *Female* images are downloaded from `http://www.imageval.com/scene-database-4-faces-3-meters/`, while the *Cabbage* and *Scene* images are from `https://sites.google.com/site/hyperspectralcolorimaging/dataset`.



(a) *Cabbage*. (b) *Scene*. (c) *Female*.

Figure 7. Hyperspectral images used in the experiment. Images are of size $1312 \times 432 \times 49$, $1312 \times 951 \times 49$ and $592 \times 409 \times 148$ respectively.