Supplemental Material for "Online Nonnegative Matrix Factorization with Outliers"

Renbo Zhao, Vincent Y. F. Tan

In this supplemental material, the indices of all the sections, definitions, lemmas and equations are prepended with an 'S' to distinguish those in the main text. The organization of this article is as follows. In section S-1, we drive the ADMM algorithms presented in Section IV-B. In Section S-2, we extend our two solvers (based on PGD and ADMM) in Section IV to the *batch* NMF problems with outliers. Then, we provide detailed proofs of the theorems and lemmas in Section V in Section S-3 to S-8. The technical lemmas used in the algorithm derivation (Section IV) and the convergence analysis (Section V) are shown in Section S-9. Finally, we show additional experiment results in Section S-10 to supplement those in Section VII. The *finite* constants c, c_1 and c_2 are used repeatedly in different sections, and their meanings depend on the context.

S-1. Derivation of the ADMM Algorithms in Section IV-B

A. Algorithms for (8)

Minimizing $\mathbf{h}$ and $\mathbf{r}$ amounts to solving the following two unconstrained problems

$$\begin{aligned} & \min_{\mathbf{h}} \quad \frac{1}{2}\left\|\mathbf{W}\mathbf{h}+(\mathbf{r}-\mathbf{v})\right\|_2^2+\boldsymbol{\alpha}^T(\mathbf{h}-\mathbf{u})+\frac{\rho_1}{2}\left\|\mathbf{h}-\mathbf{u}\right\|_2^2 \\ \iff \quad & \min_{\mathbf{h}} \quad \frac{1}{2}\mathbf{h}^T(\mathbf{W}^T\mathbf{W}+\rho_1\mathbf{I})\mathbf{h}+\left(\mathbf{W}^T(\mathbf{r}-\mathbf{v})-\rho_1\mathbf{u}+\boldsymbol{\alpha}\right)^T\mathbf{h} \end{aligned} \tag{S-1}$$

and

$$\begin{aligned} & \min_{\mathbf{r}} \quad \frac{1}{2}\left\|(\mathbf{v}-\mathbf{W}\mathbf{h})-\mathbf{r}\right\|_2^2+\frac{\widetilde{\rho}_2}{2}\left\|\mathbf{r}-\mathbf{q}\right\|_2^2+\boldsymbol{\beta}^T(\mathbf{r}-\mathbf{q})+\lambda\left\|\mathbf{r}\right\|_1 \\ \iff \quad & \min_{\mathbf{r}} \quad \frac{1+\widetilde{\rho}_2}{2}\mathbf{r}^T\mathbf{r}+(\boldsymbol{\beta}-\mathbf{v}+\mathbf{W}\mathbf{h}-\widetilde{\rho}_2\mathbf{q})^T\mathbf{r}+\lambda\left\|\mathbf{r}\right\|_1 \\ \iff \quad & \min_{\mathbf{r}} \quad \frac{1+\widetilde{\rho}_2}{2}\left\|\mathbf{r}-\frac{\widetilde{\rho}_2\mathbf{q}+\mathbf{v}-\boldsymbol{\beta}-\mathbf{W}\mathbf{h}}{1+\widetilde{\rho}_2}\right\|_2^2+\lambda\left\|\mathbf{r}\right\|_1. \end{aligned} \tag{S-2}$$

We notice that (S-1) is a standard strongly convex quadratic minimization problem and (S-2) is a standard proximal minimization problem with ℓ_1 norm, thus the closed-form optimal solutions for (S-1) and (S-2) are

$$\begin{aligned} \mathbf{h}^* &= (\mathbf{W}^T\mathbf{W}+\rho_1\mathbf{I})^{-1}(\mathbf{W}^T(\mathbf{v}-\mathbf{r})+\rho_1\mathbf{u}-\boldsymbol{\alpha}) \\ \mathbf{r}^* &= \mathcal{S}_{\lambda/(1+\widetilde{\rho}_2)}\left(\frac{\widetilde{\rho}_2\mathbf{q}+\mathbf{v}-\boldsymbol{\beta}-\mathbf{W}\mathbf{h}}{1+\widetilde{\rho}_2}\right)=\frac{\mathcal{S}_{\lambda}(\widetilde{\rho}_2\mathbf{q}+\mathbf{v}-\boldsymbol{\beta}-\mathbf{W}\mathbf{h})}{1+\widetilde{\rho}_2}. \end{aligned}$$

Minimizing $\mathbf{u}$ and $\mathbf{q}$ amounts to solving the following two constrained problems

$$\min_{\mathbf{u}\geq 0} \quad \frac{\rho_1}{2}\left\|\mathbf{h}-\mathbf{u}\right\|_2^2-\boldsymbol{\alpha}^T\mathbf{u} \tag{S-3}$$

$$\min_{\left\|\mathbf{q}\right\|_\infty\leq M} \quad \frac{\widetilde{\rho}_2}{2}\left\|\mathbf{r}-\mathbf{q}\right\|_2^2-\boldsymbol{\beta}^T\mathbf{q}. \tag{S-4}$$

Since both constraints $\mathbf{u}\geq 0$ and $\left\|\mathbf{q}\right\|_\infty\leq M$ are separable across coordinates, we can simply solve the unconstrained quadratic minimization problems and then project the optimal solutions to the feasible sets.

B. Algorithms for (10)

Minimizing $\mathbf{W}$ amounts to solving the unconstrained quadratic minimization problem

$$\begin{aligned} & \min_{\mathbf{W}} \quad \frac{1}{2}\operatorname{tr}\left(\mathbf{W}^T\mathbf{W}\mathbf{A}_t\right)-\operatorname{tr}\left(\mathbf{W}^T\mathbf{B}_t\right)+\langle\mathbf{D},\mathbf{W}-\mathbf{Q}\rangle+\frac{\widetilde{\rho}_3}{2}\left\|\mathbf{W}-\mathbf{Q}\right\|_F^2 \\ \iff \quad & \min_{\mathbf{W}} \quad \frac{1}{2}\operatorname{tr}\left(\mathbf{W}\left(\mathbf{A}_t+\widetilde{\rho}_3\mathbf{I}\right)\mathbf{W}^T\right)-\operatorname{tr}\left(\left(\mathbf{B}_t-\mathbf{D}+\widetilde{\rho}_3\mathbf{Q}\right)^T\mathbf{W}\right) \end{aligned}$$

so

$$\mathbf{W}^*=\left(\mathbf{B}_t-\mathbf{D}+\widetilde{\rho}_3\mathbf{Q}\right)\left(\mathbf{A}_t+\widetilde{\rho}_3\mathbf{I}\right)^{-1}. \tag{S-5}$$

Minimizing $\mathbf{Q}$ amounts to solving the constrained quadratic minimization problem

$$\begin{aligned} & \min_{\mathbf{Q}\in\mathcal{C}} \quad \frac{\widetilde{\rho}_3}{2}\left\|\mathbf{W}-\mathbf{Q}\right\|_F^2-\langle\mathbf{D},\mathbf{Q}\rangle \\ \Longleftrightarrow \quad & \min_{\mathbf{Q}\in\mathcal{C}} \quad \frac{\widetilde{\rho}_3}{2}\left\|\mathbf{Q}-(\mathbf{W}+\mathbf{D}/\widetilde{\rho}_3)\right\|_F^2. \end{aligned}$$

Then we have

$$\mathbf{Q}^*=\mathcal{P}_{\mathcal{C}}(\mathbf{W}+\mathbf{D}/\widetilde{\rho}_3). \tag{S-6}$$

S-2. Extension to the Batch NMF Problem with Outliers

A. Problem Formulation

As usual, we denote the data matrix (with outliers) as $\mathbf{V}$, basis matrix as $\mathbf{W}$, coefficient matrix as $\mathbf{H}$ and outlier matrix as $\mathbf{R}$. The number of total data samples is denoted as N. Then, the batch counterpart for the online NMF problem with outliers can be formulated as follows

$$\begin{aligned} \min \quad & \frac{1}{2}\left\|\mathbf{V}-\mathbf{W}\mathbf{H}-\mathbf{R}\right\|_F^2+\lambda\left\|\mathbf{R}\right\|_{1,1} \\ \text{s.t.} \quad & \mathbf{H}\in\mathbb{R}_+^{K\times N},\mathbf{R}\in\widetilde{\mathcal{R}},\mathbf{W}\in\mathcal{C}, \end{aligned} \tag{S-7}$$

where $\widetilde{\mathcal{R}}=\{\mathbf{R}\in\mathbb{R}^{F\times N}\mid|r_{i,j}|\leq M,\forall\,(i,j)\in[F]\times[N]\}$.

B. Notations

In the sequel, we overload soft-thresholding operators $\widetilde{\mathcal{S}}_{\lambda,M}$ and $\mathcal{S}_\lambda$. When these two operators are applied to matrices, each operator denotes entrywise soft-thresholding. The updated variables are denoted with superscripts '+'.

C. Batch algorithm based on PGD (BPGD)

Based on the principle of block coordinate descent, we update $\mathbf{H}$, $\mathbf{R}$ and $\mathbf{W}$ sequentially as follows.

$$\mathbf{H}^+:=\mathcal{P}_+\left(\mathbf{H}-\eta_1(\mathbf{W})\mathbf{W}^T(\mathbf{W}\mathbf{H}+\mathbf{R}-\mathbf{V})\right) \tag{S-8}$$

$$\mathbf{R}^+:=\widetilde{\mathcal{S}}_{\lambda,M}\left(\mathbf{V}-\mathbf{W}\mathbf{H}^+\right) \tag{S-9}$$

$$\mathbf{W}_{:j}^+:=\frac{\mathcal{P}_+(\mathbf{W}-\eta_2(\mathbf{H}^+)\mathbf{G})_{:j}}{\max\{1,\|\mathcal{P}_+(\mathbf{W}-\eta_2(\mathbf{H}^+)\mathbf{G})_{:j}\|_2\}},\ \forall\, j\in[K], \tag{S-10}$$

where $\mathbf{G}=(\mathbf{W}\mathbf{H}^+ +\mathbf{R}^+-\mathbf{V})\mathbf{H}^{+T}$, $\eta_1(\mathbf{W})\in\left(0,\|\mathbf{W}\|_2^{-2}\right]$ and $\eta_2(\mathbf{H}^+)\in\left(0,\left\|\mathbf{H}^+\mathbf{H}^{+T}\right\|_F^{-1}\right]$.

D. Batch algorithm based on ADMM (BADMM)

(S-7) can be reformulated as

$$\begin{aligned} \min \quad & \frac{1}{2}\left\|\mathbf{V}-\mathbf{W}\mathbf{H}-\mathbf{R}\right\|_F^2+\lambda\left\|\mathbf{R}\right\|_{1,1} \\ \text{s.t.} \quad & \mathbf{H}=\mathbf{U},\mathbf{R}=\mathbf{Q},\mathbf{W}=\mathbf{\Psi},\mathbf{U}\in\mathbb{R}_+^{K\times N},\mathbf{Q}\in\widetilde{\mathcal{R}},\mathbf{\Psi}\in\mathcal{C}. \end{aligned}$$

Thus the augmented Lagrangian is

$$\begin{aligned} \widetilde{\mathcal{L}}(\mathbf{H},\mathbf{R},\mathbf{W},\mathbf{U},\mathbf{Q},\mathbf{\Psi},\mathbf{A},\mathbf{B},\mathbf{\Psi})=&\frac{1}{2}\left\|\mathbf{V}-\mathbf{W}\mathbf{H}-\mathbf{R}\right\|_F^2+\lambda\left\|\mathbf{R}\right\|_{1,1}+\langle\mathbf{A},\mathbf{H}-\mathbf{U}\rangle+\langle\mathbf{B},\mathbf{R}-\mathbf{Q}\rangle+\langle\mathbf{D},\mathbf{W}-\mathbf{\Psi}\rangle \\ &+\frac{\widetilde{\rho}_1}{2}\left\|\mathbf{H}-\mathbf{U}\right\|_F^2+\frac{\widetilde{\rho}_2}{2}\left\|\mathbf{R}-\mathbf{Q}\right\|_F^2+\frac{\widetilde{\rho}_3}{2}\left\|\mathbf{W}-\mathbf{\Psi}\right\|_F^2, \end{aligned} \tag{S-11}$$

where $\mathbf{A}$, $\mathbf{B}$ and $\mathbf{D}$ are dual variables and $\widetilde{\rho}_1$, $\widetilde{\rho}_2$ and $\widetilde{\rho}_3$ are positive penalty parameters. Therefore we can derive the following update rules

$$\mathbf{H}^+ := (\mathbf{W}^T\mathbf{W} + \widetilde{\rho}_1\mathbf{I})^{-1}\left(\mathbf{W}^T(\mathbf{V}-\mathbf{R}) + \widetilde{\rho}_1\mathbf{U} - \mathbf{A}\right) \tag{S-12}$$

$$\mathbf{R}^+ := \mathcal{S}_\lambda(\widetilde{\rho}_2\mathbf{Q} + \mathbf{V} - \mathbf{B} - \mathbf{W}\mathbf{H}^+)/(1+\widetilde{\rho}_2) \tag{S-13}$$

$$\mathbf{W}^+ := \left((\mathbf{V}-\mathbf{R}^+)\mathbf{H}^{+T} - \mathbf{D} + \widetilde{\rho}_3\mathbf{\Psi}\right)\left(\mathbf{H}^+\mathbf{H}^{+T} + \widetilde{\rho}_3\mathbf{I}\right)^{-1} \tag{S-14}$$

$$\mathbf{U}^+ := \mathcal{P}_+\left(\mathbf{H}^+ + \mathbf{A}/\widetilde{\rho}_1\right) \tag{S-15}$$

$$\mathbf{Q}^+ := \mathcal{P}_{\widetilde{\mathcal{R}}}\left(\mathbf{R}^+ + \mathbf{B}/\widetilde{\rho}_2\right) \tag{S-16}$$

$$\mathbf{\Psi}^+ := \mathcal{P}_{\mathcal{C}}\left(\mathbf{W}^+ + \mathbf{D}/\widetilde{\rho}_3\right) \tag{S-17}$$

$$\mathbf{A}^+ := \mathbf{A} + \widetilde{\rho}_1(\mathbf{H}^+ - \mathbf{U}^+) \tag{S-18}$$

$$\mathbf{B}^+ := \mathbf{B} + \widetilde{\rho}_2(\mathbf{R}^+ - \mathbf{Q}^+) \tag{S-19}$$

$$\mathbf{D}^+ := \mathbf{D} + \widetilde{\rho}_3\left(\mathbf{W}^+ - \mathbf{\Psi}^+\right), \tag{S-20}$$

S-3. Proof of Lemma 1

Before proving Lemma 1, we present two lemmas which will be used in the proof. Both lemmas can be proved using straightforward calculations. See Section S-7 and S-8 for detailed proofs.

Lemma S-1. *If for each $\mathbf{v} \in \mathcal{V}$, both $\mathbf{h}^*(\mathbf{v},\mathbf{W})$ and $\mathbf{r}^*(\mathbf{v},\mathbf{W})$ in* (S-21) *are Lipschitz on $\mathcal{C}$, with Lipschitz constants c_1 and c_2 (independent of $\mathbf{v}$) respectively, then $\mathbf{W} \mapsto \nabla_{\mathbf{W}}\ell(\mathbf{v},\mathbf{W})$ is Lipschitz on $\mathcal{C}$ with Lipschitz constant c_3 (independent of $\mathbf{v}$). Consequently, $\nabla f(\mathbf{W})$ in* (S-22) *is Lipschitz on $\mathcal{C}$ with Lipschitz constant c_3.*

Lemma S-2. *Let $\mathbf{z}$, $\mathbf{z}' \in \mathcal{Z} \subseteq \mathbb{R}^m$ and $\mathbf{A}$, $\mathbf{A}' \in \mathcal{A} \subseteq \mathbb{R}^{m\times n}$, where both $\mathcal{Z}$ and $\mathcal{A}$ are compact sets. Let $\mathcal{B}$ be a compact set in $\mathbb{R}^n$, and define $g : \mathcal{B} \to \mathbb{R}$ as $g(\mathbf{b}) = 1/2\left\|\mathbf{z}-\mathbf{A}\mathbf{b}\right\|_2^2 - 1/2\left\|\mathbf{z}'-\mathbf{A}'\mathbf{b}\right\|_2^2$. Then g is Lipschitz on $\mathcal{B}$ with Lipschitz constant $c_1\left\|\mathbf{z}-\mathbf{z}'\right\|_2 + c_2\left\|\mathbf{A}-\mathbf{A}'\right\|_2$, where c_1 and c_2 are two positive constants. In particular, when both $\mathbf{z}'$ and $\mathbf{A}'$ are zero, we have that $\widetilde{g}(\mathbf{b}) = 1/2\left\|\mathbf{z}-\mathbf{A}\mathbf{b}\right\|_2^2$ is Lipschitz on $\mathcal{B}$ with Lipschitz constant c independent of $\mathbf{z}$ and $\mathbf{A}$.*

It is easy to verify that the following conditions hold

1) $\widetilde{\ell}(\mathbf{v},\mathbf{W},\mathbf{h},\mathbf{r})$ is differentiable on $\mathcal{V}\times\mathcal{C}$, for each $(\mathbf{h},\mathbf{r}) \in \mathcal{H}\times\mathcal{R}$,
2) $\widetilde{\ell}(\mathbf{v},\mathbf{W},\mathbf{h},\mathbf{r})$ and $\nabla_{(\mathbf{v},\mathbf{W})}\widetilde{\ell}(\mathbf{v},\mathbf{W},\mathbf{h},\mathbf{r})$ are continuous on $\mathcal{V}\times\mathcal{C}\times\mathcal{H}\times\mathcal{R}$,
3) (31) has unique minimizer $(\mathbf{h}^*(\mathbf{v},\mathbf{W}),\mathbf{r}^*(\mathbf{v},\mathbf{W}))$ for each $(\mathbf{v},\mathbf{W}) \in \mathcal{V}\times\mathcal{C}$, due to Assumption 2.

Thus, we can invoke Danskin's theorem (see Lemma S-5) to conclude that $\ell(\mathbf{v},\mathbf{W})$ is differentiable on $\mathcal{V}\times\mathcal{C}$ and

$$\nabla_{\mathbf{W}}\ell(\mathbf{v},\mathbf{W}) = \left(\mathbf{W}\mathbf{h}^*(\mathbf{v},\mathbf{W}) + \mathbf{r}^*(\mathbf{v},\mathbf{W}) - \mathbf{v}\right)\mathbf{h}^*(\mathbf{v},\mathbf{W})^T. \tag{S-21}$$

Furthermore, we can show $(\mathbf{h}^*(\mathbf{v},\mathbf{W}),\mathbf{r}^*(\mathbf{v},\mathbf{W}))$ is continuous on $\mathcal{V}\times\mathcal{C}$ by the maximum theorem (see Lemma S-6), since the conditions in this theorem are trivially satisfied in our case. Thus, $\nabla_{\mathbf{W}}\ell(\mathbf{v},\mathbf{W})$ is continuous on $\mathcal{V}\times\mathcal{C}$.

Leveraging the regularity of $\ell(\mathbf{v},\mathbf{W})$, we proceed to show the regularity of $f(\mathbf{W})$. Since for all $\mathbf{v} \in \mathcal{V}$, both $\ell(\mathbf{v},\mathbf{W})$ and $\nabla_{\mathbf{W}}\ell(\mathbf{v},\mathbf{W})$ are continuous on $\mathcal{C}$, by Leibniz integral rule (see Lemma S-7), we conclude that $f(\mathbf{W})$ is differentiable on $\mathcal{C}$ and

$$\nabla f(\mathbf{W}) = \mathbb{E}_{\mathbf{v}}[\nabla_{\mathbf{W}}\ell(\mathbf{v},\mathbf{W})]. \tag{S-22}$$

By Lemma S-1, to show both $\mathbf{W} \mapsto \nabla_{\mathbf{W}}\ell(\mathbf{v},\mathbf{W})$ and $\nabla f(\mathbf{W})$ are Lipschitz on $\mathcal{C}$, it suffices to show both $\mathbf{h}^*(\mathbf{v},\mathbf{W})$ and $\mathbf{r}^*(\mathbf{v},\mathbf{W})$ are Lipschitz on $\mathcal{C}$, for all $\mathbf{v} \in \mathcal{V}$. Fix arbitrary $\mathbf{v}_1,\mathbf{v}_2 \in \mathcal{V}$ and $\mathbf{W}_1,\mathbf{W}_2 \in \mathcal{C}$. Define

$$\begin{aligned} d(\mathbf{h},\mathbf{r}) &\triangleq \widetilde{\ell}(\mathbf{v}_1,\mathbf{W}_1,\mathbf{h},\mathbf{r}) - \widetilde{\ell}(\mathbf{v}_2,\mathbf{W}_2,\mathbf{h},\mathbf{r}) \\ &= \frac{1}{2}\left\|\mathbf{v}_1 - \mathbf{Y}_1\mathbf{b}(\mathbf{h},\mathbf{r})\right\|_2^2 - \frac{1}{2}\left\|\mathbf{v}_2 - \mathbf{Y}_2\mathbf{b}(\mathbf{h},\mathbf{r})\right\|_2^2 \end{aligned}$$

where $\mathbf{Y}_i = [\mathbf{W}_i \ \ \mathbf{I}]$, $i = 1,2$ and $\mathbf{b}(\mathbf{h},\mathbf{r}) = [\mathbf{h}^T \ \ \mathbf{r}^T]^T$. By Lemma S-2, we have for all $(\mathbf{h}_1,\mathbf{r}_1),(\mathbf{h}_2,\mathbf{r}_2) \in \mathcal{H}\times\mathcal{R}$,

$$|d(\mathbf{h}_1,\mathbf{r}_1) - d(\mathbf{h}_2,\mathbf{r}_2)| \le \left(c_1\left\|\mathbf{v}_1-\mathbf{v}_2\right\|_2 + c_2\left\|\mathbf{Y}_1-\mathbf{Y}_2\right\|_2\right)\left\|\mathbf{b}(\mathbf{h}_1,\mathbf{r}_1) - \mathbf{b}(\mathbf{h}_2,\mathbf{r}_2)\right\|_2 \tag{S-23}$$

where c_1 and c_2 are positive constants. In particular, we have

$$|d(\mathbf{h}_1^*,\mathbf{r}_1^*) - d(\mathbf{h}_2^*,\mathbf{r}_2^*)| \le \left(c_1\left\|\mathbf{v}_1-\mathbf{v}_2\right\|_2 + c_2\left\|\mathbf{Y}_1-\mathbf{Y}_2\right\|_2\right)\left\|\mathbf{b}(\mathbf{h}_1^*,\mathbf{r}_1^*) - \mathbf{b}(\mathbf{h}_2^*,\mathbf{r}_2^*)\right\|_2 \tag{S-24}$$

where $\mathbf{h}_i^* = \mathbf{h}^*(\mathbf{v}_i, \mathbf{W}_i)$ and $\mathbf{r}_i^* = \mathbf{r}^*(\mathbf{v}_i, \mathbf{W}_i)$, $i = 1, 2$. On the other hand, by Assumption 2,

$$\begin{aligned}|d(\mathbf{h}_2^*, \mathbf{r}_2^*) - d(\mathbf{h}_1^*, \mathbf{r}_1^*)| &= \left|\widetilde{\ell}(\mathbf{v}_1, \mathbf{W}_1, \mathbf{h}_2^*, \mathbf{r}_2^*) - \widetilde{\ell}(\mathbf{v}_2, \mathbf{W}_2, \mathbf{h}_2^*, \mathbf{r}_2^*) - \widetilde{\ell}(\mathbf{v}_1, \mathbf{W}_1, \mathbf{h}_1^*, \mathbf{r}_1^*) + \widetilde{\ell}(\mathbf{v}_2, \mathbf{W}_2, \mathbf{h}_1^*, \mathbf{r}_1^*)\right| \\ &= (\widetilde{\ell}(\mathbf{v}_1, \mathbf{W}_1, \mathbf{h}_2^*, \mathbf{r}_2^*) - \widetilde{\ell}(\mathbf{v}_1, \mathbf{W}_1, \mathbf{h}_1^*, \mathbf{r}_1^*)) + (\widetilde{\ell}(\mathbf{v}_2, \mathbf{W}_2, \mathbf{h}_1^*, \mathbf{r}_1^*) - \widetilde{\ell}(\mathbf{v}_2, \mathbf{W}_2, \mathbf{h}_2^*, \mathbf{r}_2^*)) \\ &\geq m_1 \left\|\mathbf{b}(\mathbf{h}_1^*, \mathbf{r}_1^*) - \mathbf{b}(\mathbf{h}_2^*, \mathbf{r}_2^*)\right\|_2^2. \end{aligned} \tag{S-25}$$

Combining (S-24) and (S-25), we have

$$\max\{\|\mathbf{h}_1^* - \mathbf{h}_2^*\|_2, \|\mathbf{r}_1^* - \mathbf{r}_2^*\|_2\} \leq \left\|\mathbf{b}(\mathbf{h}_1^*, \mathbf{r}_1^*) - \mathbf{b}(\mathbf{h}_2^*, \mathbf{r}_2^*)\right\|_2 \leq c_1' \left\|\mathbf{v}_1 - \mathbf{v}_2\right\|_2 + c_2' \left\|\mathbf{W}_1 - \mathbf{W}_2\right\|_2 \tag{S-26}$$

where $c_i' = c_i/m_1$, $i = 1, 2$. This indeed shows both $\mathbf{h}^*(\mathbf{v}, \mathbf{W})$ and $\mathbf{r}^*(\mathbf{v}, \mathbf{W})$ are Lipschitz on $\mathcal{V} \times \mathcal{C}$ since

$$c_1' \left\|\mathbf{v}_1 - \mathbf{v}_2\right\|_2 + c_2' \left\|\mathbf{W}_1 - \mathbf{W}_2\right\|_2 \leq 2\max(c_1', c_2') \left\|[\mathbf{v}_1 \;\; \mathbf{W}_1] - [\mathbf{v}_2 \;\; \mathbf{W}_2]\right\|_F. \tag{S-27}$$

Hence we complete the proof.

S-4. Proof of Lemma 2

By Assumption 3, for all $t \geq 1$, we have

$$\widetilde{f}_t(\mathbf{W}_{t+1}) - \widetilde{f}_t(\mathbf{W}_t) \geq \frac{m_2}{2} \left\|\mathbf{W}_{t+1} - \mathbf{W}_t\right\|_F^2, \tag{S-28}$$

since $\mathbf{W}_t = \arg\min_{\mathbf{W}} \widetilde{f}_t(\mathbf{W})$. On the other hand,

$$\begin{aligned}\widetilde{f}_t(\mathbf{W}_{t+1}) - \widetilde{f}_t(\mathbf{W}_t) &= \widetilde{f}_t(\mathbf{W}_{t+1}) - \widetilde{f}_{t+1}(\mathbf{W}_{t+1}) + \widetilde{f}_{t+1}(\mathbf{W}_{t+1}) - \widetilde{f}_{t+1}(\mathbf{W}_t) + \widetilde{f}_{t+1}(\mathbf{W}_t) - \widetilde{f}_t(\mathbf{W}_t) \\ &\leq \widetilde{d}_t(\mathbf{W}_{t+1}) - \widetilde{d}_t(\mathbf{W}_t)\end{aligned}$$

where $\widetilde{d}_t(\mathbf{W}) \triangleq \widetilde{f}_t(\mathbf{W}) - \widetilde{f}_{t+1}(\mathbf{W})$, for all $\mathbf{W} \in \mathcal{C}$. We aim to show $\widetilde{d}_t$ is Lipschitz on $\mathcal{C}$ with Lipschitz constant only dependent on t. For all $\mathbf{W} \in \mathcal{C}$, we have

$$\begin{aligned}\widetilde{d}_t(\mathbf{W}) &= \frac{1}{t}\sum_{i=1}^{t} \frac{1}{2}\left\|\mathbf{v}_i - \mathbf{W}\mathbf{h}_i - \mathbf{r}_i\right\|_2^2 + \lambda\left\|\mathbf{r}_i\right\|_1 - \frac{1}{t+1}\sum_{i=1}^{t+1} \frac{1}{2}\left\|\mathbf{v}_i - \mathbf{W}\mathbf{h}_i - \mathbf{r}_i\right\|_2^2 + \lambda\left\|\mathbf{r}_i\right\|_1 \\ &= \frac{1}{t(t+1)}\left((t+1)\sum_{i=1}^{t} \frac{1}{2}\left\|\mathbf{v}_i - \mathbf{W}\mathbf{h}_i - \mathbf{r}_i\right\|_2^2 + \lambda\left\|\mathbf{r}_i\right\|_1 - t\sum_{i=1}^{t+1} \frac{1}{2}\left\|\mathbf{v}_i - \mathbf{W}\mathbf{h}_i - \mathbf{r}_i\right\|_2^2 + \lambda\left\|\mathbf{r}_i\right\|_1\right) \\ &\overset{\mathrm{c}}{=} \frac{1}{t(t+1)}\left(\sum_{i=1}^{t} \frac{1}{2}\left\|\mathbf{v}_i - \mathbf{W}\mathbf{h}_i - \mathbf{r}_i\right\|_2^2 - \frac{t}{2}\left\|\mathbf{v}_{t+1} - \mathbf{W}\mathbf{h}_{t+1} - \mathbf{r}_{t+1}\right\|_2^2\right) \\ &= \frac{1}{t(t+1)}\sum_{i=1}^{t}\left(\frac{1}{2}\left\|\mathbf{v}_i - \mathbf{W}\mathbf{h}_i - \mathbf{r}_i\right\|_2^2 - \frac{1}{2}\left\|\mathbf{v}_{t+1} - \mathbf{W}\mathbf{h}_{t+1} - \mathbf{r}_{t+1}\right\|_2^2\right),\end{aligned}$$

where $\overset{\mathrm{c}}{=}$ denotes equality up to an additive constant (independent of $\mathbf{W}$). By a reasoning similar to the one for Lemma S-2, we can show there exist some positive constants c_1 and c_2 such that

$$\begin{aligned}\left|\widetilde{d}_t(\mathbf{W}_{t+1}) - \widetilde{d}_t(\mathbf{W}_t)\right| &\leq \frac{1}{t(t+1)}\sum_{i=1}^{t}\left(c_1\left\|\mathbf{h}_i - \mathbf{h}_{t+1}\right\|_2 + c_2\left\|\mathbf{v}_i - \mathbf{v}_{t+1}\right\|_2 + c_2\left\|\mathbf{r}_i - \mathbf{r}_{t+1}\right\|_2\right)\left\|\mathbf{W}_{t+1} - \mathbf{W}_t\right\|_2 \\ &\leq \frac{c_3}{t+1}\left\|\mathbf{W}_{t+1} - \mathbf{W}_t\right\|_F,\end{aligned} \tag{S-29}$$

where $c_3 > 0$ is a constant since for all $i \geq 1$, $\mathbf{v}_i$, $\mathbf{h}_i$ and $\mathbf{r}_i$ are bounded a.s.. Combining (S-28) and (S-29), we have with probability one,

$$\left\|\mathbf{W}_{t+1} - \mathbf{W}_t\right\|_F \leq \frac{c_3'}{t+1}, \tag{S-30}$$

where $c_3' = c_3/m_2$. Hence we complete the proof.

S-5. Proof of Theorem 1

We prove that $\{\widetilde{f}_t(\mathbf{W}_t)\}_{t\geq 1}$ converges a.s. by the quasi-martingale convergence theorem (see Lemma S-8). Let us first define a filtration $\{\mathcal{F}_t\}_{t\geq 1}$ where $\mathcal{F}_t \triangleq \sigma\left\{\mathbf{v}_i, \mathbf{W}_i, \mathbf{h}_i, \mathbf{r}_i\right\}_{i\in[t]}$ is the minimal σ-algebra such that $\left\{\mathbf{v}_i, \mathbf{W}_i, \mathbf{h}_i, \mathbf{r}_i\right\}_{i\in[t]}$ are measurable. Also define

$$u_t \triangleq \widetilde{f}_t(\mathbf{W}_t) \quad \text{and} \quad \delta_t \triangleq \begin{cases} 1, & \text{if } \mathbb{E}[u_{t+1} - u_t|\mathcal{F}_t] > 0 \\ 0, & \text{otherwise} \end{cases}, \tag{S-31}$$

then it is easy to see $\{u_t\}_{t\geq 1}$ is adapted to $\{\mathcal{F}_t\}_{t\geq 1}$. According to the quasi-martingale convergence theorem, it suffices to show $\sum_{t=1}^{\infty}\mathbb{E}[\delta_t(u_{t+1}-u_t)]<\infty$. To bound $\mathbb{E}[\delta_t(u_{t+1}-u_t)]$, we decompose $u_{t+1}-u_t$ as

$$\begin{aligned}
u_{t+1}-u_t &= \widetilde{f}_{t+1}(\mathbf{W}_{t+1})-\widetilde{f}_{t+1}(\mathbf{W}_t)+\widetilde{f}_{t+1}(\mathbf{W}_t)-\widetilde{f}_t(\mathbf{W}_t)\\
&= \widetilde{f}_{t+1}(\mathbf{W}_{t+1})-\widetilde{f}_{t+1}(\mathbf{W}_t)+\frac{1}{t+1}\widetilde{\ell}(\mathbf{v}_{t+1},\mathbf{h}_{t+1},\mathbf{r}_{t+1},\mathbf{W}_t)+\frac{t}{t+1}\widetilde{f}_t(\mathbf{W}_t)-\widetilde{f}_t(\mathbf{W}_t)\\
&= \widetilde{f}_{t+1}(\mathbf{W}_{t+1})-\widetilde{f}_{t+1}(\mathbf{W}_t)+\frac{\ell(\mathbf{v}_{t+1},\mathbf{W}_t)-\widetilde{f}_t(\mathbf{W}_t)}{t+1} \qquad \text{(S-32)}\\
&= \underbrace{\widetilde{f}_{t+1}(\mathbf{W}_{t+1})-\widetilde{f}_{t+1}(\mathbf{W}_t)}_{\langle 1\rangle}+\frac{\ell(\mathbf{v}_{t+1},\mathbf{W}_t)-f_t(\mathbf{W}_t)}{t+1}+\underbrace{\frac{f_t(\mathbf{W}_t)-\widetilde{f}_t(\mathbf{W}_t)}{t+1}}_{\langle 2\rangle}. \qquad \text{(S-33)}
\end{aligned}$$

By definition, it is easy to see both $\langle 1\rangle, \langle 2\rangle \leq 0$. In (S-33), we insert the term $f_t(\mathbf{W}_t)$ in order to invoke Donsker's theorem (see Lemma S-9). Thus,

$$\mathbb{E}[u_{t+1}-u_t\,|\,\mathcal{F}_t]\leq \frac{\mathbb{E}[\ell(\mathbf{v}_{t+1},\mathbf{W}_t)-f_t(\mathbf{W}_t)\,|\,\mathcal{F}_t]}{t+1}=\frac{f(\mathbf{W}_t)-f_t(\mathbf{W}_t)}{t+1}$$

and using the definition of δ_t,

$$\mathbb{E}[\delta_t\mathbb{E}[u_{t+1}-u_t\,|\,\mathcal{F}_t]]\leq \frac{\mathbb{E}\left[|f(\mathbf{W}_t)-f_t(\mathbf{W}_t)|\right]}{t+1}\leq \frac{\mathbb{E}\left[\|f-f_t\|_{\mathcal{C}}\right]}{t+1}=\frac{\mathbb{E}\left[\left\|\sqrt{t}(f-f_t)\right\|_{\mathcal{C}}\right]}{\sqrt{t}(t+1)}, \qquad \text{(S-34)}$$

where $\|\cdot\|_{\mathcal{C}}$ denotes the uniform norm on $\mathcal{C}$. By Lemma 1, we know for all $\mathbf{v}\in\mathcal{V}$, $\ell(\mathbf{v},\cdot)$ is Lipschitz on $\mathcal{C}$ with Lipschitz constant independent of $\mathbf{v}$. Thus, by Lemma S-11, the measurable function class $\{\ell(\cdot,\mathbf{W}):\mathbf{W}\in\mathcal{C}\}$ is $\mathbb{P}$-Donsker. Consequently $\mathbb{E}\left[\left\|\sqrt{t}(f-f_t)\right\|_{\mathcal{C}}\right]$ is bounded by a constant $c>0$. Thus, $\mathbb{E}[\delta_t\mathbb{E}[u_{t+1}-u_t\,|\,\mathcal{F}_t]]\leq c/t^{3/2}$. Since $\mathbb{E}[\delta_t(u_{t+1}-u_t)]=\mathbb{E}[\mathbb{E}[\delta_t(u_{t+1}-u_t)]\,|\,\mathcal{F}_t]]=\mathbb{E}[\delta_t\mathbb{E}[u_{t+1}-u_t\,|\,\mathcal{F}_t]]$, we have $\sum_{t=1}^{\infty}\mathbb{E}[\delta_t(u_{t+1}-u_t)]<\infty$. Thus $\{\widetilde{f}_t(\mathbf{W}_t)\}_{t\geq 1}$ converges a.s.. Moreover, by Lemma S-8, we also have $\sum_{t=1}^{\infty}\mathbb{E}[|\mathbb{E}[u_{t+1}-u_t|\mathcal{F}_t]|]<\infty$.

Leveraging this result, we proceed to show the almost sure convergence of $\{f(\mathbf{W}_t)\}_{t\geq 1}$. By Lemma S-11, the measurable function class $\{\ell(\cdot,\mathbf{W}):\mathbf{W}\in\mathcal{C}\}$ is also $\mathbb{P}$-Glivenko-Cantelli. Thus by Glivenko-Cantelli theorem (see Lemma S-10), we have $\|f_t-f\|_{\mathcal{C}}\xrightarrow{\text{a.s.}}0$. Hence it suffices to show $\{f_t(\mathbf{W}_t)\}_{t\geq 1}$ converges a.s.. We show this by proving $f_t(\mathbf{W}_t)-\widetilde{f}_t(\mathbf{W}_t)\xrightarrow{\text{a.s.}}0$. First, from (S-33), we have

$$\begin{aligned}
\frac{\widetilde{f}_t(\mathbf{W}_t)-f_t(\mathbf{W}_t)}{t+1} &= \mathbb{E}\left[\frac{\widetilde{f}_t(\mathbf{W}_t)-f_t(\mathbf{W}_t)}{t+1}\,\middle|\,\mathcal{F}_t\right]\\
&= \mathbb{E}[\widetilde{f}_{t+1}(\mathbf{W}_{t+1})-\widetilde{f}_{t+1}(\mathbf{W}_t)\,|\,\mathcal{F}_t]+\mathbb{E}\left[\frac{\ell(\mathbf{v}_{t+1},\mathbf{W}_t)-f_t(\mathbf{W}_t)}{t+1}\,\middle|\,\mathcal{F}_t\right]-\mathbb{E}[u_{t+1}-u_t\,|\,\mathcal{F}_t]\\
&\leq \frac{\mathbb{E}\left[\ell(\mathbf{v}_{t+1},\mathbf{W}_t)\,|\,\mathcal{F}_t\right]-f_t(\mathbf{W}_t)}{t+1}-\mathbb{E}[u_{t+1}-u_t\,|\,\mathcal{F}_t]\\
&\leq \frac{f(\mathbf{W}_t)-f_t(\mathbf{W}_t)}{t+1}-\mathbb{E}[u_{t+1}-u_t\,|\,\mathcal{F}_t]\\
&\leq \frac{\|f-f_t\|_{\mathcal{C}}}{t+1}-\mathbb{E}[u_{t+1}-u_t\,|\,\mathcal{F}_t].
\end{aligned}$$

Since both $\sum_{t=1}^{\infty}\frac{\|f-f_t\|_{\mathcal{C}}}{t+1}$ and $\sum_{t=1}^{\infty}|\mathbb{E}[u_{t+1}-u_t\,|\,\mathcal{F}_t]|$ converge a.s., we conclude $\sum_{t=1}^{\infty}\frac{\widetilde{f}_t(\mathbf{W}_t)-f_t(\mathbf{W}_t)}{t+1}$ converges a.s.. Define $b_t\triangleq\widetilde{f}_t(\mathbf{W}_t)-f_t(\mathbf{W}_t)$, we show $|b_{t+1}-b_t|=O(1/t)$ a.s. by proving both $|\widetilde{f}_{t+1}(\mathbf{W}_{t+1})-\widetilde{f}_t(\mathbf{W}_t)|=O(1/t)$ and $|f_{t+1}(\mathbf{W}_{t+1})-f_t(\mathbf{W}_t)|=O(1/t)$ a.s.. First, from (S-32) and the Lipschitz continuity of $\widetilde{f}_t$ on $\mathcal{C}$,

$$\begin{aligned}
|\widetilde{f}_{t+1}(\mathbf{W}_{t+1})-\widetilde{f}_t(\mathbf{W}_t)| &\leq |\widetilde{f}_{t+1}(\mathbf{W}_{t+1})-\widetilde{f}_{t+1}(\mathbf{W}_t)|+\left|\frac{\ell(\mathbf{v}_{t+1},\mathbf{W}_t)-\widetilde{f}_t(\mathbf{W}_t)}{t+1}\right|\\
&\leq c\,\|\mathbf{W}_{t+1}-\mathbf{W}_t\|_F+\frac{|\ell(\mathbf{v}_{t+1},\mathbf{W}_t)|+|\widetilde{f}_t(\mathbf{W}_t)|}{t+1},
\end{aligned}$$

where $c>0$ is a constant independent of t. Since both $|\ell(\mathbf{v}_{t+1},\mathbf{W}_t)|$ and $|\widetilde{f}_t(\mathbf{W}_t)|$ are bounded on $\mathcal{C}$ a.s. and $\|\mathbf{W}_{t+1}-\mathbf{W}_t\|_F=O(1/t)$ a.s. (by Lemma 2), we have $|\widetilde{f}_{t+1}(\mathbf{W}_{t+1})-\widetilde{f}_t(\mathbf{W}_t)|=O(1/t)$ a.s.. Similarly, by the Lipschitz continuity of f_t on $\mathcal{C}$, we also have $|f_{t+1}(\mathbf{W}_{t+1})-f_t(\mathbf{W}_t)|=O(1/t)$ a.s.. Now we invoke Lemma S-12 to conclude

$$f_t(\mathbf{W}_t)-\widetilde{f}_t(\mathbf{W}_t)\xrightarrow{\text{a.s.}}0. \qquad \text{(S-35)}$$

Since $f_t(\mathbf{W}_t)-f(\mathbf{W}_t)\xrightarrow{\text{a.s.}}0$, both $\{f(\mathbf{W}_t)\}_{t\geq 1}$ and $\{\widetilde{f}_t(\mathbf{W}_t)\}_{t\geq 1}$ converge to the same almost sure limit.

S-6. Proof of Theorem 2

By (S-35), it suffices to show that for every realization of $\{\mathbf{v}_t\}_{t\geq 1}$ such that $\widetilde{f}_t(\mathbf{W}_t)-f_t(\mathbf{W}_t)\to 0$, each subsequential limit of $\{\mathbf{W}_t\}_{t\geq 1}$ is a stationary point of f. We focus on such a realization, then all the variables in the sequel become deterministic. (With a slight abuse of notations we use the same notations to denote the deterministic variables.) By the compactness of $\mathcal{V}$, $\mathcal{H}$ and $\mathcal{R}$, both sequences $\{\mathbf{A}_t\}_{t\geq 1}$ and $\{\mathbf{B}_t\}_{t\geq 1}$ are bounded. Thus there exist compact sets $\mathcal{A}$ and $\mathcal{B}$ such that $\{\mathbf{A}_t\}_{t\geq 1}\subseteq\mathcal{A}$ and $\{\mathbf{B}_t\}_{t\geq 1}\subseteq\mathcal{B}$. Similar reasoning shows that the sequence $\{\widetilde{f}_t(\mathbf{0})\}_{t\geq 1}$ resides in a compact set $\widetilde{\mathcal{F}}$. By the compactness of $\mathcal{C}$, there exists a convergent subsequence $\{\mathbf{W}_{t_m}\}_{m\geq 1}$ in $\{\mathbf{W}_t\}_{t\geq 1}$. Also, by the compactness of $\mathcal{A}$, $\mathcal{B}$ and $\widetilde{\mathcal{F}}$, it is possible to find convergent subsequences $\{\mathbf{A}_{t_k}\}_{k\geq 1}$, $\{\mathbf{B}_{t_k}\}_{k\geq 1}$ and $\{f_{t_k}(\mathbf{0})\}_{k\geq 1}$ such that $\{t_k\}_{k\geq 1}\subseteq\{t_m\}_{m\geq 1}$. Thus, we focus on the convergent sequences $\{\mathbf{W}_{t_k}\}_{k\geq 1}$, $\{\mathbf{A}_{t_k}\}_{k\geq 1}$, $\{\mathbf{B}_{t_k}\}_{k\geq 1}$ and $\{\widetilde{f}_{t_k}(\mathbf{0})\}_{k\geq 1}$ and drop the subscript k to make notations uncluttered. We denote the limits of the sequences $\{\mathbf{W}_t\}_{t\geq 1}$, $\{\mathbf{A}_t\}_{t\geq 1}$ and $\{\mathbf{B}_t\}_{t\geq 1}$ as $\overline{\mathbf{W}}$, $\overline{\mathbf{A}}$ and $\overline{\mathbf{B}}$ respectively.

First, we show the sequence of differentiable functions $\{\widetilde{f}_t\}_{t\geq 1}$ converges uniformly to a differentiable function $\overline{f}$. Since $\{\widetilde{f}_t(\mathbf{0})\}_{t\geq 1}$ converges, it suffices to show the sequence $\{\nabla\widetilde{f}_t\}_{t\geq 1}$ converges uniformly to a function $\overline{h}$. Since $\nabla\widetilde{f}_t(\mathbf{W})=\mathbf{W}\mathbf{A}_t-\mathbf{B}_t$, for any $t,t'\geq 1$ and any $\mathbf{W}\in\mathcal{C}$, we have

$$\begin{aligned}\left\|\nabla\widetilde{f}_t(\mathbf{W})-\nabla\widetilde{f}_{t'}(\mathbf{W})\right\|_F&=\left\|\mathbf{W}\left(\mathbf{A}_t-\mathbf{A}_{t'}\right)-\left(\mathbf{B}_t-\mathbf{B}_{t'}\right)\right\|_F\\&\leq\|\mathbf{W}\|_F\left\|\mathbf{A}_t-\mathbf{A}_{t'}\right\|_F+\left\|\mathbf{B}_t-\mathbf{B}_{t'}\right\|_F\\&\leq\sqrt{K}\left(\left\|\mathbf{A}_t-\mathbf{A}_{t'}\right\|_F+\left\|\mathbf{B}_t-\mathbf{B}_{t'}\right\|_F\right).\end{aligned}\tag{S-36}$$

From (S-36), it is easy to see $\overline{h}(\mathbf{W})=\mathbf{W}\overline{\mathbf{A}}-\overline{\mathbf{B}}$ since $\sup_{\mathbf{W}\in\mathcal{C}}\|\nabla\widetilde{f}_t(\mathbf{W})-\overline{h}(\mathbf{W})\|_F\leq\sqrt{K}\left(\|\mathbf{A}_t-\overline{\mathbf{A}}\|_F+\|\mathbf{B}_t-\overline{\mathbf{B}}\|_F\right)$.

Next, define $g_t\triangleq\widetilde{f}_t-f_t$, for all $t\geq 1$. By definition, we have $g_t(\mathbf{W})\geq 0$, for any $\mathbf{W}\in\mathcal{C}$. Since $\widetilde{f}_t\overset{\text{u}}{\to}\overline{f}$ and $f_t\overset{\text{u}}{\to}f$ (by Glivenko-Cantelli theorem), we have $g_t\overset{\text{u}}{\to}\overline{g}\triangleq\overline{f}-f$. Since both $\overline{f}$ and f are differentiable, $\overline{g}$ is differentiable and $\nabla f=\nabla\overline{f}-\nabla\overline{g}$. To show $\overline{\mathbf{W}}$ is a stationary point of f, it suffices to show for any $\mathbf{W}\in\mathcal{C}$, the directional derivative $\left\langle\nabla f(\overline{\mathbf{W}}),\mathbf{W}-\overline{\mathbf{W}}\right\rangle\geq 0$. We show this by proving $\left\langle\nabla\overline{f}(\overline{\mathbf{W}}),\mathbf{W}-\overline{\mathbf{W}}\right\rangle\geq 0$ and $\left\langle\nabla\overline{g}(\overline{\mathbf{W}}),\mathbf{W}-\overline{\mathbf{W}}\right\rangle=0$ for any $\mathbf{W}\in\mathcal{C}$.

By definition, for any $\mathbf{W}\in\mathcal{C}$ and $t\geq 1$, we have $\widetilde{f}_t(\mathbf{W}_t)\leq\widetilde{f}_t(\mathbf{W})$. First consider

$$\begin{aligned}\left|\widetilde{f}_t(\mathbf{W}_t)-\overline{f}(\overline{\mathbf{W}})\right|&=\left|\widetilde{f}_t(\mathbf{W}_t)-\overline{f}(\mathbf{W}_t)+\overline{f}(\mathbf{W}_t)-\overline{f}(\overline{\mathbf{W}})\right|\\&\leq\left\|\widetilde{f}_t-\overline{f}\right\|_{\mathcal{C}}+\left|\overline{f}(\mathbf{W}_t)-\overline{f}(\overline{\mathbf{W}})\right|.\end{aligned}$$

Since $\widetilde{f}_t\overset{\text{u}}{\to}\overline{f}$ and $\overline{f}$ is continuous, we have $\widetilde{f}_t(\mathbf{W}_t)\to\overline{f}(\overline{\mathbf{W}})$ as $t\to\infty$. Thus $\overline{f}(\overline{\mathbf{W}})\leq\overline{f}(\mathbf{W})$, for any $\mathbf{W}\in\mathcal{C}$. This implies $\left\langle\nabla\overline{f}(\overline{\mathbf{W}}),\mathbf{W}-\overline{\mathbf{W}}\right\rangle\geq 0$.

Next we show $\left\langle\nabla\overline{g}(\overline{\mathbf{W}}),\mathbf{W}-\overline{\mathbf{W}}\right\rangle=0$ for any $\mathbf{W}\in\mathcal{C}$. It suffices to show $\nabla\overline{g}(\overline{\mathbf{W}})=\mathbf{0}$. Since both $\widetilde{f}_t$ and f_t are differentiable, g_t is differentiable and $\nabla g_t=\nabla\widetilde{f}_t-\nabla f$. First it is easy to see ∇g_t is Lipschitz on $\mathcal{C}$ with constant $L>0$ independent of t since both $\nabla\widetilde{f}_t$ and ∇f_t are Lipschitz on $\mathcal{C}$ with constants independent of t. It is possible to construct another differentiable function $\widetilde{g}_t$ with domain $\mathbb{R}^{F\times K}$ such that $\widetilde{g}_t$ is nonnegative with a L-Lipschitz gradient on $\mathbb{R}^{F\times K}$ and $\widetilde{g}_t(\mathbf{W})=g_t(\mathbf{W})$ for all $\mathbf{W}\in\mathcal{C}$. Thus

$$\frac{1}{2L}\|\nabla g_t(\mathbf{W}_t)\|_F^2=\frac{1}{2L}\|\nabla\widetilde{g}_t(\mathbf{W}_t)\|_F^2\leq\widetilde{g}_t(\mathbf{W}_t)-\widetilde{g}_t^*\leq g_t(\mathbf{W}_t)=\widetilde{f}_t(\mathbf{W}_t)-f_t(\mathbf{W}_t),$$

where $\widetilde{g}_t^*=\inf_{\mathbf{W}\in\mathbb{R}^{F\times K}}\widetilde{g}_t(\mathbf{W})\geq 0$. Since $\widetilde{f}_t(\mathbf{W}_t)-f_t(\mathbf{W}_t)\to 0$, we have $\nabla g_t(\mathbf{W}_t)\to\mathbf{0}$ as $t\to\infty$. Now consider the first-order Taylor expansion of g_t at $\mathbf{W}_t$

$$g_t(\mathbf{W})=g_t(\mathbf{W}_t)+\langle\nabla g_t(\mathbf{W}_t),\mathbf{W}-\mathbf{W}_t\rangle+o\left(\|\mathbf{W}-\mathbf{W}_t\|_F\right),\ \ \forall\,\mathbf{W}\in\mathcal{C}.\tag{S-37}$$

As $t\to\infty$, we have

$$\overline{g}(\mathbf{W})=\overline{g}(\overline{\mathbf{W}})+o\left(\left\|\mathbf{W}-\overline{\mathbf{W}}\right\|_F\right),\ \ \forall\,\mathbf{W}\in\mathcal{C}.\tag{S-38}$$

On the other hand, we have

$$\overline{g}(\mathbf{W})=\overline{g}(\overline{\mathbf{W}})+\left\langle\nabla\overline{g}(\overline{\mathbf{W}}),\mathbf{W}-\overline{\mathbf{W}}\right\rangle+o\left(\left\|\mathbf{W}-\overline{\mathbf{W}}\right\|_F\right),\ \ \forall\,\mathbf{W}\in\mathcal{C}.\tag{S-39}$$

Comparing (S-38) and (S-39), we have

$$\left\langle\nabla\overline{g}(\overline{\mathbf{W}}),\mathbf{W}-\overline{\mathbf{W}}\right\rangle+o\left(\left\|\mathbf{W}-\overline{\mathbf{W}}\right\|_F\right)=0,\ \ \forall\,\mathbf{W}\in\mathcal{C}.\tag{S-40}$$

Therefore we conclude $\nabla\overline{g}(\overline{\mathbf{W}})=\mathbf{0}$.

S-7. Proof of Lemma S-1

Since $\mathbf{v}$, $\mathbf{W}$, $\mathbf{h}^*(\mathbf{v},\mathbf{W})$ and $\mathbf{r}^*(\mathbf{v},\mathbf{W})$ are all bounded, there exist positive constants M_1, M_2, M_3 and M_4 that upper bound $\|\mathbf{v}\|$, $\|\mathbf{W}\|$, $\|\mathbf{h}^*(\mathbf{v},\mathbf{W})\|$ and $\|\mathbf{r}^*(\mathbf{v},\mathbf{W})\|$ respectively. Here the matrix norm is the one induced by the (general) vector norm.

Take arbitrary $\mathbf{W}_1$ and $\mathbf{W}_2$ in $\mathcal{C}$, we have

$$\begin{aligned}\|\nabla f(\mathbf{W}_1) - \nabla f(\mathbf{W}_2)\| &= \|\mathbb{E}_{\mathbf{v}}[\nabla_{\mathbf{W}}\ell(\mathbf{v},\mathbf{W}_1) - \nabla_{\mathbf{W}}\ell(\mathbf{v},\mathbf{W}_2)]\| \\ &\le \mathbb{E}_{\mathbf{v}}\|\nabla_{\mathbf{W}}\ell(\mathbf{v},\mathbf{W}_1) - \nabla_{\mathbf{W}}\ell(\mathbf{v},\mathbf{W}_2)\|.\end{aligned}$$

Fix an arbitrary $\mathbf{v}\in\mathcal{V}$ we have

$$\begin{aligned}\|\nabla_{\mathbf{W}}\ell(\mathbf{v},\mathbf{W}_1) - \nabla_{\mathbf{W}}\ell(\mathbf{v},\mathbf{W}_2)\| \le{}& \left\|\mathbf{W}_1\mathbf{h}^*(\mathbf{v},\mathbf{W}_1)\mathbf{h}^*(\mathbf{v},\mathbf{W}_1)^T - \mathbf{W}_2\mathbf{h}^*(\mathbf{v},\mathbf{W}_2)\mathbf{h}^*(\mathbf{v},\mathbf{W}_2)^T\right\| \\ &+ \left\|\mathbf{r}^*(\mathbf{v},\mathbf{W}_1)\mathbf{h}^*(\mathbf{v},\mathbf{W}_1)^T - \mathbf{r}^*(\mathbf{v},\mathbf{W}_2)\mathbf{h}^*(\mathbf{v},\mathbf{W}_2)^T\right\| \\ &+ \left\|\mathbf{v}\mathbf{h}^*(\mathbf{v},\mathbf{W}_1)^T - \mathbf{v}\mathbf{h}^*(\mathbf{v},\mathbf{W}_2)^T\right\|.\end{aligned} \tag{S-41}$$

We bound each term on the RHS of (S-41) as follows

$$\begin{aligned}&\left\|\mathbf{W}_1\mathbf{h}^*(\mathbf{v},\mathbf{W}_1)\mathbf{h}^*(\mathbf{v},\mathbf{W}_1)^T - \mathbf{W}_2\mathbf{h}^*(\mathbf{v},\mathbf{W}_2)\mathbf{h}^*(\mathbf{v},\mathbf{W}_2)^T\right\| \\ \le{}& \|\mathbf{W}_1\|\,\|\mathbf{h}^*(\mathbf{v},\mathbf{W}_1)\|\,\|\mathbf{h}^*(\mathbf{v},\mathbf{W}_1) - \mathbf{h}^*(\mathbf{v},\mathbf{W}_2)\| + \|\mathbf{W}_1\mathbf{h}^*(\mathbf{v},\mathbf{W}_1) - \mathbf{W}_2\mathbf{h}^*(\mathbf{v},\mathbf{W}_2)\|\,\|\mathbf{h}^*(\mathbf{v},\mathbf{W}_2)\| \\ \le{}& M_2M_3c_1\|\mathbf{W}_1-\mathbf{W}_2\| + (\|\mathbf{W}_1\|\,\|\mathbf{h}^*(\mathbf{v},\mathbf{W}_1) - \mathbf{h}^*(\mathbf{v},\mathbf{W}_2)\| + \|\mathbf{W}_1-\mathbf{W}_2\|\,\|\mathbf{h}^*(\mathbf{v},\mathbf{W}_2)\|)\,\|\mathbf{h}^*(\mathbf{v},\mathbf{W}_2)\| \\ \le{}& M_2M_3c_1\|\mathbf{W}_1-\mathbf{W}_2\| + M_3\left(M_2c_1\|\mathbf{W}_1-\mathbf{W}_2\| + M_3\|\mathbf{W}_1-\mathbf{W}_2\|\right) \\ ={}& (2c_1M_2M_3 + M_3^2)\|\mathbf{W}_1-\mathbf{W}_2\|,\end{aligned}$$

$$\begin{aligned}&\left\|\mathbf{r}^*(\mathbf{v},\mathbf{W}_1)\mathbf{h}^*(\mathbf{v},\mathbf{W}_1)^T - \mathbf{r}^*(\mathbf{v},\mathbf{W}_2)\mathbf{h}^*(\mathbf{v},\mathbf{W}_2)^T\right\| \\ \le{}& \|\mathbf{r}^*(\mathbf{v},\mathbf{W}_1)\|\,\|\mathbf{h}^*(\mathbf{v},\mathbf{W}_1) - \mathbf{h}^*(\mathbf{v},\mathbf{W}_2)\| + \|\mathbf{r}^*(\mathbf{v},\mathbf{W}_1) - \mathbf{r}^*(\mathbf{v},\mathbf{W}_2)\|\,\|\mathbf{h}^*(\mathbf{v},\mathbf{W}_2)\| \\ \le{}& c_1M_4\|\mathbf{W}_1-\mathbf{W}_2\| + c_2M_3\|\mathbf{W}_1-\mathbf{W}_2\| \\ \le{}& (c_1M_4 + c_2M_3)\|\mathbf{W}_1-\mathbf{W}_2\|,\end{aligned}$$

and

$$\left\|\mathbf{v}\mathbf{h}^*(\mathbf{v},\mathbf{W}_1)^T - \mathbf{v}\mathbf{h}^*(\mathbf{v},\mathbf{W}_2)^T\right\| \le c_1M_1\|\mathbf{W}_1-\mathbf{W}_2\|.$$

Thus, take $c_3 = 2c_1M_2M_3 + M_3^2 + c_1M_4 + c_2M_3 + c_1M_1$ and we finish the proof.

S-8. Proof of Lemma S-2

It suffices to show $\|\nabla g(\mathbf{b})\|_2 \le c_1\|\mathbf{z}-\mathbf{z}'\|_2 + c_2\|\mathbf{A}-\mathbf{A}'\|_2$ for any $\mathbf{b}\in\mathcal{B}$ and some positive constants c_1 and c_2 (independent of $\mathbf{b}$). We write $\|\nabla g(\mathbf{b})\|_2$ as

$$\begin{aligned}\|\nabla g(\mathbf{b})\|_2 &= \|\mathbf{A}'^T(\mathbf{A}'\mathbf{b}-\mathbf{z}') - \mathbf{A}^T(\mathbf{A}\mathbf{b}-\mathbf{z})\|_2 \\ &= \|(\mathbf{A}'^T\mathbf{A}' - \mathbf{A}^T\mathbf{A})\mathbf{b} - (\mathbf{A}'^T\mathbf{z}' - \mathbf{A}^T\mathbf{z})\|_2 \\ &\le \underbrace{\|\mathbf{A}'^T\mathbf{A}' - \mathbf{A}^T\mathbf{A}\|_2\|\mathbf{b}\|_2}_{\langle 1\rangle} + \underbrace{\|\mathbf{A}'^T\mathbf{z}' - \mathbf{A}^T\mathbf{z}\|_2}_{\langle 2\rangle}.\end{aligned}$$

By the compactness of $\mathcal{Z}$, $\mathcal{A}$ and $\mathcal{B}$, there exist positive constants M_1, M_2 and M_3 such that $\|\mathbf{z}\|_2\le M_1$, $\|\mathbf{A}\|_2\le M_2$ and $\|\mathbf{b}\|_2\le M_3$, for any $\mathbf{z}\in\mathcal{Z}$, $\mathbf{A}\in\mathcal{A}$ and $\mathbf{b}\in\mathcal{B}$. Thus,

$$\begin{aligned}\langle 1\rangle &\le M_3\|\mathbf{A}^T\mathbf{A} - \mathbf{A}^T\mathbf{A}' + \mathbf{A}^T\mathbf{A}' - \mathbf{A}'^T\mathbf{A}'\|_2 \\ &\le M_3\left(\|\mathbf{A}^T(\mathbf{A}-\mathbf{A}')\|_2 + \left\|(\mathbf{A}-\mathbf{A}')^T\mathbf{A}'\right\|_2\right) \\ &\le 2M_2M_3\|\mathbf{A}-\mathbf{A}'\|_2.\end{aligned}$$

Similarly for $\langle 2\rangle$ we have

$$\begin{aligned}\langle 2\rangle &= \left\|\mathbf{A}^T\mathbf{z} - \mathbf{A}^T\mathbf{z}' + \mathbf{A}^T\mathbf{z}' - \mathbf{A}'^T\mathbf{z}'\right\|_2 \\ &\le \left\|\mathbf{A}^T(\mathbf{z}-\mathbf{z}')\right\|_2 + \left\|(\mathbf{A}-\mathbf{A}')^T\mathbf{z}'\right\|_2 \\ &\le M_2\|\mathbf{z}-\mathbf{z}'\|_2 + M_1\|\mathbf{A}-\mathbf{A}'\|_2.\end{aligned}$$

Hence $\langle 1\rangle + \langle 2\rangle \le M_2\|\mathbf{z}-\mathbf{z}'\|_2 + (M_1 + 2M_2M_3)\|\mathbf{A}-\mathbf{A}'\|_2$. We now take $c_1 = M_2$ and $c_2 = M_1 + 2M_2M_3$ to complete the proof.

S-9. TECHNICAL LEMMAS

Lemma S-3 ([1, Lemma 5]). *Let I be a closed interval in $\mathbb{R}$. Define $g_{\tau,I}(t) = \tau\,|t| + \delta_I(t)$, where δ_I is the indicator function for the interval I. Then the proximal operator for $g_{\tau,I}$ is given by*

$$\mathbf{prox}_{g_{\tau,I}}(q) = \Pi_I(\mathcal{S}_\tau(q)), \tag{S-42}$$

where $q \in \mathbb{R}$, $\mathcal{S}_\tau$ is the soft-thresholding operator with threshold τ and Π_I is the Euclidean projector onto the interval I.

Lemma S-4 (Projection onto nonnegative ℓ_2 balls). *Let $\mathcal{C}' \triangleq \{\mathbf{x} \in \mathbb{R}^n_+ \mid \|\mathbf{x}\| \le 1\}$. Then for all $\mathbf{y} \in \mathbb{R}^n$,*

$$\Pi_{\mathcal{C}'}(\mathbf{y}) = \frac{\mathbf{y}_+}{\max\{1, \|\mathbf{y}_+\|_2\}}, \tag{S-43}$$

where $(\mathbf{y}_+)_i = \max\{0, y_i\},\ \forall\, i \in [n]$.

Proof. The KKT conditions for

$$\begin{aligned} \min \quad & \|\mathbf{y} - \mathbf{x}\|_2^2 \\ \text{s.t.} \quad & \mathbf{x} \ge 0,\ \|\mathbf{x}\|_2 \le 1 \end{aligned}$$

are given by

$$\mathbf{x}^* \ge 0,\ \|\mathbf{x}^*\|_2 \le 1,\ \lambda^* \ge 0 \tag{S-44}$$

$$(\lambda^* + 1)\mathbf{x}^* \ge \mathbf{y}, \tag{S-45}$$

$$\lambda^*(\|\mathbf{x}^*\|_2^2 - 1) = 0, \tag{S-46}$$

$$(\lambda^* + 1)(x_i^*)^2 = x_i^* y_i,\ \forall\, i \in [n]. \tag{S-47}$$

We fix an $i \in [n]$. Define $\mathcal{I} \triangleq \{i \in [n] \,|\, y_i > 0\}$. For any $\mathbf{z} \in \mathbb{R}^n$, let $\mathbf{z}_\mathcal{I}$ to be the subvector of $\mathbf{z}$ with indices from $\mathcal{I}$. If $y_i \le 0$, then by (S-47) $x_i^* = 0$. If $y_i > 0$, then by (S-45) $x_i^* > 0$. Thus by (S-47) we have $x_i^* = y_i/(\lambda^*+1)$. If $\lambda^* = 0$, then $x_i^* = y_i$. In this case $\|\mathbf{y}_\mathcal{I}\|_2 = \|\mathbf{x}_\mathcal{I}^*\|_2 = \|\mathbf{x}^*\|_2 \le 1$. If $\lambda^* > 0$, then by (S-46) $\|\mathbf{x}^*\|_2^2 = 1$. Then $\|\mathbf{x}^*\|_2^2 = \|\mathbf{x}_\mathcal{I}^*\|_2^2 = \|\mathbf{y}_\mathcal{I}\|_2^2/(\lambda^*+1)^2 = 1$. This means $\lambda^* + 1 = \|\mathbf{y}_\mathcal{I}\|_2$ so $x_i^* = y_i/\|\mathbf{y}_\mathcal{I}\|_2$. Also, we notice in such case $y_i > x_i^* > 0$ so $\|\mathbf{y}_\mathcal{I}\|_2 > 1$. Combining both cases where $\lambda^* = 0$ and $\lambda^* > 0$, we have $x_i^* = y_i/\max\{1, \|\mathbf{y}_\mathcal{I}\|_2\}$, for all $i \in \mathcal{I}$. □

Lemma S-5 (Danskin's Theorem; [2, Theorem 4.1]). *Let $\mathcal{X}$ be a metric space and $\mathcal{U}$ be a normed vector space. Let $f : \mathcal{X} \times \mathcal{U} \to \mathbb{R}$ have the following properties*

1) *$f(x,\cdot)$ is differentiable on $\mathcal{U}$, for any $x \in \mathcal{X}$.*
2) *$f(x,u)$ and $\nabla_u f(x,u)$ are continuous on $\mathcal{X} \times \mathcal{U}$.*

Let Φ be a compact set in $\mathcal{X}$. Define $v(u) = \inf_{x\in\Phi} f(x,u)$ and $S(u) = \arg\min_{x\in\Phi} f(x,u)$, then $v(u)$ is directionally differentiable and its directional derivative along $d \in \mathcal{U}$, $v'(u,d)$ is given by

$$v'(u,d) = \min_{x \in S(u)} \nabla_u f(x,u)^T d. \tag{S-48}$$

In particular, if for some $u_0 \in \mathcal{U}$, $S(u_0) = \{x_0\}$, then v is differentiable at $u = u_0$ and $\nabla v(u_0) = \nabla_u f(x_0, u_0)$.

Lemma S-6 (The Maximum Theorem; [3, Theorem 14.2.1 & Example 2]). *Let $\mathcal{P}$ and $\mathcal{X}$ be two metric spaces. Consider a maximization problem*

$$\max_{x \in B(p)} f(p,x), \tag{S-49}$$

where $B : \mathcal{P} \twoheadrightarrow \mathcal{X}$ is a correspondence and $f : \mathcal{P} \times \mathcal{X} \to \mathbb{R}$ is a function. If B is compact-valued and continuous on $\mathcal{P}$ and f is continuous on $\mathcal{P} \times \mathcal{X}$, then the correspondence $S(p) = \arg\max_{x\in B(p)} f(p,x)$ is compact-valued and upper hemicontinuous, for any $p \in \mathcal{P}$. In particular, if for some $p_0 \in \mathcal{P}$, $S(p_0) = \{s(p_0)\}$, where $s : \mathcal{P} \to \mathcal{X}$ is a function, then s is continuous at $p = p_0$. Moreover, we have the same conclusions if the maximization in (S-49) *is replaced by minimization.*

Lemma S-7 (Leibniz Integral Rule). *Let $\mathcal{X}$ be an open set in $\mathbb{R}^n$ and let $(\Omega, \mathcal{A}, \mu)$ be a measure space. If $f : \mathcal{X} \times \Omega \to \mathbb{R}$ satisfies*

1) *For all $x \in \mathcal{X}$, the mapping $\omega \mapsto f(x,\omega)$ is Lebesgue integrable.*
2) *For all $\omega \in \Omega$, $\nabla_x f(x,\omega)$ exists on $\mathcal{X}$.*
3) *For all $x \in \mathcal{X}$, the mapping $\omega \mapsto \nabla_x f(x,\omega)$ is Lebesgue integrable.*

Then $\int_\Omega f(x,\omega)\,d\mu(\omega)$ is differentiable and

$$\nabla_x \int_\Omega f(x,\omega)\,d\mu(\omega) = \int_\Omega \nabla_x f(x,\omega)\,d\mu(\omega). \tag{S-50}$$

Remark 1. This is a simplified version of the Leibniz Integral Rule. See [4, Theorem 16.8] for weaker conditions on f.

Definition S-1 (Quasi-martingale; [5, Definition 1.4]). Let $\{X_t\}_{t\in\mathcal{T}}$ be a stochastic process and let $\{\mathcal{F}_t\}_{t\in\mathcal{T}}$ be the filtration to which $\{X_t\}_{t\in\mathcal{T}}$ is adapted, where $\mathcal{T}$ is a subset of the real line. We call $\{X_t\}_{t\in\mathcal{T}}$ a quasi-martingale if there exist two stochastic processes $\{Y_t\}_{t\in\mathcal{T}}$ and $\{Z_t\}_{t\in\mathcal{T}}$ such that

i) both $\{Y_t\}_{t\in\mathcal{T}}$ and $\{Z_t\}_{t\in\mathcal{T}}$ are adapted to $\{\mathcal{F}_t\}_{t\in\mathcal{T}}$,

ii) $\{Y_t\}_{t\in\mathcal{T}}$ is a martingale and $\{Z_t\}_{t\in\mathcal{T}}$ has bounded variations on $\mathcal{T}$ a.s.,

iii) $X_t = Y_t + Z_t$, for all $t \in \mathcal{T}$ a.s..

Lemma S-8 (The Quasi-martingale Convergence Theorem; [6, Theorem 9.4 & Proposition 9.5]). *Let $(u_t)_{t\geq 1}$ be a nonnegative discrete-time stochastic process on a probability space $(\Omega, \mathcal{A}, \mathbb{P})$, i.e., $u_t \geq 0$ a.s., for all $t \geq 1$. Let $\{\mathcal{F}_t\}_{t\geq 1}$ be a filtration to which $(u_t)_{t\geq 1}$ is adapted. Define another binary stochastic process $(\delta_t)_{t\geq 1}$ as*

$$\delta_t = \begin{cases} 1, & \text{if } \mathbb{E}[u_{t+1} - u_t|\mathcal{F}_t] > 0 \\ 0, & \text{otherwise} \end{cases}. \tag{S-51}$$

If $\sum_{t=1}^{\infty} \mathbb{E}[\delta_t(u_{t+1} - u_t)] < \infty$, then $u_t \xrightarrow{\text{a.s.}} u$, where u is integrable on $(\Omega, \mathcal{F}, \mathbb{P})$ and nonnegative a.s.. Furthermore, $(u_t)_{t\geq 1}$ is a quasi-martingale and

$$\sum_{t=1}^{\infty} \mathbb{E}[|\mathbb{E}[u_{t+1} - u_t|\mathcal{F}_t]|] < \infty. \tag{S-52}$$

Lemma S-9 (Donsker's Theorem; [7, Section 19.2]). *Let $X_1, \ldots, X_n$ be i.i.d. generated from a distribution $\mathbb{P}$. Define the empirical distribution $\mathbb{P}_n \triangleq \frac{1}{n}\sum_{i=1}^{n} \delta_{X_i}$. For a measurable function f, define $\mathbb{P}_n f$ and $\mathbb{P}f$ as the expectations of f under the distributions $\mathbb{P}_n$ and $\mathbb{P}$ respectively. Define an empirical process $G_n(f) \triangleq \sqrt{n}(\mathbb{P}_n f - \mathbb{P}f)$, $f \in \mathcal{F}$, where $\mathcal{F}$ is a class of measurable functions. $\mathcal{F}$ is $\mathbb{P}$-Donsker if and only if the sequence of empirical processes $\{G_n\}_{n\geq 1}$ converges in distribution to a zero-mean Gaussian process G tight in $\ell^\infty(\mathcal{F})$, where $\ell^\infty(\mathcal{F})$ is the space of all real-valued and bounded functionals defined on $\mathcal{F}$ equipped with the uniform norm on $\mathcal{F}$, denoted as $\|\cdot\|_{\mathcal{F}}$. Moreover, in such case, we have $\mathbb{E}\|G_n\|_{\mathcal{F}} \to \mathbb{E}\|G\|_{\mathcal{F}}$.*

Lemma S-10 (Glivenko-Cantelli theorem; [7, Section 19.2]). *Let $\mathbb{P}_n$ and $\mathbb{P}$ be the distributions defined as in Lemma S-9. A class of measurable functions $\mathcal{F}$ is $\mathbb{P}$-Glivenko-Cantelli if and only if $\sup_{f\in\mathcal{F}} |\mathbb{P}_n f - \mathbb{P}f| \xrightarrow{\text{a.s.}} 0$.*

Lemma S-11 (A sufficient condition for $\mathbb{P}$-Glivenko-Cantelli and $\mathbb{P}$-Donsker classes; [7, Example 19.7]). *Define a probability space $(\mathcal{X}, \mathcal{A}, \mathbb{P})$. Let $\mathcal{F} = \{f_\theta : \mathcal{X} \to \mathbb{R} \,|\, \theta \in \Theta\}$ be a class of measurable functions, where Θ is a bounded subset in $\mathbb{R}^d$. If there exists a universal constant $K > 0$ such that*

$$|f_{\theta_1}(x) - f_{\theta_2}(x)| \leq K\|\theta_1 - \theta_2\|, \ \forall\, \theta_1, \theta_2 \in \Theta, \ \forall\, x \in \mathcal{X}, \tag{S-53}$$

where $\|\cdot\|$ is a general vector norm in $\mathbb{R}^d$, then $\mathcal{F}$ is both $\mathbb{P}$-Glivenko-Cantelli and $\mathbb{P}$-Donsker.

Lemma S-12 ([8, Lemma 8]). *Let (a_n), (b_n) be two nonnegative sequences. Suppose $\sum_{n=1}^{\infty} a_n = \infty$ and $\sum_{n=1}^{\infty} a_n b_n < \infty$, and $\exists\, N \in \mathbb{N}$ and $K > 0$ such that for all $n \geq N$, $|b_{n+1} - b_n| \leq K a_n$. Then (b_n) converges and $\lim_{n\to\infty} b_n = 0$.*

S-10. Additional experiment results

This section consists of two parts. In the first part, we show the convergence speeds of all the online and batch algorithms on the (contaminated) CBCL face dataset for different values of the mini-batch size τ, the latent dimension K, the penalty parameter ρ in the ADMM-based algorithms, the step-size parameter κ in the PGD-based algorithms and the (salt and pepper) noise density parameters ν and $\widetilde{\nu}$ in Figure 1 to 6. As mentioned in Section VII-C, all the convergence results on the CBCL face dataset agree with those on the synthetic dataset. In the second part, we show the quality of the denoised images of all the algorithms on the CBCL face dataset for $K = 25$ and $K = 100$ in Table I (a) and (b) respectively. The corresponding running times of all the algorithms for $K = 25$ and $K = 100$ are shown in Table II (a) and (b) respectively. As stated in Section VII-E, the results when $K = 25$ and $K = 100$ are similar to those when $K = 49$.

References

[1] H. Zhang and L. Cheng, "Projected shrinkage algorithm for box-constrained ℓ_1-minimization," *Optim. Lett.*, 2015.

[2] J. F. Bonnans and A. Shapiro, "Optimization problems with perturbations: A guided tour," *SIAM Review*, 1998.

[3] K. Sydsaeter, P. Hammond, A. Seierstad, and A. Strom, *Further Mathematics for Economic Analysis*. 2005.

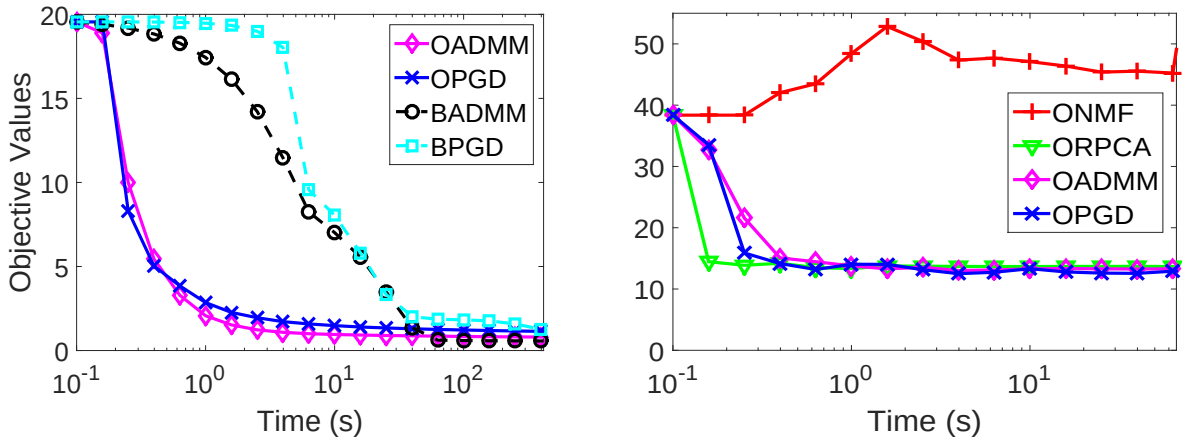
[4] P. Billingsley, *Probability and Measure*. John Wiley & Sons, 2nd ed., 1986.

[5] D. L. Fisk, "Quasi-martingales," *Trans. Amer. Math. Soc.*, 1965.

[6] M. Métivier, *Semimartingales: A Course on Stochastic Processes*. 1982.

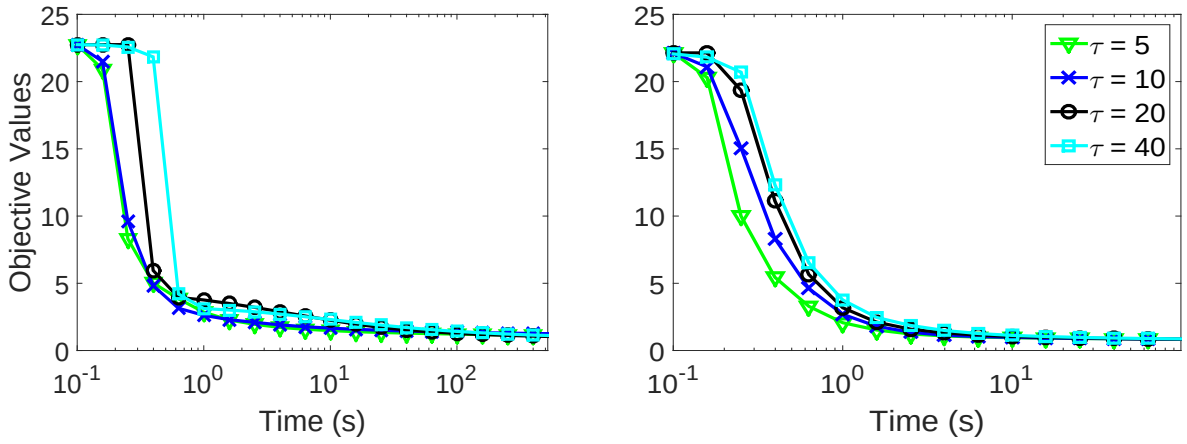
[7] A. W. van der Vaart, *Asymptotic Statistics*. Cambridge Press, 2000.

[8] J. Mairal, F. Bach, J. Ponce, and G. Sapiro, "Online learning for matrix factorization and sparse coding," *J. Mach. Learn. Res.*, 2010.



(a) (b)

Fig. 1. The objective values (as a function of time) of (a) our online algorithms and their batch counterparts (b) our online algorithms and other online algorithms on the CBCL face dataset. The parameters are set according to the canonical setting.



(a) OPGD (b) OADMM

Fig. 2. The objective values (as a function of time) of (a) OPGD and (b) OADMM for different values of τ on the CBCL face dataset. All the other parameters are set according to the canonical setting.

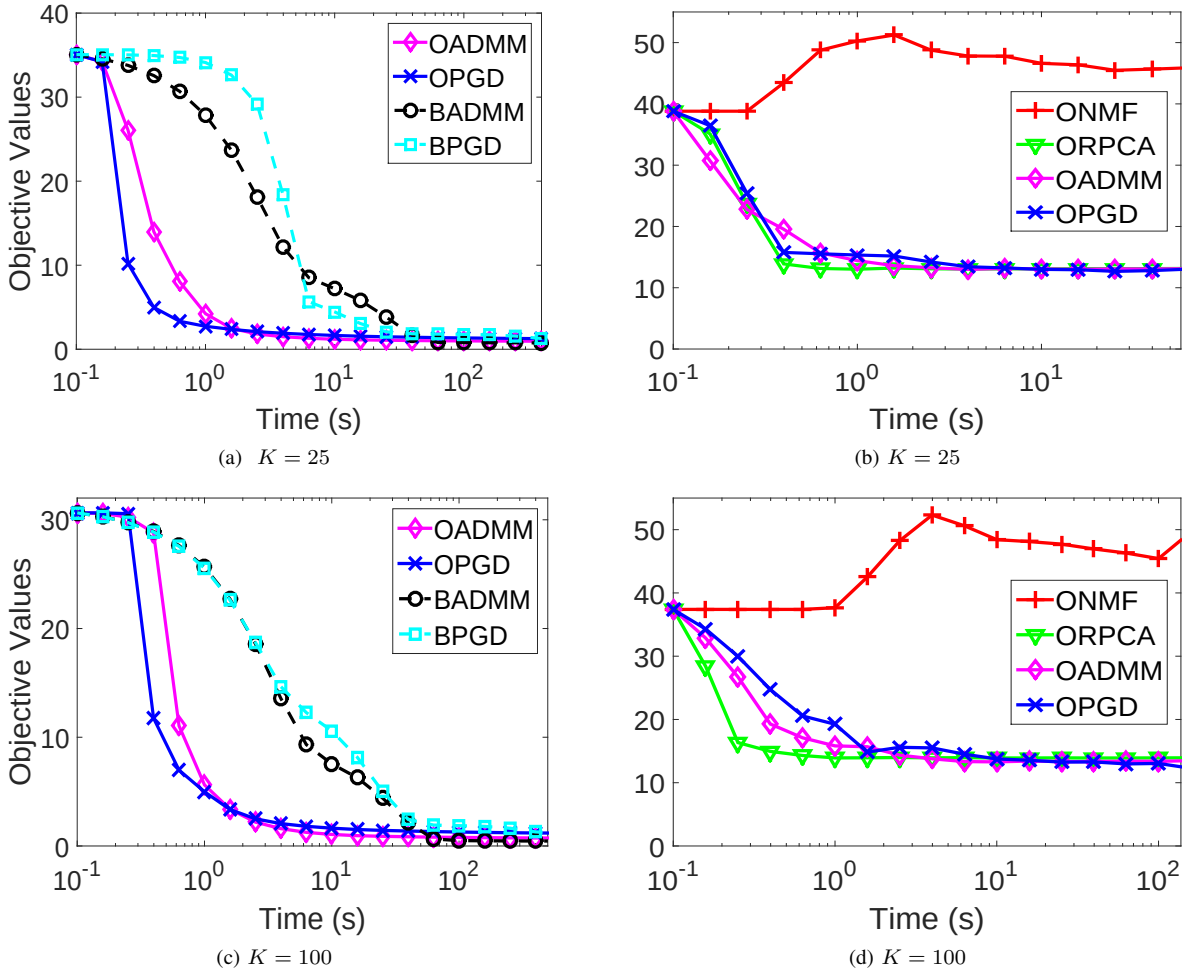
TABLE I
PSNRS (IN DB) OF ALL THE ALGORITHMS ON THE CBCL FACE DATASET WITH DIFFERENT NOISE DENSITY.

	Setting 1	Setting 2	Setting 3
OADMM	11.39 ± 0.16	11.37 ± 0.12	11.35 ± 0.18
OPGD	11.49 ± 0.05	11.43 ± 0.09	11.38 ± 0.06
BADMM	11.51 ± 0.19	11.46 ± 0.07	11.41 ± 0.15
BPGD	11.54 ± 0.07	11.46 ± 0.17	11.42 ± 0.15
ONMF	5.99 ± 0.04	5.97 ± 0.12	5.97 ± 0.08
ORPCA	11.26 ± 0.05	11.24 ± 0.11	11.22 ± 0.11

(a) $K = 25$

	Setting 1	Setting 2	Setting 3
OADMM	11.39 ± 0.02	11.35 ± 0.11	11.35 ± 0.06
OPGD	11.51 ± 0.01	11.45 ± 0.09	11.44 ± 0.11
BADMM	11.51 ± 0.11	11.47 ± 0.00	11.45 ± 0.03
BPGD	11.52 ± 0.16	11.46 ± 0.07	11.45 ± 0.12
ONMF	5.99 ± 0.19	5.97 ± 0.03	5.95 ± 0.05
ORPCA	11.26 ± 0.03	11.24 ± 0.16	11.20 ± 0.13

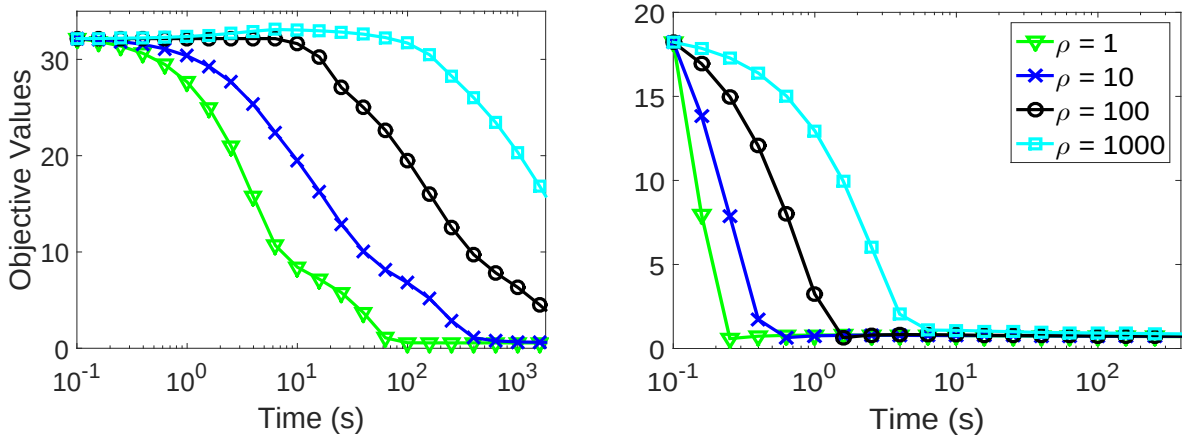
(b) $K = 100$



(a) $K = 25$ (b) $K = 25$

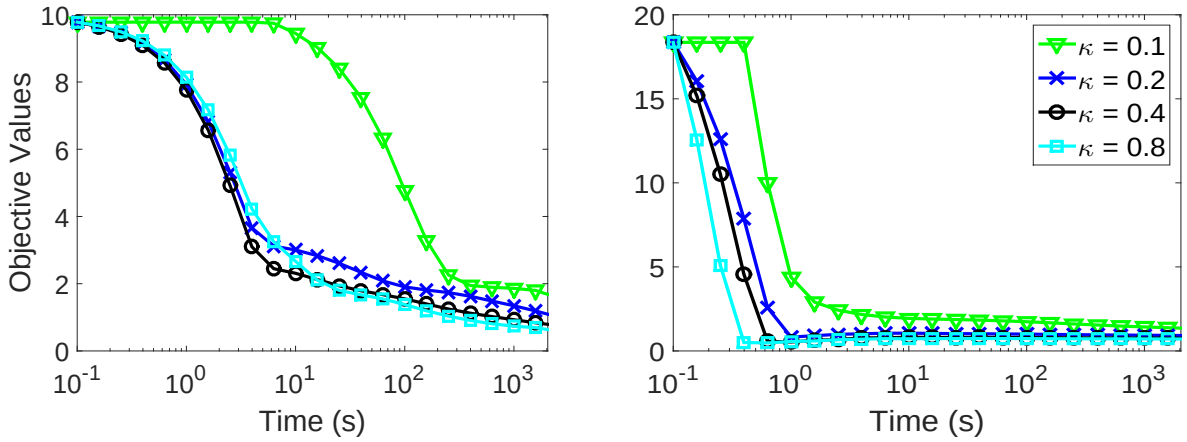
(c) $K = 100$ (d) $K = 100$

Fig. 3. The objective values (as a function of time) of all the algorithms for different values of K on the CBCL face dataset. In (a) and (b), $K = 25$. In (c) and (d), $K = 100$. All the other parameters are set according to the canonical setting.



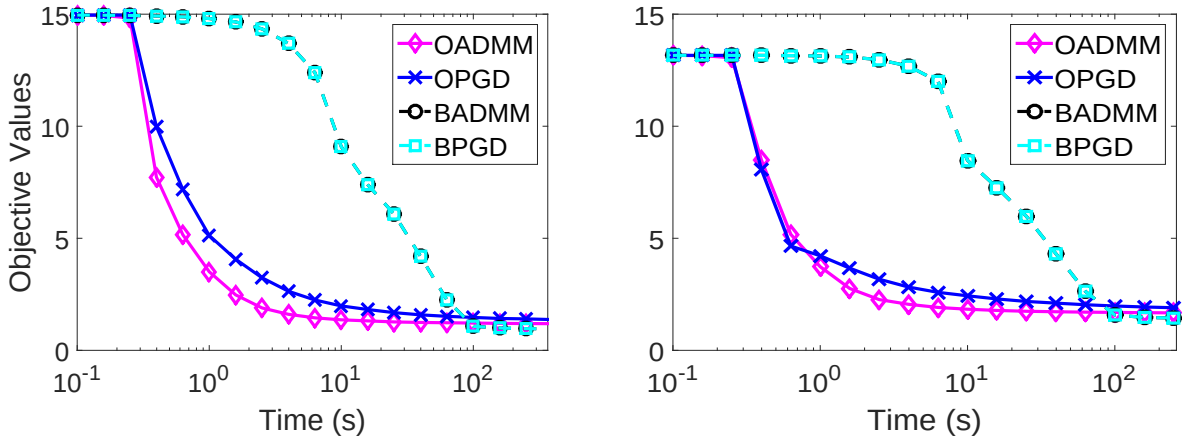
(a) BADMM (b) OADMM

Fig. 4. The objective values (as a function of time) of (a) BADMM and (b) OADMM for different values of ρ on the CBCL face dataset. All the other parameters are set according to the canonical setting.



(a) BPGD (b) OPGD

Fig. 5. The objective values (as a function of time) of (a) BPGD and (b) OPGD for different values of κ on the CBCL face dataset. All the other parameters are set according to the canonical setting.



(a) $\nu = 0.8$, $\widetilde{\nu} = 0.2$ (b) $\nu = 0.9$, $\widetilde{\nu} = 0.3$

Fig. 6. The objective values (as a function of time) of our online and batch algorithms on the CBCL face dataset with a larger proportion of outliers.

TABLE II
RUNNING TIMES (IN SECONDS) OF ALL THE ALGORITHMS ON THE CBCL FACE DATASET WITH DIFFERENT NOISE DENSITY.

	Setting 1	Setting 2	Setting 3
OADMM	420.58 $\pm$ 2.59	427.66 $\pm$ 4.80	430.02 $\pm$ 2.93
OPGD	431.66 $\pm$ 2.49	455.15 $\pm$ 1.70	463.67 $\pm$ 1.12
BADMM	1009.45 $\pm$ 11.27	1184.29 $\pm$ 10.49	1240.91 $\pm$ 8.21
BPGD	1125.58 $\pm$ 12.83	1185.64 $\pm$ 13.36	1279.07 $\pm$ 9.08
ONMF	2384.70 $\pm$ 9.59	2588.29 $\pm$ 14.39	2698.57 $\pm$ 10.24
ORPCA	365.98 $\pm$ 5.29	382.49 $\pm$ 4.20	393.10 $\pm$ 4.27

(a) $K = 25$

	Setting 1	Setting 2	Setting 3
OADMM	422.97 $\pm$ 2.38	424.67 $\pm$ 2.65	434.17 $\pm$ 4.67
OPGD	430.19 $\pm$ 2.27	448.72 $\pm$ 3.90	454.30 $\pm$ 4.65
BADMM	1009.04 $\pm$ 8.53	1187.58 $\pm$ 5.06	1250.53 $\pm$ 4.67
BPGD	1131.89 $\pm$ 7.04	1192.46 $\pm$ 7.43	1280.55 $\pm$ 7.93
ONMF	2379.86 $\pm$ 15.18	2591.09 $\pm$ 11.91	2693.08 $\pm$ 12.48
ORPCA	363.37 $\pm$ 3.01	390.58 $\pm$ 5.31	401.21 $\pm$ 3.27

(b) $K = 100$