Supplemental Material for "Online Nonnegative Matrix Factorization with General Divergences"

S-1. Related Works

A. Online Matrix Factorization Beyond the Squared-ℓ_2 Loss

In the literature of online matrix factorization [1]–[5], it is assumed that i.i.d. (independent and identically distributed) data samples $\{\mathbf{v}_t\}_{t\in\mathbb{N}}$ (drawn from a common distribution $\mathbb{P}$) arrive in a streaming manner, and the storage space does not scale with time. Under such setting, we aim to solve the following *stochastic program*, i.e., minimize the expected loss

$$\min_{\mathbf{W}\in\mathcal{C}} \mathbb{E}_{\mathbf{v}\sim\mathbb{P}}[\widetilde{\ell}(\mathbf{v},\mathbf{W})], \tag{S-1}$$

where $\mathbf{v}$ is the (random) data vector with distribution $\mathbb{P}$, $\mathbf{W}$ the basis matrix constrained in the set $\mathcal{C}$ and $\widetilde{\ell}$ the loss function with respect to (w.r.t.) a single data sample $\mathbf{v}$. Most of the literature on online matrix factorization (including online NMF [2], online dictionary learning [1] and online low-rank representation [5]) focus on the case where the data fidelity term is the squared ℓ_2 loss, i.e., $\widetilde{\ell}$ is defined as

$$\widetilde{\ell}(\mathbf{v},\mathbf{W}) \triangleq \min_{\mathbf{h}\in\mathcal{H}} \|\mathbf{v}-\mathbf{W}\mathbf{h}\|^2 + \lambda\psi(\mathbf{h}), \tag{S-2}$$

where $\mathbf{h}$ is the coefficient vector constrained in the set $\mathcal{H}$ and $\lambda\psi(\mathbf{h})$ is some regularizer on $\mathbf{h}$ with penalty parameter $\lambda>0$. However, the literature with other forms of data fidelity terms is relative scarce. Among them, some works on real-time music signal processing [6], [7] consider minimizing the IS divergence in an online manner. Other works on visual tracking [8], [9] consider the online minimization of the Huber loss. However, almost all of methods proposed in these works are *heuristic* in nature, in the sense that the global convergence of the sequence (or any subsequence) of the dictionaries $\{\mathbf{W}_t\}_{t\in\mathbb{N}}$ cannot be guaranteed (either a.s. or with high probability). Furthermore, since most of these works are conducted on an *ad hoc* basis, the approaches therein cannot be easily generalized to other divergences in a straightforward manner. As different divergences are suited to different applications in practice (see Section 2.1), a unified framework is needed to systematically study the convergence properties of NMF for various divergences.

B. Stochastic Projected Subgradient Descent (SPSGD) Applied to Online Matrix Factorization

As discussed in Section S-1-A, only *differentiable* data fidelity terms (the IS divergence and the Huber loss) are considered in the literature of online matrix factorization. Thus, only the stochastic projected gradient descent (SPGD) method has been employed in the prior works [1], [2], [8]–[10]. In particular, the efficacy of such method with the squared-ℓ_2 loss and the Huber loss has been empirically verified in [1] and [8], [9] respectively. In [10], SPGD was employed on online dictionary learning over distributed models, with both squared-ℓ_2 loss and the Huber loss. In [2], the authors leverage the robust stochastic approximation method [11], a variant of SPGD, and consider both the squared ℓ_2 loss and the IS divergence. However, for all the abovementioned works, convergence guarantees on the sequence (or any subsequence) of the dictionaries $\{\mathbf{W}_t\}_{t\in\mathbb{N}}$ generated by the SPGD algorithm have not been established.

S-2. Algorithms

A. Implementations of $\Pi_{\mathcal{C}}$ and $\Pi_{\mathcal{H}}$

The projection operator $\Pi_{\mathcal{C}}$ in (5) can be implemented in a straightforward manner if the data point lies in $\mathcal{C}' \triangleq \{\mathbf{W}\in\mathbb{R}_+^{F\times K} \mid \|\mathbf{W}_{i:}\|_1 \geq \epsilon, \forall\, i\in[F]\}$. Otherwise, if there exists $i\in[F]$ such that $\|\mathbf{W}_{i:}\|_1<\epsilon$, $\Pi_{\mathcal{C}}$ amounts to projecting $\mathbf{W}_{i:}$ onto the probability simplex in $\mathbb{R}_+^K$. Efficient algorithms have been extensively discussed in the literature, for e.g., [12, Section 3]. Since $\epsilon<1$, the constraint $w_{ij}\leq 1$, for any $j\in[K]$ is automatically satisfied after such projection. The projection onto the set $\mathcal{H}$ simply involves entrywise thresholding.

B. Choice of step sizes $\{\beta_t^k\}_{k\in\mathbb{N}}$ in Algorithm 2

For the divergences $d(\cdot\|\cdot)\in\mathcal{D}_1$, the corresponding function $\overline{d}_t$ is differentiable on $\mathcal{H}$ and $\nabla\overline{d}_t$ is Lipschitz on $\mathcal{H}$ with Lipschitz constant $L_t>0$. For these cases, there are two ways to choose the step sizes $\{\beta_t^k\}_{k\in\mathbb{N}}$ such that the sequence of iterates $\{\mathbf{h}_t^k\}_{k\in\mathbb{N}}$ converges to the set of critical points of (4) as $k\to\infty$.[1] The first approach is the well-known Armijo rule, which applies to all the continuously differentiable g_t (see [13, Theorem 2.4] for details). The implementation of Armijo rule

[1]Given a finite-dimensional Banach space $\mathcal{X}$, a sequence (x_n) in $\mathcal{X}$ is said to converge to a set $\mathcal{A}\subseteq\mathcal{X}$ if $\lim_{n\to\infty}\inf_{a\in\mathcal{A}}\|x_n-a\|=0$.

Algorithm S-1 Armijo rule for step size selection

Input: Dictionary matrix $\mathbf{W}_{t-1}$, data sample $\mathbf{v}_t$, coefficient vector $\mathbf{h}_t^k$, maximum number of iterations q
Initialize $\alpha \in (0, 0.5), \gamma \in (0,1), i := 0, \xi^0 := 1$
while $\overline{d}_t(\mathbf{h}_t^k - \xi^i \nabla \overline{d}_t(\mathbf{h}_t^k)) > \overline{d}_t(\mathbf{h}_t^k) - \alpha \xi^i \|\nabla \overline{d}_t(\mathbf{h}_t^k)\|^2$ **and** $i \leq q$

$$\xi^{i+1} := \gamma \xi^i, \quad i := i+1$$

end
Output: Final step size $\beta_t^k \triangleq \xi^i$

is shown in Algorithm S-1, where we set $\alpha = 0.01$ and $\gamma = 0.1$ following the suggestions in [14]. We also set $q = 10$. The second approach is to use constant step sizes, i.e., $\beta_t^k = \beta_t$, for all $k \in \mathbb{N}$. If $\beta_t \in (0, 1/L_t]$, then Algorithm 2 can be interpreted as an MM algorithm [15], and the convergence is guaranteed by [16, Theorem 1]. (In this work, we set $\beta_t = 1/L_t$ for simplicity.) We now provide some guidelines for choosing between these two approaches. The second approach is suitable for the functions $\nabla \overline{d}_t$ whose smallest Lipschitz constant on any subset $\mathcal{U} \subseteq \mathcal{H}$,[2] $L_t^*(\mathcal{U})$ does not vary much across all the subsets of $\mathcal{H}$. Examples of the corresponding divergences include the Huber loss and the squared ℓ_2 loss. However, the gradients $\nabla \overline{d}_t$ corresponding to some other divergences (e.g., the IS and the KL divergences) in general have much larger $L_t^*(\mathcal{U})$ when $\mathcal{U}$ is in the vicinity of $\mathbf{bd}\,\mathcal{H}$ than elsewhere. Since $L_t \geq \sup_{\mathcal{U} \subseteq \mathcal{H}} L_t^*(\mathcal{U})$, the constant step size β_t will be very small even when $\mathbf{h}_t^k$ lies in the "center" of $\mathcal{H}$, where $\overline{d}_t$ is relatively smooth. Under such scenario, it is more appropriate to use Armijo rule especially when the evaluation of $\overline{d}_t$ is not expensive. Now we consider the divergences $d(\cdot\|\cdot) \notin \mathcal{D}_1$, i.e., the ℓ_1 and ℓ_2 losses. For the ℓ_2 loss, the first approach above is still applicable since $\|\cdot\|$ is non-differentiable only at $\mathbf{0}$. For the ℓ_1 loss, we employ the *modified Polyak's step size policy* with tolerance parameter δ_{tol} (set to 0.01 in this work) [17], [18, Section 6.3.1] due to efficiency considerations. Although this step size policy can only guarantee $\liminf_{k\to\infty} \overline{d}_t(\mathbf{h}_t^k) \leq \min_{\mathbf{h}\in\mathcal{H}} \overline{d}_t(\mathbf{h}) + \delta_{\text{tol}}$, as shown in Section 6, it performs reasonably well empirically.

S-3. Convergence Analysis

A. Proof of Lemma 4

First we rewrite (11) as

$$\frac{d}{ds} W(s) = -\nabla f(W(s)) + z(s),\ W(0) = \mathbf{W}_0,\ s \geq 0, \tag{S-3}$$

where

$$z(s) \triangleq \pi_{\mathcal{C}}\Big[W(s), -\nabla f(W(s))\Big] + \nabla f(W(s)), s \in \mathbb{R}_+. \tag{S-4}$$

From Lemma 3 and Lemma S-11, there exists an almost sure set $\mathcal{A} \in \Omega$ such that for each $\omega \in \mathcal{A}$, $\{W^t(\omega,\cdot)\}_{t\in\mathbb{N}}$ and $\{Z^t(\omega,\cdot)\}_{t\in\mathbb{N}}$ are asymptotically equicontinuous on $\mathbb{R}_+$. Due to the compactness of $\mathcal{C}$, $\{W^t(\omega,\cdot)\}_{t\in\mathbb{N}}$ and $\{Z^t(\omega,\cdot)\}_{t\in\mathbb{N}}$ are also uniformly bounded. Fix $S \in (0,\infty)$. By the (generalized) Arzelà-Ascoli Theorem (see Lemma S-13), there exists a sequence $\{t_k\}_{k\in\mathbb{N}}$ such that $t_k \uparrow \infty$, a continuous $\overline{W}(\omega,\cdot)$ and a continuous $\overline{Z}(\omega,\cdot)$ such that $W^{t_k}(\omega,\cdot) \xrightarrow{\text{u}} \overline{W}(\omega,\cdot)$ and $Z^{t_k}(\omega,\cdot) \xrightarrow{\text{u}} \overline{Z}(\omega,\cdot)$ on $[0,S]$. (Note that $\{t_k\}_{k\in\mathbb{N}}$, $\overline{W}(\omega,\cdot)$ and $\overline{Z}(\omega,\cdot)$ may depend on S.) Define

$$\overline{G}(s) \triangleq -\int_0^s \nabla f\left(\overline{W}(\tau)\right)\, d\tau,\ s \in [0,S]. \tag{S-5}$$

We now show $G^{t_k}(\omega,\cdot) \xrightarrow{\text{u}} \overline{G}(\omega,\cdot)$ on $[0,S]$. By Lemma S-12 and continuity of ∇f, we have $\nabla f(W^{t_k}(\omega,\cdot)) \xrightarrow{\text{u}} \nabla f\left(\overline{W}(\omega,\cdot)\right)$ on $[0,S]$. Thus

$$\begin{aligned}
\lim_{t\to\infty} \sup_{s\in[0,S]} \left\|G^{t_k}(\omega,s) - \overline{G}(\omega,s)\right\| &\leq \lim_{t\to\infty} \sup_{s\in[0,S]} \int_0^s \left\|\nabla f(\overline{W}(\omega,\tau)) - \nabla f(W^{t_k}(\omega,\tau))\right\| d\tau \\
&\leq S \lim_{t\to\infty} \sup_{s\in[0,S]} \left\|\nabla f(\overline{W}(\omega,s)) - \nabla f(W^{t_k}(\omega,s))\right\| \\
&= 0.
\end{aligned}$$

From Lemma 3, we also have $\Delta_1^{t_k}(\omega,\cdot) \xrightarrow{\text{u}} \mathbf{0}$ and $N^{t_k}(\omega,\cdot) \xrightarrow{\text{u}} \mathbf{0}$ on $[0,S]$. Therefore, from (8), we have

$$\overline{W}(\omega,s) = \overline{W}(\omega,0) - \int_0^s \nabla f\left(\overline{W}(\omega,\tau)\right)\, d\tau + \overline{Z}(\omega,s),\ s \in [0,S]. \tag{S-6}$$

[2]For any $t \in \mathbb{N}$, the smallest Lipschitz constant of $\nabla \overline{d}_t$ on $\mathcal{U}$, $L_t^*(\mathcal{U}) \triangleq \inf\{L \,|\, \|\nabla \overline{d}_t(\mathbf{h}_1) - \nabla \overline{d}_t(\mathbf{h}_2)\| \leq L\,\|\mathbf{h}_1 - \mathbf{h}_2\|, \forall\, \mathbf{h}_1, \mathbf{h}_2 \in \mathcal{U}\}$.

Thus, to show $\left\{\overline{W}(\omega,\cdot),\overline{Z}(\omega,\cdot)\right\}$ satisfies (the integral form) of (S-3) (on $[0,S]$), it remains to show $\overline{Z}(\omega,s)=\int_0^s z(\tau)\,d\tau$, $s\in[0,S]$. By the definition of $\{Z^t(\omega,\cdot)\}_{t\in\mathbb{N}}$, we have $\overline{Z}(\omega,0)=0$. Also, by the closedness of $\mathcal{C}$, we have $\overline{W}(\omega,s)\in\mathcal{C}$, for all $s\geq 0$. First we define the inward normal set at $\mathbf{W}\in\mathcal{C}$, $\mathcal{N}(\mathbf{W})$ as

$$\mathcal{N}(\mathbf{W})\triangleq\begin{cases}\left\{\mathbf{N}\in\mathbb{R}^{F\times K}\mid\|\mathbf{N}\|\leq M,\,\langle\mathbf{N},\mathbf{W}'-\mathbf{W}\rangle\geq 0,\,\forall\,\mathbf{W}'\in\mathcal{C}\right\}, & \mathbf{W}\in\mathbf{bd}\,\mathcal{C}\\ \left\{\mathbf{0}\in\mathbb{R}^{F\times K}\right\}, & \mathbf{W}\in\mathbf{int}\,\mathcal{C}\end{cases}. \tag{S-7}$$

From (S-7), we notice that $\mathcal{N}(\mathbf{W})$ is compact and convex for any $\mathbf{W}\in\mathcal{C}$. By the definition of $Z^t(\omega,\cdot)$, it is also obvious that for any $t\in\mathbb{N}$ and $s\geq 0$, $Z^t(\omega,s)\in\mathcal{N}(W^t(\omega,s))$.

By Lemma S-14, it suffices to show $\overline{W}(\omega,\cdot)$ is Lipschitz on $[0,S]$ and for any $\tau\in[0,S]$

1) $\overline{Z}(\omega,\tau)=\mathbf{0}$ if $\overline{W}(\omega,s)\in\mathbf{int}\,\mathcal{C}$ for almost all $s\in[0,\tau]$ (in the sense of Lebesgue measure),
2) $\overline{Z}(\omega,\tau)\in\overline{\mathbf{conv}}\left[\bigcup_{s\in[0,\tau]}\mathcal{N}\left(\overline{W}(\omega,s)\right)\right]$.

First, we show $\overline{Z}(\omega,\cdot)$ is Lipschitz on $[0,S]$. By Lemma 3, we have for any $s_0,s_1\in[0,S]$,

$$\begin{aligned}\left\|\overline{Z}(\omega,s_0)-\overline{Z}(\omega,s_1)\right\|&=\lim_{k\to\infty}\left\|Z^{t_k}(\omega,s_0)-Z^{t_k}(\omega,s_1)\right\|\\&\leq\lim_{k\to\infty}\left\|Y^{t_k}(\omega,s_0)-Y^{t_k}(\omega,s_1)\right\|+\left\|\Delta_2^{t_k}(\omega,s_0)-\Delta_2^{t_k}(\omega,s_1)\right\|\\&\leq\lim_{k\to\infty}\left\|\int_{s_0}^{s_1}Z^{t_k}(\omega,\tau)\,d\tau\right\|+2\sup_{s\in\mathbb{R}_+}\left\|\Delta_2^{t_k}(\omega,s)\right\|\\&\leq\lim_{k\to\infty}|s_0-s_1|\sup_{\tau\in[s_0,s_1]}\left\|Z^{t_k}(\omega,\tau)\right\|\\&\leq M\,|s_0-s_1|\,.\end{aligned}$$

Since $\nabla f\left(\overline{W}(\omega,\cdot)\right)$ is bounded on $[0,S]$, by (S-6), we conclude $\overline{W}(\omega,\cdot)$ is Lipschitz on $[0,S]$. Next, since $\overline{W}(\omega,s)\in\mathbf{int}\,\mathcal{C}$ for almost all $s\in[0,\tau]$, there exists $\{s_n\}_{n\in\mathbb{N}}$ in $[0,\tau]$ such that $s_n\uparrow\tau$ and $\overline{W}(\omega,s_n)\in\mathbf{int}\,\mathcal{C}$ for all $n\in\mathbb{N}$. Hence $\overline{Z}(\omega,s_n)=\mathbf{0}$ for all $n\in\mathbb{N}$. The continuity of $\overline{Z}(\omega,\cdot)$ implies $\overline{Z}(\omega,\tau)=\mathbf{0}$. To show the last claim, we leverage the upper semicontinuity of the correspondence $\mathcal{N}$ (see Definition S-2). We first show $\mathcal{N}$ is upper semicontinuous on $\mathcal{C}$ by Lemma S-17. It suffices to show

$$\bigcap_{\delta>0}\overline{\mathbf{conv}}\left(\bigcup_{\mathbf{W}'\in\mathcal{B}_\delta(\mathbf{W})}\mathcal{N}(\mathbf{W}')\right)\subseteq\mathcal{N}(\mathbf{W}),\,\forall\,\mathbf{W}\in\mathcal{C}, \tag{S-8}$$

where $\mathcal{B}_\delta(\mathbf{W})\triangleq\{\mathbf{W}'\in\mathcal{C}\mid\|\mathbf{W}-\mathbf{W}'\|<\delta\}$. Suppose (S-8) is false, then for any $\delta>0$, there exists $\mathbf{W}_0\in\mathcal{C}$ and $\mathbf{N}_0$ such that $\mathbf{N}_0\in\overline{\mathbf{conv}}\left(\bigcup_{\mathbf{W}'\in\mathcal{B}_\delta(\mathbf{W}_0)}\mathcal{N}(\mathbf{W}')\right)$ and $\mathbf{N}_0\notin\mathcal{N}(\mathbf{W}_0)$. For any $\epsilon>0$, there exists $\mathbf{N}'\in\mathcal{B}_\epsilon(\mathbf{N}_0)$, $\lambda\in[0,1]$ and $\mathbf{W}_1,\mathbf{W}_2\in\mathcal{B}_\delta(\mathbf{W}_0)$ such that $\mathbf{N}'=\lambda\mathbf{N}_1+(1-\lambda)\mathbf{N}_2$, where $\mathbf{N}_i\in\mathcal{N}(\mathbf{W}_i)$, $i=1,2$. Hence for any $\mathbf{W}'\in\mathcal{C}$,

$$\begin{aligned}\langle\mathbf{N}_0,\mathbf{W}'-\mathbf{W}_0\rangle&=\langle\mathbf{N}',\mathbf{W}'-\mathbf{W}_0\rangle+\langle\mathbf{N}_0-\mathbf{N}',\mathbf{W}'-\mathbf{W}_0\rangle\\&=\lambda\langle\mathbf{N}_1,\mathbf{W}'-\mathbf{W}_0\rangle+(1-\lambda)\langle\mathbf{N}_2,\mathbf{W}'-\mathbf{W}_0\rangle+\langle\mathbf{N}_0-\mathbf{N}',\mathbf{W}'-\mathbf{W}_0\rangle\\&=\lambda\langle\mathbf{N}_1,\mathbf{W}'-\mathbf{W}_1\rangle+\lambda\langle\mathbf{N}_1,\mathbf{W}_1-\mathbf{W}_0\rangle+(1-\lambda)\langle\mathbf{N}_2,\mathbf{W}'-\mathbf{W}_2\rangle\\&\qquad+(1-\lambda)\langle\mathbf{N}_2,\mathbf{W}_2-\mathbf{W}_0\rangle+\langle\mathbf{N}_0-\mathbf{N}',\mathbf{W}'-\mathbf{W}_0\rangle\\&\geq-\lambda\|\mathbf{N}_1\|\|\mathbf{W}_1-\mathbf{W}_0\|-(1-\lambda)\|\mathbf{N}_2\|\|\mathbf{W}_2-\mathbf{W}_0\|-\|\mathbf{N}_0-\mathbf{N}'\|\|\mathbf{W}'-\mathbf{W}_0\|\\&\geq-\lambda\delta\|\mathbf{N}_1\|-(1-\lambda)\delta\|\mathbf{N}_2\|-\epsilon\|\mathbf{W}'-\mathbf{W}_0\|\\&\geq-(\delta M+\epsilon\mathbf{diam}\,\mathcal{C}),\end{aligned}$$

where $\mathbf{diam}\,\mathcal{C}\triangleq\max_{\mathbf{X},\mathbf{Y}\in\mathcal{C}}\|\mathbf{X}-\mathbf{Y}\|$. The compactness of $\mathcal{C}$ implies $\mathbf{diam}\,\mathcal{C}<\infty$. Let both $\delta\to 0$ and $\epsilon\to 0$ we have $\langle\mathbf{N}_0,\mathbf{W}'-\mathbf{W}_0\rangle\geq 0$, for any $\mathbf{W}'\in\mathcal{C}$. This contradicts $\mathbf{N}_0\notin\mathcal{N}(\mathbf{W}_0)$. Thus we conclude that $\mathcal{N}$ is upper semicontinuous on $\mathcal{C}$. Since $\mathcal{N}$ is compact-valued, $\mathcal{G}(\mathcal{N})$ is closed by Lemma S-16. Again, take a sequence $\{s_n\}_{n\in\mathbb{N}}$ in $[0,\tau]$ such that $s_n\uparrow\tau$. For any $n\in\mathbb{N}$, since $Z^{t_k}(\omega,s_n)\to\overline{Z}(\omega,s_n)$, $Z^{t_k}(\omega,s_n)\in\mathcal{N}(W^{t_k}(\omega,s_n))$ and $W^{t_k}(\omega,s_n)\to\overline{W}(\omega,s_n)$, by the closedness of $\mathcal{G}(\mathcal{N})$, we have $\overline{Z}(\omega,s_n)\in\mathcal{N}(\overline{W}(\omega,s_n))$. Since $\overline{Z}(\omega,s_n)\to\overline{Z}(\omega,\tau)$, $\overline{W}(\omega,s_n)\to\overline{W}(\omega,\tau)$, we have $\overline{Z}(\omega,\tau)\in\mathcal{N}(\overline{W}(\omega,\tau))\subseteq\overline{\mathbf{conv}}\left[\bigcup_{s\in[0,\tau]}\mathcal{N}\left(\overline{W}(\omega,s)\right)\right]$.

Now, fix a sequence $\{S_n\}_{n\in\mathbb{N}}\subseteq(0,\infty)$ such that $S_n\uparrow\infty$, and let $\left\{\overline{W}_n(\omega,\cdot),\overline{Z}_n(\omega,\cdot)\right\}$ be the (continuous) limit functions corresponding to S_n. Fix $i\in\mathbb{N}$. For any $S_i<S_j$, there exist $\left\{\overline{W}_j(\omega,\cdot),\overline{Z}_k(\omega,\cdot)\right\}$ such that $\overline{W}_i(\omega,\cdot)=\overline{W}_j(\omega,\cdot)$ and $\overline{Z}_i(\omega,\cdot)=\overline{Z}_j(\omega,\cdot)$ on $[0,S_i]$. Thus there exists $\{\bar{t}_k\}_{k\in\mathbb{N}}$ such that $W^{\bar{t}_k}(\omega,\cdot)\xrightarrow{\mathrm{u}}\overline{W}_\infty(\omega,\cdot)$ and $Z^{\bar{t}_k}(\omega,\cdot)\xrightarrow{\mathrm{u}}\overline{Z}_\infty(\omega,\cdot)$ on $\mathbb{R}_+$. Moreover, the continuous limit functions $\left\{\overline{W}_\infty(\omega,\cdot),\overline{Z}_\infty(\omega,\cdot)\right\}$ satisfy (S-3) on $\mathbb{R}_+$.

The above implies the solution set of (S-3) is nonempty. Moreover, the compactness of $\mathcal{C}$ implies the limit set of (S-3), $\mathcal{L}(-\nabla f,\mathcal{C},\mathbf{W}_0)\neq\emptyset$. For any convergent subsequence $\{\mathbf{W}_{t_l}(\omega)\}_{l\in\mathbb{N}}$, there exist a non-decreasing sequence $\{t'_l\}_{l\in\mathbb{N}}\subseteq\{\bar{t}_k\}_{k\in\mathbb{N}}$

with $t_l' \uparrow \infty$ and $\{\tau_l\}_{l\in\mathbb{N}} \uparrow \infty$ such that $t_l = m(\tau_l + s_{t_l'})$, for all $l \in \mathbb{N}$. Therefore,

$$\begin{aligned}
\lim_{l\to\infty} \mathbf{dist}\left(\mathbf{W}_{t_l}(\omega), \mathcal{L}(-\nabla f, \mathcal{C}, \mathbf{W}_0)\right) &= \lim_{l\to\infty} \inf_{\mathbf{W}\in\mathcal{L}(-\nabla f,\mathcal{C},\mathbf{W}_0)} \left\|W^{t_l'}(\omega,\tau_l) - \mathbf{W}\right\| \\
&\le \lim_{l\to\infty} \inf_{\mathbf{W}\in\mathcal{L}(-\nabla f,\mathcal{C},\mathbf{W}_0)} \left\|W^{t_l'}(\omega,\tau_l) - \overline{W}_\infty(\omega,\tau_l)\right\| + \left\|\overline{W}_\infty(\omega,\tau_l) - \mathbf{W}\right\| \\
&\le \lim_{l\to\infty} \sup_{s\ge 0} \left\|W^{t_l'}(\omega,s) - \overline{W}_\infty(\omega,s)\right\| + \lim_{l\to\infty} \inf_{\mathbf{W}\in\mathcal{L}(-\nabla f,\mathcal{C},\mathbf{W}_0)} \left\|\overline{W}_\infty(\omega,\tau_l) - \mathbf{W}\right\| \\
&\overset{(a)}{=} \lim_{l\to\infty} \mathbf{dist}\left(\overline{W}_\infty(\omega,\tau_l), \mathcal{L}(-\nabla f, \mathcal{C}, \mathbf{W}_0)\right) \\
&\overset{(b)}{=} 0,
\end{aligned}$$

where in (a) we use the fact that $W^{\bar{t}_k}(\omega,\cdot) \xrightarrow{\text{u}} \overline{W}_\infty(\omega,\cdot)$ on $\mathbb{R}_+$ and in (b) we use the definition of $\mathcal{L}(-\nabla f, \mathcal{C}, \mathbf{W}_0)$. Thus we conclude that $\lim_{l\to\infty} \mathbf{W}_{t_l}(\omega) \in \mathbf{cl}\,\mathcal{L}(-\nabla f, \mathcal{C}, \mathbf{W}_0)$. Hence we prove $\mathbf{W}_t(\omega) \to \mathcal{L}(-\nabla f, \mathcal{C}, \mathbf{W}_0)$ as $t \to \infty$.

B. Proof of Lemma 5

We leverage the Lyapunov stability theory [19, Section 6.6] to prove the lemma. First, define $L : \mathcal{C} \to \mathbb{R}$ such that $L(\mathbf{W}) \triangleq f(\mathbf{W}) - \min_{\mathbf{W}\in\mathcal{C}} f(\mathbf{W})$, $\mathbf{W} \in \mathcal{C}$. By Definition S-1, we have that L is a Lyapunov function (with possibly non-unique zeros on $\mathcal{C}$). By [19, Theorem 6.15] (see Lemma S-15),

$$\mathcal{L}(-\nabla f, \mathcal{C}, \mathbf{W}_0) \subseteq \bigcup_{W(\cdot)\in\mathcal{P}(-\nabla f,\mathcal{C},\mathbf{W}_0)} \Big\{W(s) \,\Big|\, \frac{d}{ds} L(W(s)) = 0\Big\},$$

where

$$\frac{d}{ds} L(W(s)) = \Big\langle \nabla f(W(s)), \pi_{\mathcal{C}}\Big[W(s), -\nabla f(W(s))\Big]\Big\rangle,\ s \ge 0.$$

Given $\langle \nabla f(\mathbf{W}), \pi_{\mathcal{C}}[\mathbf{W}, -\nabla f(\mathbf{W})]\rangle = 0$, it is obvious that $\pi_{\mathcal{C}}[W(s), -\nabla f(W(s))] = \mathbf{0}$, if there exists $\delta > 0$ such that $\mathbf{W} - \delta\nabla f(\mathbf{W}) \in \mathcal{C}$. Otherwise, by the convexity of $\mathcal{C}$,

$$\begin{aligned}
&\mathbf{dist}^{\,2}(\mathbf{W} - \nabla f(\mathbf{W}), \mathcal{C}) \\
&\ge \left\|\pi_{\mathcal{C}}\left[\mathbf{W}, -\nabla f(\mathbf{W})\right] + \nabla f(\mathbf{W})\right\|^2 \\
&= \left\|\pi_{\mathcal{C}}\left[\mathbf{W}, -\nabla f(\mathbf{W})\right]\right\|^2 + \left\|\nabla f(\mathbf{W})\right\|^2 \\
&= \left\|\pi_{\mathcal{C}}\left[\mathbf{W}, -\nabla f(\mathbf{W})\right]\right\|^2 + \left\|(\mathbf{W} - \nabla f(\mathbf{W})) - \mathbf{W}\right\|^2 \\
&\ge \left\|\pi_{\mathcal{C}}\left[\mathbf{W}, -\nabla f(\mathbf{W})\right]\right\|^2 + \mathbf{dist}^{\,2}(\mathbf{W} - \nabla f(\mathbf{W}), \mathcal{C}),
\end{aligned}$$

where for any $\mathbf{Y} \in \mathbb{R}^{F\times K}$, $\mathbf{dist}\,(\mathbf{Y}, \mathcal{C}) \triangleq \|\Pi_{\mathcal{C}}\mathbf{Y} - \mathbf{Y}\|$. Hence we conclude $\pi_{\mathcal{C}}\left[\mathbf{W}, -\nabla f(\mathbf{W})\right] = \mathbf{0}$. Thus we conclude $\mathcal{L}(-\nabla f, \mathcal{C}, \mathbf{W}_0) \subseteq \mathcal{S}(-\nabla f, \mathcal{C})$.

We show the second claim in a similar way. Before we proceed, let us first define a supporting hyperplane (see [20, Section 2.5.2]) at any $\mathbf{W} \in \mathbf{bd}\,\mathcal{C}$, $\mathcal{T}_{\mathbf{W}}$ as[3]

$$\mathcal{T}_{\mathbf{W}} \triangleq \{\mathbf{W}' \in \mathbb{R}^{F\times K} \mid \langle \mathbf{T}, \mathbf{W}' - \mathbf{W}\rangle = 0\}, \tag{S-9}$$

where the (outward) normal $\mathbf{T} \in \mathbb{R}^{F\times K}$ of $\mathcal{T}_{\mathbf{W}}$ satisfies $\langle \mathbf{T}, \mathbf{W}' - \mathbf{W}\rangle \le 0$, for all $\mathbf{W}' \in \mathcal{C}$. Given $\pi_{\mathcal{C}}\left[\mathbf{W}, -\nabla f(\mathbf{W})\right] = \mathbf{0}$, we only focus on the case where $\mathbf{W} \in \mathbf{bd}\,\mathcal{C}$ and for any $\delta > 0$, $\mathbf{W} - \delta\nabla f(\mathbf{W}) \notin \mathcal{C}$, otherwise the claim trivially holds. By the definition of $\pi_{\mathcal{C}}\left[\mathbf{W}, -\nabla f(\mathbf{W})\right]$ and convexity of $\mathcal{C}$, there exists a supporting hyperplane $\mathcal{T}_{\mathbf{W}}$ such that

$$\pi_{\mathcal{C}}\left[\mathbf{W}, -\nabla f(\mathbf{W})\right] = \Pi_{\mathcal{T}_{\mathbf{W}}}(\mathbf{W} - \nabla f(\mathbf{W})) - \mathbf{W}. \tag{S-10}$$

Since $\pi_{\mathcal{C}}\left[\mathbf{W}, -\nabla f(\mathbf{W})\right] = \mathbf{0}$, we have $\Pi_{\mathcal{T}_{\mathbf{W}}}(\mathbf{W} - \nabla f(\mathbf{W})) = \mathbf{W}$. This implies that $-\nabla f(\mathbf{W})$ is the (outward) normal of $\mathcal{T}_{\mathbf{W}}$. The definition of $\mathcal{T}_{\mathbf{W}}$ implies (12).

C. Proof of Lemma 3

We first present two corollaries of Lemma 1 and 2.

Corollary S-1. *We have* $\sup_{t\in\mathbb{N}} \mathbb{E}\left[\|\nabla_{\mathbf{W}}\ell(\mathbf{v}_t, \mathbf{W}_{t-1})\|^2\right] \le M^2$. *In addition, there exists a real constant* $M' > 0$ *such that for each* $\omega \in \Omega$, $\sup_{t\in\mathbb{N}} \|\nabla f(\mathbf{W}_{t-1}(\omega))\| \le M'$.

[3]Note that more than one supporting hyperplanes may exist at $\mathbf{W} \in \mathbf{bd}\,\mathcal{C}$. The supporting hyperplane that $\mathcal{T}_{\mathbf{W}}$ refers to depends on the context.

Corollary S-2. *$\{\mathbf{N}_t\}_{t\in\mathbb{N}}$ is a martingale difference sequence adapted to $\{\mathscr{F}_t\}_{t\geq 0}$. Moreover, there exists a real constant $M''>0$ such that $\sup_{t\in\mathbb{N}}\mathbb{E}\left[\|\mathbf{N}_t\|^2\right]\leq M''^2$.*

We first show $N^t \xrightarrow{\text{u}} \mathbf{0}$ on $\mathbb{R}_+$ a.s.. Fix $t\in\mathbb{N}$. Since $\{\eta_t\mathbf{N}_t\}_{t\in\mathbb{N}}$ is a martingale difference sequence (adapted to $\{\mathscr{F}_t\}_{t\geq 0}$), $\left\{\mathbf{M}_t \triangleq \sum_{l=1}^t \eta_l\mathbf{N}_l\right\}_{t\in\mathbb{N}}$ is a martingale. We shall prove $\{\mathbf{M}_t\}_{t\in\mathbb{N}}$ converges a.s. to a random variable $\mathbf{M}$. First, we see $\{\mathbf{M}_t\}_{t\in\mathbb{N}}$ is square-integrable since

$$\begin{aligned}
\sup_{t\in\mathbb{N}}\mathbb{E}\left[\|\mathbf{M}_t\|^2\right] &= \sup_{t\in\mathbb{N}}\mathbb{E}\left[\left\|\sum_{l=1}^t \eta_l\mathbf{N}_l\right\|^2\right] \\
&= \sup_{t\in\mathbb{N}}\sum_{l=1}^t \eta_l^2\mathbb{E}\left[\|\mathbf{N}_l\|^2\right] + \sum_{k\neq l}\eta_k\eta_l\mathbb{E}\left[\langle\mathbf{N}_k,\mathbf{N}_l\rangle\right] \\
&\overset{(a)}{=} \sup_{t\in\mathbb{N}}\sum_{l=1}^t \eta_l^2\mathbb{E}\left[\|\mathbf{N}_l\|^2\right] \\
&\overset{(b)}{\leq} M''^2\sup_{t\in\mathbb{N}}\sum_{l=1}^t\eta_l^2 \\
&< \infty,
\end{aligned}$$

where (a) follows the orthogonality of the martingale difference sequence and (b) follows from Corollary S-2. Moreover, by the continuities of $(\mathbf{v},\mathbf{W})\mapsto\nabla_{\mathbf{W}}\ell(\mathbf{v},\mathbf{W})$ (on $\mathcal{V}\times\mathcal{C}$) and ∇f (on $\mathcal{C}$) and compactness of $\mathcal{V}$ and $\mathcal{C}$, there exists a constant $C\in(0,\infty)$ such that $\sup_{t\in\mathbb{N}}\mathbb{E}\left[\|\mathbf{N}_t\|^2|\mathscr{F}_{t-1}\right]\leq C^2$ a.s.. Therefore,

$$\begin{aligned}
\sum_{t=2}^\infty\mathbb{E}\left[\|\mathbf{M}_t-\mathbf{M}_{t-1}\|^2|\mathscr{F}_{t-1}\right] &= \sum_{t=2}^\infty\eta_t^2\,\mathbb{E}\left[\|\mathbf{N}_t\|^2|\mathscr{F}_{t-1}\right] \\
&\leq C^2\sum_{t=2}^\infty\eta_t^2 \\
&< \infty \text{ a.s.}
\end{aligned}$$

Thus by Lemma S-7, there exists a (finite) random variable $\mathbf{M}$ such that $\mathbf{M}_n\xrightarrow{\text{a.s.}}\mathbf{M}$. Then there exists an almost sure set $\mathcal{A}\in\Omega$ such that for all $\omega\in\mathcal{A}$,

$$\begin{aligned}
\lim_{t\to\infty}\sup_{s\geq 0}\left\|N^t(\omega,s)\right\| &= \lim_{t\to\infty}\sup_{s>0}\left\|\sum_{i=t+1}^{m(s_t+s)}\eta_i\mathbf{N}_i(\omega)\right\| \\
&= \lim_{t\to\infty}\sup_{j\geq t+1}\|\mathbf{M}_j(\omega)-\mathbf{M}_t(\omega)\| \\
&\leq \lim_{t\to\infty}\sup_{j\geq t+1}\|\mathbf{M}_j(\omega)-\mathbf{M}(\omega)\|+\|\mathbf{M}_t(\omega)-\mathbf{M}(\omega)\| \\
&\leq 2\lim_{t\to\infty}\sup_{j\geq t}\|\mathbf{M}_j(\omega)-\mathbf{M}(\omega)\| \\
&= 0.
\end{aligned}$$

This implies $N^t\xrightarrow{\text{u}}\mathbf{0}$ on $\mathbb{R}_+$ a.s.. Moreover, by Lemma S-9, $\{N^t\}_{t\in\mathbb{N}}$ is asymptotically equicontinuous on $\mathbb{R}_+$ a.s..

We have $\Delta_1^t\xrightarrow{\text{u}}\mathbf{0}$ on $\mathbb{R}_+$ a.s. because for all $\omega\in\mathcal{A}$,

$$\begin{aligned}
\lim_{t\to\infty}\sup_{s\geq 0}\left\|\Delta_1^t(\omega,s)\right\| &= \lim_{t\to\infty}\sup_{s\geq 0}\left\|\int_0^s\nabla f(W^t(\omega,\tau))\,d\tau-\sum_{i=t}^{m(s_t+s)-1}\eta_{i+1}\nabla f(\mathbf{W}_i(\omega))\right\| \\
&\leq \lim_{t\to\infty}\sup_{j\geq t}\sup_{s'\in[s_j,s_{j+1}]}\left\|\int_{s'}^{s_{j+1}}\nabla f(W^t(\omega,\tau))\,d\tau\right\| \\
&\leq \lim_{t\to\infty}\sup_{j\geq t}\eta_{j+1}\|\nabla f(\mathbf{W}_j(\omega))\| \\
&\leq M'\limsup_{t\to\infty}\eta_t \\
&= 0,
\end{aligned}$$

where the second last step follows from Corollary S-1. By Lemma S-9, $\{\Delta_1^t\}_{t\in\mathbb{N}}$ is asymptotically equicontinuous on $\mathbb{R}_+$ a.s..

By the definition of G^t in (9), we observe for each $t \in \mathbb{N}$ and $\omega \in \mathcal{A}$, $G^t(\omega,\cdot)$ is continuous on $\mathbb{R}_+$ and continuously differentiable on $\mathbb{R}_+ \setminus \mathcal{Q}$ with $\frac{d}{ds}G^t(\omega,s) = -\nabla f(W^t(\omega,s))$, $s \in \mathbb{R}_+ \setminus \mathcal{Q}$, where $\mathcal{Q} \triangleq \{s_t\}_{t\geq 0}$. By Corollary S-1, we have $\sup_{t\in\mathbb{N}}\sup_{s\geq 0}\|\nabla f(W^t(\omega,s))\| \leq M'$. This implies each $G^t(\omega,\cdot)$ is Lipschitz with Lipschitz constant L_t and $\{L_t\}_{t\in\mathbb{N}}$ is bounded. Then by Lemma S-10, we conclude that $\{G^t(\omega,\cdot)\}_{t\in\mathbb{N}}$ is equicontinuous on $\mathbb{R}_+$. Since for each $t \in \mathbb{N}$ and $\omega \in \mathcal{A}$, $F^t(\omega,\cdot) = G^t(\omega,\cdot) + \Delta^t(\omega,\cdot)$, by Lemma S-11, $\{F^t\}_{t\in\mathbb{N}}$ is asymptotically equicontinuous on $\mathbb{R}_+$ a.s..

Using a similar argument, we can show for all $\omega \in \mathcal{A}$, $\Delta_2^t(\omega,\cdot) \xrightarrow{\text{u}} \mathbf{0}$ on $\mathbb{R}_+$. By the definition of $\mathbf{Z}_t$, we have for any $t \in \mathbb{N}$ and $\omega \in \mathcal{A}$, $\|\mathbf{Z}_t(\omega)\| \leq \|\nabla_{\mathbf{W}}\ell(\mathbf{v}_t(\omega),\mathbf{W}_{t-1}(\omega))\| \leq M$. Hence each $Y^t(\omega,\cdot)$ is Lipschitz with Lipschitz constant L_t' and $\{L_t'\}_{t\in\mathbb{N}}$ is bounded. Thus again by Lemma S-10, $\{Y^t(\omega,\cdot)\}_{t\in\mathbb{N}}$ is equicontinuous on $\mathbb{R}_+$. Consequently, we have $\{Z^t(\omega,\cdot)\}_{t\in\mathbb{N}}$ is asymptotically equicontinuous on $\mathbb{R}_+$.

D. Proof of Lemma 1

It is easy to check that i) $(\mathbf{v},\mathbf{W}) \mapsto d(\mathbf{v}\|\mathbf{W}\mathbf{h})$ is differentiable on $\mathcal{V}\times\mathcal{C}$, for each $\mathbf{h} \in \mathcal{H}$, ii) $(\mathbf{v},\mathbf{W},\mathbf{h}) \mapsto d(\mathbf{v}\|\mathbf{W}\mathbf{h})$ is continuous on $\mathcal{V}\times\mathcal{C}\times\mathcal{H}$ and iii) $(\mathbf{v},\mathbf{W},\mathbf{h}) \mapsto \nabla_{\mathbf{W}}d(\mathbf{v}\|\mathbf{W}\mathbf{h})$ and $(\mathbf{v},\mathbf{W},\mathbf{h}) \mapsto \nabla_{\mathbf{v}}d(\mathbf{v}\|\mathbf{W}\mathbf{h})$ are both continuous on $\mathcal{V}\times\mathcal{C}\times\mathcal{H}$. Furthermore, Assumption 2 implies $\mathbf{h}^*(\mathbf{v},\mathbf{W})$ is a unique minimizer of (4) for each $(\mathbf{v},\mathbf{W}) \in \mathcal{V}\times\mathcal{C}$. Then by the compactness of $\mathcal{H}$ and the maximum theorem (see Lemma S-3), $\mathbf{h}^*(\mathbf{v},\mathbf{W})$ is continuous on $\mathcal{V}\times\mathcal{W}$. By Danskin's theorem (see Lemma S-4) and again by the compactness of $\mathcal{H}$, $\ell(\mathbf{v},\mathbf{W})$ is differentiable on $\mathcal{V}\times\mathcal{W}$ and $\nabla_{\mathbf{W}}\ell(\mathbf{v},\mathbf{W}) = \nabla_{\mathbf{W}}d(\mathbf{v}\|\mathbf{W}\mathbf{h}^*(\mathbf{v},\mathbf{W}))$, which is continuous on $\mathcal{V}\times\mathcal{W}$. Since $\mathcal{V}\times\mathcal{W}$ is compact (by Assumption 1), there exists $M \in (0,\infty)$ such that $\|\nabla_{\mathbf{W}}\ell(\mathbf{v},\mathbf{W})\| \leq M$, for all $(\mathbf{v},\mathbf{W}) \in \mathcal{V}\times\mathcal{W}$.

E. Proof of Lemma 2

Since both $\mathbf{v} \mapsto \ell(\mathbf{v},\mathbf{W})$ and $\mathbf{v} \mapsto \nabla_{\mathbf{W}}\ell(\mathbf{v},\mathbf{W})$ are continuous on $\mathcal{V}$ (by Lemma 1) and $\mathcal{V}$ is compact, both of them are Lebesgue integrable. Thus, by Leibniz integral rule (see Lemma S-6), we have $\nabla f(\mathbf{W}) = \mathbb{E}_{\mathbf{v}}\left[\nabla_{\mathbf{W}}\ell(\mathbf{v},\mathbf{W})\right]$ for each $\mathbf{W} \in \mathcal{C}$. The continuity of ∇f on $\mathcal{C}$ is implied by the continuity of $\nabla_{\mathbf{W}}\ell(\mathbf{v},\mathbf{W})$ on $\mathcal{V}\times\mathcal{W}$.

F. Discussions

We first remark that for the divergences in class $\mathcal{D}_1 \cap \mathcal{D}_2$, it might be possible to analyze the convergence of Algorithm 1 under the stochastic MM framework, by choosing the quadratic majorant of the sample average of f, as per discussion in [21, Section 4]. However, our analysis based on stochastic approximation theory and projected dynamical systems [19], [22] serve as a more direct approach, since we need not transform Algorithm 1 as a stochastic MM algorithm a priori. Next, we discuss the difficulties to tackle the divergences in class $\mathcal{D}_1\Delta\mathcal{D}_2$. In such case f may be nonsmooth and nonconvex. Without additional assumptions,[4] proving asymptotic convergence guarantees to stationary points is still an open question in the literature. Moreover, nonconvexity makes solving (4) NP-hard. If we assume there exists an oracle that can solve (4), a possible approach to prove convergence to critical points would be to generalize the convergence analysis for the SPSGD method to the nonconvex problems.[5]

S-4. Applications and Experiments

A. Discussions on Choices of Parameters

The mini-batch size τ controls the frequency of dictionary update. In the online NMF literature there are no principled ways to select τ, since this parameter is typically data dependent [1]. In our experiments, we set $\tau = 20$ since it yielded satisfactory performances. For the latent dimension K, there are several ways to choose it. The most direct way leverages domain knowledge. For example, if the data matrix is the term-document matrix, then K corresponds to the number of topics (categories) that the documents belong to (such that each document can be viewed as a linear combination of keywords in each topic). Since this number is known for most text datasets, the value of K can be directly obtained. Otherwise, some works [27], [28] propose to choose K using Bayesian modeling. However, the computational burden introduced by the complex modeling is prohibitive especially for large-scale data. Hence, we set $K = 40$ unless a more accurate estimate can be obtained from the domain knowledge. Lastly we discuss the choice of the step size η_t. From Section 3.2, a straightforward expression for η_t would be $\eta_t = a/(\tau t + b)$, where a and b are both positive numbers. In the initial phase where t is small, the step size approximately equals the constant a/b. Therefore the value of b determines the duration of this phase. Similar to τ, the choice of b is also data-dependent and lacks clear guidelines. As such, we fixed $b = 1\times10^4$ as we found this value gave us satisfactory results in practice. Moreover, we also set $a = 1\times10^4$. We will show that our algorithms are insensitive to the values of these parameters in Section S-4-C.

[4]For example, in [21], [23]–[25], the authors assume the objective function f in (3) can be decomposed into two parts, one being nonconvex but smooth and the other being nonsmooth but convex. However, such assumption does not cover our case.

[5]In such case, the subgradient should be defined as in the context of nonconvex analysis, e.g., see [26, Section 3.1].

TABLE S-1
DIVERGENCES IN CLASS $\overline{\mathcal{D}}$ AND THEIR CORRESPONDING NOISE GENERATION PROCEDURES

	Data Generation	Expressions of Distributions[a]	Parameter Value
IS	$\overline{v}_{ij} \sim \mathscr{G}(v_{ij}; \kappa, v_{ij}^o/\kappa)$	$\mathscr{G}(x;\kappa,\theta) \triangleq x^{\kappa-1}e^{-x/\theta}/(\theta^\kappa\Gamma(\kappa)), x \in \mathbb{R}_+$	$\kappa = 1000$
KL	$\overline{v}_{ij} \sim \mathscr{P}(v_{ij}; v_{ij}^o)$	$\mathscr{P}(k;\lambda) \triangleq \lambda^k e^{-\lambda}/k!,\ k \in \mathbb{N} \cup \{0\}$	—
Squared-ℓ_2	$\overline{v}_{ij} \sim \mathscr{N}(v_{ij}; v_{ij}^o, \varsigma^2)$	$\mathscr{N}(x;\mu,\varsigma^2) \triangleq 1/\sqrt{2\pi\varsigma^2}\exp\{-(x-\mu)^2/(2\varsigma^2)\}, x \in \mathbb{R}$	$\varsigma = 30$
Huber, ℓ_1, ℓ_2	$\overline{v}_{ij} \sim \mathscr{U}(v_{ij}; v_{ij}^o, \lambda), (i,j) \in \mathcal{Q}$	$\mathscr{U}(v_{ij}; v_{ij}^o, \lambda) \triangleq 1/(2\lambda), x \in [v_{ij}^o - \lambda, v_{ij}^o + \lambda]$	$\lambda = 2000$

[a] The function $\Gamma(\cdot)$ in the expressions of distributions denotes the Gamma function.

B. Synthetic Data Generation

To generate the (noisy) data matrix $\mathbf{V}$, we first generated the ground-truth data matrix $\mathbf{V}^o \triangleq \mathbf{W}^o\mathbf{H}^o$, where $\mathbf{W}^o \in \mathbb{R}_+^{F\times K^o}$ and $\mathbf{H}^o \in \mathbb{R}_+^{K^o\times N}$ denotes the ground-truth dictionary and coefficient matrices respectively. We set $F = 2\times 10^3$, $N = 1\times 10^5$ and $K^o = 40$.[6] The entries of $\mathbf{W}^o$ and $\mathbf{H}^o$ were generated i.i.d. from the shifted half-normal distribution $\mathscr{HN}_\varkappa(\sigma^2)$.[7] We set $\varkappa = 1$ to prevent entries of $\mathbf{W}^o$ and $\mathbf{H}^o$ from being arbitrarily small. Next we contaminated $\mathbf{V}^o$ with entrywise i.i.d. noise to obtain $\overline{\mathbf{V}}$. For the IS, KL and squared-ℓ_2 divergences, the distributions of the noise were chosen to be multiplicative Gamma, Poisson and additive Gaussian respectively, so that solving the (batch) NMF problem (1)is equivalent to the ML estimation of $\mathbf{V}^o$ from $\overline{\mathbf{V}}$ [29]. The parameters of these distributions were chosen such that the signal-to-noise ratio (SNR), $\mathrm{SNR} \triangleq 20\log_{10}(\|\mathbf{V}^o\|/\|\overline{\mathbf{V}} - \mathbf{V}^o\|)$ was approximately 30 dB. In particular, we chose $\sigma = 5$ to ensure the SNR for the Poisson noise satisfied the condition.[8] Since the other divergences considered (Huber, ℓ_1 and ℓ_2) are mainly used in the robust NMF, we added outliers to the ground-truth $\mathbf{V}^o$ as follows. We first randomly selected an index set $\mathcal{Q} \triangleq \Pi_{i\in[N]}\mathcal{Q}_i$ such that for any $i \in [N]$, $\mathcal{Q}_i \in [F]\times\{i\}$ and $|\mathcal{Q}| = 0.3F$. Then each entry v_{ij}^o with $(i,j) \in \mathcal{Q}$ was contaminated with (symmetric) uniform noise with magnitude λ. We chose $\lambda = 2\mathbb{E}[v_{ij}^o] = 2K^o(\varkappa + \sigma\sqrt{2/\pi})^2 \approx 2000$. The noise generation procedures for all the abovementioned divergences are summarized in Table S-1. The final data matrix $\mathbf{V}$ was obtained by projecting $\overline{\mathbf{V}}$ onto a compact set $\mathcal{V} \triangleq [0, 4000]^{F\times N}$ since $4000 \approx 4\mathbb{E}[v_{ij}^o]$.

C. Insensitivity to Key Parameters

To examine the sensitivity of OL to the key parameters τ, K and a, for each of the six divergences, we varied one parameter at each time in log-scale while keeping the other two fixed as in the canonical setting. From the plots of objective values versus time with different values of τ, K and a (shown in Figures S-1, S-2 and S-3 respectively), we observe that the convergence speeds of OL for all the divergences exhibit small (or even unnoticeable) variations across different values of τ, K and a. This shows the performance of our online algorithm is relatively insensitive to these key parameters. Therefore, in the following experiments on real data, we will use the canonical values of τ, K and a unless mentioned otherwise. Note that similarly to Figure 1, the results shown in Figures S-1, S-2 and S-3 are also relatively insensitive to different initializations of $\mathbf{W}$.

D. Additional results for Section 6.4

The topics learned from the `Guardian` dataset by OL-KL, B-KL and OL-Wang2 (ten representative words per topic) are shown in Table S-2. The topics learned from the `Wikipedia` dataset by these algorithms are shown in Table S-3. The average document clustering accuracies and running times of all the three algorithms (with standard deviations) on the `Wikipedia` dataset are shown in Table S-4. The results obtained from the `Wikipedia` dataset convey similar messages as those on the `Guardian` dataset.

E. Additional results for Section 6.5

The additional foreground-background separation results on the `Hall` dataset with four algorithmsOL-Huber, OL-Wang, B-Huber and OL-Guan are shown in Figure S-4. The foreground-background separation results on the `Escalator` dataset with these algorithms are shown in Figure S-5. The average running times (with standard deviations) of all the four algorithms on the `Escalator` dataset are shown in Table S-5. The results obtained from the `Escalator` dataset convey similar messages as those on the `Hall` dataset.

[6] Note that in general $K \neq K^o$, i.e., the latent dimension K given a priori in the algorithm may not match the ground-truth K^o.

[7] $\mathscr{HN}_\varkappa(\sigma^2)$ denotes the shifted half-normal distribution with scale parameter σ^2 and offset $\varkappa > 0$, i.e., $\mathscr{HN}_\varkappa(y;\sigma^2) = \frac{\sqrt{2}}{\sigma\sqrt{\pi}}\exp(-\frac{(y-\varkappa)^2}{2\sigma^2})$ for $y \geq \varkappa$ and 0 otherwise.

[8] By assuming the entries of $\mathbf{V}^o$ are i.i.d. and using the law of large numbers, all the distribution parameters (and σ) can be analytically estimated. See [27] for details.

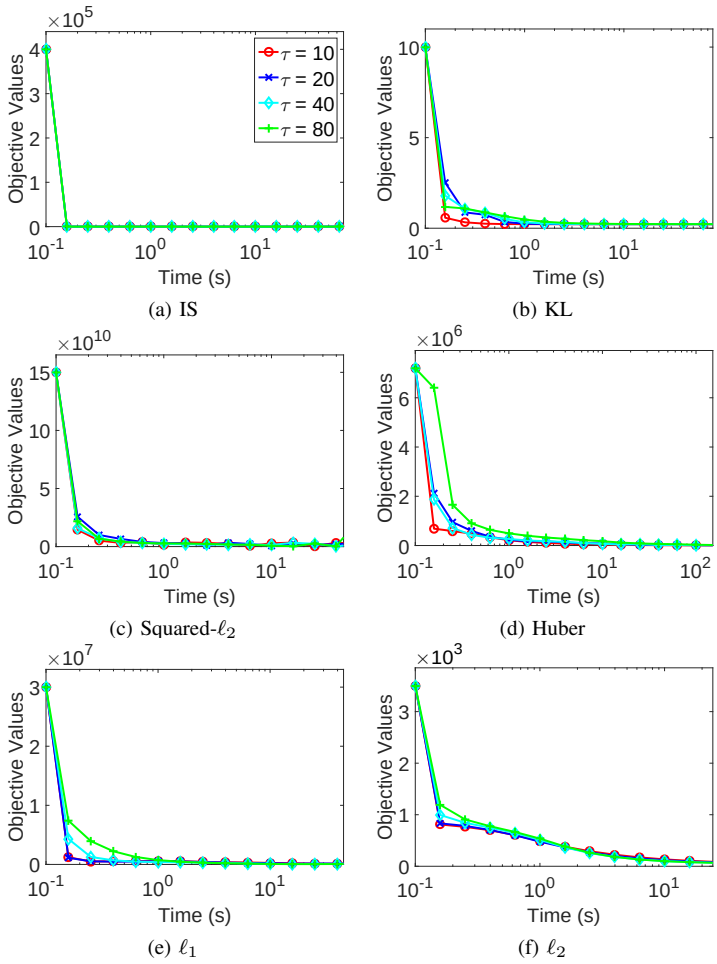

Fig. S-1. Objective values versus time (in seconds) of our online algorithms with different values of τ for all the six divergences. K and a are in the canonical setting.

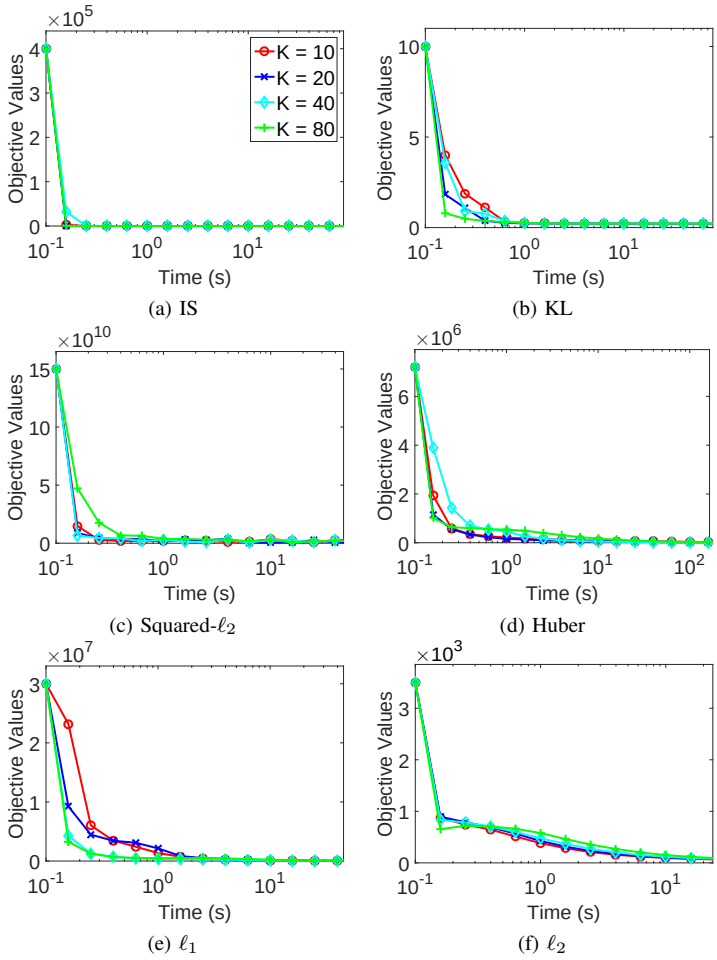

Fig. S-2. Objective values versus time (in seconds) of our online algorithms with different values of K for all the six divergences. τ and a are in the canonical setting.

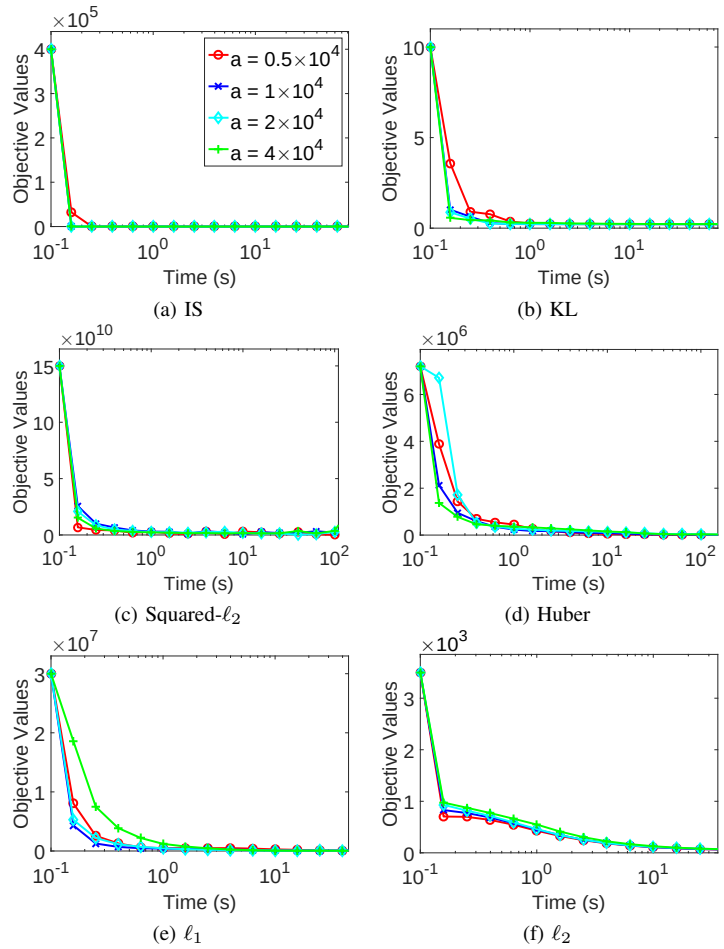

Fig. S-3. Objective values versus time (in seconds) of our online algorithms with different values of a for all the six divergences. τ and K are in the canonical setting.

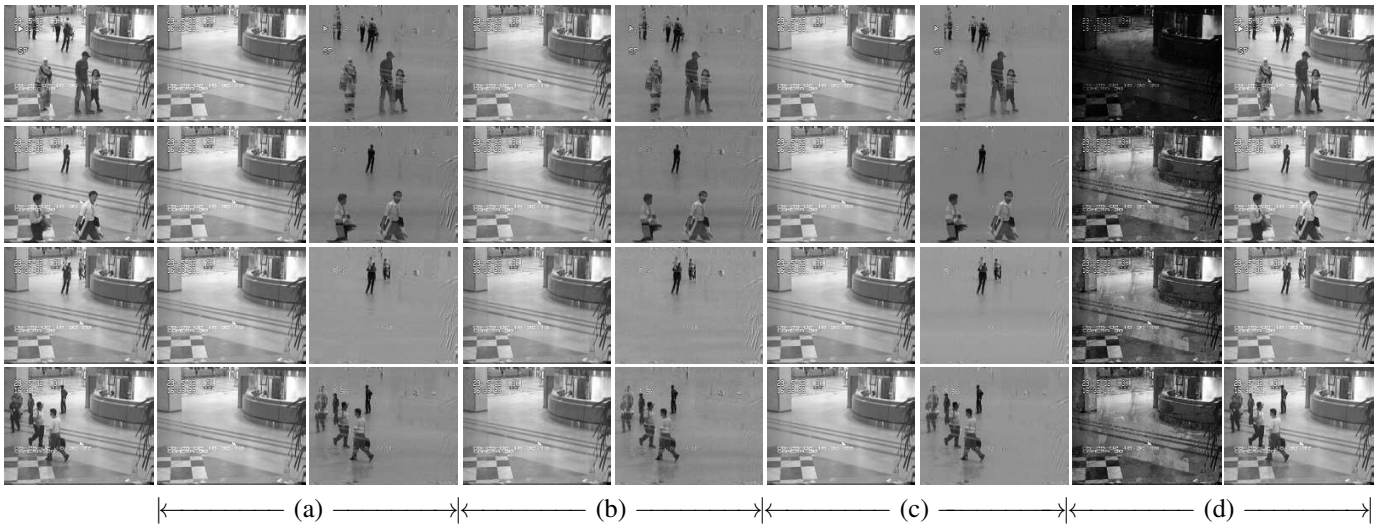

Fig. S-4. Additional foreground-background separation results on the `Hall` dataset with four algorithms: (a) OL-Huber, (b) OL-Wang, (c) B-Huber and (d) OL-Guan. The leftmost column shows the original video frames.

TABLE S-2
TOPICS LEARNED FROM THE Guardian DATASET BY THREE ALGORITHMS: OL-KL, B-KL AND OL-Wang2.

Business	Politics	Music	Fashion	Football
company	labour	music	fashion	league
sales	ultimately	album	wonder	club
market	party	band	weaves	universally
shares	government	songs	week	welsh
business	unions	vogue	war	team
group	bank	song	look	season
price	cameron	track	woolf	players
ultimately	voluntary	pop	wealthy	manager
growth	minister	workings	style	game
bank	workings	sound	hair	football

(a) OL-KL

Business	Politics	Music	Fashion	Football
bank	labour	music	fashion	league
company	party	album	wonder	club
ultimately	cameron	band	weaves	universally
growth	ultimately	vogue	week	team
market	unions	songs	look	welsh
business	voluntary	song	wealthy	season
sales	people	pop	war	players
government	minister	workings	woolf	manager
tax	war	rock	style	game
economy	miliband	sound	hair	football

(b) B-KL

Business	Politics	Music	Fashion	Football
bank	labour	music	fashion	league
growth	party	album	week	club
shares	unions	band	wonder	welsh
company	miliband	vogue	weaves	season
market	voluntary	songs	war	universally
sales	ultimately	song	wealthy	team
economy	cameron	pop	woolf	players
group	government	rock	look	manager
business	minister	sound	clothes	game
price	tory	singer	workings	winding

(c) OL-Wang2

TABLE S-3
TOPICS LEARNED FROM THE Wikipedia DATASET BY THREE ALGORITHMS: OL-KL, B-KL AND OL-Wang2.

Music	Military	Space	Medicine	Sports	Transportation
album	raids	orbited	patricia	player	stationary
songwriters	navigation	stanford	treating	racer	rains
muse	army	plains	tumors	season	trainer
recordings	unit	observatory	disease	teaching	line
band	motors	sunglasses	syndromes	winner	services
releasing	battle	planners	protective	omar	route
singles	spurs	earth	thermonuclear	league	warehouses
vienna	shines	moons	systematic	scores	shipyard
performs	air	mars	problem	stafford	canal
player	operates	sold	rna	championships	pass

(a) OL-KL

Music	Military	Space	Medicine	Sports	Transportation
songwriters	mission	space	patricia	player	stationary
album	warehouses	orbited	treating	stanford	rains
recordings	unit	plains	cells	season	services
muse	army	moons	disease	teaching	shines
releasing	air	observatory	syndromes	omar	trainer
band	spurs	stafford	specifically	scores	line
singles	aircraft	earth	blood	league	shipyard
vienna	battle	sold	muse	rotten	operates
performs	offerings	planners	thermonuclear	game	pass
tough	navigation	racer	symphony	games	route

(b) B-KL

Music	Military	Space	Medicine	Sports	Transportation
album	racer	orbited	patricia	player	stationary
songwriters	navigation	stanford	treating	season	rains
band	warehouses	plains	symphony	teaching	line
releasing	shipyard	space	disease	league	trainer
recordings	shines	planners	syndromes	game	services
muse	unit	observatory	muse	scores	raids
singles	spurs	earth	cells	games	pass
vienna	winner	sunglasses	blood	tourists	route
chart	aircraft	moons	thermonuclear	winner	london
tracks	army	mars	surgically	cup	platforms

(c) OL-Wang2

TABLE S-4
CLUSTERING ACCURACIES AND RUNNING TIMES OF OL-KL, B-KL AND OL-Wang2 ON THE Wikipedia DATASET.

Algorithms	Accuracy	Time (s)
OL-KL	0.712 ± 0.01	53.18 ± 0.51
B-KL	0.716 ± 0.01	303.75 ± 0.61
OL-Wang2	0.656 ± 0.04	55.08 ± 0.41

TABLE S-5
RUNNING TIMES OF OL-Huber, OL-Wang, B-Huber AND OL-Guan ON THE Escalator DATASET.

Algorithms	Time (s)	Algorithms	Time (s)
OL-Huber	63.35 ± 0.74	OL-Wang	71.73 ± 0.97
B-Huber	375.22 ± 2.48	OL-Guan	127.12 ± 1.35

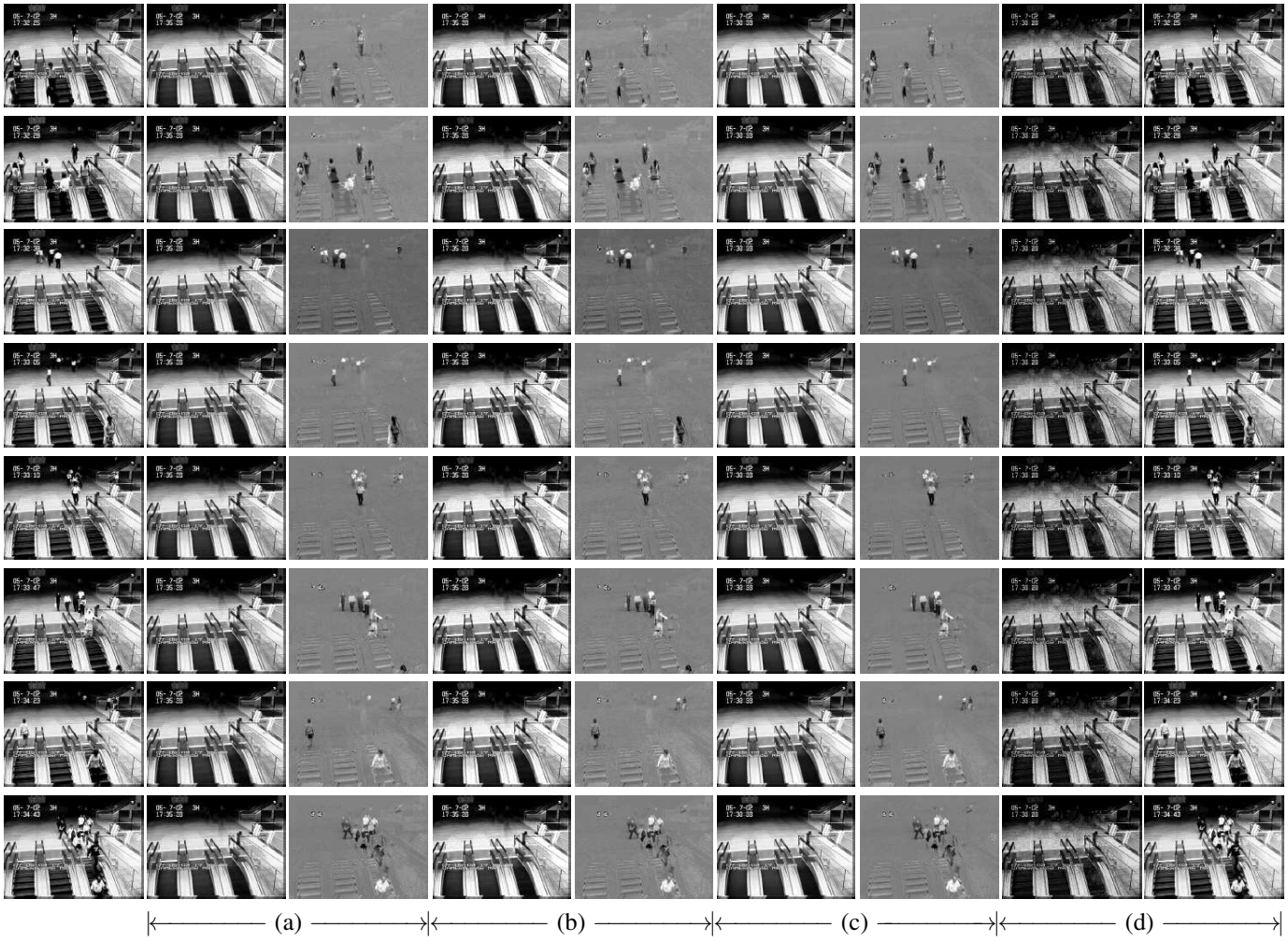

Fig. S-5. Foreground-background separation results on the `Escalator` dataset with four algorithms: (a) OL-Huber, (b) OL-Wang, (c) B-Huber and (d) OL-Guan. The leftmost column shows the original video frames.

S-5. Technical Lemmas

A. Convergence of PGD and MM algorithms

Lemma S-1 (Adapted from [16, Theorem 1])**.** *Given a real Hilbert space $\mathcal{Y}$ and a function $f : \mathcal{Y} \to \mathbb{R}$, consider the following optimization problem*

$$\min_{x\in\mathcal{X}} f(x), \tag{S-11}$$

where $\mathcal{X} \subseteq \mathcal{Y}$ is nonempty, closed and convex and f is differentiable on $\mathcal{X}$. For any $x \in \mathcal{X}$, define a differentiable function $u(x,\cdot) : \mathcal{X} \to \mathbb{R}$ such that $u(x,\cdot)$ is a majorant for f at x.[9] Fix an arbitrary initial point $x_0 \in \mathcal{X}$ and consider the sequence of iterates $\{x^k\}_{k\in\mathbb{N}}$ generated by the following MM algorithm

$$x^k := \min_{y\in\mathcal{X}} u(x^{k-1}, y),\ \forall\, k \in \mathbb{N}. \tag{S-12}$$

Then $\{x^k\}_{k\in\mathbb{N}}$ has at least one limit point and moreover, the any limit point of $\{x^k\}_{k\in\mathbb{N}}$ is a stationary point of (S-11)*.*

Lemma S-2 (Adapted from [13, Theorem 2.4])**.** *Consider a real Hilbert space $\mathcal{Y}$. Let $\mathcal{X} \subseteq \mathcal{Y}$ be a nonempty compact convex set and $f : \mathcal{Y} \to \mathbb{R}$ be continuously differentiable on $\mathcal{Y}$. Fix an arbitrary initial point $x_0 \in \mathcal{X}$ and consider the sequence of iterates $\{x^k\}_{k\in\mathbb{N}}$ generated by the following projected gradient algorithm*

$$x^k := \Pi_{\mathcal{X}}\Big\{x^{k-1} - \beta^k \nabla f(x^{k-1})\Big\},\ \forall\, k \in \mathbb{N}, \tag{S-13}$$

where the sequence of step sizes $\{\beta^k\}_{k\in\mathbb{N}}$ is chosen according to the Armijo rule [30]. Then $\{x^k\}_{k\in\mathbb{N}}$ has at least one limit point and moreover, the any limit point[10] of $\{x^k\}_{k\in\mathbb{N}}$ is a stationary point of the optimization problem $\min_{x\in\mathcal{X}} f(x)$.

B. Optimal-value functions

Lemma S-3 (The Maximum Theorem; [31, Theorem 14.2.1 & Example 2])**.** *Let $\mathcal{P}$ and $\mathcal{X}$ be two metric spaces. Consider a maximization problem*

$$\max_{x\in B(p)} f(p, x), \tag{S-14}$$

where $B : \mathcal{P} \rightrightarrows \mathcal{X}$ 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-14) *is replaced by minimization.*

Lemma S-4 (Danskin's Theorem; [32, 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 (Hadamard) directionally differentiable and its directional derivative along $d \in \mathcal{U}$, $v'(u,d)$ is given by

$$v'(u,d) = \min_{x\in S(u)} \langle \nabla_u f(x,u), d\rangle\,. \tag{S-15}$$

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

Lemma S-5 (Minimization of convex functions; [20, Section 3.2.5], [33])**.** *Let $\mathcal{X}$ and $\mathcal{Y}$ be two inner product spaces and $\mathcal{X} \times \mathcal{Y}$ be their product space such that for any (x,y) and (x',y') in $\mathcal{X} \times \mathcal{Y}$, $\langle (x,y),(x',y')\rangle = \langle x, x'\rangle + \langle y, y'\rangle$. Consider functions $h : \mathcal{X} \times \mathcal{Y} \to \mathbb{R}$ and $f : \mathcal{X} \to \mathbb{R}$ such that*

$$f(x) \triangleq \inf_{y\in\mathcal{Y}} h(x,y),\ \forall\, x \in \mathcal{X}. \tag{S-16}$$

If h is convex on $\mathcal{X} \times \mathcal{Y}$ and $\mathcal{Y}$ is convex, then f is convex on $\mathcal{X}$. If we further assume $\mathcal{S}(x_0) \triangleq \arg\min_{y\in\mathcal{Y}} h(x_0, y) \neq \emptyset$, then the subdifferential of f at $x_0 \in \mathcal{X}$,

$$\partial f(x_0) = \bigcup_{y_0\in\mathcal{S}(x_0)} \{g \in \mathcal{X} \,|\, (g,g') \in \partial h(x_0,y_0), \text{ where } \langle g', y - y_0\rangle = 0, \forall\, y \in \mathcal{Y}\}\,. \tag{S-17}$$

[9] By this, we mean $u(x,x) = f(x)$ and $u(x,y) \geq f(y)$ for any $y \in \mathcal{X}$.

[10] The limit point is defined in the topological sense, i.e., $\overline{x} \in \mathcal{X}$ is a limit point of $\{x^k\}_{k\in\mathbb{N}}$ if for any neighborhood $\mathcal{U}$ of $\overline{x}$, there are infinitely many elements of $\{x^k\}_{k\in\mathbb{N}}$ in $\mathcal{U}$.

Proof.

$$\begin{aligned} g \in f(x_0) &\iff f(x) \geq f(x_0) + \langle g, x - x_0 \rangle, \forall\, x \in \mathcal{X} \\ &\iff h(x,y) \geq h(x_0,y_0) + \langle (g,g'), (x-x_0, y'-y_0) \rangle, \forall\, x \in \mathcal{X}, \forall\, y \in \mathcal{S}(x), \forall\, y_0 \in \mathcal{S}(x_0), \forall\, y' \in \mathcal{Y}, \\ &\qquad\qquad \forall\, g' \in \mathcal{X} \text{ s.t. } \langle g', y' - y_0 \rangle = 0 \\ &\iff h(x,y) \geq h(x_0,y_0) + \langle (g,g'), (x-x_0, y-y_0) \rangle, \forall\, x \in \mathcal{X}, \forall\, y \in \mathcal{Y}, \forall\, y_0 \in \mathcal{S}(x_0), \forall\, g' \text{ s.t. } \langle g', y - y_0 \rangle = 0 \\ &\iff (g,g') \in \partial h(x_0,y_0), \forall\, y_0 \in \mathcal{S}(x_0), \forall\, g' \text{ s.t. } \langle g', y - y_0 \rangle = 0, \forall\, y \in \mathcal{Y}. \end{aligned}$$

□

C. Miscellaneous

Lemma S-6 (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 on $\mathcal{X}$ and for each $x \in \mathcal{X}$,

$$\nabla_x \int_\Omega f(x,\omega)\, d\mu(\omega) = \int_\Omega \nabla_x f(x,\omega)\, d\mu(\omega). \tag{S-18}$$

Remark S-1. This is a simplified version of the Leibniz Integral Rule. See [34, Theorem 16.8] for weaker conditions on f.

Lemma S-7 (Almost sure convergence of square-integrable martingales; [35, Theorem 5.4.9])**.** *Let $\{X_n\}_{n\geq 1}$ be a martingale in a normed space $\mathcal{X}$ adapted to the filtration $\{\mathscr{F}_n\}_{n\geq 0}$ such that $\sup_{n\in\mathbb{N}} \mathbb{E}\left[\|X_n\|^2\right] < \infty$. Define the* quadratic variation process *$\{\langle X\rangle_n\}_{n\geq 2}$ as*

$$\langle X\rangle_n \triangleq \sum_{i=2}^{n} \mathbb{E}\left[\|X_i - X_{i-1}\|^2 \,|\mathscr{F}_{i-1}\right],\ \forall\, n \geq 2. \tag{S-19}$$

Then there exists a random variable X such that on the set $\{\lim_{n\to\infty}\langle X\rangle_n < \infty\}$, the sequence $\{X_n\}_{n\geq 1}$ converges a.s. to X and $\|X\| < \infty$ a.s..

Lemma S-8 (Expectation of convex functions; [33])**.** *Let $(\mathcal{U}, \mathscr{A}, \nu)$ be a probability space and $h : \mathcal{X} \times \mathcal{U}$ be a function such that for each $u \in \mathcal{U}$, $x \mapsto h(x,u)$ is convex on $\mathcal{X}$, where $\mathcal{X}$ is a convex set equipped with an inner product $\langle\cdot,\cdot\rangle$. Define $f(x) \triangleq \mathbb{E}_u(x,u)$, for any $x \in \mathcal{X}$. Then f is convex on $\mathcal{X}$. Fix any $x_0 \in \mathcal{X}$. Then for any $g_{x_0}(u) \in \partial_x h(x_0,u)$, $\mathbb{E}_u[g_{x_0}(u)] \in \partial f(x_0)$.*

D. Asymptotic Equicontinuity and Uniform Convergence

In this section, unless otherwise mentioned, we assume the sequences of functions $\{f_n\}_{n\in\mathbb{N}}$ and $\{g_n\}_{n\in\mathbb{N}}$ are defined on a common metric space $(\mathcal{X}, d)$ and mapped to a common metric space $(\mathcal{Y}, \rho)$.

Lemma S-9 (Uniform convergence implies asymptotic equicontinuity)**.** *If a sequence of functions $\{f_n\}_{n\in\mathbb{N}}$ converges uniformly to a continuous function f on $\mathcal{X}$, then it is asymptotically equicontinuous on $\mathcal{X}$.*

Proof. Fix $\epsilon > 0$. Since $f_n \xrightarrow{\text{u}} f$, there exists $N \in \mathbb{N}$ such that for all $n \geq N$, $\sup_{x\in\mathcal{X}} |f_n(x) - f(x)| < \epsilon/6$. Fix $x_0 \in \mathcal{X}$. Then there exists $\delta > 0$ such that $\sup_{x'\in\mathcal{N}_\delta(x)} \rho(f(x), f(x')) < \epsilon/6$, where $\mathcal{N}_\delta(x) \triangleq \{x' \in \mathcal{X} : d(x,x') < \delta\}$. Thus for all $n \geq N$, $\sup_{x'\in\mathcal{N}_\delta(x)} \rho(f_n(x), f_n(x')) \leq \rho(f_n(x), f(x)) + \sup_{x'\in\mathcal{N}_\delta(x)} \rho(f(x), f(x')) + \sup_{x'\in\mathcal{N}_\delta(x)} \rho(f_n(x'), f(x')) < \epsilon/2$. This shows $\limsup_{n\to\infty}\ \sup_{x'\in\mathcal{N}_\delta(x)} \rho(f_n(x), f_n(x')) < \epsilon$. Since this holds for all $x \in \mathcal{X}$, we complete the proof. □

Lemma S-10 (Lipschitzness implies equicontinuity)**.** *Given a sequence of continuous functions $\{f_n\}_{n\in\mathbb{N}}$. If each f_n is Lipschitz on $\mathcal{X}$ with Lipschitz constant L_n and there exists $M \in (0,\infty)$ such that $\sup_{n\geq 1} L_n \leq M$, then $\{f_n\}_{n\in\mathbb{N}}$ is equicontinuous on $\mathcal{X}$.*

Lemma S-11 (Finite sum preserves asymptotic equicontinuity)**.** *Let $\{f_n\}_{n\in\mathbb{N}}$ and $\{g_n\}_{n\in\mathbb{N}}$ be both asymptotically equicontinuous on $\mathcal{X}$. Assume the metric ρ is translation-invariant (for e.g., induced by a norm). Then $\{f_n + g_n\}_{n\in\mathbb{N}}$ is asymptotically equicontinuous on $\mathcal{X}$.*

Proof. Fix an $\epsilon > 0$ and $x \in \mathcal{X}$, there exist $\delta_1 > 0$ and $\delta_2 > 0$ respectively such that

$$\begin{aligned} &\limsup_{n\to\infty} \sup_{x'\in\mathcal{X}: d(x,x')<\delta_1} \rho(f_n(x), f_n(x')) < \epsilon/2, \\ &\limsup_{n\to\infty} \sup_{x'\in\mathcal{X}: d(x,x')<\delta_2} \rho(g_n(x), g_n(x')) < \epsilon/2. \end{aligned}$$

Take $\delta = \min(\delta_1, \delta_2)$, we have

$$\begin{aligned}
&\limsup_{n\to\infty} \sup_{x'\in\mathcal{X}:d(x,x')<\delta} \rho((f_n+g_n)(x),(f_n+g_n)(x')) \\
\leq\ &\limsup_{n\to\infty} \sup_{x'\in\mathcal{X}:d(x,x')<\delta} \rho((f_n+g_n)(x),f_n(x')+g_n(x)) + \rho(f_n(x')+g_n(x),f_n(x')+g_n(x')) \\
\leq\ &\limsup_{n\to\infty} \sup_{x'\in\mathcal{X}:d(x,x')<\delta} \rho(f_n(x),f_n(x')) + \rho(g_n(x),g_n(x')) \\
<\ &\epsilon.
\end{aligned}$$

□

Lemma S-12 (Continuous transformation preserves uniform convergence)**.** *Assume $\mathcal{X}$ to be compact. Let $g : \mathcal{Y} \to \mathcal{Z}$ be a continuous function, where $(\mathcal{Z}, r)$ is a metric space. If $\{f_n\}_{n\in\mathbb{N}}$ uniformly converges to a continuous function f on $\mathcal{X}$, then $\{g \circ f_n\}_{n\in\mathbb{N}}$ uniformly converges to $g \circ f$ on $\mathcal{X}$.*

Proof. First, since $\mathcal{X}$ is compact and f is continuous on $\mathcal{X}$, $f(\mathcal{X})$ is compact in $\mathcal{Y}$. Since g is continuous on $\mathcal{Y}$, g is uniformly continuous on $f(\mathcal{X})$. Fix $\epsilon > 0$. there exists a $\delta > 0$ such that for all $y, y' \in f(\mathcal{X})$ and $\rho(y, y') < \delta$, $r(g(y), g(y')) < \epsilon$. Since $f_n \xrightarrow{\text{u}} f$ on $\mathcal{X}$, there exits a $K \in \mathbb{N}$ such that for all $n \geq K$ and $x \in \mathcal{X}$, $\rho(f(x), f_n(x)) < \delta$. Consequently, $r(g(f_n(x)), g(f(x))) < \epsilon$. This implies $g \circ f_n \xrightarrow{\text{u}} g \circ f$ on $\mathcal{X}$. □

Lemma S-13 (Generalized Arzelà-Ascoli Theorem [36])**.** *If the sequence of functions $\{f_n\}_{n\geq 1}$ is asymptotically equicontinuous and uniformly bounded on $\mathcal{X}$ (assumed to be compact), then there exists a subsequence $\{f_{n_k}\}_{k\geq 1}$ that converges uniformly to a continuous function f on $\mathcal{X}$.*

E. Projected Dynamical Systems and Lyapunov Stability Theory

Lemma S-14 (Adapted from [37, Theorem 3.1, Chapter 4])**.** *Assume* (S-6) *holds with $\overline{Z}(\omega, 0) = 0$ and $\overline{W}(\omega, s) \in \mathcal{C}$, for all $s \geq 0$. Denote λ as the Lebesgue measure on $\mathbb{R}$. If $\overline{W}(\omega, \cdot)$ is Lipschitz on $\mathbb{R}_+$ and for any $\tau > 0$,*

1) $\overline{Z}(\tau) = \mathbf{0}$ if $\overline{W}(\omega, s) \in \mathbf{int}\,\mathcal{C}$ for all $s \in \mathcal{T}$, where $\mathcal{T}$ is any set in $[0, \tau]$ with $\lambda(\mathcal{T}) = \tau$,

2) $\overline{Z}(\tau) \in \overline{\mathbf{conv}}\left[\bigcup_{s\in[0,\tau]} \mathcal{N}\left(\overline{W}(\omega, s)\right)\right]$,

where $\mathcal{N}$ is defined in (S-7)*, then*

$$\overline{Z}(s) = \int_0^s z(\tau)\, d\tau, \tag{S-20}$$

where $z : \mathbb{R}_+ \to \mathbb{R}^{F\times K}$ is defined in (S-4)*.*

Definition S-1 (Lyapunov function and its Lie derivative; [19, Section 6.6])**.** Consider the PDS given in (6). Assume the normed space $(\mathcal{X}, \|\cdot\|)$ is equipped with the inner product $\langle\cdot,\cdot\rangle$. Fix $x_0 \in \mathcal{K}$ and choose a neighborhood of x_0 in $\mathcal{K}$, denoted as $\mathcal{U}(x_0)$. A continuously differentiable function $L : \mathcal{U}(x_0) \to \mathbb{R}_+$ is called a Lyapunov function if $L(x_0) = 0$, $L(x) > 0$ for any $x \in \mathcal{U}(x_0) \setminus \{x_0\}$ and for any $x(\cdot) \in \mathcal{P}(g, \mathcal{K}, x_0)$,

$$L(x(t_1)) \leq L(x(t_0)),\ \forall\, t_0, t_1 \in \mathcal{I}, t_0 < t_1,\ \text{s.t. } \{x(t_0), x(t_1)\} \subseteq \mathcal{U}(x_0) \setminus \{x_0\}. \tag{S-21}$$

Moreover, for any $x(\cdot) \in \mathcal{P}(g, \mathcal{K}, x_0)$, the Lie derivative of L on $\mathcal{I}$, $\frac{d}{ds}L(x(s))$ is given by

$$\frac{d}{ds}L(x(s)) = \langle \nabla_x L(x(s)), x'(s)\rangle\,,\ \forall\, s \in \mathcal{I}. \tag{S-22}$$

Lemma S-15 (All limit points are stationary; [19, Theorem 6.15])**.** *Consider the PDS given in* (6)*. Let $L : \mathcal{U} \subseteq \mathcal{K} \to \mathbb{R}_+$ be a Lyapunov function with possibly non-unique zeros (i.e., L may only be positive semidefinite on $\mathcal{U}$). Suppose each solution $x(\cdot) \in \mathcal{P}(g, \mathcal{K}, x_0)$ is contained in $\mathcal{U}$, then L is constant on $\mathcal{L}(g, \mathcal{K}, x_0) \cap \mathcal{U}$. In other words, the Lie derivative of L vanishes on $\mathcal{L}(g, \mathcal{K}, x_0) \cap \mathcal{U}$.*

F. Correspondence and Upper Semicontinuity

For further details, see [38, Chapter 1].

Definition S-2 (Correspondence and its graph)**.** Given two metric spaces $(\mathcal{X}, d)$ and $(\mathcal{Y}, \rho)$, a correspondence $\mathcal{F} : \mathcal{X} \rightrightarrows \mathcal{Y}$ maps esch $x \in \mathcal{X}$ to a subset $\mathcal{F}(x)$ in $\mathcal{Y}$. The graph of $\mathcal{F}$, $\mathcal{G}(\mathcal{F})$ is defined as

$$\mathcal{G}(\mathcal{F}) \triangleq \{(x, y) \in \mathcal{X} \times \mathcal{Y} \,|\, y \in \mathcal{F}(y)\}\,. \tag{S-23}$$

Definition S-3 (Upper semicontinuous correspondence)**.** A correspondence $\mathcal{F}$ as defined in Definition S-2 is called upper semicontinuous at $x_0 \in \mathcal{X}$ if for any open set $\mathcal{U} \subseteq \mathcal{Y}$ such that $\mathcal{F}(x_0) \subseteq \mathcal{U}$, there exists an open set $\mathcal{V} \subseteq \mathcal{X}$ such that $x_0 \in \mathcal{V}$ and $\mathcal{F}(x) \subseteq \mathcal{U}$ for any $x \in \mathcal{V}$.

Lemma S-16 (Closed graph property; [38, Proposition 2 & Colrollary 1])**.** *Let the correspondence $\mathcal{F}$ be given in Definition S-2.*

1) *If for each $x \in \mathcal{X}$, $\mathcal{F}(x)$ is closed, and $\mathcal{F}$ is upper semicontinuous, then $\mathcal{G}(\mathcal{F})$ is closed (in $\mathcal{X} \times \mathcal{Y}$).*
2) *If $(\mathcal{Y}, \rho)$ is a compact metric space, and $\mathcal{G}(\mathcal{F})$ is closed, then $\mathcal{F}$ is upper semicontinuous.*

Lemma S-17 (Sufficient conditions for upper semicontinuity; [38, Section 1.1])**.** *Let the correspondence $\mathcal{F}$ be given in Definition S-2. Assume $\mathcal{F}$ is compact-valued on $\mathcal{X}$. Fix $x \in \mathcal{X}$. If $\mathcal{F}$ satisfies*

$$\bigcap_{\delta>0} \overline{\mathbf{conv}} \left(\bigcup_{z \in \mathcal{B}_\delta(x)} \mathcal{F}(z) \right) = \mathcal{F}(x), \tag{S-24}$$

then $\mathcal{F}$ is upper semicontinuous at x. Here $\overline{\mathbf{conv}}\,\mathcal{S}$ denotes the closed convex hull of a set $\mathcal{S}$ and $\mathcal{B}_\delta(x) \triangleq \{z \in \mathcal{X} \mid d(x,z) < \delta\}$.

REFERENCES

[1] J. Mairal, F. Bach, J. Ponce, and G. Sapiro, "Online learning for matrix factorization and sparse coding," *J. Mach. Learn. Res.*, vol. 11, pp. 19–60, Mar 2010.
[2] N. Guan, D. Tao, Z. Luo, and B. Yuan, "Online nonnegative matrix factorization with robust stochastic approximation," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 23, no. 7, pp. 1087–1099, Jul. 2012.
[3] J. Feng, H. Xu, and S. Yan, "Online robust PCA via stochastic optimization," in *Proc. NIPS*, Lake Tahoe, USA, Dec. 2013, pp. 404–412.
[4] B. Shen, B. Liu, Q. Wang, and R. Ji, "Robust nonnegative matrix factorization via l_1 norm regularization by multiplicative updating rules," in *Proc. ICIP*, Paris, France, Oct. 2014, pp. 5282–5286.
[5] J. Shen, P. Li, and H. Xu, "Online low-rank subspace clustering by explicit basis modeling," in *Proc. ICML*, New York, USA, Jul. 2016, pp. 1–10.
[6] A. Lefèvre, F. Bach, and C. Févotte, "Online algorithms for nonnegative matrix factorization with the itakura-saito divergence," in *Proc. WASPAA*, New Paltz, New York, USA, Oct 2011, pp. 313–316.
[7] A. Dessein, A. Cont, and G. Lemaitre, "Real-time detection of overlapping sound events with non-negative matrix factorization," *Matrix Information Geometry*, pp. 341–371, 2011.
[8] N. Wang, J. Wang, and D.-Y. Yeung, "Online robust non-negative dictionary learning for visual tracking," in *Proc. ICCV*, Sydney, Australia, Dec. 2013, pp. 657–664.
[9] X. Zhang, N. Guan, D. Tao, X. Qiu, and Z. Luo, "Online multi-modal robust non-negative dictionary learning for visual tracking," *PLoS ONE*, vol. 10, no. 5, pp. 1–17, 2015.
[10] J. Chen, Z. J. Towfic, and A. H. Sayed, "Dictionary learning over distributed models," *IEEE Trans. Signal Process.*, vol. 63, no. 4, pp. 1001–1016, 2015.
[11] A. Nemirovski, A. Juditsky, G. Lan, and A. Shapiro, "Robust stochastic approximation approach to stochastic programming," *SIAM J. Optimiz.*, vol. 19, no. 4, pp. 1574–1609, Jan. 2009.
[12] J. Duchi, S. Shalev-Shwartz, Y. Singer, and T. Chandra, "Efficient projections onto the l1-ball for learning in high dimensions," in *Proc. ICML*, Helsinki, Finland, Jul. 2008, pp. 272–279.
[13] P. H. Calamai and J. J. Moré, "Projected gradient methods for linearly constrained problems," *Math. Program.*, vol. 39, no. 1, pp. 93–116, 1987.
[14] C.-J. Lin, "On the convergence of multiplicative update algorithms for nonnegative matrix factorization," *IEEE Trans. Neural Netw.*, vol. 18, no. 6, pp. 1589–1596, 2007.
[15] N. Parikh and S. Boyd, "Proximal algorithms," *Found. Trends Optim.*, vol. 1, no. 3, pp. 123–231, 2014.
[16] M. Razaviyayn, M. Hong, and Z.-Q. Luo, "A unified convergence analysis of block successive minimization methods for nonsmooth optimization," *SIAM J. Optim.*, vol. 23, no. 2, pp. 1126–1153, 2013.
[17] A. Nedić, "Subgradient projection method," http://www.ifp.illinois.edu/~angelia/sgd_notes.pdf, 2008.
[18] D. P. Bertsekas, *Nonlinear Programming*. Athena Scitific, 1999.
[19] G. Teschl, *Ordinary Differential Equations and Dynamical Systems*. Amer. Math. Soc., 2012.
[20] S. Boyd and L. Vandenberghe, *Convex Optimization*. Cambridge University Press, 2004.
[21] M. Razaviyayn, M. Sanjabi, and Z.-Q. Luo, "A stochastic successive minimization method for nonsmooth nonconvex optimization with applications to transceiver design in wireless communication networks," *Math. Program.*, vol. 157, no. 2, pp. 515–545, 2016.
[22] P. Dupuis and A. Nagurney, "Dynamical systems and variational inequalities," *Ann. Oper. Res.*, vol. 44, no. 1, pp. 7–42, 1993.
[23] S. Ghadimi, G. Lan, and H. Zhang, "Mini-batch stochastic approximation methods for nonconvex stochastic composite optimization," *Math. Program.*, vol. 155, no. 1-2, pp. 267–305, Jan. 2016.
[24] S. Ghadimi and G. Lan, "Accelerated gradient methods for nonconvex nonlinear and stochastic programming," *Math. Program.*, vol. 156, no. 1, pp. 59–99, 2016.
[25] S. J. Reddi, S. Sra, B. Poczos, and A. Smola, "Fast stochastic methods for nonsmooth nonconvex optimization," in *Proc. NIPS*, Barcelona, Spain, Dec. 2016.
[26] C. Bao, H. Ji, Y. Quan, and Z. Shen, "Dictionary learning for sparse coding: Algorithms and analysis," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 38, no. 7, pp. 1356–1369, July 2015.
[27] V. Y. F. Tan and C. Févotte, "Automatic relevance determination in nonnegative matrix factorization with the β-divergence," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 35, no. 7, pp. 1592–1605, Jul. 2013.
[28] C. M. Bishop, "Bayesian PCA," in *Proc. NIPS*, Denver, USA, Jan. 1999, pp. 382–388.
[29] C. Févotte, N. Bertin, and J.-L. Durrieu, "Nonnegative matrix factorization with the Itakura-Saito divergence. With application to music analysis," *Neural Comput.*, vol. 21, no. 3, pp. 793–830, Mar. 2009.
[30] L. Armijo, "Minimization of functions having lipschitz continuous first partial derivatives," *Pacific J. Math.*, vol. 16, no. 1, pp. 1–3, 1966.
[31] K. Sydsaeter, P. Hammond, A. Seierstad, and A. Strom, *Further Mathematics for Economic Analysis*. Pearson, 2005.
[32] J. F. Bonnans and A. Shapiro, "Optimization problems with perturbations: A guided tour," *SIAM Rev.*, vol. 40, no. 2, pp. 228–264, 1998.
[33] S. Boyd, J. Duchi, and L. Vandenberghe, "Subgradients," http://web.stanford.edu/class/ee364b/lectures/subgradients_notes.pdf, 2015.
[34] P. Billingsley, *Probability and Measure*, 2nd ed. John Wiley & Sons, 1986.
[35] R. Durrett, *Probability: theory and examples*. Duxbury Press, 2013.
[36] G. G. Yin and Q. Zhang, *Discrete-Time Markov Chains: Two-Time-Scale Methods and Applications*. Springer, 2005.
[37] H. J. Kushner and G. Yin, *Stochastic approximation and recursive algorithms and applications*. Springer, 2003.
[38] J.-P. Aubin and A. Cellina, *Differential inclusions: set-valued maps and viability theory*. Springer-Verlag, 1984.