Online Nonnegative Matrix Factorization with General Divergences

Vincent Y. F. Tan (ECE, Mathematics, NUS)

Joint work with Renbo Zhao (NUS) and Huan Xu (GeorgiaTech)

IWCT, Shanghai Jiaotong University

December 13, 2016

Nonnegative Matrix Factorization (NMF)

- Given a data matrix $\mathbf{V} \geq 0$, find basis matrix $\mathbf{W} \geq 0$ and a nonnegative coefficient matrix $\mathbf{H} \geq 0$ such that

$$\mathbf{V} \approx \mathbf{WH},$$

by solving a nonconvex problem

$$\min_{\mathbf{W} \geq 0, \mathbf{H} \geq 0} \left[D(\mathbf{V} \| \mathbf{WH}) \triangleq \sum_{n=1}^{N} d(\mathbf{v}_n \| \mathbf{W}\mathbf{h}_n) \right]$$

Nonnegative Matrix Factorization (NMF)

- Given a data matrix $\mathbf{V} \geq 0$, find basis matrix $\mathbf{W} \geq 0$ and a nonnegative coefficient matrix $\mathbf{H} \geq 0$ such that

$$\mathbf{V} \approx \mathbf{WH},$$

by solving a nonconvex problem

$$\min_{\mathbf{W} \geq 0, \mathbf{H} \geq 0} \left[D(\mathbf{V} \| \mathbf{WH}) \triangleq \sum_{n=1}^{N} d(\mathbf{v}_n \| \mathbf{W}\mathbf{h}_n) \right]$$

- The non-subtractive, parts-based basis representation makes it attractive for data analysis.

Nonnegative Matrix Factorization (NMF)

- NMF provides an unsupervised linear representation of data

$$\mathbf{V} \approx \mathbf{WH}$$

- Other methods include
  - Principal component analysis (PCA)
  - Independent component analysis (ICA)

Nonnegative Matrix Factorization (NMF)

- NMF provides an unsupervised linear representation of data

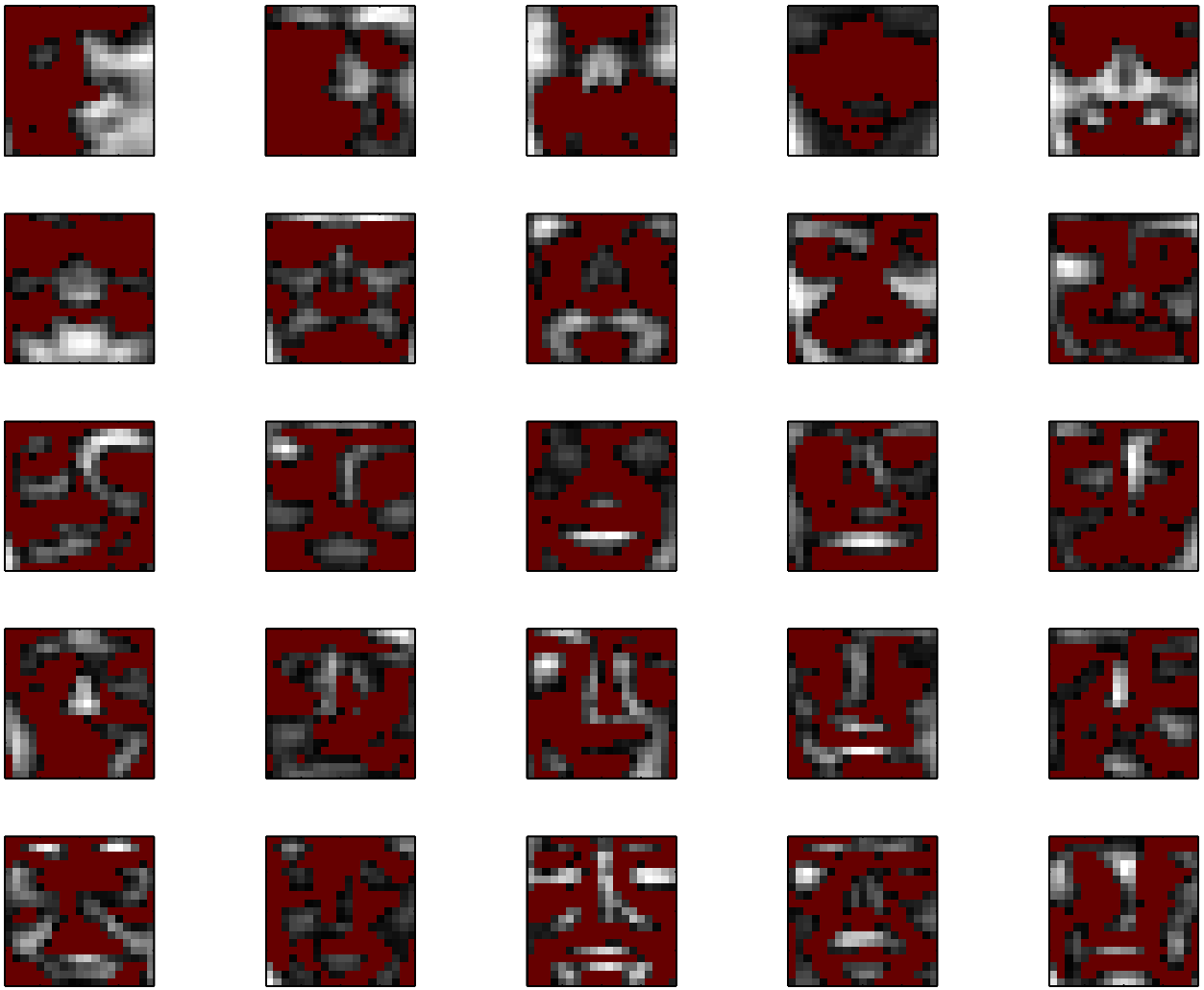
$$\mathbf{V} \approx \mathbf{WH}$$

- Other methods include
  - Principal component analysis (PCA)
  - Independent component analysis (ICA)

- Nonnegative matrices $\mathbf{W}$ and $\mathbf{H}$ are both learned from the set of data vectors $\mathbf{V} = [\mathbf{v}_1, \ldots, \mathbf{v}_n, \ldots, \mathbf{v}_N]$
  - Nonnegativity of $\mathbf{W}$ ensures interpretability of dictionary
  - Nonnegativity of $\mathbf{H}$ tends to produce parts-based representations because subtractive combinations are forbidden

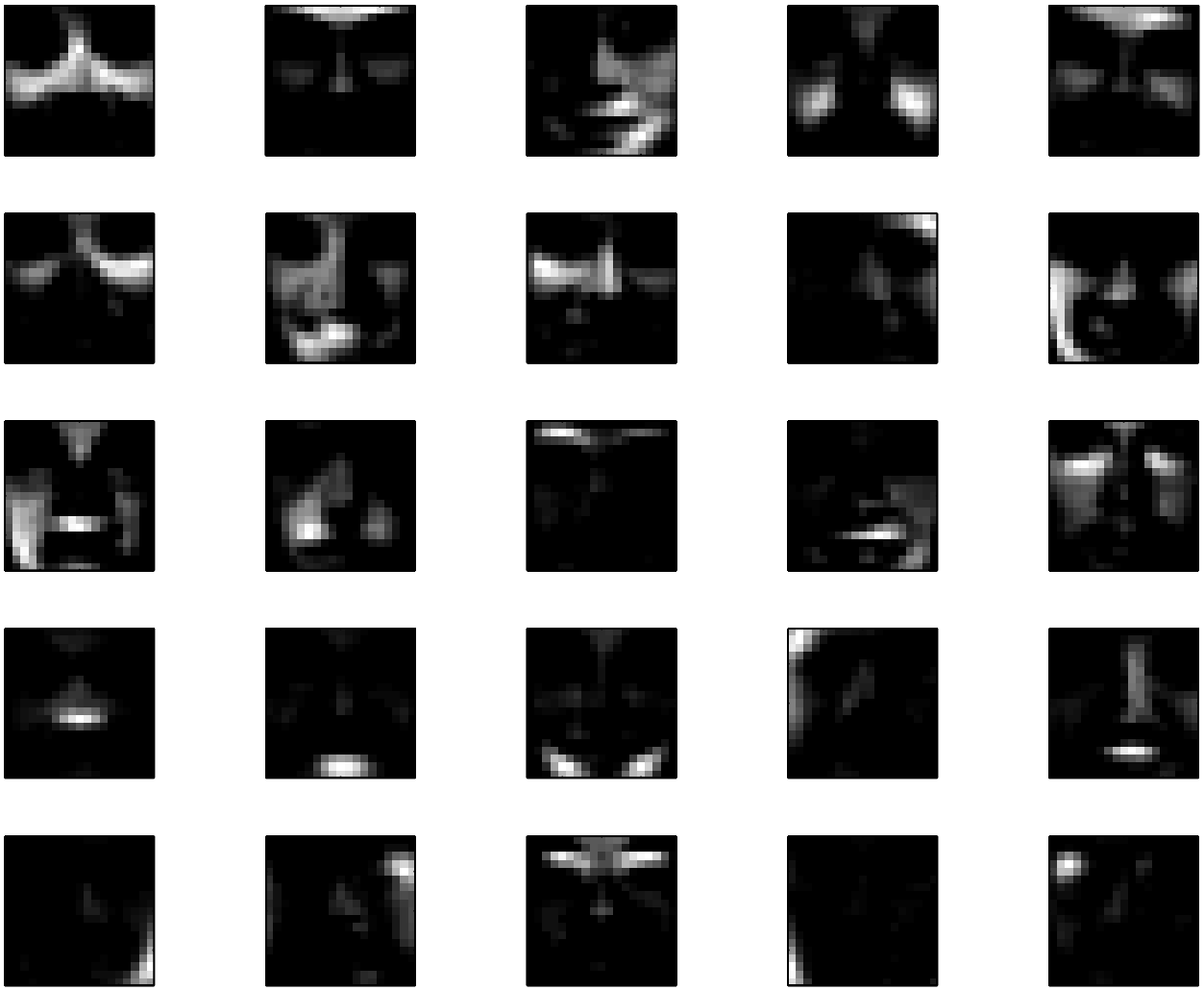
49 images among 2429 from MIT's CBCL face dataset

PCA dictionary with $K = 25$



Red pixels indicate negative values

NMF dictionary with $K = 25$



From Lee and Seung's seminal 1999 paper on NMF

Motivation

- NMF algorithms with squared-ℓ_2 loss have limited applications.
  - When the observation noise is non-Gaussian, other divergence should be used for purpose of ML estimation.
  - When outliers exist in the data matrix $\mathbf{V}$, robust loss functions should be used.

Motivation

- NMF algorithms with squared-ℓ_2 loss have limited applications.
  - When the observation noise is non-Gaussian, other divergence should be used for purpose of ML estimation.
  - When outliers exist in the data matrix $\mathbf{V}$, robust loss functions should be used.

- In Batch NMF, the entire data matrix $\mathbf{V}$ is available all at once. Algorithms have been proposed for many divergences, but not suitable for large-scale data.

Motivation

- NMF algorithms with squared-ℓ_2 loss have limited applications.
  - When the observation noise is non-Gaussian, other divergence should be used for purpose of ML estimation.
  - When outliers exist in the data matrix $\mathbf{V}$, robust loss functions should be used.

- In Batch NMF, the entire data matrix $\mathbf{V}$ is available all at once. Algorithms have been proposed for many divergences, but not suitable for large-scale data.

- In Online NMF, data points $\mathbf{v}_n$ arrive sequentially. Algorithms developed so far are mainly confined to the squared-ℓ_2 loss. Exceptions include some ad hoc works that
  1. have no convergence guarantees;
  2. are not easily generalizable

Motivation

- NMF algorithms with squared-ℓ_2 loss have limited applications.
  - When the observation noise is non-Gaussian, other divergence should be used for purpose of ML estimation.
  - When outliers exist in the data matrix $\mathbf{V}$, robust loss functions should be used.

- In Batch NMF, the entire data matrix $\mathbf{V}$ is available all at once. Algorithms have been proposed for many divergences, but not suitable for large-scale data.

- In Online NMF, data points $\mathbf{v}_n$ arrive sequentially. Algorithms developed so far are mainly confined to the squared-ℓ_2 loss. Exceptions include some ad hoc works that
  1. have no convergence guarantees;
  2. are not easily generalizable

- How to develop an online NMF algorithm that is applicable to a wide variety of divergences?

Challenges and Contributions

- Many online NMF algorithms (with the squared-ℓ_2 loss) leverage the stochastic majorization-minimization framework.

Challenges and Contributions

- Many online NMF algorithms (with the squared-ℓ_2 loss) leverage the stochastic majorization-minimization framework.
- This framework crucially relies on that some sufficient statistics can be formed, but this condition does not hold for most divergences other than the squared-ℓ_2 loss.

Challenges and Contributions

- Many online NMF algorithms (with the squared-ℓ_2 loss) leverage the stochastic majorization-minimization framework.
- This framework crucially relies on that some sufficient statistics can be formed, but this condition does not hold for most divergences other than the squared-ℓ_2 loss.
- We leverage another framework called stochastic approximation framework to develop an online NMF algorithm that is applicable to a wide variety of divergences.

Divergences in Consideration

In this work, we consider a wide range of divergences $\mathcal{D} \triangleq \mathcal{D}_1 \cup \mathcal{D}_2$, where

$$\mathcal{D}_1 \triangleq \{d(\cdot\|\cdot) \,|\, \forall\, \mathbf{x} \in \mathbb{R}^F_{++}, d(\mathbf{x}\|\cdot) \text{ is differentiable on } \mathbb{R}^F_{++}\}.$$

Divergences in Consideration

In this work, we consider a wide range of divergences $\mathcal{D} \triangleq \mathcal{D}_1 \cup \mathcal{D}_2$, where

$$\mathcal{D}_1 \triangleq \{d(\cdot\|\cdot) \,|\, \forall\, \mathbf{x} \in \mathbb{R}^F_{++}, d(\mathbf{x}\|\cdot) \text{ is differentiable on } \mathbb{R}^F_{++}\}.$$

and

$$\mathcal{D}_2 \triangleq \{d(\cdot\|\cdot) \,|\, \forall\, \mathbf{x} \in \mathbb{R}^F_{++}, d(\mathbf{x}\|\cdot) \text{ is convex on } \mathbb{R}^F_{++}\}.$$

Examples of Divergences

- $\mathcal{D}_1$ contains the families of α, β, α-β and γ divergences;
- $\mathcal{D}_2$ contains the α-divergences, β-divergences with $\beta \in [1, 2]$, several robust metrics including ℓ_1, ℓ_2 and Huber loss

Examples of Divergences

- $\mathcal{D}_1$ contains the families of α, β, α-β and γ divergences;
- $\mathcal{D}_2$ contains the α-divergences, β-divergences with $\beta \in [1, 2]$, several robust metrics including ℓ_1, ℓ_2 and Huber loss

Table 1: Expressions of some important cases of Csiszár f-divergence

ℓ_1 distance	$\sum_i \lvert x_i - y_i \rvert$
α-divergence ($\alpha \in \mathbb{R} \setminus \{0, 1\}$)	$\frac{1}{\alpha(\alpha-1)} \sum_i \left(y_i \left[\left(\frac{x_i}{y_i} \right)^{\alpha} - 1 \right] - \alpha(x_i - y_i) \right)$
Hellinger distance ($\alpha = \frac{1}{2}$)	$2 \sum_i (\sqrt{x_i} - \sqrt{y_i})^2$
KL divergence ($\alpha \to 1$)	$\sum_i x_i \log(x_i / y_i) - x_i + y_i$

Examples of Divergences

- $\mathcal{D}_1$ contains the families of α, β, α-β and γ divergences;
- $\mathcal{D}_2$ contains the α-divergences, β-divergences with $\beta \in [1, 2]$, several robust metrics including ℓ_1, ℓ_2 and Huber loss

Table 1: Expressions of some important cases of Csiszár f-divergence

ℓ_1 distance	$\sum_i \lvert x_i - y_i \rvert$
α-divergence ($\alpha \in \mathbb{R} \setminus \{0, 1\}$)	$\frac{1}{\alpha(\alpha-1)} \sum_i \left(y_i \left[\left(\frac{x_i}{y_i} \right)^{\alpha} - 1 \right] - \alpha(x_i - y_i) \right)$
Hellinger distance ($\alpha = \frac{1}{2}$)	$2 \sum_i (\sqrt{x_i} - \sqrt{y_i})^2$
KL divergence ($\alpha \to 1$)	$\sum_i x_i \log(x_i / y_i) - x_i + y_i$

Table 2: Expressions of some important cases of Bregman divergence

Mahalanobis distance	$(\mathbf{x} - \mathbf{y})^T \mathbf{A} (\mathbf{x} - \mathbf{y})/2$
β-divergence ($\beta \in \mathbb{R} \setminus \{0, 1\}$)	$\frac{1}{\beta(\beta-1)} \sum_i \left(x_i^{\beta} - y_i^{\beta} - \beta y_i^{\beta-1} (x_i - y_i) \right)$
IS divergence ($\beta \to 0$)	$\sum_i (\log(y_i / x_i) + x_i / y_i - 1)$
KL divergence ($\beta \to 1$)	$\sum_i x_i \log(x_i / y_i) - x_i + y_i$
Squared ℓ_2 distance ($\beta = 2$)	$\lVert \mathbf{x} - \mathbf{y} \rVert_2^2 / 2$

Problem Formulation

- Consider a loss function

$$\ell(\mathbf{v}, \mathbf{W}) \triangleq \min_{\mathbf{h} \in \mathcal{H}} d(\mathbf{v} \| \mathbf{W}\mathbf{h}).$$

where

$$\mathcal{H} \triangleq \{\mathbf{h} \in \mathbb{R}_+^K \mid \epsilon' \leq h_i \leq U', \forall\, i \in [K]\}.$$

Problem Formulation

- Consider a loss function

$$\ell(\mathbf{v}, \mathbf{W}) \triangleq \min_{\mathbf{h} \in \mathcal{H}} d(\mathbf{v} \| \mathbf{W}\mathbf{h}).$$

where

$$\mathcal{H} \triangleq \{\mathbf{h} \in \mathbb{R}_+^K \mid \epsilon' \leq h_i \leq U', \forall\, i \in [K]\}.$$

- Main problem (a stochastic program)

$$\min_{\mathbf{W} \in \mathcal{C}} \left[f(\mathbf{W}) \triangleq \mathbb{E}_{\mathbf{v} \sim \mathbb{P}}[\ell(\mathbf{v}, \mathbf{W})] \right],$$

where

$$\mathcal{C} \triangleq \{\mathbf{W} \in \mathbb{R}_+^{F \times K} \mid \|\mathbf{W}_{i:}\|_1 \geq \epsilon, \|\mathbf{W}_{:j}\|_\infty \leq U, \forall\, (i,j) \in [F] \times [K]\}$$

is the constraint set on $\mathbf{W}$.

Definitions

Let $\mathcal{X}$ be a finite-dimensional real Banach space. Let $f : \mathcal{X} \to \mathbb{R}$.

Definition (Fréchet subdifferential)

The *Fréchet subdifferential* at $x \in \mathcal{X}$, $\hat{\partial} f(x)$ is defined as

$$\hat{\partial} f(x) \triangleq \left\{ g \in \mathcal{X}^* \,\middle|\, \liminf_{y \to x, y \in \mathcal{X}} \frac{f(y) - f(x) - g(y - x)}{\|y - x\|} \geq 0 \right\},$$

where $\mathcal{X}^*$ is the topological dual space of $\mathcal{X}$.

Definitions

Let $\mathcal{X}$ be a finite-dimensional real Banach space. Let $f : \mathcal{X} \to \mathbb{R}$.

Definition (Fréchet subdifferential)

The *Fréchet subdifferential* at $x \in \mathcal{X}$, $\hat{\partial} f(x)$ is defined as

$$\hat{\partial} f(x) \triangleq \left\{ g \in \mathcal{X}^* \,\middle|\, \liminf_{y \to x, y \in \mathcal{X}} \frac{f(y) - f(x) - g(y - x)}{\|y - x\|} \geq 0 \right\},$$

where $\mathcal{X}^*$ is the topological dual space of $\mathcal{X}$.

Definition (Directional derivative)

The *(Gâteaux) directional derivative* of f at $x \in \mathcal{X}$ along direction $d \in \mathcal{X}$, $f'(x; d)$ is defined as

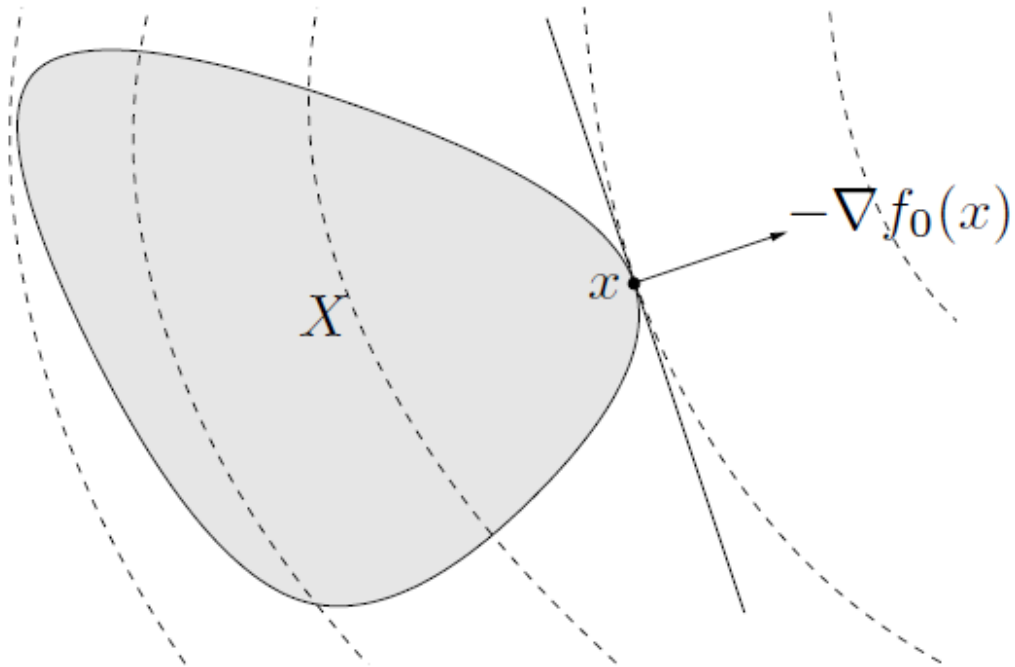
$$f'(x; d) \triangleq \lim_{\delta \downarrow 0} \frac{f(x + \delta d) - f(x)}{\delta}.$$

f is called directionally differentiable if $f'(x; d)$ exists $\forall\, x \in \mathcal{X}, d \in \mathcal{X}$.

Illustration of Critical Points

x is optimal for the optimization probl. $\min_{x \in X} f(x)$ if $x \in X$ and

$$\nabla f(x)^T (y - x) \geq 0 \quad \forall\, y \in X$$



$f'(x; d)$ is a generalization of $\nabla f(x)^T d$

Other Definitions

- Also define $\tilde{d}_t : \mathbb{R}_{++}^{F \times K} \to \mathbb{R}$ and $\bar{d}_t : \mathbb{R}_{++}^{K} \to \mathbb{R}$ as

$$\tilde{d}_t(\mathbf{W}) \triangleq d(\mathbf{v}_t \| \mathbf{W}\mathbf{h}_t),$$

and

$$\bar{d}_t(\mathbf{h}) \triangleq d(\mathbf{v}_t \| \mathbf{W}_{t-1}\mathbf{h}),$$

where $\{\mathbf{v}_t, \mathbf{W}_t, \mathbf{h}_t\}_{t \in \mathbb{N}}$ are generated per Algorithm I.

- Our algorithm will be defined in terms of these divergence functions

Algorithm I

Input: Initial basis matrix $\mathbf{W}_0 \in \mathcal{C}$, number of iterations T, sequence of step sizes $\{\eta_t\}_{t\in\mathbb{N}}$ ($\sum_{t=1}^{\infty}\eta_t = \infty$ and $\sum_{t=1}^{\infty}\eta_t^2 < \infty$)

for $t = 1$ to T **do**

1) Draw a data sample $\mathbf{v}_t$ from $\mathbb{P}$.

2) Learn the coefficient vector $\mathbf{h}_t$ such that $\mathbf{h}_t$ is a critical point of

$$\min_{\mathbf{h}\in\mathcal{H}} \left[\bar{d}_t(\mathbf{h}) \triangleq d(\mathbf{v}_t \| \mathbf{W}_{t-1}\mathbf{h})\right].$$

3) Update the basis matrix from $\mathbf{W}_{t-1}$ to $\mathbf{W}_t$

$$\mathbf{W}_t := \Pi_{\mathcal{C}}\Big\{\mathbf{W}_{t-1} - \eta_t \mathbf{G}_t\Big\},$$

where $\mathbf{G}_t$ is any element in $\hat{\partial}\tilde{d}_t(\mathbf{W}_{t-1})$.

end for

Output: Final basis matrix $\mathbf{W}_T$

Subroutine: Learning $\mathbf{h}_t$

Input: initial coefficient vector $\mathbf{h}_t^0 \in \mathcal{H}$, basis matrix $\mathbf{W}_{t-1}$, data sample $\mathbf{v}_t$, step sizes $\{\beta_t^k\}_{k\in\mathbb{N}}$
Initialize $k := 0$
repeat

$$\mathbf{h}_t^k := \Pi_{\mathcal{H}}\left\{\mathbf{h}_t^{k-1} - \beta_t^k \mathbf{g}_t^k\right\}, \quad \text{where} \quad \mathbf{g}_t^k \in \hat{\partial}\bar{d}_t(\mathbf{h}_t^{k-1})$$

$$k := k + 1$$

until some convergence criterion is met
Output: Final coefficient vector $\mathbf{h}_t$

Main convergence theorem

Assumptions

1. *The support set* $\mathcal{V} \subseteq \mathbb{R}^F_{++}$ *for the data generation distribution* $\mathbb{P}$ *is compact.*
2. *For all* $(\mathbf{v}, \mathbf{W}) \in \mathcal{V} \times \mathcal{C}$ *and* $d(\cdot\|\cdot) \in \mathcal{D}_2$, $d(\mathbf{v}\|\mathbf{W}\mathbf{h})$ *is m-strongly convex in* $\mathbf{h}$ *for some constant* $m > 0$.

Main convergence theorem

Assumptions

1. *The support set* $\mathcal{V} \subseteq \mathbb{R}^F_{++}$ *for the data generation distribution* $\mathbb{P}$ *is compact.*
2. *For all* $(\mathbf{v}, \mathbf{W}) \in \mathcal{V} \times \mathcal{C}$ *and* $d(\cdot\|\cdot) \in \mathcal{D}_2$, $d(\mathbf{v}\|\mathbf{W}\mathbf{h})$ *is m-strongly convex in* $\mathbf{h}$ *for some constant* $m > 0$.

Theorem

As $t \to \infty$, *the sequence of dictionaries* $\{\mathbf{W}_t\}_{t\in\mathbb{N}}$ *converges almost surely to the set of critical points of*

$$\min_{\mathbf{W}\in\mathcal{C}} \left[f(\mathbf{W}) \triangleq \mathbb{E}_{\mathbf{v}\sim\mathbb{P}}[\ell(\mathbf{v}, \mathbf{W})] \right]$$

formulated with any divergence in $\mathcal{D}_2$ *(convex class).*

Proof Ideas

- The following update of the basis matrix is at the core of our algorithm:

$$\mathbf{W}_t := \Pi_{\mathcal{C}}\Big\{\mathbf{W}_{t-1} - \eta_t \mathbf{G}_t\Big\}.$$

Proof Ideas

- The following update of the basis matrix is at the core of our algorithm:

$$\mathbf{W}_t := \Pi_{\mathcal{C}}\Big\{\mathbf{W}_{t-1} - \eta_t \mathbf{G}_t\Big\}.$$

- Model the projection $\Pi_{\mathcal{C}}$ as an additive noise term $\mathbf{N}_t$:

$$\mathbf{W}_t := \mathbf{W}_{t-1} - \eta_t \nabla f(\mathbf{W}_{t-1}) - \eta_t \mathbf{N}_t + \eta_t \mathbf{Z}_t,$$

where

$$\mathbf{Z}_t \triangleq \frac{1}{\eta_t}\Pi_{\mathcal{C}}\Big\{\mathbf{W}_{t-1} - \eta_t \nabla_{\mathbf{W}} \widetilde{\ell}(\mathbf{v}_t, \mathbf{W}_{t-1})\Big\} - \frac{1}{\eta_t}\Big\{\mathbf{W}_{t-1} - \eta_t \nabla_{\mathbf{W}} \widetilde{\ell}(\mathbf{v}_t, \mathbf{W}_{t-1})\Big\}.$$

Proof Ideas

- The following update of the basis matrix is at the core of our algorithm:

$$\mathbf{W}_t := \Pi_{\mathcal{C}}\Big\{\mathbf{W}_{t-1} - \eta_t \mathbf{G}_t\Big\}.$$

- Model the projection $\Pi_{\mathcal{C}}$ as an additive noise term $\mathbf{N}_t$:

$$\mathbf{W}_t := \mathbf{W}_{t-1} - \eta_t \nabla f(\mathbf{W}_{t-1}) - \eta_t \mathbf{N}_t + \eta_t \mathbf{Z}_t,$$

where

$$\mathbf{Z}_t \triangleq \frac{1}{\eta_t}\Pi_{\mathcal{C}}\Big\{\mathbf{W}_{t-1} - \eta_t \nabla_{\mathbf{W}} \widetilde{\ell}(\mathbf{v}_t, \mathbf{W}_{t-1})\Big\} - \frac{1}{\eta_t}\Big\{\mathbf{W}_{t-1} - \eta_t \nabla_{\mathbf{W}} \widetilde{\ell}(\mathbf{v}_t, \mathbf{W}_{t-1})\Big\}.$$

- Perform a continuous-time interpolation

$$\mathbf{W}^t(s) = \mathbf{W}^t(0) + \mathbf{F}^{t-1}(s) + \mathbf{N}^{t-1}(s) + \mathbf{Z}^{t-1}(s).$$

Illustration of Continuous-Time Interpolation

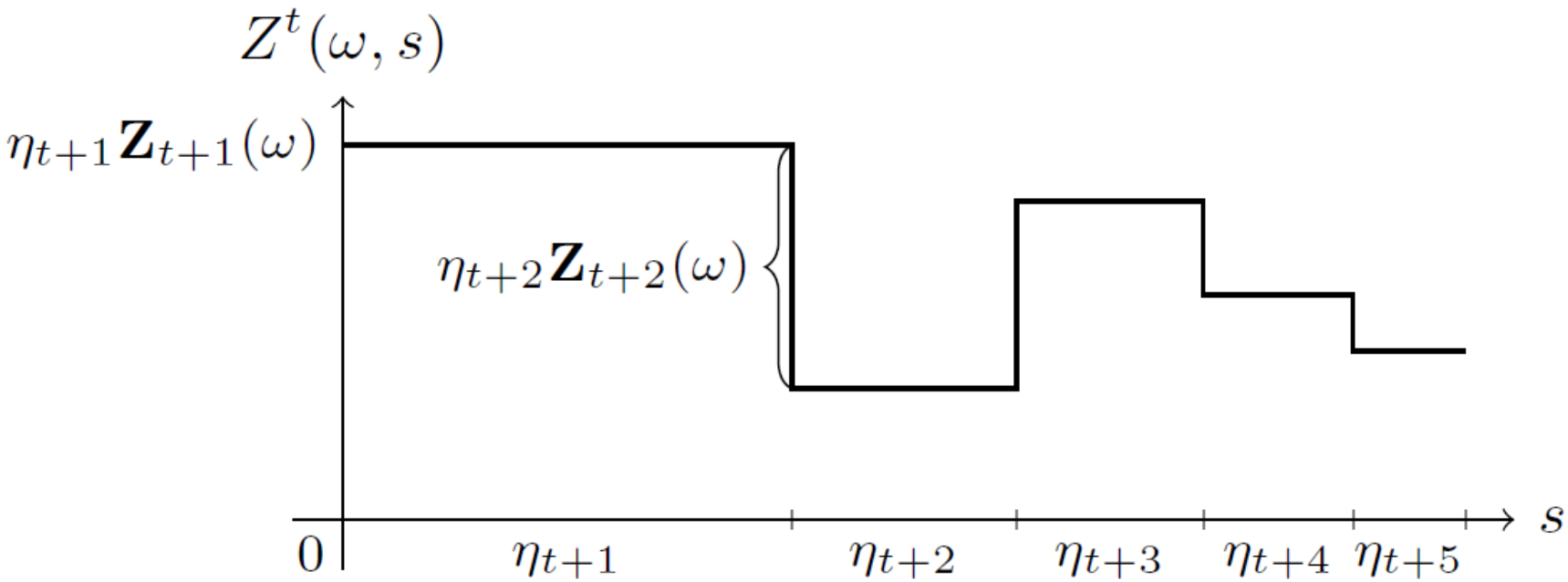


Figure 1: Plot of $Z^t(\omega, \cdot)$ on $\mathbb{R}_+$ for some $\omega \in \Omega$.

Key Lemmas

Lemma (Almost sure convergence to the limit set)

The stochastic process $\{\mathbf{W}_t\}_{t\in\mathbb{N}}$ *generated by the algorithm converges almost surely to* $\mathcal{L}(-\nabla f, \mathcal{C}, \mathbf{W}_0)$*, the limit set of the following projected dynamical system*

$$\frac{\mathrm{d}}{\mathrm{d}s}\mathbf{W}(s) = \pi_{\mathcal{C}}\Big[\mathbf{W}(s), -\nabla f(\mathbf{W}(s))\Big], \quad \mathbf{W}(0) = \mathbf{W}_0, \quad s \geq 0.$$

Key Lemmas

Lemma (Almost sure convergence to the limit set)

The stochastic process $\{\mathbf{W}_t\}_{t\in\mathbb{N}}$ *generated by the algorithm converges almost surely to* $\mathcal{L}(-\nabla f, \mathcal{C}, \mathbf{W}_0)$*, the limit set of the following projected dynamical system*

$$\frac{\mathrm{d}}{\mathrm{d}s}\mathbf{W}(s) = \pi_{\mathcal{C}}\Big[\mathbf{W}(s), -\nabla f(\mathbf{W}(s))\Big], \quad \mathbf{W}(0) = \mathbf{W}_0, \quad s \geq 0.$$

Lemma (Characterization of the limit set)

In the above dynamical system, $\mathcal{L}(-\nabla f, \mathcal{C}, \mathbf{W}_0) \subseteq \mathcal{S}(-\nabla f, \mathcal{C})$*, i.e., every limit point is a stationary point associated with* $-\nabla f$ *and* $\mathcal{C}$*. Moreover, each* $\mathbf{W} \in \mathcal{S}(-\nabla f, \mathcal{C})$ *satisfies*

$$\langle \nabla f(\mathbf{W}), \mathbf{W}' - \mathbf{W} \rangle \geq 0, \quad \forall \mathbf{W}' \in \mathcal{C}.$$

This implies each stationary point in $\mathcal{S}(-\nabla f, \mathcal{C})$ *is a critical point.*

Applications and Experiments

- Focus on six important divergences from $\mathcal{D}_1 \cup \mathcal{D}_2$: IS, KL, squared-$\ell_2$, Huber, ℓ_1, ℓ_2.

Applications and Experiments

- Focus on six important divergences from $\mathcal{D}_1 \cup \mathcal{D}_2$: IS, KL, squared-$\ell_2$, Huber, ℓ_1, ℓ_2.
- Test our online algorithms with all the six divergences on a synthetic dataset.

Applications and Experiments

- Focus on six important divergences from $\mathcal{D}_1 \cup \mathcal{D}_2$: IS, KL, squared-$\ell_2$, Huber, ℓ_1, ℓ_2.
- Test our online algorithms with all the six divergences on a synthetic dataset.
- Test our online algorithm with the KL-divergence on topic modeling and document clustering.

Applications and Experiments

- Focus on six important divergences from $\mathcal{D}_1 \cup \mathcal{D}_2$: IS, KL, squared-$\ell_2$, Huber, ℓ_1, ℓ_2.
- Test our online algorithms with all the six divergences on a synthetic dataset.
- Test our online algorithm with the KL-divergence on topic modeling and document clustering.
- Test our online algorithm with the Huber loss on a foreground-background separation task.

Heuristics and Parameter Settings

- Heuristics: mini-batch input and multi-pass extension.

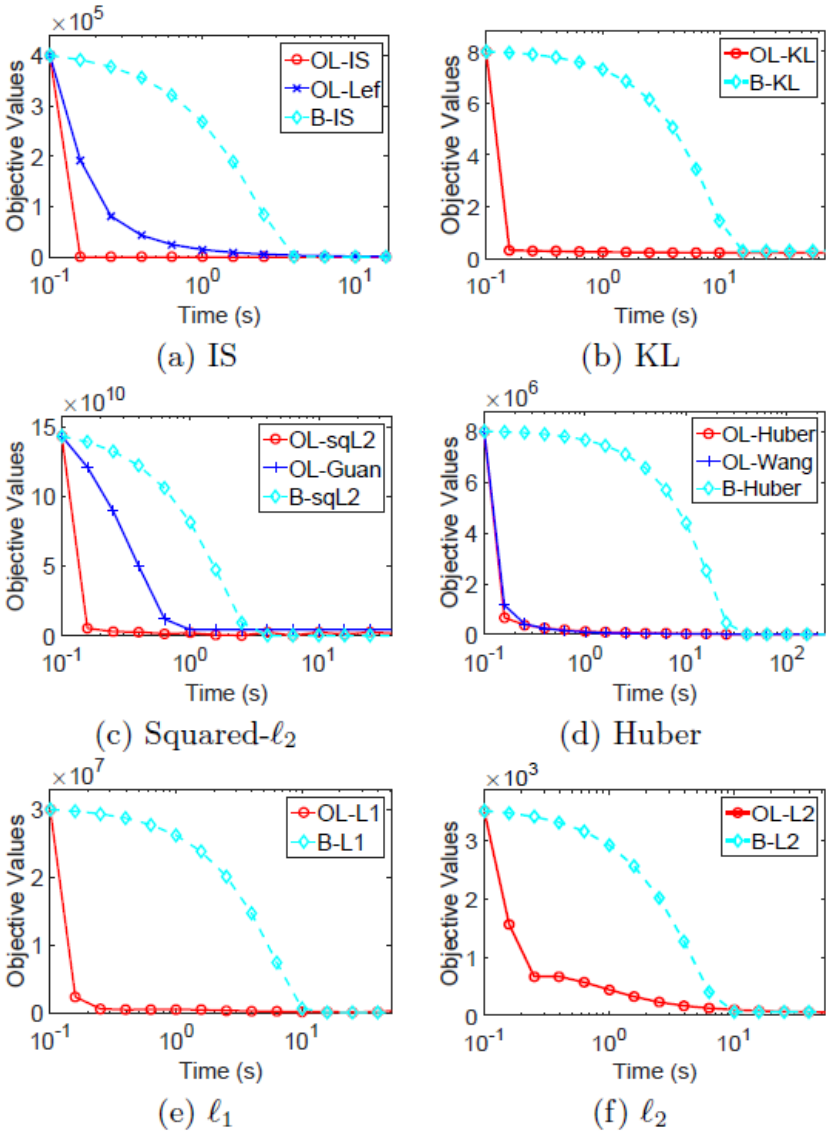
Heuristics and Parameter Settings

- Heuristics: mini-batch input and multi-pass extension.
- Parameters:
  - The mini-batch size $\tau = 20$.
  - The step size
$$\eta_t = \frac{a}{b + \tau t},$$
where $a = b = 1 \times 10^4$.
  - The latent dimension K was determined from the domain knowledge or fixed to 40.

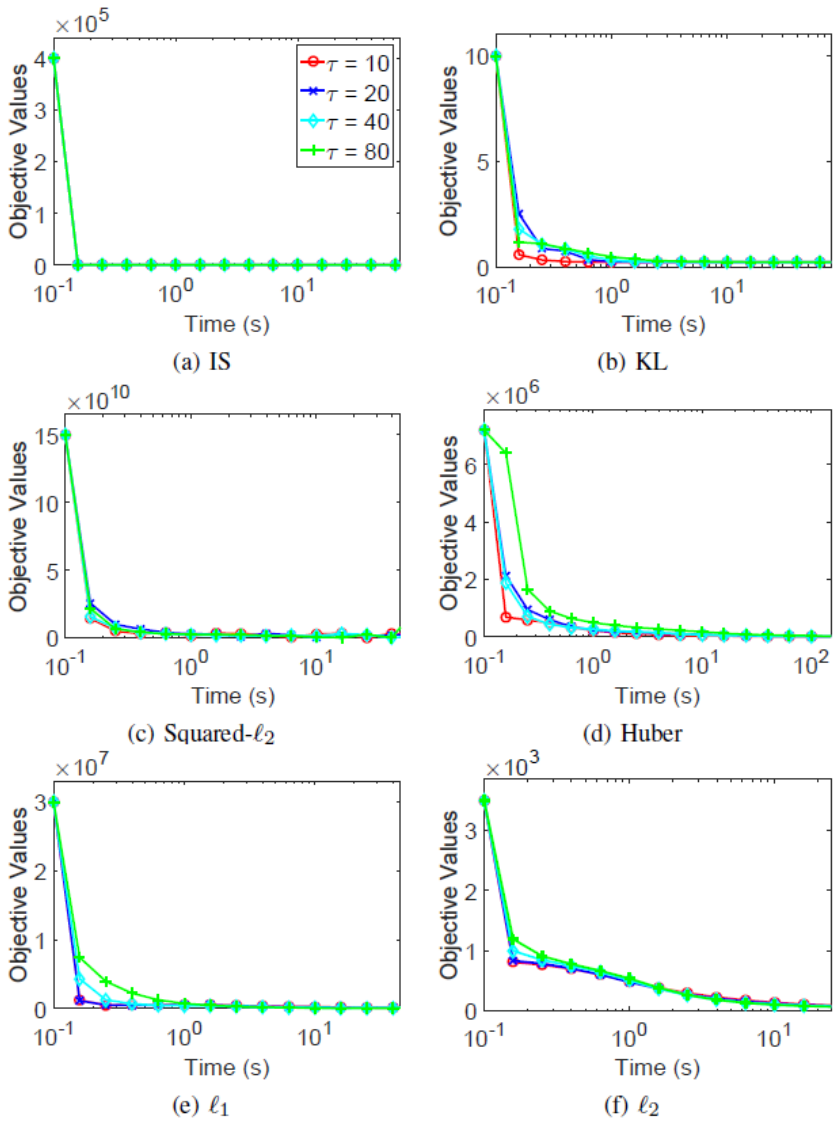
Heuristics and Parameter Settings

- Heuristics: mini-batch input and multi-pass extension.
- Parameters:
  - The mini-batch size $\tau = 20$.
  - The step size
$$\eta_t = \frac{a}{b + \tau t},$$
where $a = b = 1 \times 10^4$.
  - The latent dimension K was determined from the domain knowledge or fixed to 40.
- Our algorithms are insensitive to the value of these parameters.

Synthetic Dataset

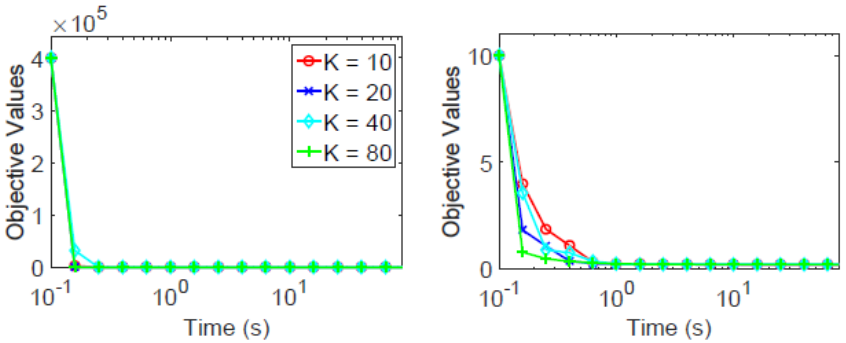


(a) IS (b) KL (c) Squared-ℓ_2 (d) Huber (e) ℓ_1 (f) ℓ_2



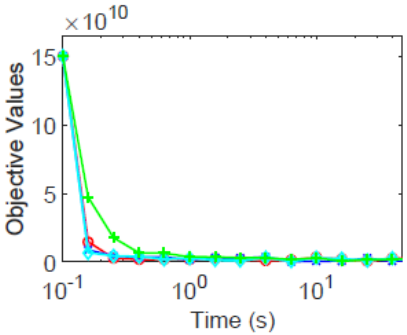
(a) IS (b) KL (c) Squared-ℓ_2 (d) Huber (e) ℓ_1 (f) ℓ_2

Synthetic Dataset

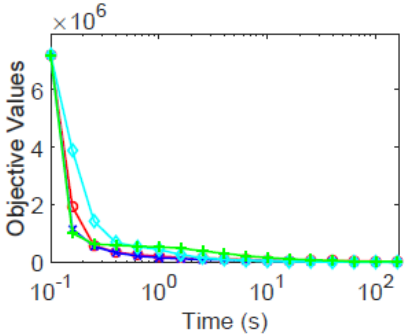


(a) IS

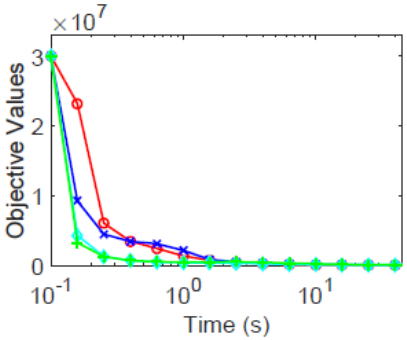
(b) KL



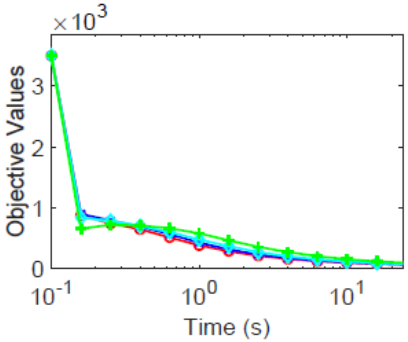
(c) Squared-ℓ_2



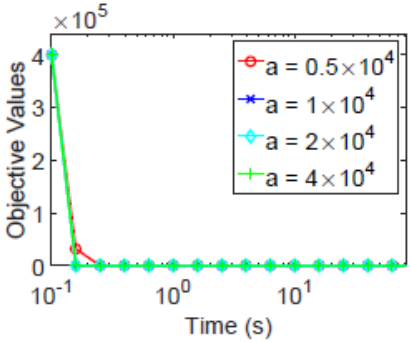
(d) Huber



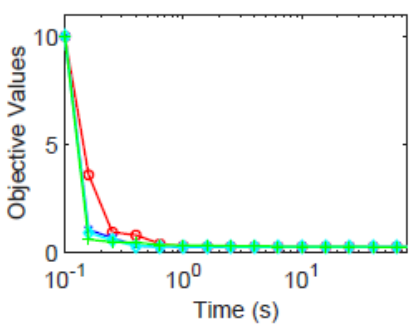
(e) ℓ_1



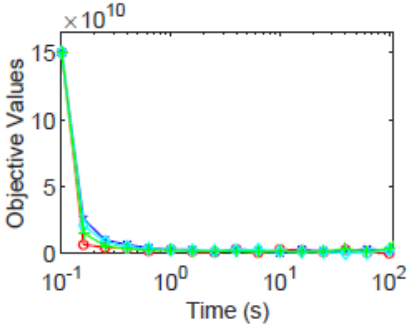
(f) ℓ_2



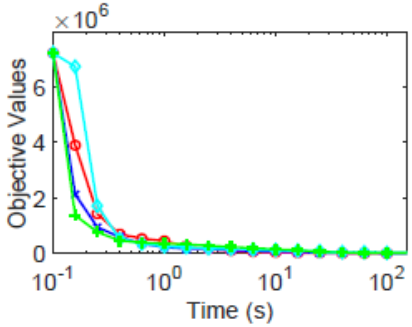
(a) IS



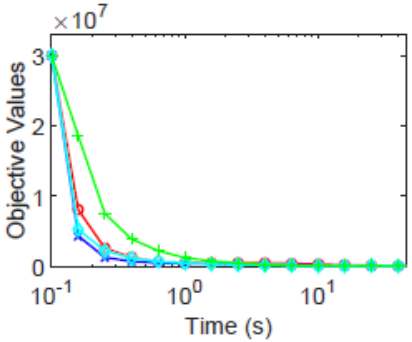
(b) KL



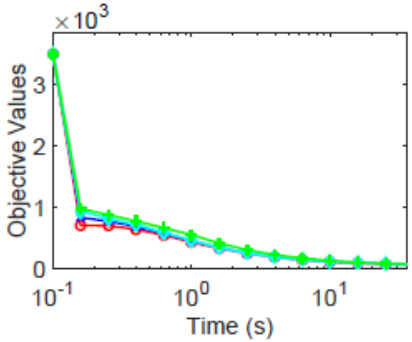
(c) Squared-ℓ_2



(d) Huber



(e) ℓ_1



(f) ℓ_2

Topic Learning I

- K was set to the number of topics in the dataset.
- Experiments were run using 20 initializations of $\mathbf{W}_0$.
- Application of our online algorithm with KL-divergence.
- Topic learned from the columns of $\mathbf{W}$ (basis vectors).
- Two datasets: `Guardian` (10801×5413, five topics) and `Wikipedia` (17311×5738, six topics).

Table 3: 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

(a) OL-KL

Topic Learning II

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

(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

(c) OL-Wang2

Document clustering

Simple clustering rule: we assign the j-th document to the k-th topic if

$$k \in \arg\max_{k' \in [K]} h_{k'j}.$$

Document clustering

Simple clustering rule: we assign the j-th document to the k-th topic if

$$k \in \arg\max_{k' \in [K]} h_{k'j}.$$

Table 4: Average document clustering accuracies and running times of OL-KL, B-KL and OL-Wang2 on the `Guardian` dataset.

Algorithms	**Accuracy**	**Time (s)**
OL-KL	0.697 ± 0.01	29.25 ± 0.58
B-KL	0.701 ± 0.01	183.32 ± 2.09
OL-Wang2	0.643 ± 0.03	32.46 ± 0.68

Foreground-Background Separation I

- K was set to 40.
- Experiments were run using 20 initializations of $\mathbf{W}_0$.
- Application of our online algorithm with Huber loss.
- Background in the t-th frame reconstructed as $\overline{\mathbf{W}}\mathbf{h}_t$, foreground obtained by subtraction.
- Two datasets: `Hall` (25344×1250) and `Wikipedia` (20800×2000).

Foreground-Background Separation II

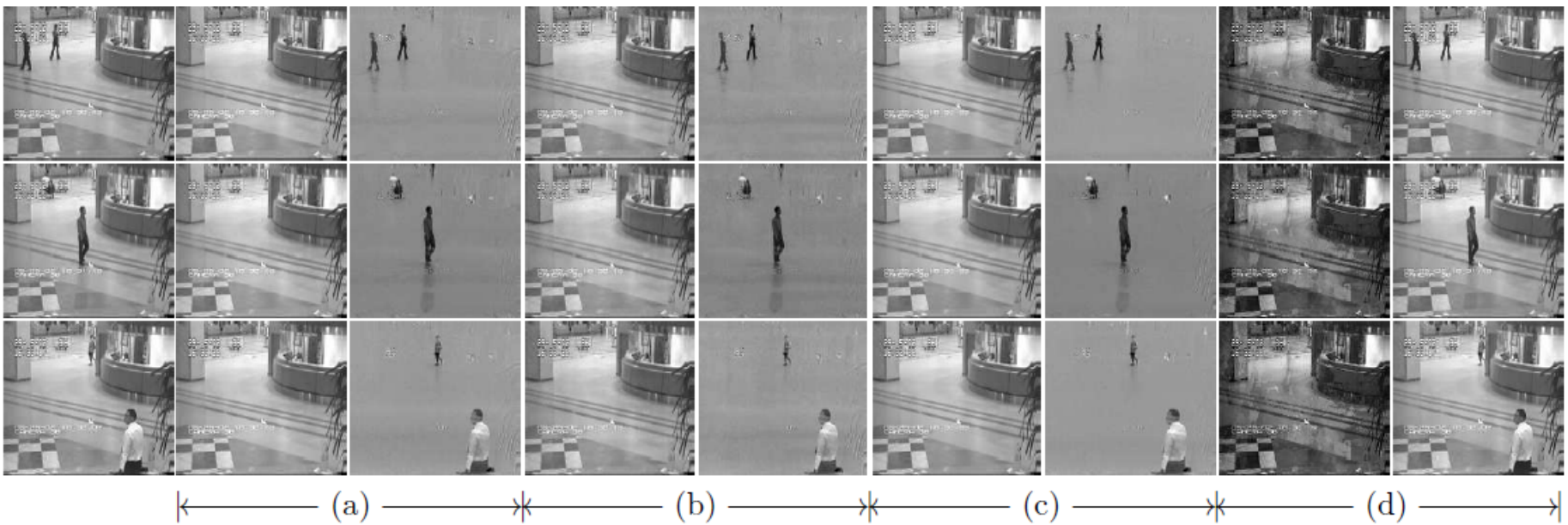


Figure 2: Foreground-background separation results on the `Hall` dataset with four algorithms: (a) OL-Huber, (b) OL-Wang, (c) B-Huber and (d) OL-Guan. The leftmost column shows the original video frames.

Foreground-Background Separation III

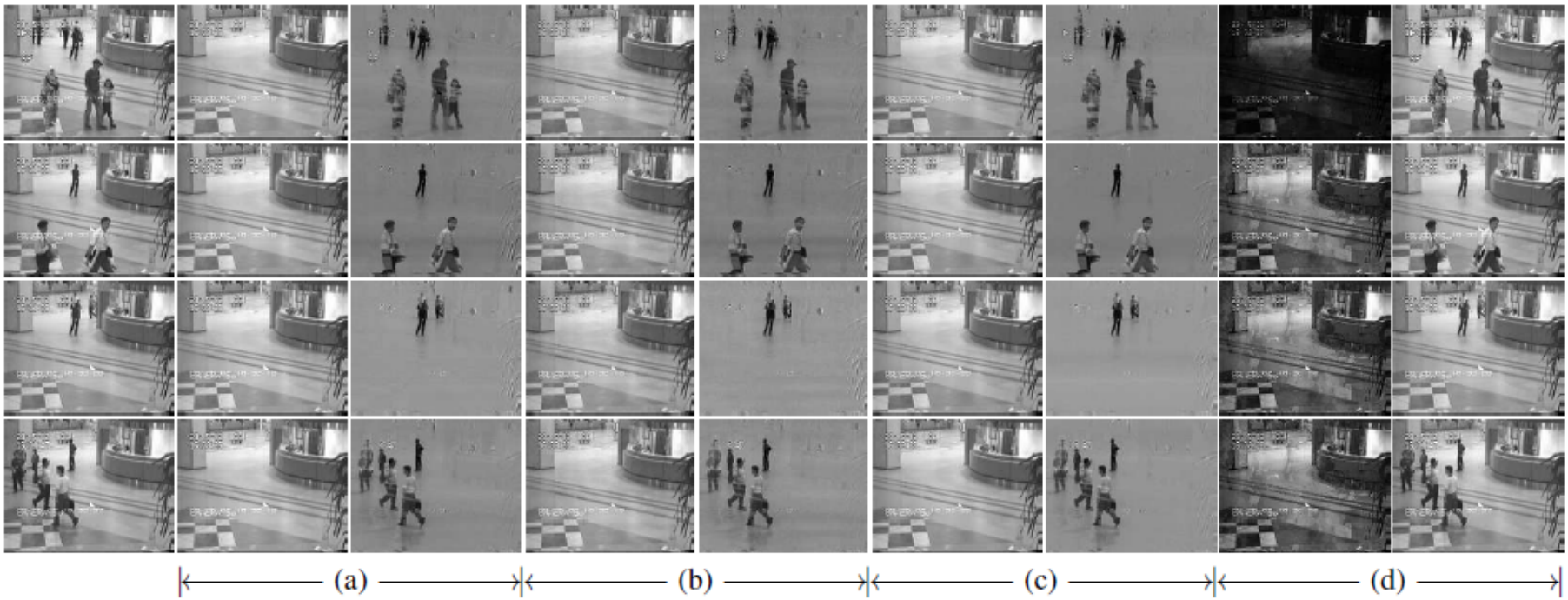


Fig. S-4. Additional foreground-background separation results on the `Hall` dataset with four algorithms: (a) OL-Huber, (b) OL-Wang, (c) B-Huber and (d) OL-Guan. The leftmost column shows the original video frames.

Foreground-Background Separation IV

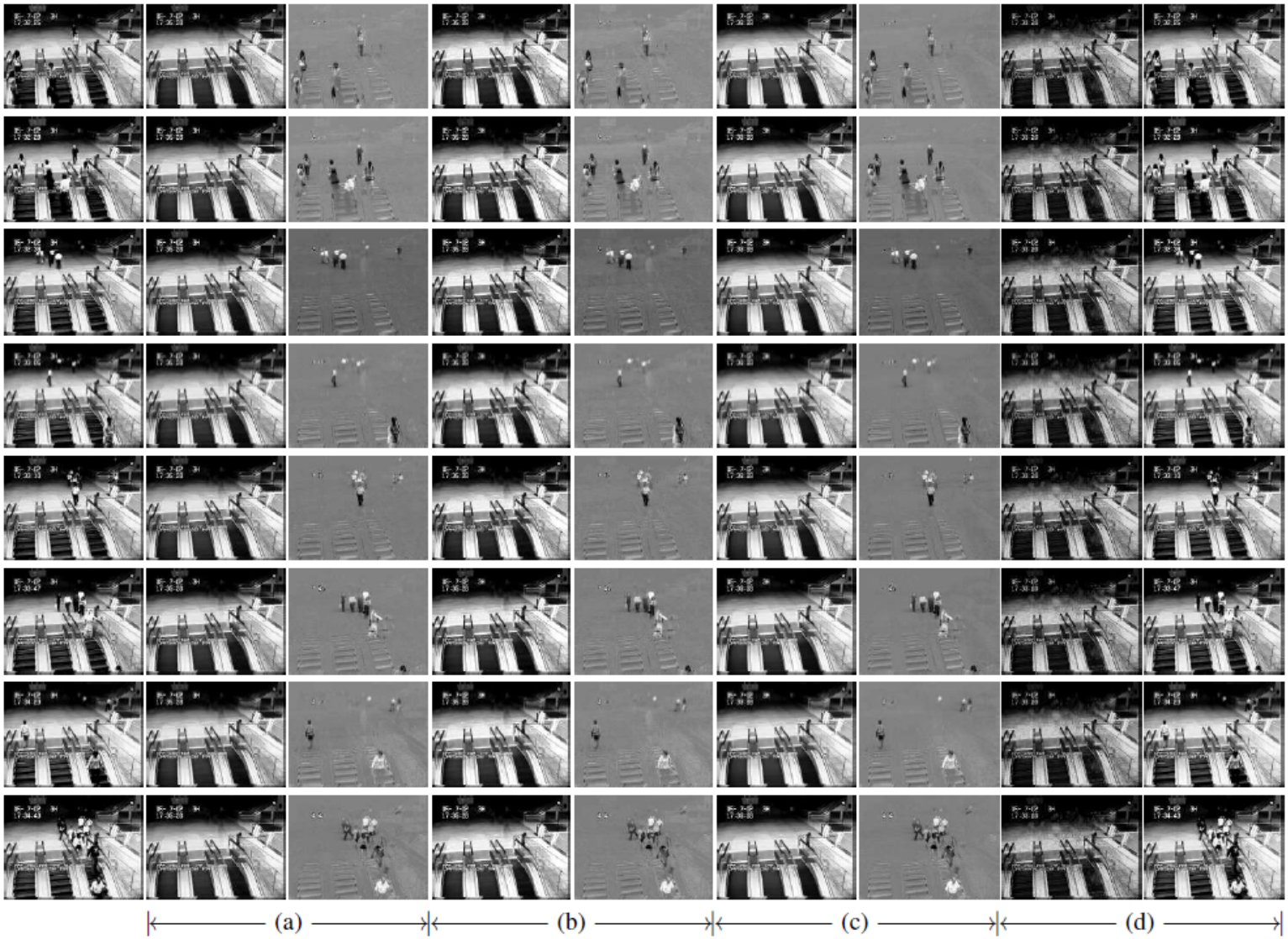


Fig. S-5. Foreground-background separation results on the `Escalator` dataset with four algorithms: (a) OL-Huber, (b) OL-Wang, (c) B-Huber and (d) OL-Guan. The leftmost column shows the original video frames.

Running times

Table 5: Average running times of OL-Huber, OL-Wang, B-Huber and OL-Guan on the `Hall` dataset.

Algorithms	**Time (s)**	**Algorithms**	**Time (s)**
OL-Huber	38.79 ± 0.45	OL-Wang	45.36 ± 0.59
B-Huber	276.66 ± 1.93	OL-Guan	95.85 ± 0.82

Table 6: 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

Conclusion

- Proposed a general framework for doing online NMF with general divergences
- Proved that the sequence of iterates converges to the set of critical points
- Validated the algorithm on several real datasets
- Please visit https://arxiv.org/abs/1608.00075 for more details