PRIMAL-DUAL ALGORITHMS FOR NON-NEGATIVE MATRIX FACTORIZATION WITH THE KULLBACK-LEIBLER DIVERGENCE

Felipe Yanez
INSEAD

Joint work with Francis Bach, done while at INRIA/École normale supérieure

42nd IEEE International Conference on Acoustics, Speech and Signal Processing

March 7, 2017, New Orleans, LA, USA

Goal for this presentation

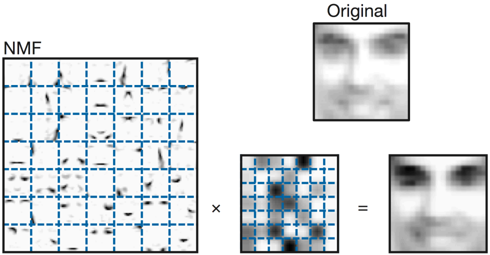
Share how we developed a first-order method for non-negative matrix factorization (NMF) with the Kullback-Leibler (KL) loss.

Agenda

1. Context of the problem.
2. Formulation of the proposed method.
3. Experimental results on synthetic and real-world data.

MOTIVATION

What is non-negative matrix factorization?



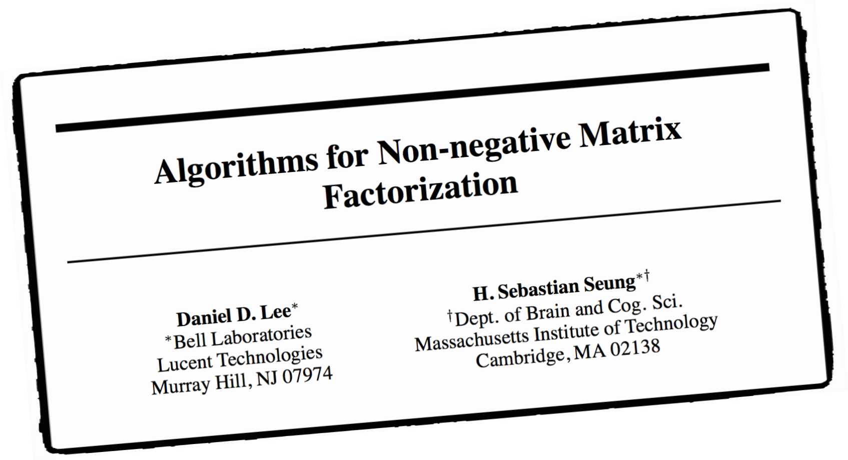
(Lee & Seung, Nature 1999)

Given a matrix $\mathbf{V}$, find $\mathbf{W}$ and $\mathbf{H}$ such that

$$\mathbf{V} \approx \mathbf{W}\mathbf{H},$$

where $\mathbf{V} \in \mathbb{R}_+^{n \times m}$, $\mathbf{W} \in \mathbb{R}_+^{n \times r}$, $\mathbf{H} \in \mathbb{R}_+^{r \times m}$, with $r \leq \min(n, m)$.

Multiplicative updates algorithms


Algorithms for Non-negative Matrix Factorization

Daniel D. Lee*
*Bell Laboratories
Lucent Technologies
Murray Hill, NJ 07974

H. Sebastian Seung*†
†Dept. of Brain and Cog. Sci.
Massachusetts Institute of Technology
Cambridge, MA 02138

Multiplicative updates algorithms

Advantages: stability, ease of implementation, and linear complexity per iteration.

Disadvantages: slow convergence, asymptotic convergence to zeros, and poor local optima.

(Sun & Févotte, ICASSP 2014)

Multiplicative updates algorithms

Advantages: stability, ease of implementation, and linear complexity per iteration.

Disadvantages: slow convergence, asymptotic convergence to zeros, and poor local optima.

Is it possible to address these shortcomings?

Gradient-based methods have better behavior

... but only apply to smooth losses

To find $\mathbf{W}$ and $\mathbf{H}$ with loss $d(x|y)$ we solve

$$\underset{\mathbf{W},\mathbf{H}\geq 0}{\text{minimize}} \quad D(\mathbf{V}|\mathbf{W}\mathbf{H}) = \sum_{ij} d\left(\mathbf{V}_{ij}|(\mathbf{W}\mathbf{H})_{ij}\right).$$

Euclidean (smooth): $d_{EUC}(x|y) = \frac{1}{2}(y-x)^2$

Kullback-Leibler (non-smooth): $d_{KL}(x|y) = x\log(x/y) + (y-x)$

Gradient-based methods have better behavior

... but only apply to smooth losses

To find $\mathbf{W}$ and $\mathbf{H}$ with loss $d(x|y)$ we solve

$$\underset{\mathbf{W},\mathbf{H}\geq 0}{\text{minimize}} \quad D(\mathbf{V}|\mathbf{WH}) = \sum_{ij} d\left(\mathbf{V}_{ij}|(\mathbf{WH})_{ij}\right).$$

Euclidean (smooth): $d_{EUC}(x|y) = \frac{1}{2}(y-x)^2$

Lin (2007) and Kim et al. (2008), between others.

Kullback-Leibler (non-smooth): $d_{KL}(x|y) = x\log(x/y) + (y-x)$

Gradient-based methods have better behavior

... but only apply to smooth losses

To find $\mathbf{W}$ and $\mathbf{H}$ with loss $d(x|y)$ we solve

$$\underset{\mathbf{W},\mathbf{H} \geq 0}{\text{minimize}} \quad D(\mathbf{V}|\mathbf{W}\mathbf{H}) = \sum_{ij} d\left(\mathbf{V}_{ij}|(\mathbf{W}\mathbf{H})_{ij}\right).$$

Euclidean (smooth): $d_{EUC}(x|y) = \frac{1}{2}(y-x)^2$

Lin (2007) and Kim et al. (2008), between others.

Kullback-Leibler (non-smooth): $d_{KL}(x|y) = x\log(x/y) + (y-x)$

The goal is to provide a similar first-order method for the KL loss.

PROPOSED METHOD

The saddle-point problem offers flexibility

... it does not require a smooth loss

$$\min_{x \in \mathcal{X}} \max_{y \in \mathcal{Y}} \langle Kx, y \rangle + G(x) - F^*(y)$$

$$\underbrace{\min_{x \in \mathcal{X}} F(Kx) + G(x)}_{\text{PRIMAL}} \qquad \underbrace{\max_{y \in \mathcal{Y}} -F^*(y) - G^*(-K^\top y)}_{\text{DUAL}}$$

- $\mathcal{X}$ and $\mathcal{Y}$ are two real vector spaces, with $\dim(\mathcal{X}) = p$ and $\dim(\mathcal{Y}) = q$.
- $G : \mathcal{X} \to \mathbb{R} \cup [+\infty]$ and $F^* : \mathcal{Y} \to \mathbb{R} \cup [+\infty]$ are proper, convex, and lower-semicontinuous functions. F^* is the convex conjugate of F.
- $K : \mathcal{X} \to \mathcal{Y}$ is a continuous linear operator with induced norm $\|K\| = \max\{\|Kx\| : x \in \mathcal{X} \textit{ with } \|x\| \leq 1\}$.

(Modified from: Chambolle & Pock, J Math Imaging Vis 2011)

Non-negative decomposition (convex)

$$\underset{\mathbf{W},\mathbf{H}\geq 0}{\text{minimize}} \quad D_{KL}(\mathbf{V}|\mathbf{W}\mathbf{H})$$

$$\underset{\mathbf{H}\geq 0}{\text{minimize}}\, D_{KL}(\mathbf{V}|\mathbf{W}\mathbf{H}) \qquad \underset{\mathbf{W}\geq 0}{\text{minimize}}\, D_{KL}(\mathbf{V}^\top|\mathbf{H}^\top\mathbf{W}^\top)$$

$a \in \mathbb{R}_+^p$	is the given data
$K \in \mathbb{R}_+^{p\times q}$	is the fixed factor
$x \in \mathbb{R}_+^q$	is to be estimated

$$\underset{x\succeq 0}{\text{minimize}}\, D_{KL}(a|Kx)$$

Primal and dual formulation

The non-negative decomposition problem with the KL loss

$$\underset{x \succeq \mathbf{0}}{\text{minimize}} \;\; a^\top \log\left(a \oslash (Kx)\right) + \mathbf{1}^\top (Kx - a)$$

is equivalent to the primal problem $\min_{x \in \mathcal{X}} \; F(Kx) + G(x)$ with $F(y) = a^\top \log(a \oslash y) - \mathbf{1}^\top a$ and $G(x) = \mathbb{1}_{x \succeq \mathbf{0}} + \mathbf{1}^\top K x$.

Then, the dual problem $\max_{y \in \mathcal{Y}} \; -F^*(y) - G^*(-K^\top y)$ with $F^*(y) = -a^\top \log(-y)$ and $G^*(x) = \mathbb{1}_{x \preceq K^\top \mathbf{1}}$ is

$$\underset{K^\top(-y) \preceq K^\top \mathbf{1}}{\text{maximize}} \;\; a^\top \log(-y)\,.$$

Note: $\oslash$ represents the entry-wise division operator.

First-order primal-dual algorithm

Select $K \in \mathbb{R}_+^{p \times q}$, $x \in \mathbb{R}_+^q$, and $\sigma, \tau > 0$;

Set $x = \bar{x} = x_{old} = x_0$, and $y = y_0$;

while *stopping criteria not reached* **do**

$y \leftarrow \mathbf{prox}_{\sigma F^*}(y + \sigma K \bar{x})$;

$x \leftarrow \mathbf{prox}_{\tau G}(x - \tau K^\top y)$;

$\bar{x} \leftarrow 2x - x_{old}$;

$x_{old} \leftarrow x$;

end

return $x^\star = x$ *and* $y^\star = y$

(Modified from: Chambolle & Pock, J Math Imaging Vis 2011)

First-order primal-dual algorithm

Select $K \in \mathbb{R}_+^{p\times q}$, $x \in \mathbb{R}_+^q$, and $\sigma, \tau > 0$;

Set $x = \bar{x} = x_{old} = x_0$, and $y = y_0$;

while *stopping criteria not reached* **do**

$y \leftarrow \mathbf{prox}_{\sigma F^*}(y + \sigma K\bar{x})$;

$x \leftarrow \mathbf{prox}_{\tau G}(x - \tau K^\top y)$;

$\bar{x} \leftarrow 2x - x_{old}$;

$x_{old} \leftarrow x$;

end

return $x^\star = x$ *and* $y^\star = y$

$$\mathbf{prox}_{\sigma F^*}(y) = \frac{1}{2}\left(y - \sqrt{y \circ y + 4\sigma a}\right)$$

$$\mathbf{prox}_{\tau G}(x) = \left(x - \tau K^\top \mathbf{1}\right)_+$$

The proximal operator is defined as

$$\mathbf{prox}_{\tau F}(x) = \arg\min_y \left\{\frac{\|x-y\|^2}{2\tau} + F(y)\right\}$$

(Modified from: Chambolle & Pock, J Math Imaging Vis 2011)

First-order primal-dual algorithm

Select $K \in \mathbb{R}_+^{p \times q}$, $x \in \mathbb{R}_+^q$, and $\sigma, \tau > 0$;

Set $x = \bar{x} = x_{old} = x_0$, and $y = y_0$;

while *stopping criteria not reached* **do**

$y \leftarrow \mathbf{prox}_{\sigma F^*}(y + \sigma K\bar{x})$; $\mathbf{prox}_{\sigma F^*}(y) = \frac{1}{2}\left(y - \sqrt{y \circ y + 4\sigma a}\right)$

$x \leftarrow \mathbf{prox}_{\tau G}(x - \tau K^\top y)$; $\mathbf{prox}_{\tau G}(x) = \left(x - \tau K^\top \mathbf{1}\right)_+$

$\bar{x} \leftarrow 2x - x_{old}$;

$x_{old} \leftarrow x$;

We need to set the step-sizes !

end

return $x^\star = x$ *and* $y^\star = y$

(Modified from: Chambolle & Pock, J Math Imaging Vis 2011)

Automatic heuristic selection of step-sizes

Based on the convergence proofs, we know that

1. the step-sizes have to satisfy $\tau\sigma\|K\|^2 \leq 1$, and
2. the convergence rate is controlled by the quantity C.

Automatic heuristic selection of step-sizes

Based on the convergence proofs, we know that

1. the step-sizes have to satisfy $\tau\sigma\|K\|^2 \leq 1$, and
2. the convergence rate is controlled by the quantity C.

We formulate an optimization problem to estimate σ and τ

$$\begin{aligned} \underset{\sigma,\tau}{\text{minimize}} \quad & C = \frac{\|y_0 - y^\star\|^2}{2\sigma} + \frac{\|x_0 - x^\star\|^2}{2\tau} \\ \text{subject to} \quad & \tau\sigma\|K\|^2 \leq 1. \end{aligned}$$

Automatic heuristic selection of step-sizes

Based on the convergence proofs, we know that

1. the step-sizes have to satisfy $\tau\sigma\|K\|^2 \leq 1$, and
2. the convergence rate is controlled by the quantity C.

We formulate an optimization problem to estimate σ and τ

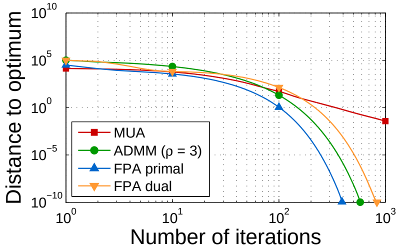
$$\begin{aligned} \underset{\sigma,\tau}{\text{minimize}} \quad & C = \frac{\|y_0 - y^\star\|^2}{2\sigma} + \frac{\|x_0 - x^\star\|^2}{2\tau} \\ \text{subject to} \quad & \tau\sigma\|K\|^2 \leq 1. \end{aligned}$$

Using heuristic replacements, $(x^\star, y^\star) = (\alpha\mathbf{1}, \beta\mathbf{1})$, we obtain

$$\sigma = \frac{\sqrt{p}\,\mathbf{1}^\top K\,\mathbf{1}}{\sqrt{q}\,\|K\|\mathbf{1}^\top a} \quad \text{and} \quad \tau = \frac{\sqrt{q}\,\mathbf{1}^\top a}{\sqrt{p}\,\|K\|\mathbf{1}^\top K\,\mathbf{1}}.$$

RESULTS

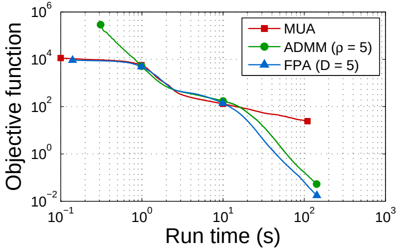
Experiments on synthetic data



Non-negative decomposition

$(n, m, r) = (200, 500, 10)$

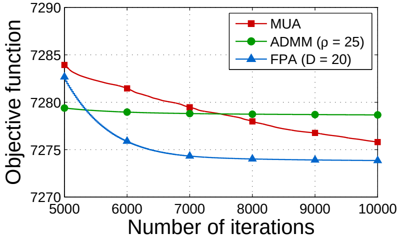
Estimate $\mathbf{H}$ given $\mathbf{W}^\star$



Non-negative matrix factorization

$(n, m, r) = (200, 500, 10)$

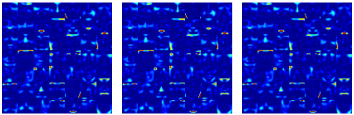
Non-negative matrix factorization on real-world data



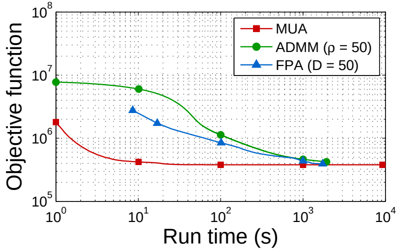
MIT-CBCL Face Database #1

$$(n, m, r) = (361, 2429, 10)$$

$\mathbf{W}$



MUA (22.4 min) ADMM (21.2 min) FPA (20.9 min)



Spectrogram of a jazz song

$$(n, m, r) = (129, 9312, 20)$$

MUA (153 min) **ADMM (32 min)** **FPA (28 min)**

Future work: extension to topic models

- The formulations of probabilistic latent semantic analysis or latent Dirichlet allocation relate to ours.
- If we include the constraint $\mathbf{1}^\top x = 1$ to G:

$$G(x) = \mathbb{1}_{\{\mathbf{1}^\top x=1;\, x\succeq 0\}} + 1^\top K x,$$

  we can use our method to find the latent topics.
- Note that in this case $\mathbf{prox}_{\tau G}(x)$ does not have a closed solution, but can be efficiently solved with dedicated methods for orthogonal projections on the simplex.

Summary

- We proposed a first-order primal-dual algorithm for non-negative decomposition problems with the Kullback-Leibler loss.
- By using alternating optimization, our algorithm readily extends to non-negative matrix factorization.
- All required computations may be obtained in closed form. We provided an efficient heuristic way to select step-sizes.
- On synthetic or real-world data, our method is either faster than existing algorithms, or leads to improved local optima, or both.