New numerical tools for optimal transport and their machine learning applications

Jianbo Ye

Pennsylvania State University

jxy198@ist.psu.edu

(joint work with my advisors James Z. Wang and Jia Li)

Oaxaca. May 4, 2017

Overview

1. Wasserstein Barycenter with Bregman ADMM
   - B-ADMM based Algorithm
   - Discrete Distribution Clustering

2. A Simulated Annealing based Inexact Oracle for Wasserstein Loss Minimization
   - The Gibbs-OT Sampler
   - Wasserstein Non-negative Matrix Factorization

Optimal Transport with Bregman ADMM

The classical setup of optimal transport is to solve

$$W(\mathbf{p},\mathbf{q}) = \min_{Z\in\Pi(\mathbf{p},\mathbf{q})} \langle Z, M\rangle$$

where $\mathbf{p},\mathbf{q}$ belongs to the probabilistic simplex $\Delta_{m_1}, \Delta_{m_2}$, the coupling set $\Pi(\mathbf{p},\mathbf{q}) = \{Z \in \mathbb{R}_+^{m_1\times m_2} : Z\cdot \mathbb{1}_{m_2} = \mathbf{p}; Z^T\cdot \mathbb{1}_{m_1} = \mathbf{q};\}$ and let $M \in \mathbb{R}_+^{m_1\times m_2}$ be the matrix of costs.

Our goal: find an efficient approximate solution $\widetilde{Z}$ with some justifications.
Previous work: Entropic Regularization (e.g. [Cuturi 2013] [Cuturi & Doucet 2014] [Benamou et al. 2015] [Cuturi & Peyré 2016])

ADMM-type treatment

Rewrite the problem as

$$W(\mathbf{p},\mathbf{q}) = \min_{\substack{Z_1\in\Pi_1(\mathbf{p})\\ Z_2\in\Pi_2(\mathbf{q})}} \langle Z_1, M\rangle \quad \text{s.t.} \underbrace{Z_1 = Z_2}_{\Lambda\ :\ \text{multiplier}}$$

where

$$\Pi_1(\mathbf{p}) = \{Z \in \mathbb{R}^{m_1\times m_2} : Z\cdot \mathbb{1}_{m_2} = \mathbf{p};\}$$

and

$$\Pi_2(\mathbf{q}) = \{Z \in \mathbb{R}^{m_1\times m_2} : Z^T\cdot \mathbb{1}_{m_1} = \mathbf{q};\}.$$

Indeed, one can further convert this into a saddle point formulation w.r.t. (Z_1, Z_2) and Λ and use the off-the-shelf dimension-free algorithms (e.g. mirror descent or mirror prox algorithms). But the speed is practically slow and the per-iteration computational cost is high.

Bregman ADMM [Wang & Banerjee 2014]

Iterations:

$$
\begin{aligned}
Z_1 &:= \underset{Z_1 \in \Pi_1(\mathbf{p})}{\arg\min} \langle Z_1, M \rangle + \langle \Lambda, Z_1 \rangle + \underbrace{\rho \cdot \mathrm{KL}(Z_1, Z_2)}_{\text{replace } |\cdot|^2 \text{ with } B_\Phi(\cdot,\cdot)} \\
Z_2 &:= \underset{Z_2 \in \Pi_2(\mathbf{q})}{\arg\min} -\langle \Lambda, Z_2 \rangle + \rho \cdot \mathrm{KL}(Z_2, Z_1) \\
\Lambda &:= \Lambda + \rho(Z_1 - Z_2)
\end{aligned}
$$

Here we recommend $\rho = \rho_0 \cdot \mathrm{median}(M)$ and $\rho_0 \in [1, 10]$.

Bregman ADMM [Wang & Banerjee 2014]

Implementation (use caching to avoid repeated calculation of $\exp(\cdot)$):

$$
\begin{aligned}
Z_1 &:= Z_2 \odot \exp\left\{\frac{M+\Lambda}{\rho}\right\} \\
Z_1 &:= P_{\Pi_1(\mathbf{p})}(Z_1) \\
Z_2 &:= Z_1 \odot \exp\left\{\frac{-\Lambda}{\rho}\right\} \\
Z_2 &:= P_{\Pi_2(\mathbf{q})}(Z_2) \\
\Lambda &:= \Lambda + \rho(Z_1 - Z_2)
\end{aligned}
$$

Convergence [Wang & Banerjee 2014]

The main result presents the convergence of iterative solutions: Let $D(W^*, W^t) = \mathrm{KL}(Z^*, Z_2^t) + \frac{1}{\rho^2}\|\Lambda^* - \Lambda^t\|^2$,

$$KL(Z_1^{t+1}, Z_2^t) \leq \underbrace{D(W^*, W^t)}_{\text{monotonic nonincreasing}} - D(W^*, W^{t+1})$$

And it also presents guaranteed optimality:

$$\langle \bar{Z}_1^T, M\rangle - \langle Z^*, M\rangle \leq \frac{\rho \mathrm{KL}(Z^*, Z_2^0)}{T},$$

$$\|\bar{Z}_1^T - \bar{Z}_2^T\|^2 \leq \frac{2D(W^*, W^0)}{T},$$

where $\bar{Z}_j^T = \frac{1}{T}\sum_{t=1}^T Z_j^t$, $j = 1, 2$.

Algorithm for Computing W- Barycenter [Ye et al. 2017a]

Consider we are to find a barycenter probability $\mathbf{q} \in \Delta_m$ along with N OTs to solve together. The problem becomes

$$\min_{\mathbf{q}} \sum_{i=1}^{N} W(\mathbf{p}^{(i)}, \mathbf{q})$$

You come with a similar reformulation:

$$W(\mathbf{p}^{(k)}, \mathbf{q}) = \min_{\substack{Z_1^{(k)} \in \Pi_1(\mathbf{p}^{(k)}) \\ Z_2^{(k)} \in \Pi_2(\mathbf{q})}} \langle Z_1^{(k)}, M^{(k)} \rangle \quad \text{s.t.} \underbrace{Z_1^{(k)} = Z_2^{(k)}}_{\Lambda^{(k)} \text{ : multiplier}}$$

Iterations:

$$
\begin{aligned}
Z_1^{(k)} &:= Z_2^{(k)} \odot \exp\left\{\frac{M^{(k)} + \Lambda^{(k)}}{\rho}\right\} \\
Z_1^{(k)} &:= P_{\Pi_1(\mathbf{p}^{(k)})}(Z_1^{(k)}) \\
Z_2^{(k)} &:= Z_1^{(k)} \odot \exp\left\{\frac{-\Lambda^{(k)}}{\rho}\right\} \\
\mathbf{q}^{(k)} &:= P_{\Delta_{m_2}}\left((Z_2^{(k)})^T \cdot \mathbb{1}_{m_1}\right) \\
\mathbf{q}_i^p &:\propto \frac{1}{N}\sum_{k=1}^{N}(\mathbf{q}_i^{(k)})^p, \quad p = 1 \text{ or } \frac{1}{2} \text{ (a heuristic!)} \\
Z_2^{(k)} &:= P_{\Pi_2(\mathbf{q})}(Z_2^{(k)}) \\
\Lambda^{(k)} &:= \Lambda^{(k)} + \rho(Z_1^{(k)} - Z_2^{(k)})
\end{aligned}
$$

Discrete Distribution Clustering [Ye et al. 2017a]

D2-Clustering uses 2-Wasserstein barycenter as a kind of "mean" in K-means clustering with the following remarks:

- Given the Wasserstein distance is a metric, one can use **triangle inequality** to prune a significant portion of pairwise distance calculations [Elkan 2003].
- Every outer iteration, the B-ADMM algorithm (to compute the barycenter) warmly starts from the previous Z_1, Z_2 but reset $\Lambda = 0$.
- A finite number of iterations are used for each B-ADMM loop.
- The support points of Wasserstein barycenter are updated (least square estimates) simultaneously every τ B-ADMM iterations using computed Z_1 (e.g. $\tau = 10, 50$).
- The size of support points of Wasserstein barycenter is set to be small.

Document Clustering [Ye et al. 2017b]

Idea: every document is a weighted collection of words. Every word has a Euclidean embedding vector representing its semantic meaning.

- Take advantages of both BoW representation and Word Embedding, yet ignoring the order of words.
- No vector representation for each document.
- Perform robustly well even with a randomly sampled word embedding.

Best AMIs of compared methods on different datasets and their averaging. The best results are marked in bold font for each dataset, the 2nd and 3rd are marked by blue and magenta colors respectively.

	regular				domain-specific		
	BBCNews abstract	Wik events	Reuters	Newsgroups	BBCSport	Ohsumed	Avg.
Tfidf-N	0.389	0.448	0.470	0.388	**0.883**	0.210	0.465
Tfidf	0.376	0.446	0.456	0.417	0.799	0.235	0.455
Laplacian	0.538	0.395	0.448	0.385	0.855	0.223	0.474
LSI	0.454	0.379	0.400	0.398	0.840	0.222	0.448
LPP	0.521	0.462	0.426	**0.515**	0.859	**0.284**	0.511
NMF	0.537	0.395	0.438	0.453	0.809	0.226	0.476
LDA	0.151	0.280	0.503	0.288	0.616	0.132	0.328
AvgDoc	**0.753**	0.312	0.413	0.376	0.504	0.172	0.422
PV	0.428	0.289	0.471	0.275	0.553	0.233	0.375
D2C (*)	**0.759**	**0.545**	**0.534**	0.493	0.812	0.260	**0.567**

Comparison between *random* word embeddings (upper row) and meaningful *pre-trained* word embeddings (lower row), based on their best ARI, AMI, and V-measures. The improvements by percentiles are also shown in the subscripts.

	ARI	AMI	V-measure
BBCNews	.146	.187	.190
abstract	$.792_{+442\%}$	$.759_{+306\%}$	$.762_{+301\%}$
Wiki events	.194	.369	.463
	$.277_{+43\%}$	$.545_{+48\%}$	$.611_{+32\%}$
Reuters	.498	.524	.588
	$.515_{+3\%}$	$.534_{+2\%}$	$.594_{+1\%}$
Newsgroups	.194	.358	.390
	$.305_{+57\%}$	$.493_{+38\%}$	$.499_{+28\%}$
BBCSport	.755	.740	.760
	$.801_{+6\%}$	$.812_{+10\%}$	$.817_{+8\%}$
Ohsumed	.080	.204	.292
	$.116_{+45\%}$	$.260_{+27\%}$	$.349_{+20\%}$

The Dual Formulation of OT

Consider the following LP problem,

$$W(\mathbf{p},\mathbf{q}) = \max_{\mathbf{f}\in\Omega(M)} \langle \mathbf{p},\mathbf{g}\rangle - \langle \mathbf{q},\mathbf{h}\rangle \ .$$

where

$$\Omega(M) \stackrel{\text{def.}}{=} \Big\{ \mathbf{f} = [\mathbf{g};\mathbf{h}] \in \mathbb{R}^{m_1+m_2} \ \Big| \ -C_M < g_i - h_j \le M_{i,j}, 1 \le i \le m_1, 1 \le j \le m_2 \Big\} \ .$$

For sufficiently large C_M, $W_{dual}(\mathbf{p},\mathbf{q}) = W_{primal}(\mathbf{p},\mathbf{q})$ if $\mathbf{p},\mathbf{q}$ are strictly positive.

The Boltzmann distribution

Let $\Omega_0(M) = \{\mathbf{f} = [\mathbf{g}; \mathbf{h}] \in \Omega(M) \mid g_1 = 0\}$, consider a probability density on $\Omega_0 \in \mathbb{R}^{m_1+m_2-1}$ with

$$p(\mathbf{f}; \mathbf{p}, \mathbf{q}) \propto \exp\left[\frac{1}{T}\left(\langle \mathbf{p}, \mathbf{g}\rangle - \langle \mathbf{q}, \mathbf{h}\rangle\right)\right],$$

The samples from the Boltzmann distribution will eventually concentrate at the optimum set of its deriving problem (*e.g.* $W_{dual}(\mathbf{p}, \mathbf{q})$) as $T \to 0$.

Given any $\mathbf{f} = (\mathbf{g}; \mathbf{h}) \in \Omega_0(M)$ and any $C_M > 0$, we have for any i and j,

$$g_i \leq U_i(\mathbf{h}) \stackrel{\text{def.}}{=} \min_{1 \leq j \leq m_2} (M_{i,j} + h_j),$$

$$h_j \geq L_j(\mathbf{g}) \stackrel{\text{def.}}{=} \max_{1 \leq i \leq m_1} (g_i - M_{i,j}).$$

and

$$g_i > \widehat{L}_i(\mathbf{h}) \stackrel{\text{def.}}{=} \max_{1 \leq j \leq m_2} (-C_M + h_j),$$

$$h_j < \widehat{U}_j(\mathbf{g}) \stackrel{\text{def.}}{=} \max_{1 \leq i \leq m_1} (C_M + g_i).$$

Here $U_i = U_i(\mathbf{h})$ and $L_j = L_j(\mathbf{g})$ are auxiliary variables. g_i's are conditionally independent given $\mathbf{h}$, and likewise h_j's are also conditionally independent given $\mathbf{g}$.

Furthermore, each of their conditional probabilities within its feasible region (subject to C_M) satisfies

$$p(g_i|\mathbf{h}) \propto \exp\left(\frac{g_i p_i}{T}\right),\ \widehat{L}_i(\mathbf{h}) < g_i \leq U_i(\mathbf{h}),$$

$$p(h_j|\mathbf{g}) \propto \exp\left(-\frac{h_j q_j}{T}\right),\ L_j(\mathbf{g}) \leq h_j < \widehat{U}_j(\mathbf{g}),$$

where $2 \leq i \leq m_1$ and $1 \leq j \leq m_2$. As $C_M \to +\infty$, $\widehat{U}_j(\mathbf{g}) \to +\infty$ and $\widehat{L}_i(\mathbf{h}) \to -\infty$.

The Gibbs-OT Sampler [Ye, Wang, Li 2016]

Given $\mathbf{f}^{(0)} \in \Omega_0(M)$, $\mathbf{p} \in \Delta_{m_1}$ and $\mathbf{q} \in \Delta_{m_2}$, and $T^{(1)}, \ldots, T^{(2N)} > 0$, for $t = 1, \ldots, N$, we define the following Markov chain

1. Randomly sample

$$\theta_1, \ldots, \theta_{m_2} \overset{i.i.d.}{\sim} \text{Exponential}(1).$$

For $j = 1, 2, \ldots, m_2$, let

$$\begin{cases} L_j^{(t)} := & \max_{1 \le i \le m_1} \left(g_i^{(t-1)} - M_{i,j} \right) \\ h_j^{(t)} := & L_j^{(t)} + \theta_j \cdot T^{(2t-1)} / q_j \end{cases}$$

2. Randomly sample

$$\theta_2, \ldots, \theta_{m_1} \overset{i.i.d.}{\sim} \text{Exponential}(1).$$

For $i = (1), 2, \ldots, m_1$, let

$$\begin{cases} U_i^{(t)} := & \min_{1 \le j \le m_2} \left(M_{i,j} + h_j^{(t)} \right) \\ g_i^{(t)} := & U_i^{(t)} - \theta_i \cdot T^{(2t)} / p_i \end{cases}$$

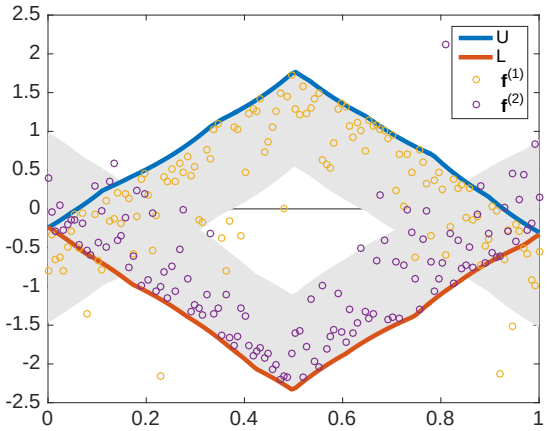
Wasserstein Loss Minimization

Consider the following loss function to minimize w.r.t θ:

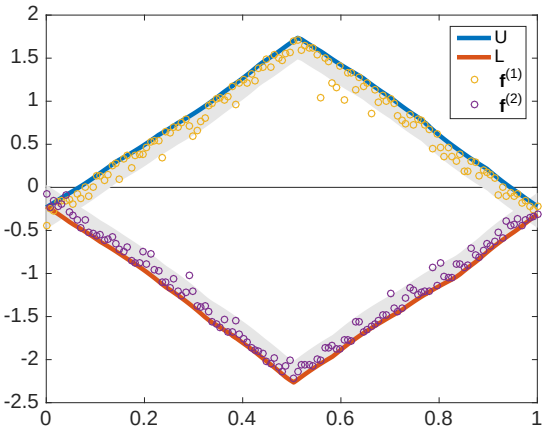
$$\mathfrak{R}(\theta) := \sum_{i=1}^{|\mathcal{D}|} W(\mathbf{p}_i(\theta), \mathbf{q}_i(\theta))$$

Define auxiliary function $V(\mathbf{x}, \mathbf{y}) \stackrel{\text{def.}}{=} \langle \mathbf{p}, \mathbf{x} \rangle - \langle \mathbf{q}, \mathbf{y} \rangle$. To minimize the Wasserstein losses $W(\mathbf{p}, \mathbf{q})$ approximately in such WLMs, we propose to instead optimize its **asymptotically consistent** upper bound $\mathbb{E}[V(\mathbf{U}, \mathbf{V})]$ at equilibrium of Boltzmann distribution $p(\mathbf{f}; \mathbf{p}, \mathbf{q})$ using its stochastic sub-gradients: $\mathbf{U} \in \partial V(\mathbf{U}, \mathbf{V}))/\partial \mathbf{p}$ and $-\mathbf{L} \in \partial V(\mathbf{U}, \mathbf{V}))/\partial \mathbf{q}$.

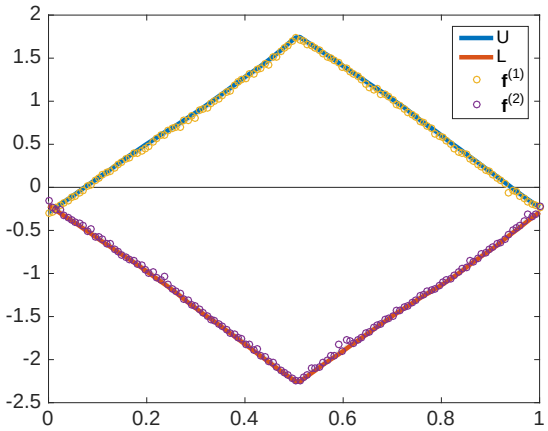
An illustrative example of a simple 1D optimal transportation problem with Coulomb cost



(a) 20 iterations



(b) 40 iterations



(c) 60 iterations

Some Discussions

- (Intuitively) A suitable annealing schedule should let $V(\mathbf{U}, \mathbf{V})$ be supermartingale, which gives a (calculable) upper bound for temperature T at every iteration.
- An approximate primal solution can be quickly recovered from MCMC samples. Similarly, the procedure can be naturally extended to estimate barycentric mapping at zero-mass points using MCMC (related to [Perrot et al. 2016]).
- The analysis of $\mathbf{g}, \mathbf{h}$ chain can be achieved by the analysis of $\mathbf{U}, \mathbf{V}$ chain. The transitive kernel underpinning $\mathbf{U}, \mathbf{V}$ chain admits a closed form, and can be of interest (e.g. for constructing Hilbert spaces for Lipschitz continuous functions).
- The solution obtained from Gibbs-OT seems to possess very different properties than other regularized approaches.

Problem Setup

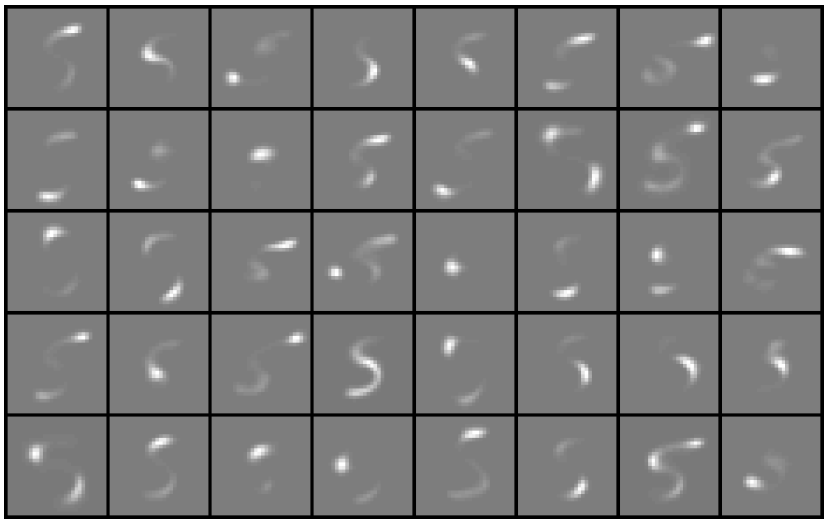
The Wasserstein NMF models each distribution as a linear additive combination of K base distributions:

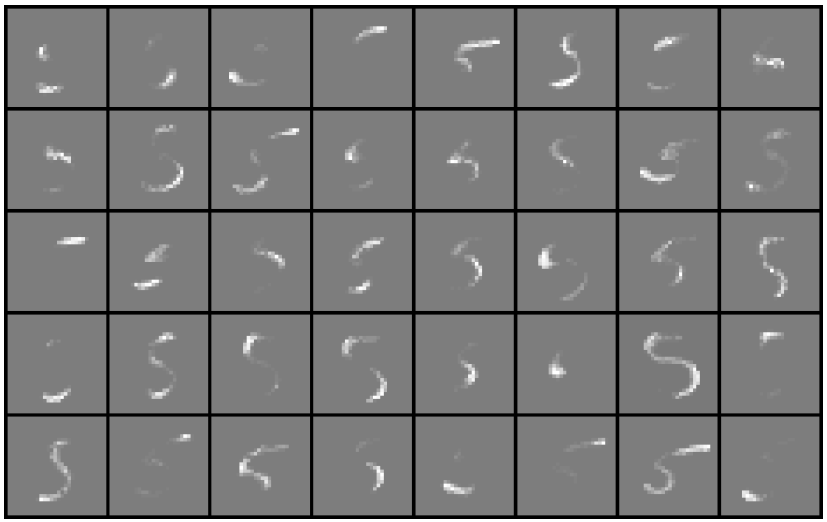
$$\min_{\{\beta^{(i)}\},\{\Psi_k\}} \sum_{i=1}^{n} W(\Phi_i, \sum_{k=1}^{K} \beta_k^{(i)} \Psi_k),$$

where we solve linear coefficients $\{\beta^{(i)} \in \Delta_k\}$ and base distributions $\{\Psi_k\}$.

Graphical Comparison of Base Distributions

Left: Entropic Regularization [Rolet & Cuturi 2016];
Right: Gibbs-OT Sampler [Ye, Wang, Li 2016]





Graphical Comparison of Base Distributions

Left: Entropic Regularization [Rolet & Cuturi 2016];
Right: Gibbs-OT Sampler [Ye, Wang, Li 2016]