Mirror Descent on Relaxed Optimal Transport

Su Xun

笠井研究室 修士 2 年

2022 年 7 月 20 日

Outline

- UOT and NMF/Inverse Problem
- Mirror Descent and Acceleration
- Algorithms for ROT and SROT
- Experiments
- Ideas and Future Plan

Background of OT

Optimal Transport

$$W(\alpha, \beta) := \min_{\mathbf{T} \in \mathbb{R}_{+}^{n \times n}} \langle \mathbf{C}, \mathbf{T} \rangle$$

$$\mathbf{T}\mathbb{1} = \alpha, \mathbf{T}^T \mathbb{1} = \beta, \mathbf{T}_{ij} > 0$$

- Applications on GAN, Retrieving information, Domain adaptation and so on.

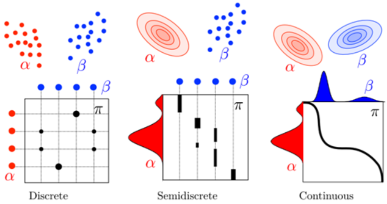


Figure: Different forms of Optimal Transport

Inbalanced OT problem

Relaxed Optimal Transport

$$W(\alpha, \beta) := \min_{\mathbf{T} \in \mathbb{R}_{+}^{n \times n}} \langle \mathbf{C}, \mathbf{T} \rangle + D(\mathbf{T}\mathbb{1}, \alpha) + D(\mathbf{T}^T \mathbb{1}, \beta)$$

- Optimal Transport could only deal with balanced samples, a relaxed version is required for a more general application.
- the most famous ROT solver is the Sinkhorn, which add an entropy part $\eta H(\mathbf{T})$ onto the problem and has a complexity $O(\frac{n^2}{\epsilon})$ [Pham et al., 2020]

Background of OT

- Relaxed Optimal Transport

$$f(t) = c^T t + D(At, b), t \in \mathbb{R}^{n^2}$$

- Regularized Non-negative Matrix Factorization

$$f(t) = R(t) + D(At, b)$$

- Semi-relaxed Optimal Transport

$$f(t) = c^T t + D(At, b), Bt = a$$

- D is a divergence.

Bregman Proximal Descent

Composite convex problem

$$\min_{x \in \mathbb{R}^n} \{f(x) + \psi(x)\}$$

$f(x)$ is convex and differentiable and $\psi(x)$ is convex.

- Proximal Gradient

$$\mathrm{Prox}_\gamma(x) := \operatorname*{argmin}_{z \in \mathbb{R}^n} \left\{ \frac{1}{2\gamma} \|z - x\|^2 + \psi(z) \right\}$$

- Bregman Proximal

$$\mathrm{Prox}_\gamma(x) := \operatorname*{argmin}_{z \in \mathbb{R}^n} \left\{ \frac{1}{2\gamma} D(z, x) + \psi(z) \right\}$$

Bregman Proximal Descent

Mirror Descent

$$t_k = \mathrm{Prox}_{t,\psi}(t_{k-1} - \alpha_k \nabla f(t_{k-1}))$$

- Accelerated Proximal Gradient with nestrov

$$y_k = t_k + \beta_k(x_k - x_{k-1})$$

$$t_{k+1} = \mathrm{Prox}_{\gamma,\psi}(y_k - \alpha_k \nabla f(y_k))$$

Benefits of Bregman Proximal Descent

- The Bregman proximal template provides a lot more flexibility.
- Cheap Projection When the constrain is simplex, using KL divergence could attain a closed form projection.
- Better convergence.

$$f\left(x_k\right)+\psi\left(x_k\right)-(f(\bar{x})+\psi(\bar{x})) \leq \frac{L \cdot \left\|x_0-\bar{x}\right\|^2}{2k}$$

$$f\left(x_k\right)+\psi\left(x_k\right)-(f(\bar{x})+\psi(\bar{x})) \leq \frac{L D_h\left(\bar{x}, x_0\right)}{k}.$$

- $L-Relatively\ smoothness$ holds more broadly than $L-smoothness$

BPD on ROT

Composite convex problem

$$\min_{x\in\mathbb{R}^n}\{f(x)+\psi(x)\}$$

- On the Regularizor

$$f(x)=D(At,b),\psi(x)=c^Tt$$

- On the constrain

$$f(x)=D(At,b)+c^Tt,\psi(x)=\begin{cases}1, & Bt=a\\ 0, & Bt\neq a\end{cases}$$

- The choice of $\psi(x)$ had better to be sure there exists a closed form solution.

ROT

Relaxed Optimal Transport

$$W(a,b) = \min_{t \in \mathbb{R}_+^{n^2}} c^\mathsf{T} t + \tau D_h(Mt, b) + \tau D_h(Nt, a)$$

The update formula for mirror descent

$$\operatorname{prox}_{h,\gamma}(x_k - \gamma \nabla f(x_k)) = \frac{x_k}{e^{\gamma(\frac{c}{\tau} + \nabla f(x_k))}}$$

reorganize it as:

$$T_{k+1} = (\operatorname{diag}(\frac{a}{T_k \mathbb{1}}))^\gamma (T_k \odot \exp(-\frac{\gamma}{\tau} c))(\operatorname{diag}(\frac{b}{T_k^\mathsf{T} \mathbb{1}}))^\gamma$$

when $\gamma = \frac{1}{L} = \frac{1}{2}$, it is equal to the method in [Chapel et al., 2021], however, large τ causes a really slow convergence speed, which could be alleviated by gradually increasing τ

SROT

Semi-relaxed Optimal Transport

$$W(a,b)_{Nt=a} = \min_{t \in \mathbb{R}_+^{n^2}} c^\top t + \tau D_h(Mt, b)$$

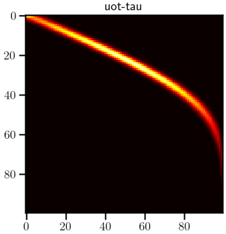
We could use Bregman projection with nestrov acceleration.

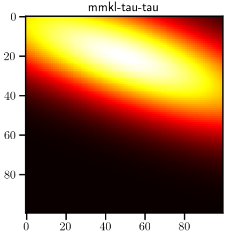
$$y_k = t_k \times \exp\left(-\alpha \nabla (f(t_k) + \tau D_h(Mt_k, b))\right) \tag{1}$$

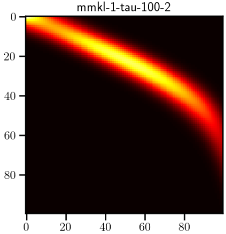
$$x_{k+1} = y_k - \frac{t_k - 1}{t_{k+1}}(y_k - x_{k-1}) \tag{2}$$

This could speed up the projection process from $O(n \ln n)$ to $O(n)$.

Gaussian Transport(ROT)







Convergence Speed(ROT)

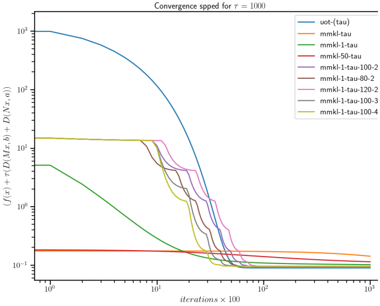


Figure: Marginal error

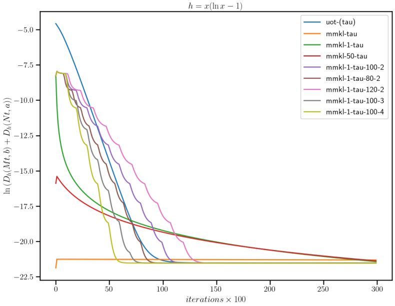
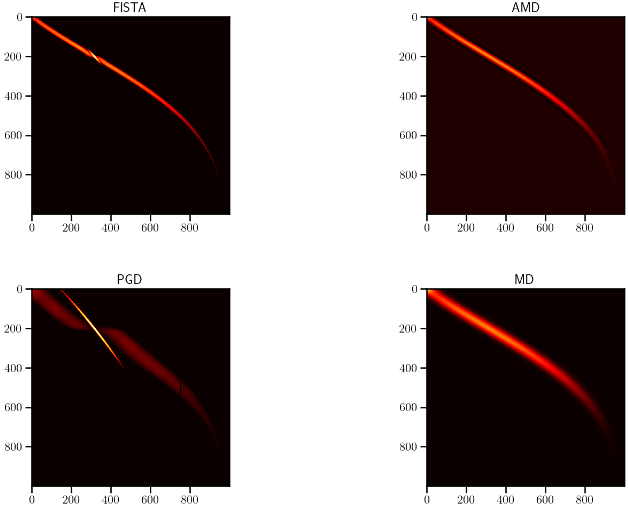


Figure: ROT function

Gaussian Transport(SROT)



Convergence Speed(SROT)

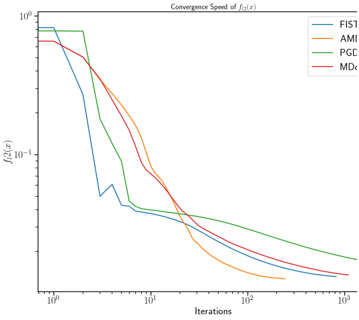


Figure: Iterations

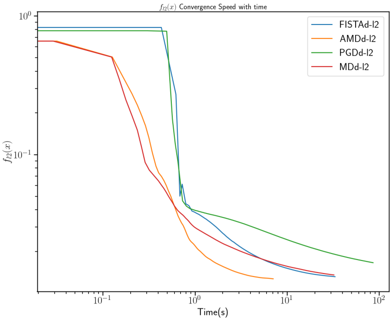


Figure: Time

Outcomes and Questions

- The τ largely influenced the convergence speed of the ROT problem, doubling could speed up the computational process, however, it is unclear how to change the τ and a frequent change of τ causes computational burden. Changing τ also make nestrov less effective.
- The same dynamic τ method didn't work on L_2 ROT problems, which is quite different from KL ROT.
- It is possible to apply Bregman proximal onto the SROT problem and use Bregman projection at the same time.

Future Plan

- Theoretically analyzing on the convergence of the method and find a reasonable convergence scheme.
- Analyzing the nestrov and why it failed on a dynamic τ scheme.
- Expanding the experiments on L_2 and other condition.

References I

- Chapel, L., Flamary, R., Wu, H., F 辿 votte, C., and Gasso, G. (2021).
  Unbalanced optimal transport through non-negative penalized linear regression.
- Pham, K., Le, K., Ho, N., Pham, T., and Bui, H. (2020).
  On unbalanced optimal transport: An analysis of sinkhorn algorithm.
  CoRR, abs/2002.03293.

ご清聴ありがとうございました.

Thank you for listening.