Mirror Descent on Unbalanced Optimal Transport and Acceleration

Su Xun

笠井研究室 修士 2 年

2022 年 9 月 1 日

Outline

- Backgrounds of Optimal Transport and Unbalanced Optimal Transport problems
- The Lasso problem and Mirror Descent Algorithms
- Acceleration
- Shifting projection
- Prospect and Plan

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.

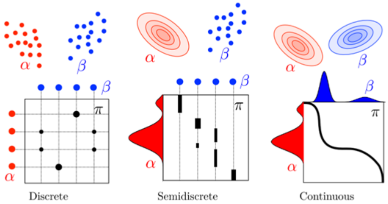

Figure: Different forms of Optimal Transport

Unbalanced Optimal Transport (UOT)

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

- Optimal Transport can only deal with balanced samples, a relaxed version with divergence function D_h is required for more general applications.
- the most famous UOT solver is the Sinkhorn, which uses Kullback-Leibler divergence to penalize and add an entropy part $\eta H(\mathbf{T})$ onto the problem, its complexity is $O(\frac{n^2}{\epsilon})$ [Pham et al., 2020]
- It is natural to consider whether other powerful optimizers exist.

Lasso problem and Mirror Descent

- UOT has a similar structure to the Lasso problem:

$$f(t) = g(t) + D_h(Xt, b), t \in \mathbb{R}^{n^2}$$

- Lasso Problem:

$$f(t) = \lambda\|t\| + \|Xt - b\|_2^2$$

- L_2 or Kullback-Leibler divergence penalized UOT

$$f(t) = \lambda c^\mathsf{T} t + \|Xt - b\|_2^2$$
$$f(t) = \lambda c^\mathsf{T} t + KL(Xt, b)$$

$b = [\alpha^\mathsf{T} \quad \beta^\mathsf{T}]^\mathsf{T}$ and X, for example, when $n = 3$, is:

$$X = \begin{pmatrix} 1 & 1 & 1 & & & & & & \\ & & & 1 & 1 & 1 & & & \\ & & & & & & 1 & 1 & 1 \\ 1 & & & 1 & & & 1 & & \\ & 1 & & & 1 & & & 1 & \\ & & 1 & & & 1 & & & 1 \end{pmatrix} \tag{1}$$

Lasso problem and Mirror Descent

The problem with a similar structure is suitable for Mirror Descent Algorithm.

Composite convex problem

$$\min_{x\in\mathbb{R}^n}\{d(x)+g(x)\}$$

$d(x)$ is convex and differentiable and $g(x)$ is convex.

- Proximal Gradient

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

- Bregman Proximal

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

Lasso problem and Mirror Descent

Composite convex problem

$$\min_{x \in \mathbb{R}^n} \{d(x) + g(x)\}$$

- For Lasso:

$$d(x) = \|At - b\|_2^2, \quad g(x) = \lambda \|t\|$$

- For L_2 penalized UOT:

$$d(x) = \|At - b\|_2^2, \quad g(x) = \lambda c^\mathsf{T} t$$

- For Kullback-Leibler divergence penalized UOT:

$$d(x) = KL(At, b), \quad g(x) = \lambda c^\mathsf{T} t$$

Lasso problem and Mirror Descent

- For L_2 UOT, we get

$$T^{k+1} = \max(T^k - \gamma(T^k \mathbb{1}\mathbb{1}^\mathsf{T} + \mathbb{1}\mathbb{1}^\mathsf{T} T^k) - \alpha \mathbb{1}^\mathsf{T} + \mathbb{1}\beta^\mathsf{T} - \lambda \mathbf{C}, 0)$$

  It is the ISTA algorithm.

- For KL UOT, we get

$$T_{k+1} = (\mathrm{diag}\,(\frac{\alpha}{T_k \mathbb{1}}))^\gamma (T_k \odot \exp\,(-\gamma\lambda\mathbf{C}))(\mathrm{diag}\,(\frac{\beta}{T_k^\mathsf{T} \mathbb{1}}))^\gamma$$

  when $\gamma = \frac{1}{L} = \frac{1}{2}$ ($f(x)$ is an L-strongly convex function), it is equal to the Majorization-Minimization (MM) algorithm in [Chapel et al., 2021].

Acceleration Methods

- Lucky, we can borrow the accelerating methods from the Lasso problem:
  - Nesterov Acceleration
  - Path-following Algorithm [Tibshirani and Taylor, 2011]
  - Screening [Ghaoui et al., 2010]
- I am focusing on the Screening method and I revised this method for the UOT problem to take advantage of its sparse X matrix, which is rare in the normal Lasso problem.

Screening

Motivation:

Lasso-like regularizations cause a sparse solution $\mathrm{card}(t_{ij} \| t_{ij} = 0) \approx n^2$, for $t \in \mathbb{R}^{n \times n}$. We identify the elements equal to zero theoretically and freeze them to save computational time.

- As the UOT could be regarded as a Lasso-like problem, this technology can handle UOT as well.

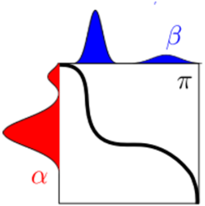

Figure: The typical sparse solution for OT problem

Screening

Dynamic Screening Framework [Yamada and Yamada, 2021]

$$P(t) = \min_t d(Xt) + g(t)$$

$$D(\theta) = \max_\theta -d^*(-\theta) + g^*(X^T\theta)$$

- $d(Xt)$ is the distance measure like L_2 function and KL divergence.
- $g(\beta)$ is the Lasso-like sparse regularization such as L_1 penalty or optimal transport problem, we can convert it to constraints, then the dual problem is:

$$D(\theta) = \max_\theta -d^*(-\theta)$$
$$\text{s.t.} \quad \forall i, \quad h_i(\theta) \le 0$$

Screening

- Relying on the KKT condition, we can assert the existence of a series of dual constraints $h_i(\theta)$, that for optimum $\hat{\theta}$, if $h_i(\hat{\theta}) < 0$, then $t_i = 0$.
- For Lasso, the dual constraints are:

$$h_i(\theta) = \|x_i^T \theta\| - 1 \leq 0$$

- For UOT, the dual constraints are:

$$h_i(\theta) = x_i^T \theta - c_i \leq 0$$

- For $\hat{\theta}$, if the the $\leq$ could be replaced by $<$, then we have $\hat{t_i} = 0$
- However, we don't know the value of the optimum solution $\hat{\theta}$ at first.

Screening

- If we can find a $\tilde{\theta}$ that satisefied with the dual constrains, then we can construct an area $R^{DS}(\tilde{\theta})$, for L_2 penalized problem:

$$\frac{1}{2}\|\theta - \tilde{\theta}\|_2^2 + D(\tilde{\theta}) \leq D(\theta) \leq -d^*(-\theta)$$
$$\text{s.t. } g(\tilde{t}) - \theta^T X \tilde{t} < 0$$

- The left part is strongly concave inequality and the right part is the dual inequality.
- This area contains $\tilde{\theta}$ and the optimum $\hat{\theta}$
- If we could prove the $\max_{\theta \in R^{DS}(\tilde{\theta})} h_i(\theta) < 0$, then it holds for the optimum $\hat{\theta}$. It indicates that $h_i(\hat{\theta}) < 0$, and the element i of the primal optimal solution $\hat{t}$ must be zero.

Screening

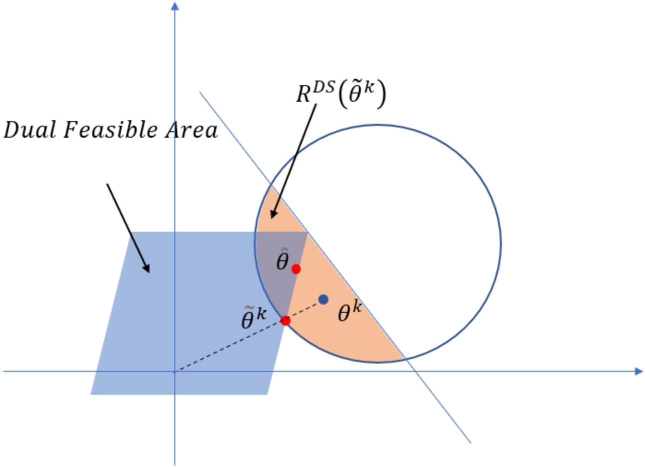

Figure: Projection in Screening

Screening

- We can dynamically compute an approximate solution θ^k by any algorithm and project it onto the dual constraints as $\tilde{\theta}^k$.
- We hope the projected $\tilde{\theta}^k$ could be closed enough to $\hat{\theta}$ to produce a smaller $R^{DS}(\tilde{\theta}^k)$
- A smaller area can help us screen more variables as

$$\max_{\theta \in \tilde{R} \in R^{DS}(\tilde{\theta})} \|x_i^T \theta\| \leq \max_{\theta \in R^{DS}(\tilde{\theta})} \|x_i^T \theta\|$$

always holds.

Projection methods

- The Lasso method is to shrink all $\tilde{\theta}$ together.

$$\tilde{\theta} = \frac{\theta}{\|\frac{X^T\theta}{c}\|_\infty}$$

- It is not suitable for the UOT problem as the cost value c_i might be small and even zero.
- We propose to use a shifting method, as the x_i has a specific sparse structure which could rewrite the problem as:

$$\theta_{i_1} + \theta_{i_1} < c_i$$

we decide to shift θ_j according to the maximum positive difference of $\frac{\theta_{i_1}+\theta_{i_1}-c_i}{2}$

Screening

Shifting Screening method:

$$\tilde{\theta}_i = \begin{cases} \theta_i - \max\limits_{j \bmod n = i} \left(\dfrac{\theta_{j_1} + \theta_{j_2} - c_j}{2}\right) & 0 \le i < n \\ \theta_i - \max\limits_{in \le j < i(n+1)} \left(\dfrac{\theta_{j_1} + \theta_{j_2} - c_j}{2}\right) & n \le i < 2n \end{cases}$$

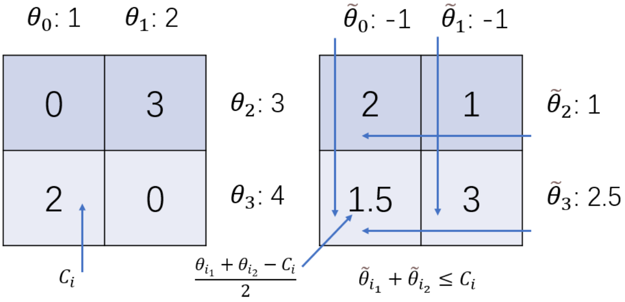

Figure: Shifting on a 2×2 matrix

Screening

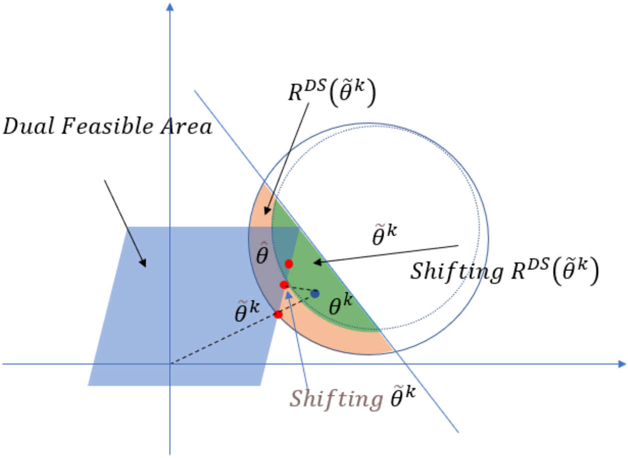

Figure: Difference of the projection method in Screening

Screening

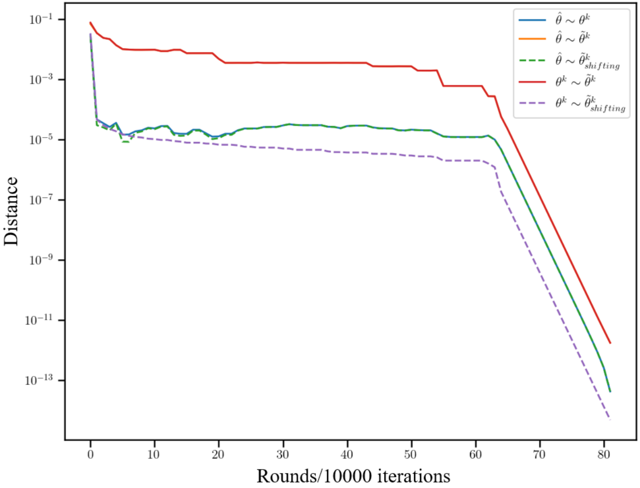

Figure: Distance between the projected point with $\hat{\theta}$ or θ^k

Screening

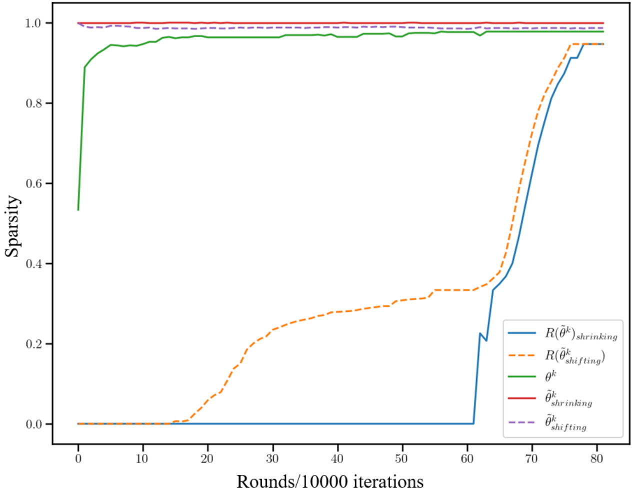

Figure: The Screening Ratio

Potential and defects

- The UOT problem has the potential to screen out better due to its specific sparse structure of matrix X
- Screening is irrelevant to the optimization method you use and especially effective for the MM algorithm (which could be regarded as one kind of Mirror Descent)
- KL penalized Lasso problem also has a screening method [Dantas et al., 2021], which could be applied to the KL penalized UOT problem, we might accelerate Sinkhorn Algorithm, which is only suitable for KL penalized UOT, with the Screening method.
- However, Screening needs too many iterations to start even after the revision. There might exist a better method to find a smaller area for the UOT problem

Future Plan

- Thinking of revising the screening method from the perspective of constructing area.
- Combining the Screening method with Mirror descent and other algorithms to test its speed-up ratio.
- Generalizing the screening method to KL penalized the UOT problem and the Sinkhorn Algorithm.

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.
- Dantas, C. F., Soubies, E., and F 辿 votte, C. (2021).
  Safe screening for sparse regression with the kullback-leibler divergence.
  In ICASSP 2021 - 2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pages 5544–5548.
- Ghaoui, L. E., Viallon, V., and Rabbani, T. (2010).
  Safe feature elimination for the lasso and sparse supervised learning problems.
  arXiv preprint arXiv:1009.4219.
- 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.

References II

- Tibshirani, R. J. and Taylor, J. (2011).
  The solution path of the generalized lasso.
  The Annals of Statistics, 39(3).
- Yamada, H. and Yamada, M. (2021).
  Dynamic sasvi: Strong safe screening for norm-regularized least squares.
  In Ranzato, M., Beygelzimer, A., Dauphin, Y., Liang, P., and Vaughan, J. W., editors, Advances in Neural Information Processing Systems, volume 34, pages 14645–14655. Curran Associates, Inc.

ご清聴ありがとうございました.

Thank you for listening.