Dynamic Screening for L_2 Penalied Unbalanced Optimal Transport

Su Xun

笠井研究室 修士 2 年

2022 年 10 月 5 日

Outline

- Backgrounds of Optimal Transport and Unbalanced Optimal Transport problems
- The Lasso Problem and Dynamic Screening
- Shifting Projection
- Two-Plane Screening
- Prospect and Plan

Optimal Transport

$$\mathrm{OT}(\mathbf{a}, \mathbf{b}) := \min_{\mathbf{T} \in \mathbb{R}_+^{m \times n}} \langle \mathbf{C}, \mathbf{T} \rangle$$
$$\mathbf{T}\mathbb{1}_m = \mathbf{a}, \mathbf{T}^T \mathbb{1}_n = \mathbf{b} \tag{1}$$

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

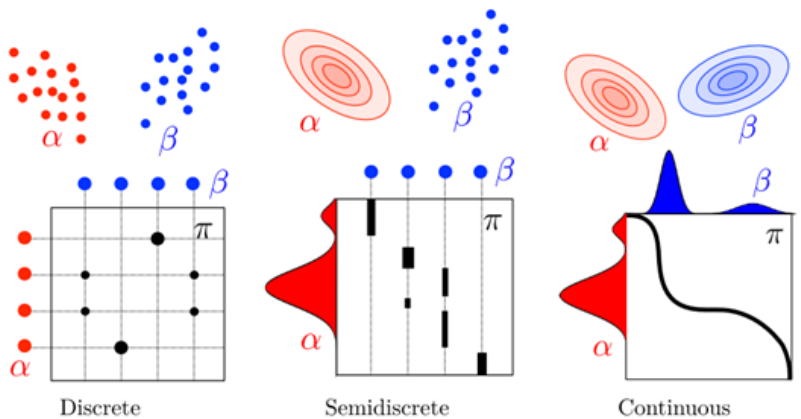

Figure: Different forms of Optimal Transport

Unbalanced Optimal Transport (UOT)

$$\mathrm{UOT}(\mathbf{a}, \mathbf{b}) := \min_{\mathbf{T} \in \mathbb{R}_+^{m \times n}, \lambda > 0} \lambda \langle \mathbf{C}, \mathbf{T} \rangle + D_h(\mathbf{T}\mathbb{1}_m, \mathbf{a}) + D_h(\mathbf{T}^\mathsf{T}\mathbb{1}_n, \mathbf{b}) \tag{2}$$

- 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.

The Lasso problem and the UOT problem

- UOT has a similar structure to the Lasso problem:

$$\min_{\mathbf{t}\in\mathbb{R}_+^{mn}} f(t) := \min_{\mathbf{t}\in\mathbb{R}_+^{mn}} \lambda \mathbf{c}^\mathsf{T}\mathbf{t} + D_h(\mathbf{X}\mathbf{t}, \mathbf{y})$$

- Lasso Problem:

$$\min_{\mathbf{t}\in\mathbb{R}^{mn}} f(t) := \min_{\mathbf{t}\in\mathbb{R}^{mn}} \lambda\|\mathbf{t}\| + D_h(\mathbf{X}\mathbf{t}, \mathbf{y}) \tag{3}$$

- L_2 or Kullback-Leibler divergence penalized UOT

$$f(t) := \lambda \mathbf{c}^\mathsf{T}\mathbf{t} + \|\mathbf{X}\mathbf{t} - \mathbf{y}\|_2^2$$
$$f(t) := \lambda \mathbf{c}^\mathsf{T}\mathbf{t} + KL(\mathbf{X}\mathbf{t}, \mathbf{y})$$

$y = [\mathbf{a}^\mathsf{T}\mathbf{b}^\mathsf{T}]^\mathsf{T}$ and $\mathbf{X}$, for example, when $n = 3$, is:

$$\mathbf{X} = \begin{pmatrix} 1 & 1 & 1 & & & & & & \\ & & & 1 & 1 & 1 & & & \\ & & & & & & 1 & 1 & 1 \\ 1 & & & 1 & & & 1 & & \\ & 1 & & & 1 & & & 1 & \\ & & 1 & & & 1 & & & 1 \end{pmatrix} \tag{4}$$

The Lasso Problem and Dynamic Screening

Composite convex problem

$$\min_{\mathbf{t}\in\mathbb{R}^n}\{d(\mathbf{t})+g(\mathbf{t})\} \tag{5}$$

- For Lasso:
$$d(\mathbf{t})=\|\mathbf{X}\mathbf{t}-\mathbf{y}\|_2^2, \quad g(\mathbf{t})=\lambda\|\mathbf{t}\|$$
- For L_2 penalized UOT:
$$d(\mathbf{t})=\|\mathbf{X}\mathbf{t}-\mathbf{y}\|_2^2, \quad g(\mathbf{t})=\lambda\mathbf{c}^\mathsf{T}\mathbf{t}$$
- For Kullback-Leibler divergence penalized UOT:
$$d(\mathbf{t})=KL(\mathbf{X}\mathbf{t},\mathbf{t}), \quad g(\mathbf{t})=\lambda\mathbf{c}^\mathsf{T}\mathbf{t}$$

Screening

Motivation:
Lasso-like regularizations cause a sparse solution $\mathrm{card}(\mathbf{T}_{ij} \| \mathbf{T}_{ij} = 0) \approx n^2$, for $\mathbf{T} \in \mathbb{R}^{m \times n}$. 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.

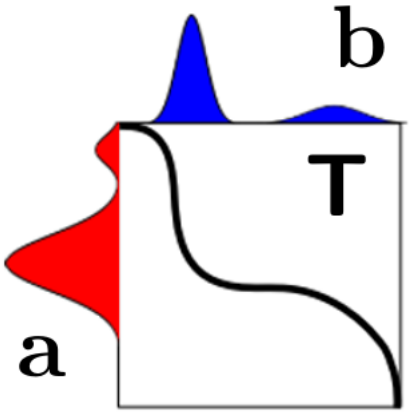

Figure: The typical sparse solution for OT problem

Dynamic Screening Framework

Frenchel-Rockafellar Duality [Yamada and Yamada, 2021]

$$P(t) = \min_t d(\mathbf{X}\mathbf{t}) + g(\mathbf{t})$$
$$D(\theta) = \max_\theta -d^*(-\theta) + g^*(\mathbf{X}^\mathsf{T}\theta) \tag{6}$$

- $d(\mathbf{X}\mathbf{t})$ is the distance measure like L_2 function and KL divergence.
- $g(\mathbf{t})$ 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:

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

Screening

- Relying on the KKT condition, we know that, for the optimum $\hat{\theta}$, we must have $\hat{\mathbf{t}}_i h_i(\hat{\theta}) = 0$. which indicates if $h_i(\hat{\theta}) < 0$, then $\hat{\mathbf{t}}_i = 0$.
- For Lasso, the KKT condition is:

$$h_i(\hat{\theta}) = \|\mathbf{x}_i^T \hat{\theta}\| - 1 < 0 \Rightarrow \hat{\mathbf{t}}_i = 0$$

- For UOT:

$$h_i(\hat{\theta}) = \mathbf{x}_i^T \hat{\theta} - \lambda \mathbf{c}_i \leq 0 \Rightarrow \hat{\mathbf{t}}_i = 0$$

- However, we don't know the value of the optimum solution $\hat{\theta}$ at first.
- If we could find an area R^{DS} containing the $\hat{\theta}$, we know that:

$$h_i(\hat{\theta}) \leq \max_{\theta \in R^{DS}} h_i(\theta)$$

- And if found that $\max_{\theta \in R^{DS}} h_i(\theta) < 0$, we can freeze the elements $\hat{\mathbf{t}}_i$ before we know the optimum solution.

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:

$$\begin{aligned} &\frac{1}{2}\|\theta - \tilde{\theta}\|_2^2 + D(\tilde{\theta}) \leq D(\theta) \leq -d^*(-\theta) \\ &\text{s.t. } g(\tilde{\mathbf{t}}) - \theta^T \mathbf{X}\tilde{\mathbf{t}} < 0 \end{aligned} \tag{7}$$

- 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 $\hat{\mathbf{t}}$ must be zero.

Screening

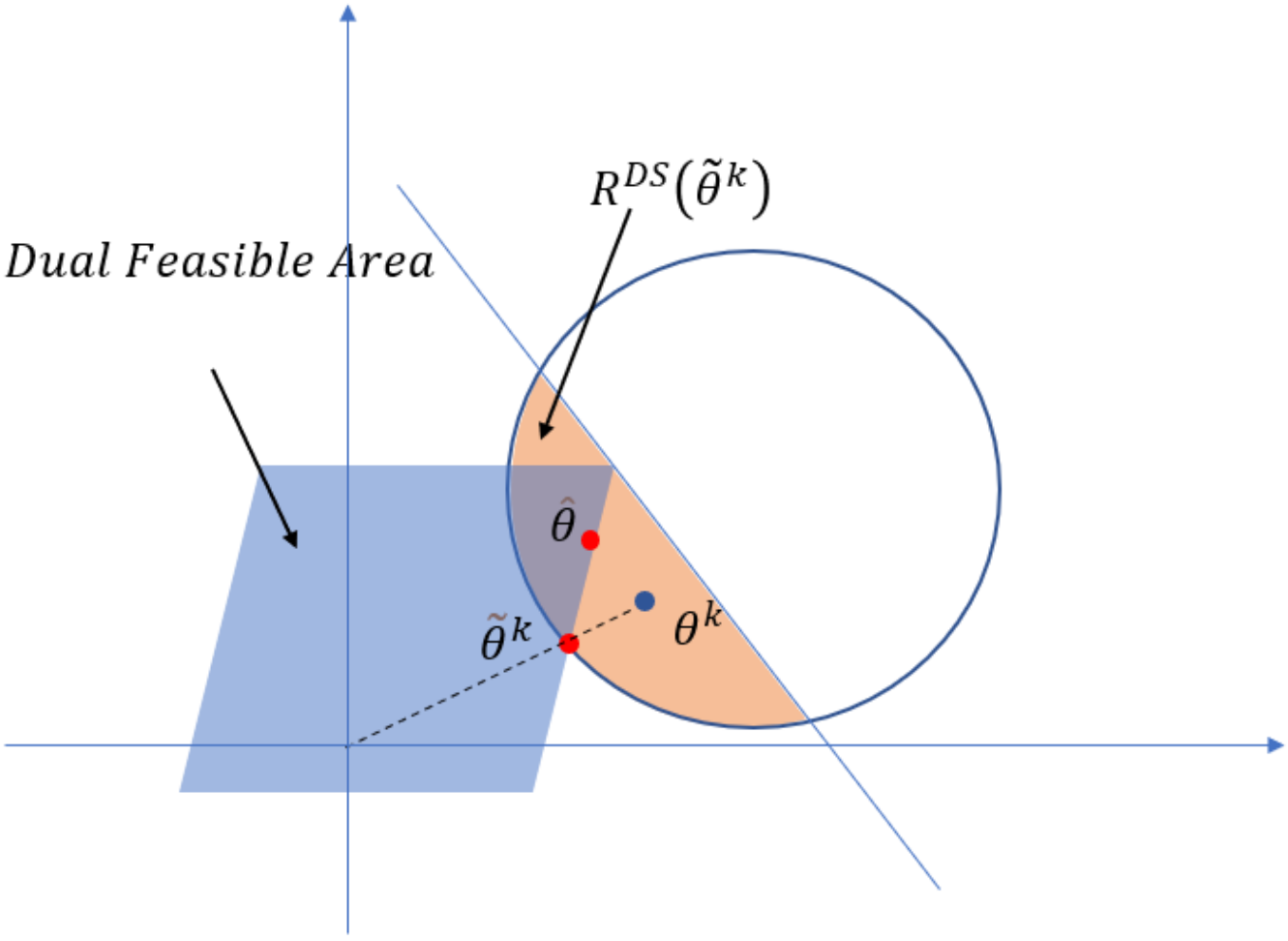

Figure: Projection in Screening

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

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

always holds.

The Projection method

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

$$\tilde{\theta} = \frac{\theta}{\max(1, \| \frac{\mathbf{X}^T \theta}{\mathbf{c}} \|_\infty)}$$

- 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:

$$\theta_{i_1} + \theta_{i_1} < \mathbf{c}_i$$

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

The Projection method

Shifting Screening method:

$$\tilde{\theta}_i = \begin{cases} \theta_i - \max\limits_{j \atop \text{mod } m=i} & (\frac{\theta_{j_1} + \theta_{j_2} - \mathbf{c}_j}{2}) 0 \le i < m \\ \theta_i - \max\limits_{j|m=i} & (\frac{\theta_{j_1} + \theta_{j_2} - \mathbf{c}_j}{2}) m \le i < m+n \end{cases}$$

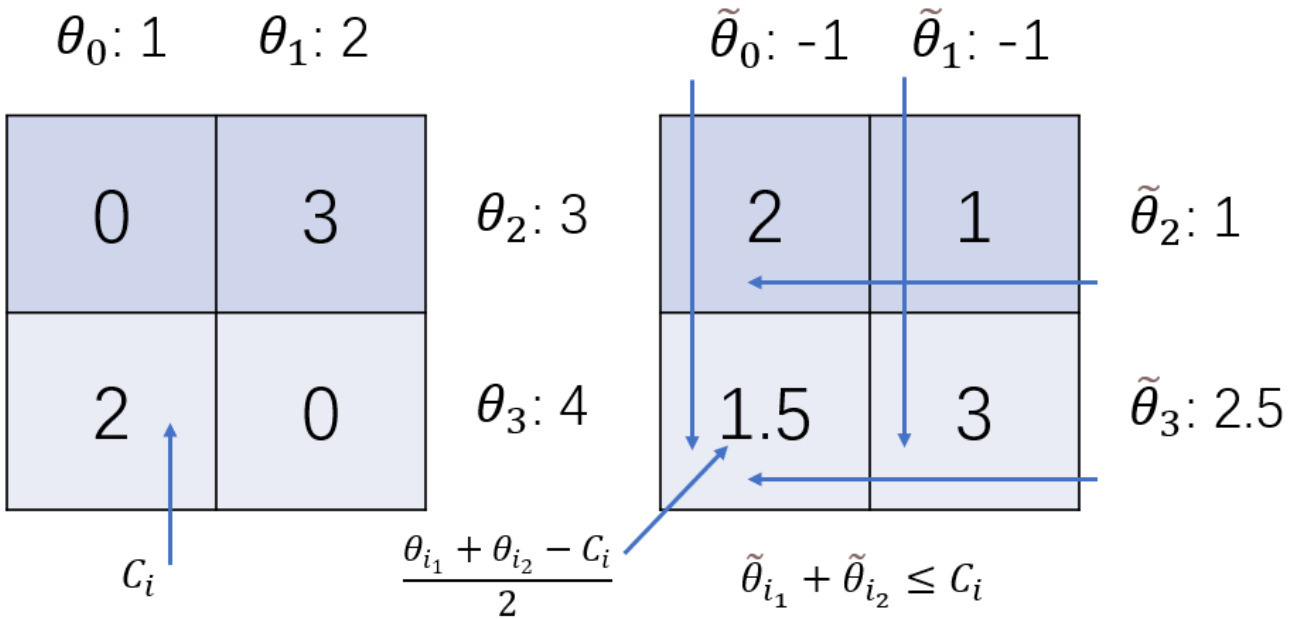

Figure: Shifting on a 2×2 matrix

The Projection Method

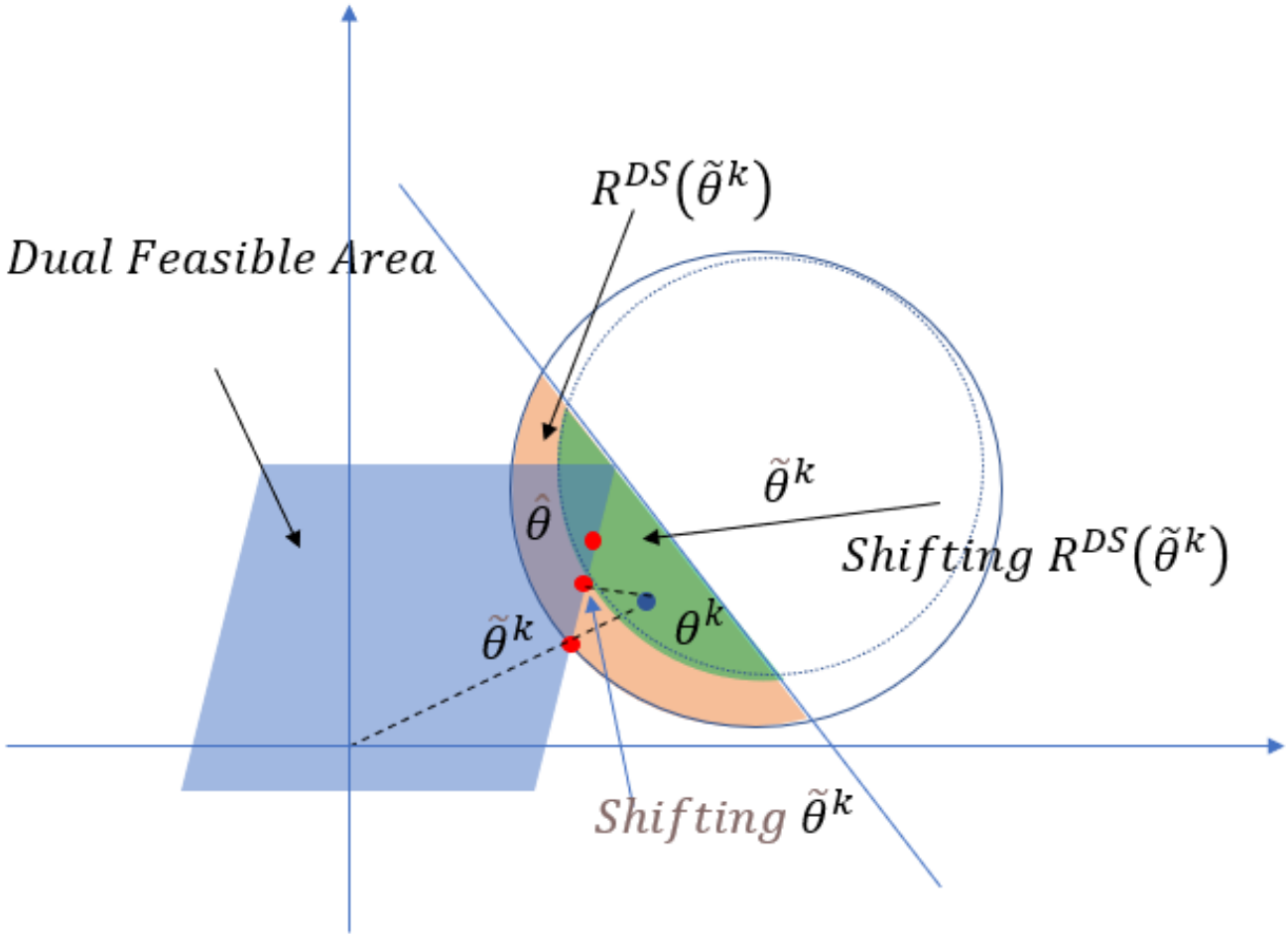

Figure: Difference of the projection method in Screening

Experiments

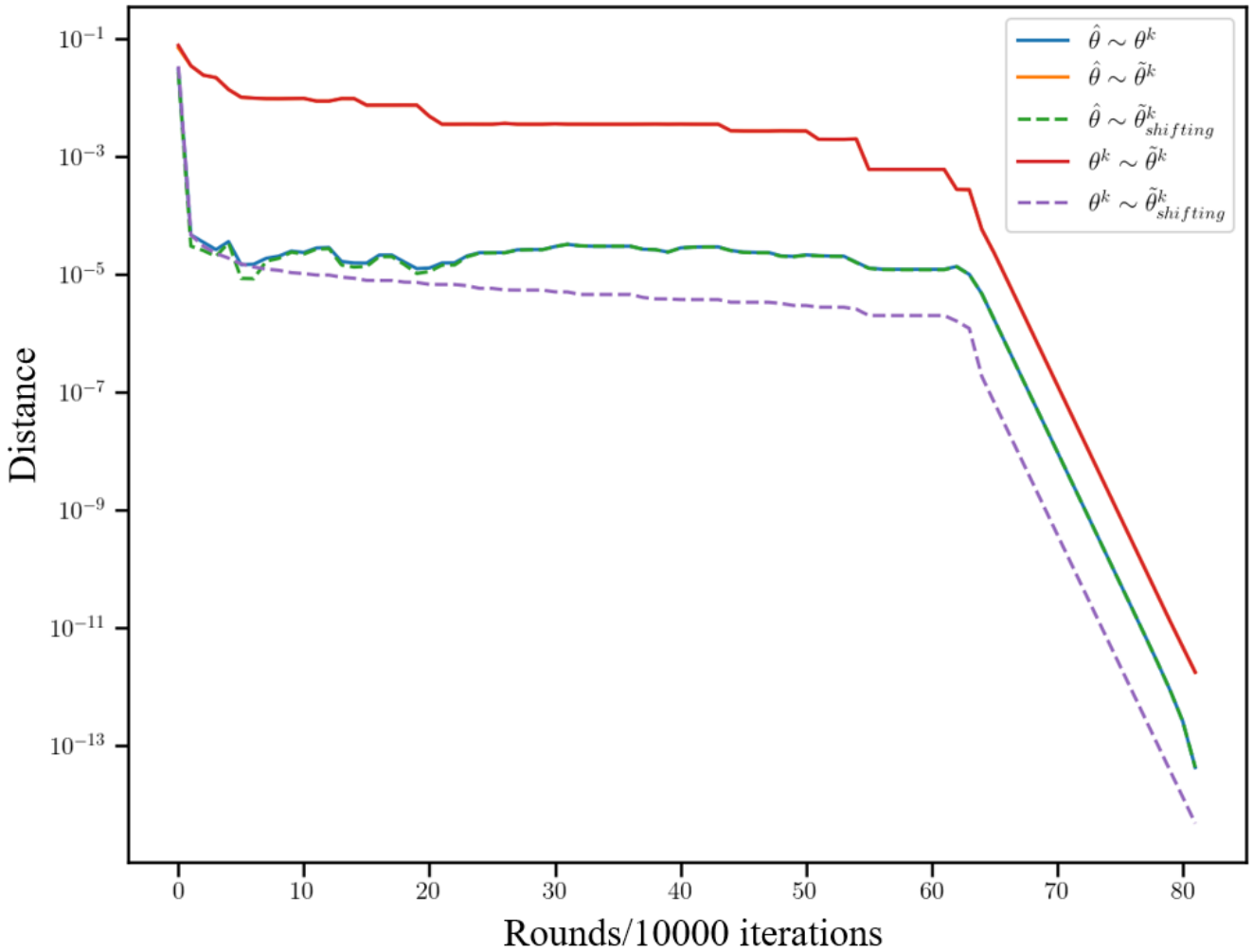

Figure: Distance between the projected point with $\hat{\theta}$ or θ^k

Experiments

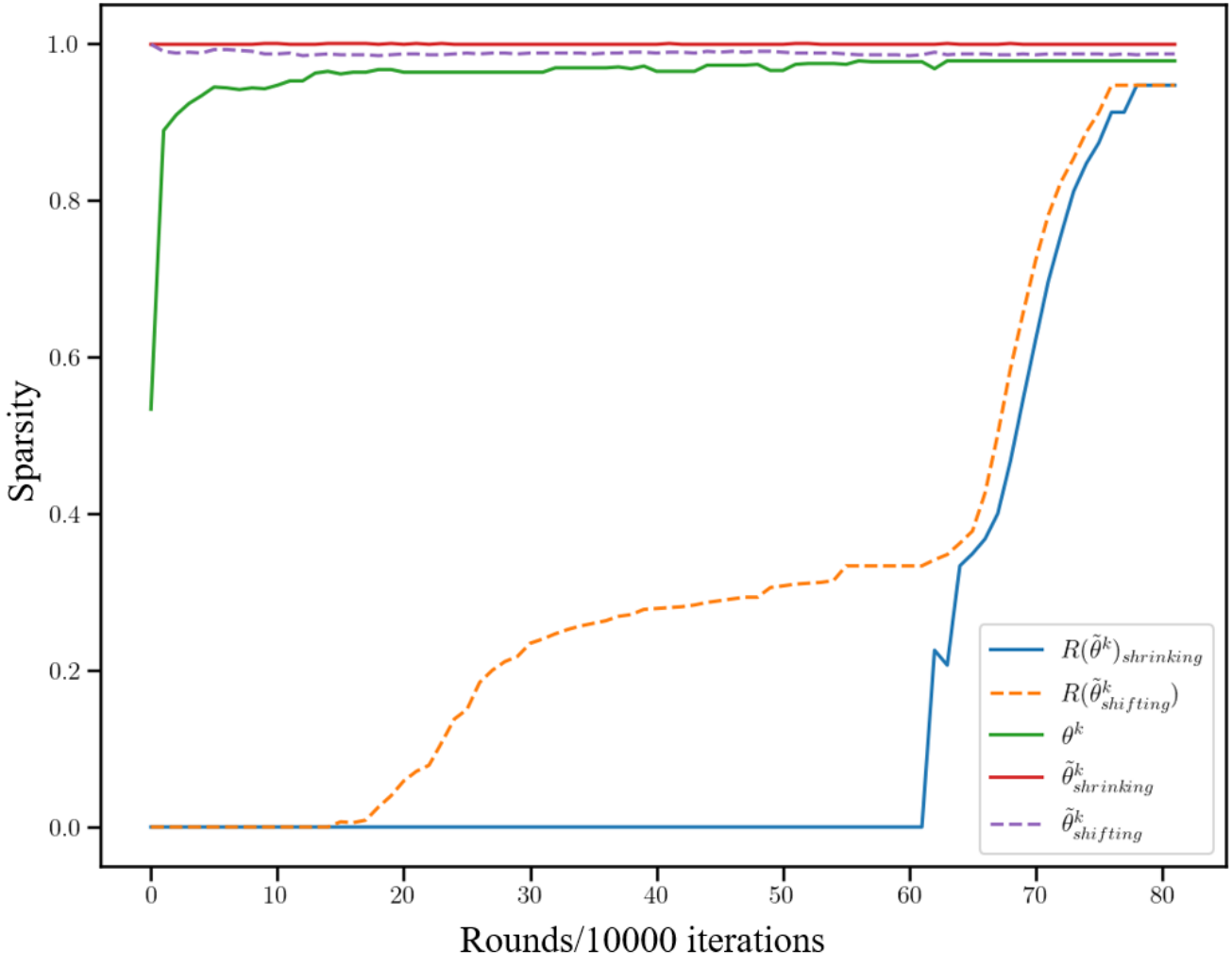

Figure: The Screening Ratio

Potential and defects

- The UOT problem has the potential to screen out better due to its specific sparse structure of matrix $\mathbf{X}$.
- 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.

Future Plan

- We designed a new Two Planes Screening method to construct a new and smaller area $R^{T-DS}(\tilde{\theta})$, which further improve the screening ratio without adding much computational burden. We are writing papers and orgnizing experiments for AISTATS.
- We hope to generalizing the screening method to the KL penalized 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.