The Sparse-Plus-Low-Rank Quasi-Newton Method for Entropic-Regularized Optimal Transport

Chenrui Wang[1] Yixuan Qiu[1]

[1]School of Statistics and Data Science, Shanghai University of Finance and Economics

June 16, 2025

The Primal Problem

Definition 1 (The Entropic-Regularized Optimal Transport)

The entropic-regularized OT problem[a] has the following form:

$$\min_{P \in \Pi(a,b)} \langle P, M \rangle - \eta h(P), \tag{1}$$

- $h(P) = \sum_{i=1}^{n} \sum_{j=1}^{m} P_{ij}(1 - \log P_{ij})$
- $a^T \mathbf{1}_n = b^T \mathbf{1}_m = 1$
- $\Pi(a, b) = \{P \in \mathbb{R}^{n \times m} : P\mathbf{1}_m = a, P^T \mathbf{1}_n = b, P \geq 0\}$

[a] Marco Cuturi. "Sinkhorn distances: Lightspeed computation of optimal transport". In: *Advances in Neural Information Processing Systems*. Vol. 26. 2013.

The Dual Problem

The dual problem of (1):

$$\mathcal{L}(\alpha, \beta) = \alpha^T a + \beta^T b \\ - \eta \sum_{i=1}^{n} \sum_{j=1}^{m} \exp\{\eta^{-1}(\alpha_i + \beta_j - M_{ij})\}. \qquad (2)$$

- $\alpha \in \mathbb{R}^n$, $\beta \in \mathbb{R}^m$ are free variables
- $\mathcal{L}(\alpha, \beta) = \mathcal{L}(\alpha + c\mathbf{1}_n, \beta - c\mathbf{1}_m), \forall c \in \mathbb{R}$, so we remove the redundant degree of freedom by setting $\beta_m = 0$ globally.

The Main Objective

The main objective:

$$\min_{x \in \mathbb{R}^{n+m-1}} f(x) := \min_{\alpha \in \mathbb{R}^n, \beta \in \mathbb{R}^m} -\mathcal{L}(\alpha, \beta). \tag{3}$$

- $f(x)$ is strongly convex
- $\nabla f(x), \nabla^2 f(x)$ both have closed-form expressions:

$$g(x) = \begin{bmatrix} T\mathbf{1}_m - a \\ \tilde{T}^T \mathbf{1}_n - \tilde{b} \end{bmatrix}, \; H(x) = \eta^{-1} \begin{bmatrix} \mathbf{diag}(T\mathbf{1}_m) & \tilde{T} \\ \tilde{T}^T & \mathbf{diag}(\tilde{T}^T \mathbf{1}_n) \end{bmatrix}.$$

- Given an optimal solution (α^*, β^*), the primal optimal solution can be obtained as $T^*_{ij} = \exp\{\eta^{-1}(\alpha^*_i + \beta^*_j - M_{ij})\}$

Overview

We solve (3) by introducing the Sparse-Plus-Low-Rank approach:

1. The algorithm is based on a quasi-Newton framework
2. **Sparse**: we obtain an approximation of $H(x)$ by sparsification
3. **Low-Rank**: we incorporate a low-rank correction term $auu^T + bvv^T$ to enhance the approximation quality
4. The update rule is:

$$x_{k+1} = x_k - \alpha_k B_k^{-1} g_k,$$

where α_k is the step size, g_k is the gradient and B_k is the approximated Hessian matrix

Sparsification Scheme

Definition 2 (Sparsification scheme)

A sparsification scheme is defined by a set of coordinates $\Omega \subseteq \bar{\Omega} = \{(i,j) : i \in [n], j \in [m-1]\}$. In particular, the sparsified matrix $\tilde{T}_\Omega$ has elements

$$(\tilde{T}_\Omega)_{ij} = \begin{cases} \tilde{T}_{ij}, & (i,j) \in \Omega, \\ 0, & (i,j) \notin \Omega, \end{cases}$$

and the sparsified Hessian matrix is given by

$$H_\Omega = H_\Omega(x) = \eta^{-1} \begin{bmatrix} \mathbf{diag}(T\mathbf{1}_m) & \tilde{T}_\Omega \\ \tilde{T}_\Omega^T & \mathbf{diag}(\tilde{T}^T\mathbf{1}_n) \end{bmatrix}.$$

Sparsifying the Hessian Matrix

Sparsification at each iteration:

- $\Omega^* = \{(i,j) : i = 1 \text{ or } j = 1, i \in [n], j \in [m-1]\}$
- $\Omega(\rho)$: coordinates of the largest $100\rho\%$ elements of $\tilde{T}$
- $\Omega = \Omega^* \cup \Omega(\rho)$

The Low-Rank Terms

At the $(k+1)^{th}$ iteration of the Newton-type optimization procedure, the approximated Hessian matrix is:

$$H_{k+1} \approx B_{k+1} := H_{\Omega}^{k+1} + auu^T + bvv^T + \tau_{k+1} I,$$

- τ_{k+1} is a shift parameter for numerical stability
- Motivated by the BFGS algorithm, a, b, u, v are determined by the secant equation:

$$\begin{aligned} u &= y_k, \quad v = (H_{\Omega}^{k+1} + \tau_{k+1} I) s_k, \\ a &= \frac{1}{y_k^T s_k}, \quad b = -\frac{1}{s_k^T (H_{\Omega}^{k+1} + \tau_{k+1} I) s_k}. \end{aligned} \tag{4}$$

where $s_k = x_{k+1} - x_k, y_k = g_{k+1} - g_k$

Eigenvalue Structure

Theorem 3 (Eigenvalue Guarantees)

$\forall \Omega \subseteq \bar{\Omega} : \exists k, s.t. (H_\Omega)^k > 0$, *$H_\Omega$ has the following properties:*

$$\lambda_{\max}(H_\Omega) \leq \lambda_{\max}(H),$$
$$\lambda_{\min}(H_\Omega) \geq \lambda_{\min}(H),$$

where $H = H_{\bar{\Omega}} = H(x)$. The equalities hold if and only if $\Omega = \bar{\Omega}$.

Theorem 3 shows that:

- Positive definiteness is maintained after sparsification
- The sparsified Hessian has a smaller condition number
- Such theorem allows for highly flexible algorithm designs

Convergence Analysis

Theorem 4 (Global Convergence)

Let x_0 be an arbitrary initial value, and $\{x_k\}$ be generated by the SPLR algorithm. Then

$$\lim_{k \to \infty} \|g(x_k)\| = 0.$$

Theorem 5 (Linear Convergence)

Let f^* be the optimal value of $f(x)$. Then for all $k \geq 1$, there is a constant $0 < r < 1$ such that

$$f(x_{k+1}) - f^* \leq r[f(x_k) - f^*].$$

Settings

We compare SPLR with the following algorithms:

1. The Sinkhorn algorithm (equivalent to block coordinate descent, BCD);
2. The adaptive primal-dual accelerated gradient descent (APDAGD[1]);
3. L-BFGS;
4. The Newton method;
5. the SSNS algorithm[2]

[1] Pavel Dvurechensky, Alexander Gasnikov, and Alexey Kroshnin. "Computational Optimal Transport: Complexity by Accelerated Gradient Descent Is Better Than by Sinkhorn's Algorithm". In: *Proceedings of the 35th International Conference on Machine Learning*. 2018, pp. 1367–1376.

[2] Zihao Tang and Yixuan Qiu. "Safe and Sparse Newton Method for Entropic-Regularized Optimal Transport". In: *Advances in Neural Information Processing Systems*. Vol. 38. 2024.

Synthetic I

$M_{ij} \overset{iid}{\sim} \mathrm{Unif}(0,1)$, and $a = n^{-1}\mathbf{1}_n$, $b = m^{-1}\mathbf{1}_m$.

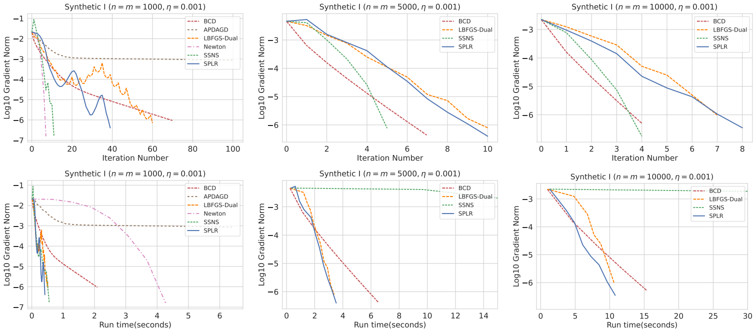

Figure: Top: Gradient norm vs. iteration number. Bottom: Gradient norm vs. run time.

Synthetic II

$$M_{ij} = (x_i - y_j)^2, a \sim \exp(1), b \sim 0.2 \cdot N(1, 0.2) + 0.8 \cdot N(3, 0.5).$$

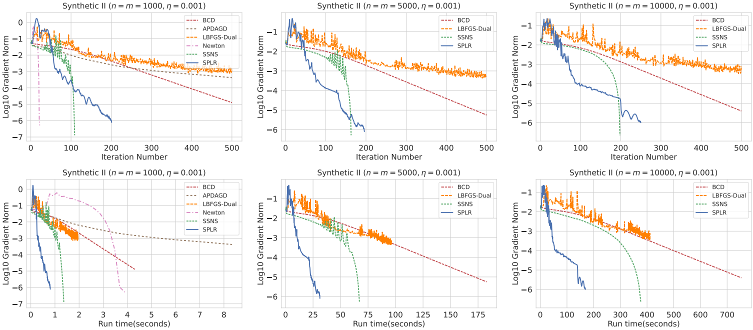

Figure: Top: Gradient norm vs. iteration number. Bottom: Gradient norm vs. run time.

MNIST and Fashion-MNIST

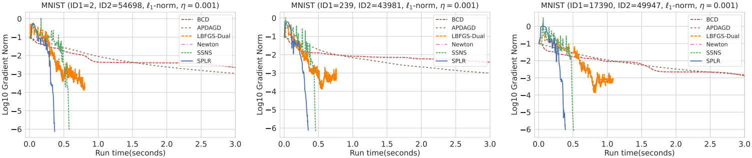

Figure: Performance of different algorithms on the MNIST data.

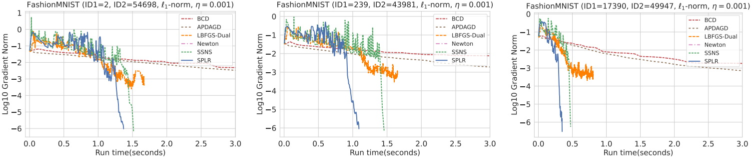

Figure: Performance of different algorithms on the Fashion-MNIST data.

ImageNet

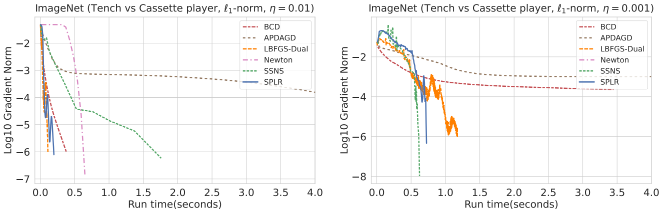

Figure: Performance of different algorithms on the ImageNet data. Left: $\eta = 0.01$. Right: $\eta = 0.001$.

Ablation Study

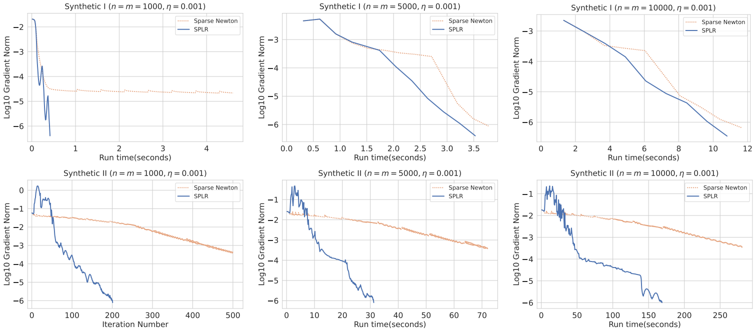

Figure: Top: Synthetic I. Bottom: Synthetic II.

Summary

1. We proposed a new efficient quasi-Newton method for solving entropic-regularized optimal transport problems.
2. We provided theoretical results on how the sparsification process affects the eigenvalues.
3. We proved both the global convergence and the linear convergence rate of the SPLR method.

Thank You!