Safe and Sparse Newton Method for Entropic-Regularized Optimal Transport

Zihao Tang, Yixuan Qiu

Shanghai University of Finance and Economics

NeurIPS 2024 @ Vancouver, December 11

Outline

Motivation and Problem Setting

Optimal Transport in Machine Learning

- Optimal Transport (OT) provides a mathematical framework for measuring and minimizing the difference between two probability distributions.

Optimal Transport in Machine Learning

- **Domain Adaptation:** Matching distributions from source and target domains.
- **Generative Model:** Modeling data distributions, especially in generative adversarial networks (GANs).
- **Metric:** Used to define Wasserstein distances in deep learning and other fields.
- ...

Importance of Entropic-Regularization

- Entropic Regularization introduces an entropy term to the standard OT problem, turning the original problem into a smooth approximation.

Importance of Entropic-Regularization

- The regularization ensures that the optimal transport plan is computationally feasible for large-scale problems, at the cost of some accuracy and optimality.

Formulation of Entropic-Regularized OT

- Objective function

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

  - M is the cost matrix.
  - $\Pi(a,b) = \{T \in \mathbb{R}^{n \times m} : T1_m = a, T^T 1_n = b, T \geq 0\}$.
  - $h(T) = \sum_i \sum_j T_{ij}(1 - \log T_{ij})$.
  - η controls the level of regularization (smoothness).

Safe and Sparse Newton Method

Newton’s Method for Entropic-Regularized OT

- Advantages
  - **Quadratic Convergence:** Newton’s method converges quickly for smooth problems, if the initial value is sufficiently close to the optimum.
- Limitations
  - **Sensitivity to Initial Conditions:** The algorithm can struggle with ill conditioned problems and poor initial guesses.
  - **Computationally Expensive:** Calculating Hessians and solving large linear systems may become prohibitively expensive for very high-dimensional problems.

Algorithm 1: Sparsifying the Hessian Matrix

- **Fundamental Reason For Sparsification:** Sparse linear systems solve Newton directions faster.
- **Good Approximation:** The density of the Hessian matrix H originates from the approximately sparse entropic-regularized optimal transport plan T. We sparsify it using Algorithm 1, obtaining the sparse Hessian matrix H_δ, and theoretically prove that it provides a good approximation.

Algorithm 1 Sparsifying the Hessian matrix.

Input: Dual variable vector $x = (\alpha^T, \tilde{\beta}^T)^T$, threshold parameter $\delta \geq 0$
Output: Sparsified Hessian matrix H_δ

Initialize a zero matrix $\Delta \in \mathbb{R}^{n \times m}$ and compute $T = \tau(\alpha, \beta)$
for $j = 1, 2, \dots, m-1$ **do**
  $\phi \leftarrow \texttt{select\_small}(T_{\cdot j}, \delta)$, $\quad \Delta_{\cdot j} \leftarrow \texttt{apply\_mask}(T_{\cdot j}, \phi)$
for $i = 1, 2, \dots, n$ **do**
  $\phi \leftarrow \texttt{select\_small}(\Delta_{i \cdot}, \delta)$, $\quad \Delta_{i \cdot} \leftarrow \texttt{apply\_mask}(\Delta_{i \cdot}, \phi)$
$T_\delta \leftarrow T - \Delta$
$H_\delta \leftarrow \eta^{-1} \begin{bmatrix} \mathbf{diag}(T\mathbf{1}_m) & \tilde{T}_\delta \\ \tilde{T}_\delta^T & \mathbf{diag}(\tilde{T}^T\mathbf{1}_n) \end{bmatrix}$

Density stems from the optimal transport plan

Algorithm 2: SSNS

- **Positive Definite:** Ensuring the sparsified approximate Hessian matrix H_δ remains positive definite, thus safe to compute p_k

---

Algorithm 2 Safe and sparse Newton method for Sinkhorn-type optimal transport.

Input: Initial point x_0, parameters $\{\mu_0, \nu_0, c_l, c_u, \kappa\} > 0$, $\gamma \geq 1$, $\rho_0 \in (0, \frac{1}{2})$, $\varepsilon_{tol} > 0$
Default values: $\mu_0 = 1$, $\nu_0 = 0.01$, $c_l = 0.1$, $c_u = 1$, $\kappa = 0.001$, $\gamma = 1$, $\rho_0 = \frac{1}{4}$
Output: x_k

1: **for** $k = 0, 1, 2, \dots$ **do**
2:     Compute $g_k = g(x_k)$, $\delta_k = \nu_0 \|g_k\|^\gamma$
3:     **if** $\|g_k\| < \varepsilon_{tol}$ **then**
4:         **return** x_k
5:     Compute H_{δ_k} according to Algorithm 1 with $x \leftarrow x_k$
6:     Compute $p_k = -(H_{\delta_k} + \mu_k \|g_k\| I)^{-1} g_k$
7:     Select any $\xi_k \in [c_l, c_u]$
8:     Compute $\rho_k = \dfrac{f(x_k) - f(x_k + \xi_k p_k)}{m_k(0) - m_k(\xi_k p_k)}$, $m_k(\cdot)$ is defined in (6)
9:     Update $\mu_{k+1} = \begin{cases} 4\mu_k, & \text{if } \rho_k < \rho_0 \\ \max\{\mu_k/2, \kappa\}, & \text{if } \rho_k \geq 1 - \rho_0 \\ \mu_k, & \text{otherwise} \end{cases}$
10:     **if** $\rho_k > 0$ **then**
11:         $x_{k+1} = x_k + \xi_k p_k$
12:     **else**
13:         $x_{k+1} = x_k$

5: For sparse

6: For safe

Global convergence

Theorem (Global convergence guarantee)
Let $\{x_k\}$ be generated by Algorithm 2, and x^* is an optimal point. Then either Algorithm 2 terminates in finite iterations, or x_k satisfies $\lim_{k\to\infty}\|g(x_k)\| = 0$, $\lim_{k\to\infty}\|x_k - x^*\| = 0$.

- **Convergence from Any Initial Point:** Starting from any arbitrary initial point x_0 the iterates generated by the algorithm converge to the unique global optimum x^*.
- **End-to-End Efficiency:** The method eliminates the need for warm initialization with the Sinkhorn algorithm, enabling a more streamlined, end-to-end process.

Quadratic local convergence rate

Theorem (Quadratic local convergence rate)
Fix $\xi_k \equiv 1$. Then there exists an integer $K' > 0$ and a constant $L > 0$ such that for all $k \geq K'$,

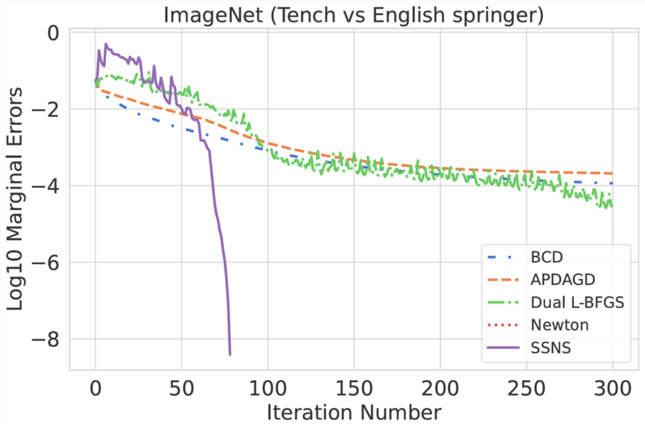
$$\|x_{k+1} - x^*\| \leq L\|x_k - x^*\|^2.$$

- **Convergence Rate Comparable to Newton:** SSNS achieves a quadratic local convergence rate that aligns with the Newton method using a genuine and dense Hessian matrix.

Numerical Experiments

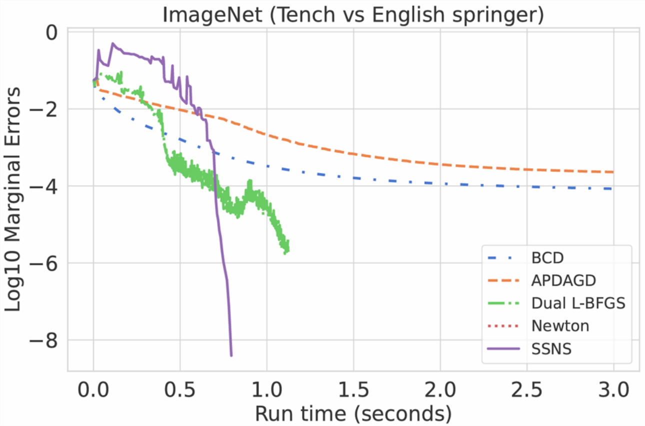
Numerical Experiments

- Iteration v.s. Log10 marginal errors



Numerical Experiments

- Runtime v.s. Log10 marginal errors

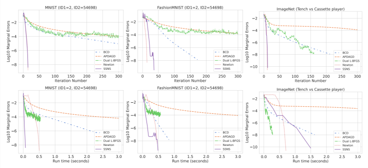


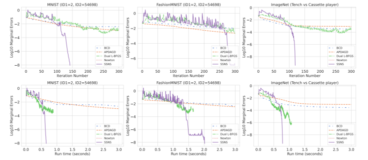
Numerical Experiments

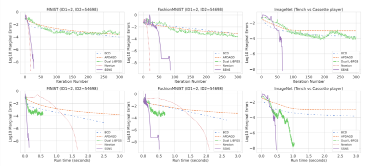
- The experimental results show that SSNS has advantages in both the number of iterations and runtime in most scenarios.

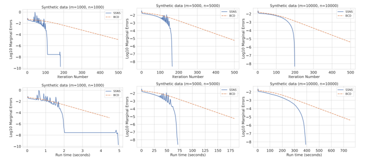
Numerical Experiments

- More experiments are in the article.









Summary

- We propose a Hessian sparsification scheme with strict control over approximation error.
- Based on this scheme, we prove that the sparsified Hessian matrix is always positive definite, enabling a safe Newton-type method that avoids singularities.
- The algorithm is easy to implement, avoids most hyperparameter tuning, and is included in the **RegOT** Python package.
- We provide rigorous global and local convergence analysis for the algorithm, which is lacking in current literature.

THANK YOU!