A Truncated Newton Method for Optimal Transport

Mete Kemertas, Amir-massoud Farahmand, Allan D. Jepson

Paper

github.com/metekemertas/mdot_tnt

Computer Science
UNIVERSITY OF TORONTO

POLYTECHNIQUE
MONTRÉAL

VECTOR
INSTITUTE

Mila

Problem & Motivation

- Existing practical solvers for the discrete OT problem either:
  a. don't scale well with problem size n,
  b. don't leverage GPU parallelization,
  c. sacrifice accuracy for scalability (e.g., entropic regularization methods don't converge quickly enough in the weak regularization regime to be practical).
- We need better solvers, not workarounds.

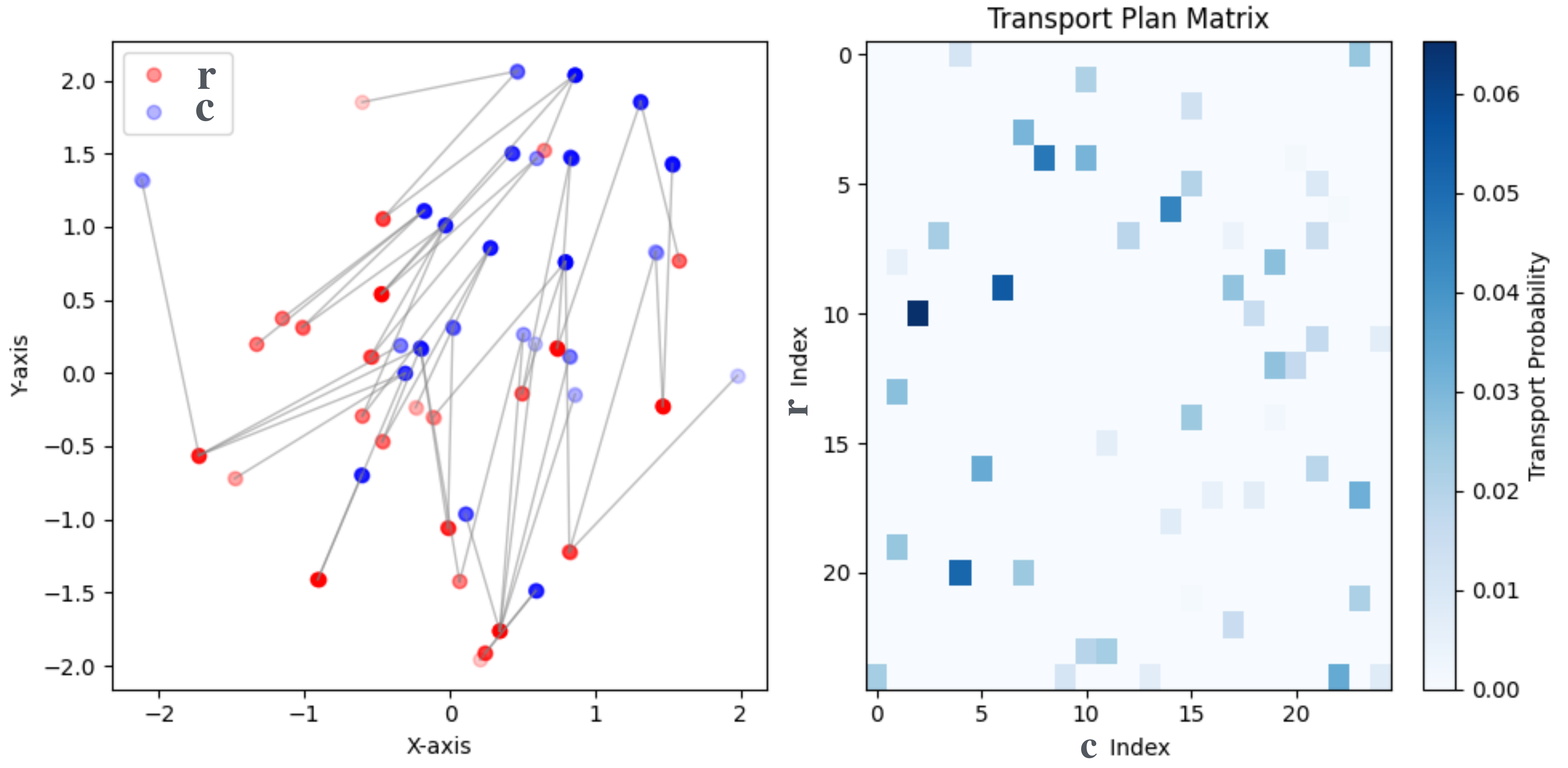
The Discrete Optimal Transport Problem

Feasible set:

$$U(\mathbf{r},\mathbf{c}) = \{P \in \mathbb{R}^{n\times n}_{\geq 0} \mid P\mathbf{1} = \mathbf{r}, P^\top\mathbf{1} = \mathbf{c}\}$$

Optimization problem given cost matrix C:

$$\text{minimize}_{P\in U(\mathbf{r},\mathbf{c})}\langle P, C\rangle$$



- This is a well-studied linear program, with many existing specialized solvers.
- Best practical <u>exact</u> solvers have $O(n^3)$ complexity [1], and theoretical solvers $O(n^{2.5})$ [2].

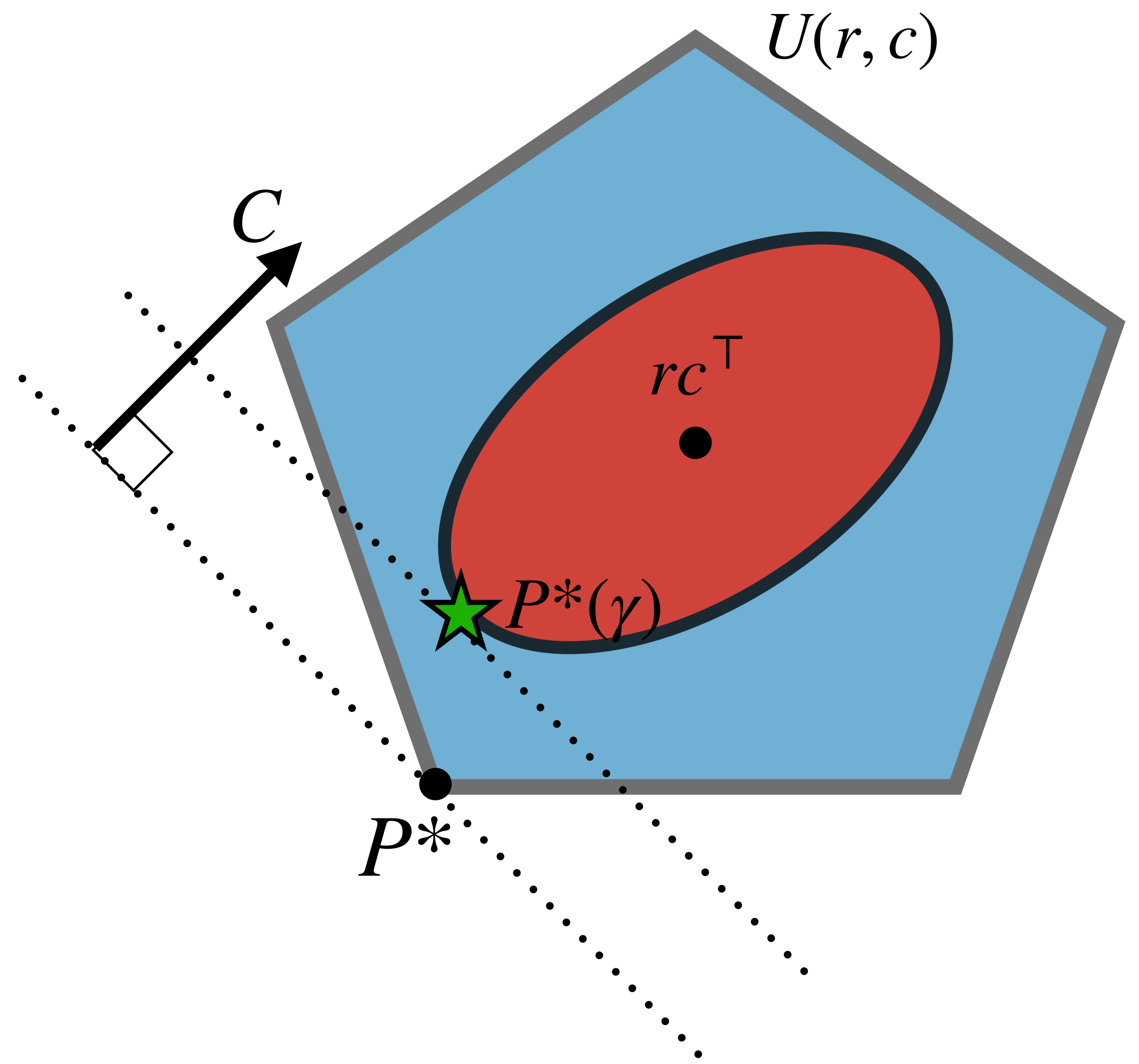
[1] Pele, O., & Werman, M. *Fast and robust earth mover's distances.* ICCV, 2009.

[2] Lee, Y. T., & Sidford, A. *Path finding methods for linear programming: Solving linear programs in o (vrank) iterations and faster algorithms for maximum flow.* FOCS, 2014.

Entropic-regularized Optimal Transport

Favour high-entropy solutions:

$$\text{minimize}_{P \in U(\mathbf{r},\mathbf{c})} \langle P, C \rangle - \frac{1}{\gamma} H(P)$$



GPU-parallel Sinkhorn's algorithm has $\tilde{O}(n^2)$ dependence on problem size [3].

Very slow when γ is large; guarantees $\langle P - P^*, C \rangle \leq \varepsilon$ in $\tilde{O}(n^2/\varepsilon^2)$ time [4]. Best alternative theoretically $\tilde{O}(n^2/\varepsilon)$, but Sinkhorn still outperforms many existing alternatives (at worst with some tuning) [5].

[3] Cuturi, M. Sinkhorn distances: *Lightspeed computation of optimal transport*. NeurIPS, 2013.
[4] Dvurechensky, P., Gasnikov, A., and Kroshnin, A. *Computational optimal transport: Complexity by accelerated gradient descent is better than by Sinkhorn's algorithm*. ICML, 2018.
[5] Jambulapati, A., Sidford, A., and Tian, K. *A direct $\tilde{O}(1/\varepsilon)$ iteration parallel algorithm for optimal transport*. NeurIPS, 2019.

Prior Work: Temperature Annealing

<u>Recall</u> the EOT dual objective (convex):

$$g(\mathbf{u}, \mathbf{v}; \gamma) = \sum_{ij} P(\mathbf{u}, \mathbf{v}; \gamma)_{ij} - \langle \mathbf{u}, \mathbf{r} \rangle - \langle \mathbf{v}, \mathbf{c} \rangle$$

where $P(\mathbf{u}, \mathbf{v}; \gamma)_{ij} = \exp(u_i + v_j - \gamma C_{ij})$. Then, for optimal $\mathbf{u}^*, \mathbf{v}^* \in \mathbb{R}^n$ we have:

$$P(\mathbf{u}^*, \mathbf{v}^*; \gamma) = \underset{P \in U(\mathbf{r},\mathbf{c})}{\operatorname{argmin}} \ \langle P, C \rangle - \frac{1}{\gamma} H(P)$$

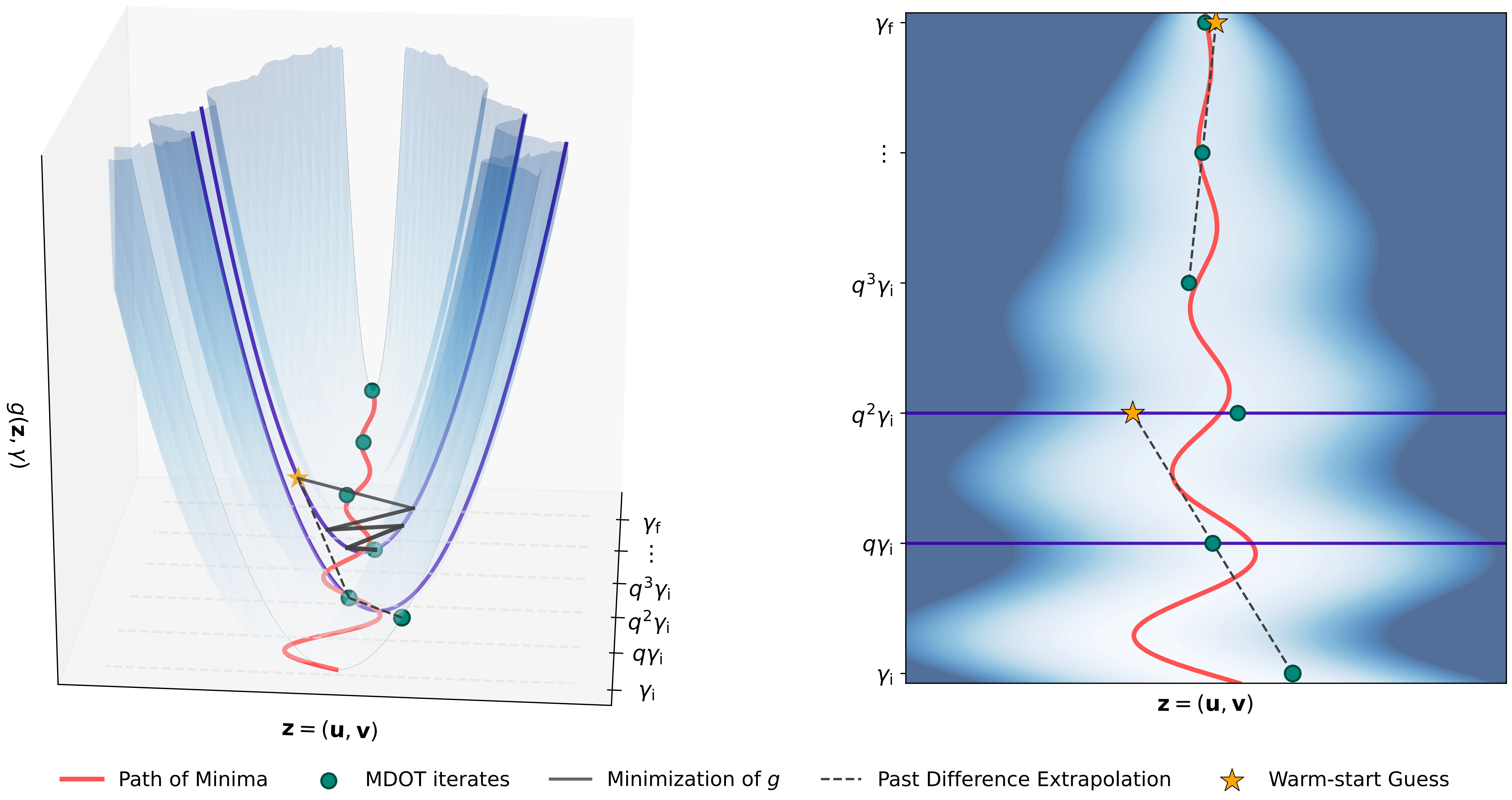
Several prior works minimize the dual while progressively increasing γ.

This is known as the temperature annealing or ε-scaling *heuristic*.

Prior Work: Mirror Descent Optimal Transport (MDOT)

In our recent work, we introduced the MDOT framework [6];
temperature annealing is a certain kind of inexact mirror descent on the OT problem.

Visualization of MDOT in Dual Space



[6] Kemertas, M., Jepson, A. D., & Farahmand, A. M. (2025). *Efficient and accurate optimal transport with mirror descent and conjugate gradients.*

Accelerating MDOT

Goal: 🎯

To overcome the ill-conditioning for large γ, develop a second-order minimizer for the convex dual objective that is

a) GPU parallelizable,

b) Numerically stable (e.g., in the weak regularization regime),

c) Scales to high dimensions.

Designing a New OT Solver with MDOT

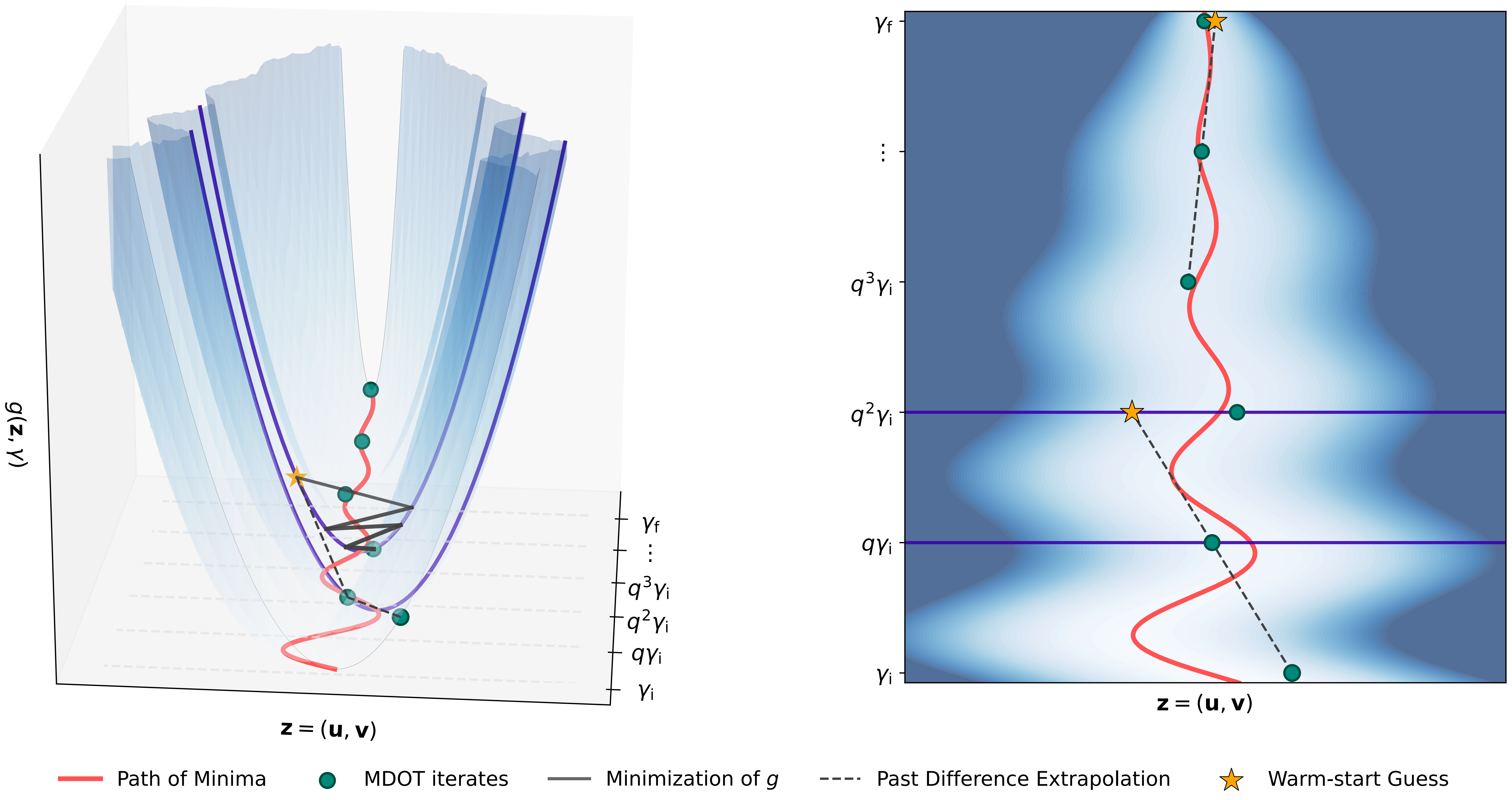
Newton's method converges quadratically near the solution.

<u>Idea</u>:

1. Adapt q to initialize each problem near the quadratic convergence zone.

2. Solve the Newton system approximately (*truncated* Newton) using GPU-parallel conjugate gradients (1st-order optimal).

Visualization of MDOT in Dual Space



Transforming the Newton System

The Hessian of the dual has a zero eigenvalue and can be ill-conditioned:

$$\nabla^2 g = \begin{pmatrix} \mathbf{D}(\boldsymbol{r}(P)) & P \\ P^\top & \mathbf{D}(\boldsymbol{c}(P)) \end{pmatrix}$$

"Discount" the Hessian to make it positive-definite, given some $\rho \in [0,1)$:

$$\nabla^2 g(\rho) := \begin{pmatrix} \mathbf{D}(\boldsymbol{r}(P)) & \sqrt{\rho}P \\ \sqrt{\rho}P^\top & \mathbf{D}(\boldsymbol{c}(P)) \end{pmatrix}$$

Start with $\rho = 0$ and anneal $1 - \rho$ until approximate solution of the "discounted Newton system" satisfies conditions for quadratic convergence.

$$\nabla^2 g \, \boldsymbol{d} = -\nabla g$$

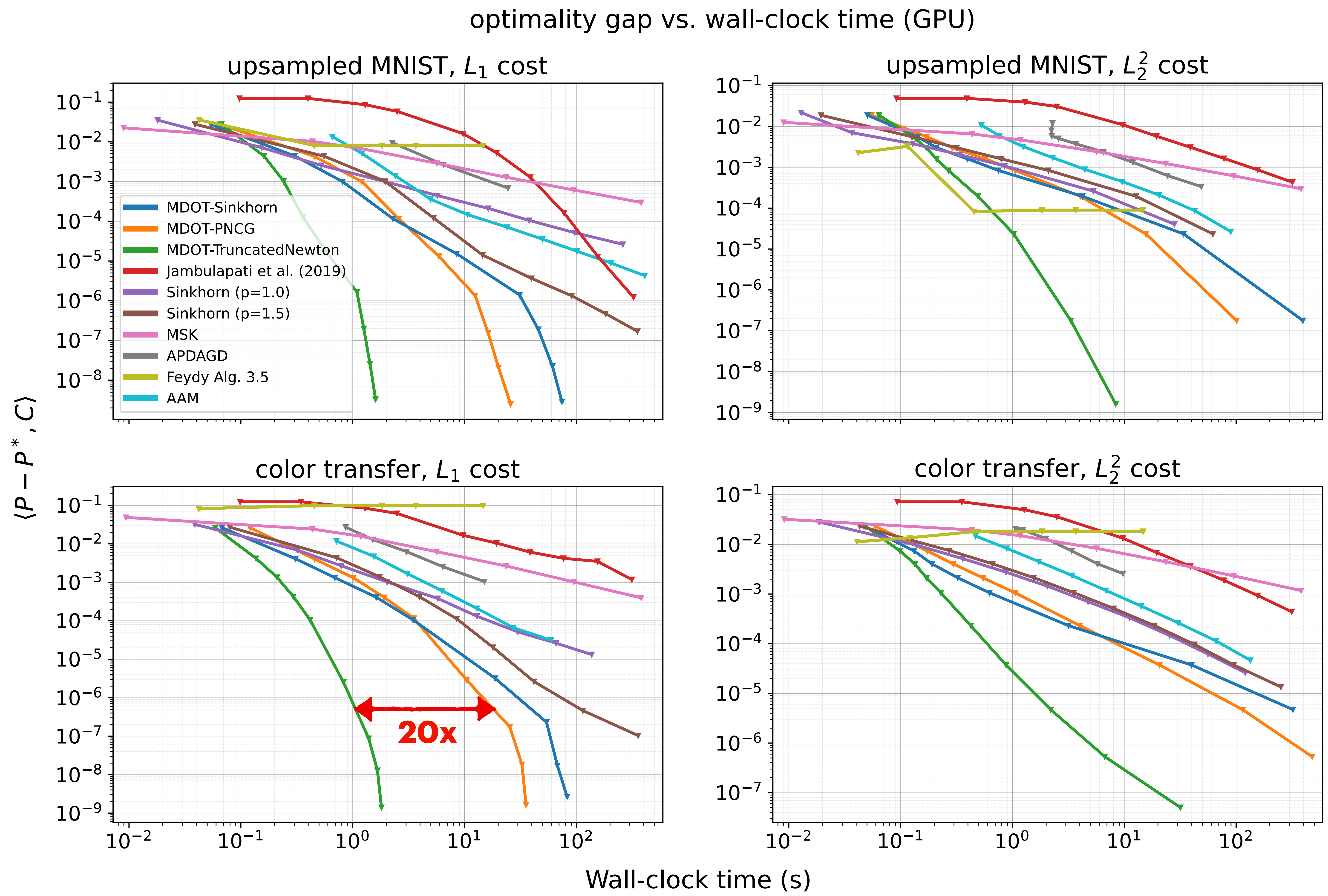
$$\Downarrow$$

$$\nabla^2 g(\rho) \, \boldsymbol{d} = -\nabla g$$

Benchmarking: A Better Solver

Quadratic local convergence guarantee, combined with linear algebra tricks for efficiency and numerical stability yields **orders of magnitude speedup on 12 datasets x 2 cost functions.**

4-6 decimal accuracy, returning a strictly feasible plan in $U(\mathbf{r}, \mathbf{c})$ in less than a second on a 2018-era GPU (n=4096).



Scalability

In principle, can solve **n = 1 million** dimensional OT problem (color transfer) to high precision.

Supports $O(n)$ memory footprint implementation.

Image A

Image B

A → B

B → A