A Sinkhorn-Newton method for entropic optimal transport

Christoph Brauer* Christian Clason† Dirk Lorenz* Benedikt Wirth*

* TU Braunschweig, Germany † University Duisburg-Essen, Germany * University of Münster, Germany

Entropically regularized optimal transport

- We consider the entropically regularized Kantorovich problem of optimal mass transport between two given probability measures $\mu \in \mathbb{R}^M$ and $\nu \in \mathbb{R}^N$, with non-negative cost $c \in \mathbb{R}^{M\times N}$ and regularization strength $\gamma > 0$,

$$\begin{aligned} \inf_{\pi\in\mathbb{R}^{M\times N}} \quad & \langle c,\pi\rangle + \gamma\langle \pi, \log\pi - \mathbb{1}\rangle \\ \text{s.t.} \quad & \pi\mathbf{1} = \mu \\ & \pi^\top\mathbf{1} = \nu, \end{aligned} \tag{P}$$

where $\mathbf{1}$ and $\mathbb{1}$ denote vectors and matrices of all ones.

Dual problem

- By Fenchel-Rockafellar duality, the primal problem (P) is associated with the dual problem

$$\sup_{\alpha\in\mathbb{R}^M,\ \beta\in\mathbb{R}^N} \quad -\langle\mu,\alpha\rangle - \langle\nu,\beta\rangle - \gamma\langle e^{-\frac{c}{\gamma}}, e^{-\frac{\alpha\mathbf{1}^\top + \mathbf{1}\beta^\top}{\gamma}}\rangle, \tag{D}$$

where the exponential function is applied componentwise to the respective matrices.

- In the following, we abbreviate $K := e^{-\frac{c}{\gamma}} \in \mathbb{R}^{M\times N}$ and use the symbol $\odot$ to denote the Hadamard product.

Optimality conditions

- The primal and the dual problems are connected via the optimality conditions

$$\pi = K \odot e^{-\frac{\alpha\mathbf{1}^\top + \mathbf{1}\beta^\top}{\gamma}} \tag{1}$$

$$\mu = e^{-\frac{\alpha}{\gamma}} \odot K e^{-\frac{\beta}{\gamma}} \tag{2}$$

$$\nu = e^{-\frac{\beta}{\gamma}} \odot K^\top e^{-\frac{\alpha}{\gamma}}. \tag{3}$$

- The first condition implies that the optimal transport plan π is the componentwise product of K and a low-rank matrix induced by the dual variables α and β.
- The second and third conditions are the constraints of (P) with π replaced by the right-hand side of (1).

Newton step

- Our approach is to solve (2) and (3) simultaneously by applying Newton's method to the function

$$G(\alpha,\beta) := \begin{pmatrix} \mu - e^{-\frac{\alpha}{\gamma}} \odot K e^{-\frac{\beta}{\gamma}} \\ \nu - e^{-\frac{\beta}{\gamma}} \odot K^\top e^{-\frac{\alpha}{\gamma}} \end{pmatrix}. \tag{4}$$

- The associated Newton step can be written in the form of

$$\underbrace{\frac{1}{\gamma}\begin{bmatrix} \mathrm{Diag}(\pi\mathbf{1}) & \pi \\ \pi^\top & \mathrm{Diag}(\pi^\top\mathbf{1}) \end{bmatrix}}_{=DG(\alpha,\beta)} \begin{pmatrix} \delta\alpha \\ \delta\beta \end{pmatrix} = -\underbrace{\begin{pmatrix} \mu - \pi\mathbf{1} \\ \nu - \pi^\top\mathbf{1} \end{pmatrix}}_{=G(\alpha,\beta)}, \tag{5}$$

where (1) is used to simplify both $G(\alpha,\beta)$ and $DG(\alpha,\beta)$.

Properties

- For $\alpha, \beta > -\infty$, $DG(\alpha,\beta)$ is symmetric positive semi-definite, and its kernel is $\ker(DG(\alpha,\beta)) = \operatorname{span}\{\binom{\mathbf{1}}{-\mathbf{1}}\}$. Hence, we can use a (preconditioned) conjugate gradient method to solve (5), which operates on the orthogonal complement of the kernel as long as the initial point satisfies $\sum_i \alpha_i^0 = \sum_j \beta_j^0$.
- A cheap diagonal preconditioner is provided by $DG(\alpha,\beta)$ without the off-diagonal blocks.
- If the initial point (α^0,β^0) is chosen sufficiently close to a solution of (2)–(3) and if the optimal transport plan satisfies $\pi \geq \varepsilon\mathbb{1}$ for some $\varepsilon > 0$, then the Newton iteration converges quadratically.
- After a substitution, a Sinkhorn-Knopp step approximates (5) by neglecting the off-diagonal blocks of $DG(\alpha,\beta)$ and solving separately for both variables.

Numerical Experiments

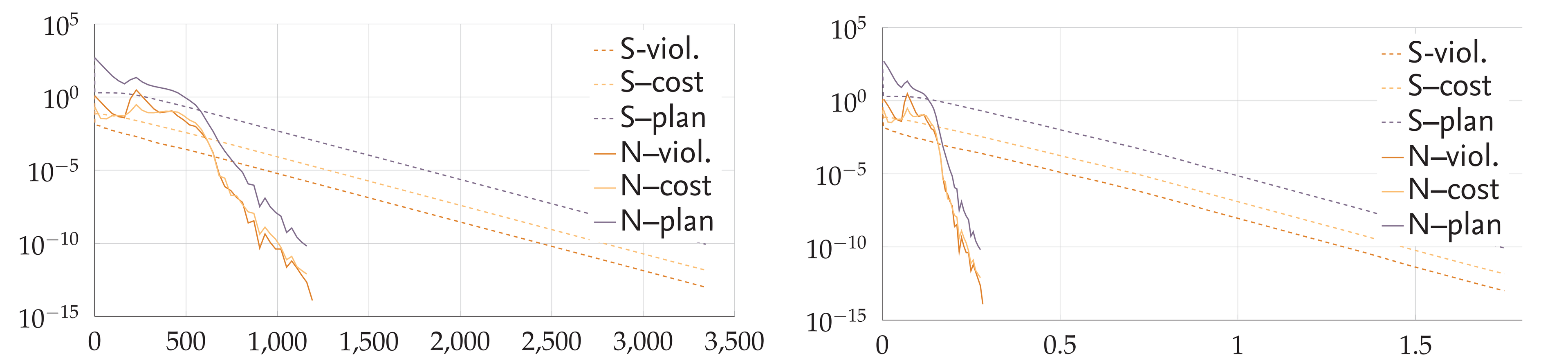

Figure 1: Exemplary performance of Sinkhorn (S) and Newton (N) iterations measured by constraint violation (viol.), distance to optimal transport costs (cost) and distance to optimal transport plan (plan). Left: Erros over CG iterations. Right: Errors over run time in seconds.

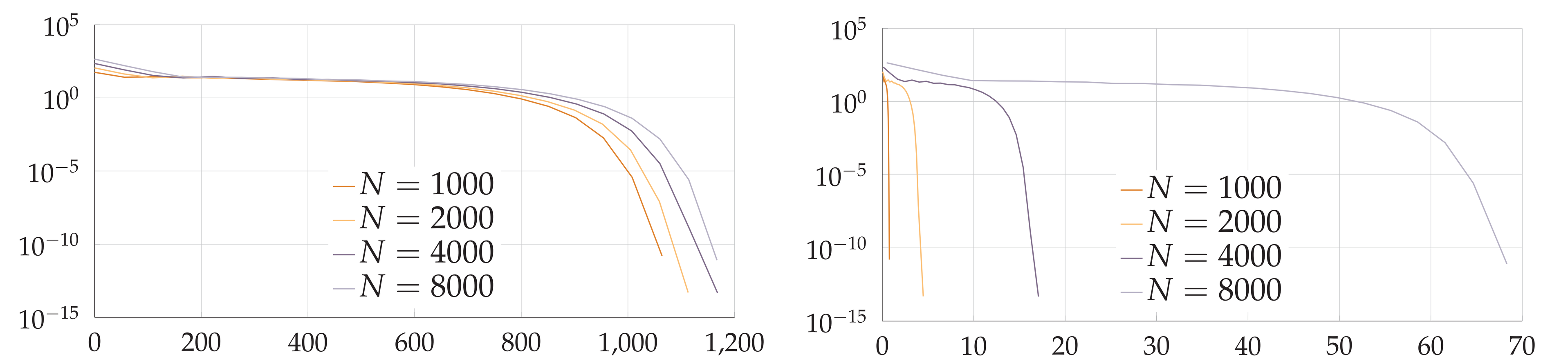

Figure 2: Convergence behavior of Newton for different mesh sizes N (with $M = N$), measured by constraint violation. Left: Errors over CG iterations. Right: Errors over run time in seconds.

UNIVERSITÄT DUISBURG ESSEN

WWU MÜNSTER

Technische Universität Braunschweig