Sharper Exponential Convergence Rates for Sinkhorn's Algorithm in Continuous Settings

Alex Delalande

Joint work with Lénaïc Chizat and Tomas Vaškevičius

EPFL

November 2024

Introduction

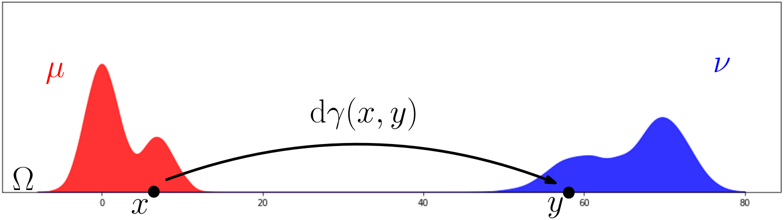
Optimal Transport problem

Optimal Transport problem (Monge, 1781; Kantorovich, 1942):

- Given $\mu, \nu \in \mathcal{P}(\Omega)$ and $c : \Omega \times \Omega \to \mathbb{R}$, solve

$$\boxed{\inf_{\gamma \in \Gamma(\mu,\nu)} \int_{\Omega\times\Omega} c(x,y) \mathrm{d}\gamma(x,y),}$$

where $\Gamma(\mu,\nu) = \{\gamma \in \mathcal{P}(\Omega \times \Omega) \mid \forall A \subset \Omega, \gamma(A \times \Omega) = \mu(A), \gamma(\Omega \times A) = \nu(A)\}$.



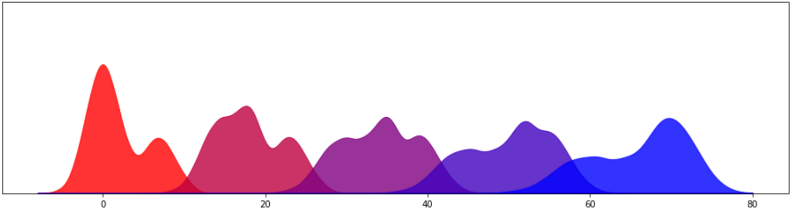
Introduction

Optimal Transport problem

- p-***Wasserstein distance*** between μ and ν when $\Omega \subset \mathbb{R}^d$:

$$\boxed{\mathrm{W}_p(\mu, \nu) := \left(\inf_{\gamma \in \Gamma(\mu,\nu)} \int_{\Omega \times \Omega} \|x - y\|^p \, \mathrm{d}\gamma(x, y) \right)^{1/p}.}$$

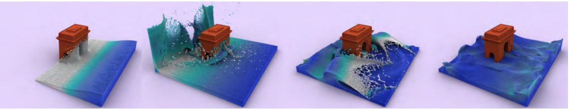
- Geodesic distance, interpolations, barycenters, gradient flows, Riemannian interpretation of the 2-*Wasserstein space* $(\mathcal{P}_2(\mathbb{R}^d), \mathrm{W}_2)$...
(Otto, 2001; Ambrosio, Gigli, Savaré, 2004)



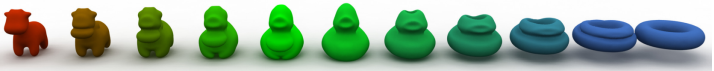
Introduction

Optimal Transport applications

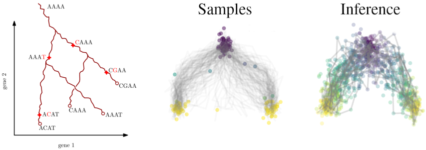
- **Euler equations**: (de Goes et al., 2015)



- **Computer graphics**: (Salomon et al., 2015)



- **Trajectory inference** for single cell RNA-sequencing data: (Forrow et al., 2021; Chizat et al., 2022)



- **Cosmology**, **quantum chemistry**, **meteorology**, **economics**, **image processing**, **machine learning**...

Introduction

"Entropy-regularized" Optimal Transport problem

$$\inf_{\gamma \in \Gamma(\mu,\nu)} \int_{\Omega\times\Omega} c(x,y)\mathrm{d}\gamma(x,y).$$

- In practice, optimal transport value can be:
  - Difficult to compute numerically:
    $\tilde{O}(n^3)$ numerical complexity when μ, ν have n support points.
  - Difficult to estimate statistically:
    $O(n^{-1/d})$ sample complexity when μ, ν are supported over $\mathbb{R}^d$.

Introduction

"Entropy-regularized" Optimal Transport problem

$$\boxed{\inf_{\gamma \in \Gamma(\mu,\nu)} \int_{\Omega\times\Omega} c(x,y)\mathrm{d}\gamma(x,y).}$$

- In practice, optimal transport value can be:
  - Difficult to compute numerically:
    $\tilde{O}(n^3)$ numerical complexity when μ, ν have n support points.
  - Difficult to estimate statistically:
    $O(n^{-1/d})$ sample complexity when μ, ν are supported over $\mathbb{R}^d$.

"Entropy-regularized" Optimal Transport problem:

- Given $\mu, \nu \in \mathcal{P}(\Omega)$, $c : \Omega \times \Omega \to \mathbb{R}$ and $\lambda > 0$, solve

$$\boxed{\inf_{\gamma \in \Gamma(\mu,\nu)} \int_{\Omega\times\Omega} c(x,y)\mathrm{d}\gamma(x,y) + \lambda \mathrm{KL}(\gamma | \mu \otimes \nu).}$$

Equivalent to the static Schrödinger problem (Schrödinger, 1931; Léonard, 2014).

Introduction

"Entropy-regularized" Optimal Transport problem

$$\inf_{\gamma \in \Gamma(\mu,\nu)} \int_{\Omega \times \Omega} c(x,y) \mathrm{d}\gamma(x,y) + \lambda \mathrm{KL}(\gamma | \mu \otimes \nu).$$

Introduction

"Entropy-regularized" Optimal Transport problem

$$\boxed{\inf_{\gamma\in\Gamma(\mu,\nu)} \int_{\Omega\times\Omega} c(x,y)\mathrm{d}\gamma(x,y) + \lambda \mathrm{KL}(\gamma|\mu\otimes\nu).}$$

- Dual problem:

$$\sup_{\phi\in L^1(\mu),\psi\in L^1(\nu)} \int \phi \mathrm{d}\mu + \int \psi \mathrm{d}\nu + \lambda\left(1 - \int\int \exp\left(\frac{\phi\oplus\psi - c}{\lambda}\right)\mathrm{d}\mu\mathrm{d}\nu\right).$$

Introduction

"Entropy-regularized" Optimal Transport problem

$$\inf_{\gamma \in \Gamma(\mu,\nu)} \int_{\Omega \times \Omega} c(x,y) \mathrm{d}\gamma(x,y) + \lambda \mathrm{KL}(\gamma | \mu \otimes \nu).$$

- Dual problem:

$$\sup_{\phi \in L^1(\mu), \psi \in L^1(\nu)} \int \phi \mathrm{d}\mu + \int \psi \mathrm{d}\nu + \lambda \left(1 - \int \int \exp\left(\frac{\phi \oplus \psi - c}{\lambda}\right) \mathrm{d}\mu \mathrm{d}\nu\right).$$

- Primal-dual relation:

$$\gamma^* = \exp\left(\frac{\phi^* \oplus \psi^* - c}{\lambda}\right).$$

Introduction

"Entropy-regularized" Optimal Transport problem

$$\boxed{\sup_{\phi \in L^1(\mu), \psi \in L^1(\nu)} \int \phi \mathrm{d}\mu + \int \psi \mathrm{d}\nu + \lambda \left(1 - \int \int \exp\left(\frac{\phi \oplus \psi - c}{\lambda} \right) \mathrm{d}\mu \mathrm{d}\nu \right).}$$

Introduction

"Entropy-regularized" Optimal Transport problem

$$\sup_{\phi\in L^1(\mu),\psi\in L^1(\nu)} \int \phi \mathrm{d}\mu + \int \psi \mathrm{d}\nu + \lambda\left(1 - \int\int \exp\left(\frac{\phi\oplus\psi - c}{\lambda}\right)\mathrm{d}\mu\mathrm{d}\nu\right).$$

- **Optimality conditions** yield the *Schrödinger system*:

$$\begin{cases} \phi^*(x) = -\lambda \log \int \exp\left(\frac{\psi^*(y)-c(x,y)}{\lambda}\right)\mathrm{d}\nu(y) & \text{for } \mu\text{-a.e. } x, \\ \psi^*(y) = -\lambda \log \int \exp\left(\frac{\phi^*(x)-c(x,y)}{\lambda}\right)\mathrm{d}\mu(x) & \text{for } \nu\text{-a.e. } y. \end{cases}$$

Introduction

"Entropy-regularized" Optimal Transport problem

$$\boxed{\sup_{\phi\in L^1(\mu),\psi\in L^1(\nu)} \int \phi \mathrm{d}\mu + \int \psi \mathrm{d}\nu + \lambda \left(1 - \int\int \exp\left(\frac{\phi \oplus \psi - c}{\lambda}\right) \mathrm{d}\mu \mathrm{d}\nu\right).}$$

- **Optimality conditions** yield the *Schrödinger system*:

$$\begin{cases} \phi^*(x) = -\lambda \log \int \exp\left(\frac{\psi^*(y) - c(x,y)}{\lambda}\right) \mathrm{d}\nu(y) & \text{for } \mu\text{-a.e. } x, \\ \psi^*(y) = -\lambda \log \int \exp\left(\frac{\phi^*(x) - c(x,y)}{\lambda}\right) \mathrm{d}\mu(x) & \text{for } \nu\text{-a.e. } y. \end{cases}$$

Sinkhorn's algorithm: starting from arbitraty $\psi_0 \in L^1(\nu)$, set $\forall t \in \mathbb{N}$

$$\begin{cases} \phi_{t+\frac{1}{2}}(x) = -\lambda \log \int \exp\left(\frac{\psi_t(y) - c(x,y)}{\lambda}\right) \mathrm{d}\nu(y) & \text{for } \mu\text{-a.e. } x, \\ \psi_{t+1}(y) = -\lambda \log \int \exp\left(\frac{\phi_{t+1/2}(x) - c(x,y)}{\lambda}\right) \mathrm{d}\mu(x) & \text{for } \nu\text{-a.e. } y. \end{cases}$$

Introduction

Sinkhorn's algorithm

$$\begin{cases} \phi_{t+\frac{1}{2}}(x) = -\lambda \log \int \exp\left(\frac{\psi_t(y) - c(x,y)}{\lambda}\right) \mathrm{d}\nu(y) & \text{for } \mu\text{-a.e. } x, \\ \psi_{t+1}(y) = -\lambda \log \int \exp\left(\frac{\phi_{t+1/2}(x) - c(x,y)}{\lambda}\right) \mathrm{d}\mu(x) & \text{for } \nu\text{-a.e. } y. \end{cases}$$

- Also known as:
    - Sinkhorn-Knopp algorithm,
    - Iterative Proportional Fitting Procedure (IPFP),
    - RAS algorithm,
    - Fortet's iterations,
    - Bregman alternative projection,
    - Matrix scaling algorithm...

Introduction

Sinkhorn's algorithm

$$\begin{cases} \phi_{t+\frac{1}{2}}(x) = -\lambda \log \int \exp\left(\frac{\psi_t(y)-c(x,y)}{\lambda}\right) \mathrm{d}\nu(y) & \text{for } \mu\text{-a.e. } x, \\ \psi_{t+1}(y) = -\lambda \log \int \exp\left(\frac{\phi_{t+1/2}(x)-c(x,y)}{\lambda}\right) \mathrm{d}\mu(x) & \text{for } \nu\text{-a.e. } y. \end{cases}$$

- Link with **matrix scaling**: when $\mu = \sum_{i=1}^n \mu_i \delta_{x_i}$ and $\nu = \sum_{j=1}^n \nu_j \delta_{y_j}$, set:

$$\begin{cases} \mu = (\mu_i)_{1\le i\le n} \in \mathbb{R}^n, \\ \nu = (\nu_j)_{1\le j\le n} \in \mathbb{R}^n, \\ u_{t+\frac{1}{2}} = (e^{\frac{\phi_{t+1/2}(x_i)}{\lambda}} \mu_i)_{1\le i\le n} \in \mathbb{R}^n, \\ v_t = (e^{\frac{\psi_t(y_j)}{\lambda}} \nu_j)_{1\le j\le n} \in \mathbb{R}^n, \\ \text{and } K = (e^{-c(x_i,y_j)/\lambda})_{1\le i,j\le n} \in \mathbb{R}^{n\times n}. \end{cases} \quad \text{Then: } \begin{cases} u_{t+\frac{1}{2}} = \mu \oslash K v_t, \\ v_{t+1} = \nu \oslash K^\top u_{t+\frac{1}{2}}. \end{cases}$$

Introduction

Sinkhorn's algorithm

$$\begin{cases} \phi_{t+\frac{1}{2}}(x) = -\lambda \log \int \exp\left(\frac{\psi_t(y)-c(x,y)}{\lambda}\right) \mathrm{d}\nu(y) & \text{for } \mu\text{-a.e. } x, \\ \psi_{t+1}(y) = -\lambda \log \int \exp\left(\frac{\phi_{t+1/2}(x)-c(x,y)}{\lambda}\right) \mathrm{d}\mu(x) & \text{for } \nu\text{-a.e. } y. \end{cases}$$

- Link with **matrix scaling**: when $\mu = \sum_{i=1}^n \mu_i \delta_{x_i}$ and $\nu = \sum_{j=1}^n \nu_j \delta_{y_j}$, set:

$$\begin{cases} \mu = (\mu_i)_{1\le i\le n} \in \mathbb{R}^n, \\ \nu = (\nu_j)_{1\le j\le n} \in \mathbb{R}^n, \\ u_{t+\frac{1}{2}} = (e^{\frac{\phi_{t+1/2}(x_i)}{\lambda}} \mu_i)_{1\le i\le n} \in \mathbb{R}^n, \\ v_t = (e^{\frac{\psi_t(y_j)}{\lambda}} \nu_j)_{1\le j\le n} \in \mathbb{R}^n, \\ \text{and } K = (e^{-c(x_i,y_j)/\lambda})_{1\le i,j\le n} \in \mathbb{R}^{n\times n}. \end{cases} \quad \text{Then: } \begin{cases} u_{t+\frac{1}{2}} = \mu \oslash K v_t, \\ v_{t+1} = \nu \oslash K^\top u_{t+\frac{1}{2}}. \end{cases}$$

Theorem (Sinkhorn, 1964): The sequences $(u_t)_t, (v_t)_t$ converge to the *unique scalings* u^*, v^* of the matrix K that satisfy

$$\boxed{\gamma^* := \mathrm{diag}(u^*) K \mathrm{diag}(v^*) \in \Gamma(\mu, \nu).}$$

Introduction

Sinkhorn's algorithm

$$\begin{cases} \phi_{t+\frac{1}{2}}(x) = -\lambda \log \int \exp\left(\frac{\psi_t(y)-c(x,y)}{\lambda}\right) \mathrm{d}\nu(y) & \text{for } \mu\text{-a.e. } x, \\ \psi_{t+1}(y) = -\lambda \log \int \exp\left(\frac{\phi_{t+1/2}(x)-c(x,y)}{\lambda}\right) \mathrm{d}\mu(x) & \text{for } \nu\text{-a.e. } y. \end{cases}$$

- Link with **matrix scaling**: when $\mu = \sum_{i=1}^n \mu_i \delta_{x_i}$ and $\nu = \sum_{j=1}^n \nu_j \delta_{y_j}$, set:

$$\begin{cases} \mu = (\mu_i)_{1\le i \le n} \in \mathbb{R}^n, \\ \nu = (\nu_j)_{1\le j\le n} \in \mathbb{R}^n, \\ u_{t+\frac{1}{2}} = (e^{\frac{\phi_{t+1/2}(x_i)}{\lambda}} \mu_i)_{1\le i\le n} \in \mathbb{R}^n, \\ v_t = (e^{\frac{\psi_t(y_j)}{\lambda}} \nu_j)_{1\le j\le n} \in \mathbb{R}^n, \\ \text{and } K = (e^{-c(x_i,y_j)/\lambda})_{1\le i,j\le n} \in \mathbb{R}^{n\times n}. \end{cases}$$

Then: $$\begin{cases} u_{t+\frac{1}{2}} = \mu \oslash K v_t, \\ v_{t+1} = \nu \oslash K^\top u_{t+\frac{1}{2}}. \end{cases}$$

Theorem (Sinkhorn, 1964): The sequences $(u_t)_t, (v_t)_t$ converge to the *unique scalings* u^*, v^* of the matrix K that satisfy

$$\boxed{\gamma^* := \mathrm{diag}(u^*) K \mathrm{diag}(v^*) \in \Gamma(\mu,\nu).}$$

$\rightarrow$ **What is the speed of this convergence?**

Introduction

Sinkhorn's algorithm - Known convergence rates

- **Hilbert's projective metric on $(\mathbb{R}_+^*)^n$:**

$$\forall u, \tilde{u} \in (\mathbb{R}_+^*)^n, \quad d_{\mathcal{H}}(u, \tilde{u}) = \log \max_{i,j} \frac{u_i \tilde{u}_j}{u_j \tilde{u}_i} = \|\log u - \log \tilde{u}\|_{osc}.$$

Theorem (Birkhoff, 1957; Samelson et al., 1957):
Any matrix $K \in (\mathbb{R}_+^*)^{n \times n}$ is a contraction on $(\mathbb{R}_+^*)^n$ with respect to $d_{\mathcal{H}}$:

$$\forall u, \tilde{u} \in (\mathbb{R}_+^*)^n, \quad d_{\mathcal{H}}(Ku, K\tilde{u}) \leq \kappa(K) d_{\mathcal{H}}(u, \tilde{u}).$$

Introduction

Sinkhorn's algorithm - Known convergence rates

- **Hilbert's projective metric on $(\mathbb{R}_+^*)^n$:**

$$\forall u, \tilde{u} \in (\mathbb{R}_+^*)^n, \quad d_{\mathcal{H}}(u, \tilde{u}) = \log \max_{i,j} \frac{u_i \tilde{u}_j}{u_j \tilde{u}_i} = \|\log u - \log \tilde{u}\|_{osc}.$$

Theorem (Birkhoff, 1957; Samelson et al., 1957):
Any matrix $K \in (\mathbb{R}_+^*)^{n \times n}$ is a contraction on $(\mathbb{R}_+^*)^n$ with respect to $d_{\mathcal{H}}$:

$$\forall u, \tilde{u} \in (\mathbb{R}_+^*)^n, \quad d_{\mathcal{H}}(Ku, K\tilde{u}) \leq \kappa(K) d_{\mathcal{H}}(u, \tilde{u}).$$

Corollary (Franklin and Lorenz, 1989):
The Sinkhorn sequences satisfy:

$$\begin{cases} \|\phi_t - \phi_*\|_{osc} \leq (1 - e^{-c_\infty/\lambda})^t \|\phi_0 - \phi_*\|_{osc}, \\ \|\psi_t - \psi_*\|_{osc} \leq (1 - e^{-c_\infty/\lambda})^t \|\psi_0 - \psi_*\|_{osc}, \end{cases}$$

where $c_\infty = \|c\|_{osc} = \sup c - \inf c$.

Introduction

Sinkhorn's algorithm - Known convergence rates

- **Hilbert's projective metric on $(\mathbb{R}_+^*)^n$:**

$$\forall u, \tilde{u} \in (\mathbb{R}_+^*)^n, \quad d_{\mathcal{H}}(u, \tilde{u}) = \log \max_{i,j} \frac{u_i \tilde{u}_j}{u_j \tilde{u}_i} = \|\log u - \log \tilde{u}\|_{osc}.$$

Theorem (Birkhoff, 1957; Samelson et al., 1957):
Any matrix $K \in (\mathbb{R}_+^*)^{n \times n}$ is a contraction on $(\mathbb{R}_+^*)^n$ with respect to $d_{\mathcal{H}}$:

$$\forall u, \tilde{u} \in (\mathbb{R}_+^*)^n, \quad d_{\mathcal{H}}(Ku, K\tilde{u}) \leq \kappa(K) d_{\mathcal{H}}(u, \tilde{u}).$$

Corollary (Franklin and Lorenz, 1989):
The Sinkhorn sequences satisfy:

$$\begin{cases} \|\phi_t - \phi_*\|_{osc} \leq (1 - e^{-c_\infty/\lambda})^t \|\phi_0 - \phi_*\|_{osc}, \\ \|\psi_t - \psi_*\|_{osc} \leq (1 - e^{-c_\infty/\lambda})^t \|\psi_0 - \psi_*\|_{osc}, \end{cases}$$

where $c_\infty = \|c\|_{osc} = \sup c - \inf c$.

Problem: The constant $e^{-c_\infty/\lambda}$ is *very small* when λ is *small*.

Introduction

Sinkhorn's algorithm - Known convergence rates

- **Sub-optimality gap:** $\forall t, \quad \boxed{\delta_t = F(\phi_*, \psi_*) - F(\phi_{t+1/2}, \psi_t),}$
  where $F(\phi, \psi) = \langle \phi | \mu \rangle + \langle \psi | \nu \rangle + \lambda(1 - \langle e^{\frac{\phi \oplus \psi - c}{\lambda}} | \mu \otimes \nu \rangle)$.

Theorem (Dvurechensky, Gasnikov and Kroshnin, 2018):
The sub-optimality satisfies:

$$\delta_t \leq \frac{2c_\infty^2}{\lambda t}.$$

Introduction

Sinkhorn's algorithm - Known convergence rates

- **Sub-optimality gap:** $\forall t, \quad \boxed{\delta_t = F(\phi_*, \psi_*) - F(\phi_{t+1/2}, \psi_t),}$
  where $F(\phi, \psi) = \langle \phi | \mu \rangle + \langle \psi | \nu \rangle + \lambda(1 - \langle e^{\frac{\phi \oplus \psi - c}{\lambda}} | \mu \otimes \nu \rangle)$.

Theorem (Dvurechensky, Gasnikov and Kroshnin, 2018):
The sub-optimality satisfies:

$$\delta_t \leq \frac{2c_\infty^2}{\lambda t}.$$

Problem: Polynomial convergence rate instead of exponential convergence rate.

Main result

Exponential convergence rates with robust contraction constants.

- Case 1: log-concave source measure.

Theorem (Chizat, D. and Vaškevičius, 2024):

- Let $c(x, y) = -\langle x, y \rangle$.
- Let $\mathcal{X} \subset \mathbb{R}^d$ be compact and convex, let $\mu \in \mathcal{P}_{a.c.}(\mathcal{X})$ with log-concave density.
- Let $\mathcal{Y} \subset \mathbb{R}^d$ be compact and $\nu \in \mathcal{P}(\mathcal{Y})$.

If $\lambda \leq c_\infty$, then

$$\forall t \geq 0, \qquad \delta_t \leq \delta_0 \left(1 - \frac{\lambda}{2^9 c_\infty}\right)^t.$$

Main result

Exponential convergence rates with robust contraction constants.

- Case 2: source measure with bounded density.

Theorem (Chizat, D. and Vaškevičius, 2024):

- Let $c(x, y) = -\langle x, y\rangle$.
- Let $\mathcal{X} \subset \mathbb{R}^d$ be compact and convex, let $\mu \in \mathcal{P}_{a.c.}(\mathcal{X})$ and assume that the density of μ satisfies

$$0 < m \leq f_\mu \leq M < +\infty.$$

- Let $\mathcal{Y} \subset \mathbb{R}^d$ be compact and $\nu \in \mathcal{P}(\mathcal{Y})$.

If $\lambda \leq c_\infty$, then

$$\forall t \geq 0, \qquad \delta_t \leq \delta_0 \left(1 - \frac{m}{2^{10} M} \frac{\lambda^2}{c_\infty^2}\right)^t.$$

Main result

Exponential convergence rates with robust contraction constants.

- Case 2: source measure with bounded density.

Theorem (Chizat, D. and Vaškevičius, 2024):

- Let $c(x, y) = -\langle x, y \rangle$.
- Let $\mathcal{X} \subset \mathbb{R}^d$ be compact and convex, let $\mu \in \mathcal{P}_{a.c.}(\mathcal{X})$ and assume that the density of μ satisfies

$$0 < m \leq f_\mu \leq M < +\infty.$$

- Let $\mathcal{Y} \subset \mathbb{R}^d$ be compact and $\nu \in \mathcal{P}(\mathcal{Y})$.

If $\lambda \leq c_\infty$, then

$$\forall t \geq 0, \qquad \delta_t \leq \delta_0 \left(1 - \frac{m}{2^{10} M} \frac{\lambda^2}{c_\infty^2}\right)^t .$$

Remarks:

- Convexity of $\mathcal{X}$ may be relaxed.
- Cost c may be any $\mathcal{C}^2$ function.
- In certain settings, $\frac{\lambda^2}{c_\infty^2}$ may be replaced with $\frac{\lambda}{c_\infty}$ for t *large enough*.

Elements of proof

Preamble: semi-dual functional

- Recall we want to solve

$$\boxed{\sup_{\phi \in L^1(\mu), \psi \in L^1(\nu)} F(\phi, \psi),}$$

where $F(\phi, \psi) = \int \phi \mathrm{d}\mu + \int \psi \mathrm{d}\nu + \lambda \left(1 - \int \int \exp\left(\frac{\phi \oplus \psi - c}{\lambda}\right) \mathrm{d}\mu \mathrm{d}\nu\right).$

Elements of proof

Preamble: semi-dual functional

- Recall we want to solve

$$\boxed{\sup_{\phi\in L^1(\mu),\psi\in L^1(\nu)} F(\phi,\psi),}$$

where $F(\phi,\psi)=\int \phi \mathrm{d}\mu + \int \psi \mathrm{d}\nu + \lambda\left(1-\int\int \exp\left(\frac{\phi\oplus\psi-c}{\lambda}\right)\mathrm{d}\mu\mathrm{d}\nu\right).$

- **Semi-dual functional:** for any $\psi \in L^1(\nu)$, define

$$\begin{aligned} E(\psi) &= \sup_{\phi\in L^1(\mu)} F(\phi,\psi) \\ &= \int \psi^{c,\lambda}\mathrm{d}\mu + \int \psi \mathrm{d}\nu. \end{aligned}$$

where $\psi^{c,\lambda}(x) = -\lambda \log \int e^{\frac{\psi(y)-c(x,y)}{\lambda}}\mathrm{d}\nu(y).$

Elements of proof

Preamble: semi-dual functional

- Recall we want to solve

$$\boxed{\sup_{\phi\in L^1(\mu),\psi\in L^1(\nu)} F(\phi,\psi),}$$

where $F(\phi,\psi)=\int\phi\mathrm{d}\mu+\int\psi\mathrm{d}\nu+\lambda\left(1-\int\int\exp\left(\frac{\phi\oplus\psi-c}{\lambda}\right)\mathrm{d}\mu\mathrm{d}\nu\right).$

- **Semi-dual functional:** for any $\psi\in L^1(\nu)$, define

$$\begin{aligned}E(\psi)&=\sup_{\phi\in L^1(\mu)} F(\phi,\psi)\\&=\int\psi^{c,\lambda}\mathrm{d}\mu+\int\psi\mathrm{d}\nu.\end{aligned}$$

where $\psi^{c,\lambda}(x)=-\lambda\log\int e^{\frac{\psi(y)-c(x,y)}{\lambda}}\mathrm{d}\nu(y).$

- **New problem:** solve

$$\boxed{\sup_{\psi\in L^1(\nu)} E(\psi).}$$

Elements of proof

Preamble: semi-dual functional

$$E : \psi \mapsto \int \psi^{c,\lambda} \mathrm{d}\mu + \int \psi \mathrm{d}\nu, \quad \text{where} \quad \psi^{c,\lambda}(x) = -\lambda \log \int e^{\frac{\psi(y) - c(x,y)}{\lambda}} \mathrm{d}\nu(y).$$

Key properties:

Elements of proof

Preamble: semi-dual functional

$$E : \psi \mapsto \int \psi^{c,\lambda} \mathrm{d}\mu + \int \psi \mathrm{d}\nu, \quad \text{where} \quad \psi^{c,\lambda}(x) = -\lambda \log \int e^{\frac{\psi(y) - c(x,y)}{\lambda}} \mathrm{d}\nu(y).$$

Key properties:

1. Sub-optimality:

$$\boxed{\delta_t = E(\psi_*) - E(\psi_t).}$$

Elements of proof

Preamble: semi-dual functional

$$E : \psi \mapsto \int \psi^{c,\lambda} \mathrm{d}\mu + \int \psi \mathrm{d}\nu, \quad \text{where} \quad \psi^{c,\lambda}(x) = -\lambda \log \int e^{\frac{\psi(y) - c(x,y)}{\lambda}} \mathrm{d}\nu(y).$$

Key properties:

1. Sub-optimality:

$$\boxed{\delta_t = E(\psi_*) - E(\psi_t).}$$

2. One-step-improvement:

$$\boxed{\delta_{t+1} \le \delta_t - \lambda \mathrm{KL}(\nu | \nu[\psi_t]),}$$

   where $\nu[\psi](y) = \int e^{\frac{\psi^{c,\lambda}(x) + \psi(y) - c(x,y)}{\lambda}} \mathrm{d}\mu(x)$.

Elements of proof

Preamble: semi-dual functional

$$E : \psi \mapsto \int \psi^{c,\lambda} \mathrm{d}\mu + \int \psi \mathrm{d}\nu, \quad \text{where} \quad \psi^{c,\lambda}(x) = -\lambda \log \int e^{\frac{\psi(y)-c(x,y)}{\lambda}} \mathrm{d}\nu(y).$$

Key properties:

1. Sub-optimality:

$$\delta_t = E(\psi_*) - E(\psi_t).$$

2. One-step-improvement:

$$\delta_{t+1} \leq \delta_t - \lambda \mathrm{KL}(\nu | \nu[\psi_t]),$$

where $\nu[\psi](y) = \int e^{\frac{\psi^{c,\lambda}(x)+\psi(y)-c(x,y)}{\lambda}} \mathrm{d}\mu(x)$.

3. E is concave and its *gradient* is

$$\nabla E(\psi) = \nu - \nu[\psi].$$

Elements of proof

One-step-improvement bound

- By concavity of E,

$$\delta_t \leq \langle \psi^* - \psi_t | \nu - \nu[\psi_t] \rangle.$$

Elements of proof

One-step-improvement bound

- By concavity of E,

$$\delta_t \leq \langle \psi^* - \psi_t | \nu - \nu[\psi_t] \rangle.$$

- For all $\eta > 0$,

$$\begin{aligned}
\delta_t &= \eta^{-1}\{\eta\langle \psi^* - \psi_t | \nu - \nu[\psi_t]\rangle - \mathrm{KL}(\nu|\nu[\psi_t])\} + \eta^{-1}\mathrm{KL}(\nu|\nu[\psi_t]) \\
&\leq \eta^{-1} \sup_{\nu' \in \mathcal{P}(\mathbb{R}^d)} \{\eta\langle \psi^* - \psi_t | \nu' - \nu[\psi_t]\rangle - \mathrm{KL}(\nu'|\nu[\psi_t])\} + \eta^{-1}\mathrm{KL}(\nu|\nu[\psi_t]) \\
&= \eta^{-1} \log \mathbb{E}_{\nu[\psi_t]} \exp(\eta f) + \eta^{-1}\mathrm{KL}(\nu|\nu[\psi_t]),
\end{aligned}$$

where $f = \psi^* - \psi_t - \mathbb{E}_{\nu[\psi_t]}[\psi^* - \psi_t]$.

$$\delta_t \leq \eta^{-1} \log \mathbb{E}_{\nu[\psi_t]} \exp(\eta f) + \eta^{-1}\mathrm{KL}(\nu|\nu[\psi_t]).$$

Elements of proof

One-step-improvement bound

$$\delta_t \leq \eta^{-1} \log \mathbb{E}_{\nu[\psi_t]} \exp(\eta f) + \eta^{-1} \mathrm{KL}(\nu | \nu[\psi_t]).$$

- Recovering the polynomial rate:

Elements of proof

One-step-improvement bound

$$\delta_t \leq \eta^{-1} \log \mathbb{E}_{\nu[\psi_t]} \exp(\eta f) + \eta^{-1}\mathrm{KL}(\nu|\nu[\psi_t]).$$

- Recovering the polynomial rate:

  1. Using $\|f\|_{osc} = \|\psi^* - \psi_t\|_{osc} \leq 2c_\infty$, Hoeffding's inequality yields
  $$\mathbb{E}_{\nu[\psi_t]} \exp(\eta f) \leq \exp(2\eta^2 c_\infty^2).$$
  2. Injecting and optimizing in η yields
  $$\delta_t \leq c_\infty \sqrt{2\mathrm{KL}(\nu|\nu[\psi_t])}.$$
  3. Combining with the one-step-improvement $\delta_{t+1} \leq \delta_t - \lambda \mathrm{KL}(\nu|\nu[\psi_t])$,
  $$\delta_t \leq c_\infty \sqrt{2\lambda^{-1}(\delta_t - \delta_{t+1})}.$$
  4. Re-arranging leads to $\frac{\lambda}{2c_\infty^2} \leq \frac{1}{\delta_{t+1}} - \frac{1}{\delta_t}$, which yields
  $$\delta_t \leq \frac{2c_\infty^2}{\lambda t}.$$

Elements of proof

One-step-improvement bound

$$\delta_t \leq \eta^{-1} \log \mathbb{E}_{\nu[\psi_t]} \exp(\eta f) + \eta^{-1} \mathrm{KL}(\nu | \nu[\psi_t]).$$

- Using $\|f\|_{osc} \leq 2c_\infty$, **Bernstein's inequality** yields

$$\mathbb{E}_{\nu[\psi_t]} \left[\exp\left(\eta f\right)\right] \leq \exp\left(\frac{\eta^2 \mathbb{V}\mathrm{ar}_{\nu[\psi_t]}(\psi^* - \psi_t)}{2(1 - \eta \frac{2c_\infty}{3})}\right).$$

Elements of proof

One-step-improvement bound

$$\delta_t \leq \eta^{-1} \log \mathbb{E}_{\nu[\psi_t]} \exp(\eta f) + \eta^{-1} \mathrm{KL}(\nu|\nu[\psi_t]).$$

- Using $\|f\|_{osc} \leq 2c_\infty$, **Bernstein's inequality** yields

$$\mathbb{E}_{\nu[\psi_t]} [\exp(\eta f)] \leq \exp\left(\frac{\eta^2 \mathbb{V}\mathrm{ar}_{\nu[\psi_t]}(\psi^* - \psi_t)}{2(1 - \eta\frac{2c_\infty}{3})}\right).$$

- Consequence:

Proposition (Chizat, D. and Vaškevičius, 2024):

$$\delta_t \leq 2\sqrt{\lambda^{-1}\mathbb{V}\mathrm{ar}_\nu(\psi^* - \psi_t)(\delta_t - \delta_{t+1})} + \frac{14c_\infty}{3}\lambda^{-1}(\delta_t - \delta_{t+1}).$$

Elements of proof

One-step-improvement bound

$$\delta_t \leq \eta^{-1} \log \mathbb{E}_{\nu[\psi_t]} \exp(\eta f) + \eta^{-1} \mathrm{KL}(\nu | \nu[\psi_t]).$$

- Using $\|f\|_{osc} \leq 2c_\infty$, **Bernstein's inequality** yields

$$\mathbb{E}_{\nu[\psi_t]} [\exp(\eta f)] \leq \exp\left(\frac{\eta^2 \mathbb{V}\mathrm{ar}_{\nu[\psi_t]}(\psi^* - \psi_t)}{2(1 - \eta \frac{2c_\infty}{3})} \right).$$

- Consequence:

Proposition (Chizat, D. and Vaškevičius, 2024):

$$\delta_t \leq 2\sqrt{\lambda^{-1} \mathbb{V}\mathrm{ar}_\nu(\psi^* - \psi_t)(\delta_t - \delta_{t+1})} + \frac{14 c_\infty}{3} \lambda^{-1} (\delta_t - \delta_{t+1}).$$

$\rightarrow$ **To conclude, need to relate $\mathbb{V}\mathrm{ar}_\nu(\psi^* - \psi_t)$ back to δ_t and δ_{t+1}.**

Elements of proof

Strong-concavity estimate

- With $v = \psi^* - \psi_t$, sub-optimality satisfies

$$\delta_t = E(\psi^*) - E(\psi_t) = -\int_{\varepsilon=0}^{1} \int_{s=\varepsilon}^{1} \frac{\mathrm{d}^2}{\mathrm{d}s^2} E(\psi_t + sv) \mathrm{d}s \mathrm{d}\varepsilon.$$

Elements of proof

Strong-concavity estimate

- With $v = \psi^* - \psi_t$, sub-optimality satisfies

$$\delta_t = E(\psi^*) - E(\psi_t) = -\int_{\varepsilon=0}^{1}\int_{s=\varepsilon}^{1} \frac{\mathrm{d}^2}{\mathrm{d}s^2} E(\psi_t + sv)\mathrm{d}s\mathrm{d}\varepsilon.$$

- Second-order derivative of E: $\forall \psi, v \in L^1(\nu), \varepsilon \in \mathbb{R}$,

$$\boxed{\frac{\mathrm{d}^2}{\mathrm{d}\varepsilon^2} E(\psi + sv) = -\frac{1}{\lambda}\int \mathbb{V}\mathrm{ar}_{\nu_x[\psi+sv]}(v)\mathrm{d}\mu(x),}$$

where $\nu_x[\psi](y) = e^{\frac{\psi^{c,\lambda}(x)+\psi(y)-c(x,y)}{\lambda}}$ is s.t. $\nu[\psi] = \int \nu_x[\psi]\mathrm{d}\mu(x)$.

Elements of proof

Strong-concavity estimate

- With $v = \psi^* - \psi_t$, sub-optimality satisfies

$$\delta_t = E(\psi^*) - E(\psi_t) = -\int_{\varepsilon=0}^{1}\int_{s=\varepsilon}^{1} \frac{\mathrm{d}^2}{\mathrm{d}s^2} E(\psi_t + sv)\mathrm{d}s\mathrm{d}\varepsilon.$$

- Second-order derivative of E: $\forall \psi, v \in L^1(\nu), \varepsilon \in \mathbb{R}$,

$$\boxed{\frac{\mathrm{d}^2}{\mathrm{d}\varepsilon^2} E(\psi + sv) = -\frac{1}{\lambda}\int \mathbb{V}\mathrm{ar}_{\nu_x[\psi+sv]}(v)\mathrm{d}\mu(x),}$$

where $\nu_x[\psi](y) = e^{\frac{\psi^{c,\lambda}(x)+\psi(y)-c(x,y)}{\lambda}}$ is s.t. $\nu[\psi] = \int \nu_x[\psi]\mathrm{d}\mu(x)$.

$$\Longrightarrow \boxed{\delta_t = \frac{1}{\lambda}\int_{\varepsilon=0}^{1}\int_{s=\varepsilon}^{1}\int \mathbb{V}\mathrm{ar}_{\nu_x[\psi+sv]}(v)\mathrm{d}\mu(x)\mathrm{d}s\mathrm{d}\varepsilon.}$$

But $\int \mathbb{V}\mathrm{ar}_{\nu_x[\psi+sv]}(v)\mathrm{d}\mu(x) \leq \mathbb{V}\mathrm{ar}_{\nu[\psi+sv]}(v)$, and we need a reverse inequality.

Elements of proof

Strong-concavity estimate

- We need a way to upper bound $\frac{\mathrm{d}^2}{\mathrm{d}s^2} E(\psi + sv)$ in terms of $\mathbb{V}\mathrm{ar}_{\nu[\psi+sv]}(v)$.

Elements of proof

Strong-concavity estimate

- We need a way to upper bound $\frac{\mathrm{d}^2}{\mathrm{d}s^2} E(\psi + sv)$ in terms of $\mathbb{V}\mathrm{ar}_{\nu[\psi+sv]}(v)$.

- **Log-partition function:** for any $\psi \in L^1(\nu)$, define

$$I(\psi) = \log \int \exp(\psi^{c,\lambda}) \mathrm{d}\mu.$$

- I is *twice-differentiable* and satisfies

$$\frac{\mathrm{d}^2}{\mathrm{d}s^2} I(\psi + sv) \geq C(\lambda) \frac{\mathrm{d}^2}{\mathrm{d}s^2} E(\psi + sv) + \tilde{C}(\lambda) \mathbb{V}\mathrm{ar}_{\nu[\psi+sv]}(v).$$

Elements of proof

Strong-concavity estimate

Theorem (Prékopa, 1971/73; Leindler, 1972; Cordero-Erausquin et al., 2006): *Weighted Prékopa-Leindler inequality*.

Let $\xi \geq 0$ and ρ be a measure on $\mathbb{R}^d$ of the form $\mathrm{d}\rho = e^{-W}$ where $\nabla^2 W \succeq \xi$. Let $\alpha \in [0,1]$ and let $f, g, h : \mathbb{R}^d \to \mathbb{R}_+$ be such that for all $x, y \in \mathbb{R}^d$,

$$h((1-\alpha)x + \alpha y) \geq e^{-\xi\alpha(1-\alpha)\|x-y\|^2/2} f(x)^{1-\alpha} g(y)^{\alpha}.$$

Then,

$$\int_{\mathbb{R}^d} h \mathrm{d}\rho \geq \left(\int_{\mathbb{R}^d} f \mathrm{d}\rho\right)^{1-s} \left(\int_{\mathbb{R}^d} g \mathrm{d}\rho\right)^{s}.$$

Elements of proof

Strong-concavity estimate

Theorem (Prékopa, 1971/73; Leindler, 1972; Cordero-Erausquin et al., 2006): *Weighted Prékopa-Leindler inequality*.

Let $\xi \geq 0$ and ρ be a measure on $\mathbb{R}^d$ of the form $\mathrm{d}\rho = e^{-W}$ where $\nabla^2 W \succeq \xi$. Let $\alpha \in [0,1]$ and let $f, g, h : \mathbb{R}^d \to \mathbb{R}_+$ be such that for all $x, y \in \mathbb{R}^d$,

$$h((1-\alpha)x + \alpha y) \geq e^{-\xi\alpha(1-\alpha)\|x-y\|^2/2} f(x)^{1-\alpha} g(y)^{\alpha}.$$

Then,

$$\int_{\mathbb{R}^d} h \mathrm{d}\rho \geq \left(\int_{\mathbb{R}^d} f \mathrm{d}\rho\right)^{1-s} \left(\int_{\mathbb{R}^d} g \mathrm{d}\rho\right)^{s}.$$

$$I(\psi) = \log \int \exp(\psi^{c,\lambda}) \mathrm{d}\mu.$$

Lemma (Chizat, D. and Vaškevičius, 2024): *I* is a concave functional.

Elements of proof

Strong-concavity estimate

- From the concavity of I:

Proposition (Chizat, D. and Vaškevičius, 2024):

$$\frac{\mathrm{d}^2}{\mathrm{d}s^2} E(\psi + sv) \leq -C(\lambda)\mathbb{V}\mathrm{ar}_{\nu[\psi+sv]}(v).$$

Elements of proof

Strong-concavity estimate

- From the concavity of I:

Proposition (Chizat, D. and Vaškevičius, 2024):

$$\frac{\mathrm{d}^2}{\mathrm{d}s^2} E(\psi + sv) \leq -C(\lambda)\mathbb{V}\mathrm{ar}_{\nu[\psi+sv]}(v).$$

Remarks:
- Case $\lambda = 0$ and $c(x, y) = -\langle x|y\rangle$ can be deduced from the Brascamp-Lieb inequality.
- Valid for any semi-concave cost c (e.g. $\mathcal{C}^2$ cost).
- In the $\lambda \to 0$ regime, yields a novel estimate of the strong-concavity of the dual Kantorovich problem in OT.

Elements of proof

Conclusion

- The strong-concavity estimate yields

$$\delta_t = -\int_{\varepsilon=0}^{1}\int_{s=\varepsilon}^{1}\frac{\mathrm{d}^2}{\mathrm{d}s^2}E(\psi_t+sv)\mathrm{d}s\mathrm{d}\varepsilon \geq C(\lambda)\mathbb{V}\mathrm{ar}_\nu(\psi^*-\psi_t).$$

- Together with the one-step-improvement bound, this entails

$$\delta_t \leq 2\sqrt{C(\lambda)\delta_t(\delta_t-\delta_{t+1})}+\frac{14c_\infty}{3}\lambda^{-1}(\delta_t-\delta_{t+1}).$$

- Conclusion:

$$\boxed{\delta_{t+1} \leq \kappa(\lambda)\delta_t.}$$

Lower bound

Tightness of the the $1 - \Theta(\frac{\lambda}{c_\infty})$ contraction constant.

Theorem (Chizat, D. and Vaškevičius, 2024):

- On $\mathbb{R}$, let $\mu = \mathcal{N}(0,1)$ and $\nu = \mathcal{N}(0,\sigma^2)$ with $\sigma > 0$.
- Let $c(x,y) = -xy$ and $\psi_0 = 0$.

If $\lambda \leq \sigma/5$, then

$$\delta_t \geq \frac{\sigma}{20}\left(1 - \frac{5\lambda}{\sigma}\right)^t.$$

Main result: general statement

Theorem (Chizat, D. and Vaškevičius, 2024): Assume that $\mathcal{X}$ is convex, $\exists \xi \in \mathbb{R}_+$ s.t. $\forall y \in \mathcal{Y}$, $x \mapsto c(x,y)$ is ξ-semi-concave, and $\|c\|_{\mathrm{osc}} = c_\infty < \infty$. Then, for any integer $t \geq 0$, the Sinkhorn iterates $(\psi_t)_{t\geq 0}$ satisfy

$$\boxed{E(\psi^*) - E(\psi_{t+1}) \leq (1-\alpha^{-1})(E(\psi^*) - E(\psi_t))}$$

provided either one of the following additional assumption holds:

1. The domain $\mathcal{X}$ is compact and included in $\{x : \|x\| \leq R_{\mathcal{X}}\}$, the measure μ admits a density $f_\mu(x)$ such that $\frac{\sup_{x\in\mathcal{X}} f_\mu(x)}{\inf_{x'\in\mathcal{X}} f_\mu(x')} = \kappa < \infty$, and
$$\alpha = 176\{1 + (c_\infty + \frac{\xi}{2}R_{\mathcal{X}}^2)\kappa\lambda^{-1} + c_\infty^2\lambda^{-2}\}.$$
2. There exists a ξ-strongly convex function $V : \mathcal{X} \to \mathbb{R}$ such that the density of μ reads $f_\mu(x) = e^{-V(x)}$, and
$$\alpha = 176\{1 + c_\infty\lambda^{-1} + c_\infty^2\lambda^{-2}\}.$$
3. There exists $\zeta \in \mathbb{R}_+$ such that for all $y \in \mathcal{Y}$, $x \mapsto c(x,y)$ is ζ-semi-convex, there exists a $\max(\xi, (\xi+\zeta)/\lambda)$-strongly convex function $V : \mathcal{X} \to \mathbb{R}$ such that the density of μ reads $f_\mu(x) = e^{-V(x)}$, and
$$\alpha = 176\{1 + c_\infty\lambda^{-1}\}.$$

Thank you for your attention!