Primer on Entropic Optimal Transport

Marcel Nutz
Columbia University

June 2025

Outline

Entropic Optimal Transport

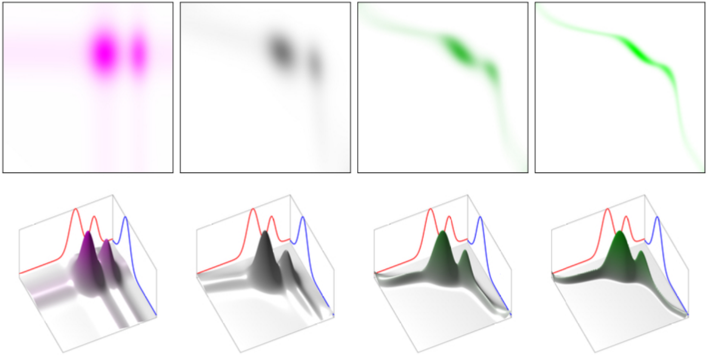
- Regularization parameter $\varepsilon > 0$
- Entropic optimal transport (EOT) problem:

$$\boxed{\mathrm{EOT}_\varepsilon := \inf_{\pi \in \Pi(\mu,\nu)} \int_{X \times Y} c \, d\pi + \varepsilon H(\pi|\mu \otimes \nu)}$$

- $H(\cdot|\mu \otimes \nu)$ is relative entropy (Kullback–Leibler divergence) wrt. $\mu \otimes \nu$
- Unique minimizer π_ε, and $\pi_\varepsilon \sim \mu \otimes \nu$

- EOT is tradeoff between transport cost and entropy
- "Interpolates" between optimal transport and $\mu \otimes \nu$

Entropic Regularization



$\varepsilon = 10$ $\varepsilon = 1$ $\varepsilon = 10^{-1}$ $\varepsilon = 10^{-2}$

(Figure from Peyré–Cuturi 2019)

Properties of EOT

- Generally an easier problem than OT
- We'll assume throughout that $c \in L^1(\mu \otimes \nu)$ and bounded from below
- And that is enough for existence and uniqueness of optimal coupling π_ε

- Computation through Sinkhorn's algorithm (IPFP)
- Solve EOT_ε for small ε to approximate OT

- EOT has many desirable properties related to smoothness

Outline

Definition of Relative Entropy (KL Divergence)

Let $(\Omega, \mathcal{F})$ be a measurable space

Definition

Given $Q, R \in \mathcal{P}(\Omega)$, the *relative entropy* (Kullback–Leibler divergence) is defined as

$$H(Q|R) = \begin{cases} E^Q\left[\log \frac{dQ}{dR}\right], & Q \ll R, \\ \infty, & \text{otherwise.} \end{cases}$$

- Or: $H(Q|R) = E^R\left[h\left(\frac{dQ}{dR}\right)\right]$ for $h(z) = z \log z$
- $h(z) = z \log z$ is strictly convex on $(0, \infty)$
- From Jensen's inequality:

$$Q \mapsto H(Q|R) \text{ is convex and nonnegative}$$

- Strictly convex where finite
- $H(Q|R) = 0$ if and only if $Q = R$

Pinsker's Inequality with Total Variation Distance

Let $P, Q \in \mathcal{P}(\Omega)$ and $R \gg P, Q$. Define

$$\|P - Q\|_{TV} := \int \left| \frac{dP}{dR} - \frac{dQ}{dR} \right| dR$$

Alternative definitions:

$$\|P - Q\|_{TV} = \sup_{|\phi| \leq 1} \int \phi \, d(P - Q) = 2 \sup_{A \subset \Omega} |P(A) - Q(A)|$$

Lemma (Pinsker's Inequality)

For $Q, R \in \mathcal{P}(\Omega)$,

$$\|Q - R\|_{TV} \leq \sqrt{2H(Q|R)}$$

Donsker–Varadhan Variational Formula

Lemma (Donsker–Varadhan)

$$H(Q|R) = \sup_{\phi \text{ bounded, measurable}} \left(E^Q[\phi] - \log E^R[e^\phi] \right)$$

Or:

$$H(Q|R) = \sup_{\phi : E^R[e^\phi] < \infty} \left(E^Q[\phi] - \log E^R[e^\phi] \right)$$

If Ω *Polish:*

$$H(Q|R) = \sup_{\phi \text{ bounded, continuous}} \left(E^Q[\phi] - \log E^R[e^\phi] \right)$$

Implies:

- $(Q, R) \mapsto H(Q|R)$ is jointly convex and lower semicontinuous (in TV, and also weakly in the Polish setting)

Additivity w.r.t. Conditioning

Suppose both Q and R are written as disintegrations w.r.t. a marginal:

$$Q = Q_0(dx) \otimes Q^x(dy), \quad R = R_0(dx) \otimes R^x(dy)$$

If $Q \ll R$, then

$$\frac{dQ}{dR}(x,y) = \frac{dQ_0}{dR_0}(x)\frac{dQ^x}{dR^x}(y)$$

and hence

$$\begin{aligned} H(Q|R) &= \int \log\left(\frac{dQ}{dR}\right)\, dQ \\ &= \int \left[\log\left(\frac{dQ_0}{dR_0}\right) + \log\left(\frac{dQ^x}{dR^x}\right)\right]\, Q^x(dy)Q_0(dx) \\ &= H(Q_0|R_0) + \int H(Q^x|R^x)\, Q_0(dx) \end{aligned}$$

Data Processing Inequality

Lemma (Data Processing Inequality)

Let $Q, R \in \mathcal{P}(\Omega)$ and let $T : \Omega \to \Omega'$ be measurable.

Then the pushforwards $T_\# Q, T_\# R \in \mathcal{P}(\Omega')$ satisfy

$$H(T_\# Q | T_\# R) \leq H(Q|R).$$

More generally, map T can be replaced by a stochastic kernel.

Interpretation: Transformations cannot increase relative entropy

Example: $T =$ Projection to a marginal (cf. preceding slide)

Outline

Existence and Uniqueness of Entropic "Projection"

Theorem

Let $\emptyset \neq \mathcal{Q} \subseteq \mathcal{P}(\Omega)$ be convex and closed w.r.t. TV, and

$$\mathcal{Q}_{fin} := \{Q \in \mathcal{Q} : \ H(Q|R) < \infty\} \neq \emptyset.$$

(a) *There exists a unique $Q_* \in \mathcal{Q}$ such that*

$$H(Q_*|R) = \inf_{Q \in \mathcal{Q}} H(Q|R) \in [0, \infty).$$

Moreover, $Q_* \gg Q$ for any $Q \in \mathcal{Q}_{fin}$.

(b) *$Q_0 \in \mathcal{Q}$ is the minimizer Q_* if and only if $Z_0 := dQ_0/dR$ exists and*

$$E^Q[\log Z_0] \geq H(Q_0|R) \quad \text{for all} \quad Q \in \mathcal{Q}_{fin}$$

- If $\mathcal{Q}$ is weakly compact, existence also follows $H(\cdot|R)$ being LSC
- (b) is the first-order condition of optimality. Often holds with equality

Outline

Static Schrödinger Bridge Problem

Given a reference measure $R \in \mathcal{P}(X \times Y)$, the static Schrödinger bridge problem is

$$\min_{\pi \Pi(\mu,\nu)} H(\pi | R)$$

- EOT is a special case:

$$\pi_\varepsilon = \operatorname*{arg\,min}_{\Pi(\mu,\nu)} H(\,\cdot\,|R) \qquad \text{for} \qquad dR := \frac{e^{-c/\varepsilon}}{E^P[e^{-c/\varepsilon}]} d(\mu \otimes \nu)$$

  Note that $R \sim \mu \otimes \nu$ in this case

→ General result on entropy minimization implies existence and uniqueness of π_ε

- As $\Pi(\mu, \nu)$ is weakly compact (assuming Polish spaces), this also follows directly from LSC and strict convexity
- $\pi_\varepsilon \sim \mu \otimes \nu$ as $\mu \otimes \nu \in \Pi(\mu, \nu)$
- Conversely, we can define c from R if $R \ll \mu \otimes \nu$

Detour: Dynamic Schrödinger Bridge Problem

Given: $\mu, \nu \sim \mathcal{L}^d$ on $\mathbb{R}^d$ and a reference probability measure $\mathbb{P}$ on $\mathcal{C}([0,1];\mathbb{R}^d)$

Problem: Find law $\mathbb{Q}$ on $\mathcal{C}([0,1];\mathbb{R}^d)$

- with marginals $\mathbb{Q}_0 = \mu$ at $t = 0$ and $\mathbb{Q}_1 = \nu$ at $t = 1$

$\rightarrow$ joint marginal $\mathbb{Q}_{01}$ at $t = (0,1)$ is element of $\Pi(\mu, \nu)$

- minimizing $H(\mathbb{Q}|\mathbb{P})$

- For simplicity, we take $\mathbb{P} = \mathbb{W}$ the law of a Brownian motion with some[1] initial distribution $\tilde{\mu} \sim \mathcal{L}^d$
- Let also $\tilde{\nu} \in \mathcal{P}(\mathbb{R})$ be equivalent to Lebesgue, e.g., $\mathcal{N}(0,1)$
- $\mathbb{W}_{01}(dx, dy) \propto e^{-\|x-y\|^2/2}\tilde{\mu}(dx)dy = e^{-\|x-y\|^2/2-g(y)}\tilde{\mu}(dx)\tilde{\nu}(dy)$ where $g = \log(d\tilde{\nu}/dy)$

[1] Instead of a "reversible Brownian motion" which has time-t marginals = Lebesgue measure. See Léonard (2014), *A survey of the Schrödinger problem and some of its connections with optimal transport*

Solution:

- Disintegrate any admissible law $\mathbb{Q}$ as

$$\mathbb{Q} = \mathbb{Q}_{01}(dx, dy) \otimes \mathbb{Q}^{x,y}$$

  where
  - $\mathbb{Q}_{01}(dx, dy) \in \Pi(\mu, \nu)$ joint marginal at $(0, 1)$
  - conditional law $\mathbb{Q}^{x,y}$, i.e., law of a process starting at x and ending at y

- Recall additivity property

$$H(\mathbb{Q}|\mathbb{W}) = H(\mathbb{Q}_{01}|\mathbb{W}_{01}) + \int H(\mathbb{Q}^{x,y}|\mathbb{W}^{x,y})\, d\mathbb{Q}_{01}$$

- Second term minimized by taking $\mathbb{Q}^{x,y} = \mathbb{W}^{x,y} =$ Brownian bridge from x to y
- Minimizing first term = static SB problem with

$$R = \mathbb{W}_{01} \propto e^{-\|x-y\|^2/2}\mathbb{W}_0(dx)\mathbb{W}_1(dy)$$

- Argmin is the same as for $R \propto e^{-\|x-y\|^2/2}\mu(dx)\nu(dy)$
- I.e., the EOT problem with cost $c(x, y) = \|x - y\|^2/2$

Outline

Form of the Optimal Coupling

(a) Optimal density is given by potentials:

- There exist measurable functions $\varphi_\varepsilon : \mathsf{X} \to \mathbb{R}$, $\psi_\varepsilon : \mathsf{Y} \to \mathbb{R}$ such that

$$\frac{d\pi_\varepsilon}{d(\mu \otimes \nu)} = \exp\left(\frac{\varphi_\varepsilon \oplus \psi_\varepsilon - c}{\varepsilon}\right) \quad \mu \otimes \nu\text{-a.s.}$$

- The functions $\varphi_\varepsilon, \psi_\varepsilon$ (called **EOT** or **Schrödinger potentials**) are a.s. unique up to an additive constant: if $\tilde\varphi_\varepsilon, \tilde\psi_\varepsilon$ are potentials, then

$$\tilde\varphi_\varepsilon = \varphi_\varepsilon + a, \quad \tilde\psi_\varepsilon = \psi_\varepsilon - a \text{ for some } a \in \mathbb{R}.$$

- Our assumption $c \in L^1(\mu \otimes \nu)$ implies $\varphi_\varepsilon \in L^1(\mu)$ and $\psi_\varepsilon \in L^1(\nu)$.

(b) Form implies optimality:

- Let $\pi \in \Pi(\mu, \nu)$ have the form

$$\frac{d\pi}{d(\mu \otimes \nu)} = \exp\left(\frac{\varphi \oplus \psi - c}{\varepsilon}\right)$$

  for some measurable $\varphi : \mathsf{X} \to [-\infty, \infty)$, $\psi : \mathsf{Y} \to [-\infty, \infty)$.

  Then $\pi = \pi_\varepsilon$ and φ, ψ are EOT potentials.

EOT Cost and Potentials

$$\begin{aligned}
\mathrm{EOT}_\varepsilon &= \int c\, d\pi_\varepsilon + \varepsilon H(\pi_\varepsilon|\mu \otimes \nu) \\
&= \int c\, d\pi_\varepsilon + \varepsilon \int \log \exp\left(\frac{\varphi_\varepsilon \oplus \psi_\varepsilon - c}{\varepsilon}\right) d\pi_\varepsilon \\
&= \int c\, d\pi_\varepsilon + \int \varphi_\varepsilon \oplus \psi_\varepsilon\, d\pi_\varepsilon - \int c\, d\pi_\varepsilon \\
&= \int \varphi_\varepsilon\, d\mu + \int \psi_\varepsilon\, d\nu
\end{aligned}$$

Dual Problem

Unconstrained dual problem of EOT:

$$\mathrm{DUAL}_\varepsilon = \sup_{\varphi\in L^1(\mu),\,\psi\in L^1(\nu)} \int \varphi\, d\mu + \int \psi\, d\nu - \varepsilon \int e^{\frac{\varphi\oplus\psi - c}{\varepsilon}}\, d(\mu\otimes\nu) + \varepsilon$$

$$= \sup_{\varphi\in L^1(\mu),\,\psi\in L^1(\nu)} \int \left(\varphi \oplus \psi - \varepsilon e^{\frac{\varphi\oplus\psi - c}{\varepsilon}} + \varepsilon\right)\, d(\mu\otimes\nu)$$

Note: $f(z) = \varepsilon(z\log z - z + 1)$ has convex conjugate $f^*(y) = \varepsilon e^{y/\varepsilon} - \varepsilon$

Theorem:

- Dual problem has a unique (mod $\mathbb{R}$) solution $(\varphi_\varepsilon, \psi_\varepsilon) \in L^1(\mu) \times L^1(\nu)$
- Solution yields EOT potentials, and vice versa
- Strong duality holds:

$$\mathrm{EOT}_\varepsilon = \mathrm{DUAL}_\varepsilon = \int \varphi_\varepsilon\, d\mu + \int \psi_\varepsilon\, d\nu$$

Schrödinger System

Measurable functions $\varphi_\varepsilon : \mathsf{X} \to \mathbb{R}$, $\psi_\varepsilon : \mathsf{Y} \to \mathbb{R}$ are potentials if and only if they satisfy the **Schrödinger System**

$$\varphi_\varepsilon(x) = -\varepsilon \log \int e^{\frac{\psi_\varepsilon(y) - c(x,y)}{\varepsilon}} \, \nu(dy) \quad \mu\text{-a.s.},$$
$$\psi_\varepsilon(y) = -\varepsilon \log \int e^{\frac{\varphi_\varepsilon(x) - c(x,y)}{\varepsilon}} \, \mu(dx) \quad \nu\text{-a.s.}$$

or equivalently

$$\int \exp\left(\frac{\varphi_\varepsilon \oplus \psi_\varepsilon - c}{\varepsilon}\right) d\nu = 1 \quad \mu\text{-a.s.}, \tag{1}$$
$$\int \exp\left(\frac{\varphi_\varepsilon \oplus \psi_\varepsilon - c}{\varepsilon}\right) d\mu = 1 \quad \nu\text{-a.s.} \tag{2}$$

Btw: shows how to get a version of the potentials defined at every (x, y)

Schrödinger System: Interpretation/Proof

$$\int \exp\left(\frac{\varphi_\varepsilon \oplus \psi_\varepsilon - c}{\varepsilon}\right) d\nu = 1 \quad \mu\text{-a.s.}, \tag{1}$$

$$\int \exp\left(\frac{\varphi_\varepsilon \oplus \psi_\varepsilon - c}{\varepsilon}\right) d\mu = 1 \quad \nu\text{-a.s.} \tag{2}$$

Interpretation as Marginal Density Condition:

- Read $1 = \frac{d\mu}{d\mu}$ in (1). Then (1) $\Leftrightarrow \exp\left(\frac{\varphi_\varepsilon \oplus \psi_\varepsilon - c}{\varepsilon}\right)$ is density (w.r.t. $\mu \otimes \nu$) of a measure with 1st marginal $= \mu$
- Thus Schrödinger System $\Leftrightarrow \exp\left(\frac{\varphi_\varepsilon \oplus \psi_\varepsilon - c}{\varepsilon}\right)$ is density of a coupling

Interpretation as Dual FOC:

- Since potentials are dual maximizers of the (unconstrained, concave) dual problem, they satisfy the first-order condition with equality
- Conversely, FOC implies optimality in the dual problem

c-Conjugate for EOT

- Let $\psi : \mathsf{Y} \to \mathbb{R}$ be measurable (and the integral below finite)
- In analogy to OT, consider the (c, ε)-conjugate

$$\psi^{c,\varepsilon}(x) := -\varepsilon \log \int e^{\frac{\psi(y) - c(x,y)}{\varepsilon}} \, \nu(dy)$$

- Symmetrically, $\varphi(x)$ determines a function $\varphi^{c,\varepsilon}(y)$

Dual Interpretation:

- Write $\mathcal{D}$ for the dual objective
- Then $\psi^{c,\varepsilon} = \arg\max_{\varphi} \mathcal{D}(\varphi, \psi)$
- In particular, $\mathcal{D}(\psi^{c,\varepsilon}, \psi) \geq \mathcal{D}(\varphi, \psi)$
- Iterate this? Later: Sinkhorn–Knopp (IPFP) algorithm

Primal Interpretation:

- Fit one marginal: $\psi^{c,\varepsilon}$ is such that $\exp\left(\frac{\psi^{c,\varepsilon} \oplus \psi - c}{\varepsilon}\right)$ is density (w.r.t. $\mu \otimes \nu$) of a measure with 1st marginal $= \mu$

Regularity of (c, ε)-Conjugate

Recall:

$$\psi^{c,\varepsilon}(x) := -\varepsilon \log \int e^{\frac{\psi(y)-c(x,y)}{\varepsilon}} \, \nu(dy)$$

Example 1: If c is L-Lipschitz on $\mathbb{R}^d \times \mathbb{R}^d$, then $\psi^{c,\varepsilon}$ is also L-Lipschitz:

$$\begin{aligned}
\psi^{c,\varepsilon}(x) - \psi^{c,\varepsilon}(\tilde{x}) &= -\varepsilon \log \int e^{\frac{\psi(y)-c(x,y)}{\varepsilon}} \, \nu(dy) + \varepsilon \log \int e^{\frac{\psi(y)-c(\tilde{x},y)}{\varepsilon}} \, \nu(dy) \\
&\leq -\varepsilon \log \int e^{\frac{\psi(y)-c(x,y)}{\varepsilon}} \, \nu(dy) + \varepsilon \log \int e^{\frac{\psi(y)-c(x,y)+L\|x-\tilde{x}\|}{\varepsilon}} \, \nu(d \\
&\leq L\|x - \tilde{x}\|
\end{aligned}$$

More generally: $c(\cdot, y)$ $\omega(\cdot)$-continuous $\forall y$ implies $\psi^{c,\varepsilon}$ ω-continuous

Example 2: If c is differentiable (and sufficiently integrable), get derivatives of $\psi^{c,\varepsilon}(x)$ by differentiating under the integral

Regularity of Potentials

Apply this to the potentials:

- If $\psi = \psi_\varepsilon$, then $\psi^{c,\varepsilon} = \varphi_\varepsilon$ a.s. by Schrödinger system
- Similarly for ψ_ε
- Hence, $(\varphi_\varepsilon, \psi_\varepsilon)$ are conjugates of one another (after choosing versions)

- In particular, we can get regularity estimates for the potentials from the properties of c
- E.g., if $c(x,y) = \|x-y\|^2$ and μ, ν are supported in unit ball, potentials are smooth with C^k norms independent of μ, ν (for any k)
- Many applications (later)

Proof of Existence of Potentials by Solving Dual Problem

- Suppose μ, ν compactly supported, c continuous (hence bounded, UC)
- Consider $(\varphi_n, \psi_n)_{n \geq 1}$ maximizing sequence of dual problem:

$$\mathcal{D}(\varphi_n, \psi_n) \uparrow \mathrm{DUAL}_\varepsilon$$

- Normalize $\int \varphi_n \, d\mu = \int \psi_n \, d\nu$ (or similar)
- "Improved" sequence

$$(\tilde{\varphi}_n, \tilde{\psi}_n) := (\psi_n^{c,\varepsilon}, (\psi_n^{c,\varepsilon})^{c,\varepsilon})$$

  is again maximizing
- Equicontinuous and uniformly bounded (latter uses normalization)
- Arzelà–Ascoli yields a uniformly convergent subsequence
- Limit (φ_*, ψ_*) is an optimizer in the dual problem
- FOC shows (φ_*, ψ_*) are potentials

Outline

Convergence of EOT to OT

As the regularization parameter $\varepsilon \to 0$, we may hope that (under regularity conditions, including c continuous):

- Optimal costs converge: $\mathrm{EOT}_\varepsilon \to \mathrm{OT}$
- Optimal couplings converge: $\pi_\varepsilon \to \pi_0$, where π_0 is an OT
- Dual potentials converge: $\varphi_\varepsilon \to \varphi_0$ and $\psi_\varepsilon \to \psi_0$

Convergence Rate of the EOT Cost

- For c Lipschitz:

$$\mathrm{EOT}_\varepsilon - \mathrm{OT} \leq d\varepsilon \log\left(\frac{1}{\varepsilon}\right) + K\varepsilon$$

  Leading term sharp for $c(x,y) = \|x-y\|$ and particular marginals
- For c with bounded second derivative, factor d improved to $d/2$:

$$\mathrm{EOT}_\varepsilon - \mathrm{OT} \leq \frac{d}{2}\varepsilon \log\left(\frac{1}{\varepsilon}\right) + K\varepsilon$$

  Leading term is exact for certain c including $c(x,y) = \|x-y\|^2$ when marginals are regular

Sketch of Proof of Lipschitz Rate (using "Shadows")[2]

- Let $\mu, \nu \in \mathcal{P}(\mathbb{R}^d)$ have finite first moment; c L-Lipschitz
- Idea: start with OT optimizer π_0, build π^ε that is close to optimal for OT but has controlled relative entropy
- Given $\pi \in \Pi(\mu, \nu)$ and $(\tilde{\mu}, \tilde{\nu})$, define the **shadow** $\tilde{\pi} \in \Pi(\tilde{\mu}, \tilde{\nu})$ by

$$\begin{array}{ccc} \mu & \xrightarrow{\pi} & \nu \\ \mathcal{W}_1\text{-opt}\uparrow & & \downarrow\mathcal{W}_1\text{-opt} \\ \tilde{\mu} & \xrightarrow{\tilde{\pi}} & \tilde{\nu} \end{array}$$

- Construction and Data Processing Inequality imply

$$\mathcal{W}_1(\pi, \hat{\pi}) = \mathcal{W}_1(\mu, \tilde{\mu}) + \mathcal{W}_1(\nu, \tilde{\nu}),$$
$$H(\tilde{\pi}|\tilde{\mu} \otimes \tilde{\nu}) \leq H(\pi|\mu \otimes \nu)$$

[2] Details: Eckstein, Nutz (2024), *Convergence Rates for Regularized Optimal Transport via Quantization*. A version of the "block approximation;" see Carlier, Pegon, Tamanini (2023), *Convergence rate of general entropic optimal transport costs*

Sketch of Proof of Lipschitz Rate

- **Fact 1:** If ν_n is supported on n points and $\pi \in \Pi(\mu \otimes \nu_n)$, then $H(\pi|\mu \otimes \nu_n) \leq \log n$
- **Fact 2:** $\exists\, \nu_n$ supported on n points with $W_1(\nu, \nu_n) \leq Cn^{-1/d}$
- Start with optimal transport $\pi_0 \in \Pi(\mu, \nu)$. Consider ν_n as in Fact 2.
- Take $\hat{\pi} :=$ shadow of π in $\Pi(\mu, \nu_n)$
- Take $\pi_n :=$ shadow of $\hat{\pi}$ in $\Pi(\mu, \nu)$
- Then

$$H(\pi_n|\mu \otimes \nu) \leq H(\hat{\pi}|\mu \otimes \nu_n) \leq \log n,$$

$$\int c\, d\pi_n \leq \int c\, d\pi_0 + L\mathcal{W}_1(\nu, \nu_n) \leq \mathrm{OT} + LCn^{-1/d}$$

- Using π_n for EOT $\Rightarrow$ $\mathrm{EOT}_\varepsilon \leq \mathrm{OT} + LCn^{-1/d} + \varepsilon \log n$
- Choosing $n := (1/\varepsilon)^d$ yields $\mathrm{EOT}_\varepsilon \leq \mathrm{OT} + LC\varepsilon + d\varepsilon \log(1/\varepsilon)$