Estimating Wasserstein distances, I

Jonathan Niles-Weed

Center for Data Science and Courant Institute, NYU

Statistical Optimal Transport

OT gives us **distances**, **couplings**, **maps** between distributions

Basic statistical question: can I **estimate** these objects from data?

E.g., distributionally robust optimization: [Kuhn et al. '19; Blanchet et al. '21]

$$\inf_{\theta \in \Theta} \mathbb{E}_{\mu} \ell(X, \theta) \longrightarrow \inf_{\theta \in \Theta} \sup_{\nu \in \mathcal{U}_\delta(\mu)} \mathbb{E}_{\nu} \ell(X, \theta)$$

$\mathcal{U}_\delta(\mu) = \{\nu : W_p(\mu, \nu) \leq \delta\}$ is "ambiguity set."

But all we have is $X_1, \ldots, X_n \sim \mu$! Can we estimate ambiguity set?

Law of large numbers

Classic LLN: $\frac{1}{n}\sum_{i=1}^{n}\ell(X_i,\theta) \to \mathbb{E}_\mu \ell(X,\theta)$ at $n^{-1/2}$ rate.
Implication: Empirical average is a good proxy for risk

"Wasserstein" LLN: Let $\mu_n := \frac{1}{n}\sum_{i=1}^{n}\delta_{X_i}$. Does $W_p(\mu_n,\nu) \to W_p(\mu,\nu)$?
How fast?

Law of large numbers

Classic LLN: $\frac{1}{n}\sum_{i=1}^{n} \ell(X_i, \theta) \to \mathbb{E}_\mu \ell(X, \theta)$ at $n^{-1/2}$ rate.
Implication: Empirical average is a good proxy for risk

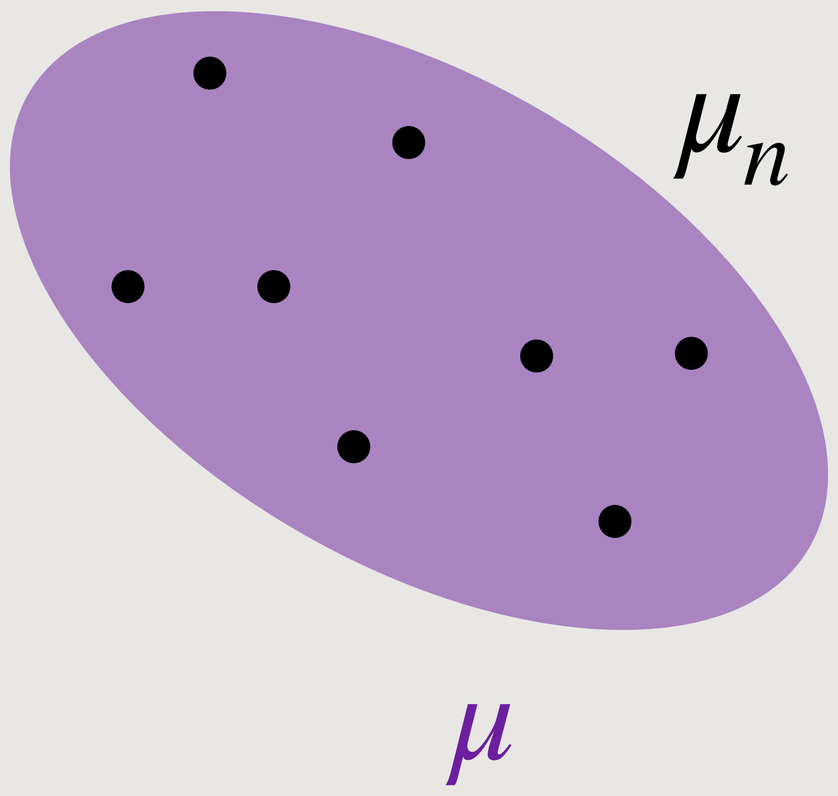
"Wasserstein" LLN: Let $\mu_n := \frac{1}{n}\sum_{i=1}^{n} \delta_{X_i}$. Does $W_p(\mu_n, \nu) \to W_p(\mu, \nu)$?
How fast?

μ

Law of large numbers

Classic LLN: $\frac{1}{n}\sum_{i=1}^{n} \ell(X_i, \theta) \to \mathbb{E}_\mu \ell(X, \theta)$ at $n^{-1/2}$ rate.
Implication: Empirical average is a good proxy for risk

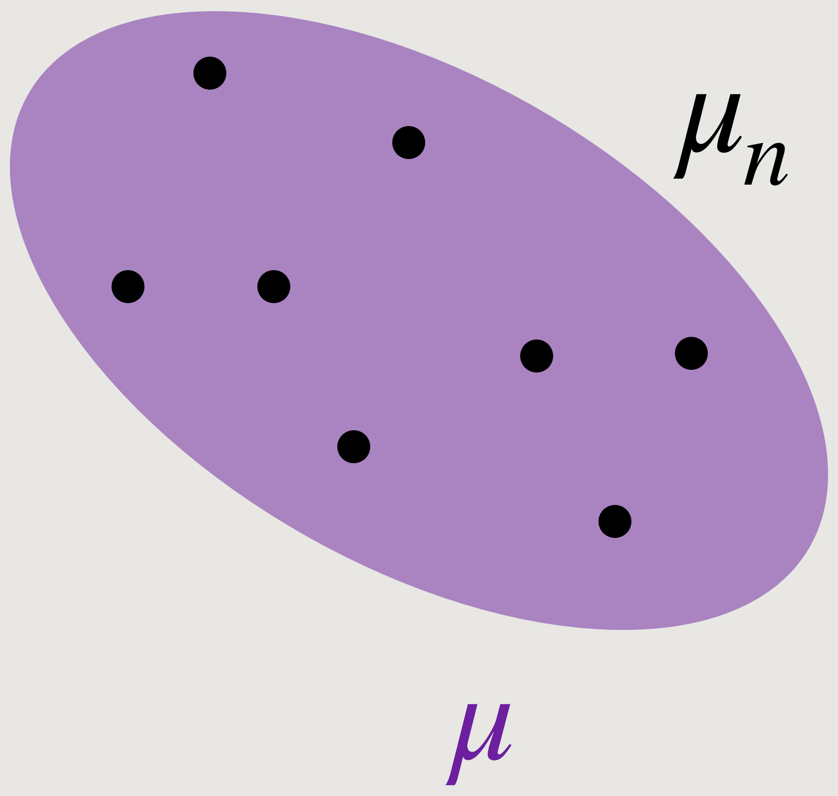
"Wasserstein" LLN: Let $\mu_n := \frac{1}{n}\sum_{i=1}^{n} \delta_{X_i}$. Does $W_p(\mu_n, \nu) \to W_p(\mu, \nu)$?
How fast?



Law of large numbers

Classic LLN: $\frac{1}{n}\sum_{i=1}^{n} \ell(X_i, \theta) \to \mathbb{E}_\mu \ell(X, \theta)$ at $n^{-1/2}$ rate.
Implication: Empirical average is a good proxy for risk

"Wasserstein" LLN: Let $\mu_n := \frac{1}{n}\sum_{i=1}^{n} \delta_{X_i}$. Does $W_p(\mu_n, \nu) \to W_p(\mu, \nu)$?
How fast?



Analogous question arises whenever we want to use Wasserstien distances for statistics or ML task

Law of large numbers

Lecture I: Wasserstein law of large numbers

Goal: Explore Wasserstein LLN with goal of learning two main techniques—**transport method** and **empirical process method.**

Lecture II: Refined arguments for measures with special structure

Lecture III: Benefits and drawbacks of alternative distances

Note: ideally want much more than just LLN (distributional limits, optimal tail bounds, etc.), but many aspects still **open**.

Law of large numbers

Lecture I: Wasserstein law of large numbers
Goal: Explore Wasserstein LLN with goal of learning two main techniques—**transport method** and **empirical process method.**

Lecture II: Refined arguments for measures with special structure

Lecture III: Benefits and drawbacks of alternative distances

Note: ideally want much more than just LLN (distributional limits, optimal tail bounds, etc.), but many aspects still **open**.

Wasserstein LLN

Question: if μ_n consists of n i.i.d. samples from $\mu \in \mathscr{P}(\mathbb{R}^d)$, how does $W_p(\mu_n, \nu) - W_p(\mu, \nu)$ behave?

Observation I: random, need to understand bias and fluctuations

Observation II: since W_p satisfies triangle inequality,

$$\sup_{\nu} |W_p(\mu_n, \nu) - W_p(\mu, \nu)| = W_p(\mu_n, \mu)$$

Control of $W_p(\mu_n, \mu) \iff$ **uniform control**

A simple lower bound

How large is $\mathbb{E} W_p(\mu_n, \mu)$?

Claim: No smaller than $n^{-1/2p}$ in general.

Proof: Suppose μ is uniform measure on $\{x, y\}$. Then μ_n is supported on same two points, and any coupling between μ and μ_n must move mass $|\mu(\{x\}) - \mu_n(\{x\})|$ a distance of $|x - y|$.

Therefore $\mathbb{E} W_p(\mu_n, \mu) \geq |x - y| \mathbb{E} |\mu(\{x\}) - \mu_n(\{x\})|^{1/p} \gtrsim n^{-1/2p}$.

A more complicated lower bound

How large is $\mathbb{E}W_p(\mu_n, \mu)$?

Claim: If μ is absolutely continuous, no smaller than $n^{-1/d}$. [Dudley '69]

Proof: If $\mu \ll \lambda$, then there exists $c_\mu > 0$ such that any set with $\mu(S) \geq 1/2$ satisfies $\lambda(S) \geq c_\mu$. A ball of radius ε around any point has λ mass at most $C_d \varepsilon^d$. If $C_d n \varepsilon^d < c_\mu$, then any coupling must move μ mass $1/2$ a distance ε.

Therefore $W_p^p(\mu_n, \mu) \geq \varepsilon^p/2$ if $\varepsilon = (c_\mu^{-1} C_d n)^{-1/d}$. *(Almost sure bound!)*

Basic rates of convergence

Theorem [Boissard & Le Gouic '14, Fournier & Guillin '15]: If μ has finite moments of all orders, then

$$\mathbb{E} W_p(\mu_n, \mu) \leq (\mathbb{E} W_p^p(\mu_n, \mu))^{1/p} \lesssim_{p,d} \begin{cases} n^{-1/d} & d > 2p \\ n^{-1/2p}(\log n)^{1/p} & d = 2p \\ n^{-1/2p} & d < 2p \end{cases}$$

Moment restrictions can be relaxed, constants improved [Lei '20]

By our earlier arguments, these rates are essentially best possible.

Idea of proof

"Dyadic partitioning argument": suppose $\mathrm{supp}(\mu) \subseteq [0,1]^d$.

Observation: $W_p^p(\mu, \mu_n) \leq d^{p/2}$.

Idea of proof

“Dyadic partitioning argument”: suppose $\mathrm{supp}(\mu) \subseteq [0,1]^d$.

Observation: $W_p^p(\mu, \mu_n) \leq d^{p/2}$.

And: if sub-cubes were all balanced, could improve by factor of 2^p.

Idea of proof

"Dyadic partitioning argument": suppose $\operatorname{supp}(\mu) \subseteq [0,1]^d$.

Observation: $W_p^p(\mu, \mu_n) \leq d^{p/2} \cdot \left(\frac{1}{2} \sum_{i=1}^{2^d} |\mu_n(Q) - \mu(Q)| + 2^{-p} \right)$.

And: if sub-cubes were all balanced, could improve by factor of 2^p.

Idea of proof

"Dyadic partitioning argument": suppose $\operatorname{supp}(\mu) \subseteq [0,1]^d$.

Observation: $W_p^p(\mu, \mu_n) \leq d^{p/2} \cdot \left(\frac{1}{2} \sum_{i=1}^{2^d} |\mu_n(Q) - \mu(Q)| + 2^{-p} \right)$.

And: if sub-cubes were all balanced, could improve by factor of 2^p.

Idea: recurse!

Idea of proof

Let $\{\mathcal{Q}_j\}_{j\geq 0}$ be *dyadic partition* of $[0,1]^d$, i.e., $\mathcal{Q}_j$ contains all 2^{dj} sub-cubes of $[0,1]^d$ with side length 2^{-j}, corners at $2^{-j}\cdot v$, $v \in \mathbb{Z}^d$.

Theorem [Weed & Bach '19]: For *any* $\mu, \nu \in \mathcal{P}([0,1]^d)$,

$$W_p^p(\mu,\nu) \leq d^{p/2}\left(\sum_{j=0}^{J-1} 2^{-jp} \sum_{Q\in\mathcal{Q}_{j+1}} |\mu(Q)-\nu(Q)| + 2^{-J}\right)$$

Idea of proof

Theorem: For *any* $\mu, \nu \in \mathscr{P}([0,1]^d)$,

$$W_p^p(\mu,\nu) \leq d^{p/2}\left(\sum_{j=0}^{J-1} 2^{-jp} \sum_{Q \in \mathcal{Q}_{j+1}} |\mu(Q) - \nu(Q)| + 2^{-J}\right)$$

Idea of proof

We obtain

$$\mathbb{E} W_p^p(\mu, \mu_n) \lesssim_d \sum_{j=0}^{J-1} 2^{-jp} \sqrt{2^{dj}/n} + 2^{-J}$$

If $d < 2p$, first term dominates ("large scale" behavior dominates)

If $d > 2p$, last term dominates ("small scale" behavior dominates)

If $d = 2p$, all terms are of the same order (all scales contribute)

Basic rates of convergence

Theorem [Boissard & Le Gouic '14, Fournier & Guillin '15]: If μ has finite moments of all orders, then

$$\mathbb{E}W_p(\mu_n, \mu) \leq (\mathbb{E}W_p^p(\mu_n, \mu))^{1/p} \lesssim_{p,d} \begin{cases} n^{-1/d} & d > 2p \\ n^{-1/2p}(\log n)^{1/p} & d = 2p \\ n^{-1/2p} & d < 2p \end{cases}$$

Basic rates of convergence

Theorem [Boissard & Le Gouic '14, Fournier & Guillin '15]: If μ has finite moments of all orders, then

$$\mathbb{E}W_p(\mu_n, \mu) \le (\mathbb{E}W_p^p(\mu_n, \mu))^{1/p} \lesssim_{p,d} \boxed{\begin{cases} n^{-1/d} & d > 2p \\ n^{-1/2p}(\log n)^{1/p} & d = 2p \\ n^{-1/2p} & d < 2p \end{cases}}$$

$$=: r_{d,p}(n)$$

Any improvements?

“Worst case” examples suggest:

- $n^{-1/2p}$ rate may improve if support is connected
  (some results support this, but general story not clear) [Ledoux ’19, Divol ’21]

- $n^{-1/d}$ rate may improves if measure is singular
  (stay tuned)

- If μ has a density, then empirical measure is always “cursed”

Intrinsic dimension

Definition: Let $N_\epsilon(S)$ be the smallest number of closed ϵ-balls needed to cover S. The **(upper) Minkowski dimension** $d_M(S)$ is

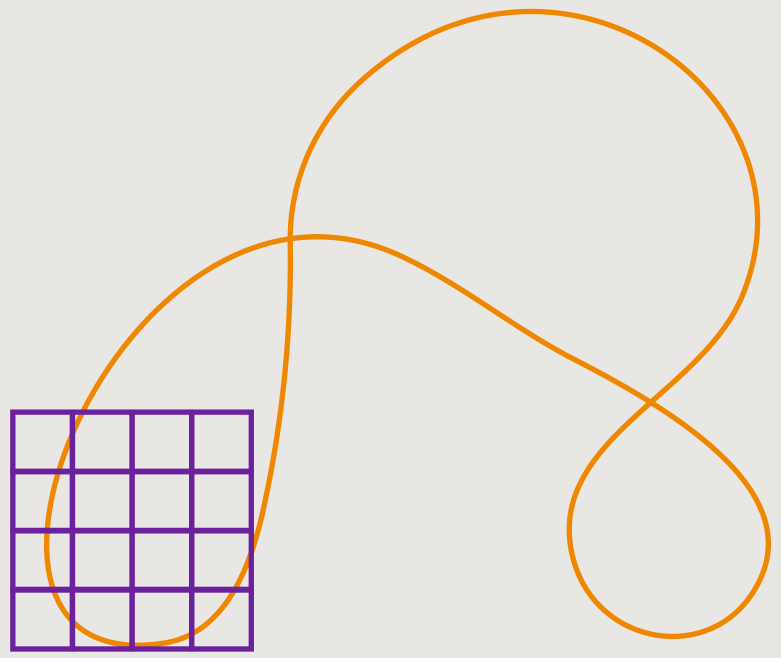
$$\limsup_{\epsilon \to 0} \frac{\log N_\epsilon(S)}{\log \epsilon^{-1}}$$

Theorem [Weed & Bach, '19]: If $d_M(\mathrm{supp}(\mu)) < k$, then

$$\mathbb{E} W_p(\mu, \mu_n) \lesssim r_{k,p}(n)$$

Idea: only $\approx 2^{jk}$ elements of $\mathcal{Q}_j$ are non-empty

$$\sum_{Q \in \mathcal{Q}_{j+1}} \mathbb{E}\,|\,\mu(Q) - \mu_n(Q)\,| = \sum_{Q \in \mathcal{Q}_{j+1},\ \mu(Q)>0} \mathbb{E}\,|\,\mu(Q) - \mu_n(Q)\,| \lesssim \sqrt{2^{jk}/n}$$



Alternate proof methods

Dyadic partitioning argument is a "**transportation method**": prove $W_p(\mu_n, \mu)$ is small by constructing a candidate coupling.

Downside: challenging to prove non-uniform bounds

Theorem [Hundrieser et al. '22a] If $\mu, \nu \in \mathscr{P}([0,1]^d)$ and $d_M(\mathrm{supp}(\nu)) < k$ then,

$$\mathbb{E}\,|\,W_1(\mu_n, \nu) - W_1(\mu, \nu)\,| \lesssim r_{k,1}(n)$$

Proof idea: **empirical process method**, based on duality

Empirical process method

Two prerequisites:

- **Duality:** $W_1(\mu, \nu) = \sup_{f \in \mathrm{Lip}} \int f(d\mu - d\nu)$, where $\mathrm{Lip} = \mathrm{Lip}(\mathbb{R}^d)$
- **Chaining:** for any set $\mathcal{F} \subseteq L^\infty$ and $\delta > 0$,

$$\mathbb{E} \sup_{f \in \mathcal{F}} \left| \int f(d\mu_n - d\mu) \right| \lesssim \delta + n^{-1/2} \int_\delta^1 \sqrt{\log N_\epsilon(\mathcal{F})} d\epsilon$$

If $\mathcal{F} = \mathrm{Lip}(S)$ with $d_M(S) < k$, then $\log N_\epsilon(\mathcal{F}) \lesssim \epsilon^{-k}$.

Empirical process method

$$W_1(\mu_n, \nu) - W_1(\mu, \nu) = \sup_{f_n \in \mathrm{Lip}} \int f_n(d\mu_n - d\nu) - \sup_{f \in \mathrm{Lip}} \int f(d\mu - d\nu)$$

$$\leq \sup_{f \in \mathrm{Lip}} \left(\int f(d\mu_n - d\nu) - \int f(d\mu - d\nu) \right)$$

$$= \sup_{f \in \mathrm{Lip}} \int f(d\mu_n - d\mu)$$

Taking sup over $\mathrm{Lip}([0,1]^d)$ yields $\inf_{\delta>0} \left(\delta + n^{-1/2} \int_\delta^1 \epsilon^{-d/2} d\epsilon \right) = r_{d,1}(n)$

Empirical process method

$$W_1(\mu_n, \nu) - W_1(\mu, \nu) = \sup_{f_n \in \mathrm{Lip}} \int f_n(d\mu_n - d\nu) - \sup_{f \in \mathrm{Lip}} \int f(d\mu - d\nu)$$

$$\leq \sup_{f \in \mathrm{Lip}} \left(\int f(d\mu_n - d\nu) - \int f(d\mu - d\nu) \right)$$

$$= \sup_{f \in \mathrm{Lip}} \int f(d\mu_n - d\mu) = \boxed{W_1(\mu_n, \mu)}$$

Taking sup over $\mathrm{Lip}([0,1]^d)$ yields $\inf_{\delta>0} \left(\delta + n^{-1/2} \int_\delta^1 \epsilon^{-d/2} d\epsilon \right) = r_{d,1}(n)$

Empirical process method

$$W_1(\mu_n, \nu) - W_1(\mu, \nu) = \sup_{f_n \in \mathrm{Lip}} \int f_n (d\mu_n - d\nu) - \sup_{f \in \mathrm{Lip}} \int f (d\mu - d\nu)$$

$$\leq \sup_{f \in \mathrm{Lip}} \left(\int f (d\mu_n - d\nu) - \int f (d\mu - d\nu) \right)$$

$$= \sup_{f \in \mathrm{Lip}} \int f (d\mu_n - d\mu)$$

Empirical process method

$$
\begin{aligned}
W_1(\mu_n,\nu) - W_1(\mu,\nu) &= \sup_{f_n \in \mathrm{Lip}} \int f_n (d\mu_n - d\nu) - \sup_{f \in \mathrm{Lip}} \int f (d\mu - d\nu) \\
&\le \sup_{f \in \mathrm{Lip}} \left(\int f (d\mu_n - d\nu) - \int f (d\mu - d\nu) \right) \\
&= \sup_{f \in \mathrm{Lip}} \int f (d\mu_n - d\mu)
\end{aligned}
$$

Idea: if ν is low-dimensional, we can guarantee sup lies in $\mathscr{F} \subsetneq \mathrm{Lip}$

Empirical process method

$$W_1(\mu_n,\nu) - W_1(\mu,\nu) = \sup_{f_n\in\mathscr{F}} \int f_n(d\mu_n - d\nu) - \sup_{f\in\mathscr{F}} \int f(d\mu - d\nu)$$

$$\le \sup_{f\in\mathscr{F}} \left(\int f(d\mu_n - d\nu) - \int f(d\mu - d\nu) \right)$$

$$= \sup_{f\in\mathscr{F}} \int f(d\mu_n - d\mu) \overset{?}{\ll} W_1(\mu_n,\mu)$$

What is $\mathscr{F}$?

Lipschitz extension

Let $S = \mathrm{supp}(\nu)$. Consider specifying $f \in \mathrm{Lip}$ by first choosing $f|_S$.

$$W_1(\rho, \nu) = \sup_{f \in \mathrm{Lip}} \int f(d\rho - d\nu) = \sup_{g \in \mathrm{Lip}(S)} \sup_{f \in \mathrm{Lip}, f|_S = g} \int f(d\rho - d\nu)$$

$$= \sup_{g \in \mathrm{Lip}(S)} \sup_{f \in \mathrm{Lip}, f|_S = g} \int f d\rho - \int g d\nu$$

Lipschitz extension

Let $S = \mathrm{supp}(\nu)$. Consider specifying $f \in \mathrm{Lip}$ by first choosing $f|_S$.

$$W_1(\rho, \nu) = \sup_{f \in \mathrm{Lip}} \int f(d\rho - d\nu) = \sup_{g \in \mathrm{Lip}(S)} \sup_{f \in \mathrm{Lip}, f|_S = g} \int f(d\rho - d\nu)$$

$$= \sup_{g \in \mathrm{Lip}(S)} \sup_{f \in \mathrm{Lip}, f|_S = g} \int f d\rho - \int g d\nu$$

Given $g \in \mathrm{Lip}(S)$, let $\bar{g} \in \mathrm{Lip}$ be largest Lipschitz function with $\bar{g}|_S = g$.

Lipschitz extension

Let $S = \mathrm{supp}(\nu)$. Consider specifying $f \in \mathrm{Lip}$ by first choosing $f|_S$.

$$W_1(\rho, \nu) = \sup_{f \in \mathrm{Lip}} \int f(d\rho - d\nu) = \sup_{g \in \mathrm{Lip}(S)} \sup_{f \in \mathrm{Lip}, f|_S = g} \int f(d\rho - d\nu)$$

$$= \sup_{g \in \mathrm{Lip}(S)} \sup_{f \in \mathrm{Lip}, f|_S = g} \int f d\rho - \int g d\nu$$

Given $g \in \mathrm{Lip}(S)$, let $\bar{g} \in \mathrm{Lip}$ be largest Lipschitz function with $\bar{g}|_S = g$.

$$= \sup_{g \in \mathrm{Lip}(S)} \int \bar{g} d\rho - \int g d\nu$$

Lipschitz extension

$$W_1(\mu_n, \nu) - W_1(\mu, \nu) = \sup_{g \in \mathrm{Lip}(S)} \int \bar{g} d\mu_n - \int g d\nu - \sup_{g \in \mathrm{Lip}(S)} \int \bar{g} d\mu - \int g d\nu$$

$$\leq \sup_{g \in \mathrm{Lip}(S)} \left(\int \bar{g} d\mu_n - \int g d\nu - \int \bar{g} d\mu + \int g d\nu) \right)$$

$$= \sup_{g \in \mathrm{Lip}(S)} \int \bar{g}(d\mu_n - d\mu)$$

Fact: $\mathscr{F} := \{f : f = \bar{g}, g \in \mathrm{Lip}(S)\}$ is much smaller than Lip. If $d_M(S) < k < d$,

$$\log N_\epsilon(\mathscr{F}) \lesssim \epsilon^{-k} \ll \epsilon^{-d} = \log N_\epsilon(\mathrm{Lip})$$

Empirical process method

Benefits: easy to exploit $\mu \neq \nu$, partial extension to W_p for $p > 1$ via "c-concavity", easiest path to tail bounds on W_p^p [Weed & Bach '19], doorway to distributional limits [Hundrieser et al. '22b]

Challenges: primal interpretation less clear, fails to capture correct rate very near null, difficult to extend to unbounded setting [Manole & Niles-Weed '21]

Unified interpretation: [Sommerfeld et al. '19] transportation method *is* an empirical process method with tree-metric approximation of $\|\cdot\|^p$.

Optimality?

Is rate $r_{d,p}(n)$ optimal?

Yes, in two senses:

1. Examples where $\mathbb{E}W_p(\mu_n, \mu) \gtrsim n^{-1/2p}$ and $\mathbb{E}W_p(\mu_n, \mu) \gtrsim n^{-1/d}$

2. **Minimax lower bound** [Singh & Póczos '18]

$$\inf_{\hat{\mu}} \sup_{\mu \in \mathscr{P}([0,1]^d)} \mathbb{E}_\mu W_p(\hat{\mu}(X_1, \ldots, X_n), \mu) \gtrsim \max\{n^{-1/d}, n^{-1/2p}\}$$

Optimality?

Minimax estimator implies that empirical measure μ_n is essentially the best possible estimate of μ in W_p **without further assumptions**

But: improvements possible when we assume μ has smooth density

Smooth densities

Idea: if μ is smooth, should use a smoother estimator $\hat{\mu}$

Example: [Liang '17, Singh et al. '18] If the density of μ lies in *Sobolev space* H^s, then an *orthogonal series estimator* $\hat{\mu}$ achieves

$$\mathbb{E}W_1(\mu,\hat{\mu}) \lesssim \begin{cases} n^{-\frac{1+s}{d+2s}} & d \geq 3 \\ n^{-1/2}\sqrt{\log n} & d = 2 \\ n^{-1/2} & d = 1 \end{cases}$$

Proof: Refined duality method

Orthogonal series estimators

Fix an orthogonal basis of L^2. E.g., for densities on $[0,1]^d$ (actually, torus), **Fourier basis:**

$$\phi_z(x) = e^{2\pi i \langle z, x\rangle} \qquad\qquad z \in \mathbb{Z}^d$$

Write $\mu(x) = \sum_{z \in \mathbb{Z}^d} \beta_z \phi_z(x)$. Then $\mu \in H^s \iff \sum_{z \in \mathbb{Z}^d} |z|^{2s} |\beta_z|^2 < \infty$.

Important fact: $\hat{\beta}_z = \frac{1}{n} \sum_{i=1}^n \phi_z(X_i)$ is unbiased estimator of β_z.

Orthogonal series estimators

Fix a threshold M. The orthogonal series estimator is

$$\hat{\mu} := \sum_{z \in \mathbb{Z}^d, |z| \leq M} \hat{\beta}_z \phi_z$$

Then

$$\mu(x) - \hat{\mu}(x) = \sum_{|z| \leq M} (\beta_z - \hat{\beta}_z)\phi_z(x) + \sum_{|z| > M} \beta_z \phi_z(x)$$

estimation error

approximation error

$O(M^d)$ terms of size $O(n^{-1/2})$

L^2 norm at most M^{-s}

W_1 bounds

If $f \in \mathrm{Lip}([0,1]^d)$, then $f(x) = \sum_z \alpha_z(f)\phi_z(x)$ with $\left(\sum_z |z|^2 |\alpha_z(f)|^2\right)^{1/2} \lesssim 1.$

$$
\begin{aligned}
W_1(\mu, \hat{\mu}) &= \sup_{f \in \mathrm{Lip}} \int f(x)(\mu(x) - \hat{\mu}(x))dx \\
&= \sup_{f \in \mathrm{Lip}} \int \left(\sum_z \alpha_z(f)\phi_z(x)\right)\left(\sum_{|z| \le M} (\beta_z - \hat{\beta}_z)\phi_z(x) + \sum_{|z|>M} \beta_z \phi_z(x)\right) dx \\
&= \sup_{f \in \mathrm{Lip}} \sum_{|z| \le M} \alpha_z(f)(\beta_z - \hat{\beta}_z) + \sum_{|z|>M} \alpha_z(f)\beta_z \\
&\lesssim \left(\sum_{|z| \le M} |z|^{-2} |\beta_z - \hat{\beta}_z|^2\right)^{1/2} + \left(\sum_{|z|>M} |z|^{-2} |\beta_z|^2\right)^{1/2}
\end{aligned}
$$

W_1 bounds

$$W_1(\mu, \hat{\mu}) \lesssim \left(\sum_{|z| \le M} |z|^{-2} |\beta_z - \hat{\beta}_z|^2 \right)^{1/2} + \left(\sum_{|z| > M} |z|^{-2} |\beta_z|^2 \right)^{1/2}$$

Second term bounded by $M^{-(s+1)}$

First term of order $n^{-1/2} \left(\sum_{|z| \le M} |z|^{-2} \right)^{1/2} \approx n^{-1/2} \left(\sum_{1 \le r \le M} r^{-2} \cdot r^{d-1} \right)^{1/2}$

$$\lesssim n^{-1/2} \begin{cases} 1 & d = 1 \\ \sqrt{\log M} & d = 2 \\ M^{d/2-1} & d \ge 3 \end{cases}$$

W_1 bounds

Message:

- chose truncation to balance estimation & approximation error
- again, small scales dominate in high dimension, large scales in low
- $W_1(\mu, \hat{\mu})$ is small because $\|\mu - \hat{\mu}\|$ is small
- similar bound achievable via transportation method (worse log)

Surprisingly, this method does *not* work to get optimal rate for W_p!

W_p bounds

Consider $\mu \in \mathscr{P}([0,1]^d)$ with densities in Hölder class C^s

Subset of densities **bounded below**: $C^s(L; m) = C^s(L) \cap \{f : f \geq m\}$

Theorem: [Niles-Weed & Berthet '22] For any $p \geq 1$, $s \geq 0$, there exists estimator $\hat{f}$ satisfying

$$\sup_{f \in C^s(L;m)} \mathbb{E} W_p(f, \hat{f}) \lesssim \begin{cases} n^{-\frac{1+s}{d+2s}} & d \geq 3 \\ n^{-1/2} \log n & d = 2 \\ n^{-1/2} & d = 1 \end{cases}$$

Matches $p = 1$ result up to logarithm.

W_p bounds

Consider $\mu \in \mathscr{P}([0,1]^d)$ with densities in Hölder class C^s

Subset of densities **bounded below**: $C^s(L; m) = C^s(L) \cap \{f : f \geq m\}$

Theorem: [Niles-Weed & Berthet '22] For any $p \geq 1$, $s \geq 0$, there exists estimator $\hat{f}$ satisfying

$$\sup_{f \in C^s(L;m)} \mathbb{E} W_p(f, \hat{f}) \lesssim \begin{cases} n^{-\frac{1+s}{d+2s}} & d \geq 3 \\ n^{-1/2} \log n & d = 2 \\ n^{-1/2} & d = 1 \end{cases}$$

Matches $p = 1$ result up to logarithm.

W_p bounds

Consider $\mu \in \mathscr{P}([0,1]^d)$ with densities in Hölder class C^s

Subset of densities **bounded below**: $C^s(L; m) = C^s(L) \cap \{f : f \geq m\}$

Theorem: [Niles-Weed & Berthet '22] For any $p \geq 1$, $s \geq 0$, there exists estimator $\hat{f}$ satisfying

$$\sup_{f \in C^s(L; m)} \mathbb{E} W_p(f, \hat{f}) \lesssim m^{-1/p'} \begin{cases} n^{-\frac{1+s}{d+2s}} & d \geq 3 \\ n^{-1/2} \log n & d = 2 \\ n^{-1/2} & d = 1 \end{cases}$$

$$\frac{1}{p} + \frac{1}{p'} = 1$$

Matches $p = 1$ result up to logarithm.

W_p bounds

Theorem: [Niles-Weed & Berthet '22] For any $p \geq 1$, $s \in [0,1)$,

$$\sup_{f \in C^s(L)} \mathbb{E} W_p(f, \hat{\mu}) \lesssim \begin{cases} n^{-\frac{1+s/p}{d+s}} & d-s > 2p \\ n^{-1/2p} \log n & d-s = 2p \\ n^{-1/2p} & d-s < 2p \end{cases}$$

Remarks:

- strictly worse than rate for $m > 0$ for all $p > 1, s > 0$
- elbow depending on p reappears
- both rates minimax optimal

Why different rates?

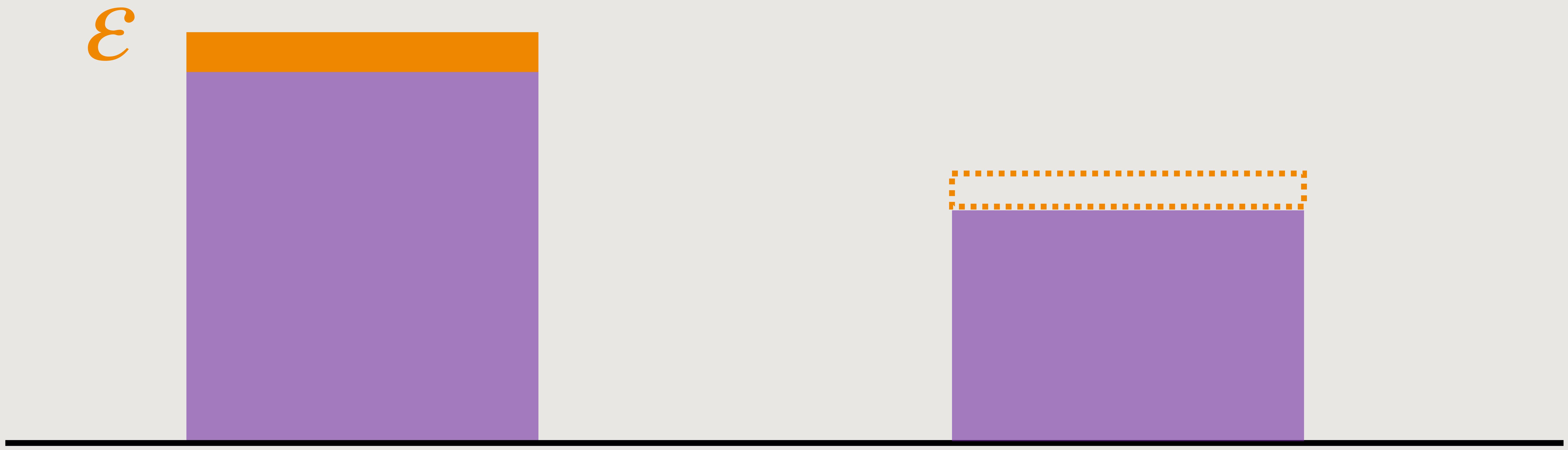
Intuition:

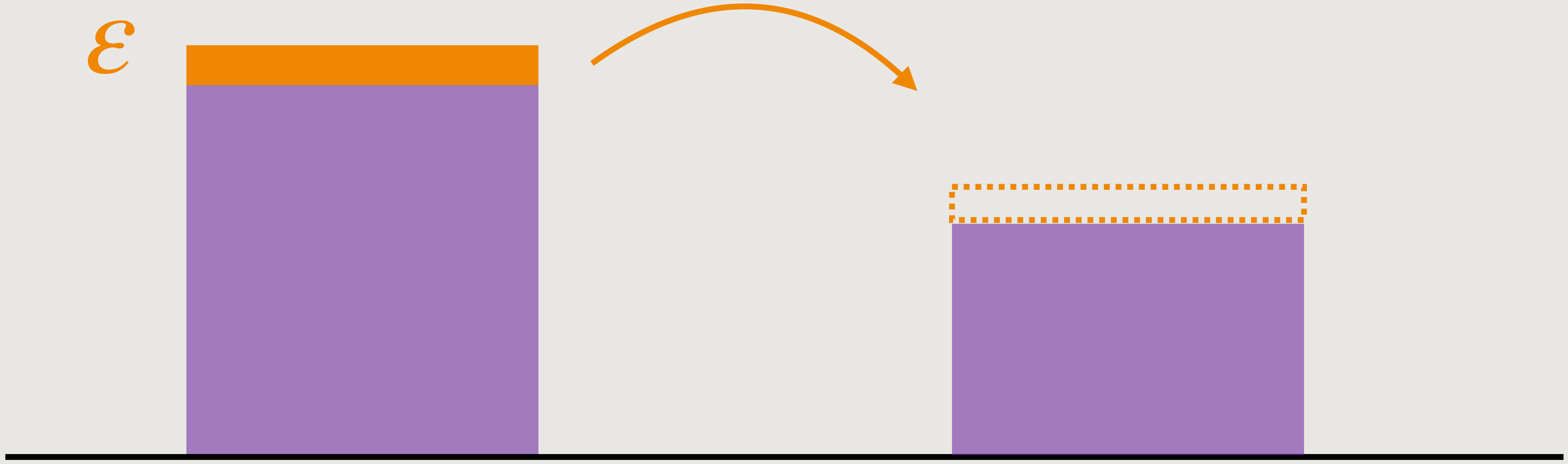
when density is bounded below, W_p is "norm-like"

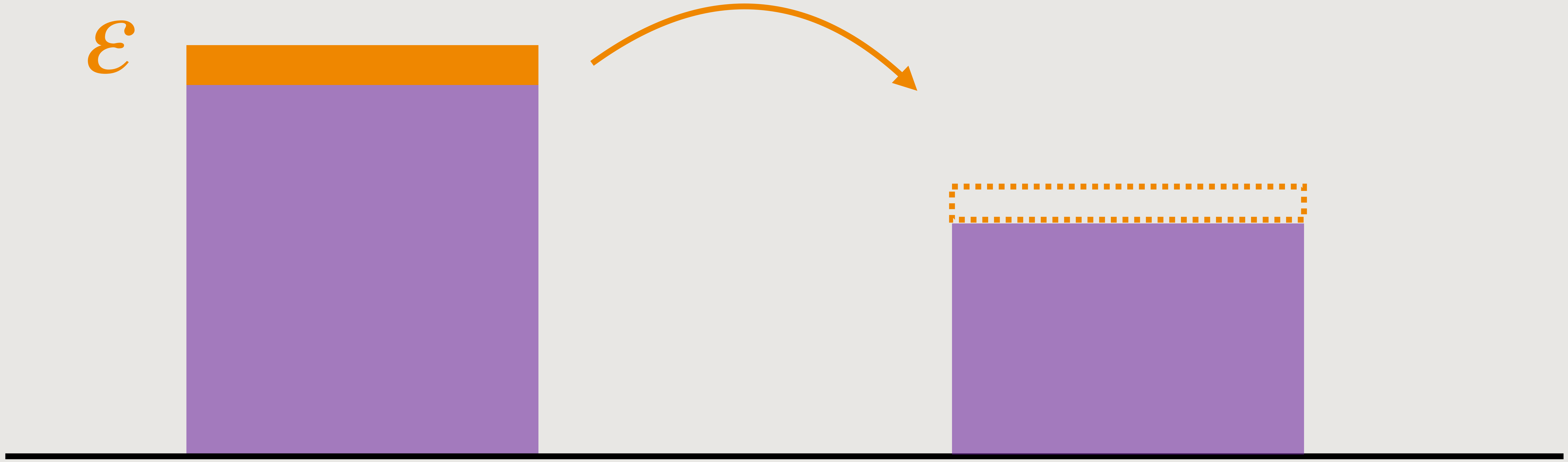
when density is not bounded below, W_p^p is "norm-like"


f

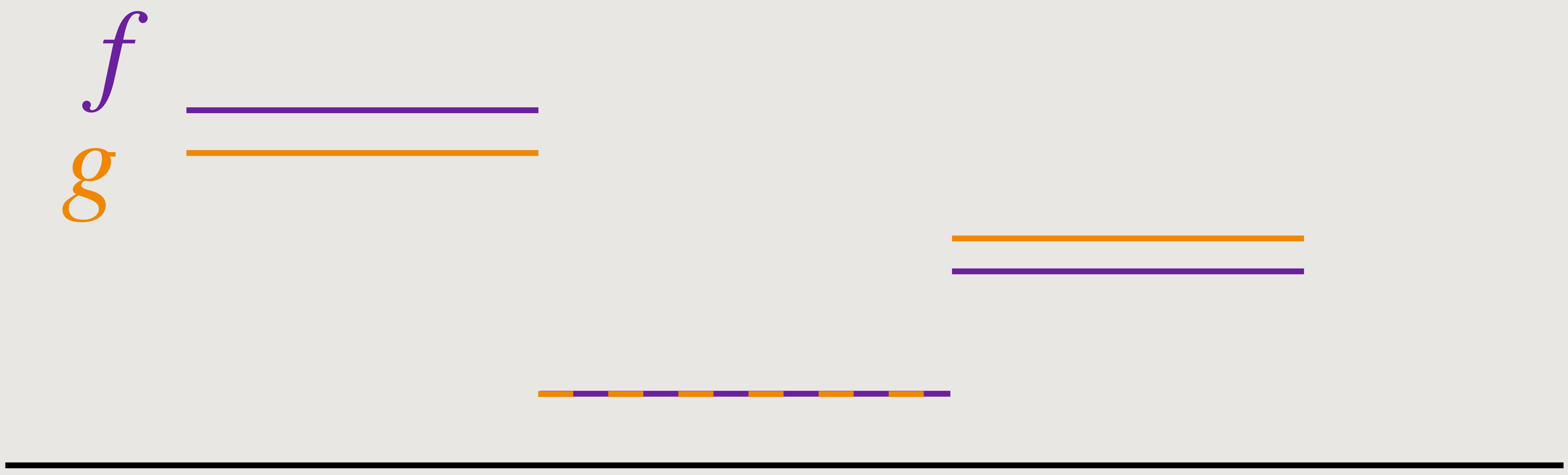

f
g

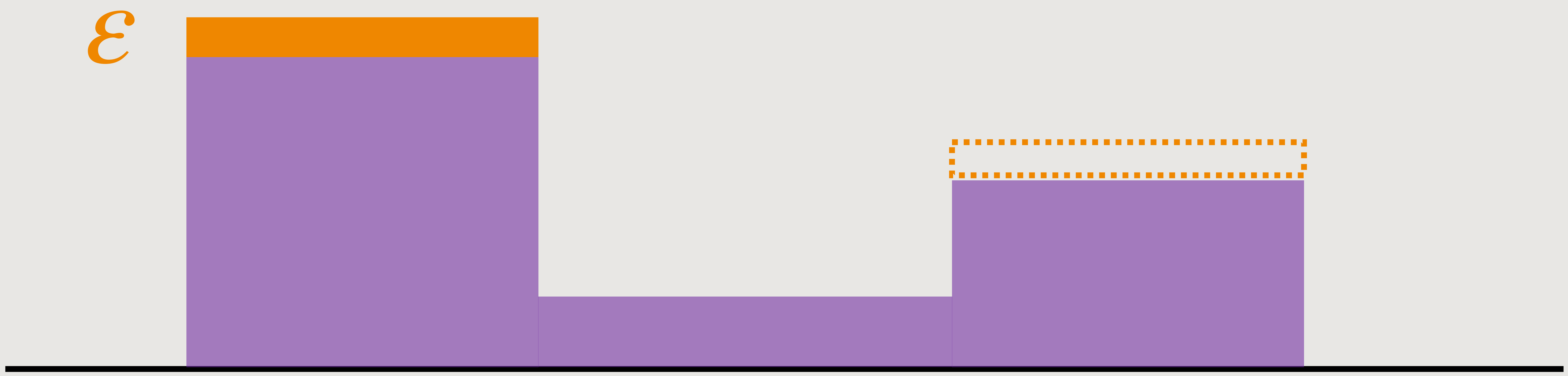

ε

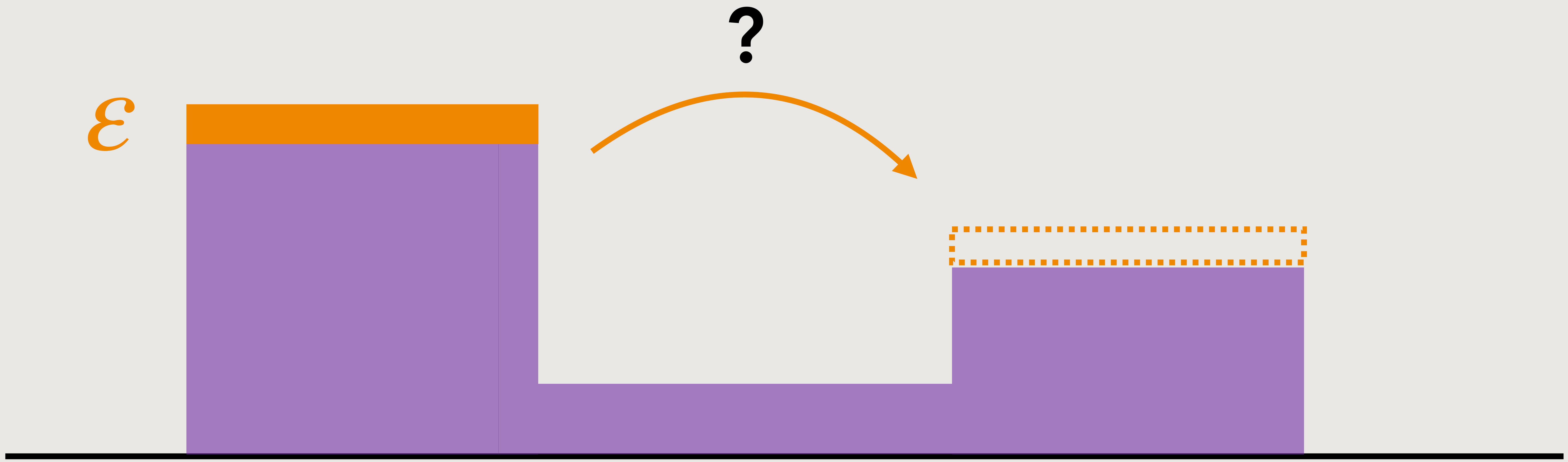

ε



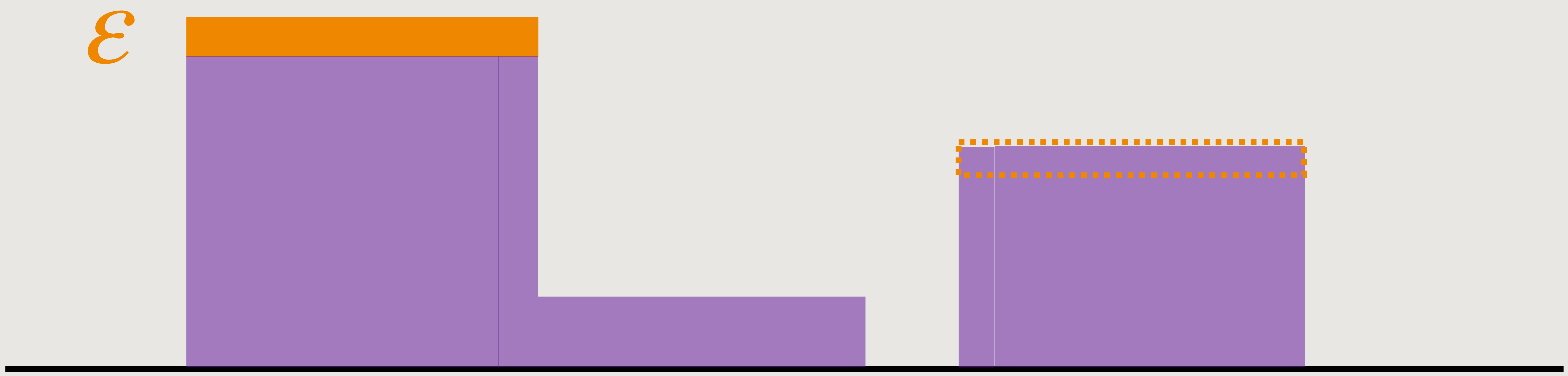
$$W_p^p = \int \|x - y\|^p \, \mathrm{d}\pi(x, y) \approx \varepsilon \cdot 1^p = \varepsilon$$

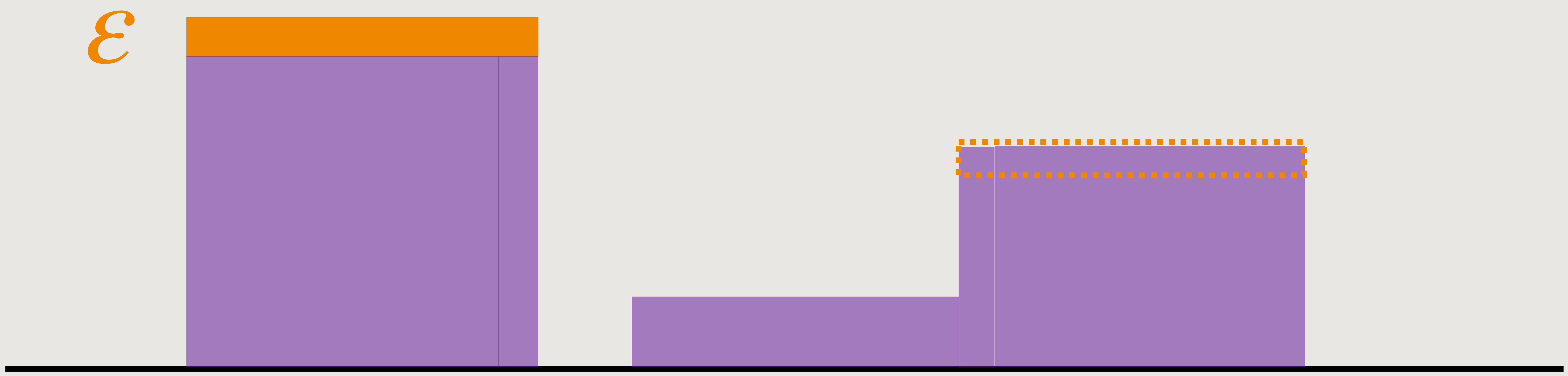

f
g

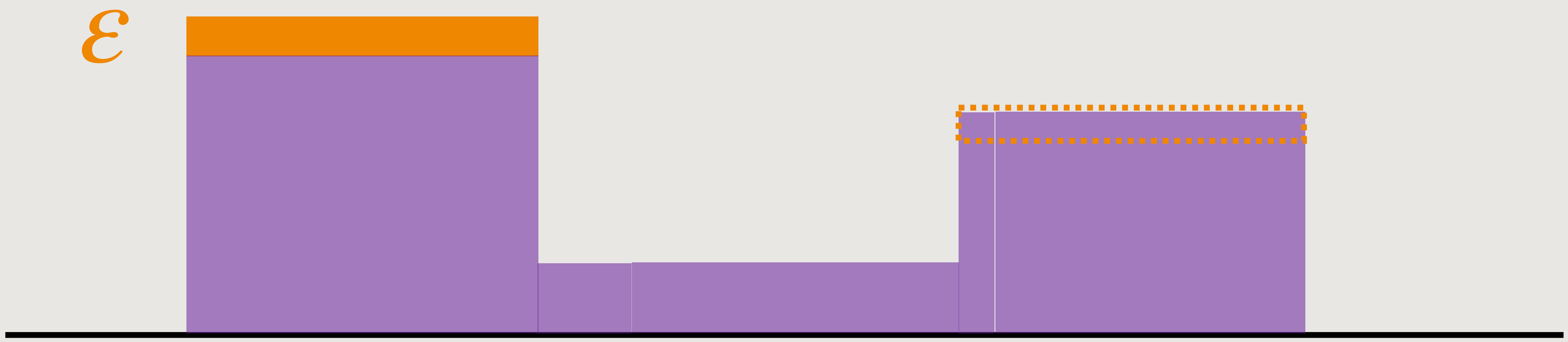

ε

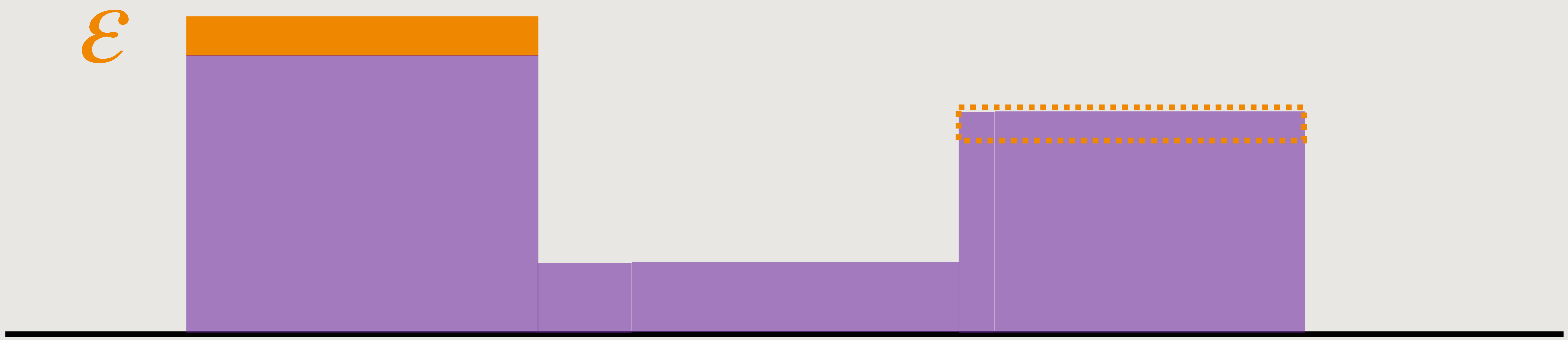


ε
?

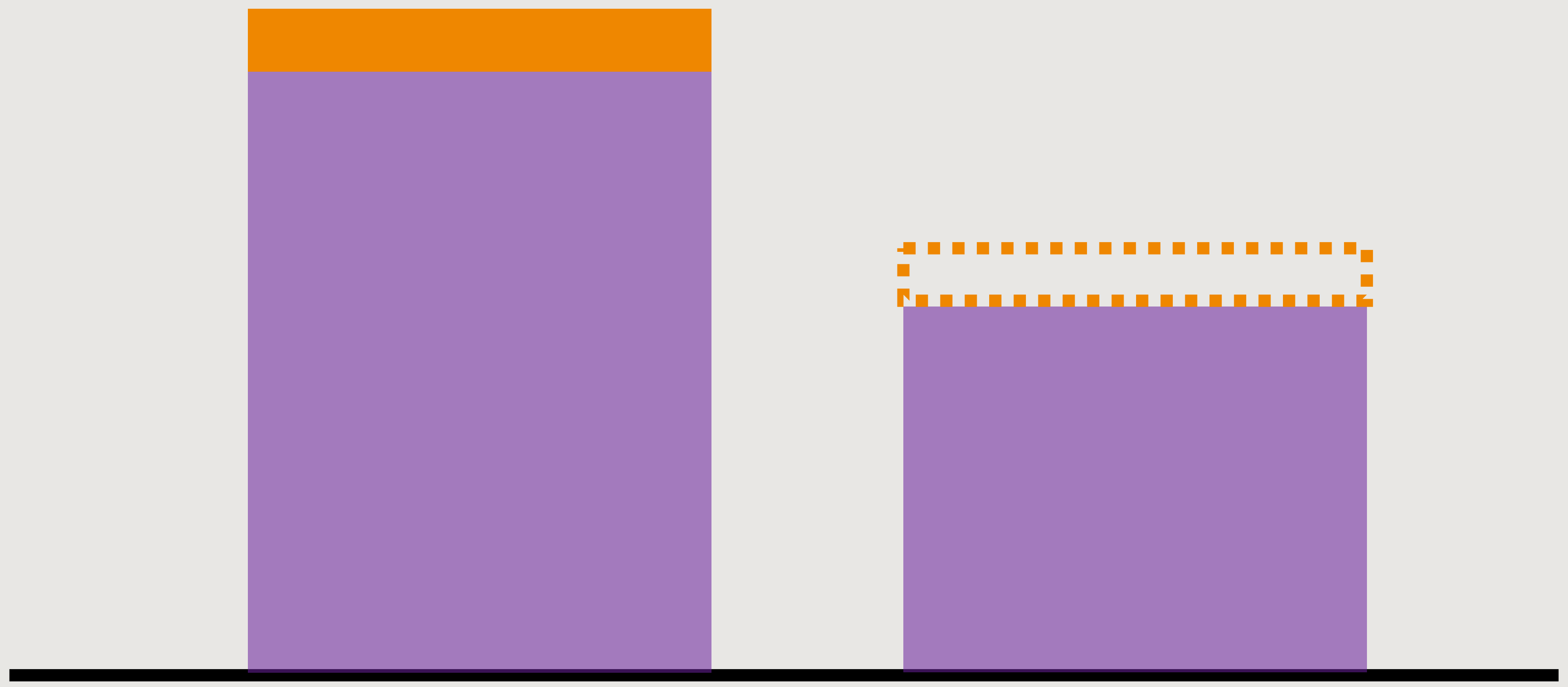

ε


ε

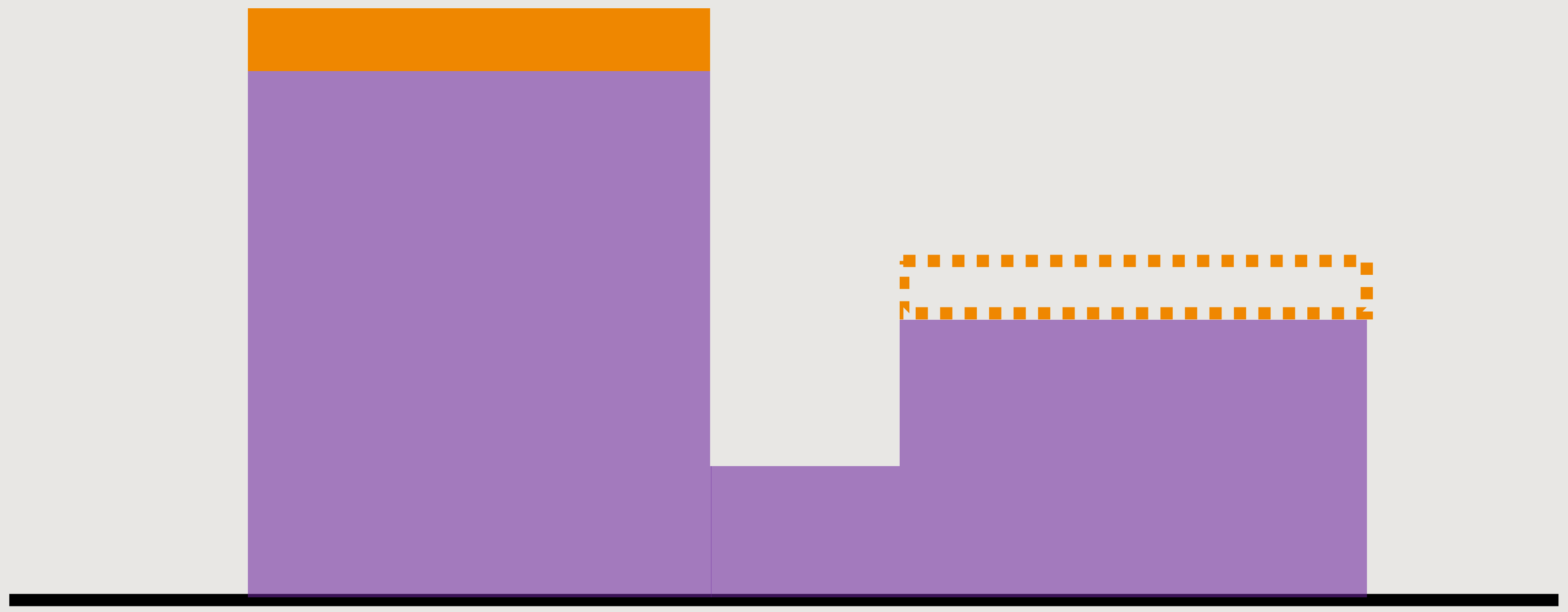

ε



$$W_p^p = \int \|x - y\|^p \, \mathrm{d}\pi(x, y) \approx \varepsilon^p \cdot 1 = \varepsilon^p$$



$$W_p^p \asymp \|f - g\|$$



$$W_p \asymp \|f - g\|$$

Proof technique

Main tool: refined *continuous* transportation method, exploiting *dynamic* characterization of W_p + **wavelet** decomposition.

Theorem: [Benamou & Brenier, '00]

$$W_p^p(\mu,\nu) = \inf_{\rho_t, E_t} \int_0^1 \int_\Omega \|v_t\|^p d\rho_t dt$$

such that $\rho_0 = \mu, \rho_1 = \nu$ and $\partial_t \rho_t + \nabla \cdot (v_t \rho_t) = 0$ (continuity eq).

Transport comes from dynamics $X_0 \sim \mu$, $\dot{X}_t = v_t(X_t)$.

Estmating Wasserstein distances, III

Jonathan Niles-Weed
Center for Data Science and Courant Institute, NYU

Review from Lectures I & II

Estimation of the Wasserstein distance is generally **slow** in high dimensions, unless very strong assumptions are made.

Questions:

- Are other distances better?
- Does it matter?

Other distances?

We will survey some proposals for modifications of Wasserstein distance, and show why they can be estimated more easily than W_p

Benefits and drawbacks compared to W_p

Main theme: inherent tension between **ease of estimation** and **discriminative power**

Discriminative power

A blessing and curse of Wasserstein distances is that they are strong: if $W_p(\mu, \nu)$ small, then μ and ν are *very* similar

Formalization: [Weaver, '18; Bing et al., '22] W_1 is **largest** (most discriminative) jointly convex metric on $\mathscr{P}(\mathbb{R}^d)$ satisfying $W_1(\delta_x, \delta_y) = \|x - y\|$.

Modifications of the Wasserstein distance with better statistical properties are necessarily less sensitive. (Possibly this is good!)

Entropic regularization

Blockbuster idea with many computational benefits. [Cuturi, '13; Peyré & Cuturi '19]

Gives rise to entropic OT cost:

$$OT_\epsilon(\mu, \nu) = \inf_{\pi \in \Pi(\mu,\nu)} \int \|x - y\|^2 d\pi(x, y) + \epsilon D(\pi \| \mu \otimes \nu)$$

and Sinkhorn divergence [Genevay et al., '18; Feydy et al. '19]:

$$S_\epsilon(\mu, \nu) = OT_\epsilon(\mu, \nu) - \frac{1}{2}(OT_\epsilon(\mu, \mu) + OT_\epsilon(\nu, \nu))$$

$(S_\epsilon(\mu, \nu) = 0 \iff \mu = \nu)$

Entropic regularization

Empirical estimation is substantially easier with entropic regularization!

Theorem: [Genevay et al., '19; Mena & Niles-Weed, '19; Rigollet & Stromme, '22; del Barrio et al., '22] If μ and ν have compact support, then

$$\mathbb{E}\,|\,OT_\epsilon(\mu_n, \nu) - OT_\epsilon(\mu, \nu)\,| \lesssim n^{-1/2}$$

$$|\,\mathbb{E}OT_\epsilon(\mu_n, \nu) - OT_\epsilon(\mu, \nu)\,| \lesssim n^{-1}$$

$$\mathbb{E}S_\epsilon(\mu_n, \mu) \lesssim n^{-1}$$

Constant depends very poorly on ϵ. Proof: empirical process method.

Entropic regularization

Empirical estimation is substantially easier with entropic regularization!

Downside: From theory perspective, statistical benefits only visible when ϵ is large—in which case optimal coupling is very blurry.

(Heuristic: minimizing entropic OT cost corresponds to deconvolution problem with Gaussian of variance ϵI. [Rigollet & Weed, '18])

Gaussian-smoothed Wasserstein

Given $t > 0$, define $W_p^{(t)}(\mu, \nu) = W_p(\mu * \rho_t, \nu * \rho_t)$

Gaussian measure with covariance $t \cdot \mathrm{I}$

Common trick in analysis: compared with μ, the convolved measure has bounded, smooth density, nice characteristic function, …

Investigated by [Weed '18; Goldfeld et al., '20; Goldfeld & Greenewald '20; Zhang et al., '21]. $W_p^{(t)}$ is a metric, recovers W_p in $t \to 0$ limit.

Gaussian-smoothed Wasserstein

Theorem [Goldfeld et al., '20]: If μ is compactly supported and $p \leq 2$, then

$$\mathbb{E} W_p^{(t)}(\mu, \mu_n) \lesssim n^{-1/2}$$

Proof idea: use that $\mu * \rho_t$ enjoys *log-Sobolev inequality*, which implies

$$W_2^2(\mu * \rho_t, \mu_n * \rho_t) \lesssim \mathrm{D}(\mu_n * \rho_t \| \mu * \rho_t)$$

Gaussian-smoothed Wasserstein

Theorem [Goldfeld et al., '20]**:** If μ is compactly supported and $p \leq 2$, then

$$\mathbb{E} W_p^{(t)}(\mu, \mu_n) \lesssim n^{-1/2}$$

Downside: No free lunch (implicit constant is exponential in d), not computationally feasible in high dimensions, no coupling

Integral probability metrics

Original "Wasserstein GAN" [Arjovsky et al., '17] consists of two neural nets: a generator f_G and discriminator f_D, trained with the objective,

$$\min_{f_G \in \mathscr{G}} \max_{f_D \in \mathscr{D}} \int f_D(x) d\mu(x) - \int f_D(f_G(z)) d\rho(z)$$

where ρ is a "base" measure, e.g., $\mathcal{N}(0, I)$. Enforce $\mathscr{D} \approx \mathrm{Lip}$ by "weight clipping", so that GAN minimizes $W_1(\mu, (f_G)_\sharp \rho)$.

Observations: [Arora et al., '17]

1. Need a lot of data in high dimensions ("fail to generalize")
2. But for real neural nets, $\mathscr{D}$ is much smaller than Lip!

Integral probability metrics

Slow convergence of W_1 because $\log N_\epsilon(\mathrm{Lip}) \asymp \epsilon^{-d}$.

But if $\mathscr{D}$ consists of neural networks with p parameters, then under sufficient regularity assumptions $\log N_\epsilon(\mathscr{D}) \asymp p \log \epsilon^{-1}$.

Benefits in high dimensions: if $d_{\mathscr{D}}(\mu, \nu) = \sup_{f_D \in \mathscr{D}} \int f_D(d\mu - d\nu)$,

$$\mathbb{E} d_{\mathscr{D}}(\mu, \mu_n) \asymp \sqrt{p/n} \ll \mathbb{E} W_1(\mu, \mu_n) \asymp n^{-1/d}$$

(**Still hopeless:** StyleGAN3 [Karras et al., '21] has 20m parameters.)

Integral probability metrics

More generally, an integral probability metric [Müller, '97] is any (pseudo)metric of the form

$$d_{\mathscr{F}}(\mu, \nu) = \sup_{f \in \mathscr{F}} \int f(d\mu - d\nu)$$

Smaller $\mathscr{F}$: less discriminative, but fewer samples to estimate

E.g., in maximum mean discrepancy (MMD) [Gretton et al., '06], $\mathscr{F}$ is unit ball in RKHS H, for which $\log N_\epsilon(\mathscr{F}) \leq C_H \log \epsilon^{-1}$.

Downsides: not usually computationally efficient, no coupling

Sliced Wasserstein distances

"Low-dimensional" variant of W_1, proposed for computational reasons [Rabin et al., '11]

$$SW_1(\mu,\nu) = \int W_1(\mu_\theta, \nu_\theta) d\sigma(\theta)$$

projection to $\mathrm{span}(\theta)$

uniform distribution on $\mathbb{S}_{d-1}$

Inherits $n^{-1/2}$ rate from the $d = 1$ case. [Nadjahi et al., '20]

Other variants based on different ways of aggregating $W_1(\mu_\theta, \nu_\theta)$. [Xi & Niles-Weed, '22]

Sliced Wasserstein distances

Downsides: Extremely weak in high dimensions [Huber, '85]

$$V_{j+1} = 65539 \cdot V_j \pmod{2^{31}}$$

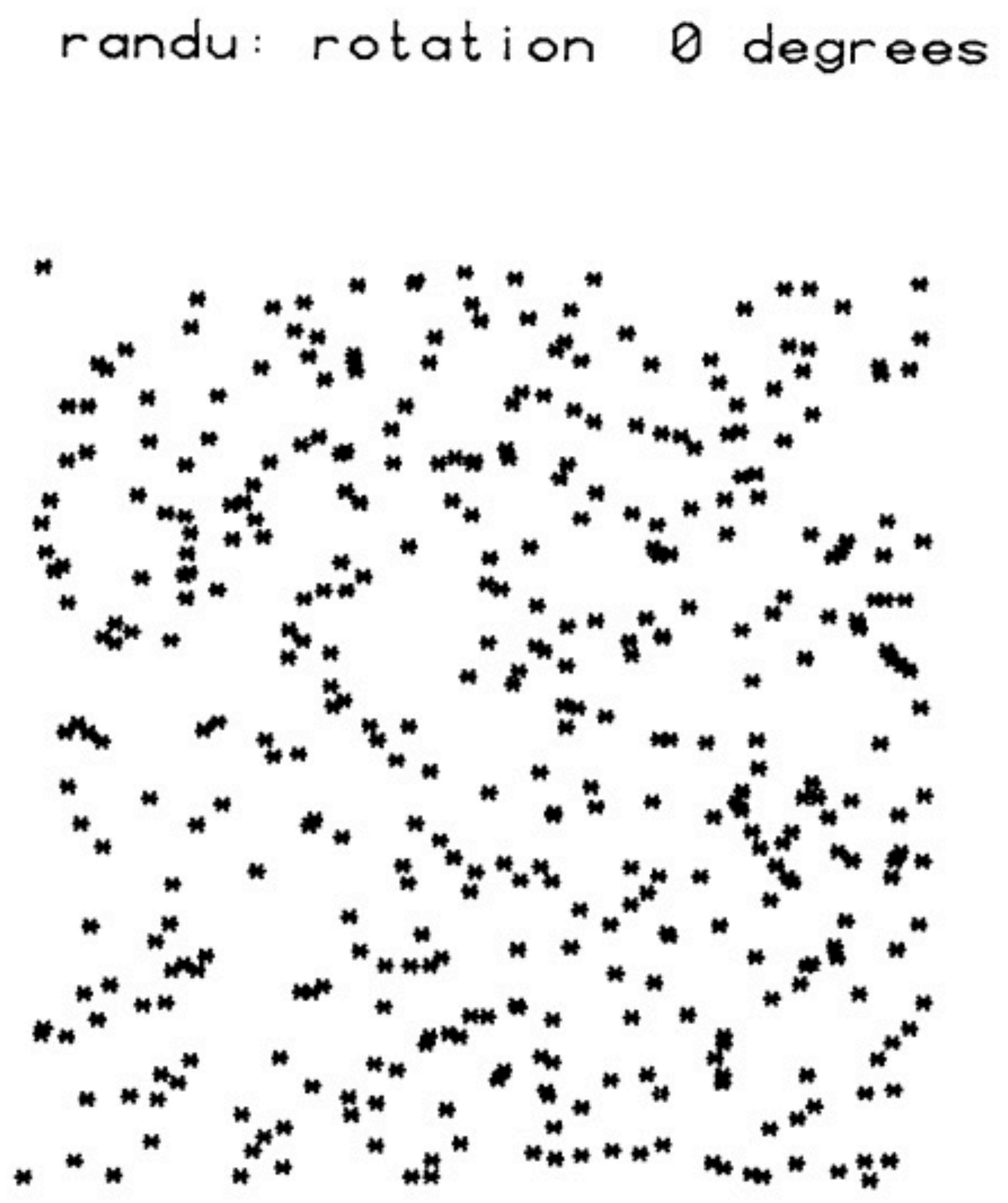


FIG. 2.1.

Sliced Wasserstein distances

Downsides: Extremely weak in high dimensions [Huber, '85]

$$V_{j+1} = 65539 \cdot V_j \pmod{2^{31}}$$

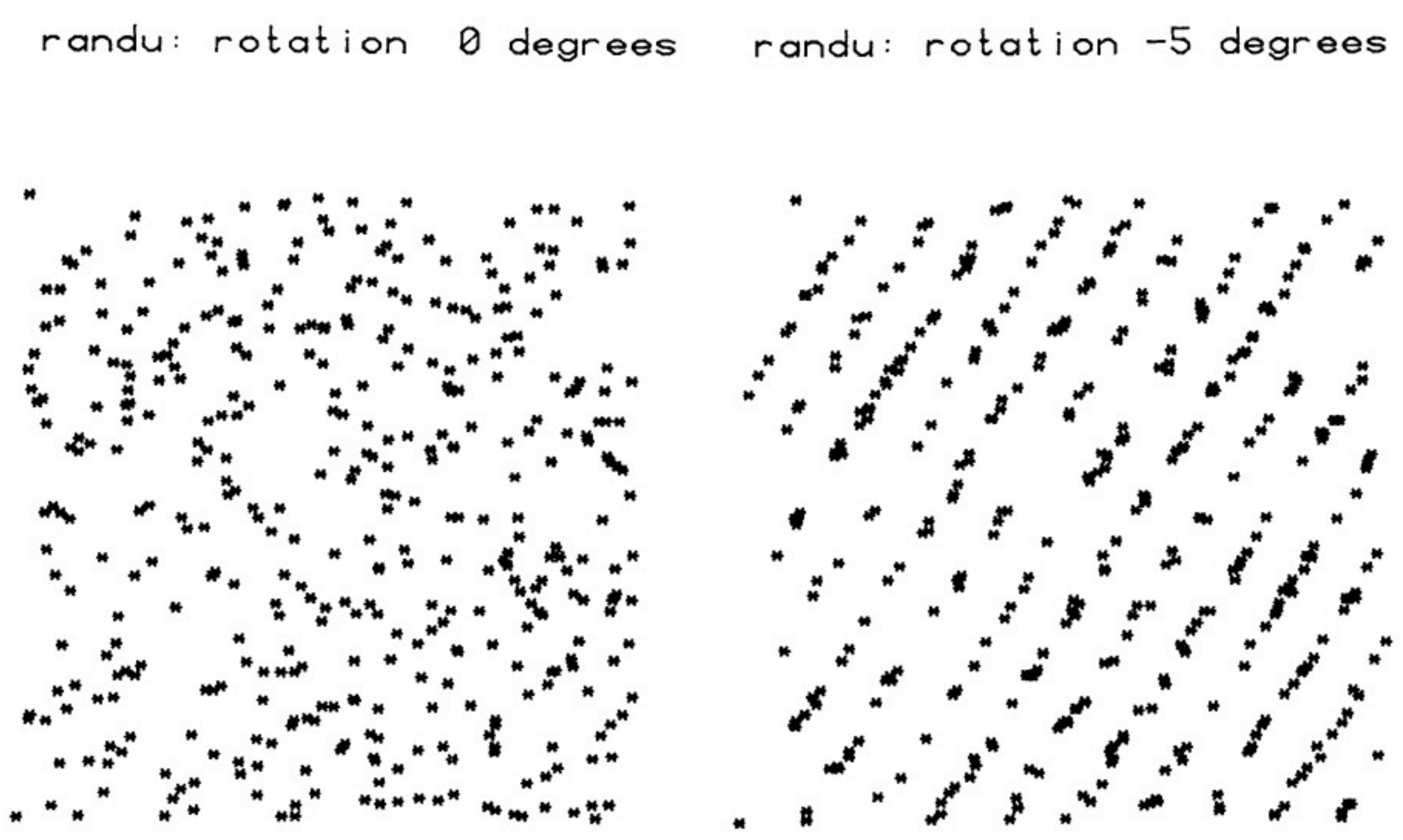
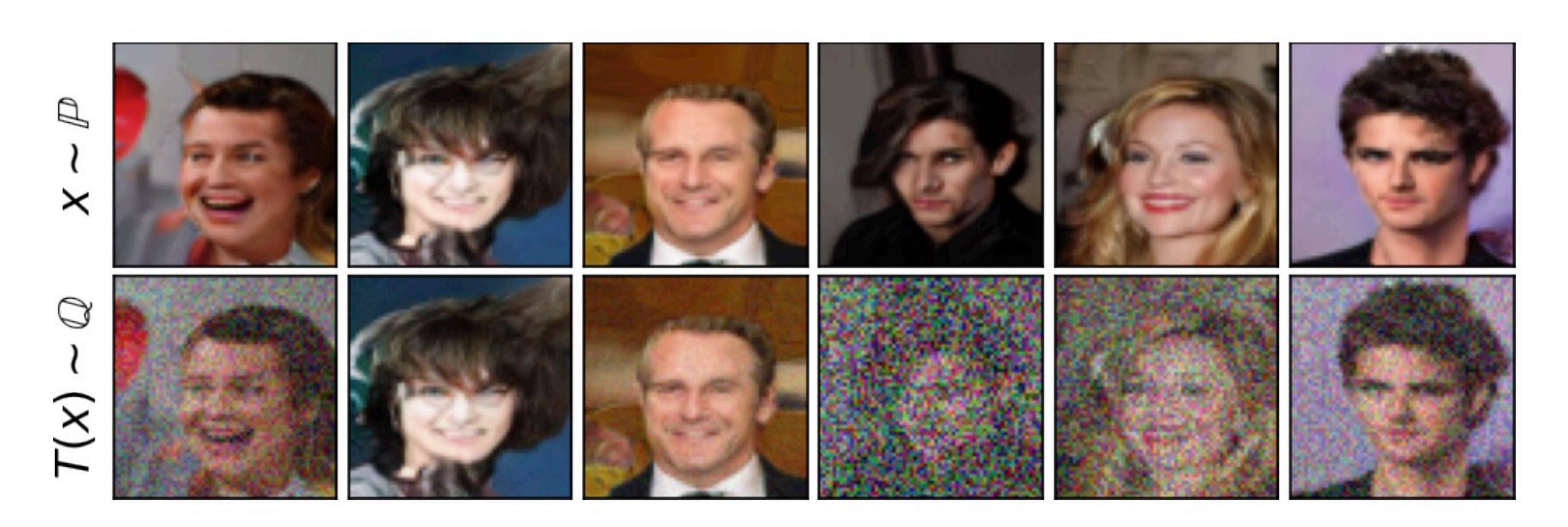


FIG. 2.1.

Does it matter?

Cuturi: "distances matter less than couplings and gradients"

Poor estimates of $W_p(\mu, \nu)$ can still contain useful information! [Korotin et al., '22]



synthetic $d = 2^{12}$ dimensional benchmark data set

Does it matter?

Trained 10 standard neural network-based W_1 discriminators

Does it matter?

Trained 10 standard neural network-based W_1 discriminators

N\Solver	$\lfloor$WC$\rceil$	$\lfloor$GP$\rceil$	$\lfloor$LP$\rceil$	$\lfloor$SN$\rceil$	$\lfloor$LS$\rceil$	$\lfloor$MM:B$\rceil$	$\lfloor$MM:Bv2$\rceil$	$\lfloor$MM$\rceil$	$\lfloor$MM:R$\rceil$	$\lfloor$DOT$\rceil$	True
1	$\gg$100	212.07	211.62	178.77	33.22	15.17	86.06	$\gg$100	37.92	66.03	58.24
16	$\gg$100	64.38	66.01	78.77	3.15	2.60	51.05	$\gg$100	41.24	53.54	29.78

$\widehat{W}_1$ versus W_1

Does it matter?

Trained 10 standard neural network-based W_1 discriminators

N\Solver	⌊**WC**⌉	⌊**GP**⌉	⌊**LP**⌉	⌊**SN**⌉	⌊**LS**⌉	⌊**MM:B**⌉	⌊**MM:Bv2**⌉	⌊**MM**⌉	⌊**MM:R**⌉	⌊**DOT**⌉	True
1	$\gg$100	212.07	211.62	178.77	33.22	15.17	86.06	$\gg$100	37.92	66.03	58.24
16	$\gg$100	64.38	66.01	78.77	3.15	2.60	51.05	$\gg$100	41.24	53.54	29.78

$\widehat{W}_1$ versus W_1

N\Solver	⌊**WC**⌉	⌊**GP**⌉	⌊**LP**⌉	⌊**SN**⌉	⌊**LS**⌉	⌊**MM:B**⌉	⌊**MM:Bv2**⌉	⌊**MM**⌉	⌊**MM:R**⌉	⌊**DOT**⌉	True
1	0.35	0.86	0.95	0.36	0.43	0.43	0.32	0.01	0.97	0.64	1.00
16	0.48	0.92	0.92	0.42	0.11	0.14	0.01	0.20	0.92	0.25	1.00

Correlation between $\nabla \widehat{W}_1$ and ∇W_1

Does it matter?

More broadly, what works best for actual applications? [Korotin et al., '21]

discrepancy, we further test the solvers in a setting of image generation. Our study reveals crucial limitations of existing solvers and shows that increased OT accuracy does not necessarily correlate to better results downstream.

Moral?

Existing statistical results are precise and mostly tight, but **fail to capture** important phenomena visible in practice.

Positive spin: much more to do!

References, I

- Arora, Ge, Liang, Ma, Zhang. "Generalization and Equilibrium in Generative Adversarial Nets (GANs)," 2017.
- Arjovsky, Chintala, Bottou. "Wasserstein GAN," 2017.
- Benamou, Brenier. "A computational fluid mechanics solution to the Monge-Kantorovich mass transfer problem," 2000.
- Bing, Bunea, Niles-Weed. "The Sketched Wasserstein Distance for mixture distributions," 2022.
- Blanchet, Murthy, Nguyen. "Statistical Analysis of Wasserstein Distributionally Robust Estimators," 2021.
- Boissard, Le Gouic. "On the mean speed of convergence of empirical and occupation measures in Wasserstein distance," 2014.
- Cuturi. "Sinkhorn Distances: Lightspeed Computation of Optimal Transport," 2013.
- Divol. "A short proof on the rate of convergence of the empirical measure for the Wasserstein distance," 2021.
- Dudley. "The Speed of Mean Glivenko-Cantelli Convergence," 1969.
- Feydy, Séjourné, Vialard, Amari, Trouve, Peyré. "Interpolating between Optimal Transport and MMD using Sinkhorn Divergences," 2019.
- Fournier, Guillin. "On the rate of convergence in Wasserstein distance of the empirical measure," 2015.
- Gangbo. "An elementary proof of the polar factorization of vector-valued functions," 1994.
- Genevay, Peyré, Cuturi. "Learning Generative Models with Sinkhorn Divergences," 2018.
- Goldfeld, Greenewald. "Gaussian-smoothed optimal transport: Metric structure and statistical efficiency," 2020.
- Goldfeld, Greenewald, Niles-Weed, Polyanskiy. "Convergence of smoothed empirical measures with applications to entropy estimation," 2020.
- Gretton, Borgwardt, Rasch, Schölkopf, Smola. "A kernel method for the two sample problem," 2006.
- Karras, Aittala, Laine, Härkönen, Hellsten, Lehtinen, Aila. "Alias-Free Generative Adversarial Networks," 2021.
- Korotin, Kolesov, Burnaev. "Kantorovich Strikes Back! Wasserstein GANs are not Optimal Transport?," 2022.
- Korotin, Li, Genevay, Solomon, Filippov, Burnaev. "Do neural optimal transport solvers work? A continuous Wasserstein-2 benchmark," 2021.
- Kuhn, Esfahani, Nguyen, Shafieezadeh-Abadeh."Wasserstein Distributionally Robust Optimization: Theory and Applications in Machine Learning," 2019.

References, II

- Huber. "Projection Pursuit," 1985.
- Hütter, Rigollet. "Minimax rates of estimation for smooth optimal transport maps," 2021.
- Hundrieser, Staudt, Munk. "Empirical Optimal Transport between Different Measures Adapts to Lower Complexity," 2022a.
- Hundrieser, Klatt, Staudt, Munk. "A Unifying Approach to Distributional Limits for Empirical Optimal Transport," 2022b.
- Ledoux. "On optimal matching of Gaussian samples," 2019.
- Lei. "Convergence and Concentration of Empirical Measures under Wasserstein Distance in Unbounded Functional Spaces," 2020.
- Liang. "How Well Generative Adversarial Networks Learn Distributions," 2021.
- Manole, Niles-Weed. "Sharp Convergence Rates for Empirical Optimal Transport with Smooth Costs," 2021.
- Müller. "Integral probability metrics and their generating classes of functions," 1997.
- Nadjahi, Durmus, Chizat, Kolouri, Shahrampour, Simsekli. "Statistical and topological properties of sliced probability divergences," 2020.
- Niles-Weed, Rigollet. "Estimation of Wasserstein distances in the spiked transport model," 2022.
- Peyré, Cuturi. *Computational Optimal Transport*, 2019.
- Rabin, Peyré, Delon, Bernot. "Wasserstein barycenter and its application to texture mixing," 2011.
- Rigollet, Weed. "Entropic optimal transport is maximum-likelihood deconvolution," 2018.
- Singh, Póczos. "Minimax Distribution Estimation in Wasserstein Distance," 2018.
- Singh, Uppal, Li, Li, Zaheer, Póczos. "Nonparametric Density Estimation under Adversarial Losses," 2018.
- Sommerfeld, Schrieber, Zemel, Munk. "Optimal Transport: Fast Probabilistic Approximation with Exact Solvers," 2019.
- Weaver. *Lipschitz Algebras*, 2018.
- Weed. "Sharper rates for estimating differential entropy under gaussian convolutions," 2018.
- Weed, Bach. "Sharp asymptotic and finite-sample rates of convergence of empirical measures in Wasserstein distance," 2019.
- Xi, Niles-Weed. "Distributional Convergence of the Sliced Wasserstein Process," 2022.
- Zhang, Cheng, Reeves. "Convergence of Gaussian-smoothed optimal transport distance with sub-gamma distributions and dependent samples," 2021.