Optimal Transport Map Estimation in General Function Spaces

Aram-Alexandre Pooladian
New York University

Simons Institute (UC Berkeley)
GMOS Reunion Workshop

in collaboration with

Vincent Divol

Jon Niles-Weed

Dirt Moving

Dirt Moving

Dirt Moving

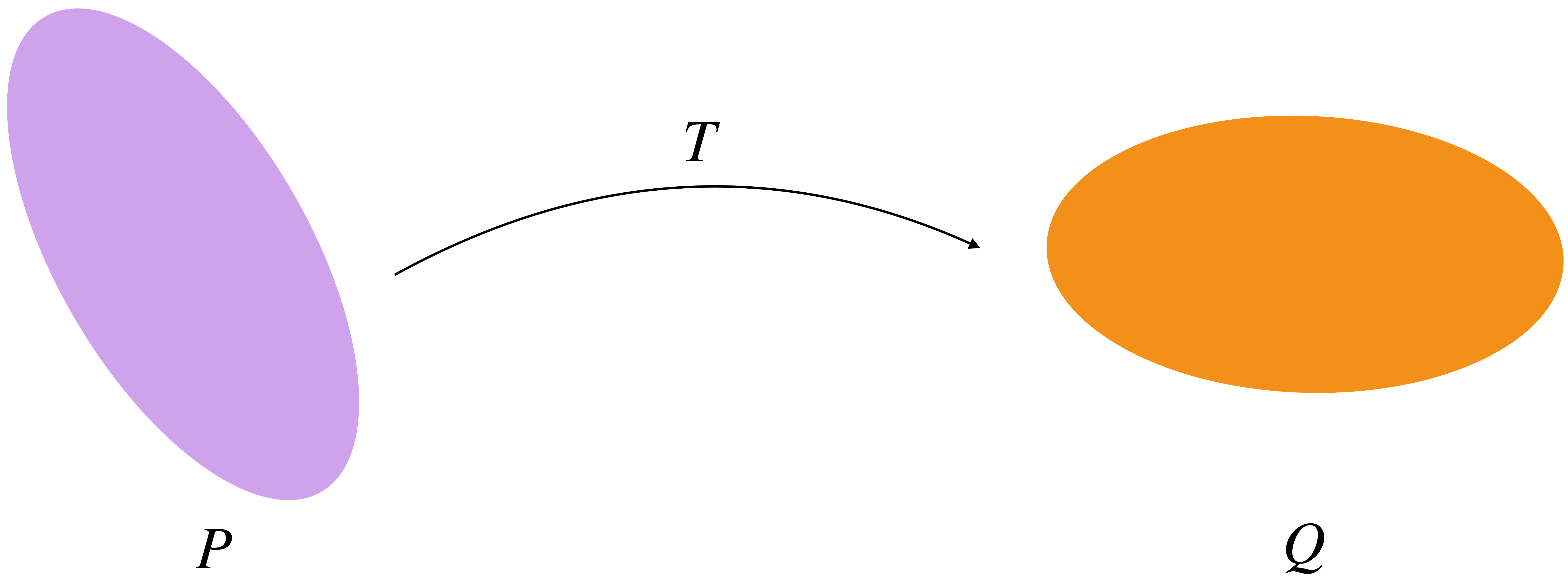
Dirt Moving

Dirt Moving

Transport maps

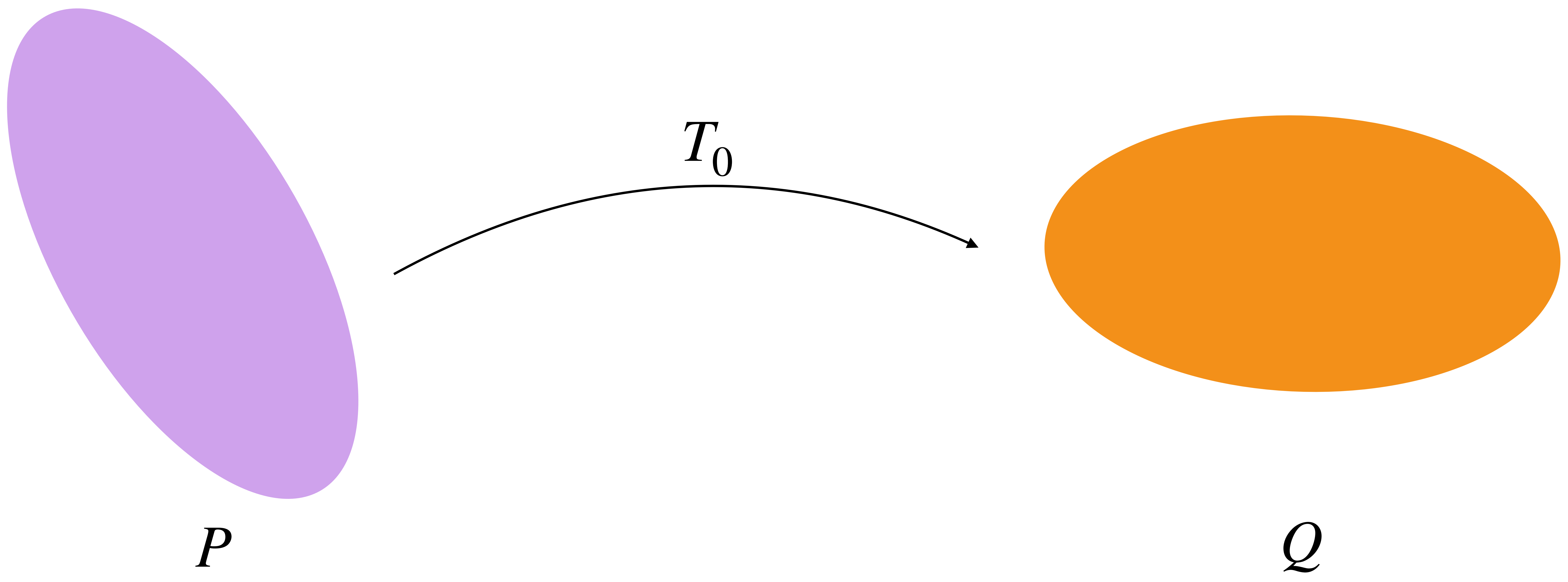
Transport maps

Transport maps

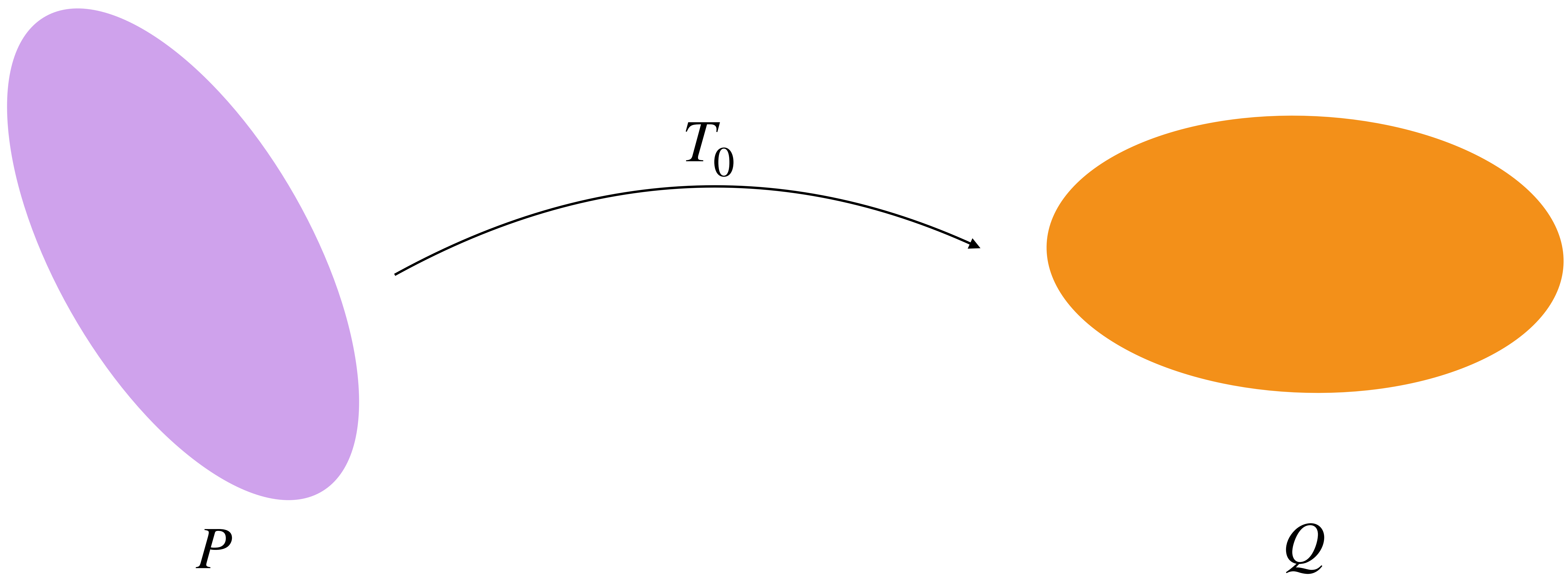


Call T a *transport map* if $T_\sharp P = Q$ i.e. $X \sim P, T(X) \sim Q$

<u>Optimal</u> transport maps



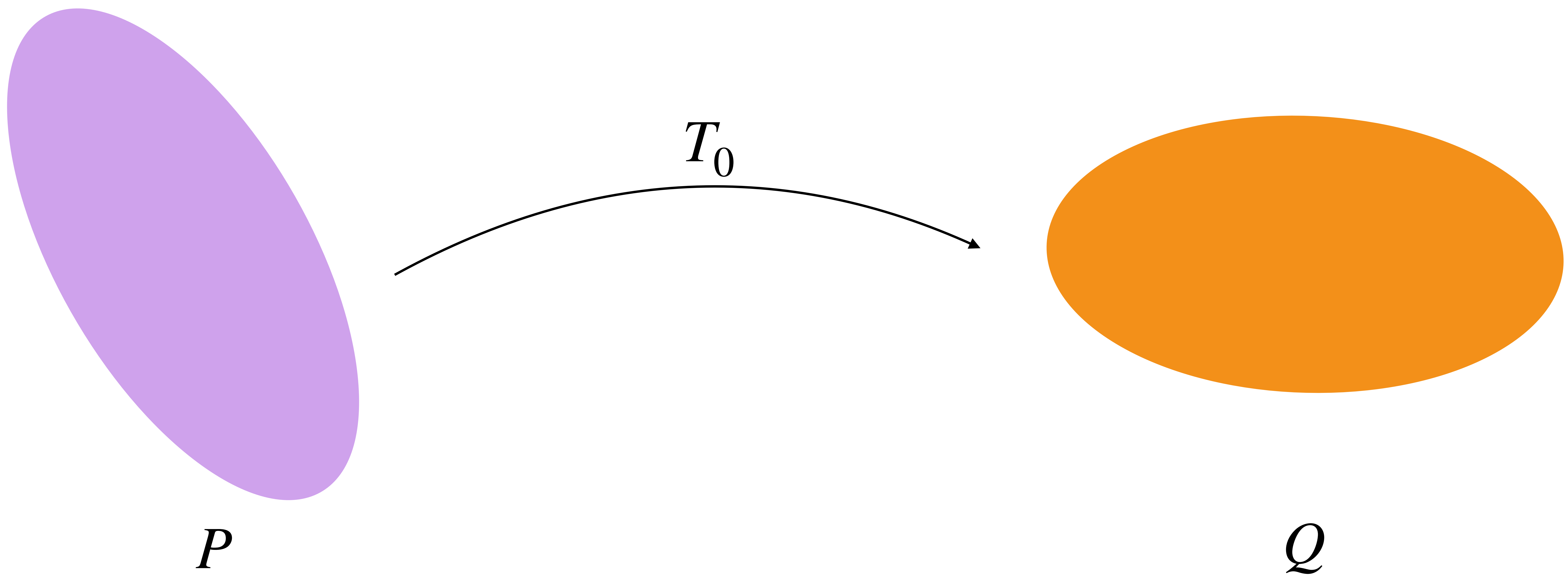
Optimal transport maps



Monge Problem

$$\boxed{T_0 := \operatorname*{argmin}_{T\,:\,T_\sharp P = Q} \int \frac{1}{2}\|x - T(x)\|_2^2 \, \mathrm{d}P(x)}$$

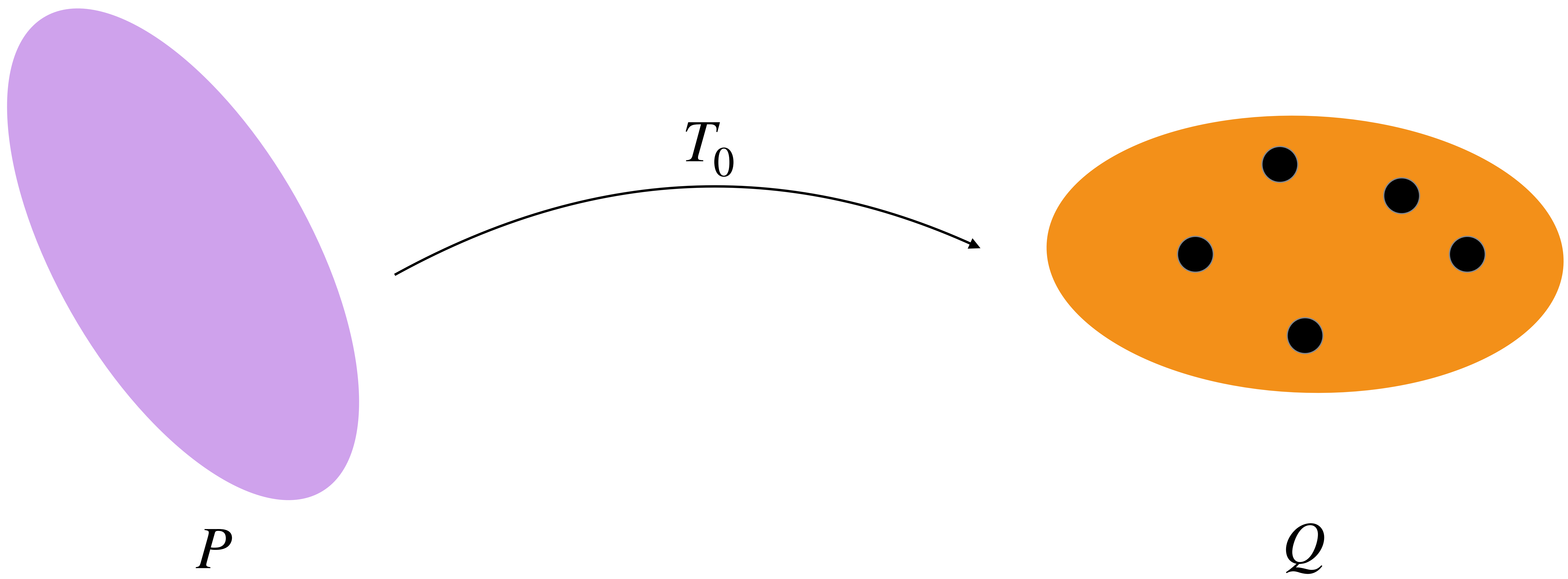
<u>Optimal</u> transport maps



$$\boxed{T_0 := \underset{T\,:\,T_\sharp P = Q}{\mathrm{argmin}} \int \frac{1}{2}\|x - T(x)\|_2^2 \,\mathrm{d}P(x)}$$

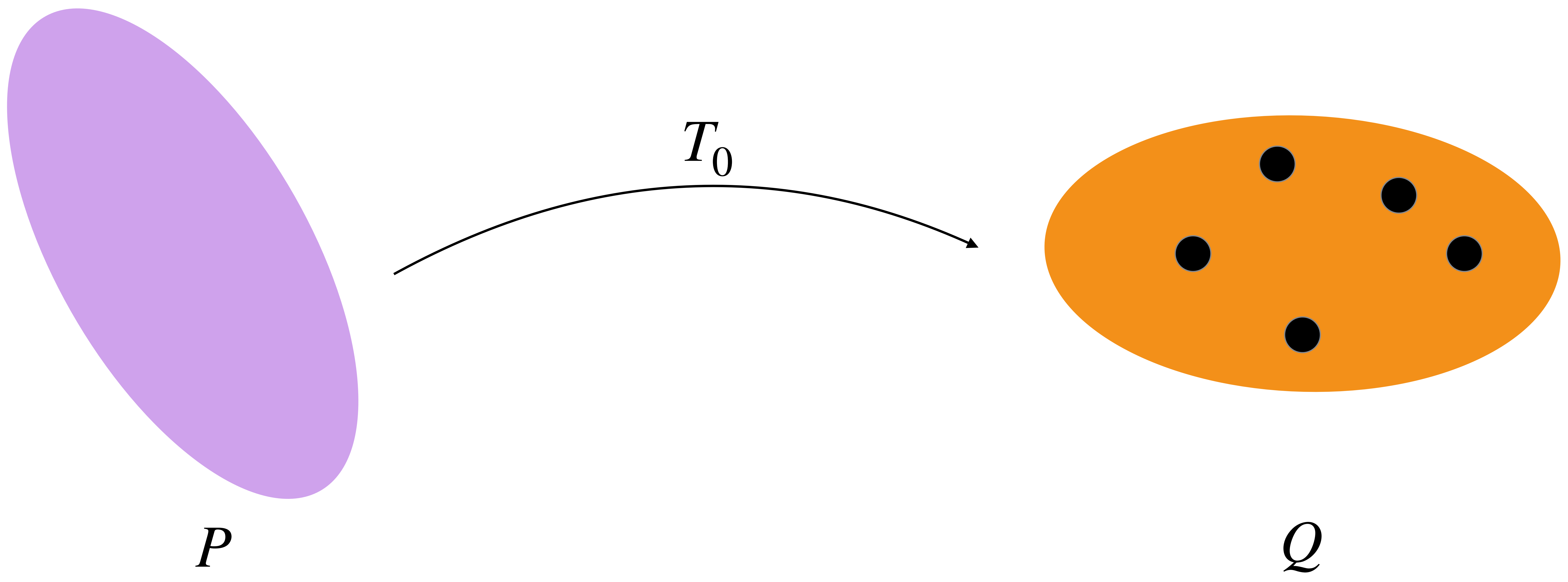
Brenier's Theorem: $T_0 = \nabla \varphi_0$ for some convex function φ_0

Statistical estimation of OT maps



Given P (e.g. standard Normal) and i.i.d samples $Y_1, \ldots, Y_n \sim Q$

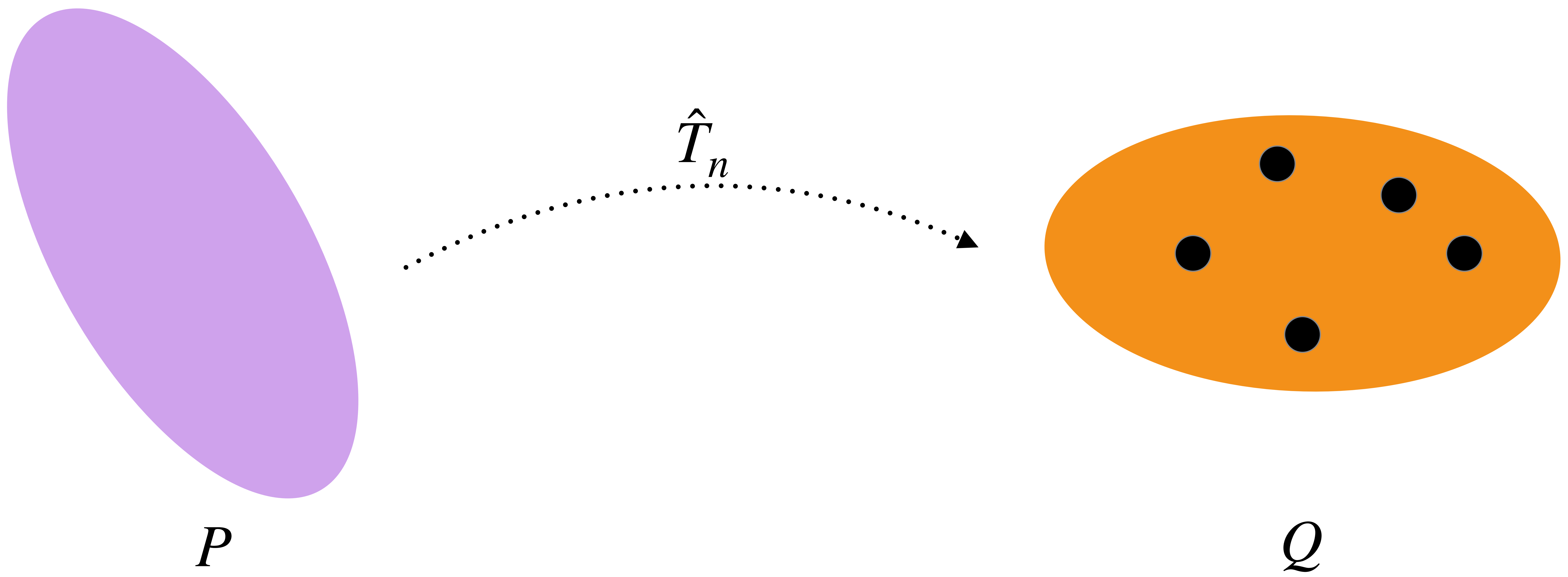
Statistical estimation of OT maps



Given P (e.g. standard Normal) and i.i.d samples $Y_1, \ldots, Y_n \sim Q$

Question: How to estimate T_0 on the basis of samples?

Statistical estimation of OT maps



Goal: Define estimator $\hat{T}_n$ s.t. under appropriate assumptions, $\mathbb{E}\|\hat{T}_n - T_0\|^2_{L^2(P)} \lesssim$?

Prior work

Assumptions (prior work):

- P and Q have compact support, with densities bounded above and below
- $T_0 \in C^s$ (s-times differentiable)
- T_0 is bi-Lipschitz, equivalently $I\alpha \preceq \nabla^2 \varphi_0 \preceq \beta I$

Results (prior work):

- [HR21] proposed a *wavelet* based estimator
- [MB+21] proposed the *1-Nearest-Neighbor* estimator
- [PNW21] proposed the *entropic map* estimator
- among others

Prior work

Results (prior work):

- [HR21] proposed a *wavelet* based estimator, $\nabla \hat{\varphi}_W$
- [MB+21] proposed a *1-Nearest-Neighbor* estimator
- [PNW21] proposed the *entropic map* estimator

Method: estimate φ_0 with wavelet class $W_J^{\alpha,\beta}$, need $0 < P_{\min} \leq P(x) \leq P_{\max}$

Prior work

Results (prior work):

- [HR21] proposed a *wavelet* based estimator, $\nabla \hat{\varphi}_W$
- [MB+21] proposed a *1-Nearest-Neighbor* estimator
- [PNW21] proposed the *entropic map* estimator

$$\mathbb{E}\|\nabla \hat{\varphi}_W - \nabla \varphi_0\|^2_{L^2(P)} \lesssim_{\log(n)} n^{-\frac{2s}{2s+d-2}}$$

Method: estimate φ_0 with wavelet class $W^{\alpha,\beta}_J$, need $0 < P_{\min} \leq P(x) \leq P_{\max}$

Prior work

Results (prior work):

- [HR21] proposed a *wavelet* based estimator, $\nabla \hat{\varphi}_W$
- [MB+21] proposed a *1-Nearest-Neighbor* estimator (s=1), $\hat{T}_{\text{1NN}}$
- [PNW21] proposed the *entropic map* estimator

Method: compute OT coupling $(X_i, Y_{\sigma(i)})$, match to closest $Y_{\sigma(i)}$ from training set

Prior work

Results (prior work):

- [HR21] proposed a *wavelet* based estimator, $\nabla \hat{\varphi}_W$
- [MB+21] proposed a *1-Nearest-Neighbor* estimator (s=1), $\hat{T}_{1\mathrm{NN}}$
- [PNW21] proposed the *entropic map* estimator

$$\mathbb{E}\|\hat{T}_{1\mathrm{NN}} - \nabla \varphi_0\|^2_{L^2(P)} \lesssim_{\log(n)} n^{-\frac{2}{d}}$$

Method: compute OT coupling $(X_i, Y_{\sigma(i)})$, match to closest $Y_{\sigma(i)}$ from training set

Prior work

Results (prior work):

- [HR21] proposed a *wavelet* based estimator, $\nabla \hat{\varphi}_W$
- [MB+21] proposed a *1-Nearest-Neighbor* estimator (s=1), $\hat{T}_{\mathrm{1NN}}$
- [PNW21] proposed the *entropic map* estimator (s=1), $\nabla \hat{\varphi}_\varepsilon$

Method: entropic optimal transport

Prior work

Results (prior work):

- [HR21] proposed a *wavelet* based estimator, $\nabla \hat{\varphi}_W$
- [MB+21] proposed a *1-Nearest-Neighbor* estimator (s=1), $\hat{T}_{1\mathrm{NN}}$
- [PNW21] proposed the *entropic map* estimator (s=1), $\nabla \hat{\varphi}_\varepsilon$

$$\mathbb{E}\|\nabla \hat{\varphi}_\varepsilon - \nabla \varphi_0\|^2_{L^2(P)} \lesssim_{\log(n)} n^{-\frac{1}{d+2}}$$

Method: entropic optimal transport

Prior work

Assumptions (prior work):

- P and Q have compact support, with densities bounded above and below
- $T_0 \in C^s$ (s-times differentiable)
- T_0 is bi-Lipschitz, equivalently $I\alpha \preceq \nabla^2 \varphi_0 \preceq \beta I$

This talk:

Prior work

Assumptions (prior work):

- P and Q have compact support, with densities bounded above and below
- $T_0 \in C^s$ (s-times differentiable)
- T_0 is bi-Lipschitz, equivalently $I\alpha \preceq \nabla^2 \varphi_0 \preceq \beta I$

This talk: extend assumptions to include

Prior work

Assumptions (prior work):

- P and Q have compact support, with densities bounded above and below
- $T_0 \in C^s$ (s-times differentiable)
- T_0 is bi-Lipschitz, equivalently $I\alpha \preceq \nabla^2 \varphi_0 \preceq \beta I$

This talk: extend assumptions to include

- P and Q not having compact support
- φ_0 can exist in more general function spaces

Semidual formulation

Semidual formulation

$$\frac{1}{2}W_2^2(P,Q) = \min_{T\,:\,T_\sharp P = Q} \int \frac{1}{2}\|x - T(x)\|_2^2 \,\mathrm{d}P(x)$$

Semidual formulation

$$\tfrac{1}{2}W_2^2(P,Q) = \min_{T\,:\,T_\sharp P = Q} \int \frac{1}{2}\|x - T(x)\|_2^2 \,\mathrm{d}P(x) \; = \tfrac{1}{2}(M_2(P) + M_2(Q)) - S(\varphi_0)$$

with $$S(\varphi_0) = \min_{\varphi} \int \varphi(x)\mathrm{d}P(x) + \int \varphi^*(y)\mathrm{d}Q(y)$$

Semidual formulation

$$\frac{1}{2}W_2^2(P,Q) = \min_{T\,:\,T_\sharp P = Q} \int \frac{1}{2}\|x - T(x)\|_2^2 \, \mathrm{d}P(x) = \frac{1}{2}(M_2(P) + M_2(Q)) - S(\varphi_0)$$

with $\quad S(\varphi_0) = \min_{\varphi} \int \varphi(x)\mathrm{d}P(x) + \int \varphi^*(y)\mathrm{d}Q(y)$

Why? φ_0 is the optimal Brenier potential, and $T_0 = \nabla \varphi_0$

Semidual formulation

Established that map estimation is equivalent to solving:

$$\operatorname{argmin}_{\varphi} S(\varphi) = \int \varphi(x)\mathrm{d}P(x) + \int \varphi^*(y)\mathrm{d}Q(y)$$

Semidual formulation

Established that map estimation is equivalent to solving:

$$\mathrm{argmin}_{\varphi}\, S(\varphi) = \int \varphi(x)\mathrm{d}P(x) + \int \varphi^*(y)\mathrm{d}Q(y)$$

Idea from [HR21]: study properties of the minimizer to the *empirical semidual*

Semidual formulation

Established that map estimation is equivalent to solving:

$$\text{argmin}_{\varphi}\ S(\varphi) = \int \varphi(x)\mathrm{d}P(x) + \int \varphi^*(y)\mathrm{d}Q(y)$$

Idea from [HR21]: study properties of the minimizer to the *empirical semidual*

$$\hat{\varphi}_{\mathcal{F}} = \underset{\varphi \in \mathcal{F}}{\text{argmin}}\ S_n(\varphi) := \int \varphi(x)\mathrm{d}P(x) + \frac{1}{n}\sum_{i=1}^{n} \varphi^*(Y_i)$$

for some function class $\mathcal{F}$ that φ_0 lies in or is close to.

Semidual formulation

Established that map estimation is equivalent to solving:

$$\text{argmin}_{\varphi}\ S(\varphi) = \int \varphi(x)\mathrm{d}P(x) + \int \varphi^*(y)\mathrm{d}Q(y)$$

Idea from [HR21]: study properties of the minimizer to the <u>*empirical semidual*</u>

$$\hat{\varphi}_{\mathscr{F}} = \underset{\varphi \in \mathscr{F}}{\text{argmin}}\ S_n(\varphi) := \int \varphi(x)\mathrm{d}P(x) + \frac{1}{n}\sum_{i=1}^{n} \varphi^*(Y_i)$$

for some function class $\mathscr{F}$ that φ_0 lies in or is close to.

Our final estimator is then $\hat{T} = \nabla \hat{\varphi}_{\mathscr{F}}$

Potential function classes

Examples of non-parametric classes:

- s-Hölder smooth functions (prior work)
- Reproducing Kernel Hilbert Spaces **(new!)**
- Shallow Neural Networks (a.k.a Barron space) **(new!)**
- "Low-dimensional" potential functions **(new!)**

Examples of parametric classes:

- Finite set **(new!)**
- Quadratics potentials **(new!)**
- Input Convex Neural Networks (ICNNs) **(new!)**

Assumptions

- **(A1)** P satisfies a Poincaré inequality (with bounded or **unbounded domain!**)
- **(A2)** All $\varphi \in \mathscr{F}$ are β-smooth — $\nabla^2 \varphi \preceq \beta I$
- **(A3)** φ_0 is α-strongly convex and β-smooth — $\alpha I \preceq \nabla^2 \varphi_0 \preceq \beta I$
- **(A4)** Metric entropy condition on $\mathscr{F}$

“Meta” theorems

[Theorem 2+3, (Divol, Niles-Weed, P. 2022)]

"Meta" theorems

[Theorem 2+3, (Divol, Niles-Weed, P. 2022)]

$$\mathbb{E}\|\nabla\hat{\varphi}_{\mathscr{F}} - \nabla\varphi_0\|^2_{L^2(P)} \lesssim_{\log(n),\log(d)} \mathrm{Rate}(\mathscr{F}, n)$$

"Meta" theorems

[Theorem 2+3, (Divol, Niles-Weed, P. 2022)]

$$\mathbb{E}\|\nabla\hat{\varphi}_{\mathscr{F}} - \nabla\varphi_0\|^2_{L^2(P)} \lesssim_{\log(n),\log(d)} \text{Rate}(\mathscr{F}, n)$$

Today: **second** of two "meta" theorems:

- Theorem 2 has <u>suboptimal</u> rates but weaker conditions
- To have improved rates: need strong convexity, Poincaré inequality, and P having a nice density

Sanity checks

Sanity check: $\mathcal{F}$ is a finite set

Suppose:

- $\mathcal{F} = \{\varphi_1, \ldots, \varphi_K\}$ is a set of strongly convex, smooth potentials
- You know that $\varphi_0 \in \mathcal{F}$

$$\mathbb{E}\|\nabla \hat{\varphi}_{\mathcal{F}} - \nabla \varphi_0\|^2_{L^2(P)} \lesssim_{\log(n)} n^{-1}$$

Improves upon the work of [VV21]; they don't assume Poincaré

Sanity check: $\mathcal{F}$ is the set of Quadratics

Suppose:

- $\mathcal{F} = \{x \mapsto \frac{1}{2}x^\top A^{1/2}x + b^\top x \ : \ A \in \mathbb{S}_+^d, b \in \mathbb{R}^d\}$
- You know that $\varphi_0 \in \mathcal{F}$ i.e. $T_0(x) = A^{1/2}x + b$

$$\mathbb{E}\|\nabla \hat{\varphi}_{\mathcal{F}} - \nabla \varphi_0\|^2_{L^2(P)} \lesssim_{\log(n)} n^{-1}$$

Recovers the work of [FLF19] where they use the plug-in estimator

Sanity check: Parametric family

Let $\Theta \subseteq \mathbb{R}^m$ and consider potentials s.t. $|\varphi_\theta(x) - \varphi_{\theta'}(x)| \leq L\|\theta - \theta'\|(1 + \|x\|)^p$

Example: φ_0 can be represented as an ICNN with m parameters

$$\mathbb{E}\|\nabla \hat{\varphi}_{\mathscr{F}} - \nabla \varphi_0\|^2_{L^2(P)} \lesssim_{\log(n)} \frac{m}{n}$$

Example 1: RKHS

Suppose $f \in \mathscr{H}$ with $f(x) = \langle f, \mathscr{K}(\,\cdot\,, x)\rangle_{\mathscr{H}}$ and $\mathscr{K}$ is sufficiently nice

- $\mathscr{K}$ has finite spectrum
- $\mathscr{K}$ has exponentially decaying spectrum (e.g. Gaussian Kernel)

$$\mathbb{E}\|\nabla \hat{\varphi}_{\mathscr{F}} - \nabla \varphi_0\|^2_{L^2(P)} \lesssim_{\log(n)} n^{-1}$$

Example 2: Hölder-smooth functions

Suppose $\varphi_0 \in C_L^{s+1}(\Omega)$ and let $\mathscr{F} = W_J(\square_R)$ (finite wavelets over cube)

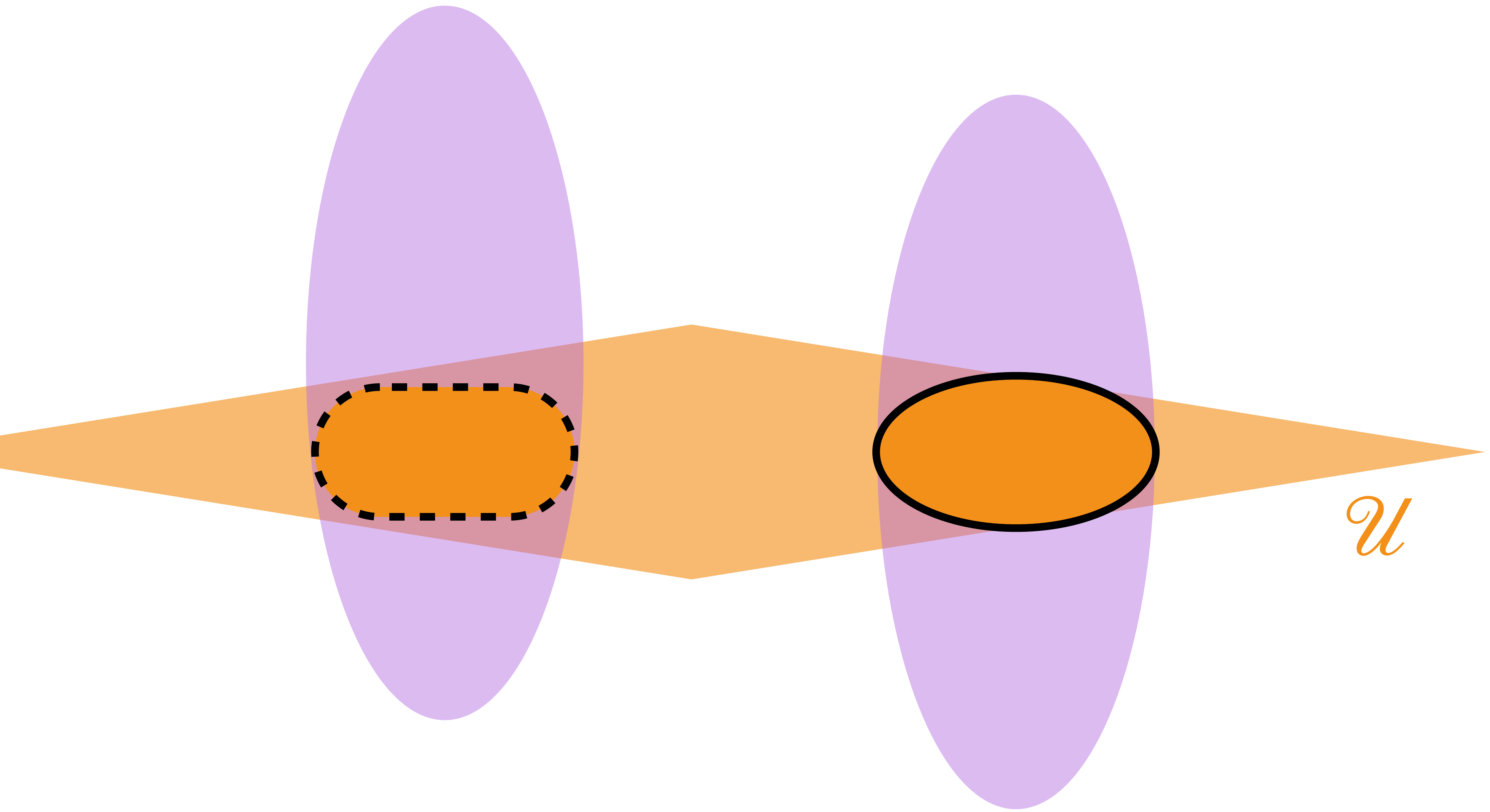
$$\mathbb{E}\|\nabla\hat{\varphi}_{\mathscr{F}} - \nabla\varphi_0\|^2_{L^2(P)} \lesssim_{\log(n)} n^{-\frac{2s}{2s+d-2}}$$

Special case: when both P and Q are log-smooth, and log-strongly concave, Caffarelli contraction kicks in (see [Chewi, **P.** 2022]):

$$\mathbb{E}\|\nabla\hat{\varphi}_{\mathscr{F}} - \nabla\varphi_0\|^2_{L^2(P)} \lesssim_{\log(n)} n^{-2/d}$$

Example 3: "Low-dimensional" potentials

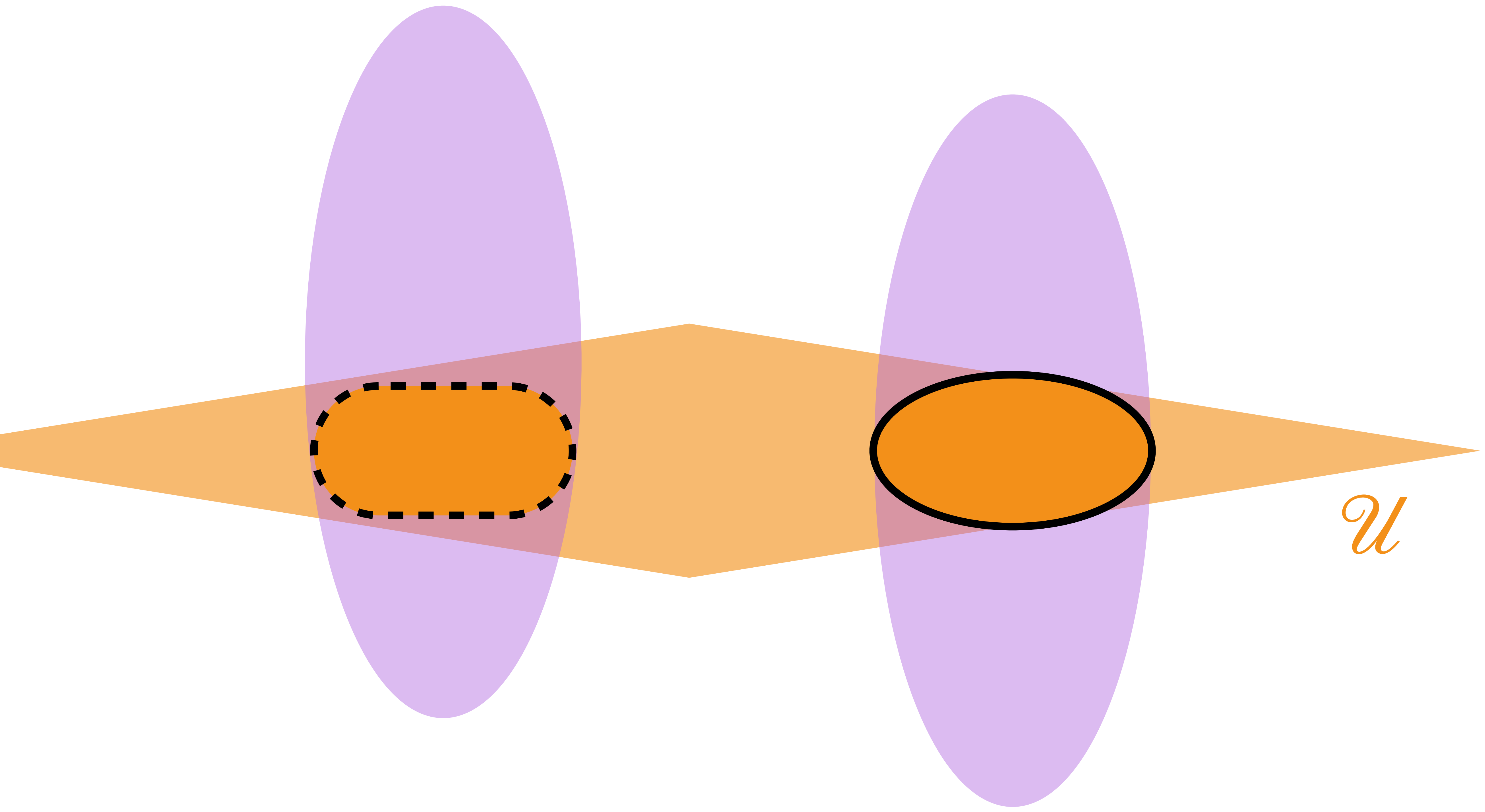
Potential functions that resemble the "Spiked Transport Model" [NWR21]



- P and Q live on $\mathscr{U}$
- Support $\mathscr{U}$ with $\dim(\mathscr{U}) = k$
- Noise outside i.e. on $\mathscr{U}^\perp$
- Only pay for underlying dimension $k \ll d$

Example 3: "Low-dimensional" potentials

Potential functions that resemble the "Spiked Transport Model" [NWR21]



- P and Q live on $\mathcal{U}$
- Support $\mathcal{U}$ with $\dim(\mathcal{U}) = k$
- Noise outside i.e. on $\mathcal{U}^\perp$
- Only pay for underlying dimension $k \ll d$

Final rate: $n^{-\frac{2s}{2s+k-2}} \ll n^{-\frac{2s}{2s+d-2}}$

Example 4: Barron Spaces

Example 4: Barron Spaces

We now say $\varphi \in \mathscr{F}_\sigma$ if we can write $\varphi(x) = \int \sigma(x, v) \, \mathrm{d}\theta(v)$ where

Example 4: Barron Spaces

We now say $\varphi \in \mathcal{F}_\sigma$ if we can write $\varphi(x) = \int \sigma(x, v)\ \mathrm{d}\theta(v)$ where

- $x \mapsto \sigma(x, v)$ is convex, with $\sigma(0,v) = 0$, and β-smooth
- $v \mapsto \sigma(x, v) \in C^s(\mathcal{M})$

Example 4: Barron Spaces

We now say $\varphi \in \mathcal{F}_\sigma$ if we can write $\varphi(x) = \int \sigma(x, v)\ \mathrm{d}\theta(v)$ where

- $x \mapsto \sigma(x, v)$ is convex, with $\sigma(0,v) = 0$, and β-smooth
- $v \mapsto \sigma(x, v) \in C^s(\mathcal{M})$

See e.g.
- [EMW22], [Bach17] for theory
- [Mak+20], [Hua+21], [BKC22] for practice

Example 4: Barron Spaces

We now say $\varphi \in \mathscr{F}_\sigma$ if we can write $\varphi(x) = \int \sigma(x, v)\, \mathrm{d}\theta(v)$ where

- $x \mapsto \sigma(x, v)$ is convex, with $\sigma(0,v) = 0$, and β-smooth
- $v \mapsto \sigma(x, v) \in C^s(\mathscr{M})$

Example: $\sigma(\langle x, v\rangle) = \langle x, v\rangle_+^2$ i.e. $\nabla \varphi_0$ is a shallow NN with ReLU activation

Example 4: Barron Spaces

We now say $\varphi \in \mathcal{F}_\sigma$ if we can write $\varphi(x) = \displaystyle\int \sigma(x, v)\ \mathrm{d}\theta(v)$ where

- $x \mapsto \sigma(x, v)$ is convex, with $\sigma(0,v) = 0$, and β-smooth
- $v \mapsto \sigma(x, v) \in C^s(\mathcal{M})$

Example: $\sigma(\langle x, v \rangle) = \langle x, v \rangle_+^2$ i.e. $\nabla \varphi_0$ is a shallow NN with ReLU activation

$$\mathbb{E}\|\nabla \hat{\varphi}_{\mathcal{F}_\sigma^1} - \nabla \varphi_0\|_{L^2(P)}^2 \lesssim_{\log(n)} n^{-\frac{1}{2}-\frac{1}{d}}$$

Example 4: Barron Spaces

We now say $\varphi \in \mathscr{F}_\sigma$ if we can write $\varphi(x) = \int \sigma(x, v)\ \mathrm{d}\theta(v)$ where

- $x \mapsto \sigma(x, v)$ is convex, with $\sigma(0,v) = 0$, and β-smooth
- $v \mapsto \sigma(x, v) \in C^s(\mathscr{M})$

Example: $\sigma(\langle x, v\rangle) = \langle x, v\rangle_+^2$ i.e. $\nabla\varphi_0$ is a shallow NN with ReLU activation

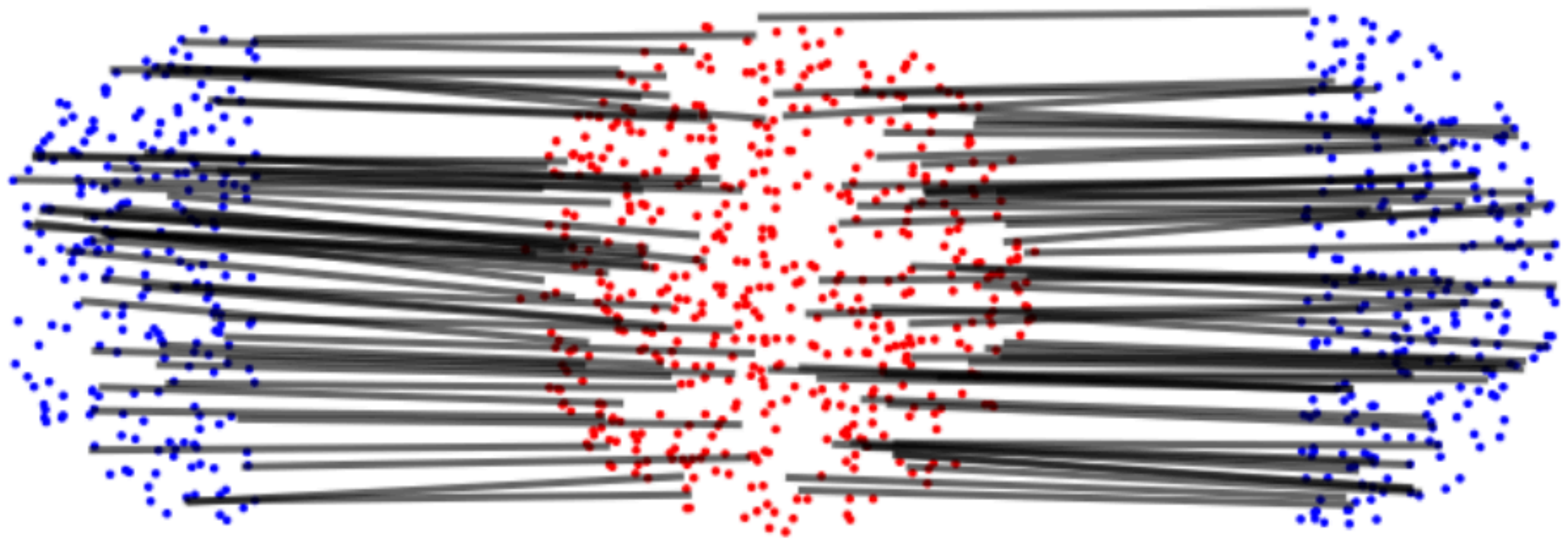
$$\mathbb{E}\|\nabla\hat{\varphi}_{\mathscr{F}_\sigma^1} - \nabla\varphi_0\|_{L^2(P)}^2 \lesssim_{\log(n)} n^{-\frac{1}{2}-\frac{1}{d}}$$

(Can handle more smooth activation functions of this form!)

Future directions:

Hard question: estimation *discontinuous* transport map e.g.

$$\varphi_0(x) = 2\,|\,x_1\,| + \frac{1}{2}\|x\|^2$$



Thanks!

Bibliography

- [HR21] J-C. Hütter, and P. Rigollet. *Minimax rates of estimation for smooth optimal transport maps.* Annals of Statistics
- [DGS21] N. Deb, P. Ghosal, and B. Sen. *Rates of Estimation of Optimal Transport Maps using Plug-in Estimators via Barycentric Projections.* NeurIPS 2021
- [MB+21] T. Manole, S. Balakrishnan, J. Niles-Weed, and L. Wasserman. *Plugin Estimation of Smooth Optimal Transport Maps.* ArXiv 2021
- [Gen19] A. Genevay. *Entropy-regularized optimal transport for machine learning*. PhD Thesis, 2019
- [SDF+18] V. Seguy, B. Damodaran, R. Flamary, N. Courty, A. Rolet, and M. Blondel. *Large-scale optimal transport and mapping estimation.* ICLR 2018
- [Pal19] S. Pal. *On the difference between entropic cost and the optimal transport cost.* ArXiv 2019
- [C13] M. Cuturi. *Sinkhorn distances: Lightspeed computation of optimal transport*. NIPS, 2013
- [CRL+2020] L. Chizat, P. Roussillon, F. Léger, F-X. Vialard, G. Peyré. *Faster Wasserstein Distance Estimation with the Sinkhorn Divergence.* NeurIPS, 2020