Statistical Optimal Transport

Rathindra Nath Karmakar

November 8, 2024

Overview

1. **Chapter 1** - Optimal Transport
2. **Chapter 2** - Estimation of Wasserstein distances
3. **Chapter 3** - Estimation of transport maps
   - Problem formulation
   - Naïve approach and limitations
   - The semidual problem
   - A special case: affine transport maps
   - Stability of the semidual problem
   - Obtaining the slow rate
   - One - shot localization
   - Obtaining the fast rate

Recall

Theorem (Fundamental theorem of OT)

Let $\mu, \nu \in \mathcal{P}_2(\mathbb{R}^d)$, $X \sim \mu$ and assume that μ has a density. TFAE:

1. $\bar{\gamma} \in \Gamma_{\mu,\nu}$ is an optimal coupling for W_2^2.
2. There exists a convex function φ such that $(X, \nabla\varphi(X)) \sim \bar{\gamma}$.
3. (Kantorovich duality)

$$\int \|x-y\|^2 \, d\bar{\gamma}(x,y) = \sup_{f \in L^1(\mu), g \in L^1(\nu)} \left(\int f \, d\mu + \int g \, d\nu \right),$$

where $f(x) + g(y) \leq \|x-y\|^2$.

Recall

Semidual formulation

If $\nabla\varphi$ is the optimal transport map, then φ solves

$$\min_{\phi} \mathcal{S}(\phi) := \int \phi \, d\mu + \int \phi^* \, d\nu$$

Problem Formulation

- **Given:**
  - Samples $X_1, \ldots, X_n \overset{\text{i.i.d.}}{\sim} \mu$
  - Samples $Y_1, \ldots, Y_n \overset{\text{i.i.d.}}{\sim} \nu$
- **Goal:** Estimate the optimal transport map T from μ to ν.
- **Performance Measure:**

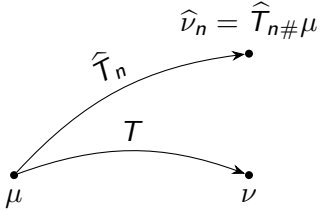
$$\int \|\widehat{T}_n(x) - T(x)\|^2 \, \mu(dx)$$

Relation to Wasserstein Distance

- **Integrated L^2 Error Controls Wasserstein Distance:**

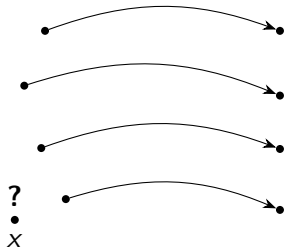
$$W_2^2(\widehat{\nu}_n, \nu) \leq \int \| \widehat{T}_n(x) - T(x) \|^2 \, \mu(dx)$$

- **Illustration:**



Naïve Approach and Limitations

- **Naïve Approach:** Compute optimal coupling between empirical measures μ_n and ν_n.
- **Issue:** Resulting map $\widehat{T}_n$ is only defined on sample points $\{X_i\}$.
- **Challenge:** Extending $\widehat{T}_n$ to all of $\mathbb{R}^d$ is non-trivial.



Interpolate?

Naïve Approach and Limitations - cont.

- **Setting:** Assume $d = 1$. $X_{(1)} < \cdots < X_{(n)}, Y_{(1)} < \cdots < Y_{(n)}$ - samples in sorted order.

Naïve Approach and Limitations - cont.

- **Setting:** Assume $d = 1$. $X_{(1)} < \cdots < X_{(n)}$, $Y_{(1)} < \cdots < Y_{(n)}$ - samples in sorted order.
- **One nearest neighbour estimator:** Given $x \in [0, 1]$, let $X_{(i)}$ denote the largest sample from μ such that $X_{(i)} \leq x$. Set $\widehat{T}_n(x) := Y_{(i)}$ (if no such $X_{(i)}$ exists, then output $\widehat{T}_n(x) := 0$).

Naïve Approach and Limitations - cont.

- **Setting:** Assume $d = 1$. $X_{(1)} < \cdots < X_{(n)}$, $Y_{(1)} < \cdots < Y_{(n)}$ - samples in sorted order.
- **One nearest neighbour estimator:** Given $x \in [0, 1]$, let $X_{(i)}$ denote the largest sample from μ such that $X_{(i)} \leq x$. Set $\widehat{T}_n(x) := Y_{(i)}$ (if no such $X_{(i)}$ exists, then output $\widehat{T}_n(x) := 0$).

(Pointwise) Fast rate for 1NN estimator

Assume μ and ν have densities supported on $[0, 1]$, bounded away from 0 and ∞. For $x \in [0, 1]$,

$$\mathbb{E}|\widehat{T}_n(x) - T(x)|^2 \lesssim \frac{1}{n}$$

where T is the true optimal transport map $F_\nu^\dagger \circ F_\mu$ from μ to ν.

- **Sketch of proof:**
- The argument can be extended to show fast rate for the integrated error $\mathbb{E}\|\widehat{T}_n - T\|^2_{L^2(\mu)}$
- **Higher dimensions:** See **"Plugin Estimation of Smooth Optimal Transport Maps"** by Manole, Balakrishnan, Niles-Weed and Wasserman, 2021. (Slightly different definition using Voronoi cells)

Naïve Approach and Limitations - cont.

Rate of 1NN estimator [**MBNWW'21, Proposition 15**]

Let $\mu, \nu \in \mathcal{P}_{ac}([0,1]^d)$ admit densities p, q (respectively) such that $\gamma^{-1} \leq p \leq \gamma$ over $[0,1]^d$, for some $\gamma > 0$. Also, assume that the optimal transport plan is $\frac{1}{\lambda}-$strongly convex and $\lambda-$smooth then

$$\mathbb{E}\left\| \widehat{T}_n^{1\mathrm{NN}} - T \right\|^2_{L_2^2(P)} \lesssim_\lambda (\log n)^2 \begin{cases} n^{-1}, & d \leq 3 \\ n^{-1}(\log n)^2, & d = 4 \\ n^{-2/d}, & d \geq 5. \end{cases}$$

Naïve Approach and Limitations - cont.

Rate of 1NN estimator [MBNWW'21, Proposition 15]

Let $\mu, \nu \in \mathcal{P}_{ac}([0,1]^d)$ admit densities p, q (respectively) such that $\gamma^{-1} \leq p \leq \gamma$ over $[0,1]^d$, for some $\gamma > 0$. Also, assume that the optimal transport plan is $\frac{1}{\lambda}-$strongly convex and $\lambda-$smooth then

$$\mathbb{E}\left\| \widehat{T}_n^{1\mathrm{NN}} - T \right\|^2_{L_2^2(P)} \lesssim_\lambda (\log n)^2 \begin{cases} n^{-1}, & d \leq 3 \\ n^{-1}(\log n)^2, & d = 4 \\ n^{-2/d}, & d \geq 5. \end{cases}$$

CURSE OF DIMENSIONALITY

Alternative Estimation Strategy

- **Semidual Formulation:**
  - Offers a principled way to estimate T over $\mathbb{R}^d$.
- **Advantages:**
  - Incorporates additional assumptions (e.g., smoothness).
  - Dimension-free fast rates

The Semidual Problem

- **Semidual Functional:**

$$\mathcal{S}(\phi) = \int \phi \, d\mu + \int \phi^* \, d\nu$$

The Semidual Problem

- **Semidual Functional:**

$$\mathcal{S}(\phi) = \int \phi \, d\mu + \int \phi^* \, d\nu$$

- **Optimal Transport Map:**
  - $T = \nabla\varphi$, where φ minimizes $\mathcal{S}(\phi)$ among all $\phi \in \mathcal{F}$, a subclass of convex $\mu-$ae differentiable functions in $L_1(\mu)$.
  - Smoothness assumptions on $\mathcal{F}$ necessary to derive rates.
  - In applications, we generally know $\mathcal{F}$ beforehand.

Estimation via the Semidual

- **Empirical Semidual Functional:**

$$\widehat{\varphi} = \arg\min_{\phi \in \mathcal{F}} \mathcal{S}_n(\phi)$$

where

$$\mathcal{S}_n(\phi) = \int \phi \, d\mu_n + \int \phi^* \, d\nu_n$$

- **Estimator:** $\widehat{T}_n = \nabla \widehat{\varphi}$.

Recall (Gaussian OT)

Problem

Let $m_i \in \mathbb{R}^d$ and Σ_i be positive definite $d \times d$ matrices and let $\mu_i := \mathcal{N}(m_i, \Sigma_i)$, $i = 1, 2$. Assume further that Σ_1 and Σ_2 commute. We consider the (W_2^2)-OT problem from μ_1 to μ_2.

Recall (Gaussian OT)

Problem

Let $m_i \in \mathbb{R}^d$ and Σ_i be positive definite $d \times d$ matrices and let $\mu_i := \mathcal{N}(m_i, \Sigma_i)$, $i = 1, 2$. Assume further that Σ_1 and Σ_2 commute. We consider the (W_2^2)-OT problem from μ_1 to μ_2. In light of the improved Brenier theorem, it is enough to construct a map $T : \mathbb{R}^d \to \mathbb{R}^d$, such that

1. $T_{\#}(\mu_1) = \mu_2$;
2. T is the gradient of a convex function on $\mathbb{R}^d$.

Recall (Gaussian OT - cont.)

The naive choice

$$T(x) = \Sigma_2^{1/2}\Sigma_1^{-1/2}(x - m_1) + m_2.$$

works if Σ_1 and Σ_2 commute. With this choice, we obtain

$$W_2^2(\mu_1, \mu_2) = \|m_1 - m_2\|^2 + \|\Sigma_1^{1/2} - \Sigma_2^{1/2}\|_{\mathrm{HS}}^2.$$

Recall (Gaussian OT - cont.)

The naive choice

$$T(x) = \Sigma_2^{1/2}\Sigma_1^{-1/2}(x - m_1) + m_2.$$

works if Σ_1 and Σ_2 commute. With this choice, we obtain

$$W_2^2(\mu_1, \mu_2) = \|m_1 - m_2\|^2 + \|\Sigma_1^{1/2} - \Sigma_2^{1/2}\|_{\mathrm{HS}}^2.$$

Remark: This only gives an upper bound if Σ_1 and Σ_2 don't commute.

A Special Case: Affine Transport Maps

- **Scenario:**
  - $\mu = \mathcal{N}(0, I)$ is known, and we obtain samples from $\nu = (\nabla\varphi)_{\#}\mu$ for some $\varphi \in \mathcal{F}$, the class of all convex quadratic functions $x \mapsto \frac{1}{2}x^\top Ax + b^\top x$ with $A \succeq 0$.

A Special Case: Affine Transport Maps

- **Scenario:**
  - $\mu = \mathcal{N}(0, I)$ is known, and we obtain samples from $\nu = (\nabla\varphi)_{\#}\mu$ for some $\varphi \in \mathcal{F}$, the class of all convex quadratic functions $x \mapsto \frac{1}{2}x^\top Ax + b^\top x$ with $A \succeq 0$.
- **Estimator:**
  - $\nu = \mathcal{N}(b, A^2)$ and $\nabla\varphi = Ax + b = Cov(\nu)^{1/2}x + Mean(\mu)$
  - Use sample mean $\widehat{m}$ and covariance $\widehat{\Sigma}$.

$$\widehat{T}_n(x) = \widehat{\Sigma}^{1/2}x + \widehat{m}$$

Affine Transport Maps - cont.

Proposition

Let $\mathcal{F}$ be the set of all convex quadratic functions on $\mathbb{R}^d$. Let $\mu = \mathcal{N}(0, I)$, and write ν_n for an empirical measure consisting of i.i.d. samples from a probability measure ν. If

$$\hat{\varphi} = \arg\min_{\varphi \in \mathcal{F}} \left\{ \int \varphi \, d\mu + \int \varphi^* \, d\nu_n \right\},$$

then

$$\nabla \hat{\varphi}(x) = \hat{\Sigma}^{1/2} x + \hat{m},$$

where $\hat{m}$ and $\hat{\Sigma}$ are the mean and covariance of ν_n, respectively.

Sketch of proof

By definition, $\hat{\varphi} = \frac{1}{2} x^\top \hat{A} x + \hat{b}^\top x$, where $(\hat{A}, \hat{b})$ solve

$$\min_{A \succeq 0,\, b \in \mathbb{R}^d} \left[\int \left(\frac{1}{2} x^\top A x + b^\top x \right) \mu(dx) + \int \left(\frac{1}{2} y^\top A y + b^\top y \right)^* \nu_n(dy) \right].$$

Sketch of proof

By definition, $\hat{\varphi} = \frac{1}{2} x^\top \hat{A} x + \hat{b}^\top x$, where $(\hat{A}, \hat{b})$ solve

$$\min_{A \succeq 0,\, b \in \mathbb{R}^d} \left[\int \left(\frac{1}{2} x^\top A x + b^\top x \right) \mu(dx) + \int \left(\frac{1}{2} y^\top A y + b^\top y \right)^* \nu_n(dy) \right].$$

- $\left(\frac{1}{2} y^\top A y + b^\top y \right)^* = \frac{1}{2} (y - b)^\top A^{-1} (y - b)$

Sketch of proof

By definition, $\hat{\varphi} = \frac{1}{2} x^\top \hat{A} x + \hat{b}^\top x$, where $(\hat{A}, \hat{b})$ solve

$$\min_{A \succeq 0,\, b \in \mathbb{R}^d} \left[\int \left(\frac{1}{2} x^\top A x + b^\top x \right) \mu(dx) + \int \left(\frac{1}{2} y^\top A y + b^\top y \right)^* \nu_n(dy) \right].$$

- $\left(\frac{1}{2} y^\top A y + b^\top y \right)^* = \frac{1}{2} (y - b)^\top A^{-1} (y - b)$
- Can replace ν_n by $\mathcal{N}(\hat{m}, \hat{\Sigma})$

Sketch of proof

By definition, $\hat{\varphi} = \frac{1}{2} x^\top \hat{A} x + \hat{b}^\top x$, where $(\hat{A}, \hat{b})$ solve

$$\min_{A \succeq 0,\, b \in \mathbb{R}^d} \left[\int \left(\frac{1}{2} x^\top A x + b^\top x \right) \mu(dx) + \int \left(\frac{1}{2} y^\top A y + b^\top y \right)^* \nu_n(dy) \right].$$

- $\left(\frac{1}{2} y^\top A y + b^\top y\right)^* = \frac{1}{2} (y - b)^\top A^{-1} (y - b)$
- Can replace ν_n by $\mathcal{N}(\hat{m}, \hat{\Sigma})$
- The convex function $\hat{\varphi} = \frac{1}{2} x^\top \hat{\Sigma} x + \hat{m}^\top x$ which solves the **semidual** problem for Gaussian OT lies in $\mathcal{F}$.

Fast rate for Gaussian OT (Sketch)

Let $\hat{A} = \hat{\Sigma}$ and $\hat{b} = \hat{m}$.

- **Simplification:**

$$\begin{aligned}
\mathbb{E}_x \left[\| T(x) - \hat{T}(x) \|^2 \right] &= \mathbb{E}_x \left[\| (A - \hat{A})x + (b - \hat{b}) \|^2 \right] \\
&= \mathbb{E}_x \left[x^\top (A - \hat{A})^\top (A - \hat{A}) x \right] + \| b - \hat{b} \|^2 \\
&= \operatorname{trace} \left((A - \hat{A})^\top (A - \hat{A}) \right) + \| b - \hat{b} \|^2 \\
&= \| A - \hat{A} \|_{\mathrm{HS}}^2 + \| b - \hat{b} \|^2 .
\end{aligned}$$

Fast rate for Gaussian OT (Sketch)

Let $\hat{A} = \hat{\Sigma}$ and $\hat{b} = \hat{m}$.

- **Simplification:**

$$\begin{aligned}
\mathbb{E}_x\left[\|T(x) - \hat{T}(x)\|^2\right] &= \mathbb{E}_x\left[\|(A-\hat{A})x + (b-\hat{b})\|^2\right] \\
&= \mathbb{E}_x\left[x^\top (A-\hat{A})^\top (A-\hat{A})x\right] + \|b-\hat{b}\|^2 \\
&= \operatorname{trace}\left((A-\hat{A})^\top (A-\hat{A})\right) + \|b-\hat{b}\|^2 \\
&= \|A-\hat{A}\|_{\mathrm{HS}}^2 + \|b-\hat{b}\|^2.
\end{aligned}$$

$$\mathbb{E}\left[\mathbb{E}_x\left[\|T(x) - \hat{T}(x)\|^2\right]\right] = \mathbb{E}\left[\|A-\hat{A}\|_{\mathrm{HS}}^2\right] + \mathbb{E}\left[\|b-\hat{b}\|^2\right].$$

Fast Rate for Gaussian OT (Sketch) - cont.

- **Rate for mean:**

$$\hat{b} \sim \mathcal{N}\left(b, \frac{A^2}{n}\right) \implies \mathbb{E}\left[\|b - \hat{b}\|^2\right] = \frac{1}{n}\,\mathrm{trace}(A^2)$$

Fast Rate for Gaussian OT (Sketch) - cont.

- **Rate for mean:**

$$\hat{b} \sim \mathcal{N}\left(b, \frac{A^2}{n}\right) \implies \mathbb{E}\left[\|b - \hat{b}\|^2\right] = \frac{1}{n}\operatorname{trace}(A^2)$$

- **Rate for covariance term:**

$$\mathbb{E}\|A - \hat{A}\|_{\mathrm{HS}}^2 \sim \mathbb{E}\left\|\frac{1}{2}A^{-1}\left(\hat{\Sigma}^{\frac{1}{2}} - \Sigma^{\frac{1}{2}}\right)\right\|_{\mathrm{HS}}^2$$

Fast Rate for Gaussian OT (Sketch) - cont.

- **Rate for mean:**

$$\hat{b} \sim \mathcal{N}\left(b, \frac{A^2}{n}\right) \implies \mathbb{E}\left[\|b - \hat{b}\|^2\right] = \frac{1}{n}\operatorname{trace}(A^2)$$

- **Rate for covariance term:**

$$\mathbb{E}\|A - \hat{A}\|_{\mathrm{HS}}^2 \sim \mathbb{E}\left\|\frac{1}{2}A^{-1}\left(\hat{\Sigma}^{\frac{1}{2}} - \Sigma^{\frac{1}{2}}\right)\right\|_{\mathrm{HS}}^2$$

As $n \times \hat{\Sigma} \sim W_p(A^2, n-1)$, letting $\Sigma = A^2$, and $\Delta = \hat{\Sigma}^{\frac{1}{2}} - \Sigma^{\frac{1}{2}}$ we have

$$\frac{1}{4}\mathbb{E}\operatorname{trace}\left(\Sigma^{-1}\Delta\Delta^\top\right) \sim \frac{1}{4}\operatorname{tr}\left(\Sigma^{-1} \cdot \frac{1}{n-1}\left(\Sigma \circ \Sigma + \Sigma \otimes \Sigma\right)\right)$$

Slow rate for the semidual estimator

Hence, for the Gaussian OT, we get a rate of $\Theta_d\left(\frac{1}{n}\right)$.

Can we generalize this?

Look at **"Concentration bounds for linear Monge mapping estimation and optimal transport domain adaptation"** by Flamary, Lounici and Ferrari, 2020, for extension to subGaussian case and different sample sizes.

Slow rate for the semidual estimator

What about other distributions?

Slow rate for the semidual estimator

What about other distributions?

Theorem 2

Assume $\mathcal{F}$ is "nice". The semidual estimator $\hat{\varphi}$ satisfies the bound

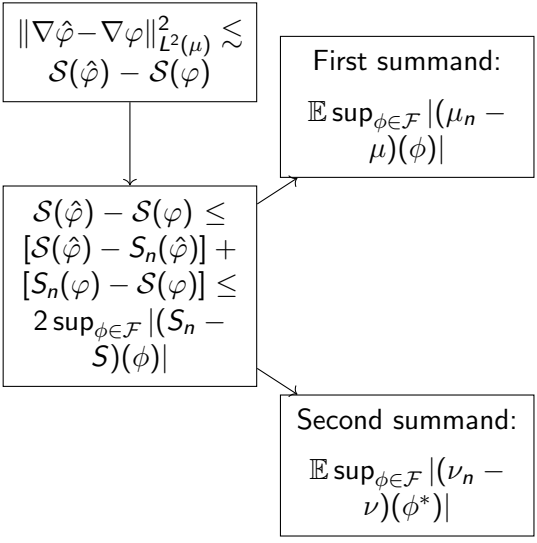
$$\mathbb{E}\|\nabla\hat{\varphi} - \nabla\varphi\|^2_{L_2(\mu)} \lesssim n^{-1/2}.$$

Slow rate for the semidual estimator - cont.

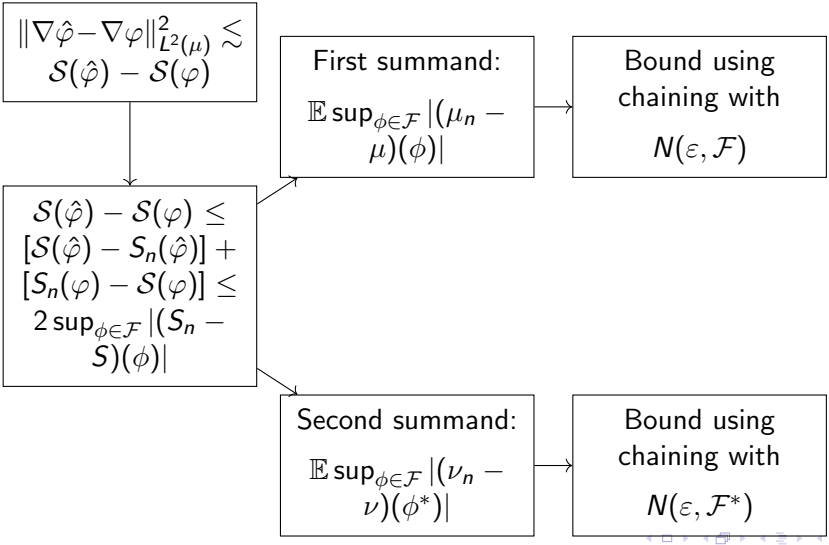
$$\|\nabla\hat{\varphi}-\nabla\varphi\|^2_{L^2(\mu)} \lesssim \mathcal{S}(\hat{\varphi}) - \mathcal{S}(\varphi)$$

$$\mathcal{S}(\hat{\varphi}) - \mathcal{S}(\varphi) \leq [\mathcal{S}(\hat{\varphi}) - S_n(\hat{\varphi})] + [S_n(\varphi) - \mathcal{S}(\varphi)] \leq 2\sup_{\phi\in\mathcal{F}} |(S_n - \mathcal{S})(\phi)|$$

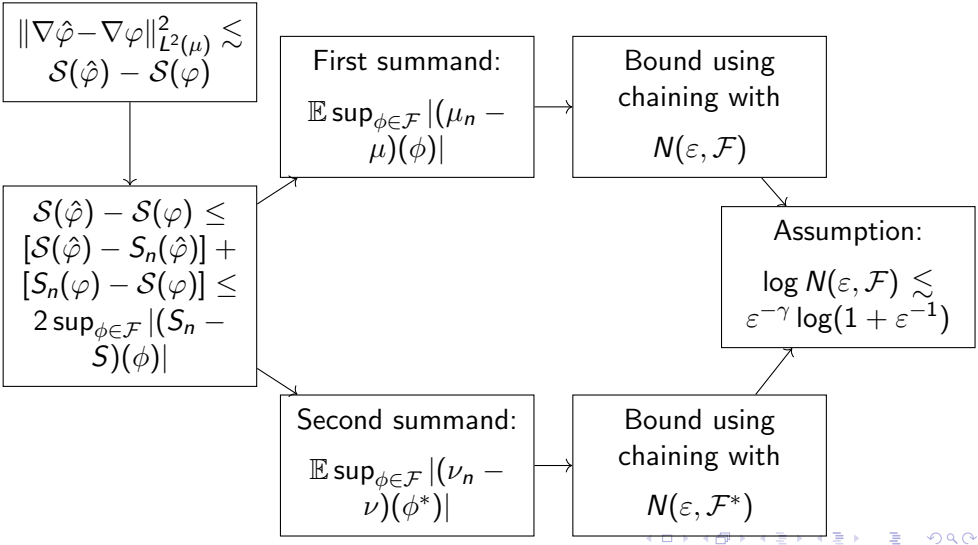
Slow rate for the semidual estimator - cont.



Slow rate for the semidual estimator - cont.



Slow rate for the semidual estimator - cont.



Next time:

Obtaining the slow rate

Recap

- **Problem formulation :** Given samples from μ and $\nu = T_{\#}\mu$, (T - optimal transport with quadratic costs), estimate T.

Recap

- **Problem formulation :** Given samples from μ and $\nu = T_{\#}\mu$, (T - optimal transport with quadratic costs), estimate T.
- **1-NN estimator:** $X_{(i)} \mapsto Y_{(i)}$ and interpolate.

Recap

- **Problem formulation :** Given samples from μ and $\nu = T_{\#}\mu$, (T - optimal transport with quadratic costs), estimate T.
- **1-NN estimator:** $X_{(i)} \mapsto Y_{(i)}$ and interpolate.
- **Limitation:** Curse of dimensionality

Recap

- **Problem formulation :** Given samples from μ and $\nu = T_{\#}\mu$, (T - optimal transport with quadratic costs), estimate T.
- **1-NN estimator:** $X_{(i)} \mapsto Y_{(i)}$ and interpolate.
- **Limitation:** Curse of dimensionality
- **Empirical Semidual Functional:**

$$\widehat{\varphi} = \arg\min_{\phi \in \mathcal{F}} \mathcal{S}_n(\phi)$$

where

$$\mathcal{S}_n(\phi) = \int \phi \, d\mu_n + \int \phi^* \, d\nu_n$$

- **Semidual Estimator:** $\widehat{T}_n = \nabla \widehat{\varphi}$.

Recap

- **Problem formulation :** Given samples from μ and $\nu = T_{\#}\mu$, (T - optimal transport with quadratic costs), estimate T.
- **1-NN estimator:** $X_{(i)} \mapsto Y_{(i)}$ and interpolate.
- **Limitation:** Curse of dimensionality
- **Empirical Semidual Functional:**

$$\widehat{\varphi} = \arg\min_{\phi \in \mathcal{F}} \mathcal{S}_n(\phi)$$

where

$$\mathcal{S}_n(\phi) = \int \phi \, d\mu_n + \int \phi^* \, d\nu_n$$

- **Semidual Estimator:** $\widehat{T}_n = \nabla \widehat{\varphi}$.
- **Example of Gaussian OT:** $\widehat{T}_n = \hat{\Sigma}^{1/2} x + \hat{m}$

Recap

- **Problem formulation :** Given samples from μ and $\nu = T_{\#}\mu$, (T - optimal transport with quadratic costs), estimate T.
- **1-NN estimator:** $X_{(i)} \mapsto Y_{(i)}$ and interpolate.
- **Limitation:** Curse of dimensionality
- **Empirical Semidual Functional:**

$$\widehat{\varphi} = \arg\min_{\phi \in \mathcal{F}} \mathcal{S}_n(\phi)$$

where

$$\mathcal{S}_n(\phi) = \int \phi \, d\mu_n + \int \phi^* \, d\nu_n$$

- **Semidual Estimator:** $\widehat{T}_n = \nabla\widehat{\varphi}$.
- **Example of Gaussian OT:** $\widehat{T}_n = \hat{\Sigma}^{1/2}x + \hat{m}$
- **Slow rate:** $\mathbb{E}\|\nabla\hat{\varphi} - \nabla\varphi\|^2_{L_2(\mu)} \lesssim n^{-1/2}$

Stability of the Semidual Problem

Theorem 1 (Stability):

If ϕ is strongly convex and smooth, satisfying $\frac{1}{2}I \preceq \nabla^2\phi(x) \preceq 2I$ for $\|x\| < 1$, then

$$\frac{1}{4}\|\nabla\phi - \nabla\varphi\|^2_{L^2(\mu)} \leq \mathcal{S}(\phi) - \mathcal{S}(\varphi) \leq \|\nabla\phi - \nabla\varphi\|^2_{L^2(\mu)}.$$

Stability of the Semidual Problem - Proof

We will use, without proof, the following results from convex analysis (vide **Appendix 9** of **Statistical Optimal Transport** by Chewi, Niles-Weed, Rigollet):

- For $\alpha > 0$, ϕ is $\alpha-$strongly convex iff ϕ^* is $\frac{1}{\alpha}-$smooth.
- $\nabla\phi$ is a diffeomorphism with inverse $\nabla\phi^*$
- $\phi(x) + \phi^*(\nabla\phi(x)) = \langle x, \nabla\phi(x)\rangle$

Stability of the Semidual Problem - Proof

We will use, without proof, the following results from convex analysis (vide **Appendix 9** of **Statistical Optimal Transport** by Chewi, Niles-Weed, Rigollet):

- For $\alpha > 0$, ϕ is $\alpha-$strongly convex iff ϕ^* is $\frac{1}{\alpha}-$smooth.
- $\nabla\phi$ is a diffeomorphism with inverse $\nabla\phi^*$
- $\phi(x) + \phi^*(\nabla\phi(x)) = \langle x, \nabla\phi(x)\rangle$

Easy to see for $\phi(x) = \frac{1}{2}\langle x, Ax\rangle + \langle b, x\rangle + c$

Stability of the Semidual Problem - Proof

- **Pushforward:**

$$\begin{aligned}\mathcal{S}(\phi) &= \int \phi(x)\,\mu(dx) + \int \phi^*(y)\,\nu(dy) \\ &= \int \left(\phi(x) + \phi^*(\nabla\varphi(x))\right)\mu(dx).\end{aligned}$$

Stability of the Semidual Problem - Proof

- **Pushforward:**

$$\begin{aligned}\mathcal{S}(\phi) &= \int \phi(x)\,\mu(dx) + \int \phi^*(y)\,\nu(dy) \\ &= \int \left(\phi(x) + \phi^*(\nabla\varphi(x))\right)\mu(dx).\end{aligned}$$

- **Strong convexity of ϕ^*:**

$$\begin{aligned}\phi^*(\nabla\varphi(x)) \geq & \phi^*(\nabla\phi(x)) + \langle \nabla\phi^*(\nabla\phi(x)), \nabla\varphi(x) - \nabla\phi(x)\rangle \\ & + \frac{1}{4}\|\nabla\varphi(x) - \nabla\phi(x)\|^2,\end{aligned}$$

Stability of the Semidual Problem - Proof

- **Pushforward:**

$$\begin{aligned}\mathcal{S}(\phi) &= \int \phi(x)\,\mu(dx) + \int \phi^*(y)\,\nu(dy) \\ &= \int (\phi(x) + \phi^*(\nabla\varphi(x)))\,\mu(dx).\end{aligned}$$

- **Strong convexity of ϕ^*:**

$$\begin{aligned}\phi^*(\nabla\varphi(x)) \geq & \phi^*(\nabla\phi(x)) + \langle \nabla\phi^*(\nabla\phi(x)), \nabla\varphi(x) - \nabla\phi(x)\rangle \\ & + \frac{1}{4}\|\nabla\varphi(x) - \nabla\phi(x)\|^2,\end{aligned}$$

- **Using properties outlined earlier:**

$$\begin{aligned}\phi(x) + \phi^*(\nabla\varphi(x)) &\geq \langle x, \nabla\varphi(x)\rangle + \frac{1}{4}\|\nabla\varphi(x) - \nabla\phi(x)\|^2 \\ &= \varphi(x) + \varphi^*(\nabla\varphi(x)) + \frac{1}{4}\|\nabla\varphi(x) - \nabla\phi(x)\|^2\end{aligned}$$

Stability of the Semidual Problem - Proof

$$\implies \mathcal{S}(\phi) \geq \mathcal{S}(\varphi) + \frac{1}{4}\|\nabla\varphi - \nabla\phi\|^2_{L^2(\mu)}.$$

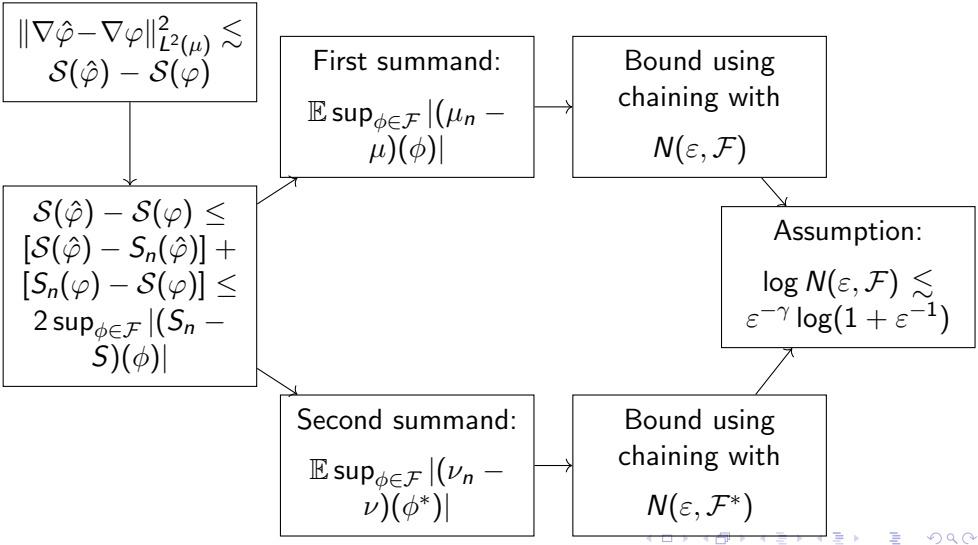
Stability of the Semidual Problem - Proof

$$\implies \mathcal{S}(\phi) \geq \mathcal{S}(\varphi) + \frac{1}{4}\|\nabla\varphi - \nabla\phi\|^2_{L^2(\mu)}.$$

- **Now using smoothness:** Similarly, the other direction follows by upper bounding $S(\phi)$ using the smoothness of ϕ^*

$$\mathcal{S}(\phi) \leq \mathcal{S}(\varphi) + \|\nabla\varphi - \nabla\phi\|^2_{L^2(\mu)}.$$

Obtaining the slow rate - cont.



Obtaining the slow rate - cont.

Chaining bound **[vH14, Theorem 5.31]**

If $\mathcal{F}$ is a set of real-valued functions on Ω such that $\|f\|_{L^\infty(\Omega)} \leq R$ for all $f \in \mathcal{F}$, then

$$\mathbb{E} \sup_{f \in \mathcal{F}} \frac{1}{n} \sum_{i=1}^{n} \{f(X_i) - \mathbb{E} f(X_i)\} \lesssim \inf_{\tau > 0} \left\{ \tau + \frac{1}{\sqrt{n}} \int_{\tau}^{R} \sqrt{\log N(\varepsilon, \mathcal{F})}\, \mathrm{d}\varepsilon \right\}.$$

Obtaining the slow rate - cont.

So far, we have

$$\mathbb{E}\|\nabla\hat{\varphi} - \nabla\varphi\|^2_{L^2(\mu)} \lesssim \mathbb{E}\sup_{\phi\in\mathcal{F}}|(\mu_n - \mu)(\phi)| + \mathbb{E}\sup_{\phi\in\mathcal{F}}|(\nu_n - \nu)(\phi^*)|.$$

- **Assumption:** there exists a positive constant R such that $\|\phi\|_{L^\infty(\Omega)} \leq R$ for all $\phi \in \mathcal{F}$.

Obtaining the slow rate - cont.

So far, we have

$$\mathbb{E}\|\nabla\hat{\varphi} - \nabla\varphi\|^2_{L^2(\mu)} \lesssim \mathbb{E}\sup_{\phi\in\mathcal{F}}|(\mu_n - \mu)(\phi)| + \mathbb{E}\sup_{\phi\in\mathcal{F}}|(\nu_n - \nu)(\phi^*)|.$$

- **Assumption:** there exists a positive constant R such that $\|\phi\|_{L^\infty(\Omega)} \leq R$ for all $\phi \in \mathcal{F}$.
- **Chaining bound:** Putting $\tau = 0$ yields

$$\mathbb{E}\sup_{\phi\in\mathcal{F}}|(\mu_n - \mu)(\phi)| \lesssim \frac{1}{\sqrt{n}}\int_0^R \sqrt{\log N(\varepsilon, \mathcal{F})}\,\mathrm{d}\varepsilon.$$

Obtaining the slow rate - cont.

Assumption for Covering Numbers

For simplicity, we focus on the case where the class of functions is small enough that the covering numbers satisfy

$$\log N(\varepsilon, \mathcal{F}) \lesssim \varepsilon^{-\gamma} \log(1 + \varepsilon^{-1}), \quad \gamma \in [0, 1) \tag{3.5}$$

Obtaining the slow rate - cont.

Example (Lipschitz in the parameter)

Assume that $\mathcal{F} = \{\phi_\theta\}_{\theta \in \Theta}$, where $\Theta \subset \mathbb{R}^M$ is bounded, and the potentials satisfy $\|\phi_\theta - \phi_{\theta'}\|_{L^\infty(\Omega)} \lesssim \|\theta - \theta'\|$. Then there exists a positive constant C such that the covering numbers of $\mathcal{F}$ satisfy $\log \mathcal{N}(\varepsilon, \mathcal{F}) = 0$ if $\varepsilon \geq C$ and

$$\log \mathcal{N}(\varepsilon, \mathcal{F}) \lesssim \log(1 + \varepsilon^{-1})$$

otherwise.

Obtaining the slow rate - cont.

Example (Lipschitz in the parameter)

Assume that $\mathcal{F} = \{\phi_\theta\}_{\theta \in \Theta}$, where $\Theta \subset \mathbb{R}^M$ is bounded, and the potentials satisfy $\|\phi_\theta - \phi_{\theta'}\|_{L^\infty(\Omega)} \lesssim \|\theta - \theta'\|$. Then there exists a positive constant C such that the covering numbers of $\mathcal{F}$ satisfy $\log \mathcal{N}(\varepsilon, \mathcal{F}) = 0$ if $\varepsilon \geq C$ and

$$\log \mathcal{N}(\varepsilon, \mathcal{F}) \lesssim \log(1 + \varepsilon^{-1})$$

otherwise.

For instance, convex quadratic functions are Lipschitz in the parameter (A, b), provided we restrict to $\Omega = B_1(0)$

Obtaining the slow rate - cont.

Sketch of Proof:

- Say, $\mathcal{F} \subseteq B_R(0)$ for some $R > 0$.
- For any $\delta > 0$, there exists $\theta_1, \ldots, \theta_N$ with $N \leq (1 + 2R\delta^{-1})^M$ such that $\bigcup_{i=1}^{N} B_\delta(\theta_i) \supset \Theta$.

  (For instance, take a δ packing of Θ. The balls $B_{\frac{\delta}{2}}(\theta_i)$ then are all disjoint and lie inside $B_{R+\frac{\delta}{2}}(0)$. This bounds N by volumetric computation.)

Obtaining the slow rate - cont.

Sketch of Proof:

- Say, $\mathcal{F} \subseteq B_R(0)$ for some $R > 0$.
- For any $\delta > 0$, there exists $\theta_1, \ldots, \theta_N$ with $N \leq (1 + 2R\delta^{-1})^M$ such that $\bigcup_{i=1}^{N} B_\delta(\theta_i) \supset \Theta$.

  (For instance, take a δ packing of Θ. The balls $B_{\frac{\delta}{2}}(\theta_i)$ then are all disjoint and lie inside $B_{R+\frac{\delta}{2}}(0)$. This bounds N by volumetric computation.)
- $\phi_{\theta_1}, \ldots, \phi_{\theta_N}$ is an $O(\delta)$-cover of $\mathcal{F}$.

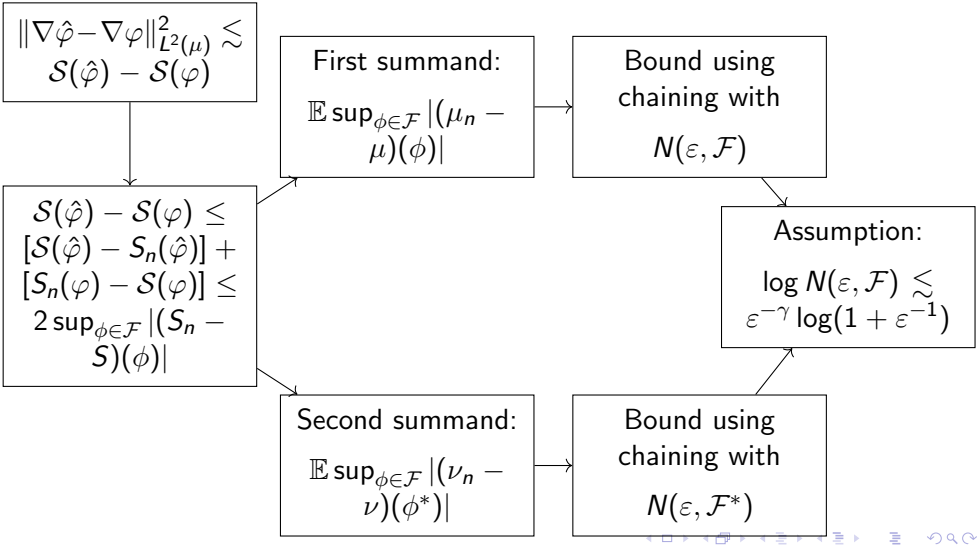
Obtaining the slow rate - cont.

Sketch of Proof:

- Say, $\mathcal{F} \subseteq B_R(0)$ for some $R > 0$.
- For any $\delta > 0$, there exists $\theta_1, \ldots, \theta_N$ with $N \leq (1 + 2R\delta^{-1})^M$ such that $\bigcup_{i=1}^{N} B_\delta(\theta_i) \supset \Theta$.

  (For instance, take a δ packing of Θ. The balls $B_{\frac{\delta}{2}}(\theta_i)$ then are all disjoint and lie inside $B_{R+\frac{\delta}{2}}(0)$. This bounds N by volumetric computation.)
- $\phi_{\theta_1}, \ldots, \phi_{\theta_N}$ is an $O(\delta)$-cover of $\mathcal{F}$.
- Taking $\delta = c\varepsilon$ for a sufficiently small positive constant c yields the claim.

Obtaining the slow rate - cont.



Obtaining the slow rate - cont.

- So far, we have

$$\mathbb{E} \sup_{\phi \in \mathcal{F}} |(\mu_n - \mu)(\phi)| \lesssim \frac{1}{\sqrt{n}} \int_0^R \sqrt{\log N(\varepsilon, \mathcal{F})}\, \mathrm{d}\varepsilon.$$

Obtaining the slow rate - cont.

- So far, we have

$$\mathbb{E} \sup_{\phi \in \mathcal{F}} |(\mu_n - \mu)(\phi)| \lesssim \frac{1}{\sqrt{n}} \int_0^R \sqrt{\log N(\varepsilon, \mathcal{F})}\, \mathrm{d}\varepsilon.$$

- **Bounding the Covering Number:**

$$\frac{1}{\sqrt{n}} \int_0^R \sqrt{\log N(\varepsilon, \mathcal{F})}\, \mathrm{d}\varepsilon \lesssim \frac{1}{\sqrt{n}} \int_0^R \varepsilon^{-\gamma/2} \sqrt{\log(1 + \varepsilon^{-1})}\, \mathrm{d}\varepsilon \lesssim n^{-1/2},$$

- **Conjugation is a contraction:**

$$|\phi^*(y) - \psi^*(y)| = \left| \sup_{x \in \Omega} \{\langle x, y \rangle - \phi(x)\} - \sup_{x' \in \Omega} \{\langle x', y \rangle - \psi(x')\} \right|$$

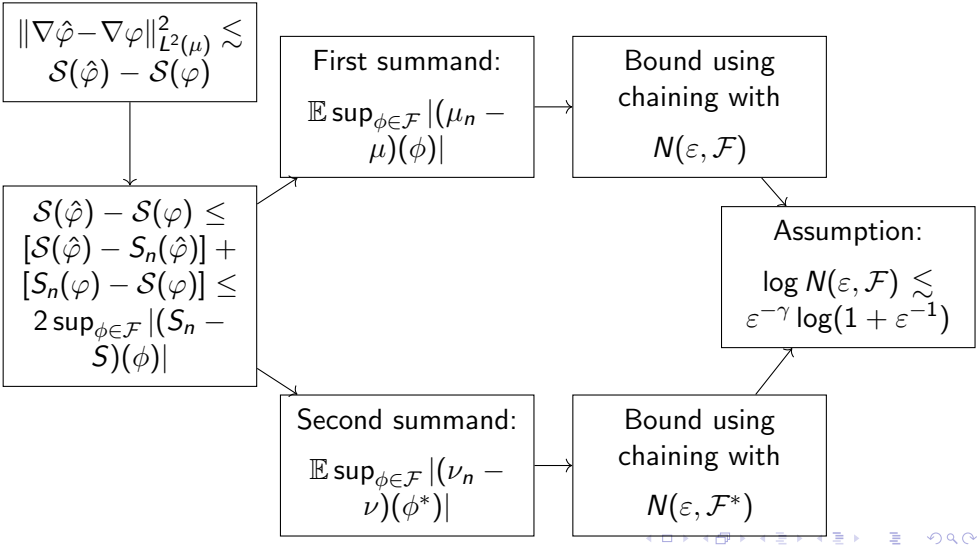
$$\leq \sup_{x \in \Omega} |\phi(x) - \psi(x)| = \|\phi - \psi\|_{L^\infty(\Omega)}.$$

- For any $\varepsilon > 0$,

$$N(\varepsilon, \mathcal{F}^*) \leq N(\varepsilon, \mathcal{F}).$$

- Bound as before

Obtaining the slow rate - cont.



Obtaining the slow rate - cont.

Theorem 2 (Slow rate for semi-dual estimator)

Adopt the assumption for covering numbers. Further, assume that

- The supports of μ and ν lie in $\Omega = B_1(0)$.
- The set $\mathcal{F}$ is bounded in L^∞ on Ω, i.e., $\sup_{\phi \in \mathcal{F}} \|\phi\|_{L^\infty(\Omega)} \leq R$ for some $R > 0$.
- The potentials satisfy $\phi(0) = 0$ and $\phi(x) = +\infty$ if $x \notin \Omega$.
- The potentials are lower-semicontinuous, smooth, and strongly convex on Ω: $\frac{1}{2} I \preceq \nabla^2 \phi(x) \preceq 2I$ if $\|x\| < 1$.

Then the semidual estimator $\hat{\varphi}$ satisfies the bound

$$\mathbb{E}\|\nabla \hat{\varphi} - \nabla \varphi\|^2_{L_2(\mu)} \lesssim_{R,\gamma} n^{-1/2}$$

One-shot Localization

The Localization Argument

Define

$$\varphi_\varepsilon = (1-\lambda)\varphi + \lambda\hat{\varphi}, \quad \lambda = \frac{\varepsilon}{\varepsilon + \|\nabla\hat{\varphi} - \nabla\varphi\|_{L^2(\mu)}}.$$

Then,

$$\|\nabla\varphi_\varepsilon - \nabla\varphi\|_{L^2(\mu)} = \lambda\|\nabla\hat{\varphi} - \nabla\varphi\|_{L^2(\mu)} = \varepsilon\left(\frac{\|\nabla\hat{\varphi} - \nabla\varphi\|_{L^2(\mu)}}{\varepsilon + \|\nabla\hat{\varphi} - \nabla\varphi\|_{L^2(\mu)}}\right) \leq \varepsilon.$$

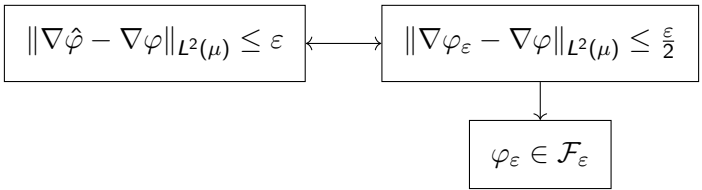
The Localization Argument

Define

$$\varphi_\varepsilon = (1-\lambda)\varphi + \lambda\hat{\varphi}, \quad \lambda = \frac{\varepsilon}{\varepsilon + \|\nabla\hat{\varphi} - \nabla\varphi\|_{L^2(\mu)}}.$$

Then,

$$\|\nabla\varphi_\varepsilon - \nabla\varphi\|_{L^2(\mu)} = \lambda\|\nabla\hat{\varphi} - \nabla\varphi\|_{L^2(\mu)} = \varepsilon\left(\frac{\|\nabla\hat{\varphi} - \nabla\varphi\|_{L^2(\mu)}}{\varepsilon + \|\nabla\hat{\varphi} - \nabla\varphi\|_{L^2(\mu)}}\right) \leq \varepsilon.$$



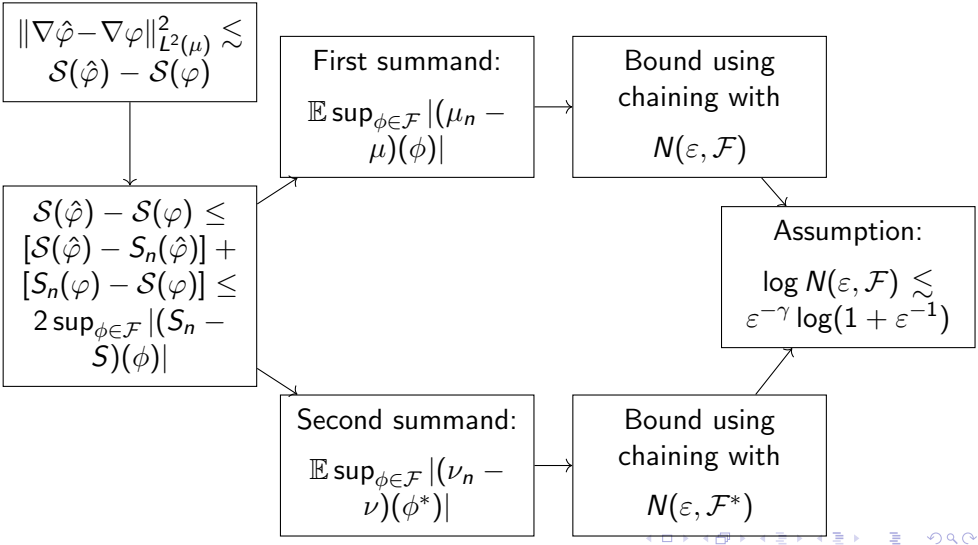
Fast rate for the semidual estimator

Theorem 2

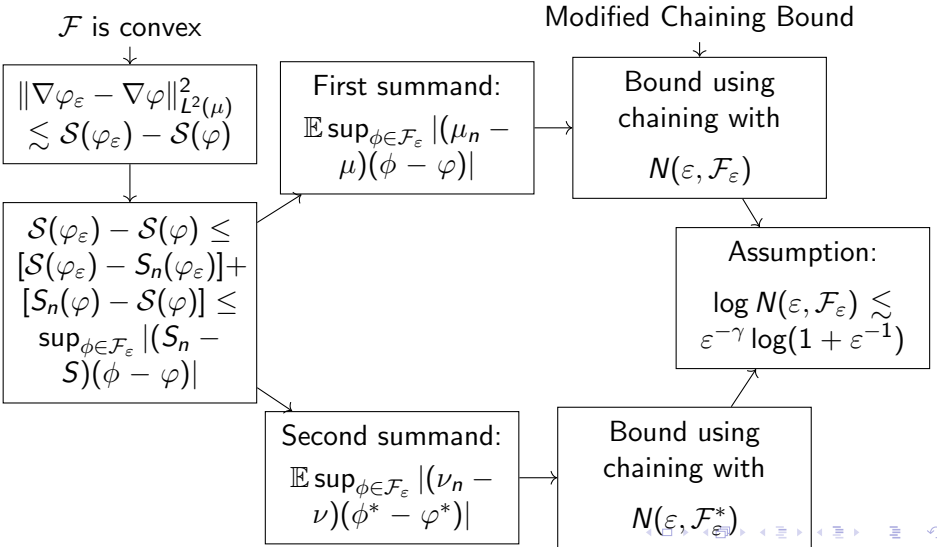
Assume $\mathcal{F}$ and μ are "nice". The semidual estimator $\hat{\varphi}$ satisfies the bound

$$\mathbb{E}\|\nabla\hat{\varphi} - \nabla\varphi\|^2_{L_2(\mu)} \lesssim \left(\frac{\log n}{n}\right)^{\frac{2}{2+\gamma}}.$$

Looking back : Obtaining the slow rate



Obtaining the fast rate - cont.



Obtaining the fast rate

Next time:

Modified Chaining Bound

Questions from last day

1. Motivating the 1-NN construction
2. Proof of pointwise fast rate
3. The restricted class $\mathcal{F}$
4. The γ in covering number bounds
5. Adaptive estimators

Motivating 1-NN

- **Setting:** Optimal Transport (OT) between counting measures on $\{x_1, x_2, x_3\}$ and $\{y_1, y_2, y_3\}$, where $x_1 < x_2 < x_3$ and $y_1 < y_2 < y_3$.

Motivating 1-NN

- **Setting:** Optimal Transport (OT) between counting measures on $\{x_1, x_2, x_3\}$ and $\{y_1, y_2, y_3\}$, where $x_1 < x_2 < x_3$ and $y_1 < y_2 < y_3$.
- **Equivalent Question:**

$$\min_{\sigma} \sum_i (x_i - y_{\sigma(i)})^2 = \sum_i x_i^2 + \sum_i y_i^2 - 2 \max_{\sigma} \sum_i x_i y_{\sigma(i)}$$

Motivating 1-NN

- **Setting:** Optimal Transport (OT) between counting measures on $\{x_1, x_2, x_3\}$ and $\{y_1, y_2, y_3\}$, where $x_1 < x_2 < x_3$ and $y_1 < y_2 < y_3$.
- **Equivalent Question:**

$$\min_{\sigma} \sum_i (x_i - y_{\sigma(i)})^2 = \sum_i x_i^2 + \sum_i y_i^2 - 2 \max_{\sigma} \sum_i x_i y_{\sigma(i)}$$

- **Local changes:** (Cyclical monotonicity type argument)
  $x_1y_1 + x_2y_2 + x_3y_3 > x_1y_2 + x_2y_1 + x_3y_3 > x_1y_3 + x_2y_1 + x_3y_2$

Motivating 1-NN

- **Setting:** Optimal Transport (OT) between counting measures on $\{x_1, x_2, x_3\}$ and $\{y_1, y_2, y_3\}$, where $x_1 < x_2 < x_3$ and $y_1 < y_2 < y_3$.
- **Equivalent Question:**

$$\min_{\sigma} \sum_i (x_i - y_{\sigma(i)})^2 = \sum_i x_i^2 + \sum_i y_i^2 - 2 \max_{\sigma} \sum_i x_i y_{\sigma(i)}$$

- **Local changes:** (Cyclical monotonicity type argument)
  $x_1y_1 + x_2y_2 + x_3y_3 > x_1y_2 + x_2y_1 + x_3y_3 > x_1y_3 + x_2y_1 + x_3y_2$
  $\sigma = Id$ is the optimum.

Motivating 1-NN - cont.

The Rearrangement Inequality

Given $x_1 < x_2 \ldots < x_n$ and $y_1 < y_2 \ldots < y_n$, we have

$$\sum x_i y_i > \sum x_i y_{\sigma(i)} > \sum x_i y_{n+1-i}$$

Motivating 1-NN - cont.

The Rearrangement Inequality

Given $x_1 < x_2 \ldots < x_n$ and $y_1 < y_2 \ldots < y_n$, we have

$$\sum x_i y_i > \sum x_i y_{\sigma(i)} > \sum x_i y_{n+1-i}$$

OT for the counting measure

Given $x_1 < x_2 \ldots < x_n$ and $y_1 < y_2 \ldots < y_n$, the optimal transport map between the counting measures on them is the **Id (minimizes the quadratic cost)** and the most suboptimal transport map is $\mathbf{i} \mapsto \mathbf{n} + \mathbf{1} - \mathbf{i}$ **(maximizes the quadratic cost)**.

Motivating the 1-NN - cont.

- **1-NN estimator:** $X_{(i)} \mapsto Y_{(i)}$

Motivating the 1-NN - cont.

- **1-NN estimator:** $X_{(i)} \mapsto Y_{(i)}$
- **Discrete Brenier:** Note that this map is an increasing function. (One dimensional Brenier: the optimal transport map is the derivative of a convex function)

Motivating the 1-NN - cont.

- **1-NN estimator:** $X_{(i)} \mapsto Y_{(i)}$
- **Discrete Brenier:** Note that this map is an increasing function. (One dimensional Brenier: the optimal transport map is the derivative of a convex function)
- **Special cases:** In particular if $Y_i = f(X_i)$, where f is increasing, then f itself is the optimal transport map.

Motivating the 1-NN - cont.

- **1-NN estimator:** $X_{(i)} \mapsto Y_{(i)}$
- **Discrete Brenier:** Note that this map is an increasing function. (One dimensional Brenier: the optimal transport map is the derivative of a convex function)
- **Special cases:** In particular if $Y_i = f(X_i)$, where f is increasing, then f itself is the optimal transport map.

  But if f is decreasing, it is the most suboptimal transport map.

Motivating the 1-NN - cont.

- **1-NN estimator:** $X_{(i)} \mapsto Y_{(i)}$
- **Discrete Brenier:** Note that this map is an increasing function. (One dimensional Brenier: the optimal transport map is the derivative of a convex function)
- **Special cases:** In particular if $Y_i = f(X_i)$, where f is increasing, then f itself is the optimal transport map.

  But if f is decreasing, it is the most suboptimal transport map.

  So, in the case of $Y = -X$, the 1-NN estimator is not trying to estimate $-Id$.

Proof of the pointwise fast rate

(Pointwise) Fast rate for 1NN estimator

Assume μ and ν have densities supported on $[0,1]$, bounded away from 0 and ∞. For $x \in [0,1]$,

$$\mathbb{E}|\widehat{T}_n(x) - T(x)|^2 \lesssim \frac{1}{n}$$

where T is the true optimal transport map $F_\nu^\dagger \circ F_\mu$ from μ to ν.

Proof of the pointwise fast rate - cont.

- Let $N_x := \sum_{i=1}^n \mathbf{1}\{X_i \le x\}$ and $N'_y := \sum_{i=1}^n \mathbf{1}\{Y_i \le y\}$. Argue that

$$\mathbb{E}\|\hat{T}(x) - T(x)\|^2 \le \int_0^A \mathbb{P}\left(N_x \ge N'_{T(x)+t}\right) 2t dt + \int_0^A \mathbb{P}\left(N_x \le N'_{T(x)-t}\right) 2t dt$$

Proof of the pointwise fast rate - cont.

- Let $N_x := \sum_{i=1}^n \mathbf{1}\{X_i \le x\}$ and $N'_y := \sum_{i=1}^n \mathbf{1}\{Y_i \le y\}$. Argue that

$$\mathbb{E}\|\hat{T}(x) - T(x)\|^2 \le \int_0^A \mathbb{P}\left(N_x \ge N'_{T(x)+t}\right) 2t dt + \int_0^A \mathbb{P}\left(N_x \le N'_{T(x)-t}\right) 2t dt$$

- The **limiting event** $\{F_X(x) \ge F_Y(T(x)+t)\}$ is contained in $\{F_Y(T(x)) \ge F_Y(T(x)+t)\}$, which is empty if the density of ν is bounded away from 0.

Proof of the pointwise fast rate - cont.

- Let $N_x := \sum_{i=1}^n \mathbf{1}\{X_i \le x\}$ and $N'_y := \sum_{i=1}^n \mathbf{1}\{Y_i \le y\}$. Argue that

$$\mathbb{E}\|\hat{T}(x) - T(x)\|^2 \le \int_0^A \mathbb{P}\left(N_x \ge N'_{T(x)+t}\right) 2t dt + \int_0^A \mathbb{P}\left(N_x \le N'_{T(x)-t}\right) 2t dt$$

- The **limiting event** $\{F_X(x) \ge F_Y(T(x)+t)\}$ is contained in $\{F_Y(T(x)) \ge F_Y(T(x)+t)\}$, which is empty if the density of ν is bounded away from 0.
  We used: $\mathbb{P}(X \le x) \le \mathbb{P}(T(X) \le T(x))$, as T must be increasing.

Proof of the pointwise fast rate - cont.

- Let $N_x := \sum_{i=1}^n \mathbf{1}\{X_i \le x\}$ and $N'_y := \sum_{i=1}^n \mathbf{1}\{Y_i \le y\}$. Argue that

$$\mathbb{E}\|\hat{T}(x) - T(x)\|^2 \le \int_0^A \mathbb{P}\left(N_x \ge N'_{T(x)+t}\right) 2t dt + \int_0^A \mathbb{P}\left(N_x \le N'_{T(x)-t}\right) 2t dt$$

- The **limiting event** $\{F_X(x) \ge F_Y(T(x)+t)\}$ is contained in $\{F_Y(T(x)) \ge F_Y(T(x)+t)\}$, which is empty if the density of ν is bounded away from 0.
  We used: $\mathbb{P}(X \le x) \le \mathbb{P}(T(X) \le T(x))$, as T must be increasing.
- Using the Dvoretzky–Kiefer–Wolfowitz inequality, argue that the finite n probability and the limiting probability are close.

The restricted class $\mathcal{F}$

- We got the slow rate $\frac{1}{\sqrt{n}}$ and will derive the fast rate $\left(\frac{\log n}{n}\right)^{\frac{2}{2+\gamma}}$ under very general regularity conditions on $\mathcal{F}$ (lower-semicontinuity, smoothness, strong convexity, a.s. boundedness).
  Multivariate normal achieves $\frac{1}{n}$ rate.

The restricted class $\mathcal{F}$

- We got the slow rate $\frac{1}{\sqrt{n}}$ and will derive the fast rate $\left(\frac{\log n}{n}\right)^{\frac{2}{2+\gamma}}$ under very general regularity conditions on $\mathcal{F}$ (lower-semicontinuity, smoothness, strong convexity, a.s. boundedness).
  Multivariate normal achieves $\frac{1}{n}$ rate.
- Given a "smaller" class $\mathcal{F}$, faster rates can be derived.

The restricted class $\mathcal{F}$

- We got the slow rate $\frac{1}{\sqrt{n}}$ and will derive the fast rate $\left(\frac{\log n}{n}\right)^{\frac{2}{2+\gamma}}$ under very general regularity conditions on $\mathcal{F}$ (lower-semicontinuity, smoothness, strong convexity, a.s. boundedness).
  Multivariate normal achieves $\frac{1}{n}$ rate.
- Given a "smaller" class $\mathcal{F}$, faster rates can be derived.
- **Situation 1:** We know the functional form of the measures μ and ν are able to derive the functional form of $\varphi \in \mathcal{F}$.

The restricted class $\mathcal{F}$

- We got the slow rate $\frac{1}{\sqrt{n}}$ and will derive the fast rate $\left(\frac{\log n}{n}\right)^{\frac{2}{2+\gamma}}$ under very general regularity conditions on $\mathcal{F}$ (lower-semicontinuity, smoothness, strong convexity, a.s. boundedness).
  Multivariate normal achieves $\frac{1}{n}$ rate.
- Given a "smaller" class $\mathcal{F}$, faster rates can be derived.
- **Situation 1:** We know the functional form of the measures μ and ν are able to derive the functional form of $\varphi \in \mathcal{F}$.

  Look at **"Optimal transport map estimation in general function spaces"** by Divol, Niles-Weed, Pooladian, 2024.

The restricted class $\mathcal{F}$

- **Situation 2:** We don't know the functional form and would like to approximate it (say, using neural networks).

The restricted class $\mathcal{F}$

- **Situation 2:** We don't know the functional form and would like to approximate it (say, using neural networks).

Approx. by large parametric classes **[DNWP24, Theorem 3]**

Consider μ and $\mathcal{F}$ that satisfy regularity assumptions as before. Assume that φ is both "smooth" and "strongly convex". Let $(\mathcal{F}_j)_j$ be an approximating family of the class of potentials that is "smooth". Then, for J such that $2^J \asymp n^{1/(2s+d-4)}$, it holds that

$$\mathbb{E}\|\nabla\hat{\varphi}_{\mathcal{F}_J} - \nabla\varphi\|^2_{L^2(\mu)} \lesssim_{\log n} n^{-\frac{2(s-1)}{2s+d-4}}.$$

Here, d is approximately linear in the metric entropy of $\mathcal{F}$ and s is approximately the infimum of the quantity $\frac{-1}{j}\log_2\|\nabla\varphi - \nabla\varphi_j\|$

The γ in covering number bounds

- **In the book:** $\gamma \in [0, 1)$
- Let $u = \log(x^{-1}) = -\log x$. Then:

$$
\begin{aligned}
I &= \int_0^A x^{-\frac{\gamma}{2}} \sqrt{\log(1 + x^{-1})}\, dx \\
&= \int_{\log(A^{-1})}^{\infty} e^{\frac{\gamma}{2} u} \sqrt{\log(1 + e^{u})} \cdot e^{-u}\, du \\
&= \int_{\log(A^{-1})}^{\infty} e^{-\left(1 - \frac{\gamma}{2}\right) u} \sqrt{\log(1 + e^{u})}\, du.
\end{aligned}
$$

  Diverges for $\gamma \geq 2$

- **"Large" function classes:** Lipschitz functions in dimensions ≥ 2

Adaptive Estimators

- **Semidual estimator is not adaptive:** The rates in our estimator depend on nuisance parameters like R and γ of the function class $\mathcal{F}$ itself.

Adaptive Estimators

- **Semidual estimator is not adaptive:** The rates in our estimator depend on nuisance parameters like R and γ of the function class $\mathcal{F}$ itself.
- **An adaptive estimator:** Wait for chapter 4, and then read:

  "Entropic estimation of optimal transport map" by Pooladian, Niles-Weed, 2024.

Recap

- **Problem formulation :** Given samples from μ and $\nu = T_{\#}\mu$, (T - optimal transport with quadratic costs), estimate T.

Recap

- **Problem formulation :** Given samples from μ and $\nu = T_{\#}\mu$, (T - optimal transport with quadratic costs), estimate T.
- **Empirical Semidual Formulation:**

$$\widehat{\varphi} = \arg\min_{\phi \in \mathcal{F}} \int \phi \, d\mu_n + \int \phi^* \, d\nu_n$$

- **Semidual Estimator:** $\widehat{T}_n = \nabla \widehat{\varphi}$.

Recap

- **Problem formulation :** Given samples from μ and $\nu = T_{\#}\mu$, (T - optimal transport with quadratic costs), estimate T.
- **Empirical Semidual Formulation:**

$$\widehat{\varphi} = \arg\min_{\phi \in \mathcal{F}} \int \phi \, d\mu_n + \int \phi^* \, d\nu_n$$

- **Semidual Estimator:** $\widehat{T}_n = \nabla\widehat{\varphi}$.
- **Slow rate:** $\mathbb{E}\|\nabla\hat{\varphi} - \nabla\varphi\|^2_{L_2(\mu)} \lesssim n^{-1/2}$

Recap

- **Problem formulation :** Given samples from μ and $\nu = T_{\#}\mu$, (T - optimal transport with quadratic costs), estimate T.
- **Empirical Semidual Formulation:**

$$\widehat{\varphi} = \arg\min_{\phi \in \mathcal{F}} \int \phi \, d\mu_n + \int \phi^* \, d\nu_n$$

- **Semidual Estimator:** $\widehat{T}_n = \nabla \widehat{\varphi}$.
- **Slow rate:** $\mathbb{E}\|\nabla \hat{\varphi} - \nabla \varphi\|^2_{L_2(\mu)} \lesssim n^{-1/2}$
- **Issue:** Minimizing over the entire class $\mathcal{F}$.

  A priori no control over $\|\nabla \phi - \nabla \varphi\|_{L^2(\mu)}$ for $\phi \in \mathcal{F}$

Recap - cont.

- **Localised class:** $\mathcal{F}_\varepsilon = \{\phi \in \mathcal{F} \mid \|\nabla\phi - \nabla\varphi\|_{L^2(\mu)} \leq \varepsilon\}$

Recap - cont.

- **Localised class:** $\mathcal{F}_\varepsilon = \{\phi \in \mathcal{F} \mid \|\nabla\phi - \nabla\varphi\|_{L^2(\mu)} \leq \varepsilon\}$
- **Localised estimator:**

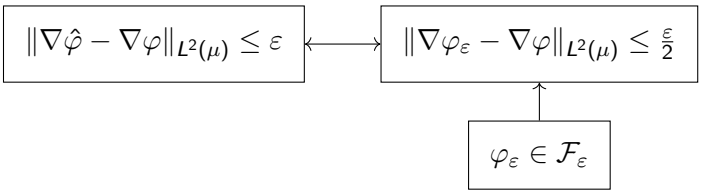
$$\varphi_\varepsilon = (1-\lambda)\varphi + \lambda\hat{\varphi}, \quad \lambda = \frac{\varepsilon}{\varepsilon + \|\nabla\hat{\varphi} - \nabla\varphi\|_{L^2(\mu)}}.$$

Recap - cont.

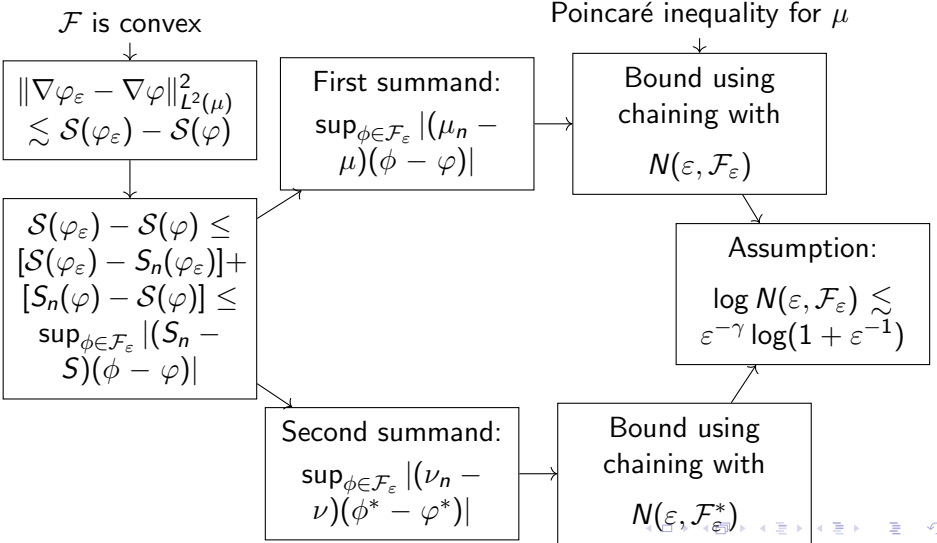
- **Localised class:** $\mathcal{F}_\varepsilon = \{\phi \in \mathcal{F} \mid \|\nabla\phi - \nabla\varphi\|_{L^2(\mu)} \leq \varepsilon\}$
- **Localised estimator:**

$$\varphi_\varepsilon = (1-\lambda)\varphi + \lambda\hat{\varphi}, \quad \lambda = \frac{\varepsilon}{\varepsilon + \|\nabla\hat{\varphi} - \nabla\varphi\|_{L^2(\mu)}}.$$

- **Justification:**



Obtaining the fast rate - cont.



Obtaining the fast rate - cont.

Modified Chaining Bound **[vdVW23, Theorem 2.14.21]**

Let P be a probability measure on a set $\Omega \subseteq \mathbb{R}^d$. Let $X_1, \ldots, X_n \overset{\text{i.i.d.}}{\sim} P$. If $\mathcal{F}$ is a set of real-valued functions such that $\|f\|_{L^2(P)} \leq \sigma$ and $\|f\|_{L^\infty(\Omega)} \leq R$ for all $f \in \mathcal{F}$, then

$$\mathbb{E} \sup_{f \in \mathcal{F}} \frac{1}{n} \sum_{i=1}^{n} \{f(X_i) - \mathbb{E}f(X_i)\} \lesssim \frac{1}{\sqrt{n}} \int_0^\sigma \sqrt{\log N(\epsilon, \mathcal{F})}\, d\epsilon + \frac{1}{n} \int_0^R \log N(\epsilon, \mathcal{F})\, d\epsilon.$$

Obtaining the fast rate - cont.

Poincare inequality $\implies$ L^2 bound

In addition to previous assumptions, suppose μ satisfies a Poincaré inequality. Then

$$\|\phi - \varphi - \mu(\phi - \varphi)\|_{L^2(\mu)} \lesssim \varepsilon$$

$$\|\phi^* - \varphi^* - \nu(\phi^* - \varphi^*)\|_{L^2(\nu)} \lesssim \varepsilon$$

for all $\phi \in \mathcal{F}_\varepsilon$.

Obtaining the fast rate - cont.

- **Poincaré Inequality $\implies$ L^2 Bound for μ:**

$$\|\phi - \varphi - \mu(\phi - \varphi)\|^2_{L^2(\mu)} \leq C\|\nabla(\phi - \varphi)\|^2_{L^2(\mu)} \lesssim \varepsilon^2.$$

Obtaining the fast rate - cont.

- **Poincaré Inequality $\implies$ L^2 Bound for μ:**

$$\|\phi - \varphi - \mu(\phi - \varphi)\|^2_{L^2(\mu)} \leq C\|\nabla(\phi - \varphi)\|^2_{L^2(\mu)} \lesssim \varepsilon^2.$$

- **Poincaré Inequality $\implies$ L^2 Bound for ν:**

$$\|\phi^* - \varphi^* - \nu(\phi^* - \varphi^*)\|^2_{L^2(\nu)} \leq C\|\nabla(\phi^* - \varphi^*)\|^2_{L^2(\nu)} \lesssim \varepsilon^2.$$

Obtaining the fast rate - cont.

- **Poincaré Inequality $\implies$ L^2 Bound for μ:**

$$\|\phi - \varphi - \mu(\phi - \varphi)\|^2_{L^2(\mu)} \leq C\|\nabla(\phi - \varphi)\|^2_{L^2(\mu)} \lesssim \varepsilon^2.$$

- **Poincaré Inequality $\implies$ L^2 Bound for ν:**

$$\|\phi^* - \varphi^* - \nu(\phi^* - \varphi^*)\|^2_{L^2(\nu)} \leq C\|\nabla(\phi^* - \varphi^*)\|^2_{L^2(\nu)} \lesssim \varepsilon^2.$$

- But we haven't proved (i) a Poincaré inequality for ν (ii) a bound for $\|\nabla(\phi^* - \varphi^*)\|^2_{L^2(\nu)}$

Obtaining the fast rate - cont.

- **Poincaré Inequality ν:** The Poincaré inequality for μ implies that for any $f \in L^2(\nu)$,

$$\int \left(f - \int f \, d\nu \right)^2 d\nu = \int \left(f \circ \nabla\varphi - \int f \circ \nabla\varphi \, d\mu \right)^2 d\mu$$

$$\leq C \int \|\nabla(f \circ \nabla\varphi)\|^2 d\mu \leq 4C \int \|\nabla f(\nabla\varphi(x))\|^2 d\mu \leq 4C \int \|\nabla f\|^2 d\nu$$

Obtaining the fast rate - cont.

- **Poincaré Inequality ν:** The Poincaré inequality for μ implies that for any $f \in L^2(\nu)$,

$$\int \left(f - \int f \, d\nu \right)^2 d\nu = \int \left(f \circ \nabla\varphi - \int f \circ \nabla\varphi \, d\mu \right)^2 d\mu$$

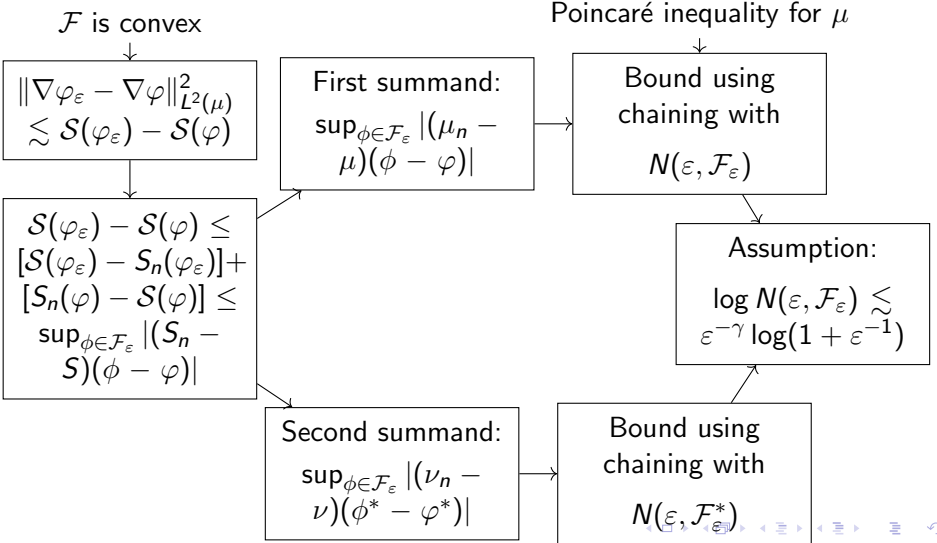
$$\leq C \int \|\nabla(f \circ \nabla\varphi)\|^2 d\mu \leq 4C \int \|\nabla f(\nabla\varphi(x))\|^2 d\mu \leq 4C \int \|\nabla f\|^2 d\nu$$

- **Duality Implies Bound for $\|\nabla(\phi^* - \varphi^*)\|^2_{L^2(\nu)}$:** Define T such that $T(\phi^*) = S(\phi)$. Then, for φ^* is the minimizer of T. The stability result for T implies that

$$\frac{1}{4}\|\nabla\phi^* - \nabla\varphi^*\|^2_{L^2(\nu)} \leq \mathcal{T}(\phi^*) - \mathcal{T}(\varphi^*) = \mathcal{S}(\phi) - \mathcal{S}(\varphi) \leq \|\nabla\phi - \nabla\varphi\|^2_{L^2(\mu)}.$$

Therefore, $\|\nabla\phi^* - \nabla\varphi^*\|^2_{L^2(\nu)} \lesssim \varepsilon^2$ for all $\phi \in \mathcal{F}_\varepsilon$.

Obtaining the fast rate - cont.



Obtaining the fast rate - cont.

Theorem 3 (Fast rate for semi-dual estimator)

Adopt the assumption for covering numbers. Further, assume that

- The supports of μ and ν lie in $\Omega = B_1(0)$.
- $\sup_{\phi \in \mathcal{F}} \|\phi\|_{L^\infty(\Omega)} \leq R$ for some $R > 0$.
- The potentials satisfy $\phi(0) = 0$ and $\phi(x) = +\infty$ if $x \notin \Omega$.
- The potentials are lower-semicontinuous, smooth, and strongly convex on Ω: $\frac{1}{2} I \preceq \nabla^2 \phi(x) \preceq 2I$ if $\|x\| < 1$.
- $\mathcal{F}$ is convex.
- μ satisfies a Poincaré inequality.

Then the semidual estimator $\hat{\varphi}$ satisfies the bound

$$\mathbb{E}\|\nabla\hat{\varphi} - \nabla\varphi\|^2_{L^2(\mu)} \lesssim_{R,\gamma} \left(\frac{\log n}{n}\right)^{\frac{2}{2+\gamma}}$$

Questions from last day

1. Approximation by large parametric class
2. How does the rate depend on the parameters of the Neural Network?

The restricted class $\mathcal{F}$

Approx. by large parametric classes **[DNWP24, Theorem 3]**

Consider μ and $\mathcal{F}$ that satisfy regularity assumptions as before. Assume that φ is both "smooth" and "strongly convex". Let $(\mathcal{F}_j)_j$ be an approximating family of the class of potentials that is "smooth". Then, for J such that $2^J \asymp n^{1/(2s+d-4)}$, it holds that

$$\mathbb{E}\|\nabla\hat{\varphi}_{\mathcal{F}_J} - \nabla\varphi\|^2_{L^2(\mu)} \lesssim_{\log n} n^{-\frac{2(s-1)}{2s+d-4}}.$$

Here, d is approximately linear in the **log of the metric entropy** of $\mathcal{F}$ and s is approximately the infimum of the quantity $\frac{-1}{j}\log_2\|\nabla\varphi - \nabla\varphi_j\|$.

The restricted class $\mathcal{F}$

The dimension d

For every $\tau > 0$ the ball B_τ centered at φ (resp. the ball B_τ^* centered at φ^*) of radius τ in $\mathcal{F}$ for the pseudo-norm $g \mapsto \|\nabla g\|_{L^2(P)}$ (resp. in $\mathcal{F}^*$ for the pseudo-norm $g^* \mapsto \|\nabla g^*\|_{L^2(Q)}$) satisfies for all $0 < \epsilon < \tau$,

$$\log \mathcal{N}(\epsilon, B_\tau, L^2(P)) \lesssim_{\log_+(1/\epsilon), \log_+(1/\tau)} \left(\frac{\epsilon}{\tau}\right)^{-d}$$

$$\log \mathcal{N}(\epsilon, B_\tau^*, L^2(Q)) \lesssim_{\log_+(1/\epsilon), \log_+(1/\tau)} \left(\frac{\epsilon}{\tau}\right)^{-d}$$

The restricted class $\mathcal{F}$

The dimension d

For every $\tau > 0$ the ball B_τ centered at φ (resp. the ball B_τ^* centered at φ^*) of radius τ in $\mathcal{F}$ for the pseudo-norm $g \mapsto \|\nabla g\|_{L^2(P)}$ (resp. in $\mathcal{F}^*$ for the pseudo-norm $g^* \mapsto \|\nabla g^*\|_{L^2(Q)}$) satisfies for all $0 < \epsilon < \tau$,

$$\log \mathcal{N}(\epsilon, B_\tau, L^2(P)) \lesssim_{\log_+(1/\epsilon), \log_+(1/\tau)} \left(\frac{\epsilon}{\tau}\right)^{-d}$$

$$\log \mathcal{N}(\epsilon, B_\tau^*, L^2(Q)) \lesssim_{\log_+(1/\epsilon), \log_+(1/\tau)} \left(\frac{\epsilon}{\tau}\right)^{-d}$$

What about **s** ?

Questions from last day

1. Approximation by large parametric class
2. How does the rate depend on the parameters of the Neural Network?

The restricted class $\mathcal{F}$ - cont.

Example (ReQU Neural Networks)

Assume regularity conditions as before. Further, suppose $\mathcal{F} \subseteq \mathcal{C}_L^s([0,1]^d)$ for some $L \geq 0$ and $s > 2$. There exists a family $\mathcal{F}_j$ of ReQU neural networks with

- **Depth (Number of Layers):** $D_j = O\big(\log d + \lfloor s \rfloor + \log\log L\big)$.
- **Width (Number of Neurons per Layer):** $W_j = O\big(d(2^j + \lfloor s \rfloor)^d\big)$
- **Total Number of Parameters:** $N_j = O\big((ds + d^2 + \log\log L)(2^j + \lfloor s \rfloor)^d\big)$.

which approximates $\mathcal{F}$ well in Hölder norm. We then modify the family of potentials to be equal to $+\infty$ outside $B(0, 2R)$. Then for a certain choice of J

$$\mathbb{E}\|\nabla\widehat{\varphi}_{\tilde{\mathcal{F}}_J} - \nabla\varphi_0\|_{L^2(P)}^2 \lesssim_{\log n} n^{-\frac{2(s-1)}{2s+d-4}}.$$

Next time:

Entropic Optimal Transport