Effect of Dependence on the Convergence of Empirical Wasserstein Distance

Nabarun Deb
Mathematics Department, University of British Columbia
Kantorovich Initiative Postdoc

Joint work with Debarghya Mukherjee (Princeton University)

Kantorovich Initiative Seminar Series, 2022-23

Introduction to Wasserstein Distance

Let $\mathcal{X}$, $\mathcal{Y}$ be subsets of $\mathbb{R}^d$. Given two measures μ and ν supported on $\mathcal{X}$ and $\mathcal{Y}$,

$$W_p^p(\mu, \nu) := \min_{\pi \in \Gamma(\mu,\nu)} \int \|x - y\|^p \, d\pi(x, y)$$

for $p \geq 1$, where $\Gamma(\mu, \nu)$ is the space of probability measures on $\mathcal{X} \times \mathcal{Y}$ with marginals μ and ν.

Introduction to Wasserstein Distance

Let $\mathcal{X}$, $\mathcal{Y}$ be subsets of $\mathbb{R}^d$. Given two measures μ and ν supported on $\mathcal{X}$ and $\mathcal{Y}$,

$$W_p^p(\mu, \nu) := \min_{\pi \in \Gamma(\mu,\nu)} \int \|x - y\|^p \, d\pi(x, y)$$

for $p \geq 1$, where $\Gamma(\mu, \nu)$ is the space of probability measures on $\mathcal{X} \times \mathcal{Y}$ with marginals μ and ν.

For this talk ...

- $\mathcal{X}$ and $\mathcal{Y}$ are compact subsets of $\mathbb{R}^d$
- Take $d \geq 5$

Introduction to Wasserstein Distance

Let $\mathcal{X}$, $\mathcal{Y}$ be subsets of $\mathbb{R}^d$. Given two measures μ and ν supported on $\mathcal{X}$ and $\mathcal{Y}$,

$$W_p^p(\mu, \nu) := \min_{\pi \in \Gamma(\mu,\nu)} \int \|x - y\|^p \, d\pi(x, y)$$

for $p \geq 1$, where $\Gamma(\mu, \nu)$ is the space of probability measures on $\mathcal{X} \times \mathcal{Y}$ with marginals μ and ν.

For this talk ...

- $\mathcal{X}$ and $\mathcal{Y}$ are compact subsets of $\mathbb{R}^d$
- Take $d \geq 5$
- Objective is to estimate $W_p(\mu, \nu)$. Applications in —
  1. Computational biology: Schiebinger et al., 2019, Tameling et al., 2021
  2. Signal and image processing: Bonneel et al., 2011; Kolouri et al., 2017
  3. Also see Panaretos and Zemel (2019), Santambrogio (2015), Peyré and Cuturi (2019) for surveys.

Estimation of $W_p(\mu, \nu)$ — Plug-in principle

- Usual setting: $X_1, X_2, \ldots, X_n \overset{i.i.d}{\sim} \mu$ and $Y_1, Y_2, \ldots, Y_n \overset{i.i.d.}{\sim} \nu$.

Estimation of $W_p(\mu, \nu)$ — Plug-in principle

- Usual setting: $X_1, X_2, \ldots, X_n \overset{i.i.d}{\sim} \mu$ and $Y_1, Y_2, \ldots, Y_n \overset{i.i.d.}{\sim} \nu$.
- Plug-in idea: Replace μ by μ_n and ν by ν_n, where

$$\mu_n = \frac{1}{n}\sum_{i=1}^{n} \delta_{X_i}, \qquad \nu_n = \frac{1}{n}\sum_{j=1}^{n} \delta_{Y_j}.$$

Estimation of $W_p(\mu, \nu)$ — Plug-in principle

- Usual setting: $X_1, X_2, \dots, X_n \overset{i.i.d}{\sim} \mu$ and $Y_1, Y_2, \dots, Y_n \overset{i.i.d.}{\sim} \nu$.
- Plug-in idea: Replace μ by μ_n and ν by ν_n, where

$$\mu_n = \frac{1}{n}\sum_{i=1}^{n} \delta_{X_i}, \qquad \nu_n = \frac{1}{n}\sum_{j=1}^{n} \delta_{Y_j}.$$

- Estimate $W_p(\mu, \nu)$ by $W_p(\mu_n, \nu_n)$.
- Can be computed exactly using the Hungarian algorithm; parallel computing Date and Nagi (2016)
- Extensively studied estimator, including rates of convergence, tail bounds, lower bounds, central limit theorems (appropriate centering), ...
- See Dudley (1969), Boissard and Le Gouic (2014), Fournier and Guillin (2015), Singh and Póczos (2018), Liang (2019), Niles-Weed and Rigollet (2019), Manole and Niles-Weed (2021), Chizat et al. (2020), Hundrieser et al. (2021), Hundrieser et al. (2022), ...

Plug-in principle continued ...

- By triangle inequality

$$\sup_{(\mu,\nu)} \mathbb{E}|W_p(\mu_n,\nu_n) - W_p(\mu,\nu)| \leq \sup_{(\mu,\nu)} \left(\mathbb{E}W_p(\mu_n,\mu) + \mathbb{E}W_p(\nu_n,\nu)\right) \lesssim n^{-1/d},$$

for $2p < d$, see Fournier and Guillin (2015).

Plug-in principle continued ...

- By triangle inequality

$$\sup_{(\mu,\nu)} \mathbb{E}|W_p(\mu_n,\nu_n) - W_p(\mu,\nu)| \leq \sup_{(\mu,\nu)} \left(\mathbb{E}W_p(\mu_n,\mu) + \mathbb{E}W_p(\nu_n,\nu)\right) \lesssim n^{-1/d},$$

  for $2p < d$, see Fournier and Guillin (2015).

- This bound cannot be improved in general, see Liang (2019), Niles-Weed and Rigollet (2019)

Plug-in principle continued ...

- By triangle inequality

$$\sup_{(\mu,\nu)} \mathbb{E}|W_p(\mu_n,\nu_n) - W_p(\mu,\nu)| \leq \sup_{(\mu,\nu)} \left(\mathbb{E}W_p(\mu_n,\mu) + \mathbb{E}W_p(\nu_n,\nu)\right) \lesssim n^{-1/d},$$

  for $2p < d$, see Fournier and Guillin (2015).

- This bound cannot be improved in general, see Liang (2019), Niles-Weed and Rigollet (2019)

- Crucially the worst case rate comes from measures μ and ν which are close.

- If $W_p(\mu,\nu)$ is bounded away from 0, faster rates (see Chizat et al. (2020), Manole and Niles-Weed (2021), Hundreiser et al. (2021)),

$$\sup_{(\mu,\nu):\ W_p(\mu,\nu)>\delta} \mathbb{E}|W_p(\mu_n,\nu_n) - W_p(\mu,\nu)| \lesssim n^{-\frac{\min(p,2)}{d}}.$$

Plug-in principle continued ...

- By triangle inequality

$$\sup_{(\mu,\nu)} \mathbb{E}|W_p(\mu_n,\nu_n)-W_p(\mu,\nu)| \leq \sup_{(\mu,\nu)} (\mathbb{E}W_p(\mu_n,\mu)+\mathbb{E}W_p(\nu_n,\nu)) \lesssim n^{-1/d},$$

  for $2p < d$, see Fournier and Guillin (2015).

- This bound cannot be improved in general, see Liang (2019), Niles-Weed and Rigollet (2019)

- Crucially the worst case rate comes from measures μ and ν which are close.

- If $W_p(\mu,\nu)$ is bounded away from 0, faster rates (see Chizat et al. (2020), Manole and Niles-Weed (2021), Hundreiser et al. (2021)),

$$\sup_{(\mu,\nu):\ W_p(\mu,\nu)>\delta} \mathbb{E}|W_p(\mu_n,\nu_n)-W_p(\mu,\nu)| \lesssim n^{-\frac{\min(p,2)}{d}}.$$

In this talk ...

This case is our focus, where we change the i.i.d. assumption.

Why dependence?

Dependence can arise in many natural settings:

- Time series data in economics and finance (e.g. stock market data, weather data)
- Markov chains, hidden markov models
- Online learning, where data comes in stream (e.g. object tracking, strategic classification, reinforcement learning etc.)
- Longitudinal medical data (e.g. sequence of data of a patient over a time horizon)

Dependence and Wasserstein distance

- The rate of convergence of the empirical measure under $W_p(\mu_n, \mu)$ — **Fournier and Guillin, 2015**. The rate slows down under long range dependence (more on this later)

Dependence and Wasserstein distance

- The rate of convergence of the empirical measure under $W_p(\mu_n, \mu)$ — **Fournier and Guillin, 2015**. The rate slows down under long range dependence (more on this later)

- Suppose $X_1, X_2, \ldots$ and $Y_1, Y_2, \ldots$ are stationary with marginals μ and ν. Then for $d = 1, 2, 3$ and under short range dependence (other technical assumptions), Hundreiser et al. (2022) proved that

$$\sqrt{n}(W_p(\mu_n, \nu_n) - W_p(\mu, \nu)) \xrightarrow{d} \mathcal{N}(0, \sigma^2_{\mu,\nu}).$$

Here $\sigma^2_{\mu,\nu} > 0$ if $\mu \neq \nu$.

Dependence and Wasserstein distance

- The rate of convergence of the empirical measure under $W_p(\mu_n, \mu)$ — **Fournier and Guillin, 2015**. The rate slows down under long range dependence (more on this later)

- Suppose $X_1, X_2, \ldots$ and $Y_1, Y_2, \ldots$ are stationary with marginals μ and ν. Then for $d = 1, 2, 3$ and under short range dependence (other technical assumptions), Hundreiser et al. (2022) proved that

$$\sqrt{n}(W_p(\mu_n, \nu_n) - W_p(\mu, \nu)) \xrightarrow{d} \mathcal{N}(0, \sigma^2_{\mu,\nu}).$$

  Here $\sigma^2_{\mu,\nu} > 0$ if $\mu \neq \nu$.

- CLTs for regularized Wasserstein distances — Goldfeld et al. (2022) — same flavor as above

Dependence and Wasserstein distance (Continued)

- CLTs for parameter estimators via Wasserstein minimization — Bernton et al. (2019). Consider $\mu - \mathbb{P}_{\theta^*}$ where $d = 1$ and $\theta^* \in \mathbb{R}^r$. Consider

$$\hat{\theta}_n \in \arg\min W_1(\mu_n, \mathbb{P}_{\theta^*}).$$

Then under short range dependence,

$$\sqrt{n}(\hat{\theta}_n - \theta^*) \xrightarrow{d} \arg\min_u \int |G^*(t) - \langle u, D_{\theta^*}(t)\rangle| \, dt.$$

Here $D_{\theta^*}(\cdot) : \mathbb{R} \to \mathbb{R}^r$ is a smooth map depending on $\mathbb{P}_\theta$.

Dependence and Wasserstein distance (Continued)

- CLTs for parameter estimators via Wasserstein minimization — Bernton et al. (2019). Consider $\mu - \mathbb{P}_{\theta^*}$ where $d = 1$ and $\theta^* \in \mathbb{R}^r$. Consider

$$\hat{\theta}_n \in \arg\min W_1(\mu_n, \mathbb{P}_{\theta^*}).$$

  Then under short range dependence,

$$\sqrt{n}(\hat{\theta}_n - \theta^*) \xrightarrow{d} \arg\min_u \int |G^*(t) - \langle u, D_{\theta^*}(t)\rangle|\, dt.$$

  Here $D_{\theta^*}(\cdot) : \mathbb{R} \to \mathbb{R}^r$ is a smooth map depending on $\mathbb{P}_\theta$.

- Comparing two (stationary) time series using spectral densities and a Wasserstein (Fourier) distance — Cazelles et al. (2020)

Dependence and Wasserstein distance (Continued)

- CLTs for parameter estimators via Wasserstein minimization — Bernton et al. (2019). Consider $\mu - \mathbb{P}_{\theta^*}$ where $d = 1$ and $\theta^* \in \mathbb{R}^r$. Consider

  $$\hat{\theta}_n \in \arg\min W_1(\mu_n, \mathbb{P}_{\theta^*}).$$

  Then under short range dependence,

  $$\sqrt{n}(\hat{\theta}_n - \theta^*) \xrightarrow{d} \arg\min_u \int |G^*(t) - \langle u, D_{\theta^*}(t)\rangle| \, dt.$$

  Here $D_{\theta^*}(\cdot) : \mathbb{R} \to \mathbb{R}^r$ is a smooth map depending on $\mathbb{P}_\theta$.

- Comparing two (stationary) time series using spectral densities and a Wasserstein (Fourier) distance — Cazelles et al. (2020)

- Constrained optimal transport on markov chains O'Connor et al. (2022)

- Using Wasserstein distances to analyze — visualize+synchronize non-linear time series Muskulus and Verduyn-Lunel (2011)

A simple example: MA(∞) model

- Consider the following moving average model:

$$X_i = \sum_{k=0}^{\infty} a_k \epsilon_{i-k}$$

where $\epsilon_k \overset{i.i.d.}{\sim} \mathcal{N}(0, \sigma^2 I_2)$.

A simple example: MA(∞) model

- Consider the following moving average model:

$$X_i = \sum_{k=0}^{\infty} a_k \epsilon_{i-k}$$

where $\epsilon_k \overset{i.i.d.}{\sim} \mathcal{N}(0, \sigma^2 I_2)$.

- Assume

$$a_k = k^{-\rho}$$

for some $\rho > 1/2$.

A simple example: MA(∞) model

- Consider the following moving average model:

$$X_i = \sum_{k=0}^{\infty} a_k \epsilon_{i-k}$$

  where $\epsilon_k \overset{i.i.d.}{\sim} \mathcal{N}(0, \sigma^2 I_2)$.

- Assume

$$a_k = k^{-\rho}$$

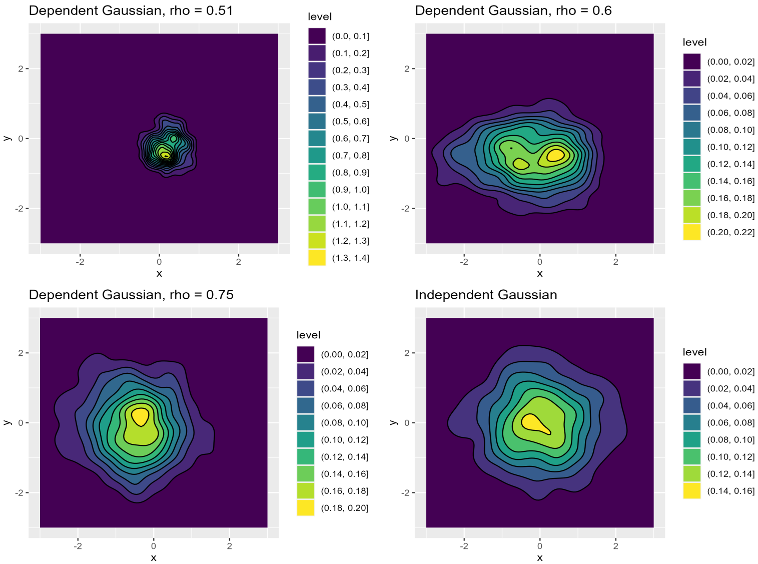
  for some $\rho > 1/2$.

- Easy to check that the series converges a.s. and

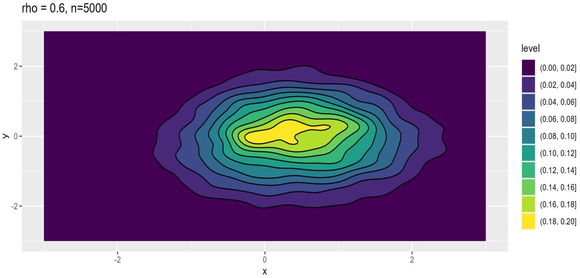
$$X_i \sim \mathcal{N}(0, I_2 \underbrace{\sum_{j=1}^{\infty} j^{-2\rho}}_{\sigma_\rho^2}).$$

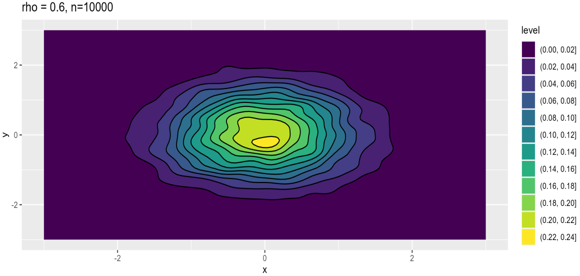
- Set $Y_i = \frac{X_i}{\sigma_\rho} \sim \mathcal{N}(0, I_2)$ but the joint distribution of $(Y_1, \ldots, Y_n)$ depends heavily on ρ.

Kernel density contours across mixing



Kernel density contours with n





A log-log plot in the two-sample case

- Consider $\{X_i\}_{i\geq 1}$ and $\{Z_i\}_{i\geq 1}$ be two MA(∞) sequences with σ^2 equals 1 and 4 respectively.
- $W_2(\mu, \nu)$ has closed forms as they are both Gaussian.

A log-log plot in the two-sample case

- Consider $\{X_i\}_{i\geq 1}$ and $\{Z_i\}_{i\geq 1}$ be two MA(∞) sequences with σ^2 equals 1 and 4 respectively.
- $W_2(\mu, \nu)$ has closed forms as they are both Gaussian.
- Want to study $|W_2(\mu_n, \nu_n) - W_2(\mu, \nu)|$ empirically

A log-log plot in the two-sample case

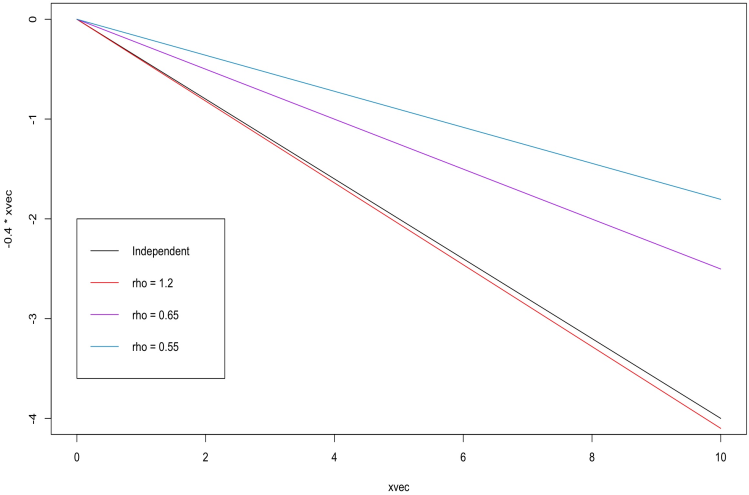
- Consider $\{X_i\}_{i\geq 1}$ and $\{Z_i\}_{i\geq 1}$ be two MA(∞) sequences with σ^2 equals 1 and 4 respectively.
- $W_2(\mu, \nu)$ has closed forms as they are both Gaussian.
- Want to study $|W_2(\mu_n, \nu_n) - W_2(\mu, \nu)|$ empirically
- Choose the number of samples n varying in a grid between 2^9 — 2^{12}
- Compute $W_2(\mu_n, \nu_n)$ for each n in the grid. Replicate the experiment 1000 times
- Look at the slope of the regression line of

$$\log_2\left(\text{av}|W_2(\mu, \nu_n) - W_2(\mu, \nu)|\right) \quad \text{on} \quad \log_2(n)$$

- The slope of the line is expected to indicate the rate of convergence

Plots of rates

Under independence between $\{X_i\}$ and $\{Z_i\}$, rate $2/d = 2/5 = 0.4$.



Outline

Outline

Notions of strong mixing for dependence

- Given a strictly stationary sequence of random variables $\{X_t\}_{t\in\mathbb{N}}$ on a probability space $(\Omega, \mathcal{A}, \mathbb{P})$, the we understand dependence/independence as a property of the underlying σ-field.

Notions of strong mixing for dependence

- Given a strictly stationary sequence of random variables $\{X_t\}_{t\in\mathbb{N}}$ on a probability space $(\Omega, \mathcal{A}, \mathbb{P})$, the we understand dependence/independence as a property of the underlying σ-field.

- Four (arguably) most popular used notion of dependence:
  1. $\alpha(n) = \sup_{k\geq 1} \sup_{\substack{A\in\sigma(X_{1:k}) \\ B\in\sigma(X_{k+n+1:\infty})}} |\mathbb{P}(A\cap B) - \mathbb{P}(A)\mathbb{P}(B)|$

Notions of strong mixing for dependence

- Given a strictly stationary sequence of random variables $\{X_t\}_{t\in\mathbb{N}}$ on a probability space $(\Omega, \mathcal{A}, \mathbb{P})$, the we understand dependence/independence as a property of the underlying σ-field.

- Four (arguably) most popular used notion of dependence:
  1. $\alpha(n) = \sup_{k\geq 1} \sup_{\substack{A\in\sigma(X_{1:k}) \\ B\in\sigma(X_{k+n+1:\infty})}} |\mathbb{P}(A\cap B) - \mathbb{P}(A)\mathbb{P}(B)|$
  2. $\beta(n) = \sup_{k\geq 1} \mathbb{E}\left[\sup_{A\in\sigma(X_{1:k})} |\mathbb{P}(A \mid \sigma(X_{k+n+1:\infty})) - \mathbb{P}(A)|\right]$

Notions of strong mixing for dependence

- Given a strictly stationary sequence of random variables $\{X_t\}_{t\in\mathbb{N}}$ on a probability space $(\Omega, \mathcal{A}, \mathbb{P})$, the we understand dependence/independence as a property of the underlying σ-field.

- Four (arguably) most popular used notion of dependence:
  1. $\alpha(n) = \sup_{k\geq 1} \sup_{\substack{A\in\sigma(X_{1:k}) \\ B\in\sigma(X_{k+n+1:\infty})}} |\mathbb{P}(A\cap B) - \mathbb{P}(A)\mathbb{P}(B)|$
  2. $\beta(n) = \sup_{k\geq 1} \mathbb{E}\left[\sup_{A\in\sigma(X_{1:k})} |\mathbb{P}(A \mid \sigma(X_{k+n+1:\infty})) - \mathbb{P}(A)|\right]$
  3. $\rho(n) = \sup_{k\geq 1} \sup_{\substack{f\in L_2(\sigma(X_{1:k})) \\ g\in\sigma(X_{k+n+1:\infty})}} |\mathrm{cor}(f, g)|$

Notions of strong mixing for dependence

- Given a strictly stationary sequence of random variables $\{X_t\}_{t\in\mathbb{N}}$ on a probability space $(\Omega, \mathcal{A}, \mathbb{P})$, the we understand dependence/independence as a property of the underlying σ-field.

- Four (arguably) most popular used notion of dependence:
  1. $\alpha(n) = \sup_{k\geq 1} \sup_{\substack{A\in\sigma(X_{1:k}) \\ B\in\sigma(X_{k+n+1:\infty})}} |\mathbb{P}(A\cap B) - \mathbb{P}(A)\mathbb{P}(B)|$
  2. $\beta(n) = \sup_{k\geq 1} \mathbb{E}\left[\sup_{A\in\sigma(X_{1:k})} |\mathbb{P}(A \mid \sigma(X_{k+n+1:\infty})) - \mathbb{P}(A)|\right]$
  3. $\rho(n) = \sup_{k\geq 1} \sup_{\substack{f\in L_2(\sigma(X_{1:k})) \\ g\in\sigma(X_{k+n+1:\infty})}} |\mathrm{cor}(f,g)|$
  4. $\phi(n) = \sup_{k\geq 1} \sup_{\substack{A\in\sigma(X_{1:k}) \\ B\in\sigma(X_{k+n+1:\infty})}} |\mathbb{P}(A \mid B) - \mathbb{P}(A)|$

- Relation between the notions:

$$2\alpha(n) \leq \beta(n) \leq \phi(n), \quad 4\alpha(n) \leq \rho(n) \leq 2\sqrt{\phi(n)}$$

Notions of strong mixing for dependence

- Given a strictly stationary sequence of random variables $\{X_t\}_{t\in\mathbb{N}}$ on a probability space $(\Omega, \mathcal{A}, \mathbb{P})$, the we understand dependence/independence as a property of the underlying σ-field.

- Four (arguably) most popular used notion of dependence:
  1. $\alpha(n) = \sup_{k\geq 1} \sup_{\substack{A\in\sigma(X_{1:k}) \\ B\in\sigma(X_{k+n+1:\infty})}} |\mathbb{P}(A\cap B) - \mathbb{P}(A)\mathbb{P}(B)|$
  2. $\beta(n) = \sup_{k\geq 1} \mathbb{E}\left[\sup_{A\in\sigma(X_{1:k})} |\mathbb{P}(A \mid \sigma(X_{k+n+1:\infty})) - \mathbb{P}(A)|\right]$
  3. $\rho(n) = \sup_{k\geq 1} \sup_{\substack{f\in L_2(\sigma(X_{1:k})) \\ g\in\sigma(X_{k+n+1:\infty})}} |\mathrm{cor}(f,g)|$
  4. $\phi(n) = \sup_{k\geq 1} \sup_{\substack{A\in\sigma(X_{1:k}) \\ B\in\sigma(X_{k+n+1:\infty})}} |\mathbb{P}(A \mid B) - \mathbb{P}(A)|$

- Relation between the notions:

$$2\alpha(n) \leq \beta(n) \leq \phi(n), \quad 4\alpha(n) \leq \rho(n) \leq 2\sqrt{\phi(n)}$$

Goal

Bound $\mathbb{E}|W_p(\mu_n, \nu_n) - W_p(\mu, \nu)|$ in terms of β-mixing coefficients

β-mixing and Berbee's Coupling

β-mixing is typically regarded as second most general notion:

1. (Eberlein, (1984)) established CLT for β-mixing sequence under the condition $\beta(n) = n^{-(1+\epsilon)(1+2/\delta)}$.
2. (Yu (1994)), (Doukhan et.al. (1994), (1995)) extended some results of standard empirical process theory for β-mixing sequence.
3. (Karandikar et.al. (2009)) extended some aspects of Bayesian learning to β-mixing sequences.
4. (Bernton et al. (2019), Goldfeld et al. (2022)) show $\sqrt{n}$ rates for parameter estimation and regularized OT under β-mixing

β-mixing and Berbee's Coupling

β-mixing is typically regarded as second most general notion:

1. (Eberlein, (1984)) established CLT for β-mixing sequence under the condition $\beta(n) = n^{-(1+\epsilon)(1+2/\delta)}$.
2. (Yu (1994)), (Doukhan et.al. (1994), (1995)) extended some results of standard empirical process theory for β-mixing sequence.
3. (Karandikar et.al. (2009)) extended some aspects of Bayesian learning to β-mixing sequences.
4. (Bernton et al. (2019), Goldfeld et al. (2022)) show $\sqrt{n}$ rates for parameter estimation and regularized OT under β-mixing

Theorem (Berbee's Coupling)

Given (X, Y) and an independent $U \sim Unif\,(0,1)$ on the same probability space, one can construct $Y^* = f(X, Y, U)$ such that:

1. $Y^* \stackrel{\mathscr{L}}{=} Y$ *and* $Y^* \perp\!\!\!\perp X$.
2. $\mathbb{P}(Y \neq Y^*) = \beta(\sigma(X), \sigma(Y))$.

Outline

An ambiguous definition

- Using β-mixing as a *proxy*, short range and long range dependencies typically mean

$$\sum_k \beta(k) < \infty \quad \text{Short range},$$

$$\sum_k \beta(k) = \infty \quad \text{Long range}.$$

- Same with other mixing coefficients.

An ambiguous definition

- Using β-mixing as a *proxy*, short range and long range dependencies typically mean

$$\sum_k \beta(k) < \infty \quad \text{Short range},$$

$$\sum_k \beta(k) = \infty \quad \text{Long range}.$$

- Same with other mixing coefficients.
- By Rio (1995), Dedecker (2003), say $\{X_t\}_t$ is a strictly stationary β-mixing sequence, then

$$\mathrm{Var}(\sum_{t=1}^{n} X_t) \lesssim n(1 + \sum_{k=0}^{\infty} \beta(k)).$$

Under long range dependence, behavior of $\sum_{t=1}^{n} X_t$ can be very different from i.i.d. case.

Long range and short range dependency (continued)

- Standard properties like WLLN, CLT continues to hold under SRD:
  1. A general version of CLT was proved in Peligrad, (1990)
  2. Consistency for non-parametric kernel density estimation was established in (Roussas, (1990)).
  3. Bernstein type concentration inequality was established in (Merlevede, Peligrad and Rio, (1990)).
  4. In OT, Bernton et al. (2019), Goldfeld et al. (2022) obtain limit theory under SRD with β-mixing

Long range and short range dependency (continued)

- Standard properties like WLLN, CLT continues to hold under SRD:
  1. A general version of CLT was proved in Peligrad, (1990)
  2. Consistency for non-parametric kernel density estimation was established in (Roussas, (1990)).
  3. Bernstein type concentration inequality was established in (Merlevede, Peligrad and Rio, (1990)).
  4. In OT, Bernton et al. (2019), Goldfeld et al. (2022) obtain limit theory under SRD with β-mixing
  5. In Fournier and Guillin (2015), rates were obtained for SRD with ρ-mixing (same as i.i.d. case)

Long range and short range dependency (continued)

- Standard properties like WLLN, CLT continues to hold under SRD:
  1. A general version of CLT was proved in Peligrad, (1990)
  2. Consistency for non-parametric kernel density estimation was established in (Roussas, (1990)).
  3. Bernstein type concentration inequality was established in (Merlevede, Peligrad and Rio, (1990)).
  4. In OT, Bernton et al. (2019), Goldfeld et al. (2022) obtain limit theory under SRD with β-mixing
  5. In Fournier and Guillin (2015), rates were obtained for SRD with ρ-mixing (same as i.i.d. case)
- Properties under LRD is much less explored: a noteworthy example is (Yu, 1994) where some properties of expected suprema of an empirical process is established under LRD.

Effect of dependence on estimation of Wasserstein distance

Recall the result of (Fournier and Guillin, 2015):

One of their main results

If $\{X_i\}_{i=1}^n$ is a sequence of stationary random variable with summable ρ-mixing sequence, i.e. $\sum_k \rho(k) < \infty$. Then:

$$\mathbb{E}\left[W_p(\mu_n, \mu)\right] \lesssim \begin{cases} n^{-\frac{1}{2p}}\,, & \text{if } p > d/2 \\ n^{-\frac{1}{2p}} \log\left(1+n\right), & \text{if } p = d/2 \\ n^{-\frac{1}{d}}\,, & \text{if } p < d/2\,. \end{cases}$$

Effect of dependence on estimation of Wasserstein distance

Recall the result of (Fournier and Guillin, 2015):

One of their main results

If $\{X_i\}_{i=1}^n$ is a sequence of stationary random variable with summable ρ-mixing sequence, i.e. $\sum_k \rho(k) < \infty$. Then:

$$\mathbb{E}\left[W_p(\mu_n, \mu)\right] \lesssim \begin{cases} n^{-\frac{1}{2p}}\,, & \text{if } p > d/2 \\ n^{-\frac{1}{2p}} \log\left(1+n\right), & \text{if } p = d/2 \\ n^{-\frac{1}{d}}\,, & \text{if } p < d/2\,. \end{cases}$$

Proposition (Directly applying the $W_p(\mu_n, \mu)$ bounds)

If $\{X_i\}_{i=1}^n$ is a sequence of compactly supported stationary random variable with $\rho(k) = k^{-\rho}$ for some $\rho > 0$. Then:

$$\mathbb{E}\left[W_p(\mu_n, \mu)\right] \lesssim \begin{cases} n^{-\frac{(\rho\wedge 1)}{2p}}\,, & \text{if } p > d/2 \\ n^{-\frac{(\rho\wedge 1)}{2p}} \log\left(1+n\right), & \text{if } p = d/2 \\ n^{-\frac{(\rho\wedge 1)}{d}}\,, & \text{if } p < d/2\,. \end{cases}$$

Outline

Main result

Define

$$p^* = \min(p, 2), \qquad \beta^* := \frac{p*}{d - p*} < 1.$$

Main result

Suppose $X_1, \ldots, X_n$ and $Y_1, \ldots, Y_n$ are drawn from strictly stationary sequences with common marginals μ and ν respectively. Say both sequences have a β-mixing coefficient $\beta(k) = k^{-\beta}$ for some $\beta > 0$. Then under the usual assumptions:

$$\mathbb{E}|W_p(\mu_n, \nu_n) - W_p(\mu, \nu)| \lesssim \begin{cases} n^{-\frac{p^*}{d}} & \text{if } \beta > \beta^* \\ n^{-\frac{\beta}{1+\beta}} & \text{if } \beta < \beta^*. \end{cases}$$

Main result

Define

$$p^* = \min(p, 2), \qquad \beta^* := \frac{p*}{d - p*} < 1.$$

Main result

Suppose $X_1, \ldots, X_n$ and $Y_1, \ldots, Y_n$ are drawn from strictly stationary sequences with common marginals μ and ν respectively. Say both sequences have a β-mixing coefficient $\beta(k) = k^{-\beta}$ for some $\beta > 0$. Then under the usual assumptions:

$$\mathbb{E}|W_p(\mu_n, \nu_n) - W_p(\mu, \nu)| \lesssim \begin{cases} n^{-\frac{p^*}{d}} & \text{if } \beta > \beta^* \\ n^{-\frac{\beta}{1+\beta}} & \text{if } \beta < \beta^*. \end{cases}$$

- Short range ($\beta > 1$) Rate always same as in the i.i.d. case.

Main result

Define

$$p^* = \min(p, 2), \qquad \beta^* := \frac{p*}{d - p*} < 1.$$

Main result

Suppose $X_1, \ldots, X_n$ and $Y_1, \ldots, Y_n$ are drawn from strictly stationary sequences with common marginals μ and ν respectively. Say both sequences have a β-mixing coefficient $\beta(k) = k^{-\beta}$ for some $\beta > 0$. Then under the usual assumptions:

$$\mathbb{E}|W_p(\mu_n, \nu_n) - W_p(\mu, \nu)| \lesssim \begin{cases} n^{-\frac{p^*}{d}} & \text{if } \beta > \beta^* \\ n^{-\frac{\beta}{1+\beta}} & \text{if } \beta < \beta^*. \end{cases}$$

- Short range ($\beta > 1$) Rate always same as in the i.i.d. case.
- Long range ($\beta < 1$) Up to a dimension factor (inversely proportional to d), you do not see the effect of dependence — same rates as i.i.d.

Main result

Define

$$p^* = \min(p, 2), \qquad \beta^* := \frac{p*}{d - p*} < 1.$$

Main result

Suppose $X_1, \ldots, X_n$ and $Y_1, \ldots, Y_n$ are drawn from strictly stationary sequences with common marginals μ and ν respectively. Say both sequences have a β-mixing coefficient $\beta(k) = k^{-\beta}$ for some $\beta > 0$. Then under the usual assumptions:

$$\mathbb{E}|W_p(\mu_n, \nu_n) - W_p(\mu, \nu)| \lesssim \begin{cases} n^{-\frac{p^*}{d}} & \text{if } \beta > \beta^* \\ n^{-\frac{\beta}{1+\beta}} & \text{if } \beta < \beta^*. \end{cases}$$

- Short range ($\beta > 1$) Rate always same as in the i.i.d. case.
- Long range ($\beta < 1$) Up to a dimension factor (inversely proportional to d), you do not see the effect of dependence — same rates as i.i.d.
- Certain decoupling effect — $n^{-\frac{\beta}{\beta+1}}$ and $n^{-\frac{p^*}{d}}$, none of the terms depend on β and d simultaneously (different from Fournier and Guillin (2015))

A related result

What happens in the absence of the curse of dimensionality? — semi-discrete problem.

A related result

What happens in the absence of the curse of dimensionality? — semi-discrete problem.

Finitely supported measure

Suppose $X_1, \ldots, X_n$ and $Y_1, \ldots, Y_n$ are drawn from strictly stationary sequences with common marginals μ and ν respectively, where one of the measures is finitely supported. Say both sequences have a β-mixing coefficient $\beta(k) = k^{-\beta}$ for some $\beta > 0$. Then under the usual assumptions:

$$\mathbb{E}|W_p(\mu_n, \nu_n) - W_p(\mu, \nu)| \lesssim \begin{cases} n^{-\frac{1}{2}} & \text{if } \beta > 1 \\ n^{-\frac{\beta}{1+\beta}} & \text{if } \beta < 1. \end{cases}$$

A related result

What happens in the absence of the curse of dimensionality? — semi-discrete problem.

Finitely supported measure

Suppose $X_1, \dots, X_n$ and $Y_1, \dots, Y_n$ are drawn from strictly stationary sequences with common marginals μ and ν respectively, where one of the measures is finitely supported. Say both sequences have a β-mixing coefficient $\beta(k) = k^{-\beta}$ for some $\beta > 0$. Then under the usual assumptions:

$$\mathbb{E}|W_p(\mu_n, \nu_n) - W_p(\mu, \nu)| \lesssim \begin{cases} n^{-\frac{1}{2}} & \text{if } \beta > 1 \\ n^{-\frac{\beta}{1+\beta}} & \text{if } \beta < 1. \end{cases}$$

- The rates under the empirical measure adapt to the fact that one of the measures is inherently less complex
- Under independence, the adaptation was proved in Hundrieser et al. (2021). For related results, see Niles-Weed and Bach(2022)

Outline

Proof ideas: Preliminary

- Let's start with the dual formulation for $W_2^2(\mu, \nu)$:

$$\sup_{\substack{f:\mathcal{X}\to\mathcal{Y} \\ f\in\mathrm{CVX}, \|f\|_\infty \le 1}} \left\{ \int (\|x\|^2 - 2f(x))\, d\mu(x) + \int (\|y\|^2 - 2f^*(y))\, d\nu(y) \right\}$$

Proof ideas: Preliminary

- Let's start with the dual formulation for $W_2^2(\mu, \nu)$:

$$\sup_{\substack{f:\mathcal{X}\to\mathcal{Y} \\ f\in \mathrm{CVX}, \|f\|_\infty \le 1}} \left\{ \int (\|x\|^2 - 2f(x))\, d\mu(x) + \int (\|y\|^2 - 2f^*(y))\, d\nu(y) \right\}$$

- From OT to empirical process (Chizat et al. (2020), Mena and Niles-Weed (2019), Manole and Weed (2021), ...):

$$\begin{aligned} &\mathbb{E}\left[\left|W_2^2(\mu_n, \nu_n) - W_2^2(\mu, \nu)\right|\right] \\ &\quad \le \mathbb{E}\left[\sup_{f\in\mathcal{F}} \int f\ (d\mu_n - d\nu)\right] + \mathbb{E}\left[\sup_{f\in\mathcal{F}} \int f\ (d\nu_n - d\nu)\right] \end{aligned}$$

Proof ideas: Preliminary

- Let's start with the dual formulation for $W_2^2(\mu, \nu)$:

$$\sup_{\substack{f:\mathcal{X}\to\mathcal{Y} \\ f\in \mathrm{CVX}, \|f\|_\infty \le 1}} \left\{ \int (\|x\|^2 - 2f(x)) \, d\mu(x) + \int (\|y\|^2 - 2f^*(y)) \, d\nu(y) \right\}$$

- From OT to empirical process (Chizat et al. (2020), Mena and Niles-Weed (2019), Manole and Weed (2021), ...):

$$\begin{aligned} &\mathbb{E}\left[\left| W_2^2(\mu_n, \nu_n) - W_2^2(\mu, \nu) \right|\right] \\ &\qquad \le \mathbb{E}\left[\sup_{f\in\mathcal{F}} \int f \, (d\mu_n - d\nu)\right] + \mathbb{E}\left[\sup_{f\in\mathcal{F}} \int f \, (d\nu_n - d\nu)\right] \end{aligned}$$

- Expected suprema of an empirical process but with respect to dependent data!

Proof ideas: Essential tools

- Three key techniques for our proof is:
  1. Berbee's coupling Theorem (showed few slides before).

Proof ideas: Essential tools

- Three key techniques for our proof is:
  1. Berbee's coupling Theorem (showed few slides before).
  2. Blocking technique of Bernstein. (In a sequence of dependent data, if two blocks are far away, the dependence between them is meager, goes back to (Bernstein, 1927)).

Proof ideas: Essential tools

- Three key techniques for our proof is:
  1. Berbee's coupling Theorem (showed few slides before).
  2. Blocking technique of Bernstein. (In a sequence of dependent data, if two blocks are far away, the dependence between them is meager, goes back to (Bernstein, 1927)).
  3. Chaining method with truncation (for non-Donsker class of function, as integral of log covering number diverges near 0, c.f. (Wainwright, 2019))

Proof ideas: Essential tools

- Three key techniques for our proof is:
  1. Berbee's coupling Theorem (showed few slides before).
  2. Blocking technique of Bernstein. (In a sequence of dependent data, if two blocks are far away, the dependence between them is meager, goes back to (Bernstein, 1927)).
  3. Chaining method with truncation (for non-Donsker class of function, as integral of log covering number diverges near 0, c.f. (Wainwright, 2019))

Proposition: Maximal inequality for finitely many functions

Suppose $\{X_i\}_{1 \le i \le n}$ a stationary sequence and let $\mathcal{F}$ be finite collection of functions with $\|f\|_\infty \le b$ and $\pi_q = \sqrt{4 \sum_{j=0}^{q-1} \beta(j)}$. Then:

$$\mathbb{E}\left[\max_{f \in \mathcal{F}} \left| \frac{1}{\sqrt{n}} \sum_{i=1}^{n} (f(X_i) - Pf) \right| \right] \lesssim b \inf_{1 \le q \le n} \left(\pi_q \sqrt{\log |\mathcal{F}|} + q \frac{\log |\mathcal{F}|}{\sqrt{n}} + \beta(q)\sqrt{n} \right)$$

Proof ideas: A general maximal inequalty

- The bound is:

$$b \inf_{1 \leq q \leq n} \left(\underbrace{\pi_q \sqrt{\log |\mathcal{F}|} + q \frac{\log |\mathcal{F}|}{\sqrt{n}}}_{\uparrow \text{ with } q} + \underbrace{\beta_q \sqrt{n}}_{\downarrow \text{ with } q} \right)$$

Therefore, q should be chosen carefully to balance these terms!

Proof ideas: A general maximal inequalty

- The bound is:

$$b \inf_{1\leq q\leq n}\left(\underbrace{\pi_q\sqrt{\log|\mathcal{F}|}+q\frac{\log|\mathcal{F}|}{\sqrt{n}}}_{\uparrow \text{ with } q}+\underbrace{\beta_q\sqrt{n}}_{\downarrow \text{ with } q}\right)$$

  Therefore, q should be chosen carefully to balance these terms!

- An example: if $\beta(j)\sim j^{-\beta}$ then:

$$b \inf_{1\leq q\leq n}\left(q^{1-\beta}\sqrt{\log|\mathcal{F}|}+q\frac{\log|\mathcal{F}|}{\sqrt{n}}+\beta_q\sqrt{n}\right)$$

Proof ideas: A general maximal inequalty

- The bound is:

$$b \inf_{1 \le q \le n} \left(\underbrace{\pi_q \sqrt{\log |\mathcal{F}|} + q \frac{\log |\mathcal{F}|}{\sqrt{n}}}_{\uparrow \text{ with } q} + \underbrace{\beta_q \sqrt{n}}_{\downarrow \text{ with } q} \right)$$

  Therefore, q should be chosen carefully to balance these terms!

- An example: if $\beta(j) \sim j^{-\beta}$ then:

$$b \inf_{1 \le q \le n} \left(q^{1-\beta} \sqrt{\log |\mathcal{F}|} + q \frac{\log |\mathcal{F}|}{\sqrt{n}} + \beta_q \sqrt{n} \right)$$

  Simple algebra yields:

  1. (i) = (ii) when $q = (n/\log \mathcal{F})^{1/2\beta}$.
  2. (ii) = (iii) when $q = (n/\log \mathcal{F})^{1/(1+\beta)}$
  3. (iii) = (i) when $q = (n/\log \mathcal{F})^{1/2}$.

A theorem for maximal inequality over infinite set

Theorem

Suppose $\mathcal{F}$ be class of function satisfies the following covering number condition:

$$\log \mathcal{N}(\mathcal{F}, \|\cdot\|_\infty, \epsilon) \lesssim \epsilon^{-\alpha} \quad \alpha > 2\,.$$

If $\beta_j \sim j^{-\beta}$ for some $\beta > 0$ then we have:

$$\mathbb{E}\left[\sup_{f \in \mathcal{F}} \left| \int f \; (d\mu_n - \mu) \right| \right] \lesssim n^{-\left(\frac{\beta}{\beta+1} \wedge \frac{1}{2}\right)} + n^{-\frac{1}{\alpha}}\,.$$

A theorem for maximal inequality over infinite set

Theorem

Suppose $\mathcal{F}$ be class of function satisfies the following covering number condition:

$$\log \mathcal{N}(\mathcal{F}, \|\cdot\|_\infty, \epsilon) \lesssim \epsilon^{-\alpha} \quad \alpha > 2\,.$$

If $\beta_j \sim j^{-\beta}$ for some $\beta > 0$ then we have:

$$\mathbb{E}\left[\sup_{f \in \mathcal{F}} \left| \int f \; (d\mu_n - \mu) \right| \right] \lesssim n^{-\left(\frac{\beta}{\beta+1} \wedge \frac{1}{2}\right)} + n^{-\frac{1}{\alpha}}\,.$$

- In case of W_2^2, the value of $\alpha = d/2$ and $\alpha > 2$ for $d > 4$.

A theorem for maximal inequality over infinite set

Theorem

Suppose $\mathcal{F}$ be class of function satisfies the following covering number condition:

$$\log \mathcal{N}(\mathcal{F}, \|\cdot\|_\infty, \epsilon) \lesssim \epsilon^{-\alpha} \quad \alpha > 2\,.$$

If $\beta_j \sim j^{-\beta}$ for some $\beta > 0$ then we have:

$$\mathbb{E}\left[\sup_{f \in \mathcal{F}} \left| \int f \ (d\mu_n - \mu) \right| \right] \lesssim n^{-\left(\frac{\beta}{\beta+1} \wedge \frac{1}{2}\right)} + n^{-\frac{1}{\alpha}}\,.$$

- In case of W_2^2, the value of $\alpha = d/2$ and $\alpha > 2$ for $d > 4$.
- Our proof relies on the techniques developed in a series of work of Doukhan, Massart and Rio (*e.g.* (Rio, 1993), (DMR, 1994), (DMR, 1995), whilst the main difference is that our result generalizes to the case when $\beta < 1$ at the expense of stronger (here $\|\cdot\|_\infty$) norm on the covering number.

Comparison with Yu (1994)

- The exponent $\frac{\beta}{\beta+1}$ is not new/unexpected as it "almost" occurs in Yu (1994).

Comparison with Yu (1994)

- The exponent $\frac{\beta}{\beta+1}$ is not new/unexpected as it "almost" occurs in Yu (1994).
- To be more precise, for $0 < \beta < 1$, (Yu, 1994) obtained a bound of the form

$$o_p(n^{-\frac{s}{s+1}}), \quad \text{for all } 0 < s < \beta$$

when the function class is "small", i.e.,

$$\log \mathcal{N}(\mathcal{F}, \|\cdot\|_\infty, \epsilon) \lesssim -\log \epsilon.$$

Comparison with Yu (1994)

- The exponent $\frac{\beta}{\beta+1}$ is not new/unexpected as it "almost" occurs in Yu (1994).
- To be more precise, for $0 < \beta < 1$, (Yu, 1994) obtained a bound of the form
$$o_p(n^{-\frac{s}{s+1}}), \quad \text{for all } 0 < s < \beta$$
when the function class is "small", i.e.,
$$\log \mathcal{N}(\mathcal{F}, \|\cdot\|_\infty, \epsilon) \lesssim -\log \epsilon.$$
- Three key differences:
  1. Our function classes of interest have larger size
  2. Choosing $s = \beta$, which replaces $o(\cdot)$ by $O(\cdot)$.
  3. Translating the asymptotic bound to bounds on finite sample error bounds

Summary

- Our maximal inequality for β-mixing sequence can be used in various applications, *e.g.* Non-parametric regression, Regularized optimal transport, convergence of optimal transport maps, etc.

Summary

- Our maximal inequality for β-mixing sequence can be used in various applications, *e.g.* Non-parametric regression, Regularized optimal transport, convergence of optimal transport maps, etc. All of these can be related to bounding expected supremum of empirical processes, even under dependence (see Mena and Niles-Weed (2019), Deb et al. (2021), Manole et al. (2021))

Summary

- Our maximal inequality for β-mixing sequence can be used in various applications, *e.g.* Non-parametric regression, Regularized optimal transport, convergence of optimal transport maps, etc. All of these can be related to bounding expected supremum of empirical processes, even under dependence (see Mena and Niles-Weed (2019), Deb et al. (2021), Manole et al. (2021))

- Our analysis indicates that the *threshold* on β (when $\beta(j) \sim j^{-\beta}$), below which we get slower rate (in comparison to i.i.d. setup) relies on the underlying dimension.

Summary

- Our maximal inequality for β-mixing sequence can be used in various applications, *e.g.* Non-parametric regression, Regularized optimal transport, convergence of optimal transport maps, etc. All of these can be related to bounding expected supremum of empirical processes, even under dependence (see Mena and Niles-Weed (2019), Deb et al. (2021), Manole et al. (2021))

- Our analysis indicates that the *threshold* on β (when $\beta(j) \sim j^{-\beta}$), below which we get slower rate (in comparison to i.i.d. setup) relies on the underlying dimension.

- Ongoing work:

  1. Relax the mixing condition to $\alpha(j)$.

  2. Tail bound and asymptotic limit theorem, especially when $\beta < 1$.

Summary

- Our maximal inequality for β-mixing sequence can be used in various applications, *e.g.* Non-parametric regression, Regularized optimal transport, convergence of optimal transport maps, etc. All of these can be related to bounding expected supremum of empirical processes, even under dependence (see Mena and Niles-Weed (2019), Deb et al. (2021), Manole et al. (2021))

- Our analysis indicates that the *threshold* on β (when $\beta(j) \sim j^{-\beta}$), below which we get slower rate (in comparison to i.i.d. setup) relies on the underlying dimension.

- Ongoing work:
  1. Relax the mixing condition to $\alpha(j)$.
  2. Tail bound and asymptotic limit theorem, especially when $\beta < 1$.

Thank you. Questions?