Trade-off Between Dependence and Complexity in Empirical Processes

Nabarun Deb
University of Chicago Booth School of Business
https://nabarund.github.io/

IISA 2024

September 4, 2025

Collaborator

Debarghya Mukherjee, Department of Mathematics and Statistics, Boston University

https://debarghya-mukherjee.github.io/

Maximal Inequalities for empirical processes

- Consider $X_1, X_2, \dots, X_n$ from some distribution μ (not necessarily independent) on $\mathbb{R}^d$
- Define the empirical measure

$$\mu_n = \frac{1}{n} \sum_{i=1}^{n} \delta_{X_i}.$$

Maximal Inequalities for empirical processes

- Consider $X_1, X_2, \ldots, X_n$ from some distribution μ (not necessarily independent) on $\mathbb{R}^d$
- Define the empirical measure

$$\mu_n = \frac{1}{n} \sum_{i=1}^{n} \delta_{X_i}.$$

In particular, $\mathbb{E}_{\mu_n} f = \int f \, d\mu_n = \frac{1}{n} \sum_{i=1}^{n} f(X_i)$.

Maximal Inequalities for empirical processes

- Consider $X_1, X_2, \ldots, X_n$ from some distribution μ (not necessarily independent) on $\mathbb{R}^d$
- Define the empirical measure

$$\mu_n = \frac{1}{n} \sum_{i=1}^{n} \delta_{X_i}.$$

In particular, $\mathbb{E}_{\mu_n} f = \int f \, d\mu_n = \frac{1}{n} \sum_{i=1}^{n} f(X_i).$

- An empirical process is typically

$$\left\{ \int f \, d(\mu_n - \mu) : \; f \in \mathcal{F} \right\}.$$

Maximal Inequalities for empirical processes

- Consider $X_1, X_2, \ldots, X_n$ from some distribution μ (not necessarily independent) on $\mathbb{R}^d$
- Define the empirical measure
$$\mu_n = \frac{1}{n}\sum_{i=1}^{n} \delta_{X_i}.$$
In particular, $\mathbb{E}_{\mu_n} f = \int f \, d\mu_n = \frac{1}{n}\sum_{i=1}^{n} f(X_i)$.
- An empirical process is typically
$$\left\{ \int f \, d(\mu_n - \mu) : \ f \in \mathcal{F} \right\}.$$

Our goal — maximal inequality

Assuming some *mixing conditions*, get an upper bound of
$$\mathbb{E} \sup_{f \in \mathcal{F}} \left| \int f \, d(\mu_n - \mu) \right|.$$

Why do we care?

- Consider $d = 1$ and $\mathcal{F} := \{\mathbf{1}(-\infty, x] : \ x \in \mathbb{R}\}$, then

$$\sup_{f \in \mathcal{F}} \left| \int f \, d(\mu_n - \mu) \right| = \sup_{x \in \mathbb{R}} |F_n(x) - F(x)|,$$

where

$$F_n(x) = \frac{1}{n} \sum_{i=1}^{n} \mathbf{1}(X_i \leq x), \quad \text{and} \quad F(x) = P(X \leq x).$$

Why do we care?

- Consider $d = 1$ and $\mathcal{F} := \{\mathbf{1}(-\infty, x] : \ x \in \mathbb{R}\}$, then

$$\sup_{f \in \mathcal{F}} \left| \int f \, d(\mu_n - \mu) \right| = \sup_{x \in \mathbb{R}} |F_n(x) - F(x)|,$$

  where

$$F_n(x) = \frac{1}{n} \sum_{i=1}^{n} \mathbf{1}(X_i \le x), \quad \text{and} \quad F(x) = P(X \le x).$$

- Applications:
  - Kolmogorov-Smirnov goodness of fit (i.i.d. setting)

$$\sqrt{n} \sup_x |F_n(x) - F(x)| = O_p(1).$$

    Also see *DKW inequality* — Dvoretzsky, Kiefer, Wolfowitz (1956), Massart (1990)

Why do we care?

- Consider $d = 1$ and $\mathcal{F} := \{\mathbf{1}(-\infty, x] : \ x \in \mathbb{R}\}$, then

$$\sup_{f \in \mathcal{F}} \left| \int f \, d(\mu_n - \mu) \right| = \sup_{x \in \mathbb{R}} |F_n(x) - F(x)|,$$

  where

$$F_n(x) = \frac{1}{n} \sum_{i=1}^{n} \mathbf{1}(X_i \leq x), \quad \text{and} \quad F(x) = P(X \leq x).$$

- Applications:
  - Kolmogorov-Smirnov goodness of fit (i.i.d. setting)

$$\sqrt{n} \sup_x |F_n(x) - F(x)| = O_p(1).$$

    Also see *DKW inequality* — Dvoretzky, Kiefer, Wolfowitz (1956), Massart (1990)
  - Extensions to two-sample testing, independence testing, etc.
  - Multivariate extensions with coordinatewise ordering Naaman (2021)

Other applications (i.i.d. case)

- Nonparametric least squares regression

$$Y_i = f^*(X_i) + \epsilon_i, \quad \mathbb{E}[\epsilon_i | X_i] = 0.$$

Estimate f^* using

$$\hat{f}_n = \arg\min_{f \in \mathcal{F}} \sum_{i=1}^{n} (Y_i - f(X_i))^2.$$

Maximal inequalities govern $\frac{1}{n} \sum_{i=1}^{n} (\hat{f}_n(X_i) - f^*(X_i))^2$ (see Vaart and Wellner (1996), Sara van de Geer (2009))

Other applications (i.i.d. case)

- Nonparametric least squares regression

$$Y_i = f^*(X_i) + \epsilon_i, \quad \mathbb{E}[\epsilon_i | X_i] = 0.$$

  Estimate f^* using

$$\hat{f}_n = \arg\min_{f \in \mathcal{F}} \sum_{i=1}^{n} (Y_i - f(X_i))^2.$$

  Maximal inequalities govern $\frac{1}{n}\sum_{i=1}^{n}(\hat{f}_n(X_i) - f^*(X_i))^2$ (see Vaart and Wellner (1996), Sara van de Geer (2009))

- Function fitting with non convex optimization such as deep neural nets (Schmidt-Hieber (2020), Ohn and Kim (2022))

Other applications (i.i.d. case)

- Nonparametric least squares regression

$$Y_i = f^*(X_i) + \epsilon_i, \quad \mathbb{E}[\epsilon_i | X_i] = 0.$$

  Estimate f^* using

$$\hat{f}_n = \arg\min_{f \in \mathcal{F}} \sum_{i=1}^{n} (Y_i - f(X_i))^2.$$

  Maximal inequalities govern $\frac{1}{n}\sum_{i=1}^{n}(\hat{f}_n(X_i) - f^*(X_i))^2$ (see Vaart and Wellner (1996), Sara van de Geer (2009))

- Function fitting with non convex optimization such as deep neural nets (Schmidt-Hieber (2020), Ohn and Kim (2022))
- Optimal transport distance and map estimation (see Hütter and Rigollet (2021), Manole and Weed (2021), Deb, Ghosal, and Sen (2021))

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)

Some related work

- Nonparametric least squares under mixing conditions (see Mohri and Rostamizadeh (2008), Zhang, Cao, and Yan (2012), Roy, Balasubramanian, and Erdogdu (2021))
- Function fitting with deep neural nets under mixing conditions (see Ma and Safikhani (2022), Kengne and Modou (2023), Kurisu, Fukami, and Koike (2023))
- "Wasserstein" distance (optimal transport) estimation under mixing conditions (see Fournier and Guillin (2015), Bernton et al. (2019), Cazelles et al. (2020))

Some related work

- Nonparametric least squares under mixing conditions (see Mohri and Rostamizadeh (2008), Zhang, Cao, and Yan (2012), Roy, Balasubramanian, and Erdogdu (2021))
- Function fitting with deep neural nets under mixing conditions (see Ma and Safikhani (2022), Kengne and Modou (2023), Kurisu, Fukami, and Koike (2023))
- "Wasserstein" distance (optimal transport) estimation under mixing conditions (see Fournier and Guillin (2015), Bernton et al. (2019), Cazelles et al. (2020))

In this talk ...

— Most existing work focuses on exponentially fast mixing or simple function classes $\mathcal{F}$

— We focus on much stronger dependence (including sub-polynomial mixing) and complex function classes. We examine if i.i.d. like rates can still be recovered

Outline

1. General empirical process bounds
   - Main mixing assumptions — Formal Problem Statement
   - Long and Short Range Dependence
   - General maximal inequalities
   - Proof ideas
2. Shape restricted convex regression
   - Bounded convex Least squares (LS) estimator
   - Faster rates and localization
3. Conclusion

Outline

Outline

Notions of mixing for dependence

- Given a strictly stationary sequence of random variables $\{X_t\}_{t\in\mathbb{N}}$ on a probability space $(\Omega, \mathcal{A}, \mathbb{P})$

Notions of mixing for dependence

- Given a strictly stationary sequence of random variables $\{X_t\}_{t\in\mathbb{N}}$ on a probability space $(\Omega, \mathcal{A}, \mathbb{P})$
- 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 mixing for dependence

- Given a strictly stationary sequence of random variables $\{X_t\}_{t\in\mathbb{N}}$ on a probability space $(\Omega, \mathcal{A}, \mathbb{P})$
- 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 mixing for dependence

- Given a strictly stationary sequence of random variables $\{X_t\}_{t\in\mathbb{N}}$ on a probability space $(\Omega, \mathcal{A}, \mathbb{P})$
- 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 mixing for dependence

- Given a strictly stationary sequence of random variables $\{X_t\}_{t\in\mathbb{N}}$ on a probability space $(\Omega, \mathcal{A}, \mathbb{P})$
- 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:

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

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

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

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.

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.
- Also note that expected supremum of empirical processes don't just depend on covariance bounds but on the "size" of the function class

Outline

General maximal inequality with bracketing

- Recall our goal: To bound

$$\mathbb{E} \sup_{f \in \mathcal{F}} \left| \int f \, d(\mu_n - \mu) \right|$$

General maximal inequality with bracketing

- Recall our goal: To bound

$$\mathbb{E} \sup_{f \in \mathcal{F}} \left| \int f \, d(\mu_n - \mu) \right|$$

- Size of $\mathcal{F}$: Bracketing number $N(u, \|\cdot\|, \mathcal{F})$ is the number of pairs $[L_j, U_j]$ of functions such that $\|U_j - L_j\| \le u$ and given any $f \in \mathcal{F}$, there exists j_f satisfying

$$L_{j_f} \le f \le U_{j_f}$$

General maximal inequality with bracketing

- Recall our goal: To bound

$$\mathbb{E} \sup_{f \in \mathcal{F}} \left| \int f \, d(\mu_n - \mu) \right|$$

- Size of $\mathcal{F}$: Bracketing number $N(u, \|\cdot\|, \mathcal{F})$ is the number of pairs $[L_j, U_j]$ of functions such that $\|U_j - L_j\| \leq u$ and given any $f \in \mathcal{F}$, there exists j_f satisfying

$$L_{j_f} \leq f \leq U_{j_f}$$

- An important function on the space of positive integers

$$\Lambda(q) := \sum_{k=0}^{q-1} \beta_k.$$

Maximal inequality with L_∞ bracketing

Given $u > 0$, solve the following equation on positive integers:

$$\beta(q) \approx \frac{q}{n}(1 + \log N(u, \mathcal{F}, \|\cdot\|_\infty))$$

to get $q_n(u)$.

Maximal inequality with L_∞ bracketing

Given $u > 0$, solve the following equation on positive integers:

$$\beta(q) \approx \frac{q}{n}(1 + \log N(u, \mathcal{F}, \|\cdot\|_\infty))$$

to get $q_n(u)$.

Informal bound

Suppose $\mathcal{F}$ has a L_∞ diameter σ (bounded above and below in n), then

$$\mathbb{E} \sup_{f \in \mathcal{F}} \left| \int f \, d(\mu_n - \mu) \right| \lesssim n^{-1/2} a,$$

where

$$a \geq \int_{\frac{a}{\sqrt{n}}}^{\sigma} \sqrt{\Lambda(q_n(u)) \log N(u, \mathcal{F}, \|\cdot\|_\infty)} \, du$$

For i.i.d. data $q_n(u) = 1$, $\Lambda(q_n(u)) = 1$ and we get back usual bound with integral of square root of log bracketing number

Maximal inequality with L_r bracketing, $r > 2$

- Given $u > 0$, the definition of $q_n(u)$ stays the same with $\|\cdot\|_\infty$ replaced with $\|\cdot\|_r$.

Maximal inequality with L_r bracketing, $r > 2$

- Given $u > 0$, the definition of $q_n(u)$ stays the same with $\|\cdot\|_\infty$ replaced with $\|\cdot\|_r$.
- Consider
$$\Lambda_r(q) := \sum_{k=0}^{q-1} \beta_k^{1-\frac{2}{r}}.$$

Maximal inequality with L_r bracketing, $r > 2$

- Given $u > 0$, the definition of $q_n(u)$ stays the same with $\|\cdot\|_\infty$ replaced with $\|\cdot\|_r$.
- Consider

$$\Lambda_r(q) := \sum_{k=0}^{q-1} \beta_k^{1-\frac{2}{r}}.$$

Informal bound

Suppose $\mathcal{F}$ has a L_r diameter σ (bounded above and below in n), then

$$\mathbb{E} \sup_{f \in \mathcal{F}} \left| \int f \, d(\mu_n - \mu) \right| \lesssim n^{-1/2} a,$$

where

$$a \geq \int_{\frac{a}{\sqrt{n}}}^{\sigma} \sqrt{\Lambda_r(q_n(u)) \log N(u, \mathcal{F}, \|\cdot\|_\infty)} \, du$$

Maximal inequality with L_r bracketing, $r > 2$

- Given $u > 0$, the definition of $q_n(u)$ stays the same with $\|\cdot\|_\infty$ replaced with $\|\cdot\|_r$.
- Consider
$$\Lambda_r(q) := \sum_{k=0}^{q-1} \beta_k^{1-\frac{2}{r}}.$$

Informal bound

Suppose $\mathcal{F}$ has a L_r diameter σ (bounded above and below in n), then

$$\mathbb{E} \sup_{f \in \mathcal{F}} \left| \int f \, d(\mu_n - \mu) \right| \lesssim n^{-1/2} a,$$

where

$$a \geq \int_{\frac{a}{\sqrt{n}}}^{\sigma} \sqrt{\Lambda_r(q_n(u)) \log N(u, \mathcal{F}, \|\cdot\|_\infty)} \, du$$

Note the degeneracy for $r = 2$. We will come back to this.

Example

- Suppose $\alpha > 2$ and $\mathcal{F}$ is a class of functions satisfying

$$\log N(u, \mathcal{F}, \|\cdot\|_\infty) \lesssim u^{-\frac{1}{\alpha}}.$$

Further assume $\beta_k \leq (1+k)^{-\beta}$ for some $\beta > 0$

Example

- Suppose $\alpha > 2$ and $\mathcal{F}$ is a class of functions satisfying
$$\log N(u, \mathcal{F}, \|\cdot\|_\infty) \lesssim u^{-\frac{1}{\alpha}}.$$
Further assume $\beta_k \le (1+k)^{-\beta}$ for some $\beta > 0$
- This will imply
$$q_n(u) = (nu^{\alpha})^{\frac{1}{1+\beta}}.$$

Example

- Suppose $\alpha > 2$ and $\mathcal{F}$ is a class of functions satisfying

$$\log N(u, \mathcal{F}, \|\cdot\|_\infty) \lesssim u^{-\frac{1}{\alpha}}.$$

Further assume $\beta_k \leq (1+k)^{-\beta}$ for some $\beta > 0$

- This will imply

$$q_n(u) = (nu^\alpha)^{\frac{1}{1+\beta}}.$$

- Plugging into the previous theorem gives (for $d \geq 2s+1$),

$$\frac{1}{n}\sum_{i=1}^n \mathbb{E}(\hat{f}_n(X_i) - f^*(X_i))^2 \lesssim \begin{cases} n^{-\frac{1}{\alpha}} & \text{if } \beta > \frac{1}{\alpha-1} \\ n^{-\frac{\beta}{\beta+1}} & \text{otherwise} \end{cases}.$$

Example

- Suppose $\alpha > 2$ and $\mathcal{F}$ is a class of functions satisfying

$$\log N(u, \mathcal{F}, \|\cdot\|_\infty) \lesssim u^{-\frac{1}{\alpha}}.$$

Further assume $\beta_k \leq (1+k)^{-\beta}$ for some $\beta > 0$

- This will imply

$$q_n(u) = (nu^\alpha)^{\frac{1}{1+\beta}}.$$

- Plugging into the previous theorem gives (for $d \geq 2s+1$),

$$\frac{1}{n}\sum_{i=1}^{n} \mathbb{E}(\hat{f}_n(X_i) - f^*(X_i))^2 \lesssim \begin{cases} n^{-\frac{1}{\alpha}} & \text{if } \beta > \frac{1}{\alpha-1} \\ n^{-\frac{\beta}{\beta+1}} & \text{otherwise} \end{cases}.$$

Potential optimality

- The $n^{-\frac{1}{\alpha}}$ rate is not improvable in general; Birge and Massart, 1993

Example

- Suppose $\alpha > 2$ and $\mathcal{F}$ is a class of functions satisfying

$$\log N(u, \mathcal{F}, \|\cdot\|_\infty) \lesssim u^{-\frac{1}{\alpha}}.$$

Further assume $\beta_k \leq (1+k)^{-\beta}$ for some $\beta > 0$

- This will imply

$$q_n(u) = (nu^\alpha)^{\frac{1}{1+\beta}}.$$

- Plugging into the previous theorem gives (for $d \geq 2s+1$),

$$\frac{1}{n}\sum_{i=1}^{n}\mathbb{E}(\hat{f}_n(X_i) - f^*(X_i))^2 \lesssim \begin{cases} n^{-\frac{1}{\alpha}} & \text{if } \beta > \frac{1}{\alpha-1} \\ n^{-\frac{\beta}{\beta+1}} & \text{otherwise} \end{cases}.$$

Potential optimality

- The $n^{-\frac{1}{\alpha}}$ rate is not improvable in general; Birge and Massart, 1993
- If $\alpha > 2$ then in the long range dependence regime $(1/(\alpha-1), 1)$, we get the optimal $n^{-\frac{1}{\alpha}}$ rates

Outline

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 adaptive truncation (for non-Donsker class of function, as integral of log bracketing number diverges near 0, c.f. Ossiander (1987), Pollard (2002)

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 adaptive truncation (for non-Donsker class of function, as integral of log bracketing number diverges near 0, c.f. Ossiander (1987), Pollard (2002)
- Our proof relies on the techniques developed in a series of works by Doukhan, Massart and Rio (*e.g.* Rio (1993), DMR (1994, 1995)), whilst the main difference is that our result generalizes to the case when $\beta < 1$

Outline

Outline

An illustration: Multivariate convex regression

- Consider the least squares regression with stationary β-mixing data, $(X_1, Y_1), \ldots, (X_n, Y_n)$, assume compact (polytopal) supports. Goal is to estimate $f^*(x) = E[Y|X = x] \in \mathcal{F}$ with the estimator

$$\hat{f}_n := \arg\min_{f \in \mathcal{F}} \sum_{i=1}^{n} (Y_i - f(X_i))^2.$$

An illustration: Multivariate convex regression

- Consider the least squares regression with stationary β-mixing data, $(X_1, Y_1), \ldots, (X_n, Y_n)$, assume compact (polytopal) supports. Goal is to estimate $f^*(x) = E[Y|X = x] \in \mathcal{F}$ with the estimator

$$\hat{f}_n := \arg\min_{f \in \mathcal{F}} \sum_{i=1}^{n} (Y_i - f(X_i))^2.$$

- Suppose $\mathcal{F}$ is the class of convex functions on $\mathbb{R}^d$ for $d \geq 5$ which are bounded by 1, then

$$\log N(u, \mathcal{F}, \|\cdot\|_r) \lesssim C_r u^{-\frac{d}{2}}.$$

An illustration: Multivariate convex regression

- Consider the least squares regression with stationary β-mixing data, $(X_1, Y_1), \ldots, (X_n, Y_n)$, assume compact (polytopal) supports. Goal is to estimate $f^*(x) = E[Y|X = x] \in \mathcal{F}$ with the estimator

$$\hat{f}_n := \arg\min_{f \in \mathcal{F}} \sum_{i=1}^{n} (Y_i - f(X_i))^2.$$

- Suppose $\mathcal{F}$ is the class of convex functions on $\mathbb{R}^d$ for $d \geq 5$ which are bounded by 1, then

$$\log N(u, \mathcal{F}, \|\cdot\|_r) \lesssim C_r u^{-\frac{d}{2}}.$$

- Plugging into the previous theorem gives (for $d \geq 5$),

$$\mathbb{E}(\hat{f}_n(X) - f^*(X))^2 \lesssim \begin{cases} n^{-\frac{2}{d}} & \text{if } \beta > \frac{2}{d-2} \\ n^{-\frac{\beta}{\beta+1}} & \text{otherwise} \end{cases}.$$

An illustration: Multivariate convex regression

- Consider the least squares regression with stationary β-mixing data, $(X_1, Y_1), \ldots, (X_n, Y_n)$, assume compact (polytopal) supports. Goal is to estimate $f^*(x) = E[Y|X = x] \in \mathcal{F}$ with the estimator

$$\hat{f}_n := \arg\min_{f \in \mathcal{F}} \sum_{i=1}^{n} (Y_i - f(X_i))^2.$$

- Suppose $\mathcal{F}$ is the class of convex functions on $\mathbb{R}^d$ for $d \geq 5$ which are bounded by 1, then

$$\log N(u, \mathcal{F}, \|\cdot\|_r) \lesssim C_r u^{-\frac{d}{2}}.$$

- Plugging into the previous theorem gives (for $d \geq 5$),

$$\mathbb{E}(\hat{f}_n(X) - f^*(X))^2 \lesssim \begin{cases} n^{-\frac{2}{d}} & \text{if } \beta > \frac{2}{d-2} \\ n^{-\frac{\beta}{\beta+1}} & \text{otherwise} \end{cases}.$$

- Rate is not improvable for LS estimator even under independence

Outline

Is this rate improvable?

- Bounded convex LS estimator enjoys some tuning-free adaptation when f^* is affine, in the i.i.d. setting

Is this rate improvable?

- Bounded convex LS estimator enjoys some tuning-free adaptation when f^* is affine, in the i.i.d. setting
- The rate of convergence is

$$\mathbb{E}(\hat{f}_n(X) - f^*(X))^2 \lesssim n^{-\frac{4}{d}}$$

for $d > 4$

Is this rate improvable?

- Bounded convex LS estimator enjoys some tuning-free adaptation when f^* is affine, in the i.i.d. setting
- The rate of convergence is

$$\mathbb{E}(\hat{f}_n(X) - f^*(X))^2 \lesssim n^{-\frac{4}{d}}$$

  for $d > 4$

- This rate is known to not be improvable

Is this rate improvable?

- Bounded convex LS estimator enjoys some tuning-free adaptation when f^* is affine, in the i.i.d. setting
- The rate of convergence is

$$\mathbb{E}(\hat{f}_n(X) - f^*(X))^2 \lesssim n^{-\frac{4}{d}}$$

  for $d > 4$

- This rate is known to not be improvable
- The rate comes from solving the following equation:

$$\delta_n^2 \sim \mathbb{E} \sup_{f \in \mathcal{F}:\ \|f - f^*\|_{L_2} \le \delta_n} \left| \int d\, d(\mu_n - \mu) \right|$$

Is this rate improvable?

- Bounded convex LS estimator enjoys some tuning-free adaptation when f^* is affine, in the i.i.d. setting
- The rate of convergence is

$$\mathbb{E}(\hat{f}_n(X) - f^*(X))^2 \lesssim n^{-\frac{4}{d}}$$

for $d > 4$

- This rate is known to not be improvable
- The rate comes from solving the following equation:

$$\delta_n^2 \sim \mathbb{E} \sup_{f \in \mathcal{F}:\ \|f - f^*\|_{L_2} \leq \delta_n} \left| \int d\, d(\mu_n - \mu) \right|$$

- *Note the occurence of L_2 norm which is not covered by our earlier result*

Maximal inequality with L_2 bracketing

- Stronger mixing condition $\gamma_k = \beta_k \vee \rho_k$

Maximal inequality with L_2 bracketing

- Stronger mixing condition $\gamma_k = \beta_k \vee \rho_k$
- Given $u > 0$, the definition of $q_n(u)$ stays the same with $\|\cdot\|_\infty$ replaced with $\|\cdot\|_2$ and β_k with γ_k.

Maximal inequality with L_2 bracketing

- Stronger mixing condition $\gamma_k = \beta_k \vee \rho_k$
- Given $u > 0$, the definition of $q_n(u)$ stays the same with $\|\cdot\|_\infty$ replaced with $\|\cdot\|_2$ and β_k with γ_k.
- Consider

$$\Lambda_2(q) := \sum_{k=0}^{q-1} \gamma_k.$$

Maximal inequality with L_2 bracketing

- Stronger mixing condition $\gamma_k = \beta_k \vee \rho_k$
- Given $u > 0$, the definition of $q_n(u)$ stays the same with $\|\cdot\|_\infty$ replaced with $\|\cdot\|_2$ and β_k with γ_k.
- Consider

$$\Lambda_2(q) := \sum_{k=0}^{q-1} \gamma_k.$$

Informal bound

Suppose $\mathcal{F}$ has a L_2 diameter σ (bounded above and below in n), then

$$\mathbb{E} \sup_{f \in \mathcal{F}} \left| \int f \, d(\mu_n - \mu) \right| \lesssim n^{-1/2} a,$$

where

$$a \geq \int_{\frac{a}{\sqrt{n}}}^{\sigma} \sqrt{\Lambda_2(q_n(u)) \log N(u, \mathcal{F}, \|\cdot\|_2)} \, du$$

Faster rates with stronger mixing

- Assume stronger mixing $\gamma_k = \beta_k \vee \rho_k \lesssim (1+k)^{-\gamma}$

Faster rates with stronger mixing

- Assume stronger mixing $\gamma_k = \beta_k \vee \rho_k \lesssim (1+k)^{-\gamma}$
- Then we can provide a bound for localized empirical processes with respect to L_2-norm

Faster rates with stronger mixing

- Assume stronger mixing $\gamma_k = \beta_k \vee \rho_k \lesssim (1+k)^{-\gamma}$
- Then we can provide a bound for localized empirical processes with respect to L_2-norm

Rates for adaptation

Consider the multivariate shape-restricted regression setting from before. Suppose that f^* is k-piece affine, i.e., there exists k simplices in dimension d such that f^* is affine on all of them. Then under the stronger mixing assumption, we have:

$$\mathbb{E}(\hat{f}_n(X) - f^*(X))^2 \lesssim n^{-\frac{4}{d}}$$

for $d > 4(1+\gamma^{-1})$

Faster rates with stronger mixing

- Assume stronger mixing $\gamma_k = \beta_k \vee \rho_k \lesssim (1+k)^{-\gamma}$
- Then we can provide a bound for localized empirical processes with respect to L_2-norm

Rates for adaptation

Consider the multivariate shape-restricted regression setting from before. Suppose that f^* is k-piece affine, i.e., there exists k simplices in dimension d such that f^* is affine on all of them. Then under the stronger mixing assumption, we have:

$$\mathbb{E}(\hat{f}_n(X) - f^*(X))^2 \lesssim n^{-\frac{4}{d}}$$

for $d > 4(1+\gamma^{-1})$

In particular, if $d > 8$, then there exists an interval in the long range dependence regime $(4/(d-4), 1)$ where optimal i.i.d. like rates are recovered

Outline

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

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

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

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

  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 $t = \beta$, which replaces $o(\cdot)$ by $O(\cdot)$.
  3. Translating the asymptotic bound to bounds on finite sample error bounds

Summary

- Our maximal inequalities can be used in various applications, *e.g.*
  - Non-parametric regression with adaptation
  - Regularized and unregularized optimal transport
  - Function fitting with deep neural nets in both low and high dimensions
  - Classification under non-convex loss function

Summary

- Our maximal inequalities can be used in various applications, *e.g.*
  - Non-parametric regression with adaptation
  - Regularized and unregularized optimal transport
  - Function fitting with deep neural nets in both low and high dimensions
  - Classification under non-convex loss function

- Our analysis indicates a new *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/complexity of function classes.

Summary

- Our maximal inequalities can be used in various applications, *e.g.*
  - Non-parametric regression with adaptation
  - Regularized and unregularized optimal transport
  - Function fitting with deep neural nets in both low and high dimensions
  - Classification under non-convex loss function

- Our analysis indicates a new *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/complexity of function classes.

- Ongoing work:
  1. Relax the mixing condition to $\alpha(j)$ (strong mixing).
  2. Tail bound and asymptotic limit theorem, especially when $\beta < 1$.
  3. Improve localization bounds
  4. Minimax lower bounds

Summary

- Our maximal inequalities can be used in various applications, *e.g.*
  - Non-parametric regression with adaptation
  - Regularized and unregularized optimal transport
  - Function fitting with deep neural nets in both low and high dimensions
  - Classification under non-convex loss function

- Our analysis indicates a new *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/complexity of function classes.

- Ongoing work:
  1. Relax the mixing condition to $\alpha(j)$ (strong mixing).
  2. Tail bound and asymptotic limit theorem, especially when $\beta < 1$.
  3. Improve localization bounds
  4. Minimax lower bounds

Thank you. Questions?