Refined Maximal Inequalities: Some Questions and Answers

Arun Kumar Kuchibhotla

2 July, 2022

Carnegie Mellon University

Table of contents

Maximal Inequalities

Maximal Inequalities

- In this talk, maximal inequalities refer to bounds on quantities like $\mathbb{E}[\sup_{t\in T} X_t]$ for a stochastic process $\{X_t\}_{t\in T}$.
- This expected supremum arises naturally in finding rates of convergence of empirical risk minimizers, including the least squares estimator.
- Formally, suppose $(X_i, Y_i), 1 \le i \le n$ are iid with $Y_i = f_0(X_i) + \varepsilon_i$ $(\mathbb{E}[\varepsilon_i|X_i] = 0)$. For a function class $\mathcal{F}$, set

$$\widehat{f}_n \;:=\; \operatorname*{argmin}_{f\in\mathcal{F}} \sum_{i=1}^{n} (Y_i - f(X_i))^2.$$

- It is well-known that $\|\widehat{f}_n - f_0\|_2 = O_p(r_n)$ for r_n satisfying $\phi_n(r_n) \le \sqrt{n} r_n^2$, where

$$\phi_n(\delta) = \mathbb{E}\left[\sup_{\|f-f_0\|_2\le\delta} \left|\frac{1}{\sqrt{n}}\sum_{i=1}^{n} \varepsilon_i (f - f_0)(X_i)\right|\right],$$

which is a expected supremum of a stochastic process.

Chaining for Maximal Inequalities

- One of the classical techniques for bounding the expected supremum is chaining. There are two prominent chaining version: Dudley's chaining and generic chaining.
- The general idea is as follows: construct sets $T_0 \subset T_1 \subset \cdots \subset T$ such that T_j is a finite cardinality set for every $j < \infty$.
- For any $t \in T$, let $\pi_j(t) \in T_j$ denote an element of T_j that is closest to t in T_j. Then we get

$$X_t = X_{\pi_0(t)} + \sum_{j=1}^{\infty} (X_{\pi_j(t)} - X_{\pi_{j-1}(t)}).$$

- Taking the supremum over all $t \in T$ and the expectation, we get

$$\mathbb{E}\left[\sup_{t \in T} X_t\right] \le \mathbb{E}\left[\sup_{s \in T_0} X_s\right] + \sum_{j=1}^{\infty} \mathbb{E}\left[\sup_{t \in T} |X_{\pi_j(t)} - X_{\pi_{j-1}(t)}|\right].$$

- For each fixed $j \ge 1$, the right hand side term is expected maximum of finite number of random variables.

Chaining for Maximal Inequalities

- As shown for any sequence of nested sets $T_0 \subset T_1 \subset \cdots \subset T$,

$$\mathbb{E}\left[\sup_{t\in T} X_t\right] \le \mathbb{E}\left[\sup_{s\in T_0} X_s\right] + \sum_{j=1}^{\infty} \mathbb{E}\left[\sup_{t\in T} |X_{\pi_j(t)} - X_{\pi_{j-1}(t)}|\right].$$

- If the nested sets T_j's are chosen so as to ensure $\sup_{t\in T} \|X_t - X_{\pi_j(t)}\|_2 \le 2^{-j}$, then we get Dudley's chaining.
- If the nested sets T_j's are chosen so that $\mathrm{Card}(T_j) \le 2^{2^j}$, then we get generic chaining.
- For some specific stochastic processes $X_t, t \in T$, it is known that the bound obtained via generic chaining cannot be improved. There are examples where Dudley's chaining is sub-optimal.
- Irrespective of the optimality, the chaining methods reduce the problem of controlling expected supremum of a stochastic process to that of expected maximum of a finite number of random variables.

Chaining for Maximal Inequalities

- Recall for any sequence of nested sets $T_0 \subset T_1 \subset \cdots \subset T$,

$$\mathbb{E}\left[\sup_{t \in T} X_t\right] \le \mathbb{E}\left[\sup_{s \in T_0} X_s\right] + \sum_{j=1}^{\infty} \mathbb{E}\left[\sup_{t \in T} |X_{\pi_j(t)} - X_{\pi_{j-1}(t)}|\right].$$

- To control the expected maximum's on the right hand side, one of the common assumptions used is sub-Gaussian increments, i.e., for some distance measure $d(\cdot,\cdot)$, $\|X_t - X_s\|_{\psi_2} \le Cd(s,t)$ $\forall s,t \in T$.
- Exponential moment control is what is more important here than sub-Gaussianity. Such exponential moment control allows one to conclude that

$$\begin{aligned}
&\mathbb{E}\left[\sup_{t \in T} |X_{\pi_j(t)} - X_{\pi_{j-1}(t)}|\right] \\
&\le \max_{t \in T} d_1(t, \pi_j(t))\sqrt{\log(|T_j|)} + \max_{t \in T} d_2(t, \pi_j(t))(\log(|T_j|))^{\beta},
\end{aligned}$$

for some distance measure d_1, d_2 and some $\beta > 0$.

Logarithmic dependence on $|T_j|$ implies good bounds.

Maximal Inequality for Finite Maximum

Finite Maximums

- Exponential moment control for increments of the stochastic process implies logarithmic dependence on $|T_j|$. The converse is not true.
- For example, in case of the least squares estimator, $X_f = n^{-1/2} \sum_{i=1}^{n} \varepsilon_i (f - f_0)(X_i)$ and even if ε only has $2 + \eta$ moments, one can obtain logarithmic dependence.
- The reason simply is that ε_i is a common factor when considering the supremum over f and a simple truncation arguments yields this result.
- Are there general maximal inequalities for finite maximums that yield logarithmic dependence?

Finite Maximums

Theorem (K. and Patra (2022, AoS))
Suppose $X_1, \ldots, X_n$ are iid random variables in some measurable space $\mathcal{X}$ and $f_1, \ldots, f_N$ are arbitrary mean zero functions from $\mathcal{X}$ to $\mathbb{R}$ with $\|f_j\|_2 \le \delta$. Then for any $q \ge 2$,

$$\mathbb{E}\left[\max_{1\le j\le N}\left|\frac{1}{\sqrt{n}}\sum_{i=1}^{n} f_j(X_i)\right|\right] \le \delta\sqrt{6\log(2N)} + \frac{2^{1/2+1/q}}{n^{1/2-1/q}}(3\log(2N))^{1-1/q}\|F\|_q,$$

where $F(x) = \max_{1\le j\le N}|f_j(x)|$.

- Always a logarithmic dependence on N as long as the envelope F has finite q-th moment.
- Even with $q = 2$, this bound implies a rate of $\sqrt{\log(N)}$ which is the optimal dependence under Gaussianity.
- This bound improves upon a result of Chernozhukov et al. (2015).
- For motivation, recall

$$\mathbb{E}\left[\sup_{\|f-f_0\|_2\le\delta}\left|\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\varepsilon_i(f-f_0)(X_i)\right|\right]$$

Finite maximums: optimality

- Is the bound provided optimal?
- Set $\Delta_n = \max_{1\le j\le N} |n^{-1/2}\sum_{i=1}^n f_j(X_i)|$. Recall the bound

$$\mathbb{E}\left[\Delta_n\right] \;\lesssim\; \delta\sqrt{\log(N)} + \frac{(\log(N))^{1-1/q}}{n^{1/2-1/q}}\|F\|_q. \qquad (1)$$

- **Answer:** No! In general, this bound is not optimal.
- For $q=\infty$, this bound is provably sub-optimal. In this case (1) becomes

$$\mathbb{E}[\Delta_n] \;\lesssim\; \delta\sqrt{\log(N)} + \frac{\log(N)}{\sqrt{n}}\|F\|_\infty.$$

- The sub-optimality can be seen easily by noting that (1) for $q=\infty$ can be obtained via Bernstein's inequality which can be improved via Bennett's inequality:

$$\mathbb{E}\left[\Delta_n\right] \;\lesssim\; \delta\sqrt{\log(N)} + \frac{\|F\|_\infty \log(N)/\sqrt{n}}{\log(3\|F\|_\infty^2\log(2N)/(n\delta^2)\vee 1}, \qquad (2)$$

which is as good as (1).

Some problems in maximal inequalities

Finite maximums: optimal bound for $q = \infty$

- Even with this correction based on Bennett's inequality, one cannot claim optimality for each collection of f_j's.
- The formulation of optimality requires certain uniformity over a collection of f_j's.
- Define $\Delta_n(\{f_j\}) = \max_{1\le j\le N} |n^{-1/2}\sum_{i=1}^n f_j(X_i)|$ and
  $\mathcal{E}^{\circ}_{\infty}(A, B) = \sup\{\mathbb{E}[\Delta_n(\{f_j\})] : \mathbb{E}[f_j(X_i)] = 0, \|f_j\|_2 \le A, \|f_j\|_\infty \le B\}$.
- Note that $\mathcal{E}^{\circ}_{\infty}(A, B)$ depends on n, N, A, B.
- Using the optimality of Bennett's inequality over all bounded random variables (Major, 2005, Prob. Surveys), one can show that the proposed bound before is optimal for $\mathcal{E}^{\circ}_{\infty}(A, B)$.
- But bounded random variables are sub-exponential and not the most interesting practical case.

Finite maximums: optimality for $q < \infty$

- Recall $\Delta_n(\{f_j\}) = \max_{1 \le j \le N} |n^{-1/2} \sum_{i=1}^n f_j(X_i)|$.
- With $F(x) = \max_{1 \le j \le N} |f_j(x)|$. Define

  $$\mathcal{E}_q^\circ(A, B) = \sup\{\mathbb{E}[\Delta_n(\{f_j\})] : \mathbb{E}[f_j(X_i)] = 0, \|f_j\|_2 \le A, \|F\|_q \le B\}.$$

- $\mathcal{E}_q^\circ$ is the largest expected value when given variance bound on individual functions and L_q control on the envelope.
- We know $\mathcal{E}_q^\circ(A, B) \lesssim A\sqrt{\log(N)} + B(\log(N))^{1-1/q}/n^{1/2-1/q}$.
- This bound is already logarithmic in N. But this cannot be optimal as $q \to \infty$. What is the optimal bound?
- Answer currently unknown. Using some classical results, one can obtain some reductions.

Reductions for Optimality

- Recall $\Delta_n(\{f_j\}) = \max_{1\le j\le N} |n^{-1/2}\sum_{i=1}^n f_j(X_i)|$ and $F(x) = \max_{1\le j\le N}|f_j(x)|$.
- Define

$$\mathcal{E}^{\circ}_{q,\infty}(A, B_q, B_\infty) = \sup\{\mathbb{E}[\Delta_n(\{f_j\})] : \\ \mathbb{E}[f_j(X_i)] = 0,\ \|f_j\|_2 \le A,\ \|F\|_k \le B_k,\ k = q,\infty\}.$$

- In comparison to $\mathcal{E}^{\circ}_q$, $\mathcal{E}^{\circ}_{q,\infty}$ has an additional control on $\|F\|_\infty$.
- It can be proved using results of de la Pena and Gine (1999) that there exists $T_q(A, B)$ such that

$$\mathcal{E}^{\circ}_q(A) \asymp \mathcal{E}^{\circ}_{q,\infty}(A, B, T_q(A, B)) + \mathcal{M}_q(A, B),$$

where

$$\mathcal{M}_q(A, B) = \sup\left\{\mathbb{E}\left[\max_{1\le i\le n} F(X_i)\right] : \mathbb{E}[f_j(X_i)] = 0,\ \|f_j\|_2 \le A,\ \|F\|_q \le B\right\}$$

Conclusions

Conclusions

- Maximal inequalities are in shortage for heavy-tailed data. Such maximal inequalities yield better understanding of ERMs under weaker assumptions.
- In the context of non-parametric least squares estimator, the proposed maximal inequalities for finite maximums yield new results under heavy-tailed data.
- The study of optimal maximal inequalities is non-existent to the best of the author's knowledge. Some results due to Pinelis do exist for smooth Banach spaces which do not readily apply to the problem at hand.
- Optimal maximal inequalities for finite maximums can pave way for deriving optimal maximal inequalities for supremum of empirical processes.

References

- Chernozhukov, V., Chetverikov, D., and Kato, K. (2015). Comparison and anti-concentration bounds for maxima of Gaussian random vectors. *Prob. Theory Related Fields, 162(1-2):47–70*.
- de la Pena, V. H. and Gine, E. (1999). *Decoupling*. Springer-Verlag (New York).
- Kuchibhotla and Patra (2022). On Least Squares Estimation Under Heteroscedastic and Heavy-Tailed Errors. *Annals of Statistics.*

References

- Chernozhukov, V., Chetverikov, D., and Kato, K. (2015). Comparison and anti-concentration bounds for maxima of Gaussian random vectors. *Prob. Theory Related Fields, 162(1-2):47–70*.
- de la Pena, V. H. and Gine, E. (1999). *Decoupling*. Springer-Verlag (New York).
- Kuchibhotla and Patra (2022). On Least Squares Estimation Under Heteroscedastic and Heavy-Tailed Errors. *Annals of Statistics.*

Thank You!