Basics of Concentration Inequalities

John Duchi

Stats 300b – Winter Quarter 2021

Outline

- Sub-Gaussian and sub-exponential random variables
- Symmetrization
- Applications to uniform laws
- Azuma-Hoeffding inequalities
- Doob martingales and bounded differences inequality

Reading: (this is more than sufficient)

- Wainwright, *High Dimensional Statistics*, Chapters 2.1–2.2
- Vershynin, *High Dimensional Probability*, Chapters 1–2.
- Additional perspective: van der Vaart, *Asymptotic Statistics*, Chapter 19.1–19.2

Concentration inequalities

inequalities of the form

$$\mathbb{P}(X \geq t) \leq \phi(t)$$

where ϕ goes to zero (quickly) as $t \to \infty$

often, want to deal with sums, so instead (e.g.)

$$\mathbb{P}(\overline{X}_n \geq t) \leq \phi_n(t)$$

- underpin many ULLNs
- key in high-dimensional statistics (concentration of measure)

The familiar Markov bounds

Proposition (Markov's inequality)

If $X \geq 0$*, then* $\mathbb{P}(X \geq t) \leq \frac{\mathbb{E}[X]}{t}$ *for all* $t \geq 0$*.*

Proposition (Chebyshev's inequality)

For any $t \geq 0$*,* $\mathbb{P}(|X - \mathbb{E}[X]| \geq t) \leq \frac{\mathsf{Var}(X)}{t^2}$

Sub-gaussian random variables

A mean-zero random variable X is σ^2-*sub-Gaussian* if

$$\mathbb{E}[\exp(\lambda X)] \leq \exp\left(\frac{\lambda^2\sigma^2}{2}\right) \quad \text{for all } \lambda \in \mathbb{R}.$$

(many equivalent definitions; see Vershynin or exercises)

Example

If $X \sim \mathcal{N}(\mu, \sigma^2)$, then

$$\mathbb{E}[\exp(\lambda(X - \mathbb{E}[X]))] =$$

Example

If $X \in [a, b]$, then

$$\mathbb{E}[\exp(\lambda(X - \mathbb{E}[X]))] \leq \exp\left(\frac{\lambda^2(b-a)^2}{8}\right).$$

Tensorization identities

- variance inequality familiar: if X_i are independent,

$$\mathrm{Var}\left(\sum_{i=1}^{n} X_i\right) = \sum_{i=1}^{n} \mathrm{Var}(X_i)$$

Proposition

If X_i are independent and σ_i^2-sub-Gaussian, then $\sum_{i=1}^{n} X_i$ is $\sum_{i=1}^{n} \sigma_i^2$ sub-Gaussian.

Chernoff and Hoeffding Inequalities

Corollary (Chernoff bounds for sub-Gaussian random variables)

Let X be σ^2-sub-Gaussian. Then

$$\mathbb{P}\left(X - \mathbb{E}[X] \geq t\right) \leq \exp\left(-\frac{t^2}{2\sigma^2}\right).$$

Corollary (Hoeffding bounds)

If X_i are independent σ_i^2-sub-Gaussian random variables,

$$\mathbb{P}\left(\frac{1}{n}\sum_{i=1}^{n}(X_i - \mathbb{E}[X_i]) \geq t\right) \leq \exp\left(-\frac{nt^2}{\frac{2}{n}\sum_{i=1}^{n}\sigma_i^2}\right).$$

- usually stated as $X_i \in [a, b]$, so bound is $\exp(-\frac{2nt^2}{(b-a)^2})$

Maxima of sub-Gaussian random variables

- often want to control deviations of maximum (supremum in ULLNs)

Proposition

Let $\{Z_i\}_{i=1}^N$ be σ^2-sub-Gaussian (not necessarily independent). Then

$$\mathbb{E}\left[\max_i Z_i\right] \leq \sqrt{2\sigma^2 \log N}.$$

Sub-exponential random variables

- more nuanced control if variance small, or sub-gaussian parameter unavailable

Definition (Sub-exponential)

A random variable X is (τ^2, b)-sub-exponential if

$$\mathbb{E}[\exp(\lambda(X - \mathbb{E}[X]))] \le \exp\left(\frac{\lambda^2\tau^2}{2}\right) \quad \text{for } |\lambda| \le \frac{1}{b}$$

Proposition (Tail bounds for sub-exponentials)

If X is (τ^2, b)-sub-exponential, then

$$\mathbb{P}\left(|X - \mathbb{E}[X]| \ge t\right) \le 2\exp\left(-\min\left\{\frac{t^2}{2\tau^2}, \frac{t}{b}\right\}\right)$$

Examples

Example (Bounded random variables)

If $X \in [-b, b]$, $\mathbb{E}[X] = 0$, and $\mathrm{Var}(X) = \sigma^2$, X is $(2\sigma^2, b)$-sub-exponential.

- see also Vershynin, Ch. 2

Tensorization

Proposition (Tensorization)

Let X_i be independent (τ_i^2, b_i)-sub-exponential. Then $\sum_{i=1}^n X_i$ is $(\sum_{i=1}^n \tau_i^2, \max_{i\le n} b_i)$-sub-exponential.

Corollary (Bernstein-type bounds)

If $|X_i| \le b$ and $\mathrm{Var}(X_i) \le \sigma^2$, then

$$\mathbb{P}(|\overline{X}_n - \mathbb{E}[\overline{X}_n]| \ge t) \le 2\exp\left(-c\min\left\{\frac{nt^2}{\sigma^2}, \frac{nt}{b}\right\}\right) \quad \textit{for } t \ge 0.$$

Symmetrization

- important idea in uniform laws of large numbers and concentration
- Banach space theory (surprisingly) develops many of these ideas

motivation: for ULLNs, Markov's inequality gives

$$\mathbb{P}\left(\sup_{f\in\mathcal{F}}(P_n - P)f \geq t\right) \leq \frac{\mathbb{E}[\sup_{f\in\mathcal{F}}(P_n - P)f]}{t}$$

sometimes if $P_n - P$ is symmetric, it's easier to deal with

Symmetrization in a vector space

- X_i are arbitrary vectors in a normed space with norm $\|\cdot\|$
- $\varepsilon_i \in \{\pm 1\}$ are i.i.d. uniform signs (*Rademacher variables*)

Theorem

Let $F : \mathbb{R}_+ \to \mathbb{R}_+$ be convex, increasing, and X_i be independent. Then

$$\mathbb{E}\left[F\left(\left\|\sum_{i=1}^{n}(X_i - \mathbb{E}[X_i])\right\|\right)\right] \le \mathbb{E}\left[F\left(2\left\|\sum_{i=1}^{n}\varepsilon_i X_i\right\|\right)\right].$$

Consequences

Corollary

If $\mathbb{E}[X_i] = 0$*, for any norm* $\|\cdot\|$ *and* $p \geq 1$*, we have*

$$\mathbb{E}\left[\left\|\sum_{i=1}^{n} X_i\right\|^p\right] \leq 2^p \mathbb{E}\left[\left\|\sum_{i=1}^{n} \varepsilon_i X_i\right\|^p\right]$$

Consequences

- treat measures as vectors (linear mappings from $\mathcal{F}$ to $\mathbb{R}$)
- norm $\|\mu\|_{\mathcal{F}} = \sup_{f \in \mathcal{F}} |\int f d\mu|$
- (ignore measurability, completeness, etc.)

Corollary

Let P_n^0 be shorthand for random measure

$$P_n^0 f := \frac{1}{n} \sum_{i=1}^{n} \varepsilon_i f(X_i).$$

Then

$$\mathbb{E}\left[\|P_n - P\|_{\mathcal{F}}^p\right] \leq 2^p \mathbb{E}\left[\left\|P_n^0\right\|_{\mathcal{F}}\right].$$

Uses of symmetrization

- often easier to deal with symmetric random variables
- can give (much) more precise bounds on these quantities
- easy proofs of ULLNs
- quantity $\sum_{i=1}^{n} \varepsilon_i X_i$ is $\sum_{i=1}^{n} X_i^2$-sub-Gaussian (conditional on X_is)

Rademacher complexities

Definition

The *empirical Rademacher complexity* of a class $\mathcal{F}$ is

$$R_n(\mathcal{F} \mid X_1^n) := \mathbb{E}\left[\sup_{f \in \mathcal{F}} \left|\sum_{i=1}^n \varepsilon_i f(X_i)\right| \mid X_1^n\right] = n\mathbb{E}\left[\left\|P_n^0\right\|_{\mathcal{F}} \mid X_1^n\right].$$

The *Rademacher complexity* is $R_n(\mathcal{F}) := \mathbb{E}[R_n(\mathcal{F} \mid X_1^n)]$.

Corollary

$$\mathbb{P}\left(\sup_{f \in \mathcal{F}} |P_n f - Pf| \geq t\right) \leq \frac{2}{nt} R_n(\mathcal{F}),$$

so if $R_n(\mathcal{F}) = o(n)$ *then* $\|P_n - P\|_{\mathcal{F}} \xrightarrow{p} 0$.

Metric entropies and symmetrization give a ULLN

- typical to have an *envelope function*, i.e. if $\mathcal{F} \subset \{\mathcal{X} \to \mathbb{R}\}$ there exists F such that

$$|f(x)| \le F(x) \text{ for all } f \in \mathcal{F} \text{ and } PF < \infty$$

- Define truncated class for $M \in \mathbb{R}_+$ by

$$f_M(x) := \begin{cases} f(x) & \text{if } |f(x)| \le M \\ 0 & \text{otherwise} \end{cases}$$

and $\mathcal{F}_M := \{f_M : f \in \mathcal{F}\}$

Theorem
Let $\mathcal{F}$ have envelope $F \in L^1(P)$. If $\log N(\mathcal{F}_M, L^1(P_n), \epsilon) = o(n)$ for all $M < \infty$ and $\epsilon > 0$, then $\|P_n - P\|_{\mathcal{F}} \xrightarrow{p} 0$.

Proof of ULLN

Lemma (Metric entropies bound Rademacher complexity)

For any class of functions $\mathcal{G} \subset \{\mathcal{X} \to \mathbb{R}\}$, *for* $\sigma_n^2 = \sup_{g \in \mathcal{G}} \frac{1}{n} \sum_{i=1}^n g(X_i)^2$ *we have*

$$R_n(\mathcal{G} \mid X_1^n) \lesssim \sqrt{n\sigma_n^2 \log N(\mathcal{G}, L^1(P_n), \epsilon)} + \epsilon.$$

Examples:

Example (Lipschitz functions)

If $\mathcal{F}$ is the collection of 1-Lipschitz functions on $[0,1]$ with $f(0) = 0$, then

$$\log N(\mathcal{F}, \|\cdot\|_\infty, \epsilon) \asymp \frac{1}{\epsilon}$$

and

$$\mathbb{E}\left[\left\|P_n^0 f\right\|_\mathcal{F}\right] \lesssim \epsilon + \frac{1}{\sqrt{n\epsilon}}$$

Revisiting concentration

goal: often we'd like to show concentration of more complex objects than averages, e.g.

$$\sup_{f \in \mathcal{F}} \frac{1}{n} \sum_{i=1}^{n} f(X_i)$$

major tool: martingales

Martingales

Definition (Non-measure theoretic version)

Let $X_1, X_2, \ldots$ be a sequence of random variables and $Z_1, Z_2, \ldots$ be another, where X_i and Z_{i-1} are functions of Z_i. Then X_i is a *martingale difference sequence* adapted to Z_i if

$$\mathbb{E}[X_i \mid Z_{i-1}] = 0 \text{ for all } i,$$

and $M_n := \sum_{i=1}^{n} X_i$ is the associated *martingale*

(converse definition: given M_n such that $\mathbb{E}[M_n \mid Z_{n-1}] = M_{n-1}$ and M_n is a function of Z_n, $X_n = M_n - M_{n-1}$ is the difference sequence)

Sub-Gaussian Martingales

A martingale difference sequence $\{X_i\}$ is σ^2-sub-Gaussian if

$$\mathbb{E}[\exp(\lambda X_i) \mid Z_{i-1}] \leq \exp\left(\frac{\lambda^2 \sigma^2}{2}\right) \quad \text{for all } i, Z_1^{i-1}.$$

Theorem (Azuma-Hoeffding)

Let X_i be σ_i^2-sub-Gaussian martingale differences. Then for $t \geq 0$,

$$P\left(\sum_{i=1}^n X_i \geq t\right) \leq \exp\left(-\frac{t^2}{2\sum_{i=1}^n \sigma_i^2}\right).$$

Doob martingales and functions of independent variables

- $X_i \in \mathcal{X}$ are independent random variables
- $f : \mathcal{X}^n \to \mathbb{R}$
- to control $f(X_1^n) - \mathbb{E}[f(X_1^n)]$ construct *Doob martingale*

construction: set $Z_i = \{X_1^{i-1}\}$ and define *differences*

$$D_i := \mathbb{E}[f(X_1^n) \mid Z_i] - \mathbb{E}[f(X_1^n) \mid Z_{i-1}],$$

so

$$\sum_{i=1}^n D_i = f(X_1^n) - \mathbb{E}[f(X_1^n)]$$

observation: D_i are martingale differences adapted to Z_i

Bounded differences (McDiarmid's) inequality

Theorem (Bounded differences)

Let $f : \mathcal{X}^n \to \mathbb{R}$ satisfy c_i-bounded differences,

$$|f(x_1^{i-1}, x_i, x_{i+1}^n) - f(x_1^{i-1}, x_i', x_{i+1}^n)| \le c_i \quad \text{all } x, x' \in \mathcal{X}^n.$$

Then $f - Pf$ is $\frac{1}{4}\sum_{i=1}^n c_i^2$-sub-Gaussian.

Corollary

Let $f : \mathcal{X}^n \to \mathbb{R}$ have c_i-bounded differences and X_i be independent. Then

$$\mathbb{P}\left(f(X_1^n) - Pf \ge t\right) \le \exp\left(-\frac{2t^2}{\sum_{i=1}^n c_i^2}\right) \quad \text{for } t \ge 0.$$

Rademacher complexities and bounded differences

- the *empirical process* often satisfies bounded differences

Proposition

Let $\mathcal{F} \subset \{\mathcal{X} \to \mathbb{R}\}$ *satisfy* $|f(x) - f(x')| \le B$ *for* $x, x' \in \mathcal{X}$*. Then*

$$\sup_{f \in \mathcal{F}} \frac{1}{n} \sum_{i=1}^{n} (f(X_i) - Pf) \quad \text{and} \quad \|P_n - P\|_{\mathcal{F}} = \sup_{f \in \mathcal{F}} \left| \frac{1}{n} \sum_{i=1}^{n} (f(X_i) - Pf) \right|$$

satisfy $\frac{B}{n}$ *bounded differences.*

Concentration of the empirical process

Corollary

Let $\mathcal{F} \subset \{\mathcal{X} \to \mathbb{R}\}$ satisfy $|f(x) - f(x')| \le B$ for all $x, x' \in \mathcal{X}$. Then

$$\mathbb{P}\left(\sup_{f \in \mathcal{F}}(P_n f - Pf) \ge \mathbb{E}[\sup_{f \in \mathcal{F}}(P_n f - Pf)] \ge t\right) \le \exp\left(-\frac{2nt^2}{B^2}\right)$$

$$\mathbb{P}\left(\|P_n - P\|_{\mathcal{F}} \ge \mathbb{E}[\|P_n - P\|_{\mathcal{F}}] + t\right) \le \exp\left(-\frac{2nt^2}{B^2}\right)$$

for all $t \ge 0$.

Preview: by symmetrization,

$$\mathbb{E}\left[\|P_n - P\|_{\mathcal{F}}\right] \le 2\mathbb{E}\left[\left\|P_n^0\right\|_{\mathcal{F}}\right] = 2\frac{R_n(\mathcal{F})}{n},$$

so controlling expectations evidently important