Warning: these notes may contain factual errors

Outline:
- concentration inequalities for functions with bounded differences
- ULLN for bounded class via concentration and chaining
- growth rates, moduli of continuity

Reading: VDV 18-19, HDP 8

Recap: Recall that if a process \( \{X_t\}_{t \in T} \) is sub-Gaussian, i.e.
\[
\mathbb{E} \exp(\lambda (X_s - X_t)) \leq \exp\left(\frac{\lambda^2 d(s, t)^2}{2}\right), \forall s, t \in T
\]
then \( \exists C < \infty \) such that
\[
\mathbb{E}[\sup_{t \in T} X_t] \leq C \int_0^{\text{diam}(T)} \sqrt{\log N(T, d, \varepsilon)} d\varepsilon
\]
where \( N \) is the covering number. As a corollary, if we define the entropy integral
\[
J(T; \delta) = \int_\delta^\infty \sqrt{\log N(T, d, \varepsilon)} d\varepsilon
\]
Then
\[
\mathbb{E}[\sup_{t \in T} X_t] \leq C(\mathbb{E} \sup_{d(t, s) \leq \delta} |X_t - X_s|) + J(T; \delta)
\]
where we note that the integral in \( J(T; \delta) \) has upper limit \( \text{diam}(T) \) since for \( \varepsilon > \text{diam}(T) \), covering number is 1.

Example: Let \( \mathcal{F} \subset \{X \rightarrow \mathbb{R}\} \) be a VC class, with \( \|f\|_\infty \leq b \) for all \( f \in \mathcal{F} \). Then
\[
\sqrt{n} P^0 f := \frac{1}{\sqrt{n}} \sum_i \varepsilon_i f(X_i)
\]
where \( \varepsilon_i \) are iid Rademacher, is sub-Gaussian for fixed \( X_{1:n} \), in terms of the \( \| \cdot \|_{L_2(P_n)} \) norm. Applying chaining, we obtain
\[
\sqrt{n} \mathbb{E}[\sup_{f \in \mathcal{F}} |P^0 f|] \leq C \int_0^\infty \sqrt{\log N(F, \| \cdot \|_{L_2(P_n)}, \varepsilon)} d\varepsilon = *
\]
Recall the following bound on the covering number for uniformly bounded VC class $\mathcal{F}$:

$$\sup_P N(F, \| \cdot \|_{L_r(P)}, \varepsilon) \leq c_r \left( \frac{b}{\varepsilon} \right)^{r\text{VC}(\mathcal{F})} \leq c_r (1 + \frac{b}{\varepsilon})^{r\text{VC}(\mathcal{F})}$$

Applying this we have

$$\ast \leq C \int_0^b \sqrt{C + \text{VC}(\mathcal{F}) \log(1 + \frac{b}{\varepsilon})}d\varepsilon \leq C \int_0^b \frac{1 + \text{VC}(\mathcal{F})}{\varepsilon} \frac{b}{\varepsilon}d\varepsilon \leq C\sqrt{\text{VC}(\mathcal{F})} \cdot b$$

which gives the bound

$$\mathbb{E}[\sup_{f \in \mathcal{F}} |P^0_n f|] \leq Cb \sqrt{\frac{\text{VC}(\mathcal{F})}{n}}$$

1 Concentration Inequalities (revisited)

**Remark** Often we want to understand concentration of more sophisticated things than iid sums, e.g. $\sup_{f \in \mathcal{F}} |P_n f - Pf|$, which is what we care about for ULLN. We want to answer the following question: If $X_1:n$ are independent, when does $f(X_1:n)$ concentrate around $\mathbb{E}f(X_1:n)$, where $f : \mathcal{X}^n \to \mathbb{R}$? The idea is that if $f$ depends “little” on individual $X_i$, there should be concentration. We use bounded differences and martingale methods to show this.

**Definition 1.1.** A sequence $\{X_i\}$ adapted to a filtration $\mathcal{F}_1 \subset \mathcal{F}_2 \subset \ldots$ (increasing sequence of $\sigma$-fields) is a martingale difference sequence (MGD) if

- $X_i \in \mathcal{F}_i$ for any $i \in \mathbb{N}$
- $\mathbb{E}[X_i | \mathcal{F}_{i-1}] = 0$ for any $i \in \mathbb{N}$.

Recall $M_n = \sum_{i=1}^n X_i$ is associated martingale ($X_i = M_i - M_{i-1}$) and note that $\mathbb{E}[M_n | \mathcal{F}_{i-1}] = M_{n-1}$.

**Definition 1.2.** Let $X_i$ be a MGD, it is $\delta^2_i$-sub-Gaussian if

$$\mathbb{E}[\exp(\lambda X_i) | \mathcal{F}_{i-1}] \leq \exp\left(\frac{\lambda^2 \delta^2_i}{2}\right)$$

for all $i \in \mathbb{N}$.

**Theorem 1.** If $\{X_i\}$ are $\sigma^2_i$-sub-Gaussian MGD, then

$$M_n := \sum_{i=1}^n X_i$$

is $\sum_{i=1}^n \sigma^2_i$-sub-Gaussian.
Proof. We have
\[ E[\exp(\lambda \sum_{i=1}^{n} X_i)] = E \left[ E \left[ e^{\lambda X_n} \mid \mathcal{F}_{n-1} \right] \cdot E \left[ \exp^{\lambda \sum_{i=1}^{n-1} X_i} \mid \mathcal{F}_{n-1} \right] \right] \]
\[ \leq \exp\left(\frac{\lambda^2 \sigma_n^2}{2}\right) \cdot E \left[ \exp(\lambda \sum_{i=1}^{n-1} X_i) \right] \]
and proof follows by induction.

\[ \square \]

Corollary 2. (Azuma-Hoeffding) Under conditions of the previous theorem, we have the bound
\[ \mathbb{P}(\frac{1}{n} \sum_{i=1}^{n} X_i \geq t) \leq \exp\left(-\frac{nt^2}{2\sum_{i} \sigma_i^2}\right) \]

Example: Recall that, if \(|X_i| \leq c_i\), then \(\sigma_i^2 \leq c_i^2\), so the previous bound implies
\[ \mathbb{P}(\frac{1}{n} \sum_{i=1}^{n} X_i \geq t) \leq \exp\left(-\frac{nt^2}{2\sum_{i} c_i^2}\right) \]

2 Martingales and Bounded Differences

Let \(\{X_i\}_{i=1}^{n}\) be independent, \(X_i \in \mathcal{X}\). Let \(f : \mathcal{X}^n \to \mathbb{R}\). How to use the previous results about martingale to control \(f(X_{1:n}) - \mathbb{E}[f(X_{1:n})]\) Doob martingale provides a useful construction for transforming \(f(X_{1:n}) - \mathbb{E}[f(X_{1:n})]\) with \(f : \mathcal{X}^n \to \mathbb{R}\) into a sum of MGDs.

2.1 Doob martingale

Definition 2.1. Let \(f : \mathcal{X}^n \to \mathbb{R}\) and \(X_i\) be random variables. Let \(\mathcal{F}_i = \sigma(X_1, \ldots, X_i)\). Define
\[ D_i := \mathbb{E}[f(X_{1:n}) \mid \mathcal{F}_i] - \mathbb{E}[f(X_{1:n}) \mid \mathcal{F}_{i-1}] \]
Then \(D_i\)'s are called the Doob MGDs.

Note that
\[ \mathbb{E}[D_i \mid \mathcal{F}_{i-1}] = 0 \]
\[ \sum_{i=1}^{n} D_i = f(X_{1:n}) - \mathbb{E}[f(X_{1:n})] \]

By the previous theorem, we see that if \(D_i\)'s are sub-Gaussian, so is \(f(X_{1:n}) - \mathbb{E}[f(X_{1:n})]\). The question becomes: for what \(f\)'s are \(D_i\)'s small? The answer is the class of functions \(f\) with bounded differences.

2.2 Bounded differences

Definition 2.2. A function \(f : \mathcal{X}^n \to \mathbb{R}\) has bounded differences if
\[ \sup_{X_{1:n} \in \mathcal{X}^n, X'_i \in \mathcal{X}'} |f(X_{1:n}) - f(X_{1:n-1}, X'_i, X_{i+1:n})| \leq c_i \]
Example: Let $X \in [-1,1]$ and $f(X_{1:n}) = \frac{1}{n} \sum X_i$, then

$$|f(X_{1:n}) - f(X_{1:n-1}, X'_i, X_{i+1:n})| \leq \frac{1}{n} |X_i - X'_i| \leq \frac{2}{n}$$

Theorem 3. (McDiarmid’s inequality) If $X_i$ are independent, $f$ has bounded differences, then

$$P(f(X_{1:n}) - \mathbb{E}f(X_{1:n}) \geq t) \leq \exp\left(-\frac{2t^2}{\sum_{i=1}^n c_i^2}\right)$$

and similarly for lower tail.

Proof It suffices to show that bounded differences implies $D_i$’s are $\frac{c_i^2}{n}$-sub-Gaussian, since then the Azuma-Hoeffding bound will imply the desired bound. We have

$$D_i = \mathbb{E}[f(X_{1:n}) | \mathcal{F}_i] - \mathbb{E}[f(X_{1:n}) | \mathcal{F}_{i-1}]$$

$$= \int f(X_{i:i}, X_{i+1:n}) dP_i - \int f(X_{i:i-1}, X_i, X_{i+1:n}) dP_i dP^n_i(X_{i+1:n})$$

$$= \int \int [f(X_{i:i-1}, X'_i, X_{i+1:n}) - f(X_{i:i-1}, X_i, X_{i+1:n})] dP_i dP^n_i(X_{i+1:n})$$

The term in the integrand is bounded above by $c_i$, so that $D_i$ is $\frac{c_i^2}{n}$-sub-Gaussian.

Example Supremum of bounded function classes Let $\mathcal{F} \subset \{X \to \mathbb{R}\}$ and assume $|f(X)| \in [a,b]$. Suppose $P_n, P'_n$ differ only in $X_i$ and $X'_i$. Then the supremum $\sup_f |P_n f - Pf|$ has bounded differences:

$$\sup_f |P_n f - Pf| - \sup_f |P'_n f - Pf|$$

$$\leq \sup_f |P_n f - Pf| - |P'_n f - Pf|$$

$$\leq \sup_f |P_n f - P_n f|$$

$$= \sup_{f \in \mathcal{F}} |f(X_i) - f(X'_i)|/n \leq \frac{b-a}{n}$$

Corollary 4. Let $\mathcal{F} \subset \{X \to [a,b]\}$. Then

$$P\left(\sup_{f \in \mathcal{F}} |P_n f - Pf| - \mathbb{E}\left[\sup_{f \in \mathcal{F}} |P_n f - Pf|\right] \geq t\right) \leq \exp\left(-\frac{2nt^2}{(b-a)^2}\right)$$

Proof Set $c_i^2 = \frac{(b-a)^2}{n^2}$ in McDiarmid.

If we want ULLN for bounded class $\mathcal{F}$, all we need is control over $\mathbb{E}\|P_n - P\|_{\mathcal{F}}$. But this is precisely what we can do with chaining. Applying the bound on $\mathbb{E}\|P_n - P\|_{\mathcal{F}}$, we get the following convergence rate result.
Corollary 5. Let \( \mathcal{F} \) be a bounded VC class, \( f(x) \in [a, b] \). Then

\[
\mathbb{P}
\left(
\|P_n - P\|_F \geq C \sqrt{\frac{VC(\mathcal{F})}{n}} + t
\right)
\leq \exp\left(-\frac{2nt^2}{(b-a)^2}\right)
\]

where \( C \) depends on the VC class bound.

As a consequence, letting \( \mathcal{F} = \{1(X \leq t), t \in \mathbb{R}^d\} \), then

\[
\mathbb{P}(\sup_{t \in \mathbb{R}^d} |P_n(X \leq t) - P(X \leq t)| \geq C \sqrt{\frac{d}{n}} + \varepsilon) \leq \exp(-2n\varepsilon^2)
\]

which is the DKW inequality, up to sharp constants.

3 Convergence Rates

Next we move on to rates of convergence for model parameters, which are solutions of optimization problems. Our setting is empirical minimization (M-estimation).

Let \( \ell : \Theta \times \mathcal{X} \to \mathbb{R} \) be the loss function and

\[
L(\theta) := \mathbb{E}(\ell(\theta; X))
\]

\[
L_n(\theta) := P_n\ell(\theta; X)
\]

If

\[
\hat{\theta}_n = \arg\min_{\theta} L_n(\theta)
\]

\[
\theta^* = \arg\min_{\theta} L(\theta)
\]

How quickly does \( \hat{\theta}_n \to \theta^* \)? We hope that the growth in \( L \) near \( \theta^* \) \( \gg \) variation of \( L_n(\theta) - L_n(\theta^*) \) for \( \theta \) near \( \theta^* \). Our goal is to show \( L_n(\theta) > L_n(\theta^*) \) for \( \theta \) “far enough” from \( \theta^* \).