27.1 Concentration Inequalities for U-statistics

Now we consider finite sample bounds for U-statistics. We first look at a result from [H63]. Let

\[ U_n = \frac{1}{\binom{n}{m}} \sum_{i_1 < \ldots < i_m} h(X_{i_1}, \ldots, X_{i_m}) , \]

and assume that \( |h(x_1, \ldots, x_m)| \leq B \) and \( B = \|h\|_\infty \). Recall that using the bounded difference inequality, we can get

\[ \mathbb{P}[|U_n - \theta| \geq t] \leq 2e^{-\frac{t^2}{2B^2 n m^2}} . \]

As a reminder we are now allowing \( m \) to vary with \( n \). Now we prove a better dependence between \( n \) and \( m \), where instead of an order of \( \frac{n}{m^2} \), we can get \( \frac{n}{m} \), i.e. the order of the U-statistic grows with \( m \) at a better rate:

\[ \mathbb{P}[|U_n - \theta| \geq t] \leq 2e^{-\frac{t^2}{2B^2 \frac{n}{m}}} . \]

**Proof:** Let \( k = \lceil \frac{n}{m} \rceil \), the smallest integer \( \geq \frac{n}{m} \), and set

\[ W(x_1, \ldots, x_n) = \frac{h(x_1, \ldots, x_m) + h(x_{m+1}, \ldots, x_{2m}) + \ldots + h(x_{m(k-1)+1}, \ldots, x_{mk})}{k} , \]

i.e. we break our sample into \( k \) non-overlapping blocks of size \( m \). Then

\[ k \sum_{\sigma \in S^n} W(X_{\sigma_1}, \ldots, X_{\sigma_n}) = km!(n-m)! \sum_{i_1 < \ldots < i_m} h(X_{i_1}, \ldots, X_{i_m}) \]

Here \( S^n \) is the permutation group on \( \{1, \ldots, n\} \). Then we have

\[ \sum_{\sigma \in S^n} W(X_{\sigma_1}, \ldots, X_{\sigma_n}) = m!(n-m)! \binom{n}{m} U_n \]

which gives

\[ U_n = \frac{1}{n!} \sum_{\sigma \in S^n} W(X_{\sigma_1}, \ldots, X_{\sigma_n}) . \]
Notice that for a fixed $\sigma$, $W(X_{\sigma_1},...,X_{\sigma_n})$ is an average of $k$ iid random variables bounded by $B$ in absolute value (by assumption), so we can define

$$T_{\sigma} = W(X_{\sigma_1},...,X_{\sigma_n}) - \theta \in SG\left(\frac{B^2}{k}\right).$$

The above was for one fixed $\sigma$. Now we sum over all possible values of $\sigma$:

$$U_n - \theta = \frac{1}{n!} \sum_{\sigma} T_{\sigma}.$$ 

For $t > 0$,

$$\mathbb{P}[U_n - \theta \geq t] \leq e^{-\lambda t} \mathbb{E}[e^{\lambda(U_n - \theta)}]$$

with $\lambda > 0$ by Markov’s inequality.

$$= e^{-\lambda t} \mathbb{E}[e^{\lambda \frac{k}{B^2} \sum_{\sigma} T_{\sigma}}]$$

$$\leq e^{-\lambda t} \frac{1}{n!} \sum_{\sigma} \mathbb{E}[e^{\lambda T_{\sigma}}]$$

by Jensen’s inequality.

$$\leq e^{-\lambda t} e^{\frac{\lambda^2 k}{B^2}}$$

since $T_{\sigma} \in SG\left(\frac{B^2}{k}\right)$.

$$= e^{-\frac{t^2 n}{2\pi^2 k}}$$

taking $\lambda = \frac{kt}{B^2}$ (minimizing over $\lambda$).

$$\leq e^{-\frac{t^2}{2\pi^2}(\frac{\pi}{\bar{a}})}.$$ 

Repeating for the other side gives us the desired result.

27.1.1 Sharpest Known Result

Now, we assume that $\text{Range}(h) = [0,1]$. As a generalization of the earlier result we can show that

$$\mathbb{P}[|U_n - \theta| \geq t] \leq a e^{-\frac{t^2}{2\pi^2(\bar{a} + b \bar{a})}}.$$ 

In our earlier result [H63], we have $a = 2$, $\sigma^2 = \mathbb{V}[h(X_{1},...,X_{m})]$, and $b = \frac{2}{3}$.

By considering the variance, [A95] showed this for $a = 4$, $\sigma^2 = m\zeta_1$, $b = 2m^3m^{m-1} + \frac{4}{3}m^{-2}$. Here $\sigma^2$ is exactly the asymptotic variance of a non-degenerate U-statistic. However, as we can see, the dependence on $m$ here is not great due to the $m^{m-1}$ term, and this result is not useful in a high-dimensional setting if $m$ is growing with $n$.

27.2 Review

27.2.1 High-dimensional Statistical Model

A statistical model can be described by $\mathcal{P}$ on $\mathcal{X}$, where $X_1,...,X_n \overset{iid}{\sim} P \in \mathcal{P}$ and $\theta : \mathcal{P} \to \mathbb{R}^d$. In a parametric model, $\mathcal{P}, \theta$ and $d$ are fixed, whereas in a high-dimensional statistical model, $\mathcal{P}, \theta$ and/or $d$ can change with $n$. There are two regimes in high-dimensional statistics: $d \gg n$ and $d \ll n$ (where $d = d(n)$). In the first regime we have to make strong structural assumptions such as sparsity. Here we focus on the second regime.
27.2.2 Concentration Inequalities

We are interested in finite sample results, and in particular we are interested in deviations of a function from its mean (or median), i.e. bounds on
\[ \mathbb{P}[|f(X_1, ..., X_n) - \mathbb{E}[f(X_1, ..., X_n)]| \geq t]. \]

The canonical choice for \( f(X_1, ..., X_n) \) is \( \sum X_i \), where the \( X \)'s are independent. To get such bounds we assume that the \( X \)'s are either sub-Gaussian or sub-exponential.

We can derive the following result:
\[ X \in SG(\sigma^2) \implies X^2 \in SG(5\sigma^2, 10\sigma). \]

27.2.2.1 Hoeffding, Bernstein, Bounded Difference Inequality

If \( |X - \mathbb{E}[X]| \leq b \) a.e. and \( \sigma^2 = \mathbb{V}[X] \), we can derive the following two results:
\[ \mathbb{P}[|X - \mathbb{E}[X]| \geq t] \leq 2e^{-\frac{t^2}{2(\sigma^2 + bt)}} \text{ using Bernstein's, and} \]
\[ \mathbb{P}[|X - \mathbb{E}[X]| \geq t] \leq 2e^{-\frac{t^2}{2\sigma^2}} \text{ using Hoeffding's.} \]

We use Bernstein’s when we would like to incorporate the variance. Hoeffding’s is an off-the-shelf result for bounded random variables.

If we are dealing with a function of the \( X \)'s, where \( f \) is not a sum, and might be complicated, we can use the bounded difference inequality. To use this we need to check that the bounded difference property holds. If \( X_1, ..., X_n \) are independent and \( f(X_1, ..., X_n) \) satisfies the bounded difference property with parameters \((L_1, ..., L_n)\), i.e.
\[ |f(x_1, ..., x_{k-1}, x_k, x_{k+1}, ..., x_n) - f(x_1, ..., x_{k-1}, x'_k, x_{k+1}, ..., x_n)| \leq L_k \]
for all \( k \) from 1 to \( n \), we get the following bound
\[ \mathbb{P}[|f(X) - \mathbb{E}[f(X)]| \geq t] \leq 2e^{-\frac{t^2}{\sum_{k=1}^{n} L_k^2}}. \]

27.2.2.2 Maximal Inequalities

We have this important result for maxima of sub-Gaussian random variables, not necessarily independent.
\[ \mathbb{E}[\max_i X_i] \leq \sqrt{2\sigma^2 \log n} \]

There is an extension to sub-exponential random variables. We can also use the union bound to derive the following result:
\[ \mathbb{P}[\max_i X_i > t] \leq e^{-\frac{t^2}{2\sigma^2} + \log n}. \]
27.2.2.3 Lipschitz Functions of Gaussians

We can derive the following strong concentration property when we are dealing with Lipschitz functions of Gaussian random variables. This is dimension-free.

If \( X_1, \ldots, X_n \sim N(0, \sigma^2 I) \), and \( f : \mathbb{R}^n \to \mathbb{R} \) is \( L \)-Lipschitz with respect to the Euclidean norm, we have

\[
P(|f(X_1, \ldots, X_n) - \mathbb{E}[f(X_1, \ldots, X_n)]| \geq t) \leq 2e^{-\frac{t^2}{2L^2\sigma^2}}.
\]

27.2.3 Packing and Covering Numbers

We derived the following useful result. In \( \mathbb{R}^d \), let \( \|\cdot\| \) be a norm and \( B \) be its unit ball. Then for \( \epsilon > 0 \),

\[
\left(\frac{1}{\epsilon}\right)^d \leq N(\epsilon, B, \|\cdot\|) \leq \left(1 + \frac{2}{\epsilon}\right)^d.
\]

27.2.3.1 Discretization Argument

Previously, we derived bounds on \( \mathbb{E}[\max_i X_i] \), where the maximum is over a finite set. We can use a discretization argument to derive bounds when the maximum is taken over an infinite set.

When \( X \) is a vector in \( \mathbb{R}^d \), we say that \( X \in SG_d(\sigma^2) \) if \( v^T X \in SG(\sigma^2), \forall v \) where \( \|v\| = 1 \) (this also holds if \( \|v\| \leq 1 \), with parameter \( \leq \sigma^2 \)).

Now we can derive bounds on \( \mathbb{P}[\max_{\theta \in B} \theta^T X \geq t] \) and \( \mathbb{E}[\max_{\theta \in B} \theta^T X] \) using a covering of the unit ball.

Remark As a reminder, we can write \( \|X\|_2 = \sup_{\theta \in B} \theta^T X \).

27.2.3.2 Applications

Let \( X_1, \ldots, X_n \overset{iid}{\sim} (0, \Sigma) \) in \( \mathbb{R}^d \), and \( \in SG(\sigma^2) \). Then we have

\[
P\left(\|\Sigma - \hat{\Sigma}\|_{op} \geq \sigma^2 c \min\left\{\sqrt{\frac{d + \log(2/\delta)}{n}}, \frac{d + \log(2/\delta)}{n}\right\}\right) \leq \delta.
\]

This was derived using a discretization argument, where using the definition of the operator norm, the LHS can be written as \( \max_{x \in S^{d-1}} |x^T (\Sigma - \hat{\Sigma}) x| \), and we can bound this with \( 2 \max_{y \in N_{1/\delta}} |y^T (\Sigma - \hat{\Sigma}) y| \).

If we use a less restrictive distance such as \( \|\cdot\|_\infty \), we can get much better rates.

We also derived Weyl’s Inequality:

\[
\max_i \left|\lambda_i(\Sigma) - \lambda_i(\hat{\Sigma})\right| \leq \|\Sigma - \hat{\Sigma}\|_{op}.
\]

27.2.4 Linear Models

We use the following set-up for our discussion of linear models. Let

\[
Y = X\beta^* + \epsilon,
\]
where \(\epsilon_1, \ldots, \epsilon_n \overset{iid}{\sim} N(0, \sigma^2)\). \(X\) is a \(n \times d\) matrix and we treat the \(X\)'s as fixed. We first estimate \(\hat{\beta}\) using least squares. We are interested in \(\frac{1}{n} \|\hat{\beta} - \beta^*\|^2\) and \(\frac{1}{n} \|X(\hat{\beta} - \beta^*)\|^2\).

Bounding the first quantity requires \(\lambda_{\min}(X^TX) > 0\). For the second quantity, we derived the following result:

\[
P\left[\frac{1}{n} \|X(\hat{\beta} - \beta^*)\|^2 \leq c \frac{\sigma^2 r + \log \left(\frac{1}{n}\right)}{n}\right] \geq 1 - \delta,
\]

where \(r = \text{rank}(X^TX)\). We can think of \(\delta\) as \(\frac{1}{n}\).

We can derive the following basic inequality:

\[
\|X(\hat{\beta} - \beta^*)\|^2 \leq \frac{1}{2n} \|X^T \hat{\epsilon}\|^2 - \frac{\sigma^2}{n} \|\hat{\beta}\|^2 - \frac{\lambda_n}{\|\beta^*\|_1 - \|\hat{\beta}\|_1}.
\]

from which we can get

\[
\|X(\hat{\beta} - \beta^*)\|^2 \leq c \frac{\|\beta^*\|_1 \lambda_n}{\sigma^2/n}.
\]

if \(\lambda_n \geq \frac{\|\epsilon X\|_\infty}{\sigma^2/n}\). We choose \(\lambda_n = \sqrt{\frac{2\sigma^2}{n} (\log(1/\delta) + \log d)}\).

In order to get better rates we need conditions on the design matrix \(X\), in particular we covered the restricted eigenvalue condition.

### 27.2.4.2 Oracle Inequalities and Persistence

We covered oracle inequalities, which allow for a misspecified model. We also covered persistence, which allows us to tackle sparse regression problems without making any assumptions.

### 27.2.5 PCA

We looked at both PCA and sparse PCA.
27.2.5.1 Distance Between Linear Spaces

Here we are interested in the distance between linear spaces $\mathcal{E}$ and $\mathcal{F}$, which we measure using $\|\sin \Theta(\mathcal{E}, \mathcal{F})\|_F$. The principal or canonical angles between $\mathcal{E}$ and $\mathcal{F}$ are $\cos^{-1}(\sigma_1), ..., \cos^{-1}(\sigma_d)$, where $\sigma_1 \geq \sigma_2 \geq ... \geq \sigma_d$ are the singular values of $E^T F$ or $F^T E$.

27.2.5.2 Davis-Kahan Theorem

The Davis-Kahan Theorem is due to Yu, Wang and Samworth [YWS14]. We have the following result:

$$\|\sin \Theta(\mathcal{E}, \mathcal{F})\|_F \leq \frac{2 \min \left\{ \sqrt{d} \| \Sigma - \tilde{\Sigma} \|_{op}, \| \Sigma - \tilde{\Sigma} \|_F \right\}}{\delta},$$

where $\delta = \lambda_d(\Sigma) - \lambda_{d+1}(\Sigma)$, the eigengap. We apply this to the spiked covariance model, defined by $\Sigma = \theta vv^T + I_d$, where $\|v\| = 1, \theta > 0$.

27.2.6 Uniform Law of Large Numbers

Here we defined

$$\|P_n - P\|_F = \sup_{f \in \mathcal{F}} \frac{1}{n} \left| \sum_{i=1}^{n} f(x_i) - \mathbb{E}[f(x_i)] \right|.$$

This quantity is bounded by the Rademacher complexity, defined as

$$\mathcal{R}_n(\mathcal{F}) = \sup_{f \in \mathcal{F}} \mathbb{E}_{X, \epsilon} \left[ \left| \sum_{i=1}^{n} \epsilon_i f(x_i) \right| \right].$$

27.2.7 Other Topics

There were some other topics that were covered in the course but we did not have time to review all of them in class.

References

