0.0.1 Applications of concentration inequalities

Having developed the machinery above, we next apply it to a few concrete problems to give a sense of how to use it. A key lemma which we will use repeatedly is the union bound, which states that if \(E_1, \ldots, E_n\) are events with probabilities \(\pi_1, \ldots, \pi_n\), then the probability of \(E_1 \cup \cdots \cup E_n\) is at most \(\pi_1 + \cdots + \pi_n\). A corollary is that if \(n\) events each have probability \(\ll 1/n\), then there is a large probability that none of the events occur.

Maximum of sub-Gaussians. Suppose that \(X_1, \ldots, X_n\) are mean-zero sub-Gaussian with parameter \(\sigma\), and let \(Y = \max_{i=1}^n X_i\). How large is \(Y\)? We will show the following:

**Lemma 0.1.** The random variable \(Y\) is \(\mathcal{O}(\sigma \sqrt{\log(n/\delta)})\) with probability \(1 - \delta\).

**Proof.** By the Chernoff bound for sub-Gaussians, we have that \(\mathbb{P}[X_i \geq \sigma \sqrt{6 \log(n/\delta)}] \leq \exp(-\log(n/\delta)) = \delta/n\). Thus by the union bound, the probability that any of the \(X_i\) exceed \(\sigma \sqrt{6 \log(n/\delta)}\) is at most \(\delta\). Thus with probability at least \(1 - \delta\) we have \(Y \leq \sigma \sqrt{6 \log(n/\delta)}\), as claimed.

Lemma 0.1 illustrates a typical proof strategy: We first decompose the event we care about as a union of simpler events, then show that each individual event holds with high probability by exploiting independence.

As long as the “failure probability” of a single event is much small than the inverse of the number of events, the quantity inside the sup is attractive to analyze because it is an average of independent random variables. Indeed, we have

\[
\frac{1}{n} \sum_{i=1}^n \langle X_i, v \rangle^2 \leq \mathbb{E}[\exp(\frac{n}{\sigma^2} v^T M v)] = \prod_{i=1}^n \mathbb{E}[\exp(|\langle X_i, v \rangle|^2/\sigma^2)] \leq 2^n, 
\]

where the last step follows by sub-Gaussianity if \(\langle X_i, v \rangle\). The Chernoff bound then gives \(\mathbb{P}[v^T M v \geq t] \leq 2^n \exp(-nt/\sigma^2)\).

If we were to follow the same strategy as Lemma 0.1, the next step would be to union bound over \(v\). Unfortunately, there are infinitely many \(v\) so we cannot do this directly. Fortunately, we can get by with only considering a large but finite number of \(v\); we will construct a finite subset \(\mathcal{N}_{1/4}\) of the unit ball such that

\[
\sup_{v \in \mathcal{N}_{1/4}} v^T M v \geq \frac{1}{2} \sup_{\|v\|_2 \leq 1} v^T M v. 
\]

Our construction follows Section 5.2.2 of ?.

Eigenvalue of random matrix. Let \(X_1, \ldots, X_n\) be independent zero-mean sub-Gaussian variables in \(\mathbb{R}^d\) with parameter \(\sigma\), and let \(M = \frac{1}{n} \sum_{i=1}^n X_i X_i^\top\). How large is \(\|M\|\), the maximum eigenvalue of \(M\)? We will show:

**Lemma 0.2.** The maximum eigenvalue \(\|M\|\) is \(\mathcal{O}(\sigma^2 \cdot (1 + d/n + \log(1/\delta)/n))\) with probability \(1 - \delta\).

**Proof.** The maximum eigenvalue can be expressed as

\[
\|M\| = \sup_{\|v\|_2 \leq 1} v^\top M v = \sup_{\|v\|_2 \leq 1} \frac{1}{n} \sum_{i=1}^n |\langle X_i, v \rangle|^2.
\]

The quantity inside the sup is attractive to analyze because it is an average of independent random variables. Indeed, we have

\[
\mathbb{E}[\exp(\frac{n}{\sigma^2} v^\top M v)] = \prod_{i=1}^n \mathbb{E}[\exp(|\langle X_i, v \rangle|^2/\sigma^2)] \leq 2^n, 
\]

for all distinct \(x, y \in \mathcal{N}_{1/4}\). We observe that \(|\mathcal{N}_{1/4}| \leq 9^d\); this is because the balls of radius 1/8 around each point in \(\mathcal{N}_{1/4}\) are disjoint and contained in a ball of radius 9/8.
To establish (4), let \( v \) maximize \( v^\top Mv \) over \( \|v\|_2 \leq 1 \) and let \( u \) maximize \( v^\top Mv \) over \( \mathcal{N}_{1/4} \). Then

\[
|v^\top Mv - u^\top Mu| = |v^\top M(v-u) + u^\top M(v-u)| \leq (\|v\|_2 + \|u\|_2)\|M\|\|v-u\|_2 \leq 2\cdot\|M\| \cdot (1/4) = \|M\|/2.
\]

Since \( v^\top Mv = \|M\| \), we obtain \( \|M\| - u^\top Mu \leq \|M\|/2 \), whence \( u^\top Mu \geq \|M\|/2 \), which establishes (4). We are now ready to apply the union bound: Recall that from the Chernoff bound on \( v^\top Mv \), we had \( \mathbb{P}[v^\top Mv \geq t] \leq 2^n \exp(-nt/\sigma^2) \), so

\[
\mathbb{P}[\sup_{v \in \mathcal{N}_{1/4}} v^\top Mv \geq t] \leq 9^d 2^n \exp(-nt/\sigma^2).
\]

Solving for this quantity to equal \( \delta \), we obtain

\[
t = \frac{\sigma^2}{n} \cdot (n \log(2) + d \log(9) + \log(1/\delta)) = \mathcal{O}(\sigma^2 \cdot (1 + d/n + \log(1/\delta)/n)),
\]

as was to be shown. \( \square \)

**VC dimension.** Our final example will be important in the following section: it concerns how quickly a family of events with certain geometric structure converges to its expectation. Let \( \mathcal{H} \) be a collection of functions \( f : \mathcal{X} \to \{0,1\} \), and define the **VC dimension** \( \text{vc}(\mathcal{H}) \) to be the maximum \( d \) for which there are points \( x_1, \ldots, x_d \) such that \( (f(x_1), \ldots, f(x_d)) \) can take on all \( 2^d \) possible values. For instance:

- If \( \mathcal{X} = \mathbb{R} \) and \( \mathcal{H} = \{1[x \geq \tau] \mid \tau \in \mathbb{R}\} \) is the family of threshold functions, then \( \text{vc}(\mathcal{H}) = 1 \).
- If \( \mathcal{X} = \mathbb{R}^d \) and \( \mathcal{H} = \{1[y \cdot v \geq \tau] \mid v \in \mathbb{R}^d, \tau \in \mathbb{R}\} \) is the family of half-spaces, then \( \text{vc}(\mathcal{H}) = d + 1 \).

Additionally, for a point set \( S = \{x_1, \ldots, x_n\} \), let \( V_{\mathcal{H}}(S) \) denote the number of distinct values of \( (f(x_1), \ldots, f(x_n)) \) and \( V_{\mathcal{H}}(n) = \max\{V_{\mathcal{H}}(S) \mid |S| = n\} \). Thus the VC dimension is exactly the maximum \( n \) such that \( V_{\mathcal{H}}(n) = 2^n \).

We will show the following:

**Proposition 0.3.** Let \( \mathcal{H} \) be a family of functions with \( \text{vc}(\mathcal{H}) = d \), and let \( X_1, \ldots, X_n \sim p \) be i.i.d. random variables over \( \mathcal{X} \). For \( f : \mathcal{X} \to \{0,1\} \), let \( \nu_n(f) = \frac{1}{n} |\{i \mid f(X_i) = 1\}| \) and let \( \nu(f) = p(f(X) = 1) \). Then

\[
\sup_{f \in \mathcal{H}} |\nu_n(f) - \nu(f)| \leq \mathcal{O}\left(\sqrt{\frac{d + \log(1/\delta)}{n}}\right)
\]

with probability \( 1 - \delta \).

We will prove a weaker result that has a \( d \log(n) \) factor instead of \( d \), and which bounds the expected value rather than giving a probability \( 1 - \delta \) bound. The \( \log(1/\delta) \) tail bound follows from **McDiarmid’s inequality**, which is a standard result in a probability course but requires tools that would take us too far afield. Removing the \( \log(n) \) factor is slightly more involved and uses a tool called **chaining**.

**Proof of Proposition 0.3.** The importance of the VC dimension for our purposes lies in the Sauer-Shelah lemma:

**Lemma 0.4** (Sauer-Shelah). Let \( d = \text{vc}(\mathcal{H}) \). Then \( V_{\mathcal{H}}(n) \leq \sum_{k=0}^d \binom{n}{k} \leq 2n^d \).

It is tempting to union bound over the at most \( V_{\mathcal{H}}(n) \) distinct values of \( (f(X_1), \ldots, f(X_n)) \); however, this doesn’t work because revealing \( X_1, \ldots, X_n \) uses up all of the randomness in the problem and we have no randomness left from which to get a concentration inequality! We will instead have to introduce some new randomness using a technique called **symmetrization**.
Regarding the expectation, let \(X_1', \ldots, X_n'\) be independent copies of \(X_1, \ldots, X_n\) and let \(\nu_n'(f)\) denote the version of \(\nu_n(f)\) computed with the \(X_i'\). Then we have

\[
\mathbb{E}_{X} \left[ \sup_{f \in \mathcal{H}} |\nu_n(f) - \nu(f)| \right] \leq \mathbb{E}_{X, X'} \left[ \sup_{f \in \mathcal{H}} |\nu_n(f) - \nu_n'(f)| \right]
\]

(11)

\[
= \frac{1}{n} \mathbb{E}_{X, X'} \left[ \sup_{f \in \mathcal{H}} \sum_{i=1}^{n} f(X_i) - f(X_i') \right].
\]

(12)

We can create our new randomness by noting that since \(X_i\) and \(X_i'\) are identically distributed, \(f(X_i) - f(X_i')\) has the same distribution as \(s_i (f(X_i) - f(X_i'))\), where \(s_i\) is a random sign variable that is \(\pm 1\) with equal probability. Introducing these variables and continuing the inequality, we thus have

\[
\frac{1}{n} \mathbb{E}_{X, X'} \left[ \sup_{f \in \mathcal{H}} \left| \sum_{i=1}^{n} f(X_i) - f(X_i') \right| \right] = \frac{1}{n} \mathbb{E}_{X, X', s} \left[ \sup_{f \in \mathcal{H}} \left| \sum_{i=1}^{n} s_i (f(X_i) - f(X_i')) \right| \right].
\]

(13)

We now have enough randomness to exploit the Sauer-Shelah lemma. If we fix \(X\) and \(X'\), note that the quantities \(f(X_i) - f(X_i')\) take values in \([-1, 1]\) and collectively can take on at most \(V_H(n)^2 = \mathcal{O}(n^{2d})\) values. But for fixed \(X, X'\), the quantities \(s_i (f(X_i) - f(X_i'))\) are independent, zero-mean, bounded random variables and hence for fixed \(f\) we have \(\mathbb{P}[\sum_i s_i (f(X_i) - f(X_i')) \geq t] \leq \exp(-t^2/9n)\) by Hoeffding’s inequality. Union bounding over the \(\mathcal{O}(n^{2d})\) effectively distinct \(f\), we obtain

\[
\mathbb{P}_{s} \left[ \sup_{f \in \mathcal{H}} \left| \sum_{i=1}^{n} s_i (f(X_i) - f(X_i')) \right| \geq t \mid X, X' \right] \leq \mathcal{O}(n^{2d}) \exp(-t^2/9n).
\]

(14)

This is small as long as \(t \gg \sqrt{nd \log n}\), so (13) is \(\mathcal{O}(\sqrt{d \log n/n})\), as claimed.

A particular consequence of Proposition 0.3 is the Dvoretzky-Kiefer-Wolfowitz inequality:

**Proposition 0.5 (DKW inequality).** For a distribution \(p\) on \(\mathbb{R}\) and i.i.d. samples \(X_1, \ldots, X_n \sim p\), define the empirical cumulative density function as \(F_n(x) = \frac{1}{n} \sum_{i=1}^{n} \mathbb{1}[X_i \leq x]\), and the population cumulative density function as \(F(x) = p(X \leq x)\). Then \(\mathbb{P} [\sup_{x \in \mathbb{R}} |F_n(x) - F(x)| \geq t] \leq 2e^{-2nt^2}\).

This follows from applying Proposition 0.3 to the family of threshold functions.