

1 Wrap up of concentration inequalities

Let $Z = \sum_i X_i$ with X_i independent. Then we can bound the log of the tail probabilities via the following inequality:

$$\log \mathbb{P}[Z - \mathbb{E}[Z] \geq t] \leq \inf_{\lambda \geq 0} \left(\sum_i \log \mathbb{E} \left[e^{\lambda(X_i - \mathbb{E}[X_i])} \right] - \lambda t \right), \quad (1)$$

Where $\mathbb{E} \left[e^{\lambda(X_i - \mathbb{E}[X_i])} \right]$ is our moment-generating function. The proof follows by taking the exponential of the inequality in the probability and using Markov's inequality. This inequality therefore depends crucially on the MGF. We can get a rough idea of how this bound behaves by taking the Taylor series:

$$\log \mathbb{P}[Z - \mathbb{E}[Z] \geq t] \leq \inf_{\lambda \geq 0} \left(\sum_i \log \underbrace{\mathbb{E} \left[e^{\lambda(X_i - \mathbb{E}[X_i])} \right]}_{\approx 1 + \lambda^2 \text{Var}[X_i]} - \lambda t \right),$$

Let's derive these bounds for some examples (MGF).

1.1 Example 1: Gaussian

Let $X_i \sim N(\mu, \sigma^2)$. This bound is relatively straightforward to prove since the Gaussian is a stable distribution (see HW 0). However, we can also use Equation 1 to establish a tail bound. First, let's look at the MGF in this case:

$$M(\lambda) = \mathbb{E} \left[e^{\lambda(X - \mathbb{E}[X])} \right].$$

Since $X - \mathbb{E}[X] \sim N(0, \sigma^2)$, by taking the integral, we get:

$$\mathbb{E} \left[e^{\lambda(X - \mathbb{E}[X])} \right] = \int_{\mathbb{R}} e^{\lambda x} e^{-x^2/(2\sigma^2)} dx = e^{\sigma^2 \lambda^2 / 2}$$

so the log is just

$$\log \mathbb{E} \left[e^{\lambda(X - \mathbb{E}[X])} \right] = \frac{\sigma^2 \lambda^2}{2}.$$

This matches the result we expect from the heuristic up to constant scaling, since

$$\mathbb{E} \left[e^{\lambda(X_i - \mathbb{E}[X_i])} \right] \approx 1 + \lambda^2 \text{Var}[X] \implies \log \mathbb{E} \left[e^{\lambda(X_i - \mathbb{E}[X_i])} \right] \approx \lambda^2 \sigma^2.$$

Plugging back into the given bound at (1), then yields

$$\log \mathbb{P}[Z - \mathbb{E}[Z] \geq t] \leq \inf_{\lambda \geq 0} \left[\lambda^2 \sum_i \sigma_i^2 / 2 - \lambda t \right],$$

and, computing the minimum yields

$$\inf_{\lambda \geq 0} \left[\lambda^2 \sum_i \sigma_i^2 / 2 - \lambda t \right] = -\frac{t^2}{2 \sum_i \sigma_i^2},$$

noting that $\text{Var}[Z] = 2 \sum_i \sigma_i^2$. Therefore

$$\mathbb{P}[Z - \mathbb{E}[Z] \geq t] \leq \exp \left(-\frac{t^2}{2 \sum_i \sigma_i^2} \right),$$

1.2 Example 2: Bounded random variables

Let $X_i \in [a_i, b_i]$, w.p. 1, then necessarily we have that

$$|X_i - \mathbb{E}[X_i]| \leq |b_i - a_i|.$$

A first attempt at writing out the MGF yields

$$\log \mathbb{E} \left[e^{\lambda(X_i - \mathbb{E}[X_i])} \right] \leq \log \mathbb{E} \left[e^{\lambda(b_i - a_i)} \right] = \lambda(b_i - a_i).$$

However, taking the inf then yields a term linear in λ since

$$\log \mathbb{P}[Z - \mathbb{E}[Z] \geq t] \leq \inf_{\lambda \geq 0} \left(\lambda \sum_i (b_i - a_i) - \lambda t \right)$$

which leads to a trivial bound. Let $\sigma_i \equiv b_i - a_i$. We can try a slightly better bound by using the Taylor expansion:

$$\begin{aligned} \log \mathbb{E} \left[e^{\lambda(X - \mathbb{E}[X])} \right] &= \log \left(1 + \lambda \mathbb{E}[X_i - \mathbb{E}[X_i]] + \frac{\lambda^2}{2} \mathbb{E}[(X_i - \mathbb{E}[X_i])^2] + \dots \right) \\ &\leq \log \left(1 + \frac{\lambda^2}{2} \mathbb{E}[(X_i - \mathbb{E}[X_i])^2] + \dots \right) \\ &\leq \log \left(e^{\lambda^2 \sigma_i^2} - \lambda^2 \sigma_i^2 \right) \\ &\leq \lambda^2 \sigma_i^2 \end{aligned}$$

where the last line comes from knowing that $e^t - t \leq e^{t^2}$. In general, this doesn't quite yield optimal results (the inequality on the right can be improved to at least $\lambda^2 \sigma_i^2 / 4$, but it's good up to constants). Note that we really mostly care that the inequality is smaller than some *quadratic* in λ . This is the general definition of a sub-gaussian random variable.

Definition 1. X is (σ -)subgaussian r.v. with finite first moment and variance proxy σ^2 (note: this parameter is, in general, at least as large as the variance, but not necessarily equal) if

$$\mathbb{E} \left[e^{\lambda(X - \mathbb{E}[X])} \right] \leq e^{\sigma^2 \lambda^2 / 2}, \quad \lambda \in \mathbb{R}. \quad (2)$$

Subgaussian variables have a nice tail bound: if X is σ -subgaussian then this implies that (note that this only requires that (2) hold for $\lambda \geq 0$),

$$\mathbb{P}[X - \mathbb{E}[X] \geq t] \leq e^{-t^2 / (2\sigma^2)}.$$

In the case where (2) holds not just for positive λ but for all λ , we get that

$$\mathbb{P}[X - \mathbb{E}[X] \leq -t] \leq e^{-t^2/(2\sigma^2)},$$

where the proof is identical to the case of the positive tail.

Theorem 1. *If X_i are independent and σ_i -subgaussian, then the variance proxy for $Z = \sum_i X_i$ is also subgaussian with proxy $\sum_i \sigma_i^2$. Putting this together, this yields*

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

Proof.

$$\mathbb{E}\left[e^{\lambda(Z - \mathbb{E}[Z])}\right] = \prod_i \mathbb{E}\left[e^{\lambda(X_i - \mathbb{E}[X_i])}\right] \leq \prod_i e^{\lambda^2 \sigma_i^2 / 2} = \exp\left(\frac{\lambda^2 \sum_i \sigma_i^2}{2}\right)$$

□

2 Rademacher Complexity

Let's briefly recall the setup: we care about uniform convergence. That is, we care about

$$\sup_{h \in H} |\hat{L}(h) - L(h)|.$$

Which lets us prove our “final goal,” which is that w.p. $1 - \delta$, we have that

$$\sup_{h \in H} |\hat{L}(h) - L(h)| \leq \text{something}.$$

For a moment, let's consider a weaker goal: what if we only care about the expectation, defining, as usual, that z_i are the complete labelled points (e.g. $z_i = (x_i, y_i)$) for simpler notation:

$$\mathbb{E}_{\{z_i\}} \left[\sup_{h \in H} |\hat{L}(h) - L(h)| \right] \leq \text{something}.$$

If we do have this, then we can just apply Markov's inequality to talk about bounds.

Definition 2. *Let F be a family of real-valued functions $f : Z \rightarrow \mathbb{R}$, where $Z = X \times Y$ (in the set sense). Then the Rademacher complexity of F is defined as*

$$R_n(F) \equiv \mathbb{E}_{\{z_i\}} \left[\mathbb{E}_{\{\sigma_i\} \sim \{-1, +1\}^n} \left[\sup_{f \in F} \frac{1}{n} \sum_i \sigma_i f(z_i) \right] \right].$$

In some sense, this is how much the output $f(z_1), \dots, f(z_n)$ can be correlated with a random pattern $\{\sigma_i\} \sim \{-1, +1\}^n$, if we're able to pick any function f from the family of functions.¹

Theorem 2.

$$\mathbb{E}_{\{z_i\}} \left[\sup_{f \in F} \frac{1}{n} \sum_i (f(z_i) - \mathbb{E}_{z \sim p} [f(z)]) \right] \leq 2R_n(F)$$

¹If we can correlate arbitrary patterns, given i.i.d. inputs (over our space of inputs), this indicates that our class of functions, F , is fairly complete w.r.t. the distribution of inputs.

Corollary 1. *Let $F = \{z = (x, y) \mapsto \ell(z, h) : h \in H\}$, which is a family of losses, then, applying the statement above yields*

$$\begin{aligned} \mathbb{E}_{\{z_i\}} \left[\sup_{h \in H} \left| \frac{1}{n} \sum_i (\ell(z_i, h) - \mathbb{E}_{z \sim p} [\ell(z, h)]) \right| \right] &= \mathbb{E}_{\{z_i\}} \left[\sup_{h \in H} |\hat{L}(h) - L(h)| \right] \\ &\leq 2R_n(F \cup F^-) = 2\mathbb{E}_{\{\sigma_i, z_i\}} \left[\sup_{h \in H} \left| \frac{1}{n} \sum_{i=1}^n \sigma_i \ell(z_i, h) \right| \right], \end{aligned}$$

where $F^- = \{-f \mid f \in F\}$ is the negative family of functions.

This variant is obtained by applying Theorem 2 to the family $F \cup F^-$. See Section 2.1 for a detailed version of this argument. We also note that this Corollary is not necessarily needed for the final bound for the excess risk. See the proof in Theorem 3 of scribe notes 6 for a way to avoid bounding $\mathbb{E}_{\{z_i\}} \left[\sup_{h \in H} |\hat{L}(h) - L(h)| \right]$.

This implies that

generalization error \lesssim how well losses can be correlated with random sign patterns.

Of course, the correlation in this case depends heavily on the distribution that we draw samples from (e.g. a simple counterexample to our intuitive notion of ‘generalization error’ is to set p to be a single point mass at any one point, then $R_n(F) \rightarrow 0$ as $n \rightarrow \infty$ in a way that does not depend on the particular choice of p). Returning to Theorem 2, we can give a proof by symmetrization.

Proof. Fix $z = \{z_1, \dots, z_n\}$, and let’s look at

$$\begin{aligned} \sup_{f \in F} \left(\frac{1}{n} \sum_i f(z_i) - \mathbb{E}[f] \right) &= \sup_{f \in F} \left(\frac{1}{n} \sum_i f(z_i) - \mathbb{E}_{z'} \left[\frac{1}{n} \sum_i f(z'_i) \right] \right) \\ &= \sup_{f \in F} \mathbb{E}_{z'} \left[\frac{1}{n} \sum_i f(z_i) - \frac{1}{n} \sum_i f(z'_i) \right] \\ &\leq \mathbb{E}_{z'} \left[\sup_{f \in F} \left(\frac{1}{n} \sum_i f(z_i) - \frac{1}{n} \sum_i f(z'_i) \right) \right], \end{aligned}$$

where the inequality is by Fatou’s lemma. Therefore, taking the expectation of both sides yields

$$\begin{aligned} \mathbb{E}_z \left[\sup_{f \in F} \frac{1}{n} \sum_i f(z_i) - \mathbb{E}[f] \right] \\ \leq \mathbb{E}_z \left[\mathbb{E}_{z'} \left[\sup_{f \in F} \frac{1}{n} \sum_i (f(z_i) - f(z'_i)) \right] \right]. \end{aligned}$$

Here, note that $f(z_i) - f(z'_i) \stackrel{d}{=} \sigma_i(f(z_i) - f(z'_i))$, for $\sigma_i \sim \{\pm 1\}$, so the expectations must all be equivalent, e.g.

$$\mathbb{E}_{z, z'} \left[\sup_{f \in F} \left(\frac{1}{n} \sum_i (f(z_i) - f(z'_i)) \right) \right] = \mathbb{E}_{\sigma, z, z'} \left[\sup_{f \in F} \left(\frac{1}{n} \sum_i \sigma_i (f(z_i) - f(z'_i)) \right) \right].$$

Finally, noting that, for any functions a, b

$$\sup_z (a(z) - b(z)) \leq \sup_z a(z) + \sup_{z'} (-b(z')),$$

and that $-\sigma_i \stackrel{d}{=} \sigma_i$, then

$$\begin{aligned} & \mathbb{E}_{\sigma, z, z'} \left[\sup_{f \in F} \left(\frac{1}{n} \sum_i (f(z_i) - f(z'_i)) \right) \right] \\ & \leq \mathbb{E}_{\sigma, z, z'} \left[\sup_{f \in F} \left(\frac{1}{n} \sum_i \sigma_i f(z_i) \right) \right] + \mathbb{E}_{\sigma, z, z'} \left[\sup_{f \in F} \left(\frac{1}{n} \sum_i -\sigma_i f(z_i) \right) \right] \\ & = 2 \mathbb{E}_{\sigma, z, z'} \left[\sup_{f \in F} \left(\frac{1}{n} \sum_i \sigma_i f(z_i) \right) \right] \\ & = 2R_n(F). \end{aligned}$$

□

2.1 Dealing with the absolute value

We now show Corollary 1 from Theorem 2. Let us define $F^- = \{-f \mid f \in F\}$, then we have

$$\begin{aligned} & \mathbb{E}_z \left[\sup_{f \in F} |\hat{L}(f) - L(f)| \right] = \mathbb{E}_z \left[\max \left\{ \sup_{f \in F} \hat{L}(f) - L(f), \sup_{f \in F} \hat{L}(-f) - L(-f) \right\} \right] \\ & = \mathbb{E}_z \left[\max \left\{ \sup_{f \in F} \hat{L}(f) - L(f), \sup_{f \in F^-} \hat{L}(f) - L(f) \right\} \right] \\ & \leq 2R_n(F \cup F^-) \\ & = 2 \mathbb{E}_{\sigma, z} \left[\sup_{f \in F \cup F^-} \frac{1}{n} \sum_{i=1}^n \sigma_i f(z_i) \right] = 2 \mathbb{E}_{\sigma, z} \left[\sup_{f \in F} \left| \frac{1}{n} \sum_{i=1}^n \sigma_i f(z_i) \right| \right], \end{aligned}$$

where we have applied Theorem 2 on the class $F \cup F^-$.

If F contains the all 0 function, it turns out that we can further bound this by the standard Rademacher complexity of f as shown in the following lemma. The lemma is also useful for Lecture 8 in week 4 when we compute the Rademacher complexity of neural networks. We also note that the lemma below is not true without the assumption that $f_0 \in F$. (In the lecture, I (Tengyu) claimed this incorrectly.)

Lemma 1. *Let f_0 denote the all 0 function, and suppose $f_0 \in F$. Then*

$$\mathbb{E}_{\sigma, z} \left[\sup_{f \in F} \left| \frac{1}{n} \sum_{i=1}^n \sigma_i f(z_i) \right| \right] \leq 2R_n(F)$$

Proof. By the definition of supremum and infimum, and since $\mathbf{0} \in F$, we have

$$\begin{aligned} \mathbb{E}_{\sigma,z} \left[\sup_{f \in F} \left| \frac{1}{n} \sum_{i=1}^n \sigma_i f(z_i) \right| \right] &\leq \mathbb{E}_{\sigma,z} \left[\left| \frac{1}{n} \sum_{i=1}^n f_0(z_i) \right| \right] + \\ &\quad \mathbb{E}_{\sigma,z} \left[\sup_{f \in F} \left| \frac{1}{n} \sum_{i=1}^n \sigma_i f(z_i) \right| - \inf_{f \in F} \left| \frac{1}{n} \sum_{i=1}^n \sigma_i f(z_i) \right| \right] \\ &\leq \mathbb{E}_{\sigma,z} \left[\sup_{f \in F} \left| \frac{1}{n} \sum_{i=1}^n \sigma_i f(z_i) \right| - \inf_{f \in F} \left| \frac{1}{n} \sum_{i=1}^n \sigma_i f(z_i) \right| \right] \end{aligned}$$

Next, we argue that for any choice of σ, z ,

$$\sup_{f \in F} \left| \frac{1}{n} \sum_{i=1}^n \sigma_i f(z_i) \right| - \inf_{f \in F} \left| \frac{1}{n} \sum_{i=1}^n \sigma_i f(z_i) \right| \leq \sup_{f \in F} \frac{1}{n} \sum_{i=1}^n \sigma_i f(z_i) - \inf_{f \in F} \frac{1}{n} \sum_{i=1}^n \sigma_i f(z_i)$$

The simplest way to see this is casework on the signs of $\frac{1}{n} \sum_{i=1}^n \sigma_i f(z_i)$ achieving the supremum and infimum. Let s_1 be the sign of $\frac{1}{n} \sum_{i=1}^n \sigma_i f(z_i)$ that achieves the supremum, and s_2 be the sign of the value achieving the infimum.

Case 1) $s_1 < 0, s_2 < 0$. In this case, $\inf_{f \in F} \frac{1}{n} \sum_{i=1}^n \sigma_i f(z_i) \leq -\sup_{f \in F} \left| \frac{1}{n} \sum_{i=1}^n \sigma_i f(z_i) \right|$, and $\sup_{f \in F} \frac{1}{n} \sum_{i=1}^n \sigma_i f(z_i) \geq -\inf_{f \in F} \left| \frac{1}{n} \sum_{i=1}^n \sigma_i f(z_i) \right|$. Subtracting the former from the latter gives the desired statement.

Case 2) $s_1 < 0, s_2 \geq 0$. Again, $\inf_{f \in F} \frac{1}{n} \sum_{i=1}^n \sigma_i f(z_i) \leq -\sup_{f \in F} \left| \frac{1}{n} \sum_{i=1}^n \sigma_i f(z_i) \right|$. Also, $\sup_{f \in F} \frac{1}{n} \sum_{i=1}^n \sigma_i f(z_i) \geq 0$. Again subtracting the former from the latter gives the desired statement.

Case 3) $s_1 \geq 0, s_2 < 0$. Now $\sup_{f \in F} \frac{1}{n} \sum_{i=1}^n \sigma_i f(z_i) \geq \sup_{f \in F} \left| \frac{1}{n} \sum_{i=1}^n \sigma_i f(z_i) \right|$. Furthermore, $\inf_{f \in F} \frac{1}{n} \sum_{i=1}^n \sigma_i f(z_i) \leq 0$. Subtracting the latter from the former gives the desired statement.

Case 4) $s_1 \geq 0, s_2 \geq 0$. Now $\sup_{f \in F} \frac{1}{n} \sum_{i=1}^n \sigma_i f(z_i) \geq \sup_{f \in F} \left| \frac{1}{n} \sum_{i=1}^n \sigma_i f(z_i) \right|$. Furthermore, $\inf_{f \in F} \frac{1}{n} \sum_{i=1}^n \sigma_i f(z_i) \leq \inf_{f \in F} \left| \frac{1}{n} \sum_{i=1}^n \sigma_i f(z_i) \right|$. Subtracting the latter from the former gives the desired statement.

Thus, it follows that

$$\mathbb{E}_{\sigma,z} \left[\sup_{f \in F} \left| \frac{1}{n} \sum_{i=1}^n \sigma_i f(z_i) \right| \right] \leq \mathbb{E}_{\sigma,z} \left[\sup_{f \in F} \frac{1}{n} \sum_{i=1}^n \sigma_i f(z_i) - \inf_{f \in F} \frac{1}{n} \sum_{i=1}^n \sigma_i f(z_i) \right]$$

Now we note that $\inf_{f \in F} \frac{1}{n} \sum_{i=1}^n \sigma_i f(z_i) = -\sup_{f \in F} -\frac{1}{n} \sum_{i=1}^n \sigma_i f(z_i)$. Since $-\sigma_i$ and σ_i follow the same distribution, we conclude that

$$\mathbb{E}_{\sigma,z} \left[\inf_{f \in F} \frac{1}{n} \sum_{i=1}^n \sigma_i f(z_i) \right] = -\mathbb{E}_{\sigma,z} \left[\sup_{f \in F} \frac{1}{n} \sum_{i=1}^n \sigma_i f(z_i) \right]$$

Plugging this into before,

$$\mathbb{E}_{\sigma,z} \left[\sup_{f \in F} \left| \frac{1}{n} \sum_{i=1}^n \sigma_i f(z_i) \right| \right] \leq 2\mathbb{E}_{\sigma,z} \left[\sup_{f \in F} \frac{1}{n} \sum_{i=1}^n \sigma_i f(z_i) \right] = 2R_n(F)$$

□