

Lecture 3: September 1

Lecturer: Siva Balakrishnan

This lecture will be mostly a summary of other useful exponential tail bounds. We will not prove any of these in lecture, some of them follow similar lines of using Chernoff's method in clever ways. I can provide references if you are curious. In particular, we will go through:

1. Bernstein's inequality: sharper concentration for bounded random variables
2. Azuma's inequality: Concentration of Lipschitz functions of bounded random variables
3. Levy's inequality/Tsirelson's inequality: Concentration of Lipschitz functions of Gaussian random variables
4. χ^2 tail bound

Finally, we will see an application of the χ^2 tail bound in proving the Johnson-Lindenstrauss lemma.

3.1 Bernstein's inequality

One nice thing about the Gaussian tail inequality was that it explicitly depended on the variance of the random variable X , i.e. roughly the inequality guaranteed us that the deviation from the mean was at most $\sigma\sqrt{\log(2/\delta)}/n$ with probability at least $1 - \delta$.

On the other hand Hoeffding's bound depended only on the bounds of the random variable but not explicitly on the variance of the RVs. The bound $b - a$, provides a (possibly loose) upper bound on the variance. One might at least hope that if the random variables were bounded, and additionally had *small variance* we might be able to improve Hoeffding's bound.

This is indeed the case. Such inequalities are typically known as Bernstein inequalities. As a concrete example, suppose we had X_1, \dots, X_n which were i.i.d from a distribution with mean zero, bounded support $[a, b]$, with variance $\mathbb{E}[X^2] = \sigma^2$. Then,

$$\mathbb{P}(|\hat{\mu} - \mu| \geq t) \leq 2 \exp \left(-\frac{nt^2}{2(\sigma^2 + (b-a)t)} \right).$$

Roughly this inequality says with probability at least $1 - \delta$,

$$|\hat{\mu} - \mu| \leq 4\sigma \sqrt{\frac{\ln(2/\delta)}{n}} + \frac{4(b-a) \ln(2/\delta)}{n}.$$

As an exercise work through the above algebra.

Upto some small constants this is never worse than Hoeffding's bound, which just comes from using the worst-case upper bound of $\sigma \leq b - a$. When the RVs have small variance, i.e. σ is small, this bound can be much sharper than Hoeffding's bound. These are cases where one has a random variable that occasionally takes large values (so the bounds are not great) but has much smaller variance.

Intuitively, it captures more of the Chebyshev effect, i.e. that random variables with small variance should be tightly concentrated around their mean.

3.2 Azuma's inequality

So far we have mostly been focused on the concentration of averages. A natural question is whether other functions of i.i.d. random variables also show exponential concentration. It turns out that many other functions do concentrate sharply, and roughly the main property of the function that we need is that if we change the value of one random variable the function does not change dramatically.

Formally, we have i.i.d. RVs X_1, \dots, X_n , where each $X_i \in \mathbb{R}$. We have a function $f : \mathbb{R}^n \mapsto \mathbb{R}$, that satisfies the property that:

$$|f(x_1, \dots, x_n) - f(x_1, \dots, x_{k-1}, x'_k, x_{k+1}, \dots, x_n)| \leq L_k,$$

for every $x, x' \in \mathbb{R}^n$, i.e. the function changes by at most L_k if its k -th co-ordinate is changed. This is known as the bounded difference condition.

If the random variables X_1, \dots, X_n are i.i.d then for all $t \geq 0$

$$\mathbb{P}(|f(X_1, \dots, X_n) - \mathbb{E}[f(X_1, \dots, X_n)]| \geq t) \leq 2 \exp \left(- \frac{2t^2}{\sum_{k=1}^n L_k^2} \right).$$

Example 1: A simple example of this inequality in action is to see that it directly implies the Hoeffding bound. In this case the function of interest is the average:

$$f(X_1, \dots, X_n) = \frac{1}{n} \sum_{i=1}^n X_i,$$

and since the random variables are bounded we have that each $L_k \leq (b - a)/n$. This in turn directly yields Hoeffding's bound (with slightly better constants).

Example 2: A perhaps more interesting example is that of U -statistics. A U -statistic is defined by a kernel, which is just a function of two random variables, i.e. $g : \mathbb{R}^2 \mapsto \mathbb{R}$. The

U-statistic is then given as:

$$U(X_1, \dots, X_n) := \frac{1}{\binom{n}{2}} \sum_{j < k} g(X_j, X_k).$$

There are many examples of U -statistics, for instance:

1. **Variance:** The usual estimator of the sample variance:

$$\hat{\sigma} = \frac{1}{n-1} \sum_{i=1}^n (X_i - \hat{\mu})^2,$$

is the U -statistic that arises from taking $g(X_j, X_k) = \frac{1}{2}(X_j - X_k)^2$.

2. **Mean absolute deviation:** If we take $g(X_j, X_k) = |X_j - X_k|$, this leads to a U -statistic that is an unbiased estimator of the mean absolute deviation $\mathbb{E}|X_1 - X_2|$.

For bounded U -statistics, i.e. if $g(X_i, X_j) \leq b$, we can apply Azuma's inequality to obtain a concentration bound. Note that since each random variable X_i participates in $(n-1)$ terms we have that,

$$|U(X_1, \dots, X_n) - U(X_1, \dots, X'_i, \dots, X_n)| \leq \frac{1}{\binom{n}{2}} (n-1)(2b) = \frac{4b}{n}.$$

So that Azuma's inequality tells us that,

$$\mathbb{P}(|U(X_1, \dots, X_n) - \mathbb{E}[U(X_1, \dots, X_n)]| \geq t) \leq 2 \exp(-nt^2/(8b^2)).$$

3.3 Levy's inequality

There is a similar concentration inequality that applies to functions of Gaussian random variables that are sufficiently smooth. In this case, the assumption is quite different. We assume that:

$$|f(X_1, \dots, X_n) - f(Y_1, \dots, Y_n)| \leq L \sqrt{\sum_{i=1}^n (X_i - Y_i)^2},$$

for all $X_1, \dots, X_n, Y_1, \dots, Y_n \in \mathbb{R}$.

For such functions we have that if $X_1, \dots, X_n \sim N(0, 1)$ then,

$$\mathbb{P}(|f(X_1, \dots, X_n) - \mathbb{E}[f(X_1, \dots, X_n)]| \geq t) \leq 2 \exp\left(-\frac{t^2}{2L^2}\right).$$

3.4 χ^2 tail bounds

A χ^2 random variable with n degrees of freedom, denoted by $Y \sim \chi_n^2$, is a RV that is a sum of n i.i.d. standard Gaussian RVs, i.e. $Y = \sum_{i=1}^n X_i^2$ where each $X_i \sim N(0, 1)$. The expected value $\mathbb{E}[X_i^2] = 1$, and we have the χ^2 tail bound:

$$\mathbb{P}\left(\left|\frac{1}{n} \sum_{k=1}^n Z_k^2 - 1\right| \geq t\right) \leq 2 \exp(-nt^2/8) \quad \text{for all } t \in (0, 1).$$

You will derive this in your HW using the Chernoff method. Analogous to the class of sub-Gaussian RVs, χ^2 random variables belong to a class of what are known as *sub-exponential* random variables. The main note-worthy difference is that the Gaussian-type behaviour of the tail only holds for small values of the deviation t .

Detour: The union bound. This is also known as Boole's inequality. It says that if we have events A_1, \dots, A_n then

$$\mathbb{P}\left(\bigcup_{i=1}^n A_i\right) \leq \sum_{i=1}^n \mathbb{P}(A_i).$$

In particular, if we consider a case when each event A_i is a failure of some type, then the above inequality says that the probability that even a single failure occurs is at most the sum of the probabilities of each failure.

Example: The Johnson-Lindenstrauss Lemma. One very nice application of χ^2 tail bounds is in the analysis of what are known as “random projections”. Suppose we have a data set $X_1, \dots, X_n \in \mathbb{R}^d$ where d is quite large. Storing such a dataset might be expensive and as a result we often resort to “sketching” or “random projection” where the goal is to create a map $F : \mathbb{R}^d \mapsto \mathbb{R}^m$, with $m \ll d$. We then instead store the mapped dataset $\{F(X_1), \dots, F(X_n)\}$. The challenge is to design this map F in a way that preserves essential features of the original dataset. In particular, we would like that for every pair (X_i, X_j) we have that,

$$(1 - \epsilon) \|X_i - X_j\|_2^2 \leq \|F(X_i) - F(X_j)\|_2^2 \leq (1 + \epsilon) \|X_i - X_j\|_2^2,$$

i.e. the map preserves all the pair-wise distances up to a $(1 \pm \epsilon)$ factor. Of course, if m is large we might expect this is not too difficult.

The Johnson-Lindenstrauss lemma is quite stunning: it says that a simple randomized construction will produce such a map with probability at least $1 - \delta$ provided that,

$$m \geq \frac{16 \log(n/\delta)}{\epsilon^2}.$$

Notice that this is completely independent of the original dimension d and depends on logarithmically on the number of points n . This map can result in huge savings in storage cost while still essentially preserving all the pairwise distances.

The map itself is quite simple: we construct a matrix $Z \in \mathbb{R}^{m \times d}$, where each entry of Z is i.i.d $N(0, 1)$. We then define the map as:

$$F(X_i) = \frac{ZX_i}{\sqrt{m}}.$$

Now let us fix a pair (X_j, X_k) and consider,

$$\begin{aligned} \frac{\|F(X_j) - F(X_k)\|_2^2}{\|X_j - X_k\|_2^2} &= \left\| \frac{Z(X_j - X_k)}{\sqrt{m}\|X_j - X_k\|_2} \right\|_2^2 \\ &= \frac{1}{m} \sum_{i=1}^m \underbrace{\langle Z_i, \frac{X_j - X_k}{\|X_j - X_k\|_2} \rangle^2}_{T_i}. \end{aligned}$$

Now, for some fixed numbers a_j the distribution of $\sum_{j=1}^d a_j Z_{ij}$ is Gaussian with mean 0 and variance $\sum_{j=1}^d a_j^2$. So each term T_i is an independent χ^2 random variable. Now applying the χ^2 tail bound, we obtain that,

$$\mathbb{P} \left(\left| \frac{\|F(X_j) - F(X_k)\|_2^2}{\|X_j - X_k\|_2^2} - 1 \right| \geq \epsilon \right) \leq 2 \exp(-m\epsilon^2/8).$$

Thus for the fixed pair (X_i, X_j) the probability that our map fails to preserve the distance is exponentially small, i.e. is at most $2 \exp(-m\epsilon^2/8)$. Now, to find the probability that our map fails to preserve *any* of our $\binom{n}{2}$ pairwise distances we simply apply the union bound to conclude that, the probability of any failure is at most:

$$\mathbb{P}(\text{failure}) \leq 2 \binom{n}{2} \exp(-m\epsilon^2/8).$$

Now, it is straightforward to verify that if

$$m \geq \frac{16 \log(n/\delta)}{\epsilon^2},$$

then this probability is at most δ as desired. An important point to note is that the *exponential concentration* is what leads to such a small value for m (i.e. it only needs to grow logarithmically with the sample size).