ECE 901 Spring 2014 Statistical Learning Theory

instructor: R. Nowak

Concentration Inequalities

1 Convergence of Sums of Independent Random Variables

The most important form of statistic considered in this course is a sum of independent random variables.

Example 1. *A biologist is studying the new artificial lifeform called* synthia. *She is interested to see if the synthia cells can survive in cold conditions. To test synthia's hardiness, the biologist will conduct n independent experiments. She has grown n cell cultures under ideal conditions and then exposed each to cold conditions. The number of cells in each culture is measured before and after spending one day in cold conditions. The fraction of cells surviving the cold is recorded. Let $x_1, \ldots, x_n$ denote the recorded fractions. The average $\widehat{p} := \frac{1}{n}\sum_{i=1}^n x_i$ is an estimator of the survival probability.*

Understanding behavior of sums of independent random variables is extremely important. For instance, the biologist in the example above would like to know that the estimator is reasonably accurate. Let $X_1, \ldots, X_n$ be independent and identically distributed random variables with variance $\sigma^2 < \infty$ and consider the average $\widehat{\mu} := \frac{1}{n}\sum_{i=1}^n X_i$. First note that $\mathbb{E}[\widehat{\mu}] = \mathbb{E}[X]$. An easy calculation shows that the variance of $\widehat{\mu}$ is σ^2/n. So the average has the same mean value as the random variables and the variance is reduced by a factor of n. Lower variance means less uncertainty. So it is possible to reduce uncertainty by averaging. The more we average, the less the uncertainty (assuming, as we are, that the random variables are independent, which implies they are uncorrelated).

The argument above quantifies the effect of averaging on the variance, but often we would like to say more about the distribution of the average. The *Central Limit Theorem* is a classic result showing that the probability distribution of the average of n independent and identically distributed random variables with mean μ and variance $\sigma^2 < \infty$ tends to a Gaussian distribution with mean μ and variance σ^2/n, regardless of the form of the distribution of the variables. By 'tends to' we mean in the limit as n tends to infinity.

In many applications we would like to say something more about the distributional characteristics for finite values of n. One approach is to calculate the distribution of the average explicitly. Recall that if the random variables have a density p_X, then the density of the sum $\sum_{i=1}^n X_i$ is the n-fold convolution of the density p_X with itself (again this hinges on the assumption that the random variables are independent; it is easy to see by considering the characteristic function of the sum and recalling that multiplication of Fourier transforms is equivalent to convolution in the inverse domain). However, this exact calculation can be sometimes difficult or impossible, if for instance we don't know the density p_X, and so sometimes probability bounds are more useful.

Let Z be a non-negative random variable and take $t > 0$. Then

$$\begin{aligned} \mathbb{E}[Z] &\geq \mathbb{E}[Z\,\mathbf{1}_{Z\geq t}] \\ &\geq \mathbb{E}[t\,\mathbf{1}_{Z\geq t}] \;=\; t\,\mathbb{P}(Z \geq t) \end{aligned}$$

The result $\mathbb{P}(Z \geq t) \leq \mathbb{E}[Z]/t$ is called *Markov's Inequality*. We can generalize this inequality as follows. Let ϕ be any non-decreasing, non-negative function. Then

$$\mathbb{P}(Z \geq t) \;=\; \mathbb{P}(\phi(Z) \geq \phi(t)) \;\leq\; \frac{\mathbb{E}[\phi(Z)]}{\phi(t)}\ .$$

We can use this to get a bound on the probability 'tails' of any random variable Z. Let $t > 0$

$$\begin{aligned}\mathbb{P}(|Z - \mathbb{E}[Z]| \geq t) &= P((Z - \mathbb{E}[Z])^2 \geq t^2) \\ &\leq \frac{\mathbb{E}[(Z - \mathbb{E}[Z])^2]}{t^2} \\ &= \frac{\mathrm{Var(Z)}}{t^2} ,\end{aligned}$$

where Var(Z) denotes the variance of Z. This inequality is known as *Chebyshev's Inequality*. If we apply this to the average $\widehat{\mu} = \frac{1}{n}\sum_{i=1}^{n} X_i$, then we have

$$\mathbb{P}(|\widehat{\mu} - \mu| \geq t) \leq \frac{\sigma^2}{nt^2}$$

where μ and σ^2 are the mean and variance of the random variables $\{X_i\}$. This shows that not only is the variance reduced by averaging, but the tails of the distribution (probability of observing values a distance of more than t from the mean) are smaller.

The tail bound given by Chebyshev's Inequality is loose, and much tighter bounds are possible under slightly stronger assumptions. For example, if $X_i \overset{iid}{\sim} \mathcal{N}(\mu, 1)$, then $\widehat{\mu} \sim \mathcal{N}(\mu, 1/n)$. The following tail-bound for the Gaussian density shows that in this case $\mathbb{P}(|\widehat{\mu} - \mu| \geq t) \leq e^{-nt^2/2}$.

Theorem 1. *The tail of the standard Gaussian $\mathcal{N}(0,1)$ distribution satisfies the bound for any $t \geq 0$,*

$$\frac{1}{\sqrt{2\pi}} \int_t^\infty e^{\frac{-x^2}{2}} dx \leq \min\left\{ \frac{1}{2} e^{\frac{-t^2}{2}} , \frac{1}{\sqrt{2\pi}\, t^2} e^{\frac{-t^2}{2}} \right\}$$

Proof. Consider

$$R := \frac{\frac{1}{\sqrt{2\pi}} \int_t^\infty e^{\frac{-x^2}{2}} dx}{e^{-\frac{t^2}{2}}} = \frac{1}{\sqrt{2\pi}} \int_t^\infty e^{\frac{-(x^2 - t^2)}{2}} dx = \frac{1}{\sqrt{2\pi}} \int_t^\infty e^{\frac{-(x-t)(x+t)}{2}} dx$$

For the first bound, let $y = x - t$,

$$R = \frac{1}{\sqrt{2\pi}} \int_0^\infty e^{\frac{-y(y+2t)}{2}} dy \leq \frac{1}{\sqrt{2\pi}} \int_0^\infty e^{\frac{-y^2}{2}} dy = \frac{1}{2}$$

For the second bound, note that

$$R \leq \frac{1}{\sqrt{2\pi}} \int_t^\infty e^{\frac{-2t(x-t)}{2}} dx = \frac{1}{\sqrt{2\pi}} e^{t^2} \int_t^\infty e^{-tx} dx = \frac{1}{\sqrt{2\pi}} e^{t^2} \frac{e^{-t^2}}{t} = \frac{1}{\sqrt{2\pi t^2}}$$

□

1.1 The Chernoff Method

More generally, if the random variables $\{X_i\}$ are bounded or *sub-Gaussian* (meaning the tails of the probability distribution decay at least as fast as Gaussian tails), then the tails of the average converge exponentially fast in n. The key to this sort of result is the so-called *Chernoff bounding method*, based on Markov's inequality and the exponential function (non-decreasing, non-negative). If Z is any real-valued random variable and $s > 0$, then

$$\mathbb{P}(Z > t) = \mathbb{P}(e^{sZ} > e^{st}) \leq e^{-st}\, \mathbb{E}[e^{sZ}] .$$

We can choose $s > 0$ to minimize this upper bound. In particular, if we define the function

$$\psi^*(t) \;=\; \max_{s>0} \left\{ st - \log \mathbb{E}[e^{sZ}] \right\} \;,$$

then $\mathbb{P}(Z > t) \leq e^{-\psi^*(t)}$.

Exponential bounds of this form can be obtained explicitly for many classes of random variables. One of the most important is the class of sub-Gaussian random variables. A random variable X is said to be sub-Gaussian if there exists a constant $c > 0$ such that $\mathbb{E}[e^{sX}] \leq e^{cs^2/2}$ for all $s \in \mathbb{R}$.

Theorem 2. *Let $X_1, X_2, ..., X_n$ be independent sub-Gaussian random variables such that $\mathbb{E}[e^{s(X_1 - \mathbb{E}[X_1])}] \leq e^{cs^2/2}$ for a constant $c > 0$. Let $S_n = \sum_{i=1}^n X_i$. Then for any $t > 0$, we have*

$$\mathbb{P}(|S_n - \mathbb{E}[S_n]| \geq t) \leq 2\, e^{-t^2/(2nc)}$$

and equivalently if $\widehat{\mu} := \frac{1}{n} S_n$ we have

$$\mathbb{P}(|\widehat{\mu} - \mu| \geq t) \leq 2\, e^{-nt^2/(2c)}$$

Proof.

$$\begin{aligned}
\mathbb{P}(\sum_{i=1}^n X_i - \mathbb{E}[X_i] \geq t) &\leq e^{-st} \mathbb{E}\left[e^{s\left(\sum_{i=1}^n X_i - E[X_i]\right)} \right] \\
&= e^{-st} \mathbb{E}\left[\prod_{i=1}^n e^{s(X_i - \mathbb{E}[X_i])} \right] \\
&= e^{-st} \prod_{i=1}^n \mathbb{E}\left[e^{s(X_i - \mathbb{E}[X_i])} \right] \\
&= e^{-st} e^{ncs^2/2} \\
&= e^{-t^2/(2nc)}
\end{aligned}$$

where the last step follows by taking $s = t/(nc)$. □

To apply the result above we need to verify that the sub-Gaussian condition, $\mathbb{E}\left[e^{s(X_i - \mathbb{E}[X_i])}\right] \leq e^{cs^2/2}$, holds for some $c > 0$. As the name suggests, the condition holds if the tails of the probability distribution decay like $e^{-t^2/2}$ (or faster).

Theorem 3. *If $\mathbb{P}(|X_i - \mathbb{E}[X_i]| \geq t) \leq a e^{-bt^2/2}$ holds for constants $a, b > 0$ and all $t \geq 0$, then*

$$\mathbb{E}[e^{s(X_i - \mathbb{E}[X_i])}] \leq e^{4as^2/b} \;.$$

Proof. Let X be a zero-mean random variable satisfying $\mathbb{P}(|X| \geq t) \leq a e^{-bt^2/2}$. First note since X has mean zero, Jensen's inequality implies $\mathbb{E}[e^{sX}] \geq e^{s\mathbb{E}X} = 1$ for all $s \in \mathbb{R}$. Thus, if X_1 and X_2 are two independent copies of X, then

$$\mathbb{E}[e^{s(X_1 - X_2)}] = \mathbb{E}[e^{sX_1}]\mathbb{E}[e^{-sX_2}] \geq \mathbb{E}[e^{sX_1}] = \mathbb{E}[e^{sX}] \;.$$

Thus, we can write

$$\mathbb{E}[e^{sX}] \;\leq\; \mathbb{E}[e^{s(X_1 - X_2)}] \;=\; 1 + \sum_{\ell \geq 1} \frac{s^\ell\, \mathbb{E}[(X_1 - X_2)^\ell]}{\ell\,!} \;.$$

Also, since $\mathbb{E}[(X_1 - X_2)^\ell] = 0$ for ℓ odd, we have

$$\mathbb{E}[e^{sX}] \;\leq\; 1 + \sum_{\ell \geq 1} \frac{s^{2\ell}\, \mathbb{E}[(X_1 - X_2)^{2\ell}]}{(2\ell)!} \;.$$

Next note that since $x^{2\ell}$ is convex in x by Jensen's inequality we have

$$\mathbb{E}[(X_1 - X_2)^{2\ell}] \;=\; \mathbb{E}[2^\ell (X_1/2 - X_2/2)^{2\ell}] \leq \; 2^{2\ell-1}\left(\mathbb{E}[X_1^{2\ell}] + \mathbb{E}[X_2^{2\ell}]\right) = 2^{2\ell}\mathbb{E}[X^{2\ell}] \; .$$

Next note that $\mathbb{E}[X^{2\ell}] = \int_0^\infty \mathbb{P}(X^{2\ell} > t)\, dt$ and by the change of variables $t = x^{2\ell}$ we have

$$\mathbb{E}[X^{2\ell}] \quad = \quad 2\ell \int_0^\infty x^{2\ell-1}\mathbb{P}(|X| > x)\, dx \;\leq\; 2\ell a \int_0^\infty x^{2\ell-1} e^{-bx^2/2}\, dx \; .$$

Now substitute $x = \sqrt{2y/b}$ to get

$$\mathbb{E}[X^{2\ell}] \quad \leq \quad (2/b)^\ell \ell a \int_0^\infty y^{\ell-1} e^{-y}\, dy \;=\; (2/b)^\ell a\, \ell! \; .$$

So we have $\mathbb{E}[(X_1 - X_2)^{2\ell}] \leq 2^{2\ell+1} b^{-\ell} a \ell! \leq (8a/b)^\ell \ell!$ since a must be at least 1. Now plugging this into the bound for $\mathbb{E}[e^{sX}]$ above, we see that each term in the sum is bounded by $s^{2\ell}(8a/b)^\ell \ell!/(2\ell)!$. Since $(2\ell)! \geq 2^\ell(\ell!)^2$ each term can be bounded by $(4as^2/b)^\ell/\ell!$, and so $\mathbb{E}[e^{sX}] \;\leq\; e^{4as^2/b}$. □

The simplest result of this form is for bounded random variables.

Theorem 4. *(Hoeffding's Inequality). Let $X_1, X_2, ..., X_n$ be independent bounded random variables such that $X_i \in [a_i, b_i]$ with probability 1. Let $S_n = \sum_{i=1}^n X_i$. Then for any $t > 0$, we have*

$$\mathbb{P}(|S_n - \mathbb{E}[S_n]| \geq t) \leq 2\, e^{-\frac{2t^2}{\sum_{i=1}^n (b_i - a_i)^2}}$$

Proof. We prove a special case. The more general result above also has slightly better constants, and its proof is later in the notes. Here, assume that $a \leq X_i \leq b$ with probability 1 for all i. Then the following bound

$$\mathbb{P}(|S_n - \mathbb{E}[S_n]| \geq t) \leq 2\, e^{-\frac{t^2}{2\sum_{i=1}^n (b_i - a_i)^2}}$$

follows from Theorem 3 above by noting that if $a \leq X_1, X_2 \leq b$ with probability 1, then $\mathbb{E}[(X_1 - X_2)^{2\ell}] \leq (b-a)^{2\ell}$. □

If the random variables $\{X_i\}$ are binary-valued, then this result is usually referred to as the *Chernoff Bound*. Another proof of Hoeffding's Inequality, which relies Markov's inequality and some elementary concepts from convex analysis, is given in the next section. Note that if the random variables in the average $\widehat{\mu} = \frac{1}{n}\sum_{i=1}^n X_i$ are bounded according to $a \leq X_i \leq b$. Let $c = (b-a)^2$. Then Hoeffding's Inequality implies

$$\mathbb{P}(|\widehat{\mu} - \mu| \geq t) \quad \leq \quad 2\, e^{-\frac{2nt^2}{c}} \tag{1}$$

In other words, the tails of the distribution of the average are tending to zero at an exponential rate in n, much faster than indicated by Chebeyshev's Inequality.

Example 2. *Let us revisit the synthia experiments. The biologist has collected n observations, $x_1, \ldots, x_n$, each corresponding to the fraction of cells that survived in a given experiment. Her estimator of the survival rate is $\frac{1}{n}\sum_{i=1}^n x_i$. How confident can she be that this is an accurate estimator of the true survival rate? Let us model her observations as realizations of n iid random variables $X_1, \ldots, X_n$ with mean p and define $\widehat{p} = \frac{1}{n}\sum_{i=1}^n X_i$. We say that her estimator is probability approximately correct with non-negative parameters (ϵ, δ) if*

$$\mathbb{P}(|\widehat{p} - p| > \epsilon) \;\leq \delta$$

The random variables are bounded between 0 and 1 and so the value of c in (1) above is equal to 1. For desired accuracy $\epsilon > 0$ and confidence $1 - \delta$, how many experiments will be sufficient? From (1) we equate $\delta = 2\,\exp(-2n\epsilon^2)$ which yields $n \geq \frac{1}{2\epsilon^2}\log(2/\delta)$. Note that this requires no knowledge of the distribution of the $\{X_i\}$ apart from the fact that they are bounded. The result can be summarized as follows. If $n \geq \frac{1}{2\epsilon^2}\log(2/\delta)$, then the probability that her estimate is off the mark by more than ϵ is less than δ.

2 Proof of Hoeffding's Inequality

Let X be any random variable and $s > 0$. Note that $\mathbb{P}(X \geq t) = \mathbb{P}(e^{sX} \geq e^{st}) \leq e^{-st}\mathbb{E}[e^{sX}]$, by using Markov's inequality, and noting that e^{sx} is a non-negative monotone increasing function. For clever choices of s this can be quite a good bound.

Let's look now at $\sum_{i=1}^{n} X_i - \mathbb{E}[X_i]$. Then

$$\begin{aligned}
\mathbb{P}(\sum_{i=1}^{n} X_i - \mathbb{E}[X_i] \geq t) &\leq e^{-st}\mathbb{E}\left[e^{s\left(\sum_{i=1}^{n} X_i - E[X_i]\right)}\right] \\
&= e^{-st}\mathbb{E}\left[\prod_{i=1}^{n} e^{s(X_i - \mathbb{E}[X_i])}\right] \\
&= e^{-st}\prod_{i=1}^{n}\mathbb{E}\left[e^{s(X_i - \mathbb{E}[X_i])}\right] ,
\end{aligned}$$

where the last step follows from the independence of the X_i's. To complete the proof we need to find a good bound for $\mathbb{E}\left[e^{s(X_i - E[X_i])}\right]$.

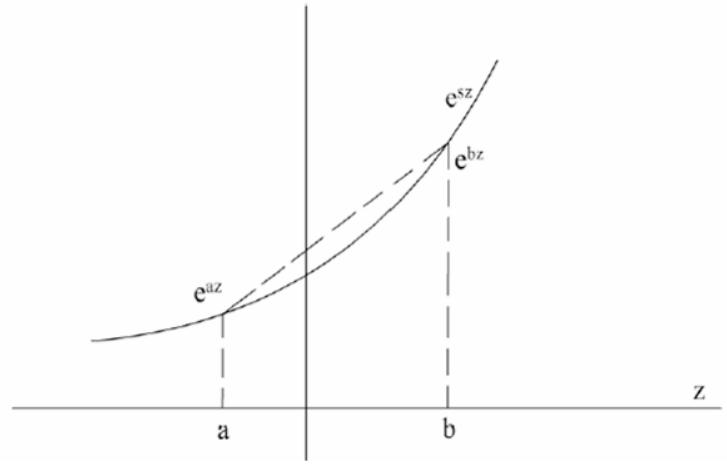

Figure 1: Convexity of exponential function.

Lemma 1. *Let Z be a r.v. such that $\mathbb{E}[Z] = 0$ and $a \leq Z \leq b$ with probability one. Then*

$$\mathbb{E}\left[e^{sZ}\right] \leq e^{\frac{s^2(b-a)^2}{8}} .$$

This upper bound is derived as follows. By the convexity of the exponential function (see Fig. 1),

$$e^{sz} \leq \frac{z-a}{b-a}e^{sb} + \frac{b-z}{b-a}e^{sa}, \text{ for } a \leq z \leq b .$$

Thus,

$$\begin{aligned}
\mathbb{E}[e^{sZ}] &\leq \mathbb{E}\left[\frac{Z-a}{b-a}\right]e^{sb} + \mathbb{E}\left[\frac{b-Z}{b-a}\right]e^{sa} \\
&= \frac{b}{b-a}e^{sa} - \frac{a}{b-a}e^{sb} , \text{ since } \mathbb{E}[Z] = 0 \\
&= (1 - \lambda + \lambda e^{s(b-a)})e^{-\lambda s(b-a)} , \text{ where } \lambda = \frac{-a}{b-a}
\end{aligned}$$

Now let $u = s(b-a)$ and define

$$\phi(u) \equiv -\lambda u + \log(1 - \lambda + \lambda e^u) \ ,$$

so that

$$\mathbb{E}[e^{sZ}] \leq (1 - \lambda + \lambda e^{s(b-a)})e^{-\lambda s(b-a)} = e^{\phi(u)} \ .$$

We want to find a good upper-bound on $e^{\phi(u)}$. Let's express $\phi(u)$ as its Taylor series with remainder:

$$\phi(u) = \phi(0) + u\phi'(0) + \frac{u^2}{2}\phi''(v) \text{ for some } v \in [0, u] \ .$$

$$\begin{aligned}
\phi'(u) &= -\lambda + \frac{\lambda e^u}{1 - \lambda + \lambda e^u} \Rightarrow \phi'(0) = 0 \\
\phi''(u) &= \frac{\lambda e^u}{1 - \lambda + \lambda e^u} - \frac{\lambda^2 e^{2u}}{(1 - \lambda + \lambda e^u)^2} \\
&= \frac{\lambda e^u}{1 - \lambda + \lambda e^u}(1 - \frac{\lambda e^u}{1 - \lambda + \lambda e^u}) \\
&= \rho(1 - \rho) \ ,
\end{aligned}$$

where $\rho = \frac{\lambda e^u}{1-\lambda+\lambda e^u}$. Now note that $\rho(1-\rho) \leq 1/4$, for any value of ρ (the maximum is attained when $\rho = 1/2$, therefore $\phi''(u) \leq 1/4$. So finally we have $\phi(u) \leq \frac{u^2}{8} = \frac{s^2(b-a)^2}{8}$, and therefore

$$\mathbb{E}[e^{sZ}] \leq e^{\frac{s^2(b-a)^2}{8}} \ .$$

Now, we can apply this upper bound to derive Hoeffding's inequality.

$$\begin{aligned}
\mathbb{P}(S_n - \mathbb{E}[S_n] \geq t) &\leq e^{-st} \prod_{i=1}^{n} \mathbb{E}[e^{s(X_i - \mathbb{E}[X_i])}] \\
&\leq e^{-st} \prod_{i=1}^{n} e^{\frac{s^2(b_i-a_i)^2}{8}} \\
&= e^{-st} e^{s^2 \sum_{i=1}^{n} \frac{(b_i-a_i)^2}{8}} \\
&= e^{\frac{-2t^2}{\sum_{i=1}^{n}(b_i-a_i)^2}} \\
&\quad \text{by choosing } \ s = \frac{4t}{\sum_{i=1}^{n}(b_i - a_i)^2}
\end{aligned}$$

The same result applies to the r.v.'s $-X_1, \ldots, -X_n$, and combining these two results yields the claim of the theorem.