[Lecture 3]

0.1 Concentration Inequalities

So far we have only considered the infinite-data limit where we directly observe $\tilde{p}$; but in general we would like to analyze what happens in finite samples where we only observe $X_1, \ldots, X_n$ sampled independently from $\tilde{p}$. In order to do this, we will want to be able to formalize statements such as "if we take the average of a large number of samples, it converges to the population mean". In order to do this, we will need a set of mathematical tools called *concentration inequalities*. A proper treatment of concentration could itself occupy an entire course, but we will cover the ideas here that are most relevant for our later analyses. See **?**, **?**, or **?** for more detailed expositions. Terence Tao also has some well-written lectures notes.

Concentration inequalities usually involve two steps:

1. We establish concentration for a single random variable, in terms of some property of that random variable.
2. We show that the property composes nicely for products of independent random variables.

A prototypical example (covered below) is showing that (1) a random variable has at most a $1/t^2$ probability of being t standard deviations from its mean; and (2) the standard deviation of a sum of n i.i.d. random variables is $\sqrt{n}$ times the standard deviation of a single variable.

The simplest concentration inequality is *Markov's inequality*. Consider the following question:

> A slot machine has an expected pay-out of \$5 (and its payout is always non-negative). What can we say about the probability that it pays out at least \$100?

We observe that the probability must be at most 0.05, since a 0.05 chance of a \$100 payout would by itself already contribute \$5 to the expected value. Moreover, this bound is achievable by taking a slot machine that pays \$0 with probability 0.95 and \$100 with probability 0.05. Markov's inequality is the generalization of this observation:

Theorem 0.1 (Markov's inequality)**.** *Let X be a non-negative random variable with mean μ. Then, $\mathbb{P}[X \geq t \cdot \mu] \leq \frac{1}{t}$.*

Markov's inequality accomplishes our first goal of establishing concentration for a single random variable, but it has two issues: first, the $\frac{1}{t}$ tail bound decays too slowly in many cases (we instead would like exponentially decaying tails); second, Markov's inequality doesn't compose well and so doesn't accomplish our second goal.

We can address both issues by applying Markov's inequality to some transformed random variable. For instance, applying Markov's inequality to the random variable $Z = (X - \mu)^2$ yields the stronger *Chebyshev inequality*:

Theorem 0.2 (Chebyshev's inequality)**.** *Let X be a real-valued random variable with mean μ and variance σ^2. Then, $\mathbb{P}[|X - \mu| \geq t \cdot \sigma] \leq \frac{1}{t^2}$.*

Proof. Since $Z = (X - \mu)^2$ is non-negative, we have that $\mathbb{P}[Z \geq t^2 \cdot \sigma^2] \leq \frac{1}{t^2}$ by Markov's inequality. Taking the square-root gives $\mathbb{P}[|X - \mu| \geq t \cdot \sigma] \leq \frac{1}{t^2}$, as was to be shown. □

Chebyshev's inequality improves the $1/t$ dependence to $1/t^2$. But more importantly, it gives a bound in terms of a quantity (the variance σ^2) that composes nicely:

Lemma 0.3 (Additivity of variance)**.** *Let $X_1, \ldots, X_n$ be pairwise independent random variables, and let $\mathsf{Var}[X]$ denote the variance of X. Then,*

$$\mathsf{Var}[X_1 + \cdots + X_n] = \mathsf{Var}[X_1] + \cdots + \mathsf{Var}[X_n]. \tag{1}$$

Proof. It suffices by induction to prove this for two random variables. Without loss of generality assume that both variables have mean zero. Then we have $\mathsf{Var}[X + Y] = \mathbb{E}[(X + Y)^2] = \mathbb{E}[X^2] + \mathbb{E}[Y^2] + 2\mathbb{E}[XY] = \mathsf{Var}[X] + \mathsf{Var}[Y] + 2\mathbb{E}[X]\mathbb{E}[Y] = \mathsf{Var}[X] + \mathsf{Var}[Y]$, where the second-to-last step uses pairwise independence. □

Chebyshev's inequality together with Lemma 0.3 together allow us to show that an average of i.i.d. random variables converges to its mean at a $1/\sqrt{n}$ rate:

Corollary 0.4. *Suppose $X_1, \dots, X_n$ are drawn i.i.d. from p, where p has mean μ and variance σ^2. Also let $S = \frac{1}{n}(X_1 + \cdots + X_n)$. Then, $\mathbb{P}[|S - \mu|/\sigma \geq t/\sqrt{n}] \leq 1/t^2$.*

Proof. Lemma 0.3 implies that $\mathsf{Var}[S] = \sigma^2/n$, from which the result follows by Chebyshev's inequality. □

Higher moments. Chebyshev's inequality gives bounds in terms of the second moment of $X - \mu$. Can we do better by considering higher moments such as the 4th moment? Supposing that $\mathbb{E}[(X-\mu)^4] \leq \tau^4$, we do get the analogous bound $\mathbb{P}[|X - \mu| \geq t \cdot \tau] \leq 1/t^4$. However, the 4th moment doesn't compose as nicely as the variance; if X and Y are two independent mean-zero random variables, then we have

$$\mathbb{E}[(X+Y)^4] = \mathbb{E}[X^4] + \mathbb{E}[Y^4] + 6\mathbb{E}[X^2]\mathbb{E}[Y^2], \tag{2}$$

where the $\mathbb{E}[X^2]\mathbb{E}[Y^2]$ can't be easily dealt with. It is possible to bound higher moments under composition, for instance using the *Rosenthal inequality* which states that

$$\mathbb{E}[|\sum_i X_i|^p] \leq \mathcal{O}(p)^p \sum_i \mathbb{E}[|X_i|^p] + \mathcal{O}(\sqrt{p})^p (\sum_i \mathbb{E}[X_i^2])^{p/2} \tag{3}$$

for independent random variables X_i. Note that the first term on the right-hand-side typically grows as $n \cdot \mathcal{O}(p)^p$ while the second term typically grows as $\mathcal{O}(\sqrt{pn})^p$.

We will typically not take the Rosenthal approach and instead work with an alternative, nicer object called the *moment generating function*:

$$m_X(\lambda) \stackrel{\text{def}}{=} \mathbb{E}[\exp(\lambda(X - \mu))]. \tag{4}$$

For independent random variables, the moment generating function composes via the identity $m_{X_1+\cdots+X_n}(\lambda) = \prod_{i=1}^n m_{X_i}(\lambda)$. Applying Markov's inequality to the moment generating function yields the *Chernoff bound*:

Theorem 0.5 (Chernoff bound)**.** *For a random variable X with moment generating $m_X(\lambda)$, we have*

$$\mathbb{P}[X - \mu \geq t] \leq \inf_{\lambda \geq 0} m_X(\lambda) e^{-\lambda t}. \tag{5}$$

Proof. By Markov's inequality, $\mathbb{P}[X - \mu \geq t] = \mathbb{P}[\exp(\lambda(X-\mu)) \geq \exp(\lambda t)] \leq \mathbb{E}[\exp(\lambda(X-\mu))]/\exp(\lambda t)$, which is equal to $m_X(\lambda)e^{-\lambda t}$ by the definition of m_X. Taking inf over λ yields the claimed bound. □

Sub-exponential and sub-Gaussian distributions. An important special case is sub-exponential random variables; recall these are random variables satisfying $\mathbb{E}[\exp(|X - \mu|/\sigma)] \leq 2$. For these, applying the Chernoff bound with $\lambda = 1/\sigma$ yields $\mathbb{P}[X - \mu \geq t] \leq 2e^{-t/\sigma}$.

Another special case is sub-Gaussian random variables (those satisfying $\mathbb{E}[\exp((X-\mu)^2/\sigma^2)] \leq 2$). In this case, using the inequality $ab \leq a^2/4 + b^2$, we have

$$m_X(\lambda) = \mathbb{E}[\exp(\lambda(X-\mu))] \leq \mathbb{E}[\exp(\lambda^2\sigma^2/4 + (X-\mu)^2/\sigma^2)] \leq 2\exp(\lambda^2\sigma^2/4). \tag{6}$$

The factor of 2 is pesky and actually we can get the more convenient bound $m_X(\lambda) \leq \exp(3\lambda^2\sigma^2/2)$ (**?**). Plugging this into the Chernoff bound yields $\mathbb{P}[X - \mu \geq t] \leq \exp(3\lambda^2\sigma^2/2 - \lambda t)$; minimizing over λ gives the optimized bound $\mathbb{P}[X - \mu \geq t] \leq \exp(-t^2/6\sigma^2)$.

Sub-Gaussians are particularly convenient because the bound $m_X(\lambda) \leq \exp(3\lambda^2\sigma^2/2)$ composes well. Let $X_1, \dots, X_n$ be independent sub-Gaussians with constants $\sigma_1, \dots, \sigma_n$. Then we have $m_{X_1+\cdots+X_n}(\lambda) \leq \exp(3\lambda^2(\sigma_1^2 + \cdots + \sigma_n^2)/2)$. We will use this to bound the behavior of sums of bounded random variables using *Hoeffding's inequality*:[1]

[1] Most of the constants presented here are suboptimal; we have focused on giving simpler proofs at the expense of sharp constants.

Theorem 0.6 (Hoeffding's inequality)**.** *Let $X_1, \ldots, X_n$ be zero-mean random variables lying in $[-M, M]$, and let $S = \frac{1}{n}(X_1 + \cdots + X_n)$. Then, $\mathbb{P}[S \geq t] \leq \exp(-\ln(2)nt^2/6M^2) \leq \exp(-nt^2/9M^2)$.*

Proof. First, note that each X_i is sub-Gaussian with parameter $\sigma = M/\sqrt{\ln 2}$, since $\mathbb{E}[\exp(X_i^2/\sigma^2)] \leq \exp(M^2/\sigma^2) = \exp(\ln(2)) = 2$. We thus have $m_{X_i}(\lambda) \leq \exp(3\lambda^2 M^2/2\ln 2)$, and so by the multiplicativity of moment generating functions we obtain $m_S(\lambda) \leq \exp(3\lambda^2 M^2/(2n\ln 2))$. Plugging into Chernoff's bound and optimizing λ as before yields $\mathbb{P}[S \geq t] \leq \exp(-\ln(2)nt^2/6M^2)$ as claimed. □

Hoeffding's inequality shows that a sum of independent random variables converges to its mean at a $1/\sqrt{n}$ rate, with tails that decay as fast as a Gaussian as long as each of the individual variables is bounded. Compare this to the $1/t^2$ decay that we obtained earlier through Chebyshev's inequality.

Cumulants. The moment generating function is a convenient tool because it multiplies over independent random variables. However, its existence requires that X already have thin tails, since $\mathbb{E}[\exp(\lambda X)]$ must be finite. For heavy-tailed distributions a (laborious) alternative is to use *cumulants*.

The cumulant function is defined as

$$K_X(\lambda) \stackrel{\text{def}}{=} \log \mathbb{E}[\exp(\lambda X)]. \tag{7}$$

Note this is the log of the moment-generating function. Taking the log is convenient because now we have additivity: $K_{X+Y}(\lambda) = K_X(\lambda) + K_Y(\lambda)$ for independent X, Y. Cumulants are obtained by writing $K_X(\lambda)$ as a power series:

$$K_X(\lambda) = 1 + \sum_{n=1}^{\infty} \frac{\kappa_n(X)}{n!}\lambda^n. \tag{8}$$

When $\mathbb{E}[X] = 0$, the first few values of κ_n are:

$$\kappa_1(X) = 0, \tag{9}$$
$$\kappa_2(X) = \mathbb{E}[X^2], \tag{10}$$
$$\kappa_3(X) = \mathbb{E}[X^3], \tag{11}$$
$$\kappa_4(X) = \mathbb{E}[X^4] - 3\mathbb{E}[X^2]^2, \tag{12}$$
$$\kappa_5(X) = \mathbb{E}[X^5] - 10\mathbb{E}[X^3]\mathbb{E}[X^2], \tag{13}$$
$$\kappa_6(X) = \mathbb{E}[X^6] - 16\mathbb{E}[X^4]\mathbb{E}[X^2] - 10\mathbb{E}[X^3]^2 + 30\mathbb{E}[X^2]^3. \tag{14}$$

Since K is additive, each of the κ_n are as well. Thus while we ran into the issue that $\mathbb{E}[(X+Y)^4] \neq \mathbb{E}[X^4] + \mathbb{E}[Y^4]$, it is the case that $\kappa_4(X+Y) = \kappa_4(X) + \kappa_4(Y)$ as long as X and Y are independent. By going back and forth between moments and cumulants it is possible to obtain tail bounds even if only some of the moments exist. However, this can be arduous and Rosenthal's inequality is probably the better route in such cases.