Lecture 2: Concentration of Measure

Akshay Krishnamurthy
akshay@cs.umass.edu

September 7, 2017

1 Recap

Last time we introduced the PAC-learning setting, which focused on binary classification with 0/1 loss, and importantly made the realizability assumption. We defined the sample complexity function $n_{\mathcal{H}} : (0,1)^2 \to \mathbb{N}$ and proved that finite hypothesis classes are PAC learnable with sample complexity

$$n_{\mathcal{H}}(\epsilon, \delta) \leq \left\lceil \frac{\log(|\mathcal{H}|/\delta)}{\epsilon} \right\rceil.$$

In the proof, we used both finite-ness of $\mathcal{H}$ and the realizability assumptions in a crucial way. But we already saw one problem where the theorem does not apply, namely learning linear separators, where $|\mathcal{H}| = \infty$. Of course it is more common than not that realizability is not satisfied, so we'd also like some sort of guarantee in the absence of realizability. In the next few lectures we will build up the tools to relax both the realizability and the finite-ness assumptions. This requires more powerful probabilistic tools, in the form of *concentration inequalities*, and also takes us into the field of *empirical process theory*.

2 Probability Background

On our way to building the probabilistic tools, we first review some basic concepts. This is by no means exhaustive.

Random Variables. A probability space Ω is just a set of possible outcomes, and a *random variable* $X : \Omega \to \mathbb{R}$ is a mapping from outcomes to the reals. If P is a distribution over outcomes Ω, we write $P(X \in A) = P(\{\omega \in \Omega : X(\omega) \in A\})$, and we often $X \sim P$ to denote that P is the distribution on Ω. If X takes discrete values, it has a *probability mass function* defined as $p(x) = P(X = x)$, and if it is continuous it has a *probability density function* p that satisfies $P(X \in A) = \int_A p(x)dx$. A pair of random variables X, Y, have a joint distribution $P(X \in A, Y \in B)$ and a joint probability mass/density function $p(x, y)$. The marginal density is $p(x) = \int p(x, y)dy$ and the conditional density is $p(y|x) = p(x, y)/p(x)$.

For our purposes we will often be more informal with the definitions of random variables.

Expectations. If X is a random variable and $g : \mathbb{R} \to \mathbb{R}$ is some function, the expected value is $\mathbb{E}[g(X)] = \int g(x)dP(x) = \int g(x)p(x)dx$ where the integral is replaced by a sum in the discrete case. For a pair of random variables, the conditional expectation $\mathbb{E}[Y|X]$ is a random variable, whose value when $X = x$ is $\mathbb{E}[Y|X = x] = \int yp(y|x)dy$. Expectations satisfy many useful properties:

1. *Linearity.* $\mathbb{E}[\sum_j c_j g_j(X)] = \sum_j c_j \mathbb{E} g_j(X)$.
2. *Independence.* If X, Y are independent random variables, meaning $P(X \in A, Y \in B) = P(X \in A)P(Y \in B)$ for all A, B, then $\mathbb{E}[XY] = \mathbb{E}[X]\mathbb{E}[Y]$. This of course generalizes to more random variables.
3. *Iterated Expectation.*

$$\mathbb{E}[Y] = \mathbb{E}[\mathbb{E}[Y|X]] = \int \mathbb{E}[Y|X = x]p(x)dx.$$

Variance. The variance of a random variable is simply

$$\mathrm{Var}[X] = \mathbb{E}[(X - \mathbb{E}[X])^2] = \mathbb{E}[X^2] - \mathbb{E}[X]^2.$$

Variance also inherits some useful properties from the expectation

1. *Independence.* If $X_1, \ldots, X_n$ are independent, then $\mathrm{Var}[\sum a_i X_i] = \sum_i a_i^2 \mathrm{Var}[X_i]$
2. *Total Variation.* $\mathrm{Var}[Y] = \mathrm{Var}[\mathbb{E}[Y|X]] + \mathbb{E}[\mathrm{Var}[Y|X]]$.

Union Bound. A useful property is that if A and B are events defined on a probabiity space Ω.

$$\mathbb{P}[A \cup B] \leq \mathbb{P}[A] + \mathbb{P}[B]$$

This can be proved by drawing a picture where A and B are regions in the probability space. The right hand side double counts the intersection, which must have positive probability.

Moment Generating Function. The *moment generating function* of a random variable X is $M_X(t) = \mathbb{E}[e^{tX}]$. As an exercise, show that

$$\left.\frac{\partial^k M_X(t)}{\partial t^k}\right|_{t=0} = \mathbb{E}[X^k],$$

which reveals where the name comes from.

3 Concentration of Measure

Question: Let $\{X_1, X_2, \ldots, \}$ be a sequence of iid random variables with $\mathbb{E}[X_i] = \mu$ and $\mathrm{Var}[X_i] = \sigma^2 < \infty$. Let $\bar{X}_n = \frac{1}{n}\sum_{i=1}^n X_i$. How close is $\bar{X}_n$ to μ?

Central Limit Theorem. An asymptotic answer is given by the Central Limit Theorem, which in some sense is the best we can hope for.

Theorem 1 (Central Limit Theorem (Lindberg-Levy))**.**

$$\sqrt{n}(\bar{X}_n - \mu) \xrightarrow{d} \mathcal{N}(0, \sigma^2)$$

The CLT shows that the sample mean concentrates around the true mean at an $O(1/\sqrt{n})$ rate. However, it is an asymptotic statement, and for proving sample complexity bounds, we really need non-asymptotic results. Thus, we would like a non-asymptotic version of the CLT, which is where concentration inequalities come in.

Gaussian Tail Inequality. Since according to the CLT sample averages look like Gaussians, it is also worth thinking about what we can show for Gaussians in finite sample.

Theorem 2 (Gaussian Tail inequality)**.** *If* $X_1, \ldots, X_n \overset{iid}{\sim} \mathcal{N}(0,1)$ *then*

$$\mathbb{P}[|\bar{X}_n| > \epsilon] \leq \frac{2}{\sqrt{n}\epsilon} \exp(-n\epsilon^2/2)$$

Re-arranging, for any $\delta \in (0, 2/e)$ *with probability at least* $1 - \delta$

$$|\bar{X}_n| \leq \sqrt{\frac{2\log(2/\delta)}{n}}$$

Observe that this re-arrangement reveals two things. First the sample mean converges to the true mean at a $1/\sqrt{n}$-rate, which is just like in the central limit theorem. Moreover the dependence on the failure probability is only logarithmic, so we can make δ extremely small with minimal consequence. Due to this logarithmic dependence, this is usually called a *exponential concentration inequality.*

Proof. Consider just a single Gaussian X, which has density $\phi(x) = (2\pi)^{-1/2}e^{-x^2/2}$. Observe that the derivative is $\phi'(x) = (2\pi)^{-1/2}e^{-x^2/2}(-x) = -x\phi(x)$. Now

$$\mathbb{P}[X > \epsilon] = \int_\epsilon^\infty \phi(s)ds \leq \frac{1}{\epsilon}\int_\epsilon^\infty s\phi(s)ds = \frac{-1}{\epsilon}\int_\epsilon^\infty \phi'(s)ds = \frac{\phi(\epsilon)}{\epsilon} \leq \frac{e^{-\epsilon^2/2}}{\epsilon}.$$

For the sample mean, use the fact that $\sqrt{n}\bar{X}_n \sim \mathcal{N}(0,1)$ to simply rescale the inequality. Finally to re-arrange, the choice of δ ensures that $\epsilon \geq 1/\sqrt{n}$, so we can eliminate the term outside of the exponential. □

Of course, we aren't always lucky enough to have Gaussian random variables and the above proof seems quite sensitive to Gaussianity, so we need something more general. For example, with the 0/1 loss, we have random variables that take values in $\{0,1\}$, or more generally they are bounded. Can we prove a similar inequality for bounded random variables? This will require several steps, which is often known as the Chernoff method.

Step 1. The first step, which is also useful elsewhere is Markov's inequality.

Proposition 3 (Markov's inequality)**.** *Let X be a non-negative random variable, then*

$$\mathbb{P}[X > \epsilon] \leq \frac{\mathbb{E}[X]}{\epsilon}$$

Proof. The proof is actually essentially the same as the first part of the Gaussian tail bound proof

$$\mathbb{P}[X > \epsilon] = \int_\epsilon^\infty p(x)dx = \int_\epsilon^\infty \frac{x}{x}p(x)dx \leq \frac{1}{\epsilon}\int_\epsilon^\infty xp(x)dx \leq \frac{1}{\epsilon}\int_0^\infty xp(x)dx = \frac{\mathbb{E}[X]}{\epsilon}$$ □

Markov's inequality requires non-negativity, but an easy consequence is to work with the variance.

Corollary 4 (Chebyshev's inequality)**.** *Let X be a random variable with mean μ and variance σ^2, then*

$$\mathbb{P}[|X - \mu| \geq \epsilon] \leq \frac{\sigma^2}{\epsilon^2}.$$

The re-arrangement here is that with probability $1 - \delta$

$$|X - \mu| \leq \sqrt{\frac{\sigma^2}{\delta}}$$

As a consequence, we recovered the $1/\sqrt{n}$ rate for the sample mean, but we did not get the logarithmic dependence on δ. The logarithmic dependence is really important, because in our applications, we will often take a union bound over exponentially many events (in fact we already did this in the PAC-learning theorem). If δ appears polynomially, this union bound will critically damage our sample complexity bound.

On the other hand, we only made very weak assumptions on the random variable X, namely that it has finite variance. In general we have much more information to leverage.

Step 2. Rather than apply Markov's inequality to the variance, it is better to apply to the MGF.

Proposition 5 (Chernoff method)**.**

$$\mathbb{P}[X \geq \epsilon] \leq \inf_{t>0} \exp(-t\epsilon)\mathbb{E}[e^{tX}].$$

So now we need to bound the moment generating function. Before that, to see why this was better than Chebyshev's inequality, for $X \sim \mathcal{N}(0,1)$, we have $M_X(t) = \exp(t^2/2)$ and so we get

$$\mathbb{P}[X \geq \epsilon] \leq \inf_{t>0} \exp(-t\epsilon + t^2/2)$$

This is minimized when $t = \epsilon$ and gives a tail bound of $\exp(-\epsilon^2/2)$, *exactly* what we saw in Theorem 2, except without the restriction on δ, so it's even better!

Lemma 6 (Hoeffding's Lemma (worse constant))**.** *Let X be a random variable with mean 0 and $a \le X \le b$ almost surely, then*

$$\mathbb{E}e^{tX} \le \exp(t^2(b-a)^2/2).$$

Proof. Let us first study a rademacher random variable $\sigma \sim \text{Unif}(\{-1,1\})$.

$$\mathbb{E}e^{t\sigma} = \frac{1}{2}[e^t + e^{-t}] = \frac{1}{2}\left[\sum_{k=0}^{\infty}\frac{(-t)^k}{k!} + \frac{t^k}{k!}\right] = \sum_{k=0}^{\infty}\frac{t^{2k}}{(2k)!} \le \sum_{k=0}^{\infty}\frac{t^{2k}}{2^k k!} = \exp(t^2/2).$$

Thus we have proved something stronger for Rademacher random variables (by a factor of 4). For general random variable X we use *symmetrization*. Let X' denote an independent copy of X, which is mean 0.

$$\begin{aligned}\mathbb{E}e^{tX} &= \mathbb{E}\exp(t(X - \mathbb{E}X')) \le \mathbb{E}_{X,X'}\exp(t(X-X')) = \mathbb{E}_{X,X',\sigma}\exp(t\sigma(X-X')) \\ &\le \mathbb{E}_{X,X'}\exp(t^2(X-X')^2/2) \le \exp(t^2(b-a)^2/2).\end{aligned}$$

The first inequality is Jensen's inequality, due to the convexity of $\exp(\cdot)$. Then we use that $X - X'$ and $X' - X$ have the same distribution, so introducing the Rademacher random variable σ has no effect. Finally, thinking of X, X' as fixed, we use the MGF bound for Rademachers and then finally plug in the worst case values for X, X'. □

The original version of the lemma has sharper constants, but the above proof is much cleaner and uses the symmetrization idea that we'll see repeatedly. However we'll use the sharper statement going forward.

Lemma 7 (Hoeffding's Lemma)**.** *Let X be a random variable with mean 0 and $a \le X \le b$ almost surely, then*

$$\mathbb{E}e^{tX} \le \exp(t^2(b-a)^2/8).$$

Proof. Note that e^{tx} is convex in x. We will apply Jensen's inequality which states that for a convex function ϕ and a random variable X, $\phi(\mathbb{E}X) \le \mathbb{E}\phi(X)$. Consider the random variable that takes value a with probability $\frac{b-x}{b-a}$ and takes value b with probability $\frac{x-a}{b-a}$. Clearly this random variable has expectation x and hence,

$$\exp(tx) \le \frac{b-x}{b-a}e^{ta} + \frac{x-a}{b-a}e^{tb}.$$

Taking expectation on both sides yields (since $\mathbb{E}X = 0$)

$$\mathbb{E}e^{tX} \le \mathbb{E}\frac{b-X}{b-a}e^{ta} + \frac{X-a}{b-a}e^{tb} = \frac{be^{ta} - ae^{tb}}{b-a}.$$

The last step is to upper bound this function using a Taylor expansion. Recall that Taylor's theorem states that for a function f and a value t, there exists a $\xi \in [0,t]$ such that

$$f(t) = f(0) + f'(0)t + f''(\xi^2)t^2/2.$$

We will apply this to the function $f(t) = \log\left(\frac{be^{ta} - ae^{tb}}{b-a}\right)$. This function has $f(0) = 0$, $f'(0) = 0$, and

$$f''(\xi) = \frac{-ab(a-b)^2e^{\xi(a+b)}}{(ae^{\xi b} - be^{\xi a})^2} \le \frac{-ab(a-b)^2e^{\xi(a+b)}}{-4abe^{\xi(a+b)}} = (a-b)^2/4.$$

Plugging into Taylor's theorem proves the result. □

To summarize, we now have the main finite sample concentration inequality for bounded random variables.

Theorem 8 (Hoeffding's inequality)**.** *Let $X_1 \ldots, X_n$ be iid random variables with mean μ and such that $a_i \le X_i \le b_i$ for all i. Then, with $\bar{X}_n = \frac{1}{n}\sum_{i=1}^n X_i$*

$$\mathbb{P}[\bar{X}_n - \mu \ge \epsilon] \le \exp(-2n^2\epsilon^2/\sum_{i=1}^n (b_i - a_i)^2).$$

Proof. Using the Chernoff method and Hoeffding's lemma, we immediately get

$$\mathbb{P}[\bar{X}_n - \mu \geq \epsilon] \leq \inf_{t>0} \exp(-t\epsilon)\mathbb{E}\exp(\bar{X}_n - \mu) = \inf_{t>0} \exp(-t\epsilon)\prod_{i=1}^{n} \mathbb{E}\exp((X_i - \mu)/n)$$
$$\leq \inf_{t>0} \exp(-t\epsilon)\exp\left(\frac{t^2\sum_{i=1}^{n}(b_i - a_i)^2}{8n^2}\right)$$

This one is optimized when $t = \frac{4n^2\epsilon}{\sum_{i=1}^{n}(b_i - a_i)^2}$ and it becomes

$$\exp(-2n^2\epsilon^2/\sum_{i=1}^{n}(b_i - a_i)^2). \qquad \square$$

This is just what we want. If each random variable has bounded range, then we get a $\exp(-n\epsilon^2)$ tail bound, just like in the Gaussian case.

This might seem like a tedious exercise, but I also want to illustrate the techniques used to prove concentration inequalities. Concentration inequalities play a central role in machine learning theory, and you may not be able to get away with using standard results, so it is important to understand the tools used to derive them.

In the homework you'll see an even more powerful inequality, but let me wrap up with a generalization of Hoeffding's inequality, which we will use in a few lectures. You might also see this referred to as Azuma's inequality.

Theorem 9 (McDiarmid's inequality)**.** *Let $X_1, \ldots, X_n$ be independent random variables taking values in a space $\mathcal{X}$ and let $f : \mathcal{X}^n \to \mathbb{R}$ be some function. If f satisfies the* bounded differences property*, for all i and $x_1, \ldots, x_n, x_i'$*

$$|f(x_1, \ldots, x_{i-1}, x_i, x_{i+1}, \ldots, x_n) - f(x_1, \ldots, x_{i-1}, x_i', x_{i+1}, \ldots, x_n)| \leq c_i$$

then

$$\mathbb{P}[f(X_1, \ldots, X_n) - \mathbb{E}f(X_1, \ldots, X_n) > \epsilon] \leq \exp\left(\frac{-2t^2}{\sum_{i=1}^{n} c_i^2}\right).$$