Machine Learning Theory
Lecture 4

Nicholas Harvey

April 28, 2020

1 Basic Probability

One of the first concentration bounds that you learn in probability theory is Markov's inequality. It bounds the right-tail of a random variable, using very few assumptions.

Theorem 1.1 (Markov's Inequality). Let Y be a real-valued random variable that assumes only nonnegative values. Then, for all $a > 0$,

$$\Pr[\,Y \geq a\,] \;\leq\; \frac{\mathrm{E}[\,Y\,]}{a}.$$

References. Wikipedia, [1, Equation B.3], Grimmett-Stirzaker Lemma 7.2.7, Durrett Theorem 1.6.4.

1.1 Unions and Intersections

Another useful tool is the "union bound". We typically use this to show that no "bad events" should happen.

Fact 1.2 (Union Bound). Let $\mathcal{E}_1$ and $\mathcal{E}_2$ be arbitrary events, not necessarily independent. Then $\Pr[\,\mathcal{E}_1 \cup \mathcal{E}_2\,] \leq \Pr[\,\mathcal{E}_1\,] + \Pr[\,\mathcal{E}_2\,]$.

Often we use it in the reverse direction, to show all "good events" should happen.

Fact 1.3 (Reverse Union Bound). Let $\mathcal{F}_1$ and $\mathcal{F}_2$ be arbitrary events, not necessarily independent. Suppose that $\Pr[\,\mathcal{F}_1\,] \geq 1 - p_1$ and $\Pr[\,\mathcal{F}_2\,] \geq 1 - p_2$. Then $\Pr[\,\mathcal{F}_1 \cap \mathcal{F}_2\,] \geq 1 - (p_1 + p_2)$.

Proof. Let $\mathcal{E}_i = \overline{\mathcal{F}_i}$. Then $\Pr[\,\mathcal{E}_i\,] \leq p_i$. Then

$$\begin{aligned}
\Pr[\,\mathcal{F}_1 \cap \mathcal{F}_2\,] &= 1 - \Pr\left[\,\overline{\mathcal{F}_1 \cap \mathcal{F}_2}\,\right] && \text{(complementary event)} \\
&= 1 - \Pr\left[\,\overline{\mathcal{F}_1} \cup \overline{\mathcal{F}_2}\,\right] && \text{(De Morgan's law)} \\
&= 1 - \Pr[\,\mathcal{E}_1 \cup \mathcal{E}_2\,] && \text{(definition of } \mathcal{E}_i\text{)} \\
&\geq 1 - (p_1 + p_2) && \text{(union bound).}
\end{aligned}$$

□

Another useful trick concerns the union of *independent* events. If $\mathcal{F}_1$ and $\mathcal{F}_2$ are each likely to happen, and independent, then their union is *even more likely* to happen.

Fact 1.4. Let $\mathcal{F}_1$ and $\mathcal{F}_2$ be independent events. Suppose that $\Pr[\mathcal{F}_1] \geq 1 - p_1$ and $\Pr[\mathcal{F}_2] \geq 1 - p_2$. Then $\Pr[\mathcal{F}_1 \cup \mathcal{F}_2] \geq 1 - p_1 p_2$.

Proof. Observe that $\Pr\left[\overline{\mathcal{F}_i}\right] \leq p_i$. So

$$\begin{aligned}
\Pr[\mathcal{F}_1 \cup \mathcal{F}_2] &= 1 - \Pr\left[\overline{\mathcal{F}_1} \cap \overline{\mathcal{F}_2}\right] && \text{(De Morgan's law)} \\
&= 1 - \Pr\left[\overline{\mathcal{F}_1}\right] \Pr\left[\overline{\mathcal{F}_2}\right] && \text{(independence)} \\
&\geq 1 - p_1 p_2.
\end{aligned}$$

□

2 Hoeffding's Inequality

Theorem 2.1. Let $X_1, ..., X_n$ be independent random variables such that X_i always lies in the interval $[0, 1]$. Define $X = \sum_{i=1}^n X_i$. Then

$$\Pr[|X - \mathrm{E}[X]| \geq t] \;\leq\; 2\exp(-2t^2/n) \qquad \forall t \geq 0.$$

References. Wikipedia, [1, Lemma B.6].

We will prove a weaker result where exponent is decreased from 2 to 1/2.

Simplifications. First of all, we will "center" the random variables, which cleans up the inequality by eliminating the expectation. Define $\hat{X}_i = X_i - \mathrm{E}[X_i]$ and $\hat{X} = \sum_{i=1}^n \hat{X}_i$. Note that[1] $\hat{X}_i \in [-1, 1]$. Our main argument is to prove that

$$\Pr\left[\hat{X} \geq t\right] \;\leq\; \exp(-t^2/2n). \tag{2.1}$$

The same argument also applies to $-\hat{X}$, so we get that

$$\Pr\left[-\hat{X} \geq t\right] \;=\; \Pr\left[\hat{X} \leq -t\right] \;\leq\; \exp(-t^2/2n).$$

Combining them with a union bound, we get

$$\Pr[|X - \mathrm{E}[X]| \geq t] \;=\; \Pr\left[|\hat{X}| \geq t\right] \;\leq\; \Pr\left[\hat{X} \geq t\right] + \Pr\left[-\hat{X} \geq t\right] \;\leq\; 2\exp(-t^2/2n).$$

This proves the theorem (with the weaker exponent).

[1] This step is where the argument is not careful enough to obtain the optimal exponent: $\hat{X}_i$ is actually supported on an interval of length 1, although our argument only assumes that it is supported on an interval of length 2.

Proof (of (2.1)). The Hoeffding inequality crucially relies on mutual independence of the $\hat{X}_i$ random variables. How can we exploit independence in the proof? What special properties to independent random variables have? One basic property is that

$$\mathrm{E}[\,A \cdot B\,] \;=\; \mathrm{E}[\,A\,] \cdot \mathrm{E}[\,B\,] \tag{2.2}$$

for any independent random variables A and B.

Key Idea #1: The Hoeffding inequality has nothing to do with *products* of random variables, it is about *sums* of random variables. So one trick we could try is to convert sums into products using the exponential function. Fix some parameter $\lambda > 0$ whose value we will choose later. Define

$$\begin{aligned} Y_i &= \exp(\lambda \hat{X}_i) \\ Y &= \exp(\lambda \hat{X}) \;=\; \exp\Big(\lambda \sum_{i=1}^n \hat{X}_i\Big) \;=\; \prod_{i=1}^n \exp(\lambda \hat{X}_i) \;=\; \prod_{i=1}^n Y_i. \end{aligned}$$

It is easy to check that, since $\{X_1, ..., X_n\}$ is mutually independent, so is $\left\{\hat{X}_1, ..., \hat{X}_n\right\}$ and $\{Y_1, ..., Y_n\}$. Therefore, by (2.2),

$$\mathrm{E}[\,Y\,] \;=\; \prod_{i=1}^n \mathrm{E}[\,Y_i\,]. \tag{2.3}$$

So far this all seems quite good. We want to prove that $\hat{X}$ is small, which is equivalent to proving Y is small. Using (2.3), we can do this by showing that the $\mathrm{E}[\,Y_i\,]$ terms are small. Doing so involves an extremely useful tool.

Key Idea #2: The second main idea is a clever trick to bound terms of the form $\mathrm{E}[\,\exp(\lambda A)\,]$, where A is a mean-zero random variable. We discuss this idea in more detail in the next subsection. We will use[2] Claim 2.2 to show

$$\mathrm{E}[\,Y_i\,] \;=\; \mathrm{E}\Big[\exp(\lambda \hat{X}_i)\Big] \;\le\; \exp(\lambda^2/2). \tag{2.4}$$

Thus, combining this with (2.3),

$$\mathrm{E}[\,Y\,] \;\le\; \prod_{i=1}^n \exp(\lambda^2/2) \;=\; \exp(\lambda^2 n/2). \tag{2.5}$$

Now we are ready to prove Hoeffding's inequality:

$$\begin{aligned} \Pr\Big[\hat{X} \ge t\Big] &= \Pr\Big[\exp(\lambda \hat{X}) \ge \exp\big(\lambda t\big)\Big] \quad \text{(by monotonicity of } e^x) \\ &\le \frac{\mathrm{E}\Big[\exp(\lambda \hat{X})\Big]}{\exp(\lambda t)} \quad \text{(by Markov's inequality (Theorem 1.1))} \\ &= \mathrm{E}[\,Y\,] \cdot \exp(-\lambda t) \\ &\le \exp(\lambda^2 n/2 - \lambda t) \quad \text{(by (2.5))} \\ &= \exp(-t^2/2n), \end{aligned}$$

by optimizing to get $\lambda = t/n$. $\square$

[2] If we were more careful here and instead used Lemma 2.4, we could improve the constant in the exponent in (2.4) from 1/2 to 1/8. This would improve the constant in the exponent in (2.1) from 1/2 to 2.

2.1 Exponentiated Mean-Zero RVs

The second main idea of Hoeffding's inequality is the following claim.

Claim 2.2. Let A be a random variable such that $|A| \leq 1$ with probability 1 and $\mathrm{E}[\,A\,] = 0$. Then for any $\lambda > 0$, we have $\mathrm{E}[\,\exp(\lambda A)\,] \leq \exp(\lambda^2/2)$.

Intuitively, the expectation should be maximized by the random variable A that is uniform on $\{-1, +1\}$. In this case,

$$\mathrm{E}[\,\exp(\lambda A)\,] \;=\; \frac{1}{2}e^{\lambda} - \frac{1}{2}e^{-\lambda} \;\leq\; e^{\lambda^2/2}.$$

This inequality is a nice bound on the hyperbolic cosine function (Claim 2.3). The full proof of Claim 2.2 basically reduces to the case of $A \in \{-1, 1\}$ using convexity of e^x.

Proof. Define $p = (1 + A)/2$ and $q = (1 - A)/2$. Observe that $p, q \geq 0$, $p + q = 1$, and $p - q = A$. By convexity,

$$\exp(\lambda A) = \exp\big(\lambda(p-q)\big) = \exp\big(\lambda p + (-\lambda)q\big) \leq p \cdot \exp(\lambda) + q \cdot \exp(-\lambda) = \frac{e^{\lambda} + e^{-\lambda}}{2} + \frac{A}{2}(e^{\lambda} - e^{-\lambda}).$$

Thus,

$$\mathrm{E}[\,\exp(\lambda A)\,] \;\leq\; \mathrm{E}\left[\frac{e^{\lambda} + e^{-\lambda}}{2} + \frac{A}{2}(e^{\lambda} - e^{-\lambda})\right] \;=\; \frac{e^{\lambda} + e^{-\lambda}}{2},$$

since $\mathrm{E}[\,A\,] = 0$. This last quantity is bounded by the following technical claim. $\square$

Claim 2.3 (Approximation of Cosh). For any real x, we have $(e^x + e^{-x})/2 \leq \exp(x^2/2)$.

References. A more general result can be found in Alon & Spencer Lemma A.1.5.

Proof. First observe that the product of all the even numbers at most $2n$ does not exceed the product of all numbers at most $2n$. In symbols,

$$2^n(n!) \;=\; \prod_{i=1}^{n}(2i) \;\leq\; \prod_{i=1}^{2n} i \;=\; (2n)!$$

Now to bound $(e^x + e^{-x})/2$, we write it as a Taylor series and observe that the odd terms cancel.

$$\frac{e^x + e^{-x}}{2} \;=\; \sum_{n \geq 0} \frac{x^n}{n!} + \sum_{n \geq 0} \frac{(-x)^n}{n!} \;=\; \sum_{n \geq 0} \frac{x^{2n}}{(2n)!} \;\leq\; \sum_{n \geq 0} \frac{x^{2n}}{2^n(n!)} \;=\; \sum_{n \geq 0} \frac{(x^2/2)^n}{n!} \;=\; \exp(x^2/2).$$

$\square$

A common scenario is that A is mean-zero, but lies in an "asymmetric" interval $[a, b]$, where $a < 0 < b$. A slightly tighter version of these MGF bounds can be derived for this scenario.

Lemma 2.4 (Hoeffding's Lemma). Let A be a random variable such that $A \in [a, b]$ with probability 1 and $\mathrm{E}[\,A\,] = 0$. Then for any $\lambda > 0$, we have $\mathrm{E}[\,\exp(\lambda A)\,] \leq \exp\big(\lambda^2(b-a)^2/8\big)$.

References. Wikipedia, [1, Lemma B.7].

The proof uses ideas similar to the proof of Claim 2.2, except we cannot use Claim 2.3 and must instead use an ad-hoc calculus argument.

References

[1] Shai Shalev-Shwartz and Shai Ben-David. *Understanding Machine Learning: From Theory to Algorithms*. Cambridge University Press, 2014.