Details and Complements for Lecture 6

36-465/665, Spring 2021

18 February 2021

Contents

1 Notation

- Upper case = random variable, lower case = realization
- When $X_1, \ldots X_n, \ldots$ is a sequence of random variables, $X_{1:n}$ abbreviates the finite sequence $X_1, X_2, \ldots X_{n-1}, X_n$, and X indicates the entire infinite random sequence. This may suggest that X_n is just a coordinate of X; it should.

2 Bound on the variance of a bounded-range random variable

Proposition 2.1. *Suppose Z is a scalar random variable confined to some interval, so that $\mathbb{P}(a \leq Z \leq b) = 1$. Then $\mathrm{Var}[Z] \leq (b-a)^2/4$.*

Proof. First, we will maximize the variance of bounded-range random variables which are supported only at the end-points of the range. Then we will argue that shifting probability into the interior of the range can only reduce the variance, so the apparently-special case we considered first is all we really need.

So, consider random variables where $Z = b$ with probability p, and $Z = a$ with probability $1 - p$. The expectation is clearly

$$\mathbb{E}[Z] = bp + (1-p)a = a + (b-a)p \tag{1}$$

and the variance is

$$\begin{aligned}
\mathbb{E}\left[(Z - \mathbb{E}[Z])^2\right] &= p(b - [a + (b-a)p])^2 + (1-p)(a - [a + (b-a)p])^2 & (2)\\
&= p(b - a - (b-a)p)^2 + (1-p)(a - a - (b-a)p)^2 & (3)\\
&= p((b-a)(1-p))^2 + (1-p)((b-a)p)^2 & (4)\\
&= p(b-a)^2(1-p)^2 + (1-p)(b-a)^2p^2 & (5)\\
&= (b-a)^2p(1-p)[(1-p)+p] & (6)\\
&= (b-a)^2p(1-p) & (7)
\end{aligned}$$

This is clearly maximized when $p = 1/2$, yielding a maximum value for $\mathrm{Var}[Z]$ of $(b-a)^2/4$.

The above result was for distributions which divide their probability between $Z = a$ and $Z = b$; let us call these the "boundary distributions". Fix any one such distribution, with mean $\mu = a + (b-a)p$. Now consider any other, non-boundary distribution on $[a, b]$ with the same expectation μ. Clearly, this distribution must put more probability mass closer to μ than the boundary distribution we started with. But since $\mathrm{Var}[Z] = \mathbb{E}\left[(Z - \mathbb{E}[Z])^2\right]$, this distribution must have a smaller variance than the two-point distribution with the same μ. Moreover, every non-boundary distribution has the same expectation as some boundary distribution. (This is because every distribution on a finite interval $[a, b]$ (i) has a well-defined expected value, and (ii) that expected value is also in the interval $[a, b]$, which (iii) means it equals $a + (b-a)p$ for some $p \in [0, 1]$.) So every non-boundary distribution has less variance than the boundary distribution with the same expectation. Hence the distribution of maximum variance must be a boundary distribution. But we have just found the distribution which maximizes variance within the boundary distributions. □

3 Hoeffding's bound on the moment generating function

Proposition 3.1. *Suppose Z is confined to the interval $[a, b]$, with expected value μ. The $M_Z(t) \leq e^{t\mu}e^{t^2(b-a)^2/8}$, i.e., Z is sub-Gaussian with scale of at most $(b-a)^2/4$.*

Proof. The factor of $e^{t\mu}$ in the result just arises from "pulling out" the expectation of Z, so the real assertion is that $M_{Z-\mathbb{E}[Z]}(t) \leq e^{t^2(b-a)^2/8}$. One might think that this would require calculations to bound the third, fourth, etc., central moments, rather in the fashion that our previous result bounded the variance. However, a somewhat indirect approach, that goes back to Hoeffding, avoids this at the cost of look a bit strange at first.

First, let's define $C \equiv Z - \mathbb{E}[Z]$, and $K(t) \equiv \log M_C(t)$. The result we care about is equivalent to the statement that $K(t) \leq t^2(b-a)^2/8$, so if we can establish the latter we've won.

Hoeffding's trick for doing so is to consider a *broader* family of distributions than just that of C. Say that C has pdf[1] $p(z)$, and, for each θ, define C_θ as the random variable with pdf

$$p_\theta(z) = p(z)\exp(\theta z - K(\theta)) \tag{8}$$

Notice that $p_\theta(z) \geq 0$ and that $\int p_\theta(z)dz = \int p(z)e^{\theta z}dze^{-K(\theta)} = \frac{M_C(\theta)}{M_C(\theta)} = 1$, so $p_\theta(z)$ is a valid probability density function for each θ.

This is an example of what's called an "exponential family" distribution, with "base distribution" $p(z)$. One

[1] If you're worried about non-continuous distributions, you have a point (pardon the pun), but this can be handled by using measure-theoretic probability. Let P be the probability measure of C, and say that P_θ is the measure absolutely continuous with respect to P with density (=Radon-Nikodym derivative) $\frac{dP_\theta}{dP} = \exp(\theta z - K(\theta))$. The arguments in the text then go through unchanged.

of the basic properties of such a distribution is that we can recover its moments from the function $K(\theta)$. Thus

$$\begin{aligned}
\frac{dK}{d\theta} &= \frac{d}{d\theta}\log \mathbb{E}\left[e^{\theta C}\right] && (9)\\
&= \frac{1}{\mathbb{E}\left[e^{\theta C}\right]}\mathbb{E}\left[\frac{d}{d\theta}e^{\theta C}\right] && (10)\\
&= e^{-K(\theta)}\mathbb{E}\left[Ce^{\theta C}\right] && (11)\\
&= \mathbb{E}\left[Ce^{\theta C - K(\theta)}\right] && (12)
\end{aligned}$$

which is $\mathbb{E}\left[C_\theta\right]$

Similarly,

$$\begin{aligned}
\frac{d^2K}{d\theta^2} &= \mathrm{Var}\left[C_\theta\right] && (13)\\
&\leq \frac{(b-a)^2}{4} && (14)
\end{aligned}$$

using the result of the previous section.

Finally, let us Taylor expand $K(t)$ around 0. Taylor's theorem in this case becomes exact, not an approximation, for some point between the expansion point and t, substituted into the last derivative, thus:

$$K(t) = K(0) + t\frac{dK}{d\theta}(0) + \frac{t^2}{2}\frac{d^2K}{d\theta^2}(\theta) \quad (15)$$

where $0 \leq \theta \leq t$. But $K(0) = 0$, $\frac{dK}{d\theta}(0) = \mathbb{E}\left[C\right] = 0$, and $\frac{d^2K}{d\theta^2}(\theta) \leq (b-a)^2/4$, so

$$K(t) \leq \frac{t^2}{2}\frac{(b-a)^2}{4} = t^2\frac{(b-a)^2}{8} \quad (16)$$

□

4 Martingales and Martingale Differences

I don't know of any way of proving results like the Efron-Stein inequality or the bounded difference / McDiarmid inequality without using ideas from probability theory about types of random sequences called "martingales" and "martingale difference sequences".

A sequence of random variables Z is a **martingale**[2] when the conditional expectation of the next value is the present value:

$$\mathbb{E}\left[Z_{n+1}|Z_1, Z_2, \ldots Z_n\right] = Z_n \quad (17)$$

By extension, the sequence Z is a martingale **with respect to the sequence** X when

- Z_n is a function[3] of $X_{1:n}$, and

[2]In English, the word "martingale" original meant a set of straps attached to a horse's head to keep the animal from raising its head too high. The word came to English from French, which got it from Spanish, which got it from the Arabic word *al marta'a*, "the fastening". It then somehow became the name of a gambling scheme where you keep doubling your bet every time you lose, in the hope of coming out ahead in one throw. (Maybe with the idea that this rule was restraining how high you bet?) In the 1930s, French mathematicians trying to define the notion of probability in terms of betting needed a name for processes that represented the stake of a gambler in a fair game, and somehow seized on "martingale". Random processes whose next increment averages to zero have since turned out to be extremely useful technical tools in probability theory, even in contexts that have nothing to do with gambling.

[3]Strictly speaking, Z_n must be a "measurable" function of $X_{1:n}$. A very rough explanation of "measurable function" would go as follows. Suppose we know that the argument X to a function f is in some set A, in stmbols $X \in A$ for some set A. Then we know that $f(X) \in f(A)$, the "image" of A under the function f. Now consider what happens if we know that X is in a

- the conditional expectation of the next value given the history is the current value:

$$\mathbb{E}\left[Z_{n+1}|X_1, \ldots X_n\right] = Z_n \tag{18}$$

A martingale, in the simple sense, is a martingale with respect to itself.

Theorem 4.1. *If Z is a martingale with respect to X, then for any $m > n$,*

$$\mathbb{E}\left[Z_m|X_{1:n}\right] = Z_n \tag{19}$$

Proof. An equivalent statement is that for any $h > 0$, $\mathbb{E}\left[Z_{n+h}|X_{1:t}\right] = Z_n$. We will prove this by mathematical induction, since it's clearly true for $h = 1$ (by the definition of a martingale).

$$\begin{aligned}
\mathbb{E}\left[Z_{n+h+1}|X_{1:t}\right] &= \mathbb{E}\left[\mathbb{E}\left[Z_{n+h+1}|X_{1:n+h}\right]|X_{1:t}\right] &(20)\\
&= \mathbb{E}\left[Z_{n+h}|X_{1:t}\right] &(21)\\
&= Z_n &(22)
\end{aligned}$$

using the martingale property in the middle line, and the inductive hypothesis in the last line. □

The random sequence D_n is a **martingale difference sequence** with respect to X when - D_n is a function of $X_1, \ldots X_n$, and - the conditional expectation of the next value give the past is always zero:

$$\mathbb{E}\left[D_{n+1}|X_1, \ldots X_n\right] = 0 \tag{23}$$

Our next result justifies the name "martingale difference sequence":

Proposition 4.1. *If Z is a martingale with respect to X, and we define $D_n \equiv Z_n - Z_{n-1}$, then D is a martingale difference sequence with respect to X.*

Proof.

$$\begin{aligned}
\mathbb{E}\left[D_{n+1}|X_{1:n}\right] &= \mathbb{E}\left[Z_{n+1} - Z_n|X_{1:n}\right] &(24)\\
&= \mathbb{E}\left[Z_{n+1}|X_{1:n}\right] - \mathbb{E}\left[Z_n|X_{1:n}\right] &(25)\\
&= Z_n - Z_n = 0 &(26)
\end{aligned}$$

where in the last step we use the facts that Z_n is a function of $X_{1:n}$, and that $\mathbb{E}\left[Z_{n+1}|X_{1:n}\right] = Z_n$. □

We can also go the other way:

Proposition 4.2. *If D is a martingale difference sequence with respect to X, then $Y_n = \sum_{i=1}^{n} D_i$ defines a martingale with respect to X.*

Proof. $\mathbb{E}\left[Y_{n+1}|X_{1:n}\right] = \mathbb{E}\left[Y_n + D_{n+1}|X_{1:n}\right] = Y_n$. □

Proposition 4.3. *If D is a martingale difference sequence, then $\mathbb{E}\left[D_n\right] = 0$ for all t.*

Proposition 4.4. *By mathematical induction. Suppose $\mathbb{E}\left[D_n\right] = 0$. By the law of total expectation, $\mathbb{E}\left[D_{n+1}\right] = \mathbb{E}\left[\mathbb{E}\left[D_{n+1}|X_1, \ldots X_n\right]\right] = \mathbb{E}\left[D_n\right] = 0$.*

Proposition 4.5. *If D is a martingale difference sequence, and $s < t$, then $\mathbb{E}\left[D_n|X_1, \ldots X_s\right] = 0$.*

Proof. By mathematical induction and the law of total expectation, as before. □

sequence of nested, increasingly small sets A_n, which narrow down to a particular point x. The function f is "measurable" if the image sets $f(A_n)$ also narrow down to the single point $f(x)$. This is a somewhat weaker requirement that f being continuous. (For instance, the indicator function for any interval on the real line is measurable, but discontinuous.) The point of requiring a measurable function, rather than just any function, is to rule out very erratic, "pathological" functions which do not play nicely with probability distributions. — If you know enough measure theory, or measure-theoretic probability, to poke holes in this explanation, you also know the real definition of "measurable function" as well as I do. If this is your first exposure to such concepts and are curious to learn more, I recommend either Grimmett and Stirzaker (1992), or Pollard (2002). (The classic book by Halmos (1950) is great but a bit abstract.)

Proposition 4.6. *If D is a martingale difference sequence, and $s \neq t$, then $\mathrm{Cov}[D_n, D_m] = 0$. That is, martingale difference sequences are uncorrelated.*

Proof. Without loss of generality, say that $m < n$. Because $\mathbb{E}[D_n] = \mathbb{E}[D_m] = 0$ (by the previous proposition), we're done if we can show $\mathbb{E}[D_m D_n] = 0$. To do this, use the law of total expectation again:

$$\begin{aligned} \mathbb{E}[D_m D_n] &= \mathbb{E}[D_m \mathbb{E}[D_n | X_1, \ldots X_s]] & (27) \\ &= \mathbb{E}[D_m 0] = 0 & (28) \end{aligned}$$

□

Because uncorrelated variables are sometimes called "orthogonal", we sometimes say martingale difference sequences are "orthogonal sequences"[4].

4.1 Sub-Gaussian martingale difference sequences

Proposition 4.7. *Suppose D is a martingale difference sequence with respect to X, and each D_n is conditionally sub-Gaussian, meaning that $\mathbb{E}\left[e^{tD_n} | X_1, \ldots X_{n-1}\right] \leq \exp\left(t^2\sigma_n^2/2\right)$ for some $\sigma_n^2 < \infty$. Then the sum of the D_ns is sub-Gaussian, and the sub-Gaussian constants add:*

$$\mathbb{E}\left[\exp\left(t\sum_{i=1}^{n} D_i\right)\right] \leq \exp\left(-\frac{t^2}{2}\sum_{i=1}^{n}\sigma_i^2\right) \tag{29}$$

Proof. By mathematical induction. To get the inductive step, assume this holds for $n-1$, and then use the law of total expectation and the conditional sub-Gaussian bound:

$$\begin{aligned} \mathbb{E}\left[\exp\left(t\sum_{i=1}^{n} D_i\right)\right] &= \mathbb{E}\left[\mathbb{E}\left[\exp\left(t\sum_{i=1}^{n} D_i\right) | X_{1:n-1}\right]\right] & (30) \\ &= \mathbb{E}\left[\left(\exp\left(t\sum_{i=1}^{n-1} D_i\right)\right)\mathbb{E}\left[\exp\left(tD_n\right) | X_{1:n-1}\right]\right] & (31) \\ &\leq \mathbb{E}\left[\left(\exp\left(t\sum_{i=1}^{n-1} D_i\right)\right)\exp\left(\frac{t^2}{2\sigma_n^2}\right)\right] & (32) \\ &= \exp\left(\frac{t^2}{2}\sigma_n^2\right)\mathbb{E}\left[\exp\left(t\sum_{i=1}^{n-1} D_i\right)\right] & (33) \end{aligned}$$

The $n = 1$ case is guaranteed by the assumption, since when $n = 1$ we can't condition on earlier Xs, hence $\mathbb{E}\left[e^{tD_1}\right] \leq \exp\left(t^2\sigma_1^2/2\right)$ for some σ_1^2. □

Proposition 4.8. *If D is a martingale difference sequence with respect to X, and $\mathbb{P}(a_n \leq D_n \leq b_n) = 1$, then*

$$\mathbb{E}\left[\exp\left(t\sum_{i=1}^{n} D_i\right)\right] \leq \exp\left(\frac{t^2}{2}\sum_{i=1}^{n}\frac{(b_i - a_i)^2}{4}\right) \tag{34}$$

Proof. Combine the previous proposition with Hoeffding's bound on the moment generating function of a limited-range random variable. □

[4]Suppose the sequence O has $\mathbb{E}[O_n] = 0$ for all n, and $\mathrm{Cov}[O_n, O_m] = 0$ for all $m \neq n$. Is this **orthogonal sequence** O necessarily a martingale difference sequence? If so, can you prove it? If not, can you find a counter-example, a sequence that is orthogonal but is not a martingale difference sequence?

Both of the last two propositions are sometimes called "the Azuma-Hoeffding bound", as are a number of other closely related results about the moment generating functions of sums of martingale differences, of martingales, on the deviations of sums of martingale differences from their expectations, etc. (Hoeffding's original paper indicated that his result should apply not just to independent variables but also to martingale differences, and Azuma seems to have been the first to work through the details.)

4.2 Turning ordinary functions into martingales, and martingale differences, a.k.a. Doob's martingale

Take any function f of n arguments, say $f(X_1, \ldots X_n)$. A classic, widely-used construction relates this to a martingale. Define

$$Z_k = \mathbb{E}\left[f(X_{1:n})|X_{1:k}\right] \tag{35}$$

with the understanding that $Z_n = f(X_{1:n})$, the random function itself, and that $Z_0 = \mathbb{E}\left[f(X_{1:n})\right]$, the unconditional expectation. Let's check that this is a martingale with respect to X: - For each k, Z_k is clearly a function of $X_{1:k}$. - We thus need to check the conditional expectation property, that $\mathbb{E}\left[Z_{k+1}|X_{1:k}\right] = Z_k$. But this follows by the law of total expectation:

$$\begin{aligned}
\mathbb{E}\left[Z_{k+1}|X_{1:k}\right] &= \mathbb{E}\left[\mathbb{E}\left[f(X_1, \ldots X_n)|X_{1:k+1}\right]|X_{1:k}\right] &(36)\\
&= \mathbb{E}\left[f(X_1, \ldots X_n)|X_{1:k}\right] &(37)\\
&= Z_k &(38)
\end{aligned}$$

by the definition of Z_k.

If we define $D_k = Z_k - Z_{k-1}$, then D is a martingale difference sequence. We then have the **Doob representation**

$$f(X_1, \ldots X_n) = \mathbb{E}\left[f(X_1, \ldots X_n)\right] + \sum_{i=1}^{n} D_i \tag{39}$$

which shows how the function is equal to its expected value plus a sum of expectation-0, uncorrelated random terms.

5 The Efron-Stein inequality

Let us begin with the Doob representation, which to recall is

$$f(X_1, \ldots X_n) = \mathbb{E}\left[f(X_1, \ldots X_n)\right] + \sum_{i=1}^{n} D_i \tag{40}$$

where the martingale difference sequence D is defined by $D_k = \mathbb{E}\left[f(X_{1:n})|X_{1:k}\right] - \mathbb{E}\left[f(X_{1:n})|X_{1:k-1}\right]$. Because $\mathbb{E}\left[f(X_1, \ldots X_n)\right]$ is non-random,

$$\mathrm{Var}\left[f(X_1, \ldots X_n)\right] = \mathrm{Var}\left[\sum_{i=1}^{n} D_i\right] \tag{41}$$

Since, as we've seen, the terms in a martingale difference sequence are uncorrelated,

$$\mathrm{Var}\left[f(X_1, \ldots X_n)\right] = \sum_{i=1}^{n} \mathrm{Var}\left[D_i\right] = \sum_{i=1}^{n} \mathbb{E}\left[D_i^2\right] \tag{42}$$

where the last equality holds because $\mathbb{E}\left[D_k\right] = 0$ for all k.

Nothing in the last paragraph relies on the X variables being independent, but the next steps in the argument rely crucially on this.

6 The bounded difference / McDiarmid inequality

Recall from the slides that we say a function $f(x_1, \ldots x_n)$ has the **bounded difference property**, with bounds d_i, if changing argument x_i can change the value of the function by at most d_i:

$$\max_{x_i, y_i} |f(x_1, \ldots x_{i-1}, x_i, x_{i+1}, \ldots x_n) - f(x_1, \ldots x_{i-1}, y_i, x_{i+1}, \ldots x_n)| \leq d_i \tag{43}$$

We want to use this to show that f must be close to its expectation. Doob's construction tells us how to re-write f as a sum of martingale differences, and the Azuma-Hoeffding result tells us sums of martingale differences are close to their expectations. So we're almost there, and just need to prove the following:

Proposition 6.1. *If f has the bounded difference property with bounds d, then the martingale differences $D_k \equiv \mathbb{E}\left[f(X_{1:n})|X_{1:k}\right] - \mathbb{E}\left[f(X_{1:n})|X_{1:k-1}\right]$ is bounded to an interval of the same length d_k.*

Proof. Let's define upper and lower limits for D_k:

$$L_k \equiv \min_x \mathbb{E}\left[f(X_{1:n})|X_1, \ldots X_{k-1}, X_k = x\right] - \mathbb{E}\left[f(X_{1:n})|X_1, \ldots X_{k-1}\right] \tag{44}$$

$$U_k \equiv \max_x \mathbb{E}\left[f(X_{1:n})|X_1, \ldots X_{k-1}, X_k = x\right] - \mathbb{E}\left[f(X_{1:n})|X_1, \ldots X_{k-1}\right] \tag{45}$$

Notice that L_k and U_k are random variables, since they are functions of $X_{1:k-1}$ (hence they're written with upper-case letters). Nonetheless it should be clear that

$$L_k \leq D_k \leq U_k \tag{46}$$

so we need to show that $U_k - L_k \leq d_k$.

$$\begin{aligned}
U_k - L_k &= \max_x \mathbb{E}\left[f(X_{1:n})|X_{1:k-1}, X_k = x\right] - \min_{x'} \mathbb{E}\left[f(X_{1:n})|X_{1:k-1}, X_k = x'\right] && (47)\\
&= \max_{x,x'} \mathbb{E}\left[f(X_{1:n})|X_{1:k-1}, X_k = x\right] - \mathbb{E}\left[f(X_{1:n})|X_{1:k-1}, X_k = x'\right] && (48)\\
&= \max_{x,x'} \int f(X_1, \ldots X_{k-1}, x, x_{k+1}, \ldots x_n) p(x_{k+1:n}|X_{1:k-1}, X_k = x) dx_{k+1:n} && (49)\\
&\quad - \int f(X_1, \ldots X_{k-1}, x, x_{k+1}, \ldots x_n) p(x_{k+1:n}|X_{1:k-1}, X_k = x') dx_{k+1:n}
\end{aligned}$$

(Of course if the Xs are discrete we do a sum instead of this integral.) Here, finally, we use the fact that the X_k are independent, so that the distribution of $X_{k+1}, \ldots X_n$ is the same no matter what $X_{1:k}$ might be, and in particular whether $X_k = x$ or $X_k = x'$ makes no difference. This lets us combine the integrals:

$$\begin{aligned}
&U_k - L_k && (50)\\
&= \max_{x,x'} \int \left(f(X_1, \ldots X_{k-1}, x, x_{k+1}, \ldots x_n) - f(X_1, \ldots X_{k-1}, x, x_{k+1}, \ldots x_n)\right) p(x_{k+1:n}) dx_{k+1:n} &&\\
&\leq \int d_k p(x_{k+1:n}) dx_{k+1:n} && (51)\\
&= d_k && (52)
\end{aligned}$$

using the bounded difference property of f. So we have shown that $\mathbb{P}\left(U_k \geq D_k \geq L_k\right) = 1$ and $U_k - L_k \leq d_k$, as desired. □

Combining this with our previous result about bounded martingale differences, we get

Proposition 6.2. *If f has the bounded difference property with bounds d, then*

$$\mathbb{E}\left[\exp\left(tf(X_{1:n})\right)\right] \leq \exp\left(\frac{t^2}{2} \sum_{i=1}^{n} \frac{d_i^2}{4}\right) \tag{53}$$

and the bounded-difference or McDiarmid inequality in the slides follows by the now-familiar sub-Gaussian deviation bound.

References

Grimmett, G. R., and D. R. Stirzaker. 1992. *Probability and Random Processes*. 2nd ed. Oxford: Oxford University Press.

Halmos, Paul R. 1950. *Measure Theory*. New York: Van Nostrand.

Pollard, David. 2002. *A User's Guide to Measure Theoretic Probability*. Cambridge, England: Cambridge University Press.