Concentration inequalities

Maxim Raginsky

September 1, 2015

In the previous lecture, the following result was stated without proof. If $X_1, \ldots, X_n$ are independent Bernoulli(θ) random variables representing the outcomes of a sequence of n tosses of a coin with bias (probability of HEADS) θ, then for any $\varepsilon \in (0,1)$

$$\mathbb{P}\left(\left|\widehat{\theta}_n - \theta\right| \geq \varepsilon\right) \leq 2e^{-2n\varepsilon^2} \tag{1}$$

where

$$\widehat{\theta}_n = \frac{1}{n}\sum_{i=1}^{n} X_i$$

is the fraction of HEADS in $X^n = (X_1, \ldots, X_n)$. Since $\theta = \mathbb{E}\widehat{\theta}_n$, (1) says that the *sample* (or *empirical*) *average* of the X_i's *concentrates sharply* around the statistical average $\theta = \mathbb{E}X_1$. Bounds like these are fundamental in statistical learning theory. In the next few lectures, we will learn the techniques needed to derive such bounds for settings much more complicated than coin tossing. This is not meant to be a complete picture; more details and additional results can be found in the excellent survey by Boucheron et al. [BBL04].

1 The basic tools

We start with *Markov's inequality*: Let $X \in \mathbb{R}$ be a nonnegative random variable. Then for any $t > 0$ we have

$$\mathbb{P}(X \geq t) \leq \frac{\mathbb{E}X}{t}. \tag{2}$$

The proof is simple:

$$\begin{aligned}
\mathbb{P}(X \geq t) &= \mathbb{E}[\mathbf{1}_{\{X \geq t\}}] && (3)\\
&\leq \frac{\mathbb{E}[X\mathbf{1}_{\{X \geq t\}}]}{t} && (4)\\
&\leq \frac{\mathbb{E}X}{t}, && (5)
\end{aligned}$$

where:

- (3) uses the fact that the probability of an event can be expressed as the expectation of its indicator function:
$$\mathbb{P}(X \in A) = \int_A P_X(dx) = \int_{\mathsf{X}} \mathbf{1}_{\{x \in A\}} P_X(dx) = \mathbb{E}[\mathbf{1}_{\{X \in A\}}]$$

- (4) uses the fact that
$$X \geq t > 0 \qquad \Longrightarrow \qquad \frac{X}{t} \geq 1$$
- (5) uses the fact that
$$X \geq 0 \qquad \Longrightarrow \qquad X\mathbf{1}_{\{X \geq t\}} \leq X,$$
so consequently $\mathbb{E}[X\mathbf{1}_{\{X\geq t\}}] \leq \mathbb{E}X$.

Markov's inequality leads to our first bound on the probability that a random variable deviates from its expectation by more than a given amount: *Chebyshev's inequality*. Let X be an arbitrary real random variable. Then for any $t > 0$
$$\mathbb{P}(|X - \mathbb{E}X| \geq t) \leq \frac{\mathrm{Var}[X]}{t^2}, \tag{6}$$
where $\mathrm{Var}\, X \triangleq \mathbb{E}[|X - \mathbb{E}X|^2] = \mathbb{E}X^2 - (\mathbb{E}X)^2$ is the variance of X. To prove (6), we apply Markov's inequality (2) to the nonnegative random variable $|X - \mathbb{E}X|^2$:
$$\mathbb{P}(|X - \mathbb{E}X| \geq t) = \mathbb{P}\left(|X - \mathbb{E}X|^2 \geq t^2\right) \tag{7}$$
$$\leq \frac{\mathbb{E}|X - \mathbb{E}X|^2}{t^2}, \tag{8}$$
where the first step uses the fact that the function $\phi(x) = x^2$ is monotonically increasing on $[0, \infty)$, so that $a \geq b \geq 0$ if and only if $a^2 \geq b^2$.

Now let's apply these tools to the problem of bounding the probability that, for a coin with bias θ, the fraction of HEADS in n trials differs from θ by more than some $\varepsilon > 0$. To that end, let us represent the outcomes of the n tosses by n independent Bernoulli(θ) random variables $X_1, \ldots, X_n \in \{0, 1\}$, where $\mathbb{P}(X_i = 1) = \theta$ for all i. Let
$$\widehat{\theta}_n = \frac{1}{n}\sum_{i=1}^{n} X_i.$$
Then
$$\mathbb{E}\widehat{\theta}_n = \mathbb{E}\left[\frac{1}{n}\sum_{i=1}^{n} X_i\right] = \frac{1}{n}\sum_{i=1}^{n} \underbrace{\mathbb{E}X_i}_{=\mathbb{P}(X_i=1)} = \theta$$
and
$$\mathrm{Var}[\widehat{\theta}_n] = \mathrm{Var}\left[\frac{1}{n}\sum_{i=1}^{n} X_i\right] = \frac{1}{n^2}\sum_{i=1}^{n} \mathrm{Var}[X_i] = \frac{\theta(1-\theta)}{n},$$
where we have used the fact that the X_i's are i.i.d., so $\mathrm{Var}[X_1 + \ldots + X_n] = \sum_{i=1}^{n} \mathrm{Var}\, X_i = n\,\mathrm{Var}\, X_1$. Now we are in a position to apply Chebyshev's inequality:
$$\mathbb{P}\left(\left|\widehat{\theta}_n - \theta\right| \geq \varepsilon\right) \leq \frac{\mathrm{Var}[\widehat{\theta}_n]}{\varepsilon^2} = \frac{\theta(1-\theta)}{n\varepsilon^2}. \tag{9}$$

At the very least, (9) shows that the probability of getting a bad sample decreases with sample size. Unfortunately, it does not decrease fast enough. To see why, we can appeal to the Central Limit Theorem, which (roughly) states that

$$\mathbb{P}\left(\sqrt{\frac{n}{\theta(1-\theta)}}\left(\widehat{\theta}_n-\theta\right)\geq t\right)\xrightarrow{n\to\infty}1-\Phi(t)\leq\frac{1}{\sqrt{2\pi}}\frac{e^{-t^2/2}}{t},$$

where $\Phi(t)=(1/\sqrt{2\pi})\int_{-\infty}^{t}e^{-x^2/2}dx$ is the standard Gaussian CDF. This would suggest something like

$$\mathbb{P}\left(\widehat{\theta}_n-\theta\geq\varepsilon\right)\approx\exp\left(-\frac{n\varepsilon^2}{2\theta(1-\theta)}\right),$$

which decays with n much faster than the right-hand side of (9),

2 The Chernoff bounding trick and Hoeffding's inequality

To fix (9), we will use a very powerful technique, known as the *Chernoff bounding trick* [Che52]. Let X be real-valued random variable. Suppose we are interested in bounding the probability $\mathbb{P}(X\geq\mathbb{E}X+t)$ for some particular $t>0$. Observe that for any $s>0$ we have

$$\mathbb{P}(X\geq\mathbb{E}X+t)=\mathbb{P}\left(e^{s(X-\mathbb{E}X)}\geq e^{st}\right)\leq e^{-st}\mathbb{E}\left[e^{s(X-\mathbb{E}X)}\right],\tag{10}$$

where the first step is by monotonicity of the function $\phi(x)=e^{sx}$ and the second step is by Markov's inequality (2). The Chernoff trick is to choose an $s>0$ that would make the right-hand side of (10) suitably small. In fact, since (10) holds simultaneously for *all* $s>0$, the optimal thing to do is to take

$$\mathbb{P}(X\geq\mathbb{E}X+t)\leq\inf_{s>0}e^{-st}\mathbb{E}\left[e^{s(X-\mathbb{E}X)}\right].$$

However, often a good upper bound on the *moment-generating function* $\mathbb{E}\left[e^{s(X-\mathbb{E}X)}\right]$ is enough. One such bound was developed by Hoeffding [Hoe63] for the case when X is bounded with probability one:

Lemma 1 (Hoeffding)**.** *Let X be a random variable, such that $\mathbb{P}(a\leq X\leq b)=1$ for some $-\infty<a\leq b<\infty$. Then for all $s>0$*

$$\mathbb{E}\left[e^{s(X-\mathbb{E}X)}\right]\leq e^{s^2(b-a)^2/8}.\tag{11}$$

To prove the lemma, we first start with a useful bound on the *variance* of a bounded random variable:

Lemma 2. *If U is a random variable such that $\mathbb{P}(a\leq U\leq b)$, then*

$$\mathrm{Var}[U]\leq\frac{(b-a)^2}{4}.\tag{12}$$

Proof. We use the fact that, for any real-valued random variable U,

$$\mathrm{Var}[U]\leq\mathbb{E}[(U-c)^2],\qquad\forall c\in\mathbb{R}.\tag{13}$$

(In particular $c=\mathbb{E}U$ achieves equality in the above bound.) Now let $c=\frac{a+b}{2}$, the midpoint of the interval $[a,b]$. Then, since $a\leq U\leq b$ almost surely, we know that

$$|U-c|\leq\frac{b-a}{2}.$$

Using this c in (13), we obtain $\mathrm{Var}[U]\leq\mathbb{E}[(U-c)^2]\leq\frac{(b-a)^2}{4}$, as claimed. □

Remark 1. The bound of Lemma 2 is actually sharp: consider

$$U = \begin{cases} a, & \text{with prob. } 1/2 \\ b, & \text{with prob. } 1/2 \end{cases}.$$

Then

$$\mathrm{Var}[U] = \mathbb{E}U^2 - (\mathbb{E}U)^2 = \frac{a^2+b^2}{2} - \left(\frac{a+b}{2}\right)^2 = \frac{(b-a)^2}{4}.$$

Now we can prove Hoeffding's lemma:

Proof (of Lemma 1). Without loss of generality, we may assume that $\mathbb{E}X = 0$. Thus, we are interested in bounding $\mathbb{E}[e^{sX}]$. Let's consider instead the *logarithmic moment-generating function*

$$\psi(s) \triangleq \log \mathbb{E}[e^{sX}].$$

Then

$$\psi'(s) = \frac{\mathbb{E}[Xe^{sX}]}{\mathbb{E}[e^{sX}]}, \qquad \psi''(s) = \frac{\mathbb{E}[X^2e^{sX}]}{\mathbb{E}[e^{sX}]} - \left[\frac{\mathbb{E}[Xe^{sX}]}{\mathbb{E}[e^{sX}]}\right]^2. \tag{14}$$

(I am being a bit loose here, assuming that we can interchange the order of differentiation and expectation, but in this case everything can be confirmed rigorously. I won't bore you with the details.) Now consider another random variable U whose distribution is related to X by

$$\mathbb{E}[f(U)] = \frac{\mathbb{E}[f(X)e^{sX}]}{\mathbb{E}[e^{sX}]} \tag{15}$$

for any real-valued function $f : \mathbb{R} \to \mathbb{R}$. To convince ourselves that this is a legitimate construction, let's plug in an indicator function of any event A:

$$\mathbb{P}[U \in A] = \mathbb{E}[\mathbf{1}_{\{U\in A\}}] = \frac{\mathbb{E}[\mathbf{1}_{\{X\in A\}}e^{sX}]}{\mathbb{E}[e^{sX}]}. \tag{16}$$

It is then not hard to show that this is indeed a valid probability measure. This construction is known as the *twisting* (or *tilting*) technique or as *exponential change of measure*.

We note two things:

1. Using (16) with $A = [a,b]$, we get

$$\mathbb{P}[a \le U \le b] = \frac{\mathbb{E}[\mathbf{1}_{\{a\le X\le b\}}e^{sX}]}{\mathbb{E}[e^{sX}]} = 1, \tag{17}$$

since $a \le X \le b$. Moreover, if A is any event in the complement of $[a,b]$, then $\mathbb{P}[U \in A] = 0$, since $\mathbf{1}_{\{X\in A\}} = 0$. That is, U is bounded between a and b with probability one, just like X.

2. Using (15) first with $f(U) = U$ and then with $f(U) = U^2$, we get

$$\mathbb{E}[U] = \frac{\mathbb{E}[Xe^{sX}]}{\mathbb{E}[e^{sX}]}, \qquad \mathbb{E}[U^2] = \frac{\mathbb{E}[X^2e^{sX}]}{\mathbb{E}[e^{sX}]}. \tag{18}$$

Comparing the expressions in (18) with (14), we observe that $\psi''(s) = \mathrm{Var}[U]$. Now, since $a \le U \le B$, it follows from Lemma 2 that $\psi''(s) \le \frac{(b-a)^2}{4}$. Therefore,

$$\psi(s) = \int_0^s \int_0^t \psi''(v)\mathrm{d}v\mathrm{d}t \le \frac{s^2(b-a)^2}{8},$$

where we have used the fact that $\psi'(0) = \psi(0) = 0$. Exponentiating both sides, we are done. □

We will now use the Chernoff method and the above lemma to prove the following

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

$$\mathbb{P}(S_n - \mathbb{E}S_n \ge t) \le \exp\left(-\frac{2t^2}{\sum_{i=1}^n (b_i - a_i)^2}\right); \tag{19}$$

$$\mathbb{P}(S_n - \mathbb{E}S_n \le -t) \le \exp\left(-\frac{2t^2}{\sum_{i=1}^n (b_i - a_i)^2}\right). \tag{20}$$

Consequently,

$$\mathbb{P}(|S_n - \mathbb{E}S_n| \ge t) \le 2\exp\left(-\frac{2t^2}{\sum_{i=1}^n (b_i - a_i)^2}\right). \tag{21}$$

Proof. By replacing each X_i with $X_i - \mathbb{E}X_i$, we may as well assume that $\mathbb{E}X_i = 0$. Then $S_n = \sum_{i=1}^n X_i$. Using Chernoff's trick, we write

$$\mathbb{P}(S_n \ge t) = \mathbb{P}\left(e^{sS_n} \ge e^{st}\right) \le e^{-st}\mathbb{E}\left[e^{sS_n}\right]. \tag{22}$$

Since the X_i's are independent,

$$\mathbb{E}\left[e^{sS_n}\right] = \mathbb{E}\left[e^{s(X_1+\ldots+X_n)}\right] = \mathbb{E}\left[\prod_{i=1}^n e^{sX_i}\right] = \prod_{i=1}^n \mathbb{E}\left[e^{sX_i}\right]. \tag{23}$$

Since $X_i \in [a_i, b_i]$, we can apply Lemma 1 to write $\mathbb{E}\left[e^{sX_i}\right] \le e^{s^2(b_i-a_i)^2/8}$. Substituting this into (23) and (22), we obtain

$$\begin{aligned}\mathbb{P}(S_n \ge t) &\le e^{-st}\prod_{i=1}^n e^{s^2(b_i-a_i)^2/8} \\ &= \exp\left(-st + \frac{s^2}{8}\sum_{i=1}^n (b_i - a_i)^2\right)\end{aligned}$$

If we choose $s = \frac{4t}{\sum_{i=1}^n (b_i-a_i)^2}$, then we obtain (19). The proof of (20) is similar. □

Now we will apply Hoeffding's inequality to improve our crude concentration bound (9) for the sum of n independent Bernoulli(θ) random variables, $X_1, \ldots, X_n$. Since each $X_i \in \{0, 1\}$, we can apply Theorem 1 to get, for any $t > 0$,

$$\mathbb{P}\left(\left|\sum_{i=1}^n X_i - n\theta\right| \ge t\right) \le 2e^{-2t^2/n}.$$

Therefore,

$$\mathbb{P}\left(\left|\widehat{\theta}_n - \theta\right| \ge \varepsilon\right) = \mathbb{P}\left(\left|\sum_{i=1}^n X_i - n\theta\right| \ge n\varepsilon\right) \le 2e^{-2n\varepsilon^2},$$

which gives us the claimed bound (1).

3 From bounded variables to bounded differences: McDiarmid's inequality

Hoeffding's inequality applies to sums of independent random variables. We will now develop its generalization, due to McDiarmid [McD89], to *arbitrary* real-valued functions of independent random variables that satisfy a certain condition.

Let X be some set, and consider a function $g : \mathsf{X}^n \to \mathbb{R}$. We say that g has *bounded differences* if there exist nonnegative numbers $c_1, \ldots, c_n$, such that

$$\sup_{x \in \mathsf{X}} g(x_1, \ldots, x_{i-1}, x, x_{i+1}, \ldots, x_n) - \inf_{x \in \mathsf{X}} g(x_1, \ldots, x_{i-1}, x, x_{i+1}, \ldots, x_n) \leq c_i \tag{24}$$

for all $i = 1, \ldots, n$ and all $x_1, \ldots, x_{i-1}, x_{i+1}, \ldots, x_n \in \mathsf{X}$. In words, if we change the ith variable while keeping all the others fixed, the value of g will not change by more than c_i.

Theorem 2 (McDiarmid's inequality [McD89])**.** *Let $X^n = (X_1, \ldots, X_n) \in \mathsf{X}^n$ be an n-tuple of independent X-valued random variables. If a function $g : \mathsf{X}^n \to \mathbb{R}$ has bounded differences, as in (24), then, for all $t > 0$,*

$$\mathbb{P}\left(g(X^n) - \mathbb{E}g(X^n) \geq t\right) \leq \exp\left(-\frac{2t^2}{\sum_{i=1}^n c_i^2}\right); \tag{25}$$

$$\mathbb{P}\left(\mathbb{E}g(X^n) - g(X^n) \geq t\right) \leq \exp\left(-\frac{2t^2}{\sum_{i=1}^n c_i^2}\right). \tag{26}$$

Proof. Let me first sketch the general idea behind the proof. Let $Z = g(X^n)$ and $V = Z - \mathbb{E}Z$. The first step will be to write V as a sum $\sum_{i=1}^n V_i$, where the terms V_i are constructed so that:

1. V_i is a function only of $X^i = (X_1, \ldots, X_i)$, and $\mathbb{E}[V_i | X^{i-1}] = 0$.
2. There exists a function $\Psi_i : \mathsf{X}^{i-1} \to \mathbb{R}$ such that, conditionally on X^{i-1},
$$\Psi_i(X^{i-1}) \leq V_i \leq \Psi_i(X^{i-1}) + c_i.$$

Provided we can arrange things in this way, we can apply Lemma 1 to V_i conditionally on X^{i-1}:

$$\mathbb{E}\left[e^{sV_i} | X^{i-1}\right] \leq e^{s^2 c_i^2/8}. \tag{27}$$

Then, using Chernoff's method, we have

$$\begin{aligned}
\mathbb{P}(Z - \mathbb{E}Z \geq t) &= \mathbb{P}(V \geq t) \\
&\leq e^{-st}\mathbb{E}\left[e^{sV}\right] \\
&= e^{-st}\mathbb{E}\left[e^{s\sum_{i=1}^n V_i}\right] \\
&= e^{-st}\mathbb{E}\left[e^{s\sum_{i=1}^{n-1} V_i} e^{sV_n}\right] \\
&= e^{-st}\mathbb{E}\left[e^{s\sum_{i=1}^{n-1} V_i}\mathbb{E}\left[e^{sV_n}\middle| X^{n-1}\right]\right] \\
&\leq e^{-st}e^{s^2c_n^2/8}\mathbb{E}\left[e^{s\sum_{i=1}^{n-1} V_i}\right],
\end{aligned}$$

where in the next-to-last step we used the fact that $V_1, \ldots, V_{n-1}$ depend only on X^{n-1}, and in the last step we used (27) with $i = n$. If we continue peeling off the terms involving $V_{n-1}, V_{n-2}, \ldots, V_1$, we will get

$$\mathbb{P}(Z - \mathbb{E}Z \geq t) \leq \exp\left(-st + \frac{s^2}{8}\sum_{i=1}^n c_i^2\right).$$

Taking $s = 4t/\sum_{i=1}^n c_i^2$, we end up with (25).

It remains to construct the V_i's with the desired properties. To that end, let

$$H_i(X^i) = \mathbb{E}[Z|X^i] \qquad \text{and} \qquad V_i = H_i(X^i) - H_{i-1}(X^{i-1}).$$

Then

$$\sum_{i=1}^n V_i = \sum_{i=1}^n \Big\{\mathbb{E}[Z|X^i] - \mathbb{E}[Z|X^{i-1}]\Big\} = \mathbb{E}[Z|X^n] - \mathbb{E}Z = Z - \mathbb{E}Z = V.$$

Note that V_i depends only on X^i by construction, and that

$$\begin{aligned}
\mathbb{E}[V_i|X^{i-1}] &= \mathbb{E}\Big[H_i(X^i) - H_{i-1}(X^{i-1})\Big|X^{i-1}\Big] \\
&= \mathbb{E}\Big[H_i(X^i)\Big|X^{i-1}\Big] - H_{i-1}(X^{i-1}) \\
&= \mathbb{E}\Big[\mathbb{E}[Z|X^{i-1}, X_i]\Big|X^{i-1}\Big] - H_{i-1}(X^{i-1}) \\
&= \mathbb{E}[Z|X^{i-1}] - H_{i-1}(X^{i-1}) \\
&= H_{i-1}(X^{i-1}) - H_{i-1}(X^{i-1}) \\
&= 0,
\end{aligned}$$

where we have used the law of iterated expectation in the conditional form $\mathbb{E}[\mathbb{E}[U|V,W]|V] = \mathbb{E}[U|V]$. Moreover, let

$$\begin{aligned}
\Psi_i(X^{i-1}) &= \inf_{x\in\mathsf{X}}\Big(H_i(X^{i-1}, x) - H_{i-1}(X^{i-1})\Big) \\
\Psi_i'(X^{i-1}) &= \sup_{x'\in\mathsf{X}}\Big(H_i(X^{i-1}, x') - H_{i-1}(X^{i-1})\Big),
\end{aligned}$$

where, owing to the fact that the X_i's are independent, we have

$$H_i(X^{i-1}, x) = \mathbb{E}[Z|X^{i-1}, X_i = x] = \int g(X^{i-1}, x, x_{i+1}^n) P_{X_{i+1}^n}(\mathrm{d}x_{i+1}^n)$$

x_{i+1}^n denoting the tuple $(x_{i+1}, \ldots, x_n)$. Then

$$\begin{aligned}
\Psi_i'(X^{i-1}) - \Psi_i(X^{i-1}) &= \sup_{x'\in\mathsf{X}}\Big(H_i(X^{i-1}, x') - H_{i-1}(X^{i-1})\Big) - \inf_{x\in\mathsf{X}}\Big(H_i(X^{i-1}, x) - H_{i-1}(X^{i-1})\Big) \\
&= \sup_{x\in\mathsf{X}}\sup_{x'\in\mathsf{X}}\Big(H_i(X^{i-1}, x) - H_i(X^{i-1}, x')\Big) \\
&= \sup_{x\in\mathsf{X}}\sup_{x'\in\mathsf{X}}\Big(\mathbb{E}[Z|X^{i-1}, X_i = x] - \mathbb{E}[Z|X^{i-1}, X_i = x']\Big) \\
&= \sup_{x\in\mathsf{X}}\sup_{x'\in\mathsf{X}}\left(\int\Big[g(X^{i-1}, x, x_{i+1}^n) - g(X^{i-1}, x', x_{i+1}^n)\Big]P(\mathrm{d}x_{i+1}^n)\right) \\
&\le \int \sup_{x\in\mathsf{X}}\sup_{x'\in\mathsf{X}'}\Big|g(X^{i-1}, x, x_{i+1}^n) - g(X^{i-1}, x', x_{i+1}^n)\Big|P(\mathrm{d}x_{i+1}^n) \\
&\le c_i,
\end{aligned}$$

where the last step follows from the bounded difference property. Thus, we can write $\Psi_i'(X^{i-1}) \le \Psi_i(X^{i-1}) + c_i$, which implies that, indeed,

$$\Psi_i(X^{i-1}) \le V_i \le \Psi_i(X^{i-1}) + c_i$$

conditionally on X^{i-1}. □

4 McDiarmid's inequality in action

McDiarmid's inequality is an extremely powerful and often used tool in statistical learning theory. We will now discuss several examples of its use. To that end, we will first introduce some notation and definitions.

Let X be some (measurable) space. If Q is a probability distribution of an X-valued random variable X, then we can compute the expectation of any (measurable) function $f : \mathsf{X} \to \mathbb{R}$ w.r.t. Q. So far, we have denoted this expectation by $\mathbb{E}f(X)$ or by $\mathbb{E}_Q f(X)$. We will often find it convenient to use an alternative notation, $Q(f)$.

Let $X^n = (X_1, \ldots, X_n)$ be n independent identically distributed (i.i.d.) X-valued random variables with common distribution P. The main object of interest to us is the *empirical distribution* induced by X^n, which we will denote by P_{X^n}. The empirical distribution assigns the probability $1/n$ to each X_i, i.e.,

$$\mathsf{P}_{X^n} = \frac{1}{n} \sum_{i=1}^{n} \delta_{X_i}.$$

Here, δ_x denotes a unit mass concentrated at a point $x \in \mathsf{X}$, i.e., the probability distribution on X that assigns each event A the probability

$$\delta_x(A) = \mathbf{1}_{\{x \in A\}}, \qquad \forall \text{ measurable } A \subseteq \mathsf{X}.$$

We note the following important facts about P_{X^n}:

1. Being a function of the sample X^n, P_{X^n} is a *random variable* taking values in the space of probability distributions over X.

2. The probability of a set $A \subseteq \mathsf{X}$ under P_{X^n},

   $$\mathsf{P}_{X^n}(A) = \frac{1}{n} \sum_{i=1}^{n} \mathbf{1}_{\{X_i \in A\}},$$

   is the *empirical frequency* of the set A on the sample X^n. The expectation of $\mathsf{P}_{X^n}(A)$ is equal to $P(A)$, the P-probability of A. Indeed,

   $$\mathbb{E}\mathsf{P}_{X^n}(A) = \mathbb{E}\left[\frac{1}{n} \sum_{i=1}^{n} \mathbf{1}_{\{X_i \in A\}}\right] = \frac{1}{n} \sum_{i=1}^{n} \mathbb{E}[\mathbf{1}_{\{X_i \in A\}}] = \frac{1}{n} \sum_{i=1}^{n} \mathbb{P}(X_i \in A) = P(A).$$

   (Think back to our coin-tossing example – this is a generalization of that idea, where we approximate actual probabilities of events by their relative frequencies in a series of independent trials.)

3. Given a function $f : \mathsf{X} \to \mathbb{R}$, we can compute its expectation w.r.t. P_{X^n}:

   $$\mathsf{P}_{X^n}(f) = \frac{1}{n} \sum_{i=1}^{n} f(X_i),$$

   which is just the sample mean of f on X^n. It is also referred to as the *empirical expectation* of f on X^n. We have

   $$\mathbb{E}\mathsf{P}_{X^n}(f) = \mathbb{E}\left[\frac{1}{n} \sum_{i=1}^{n} f(X_i)\right] = \frac{1}{n} \sum_{i=1}^{n} \mathbb{E}f(X_i) = \mathbb{E}f(X) = P(f).$$

We can now proceed to our examples.

4.1 Sums of bounded random variables

In the special case when $\mathsf{X} = \mathbb{R}$, P is a probability distribution supported on a finite interval, and $g(X^n)$ is the sum

$$g(X^n) = \sum_{i=1}^{n} X_i,$$

McDiarmid's inequality simply reduces to Hoeffding's. Indeed, for any $x^n \in [a,b]^n$ and $x_i' \in [a,b]$ we have

$$g(x^{i-1}, x_i, x_{i+1}^n) - g(x^{i-1}, x_i', x_{i+1}^n) = x_i - x_i' \le b - a.$$

Interchanging the roles of x_i' and x_i, we get

$$g(x^{i-1}, x_i', x_{i+1}^n) - g(x^{i-1}, x_i, x_{i+1}^n) = x_i' - x_i \le b - a.$$

Hence, we may apply Theorem 2 with $c_i = b - a$ for all i to get

$$\mathbb{P}\left(|g(X^n) - \mathbb{E}g(X^n)| \ge t\right) \le 2\exp\left(-\frac{2t^2}{n(b-a)^2}\right).$$

4.2 Uniform deviations

Let $X_1, \ldots, X_n$ be n i.i.d. X-valued random variables with common distribution P. By the Law of Large Numbers, for any $A \subseteq \mathsf{X}$ and any $\varepsilon > 0$

$$\lim_{n\to\infty} \mathbb{P}\left(|\mathsf{P}_{X^n}(A) - P(A)| \ge \varepsilon\right) = 0.$$

In fact, we can use Hoeffding's inequality to show that

$$\mathbb{P}\left(|\mathsf{P}_{X^n}(A) - P(A)| \ge \varepsilon\right) \le 2e^{-2n\varepsilon^2}.$$

This probability bound holds for each A *separately*. However, in learning theory we are often interested in the deviation of empirical frequencies from true probabilities simultaneously over some *collection* of subsets of X. To that end, let $\mathscr{A}$ be such a collection and consider the function

$$g(X^n) \triangleq \sup_{A\in\mathscr{A}} |\mathsf{P}_{X^n}(A) - P(A)|. \tag{28}$$

Later in the course we will see that, for certain choices of $\mathscr{A}$, $\mathbb{E}g(X^n) = O(1/\sqrt{n})$. However, regardless of what $\mathscr{A}$ is, it is easy to see that, by changing only one X_i, the value of $g(X^n)$ can change at most by $1/n$. Let $x^n = (x_1, \ldots, x_n)$, choose some other $x_i' \in \mathsf{X}$, and let $x_{(i)}^n$ denote x^n with x_i replaced by x_i':

$$x^n = (x^{i-1}, x_i, x_{i+1}^n), \qquad x_{(i)}^n = (x^{i-1}, x_i', x_{i+1}^n).$$

Then

$$\begin{aligned}
g(x^n) - g(x^n_{(i)}) &= \sup_{A\in\mathscr{A}} \left|\mathsf{P}_{x^n}(A) - P(A)\right| - \sup_{A'\in\mathscr{A}} \left|\mathsf{P}_{x^n_{(i)}}(A') - P(A')\right| \\
&= \sup_{A\in\mathscr{A}} \inf_{A'\in\mathscr{A}'} \left\{\left|\mathsf{P}_{x^n}(A) - P(A)\right| - \left|\mathsf{P}_{x^n_{(i)}}(A') - P(A')\right|\right\} \\
&\le \sup_{A\in\mathscr{A}} \left\{\left|\mathsf{P}_{x^n}(A) - P(A)\right| - \left|\mathsf{P}_{x^n_{(i)}}(A) - P(A)\right|\right\} \\
&\le \sup_{A\in\mathscr{A}} \left|\mathsf{P}_{x^n}(A) - \mathsf{P}_{x^n_{(i)}}(A)\right| \\
&= \frac{1}{n} \sup_{A\in\mathscr{A}} \left|\mathbf{1}_{\{x_i\in A\}} - \mathbf{1}_{\{x'_i\in A\}}\right| \\
&\le \frac{1}{n}.
\end{aligned}$$

Interchanging the roles of x^n and $x^n_{(i)}$, we obtain

$$g(x^n_{(i)}) - g(x^n) \le \frac{1}{n}.$$

Thus,

$$\left|g(x^n) - g(x^n_{(i)})\right| \le \frac{1}{n}.$$

Note that this bound holds for all i and all choices of x^n and $x^n_{(i)}$. This means that the function g defined in (28) has bounded differences with $c_1 = \ldots = c_n = 1/n$. Consequently, we can use Theorem 2 to get

$$\mathbb{P}\left(|g(X^n) - \mathbb{E}g(X^n)| \ge \varepsilon\right) \le 2e^{-2n\varepsilon^2}.$$

This shows that the *uniform deviation* $g(X^n)$ concentrates sharply around its mean $\mathbb{E}g(X^n)$.

4.3 Uniform deviations continued

The same idea applies to arbitrary real-valued functions over X. Let $X^n = (X_1, \ldots, X_n)$ be as in the previous example. Given any function $f : \mathsf{X} \to [0,1]$, Hoeffding's inequality tells us that

$$\mathbb{P}\left(\left|\mathsf{P}_{X^n}(f) - \mathbb{E}f(X)\right| \ge \varepsilon\right) \le 2e^{-2n\varepsilon^2}.$$

However, just as in the previous example, in learning theory we are primarily interested in controlling the deviations of empirical means from true means simultaneously over whole classes of functions. To that end, let $\mathscr{F}$ be such a class consisting of functions $f : \mathsf{X} \to [0,1]$ and consider the *uniform deviation*

$$g(X^n) \triangleq \sup_{f\in\mathscr{F}} \left|\mathsf{P}_{X^n}(f) - P(f)\right|.$$

An argument entirely similar to the one in the previous example[1] shows that this g has bounded differences with $c_1 = \ldots = c_n = 1/n$. Therefore, applying McDiarmid's inequality, we obtain

$$\mathbb{P}\left(\left|g(X^n) - \mathbb{E}g(X^n)\right| \ge \varepsilon\right) \le 2e^{-2n\varepsilon^2}.$$

We will see later that, for certain function classes $\mathscr{F}$, we will have $\mathbb{E}g(X^n) = O(1/\sqrt{n})$.

[1] Exercise: verify this!

4.4 Kernel density estimation

For our final example, let $X^n = (X_1, \ldots, X_n)$ be an n-tuple of i.i.d. real-valued random variables whose common distribution P has a probability density function (pdf) f, i.e.,

$$P(A) = \int_A f(x)\mathrm{d}x$$

for any measurable set $A \subseteq \mathbb{R}$. We wish to estimate f from the sample X^n. A popular method is to use a *kernel estimate* (the book by Devroye and Lugosi [DL01] has plenty of material on density estimation, including kernel methods, from the viewpoint of statistical learning theory). To that end, we pick a non-negative function $K : \mathbb{R} \to \mathbb{R}$ that integrates to one, $\int K(x)\mathrm{d}x = 1$ (such a function is called a *kernel*), as well as a positive *bandwidth* (or *smoothing constant*) $h > 0$ and form the estimate

$$\widehat{f}_n(x) = \frac{1}{nh}\sum_{i=1}^{n} K\left(\frac{x - X_i}{h}\right).$$

It is not hard to verify[2] that $\widehat{f}_n$ is a valid pdf, i.e., that it is nonnegative and integrates to one. A common way of quantifying the performance of a density estimator is to use the L_1 distance to the true density f:

$$\|\widehat{f}_n - f\|_{L_1} = \int_{\mathbb{R}} \left|\widehat{f}_n(x) - f(x)\right| \mathrm{d}x.$$

Note that $\|\widehat{f}_n - f\|_{L_1}$ is a random variable since it depends on the random sample X^n. Thus, we can write it as a function $g(X^n)$ of the sample X^n. Leaving aside the problem of actually bounding $\mathbb{E}g(X^n)$, we can easily establish a concentration bound for it using McDiarmid's inequality. To do that, we need to check that g has bounded differences. Choosing x^n and $x^n_{(i)}$ as before, we have

$$\begin{aligned}
&g(x^n) - g(x^n_{(i)}) \\
&= \int_{\mathbb{R}} \left| \frac{1}{nh}\sum_{j=1}^{i-1} K\left(\frac{x - x_j}{h}\right) + \frac{1}{nh}K\left(\frac{x - x_i}{h}\right) + \frac{1}{nh}\sum_{j=i+1}^{n} K\left(\frac{x - x_j}{h}\right) - f(x) \right| \mathrm{d}x \\
&\quad - \int_{\mathbb{R}} \left| \frac{1}{nh}\sum_{j=1}^{i-1} K\left(\frac{x - x_j}{h}\right) + \frac{1}{nh}K\left(\frac{x - x'_i}{h}\right) + \frac{1}{nh}\sum_{j=i+1}^{n} K\left(\frac{x - x_j}{h}\right) - f(x) \right| \mathrm{d}x \\
&\leq \frac{1}{nh}\int_{\mathbb{R}} \left| K\left(\frac{x - x_i}{h}\right) - K\left(\frac{x - x'_i}{h}\right) \right| \mathrm{d}x \\
&\leq \frac{2}{nh}\int_{\mathbb{R}} K\left(\frac{x}{h}\right) \mathrm{d}x \\
&= \frac{2}{n}.
\end{aligned}$$

Thus, we see that $g(X^n)$ has the bounded differences property with $c_1 = \ldots = c_n = 2/n$, so that

$$\mathbb{P}\left(\left|g(X^n) - \mathbb{E}g(X^n)\right| \geq \varepsilon\right) \leq 2e^{-n\varepsilon^2/2}.$$

[2] Another exercise!

References

[BBL04] S. Boucheron, O. Bousquet, and G. Lugosi. Concentration inequalities. In O. Bousquet, U. von Luxburg, and G. Rätsch, editors, *Advanced Lectures in Machine Learning*, pages 208–240. Springer, 2004.

[Che52] H. Chernoff. A meausre of asymptotic efficiency of tests of a hypothesis based on the sum of observations. *Annals of Mathematical Statistics*, 23:493–507, 1952.

[DL01] L. Devroye and G. Lugosi. *Combinatorial Methods in Density Estimation*. Springer, 2001.

[Hoe63] W. Hoeffding. Probability inequalities for sums of bounded random variables. *Journal of the American Statistical Association*, 58:13–30, 1963.

[McD89] C. McDiarmid. On the method of bounded differences. In *Surveys in Combinatorics*, pages 148–188. Cambridge University Press, 1989.