Hoeffding, Azuma, McDiarmid

Karl Stratos

1 Hoeffding (sum of independent RVs)

Hoeffding's lemma. If $X \in [a, b]$ and $\mathbf{E}[X] = 0$, then for all $t > 0$:

$$\mathbf{E}[e^{tX}] \leq e^{t^2(b-a)^2/8}$$

Proof. Since e^{tx} is convex, for all $x \in [a, b]$:

$$e^{tx} \leq \frac{b-x}{b-a}e^{ta} + \frac{x-a}{b-a}e^{tb}$$

This means:

$$\mathbf{E}[e^{tX}] \leq \frac{b}{b-a}e^{ta} - \frac{a}{b-a}e^{tb} = \left(\frac{b}{b-a} - \frac{a}{b-a}e^{t(b-a)}\right)e^{ta} = e^{\phi(t)}$$

where $\phi(t) := ta + \ln\left(\frac{b}{b-a} - \frac{a}{b-a}e^{t(b-a)}\right)$. We did the second step because we want the form $(b-a)$. Look at the derivatives of ϕ:

$$\begin{aligned}
\phi'(x) &= a - \frac{a}{\frac{b}{b-1}e^{-t(b-a)} - \frac{a}{b-a}} \\
\phi''(x) &= \frac{-abe^{-t(b-a)}}{\left(\frac{b}{b-a}e^{-t(b-a)} - \frac{a}{b-a}\right)^2} \\
&= \frac{\alpha(1-\alpha)e^{-t(b-a)}(b-a)^2}{\left((1-\alpha)e^{-t(b-a)} + \alpha\right)^2} \qquad \text{for } \alpha := \tfrac{-a}{b-a} \\
&= \underbrace{\frac{\alpha}{\left((1-\alpha)e^{-t(b-a)} + \alpha\right)}}_{u}\underbrace{\frac{(1-\alpha)e^{-t(b-a)}}{\left((1-\alpha)e^{-t(b-a)} + \alpha\right)}}_{1-u}(b-a)^2 \leq \frac{(b-a)^2}{4}
\end{aligned}$$

We used the fact that the concave function $u(1-u) = u - u^2$ achieves its maximum of $1/4$ at $u = 1/2$.

Now we approximate $\phi(t)$ at $t = 0$ with the first-degree Taylor polynomial. The **Remainder theorem** gives us that

$$\begin{aligned}
\phi(t) &= \phi(0) + \frac{1}{t}\phi'(0) + R_1(\theta) \qquad \text{for some } \theta \in [0, t] \\
&= \frac{t^2}{2}\phi''(\theta) \leq \frac{t^2(b-a)^2}{8}
\end{aligned}$$

□

Hoeffding's inequality. Given iid random variables $X_1 \dots X_m$ where $X_i \in [a_i, b_i]$, let $S_m := \sum_{i=1}^m X_i$. Then for any $\epsilon > 0$:

$$P(S_m - \mathbf{E}[S_m] \geq \epsilon) \leq e^{-2\epsilon^2 / \sum_{i=1}^m (b_i - a_i)^2}$$

Proof. Using the **Chernoff bounding technique**, we write for all $t \geq 0$:

$$\begin{aligned}
P(S_m - \mathbf{E}[S_m] \geq \epsilon) &= P(e^{t(S_m - \mathbf{E}[S_m])} \geq e^{t\epsilon}) \\
&\leq \mathbf{E}[e^{t(S_m - \mathbf{E}[S_m])}] e^{-t\epsilon} && \text{by Markov} \\
&= \mathbf{E}\left[\prod_{i=1}^m e^{t(X_i - \mathbf{E}[X_i])}\right] e^{-t\epsilon} \\
&= \prod_{i=1}^m \mathbf{E}\left[e^{t(X_i - \mathbf{E}[X_i])}\right] e^{-t\epsilon} && \text{by independence} \\
&\leq \prod_{i=1}^m e^{\frac{t^2(b_i - a_i)^2}{8}} e^{-t\epsilon} && \text{by Hoeffding's lemma} \\
&= \prod_{i=1}^m e^{\frac{t^2(b_i - a_i)^2}{8} - t\epsilon}
\end{aligned}$$

Since $\frac{t^2(b_i-a_i)^2}{8} - t\epsilon$ is convex, we minimize it with $t = \frac{4\epsilon}{(b_i - a_i)^2}$, yielding the bound $\prod_{i=1}^m e^{\frac{-2\epsilon^2}{(b_i-a_i)^2}} = e^{\frac{-2\epsilon^2}{\sum_{i=1}^m (b_i-a_i)^2}}$.[1] □

The proof suggests that the result can be generalized to variables that are not necessarily independent, since we just need the expectation to break over a product.

2 Azuma (sum of martingale differences)

Conditional Hoeffding's lemma. If $V \in [f(Z), f(Z) + c]$ and $\mathbf{E}[V|Z] = 0$, then for all $t > 0$:

$$\mathbf{E}[e^{tV}|Z] \leq e^{t^2c^2/8}$$

Note that $\mathbf{E}[e^{tV}|Z]$ is a random variable in Z.

Proof. Similar to the proof of Hoeffding's lemma. Use $a = f(Z), b = f(Z) + c$ and use $\mathbf{E}[\cdot|Z]$ instead of $\mathbf{E}[\cdot]$. □

$V_1, V_2, \dots$ is called a **martingale difference sequence wrt.** $X_1, X_2, \dots$ if

- V_i is a function of $X_1 \dots X_i$.
- $\mathbf{E}[|V_i|] < \infty$
- $\mathbf{E}[V_{i+1}|X_1 \dots X_i] = 0$

[1] Without Hoeffding's lemma, we could handle the case $X_i \in \{0, 1\}$ by explicitly bounding the non-centered quantity $\mathbf{E}[e^{tX_i}] = p_i e^t + (1 - p_i) = 1 - p_i(e^t + 1) \leq \exp(-p_i(e^t + 1))$ (here $p_i := \mathbf{E}[X_i]$) and observing $\prod_{i=1}^m \mathbf{E}[e^{tX_i}] \leq \exp(-\mathbf{E}[S_m](e^t + 1))$.

Azuma's inequality. Given a martingale difference sequence $V_1, V_2, \ldots$ wrt. $X_1, X_2, \ldots$ where $V_i \in [f_i(X_1 \ldots X_{i-1}), f_i(X_1 \ldots X_{i-1}) + c_i]$ for some f_i and $c_i \geq 0$, for all $\epsilon > 0$:

$$\mathbf{E}\left[\sum_{i=1}^{m} V_i \geq \epsilon\right] \leq e^{-2\epsilon^2/\sum_{i=1}^{m} c_i^2}$$

Proof. For each $k \in [m]$, define $S_k := \sum_{i=1}^{k} V_i$. By the law of iterated expectations (LIE) $\mathbf{E}_X[X] = \mathbf{E}_Z[\mathbf{E}_{X|Z}[X|Z]]$ (see the appendix):

$$\mathbf{E}[e^{tS_k}] = \mathbf{E}[\mathbf{E}[e^{tS_k}|X_1 \ldots X_{k-1}]]$$

where

$$\begin{aligned}
\mathbf{E}[e^{tS_k}|X_1 \ldots X_{k-1}] &= \mathbf{E}[e^{tS_{k-1}} e^{tV_k}|X_1 \ldots X_{k-1}] \\
&= \mathbf{E}[e^{tS_{k-1}}|X_1 \ldots X_{k-1}]\mathbf{E}[e^{tV_k}|X_1 \ldots X_{k-1}] \\
&\leq \mathbf{E}[e^{tS_{k-1}}|X_1 \ldots X_{k-1}]e^{t^2 c_k^2/8}
\end{aligned}$$

The second step holds because S_{k-1} only depends on $X_1 \ldots X_{k-1}$. The third step holds by conditional Hoeffding's lemma. Thus

$$\mathbf{E}[e^{tS_m}] \leq e^{t^2 c_m^2/8}\mathbf{E}[e^{tS_{m-1}}] \leq \cdots \leq e^{\frac{t^2 \sum_{i=1}^{m} c_i^2}{8}}$$

Use the Chernoff bounding technique on S_m:

$$\begin{aligned}
P(S_m \geq \epsilon) &= P(e^{tS_m} \geq e^{t\epsilon}) \\
&\leq \mathbf{E}\left[e^{tS_m}\right] e^{-t\epsilon} && \text{by Markov} \\
&\leq e^{\frac{t^2 \sum_{i=1}^{m} c_i^2}{8} - t\epsilon} && \text{by the above argument}
\end{aligned}$$

By minimizing the convex function $\frac{t^2 \sum_{i=1}^{m} c_i^2}{8} - t\epsilon$ with $t = 4\epsilon/\sum_{i=1}^{m} c_i^2$, we get the bound $e^{-2\epsilon^2/\sum_{i=1}^{m} c_i^2}$. □

3 McDiarmid (**"Lipschitz" function of independent RVs**)

McDiarmid's inequality. Given iid random variables $X_1 \ldots X_m \in \mathcal{X}$, let $f : \mathcal{X}^m \to \mathbb{R}$ be function bounded in a Lipschitz-like manner as follows: for all $x_1 \ldots x_m, x_i' \in \mathcal{X}$, there is some $c_i \geq 0$ such that

$$|f(x_1 \ldots x_i \ldots x_m) - f(x_1 \ldots x_i' \ldots x_m)| \leq c_i$$

Let $f(S) := f(X_1 \ldots X_m)$. Then

$$P(f(S) - \mathbf{E}[f(S)] \geq \epsilon) \leq e^{-2\epsilon^2/\sum_{i=1}^{m} c_i^2}$$

Proof. Define $V := f(S) - \mathbf{E}[f(S)]$. Will show $V = \sum_{i=1}^{m} V_i$ is a sum of bounded margingale differences $V_i \in [f_i(X_1 \ldots X_{i-1}), f_i(X_1 \ldots X_{i-1}) + c_i]$. Then Azuma's inequality gives the desired result.

Define $V_i := \mathbf{E}[V|X_1 \ldots X_i] - \mathbf{E}[V|X_1 \ldots X_{i-1}]$. Note that each V_i is a function of $X_1 \ldots X_i$ and the telescoping sum gives

$$\sum_{i=1}^{m} V_i = \mathbf{E}[V|X_1 \ldots X_m] = V$$

In addition, $\mathbf{E}[\mathbf{E}[V|X_1 \dots X_i]|X_1 \dots X_{i-1}] = \mathbf{E}[V|X_1 \dots X_{i-1}]$ (by LIE), so we have

$$\mathbf{E}[V_i|X_1 \dots X_{i-1}] = \mathbf{E}[\mathbf{E}[V|X_1 \dots X_i] - V|X_1 \dots X_{i-1}] = 0$$

Thus $V_1 \dots V_m$ is a martingale difference sequence wrt. $X_1 \dots X_m$.[2]

Now bound V_i in terms of $X_1 \dots X_{i-1}$:

$$\begin{aligned}
V_i &\leq \sup_{x \in \mathcal{X}} \mathbf{E}[V|X_1 \dots X_i = x] - \mathbf{E}[V|X_1 \dots X_{i-1}] =: W_i \\
V_i &\geq \inf_{x \in \mathcal{X}} \mathbf{E}[V|X_1 \dots X_i = x] - \mathbf{E}[V|X_1 \dots X_{i-1}] =: U_i
\end{aligned}$$

Using the "Lipschitz" condition on f:

$$\begin{aligned}
W_i - U_i &= \sup_{x,x' \in \mathcal{X}} \mathbf{E}[V|X_1 \dots X_i = x] - \mathbf{E}[V|X_1 \dots X_i = x'] \\
&= \sup_{x,x' \in \mathcal{X}} \mathbf{E}[f(S)|X_1 \dots X_i = x] - \mathbf{E}[f(S)|X_1 \dots X_i = x'] \\
&\leq c_i
\end{aligned}$$

Thus $W_i \leq U_i + c_i$ and it follows $V_i \in [U_i, U_i + c_i]$ where U_i is a function of $X_1 \dots X_{i-1}$. $\square$

References. Appendix *D* of *Foundations of Machine Learning* (MRT), Chapter 12 of *Probability and Computing* (MU)

[2] We've constructed a doob martingale $Z_0, Z_1, \dots, Z_m$ wrt. $X_0 = \text{constant}, X_1, \dots, X_m$ for the target quantity V. That is, $Z_i := \mathbf{E}[V|X_0 \dots X_m]$ which gives $V_i = Z_i - Z_{i-1}$.

4 Appendies

4.1 Crash Course on Conditional RVs

The proof of Azuma's and McDiarmid's inequality makes heavy use of *conditional* expectations.

- Let's say X is a random variable.
- Then $\mathbf{E}_X[X]$ is a constant.
- However, $\mathbf{E}_{X|Y}[X|Y]$ *is a random variable* (random over Y)! We can only compute a value for a specific $y \in Y$:

  $$\mathbf{E}_{X|Y}[X|Y=y] = \int_x P_{X|Y}(X=x|Y=y) \times x \ \ dx$$

  is a constant.

The **law of iterated expectations** (LIE)[3] states that

$$\mathbf{E}_Y[\underbrace{\mathbf{E}_{X|Y}[X|Y]}_{\text{fnc of } Y}] = \underbrace{\mathbf{E}_X[X]}_{\text{constant}}$$

Now that we know the definition, it's pretty easy to show:

$$\begin{aligned}
\mathbf{E}_Y[\mathbf{E}_{X|Y}[X|Y]] &= \int_y P_Y(Y=y) \times \mathbf{E}_{X|Y}[X|Y=y] \ \ dy \\
&= \int_y P_Y(Y=y) \times \left(\int_x P_{X|Y}(X=x|Y=y) \times x \ \ dx \right) \ \ dy \\
&= \int_x \left(\int_y P_Y(Y=y) \times P_{X|Y}(X=x|Y=y) \ \ dy \right) \times x \ \ dx \\
&= \int_x P_X(X=x) \times x \ \ dx \\
&= \mathbf{E}_X[X]
\end{aligned}$$

The same principle holds when we work with more than two variables:

$$\mathbf{E}_{Y|Z}[\underbrace{\mathbf{E}_{X|Y,Z}[X|Y,Z]}_{\text{fnc of } Y,Z}|Z] = \underbrace{\mathbf{E}_{X|Z}[X|Z]}_{\text{fnc of } Z}$$

It basically says we're free to condition on anything as long as we eventually take expectation over it.

4.2 Martingales

A sequence $Z_0, Z_1 \ldots$ is a **martingale** wrt. $X_0, X_1 \ldots$ if

- Z_i is a function of $X_0 \ldots X_i$.
- $\mathbf{E}[|Z_i|] \leq \infty$
- $\mathbf{E}[Z_{i+1}|X_0 \ldots X_i] = Z_i$

[3] Also called the law of total expectation, the tower rule, the smoothing theorem, Adam's Law.

A **doob martingale** is a martingale constructed as follows. Let $X_0 \dots X_n$ be any sequence. We are interested in Y that depends on all $X_0 \dots X_n$; we assume $\mathbf{E}[|Y|] \leq \infty$. We define Z_i to be the expectation of Y given $X_0 \dots X_i$:

$$Z_i := \mathbf{E}[Y|X_0 \dots X_i]$$

To verify $Z_0 \dots Z_n$ is a martingale, we need to check the third condition:

$$\begin{aligned}
\mathbf{E}[Z_{i+1}|X_0 \dots X_i] &= \mathbf{E}[\mathbf{E}[Y|X_0 \dots X_{i+1}]|X_0 \dots X_i] && \text{by def} \\
&= \mathbf{E}[Y|X_0 \dots X_i] && \text{by LIE} \\
&= Z_i
\end{aligned}$$

For instance, consider a sequence of rewards in n independent fair gambles: $X_1 \dots X_n$ where $\mathbf{E}[X_i] = 0$. We are interested in the total reward $Y = \sum_{i=1}^n X_i$. Then our doob martingale is given by

$$Z_i = \sum_{j=1}^{n} \mathbf{E}[X_j|X_1 \dots X_i] = \sum_{j=1}^{i} X_j$$

since $\mathbf{E}[X_j|X_1 \dots X_i] = \mathbf{E}[X_j] = 0$ for $j > i$. I.e., the refined estimate of the total reward at time i is simply the sum up to that time.

By construction, if $Z_0, Z_1, \dots$ is a martingale wrt. $X_0, X_1, \dots$, then $V_1, V_2, \dots$ defined by

$$V_i := Z_i - Z_{i-1}$$

is a **martingle difference sequence** defined before since

- $V_i = Z_i - Z_{i-1}$ is a function of $X_1 \dots X_i$.
- $\mathbf{E}[|V_i|] = \mathbf{E}[|Z_i - Z_{i-1}|] < \infty$
- $\mathbf{E}[V_{i+1}|X_1 \dots X_i] = \mathbf{E}[Z_{i+1}|X_1 \dots X_i] - Z_i = 0$