Hoeffding, Azuma, McDiarmid

Karl Stratos

1 Hoeffding (sum of independent RVs)

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

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

**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:

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

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 $\phi$:

$$\phi'(x) = a - \frac{a}{b-a} e^{-t(b-a)} - \frac{a}{b-a}$$

$$\phi''(x) = \frac{-abe^{-t(b-a)}}{\left( \frac{b}{b-a} e^{-t(b-a)} - \frac{a}{b-a} \right)^2}$$

$$= \frac{\alpha(1-\alpha)e^{-t(b-a)}(b-a)^2}{(1-\alpha)e^{-t(b-a)} + \alpha}$$

for $\alpha := \frac{-a}{b-a}$

$$= \frac{\alpha}{(1-\alpha)e^{-t(b-a)} + \alpha} \frac{(1-\alpha)e^{-t(b-a)}(b-a)^2}{(1-\alpha)e^{-t(b-a)} + \alpha} \leq \frac{(b-a)^2}{4}$$

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

$$\phi(t) = \phi(0) + \frac{1}{t} \phi'(0) + R_1(\theta)$$

for some $\theta \in [0,t]$

$$= \frac{t^2}{2} \phi''(\theta) \leq \frac{t^2(b-a)^2}{8}$$

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

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

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

$$P(S_m - E[S_m] \geq \epsilon) = P(e^{t(S_m - E[S_m])} \geq e^{t\epsilon})$$

$$\leq E[e^{t(S_m - E[S_m])}e^{-t\epsilon}] \text{ by Markov}$$

$$= E \left( \prod_{i=1}^{m} e^{t(X_i - E[X_i])} \right) e^{-t\epsilon} \text{ by independence}$$

$$\leq \prod_{i=1}^{m} e^{\frac{t^2(b_i - a_i)^2}{8}} e^{-t\epsilon} \text{ by Hoeffding’s lemma}$$

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

$$\prod_{i=1}^{m} e^{\frac{t^2(b_i - a_i)^2}{8}} = e^{\sum_{i=1}^{m} \frac{t^2(b_i - a_i)^2}{8}}.$$ 

\[ \square \]

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 $E[V|Z] = 0$, then for all $t > 0$:

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

Note that $E[e^{tV}|Z]$ 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 $E[|Z]$ instead of $E[\cdot]$.

\[ \square \]

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

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

\[ 1 \text{Without Hoeffding’s lemma, we could handle the case } X_i \in \{0, 1\} \text{ by explicitly bounding the non-centered quantity } E[e^{X_i}] = p_i e^1 + (1 - p_i) = 1 - p_i(e^1 + 1) \leq \exp(-p_i(e^1 + 1)) \text{(here } p_i := E[X_i]) \text{ and observing } \prod_{i=1}^{m} E[e^{X_i}] \leq \exp(-E[S_m](e^1 + 1)). \]

2
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 $\epsilon > 0$:

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

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

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

where

$$E[e^{tS_k} | X_1 \ldots X_{k-1}] = E[e^{tS_{k-1}} e^{tV_k} | X_1 \ldots X_{k-1}] = E[e^{tS_{k-1}} | X_1 \ldots X_{k-1}] E[e^{tV_k} | X_1 \ldots X_{k-1}] \leq E[e^{tS_{k-1}} | X_1 \ldots X_{k-1}] e^{2\epsilon^2 / 8}$$

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

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

Use the Chernoff bounding technique on $S_m$:

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

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 $e^{-2\epsilon^2 / \sum_{i=1}^{m} c_i^2}$. \hfill \square

3 McDiarmid (“Lipschitz” function of independent RVs)

McDiarmid’s inequality. Given iid random variables $X_1 \ldots X_m \in \mathcal{X}$, let $f : \mathcal{X}^m \rightarrow \mathbb{R}$ 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

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

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

$$\sum_{i=1}^{m} V_i = E[V|X_1 \ldots X_m] = V$$
In addition, $E[E[V|X_1 \ldots X_i]|X_1 \ldots X_{i-1}] = E[V|X_1 \ldots X_{i-1}]$ (by LIE), so we have

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

Thus $V_1 \ldots V_m$ is a martingale difference sequence wrt. $X_1 \ldots X_m$.\(^2\)

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

$$V_i = \sup_{x \in \mathcal{X}} E[V|X_1 \ldots X_i = x] - E[V|X_1 \ldots X_{i-1}] =: W_i$$

$$V_i \geq \inf_{x \in \mathcal{X}} E[V|X_1 \ldots X_i = x] - E[V|X_1 \ldots X_{i-1}] =: U_i$$

Using the “Lipschitz” condition on $f$:

$$W_i - U_i = \sup_{x,x' \in \mathcal{X}} E[V|X_1 \ldots X_i = x] - E[V|X_1 \ldots X_{i-1} = x']$$

$$= \sup_{x,x' \in \mathcal{X}} E[f(S)|X_1 \ldots X_i = x] - E[f(S)|X_1 \ldots X_{i-1} = x']$$

$$\leq c_i$$

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 $X_1 \ldots X_{i-1}$.

\[\square\]

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

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

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 $\mathbb{E}_X[X]$ is a constant.
• However, $\mathbb{E}_{X|Y}[X|Y]$ is a random variable (random over $Y$)! We can only compute a value for a specific $y \in Y$:

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

is a constant.

The law of iterated expectations (LIE)$^3$ states that

$$
\mathbb{E}_Y[\mathbb{E}_{X|Y}[X|Y]] = \mathbb{E}_X[X]
$$

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

$$
\mathbb{E}_Y[\mathbb{E}_{X|Y}[X|Y]] = \int_y P_Y(Y = y) \times \mathbb{E}_{X|Y}[X|Y = y] \, dy
$$

$$
= \int_y P_Y(Y = y) \times \left( \int_x P_{X|Y}(X = x|Y = y) \times x \, dx \right) \, dy
$$

$$
= \int_x \left( \int_y P_Y(Y = y) \times P_{X|Y}(X = x|Y = y) \, dy \right) \times x \, dx
$$

$$
= \int_x P_X(X = x) \times x \, dx
$$

$$
= \mathbb{E}_X[X]
$$

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

$$
\mathbb{E}_{Y,Z}[\mathbb{E}_{X|Y,Z}[X|Y,Z]|Z] = \mathbb{E}_{X|Z}[X|Z]
$$

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 $X_0 \ldots X_i$.
• $\mathbb{E}[|Z_i|] \leq \infty$
• $\mathbb{E}[Z_{i+1}|X_0 \ldots X_i] = Z_i$

$^3$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 $X_0 \ldots X_n$ be any sequence. We are interested in $Y$ that depends on all $X_0 \ldots X_n$; we assume $E[|Y|] \leq \infty$. We define $Z_i$ to be the expectation of $Y$ given $X_0 \ldots X_i$:

$$Z_i := E[Y|X_0 \ldots X_i]$$

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

$$E[Z_{i+1}|X_0 \ldots X_i] = E[E[Y|X_0 \ldots X_{i+1}]|X_0 \ldots X_i]$$

by def

$$= E[Y|X_0 \ldots X_i]$$

by LIE

$$= Z_i$$

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

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

since $E[X_j|X_1 \ldots X_i] = E[X_j] = 0$ 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 $Z_0, Z_1, \ldots$ is a martingale wrt. $X_0, X_1, \ldots$, then $V_1, V_2, \ldots$ 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 $X_1 \ldots X_i$.
- $E[|V_i|] = E[|Z_i - Z_{i-1}|] < \infty$
- $E[V_{i+1}|X_1 \ldots X_i] = E[Z_{i+1}|X_1 \ldots X_i] - Z_i = 0$