Concentration Inequalities for Martingales and Based on Martingales

ECE 534, UIUC, Fall 2026

1 Hoeffding's lemma and Hoeffding's inequality

Lemma 1 (Hoeffding's lemma)**.** *Let X be a random variable with $\mathbb{E}\boldsymbol{X} = \mathbf{0}$ and $\boldsymbol{a} \le \boldsymbol{X} \le \boldsymbol{b}$. Then for every $s \in \mathbb{R}$,*

$$\mathbb{E}\big[e^{sX}\big] \le \exp\Big(\frac{s^2(b-a)^2}{8}\Big).$$

Corollary 2 (Hoeffding's lemma for independent random variables)**.** *Let $X_1, \dots, X_n$ be independent random variables with $a_i \le X_i \le b_i$, and let $S_n = \sum_{i=1}^n X_i$. Then for every $s \in \mathbb{R}$,*

$$\mathbb{E}\big[e^{s(S_n - \mathbb{E}S_n)}\big] = \prod_{i=1}^n \mathbb{E}\big[e^{s(X_i - \mathbb{E}X_i)}\big] \le \exp\Big(\frac{s^2}{8}\sum_{i=1}^n (b_i - a_i)^2\Big).$$

(The equality is independence; the inequality is Lemma 1 applied to each $X_i - \mathbb{E}X_i$, which has mean zero and lies in an interval of length $b_i - a_i$.)

Theorem 3 (Hoeffding's inequality)**.** *Let $X_1, \dots, X_n$ be independent random variables with $a_i \le X_i \le b_i$, and let $S_n = \sum_{i=1}^n X_i$. Then for every $t > 0$,*

$$\mathbb{P}\big(S_n - \mathbb{E}S_n \ge t\big) \le \exp\Big(-\frac{2t^2}{\sum_{i=1}^n (b_i - a_i)^2}\Big), \qquad \mathbb{P}\big(|S_n - \mathbb{E}S_n| \ge t\big) \le 2\exp\Big(-\frac{2t^2}{\sum_{i=1}^n (b_i - a_i)^2}\Big).$$

Remark 4. Hoeffding's inequality follows from Corollary 2 by the Chernoff bound $\mathbb{P}(S_n - \mathbb{E}S_n \ge t) \le e^{-st}\,\mathbb{E}[e^{s(S_n - \mathbb{E}S_n)}]$ and the choice $s = 4t/\sum_i (b_i - a_i)^2$. The two results below extend it in two directions: Azuma's inequality drops independence and replaces the sum by a martingale, and McDiarmid's inequality replaces the sum S_n by a general function $f(X_1, \dots, X_n)$ of independent variables with bounded differences.

When is Hoeffding's inequality useful? How should $b_i - a_i$ scale?

Write $L^2 \coloneqq \sum_{i=1}^n (b_i - a_i)^2$, so the bound reads $\mathbb{P}(|S_n - \mathbb{E}S_n| \ge t) \le 2e^{-2t^2/L^2}$. The bound is trivial (at least 1) when $t \lesssim L$ and becomes small only when $t \gg L$. So L is the scale of the fluctuations that Hoeffding's inequality "sees": it guarantees $|S_n - \mathbb{E}S_n| = O(L)$ with high probability. For the inequality to say something useful, L must be small compared with the deviations we care about.

Sample means. The most common use is the sample mean $\bar{X}_n = S_n/n$. Taking $t = n\varepsilon$,

$$\mathbb{P}\big(|\bar{X}_n - \mathbb{E}\bar{X}_n| \ge \varepsilon\big) \le 2\exp\Big(-\frac{2n^2\varepsilon^2}{\sum_{i=1}^n (b_i - a_i)^2}\Big).$$

- If the ranges are bounded by a constant, $b_i - a_i \le R$, then $\sum_i (b_i - a_i)^2 \le nR^2$ and

$$\mathbb{P}\big(|\bar{X}_n - \mathbb{E}\bar{X}_n| \ge \varepsilon\big) \le 2e^{-2n\varepsilon^2/R^2} :$$

the probability of an ε-deviation decays *exponentially in* n. Equivalently, the sample mean is within ε of its mean with probability at least $1 - \delta$ as soon as

$$n \ge \frac{R^2 \ln(2/\delta)}{2\varepsilon^2}.$$

The typical deviation shrinks like $R/\sqrt{n}$, the same $1/\sqrt{n}$ rate as in the central limit theorem.

- More generally, the bound tends to 0 for every fixed $\varepsilon > 0$ exactly when $\sum_{i=1}^n (b_i - a_i)^2 = o(n^2)$. If all ranges equal some R_n that may grow with n, this means $R_n = o(\sqrt{n})$: the ranges may grow, but more slowly than $\sqrt{n}$.

No single term should dominate. Because the ranges enter through the sum of squares, one variable with a huge range can ruin the bound even if all the others are tiny. For example, if $b_1 - a_1 = n$ and all other ranges are 1, then $L^2 \ge n^2$ and the bound on $\mathbb{P}(|\bar{X}_n - \mathbb{E}\bar{X}_n| \ge \varepsilon)$ never drops below $2e^{-2\varepsilon^2}$, no matter how large n is. This is the right behavior: one variable whose range is comparable to the whole sum can by itself move the average by a constant amount.

Range versus variance. Hoeffding's inequality uses only the ranges, not the variances. Since a variable in $[a_i, b_i]$ has variance at most $(b_i - a_i)^2/4$, the bound is sharp when the variance is comparable to the squared range (for example, fair ± 1 coin flips). It is loose when the variance is much smaller than the range: for Bernoulli(p) variables with small p, the range is 1 but the variance is only $p(1-p)$, and the true fluctuations of S_n are of order $\sqrt{np}$ rather than $\sqrt{n}$. In that regime, inequalities that use the variance (such as Bernstein's inequality or the Chernoff bound for Bernoulli sums) give much better results.

2 Azuma's inequality

Hoeffding's inequality bounds a sum of independent bounded terms. Azuma's inequality keeps the boundedness of the terms but replaces independence by a much weaker requirement: the terms are the increments of a martingale.

Theorem 5 (Azuma's inequality)**.** *Let $Z_0, Z_1, \dots, Z_n$ be a martingale with respect to a filtration $\mathcal{F}_0 \subseteq \mathcal{F}_1 \subseteq \cdots \subseteq \mathcal{F}_n$, and suppose there are constants $c_1, \dots, c_n$, not all zero, such that*

$$|Z_k - Z_{k-1}| \le c_k \qquad \text{for } k = 1, \dots, n.$$

Then for every $t > 0$,

$$\mathbb{P}\big(Z_n - Z_0 \ge t\big) \le \exp\Big(-\frac{t^2}{2\sum_{k=1}^n c_k^2}\Big), \qquad \mathbb{P}\big(|Z_n - Z_0| \ge t\big) \le 2\exp\Big(-\frac{t^2}{2\sum_{k=1}^n c_k^2}\Big).$$

Proof of Theorem 5. Let $D_k := Z_k - Z_{k-1}$ be the increments, so $Z_n - Z_0 = \sum_{k=1}^n D_k$. We know two things about D_k:

- D_k has conditional mean zero given the past: $\mathbb{E}[D_k \mid \mathcal{F}_{k-1}] = \mathbb{E}[Z_k \mid \mathcal{F}_{k-1}] - Z_{k-1} = Z_{k-1} - Z_{k-1} = 0$, by the martingale property and because Z_{k-1} is $\mathcal{F}_{k-1}$-measurable;

- $-c_k \le D_k \le c_k$.

Step 1: conditional moment generating function of each increment. Fix $s \in \mathbb{R}$. Given $\mathcal{F}_{k-1}$, the increment D_k has mean zero and lies in $[-c_k, c_k]$, so the argument of Hoeffding's lemma with $a = -c_k$, $b = c_k$ gives the cosh bound

$$\mathbb{E}\big[e^{sD_k} \,\big|\, \mathcal{F}_{k-1}\big] \le \cosh(sc_k) \le e^{s^2 c_k^2/2}.$$

Step 2: the moment generating function of the sum. Let $S_k := D_1 + \cdots + D_k$, with $S_0 = 0$. Since S_{k-1} is $\mathcal{F}_{k-1}$-measurable, the tower rule and pulling out known factors give

$$\mathbb{E}\big[e^{sS_k}\big] = \mathbb{E}\Big[e^{sS_{k-1}}\,\mathbb{E}\big[e^{sD_k} \,\big|\, \mathcal{F}_{k-1}\big]\Big] \le e^{s^2 c_k^2/2}\,\mathbb{E}\big[e^{sS_{k-1}}\big].$$

Applying this for $k = n, n-1, \ldots, 1$ in turn,

$$\mathbb{E}\big[e^{s(Z_n - Z_0)}\big] = \mathbb{E}\big[e^{sS_n}\big] \le \exp\Big(\frac{s^2}{2}\sum_{k=1}^{n} c_k^2\Big).$$

Step 3: Chernoff bound. For $s > 0$, Markov's inequality applied to e^{sS_n} gives

$$\mathbb{P}\big(Z_n - Z_0 \ge t\big) \le e^{-st}\,\mathbb{E}\big[e^{sS_n}\big] \le \exp\Big(-st + \frac{s^2}{2}\sum_{k=1}^{n} c_k^2\Big).$$

The exponent is a quadratic in s, minimized at $s = t/\sum_k c_k^2$, where it equals $-t^2/(2\sum_k c_k^2)$. This proves the one-sided bound. Applying it to the martingale $-Z_k$, which has the same increment bounds, and adding the two bounds gives the two-sided version. □

Remark 6 (Comparison with Hoeffding; looking ahead to McDiarmid).

- The condition $|D_k| \le c_k$ places D_k in an interval of length $2c_k$. Hoeffding's inequality, for an increment in an interval of length ℓ_k, gives exponent $-2t^2/\sum_k \ell_k^2$; with $\ell_k = 2c_k$ this is exactly $-t^2/(2\sum_k c_k^2)$, so Azuma's bound is what one expects.
- If instead each D_k is known to lie, given $\mathcal{F}_{k-1}$, in an interval of length c_k (whose endpoints may depend on the past), the same proof with Hoeffding's lemma in Step 1 gives the sharper bound $\exp\big(-2t^2/\sum_k c_k^2\big)$. McDiarmid's inequality in Section 3 is this version applied to the Doob martingale $Z_k = \mathbb{E}[f(X) \mid \mathcal{F}_k]$.

3 McDiarmid's inequality

McDiarmid's inequality applies the martingale method of Section 2 to a function of independent random variables. Independence is used to show that the Doob martingale of the function has increments that lie, given the past, in intervals of length c_k; Steps 4 and 5 of the proof are then the argument of Azuma's inequality.

Definition 7 (Bounded differences)**.** A function $f(x_1, \ldots, x_n)$ has the *bounded differences property* with constants $c_1, \ldots, c_n \ge 0$ if changing only the i-th input changes the output by at most c_i:

$$\big|f(x_1, \ldots, x_i, \ldots, x_n) - f(x_1, \ldots, x_i', \ldots, x_n)\big| \le c_i \qquad \text{for all inputs and all } i.$$

Theorem 8 (McDiarmid's inequality)**.** *Let $X_1, \dots, X_n$ be independent random variables and let f have the bounded differences property with constants $c_1, \dots, c_n$, not all zero. Write $X = (X_1, \dots, X_n)$. Then for every $t > 0$,*

$$\mathbb{P}\big(f(X) - \mathbb{E}f(X) \geq t\big) \leq \exp\Big(-\frac{2t^2}{\sum_{i=1}^n c_i^2}\Big), \qquad \mathbb{P}\big(|f(X) - \mathbb{E}f(X)| \geq t\big) \leq 2\exp\Big(-\frac{2t^2}{\sum_{i=1}^n c_i^2}\Big).$$

With $f(x) = \sum_i x_i$ and $X_i \in [a_i, b_i]$ we have $c_i = b_i - a_i$, and McDiarmid's inequality reduces to Hoeffding's inequality.

Proof. Step 0: everything is finite. Changing the inputs of f one at a time from any point x to a fixed point y shows $|f(x) - f(y)| \leq \sum_i c_i$. So f is bounded, and all the expectations below are finite.

Step 1: the Doob martingale. Let $\mathcal{F}_0 = \{\emptyset, \Omega\}$ and $\mathcal{F}_k = \sigma(X_1, \dots, X_k)$, and let

$$Z_k \coloneqq \mathbb{E}[f(X) \mid \mathcal{F}_k], \qquad k = 0, 1, \dots, n.$$

This is a martingale with respect to $(\mathcal{F}_k)$ (the Doob martingale of $f(X)$): by the tower rule, $\mathbb{E}[Z_{k+1} \mid \mathcal{F}_k] = \mathbb{E}\big[\mathbb{E}[f(X) \mid \mathcal{F}_{k+1}] \,\big|\, \mathcal{F}_k\big] = \mathbb{E}[f(X) \mid \mathcal{F}_k] = Z_k$. Its endpoints are $Z_0 = \mathbb{E}f(X)$ and $Z_n = f(X)$. With increments $D_k \coloneqq Z_k - Z_{k-1}$, the sum telescopes:

$$f(X) - \mathbb{E}f(X) = Z_n - Z_0 = \sum_{k=1}^n D_k.$$

Step 2: a formula for Z_k. For each k define the "partially averaged" function

$$g_k(x_1, \dots, x_k) \coloneqq \mathbb{E}\big[f(x_1, \dots, x_k, X_{k+1}, \dots, X_n)\big],$$

which fixes the first k inputs and averages over the rest; $g_n = f$ and $g_0 = \mathbb{E}f(X)$. Because $X_{k+1}, \dots, X_n$ are independent of $\mathcal{F}_k$, conditioning on $\mathcal{F}_k$ amounts to plugging in the observed values of $X_1, \dots, X_k$ and averaging over the rest, so

$$Z_k = g_k(X_1, \dots, X_k).$$

Two properties of g_k will be used:

$$g_{k-1}(x_1, \dots, x_{k-1}) = \mathbb{E}\big[g_k(x_1, \dots, x_{k-1}, X_k)\big], \tag{1}$$

which holds because, by independence, averaging over $X_k, \dots, X_n$ can be done by first averaging over $X_{k+1}, \dots, X_n$ and then over X_k; and

$$\big|g_k(x_1, \dots, x_{k-1}, x) - g_k(x_1, \dots, x_{k-1}, x')\big| \leq c_k \qquad \text{for all } x, x', \tag{2}$$

that is, g_k inherits the bounded differences property in its last input. Property (2) is the key step of the proof; it is explained in detail in Section 3.1 below.

Step 3: each increment is a bounded, zero-mean random variable given $\mathcal{F}_{k-1}$. Fix $s \in \mathbb{R}$ and $1 \leq k \leq n$, and fix observed values $u = (x_1, \dots, x_{k-1})$ of the first $k-1$ variables. With these values plugged in, the increment becomes

$$\Delta_u \coloneqq g_k(u, X_k) - g_{k-1}(u),$$

which depends only on the single random variable X_k. By (1), $\mathbb{E}\Delta_u = 0$. By (2), all values of Δ_u lie within c_k of each other, so Δ_u lies in an interval $[a_u, a_u + c_k]$, where a_u is the smallest value of $g_k(u,x) - g_{k-1}(u)$ over x. Hoeffding's lemma therefore gives $\mathbb{E}[e^{s\Delta_u}] \le e^{s^2 c_k^2/8}$, whatever the observed values u are. Since X_k is independent of $\mathcal{F}_{k-1}$, conditioning on $\mathcal{F}_{k-1}$ amounts to plugging in the values of $X_1, \dots, X_{k-1}$, so

$$\mathbb{E}\big[e^{sD_k} \mid \mathcal{F}_{k-1}\big] = \mathbb{E}\big[e^{s\Delta_u}\big]\Big|_{u=(X_1,\dots,X_{k-1})} \le e^{s^2 c_k^2/8}. \tag{3}$$

Step 4: the moment generating function of the sum. Let $S_k := D_1 + \cdots + D_k$, with $S_0 = 0$. Since S_{k-1} is $\mathcal{F}_{k-1}$-measurable, the tower rule and pulling out known factors give, with (3),

$$\mathbb{E}\big[e^{sS_k}\big] = \mathbb{E}\Big[e^{sS_{k-1}}\,\mathbb{E}\big[e^{sD_k} \mid \mathcal{F}_{k-1}\big]\Big] \le e^{s^2 c_k^2/8}\,\mathbb{E}\big[e^{sS_{k-1}}\big].$$

Applying this for $k = n, n-1, \dots, 1$ in turn, $\mathbb{E}\big[e^{s(f(X)-\mathbb{E}f(X))}\big] \le \exp\big(\frac{s^2}{8}\sum_k c_k^2\big)$.

Step 5: Chernoff bound. For $s > 0$, Markov's inequality gives

$$\mathbb{P}\big(f(X) - \mathbb{E}f(X) \ge t\big) \le e^{-st}\,\mathbb{E}\big[e^{sS_n}\big] \le \exp\Big(-st + \frac{s^2}{8}\sum_{k=1}^{n} c_k^2\Big),$$

and the choice $s = 4t/\sum_k c_k^2$ gives the exponent $-2t^2/\sum_k c_k^2$. Applying the result to $-f$ and adding the two bounds gives the two-sided version. □

3.1 A closer look at Step 2: why g_k inherits the bounded differences property

We now justify (2), stated as a proposition.

Proposition 9. *For any fixed $x_1, \dots, x_{k-1}$ and any two values x, x' of the k-th input,*

$$\big|g_k(x_1, \dots, x_{k-1}, x) - g_k(x_1, \dots, x_{k-1}, x')\big| \le c_k.$$

To keep notation short, write $u = (x_1, \dots, x_{k-1})$ for the fixed earlier inputs and $W = (X_{k+1}, \dots, X_n)$ for the variables not yet revealed. Then

$$g_k(u,x) = \mathbb{E}\big[f(u,x,W)\big].$$

In words, $g_k(u,x)$ plugs in the known values u and x and averages f over the randomness that remains.

Proof. The argument has three steps.

Step (i): the difference of two averages is the average of the difference. Both $g_k(u,x)$ and $g_k(u,x')$ are averages over the same random vector W. The only thing that changes between them is the constant plugged into position k; the distribution of W does not depend on that choice. So by linearity of expectation,

$$g_k(u,x) - g_k(u,x') = \mathbb{E}\big[f(u,x,W)\big] - \mathbb{E}\big[f(u,x',W)\big] = \mathbb{E}\big[f(u,x,W) - f(u,x',W)\big].$$

Step (ii): the difference inside the average is always small. Fix any possible outcome w of W. Then $f(u,x,w)$ and $f(u,x',w)$ are two evaluations of f whose inputs agree everywhere except in position k. The bounded differences property of f applies directly:

$$\big|f(u,x,w) - f(u,x',w)\big| \le c_k \qquad \text{for every outcome } w.$$

This holds for every outcome, not just typical ones, because the property of f is assumed for all input values. In other words, the random variable $V := f(u, x, W) - f(u, x', W)$ always lies in $[-c_k, c_k]$.

Step (iii): averaging cannot leave an interval. If a random variable always lies in $[-c_k, c_k]$, so does its average: $|\mathbb{E}V| \le \mathbb{E}|V| \le c_k$. Combining with Step (i),

$$\big|g_k(u, x) - g_k(u, x')\big| = \big|\mathbb{E}V\big| \le c_k.$$

□

A concrete picture. Take $n = 2$, $k = 1$, and suppose X_2 takes values $w_1, \dots, w_m$ with probabilities $p_1, \dots, p_m$. Then $g_1(x) = \sum_j p_j\, f(x, w_j)$, so

$$g_1(x) - g_1(x') = \sum_j p_j \big[f(x, w_j) - f(x', w_j)\big].$$

Each bracket lies in $[-c_1, c_1]$, and the weights p_j are nonnegative and sum to 1. A weighted average of numbers in an interval stays in that interval, so the result is in $[-c_1, c_1]$. The general case is the same idea, with the sum replaced by an expectation.

Where independence comes in. The argument above only uses the definition of g_k as an average over W with the first k inputs held fixed. Independence is what makes this definition the right one. Because W is independent of $X_1, \dots, X_k$, conditioning on the first k variables does not change the distribution of W; that is why $Z_k = \mathbb{E}[f(X) \mid \mathcal{F}_k] = g_k(X_1, \dots, X_k)$. If the later variables depended on X_k, then changing x to x' would also change the distribution being averaged over, Step (i) would fail, and the bound could break.

Why we need it. In Step 3 of the proof of Theorem 8, Proposition 9 shows that the increment $\Delta_u = g_k(u, X_k) - g_{k-1}(u)$ takes values that are all within c_k of each other. So Δ_u lies in an interval of length c_k, which is exactly what Hoeffding's lemma needs.