Randomized Algorithms

Lecture 19: Martingales and concentration

Alex Gittens
Department of Computer Science, Rensselaer Polytechnic Institute

Fall 2026

From sums to functions

For independent bounded variables, we have bounded deviations of sums by controlling their exponential moments.

$$S = \sum_{i=1}^{m} X_i.$$

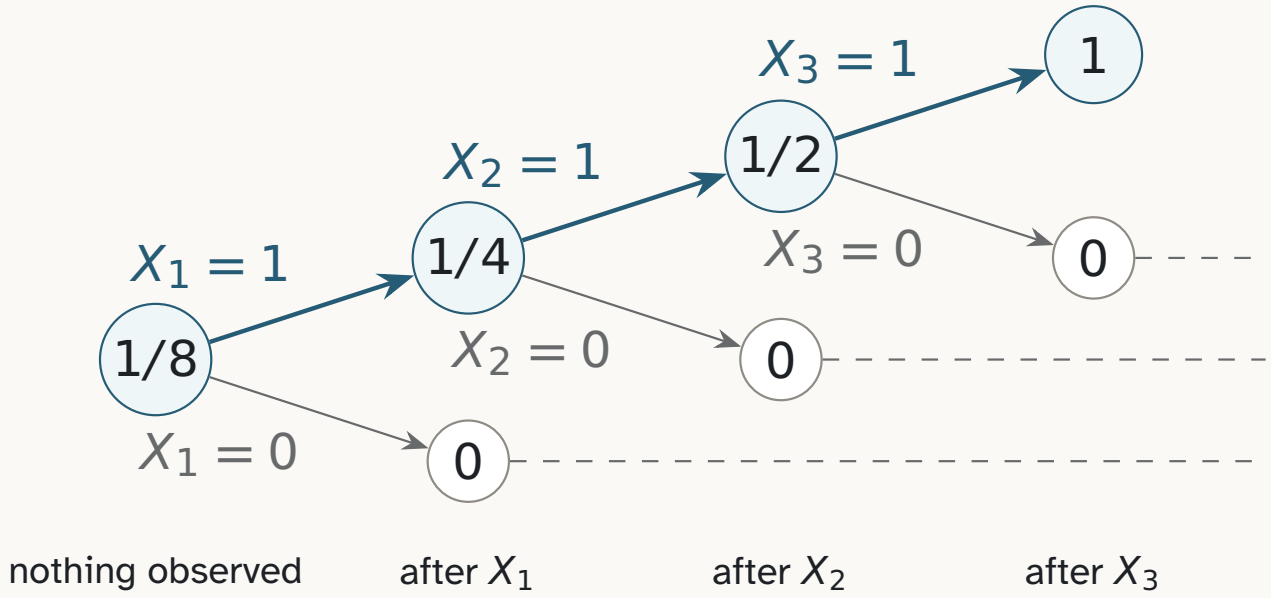
Now suppose the answer is a nonlinear function,

$$Z = f(X_1, \ldots, X_m).$$

Can we still express $Z - \mathbb{E}Z$ as a sum? Reveal the inputs one at a time and track the expected final answer.

Expected answers as inputs are revealed

Let X_1, X_2, X_3 be independent fair bits and $Z = X_1 X_2 X_3$. The answer is 1 only if all three bits are 1. What is the expected answer before and after each observation?



Each node shows the expected final product. Once a zero appears, the product is known; the dashed branches stay at zero.

Add the changes along a path

Consider the outcome $(X_1, X_2, X_3) = (1, 0, 1)$. The successive changes in the expected product are

$$\begin{aligned}
&\mathbb{E}[Z \mid X_1 = 1] - \mathbb{E}Z, \\
&\mathbb{E}[Z \mid X_1 = 1, X_2 = 0] - \mathbb{E}[Z \mid X_1 = 1], \\
&\mathbb{E}[Z \mid X_1 = 1, X_2 = 0, X_3 = 1] - \mathbb{E}[Z \mid X_1 = 1, X_2 = 0].
\end{aligned}$$

Substituting the expected values and adding gives

$$\left(\frac{1}{4} - \frac{1}{8}\right) + \left(0 - \frac{1}{4}\right) + (0 - 0) = 0 - \frac{1}{8} = -\frac{1}{8}.$$

The intermediate conditional expectations cancel, leaving

$$\mathbb{E}[Z \mid X_1 = 1, X_2 = 0, X_3 = 1] - \mathbb{E}Z = -\frac{1}{8}.$$

All inputs are observed, so this is the realized deviation $Z - \mathbb{E}Z$.

Write the deviation as a sum

For independent inputs $X_1, \ldots, X_m$ and an integrable answer $f(X)$, let

$$M_i = \mathbb{E}[f(X) \mid X_1, \ldots, X_i], \qquad D_i = M_i - M_{i-1}.$$

M_i is the expected final answer after the first i observations. D_i is the change caused by learning X_i.

$$\begin{aligned}\sum_{i=1}^{m} D_i &= (M_1 - M_0) + (M_2 - M_1) + \cdots + (M_m - M_{m-1}) \\ &= M_m - M_0 \\ &= f(X) - \mathbb{E}f(X).\end{aligned}$$

We have a sum, even though the answer is nonlinear.

What can we say about the next change?

For $Z = X_1X_2X_3$, consider learning the second bit.

Observed	Expected answer	Next change D_2
$X_1 = 0$	0	0
$X_1 = 1$	$1/4$	$\pm 1/4$, each with probability $1/2$

The distribution of D_2 depends on the first observation, so the changes need not be independent. Given the past, however, the next change has mean zero and lies in a short interval. For independent sums we

multiplied bounds on individual MGFs. Here we will condition on the past and bound

$$\mathbb{E}[\, e^{\lambda D_i} \mid X_1, \ldots, X_{i-1}].$$

Martingales

The information revealed so far is $\mathcal{F}_i = \sigma(X_1, \ldots, X_i)$. These increasing collections of information form a **filtration**.

Martingale

An integrable sequence (M_i) is a martingale with respect to $(\mathcal{F}_i)$ if, almost surely,

$$\mathbb{E}[M_i \mid \mathcal{F}_i] = M_i, \qquad \mathbb{E}[M_i \mid \mathcal{F}_{i-1}] = M_{i-1}.$$

For $M_i = \mathbb{E}[f(X) \mid \mathcal{F}_i]$, the tower property gives

$$\mathbb{E}[M_i \mid \mathcal{F}_{i-1}] = \mathbb{E}[\mathbb{E}[f(X) \mid \mathcal{F}_i] \mid \mathcal{F}_{i-1}] = M_{i-1}. \quad \square$$

Thus $\mathbb{E}[D_i \mid \mathcal{F}_{i-1}] = 0$: the next change averages to zero given the past, just as in the three-bit example.

Absolute changes and interval widths

For $Z = X_1X_2X_3$, the largest conditional ranges occur when all previous bits are 1:

Change	Possible values	Bound on $\lvert D_i \rvert$	Interval width
D_1	$-1/8,\ +1/8$	$1/8$	$1/4$
D_2	$-1/4,\ +1/4$	$1/4$	$1/2$
D_3	$-1/2,\ +1/2$	$1/2$	1

If a previous bit is zero, the next change is zero. For example, after $X_1 = 1$, the next expectation is 0 or $1/2$, and the current expectation is $1/4$:

$$D_2 \in \{0 - 1/4,\ 1/2 - 1/4\} = \{-1/4, +1/4\}.$$

The interval width is $1/4 - (-1/4) = 1/2$.

Azuma–Hoeffding

Let $(M_i, \mathcal{F}_i)_{i=0}^m$ be a martingale and $D_i = M_i - M_{i-1}$.

Azuma–Hoeffding

Suppose A_i, B_i are $\mathcal{F}_{i-1}$-measurable and, almost surely,

$$A_i \le D_i \le B_i, \qquad B_i - A_i \le c_i,$$

where the c_i are deterministic. For $V = \sum_i c_i^2 > 0$,

$$\boxed{\Pr(|M_m - M_0| \ge t) \le 2e^{-2t^2/V}, \quad t > 0.}$$

The interval endpoints may depend on the observations so far; the width bounds c_i are fixed.

Recall Hoeffding's lemma

Hoeffding's lemma

If $\mathbb{E}Y = 0$ and $Y \in [a, b]$ almost surely, then

$$\mathbb{E}e^{\lambda Y} \le e^{\lambda^2 (b-a)^2/8}, \qquad \lambda \in \mathbb{R}.$$

Fix the past $\mathcal{F}_{i-1}$. Under this conditional law, D_i has mean zero and lies in $[A_i, B_i]$, of width at most c_i. Apply the same lemma:

$$\boxed{\mathbb{E}[\, e^{\lambda D_i} \mid \mathcal{F}_{i-1}] \le e^{\lambda^2 c_i^2/8}.}$$

Azuma–Hoeffding: bound the MGF

Write $V = \sum_i c_i^2$. Since $M_{m-1} - M_0$ is known given $\mathcal{F}_{m-1}$,

$$\begin{aligned}\mathbb{E}e^{\lambda(M_m - M_0)} &= \mathbb{E}\big[e^{\lambda(M_{m-1}-M_0)}\mathbb{E}[\,e^{\lambda D_m} \mid \mathcal{F}_{m-1}]\big] \\ &\le e^{\lambda^2 c_m^2/8}\mathbb{E}e^{\lambda(M_{m-1}-M_0)}.\end{aligned}$$

Repeating the same step gives

$$\mathbb{E}e^{\lambda(M_m - M_0)} \le e^{\lambda^2 V/8}.$$

Conditioning replaces the independence factorization in the Hoeffding proof.

Azuma–Hoeffding: tail bound

For $\lambda > 0$, exponential Markov and the moment bound give

$$\Pr(M_m - M_0 \geq t) \leq e^{-\lambda t}\mathbb{E}e^{\lambda(M_m - M_0)} \leq e^{-\lambda t + \lambda^2 V/8}.$$

The exponent has derivative $-t + \lambda V/4$. Set $\lambda = 4t/V$ to obtain

$$\Pr(M_m - M_0 \geq t) \leq e^{-2t^2/V}.$$

Apply the argument to $-M_i$, whose increments have the same conditional interval widths. Taking a union bound over the two tails gives

$$\boxed{\Pr(|M_m - M_0| \geq t) \leq 2e^{-2t^2/V}.}$$

□

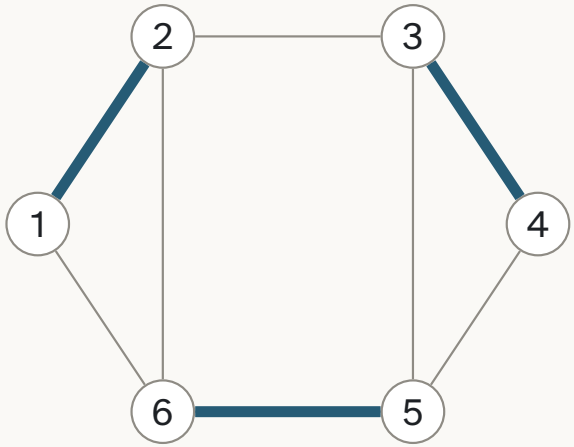
PART 1

Matching in a random graph

1

Maximum matching in a random graph

A matching is a collection of edges with no shared endpoint. Write $\nu(G)$ for the maximum possible number of edges in a matching.



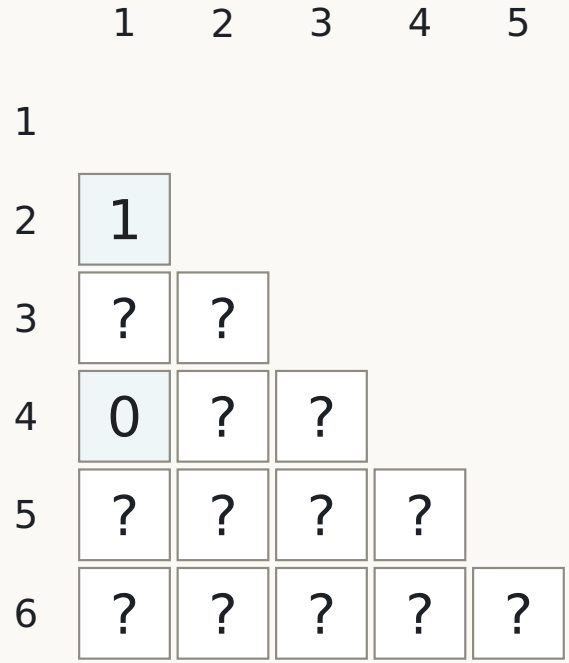
thick edges: three pairs

$\nu(G) = 3$

Let $G \sim G(n, p)$: on n labeled vertices, every potential edge is present independently with probability p. We will control fluctuations of $\nu(G)$ across these random graphs.

Independent entries encode the graph

For $G \sim G(n, p)$, the entries A_{ij} with $i > j$ are independent Bernoulli(p). The upper triangle duplicates them.



1: present 0: absent ?: unrevealed
There are $m = \binom{n}{2}$ independent entries.

The matching number $\nu(G)$ is a highly nonlinear function of these entries.

Changing one edge

Let H be the graph with that edge removed. Then

$$\nu(H) \leq \nu(H + e) \leq \nu(H) + 1.$$

Adding an edge enlarges the set of possible matchings, giving the first inequality. A matching using the new edge loses only that edge when we remove it, giving the second. Replacing one edge indicator changes the answer by at most 1.

□

Matching: verify the hypothesis

Reveal the edge indicators $X_1, \dots, X_m$, and let

$$M_i = \mathbb{E}[\,\nu(G) \mid X_1, \dots, X_i].$$

Fix the observed entries $x_{<i} = (x_1, \dots, x_{i-1})$. Let G_1 and G_0 be completions with the next edge present or absent and the same remaining entries $X_{>i}$. Independence leaves those remaining entries with the same distribution in both cases. Thus

$$\begin{aligned} &\big|\mathbb{E}[\,\nu(G) \mid X_{<i} = x_{<i}, X_i = 1] \\ &\qquad - \mathbb{E}[\,\nu(G) \mid X_{<i} = x_{<i}, X_i = 0]\big| \\ &\quad = |\mathbb{E}[\,\nu(G_1) - \nu(G_0)]| \le \mathbb{E}|\nu(G_1) - \nu(G_0)| \le 1. \end{aligned}$$

The two possible values of M_i differ by at most 1. Subtracting M_{i-1} preserves the width, so $c_i = 1$.

□

The edge exposure bound

There are $m = \binom{n}{2}$ conditional intervals of width at most 1.
Azuma–Hoeffding gives

$$\Pr(|\nu(G) - \mathbb{E}\nu(G)| \geq t) \leq 2e^{-2t^2/m}.$$

For tail probability δ, the resulting scale of likely deviations is

$$t_\delta = \sqrt{\frac{m}{2} \log \frac{2}{\delta}} = \frac{1}{2}\sqrt{n(n-1) \log \frac{2}{\delta}}.$$

This has scale n, the same scale as the entire range

$$0 \leq \nu(G) \leq \lfloor n/2 \rfloor.$$

Why did the proof accumulate $\binom{n}{2}$ contributions?

PART 2

Vertex exposure

2

Reveal columns instead of entries

	1	2	3	4	5	6
1						
2	1					
3	0	?				
4	0	?	?			
5	0	?	?	?		
6	1	?	?	?	?	

Reveal the strict lower triangle one column at a time. Column j reveals all edges from j to vertices $j+1, \ldots, n$. There are $n-1$ independent columns, since they contain disjoint sets of independent entries.

Replacing a column changes several edges. How much can it change the matching number?

Changing all the edges in one column

Let G and $G\prime$ agree in every lower-triangular column except column j. They differ only in edges incident to j. Delete that vertex from both graphs:

$$G - j = G' - j = H.$$

A matching of H works in either graph. Deleting j from any matching loses at most one edge. Therefore

$$\nu(G), \nu(G') \in [\,\nu(H), \nu(H) + 1\,].$$

Thus $|\nu(G) - \nu(G')| \leq 1$, even though many edges may have changed. □

The improved matching bound

Write the independent lower-triangular columns as

$$B_j = (A_{j+1,j}, \ldots, A_{n,j}), \qquad j = 1, \ldots, n-1.$$

Define $M_j = \mathbb{E}[\,\nu(G) \mid B_1, \ldots, B_j\,]$. Its endpoints are $M_0 = \mathbb{E}\nu(G)$ and $M_{n-1} = \nu(G)$. The same completion comparison gives conditional

interval width at most 1. Azuma applies to this filtration, including vector observations:

$$\boxed{\Pr(|\nu(G) - \mathbb{E}\nu(G)| \geq t) \leq 2e^{-2t^2/(n-1)}.}$$

For tail probability δ, the resulting scale of likely deviations is

$$\sqrt{\frac{n-1}{2} \log \frac{2}{\delta}}.$$

Vertex exposure improves the scale from n to $\sqrt{n}$. Vector observations: construction and proof in the appendix.

Exercise: how many time slots?

Vertices are activities; an edge joins two activities that cannot use the same time slot. The chromatic number $\chi(G)$ is the minimum number of slots needed.

Exercise. Show that

$$\chi(G - v) \leq \chi(G) \leq \chi(G - v) + 1.$$

Use a vertex exposure process to obtain a concentration bound for $\chi(G)$ when $G \sim G(n, p)$.

Concentration of the chromatic number

Restrict a coloring of G to $G - v$:

$$\chi(G - v) \le \chi(G).$$

Color $G - v$ optimally, then give v a new color:

$$\chi(G) \le \chi(G - v) + 1.$$

Replacing one exposed column leaves the same graph after deleting its vertex. The two chromatic numbers differ by at most 1, so

$$\boxed{\Pr(|\chi(G) - \mathbb{E}\chi(G)| \ge t) \le 2e^{-2t^2/(n-1)}.}$$

□

For $p = 1/2$, $\mathbb{E}\chi(G) \sim n/(2\log_2 n)$. Our $\sqrt{n}$ deviation scale is small compared with the mean, but sharper concentration bounds are known. For any $\omega(n) \to \infty$, an interval of length $\omega(n)\sqrt{n}/\log n$ contains $\chi(G)$ with probability tending to one.

Exercise: reveal the summands

Let $X_1, \ldots, X_m$ be independent, with $X_i \in [a_i, b_i]$, and let $S = \sum_i X_i$. Reveal the variables one at a time:

$$M_i = \mathbb{E}[S \mid X_1, \ldots, X_i].$$

1. Compute M_i in terms of the observed variables and the expectations of the remaining variables.
2. Find $D_i = M_i - M_{i-1}$. What is its conditional interval width?
3. Apply Azuma–Hoeffding. Which inequality do you recover?

Hoeffding from Azuma–Hoeffding

Independence gives

$$M_i = \sum_{j \le i} X_j + \sum_{j > i} \mathbb{E}X_j.$$

Consequently,

$$D_i = X_i - \mathbb{E}X_i \in [\,a_i - \mathbb{E}X_i,\ b_i - \mathbb{E}X_i\,].$$

The interval width is $b_i - a_i$. Since $M_m - M_0 = S - \mathbb{E}S$, Azuma–Hoeffding gives

$$\boxed{\Pr(|S - \mathbb{E}S| \ge t) \le 2\exp\left(-\frac{2t^2}{\sum_i (b_i - a_i)^2}\right).}$$

This is Hoeffding's inequality.

□

PART 3

Functions with bounded differences

3

From conditional ranges to sensitivity

For fixed previous inputs $x_{<i}$, write

$$g_i(a) = \mathbb{E}[f(x_{<i}, a, X_{>i})].$$

Azuma needs a bound on the range of these conditional expectations:

$$\sup_{a,b} |g_i(a) - g_i(b)|.$$

We bounded it before averaging:

$$|g_i(a) - g_i(b)| \le \mathbb{E}|f(x_{<i}, a, X_{>i}) - f(x_{<i}, b, X_{>i})| \le c_i.$$

This follows if replacing input i changes f by at most c_i, for every choice of the other inputs. That sensitivity bound is often easier to prove. It may exceed the conditional range that Azuma actually needs.

McDiarmid's inequality

Bounded differences

Let $X_1, \ldots, X_m$ be independent and $f(X)$ integrable. If replacing input i changes f by at most c_i, then

$$\Pr(|f(X) - \mathbb{E}f(X)| \geq t) \leq 2\exp\left(-\frac{2t^2}{\sum_i c_i^2}\right).$$

The sensitivity condition holds for every pair in the product of the input spaces. Assume $\sum_i c_i^2 > 0$; if it is zero, f is constant. The proof uses the conditional expectations constructed for the graph examples.

Proof of McDiarmid

The **Doob martingale** associated with $f(X)$ is

$$M_i = \mathbb{E}[f(X) \mid X_1, \ldots, X_i].$$

Fix the previous inputs $x_1, \ldots, x_{i-1}$ and compare next values a, b.

Average over the same remaining random inputs:

$$\begin{aligned}\Big|&\mathbb{E}f(x_1, \ldots, x_{i-1}, a, X_{i+1}, \ldots, X_m) \\ &- \mathbb{E}f(x_1, \ldots, x_{i-1}, b, X_{i+1}, \ldots, X_m)\Big| \le c_i.\end{aligned}$$

Because the inputs are independent, observing the prefix and the next input does not change the distribution of the remaining inputs. The displayed expectations are therefore the possible next values of M_i; any two differ by at most c_i. Thus D_i has conditional range of width at most c_i.

Apply the conditional interval form of Azuma–Hoeffding.

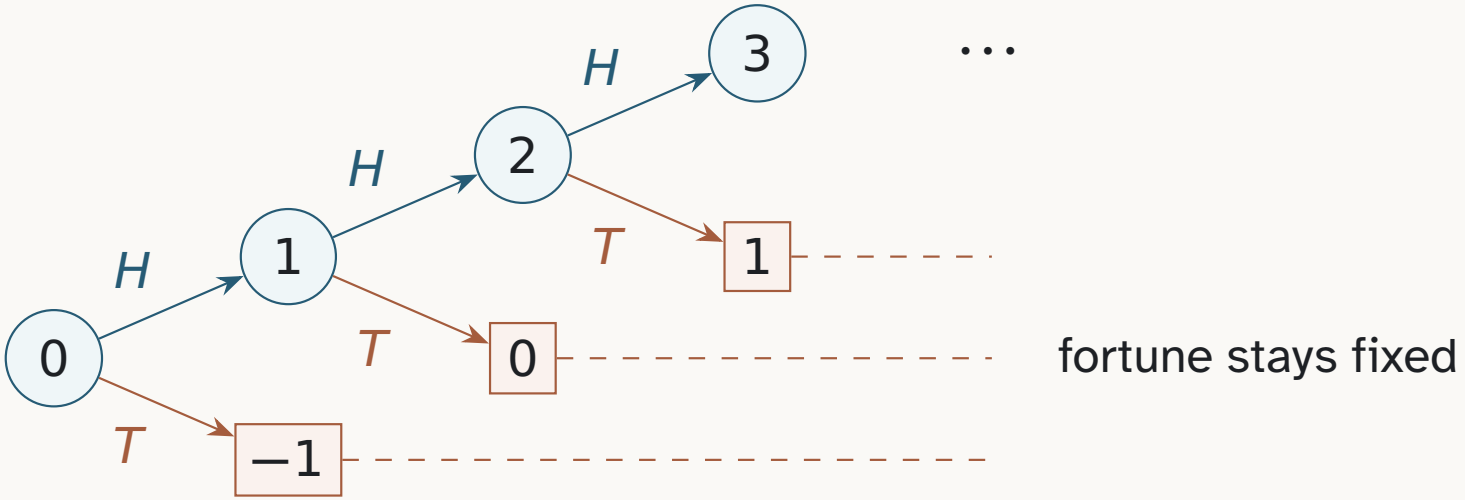
□

When can McDiarmid lose information?

McDiarmid bounds how much the *final answer* changes when one input is replaced. In a process that responds to its observations, an early input can change many later actions. Its effect on the final answer may be large, even when each step changes the current value by only one. We will compare the two arguments for fair unit bets that stop at the first loss.

Fair bets until the first loss

Start with fortune 0. Bet one unit on each fair coin flip, stopping after the first loss or after m rounds.



Presample independent flips $X_1, \ldots, X_m$ and write $F(X)$ for the final fortune. How sensitive is F to one flip?

Apply McDiarmid to the final fortune

If the first $i-1$ flips are heads, compare

Flips from round i	Final fortune
$H, H, \ldots, H$	m
$T, H, \ldots, H$	$i-2$

Thus the worst change at coordinate i is $c_i = m - i + 2$. Set

$$V = \sum_{i=1}^{m} (m - i + 2)^2 = \Theta(m^3).$$

Each round has expected gain zero: whether we bet depends only on earlier flips. Hence $\mathbb{E}F = 0$, and McDiarmid gives

$$\Pr(|F| \geq t) \leq 2e^{-2t^2/V}.$$

The denominator grows as m^3.

□

Apply Azuma to the evolving fortune

Let M_i be the fortune after round i, held constant after stopping. Given the previous flips,

$$D_i = \begin{cases} +1 \text{ or } -1, \text{ each with probability } 1/2, & \text{still betting,} \\ 0, & \text{already stopped.} \end{cases}$$

Thus $\mathbb{E}[D_i \mid \mathcal{F}_{i-1}] = 0$, so (M_i) is a martingale and $\mathbb{E}F = \mathbb{E}M_m = 0$. Its conditional interval widths are at most 2. Azuma gives

$$\boxed{\Pr(|F| \geq t) \leq 2e^{-t^2/(2m)}.}$$

□

Azuma gives a smaller denominator here. This comparison does not establish that its bound is sharp.

PART 4

Concentration of an optimal value

4

Fitting a model from data

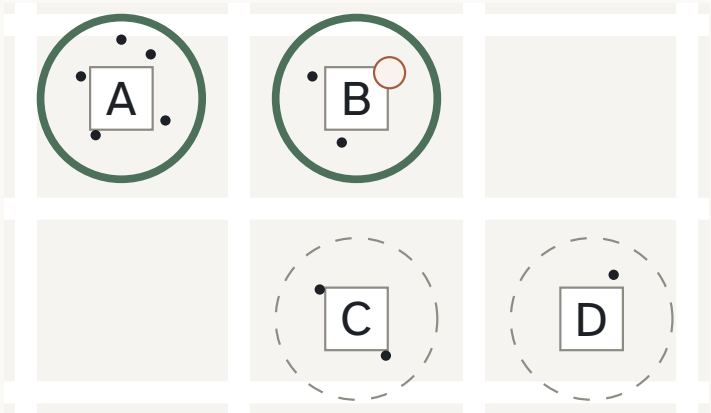
We often choose a model by optimizing its performance on observed data. A facility selection is one example: the selected sites form the model, and the covered fraction measures its performance. Two questions arise:

1. How much does the best training score change when we draw a different sample?
2. How well does the selected model perform on new data?

We first study the training optimum. The appendix takes up generalization.

Choosing facilities from observed demand

A city has a fixed list $\mathcal{L}$ of candidate sites and funding for k facilities. It observes s independent request locations, then selects sites to cover as many sampled requests as possible.



A request is covered within distance r of a selected site. Squares are sites; dots are requests; solid circles show the selected coverage regions.

Training performance of a facility selection

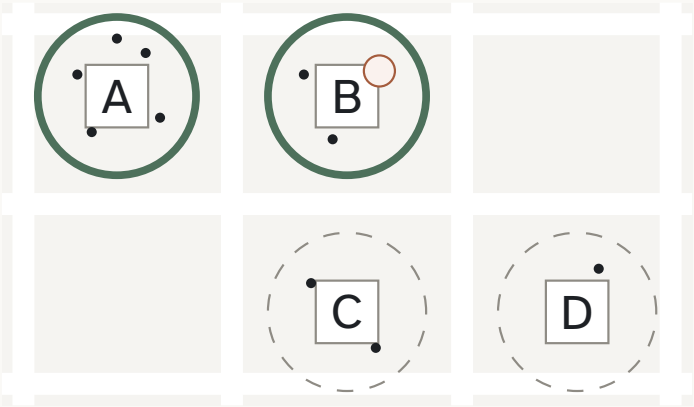
For the independent request locations $X_1, \ldots, X_s$, define

$$C_S(X) = \sum_{j=1}^{s} \mathbf{1}\{X_j \text{ is covered by } S\}, \qquad F(X) = \max_{\substack{S \subseteq \mathcal{L} \\ |S| \le k}} C_S(X).$$

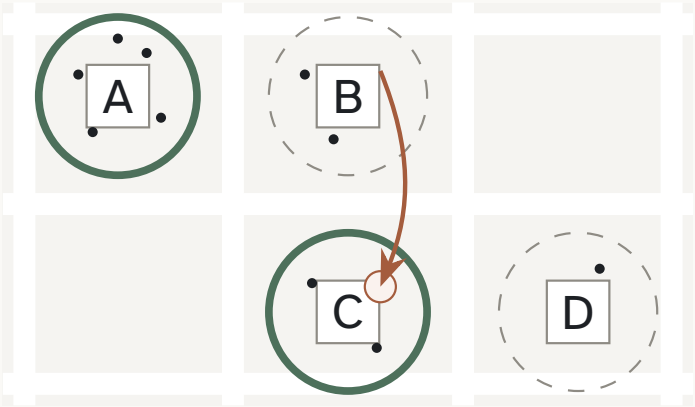
Model fitting	Facility placement
Model	Selected sites S
Training example	Request location X_j
Training score	Covered fraction $C_S(X)/s$
Best training score	$F(X)/s$

The selected sites depend on the sample. We want to bound fluctuations of the best achievable covered fraction.

Changing one request



original: A,B cover 8



reoptimized: A,C cover 7

With the facilities fixed, only that request can change its coverage status:

$$|C_S(x) - C_S(x')| \leq 1 \quad \text{for every selection } S.$$

Reoptimizing may choose different facilities: here A,C cover 8. Can the optimal count change by more than one? Squares: candidate sites. Circles: coverage regions. Dots: requests.

The optimal count changes by at most one

Let S^* be optimal for the original sample x. On the new sample x', the same selection loses at most one covered request:

$$\begin{aligned} F(x) &= C_{S^*}(x) \\ &\leq C_{S^*}(x') + 1 \\ &\leq F(x') + 1. \end{aligned}$$

Reverse the roles of x, x' to get

$$\boxed{|F(x) - F(x')| \leq 1.}$$

□

The same argument works for a maximum of any family of functions sharing a common sensitivity bound.

Concentration of optimal coverage

There are s independent requests, each with sensitivity 1. McDiarmid gives

$$\boxed{\Pr(|F - \mathbb{E}F| \geq t) \leq 2e^{-2t^2/s}.}$$

Equivalently, for the optimal covered fraction,

$$\Pr\left(\left|\frac{F}{s} - \mathbb{E}\frac{F}{s}\right| \geq \varepsilon\right) \leq 2e^{-2s\varepsilon^2}.$$

No union bound over facility selections is needed: we bounded the sensitivity of the optimized value itself. Across independent samples, the optimal covered fraction concentrates around its expectation, with deviations of order $s^{-1/2}$.

The same argument applies to fitting models

Let $\mathcal{H}$ be a model class fixed before observing the sample. Model h earns score $g_h(x) \in [0, 1]$ on example x. For s independent examples, define

$$\widehat{R}(h) = \frac{1}{s}\sum_{j=1}^{s} g_h(X_j), \qquad \widehat{R}^* = \sup_{h\in\mathcal{H}} \widehat{R}(h).$$

Replacing one example changes every model's score by at most $1/s$. The inequality

$$|\sup_h a_h - \sup_h b_h| \le \sup_h |a_h - b_h|$$

preserves this bound for the best training score. Hence

$$\boxed{\Pr(|\widehat{R}^* - \mathbb{E}\widehat{R}^*| \ge \varepsilon) \le 2e^{-2s\varepsilon^2}.}$$

□

The best training score concentrates even if the selected model changes. The bound has no model-class size term. Assume the supremum is measurable.

What to check in an application

- Identify the random inputs or evolving process.
- Choose what to reveal. The exposure process can change the strength of the bound.
- Bound conditional increments, or prove coordinate sensitivity and use McDiarmid.
- Compare the resulting deviation scale with the range of the answer.

For matching, changing the exposure process reduced the number of unit contributions from $\binom{n}{2}$ to $n - 1$.

Questions?

Appendix

1. Revealing several inputs together
   Why independent random vectors give the same martingale construction.

2. How many servers receive work?
   Independent assignment inputs, dependent occupancy indicators, and concentration around an explicit expectation.

3. Training performance and generalization
   What changes when we ask how a model selected from data performs on new examples?

Revealing several inputs together

Let $B_1, \ldots, B_r$ be independent random vectors, and let $Z = f(B_1, \ldots, B_r)$ be integrable. Suppose replacing B_j, with all other vectors fixed, changes f by at most a deterministic c_j. For $\sum_j c_j^2 > 0$,

$$\boxed{\Pr(|Z - \mathbb{E}Z| \geq t) \leq 2 \exp\left(-\frac{2t^2}{\sum_j c_j^2}\right), \quad t > 0.}$$

Use the filtration $\mathcal{F}_j = \sigma(B_1, \ldots, B_j)$ and the martingale $M_j = \mathbb{E}[Z \mid \mathcal{F}_j]$.

A vector observation reveals several coordinates at once; Azuma uses only the filtration and conditional interval widths.

Why vector observations work

Let $B_1, \dots, B_r$ be independent random vectors, and let $Z = f(B_1, \dots, B_r)$ be integrable. Set

$$\mathcal{F}_j = \sigma(B_1, \dots, B_j), \qquad M_j = \mathbb{E}[Z \mid \mathcal{F}_j].$$

The tower property gives $\mathbb{E}[M_j \mid \mathcal{F}_{j-1}] = M_{j-1}$. Also $M_0 = \mathbb{E}Z$ and $M_r = Z$. Suppose replacing vector B_j, with all others fixed, changes f by at most c_j. Fix the previous vectors $b_{<j}$ and compare any two next vectors a, b:

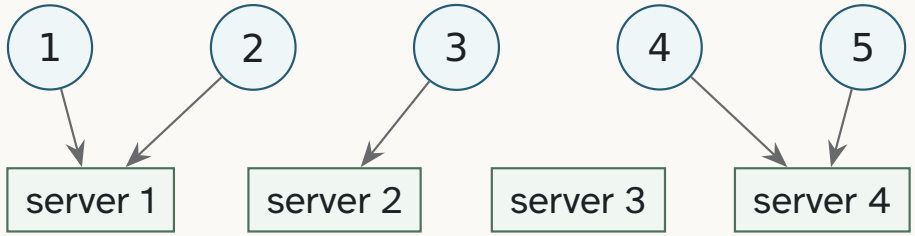
$$\begin{aligned} &|\mathbb{E}f(b_{<j}, a, B_{>j}) - \mathbb{E}f(b_{<j}, b, B_{>j})| \\ &\qquad \le \mathbb{E}|f(b_{<j}, a, B_{>j}) - f(b_{<j}, b, B_{>j})| \le c_j. \end{aligned}$$

The expectations use the same remaining-vector distribution. Thus the conditional width of $M_j - M_{j-1}$ is at most c_j. Apply Azuma.

□

How many servers receive work?

A batch of m jobs is assigned independently and uniformly to b servers. How many servers receive at least one job?



Let Z be the number of occupied servers. By linearity of expectation,

$$\mathbb{E}Z = \sum_{\ell=1}^{b} \Pr(\text{server } \ell \text{ receives a job}) = b\left[1 - \left(1 - \frac{1}{b}\right)^{m}\right].$$

Occupancy indicators are dependent; job destinations are independent.

Changing one job’s destination

Use the independent job destinations as the inputs to McDiarmid. Moving one job can empty its old server and occupy its new server. These changes have opposite signs: the net change in Z is at most 1. Thus

$$\boxed{\Pr\left(\left|Z - b\left[1 - \left(1 - \frac{1}{b}\right)^{m}\right]\right| \geq t\right) \leq 2e^{-2t^2/m}.}$$

□ We obtained concentration around an explicit expectation by exposing assignments rather than occupied-server indicators.

Training performance and generalization

Now assume $X_1, \ldots, X_s$ are iid, and X_{new} is an independent example from the same distribution. For model h, define

$$\widehat{R}(h) = \frac{1}{s} \sum_{j=1}^{s} g_h(X_j), \qquad R(h) = \mathbb{E}[\, g_h(X_{\text{new}})].$$

Let $\mathcal{H}$ be fixed before sampling. Assume a maximizer exists and choose $\widehat{h}$ to maximize the training score.

Main lecture	New question
How much does $\widehat{R}^*$ vary across samples?	How close is $R(\widehat{h})$ to $\widehat{R}(\widehat{h})$?

For facilities, $R(S)$ is the probability that a new request is covered by the selected sites S.

Perfect training coverage can be misleading

Suppose requests are uniform on N separated locations. Each location has a candidate facility covering only that location. In this example the budget is s facilities, one for each sampled request. Select a facility at each distinct sampled location. Every sample of s requests is covered completely:

$$\widehat{R}^* = 1 \quad \text{for every sample.}$$

But any selection of at most s sites covers a new request with probability at most

$$R(\widehat{S}) \le s/N.$$

When $N \gg s$, training coverage is perfect and coverage of new requests is poor. Concentration of the best training score does not compare training and population performance.

Control every model at once

For a fixed model h, Hoeffding gives

$$\Pr(|\widehat{R}(h) - R(h)| > \varepsilon) \leq 2e^{-2s\varepsilon^2}.$$

The selected model $\widehat{h}$ depends on the sample. It cannot be substituted for a fixed h in this argument. Instead, bound the **uniform discrepancy**

$$U(X) = \sup_{h \in \mathcal{H}} |\widehat{R}(h) - R(h)|.$$

For every sample, this controls its selected model too:

$$|\widehat{R}(\widehat{h}) - R(\widehat{h})| \leq U(X).$$

A finite class: Hoeffding and a union bound

Suppose there are H candidate models. A union bound gives

$$\Pr(U > \varepsilon) \le 2He^{-2s\varepsilon^2}.$$

With probability at least $1 - \delta$, simultaneously for every model,

$$|\widehat{R}(h) - R(h)| \le \sqrt{\frac{\log(2H/\delta)}{2s}}.$$

In particular,

$$\boxed{R(\widehat{h}) \ge \widehat{R}(\widehat{h}) - \sqrt{\frac{\log(2H/\delta)}{2s}}.}$$

□ The class size enters because we compare training and population scores for all candidates, including the one selected from the data.

Apply the bound to facility selections

With L candidate sites and a budget of k, the number of feasible selections is

$$H = \sum_{j=0}^{\min(k,L)} \binom{L}{j}.$$

For $S \subseteq \mathcal{L}$, let $q(S)$ be its probability of covering a new request and $\widehat{q}(S) = C_S(X)/s$ its sampled covered fraction. For the sample-selected sites $\widehat{S}$, with probability at least $1 - \delta$,

$$\boxed{q(\widehat{S}) \geq \widehat{q}(\widehat{S}) - \sqrt{\frac{\log(2H/\delta)}{2s}}.}$$

This guarantees coverage on new demand from the same distribution. The main lecture instead bounded fluctuations of the best sample coverage around its expectation.

McDiarmid also controls uniform discrepancy

Recall $U(X) = \sup_h |\widehat{R}(h) - R(h)|$. For samples x, x' differing in one example, the population scores stay fixed, so

$$|U(x) - U(x')| \le \sup_h |\widehat{R}_x(h) - \widehat{R}_{x'}(h)| \le 1/s.$$

McDiarmid therefore gives the one-sided bound

$$\boxed{\Pr(U \ge \mathbb{E}U + \varepsilon) \le e^{-2s\varepsilon^2}.}$$

□ To conclude that generalization error is small, we must also show that $\mathbb{E}U$ is small. That requires information about the model class, beyond bounded scores. For a finite class, the preceding union bound already controls U directly.

Two uses of bounded differences

Best training score $\widehat{R}^*$	Uniform discrepancy U
Measures the achievable optimum on the sample. Replacing one example changes it by at most $1/s$. McDiarmid controls variation around $\mathbb{E}\widehat{R}^*$.	Compares training and population scores for all models. Replacing one example changes it by at most $1/s$. McDiarmid controls variation around $\mathbb{E}U$.

A reproducible training optimum can still coexist with poor generalization.

Generalization needs control of the discrepancy itself, including its expectation. Assume the suprema are measurable; a finite model class suffices.