1 Some definitions

Recall that a *martingale* is a sequence of r.v.s $Z_0, Z_1, Z_2, \ldots$ (denoted by (Z_i)) if each Z_i satisfies $E[|Z_i|] < \infty$, and

$$E[Z_i \mid Z_0, ..., Z_{i-1}] = Z_{i-1}.$$

Somewhat more generally, given a sequence (X_i) of random variables, a *martingale with respect to* (X_i) is another sequence of r.v.s $Z_0, Z_1, Z_2, \ldots$ (denoted by (Z_i)) if each Z_i satisfies

- $E[|Z_i|] < \infty$,
- there exists functions g_i such that $Z_i = g_i(X_1, X_2, \ldots, X_i)$, and
- $E[Z_i \mid X_1, \ldots, X_{i-1}] = Z_{i-1}$.

One can define things even more generally, but for the purposes of this course, let's just proceed with this. If you'd like more details, check out, say, books by Grimmett and Stirzaker, or Durett, or many others.)

1.1 The Azuma-Hoeffding Inequality

Theorem 1 (Azuma-Hoeffding) *If (Z_i) is a martingale such that for each i, $|Z_i - Z_{i-1}| < c_i$. Then*

$$\Pr[|Z_n - Z_0| \geq \lambda] \leq 2 \exp\left\{ -\frac{\lambda^2}{2\sum_i c_i^2} \right\}.$$

(Apparently Bernstein had essentially figured this one out as well, in addition to the Chernoff-Hoeffding bounds, back in 1937.) The proof of this bound can be found in most texts, we'll skip it here. BTW, if you just want the upper or lower tail, replace $2e^{blah}$ by e^{blah} on the right hand side.

2 The Doob Martingale

Most often, the case we will be concerned with is where our entire space is defined by a sequence of random variables $X_1, X_2, \ldots, X_n$, where each X_i takes values in the set Ω. Moreover, we will be interested in some *bounded* function $f : \Omega^n \to \mathbb{R}$, and will want to understand how $f(X_1, X_2, \ldots, X_n)$ behaves, when (X_i) is drawn from the underlying distribution. (Very often these X_i's will be drawn from a "product distribution"—i.e., they will be independent of each other, but they need not be.) Specifically, we ask:

How concentrated is f around its mean $E[f] := E_{X_1,X_2,\ldots,X_n}[f(X_1, X_2, \ldots, X_n)]$?

To this end, define for every $i \in \{0, 1, \ldots, n\}$, the random variable

$$Z_i := E[f(X) \mid X_1, X_2, \ldots, X_i].$$

(At this point, it is useful to remember the definition of a random variable as a function from the sample space to the reals: so this r.v. Z_i is also such a function, obtained by taking averages of f over parts of the sample space.)

How does the random variable Z_0 behave? It's just the constant $E[f]$: the expected value of the function f given random settings for X_1 through X_n. What about Z_1? It is a function that depends only on its first variable, namely $Z_1(x_1) = E_{X_2,\ldots,X_n}[f(x_1, X_2, \ldots, X_n)]$—instead of averaging f over the entire sample space, we partition Ω according to value of the first variable, and average over each part in the partition. And Z_2 is a function of x_1, x_2, averages over the other variables. And so on to Z_n, which is the same as the function f. So, as we go from 0 to n, the random variables Z_i go from the constant function $E[f]$ to the function f.

Picture here

Of course, we're defining this for a reason: (Z_i) is a martingale with respect to (X_i).

Lemma 2 *For a bounded function f, the sequence $(Z_i)_{i=0}^n$ is a martingale with respect to (X_i). (It's called the Doob martingale for f.)*

PROOF: The first two properties of (Z_i) being a martingale with respect to (X_i) follow from f being bounded, and the definition of Z_i itself. For the last property,

$$\begin{aligned} E[Z_i \mid X_1, \ldots X_{i-1}] &= E[\,E[f \mid X_1, X_2, \ldots, X_i] \mid X_1, \ldots X_{i-1}] \\ &= E[f \mid X_1, \ldots X_{i-1}] = Z_{i-1}. \end{aligned}$$

The first equaility is the definition of Z_i, the second from the fact that $E[U \mid V] = E[E[U \mid V, W] \mid V]$ for random variables U, V, W, and the last from the definition of Z_{i-1}. $\square$

Assuming that f was bounded was not necessary, one can work with weaker assumptions—see the texts for more details.

Before we continue on this thread, let us show some Doob martingales which arise in CS/Math-y applications.

1. Throw m balls into n bins, and let f be some function of the load: the number of empty bins, the max load, the second-highly loaded bin, or some similar function. Let $\Omega = [n]$, and X_i be the index of the bin into which ball i lands. For $Z_i = E[f \mid X_1, \ldots, X_i]$, (Z_i) is a martingale with respect to (X_i).

2. Consider the random graph $G_{n,p}$: n vertices, each of the $\binom{n}{2}$ edges chosen independently with probability p. Let χ be the chromatic number of the graph, the minimum number of colors to properly color the graph. There are two natural Doob martingales associated with this, depending on how we choose the variables X_i.

   In the first one, let X_i be the i^{th} edge, and which gives us a martingle sequence of length $\binom{n}{2}$. This is called the *edge-exposure martingale*. For the second one, let X_i be the collection of edges going from the vertex i to vertices $1, 2, \ldots, i-1$: the new martingale has length n and is called the *vertex exposure* martingale.

3. Consider a run of quicksort on a particular input: let Q be the number of comparisons. Let X_1 be the first pivot, X_2 the second, etc. Then $Z_i = E[Q \mid X_1, \ldots, X_i]$ is a Doob martingale with respect to (X_i).

   BTW, are these X_i's independent of each other? Naively, they might depend on the size of the current set, which makes it dependent on the past. One way you can make these independent is by letting these X_i's be, say, random independent permutations on all n elements, and when you want to choose the i^{th} pivot, pick the first element from the current set according to the permutation X_i. (Or, you could let X_i be a random independent real in $[0, 1]$ and use that to pick a random element from the current set, etc.)

4. Suppose we have r red and b blue balls in a bin. We draw n balls *without replacement* from this bin: what is the number of red balls drawn? Let X_i be the indicator for whether the i^{th} ball is red, and let $f = X_1 + X_2 + \ldots + X_n$ is the number of red balls. Then $Z_i = E[f \mid X_1, \ldots, X_i]$ is a martingale with respect to (X_i).

   However, in this example, the X_i's are not independent. Nonetheless, the sequence is a Doob martingale. (As in the quicksort example, one can define it with respect to a different set of variables which are independent of each other.)

So yeah, if we want to study the concentration of f around $E[f]$, we can now apply Azuma-Hoeffding to the Doob martingale, which gives us the concentration of Z_n (i.e., f) around Z_0 (i.e., $E[f]$). Good, good.

Next step: to apply Azuma-Hoeffding to the Doob martingale (Z_i), we need to bound $|Z_i - Z_{i-1}|$ for all i. Which just says that if we can go from f to Ef in a "small" number of steps (n), and each time we're not smoothing out "too agressively" ($|Z_i - Z_{i-1}| \leq c_i$), then f is concentrated about its mean.

2.1 Indepedence and Lipschitz-ness

One case when it's easy to bound the $|Z_i - Z_{i-1}|$'s is when the X_i's are independent of each other, and also f is not too sensitive in any coordinate—namely, changing any coordinate does not change the value of f by much. Let's see this in detail.

Definition 3 *Given values $(c_i)_{i=1}^n$, the function f is $\underline{(c_i)\text{-Lipschitz}}$ if for all j and $x_j \in \Omega$, for all $i \in [n]$ and for all $x_i' \in \Omega$, it holds that*

$$|f(x_1, \ldots, x_{i-1}, x_i, x_{i+1}, \ldots, x_n) - f(x_1, \ldots, x_{i-1}, x_i', x_{i+1}, \ldots, x_n)| \leq c_i.$$

If $c_i = c$ for all i, then we just say f is c-Lipschitz.

Lemma 4 *If f is (c_i)-Lipschitz and $X_1, X_2, \ldots, X_n$ are independent, then the Doob martingale of f with respect to (X_i) satisfies*

$$|Z_i - Z_{i-1}| \leq c_i.$$

Proof: Let us use $X_{(i:j)}$ to denote the sequence $X_i, \ldots, X_j$, etc. Recall that

$$Z_i = E[f \mid X_{(1:i)}] = \sum_{a_{i+1}, \ldots, a_n} f(X_{(1:i)}, a_{(i+1:n)}) \Pr[X_{(i+1:n)} = a_{(i+1:n)} \mid X_{(1:i)}]$$

$$= \sum_{a_{i+1}, \ldots, a_n} f(X_{(1:i)}, a_{(i+1:n)}) \Pr[X_{(i+1:n)} = a_{(i+1:n)}]$$

where the last equality is from independence. Similarly for Z_{i-1}. Hence

$$
\begin{aligned}
|Z_i - Z_{i-1}| &= \sum_{a_{i+1},\dots,a_n} \left| f(X_{(1:i)}, a_{(i+1:n)}) - \sum_{a_i'} \Pr[X_i = a_i'] f(X_{(1:i-1)}, a_i', a_{(i+1:n)}) \right| \cdot \Pr[X_{(i+1:n)} = a_{(i+1:n)}] \\
&\leq \sum_{a_{i+1},\dots,a_n} c_i \cdot \Pr[X_{(i+1:n)} = a_{(i+1:n)}] = c_i.
\end{aligned}
$$

where the inequality is from the fact that changing the i^{th} coordinate from a_i to a_i' cannot change the function value by more than c_i, and that $\sum_{a_i'} \Pr[X_i = a_i'] = 1$. □

Now applying Azuma-Hoeffding, we immediately get:

Corollary 5 (McDiarmid's Inequality) *If f_i is c_i-Lipschitz for each i, and $X_1, X_2, \dots, X_n$ are independent, then*

$$
\Pr[f - E[f] \geq \lambda] \leq \exp\left(-\frac{\lambda^2}{2\sum_i c_i^2} \right),
$$
$$
\Pr[f - E[f] < \lambda] \leq \exp\left(-\frac{\lambda^2}{2\sum_i c_i^2} \right).
$$

(Disclosure: I am cheating. McDiarmid's inequality has better constants, the constant 2 in the denominator moves to the numerator.) And armed with this inequality, we can give concentration results for some applications we mentioned above.

1. For the m balls and n bins example, say f is the number of empty bins: hence $Ef = n(1 - 1/n)^m \approx n\, e^{-m/n}$. Also, changing the location of the i^{th} ball changes f by at most 1. So f is 1-Lipschitz, and hence

   $$\Pr[|f - Ef| \geq \lambda] \leq 2 \exp\left(-\frac{\lambda^2}{2m} \right).$$

   Hence, whp, $f \approx n\, e^{-m/n} \pm O(\sqrt{m \log n})$.

2. For the case where χ is the chromatic number of a random graph $G_{n,p}$, and we define the edge-exposure martingale $Z_i = E[\chi \mid E_1, E_2, \dots, E_i]$, clearly χ is 1-Lipschitz. Hence

   $$\Pr[|\chi - E\chi| \geq \lambda] \leq 2 \exp\left(-\frac{\lambda^2}{2\binom{n}{2}} \right)$$

   This is not very interesting, since the right hand side is < 1 only when $\lambda \approx n$—but the chromatic number itself lies in $[1, n]$, so we get almost no concentration at all.

   Instead, we could use a vertex-exposure martingale, where at the i^{th} step we expose the vertex i and its edges going to vertices $1, 2, \dots, i-1$. Even with respect to these variables, the function χ is 1-Lipschitz, and hence

   $$\Pr[|\chi - E\chi| \geq \lambda] \leq 2 \exp\left(-\frac{\lambda^2}{2n} \right)$$

   And hence the chromatic number of the random graph $G_{n,p}$ is concentrated to within $\approx \sqrt{n}$ around its mean.

3 Concentration for Random Geometric TSP

McDiarmid's inequality is convenient to use, but Lipschitz-ness often does not get us as far as we'd like (even with independence). Sometimes you need to bound $|Z_i - Z_{i-1}|$ directly to get the full power of Azuma-Hoeffding. Here's one example:

> *Let $X_1, X_2, \ldots, X_n$ be n points picked independently and uniformly at random from the unit square $[0,1]^2$. Let $\tau : ([0,1]^2)^n \to \mathbb{R}$ be the length of the shortest traveling salesman tour on n points. How closely is τ concentrated around its mean $E[\tau(X_1, X_2, \ldots, X_n)]$?*

In the HW, you will show that $E\tau = \Theta(n^{1/2})$; in fact, one can pin down $E\tau$ up to the leading constant. (See the work of Rhee and others.)

3.1 Using McDiarmid: a weak first bound

Note that τ is $2\sqrt{2}$-Lipschitz. By Corollary 5 we get that

$$\Pr[|\tau - E\tau| \geq \lambda] \leq 2\exp(-\frac{\lambda^2}{16n}).$$

If we want the deviation probability to be $1/poly(n)$, we would have to set $\lambda = \Omega(\sqrt{n \log n})$. Not so great, since this is pretty large compared to the expectation itself—we'd like a tighter bound.

3.2 So let's be more careful: an improved bound

And in fact, we'll get a better bound using the very same Doob martingale (Z_i) associated with τ:

$$Z_i = E[\tau(X_1, X_2, \ldots, X_n) \mid X_1, X_2, \ldots, X_i].$$

But instead of just using the $O(1)$-Lipschitzness of τ, let us bound $|Z_i - Z_{i-1}|$ better.

Lemma 6

$$|Z_i - Z_{i-1}| \leq \min\left\{2\sqrt{2}, \frac{O(1)}{\sqrt{n-i}}\right\}.$$

Before we prove this lemma, let us complete the concentration bound for TSP using this. Setting $c_i = O(1/\sqrt{n-i})$ gives us $\sum_i c_i^2 = O(\log n)$, and hence Azuma-Hoeffding gives:

$$\Pr[|\tau - E\tau| \geq \lambda] \leq 2\exp\left(-\frac{\lambda^2}{2\sum_i c_i^2}\right) \leq 2\exp\left(-\frac{\lambda^2}{O(\log n)}\right).$$

So

$$\Pr[|\tau - E\tau| \leq O(\log n)] \geq 1 - 1/poly(n).$$

Much better!

3.3 Some useful lemmas

To prove Lemma 6, we'll need a simple geometric lemma:

Lemma 7 *Let $x \in [0,1]^2$. Pick k random points A from $[0,1]^2$, the expected distance of point x to its closest point in A is $O(1/\sqrt{k})$.*

PROOF: Define the random variable $W = d(x, A)$. Hence, $W \geq r$ exactly when $B(x,r) \cap A = \emptyset$. For $r \in [0, \sqrt{2}]$, the area of $B(x,r) \cap [0,1]^2$ is at least $c_0 r^2$ for some constant c_0.

Define $r_0 = \sqrt{c_0/k}$. For some $r = \lambda r_0 \in [0, \sqrt{2}]$, the chance that k points all miss this ball, and hence $\Pr[W \geq r = \lambda r_0]$ is at most

$$(1 - c_0 r^2)^k = (1 - \lambda^2/k)^k \leq e^{-\lambda^2}.$$

Of course, for $r > \sqrt{2}$, $\Pr[W \geq r] = 0$.

And hence

$$E[W] = \int_{r \geq 0} \Pr[W \geq r] dr = \sum_{\lambda \in \mathbb{Z}_{\geq 0}} \int_{r \in [\lambda r_0, (\lambda+1) r_0]} \Pr[W \geq r] dr \leq \sum_{\lambda \in \mathbb{Z}_{\geq 0}} (\lambda + 1) r_0 \cdot e^{-\lambda^2} \leq O(r_0).$$

□

Secondly, here is another lemma about how the TSP behaves:

Lemma 8 *For any set of $n-1$ points, $A = \{x_1, x_2, \ldots, x_{i-1}, x_{i+1}, \ldots, x_n\}$, we get*

$$|\tau(A + x_i) - \tau(A + x_i')| \leq 2(d(x_i, A) + d(x_i', A)).$$

PROOF: Follows from the fact that $\tau(A + x) \in [TSP(A), TSP(A) + 2d(x, A)]$, for any x. □

3.4 Proving Lemma 6

OK, now to the proof of Lemma 6. Recall that we want to bound $|Z_i - Z_{i-1}|$; since τ is $2\sqrt{2}$-Lipschitz, we get $|Z_i - Z_{i-1}| \leq 2\sqrt{2}$ immediately. For the second bound of $O(1/\sqrt{k-i})$, note that

$$\begin{aligned} Z_{i-1} &= E[\tau(X_1, X_2, \ldots, X_{i-1}, X_i, \ldots, X_n) \mid X_1, X_2, \ldots, X_{i-1}] \\ &= E[\tau(X_1, X_2, \ldots, X_{i-1}, \widehat{X}_i, \ldots, X_n) \mid X_1, X_2, \ldots, X_{i-1}] \\ &= E[\tau(X_1, X_2, \ldots, X_{i-1}, \widehat{X}_i, \ldots, X_n) \mid X_1, X_2, \ldots, X_i] \end{aligned}$$

where $\widehat{X}_i$ is a independent copy of the random variable X_i. Hence

$$|Z_i - Z_{i-1}| = E[\tau(X_1, \ldots, X_{i-1}, X_i, \ldots, X_n) - \tau(X_1, \ldots, X_{i-1}, \widehat{X}_i, \ldots, X_n) \mid X_1, X_2, \ldots, X_i].$$

Then, if we define the set $S = X_1, X_2, \ldots, X_{i-1}$ and $T = X_{i+1}, \ldots, X_n$, then we get

$$\begin{aligned} |Z_i - Z_{i-1}| &= E[TSP(S \cup T \cup \{X_i\}) - TSP(S \cup T \cup \{\widehat{X}_i\}) \mid X_1, X_2, \ldots, X_i] \\ &\leq E[2(d(X_i, S \cup T) + d(\widehat{X}_i, S \cup T)) \mid X_1, X_2, \ldots, X_i] \\ &\leq E_{\widehat{X}_i, T}[2(d(X_i, T) + d(\widehat{X}_i, T)) \mid X_i]. \end{aligned}$$

where the first inequality uses Lemma 8 and the second uses the fact that the minimum distance to a set only increses when the set gets smaller. But now we can invoke Lemma 7 to bound each of the terms by $O(1)/\sqrt{|T|} = O(1)/\sqrt{n-i}$. This completes the proof of Lemma 6.

3.5 Some more about Geometric TSP

For constant dimension $d > 2$, one can consider the same problems in $[0,1]^d$: the expected TSP length is now $\Theta(n^{1-1/d})$, and using similar arguments, you can show that devations of $\Omega(tn^{1/2-1/d})$ have probability $\leq e^{-t^2}$.

The result we just proved was by Rhee and Talagrand, but it was not the last result about TSP concentration. Rhee and Talagrand subsequently improved this bound to the TSP has subgaussian tails!

$$\Pr[|\tau - E\tau| \geq \lambda] \leq ce^{-\lambda^2/O(1)}.$$

We'll show a proof of this using Talagrand's inequality, in a later lecture.

If you're interested in this line of research, here is a survey article by Michael Steele on concentration properties of optimization problems in Euclidean space, and another one by Alan Frieze and Joe Yukich on many aspects of probabilistic TSP.

4 Citations

As mentioned in a previous post, McDiarmid and Hayward use martingales to give extremely strong concentration results for QuickSort . The book by Dubhashi and Panconesi (preliminary version here) sketches this result, and also contains many other examples and extensions of the use of martingales.

Other resources for concentration using martingales: this survey by Colin McDiarmid, or this article by Fan Chung and Linyuan Lu.

Apart from giving us powerful concentration results, martingales and "stopping times" combine to give very surprising and powerful results: see this survey by Yuval Peres at SODA 2010, or these course notes by Yuval and Eyal Lubetzky.