Lecturer: Goran Zuzic Scribe: Ziye Tang and Taisuke Yasuda

In this lecture, we will study two problems using concentration inequalities: set balancing and random walks. Along the way, we will build intuition on concentration inequalities and see new powerful tools for studying these problems by introducing martingales.

1 Review of Concentration Inequalities

We recall two basic inequalities that we will refer to extensively throughout these notes: the union bound and the Chernoff bound. The union bound states that for two events A and B, $\mathbf{Pr}(A \cup B) \leq \mathbf{Pr}(A) + \mathbf{Pr}(B)$. The typical way we use the union bound is to upper bound the event that any one of many bad events happen. The Chernoff bound states the following: if $X_1, X_2, \ldots, X_n \in [0,1]$ are IRVs (independent random variables) and $\mu \coloneqq \mathbf{E}(\sum_{i=1}^n X_i)$ is the expectation of their sum, then the sum cannot deviate too much from its expectation:

$$\mathbf{Pr}\left(\left|\sum_{i=1}^n X_i - \mu\right| \geq \lambda\right) \leq 2\exp\left(\frac{-c\lambda^2}{\mu+\lambda}\right)$$

where c is some positive constant.

Example 14.1. Let $X_1, \ldots, X_n \sim \text{Rad}$ be i.i.d. Rademacher variables, i.e. they take on values ± 1 with probability $1/2$ each. Then,

$$\mathbf{Pr}\left(\left|\sum_{i=1}^n X_i\right| \geq \lambda\right) \leq 2\exp\left(\frac{-c'\lambda^2}{n+\lambda}\right).$$

Proof. Let $Y_i = (X_i + 1)/2$, $S_n = \sum_{i=1}^n Y_i$. By applying Chernoff to $\{Y_i\}_{i=1}^n$, we have

$$\mathbf{Pr}(|S_n - \mathbf{E}(S_n)| \geq \lambda) \leq 2\exp\left(\frac{-c\lambda^2}{n/2+\lambda}\right).$$

Notice $S_n - \mathbf{E}(S_n) = \sum_{i=1}^n X_i/2$ which concludes the proof. □

2 Set Balancing Problem

Now we consider the set balancing problem as a classical application of Chernoff and union bound. One motivation for the set balancing problem comes from the experiment design where a group of subjects are divided into a treatment group and a control group. Each subject has several features. The goal is to find a division such that each feature has roughly the same number of subjects from the treatment and the control group. This is modeled as follows:

Set Balancing Problem: Given n items $V = \{v_i\}_{i=1}^n$ and n sets $\{S_i\}_{i=1}^n$ where $S_i \subset V$. Assign ± 1 to each item in V. Let M be the maximum absolute sum over all sets. The goal is to minimize M.

The absolute sum of a set S_i is called the *discrepancy* of S_i which we denote by $\text{disc}(S_i)$. If we take a probabilistic view, then it suffices to display a distribution of assignment with small $\mathbf{E}(\text{disc}(S_i))$ and bound the concentration around it together with a union bound. Luckily a uniform random assignment achieves a zero expectation. Therefore we have the following:

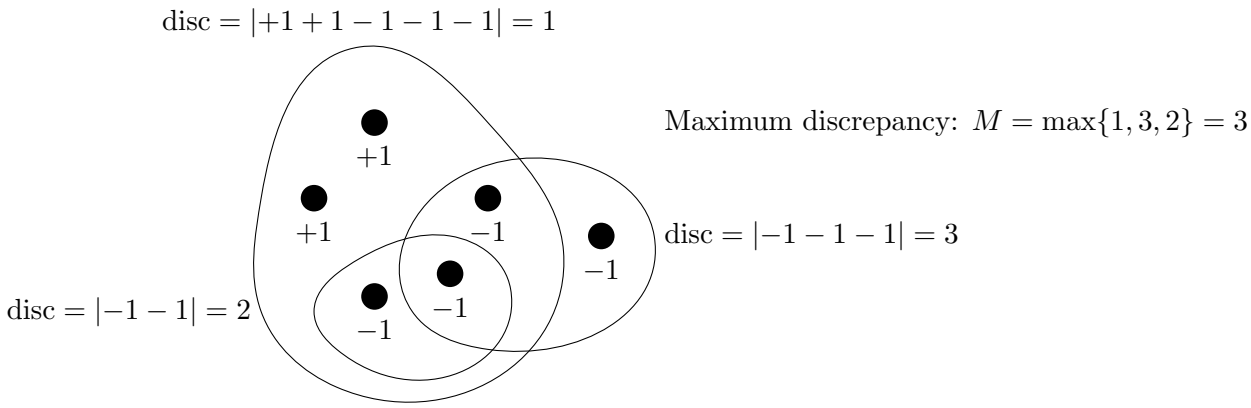

Figure 14.1: Set balancing problem

Theorem 14.2. *Uniform random assignment gives that for any set S, $\text{disc}(S) \le O(\sqrt{n \log n})$ with high probability.*

Proof. Let item v_i take on value $X_i \sim \text{Rad}$. Then $\text{disc}(S) = \sum_{i=1}^{n} \mathbb{1}\{X_i \in S\} \cdot X_i$, where $\mathbb{1}\{X_i \in S\}$ is the indicator for $X_i \in S$. By Chernoff,

$$\mathbf{Pr}\left(|\text{disc}(S)| \ge C\sqrt{n \log n}\right) \le 2 \exp\left(-c \frac{C^2 n \log n}{n + C\sqrt{n \log n}}\right) \le n^{-C'}.$$

Up to a constant we can ignore the $\sqrt{n \log n}$ on denominator, and choose the constant C to be big enough, the above is at most $n^{-C'}$. This completes the analysis for one group. Finally, we just union bound over n groups, namely, $\mathbf{Pr}[\max_i \text{disc}\, S_i \ge C\sqrt{n \log n}] \le n \cdot n^{-C'} \le n^{-C'+1} < 1$.

□

2.1 History

With the algorithm described above, you cannot do better than a discrepancy of $O(\sqrt{n \log n})$ (section 3). However, one can use more sophisticated constructions to obtain $O(\sqrt{n})$. This was first done by Spencer in 1985 in his famous six standard deviations paper [Spe85] (paper). This gets us $\text{disc}(S) \le 6\sqrt{n}$, but it is highly nonconstructive, and it was an open problem for a long time how actually to solve this problem constructively with $O(\sqrt{n})$ discrepancy. This was finally resolved in 2010 by Bansal [Ban10] (paper).

3 From Tail Bounds to Expectation

Now we take a different viewpoint to understand where $O(\sqrt{n \log n})$ comes from in the set balancing problem and show that the bound is essentially tight for uniform random assignment.

Recall $M = \max_i |\text{disc}(S_i)|$. We compute $\mathbf{E}(M)$ using tail bounds from section 2. To do so, we need the following lemma which can be shown by switching the order of integration (Fubini's theorem).

Lemma 14.3. *For a non-negative random variable $X \ge 0$, we have*

$$\mathbf{E}(X) = \int_0^\infty \mathbf{Pr}(X \ge \lambda)\, d\lambda.$$

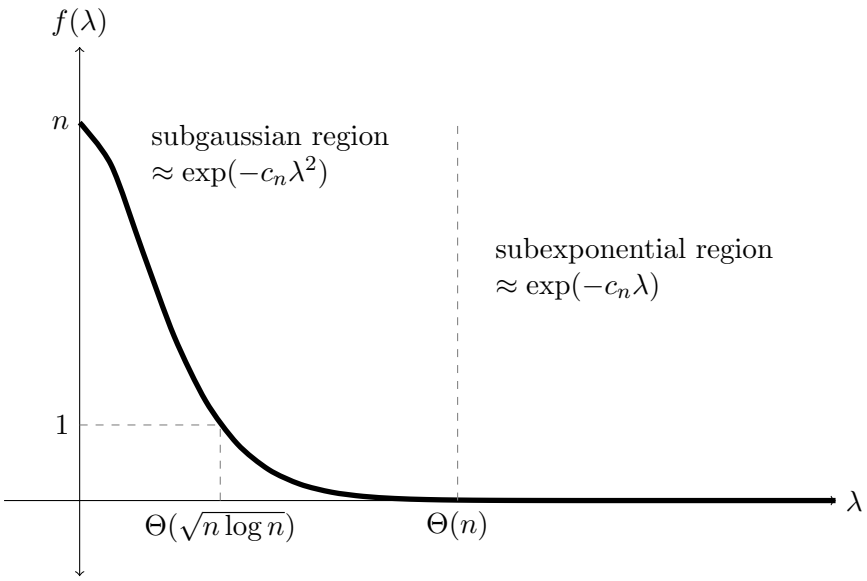

Figure 14.2: Different behaviors of the Chernoff bound

In section 2 we showed

$$\mathbf{Pr}(M \geq \lambda) \leq 2n \exp\left(-c\frac{\lambda^2}{n+\lambda}\right).$$

Let $f(\lambda) = 2n\exp(-c\frac{\lambda^2}{n+\lambda})$. We make several observations about $f(\lambda)$: it is a decreasing function. When $\lambda = \Theta(\sqrt{n\log n})$, $f(\lambda) \approx 1$. Moreover there is a phase transition at $\lambda = n$: when $\lambda \leq n$, it behaves like a subgaussian; when $\lambda \geq n$ it behaves like a subexponential. When $\lambda = \Theta(\sqrt{n\log n})$, we are in the subguassian region.

To calculate $\mathbf{E}(M)$, notice the tail bound $f(\lambda)$ is meaningless when $\lambda \leq \Theta(\sqrt{n\log n})$ and therefore we set it to be 1. Applying lemma 14.3, we have

$$\begin{aligned}\mathbf{E}(M) &= \int_0^\infty f(\lambda) \leq \int_0^{c\sqrt{n\log n}} 1 + \int_{c\sqrt{n\log n}}^\infty O\left((\frac{\lambda}{\sqrt{n}})^{-2}\right) \\ &= O\left(\sqrt{n\log n}\right) + \sqrt{n}\cdot O(1) = O(\sqrt{n\log n}).\end{aligned}$$

4 Random Walks

Consider unbiased random walk on a line. We have a sequence of random variables $X_0, X_1, X_2, \ldots$ where $X_{i+1} = X_i + R_i$ for independent Rademacher RVs $R_i \sim$ Rad. Basically, we go up or down with equal probability. First note that the expectation at time step n is $\mathbf{E}(X_n) = 0$ by linearity. Second, we have the result that $\mathbf{E}(|X_n|) = \Theta(\sqrt{n})$. Now, what about the maximum deviation from 0, i.e. what is $\mathbf{E}(\max_i|X_i|)$?

Notice $\max_i|X_i|$ corresponds exactly to the max discrepancy M in section 3. If we repeat the above analysis, we get $\mathbf{E}(\max_i|X_i|) = O(\sqrt{n\log n})$. On the other hand, a lower bound of $\Omega(\sqrt{n})$ is observed from the fact that $|X_n| \leq \max_i|X_i|$. Next we exploit the problem structure to show that the correct bound is indeed $\Theta(\sqrt{n})$.

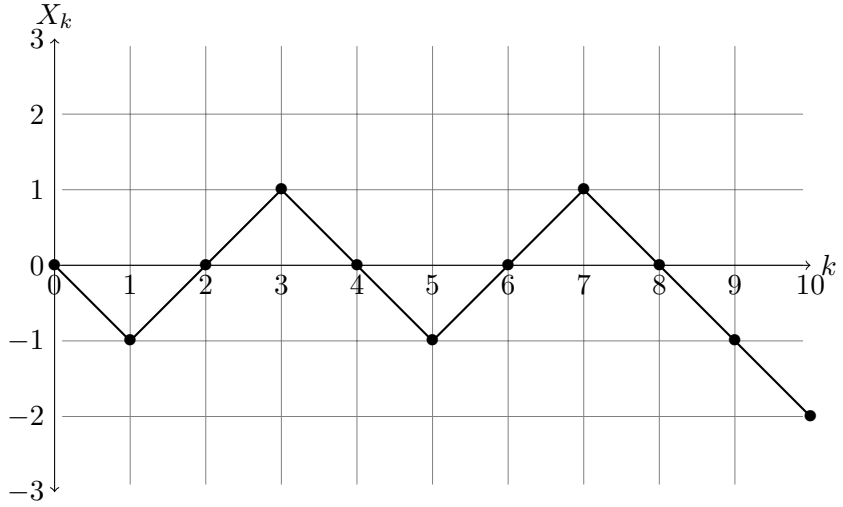

Figure 14.3: A random walk.

Proposition 14.4. *Let X_i be the random walk described above. Then,*

$$\mathbf{E}\left(\max_{i=1}^{n}|X_i|\right) = O(\sqrt{n}).$$

Proof. Notice that $O(\sqrt{\log n})$ is needed in the above bound to cancel out the factor of n introduced by the union bound. Therefore the idea is to avoid using such a crude union bound. This motivates us to decompose the sum $X_n = \sum_{j=1}^{n} R_j$ into several components. We consider $\log n$ ways to group the n Rademachers, $\{R_j\}_{j=1}^{n}$. In the kth grouping, which we call the kth level, we group the R_j into 2^k sums of length $n/2^k$. That is, we write X_n as

$$X_n = \sum_{j=1}^{n} R_j = \sum_{i=1}^{2^k} L_k^{(i)}, \qquad L_k^{(i)} \coloneqq \sum_{j=(i-1)(n/2^k+1)}^{i(n/2^k)} R_j.$$

Then, note that for any i, we can write X_i as the sum of at most one term $L_k^{(i)}$ from each level L_k for $0 \le k \le \log n$, which can be seen by expanding the number i in base 2. Thus, we may bound

$$\max_{i=1}^{n}|X_i| \le \sum_{k=0}^{\log n} \max_{i=1}^{2^k}|L_k^{(i)}|.$$

Now at level k, we have by the additive Chernoff bound and the union bound that

$$\mathbf{Pr}\left(\max_{i=1}^{2^k}|L_k^{(i)}| \ge \lambda\right) \le 2^k \exp\left(-\frac{c\lambda^2}{n/2^k}\right) \le \exp\left(k - \frac{c\lambda^2}{n/2^k}\right).$$

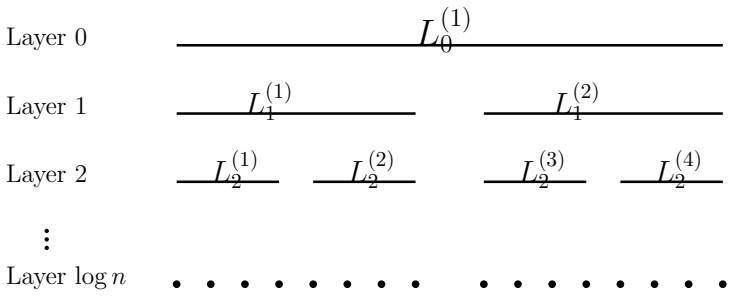

Figure 14.4: Segmentation tree with $\log n$ layer. Layer L_j contains $\frac{n}{2^j}$ IRVs.

We then convert the taill bound to an expectation bound via lemma 14.3. Then, we may obtain

$$\begin{aligned}
\mathbf{E}\left(\max_{i=1}^{2^k}|L_k^{(i)}|\right) &= \int_0^\infty \mathbf{Pr}\left(\max_{i=1}^{2^k}|L_k^{(i)}| \geq \lambda\right) d\lambda \\
&\leq \int_0^{c'\sqrt{nk/2^k}} 1 \; d\lambda + \int_{c'\sqrt{nk/2^k}}^\infty \exp\left(k - \frac{c\lambda^2}{n/2^k}\right) d\lambda \\
&= c'\sqrt{\frac{nk}{2^k}} + \int_0^\infty \exp\left(k - \frac{c(\lambda + c'\sqrt{nk/2^k})^2}{n/2^k}\right) d\lambda \\
&\leq c'\sqrt{\frac{nk}{2^k}} + \int_0^\infty \exp\left(-\frac{c\lambda^2}{n/2^k}\right) d\lambda \\
&\leq c'\sqrt{\frac{nk}{2^k}} + \frac{1}{2}\sqrt{\frac{n}{c2^k}} = O\left(\sqrt{nk/2^k}\right)
\end{aligned}$$

if we choose $c' = 1/\sqrt{c}$. Then by summing over $0 \leq k \leq \log n$, we conclude that

$$\mathbf{E}\left(\max_{i=1}^{n}|X_i|\right) \leq \sum_{k=0}^{\log n} \mathbf{E}\left(\max_{i=1}^{2^k}\left|L_k^{(i)}\right|\right) \leq \sum_{k=0}^{\log n} O\left(\sqrt{n}\frac{\sqrt{k}}{2^{k/2}}\right) = O(\sqrt{n})\sum_{k=0}^{\log n}\frac{\sqrt{k}}{2^{k/2}} = O(\sqrt{n}). \qquad \square$$

5 Martingales

In the final section of this lecture, we will introduce martingales, some of their nice results, and a simple proof of proposition 14.4 using these results.

Definition 14.5. A *martingale* is a sequence of random variables $0 = X_0, X_1, X_2, \ldots$, such that

1. $\mathbf{E}(|X_k|) < \infty$ for all k.
2. $\mathbf{E}(X_{k+1} \mid X_1, \ldots, X_k) = X_k$.

A few remarks are in order for the above definition. First of all, although we take martingales to always start at 0 for simplicity of notation in this lecture, there is no reason to restrict this in general. The first condition is a technical one that we do not really care about—it allows for the proof of nice properties of martingales that are beyond the scope of these notes. The second condition is the defining feature of martingales—the intuition of martingales is that they model the outcome of fair games, where once you see the first k outcomes, the expectation of the outcome of the next time step is just the current outcome.

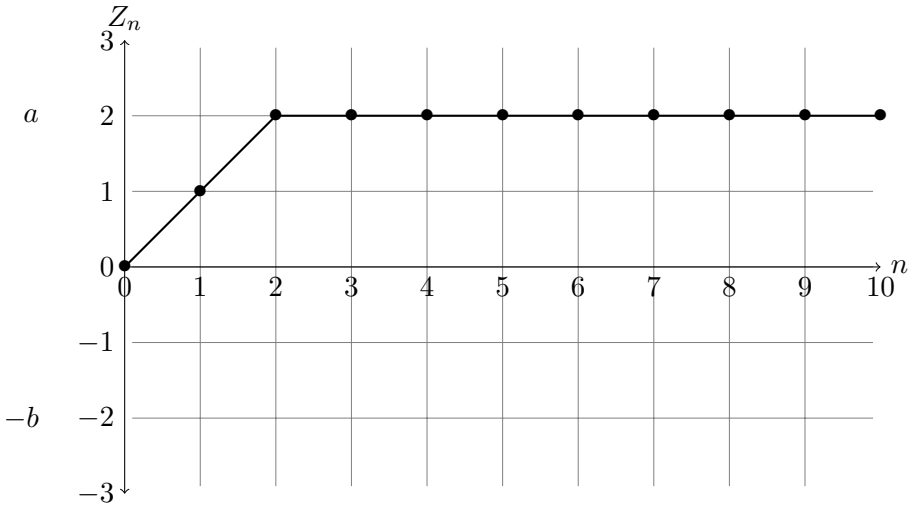

Figure 14.5: A stopped random walk.

Example 14.6. $X_{k+1} = X_k + R_k$ is a martingale, where $R_k \sim \text{Rad}$ are drawn independently for every k.

Example 14.7. Let X_k be defined as above. Then, $Y_k = X_k^2 - k$ is a martingale.

Proof. Condition 1 holds since $-k \leq X_k \leq k$ for every k. To check condition 2, we compute by conditioning on the Rademacher variable:

$$\mathbf{E}(Y_{k+1} \mid Y_k) = \frac{1}{2}(X_k + 1)^2 + \frac{1}{2}(X_k - 1)^2 - (k+1) = X_k^2 + 1 - (k+1) = X_k^2 - k = Y_k.$$

We thus conclude as desired. □

We also state the following simple result for later use.

Proposition 14.8. *Let $0 = X_0, X_1, X_2, \ldots$ be a martingale. Then $\mathbf{E}(X_k) = 0$.*

Proof. We have

$$\mathbf{E}(X_k) = \mathop{\mathbf{E}}_{X_1,\ldots,X_{k-1}}\left(\mathop{\mathbf{E}}_{X_k}(X_k \mid X_1, \ldots, X_{k-1})\right) = \mathbf{E}(X_{k-1})$$

and then we may just induct down to $k = 0$. □

5.1 Stopped random walks

We now show a more interesting example of a martingale.

Example 14.9 (Stopped random walk)**.** Define a sequence of random variables $\{Z_n\}_{n=0}^{\infty}$ by $Z_0 = 0$ and

$$Z_{n+1} = \begin{cases} Z_n & Z_n \in \{a, -b\} \\ Z_n + R_n & \text{otherwise} \end{cases}$$

where $R_n \sim \text{Rad}$ are drawn independently for each n. Then $\{Z_n\}_{n=0}^{\infty}$ is a martingale.

We will take the following proposition without proof. We refer the motivated reader to chapter 1 of these notes for a proof.

Proposition 14.10. *Let $\{Z_n\}_{n=0}^{\infty}$ be a random walk stopped at $\{a, -b\}$ and let τ be the random variable indicating the stopping time, i.e. the number of time steps that the martingale takes to reach a or $-b$. Then, $\mathbf{E}(\tau) < \infty$.*

We now prove a simple yet powerful result for stopped martingales.

Proposition 14.11 (Gambler's ruin)**.** *For a martingale $\{Z_n\}_{n=0}^{\infty}$ stopped at $\{a, -b\}$,*

$$p_a \coloneqq \mathbf{Pr}(\{Z_n\}_{n=0}^{\infty} \text{ stops at } a) = \frac{b}{a+b}, \qquad p_b \coloneqq \mathbf{Pr}(\{Z_n\}_{n=0}^{\infty} \text{ stops at } -b) = \frac{a}{a+b}.$$

Proof. Let $\varepsilon > 0$. Then we may find N large enough such that $\mathbf{Pr}(\tau \geq N) < \varepsilon$, since by proposition 14.10,

$$\mathbf{E}(\tau) = \sum_{n=0}^{\infty} \mathbf{Pr}(\tau \geq n) < \infty.$$

Now define

$$p_a^N \coloneqq \mathbf{Pr}(\{Z_n\}_{n=0}^{\infty} \text{ stops at } a \mid \tau < N)$$
$$p_b^N \coloneqq \mathbf{Pr}(\{Z_n\}_{n=0}^{\infty} \text{ stops at } -b \mid \tau < N)$$

i.e. the probabilities that $\{Z_n\}_{n=0}^{\infty}$ stops at a and $-b$ respectively, conditioned on the event that $\{Z_n\}_{n=0}^{\infty}$ stops before time N. Now using proposition 14.8, we compute

$$\begin{aligned} 0 &= \mathbf{E}(Z_N) \\ &= \mathbf{E}(Z_N \mid \tau < N)(1 - \mathbf{Pr}(\tau \geq N)) + \mathbf{E}(Z_N \mid \tau \geq N)\,\mathbf{Pr}(\tau \geq N) \\ &= ap_a^N + (-b)p_b^N + \left(\mathbf{E}(Z_N \mid \tau \geq N) - ap_a^N - (-b)p_b^N\right)\mathbf{Pr}(\tau \geq N). \end{aligned}$$

We also know $p_a^N + p_b^N = 1 - \mathbf{Pr}(\tau \geq N)$. Solving for p_a^N using these two equations and noting that $-b \leq \mathbf{E}(Z_N \mid \tau \geq N) \leq a$ yields

$$p_a^N = \frac{b}{a+b} \pm O(\epsilon)$$
$$p_b^N = \frac{a}{a+b} \pm O(\epsilon)$$

In view of a theorem[1] in probability theory, we have

$$p_a = \mathbf{Pr}(\{Z_n\}_{n=0}^{\infty} \text{ stops at } a) = \lim_{N\to\infty} \mathbf{Pr}(\{Z_n\}_{n=0}^{\infty} \text{ stops at } a \mid \tau < N) = \lim_{N\to\infty} p_a^N$$
$$p_b = \mathbf{Pr}(\{Z_n\}_{n=0}^{\infty} \text{ stops at } -b) = \lim_{N\to\infty} \mathbf{Pr}(\{Z_n\}_{n=0}^{\infty} \text{ stops at } -b \mid \tau < N) = \lim_{N\to\infty} p_b^N$$

Then by sending $N \to \infty$, we can send $\varepsilon \to 0$, so the above are just

$$p_a = \frac{b}{a+b}, \qquad p_b = \frac{a}{a+b},$$

as claimed. □

In the above result, we used the fact that the stopping time has finite expectation. We can now bootstrap this to compute the expected waiting time exactly, using proposition 14.11. We will not prove this and refer again to chapter 1 of these notes for a proof.

Proposition 14.12. *Let $\{Z_n\}_{n=0}^{\infty}$ be a random walk stopped at $\{a, -b\}$ and let τ be the random variable indicating the stopping time. Then, $\mathbf{E}(\tau) = ab$.*

Proof sketch. The idea is to use $0 = \mathbf{E}(Y_k)$ on the martingale $Y_k = X_k^2 - k$. □

[1] The theorem is the monotone convergence theorem for those interested in the gory details.

5.2 A martingale proof of the segment tree result

Finally, we will give a super slick martingale proof of proposition 14.4.

We first introduce the following Chernoff-type inequality for martingales with bounded differences. The proof of the result is distracting from the theme of this lecture, so we refer those interested to these notes.

Theorem 14.13 (Azuma's inequality)**.** *If $\{Z_n\}_{n=0}^{\infty}$ is a martingale with bounded differences, i.e. $|Z_{n+1} - X_n| \leq 1$ almost surely (with probability 1) for all n, then we have the following Chernoff-type bound:*

$$\mathbf{Pr}(|Z_n| \geq \lambda) \leq 2\exp\left(-\frac{\lambda^2}{2n}\right).$$

Essentially, the theorem allows us to say that all the Chernoff stuff we did previously applies to martingales. We may think of this result as a relaxation of the independence (note that conversely, it is easy to see that sums of independent 0-mean random variables are martingales). Now as promised, the proof.

Martingale proof of proposition 14.4. Fix $\lambda > 0$. Then, we may consider the random walk $\{Z_n\}_{n=0}^{\infty}$, given by stopping X_k at $\{-\lambda, \lambda\}$. Note then that $\max_{k=1}^{n} |X_k| \geq \lambda$ if and only if $|X_k| \geq \lambda$ at some $1 \leq k \leq n$. This in turn happens if and only if $|Z_n| \geq \lambda$. Then,

$$\mathbf{Pr}\left(\max_{k=1}^{n} |X_k| \geq \lambda\right) = \mathbf{Pr}(|Z_n| \geq \lambda) \leq 2\exp\left(-\frac{\lambda^2}{n}\right)$$

by Azuma's inequality, since $\{Z_n\}_{n=0}^{\infty}$ is a martingale. Finally by lemma 14.3, we have

$$\begin{aligned}\mathbf{E}\left(\max_{k=1}^{n} |X_k|\right) &= \int_0^{\infty} \mathbf{Pr}\left(\max_{k=1}^{n} |X_k| \geq \lambda\right) d\lambda \\ &\leq \int_0^{\infty} 2\exp\left(-\frac{\lambda^2}{n}\right) d\lambda = \sqrt{\pi n} = O(\sqrt{n}).\end{aligned}$$

□

Acknowledgments

These lecture notes were scribed by Ziye Tang and Taisuke Yasuda, based on previous scribe notes of Ziye Tang and Taisuke Yasuda.

References

[Ban10] Nikhil Bansal. Constructive algorithms for discrepancy minimization. In *Foundations of Computer Science (FOCS), 2010 51st Annual IEEE Symposium on*, pages 3–10. IEEE, 2010. 2.1

[Spe85] Joel Spencer. Six standard deviations suffice. *Transactions of the American mathematical society*, 289(2):679–706, 1985. 2.1