6.7720/18.619/15.070 Lecture 9
Introduction to Martingales

Kuikui Liu

March 3, 2025

Acknowledgements & Disclaimers *In the process of writing these notes, we consulted materials by Steven Lalle, Yury Polyanskiy, Shayan Oveis Gharan, James R. Lee, Anna Karlin, and Alistair Sinclair. Please be advised that these notes have not been subjected to the usual scrutiny reserved for formal publications. If you do spot an error, please contact the instructor.*

1 Martingales

In this lecture, we extend Chernoff-type bounds along two important axes:

- Previously, we only considered linear functionals or norms of bounded/sub-Gaussian/sub-exponential random vectors. In this lecture, we will derive concentration inequalities which work for arbitrary functions which are *Lipschitz* in a very natural sense.
- Up until now, nearly all our stochastic processes have been given by collections of *independent* random variables. In this lecture, we move far beyond the independent setting, and study *martingales*.

Definition 1 (Martingale)**.** *We say a (possibly finite) sequence of random variables $\{Y_n\}_{n\geq 0}$ is a* martingale *with respect to another sequence of random variables $\{X_n\}_{n\geq 0}$ if for every n,*

- $\mathbb{E}\left[|Y_n|\right] < \infty$*,*
- Y_n *is a function of* $X_0, \dots, X_n$*, and*
- $\mathbb{E}[Y_{n+1} \mid X_0, \dots, X_n] = Y_n$*.*

We simply say $\{Y_n\}_{n\geq 0}$ is a martingale if it is a martingale with respect to itself.

The key martingale criterion says that the difference $Y_{n+1} - Y_n$ is unbiased (has expectation zero) conditioned the past information. Note that the proper level of generality in which a martingale should be defined is via filtrations and measure theory. Since we won't need this in the bulk of this course, we relegate a formal discussion to Appendix A.

At a high level, one can view the random variables $X_0, \dots, X_n$ as the "information" one has "up to time n". This intuition is perhaps best illustrated through a representative class of examples.

1.1 Doob Martingales

As a thought experiment, consider the following statistical estimation "game" played between two players Alice and Bob. To set the parameters of the game before they play, Alice and Bob agree on some fixed *deterministic* function $f : \mathbb{R}^n \to \mathbb{R}$ (e.g. the sum $f(x_1, \dots, x_n) = \sum_{i=1}^n x_i$), and some probability distribution μ over $\mathbb{R}^n$ such that $\mathbb{E}_\mu\left[|f|\right] < \infty$. Note that the coordinates of a random sample from μ may be *arbitrarily correlated*. Alice and Bob are also allowed unbounded computational resources.

Having agreed on f and μ, to play the game, Alice samples a random vector $X = (X_1, \dots, X_n) \sim \mu$ and computes $Y = f(X_1, \dots, X_n)$, which is a random variable whose randomness depends on X. Bob's goal is to estimate the value of Y that Alice has sampled. Given zero information about the inputs X Alice sampled, one possible strategy for Bob is to simply output the expectation $\mathbb{E}_\mu[f]$

of f with respect to μ. This makes a lot of intuitive sense, especially if μ is concentrated, and in fact, it is easy to check that it is optimal with respect to the expected squared distance to Y, i.e.

$$\mathbb{E}_\mu[f] = \arg\min_{s\in\mathbb{R}} \mathbb{E}_Y\left[(s-Y)^2\right].$$

What if Alice decides to be generous, and reveals the first k coordinates $X_1,\dots,X_k$ for some $0 \le k \le n$? Certainly, if $k = n$, then Bob could just compute $Y = f(X)$. If $k < n$, then the natural generalization of the $k = 0$ strategy we just mentioned is for Bob to compute

$$Y_k = \mathbb{E}_\mu\left[f(X_1,\dots,X_n) \mid X_1,\dots,X_k\right].$$

Once the values of $X_1,\dots,X_k$ are fixed and revealed to Bob, the estimator Y_k that Bob outputs becomes some fixed number. However, we can also think of Y_k as a random variable whose randomness comes from the $X_1,\dots,X_k$ Alice sampled.

Lemma 1.1. *The random variables $\{Y_k\}_{k=0}^n$ form a martingale with respect to $\{X_k\}_{k=1}^n$.*[1]

Proof. Observe that for each $0 \le k \le n-1$,

$$\begin{aligned}
\mathbb{E}\left[Y_{k+1} \mid X_1,\dots,X_k\right] &= \mathbb{E}\left[\mathbb{E}\left[f(X) \mid X_1,\dots,X_{k+1}\right] \mid X_1,\dots,X_k\right] && \text{(Definition of } Y_{k+1}\text{)}\\
&= \mathbb{E}\left[f(X) \mid X_1,\dots,X_k\right] && (*)\\
&= Y_k.
\end{aligned}$$

Here, the crucial step $(*)$ follows from the *Tower Property of Conditional Expectations*.[2] One way to think about it is that in the expression involving nested expectations, the inner one is over the unrevealed coordinates $X_{k+2},\dots,X_n$, and the outer one introduces an additional expectation over X_{k+1}. □

The scenario we just described is a concrete example of how martingales model situations in which information is gradually accumulated over time. This is perhaps the primary reason why martingales are so powerful and well-studied. Lemma 1.1 is a special case of a much more general class of martingales called *Doob martingales* (or *Lévy martingales*), all of which have this intuitive interpretation. Let us see a few concrete examples of Doob martingales.

Balls in Bins Suppose we have m balls and n bins. In the kth step, we throw the kth ball independently into a uniformly random bin $X_k \in \{1,\dots,n\}$. This type of process commonly appears in the analysis of load balancing protocols in parallel computing, and data structures based on hashing. After throwing all the balls, we are interested in the statistics of the distribution of balls among the bins. An example of such a statistic is the number of empty bins

$$f(X_1,\dots,X_m) = \#\{i \in [n] : X_k \neq i, \forall k\}.$$

This particular statistic can actually be expressed as a sum of random variables, namely

$$\#\{\text{empty bins}\} = \sum_{i=1}^n \mathbf{1}[\text{bin } i \text{ is empty}].$$

However, these indicator random variables are *not independent*.[3] To get around this, one can instead build an associated Doob martingale by letting $Y_k = \mathbb{E}[f(X_1,\dots,X_m) \mid X_1,\dots,X_k]$, and then use this martingale to study e.g. concentration of $Y_m = f(X_1,\dots,X_m)$ around its expectation.

Edge/Vertex Exposure in Random Graphs Let f be any function mapping simple undirected graphs G to real numbers. A fruitful technique for studying the random variable $f(G)$ for a randomly chosen graph G (e.g. according to Erdös–Rényi $\boldsymbol{G}(n,p)$) is to iteratively reveal information about G and study the associated Doob martingale. Two natural ways are the following:

[1] One can artificially set $X_0 = \emptyset$, which represents Alice "adding no new information" about X.
[2] This is just a fancier term for the Law of Total Expectation or the Law of Total Probability.
[3] One can show, however, that these random variables are *negatively correlated*.

- **Edge Exposure Filtration:** One can fix some ordering $e_1, \dots, e_{\binom{V}{2}}$ of the set of pairs $\binom{V}{2}$, and then reveal whether or not $e_t \in G$ as one iterates over t. For each $0 \le t \le \binom{V}{2}$, one gets a random set of edges $E_t \subseteq \{e_1, \dots, e_t\}$, and a corresponding Doob martingale $Y_t = \mathbb{E}[f(G) \mid E_0, \dots, E_t]$.
- **Vertex Exposure Filtration:** One can also fix some ordering $v_1, \dots, v_n$, and then reveal the set of edges between v_t and $v_1, \dots, v_{t-1}$ in G as one iterates over t. Hence, for each $0 \le t \le n$, one gets the (random) induced subgraph G_t of G on vertex set $\{v_1, \dots, v_t\}$, and a corresponding Doob martingale $Y_t = \mathbb{E}[f(G) \mid G_0, \dots, G_t]$.

Note that f can be an incredibly complicated function. For instance, if f is the chromatic number, then even approximating f to within a multiplicative factor of $n^{\Omega(1)}$ is as hard as solving general instances of SAT (i.e. it is NP-hard). Nonetheless, martingale techniques can allow one to deduce strong concentration around its expectation without even knowing what the expectation is; note, however, that establishing matching upper and lower bounds for the latter task is nontrivial. We discuss the chromatic number in greater depth in Section 4.

1.2 Other Simple Examples of Martingales

Gambling Let $\{X_n\}_{n=0}^{\infty}$ be a sequence of i.i.d. $\mathsf{Unif}\{\pm 1\}$ random variables, and for a sequence of deterministic functions $\{f_n : \{\pm 1\}^n \to \mathbb{R}_{\ge 0}\}_{n=1}^{\infty}$, define the random variables

$$W_n = W_{n-1} + f_n(X_0, \dots, X_{n-1}) \cdot X_n, \qquad \forall n \in \mathbb{N}.$$

Then $\{W_n\}_{n=0}^{\infty}$ is a martingale with respect to $\{X_n\}_{n=0}^{\infty}$ by independence and the fact that each X_n has mean-zero. If $\{X_n\}_{n=0}^{\infty}$ are the win/loss outcomes of a (fair) game at a casino, one can then interpret the sequence of functions $\{f_n\}_{n=1}^{\infty}$ as a gambling strategy, where the player bets $f_n(X_0, \dots, X_{n-1})$ units of currency based on looking at the history of outcomes $X_0, \dots, X_{n-1}$. The random variables $\{W_n\}_{n=0}^{\infty}$ then represent the evolution of the player's wealth.

Remark 1. Later in the course, we will discuss another class of dependent stochastic processes called *Markov chains*, whose definition has perhaps a similar flavor to Definition 1. However, these are distinct (although related) concepts. For example, one can view the martingale $\{W_n\}_{n=0}^{\infty}$ as an unbiased random walk which is *not Markovian*, because the size of the next step depends on the entire history, not only the current state. It is also easy to construct Markov chains which are not martingales (e.g. biased random walks on $\mathbb{Z}$).

Galton–Watson Branching Processes Suppose $\{Z_\ell\}_{\ell \in \mathbb{N}}$ is a Galton–Watson branching process with offspring distribution ξ having finite mean μ and variance σ^2. Then the random variables $\left\{Z_\ell / \mu^\ell\right\}_{\ell \in \mathbb{N}}$ form a martingale, since

$$\mathbb{E}\left[\frac{Z_{k+1}}{\mu^{k+1}} \,\middle|\, \frac{Z_0}{\mu^0}, \dots, \frac{Z_k}{\mu^k}\right] = \mathbb{E}\left[\frac{Z_{k+1}}{\mu^{k+1}} \,\middle|\, \frac{Z_k}{\mu^k}\right] = \frac{Z_k \cdot \mu}{\mu^{k+1}} = \frac{Z_k}{\mu^k}.$$

Second Moment Martingales Let $\{X_n\}_{n=0}^{\infty}$ be a sequence of independent zero-mean random variables with variance $\sigma^2 < \infty$. Then the sequence of random variables $\{Y_n\}_{n=0}^{\infty}$ given by

$$Y_n \stackrel{\mathsf{def}}{=} \left(\sum_{k=0}^{n} X_k\right)^2 - \sigma^2 n$$

is a martingale with respect to $\{X_n\}_{n=0}^{\infty}$. Indeed,

$$\begin{aligned}
\mathbb{E}\left[Y_{n+1} \mid X_0, \ldots, X_n\right] &= \mathbb{E}\left[\left(X_{n+1} + \sum_{k=0}^{n} X_k\right)^2 - \sigma^2(n+1) \,\middle|\, X_0, \ldots, X_n\right] \\
&= \mathbb{E}\left[X_{n+1}^2 + 2X_{n+1}\sum_{k=0}^{n} X_k + \left(\sum_{k=0}^{n} X_k\right)^2 - \sigma^2(n+1) \,\middle|\, X_0, \ldots, X_n\right] \\
&= \left(\sum_{k=0}^{n} X_k\right)^2 - \sigma^2 n \qquad \text{(Independence and linearity of expectation)} \\
&= Y_n.
\end{aligned}$$

Actually, if one were to cut off these sequences at some finite N, then by using independence of $\{X_n\}_{n=0}^{N}$, we can write $\{Y_n\}_{n=0}^{N}$ as a Doob martingale with respect to $\{X_n\}_{n=0}^{N}$ in the form given in Lemma 1.1 by letting $f(X_0, \ldots, X_N) = \left(\sum_{k=0}^{N} X_k\right)^2 - \sigma^2 N$.

2 Concentration for Martingales

Let us begin with a basic observation concerning the mean and variance of a martingale.

Lemma 2.1. *Let $\{Y_n\}_{n\geq 0}$ be a martingale with respect to another sequence of random variables $\{X_n\}_{n\geq 0}$. Then for all n, we have*

$$\mathbb{E}[Y_n] = \mathbb{E}[Y_0] \qquad \text{and} \qquad \mathbb{E}\left[(Y_n - Y_0)^2\right] = \sum_{k=1}^{n} \mathbb{E}\left[(Y_k - Y_{k-1})^2\right].$$

Proof. The first claim is an immediate consequence of the Tower Property of Conditional Expectations. Using this, we can rewrite the second claim as

$$\operatorname{Var}(Y_n - Y_0) = \sum_{k=1}^{n} \operatorname{Var}(Y_k - Y_{k-1}).$$

Since $Y_n - Y_0 = \sum_{k=1}^{n}(Y_k - Y_{k-1})$ (by telescoping), it suffices to prove that $Y_k - Y_{k-1}$ is *uncorrelated* with $Y_\ell - Y_{\ell-1}$ for any $k < \ell$. For this, we have

$$\begin{aligned}
\mathbb{E}\left[(Y_\ell - Y_{\ell-1})(Y_k - Y_{k-1})\right] &= \mathbb{E}\left[\mathbb{E}\left[(Y_\ell - Y_{\ell-1})(Y_k - Y_{k-1}) \mid X_0, \ldots, X_k\right]\right] \\
&= \mathbb{E}\left[(Y_k - Y_{k-1}) \cdot \mathbb{E}\left[Y_\ell - Y_{\ell-1} \mid X_0, \ldots, X_k\right]\right] \\
&\qquad\qquad (Y_k - Y_{k-1} \text{ is a function of } X_0, \ldots, X_k) \\
&= 0,
\end{aligned}$$

where we used $k < \ell$ so that $\mathbb{E}[Y_\ell - Y_{\ell-1} \mid X_0, \ldots, X_k] = 0$. □

In particular, if $Y_0 \equiv 0$, then Lemma 2.1 shows that the variance of Y_n can be decomposed as a sum of the variances of the *increments* $Y_k - Y_{k-1}$. Now, similar to Chernoff–Hoeffding, if we ensure almost-sure boundedness of these increments, then we can obtain Chernoff-type concentration.

Theorem 2.2 (Azuma–Hoeffding Inequality)**.** *Let $\{Y_n\}_{n\geq 0}$ be a martingale with respect to another sequence of random variables $\{X_n\}_{n\geq 0}$. If for each n, there exists a positive constant $c_n > 0$ such that $|Y_n - Y_{n-1}| \leq c_n$ almost surely, then*

$$\Pr\left[Y_n - Y_0 \geq t\right] \leq \exp\left(-\frac{t^2}{2\sum_{k=1}^{n} c_k^2}\right), \qquad \forall n, \forall t \geq 0.$$

Remark 2. Note that if Y_0 is identically equal to some constant, e.g. 0, then $Y_0 = \mathbb{E}[Y_n]$ and the above yields a bound on the probability that Y_n deviates from its expectation.

Remark 3. One can easily generalize the proof to the setting where the increments are sub-Gaussian or sub-exponential (almost surely conditioned on the past).

Proof. As usual, we consider the moment generating function. For a parameter $s > 0$, observe that

$$\begin{aligned}
&\mathbb{E}\left[\exp\left(s \cdot (Y_n - Y_0)\right) \mid X_0, \ldots, X_{n-1}\right] \\
&\quad = \mathbb{E}\left[\exp\left(s \cdot (Y_{n-1} - Y_0)\right) \cdot \exp\left(s \cdot (Y_n - Y_{n-1})\right) \mid X_0, \ldots, X_{n-1}\right] \\
&\quad = \exp\left(s \cdot (Y_{n-1} - Y_0)\right) \cdot \mathbb{E}\left[\exp\left(s \cdot (Y_n - Y_{n-1})\right) \mid X_0, \ldots, X_{n-1}\right] \\
&\quad \leq \exp\left(s \cdot (Y_{n-1} - Y_0)\right) \cdot \exp\left(\frac{c_n^2 s^2}{2}\right). \qquad \text{(Hoeffding's Lemma)}
\end{aligned}$$

Taking expectations of both sides, we see that

$$\mathbb{E}\left[\exp\left(s \cdot (Y_n - Y_0)\right)\right] \leq \exp\left(\frac{c_n^2 s^2}{2}\right) \cdot \mathbb{E}\left[\exp\left(s \cdot (Y_{n-1} - Y_0)\right)\right].$$

Since this holds for all n, by induction, it follows that the moment generating function of $Y_n - Y_0$ is bounded as

$$\mathbb{E}\left[\exp\left(s \cdot (Y_n - Y_0)\right)\right] \leq \exp\left(\frac{s^2 \cdot \sum_{k=1}^n c_k^2}{2}\right), \qquad \forall s \in \mathbb{R}.$$

Applying Markov's Inequality in the usual way and optimizing over $s \in \mathbb{R}$ yields the desired bound. □

3 Concentration for Lipschitz Functions

Let us see how to use Theorem 2.2 to derive concentration for general functions of independent random variables which are Lipschitz in a precise sense.

Theorem 3.1 (McDiarmid's Inequality)**.** *Let $X_1, \ldots, X_n$ be a collection of independent real-valued random variables. Suppose $f : \mathbb{R}^n \to \mathbb{R}$ which is L-Lipschitz with respect* Hamming distance *on $\mathbb{R}^n$, i.e. for all $\boldsymbol{x}, \boldsymbol{y} \in \mathbb{R}^n$*

$$|f(\boldsymbol{x}) - f(\boldsymbol{y})| \leq L \cdot \#\{i \in [n] : \boldsymbol{x}_i \neq \boldsymbol{y}_i\}.$$

Then for every $t \geq 0$,

$$\Pr\left[f(X) - \mathbb{E}[f(X)] \geq t\right] \leq \exp\left(-\frac{t^2}{2L^2 n}\right).$$

Remark 4. This result can be generalized to allow for arbitrary *(Cartesian) product spaces* $\mathcal{X}_1 \times \cdots \times \mathcal{X}_n$, where X_k is independently sampled from $\mathcal{X}_k$ for each k (e.g. $\{\pm 1\}, \mathbb{R}, \mathbb{N}, \mathbb{F}_p$, etc.). Furthermore, Lipschitzness with respect to Hamming distance can be relaxed to the *bounded differences property*, which is essentially Lipschitzness with respect to *weighted* Hamming metrics:

$$|f(\boldsymbol{x}) - f(\boldsymbol{y})| \leq \sum_{i=1}^n c_i \cdot \mathbf{1}[\boldsymbol{x}_i \neq \boldsymbol{y}_i],$$

for some positive constants $c_1, \ldots, c_n > 0$. In this more general setting, one simply replaces $L^2 n$ in the tail bound with $\sum_{i=1}^n c_i^2$. Moreover, one can replace the factor of $\frac{1}{2}$ by 2 in the exponential.

Remark 5. There are various other concentration inequalities which hold with a different notion of Lipschitzness. For instance, the *Gaussian concentration inequality* states that if X is a vector of i.i.d. standard Gaussians in $\mathbb{R}^n$, and $f : \mathbb{R}^n \to \mathbb{R}$ is L-Lipschitz with respect to the standard Euclidean norm on $\mathbb{R}^n$, then

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

We will treat such inequalities much more systematically later in the course.

Remark 6. McDiarmid's Inequality is sharp for linear Lipschitz functionals (e.g. sums of bounded random variables). However, it becomes lossy for nonlinear Lipschitz functions. One example we'll see in a future lecture is triangle counts in $\boldsymbol{G}(n,p)$. In our application to stochastic Euclidean TSP below, we'll also see that a direct application of Theorem 3.1 is suboptimal. We note that there has been beautiful recent developments in the area of *nonlinear large deviations* [**???**], which aims to develop sharp tail estimates.

Proof of Theorem 3.1. Consider the Doob martingale $Y_k = \mathbb{E}[f(X_1, \dots, X_n) \mid X_1, \dots, X_k]$ for $k = 0, \dots, n$; this is the same martingale as the one we discussed in Section 1.1 and Lemma 1.1. We have $Y_n = f(X)$ while $Y_0 = \mathbb{E}[f(X)]$. Moreover, L-Lipschitzness of f with respect to Hamming distance, combined with independence of $X_1, \dots, X_n$, implies that the increments are bounded as $|Y_k - Y_{k-1}| \leq L$ almost surely for all $k = 1, \dots, n$. The claim is then an immediate consequence of Theorem 2.2. □

4 Concentration for the Chromatic Number of Erdös–Rényi

Recall that for a graph $G = (V, E)$ and a positive integer $q \in \mathbb{N}$, a *(proper) q-coloring* is a mapping $\sigma : V \to [q] = \{1, \dots, q\}$ such that for every edge $uv \in E$, we have $\sigma(u) \neq \sigma(v)$. We write $\chi(G)$ for the smallest q such that G admits a proper q-coloring. In this section, we consider the dense Erdös–Rényi random graph, i.e. $\boldsymbol{G}(n, p)$ where p is a constant independent of n. We will establish sharp concentration estimates for its chromatic number. We begin with a simple lower bound on its expectation.

Fact 4.1. *For any constant $p > 0$, we have the lower bound $\mathbb{E}\left[\chi\left(\boldsymbol{G}(n,p)\right)\right] \geq \Omega_p\left(\frac{n}{\log n}\right)$.*

Here, by $\Omega_p(\cdot)$, we mean the lower bound is up to constants depending on p. The key observation behind its proof is that in any proper q-coloring $\sigma : V \to [q]$, every color class $\sigma^{-1}(\mathfrak{c})$, where $\mathfrak{c} \in [q]$, must be an independent set. Hence, we must have the lower bound

$$q \geq \frac{n}{|\mathsf{Maximum\ Independent\ Set}|}.$$

In $\boldsymbol{G}(n, p)$, one can upper bound the size of the maximum independent set by $O_p(\log n)$ using the first moment method. In a future lecture, we will discuss the size of the maximum independent set of dense Erdös–Rényi in greater depth. We also note that sharper results are known for the chromatic number. Again, writing $A = (1 \pm \epsilon)B$ to mean $(1 - \epsilon)B \leq A \leq (1 + \epsilon)B$ for positive quantities $A, B > 0$, we have

$$\mathbb{E}\left[\chi\left(\boldsymbol{G}(n,p)\right)\right] = (1 \pm o(1)) \cdot \frac{n}{2\log_{\frac{1}{1-p}} n}.$$

Now, let us turn to concentration around its expectation.

Theorem 4.2. *For any constant $p > 0$ and any $0 < \epsilon < 1$, we have*

$$\Pr\left[\chi\left(\boldsymbol{G}(n,p)\right) \notin (1 \pm \epsilon) \cdot \mathbb{E}\left[\chi\left(\boldsymbol{G}(n,p)\right)\right]\right] \leq 2 \cdot \exp\left(-O_p\left(\frac{\epsilon^2 \cdot n}{\log^2 n}\right)\right).$$

Proof. Our goal is to leverage McDiarmid's Inequality (see Theorem 3.1). Perhaps the most immediate way to do this is to use the fact that $\chi : 2^{\binom{[n]}{2}} \to \mathbb{N}$ is 1-Lipschitz with respect to Hamming distance on $2^{\binom{[n]}{2}}$; this corresponds to the edge exposure martingale we discussed above. Indeed, if one adds an edge to a graph, then the chromatic number can increase by at most 1, since one could take any previously valid coloring, and color one of the endpoints of the new edge with a new color. This same argument shows that by removing an edge, the chromatic number can decrease by at most 1. However, because there are $\binom{n}{2}$ independent $\{0, 1\}$-valued random variables (corresponding to possible edges), McDiarmid's Inequality combined with Fact 4.1 only yields a tail bound of the form

$$2 \cdot \exp\left(-O_p\left(\frac{\epsilon^2}{\log^2 n}\right)\right),$$

which tends to 1 as $n \to \infty$. Unfortunately, this is not a very meaningful bound.

The main reason the above approach failed was because we were viewing χ as a function of too many independent random variables. To reduce the number of inputs, we use the vertex exposure martingale instead. Arbitrarily fix some ordering of the vertices $v_1, \dots, v_n$, and for each $1 \leq t \leq n$, let $E_t = \{v_i v_t \in E : 1 \leq i \leq t - 1\}$ denote the set of edges connecting v_t to the preceding vertices $v_1, \dots, v_{t-1}$. $E_1, \dots, E_n$ are independent because they are determined by statuses of disjoint sets

of vertex pairs. Moreover, fixing the values of $E_1, \dots, E_n$ uniquely determines the graph, and so we may view χ as a function of these variables.

Now, χ remains 1-Lipschitz with respect this choice of inputs, i.e. by arbitrarily changing E_t for any single $1 \le t \le n$, the chromatic number can change by at most 1. To see this, observe that the chromatic number of G is always upper bounded (resp. lower bounded) by the chromatic number of the graph formed by taking G and adding (resp. removing) all possible edges to v_t. But these upper and lower bounds can differ by at most one, since we can take a proper coloring of the graph with fewer edges, and introduce a single new color just for the vertex v_t. Since we now only have n independent random variables in the input to χ, McDiarmid's Inequality combined with Fact 4.1 immediately yields the desired conclusion. □

A A Brief Measure-Theoretic Treatment of Martingales

Recall that a *probability space* is a triple $(\Omega, \mathcal{F}, \mu)$, where

- Ω is some state space (e.g. $\{\pm 1\}^n$ or $\mathbb{R}^n$),
- $\mathcal{F}$ is a *σ-algebra*, i.e. a family of subsets of Ω (called the *measurable sets* or *events*) which is closed under complements and countable unions, and
- $\mu : \mathcal{F} \to [0, 1]$ is a probability measure, i.e. a function such that $\mu(\Omega) = 1$ and $\mu\left(\bigcup_{k=0}^{\infty} E_k\right) = \sum_{k=0}^{\infty} \mu(E_k)$ for any countable collection of mutually exclusive events $\{E_k\}_{k=0}^{\infty}$.

Also recall that a *random variable* is a function $X : \Omega \to \Sigma$, where $(\Sigma, \mathcal{G})$ is some other measurable space (often $\mathbb{R}$ equipped with Lebesgue measurable sets), which is *measurable*, i.e. the preimage $X^{-1}(B)$ is $\mathcal{F}$-measurable for any $\mathcal{G}$-measurable B. Note that given a measurable space $(\Sigma, \mathcal{G})$ and a function $X : \Omega \to \Sigma$, one can define the *σ-algebra $\sigma(X)$ generated by X* as the smallest σ-algebra containing the sets $X^{-1}(B)$ for every $B \in \mathcal{G}$.

Finally, recall that if $\mathcal{F}' \subseteq \mathcal{F}$ is a sub-σ-algebra and $X : \Omega \to \mathbb{R}$ is $\mathcal{F}$-measurable, then we say a $\mathcal{F}'$-measurable function $X' : \Omega \to \mathbb{R}$ is a *conditional expectation* if it satisfies

$$\int_A X' \, d\mu = \int_A X \, d\mu, \qquad \forall A \in \mathcal{F}'.$$

Often, we will simply write $\mathbb{E}[X \mid \mathcal{F}']$ for such a conditional expectation when it is clear from context.

Definition 2 (Filtration)**.** *For a probability space $(\Omega, \mathcal{F}, \mu)$, a* filtration *is a sequence of σ-algebras $\{\mathcal{F}_n\}_{n \ge 0}$ which is* nested *(or* increasing*):*

$$\mathcal{F}_0 \subseteq \mathcal{F}_1 \subseteq \cdots \subseteq \mathcal{F}_\infty \stackrel{\text{def}}{=} \sigma\left(\bigcup_{n=0}^{\infty} \mathcal{F}_n\right) \subseteq \mathcal{F}.$$

For instance, if $\{X_n\}_{n=0}^{\infty}$ is a sequence of real-valued random variables, then they generate a filtration via $\mathcal{F}_n \stackrel{\text{def}}{=} \sigma(X_0, \dots, X_n)$ for every n.

It is easiest to get handle on all this abstract nonsense by looking at a concrete example. Suppose $\Omega = \{\pm 1\}^n$, $\mathcal{F}$ is the collection of all subsets of Ω, and μ is any probability measure. If we let $X_k : \{\pm 1\}^n \to \mathbb{R}$ be the function which outputs the kth coordinate of its input, then the σ-algebra $\mathcal{F}_k = \sigma(X_1, \dots, X_k)$ consists of *subcubes*:

$$\{(x_1, \dots, x_k, x_{k+1}, \dots, x_n) \in \{\pm 1\}^n : x_{k+1}, \dots, x_n \in \{\pm 1\}\}, \qquad \forall (x_1, \dots, x_k) \in \{\pm 1\}^k.$$

Thus, the filtration $\{\mathcal{F}_k\}_{k=0}^{n}$ describes the process of revealing the values of the first k coordinates for some k.

In general, if $X \in \Omega$ is drawn from μ, then we can "reveal information about X" by enforcing that X land in some set which is measurable with respect to a sub-σ-algebra $\mathcal{F}'$ (e.g. the first k coordinates of X equal some fixed $x_1, \dots, x_k$). A filtration then gives a formal way of "zooming in" on the full description of X. If we additionally have a function $f : \Omega \to \mathbb{R}$, then looking at conditional expectations of f with respect to successive σ-algebras in a filtration gives a systematic way of "zooming in" on $f(X)$.

We now define martingales in the language of measure theory.

Definition 3 (Martingale)**.** *Fix a probability space* $(\Omega, \mathcal{F}, \mu)$ *and a filtration* $\{\mathcal{F}_n\}_{n\geq 0}$*. We say a sequence of real-valued random variables* $\{Y_n\}_{n\geq 0}$ *is* adapted *to* $\{\mathcal{F}_n\}_{n\geq 0}$ *if* Y_n *is a* $\mathcal{F}_n$*-measurable function for every* n*. We say* $\{Y_n\}_{n\geq 0}$ *is a* martingale *with respect to* $\{\mathcal{F}_n\}_{n\geq 0}$ *if*

- $\{Y_n\}_{n\geq 0}$ *is adapted to* $\{\mathcal{F}_n\}_{n\geq 0}$*,*
- $\mathbb{E}\left[|Y_n|\right] < \infty$ *for all* n*, and*
- $\mathbb{E}\left[Y_n \mid \mathcal{F}_{n-1}\right] = Y_{n-1}$ *for all* n*.*

Note that if Y_n is measurable with respect to $\sigma(X_0, \ldots, X_n)$, then it means that specifying the values of $X_0, \ldots, X_n$ completely determines the value of Y_n.