P2P and Martingales

Minseok Kwon

1 Martingales

1.1 Conditional Expectation

Let X be a random variable and $\mathcal{E}$ any event that occurs with a non-zero probability. The *conditional density function* of X given $\mathcal{E}$ is given by $\mathbf{Pr}[X = x|\mathcal{E}]$. In particular, $\mathcal{E}$ can be the event that some other random variable Y takes on a specific value y. Denoting the joint density function of X and Y by $p(x, y)$, we have

$$\mathbf{Pr}[X = x|Y = y] = \frac{p(x, y)}{\mathbf{Pr}[Y = y]} = \frac{p(x, y)}{\sum_x p(x, y)}$$

and

$$E[X|Y = y] = \frac{\sum_x xp(x, y)}{\sum_x p(x, y)}$$

where $E[X|Y = y]$ is the *conditional expectation* of X given that Y equals y.

Definition 1.1 *The random variable $E[X|Y]$ is defined to be the random variable $f(Y)$ such that $f(y) = E[X|Y = y]$.*

Example: Consider independent throws of an unbiased 6-sided die. For $1 \leq i \leq 6$, let X_i denote the number of times the value i appears in n throws of the die. Then

$$E[X_1|X_2] = \frac{n - X_2}{5}$$

$$E[X_1|X_2, X_3] = \frac{n - X_2 - X_3}{4}$$

If we knew that there are α occurrences of 2, we can compute the expected value of X_1 as $(n-\alpha)/5$; given the further information that there are β occurrences of 3, we can compute the expected value of X_1 as $(n-\alpha-\beta)/4$:

$$E[X_1|X_2=\alpha] = \frac{n-\alpha}{5}$$

$$E[X_1|X_2=\alpha, X_3=\beta] = \frac{n-\alpha-\beta}{4}$$

Lemma 1.1 *Property of Conditional Expectation:*

$$E[E[X|Y]] = E[X]$$

Proof:

$$\begin{aligned}
E[E[X|Y]] &= \sum_y E[X|Y=y]P\{Y=y\} \\
&= \sum_y \sum_x xP\{X=x|Y=y\}P\{Y=y\} \\
&= \sum_y \sum_x xP\{X=x, Y=y\} \\
&= \sum_x x \sum_y P\{X=x, Y=y\} \\
&= \sum_x xP\{X=x\} \\
&= E[X]
\end{aligned}$$

□

1.2 Martingales

Martingales are useful in handling sums of random variables which are not totally independent.

Definition 1.2 *A sequence of random variables $X_0, X_1, \ldots,$ is said to be a* martingale sequence *if for all $i > 0$,*

$$E[X_i|X_0, \ldots, X_{i-1}] = X_{i-1}.$$

Theorem 1.1 *Let $X_0, X_1, \ldots$ be a martingale sequence such that for each k,*

$$|X_k - X_{k-1}| \leq c_k$$

where c_k may depend on k. Then Azuma's inequality *gives for all $t \geq 0$ and any $\lambda > 0$,*

$$\Pr(|X_t - X_0| \geq \lambda) \leq 2e^{-\frac{\lambda^2}{2\sum_{k=1}^{t} c_k^2}} \tag{1}$$

It is easy to see the connection between this bound and the Chernoff bound for the sum of Poisson trials. Let $Z_1, \ldots, Z_n$ be independent variables that take values 0 or 1 each with probability 1/2. The random variable $S = \sum_{i=1}^{n} Z_i$ has the binomial distribution with parameters n and $p = 1/2$. Define a martingale sequence X_0, X_1, ..., X_n by setting $X_0 = E[S]$, and, for $1 \leq i \leq n$, $X_i = E[S|Z_1, \ldots, Z_i]$. It is clear that for $1 \leq i \leq n$, $|X_i - X_{i-1}| \leq 1$, since fixing the value of any one variable Z_i can only affect the expected value of the sum S by at most 1. It follows that the probability that S deviates from its expected value $X_0 = E[S] = n/2$ by more than λ is bounded by $2e^{-\lambda^2/2n}$, a slightly weaker result than can be inferred from the Chernoff bound for binomial distributions.

Corollary 1.1 *If $|X_k - X_{k-1}| \leq c$ where c is independent of k, then for all $t \geq 0$ and any $\lambda > 0$,*

$$\Pr(|X_t - X_0| \geq \lambda c\sqrt{t}) \leq 2e^{-\lambda^2/2}$$

The application of Azuma's inequality is sometimes called "the method of bounded differences." In applying this method to a martingale sequence, it is essential to set up the martingale in such a way as to guarantee the "bounded difference" property.

1.3 Doob Martingale

Let $Z_1, Z_2, \ldots, Z_n$ be *any* sequence of random variables and let X be any random variable. Define the random variable $X_i = E[X|Z_1, \ldots, Z_i]$, i.e., the conditional expectation of X conditioned on the variables Z_1 to Z_i. Then $X_0 = E[X], X_1, \ldots, X_n$ form a martingale sequence (Doob martingale).

1.4 Lipschitz condition

Definition 1.3 *Let $f : D_1 \times \cdots \times D_n \to R$ be a real-valued function with n arguments from possibly distinct domains. The function f is said to satisfy the* Lipschitz condition *if for any $x_1 \in D_1, \ldots, x_n \in D_n$, any $i \in \{1, \ldots, n\}$, and any $y_i \in D_i$, $|f(x_1, \ldots x_{i-1}, x_i, x_{i+1}, \ldots, x_n) - f(x_1, \ldots, x_{i-1}, y_i, x_{i+1}, \ldots, x_n)| \leq 1$*

Basically, a function satisfies the Lipschitz condition if an arbitrary change in the value of any one argument does not change the value of the function by more than 1. Suppose we have a sequence of random variables $X_1, \ldots, X_n$, and a function $f(X_1, \ldots, X_n)$ defined over them such that f satisfies the Lipschitz condition. Define the Doob martingale sequence Y_0, Y_1, ..., Y_n by setting $Y_0 = E[f(X_1, \ldots, X_n)]$ and, for $1 \leq i \leq n$, $Y_i = E[f(X_1, \ldots, X_n)|X_1, \ldots, X_i]$. It is easy to verify that the Lipschitz condition implies that for $1 \leq i \leq n$, $|Y_i - Y_{i-1}| \leq 1$.

2 Diameter

2.1 Expansion Lemma

Lemma 2.1 *If $|\Gamma_{i-1}(v)| = o(N)$,*

$$Pr\{|\Gamma_i(v)| \geq 2|\Gamma_{i-1}(v)|\} \geq 1 - 1/N^5.$$

Proof: We use an exposure martingale to prove that $\Gamma_i(v)$ is concentrated around its mean with high probability.

Let $Y = |\Gamma_i(v)|$ be the number of c-nodes outside W that are connected to W by blue edges. $E[Y] = \frac{f}{4}(1 - o(1))$. Let $w_1, w_2, \ldots.$ be an enumeration of the nodes in W. Define an martingale $Z_0, Z_1, \ldots.$, such that $Z_0 = E[Y]$, $Z_i = E[Y \mid N(w_1), \ldots., N(w_i)]$, $Z_w = Y$. Since the degree of all nodes is bounded by C, a node w_i can connect to no more than C nodes outside W. Thus, $|Z_i - Z_{i-1}| < C$.

Using Azuma's inequality it follows that that for sufficiently large constant d,

$$Pr\{|Y - E[Y]| \geq \frac{f}{8}\frac{\sqrt{w}}{C}C\sqrt{w}\} \leq 2e^{-\frac{f^2}{128C^2}w} \leq 1/N^5.$$

□

2.2 Diameter

Our goal is to show that w.h.p the distance between any two c-nodes is $O(\log N)$.

Consider any two c-nodes v and u. By applying expansion lemma repeatedly $O(\log N)$ times we have with probability $1 - O(\frac{\log N}{N^5})$, for some $k_v, k_u = O(\log N)$, $|\Gamma_{k_v}(v)| \geq \sqrt{N}\log N$ and $|\Gamma_{k_u}(u)| \geq \sqrt{N}\log N$. The probability that $\Gamma_{k_v}(v)$ and $\Gamma_{k_u}(u)$ are disjoint and not connected by an edge is bounded by $(1 - f/2N)^{N\log^2 N}$.

Thus with probability $1 - O(\frac{\log N}{N^5})$ an arbitrary pair of nodes u and v are connected by a path of length $O(\log N)$ in G_t. Summing the failure probability over all $\binom{n}{2}$ pairs it follows that w.h.p. any pair of nodes in G_t is connected by a path of length $O(\log N)$.