T-79.7003: Graphs and Networks **Fall 2013**

Lecture 5: Probabilistic tools and Applications II

Lecturer: Charalampos E. Tsourakakis *Oct. 11, 2013*

5.1 Overview

- In the first part of today's lecture we will practice Chernoff bounds. Last time, we saw three problems on which we applied Chernoff bounds: the simple coin tossing problem and two applications on random graphs. Today, we will see three more applications of Chernoff bounds. They illustrate a simple parameter learning problem, approximation algorithm design and the probabilistic method respectively.
- In the second part of the lecture, we will see the Azuma-Hoeffding inequality in the context of discrete time martingales. We will use it in the lecture after the midterm to analyze the degree sequence of the preferential attachment model. Today we will see an application of it on the chromatic number of $G(n, 1/2)$ [Shamir and Spencer, 1987].

5.2 Chernoff Applications

5.2.1 A simple learning problem

Suppose we have a thumbtack and we are interested into the probability of falling with the nail up when we flip it. Our goal is to estimate this probability accurately. Specifically, let p be the probability that we want to estimate, and $\tilde{p}$ the estimator we will use. We will flip the thumbtack n times and we will output as our estimator the average number of times that the thumbtack landed with the nail up. The estimator $\tilde{p}$ is unbiased. To see why notice that after n flips the expected number $\mathbb{E}\left[X_u\right]$ of nail-up tosses X_u is np. It is worth mentioning that this is also the estimator you obtain by maximizing the log-likelihood. Let D be the observed sequence of the n tosses. Let n_u, n_d be the number of tosses with the nail up and down respectively, $n_u + n_d = n$. The principle of maximum likelihood estimation (MLE) picks as the estimator

$$\tilde{p} = \arg\max_p \mathbf{Pr}\left[D|p\right] = \arg\max_p \log \mathbf{Pr}\left[D|p\right].$$

In our setting $\tilde{p} = \arg\max_p \log\left(p^{n_u}(1-p)^{n_d}\right) = \arg\max_p n_u \log p + n_d \log(1-p)$. As you expect, this objective is maximized for $\tilde{p} = \frac{n_u}{n}$.

Our goal is to find how many tosses n we need to perform such that $p, \tilde{p}$ are within distance ϵ with probability at least $1-\delta$. We can use the Chernoff bound to find a lower bound on n.

$$\mathbf{Pr}\left[|p-\tilde{p}| \geq \epsilon\right] \leq \delta = \mathbf{Pr}\left[|X_u - np| \geq \frac{\epsilon}{p} np\right] \leq 2e^{-np\frac{\epsilon^2}{3p^2}} \leq 2e^{-\epsilon^2 n/3}.$$

Hence, it suffices to set n to the smallest possible value that satisfies $2e^{-\epsilon^2 n/3} \leq \delta$, namely $n = \lceil \frac{3}{\epsilon^2} \log\left(\frac{2}{\delta}\right) \rceil$.

5.2.2 Approximation algorithm design

We will see an example of an important technique for designing approximation algorithms. The analysis is based on Chernoff bounds. Let $\mathcal{F} = \{a_1, \dots, a_n\}$ be a family of n binary strings with n bits, i.e., $a_i \in \{0,1\}^n$ for $i = 1, \dots, n$. Our goal is to find a string $s \in \{0,1\}^n$ such that it maximizes the minimum distance from all strings in $\mathcal{F}$. We define the distance of two strings to be their Hamming distance, i.e., the number of coordinates in which they disagree. Our goal is to find

$$s^* = \arg \max_{x \in \{0,1\}^n} \min_{1 \leq i \leq n} dist(x, a_i).$$

We can formulate the problem as an integer program (IP) as follows.

$$\boxed{\begin{array}{rl} & d^* = \textbf{min} \qquad d \\ \textbf{s.t.} & d \geq \displaystyle\sum_{i:a_j(i)=1} (1 - x_i) + \sum_{i:a_j(i)=0} x_i \quad \forall j \in \{1, .., n\} \\ & \textbf{s.t.} \quad x_i \in \{0,1\} \quad \forall i \in \{1, .., n\} \end{array}} \tag{5.1}$$

Due to the hardness of the problem, we relax the constraint $x \in \{0,1\}^n$ to the constraint $x \in [0,1]^n$[1] and therefore we obtain is a linear program (LP).

$$\boxed{\begin{array}{rl} & \textbf{min} \qquad d \\ \textbf{s.t.} & d \geq \displaystyle\sum_{i:a_j(i)=1} (1 - x_i) + \sum_{i:a_j(i)=0} x_i \quad \forall j \in \{1, .., n\} \\ & \textbf{s.t.} \quad 0 \leq x_i \leq 1 \quad \forall i \in \{1, .., n\} \end{array}} \tag{5.2}$$

Relaxation 5.2 is called the *linear programming relaxation* of the integer program 5.1. Notice that s is a feasible solution of the relaxation and therefore the optimal value of the IP is lower bounded by the optimal value of the LP, i.e., $OPT_{LP} \leq OPT_{IP}$. Let's call the optimal vector of the LP x^*. However, the solution of the LP is not going to be a valid solution to the IP: in general some of the coordinates of x^* are not going to be 0 or 1. Therefore, we create a solution $\bar{s}$ by rounding x^*. Let $\bar{d}$ be the resulting objective value. Clearly, since $\bar{s} \in \{0,1\}^n$, $\bar{d} \geq d^*$. To prove an approximation gurantee, we need to upper bound $\bar{d}$ as a function of d^*.

Theorem 5.1 *Consider the following rounding scheme. For each i we flip a coin and with probability x_i^* we set $\bar{s}_i = 1$. With the remaining probability we set $\bar{s}_i^* = 0$. Then,* **whp**

[1] This notation means that the i-th coordinate of x satisfies $0 \leq x_i \leq 1$ for all i.

$$\bar{d} \leq d^* + 3\sqrt{n \log n},$$

where C is a sufficiently large constant.

Proof:

Consider the distance between a fixed string $a_j \in \mathcal{F}$ and the random string $\bar{s}$. For coordinate i we define an indicator variable Y_i which equals 1 if $\bar{s}(i) \neq a_j(i)$. There are two cases when they disagree. If $a_j(i) = 1$ and $\bar{s}(i) = 0$ which happens with probability $1 - x_i^*$, and if $a_j(i) = 0$ and $\bar{s}(i) = 1$ which happens with probability x_i^*. Therefore,

$$\mathbb{E}\left[|\bar{s} - a_j|_1\right] = \sum_{i=1}^{n} \mathbb{E}\left[Y_i\right] = \sum_{i:a_j(i)=1} (1 - x_i^*) + \sum_{i:a_j(i)=0} x_i^*.$$

Notice that since x^* is a feasible solution to the LP, $OPT_{LP} = \bar{d} \geq \mathbb{E}\left[|\bar{s} - a_j|_1\right]$ for all $j = 1, \ldots, m$ by the linear contraints. The Chernoff bound applies to the random variable $Z_j = |\bar{s} - a_j|_1{}^2$.

$$\mathbf{Pr}\left[Z_j \geq OPT_{IP} + C\sqrt{n \log n}\right] \leq \mathbf{Pr}\left[Z_j \geq \mathbb{E}\left[Z_j\right] + C\sqrt{n \log n}\right] \leq n^{-C^2/3}.$$

For $C = 3$ and a union bound over the n strings results in

$$\mathbf{Pr}\left[\exists j : Z_j \geq OPT_{IP} + 3\sqrt{n \log n}\right] = o(1).$$

The claim follows directly.

■

5.2.3 Discrepancy

This section illustrates the probabilistic method and randomized algorithm design in the context of a neat problem. Consider a set system, a.k.a. hypergraph, $(V, \mathcal{F})$ where $V = [n]$ is the ground set and $\mathcal{F} = \{A_1, \ldots, A_m\}$ where $A_i \subseteq V$. We wish to color the ground set V with two colors, say red and blue, in such way that all sets in the family are colored in a "balanced" way, i.e., each set has nearly the same number of red and blue points. As it can be seen from the family $\mathcal{F} = 2^{[n]}$ this is not possible, since by the pidgeonhole principle at least one color will appear at least $n/2$ times and all the possible subsets of those points will be monochromatic. We formalize the above ideas immediately. It shall be convenient to use in the place of red/blue colorings, the coloring

$$\chi : V \rightarrow \{-1, +1\}.$$

For any $A \subseteq V$ define

[2] The previous time we proved the Chernoff bound when all indicator variables were identically distributed as Bernoulli variables with parameter p. This is not the case where, since we have two types of Bernoulli variables appearing in the summation. The Chernoff bound still applies. You are going to prove this in homework 2.

$$\chi(A) = \sum_{i \in A} \chi(i).$$

Define the discrepancy of $\mathcal{F}$ with respect to χ by

$$\text{disc}_{\chi}(\mathcal{F}) = \max_{A_i \in \mathcal{F}} |\chi(A_i)|.$$

The discrepancy of $\mathcal{F}$ is

$$\text{disc}(\mathcal{F}) = \min_{\chi} \text{disc}_{\chi}(\mathcal{F}).$$

We will not make any use of the following observation but it is worth outlining that the discrepancy can be defined in a linear algebraic way. Specifically, let A be the $m \times n$ incidence matrix of $\mathcal{F}$. Then,

$$\text{disc}(\mathcal{F}) = \min_{x \in \{-1,+1\}} ||Ax||_{+\infty}.$$

Let's prove the next theorem by applying union and Chernoff bounds.

Theorem 5.2

$$\mathit{disc}(\mathcal{F}) \leq \sqrt{2n \log{(2m)}}.$$

Proof: Select a coloring χ uniformly at random from the set of all possible random colorings. Let us call A_i bad if its discrepancy exceeds $t = \sqrt{2n \log 2m}$. Applying the Chernoff-Hoeffding bound for set A_i we obtain:

$$\mathbf{Pr}\left[A_i \text{ is bad}\right] = \mathbf{Pr}\left[|\chi(A_i)| > t\right] < 2\exp\Big(-\frac{t^2}{2|A_i|}\Big) \leq 2\exp\Big(-\frac{t^2}{2n}\Big) = \frac{1}{m}.$$

Using a simple union bound we see that

$$\mathbf{Pr}\left[\text{disc}(\mathcal{F}) > t\right] = \mathbf{Pr}\left[\exists \text{ bad } A_i\right] < m \times \frac{1}{m}.$$

■

Theorem 5.2 serves as our basis for a randomized algorithm that succeeds with as high probability as we want. Let $t = \sqrt{2n \log 2m}$. Since the probability of obtaining a coloring that gives discrepancy larger than t is less than $\frac{1}{\sqrt{m}}$, we can boost the success probability by repeating the random coloring k times. The failure probability is at most $\frac{1}{m^{k/2}}$. Assume $m = n$. We have proved that the discrepancy is $O(\sqrt{n \log n})$. Again, for the sake of completeness, a famous result of Joel Spencer states that $\text{disc}(F) = O(\sqrt{n})$.

5.3 Azuma-Hoeffding Inequality

In one of the next lectures, where we are going to work out the degree sequence of the preferential attachment model. The theory of discrete time martingales will be the key to establish concentration.

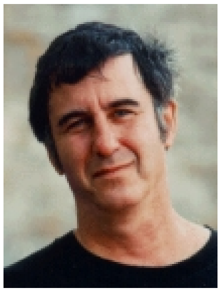

Figure 5.1: Joel Spencer proved his favorite result [Spencer, 1985] known as "six standard deviations suffice" while being in the audience of a talk.

Definition 5.3 (Martingale sequence) *A martingale is a sequence $X_0, \dots, X_n$ of random variables so that for $0 \leq i < n$*

$$\mathbb{E}\left[X_{i+1} | X_0, \dots, X_i\right] = X_i.$$

An easy example of a martingale is related to gambling. Imagine a player who plays a fair game and let X_i be the amount of money he/she has after i rounds, $i \geq 0$. Initially, the player has X_0 dollars. No matter what playing strategy the player will follow, the expected money after $i+1$ rounds is equal to the money after i rounds.

Martingales satisfy concentration inequalities similar to the Chernoff bounds. However, as we saw in the previous lecture, Chernoff bounds apply to a random variable which is the sum of independent random variables. Here, the increments $X_{i+1} - X_i$ are allowed to be dependent.

Theorem 5.4 (Azuma-Hoeffding inequality) *Let $X_0, X_1, \dots, X_n$ be a martingale with $|X_i - X_{i-1}| \leq 1$ for each i. Then*

$$\mathbf{Pr}\left[|X_n - X_0| \geq \alpha\right] \leq 2e^{-\frac{\alpha^2}{2n}}.$$

First, notice that by symmetry ($-X_0, -X_1, \dots$ is also a martingale) it suffices to prove $\mathbf{Pr}\left[X_n - X_0 \geq \alpha\right] \leq e^{-\frac{\alpha^2}{2n}}$. The proof is based on the exponential moment method, but we will control the moment generating function using inductively conditional expectations.

Proof Sketch 5.5 *Using the exponential moment method we get*

$$\mathbf{Pr}\left[X_n - X_0 \geq \alpha\right] \leq \mathbf{Pr}\left[e^{t(X_n - X_0)} \geq e^{t\alpha}\right] \leq e^{-t\alpha}\mathbb{E}\left[e^{t(X_n - X_0)}\right].$$

We obtain a good estimate of $\mathbb{E}\left[e^{t(X_n - X_0)}\right]$ *using the conditional expectation*

$$\mathbb{E}\left[e^{t(X_n - X_0)}\right] = \mathbb{E}\left[e^{t(X_{n-1} - X_0)}\mathbb{E}\left[e^{t(X_n - X_{n-1})} | X_0, \dots, X_{n-1}\right]\right].$$

Then we use the convexity of the function e^{tx} to upper bound $e^{tx} \leq \frac{1+x}{2}e^t + \frac{1-x}{2}e^{-t}$ for $|x| \leq 1$. As you expect, we use $x = X_n - X_{n-1}$. Using the martingale hypothesis, we obtain $\mathbb{E}\left[e^{t(X_n - X_{n-1})} | X_0, \dots, X_{n-1}\right] \leq \cosh(t) \leq e^{t^2/2}$. *By induction on n we get* $\mathbb{E}\left[e^{t(X_n - X_0)}\right] \leq e^{nt^2/2}$. *The claim follows by setting $t = \frac{\alpha}{n}$.*

5.3.1 Vertex and edge exposure martingales

There are two types of martingales we are going to use in this class[3] . Consider the random Erdös-Rényi graph as a vector with $\binom{n}{2}$ coordinates, where the i-th coordinate corresponds to the i-th edge (assume an arbitrary ordering of the edges of the complete graph K_n). Our probability space Ω consists of all possible such vectors. Well, this is not 100% exact. Formally speaking, a probability space is defined by a triple of things: Ω which is the sample space and which we sometimes call probability space when no confusion is caused by this abuse of terminology, $\mathcal{F}$ which is the algebra of all subsets of Ω and the probability measure $\mathbb{P}$. We denote the probability space as $(\Omega, \mathcal{F}, \mathbb{P})$.

In general when we have a probability space Ω and a partition of it $\mathcal{P} = \{P_1, \ldots, P_k\}$, an algebra $\mathcal{A}(\mathcal{P})$ is naturally defined as the family of all unions of the events from $\mathcal{P}$. Vice versa, any algebra of subsets of Ω induces a partition. We will call a partition $\mathcal{P}$ finer than a partition $\mathcal{P}'$ of Ω if $\mathcal{A}(\mathcal{P}') \subseteq \mathcal{A}(\mathcal{P})$. We write $\mathcal{P} < \mathcal{P}'$. A key property is the tower property of conditional expectations: if $\mathcal{P} < \mathcal{P}'$ and X is a random variable defined on Ω then

$$\mathbb{E}\left[X|\mathcal{P}\right] = \mathbb{E}\left[\mathbb{E}\left[X|\mathcal{P}'\right]|\mathcal{P}\right].$$

Suppose we are interested in some random variable $X : \Omega \to \mathbb{R}$. Typically, X will be a graph theoretic invariant such as the chromatic number, see Section 5.3.2. In the case of the vertex exposure martingale, we will *expose* the graph by revealing vertices with the edges incident to them in an arbitrary order. As we expose vertices, the partition of the sample space gets *refined*. This creates a *filtration*, a sequence of refined partitions $\mathcal{F} = \mathcal{P}_0 < \mathcal{P}_1 < \ldots$. We start from the trivial partition $\mathcal{P}_0$ and we end up in the $\mathcal{P}_n$ partition where all the information has been revealed. The sequence $X_0, X_1, \ldots, X_k, \ldots, X_n$ of random variables is a martingale, where $X_k = \mathbb{E}\left[X|\mathcal{P}_k\right]$. This is known as the *exposure martingale* or *Doob martingale*. Intuitively, we obtain in each step information about our graph, the partition of the sample space gets refined, and we are interested in the expectation of our graph invariant during this process. Initially, since we know nothing about the graph $X_0 = \mathbb{E}\left[X|\mathcal{P}_0\right] = \mathbb{E}\left[X\right]$ and when everything has been revealed $X_n = \mathbb{E}\left[X|\mathcal{P}_n\right] = X$. The idea of edge exposure martingale as you expect is the same but rather than revealing all edges incident to each vertex we reveal one edge per time.

The following definition is useful.

Definition 5.6 *A function $f(Z_1, \ldots, Z_n)$ is called c-Lipschitz if when changing the coordinate of any coordinate of f causes f to change by at most $\pm c$.*

The following lemma holds for c-Lipschitz functions.

Lemma 5.7 *If f is c-Lipschitz function and Z_i is independent of $Z_{i+1}, \ldots, Z_n$ conditioned on $Z_1, \ldots, Z_{i-1}$, then the Doob martingale X_i of f with respect to Z_i satisfies $|X_i - X_{i-1}| \leq c$.*

This lemma is useful since if we the conditions of the lemma hold, we can invoke the Azuma-Hoeffding inequality.

5.3.2 Chromatic number of $G(n, \frac{1}{2})$

We start with a reminder. A proper vertex coloring of graph G or just proper coloring is an assignment of colors to the vertex set $V(G)$ in such way that any two adjacent vertices receive different colors. The

[3] Note: this section is presented deliberately in an informal way, insisting more on the intuition rather than the formal details of a filtration.

chromatic number $\chi(G)$ of a graph G is the minimum number of colors required to properly color G.

Theorem 5.8 ([Shamir and Spencer, 1987]) *Let X be the chromatic number of $G \sim G(n, \frac{1}{2})$. Then,*

$$\mathbf{Pr}\left[|X - \mathbb{E}[X]| \geq \lambda\right] \leq e^{-\frac{\lambda^2}{2n}}.$$

Proof: Use the vertex exposure martingale where $X = X(Z_1, \ldots, Z_n)$ where

$$Z_j = \{(i, j) \in E(G) : i < j\}.$$

Then X is 1-Lipschitz. The result follows directly from Azuma-Hoeffding. ■

References

[Shamir and Spencer, 1987] Shamir, E. and Spencer, J. (1987). Sharp concentration of the chromatic number on random graphs g(n,p). *Combinatorica*, 7(1):121–129.

[Spencer, 1985] Spencer, J. (1985). Six standard deviations suffice. *Transactions of the American Mathematical Society*, 289(2):679–706.