PACS Part 2, Lecture 2

Moment-Generating Functions

- Definition: $M_X(t) = \mathbb{E}e^{tX}$
- the expected value always exists for $t = 0$, but is not guaranteed to exist for other t
- if it does exist in a neighborhood of $t = 0$, then we have the following formula, which justifies the name:

$$M_X(t) = \sum_{n=0}^{\infty} \frac{t^n}{n!} \mathbb{E}X^n$$

- in this case, the MGF uniquely determines the distribution of X
- Bernoulli: $M_X(t) = 1 - p + pe^t$
- Geometric: $M_X(t) = pe^t/(1-(1-p)e^t)$ for $t < \log \frac{1}{1-p}$
- if X and Y are independent, then $M_{X+Y} = M_X \cdot M_Y$

Chernoff Bounds

- the general form $\mathbb{P}(X \geq a) = \mathbb{P}(e^{tX} \geq e^{ta}) \leq \mathbb{E}e^{tX}/e^{ta}$ follows from Markov's inequality
- the parameter $t > 0$ is free, so can be optimized upon
- most often used for sums of independent Bernoulli trials $X = \sum_{i=1}^{n} X_i$ with $p_i = \mathbb{E}X_i$ and $\mu = \mathbb{E}X$
- Upper tail bound: $\mathbb{P}(X \geq (1+\delta)\mu) \leq e^{-\mu\delta^2/3}$ for $0 < \delta < 1$

  Proof: general Chernoff bound with $t = \log(1+\delta)$ and then show $e^{\delta}/(1+\delta)^{(1+\delta)} \leq e^{-\delta^2/3}$ for $0 < \delta < 1$
- Lower tail bound: $\mathbb{P}(X \leq (1-\delta)\mu) \leq e^{-\mu\delta^2/2}$ for $0 < \delta < 1$

  Proof: general Chernoff bound with $t = \log(1-\delta)$ and then show $e^{-\delta}/(1-\delta)^{(1-\delta)} \leq e^{-\delta^2/2}$ for $0 < \delta < 1$
- Combined tail bound: $\mathbb{P}(|X - \mu| \geq \delta\mu) \leq 2e^{-\mu\delta^2/3}$
- Application: Parameter estimation

  Let us take n DNA samples and test them for the presence of a certain mutation. We observe $X = \tilde{p}n$ mutations among the samples. Assuming that the presence of the mutation is i.i.d. for each sample, we seek to estimate the real probability p of having the mutation. We would want a confidence interval of the form $\mathbb{P}(|p - \tilde{p}| \leq \delta) \geq 1 - \gamma$. Using the Chernoff bound, we can estimate the error as $\mathbb{P}(|p - \tilde{p}| \geq \delta) = \mathbb{P}(|np - X| \geq \frac{\delta}{p} np) \leq 2e^{-np\delta^2/3p^2} \leq 2e^{-n\delta^2/3} = \gamma$.

- Hoeffding bound: Replaces the Bernoulli assumption by a boundedness assumption. If $a \leq X_i \leq b$ and $\mathbb{E}X_i = \mu/n$ for all i, then:

$$\mathbb{P}(|X - \mu| \geq \delta) \leq 2e^{-2\delta^2/n(b-a)^2}$$

Balls into Bins

- Basic setup: We throw m balls into n bins, independently and uniformly.
- Example: Birthday paradox

  Let us take $m = 30$ people and $n = 365$ possible birthdays. The probability that none of the 30 people have a common birthday is:

$$\prod_{j=1}^{m-1} \left(1 - \frac{j}{n}\right) \approx \prod_{j=1}^{m-1} e^{-j/n} = \exp\left(-\frac{(m-1)m}{2n}\right) \approx e^{-m^2/2n}$$

  To get a constant probability of two people sharing a birthday, it thus suffices to choose $m = \Omega(\sqrt{n})$.

- Poisson approximation: The probability of a given bin having r balls is:

$$\binom{m}{r} \left(\frac{1}{n}\right)^r \left(1 - \frac{1}{n}\right)^{m-r} \approx \frac{e^{-m/n}(m/n)^r}{r!}$$

  In other words, it is approximately a Poisson random variable with parameter $\mu = m/n$.

- We make an error when assuming that the number of balls in every bin is i.i.d. Poisson with parameter $\mu = m/n$, but sometimes that error is not too big.
- Denote by $X_1, \ldots, X_n$ the loads of each bin in the exact case and by $Y_1, \ldots, Y_n$ the loads of each bin in the Poisson case. Then, for every nonnegative function f we have:

$$\mathbb{E}f(X_1, \ldots, X_n) \leq e\sqrt{m}\mathbb{E}f(Y_1, \ldots, Y_n)$$

- The proof uses the law of total probability using the fact that the conditional distribution of $(Y_1, \ldots, Y_n)$ under the condition $\sum_{i=1}^n Y_i = m$ is equal to the distribution of $(X_1, \ldots, X_n)$.
- The special case of an indicator function f is particularly important.

Power of Two Choices

- Question: What is the maximum load of a bin if $m = n$?
- In the basic setup, it is at least $\Omega(\log n / \log\log n)$ with high probability.

  Proof: We first work in the Poisson case. The probability that a bin has load at least $M = \log n / \log\log n$ is at least $1/eM!$. Thus, the probability that all bins have load $< M$ is at most:

  $$p = \left(1 - \frac{1}{eM!}\right)^n \leq e^{-n/eM!}$$

  We verify that, for sufficiently large n, we have $\log M! \leq M \log M + \log M \leq \log n - \log n / \log\log n$, and hence $M! \leq n/2e \log n$ and $p \leq 1/n^2$. Transferring this to the exact case, we get that the probability of a maximum load is at most $e/n^{3/2} \leq 1/n$.

- Next we study d-balanced allocations: Each ball samples d possible bins and is placed in the bin among them with the lowest load. The classical setup is $d = 1$.
- The maximum load in with d-balanced allocations with $d \geq 2$ is at most $O(\log\log n)$ with high probability.

  Proof: Let $h(t)$ be the height of ball t in its bin. Let $\nu_i(t)$ be the number of bins with lead at least i after t throws, $\mu_i(t)$ the number of balls of height at least i after t throws. We have $\nu_i(t) \leq \mu_i(t)$. We want to find β_i such that $\nu_i(n) \leq \beta_i$ with high probability and $\beta_j < 1$ for some $j = O(\log\log n)$.

  Let $\mathcal{E}_i$ be the event $\nu_i(n) \leq \beta_i$. Let $Y_i(t)$ be the following indicator variable:

  $$Y_i(t) = 1 \iff h(t) \geq i + 1 \wedge \nu_i(t-1) \leq \beta_i$$

  Whatever the bin choices $\omega_1, \ldots \omega_{t-1}$ in the first $t-1$ throws, we have

  $$\mathbb{P}(Y_i(t) = 1 \mid \omega_1, \ldots, \omega_{t-1}) \leq \frac{\beta_i^d}{n^d}$$

  since for $h(t) \geq i + 1$ to hold, all d samples need to be taken from the at most β_i suitable bins out of the n bins. To keep one power of n, we set $\beta_{i+1} = 2\beta_i^d / n^{d-1}$ and initialize $\beta_4 = n/4$. With this initialization, we have $\mathbb{P}(\mathcal{E}_4) = 1$.

  By the law of total probability, we have $\mathbb{P}\left(\sum_{t=1}^n Y_i(t) > k\right) \leq \mathbb{P}\left(\sum_{t=1}^n Z_t > k\right)$ where the Z_t are i.i.d. Bernoulli trials with success parameter $p_i = \beta_i^d / n^d$. The event $\mathcal{E}_i$ implies $\sum_{t=1}^n Y_i(t) = \mu_{i+1}(n) \geq \nu_{i+1}(n)$, and thus

  $$\mathbb{P}(\neg\mathcal{E}_{i+1} \mid \mathcal{E}_i) \leq \frac{\mathbb{P}\left(\sum_{t=1}^n Z_t > 2np_i\right)}{\mathbb{P}(\mathcal{E}_i)} \leq \frac{e^{-p_i n/3}}{\mathbb{P}(\mathcal{E}_i)} \leq \frac{1}{n^2 \mathbb{P}(\mathcal{E}_i)}$$

by the Chernoff bound if $p_i n \geq 6 \log n$. This then implies $\mathbb{P}(\neg\mathcal{E}_{i+1}) \leq \mathbb{P}(\neg\mathcal{E}_i) + 1/n^2$, i.e., $\mathcal{E}_i$ holds with high probability.

The condition $p_i n \geq 6 \log n$ restricts i to be $O(\log \log n)$ since $\beta_{i+4} \leq n/2^{d^i}$ as long as it is true. Let i^* be the smallest value that violates the condition. It is $\beta_{i^*} = O(\log n)$. One can prove that $\mathbb{P}(\nu_{i^*+3}(n) \geq 1) = O(1/n)$ by another Chernoff bound and the union bound.

But then, with high probability, there is no bin with load higher than $i^* + 3 = O(\log \log n)$.