Chapter 7

Concentration of Random Variables – Chernoff's Inequality

By Sariel Har-Peled, April 14, 2016[①]

7.1. Concentration of mass and Chernoff's inequality

7.1.1. Example: Binomial distribution

Consider the binomial distribution $\mathrm{Bin}(n, 1/2)$ for various values of n as depicted in Figure 7.1 – here we think about the value of the variable as the number of heads in flipping a fair coin n times. Clearly, as the value of n increases the probability of getting a number of heads that is significantly smaller or larger than $n/2$ is tiny. Here we are interested in quantifying exactly how far can we divert from this expected value. Specifically, if $X \sim \mathrm{Bin}(n, 1/2)$, then we would be interested in bounding the probability $\mathbf{Pr}[X > n/2 + \Delta]$, where $\Delta = t\sigma_X = t\sqrt{n}/2$ (i.e., we are t standard deviations away from the expectation). For $t > 2$, this probability is roughly 2^{-t}, which is what we prove here.

More surprisingly, if you look only on the middle of the distribution, it looks the same after clipping away the uninteresting tails, see Figure 7.2; that is, it looks more and more like the normal distribution. This is a universal phenomena known the *central limit theorem* – every sum of nicely behaved random variables behaves like the normal distribution. We unfortunately need a more precise quantification of this behavior, thus the following.

7.1.2. A restricted case of Chernoff inequality via games

7.1.2.1. Chernoff games

7.1.2.1.1. The game. Consider the game where a player starts with $Y_0 = 1$ dollars. At every round, the player can bet a certain amount x (fractions are fine). With probability half she loses her bet, and with probability half she gains an amount equal to her bet. The player is not allowed to go all in – because if she looses then the game is over. So it is natural to ask what her optimal betting strategy is, such that in the end of the game she has as much money as possible.

[①]This work is licensed under the Creative Commons Attribution-Noncommercial 3.0 License. To view a copy of this license, visit http://creativecommons.org/licenses/by-nc/3.0/ or send a letter to Creative Commons, 171 Second Street, Suite 300, San Francisco, California, 94105, USA.

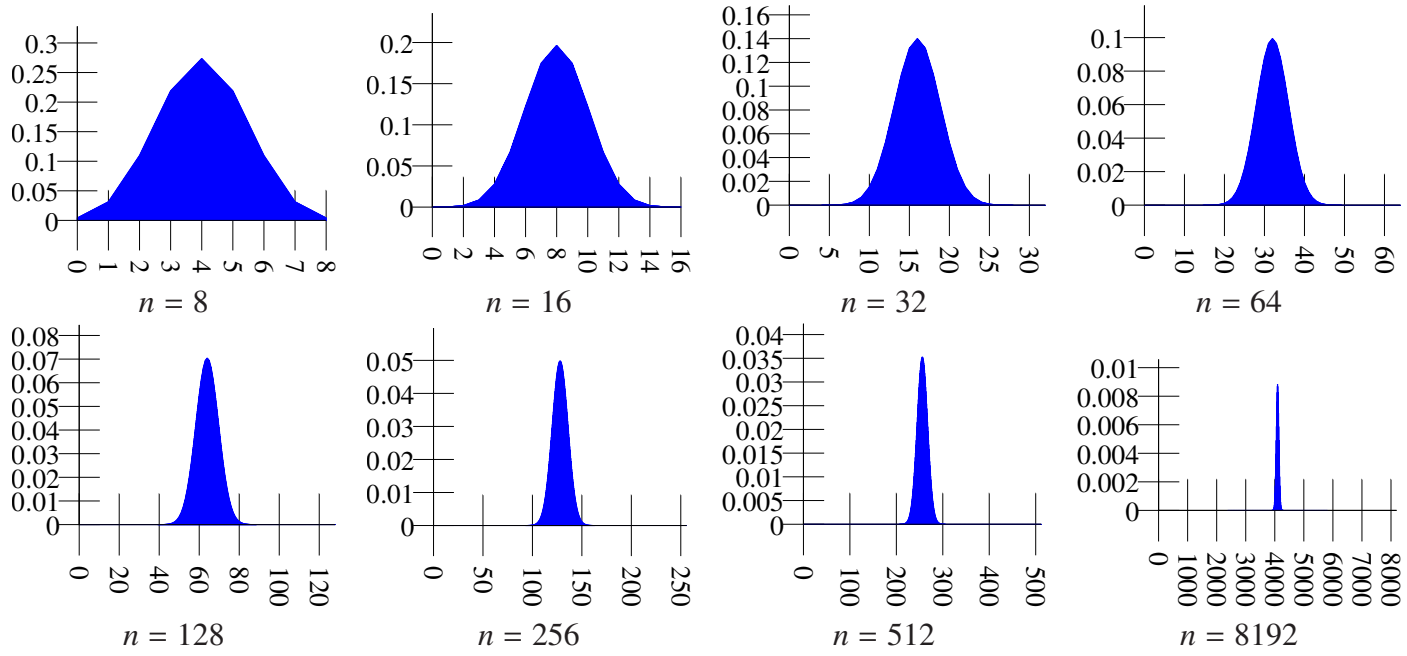

Figure 7.1: The binomial distribution for different values of n. It pretty quickly concentrates around its expectation.

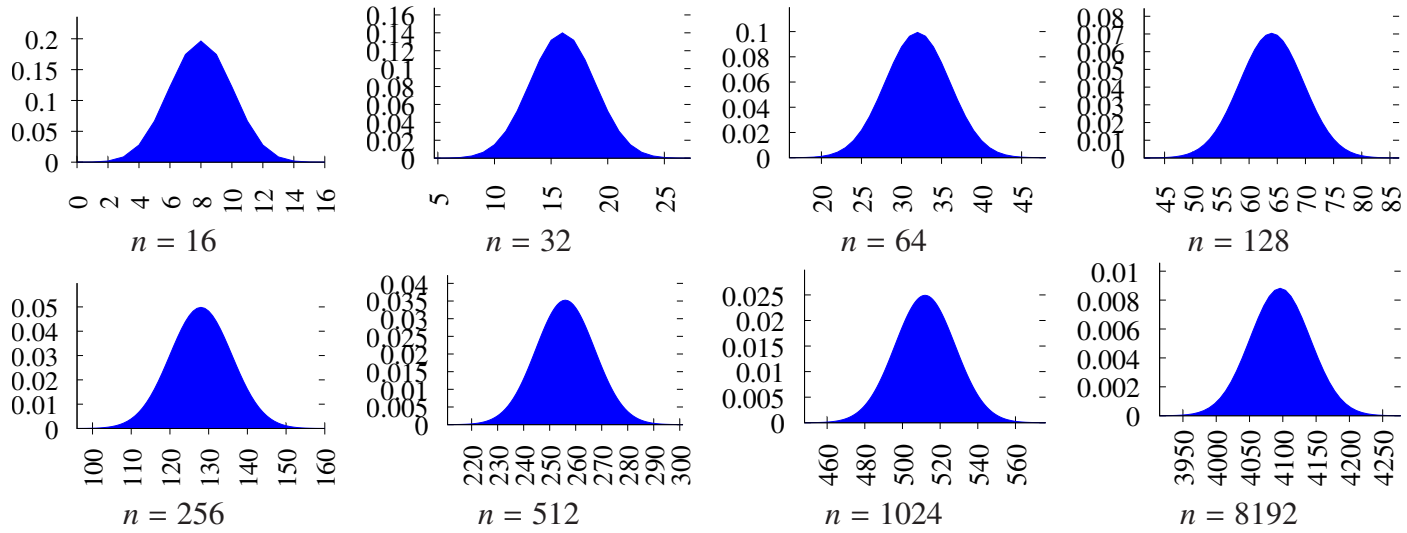

Figure 7.2: The "middle" of the binomial distribution for different values of n. It very quickly converges to the normal distribution (under appropriate rescaling and translation.

Values	Probabilities	Inequality	Ref
$-1, +1$	$\mathbf{Pr}[X_i = -1] =$ $\mathbf{Pr}[X_i = 1] = 1/2$	$\mathbf{Pr}\Big[Y \geq \Delta\Big] \leq \exp\Big(-\Delta^2/2n\Big)$ $\mathbf{Pr}\Big[Y \leq -\Delta\Big] \leq \exp\Big(-\Delta^2/2n\Big)$ $\mathbf{Pr}\Big[\lvert Y\rvert \geq \Delta\Big] \leq 2\exp\Big(-\Delta^2/2n\Big)$	Theorem 7.1.7$_{\text{p6}}$ Theorem 7.1.7$_{\text{p6}}$ Corollary 7.1.8$_{\text{p7}}$
$0, 1$	$\mathbf{Pr}[X_i = 0] =$ $\mathbf{Pr}[X_i = 1] = 1/2$	$\mathbf{Pr}\Big[\Big\lvert Y - \frac{n}{2}\Big\rvert \geq \Delta\Big] \leq 2\exp\Big(-2\Delta^2/n\Big)$	Corollary 7.1.9$_{\text{p7}}$
0,1	$\mathbf{Pr}[X_i = 0] = 1 - p_i$ $\mathbf{Pr}[X_i = 1] = p_i$	$\mathbf{Pr}\Big[Y > (1+\delta)\mu\Big] < \Big(\frac{e^\delta}{(1+\delta)^{1+\delta}}\Big)^\mu$	Theorem 7.3.2$_{\text{p12}}$
	For $\delta \leq 2e - 1$ $\delta \geq 2e - 1$ $\delta \geq e^2$	$\mathbf{Pr}[Y > (1+\delta)\mu] < \exp\Big(-\mu\delta^2/4\Big)$ $\mathbf{Pr}[Y > (1+\delta)\mu] < 2^{-\mu(1+\delta)}$ $\mathbf{Pr}[Y > (1+\delta)\mu] < \exp\Big(-(\mu\delta/2)\ln\delta\Big)$	Theorem 7.3.2$_{\text{p12}}$
	For $\delta \geq 0$	$\mathbf{Pr}[Y < (1-\delta)\mu] < \exp\Big(-\mu\delta^2/2\Big)$	Theorem 7.3.5$_{\text{p13}}$
$X_i \in [0, 1]$	X_is have arbitrary independent distributions.	$\mathbf{Pr}\Big[Y - \mu \geq \varepsilon\mu\Big] \leq \exp\Big(-\varepsilon^2\mu/4\Big)$ $\mathbf{Pr}\Big[Y - \mu \leq -\varepsilon\mu\Big] \leq \exp\Big(-\varepsilon^2\mu/2\Big).$	Theorem 7.4.5$_{\text{p15}}$
$X_i \in [a_i, b_i]$	X_is have arbitrary independent distributions.	$\mathbf{Pr}\Big[\Big\lvert Y - \mu\Big\rvert \geq \eta\Big] \leq 2\exp\Big(-\frac{2\eta^2}{\sum_{i=1}^n (b_i - a_i)^2}\Big)$	Theorem 7.5.3$_{\text{p18}}$

Table 7.1: Summary of Chernoff type inequalities covered. Here we have n independent random variables $X_1, \ldots, X_n$, $Y = \sum_i X_i$ and $\mu = \mathbf{E}[Y]$.

7.1.2.1.2. Is the game pointless? So, let Y_{i-1} be the money the player has in the end of the $(i-1)$th round, and she bets an amount $\psi_i \le Y_{i-1}$ in the ith round. As such, in the end of the ith round, she has

$$Y_i = \begin{cases} Y_{i-1} - \psi_i & \text{LOSE: probability half} \\ Y_{i-1} + \psi_i & \text{WIN: probability half} \end{cases}$$

dollars. This game, in expectation, does not change the amount of money the player has. Indeed, we have

$$\mathbf{E}\Big[Y_i \;\Big|\; Y_{i-1}\Big] = \frac{1}{2}(Y_{i-1} - \psi_i) + \frac{1}{2}(Y_{i-1} + \psi_i) = Y_{i-1}.$$

And as such, we have that $\mathbf{E}[Y_i] = \mathbf{E}\Big[\mathbf{E}\Big[Y_i \Big| Y_{i-1}\Big]\Big] = \mathbf{E}[Y_{i-1}] = \cdots = \mathbf{E}[Y_0] = 1$. In particular, $\mathbf{E}[Y_n] = 1$ – namely, on average, independent of the player strategy she is not going to make any money in this game (and she is allowed to change her bets after every round). Unless, she is lucky[②]...

7.1.2.1.3. What about a lucky player? The player believes she will get lucky and wants to develop a strategy to take advantage of it. Formally, she believes that she can win, say, at least $(1+\delta)/2$ fraction of her bets (instead of the predicted $1/2$) – for example, if the bets are in the stock market, she can improve her chances by doing more research on the companies she is investing in[③]. Unfortunately, the player does not know which rounds she is going to be lucky in – so she still needs to be careful.

7.1.2.1.4. In a search of a good strategy. Of course, there are many safe strategies the player can use, from not playing at all, to risking only a tiny fraction of her money at each round. In other words, our quest here is to find the best strategy that extracts the maximum benefit for the player out of her inherent luck.

Here, we restrict ourselves to a simple strategy – at every round, the player would bet β fraction of her money, where β is a parameter to be determined. Specifically, in the end of the ith round, the player would have

$$Y_i = \begin{cases} (1-\beta)Y_{i-1} & \text{LOSE} \\ (1+\beta)Y_{i-1} & \text{WIN.} \end{cases}$$

By our assumption, the player is going to win in at least $M = (1+\delta)n/2$ rounds. Our purpose here is to figure out what the value of β should be so that player gets as rich as possible[④]. Now, if the player is successful in $\ge M$ rounds, out of the n rounds of the game, then the amount of money the player has, in the end of the game, is

$$\begin{aligned} Y_n &\ge (1-\beta)^{n-M}(1+\beta)^M = (1-\beta)^{n/2-(\delta/2)n}(1+\beta)^{n/2+(\delta/2)n} = \Big((1-\beta)(1+\beta)\Big)^{n/2-(\delta/2)n}(1+\beta)^{\delta n} \\ &= \Big(1-\beta^2\Big)^{n/2-(\delta/2)n}(1+\beta)^{\delta n} \ge \exp\Big(-2\beta^2\Big)^{n/2-(\delta/2)n}\exp(\beta/2)^{\delta n} = \exp\Big(\Big(-\beta^2+\beta^2\delta+\beta\delta/2\Big)n\Big). \end{aligned}$$

To maximize this quantity, we choose $\beta = \delta/4$ (there is a better choice, see Lemma 7.1.6, but we use this value for the simplicity of exposition). Thus, we have that $Y_n \ge \exp\left(\left(-\frac{\delta^2}{16} + \frac{\delta^3}{16} + \frac{\delta^2}{8}\right)n\right) \ge \exp\left(\frac{\delta^2}{16}n\right)$, proving the following.

Lemma 7.1.1. *Consider a Chernoff game with n rounds, starting with one dollar, where the player wins in $\ge (1+\delta)n/2$ of the rounds. If the player bets $\delta/4$ fraction of her current money, at all rounds, then in the end of the game the player would have at least $\exp\big(n\delta^2/16\big)$ dollars.*

[②] "I would rather have a general who was lucky than one who was good." – Napoleon Bonaparte.

[③] "I am a great believer in luck, and I find the harder I work, the more I have of it." – Thomas Jefferson.

[④] This optimal choice is known as Kelly criterion, see Remark 7.1.3.

Remark 7.1.2. Note, that Lemma 7.1.1 holds if the player wins any $\geq (1+\delta)n/2$ rounds. In particular, the statement does not require randomness by itself – for our application, however, it is more natural and interesting to think about the player wins as being randomly distributed.

Remark 7.1.3. Interestingly, the idea of choosing the best fraction to bet is an old and natural question arising in investments strategies, and the right fraction to use is known as *Kelly criterion*, going back to Kelly's work from 1956 [Kel56].

7.1.2.2. Chernoff's inequality

The above implies that if a player is lucky, then she is going to become filthy rich[5]. Intuitively, this should be a pretty rare event – because if the player is rich, then (on average) many other people have to be poor. We are thus ready for the kill.

Theorem 7.1.4 (Chernoff's inequality). *Let $X_1, \ldots, X_n$ be n independent random variables, where $X_i = 0$ or $X_i = 1$ with equal probability. Then, for any $\delta \in (0, 1/2)$, we have that*

$$\mathbf{Pr}\left[\sum_i X_i \geq (1+\delta)\frac{n}{2}\right] \leq \exp\left(-\frac{\delta^2}{16}n\right).$$

Proof: Imagine that we are playing the Chernoff game above, with $\beta = \delta/4$, starting with 1 dollar, and let Y_i be the amount of money in the end of the ith round. Here $X_i = 1$ indicates that the player won the ith round. We have, by Lemma 7.1.1 and Markov's inequality, that

$$\mathbf{Pr}\left[\sum_i X_i \geq (1+\delta)\frac{n}{2}\right] \leq \mathbf{Pr}\left[Y_n \geq \exp\left(\frac{n\delta^2}{16}\right)\right] \leq \frac{\mathbf{E}[Y_n]}{\exp(n\delta^2/16)} = \frac{1}{\exp(n\delta^2/16)} = \exp\left(-\frac{\delta^2}{16}n\right),$$

as claimed. ■

7.1.2.2.1. This is crazy – so intuition maybe? If the player is $(1+\delta)/2$-lucky then she can make a lot of money; specifically, at least $f(\delta) = \exp\left(n\delta^2/16\right)$ dollars by the end of the game. Namely, beating the odds has significant monetary value, and this value grows quickly with δ. Since we are in a "zero-sum" game settings, this event should be very rare indeed. Under this interpretation, of course, the player needs to know in advance the value of δ – so imagine that she guesses it somehow in advance, or she plays the game in parallel with all the possible values of δ, and she settles on the instance that maximizes her profit.

7.1.2.2.2. Can one do better? No, not really. Chernoff inequality is tight (this is a challenging homework exercise) up to the constant in the exponent. The best bound I know for this version of the inequality has $1/2$ instead of $1/16$ in the exponent. Note, however, that no real effort was taken to optimize the constants – this is not the purpose of this write-up.

7.1.2.3. Some low level boring calculations

Above, we used the following well known facts.

[5] Not that there is anything wrong with that – many of my friends are filthy,

Lemma 7.1.5. *(A) Markov's inequality. For any positive random variable X and $t > 0$, we have* $\mathbf{Pr}[X \geq t] \leq \mathbf{E}[X]/t$.

(B) For any two random variables X and Y, we have that $\mathbf{E}\Big[X\Big] = \mathbf{E}\Big[\mathbf{E}\Big[X\,\Big|Y\Big]\Big]$.

(C) For $x \in (0,1)$, $1 + x \geq e^{x/2}$.

(D) For $x \in (0, 1/2)$, $1 - x \geq e^{-2x}$.

Lemma 7.1.6. *The quantity* $\exp\Big(\Big(-\beta^2 + \beta^2\delta + \beta\delta/2\Big)n\Big)$ *is maximal for* $\beta = \frac{\delta}{4(1-\delta)}$.

Proof: We have to maximize $f(\beta) = -\beta^2 + \beta^2\delta + \beta\delta/2$ by choosing the correct value of β (as a function of δ, naturally). $f'(\beta) = -2\beta + 2\beta\delta + \delta/2 = 0 \quad\Longleftrightarrow\quad 2(\delta-1)\beta = -\delta/2 \quad\Longleftrightarrow\quad \beta = \frac{\delta}{4(1-\delta)}$. ■

7.1.3. Chernoff Inequality - A Special Case – the classical proof

Theorem 7.1.7. *Let $X_1, \ldots, X_n$ be n independent random variables, such that* $\mathbf{Pr}[X_i = 1] = \mathbf{Pr}[X_i = -1] = \frac{1}{2}$, *for $i = 1, \ldots, n$. Let $Y = \sum_{i=1}^{n} X_i$. Then, for any $\Delta > 0$, we have*

$$\mathbf{Pr}\Big[Y \geq \Delta\Big] \leq \exp\Big(-\Delta^2/2n\Big).$$

Proof: Clearly, for an arbitrary t, to specified shortly, we have

$$\mathbf{Pr}[Y \geq \Delta] = \mathbf{Pr}[\exp(tY) \geq \exp(t\Delta)] \leq \frac{\mathbf{E}[\exp(tY)]}{\exp(t\Delta)},$$

the first part follows by the fact that $\exp(\cdot)$ preserve ordering, and the second part follows by the Markov inequality.

Observe that

$$\begin{aligned}
\mathbf{E}[\exp(tX_i)] &= \frac{1}{2}e^{t} + \frac{1}{2}e^{-t} = \frac{e^t + e^{-t}}{2} \\
&= \frac{1}{2}\left(1 + \frac{t}{1!} + \frac{t^2}{2!} + \frac{t^3}{3!} + \cdots\right) \\
&+ \frac{1}{2}\left(1 - \frac{t}{1!} + \frac{t^2}{2!} - \frac{t^3}{3!} + \cdots\right) \\
&= \left(1 + \quad + \frac{t^2}{2!} + \quad + \cdots + \frac{t^{2k}}{(2k)!} + \cdots\right),
\end{aligned}$$

by the Taylor expansion of $\exp(\cdot)$. Note, that $(2k)! \geq (k!)2^k$, and thus

$$\mathbf{E}[\exp(tX_i)] = \sum_{i=0}^{\infty} \frac{t^{2i}}{(2i)!} \leq \sum_{i=0}^{\infty} \frac{t^{2i}}{2^i(i!)} = \sum_{i=0}^{\infty} \frac{1}{i!}\left(\frac{t^2}{2}\right)^i = \exp\Big(t^2/2\Big),$$

again, by the Taylor expansion of $\exp(\cdot)$. Next, by the independence of the X_is, we have

$$\mathbf{E}[\exp(tY)] = \mathbf{E}\left[\exp\left(\sum_i tX_i\right)\right] = \mathbf{E}\left[\prod_i \exp(tX_i)\right] = \prod_{i=1}^{n} \mathbf{E}[\exp(tX_i)] \leq \prod_{i=1}^{n} e^{t^2/2} = e^{nt^2/2}.$$

We have $\mathbf{Pr}[Y \geq \Delta] \leq \dfrac{\exp\Big(nt^2/2\Big)}{\exp(t\Delta)} = \exp\Big(nt^2/2 - t\Delta\Big)$.

Next, by minimizing the above quantity for t, we set $t = \Delta/n$. We conclude,

$$\mathbf{Pr}[Y \geq \Delta] \leq \exp\left(\frac{n}{2}\left(\frac{\Delta}{n}\right)^2 - \frac{\Delta}{n}\Delta\right) = \exp\left(-\frac{\Delta^2}{2n}\right).$$ ■

By the symmetry of Y, we get the following:

Corollary 7.1.8. *Let $X_1, \ldots, X_n$ be n independent random variables, such that $\mathbf{Pr}[X_i = 1] = \mathbf{Pr}[X_i = -1] = \frac{1}{2}$, for $i = 1, \ldots, n$. Let $Y = \sum_{i=1}^{n} X_i$. Then, for any $\Delta > 0$, we have*

$$\mathbf{Pr}[|Y| \geq \Delta] \leq 2e^{-\Delta^2/2n}.$$

Corollary 7.1.9. *Let $X_1, \ldots, X_n$ be n independent* coin flips*, such that $\mathbf{Pr}[X_i = 0] = \mathbf{Pr}[X_i = 1] = \frac{1}{2}$, for $i = 1, \ldots, n$. Let $Y = \sum_{i=1}^{n} X_i$. Then, for any $\Delta > 0$, we have*

$$\mathbf{Pr}\left[\left|Y - \frac{n}{2}\right| \geq \Delta\right] \leq 2e^{-2\Delta^2/n}.$$

Remark 7.1.10. Before going any further, it is might be instrumental to understand what this inequalities imply. Consider then case where X_i is either zero or one with probability half. In this case $\mu = \mathbf{E}[Y] = n/2$. Set $\delta = t\sqrt{n}$ ($\sqrt{\mu}$ is approximately the standard deviation of X if $p_i = 1/2$). We have by

$$\mathbf{Pr}\left[\left|Y - \frac{n}{2}\right| \geq \Delta\right] \leq 2\exp\left(-2\Delta^2/n\right) = 2\exp\left(-2(t\sqrt{n})^2/n\right) = 2\exp\left(-2t^2\right).$$

Thus, Chernoff inequality implies exponential decay (i.e., $\leq 2^{-t}$) with t standard deviations, instead of just polynomial (i.e., $\leq 1/t^2$) by the Chebychev's inequality.

7.2. Applications of Chernoff's inequality

There is a zoo of Chernoff type inequalities, and prove some of them later on the chapter – while being very useful and technically interesting, they tend to numb to reader into boredom and submission. As such, we discuss applications of Chernoff's inequality here, and the interested reader can read the proofs of the more general forms only if they are interested in them.

7.2.1. QuickSort is Quick

We revisit **QuickSort**. We remind the reader that the running time of **QuickSort** is proportional to the number of comparisons performed by the algorithm. Next, consider an arbitrary element u being sorted. Consider the ith level recursive subproblem that contains u, and let S_i be the set of elements in this subproblems. We consider u to be *successful* in the ith level, if $|S_{i+1}| \leq |S_i|/2$. Namely, if u is successful, then the next level in the recursion involving u would include a considerably smaller subproblem. Let X_i be the indicator variable which is 1 if u is successful.

We first observe that if **QuickSort** is applied to an array with n elements, then u can be successful at most $T = \lceil \lg n \rceil$ times, before the subproblem it participates in is of size one, and the recursion stops. Thus, consider the indicator variable X_i which is 1 if u is successful in the ith level, and zero otherwise. Note that the X_is are independent, and $\mathbf{Pr}[X_i = 1] = 1/2$.

If u participates in v levels, then we have the random variables $X_1, X_2, \ldots, X_v$. To make things simpler, we will extend this series by adding independent random variables, such that $\mathbf{Pr}[`]X_i = 1 = 1/2$, for $i \geq v$. Thus, we have an infinite sequence of independent random variables, that are 0/1 and get 1 with probability $1/2$. The question is how many elements in the sequence we need to read, till we get T ones.

Lemma 7.2.1. *Let $X_1, X_2, \ldots$ be an infinite sequence of independent random 0/1 variables. Let M be an arbitrary parameter. Then the probability that we need to read more than $2M + 4t\sqrt{M}$ variables of this sequence till we collect M ones is at most $2\exp\left(-t^2\right)$, for $t \leq \sqrt{M}$. If $t \geq \sqrt{M}$ then this probability is at most $2\exp\left(-t\sqrt{M}\right)$.*

Proof: Consider the random variable $Y = \sum_{i=1}^{L} X_i$, where $L = 2M + 4t\sqrt{M}$. Its expectation is $L/2$, and using the Chernoff inequality, we get

$$\begin{aligned}\alpha = \mathbf{Pr}\left[Y \leq M\right] &\leq \mathbf{Pr}\left[\left|Y - \frac{L}{2}\right| \geq \frac{L}{2} - M\right] \leq 2\exp\left(-\frac{2}{L}\left(\frac{L}{2} - M\right)^2\right) \\ &\leq 2\exp\left(-\frac{2}{L}\left(M + 2t\sqrt{M} - M\right)^2\right) \leq 2\exp\left(-\frac{2}{L}\left(2t\sqrt{M}\right)^2\right) = 2\exp\left(-\frac{8t^2M}{L}\right),\end{aligned}$$

by Corollary 7.1.9. For $t \leq \sqrt{M}$ we have that $L = 2M + 4t\sqrt{M} \leq 8M$, as such in this case $\mathbf{Pr}[Y \leq M] \leq 2\exp\left(-t^2\right)$.

If $t \geq \sqrt{M}$, then $\alpha = 2\exp\left(-\dfrac{8t^2M}{2M + 4t\sqrt{M}}\right) \leq 2\exp\left(-\dfrac{8t^2M}{6t\sqrt{M}}\right) \leq 2\exp\left(-t\sqrt{M}\right)$. ■

Going back to the **QuickSort** problem, we have that if we sort n elements, the probability that u will participate in more than $L = (4 + c)\lceil \lg n \rceil = 2\lceil \lg n \rceil + 4c\sqrt{\lg n}\sqrt{\lg n}$, is smaller than $2\exp\left(-c\sqrt{\lg n}\sqrt{\lg n}\right) \leq 1/n^c$, by Lemma 7.2.1. There are n elements being sorted, and as such the probability that any element would participate in more than $(4 + c + 1)\lceil \lg n \rceil$ recursive calls is smaller than $1/n^c$.

Lemma 7.2.2. *For any $c > 0$, the probability that* **QuickSort** *performs more than $(6 + c)n \lg n$, is smaller than $1/n^c$.*

7.2.2. How many times can the minimum change?

Let $\Pi = \pi_1 \ldots \pi_n$ be a random permutation of $\{1, \ldots, n\}$. Let $\mathcal{E}_i$ be the event that π_i is the minimum number seen so far as we read Π; that is, $\mathcal{E}_i$ is the event that $\pi_i = \min_{k=1}^{i} \pi_k$. Let X_i be the indicator variable that is one if $\mathcal{E}_i$ happens. We already seen, and it is easy to verify, that $\mathbf{E}[X_i] = 1/i$. We are interested in how many times the minimum might change[6]; that is $Z = \sum_i X_i$, and how concentrated is the distribution of Z. The following is maybe surprising.

Lemma 7.2.3. *The events $\mathcal{E}_1, \ldots, \mathcal{E}_n$ are independent (as such, variables $X_1, \ldots, X_n$ are independent).*

Proof: The trick is to think about the sampling process in a different way, and then the result readily follows. Indeed, we randomly pick a permutation of the given numbers, and set the first number to be π_n. We then, again, pick a random permutation of the remaining numbers and set the first number as the penultimate number (i.e., π_{n-1}) in the output permutation. We repeat this process till we generate the whole permutation.

Now, consider $1 \leq i_1 < i_2 < \ldots < i_k \leq n$, and observe that $\mathbf{Pr}\left[\mathcal{E}_{i_k} \,\middle|\, \mathcal{E}_{i_1} \cap \ldots \cap \mathcal{E}_{i_{k-1}}\right] = \mathbf{Pr}[\mathcal{E}_{i_k}]$, since by our thought experiment, $\mathcal{E}_{i_k}$ is determined before all the other variables $\mathcal{E}_{i_{k-1}}, \ldots, \mathcal{E}_{i_1}$, and these variables are inherently not effected by this event happening or not. As such, we have

$$\begin{aligned}\mathbf{Pr}\left[\mathcal{E}_{i_1} \cap \mathcal{E}_{i_2} \cap \ldots \cap \mathcal{E}_{i_k}\right] &= \mathbf{Pr}\left[\mathcal{E}_{i_k} \,\middle|\, \mathcal{E}_{i_1} \cap \ldots \cap \mathcal{E}_{i_{k-1}}\right] \mathbf{Pr}[\mathcal{E}_{i_1} \cap \ldots \cap \mathcal{E}_{i_{k-1}}] \\ &= \mathbf{Pr}\left[\mathcal{E}_{i_k}\right] \mathbf{Pr}\left[\mathcal{E}_{i_1} \cap \mathcal{E}_{i_2} \cap \ldots \cap \mathcal{E}_{i_{k-1}}\right] = \prod_{j=1}^{k} \mathbf{Pr}\left[\mathcal{E}_{i_j}\right] = \prod_{j=1}^{k} \frac{1}{i_j},\end{aligned}$$

by induction. ■

[6] The answer, my friend, is blowing in the permutation.

Theorem 7.2.4. *Let $\Pi = \pi_1 \dots \pi_n$ be a random permutation of $1, \dots, n$, and let Z be the number of times, that π_i is the smallest number among $\pi_1, \dots, \pi_i$, for $i = 1, \dots, n$. Then, we have that for $t \geq 2e$ that $\mathbf{Pr}[Z > t \ln n] \leq 1/n^{t \ln 2}$, and for $t \in [1, 2e]$, we have that $\mathbf{Pr}[Z > t \ln n] \leq 1/n^{(t-1)^2/4}$.*

Proof: Follows readily from Chernoff's inequality, as $Z = \sum_i X_i$ is a sum of independent indicator variables, and, since by linearity of expectations, we have

$$\mu = \mathbf{E}\big[Z\big] = \sum_i \mathbf{E}\big[X_i\big] = \sum_{i=1}^{n} \frac{1}{i} \geq \int_{x=1}^{n+1} \frac{1}{x} \mathrm{d}x = \ln(n+1) \geq \ln n.$$

Next, we set $\delta = t - 1$, and use Theorem 7.3.2$_{\text{p12}}$. ■

7.2.3. Routing in a Parallel Computer

Let G be a graph of a network, where every node is a processor. The processor communicate by sending packets on the edges. Let $\big[0, \dots, N-1\big]$ denote be vertices (i.e., processors) of G, where $N = 2^n$, and G is the hypercube. As such, each processes is identified with a binary string $b_1 b_2 \dots b_n \in \{0,1\}^n$. Two nodes are connected if their binary string differs only in a single bit. Namely, G is the binary *hypercube* over n bits.

We want to investigate the best routing strategy for this topology of network. We assume that every processor need to send a message to a single other processor. This is represented by a permutation π, and we would like to figure out how to send the messages encoded by the permutation while create minimum delay/congestion.

Specifically, in our model, every edge has a FIFO queue[7] of the packets it has to transmit. At every clock tick, one message get sent. All the processors start sending the packets in their permutation in the same time.

A routing scheme is *oblivious* if every node that has to forward a packet, inspect the packet, and depending only on the content of the packet decides how to forward it. That is, such a routing scheme is local in nature, and does not take into account other considerations. Oblivious routing is of course a bad idea – it ignores congestion in the network, and might insist routing things through regions of the hypercube that are "gridlocked".

Theorem 7.2.5 ([KKT91]). *For any deterministic oblivious permutation routing algorithm on a network of N nodes each of out-degree n, there is a permutation for which the routing of the permutation takes $\Omega\Big(\sqrt{N/n}\Big)$ units of time (i.e., ticks).*

Proof: (Sketch.) The above is implied by a nice averaging argument – construct, for every possible destination, the routing tree of all packets to this specific node. Argue that there must be many edges in this tree that are highly congested in this tree (which is NOT the permutation routing we are looking for!). Now, by averaging, there must be a single edge that is congested in "many" of these trees. Pick a source-destination pair from each one of these trees that uses this edge, and complete it into a full permutation in the natural way. Clearly, the congestion of the resulting permutation is high. For the exact details see [KKT91]. ■

7.2.3.0.1. How do we send a packet? We use *bit fixing*. Namely, the packet from the i node, always go to the current adjacent node that have the first different bit as we scan the destination string $d(i)$. For example, packet from (0000) going to (1101), would pass through (1000), (1100), (1101).

7.2.3.0.2. The routing algorithm. We assume each edge have a FIFO queue. The routing algorithm is depicted in Figure 7.3.

[7] First in, first out queue. I sure hope you already knew that.

RandomRoute($v_0, \ldots, v_{N-1}$)

// v_i: Packet at node i to be routed to node $d(i)$.

(i) Pick a *random* intermediate destination $\sigma(i)$ from $[1, \ldots, N]$. Packet v_i travels to $\sigma(i)$.

// Here random sampling is done with ***replacement***.

// Several packets might travel to the same destination.

(ii) Wait till all the packets arrive to their intermediate destination.

(iii) Packet v_i travels from $\sigma(i)$ to its destination $d(i)$.

Figure 7.3: The routing algorithm

7.2.3.1. Analysis

We analyze only (i) as (iii) follows from the same analysis. In the following, let ρ_i denote the route taken by v_i in (i).

Exercise 7.2.6. Once a packet v_j that travel along a path ρ_j can not leave a path ρ_i, and then join it again later. Namely, $\rho_i \cap \rho_j$ is (maybe an empty) path.

Lemma 7.2.7. *Let the route of a message* **c** *follow the sequence of edges* $\pi = (e_1, e_2, \ldots, e_k)$. *Let* S *be the set of packets whose routes pass through at least one of* $(e_1, \ldots, e_k)$. *Then, the delay incurred by* **c** *is at most* $|S|$.

Proof: A packet in S is said to leave π at that time step at which it traverses an edge of π for the last time. If a packet is ready to follow edge e_j at time t, we define its *lag* at time t to be $t - j$. The lag of **c** is initially zero, and the delay incurred by **c** is its lag when it traverse e_k. We will show that each step at which the lag of **c** increases by one can be charged to a distinct member of S.

We argue that if the lag of **c** reaches $\ell + 1$, some packet in S leaves π with lag ℓ. When the lag of **c** increases from ℓ to $\ell + 1$, there must be at least one packet (from S) that wishes to traverse the same edge as **c** at that time step, since otherwise **c** would be permitted to traverse this edge and its lag would not increase. Thus, S contains at least one packet whose lag reach the value ℓ.

Let τ be the last time step at which any packet in S has lag ℓ. Thus there is a packet **d** ready to follow edge e_μ at τ, such that $\tau - \mu = \ell$. We argue that some packet of S leaves π at τ; this establishes the lemma since once a packet leaves π, it would never join it again and as such will never again delay **c**.

Since **d** is ready to follow e_μ at τ, some packet ω (which may be **d** itself) in S follows e_μ at time τ. Now ω leaves π at time τ; if not, some packet will follow $e_{\mu+1}$ at step $\mu + 1$ with lag still at ℓ, violating the maximality of τ. We charge to ω the increase in the lag of **c** from ℓ to $\ell + 1$; since ω leaves π, it will never be charged again. Thus, each member of S whose route intersects π is charge for at most one delay, establishing the lemma. ■

Let H_{ij} be an indicator variable that is 1 if ρ_i and ρ_j share an edge, and 0 otherwise. The total delay for v_i is at most $\leq \sum_j H_{ij}$.

Crucially, for a fixed i, the variables $H_{i1}, \ldots, H_{iN}$ are independent. Indeed, imagine first picking the destination of v_i, and let the associated path be ρ_i. Now, pick the destinations of all the other packets in the network. Since the sampling of destinations is done with replacements, whether or not, the path of v_j intersects ρ_i or not, is independent of whether v_k intersects ρ_i. Of course, the probabilities $\mathbf{Pr}\left[H_{ij} = 1\right]$ and $\mathbf{Pr}\left[H_{ik} = 1\right]$ are probably different. Confusingly, however, $H_{11}, \ldots, H_{NN}$ are not independent. Indeed, imagine k and j being close vertices on the hypercube. If $H_{ij} = 1$ then intuitively it means that ρ_i is traveling close to the vertex v_j, and as such there is a higher probability that $H_{ik} = 1$.

Let $\rho_i = (e_1, \ldots, e_k)$, and let $T(e)$ be the number of packets (i.e., paths) that pass through e. We have that

$$\sum_{j=1}^{N} H_{ij} \leq \sum_{j=1}^{k} T(e_j) \quad \text{and thus} \quad \mathbf{E}\left[\sum_{j=1}^{N} H_{ij}\right] \leq \mathbf{E}\left[\sum_{j=1}^{k} T(e_j)\right].$$

Because of symmetry, the variables $T(e)$ have the same distribution for all the edges of G. On the other hand, the expected length of a path is $n/2$, there are N packets, and there are $Nn/2$ edges. We conclude $E[T(e)] = 1$. Thus

$$\mu = \mathbf{E}\left[\sum_{j=1}^{N} H_{ij}\right] \leq \mathbf{E}\left[\sum_{j=1}^{k} T(e_j)\right] = \mathbf{E}\Big[|\rho_i|\Big] \leq \frac{n}{2}.$$

By the Chernoff inequality, we have

$$\mathbf{Pr}\left[\sum_{j} H_{ij} > 7n\right] \leq \mathbf{Pr}\left[\sum_{j} H_{ij} > (1+13)\mu\right] < 2^{-13\mu} \leq 2^{-6n}.$$

Since there are $N = 2^n$ packets, we know that with probability $\leq 2^{-5n}$ all packets arrive to their temporary destination in a delay of most $7n$.

Theorem 7.2.8. *Each packet arrives to its destination in $\leq 14n$ stages, in probability at least $1 - 1/N$ (note that this is very conservative).*

7.2.4. Faraway Strings

Consider the Hamming distance between binary strings. It is natural to ask how many strings of length n can one have, such that any pair of them, is of Hamming distance at least t from each other. Consider two random strings, generated by picking at each bit randomly and independently. Thus, $\mathbf{E}[d_H(x,y)] = n/2$, where $d_H(x,y)$ denote the hamming distance between x and y. In particular, using the Chernoff inequality, we have that

$$\mathbf{Pr}[d_H(x,y) \leq n/2 - \Delta] \leq \exp\Big(-2\Delta^2/n\Big).$$

Next, consider generating M such string, where the value of M would be determined shortly. Clearly, the probability that any pair of strings are at distance at most $n/2 - \Delta$, is

$$\alpha \leq \binom{M}{2} \exp\Big(-2\Delta^2/n\Big) < M^2 \exp\Big(-2\Delta^2/n\Big).$$

If this probability is smaller than one, then there is some probability that all the M strings are of distance at least $n/2 - \Delta$ from each other. Namely, there exists a set of M strings such that every pair of them is far. We used here the fact that if an event has probability larger than zero, then it exists. Thus, set $\Delta = n/4$, and observe that

$$\alpha < M^2 \exp\Big(-2n^2/16n\Big) = M^2 \exp(-n/8).$$

Thus, for $M = \exp(n/16)$, we have that $\alpha < 1$. We conclude:

Lemma 7.2.9. *There exists a set of $\exp(n/16)$ binary strings of length n, such that any pair of them is at Hamming distance at least $n/4$ from each other.*

This is our first introduction to the beautiful technique known as the probabilistic method — we will hear more about it later in the course.

This result has also interesting interpretation in the Euclidean setting. Indeed, consider the sphere $\mathbb{S}$ of radius $\sqrt{n}/2$ centered at $(1/2, 1/2, \ldots, 1/2) \in \mathbb{R}^n$. Clearly, all the vertices of the binary hypercube $\{0,1\}^n$ lie on this sphere. As such, let P be the set of points on $\mathbb{S}$ that exists according to Lemma 7.2.9. A pair p, q of points of P have *Euclidean* distance at least $\sqrt{d_H(p,q)} = \sqrt{n}4 = \sqrt{n}/2$ from each other. We conclude:

Lemma 7.2.10. *Consider the unit hypersphere $\mathbb{S}$ in $\mathbb{R}^n$. The sphere $\mathbb{S}$ contains a set Q of points, such that each pair of points is at (Euclidean) distance at least one from each other, and $|Q| \geq \exp(n/16)$.*

7.3. The Chernoff Bound — General Case

Here we present the Chernoff bound in a more general settings.

Question 7.3.1. *Let $X_1, \ldots, X_n$ be n independent Bernoulli trials, where*

$$\mathbf{Pr}\Big[X_i = 1\Big] = p_i, \quad \text{and} \quad \mathbf{Pr}\Big[X_i = 0\Big] = q_i = 1 - p_i.$$

(Each X_i is known as a Poisson trials.) And let $X = \sum_{i=1}^{b} X_i$. $\mu = \mathbf{E}[X] = \sum_i p_i$. We are interested in the question of what is the probability that $X > (1+\delta)\mu$?

Theorem 7.3.2. *For any $\delta > 0$, we have $\mathbf{Pr}\Big[X > (1+\delta)\mu\Big] < \left(\dfrac{e^\delta}{(1+\delta)^{1+\delta}}\right)^\mu$.*

Or in a more simplified form, we have:

$$\delta \leq 2e - 1 \qquad \mathbf{Pr}\Big[X > (1+\delta)\mu\Big] < \exp\Big(-\mu\delta^2/4\Big), \tag{7.1}$$

$$\delta > 2e - 1 \qquad \mathbf{Pr}\Big[X > (1+\delta)\mu\Big] < 2^{-\mu(1+\delta)}, \tag{7.2}$$

and $$\delta \geq e^2 \qquad \mathbf{Pr}\Big[X > (1+\delta)\mu\Big] < \exp\left(-\frac{\mu\delta\ln\delta}{2}\right). \tag{7.3}$$

Proof: We have $\mathbf{Pr}[X > (1+\delta)\mu] = \mathbf{Pr}\Big[e^{tX} > e^{t(1+\delta)\mu}\Big]$. By the Markov inequality, we have:

$$\mathbf{Pr}\Big[X > (1+\delta)\mu\Big] < \frac{\mathbf{E}\Big[e^{tX}\Big]}{e^{t(1+\delta)\mu}}$$

On the other hand,

$$\mathbf{E}\Big[e^{tX}\Big] = \mathbf{E}\Big[e^{t(X_1+X_2\ldots+X_n)}\Big] = \mathbf{E}\Big[e^{tX_1}\Big]\cdots\mathbf{E}\Big[e^{tX_n}\Big].$$

Namely,

$$\mathbf{Pr}[X > (1+\delta)\mu] < \frac{\prod_{i=1}^n \mathbf{E}\Big[e^{tX_i}\Big]}{e^{t(1+\delta)\mu}} = \frac{\prod_{i=1}^n \Big((1-p_i)e^0 + p_i e^t\Big)}{e^{t(1+\delta)\mu}} = \frac{\prod_{i=1}^n (1 + p_i(e^t-1))}{e^{t(1+\delta)\mu}}.$$

Let $y = p_i(e^t - 1)$. We know that $1 + y < e^y$ (since $y > 0$). Thus,

$$\begin{aligned}
\mathbf{Pr}[X > (1+\delta)\mu] &< \frac{\prod_{i=1}^n \exp(p_i(e^t-1))}{e^{t(1+\delta)\mu}} = \frac{\exp(\sum_{i=1}^n p_i(e^t-1))}{e^{t(1+\delta)\mu}} \\
&= \frac{\exp((e^t-1)\sum_{i=1}^n p_i)}{e^{t(1+\delta)\mu}} = \frac{\exp((e^t-1)\mu)}{e^{t(1+\delta)\mu}} = \left(\frac{\exp(e^t-1)}{e^{t(1+\delta)}}\right)^\mu \\
&= \left(\frac{\exp(\delta)}{(1+\delta)^{(1+\delta)}}\right)^\mu,
\end{aligned}$$

if we set $t = \log(1+\delta)$.

For the proof of the simplified form, see Section 7.3.1. ■

Definition 7.3.3. $F^+(\mu,\delta) = \left[\dfrac{e^\delta}{(1+\delta)^{(1+\delta)}}\right]^\mu$.

Example 7.3.4. Arkansas Aardvarks win a game with probability $1/3$. What is their probability to have a winning season with n games. By Chernoff inequality, this probability is smaller than

$$F^+(n/3, 1/2) = \left[\frac{e^{1/2}}{1.5^{1.5}}\right]^{n/3} = (0.89745)^{n/3} = 0.964577^n.$$

For $n = 40$, this probability is smaller than 0.236307. For $n = 100$ this is less than 0.027145. For $n = 1000$, this is smaller than $2.17221 \cdot 10^{-16}$ (which is pretty slim and shady). Namely, as the number of experiments is increases, the distribution converges to its expectation, and this converge is exponential.

Theorem 7.3.5. *Under the same assumptions as Theorem 7.3.2, we have:* $\mathbf{Pr}[X < (1-\delta)\mu] < \exp\left(-\mu\delta^2/2\right)$.

Definition 7.3.6. Let $F^-(\mu,\delta) = e^{-\mu\delta^2/2}$, and let $\Delta^-(\mu,\varepsilon)$ denote the quantity, which is what should be the value of δ, so that the probability is smaller than ε. We have that

$$\Delta^-(\mu,\varepsilon) = \sqrt{\frac{2\log 1/\varepsilon}{\mu}}.$$

And for large δ we have $\Delta^+(\mu,\varepsilon) < \dfrac{\log_2(1/\varepsilon)}{\mu} - 1$.

7.3.1. A More Convenient Form

Proof: **(of simplified form of Theorem 7.3.2$_{\text{p12}}$)** Eq. (7.2) is easy. Indeed, we have

$$\left[\frac{e}{1+\delta}\right]^{(1+\delta)\mu} \le \left[\frac{e}{1+2e-1}\right]^{(1+\delta)\mu} \le 2^{-(1+\delta)\mu},$$

since $\delta > 2e-1$. For the stronger version, Eq. (7.3), observe that

$$\mathbf{Pr}\Big[X > (1+\delta)\mu\Big] < \left(\frac{e^\delta}{(1+\delta)^{1+\delta}}\right)^\mu = \exp\Big(\mu\delta - \mu(1+\delta)\ln(1+\delta)\Big). \tag{7.4}$$

As such, we have

$$\mathbf{Pr}\Big[X > (1+\delta)\mu\Big] < \exp\Big(-\mu(1+\delta)(\ln(1+\delta)-1)\Big) \le \exp\left(-\mu\delta\ln\frac{1+\delta}{e}\right) \le \exp\left(-\frac{\mu\delta\ln\delta}{2}\right),$$

since for $x \ge e^2$ we have that $\dfrac{1+x}{e} \ge \sqrt{x} \iff \ln\dfrac{1+x}{e} \ge \dfrac{\ln x}{2}$.

As for Eq. (7.1), we prove this only for $\delta \le 1/2$. For details about the case $1/2 \le \delta \le 2e-1$, see [MR95]. The Taylor expansion of $\ln(1+\delta)$ is

$$\delta - \frac{\delta^2}{2} + \frac{\delta^3}{3} - \frac{\delta^4}{4} + \cdot \ge \delta - \frac{\delta^2}{2},$$

for $\delta \leq 1$. Thus, plugging into Eq. (7.4), we have

$$\begin{aligned}\mathbf{Pr}\Big[X > (1+\delta)\mu\Big] &< \exp\Big(\mu\Big[\delta - (1+\delta)(\delta - \delta^2/2)\Big]\Big) = \exp\Big(\mu\Big(\delta - \delta + \delta^2/2 - \delta^2 + \delta^3/2\Big)\Big) \\ &\leq \exp\Big(\mu\Big(-\delta^2/2 + \delta^3/2\Big)\Big) \leq \exp\Big(-\mu\delta^2/4\Big),\end{aligned}$$

for $\delta \leq 1/2$. ■

7.4. A special case of Hoeffding's inequality

In this section, we prove yet another version of Chernoff inequality, where each variable is randomly picked according to its own distribution in the range $[0, 1]$. We prove a more general version of this inequality in Section 7.5, but the version presented here does not follow from this generalization.

Theorem 7.4.1. *Let $X_1, \ldots, X_n \in [0,1]$ be n independent random variables, let $X = \sum_{i=1}^n X_i$, and let $\mu = \mathbf{E}[X]$. We have that $\mathbf{Pr}\Big[X - \mu \geq \eta\Big] \leq \left(\frac{\mu}{\mu+\eta}\right)^{\mu+\eta}\left(\frac{n-\mu}{n-\mu-\eta}\right)^{n-\mu-\eta}$.*

Proof: Let $s \geq 1$ be some arbitrary parameter. By the standard arguments, we have

$$\gamma = \mathbf{Pr}\Big[X \geq \mu + \eta\Big] = \mathbf{Pr}\Big[s^X \geq s^{\mu+\eta}\Big] \leq \frac{\mathbf{E}\Big[s^X\Big]}{s^{\mu+\eta}} = s^{-\mu-\eta}\prod_{i=1}^n \mathbf{E}\Big[s^{X_i}\Big].$$

By calculations, see Lemma 7.4.6 below, one can show that $\mathbf{E}\Big[s^{X_1}\Big] \leq 1 + (s-1)\,\mathbf{E}[X_i]$. As such, by the AM-GM inequality[8], we have that

$$\prod_{i=1}^n \mathbf{E}\Big[s^{X_i}\Big] \leq \prod_{i=1}^n \Big(1 + (s-1)\,\mathbf{E}[X_i]\Big) \leq \left(\frac{1}{n}\sum_{i=1}^n \Big(1 + (s-1)\,\mathbf{E}[X_i]\Big)\right)^n = \left(1 + (s-1)\frac{\mu}{n}\right)^n.$$

Setting $s = \dfrac{(\mu+\eta)(n-\mu)}{\mu(n-\mu-\eta)} = \dfrac{\mu n - \mu^2 + \eta n - \eta\mu}{\mu n - \mu^2 - \eta\mu}$ we have that

$$1 + (s-1)\frac{\mu}{n} = 1 + \frac{\eta n}{\mu n - \mu^2 - \eta\mu}\cdot\frac{\mu}{n} = 1 + \frac{\eta}{n-\mu-\eta} = \frac{n-\mu}{n-\mu-\eta}.$$

As such, we have that

$$\gamma \leq s^{-\mu-\eta}\prod_{i=1}^n \mathbf{E}\Big[s^{X_i}\Big] = \left(\frac{\mu(n-\mu-\eta)}{(\mu+\eta)(n-\mu)}\right)^{\mu+\eta}\left(\frac{n-\mu}{n-\mu-\eta}\right)^n = \left(\frac{\mu}{(\mu+\eta)}\right)^{\mu+\eta}\left(\frac{n-\mu}{n-\mu-\eta}\right)^{n-\mu-\eta}.$$ ■

Remark 7.4.2. Setting $s = (\mu+\eta)/\mu$ in the proof of Theorem 7.4.1, we have

$$\mathbf{Pr}\Big[X - \mu \geq \eta\Big] \leq \left(\tfrac{\mu}{\mu+\eta}\right)^{\mu+\eta}\left(1 + \left(\tfrac{\mu+\eta}{\mu} - 1\right)\tfrac{\mu}{n}\right)^n = \left(\tfrac{\mu}{\mu+\eta}\right)^{\mu+\eta}\left(1 + \tfrac{\eta}{n}\right)^n.$$

[8] The inequality between arithmetic and geometric means: $(\sum_{i=1}^n x_i)/n \geq \sqrt[n]{x_1 \cdots x_n}$.

Corollary 7.4.3. *Let $X_1, \ldots, X_n \in [0,1]$ be n independent random variables, let $\overline{X} = \sum_{i=1}^n X_i/n$, $p = \mathbf{E}\left[\overline{X}\right] = \mu/n$ and $q = 1-p$. Then, we have that $\mathbf{Pr}\left[\overline{X} - p \geq t\right] \leq \exp(nf(t))$, for*

$$f(t) = (p+t)\ln\frac{p}{p+t} + (q-t)\ln\frac{q}{q-t}. \tag{7.5}$$

Theorem 7.4.4. *Let $X_1, \ldots, X_n \in [0,1]$ be n independent random variables, let $\overline{X} = (\sum_{i=1}^n X_i)/n$, and let $p = \mathbf{E}[X]$. We have that $\mathbf{Pr}\left[\overline{X} - p \geq t\right] \leq \exp\left(-2nt^2\right)$ and $\mathbf{Pr}\left[\overline{X} - p \leq -t\right] \leq \exp\left(-2nt^2\right)$.*

Proof: Let $p = \mu/n$, $q = 1-p$, and let $f(t)$ be the function from Eq. (7.5), for $t \in (-p, q)$. Now, we have that

$$\begin{aligned} f'(t) &= \ln\frac{p}{p+t} + (p+t)\frac{p+t}{p}\left(-\frac{p}{(p+t)^2}\right) - \ln\frac{q}{q-t} - (q-t)\frac{q-t}{q}\frac{q}{(q-t)^2} = \ln\frac{p}{p+t} - \ln\frac{q}{q-t} \\ &= \ln\frac{p(q-t)}{q(p+t)}. \end{aligned}$$

As for the second derivative, we have

$$f''(t) = \frac{\cancel{q(p+t)}}{p(q-t)} \cdot \frac{p}{q} \cdot \frac{(p+t)(-1) - (q-t)}{\cancel{(p+t)^2}}. = \frac{-p-t-q+t}{(q-t)(p+t)} = -\frac{1}{(q-t)(p+t)} \leq -4.$$

Indeed, $t \in (-p, q)$ and the denominator is minimized for $t = (q-p)/2$, and as such $(q-t)(p+t) \leq \Big(2q - (q-p)\Big)\Big(2p + (q-p)\Big)/4 = (p+q)^2/4 = 1/4$.

Now, $f(0) = 0$ and $f'(0) = 0$, and by Taylor's expansion, we have that $f(t) = f(0) + f'(0)t + \dfrac{f''(x)}{2}t^2 \leq -2t^2$, where x is between 0 and t.

The first bound now readily follows from plugging this bound into Corollary 7.4.3. The second bound follows by considering the random variants $Y_i = 1 - X_i$, for all i, and plugging this into the first bound. Indeed, for $\overline{Y} = 1 - \overline{X}$, we have that $q = \mathbf{E}\left[\overline{Y}\right]$, and then $\overline{X} - p \leq -t \iff t \leq p - \overline{X} \iff t \leq 1 - q - (1 - \overline{Y}) = \overline{Y} - q$. Thus, $\mathbf{Pr}\left[\overline{X} - p \leq -t\right] = \mathbf{Pr}\left[\overline{Y} - q \geq t\right] \leq \exp\left(-2nt^2\right)$. ■

Theorem 7.4.5. *Let $X_1, \ldots, X_n \in [0,1]$ be n independent random variables, let $X = (\sum_{i=1}^n X_i)$, and let $\mu = \mathbf{E}[X]$. We have that $\mathbf{Pr}[X - \mu \geq \varepsilon\mu] \leq \exp\left(-\varepsilon^2\mu/4\right)$ and $\mathbf{Pr}[X - \mu \leq -\varepsilon\mu] \leq \exp\left(-\varepsilon^2\mu/2\right)$.*

Proof: Let $p = \mu/n$, and let $g(x) = f(px)$, for $x \in [0,1]$ and $xp < q$. As before, computing the derivative of g, we have

$$g'(x) = pf'(xp) = p\ln\frac{p(q-xp)}{q(p+xp)} = p\ln\frac{q-xp}{q(1+x)} \leq p\ln\frac{1}{1+x} \leq -\frac{px}{2},$$

since $(q-xp)/q$ is maximized for $x = 0$, and$\ln\frac{1}{1+x} \leq -x/2$, for $x \in [0,1]$, as can be easily verified[9]. Now, $g(0) = f(0) = 0$, and by integration, we have that $g(x) = \int_{y=0}^{x} g'(y)\mathrm{d}y \leq \int_{y=0}^{x}(-py/2)\mathrm{d}y = -px^2/4$. Now, plugging into Corollary 7.4.3, we get that the desired probability $\mathbf{Pr}[X - \mu \geq \varepsilon\mu]$ is

$$\mathbf{Pr}\left[\overline{X} - p \geq \varepsilon p\right] \leq \exp(nf(\varepsilon p)) = \exp\left(ng(\varepsilon)\right) \leq \exp\left(-pn\varepsilon^2/4\right) = \exp\left(-\mu\varepsilon^2/4\right).$$

[9] Indeed, this is equivalent to $\frac{1}{1+x} \leq e^{-x/2} \iff e^{x/2} \leq 1 + x$, which readily holds for $x \in [0,1]$.

As for the other inequality, set $h(x) = g(-x) = f(-xp)$. Then

$$h'(x) = -pf'(-xp) = -p \ln \frac{p(q+xp)}{q(p-xp)} = p \ln \frac{q(1-x)}{q+xp} = p \ln \frac{q-xq}{q+xp} = p \ln\left(1 - x\frac{p+q}{q+xp}\right)$$
$$= p \ln\left(1 - x\frac{1}{q+xp}\right) \leq p \ln(1-x) \leq -px,$$

since $1 - x \leq e^{-x}$. By integration, as before, we conclude that $h(x) \leq -px^2/2$. Now, plugging into Corollary 7.4.3, we get $\mathbf{Pr}[X - \mu \leq -\varepsilon\mu] = \mathbf{Pr}\left[\overline{X} - p \leq -\varepsilon p\right] \leq \exp(nf(-\varepsilon p)) \leq \exp(nh(\varepsilon)) \leq \exp\left(-np\varepsilon^2/2\right) \leq \exp\left(-\mu\varepsilon^2/2\right)$. ■

7.4.1. Some technical lemmas

Lemma 7.4.6. *Let $X \in [0,1]$ be a random variable, and let $s \geq 1$. Then $\mathbf{E}\left[s^X\right] \leq 1 + (s-1)\,\mathbf{E}[X]$.*

Proof: For the sake of simplicity of exposition, assume that X is a discrete random variable, and that there is a value $\alpha \in (0, 1/2)$, such that $\beta = \mathbf{Pr}[X = \alpha] > 0$. Consider the modified random variable X', such that $\mathbf{Pr}[X' = 0] = \mathbf{Pr}[X = 0] + \beta/2$, and $\mathbf{Pr}[X' = 2\alpha] = \mathbf{Pr}[X = \alpha] + \beta/2$. Clearly, $\mathbf{E}[X] = \mathbf{E}[X']$. Next, observe that $\mathbf{E}\left[s^{X'}\right] - \mathbf{E}\left[s^X\right] = (\beta/2)(s^{2\alpha} + s^0) - \beta s^\alpha \geq 0$, by the convexity of s^x. We conclude that $\mathbf{E}\left[s^X\right]$ achieves its maximum if takes only the values 0 and 1. But then, we have that $\mathbf{E}\left[s^X\right] = \mathbf{Pr}[X = 0]\, s^0 + \mathbf{Pr}[X = 1]\, s^1 = (1 - \mathbf{E}[X]) + \mathbf{E}[X]\, s = 1 + (s-1)\,\mathbf{E}[X]$, as claimed. ■

7.5. Hoeffding's inequality

In this section, we prove a generalization of Chernoff's inequality. The proof is considerably more tedious, and it is included here for the sake of completeness.

Lemma 7.5.1. *Let X be a random variable. If $\mathbf{E}[X] = 0$ and $a \leq X \leq b$, then for any $s > 0$, we have $\mathbf{E}\left[e^{sX}\right] \leq \exp\left(s^2(b-a)^2/8\right)$.*

Proof: Let $a \leq x \leq b$ and observe that x can be written as a convex combination of a and b. In particular, we have

$$x = \lambda a + (1-\lambda)b \quad \text{for} \quad \lambda = \frac{b-x}{b-a} \in [0,1].$$

Since $s > 0$, the function $\exp(sx)$ is convex, and as such

$$e^{sx} \leq \frac{b-x}{b-a}e^{sa} + \frac{x-a}{b-a}e^{sb},$$

since we have that $f(\lambda x + (1-\lambda)y) \leq \lambda f(x) + (1-\lambda)f(y)$ if $f(\cdot)$ is a convex function. Thus, for a random variable X, by linearity of expectation, we have

$$\mathbf{E}\left[e^{sX}\right] \leq \mathbf{E}\left[\frac{b-X}{b-a}e^{sa} + \frac{X-a}{b-a}e^{sb}\right] = \frac{b-\mathbf{E}[X]}{b-a}e^{sa} + \frac{\mathbf{E}[X]-a}{b-a}e^{sb}$$
$$= \frac{b}{b-a}e^{sa} - \frac{a}{b-a}e^{sb},$$

since $\mathbf{E}[X] = 0$.

Next, set $p = -\dfrac{a}{b-a}$ and observe that $1 - p = 1 + \dfrac{a}{b-a} = \dfrac{b}{b-a}$ and

$$-ps(b-a) = -\left(-\frac{a}{b-a}\right)s(b-a) = sa.$$

As such, we have

$$\begin{aligned}\mathbf{E}\left[e^{sX}\right] &\leq (1-p)e^{sa} + pe^{sb} = (1 - p + pe^{s(b-a)})e^{sa} \\ &= (1 - p + pe^{s(b-a)})e^{-ps(b-a)} \\ &= \exp\left(-ps(b-a) + \ln\left(1 - p + pe^{s(b-a)}\right)\right) = \exp(-pu + \ln(1 - p + pe^{u})),\end{aligned}$$

for $u = s(b-a)$. Setting

$$\phi(u) = -pu + \ln(1 - p + pe^{u}),$$

we thus have $\mathbf{E}\left[e^{sX}\right] \leq \exp(\phi(u))$. To prove the claim, we will show that $\phi(u) \leq u^2/8 = s^2(b-a)^2/8$.

To see that, expand $\phi(u)$ about zero using Taylor's expansion. We have

$$\phi(u) = \phi(0) + u\phi'(0) + \frac{1}{2}u^2\phi''(\theta) \tag{7.6}$$

where $\theta \in [0, u]$, and notice that $\phi(0) = 0$. Furthermore, we have

$$\phi'(u) = -p + \frac{pe^{u}}{1 - p + pe^{u}},$$

and as such $\phi'(0) = -p + \frac{p}{1-p+p} = 0$. Now,

$$\phi''(u) = \frac{(1 - p + pe^{u})pe^{u} - (pe^{u})^2}{(1 - p + pe^{u})^2} = \frac{(1-p)pe^{u}}{(1 - p + pe^{u})^2}.$$

For any $x, y \geq 0$, we have $(x+y)^2 \geq 4xy$ as this is equivalent to $(x-y)^2 \geq 0$. Setting $x = 1 - p$ and $y = pe^{u}$, we have that

$$\phi''(u) = \frac{(1-p)pe^{u}}{(1 - p + pe^{u})^2} \leq \frac{(1-p)pe^{u}}{4(1-p)pe^{u}} = \frac{1}{4}.$$

Plugging this into Eq. (7.6), we get that

$$\phi(u) \leq \frac{1}{8}u^2 = \frac{1}{8}(s(b-a))^2 \quad \text{and} \quad \mathbf{E}\left[e^{sX}\right] \leq \exp(\phi(u)) \leq \exp\left(\frac{1}{8}(s(b-a))^2\right),$$

as claimed. ∎

Lemma 7.5.2. *Let X be a random variable. If $\mathbf{E}[X] = 0$ and $a \leq X \leq b$, then for any $s > 0$, we have*

$$\mathbf{Pr}[X > t] \leq \frac{\exp\left(\frac{s^2(b-a)^2}{8}\right)}{e^{st}}.$$

Proof: Using the same technique we used in proving Chernoff's inequality, we have that

$$\mathbf{Pr}[X > t] = \mathbf{Pr}\left[e^{sX} > e^{st}\right] \leq \frac{\mathbf{E}\left[e^{sX}\right]}{e^{st}} \leq \frac{\exp\left(\frac{s^2(b-a)^2}{8}\right)}{e^{st}}.$$ ∎

Theorem 7.5.3 (Hoeffding's inequality). *Let $X_1, \ldots, X_n$ be independent random variables, where $X_i \in [a_i, b_i]$, for $i = 1, \ldots, n$. Then, for the random variable $S = X_1 + \cdots + X_n$ and any $\eta > 0$, we have*

$$\mathbf{Pr}\left[\left|S - \mathbf{E}[S]\right| \geq \eta\right] \leq 2\exp\left(-\frac{2\eta^2}{\sum_{i=1}^{n}(b_i - a_i)^2}\right).$$

Proof: Let $Z_i = X_i - \mathbf{E}[X_i]$, for $i = 1, \ldots, n$. Set $Z = \sum_{i=1}^{n} Z_i$, and observe that

$$\mathbf{Pr}[Z \geq \eta] = \mathbf{Pr}\left[e^{sZ} \geq e^{s\eta}\right] \leq \frac{\mathbf{E}[\exp(sZ)]}{\exp(s\eta)},$$

by Markov's inequality. Arguing as in the proof of Chernoff's inequality, we have

$$\mathbf{E}[\exp(sZ)] = \mathbf{E}\left[\prod_{i=1}^{n}\exp(sZ_i)\right] = \prod_{i=1}^{n}\mathbf{E}[\exp(sZ_i)] \leq \prod_{i=1}^{n}\exp\left(\frac{s^2(b_i - a_i)^2}{8}\right),$$

since the Z_is are independent and by Lemma 7.5.1. This implies that

$$\mathbf{Pr}[Z \geq \eta] \leq \exp(-s\eta)\prod_{i=1}^{n} e^{s^2(b_i - a_i)^2/8} = \exp\left(\frac{s^2}{8}\sum_{i=1}^{n}(b_i - a_i)^2 - s\eta\right).$$

The upper bound is minimized for $s = 4\eta/\left(\sum_i (b_i - a_i)^2\right)$, implying

$$\mathbf{Pr}[Z \geq \eta] \leq \exp\left(-\frac{2\eta^2}{\sum(b_i - a_i)^2}\right).$$

The claim now follows by the symmetry of the upper bound (i.e., apply the same proof to $-Z$). ■

7.6. Bibliographical notes

Some of the exposition here follows more or less the exposition in [MR95]. Exercise 7.7.1 (without the hint) is from [Mat99]. McDiarmid [McD89] provides a survey of Chernoff type inequalities, and Theorem 7.4.5 and Section 7.4 is taken from there (our proof has somewhat weaker constants).

Section 7.2.3 is based on Section 4.2 in [MR95]. A similar result to Theorem 7.2.8 is known for the case of the wrapped butterfly topology (which is similar to the hypercube topology but every node has a constant degree, and there is no clear symmetry). The interested reader is referred to [MU05].

A more general treatment of such inequalities and tools is provided by Dubhashi and Panconesi [DP09].

7.7. Exercises

Exercise 7.7.1 (Chernoff inequality is tight.). Let $S = \sum_{i=1}^{n} S_i$ be a sum of n independent random variables each attaining values $+1$ and -1 with equal probability. Let $P(n, \Delta) = \mathbf{Pr}[S > \Delta]$. Prove that for $\Delta \leq n/C$,

$$P(n, \Delta) \geq \frac{1}{C}\exp\left(-\frac{\Delta^2}{Cn}\right),$$

where C is a suitable constant. That is, the well-known Chernoff bound $P(n, \Delta) \leq \exp(-\Delta^2/2n))$ is close to the truth.

Exercise 7.7.2 (Chernoff inequality is tight by direct calculations.). For this question use only basic argumentation – do not use Stirling's formula, Chernoff inequality or any similar "heavy" machinery.

(A) Prove that $\sum_{i=0}^{n-k} \binom{2n}{i} \leq \frac{n}{4k^2} 2^{2n}$.

Hint: Consider flipping a coin $2n$ times. Write down explicitly the probability of this coin to have at most $n-k$ heads, and use Chebyshev inequality.

(B) Using (A), prove that $\binom{2n}{n} \geq 2^{2n}/4\sqrt{n}$ (which is a pretty good estimate).

(C) Prove that $\binom{2n}{n+i+1} = \left(1 - \frac{2i+1}{n+i+1}\right)\binom{2n}{n+i}$.

(D) Prove that $\binom{2n}{n+i} \leq \exp\left(\frac{-i(i-1)}{2n}\right)\binom{2n}{n}$.

(E) Prove that $\binom{2n}{n+i} \geq \exp\left(-\frac{8i^2}{n}\right)\binom{2n}{n}$.

(F) Using the above, prove that $\binom{2n}{n} \leq c\frac{2^{2n}}{\sqrt{n}}$ for some constant c (I got $c = 0.824...$ but any reasonable constant will do).

(G) Using the above, prove that

$$\sum_{i=t\sqrt{n}+1}^{(t+1)\sqrt{n}} \binom{2n}{n-i} \leq c2^{2n} \exp\left(-t^2/2\right).$$

In particular, conclude that when flipping fair coin $2n$ times, the probability to get less than $n - t\sqrt{n}$ heads (for t an integer) is smaller than $c' \exp\left(-t^2/2\right)$, for some constant c'.

(H) Let X be the number of heads in $2n$ coin flips. Prove that for any integer $t > 0$ and any $\delta > 0$ sufficiently small, it holds that $\mathbf{Pr}[X < (1-\delta)n] \geq \exp\left(-c''\delta^2 n\right)$, where c'' is some constant. Namely, the Chernoff inequality is tight in the worst case.

Exercise 7.7.3 (More binary strings. More!). To some extent, Lemma 7.2.9 is somewhat silly, as one can prove a better bound by direct argumentation. Indeed, for a fixed binary string x of length n, show a bound on the number of strings in the Hamming ball around x of radius $n/4$ (i.e., binary strings of distance at most $n/4$ from x). (Hint: interpret the special case of the Chernoff inequality as an inequality over binomial coefficients.)

Next, argue that the greedy algorithm which repeatedly pick a string which is in distance $\geq n/4$ from all strings picked so far, stops after picking at least $\exp(n/8)$ strings.

Exercise 7.7.4 (Tail inequality for geometric variables). Let $X_1, \ldots, X_m$ be m independent random variables with geometric distribution with probability p (i.e., $\mathbf{Pr}[X_i = j] = (1-p)^{j-1}p$). Let $Y = \sum_i X_i$, and let $\mu = \mathbf{E}[Y] = m/p$. Prove that $\mathbf{Pr}\left[Y \geq (1+\delta)\mu\right] \leq \exp\left(-m\delta^2/8\right)$.

Bibliography

[DP09] D. Dubhashi and A. Panconesi. *Concentration of Measure for the Analysis of Randomized Algorithms*. Cambridge University Press, 2009.

[Kel56] J. L. Kelly. A new interpretation of information rate. *Bell Sys. Tech. J.*, 35(4):917–926, jul 1956.

[KKT91] C. Kaklamanis, D. Krizanc, and T. Tsantilas. Tight bounds for oblivious routing in the hypercube. *Math. sys. theory*, 24(1):223–232, 1991.

[Mat99] J. Matoušek. *Geometric Discrepancy*. Springer, 1999.

[McD89] C. McDiarmid. *Surveys in Combinatorics*, chapter On the method of bounded differences. Cambridge University Press, 1989.

[MR95] R. Motwani and P. Raghavan. *Randomized Algorithms*. Cambridge University Press, Cambridge, UK, 1995.

[MU05] M. Mitzenmacher and U. Upfal. *Probability and Computing – randomized algorithms and probabilistic analysis*. Cambridge, 2005.