Chapter 7

Chernoff Inequality - Part II

By Sariel Har-Peled, May 29, 2013[①]

7.1 Tail Inequalities

7.1.1 Chernoff Inequality - A Special Case

We saw the following in the previous lecture.

Theorem 7.1.1. *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*

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

Corollary 7.1.2. *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 7.1.3. *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*

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

7.1.2 The Chernoff Bound — General Case

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

Question 7.1.4. *Let $X_1, \ldots, X_n$ be n independent Bernoulli trials, where*

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

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

Or in a more simplified form, for any $\delta \leq 2e - 1$,

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

and

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

for $\delta \geq 2e - 1$.

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

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

On the other hand,

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

Namely,

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

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 7.1.3. ∎

Definition 7.1.6. $F^{+}(\mu, \delta) = \left[\frac{e^{\delta}}{(1+\delta)^{(1+\delta)}}\right]^{\mu}$.

Example 7.1.7. 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 7.1.8. *Under the same assumptions as Theorem 7.1.5, we have:*

$$\mathbf{Pr}[X < (1-\delta)\mu] < e^{-\mu\delta^2/2}.$$

Definition 7.1.9. $F^-(\mu, \delta) = e^{-\mu\delta^2/2}$.

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

$$\Delta^+(\mu, \varepsilon) < \frac{\log_2(1/\varepsilon)}{\mu} - 1.$$

7.1.3 A More Convenient Form

Proof: **(of simplified form of Theorem 7.1.5)** Eq. (7.2) is easy. Indeed, we have

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

since $\delta > 2e - 1$.

As for Eq. (7.1), we prove this only for $\delta \leq 1/2$. For details about the case $1/2 \leq \delta \leq 2e-1$, see [MR95]. By Theorem 7.1.5, we have

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

The Taylor expansion of $\ln(1+\delta)$ is

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

for $\delta \leq 1$. Thus,

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

for $\delta \leq 1/2$. ∎

Values	Probabilities	Inequality	Ref
$-1, +1$	$\mathbf{Pr}[X_i = -1] =$ $\mathbf{Pr}[X_i = 1] = \frac{1}{2}$	$\mathbf{Pr}[Y \geq \Delta] \leq e^{-\Delta^2/2n}$ $\mathbf{Pr}[Y \leq -\Delta] \leq e^{-\Delta^2/2n}$ $\mathbf{Pr}[\lvert Y\rvert \geq \Delta] \leq 2e^{-\Delta^2/2n}$	Theorem 7.1.1 Theorem 7.1.1 Corollary 7.1.2
$0, 1$	$\mathbf{Pr}[X_i = 0] =$ $\mathbf{Pr}[X_i = 1] = \frac{1}{2}$	$\mathbf{Pr}\left[\left\lvert Y - \frac{n}{2}\right\rvert \geq \Delta\right] \leq 2e^{-2\Delta^2/n}$	Corollary 7.1.3
0,1	$\mathbf{Pr}[X_i = 0] = 1 - p_i$ $\mathbf{Pr}[X_i = 1] = p_i$	$\mathbf{Pr}[Y > (1+\delta)\mu] < \left(\frac{e^\delta}{(1+\delta)^{1+\delta}}\right)^\mu$	Theorem 7.1.5
	For $\delta \leq 2e - 1$ $\delta \geq 2e - 1$ For $\delta \geq 0$	$\mathbf{Pr}[Y > (1+\delta)\mu] < \exp\left(-\mu\delta^2/4\right)$ $\mathbf{Pr}[Y > (1+\delta)\mu] < 2^{-\mu(1+\delta)}$ $\mathbf{Pr}[Y < (1-\delta)\mu] < \exp\left(-\mu\delta^2/2\right)$	Theorem 7.1.5 Theorem 7.1.8

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

7.2 Application: 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 $[0, \ldots, N]$ 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 $b_1 b_2 \ldots b_n$. Two nodes are connected if their binary string differs only in a single bit. Namely, G is the hypercube over n vertices.

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 representation by a permutation π, and we would like to figure out how to send the permutation and create minimum delay?

In our model, every edge has a FIFO queue 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.

Theorem 7.2.1. *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 $\Omega\left(\sqrt{N/n}\right)$ time.*

Oblivious here refers to the fact that the routing of packet is determined only by inspecting only the packet, and without referring to other things in the network.

How do we sent 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).

We assume each edge have a FIFO queue. The routing algorithm is depicted in Figure 7.1.

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).

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

(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 7.1: The routing algorithm

Exercise 7.2.2. 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 7.2.3. *Let the route of a message* $\mathbf{c}$ *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* $\mathbf{c}$ *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 $\mathbf{c}$ is initially zero, and the delay incurred by $\mathbf{c}$ is its lag when it traverse e_k. We will show that each step at which the lag of $\mathbf{c}$ increases by one can be charged to a distinct member of S.

We argue that if the lag of $\mathbf{c}$ reaches $\ell + 1$, some packet in S leaves π with lag ℓ. When the lag of $\mathbf{c}$ increases from ℓ to $\ell + 1$, there must be at least one packet (from S) that wishes to traverse the same edge as $\mathbf{c}$ at that time step, since otherwise $\mathbf{c}$ 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 $\mathbf{d}$ 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 $\mathbf{c}$.

Since $\mathbf{d}$ is ready to follow e_μ at τ, some packet ω (which may be $\mathbf{d}$ 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 $\mathbf{c}$ 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 $\leq \sum_j H_{ij}$. Note, that for a fixed i, the variables $H_{i1}, \ldots, H_{iN}$ are independent (note however, that $H_{11}, \ldots, H_{NN}$ are not independent!). For $\rho_i = (e_1, \ldots, e_k)$, let $T(e)$ be the number of packets (i.e., paths) that pass through e.

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

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

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

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

7.3 Application of the Chernoff Inequality – 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

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

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

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

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

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

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

Lemma 7.3.1. *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 7.3.1. A pair p, q of points of P have *Euclidean* distance at least $\sqrt{d_H(p,q)} = \sqrt{n4} = \sqrt{n}/2$ from each other. We conclude:

Lemma 7.3.2. *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 $|Q| \geq \exp(n/16)$.*

7.4 Bibliographical notes

The exposition here follows more or less the exposition in [MR95]. Exercise 7.5.1 (without the hint) is from [Mat99].

Section 7.2 is based on Section 4.2 in [MR95]. A similar result to Theorem 7.2.4 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].

7.5 Exercises

Exercise 7.5.1. [**10 points**] 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.

[**Hint:** Use Stirling's formula. There is also an elementary solution, using estimates for the middle binomial coefficients [MN98, pages 83–84], but this solution is considerably more involved and yields unfriendly constants.**]**

Exercise 7.5.2. To some extent, Lemma 7.3.1 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.

Bibliography

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

[MN98] J. Matoušek and J. Nešetřil. *Invitation to Discrete Mathematics*. Oxford Univ Press, 1998.

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

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