Tail Inequalities

497 - Randomized Algorithms

Sariel Har-Peled

December 20, 2002

"Wir mssen wissen, wir werden wissen" (We must know, we shall know)
— David Hilbert

1 Tail Inequalities

1.1 The Chernoff Bound — Special Case

Theorem 1.1 *Let $X_1, \ldots, X_n$ be n independent random varaibles, 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*

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

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

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

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

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

We have

$$\mathbf{Pr}[Y \ge \Delta] \le \frac{\exp(nt^2/2)}{\exp(t\Delta)} = \exp\bigl(nt^2/2 - t\Delta\bigr).$$

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

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

■

By the symmetry of Y, we get the following:

Corollary 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*

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

Corollary 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*

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

1.2 The Chernoff Bound — General Case

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

Question 1.4 *Let*

1. *$X_1, \ldots, X_n$ - n independent Bernoulli trials, where*

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

   Each X_i is known as a Poisson trials.

2. *$X = \sum_{i=1}^{b} X_i$. $\mu = E[X] = \sum_i p_i$.*

Question*: Probability that $X > (1 + \delta)\mu$?*

Theorem 1.5 *For any $\delta > 0$,*

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

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

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

and

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

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

Remark 1.6 Before going any further, it is maybe instrumental to understand what this inequality implies. Set all probabilities to be $p_i = 1/2$, and set $\delta = t/\sqrt{\mu}$. ($\sqrt{\mu}$ is approximately the standard deviation of X if $p_i = 1/2$) Using *very fluffy* math, in particular $e^\delta \approx 1+\delta$), we get the following:

$$\begin{aligned}
\mathbf{Pr}[|X-\mu| > t\sigma_X] \approx \Pr[X > (1+\delta)\mu] &< \left(\frac{e^\delta}{(1+\delta)^{1+\delta}}\right)^\mu \approx \left(\frac{1+\delta}{(1+\delta)^{1+\delta}}\right)^{n/2} \\
&= \left(\frac{1}{1+\delta}\right)^{(t/\sqrt{n})n/2} \approx \left(e^{-\delta}\right)^{(t/\sqrt{n})n/2} \\
&= e^{-(t^2/n)n/2} = e^{-t^2}.
\end{aligned}$$

Thus, Chernoff inequality implies exponential decay with the standard deviation, instead of just polynomial (like the Cheby's inequality). We emphasize again that above calculation is incorrect, and should only be interpreted as an intuition of what is going on.

Proof: **(of Theorem 1.5)**

$$\Pr[X > (1+\delta)\mu] = \Pr\left[e^{tX} > e^{t(1+\delta)\mu}\right].$$

By Markov inequality, we have:

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

On the other hand,

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

Namely,

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

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

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

if we set $t = \log(1+\delta)$.
For the proof of the simplified form, see Section 1.3. ■

Definition 1.7 $F^+(\mu, \delta) = \left[\frac{e^{\delta}}{(1+\delta)^{(1+\delta)}}\right]^{\mu}$.

Example 1.8 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.

Exercise 1.9 *Prove that for $\delta > 2e - 1$, we have*

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

Theorem 1.10 *Under the same assumptions as Theorem 1.5, we have:*

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

Definition 1.11 $F^-(\mu, \delta) = e^{-\mu\delta^2/2}$.
$\Delta^-(\mu, \varepsilon)$ - what should be the value of δ, so that the probability is smaller than ε.

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

For large δ:

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

1.3 A More Convenient Form

Proof: **(of simplified form of Theorem 1.5)** Equation (2) is just Exercise 1.9. As for Equation (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 1.5, we have

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

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,

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

for $\delta \leq 1/2$. ■

2 Application of the Chernoff Inequality – Routing in a Parallel Computer

The following is based on Section 4.2 in [MR95].
G: A graph of processors. Packets can be sent on edges.
$[1,\ldots,N]$: The vertices (i.e., processors) of G.
$N = 2^n$, and G is a hypercube. Each processes is a binary string $b_1 b_2 \ldots b_n$.
Question: Given a permutation π, how to send the permutation and create minimum delay?

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

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. Here is the algorithm:

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

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

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

We analyze only (i) as (iii) follows from the same analysis. ρ_i - the route taken by v_i in (i).

Exercise 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 2.3 *Let the route of v_i follow the sequence of edges $\rho_i = (e_1, e_2, \ldots, e_k)$. 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 v_i is at most $|S|$.*

Let H_{ij} be an indicator variable that is 1 if ρ_i, ρ_j share an edge, 0 otherwise. Total delay for v_i is $\leq \sum_j H_{ij}$. Note, that for a fixed i, the variables $H_{i1}, \ldots, H_{iN}$ are independent (not 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.

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

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

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

By the Chernoff inequality (Exercise 1.9), we have

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

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 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).*

References

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