A Big Picture of Chernoff[*]

Karl Stratos

1 From Markov to Chernoff

Markov's inequality states that $\Pr(X \geq t) \leq \mathbf{E}[X]/t$ for any $X \geq 0$ and $t > 0$. Thus for any nondecreasing function ϕ,

$$\Pr(X \geq t) \leq \Pr(\phi(X) \geq \phi(t)) \leq \frac{\mathbf{E}[\phi(X)]}{\phi(t)} \qquad \forall X, t \in \mathbb{R}: \ \phi(X) \geq 0,\ \phi(t) > 0$$

This suggests natural choices for ϕ like a squaring or exponentiating function since we want ϕ to output a nonnegative number. By choosing $\lambda \geq 0$ and $\phi(z) = \exp(\lambda z)$, we have

$$\Pr(X \geq t) \leq \frac{\mathbf{E}[\exp(\lambda X)]}{\exp(\lambda t)} = \exp\left(-\left(\lambda t - \psi_X(\lambda)\right)\right) \qquad \forall X, t \in \mathbb{R}$$

where $\psi_X(\lambda) := \log \mathbf{E}[\exp(\lambda X)]$ is the *log MGF of X* which is convex.[1] We make the bound as tight as possible by maximizing the concave function $\lambda t - \psi_X(\lambda)$ over $\lambda \geq 0$. WLOG, we will assume $t \geq \mathbf{E}[X]$; then we can drop the nonnegative constraint on λ.[2] Hence we derive **Chernoff's inequality**

$$\Pr(X \geq t) \leq \exp\left(-\left(\sup_{\lambda \in \mathbb{R}} \lambda t - \psi_X(\lambda)\right)\right) = \exp\left(-\psi_X^*(t)\right) \qquad \forall X, t \geq \mathbf{E}[X]$$

where $\psi_X^*(t) := \sup_{\lambda \in \mathbb{R}} \lambda t - \psi_X(\lambda)$ is the *convex conjugate of* $\psi_X(\lambda)$. We can directly calculate $\psi_X(\lambda)$ and $\psi_X^*(t)$ when X follows a standard distribution,

	$X \sim \mathcal{N}(0,\nu)$	$X \sim \text{Poi}(\nu)$	$X \sim \text{Ber}(p)$
$\psi_X(\lambda)$	$\lambda^2\nu/2$	$\nu(\exp(\lambda) - \lambda - 1)$	$\log(p\exp(\lambda) + 1 - p) - \lambda p$
$\psi_X^*(t)$	$t^2/(2v)$	$\nu h\left(t/\nu\right)$	$D_{KL}\left(\text{Ber}(p+t) \vert\vert \text{Ber}(p)\right)$

where $h(z) := (1+z)\log(1+z) - z$ for $z \geq -1$. For instance, when X is distributed as $\mathcal{N}(0, 1/2)$, Chernoff's inequality states that $\Pr(X \geq t) \leq \exp\left(-t^2\right)$.

1.1 Upper Bounding the Log MGF

How do we use Chernoff's inequality when X does not follow a standard distribution? We generally upper bound the log MGF of X by a function $\phi_X(\lambda)$ whose corresponding conjugate $\phi_X^*(t) := \sup_{\lambda \in \mathbb{R}} \lambda t - \phi_X(\lambda)$ can be directly calculated, because then we can use

$$\Pr(X \geq t) \leq \exp\left(-\psi_X^*(t)\right) \leq \exp\left(-\phi_X^*(t)\right) \qquad \forall X, t \geq \mathbf{E}[X] \tag{1}$$

[*]Section 2 of BLM

A natural upper bound to consider is the log MGF of a standard distribution since its conjugate is known. In fact, the case with $\mathcal{N}(0,\nu)$ is so important that we have a special name for it. A random variable X is called **sub-Gaussian with variance factor** ν, denoted as $X \in \mathcal{G}(\nu)$, if its log MGF is bounded by the log MGF of $\mathcal{N}(0,\nu)$:

$$\psi_X(\lambda) \leq \frac{\lambda^2\nu}{2} \qquad\qquad \forall \lambda \in \mathbb{R}$$

This immediately gives $\Pr(X \geq t) \leq \exp(t^2/(2v))$ for $X \in \mathcal{G}(\nu)$ by (1). Noting that $\psi_{-X}(\lambda) = \psi_X(-\lambda) \leq (\lambda^2\nu)/2$, we also have $\Pr(-X \geq t) \leq \exp(t^2/(2v))$. Thus by the union bound,

$$\Pr(|X| \geq t) \leq 2\exp\left(\frac{t^2}{2v}\right) \qquad\qquad \forall X \in \mathcal{G}(\nu), t > 0 \tag{2}$$

An upper bound does not have to come from a standard distribution as long as the corresponding conjugate can be explicitly calculated. For instance, a generalization of sub-Gaussian is given by introducing a scale parameter: X is called **sub-Gamma on the right with variance factor** ν **and scale parameter** c, denoted as $X \in \Gamma_+(\nu, c)$, if

$$\psi_X(\lambda) \leq \frac{\lambda^2\nu}{2(1-c\lambda)} \qquad\qquad \forall \lambda \in \left(0, \frac{1}{c}\right)$$

Setting $\phi_X(\lambda) = \lambda^2\nu/2(1-c\lambda)$, it turns out that $\phi_X^*(t) = \sup_{\lambda\in(0,1/c)} t\lambda - \phi_X(\lambda)$ not only has a closed-form expression but also has an inverse $\phi_X^{*-1}(u) = \sqrt{2\nu u} + cu$ for $u > 0$ (p. 29, BLM). Combining it with (1), we have

$$\Pr\left(X \geq \sqrt{2\nu t} + ct\right) \leq \exp\left(-t\right) \qquad\qquad \forall X \in \Gamma_+(\nu, c), t > 0 \tag{3}$$

If $-X \in \Gamma_+(\nu, c)$, then X is called **sub-Gamma on the left** and denoted as $X \in \Gamma_-(\nu, c)$. If $X \in \Gamma_+(\nu, c) \cap \Gamma_-(\nu, c)$, then X is simply called **sub-Gamma** and denoted as $X \in \Gamma(\nu, c)$. Since $\psi_{-X}(\lambda) = \psi_X(-\lambda)$ and $\psi_X(0) = 0$, we can define $X \in \Gamma(\nu, c)$ to be

$$\psi_X(\lambda) \leq \frac{\lambda^2\nu}{2(1-c\lambda)} \qquad\qquad \forall \lambda \in \left(-\frac{1}{c}, \frac{1}{c}\right)$$

from which it is easy to see that $\Gamma(\nu, 0) = \mathcal{G}(\nu)$. Re-writing (3) for $X \in \Gamma(\nu, c)$ with the union bound, we have

$$\Pr\left(|X| \geq \sqrt{2\nu t} + ct\right) \leq 2\exp(-t) \qquad\qquad \forall X \in \Gamma(\nu, c), t > 0 \tag{4}$$

There is a good reason this generalization is called sub-"Gamma". A *centered* Gamma variable can be shown to be sub-Gamma (p. 28, BLM):

$$Y \sim \mathrm{Gamma}(a, b) \qquad \Longrightarrow \qquad X := Y - \mathbf{E}[Y] \in \Gamma(ab^2, b) \tag{5}$$

This fact is useful because we often work with a special case of the Gamma distribution: the chi-squared distribution $\chi^2(d) = \mathrm{Gamma}(d/2, 2)$.[3]

1.2 Sum of Independent Variables

Chernoff is good for analyzing a sum of independent variables because the log MGF factorizes. Let $X_1 \ldots X_n$ be independent and define $X = \sum_{i=1}^n X_i$. Then

$$\psi_X(\lambda) = \sum_{i=1}^n \psi_{X_i}(\lambda) \tag{6}$$

1.2.1 Hoeffding's Inequality

If $X_i \in [a_i, b_i]$ is bounded, Hoeffding's lemma states that $X_i - \mathbf{E}[X_i] \in \mathcal{G}((b_i - a_i)/4)$.[4] Thus $X - \mathbf{E}[X] \in \mathcal{G}(\sum_{i=1}^n (b_i - a_i)/4)$, and applying the sub-Gaussian Chernoff gives **Hoeffding's inequality**:

$$\Pr(|X - \mathbf{E}[X]| \geq t) \leq 2\exp\left(-\frac{2t^2}{\sum_{i=1}^n b_i - a_i}\right) \tag{7}$$

The case with binary variables $X_i \in \{0, 1\}$ (i.e., X is a binomial) is of special interest in machine learning because they can be used to analyze the deviation of a sample error. Let f be a target classifier and $h \in C$ be our hypothesis in some finite hypothesis space. Let $\mathrm{err}_D(h) := \Pr_{x\sim D}(h(x) \neq f(x))$ denote the "true" error of h on the the actual input distribution D, and $\mathrm{err}_S(h) := \Pr_{x\sim S}(h(x) \neq f(x))$ denote the sample error of h estimated on $S = \{x_1 \dots x_n\}$ drawn iid from D. Note that $\mathrm{err}_S(h) = (1/n)\sum_{i=1}^n X_i$ where $X_i = [[h(x_i) = f(x_i)]]$ and $\mathbf{E}_S[\mathrm{err}_S(h)] = \mathrm{err}_D(h)$. Thus for any $h \in C$, denoting $X = \sum_{i=1}^n X_i$,

$$\Pr(|\mathrm{err}_S(h) - \mathrm{err}_D(h)| > t) = \Pr\left(|X - \mathbf{E}[X]| > nt\right) \leq 2\exp\left(-2nt^2\right)$$

Combining this with the union bound, this allows us to make statements like: the chance that there is *any* hypothesis in C whose sample error estimated on S deviates from the true error by more than $t \in (0, 1)$ is at most $1 - \delta$, given that the number of samples is $|S| \geq (\log(2|C|) + \log(1/\delta))/(2t^2)$.

1.2.2 Bernstein's Inequality

One shortcoming of Hoeffding (7) is that it depends on the range rather than the actual variance of X. In cases where the variance is much smaller than the width of the range, we can benefit from inequalities that depend explicitly on the variance.

Theorem 1.1 (Bernstein)**.** *Let $X_1 \dots X_n$ be independent variables with $X_i \leq b$ for some $b > 0$. Let $X = \sum_{i=1}^n X_i$ and $\nu = \sum_{i=1}^n \boldsymbol{E}\left[X_i^2\right]$. Then for all $t > 0$,*

$$\Pr\left(X - \boldsymbol{E}[X] \geq t\right) \leq \exp\left(-\frac{t^2}{2(\nu + bt/3)}\right)$$

Proof sketch (p. 36, BLM). We can use X_i/b and fix it afterward, so assume $b = 1$ WLOG. The proof consists of upper bounding the log MGF of $X - \mathbf{E}[X]$ by the log MGF of $\mathrm{Poi}(\nu)$ so that $\Pr\left(X - \mathbf{E}[X] \geq t\right) \leq \nu h\left(t/\nu\right)$ (think "sub-Poisson") and using the inequality $h(u) \geq u^2/(2(1 + u/3))$. □

As a thought experiment, suppose we have rare event $X_i \in \{0, 1\}$, say we know $\mathbf{E}[X] \leq B$. Since $\nu = \sum_{i=1}^n \mathbf{E}\left[X_i^2\right] \leq \sum_{i=1}^n \mathbf{E}[X_i] = \mathbf{E}[X]$, Bernstein gives us

$$\Pr(X \geq \mathbf{E}[X] + B) \leq \exp\left(-\frac{B^2}{2(B + B/3)}\right) \leq \exp\left(-\frac{B}{4}\right)$$

On the other hand, Hoeffding gives us

$$\Pr(X \geq \mathbf{E}[X] + B) \leq \exp\left(-\frac{2B^2}{n}\right)$$

So for the purpose of bounding $\Pr(X \geq 2B) \leq \Pr(X \geq \mathbf{E}[X] + B)$, Bernstein can be much sharper if B is small relative to n. For instance, if $B = n^{1/4}$,

$$\Pr(X \geq 2\sqrt{n}) \leq \exp\left(-\frac{n^{1/4}}{4}\right) \qquad \xrightarrow[n\to\infty]{} 0 \qquad \text{(Bernstein)}$$
$$\Pr(X \geq 2\sqrt{n}) \leq \exp\left(-\frac{2}{\sqrt{n}}\right) \qquad \xrightarrow[n\to\infty]{} 1 \qquad \text{(Hoeffding)}$$

2 Examples

2.1 Length-Preserving Transformation

What is a *random* matrix $W \in \mathbb{R}^{m\times d}$ such that

$$\mathbf{E}\left[\left|\left|Wu\right|\right|^2\right] = 1 \qquad\qquad \forall u \in \mathbb{R}^d:\ \left|\left|u\right|\right|^2 = 1$$

If we define the i-th row of W to be $w_i/\sqrt{m}$ where $w_i \sim \mathcal{N}(0, I_{d\times d})$, then since $w_i^\top u$ is distributed as $\mathcal{N}(0, u^\top u) = \mathcal{N}(0,1)$ with $\mathbf{E}\left[\left(w_i^\top u\right)^2\right] = 1$, we have a desired matrix:

$$\mathbf{E}\left[\left|\left|Wu\right|\right|^2\right] = \frac{1}{m}\sum_{i=1}^m \mathbf{E}\left[\left(w_i^\top u\right)^2\right] = 1 \qquad\qquad \forall u \in \mathbb{R}^d:\ \left|\left|u\right|\right|^2 = 1$$

This transformation can be seen as projecting a direction in $\mathbb{R}^d$ onto a random m-dimensional subspace while maintaining its unit length. Suppose we have a finite set of directions in $\mathbb{R}^d$,

$$S = \left\{u \in \mathbb{R}^d:\ \left|\left|u\right|\right|^2 = 1\right\} \qquad\qquad |S| < \infty$$

How many dimensions m do we need to “sample” to ensure that the length of every $u \in S$ is concentrated around 1 when projected?

Sum of squared normals. Pick any $u \in S$. Since $m\left|\left|Wu\right|\right|^2$ is a sum of m squared normals, it is distributed as $\chi^2(m) = \text{Gamma}(m/2, 2)$ and thus $m\left|\left|Wu\right|\right|^2 - m \in \Gamma(2m, 2)$. Then by the union bound and the sub-Gamma Chernoff (4),

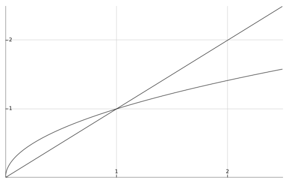
$$\begin{aligned}
\Pr\left(\exists u \in S:\ \left|\left|\left|Wu\right|\right|^2 - 1\right| \geq 2\sqrt{\frac{t}{m}} + 2\frac{t}{m}\right) &\leq \sum_{u\in S}\Pr\left(\left|\left|\left|Wu\right|\right|^2 - 1\right| \geq 2\sqrt{\frac{t}{m}} + 2\frac{t}{m}\right)\\
&= \sum_{u\in S}\Pr\left(\left|m\left|\left|Wu\right|\right|^2 - m\right| \geq 2\sqrt{mt} + 2t\right)\\
&\leq 2\left|S\right|\exp(-t)
\end{aligned}$$

Aside: solving an inequality. For any given $\epsilon > 0$, we want a simple characterization of m satisfying $\sqrt{t/m} + t/m \leq \epsilon/2$ so that we can make the statement

$$\Pr\left(\exists u \in S:\ \left|\left|\left|Wu\right|\right|^2 - 1\right| \geq \epsilon\right) \leq \Pr\left(\exists u \in S:\ \left|\left|\left|Wu\right|\right|^2 - 1\right| \geq 2\sqrt{\frac{t}{m}} + 2\frac{t}{m}\right)$$

Solving for a variable in an inequality can be messy: one such way is to substitute $x = \sqrt{t/m}$ and find $x \geq 0$ such that $x^2 + x - \epsilon/2 \leq 0$ using the quadratic formula. But the following observations greatly simplify the argument:

- We can upper bound $\sqrt{t/m} + t/m$ by a simpler function $g(m)$ and then solve for m satisfying $g(m) \leq \epsilon/2$ (since this implies $\sqrt{t/m} + t/m \leq \epsilon/2$).
- For any $x \geq 0$, $\sqrt{x}$ is an upper bound if $x \leq 1$:



- Therefore, *if we assume* $m \geq t$, then $\sqrt{t/m} + t/m \leq 2\sqrt{t/m} = g(m)$. Solving for m in $2\sqrt{t/m} \leq \epsilon/2$, we get $m \geq 16t/\epsilon^2$.
- Was that a reasonable assumption to make? It follows if we *restrict our setting to small deviation*, say we always assume $\epsilon \leq 1$, since then $\sqrt{t/m} + t/m \leq \epsilon/2$ cannot be true for $m < t$.

Setting $\delta = 2\left|S\right| \exp(-t)$ so that $t = \log(2\left|S\right|/\delta)$, we have the following result: given any $\epsilon, \delta \in (0, 1)$, if

$$m \geq \frac{16}{\epsilon^2} \log \frac{2\left|S\right|}{\delta}$$

then with probability at least $1 - \delta$, every $u \in S$ satisfies

$$1 - \epsilon < \left|\left|Wu\right|\right|^2 < 1 + \epsilon$$

In particular, note that the number of sample dimensions m does not depend on the original dimension d. This is because we never needed the information: we only worked with m random projections $w_i^\top u$ and used their Gaussian property.

Johnson-Lindenstrauss lemma Suppose we have a finite set of arbitrary vectors $S' \subset \mathbb{R}^d$. What can we say about their *pairwise* distances when projected by the length-preserving transformation W above? We construct a set of unit vectors $S := \left\{x - x' / \left|\left|x - x'\right|\right|^2 : \; x, x' \in S'\right\}$ which has at most $\left|S'\right|^2$ elements. We now apply the above result: given any $\epsilon, \delta \in (0, 1)$, if

$$m \geq \frac{32}{\epsilon^2} \log \frac{2\left|S'\right|}{\sqrt{\delta}}$$

then with probability at least $1 - \delta$, every $x, x' \in S'$ satisfies

$$(1 - \epsilon) \left|\left|x - x'\right|\right|^2 < \left|\left|Wx - Wx'\right|\right|^2 < (1 + \epsilon) \left|\left|x - x'\right|\right|^2$$

This celebrated fact is known as the Johnson-Lindenstrauss lemma.

2.2 Quadratic Polynomial

Let $X \sim \mathcal{N}(0, I_{d \times d})$ and define $Z = X^\top A X$ to be a quadratic polynomial of X for a symmetric matrix $A \in \mathbb{R}^{d \times d}$. We are interested in understanding the concentration properties of Z. First, note that if $A = I_{d \times d}$ then $Z = \sum_{i=1}^d X_i^2$ is distributed as $\chi^2(d)$

and we can just use the sub-Gamma Chernoff on $Z - d \in \Gamma(2d, 2)$. More generally, the concentration properties of Z will depend on the spectral properties of A.

Let $A = U\Lambda U^\top$ denote an eigendecomposition of A where $\Lambda = \mathrm{diag}(\lambda_1 \ldots \lambda_d)$ is a diagonal matrix of real-valued (but not necessarily non-negative) eigenvalues. We follow the example considered in BLM (Example 2.12) and use A such that $A_{i,i} = 0$ for all $i = 1 \ldots d$; this makes $\mathrm{Tr}(A) = \sum_{i=1}^d A_{i,i} = \sum_{i=1}^d \lambda_i = 0$ and Z sub-Gamma as shown by the following argument.

Define $Y = U^\top X$ to be a rotation of X, thus also distributed as $\mathcal{N}(0, I_{d\times d})$. Then

$$Z = X^\top AX = Y^\top \Lambda Y = \sum_{i=1}^d \lambda_i Y_i^2 = \sum_{i=1}^d \lambda_i Y_i^2 - \left(\sum_{i=1}^d \lambda_i\right) = \sum_{i=1}^d \lambda_i \left(Y_i^2 - 1\right)$$

which has zero mean. We can explicitly work out the log MGF of Z to incorporate λ_i thanks to the factorization of the log MGF "aligns" with λ_i. Specifically, we can show that for all $\lambda \in (0, 1/(2\max_i \lambda_i))$,

$$\psi_Z(\lambda) = \sum_{i=1}^d \psi_{\lambda_i\left(Y_i^2-1\right)}(\lambda) = \sum_{i=1}^d \frac{1}{2}\left(-\log\left(1 - 2\lambda_i\lambda\right) - 2\lambda_i\lambda\right) \leq \frac{\lambda^2 \left|\left|A\right|\right|_F^2}{1 - 2\left|\left|A\right|\right|_2 \lambda}$$

where the second equality can be verified by direct calculation; we refer to BLM (p. 39) for the inequality. The important point is that this shows $Z \in \Gamma_+(2\left|\left|A\right|\right|_F^2, 2\left|\left|A\right|\right|_2)$ and we can use the sub-Gamma Chernoff (4): for all $t > 0$,

$$\Pr\left(Z \geq 2\left|\left|A\right|\right|_F \sqrt{t} + 2\left|\left|A\right|\right|_2 t\right) \leq \exp(-t)$$

Thus the larger the matrix A is in the Frobenius norm $\left|\left|A\right|\right|_F = \sqrt{\sum_i \lambda_i^2}$ or the operator norm $\left|\left|A\right|\right|_2 = \max_i |\lambda_i|$, the looser the bound is on the concentration of Z around 0.

Reference. *Concentration Inequalities* (Boucheron, Lugosi, and Massart)

Notes

[1]Pick any $\alpha \in [0, 1]$. The key step uses Hölder's inequality $\mathbf{E}\left[|X_1 X_2|\right] \leq \mathbf{E}\left[|X_1|^p\right]^{1/p} + \mathbf{E}\left[|X_2|^q\right]^{1/q}$ with $p = 1/\alpha$ and $q = 1/(1-\alpha)$:

$$\begin{aligned}\psi_X(\alpha\lambda_1 + (1-\alpha)\lambda_2) &= \log \mathbf{E}[\exp(\alpha\lambda_1 X)\exp(\alpha\lambda_2 X)] \\ &\leq \log \mathbf{E}[\exp(\lambda_1 X)]^\alpha \mathbf{E}[\exp(\lambda_2 X)]^{1-\alpha} = \alpha\psi_X(\lambda_1) + (1-\alpha)\psi_X(\lambda_2)\end{aligned}$$

[2]To see this, note that $\lambda t - \psi_X(\lambda) \leq z(t - \mathbf{E}[X])$ by Jensen's and is negative only if $z < 0$. On the other hand, $\lambda t - \psi_X(\lambda)$ is zero at $\lambda = 0$.

[3]For instance, if we have iid $Z \sim \mathcal{N}(0, \nu I_d)$, then

$$Y := \frac{1}{\nu^2}\left|\left|Z\right|\right|^2$$

is distributed as $\chi^2(d)$. Then, by (5), $Y - d/v \in \Gamma(2d, 2)$. This allows us to use sub-Gamma tools such as (4) and derive statements such as

$$\Pr\left(\left|\left|Z\right|\right|^2 > \mathbf{E}\left|\left|Z\right|\right|^2 + 2\nu^2\left(t + \sqrt{dt}\right)\right) \leq \exp(-t) \qquad \forall t > 0$$

[4]Consider Z such that $Z = [a, b]$ and $\mathbf{E}[Z] = 0$. Then by Taylor's theorem,

$$\psi_Z(\lambda) = \psi_Z(0) + \psi_Z'(0)\lambda + \psi_Z''(\xi)\frac{\lambda^2}{2}$$

for some $\xi \in [0, \lambda]$. We have $\psi_Z(0) = 0$ and $\psi_Z'(0) = \mathbf{E}[X] = 0$, and the proof of Hoeffding's lemma consists of bounding $\psi_Z'' \leq (b-a)/4$. Then it follows $\psi_Z(\lambda) \leq \lambda^2\nu/2$ where $\nu = (b-a)/4$.