Concentration Inequalities

Garrett Thomas

November 15, 2018

1 About

This document is part of a series of notes about math and machine learning. You are free to distribute it as you wish. The latest version can be found at http://gwthomas.github.io/notes. Please report any errors to gwthomas@stanford.edu.

It is often useful to bound the probability that a random variable deviates from some other value, usually its mean. Here we present various **concentration inequalities** of this flavor.

2 Markov and Chebyshev

We first show **Markov's inequality**, which is widely applicable, and indeed used to prove several later inequalities.

Proposition 1. *(Markov's inequality) If X is a nonnegative random variable and $\epsilon > 0$, then*

$$\mathbb{P}(X \geq \epsilon) \leq \frac{\mathbb{E}[X]}{\epsilon}$$

Proof. Define a new random variable

$$Y = \epsilon 1_{\{X \geq \epsilon\}} \quad \text{i.e.} \quad Y(\omega) = \begin{cases} \epsilon & X(\omega) \geq \epsilon \\ 0 & X(\omega) < \epsilon \end{cases}$$

noting that $Y \leq X$. Then, using monotonicity and linearity of expectation,

$$\mathbb{E}[X] \geq \mathbb{E}[Y] = \mathbb{E}[\epsilon 1_{\{X \geq \epsilon\}}] = \epsilon \mathbb{P}(X \geq \epsilon)$$

Dividing through by ϵ gives the claimed inequality. □

A useful corollary is **Chebyshev's inequality**, which expresses concentration in terms of the variance.

Proposition 2. *(Chebyshev's inequality) If X is a random variable with finite mean and variance, then for any $\epsilon > 0$*

$$\mathbb{P}(|X - \mathbb{E}[X]| \geq \epsilon) \leq \frac{\mathrm{Var}(X)}{\epsilon^2}$$

In particular, for any $k > 0$,

$$\mathbb{P}\left(|X - \mathbb{E}[X]| \geq k\sqrt{\mathrm{Var}(X)}\right) \leq \frac{1}{k^2}$$

Proof. Applying Markov's inequality to the nonnegative random variable $(X - \mathbb{E}[X])^2$ gives

$$\mathbb{P}(|X - \mathbb{E}[X]| \geq \epsilon) = \mathbb{P}((X - \mathbb{E}[X])^2 \geq \epsilon^2) \leq \frac{\mathbb{E}[(X - \mathbb{E}[X])^2]}{\epsilon^2} = \frac{\mathrm{Var}(X)}{\epsilon^2}$$

as claimed. The second inequality is a special case of the first where $\epsilon = k\sqrt{\mathrm{Var}(X)}$. □

Chebyshev's inequality tells us that the probability of X falling more than k standard deviations from its mean (in either direction) is at most $1/k^2$. The power of Chebyshev's inequality is that it is widely applicable – it only requires that X have finite mean and variance. Tighter bounds can often be obtained if we know more specific information about the distribution of X.

3 Chernoff bounds, (sub-)Gaussian tails

To motivate, observe that even if a random variable X can be negative, we can apply Markov's inequality to e^X, which is always positive. Or, more generally, we can apply it to e^{tX} for $t > 0$, then optimize the bound over the choice of t.

Let $\epsilon > 0$. For any $t > 0$, we have

$$\mathbb{P}(X \geq \epsilon) = \mathbb{P}(e^{tX} \geq e^{t\epsilon}) \leq \frac{\mathbb{E}[e^{tX}]}{e^{t\epsilon}} = e^{-t\epsilon} M_X(t)$$

where $M_X(t) = \mathbb{E}[e^{tX}]$ is the **moment-generating function of** X. Since this bound holds for all $t > 0$, we have

$$\mathbb{P}(X \geq \epsilon) \leq \inf_{t>0} e^{-t\epsilon} M_X(t)$$

This is the strategy of the general **Chernoff bound**. In practice, we usually want to bound the probability that X deviates from its mean, so we will apply the results above to the random variable $X - \mathbb{E}[X]$.

We know that the Gaussian distribution has rapidly decaying tails, so let us aim for a bound similar to what one would get for a Gaussian. We compute that if $Z \sim \mathcal{N}(0, \sigma^2)$, then

$$\begin{aligned}
M_Z(t) = \mathbb{E}[e^{tZ}] &= \int_{-\infty}^{\infty} e^{tx} \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{x^2}{2\sigma^2}\right) \mathrm{d}x \\
&= \int_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{x^2 - 2\sigma^2 tx}{2\sigma^2}\right) \mathrm{d}x \\
&= \int_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{x^2 - 2\sigma^2 tx + \sigma^4 t^2}{2\sigma^2}\right) \exp\left(\frac{\sigma^4 t^2}{2\sigma^2}\right) \mathrm{d}x \\
&= \exp\left(\frac{\sigma^2 t^2}{2}\right) \underbrace{\int_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(x - \sigma^2 t)^2}{2\sigma^2}\right) \mathrm{d}x}_{1} \\
&= \exp\left(\frac{\sigma^2 t^2}{2}\right)
\end{aligned}$$

In fact, we only require an upper bound on the moment-generating function. Therefore we say that a random variable X is **sub-Gaussian with variance proxy** σ^2 (or simply σ^2**-sub-Gaussian**) if

$$M_{X - \mathbb{E}[X]}(t) = \mathbb{E}[e^{t(X - \mathbb{E}[X])}] \leq \exp\left(\frac{\sigma^2 t^2}{2}\right)$$

for all $t \in \mathbb{R}$.

Proposition 3. *Suppose X is sub-Gaussian with variance proxy σ^2. Then for any $\epsilon > 0$,*

$$\mathbb{P}(X - \mathbb{E}[X] \geq \epsilon) \leq \exp\left(-\frac{\epsilon^2}{2\sigma^2}\right)$$
$$\mathbb{P}(\mathbb{E}[X] - X \geq \epsilon) \leq \exp\left(-\frac{\epsilon^2}{2\sigma^2}\right)$$

Before giving the proof, we remark that these can be combined into a two-sided bound as follows:

$$\mathbb{P}(|X - \mathbb{E}[X]| \geq \epsilon) \leq \mathbb{P}(X - \mathbb{E}[X] \geq \epsilon) + \mathbb{P}(\mathbb{E}[X] - X \geq \epsilon) \leq 2\exp\left(-\frac{\epsilon^2}{2\sigma^2}\right)$$

Proof. Applying Markov's inequality and sub-Gaussianity,

$$\mathbb{P}(X - \mathbb{E}[X] \geq \epsilon) \leq \frac{\mathbb{E}[e^{t(X-\mathbb{E}[X])}]}{e^{t\epsilon}} \leq \exp\left(\frac{\sigma^2 t^2}{2} - t\epsilon\right)$$

Since this holds for every $t > 0$, let us make the upper bound as small as possible by optimizing over positive t. We set the derivative to zero:

$$0 = \frac{\mathrm{d}}{\mathrm{d}t}\exp\left(\frac{\sigma^2 t^2}{2} - t\epsilon\right) = \exp\left(\frac{\sigma^2 t^2}{2} - t\epsilon\right)(\sigma^2 t - \epsilon)$$

which only holds if $t = \epsilon/\sigma^2$. Plugging this back in,

$$\mathbb{P}(X - \mathbb{E}[X] \geq \epsilon) \leq \exp\left(\frac{\sigma^2}{2}\left(\frac{\epsilon}{\sigma^2}\right)^2 - \left(\frac{\epsilon}{\sigma^2}\right)\epsilon\right) = \exp\left(-\frac{\epsilon^2}{2\sigma^2}\right)$$

Similarly, since the inequality defining sub-Gaussianity holds for any t (even $t < 0$), we have

$$\mathbb{P}(\mathbb{E}[X] - X \geq \epsilon) = \mathbb{P}(e^{-t(X-\mathbb{E}[X])} \geq e^{t\epsilon}) \leq \frac{\mathbb{E}[e^{-t(X-\mathbb{E}[X])}]}{e^{t\epsilon}} \leq \exp\left(\frac{\sigma^2 t^2}{2} - t\epsilon\right)$$

for any $t > 0$. This is the same bound as before, so optimizing over $t > 0$ gives the same result. □

We now give some useful algebraic properties of sub-Gaussian variables.

Proposition 4. *(i) If X_1 and X_2 are independent sub-Gaussian random variables with respective variance proxies σ_1^2 and σ_2^2, then $X_1 + X_2$ is sub-Gaussian with variance proxy $\sigma_1^2 + \sigma_2^2$.*

(ii) If X is a sub-Gaussian random variable with variance proxy σ^2 and $c \neq 0$ is a constant, then cX is sub-Gaussian with variance proxy $c^2\sigma^2$.

Proof. (i) For any $t \in \mathbb{R}$,

$$\begin{aligned}
\mathbb{E}[e^{t(X_1+X_2-\mathbb{E}[X_1+X_2])}] &= \mathbb{E}[e^{t(X_1-\mathbb{E}[X_1])}e^{t(X_2-\mathbb{E}[X_2])}] && \text{property of exp} \\
&= \mathbb{E}[e^{t(X_1-\mathbb{E}[X_1])}]\mathbb{E}[e^{t(X_2-\mathbb{E}[X_2])}] && \text{independence} \\
&\leq \exp\left(\frac{\sigma_1^2 t^2}{2}\right)\exp\left(\frac{\sigma_2^2 t^2}{2}\right) && \text{sub-Gaussianity} \\
&= \exp\left(\frac{(\sigma_1^2 + \sigma_2^2)t^2}{2}\right)
\end{aligned}$$

so $X_1 + X_2$ is sub-Gaussian with variance proxy $\sigma_1^2 + \sigma_2^2$.

(ii) For any $t \in \mathbb{R}$,

$$\mathbb{E}[e^{t(cX - \mathbb{E}[cX])}] = \mathbb{E}[e^{(tc)(X - \mathbb{E}[X])}] \leq \exp\left(\frac{\sigma^2 (tc)^2}{2}\right) = \exp\left(\frac{(c^2\sigma^2)t^2}{2}\right)$$

so cX is sub-Gaussian with variance proxy $c^2\sigma^2$.

□

In particular, the previous result implies the following bounds for sums/means of independent sub-Gaussian random variables:

Proposition 5. *Suppose $X_1, \ldots, X_n$ are independent sub-Gaussian random variables with variance proxies $\sigma_1^2, \ldots, \sigma_n^2$. Let $S_n = \sum_{i=1}^n X_i$ and $\bar{X}_n = \frac{1}{n} S_n$. Then*

$$\mathbb{P}(|S_n - \mathbb{E}[S_n]| \geq \epsilon) \leq 2 \exp\left(-\frac{\epsilon^2}{2\sum_{i=1}^n \sigma_i^2}\right)$$

and

$$\mathbb{P}(|\bar{X}_n - \mathbb{E}[\bar{X}_n]| \geq \epsilon) \leq 2 \exp\left(-\frac{n^2\epsilon^2}{2\sum_{i=1}^n \sigma_i^2}\right)$$

Proof. By the previous result, S_n is sub-Gaussian with variance proxy $\sum_{i=1}^n \sigma_i^2$ (which implies the first inequality), and therefore $\bar{X}_n$ is sub-Gaussian with variance proxy $\frac{1}{n^2}\sum_{i=1}^n \sigma_i^2$ (which implies the second inequality). □

We remark that the one-sided versions of the inequalities above also hold without the leading factor of 2. (Indeed, these inequalities are implicitly used to prove the two-sided versions stated.)

3.1 Hoeffding

Hoeffding tells us that bounded random variables are sub-Gaussian and therefore concentrate.

Proposition 6. *(Hoeffding's lemma) Suppose X is a random variable such that $a \leq X \leq b$ almost surely. Then X is sub-Gaussian with variance proxy $\frac{(b-a)^2}{4}$, i.e. for any $t \in \mathbb{R}$*

$$\mathbb{E}[e^{t(X - \mathbb{E}[X])}] \leq \exp\left(\frac{(b-a)^2 t^2}{8}\right)$$

Proof. Let $t \in \mathbb{R}$. Since $x \mapsto e^{tx}$ is convex, it is upper bounded by the secant line: for any $x \in [a, b]$,

$$e^{tx} \leq \frac{b - x}{b - a} e^{ta} + \frac{x - a}{b - a} e^{tb}$$

Thus

$$\mathbb{E}[e^{tX}] \leq \frac{b - \mathbb{E}[X]}{b - a} e^{ta} + \frac{\mathbb{E}[X] - a}{b - a} e^{tb}$$

First assume that $\mathbb{E}[X] = 0$. Then we must have $a \leq 0$ and $0 \leq b$.

Rewrite

$$\begin{aligned}
\mathbb{E}[e^{tX}] &\le \frac{b}{b-a}e^{ta} + \frac{-a}{b-a}e^{tb} \\
&= (1-\theta)e^{ta} + \theta e^{tb} && \theta \triangleq -a/(b-a) \\
&= (1-\theta+\theta e^{t(b-a)})e^{ta} \\
&= (1-\theta+\theta e^{u})e^{-\theta u} && u \triangleq t(b-a)
\end{aligned}$$

Note that u is arbitrary because t is, but $\theta > 0$ (we will use this fact later). Define

$$\phi(u) = \log\Big((1-\theta+\theta e^{u})e^{-\theta u}\Big) = \log(1-\theta+\theta e^{u}) - \theta u$$

noting that this is well-defined because

$$\begin{aligned}
1-\theta+\theta e^{u} &= \theta\left(\frac{1}{\theta} - 1 + e^{u}\right) \\
&= \theta\left(-\frac{b}{a} + e^{u}\right) \\
&> 0
\end{aligned}$$

and that as a consequence of the previous calculations, $\mathbb{E}[e^{tX}] \le \exp\big(\phi(u)\big)$. Therefore it remains to bound $\phi(u)$.

By Taylor's theorem, for any $u \in \mathbb{R}$ there exists v between 0 and u such that

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

We compute

$$\begin{aligned}
\phi(0) &= \log(1-\theta+\theta) - 0 = 0 \\
\phi'(0) &= \frac{\theta e^{u}}{1-\theta+\theta e^{u}} - \theta\Big|_{u=0} \\
&= \theta - \theta = 0 \\
\phi''(v) &= \frac{\theta e^{v}(1-\theta+\theta e^{v}) - \theta^2 e^{2v}}{(1-\theta+\theta e^{v})^2} \\
&= \frac{\theta e^{v}}{1-\theta+\theta e^{v}}\left(1 - \frac{\theta e^{v}}{1-\theta+\theta e^{v}}\right) \\
&\le \frac{1}{4}
\end{aligned}$$

where the last inequality holds because $a(1-a) \le \frac{1}{4}$ for any $a > 0$. Thus

$$\phi(u) \le 0 + u \cdot 0 + \frac{1}{2}u^2 \cdot \frac{1}{4} = \frac{u^2}{8} = \frac{(b-a)^2t^2}{8}$$

and

$$\mathbb{E}[e^{tX}] \le \exp\big(\phi(u)\big) \le \exp\left(\frac{(b-a)^2t^2}{8}\right)$$

Now consider a general X which may have nonzero mean. The random variable $Y = X - \mathbb{E}[X]$ satisfies $\mathbb{E}[Y] = 0$ and $a - \mathbb{E}[X] \le Y \le b - \mathbb{E}[X]$ almost surely, so by the case already shown,

$$\mathbb{E}[e^{t(X-\mathbb{E}[X])}] = \mathbb{E}[e^{tY}] \le \exp\left(\frac{(b-\mathbb{E}[X]-(a-\mathbb{E}[X]))^2t^2}{8}\right) = \exp\left(\frac{(b-a)^2t^2}{8}\right)$$

□

Combining Hoeffding's lemma with the previous bound for sums/averages of independent sub-Gaussian random variables leads directly to **Hoeffding's inequality**.

Proposition 7. *(Hoeffding's inequality) Suppose $X_1, \ldots, X_n$ are independent random variables such that $a_i \leq X_i \leq b_i$ almost surely for each $i = 1, \ldots, n$. Let $S_n = \sum_{i=1}^n X_i$ and $\bar{X}_n = \frac{1}{n} S_n$. Then*

$$\mathbb{P}\left(|S_n - \mathbb{E}[S_n]| \geq \epsilon\right) \leq 2 \exp\left(-\frac{2\epsilon^2}{\sum_{i=1}^n (b_i - a_i)^2}\right)$$

and

$$\mathbb{P}\left(|\bar{X}_n - \mathbb{E}[\bar{X}_n]| \geq \epsilon\right) \leq 2 \exp\left(-\frac{2n^2\epsilon^2}{\sum_{i=1}^n (b_i - a_i)^2}\right)$$

As before, one can also use the one-sided versions without the leading factor of 2.