Exponential Tail Bounds

Mathias Winther Madsen

January 12, 2015

Here's a warm-up problem to get you started:

Problem 1. You enter the casino with 100 chips and start playing a game in which you double your capital with 75% probability and decrease it by a factor of 10 with 25% probability.

Is it a good or a bad idea? What do you expect your capital to be after $t = 1000$ games?

1 Moment-Generating Functions

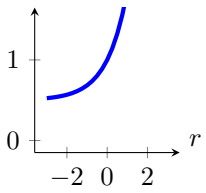
Definition 2. The **moment-generating function** of a random variable X is the function

$$G(r) \;=\; E\left[e^{rX}\right]$$

whenever these expected values are well-defined.

Example 3. Let X be a coin flip with $\Pr\{X = 1\} = p$ and $\Pr\{X = 0\} = 1 - p$. Then $\exp(rX)$ is a random variable with distribution

g	e^r	1
$\Pr\{G(r) = g\}$	p	$1-p$



The mean of this variable is

$$G(r) \;=\; pe^r + 1 - p.$$

Let X be a geomtric random variable with the distribution

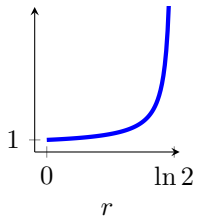
x	1	2	3	4	$\cdots$
$\Pr\{X = x\}$	$1/2$	$1/4$	$1/8$	$1/16$	$\cdots$

This describes the time you have to wait for a fair coin to come up heads the in a series of tosses. The variable $\exp(rX)$ then has the distribution

g	e^r	e^{2r}	e^{3r}	e^{4r}	$\cdots$
$\Pr\{G(r) = g\}$	$1/2$	$1/4$	$1/8$	$1/16$	$\cdots$

We can find the expected value of such a variable by using the sum formula $\sum_{k=1}^{\infty} \alpha^k = \alpha/(1-\alpha)$ for a geometric series:

$$\sum_{k=1}^{\infty} \frac{e^{kr}}{2^k} \;=\; \sum_{k=1}^{\infty} \left(\frac{e^r}{2}\right)^k \;=\; \frac{(e^r/2)}{1-(e^r/2)} \;=\; \frac{e^r}{2-e^r}. \qquad (r < \ln 2)$$



The moment-generating function for X is thus $G(r) = e^r/(2-e^r)$. The function is only defined when $r < \ln 2$. For $r \geq \ln 2$, the sum does not converge, and $G(r)$ is not defined.

Note that, in general,

$$e^{rE[X]} \;\leq\; E\left[e^{rX}\right].$$

This holds because the variation in X will be shifted to the right when it is exponentiated. Hence, a symmetric variation of X around $E[X]$ will generally map onto an asymmetric variation around $\exp(rX)$, skewed towards larger values (cf. Fig. 1).

2 Degenerate Moments

When X is a random variable, $\exp(rX)$ is also a perfectly fine random variable. In fact, if X has the probability densities $p(x)$, then $Y = \exp(rX)$ has the probability densities

$$p\left(\frac{\ln y}{r}\right) \times \frac{1}{ry}.$$

This distribution is, however, skewed significantly more to the right than the original distribution, and it might therefore not have a mean. In these situations, the moment-generating function of X is not well-defined. The following example illustrates an extreme case of this phenomenon.

Example 4. Suppose X has a probability density given by the power law

$$p(x) = \frac{1}{x^2} \qquad (x \geq 1).$$

This is an example of a **Pareto distribution**. To find its moment-generating function, we compute

$$E\left[e^{rX}\right] = \int_1^\infty e^{rx}\, dp(x) = \int_1^\infty \frac{e^{rx}}{x^2}\, dx = \infty.$$

The exponential increase in $\exp(rx)$ here far outpaces the polynomial decrease in $p(x)$, and thus $\exp(rX)$ does not have an expected value for $r > 0$. Note, however, that $\exp(rX)$ is a perfectly respectable random variable with probability density

$$p(y) = \frac{1}{y \ln(y)^2} \qquad (y \geq e).$$

This distribution is just so extremely tail-heavy that the integral of $p(y) \times y$ diverges to infinity.

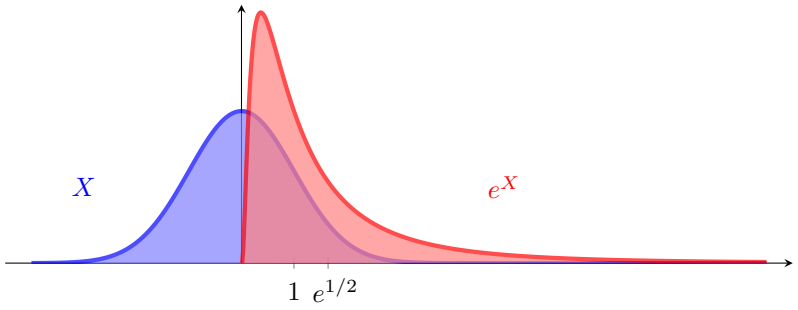


Figure 1: The probability density of a variable X with a standard normal distribution, and the more heavy-tailed density of the variable $\exp(X)$. Note that $\exp(E[X]) = e^0 = 1$, while $E[\exp(X)] = \exp(1/2)$.

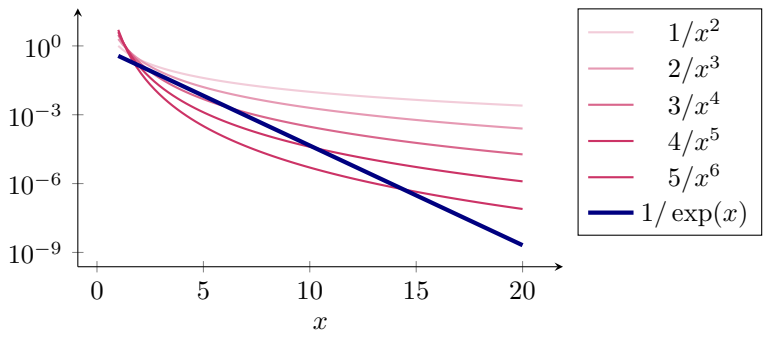


Figure 2: The exponential distribution decreases faster than power-law distribution such as the Pareto distribution.

So again, the moment-generating function of a random variable may not always exist. In fact, the exponential function has the Taylor expansion

$$e^{rX} \;=\; 1 + rx + \frac{r^2x^2}{2} + \frac{r^3x^3}{6} + \frac{r^4x^4}{24} + \cdots$$

Hence, the random variable $\exp(rX)$ only has a well-defined when it has moments of all orders — that is, when all the expected values

$$E\left[X\right], E\left[X^2\right], E\left[X^3\right], E\left[X^4\right], \ldots$$

exist. This can fail if the tails of the distribution of X decrease too slowly.

Example 5. As a generalization of the example above, consider the Pareto distribution

$$p(x) \;=\; \frac{m}{x^{m+1}} \qquad (x \geq 1)$$

does not have moments of orders higher than m, since for $k \geq m$,

$$m \int \frac{x^k}{x^{m+1}}\, dx \;\leq\; m \int \frac{1}{x}\, dx \;=\; \infty.$$

While the Pareto distribution thus has a mean and a variance if $m \geq 3$, it never has a moment-generating function. This dismal situation should be contrasted with, for instance, the rapidly decreasing probability density

$$p(x) = \frac{1}{e^x} \qquad (x \geq 1)$$

which admits of moments of any order and thus a moment-generating function.

The conclusion we can draw from this is that if you know that $G(r)$ is well-defined, then you already know that X has moments of arbitrarily high order. This observation is important to keep in mind in the following, where we use moment-generating functions to prove some results that may occasionally seem too good to be true.

3 Arithmetic and Exponential Divergence

Suppose I draw a series of samples from some probability distribution, one per time period, and record my results in the sum variable

$$S_t \;=\; X_1 + X_2 + \cdots + X_t.$$

We then know, by the linearity of expectations, that S_t grows approximately linearly in t. Thus, if we were to compare the stochastic process

$$S_1, S_2, S_3, S_4, \ldots$$

to the deterministic process

$$q, 2q, 3q, 4q, \ldots$$

we would expect the two processes to diverge at a linear speed. We could use Markov's inequality or other tools to bound the probability that they were less than $\varepsilon > 0$ apart after t steps, and this would give us the polynomials bounds we already know.

But let's instead look at the exponentiated stochastic process

$$e^{rS_1}, e^{rS_2}, e^{rS_3}, e^{rS_4}, \ldots$$

At each step, this process grows by the random factor of $\exp(rX)$. It is therefore expected to grow by a factor of $E[\exp(rX)] = G(r)$ per unit of time. This means that its central tendency is exponential rather than linear. It therefore makes sense to compare it to another exponentially increasing process of the form

$$e^{rq}, e^{2rq}, e^{3rq}, e^{4rq}, \ldots$$

This reference process grows by a factor of $\exp(rq)$ in each step.

Thus, if $\exp(rq) > G(r)$, the deterministic process is expected to outgrow the stochastic process at an exponential rate. Because of this rapid divergence, we will be able to derive exponential bounds on the probability of the two are less than $\varepsilon > 0$ apart. We shall see detailed proofs of this below, but we will first formalize the general underlying principle.

Theorem 6. *Suppose that $S_t = X_1 + X_2 + \cdots + X_t$ is a sum of independent and identically distributed random variables with a common moment-generating function G. Then for any r,*

$$\Pr\{S_t \geq qt\} \;\leq\; \left(\frac{G(r)}{e^{rq}}\right)^t.$$

Proof. We first note that $S_t \geq qt$ is equivalent to $\exp(rS_t) \geq \exp(rqt)$. We can therefore use Markov's inequality to conclude that

$$\Pr\{S_t \geq qt\} \;=\; \Pr\left\{e^{rS_t} \geq e^{rqt}\right\} \;\leq\; \frac{E\left[e^{rS_t}\right]}{e^{rqt}}.$$

In working on an artificial example, I discovered that I was using the Central Limit Theorem for large deviations where it did not apply. This led me to derive the asymptotic upper and lower bounds that were needed for the tail probabilities. [Herman] Rubin claimed he could get these bounds with much less work and I challenged him. He produced a rather simple argument, using the Markov inequality, for the upper bound. Since that seemed to be a minor lemma in the ensuing paper I published (Chernoff, 1952), I neglected to give him credit. I now consider it a serious error in judgment, especially because his result is stronger, for the upper bound, than the asymptotic result I had derived.

Herman Chernoff: "A career in statistics," in *Past, Present, and Future of Statistical Science* (2014), section 3.3.

It remains to be shown that $E\left[\exp(rS_t)\right] = G(r)^t$.

When two random variables are independent, then the expectation of their product is equal to the product of their expectations, $E[AB] = E[A]E[B]$. Since the X_i's are independent, then so are the $\exp(rX_i)$'s, and thus

$$\begin{aligned} E\left[\exp\left(S_t\right)\right] &= E\left[\exp(rX_1 + rX_2 + \cdots + rX_t)\right] \\ &= E\left[\exp(rX_1) \cdot \exp(rX_2) \cdots \exp(rX_t)\right] \\ &= E\left[\exp\left(rX_1\right)\right] \cdot E\left[\exp\left(rX_2\right)\right] \cdots E\left[\exp\left(rX_t\right)\right] \\ &= G(r)^t. \end{aligned}$$

The moment-generating function of the sum S_t is therefore equal to the tth power of the moment-generating function of X_i. □

As an example of how to apply this idea, suppose that X is a coin flip with parameter p. The theorem then tells us that

$$\Pr\left\{S_t \geq qt\right\} \leq \left(\frac{pe^r + 1 - p}{e^{rq}}\right)^t.$$

By inserting different values of r into this expression, we can thus derive different exponentially decreasing bounds on the probability that more than qt of the t coin flips came up heads. In the following, we will use this idea to prove a number of important theorems.

4 The Chernoff Bound

We first apply the exponentiation trick to a specific case where we know the distribution of each term of the sum.

Theorem 7. *(The Chernoff bound) If $S_t = X_1 + X_2 + \cdots + X_t$ is a sum of independent coin flips with $\Pr(X = 1) = p$ and $\Pr(X = 1) = 1 - p$, then*

$$\Pr\{S_t \geq qt\} \leq \left(\left(\frac{p}{q}\right)^q \left(\frac{1-p}{1-q}\right)^{1-q}\right)^t.$$

Proof. We know from the previous theorem that

$$\Pr\{S_t \geq qt\} \leq \left(\frac{pe^r + 1 - p}{e^{rq}}\right)^t.$$

It only remains for us to pick a good value of r. Since we want the smallest bound, we find the r for which

$$\frac{\partial}{\partial r}\left(\frac{pe^r + 1 - p}{e^{rq}}\right) = 0.$$

The solution to this problem is

$$r^* = \ln\frac{1-p}{1-q} - \ln\frac{p}{q}.$$

Inserting this into the numerator $pe^r + (1 - p)$, we get

$$p\left(\frac{1-p}{1-q}\right)\left(\frac{q}{p}\right) + (1-p) = (1-p)\left(\frac{q}{1-q} + 1\right) = \frac{1-p}{1-q}.$$

For the denumerator e^{rq}, we get

$$\left(\frac{q}{p} \cdot \frac{1-p}{1-q}\right)^q.$$

Doing the division, we thus find that the smallest possible bound is

$$\left(\frac{pe^{r^*} + 1 - p}{e^{r^* q}}\right)^t = \left(\frac{1-p}{1-q}\right)\left(\frac{q}{p} \cdot \frac{1-p}{1-q}\right)^{-q} = \left(\frac{p}{q}\right)^q \left(\frac{1-p}{1-q}\right)^{1-q}.$$

□

Note that the bound used in this theorem depends both on p and q. We thus select a different r depending on what distribution and what threshold we are asking about.

5 The Hoeffding Bound

We now relax the requirement that the variable X is a coin flip and instead allow it to be any variable restricted to the unit interval.

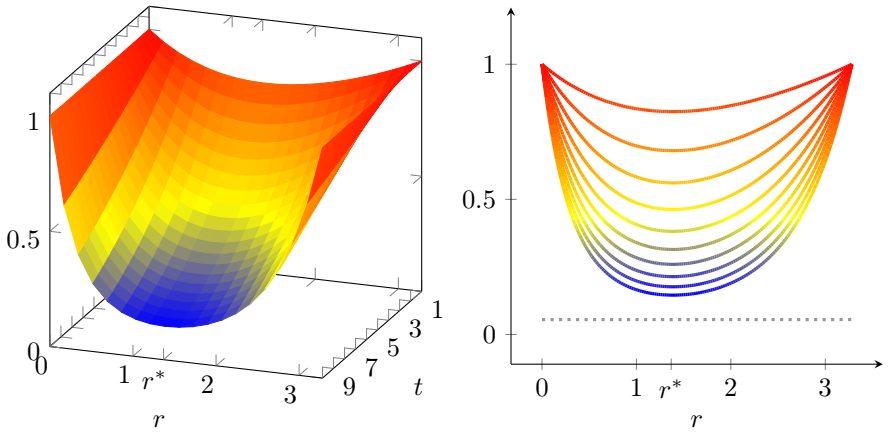
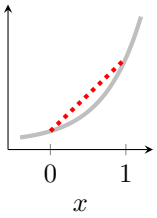


Figure 3: The Chernoff bound decreases exponentially in t for any value of r in a certain positive interval, but there is an optimal value r^* which gives the fastest rate of decrease. The example shown here has $p = 0.5$ and $q = 0.8$, which gives $r^* = 1.386$. The dotted line shows the actual probability that more than $tq = 8$ out of $t = 10$ coin flips come up heads.

Theorem 8. *(Hoeffding's inequality) If $S_t = X_1 + X_2 + \cdots + X_t$ is a sum of independent and identically distributed variables with an expected value of $E[X] = p$, and if $\Pr(0 \leq X \leq 1) = 1$, then*

$$\Pr\{S_t \geq qt\} \leq \left(\left(\frac{p}{q}\right)^q \left(\frac{1-p}{1-q}\right)^{1-q}\right)^t.$$

Proof. In order to prove this theorem, we first need to establish a bound on the moment-generating function $E[\exp(rX)]$. We do so by linear approximation.



The exponential function $\exp(rx)$ is a convex function; that is, any line segment connecting two points on its graph lie completely below the graph. Hence, since the values of X are bounded by 0 and 1, the values of $\exp(rX)$ are therefore bounded by the line connecting the points $(0, 1)$ and $(0, e^r)$.

This line has a slope of $e^r - 1$ and an intercept of 1, so its formula is

$$xe^r - x + 1.$$

We thus have $\exp(rx) \leq 1 + x\exp(r) - x$ for any x in the unit interval. The moment-generating function of X therefore satisfies

$$E\left[e^{rX}\right] \leq E[X]e^r - E[X] + 1 = pe^r - p + 1.$$

In order to select the best bound on S_t, we therefore have to solve

$$\frac{\partial}{\partial r}\left(\frac{pe^r + 1 - p}{e^{rq}}\right) = 0$$

However, this is identical to the minimization problem in the previous theorem, so we get the same bound. □

We will now further relax the restrictions on X, allowing it to be any bounded variable.

Corollary 9. *If $S_t = X_1 + X_2 + \cdots + X_t$ is a sum of independent and identically distributed variables with $E[X] = p$ and bounded by $\Pr(a \leq X \leq b) = 1$, then*

$$\Pr\{S_t \geq qt\} \leq \left(\left(\frac{p-a}{q-a} \right)^{q-a} \left(\frac{b-p}{b-q} \right)^{b-q} \right)^{t/(b-a)}.$$

Proof. We can use the previous theorem on the random variable

$$X' = \frac{X-a}{b-a}$$

and the threshold $q' = q/(b-a)$, since is restricted to the unit interval with probability 1, and since $X' \geq q'$ if and only if $X \geq q$. □

Lastly, we note that we can get the reverse the bound for free:

Corollary 10. *Under the same assumptions as in the previous corollary,*

$$\Pr\{S_t \leq qt\} \leq \left(\left(\frac{p-a}{q-a} \right)^{q-a} \left(\frac{b-p}{b-q} \right)^{b-q} \right)^{t/(b-a)}.$$

Proof. The random variable $X' = a + b - X$ satisfies $\Pr(a \leq X' \leq b) = 1$, and $X' \leq q'$ if and only if $X \geq a + b - q'$. But

$$\left(\frac{p-a}{(a+b-q)-a} \right)^{(a+b-q)-a} = -\left(\frac{p-a}{q-b} \right)^{q-b}$$

and

$$\left(\frac{b-p}{b-(a+b-q)} \right)^{b-(a+b-q)} = -\left(\frac{b-p}{q-a} \right)^{q-a}.$$

Hence, the Hoeffding bound on the probability that $X' \leq q'$ is the same as the Hoeffding bound on the probability that $X \geq q$. The bounds for the corresponding sums are consequently also the same. □