Topic 2: Probability

2.1 Probability spaces and random variables

Let Ω be a sample space, and P be a probability distribution over Ω—together, they make up a probability space \((Ω, P)\). An event \(A\) is a subset of Ω, and the probability of \(A\) is \(P(A) = \sum_{ω ∈ A} P(ω)\).

A random variable defined on \((Ω, P)\) is a real-valued function \(X: Ω → R\). We’ll also use the notation \(X \sim P\) to declare the random variable; often, we’ll even leave the probability space implicit.

The expected value (a.k.a. expectation, mean) of \(X\) is the average value that \(X\) takes on:

\[
E(X) := \sum_{ω ∈ Ω} X(ω) \cdot P(ω).
\]

We’ll often use the more intuitive notation

\[
E(X) = \sum_{x} x \cdot P(X = x),
\]
where the summation is over the range of \(X\). Here, \(P(X = x) = P(\{ω ∈ Ω : X(ω) = x\})\).

The notation above implicitly assumes that range of \(X\) is a discrete set that we can enumerate. But we’ll often encounter a continuous random variable \(X\) whose range is a continuous space, like the real line \(R\) or some subset thereof. In the cases we’ll consider, the random variable \(X\) will have a probability density function \(p: R → R_+\), a non-negative real-valued function on \(R\) satisfying

\[
\int_R p(x) \, dx = 1,
\]
where for any interval \(I ⊆ R\),

\[
P(X ∈ I) = \int_I p(x) \, dx.
\]

The expected value of \(X\) is given by

\[
E(X) = \int_R x \cdot p(x) \, dx.
\]

When we have several continuous random variables \(X_1, X_2, \ldots, X_d\), they will have a joint probability density function \(p: R^d → R_+\) that satisfies

\[
\int_{R^d} p(x) \, dx = \int_{R^d} p(x_1, x_2, \ldots, x_d) \, dx_1 \, dx_2 \cdots dx_d = 1.
\]

Each \(X_i\) has a marginal probability density function \(p_i: R → R\):

\[
p_i(x_i) := \int_{R^d} p(x_1, x_2, \ldots, x_d) \, dx_1 \cdots dx_{i-1} \, dx_{i+1} \cdots dx_d,
\]
a probability density function in its own right. We’ll collect these random variables into a vector \(X = (X_1, X_2, \ldots, X_d)\). The expected value of \(X\) is a vector \(E(X) ∈ R^d\) whose \(i\)-th entry is

\[
E(X_i) = \int_R x_i \cdot p_i(x_i) \, dx_i.
\]

The random variables \(X_1, X_2, \ldots, X_d\) are independent if \(p(x_1, x_2, \ldots, x_d) = p_1(x_1)p_2(x_2)\cdots p_d(x_d)\) for all \((x_1, x_2, \ldots, x_d) ∈ R^d\).
2.2 Linearity of expectation

Suppose \( X \) and \( Y \) are random variables. Then for any scalar \( a \geq 0 \),
\[
\mathbb{E}(aX + Y) = a \cdot \mathbb{E}(X) + \mathbb{E}(Y),
\]
regardless of whether \( X \) and \( Y \) are independent or not. This fact, called linearity of expectation, is very powerful. We are likely to encounter collections of random variables \( X_1, X_2, \ldots, X_d \) with all sorts of dependencies between them. In any of these cases, the expected value of any linear combination of random variables is the corresponding linear combination of their expected values:
\[
a_1X_1 + a_2X_2 + \cdots + a_dX_d = a_1\mathbb{E}(X_1) + a_2\mathbb{E}(X_2) + \cdots + a_d\mathbb{E}(X_d).
\]

Example: unit vectors

A unit vector in \( \mathbb{R}^d \) is a vector \( \mathbf{x} \in \mathbb{R}^d \) of unit length \( \|\mathbf{x}\|_2 = 1 \). The unit sphere
\[
S^{d-1} := \{ \mathbf{x} \in \mathbb{R}^d : \|\mathbf{x}\|_2 = 1 \}
\]
is the set of all unit vectors. (It is the \((d-1)\)-dimensional surface of the unit ball \( B^d \).)

Consider a random vector \( \mathbf{X} \) whose probability density function is the uniform density \( \sigma \) on \( S^{d-1} \), i.e., the density function that assigns equal value to each point in \( S^{d-1} \). We won’t bother writing down the formula for \( \sigma \)—it is a bit cumbersome—but we can use linearity of expectation (and some symmetry considerations) to reason about it.

What is the expected value of, say, \( X_1 \)? For any \( \mathbf{x} = (x_1, x_2, \ldots, x_d) \in S^{d-1} \), the point \((-x_1, x_2, \ldots, x_d)\) is also in \( S^{d-1} \). Hence, for any \( \mathbf{x} \in S^{d-1} \),
\[
\sigma(x_1, x_2, \ldots, x_d) = \sigma(-x_1, x_2, \ldots, x_d),
\]
which implies that \( \mathbb{E}(X_1) = 0 \). For the same reason, \( \mathbb{E}(X_1X_2) = \mathbb{E}(X_1X_2X_3) = \cdots = 0 \). And of course nothing is special about the first coordinate, or first two coordinates, etc.; the same holds for any distinct \( i_1, i_2, \ldots \in \{1, 2, \ldots, d\} \).

We next determine \( \mathbb{E}(X_i^2) \). By linearity of expectation,
\[
1 = \mathbb{E}\|\mathbf{X}\|_2^2 = \mathbb{E}\left( \sum_{i=1}^{d} X_i^2 \right) = \sum_{i=1}^{d} \mathbb{E}(X_i^2).
\]
And by symmetry, it follows that \( \mathbb{E}(X_i^2) = 1/d \) for each \( i \in \{1, 2, \ldots, d\} \).

Not only is there nothing special about any particular coordinate under \( \sigma \), there is nothing special about any particular direction. For any unit vector \( \mathbf{u} \in S^{d-1} \), we have \( \mathbb{E}(\langle \mathbf{u}, \mathbf{X} \rangle^2) = 1/d \).

This can be verified algebraically:
\[
\mathbb{E}(\langle \mathbf{u}, \mathbf{X} \rangle^2) = \mathbb{E}\left( \sum_{i=1}^{d} u_iX_i \right)^2 = \mathbb{E}\left( \sum_{i=1}^{d} u_i^2X_i^2 + \sum_{i \neq j} u_iu_jX_iX_j \right)
\]
\[
= \sum_{i=1}^{d} u_i^2 \mathbb{E}(X_i^2) + \sum_{i \neq j} u_iu_j \mathbb{E}(X_iX_j) = \frac{1}{d} \sum_{i=1}^{d} u_i^2 + 0 = \frac{1}{d}.
\]
2.3 Variance

The mean of $X$ tells us the what value $X$ takes on average. But we are also interested in how far $X$ is from its mean, on average. This is captured by the variance of $X$,

$$\text{var}(X) := \mathbb{E}(X - \mathbb{E}(X))^2.$$ 

We look at $\mathbb{E}(X - \mathbb{E}(X))^2$ rather than, say, $\mathbb{E}|X - \mathbb{E}(X)|$ simply for mathematical convenience. Unfortunately, the “units” of $X$ and var($X$) are not the same: if $X$ is measured in “meters”, then var($X$) is measured in “square meters”. Therefore, we will often also look at the square-root of variance $\sqrt{\text{var}(X)}$, which is called the standard deviation.

**Jensen’s inequality**

Another (easily checked) formula for variance is $\text{var}(X) = \mathbb{E}(X^2) - (\mathbb{E}(X))^2$. And since the variance of a random variable is always non-negative, we have the inequality

$$(\mathbb{E}(X))^2 \leq \mathbb{E}(X^2).$$

This inequality is actually a special case of a more general inequality, called Jensen’s inequality, which states that for any convex function $f : \mathbb{R} \rightarrow \mathbb{R}$, we have $f(\mathbb{E}(X)) \leq \mathbb{E}(f(X))$. A function $f$ is convex if $f(px + (1-p)y) \leq pf(x) + (1-p)f(y)$ for any $p \in [0,1]$ and $x, y \in \mathbb{R}$. The above inequality is the special case where $f(x) = x^2$, which is convex.

Applying the inequality to the random variable $|X - \mathbb{E}(X)|$, we obtain the convenient fact

$$\mathbb{E}|X - \mathbb{E}(X)| \leq \sqrt{\mathbb{E}(X - \mathbb{E}(X))^2} = \sqrt{\text{var}(X)}.$$ 

**Variance of linear combinations**

If $X$ and $Y$ are random variables, then for any scalar $a$,

$$\text{var}(aX + Y) = a^2 \text{var}(X) + \text{var}(Y) + 2a \text{cov}(X,Y)$$

where

$$\text{cov}(X,Y) := \mathbb{E}((X - \mathbb{E}(X))(Y - \mathbb{E}(Y)))$$

is the covariance between $X$ and $Y$, which need not be zero. However, if $X$ and $Y$ are independent, then

$$\text{cov}(X,Y) = \mathbb{E}((X - \mathbb{E}(X))(Y - \mathbb{E}(Y))) = \mathbb{E}(X - \mathbb{E}(X))\mathbb{E}(Y - \mathbb{E}(Y)) = 0,$$

upon which we have $\text{var}(aX + Y) = a^2 \text{var}(X) + \text{var}(Y)$. More generally, if $X_1, X_2, \ldots, X_d$ are independent random variables, then

$$\text{var}(a_1X_1 + a_2X_2 + \cdots + a_dX_d) = a_1^2 \text{var}(X_1) + a_2^2 \text{var}(X_2) + \cdots + a_d^2 \text{var}(X_d).$$

(Actually, this holds even if we just have pairwise independence—i.e., any two $X_i$ and $X_j$ are independent.)

**Example: symmetric random walk**

A symmetric random walk on $\mathbb{Z}$ is a stochastic process $(S_t)_{t \in \mathbb{Z}_+}$. Here, $S_0 := 0$, and for each time $t \geq 1$, $S_t := S_{t-1} + X_t$, where $\mathbb{P}(X_t = -1) = \mathbb{P}(X_t = 1) = 1/2$. Clearly $S_n = \sum_{i=1}^n X_i$. Each $X_t$ has mean $\mathbb{E}(X_t) = 0$, and variance $\text{var}(X_t) = 1$. By linearity of expectation, $\mathbb{E}(S_n) = \sum_{t=1}^n \mathbb{E}(X_t) = 0$. We assume $\{X_t : t \in \mathbb{N}\}$ are independent, so we have that $\text{var}(S_n) = \sum_{t=1}^n \text{var}(X_t) = n$, which in turn implies

$$\mathbb{E}|S_n| \leq \sqrt{\text{var}(S_n)} = \sqrt{n}.$$
2.4 Tail bounds from moments

In a symmetric random walk, the position after \(n\) steps is, in expectation, within \(\sqrt{n}\) of the origin. Can something similar be said about any particular realization of the random walk? To answer this question, we appeal to Markov’s inequality.

**Theorem 2.1** (Markov’s inequality). For any \(t \geq 0\), \(\mathbb{P}(|X| \geq t) \leq \mathbb{E}|X|/t\).

**Proof.** For any \(t \geq 0\),

\[
|X| \geq \mathbb{1}\{|X| \geq t\} \cdot t.
\]

(Here, \(\mathbb{1}\{\phi\} := 1\) if \(\phi\) is true, and \(\mathbb{1}\{\phi\} := 0\) otherwise.) Take expectations of both sides to obtain

\[
\mathbb{E}|X| \geq \mathbb{E}(\mathbb{1}\{|X| \geq t\}) \cdot t.
\]

Now interpret the right-hand side as \(\mathbb{P}(|X| \geq t) \cdot t\) and rearrange.

Suppose, for \(t > 0\) and \(\delta \in (0, 1)\), that \(\mathbb{P}(X = 0) = 1 - \delta\) and \(\mathbb{P}(X = t) = \delta\). Then \(\mathbb{P}(X \geq t) = \delta = \mathbb{E}(X)/t\). So Markov’s inequality is tight at least in some cases.

For the symmetric random walk, let \(t = c\sqrt{n}\) for some \(c > 0\). Applying Markov’s inequality, we obtain the inequality

\[
\mathbb{P}(|S_n| \geq c\sqrt{n}) \leq \frac{\mathbb{E}|S_n|}{c\sqrt{n}} \leq \frac{1}{c}.
\]

This means that for at least a \(1 - 1/c\) fraction of the realizations, the position after \(n\) steps is within \(c\sqrt{n}\) of the origin.

This tail bound can sometimes be improved by a more effective use of Markov’s inequality. The essence of this is idea is captured in Chebyshev’s inequality.

**Theorem 2.2** (Chebyshev’s inequality). For any \(t \geq 0\), \(\mathbb{P}(|X - \mathbb{E}(X)| \geq t) \leq \text{var}(X)/t^2\).

**Proof.** Apply Markov’s inequality to the random variable \((X - \mathbb{E}(X))^2\).

Again with the symmetric random walk, applying Chebyshev’s inequality gives

\[
\mathbb{P}(|S_n| \geq c\sqrt{n}) \leq \frac{\text{var}(S_n)}{c^2n} = \frac{1}{c^2}.
\]

The improvement comes from exploiting the variance of \(|S_n|\), whereas previously we just used its expected value. By exploiting higher-order moments (\(\mathbb{E}(X^3)\), \(\mathbb{E}(X^4)\), etc.), we may obtain further improvements.

**Weak Law of Large Numbers**

Using Chebyshev’s inequality, we can prove a simple version of the *Law of Large Numbers* (LLN). We prove the non-asymptotic form of the “Weak” LLN by applying Chebyshev’s inequality as applied to the average of \(n\) independent and identically distributed (iid) random variables.

**Theorem 2.3** (Weak Law of Large Numbers). Let \(X_1, X_2, \ldots, X_n\) be iid random variables. Let \(\mu_n := (1/n)\sum_{i=1}^n X_i\), \(\mu := \mathbb{E}(X_1)\), and \(\sigma^2 := \text{var}(X_1)\). For any \(\varepsilon > 0\), \(\mathbb{P}(|\mu_n - \mu| > \varepsilon) \leq \sigma^2/(n\varepsilon)\).

**Proof.** By linearity of expectation, \(\mathbb{E}(\mu_n) = \mu\). Therefore, applying Chebyshev’s inequality and using the fact that \(X_1, X_2, \ldots, X_n\) are iid, \(\mathbb{P}(|\mu_n - \mu| > \varepsilon) \leq \text{var}(\mu_n)/\varepsilon = \sigma^2/(n\varepsilon)\).
2.5 Chernoff bounds

There is a slick approach for using all moments simultaneously to obtain a tail bound. The method is called the Chernoff bounding method, or the Cramér-Chernoff method, although it was used much earlier by Bernstein. In computer science, the resulting bounds are often called Chernoff bounds.

Moment generating functions

A useful property of a random variable is its moment generating function (mgf) \( M_X \): for \( \lambda \in \mathbb{R} \),

\[
M_X(\lambda) := \mathbb{E}(\exp(\lambda X)) = 1 + \lambda \mathbb{E}(X) + \frac{\lambda^2}{2!} \mathbb{E}(X^2) + \frac{\lambda^3}{3!} \mathbb{E}(X^3) + \cdots.
\]

Note that it is possible that the expectation is infinite (i.e., does not exist) for some values of \( \lambda \neq 0 \). However, if \( M_X(-\lambda_1) \) and \( M_X(\lambda_2) \) exist for some \( \lambda_1, \lambda_2 > 0 \), then the following statements hold.

- \( M_X(\lambda) \) exists for all values \( \lambda \in [-\lambda_1, \lambda_2] \) and is infinitely differentiable on \(( -\lambda_1, \lambda_2 )\).
- \( \mathbb{E}(X^p) \) is finite for all \( p \in \mathbb{N} \).
- The distribution of \( X \) can be determined from \( M_X \), and if \( M_X(\lambda) = M_Y(\lambda) \) for all \( \lambda \in [-\lambda_1, \lambda_2] \), then \( X \) and \( Y \) have the same distribution.

We will often work with the natural logarithm of the mgf, which is called the cumulative generating function or logarithmic moment generating function (log mgf)

\[
\psi_X(\lambda) := \ln M_X(\lambda).
\]

It is clear that \( \psi_X(\lambda) \) exists if and only if \( M_X(\lambda) \) exists. Some simple properties of the log mgf are as follows.

- \( \psi_X(0) = 0 \) and \( \psi_{aX+b}(\lambda) = \psi_X(a\lambda) + b\lambda \).
- If \( X_1, X_2, \ldots, X_n \) are independent, and \( \psi_{X_i}(\lambda) \) exists for each \( i \), then \( \psi_{\sum_{i=1}^n X_i}(\lambda) = \sum_{i=1}^n \psi_{X_i}(\lambda) \).
- If \( \psi_X \) exists on an interval \((-\lambda_1, \lambda_2)\) for some \( \lambda_1, \lambda_2 > 0 \), then it is infinitely differentiable on the same interval.

Examples

We write \( X \sim \text{Poi}(\mu) \) to mean that the random variable \( X \) has a Poisson distribution with rate \( \mu > 0 \). The distribution of \( X \) is given by

\[
\mathbb{P}(X = k) = \frac{\exp(-\mu)\mu^k}{k!}, \quad k \in \mathbb{Z}_+.
\]

It can be checked that \( \mathbb{E}(X) = \mu \) and \( \psi_X(\lambda) = \mu(\exp(\lambda) - 1) \).

We write \( X \sim \text{N}(\mu, \sigma^2) \) to mean that the random variable \( X \) has a Gaussian (Normal) distribution with mean \( \mu \) and variance \( \sigma^2 \). The density of \( X \) is

\[
p(x) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right), \quad x \in \mathbb{R}.
\]

Furthermore, \( \psi_X(\lambda) = \mu \lambda + \sigma^2 \lambda^2/2 \).

Finally, if \( Z_1, Z_2, \ldots, Z_k \) are iid \( \text{N}(0, 1) \) random variables, then \( X := \sum_{i=1}^k Z_i^2 \) has a \( \chi^2 \) distribution with \( k \) degrees of freedom, written \( X \sim \chi^2(k) \). It is known that \( \mathbb{E}(X) = k \) and \( \psi_X(\lambda) = -(k/2)\ln(1 - 2\lambda) \) for \( \lambda < 1/2 \) (and not defined for \( \lambda \geq 1/2 \)). It can also be checked that \( X \) is independent of the uniform random unit vector \( U := (Z_1, Z_2, \ldots, Z_k)/\sqrt{k} \).
Cramér-Chernoff inequality

The Cramér-Chernoff inequality is based on applying Markov’s inequality to an exponential function of a random variable.

**Theorem 2.4** (Cramér-Chernoff inequality). For any $t \in \mathbb{R}$,

$$
P(X \geq t) \leq \exp\left(-\max_{\lambda \geq 0}\{t\lambda - \psi_X(\lambda)\}\right).
$$

If $t \geq \mathbb{E}(X)$, then $P(X \geq t) \leq \exp(-\psi^*_X(t))$, where

$$
\psi^*_X(t) := \max_{\lambda \in \mathbb{R}}\{t\lambda - \psi_X(\lambda)\}.
$$

Therefore, for any $t \geq 0$,

$$
P(X \geq \mathbb{E}(X) + t) \leq \exp\left(-\max_{\lambda \geq 0}\{t\lambda - \psi_X(\mathbb{E}(X))(\lambda)\}\right) = \exp\left(-\psi^*_X(-\mathbb{E}(X))(t)\right).
$$

**Proof.** For any $t \in \mathbb{R}$ and $\lambda \geq 0$,

$$
P(X \geq t) \leq P(\lambda X \geq \lambda t) = P(\exp(\lambda X) \geq \exp(\lambda t)).
$$

Applying Markov’s inequality to the right-hand side, we obtain

$$
P(X \geq t) \leq \frac{\mathbb{E}(\exp(\lambda X))}{\exp(\lambda t)} = \exp\left(-(t\lambda - \psi_X(\lambda))\right).
$$

Since this inequality holds for all $\lambda \geq 0$, we have

$$
P(X \geq t) \leq \exp\left(-\max_{\lambda \geq 0}\{t\lambda - \psi_X(\lambda)\}\right).
$$

It remains to show that the maximum can be extended from $\lambda \geq 0$ to $\lambda \in \mathbb{R}$. The maximum is always non-negative, since the choice $\lambda = 0$ yields the value 0. Therefore, it suffices to show that choosing $\lambda < 0$ cannot yield a positive value.

For any $\lambda \in \mathbb{R}$, the function $f(x) = \exp(\lambda x)$ is convex, and hence

$$
\exp(\lambda \mathbb{E}(X)) \leq \mathbb{E}(\exp(\lambda X))
$$

by Jensen’s inequality. Taking logarithm of both sides implies that

$$
\lambda \mathbb{E}(X) \leq \psi_X(\lambda)
$$

for all $\lambda \in \mathbb{R}$. Thus, for any $t \geq \mathbb{E}(X)$ and $\lambda < 0$,

$$
t\lambda - \psi_X(\lambda) \leq \lambda \mathbb{E}(X) - \psi_X(\lambda) \leq 0.
$$

As stated, the Cramér-Chernoff inequality only provides a bound for the “right tail”. But by applying the inequality to $\mathbb{E}(X) - X$, we may also obtain a bound for the “left tail”.

**Corollary 2.1.** For any $t \geq 0$, $P(X \leq \mathbb{E}(X) - t) \leq \exp(-\psi^*_X(-\mathbb{E}(X))(t))$.

**Proof.** It is easy to verify from definitions that for any random variable $Z$,

$$
\psi^*_Z(t) = \psi^*_Z(-t).
$$

Therefore, using $Z := X - \mathbb{E}(X)$, we obtain via the Cramér-Chernoff inequality

$$
P(X \leq \mathbb{E}(X) - t) = P(-Z \geq t) \leq \exp(-\psi^*_Z(t)) = \exp(-\psi^*_Z(-t)) = \exp(-\psi^*_X(-\mathbb{E}(X))(-t)).$$
Example: Poisson
If $X \sim \text{Poi}(\mu)$, then $\mathbb{E}(X) = \mu$, and

$$\psi_{X - \mu}(\lambda) = \mu (\exp(\lambda) - \lambda - 1).$$

Therefore

$$\psi^*_{X - \mu}(t) = \mu \left( (1 + t/\mu) \ln(1 + t/\mu) - t/\mu \right),$$

where the maximum in the definition of $\psi^*_{X - \mu}(t)$ is achieved at $\lambda = \ln(1 + t/\mu)$. Applying the Cramér-Chernoff inequality, we have for any $t \geq 0$,

$$\mathbb{P}(X \geq \mu + t) \leq \exp(-\mu \cdot h(t/\mu)),$$

where $h(x) := (1 + x) \ln(1 + x) - x$. We can similarly obtain

$$\mathbb{P}(X \leq \mu - t) \leq \exp(-\mu \cdot h(-t/\mu)).$$

Example: Gaussian
If $X \sim \text{N}(\mu, \sigma^2)$, then $\mathbb{E}(X) = \mu$, and $\psi_{X - \mu}(\lambda) = \sigma^2 \lambda^2/2$. Therefore $\psi^*_{X - \mu}(t) = t^2/(2\sigma^2)$, where the maximum in the definition of $\psi^*_{X - \mu}(t)$ is achieved at $\lambda = t/\sigma^2$. So for any $t \geq 0$,

$$\mathbb{P}(X \geq \mu + t) \leq \exp\left(-\frac{t^2}{2\sigma^2}\right).$$

And by symmetry, for any $t \geq 0$,

$$\mathbb{P}(X \leq \mu - t) \leq \exp\left(-\frac{t^2}{2\sigma^2}\right).$$

Example: $\chi^2$
If $X \sim \chi^2(k)$, then $\mathbb{E}(X) = k$, and $\psi_{X - k}(\lambda) = -((k/2) \ln(1 - 2\lambda) - k\lambda$ for $\lambda < 1/2$. Therefore,

$$\psi^*_{X - k}(t) = \frac{k}{2} \left( \frac{t}{k} - \ln \left( 1 + \frac{t}{k} \right) \right),$$

where the maximum in the definition of $\psi^*_{X - k}(t)$ is achieved at $\lambda = (1/2)(1 - k/(k + t))$. Thus, we have for any $t \geq 0$,

$$\mathbb{P}(X \geq k + t) \leq \exp \left(-\frac{k}{2} \left( \frac{t}{k} - \ln \left( 1 + \frac{t}{k} \right) \right) \right) = \exp \left(-\frac{k}{2} \left( \frac{\varepsilon^2}{3} + \frac{\varepsilon^4}{5} + \cdots \right) \right)$$

where $\varepsilon = t/k > 0$. We similarly obtain for any $t \geq 0$ (but really only $0 \leq t < k$ makes sense),

$$\mathbb{P}(X \leq k - t) \leq \exp \left(-\frac{k}{2} \left( -\ln \left( 1 - \frac{t}{k} \right) - \frac{t}{k} \right) \right) = \exp \left(-\frac{k}{2} \left( \frac{\varepsilon^2}{3} + \frac{\varepsilon^4}{4} + \frac{\varepsilon^6}{5} + \cdots \right) \right)$$

(the latter expression holds for $\varepsilon = t/k \in (0, 1)$).
Hoeffding’s inequality

In many cases, the specific form of the log mgf is unknown, but other aspects of the random variable are known. The following lemma provides a bound on the log mgf for bounded random variables.

**Lemma 2.1.** Let $X$ be a $[0, 1]$-valued random variable with $E(X) = \mu$. Let $Y$ be a $\{0, 1\}$-valued random variable with $E(Y) = \mu$. For any $\lambda \in \mathbb{R}$,

$$
\psi_{X-\mu}(\lambda) \leq \psi_{Y-\mu}(\lambda) \leq \frac{\lambda^2}{8}.
$$

**Proof.** We first show that $\psi_X(\lambda) \leq \psi_Y(\lambda)$ for $\lambda \geq 0$. We have that $M_X(0) = M_Y(0)$. Now observe that for all $\lambda \geq 0$,

$$
M_X(\lambda) = \mathbb{E}[X \exp(\lambda X)] \leq \mathbb{E}[X \exp(\lambda)] = \mu \exp(\lambda) = M_Y(\lambda).
$$

This implies $M_X(\lambda) \leq M_Y(\lambda)$ for all $\lambda \geq 0$, and thus $\psi_X(\lambda) \leq \psi_Y(\lambda)$.

Now we show that $\psi_X(-\lambda) \leq \psi_Y(-\lambda)$ for all $\lambda \geq 0$. Define $\bar{X} := 1 - X$ and $\bar{\mu} := \mathbb{E}(\bar{X}) = 1 - \mu$. Also define $\bar{Y} := 1 - Y$ so $\mathbb{E}(\bar{Y}) = \bar{\mu}$ as well. By the previous inequality, for all $\lambda \geq 0$,

$$
\lambda + \psi_{-X}(\lambda) = \psi_{1-X}(\lambda) = \psi_X(\lambda) \leq \psi_Y(\lambda) = \psi_{1-Y}(\lambda) = \lambda + \psi_{-Y}(\lambda).
$$

This implies $\psi_X(-\lambda) = \psi_{-X}(\lambda) \leq \psi_{-Y}(\lambda) = \psi_Y(-\lambda)$ for all $\lambda \geq 0$, as required. Therefore, we conclude $\psi_X(\lambda) \leq \psi_Y(\lambda)$ for all $\lambda \in \mathbb{R}$.

Fix any $\lambda \in \mathbb{R}$. Taylor’s theorem implies the existence of $\eta \in \mathbb{R}$ between 0 and $\lambda$ such that

$$
\psi_Y(\lambda) = \psi_Y(0) + \psi_Y'(0)\lambda + \frac{1}{2}\psi_Y''(\eta)\lambda^2.
$$

It can be checked that for all $\lambda \in \mathbb{R}$,

$$
\psi_Y'(\lambda) = \frac{\mu \exp(\lambda)}{1 - \mu + \mu \exp(\lambda)} \in [0, 1], \
\psi_Y''(\lambda) = \psi_Y'(\lambda)(1 - \psi_Y'(\lambda)) \leq \frac{1}{4}.
$$

Since $\psi_Y(0) = 0$ and $\psi_Y'(0) = \mu$, it follows that $\psi_Y(\lambda) \leq \mu \lambda + \lambda^2/8$. We conclude that for all $\lambda \in \mathbb{R}$, $\psi_{X-\mu}(\lambda) \leq \psi_{Y-\mu}(\lambda) \leq \lambda^2/8$. \hfill $\square$

**Theorem 2.5 (Hoeffding’s inequality).** Let $X_1, X_2, \ldots, X_n$ be independent random variables, where $X_i \in [0, \sigma_i]$ for some $\sigma_i > 0$. Let $S := \sum_{i=1}^{n} X_i$, $m := \sum_{i=1}^{n} \mathbb{E}(X_i)$, and $\sigma^2 := \sum_{i=1}^{n} \sigma_i^2$. For any $t \geq 0$, $P(S \geq m + t) \leq \exp(-2t^2/\sigma^2)$ and $P(S \leq m - t) \leq \exp(-2t^2/\sigma^2)$.

**Proof.** Since $X_1, X_2, \ldots, X_n$ are independent,

$$
\psi_{S-m}(\lambda) = \sum_{i=1}^{n} \psi_{X_i-\mathbb{E}(X_i)}(\lambda) = \sum_{i=1}^{n} \psi_{(X_i-\mathbb{E}(X_i))/\sigma_i}(\sigma_i \lambda) \leq \frac{\sigma^2 \lambda^2}{8},
$$

where the last inequality follows from Lemma 2.1. Therefore $\psi_{S-m}(t) \geq 2t^2/\sigma^2$. Similarly, $\psi_{m-S}(\lambda) \leq \sigma^2 \lambda^2/8$ and $\psi_{m-S}(t) \geq 2t^2/\sigma^2$. The conclusions now follow from the Cramér-Chernoff inequality. \hfill $\square$

Returning to the symmetric random walk example, applying Hoeffding’s inequality gives

$$
P(|S_n| \geq c\sqrt{n}) = P(S_n \geq c\sqrt{n}) + P(S_n \leq -c\sqrt{n}) \leq 2 \exp(-c^2/2).
$$

Compared to the $1/c$ and $1/c^2$ bounds from Markov’s and Chebyshev’s inequalities, the $2 \exp(-c^2/2)$ tail bound delivered by Hoeffding’s inequality is far superior (except when $c$ is very small).
Bennett’s and Bernstein’s inequalities

By leveraging more information about the random variables, we can derive different bounds on the log mgf. This can lead to improved tail bounds compared to those that do not use additional information in the log mgf bound.

Lemma 2.2. Let $X$ be a random variable satisfying $\mathbb{E}(X) = \mu$ and $X - \mu \leq 1$ for some $\mu \in \mathbb{R}$. For any $\lambda \geq 0$,

$$
\psi_{X-\mu}(\lambda) \leq (\exp(\lambda) - \lambda - 1) \var(X).
$$

Proof. Without loss of generality, we assume $\mu = 0$. Define the increasing function $g: \mathbb{R} \to \mathbb{R}$ with $g(x) := (\exp(x) - x - 1)/x^2$ for $x \neq 0$ and $g(0) := 1/2$. Because $\ln(1 + x) \leq x$ for all $x > -1$,

$$
\psi_X(\lambda) \leq \mathbb{E}(\exp(\lambda X) - 1) = \mathbb{E}(\exp(\lambda X) - \lambda X - 1) = \mathbb{E}(g(\lambda X)(\lambda X)^2).
$$

For any $\lambda \geq 0$, we have $g(\lambda X) \leq g(\lambda)$ since $\lambda X \leq \lambda$. In this case, it follows that $\psi_X(\lambda) \leq (\exp(\lambda) - \lambda - 1) \mathbb{E}(X^2)$.

Theorem 2.6 (Bennett’s inequality). Let $X_1, X_2, \ldots, X_n$ be independent random variables satisfying $X_i - \mathbb{E}(X_i) \leq b$ for all $i \in [n]$, where $b > 0$. Let $S := \sum_{i=1}^n X_i$, $m := \sum_{i=1}^n \mathbb{E}(X_i)$, and $v := \sum_{i=1}^n \var(X_i)$. For any $t \geq 0$,

$$
\mathbb{P}(S \geq m + t) \leq \exp \left( - \frac{v}{b^2} \cdot h \left( \frac{bt}{v} \right) \right) \leq \exp \left( - \frac{9v}{b^2} \cdot h_1 \left( \frac{bt}{3v} \right) \right) \leq \exp \left( - \frac{t^2}{2(v + bt/3)} \right),
$$

where $h(x) := (1 + x) \ln(1 + x) - x$ and $h_1(x) := 1 + x - \sqrt{1 + 2x}$.

Proof. Since $X_1, X_2, \ldots, X_n$ are independent, for any $\lambda \geq 0$,

$$
\psi_{S-m}(\lambda) = \sum_{i=1}^n \psi_{X_i-\mathbb{E}(X_i)}(\lambda) = \sum_{i=1}^n \psi_{(X_i-\mathbb{E}(X_i))/b\lambda} \leq \exp(h(\lambda)) - b\lambda - 1
$$

for any $\lambda \geq 0$. The first probability inequality $\mathbb{P}(S \geq m + t) \leq \exp(-(v/b^2) \cdot h(bt/v))$ now follows from the Cramér-Chernoff inequality. The second and third inequalities can be checked using calculus:

$$
h(x) \geq 9 \cdot h_1(x/3) \geq \frac{x^2}{2(1 + x/3)}
$$

for all $x \geq 0$.

A few comments on Theorem 2.6 are in order. First, observe that the function $h(x)$ also appears in the right tail bound for a Poisson random variable $X \sim \text{Po}(\mu)$: $\mathbb{P}(X \geq \mu + t) \leq \exp(-\mu \cdot h(t/\mu))$.

Second, the function $h_1$ conveniently has as its inverse the function $h_1^{-1}(y) = \sqrt{2y} + y$. Thus, we can easily find the value $t$ so that the tail bound is at most some $\delta \in (0, 1)$:

$$
\frac{9v}{b^2} \cdot h_1 \left( \frac{bt}{3v} \right) = \ln(1/\delta), \quad \text{which entails} \quad t = \sqrt{2v \ln(1/\delta)} + \frac{b \ln(1/\delta)}{3}.
$$
When the variance term \( v \) is \( o(n) \), Bennett’s inequality may give an improved bound compared to that from Hoeffding’s inequality (Theorem 2.5).

For example, if \( \text{var}(X_i) = O(1/n) \) for each \( i \in [n] \), then \( v = O(1) \), upon which we have with probability at least \( 1 - \delta \), \( S \leq m + O(\log(1/\delta)) \). This consistent with the heuristic Poisson approximation for a Binomial random variable \( S \sim \text{Bin}(n,p) \) where \( p = O(1/n) \). In this case, \( S \) approximately follows \( \text{Poi}(np) \), and indeed, Bennett’s inequality and a Chernoff bound for \( \text{Poi}(np) \) give the same right tail bound of \( \exp\left(-np \cdot (t/(np))\right) \) (since \( \text{var}(S) = np(1-p) \leq np \)).

Finally, we note that the boundedness condition \( X_i - \mathbb{E}(X_i) \leq b \) in Bennett’s inequality can be replaced by a moment condition:

\[
\sum_{i=1}^{n} \mathbb{E}\left[ \max\{X_i - \mathbb{E}(X_i), 0\}^k \right] \leq \frac{k!}{2} \left( \frac{b}{3} \right)^{k-2} \sum_{i=1}^{n} \text{var}(X_i^2) \quad \text{for all integers } k \geq 3.
\]

Under this condition, we have the tail bound

\[
\mathbb{P}(S \geq m + t) \leq \exp\left(- \frac{9v}{b^2} \cdot h_1\left( \frac{bt}{3v} \right) \right).
\]

Both this inequality and the second probability inequality in Theorem 2.6 are called *Bernstein’s inequality*. 