Exponential Tilting, Concentration, and Convex Duality

Thiago Ramos

1 Importance Sampling and Exponential Tilting

Suppose we are interested in estimating a rare probability of the form

$$p = \mathbb{P}(X > 10),$$

where $X \sim \mathcal{N}(0,1)$, with density

$$f(x) = \frac{1}{\sqrt{2\pi}} e^{-x^2/2}.$$

A natural Monte Carlo estimator is

$$\hat{p}_{\mathrm{MC}} = \frac{1}{n} \sum_{i=1}^{n} \mathbf{1}\{X_i > 10\}.$$

However, this approach fails completely when the event of interest is extremely rare.

For instance, in a simulation with $n = 10^7$ samples drawn from $\mathcal{N}(0,1)$, not a single value exceeded 10, resulting in

$$\hat{p}_{\mathrm{MC}} = 0,$$

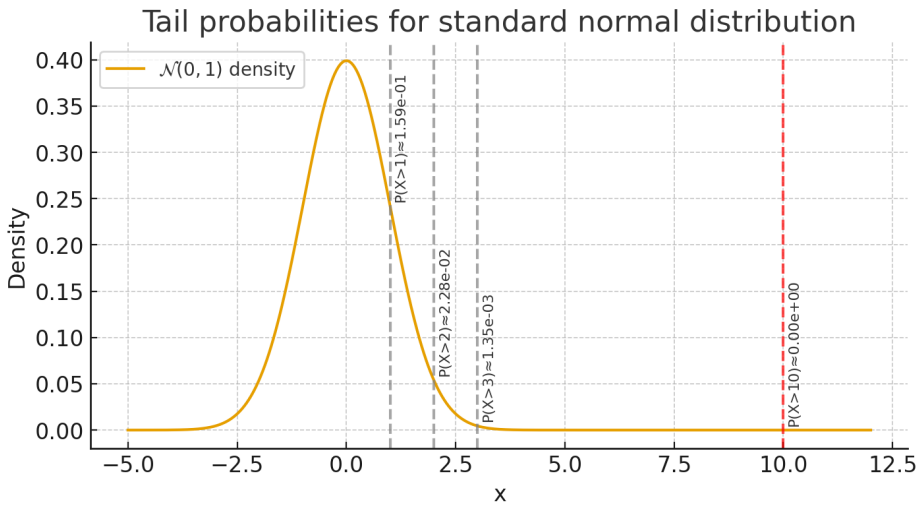
while the true probability is

$$\mathbb{P}(X > 10) = 7.62 \times 10^{-24}.$$

In a simulation with $n = 10^7$ samples drawn from $\mathcal{N}(0,1)$, not a single value exceeded 10, leading to

$$\hat{p}_{\mathrm{MC}} = 0, \qquad p = \mathbb{P}(X > 10) = 7.62 \times 10^{-24}.$$

The estimator collapses simply because the event $\{X > 10\}$ is far outside the region where the standard normal places any noticeable probability mass. To make progress, we must somehow sample more often from the region that actually contributes to the probability we wish to estimate, while keeping the result unbiased. This is precisely the idea behind *importance sampling*.



The key idea behind importance sampling is simple: if the event of interest is extremely rare under the original distribution, we can sample from a different distribution where that event is more likely to occur. We then correct for this change of measure by weighting each sample appropriately.

Let $f(x)$ denote the density of the original distribution of X, and suppose we want to estimate

$$p = \mathbb{E}_f[\,h(X)\,] = \int h(x) f(x)\, dx,$$

where $h(x) = \mathbf{1}\{x > 10\}$. Direct Monte Carlo estimation uses samples $X_i \sim f$ and computes

$$\hat{p}_{\text{MC}} = \frac{1}{n} \sum_{i=1}^{n} h(X_i),$$

which, as we have seen, is ineffective for rare events.

To improve this, we introduce an alternative density $g(x)$ satisfying $g(x) > 0$ whenever $h(x)f(x) > 0$. By multiplying and dividing the integrand by $g(x)$, we obtain

$$p = \int h(x) \frac{f(x)}{g(x)} g(x)\, dx = \mathbb{E}_g\left[h(X) \frac{f(X)}{g(X)} \right].$$

Hence, we can estimate p using samples $X_i \sim g$:

$$\hat{p}_{\text{IS}} = \frac{1}{n} \sum_{i=1}^{n} h(X_i)\, w(X_i), \qquad \text{where } w(x) = \frac{f(x)}{g(x)}.$$

The distribution g is called the *proposal distribution*, and the ratio $w(x)$ is known as the *importance weight*. Intuitively, g should be chosen so that samples in the region where $h(x)$ is large (for instance, $x > 10$) become much more frequent, while the weights $w(x)$ compensate for this oversampling.

We now return to the problem introduced at the beginning: estimating

$$p = \mathbb{P}(X > 10), \quad X \sim \mathcal{N}(0, 1).$$

The previous figure made it clear that this probability is astronomically small, making direct Monte Carlo estimation infeasible. To overcome this limitation, we can apply the idea of importance sampling introduced above.

We choose as proposal distribution

$$g(x) = \mathcal{N}(10, 1),$$

which places most of its probability mass around the region of interest $x > 10$. Under this new distribution, the event becomes common, and the estimator

$$\hat{p}_{\text{IS}} = \frac{1}{n} \sum_{i=1}^{n} \mathbf{1}\{X_i > 10\} \frac{f(X_i)}{g(X_i)}, \quad X_i \sim g,$$

provides a meaningful estimate even for moderate sample sizes.

To assess the effect of this change of measure, we simulate $n = 10^6$ samples from $g = \mathcal{N}(10, 1)$ and compute

$$\hat{p}_{\text{IS}} = \frac{1}{n} \sum_{i=1}^{n} \mathbf{1}\{X_i > 10\} \frac{f(X_i)}{g(X_i)}, \quad X_i \sim g,$$

where f is the density of $\mathcal{N}(0, 1)$. Using this approach, we obtain

$$\hat{p}_{\text{IS}} = 7.67 \times 10^{-24}, \qquad p_{\text{true}} = 7.62 \times 10^{-24}, \qquad \text{SE} \approx 2.6 \times 10^{-26}.$$

The example above shows that the efficiency of importance sampling depends crucially on the choice of the proposal distribution g. If g assigns significant probability mass to the region where $h(x)$ is large, the variance of the estimator can be dramatically reduced.

A natural question arises: how should we choose g systematically? One elegant answer comes from the idea of *exponential tilting*. Instead of choosing g arbitrarily, we construct a whole family of deformed versions of the original density f by exponential reweighting:

$$f_\lambda(x) = \frac{e^{\lambda x} f(x)}{Z(\lambda)}, \qquad Z(\lambda) = \int e^{\lambda x} f(x)\, dx.$$

The normalizing constant $Z(\lambda)$ is the *moment generating function* of X under the original distribution f. Its logarithm,

$$\psi(\lambda) = \log Z(\lambda),$$

is called the *log-partition function*. The parameter λ controls how much probability mass is shifted toward the upper or lower tail of the original distribution: positive values of λ tilt the distribution toward larger values of x, while negative values emphasize smaller ones.

As a concrete example, let us come back to the case $X \sim \mathcal{N}(0, 1)$. Applying the exponential tilting defined above gives

$$f_\lambda(x) = \frac{e^{\lambda x} f(x)}{Z(\lambda)} = \frac{1}{\sqrt{2\pi}\, Z(\lambda)}\, e^{-x^2/2 + \lambda x}.$$

Completing the square in the exponent,

$$-\frac{x^2}{2} + \lambda x = -\frac{1}{2}(x^2 - 2\lambda x) = -\frac{1}{2}(x - \lambda)^2 + \frac{\lambda^2}{2},$$

so that

$$f_\lambda(x) = \frac{1}{\sqrt{2\pi}} e^{-(x-\lambda)^2/2}.$$

The normalizing constant is therefore $Z(\lambda) = e^{\lambda^2/2}$, and we see that

$$f_\lambda(x) = \mathcal{N}(\lambda, 1).$$

Hence, exponential tilting of a standard normal simply shifts its mean by λ, leaving the variance unchanged. In this sense, our earlier choice of $g = \mathcal{N}(10, 1)$ can be interpreted as an exponential tilting of $f = \mathcal{N}(0, 1)$ with parameter $\lambda = 10$.

2 Chernoff Bound

The exponential tilting introduced above can also be understood as a change of measure applied to probabilities themselves. This is, in essence, the same idea we used in the normal example, where we replaced $f = \mathcal{N}(0, 1)$ by its tilted version $f_\lambda = \mathcal{N}(\lambda, 1)$ to make the rare event $\{X > 10\}$ typical under the new distribution.

Let X have density f, and define the tilted density

$$f_\lambda(x) = \frac{e^{\lambda x} f(x)}{Z(\lambda)}, \qquad Z(\lambda) = \mathbb{E}_f[e^{\lambda X}].$$

We can then express any probability as an expectation under this new measure:

$$\mathbb{P}(X \geq a) = \int_{x \geq a} f(x)\, dx = \int_{x \geq a} \frac{f(x)}{f_\lambda(x)} f_\lambda(x)\, dx = \mathbb{E}_\lambda \left[\frac{f(X)}{f_\lambda(X)} \mathbf{1}\{X \geq a\} \right].$$

Using the definition of f_λ, the likelihood ratio is

$$\frac{f(X)}{f_\lambda(X)} = e^{-\lambda X + \psi(\lambda)},$$

and therefore

$$\mathbb{P}(X \geq a) = \mathbb{E}_\lambda \left[e^{-\lambda X + \psi(\lambda)} \mathbf{1}\{X \geq a\} \right] = Z(\lambda) \mathbb{E}_\lambda \left[e^{-\lambda X} \mathbf{1}\{X \geq a\} \right].$$

This identity expresses the probability of a rare event as an expectation under the tilted measure f_λ. In principle, this equality could be used for estimation — one could simulate from f_λ and average the weights $e^{-\lambda X + \psi(\lambda)} \mathbf{1}\{X \geq a\}$, exactly as in importance sampling. However, if our goal is not to estimate but to *bound* the probability, we can replace the random weight $e^{-\lambda X}$ by a deterministic upper bound that holds on the event of interest.

On the event $\{X \geq a\}$, we have $e^{-\lambda X} \leq e^{-\lambda a}$. Applying this inequality inside the expectation gives

$$\mathbb{P}(X \geq a) \leq e^{-\lambda a}\, \mathbb{E}_\lambda[e^{\psi(\lambda)} \mathbf{1}\{X \geq a\}] = e^{-\lambda a + \psi(\lambda)}\, \mathbb{P}_\lambda(X \geq a).$$

Since $\mathbb{P}_\lambda(X \geq a) \leq 1$, we finally obtain

$$\mathbb{P}(X \geq a) \leq \exp\big(-\lambda a + \psi(\lambda) \big).$$

This step transforms the exact importance sampling identity into a deterministic upper bound — the *Chernoff bound*. It shows that the same exponential tilting used for variance reduction in Monte Carlo estimation also provides a clean analytical way to control rare-event probabilities.

The bound obtained above depends on the parameter λ. Different values of λ correspond to different tilted distributions f_λ, and hence to different changes of measure. To obtain the tightest bound, we minimize the exponent:

$$\mathbb{P}(X \geq a) \leq \inf_{\lambda > 0} \exp\big(-\lambda a + \psi(\lambda)\big).$$

The optimal value $\lambda^\star$ satisfies the first-order condition

$$\psi'(\lambda^\star) = a.$$

To understand the condition for the optimal $\lambda^\star$, let us compute the derivative of the log-partition function. Starting from

$$\psi(\lambda) = \log \int e^{\lambda x} f(x)\, dx,$$

we differentiate with respect to λ:

$$\psi'(\lambda) = \frac{\int x e^{\lambda x} f(x)\, dx}{\int e^{\lambda x} f(x)\, dx}.$$

This expression can be recognized as the mean of X under the tilted density $f_\lambda(x) \propto e^{\lambda x} f(x)$:

$$\psi'(\lambda) = \mathbb{E}_\lambda[X].$$

Hence, the derivative of the log-partition function coincides with the expected value of X under the exponential tilting. At the optimal value $\lambda^\star$, we have

$$\mathbb{E}_\lambda[X] = \psi'(\lambda^\star) = a,$$

which means that under the optimal tilt, the mean of X equals the threshold a. In probabilistic terms, this tells us that the distribution $f_{\lambda^\star}$ makes the event $\{X \geq a\}$ typical — its average outcome already lies at the boundary of the rare region we are trying to study.

To conclude, let us look at another example. Consider $X \sim \mathrm{Exp}(\theta)$ with density

$$f(x) = \theta e^{-\theta x}, \qquad x \geq 0.$$

We are interested in estimating

$$p = \mathbb{P}(X > a)$$

for a large threshold a. The moment generating function of X is

$$Z(\lambda) = \mathbb{E}[e^{\lambda X}] = \frac{\theta}{\theta - \lambda}, \qquad \lambda < \theta,$$

so that the log-partition function is

$$\psi(\lambda) = -\log(1 - \lambda/\theta), \qquad \psi'(\lambda) = \frac{1}{\theta - \lambda}.$$

The tilted density is then

$$f_\lambda(x) = (\theta - \lambda)\, e^{-(\theta-\lambda)x}, \qquad x \geq 0,$$

which is again an exponential distribution, but with a smaller rate parameter $\theta - \lambda$. As λ increases, the mean $1/(\theta - \lambda)$ shifts to larger values of X, concentrating probability mass in the tail.

At the optimal tilt $\lambda^\star$, the first-order condition

$$\psi'(\lambda^\star) = a$$

gives $1/(\theta - \lambda^\star) = a$, or equivalently

$$\lambda^\star = \theta - \frac{1}{a}.$$

Under this tilted distribution, the mean of X is exactly a, meaning that the rare event $\{X > a\}$ has become typical.

The Chernoff bound then follows from substituting $\lambda^\star$ into

$$\mathbb{P}(X > a) \leq \exp\big(-\lambda a + \psi(\lambda)\big),$$

which yields

$$\mathbb{P}(X > a) \leq \theta a\, e^{1-\theta a}.$$

To illustrate these results numerically, consider again the exponential case with $\theta = 1$ and a large threshold $a = 20$. We compare three approaches for estimating the probability $p = \mathbb{P}(X > a)$: direct Monte Carlo, importance sampling using the tilted distribution $f_{\lambda^\star}$, and the analytical Chernoff bound.

The exact probability is

$$p = e^{-\theta a} = e^{-20} \approx 2.06 \times 10^{-9}.$$

Using 10^6 simulated samples, the naive Monte Carlo estimator gives

$$\hat{p}_{\mathrm{MC}} = 0,$$

since no sample exceeds the threshold $a = 20$; the event is too rare to appear under the original distribution. In contrast, the importance sampling estimator based on the tilted law $f_{\lambda^\star}$ with $\lambda^\star = 0.95$ yields

$$\hat{p}_{\mathrm{IS}} = 2.05 \times 10^{-9},$$

essentially matching the true value with negligible variance. For comparison, the Chernoff bound gives

$$\mathbb{P}(X > a) \leq 1.12 \times 10^{-7},$$

a valid but much looser analytical upper bound.

A significant part of the gap between the bound and the true probability comes from the simplification

$$\mathbb{P}_\lambda(X \geq a) \leq 1.$$

Under the optimal tilt $f_{\lambda^\star}$, the mean of X is indeed equal to the threshold a, yet a considerable fraction of the probability mass remains below it.

In the case of the exponential distribution, the probability of exceeding the threshold can be computed exactly under the tilted law. For $X \sim \text{Exp}(\theta)$ and $f_{\lambda^\star}(x) = (\theta - \lambda^\star)e^{-(\theta-\lambda^\star)x}$, the tail probability under the optimal tilt is

$$\mathbb{P}_{\lambda^\star}(X \geq a) = e^{-(\theta-\lambda^\star)a} = e^{-(\theta-(\theta-1/a))a} = e^{-1}.$$

Hence, under the optimal tilting, exactly a fraction $e^{-1} \approx 0.3679$ of the probability mass of $f_{\lambda^\star}$ lies above the threshold a. Incorporating this term allows us to compare directly the corrected and uncorrected forms in simulation. For $\theta = 1$ and $a = 20$, the exact probability is $e^{-20} = 2.06 \times 10^{-9}$. When the correction factor $\mathbb{P}_{\lambda^\star}(X \geq a) = e^{-1}$ is included, the expression

$$e^{-1}\, e^{1-\theta a} = 4.13 \times 10^{-8}$$

is much closer to the true value than the uncorrected Chernoff bound

$$e^{1-\theta a} = 1.12 \times 10^{-7}.$$

This constant correction reflects how much of the tilted mass remains below the threshold. For asymmetric distributions such as the exponential, this fraction depends on the shape of the tail—here it is e^{-1}.

3 Variance under Exponential Tilting

We have already seen that the tilted measure f_λ makes the region of interest, such as $X > a$, typical. The optimal tilt is achieved when $\psi'(\lambda) = a$, since $\psi'(\lambda) = \mathbb{E}_\lambda[X]$ is the mean of X under the tilted distribution. In other words, the parameter λ shifts the distribution so that its expectation coincides with the point we want to estimate.

To assess the quality of this reweighting, one natural quantity to study is the variance of the estimator under the tilted measure. If the tilted distribution remains highly concentrated around its mean, the importance sampling weights are stable and the estimator is efficient. Conversely, if the tilted law is too dispersed, the weights fluctuate strongly and the estimator suffers from high variance. Thus, the concentration of the tilted distribution provides a direct measure of the quality of the importance sampling estimator.

This concentration is captured by the second derivative of the log-partition. Indeed, starting from

$$\psi(\lambda) = \log \mathbb{E}\left[e^{\lambda X}\right],$$

we obtain

$$\psi'(\lambda) = \frac{\mathbb{E}\left[Xe^{\lambda X}\right]}{\mathbb{E}\left[e^{\lambda X}\right]} = \mathbb{E}_\lambda[X],$$

which represents the mean of X under the tilted distribution. Differentiating once more,

$$\psi''(\lambda) = \frac{\mathbb{E}\left[X^2e^{\lambda X}\right]}{\mathbb{E}\left[e^{\lambda X}\right]} - \left(\frac{\mathbb{E}\left[Xe^{\lambda X}\right]}{\mathbb{E}\left[e^{\lambda X}\right]}\right)^2 = \mathbb{E}_\lambda\left[X^2\right] - (\mathbb{E}_\lambda[X])^2 = \text{Var}_\lambda[X].$$

Hence, the curvature of the log-partition quantifies how concentrated the tilted measure is around its mean, and therefore how efficient the importance sampling estimator will be.

In some cases, this variance can be uniformly bounded for all values of λ. For instance, when X is a bounded random variable such that $X \in [a, b]$, the variance under any tilted measure satisfies

$$\mathrm{Var}_\lambda [X] \leq \frac{(b-a)^2}{4}.$$

Indeed, note that the distance from X to the midpoint of the interval (a, b) is always smaller than half the length of the interval, that is,

$$X - \frac{a+b}{2} \leq \frac{b-a}{2}.$$

Let $m = \frac{a+b}{2}$. Then

$$(X-m)^2 \leq \left(\frac{b-a}{2}\right)^2.$$

Taking expectations under the tilted law gives

$$\mathbb{E}_\lambda \left[(X-m)^2\right] \leq \left(\frac{b-a}{2}\right)^2.$$

Moreover, since for any constant c

$$\mathrm{Var}_\lambda [X] = \min_{c \in \mathbb{R}} \mathbb{E}_\lambda \left[(X-c)^2\right] \leq \mathbb{E}_\lambda \left[(X-m)^2\right],$$

we obtain

$$\mathrm{Var}_\lambda [X] \leq \left(\frac{b-a}{2}\right)^2,$$

as claimed. This shows that, for any bounded random variable, the variance under exponential tilting remains uniformly controlled. In particular, the curvature of the log-partition function—which determines both the concentration of the tilted law and the stability of the importance sampling estimator—cannot grow without bound.

Now suppose that X is centered, so that $\mathbb{E}[X] = 0$. Then, by definition,

$$\psi(0) = \log \mathbb{E}\left[e^{0 \cdot X}\right] = 0, \qquad \psi'(0) = \mathbb{E}[X] = 0.$$

By the second-order Taylor expansion of ψ, there exists some $\theta \in (0, \lambda)$ such that

$$\psi(\lambda) = \psi(0) + \psi'(0)\lambda + \frac{\lambda^2}{2}\psi''(\theta) = \frac{\lambda^2}{2}\mathrm{Var}_\theta [X] \leq \frac{\lambda^2}{2} \sup_{\theta \in (0,\lambda)} \mathrm{Var}_\theta [X].$$

In particular, knowing how the variance behaves under tilting allows us to control the entire shape of the log-partition function. If the tilted variance remains uniformly bounded, the curvature of ψ is also bounded, and the exponential moments of X grow at most quadratically in λ.

In the special case of bounded variables, combining this with the uniform bound on the tilted variance yields

$$\psi(\lambda) \leq \frac{\lambda^2 (b-a)^2}{8}.$$

This result shows that whenever the variance under exponential tilting is uniformly bounded, the log-partition function grows at most quadratically in λ. Quadratic growth of the log-partition is precisely the hallmark of sub-Gaussian behavior. In the next section, we formalize this connection and show how it allows us to control rare-event probabilities even when the exact log-partition function is unknown.

4 Sub-Gaussian Variables and Hoeffding Inequality

Suppose we would like to apply exponential tilting for importance sampling, but the log-partition function

$$\psi(\lambda) = \log \mathbb{E}[e^{\lambda X}]$$

is unknown or too difficult to compute exactly. In this case, it is often enough to know an *upper bound* on $\psi(\lambda)$. If we can find a simple function that dominates the true log-partition, we can still control rare-event probabilities and obtain exponential tail bounds.

For instance, suppose we know that

$$\psi(\lambda) \leq \frac{\sigma^2\lambda^2}{2}, \qquad \forall \lambda \in \mathbb{R}.$$

This means that the exponential moments of X grow no faster than those of a Gaussian random variable with variance σ^2. We then say that X is *sub-Gaussian* with variance proxy σ^2.

Recall that, under exponential tilting, the probability of a rare event can be written as

$$\mathbb{P}(X \geq a) = \mathbb{E}_\lambda\left[e^{-\lambda X+\psi(\lambda)}\mathbf{1}\{X \geq a\}\right] = e^{\psi(\lambda)}\,\mathbb{E}_\lambda\left[e^{-\lambda X}\mathbf{1}\{X \geq a\}\right].$$

If the exact log-partition function is unknown, we can replace it by any valid upper bound. Using the sub-Gaussian condition above, the Chernoff argument gives

$$\mathbb{P}(X \geq a) \leq \inf_{\lambda>0} e^{-\lambda a+\psi(\lambda)} \leq \inf_{\lambda>0} e^{-\lambda a+\frac{\sigma^2\lambda^2}{2}}.$$

Minimizing the exponent with respect to λ yields $\lambda^\star = a/\sigma^2$, which gives

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

Hence, any random variable whose log-partition function is bounded by a quadratic exhibits Gaussian-type tail decay. Even though we cannot perform exact importance sampling without knowing the true normalizing constant $e^{\psi(\lambda)}$, the inequality above provides an accurate analytical estimate of the rare-event probability.

As we have seen in the previous section, bounded random variables satisfy a uniform bound on the tilted variance, and therefore their log-partition function grows at most quadratically. This means that any bounded variable is automatically sub-Gaussian, with parameter

$$\sigma^2 = \frac{(b-a)^2}{4}.$$

This observation leads directly to one of the most fundamental results in the theory of concentration inequalities, known as *Hoeffding's lemma*, which formalizes this fact and provides explicit exponential bounds for bounded variables.

Teorema 1 (Hoeffding's Lemma)**.** *Let X be a random variable such that $X \in [a,b]$ and $\mathbb{E}[X] = 0$. Then, for all $\lambda \in \mathbb{R}$,*

$$\psi(\lambda) = \log \mathbb{E}\left[e^{\lambda X}\right] \leq \frac{\lambda^2(b-a)^2}{8}.$$

In particular, X is sub-Gaussian with variance proxy $\sigma^2 = (b-a)^2/4$.

The Hoeffding lemma immediately implies an exponential tail bound for the sum of independent bounded variables.

Teorema 2 (Hoeffding's Inequality)**.** *Let $X_1, \ldots, X_n$ be independent random variables such that $X_i \in [a_i, b_i]$ and $\mathbb{E}[X_i] = 0$ for all i. Then, for any $t > 0$,*

$$\mathbb{P}\left(\sum_{i=1}^{n} X_i \geq t\right) \leq \exp\left(-\frac{2t^2}{\sum_{i=1}^{n}(b_i - a_i)^2}\right).$$

Proof. By the Hoeffding lemma, each X_i satisfies

$$\mathbb{E}\left[e^{\lambda X_i}\right] \leq \exp\left(\frac{\lambda^2 (b_i - a_i)^2}{8}\right).$$

Since the X_i are independent,

$$\mathbb{E}\left[e^{\lambda \sum_{i=1}^{n} X_i}\right] = \prod_{i=1}^{n} \mathbb{E}\left[e^{\lambda X_i}\right] \leq \exp\left(\frac{\lambda^2}{8} \sum_{i=1}^{n} (b_i - a_i)^2\right).$$

Applying Chernoff's bound,

$$\mathbb{P}\left(\sum_{i=1}^{n} X_i \geq t\right) \leq \inf_{\lambda > 0} \exp\left(-\lambda t + \frac{\lambda^2}{8} \sum_{i=1}^{n} (b_i - a_i)^2\right).$$

Minimizing the exponent with respect to λ gives

$$\lambda^\star = \frac{4t}{\sum_{i=1}^{n}(b_i - a_i)^2},$$

and substituting this value yields

$$\mathbb{P}\left(\sum_{i=1}^{n} X_i \geq t\right) \leq \exp\left(-\frac{2t^2}{\sum_{i=1}^{n}(b_i - a_i)^2}\right),$$

which concludes the proof. □

This inequality shows that sums of independent bounded random variables exhibit Gaussian-type concentration: their tails decay as fast as e^{-ct^2}, with a constant determined solely by the width of the intervals $[a_i, b_i]$. In the context of importance sampling, this means that when each component of the estimator is bounded, the overall estimator remains stable — the effective variance of the tilted measure cannot explode.

The Hoeffding inequality establishes a general sub-Gaussian bound for sums of bounded independent random variables. To illustrate how these theoretical results — exponential tilting, Chernoff bounds, and Hoeffding's inequality — are related in practice, let us consider a simple but fundamental example where all of them can be computed explicitly.

As an example, consider a random walk composed of independent Rademacher steps,

$$S_n = \sum_{i=1}^{n} \sigma_i, \qquad \sigma_i \in \{-1, +1\}, \quad \mathbb{P}(\sigma_i = \pm 1) = \tfrac{1}{2}.$$

We are interested in the rare event $S_n > a$.

For each step, the log-partition function is

$$\psi(\lambda) = \log \mathbb{E}\left[e^{\lambda \sigma_i}\right] = \log(\cosh \lambda),$$

so for the sum S_n, we have $\psi_n(\lambda) = n \log(\cosh \lambda)$. The corresponding tilted measure satisfies

$$\mathbb{P}_\lambda(\sigma_i = 1) = \frac{e^\lambda}{e^\lambda + e^{-\lambda}} = \frac{1}{1 + e^{-2\lambda}}, \qquad \mathbb{P}_\lambda(\sigma_i = -1) = \frac{1}{1 + e^{2\lambda}}.$$

Under this tilted law, the mean and variance of each step are

$$\mathbb{E}_\lambda[\sigma_i] = \tanh \lambda, \qquad \mathrm{Var}_\lambda[\sigma_i] = 1 - \tanh^2 \lambda = \mathrm{sech}^2 \lambda,$$

so that

$$\mathrm{Var}_\lambda[S_n] = n \, \mathrm{sech}^2 \lambda.$$

This expression shows that as the tilt parameter λ increases to make the rare event $S_n > a$ typical (that is, to enforce $\mathbb{E}_\lambda[S_n] = n \tanh \lambda = a$), the variance of the tilted law decreases exponentially. Hence, the tilted distribution becomes increasingly concentrated around its mean, and the importance sampling estimator becomes more stable.

To illustrate this behavior numerically, consider $n = 100$ and $a = 60$. For this example, the exact probability of the event is

$$\mathbb{P}(S_n > a) \approx 1.35 \times 10^{-10}.$$

A naive Monte Carlo simulation with 2×10^5 samples yields no occurrences of the event ($\hat{p}_{\mathrm{MC}} = 0$). In contrast, importance sampling using the optimal tilt $\lambda^\star = \tanh^{-1}(a/n)$ gives

$$\hat{p}_{\mathrm{IS}} \approx 1.34 \times 10^{-10},$$

essentially matching the true probability. For comparison, the Chernoff bound based on the same tilt yields

$$\mathbb{P}(S_n > a) \leq e^{-\lambda^\star a + n \log(\cosh \lambda^\star)} = 4.26 \times 10^{-9}.$$

Finally, applying Hoeffding's inequality with $a_i = -1$ and $b_i = 1$ gives

$$\mathbb{P}(S_n > a) \leq \exp\left(-\frac{a^2}{2n}\right) = 2.76 \times 10^{-8},$$

a looser but fully general bound that holds for all bounded variables.

5 Why the Exponential Tilt?

When performing importance sampling, one might wonder: why use the exponential tilt instead of any other distribution with the same mean? After all, many reweightings can satisfy $\mathbb{E}_g[X] = a$. What makes the exponential tilt special?

To understand this, we take a short detour into convex duality. Given any convex function $\psi : \mathbb{R}^d \to \mathbb{R} \cup \{+\infty\}$, its *convex conjugate* (or Fenchel dual) is defined as

$$\psi^*(y) = \sup_{x \in \mathbb{R}^d} \{ \langle y, x \rangle - \psi(x) \}.$$

This transformation exchanges the roles of x and y, turning linear constraints in the primal space into smooth functions in the dual space.

A central example of this construction is the *log-partition function*

$$\psi(\lambda) = \log \mathbb{E}_f \left[e^{\lambda X} \right],$$

which is convex in λ. Its convex conjugate is obtained by applying the same rule:

$$\psi^*(a) = \sup_{\lambda \in \mathbb{R}} \{ \lambda a - \psi(\lambda) \}.$$

We now compute $\psi^*(a)$ explicitly. For any density g absolutely continuous with respect to f and any $\lambda \in \mathbb{R}$,

$$\psi(\lambda) = \log \mathbb{E}_f \left[e^{\lambda X} \right] = \log \mathbb{E}_g \left[e^{\lambda X} \frac{f(X)}{g(X)} \right] \geq \mathbb{E}_g \left[\lambda X + \log f(X) - \log g(X) \right] = \lambda \mathbb{E}_g [X] - D_{\mathrm{KL}}(g \| f),$$

where the inequality follows from Jensen's inequality. If $\mathbb{E}_g [X] = a$, this yields

$$D_{\mathrm{KL}}(g \| f) \geq \lambda a - \psi(\lambda) \quad \text{for all } \lambda.$$

Since this inequality holds for every value of λ, the left-hand side must be greater than or equal to the largest possible value of the right-hand side. Applying this reasoning gives

$$D_{\mathrm{KL}}(g \| f) \geq \sup_{\lambda} \{ \lambda a - \psi(\lambda) \}.$$

Because this holds for every g satisfying $\mathbb{E}_g [X] = a$, it also holds for the smallest such value of the left-hand side, that is,

$$\inf_{g : \mathbb{E}_g [X] = a} D_{\mathrm{KL}}(g \| f) \geq \sup_{\lambda} \{ \lambda a - \psi(\lambda) \} = \psi^*(a).$$

The maximizer in the definition of the conjugate

$$\psi^*(a) = \sup_{\lambda} \{ \lambda a - \psi(\lambda) \}$$

is found by setting the derivative with respect to λ to zero:

$$\frac{d}{d\lambda} \big(\lambda a - \psi(\lambda) \big) = a - \psi'(\lambda) = a - \frac{\mathbb{E}_f \left[X e^{\lambda X} \right]}{\mathbb{E}_f \left[e^{\lambda X} \right]}.$$

This vanishes at the point $\lambda = \lambda_a$ such that

$$\mathbb{E}_{f_{\lambda_a}} [X] = a.$$

Substituting this value into the conjugate gives

$$\psi^*(a) = a\lambda_a - \psi(\lambda_a).$$

We can now verify that the same expression also emerges from the Kullback–Leibler divergence:

$$D_{\mathrm{KL}}(f_{\lambda_a} \| f) = \mathbb{E}_{f_{\lambda_a}}\left[\log \frac{f_{\lambda_a}(X)}{f(X)}\right] = \mathbb{E}_{f_{\lambda_a}}\left[\lambda_a X - \psi(\lambda_a)\right].$$

Since $\mathbb{E}_{f_{\lambda_a}}[X] = a$, this simplifies to

$$D_{\mathrm{KL}}(f_{\lambda_a} \| f) = \lambda_a a - \psi(\lambda_a) = \psi^*(a).$$

Earlier, we established the inequality

$$D_{\mathrm{KL}}(g \| f) \geq \lambda a - \psi(\lambda) \quad \text{for all } g \text{ with } \mathbb{E}_g[X] = a \text{ and all } \lambda,$$

which implies

$$\inf_{g:\mathbb{E}_g[X]=a} D_{\mathrm{KL}}(g \| f) \geq \psi^*(a).$$

By exhibiting the specific distribution $g = f_{\lambda_a}$ that achieves equality,

$$D_{\mathrm{KL}}(f_{\lambda_a} \| f) = \lambda_a a - \psi(\lambda_a) = \psi^*(a),$$

we see that the inequality is tight, and the exponential tilt f_{λ_a} attains the infimum:

$$\psi^*(a) = \inf_{g:\mathbb{E}_g[X]=a} D_{\mathrm{KL}}(g \| f).$$

In other words, among all distributions g whose mean equals a, the exponential tilt f_{λ_a} is the one closest to f in the sense of Kullback–Leibler divergence. It therefore solves the constrained optimization problem

$$\min_{g:\mathbb{E}_g[X]=a} D_{\mathrm{KL}}(g \| f),$$

showing that the exponential tilt is not an arbitrary choice, but the optimal one dictated by convex duality.