
Large Deviations and Concentration Inequalities

Hoeffding, Chernoff, the Legendre Transform, and Cramér's Theorem

1 Overview: two ways of looking at rare events

Throughout, $X_1, X_2, \dots$ are independent real-valued random variables, usually identically distributed, with mean $\mu = \mathbb{E}[X_1]$. We write

$$S_n = \sum_{i=1}^n X_i, \quad \bar{S}_n = \frac{1}{n} S_n.$$

The weak law of large numbers says $\bar{S}_n \rightarrow \mu$ in probability, i.e. $\mathbb{P}(|\bar{S}_n - \mu| \geq \varepsilon) \rightarrow 0$ for every fixed $\varepsilon > 0$. The central limit theorem describes fluctuations of size $1/\sqrt{n}$: $\sqrt{n}(\bar{S}_n - \mu) \Rightarrow \mathcal{N}(0, \sigma^2)$. Neither result says *how fast* the probability of a deviation of *fixed* size ε goes to zero. That is the question of this lecture, and it is answered in two complementary ways.

1.1 Concentration inequalities

A *concentration inequality* is a non-asymptotic bound of the form

$$\mathbb{P}(|\bar{S}_n - \mu| \geq \varepsilon) \leq C e^{-n c(\varepsilon)} \quad \text{for every } n \geq 1 \text{ and every } \varepsilon > 0,$$

with explicit constants C and an explicit function $c(\cdot)$. Typical features:

- It holds for *each* finite n — usable for designing a system with a given sample size, number of codewords, number of training samples, etc.
- It usually requires only coarse information about the distribution (boundedness, a variance bound, a sub-Gaussian parameter), and therefore holds uniformly over a whole class of distributions.
- The price for this robustness is that the exponent $c(\varepsilon)$ is usually not the best possible for any particular distribution.
- The methods extend beyond sums to smooth functions of many independent variables (McDiarmid, Talagrand, etc.).

1.2 Large deviations

A *large deviations* statement identifies the *exact* exponential rate of decay:

$$\lim_{n \rightarrow \infty} \frac{1}{n} \log \mathbb{P}(\bar{S}_n \geq a) = -I(a), \quad a > \mu,$$

where I is called the *rate function*. Typical features:

- It is asymptotic: it pins down the exponent but says nothing about subexponential prefactors, so $\mathbb{P}(\bar{S}_n \geq a)$ could equal $n^{-1/2} e^{-nI(a)}$ or $10^6 e^{-nI(a)}$ and the statement would still hold.

- It is sharp: the rate I depends on the entire distribution of X_1 (through its moment generating function) and cannot be improved.
- It comes with a description of *how* the rare event happens (exponential tilting; see Section 6).

1.3 How they differ and how they agree

	Concentration inequalities	Large deviations
Regime	every finite n	$n \rightarrow \infty$
Output	an upper bound $\leq Ce^{-nc(\varepsilon)}$	an exact exponent $I(a)$ (upper <i>and</i> lower bounds)
Information used	coarse (range, variance)	full distribution (via the MGF)
Sharpness	exponent typically loose	exponent optimal
Prefactor	explicit	ignored
Typical tool	MGF bound + Markov	MGF bound + Markov (upper), change of measure (lower)

The two viewpoints *agree* in the following sense.

- Both describe deviations of fixed size ε (not of size $1/\sqrt{n}$), and both predict probabilities that decay *exponentially* in n .
- Both rest on the same device — the *Chernoff / exponential Markov method*. The Chernoff bound we derive in Section 3 is simultaneously a concentration inequality (it holds for every n) and the upper half of the large deviations principle (its exponent is the optimal one).
- Concentration inequalities such as Hoeffding’s are obtained by replacing the exact log-MGF by a simpler upper bound valid for a class of distributions. By the order-reversing property of the Legendre transform (Section 4), this produces a *lower bound on the rate function*: Hoeffding’s exponent $\varepsilon^2/2$ never exceeds the true rate $I(\mu + \varepsilon)$. See Section 5.2.

2 Hoeffding’s inequality

We work with random variables whose values stay within distance one of their mean.

2.1 An MGF bound for bounded, zero-mean variables

Lemma 2.1 (Hoeffding’s lemma, symmetric form). *Let Y be a random variable with $\mathbb{E}[Y] = 0$ and $|Y| \leq 1$ almost surely. Then for every $\lambda \in \mathbb{R}$,*

$$\mathbb{E}[e^{\lambda Y}] \leq \cosh \lambda \leq e^{\lambda^2/2}.$$

Proof. Step 1: convexity. Fix λ . The function $y \mapsto e^{\lambda y}$ is convex. Every $y \in [-1, 1]$ is a convex combination of the endpoints,

$$y = \frac{1+y}{2} \cdot (+1) + \frac{1-y}{2} \cdot (-1),$$

with weights $\frac{1+y}{2}, \frac{1-y}{2} \in [0, 1]$ summing to one. By convexity,

$$e^{\lambda y} \leq \frac{1+y}{2} e^{\lambda} + \frac{1-y}{2} e^{-\lambda}, \quad y \in [-1, 1].$$

In words: on $[-1, 1]$ the exponential lies below its chord.

Step 2: take expectations. Substitute $y = Y$ (allowed because $|Y| \leq 1$) and use $\mathbb{E}[Y] = 0$:

$$\mathbb{E}[e^{\lambda Y}] \leq \frac{e^{\lambda} + e^{-\lambda}}{2} + \frac{\mathbb{E}[Y]}{2} (e^{\lambda} - e^{-\lambda}) = \cosh \lambda.$$

Step 3: bound cosh via its derivatives. Since $\cosh \lambda > 0$, it suffices to show that

$$h(\lambda) := \log \cosh \lambda \leq \frac{\lambda^2}{2} \quad \text{for all } \lambda \in \mathbb{R}.$$

Differentiate, using $(\cosh)' = \sinh$ and $(\sinh)' = \cosh$:

$$h'(\lambda) = \frac{\sinh \lambda}{\cosh \lambda} = \tanh \lambda, \quad h''(\lambda) = \frac{\cosh^2 \lambda - \sinh^2 \lambda}{\cosh^2 \lambda} = \frac{1}{\cosh^2 \lambda}.$$

Bound the second derivative. Since $\cosh \lambda \geq 1$ for all λ ,

$$0 < h''(\lambda) \leq 1 \quad \text{for all } \lambda \in \mathbb{R}.$$

Unwind the bound. The initial values are $h(0) = \log \cosh 0 = 0$ and $h'(0) = \tanh 0 = 0$. Because h is even ($\cosh$ is even), it suffices to consider $\lambda \geq 0$. Integrating $h'' \leq 1$ once from 0 to $t \geq 0$,

$$h'(t) = h'(0) + \int_0^t h''(s) ds \leq t.$$

Integrating once more from 0 to $\lambda \geq 0$,

$$h(\lambda) = h(0) + \int_0^{\lambda} h'(t) dt \leq \int_0^{\lambda} t dt = \frac{\lambda^2}{2}.$$

By evenness the same bound holds for $\lambda < 0$. (Equivalently, by Taylor's theorem with Lagrange remainder, $h(\lambda) = h(0) + h'(0)\lambda + \frac{1}{2}h''(\xi)\lambda^2 \leq \frac{1}{2}\lambda^2$ for some ξ between 0 and λ .) Exponentiating,

$$\cosh \lambda = e^{h(\lambda)} \leq e^{\lambda^2/2}. \quad \square$$

Remark 2.2 (An alternative proof of Step 3). One can also compare Taylor series term by term: $\cosh \lambda = \sum_{k \geq 0} \lambda^{2k}/(2k)!$ and $e^{\lambda^2/2} = \sum_{k \geq 0} \lambda^{2k}/(2^k k!)$, and $(2k)! = \prod_{j=1}^k (2j)(2j-1) \geq \prod_{j=1}^k 2j = 2^k k!$. The derivative argument above has the advantage that it generalizes: the same “bound h'' and integrate twice” strategy, applied to the log-MGF of a variable in $[a, b]$, gives the general Hoeffding lemma $\mathbb{E}e^{\lambda(Y - \mathbb{E}Y)} \leq e^{\lambda^2(b-a)^2/8}$.

Remark 2.3. The first inequality is tight: equality holds when Y is a Rademacher variable, $\mathbb{P}(Y = 1) = \mathbb{P}(Y = -1) = \frac{1}{2}$. Among all zero-mean variables in $[-1, 1]$, the one that puts all its mass on the endpoints has the largest MGF. The second inequality replaces $\cosh \lambda$ by the MGF of a standard Gaussian; this is the step that makes Y “sub-Gaussian with parameter 1” and costs sharpness for large $|\lambda|$.

2.2 The inequality

Theorem 2.4 (Hoeffding's inequality). *Let $X_1, \dots, X_n$ be i.i.d. with mean μ and $|X_i - \mu| \leq 1$ almost surely. Then for every $\varepsilon > 0$,*

$$\mathbb{P}(\bar{S}_n - \mu \geq \varepsilon) \leq e^{-n\varepsilon^2/2}, \quad \mathbb{P}(|\bar{S}_n - \mu| \geq \varepsilon) \leq 2e^{-n\varepsilon^2/2}.$$

Proof. Let $Y_i = X_i - \mu$; then the Y_i are i.i.d., $\mathbb{E}[Y_i] = 0$ and $|Y_i| \leq 1$. For any $\lambda > 0$, the map $t \mapsto e^{\lambda t}$ is increasing, so by Markov's inequality,

$$\mathbb{P}\left(\sum_{i=1}^n Y_i \geq n\varepsilon\right) = \mathbb{P}\left(e^{\lambda \sum_{i=1}^n Y_i} \geq e^{\lambda n\varepsilon}\right) \leq e^{-\lambda n\varepsilon} \mathbb{E}\left[e^{\lambda \sum_{i=1}^n Y_i}\right].$$

By independence the MGF factorizes, and Lemma 2.1 bounds each factor:

$$\mathbb{E}\left[e^{\lambda \sum_{i=1}^n Y_i}\right] = \prod_{i=1}^n \mathbb{E}[e^{\lambda Y_i}] \leq (\cosh \lambda)^n \leq e^{n\lambda^2/2}.$$

Therefore

$$\mathbb{P}(\bar{S}_n - \mu \geq \varepsilon) \leq \exp\left(-n\left(\lambda\varepsilon - \frac{\lambda^2}{2}\right)\right) \quad \text{for every } \lambda > 0.$$

The bound holds for all $\lambda > 0$, so we may choose the best one. The exponent $\lambda\varepsilon - \lambda^2/2$ is a concave quadratic maximized at $\lambda^* = \varepsilon$, with maximum value $\varepsilon^2/2$. This gives the one-sided bound.

For the lower tail, apply the same argument to $-Y_i$, which also has zero mean and absolute value at most one: $\mathbb{P}(\bar{S}_n - \mu \leq -\varepsilon) \leq e^{-n\varepsilon^2/2}$. A union bound gives the two-sided statement. $\square$

Remark 2.5 (What the proof actually used).

1. *Identical distribution is not needed.* The proof works verbatim for independent X_i with possibly different means μ_i , provided $|X_i - \mu_i| \leq 1$, with μ replaced by $\frac{1}{n} \sum_i \mu_i$.
2. *General ranges.* If $X_i \in [a_i, b_i]$, the same method (with a slightly more careful MGF lemma) gives $\mathbb{P}(S_n - \mathbb{E}S_n \geq t) \leq \exp(-2t^2 / \sum_i (b_i - a_i)^2)$. With $b_i - a_i = 2$ and $t = n\varepsilon$ this is exactly $e^{-n\varepsilon^2/2}$.
3. *Variance blindness.* The bound does not depend on $\text{Var}(X_i)$. If the variance is much smaller than 1, the bound is far from tight; Bernstein- and Bennett-type inequalities exploit the variance.
4. *The structure of the argument* — exponentiate, apply Markov, factor the MGF, bound the log-MGF, optimize over λ — is the template for everything that follows.

3 The Chernoff bound

We now repeat the argument of Theorem 2.4 *without* replacing the MGF by a convenient upper bound.

3.1 The cumulant generating function

Definition 3.1. The *log-moment generating function* (or *cumulant generating function*) of X is

$$\Lambda(\lambda) = \log \mathbb{E}[e^{\lambda X}] \in (-\infty, +\infty], \quad \lambda \in \mathbb{R},$$

with effective domain $\text{dom } \Lambda = \{\lambda : \Lambda(\lambda) < \infty\}$.

Lemma 3.2 (Basic properties of Λ).

1. $\Lambda(0) = 0$.
2. Λ is convex on $\mathbb{R}$ (as an extended-real-valued function), and $\text{dom } \Lambda$ is an interval containing 0.
3. If $\mu = \mathbb{E}[X]$ is finite, then $\Lambda(\lambda) \geq \lambda\mu$ for all λ .
4. If 0 lies in the interior of $\text{dom } \Lambda$, then Λ is infinitely differentiable there, $\Lambda'(0) = \mu$ and $\Lambda''(0) = \text{Var}(X)$.

Proof. (1) is immediate. For (2), let $t \in (0, 1)$. By Hölder's inequality with exponents $1/t$ and $1/(1-t)$,

$$\mathbb{E}[e^{(t\lambda_1 + (1-t)\lambda_2)X}] = \mathbb{E}[(e^{\lambda_1 X})^t (e^{\lambda_2 X})^{1-t}] \leq (\mathbb{E}e^{\lambda_1 X})^t (\mathbb{E}e^{\lambda_2 X})^{1-t};$$

taking logarithms gives $\Lambda(t\lambda_1 + (1-t)\lambda_2) \leq t\Lambda(\lambda_1) + (1-t)\Lambda(\lambda_2)$. (3) is Jensen's inequality: $\mathbb{E}e^{\lambda X} \geq e^{\lambda \mathbb{E}X}$. (4) follows by differentiating under the expectation (justified by dominated convergence in the interior of the domain): $\Lambda' = \mathbb{E}[Xe^{\lambda X}]/\mathbb{E}[e^{\lambda X}]$ and $\Lambda'' = \mathbb{E}[X^2 e^{\lambda X}]/\mathbb{E}[e^{\lambda X}] - (\Lambda')^2$; evaluate at $\lambda = 0$. $\square$

3.2 The bound

Theorem 3.3 (Chernoff bound). *Let $X_1, \dots, X_n$ be i.i.d. with log-MGF Λ and finite mean μ . For every $a \geq \mu$ and every $n \geq 1$,*

$$\mathbb{P}(\bar{S}_n \geq a) \leq \exp(-n\Lambda^*(a)), \quad \text{where} \quad \Lambda^*(a) = \sup_{\lambda \in \mathbb{R}} \{\lambda a - \Lambda(\lambda)\}.$$

Similarly, $\mathbb{P}(\bar{S}_n \leq a) \leq \exp(-n\Lambda^*(a))$ for every $a \leq \mu$.

Proof. Exactly as in Hoeffding's proof, for any $\lambda \geq 0$,

$$\mathbb{P}(S_n \geq na) \leq e^{-\lambda na} \mathbb{E}[e^{\lambda S_n}] = e^{-\lambda na} (\mathbb{E}e^{\lambda X_1})^n = \exp(-n[\lambda a - \Lambda(\lambda)]).$$

(If $\lambda \notin \text{dom } \Lambda$, the right-hand side is $+\infty$ and the bound is trivially true.) Optimizing over $\lambda \geq 0$,

$$\mathbb{P}(\bar{S}_n \geq a) \leq \exp\left(-n \sup_{\lambda \geq 0} \{\lambda a - \Lambda(\lambda)\}\right). \quad (1)$$

It remains to show that, for $a \geq \mu$, restricting to $\lambda \geq 0$ loses nothing. By Lemma 3.2(3), for $\lambda < 0$,

$$\lambda a - \Lambda(\lambda) \leq \lambda a - \lambda\mu = \lambda(a - \mu) \leq 0 = 0 \cdot a - \Lambda(0).$$

So negative λ never beat $\lambda = 0$, and the supremum over $\lambda \geq 0$ equals the supremum over all $\lambda \in \mathbb{R}$. The lower-tail statement follows by applying this to $-X_i$. $\square$

Remark 3.4. In Hoeffding's proof we bounded $\Lambda_Y(\lambda) \leq \lambda^2/2$ before optimizing and obtained $\sup_{\lambda} \{\lambda\varepsilon - \lambda^2/2\} = \varepsilon^2/2$. Here we keep the exact Λ and obtain the optimization problem

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

This operation, which takes a function Λ and returns a new function Λ^* , is the *Legendre(-Fenchel) transform*. Before going further with probability we pause to understand it.

4 Interlude: the Legendre transform

4.1 Definitions

Definition 4.1 (Legendre–Fenchel transform / convex conjugate). For $f : \mathbb{R}^d \rightarrow (-\infty, +\infty]$, not identically $+\infty$, define

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

f^* is called the *convex conjugate* or *Legendre–Fenchel transform* of f .

The *classical* Legendre transform is the special case where f is differentiable and strictly convex. Then the supremum is attained where the derivative of $x \mapsto xy - f(x)$ vanishes, i.e. at the point $x(y)$ with

$$f'(x(y)) = y, \quad \text{and} \quad f^*(y) = yx(y) - f(x(y)).$$

So the classical transform trades the variable x for the *slope* $y = f'(x)$. The Legendre–Fenchel version uses the supremum instead and makes sense for any function, convex or not, smooth or not. For Λ^* in the Chernoff bound we need this general version because Λ may be $+\infty$ outside an interval and the supremum need not be attained.

4.2 Geometric intuition

There are two equivalent pictures (Figure 1).

1. **Maximal gap.** For a fixed slope y , $f^*(y)$ is the largest amount by which the line $x \mapsto yx$ through the origin exceeds the graph of f .
2. **Supporting lines.** Slide the line of slope y down until it just touches the graph of f from below. The resulting supporting line is $x \mapsto yx - f^*(y)$; its intercept with the vertical axis is $-f^*(y)$.

Picture 2 is the key one. A convex function is the upper envelope of its supporting lines, and a line is determined by its slope and intercept. So f^* records, for each slope y , the intercept of the supporting line with that slope. It is a way of describing the same convex function “by its tangent lines rather than by its points.” This immediately suggests that, for convex f , no information is lost and f can be recovered from f^* — the involution property below.

4.3 The smooth case: slopes are inverse to each other

Let f be twice differentiable on $\mathbb{R}$ with $f'' > 0$, so f' is strictly increasing and has an inverse $(f')^{-1}$ on its range. Writing $x(y) = (f')^{-1}(y)$ and $f^*(y) = yx(y) - f(x(y))$, differentiate:

$$(f^*)'(y) = x(y) + yx'(y) - f'(x(y))x'(y) = x(y) + yx'(y) - yx'(y) = x(y).$$

Hence

$$\boxed{(f^*)' = (f')^{-1}}$$

The derivative maps of f and f^* are inverse functions. Applying the same computation to f^* , the transform of f^* evaluated at x is attained at y with $(f^*)'(y) = x$, i.e. $y = f'(x)$, and

$$f^{**}(x) = x f'(x) - f^*(f'(x)) = x f'(x) - (f'(x)x - f(x)) = f(x).$$

So in the smooth strictly convex case the transform is an *involution*: $f^{**} = f$. We also get the symmetric relation $f(x) + f^*(y) = xy$ whenever $y = f'(x)$.

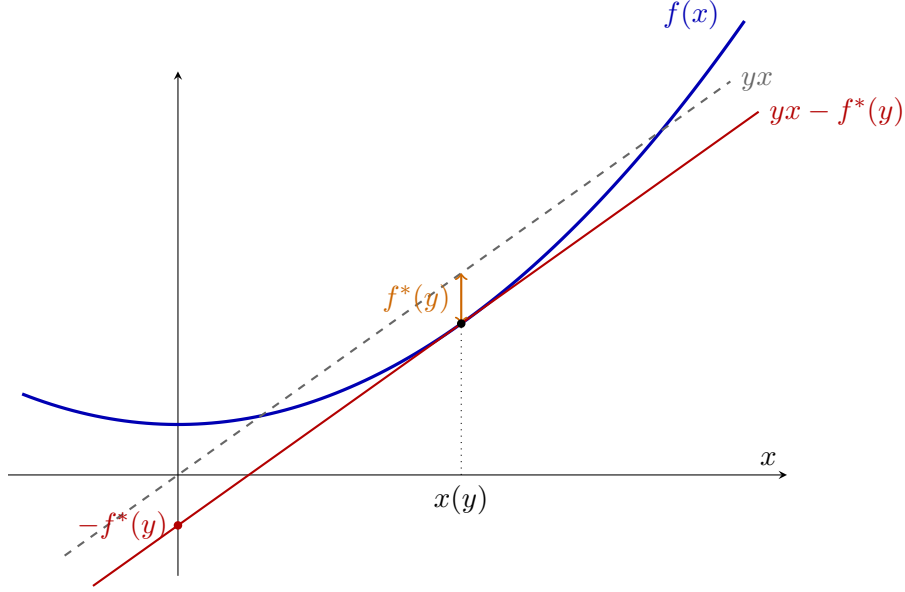


Figure 1: Geometry of the Legendre transform for a fixed slope y . The largest vertical gap between the line yx and f is $f^*(y)$; it occurs at the point $x(y)$ where $f'(x(y)) = y$. Equivalently, the supporting line of slope y has intercept $-f^*(y)$.

4.4 Examples

$f(x)$	$f^*(y)$
$\frac{1}{2}x^2$	$\frac{1}{2}y^2$ (the only self-conjugate function on $\mathbb{R}$)
$\frac{1}{p} x ^p, \ p > 1$	$\frac{1}{q} y ^q, \ \frac{1}{p} + \frac{1}{q} = 1$
e^x	$y \log y - y$ for $y > 0$; 0 for $y = 0$; $+\infty$ for $y < 0$
$-\log x \ (x > 0)$	$-1 - \log(-y)$ for $y < 0$; $+\infty$ for $y \geq 0$
$ x $	0 if $ y \leq 1$; $+\infty$ otherwise
$ax + b$	$-b$ if $y = a$; $+\infty$ otherwise
$\iota_C(x)$ (0 on C , $+\infty$ off C)	$\sigma_C(y) = \sup_{x \in C} \langle x, y \rangle$ (support function)

Notice the duality between *smoothness* and *strict convexity*, and between *growth* and *domain*: $|x|$ has a kink and grows linearly, and its conjugate is flat with bounded domain; e^x grows superlinearly on the right and is bounded on the left, and its conjugate is finite only on $[0, \infty)$.

4.5 Properties

Proposition 4.2. Let $f, g : \mathbb{R}^d \rightarrow (-\infty, +\infty]$, not identically $+\infty$.

- (a) **Convexity.** f^* is convex and lower semicontinuous, whether or not f is.
- (b) **Fenchel–Young inequality.** $f(x) + f^*(y) \geq \langle x, y \rangle$ for all x, y , with equality iff $y \in \partial f(x)$ (for differentiable convex f : iff $y = \nabla f(x)$).
- (c) **Order reversal.** If $f \leq g$ pointwise, then $f^* \geq g^*$.

(d) **Calculus rules.** For $c \in \mathbb{R}$, $\alpha > 0$, $b, x_0 \in \mathbb{R}^d$:

- $(f + c)^* = f^* - c$;
- $(\alpha f)^*(y) = \alpha f^*(y/\alpha)$;
- if $g(x) = f(x - x_0)$ then $g^*(y) = f^*(y) + \langle x_0, y \rangle$;
- if $g(x) = f(x) + \langle b, x \rangle$ then $g^*(y) = f^*(y - b)$.

(e) **Biconjugate is a minorant.** $f^{**} \leq f$. More precisely, f^{**} is the supremum of all affine functions lying below f ; equivalently, it is the largest lower semicontinuous convex function below f (the closed convex envelope).

(f) **Involution (Fenchel–Moreau).** $f^{**} = f$ if and only if f is convex and lower semicontinuous.

Proof. (a) For each fixed x , $y \mapsto \langle x, y \rangle - f(x)$ is affine. A pointwise supremum of affine functions is convex, and lower semicontinuous because its epigraph is an intersection of closed half-spaces.

(b) The inequality is the definition of the supremum. Equality means x maximizes $\langle \cdot, y \rangle - f$, i.e. $f(z) \geq f(x) + \langle y, z - x \rangle$ for all z , which is the definition of $y \in \partial f(x)$.

(c) $\langle x, y \rangle - f(x) \geq \langle x, y \rangle - g(x)$ for every x ; take suprema.

(d) Direct substitution in the definition.

(e) The affine function $\ell(x) = \langle y, x \rangle - c$ satisfies $\ell \leq f$ iff $c \geq \langle y, x \rangle - f(x)$ for all x , iff $c \geq f^*(y)$. So the best affine minorant with slope y is $x \mapsto \langle y, x \rangle - f^*(y)$, and

$$f^{**}(x) = \sup_y \{ \langle y, x \rangle - f^*(y) \}$$

is exactly the supremum of all affine minorants of f . In particular $f^{**} \leq f$. Any closed convex $h \leq f$ is itself the supremum of its affine minorants (see (f)), each of which is an affine minorant of f , so $h \leq f^{**}$.

(f) If $f^{**} = f$ then f is convex and l.s.c. by (a) applied to f^* . Conversely, if f is convex and l.s.c., its epigraph is a closed convex set. If $f^{**}(x_0) < f(x_0)$ for some x_0 , the point $(x_0, f^{**}(x_0))$ lies outside $\text{epi } f$, and the separating hyperplane theorem produces an affine minorant of f whose value at x_0 exceeds $f^{**}(x_0)$ — contradicting (e). (One needs a little care with vertical separating hyperplanes; see Rockafellar, *Convex Analysis*, Thm. 12.2.) $\square$

Example 4.3 (Non-convex f). Let $f(x) = (x^2 - 1)^2$, a double well. Then $f^{**}(x) = 0$ for $|x| \leq 1$ and $f^{**}(x) = f(x)$ for $|x| \geq 1$: the biconjugate “fills in” the non-convex part by the common tangent line (Figure 2). In thermodynamics this is the *Maxwell construction*, and the flat segment corresponds to coexistence of two phases.

4.6 The Legendre transform in physics

Classical mechanics. The Lagrangian $L(q, \dot{q})$ is a function of position and velocity. The Hamiltonian is its Legendre transform in the velocity variable,

$$H(q, p) = \sup_{\dot{q}} \{ p \dot{q} - L(q, \dot{q}) \}, \quad p = \frac{\partial L}{\partial \dot{q}}.$$

For $L = \frac{1}{2}m\dot{q}^2 - V(q)$: $p = m\dot{q}$ and $H = \frac{p^2}{2m} + V(q)$, the total energy. The slope-for-variable trade is exactly velocity $\leftrightarrow$ momentum. Involutivity means one can go back from H to L .

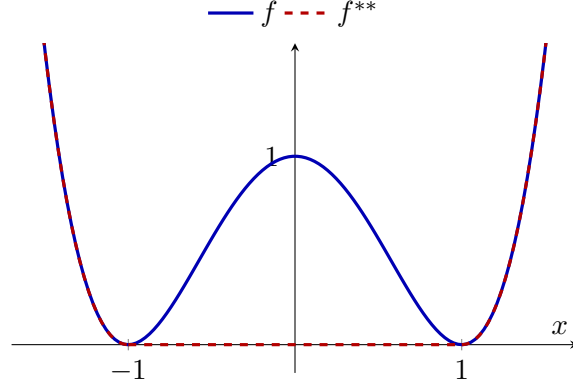


Figure 2: A non-convex function and its biconjugate (closed convex envelope).

Thermodynamics. The internal energy $U(S, V)$ is a convex function of entropy S (and volume V), and temperature is the slope $T = \partial U / \partial S$. It is often more convenient to control temperature than entropy, and the Helmholtz free energy

$$F(T, V) = \inf_S \{U(S, V) - TS\} = -U^*(T)$$

is (minus) the Legendre transform of U in S . Enthalpy and Gibbs free energy arise the same way by trading V for pressure.

Statistical mechanics — the link to this lecture. For a system of N particles with Gibbs measure at inverse temperature β , the free energy per particle $\varphi(\beta) = \lim_N \frac{1}{N} \log Z_N(\beta)$ and the microcanonical entropy $s(u)$ are related by

$$\varphi(\beta) = \sup_u \{s(u) - \beta u\}.$$

This is the same structure as Λ and Λ^* : the log-MGF Λ plays the role of a free energy (log-partition function), the rate function Λ^* plays the role of (negative) entropy, λ plays the role of an inverse temperature, and the tilted measures of Section 6 are the Gibbs (canonical) measures. Large deviations theory can be viewed as the probabilistic foundation of equilibrium statistical mechanics.

4.7 The Legendre transform in optimization

In convex optimization f^* is called the *convex conjugate*, and it is the engine of duality.

- **Weak duality from Fenchel–Young.** The primal problem $\min_x \{f(x) + g(Ax)\}$ has the Fenchel dual

$$\max_y \{-f^*(A^\top y) - g^*(-y)\}.$$

Adding the two Fenchel–Young inequalities

$$f(x) + f^*(A^\top y) \geq \langle x, A^\top y \rangle, \quad g(Ax) + g^*(-y) \geq -\langle Ax, y \rangle$$

shows that every dual value is at most every primal value.

- **Lagrange duality.** For $\min f(x)$ subject to $Ax = b$, the Lagrange dual function is

$$g(\nu) = \inf_x \{f(x) + \nu^\top (Ax - b)\} = -b^\top \nu - f^*(-A^\top \nu),$$

so computing duals amounts to computing conjugates. Convexity of f^* explains why the dual problem is always convex, even when the primal is not.

- **Strong duality** is tied to $f^{**} = f$ (Fenchel–Moreau) plus a constraint qualification.
- **A familiar pair.** The conjugate of $\log \sum_i e^{x_i}$ is the negative entropy $\sum_i p_i \log p_i$ restricted to the probability simplex. This is the same duality that turns the Bernoulli log-MGF into the KL divergence below.

5 Back to Chernoff: the rate function Λ^*

With the Legendre transform in hand, Theorem 3.3 reads

$$\mathbb{P}(\bar{S}_n \geq a) \leq e^{-n\Lambda^*(a)}, \quad \Lambda^* = \text{convex conjugate of } \Lambda.$$

Proposition 5.1 (Properties of Λ^*). *Let X have finite mean μ .*

1. Λ^* is convex and lower semicontinuous.
2. $\Lambda^*(x) \geq 0$ for all x , and $\Lambda^*(\mu) = 0$.
3. Λ^* is nondecreasing on $[\mu, \infty)$ and nonincreasing on $(-\infty, \mu]$.
4. If Λ is differentiable at λ with $\Lambda'(\lambda) = a$, then $\Lambda^*(a) = \lambda a - \Lambda(\lambda)$.
5. Since Λ is convex and l.s.c., $\Lambda^{**} = \Lambda$.

Proof. (1) Proposition 4.2(a). (2) Taking $\lambda = 0$ gives $\Lambda^* \geq 0$; by Jensen $\lambda\mu - \Lambda(\lambda) \leq 0$ for all λ , so $\Lambda^*(\mu) \leq 0$. (3) A convex function that attains its minimum at μ is monotone on each side of μ . (4) $\lambda \mapsto \lambda a - \Lambda(\lambda)$ is concave with zero derivative at λ , hence maximized there. (5) Lower semicontinuity of Λ follows from Fatou's lemma; then apply Proposition 4.2(f). $\square$

5.1 Examples of rate functions

Example 5.2 (Gaussian). $X \sim \mathcal{N}(\mu, \sigma^2)$: $\Lambda(\lambda) = \lambda\mu + \frac{1}{2}\sigma^2\lambda^2$, and

$$\Lambda^*(a) = \frac{(a - \mu)^2}{2\sigma^2}.$$

Example 5.3 (Bernoulli and relative entropy). $X \sim \text{Bern}(p)$: $\Lambda(\lambda) = \log(1 - p + pe^\lambda)$. Solving $\Lambda'(\lambda) = \frac{pe^\lambda}{1-p+pe^\lambda} = a$ gives $e^\lambda = \frac{a(1-p)}{p(1-a)}$, and for $a \in [0, 1]$,

$$\Lambda^*(a) = a \log \frac{a}{p} + (1-a) \log \frac{1-a}{1-p} = D(\text{Bern}(a) \parallel \text{Bern}(p)),$$

with $\Lambda^*(a) = +\infty$ outside $[0, 1]$. The rate function is a KL divergence — a first glimpse of Sanov's theorem.

Example 5.4 (Poisson). $X \sim \text{Poi}(\nu)$: $\Lambda(\lambda) = \nu(e^\lambda - 1)$, and $\Lambda^*(a) = a \log \frac{a}{\nu} - a + \nu$ for $a \geq 0$.

Example 5.5 (Exponential). $X \sim \text{Exp}(1)$: $\Lambda(\lambda) = -\log(1 - \lambda)$ for $\lambda < 1$ and $+\infty$ otherwise. Here $\text{dom } \Lambda = (-\infty, 1)$ is not all of $\mathbb{R}$, yet $\Lambda^*(a) = a - 1 - \log a$ for $a > 0$ is perfectly well defined. Note the linear growth as $a \rightarrow \infty$ (heavier tail, slower decay) compared to the quadratic growth in the Gaussian case.

Example 5.6 (Rademacher). $\mathbb{P}(X = \pm 1) = \frac{1}{2}$: $\Lambda(\lambda) = \log \cosh \lambda$, $\Lambda'(\lambda) = \tanh \lambda$, and for $|a| \leq 1$,

$$\Lambda^*(a) = \frac{1+a}{2} \log(1+a) + \frac{1-a}{2} \log(1-a),$$

with $\Lambda^*(\pm 1) = \log 2$ and $\Lambda^*(a) = +\infty$ for $|a| > 1$.

5.2 Hoeffding as a Chernoff bound with a majorized log-MGF

Let $|X - \mu| \leq 1$ and $Y = X - \mu$. Lemma 2.1 says $\Lambda_Y(\lambda) \leq \lambda^2/2 =: g(\lambda)$. By order reversal (Proposition 4.2(c)) and $g^*(\varepsilon) = \varepsilon^2/2$,

$$\Lambda_Y^*(\varepsilon) \geq \frac{\varepsilon^2}{2}.$$

So Hoeffding's inequality is precisely the Chernoff bound with Λ replaced by a universal quadratic majorant, and its exponent is a universal quadratic *lower bound* on the true rate function of every distribution in the class. Figure 3 shows the comparison for the Rademacher distribution, which is the extremal case for the first inequality in Lemma 2.1.

Near the mean, a second-order expansion (when Λ is smooth near 0) gives

$$\Lambda^*(\mu + \varepsilon) = \frac{\varepsilon^2}{2\sigma^2} + O(\varepsilon^3), \quad \sigma^2 = \text{Var}(X),$$

which connects with the CLT. Since $|X - \mu| \leq 1$ forces $\sigma^2 \leq 1$, this is consistent with Hoeffding's $\varepsilon^2/2$, and it shows that Hoeffding is nearly sharp for small ε exactly when $\sigma^2 \approx 1$, i.e. when X is close to Rademacher.

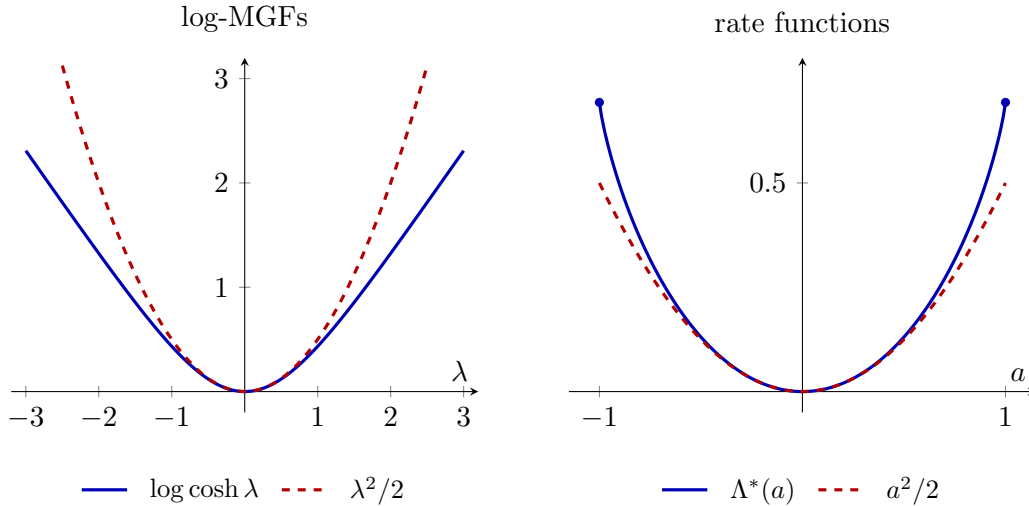


Figure 3: Rademacher variables. Left: the log-MGF lies below the Hoeffding majorant $\lambda^2/2$. Right: conjugation reverses the order, so the true rate function lies above Hoeffding's exponent $a^2/2$. The two agree to second order at $a = 0$.

6 Cramér's theorem: the converse

The Chernoff bound gives an upper bound with exponent $\Lambda^*(a)$. Cramér's theorem says this exponent cannot be improved.

Theorem 6.1 (Cramér). *Let $X_1, X_2, \dots$ be i.i.d. real random variables with log-MGF Λ finite on all of $\mathbb{R}$ and mean μ . Then for every a with $\mu < a < \text{ess sup } X_1$,*

$$\lim_{n \rightarrow \infty} \frac{1}{n} \log \mathbb{P}(\bar{S}_n \geq a) = -\Lambda^*(a).$$

The upper bound $\limsup \leq -\Lambda^*(a)$ is the Chernoff bound (Theorem 3.3), which in fact holds for every n . The content of the theorem is the lower bound

$$\liminf_{n \rightarrow \infty} \frac{1}{n} \log \mathbb{P}(\bar{S}_n \geq a) \geq -\Lambda^*(a).$$

Remark 6.2 (The general statement). In full generality (no assumption on Λ at all), Cramér's theorem is a *large deviation principle* for $\bar{S}_n$ with rate function Λ^* : for every closed set F and open set G in $\mathbb{R}$,

$$\limsup_n \frac{1}{n} \log \mathbb{P}(\bar{S}_n \in F) \leq -\inf_{x \in F} \Lambda^*(x), \quad \liminf_n \frac{1}{n} \log \mathbb{P}(\bar{S}_n \in G) \geq -\inf_{x \in G} \Lambda^*(x).$$

The version above is the special case $F = [a, \infty)$, $G = (a, \infty)$ together with continuity of Λ^* at a . See Dembo and Zeitouni, *Large Deviations Techniques and Applications*, Section 2.2.

6.1 Ingredients of the proof of the lower bound

We only outline the ideas; a full proof is in Dembo–Zeitouni, Theorem 2.2.3.

1. Exponential tilting (change of measure). For $\theta \in \mathbb{R}$ define a new distribution $\mathbb{Q}_\theta$ for X by

$$\frac{d\mathbb{Q}_\theta}{d\mathbb{P}}(x) = e^{\theta x - \Lambda(\theta)}.$$

This is a valid probability measure precisely because $e^{\Lambda(\theta)} = \mathbb{E}e^{\theta X}$. Its mean is $\mathbb{E}_{\mathbb{Q}_\theta}[X] = \Lambda'(\theta)$. Since Λ' is continuous and increasing, with $\Lambda'(0) = \mu$ and $\Lambda'(\theta) \rightarrow \text{ess sup } X$ as $\theta \rightarrow \infty$, we can choose $\theta > 0$ so that the tilted mean equals any target value in $(\mu, \text{ess sup } X)$. Under $\mathbb{Q}_\theta$, the rare event $\{\bar{S}_n \approx a\}$ becomes *typical*.

2. Likelihood ratio for the whole sample. For n i.i.d. samples the likelihood ratio depends on the data only through the sum:

$$\frac{d\mathbb{P}^{\otimes n}}{d\mathbb{Q}_\theta^{\otimes n}}(x_1, \dots, x_n) = \exp(-\theta S_n + n\Lambda(\theta)) = \exp(-n[\theta \bar{S}_n - \Lambda(\theta)]).$$

So on any event where $\bar{S}_n$ stays in a small window, the likelihood ratio is essentially constant.

3. Restrict to a window and apply the LLN under $\mathbb{Q}_\theta$. Choose θ with $\Lambda'(\theta) = a + \delta$ and consider the window $A_n = \{a \leq \bar{S}_n \leq a + 2\delta\}$. Then

$$\mathbb{P}(\bar{S}_n \geq a) \geq \mathbb{P}(A_n) = \mathbb{E}_{\mathbb{Q}_\theta} \left[e^{-n[\theta \bar{S}_n - \Lambda(\theta)]} \mathbf{1}_{A_n} \right] \geq e^{-n[\theta(a+2\delta) - \Lambda(\theta)]} \mathbb{Q}_\theta(A_n).$$

Under $\mathbb{Q}_\theta$ the mean of X is $a + \delta$, the center of the window, so by the weak law of large numbers $\mathbb{Q}_\theta(A_n) \rightarrow 1$; it contributes nothing to the exponent.

4. Identify the exponent and shrink the window. By Proposition 5.1(4), $\theta(a + \delta) - \Lambda(\theta) = \Lambda^*(a + \delta)$, so

$$\liminf_n \frac{1}{n} \log \mathbb{P}(\bar{S}_n \geq a) \geq -\Lambda^*(a + \delta) - \theta\delta.$$

Let $\delta \downarrow 0$, using continuity of Λ^* in the interior of its domain (and that $\theta = \theta(\delta)$ stays bounded).

6.2 Interpretation

- The two halves of the theorem use different tools: *Markov's inequality* for the upper bound and a *change of measure* for the lower bound. This upper-via-Chernoff, lower-via-tilting pattern recurs throughout large deviations theory (Sanov, Gärtner–Ellis, Mogulskii).
- The optimal Chernoff parameter λ^* (the one achieving the supremum in $\Lambda^*(a)$) is exactly the tilting parameter θ with $\Lambda'(\theta) = a$. The Chernoff bound is tight because tilting by λ^* makes the Markov step nearly an equality on the relevant event.
- *How rare events happen.* Conditioned on $\{\bar{S}_n \geq a\}$, the samples look, asymptotically, like i.i.d. draws from $\mathbb{Q}_\theta$ with mean a (the Gibbs conditioning principle). The rare event is realized in the “least unlikely” way, by every sample shifting a little, not by one sample being enormous — this is a light-tail phenomenon; for heavy tails the picture is reversed.
- In the statistical mechanics analogy, $\mathbb{Q}_\theta$ is the canonical ensemble at inverse temperature θ , and the Legendre duality between Λ and Λ^* is the equivalence of the canonical and micro-canonical ensembles.

7 Worked examples

Example 7.1 (Chernoff bound for i.i.d. standard Gaussians). Let $X_1, \dots, X_n$ be i.i.d. $\mathcal{N}(0, 1)$, so $\mu = 0$. Fix $\varepsilon > 0$; we bound $\mathbb{P}(\bar{S}_n \geq \varepsilon)$.

Step 1: the MGF. Completing the square in the exponent,

$$\mathbb{E}[e^{\lambda X}] = \int_{-\infty}^{\infty} e^{\lambda x} \frac{e^{-x^2/2}}{\sqrt{2\pi}} dx = e^{\lambda^2/2} \int_{-\infty}^{\infty} \frac{e^{-(x-\lambda)^2/2}}{\sqrt{2\pi}} dx = e^{\lambda^2/2},$$

because the last integrand is the $\mathcal{N}(\lambda, 1)$ density. Hence $\Lambda(\lambda) = \lambda^2/2$, finite for every $\lambda \in \mathbb{R}$.

Step 2: exponential Markov. For $\lambda > 0$, by Markov's inequality and independence,

$$\mathbb{P}(S_n \geq n\varepsilon) \leq e^{-\lambda n\varepsilon} \mathbb{E}[e^{\lambda S_n}] = e^{-\lambda n\varepsilon} (e^{\lambda^2/2})^n = \exp\left(-n\left(\lambda\varepsilon - \frac{\lambda^2}{2}\right)\right).$$

Step 3: optimize. The exponent $\lambda\varepsilon - \lambda^2/2$ is maximized at $\lambda^* = \varepsilon$ (equivalently, $\Lambda'(\lambda^*) = \lambda^* = \varepsilon$), so

$$\Lambda^*(\varepsilon) = \sup_{\lambda} \left\{ \lambda\varepsilon - \frac{\lambda^2}{2} \right\} = \frac{\varepsilon^2}{2} \quad \text{and} \quad \boxed{\mathbb{P}(\bar{S}_n \geq \varepsilon) \leq e^{-n\varepsilon^2/2}}$$

By symmetry, $\mathbb{P}(|\bar{S}_n| \geq \varepsilon) \leq 2e^{-n\varepsilon^2/2}$.

Comparison with Hoeffding. This is *exactly* the form of Theorem 2.4, yet the hypothesis of Hoeffding's inequality fails badly: a Gaussian is unbounded, so $|X_i - \mu| \leq 1$ is false (indeed $\mathbb{P}(|X_i| > 1) \approx 0.32$). The explanation is that Hoeffding's proof never used boundedness directly. It used boundedness only to obtain the MGF bound

$$\mathbb{E}[e^{\lambda(X-\mu)}] \leq e^{\lambda^2/2} \quad \text{for all } \lambda,$$

and everything after Lemma 2.1 relied on this inequality alone. A random variable satisfying it is called *sub-Gaussian with variance proxy* 1. Bounded variables with $|X - \mu| \leq 1$ are sub-Gaussian by Lemma 2.1; the standard Gaussian is sub-Gaussian with *equality*. So the two results are the same theorem for the class of 1-sub-Gaussian variables, reached from two different sufficient conditions.

Sharpness. Here $\Lambda^*(\varepsilon) = \varepsilon^2/2$ is the true Cramér rate, so the exponent cannot be improved. In fact $\bar{S}_n \sim \mathcal{N}(0, 1/n)$ exactly, and $\mathbb{P}(\bar{S}_n \geq \varepsilon) = Q(\varepsilon\sqrt{n})$, where Q is the standard Gaussian tail. The classical bounds

$$\frac{x}{1+x^2} \frac{e^{-x^2/2}}{\sqrt{2\pi}} \leq Q(x) \leq \frac{1}{x} \frac{e^{-x^2/2}}{\sqrt{2\pi}}, \quad x > 0,$$

show that $Q(\varepsilon\sqrt{n}) = \Theta(n^{-1/2}) e^{-n\varepsilon^2/2}$. The Chernoff bound loses only the polynomial prefactor. For $n = 100$, $\varepsilon = 0.5$: the exact value is $Q(5) \approx 2.87 \times 10^{-7}$, while the Chernoff bound gives $e^{-12.5} \approx 3.73 \times 10^{-6}$, about a factor $13 \approx \varepsilon\sqrt{2\pi n}$ larger.

Example 7.2 (Chernoff bound for Rademacher variables, $p = 1/2$). Let $X_1, \dots, X_n$ be i.i.d. with $\mathbb{P}(X_i = +1) = \mathbb{P}(X_i = -1) = \frac{1}{2}$. Then $\mu = 0$ and $\text{Var}(X_i) = 1$. Fix $0 < a < 1$; we compute the Chernoff bound for $\mathbb{P}(\bar{S}_n \geq a)$.

Step 1: the log-MGF.

$$\mathbb{E}[e^{\lambda X}] = \frac{1}{2}e^\lambda + \frac{1}{2}e^{-\lambda} = \cosh \lambda, \quad \Lambda(\lambda) = \log \cosh \lambda.$$

(This is the case of equality in Step 2 of Lemma 2.1.)

Step 2: find the optimal λ . The function $\lambda \mapsto \lambda a - \log \cosh \lambda$ is concave (Lemma 3.2), so its maximizer solves

$$\Lambda'(\lambda) = \tanh \lambda = a \quad \Longleftrightarrow \quad \lambda^* = \text{artanh } a = \frac{1}{2} \log \frac{1+a}{1-a}.$$

Note $\lambda^* > 0$ for $a > 0$, consistent with Theorem 3.3.

Step 3: evaluate Λ^* . Using $\cosh(\text{artanh } a) = 1/\sqrt{1-a^2}$ (from $\cosh^2 - \sinh^2 = 1$ and $\tanh = a$),

$$\Lambda^*(a) = \lambda^* a - \log \cosh \lambda^* = \frac{a}{2} \log \frac{1+a}{1-a} + \frac{1}{2} \log(1-a^2).$$

Expanding $\log(1-a^2) = \log(1+a) + \log(1-a)$ and collecting terms,

$$\boxed{\Lambda^*(a) = \frac{1+a}{2} \log(1+a) + \frac{1-a}{2} \log(1-a)}$$

and the Chernoff bound reads

$$\mathbb{P}(\bar{S}_n \geq a) \leq e^{-n\Lambda^*(a)} = \left[(1+a)^{-\frac{1+a}{2}} (1-a)^{-\frac{1-a}{2}} \right]^n.$$

Information-theoretic form. Since $X_i = 2B_i - 1$ with $B_i \sim \text{Bern}(1/2)$, the event $\bar{S}_n \geq a$ is the event that the fraction of $+1$'s is at least $q = \frac{1+a}{2}$. One checks that

$$\Lambda^*(a) = D(\text{Bern}(q) \parallel \text{Bern}(\tfrac{1}{2})) = \log 2 - H(q), \quad q = \tfrac{1+a}{2},$$

with H the binary entropy in nats — the Bernoulli example of Section 5 with $p = 1/2$.

Endpoints. As $a \uparrow 1$, $\Lambda^*(a) \rightarrow \log 2$, and the bound becomes 2^{-n} . This is exactly $\mathbb{P}(\bar{S}_n = 1) = \mathbb{P}(\text{all } X_i = +1)$, so at $a = 1$ the Chernoff bound is an equality. For $a > 1$, $\Lambda^*(a) = +\infty$ (the supremum diverges as $\lambda \rightarrow \infty$), matching $\mathbb{P}(\bar{S}_n > 1) = 0$.

Comparison with Hoeffding. The Rademacher variables satisfy Hoeffding's assumption $|X_i - \mu| \leq 1$, so Hoeffding gives $e^{-na^2/2}$. Since $(\Lambda^*)'(a) = \text{artanh } a$, integrating its power series gives

$$\Lambda^*(a) = \sum_{k=1}^{\infty} \frac{a^{2k}}{2k(2k-1)} = \frac{a^2}{2} + \frac{a^4}{12} + \frac{a^6}{30} + \cdots \geq \frac{a^2}{2},$$

so Chernoff is always at least as good, the two agree to second order near $a = 0$, and the Chernoff exponent is strictly larger for $a \neq 0$ (Figure 3). Numerically, for $n = 100$ and $a = 0.5$:

Exact: $\mathbb{P}(\text{Bin}(100, \frac{1}{2}) \geq 75)$	2.82×10^{-7}
Chernoff: $e^{-100\Lambda^*(0.5)}$, $\Lambda^*(0.5) \approx 0.1308$	2.08×10^{-6}
Hoeffding: $e^{-100 \cdot 0.125}$	3.73×10^{-6}

By Cramér's theorem the remaining gap between the exact value and the Chernoff bound is subexponential (a prefactor of order $n^{-1/2}$), whereas the gap between the Chernoff and Hoeffding exponents grows linearly in n .

References

- [1] A. Dembo and O. Zeitouni, *Large Deviations Techniques and Applications*, 2nd ed., Springer, 1998. (Ch. 2: Cramér's theorem.)
- [2] S. Boucheron, G. Lugosi, and P. Massart, *Concentration Inequalities: A Nonasymptotic Theory of Independence*, Oxford University Press, 2013.
- [3] B. Hajek, *Random Processes for Engineers*, Cambridge University Press, 2015.
- [4] R. T. Rockafellar, *Convex Analysis*, Princeton University Press, 1970. (Sec. 12: conjugate functions.)
- [5] S. Boyd and L. Vandenberghe, *Convex Optimization*, Cambridge University Press, 2004. (Sec. 3.3 and Ch. 5.)
- [6] H. Touchette, "The large deviation approach to statistical mechanics," *Physics Reports* 478 (2009), 1–69.