2 Cramér's theorem

2a A very important notion

You should be puzzled: what about the name of the notion? The answer is that this notion was introduced several times, under different names:

* "tilted measure" (in the theory of large deviations);[1]
* "canonical ensemble" (in statistical physics);
* "exponential family" (in statistics);
* "Esscher transform" (mostly, in financial mathematics and actuarial science).

Here is the simplest nontrivial example. Consider n independent copies $X_1, \dots, X_n$ of a random variable X that takes three values $-1, 0, +1$ with equal probabilities (1/3). The frequencies $\nu_{-1} = \frac{1}{n} \cdot \#\{k : X_k = -1\}$, $\nu_0 = \frac{1}{n} \cdot \#\{k : X_k = 0\}$, $\nu_{+1} = \frac{1}{n} \cdot \#\{k : X_k = +1\}$ are random; together they are the so-called empirical distribution $(\nu_{-1}, \nu_0, \nu_{+1})$; and the sample mean $\frac{1}{n}(X_1 + \dots + X_n) = \nu_{+1} - \nu_{-1}$ is also random.

For large n the event $E = \{\frac{1}{n}(X_1 + \dots + X_n) \ge \frac{3}{7}\}$ is of exponentially small probability, and nevertheless, let us consider the conditional distribution of $(\nu_{-1}, \nu_0, \nu_{+1})$ given E. We'll see that, given E,

$$(\nu_{-1}, \nu_0, \nu_{+1}) \to \Big(\frac{1}{7}, \frac{2}{7}, \frac{4}{7} \Big) \quad \text{as } n \to \infty$$

in probability; that is, for every $\varepsilon > 0$,

$$\mathbb{P}\big(|\nu_{-1} - \tfrac{1}{7}| \le \varepsilon, |\nu_0 - \tfrac{2}{7}| \le \varepsilon, |\nu_{+1} - \tfrac{4}{7}| \le \varepsilon \big) \to 1 \quad \text{as } n \to \infty \,.$$

A wonder, isn't it?

As you may guess, more generally, for arbitrary $a \in (1, \infty)$ the condition $E_a = \{\frac{1}{n}(X_1 + \dots + X_n) \ge \frac{a^2-1}{a^2+a+1}\}$ [2] leads in the limit to

$$(\nu_{-1}, \nu_0, \nu_{+1}) = \Big(\frac{1}{a^2 + a + 1}, \frac{a}{a^2 + a + 1}, \frac{a^2}{a^2 + a + 1} \Big) \,,$$

[1] Also, "Cramér-transform" (Hollander, p. 7).

[2] Thus, $E = E_2$.

that is, to

$$\frac{1}{n}(X_1 + \dots + X_n) = \nu_{+1} - \nu_{-1} = \frac{a^2-1}{a^2+a+1} \quad \text{and} \quad \frac{\nu_{+1}}{\nu_0} = \frac{\nu_0}{\nu_{-1}} \, .$$

It is easy to realize that any violation of the equality $\nu_{+1} - \nu_{-1} = \frac{a^2-1}{a^2+a+1}$ leads to an event $\frac{1}{n}(X_1 + \dots + X_n) \ge \frac{a^2-1}{a^2+a+1} + \varepsilon$ of probability exponentially smaller than that of E_a. It is less evident that any violation of the equality $\frac{\nu_{+1}}{\nu_0} = \frac{\nu_0}{\nu_{-1}}$ leads also to exponentially smaller probability. But it does, as we'll see.

This fact illustrates a key principle in large deviation theory:

> ANY LARGE DEVIATION IS DONE IN THE LEAST UNLIKELY OF ALL THE UNLIKELY WAYS!

(*Quoted from: Hollander, p. 10.*)

For $a \in (0,1)$ the same holds under the condition $\frac{1}{n}(X_1 + \dots + X_n) \le \frac{a^2-1}{a^2+a+1}$ $(= -\frac{b^2-1}{b^2+b+1}$ for $b = 1/a \in (0,\infty))$.

It is hardly possible to observe in practice the convergence $(\nu_{-1}, \nu_0, \nu_{+1}) \to \bigl(\frac{1}{7}, \frac{2}{7}, \frac{4}{7}\bigr)$, since for large n it is not feasible to see condition E satisfied even once in a long run.

Now consider a large system of n so-called spin-1 particles, described by the configuration space $\{-1,0,1\}^n$. The average spin $\frac{1}{n}(X_1 + \dots + X_n)$ has practically no chance to reach $3/7$ spontaneously, but can be forced by an external magnetic field. If a measurement shows that the average spin is (close to) $3/7$, then[1] a physicist knows that $(\nu_{-1}, \nu_0, \nu_{+1})$ is (close to) $\bigl(\frac{1}{7}, \frac{2}{7}, \frac{4}{7}\bigr)$; and in particular, $\frac{1}{n}(X_1^2 + \dots + X_n^2)$ is (close to) $5/7$.

The transition from the distribution $(\frac{1}{3}, \frac{1}{3}, \frac{1}{3})$ on the three-point set $\{-1,0,1\}$ to the distribution $\bigl(\frac{1}{a^2+a+1}, \frac{a}{a^2+a+1}, \frac{a^2}{a^2+a+1}\bigr)$ on the same set is a simple example of tilting (called also[2] twisting, or exponential change of measure, etc).

2a1 Definition. (a) Let μ be a probability measure on $\mathbb{R}$. For every $t \in \mathbb{R}$ such that $\int \mathrm{e}^{tx} \, \mu(\mathrm{d}x) = M_\mu(t) < \infty$ we define the *tilted measure* μ_t by

$$\frac{\mathrm{d}\mu_t}{\mathrm{d}\mu}(x) = \frac{1}{M_\mu(t)} \mathrm{e}^{tx} \, ;$$

that is,

$$\int f(x) \, \mu_t(\mathrm{d}x) = \frac{1}{M_\mu(t)} \int f(x) \mathrm{e}^{tx} \, \mu(\mathrm{d}x)$$

[1] Assuming thermal equilibrium in the external field.

[2] Bucklev, p. 13.

for all bounded continuous functions $f : \mathbb{R} \to \mathbb{R}$ (and therefore also for all bounded μ-measurable f). Also, the function $M_\mu : \mathbb{R} \to (0, \infty]$, $M_\mu(t) = \int \mathrm{e}^{tx} \, \mu(\mathrm{d}x)$, is called the *moment generating function* (MGF) of μ; and its logarithm $\Lambda_\mu : \mathbb{R} \to (-\infty, +\infty]$, $\Lambda_\mu(t) = \ln M_\mu(t)$, is called the *cumulant generating function.*[1]

(b) More generally, let μ be a probability measure on a measurable space. For every measurable function[2] u on this space, satisfying $\int \mathrm{e}^u \, \mathrm{d}\mu = M_\mu(u) < \infty$, we define the *tilted measure* μ_u by

$$\frac{\mathrm{d}\mu_u}{\mathrm{d}\mu}(\cdot) = \frac{1}{M_\mu(u)} \mathrm{e}^{u(\cdot)} \, .$$

Also, M is called the *moment generating functional,* and $\Lambda = \ln M$ is called the *cumulant generating functional.*

Note that the tilted measure is a probability measure.

2a2 Example (Standard normal distribution). $\frac{\mu(\mathrm{d}x)}{\mathrm{d}x} = \frac{1}{\sqrt{2\pi}} \mathrm{e}^{-x^2/2}$; $M_\mu(t) = \mathrm{e}^{t^2/2}$ (check it); $\Lambda_\mu(t) = t^2/2$; μ_t is just μ shifted by t, that is, $\int f(x-t) \, \mu_t(\mathrm{d}x) = \int f(x) \, \mu(\mathrm{d}x)$ (check it).

2a3 Example (Fair coin). $\mu(\{-1\}) = 1/2 = \mu(\{+1\})$; $M_\mu(t) = \cosh t$; $\mu_t(\{-1\}) = \frac{\mathrm{e}^{-t}}{\mathrm{e}^t + \mathrm{e}^{-t}}$, $\mu_t(\{+1\}) = \frac{\mathrm{e}^{t}}{\mathrm{e}^t + \mathrm{e}^{-t}}$ (unfair coin).

2a4 Example (Exponential distribution). $\frac{\mu(\mathrm{d}x)}{\mathrm{d}x} = \mathrm{e}^{-x}$ for $x > 0$; $M_\mu(t) = \frac{1}{1-t}$ for $-\infty < t < 1$, otherwise $+\infty$ (check it); μ_t is homothetic to μ, that is, $\int f\big((1-t)x\big) \, \mu_t(\mathrm{d}x) = \int f(x) \, \mu(\mathrm{d}x)$ for $-\infty < t < 1$ (check it).

2a5 Example (Discontinuous generating function). $\frac{\mu(\mathrm{d}x)}{\mathrm{d}x} = \frac{1}{2} \exp(-\sqrt{x})$ for $x > 0$; $M_\mu(t) \le 1$ for $-\infty < t \le 0$, but $M_\mu(t) = +\infty$ for $t > 0$, even though $\int x^k \, \mu(\mathrm{d}x) < \infty$ for all k.

2a6 Exercise.

(a) If $M_\mu(s) < \infty$ then $\forall t \;\; M_\mu(s) M_{\mu_s}(t) = M_\mu(s+t)$;

(b) if both are finite then $(\mu_s)_t = \mu_{s+t}$;

(c) if $M_\mu(u) < \infty$ then $\forall v \;\; M_\mu(u) M_{\mu_u}(v) = M_\mu(u+v)$;

(d) if both are finite then $(\mu_u)_v = \mu_{u+v}$.

Prove it.

[1] Or the logarithmic MGF; also denoted by K_μ.

[2] Real-valued.

In statistical physics (as was noted in Sect. 1b) probabilities are proportional to $\exp(-\beta H(\cdot))$, where $H(\cdot)$ is the energy, and β the inverse temperature.[1] Thus, if we add a function $h(\cdot)$ to the energy $H(\cdot)$ (which is a usual description of an external field, or another influence) then probabilities are multiplied by $\exp\bigl(-\beta h(\cdot)\bigr)$ and a normalizing constant. It means that the initial probability measure μ is replaced with the tilted measure $\mu_{-\beta h}$. Such a measure is called "canonical ensemble" (or "Gibbs measure") corresponding to H and β (or $H+h$ and β). A change of the temperature leads to tilting, too.

Tilting on $\mathbb{R}^d$ is a slight generalization of 2a1(a) toward 2a1(b); x and t run over $\mathbb{R}^d$, and $\langle t, x\rangle$ replaces tx; M and Λ are defined on $\mathbb{R}^d$ (but still real-valued, or $+\infty$).

The general case 2a1(b) boils down to the tilting on $\mathbb{R}^d$ (sometimes even to $d=1$, that is, 2a1(a)) as follows. Given μ and u as in 2a1(b), we consider the distribution of u under μ, that is, the pushforward probability measure ν on $\mathbb{R}$ (denoted often $u_*(\mu)$ or $\mu \circ u^{-1}$) defined by

$$\nu([a,b]) = \mu\bigl(u^{-1}([a,b])\bigr) = \mu\bigl(\{\omega : a \le u(\omega) \le b\}\bigr) \quad \text{for } -\infty < a < b < +\infty\,,$$

that is,[2]

$$\int f\,\mathrm{d}\nu = \int (f \circ u)\,\mathrm{d}\mu$$

for all bounded continuous functions $f : \mathbb{R} \to \mathbb{R}$ (and therefore also for all ν-integrable f). Then the distribution of u under μ_u is

$$\nu_1 = u_*(\mu_u)$$

since

$$M_\nu(1) = \int \mathrm{e}^x\,\nu(\mathrm{d}x) = \int \mathrm{e}^u\,\mathrm{d}\mu = M_\mu(u) < \infty\,;$$

$$\int f(x)\,\nu_1(\mathrm{d}x) = \frac{1}{M_\nu(1)} \int f(x)\mathrm{e}^x\,\nu(\mathrm{d}x) = \\ = \frac{1}{M_\mu(u)} \int (f\circ u)\mathrm{e}^u\,\mathrm{d}\mu = \int (f \circ u)\,\mathrm{d}\mu_u\,.$$

Likewise, given μ and u as above, and another measurable function v on the same measurable space, we consider the joint distribution $\nu = (u,v)_*(\mu)$ of u and v under μ, that is,

$$\nu([a,b]\times[c,d]) = \mu\bigl(\{\omega : a \le u(\omega) \le b, c \le v(\omega) \le b\}\bigr)\,,$$

[1]See "Canonical ensemble" and "Gibbs measure" in Wikipedia.

[2]See "Pushforward measure" in Wikipedia.

and get the joint distribution of u and v under μ_u,

$$\nu_{(1,0)} = (u,v)_*(\mu_u)\,.$$

Also,

$$\nu_{(s,t)} = (u,v)_*(\mu_{su+tv}) \quad \text{whenever } M_\mu(su+tv) < \infty\,.$$

Change of temperature is a special case. Let μ be the canonical ensemble for H and β_1; then the canonical ensemble for H and β_2 is $\mu_{(\beta_1-\beta_2)H}$; and if ν is the joint distribution of H and u at β_1 (that is, under μ), then $\nu_{(\beta_1-\beta_2,0)}$ is the joint distribution of H and u at β_2 (that is, under $\mu_{(\beta_1-\beta_2)H}$). And the distribution of u at β_2 is the corresponding marginal distribution (one-dimensional projection of the two-dimensional distribution).

Likewise, the change of the joint distribution of $u_1,\dots,u_k$ when $\beta_1 H_1$ is replaced with $\beta_2 H_2$ boils down to tilting in $\mathbb{R}^{k+2}$.

In statistics, a natural exponential family on $\mathbb{R}$ consists of probability measures, parametrized by $\theta \in \mathbb{R}$, with the density $f(\cdot|\theta)$ of the form[1]

$$f(x|\theta) = h(x)\exp\bigl(\theta x - A(\theta)\bigr)\,.$$

Clearly, $f(\cdot|\theta)$ is the tilted $f(\cdot|0)$, and $A(\cdot)$ is (up to an additive constant) the cumulant generating function.

Also the Esscher transform[2] is another name of tilting.

2b Surprisingly useful generating functions

The generating functions M_μ and $\Lambda_\mu = \ln M_\mu$, defined in 2a1, are surprisingly useful. So much useful that physicists often calculate in terms of these functions only,[3] without mentioning tilted measures!

First, let μ be a *compactly supported* probability measure on $\mathbb{R}$. Then Λ_μ is finite on the whole $\mathbb{R}$, and for all t,

$$M_\mu(t) = \int \mathrm{e}^{tx}\,\mu(\mathrm{d}x) = \int \Bigl(\sum_{k=0}^\infty \frac{1}{k!} t^k x^k\Bigr)\,\mu(\mathrm{d}x) = \sum_{k=0}^\infty \frac{1}{k!} t^k \int x^k\,\mu(\mathrm{d}x)\,,$$

which justifies the name "moment generating function".

[1] See "Exponential family" and "Natural exponential family" in Wikipedia.

[2] See "Esscher transform" in Wikipedia.

[3] Physicists call M_μ the partition function and denote it $Z_n(\beta)$; they also denote Λ_μ by $\varphi(\beta)$ and call either $\varphi(\beta)/\beta$ or $\varphi(\beta)$ the (canonical) free energy. (See page 30 in "The large deviation approach to statistical mechanics" by H. Touchette, Physics Reports 2009, **478** 1–69.)

In particular, for small t,

$$\begin{aligned}\Lambda_\mu(t) = \ln M_\mu(t) = \ln\Big(1 + tM'_\mu(0) + \frac12 t^2 M''_\mu(0) + O(t^3)\Big) = \\ = tM'_\mu(0) + \frac12 t^2 M''_\mu(0) - \frac12 t^2 \big(M'_\mu(0)\big)^2 + \mathcal{O}(t^3) = \\ = t\int x\,\mu(\mathrm{d}x) + \frac12 t^2 \bigg(\int x^2\,\mu(\mathrm{d}x) - \Big(\int x\,\mu(\mathrm{d}x)\Big)^2 \bigg) + \mathcal{O}(t^3)\,,\end{aligned}$$

that is, $\Lambda'_\mu(0)$ is the expectation of μ, and $\Lambda''_\mu(0)$ is the variance of μ. In fact, the derivatives $\Lambda^{(m)}(0)$ are the so-called cumulants of μ.

The equality

$$\Lambda_\mu(t) + \Lambda_{\mu_t}(s) = \Lambda_\mu(t+s)$$

follows from 2a6. Differentiating it in s at $s = 0$ we get

$$\Lambda^{(k)}_\mu(t) = \Lambda^{(k)}_{\mu_t}(0)\,;$$

in particular, $\Lambda'_\mu(t)$ and $\Lambda''_\mu(t)$ are the expectation and the variance of μ_t.

The variance cannot be negative, therefore Λ_μ is convex. Moreover, it is strictly convex, unless μ is a single atom. Another proof of the convexity uses Hölder's inequality: for $s, t \in \mathbb{R}$ and $\alpha, \beta > 0$ with $\alpha + \beta = 1$,

$$\begin{aligned}M_\mu(\alpha s + \beta t) = \int (\mathrm{e}^{sx})^\alpha (\mathrm{e}^{tx})^\beta\,\mu(\mathrm{d}x) \le \\ \le \Big(\int \mathrm{e}^{sx}\,\mu(\mathrm{d}x)\Big)^\alpha \Big(\int \mathrm{e}^{tx}\,\mu(\mathrm{d}x)\Big)^\beta = M^\alpha_\mu(s) M^\beta_\mu(t)\,;\end{aligned}$$

take the logarithm.

In general, a probability measure μ on $\mathbb{R}$ need not be compactly supported. Rather, $\mu_k \uparrow \mu$ for some compactly supported subprobability measures μ_k. Accordingly, $\Lambda_{\mu_k} \uparrow \Lambda_\mu$. Convexity of Λ_{μ_k} implies convexity of Λ_μ, therefore, convexity of the set $\{t : \Lambda_\mu(t) < \infty\}$. This set is an interval, containing 0, but not always of the form (a, b); it can be $[a, b]$, $[a, b)$, $(a, b]$; it can be unbounded from below, from above, or both; and it can be $\{0\}$.

2b1 Exercise. Find examples (of μ) for all these possibilities.

Consider the interior

$$G = \{t : \Lambda_\mu(t) < \infty\}^\circ = (a, b)\,, \quad -\infty \le a \le 0 \le b \le +\infty\,.$$

Leaving aside the trivial case $a = 0 = b$, we get a convex $\Lambda_\mu : (a, b) \to \mathbb{R}$.

2b2 Lemma. Λ_μ is real-analytic[1] on (a,b).

Proof. It is sufficient to prove that M_μ is real-analytic.[2] Let $a < t-\varepsilon < t < t+\varepsilon < b$; we'll prove that M_μ on $[t-\varepsilon, t+\varepsilon]$ is the sum of a power series. By 2a6(a), $M_{\mu_t}(\pm\varepsilon) < \infty$, and it is sufficient to prove that M_{μ_t} on $[-\varepsilon,\varepsilon]$ is the sum of a power series. Now we forget the original μ and rename μ_t into μ. We need to prove that M_μ on $[-\varepsilon,\varepsilon]$ is the sum of a power series, given that $M_\mu(\pm\varepsilon) < \infty$. We have

$$\sum_{k=0}^{\ell} \frac{t^k x^k}{k!} \to \mathrm{e}^{tx} \quad \text{as } \ell \to \infty\,,$$

$$\forall \ell \quad \biggl| \sum_{k=0}^{\ell} \frac{t^k x^k}{k!} \biggr| \le \mathrm{e}^{|tx|} \le \mathrm{e}^{-\varepsilon x} + \mathrm{e}^{\varepsilon x}$$

for $|t| \le \varepsilon$. By the dominated convergence theorem,

$$M_\mu(t) = \int \mathrm{e}^{tx}\,\mu(\mathrm{d}x) = \lim_\ell \int \sum_{k=0}^{\ell} \frac{t^k x^k}{k!}\,\mu(\mathrm{d}x) = \sum_{k=0}^{\infty} \frac{t^k}{k!} \int x^k\,\mu(\mathrm{d}x)\,.$$

□

Given a measure μ on $\mathbb{R}^d$ we have for $a \in \mathbb{R}^d$ and $t \in \mathbb{R}$

$$M_\mu(ta) = \int \mathrm{e}^{\langle ta,x\rangle}\,\mu(\mathrm{d}x) = \int \mathrm{e}^{ty}\,\nu(\mathrm{d}y) = M_\nu(t)$$

where ν is the distribution of $\langle a,x\rangle$ under μ. Assuming that M_μ is finite on some neighborhood of 0 we see that[3]

$$\frac{\mathrm{d}}{\mathrm{d}t}\Big|_{t=0} M_\mu(ta) = M'_\nu(0) = \int y\,\nu(\mathrm{d}y) = \int \langle a,x\rangle\,\mu(\mathrm{d}x)\,.$$

Similarly,

$$\frac{\mathrm{d}^k}{\mathrm{d}t^k}\Big|_{t=0} M_\mu(ta) = \int \langle a,x\rangle^k\,\mu(\mathrm{d}x)\,.$$

Lemma 2b2 gives

$$M_\mu(a) = \sum_{k=0}^{\infty} \frac{t^k}{k!} \int \langle a,x\rangle^k\,\mu(\mathrm{d}x)$$

[1]That is, locally a sum of a power series.

[2]However, the radius of convergence for Λ_μ may be smaller because of zeros of M_μ on the complex plane.

[3]This derivative is a linear function of a, but be careful: this fact itself does not ensure that M_μ is differentiable at 0.

for all a in a neighborhood of 0,[1] which shows that M_μ (and therefore also Λ_μ) is real-analytic near 0.

By 2a6, $M_\mu(a+b) = M_\mu(a) M_{\mu_a}(b)$ for $a, b \in \mathbb{R}^d$ such that $M_\mu(a) < \infty$. Thus, all said about M_μ and Λ_μ around 0 applies also to $M_\mu(a+\cdot)/M_\mu(a)$ and $\Lambda_\mu(a+\cdot) - \Lambda_\mu(a)$. The functions M_μ and Λ_μ are real-analytic on the interior G of the set $\{a : M_\mu(a) < \infty\}$. For all $a \in G$ and $b \in \mathbb{R}^d$,

$$\frac{\mathrm{d}}{\mathrm{d}t}\Big|_{t=0} \Lambda_\mu(a+tb) = \int \langle b, x \rangle \, \mu_a(\mathrm{d}x) \, ,$$

$$\frac{\mathrm{d}^2}{\mathrm{d}t^2}\Big|_{t=0} \Lambda_\mu(a+bt) = \int \langle b, x \rangle^2 \, \mu(\mathrm{d}x) - \Big(\int \langle b, x \rangle \, \mu(\mathrm{d}x) \Big)^2 \, ,$$

the expectation and the variance of $\langle b, \cdot \rangle$ under μ_a. It follows that Λ_μ is convex on G, and moreover, strictly convex, unless μ sits on some affine subspace of dimension $d-1$ (or less). On the other hand, by Hölder's inequality, Λ_μ is convex on the whole $\mathbb{R}^d$, thus, the set $\{a : \Lambda_\mu(a) < \infty\}$ is convex, and its interior G is also convex.

2c Independent summands

For two independent random variables X and Y, the distribution μ_{X+Y} of the sum $X+Y$ is the convolution $\mu_X * \mu_Y$ of their distributions;

(2c1) $$\int f(z) \, (\mu_X * \mu_Y)(\mathrm{d}z) = \iint f(x+y) \, \mu_X(\mathrm{d}x) \mu_Y(\mathrm{d}y) \, .$$

Taking $f(z) = \mathrm{e}^{tz}$ we have $f(x+y) = f(x)f(y)$, thus,

(2c2) $$M_{X+Y}(t) = M_X(t) M_Y(t) \, , \qquad \Lambda_{X+Y}(t) = \Lambda_X(t) + \Lambda_Y(t) \, .$$

In terms of convolution,

(2c3) $$M_{\mu*\nu}(t) = M_\mu(t) M_\nu(t) \, ; \qquad \Lambda_{\mu*\nu}(t) = \Lambda_\mu(t) + \Lambda_\nu(t) \, .$$

(However, this does not apply to $M(u), \Lambda(u)$.) Here is the tilted convolution:

(2c4) $$(\mu * \nu)_t = \mu_t * \nu_t \quad \text{whenever} M_\mu(t), M_\nu(t) < \infty \, ,$$

since

$$\int f \, \mathrm{d}(\mu * \nu)_t = \frac{1}{M_{\mu*\nu}(t)} \int f(z) \mathrm{e}^{tz} \, (\mu * \nu)(\mathrm{d}z) =$$

[1]In fact, in every ball (centered at 0) on which $M_\mu < \infty$; recall the proof of 2b2.

$$= \frac{1}{M_\mu(t)M_\nu(t)} \iint f(x+y)\mathrm{e}^{t(x+y)}\,\mu(\mathrm{d}x)\nu(\mathrm{d}y) =$$
$$= \iint f(x+y)\mu_t(\mathrm{d}x)\nu_t(\mathrm{d}y) = \int f\,\mathrm{d}(\mu_t * \nu_t)$$

for all bounded continuous f.

We turn to the sum of independent, identically distributed (i.i.d.) random variables; its distribution is

$$\mu^{*n} = \underbrace{\mu * \dots * \mu}_{n} \,.$$

By (2c4),

$$(\mu^{*n})_t = (\mu_t)^{*n} \,, \tag{2c5}$$

thus we need not hesitate writing just μ_t^{*n}.

The Legendre transform Λ_μ^* of Λ_μ will be very useful:

$$\Lambda_\mu^*(x) = \sup_{t\in\mathbb{R}} \big(tx - \Lambda_\mu(t) \big) \in [0,\infty] \,.$$

If $x = \Lambda'_\mu(t)$ for some $t \in G$ (that is, $\Lambda_\mu < \infty$ near t), then

$$\Lambda_\mu^*(x) = tx - \Lambda_\mu(t)$$

by convexity of Λ_μ.

2c6 Example (Standard normal distribution, see 2a2)**.** $\Lambda_\mu(t) = \frac12 t^2$; $x = \Lambda'_\mu(t) = t$; $\Lambda_\mu^*(x) = x \cdot x - \frac12 x^2 = \frac12 x^2$.

2c7 Example (Fair coin, see 2a3)**.** $\Lambda_\mu(t) = \ln\cosh t$; $x = \Lambda'_\mu(t) = \tanh t$; note that $\tanh^2 t + \frac{1}{\cosh^2 t} = 1$, thus $\cosh t = \frac{1}{\sqrt{1-x^2}}$ and $\Lambda_\mu(t) = -\frac12 \ln(1-x^2)$. Also, $t = \operatorname{artanh} x = \frac12 \ln\frac{1+x}{1-x}$, thus

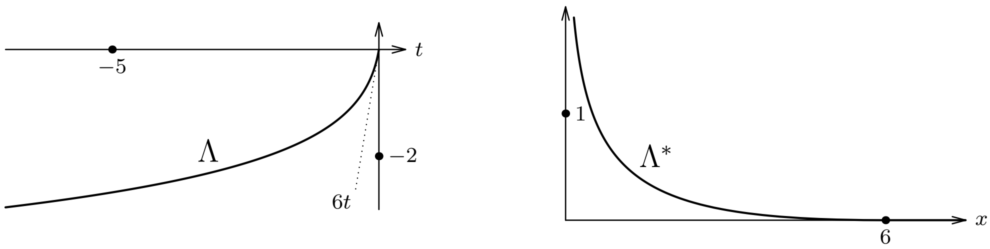
$$\Lambda_\mu^*(x) = \frac{x}{2}\ln\frac{1+x}{1-x} + \frac12\ln(1-x^2) = \frac12(1+x)\ln(1+x) + \frac12(1-x)\ln(1-x)$$

for $x \in [-1,1]$ (otherwise, ∞); just the function γ of (1a1).

2c8 Example (Exponential distribution, see 2a4)**.** $\Lambda_\mu(t) = -\ln(1-t)$; $x = \Lambda'_\mu(t) = \frac{1}{1-t}$; $t = 1 - \frac1x$; $\Lambda_\mu^*(x) = \big(1-\frac1x\big)x - \ln x = x - 1 - \ln x$.

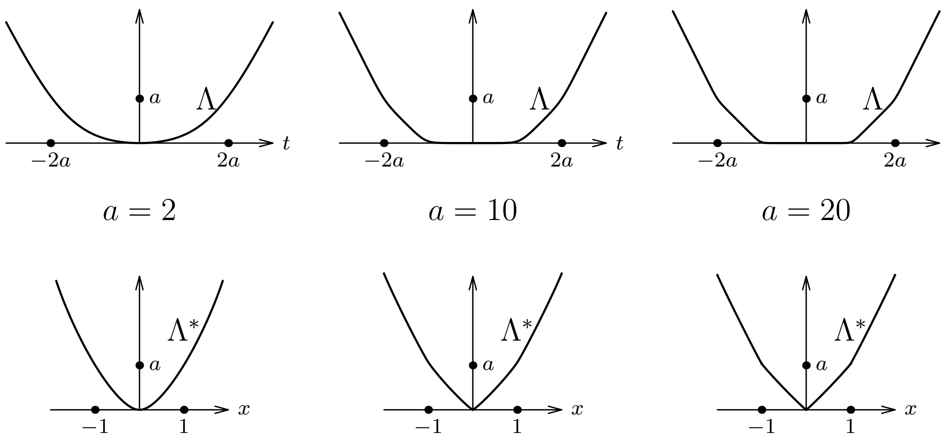
2c9 Example (Discontinuous generating function, see 2a5)**.** $\Lambda_\mu(t) \le 0$ for $-\infty < t \le 0$, but $+\infty$ for $t > 0$. Nevertheless, $\int x\,\mu(\mathrm{d}x) = 6 < \infty$, and $\Lambda'_\mu(0-) = 6$. In fact,

$$M_\mu(t) = \frac{1}{-2t}\biggl(1 - \sqrt{\frac{2\pi}{-2t}}\Phi\Bigl(-\frac{1}{\sqrt{-2t}}\Bigr)\exp\Bigl(\frac{1}{-4t}\Bigr)\biggr) \quad \text{for } t < 0\,.$$



$\Lambda^*(x) = \infty$ for $x \in (-\infty, 0]$; $0 < \Lambda^*(x) < \infty$ for $x \in (0,6)$; and $\Lambda^*(x) = 0$ for $x \in [6,\infty)$. Thus, Λ^* fails to be real-analytic near 6.

2c10 Example (Multiscale case)**.** $\mu(\{-1\}) = \frac12 \mathrm{e}^{-a} = \mu(\{+1\})$, $\mu(\{-2\}) = \frac12 \mathrm{e}^{-3a} = \mu(\{+2\})$, $\mu(\{0\}) = 1 - \mathrm{e}^{-a} - \mathrm{e}^{-3a}$; $\Lambda_\mu(t) = 1 + \mathrm{e}^{-a}(-1+\cosh t) + \mathrm{e}^{-3a}(-1+\cosh 2t)$.



For large a both functions are approximately piecewise linear. Note that the variance (and higher moments) of μ_t is not at all monotone in t.

2c11 Lemma. Let $t \in G$, $x = \Lambda'_\mu(t)$, and $\varepsilon > 0$; then $\mu_t^{*n}\bigl([nx-\varepsilon, nx+\varepsilon]\bigr) > 0$ if and only if $\mu^{*n}\bigl([nx-\varepsilon, nx+\varepsilon]\bigr) > 0$, and in this case

$$\biggl| \ln \frac{\mu_t^{*n}\bigl([nx-\varepsilon, nx+\varepsilon]\bigr)}{\mu^{*n}\bigl([nx-\varepsilon, nx+\varepsilon]\bigr)} - n\Lambda^*_\mu(x) \biggr| \le \varepsilon |t|\,.$$

Proof. Follows from the inequality

$$\left| \ln \frac{\mathrm{d}\mu_t^{*n}}{\mathrm{d}\mu^{*n}}(y) - n\Lambda_\mu^*(x) \right| \le \varepsilon |t| \quad \text{for all } y \in [nx-\varepsilon, nx+\varepsilon] \, ,$$

checked easily:

$$\ln \frac{\mathrm{d}\mu_t^{*n}}{\mathrm{d}\mu^{*n}}(y) = ty - \Lambda_{\mu^{*n}}(t) = t(y-nx) + tnx - n\Lambda_\mu(t) = t(y-nx) + n\Lambda_\mu^*(x) \, .$$

$\square$

We see that $\ln \mu^{*n}\big([nx-\varepsilon, nx+\varepsilon]\big)$ is $\varepsilon|t|$-close to $\ln \mu_t^{*n}\big([nx-\varepsilon, nx+\varepsilon]\big) - n\Lambda_\mu^*(x)$. An upper bound follows immediately:

$$\ln \mu^{*n}\big([nx-\varepsilon, nx+\varepsilon]\big) \le -n\Lambda_\mu^*(x) + \varepsilon|t| \, . \tag{2c12}$$

A lower bound needs more effort. The measure μ_t^{*n} has the expectation $n\Lambda'_\mu(t) = nx$ and the variance $n\Lambda''_\mu(t)$; by Chebyshev's inequality,

$$\mu_t^{*n}\big([nx-\varepsilon, nx+\varepsilon]\big) \ge 1 - \frac{n\Lambda''_\mu(t)}{\varepsilon^2} \, ,$$

which leads to the lower bound

$$\ln \mu^{*n}\big([nx-\varepsilon, nx+\varepsilon]\big) \ge -n\Lambda_\mu^*(x) - \varepsilon|t| + \ln\Big(1 - \frac{n\Lambda''_\mu(t)}{\varepsilon^2}\Big) \, . \tag{2c13}$$

2c14 Theorem. Let $t \in G$, $x = \Lambda'_\mu(t)$, and $\varepsilon_n > 0$ satisfy

$$\frac{\varepsilon_n}{n} \to 0 \, , \quad \frac{\varepsilon_n}{\sqrt{n}} \to \infty.$$

Then

$$\frac1n \ln \mu^{*n}\big([nx-\varepsilon_n, nx+\varepsilon_n]\big) \to -\Lambda_\mu^*(x) \quad \text{as } n \to \infty \, .$$

Proof. The upper limit is at most $-\Lambda_\mu^*(x)$ by (2c12). Taking into account that $\frac{n\Lambda''_\mu(t)}{\varepsilon_n^2} \to 0$ we see that the lower limit is at least $-\Lambda_\mu^*(x)$ by (2c13). $\square$

Here is the same result in a slightly different language.

2c15 Theorem. Let $t \in \mathbb{R}$, and $X_1, X_2, \dots$ be i.i.d. random variables such that $\ln \mathbb{E} \exp \lambda X_1 = \Lambda(\lambda) < \infty$ for all λ close enough to t. Let $\varepsilon_n > 0$ satisfy

$$\varepsilon_n \to 0 \, , \quad \sqrt{n}\,\varepsilon_n \to \infty \, .$$

Denote $x = \Lambda'(t)$. Then

$$\mathbb{P}\Big(x - \varepsilon_n \le \frac{X_1 + \dots + X_n}{n} \le x + \varepsilon_n \Big) = \exp\Big(-n\big(tx - \Lambda(t)\big) + o(n) \Big) \quad \text{as } n \to \infty \, .$$

Now we turn to moderate deviations. Here we assume that $0 \in G$, and in addition, $\Lambda'_\mu(0) = 0$, $\Lambda''_\mu(0) = 1$ (otherwise, use a linear transformation).

2c16 Theorem. Let $x_n \to 0$, $\sqrt{n}|x_n| \to \infty$, and $\varepsilon_n > 0$ satisfy

$$\frac{\varepsilon_n}{n|x_n|} \to 0\,, \quad \frac{\varepsilon_n}{\sqrt{n}} \to \infty\,.$$

Then

$$\ln \mu^{*n}\bigl([nx_n - \varepsilon_n, nx_n + \varepsilon_n]\bigr) = -\frac{1}{2} n x_n^2 \bigl(1 + o(1)\bigr) \quad \text{as } n \to \infty\,.$$

Proof. We take $t_n \to 0$ such that $x_n = \Lambda'_\mu(t_n)$ and note that $x_n \sim t_n$ (that is, their ratio converges to 1), $\Lambda_\mu(t_n) \sim \frac{1}{2}t_n^2$, and $\Lambda^*_\mu(x_n) = t_n x_n - \Lambda_\mu(t_n) \sim \frac{1}{2}x_n^2$.

By (2c12), $\ln(\dots) \le -n\Lambda^*_\mu(x_n) + \varepsilon_n|t_n| = -n \cdot \frac{1}{2}x_n^2\bigl(1+o(1)\bigr) + \varepsilon_n|x_n|\bigl(1+o(1)\bigr) = -\frac{1}{2}nx_n^2\bigl(1+o(1)\bigr)$, since $\varepsilon_n|x_n| \ll nx_n^2$.

By (2c13), taking into account that $\frac{n\Lambda''_\mu(t_n)}{\varepsilon_n^2} \sim \frac{n}{\varepsilon_n^2} \to 0$, we get $\ln(\dots) \ge -n\Lambda^*_\mu(x_n) - \varepsilon_n|t_n| + o(1) = -\frac{1}{2}nx_n^2\bigl(1+o(1)\bigr) + o(1) = -\frac{1}{2}nx_n^2\bigl(1+o(1)\bigr)$, since $nx_n^2 \to \infty$. $\square$

And the same result in the slightly different language.

2c17 Theorem. Let $X_1, X_2, \dots$ be i.i.d. random variables such that $\ln \mathbb{E} \exp \lambda X_1 < \infty$ for all λ close enough to 0, and $\mathbb{E}\, X_1 = 0$, $\mathbb{E}\, X_1^2 = 1$. Let $x_n \in \mathbb{R}$ and $\varepsilon_n > 0$ satisfy

$$|x_n| \to \infty\,, \quad x_n = o(\sqrt{n})\,, \quad \varepsilon_n \to 0\,, \quad |x_n|\varepsilon_n \to \infty\,.$$

Then

$$\mathbb{P}\Bigl(x_n(1-\varepsilon_n) \le \frac{X_1 + \dots + X_n}{\sqrt{n}} \le x_n(1+\varepsilon_n)\Bigr) = \exp\Bigl(-\frac{1}{2}x_n^2\bigl(1+o(1)\bigr)\Bigr)$$

as $n \to \infty$.

2c18 Exercise. Generalize these results (2c11, 2c14–2c17) to probability measures on $\mathbb{R}^d$; in the other language, to i.i.d. random vectors.

The condition "$\frac{\varepsilon_n}{\sqrt{n}} \to \infty$" in Theorem 2c14 may be replaced with $\frac{\varepsilon_n}{\sqrt{n}} \ge$ const with an appropriate absolute constant (think, why). The same applies to "$\sqrt{n}\varepsilon_n \to \infty$" in 2c15, "$\frac{\varepsilon_n}{\sqrt{n}} \to \infty$" in 2c16, and "$|x_n|\varepsilon_n \to \infty$" in 2c17.

Much better bounds are obtained via the Berry-Esseen bound for the central limit theorem (CLT). By CLT, the distribution of $\frac{X_1+\dots+X_n}{\sqrt{n}}$ converges weakly to the standard normal distribution (assuming $\mathbb{E}\, X_1 = 0$ and $\mathbb{E}\, X_1^2 = 1$). That is,

$$\sup_{-\infty<a<b<+\infty} \bigl| \mu^{*n}\bigl([\sqrt{n}a, \sqrt{n}b] \bigr) - \bigl(\Phi(b) - \Phi(a) \bigr) \bigr| \to 0 \quad \text{as } n \to \infty \,.$$

By the Berry-Esseen bound, this supremum never exceeds $\text{const} \cdot \mathbb{E}\, |X_1|^3 / \sqrt{n}$, the constant being absolute.[1] Using the fact that $\bigl(\mathbb{E}\, |X_1|^3 \bigr)^{1/3} \le \bigl(\mathbb{E}\, X_1^4 \bigr)^{1/4}$ we get

$$\left| \mu_t^{*n}\bigl([nx-\varepsilon, nx+\varepsilon] \bigr) - \left(2\Phi\Bigl(\frac{\varepsilon}{\sqrt{n\Lambda''_\mu(t)}} \Bigr) - 1 \right) \right| \le \frac{\text{const}}{\sqrt{n}} \cdot \left(\frac{\Lambda_\mu^{(4)}(t)}{\bigl(\Lambda''_\mu(t)\bigr)^2} + 3 \right)^{3/4} .$$

Thus, we may take ε that depends on $\Lambda_\mu^{(4)}(t)$ and $\Lambda''_\mu(t)$ but does not depend on n, and get for $\mu_t^{*n}\bigl([nx-\varepsilon, nx+\varepsilon] \bigr)$ a lower bound of order $1/\sqrt{n}$, which leads to the lower bound

$$\ln \mu^{*n}\bigl([nx-\varepsilon, nx+\varepsilon] \bigr) \ge -n\Lambda_\mu^*(x) - \mathcal{O}(|t|) - \frac12 \ln n - \mathcal{O}(1) \,.$$

The same Berry-Esseen bound gives $\mu_t^{*n}\bigl([nx-\varepsilon, nx+\varepsilon] \bigr) = \mathcal{O}(1/\sqrt{n})$, which leads to the upper bound

$$\ln \mu^{*n}\bigl([nx-\varepsilon, nx+\varepsilon] \bigr) \le -n\Lambda_\mu^*(x) + \mathcal{O}(|t|) - \frac12 \ln n + \mathcal{O}(1) \,.$$

In both cases, LDP and MDP, t is bounded (in n); also $\Lambda_\mu^{(4)}(t)$ and $\Lambda''_\mu(t)$ are bounded; thus,

$$\mu^{*n}\bigl([nx-\varepsilon, nx+\varepsilon] \bigr) = \frac{1}{\sqrt{n}} \exp\bigl(-n\Lambda_\mu^*(x) + \mathcal{O}(1) \bigr) \,.$$

Here are sLD-counterparts of the LD-theorems 2c14, 2c15.

2c19 Theorem. Let $t \in G$ and $x = \Lambda'_\mu(t)$. Then for every $\varepsilon > 0$ large enough,

$$\mu^{*n}\bigl([nx-\varepsilon, nx+\varepsilon] \bigr) = \frac{1}{\sqrt{n}} \exp\bigl(-n\Lambda_\mu^*(x) + \mathcal{O}(1) \bigr) \quad \text{as } n \to \infty \,.$$

2c20 Theorem. Let $t \in \mathbb{R}$, and $X_1, X_2, \dots$ be i.i.d. random variables such that $\ln \mathbb{E}\, \exp \lambda X_1 = \Lambda(\lambda) < \infty$ for all λ close enough to t. Denote $x = \Lambda'(t)$. Then for every $\varepsilon > 0$ large enough,

$$\mathbb{P}\left(x - \frac{\varepsilon}{n} \le \frac{X_1 + \dots + X_n}{n} \le x + \frac{\varepsilon}{n} \right) = \frac{1}{\sqrt{n}} \exp\Bigl(-n\bigl(tx - \Lambda(t) \bigr) + \mathcal{O}(1) \Bigr)$$

as $n \to \infty$.

[1] See "Berry-Esseen theorem" in Wikipedia.

Think, what happens for the fair coin case, if $\varepsilon < 1/2$.

It is possible to get an approximation up to equivalence (that is, $o(1)$ instead of $\mathcal{O}(1)$ under $\exp(\dots)$), but not easily. To this end, first of all, one has to separate lattice and non-lattice distributions, and not only in proofs but also in formulations.

Now, what about sMD? Here we assume (as before) that $0 \in G$, $\Lambda'_\mu(0) = 0$, $\Lambda''_\mu(0) = 1$, and $x_n \to 0$, $\sqrt{n}|x_n| \to \infty$. We take (again) $t_n \to 0$ such that $x_n = \Lambda'_\mu(t_n)$; still, $x_n \sim t_n$, $\Lambda_\mu(t_n) \sim \frac12 t_n^2$, and $\Lambda^*_\mu(x_n) \sim \frac12 x_n^2$. However, now this relation does not satisfy us! Now we need $\Lambda^*_\mu(x_n) = \frac12 x_n^2 + \mathcal{O}\big(\frac1n\big)$ in order to get the normal approximation $\frac{1}{\sqrt{n}} \exp\big(-\frac{n}{2}x_n^2 + \mathcal{O}(1)\big)$.

The function Λ^*_μ is real-analytic near 0, which follows from the equality $\Lambda^*_\mu(\Lambda'_\mu(t)) = t\Lambda'_\mu(t) - \Lambda_\mu(t)$, since Λ_μ is real-analytic near 0, $\Lambda'_\mu(0) = 0$, and $\Lambda''_\mu(0) = 1 \ne 0$ (indeed, the inverse function to Λ'_μ is real-analytic near 0). For small x we have $\Lambda^*_\mu(x) \sim \frac12 x^2$, thus,

$$\Lambda^*_\mu(x) = \frac12 x^2 - a_0 x^3 - a_1 x^4 - \dots$$

The numbers $a_0, a_1, \dots$ are called the coefficients of the Cramer series.[1] In particular,[2]

$$a_0 = \frac16 \Lambda^{(3)}_\mu(0)\,; \quad a_1 = \frac{1}{24}\big(\Lambda^{(4)}_\mu(0) - 3(\Lambda^{(3)}_\mu(0))^2\big)\,.$$

If $x_n = \mathcal{O}(n^{-1/3})$ then indeed $\Lambda^*_\mu(x_n) = \frac12 x_n^2 + \mathcal{O}\big(\frac1n\big)$, and we get sMD-counterparts of Theorems 2c16, 2c17.

2c21 Theorem. Let $x_n = \mathcal{O}(n^{-1/3})$. Then for every $\varepsilon > 0$ large enough,

$$\mu^{*n}\big([nx_n - \varepsilon, nx_n + \varepsilon]\big) = \frac{1}{\sqrt{n}} \exp\Big(-\frac12 n x_n^2 + \mathcal{O}(1)\Big) \quad \text{as } n \to \infty\,.$$

2c22 Theorem. Let $X_1, X_2, \dots$ be i.i.d. random variables such that $\ln \mathbb{E} \exp \lambda X_1 < \infty$ for all λ close enough to 0, and $\mathbb{E}\, X_1 = 0$, $\mathbb{E}\, X_1^2 = 1$. Let $x_n \in \mathbb{R}$ satisfy

$$x_n = \mathcal{O}(n^{1/6})\,.$$

[1] Some authors define the Cramer series as $a_0 + a_1 x + \dots$ (V.V. Petrov and J. Robinson 2008, "Large deviations for sums of independent non identically distributed random variables", Communications in Statistics **37** 2984–2990); others define it as $a_0 x^3 + a_1 x^4 + \dots$ (L.V. Rozovsky 1999, "On the Cramér series coefficients", Theory Probab. Appl. **43** 152–157).

[2] For a_2, a_3 and a formula for a_k see Rozovsky 1999.

Then for every $\varepsilon > 0$ large enough,

$$\mathbb{P}\Big(x_n - \frac{\varepsilon}{\sqrt{n}} \le \frac{X_1 + \dots + X_n}{\sqrt{n}} \le x_n + \frac{\varepsilon}{\sqrt{n}} \Big) = \frac{1}{\sqrt{n}} \exp\Big(-\frac{1}{2} x_n^2 + \mathcal{O}(1) \Big)$$

as $n \to \infty$.

If $a_0 = 0$, that is, $\mathbb{E}\, X_1^3 = 0$ (in particular, for all symmetric distributions, for example, the fair coin), then "$n^{-1/3}$" in Theorem 2c21 may be replaced with "$n^{-1/4}$", and "$n^{1/6}$" in Theorem 2c22 with "$n^{1/4}$". In general, under these conditions we get "$-\frac{1}{2} n x_n^2 + a_0 n x_n^3$" instead of "$-\frac{1}{2} n x_n^2$" in Theorem 2c21, and "$-\frac{1}{2} x_n^2 + \frac{a_0}{\sqrt{n}} x_n^3$" instead of "$-\frac{1}{2} x_n^2$" in Theorem 2c22. The new factor, being $\exp(\mathcal{O}(n^{1/4}))$, matters for sMD but does not matter for MD.

That is, under $n^{1/6}$ (in terms of 2c22) all distributions μ are served by a single, normal approximation. Between $n^{1/6}$ and $n^{1/4}$ they are not; a one-parameter family of approximations is needed. Likewise, between $n^{1/4}$ and $n^{3/10}$, two parameters are needed (a_0 and a_1; or $\mathbb{E}\, X_1^3$ and $\mathbb{E}\, X_1^4$). And generally, k parameters work between $n^{k/(2(k+2))}$ and $n^{(k+1)/(2(k+3))}$. Somehow, $k = \infty$ means $n^{1/2}$, — the LD territory; and indeed, LD uses a function Λ^*_μ that depends on all μ (rather than several parameters of μ).

In contrast, in the framework of MD (rather than sMD) the normal approximation works in the whole domain $o(n^{1/2})$; the dependence on μ appears at once when $\mathcal{O}(n^{1/2})$ is reached.

Index