3 The least unlikely way maximizes entropy

3a Cramér-Varadhan theorem

Let μ be a probability measure on $\mathbb{R}^d$ that is truly d-dimensional, that is, does not sit on an affine subspace of dimension $d-1$ (or less). Its cumulant generating function Λ_μ is real-analytic on the interior $G \subset \mathbb{R}^d$ of the convex set $\{t \in \mathbb{R}^d : \Lambda_\mu(t) < \infty\}$; we assume that $G \ne \emptyset$. For every $t \in G$ the tilted measure μ_t has the expectation vector $\nabla\Lambda_\mu(t) = \bigl(\frac{\partial}{\partial t_j}\Lambda_\mu(t)\bigr)_j$, and the covariance matrix $\bigl(\frac{\partial^2}{\partial t_i \partial t_j}\Lambda_\mu(t)\bigr)_{i,j}$, positive definite. The Legendre transform Λ^*_μ of Λ_μ is

$$\Lambda^*_\mu(x) = \sup_{t\in\mathbb{R}^d}\bigl(\langle t,x\rangle - \Lambda_\mu(t)\bigr) \in [0,\infty]\,;$$

and if $x = \nabla\Lambda_\mu(t)$ for some $t \in G$, then

$$\Lambda^*_\mu(x) = \langle t,x\rangle - \Lambda_\mu(t)\,.$$

The mapping $\nabla\Lambda_\mu : G \to \mathbb{R}^d$ is one-to-one (since Λ_μ is strictly convex on G), real-analytic, and its differential does not degenerate. It follows that the image $\nabla\Lambda_\mu(G)$ is open, and $\nabla\Lambda_\mu : G \to \nabla\Lambda_\mu(G)$ is a diffeomorphism.

3a1 Theorem. For every nonempty bounded open set $U \subset \mathbb{R}^d$ such that $\overline{U} \subset \nabla\Lambda_\mu(G)$,

$$\ln \mu^{*n}(nU) = -n \inf_{x\in U} \Lambda^*_\mu(x) + o(n) \quad \text{as } n \to \infty\,.$$

(Here $nU = \{nx : x \in U\}$.)

Proof. Theorem 2c14, generalized to $\mathbb{R}^d$ (according to 2c18), gives for $x \in U$

$$\ln \mu^{*n}(nx + [-\varepsilon_n, \varepsilon_n]^d) = -n\Lambda^*_\mu(x) + o(n) \quad \text{as } n \to \infty\,, \tag{3a2}$$

provided that $\varepsilon_n/n \to 0$ and $\varepsilon_n/\sqrt{n} \to \infty$. We take such a sequence $(\varepsilon_n)_n$.

First, we prove the lower bound:

$$\liminf_{n\to\infty} \frac{1}{n} \ln \mu^{*n}(nU) \ge - \inf_{x\in U} \Lambda^*_\mu(x) \, ,$$

that is, $\liminf(\dots) \ge -\Lambda^*_\mu(x)$ for every $x \in U$. To this end we note that, for n large enough,

$$\frac{1}{n} \ln \mu^{*n}(nU) \ge \frac{1}{n} \ln \mu^{*n}\big(nx + [-\varepsilon_n, \varepsilon_n]^d\big) \ge -\Lambda^*_\mu(x) + o(1) \, .$$

For proving the upper bound,

$$\limsup_{n\to\infty} \frac{1}{n} \ln \mu^{*n}(nU) \le - \inf_{x\in U} \Lambda^*_\mu(x) \, ,$$

we note that the $o(n)$ in (3a2) is uniform in $x \in U$ (since t and $\Lambda''_\mu(t)$ are bounded in $x \in U$; inspect the proof of 2c14). We cover[1] U by $\mathcal{O}\big((n/\varepsilon_n)^d\big)$ cubes of the form $x + [-\frac{\varepsilon_n}{n}, \frac{\varepsilon_n}{n}]^d$ with $x \in U$ and get

$$\ln \mu^{*n}(nU) \le \ln \mathcal{O}\big((n/\varepsilon_n)^d\big) + \Big(- n \inf_{x\in U} \Lambda^*_\mu(x) + o(n) \Big) \, ;$$

the upper bound follows, since $n/\varepsilon_n = o(\sqrt{n})$. □

Note that the lower bound holds whenever U (rather than its closure) is contained in $\nabla\Lambda_\mu(G)$. For arbitrary open $U \subset \mathbb{R}^n$ the lower bound still holds, but I do not prove it now; this is quite a different story, and rarely needed. Upper bounds for larger sets are more useful.

3a3 Exercise (upper bound for a half-space)**.** Let $\Lambda_\mu(t) < \infty$ and $c \ge 0$, then[2]

$$\mu^{*n}\big(\{nx : \langle t, x\rangle - \Lambda_\mu(t) \ge c\}\big) \le \mathrm{e}^{-nc} \, .$$

Prove it.

3a4 Lemma (half-space not containing the expectation)**.** Let $0 \in G$ and $c > \int \langle t, x\rangle \, \mu(\mathrm{d}x)$; then

$$\exists \varepsilon > 0 \; \forall n \;\; \mu^{*n}(\{nx : \langle t, x\rangle \ge c\}) \le \mathrm{e}^{-\varepsilon n} \, .$$

[1] For $x \notin U$ note that $U \cap \big(x + [-\frac{\varepsilon_n}{n}, \frac{\varepsilon_n}{n}]^d\big)$ can be covered by no more than 2^d cubes of the form $y + [-\frac{\varepsilon_n}{n}, \frac{\varepsilon_n}{n}]^d$ with $y \in U$.

[2] No need to assume that $G \ne \emptyset$.

Proof. We note that $c > \langle t, \nabla\Lambda_\mu(0)\rangle = \lim_{\delta\to0} \frac{\Lambda_\mu(\delta t)}{\delta}$ and take $\delta > 0$ such that $\delta t \in G$ and $\Lambda_\mu(\delta t) < \delta c$. By 3a3,

$$\mu^{*n}(\{nx : \langle t, x\rangle \ge c\} = \mu^{*n}(\{nx : \langle \delta t, x\rangle - \Lambda_\mu(\delta t) \ge \delta c - \Lambda_\mu(\delta t)\} \le \mathrm{e}^{-\varepsilon n}$$

where $\varepsilon = \delta c - \Lambda_\mu(\delta t) > 0$. □

3a5 Proposition (exponential concentration near the expectation)**.** If $0 \in G$, and $U \subset \mathbb{R}^d$ is a neighborhood of the point $\int x\,\mu(\mathrm{d}x)$, then

$$\exists \varepsilon > 0 \; \forall n \;\; \mu^{*n}(nU) \ge 1 - \mathcal{O}(\mathrm{e}^{-\varepsilon n})\,.$$

Proof. Lemma 3a4 applied to $t = \pm e_1, \dots, \pm e_d$ (where $(e_1, \dots, e_d)$ is the usual basis of $\mathbb{R}^d$) gives

$$\forall \delta > 0 \; \exists \varepsilon > 0 \; \forall n \;\; \mu^{*n}\biggl(n\Bigl(\int x\,\mu(\mathrm{d}x) + [-\delta, \delta]^d\Bigr)\biggr) \ge 1 - 2d\mathrm{e}^{-\varepsilon n}\,.$$

It remains to take δ such that $\int x\,\mu(\mathrm{d}x) + [-\delta, \delta]^d \subset U$. □

3a6 Proposition. Assume that $K \subset \mathbb{R}^d$ is a compact set, and the restriction $\Lambda^*_\mu|_K$ is continuous. Then

$$\ln \mu^{*n}(\{nx : x \in K, \Lambda^*_\mu(x) \ge c\}) \le -nc + o(n) \quad \text{as } n \to \infty$$

for every $c \in (0, \infty)$.

Proof. We'll prove that

$$\limsup_{n\to\infty} \frac{1}{n} \ln \mu^{*n}(\{nx : x \in K, \Lambda^*_\mu(x) \ge c\}) \le -(c - \varepsilon)$$

for arbitrary $\varepsilon > 0$.

For $t \in \mathbb{R}^d$ such that $\Lambda_\mu(t) < \infty$ the half-space

$$H_t = \{x \in \mathbb{R}^d : \langle t, x\rangle - \Lambda_\mu(t) > c - \varepsilon\}$$

satisfies

$$\mu^{*n}(nH_t) \le \exp\bigl(-n(c - \varepsilon)\bigr)$$

by 3a3. It is sufficient to cover the set $\{x : x \in K, \Lambda^*_\mu(x) \ge c\}$ by finitely many half-spaces H_t. The union of all H_t is an open covering of this compact set (recall the definition of Λ^*_μ); it has a finite subcovering. □

3a7 Exercise (minimum of Λ^* on half-space)**.** If $t \in G$ and $x = \nabla\Lambda_\mu(t)$, then $\Lambda^*(y) - \Lambda^*(x) \ge \langle t, y\rangle - \langle t, x\rangle$ for all y; and therefore

$$\min_{y:\langle t,y\rangle \ge \langle t,x\rangle} \Lambda^*(y) = \Lambda^*(x)\,.$$

Prove it.

3b Relative entropy

Let μ, ν be two probability measures on a measurable space Ω, such that

$$\nu = \mu_u \quad \text{for some measurable function } u : \Omega \to \mathbb{R} \,.$$

By the Radon-Nikodym theorem, this happens if and only if μ and ν are mutually absolutely continuous (in other words, equivalent), that is, have the same sets of measure zero.

By 2a6, $\mu = \nu_{-u}$, and $\Lambda_\nu(-u) = -\Lambda_\mu(u)$. Note that, for arbitrary $c \in \mathbb{R}$, $\mu_{u+c} = \mu_u$ and $\Lambda_\mu(u+c) = \Lambda_\mu(u) + c$. We may replace the given u with $u - \Lambda_\mu(u)$, getting

$$\nu = \mu_u \,, \quad \mu = \nu_{-u} \,, \quad \Lambda_\mu(u) = 0 = \Lambda_\nu(-u) \,, \quad u = \ln \frac{\mathrm{d}\nu}{\mathrm{d}\mu} = -\ln \frac{\mathrm{d}\mu}{\mathrm{d}\nu} \,.$$

This is closely related to the so-called Neyman-Pearson (statistical) test:[1] given a sample $(\omega_1, \dots, \omega_n) \in \Omega^n$ from a partially known distribution (either μ or ν), a statistician needs to decide, is it sampled from μ or ν. It is well-known (and easy to see) that the optimal test is based on the so-called observed log-likelihood ratio

$$S_n = \ln \frac{\mathrm{d}\nu^n}{\mathrm{d}\mu^n}(\omega_1, \dots, \omega_n) = \ln \frac{\mathrm{d}\nu}{\mathrm{d}\mu}(\omega_1) + \dots + \ln \frac{\mathrm{d}\nu}{\mathrm{d}\mu}(\omega_n) \,.$$

The statistician compares S_n/n with a chosen threshold γ_n; if $S_n/n \ge \gamma_n$, the decision is ν, and if[2] $S_n/n < \gamma_n$, the decision is μ. Of course, the decision can be erroneous. The decision ν is wrong with probability $\beta_n = \mu^n\{S_n/n \ge \gamma_n\}$;[3] the decision μ is wrong with probability $\alpha_n = \nu^n\{S_n/n < \gamma_n\}$. If ν is the null hypothesis and μ is the alternative hypothesis, then the wrong decision μ is called the type I error ("false positive", "false alarm"), and the wrong decision ν is called the type II error ("false negative", "miss"). Often, one chooses a small α (the size of the test) and takes the greatest γ_n such that $\alpha_n \le \alpha$.[4] The corresponding $1 - \beta_n$ is called the power of the test.

The large deviations theory can help in estimating β_n for large n. To this end, tilting on Ω may be replaced with tilting on $\mathbb{R}$ (recall Sect. 2a); that is, we replace Ω with $\mathbb{R}$, μ with $u_*(\mu)$ (renamed to μ), and ν with $u_*(\nu)$, getting

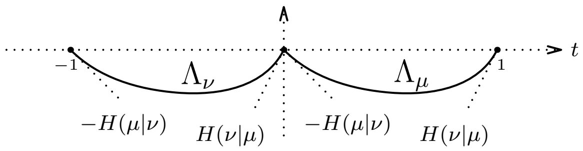
$$\nu = \mu_1 \,, \quad \mu = \nu_{-1} \,, \quad \Lambda_\mu(1) = 0 = \Lambda_\nu(-1) \,, \quad \frac{\mathrm{d}\nu}{\mathrm{d}\mu}(x) = \mathrm{e}^x \,;$$

[1] See Sect. 3.4 in Dembo and Zeitouni, and "Neyman-Pearson lemma" in Wikipedia.

[2] When $S_n/n = \gamma_n$, the decision may be randomized. Never mind.

[3] That is, $\mu^n\big(\{(\omega_1, \dots, \omega_n) \in \Omega^n : S_n(\omega_1, \dots, \omega_n)/n \ge \gamma_n\}\big)$, of course.

[4] This is the Neyman-Pearson approach. There is also Bayesian approach.



$$\Lambda_\nu(t-1) = \Lambda_\mu(t)$$

S_n is distributed either μ^{*n} or ν^{*n}, thus,[1]

$$\alpha_n = \nu^{*n}(-\infty, n\gamma_n)\,, \quad \beta_n = \mu^{*n}[n\gamma_n, \infty)\,.$$

3b1 Exercise. If $\mu \ne \nu$ then $\int x\,\mu(\mathrm{d}x) = \Lambda'_\mu(0) \in [-\infty, 0)$ and $\int x\,\nu(\mathrm{d}x) = \Lambda'_\nu(0) \in (0, \infty]$.

Prove it. Show by examples that all cases (finite or infinite) are possible.

We have

$$\int x\,\nu(\mathrm{d}x) = \int \ln\frac{\mathrm{d}\nu}{\mathrm{d}\mu}\,\mathrm{d}\nu = H(\nu|\mu) \in (0,\infty]\,;$$
$$\int x\,\mu(\mathrm{d}x) = \int \ln\frac{\mathrm{d}\nu}{\mathrm{d}\mu}\,\mathrm{d}\mu = -\int \ln\frac{\mathrm{d}\mu}{\mathrm{d}\nu}\,\mathrm{d}\mu = -H(\mu|\nu) \in [-\infty,0)\,;$$

$H(\nu|\mu)$ is the well-known *relative entropy* of ν relative to μ, called also the Kullback-Leibler divergence[2] $D(\eta||\mu)$. Note that the transition from Ω to $\mathbb{R}$ does not change the relative entropy.

By a well-known result of Chernoff,[3] β_n is exponentially small, as follows.

3b2 Proposition. If $\frac1n \ln\alpha_n \to 0$ and $\limsup_n \alpha_n < 1$, then $\frac1n \ln\beta_n \to -H(\nu|\mu)$.

That is, a large part of the measure ν^n sits on a set whose measure μ^n is $\exp\bigl(-nH(\eta|\mu) + o(n)\bigr)$ (and cannot be smaller).

In particular, when μ is the uniform distribution on a finite Ω, we have

$$H(\nu|\mu) = \ln(\#\Omega) - H(\nu)\,; \tag{3b3}$$

[1] I write just $\nu^{*n}(-\infty, n\gamma_n)$ instead of $\nu^{*n}\bigl((-\infty, n\gamma_n)\bigr)$, of course.

[2] See "Kullback-Leibler divergence" in Wikipedia.

[3] Strangely, it is well-known under the name "Chernoff-Stein lemma" or even "Stein's lemma"! Here is a quote from page 85 of "Statistical signal processing" by Don Johnson: "The attribution to statistician Charles Stein is probably incorrect. Herman Chernoff wrote a paper providing a derivation of this result. A reviewer stated that he thought Stein had derived the result in a technical report, which Chernoff had not seen. Chernoff modified his paper to include the reference without checking it. Chernoff's paper provided the link to Stein. However, Stein later denied he had proved the result; so much for not questioning reviewers! Stein's Lemma should be known as Chernoff's Lemma."

here[1]

$$H(\nu) = -\sum_{\omega} \nu(\omega) \ln \nu(\omega)$$

is the (Shannon) entropy of ν. Thus, a large part of the measure ν^n sits on a set of $\exp\bigl(nH(\nu)+o(n)\bigr)$ points (and cannot be smaller).

Proof. First we prove that $\liminf_n \gamma_n \ge H(\nu|\mu)$ (be it finite or not). Given $a < H(\nu|\mu)$, we'll prove that $\gamma_n > a$ for all n large enough. We cannot use Lemma 3a4, since $\Lambda_\nu(0+)$ may be ∞, but we can adapt that proof as follows.

We note that $\Lambda'_\nu(0-) = H(\nu|\mu) > a$, and take $\delta > 0$ such that $\Lambda_\nu(-\delta) < -\delta a$. By 3a3, for $\varepsilon = -\delta a - \Lambda_\nu(-\delta) > 0$ (and $t = -\delta$),

$$\begin{aligned} \mathrm{e}^{-\varepsilon n} \ge \nu^{*n}\bigl(\{nx : -\delta x - \Lambda_\nu(-\delta) \ge \varepsilon\}\bigr) = \\ = \nu^{*n}\bigl(\{nx : -\delta x \ge -\delta a\}\bigr) = \nu^{*n}(-\infty, na]\,, \end{aligned}$$

thus, $\frac1n \ln \nu^{*n}(-\infty, na] \le -\varepsilon < \frac1n \ln \alpha_n = \frac1n \ln \nu^{*n}(-\infty, n\gamma_n)$ for all n large enough, therefore $na < n\gamma_n$, that is, $\gamma_n > a$, which proves that $\liminf_n \gamma_n \ge H(\nu|\mu)$.

Second, we prove the upper bound: $\limsup_n \frac1n \ln \beta_n \le -H(\nu|\mu)$ (finite or not). Given $a < H(\nu|\mu)$, we'll prove that $\frac1n \ln \beta_n \le -a$, that is, $\beta_n \le \mathrm{e}^{-an}$, for all n large enough. We have $\gamma_n > a$; by 3a3 (for $t = 1$),

$$\beta_n = \mu^{*n}[n\gamma_n, \infty) \le \mu^{*n}[na, \infty) = \mu^{*n}\bigl(\{nx : x - \Lambda_\mu(1) \ge a\}\bigr) \le \mathrm{e}^{-an}\,.$$

Third, assuming $H(\nu|\mu) < \infty$ (otherwise all is done already), we prove the lower bound: $\liminf_n \frac1n \ln \beta_n \ge -H(\nu|\mu)$. Given $a > H(\nu|\mu)$, we'll prove that $\frac1n \ln \beta_n \ge -a + o(1)$ for $n \to \infty$. We have $a > \int x\,\nu(\mathrm{d}x)$; by the weak law of large numbers, $\nu^{*n}[an, \infty) \to 0$; therefore

$$\begin{aligned} \beta_n = \mu^{*n}[n\gamma_n, \infty) \ge \mu^{*n}[n\gamma_n, an] \ge \\ \ge M_{\mu^{*n}}(1)\mathrm{e}^{-an}\mu_1^{*n}[n\gamma_n, an] = \mathrm{e}^{-an}\nu^{*n}[n\gamma_n, an]\,; \end{aligned}$$

$$\begin{aligned} \liminf_n \nu^{*n}[n\gamma_n, an] = \liminf_n \bigl(\nu^{*n}[n\gamma_n, \infty) - \nu^{*n}(an, \infty)\bigr) \ge \\ \ge 1 - \limsup_n \alpha_n > 0\,; \end{aligned}$$

$\frac1n \ln \nu^{*n}[n\gamma_n, an] \to 0$; $\frac1n \ln \beta_n \ge -a + o(1)$. □

[1] Strangely, the relative entropy and the entropy differ by the sign; I do not know why.

3c Sanov's theorem

We return to the empirical distribution $(\nu_{-1}, \nu_0, \nu_{+1})$ mentioned in Sect. 2a. More generally, now we consider d possible values of a random variable, with probabilities $p_1, \dots, p_d > 0$, $p_1 + \dots + d_d = 1$. The frequencies $\eta_1, \dots, \eta_d$ satisfy $\eta_1 + \dots + \eta_d = 1$ and are distributed multinomially,[1]

$$\mathbb{P}\Big(\eta_1 = \frac{a_1}{n}, \dots, \eta_d = \frac{a_d}{n} \Big) = \frac{n!}{a_1! \dots a_d!} p_1^{a_1} \dots p_d^{a_d} \,,$$

and we could use the Stirling formula (as in Sect. 1a). But instead, we introduce such a probability measure μ on $\mathbb{R}^d$:

$$\mu\Big(\Big\{ e_j - \frac{e_1 + \dots + e_d}{d} \Big\} \Big) = p_j \quad \text{for } j = 1, \dots, d \,,$$

where $(e_1, \dots, e_d)$ is the usual basis of $\mathbb{R}^d$. That is, we use the vectors $e_j - \frac{e_1 + \dots + e_d}{d}$ as the possible values of the random variable. These vectors span a $(d-1)$-dimensional vector subspace (a hyperplane) $E = \{(x_1, \dots, x_d) : x_1 + \dots + x_d = 0\}$ in $\mathbb{R}^d$, and we may apply the Cramer-Varadhan theorem within this subspace.[2] Indeed, the measure μ^{*n} is the distribution of $\big(n\eta_1 - \frac{n}{d}, \dots, n\eta_d - \frac{n}{d} \big)$.

The convex hull of the vectors $e_1, \dots, e_d$ is the simplex $\{(x_1, \dots, x_d) \in [0, \infty)^d : x_1 + \dots + x_d = 1\}$, thus, the convex hull of the vectors $e_j - \frac{e_1 + \dots + e_d}{d}$ is the simplex $\{(x_1 - \frac{1}{d}, \dots, x_d - \frac{1}{d}) : x_1, \dots, x_d \ge 0,\ x_1 + \dots + x_d = 1\} = E \cap [-\frac{1}{d}, \infty)^d$. The expectation of μ,

$$\int x \, \mu(\mathrm{d}x) = p_1 \Big(e_1 - \frac{e_1 + \dots + e_d}{d} \Big) + \dots + p_d \Big(e_d - \frac{e_1 + \dots + e_d}{d} \Big) = \\ = \Big(p_1 - \frac{1}{d} \Big) e_1 + \dots + \Big(p_d - \frac{1}{d} \Big) e_d \in E \cap \Big(-\frac{1}{d}, \infty \Big)^d$$

is an interior point of the simplex. And every interior point of the simplex corresponds to some $p_1, \dots, p_d$, thus, to some μ.

A tilted measure μ_t also sits on these d vertices of the simplex, with probabilities $q_j = \frac{1}{M_\mu(t)} \mathrm{e}^{t_j} p_j$ $(j = 1, \dots, d)$. Interestingly, when t runs over E, the tilted measure runs over *all* strictly positive probability measures on these vertices. Indeed, given $q_1, \dots, q_d > 0$, $q_1 + \dots + q_d = 1$, we take

$$s_j = \ln q_j - \ln p_j \,, \quad t_j = s_j - \frac{s_1 + \dots + s_d}{d} \,,$$

[1] The binomial distribution, treated in Sect. 1a, is a special case.

[2] You may choose an orthonormal basis in E, thus turning E into a copy of $\mathbb{R}^{d-1}$.

then μ_t corresponds to $q_1, \dots, q_d$ (since $\mathrm{e}^{t_j} p_j = \mathrm{const} \cdot q_j$).

We have $M_\mu : E \to (0, \infty)$,

$$M_\mu(t) = \int \mathrm{e}^{\langle t,x \rangle} \, \mu(\mathrm{d}x) = p_1 \mathrm{e}^{t_1} + \dots + p_d \mathrm{e}^{t_d} \, ;$$

$\nabla \Lambda_\mu : E \to E$,

$$\nabla \Lambda_\mu(t) = \frac{1}{M_\mu(t)} \nabla M_\mu(t) = \Big(\frac{p_1 \mathrm{e}^{t_1}}{M_\mu(t)} - \frac{1}{d}, \dots, \frac{p_d \mathrm{e}^{t_d}}{M_\mu(t)} - \frac{1}{d} \Big) \, .$$

We know that $\nabla \Lambda_\mu(t) = \int x \, \mu_t(\mathrm{d}x)$ runs over the interior of the simplex (when t runs over E). Thus,

$$\nabla \Lambda_\mu(G) = \nabla \Lambda_\mu(E) = E \cap \Big(-\frac{1}{d}, \infty \Big)^d \, .$$

The Legendre transform $\Lambda_\mu^*(x)$ for $x \in E \cap \big(-\frac{1}{d}, \infty \big)^d$ is $\langle t, x \rangle - \Lambda_\mu(t)$ when $\nabla \Lambda_\mu(t) = x$; but let us treat it as a function of the tilted measure μ_t, that is, of its probabilities $q_1, \dots, q_d$ (rather than x). We have $x_j = q_j - \frac{1}{d}$,

$$q_j = \frac{1}{M_\mu(t)} \mathrm{e}^{t_j} p_j \, ; \quad \ln \frac{q_j}{p_j} = t_j - \Lambda_\mu(t) \, ;$$

$$\langle t, x \rangle = \sum_j t_j \Big(q_j - \frac{1}{d} \Big) = \sum_j t_j q_j = \sum_j q_j \ln \frac{q_j}{p_j} + \Lambda_\mu(t) \, ; \; \langle t, x \rangle - \Lambda_\mu(t) = \sum_j q_j \ln \frac{q_j}{p_j} \, ;$$

thus,

$$\Lambda_\mu^*(x) = q_1 \ln \frac{q_1}{p_1} + \dots + q_d \ln \frac{q_d}{p_d} = \int \ln \frac{\mathrm{d}\mu_t}{\mathrm{d}\mu} \, \mathrm{d}\mu_t = H(\mu_t | \mu) \, ; \tag{3c1}$$

the relative entropy again!

Ignoring the distinction between a probability measure ν on the vertices $e_j - \frac{e_1 + \dots + e_d}{d}$ of the (closed) simplex $K = E \cap [-\frac{1}{d}, \infty)^d$ and the point $\int x \, \nu(\mathrm{d}x) \in K$ of this simplex, we write

$$\Lambda_\mu^*(\nu) = H(\nu | \mu) \quad \text{for all } \mu \in K \, .$$

This equality holds on the interior of K, and therefore on the whole K, since $H(\cdot | \mu)$ is continuous on K ($0 \ln 0 = 0$, of course), and $\Lambda_\mu^*(\cdot)$ is convex and lower semicontinuous (being the supremum of affine functions), therefore continuous on K, too. Prop. 3a6 gives an upper bound for the empirical distribution η:

$$\ln \mu^{*n}(\{ n\eta : H(\eta | \mu) \ge c \}) \le -nc + o(n) \quad \text{as } n \to \infty \tag{3c2}$$

for every $c \in (0, \infty)$. On the other hand, Theorem 3a1 gives a lower bound:

$$\ln \mu^{*n}(nU) \ge -n \inf_{\nu \in U \cap K} H(\nu|\mu) + o(n) \quad \text{as } n \to \infty \tag{3c3}$$

for every open set U.

3c4 Exercise (minimum of relative entropy on half-space)**.** Let μ be a strictly positive probability measure on a finite set Ω, and $u : \Omega \to \mathbb{R}$. Then $H(\nu|\mu) - H(\mu_u|\mu) \ge \int u \, \mathrm{d}\nu - \int u \, \mathrm{d}\mu_u$ for all probability measures ν on Ω; and therefore

$$\min_{\nu : \int u \, \mathrm{d}\nu \ge \int u \, \mathrm{d}\mu_u} H(\nu|\mu) = H(\mu_u|\mu) \, .$$

Prove it.[1]

3d Conditioning and tilting

Recall the essential infimum and supremum of a measurable function u on a probability space (Ω, μ):

$$\operatorname{ess\,inf} u = \sup\{x : \mu(u^{-1}(-\infty, x)) = 0\} = \inf\{x : \mu(u^{-1}(-\infty, x)) > 0\} \in [-\infty, +\infty) \, ;$$
$$\operatorname{ess\,sup} u = \inf\{x : \mu(u^{-1}(x, \infty)) = 0\} = \sup\{x : \mu(u^{-1}(x, \infty)) > 0\} \in (-\infty, +\infty]$$

(here $\sup \emptyset = -\infty$, $\inf \emptyset = +\infty$).

3d1 Exercise. (a) $\lim_{t \to -\infty} \int u \, \mathrm{d}\mu_{tu} = \operatorname{ess\,inf} u$, $\lim_{t \to +\infty} \int u \, \mathrm{d}\mu_{tu} = \operatorname{ess\,sup} u$.

(b) If $\operatorname{ess\,inf} u < \operatorname{ess\,sup} u$, then the function $t \mapsto \int u \, \mathrm{d}\mu_{tu}$ is strictly increasing.

Prove it.[2]

Now we turn to measures and functions on a finite set; let it be $\{1, \dots, d\}$.

For every $\omega \in \{1, \dots, d\}^n$ we define frequencies $\eta_{n,\omega}(x) = \frac{1}{n} \cdot \#\{k : \omega_k = x\}$ for $x \in \{1, \dots, d\}$; these frequencies are a probability measure $\eta_{n,\omega}$ on $\{1, \dots, d\}$. Thus, η_n maps $\{1, \dots, d\}^n$ to the set of probability measures on $\{1, \dots, d\}$.

Let μ be a strictly positive probability measure on $\{1, \dots, d\}$; we endow $\{1, \dots, d\}^n$ with the probability measure μ^n, and treat η_n as a *random* probability measure on $\{1, \dots, d\}$.

Let $u : \{1, \dots, d\} \to \mathbb{R}$, and $c \in \mathbb{R}$, $\int u \, \mathrm{d}\mu < c < \max\bigl(u(1), \dots, u(d)\bigr)$. We consider events $E_n = \{\int u \, \mathrm{d}\eta_n \ge c\}$; that is, $E_n \subset \{1, \dots, d\}^n$, $E_n = \{\omega : \int u \, \mathrm{d}\eta_{n,\omega} \ge c\}$.

[1] Hint: use 3a7.

[2] Hint: the general case reduces to a measure on $\mathbb{R}$ and the function $u(x) = x$, as noted in Sect. 2a.

3d2 Proposition. Conditionally, given E_n, the random measure η_n converges (as $n \to \infty$) in probability to the tilted measure μ_{tu} where $t > 0$ is such that $\int u \, \mathrm{d}\mu_{tu} = c$.

Why just this measure, μ_{tu}? Because (as we'll see soon) this is the unique minimizer of the relative entropy $H(\nu|\mu)$ among all measures ν such that $\int u \, \mathrm{d}\nu \ge c$. In particular, when μ is the uniform measure, μ_{tu} is the unique maximizer of the Shannon entropy $H(\nu)$ (recall (3b3)).[1]

Proof. In the compact simplex K of all probability measures on $\{1, \dots, d\}$ we consider the closed set $F = \{\nu : \int u \, \mathrm{d}\nu \ge c\}$, the open set $G = \{\nu : \int u \, \mathrm{d}\nu > c\}$, note that $F = \overline{G}$, and $\mu_{tu} \in F \setminus G$. Let U be a neighborhood of μ_{tu}. The lower bound (3c3) gives

$$\ln \mu^n \{\eta_n \in G \cap U\} \ge -n \inf_{\nu \in G \cap U} H(\nu|\mu) + o(n) \ge -nH(\mu_{tu}|\mu) + o(n) \, ,$$

since μ_{tu} belongs to the closure of $G \cap U$. The upper bound (3c2) gives

$$\ln \mu^n (\{H(\eta_n|\mu) > H(\mu_{tu}|\mu) + \varepsilon\}) \le -n(H(\mu_{tu}|\mu) + \varepsilon) + o(n)$$

for all $\varepsilon > 0$. Thus, the conditional probability of the event $H(\eta_n|\mu) > H(\mu_{tu}|\mu) + \varepsilon$ given E_n (that is, $\eta_n \in F$) tends to 0. We'll prove that

$$F \setminus U \subset \{\nu : H(\nu|\mu) > H(\mu_{tu}|\mu) + \varepsilon\}$$

for some $\varepsilon > 0$. By compactness, it is sufficient to prove that $H(\nu|\mu) > H(\mu_{tu}|\mu)$ for all $\nu \in F \setminus \{\mu_{tu}\}$. The weaker inequality $H(\nu|\mu) \ge H(\mu_{tu}|\mu)$ follows from 3c4. It remains to note that the function $H(\cdot|\mu)$ on K is strictly convex, since the function $x \mapsto x \ln x$ on $[0, 1]$ is strictly convex. □

Of course, the set $\{1, \dots, d\}$ may be replaced with any finite set.

3d3 Example. Continuing an example of Sect. 2a, we take μ uniform on $\{-1, 0, 1\}$, $u(x) = x$, and $c = 3/7$. Then μ_{tu} is $\bigl(\frac{1}{a^2+a+1}, \frac{a}{a^2+a+1}, \frac{a^2}{a^2+a+1}\bigr)$ where $a = \mathrm{e}^t$; and $\int u \, \mathrm{d}\mu_{tu} = \frac{a^2-1}{a^2+a+1}$ is 3/7 for $a = 2$.

Clearly, the conditioning on $\eta_n \in F$ may be replaced with conditioning on $\eta_n \in A$ whenever $A \subset F$ and $A \supset F \cap U$ for some neighborhood U of μ_{tu}. In particular, we may condition on $c \le \int u \, \mathrm{d}\nu \le c + \varepsilon$ for arbitrary $\varepsilon > 0$.

Treating $\omega = (x_1, \dots, x_n) \in \{1, \dots, d\}^n$ as a configuration of a physical system of n particles (as in Sections 1b and 2a), and $n \int u \, \mathrm{d}\eta_{n,\omega} = u(x_1) +$

[1] "MaxEnt" (maximization of entropy) is used widely; just see "Maximum entropy" in Wikipedia. Regretfully, the strange sign of the relative entropy leads to "MinRelEnt" (??)

$\dots + u(x_n)$ as the energy[1] $H(\omega)$ of the configuration, and taking μ to be the uniform measure, we observe that most of configurations ω satisfying

$$c \le \frac{1}{n} H(\omega) \le c + \varepsilon \quad \text{(that is, } c \le \textstyle\int u \, \mathrm{d}\eta_{n,\omega} \le c + \varepsilon\text{)}$$

satisfy

$$\eta_{n,\omega} \approx \mu_{tu} \,,$$

and μ_{tu} maximizes the (Shannon) entropy among all measures ν such that $c \le \int u \, \mathrm{d}\nu \le c + \varepsilon$.

In statistical physics, usually, a system performs[2] a kind of random walk on the set $\{\omega : H(\omega) \approx c\}$. It spends most of the time in the subset $\{\omega : \eta_{n,\omega} \approx \mu_{tu}\}$; and no wonder!

Index

[1] Unfortunately, H is widely used for both the Hamiltonian and the (Shannon) entropy. In physics, entropy is usually denoted S.

[2] Since H is the Hamiltonian of an isolated system, but there is also a weak interaction with environment.