10 Cramèr theorem via Sanov's theorem

10a Sanov's theorem in general

Recall Sanov's theorem 3b4 for a finite probability space (the multinomial LDP). We want to generalize it to arbitrary (not just discrete) probability spaces.

First, consider the space of (infinite) sequences $\Omega = \{0,1\}^\infty$ and a probability measure p on Ω. Its marginals are a consistent family of probability measures on $\{0,1\}^j$.[1]

Given n points $\omega_1, \dots, \omega_n \in \Omega$, we consider the empirical measure $\frac{1}{n}\sum \delta_{\omega_k} \in P(\Omega)$,

$$\Big(\frac{1}{n} \sum_{k=1}^n \delta_{\omega_k} \Big)(A) = \frac{\#\{k : \omega_k \in A\}}{n} \, .$$

Treating $\omega_1, \dots, \omega_n$ as independent, distributed p each, we get a distribution $\mu_n \in P(P(\Omega))$ of the empirical measure,

$$\int f \, \mathrm{d}\mu_n = \int_\Omega \dots \int_\Omega f \Big(\frac{1}{n} \sum_{k=1}^n \delta_{\omega_k} \Big) \, p(\mathrm{d}\omega_1) \dots p(\mathrm{d}\omega_n) \, .$$

Indeed, Ω is a compact metrizable space (with the product topology),[2] thus, $P(\Omega)$ is also a compact metrizable space, and $P(P(\Omega))$ is well-defined.

In order to use the Dawson-Gärtner theorem 5b2 (or rather its generalization 5b4) we consider the projection (truncation) maps $\Omega \to \{0,1\}^j$ and the corresponding maps $g_j : P(\Omega) \to P(\{0,1\}^j)$. They separate points of $P(\Omega)$ (think, why). The spaces $P(\{0,1\}^j)$ depend on j, but 5b4 generalizes readily to such a situation. The image $\nu_n^{(j)}$ of the measure μ_n under g_j is the multinomial distribution (think, why) governed by the measure $p^{(j)} = g_j(p)$. By Theorem 3b4, the sequence $(\nu_n^{(j)})_n$ satisfies LDP

[1] In fact, every such family corresponds to one and only one p (Kolmogorov's theorem).
[2] Homeomorphic to the Cantor set.

with the rate function $x \mapsto H(x|p^{(j)})$ for $x \in P(\{0,1\}^j)$. By the Dawson-Gärtner theorem, the sequence $(\mu_n)_n$ satisfies LDP with the rate function $I(x) = \sup_j H\big(g_j(x)\big|g_j(p)\big)$. However, $\sup_j H\big(g_j(x)\big|g_j(p)\big) = H(x|p)$, the relative entropy $H(x|p)$ being defined by

$$H(x|p) = \int \Big(\ln \frac{\mathrm{d}x}{\mathrm{d}p}\Big)\,\mathrm{d}x = \int \Big(\frac{\mathrm{d}x}{\mathrm{d}p} \ln \frac{\mathrm{d}x}{\mathrm{d}p}\Big)\,\mathrm{d}p$$

if x is absolutely continuous w.r.t. p, otherwise $H(x|p) = \infty$. Finally,

$$I(x) = H(x|p)\,.$$

10a1 Exercise. Let p be a probability measure on $[0,1]$. Consider the distribution $\mu_n \in P(P[0,1])$ of the empirical measure $\frac1n \sum \delta_{\omega_k}$, where $\omega_1, \dots, \omega_n$ are drawn from p independently. Then $(\mu_n)_n$ satisfies LDP with the rate function $x \mapsto H(x|p)$ for $x \in P[0,1]$.

Prove it.

Hint: map $\{0,1\}^\infty$ onto $[0,1]$ using binary digits.

The same can be done for any probability measure on $\mathbb{R}^d$, and in fact, on any Polish space.

10b Cramèr theorem on a bounded interval

Every measure $p \in P[0,1]$ has its barycenter $F(p) = \int u\,p(\mathrm{d}u) \in [0,1]$. The map $F : P[0,1] \to [0,1]$ is continuous (think, why).

Given $p \in P[0,1]$ and n, we consider the corresponding distribution $\mu_n \in P(P[0,1])$ of the empirical measure $\frac1n \sum \delta_{\omega_k} \in P[0,1]$. The image $\nu_n \in P[0,1]$ of μ_n under F is nothing but the distribution of $(X_1 + \dots + X_n)/n$ where $X_1, \dots, X_n$ are independent random variables distributed p each. Indeed, $F\big(\frac1n \sum \delta_{\omega_k}\big) = (\omega_1 + \dots + \omega_n)/n$.

Combining Sanov's theorem 10a1 with the contraction principle 2b1 we conclude that (a) the sequence $(\nu_n)_n$ is LD-convergent, and (b) $(\nu_n)_n$ satisfies LDP with the rate function

$$I(y) = \min\{H(x|p) : x \in P[0,1], F(x) = y\}\,.$$

The case $y = F(p)$ is evident; here $I(y) = 0$ since $H(p|p) = 0$.

Given $\lambda \in \mathbb{R}$, we define $p_\lambda \in P[0,1]$ by

$$\frac{\mathrm{d}p_\lambda}{\mathrm{d}p}(u) = \mathrm{const}_\lambda \cdot \mathrm{e}^{\lambda u}\,, \quad \mathrm{const}_\lambda = \frac{1}{\int \mathrm{e}^{\lambda u}\,p(\mathrm{d}u)}$$

and note that

$$H(x|p_\lambda) = H(x|p) - \lambda F(x) - \ln \mathrm{const}_\lambda$$

(think, why). It means that $\min\{H(x|p_\lambda) : F(x) = y\} = \min\{H(x|p) : F(x) = y\} - \lambda y - \ln \mathrm{const}_\lambda$. The case $y = F(p_\lambda)$ is thus solved; here $\min\{H(x|p_\lambda) : F(x) = y\} = 0$, therefore $\min\{H(x|p) : F(x) = y\} = \lambda y + \ln \mathrm{const}_\lambda$, that is,

$$I(F(p_\lambda)) = \lambda F(p_\lambda) - \ln \int \mathrm{e}^{\lambda u} \, p(\mathrm{d}u) \, ;$$

here

$$F(p_\lambda) = \frac{\int u \mathrm{e}^{\lambda u} \, p(\mathrm{d}u)}{\int \mathrm{e}^{\lambda u} \, p(\mathrm{d}u)} \, .$$

The same holds for every compactly supported probability measure p on $\mathbb{R}$ (not necessarily concentrated on $[0, 1]$).

Usually one introduces the *logarithmic moment generating function*

$$\Lambda_p(\lambda) = \ln \int \mathrm{e}^{\lambda u} \, p(\mathrm{d}u)$$

(convex by Hölder's inequality) and its Legendre(-Fenchel) transform

$$\Lambda_p^*(u) = \sup_{\lambda \in \mathbb{R}} \big(\lambda u - \Lambda_p(\lambda) \big) \, .$$

Then $\mathrm{const}_\lambda = \exp(-\Lambda_p(\lambda))$ and $F(p_\lambda) = \Lambda_p'(\lambda)$ (think, why). If $u = \Lambda_p'(\lambda)$ then $\Lambda_p^*(u) = \lambda u - \Lambda_p(\lambda) = \lambda F(p_\lambda) - \Lambda_p(\lambda) = I(u)$. We see that $I(u) = \Lambda_p^*(u)$ at least for all u of the form $\Lambda_p'(\cdot)$.

Let $[a, b]$ be the smallest segment containing the support of p. Then $\Lambda_p'(-\infty) = a$ and $\Lambda_p'(+\infty) = b$. It follows that every $u \in (a, b)$ is of the form $\Lambda_p'(\cdot)$. Thus, $I(\cdot) = \Lambda_p^*(\cdot)$ on (a, b).

See also 3c (especially (3c3)).

10c A strengthened Sanov's theorem

The space $P(\mathbb{R})$ of all probability measures on $\mathbb{R}$ is endowed with the weak convergence (compare it with 2a) defined by

(10c1) $$\mu_n \to \mu \quad \Longleftrightarrow \quad \forall f \in C_{\mathrm{b}}(\mathbb{R}) \;\; \int f \, \mathrm{d}\mu_n \to \int f \, \mathrm{d}\mu$$

for $\mu, \mu_n \in P(\mathbb{R})$. It is equivalent to $\mathrm{dist}(\mu_n, \mu) \to 0$, where 'dist' is the Lévy-Prokhorov metric,

(10c2) $$\mathrm{dist}(\mu, \nu) = \inf\{\varepsilon > 0 : \forall F \;\; \mu(F) \le \nu(F_{+\varepsilon}) + \varepsilon, \, \nu(F) \le \mu(F_{+\varepsilon}) + \varepsilon\} \, ;$$

here F runs over all closed subsets of $\mathbb{R}$, or equivalently, over the rays $(-\infty, u]$, $u \in \mathbb{R}$. This metric turns $P(\mathbb{R})$ into a Polish space.

However, the barycenter $\int u\, x(\mathrm{d}u)$ cannot be treated as a continuous function of $x \in P(\mathbb{R})$ (think, why). This time we cannot combine Sanov's theorem with the contraction principle just as we did in 10b. We need the inverse contraction principle 9e1.

The space $\mathrm{M}_+(\mathbb{R})$ of all finite positive measures on $\mathbb{R}$ is also a Polish space; (10c1) and (10c2) still work. We consider the closed subset

$$\mathcal{X}_1 = \Big\{ x \in \mathrm{M}_+(\mathbb{R}) : \int \frac{x(\mathrm{d}u)}{|u|+1} = 1 \Big\}$$

and the map

$$F : \mathcal{X}_1 \to \mathcal{X}_2 = P(\mathbb{R}) \, , \quad F(x) = y \quad \Longleftrightarrow \quad \frac{\mathrm{d}y}{\mathrm{d}x}(u) = \frac{1}{|u|+1} \, ;$$

the map F is continuous and one-to-one (think, why). The barycenter of the measure $y = F(x)$ is a well-defined, continuous function of x (think, why).

Given n points $u_1, \dots, u_n \in \mathbb{R}$, we represent the empirical measure $\frac1n \sum \delta_{u_k} \in P(\mathbb{R})$ as follows:

$$\frac1n \sum_{k=1}^n \delta_{u_k} = F\Big(\frac1n \sum_{k=1}^n (|u_k|+1) \delta_{u_k} \Big) \, .$$

Given $p \in P(\mathbb{R})$, we denote by μ_n the distribution of $\frac1n \sum (|u_k|+1)\delta_{u_k}$ and by ν_n the distribution of $\frac1n \sum \delta_{u_k}$; here $u_1, \dots, u_n$ are independent, distributed p each. Thus, $\mu_n \in P(\mathcal{X}_1)$, $\nu_n \in P(\mathcal{X}_2)$, and ν_n is the image of μ_n under F.

From now on we assume that the distribution p has all exponential moments, that is,

$$\text{(10c3)} \qquad \int \mathrm{e}^{\mathrm{i}\lambda u} \, p(\mathrm{d}u) < \infty \quad \text{for all } \lambda \in \mathbb{R} \, .$$

10c4 Lemma. The sequence $(\mu_n)_n$ is exponentially tight.

Proof. (sketch) It is sufficient to prove that for every $\varepsilon > 0$ there exists $C < \infty$ such that for all n, $\mu_n(K_{+\varepsilon}) \ge 1 - \varepsilon^n$, where

$$K = \mathcal{X}_1 \cap \mathrm{M}_+[-C, C] = \{ x \in \mathcal{X}_1 : x(\mathbb{R} \setminus [-C,C]) = 0 \} \, .$$

If $x(\mathbb{R} \setminus [-C,C]) < \varepsilon$ then $x \in K_{+\varepsilon}$ (think, why); we need

$$\mathbb{P} \Big(\frac1n \sum_{k:|u_k|>C} (|u_k|+1) \ge \varepsilon \Big) \le \varepsilon^n \, .$$

We have

$$\mathbb{P}\Big(\sum_{k:|u_k|>C} (|u_k|+1) \ge n\varepsilon \Big) = \mathbb{P}\Big(\sum_k (|u_k|+1)\mathbf{1}_{(C,\infty)}(|u_k|) \ge n\varepsilon \Big) \le$$
$$\le \frac{\mathbb{E} \exp \lambda \sum_k (|u_k|+1)\mathbf{1}_{(C,\infty)}(|u_k|)}{\exp(\lambda n \varepsilon)} =$$
$$= \Big(\mathrm{e}^{-\lambda\varepsilon} \int \exp \lambda(|u|+1)\mathbf{1}_{(C,\infty)}(|u|)\, p(\mathrm{d}u) \Big)^n$$

for every $\lambda > 0$. However, $\int \exp \lambda(|u|+1)\mathbf{1}_{(C,\infty)}(|u|)\, p(\mathrm{d}u) \to 1$ as $C \to \infty$. We choose λ such that $2\mathrm{e}^{-\lambda\varepsilon} \le \varepsilon$ and then C such that $\int(\dots) \le 2$. □

By Sanov's theorem, $(\nu_n)_n$ satisfies LDP with the rate function $y \mapsto H(y|p)$. By the inverse contraction principle (Theorem 9e1) and Lemma 10c4, $(\mu_n)_n$ satisfies LDP with the rate function $x \mapsto H(F(x)|p)$ (assuming (10c3)). This is the strengthened Sanov's theorem.

10d Cramèr theorem on the line

Let $p \in P(\mathbb{R})$ satisfy (10c3), and ν_n be the distribution of $(X_1 + \dots + X_n)/n$ where $X_1, \dots, X_n$ are independent random variables distributed p each.

Applying the contraction principle (Theorem 9b1) to the continuous map $\mathcal{X}_1 \to \mathbb{R}$, $x \mapsto \operatorname{barycenter}(F(x))$ we conclude that $(\nu_n)_n$ satisfies LDP with the rate function

$$I(u) = \min\{ H(F(x)|p) : x \in \mathcal{X}_1, \operatorname{barycenter}(F(x)) = u \} =$$
$$= \min\{ H(y|p) : y \in F(\mathcal{X}_1), \operatorname{barycenter}(y) = u \} .$$

Note that

$$y \in F(\mathcal{X}_1) \quad \Longleftrightarrow \quad \int |u|\, y(\mathrm{d}u) < \infty .$$

We proceed similarly to 10b. The case $u = \operatorname{barycenter}(p)$ is evident; here $I(u) = 0$, since $H(p|p) = 0$ and $p \in F(\mathcal{X}_1)$.

Given $\lambda \in \mathbb{R}$, we define $p_\lambda \in P(\mathbb{R})$ by

$$\frac{\mathrm{d}p_\lambda}{\mathrm{d}p}(u) = \mathrm{const}_\lambda \cdot \mathrm{e}^{\lambda u}, \quad \mathrm{const}_\lambda = \frac{1}{\int \mathrm{e}^{\lambda u}\, p(\mathrm{d}u)}$$

and note that $p_\lambda \in F(\mathcal{X}_1)$ and

$$H(y|p_\lambda) = H(y|p) - \lambda \cdot \operatorname{barycenter}(y) - \ln \mathrm{const}_\lambda .$$

The rest is exactly as in 10b.

10d1 Theorem. (Cramèr) Let $X_1, X_2, \dots$ be i.i.d. random variables, and

$$\Lambda(\lambda) = \ln \mathbb{E}\, \mathrm{e}^{\lambda X_1} < \infty \quad \text{for } \lambda \in \mathbb{R} \, .$$

Then the sequence (of distributions) of random variables $(X_1 + \dots + X_n)/n$ satisfies LDP with the rate function Λ^*,

$$\Lambda^*(u) = \sup_{\lambda \in \mathbb{R}} \big(\lambda u - \Lambda(\lambda) \big) \, .$$

Many generalizations, and various proofs, are well-known. In fact, the statement remains true if the set $\{\lambda : \Lambda(\lambda) < \infty\}$ is a neighborhood of 0 (and even only $\{0\}$) rather than the whole $\mathbb{R}$. Can it still be proved via Sanov's theorem? I do not know.

See also [1, Sect. 2.2].

Finally, I formulate (without proof) a related result.

10d2 Theorem. (Gärtner) Let $\mu_1, \mu_2, \dots$ be probability distributions on $\mathbb{R}$ such that the limit

$$c(\lambda) = \lim_n \frac{1}{n} \ln \int \mathrm{e}^{n\lambda u} \mu_n(\mathrm{d}u) \in \mathbb{R}$$

exists for every $\lambda \in \mathbb{R}$, and the function $c(\cdot)$ is differentiable on $\mathbb{R}$. Then $(\mu_n)_n$ satisfies LDP with the rate function

$$I(u) = \sup_{\lambda \in \mathbb{R}} \big(\lambda u - c(\lambda) \big) \, .$$

The condition of (finiteness and) differentiability can be weakened considerably. See the Gärtner-Ellis theorem in [2, Sect. 8], [1, Sect. 2.3].

References

[1] A. Dembo, O. Zeitouni, *Large deviations techniques and applications,* Jones and Bartlett publ., 1993.

[2] R.S. Ellis, *The theory of large deviations and applications to statistical mechanics,* 2006, http://www.math.umass.edu/~rsellis/pdf-files/Dresden-lectures.pdf