1 The form of entropy-extremal distributions

Let $A$ be a finite set, let $p \in \text{Prob}(A)$, and let $E \subseteq \text{Prob}(A)$ be such that $E^c \supseteq E$. Let $X \sim p^\times n$. Then Sanov’s theorem gives that

$$P(p_X \in E) = e^{-dn + o(n)},$$

where

$$d := \inf_{q \in E} D(q \| p).$$

If $E$ is closed and convex, then the infimum $d$ is actually a minimum, achieved at a unique point $q^* \in E$. In that case we also have the conditional limit theorem, according to which

$$P(\|p_X - q^*\| < \varepsilon \mid p_X \in E) \longrightarrow 1$$
as $n \longrightarrow \infty$ for this special distribution $q$.

For some applications of this theorem, it can be important to find the minimizer $q^*$ explicitly. The most common setting is the following. Suppose that $f_1, \ldots, f_k : A \longrightarrow \mathbb{R}$ are any real-valued functions on $A$, and let $K \subseteq \mathbb{R}^k$. Then for $E$ we may take the set

$$E := \{q \in \text{Prob}(A) : \left( \int f_1 \, dq, \ldots, \int f_k \, dq \right) \in K\}$$

For instance, if $E$ is a translated quadrant $\prod_{j=1}^k [a_j, \infty)$ for some choice of thresholds $\alpha_1, \ldots, \alpha_k \in \mathbb{R}$, then

$$E := \{q \in \text{Prob}(A) : \int f_j \, dq \geq \alpha_j \ \forall j = 1, 2, \ldots, k\}. $$


On the other hand, $K$ may be a single point in $\mathbb{R}^k$, and then our set $E$ simply specifies the expectation of each of the functions $f_j$ under $q$.

In these settings, the problem of finding $q^*$ is solved by the next theorem. To formulate it, fix $f_1, \ldots, f_k$ and $\alpha_1, \ldots, \alpha_k$ as above and also $p \in \text{Prob}(A)$. For a vector of parameters $\lambda = (\lambda_1, \ldots, \lambda_k)$, let $q_\lambda \in \text{Prob}(A)$ be the distribution

$$q_\lambda(a) = \frac{1}{M(\lambda)} p(a) e^{\sum_j \lambda_j f_j(a)},$$

(1)

where

$$M(\lambda) = \sum_b p(b) e^{\sum_j \lambda_j f_j(b)}$$

is the necessary normalizing constant so that $q_\lambda$ is a probability. The function $M(\lambda)$ is actually the multivariate moment generating function of $(f_1, \ldots, f_k)$, regarded as a tuple of RVs on the probability space $(A, p)$.

**Theorem 1** (Divergence-minimizing distribution). Assume there is a $\lambda$ so that

$$\int f_j \, dq_\lambda = \alpha_j \quad \text{for all } j = 1, \ldots, k.$$  

(2)

Then $q_\lambda$ is the unique minimizer of $D(\cdot \parallel p)$ among members of $\text{Prob}(A)$ which satisfy these equations, and the minimum value is given by

$$D(q_\lambda \parallel p) = \sum_j \lambda_j \alpha_j - \log M(\lambda).$$

**Proof.** Let $q$ be another distribution which satisfies (2). We need to show that $D(q \parallel p) \geq D(q_\lambda \parallel p)$, with equality only if $q = q_\lambda$.

If $D(q \parallel p) = \infty$ then there is nothing to prove, so suppose it is finite. Then $q \ll p$. From (1) we see directly that $p \ll q_\lambda$, and so in fact $q \ll q_\lambda$. Let $\varphi$ be the Radon–Nikodym derivative, so $q(a) = \varphi(a) q_\lambda(a)$.

Now we simply compute as follows:

$$D(q \parallel p) = \sum_a q(a) \log \frac{q(a)}{p(a)}$$

$$= \sum_a q(a) \log \left[ \varphi(a) \frac{q_\lambda(a)}{p(a)} \right]$$

$$= \sum_a q(a) \log \varphi(a) + \sum_a q(a) \log \frac{q_\lambda(a)}{p(a)}$$

$$= \sum_a q(a) \log \frac{1}{M(\lambda)} + \sum_a q(a) \log \frac{\sum_j \lambda_j f_j(a)}{M(\lambda)}$$

$$= \log M(\lambda) + \sum_a q_\lambda(a) \log \frac{\sum_j \lambda_j f_j(a)}{M(\lambda)}.$$
where the first term is recognized from the definition of divergence and the second is evaluated by substituting from (1).

Since divergence is always non-negative, the expression above is at least the second term, and that term alone equals

$$ \sum_a q(a) \sum_j \lambda_j f_j(a) - \sum_a q(a) \log M(\lambda) = \sum_j \lambda_j \alpha_j - \log M(\lambda), $$

by the assumption that $q$ satisfies (2). This expression is independent of $q$. If $q = q_\lambda$, then $D(q \parallel q_\lambda) = 0$, and we obtain that $D(q_\lambda \parallel p)$ is equal to the last line above, as asserted in the theorem. Otherwise, $D(q \parallel p)$ exceeds that value by $D(q \parallel q_\lambda) > 0$.

**Remark.** The idea to consider these particular distributions $q_\lambda$ can be motivated by a careful use of Lagrange multipliers, and these can also be used to prove this theorem.

**Warning:** The equality $D(q \parallel p) = D(q \parallel q^*) + D(q^* \parallel p)$ that we obtain above holds only because of (a) the special form of $q^*$ and (b) the assumption that $q$ satisfies (2). This equality is certainly not true in general. See [CT06, Theorem 11.6.1] for some less restrictive conditions which at least give an inequality.

For Theorem 1 to be useful, we need to know that a suitable parameter vector $\lambda$ exists. We give the relevant result in case $k = 1$; the higher-dimensional case is similar but more complicated. On the other, we do now enlarge our setting to non-discrete probability spaces, for the sake of future reference.

Thus, let $(X, \mathcal{B}, \mu)$ be a probability space, and let $f : X \longrightarrow \mathbb{R}$ be an essentially bounded RV\(^1\). Assume that $f$ is not $\mu$-almost surely constant, and let $[a, b]$ be the smallest closed interval satisfying $\mu\{a \leq f \leq b\} = 1$.

For each $\lambda \in \mathbb{R}$, define a new probability measure on $(X, \mathcal{B})$ by

$$ \nu_\lambda(A) = \frac{1}{M(\lambda)} \int_A e^{\lambda f(x)} \mu(dx) \quad \text{for } A \in \mathcal{B}, $$

where

$$ M(\lambda) = \int e^{\lambda f} d\mu. $$

**Lemma 2.** The moment generating function $M$ is strictly log-convex.

\(^1\)This assumption can be weakened considerably, but we leave the necessary extra work aside.
\textit{Proof.} If $\lambda_1, \lambda_2 \in \mathbb{R}$ and $0 < t < 1$, then Hölder’s inequality gives
\[
M(t\lambda_1 + (1-t)\lambda_2) = \int e^{t\lambda_1 f} e^{(1-t)\lambda_2 f} \, d\mu \leq \left( \int e^{\lambda_1 f} \, d\mu \right)^t \left( \int e^{\lambda_2 f} \, d\mu \right)^{1-t},
\]
and so $\log M(t\lambda_1 + (1-t)\lambda_2) \leq t \log M(\lambda_1) + (1-t) \log M(\lambda_2)$. Equality holds if and only if $e^{\lambda_1 f}$ and $e^{\lambda_2 f}$ are scalar multiples of the same function away from a $\mu$-negligible set, since this when Hölder’s inequality is an equality. Since $f$ is not almost surely constant, this requires that $\lambda_1 = \lambda_2$. \hfill $\square$

\textbf{Proposition 3.} In the setting above, the following are equivalent:

i) $a < \alpha < b$;

ii) there is a $\lambda^* \in \mathbb{R}$ such that $\int f \, d\nu_{\lambda^*} = \alpha$;

iii) the function
\[
\psi : \mathbb{R} \longrightarrow \mathbb{R} : \lambda \mapsto \alpha \lambda - \log M(\lambda)
\]
has a unique maximum.

In this case, the maximum in condition (iii) occurs precisely at the value $\lambda^*$ required by condition (ii).

\textit{Proof.} First, some simple analysis shows that
\[
\int f \, d\nu_{\lambda} \longrightarrow \begin{cases} 
  b & \text{as } \lambda \longrightarrow \infty \\
  a & \text{as } \lambda \longrightarrow -\infty.
\end{cases}
\]
This integral is also continuous as a function of $\lambda$, so the equivalence of (i) and (ii) follows from the intermediate value theorem.

Turning to (iii), observe that $\psi$ is smooth, and that
\[
\psi'(\lambda) = \alpha - (\log M(\lambda))' = \alpha - \frac{M'(\lambda)}{M(\lambda)} = \alpha - \frac{1}{M(\lambda)} \int f e^{\lambda f} \, d\mu = \alpha - \int f \, d\nu_{\lambda}.
\]
So $\lambda^*$ satisfies condition (ii) if and only if $\psi'(\lambda^*) = 0$. Moreover, $\psi$ is strictly concave, by the preceding lemma, and so this value $\lambda^*$ is unique and is the location of the maximum of $\psi$. \hfill $\square$

In our previous setting with $X = A$ and with $k = 1$, we obtain the following.
Corollary 4. The distribution $q_{\lambda*}$ obtained above satisfies

$$D(q_{\lambda*} \parallel p) = \sup_{\lambda \in \mathbb{R}} [\alpha \lambda - \log M(\lambda)].$$

In fact this lemma is also fairly easy to generalize to other $(X, \mathcal{B}, \mu)$.

Theorem 1 has a natural variant for entropy, rather than divergence. To formulate this, we adjust our notation by writing

$$q_{\lambda}(a) := \frac{1}{M(\lambda)} e^{\sum_j \lambda_j f_j(a)}$$

with

$$M(\lambda) = \sum_b e^{\sum_j \lambda_j f_j(b)}.$$

Theorem 5 (Entropy-maximizing distribution). Assume there is a $\lambda$ so that

$$\int f_j d q_{\lambda} = \alpha_j \text{ for all } j = 1, \ldots, k.$$

Then $q_{\lambda}$ is the unique maximizer of $H$ among members of $\text{Prob}(A)$ which satisfy these equations, and the maximum value is given by

$$H(q_{\lambda}) = \log M(\lambda) - \sum_j \lambda_j \alpha_j.$$

This can be obtained as a corollary of Theorem 1: simply recall that if $p$ is the uniform distribution on $A$ then

$$D(q \parallel p) = \log |A| - H(q).$$

A variant of Theorem 5 will play a crucial role in our study of statistical physics later.

2 Some convex analysis and Cramér’s theorem

Sanov’s theorem is one of the two results which mark the entrance to large deviations theory; the other is Cramér’s Theorem.

Let $X_1, X_2, \ldots$ be i.i.d. real-valued RVs with common distribution $\mu \in \text{Prob}(\mathbb{R})$. Here we will assume for simplicity that $X_1$ is essentially bounded: equivalently, that $\mu([-r, r]) = 1$ for some finite $r$. Let $S_n = X_1 + \cdots + X_n$.

Finally, let

$$M(\lambda) := \int_{\mathbb{R}} e^{\lambda x} \mu(dx) = E[e^{\lambda X}] \text{ for } \lambda \in \mathbb{R}$$

be the moment generating function of $X_1$. 

5
Theorem 6 (Cramér’s Theorem). Define the function

\[ I(s) := \sup_{\lambda \in \mathbb{R}} [s\lambda - \log M(\lambda)] \quad \text{for } s \in \mathbb{R} \]

(this is called the ‘Fenchel–Legendre transform’ of \( \log M \)). For any \( U \subseteq \mathbb{R} \), we have

\[ \exp \left( - \left[ \inf_{s \in U} I(s) \right] n + o(n) \right) \leq P(S_n/n \in U) \leq \exp \left( - \left[ \inf_{s \in U} I(s) \right] n + o(n) \right) \]

The similarity to Sanov’s theorem is obvious. If time allows, I will show two proofs of Theorem 6 in class: one direct, and another using Sanov’s theorem and the results of the preceding section.

The heart of the direct proof is the use of the LLN for an altered probability measure.

Most of the work in the direct proof goes into the following special case.

Proposition 7. Assume that \( \mu \) is not a point mass (so the RVs \( X_i \) are not almost surely constant), and let \([a, b]\) be the smallest closed interval with \( \mu([a, b]) = 1 \). If \( a < s < b \) and \( \delta > 0 \), then

\[ \exp \left( - I(s)n - \Delta(\delta)n + o(n) \right) \leq P(S_n/n \in (s - \delta, s + \delta)) \leq \exp \left( - I(s)n + \Delta(\delta)n + o(n) \right). \]

This can be proved in various ways. We give two proofs: one using Sanov’s theorem, and another from first principles. We only sketch the first one for brevity.

Sketch of first proof of Proposition 7. First assume that \( \mu \) is supported on a finite subset \( A \subseteq \mathbb{R} \) with least element equal to \( a \) and greatest equal to \( b \). Write \( X \) for the string \((X_1, \ldots, X_n)\). Then we have

\[ \frac{S_n}{n} = \frac{1}{n} \sum_{i=1}^{n} X_i = \frac{1}{n} \sum_{a \in A} N(a \mid X) \cdot a = \int x \, dp_X. \]

Therefore

\[ P(S_n/n \in (s - \delta, s + \delta)) = P(p_X \in K), \]

where

\[ K := \{ \nu \in \text{Prob}(A) : \int x \, d\nu \in (s - \delta, s + \delta) \}. \]
Since $a < s < b$, the set $K$ is nonempty, convex, and contained in the closure of its interior. Therefore Sanov’s theorem gives

$$P(S_n/n \in (s - \delta, s + \delta)) = \exp\left(- \min_{\nu \in K} D(\nu \| \mu)n + o(n)\right).$$

Finally,

$$\inf_{\nu \in K} D(\nu \| \mu) = \inf_{t \in (s - \delta, s + \delta)} \inf \left\{ D(\nu \| \mu) \colon \nu \in \text{Prob}(A), \int x \, d\nu = t \right\},$$

By Theorem 1, Proposition 3, and Corollary 4, for a given $t \in (s - \delta, s + \delta)$, the inner infimum here is achieved by $\nu_{\lambda^*}$ for some $\lambda^*$ which satisfies

$$D(\nu_{\lambda^*} \| \mu) = \sup_{\lambda \in \mathbb{R}} [t\lambda - \log M(\lambda)].$$

Finally, a little analysis shows that this equals $I(s) + \Delta(\delta)$ for any $t \in (s - \delta, s + \delta)$, using that $a < s < b$. Therefore also

$$\inf_{\nu \in K} D(\nu \| \mu) = I(s) + \Delta(\delta).$$

To handle the case of a general distribution $\mu$, let $\varepsilon > 0$, and let $[X_i]$ be an $\varepsilon$-discretization of $X_i$ for each $i$:

$$[X_i] = \max\{n\varepsilon : n \in \mathbb{Z}, n\varepsilon \leq X_i\}.$$ 

Let $[S]_n := [X_1] + \cdots + [X_n]$. Then

$$\{[S]_n/n \in (s - \delta + \varepsilon, s + \delta - \varepsilon)\} \subseteq \{S_n/n \in (s - \delta, s + \delta)\}$$

$$\subseteq \{S_n/n \in (s - \delta - \varepsilon, s + \delta + \varepsilon)\}.$$

Now we apply the special case of a finite-valued random variable to $[X_i]$ for each $\varepsilon$, and then let $\varepsilon \downarrow 0$ for fixed $\delta$. We leave the remaining details to the reader. \hfill \Box

**Second proof of Proposition 7.** The key idea for our second proof is as follows:

Find a new distribution $\nu$ which has a simple relationship to $\mu$ but whose mean equals $s$. 
The use of $\nu$ in this proof resembles the way we used a good choice of probability distribution in solving the original type-counting problem.

The ‘simple relationship’ that we will use is already suggested by the work of Section 1. For $\lambda \in \mathbb{R}$, define $\nu_\lambda \in \text{Prob}(\mathbb{R})$ by

$$\nu_\lambda(dx) = \frac{1}{M(\lambda)} e^{\lambda x} \mu(dx), \quad (3)$$

observing that $M(\lambda)$ is precisely the normalizing constant needed in order that $\nu_\lambda$ still be a probability measure.

Since $a < s < b$, Proposition 3 gives a unique $\lambda^* \in \mathbb{R}$ such that

$$\int x \, d\nu_{\lambda^*} = s, \quad (4)$$

and it is characterized as the location of the unique maximum of the function

$$\lambda \mapsto s \lambda - \log M(\lambda) :$$

hence, $s \lambda^* - \log M(\lambda^*) = I(s)$. Let $\nu := \nu_{\lambda^*}$.

Having found $\lambda^*$, now let $Y_1, Y_2, \ldots$ be i.i.d.$(\nu)$-RVs on a new probability space, say with measure $Q$. Let $S_n^\nu = Y_1 + \cdots + Y_n$. In view of (4), and since each $Y_i$ still takes values in $[a, b]$ almost surely, the LLN gives

$$Q(|S_n^\nu/n - s| < \delta) \longrightarrow 1 \quad \text{as} \quad n \longrightarrow \infty \quad (5)$$

for any $\delta > 0$.

Finally, observe that it is easy to translate between probabilities of events for $S_n/n$ and the corresponding events for $S_n^\nu/n$ using the form of (3): in particular,

$$Q(|S_n^\nu/n - s| < \delta) = \nu^\times \{ (y_1, \ldots, y_n) \in [a, b] : |(y_1 + \cdots + y_n)/n - s| < \delta \}
\begin{aligned}
&= \frac{1}{M(\lambda^*)^n} \int_{\{(y_1, \ldots, y_n)/n - s| < \delta\}} e^{\lambda^*(y_1 + \cdots + y_n)} \mu^\times \times \times (dy_1, \ldots, dy_n).
\end{aligned}
$$

This integral is taken over a set where the integrand may always be estimated as $e^{\lambda^* s + \Delta(\delta)n}$, and so the integral itself may be estimated as

$$e^{\lambda^* n + \Delta(\delta)n} M(\lambda^*)^n P(|S_n/n - s| < \delta) = e^{(\lambda^* s - \log M(\lambda^*) + \Delta(\delta)n)} M(\lambda^*)^n P(|S_n/n - s| < \delta)
\begin{aligned}
&= e^{I(s) + \Delta(\delta)n} P(|S_n/n - s| < \delta)
\end{aligned}$$

8
Re-arranging and using (5), we obtain
\[ P(|S_n/n - s| < \delta) = e^{(-I(s)+\Delta(\delta))n} Q(|S_n/n - s| < \delta) = (1 - o(1))e^{(-I(s)+\Delta(\delta))n}. \]

**Remark.** The upper bound in Proposition 7 has an easier proof by applying Markov’s inequality to the RV $e^{\lambda^* S_n}$. But the lower bound really needs a more substantial idea such as the switch to the new distribution $Q$.  

**Ideas for completing the proof of Theorem 6 using Proposition 7.** For the upper bound. Cover $U$ with a finite collection of sufficiently short intervals, and apply Proposition 7 to those.

For the lower bound. Take a sufficiently short open interval inside $U^o$ which is close to where $I$ takes its infimal values on that set.

### 3 The general form of large deviations principles

Sanov’s and Cramér’s theorems are both examples of large deviation principles. A general definition can be given along the following lines. Let $M$ be a complete and separable metric space. A function $I : M \rightarrow [0, \infty]$ is lower semicontinuous if $\{ p \in M : I(p) \leq z \}$ is closed for every $z \in [0, \infty)$, and $I$ is a good rate function if these sets are actually all compact.

Let $\mu_n$ be a sequence in $\text{Prob}(M)$ (the Borel probability measures on $M$). This sequences satisfies a large deviations principle or LDP with rate function $I$ if

\[ \exp \left( - \inf_{p \in E^o} I(p) \right) n + o(n) \leq \mu_n(E) \leq \exp \left( - \inf_{p \in E} I(p) \right) n + o(n) \quad \forall E \subseteq M. \]

**Examples:**

- **Sanov:** $M = \text{Prob}(A)$, $\mu_n$ is the law of $p_X$ when $X \sim p^\times n$, and $I(q) = D(q \parallel p)$.

- **Cramér:** $M = \mathbb{R}$, $\mu_n$ is the law of $S_n/n$ when $X \sim \mu^\times n$, and $I$ is as in Theorem 6.
4 Notes and remarks

Sources for this lecture: [CT06, Chapter 12], [Csi75] and [VCC81] for the form of distributions which maximize entropy, similar to the earlier sections above; [Bil95, Section 9] and [GS01, Section 5.1] for Cramér’s theorem or slight variants. See also [Var03].

Further reading:

• See, for instance, [Csi75, Csi84, Csi06, VCC81] for the higher-dimensional generalization of Proposition 3.

• [DZ10] for much more on large deviations theory, including the abstract formalism and several other examples. A shorter overview with some advanced examples can be found in [Var]. Cramér’s theorem can be found in [DZ10, see Theorems 2.1.24 and 2.2.3].

• A natural next step: Cramér’s Theorem has a fairly natural generalization to vector-valued RVs: see [DZ10, Subsection 2.2.2]. With that in hand, one can actually use it to go back and give another proof of Sanov’s theorem.

References


TIM AUSTIN

Email: tim@cims.nyu.edu
URL: cims.nyu.edu/~tim