Math 254A Lecture 9 Notes

Daniel Raban

April 16, 2021

1 Cramér's Theorem and Recovering Entropy as the Exponent

1.1 Cramér's theorem

We have a σ-finite measure space (M, λ), and a measurable map $\varphi : M \to X$, where $X = Y^*$ is a locally convex space with the weak* topology. We found that

$$\lambda^{\times n}\left(\left\{p \in M^n : \frac{1}{n}\sum_{i=1}^n \varphi(p_i) \in U\right\}\right) = e^{n \cdot s(U) + o(n)},$$

where $s(U) = \sup_{x \in U} s(x)$ for some point function s which is upper semicontinuous and concave. To study s, we have introduced Fenchel-Legendre duality:

$$s(x) = \inf_y s^*(y) - \langle y, x \rangle,$$

where

$$s^*(y) := \sup_x s(x) + \langle y, x \rangle$$

is sometimes known as the **convex conjugate** of s. Last time, we proved a formula: if $s(x) < \infty$ for all n, then

$$s^*(y) = \log \int e^{\langle y, \varphi \rangle} \, d\lambda.$$

Remark 1.1. In the proof of this integral formula, to show $(\leq)$, we showed that $s(x) + \langle y, x \rangle \leq \text{RHS}$ for all x, y. For this, given $\varepsilon > 0$, we found $U \ni x$ such that

$$\lambda^{\times n}(\{\cdots \in U\}) \leq e^{\varepsilon n + o(n)} \left(\int e^{\langle y, \varphi \rangle} \, d\lambda\right)^n.$$

This part of the proof does not require that s is finite. In fact, it gives a way to prove $s(U) < \infty$ and hence $s(x) < \infty$. So if there is some $y \in Y$ such that $\int e^{\langle y, \varphi \rangle} \, d\lambda < \infty$, then $s < \infty$ and s^* is as in the theorem. The mantra is that $s < \infty$ everywhere iff $s^* < \infty$ somewhere.

A special case is when (M, λ) is a probability space and $X = \mathbb{R}^d$. In this case, we get the following version of the theorem we proved before:

Theorem 1.1 (Cramér, 1937)**.** *Let $\xi_1, \xi_2, \dots$ are i.i.d. random vectors in $\mathbb{R}^d$. Then*

$$\mathbb{P}\left(\frac{1}{n}\sum_{i=1}^{n} \xi \in U\right) = \exp\left(n \cdot \sup_{x \in U} s(x) + o(n)\right),$$

where

$$s(x) = \inf_{y \in \mathbb{R}^d} \Lambda(y) - \langle y, x \rangle,$$

and

$$\Lambda(y) = \log M(y) = \log \mathbb{E}[e^{\langle y, \xi_1 \rangle}]$$

*is the **cumulant generating function.***

In a number of texts, our s is denoted by $-I$ (so the inf becomes a sup, etc.).

1.2 Connection to the Kullback-Leibler divergence in the case of empirical distributions

Let K be a compact metric space, λ be a finite Borel measure, $X = M(K)$ be the space of measures on K (equal to $C(K)^*$ by Riesz representation), and $\varphi(p) = \delta_p$. In this case, $\frac{1}{n}\sum_{i=1}^{n} \varphi(p_i)$ is the empirical distribution of $(p_1, \dots, p_n)$.

Theorem 1.2. *In this setting, $s(\mu) = -\infty$ unless $\mu \in P(K)$ and $\mu \ll \lambda$, and in that case,*

$$s(\mu) = -\int \frac{d\mu}{d\lambda} \log \frac{d\mu}{d\lambda}\, d\lambda.$$

We will denote the right hand side by $\widetilde{s}(\mu)$ until we have proven the theorem; that way, the proof is to show that $s = \widetilde{s}$.

Remark 1.2. Note that

$$\widetilde{s}(\mu) = \int \eta\left(\frac{d\mu}{d\lambda}\right) d\eta, \qquad \eta(t) = \begin{cases} -t \log t & t > 0 \\ 0 & t = 0. \end{cases}$$

If $|\eta(\frac{d\mu}{d\lambda})| \in L^1(\lambda)$, then $\widetilde{s}(\mu) > -\infty$. Otherwise, we set $s(\mu) := -\infty$.

Remark 1.3. Here is an alternative formula that will be useful:

$$\widetilde{s}(\mu) = -\int \log \frac{d\mu}{d\lambda}\, d\mu.$$

This formula is useful, but it is a little harder to see the natural $-\infty$ convention with this version.

Here are 2 special cases:

Example 1.1. Let K be finite with λ being counting measure. Then $\frac{d\mu}{d\lambda}(a) = \mu(\{a\})$, and so

$$\widetilde{s}(\mu) = -\sum_a \mu(\{a\}) \log \mu(\{a\}) = H(\mu)$$

is the Shannon entropy.

Example 1.2. If $\lambda(K) = 1$, then

$$-\widetilde{s}(\mu) = \begin{cases} \int \frac{d\mu}{d\lambda} \log \frac{d\mu}{d\lambda} \, d\lambda \\ +\infty & \text{in the cases described above} \end{cases}$$

is called the **Kullback-Leibler divergence**. The standard notation for this is $D(\mu \| \lambda)$.

Lemma 1.1. *If $\lambda(K) = 1$, then $D(\mu\|\lambda) \geq 0$, with equality if $\mu = \lambda$.*

Proof.

$$\begin{aligned} D(\mu\|\lambda) = \int \frac{d\mu}{d\lambda} \log \frac{d\mu}{d\lambda} \\ &= \int -\eta\left(\frac{d\mu}{d\lambda}\right) \, d\lambda \end{aligned}$$

$-\eta$ is strictly concave, so using Jensen's inequality gives

$$\begin{aligned} &-\eta\left(\int \frac{d\mu}{d\lambda} \, d\lambda\right) \\ &= -\eta(1) \\ &= 1 \log 1 \\ &= 0. \end{aligned}$$

We get equality iff $\frac{d\mu}{d\lambda}$ is constant for λ-a.e., that is, iff $\mu = \lambda$. □

Let's prove the theorem:

Proof. We want to prove that $s = \widetilde{s}$. Using the expresion for s in terms of the Fenchel-Legendre transform and using the integral formula, we want to show that

$$\inf\left\{\log \int e^{f(p)} \, d\lambda(p) - \langle f, \mu \rangle : f \in C(K)\right\} = \widetilde{s}(\mu).$$

This is known as **Gibbs' variational formula**.

($\geq$): We want

$$\log \int e^f \, d\lambda - \langle f, \mu \rangle \geq -\int \frac{d\mu}{d\lambda} \log \frac{d\mu}{d\lambda} \, d\lambda.$$

The key object is

$$d\mu_f(p) = \frac{e^{f(p)}}{Z(f)}\, d\lambda(p), \qquad Z(f) = \int e^f \, d\lambda,$$

which is sometimes called the **Gibbs measure** of f with respect to λ. Observe that $\lambda \ll \mu_f$ and $\mu_f \ll \lambda$, so if $\mu \ll \lambda$, then $\mu \ll \mu_f$, then $\frac{d\mu}{d\lambda} = \frac{d\mu}{d\mu_f}\frac{d\mu_f}{d\lambda}$, and so

$$\begin{aligned}
\widetilde{s}(\mu) &= -\int \log \frac{d\mu}{d\lambda}\, d\mu \\
&= -\int \log \frac{d\mu}{d\mu_f}\, d\mu - \int \log \frac{d\mu_f}{d\lambda}\, d\mu \\
&= -D(\mu\|\mu_f) - \int (f - \log Z)\, d\mu \\
&= -D(\mu\|\mu_f) + \{\log Z - \langle f, \mu\rangle\}.
\end{aligned}$$

Rearrange this to get

$$\log Z - \langle f, \mu\rangle = \widetilde{s}(\mu) + D(\mu\|\mu_f) \geq \widetilde{s}(\mu),$$

with equality iff $\mu = \mu_f$.

($\leq$): We already know this if $\mu = \mu_f$ for some $f \in C(K)$. The summary of the rest of the proof is "such measures μ_f are dense as f varies." In more detail:

(a) $\inf\{\log \int e^f \, d\lambda - \langle f, \mu\rangle : f \in C(K)\}$ has the same value if we enlarge $C(K)$ to $B(K)$, the bounded Borel functions. This is because given λ and μ, $C(K)$ is dense in $L^1(\lambda + \mu)$, so for all $g \in B(K)$ (all uniformly bounded), there is some $(f_n)_n$ in $C(K)$ with $f_n \to g$ in $L^1(\lambda)$ and $L^1(\mu)$. Then $\langle f_n, \mu\rangle \to \langle g, \mu\rangle$, and $\int e^{f_n}\, d\lambda \to \int e^g \, d\lambda$.

(b) Now suppose $\mu \not\ll \lambda$. Then there is an A such that $\lambda(A) = 0$ and $\mu(A) > 0$. Let $g = c\mathbb{1}_A \in B(K)$. This gives

$$\log \int e^g \, d\lambda - \langle g, \mu\rangle = 0 - c\mu(A) \to -\infty$$

as $c \to +\infty$. So $\inf\{\cdots\} = -\infty$, as required.

(c) Lastly, suppose $d\mu = \rho\, d\lambda$. If $\rho = e^g$ with $g \in B(K)$, we are done by the previoius calculation. Otherwise, choose $(g_n)_n$ in $B(K)$ such that

$$e^{g_n} \to \rho \begin{cases} \text{from below} & \text{if } \rho > 1 \\ \text{from above} & \text{if } \rho \leq 1. \end{cases}$$

Now show that:

- $$\log \int_{\{\rho \leq 1\}} e^{g_n}\, d\lambda \to \log \int \rho\, d\lambda = \log 1 = 0,$$

- $$\log \int_{\{\rho > 1\}} e^{g_n}\, d\lambda \to \log \int \rho\, d\lambda = \log 1 = 0,$$

- $$\langle g_n, \mu \rangle \to \langle \log \rho, \mu \rangle = \widetilde{s}(\mu).$$ □