# MATH 585 – TOPICS IN MATHEMATICAL PHYSICS – FALL 2006
## *MATHEMATICS OF MEAN FIELD SPIN GLASSES AND THE REPLICA METHOD*
## LECTURE 5: SANOV'S AND CRAMÉR'S THEOREMS

S. STARR
MATHEMATICS DEPARTMENT, UNIVERSITY OF ROCHESTER

## CONTENTS

## 1. SANOV'S THEOREM

In this lecture we will review the most elementary aspects of large deviations theory. This is the large deviation theory for i.i.d. random variables taking only finitely many values. Following Dembo and Zeitouni [2], we start with Sanov's theorem. We will basically reproduce Dembo and Zeitouni's proof of Sanov's theorem, and its corollary, Cramér's theorem. The interested reader is referred to their book for many deeper results.

Let $\Omega = \{a_1, \dots, a_n\}$ be a finite set. This is simply a finite sample space. But Dembo and Zeitouni also refer to it as a finite *alphabet*, presumably because of the importance of Sanov's theorem in information theory. Let $\mathscr{M}_1(\Omega)$ be the space of probability measures on $\Omega$. Every such measure can be written as

$$\mu = \sum_{i=1}^{n} \theta_i \, \delta_{a_i} \, ,$$

where $\delta_{a_i}$ is the point-mass at $a_i \in \Omega$, and

$$\theta_1, \dots, \theta_n \geq 0 \quad \text{and} \quad \theta_1 + \dots + \theta_n = 1 \, .$$

The map $\mu \mapsto (\theta_1, \dots, \theta_n)$ is an affine bijection between $\mathscr{M}_1(\Omega)$ and $\Sigma_n := \{(\theta_1, \dots, \theta_n) : \forall i \, , \ \theta_i \geq 0 \text{ and } \sum_{i=1}^n \theta_i = 1\}$, the finite-dimensional simplex.

A natural topology on $\mathscr{M}_1(\Omega)$ is the total-variation distance. Actually, on the extended family of signed measures, the total-variation distance is a norm:

$$\|\mu - \nu\|_{\mathrm{TV}} = \max_{A \subseteq \Omega} |\mu(A) - \nu(A)| = \frac{1}{2} \sum_{i=1}^{n} |\mu\{a_i\} - \nu\{a_i\}| \, .$$

*Date*: October 6, 2006.

In particular, its restriction to $\mathscr{M}_1(\Omega)$ is a valid metric. The topology is identical to the standard topology inherited from $\mathbb{R}^n$, through the isomorphism $\mathscr{M}_1(\Omega) \cong \Sigma_n$. Modulo an overall scaling, the total-variation norm is the image of the $\ell^1$-norm on $\mathbb{R}^n$.

Given any $N$-tuple $\boldsymbol{y} = (y_1, \dots, y_N)$ with $y_1, \dots, y_N \in \Omega$, the *empirical measure* is defined as

$$L_N(\boldsymbol{y}) = L_N(y_1, \dots, y_N) = \frac{1}{N} \sum_{k=1}^{N} \delta_{y_k} .$$

This can be written as

$$L_N(\boldsymbol{y}) = \sum_{i=1}^{n} \theta_i \, \delta_{a_i} \quad \text{for} \quad \theta_i = \frac{\#\{k \in [1,N] \, : \, y_k = a_i\}}{N} ,$$

where we write $[1, N]$ for the discrete interval $\{1, \dots, N\}$. This is a measure in $\mathscr{M}_1(\Omega)$. Since $\mathscr{M}_1(\Omega)$ is, itself, a compact metric space, one can consider the large deviation problem on it.

The set-up for Sanov's theorem is that one chooses an element $\mu \in \mathscr{M}_1(\Omega)$, and then lets $\mathsf{Y}_1, \mathsf{Y}_2, \dots$ be i.i.d. random variables, distributed according to $\mu$. Let $\mathbf{P}_\mu$ denote the i.i.d product measure. Sanov's theorem describes the large deviation properties of the sequence of random measures, $L_N(\mathsf{Y}_1, \dots, \mathsf{Y}_N)$, for $N \in \mathbb{Z}_{>0}$. The following is a simple observation.

**Lemma 1.1** *Let $\mu \in \mathscr{M}_1(\Omega)$ be any measure. Suppose $\boldsymbol{y} \in \Omega^N$ and let $\nu = L_N(\boldsymbol{y})$. Then*

$$\mathbf{P}_\mu\{(\mathsf{Y}_1, \dots, \mathsf{Y}_N) = \boldsymbol{y}\} = e^{N[S(\nu|\mu) - H(\nu)]} ,$$

*where $S(\nu|\mu)$ is the relative entropy and $H(\nu)$ is just the "entropy":*

$$S(\nu|\mu) = -\sum_{i=1}^{n} \nu\{a_i\} \log\left(\frac{\nu\{a_i\}}{\mu\{a_i\}}\right) \quad \textit{and} \quad H(\nu) = -\sum_{i=1}^{n} \nu\{a_i\} \log(\nu\{a_i\}) .$$

*(As before, "$0 \log(0) = 0$".)*

*Remark* 1.2 Our sign convention for relative entropy is the opposite of Dembo and Zeitouni's choice, but is made to be consistent with our definitions from earlier lectures.

*Proof.* Using product measure, $\mathbf{P}_\mu$, on $\mathsf{Y}_1, \dots, \mathsf{Y}_N$, one has, for $\boldsymbol{y} = (y_1, \dots, y_N) \in \Omega^N$,

$$\begin{aligned} \mathbf{P}_\mu\{(\mathsf{Y}_1, \dots, \mathsf{Y}_N) = \boldsymbol{y}\} &= \prod_{k=1}^{N} \mu\{y_k\} \\ &= \prod_{i=1}^{n} (\mu\{a_i\})^{\#\{k \in [1,N] \, : \, y_k = a_i\}} \\ &= \prod_{i=1}^{n} (\mu\{a_i\})^{N\nu(\{a_i\})} . \end{aligned}$$

The last equation is because $L_N(\boldsymbol{y}) = \nu$, which gives precisely those identities. So,

$$\mathbf{P}_\mu\{(\mathsf{Y}_1, \dots, \mathsf{Y}_N) = \boldsymbol{y}\} = e^{N \sum_{i=1}^{n} \nu\{a_i\} \log(\mu\{a_i\})} .$$

But, by definition,

$$\sum_{i=1}^{n} \nu\{a_i\} \, \log(\mu\{a_i\}) \;=\; S(\nu|\mu) - H(\mu)\,.$$

□

For each $N \in \mathbb{Z}_{>0}$, Dembo and Zeitouni let $\mathscr{L}_N \subset \mathscr{M}_1(\Omega)$ be the subset consisting of all $\nu$ which can be expressed as

$$\nu \;=\; \sum_{i=1}^{n} \theta_i \, \delta_{a_i} \quad \text{such that} \quad \theta_1, \dots, \theta_n \in \left\{0, \tfrac{1}{N}, \tfrac{2}{N}, \dots, 1\right\}\,.$$

This is the set of all measures $\nu$ which can be realized as $L_N(\boldsymbol{y})$ for *some* $\boldsymbol{y} \in \Omega^N$. Dembo and Zeitouni also define $T_N(\nu)$ to be the preimage under $L_N$ of $\nu$ for each $\nu \in \mathscr{L}_N$:

$$T_N(\nu) \;:=\; \left\{\boldsymbol{y} \in \Omega^N \;:\; L_N(\boldsymbol{y}) = \nu\right\}\,.$$

The lemma shows that, for $\nu \in \mathscr{L}_N$,

$$\mathbf{P}_\mu\{L_N(\mathsf{Y}_1, \dots, \mathsf{Y}_N) = \nu\} \;=\; \sum_{\boldsymbol{y} \in T_N(\nu)} \mathbf{P}_\mu\{(\mathsf{Y}_1, \dots, \mathsf{Y}_N) = \boldsymbol{y}\} \;=\; e^{N[S(\nu|\mu) - H(\nu)]} \, \#T_N(\nu)\,.$$

Moreover, by elementary combinatorics, it is clear that

$$\#T_N(\nu) \;=\; \binom{N}{N\nu\{a_1\}, N\nu\{a_2\}, \dots, N\nu\{a_n\}}\,.$$

Therefore, one could extract the asymptotics related to $\mathbf{P}_\mu\{L_N(\mathsf{Y}_1, \dots, \mathsf{Y}_N) = \nu\}$ by resorting to Stirling's formula, if one has learned this from a course on asymptotic analysis. (See, for example, Section 1.4 in [1].) But Dembo and Zeitouni present a different and more elegant solution. It consists in bounding $\#T_N(\nu)$ by making clever use of Lemma 1.1:

**Lemma 1.3** *For every $\nu \in \mathscr{L}_N$,*

$$\frac{e^{NH(\nu)}}{\#\mathscr{L}_N} \;\leq\; \#T_N(\nu) \;\leq\; e^{NH(\nu)}\,. \tag{1.1}$$

*Remark* 1.4 Note that $H(\nu)$ is a concave function. It is well-known that, in finite dimensions, concave functions, defined on compact, convex domains, always attain their minima on the extreme points of those domains. The set $\mathscr{M}_1(\Omega)$ is compact and convex: in fact it is a simplex. The extreme points of $\mathscr{M}_1(\Omega)$ are the point-masses $\delta_{a_i}$, for $i \in \{1, \dots, n\}$. It is easy to see that $H(\delta_{a_i}) = 0$. Therefore, the "entropy" is always nonnegative.

*Proof.* The upper bound is obtained by direct application of Lemma 1.1. Recall that $S(\nu|\nu) = 0$. Therefore, taking $\nu \in \mathscr{L}_N$, we have

$$\mathbf{P}_\nu\{(\mathsf{Y}_1, \dots, \mathsf{Y}_N) = \boldsymbol{y}\} \;=\; e^{-NH(\nu)}\,,$$

for $\boldsymbol{y} \in T_N(\nu)$. This follows from Lemma 1.1, just replacing $\mu$ by $\nu$. But then clearly

$$\#T_N(\nu) \;\leq\; e^{NH(\nu)}\,,$$

because by our previous calculation

$$e^{-NH(\nu)} \, \#T_N(\nu) \;=\; \mathbf{P}_\nu\{L_N(\mathsf{Y}_1, \dots, \mathsf{Y}_N) = \nu\}\,,$$

and no probability can be greater than 1. The lower bound is slightly trickier and uses the following elementary exercise.

**Exercise.** *Given $\lambda \geq 0$, define the discrete probability mass function $f_\lambda : \mathbb{Z}_{\geq 0} \to [0,1]$ as*

$$f_\lambda(x) = e^{-\lambda} \frac{\lambda^x}{x!}\,.$$

*I.e., this is the p.m.f. of the Poisson($\lambda$) random variable. Then this distribution is "unimodal", in the sense that*

$$\begin{cases} f_\lambda(x) \leq f_\lambda(y) & \textit{if} \quad x \leq y \leq \lambda; \\ f_\lambda(x) \geq f_\lambda(y) & \textit{if} \quad \lambda \leq x \leq y. \end{cases}$$

Figure 1 shows the p.m.f. for a particular choice of $\lambda$, wherein one sees plainly that there is only a single "mode" in the graphical sense that there is a single "hump". For most values of $\lambda$ there is also only one statistical mode (absolute maximum for $f_\lambda$) for the distribution. But for $\lambda = n \in \mathbb{Z}_{>0}$ there are two statistical modes, at $n$ and $n-1$.

The consequence of this inequality when $\lambda = k \in \mathbb{Z}_{\geq 0}$ is that

$$\frac{k^k}{k!} \geq \frac{k^j}{j!}\,, \tag{1.2}$$

for any other $j \in \mathbb{Z}_{\geq 0}$ (with equality if $j = k-1$). This inequality will be used to prove that

$$\mathbf{P}_\nu\{L_N(\mathsf{Y}_1,\dots,\mathsf{Y}_N) = \nu\} \geq \mathbf{P}_\nu\{L_N(\mathsf{Y}_1,\dots,\mathsf{Y}_N) = \nu'\} \tag{1.3}$$

for any other $\nu' \in \mathscr{L}_N$, for which $\mathrm{supp}(\nu') \subseteq \mathrm{supp}(\nu)$. By Lemma 1.1, and properties of $S(\cdot|\nu)$, the right hand side of (1.3) is zero if $\mathrm{supp}(\nu') \nsubseteq \mathrm{supp}(\nu)$.

In proving (1.3) we can assume that $\mathrm{supp}(\nu) = \Omega$. Otherwise, we would just reduce our alphabet to be $\mathrm{supp}(\nu)$, which does not affect either side of (1.3) as long as $\mathrm{supp}(\nu') \subseteq \mathrm{supp}(\nu)$. Then a straightforward calculation with the multinomial distribution gives

$$\begin{aligned} \frac{\mathbf{P}_\nu\{L_N(\mathsf{Y}_1,\dots,\mathsf{Y}_N) = \nu\}}{\mathbf{P}_\nu\{L_N(\mathsf{Y}_1,\dots,\mathsf{Y}_N) = \nu'\}} &= \frac{\binom{N}{N\nu\{a_1\},\dots,N\nu\{a_n\}} \prod_{i=1}^n (\nu\{a_i\})^{N\nu\{a_i\}}}{\binom{N}{N\nu'\{a_1\},\dots,N\nu'\{a_n\}} \prod_{i=1}^n (\nu\{a_i\})^{N\nu'\{a_i\}}} \\ &= \prod_{i=1}^n \frac{(N\nu'\{a_i\})!}{(N\nu\{a_i\})!} (\nu\{a_i\})^{N[\nu\{a_i\}-\nu'\{a_i\}]}\,. \end{aligned}$$

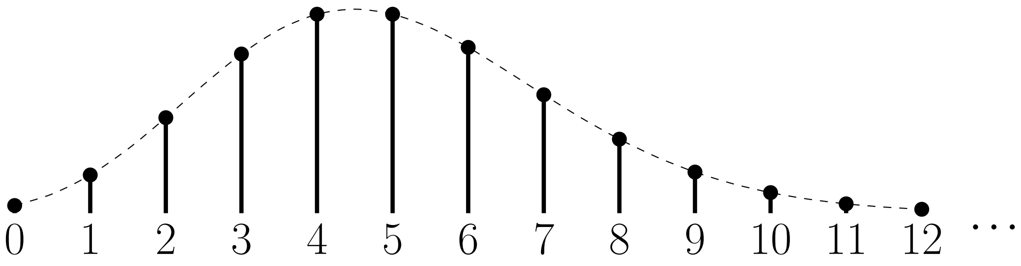

FIGURE 1. $f_\lambda(x)$ for Poisson r.v. with $\lambda = 5$, for $x = 0,\dots,12$. (Note whenever $\lambda = n$ for some $n \in \mathbb{Z}_{>0}$, one has $f_n(n-1) = f_n(n)$. This does not violate unimodality, which allows for equality as well as strict inequality.)

By (1.2), with $k = N\nu\{a_i\}$ and $j = N\nu'\{a_i\}$, we see that

$$\frac{(N\nu'\{a_i\})!}{(N\nu\{a_i\})!} \;\geq\; (N\nu\{a_i\})^{N[\nu'\{a_i\}-\nu\{a_i\}]}\,.$$

Therefore, plugging this back into the inequality for the ratio yields

$$\begin{aligned}\frac{\mathbf{P}_\nu\{L_N(\mathsf{Y}_1,\dots,\mathsf{Y}_N)=\nu\}}{\mathbf{P}_\nu\{L_N(\mathsf{Y}_1,\dots,\mathsf{Y}_N)=\nu'\}} \;&\geq\; \prod_{i=1}^n N^{N[\nu'\{a_i\}-\nu\{a_i\}]}\\ &=\; N^{N\left[\sum_{i=1}^n \nu'\{a_i\}-\sum_{i=1}^n \nu\{a_i\}\right]}\\ &=\; 1\,.\end{aligned}$$

This proves (1.3).

Now this can be used to prove the lower bound, because

$$\begin{aligned}1 \;&=\; \sum_{\nu'\in\mathscr{L}_N} \mathbf{P}_\nu\{L_N(\mathsf{Y}_1,\dots,\mathsf{Y}_N)=\nu'\}\\ &\leq\; \mathbf{P}_\nu\{L_N(\mathsf{Y}_1,\dots,\mathsf{Y}_N)=\nu\}\,\#\mathscr{L}_N\\ &=\; e^{-NH(\nu)}\,\#T_N(\nu)\,\#\mathscr{L}_N\,.\end{aligned}$$

□

Combining Lemma 1.1 and Lemma 1.3 gives:

**Corollary 1.5** *For any $\mu \in \mathscr{M}_1(\Omega)$ and any $\nu \in \mathscr{L}_N$,*

$$\frac{e^{NS(\nu|\mu)}}{\#\mathscr{L}_N} \;\leq\; \mathbf{P}_\mu\{L_N(\mathsf{Y}_1,\dots,\mathsf{Y}_N)=\nu\} \;\leq\; e^{NS(\nu|\mu)}\,.$$

*Remark* 1.6 Recall that $S(\nu|\mu)$ is always nonpositive, because for a fixed $\mu$ the maximizer of $S(\nu|\mu)$ is $\nu=\mu$, which gives 0. This was proved in the "addendum on entropy".

Note that, on the exponential scale, $\#\mathscr{L}_N$ does not contribute. The reason is that it only grows as a polynomial with $N$, as the following exercise shows.

**Exercise.** *Let $\mathscr{L}'_N = \{(k_1,\dots,k_n) \,:\, k_1,\dots,k_n \in \mathbb{Z}_{\geq 0} \quad \text{and} \quad k_1+\dots+k_n = N\}$, which is in bijection to $\mathscr{L}_N$, basically just dividing by $N$. Show that $\#\mathscr{L}'_N = \binom{N+n-1}{n-1}$.*
*(*HINT: *A common approach is to use the "stars-and-bars" representation. For example $(1,0,3,2)$ would be represented as $|*||***|**|$.)*

**Theorem 1.7** (Sanov's Theorem) *For arbitrary subsets, $\Gamma \subseteq \mathscr{M}_1(\Omega)$,*

$$\limsup_{N\to\infty}\frac{1}{N}\log\left(\mathbf{P}_\mu\{L_N(\mathsf{Y}_1,\dots,\mathsf{Y}_N)\in\Gamma\}\right) \;\leq\; \sup_{\nu\in\Gamma} S(\nu|\mu)\,.$$

*For open subsets, $U \subseteq \mathscr{M}_1(\Omega)$,*

$$\liminf_{N\to\infty}\frac{1}{N}\log\left(\mathbf{P}_\mu\{L_N(\mathsf{Y}_1,\dots,\mathsf{Y}_N)\in U\}\right) \;\geq\; \sup_{\nu\in U} S(\nu|\mu)\,.$$

We will not prove this result but here are some ideas. Because of Corollary 1.5 and the exercise,

$$\lim_{N\to\infty}\left[\frac{1}{N}\log\left(\mathbf{P}_\mu\{L_N((\mathsf{Y}_1,\dots,\mathsf{Y}_N))\in\Gamma_N\}\right)\, -\, \sup_{\nu\in\Gamma_N} S(\nu|\mu)\right]\, =\, 0\,,$$

for any sequence of sets $\Gamma_1,\Gamma_2,\dots$ with $\Gamma_N\subseteq\mathscr{L}_N$ for each $N\in\mathbb{Z}_{>0}$. Letting $\Gamma_N=\Gamma\cap\mathscr{L}_N$ leads easily to the upper bound because

$$\limsup_{N\to\infty}\ \sup_{\nu\in\Gamma\cap\mathscr{L}_N} S(\nu|\mu)\ \leq\ \sup_{\nu\in\Gamma} S(\nu|\mu)\,.$$

The lower bound is a little more tricky. One wants to prove

$$\liminf_{N\to\infty}\ \sup_{\nu\in U\cap\mathscr{L}_N} S(\nu|\mu)\ \geq\ \sup_{\nu\in U} S(\nu|\mu)\,,$$

for open sets $U$. An optimizing sequence $\nu_1,\nu_2,\dots$, for getting the $\sup$ on the right hand side of the inequality may be chosen in the subset $U\cap\mathscr{M}_1(\mathrm{supp}(\mu))$, because on the complementary subset of $U$, the relative entropy is $-\infty$. Therefore, it is of fundamental importance that $S(\cdot|\mu)$ is continuous on the set $\mathscr{M}_1(\mathrm{supp}(\mu))$. So the desired inequality basically follows by showing that, for any $\nu\in\mathscr{M}_1(\mathrm{supp}(\mu))$, there is a sequence of measures $\nu^{(1)},\nu^{(2)},\dots$ with:

- $\nu^{(N)}\in\mathscr{M}_1(\mathrm{supp}(\mu))\cap\mathcal{L}_N$ for all $N$; and
- $\lim_{N\to\infty}\|\nu-\nu^{(N)}\|_{\mathrm{TV}}=0$. (In fact one can choose $\|\nu-\nu^{(N)}\|_{\mathrm{TV}}\leq n/N$.)

These few words do not prove the theorem of course. For the real proof consult [2].

## 2. Cramér's Theorem

The following is a stronger version of Cramér's theorem than appeared in Lecture 4. (We recall that $\mathrm{cch}$ denotes "closed, convex hull".)

**Theorem 2.1** (Cramér's theorem for finite subsets of $\mathbb{R}$) *Let $\mathsf{X}_1,\mathsf{X}_2,\dots$ be i.i.d. random variables, with distribution $P_1$ such that $\mathrm{supp}(P_1)$ is a finite subset of $\mathbb{R}$. For each $N\in\mathbb{Z}_{>0}$, let $P_N$ be the probability distribution of*

$$\frac{\mathsf{X}_1+\cdots+\mathsf{X}_N}{N}\,.$$

*Let $\mathscr{X}=\mathrm{cch}(\mathrm{supp}(P_1))$. For $x\in\mathscr{X}$, define the rate function*

$$I(x)\ =\ \sup_{\lambda\in\mathbb{R}}[\lambda x-\Lambda(\lambda)]\,,$$

*where $\Lambda:\mathbb{R}\to\mathbb{R}$ is the logarithmic moment generating function of $P_1$,*

$$\Lambda(\lambda)\ :=\ \log\left(\mathbf{E}^{P_1}\left[e^{\lambda x}\right]\right)\,.$$

*Then, for every subset $\Gamma\subseteq\mathscr{X}$,*

$$\limsup_{N\to\infty}\frac{1}{N}\log(P_N(\Gamma))\ \leq\ -\inf_{x\in\Gamma} I(x)\,,$$

*and for each open subset $U\subseteq\mathscr{X}$,*

$$\liminf_{N\to\infty}\frac{1}{N}\log(P_N(U))\ \geq\ -\inf_{x\in U} I(x)\,.$$

Suppose $\mathrm{supp}(P_1) = \{r_1, \dots, r_n\} \subset \mathbb{R}$. In order to fit with the previous section, let $\Omega = \{a_1, \dots, a_n\}$ and define $f : \Omega \to \mathbb{R}$ so that $f(a_i) = r_i$. One can extend the map to $f_* : \mathscr{M}_1(\Omega) \to \mathscr{X}$ by defining

$$f_*(\nu) := \mathbf{E}^{\nu}[f] = \sum_{i=1}^{n} f(a_i)\nu(a_i) .$$

Then, defining $\mu = P_1 \circ f$, for any subset $\Gamma \subseteq \mathscr{X}$,

$$P_N(\Gamma) = \mathbf{P}_{\mu}\{L_N(\mathsf{Y}_1, \dots, \mathsf{Y}_N) \in f_*^{-1}(\Gamma)\} . \tag{2.1}$$

Indeed, taking $\mathsf{Y}_1, \mathsf{Y}_2, \dots$ to be i.i.d. random variables distributed by $\mu$, as before, we can define $\mathsf{X}_i = f(\mathsf{Y}_i)$ so that $\mathsf{X}_1, \mathsf{X}_2, \dots$ are i.i.d. random variables with distribution $P_1$. But

$$f_*(L_N(\mathsf{Y}_1, \dots, \mathsf{Y}_N)) = \frac{1}{N}\sum_{k=1}^{N} f(\mathsf{Y}_k) = \frac{\mathsf{X}_1 + \cdots + \mathsf{X}_N}{N} .$$

Therefore, with this coupling between $(\mathsf{Y}_k)_{k=1}^{\infty}$ and $(\mathsf{X}_k)_{k=1}^{\infty}$, the following events are equal

$$\{L_N(\mathsf{Y}_1, \dots, \mathsf{Y}_N) \in f_*^{-1}(\Gamma)\} = \left\{\frac{\mathsf{X}_1 + \cdots + \mathsf{X}_N}{N} \in \Gamma\right\} .$$

That certainly implies (2.1). Also, the map $f_*$ is continuous (in fact affine). So, for an open subset $U \subset \mathscr{X}$, one has $f_*^{-1}(U)$ is also open. Thus, by Sanov's theorem, one obtains: for every subset $\Gamma \subseteq \mathscr{X}$,

$$\limsup_{N\to\infty} \frac{1}{N}\log(P_N(\Gamma)) \leq \sup_{\nu \in f_*^{-1}(\Gamma)} S(\nu|\mu) ;$$

and for every open subset $U \subseteq \mathscr{X}$,

$$\liminf_{N\to\infty} \frac{1}{N}\log(P_N(U)) \geq \sup_{\nu \in f_*^{-1}(U)} S(\nu|\mu) .$$

Then, defining

$$I(x) = -\sup_{\nu \in f_*^{-1}(x)} S(\nu|\mu) ,$$

one has

$$\sup_{\nu \in f_*^{-1}(\Gamma)} S(\nu|\mu) = -\inf_{x\in\Gamma} I(x) .$$

Therefore, one obtains the result of Cramér's theorem except that the rate function is as above. The only thing left to prove is that this definition also matches the other one, namely that, for each $x \in \mathscr{X}$,

$$\sup_{\nu \in f_*^{-1}(x)} S(\nu|\mu) = -\sup_{\lambda\in\mathbb{R}}[\lambda x - \Lambda(\lambda)] = \inf_{\lambda\in\mathbb{R}}[\Lambda(\lambda) - \lambda x] .$$

*Proof.* The proof of this result will use a variant of the Gibbs variational principle. Observe first that

$$\Lambda(\lambda) = \log\left(\mathbf{E}^{P_1}[e^{\lambda x}]\right) ,$$

and this can be rewritten as

$$\Lambda(\lambda) \,=\, \log\left(\sum_{i=1}^n e^{\lambda f(a_i)} \mu\{a_i\}\right)\,.$$

This looks very much like a pressure, where $\lambda f$ is playing the role of $-\beta H$. Indeed, the most important difference is that $\mu$, which is playing the role of the *a priori* measure, need not be uniform. But that does not nullify the result of the Gibbs variational principle. Indeed, since the *a priori* measure is at least normalized, we lose the nuisance factor of $\log(|\Omega|)$ which we would have had if we took counting measure, instead. By the Gibbs variational principle, we obtain

$$\Lambda(\lambda) \,=\, \sup_{\nu\in\mathscr{M}_1(\Omega)} (S(\nu|\mu) + \lambda\mathbb{E}^\nu[f]) \,=\, \sup_{\nu\in\mathscr{M}_1(\Omega)} [S(\nu|\mu) + \lambda f_*(\nu)]\,.$$

Therefore, substituting back into the function we actually want to optimize, we have

$$\inf_{\lambda\in\mathbb{R}}[\Lambda(\lambda) - \lambda x] \,=\, \inf_{\lambda\in\mathbb{R}} \sup_{\nu\in\mathscr{M}_1(\Omega)} [S(\nu|\mu) + \lambda(f_*(\nu) - x)]\,.$$

Now we are in a position to use the Kneser-Fan theorem, which we introduced in Lecture 4, and which is proved, for example, in [3]. The function

$$\mathcal{L}(\lambda,\nu) \,=\, S(\nu|\mu) + \lambda(f_*(\nu) - x)$$

is convave in $\nu$ and convex in $\lambda$. Indeed it is linear in $\lambda$. The reason it is concave in $\nu$ is that $S(\nu|\mu)$ is concave, and $f_*(\nu)$ is linear (or more appropriately affine), therefore convex-and-concave. The set $\mathscr{M}_1(\Omega)$ is compact and convex. Therefore, by the Kneser-Fan theorem,

$$\inf_{\lambda\in\mathbb{R}} \sup_{\nu\in\mathscr{M}_1(\Omega)} [S(\nu|\mu) + \lambda(f_*(\nu) - x)] \,=\, \sup_{\nu\in\mathscr{M}_1(\Omega)} \inf_{\lambda\in\mathbb{R}} [S(\nu|\mu) + \lambda(f_*(\nu) - x)]\,.$$

But, except when $f_*(\nu) = x$, one has

$$\inf_{\lambda\in\mathbb{R}} [S(\nu|\mu) + \lambda(f_*(\nu) - x)] \,=\, -\infty\,.$$

This will surely not contribute to the supremum in $\nu\in\mathscr{M}_1(\Omega)$ (unless that supremum is $-\infty$, in which case making further restrictions will cause no harm). Therefore, one must restrict to $\nu$ such that $f_*(\nu) = x$. I.e., one must restrict to $\nu\in f_*^{-1}(x)$. Thus one has

$$\begin{aligned}\sup_{\nu\in\mathscr{M}_1(\Omega)} \inf_{\lambda\in\mathbb{R}} [S(\nu|\mu) + \lambda(f_*(\nu) - x)] \,&=\, \sup_{\nu\in f_*^{-1}(x)} \inf_{\lambda\in\mathbb{R}} [S(\nu|\mu) + \lambda(f_*(\nu) - x)] \\ &=\, \sup_{\nu\in f_*^{-1}(x)} S(\nu|\mu)\,.\end{aligned}$$

This is what we wanted to prove. □

*Remark* 2.2 This proof used the Kneser-Fan theorem, but we did not prove that theorem. A more direct proof, which is even simpler can be found in [2]. Also, one should usually mention at this point that what we are doing is taking the Legedre-Fenchel transform, which is partially involutive. One could do a much better job of stressing the connections between these points, and it usually is done in a course on large deviations, for example.

## REFERENCES

[1] G. E. Andrews, R. Askey and R. Roy, *Special Functions*, Encyclopedia of Mathematics and its Applications v. 71, Cambridge University Press, Cambridge, UK, 1999.

[2] A. Dembo and O. Zeitouni, *Large Deviation Techniques and Applications*, Applications of Mathematics v. 38, Springer Verlag, New York, Inc., 1998.

[3] M. Sion. On General Minimax Theorems. *Pacific J. Math.* **8**, 171–176, 1958.

MATHEMATICS DEPARTMENT, UNIVERSITY OF ROCHESTER, ROCHESTER, NY 14627
*E-mail address*: sstarr@math.rochester.edu