LECTURE 8: FREE ENTROPY AND LARGE DEVIATIONS

An important concept in classical probability theory is Shannon's notion of entropy. Having developed the analogy between free and classical probability theory, one hopes to find that a notion of "free entropy" exists in counterpart to the Shannon entropy. In fact there is a useful notion of free entropy. However, the development of this new concept is at present far from complete. The current state of affairs is that there are two distinct approaches to free entropy. These should give isomorphic theories, but at present we only know that they coincide in a limited number of situations.

The first approach to a theory of free entropy is via "microstates." This is rooted in the concept of large deviations. The second approach is "microstates free." This draws its inspiration from the statistical approach to classical entropy via the notion of Fisher information. The unification problem in free probability theory is to prove that these two theories of free entropy are consistent.

Let us return to the connection between random matrix theory and free probability theory which we have been developing. We know that a p-tuple

$$(A_N^{(1)}, \ldots, A_N^{(p)})$$

of $N \times N$ matrices chosen independently at random with respect to the density

$$P_N(A) = \text{const} \cdot e^{\frac{-N}{2} \operatorname{Tr}(A^2)} \tag{1}$$

on the space of $N \times N$ Hermitian matrices converges almost surely (in moments with respect to the normalized trace) to a freely independent family

$$(s_1, \ldots, s_p)$$

of semi-circular elements lying in a non-commutative probability space. The von Neumann algebra generated by p freely independent semi-circulars is the von Neumann algebra $L(\mathbb{F}_p)$ of the free group on p generators.

How likely is it to observe other distributions/operators for large N?

Date: Lecture given on Nov. 22, 2007.

Let us consider the case $p = 1$ more closely. For a random Hermitian matrix $A = A^*$ (distribution as above) with real random eigenvalues

$$\lambda_1 \leq \cdots \leq \lambda_N, \tag{2}$$

denote by

$$\mu_A = \frac{1}{N}(\delta_{\lambda_1} + \cdots + \delta_{\lambda_N}) \tag{3}$$

the eigenvalue distribution of A (also known as the "empirical eigenvalue distribution"), which is a random measure on $\mathbb{R}$. Wigner's semicirle law states that as $N \to \infty$

$$P_N(\mu_A \approx \mu_W) \to 1, \tag{4}$$

where μ_W is the (non-random) semicircular distribution (see Lecture 1). What is the rate of decay of the probability

$$P_N(\mu_A \approx \nu), \tag{5}$$

where ν is some measure (not necessarily the semicircle)? We expect that

$$P_N(\mu_A \approx \nu) \sim e^{-N^2 I(\nu)} \tag{6}$$

for some "rate function" I vanishing at μ_W. By analogy with the classical theory of large deviations, I should correspond to a suitable notion of free entropy.

1. Large Deviation Theory

Consider a sequence $X_1, X_2, \ldots$ of independent identically distributed random variables with distribution μ, and put

$$S_n = \frac{X_1 + \cdots + X_n}{n}. \tag{7}$$

Let $m = \mathbb{E}[X_i]$. Then the law of large numbers asserts that $S_n \to m$, while the central limit theorem tells us that for large n

$$S_n \approx m + \frac{\sigma^2}{\sqrt{n}} N(0,1). \tag{8}$$

For example if $\mu = N(0,1)$ then S_n has distribution $N(0, \frac{1}{n})$ and hence

$$P(S_n \approx x) = e^{-n\frac{x^2}{2}} \frac{\sqrt{n}}{\sqrt{2\pi}} \sim e^{-nI(x)}. \tag{9}$$

Thus the probability that S_n is near the value x decays exponentially in n at a rate determined by x, namely the "rate function" $I(x) = \frac{x^2}{2}$. Note that the convex function $I(x)$ has a global minimum at $x = 0$,

the minimum value there being 0, which corresponds to the fact that S_n approaches the mean 0 in probability.

This behavior is described in general by the following theorem of Cramér: Let $X_1, X_2, X_3, \ldots$ be a sequence of i.i.d random variables, with mean m, and set

$$S_n := \frac{1}{n}(X_1 + \cdots + X_n), \tag{10}$$

the empirical mean. There exists a function $I(x)$, the rate function, such that

$$P(S_n > x) \approx e^{-nI(x)}, x > m \tag{11}$$

$$P(S_n < x) \approx e^{-nI(x)}, x < m. \tag{12}$$

How does one calculate the rate function $I(x)$ for a given distribution μ? Suppose μ has mean 0. For arbitrary $x > 0$, one has for all $\lambda > 0$

$$\begin{aligned} P(S_n > x) &= P(nS_n > nx) \\ &\leq \mathbb{E}[e^{\lambda(nS_n - nx)}] \\ &= e^{-\lambda nx}\mathbb{E}[e^{\lambda(X_1+\cdots+X_n)}] \\ &= e^{-\lambda nx}\mathbb{E}[e^{\lambda X_i}]^n. \end{aligned}$$

Now put

$$\Lambda(\lambda) := \log \mathbb{E}[e^{\lambda X_i}], \tag{13}$$

the cumulant generating series of μ. Then the above reads

$$P(S_n > x) \leq e^{-\lambda nx + n\Lambda(\lambda)} = e^{-n(\lambda x - \Lambda(\lambda))}, \tag{14}$$

valid for all $\lambda > 0$. As Λ achieves its minimum at $\lambda = 0$, the above estimate is also valid for negative λ. Thus

$$P(S_n > x) \leq \inf_\lambda e^{-n(\lambda x - \Lambda(\lambda))} = e^{-n \sup_\lambda \{\lambda x - \Lambda(\lambda)\}}. \tag{15}$$

The function $\lambda \mapsto \Lambda(\lambda)$ is convex, and the "Legendre transform" of Λ defined by

$$\Lambda^*(x) := \sup_\lambda \{\lambda x - \Lambda(\lambda)\} \tag{16}$$

is also convex. Thus we have proved that

$$P(S_n > x) \leq e^{-n\Lambda^*(x)}, \tag{17}$$

where Λ^* is the Legendre transform of the cumulant generating function Λ. This gives Λ^* as a candidate for the rate function; however we also

have to check that $e^{-n\Lambda^*(x)}$ is, at least asymptotically, a lower bound; more precisely, we need to verify that

$$\liminf \frac{1}{n} \log P(x - \delta < S_n < x + \delta) \geq -\Lambda^*(x) \tag{18}$$

for all x and all $\delta > 0$. By making an appropriate shift we can reduce this to the case $x = 0$. Then $-\Lambda^*(0) = \inf_\lambda \Lambda(\lambda)$. The idea of the proof is then to perturb the distribution μ to $\tilde{\mu}$ such that $x = 0$ is the mean of $\tilde{\mu}$. Consider the case where $\Lambda(\lambda)$ has a global minimum at η, and put

$$\frac{d\tilde{\mu}}{d\mu}(x) = e^{\eta x - \Lambda(\eta)}. \tag{19}$$

Note that

$$\int_{\mathbb{R}} d\tilde{\mu} = e^{-\Lambda(\eta)} \int e^{\eta x} dx \tag{20}$$

$$= e^{-\Lambda(\eta)} \mathbb{E}[e^{\eta X_i}] \tag{21}$$

$$= e^{-\Lambda(\eta)} e^{\Lambda(\eta)} \tag{22}$$

$$= 1, \tag{23}$$

which verifies that $\tilde{\mu}$ is a probability measure. Consider now $\tilde{X}_i$ i.i.d with distribution $\tilde{\mu}$, and put

$$\tilde{S}_n = \frac{\tilde{X}_1 + \cdots + \tilde{X}_n}{n}. \tag{24}$$

We have

$$\mathbb{E}[\tilde{X}_i] = \int x d\tilde{\mu}(x) \tag{25}$$

$$= e^{-\Lambda(\eta)} \int x e^{\eta x} dx \tag{26}$$

$$= e^{-\Lambda(\eta)} \frac{d}{d\lambda} e^{\Lambda(\lambda)}|_{\lambda=\eta} \tag{27}$$

$$= e^{-\Lambda(\eta)} \Lambda'(\eta) e^{\Lambda(\eta)} \tag{28}$$

$$= \Lambda'(\eta) \tag{29}$$

$$= 0. \tag{30}$$

Now, for all $\epsilon > 0$, we have

$$P(-\epsilon < S_n < \epsilon) = \int_{|\sum_{i=1}^n X_i| < n\epsilon} \mu(dx_1) \dots \mu(dx_n) \tag{31}$$

$$\geq e^{-n\epsilon|\eta|} \int_{|\sum_{i=1}^n X_i| < n\epsilon} e^{\eta \sum X_i} \mu(dx_1) \dots \mu(dx_n) \tag{32}$$

$$= e^{-n\epsilon|\eta|} e^{n\Lambda(\eta)} \int_{|\sum_{i=1}^n X_i| < n\epsilon} \tilde{\mu}(dx_1) \dots \tilde{\mu}(dx_n) \tag{33}$$

$$= e^{-n\epsilon|\eta|} e^{n\Lambda(\eta)} P(-\epsilon < \tilde{S}_n < \epsilon). \tag{34}$$

By the law of large numbers, $\tilde{S}_n \to \mathbb{E}_{\tilde{\mu}}[\tilde{X}_i] = 0$, i.e.

$$\lim_{n\to\infty} P(-\epsilon < \tilde{S}_n < \epsilon) = 1 \tag{35}$$

for all $\epsilon > 0$. Thus for all $0 < \epsilon < \delta$

$$\liminf \frac{1}{n} \log P(-\delta < S_n < \delta) \geq \liminf \frac{1}{n} \log P(-\epsilon < S_n < \epsilon) \tag{36}$$

$$\geq \Lambda(\eta) - \epsilon|\eta|, \text{ for all } \epsilon \tag{37}$$

$$\geq \Lambda(\eta) \tag{38}$$

$$= \inf \Lambda(\lambda) \tag{39}$$

$$= -\Lambda^*(0). \tag{40}$$

This sketches the proof of Cramer's theorem for $\mathbb{R}$. A higher-dimensional form of Cramer's theorem is given in the next section.

2. Cramer's Theorem for $\mathbb{R}^d$

Let $X_1, X_2, \dots$ be a sequence of i.i.d random vectors, i.e. independent $\mathbb{R}^d$-valued random variables with common distribution μ (a probability measure on $\mathbb{R}^d$). Put

$$\Lambda(\lambda) := \mathbb{E}[e^{\langle \lambda, X_i \rangle}], \ \lambda \in \mathbb{R}^d, \tag{41}$$

and

$$\Lambda^*(x) := \sup_{\lambda \in \mathbb{R}^d} \{\langle \lambda, x \rangle - \Lambda(\lambda)\}. \tag{42}$$

Assume that $\Lambda(\lambda) < \infty$ for all $\lambda \in \mathbb{R}^d$, and put

$$S_n := \frac{1}{n}(X_1 + \dots + X_n). \tag{43}$$

Then the distribution μ_{S_n} of the random variable S_n satisfies a large deviation principle with rate function Λ^*, i.e.

- $x \mapsto \Lambda^*(x)$ is lower semicontinuous (actually convex)

- Λ^* is "good," i.e. $\{x \in \mathbb{R}^d : \Lambda^*(x) \leq \alpha\}$ is compact for all $\alpha \in \mathbb{R}$
- For any closed set $F \subset \mathbb{R}^d$, $\limsup_n \frac{1}{n} \log P(S_n \in F) \leq -\inf_{x \in F} \Lambda^*(x)$
- For any open set $G \subset \mathbb{R}^d$, $\liminf_n \log P(S_n \in G) \geq -\inf_{x \in G} \Lambda^*(x)$.

This is Cramer's Theorem for $\mathbb{R}^d$, and in an informal way it says

$$P(S_n \approx x) \sim e^{-n\Lambda^*(x)}. \tag{44}$$

Actually, we are interested not in S_n, but in the empirical distribution

$$\frac{1}{n}(\delta_{X_1} + \cdots + \delta_{X_n}). \tag{45}$$

Let us consider this in the special case of random variables taking values in a finite alphabet $A = \{a_1, \ldots, a_d\}$:

$$X_i : \Omega \to A, \tag{46}$$

with $p_k := P(X_i = a_k)$. As $n \to \infty$, the empirical distribution of the X_i's should converge to the "most likely" probability measure $(p_1, \ldots, p_d)$ on A.

Now define the vector of indicator functions $Y_i : \Omega \to \mathbb{R}^d$ by

$$Y_i := (1_{a_1}(X_i), \ldots, 1_{a_d}(X_i)), \tag{47}$$

so that in particular p_k is equal to the probability that Y_i will have a 1 in the k-th spot and 0's elsewhere. Then

$$\frac{1}{n}(Y_1 + \cdots + Y_n) \tag{48}$$

gives the relative frequency of $a_1, \ldots, a_d$:

$$\frac{1}{n}(\delta_{X_1} + \cdots + \delta_{X_n}), \tag{49}$$

i.e. the empirical distribution of $(X_1, \ldots, X_n)$.

A probability measure on A is given by a d-tuple $(q_1, \ldots, q_d)$ of positive real numbers satisfying $q_1 + \cdots + q_d = 1$. By Cramer's theorem,

$$P\{\frac{1}{n}(\delta_{X_1} + \cdots + \delta_{X_n}) \approx (q_1, \ldots, q_d)\} = P(\frac{Y_1 + \cdots + Y_n}{n} \approx (q_1, \ldots, q_d)) \sim e^{-n\Lambda^*(q_1, \ldots, q_d)}. \tag{50}$$

Here

$$\Lambda(\lambda_1, \ldots, \lambda_d) = \log \mathbb{E}[e^{\langle \lambda, Y_i \rangle}] \tag{51}$$

$$= \log(p_1 e^{\lambda_1} + \cdots + p_d e^{\lambda_d}). \tag{52}$$

Thus the Legendre transform is given by

$$\Lambda^*(q_1, \ldots, q_d) = \sup_{(\lambda_1, \ldots, \lambda_d)} \{\lambda_1 q_1 + \cdots + \lambda_d q_d - \Lambda(\lambda_1, \ldots, \lambda_d)\}. \tag{53}$$

We compute the supremum over all tuples $(\lambda_1, \ldots, \lambda_d)$ by finding the partial $\partial/\partial\lambda_i$ of

$$\lambda_1 q_1 + \cdots + \lambda_d q_d - \Lambda(\lambda_1, \ldots, \lambda_d) \tag{54}$$

to be

$$q_i - \frac{1}{p_1 e^{\lambda_1} + \cdots + p_d e^{\lambda_d}} p_i e^{\lambda_i}. \tag{55}$$

Thus the max occurs when

$$\lambda_i = \log \frac{q_i}{p_i} + \log(p_1 e^{\lambda_1} + \cdots + p_d e^{\lambda_d}), \tag{56}$$

and we compute

$$\begin{aligned}
\Lambda^*(q_1, \ldots, q_d) &= q_1 \log \frac{q_1}{p_1} + \cdots + q_d \log \frac{q_d}{p_d} + (q_1 + \cdots + q_d)\Lambda - \Lambda && (57)\\
&= q_1 \log \frac{q_1}{p_1} + \cdots + q_d \log \frac{q_d}{p_d} + \Lambda - \Lambda && (58)\\
&= q_1 \log \frac{q_1}{p_1} + \cdots + q_d \log \frac{q_d}{p_d} && (59)\\
&= H((q_1, \ldots, q_d)|(p_1, \ldots, p_d)), && (60)
\end{aligned}$$

the relative entropy of $(q_1, \ldots, q_d)$ with respect to $(p_1, \ldots, p_d)$. Note that $H((q_1, \ldots, q_d)|(p_1, \ldots, p_d)) \geq 0$, with equality holding if and only if $q_1 = p_1, \ldots, q_d = p_d$.

Thus $(p_1, \ldots, p_d)$ is the most likely realization, with other realizations exponentially unlikely; their unlikelihood is measured by the rate function Λ^*. And this rate function is indeed Shannon's relative entropy. This statement is *Sanov's theorem* for a finite alphabet; it also holds true for continuous distributions.

Sanov's Theorem: Let $X_1, X_2, \ldots$ be i.i.d real valued random variables with common distribution μ, and let

$$\nu_n = \frac{1}{n}(\delta_{X_1} + \cdots + \delta_{X_n}) \tag{61}$$

be the empirical distribution of $X_1, \ldots, X_n$, which is a random probability measure on $\mathbb{R}$. Then $\{\nu_n\}$ satisfies a large deviation principle with rate function $I(\nu) = S(\nu, \mu)$ (called the "relative entropy") given by

$$I(\nu) = \begin{cases} \int p(x) \log(p(x)) d\mu(x), & \text{if } \nu = p\mu \\ +\infty, & \text{otherwise .} \end{cases} \tag{62}$$

Concretely, this means the following. Consider the set $\mathcal{M}$ of probability measures on $\mathbb{R}$ with the weak topology (which is a metrizable

topology, e.g. by the Lévy metric). Then for closed F, open G in $\mathcal{M}$ Sanov's theorem yields

$$\limsup_{n\to\infty} \frac{1}{n} \log P(\nu_n \in F) \leq -\inf_{\nu\in F} S(\nu, \mu) \tag{63}$$

$$\liminf_{n\to\infty} \frac{1}{n} \log P(\nu_n \in G) \geq -\inf_{\nu\in G} S(\nu, \mu). \tag{64}$$

3. Back to Random Matrices

Consider again the space $\mathcal{H}_N$ of Hermitian matrices equipped with the probability measure P_N having density

$$dP_N(A) = \text{const} \cdot e^{-\frac{N}{2}\operatorname{Tr}(A^2)} dA. \tag{65}$$

The eigenvalue distribution $\tilde{P}_N$ on $\mathbb{R}^N$ is defined by

$$\tilde{P}_N(B) := P_N\{A \in \mathcal{H}_N : (\lambda_1(A), \ldots, \lambda_N(A)) \in B\}. \tag{66}$$

One knows that $\tilde{P}$ is absolutely continuous with respect to Lebesgue measure on $\mathbb{R}^N$ and has density

$$d\tilde{P}_N(\lambda_1, \ldots, \lambda_N) = C_N \cdot e^{-\frac{N}{2}\sum_{i=1}^N \lambda_i^2} \prod_{i<j} (\lambda_i - \lambda_j)^2 \prod_{i=1}^N d\lambda_i, \tag{67}$$

where

$$C_N = \frac{N^{N^2/2}}{(2\pi)^{N/2} \prod_{j=1}^N j!}. \tag{68}$$

We want to establish a large deviation principle for the empirical eigenvalue distribution

$$\mu_A = \frac{1}{N}(\delta_{\lambda_1(A)} + \cdots + \delta_{\lambda_N(A)}) \tag{69}$$

of a random matrix in $\mathcal{H}_N$.

Heuristics for the rate function are as follows. We have

(70)

$$P_N(\mu_A \approx \nu) = \tilde{P}_N(\frac{1}{N}(\delta_{\lambda_1(A)} + \cdots + \delta_{\lambda_N(A)}) \approx \nu)$$

$$= C_N \cdot \int_{\{\frac{1}{N}(\delta_{\lambda_1(A)} + \cdots + \delta_{\lambda_N(A)}) \approx \nu\}} e^{-\frac{N}{2}\sum \lambda_i^2} \prod_{i<j} (\lambda_i - \lambda_j)^2 \prod_{i=1}^N d\lambda_i. \tag{71}$$

Now for $\frac{1}{N}(\delta_{\lambda_1(A)} + \cdots + \delta_{\lambda_N(A)}) \approx \nu$,

$$-\frac{N}{2}\sum_{i=1}^N \lambda_i^2 = -\frac{N^2}{2}\frac{1}{N}\sum_{i=1}^N \lambda_i^2 \tag{72}$$

is a Riemann sum for the integral $\int x^2 d\nu(x)$. Moreover

$$\prod_{i<j}(\lambda_i - \lambda_j)^2 = \exp(\sum_{i<j} \log|\lambda_i - \lambda_j|^2) = \exp(\sum_{i\neq j} \log|\lambda_i - \lambda_j|) \tag{73}$$

is a Riemann sum for $N^2 \int\int \log|x-y| d\nu(x) d\nu(y)$.

Hence, heuristically, we expect that

$$P_N(\mu_A \approx \nu) \sim e^{-N^2 I(\nu)}, \tag{74}$$

with

(75)

$$I(\nu) = -\int\int \log|x-y| d\nu(x) d\nu(y) + \frac{1}{2}\int x^2 d\nu(x) - \lim_{N\to\infty} \frac{1}{N^2} \log C_N.$$

The value of the limit can be explicitly computed as 6/8.

This argument is made rigorous in the following theorem of Ben-Arous and Guionnet from 1997.

Put

$$I(\nu)) = -\int\int \log|x-y| d\nu(x) d\nu(y) + \frac{1}{2}\int x^2 d\nu(x) - 6/8. \tag{76}$$

Then:

(1) $I : \mathcal{M} \to [0,\infty]$ is a well-defined, convex, good function on the space of real probability measures. It has unique minimum value 0 which occurs at the Wigner semicircle distribution μ_W.
(2) The empirical eigenvalue distribution satisfies a large deviation principle with respect to $\tilde{P}_N$ with rate function I : for any open G and closed F in $\mathcal{M}$
   - $\liminf_{N\to\infty} \frac{1}{N^2} \log \tilde{P}_N(\frac{\delta_{\lambda_1}+\cdots+\delta_{\lambda_N}}{N} \in G) \geq -\inf_{\nu\in G} I(\nu)$.
   - $\limsup N \to \infty \frac{1}{N^2} \log \tilde{P}_N(\frac{\delta_{\lambda_1}+\cdots+\delta_{\lambda_N}}{N} \in F) \leq -\inf_{\nu\in F} I(\nu)$.