ICFP M2 - Statistical physics 2
Exam

Giulio Biroli and Guilhem Semerjian

June 10th 2016

The exam is made of two independent problems.

The mark will give roughly an equal weight to both.

Please write your answers to the two problems on separate pages.

No document nor calculator is allowed.

1 A variant of the Random Energy Model (REM)

We consider a variant of the Random Energy Model studied in one of the problem classes: this statistical mechanics model has $M = 2^N$ configurations $\underline{\sigma}$ indexed by N Ising spins, $\underline{\sigma} = (\sigma_1, \ldots, \sigma_N) \in \{-1, 1\}^N$. The energy of each configuration is denoted $H(\underline{\sigma})$, the system is in equilibrium with an heat bath of inverse temperature β, hence the partition function is $Z(\beta) = \sum_{\underline{\sigma}} e^{-\beta H(\underline{\sigma})}$.

The energies of the 2^N configurations are taken as independent identically distributed random variables, with a law defined by its density of probability on $]-\infty, +\infty[$ as

$$\rho(E) = C\, e^{-N^x |E|^\delta} \ ,$$

where C is a normalization constant, $\delta > 0$ an exponent defining the model, and x another exponent to be determined in the following.

The goal of the problem is to determine the quenched free-energy density

$$f_{\rm q}(\beta) = -\frac{1}{\beta} \lim_{N\to\infty} \frac{1}{N} \mathbb{E}[\ln Z(\beta)] \ ,$$

where $\mathbb{E}[\bullet]$ denotes the average over the random energies of the 2^N configurations, and to study the phase transition that occurs when β is varied.

1. What is the value of δ in the usual version of the REM with Gaussian random energies ?

2. Denote $\mathcal{N}(u, {\rm d}u)$ the random variable counting the number of configurations $\underline{\sigma}$ with intensive energies in a small interval of length ${\rm d}u$ around u, i.e. with $H(\underline{\sigma}) \in [Nu, N(u + {\rm d}u)]$. Describe the law of this random variable, and express its average value at the leading exponential order.

3. What is the value of x that ensures a good thermodynamic limit ? (i.e. such that the leading order of $\mathbb{E}[\mathcal{N}]$ is exponential in N and depends non trivially on u for finite u). In all the following you will assume that this choice of x is made.

4. Recall briefly the reasoning explained in the problem class to justify that the typical value of $\mathcal{N}$ is, at the leading exponential order,

$$\mathcal{N}_{\rm typ}(u, {\rm d}u) = \begin{cases} e^{N s_{\rm m}(u)} {\rm d}u & \text{if } \ u \in [-u_{\rm c}, u_{\rm c}] \\ 0 & \text{otherwise} \end{cases} \ , \qquad \text{with } \ s_{\rm m}(u) = \ln 2 - |u|^\delta \ .$$

   Give the value of $u_{\rm c}$.

5. Draw the shape of the function $s_{\rm m}(u)$, distinguishing carefully the cases $\delta > 1$ and $0 < \delta < 1$.

6. What is the groundstate energy density of the model ?

7. Justify that the quenched free-energy of the model is then given by the Legendre transform of $s_{\rm m}(u)$, namely
$$f_{\rm q}(\beta) = -\frac{1}{\beta} \sup_{u \in [-u_{\rm c}, u_{\rm c}]} \left[-\beta u + s_{\rm m}(u) \right] \ . \tag{1}$$

8. Consider first the case $\delta > 1$.

   (a) Draw again the shape of $s_{\rm m}(u)$, and add on the figure the graphical interpretation of the Legendre transform for different values of β (you should first interpret the value of $-\beta u + s_{\rm m}(u)$ for an arbitrary u, then explain how to maximize this quantity).

   (b) Show that the model exhibits a phase transition qualitatively similar to the Gaussian case studied in the problem class, and give the equation fixing the transition temperature $\beta_{\rm c}$.

   (c) Draw the shape of the energy and entropy as a function of the temperature.

   (d) What is the thermodynamic order of the transition ?

9. Consider now the case $0 < \delta < 1$. Study similarly the behavior of the thermodynamic quantities, and characterize the order and critical temperature of the phase transition the model undergoes. Is there a qualitative difference compared to the case $\delta > 1$?

10. A further variation of the REM has been studied in the context of the protein folding problem: in addition to the 2^N configurations with random energies one assumes the existence of an additional configuration of energy Nu_0, with $u_0 < -u_{\rm c}$. This configuration is interpreted as the perfectly folded conformation of a protein, all the other ones being associated to partially folded conformations.

    (a) Modify the formula (1) to obtain the quenched free-energy in presence of this additional configuration.

    (b) Assuming $\delta > 1$ repeat the graphical study of the Legendre transform with this additional state.

    (c) Explain graphically that there is a phase transition in this model, and give the thermodynamic order of the transition.

2 Marčenko-Pastur law for random correlation matrices

2.1 Introduction and notations

In the following we shall obtain the density of eigenvalues for random correlation matrices.
These matrices appear in many applications that involve time series of data. Henceforth we denote the set of data ξ_i^t where $t = 1, \ldots, T$ is an integer index related to time (e.g. day one, day two, ..., day T) and $i = 1, ..., N$ is an integer index related to the observable which is measured.
We assume that the ξ_i^t are independent and identically distributed random variables with zero mean and variance equal to one: $\langle \xi_i^t \xi_j^{t'} \rangle = \delta_{i,j}\delta_{t,t'}$.
Their true covariance matrix is therefore $C_{i,j} = \langle \xi_i^t \xi_j^t \rangle = \delta_{i,j}$; its spectrum consists in N eigenvalues equal to one.
In applications one tries to obtain information on the true covariance matrix using the empirical estimation:

$$E_{i,j} = \frac{1}{T} \sum_{t=1}^{T} \xi_i^t \xi_j^t \ .$$

Often the number of observables, N, and the number of observations, T, are both large and comparable. In the following we shall study the density of eigenvalues of the matrix E for $N \to \infty$, $T \to \infty$ with the ratio $\frac{N}{T} = q$ fixed. This density is defined as

$$\rho(\lambda) = \frac{1}{N} \sum_{i=1}^{N} \delta(\lambda - \lambda_i) \ ,$$

where $\lambda_1, \ldots, \lambda_N$ denote the eigenvalues of E.
We recall that the resolvent matrix is defined as $G(z) = (E - z\mathbb{1})^{-1}$. It is also useful to define the reduced resolvent matrix $G^{t'}(z) = (E^{t'} - z\mathbb{1})^{-1}$ which is constructed using the matrix

$$E_{i,j}^{t'} = \frac{1}{T} \sum_{\substack{t=1 \\ t \neq t'}}^{T} \xi_i^t \xi_j^t = E_{i,j} - \frac{1}{T} \xi_i^{t'} \xi_j^{t'} .$$

We shall also use the notation:

$$g(z) = \frac{1}{N} \mathrm{Tr}\, G(z) \qquad ; \qquad g^{t'}(z) = \frac{1}{N} \mathrm{Tr}\, G^{t'}(z)$$

for the reduced trace of the resolvents.

2.2 Self-consistent equation on the resolvent

1. Starting from the identity

$$\sum_{i,j} \xi_i^{t'} G_{i,j}^{t'}(z) (G(z)^{-1})_{j,l} = \sum_{i,j} \xi_i^{t'} G_{i,j}^{t'}(z) \left(\frac{1}{T} \xi_j^{t'} \xi_l^{t'} + E_{j,l}^{t'} - z\delta_{j,l} \right)$$

show that

$$\sum_{i,j} \xi_i^{t'} G_{i,j}^{t'}(z) (G(z)^{-1})_{j,l} = \xi_l^{t'} + \frac{1}{T} \left(\sum_{i,j} \xi_i^{t'} G_{i,j}^{t'}(z) \xi_j^{t'} \right) \xi_l^{t'} \ .$$

2. Using the previous result show that

$$\sum_i \xi_i^{t'} G_{i,k}^{t'}(z) = \left(1 + \frac{1}{T} \sum_{i,j} \xi_i^{t'} G_{i,j}^{t'}(z) \xi_j^{t'} \right) \sum_i \xi_i^{t'} G_{i,k}(z)$$

and obtain the relation:

$$\frac{1}{T} \sum_{i,j} \xi_i^{t'} G_{i,j}(z) \xi_j^{t'} = \frac{\frac{1}{T} \sum_{i,j} \xi_i^{t'} G_{i,j}^{t'}(z) \xi_j^{t'}}{1 + \frac{1}{T} \sum_{i,j} \xi_i^{t'} G_{i,j}^{t'}(z) \xi_j^{t'}} \ .$$

3. Obtain the identity:

$$\frac{1}{N}\sum_{i,j}\left(E_{i,j}-z\,\delta_{i,j}\right)G_{j,i}(z)=1$$

and show that

$$zg(z)=-1+\frac{1}{N}\sum_{t'=1}^{T}\frac{\frac{1}{T}\sum_{i,j}\xi_i^{t'}G_{i,j}^{t'}(z)\xi_j^{t'}}{1+\frac{1}{T}\sum_{i,j}\xi_i^{t'}G_{i,j}^{t'}(z)\xi_j^{t'}}\ . \tag{2}$$

4. Compute the average and the variance of $\frac{1}{T}\sum_{i,j}\xi_i^{t'}G_{i,j}^{t'}(z)\xi_j^{t'}$ for fixed $G_{i,j}^{t'}(z)$ (i.e. averaging only on the vector $\xi^{t'}$), assuming that $\xi_i^{t'}$ are Gaussian i.i.d. random variables with zero mean and unit variance (this assumption is just to simplify the analysis, the results that you will find below hold generically).

5. Assuming, as it is indeed the case, that $\mathrm{Tr}[(G^{t'}(z))^2]$ scales as N and $g^{t'}(z)\approx g(z)$ in the large N,T limit, argue why $\frac{1}{T}\sum_{i,j}\xi_i^{t'}G_{i,j}^{t'}(z)\xi_j^{t'}$ can be approximated by $qg(z)$.

6. Starting from equation (2) show that $g(z)$ verifies the self-consistent equation:

$$zqg(z)=-q+1-\frac{1}{1+qg(z)}\ .$$

7. Solve the equation and obtain that

$$g(z)=\frac{-(z+q-1)\pm\sqrt{(z+q-1)^2-4zq}}{2zq}\ .$$

2.3 From the resolvent to the density of eigenvalues

1. Recall the relationship between $g(z)$, evaluated for $z=\lambda+i0^+$, and the density of eigenvalues $\rho(\lambda)$ (λ denotes a real number).

2. Compute the values λ_- and λ_+ such that for $\lambda_-<\lambda<\lambda_+$ the trace of the resolvent $g(z)$ acquires an imaginary part when $z=\lambda+i0^+$.

3. Using the behavior of the resolvent for $z\to\pm\infty$, argue why one has to chose the plus sign in front of the square root when z is real and larger than λ_+ and the minus sign when z is real and smaller than λ_-.

4. Obtain explicitly $\rho(\lambda)$ for $q<1$.

5. What is the support of $\rho(\lambda)$ for $q<1$? Obtain its limit for $q\to 0$ and justify the result on general grounds.

6. *Facultative question.* Show that for $q>1$ the rank of the N by N matrix E is at most T. Using this result discuss why one expects that the matrix E admits at least $N-T$ zero eigenvalues for $q>1$. This leads for $q>1$ to an extra contribution, equal to $(1-\frac{1}{q})\delta(\lambda)$, to the density of eigenvalues.

7. *Facultative question.* Find the previous result directly from $g(z)$.

8. In conclusion, we have found that despite the fact that the density of eigenvalues of the true covariance matrix C is a delta function centred in one (C is the identity), its empirical estimation E has a non-trivial distribution of eigenvalues $\rho(\lambda)$. It is called the Marčenko-Pastur law in honour of the researchers who first obtained it.
Show that in the large T limit, at fixed $q=N/T$, the empirical estimation of each element of the matrix $C_{i,j}$ by $E_{i,j}$ is correct, i.e. $C_{i,j}-E_{i,j}\to 0$. Despite this convergence element per element, the density of eigenvalues of the two matrices are different. Give a reason for this result which is counterintuitive at first sight.

2.4 Conclusion

Skip it during the exam, read it after. We have shown that for $q = N/T$ not too large (as it is the case in many applications), even in absence of any correlation between the time series, the density of eigenvalues estimated from empirical data is non-trivial.
Often the density of eigenvalues of C contains important information on the problem at hand, i.e. C is not the identity and encodes the correlations between the observables. It is then crucial to make the difference between noise and information in its empirical estimation. The Marčenko-Pastur law provides a benchmark; one crude but very effective way to find out what is the the information part is to exclude from the estimation of the density of eigenvalues all that can be explained using the Marčenko-Pastur law: this would correspond to pure noise, all the rest would then correspond to true correlation. As an example we show in Fig.1 the empirical density of eigenvalues obtained for data taken from the stock market: ξ_i^t is the relative change of the price of stock i over one day.
Making the difference between noise and information in real empirical correlation matrices is an active research domain relevant for biology and finance, which is at the boundary between statistical physics and probability theory.

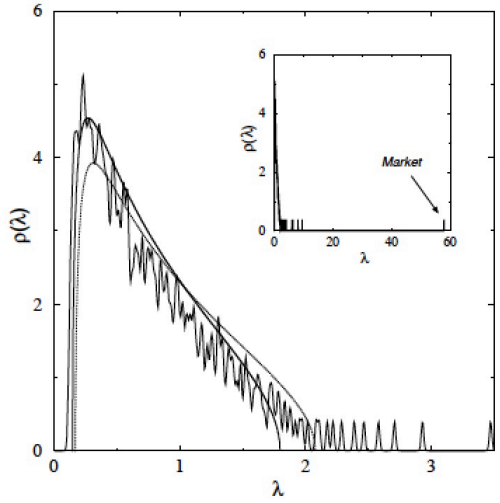

Figure 1: Smoothed density of the eigenvalues of E, where the correlation matrix E is extracted from N = 406 stocks of the Standard and Poor 500 during the years 1991-1996. For comparison we have plotted the Marčenko-Pastur law (dotted line) obtained assuming that the matrix is purely random except for its highest eigenvalue, called market mode. A better fit can be obtained assuming that not only the market mode is non-random but also a finite fraction of the highest eigenvalue. From Laloux et al. Physical Review Letters Vol. 83, 1467 (1999).