Statistical Physics of Computation 2025 - Exercises

Vittorio Erba, Emanuele Troiani

Week 4

4.1 Getting acquainted with random matrices

In this exercise, we will explore numerically the spectral properties of random Gaussian matrices

$$G_{ij} = \frac{Z_{ij} + Z_{ji}}{\sqrt{2d}}\,, \tag{1}$$

where $Z_{ij} \sim N(0,1)$ independently for each $1 \leq i,j \leq d$. This is important in the context of the spiked-Wigner model

$$y_{ij} = \sqrt{\frac{\lambda}{d}} x_i^* x_j^* + \xi_{ij} \tag{2}$$

that we defined in lecture 4. Indeed, for zero signal-to-noise ratio $\lambda = 0$, the observation is give by pure noise $y = \xi$, and ξ is proportional to the random Gaussian matrix G defined above[1]. Thus, understanding how G behaves means understanding one of the two boundary cases for the model we are studying (the other boundary case being $\lambda \gg 1$, where y is proportional a projector on the line containing x^*).

We start by studying the spectrum of G.

1. Write a code in your language of choice that generates a $d \times d$ instance of the random matrix G. Then, compute the eigenvalues of such random matrix for $d = 1000$ and plot their histogram (you may need to play with the number of bins as a function of d to obtain a nice plot). Repeat the procedure for many different instances of G at the same dimension, and at different dimensions d. What do you observe?

   We provide a simple solution in the attached notebook, and give here just the resulting figure. We observe that, modulo differences in binning and some minor variations, all the histograms look roughly the same. They have a roughly semicircular shape, with all eigenvalues bounded roughly between $(-2, 2)$, and the instance-to-instance variations become smaller at larger dimensions.

The phenomenon you observed above is called concentration of the spectrum. As $d \to +\infty$, the empirical spectral density

$$\rho(x) = \frac{1}{d}\sum_{i=1}^{d} \delta(x - \lambda_i(G)) \tag{3}$$

converges to a deterministic shape that is independent on the actual realization of G, where $\lambda_i(G)$ is the i-th eigenvalue of G. Notice that $\rho(x)$ is precisely the histogram you plotted before,

[1] Modulo a slightly different variance on the diagonal, but that plays no role in the limit $d \to \infty$, as the diagonal contains $O(d)$ elements of order $O(1)$, while the rest of the matrix contains $O(d^2)$ elements of order $O(1)$.

in the limit of very small bins (taken after sending $d \to \infty$). This implies, in particular, that the leading eigenvalue (i.e. the largest one) also concentrates to a deterministic value (roughly equal to 2 from you experiments above).

Notice also that the spectrum is composed by $O(d)$ numbers, while the original matrix by $O(d^2)$ numbers: in a sense, the spectrum is a macroscopic quantity when compared with the number of original degrees of freedom.

Finally, remark that if we did not divide the entries of G by $\sqrt{d}$, the spectrum would not have a nice limiting shape, as the boundaries of the histogram would grow proportionally to $\sqrt{d}$ (feel free to check this with your implementation!).

2. Consider now the case in which $Z_{ij} \sim P$ independently for each $1 \leq i, j \leq d$, for several examples of P with mean zero and variance equal to one. Repeat point 1. What do you observe? For example, you can consider

   - P is the uniform distribution between $(-\sqrt{3}, \sqrt{3})$.
   - P is the uniform distribution over $\{-1, +1\}$, i.e. a sample equals $+1$ with probability $1/2$, and vice-versa.

   We provide a simple solution in the attached notebook, and give here just the resulting figure. We observe that the spectrum concentration phenomenon, as well as the actual limiting spectrum, are the same as in point 1.

The phenomenon you observed above is called universality, i.e. the fact that high-dimensional deterministic properties, such as the spectrum we are studying here, do not depend too much on the properties of the microscopic degrees of freedom. Here, as long as the first two cumulants of P are the same, we obtain the same spectra.

The matrix G, in the general case of i.i.d. entries (modulo the symmetry) with zero mean and variance one, is called a *Wigner* matrix, explaining the nomenclature of the spiked-Wigner model.

It's natural to also expect a similar concentration effect to happen for eigenvectors. This is however more rare, and we can probe why in the Gaussian case $P = N(0, 1)$.

3. Show that if $P = N(0, 1)$, then G is rotationally invariant, meaning that the distribution $P(G)$ of the matrix G will not change under the action of an arbitrary rotation matrix O, or in formulas $P(G) = P(O^\top G O)$.

   The distribution of G is just the product of the distribution of all its independent entries, i.e.

   $$P(G) \propto \exp\left\{ -\sum_{i=1}^{d} \frac{G_{ii}}{4} - \sum_{i \leq j} \frac{G_{ij}}{2} \right\} = \exp\left\{ -\frac{\mathrm{Tr}[G^2]}{4} \right\} \tag{4}$$

   where the doubled variance on the diagonal is due to how we symmetrize the matrix Z to obtain G. Notice in particular that the trace of a power of a matrix is invariant under rotations, giving the result.

4. Using the previous point, argue that if $G = UDU^T$ is one eigen-decomposition of G, with U the matrix of eigenvectors and D the diagonal matrix of eigenvalues, then U is distributed uniformly over the space of rotation matrices in d dimensions.

   Given that $P(G) = P(OGO^T)$, then we see that conditioned on the eigenvalues D, all matrices of the form UDU^T with U any rotation matrix have the same probability. Hence, for any possible realization of the eigenvalues D, the eigenvectors form a uniformly distributed

random orthonormal basis of $\mathbb{R}^d$, implying that the eigen-vectors are distributed uniformly over the space of rotation matrices in d dimensions.

5. Can the eigenvectors of G concentrate?

   No, they cannot, as they are not constrained in any way in the probability distribution, which is uniform over all possible sets of orthonormal eigenvectors. As a matter of fact, in this example one can write the distribution of G as a function only the spectrum, making this fact even more clear.

It's not simple to show it, but all the examples we discussed above for generic P are approximately rotationally invariant in the limit of large d.

The moral of the story here is that in what follows, we can avoid averaging over the realization of G (i.e. over the noisy observation y of x^*) whenever we just look at eigenvalues, as eigenvalues behave deterministically for large enough d. When looking instead at eigenvectors, we should be more careful.

4.2 The spectrum of the spiked-Wigner model

Let us go back to the spiked-Wigner model (rescaled by $\sqrt{d}$ in order to have a limiting spectrum for the noise matrix as d increases)

$$Y_{ij} = \frac{\sqrt{\lambda}}{d} x_i^* x_j^* + G_{ij} \tag{5}$$

and consider $x_i^* \sim N(0,1)$ independently for $i = 1, \dots, d$.

5. Consider the spike term first, i.e.

$$S_{ij} = \frac{\sqrt{\lambda}}{d} x_i^* x_j^* \,. \tag{6}$$

   Show that the operator norm of S (i.e. its largest eigenvalue) satisfies $||S||_{\text{op}} \to \sqrt{\lambda}$ when $d \gg 1$. This is an important check to perform: if the $||S||_{\text{op}}$ diverges or vanishes as $d \to +\infty$, our signal would be either so strong to be clearly visible in the spectrum, or so small to be surely hidden in the spectrum.

   Given that S is a rank-1 spike, it has only one non-zero eigenvalue with eigenvector proportional x^* given by

$$||S||_{\text{op}} = \sqrt{\lambda} ||x^*||^2 / d \to \sqrt{\lambda} \,. \tag{7}$$

   The norm term comes from the fact that x^* is not normalized by itself, and by the central limit theorem it equals its mean, i.e. one.

6. How do you expect the spectrum of Y to look like at small and big values of λ?

   We need to study the spectrum of a rank-1 perturbation plus a Wigner matrix. For small values of λ, the spectrum should be given roughly by the spectrum of a Wigner matrix, as the strength of the spike is small. For large values of λ, the spectrum is dominated by the spike, so it should feature a leading eigenvalues $\approx \sqrt{\lambda}$ with eigenvector $\approx x^*$, plus a bulk of eigenvalues of order $O(1) \ll O(\sqrt{\lambda})$ coming from the noise in the space orthogonal to x^*.

7. Write a code in your language of choice that generates a $d \times d$ instance of the random matrix Y, for a given $\lambda \geq 0$ and d. Plot the histogram of the spectrum at $d = 1000$ for several values of λ, going from $\lambda = 0.1$ to $\lambda = 10$. What do you observe?

   The histogram of the spectrum looks exactly the same as in the non-spiked case for small enough values of λ. For large values of λ, it also looks the same, but with an additional outlier out of the main bulk.

8. Write a code in your language of choice that generates a $d \times d$ instance of the random matrix Y, for a given $\lambda \geq 0$ and $d = 2000$, and then plot the value of the 2 largest eigenvalues as a function of λ. Around which value of λ does the outlier first pop-out of the bulk?

   Around $\lambda = 1$.

The phenomenon you are observing is an important phase transition in Random Matrix Theory, the so-called Baik–Ben Arous–Péché (BBP) transition, in which a rank-1 perturbation to a Wigner matrix either does not affect the bulk, or pops out of the bulk sharply at a specific value of λ.

4.3 BBP transition and the inference problem

What are the consequences for our learning problem, i.e. the retrieval of x^* without knowledge of it other than Y and the generative model of Y? It really seems that the outlying eigenvalue/eigenvector should have at least something to do with it, as its presence and position correlates with λ growing.

9. Write a code in your language of choice that generates a $d \times d$ instance of the random matrix Y, for a given $\lambda \geq 0$ and d, and then compute the cosine similarity

$$\frac{|\hat{x}^\top x^*|}{\|\hat{x}\|\|x^*\|} \tag{8}$$

   between x^* and the leading eigenvector $\hat{x}$ of Y. What do you observe?

   For small λ the cosine similarity is essentially zero, for large values if grows to 1. We notice that there seems to be a second order phase transition, happening around $\lambda = 1$. This is also roughly the value of λ for which the leading eigenvector comes out of the bulk.

10. In the previous point, we computed the performance of an explicit estimator, the leading eigenvector of Y. It is truly an estimator, as it is a function that only takes Y as input and outputs a candidate signal $\hat{x}$. Can we say that this the Bayes-optimal estimator?

    No. First of all, we need to specify a metric w.r.t. which an estimator is Bayes-optimal, which we did not specify here. Secondly, even if we did specify a metric, we would not be able to say whether this estimator is optimal or not! We can only say that, on average over the realization of x^*, Y, the BO estimator must retrieve x^* at least as good as the estimator we considered here.

11. Wait a minute! In the previous exercise we showed that the eigenvectors of G do not concentrate, so here we should have been more careful, and we should have averaged over several instances of x^* and y! Do that, i.e. compute the cosine similarity averaged over some realizations of x^* and y, and observe that the cosine similarity of the single instance is the same as the averaged one. How do you explain this?

Only the eigenvectors of G are uniformly random. As soon as we perturb G with a spike in the direction of x^* to obtain Y, the eigenvectors of Y become a priori rotationally invariant only in the subspace orthogonal to x^*. Hence, when projecting $\hat{x}$ onto x^*, we can have a cosine similarity that concentrates (self-averages). The component of $\hat{x}$ orthogonal to x^*, if non-zero, will then be uniformly random.

To conclude, in this exercise we explored the concepts of concentration and universality for the spectra of random matrices, and their relationship with the inference problem we are studying in the lecture. Notice that it is non-trivial to compute the shape of the limiting spectral distribution of Ex 4.1, the phase transition location of Ex 4.2, and the cosine similarity of Ex 4.3, but all three can be analytically computed in the limit $d \gg 1$. We will see how in the following lectures.