Contents

1 Disordered systems, rank-one matrix estimation and Hamilton-Jacobi equations

We consider the problem of estimating a large rank-one matrix, given noisy observations. This inference problem is known to have a phase transition, in the sense that the partial recovery of the original matrix is only possible if the signal-to-noise ratio exceeds a (non-zero) value. We will present a new proof of this fact based on the study of a Hamilton-Jacobi equation. This alternative argument allows to obtain better rates of convergence, and also seems more amenable to extensions to other models such as spin glasses.

1.1 2020/5/18 Lecture 1

Students are assigned one of two dormitories. They put on a sorting hat, which decides which dorm they go in.

The students are $i \in \{1, \dots, N\}$. An assignment is $\sigma \in \{\pm 1\}^N$. The sorting had optimizes the quality of interaction between i and j, J_{ij}. Suppose (J_{ij}) are independent standard Gaussians. The larger J_{ij} is, the more that i and j like to be together. We want to maximize $\sigma \mapsto \sum J_{ij} \mathbb{1}_{\{\sigma_i = \sigma_j\}}$. By a linear transformation this is equivalent to maximizing $\sum J_{ij}\sigma_i\sigma_j$.

What is $\max_{\sigma\{\pm 1\}^N} \sum J_{ij}\sigma_i\sigma_j$ as $N \to \infty$. Because the J_{ij} can be positive or negative, we can't make all the students happy. Thus we can say there are **frustrations** in the problem. These models are called spin glasses in the literature. It's difficult to find the optimum: making local moves, you may have to decrease the objective before increasing it.

I want to consider a softer version of the maximum. We look at

$$\mathbb{E}\frac{1}{N}\log \sum_{\sigma\in\{\pm 1\}^N} \exp\left(\frac{\beta}{\sqrt{N}} \sum_{i,j=1}^{M} J_{ij}\sigma_i\sigma_j\right)$$

If β is large this is dominated by the maximum. This is like a relaxation of the problem. We should expect what's inside is order N, so we divide by N.

Parisi in the late 70's (1979) proposed an answer for what this becomes as $N \to \infty$. It's a fairly complicated formula.

Guerra 03 and Talagrand 06 proved it rigorously. I find it mysterious; I want to think about a slight variation of the problem. Instead of connections between each i, j, think of them organized in two layers; there are interactions between but not within the layers (the graph is bipartite). This seems an innocent modification, but I could not understand what to write instead of the complicated formula!

This is called the spin-glass model.

Now I consider rank-one matrix esimation/inference. The question is statistical: we only observe a noisy version of a rank-one matrix. Can we recover information about it?

A concrete setting: you are Netflix, you want to make recommendations for your customers. A simple model is that whether or not a person likes a movie is captured by a few parameters of the movie (action, introspection, sad/happy, etc.) and customer, and is a linear function of the parameters. Then you have a large low-rank matrix. A simplification is that it's a matrix of rank 1. I'll describe rank 1, but it's not hard to generalize.

Another example is community detection. The US is polarized, so to guess whether two people will be friends, maybe there is a binary variable that will predict this.

The common thread is the relation with certain partial differential equations, called Hamilton-Jacobi equations.

The Curie-Weiss model is a simple model that can be solved in many ways. I want to emphasize the method that uses intuition with Hamilton-Jacobi equations. Next when we turn to rank-1 matrix estimation, the proof will be almost the same.

Our derivation is not standard; if you want to see a more standard derivation see Friedli and Velenik `https://www.unige.ch/math/folks/velenik/smbook/index.html`.

QA:

- Can you meaningfully recover information about the rank-1 matrix? In the Ising model there is a phase transition between an ordered and disordered state. In this inference problem there is also a phase transition. When signal-to-noise ratio is too small (weak), you cannot recover meaningful information. After the threshold, you can recover partial information.

- Can we fix the number of $+1$'s and -1's? You can change the reference measure; this can be encoded as changing the reference measure.

  $\beta = 0$ is summing over reference measure. $\beta \to \infty$ recovers the maximum. β small is high temperature, β large is small temperature.

1.1.1 Definitions

We want to study the probability measure that to each $\sigma \in \{\pm 1\}^N$, associates a weight proportional to

$$\exp\left(\frac{t}{N}\sum_{i,j=1}^{N}\sigma_i\sigma_j + h\sum_{i=1}^{N}\sigma_i\right).$$

The second term doesn't have interaction; it "tilts" the σ_i, giving each a preference. Here $t > 0$ but $h \in \mathbb{R}$. Define the expected value

$$\langle f(\sigma)\rangle_{k,h} := \frac{\sum_\sigma f(\sigma)\exp\left(\frac{t}{N}\sum_{i,j=1}^{N}\sigma_i\sigma_j + h\sum_{i=1}^{N}\sigma_i\right)}{\sum_\sigma \exp\left(\frac{t}{N}\sum_{i,j=1}^{N}\sigma_i\sigma_j + h\sum_{i=1}^{N}\sigma_i\right)}.$$

The subscripts are omitted when clear.

Define the free energy

$$F_N(t,h) = \frac{1}{N}\log\sum_\sigma \exp\left(\frac{t}{N}\sum_{i,j=1}^N \sigma_i\sigma_j + h\sum_{i=1}^N \sigma_i\right)$$

You might say: this is the normalization constant, we care about the measure. This is misleading because the normalization constant is the generating function of quantities we care about. You are calculating the exponential (moment) generating function of these variables. If you understand the mgf, you understand these quantities.

Moment generating function Differentiating gives

$$\partial_h F_N = \frac{1}{N}\frac{\sum_\sigma \left(\sum_{i=1}^N \sigma_i\right)\exp\left(\frac{t}{N}\sum_{i,j=1}^N \sigma_i\sigma_j + h\sum_{i=1}^N \sigma_i\right)}{\sum_\sigma \exp\left(\frac{t}{N}\sum_{i,j=1}^N \sigma_i\sigma_j + h\sum_{i=1}^N \sigma_i\right)} = \frac{1}{N}\left\langle \sum_i \sigma_i \right\rangle \tag{1}$$

$$\partial_t F_N = \frac{1}{N}\left\langle \frac{1}{N}\sum \sigma_i\sigma_j \right\rangle = \left\langle \left(\frac{1}{N}\sum \sigma_i\right)^2 \right\rangle. \tag{2}$$

This model is simple; I can rewrite $\partial_t F_N$ in a simple way. The derivatives are all order 1. It's a good starting point to notice that

$$\partial_t F_N - (\partial_h F_N)^2 = \left\langle \left(\frac{1}{N}\sum \sigma_i\right)^2 \right\rangle - \left\langle \frac{1}{N}\sum \sigma_i \right\rangle^2.$$

This is the mean magnetization, the variance of the magnetization. Idea: The variance is lower-order, so as $N\to\infty$, F_N solves the equation with 0 on the right.

F_N is the mgf of $\sum \sigma_i$. So in particular it should encode the variance of the variable in some way. I should find a way to express it in terms of F_N. Looking at the second derivative is a good idea.

$$\partial_h^2 F_N = \frac{1}{N}\left\langle \left(\sum \sigma_i\right)^2 \right\rangle - \frac{1}{N}\left(\left\langle \sum \sigma_i \right\rangle\right)^2.$$

So we have shown

$$\partial_t F_N - (\partial_h F_N)^2 = \frac{1}{N}\partial_h^2 F_N. \tag{3}$$

This is a very important observation: everything is expressed in terms of F_N. We can forget about the probability measure, definition in terms of probability measures, and just think about what F_N satisfies this equation, and what happens when N becomes large. It also suggests that as $N\to\infty$, the RHS will vanish.

I think of this as an evolution equation; think of t as time. It will be useful to understand

what happens when $t = 0$, the initial conditions.

$$F_N(0,h) = \frac{1}{N}\log\sum_\sigma\sum_\sigma \exp\left(h\sum_i \sigma_i\right) \tag{4}$$

$$= \frac{1}{N}\log\sum_\sigma \prod_{i=1}^N \exp(h\sigma_i) \tag{5}$$

$$= \frac{1}{N}\log\left(e^h + e^{-h}\right)^N \tag{6}$$

$$F_N(0,h) = F_1(0,h) =: \psi(h). \tag{7}$$

This does not depend on N.

The most important connection is that the h-derivative is the mean magnetization (1).

What do we do with (3) and (7)?

1.1.2 Interlude on Hamilton-Jacobi equation

Let's take a step back and think about what the equation is saying. We need to thin kabout what it means to be a solution of

$$\partial_t f - (\partial_h f)^2 = 0. \tag{8}$$

The first thing to look for is a C^1 function that solves the equation pointwise. What's the problem with this?

The problem is that there is a phase transition in the Ising model. When t is small nothing impressive happens. For fixed small t, $F_N(t,h)$ will be smooth.

But for larger t, if h is positive, then the mean magnetization is positive and away from 0, and if h is tiny negative, then the mean magnetization is negative and away from 0. There will be a jump in the derivative of the function; it looks like $|h|$. The equation is not solved pointwise at $h = 0$.

QA:

- If you change the measure, you can create lots of discontinuities. Considering P with bounded support on $\mathbb{R}$, we can consider

$$\int \exp(\cdots)\, dP^{\otimes N}(\sigma).$$

  You can play with P to create more corners in the limit. This changes what this ψ function.

- What's the notion of convergence for solutions as $N \to \infty$? All functions are uniformly Lipschitz. By Arzela-Ascoli there are convergent subsequences. We can take uniform convergence as the topology. If you prove convergence for some topology, you can bring it to $C^{0,1}$ topology.

- HJ equation can be solved by characteristics. Is there a probabilistic interpretation of the PDE method of characteristics? I'll try to bypass it. Lelarge, Miolane use

different techniques: construct characteristics for finite N. The method I present is more convenient, you don't need to follow characteristics closely. Look at whether characteristics are contracting or expanding.