Contents

These are notes from the Online Open Probability School (OOPS) 2020.

- Webpage for OOPS: `https://www.math.ubc.ca/Links/OOPS/`.
- Source is available at `https://github.com/holdenlee/oops`. Contributions, corrections, and comments welcome; feel free to send a pull request. Any errors are probably due to me! You can ping me on zulip or email me at `holden.lee@duke.edu`.

  A direct link to these notes is `https://www.dropbox.com/s/u4lpayi98r06un5/oops.pdf?dl=0`.

1 Disordered systems, rank-one matrix estimation and Hamilton-Jacobi equations (Jean-Christophe Mourrat)

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.

References:

- Paper: [Mou18], `https://arxiv.org/abs/1811.01432`.
- Book for background material: [FV17], available at `https://www.unige.ch/math/folks/velenik/smbook/index.html`
- Other approaches
  - Lelarge-Miolane: [LM19], `https://arxiv.org/abs/1611.03888`
  - Barbier-Macris: [BM19], `https://arxiv.org/abs/1705.02780`

2020/5/18 Lecture 1

Suppose 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 the **spin glass model**[1]

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

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.[2]

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!

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?

Here are some concrete settings where this is useful:

- 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.

- Community detection: The US is polarized, and there is a binary variable that will predict whether two people will be friends.

[1] Note that the normalization is $\frac{1}{N}$ instead of $\frac{1}{\sqrt{N}}$ because here the $J_{ij} = 1$.

[2] 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.

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 [FV17].

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.

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)?

1.2 Interlude on Hamilton-Jacobi equation

Let's take a step back and think about what the equation is saying. We need to think about 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. Barbier and Macris [BM19] use different techniques: construct characteristics for finite N. The method I present is more convenient, you don't need to follow characteristics closely, just look at whether characteristics are contracting or expanding.

2020/5/19 Lecture 2

Recall that we showed $\partial_t F_N - (\partial_h F_N)^2 = \frac{1}{N}\partial_h^2 F_N$ (3), and hoped the limit object will solve the equation

$$\partial_t F_N - (\partial_h F_N)^2 = 0$$

We have to think about what it means to solve the equation. Because of phase transitions, we don't expect them to be C^1. We need to lower our expectations about what being a solution means.

Every Lipschitz function is differentiable almost everywhere. What if we just ask the equation to be satisfied almost everywhere? The problem is that the solution is not unique (given a initial value at $t = 0$, there are multiple possible solutions). Here are some examples.

- 0 is a solution.

- $(t,h) \mapsto t + h$ or $t - h$.

- From these 3 solutions, we can construct another solution, the tent function

$$f(t,h) = \begin{cases} t+h, & -t \le h \le 0 \\ t-h, & 0 \le h \le t \\ 0, & \text{else.} \end{cases}$$

We have an infinite number of solutions by taking combinations of these.

But these are not the solutions we care about! We know the solution should be convex in the h variable; the tent function is not convex. If we add in this condition, then we get a unique solution.

Definition 1.1: We say that a Lipschitz function $\mathbb{R}_+ \times \mathbb{R} \to \mathbb{R}$ is a **weak solution** of the HJ equation (8) if

1. (HJ) is satisfied almost everywhere.
2. For every $t \ge 0$, the mapping $h \mapsto f(t,h)$ is convex.[3]

Proposition 1.2 (Uniqueness)**:** If f and g are two weak solutions of the HJ equation with $f(0,\cdot) = g(0,\cdot)$, then $f = g$.

Why do we want this? This proposition says that two solutions with same initial condition are equal. What we want is actually a refinement of the statement: we want to compare the actual solution (to (3)) for N large to the solution of the HJ equation. Instead of two solutions being equal, we want to show the almost-solution and the solution are close.

Proof sketch. Denote $w = f - g$. We have that almost everywhere

$$\begin{aligned} \partial_t w &= (\partial_h f)^2 - (\partial_h g)^2 \\ &= \underbrace{(\partial_h f + \partial_h g)}_{=:b} \partial_h w. \end{aligned}$$

Then

$$\partial_t w - b\partial_h w = 0 \text{ (a.e.)}$$

The derivative of b is positive.[4] The rough idea is to look at how the integral $I(t)$ evolves:

$$I(t) = \int w(t,h)\, dh.$$

[3] It suffices to be semi-convex: there is a lower bound on the Hessian.

[4] For higher dimensions, the divergence of this vector field has a sign. This says that the flows of the vector field being convergent or divergent.

Suppose this is well-defined. Then integrating by parts (suppose there are no boundary terms)

$$\begin{aligned}\partial_t I(t) = \int \partial_t w &= \int b\partial_w \\ &= -\int \underbrace{\partial_h b}_{\geq 0}\, w\end{aligned}$$

This says that $I(t)$ wants to come back to 0. Some problems with this proof:

1. We pretended w has a fixed sign.
2. We integrated over whole space. This is a problem because of boundary terms when integrating by parts.

Now let's be more rigorous. Let $\phi(x) = \frac{x^2}{1+x^2}$, $v = \phi(w) = \phi(f-g)$. Showing $w = 0$ is equivalent to showing $v = 0$. We have

$$\partial_t v - b\partial_h v = 0.$$

Now v has a fixed sign. This solves item 1.

Denote $L = \|\partial_h f\|_{L^\infty} + \|\partial_h g\|_{L^\infty} + 1$. Fix $T < \infty$, and study

$$\begin{aligned}J(t) :&= \int_{-L(T-t)}^{L(T-t)} v(t,h)\, dh \\ &= \int_{-R_t}^{R_t} v\end{aligned}$$

Note J is Lipschitz in t, so it has a derivative a.e.,

$$\begin{aligned}\partial_t J &= \int_{-R_t}^{R_t} \partial_t v - L(v(t,R_t) + v(t,-R_t)) \\ \int_{-R_t}^{R_t} b\partial_h v &= -\int_{-R_t}^{R_t} (\partial_h b)v + [bv]_{-R_t}^{R_t}.\end{aligned}$$

The term $[bv]_{R_t}^{R_t}$ may have a tendency to let J increase. We hope it can be compensated by the $-$ terms: because $|b| \leq L$.

$$\begin{aligned}\partial_t J &\leq -\int_{-R_t}^{R_t} \underbrace{(\partial_h b)}_{\geq 0} \underbrace{v}_{\geq 0} \leq 0 \\ \partial_t J &\leq J.\end{aligned}$$

We know $J(0) = 0$ and $J \geq 0$. We conclude that $J \equiv 0$, and $v = 0$ a.e. Because it is Lipschitz, $v = 0$ and $f = g$. □

To be fully rigorous, we should justify interchanging the derivative and integral, and differentiating b (since f, g are only Lipschitz, and may not be twice differentiable).

Exercise 1.3 (for credit)**:** Make this proof rigorous.
Hint: convolve with a function to make it smooth. See the book [Eva10].

What makes the proof work is that the nonlinearity H is convex:

$$\partial_t f - H(\nabla f) = 0.$$

Difficulty comes in when H is not convex or concave.

Any limit point of a solution should be a solution: You can try to run the proof, show the sequence of functions is a Cauchy sequence, and the limit is a solution.

We can write down a formula for the solution.

Proposition 1.4 (Hopf-Lax formula)**:** Let ψ be convex and Lipschitz.[5] The function

$$f(t,h) = \sup_{h' \in \mathbb{R}} \left(\psi(h - h') - \frac{(h')^2}{4t} \right)$$

is the weak solution of

$$\begin{aligned} \partial_t f - (\partial_h f)^2 &= 0 \\ f(0, \cdot) &= \psi. \end{aligned}$$

[6] Note the last term is related to the convex dual of the nonlinearity: the convex dual of $p \mapsto p^2$ is $q \mapsto \frac{q^2}{4}$. See [Eva10].

Exercise 1.5: Show that for $t > 0$ small that $\partial_h f(t, 0) = 0$. You can use the Hopf-Lax formula.

For $t < \infty$ large, $\partial_h^+ f(t,0) > 0 > \partial_h^- f(t,0)$; there is a phase transition in the derivative.

1.3 Back to Curie-Weiss

It's not necessarily true that convergence of the function tells us something about the derivatives.

Proposition 1.6: If (t,h) is a point of differentiability (in h) of f, and if $F_N \to f$ (pointwise), then

$$\partial_h F_N(t,h) \to \partial_h f(t,h).$$

[5] It suffices to be locally semi-convex: for every $\delta > 0$, there exists $C_\delta < \infty$ such that for all $\geq \delta$, $h \mapsto f(t,h) + C_\delta h^2$. (Note the lower bound is degenerate as $t \to 0$.) More generally, this is true under mild regularity assumptions, though we have to define weak solution differently.

[6] Is there a modification for the pre-limit PDE in F_N? There would be some Brownian motion. f will be the some expecation of some exponential of Brownian motion. For our application, we will not have a completely closed formula for F_N, so we cannot write a formula of this form before passing to the limit.

Proof. F_N is convex in h, so

$$F_N(t,h') \geq F_N(t,h) + \underbrace{\partial_h F_N(t,h)}_{\text{bounded}}(h'-h).$$

Because the derivatives are bounded, we can take a subsequence along which $\partial_h F_N(t,h) \to p$. Then

$$f(t,h') \geq f(t,h) + p(h'-h),$$

and p must be $\partial_h f(t,h)$. □

Convergence of F_N. We have

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

We want F_N to be close to the HJ solution. Let $w = F_N - f$; then

$$\partial_t w - b\partial_h w = \frac{1}{N}\partial_h^2 F_N$$
$$\text{where } b = \partial_h F_N + \partial_h f.$$

Let $v = \phi(w)$. Before, the difference solved the same equation, but now we have to be more careful.

$$\partial_t v - b\partial_h v = \phi'(w)\frac{1}{N}\partial_h^2 F_N.$$

The argument is similar, but the RHS is different here. Define

$$J(t) = \int_{-R_t}^{R_t} v(t,h)\,dh.$$

We have an extra term

$$\partial_t J \leq \ldots + \int_{-R_t}^{R_t} \underbrace{\phi'(w)}_{\leq 1}\frac{1}{N}\underbrace{\partial_h^2 F_N(t,h)}_{\geq 0}\,dh$$
$$\partial_t J \leq \frac{1}{N}\int_{-R_t}^{R_t}\partial_h^2 F_N(t,h)\,dh = \frac{1}{N}[\partial_h F_N]_{-R_t}^{R_t} \leq \frac{2}{N}.$$

because this is the integral of a derivative. Recall $J(0) = 0$. So $J(t) \leq \frac{2t}{N}$.

Exercise 1.7: Clean up this proof! How do you get pointwise convergence?

In the rank-1 case, we find some quantities similar to F_N. We have a similar situation with a function which satisfies a similar equation, and want to show it converges to the true solution. We need some L^1 estimate.

Key point: If there are "error terms on the right hand side," we need to estimate it in L^1 in the h variable uniformly in h (locally). We want a "local L^∞_t, L^1_h estimate."

References

[BM19] Jean Barbier and Nicolas Macris. "The adaptive interpolation method: a simple scheme to prove replica formulas in Bayesian inference". In: *Probability Theory and Related Fields* 174.3-4 (2019), pp. 1133–1185. URL: https://arxiv.org/abs/1705.02780.

[Eva10] Lawrence C Evans. *Partial differential equations*. Vol. 19. American Mathematical Soc., 2010.

[FV17] Sacha Friedli and Yvan Velenik. *Statistical mechanics of lattice systems: a concrete mathematical introduction*. Cambridge University Press, 2017. URL: https://www.unige.ch/math/folks/velenik/smbook/index.html.

[LM19] Marc Lelarge and Léo Miolane. "Fundamental limits of symmetric low-rank matrix estimation". In: *Probability Theory and Related Fields* 173.3-4 (2019), pp. 859–929. URL: https://arxiv.org/abs/1611.03888.

[Mou18] Jean-Christophe Mourrat. "Hamilton-Jacobi equations for mean-field disordered systems". In: *arXiv preprint arXiv:1811.01432* (2018). URL: https://arxiv.org/abs/1811.01432.