DISORDERED SYSTEMS, RANK-ONE MATRIX ESTIMATION, AND HAMILTON-JACOBI EQUATIONS

These notes were taken during the Online Open Probability Summer School (OOPS) in 2020, based on the lectures of Jean-Christophe Mourrat called "Disordered Systems and Hamilton-Jacobi Equations" or "Rank-One Matrix Estimation and Hamilton-Jacobi Equations". The content essentially follows the papers [2] and [3] but also draws on the more general text of [1]. The flow of ideas is mostly chronological as presented by Mourrat, but, for the sake of completeness, I have inserted the proofs of some claims that he left as "exercises to the audience" during the talks.

1. Introduction

We begin with some definitions of very classical ideas in statistical mechanics. Let $N > 0$ be fixed and consider a collection of particles denoted $\{1, \dots N\}$. Now let $\{J_{ij}\}_{i,j=1}^N$ be independent standard Gaussian random variables which denote the repulsion/attraction between particles i and j, and let $\sigma \in \{+1, -1\}^N$ denote any assignment of spins to our particles. A natural question, with motivations in physics about minimizing a certain energy, is to solve the optimization problem

$$\begin{aligned} \text{maximize} \qquad & \sum_{i,j} J_{ij} \mathbf{1}\{\sigma_i = \sigma_j\} \\ \text{over} \qquad & \sigma \in \{+1, -1\}^N. \end{aligned}$$

Note that $\mathbf{1}\{\sigma_i = \sigma_j\} = \frac{1}{2}(\sigma_i \sigma_j + 1)$, so $\sum_{i,j} J_{ij} \sigma_i \sigma_j = 2 \sum_{i,j} J_{ij} \mathbf{1}\{\sigma_i = \sigma_j\} - \sum_{i,j} J_{ij}$. Since the first term just applies a positive scaling and the second term does not depend on σ, a problem with the same solution is the following:

$$\begin{aligned} \text{maximize} \qquad & \sum_{i,j} J_{ij} \mathbf{1} \sigma_i \sigma_j \\ \text{over} \qquad & \sigma \in \{+1, -1\}^N. \end{aligned}$$

Observe that the solution to either optimization problem is random since they both depend on the random weights $\{J_{ij}\}_{i,j=1}^N$.

An easier question is to take J_{ij} to be deterministically equal to 1. Then the optimal configuation is the one which assigns all particles to spin $+1$ (or, all particles to spin -1). The difference between our problem and the simpler problem is that, the optimal configuation for a diverse range of values of the J_{ij}'s will not be one that satisfies the forces of every individual pair. Rather, some pairs will be "frustrated" in order to reduce the overall objective. Such system are often called "disordered". In either case, the addition of some other term in the potential may cause the optimal configuration to be one which differs from this "ground state".

In the terminology of statistical mechanics, the model with $J_{ij} \equiv 1$ is called the *Ising model* and the model with $J_{ij} \sim_{\text{iid}} N(0, 1)$ is called a *spin glass model*. In

either case we can view these as models of particles systems on a complete graph, so obvious generalizations can be made to general graphs, a common example being the lattice $\mathbb{Z}^d \subseteq \mathbb{R}^d$ with its nearest-neighbor structure.

(Mourrat included another motivating example in his talk: Suppose that, at Hogwarts, there two houses called $+1$ and -1 and there are N students named $\{1, \dots N\}$. Let the "quality of interaction" between student i and j be denoted by J_{ij}, which we assume are iid $N(0,1)$. The job of the sorting hat is the find the optimal configuation of this spin glass model, i.e. to optimize the total quality of interaction among all students. However, I find this analogy quite strange, because, if I recall correctly, the sorting hat assigns students to houses in an online way for which this kind of global optimization seems impossible.)

If the maximum of the objective $\sigma \mapsto \sum_{i,j} J_{ij}\sigma_i\sigma_j$ is much bigger than the objective of the suboptimal configurations, then we expect a soft-max to be approximately equal to the true maximum. In othern words, for any $\beta > 0$, we expect

$$\max_{\sigma \in \{\pm 1\}^N} \sum_{i,j} J_{ij}\sigma_i\sigma_j \approx \mathbb{E}\left[\frac{1}{N}\log\left(\sum_{\sigma\in\{\pm1\}^N} \exp\left(\frac{\beta}{\sqrt{N}}\sum_{i,j} J_{ij}\sigma_i\sigma_j\right)\right)\right]. \tag{1}$$

Here the normalizing constants have been chosen to reduce each term to approximately constant order. The measure on $\{\pm 1\}^N$ whose density is

$$\exp\left(\frac{\beta}{\sqrt{N}}\sum_{i,j} J_{ij}\sigma_i\sigma_j\right) \tag{2}$$

with respect to the uniform measure on $\{\pm 1\}^N$ is called the *Gibbs measure* of this particle system, and the parameter β is called the *inverse temperature*. This family of measures (which are random since they depend on the random weights) plays a fundamental role in understanding the given particle system.

It turns out that this type of model from statistical mechanics can help in analyzing a certain problem in statistical inference, which we now outline. Suppose that $\overline{x}$ is a random vector in $\mathbb{R}^N$ with iid entries sampled from some distribution $\mathbb{P}$ of compact support. Now suppose that $W = \{W_{ij}\}_{i,j=1}^N$ is a matrix of iid $N(0,1)$ random variables. An important problem in (Bayesian) statistical inference is that of estimating the rank-one matrix $\overline{x}\overline{x}^{\mathrm{T}}$ after observing the noisy matrix

$$Y = \sqrt{\frac{2t}{N}}\overline{x}\overline{x}^{\mathrm{T}} + W. \tag{3}$$

(Here, we have the normalizing constant of $\sqrt{2/N}$ mostly for convenience of some calculations.) The paramter $t \geq 0$ represents the signal-to-noise ratio of the inference problem, whereby for large t we expect prediction to be "easy" and for small t we expect prediction to be "hard".

The fact that the spin-glass model and the rank-one matrix factorization problem have anything to do with each other is a bit surprising at first glance. As one piece of evidence, consider the relationship between the parameters β and t: It is known (from previous literature) that both problems exhibit a *phase transition* with respect to their respective parameter: In the spin-glass model, there exists

a critical inverse temperature β_c such that the optimal configuation is relatively random when $\beta < \beta_c$ and relatively ordered when $\beta > \beta_c$; Likewise, there exists a critical signal-to-noise ratio t_c such that estimation of the true matrix is impossible when $t < t_c$ and possible when $t > t_c$. As we later see, this similarity is no coincidence, and it has largely to do with the form of a certain Gibbs measure associated to both problems.

2. The Curie-Weiss Model

For this section, we're going to study a specific interacting particle system where we can develop some tools. This setting is not exactly the same as that of our spin glass model for rank-one matrix estimation, but it is relatively easy and we can get the general idea of the analysis. Our focus is the *Curie-Weiss model*, which was originally developed as a mathematical model for the physical phenomenon of ferromagnetism.

Let $N > 0$ be fixed and also consider two reals $t > 0$ and $h \in \mathbb{R}$, which we will later vary. Now consider the measure μ on $\{\pm 1\}^N$ which, to each configuration $\sigma \in \{\pm 1\}^N$ assigns the probability

$$\mu(\{\sigma\}) = \frac{\exp\left(\frac{t}{N}\sum_{i,j}\sigma_i\sigma_j + h\sum_i \sigma_i\right)}{\sum_{\sigma' \in \{\pm 1\}^N}\exp\left(\frac{t}{N}\sum_{i,j}\sigma'_i\sigma'_j + h\sum_i \sigma'_i\right)}. \tag{4}$$

Viewing the bottom term as just a normalizing constant, we can regard the probability of each configuration as consisting of two pieces: an internal interaction term akin to that of the Ising model, and an external potential term corresponding to the effect of a magnetic field on the system. That is,

$$\mu(\{\sigma\}) \propto \exp\left(\overbrace{\frac{t}{N}\sum_{i,j}\sigma_i\sigma_j}^{\text{Ising model}} + \overbrace{h\sum_{i,j}\sigma_i}^{\text{magnetic field}}\right) \tag{5}$$

Hence, the parameters t and h represent, respectively, as the inverse temperature of the particle system, and as h is the intensity of the magnetic field. Our main object of interest is actually the case of $h = 0$, but it turns out that this "enriched potential" will be a more convenient object of study.

As we will later see, it is most helpful to try to understand the normalizing constant of the probability density function of μ, so we define the following:

$$F_N(t,h) = \frac{1}{N}\log\left(\sum_{\sigma\in\{\pm 1\}^N}\exp\left(\frac{t}{N}\sum_{i,j}\sigma_i\sigma_j + h\sum_i\sigma_i\right)\right). \tag{6}$$

Observe that F_N is a just the logarithm of a finite sum of exponentials, hence we have $F_N \in C^\infty(\mathbb{R}_+ \times \mathbb{R}; \mathbb{R})$.

To see where the Hamilton-Jacobi PDE enters the picture, let us take some derivatives of $F_N(t,h)$ in order to derive some relations among them. As some useful notation, for any measurable $f : \{\pm 1\}^N \to \mathbb{R}$, let $\langle f(\sigma)\rangle_{t,h}$ denote expectation with

respect to the measure μ where t and h are the given parameters. Taking one derivative in both the t and h directions, we get:

$$\partial_t F_N = \frac{1}{N}\left\langle \frac{1}{N}\sum_{i,j}\sigma_i\sigma_j \right\rangle_{t,h} = \left\langle \left(\frac{1}{N}\sum_i \sigma_i\right)^2 \right\rangle_{t,h},$$
$$\partial_h F_N = \left\langle \frac{1}{N}\sum_i \sigma_i \right\rangle_{t,h}.$$

Note that these equations together yield

$$\partial_t F_N - (\partial_h F_N)^2 = \left\langle \left(\frac{1}{N}\sum_i \sigma_i\right)^2 \right\rangle_{t,h} - \left(\left\langle \frac{1}{N}\sum_i \sigma_i \right\rangle_{t,h}\right)^2 \tag{7}$$

where the right side is equal to the variance of the random variable $\frac{1}{N}\sum_i \sigma_i$ and is hence nonnegative. Physically, $\frac{1}{N}\sum_i \sigma_i$ represents the (empirical) mean magnetization of the system, so $\partial_h F_N = \langle \frac{1}{N}\sum_i \sigma_i \rangle_{t,h}$ represents the (population) mean magnetization. Let's take this calculation one step further: Taking another derivative in the h direction gives

$$\partial_h^2 F_N = \frac{1}{N}\left\langle \left(\sum_i \sigma_i\right)^2 \right\rangle_{t,h} - \frac{1}{N}\left(\left\langle \sum_i \sigma_i \right\rangle_{t,h}\right)^2, \tag{8}$$

and we recognize this as just N times the variance in the equation above! On the one hand, this is nonnegative, so, for each fixed $t > 0$, the function F_N is convex in h. On the other hand, this proves that the function F_N solves the PDE

$$\partial_t F_N - (\partial_h F_N)^2 = \frac{1}{N}\partial_h^2 F_N. \tag{9}$$

This is not exactly the Hamilton-Jacobi PDE, but it is not so different. Since the empirical mean magnetization $\frac{1}{N}\sum_i \sigma_i$ is bounded beteen -1 and $+1$, its variance is bounded between 0 and 1. This implies that, as $N \to \infty$, the right side converges to 0. So,we expect that the limiting behavior of the normalizing constant of the particle system F_∞ must be, in some sense, a solution to the true Hamilton-Jacobi PDE $\partial_t F_\infty - (\partial_h F_\infty)^2 = 0$. The exact sense in which F_∞ is a solution, however, is a bit subtle and will deserve more attention later.

Let's also spend a moment thinking about the initial condition of this PDE (or rather, this family of PDEs). For any $N > 0$ and any $h \in \mathbb{R}$, note that we can compute:

$$\begin{aligned} F_N(0,h) &= \frac{1}{N}\log\left(\sum_{\sigma\in\{\pm 1\}^N}\exp\left(h\sum_i \sigma_i\right)\right) \\ &= \frac{1}{N}\log\left(\sum_{\sigma\in\{\pm 1\}^N}\prod_i e^{h\sigma_i}\right) \\ &= \frac{1}{N}\log\left(\left(e^h+e^{-h}\right)^N\right) = \log\left(e^h+e^{-h}\right) = \psi(h) \end{aligned}$$

where we have defined ψ by the last equality. Interestingly, we see that the initial condition does not depend on N, so it is the same across the entire family of PDEs.

Finally, recall that our main object of interest is the function $F_N(t,0)$. Since this is just the logarithm of the moment generating function of the random variable $\frac{1}{N}\sum_{i,j}\sigma_i\sigma_j$ with respect to the uniform measure on $\{\pm 1\}^N$, we can recover everything about the distribution of this random variable from $F_N(t,0)$ alone. In the case of finite N, there are, of course, many ways to understand the distribution of $\frac{1}{N}\sum_{i,j}\sigma_i\sigma_j$; however, in the limit as $N\to\infty$, the approach is not so clear. This idea of introducting a PDE that F_N must solve will lead us to a PDE that F_∞ must solve, and hence we can understand the distribution of the (weak) limit of the random variables $\frac{1}{N}\sum_{i,j}\sigma_i\sigma_j$ as $N\to\infty$. To do this analysis, the next step is to understand the PDE that we have derived.

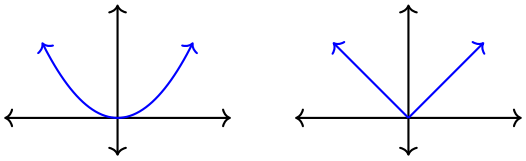
3. Hamilton-Jacobi PDE

Consider the following PDE, where f and ψ are general functions not necessarily related to the specific forms described in the previous section:

$$(*)\qquad \begin{cases} f:(\mathbb{R}_+)_t\times\mathbb{R}_h\to\mathbb{R}, \\ \partial_t f-(\partial_h f)^2=0, \\ f(0,h)=\psi(h). \end{cases}$$

This is the *Hamilton-Jacobi PDE* with initial condition ψ and it provides an important tool for analyzing the particle system(s) of the previous sections.

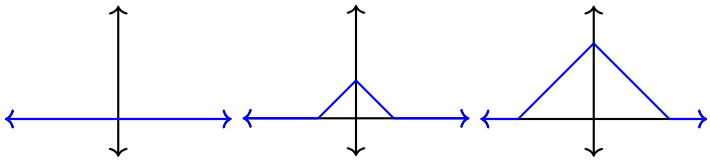
What is the sense in which a function can be a solution to $(*)$? A good first attempt is to require that f be a C^1 function which satisfies the PDE pointwise. However, we now make a heuristic argument that this notion of solution is too strong for our primary example—the large-N limit of the log moment-generating function of the Curie-Weiss model $F_\infty=\lim_{N\to\infty}F_N$—to be a solution; the argument is based on the existence of a critical threshhold t_c and the interpretation of $\partial_h F_N(t,h)$ as the mean magnetization: If $t<t_c$, then the spins $\{\sigma_i\}_{i=1}^N$ are assigned more-or-less randomly according to h, so we expect $\partial_h F_\infty(t,h)$ to vary quite smoothly in h and everything is nice. On the other hand, if $t>t_c$, then the spin configuration has more structure and is more robust to changes in h. Specifically, we expect that the mean magnetization $\partial_h F_\infty(t,h)$ remains bounded away from 0 from below as $h\downarrow 0$, and simultaneously that $\partial_h F_\infty(t,h)$ remains bounded away from 0 from above as $h\uparrow 0$. In other words, we expect that the cross-sections of the function $F_\infty(t,h)$ can be roughly visualized as follows:



$F_\infty(t,h)$ $t \ll t_c$ $\qquad$ $F_\infty(t,h)$ $t \gg t_c$

Therefore, it appears to be too restrictive to assume that our solutions are continuously differentiable.

A second guess is to require that f be Lipschitz in both t and h and that it satisfy the PDE almost everywhere. (Recall that a Lipschitz function is differentiable almost everywhere with respect to the Lebesgue measure.) However, we now argue that this notion of solution is too weak to determine a unique solution based on given initial data: If we start with the initial condition $\psi(h) = 0$ for all $h \in \mathbb{R}$, then we can construct two distinct solutions. One solution is $f(t,h) = 0$ for all (t,h), which is Lipschitz in both variables satsifies the PDE everywhere. Another solution consists of a "growing triangular pulse", which can be described best through the following figure in which we plot, for purposes of illustration, the functions $f(0,h)$, $f(\frac{1}{2},h)$, and $f(1,h)$:



$f(0,h)$ $\qquad$ $f(\frac{1}{2},h)$ $\qquad$ $f(1,h)$

This function is also Lipschitz in both t and h, satisfies the PDE almost everywhere, and has the right initial condition.

So, our notion of solution must be somewhere in between these extremes, and it turns out that the following definition strikes the right balance:

Definition 3.1. We say that a function $f : (\mathbb{R}_+)_t \times \mathbb{R}_h \to \mathbb{R}$ is a *weak solution to* $(*)$ if the following conditions are satisfied: the map $h \mapsto f(t,h)$ is convex and uniformly Lipschitz over all $t > 0$ (that is, there exists a constant $C > 0$ such that $|f(t,h) - f(t,h')| \leq C|h - h'|$ holds for all $t > 0$ and $h, h' \in \mathbb{R}$, and we let $||f||_{\mathrm{Lip},h}$ denote the smallest such constant), the map $t \mapsto f(t,h)$ is Lipschitz for all $h \in \mathbb{R}$ (but the Lipschitz constant may depend on h), the equality $f(0,h) = \psi(h)$ holds for all $h \in \mathbb{R}$, and f satisfies $(*)$ almost everywhere.

Note that this notion of weak solution is different from the notion of *viscosity solution*, the latter being slightly more standard in analysis of Hamilton-Jacobi PDE. For our purposes, this concept of weak solution is more appropriate, and the "correctness" of the definition is established by the following result:

Proposition 3.2. For any given Lipschitz, convex function $\psi : \mathbb{R} \to \mathbb{R}$, there exists a unique weak solution of the Hamilton-Jacobi PDE $(*)$, given by

$$f(t,h) = \sup_{h' \in \mathbb{R}} \left(\psi(h - h') - \frac{(h')^2}{4t} \right). \tag{10}$$

This explicit form is known as the *Hopf-Lax formula*.

Proof. For existence, we only need to verify that the Hopf-Lax formula yields a weak solution of $(*)$. This is straightforward and we omit these details for now.

For uniqueness, we will prove the result as a consequnce of a "local L^∞_t-L^1_h energy estimate" method. Suppose that f and g are both weak solutions of $(*)$ with the same initial condition. Define $L = ||f||_{\text{Lip},h} + ||g||_{\text{Lip},h} + 1$ and define the function $\phi : \mathbb{R} \to \mathbb{R}$ by $\phi(x) = x^2/(1+x^2)$, which we see has values bounded in $[0, 1)$ is equal to zero if and only if its argument is zero, and has derivatives bounded in $[0, 1/\sqrt{2})$. Now set $w(t,h) = f(t,h) - g(t,h)$ and $v(t,h) = \phi(w(t,h))$, and note that there is an almost-everywhere-defined function $b(t,h) = \partial_h f(t,h) + \partial_h g(t,h)$ such that v satisfies the PDE

$$\partial_t v(t,h) - b(t,h)\partial_h v(t,h) = 0. \tag{11}$$

In the remaining, we may omit the arguments (t,h) from the notation whenever it is clear from context. Now fix some end-time $T \in (0,\infty)$, and define the functional:

$$J(t) = \int_{-L(T-t)}^{L(T-t)} v(t,h)dh. \tag{12}$$

Note that $v(t,h)$ is almost-everywhere differeniable in t and uniformly bounded. So, we can differentiate under the integral (note that the limit of integration also depend on t), yielding:

$$\frac{dJ}{dt} = \int_{-L(T-t)}^{L(T-t)} \partial_t v(t,h)dh - Lv(t,L(T-t)) - Lv(t,-L(T-t)). \tag{13}$$

The next step would be to apply the PDE for v and use integration by parts. However, the fact that b is not differentiable in h necessitates an extra step of "smoothing" our functions. Let $\varphi \in C^\infty(\mathbb{R};\mathbb{R})$ be a *mollifier*, in the sense that $\text{supp}(\varphi) \subseteq [-1,1]$ and $\int_{\mathbb{R}} \varphi(h)dh = 1$. Now for each $\varepsilon > 0$, define $\varphi_\varepsilon(x) = \varepsilon^{-1}\varphi(x/\varepsilon)$, and then

$$f_\varepsilon(t,h) = \int_{\mathbb{R}} f(t,h-h')\varphi_\varepsilon(h')dh'. \tag{14}$$

Observe that, for any fixed $t > 0$, the function $h \mapsto f_\varepsilon(t,h)$ is both C^∞ and convex, and satisfies $\partial_h f_\varepsilon(t,h) \to \partial_h f(t,h)$ for almost all values of $h \in \mathbb{R}$; we can define the function g_ε analogously, and we get the same properties. Now we can define $b_\varepsilon = \partial_h f_\varepsilon + \partial_h g_\varepsilon$ and, for any fixed $t > 0$, this function C^∞ and non-decreasing in h, and satisfies $b_\varepsilon(t,h) \to b(t,h)$ for almost all $h \in \mathbb{R}$. The PDE solved by v can now equivalently be written as

$$\partial_t v = \partial_h(b_\varepsilon v) + (b - b_\varepsilon)\partial_h v - v\partial_h b_\varepsilon, \tag{15}$$

which holds almost everywhere. Substituting this into the integral for dJ/dt, we get:

$$\frac{dJ}{dt} = \Big(b_\varepsilon(t, L(T-t)) - L\Big)v(t, L(T-t)) + \Big(b_\varepsilon(t, -L(T-t)) - L\Big)v(t, -L(T-t)) + \int_{-L(T-t)}^{L(T-t)} ((b - b_\varepsilon)\partial_h v - v\partial_h b_\varepsilon)dh.$$

Note that, by construction, we have $b \leq L$ wherever b is defined, and hence it follows that $b_\varepsilon \leq L$. Since we have $v \geq 0$ everywhere, this implies that the first two terms above are both non-positive. So, we have

$$\frac{dJ}{dt} \leq \int_{-L(T-t)}^{L(T-t)} ((b - b_\varepsilon)\partial_h v - v\partial_h b_\varepsilon)dh, \tag{16}$$

and it only remains to analyze these remaining integral terms. For the first term, note that we have $b - b_\varepsilon \to 0$ for almost all $h \in [-L(T-t), L(T-t)]$ and that $(b - b_\varepsilon)\partial_h v$ is bounded above, so dominated convergence gives $\int_{-L(T-t)}^{L(T-t)}((b-b_\varepsilon)\partial_h v \to 0$ as $\varepsilon \to 0$. Moreover, note that we have $\partial_h b_\varepsilon \geq 0$ and $v \geq 0$, so the second term (inculding the sign) is non-positive for all $\varepsilon > 0$. Therefore, taking the limit as $\varepsilon \to 0$ gives

$$\frac{dJ}{dt} \leq 0, \tag{17}$$

as we hoped for.

Now we use the standard method to get uniqueness from the energy-estimate: Since f and g have the same initial condition, we see that $J(0) = 0$ holds. Now J, being the integral of the nonnegative function v, has $J(t) \geq 0$, so $dJ/dt \leq 0$ implies $J(t) = 0$ for all $t \in [0, T]$. Fixing $t > 0$ and taking the limit as $T \to \infty$, monotone convergence gives

$$J(t) = \int_{\mathbb{R}} v(t, h)dh = 0. \tag{18}$$

This forces $v(t, h) = 0$ for almost all $h \in \mathbb{R}$, hence $w(t, h) = 0$ for almost all $h \in \mathbb{R}$. Since $w(t, h)$ is (uniformly) continuous in h and since we have $w(t, h) = 0$ for a full-measure hence dense set of $h \in \mathbb{R}$, it follows that we have $w(t, h) = 0$ for all $h \in \mathbb{R}$. Since $t > 0$ was arbitrary, this shows that we have $f(t, h) = g(t, h)$ for all $t > 0$ and all $h \in \mathbb{R}$. □

4. Main Results

In this last section we combine these ideas from statistical mechanics, analysis of PDE, and statistical inference into the precise statement and proof sketch of the main results.

As a reminder of the setting, we assume that $\overline{x}_1, \ldots \overline{x}_N$ are iid samples from some distribution $\mathbb{P}$ on $\mathbb{R}$ of bounded support and that W is a random $N \times N$ matrix with iid $N(0, 1)$ entries. Suppose we observe

$$Y = \sqrt{\frac{2t}{N}}\overline{x}\overline{x}^{\mathrm{T}} + W. \tag{19}$$

and that our goal is to predict $\overline{x}$ from just the data Y. To do this, we consider the following error:

$$\mathrm{MSE}_N(t) = \frac{1}{N^2} \inf_{\hat{\theta}} \mathbb{E}\left[||\overline{x}\overline{x}^{\mathrm{T}} - \hat{\theta}(Y)||_F^2\right] \tag{20}$$

where the infimum is taken over all possible estimators $\hat{\theta}(Y)$ of $\overline{x}\overline{x}^{\mathrm{T}}$, and the expectation is taken over two sources of randomness: the souce of the sampling of $\overline{x}$ and the source of the noise in the data Y.

It is well-known in Bayesian statistics that the optimal estimator of $\overline{x}$ given Y is just the conditional expectation $\hat{\theta}(Y) = \mathbb{E}[\overline{x}\overline{x}^{\mathrm{T}}|Y]$, so a natural first step is to understand the conditional distribution of $\overline{x}$ given Y. To do this, define the potential

$$\begin{aligned} H_N^\circ(t,x) &= \sqrt{\frac{2t}{N}}\langle Y, xx^{\mathrm{T}}\rangle_F - \frac{t}{N}||x||_2^4 \\ &= \sqrt{\frac{2t}{N}}\langle x, Wx\rangle + \frac{2t}{N}|\langle x, \overline{x}\rangle|^2 - \frac{t}{N}||x||_2^4 \end{aligned}$$

where $\langle\cdot,\cdot\rangle_F$ denotes the Frobenius inner product. Now define the functional $\langle f(x)\rangle^\circ$ as

$$\langle f(x)\rangle^\circ = \frac{\int_{\mathbb{R}} f(x)\exp(H_N^\circ(t,x))d\mathbb{P}^{\otimes N}(x)}{\int_{\mathbb{R}}\exp(H_N^\circ(t,x))d\mathbb{P}^{\otimes N}(x)} \tag{21}$$

for any bounded measurable $f : \mathbb{R}^N \to \mathbb{R}$. By slight abuse of notation, we use the same symbol $\langle\cdot\rangle^\circ$ to denote the product measure over multiple copies of this distribution. It turns out that this definition is useful because, it can be shown, that we have $\langle f(x)\rangle^\circ = \mathbb{E}[f(\overline{x})|Y]$. Here, note that t can be seen as interpolation parameter for the conditional distribution of $\overline{x}$ given Y: If $t = 0$, then we simply draw a new copy of an independent random variable from the distribution $\mathbb{P}^{\otimes N}$; as $t \to \infty$, we simply output the exact value of $\overline{x}$. The random variable x, the statistical mechanics literature, is called a *replica*.

Now let's use this observation to make some simple manipulations leading to a highly non-trivial result. The following result can be seen as the machine that makes this entire theory work:

Lemma 4.1 (Nishimori's Property)**.** Let $x_1, \dots x_m$ be iid copies of the conditional distribution of $\overline{x}$ given Y. Then we have $\mathbb{E}\left[\langle f(x_1, \dots x_m)\rangle^\circ\right] = \mathbb{E}\left[\langle f(x_1, \dots x_{m-1}, \overline{x})\rangle^\circ\right]$ for any measurable $f : \mathbb{R}^{N\times m} \to \mathbb{R}$ for which the expectations are well-defined.

Proof. By conditional independence, the fact that $\langle f\rangle^\circ$ is $\sigma(Y)$-measurable, and the tower property, we have the following for any bounded measurable functions $f_1, \dots f_m : \mathbb{R}^N \to \mathbb{R}$:

$$\begin{aligned}\mathbb{E}\left[\langle f_1(x_1)\cdots f_m(x_m)\rangle^\circ\right] &= \mathbb{E}\left[\langle f_1(x_1)\rangle^\circ\cdots\langle f_{m-1}(x_{m-1})\langle f_m(x_m)\rangle^\circ\right]\\ &= \mathbb{E}\left[\langle f_1(x_1)\rangle^\circ\cdots\langle f_{m-1}(x_{m-1})\rangle^\circ\mathbb{E}[f_m(\overline{x})|Y]\right]\\ &= \mathbb{E}\left[\mathbb{E}[\langle f_1(x_1)\rangle^\circ\cdots\langle f_{m-1}(x_{m-1})\rangle^\circ f_m(\overline{x})|Y]\right]\\ &= \mathbb{E}\left[\mathbb{E}[\langle f_1(x_1)\cdots f_{m-1}(x_{m-1})f_m(\overline{x})\rangle^\circ|Y]\right]\\ &= \mathbb{E}\left[\langle f_1(x_1)\cdots f_{m-1}(x_{m-1})f_m(\overline{x})\rangle^\circ\right].\end{aligned}$$

This proves the result for functions that are products of functions depending on one coordinate alone, so the general result follows by applying linearity of the integral and the monotone class theorem. □

Remark 4.2. Nishimori's property can also be generalized to include functions that depend on Y, i.e. for any measurable function $f : \mathbb{R}^{N\times m+N\times N} \to \mathbb{R}$, we have $\mathbb{E}\left[\langle f(x_1,\dots x_m,Y)\rangle^\circ\right] = \mathbb{E}\left[\langle f(x_1,\dots x_{m-1},\overline{x},Y)\rangle^\circ\right]$. We will need this more general result, but we do not prove it here.

The reason Nishimori's property is so important is as follows: Taking partial derivatives of the log-normalizing constant for the Gibbs measure of a particle system will introduce new replicas into the analysis, and these can often become very hard to keep track of. However, Nishimori's property tell us that we can always "relate things back" to the original variable $\overline{x}$ and this helps keep the calculations simple.

We can already start to see where the analysis of spin-glass systems will connect to the analysis of rank-one matrix estimation: The conditional distribution of $\overline{x}$ given Y is just a (random) Gibbs measure with an exponentially-tilted probability density with respect to the measure $\mathbb{P}^{\otimes N}$ on $\mathbb{R}^N$. Moreover, the potential contains as its main term, $\langle x, Wx\rangle = \sum_{i,j} W_{ij}x_ix_j$ which is exactly the interaction potential of the spin-glass system. The remaining terms, while not negligible in order, are mathematically convenient and don't have a great physical interpretation at the moment.

Taking a lesson from the analysis of the Curie-Weiss model, we can understand the rank-one matrix estimation problem if we understand the log moment-generating function of our potential. However, we should make sure "enrich" the potential before doing this, so that there are two variables among which some differential identities can be derived. The difficulty in this seting is that we must do this in a way which preserves Nishimori's property.

To do this, let $h \geq 0$ be fixed. Suppose that, in addition to observing the matrix Y, we also observe the vector

$$y = \sqrt{2h}\overline{x} + w, \tag{22}$$

where w is a vector in $\mathbb{R}^N$ of iid $N(0,1)$ random variables, independent of $\overline{x}$ and Y. Now to discuss the conditional distribution of $\overline{x}$ given Y and y, we define

$$H_N(t,h,x) = H_N^\circ(t,x) + \sqrt{2h}\langle x, y\rangle + 2h\langle x,\overline{x}\rangle - h||x||^2. \tag{23}$$

Also set

$$\langle f(x)\rangle = \frac{\int_{\mathbb{R}} f(x)\exp(H_N(t,h,x))d\mathbb{P}^{\otimes N}(x)}{\int_{\mathbb{R}}\exp(H_N(t,h,x))d\mathbb{P}^{\otimes N}(x)}. \tag{24}$$

It can be proven, in the same manner as before, that for any bounded measurable $f:\mathbb{R}^N\to\mathbb{R}$, we have $\langle f(x)\rangle = \mathbb{E}[f(\overline{x})|Y,y]$.

Now we make some definitions about the moment-generating function of a certain random variable. Define

$$\begin{aligned} F_N(t,h) &= \frac{1}{N}\log\left(\int_{\mathbb{R}}\exp\left(H_N(t,h,x)\right)d\mathbb{P}^{\otimes N}(x)\right),\\ \overline{F}_N(t,h) &= \mathbb{E}[F_N(t,h)]. \end{aligned}$$

Since the Gibbs measure is now random, we have to define a random moment generating function and an average one. (This is not particular to matrix estimation problems. In the analysis of the spin-glass analog of the Curie-Weiss model, we also would have had a random Gibbs measure.)

Now we can finally state the main result of the relevant papers.

Theorem 4.3. There exists a weak solution f to the PDE

$$\begin{cases} f:(\mathbb{R}_+)_t\times(\mathbb{R}_{\geq 0})_h\to\mathbb{R},\\ \partial_t f-(\partial_h f)^2=0,\\ f(0,\cdot)=\overline{F}_1(0,\cdot), \end{cases} \tag{25}$$

such that the functions $\{\overline{F}_N\}_{N=1}^{\infty}$ satisfy $\overline{F}_n\to f$ locally uniformly.

Sketch of Proof. The proof consists of several steps, which we outline but do not prove here. First, use Stein's lemma (Gaussian integration by parts) and Nishimori's property to derive the relations

$$\begin{aligned} \partial_h\overline{F}_N(t,h) &= \mathbb{E}\left[\left\langle\frac{x\cdot\overline{x}}{N}\right\rangle\right]\\ \partial_t\overline{F}_N(t,h) &= \mathbb{E}\left[\left\langle\left(\frac{x\cdot\overline{x}}{N}\right)^2\right\rangle\right]. \end{aligned}$$

From these, we can show that $\overline{F}_N$ satisfies the PDE

$$\partial_t\overline{F}_N-(\partial_h\overline{F}_N)^2=\frac{1}{N^2}\mathbb{E}\left[\left\langle(\langle x,\overline{x}\rangle-\mathbb{E}[\langle x,\overline{x}\rangle])^2\right\rangle\right]. \tag{26}$$

Next we show that the right side can be bounded above by

$$\frac{1}{N^2}\mathbb{E}\left[\left\langle(\langle x,\overline{x}\rangle-\mathbb{E}[\langle x,\overline{x}\rangle])^2\right\rangle\right]\leq\frac{1}{N}\partial_h^2\overline{F}_N+\mathbb{E}\left[(\partial_h F_N-\partial_h\overline{F}_N)^2\right]. \tag{27}$$

The first is exactly the term that we had in the case of the Curie-Weiss model, and the second term arises from the noise in the internel potential arising from the random weights in W. Using some concentration inequalities (the Gaussian Poincaré inequality and the Efron-Stein inequality), we can show that the right side goes to 0 and this can be used to finish the proof. □

Given this result, we can derive properties of the large-N limiting particle system by studying the function $f(t,0)$ and its derivatives; equivalently, we can understand the statistical problem of rank one matrix inference through this same function $f(t,0)$. Using the Hopf-Lax formula, we can get a relatively explicit solution for each N, and by taking $N \to \infty$ we can use this to determine properties of the limiting system. In particular, this result can be used to show the existence of the phase transition, that there exists a parameter t_c such that for $t < t_c$ the MSE does not decay and that for $t > t_c$ the MSE decreases with t.

Finally, we remark that the paper [3] details this argument much more carefully, but for the more general setting of low-rank matrix estimation. This affects the analysis of the Hamilton-Jacobi PDE in that weak solutions only require that the marginal function $h \mapsto f(t,h)$ be "locally semiconvex" as opposed to the case of rank-one matrix estimation in which it was required to be convex. So, the argument and notation appears to be different from those appearing in this note, but the fundamental ideas are mostly the same.

References

[1] Sacha Friedli and Yvan Velenik. *Statistical Mechanics of Lattice Systems: A Concrete Mathematical Introduction*. Cambridge University Press, 2017.

[2] Jean-Christophe Mourrat. Hamilton-Jacobi Equations for Mean-Field Disordered Systems. `https://arxiv.org/pdf/1811.01432.pdf`, 2018.

[3] Jean-Christophe Mourrat. Hamilton-Jacobi Equations for Finite-Rank Matrix Inference. `https://arxiv.org/pdf/1904.05294.pdf`, 2019.