# Four lectures on Statistical Physics of Learning

Bruno Loureiro
Département d'Informatique, École Normale Supérieure - PSL & CNRS, France

Lecture notes written for the "*Statistical physics for machine learning*" course taught at the Princeton Machine Learning Theory Summer School, held in August 6 - 15, 2024.

Get in touch at: brloureiro@gmail.com

## Contents

# Part I
# Lecture 1 — Statistical physics 101

## 1 Introduction and historical context

Slides.

## 2 The Curie-Weiss model

In the introduction to these lectures we have discussed the core ideas of statistical physics, and motivated why it provides a natural framework to think about high-dimensional problems arising in different fields, and in particular in computer science. Before moving to the core of these lectures, which will be the application of these ideas to machine learning questions, we would like to illustrate some of these concepts in perhaps the simplest statistical physics problem, the *Curie-Weiss model*, defined by the following Hamiltonian:

$$\mathcal{H}(\boldsymbol{s}) = -\frac{1}{2d}\sum_{i,j=1}^{d} s_i s_j - h\sum_{i=1}^{d} s_i, \qquad \boldsymbol{s} \in \{-1,+1\}^d \tag{2.1}$$

where $h \in \mathbb{R}$ is a constant known as the *external field*.[1] Note that despite being quadratic in $\boldsymbol{s}$, the energy function eq. (2.1) is not convex due to the binary constraint in $\boldsymbol{s} \in \{-1,+1\}^d$.

Every binary variable $s_i \in \{-1,+1\}$, known as a "spin" in the physics jargon, interacts with every other spin. This means the model lacks of geometry, as there is no notion of distance between spin $s_i$ and $s_j$.[2] We say the model is *fully-connected*, *mean-field* or defined in a *complete graph*.

The probability of finding the spins at a given configuration is given by the *Gibbs-Boltzmann distribution*:

$$\mathbb{P}(\boldsymbol{S} = \boldsymbol{s}) = \frac{e^{-\beta\mathcal{H}(\boldsymbol{s})}}{Z_d(\beta, h)} \tag{2.2}$$

where $\beta \geq 0$ is a constant known as the *inverse temperature* and the normalisation constant $Z_d$ is known as the *partition function*:

$$Z_d(\beta, h) = \sum_{\boldsymbol{s}\in\{-1,+1\}^d} e^{-\beta\mathcal{H}(\boldsymbol{s})} \tag{2.3}$$

The measure in eq. (2.2) weights each configuration $\boldsymbol{s} \in \{-1,+1\}^d$ according to their energy $H(\boldsymbol{s})$. When $\beta = 0$, eq. (2.2) reduces to the uniform measure over the hypercube ($S_i = \pm 1$ with equal probability $1/2$), while in the limit $\beta \to \infty$ the measure peaks in the configurations that minimise the energy function eq. (2.1), also known as the *ground state*. Therefore, the interaction term in eq. (2.1) favours the alignment between spins, while the external field term favours configurations which are alined with $\text{sign}(h)$. This suggests also an interpretation as a model for conformist behaviour, where every citizen $i \in [d]$ pick their choice $s_i \in \{-1,+1\}$ according to the majority.

### 2.1 High-dimensional asymptotics

The central idea in the statistical physics approach to distributions such as eq. (2.2) is to find low-dimensional functions of the configurations which summarise the behaviour of the measure in the limit

[1] Note it is sometimes common to introduce a second constant $J$ for the quadratic term. Without loss of generality, we work in units of $J = 1$.

[2] In contrast to the Ising model, where the interaction is nearest-neighbours.

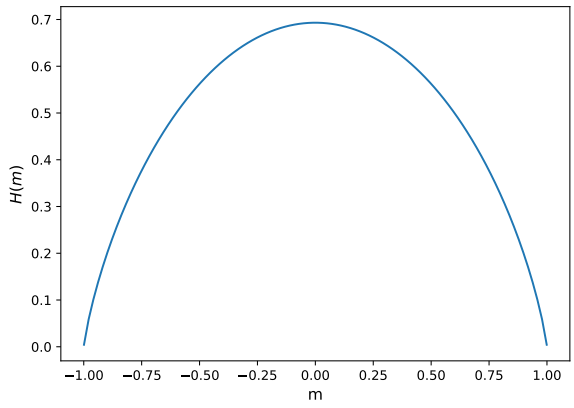


Figure 1: Binary entropy $H(m)$ defined in eq. (2.11) as a function of the magnetisation $m$.

of large $d$. In statistical physics, these are known as *macroscopic variables* or *order parameters*, and in high-dimensional statistics are known as *summary statistics*.

Often, it might not be obvious what are good order parameters for a given Hamiltonian, and solving the model might require some intuition on the system. However, for the Curie-Weiss model eq. (2.1) it is clear that the Hamiltonian is only a function of the average magnetisation:

$$\bar{s} = \frac{1}{d}\sum_{i=1}^{d} s_i \in [-1, 1] \tag{2.4}$$

Indeed, we can re-write:

$$\mathcal{H}(\boldsymbol{s}) = -\frac{1}{2}\left(\frac{1}{d}\sum_{i=1}^{d} s_i\right)^2 - dh\left(\frac{1}{d}\sum_{i=1}^{d} s_i\right) = -d\left(\bar{s}^2 + h\bar{s}\right) \tag{2.5}$$

Therefore, the probability that the system has an averaged magnetisation $\bar{S} = m$ is given be:

$$\mathbb{P}(\bar{S} = m) = \frac{\Omega(m, d)}{Z_d(\beta, h)} e^{d\beta(\frac{1}{2}m^2 + hm)} \tag{2.6}$$

where $\Omega(m, d)$ is the number of different configurations $\boldsymbol{s} \in \{-1, +1\}^d$ that correspond to an average magnetisation $\bar{s} = m$:

$$\Omega(m, d) = \binom{d}{\frac{d+dm}{2}} = \frac{d!}{(\frac{d-dm}{2})!(\frac{d+dm}{2})!} \tag{2.7}$$

Recall that $d! \sim (d/e)^d$, which means that for large $d$ there are exponentially many configurations with a given averaged magnetisation. We can also write eq. (2.6) as:

$$\mathbb{P}(\bar{S} = m) = \frac{1}{Z_d(\beta, h)} e^{d\beta(\frac{1}{2}m^2 + hm) + \log\Omega(m,d)} \tag{2.8}$$

This allow us to show the following result:

**Theorem 1** (Large deviation principle)**.** In the *high-dimensional limit* $d \to \infty$ (a.k.a. *thermodynamic limit* in physics), the distribution eq. (2.6) satisfies a *large-deviation principle*:

$$\mathbb{P}(\bar{S} = m) \underset{d\to\infty}{\sim} e^{d(\phi(m;\beta,h) - \phi(m_\star;\beta,h))} \tag{2.9}$$

with $m_\star = \max_{m\in[-1,1]} \phi(m;\beta,h)$ and:

$$\phi(m;\beta,h) = \frac{\beta}{2}m^2 + \beta h m + H(m), \tag{2.10}$$

$$H(m) = -\frac{1+m}{2}\log\left(\frac{1+m}{2}\right) - \frac{1-m}{2}\log\left(\frac{1-m}{2}\right) \tag{2.11}$$

The function $H(m)$ is also known as the *binary entropy*.

**Exercise 1.** Prove Theorem 1 by:

1. Showing that eq. (2.7) can be bounded by:

$$\frac{e^{dH(m)}}{d+1} \leq \Omega(m,d) \leq e^{dH(m)} \tag{2.12}$$

   using Stirling's approximation.

2. Then showing that eq. (2.6) can be bounded by:

$$\frac{1}{d+1}\frac{1}{Z_d(\beta,h)}e^{d\phi(m;\beta,h)} \leq \mathbb{P}(\bar{S}=m) \leq \frac{1}{Z_d(\beta,h)}e^{d\phi(m;\beta,h)} \tag{2.13}$$

3. Then, by bounding the *free entropy density* $\Phi_d(\beta,h) = {}^1\!/_d \log Z_d(\beta,h)$:

$$\phi(m_\star) - \frac{\log(d(d+1))}{d} \leq \Phi_d(\beta,h) \leq \phi(m_\star) + \frac{\log(d+1)}{d} \tag{2.14}$$

4. Finally, conclude Theorem 1 by passing to the limit.

Large deviation principles such that eq. (2.9) are at the core of statistical physics, and essentially states that the probability that a configuration $\boldsymbol{s} \in \{-1,+1\}^d$ has takes an averaged magnetisation $\bar{s}$ different from $m_\star$ is exponentially small in $d$.

Note that the key object in this discussion, the potential function $\phi(m;\beta,h)$, is directly related to the asymptotic value of the free entropy density $\Phi_d(\beta,h) = {}^1\!/_d \log Z_d(\beta,h)$. In physics, it is more common to work with the *free energy density*, a simple rescaling of $\Phi_d$:

$$f_d(\beta,h) = -\frac{1}{\beta}\Phi_d(\beta,h) = -\frac{1}{\beta d}\log Z_d(\beta,h) \tag{2.15}$$

but this is ultimately a question of taste. In the limit $d\to\infty$, we have:

$$\lim_{d\to\infty} -\beta f_d(\beta,h) = \max_{m\in[-1,1]} \phi(m;\beta,h) \tag{2.16}$$

## 2.2 Phase transitions

The large deviation principle in Theorem 1 essentially translates the study of the high-dimensional Boltzmann-Gibbs distribution eq. (2.2) to the study of a low-dimensional potential function $\phi(m)$ of the macroscopic variable $m \in [-1,1]$. We now look in detail at this problem:

$$\max_{m\in\{-1,+1\}} \phi(m;\beta,h) \tag{2.17}$$

Note that the maximisation problem above is a competition between a convex ${}^\beta\!/_2 m^2 + \beta h m$ and a concave $H(m)$ function. In physics, these are known as the *energetic* and *entropic* contributions, respectively.

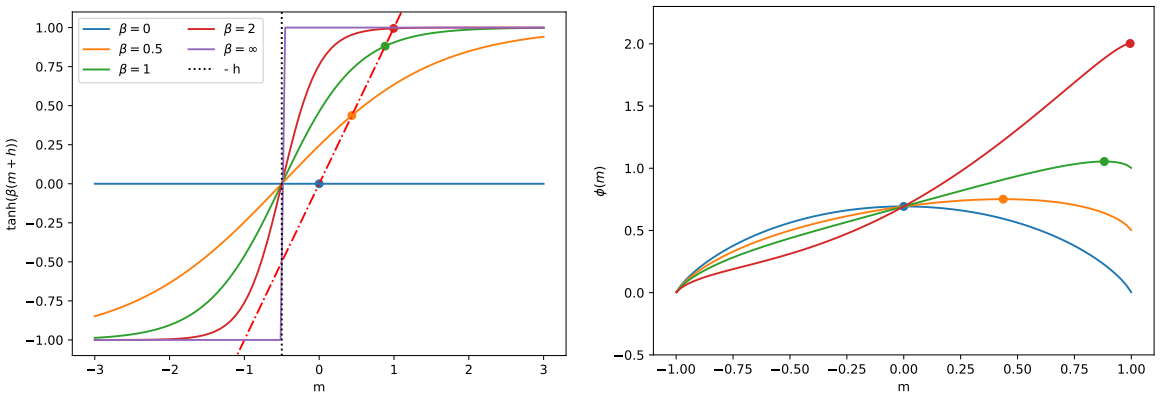


Figure 2: (Left) Right-hand side of (2.19) as a function of $m$, for fixed $h = 0.5$ and different values of the inverse temperature ($\beta$ solid lines). Solutions of (2.19) (dots) are given by the intersection of $f(m) = \tanh(\beta(h + m))$ with the line $f(m) = m$ (red dashed). (Right) Same picture in terms of the potential $\phi(m)$, where the solutions of (2.19) correspond to the global maximum of $\phi(m)$. Note that for $\beta \gg 1$ an unstable solution corresponding to a minimum of $\phi$ appear.

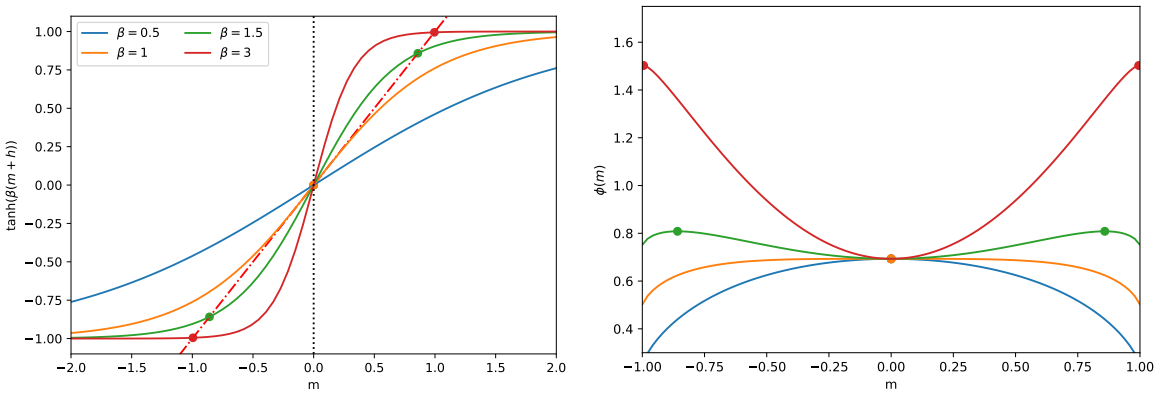


Figure 3: Same setting as Fig. 2, but for zero external field $h = 0$. For $\beta < 1$, the potential $\phi(m)$ has only one maximum corresponding to a disordered phase $m^\star = 0$. For $\beta > 1$, the system has two ordered ferromagnetic phases corresponding to the emergence of two symmetric global maxima $\pm m^\star$.

To look for the extremisers, we take the derivative and set to zero:

$$\partial_m \phi(m; \beta, h) \stackrel{!}{=} 0 \qquad \Leftrightarrow \qquad \beta(m + h) = \frac{1}{2} \log \frac{1 + m}{1 - m} \tag{2.18}$$

Recognising $\tanh^{-1}(x) = {}^1\!/_2 \log \frac{1+x}{1-x}$ we can rewrite:

$$m = \tanh(\beta(m + h)) \tag{2.19}$$

This equation, which is of the form $m = f(m)$, is known as a *self-consistent equation*, does not admit a closed-form solution. Nevertheless, it is a simple one-dimensional equation that can be easily solved in a computer by finding the intersection between $f(m) = m$ and $f(m) = tanh(\beta(m + h))$, see fig. 2.

### Zero external field and the second order transition

Note that at zero external field, the potential is a symmetric function of $m$: $\phi(m; \beta, 0) = \phi(-m; \beta, 0)$. At high-temperatures $\beta \to 0^+$, the entropic term dominates the potential, which has a single global

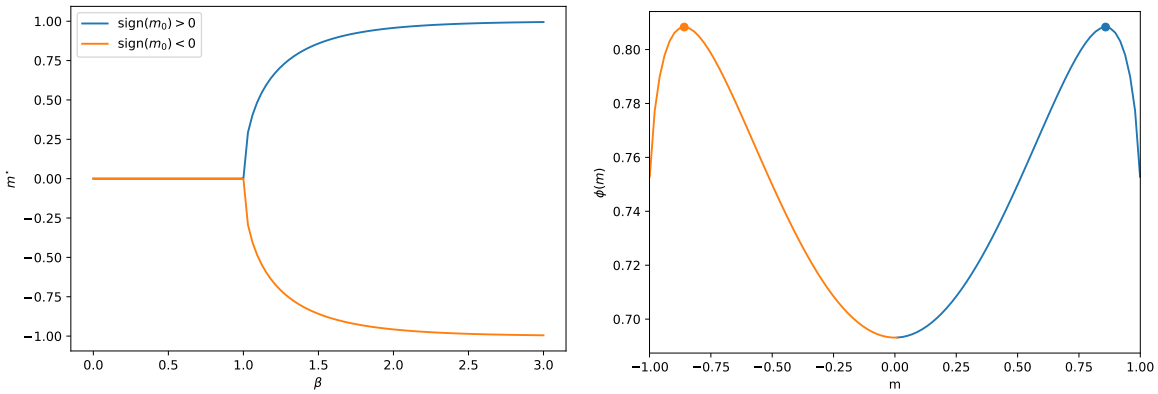


Figure 4: (Left) Fixed point $m_\star$ of the mean-field equation $m = \tanh(\beta m)$ above the critical temperature ($h = 0$, $\beta = 1.5$) as a function of the inverse temperature $\beta$. Depending on the sign of the initialisation $m_0$, we reach one of the two global maxima of $\phi(m)$ (right).

maxima at $m_\star = 0$, see fig. 1. Recalling the definition of $m$, Theorem 1 tell us that with high-probability on $d$, a typical sample from the Boltzmann-Gibbs distribution eq. (2.2) will have zero averaged magnetisation (i.e. roughly same number of $\pm 1$). In physics, this is known as the *paramagnetic* or *disordered phase*.

As the temperature $\beta^{-1}$ is lowered, $\phi(0; \beta, 0)$ continuously decrease, and at $\beta = \beta_c = 1$, $m = 0$ becomes a local minimum of $\phi(m; \beta, 0)$, with two global minimum[3] emerging, see fig. 3. This is the onset of *second order phase transition* towards a *ferromagnetic* or *ordered phase*. In the ferromagnetic phase, a typical configuration of the Boltzmann-Gibbs distribution has non-zero averaged magnetisation given by $|m_\star|$, the maximisers of $\phi(m; \beta, 0)$.

Note that the first derivative of the free energy with respect to $\beta$ (proportional to the entropy) remains a continuous function across the transition. However, we notice that the second derivative of the free energy is discontinuous, indicating this is a *second order phase transition*. This transition corresponds to a significant change in the statistical behaviour of the system at macroscopic scales: while for $\beta < 1$ a typical configuration from the Boltzmann-Gibbs distribution has no net magnetisation $m_\star = \langle \bar{S} \rangle_\beta \approx 0$ (*disordered phase*), for $\beta > 1$ a typical configuration has a net magnetisation $|m_\star| = |\langle \bar{S} \rangle_\beta| > 0$ (*ordered phase*). This is an example of an important concept in Physics known as *spontaneous symmetry breaking*: while the Hamiltonian of the system is invariant under the $\mathbb{Z}_2$ symmetry $\bar{s} \to -\bar{s}$, for $\beta > 1$ a typical draw of the Gibbs-Boltzmann distribution $\boldsymbol{S} \sim \mathbb{P}_{N,\beta}$ breaks this symmetry at the macroscopic level. Second order transitions carry a rich phenomenology. Since the transition is second order (i.e. continuous first derivative), the critical temperature can be obtained by studying the expansion of the free energy potential around $m = 0$:

$$\phi(m; \beta, 0) \underset{m \to 0}{=} \log 2 + \frac{m^2}{2}(\beta - 1) + O(m^3)$$

which give us the critical $\beta_c = 1$ as the point in which the second derivative changes sign ($m = 0$ goes from a minimum to a maximum). It is also useful to have the picture in terms of the saddle-point equation:

$$m = \tanh(\beta m).$$

The fact that $m = 0$ is always a fixed point of this equation signals it is always an extremizer of the free energy potential. From this perspective, the critical temperature $\beta_c = 1$ corresponds to

[3]Note that the symmetric $\phi(m; \beta, 0) = \phi(-m; \beta, 0)$ implies that except for $m = 0$, extremisers must always come in pairs.

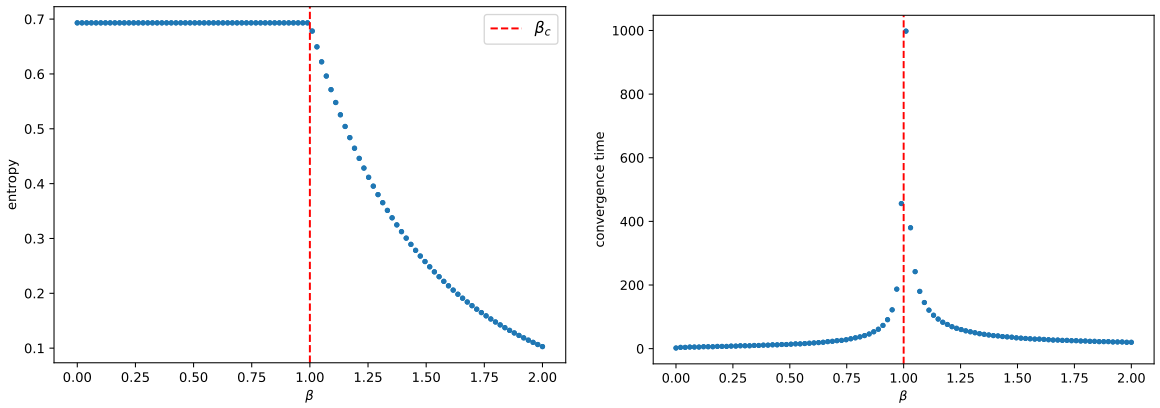


Figure 5: (Left) Entropy as a function of the inverse temperature $\beta$ at zero external field $h = 0$. Note that the entropy is a continuous function of the temperature, with a cusp at the critical point $\beta_c = 1$, indicating that its derivative (proportional to the second derivative of the free energy) has a discontinuity. (Right) Convergence time of the saddle-point equation as a function of the inverse temperature $\beta$ at zero external field $h = 0$. Note the critical slowing down close to the second order critical point $\beta_c = 1$.

a change of stability of this fixed point. Seeing the saddle-point equations as a discrete dynamical system $m^{t+1} = f(m^t)$, the stability of a fixed point can be determined by looking at the Jacobian of the update function $f : [-1, 1] \to [-1, 1]$ around the fixed point $m = 0$:

$$f(x) = \tanh(\beta x) \underset{m\to 0}{=} \beta x + O(m^3) \tag{2.20}$$

For $\beta < 1$, the fixed point is *stable* (attractor/sink of the dynamics), while for $\beta > 1$ it becomes an *unstable* (repeller/source of the dynamics). Note that this implies that close to the transition $\beta \approx 1^+$, iterating the saddle point equations starting close to zero $m^{t=0} = \epsilon \ll 1$ (but not exactly at zero) takes long to converge to a non-zero magnetisation $m > 0$, with the time diverging as we get closer to the transition. This phenomenon is known is physics as the *critical slowing down*, and together with the expansion of the free energy and the stability analysis of the equations give yet another way to characterise a second order critical point. See fig. 5 (right) for an illustration.

### Finite external field and the first order transition

Turning on the external magnetic field $h \neq 0$ can dramatically change the discussion above. First, note that the Hamiltonian loses the $\mathbb{Z}_2$ symmetry: this is known in Physics as *explicit symmetry breaking*. At high temperatures $\beta \to 0^+$, the free energy potential is convex, with a single minimum at $m = h$ aligned with the field. As temperature is lowered and we enter what previously was the ferromagnetic phase ($\beta > 1$), two behaviors are possible. For small $h$, the field simply has the effect of breaking the symmetry between the previous two global minima and making the with opposite sign a local minimum, see fig. 6 (left). In this situation, even though the equilibrium free energy is given by the now unique global minimum of the potential, the presence of a local minimum has an important effect in the dynamics. Indeed, if we initialize the saddle-point equations close to the magnetisation corresponding to the local minimum, it will converge to this local minimum, since it is also a stable fixed point of the corresponding dynamical system, see fig. 7 (left). This phenomenon is known as *metastability* in Physics. Note that metastability can be a misleading name, since in the thermodynamic limit $N \to \infty$ metastable states are stable fixed points of the free energy potential. However, at finite system size $N$, the system will dynamically reach equilibrium in a time of order $t = O(e^N)$. Metastability will play a major role in the Statistical Physics analysis of inference problems, since it is closely related to algorithmic hardness.

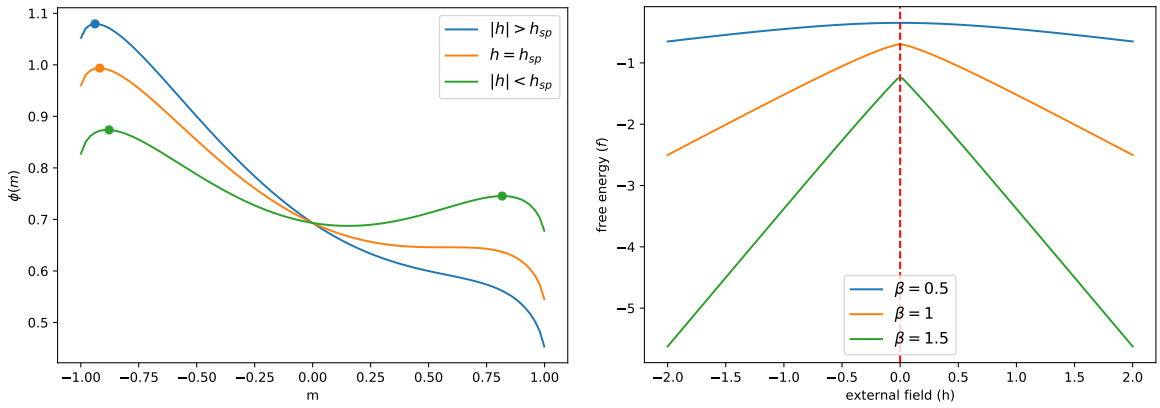


Figure 6: (Left) Free energy potential $\phi(m)$ as a function of $m$ for fixed inverse temperature $\beta = 1.5$ and varying external field $h < 0$. Note that the free energy potential has a local maximum for $|h| > h_{sp}$ that disappears at the spinodal transition $h = h_{sp}$. (Right) Free energy as a function of the external field $h$ at different temperatures. Note the non-analytical cusp at $h = 0$.

As the external field $h$ is increased, the difference in the free energy potential between the two minima increases, and eventually at a critical field $h_{\rm sp}$, known as the *spinodal point*, the local minimum disappears, making the potential convex again, see fig. 6 (left).

$$h_{\rm sp}(\beta) = \pm\sqrt{\frac{1}{\beta}\left(1-\frac{1}{\beta}\right)} \mp \frac{1}{\beta}\tanh^{-1}\left(\sqrt{1-\frac{1}{\beta}}\right), \qquad \beta > 1$$

From this discussion, it is clear that for $\beta > 1$ the magnetisation (which is the derivative of the free entropy with respect to $h$) has a discontinuity at $h = 0$, since for $h \neq 0$ we have a non-zero magnetisation and for $h = 0$ we are in the paramagnetic phase $h = 0$. This is a first order phase transition of the system with respect to the external field $h$, see fig. 6 (right). Note that as a consequence of metastability, in the region $|h| < |h_{\rm sp}|$ the system magnetisation will depend of the state in which it was initially prepared. This memory of the initial state is known as *hysteresis* in Physics, see fig. 7 (right).

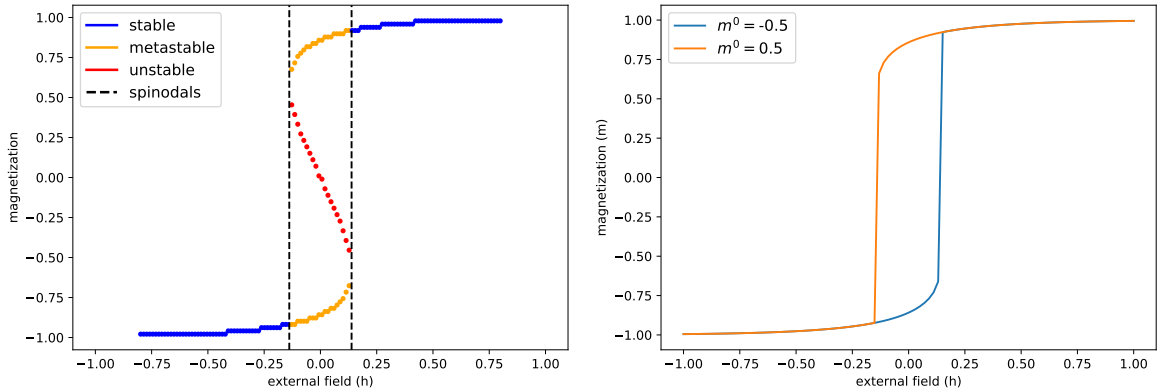


Figure 7: (Left) Stable, metastable and unstable branches of the magnetisation as a function of the external field at fixed inverse temperature $\beta = 1.5$. (Right) magnetisation obtained by iterating the saddle-point equations from different initial conditions $m^{t=0}$ as a function of the external field $h$ and fixed inverse temperature $\beta = 1.5$. Note the hysteresis loop: point at which the magnetisation discontinuously jumps from negative to positive depends on the initial state of the system.

### 2.3 Historical note

> itJe me propose de montrer ici que l'on peut fonder une théorie du ferromagnétisme sur une hypothèse extrêmement simple concernant ces actions mutuelles. Je suppose que chaque molécule éprouve de la part de l'ensemble des molécules environnantes une action égale à celle d'un champ uniforme proportionnel à l'intensité d'aimantation et de même direction qu'elle.
> Pierre Weiss, itL'hypothèse du champ moléculaire et la propriété ferromagnétique, 1907.

The Curie-Weiss model was first introduced by Pierre Weiss in (Weiss, 1907) to explain the experimental observation from Pierre Curie that magnets lose their magnetic properties when heated above a certain temperature (now known as the *Curie temperature*) (Curie, 1895) (a.k.a. the *ferromagnetic transition*): It can be seen as a fully-connected approximation of the Ising model, although it is worth noting the Curie-Weiss model precedes Ernst Ising's work. (Ising, 1925).

# Part II
# Lecture 2 — The toolbox

Last lecture have motivated how statistical physics provides a framework to think about probability distributions defined in high-dimensional spaces. We went in detail over one example, the Curie-Weiss model for the ferromagnetism, and saw how a large deviation principle allow us to characterise the high-dimensional properties of the Boltzmann-Gibbs distribution through a set of low-dimensional self-consistent equations for the magnetization, a macroscopic variable able to summarise the phases of the system across different temperatures. Finally, we did a historical tour over how these ideas, initially developed to understand the collective behaviour of matter leaked to other fields where high-dimensionality plays a central role, such as computer science, machine learning and neuroscience.

In this lecture, we turn our attention to the tools that statistical physics can offer. In particular, we will focus on two of the most useful tools: the *replica method* and *approximate message passing*. For concreteness, we introduce these two methods in the context of a model which will serve as the starting point for our discussion about neural networks: the Gaussian Covariate model.

## 3 Gaussian covariate model

Let $\boldsymbol{u}_i \in \mathbb{R}^d$ and $\boldsymbol{v}_i \in \mathbb{R}^p$ denote $i = 1, \cdots, n$ independent jointly Gaussian vectors:

$$(\boldsymbol{u}_i, \boldsymbol{v}_i) \sim \mathcal{N}\left(\mathbf{0}_{p+d}, \begin{bmatrix} {}^1\!/_d \boldsymbol{\Psi} & {}^1\!/_{\sqrt{pd}} \boldsymbol{\Phi} \\ {}^1\!/_{\sqrt{pd}} \boldsymbol{\Phi}^\top & {}^1\!/_p \boldsymbol{\Omega} \end{bmatrix}\right). \tag{3.1}$$

with positive semi-definite covariance matrices $\boldsymbol{\Psi} \in \mathbb{R}^{d\times d}$, $\boldsymbol{\Omega} \in \mathbb{R}^{p\times p}$ and $\boldsymbol{\Phi} \in \mathbb{R}^{p\times d}$ such that:

$$\operatorname{Tr} \boldsymbol{\Psi} = O(d), \qquad \operatorname{Tr} \boldsymbol{\Omega} = O(p), \qquad \operatorname{Tr}\left(\Phi^\top \Phi\right) = O(\sqrt{dp}) \tag{3.2}$$

We consider the following generalised linear estimation task:

1. **Data:** Observations $y_i \in \mathbb{R}$ are generated from the covariates $\boldsymbol{u}_i \in \mathbb{R}^d$:

$$y_i \sim P_\star\left(\cdot | \langle \boldsymbol{\beta}_\star, \boldsymbol{u}_i \rangle\right), \qquad i \in [n] \tag{3.3}$$

where $\boldsymbol{\beta}_\star \in \mathbb{R}^d$ is a signal vector (which might be random or not) of norm $||\boldsymbol{\beta}_\star||_2^2 = O(d)$ and $p_\star(y|z)$ is a given likelihood function.

2. **Model:** The goal of the statistician is to estimate the signal $\boldsymbol{\beta}_\star \in \mathbb{R}^d$ from the observation of $\mathcal{D} = \{(\boldsymbol{v}_i, y_i) \in \mathbb{R}^{p+1} : i \in [n]\}$. For that, she follows a Bayesian approach by postulating a likelihood $\psi(y|\langle \boldsymbol{\theta}, \boldsymbol{v}\rangle)$ and a prior distribution $\varphi$ over $\mathbb{R}^p$, and constructs the following posterior distribution:

$$P(\boldsymbol{\theta}|\mathcal{D}) = \frac{1}{Z_p(\mathcal{D})}\varphi(\boldsymbol{\theta}) \prod_{i\in[n]} \psi\left(y_i|\langle \boldsymbol{\theta}, v_i\rangle\right) \tag{3.4}$$

from which she can choose different estimators, e.g. the maximum-a-posteriori estimator $\hat{\boldsymbol{\theta}}_{\text{map}} = \text{argmax } P(\boldsymbol{\theta}|\mathcal{D})$ or the the posterior mean $\hat{\boldsymbol{\theta}} = \mathbb{E}[\boldsymbol{\theta}|\mathcal{D}]$.

⚠ Note that the posterior distribution introduced in eq. (3.4) is itself a random quantity, since it is a function of the data $\mathcal{D} = \{(\boldsymbol{v}_i, y_i) \in \mathbb{R}^{p+1} : i \in [n]\}$.

*Remark* 1. We assume $P_{\boldsymbol{\beta}}$ and $P_\star$ are probability densities. However, we are more flexible and don't necessarily require $\varphi$ and $\psi$ to be normalised, although the Boltzmann-Gibbs distribution is always normalised.

Although the task above is framed in a Bayesian framework, it encompasses several problems of interest in statistics and signal processing, both Bayesian and not. Below, we give a few examples.

**Example 1** (Bayes-optimal estimation)**.** In the best case scenario, the statistician knows exactly the process in eq. (3.3) that generated the observed data, such as the covariates $\boldsymbol{u}_i \in \mathbb{R}^d$, the signal distribution $P_{\boldsymbol{\beta}}$ and the likelihood $P_\star$, and the goal is to estimate the specific realisation of the signal $\boldsymbol{\beta}_\star \sim P_{\boldsymbol{\beta}}$. Using this information for inference, the statistician aligns its model with the data generating model, by doing inference directly on $(\boldsymbol{u}_i, y_i)_{i\in[n]}$ (with $p = d$ in particular) and choosing $\varphi = P_{\boldsymbol{\beta}}$ and $\psi = P_\star$, and therefore the Bayesian posterior reads:

$$P(\boldsymbol{\theta}|\mathcal{D}) = \frac{1}{Z_d}P_{\boldsymbol{\beta}}(\boldsymbol{\theta}) \prod_{i=1}^{n} P_\star(y_i|\langle \boldsymbol{\theta}, \boldsymbol{u}_i\rangle) \tag{3.5}$$

This setting is known as Bayes-optimal inference, and information theoretically corresponds to the best-case scenario for inference. In particular, in this case the posterior mean is known to achieve the best possible mean-squared error, also known as *minimum mean-squared error* (MMSE):

$$\text{mmse} = \min_{\hat{\boldsymbol{\theta}}} \mathbb{E}\left[||\boldsymbol{\beta} - \boldsymbol{\theta}(\mathcal{D})||_2^2\right] = \mathbb{E}\left[||\boldsymbol{\beta} - \mathbb{E}[\boldsymbol{\theta}|\mathcal{D}]||_2^2\right] \tag{3.6}$$

Bayes-optimality also implies an important property known as the Nishimori identity (Nishimori, 1980; Iba, 1999). Let $\boldsymbol{\theta}_1, \dots, \boldsymbol{\theta}_k \sim P(\boldsymbol{\theta}|\mathcal{D})$ denote $k$ samples from the posterior distribution in eq. (3.5), and let $\boldsymbol{\beta}_\star \sim P_{\boldsymbol{\beta}}$ denote the signal. The Nishimori identity states that:

$$\mathbb{E}\left[\mathbb{E}\left[f(y, \boldsymbol{\theta}_1, \dots, \boldsymbol{\theta}_k)|\mathcal{D}\right]\right] = \mathbb{E}[\mathbb{E}\left[(y, \boldsymbol{\theta}_1, \dots, \boldsymbol{\beta}_\star)|\mathcal{D}\right]] \tag{3.7}$$

where the inner expectation is over the posterior distribution, and the outer over the data distribution. In words: expectations over the posterior distribution are equivalent to expectation over the signal distribution. As we will see later, this identity has important consequences in the theoretical analysis.

⚠ The Nishimori identities only hold in the Bayes-optimal setting. It is false for mismatched estimation.

**Example 2** (Empirical risk minimisation)**.** Let $\ell : \mathbb{R}^2 \to \mathbb{R}_+$ denote a loss function (e.g. the square $\ell(y, \hat{y}) = {}^1\!/_2(y - \hat{y})^2$ or logistic loss $\ell(y, \hat{y}) = \log\left(1 + e^{-y\hat{y}}\right)$) and $r : \mathbb{R}^d \to \mathbb{R}_+$ denote a regulariser (e.g. the ridge $r(\boldsymbol{\theta}) = {}^\lambda\!/_2||\boldsymbol{\theta}||_2^2$ or lasso $r(\boldsymbol{\theta}) = ||\boldsymbol{\theta}||_1$ penalties). Then letting:

$$\psi(y|\langle \boldsymbol{\theta}, \boldsymbol{v}\rangle) \propto e^{\beta\ell(y, \langle \boldsymbol{\theta}, \boldsymbol{v}\rangle)}, \qquad \varphi(\boldsymbol{\theta}) \propto e^{\beta r(\boldsymbol{\theta})} \tag{3.8}$$

we see that the posterior distribution in eq. (3.4) gives more weight to configurations $\boldsymbol{\theta} \in \mathbb{R}^p$ that have lower empirical risk:

$$\hat{R}(\boldsymbol{\theta}; \mathcal{D}) = \frac{1}{n} \sum_{i=1}^{n} \ell(y_i, \langle \boldsymbol{\theta}, \boldsymbol{v}_i \rangle) + r(\boldsymbol{\theta}) \tag{3.9}$$

In particular, for $\beta \to \infty$ the MAP estimator will coincide with the empirical risk minimiser:

$$\hat{\boldsymbol{\theta}} \in \operatorname*{argmin}_{\boldsymbol{\theta} \in \mathbb{R}^p} \hat{R}(\boldsymbol{\theta}; \mathcal{D}) \tag{3.10}$$

**Example 3** (Maximum likelihood estimation)**.** Consider a well-specified setting where $p = d$ and $\boldsymbol{u} = \boldsymbol{v}$. Maximum likelihood estimation consists of taking the MAP estimator with $P(y|z) = P_\star(y|z)$ and $\varphi = 1$:

$$\hat{\boldsymbol{\theta}} = \operatorname{argmin} - \frac{1}{n} \sum_{i=1}^{n} \log P_\star(y | \langle \boldsymbol{\theta}, \boldsymbol{v}_i \rangle) \tag{3.11}$$

# 4 Tool I: The replica method

The replica method is a tool originally developed in the context of statistical physics of glasses which aims at deriving a large deviations principle of the type we have seen for the Curie-Weiss model in Equation (2.9) for random high-dimensional probability measures, such as the posterior in eq. (3.4).

Our key goal is to compute the limiting free entropy density of the model:

$$\Phi = \lim_{p \to \infty} \frac{1}{p} \mathbb{E}_{\mathcal{D}}[\log Z_d(\mathcal{D})] \tag{4.1}$$

where the expectation is over the data generating process:

$$P(\mathcal{D}) = \prod_{i=1}^{n} P(\boldsymbol{v}_i, y_i) = P_{\boldsymbol{\beta}}(\boldsymbol{\beta}_\star) \prod_{i=1}^{n} P_\star(y_i | \langle \boldsymbol{\beta}_\star, \boldsymbol{u}_i \rangle) P(\boldsymbol{u}_i, \boldsymbol{v}_i) \tag{4.2}$$

and without loss of generality we considered the case where $\boldsymbol{\beta}_\star \sim P_{\boldsymbol{\beta}}$ is random, with the deterministic case given by $P_{\boldsymbol{\beta}}(\boldsymbol{\beta}) = \delta(\boldsymbol{\beta} - \boldsymbol{\beta}_\star)$.

## 4.1 Sketch of the computation in six steps

### 4.1.1 Step 1: The replica trick

Except for very particular cases, taking the expectation explicitly in eq. (4.1) is not tractable. The replica trick avoids the explicit computation of the logarithm by using the following identity:

$$\log Z_p = \lim_{s \to 0^+} \partial_s Z_p^s \tag{4.3}$$

Switching the limit and the expectation, this reduces the computation of the free entropy density to the computation of moments of the partition function $Z_d$:

$$\phi = \lim_{p \to \infty} \frac{1}{p} \lim_{s \to 0^+} \partial_s \mathbb{E}_{\mathcal{D}}[Z_d^s] \tag{4.4}$$

### 4.1.2 Step 2: Computation of moments

Integer powers of the partition function are given by:

$$Z_d^s = \left( \int_{\mathbb{R}^p} \mathrm{d}\boldsymbol{\theta}\ \varphi(\boldsymbol{\theta}) \prod_{i=1}^n \psi\left(y_i | \langle \boldsymbol{\theta}, \boldsymbol{v}_i \rangle\right) \right)^s = \prod_{a=1}^s \int_{\mathbb{R}^p} \mathrm{d}\boldsymbol{\theta}^a\ \varphi(\boldsymbol{\theta}^a) \prod_{i=1}^n \psi\left(y_i | \langle \boldsymbol{\theta}^a, \boldsymbol{v}_i \rangle\right) \tag{4.5}$$

Computing a $s$ moment of $Z_d$ is equivalent to computing the expectation over $s$ independent copies or *replicas* of the system, hence the name of the method. For that reason, we also refer to the index $a$ as the *replica index*. Now taking the expectation with respect to the data distribution eq. (4.2):

$$\begin{aligned}
\mathbb{E}_{\mathcal{D}}[Z_d^s] &= \mathbb{E}_{\mathcal{D}} \left[ \prod_{a=1}^s \int_{\mathbb{R}^p} \mathrm{d}\boldsymbol{\theta}^a\ \varphi(\boldsymbol{\theta}^a) \prod_{i=1}^n \psi\left(y_i | \langle \boldsymbol{\theta}^a, \boldsymbol{v}_i \rangle\right) \right] \\
&= \int_{\mathbb{R}^d} \mathrm{d}\boldsymbol{\beta} P_{\boldsymbol{\beta}}(\boldsymbol{\beta}) \int_{\mathbb{R}^{p\times s}} \left( \prod_{a=1}^s \mathrm{d}\boldsymbol{\theta}^a \varphi(\boldsymbol{\theta}^a) \right) \prod_{i=1}^n \int_{\mathbb{R}} \mathrm{d}y_i\ \mathbb{E}_{(\boldsymbol{u}_i, \boldsymbol{v}_i)} \left[ P_\star\left(y_i | \langle \boldsymbol{\beta}, \boldsymbol{u}_i \rangle\right) \prod_{a=1}^n \psi\left(y_i | \langle \boldsymbol{\theta}^a, \boldsymbol{v}_i \rangle\right) \right] \\
&= \int_{\mathbb{R}^d} \mathrm{d}\boldsymbol{\beta} P_{\boldsymbol{\beta}}(\boldsymbol{\beta}) \int_{\mathbb{R}^{p\times s}} \left( \prod_{a=1}^s \mathrm{d}\boldsymbol{\theta}^a \varphi(\boldsymbol{\theta}^a) \right) \left\{ \int_{\mathbb{R}} \mathrm{d}y\ \mathbb{E}_{(\boldsymbol{u}, \boldsymbol{v})} \left[ P_\star\left(y | \langle \boldsymbol{\beta}, \boldsymbol{u} \rangle\right) \prod_{a=1}^n \psi\left(y | \langle \boldsymbol{\theta}^a, \boldsymbol{v} \rangle\right) \right] \right\}^n
\end{aligned} \tag{4.6}$$

where we used the assumption that the $n$ samples $(\boldsymbol{v}_i, y_i) \in \mathcal{D}$ are drawn independently. Note that the likelihoods $P_\star$ and $\psi$ only depend on the Gaussian covariates $(\boldsymbol{u}_i, \boldsymbol{v}_i)$ through the following inner products, also known in the statistical physics jargon as the *local-fields*:

$$\nu = \langle \boldsymbol{\beta}, \boldsymbol{u} \rangle \in \mathbb{R}, \qquad \lambda^a = \langle \boldsymbol{\theta}, \boldsymbol{v} \rangle \in \mathbb{R}, \qquad a \in [s]. \tag{4.7}$$

Conditionally on the weights $\boldsymbol{\beta} \in \mathbb{R}^d$ and $\boldsymbol{\theta} \in \mathbb{R}^p$, these $s+1$ scalar quantities are jointly Gaussian variables:

$$P(\nu, \lambda^1, \dots, \lambda^s | \boldsymbol{\theta}, \boldsymbol{\beta}) = \mathcal{N}\left( \mathbf{0}_{s+1}, \begin{bmatrix} \rho & \boldsymbol{m} \\ \boldsymbol{m} & \boldsymbol{Q} \end{bmatrix} \right) \tag{4.8}$$

where we have defined the so-called *overlaps*:

$$\rho \coloneqq \mathbb{E}[\nu^2] = {}^1\!/_d \langle \boldsymbol{\beta}, \boldsymbol{\Psi}\boldsymbol{\beta} \rangle, \qquad m^a \coloneqq \mathbb{E}[\nu \lambda^a] = {}^1\!/_{\sqrt{dp}} \langle \boldsymbol{\theta}^a, \boldsymbol{\Phi}\boldsymbol{\beta} \rangle, \qquad Q^{ab} \coloneqq \mathbb{E}[\lambda^a \lambda^b] = {}^1\!/_p \langle \boldsymbol{\theta}^a, \boldsymbol{\Omega}\boldsymbol{\theta}^b \rangle \tag{4.9}$$

Note that $\rho, m^a, q^{ab} = O(1)$ due to eq. (3.2). This allow us to rewrite the average term in eq. (4.6) as:

$$\mathbb{E}_{(\boldsymbol{u}, \boldsymbol{v})} \left[ P_\star\left(y | \langle \boldsymbol{\beta}, \boldsymbol{u} \rangle\right) \prod_{a=1}^n \psi\left(y | \langle \boldsymbol{\theta}^a, \boldsymbol{v} \rangle\right) \right] = \mathbb{E}_{(\nu, \lambda^1, \dots, \lambda^s)} \left[ P_\star\left(y | \nu\right) \prod_{a=1}^n \psi\left(y | \lambda^a\right) \middle| \boldsymbol{\beta}, \boldsymbol{\theta}^a \right] \tag{4.10}$$

Putting together, this gives:

$$\mathbb{E}_{\mathcal{D}}[Z_d^s] = \int_{\mathbb{R}^d} \mathrm{d}\boldsymbol{\beta} P_{\boldsymbol{\beta}}(\boldsymbol{\beta}) \int_{\mathbb{R}^{p\times s}} \left( \prod_{a=1}^s \mathrm{d}\boldsymbol{\theta}^a \varphi(\boldsymbol{\theta}^a) \right) \Psi_y^{(s)}(\rho, \boldsymbol{m}, \boldsymbol{Q})^n \tag{4.11}$$

where we introduced:

$$\Psi_y^{(s)}(\rho, \boldsymbol{m}, \boldsymbol{Q}) = \int_{\mathbb{R}} \mathrm{d}y\ \mathbb{E}_{(\nu, \lambda^1, \dots, \lambda^s)} \left[ P_\star\left(y | \nu\right) \prod_{a=1}^n \psi\left(y | \lambda^a\right) \middle| \boldsymbol{\beta}, \boldsymbol{\theta}^a \right] \tag{4.12}$$

*Remark* 2. It is useful to stop at this point and reflect on what exactly we have achieved with this rewriting. Initially, we had an expectation of independent replicas of the system over the high-dimensional Gaussian covariances $(\boldsymbol{u}, \boldsymbol{v}) \in \mathbb{R}^{d+p}$ given by eq. (4.6). In eq. (4.11), we have traded this high-dimensional Gaussian expectation for a low-dimensional[4] Gaussian expectation over the Gaussian local-fields $(\nu, \lambda^1, \dots, \lambda^s) \in \mathbb{R}^{s+1}$. However, this low-dimensional expectation now correlates different replicas through the matrix $\boldsymbol{Q} \in \mathbb{R}^{s \times s}$.

This is a general pattern in replica computations: initially independent, replicas become correlated after the disorder average.

### 4.1.3 Step 3: Energy-entropy decomposition

A close inspection of eq. (4.11) reveals that the dependency on the weights $\boldsymbol{\beta}, \boldsymbol{\theta}^a$ is only through the overlap parameters $(\rho, m^a, Q^{ab})$. Indeed, this is the only source of correlation between the expression inside the brackets $\{\cdot\}^n$ and the expectation over $\boldsymbol{\beta}$ and $\boldsymbol{\theta}^a$. The goal of this step is to decouple these two terms.

Recall that $X \sim p$ is a random variable and $Y = f(X)$ for a deterministic function $f$, the density of $Y$ reads:

$$p(y) = \mathbb{E}_x[\delta(y - f(x)] = \int \mathrm{d}x \; p(x)\delta(y - f(x)) \tag{4.13}$$

We now apply these identities over the overlaps, which are functions of the weights. This allow us to rewrite eq. (4.11) as:

$$\mathbb{E}_{\mathcal{D}}[Z_d^s] = \int \mathrm{d}\rho \int \prod_{a=1}^{s} \mathrm{d}m^a \int \prod_{a,b=1}^{s} \mathrm{d}Q^{ab} \; V^{(s)}(\rho, \boldsymbol{m}, \boldsymbol{Q}) \Psi_y^{(s)}(\rho, \boldsymbol{m}, \boldsymbol{Q})^n \tag{4.14}$$

where:

$$V^{(s)}(\rho, \boldsymbol{m}, \boldsymbol{Q}) = \mathbb{E}_{\boldsymbol{\beta}, \boldsymbol{\theta}^1, \dots, \boldsymbol{\theta}^s} \left[ \delta\left(\rho - {}^1\!/_d \langle \boldsymbol{\beta}, \boldsymbol{\Psi}\boldsymbol{\beta} \rangle\right) \prod_{a=1}^{s} \delta\left(m^a - {}^1\!/_{\sqrt{pd}} \langle \boldsymbol{\theta}^a, \boldsymbol{\Phi}\boldsymbol{\beta} \rangle\right) \prod_{a,b=1}^{s} \delta\left(Q^{ab} - {}^1\!/_p \langle \boldsymbol{\theta}^a, \boldsymbol{\Omega}\boldsymbol{\theta}^b \rangle\right) \right] \tag{4.15}$$

Note that $V^{(s)}$ is a volume term which counts how many configurations of $\boldsymbol{\beta}, \boldsymbol{\theta}^1, \dots, \boldsymbol{\theta}^s$ there are with a given overlap $\rho, \boldsymbol{m}, \boldsymbol{Q}$, similar to the term $\Omega$ in eq. (2.7) of our analysis of the Curie-Weiss model in Section 2. Similarly, $\Psi_y^{(s)}$ can be interpreted as an energetic term, and eq. (4.14) as a decomposition of the free entropy in an energetic and an entropic contribution.

Just as in the Curie-Weiss model, we expect the number of configurations with a given overlap to be exponential in the dimensions $d, p$. Therefore, we seek an exponential representation $V^{(s)}(\rho, \boldsymbol{m}, \boldsymbol{Q}) \sim e^{pS(\rho, \boldsymbol{m}, \boldsymbol{Q})}$ where $S$ defines the entropy. For that, we use the Fourier representation of the $\delta$-function:

$$\delta(x - a) = \int \frac{\mathrm{d}k}{2\pi} e^{-ik(x-a)}. \tag{4.16}$$

To write:

$$\begin{aligned}
\delta\left(\rho - {}^1\!/_d \langle \boldsymbol{\beta}, \boldsymbol{\Psi}\boldsymbol{\beta} \rangle\right) &= d \int \frac{\mathrm{d}\hat{\rho}}{2\pi} e^{-i\hat{\rho}(d\rho - \langle \boldsymbol{\beta}, \boldsymbol{\Psi}\boldsymbol{\beta} \rangle)} \\
\prod_{a=1}^{s} \delta\left(m^a - {}^1\!/_{\sqrt{pd}} \langle \boldsymbol{\theta}^a, \boldsymbol{\Phi}\boldsymbol{\beta} \rangle\right) &= (dp)^{s/2} \int \prod_{a=1}^{s} \frac{\mathrm{d}\hat{m}^a}{2\pi} e^{-i \sum_{a=1}^{s} \hat{m}^a (\sqrt{pd} m^a - \langle \boldsymbol{\theta}^a, \boldsymbol{\Phi}\boldsymbol{\beta} \rangle)} \\
\prod_{a,b=1}^{s} \delta\left(Q^{ab} - {}^1\!/_p \langle \boldsymbol{\theta}^a, \boldsymbol{\Omega}\boldsymbol{\theta}^b \rangle\right) &= p^{s^2} \int \prod_{a,b=1}^{s} \frac{\mathrm{d}\hat{Q}^{ab}}{2\pi} e^{-i \sum_{a,b=1}^{s} \hat{Q}^{ab} (pQ^{ab} - \langle \boldsymbol{\theta}^a, \boldsymbol{\Omega}\boldsymbol{\theta}^b \rangle)}
\end{aligned} \tag{4.17}$$

[4] Recall we eventually want to take $s \to 0^+$

Therefore, up to multiplicative constants the volume term reads:

$$V^{(s)}(\rho, \boldsymbol{m}, \boldsymbol{Q}) \propto \int \mathrm{d}\hat{\rho} e^{-id\rho\hat{\rho}} \int \prod_{a=1}^{s} \mathrm{d}\hat{m}^a e^{-i\sqrt{dp}\sum_{a=1}^{s} m^a \hat{m}^a} \int \mathrm{d}\hat{Q}^{ab} e^{-ip\sum_{a,b=1}^{s} Q^{ab}\hat{Q}^{ab}} \times$$
$$\times \mathbb{E}_{\boldsymbol{\beta},\boldsymbol{\theta}^1,\ldots,\boldsymbol{\theta}^s} \left[ e^{i\hat{\rho}\langle \boldsymbol{\beta}, \boldsymbol{\Psi}\boldsymbol{\beta}\rangle + i\sum_{a=1}^{s} \hat{m}^a \langle \boldsymbol{\theta}^a, \boldsymbol{\Phi}\boldsymbol{\beta}\rangle + i\sum_{a,b=1}^{s} \hat{Q}^{ab}\langle \boldsymbol{\theta}^a, \boldsymbol{\Omega}\boldsymbol{\theta}^b\rangle} \right] \tag{4.18}$$

Putting together, this allow us to write eq. (4.14) as the integral over an exponential:

$$\mathbb{E}_{\mathcal{D}}[Z_d^s] \propto \int \mathrm{d}\rho \mathrm{d}\hat{\rho} \int \prod_{a=1}^{s} \mathrm{d}m^a \mathrm{d}\hat{m}^a \int \prod_{a,b=1}^{s} \mathrm{d}Q^{ab} \mathrm{d}\hat{Q}^{ab} e^{p\Phi^{(s)}(\rho,\hat{\rho},\boldsymbol{m},\hat{\boldsymbol{m}},\boldsymbol{Q},\hat{\boldsymbol{Q}})} \tag{4.19}$$

with:

$$\Phi^{(s)}(\rho, \hat{\rho}, \boldsymbol{m}, \hat{\boldsymbol{m}}, \boldsymbol{Q}, \hat{\boldsymbol{Q}}) = -\frac{1}{\gamma}\rho\hat{\rho} - \frac{1}{\sqrt{\gamma}}\sum_{a=1}^{s} m^a \hat{m}^a - \sum_{a,b=1}^{s} Q^{ab}\hat{Q}^{ab} + \alpha\Psi_y^s(\rho, \boldsymbol{m}, \boldsymbol{Q}) + \Psi_\theta(\hat{\rho}, \hat{\boldsymbol{m}}, \hat{\boldsymbol{Q}}) \tag{4.20}$$

where:

$$\alpha = {}^{n}/_{p}, \qquad \gamma := {}^{p}/_{d} \tag{4.21}$$

and we have defined:

$$\Psi_\theta^{(s)}(\hat{\rho}, \hat{\boldsymbol{m}}, \hat{\boldsymbol{Q}}) = \frac{1}{p}\log \int \mathrm{d}\boldsymbol{\beta} P_{\boldsymbol{\beta}}(\boldsymbol{\beta}) \int \prod_{a=1}^{s} \mathrm{d}\boldsymbol{\theta}^a \varphi(\boldsymbol{\theta}^a) \left[ e^{\hat{\rho}\langle \boldsymbol{\beta}, \boldsymbol{\Psi}\boldsymbol{\beta}\rangle + \sum_{a=1}^{s} \hat{m}^a \langle \boldsymbol{\theta}^a, \boldsymbol{\Phi}\boldsymbol{\beta}\rangle + \sum_{a,b=1}^{s} \hat{Q}^{ab}\langle \boldsymbol{\theta}^a, \boldsymbol{\Omega}\boldsymbol{\theta}^b\rangle} \right] \tag{4.22}$$

$$\Psi_y^{(s)}(\rho, \boldsymbol{m}, \boldsymbol{Q}) = \log \int_{\mathbb{R}} \mathrm{d}y \int \mathrm{d}\nu P_\star(y|\nu) \int \prod_{a=1}^{s} \mathrm{d}\lambda^a \psi(y|\lambda^a) \mathcal{N}\left(\nu, \boldsymbol{\lambda}|\mathbf{0}_{s+1}, \begin{bmatrix} \rho & \boldsymbol{m} \\ \boldsymbol{m}^\top & \boldsymbol{Q} \end{bmatrix}\right) \tag{4.23}$$

⚠ Note that going from eq. (4.18) to eq. (4.20), we made a change of variables $\hat{\rho} \leftarrow i\hat{\rho}$, $\hat{m}^a \leftarrow i\hat{m}^a$, $\hat{Q}^{ab} \leftarrow i\hat{Q}^{ab}$. Therefore, technically speaking the integrals over the conjugate variables $(\hat{\rho}, \hat{m}^a, \hat{Q}^{ab})$ are over the imaginary axis $i\mathbb{R}$. This will not make a difference in what follows, and could be properly treated at the expense of a longer analysis.

### 4.1.4 Step 4: Saddle-point method

Although it might seem at this point that with the rewriting in eq. (4.19) we have only made the problem more complicated by introducing integrals over the variables $(\rho, \hat{\rho}, \boldsymbol{m}, \hat{\boldsymbol{m}}, \boldsymbol{Q}, \hat{\boldsymbol{Q}})$, as we will see now the strength of the method is that we will never have to compute these integrals. Indeed, under the assumption that eq. (3.2), all terms involved in eq. (4.20) are order one numbers and the dimensions $n, p, d$ only appear in eq. (4.20) through the ratios $\alpha = {}^{n}/_{p}$ and $\gamma = {}^{p}/_{d}$.

Therefore, in the scaling where the limit $p \to \infty$ in eq. (4.1) is taken with $\alpha, \gamma$ kept fixed (a.k.a. proportional scaling), thanks to the Saddle-point method the integrals in eq. (4.19) will be dominated by the configurations $(\rho, \hat{\rho}, \boldsymbol{m}, \hat{\boldsymbol{m}}, \boldsymbol{Q}, \hat{\boldsymbol{Q}})$ that extremise the potential function $\Phi^{(s)}$. In other words, we have:

$$\mathbb{E}_{\mathcal{D}}[Z_d^s] \underset{p\to\infty}{\sim} e^{p\Phi^{(s)}(\rho_\star,\hat{\rho}_\star,\boldsymbol{m}_\star,\hat{\boldsymbol{m}}_\star,\boldsymbol{Q}_\star,\hat{\boldsymbol{Q}}_\star)} \tag{4.24}$$

where $(\rho_\star, \hat{\rho}_\star, \boldsymbol{m}_\star, \hat{\boldsymbol{m}}_\star, \boldsymbol{Q}_\star, \hat{\boldsymbol{Q}}_\star)$ are the solutions of:

$$\underset{\rho,\hat{\rho},\boldsymbol{m},\hat{\boldsymbol{m}},\boldsymbol{Q},\hat{\boldsymbol{Q}}}{\textbf{extr}} \left\{ -\frac{1}{\gamma}\rho\hat{\rho} - \frac{1}{\sqrt{\gamma}} \sum_{a=1}^{s} m^a \hat{m}^a - \sum_{a,b=1}^{s} Q^{ab}\hat{Q}^{ab} + \alpha \Psi_y^s(\rho, \boldsymbol{m}, \boldsymbol{Q}) + \Psi_\theta(\hat{\rho}, \hat{\boldsymbol{m}}, \hat{\boldsymbol{Q}}) \right\} \tag{4.25}$$

In other words, the potential function $\Phi^{(s)}$ is exactly the rate function we are after! Note that the overlaps $(\rho, \hat{\rho}, \boldsymbol{m}, \hat{\boldsymbol{m}}, \boldsymbol{Q}, \hat{\boldsymbol{Q}})$ play an analogous role to the magnetisation in the Curie-Weiss model. Therefore, in principle this give us a clear path on how to compute the free entropy: we first solve the extremisation problem in eq. (4.25) and then take the limit $s \to 0^+$. However, the problem is that it is not clear how to solve eq. (4.25) in general. This seems to leave us in an *impasse*: did we just rewrote an initially hard problem in terms of another one?

⚠ When applying the saddle-point method with $p \to \infty$, we need to assume the limit of the entropic potential eq. (4.48) exists. This will be the cases in all particular examples we will look later, and can be justified ad hoc.

### 4.1.5 Step 5: Replica symmetry

Since the extremisation problem in eq. (4.25) cannot be solved explicitly, taking the $s \to 0^+$ limit requires making assumptions on the shape of the solution. At this point, we follow a standard idea in physics, which consists of searching for particular classes of solution to the problem based on our intuition. The overlaps $(\rho, \hat{\rho}, \boldsymbol{m}, \hat{\boldsymbol{m}}, \boldsymbol{Q}, \hat{\boldsymbol{Q}})$ quantify how independent samples from the posterior distribution are correlated, and therefore their shape encode the geometry of the high-dimensional measure. Therefore, in the physics jargon we need an *ansatz* that translates how we expect the measure to look like.

The simplest such *ansatz* is *replica symmetry*:[5]

$$m^a = m \qquad \hat{m}^a = \hat{m}$$
$$Q^{ab} = \begin{cases} r & \text{for } a = b \\ q & \text{for } a \neq b \end{cases}, \qquad \hat{Q}^{ab} = \begin{cases} -\frac{1}{2}\hat{r} & \text{for } a = b \\ \frac{1}{2}\hat{q} & \text{for } a \neq b \end{cases}, \tag{4.26}$$

In words, replica symmetry states that in the limit $d \to \infty$ the overlap between two independent samples of the posterior distribution $\boldsymbol{\theta}^1, \boldsymbol{\theta}^2 \sim P(\boldsymbol{\theta}|\mathcal{D})$ will concentrate at a single value $q$.

⚠ Replica symmetry will not hold in general. In particular, for some problems it can hold in a range of parameters, but not for others. There is a well-defined recipe in statistical physics to check when replica symmetry holds and to deal with cases which are not replica symmetric, known as the *replica symmetry breaking scheme* (Mézard et al., 1987).

There are two particular sub-classes of the model above for which replica symmetry can be proved to hold:

(a) In the Bayes-optimal scenario discussed in example 1. This is a consequence of the Nishimori identities eq. (3.7), see (Barbier et al., 2019; Barbier and Panchenko, 2022) for a formal proof.

(b) In convex optimisation problems, such as empirical risk minimisation example 2 with convex empirical risk and for and maximum likelihood estimation example 3 with log-concave likelihood, replica symmetry holds due to the convexity of the problem (Loureiro et al., 2021).

From now on we now assume we are in one of these scenarios.

[5] Note the factor $-1/2$ in $\hat{r}$ is just for convenience.

**Trace terms** — We start by looking at the simplest terms, the sums in eq. (4.25). Under the replica symmetric assumption, we have:

$$\sum_{a=1}^{s} \hat{m}^a m^a = s\hat{m}h$$
$$\sum_{a,b=1}^{s} \hat{Q}^{ab} Q^{ab} = \sum_{a=1}^{s} \hat{Q}^{aa} Q^{aa} + \sum_{a\neq b}^{s} \hat{Q}^{ab} Q^{ab} = -\frac{s}{2}\hat{r}r + \frac{s(s-1)}{2}\hat{q}q \tag{4.27}$$

Therefore, these are easy terms to take the limit.

**Entropic potential** — Consider now the entropic potential eq. (4.48). First, note we can write:

$$\sum_{a,b=1}^{s} \hat{Q}^{ab} \langle \boldsymbol{\theta}^a, \boldsymbol{\Omega}\boldsymbol{\theta}^b \rangle = -\frac{\hat{r}}{2} \sum_{a=1}^{s} \langle \boldsymbol{\theta}^a, \boldsymbol{\Omega}\boldsymbol{\theta}^a \rangle + \frac{\hat{q}}{2} \sum_{a\neq b}^{s} \langle \boldsymbol{\theta}^a, \boldsymbol{\Omega}\boldsymbol{\theta}^b \rangle \tag{4.28}$$
$$= -\frac{\hat{r}+\hat{q}}{2} \sum_{a=1}^{s} \langle \boldsymbol{\theta}^a, \boldsymbol{\Omega}\boldsymbol{\theta}^a \rangle + \frac{\hat{q}}{2} \sum_{a,b}^{s} \langle \boldsymbol{\theta}^a, \boldsymbol{\Omega}\boldsymbol{\theta}^b \rangle \tag{4.29}$$

where in the second equality we have added and subtracted the diagonal. Therefore:

$$\Psi_\theta = \frac{1}{p} \log \int \mathrm{d}\boldsymbol{\beta} P_{\boldsymbol{\beta}}(\boldsymbol{\beta})\ e^{\hat{\rho}\langle \boldsymbol{\beta}, \boldsymbol{\Psi}\boldsymbol{\beta}\rangle} \int \prod_{a=1}^{s} \mathrm{d}\boldsymbol{\theta}^a \varphi(\boldsymbol{\theta}^a) e^{\hat{m} \sum_{a=1}^{s} \langle \boldsymbol{\theta}^a, \boldsymbol{\Phi}\boldsymbol{\beta}\rangle - \frac{\hat{r}+\hat{q}}{2} \sum_{a=1}^{s} \langle \boldsymbol{\theta}^a, \boldsymbol{\Omega}\boldsymbol{\theta}^a\rangle + \frac{\hat{q}}{2} \sum_{a,b=1}^{s} \langle \boldsymbol{\theta}^a, \boldsymbol{\Omega}\boldsymbol{\theta}^b\rangle} \tag{4.30}$$
$$= \frac{1}{p} \log \int \mathrm{d}\boldsymbol{\beta} P_{\boldsymbol{\beta}}(\boldsymbol{\beta}) e^{\hat{\rho}\langle \boldsymbol{\beta}, \boldsymbol{\Psi}\boldsymbol{\beta}\rangle} \int \left[ \prod_{a=1}^{s} \mathrm{d}\boldsymbol{\theta}^a \varphi(\boldsymbol{\theta}^a) e^{\hat{m}^s \langle \boldsymbol{\theta}^a, \boldsymbol{\Phi}\boldsymbol{\beta}\rangle - \frac{\hat{r}+\hat{q}}{2}\langle \boldsymbol{\theta}^a, \boldsymbol{\Omega}\boldsymbol{\theta}^a\rangle} \right] e^{\frac{\hat{q}}{2} \sum_{a,b=1}^{s} \langle \boldsymbol{\theta}^a, \boldsymbol{\Omega}\boldsymbol{\theta}^b\rangle} \tag{4.31}$$

Note that all terms factorise over the replica indices, except the last. To factorise it, we use a common trick in physics known as the *Hubbard-Stratonovich transformation*:

$$e^{\frac{\hat{q}}{2} \sum_{a,b=1}^{s} \langle \boldsymbol{\theta}^a, \boldsymbol{\Omega}\boldsymbol{\theta}^b\rangle} = \mathbb{E}_{\boldsymbol{\xi}\sim\mathcal{N}(\mathbf{0}_p, \boldsymbol{I}_p)} \left[ e^{\sqrt{\hat{q}} \sum_{a=1}^{s} \langle \boldsymbol{\xi}, \boldsymbol{\Omega}^{1/2}, \boldsymbol{\theta}^a\rangle} \right] \tag{4.32}$$

where $\boldsymbol{\Omega}^{1/2}$ is the matrix square-root of $\boldsymbol{\Omega}$.[6] Putting together, we have:

$$\Psi_\theta^{(s)} = \frac{1}{p} \log \int \mathrm{d}\boldsymbol{\beta} P_{\boldsymbol{\beta}}(\boldsymbol{\beta}) e^{\hat{\rho}\langle \boldsymbol{\beta}, \boldsymbol{\Psi}\boldsymbol{\beta}\rangle} \mathbb{E}_{\boldsymbol{\xi}} \left( \int \mathrm{d}\boldsymbol{\theta} \varphi(\boldsymbol{\theta}) e^{-\frac{\hat{r}+\hat{q}}{2}\langle \boldsymbol{\theta}, \boldsymbol{\Omega}\boldsymbol{\theta}\rangle + \hat{m}\langle \boldsymbol{\theta}, \Phi\boldsymbol{\beta}\rangle + \sqrt{q}\langle \boldsymbol{\xi}, \boldsymbol{\Omega}^{1/2}\boldsymbol{\theta}\rangle} \right)^s \tag{4.33}$$

**Energetic potential** — For the energetic potential in eq. (4.49), we need to decouple a Gaussian distribution with the following block covariance matrix:

$$\boldsymbol{\Sigma} = \begin{bmatrix} \rho & \boldsymbol{m} \\ \boldsymbol{m}^\top & \boldsymbol{Q} \end{bmatrix} = \begin{bmatrix} \rho & m\mathbf{1}_s \\ m\mathbf{1}_s^\top & (r-q)\boldsymbol{I}_s + q\mathbf{1}_s\mathbf{1}_s^\top \end{bmatrix} \tag{4.34}$$

**Exercise 2.** (a) Using the block inversion formula eq. (B.4), show that we have:

$$\boldsymbol{\Sigma}^{-1} = \begin{bmatrix} \tilde{\rho} & \tilde{m}\mathbf{1}_s \\ \tilde{m}\mathbf{1}_s^\top & (\tilde{r}-\tilde{q})\boldsymbol{I}_s + \tilde{q}\mathbf{1}_s\mathbf{1}_s^\top \end{bmatrix} \tag{4.35}$$

with:

$$\tilde{\rho} = \left( \rho - \frac{sm^2}{r+(s-1)q} \right)^{-1}, \qquad \tilde{r} = \frac{1}{r-q} \left( 1 + \frac{m^2-\rho q}{\rho(r+(s-1)q) - sm^2} \right)$$
$$\tilde{m} = -\frac{m}{\rho(r+(s-1)q) - sm^2}, \qquad \tilde{q} = \frac{1}{r-q} \frac{m^2-\rho q}{\rho(r+(s-1)q) - sm^2} \tag{4.36}$$

[6] Note this is well defined since $\boldsymbol{\Omega}$ is positive semi-definite.

(b) Show that:

$$\det \boldsymbol{\Sigma} = (r-q)^{s-1}(\rho(r+(s-1)q) - sm^2) \tag{4.37}$$

(c) By using a Hubbard-Stratonovich transformation, show that you can write:

$$\Psi_y^{(s)}(\rho, \boldsymbol{m}, \boldsymbol{Q}) = \log \int \mathrm{d}y \int \mathrm{d}\nu P_\star(y|\nu) e^{-\frac{\tilde{\rho}}{2}\nu^2} \left( \int \mathrm{d}\lambda \psi(y|\lambda) e^{-\frac{\tilde{r}-\tilde{q}}{2}\lambda^2 + \left(\sqrt{-\tilde{q}}\eta + \tilde{m}\nu\right)\lambda} \right)^s - \frac{1}{2}\log\det(2\pi\boldsymbol{\Sigma}) \tag{4.38}$$

### 4.1.6 Step 6: Taking the limits

Now that all the dependency in $s$ is explicit, we can proceed in taking the $s \to 0^+$ limit. First, note that at zeroth order we have:

$$\Phi^{(s)} = -\frac{1}{\gamma}\rho\hat{\rho} + \frac{1}{p}\log \int \mathrm{d}\boldsymbol{\beta} P_{\boldsymbol{\beta}}(\boldsymbol{\beta}) e^{\hat{\rho}\langle \boldsymbol{\beta}, \boldsymbol{\Psi}\boldsymbol{\beta}\rangle} + O(s) \tag{4.39}$$

But $Z_d^0 = 1$, and therefore by consistency we must have $\lim_{s\to 0^+} \Phi^{(s)} = 0$. This implies that $\hat{\rho} = 0$, which due to the extremisation in eq. (4.25) fixes the constraint:

$$\rho = \lim_{d\to\infty} \frac{1}{d}\mathbb{E}_{\boldsymbol{\beta}}[\langle \boldsymbol{\beta}, \boldsymbol{\Psi}\boldsymbol{\beta}\rangle]. \tag{4.40}$$

This is nothing but our original definition of $\rho$. Therefore, consistency implies $\rho$ is not a fluctuating variable but simply a constant. We can now move to the first order terms. First, taking the limit on the entropy term:

$$\Phi_\theta \coloneqq \lim_{s\to 0^+} \partial_s \Psi_\theta^{(s)} = \frac{1}{p}\mathbb{E}_{\boldsymbol{\xi},\boldsymbol{\beta}} \log \int \mathrm{d}\boldsymbol{\theta} \varphi(\boldsymbol{\theta}) e^{-\frac{\hat{r}+\hat{q}}{2}\langle \boldsymbol{\theta}, \boldsymbol{\Omega}\boldsymbol{\theta}\rangle + \hat{m}\langle \boldsymbol{\theta}, \Phi\boldsymbol{\beta}\rangle + \sqrt{q}\langle \boldsymbol{\xi}, \boldsymbol{\Omega}^{1/2}\boldsymbol{\theta}\rangle} \tag{4.41}$$

**Exercise 3.** (a) show that:

$$\Phi_y \coloneqq \lim_{s\to 0^+} \partial_s \Psi_y^{(s)} = \mathbb{E}_\eta \int \mathrm{d}y \int \frac{\mathrm{d}\nu}{\sqrt{2\pi}} e^{-\frac{1}{2}\nu^2} P_\star(y|\rho\nu) \log \int \frac{\mathrm{d}\lambda}{2\pi} e^{-\frac{1}{2}\frac{\lambda^2}{r-q} + \left(m\nu + \sqrt{q - \frac{m^2}{\rho}}\eta\right)\frac{\lambda}{r-q}} \psi(y|\lambda) - \frac{1}{2}\frac{q}{r-q} \tag{4.42}$$

⚠ Be careful since the exponential inside the $(\cdot)^s$ is also a function of $s$.

(b) By making a change of variables, show we can also write the above in the following symmetric form:

$$\Phi_y(m,q,r) = \lim_{s\to 0^+} \partial_s \Psi_y^{(s)} = \mathbb{E}_\eta \int \mathrm{d}y Z_\star\left(y, \frac{m}{\sqrt{q}}\eta, \rho - \frac{m^2}{q}\right) \log Z_y(y, \sqrt{q}\eta, r-q) \tag{4.43}$$

with $\eta \sim \mathcal{N}(0,1)$ and:

$$Z_y(y,\omega,v) = \mathbb{E}_{z\sim\mathcal{N}(\omega,v)}[\psi(y|z)], \qquad Z_\star(y,\omega,v) = \mathbb{E}_{z\sim\mathcal{N}(\omega,v)}[P_\star(y|z)] \tag{4.44}$$

(c) Similarly, show that with a change of variables we can also write the entropic potential in the following symmetric form:

$$\Phi_\theta(\hat{m},\hat{q},\hat{r}) = \frac{1}{p}\mathbb{E}_{\boldsymbol{\xi}}\left[ Z_{\boldsymbol{\beta}}\left(\hat{m}(\hat{q}\boldsymbol{\Omega})^{-1/2}\boldsymbol{\xi}, \hat{m}^2(\hat{q}\boldsymbol{\Omega})^{-1}\right) \log Z_{\boldsymbol{\theta}}\left((\hat{q}\boldsymbol{\Omega})^{1/2}\boldsymbol{\xi}, (\hat{r}+\hat{q})\boldsymbol{\Omega}\right)\right] \tag{4.45}$$

with $\boldsymbol{\xi} \sim \mathcal{N}(\mathbf{0}, \boldsymbol{I}_p)$ and:

$$\mathcal{Z}_{\boldsymbol{\beta}}(\boldsymbol{b},\boldsymbol{A}) = \int \mathrm{d}\boldsymbol{\beta} P_{\boldsymbol{\beta}}(\boldsymbol{\beta}) e^{-\frac{1}{2}\langle \boldsymbol{\beta}, \boldsymbol{A}\boldsymbol{\beta}\rangle + \langle \boldsymbol{b}, \boldsymbol{\beta}\rangle}, \qquad \mathcal{Z}_{\boldsymbol{\theta}}(\boldsymbol{b},\boldsymbol{A}) = \int \mathrm{d}\boldsymbol{\theta} \varphi(\boldsymbol{\theta}) e^{-\frac{1}{2}\langle \boldsymbol{\theta}, \boldsymbol{A}\boldsymbol{\theta}\rangle + \langle \boldsymbol{b}, \boldsymbol{\theta}\rangle} \tag{4.46}$$

### 4.2 Replica symmetric free entropy

Putting together the different pieces from the previous sessions, we can conclude that in the high-dimensional limit $p \to \infty$, the replica symmetric free entropy density can be written as:

$$\phi = \lim_{p\to\infty} \frac{1}{p}\mathbb{E}\log Z_d = \underset{\substack{m,q,r\\ \hat{m},\hat{q},\hat{r}}}{\mathbf{extr}}\ \Phi := \left\{ -\frac{1}{\sqrt{\gamma}} m\hat{m} + \frac{1}{2}q\hat{q} + \frac{1}{2}r\hat{r} + \alpha\Phi_y(m,q,r) + \Phi_\theta(\hat{m},\hat{q},\hat{r}) \right\} \tag{4.47}$$

with:

$$\Phi_\theta(\hat{m},\hat{q},\hat{r}) = \lim_{p\to\infty}\frac{1}{p}\mathbb{E}_{\boldsymbol{\xi},\boldsymbol{\beta}}\log\int \mathrm{d}\boldsymbol{\theta}\varphi(\boldsymbol{\theta}) e^{-\frac{\hat{r}+\hat{q}}{2}\langle\boldsymbol{\theta},\boldsymbol{\Omega}\boldsymbol{\theta}\rangle + \hat{m}\langle\boldsymbol{\theta},\Phi\boldsymbol{\beta}\rangle + \sqrt{\hat{q}}\langle\boldsymbol{\xi},\boldsymbol{\Omega}^{1/2}\boldsymbol{\theta}\rangle} \tag{4.48}$$

$$\Phi_y(m,q,r) = \mathbb{E}_\eta \int \mathrm{d}y Z_\star\left(y, \frac{m}{\sqrt{q}}\eta, \rho - \frac{m^2}{q}\right)\log Z_y(y,\sqrt{q}\eta, r-q) \tag{4.49}$$

where $\eta \sim \mathcal{N}(0,1)$, $\boldsymbol{\xi} \sim \mathcal{N}(\mathbf{0}, \boldsymbol{I}_p)$ and $\boldsymbol{\beta} \sim P_{\boldsymbol{\beta}}$, all independent and the effective partition functions $Z_{y/\star}$ are defined in eq. (4.44).

*Remark* 3. The variables $r, \hat{r}$ only appear in eq. (4.47) through the combinations $r - q$ and $\hat{r} + \hat{q}$. Therefore, it is common to define $v = r - q$ and $\hat{v} = \hat{r} + \hat{q}$ and rewrite:

$$\phi = \underset{\substack{m,q,v\\ \hat{m},\hat{q},\hat{v}}}{\mathbf{extr}}\left\{ -\frac{1}{\sqrt{\gamma}} m\hat{m} + \frac{1}{2}v\hat{v} + \frac{1}{2}(q\hat{v} - v\hat{q}) + \alpha\Phi_y(m,q,v) + \Phi_\theta(\hat{m},\hat{q},\hat{v}) \right\} \tag{4.50}$$

The variable $v$ has also a more natural interpretation as the normalised variance of the posterior distribution. Indeed, letting $\boldsymbol{\theta} \sim P(\boldsymbol{\theta}|\mathcal{D})$, and denoting the average with respect to the posterior $P(\boldsymbol{\theta}|\mathcal{D})$ with brackets $\langle\cdot\rangle$:

$$\lim_{p\to\infty}\frac{1}{p}\mathrm{Var}(\boldsymbol{\theta}) = \lim_{p\to\infty}\frac{1}{p}\left[\langle||\boldsymbol{\theta}||_2^2\rangle - ||\langle\boldsymbol{\theta}\rangle||_2^2\right] = r - q := v \tag{4.51}$$

Let's stop to contemplate what we have achieved. We have reduced the computation of the logarithm of a high-dimensional integral to a low-dimensional extremisation problem over 6 variables $(m, q, r, \hat{m}, \hat{q}, \hat{r})$, under (almost) arbitrary prior $P_{\boldsymbol{\beta}}, \varphi$ and likelihood $P_\star, \psi$ functions, covering problems that go from empirical risk minimisation to Bayesian inference. Deriving a formula with such a scope is a quite remarkable achievement. Indeed, none of the rigorous constructive methods available in the literature - such as interpolation or CGMT - cover all these different settings under the same method.

*Remark* 4. The replica method involve a series of manipulations which might leave a mathematician uncomfortable. Most of them can be made properly justified with some additional work, except for two of them:

- The exchange of the $p \to \infty$ and the $s \to 0^+$ limit when we applied the saddle-point method in eq. (4.24)
- The fact that we computed the moments $\mathbb{E}_{\mathcal{D}}[Z_d^s]$ only for integers values of $s \in \mathbb{N}$, and then (carelessly) analytically continued to $s \in \mathbb{R}_+$ to take the $s \to 0^+$ limit.

These are the main reasons why it remains a heuristic tool and not a rigorous method.

*Remark* 5. The saddle-point method only give us an extremiser, and does not specify whether they are minima or maxima of the potential function. In particular problems where replica symmetry holds and a rigorous proof for this formula are available, such as Bayes-optimal estimation with isotropic covariates (Barbier et al., 2019) or convex empirical risk minimisation with convex risk (Loureiro et al., 2021), it can be shown that the extremisation problem take the form of sup inf problem:

$$\phi = \sup_{q,m,r\in\mathbb{R}_+}\ \inf_{\hat{m},\hat{q},\hat{v}\in\mathbb{R}_+} \Phi(m,q,v,\hat{m},\hat{q},\hat{v}) \tag{4.52}$$

### 4.3 Self-consistent equations

By definition, extremisers $(m_\star, q_\star, r_\star, \hat{m}_\star, \hat{q}_\star, \hat{r}_\star) \in \mathbb{R}^6_+$ of the potential in equation eq. (4.50) are zero-gradient points:

$$\nabla_{(m,q,v,\hat{m},\hat{q},\hat{v})}\Phi \stackrel{!}{=} 0 \tag{4.53}$$

This leads to the following equations:

$$\begin{cases} v = 2\partial_{\hat{q}}\Phi_{\boldsymbol{\theta}}(\hat{m},\hat{q},\hat{v}) \\ q = -2\left(\partial_{\hat{q}}\Phi_{\boldsymbol{\theta}}(\hat{m},\hat{q},\hat{v}) + \partial_{\hat{v}}\Phi_{\boldsymbol{\theta}}(\hat{m},\hat{q},\hat{v})\right) \\ m = \sqrt{\gamma}\partial_{\hat{m}}\Phi_{\boldsymbol{\theta}}(\hat{m},\hat{q},\hat{v}) \end{cases}, \quad \begin{cases} \hat{v} = -2\alpha\partial_q\Phi_y(m,q,v) \\ \hat{q} = 2\alpha\left(\partial_v\Phi_y(m,q,v) - \partial_q\Phi_y(m,q,v)\right) \\ \hat{m} = \sqrt{\gamma}\partial_m\Phi_y(m,q,v) \end{cases} \tag{4.54}$$

Although more cumbersome, the self-consistent equations above are of similar spirit to the self-consistent equation $m = \tanh(\beta(m+h))$ we found for the magnetisation in the Curie-Weiss model in Section 2.

**Exercise 4.** 1. From the symmetric form in eq. (4.43), show that the self-consistent equations for the conjugate variables can be explicitly written as:

$$\hat{v} = -\alpha\mathbb{E}_\eta \int \mathrm{d}y Z_\star\left(y, \frac{m}{\sqrt{q}}\eta, \rho - \frac{m^2}{q}\right)\partial_\omega f_y(y, \sqrt{q}\eta, v) \tag{4.55}$$

$$\hat{q} = \alpha\mathbb{E}_\eta \int \mathrm{d}y Z_\star\left(y, \frac{m}{\sqrt{q}}\eta, \rho - \frac{m^2}{q}\right) f_y(y, \sqrt{q}\eta, v)^2 \tag{4.56}$$

$$\hat{m} = \alpha\mathbb{E}_\eta \int \mathrm{d}y Z_\star\left(y, \frac{m}{\sqrt{q}}\eta, \rho - \frac{m^2}{q}\right) f_y(y, \sqrt{q}\eta, v) \tag{4.57}$$

where:

$$f_y(y,\omega,v) = \partial_\omega \log Z_y(y,\omega,v) \tag{4.58}$$

with $Z_{\star/y}(y,\omega,v)$ defined in eq. (4.44).

2. Similarly, show that the self-consistent equations for the overlaps can be written as:

$$v = \mathbb{E}_{\boldsymbol{\xi},\boldsymbol{\beta}}\left[\boldsymbol{\nabla}_{\boldsymbol{b}} \cdot \boldsymbol{f}_{\boldsymbol{\theta}}\left(\hat{m}\boldsymbol{\beta} + (\hat{q}\boldsymbol{\Omega})^{1/2}\boldsymbol{\xi}, \hat{v}\boldsymbol{\Omega}\right)\right] \tag{4.59}$$

$$q = \mathbb{E}_{\boldsymbol{\xi},\boldsymbol{\beta}}\left[||\boldsymbol{f}_{\boldsymbol{\theta}}\left(\hat{m}\boldsymbol{\beta} + (\hat{q}\boldsymbol{\Omega})^{1/2}\boldsymbol{\xi}, \hat{v}\boldsymbol{\Omega}\right)||_2^2\right] \tag{4.60}$$

$$m = \mathbb{E}_{\boldsymbol{\xi},\boldsymbol{\beta}}\left[\langle \boldsymbol{f}_{\boldsymbol{\theta}}\left(\hat{m}\boldsymbol{\beta} + (\hat{q}\boldsymbol{\Omega})^{1/2}\boldsymbol{\xi}, \hat{v}\boldsymbol{\Omega}\right), \boldsymbol{\beta}\rangle\right] \tag{4.61}$$

where:

$$\boldsymbol{f}_{\boldsymbol{\theta}}(\boldsymbol{b},\boldsymbol{A}) = \nabla_{\boldsymbol{b}} \log Z_{\boldsymbol{\theta}}(\boldsymbol{b},\boldsymbol{A}) \tag{4.62}$$

with $Z_{\boldsymbol{\theta}}(\boldsymbol{b},\boldsymbol{A})$ defined in eq. (4.46).

**Solving self-consistent equations numerically** — Except for very particular cases the self-consistent equations eq. (4.54) do not admit an explicit closed-form solution, and most of the times we look for solutions numerically. The most common approach to solve these equations numerically consists of iteratively applying them $x^{t+1} = f(x^t)$ from an initial condition $x^0$.

⚠ When the map $f(x)$ is contractive $f(x^t) < f(x^{t+1})$, we are guaranteed to find a unique solution to the equation $x^{t+1} = f(x^t)$, independently from $x^0$. Note this is intimately connected to

the convexity of the free entropy potential $\Phi$ in the extremisation problem in eq. (4.47). However, in general the potential will not be a convex function of the overlaps parameters, and therefore the solution is not guaranteed to be unique. In fact, since the equations eq. (4.47) are derived from gradient descent on $\Phi$, in case $\Phi$ is not convex the iterative solution will converge to the closest extremal point. As we discussed in the context of the Curie-Weiss model in Section 2, the existence of different extremal points of $\Phi$ is closely related to the existence of phase transitions in the problem.

### 4.4 Case study 1: Bayes-optimal inference

We now turn our attention to particular cases where the free entropy potential and the self-consistent equations can be further simplified. The first example if Bayes-optimal inference, which we introduced in Example 1. Recall that in this case we have $p = d$, $\boldsymbol{v} = \boldsymbol{u} \in \mathbb{R}^d$ and matched likelihood and priors:

$$\varphi = P_{\boldsymbol{\beta}}, \qquad \psi = P_\star. \tag{4.63}$$

In this case, applying the Nishimori identity eq. (3.7) to the overlaps imply:

$$q = m, \qquad \hat{q} = \hat{m}, \qquad v = \rho - m, \qquad \hat{v} = \hat{q}. \tag{4.64}$$

such that the only two independent variables $m, \hat{m}$. The free entropy extremisation problem simplifies to:

$$\phi = \underset{m,\hat{m}}{\mathbf{extr}} \left\{ -m\hat{m} + \alpha \Phi_y(m) + \Phi_{\boldsymbol{\theta}}(m) \right\} \tag{4.65}$$

with:

$$\Phi_y(m) = \mathbb{E}_\eta \int \mathrm{d}y \; Z_\star(y, \sqrt{q}\eta, \rho - m) \log Z_\star(y, \sqrt{q}\eta, \rho - m) \tag{4.66}$$

$$\Phi_{\boldsymbol{\theta}}(\hat{m}) = \mathbb{E}_{\eta,\boldsymbol{\beta}} \log Z_{\boldsymbol{\beta}} \left( \hat{m}\boldsymbol{\Omega}\boldsymbol{\beta} + (\hat{q}\boldsymbol{\Omega})^{1/2}\boldsymbol{\xi}, \hat{m}\boldsymbol{\Omega} \right) \tag{4.67}$$

where $\eta \sim \mathcal{N}(0,1)$, $\boldsymbol{\xi} \sim \mathcal{N}(\mathbf{0}, \boldsymbol{I}_d)$ and:

$$Z_\star(y, \omega, v) = \mathbb{E}_{z\sim\mathcal{N}(\omega,v)}[P_\star(y|z)], \qquad Z_{\boldsymbol{\beta}}(\boldsymbol{b}, \boldsymbol{A}) = \int \mathrm{d}\boldsymbol{\beta} P_{\boldsymbol{\beta}}(\boldsymbol{\beta}) e^{-\frac{1}{2}\langle \boldsymbol{\beta}, \boldsymbol{A}\boldsymbol{\beta}\rangle + \langle \boldsymbol{b}, \boldsymbol{\beta}\rangle} \tag{4.68}$$

Therefore, the self-consistent equations eq. (4.54) also reduce to:

$$\hat{m} = \alpha \mathbb{E}_\eta \int \mathrm{d}y \; Z_\star \left(y, \sqrt{m}\eta, \rho - m\right) f_\star(y, \sqrt{m}\eta, \rho - m) \tag{4.69}$$

$$m = \mathbb{E}_{\boldsymbol{\xi},\boldsymbol{\beta}} \left[ \langle \boldsymbol{f}_{\boldsymbol{\beta}} \left( \hat{m}\boldsymbol{\beta} + (\hat{m}\boldsymbol{\Omega})^{1/2}\boldsymbol{\xi}, \hat{m}\boldsymbol{\Omega} \right), \boldsymbol{\beta} \rangle \right] \tag{4.70}$$

The free entropy and self-consistent equations for Bayes-optimal estimation were first derived in the isotropic case $\boldsymbol{\Omega} = \boldsymbol{I}_p$ for the linear likelihood $P_\star(y|z) = \mathcal{N}(y|z, \sigma^2)$ in (Krzakala et al., 2012a,b), and later generalised and rigorously proven by (Barbier et al., 2019). This was extended in (Clarté et al., 2023) to the non-isotropic case.

### 4.5 Case study 2: Empirical risk minimisation

Consider now the empirical risk minimisation case with convex loss function $\ell$ and penalty $r$, introduced in Example 2. In this case, we have:

$$\varphi(\boldsymbol{\theta}) = e^{\beta r(\boldsymbol{\theta})}, \qquad \psi(y|z) = e^{\beta \ell(y,z)}. \tag{4.71}$$

and our interest is in the case where $\beta \to \infty$, where the posterior shrinks to the (unique) empirical risk minimiser. Focusing on the energetic terms, note that we have:

$$Z_y(y, \omega, v) = \mathbb{E}_{z\sim\mathcal{N}(\omega,v)}[P_y(y|z)] = \int \frac{\mathrm{d}z}{\sqrt{2\pi v}} e^{-\frac{(z-\omega)^2}{2v} - \beta\ell(y,z)} \tag{4.72}$$

To take the limit $\beta \to \infty$, we need to know how the parameters $(m, q, v, \hat{m}, \hat{q}, \hat{v}) \in \mathbb{R}_+^6$ scale with $\beta$. We posit the following scaling:

$$v = \frac{v_0}{\beta}, \qquad q = q_0, \qquad m = m_0 \tag{4.73}$$

$$\hat{v} = \beta\hat{v}_0, \qquad \hat{q} = \beta^2\hat{q}_0, \qquad \hat{m} = \beta\hat{m}_0 \tag{4.74}$$

where $(m_0, q_0, v_0, \hat{m}_0, \hat{q}_0, \hat{v}_0)$ are do not scale with $\beta$. Note that the scaling of the overlaps are intuitive: since $v$ is the asymptotic value of the variance of the posterior (see Remark 3) and the posterior shrinks as $\beta \to \infty$, it is reasonable that it vanishes with $\beta$. On the same note, we expect $q, m$ to not scale with $\beta$, since they parametrise the norm and correlation of a typical sample of the posterior with the signal. The scalings for the conjugate variables can then be found by inspection of the self-consistent equations eq. (4.55). With these in hand, we note that:

$$Z_y(y, \sqrt{q}\eta, v) = \int \frac{\mathrm{d}z}{\sqrt{2\pi v_0/\beta}} e^{-\beta\left(\frac{(z-\sqrt{q_0}\eta)^2}{2v_0} + \ell(y,z)\right)} \tag{4.75}$$

Applying Laplace's method, we have:

$$Z_y(y, \sqrt{q}\eta, v) \underset{\beta\to\infty}{\sim} e^{-\beta\mathcal{M}_{v_0\ell(y,\cdot)}(\sqrt{q_0}\eta)} \tag{4.76}$$

where $\mathcal{M}_{\tau f}(x)$ is the *Moreau envelope*:

$$\mathcal{M}_{\tau f}(x) = \min_{z\in\mathbb{R}} \left\{ \frac{(z-x)^2}{2\tau} + \ell(y,z) \right\} \tag{4.77}$$

Inserting this in the energetic potential eq. (4.49):

$$\Phi_y(m, q, r) \underset{\beta\to\infty}{\sim} -\beta\, \mathbb{E}_\eta \int \mathrm{d}y Z_\star \left( y, \frac{m_0}{\sqrt{q_0}}\eta, \rho - \frac{m_0^2}{q_0} \right) \mathcal{M}_{v_0\ell(y,\cdot)}\left(\sqrt{q_0}\eta\right) \tag{4.78}$$

The entropic term requires a bit more work. First, we complete the squares and write:

$$Z_{\boldsymbol{\theta}}(\boldsymbol{b}, \boldsymbol{A}) = e^{-\frac{1}{2}\langle \boldsymbol{b}, \boldsymbol{A}^{-1}\boldsymbol{b}\rangle} \int \mathrm{d}\boldsymbol{\theta}\varphi(\boldsymbol{\theta}) e^{-\frac{1}{2}||\boldsymbol{A}^{1/2}\boldsymbol{\theta} - \boldsymbol{A}^{-1/2}\boldsymbol{b}||_2^2} \tag{4.79}$$

Since in eq. (4.49) we need to evaluate:

$$\begin{aligned} \boldsymbol{b} &= \hat{m}\boldsymbol{\beta} + (\hat{q}\boldsymbol{\Omega})^{1/2}\boldsymbol{\xi} = \beta\left(\hat{m}_0\boldsymbol{\beta} + (\hat{q}_0\boldsymbol{\Omega})^{1/2}\boldsymbol{\xi}\right) \coloneqq \beta\boldsymbol{b}_0 \\ \boldsymbol{A} &= \hat{v}\boldsymbol{\Omega} = \beta(\hat{v}_0\boldsymbol{\Omega}) \coloneqq \beta\boldsymbol{A}_0 \end{aligned} \tag{4.80}$$

therefore:

$$Z_{\boldsymbol{\theta}}(\hat{m}\boldsymbol{\beta} + (\hat{q}\boldsymbol{\Omega})^{\beta/2}\boldsymbol{\xi}, \hat{q}\boldsymbol{\Omega}) = e^{-\frac{1}{2}\langle \boldsymbol{b}_0, \boldsymbol{A}_0^{-1}\boldsymbol{b}_0\rangle} \int \mathrm{d}\boldsymbol{\theta} e^{-\beta\left(r(\boldsymbol{\theta}) + \frac{1}{2}||\boldsymbol{A}_0^{1/2}\boldsymbol{\theta} - \boldsymbol{A}_0^{-1/2}\boldsymbol{b}||_2^2\right)} \tag{4.81}$$

$$\underset{\beta\to\infty}{\sim} e^{-\frac{\beta}{2}\langle \boldsymbol{b}_0, \boldsymbol{A}_0^{-1}\boldsymbol{b}_0\rangle} e^{-\beta\mathcal{M}_{r(\boldsymbol{A}_0^{-1/2}\cdot)}(\boldsymbol{A}_0^{-1/2}\boldsymbol{b}_0)} \tag{4.82}$$

where $\mathcal{M}_{\tau f}(\boldsymbol{x})$ is the vector valued Moreau envelope:

$$\mathcal{M}_{\tau f}(\boldsymbol{x}) = \min_{\boldsymbol{z}\in\mathbb{R}^p} \left\{ -\frac{1}{2\tau}||\boldsymbol{z}-\boldsymbol{x}||_2^2 + f(\boldsymbol{z}) \right\} \tag{4.83}$$

Inserting this in the entropic potential eq. (4.48):

$$\begin{aligned}
\Phi_{\boldsymbol{\theta}}(\hat{m},\hat{q},\hat{v}) \underset{\beta\to\infty}{\sim} & -\frac{\beta}{p}\mathbb{E}_{\boldsymbol{\xi},\boldsymbol{\beta}}\left[\mathcal{M}_{r((\hat{v}_0\boldsymbol{\Omega})^{-1/2}\cdot)}\left((\hat{v}_0\boldsymbol{\Omega})^{-1/2}\left((\hat{q}_0\boldsymbol{\Omega})^{1/2}\boldsymbol{\xi}+\hat{m}_0\boldsymbol{\beta}\right)\right)\right] \\
& -\frac{\beta}{2p}\mathbb{E}_{\boldsymbol{\beta},\boldsymbol{\xi}}\left[\langle(\hat{q}_0\boldsymbol{\Omega})^{1/2}\boldsymbol{\xi}+\hat{m}_0\boldsymbol{\beta},(\hat{v}_0\boldsymbol{\Omega})^{-1}((\hat{q}_0\boldsymbol{\Omega})^{1/2}\boldsymbol{\xi}+\hat{m}_0\boldsymbol{\beta})\right] \qquad (4.84) \\
= & -\frac{\beta}{p}\mathbb{E}_{\boldsymbol{\xi},\boldsymbol{\beta}}\left[\mathcal{M}_{r((\hat{v}_0\boldsymbol{\Omega})^{-1/2}\cdot)}\left((\hat{v}_0\boldsymbol{\Omega})^{-1/2}\left((\hat{q}_0\boldsymbol{\Omega})^{1/2}\boldsymbol{\xi}+\hat{m}_0\boldsymbol{\beta}\right)\right)\right] \\
& -\frac{\beta}{2}\left(\frac{\hat{q}_0}{\hat{v}_0}+\frac{\hat{m}_0^2}{\hat{v}_0}\mathbb{E}_{\boldsymbol{\beta}}\left[\frac{\langle\boldsymbol{\beta},\boldsymbol{\Omega}^{-1}\boldsymbol{\beta}\rangle}{p}\right]\right) \qquad (4.85)
\end{aligned}$$

Putting together, we see that the free entropy scales as $\beta$ as $\beta\to\infty$. For this reason, it is more convenient to look at the free energy instead:

$$f_\beta = -\frac{\phi}{\beta} \tag{4.86}$$

Which at zero temperature is given by:

$$\lim_{\beta\to\infty} f_\beta = \underset{\substack{m,q,v\\ \hat{m},\hat{q},\hat{v}}}{\mathbf{extr}} \left\{ \frac{1}{\sqrt{\gamma}}m\hat{m} - \frac{1}{2}(q\hat{v}-v\hat{q}) + \alpha\Phi_y(m,q,v) + \Phi_\theta(\hat{m},\hat{q},\hat{v}) \right\} \tag{4.87}$$

where we have dropped the $_0$ subscript to alleviate the notation and have:

$$\begin{aligned}
\Phi_y(m,q,v) &= \mathbb{E}_\eta \int \mathrm{d}y Z_\star\left(y,\frac{m}{\sqrt{q}}\eta,\rho-\frac{m^2}{q}\right)\mathcal{M}_{v_0\ell(y,\cdot)}\left(\sqrt{q}\eta\right) \\
\Phi_{\boldsymbol{\theta}}(\hat{m},\hat{q},\hat{v}) &= \mathbb{E}_{\boldsymbol{\xi},\boldsymbol{\beta}}\left[\mathcal{M}_{r((\hat{v}_0\boldsymbol{\Omega})^{-1/2}\cdot)}\left(\sqrt{\frac{\hat{q}}{\hat{v}}}\boldsymbol{\xi}+\frac{\hat{m}}{\sqrt{\hat{v}}}\boldsymbol{\Omega}^{-1/2}\boldsymbol{\beta}\right)\right]+\frac{\hat{q}}{2\hat{v}}+\frac{\hat{m}^2}{2\hat{q}}\mathbb{E}_{\boldsymbol{\beta}}\left[\frac{\langle\boldsymbol{\beta},\boldsymbol{\Omega}^{-1}\boldsymbol{\beta}\rangle}{p}\right]
\end{aligned} \tag{4.88}$$

Note that for notational convenience we have absorbed the sign in the energetic and entropic potentials in the extremisation problem.

⚠ The term $v\hat{v}$ does not scale with $\beta$, and therefore is subleading in the free energy density eq. (4.87).

**Exercise 5.** Show, either by taking the $\beta\to\infty$ limit of the self-consistent equations eq. (4.54) or by using the following identities:

$$\nabla_{\boldsymbol{x}}\mathcal{M}_{\tau f}(\boldsymbol{x}) = \frac{1}{\tau}\left(\boldsymbol{x}-\mathrm{prox}_{\tau f}(\boldsymbol{x})\right), \qquad \partial_\tau\mathcal{M}_{\tau f}(\boldsymbol{x}) = -\frac{1}{2\tau^2}||\boldsymbol{x}-\mathrm{prox}_{\tau f}(\boldsymbol{x})||_2^2 \tag{4.89}$$

with $\mathrm{prox}_{\tau f}(\boldsymbol{x})$ the proximal operator:

$$\mathrm{prox}_{\tau f}(\boldsymbol{x}) = \underset{\boldsymbol{z}\in\mathbb{R}^p}{\mathrm{argmin}}\left(\frac{1}{2\tau}||\boldsymbol{z}-\boldsymbol{x}||_2^2 + f(\boldsymbol{z})\right) \tag{4.90}$$

That the self-consistent equations for (convex) empirical risk minimisation can be written as:

$$\begin{cases} \hat{v} &= -\alpha \mathbb{E}_\eta \int \mathrm{d}y Z_\star \left( y, \frac{m}{\sqrt{q}}\eta, \rho - \frac{m^2}{q} \right) \partial_\omega f_y(y, \sqrt{q}\eta, v) \\ \hat{q} &= \alpha \mathbb{E}_\eta \int \mathrm{d}y Z_\star \left( y, \frac{m}{\sqrt{q}}\eta, \rho - \frac{m^2}{q} \right) f_y(y, \sqrt{q}\eta, v)^2 \\ \hat{m} &= \alpha \mathbb{E}_\eta \int \mathrm{d}y Z_\star \left( y, \frac{m}{\sqrt{q}}\eta, \rho - \frac{m^2}{q} \right) f_y(y, \sqrt{q}\eta, v) \end{cases} \tag{4.91}$$

$$\begin{cases} v &= \mathbb{E}_{\boldsymbol{\xi},\boldsymbol{\beta}} \left[ \boldsymbol{\nabla_b} \cdot \boldsymbol{f_\theta} \left( \hat{m}\boldsymbol{\beta} + (\hat{q}\boldsymbol{\Omega})^{1/2}\boldsymbol{\xi}, \hat{v}\boldsymbol{\Omega} \right) \right] \\ q &= \mathbb{E}_{\boldsymbol{\xi},\boldsymbol{\beta}} \left[ || \boldsymbol{f_\theta} \left( \hat{m}\boldsymbol{\beta} + (\hat{q}\boldsymbol{\Omega})^{1/2}\boldsymbol{\xi}, \hat{v}\boldsymbol{\Omega} \right) ||_2^2 \right] \\ m &= \mathbb{E}_{\boldsymbol{\xi},\boldsymbol{\beta}} \left[ \langle \boldsymbol{f_\theta} \left( \hat{m}\boldsymbol{\beta} + (\hat{q}\boldsymbol{\Omega})^{1/2}\boldsymbol{\xi}, \hat{v}\boldsymbol{\Omega} \right), \boldsymbol{\beta} \rangle \right] \end{cases} \tag{4.92}$$

with:

$$f_y(y, \omega, v) = \frac{1}{v} \left( \omega - \mathrm{prox}_{v\ell(y,\cdot)}(\omega) \right) \tag{4.93}$$

$$f_{\boldsymbol{\theta}}(\boldsymbol{b}, \boldsymbol{A}) = \mathrm{prox}_{r(\boldsymbol{A}^{-1/2}\cdot)}(\boldsymbol{A}^{-1/2}\boldsymbol{b}) \tag{4.94}$$

The asymptotic free energy density and self-consistent equations for convex (empirical) risk minimisation on the Gaussian covariate model were derived and proven by Loureiro et al. (2021).

*Remark* 6. The Moreau envelope and the proximal operator are standard objects in the field of convex optimisation, where they are related to an optimisation algorithm known as the *proximal method*, see e.g. (Boyd and Vandenberghe, 2004). Although this might sound like a curiosity at this point, their appearance in the context of the replica method will become clear when we discuss Approximate Message Passing.

### Bibliographical notes

It is unclear when exactly the replica trick was first used. In the context of spin glasses, it was introduced by Edwards and Anderson (1975). The first replica computation for the teacher-student generalised linear models is due to Gardner and Derrida (1989). The computation discussed here for a general Gaussian Covariate model appeared in (Loureiro et al., 2021).

## 5 Tool II: Approximate Message Passing

So far our discussion focused on the mathematical study of high-dimensional probability distributions, be it the Curie-Weiss Boltzmann-Gibbs distribution eq. (2.2) or the posterior distribution eq. (3.4) of the Gaussian covariate model. In the statistical physics jargon, we were interest in the *equilibrium properties* of the distribution. The key idea was to identify a set of summary statistics of the problem which can be asymptotically determined in a self-consistent way, without ever having to sample from the high-dimensional measure itself.

However, in many cases we might want to sample a configuration from the Boltzmann-Gibbs or posterior distribution. In other words, an algorithm that given the likelihood and priors (or Hamiltonian) of the problem, returns a sample $\boldsymbol{\theta} \sim P(\boldsymbol{\theta}|\mathcal{D})$.

Sampling from a high-dimensional probability measure is in general a hard problem, which in the worst case requires exponential computational cost in the dimension of the problem. And indeed, developing efficient algorithms for sampling is an active research field. However, can we leverage the asymptotic equilibrium properties we have learned from our analysis to help sampling? The answer is yes, and this is the key idea behind the Approximate Message Passing (AMP) algorithm.

But before delving into AMP, it will be instructive to go look back at the Curie-Weiss model.

### 5.1 Sampling from the Curie-Weiss model

Of course, not all measures are amenable to such a treatment. Indeed, the problems we have looked so far are *mean-field*: the interactions between different variables in the problem is both homogeneous and weak. As a concrete example, consider the Curie-Weiss Hamiltonian eq. (2.1):

$$\begin{aligned}\mathcal{H}(\boldsymbol{s}) &= -\frac{1}{2d} - \left(\frac{1}{d}\sum_{j=2}^{d} s_j + h\right) s_1 - \frac{1}{2d}\sum_{i,j=2}^{d} s_i s_j - h\sum_{i=2}^{d} s_i \\ &= -\left(\frac{1}{d}\sum_{j=2}^{d} s_j + h\right) s_1 + \mathcal{H}_{-1}(\boldsymbol{s}) - \frac{1}{2d}\end{aligned} \tag{5.1}$$

where we decomposed in the terms interacting with spin $i = 1$ and the rest. Fist, note that spin $s_1$ interacts with all the other spins only through the averaged magnetisation $1/d\sum_{j=2}^{d} s_j = \bar{s} + O(1/d)$, which plays a similar role to the external field $h$. Moreover, each spin contributes to the sum with a $O(1/d)$ - but their sum that creates a $O(1)$ effect. Hence the name *mean-field*. The weak correlation between variables imply that the Boltzmann-Gibbs distribution *almost* factorises. To see this, apply a Hubbard-Stratonovich transformation:

$$\mathbb{P}(\boldsymbol{S} = \boldsymbol{s}) = \frac{1}{Z_d} e^{\frac{\beta}{2d}\left(\sum_{i=1}^{d} s_i\right)^2 + \beta h \sum_{i=1}^{d} s_i} = \frac{1}{Z_d}\int \frac{\mathrm{d}\xi}{\sqrt{2\pi/\beta d}} e^{-\frac{\beta d}{2}\xi^2} \prod_{i=1}^{d} e^{\beta(\xi+h)s_i} \tag{5.2}$$

Defining the following conditional distribution:

$$\mathbb{P}(S = s|\xi) = \frac{e^{\beta(\xi+h)s}}{\sum_{s\in\{-1,+1\}} e^{\beta(\xi+h)s}} = \frac{e^{\beta(\xi+h)s}}{2\cosh(\beta(\xi+h))} \tag{5.3}$$

we can exactly rewrite the above as:

$$\mathbb{P}(\boldsymbol{S} = \boldsymbol{s}) = \int \mathrm{d}\xi\ \pi(\xi) \prod_{i=1}^{d} \mathbb{P}(S_i = s_i|\xi) \tag{5.4}$$

with:

$$\pi(\xi) = \frac{1}{Z_d}\sqrt{\frac{\beta d}{2\pi}} e^{-\frac{\beta d}{2}\xi^2 + d\log 2\cosh(\beta(\xi+h))} \tag{5.5}$$

Therefore, conditionally on the field Hubbard-Stratonovich field $\xi \sim \pi$, the Boltzmann-Gibbs distribution factorises $\mathbb{P}(\boldsymbol{S} = \boldsymbol{s}|\xi) = \prod_{i=1}^{d} \mathbb{P}(\boldsymbol{S}_i = \boldsymbol{s}_i|\xi)$. This means that the correlation between the spins are parametrised by a single Gaussian variable. Moreover, note that the variance of $\xi$ is of order $O(1/d)$, and therefore in the high-dimensional limit $d \to \infty$, $\xi$ concentrates:

$$\xi_\star \in \operatorname*{argmin}_{\xi\in\mathbb{R}} \left\{\frac{\beta}{2}\xi^2 - \log 2\cosh(\beta(\xi+h))\right\} \tag{5.6}$$

Looking for zero gradient points recovers the self-consistent equation $\xi = \tanh(\beta(\xi + h))$ that we had found for the magnatisation in Section 2.1. This show us explicitly that the magnetisation is enough to fully characterise the correlations between spins in the high-dimensional limit. Moreover, it give us a very efficient algorithm to sample from the high-dimensional Boltzmann-Gibbs distribution:

1. Solve eq. (5.6) and find $\xi_\star$.
2. Sample every spin $S_i \sim \mathrm{Rad}(p)$ i.i.d. with probability:

$$\mathbb{P}(S = +1|\xi_\star) = \frac{e^{\beta}(\xi_\star + h)}{2\cosh(\beta(\xi_\star + h))} \tag{5.7}$$

**Algorithm 1:** GAMP

**Input:** Data $\boldsymbol{V} \in \mathbb{R}^{n\times p}$, $\boldsymbol{y} \in \mathbb{R}^n$, denoisers $f_y, f_{\boldsymbol{\theta}}$, initial condition $\hat{\boldsymbol{\theta}}^0$.
Define $\boldsymbol{V}^2 = \boldsymbol{V} \odot \boldsymbol{V} \in \mathbb{R}^{n\times p}$ and Initialize $\hat{\boldsymbol{\theta}}^{t=0} = \hat{\boldsymbol{\theta}}^0$, $\hat{\boldsymbol{c}}^{t=0} = \mathbf{1}_d$, $\boldsymbol{g}^{t=0} = \mathbf{0}_n$.
**for** $t \leq T$ **do**
  $\boldsymbol{v}^t = \boldsymbol{V}^2\hat{\boldsymbol{c}}^t$ ; $\boldsymbol{\omega}^t = \boldsymbol{V}\hat{\boldsymbol{\theta}}^t - \boldsymbol{v}^t \odot \boldsymbol{g}^{t-1}$ ; ** Update likelihood mean and variance **
  $\boldsymbol{g}^t = f_y(\boldsymbol{y}, \boldsymbol{\omega}^t, \boldsymbol{v}^t)$ ; $\partial \boldsymbol{g}^t = \partial_\omega f_y(\boldsymbol{y}, \boldsymbol{\omega}^t, \boldsymbol{v}^t)$ ; ** Update likelihood **
  $\boldsymbol{A}^t = -\boldsymbol{V}^{2^\top}\partial \boldsymbol{g}^t$ ; $\boldsymbol{b}^t = \boldsymbol{V}^\top \boldsymbol{g}^t + \boldsymbol{A}^t \odot \hat{\boldsymbol{\theta}}^t$ ; ** Update prior mean and variance*
  */* Update marginals */*
  $\hat{\boldsymbol{\theta}}^{t+1} = f_{\boldsymbol{\theta}}(\boldsymbol{b}^t, \boldsymbol{A}^t)$ ; $\hat{\boldsymbol{c}}^{t+1} = \mathrm{diag}(\nabla_{\boldsymbol{b}} \boldsymbol{f}_{\boldsymbol{\theta}}(\boldsymbol{b}^t, \boldsymbol{A}^t))$
**end for**
**Return:** Estimators $(\hat{\boldsymbol{\theta}}_{\mathrm{amp}}, \hat{\boldsymbol{c}}_{\mathrm{amp}}) := (\hat{\boldsymbol{\theta}}^T, \hat{\boldsymbol{c}}^T)$.

## 5.2 Generalised approximate message passing

The Generalised Approximate Message Passing (GAMP) algorithm 1 is an iterative algorithm specially tailored for the GLM posterior distribution eq. (3.4). It takes as an input the training data $(\boldsymbol{V}, \boldsymbol{y}) \in \mathbb{R}^{n\times p} \times \mathbb{R}^n$ and returns $(\hat{\boldsymbol{\theta}}_{\mathrm{amp}}, \hat{\boldsymbol{c}}_{\mathrm{amp}})$, which corresponds to an estimation of the posterior mean $\mathbb{E}[\boldsymbol{\theta}|\mathcal{D}]$ and variance $\mathrm{Var}(\boldsymbol{\theta}|\mathcal{D})$. It belongs to the class of first-order methods, a class of algorithms that involve only matrix multiplication and entry-wise operations. Therefore, its a very efficient algorithm with running time complexity is $O(np)$, linear in the size of the matrix. The functions $f_{\boldsymbol{\theta}}, f_y$, also refereed to as *denoisers*, are connected to the likelihood and prior:

$$f_y(y,\omega,v) = \frac{\mathbb{E}_{z\sim\mathcal{N}(\omega,v)}\left[\frac{(z-\omega)}{v}\psi(y|z)\right]}{\mathbb{E}_{z\sim\mathcal{N}(\omega,v)}\left[\psi(y|z)\right]}, \qquad f_{\boldsymbol{\theta}}(\boldsymbol{b},\boldsymbol{A}) = \frac{\int \mathrm{d}\boldsymbol{\theta}\varphi(\boldsymbol{\theta})\boldsymbol{\theta}e^{-\frac{1}{2}\langle\boldsymbol{\theta},\boldsymbol{A}\boldsymbol{\theta}\rangle+\langle\boldsymbol{b},\boldsymbol{\theta}\rangle}}{\int \mathrm{d}\boldsymbol{\theta}\varphi(\boldsymbol{\theta})e^{-\frac{1}{2}\langle\boldsymbol{\theta},\boldsymbol{A}\boldsymbol{\theta}\rangle+\langle\boldsymbol{b},\boldsymbol{\theta}\rangle}} \tag{5.8}$$

*Remark* 7. One of the central properties of AMP we will discuss below, the state evolution equations, will follow under minimal assumptions on the denoisers. Therefore, one can in principle consider AMP iterations with other denoiser functions, in which case the estimators $\hat{\boldsymbol{\theta}}_{\mathrm{amp}}, \hat{c}_{\mathrm{amp}}$ will not necessarily correspond to an estimate of the posterior mean and variance.

Although the GAMP algorithm 1 might look mysterious at a first sight, it can be derived from first principles from a reduction of Belief Propagation, a well-known algorithm for inference on graphical models. We refer the interested reader to Appendix D for a detailed discussion of this derivation. Instead, we dedicate this section to building intuition on GAMP.

GAMP approaches the generalised linear estimation problem, i.e. the problem of estimating a signal $\boldsymbol{\beta}_\star \sim P_{\boldsymbol{\beta}}$ from observations $y_i \sim P(y|\langle\boldsymbol{\beta}_\star, \boldsymbol{u}_i\rangle)$, by decomposing it in two parts:

(a) Estimating the pre-activations $\nu_i = \langle\boldsymbol{\beta}_\star, \boldsymbol{u}_i\rangle$ from the observations $y_i \sim P(\cdot|\nu_i)$. Note this is a one-dimensional denoising problem.

(b) The estimation of signal $\boldsymbol{\beta}_\star \in \mathbb{R}^d$ from the pre-activations $\boldsymbol{\nu} = \boldsymbol{U}\boldsymbol{\beta}_\star \in \mathbb{R}^n$. Note this is a high-dimensional but linear inverse problem.

A forward pass of GAMP corresponds precisely that. Given a current estimate $(\hat{\boldsymbol{\theta}}^t, \hat{\boldsymbol{c}}^t)$ of the posterior mean and variances, it starts by computing an estimate for the mean and variance of the pre-activations:

$$\boldsymbol{\omega}^t = \boldsymbol{V}\hat{\boldsymbol{\theta}}^t - \boldsymbol{v}^t \odot \boldsymbol{g}^{t-1}, \qquad \boldsymbol{v}^t = \boldsymbol{V}^2\hat{\boldsymbol{c}}^t. \tag{5.9}$$

The term $\boldsymbol{v}^t \odot \boldsymbol{g}^{t-1}$ is known as the *Onsager term*, and plays a fundamental role in the algorithm. Indeed, a key property of GAMP is that at every step $t < T$, the pre-activations $\boldsymbol{\omega}^t$ are jointly Gaussian

variables with the ground truth pre-activations $\boldsymbol{\nu} = \boldsymbol{V}\boldsymbol{\beta}_\star$. This non-trivial property holds thanks to the presence of the Onsager term. To see this, consider the naive estimator for the pre-activation $\tilde{\boldsymbol{\omega}}^t = \boldsymbol{V}\hat{\boldsymbol{\theta}}^t$. At step $t = 0$, assuming that $\hat{\boldsymbol{\theta}}^0$ is initialised independently from $\boldsymbol{V}$, $\boldsymbol{\omega}^0$ is a sum of independent random variables with variance $O(1/p)$, and therefore it is asymptotically Gaussian by the central limit theorem. However, at step $t = 1$ we have:

$$\tilde{\boldsymbol{\omega}}^1 = \boldsymbol{V}\hat{\boldsymbol{\theta}}^1 \tag{5.10}$$

Following the steps of the algorithm, we see that $\hat{\boldsymbol{\theta}}^1 = \hat{\boldsymbol{\theta}}^1(\boldsymbol{V})$ and hence $\hat{\boldsymbol{\theta}}^1$ is not independent from $\boldsymbol{V}$, hence we cannot iterate the CLT argument. The fact that the algorithm "sees" the same covariates $\boldsymbol{V}$ at every step builds correlations over the time. As it can be shown by carefully tracking the correlations through the updates, the role of the Onsager term is precisely to remove them at every step.

### 5.2.1 State evolution

The joint Gaussianity of the pre-activations $(\boldsymbol{\nu}, \boldsymbol{\omega}^t)$ at every step can be propagated through the algorithm, implying that we can also characterise the distribution of the AMP estimation of the marginals mean $\hat{\boldsymbol{\theta}}^t$ and variance $\hat{\boldsymbol{c}}^t$. Remarkably, it can be shown that in the proportional high-dimensional limit the AMP iterations can be consistently closed on a set of 6 statistics that fully characterise the distribution of the AMP variables:

$$\begin{cases} \hat{v}^t = -\alpha \mathbb{E}_{(\nu,\omega^t)}[\partial_\omega f_y(f_\star(\nu), \omega^t, v^t)] \\ \hat{q}^t = \alpha \mathbb{E}_{(\nu,\omega^t)}[f_y(f_\star(\nu), \omega^t, v^t)^2] \\ \hat{m}^t = \alpha \mathbb{E}_{(\nu,\omega^t)}[\partial_\nu f_y(f_\star(\nu), \omega^t, v^t)] \end{cases}, \qquad \begin{cases} v^{t+1} &= \mathbb{E}_{\boldsymbol{\xi},\boldsymbol{\beta}_\star}\left[\nabla_b \cdot \boldsymbol{f}_\theta(\sqrt{\boldsymbol{\Omega}\hat{q}^t}\boldsymbol{\xi} + \hat{m}^t\boldsymbol{\beta}_\star, \boldsymbol{\Omega}\hat{v}^t)\right] \\ q^{t+1} &= \mathbb{E}_{\boldsymbol{\xi},\boldsymbol{\beta}_\star}\left[||\boldsymbol{f}_\theta(\sqrt{\boldsymbol{\Omega}\hat{q}^t}\boldsymbol{\xi} + \hat{m}^t\boldsymbol{\beta}_\star, \boldsymbol{\Omega}\hat{v}^t)||^2\right], \\ m^{t+1} &= \mathbb{E}_{\boldsymbol{\xi},\boldsymbol{\beta}_\star}\left[\langle \boldsymbol{f}_\theta(\sqrt{\boldsymbol{\Omega}\hat{q}^t}\boldsymbol{\xi} + \hat{m}^t\boldsymbol{\beta}_\star, \boldsymbol{\Omega}\hat{v}^t), \boldsymbol{\beta}_\star\rangle\right] \end{cases} \tag{5.11}$$

where $\boldsymbol{\beta}_\star \sim P_{\boldsymbol{\beta}}$, $\boldsymbol{\xi} \sim \mathcal{N}(\boldsymbol{0}_p, \boldsymbol{I}_p)$ and the likelihood $P_\star$ is parametrised by $y = f_\star(\nu)$ with:

$$(\nu, \omega^t) \sim \mathcal{N}\left(\boldsymbol{0}_2, \begin{bmatrix} \rho & m^t \\ m^t & q^t \end{bmatrix}\right) \tag{5.12}$$

These equations are known as the *state evolution equations*. The quantities $(m^t, q^t, v^t, \hat{m}^t, \hat{q}^t, \hat{v}^t)$ are directly related to the statistics of the GAMP iterates:

- The quantities $(m^t, q^t, v^t)$ track the following statistics:

$$m^t = \lim_{p\to\infty} \mathbb{E}[\nu_i \omega_i^t] = \lim_{p\to\infty} \mathbb{E}\left[\frac{\langle \hat{\boldsymbol{\theta}}^t, \boldsymbol{\beta}_\star\rangle}{p}\right], \tag{5.13}$$

$$q^t = \lim_{p\to\infty} \mathbb{E}[(\omega_i^t)^2] = \lim_{p\to\infty} \mathbb{E}\left[\frac{||\hat{\boldsymbol{\theta}}^t||_2^2}{p}\right], \tag{5.14}$$

$$v^t = \lim_{p\to\infty} \mathbb{E}\left[\frac{\langle \boldsymbol{1}_p, \hat{\boldsymbol{c}}^t\rangle}{p}\right] \tag{5.15}$$

- The quantities $(\hat{m}^t, \hat{q}^t, \hat{v}^t)$ are related to the statistics of the GAMP messages $\boldsymbol{b}_k^t, \boldsymbol{A}_k^t$. More specifically, we can show that asymptotically in $p \to \infty$:

$$\boldsymbol{b}_k^t \sim \hat{m}^t\boldsymbol{\beta}_\star + \sqrt{\boldsymbol{\Omega}\hat{q}}\boldsymbol{\xi}, \qquad \boldsymbol{A}^t \sim \hat{v}^t\boldsymbol{\Omega} \tag{5.16}$$

  with $\boldsymbol{\xi} \sim \mathcal{N}(\boldsymbol{0}, \boldsymbol{I}_p)$ and $\boldsymbol{\beta}_\star \sim P_{\boldsymbol{\beta}}$ independently.

- Together, this implies that we asymptotically have the following characterisation of the AMP estimation of the posterior mean and variance:

$$\hat{\boldsymbol{\theta}}_{\text{amp}} = \boldsymbol{f}_{\boldsymbol{\theta}}(\hat{m}^t \boldsymbol{\beta}_\star + \sqrt{\boldsymbol{\Omega}\hat{q}}\boldsymbol{\xi}, \hat{v}^t \boldsymbol{\Omega}) \in \mathbb{R}^p \tag{5.17}$$

$$\hat{\boldsymbol{c}}_{\text{amp}} = \text{diag}(\nabla_{\boldsymbol{b}} \boldsymbol{f}_{\boldsymbol{\theta}}(\hat{m}^t \boldsymbol{\beta}_\star + \sqrt{\boldsymbol{\Omega}\hat{q}}\boldsymbol{\xi}, \hat{v}^t \boldsymbol{\Omega})) \in \mathbb{R}^p \tag{5.18}$$

  where $\text{diag}(\cdot)$ is the diagonal operator taking a matrix $\boldsymbol{A} \in \mathbb{R}^{p\times p}$ to its diagonal $\text{diag}(\boldsymbol{A}) \in \mathbb{R}^p$

One should stop and appreciate that the fact that the statistics of GAMP can be asymptotically tracked by a low-dimensional, deterministic dynamical system without ever having to run it is quite remarkable. Indeed, most algorithms in machine learning, such as gradient descent or Langevin dynamics, don't admit such a simple descriptions, even for simple mean-field models.

### 5.2.2 Relationship with replicas

The attentive reader probably already realised that the state evolution eq. (5.11) are very similar to the replica self-consistent equations in eq. (4.91). Indeed, a simple change of variable reveals that these equations are identical up to the presence of time indices.

This simple observation has an important consequence: the trajectory of the AMP statistics $(m^t, q^t, v^t, \hat{m}^t, \hat{q}^t, \hat{v}^t)$ is performing gradient descent on the replica symmetric free energy potential eq. (4.50). In other words, on the space of the overlaps the AMP minimisation landscape the replica symmetric free energy landscape. This connection between AMP and the Bethe-Peierls approximation for the free energy (Bethe, 1935).

As we will discuss in Section 6, this relationship will have important consequences to Bayes-optimal inference, where the posterior mean is the minimum mean-squared error (MMSE). This will imply that, whenever AMP is initialised in the basin of the global minima of the free energy, it will achieve the information theoretical best performance (in terms of squared error).

## 5.3 To go further

Both the replica equations and the GAMP algorithm can be readily generalised to multi-index models of the type:

$$y \sim P_\star(\cdot|\boldsymbol{W}_\star \boldsymbol{x}_i) \tag{5.19}$$

for $\boldsymbol{W}_\star \in \mathbb{R}^{k\times d}$ with $k = \Theta_d(1)$ from a prior $P_{\boldsymbol{W}}$ in $\mathbb{R}^{k\times d}$. In particular, this includes narrow two-layer neural networks:

$$f_\theta(\boldsymbol{x}) = \sum_{j=1}^{k} a_k \sigma(\langle \boldsymbol{w}_j, \boldsymbol{x}\rangle) \tag{5.20}$$

where $k = \Theta_d(1)$. These class of models are also known as *committee machines* in the statistical physics literature. The storage capacity of committee machines were studied using the replica method in (Schwarze and Hertz, 1992; Monasson and Zecchina, 1995; Xiong et al., 1997). The Bayes-optimal replica symmetric formula was derived and rigorously proven by Aubin et al. (2018), who also extended the GAMP algorithm for this problem.

## Bibliographical notes

- The approximate message passing algorithm has its roots in investigation of the so-called Thouless-Anderson-Palmer (TAP) equations for the Sherrington-Kirkpatrick (SK) model. These are iterative equations introduced by Thouless et al. (1977) to compute marginals (a.k.a. local magnetisation) of the SK model at high-temperatures. The main feature of these equations is the

presence of an Onsager term introduced to mitigate self-interactions from a naive mean-field approximation. Despite having the correct fixed points at high-temperatures, the original schedule to update the TAP equations was unstable and suffered from convergence issues (Kabashima, 2003a,b). A solution to this problem was given by Bolthausen (2014), who introduced a delayed update for the Onsager term, similar to the update of $\omega^t$ in algorithm 1.

- The relationship between the TAP equations and Pearl's Belief Propagation Pearl (1982) algorithm for Bayesian inference on graphical models was first drawn by Kabashima and Saad (1998a,b) in the context of error correcting codes. However, it gained in popularity with the introduction of the *survey propagation* algorithm for the random K-SAT problem by Mézard et al. (2002).

- The idea of deriving a message passing scheme for compressive sensing from BP appeared in Sarvotham et al. (2006). The Gaussian approximation was proposed by (Donoho et al., 2009), who also coined the name *approximate message passing*. The analysis of the Bayes-optimal case was done in Krzakala et al. (2012b,a).

- The state evolution equations were proved by Bayati and Montanari (2011) for the linear case, and generalised to GLMs by Rangan (2011). The non-separable case was proven by (Berthier et al., 2020) and extended by Gerbelot and Berthier (2023) to message passing schemes defined in more general graphs.

- Optimality of AMP among first order methods was proven in (Celentano et al., 2020).

# Part III
# Lecture 3 — Lessons from simple models

## 6 Statistical-to-computational gaps

Consider the problem of Bayes-optimal inference for an isotropic generalised linear model, discussed in example 1. Given $\mathcal{D} = \{(\boldsymbol{x}_i, y_i) \in \mathbb{R}^{d+1} : i \in [n]\}$ independent observations from the model:

$$y_i \sim P_\star(\cdot|\langle \boldsymbol{\beta}_\star, \boldsymbol{x}_i\rangle) \tag{6.1}$$

with $\boldsymbol{\beta}_\star \sim P_{\boldsymbol{\beta}}$ and $\boldsymbol{x}_i \sim \mathcal{N}(\mathbf{0}_p, {}^1\!/_d \boldsymbol{I}_d)$, the goal is to estimate $\boldsymbol{\beta}_\star$ under the assumption that $P_\star, P_{\boldsymbol{\beta}}$ are known. In this case, the minimum mean-squared error is given by the posterior mean:

$$\hat{\boldsymbol{\theta}}_{\text{mmse}}(\mathcal{D}) = \mathbb{E}[\boldsymbol{\theta}|\mathcal{D}] \tag{6.2}$$

where the expectation is over the posterior distribution:

$$P(\boldsymbol{\theta}|\mathcal{D}) = \frac{1}{Z_d(\mathcal{D})} \prod_{i=1}^{k} P_\beta(\theta_k) \prod_{i=1}^{n} P_\star(y_i|\langle \boldsymbol{\theta}, \boldsymbol{x}_i\rangle) \tag{6.3}$$

where for simplicity we assumed the prior distribution $P_{\boldsymbol{\beta}}(\boldsymbol{\beta}) = \prod_{k=1} P_\beta(\beta_k)$ factorises.

### 6.1 High-dimensional asymptotics

The asymptotic (normalised) mmse is given by:

$$\lim_{d\to\infty} \frac{1}{d} \mathbb{E}\left[||\boldsymbol{\theta}_{\text{mmse}}(\mathcal{D}) - \boldsymbol{\beta}_\star||\right] = \rho - m_\star \tag{6.4}$$

where we used the Nishimori identity eq. (3.7) and defined:

$$\rho = \lim_{d\to\infty} \mathbb{E}\left[\frac{||\boldsymbol{\beta}_\star||_2^2}{d}\right], \qquad m_\star = \lim_{d\to\infty} \mathbb{E}\left[\frac{\langle \hat{\boldsymbol{\theta}}_{\text{mmse}}, \boldsymbol{\beta}_\star\rangle}{d}\right] \tag{6.5}$$

As discussed in section 4.4, the asymptotic overlap $m_\star$ is the extremiser of the following problem:

$$\phi(\rho,\alpha) = \underset{m,\hat{m}}{\textbf{extr}} \left\{-m\hat{m} + \alpha\Phi_y(m) + \Phi_{\boldsymbol{\theta}}(m)\right\} \tag{6.6}$$

with:

$$\Phi_y(m) = \mathbb{E}_\eta \int \mathrm{d}y \; Z_\star(y, \sqrt{q}\eta, \rho - m) \log Z_\star(y, \sqrt{q}\eta, \rho - m) \tag{6.7}$$

$$\Phi_{\boldsymbol{\theta}}(\hat{m}) = \mathbb{E}_{\eta,\beta} \log Z_\beta\left(\hat{m}\beta + \sqrt{\hat{q}}\xi, \hat{m}\right) \tag{6.8}$$

where $\eta \sim \mathcal{N}(0,1)$, $\xi \sim \mathcal{N}(0,1)$ and:

$$Z_\star(y,\omega,v) = \mathbb{E}_{z\sim\mathcal{N}(\omega,v)}[P_\star(y|z)], \qquad Z_\beta(b,A) = \int \mathrm{d}\boldsymbol{\beta} P_\beta(\beta) e^{-\frac{1}{2}A\beta^2 + b\beta} \tag{6.9}$$

associated with the following self-consistent equations:

$$\hat{m}^t = \alpha\mathbb{E}_\eta \int \mathrm{d}y \; Z_\star\left(y, \sqrt{m^t}\eta, \rho - m^t\right) f_\star(y, \sqrt{m^t}\eta, \rho - m^t) \tag{6.10}$$

$$m^{t+1} = \mathbb{E}_{\xi,\beta}\left[f_{\boldsymbol{\beta}}\left(\hat{m}^t\beta + \sqrt{\hat{m}^t}\xi, \hat{m}^t\beta\right)\beta\right] \tag{6.11}$$

where $f_\beta(b,A) = \partial_b \log Z_\beta$, $f_y(y,\omega,v) = \partial_\omega \log Z_y(y,\omega,v)$ and we the time indices to stress the relationship with the GAMP state evolution. In this case, replica symmetry holds and therefore this result is exact. Nevertheless, the potential function in eq. (6.6) might not be convex. As we will see in the following, this will have important algorithmic consequences.

*Remark* 8. Before moving to the discussion of an example, two remarks:

- Note that when no data is observed $\alpha = 0$, we have $\hat{m} = 0$ and:

$$m = \mathbb{E}_\beta[f_\beta(0,0)\beta] = (\mathbb{E}_\beta[\beta])^2 \tag{6.12}$$

  Therefore, even if we have observed no data we can still have a meaningful overlap with the signal if the prior has a non-zero mean. This makes sense, since in the Bayes-optimal we have access to the prior distribution $P_\beta$, which can be informative if its mean is non-zero.

- When running an algorithm such as GAMP, the initial condition $\boldsymbol{\theta}^0$ is typically independent of the signal $\boldsymbol{\beta}_\star$. In the Bayes-optimal setting, the best initialisation which is agnostic to $\boldsymbol{\beta}_\star$ but still leverages the information available is an independent draw from the prior $\boldsymbol{\theta}^0 \sim P_{\boldsymbol{\beta}}$. When the prior has zero-mean (uninformative), this implies $m^0 = 0$, i.e. zero asymptotic correlation with the signal. This corresponds to maximal mean-squared error $\text{mse} = \rho$.

### 6.2 Phase retrieval and the statistical-to-computational gap

In this section we study the phase retrieval problem with Gaussian signal, a particular example that illustrates a phenomenology shared by most of generalised linear estimation problems. It is defined by:

$$P_y(y|z) = \delta(y - |z|), \qquad P_\beta = \mathcal{N}(0,1) \tag{6.13}$$

In this case, we the effective partition functions eq. (6.9) are given by:

$$Z_\beta(b, A) = \int \mathrm{d}\theta \mathcal{N}(\theta|0,1) e^{-\frac{1}{2}A^2\theta^{2+b\theta}} = \frac{e^{\frac{1}{2}\frac{b^2}{1+A}}}{\sqrt{1+A}} \tag{6.14}$$

$$Z_y(y, \omega, v) = \mathbb{E}_{z\sim\mathcal{N}(\omega,v)}[\delta(y - |z|)] = \theta(y)\left[\mathcal{N}(y|\omega, v) + \mathcal{N}(y, -\omega, v)\right] \tag{6.15}$$

where $\theta(t) = \max(0, t)$ is the Heavyside step function. Computing the denoisers, inserting at the self-consistent eq. (6.10) and doing some massage give us:

$$q^{t+1} = \frac{\hat{q}^t}{1+\hat{q}^t}, \qquad \hat{q}^t = \frac{\alpha}{(1-q^t)^2}\mathbb{E}_{\xi,\eta}\left[\left(\sqrt{1-q^t}\eta + \sqrt{q^t}\xi\right)^2 \tanh\left(\sqrt{\frac{q^t}{1-q^t}}\xi\eta + \frac{q^t}{1-q^t}\xi^2\right)^2 - q^t\right] \tag{6.16}$$

with $\xi, \eta \sim \mathcal{N}(0,1)$ independent. Despite being more cumbersome than the Curie-Weiss self-consistent eq. (2.19), it shares two features with the Curie-Weiss model at zero external field $h = 0$:

- Since $y_i = |\langle \boldsymbol{\beta}_\star, \boldsymbol{x}_i\rangle|$ is symmetric under $\boldsymbol{\beta}_\star \to -\boldsymbol{\beta}_\star$, it is information theoretically impossible to distinguish between the $\boldsymbol{\beta}_\star$ and $-\boldsymbol{\beta}_\star$ only from the labels $y_i$. This implies that if $q_\star$ is a solution, $-q_\star$ should also be a solution, and without loss of generality we have restricted to $q \geq 0$ in eq. (6.16).
- The $(\hat{q}_\star, q_\star) = (0, 0)$ is always a fixed point of eq. (6.16).

The second observation has important computational consequences. As we discussed in remark 8, an uninformative initialisation will have asymptotic vanishing overlap, meaning that GAMP is initialised close to the fixed point $(\hat{m}, m) = (0, 0)$. If this fixed point is stable, GAMP will get stuck at initialisation. Intuitively, we expect this fixed point to become unstable as a certain quantity of data is observed, since more data means more information about the signal. In other words, we expect there is a critical sample complexity threshold $\alpha_c$ above which $(\hat{m}, m) = (0, 0)$ is unstable, such that GAMP initialised from a random initial condition will flow away from $(\hat{m}, m) = (0, 0)$ and develop a non-zero overlap $m > 0$ with the signal. This transition is known as the computational *weak recovery* transition.

 The weak recovery transition is computational since even if $(\hat{m}, m) = (0, 0)$ is stable, there could be a lower free energy minima $m_\star > 0$ which would correspond to the mmse. Nevertheless, with high-probability in the dimension $d$, an uninformed initialisation for GAMP will not land in the basin of attraction of this minima. As we will see, this is not the case for phase retrieval, for which $(\hat{m}, m) = (0, 0)$ is the only minima of eq. (6.6) - but this is specific to this problem.

To determine the location of the computational weak recovery transition, we look at the stability of this fixed point $(\hat{q}_\star, q_\star) = (0, 0)$.[7] Recall the following result from dynamical systems:

**Lemma 1.** Consider a discrete dynamical system $x^{t+1} = f(x^t)$ and let $x_\star$ be a fixed point $x_\star = f(x_\star)$ Then, $x_\star$ is stable if $f'(x_\star) < 1$ and unstable if $f'(x_\star) > 1$.

To apply this, we consider the expansion of eq. (6.16) to leading order in $(\hat{q}_\star, q_\star) = (0, 0)$:

$$q^{t+1} = \hat{q}^t + O(q^2) \tag{6.17}$$

$$\hat{q}^t = \alpha\mathbb{E}_{\xi,\eta}\left[(\eta^4\xi^2 - 1)q^t + O(q^2)\right] = 2q^t + O(q^2) \tag{6.18}$$

Therefore, at leading order:

$$q^{t+1} = 2\alpha q^t + O(q^2) \tag{6.19}$$

[7] Which is equivalent to looking at the second derivative of the potential $\Phi$.

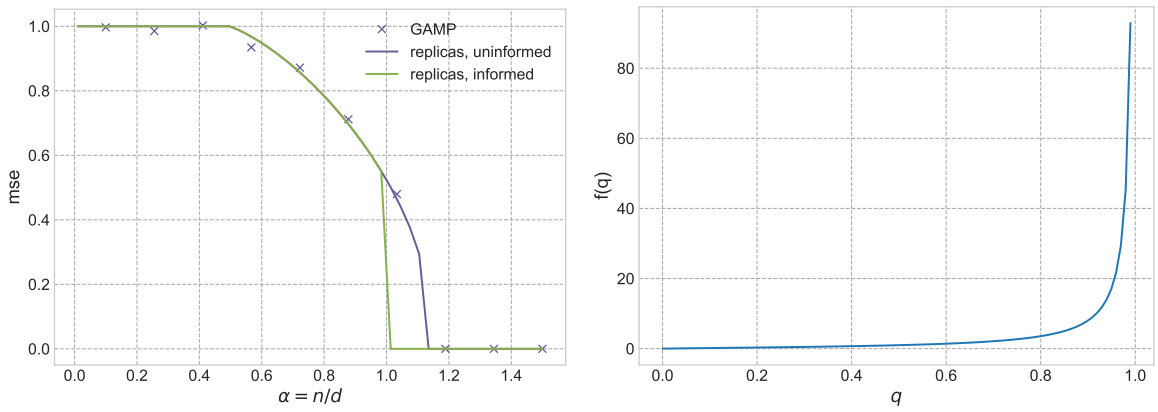


Figure 8: (**Left**) Mean squared error as a function of the sample complexity $\alpha = {}^{n}/_{d}$. The solid curves are obtained from solving the phase retrieval self-consistent equations eq. (6.16) from informed ($m^0 \approx 1$) and uninformed $m^0 \approx 0$ initialisation. Crosses denote finite-size runs of the GAMP algorithm 1 with $d = 1000$. (**Right**) Update function $\hat{q}^t = f(q^t)$ in eq. (6.16) with $\alpha = 1$.

which implies that $(\hat{q}_\star, q_\star) = (0, 0)$ is stable for $\alpha < \alpha_c = {}^{1}/_{2}$. This is the weak recovery sample complexity threshold.

On the opposite side, we can look what happens when a lot of data is available: $\alpha \to \infty$. In this case, $\hat{q} \to \infty$, which implies $q = 1$. This corresponds to a perfect alignment with the signal, also known as the *full-recovery* fixed point. When it exists, we expect this fixed point to be the global minimum of the free energy potential since by definition there cannot be better recovery than full recovery of the signal. For $\alpha = 0$, this is not a fixed point of the self-consistent equations. Indeed, for this to be a fixed point we need $\hat{q}$ to diverge, and since the update function of $\hat{q}$ is a continuous, increasing function of $q^t$ which only diverges at $q = 1$ (see fig. 8 (right)), it is not clear for which $\alpha$ this is the case.

At this point we turn to numerically solving eq. (6.16). As discussed in Section 4.3, this is done by initialising $m^{t=0}$ and numerically iterating. Recall that eq. (6.16) is doing gradient descent on the free energy potential, and therefore when more than one minima is present the iterations will converge to the minima closest to the initial condition $m^{t=0}$. There are two initial conditions which are particularly relevant:

- **Uninformed initial condition:** This corresponds to a random initialisation, independent from the signal $m^{t=0} = \varepsilon \ll 1$. As discussed above, this would be the typical initialisation of an algorithm which is agnostic to the data generating process. Note that we need to initialise away from $m = 0$ when this is a fixed point, otherwise the iterations won't move. The exact size of $\varepsilon$ is problem dependent, but heuristically we want it to be small enough such that $m^{t=0} = \varepsilon$ is on the basin of attraction of 0 when this is a fixed point.

  ⚠ Technically, a random vector $\boldsymbol{v} \sim \mathcal{N}(\mathbf{0}, \boldsymbol{I}_d)$ will have ${}^{1}/_{d}\langle \boldsymbol{v}, \boldsymbol{\beta}_\star \rangle = O({}^{1}/_{\sqrt{d}})$, so in the high-dimensional limit this corresponds exactly to $m^{t=0} = 0$. But in practice $d$ is always finite, and $m^{t=0} = \varepsilon$ is a heuristic to mimic the initial $O({}^{1}/_{\sqrt{d}})$ correlation with the signal. However, this heuristic is not rigorous since eq. (6.16) are only valid asymptotically. Proving when this heuristic is correct is a tour de force and subject research, see e.g. (Rush and Venkataramanan, 2018; Li et al., 2023; Li and Wei, 2024).

- **Informed initial condition:** This corresponds to an initialisation which is has strong correlation with the signal $m^{t=0} = 1 - \varepsilon$ for $\varepsilon \ll 1$. Since by definition the Bayes-optimal posterior mean achieves the mmse, it should correspond to the fixed point with lowest energy and highest

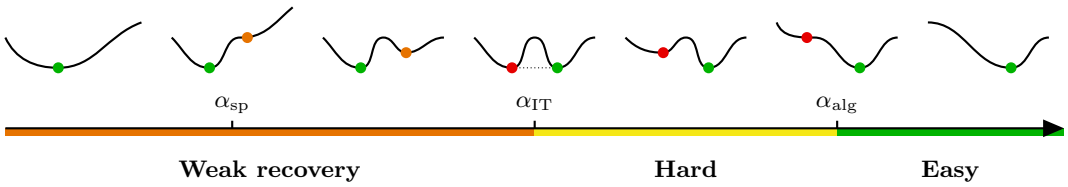


Figure 9: Illustration of the free energy landscape as a function of the overlap with the ground truth solution, when one increases $\alpha$. For small $\alpha < \alpha_{\rm sp}$, there exists a unique global minimum, whose overlap with the solution is small (high MSE). At $\alpha = \alpha_{\rm sp}$, a *local* minimum (orange dot) with higher overlap (small MSE) appears. By definition, the global minimum corresponds to the MMSE of the problem, which is the MSE attained by the Bayes-optimal estimator (green dot). For $\alpha < \alpha_{\rm IT}$ the accessible solution, i.e the global minimum (green dot) has a high MSE while a better solution exists but has a higher free energy (weak recovery phase). At $\alpha = \alpha_{\rm IT}$ the two minima are global and have the same free energy. Between $\alpha_{\rm IT} < \alpha < \alpha_{\rm alg}$ (hard phase), the local minimum with higher MSE corresponds to the performance of the AMP estimator (red dot). Above $\alpha_{\rm alg}$ only the small MSE minima survive and the AMP estimator is able to achieve the Bayes-optimal performance (easy phase).

overlap. This initialisation therefore probes what is the fixed point closest to the perfect recovery of the signal.

With these in mind, in Figure 8 (left) we show the numerical solution of the self-consistent eq. (6.16) from both uninformed and informed initialisations, comparing them with a finite-size run of the GAMP algorithm 1 from a random initial condition. This plot has three important points. From this this

- **Computational weak recovery:** This is the sample complexity threshold below which $\alpha < \alpha_c$ the uninformed fixed point is stable.
- **Information theoretical full recovery:** This is the sample complexity threshold $\alpha > \alpha_{\rm it}$ above which the Bayes-optimal estimator fully aligns with the signal, i.e. the minimum mean-squared error is exactly zero. In other words, this is the minimum amount of data that needs to be observed to information theoretically being able to fully reconstruct the signal. From the self-consistent eq. (6.16) perspective, this is the sample complexity threshold above which iterating from the informed initialisation $m^{t=0} = 1 - \varepsilon$ converges to $m = 1$ for $\varepsilon$ small enough.
- **Computational threshold ($\alpha_{\rm amp}$):** This is the sample complexity threshold $\alpha > \alpha_{\rm amp}$ above which the Bayes-optimal estimator fully aligns with the signal, i.e. the mean square error achieved by GAMP is exactly zero. In other words, this is the minimum amount of data that needs to be observed for GAMP to fully reconstruct the signal. From the self-consistent eq. (6.16) perspective, this is the sample complexity threshold above which iterating from the uninformed initialisation $m^{t=0} = \varepsilon$ converges to $m = 1$ for $\varepsilon$ small enough. Note that by definition we have $\alpha_c < \alpha_{\rm amp}$ and $\alpha_{\rm it} < \alpha_{\rm amp}$.

In Figure 8 (left), we have $\alpha_c = 1/2$, $\alpha_{\rm it} = 1$ and $\alpha_{\rm amp} \approx 1.18$. The region $\alpha \in [\alpha_{\rm it}, \alpha_{\rm amp}]$ is known as the *hard phase*. In this region, it is information theoretically possible to fully reconstruct the signal, but GAMP fails to do so.

It is useful to have also a mental picture of these thresholds in terms of the free energy potential $\Phi(m) = \Phi(m, \hat{m}_\star)$, which is summarised in fig. 9:

- The spinodal threshold $\alpha_{\rm sp}$ is defined as the threshold below which the free energy potential has only a single minima. We will denote it by $m_{1,\star}$.
- For $\alpha < \alpha_c < \alpha_{\rm sp}$, this minimum corresponds to zero overlap with the signal $m_{\star,1} = 0$ and therefore maximal mmse $= 1$.

- The $m = 0$ minimum turns into a maximum at $\alpha = \alpha_c$, with a $m_{\star,1} > 0$ minimum developing nearby. This is the point in which AMP from random initialisation is able to develop small but non-zero correlation with the signal. Note that for the phase retrieval problem, this is also the point in which the Bayes-optimal posterior mean is able to develop correlation with the signal.
- At $\alpha = \alpha_{\rm sp}$, a second minimum $m_{\star,2} > m_{\star,1}$ discontinuously appear, corresponding to higher overlap but higher energy $\Phi(m_{\star,2}) > \Phi(m_{\star,2})$.
- As $\alpha$ is increased, $m_{\star,2}$ continuously lower in energy. At $\alpha_{\rm it}$, the two minima cross $\Phi(m_{\star,1}) = \Phi(m_{\star,2})$, and for $\alpha > \alpha_{\rm it}$ we have $\Phi(m_{1,\star}) < \Phi(m_{\star,2})$. Therefore, up to $\alpha < \alpha_{\rm it}$, the mmse is given by $\text{mmse} = 1 - m_{\star,1}$, while above $\alpha > \alpha_{\rm it}$ it is given by $\text{mmse} = 1 - m_{\star,2}$. In the region $\alpha \in [\alpha_{\rm it}, \alpha_{\rm amp}]$ GAMP initialised from a random initial condition is not sub-optimal, and does not achieve the mmse. This region is the hard phase. Note that for the phase retrieval problem, we have $m_{\star,2} = 1$.
- As $\alpha$ is further increased, the first minimum $m_{1,\star}$ (which is now local), continuously rise in energy, until disappearing at $\alpha = \alpha_{\rm amp}$. Above $\alpha > \alpha_{\rm amp}$, $m_{\star,2}$ is the only minimum, and GAMP initialised from a random initial condition is able to achieve the mmse.

⚠ The picture above arises in many different random estimation problems. However, some parts of it are specific to the phase retrieval problem. For instance, in the phase retrieval problem the weak recovery threshold is both the point in which the mmse and $\text{mse}_{\rm amp}$ becomes non-zero. More generally, $\alpha_c$ is a computational threshold, and nothing prevents $\alpha_c$ to be above $\alpha_{\rm it}$ - in other words, there are problems in which the information theoretical transition occurs earlier than the computational weak recovery transition. This is the case for instance in the sparse subspace clustering problem (Pesce et al., 2022).

### 6.3 Weak recovery for general likelihood

The derivation of the computational weak recovery threshold $\alpha_c$ for the phase retrieval problem in Section 6.2 can actually be carried over for generic likelihood. For simplicity, we focus on the Gaussian prior $\beta \sim \mathcal{N}(0, 1)$

**Existence of the uninformative fixed point** — The first step is to find when $(\hat{m}, m) = (0, 0)$ is a fixed point. Defining the update functions from eq. (6.10):

$$\Lambda_y(t) = \mathbb{E}_\eta \int \mathrm{d}y \; Z_\star\left(y, \sqrt{t}\eta, \rho - t\right) f_\star(y, \sqrt{t}\eta, \rho - t), \qquad \Lambda_\theta(t) = \frac{\hat{m}}{1 + \hat{m}} \tag{6.20}$$

Note that $\hat{m} = 0$ always imply $\hat{m} = 0$. Therefore, it is sufficient to check when $\hat{m} = 0$ is a fixed point. We have:

$$f_\star(y, 0, \rho) = \frac{1}{\rho} \frac{\mathbb{E}_{z\sim\mathcal{N}(0,\rho)}[z P_\star(y|z)]}{\mathbb{E}_{z\sim\mathcal{N}(0,\rho)}[P_\star(y|z)]} \tag{6.21}$$

$$\tag{6.22}$$

Therefore, a sufficient condition for $\Lambda_\theta(0) \overset{!}{=} 0$ is for $P_\star(y|z)$ to be a symmetric function of $z$: $P_\star(y|-z) = P_\star(y|z)$. Note both this condition is satisfied for the phase retrieval likelihood.

**Stability of the uninformative fixed point** — Now, we assume $(\hat{m}, m) = (0, 0)$ is a fixed point. When is it stable? Lemma 1 says we should look for the Jacobian of $(\Lambda_\theta(t), \Lambda_y(t))$. Starting by the $\Lambda_\theta$, it is easy to see that:

$$\Lambda'_\theta(0) = 1. \tag{6.23}$$

Turning now to $\Lambda_y$:

$$\Lambda_y'(0) = \int \mathrm{d}y \frac{\mathbb{E}[((z^2-1)P_\star(y|\sqrt{\rho}z))^2]}{\mathbb{E}[P_\star(y|\sqrt{\rho}z)]} \tag{6.24}$$

Therefore, using lemma 1, the computational weak recovery threshold is given by:

$$\frac{1}{\alpha_c} = \int \mathrm{d}y \frac{\mathbb{E}[((z^2-1)P_\star(y|\sqrt{\rho}z))^2]}{\mathbb{E}[P_\star(y|\sqrt{\rho}z)]} \tag{6.25}$$

### 6.4 To go further

**Optimality of AMP** — As we discussed in Section 5.2.2, state evolution shows that GAMP is effectively performing gradient descent on the same free energy potential than the mmse estimator: the Bayes-optimal posterior mean. This highly non-trivial fact led to a conjecture that GAMP is the optimal polynomial time algorithm for a large class of random estimation problems, implying that the hard phase discussed above is a fundamental computational barrier.

This conjecture was partially proven by Celentano et al. (2020) for the more restrictive computational class of first order methods, algorithms that only perform matrix multiplication and apply entry-wise non-linear functions.

However, for general polynomial time algorithms the conjecture is known to be false. For instance, in the noiseless phase retrieval problem we discussed in section 6.2, Zadik et al. (2022) has shown that a Lenstra-Lenstra-Lovasz lattice basis reduction method is able to achieve perfect reconstruction at the information theoretical threshold $\alpha_{\text{it}} = 1$ with complexity $O(d^6)$. This algorithm exploits the specific geometry of the noiseless phase retrieval problem, and is not robust to noise.

Nevertheless, GAMP remains the best polynomial time algorithm which is robust to noise for most random estimation problems, and it is believed that the computational barriers described in this lectures do capture some fundamental notion of computational hardness. But this is the subject of ongoing research. See (Bandeira et al., 2022; Gamarnik, 2021) that go in this direction.

**Relationship to the low-degree method** — Interestingly, the computational weak recovery threshold derived in eq. (6.25) coincides exactly with the threshold derived from the low-degree method lower bound for functions with so-called generative exponent $k^\star = 2$ (Damian et al., 2024).

**Multi-index model** — The computational weak recovery phase transition was studied in the context of Gaussian multi-index models by Troiani et al. (2024), who provided a full classification of which subspaces of the target span($\boldsymbol{W}_\star$) are trivial, easy or hard to learn in the proportional high-dimensional limit.

**Multi-layer networks** — A multi-layer extension of GAMP was introduced and studied by (Manoel et al., 2017; Gabrié et al., 2018; Aubin et al., 2019, 2020).

### Bibliographical notes

For a review of computational-to-statistical gaps in the context of inference problems, see Zdeborová and Krzakala (2016).

The weak recovery threshold for generalised linear estimation with Gaussian design was derived by Mondelli and Montanari (2018), both in the real and complex cases. The GAMP full-recovery threshold was computed by Barbier et al. (2019) for different real channels, including phase retrieval. The complex case was studied by Maillard et al. (2020), who also generalised this discussion to right-invariant orthogonal matrices.

# 7 Well-specified ridge regression

Let $\mathcal{D} = \{(\boldsymbol{x}_i, y_i)_i \in \mathbb{R}^{d+1} : i \in [n]\}$ denote training data. In this section, we consider well-specified ridge regression under a Gaussian design. This consists of the following problem empirical risk minimisation problem:

$$\min_{\boldsymbol{\theta} \in \mathbb{R}^d} \frac{1}{2} \sum_{i=1}^{n} (y_i - \langle \boldsymbol{\theta}, \boldsymbol{x}_i \rangle)^2 + \frac{\lambda}{2} ||\boldsymbol{\theta}||_2^2 \tag{7.1}$$

under the assumption that the data was generated from a linear model with Gaussian covariates:

$$y_i = \langle \boldsymbol{\beta}_\star, \boldsymbol{x}_i \rangle + z_i, \quad \boldsymbol{x}_i \sim \mathcal{N}(\mathbf{0}_d, {}^1\!/_d \boldsymbol{\Omega}), \quad z_i \sim \mathcal{N}(0, \sigma^2) \tag{7.2}$$

This is a particular case of the Gaussian covariate model introduced in Section 3 with $\boldsymbol{u} = \boldsymbol{v}$ ($p = d$), squared loss $\ell(y, z) = {}^1\!/_2 (y - z)^2$ and $\ell_2$ penalty $r(\boldsymbol{\theta}) = {}^\lambda\!/_2 ||\boldsymbol{\theta}||_2^2$.

Before looking at the asymptotic solution, note that since this is a quadratic problem it admits a closed-form solution for the estimator:

$$\hat{\boldsymbol{\theta}}_\lambda(\mathcal{D}) = \left( \boldsymbol{X}^\top \boldsymbol{X} + \lambda \boldsymbol{I}_d \right)^{-1} \boldsymbol{X}^\top \boldsymbol{y} \tag{7.3}$$

where $\boldsymbol{X} \in \mathbb{R}^{n \times d}$ is the matrix obtained by stacking $\boldsymbol{x}_i \in \mathbb{R}^d$ row-wise and $\boldsymbol{y} \in \mathbb{R}^n$ is the vector of labels. Our goal is to get a high-dimensional characterisation of the empirical and population risks:

$$R(\hat{\boldsymbol{\theta}}_\lambda) = \mathbb{E}_{(\boldsymbol{x}, y)} \left[ (y - \langle \hat{\boldsymbol{\theta}}, \boldsymbol{x} \rangle)^2 \right], \qquad \hat{R}_n(\hat{\boldsymbol{\theta}}_\lambda) = \frac{1}{n} \sum_{i=1}^{n} \left( y_i - \langle \hat{\boldsymbol{\theta}}_\lambda, \boldsymbol{x}_i \rangle \right)^2 \tag{7.4}$$

In the limit $n, d \to \infty$ with fixed $\alpha = {}^n\!/_d$. As the exercise below illustrates, this problem naturally leads to a random matrix theory problem, and could be naturally approached using results from this field.

**Exercise 6.** Show that the excess risk:

$$r(\hat{\boldsymbol{\theta}}_\lambda) = \mathbb{E}_{\boldsymbol{z}} \mathbb{E}_{(\boldsymbol{x}, y)} \left[ (y - \langle \hat{\boldsymbol{\theta}}_\lambda, \boldsymbol{x} \rangle)^2 \right] - \sigma^2 \tag{7.5}$$

can be decomposed in terms of a bias and variance term:

$$r(\hat{\boldsymbol{\theta}}_\lambda) = B + V \tag{7.6}$$

with:

$$B = \lambda^2 \langle \boldsymbol{\beta}_\star, \left( \boldsymbol{X}^\top \boldsymbol{X} + \lambda \boldsymbol{I}_d \right)^{-1} \boldsymbol{\Omega} \left( \boldsymbol{X}^\top \boldsymbol{X} + \lambda \boldsymbol{I}_d \right)^{-1} \boldsymbol{\beta}_\star \rangle \tag{7.7}$$

$$V = \frac{\sigma^2}{d} \operatorname{Tr} \left\{ \boldsymbol{\Omega} \left( \boldsymbol{X}^\top \boldsymbol{X} + \lambda \boldsymbol{I}_d \right)^{-2} \boldsymbol{X}^\top \boldsymbol{X} \right\} \tag{7.8}$$

Therefore, show that in the classical statistical limit $n \to \infty$ at fixed $d = O(1)$, we have:

$$\lim_{n \to \infty} B = \frac{({}^\lambda\!/_\alpha)^2}{d} \langle \boldsymbol{\beta}_\star, \boldsymbol{\Omega} \left( \boldsymbol{\Omega} + {}^\lambda\!/_\alpha \boldsymbol{I}_d \right)^{-2} \boldsymbol{\beta}_\star \rangle \tag{7.9}$$

$$\lim_{n \to \infty} V = \frac{\sigma^2 \alpha}{d} \operatorname{Tr} \left\{ \boldsymbol{\Omega}^2 \left( \boldsymbol{\Omega} + {}^\lambda\!/_\alpha \boldsymbol{I}_d \right)^{-2} \right\} \tag{7.10}$$

However, to illustrate our results from Section 4 we approach this problem the statistical physics way.

### 7.1 Asymptotic solution

Well-specified ridge regression is perhaps the simplest empirical risk minimisation problem, and in this case all the integrals in the self-consistent equations eq. (4.91) can be done analytically:

$$\begin{cases} \hat{v} = \frac{\alpha}{1+v} \\ \hat{q} = \alpha \frac{\rho+\sigma^2+q-2m}{(1+v)^2} \\ \hat{m} = \frac{\alpha}{1+v} \end{cases} \qquad \begin{cases} v = \frac{1}{d} \operatorname{Tr} \boldsymbol{\Omega}(\lambda \boldsymbol{I}_d + \hat{v}\boldsymbol{\Omega})^{-1} \\ q = \frac{1}{d} \operatorname{Tr} \left(\hat{q}\boldsymbol{\Omega} + \hat{m}^2 \boldsymbol{\Phi}\boldsymbol{\beta}_\star\boldsymbol{\beta}_\star^\top\boldsymbol{\Phi}^\top\right) \boldsymbol{\Omega}(\lambda \boldsymbol{I}_d + \hat{v}\boldsymbol{\Omega})^{-2} \\ m = \frac{\hat{m}}{d} \operatorname{Tr} \boldsymbol{\Phi}\boldsymbol{\beta}_\star\boldsymbol{\beta}_\star^\top\boldsymbol{\Phi}^\top(\lambda \boldsymbol{I}_d + \hat{v}\boldsymbol{\Omega})^{-1} \end{cases} \tag{7.11}$$

with the asymptotic population and empirical risks given by:

$$\lim_{d\to\infty} \mathbb{E}_{\mathcal{D}}[R(\hat{\boldsymbol{\theta}}_\lambda(\mathcal{D}))] = \rho + q_\star - 2m_\star \tag{7.12}$$

$$\lim_{d\to\infty} \mathbb{E}_{\mathcal{D}}[\hat{R}_n(\hat{\boldsymbol{\theta}}_\lambda(\mathcal{D}))] = \hat{q}_\star = \frac{\rho + q_\star - 2m_\star}{(1+\hat{v}_\star)^2} \tag{7.13}$$

with $q_\star, m_\star, v_\star$ solutions of the self-consistent equations eq. (7.11). Massaging these 6 equations, we can obtain a compact characterisation of the excess risk in terms of a single equation, see Appendix Appendix E.1 for the details. Indeed, letting:

$$\nu = \frac{\lambda}{\hat{v}} \tag{7.14}$$

we can show that the bias and variance decomposition of the excess risk can asymptotically written as:

$$B(\hat{\boldsymbol{\theta}}_\lambda) \underset{d\to\infty}{\sim} \mathcal{B}(\alpha,\lambda) = \frac{\alpha\nu_\star^2 {}^1\!/_d \langle \boldsymbol{\beta}_\star, \boldsymbol{\Omega}\left(\nu_\star \boldsymbol{I}_p + \boldsymbol{\Omega}\right)^{-2}\boldsymbol{\beta}_\star\rangle}{\alpha - {}^1\!/_d \operatorname{Tr}\boldsymbol{\Omega}^2(\boldsymbol{\Omega}+\nu_\star\boldsymbol{I}_d)^{-2})} \tag{7.15}$$

$$V(\hat{\boldsymbol{\theta}}_\lambda) \underset{d\to\infty}{\sim} \mathcal{V}(\alpha,\lambda) = \sigma^2 \frac{{}^1\!/_d \operatorname{Tr}\boldsymbol{\Omega}^2(\boldsymbol{\Omega}+\nu_\star\boldsymbol{I}_d)^{-2}}{\alpha - {}^1\!/_d \operatorname{Tr}\boldsymbol{\Omega}^2(\boldsymbol{\Omega}+\nu_\star\boldsymbol{I}_d)^{-2})} \tag{7.16}$$

where $\nu_\star(\alpha,\lambda)$ is the solution of the following self-consistent equation:

$$\alpha\nu - \lambda = \frac{\nu}{d} \operatorname{Tr}\boldsymbol{\Omega}\left(\boldsymbol{\Omega} + \nu\boldsymbol{I}_d\right)^{-1}. \tag{7.17}$$

This result agrees with the random matrix theory derivations in the literature, e.g. (Hastie et al., 2022; Bach, 2024) - see bibliographical note below.

⚠ Note that in our replica derivation in Section 4, we worked with the unormalised empirical risk eq. (7.1), which differs from the convention adopted in other works such as (Bach, 2024), where the risk is normalised. These are related by constant factors of the sample complexity $\alpha$ and scaling of the regularisation magnitude $\lambda$.

*Remark* 9. The following traces appearing in the asymptotic expressions:

$$\mathrm{df}_a(\nu) = \operatorname{Tr}\boldsymbol{\Omega}^a(\boldsymbol{\Omega}+\lambda\boldsymbol{I}_d)^{-a}, \qquad a = 1,2 \tag{7.18}$$

are known as degrees-of-freedom, and are classical quantities in the statistical learning analysis of ridge regression and kernel methods (Zhang, 2005; Caponnetto and De Vito, 2007). They are decreasing functions of $\lambda$, and quantify the effective dimensionality of the matrix $\boldsymbol{\Omega}$, and satisfy:

$$0 \le \mathrm{df}_2(\lambda) \le \mathrm{df}_1(\lambda) \le \operatorname{rank}(\boldsymbol{\Omega}_n) \tag{7.19}$$

with equality on the right at $\lambda = 0^+$. The interpretation of the asymptotic results eq. (7.15) and (7.16) in terms of degrees-of-freedom was drawn by Bach (2024).

Comparing the classical statistical limit of the bias and variance eq. (7.9) and eq. (7.10) with the proportional asymptotic expression eq. (7.15) and (7.16), we remark that:

- In the proportional asymptotics, we have a self-induced ridge regularisation $\lambda/\alpha \to \nu_\star(\alpha, \lambda)$ given by the solution of the self-consistent equation eq. (7.17). In particular, since $\lambda \mapsto \nu_\star(\alpha, \lambda)$ is an increasing function, the self-induced regularisation $\nu_\star(\alpha, \lambda) \geq \lambda/\alpha$ is larger than the original regularisation (with equality when $\alpha \to \infty$).
- We have an additional multiplicative factor in both the bias and variance proportional to:

$$\frac{1}{\alpha} \leq \frac{1}{\alpha - 1/d \operatorname{Tr} \boldsymbol{\Omega}^2 (\boldsymbol{\Omega} + \nu \boldsymbol{I}_d)^2} \leq \frac{1}{\alpha - 1} \tag{7.20}$$

  In particular, note that this term diverges if:

$$\mathrm{df}_2(\nu) = \operatorname{Tr} \boldsymbol{\Omega}^2 (\boldsymbol{\Omega} + \nu_\star \boldsymbol{I}_d)^2 = n \tag{7.21}$$

## 7.2 Interpolator ($\lambda = 0^+$)

We now assume that $\boldsymbol{\Omega}$ is full-rank[8], and consider the ridgeless or least-squares limit $\lambda \to 0^+$. In this case, the self-consistent equation reads:

$$\alpha\nu = \frac{\nu}{d} \operatorname{Tr} \boldsymbol{\Omega} \left(\boldsymbol{\Omega} + \nu \boldsymbol{I}_d\right)^{-1} \tag{7.22}$$

Since $\operatorname{Tr} \boldsymbol{\Omega}(\boldsymbol{\Omega} + \nu \boldsymbol{I}_p) \leq d$, for $\alpha > 1$ ($n > d$) this equation can only be satisfied by $\nu_\star(\alpha) = 0$ - i.e. no self-induced regularisation. Inserting this in eq. (7.15) and (7.16):

$$\mathcal{B}(\alpha) = 0, \qquad \mathcal{V}(\alpha) = \frac{\sigma^2}{\alpha - 1}, \qquad \alpha > 1 \tag{7.23}$$

In the $\alpha < 1$ regime, eq. (7.22) has a single solution non-zero solution $\nu_\star(\alpha) > 0$ and such that:

$$\frac{1}{d} \operatorname{Tr} \boldsymbol{\Omega} \left(\boldsymbol{\Omega} + \nu_\star(\alpha) \boldsymbol{I}_d\right)^{-1} = \alpha. \tag{7.24}$$

The exact solution depends on the details of $\boldsymbol{\Omega} \in \mathbb{R}^{d \times d}$. Note that in this case, since:

$$0 \leq \mathrm{df}_2(\nu_\star(\alpha)) \leq \mathrm{df}_1(\nu_\star(\alpha)) = n \tag{7.25}$$

we will have a divergence of the excess risk only if the two degrees of freedom coincide: $\mathrm{df}_2(\nu_\star(\alpha)) = n$. For example, for an isotropic covariance $\boldsymbol{\Omega} = \boldsymbol{I}_d$, we have:

$$\nu_\star(\alpha) = \frac{1}{\alpha} - 1, \qquad \alpha > 1 \tag{7.26}$$

ad therefore the bias and variables are given by:

$$\mathcal{B}(\alpha) = \frac{\alpha^2 \rho}{1 - \alpha}, \qquad \mathcal{V}(\alpha) = \frac{\alpha \sigma^2}{1 - \alpha}. \tag{7.27}$$

where $\rho = 1/d ||\boldsymbol{\beta}_\star||_2^2$. Therefore, we have a divergence of both the bias and variance at $\alpha = 1$.

[8] $\operatorname{rank}(\boldsymbol{\Omega}) = d$

**Bibliographical note**

There is a vast literature on ridge regression under Gaussian design. On the physics side, one of the earliest derivations we are aware for the exact asymptotic risk appeared in (Hertz, 1991; Krogh and Hertz, 1991, 1992), under isotropic covariates and isotropic random signal. This same case was rigorously studied by (Karoui, 2013; Dicker, 2016), and later generalised to anisotropic covariance by Dobriban and Wager (2018) and anisotropic random signal by Wu and Xu (2020). The case of deterministic signal was recently treated in Hastie et al. (2022). Non-asymptotic multiplicative bounds for the risk were proven by Cheng and Montanari (2022). All the aforementioned results are based on a RMT treatment of the problem. An analogous result for the isotropic case was proven with Convex Gaussian min-max theorem (CFMT) in Thrampoulidis et al. (2015).

# Part IV
# Lecture 4 — Shallow networks

## 8 The random features model

In this lecture we turn our attention to perhaps the simplest of the neural network models: two-layer neural networks at initialisation, also known as the *random features* (RF) model. This consists of parametric functions of the type:

$$f(\boldsymbol{x}) = \frac{1}{\sqrt{p}} \sum_{j=1}^{p} a_j \sigma(\langle \boldsymbol{w}_j^0, \boldsymbol{x} \rangle) \tag{8.1}$$

where the first-layer weights $\boldsymbol{w}_j^0$ are randomly drawn and fixed at their initial value, and the first layer weights $a_j$ are trained via empirical risk minimisation with training data $\mathcal{D} = \{(\boldsymbol{x}_i, y_i) : i \in [n]\}$:

$$\min_{\boldsymbol{a} \in \mathbb{R}^p} \frac{1}{2n} \sum_{i=1}^{n} \left( y_i - \frac{1}{\sqrt{p}} \langle \boldsymbol{a}, \sigma(\boldsymbol{W}_0, \boldsymbol{x}_i) \rangle \right)^2 + \frac{\lambda}{2} ||\boldsymbol{a}||_2^2. \tag{8.2}$$

where we have defined the first-layer weight matrix $\boldsymbol{W}_0 \in \mathbb{R}^{p \times d}$ with rows $\boldsymbol{w}_{0,j} \in \mathbb{R}^p$ and the second layer vector $\boldsymbol{a} \in \mathbb{R}^p$.

The RF model was first introduced by Balcan et al. (2006); Rahimi and Recht (2007, 2008) as a computationally efficient approximation scheme to kernel methods[9]. Indeed, assuming the rows of $\boldsymbol{W}_0 \in \mathbb{R}^{p \times d}$ to be independently drawn $\boldsymbol{w}_{0,j} \sim p_{\boldsymbol{w}}$ from some distribution, the (random) features map can be seen as an empirical approximation of a kernel:

$$K(\boldsymbol{x}, \boldsymbol{x}') = \mathbb{E}_{\boldsymbol{w}} \left[ \sigma(\langle \boldsymbol{w}, \boldsymbol{x} \rangle) \sigma(\langle \boldsymbol{w}, \boldsymbol{x}' \rangle) \right] \approx \frac{1}{p} \sum_{j=1}^{p} \sigma(\langle \boldsymbol{w}_j, \boldsymbol{x} \rangle) \sigma(\langle \boldsymbol{w}_j, \boldsymbol{x}' \rangle) \tag{8.3}$$

The simplest example are translationally-invariant kernel $K(\boldsymbol{x}, \boldsymbol{x}') = \kappa(||\boldsymbol{x} - \boldsymbol{x}'||_2)$, which can be explicitly constructed with random Fourier features (Rahimi and Recht, 2007) by choosing $p_{\boldsymbol{w}}$ accordingly. Defining the feature matrix $\Phi \in \mathbb{R}^{n \times p}$ obtained by stacking $\boldsymbol{\varphi}(\boldsymbol{x}_i)$ row-wise and the response vector $\boldsymbol{y} \in \mathbb{R}^n$, eq. (8.2) admits an unique closed-form solution given by:

$$\begin{aligned} \hat{\boldsymbol{a}}_\lambda(\boldsymbol{\Phi}, \boldsymbol{y}) &= \frac{1}{n} \left( \frac{1}{n} \boldsymbol{\Phi}^\top \boldsymbol{\Phi} + \lambda \boldsymbol{I}_p \right)^{-1} \boldsymbol{\Phi}^\top \boldsymbol{y} \\ &= \frac{1}{n} \boldsymbol{\Phi}^\top \left( \frac{1}{n} \boldsymbol{\Phi} \boldsymbol{\Phi}^\top + \lambda \boldsymbol{I}_n \right)^{-1} \boldsymbol{y}. \end{aligned} \tag{8.4}$$

[9] Using the kernel trick, a kernel predictor $f(\boldsymbol{x}) = \sum_{i=1}^{n} c_i K(\boldsymbol{x}, \boldsymbol{x}_i)$ involves $O(nd)$ operations, while a RF approximation $f(\boldsymbol{x}) = \langle \boldsymbol{a}, \sigma(\boldsymbol{W}_0 \boldsymbol{x}) \rangle$ involves $O(p + d)$ operations.

Note the two expressions are related by a simple matrix identity. This is a direct consequence of the fact the (regularised) empirical risk is a quadratic, strongly convex problem for all $\lambda > 0$. In particular, in the limit $\lambda \to 0^+$, we have:

$$\hat{\boldsymbol{a}}_0(\boldsymbol{\Phi}, \boldsymbol{y}) = \boldsymbol{\Phi}^\dagger \boldsymbol{y} \tag{8.5}$$

where $(\cdot)^\dagger$ is the Moore-Penrose pseudo-inverse[10]. Note this is equivalent to the least-squares predictor when $n \geq p$.

## 8.1 Setting & Assumptions

Our main working assumption concerning the data distribution are the following:

**Assumption 1** (Gaussian covariates)**.** We assume the covariates are i.i.d. Gaussian variables with zero mean and unit variance, in other words: $\boldsymbol{x}_i \sim \mathcal{N}(\boldsymbol{0}, {}^1\!/_d \boldsymbol{I}_d)$ independently for all $i \in [n]$.

**Assumption 2** (Target function)**.** Given the covariates $\boldsymbol{x}_i \in \mathbb{R}^d$, we assume the responses are given by:

$$y_i = f_\star(\boldsymbol{x}_i) + \varepsilon_i, \tag{8.6}$$

with $f_\star \in L^2(\mathbb{R}^d, \gamma_d)$ and $\varepsilon \sim \mathcal{N}(0, \sigma^2)$. For simplicity, we will also assume $\mathbb{E}[f(\boldsymbol{x})] = 0$ so that the responses are centered $\mathbb{E}[y] = 0$.

As in our previous discussion on high-dimensional asymptotics, in the following we will mostly focus on a proportional scaling.

**Assumption 3** (Proportional asymptotics)**.** We assume $p, d, n \to \infty$ at fixed rates $\alpha \coloneqq {}^n\!/_p$, $\gamma \coloneqq {}^p\!/_d$.

A few comments about these assumptions are in place:

- The function $f_\star : \mathbb{R}^d \to \mathbb{R}$ is often refereed as the *target function* in the learning theory literature and as *teacher function* in the Statistical Physics of learning literature. Indeed, the idea of studying the typical properties of a predictor $f \in \mathcal{H}$ in learning a random target $f_\star$ function dates back to (Gardner and Derrida, 1989) in the Statistical Physics literature, where it is known as the *teacher-student framework*.

- Assumption 1 implies that the covariates are isotropic, and therefore the structure in the training data lies entirely in the target function $f_\star$. Of course, real data is structured, with important directions in the covariates typically correlating with the responses $y_i$. For instance, a simple principal component analysis (PCA) on MNIST is enough to reveal some information about the clustered structure of the digits. Therefore, we intuitively expect the results derived under the Gaussian assumption on the covariates 1 to represent an upper bound on the generalisation properties of networks trained on real data.

- Note that the assumption on the existence of a target function $f_\star$ in Assumption 2 is not really restrictive, since we can always take $f_\star$ to be the Bayes predictor. The restrictive aspect of Assumption 2 is to assume that $f_\star \in L^2(\mathbb{R}^d, \gamma_d)$ and the noise is additive Gaussian.

- Note that in practice it is hard to distinguish what is a natural scaling assumption for $n, p, d$. For instance, say we want to fit MNIST, a data set with $n = 70000$ samples of dimension $d = 784$, do we have $n = \Theta(d)$, $n = \Theta(d^{1.5})$ or $n = \Theta(d^2)$? As we will see later, in many different problems this distinction is not important, as the asymptotic results derived under a tractable scaling are close to their non-asymptotic counterpart.

[10] A simple characterisation of the Moore-Penrose pseudo-inverse of retangular matrix $\boldsymbol{A} \in \mathbb{R}^{m \times n}$ is given in terms its SVD. Letting $\boldsymbol{A} = \sum_{j=1}^r \lambda_j \boldsymbol{u}_j \boldsymbol{v}_j^\top$ with $r = \mathrm{rank}(A)$, we have $\boldsymbol{A}^\dagger = \sum_{j=1}^r \lambda_j^{-1} \boldsymbol{v}_j \boldsymbol{u}_j^\top \in \mathbb{R}^{n \times m}$.

## 8.2 Fundamental limitations in the high-dimensional regime

The ridge operator in eq. (8.4):

$$\boldsymbol{y} \in \mathbb{R}^n \mapsto \frac{1}{n}\left(\frac{1}{n}\boldsymbol{\Phi}^\top\boldsymbol{\Phi} + \lambda \boldsymbol{I}_p\right)^{-1}\boldsymbol{\Phi}^\top \boldsymbol{y} \tag{8.7}$$

projects the response onto the column-space of $\text{Image}(\Phi^\top) \subset \mathbb{R}^p$, which is a linear subspace of the feature space. To see this mathematically, denote by $\boldsymbol{\Phi} = \sum_{j=1}^r \lambda_j \boldsymbol{u}_j \boldsymbol{v}_j^\top$ the singular-value decomposition of the features $\boldsymbol{\Phi}$ with $r := \text{rank}(\boldsymbol{\Phi}) \leq \min(n, p)$. Then, we can re-write eq. (8.5) as:

$$\hat{\boldsymbol{a}}_\lambda(\Phi, \boldsymbol{y}) = \sum_{j=1}^{r} \frac{\lambda_j}{\lambda_j^2 + n\lambda}\langle \boldsymbol{u}_j, \boldsymbol{y}\rangle \boldsymbol{v}_j \tag{8.8}$$

Therefore, the predictor $f(\boldsymbol{x}; \hat{\boldsymbol{a}}_\lambda) = \langle \hat{\boldsymbol{a}}_\lambda, \boldsymbol{\varphi}(\boldsymbol{x})\rangle$ can learn at best a linear component of the target function $f_\star$ in the space spanned by the features $\boldsymbol{\varphi}(\boldsymbol{x})$. For instance, in the vanilla ridge case $\boldsymbol{\varphi}(\boldsymbol{x}) = \boldsymbol{x}$ this would imply that only a linear component of the target can be learned: $f_\star(\boldsymbol{x}) = \langle \boldsymbol{\beta}_\star, \boldsymbol{x}\rangle + f_\star^{>1}(\boldsymbol{x})$, with the non-linear component $f_\star^{>1}$ effectively behaving as part of the label noise when projected on $\hat{\boldsymbol{a}}_\lambda$. A non-linear feature map $\boldsymbol{\varphi}(\boldsymbol{x})$ therefore allows, in principle, to learn higher order, non-linear components.

To make this discussion more concrete, it is useful to decompose the target function in an orthonormal basis with respect to the distribution of the covariates. Since we assume $\boldsymbol{x}_i \sim \mathcal{N}(0, {}^1\!/_d \boldsymbol{I}_d)$, this is given by the Hermite polynomials:

$$f_\star(\boldsymbol{x}) = \sum_{\boldsymbol{\alpha} \in \mathbb{N}^d} c_{\boldsymbol{\alpha}} h_{\boldsymbol{\alpha}}(\boldsymbol{x}) \tag{8.9}$$

where $h_{\boldsymbol{\alpha}}(\boldsymbol{x})$ are the Hermite tensors, which form an orthonormal basis of $L^2(\mathbb{R}^d, \gamma_d)$. See Appendix A for a detailed introduction. This basis induces an orthogonal decomposition of $L^2(\mathbb{R}^d, \gamma_d) = \bigoplus_{\ell \geq 1} V_\kappa$, where $V_\kappa$ is the linear space spanned by polynomials of degree $\ell = |\boldsymbol{\alpha}|$. The coefficients $c_{\boldsymbol{\alpha}}$ quantify how much of the total energy of the target $||f_\star||_{\gamma_d}^2 = \sum_{\boldsymbol{\alpha}} c_{\boldsymbol{\alpha}}^2$ lies in each subspace.

Assuming the features $\boldsymbol{\Phi}$ are full-rank $r = \min(n, p)$[11], since the ridge predictor in eq. (8.8) spans a linear subspace of dimension $r$, a naive power counting suggests that to learn the component of the target in subspace $V_\ell$ requires $r = O(d^\ell)$, with the minimum between the number of samples $n$ and the width $p$ being the bottleneck for approximating $V_\ell$. Therefore, in a polynomial scaling regime $n, p = \Theta(d^\ell)$, we can learn at best a degree $\ell$ polynomial approximation of the target function $f_\star$. In particular, under the proportional asymptotics Assumption 3 which will be our focus in the following, it is enough to consider a linear target function $f_\star(\boldsymbol{x}) = \langle \boldsymbol{\beta}_\star, \boldsymbol{x}\rangle$.

⚠ It is important to keep in mind the discussion in this section is specific to ridge regression.

## 8.3 Gaussian universality

An important consequence of the discussion in Section 8.2 is that in the high-dimensional limit a random features map sees the target function at a limited resolution. This discussion can be made more quantitative, and is at the heart of the exact asymptotic characterisation of the generalisation error that will be discussed in Section 8.4. Considering the expansion of the feature map in the Hermite basis:

$$\varphi_j(\boldsymbol{x}) = \sigma\left(\langle \boldsymbol{w}_{0,j}, \boldsymbol{x}\rangle\right) = \sum_{\ell \geq 0} b_\ell h_\ell(\langle \boldsymbol{w}_{0,j}, \boldsymbol{x}\rangle) \tag{8.10}$$

[11] For instance, for $\boldsymbol{x}_i \sim \mathcal{N}(\boldsymbol{0}, {}^1\!/_d \boldsymbol{I}_d)$ and $\boldsymbol{w}_{0,j} \sim \mathcal{N}(\boldsymbol{0}, \boldsymbol{I}_d)$, $\boldsymbol{\Phi} = {}^1\!/_{\sqrt{p}}\sigma(\boldsymbol{X}\boldsymbol{W}_0^\top)$ will be a full-rank matrix with high-probability with respect to $\boldsymbol{X}, \boldsymbol{W}$.

Its first and second moments are given by:

$$\mathbb{E}_{\boldsymbol{x}}[\sigma\left(\langle \boldsymbol{w}_{0,j}, \boldsymbol{x}\rangle\right)] = b_0 \tag{8.11}$$

$$\mathbb{E}_{\boldsymbol{x}}[\sigma\left(\langle \boldsymbol{w}_{0,j}, \boldsymbol{x}\rangle\right)\sigma\left(\langle \boldsymbol{w}_{0,k}, \boldsymbol{x}\rangle\right)] = \sum_{\ell\geq 0} b_\ell^2 \left(\frac{\langle \boldsymbol{w}_{0,j}, \boldsymbol{w}_{0,k}\rangle}{d}\right)^\ell \tag{8.12}$$

In particular, note that if $\boldsymbol{w}_{0,j} \sim \mathcal{N}(\boldsymbol{0}, \boldsymbol{I}_d)$, with high-probability $1/d\langle \boldsymbol{w}_{0,j}, \boldsymbol{w}_{0,k}\rangle = O(d^{-1/2})$ for $j \neq k$ and $1/d||\boldsymbol{w}_{0,j}||^2 = 1$, meaning that to leading order in $d$, the features population covariance $\boldsymbol{\Omega} = \mathbb{E}_{\boldsymbol{x}}[\boldsymbol{\varphi}(\boldsymbol{x})\boldsymbol{\varphi}(\boldsymbol{x})^\top]$ is given by[12]:

$$\boldsymbol{\Omega} = b_0^2 \boldsymbol{1}_p \boldsymbol{1}_p^\top + b_1^2 \frac{\boldsymbol{W}_0 \boldsymbol{W}_0^\top}{d} + b_\star^2 \boldsymbol{I}_p + o_{\mathbb{P},d}(1) \tag{8.13}$$

where we have defined:

$$b_\star^2 = \sum_{\ell\geq 2} b_\ell^2 = \mathbb{E}_{z\sim\mathcal{N}(0,1)}\left[\sigma(z)^2\right] - b_0^2 - b_1^2 \tag{8.14}$$

This implies that under the proportional high-dimensional limit in Assumption 3, the features $\boldsymbol{\varphi}(\boldsymbol{x}) = \sigma(\boldsymbol{W}_0\boldsymbol{x})$ have the same first and second moments as the following Gaussian covariates:

$$\boldsymbol{g} = b_0 \boldsymbol{1}_p + b_1 \boldsymbol{W}_0 \boldsymbol{x} + b_\star \boldsymbol{z}, \qquad \boldsymbol{z} \sim \mathcal{N}(\boldsymbol{0}, \boldsymbol{I}_p) \tag{8.15}$$

It is not hard to show this equivalence also holds for higher moments. This suggests that in the proportional high-dimensional limit, we can trade the study of the original non-linear random features model in eq. (8.2) for the study of an equivalent Gaussian covariate model. This is an instance of a more general universality phenomenon, known as a Gaussian equivalence.

**Definition 1** (Gaussian equivalence principle)**.** Let $\mathcal{D} = \{(\boldsymbol{x}_i, y_i) \in \mathbb{R}^{d+1} : i \in [n]\}$ denote training data generated as follows:

$$y_i = f_\star(\boldsymbol{x}_i) + \varepsilon_i, \qquad \boldsymbol{x}_i \sim \mathcal{N}(\boldsymbol{0}, 1/d\boldsymbol{I}_d) \quad \text{independently} \tag{8.16}$$

And consider the following empirical risk minimisation problem:

$$\hat{\boldsymbol{a}}_\lambda(\boldsymbol{y}, \boldsymbol{\Phi}) = \underset{\boldsymbol{a}\in\mathbb{R}^p}{\operatorname{argmin}} \frac{1}{n}\sum_{i=1}^n \ell\left(y_i, \langle \boldsymbol{a}, \boldsymbol{\varphi}(\boldsymbol{x}_i)\rangle\right) + \lambda r(\boldsymbol{a}). \tag{8.17}$$

where $\boldsymbol{\varphi} : \mathbb{R}^d \to \mathbb{R}^p$ is a feature map, $\ell$ and $r$ convex loss and regularisation functions, respectively. Define the equivalent Gaussian covariate model:

$$\boldsymbol{g}_i \sim \mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Omega}), \qquad \text{with} \qquad \boldsymbol{\mu} = \mathbb{E}_{\boldsymbol{x}}[\boldsymbol{\varphi}(\boldsymbol{x})], \quad \boldsymbol{\Omega} = \mathrm{Cov}_{\boldsymbol{x}}(\boldsymbol{\varphi}(\boldsymbol{x})) \tag{8.18}$$

Then, we say a *Gaussian equivalence principle* (GEP) holds if under a high-dimensional asymptotics $d \to \infty$ the following hold:

$$\mathbb{E}_{\boldsymbol{X}}|R(\hat{\boldsymbol{a}}_\lambda(\boldsymbol{\Phi}, \boldsymbol{y})) - R(\hat{\boldsymbol{a}}_\lambda(\boldsymbol{G}, \boldsymbol{y}))| \to 0 \tag{8.19}$$

$$\mathbb{E}_{\boldsymbol{X}}|\hat{R}(\hat{\boldsymbol{a}}_\lambda(\boldsymbol{\Phi}, \boldsymbol{y}); \mathcal{D}) - \hat{R}(\hat{\boldsymbol{a}}_\lambda(\boldsymbol{G}, \boldsymbol{y}); \mathcal{D})| \to 0 \tag{8.20}$$

A few remarks about in place.

[12] Note this is normalised such that $\operatorname{Tr}\boldsymbol{\Omega} = \Theta(p)$

- As the name suggests, we view Gaussian equivalence as a principle rather than a theorem. The reason is that different Gaussian equivalence results holding for different hypotheses, loss functions, regularisations and scaling limits co-exist in the literature. Each has its particular assumptions and rather than discussing a problem-specific theorem, we take the broader viewpoint of Gaussian universality as a principle.

- Note that the GEP stated above goes beyond the equivalence of the population covariances discussed above. Indeed, the equivalence on the level of the errors is stronger than equivalence on the level of the population covariance. In particular, for ridge regression this also requires an equivalence on the level of the resolvent of the empirical covariance matrix.

- Note that the GEP implies that the empirical feature matrix decomposes in two pieces:

$$\Phi = b_0 \mathbf{1}_n \mathbf{1}_p^\top + b_1 \boldsymbol{X} \boldsymbol{W}_0^\top + b_\star \boldsymbol{Z} \tag{8.21}$$

  The first term is a just a rank-one spike due to the mean of the features. Assuming $n, p \geq d$, the second term consists of a rank $d$ matrix, and comes from the low-frequency components of the feature map. Finally, the last term is a random Gaussian matrix of rank $\min(n, p)$ and variance given by the remaining, high-frequency components of the feature map. Note this decomposition generalises to a polynomial scaling $n, d = \Theta(d^\ell)$, see Section 4.4 in (Misiakiewicz and Montanari, 2023) for a discussion.

## 8.4 High-dimensional asymptotics

To simplify the discussion, in this section we will make the following additional assumptions on the top of Assumption 1 and 2:

**Assumption 4.** Assume that:

- The activation function $\sigma$ has zero mean $b_0 = 0$. Note this holds for odd activation functions such as tanh.

- The target coefficients are i.i.d. Gaussian $\boldsymbol{\beta}_\star \sim \mathcal{N}(\mathbf{0}_d, \boldsymbol{I}_d)$.

The discussion in Section 8.2 and 8.3 suggests that in the proportional limit, the random features ridge regression defined in eq. (8.2) is equivalent to ridge regression on the following Gaussian covariate model:

$$\min_{\boldsymbol{a} \in \mathbb{R}^p} \frac{1}{2n} \sum_{i=1}^n \left( \langle \boldsymbol{\beta}_\star, \boldsymbol{x}_i \rangle + \varepsilon_i - \langle \boldsymbol{a}, \boldsymbol{g}_i \rangle \right)^2 + \frac{\lambda}{2} ||\boldsymbol{a}||_2^2. \tag{8.22}$$

where $(\boldsymbol{x}, \boldsymbol{g}) \in \mathbb{R}^{d+p}$ are jointly Gaussian vectors:

$$(\boldsymbol{x}, \boldsymbol{g}) \sim \mathcal{N}\left( \mathbf{0}_{d+p}, \begin{bmatrix} 1/d \boldsymbol{I}_d & 1/\sqrt{dp} \boldsymbol{\Phi}^\top \\ 1/\sqrt{dp} \boldsymbol{\Phi} & 1/p \boldsymbol{\Omega} \end{bmatrix} \right) \tag{8.23}$$

with covariances given by:

$$\boldsymbol{\Phi} = b_0 \boldsymbol{W}_0 \in \mathbb{R}^{p \times d}, \qquad \boldsymbol{\Omega} = b_1^2 \frac{\boldsymbol{W}_0 \boldsymbol{W}_0^\top}{d} + b_\star^2 \boldsymbol{I}_p \in \mathbb{R}^{p \times p} \tag{8.24}$$

This is a particular case of the model introduced in Section 3! Therefore, we can readily apply the equations we have derived for the Gaussian covariate model to derive the asymptotic excess risk for the random features model. This is given by:

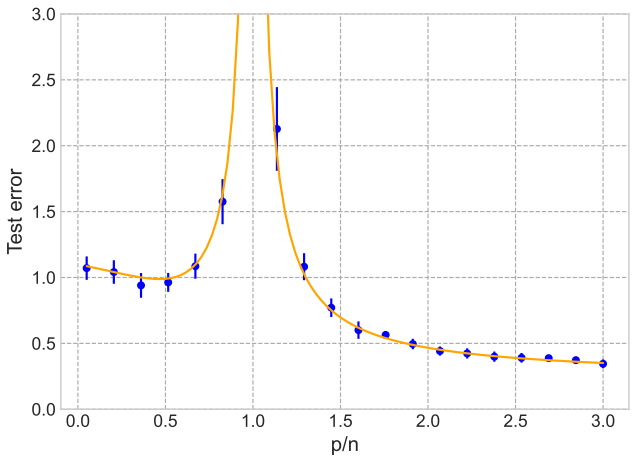


Figure 10: Test error of the random features ridge regressor as function the the number of the normalised width $p/n$ at fixed $n/d = 1.5$ and $\lambda = 0^+$. The solid line denote the theoretical result obtained from iterating the self-consistent eq. (8.29), and points denote finite-size simulations with $d = 500$.

**Result 1** (Asymptotic error for RFRR)**.** Consider the random features ridge regression problem in eq. (8.22) under Gaussian design (1). Then, under Assumption 4, the asymptotic excess risk in the limit $n, p, d \to \infty$ at fixed rates $\alpha = n/d$ and $\gamma = p/d$ is given by:

$$\lim_{d\to\infty} R(\hat{\boldsymbol{a}}_\lambda) - \sigma^2 = \mathcal{B}(\alpha, \gamma, \lambda) + \mathcal{V}(\alpha, \gamma, \lambda) \tag{8.25}$$

where the limiting *bias* $\mathcal{B}$ and *variance* $\mathcal{V}$ are given by:

$$\begin{aligned}
\mathcal{B}(\alpha, \gamma, \lambda) &= \frac{n/d\, \mathrm{Tr}\left\{ \left(b_\star^2 + \nu(\nu + b_\star^2)\boldsymbol{R}(\nu + b_\star^2; b_1^2/d\boldsymbol{W}_0^\top \boldsymbol{W}_0)\right) \boldsymbol{R}(\nu + b_\star^2; b_1^2/d\boldsymbol{W}_0^\top \boldsymbol{W}_0)\right\}}{n - \mathrm{df}_2(\nu + b_\star^2; b_1^2/d\boldsymbol{W}_0\boldsymbol{W}_0^\top)} \\
\mathcal{V}(\alpha, \gamma, \lambda) &= \sigma^2 \frac{\mathrm{df}_2(\nu + b_\star^2; b_1^2/d\boldsymbol{W}_0\boldsymbol{W}_0^\top)}{n - \mathrm{df}_2(\nu + b_\star^2; b_1^2/d\boldsymbol{W}_0\boldsymbol{W}_0^\top)}
\end{aligned} \tag{8.26}$$

where $\boldsymbol{R}$ is the resolvent matrix:

$$\boldsymbol{R}(\nu; \boldsymbol{A}) = (\nu + \boldsymbol{A})^{-1} \tag{8.27}$$

$\mathrm{df}_\alpha$ are the degrees of freedom, also known as effective dimensions:

$$\mathrm{df}_\alpha(\nu; \boldsymbol{A}) = \mathrm{Tr}\, \boldsymbol{A}^\alpha (\nu + \boldsymbol{A})^{-\alpha}, \qquad \alpha \in \{1, 2\} \tag{8.28}$$

and $\nu$ is the solution of the following self-consistent equation:

$$n - \frac{p\lambda}{\nu} = \mathrm{df}_1(\nu + b_\star^2; b_1^2/d\boldsymbol{W}_0\boldsymbol{W}_0^\top) \tag{8.29}$$

See appendix E.3 for the derivation from the GCM. This result allow us to fully characterise the risk as a function of the random features activation function $\sigma$ (which define $b_1, b_\star$) and the first-layer weights $\boldsymbol{W}_0 \in \mathbb{R}^{p\times d}$. In particular, if $1/d\boldsymbol{W}_0\boldsymbol{W}_0^\top$ admits an asymptotic spectral distribution $\mu$, all the traces above can be written as one-dimensional integrals with respect to $\mu$. Figure 10 illustrates the test error of the random features interpolator as a function of the number of the network width, obtained from eq. (8.26). For $p/n < 1$, we observe the U-shapped curved from the classical bias-variance tradeoff. Note the interpolating peak at $n = p$. Differently from the well-specified ridge regression case discussed in section 7, the error keeps decreasing beyond the interpolation point $p/n > 1$. This is known as the *double descent* behaviour (Belkin et al., 2019).

### 8.5 To go further

**Beyond proportional regime** — A GEP for the spectrum of kernel matrices in the power-law scaling was proven in (Dubova et al., 2023; Lu and Yau, 2022). Similarly, universality results for the performance of kernel ridge regression in the polynomial scaling appeared (Canatar et al., 2021; Xiao et al., 2023; Misiakiewicz and Saeed, 2024), for support vector machines in Opper and Urbanczik (2001); Dietrich et al. (1999) and for random features ridge regression in (Hu et al., 2024; Aguirre-López et al., 2024)

**Deep random features** — Fan and Wang (2020) provided a characterisation of the resolvent of deep random feature maps with i.i.d. Gaussian weights in the proportional regime. A GEP for the error was proven by (Schröder et al., 2023; Bosch et al., 2023a), and extended to correlated weights by Schröder et al. (2024).

**Mixture distributions** — Refinetti et al. (2021) derived a GEP for random features on Gaussian mixture distribution. Dandi et al. (2024) proved a GEP for mixture models.

**Limitations of universality** — Finally, several authors have discussed the limitations of GEP in different contexts, see (Gerace et al., 2024; Pesce et al., 2023; Tomasini et al., 2022; Cheng et al., 2024)

### Bibliographical notes

- Universality is a rich topic that crosses different disciplines, such as Statistical Physics (Marinari et al., 1994; Parisi and Potters, 1995; Parisi and Rizzo, 2010; Franz et al., 2017), Random Matrix Theory (Johansson, 2001; Ben Arous and Péché, 2005; Tao and Vu, 2011; Erdős et al., 2012), Signal Processing (Donoho and Tanner, 2009; Korada and Montanari, 2011) and Statistics (Panahi and Hassibi, 2017; Montanari and Nguyen, 2017; Abbasi et al., 2019).
- A precursor of Gaussian equivalence for the random features model is (Karoui, 2010), who showed that in the proportional asymptotics the spectrum of a random kernel matrix is equivalent to the spectrum of a shifted linear kernel. Other hints also appeared in (Pennington and Worah, 2017; Louart et al., 2018), where the resolvent of the empirical random features covariance was characterised, also under the proportional limit.
- The GEP for the RF model, as formulated in Definition 1, appeared around the same time in (Mei and Montanari, 2022; Goldt et al., 2020). Mei and Montanari (2022) provided a rigorous random matrix proof for random features ridge regression in the proportional asymptotics. This discussion was generalised by Gerace et al. (2020) to general convex loss functions. A central limit for one-dimensional projections of RF was proven in (Goldt et al., 2022; Hu and Lu, 2022). Hu and Lu (2022) combined this result with a Lindeberg argument to establish the universality of the test and training errors under general convex losses. This was extended to general convex regularisers by Bosch et al. (2023b). A GEP for general feature maps was formulated by Loureiro et al. (2021), who also studied its validity in real data. Montanari and Saeed (2022) proposed a refined interpolation scheme and proved universality for sub-Gaussian and NTK features.

# Appendix

# A Basics of Hermite polynomials

## A.1 Scalar Hermite polynomials

Let $\gamma : \mathbb{R} \to \mathbb{R}$ denote the standard normal probability density function (pdf):

$$\gamma(z) = \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}z^2} \tag{A.1}$$

We denote by $L^2(\mathbb{R}, \gamma)$ the space of functions which are square-integrable with respect to $\gamma$:

$$L^2(\mathbb{R}, \gamma) := \left\{ f : \mathbb{R} \to \mathbb{R} : \mathbb{E}_{Z\sim\gamma}\left[|f(Z)|^2\right] = \int_{\mathbb{R}} \gamma(\mathrm{d}z)|f(z)|^2 < \infty \right\} \tag{A.2}$$

Note that this coincides with the space of real functions which have finite-second moment with respect to the Gaussian distribution. Recall that $L^2(\mathbb{R}, \gamma)$ defines a separable[13] Hilbert space with inner-product given by:

$$\langle f, g\rangle_\gamma = \mathbb{E}_{Z\sim\gamma}\left[f(Z)g(Z)\right] = \int_{\mathbb{R}} \gamma(\mathrm{d}z) f(z)g(z) \tag{A.3}$$

As a separable Hilbert space, $L^2(\mathbb{R}, \gamma)$ admits a countable basis. There are many possible choices of bases, and the motivation behind the construction of Hermite polynomials is precisely to build a "convenient" basis for $L^2(\mathbb{R}, \gamma)$, given by polynomials which are orthogonal with respect to the inner-product $\langle \cdot, \cdot \rangle_\gamma$.

**Definition 2** (Hermite polynomials)**.** The Hermite polynomials $\mathrm{He}_j(z)$, $j \geq 0$ are a real-valued family of polynomials defined by:

$$\mathrm{He}_j(z) := \frac{(-1)^j}{\gamma(z)} \frac{\mathrm{d}^j}{\mathrm{d}z^j} \gamma(z). \tag{A.4}$$

Moreover, the family $(\mathrm{He}_j)_{j\geq 0}$ satisfy the following useful properties:

(a) **Degree:** $\mathrm{He}_j(z)$ is a polynomial of degree $j$.

(b) **Orthogonality:** For any $j, k \geq 0$, we have:

$$\langle \mathrm{He}_j, \mathrm{He}_k\rangle_\gamma = \mathbb{E}_{Z\sim\gamma}\left[\mathrm{He}_j(Z)\mathrm{He}_k(Z)\right] = j!\delta_{jk} \tag{A.5}$$

(c) **Correlations:** Let $Z, Z' \sim \mathcal{N}(0, 1)$ with $\mathbb{E}[ZZ'] = \rho \in [-1, 1]$. Then:

$$\mathbb{E}[\mathrm{He}_j(Z)\mathrm{He}_k(Z')] = j!\delta_{jk}\rho^j \tag{A.6}$$

(d) **Derivatives:**

$$\frac{\mathrm{d}}{\mathrm{d}z}\mathrm{He}_j(z) = j\mathrm{He}_{j-1}(z) \tag{A.7}$$

(e) **Completeness:** Any $f \in L^2(\mathbb{R}, \gamma)$ can be decomposed as a sum of Hermite polynomials:

$$f(z) = \sum_{j\geq 0} \frac{\hat{f}_j}{j!}\mathrm{He}_j(z), \qquad \hat{f}_j := \langle f, \mathrm{He}_j\rangle_\gamma \tag{A.8}$$

[13] A separable Hilbert space is a Hilbert space that has a countable basis.

The Hermite polynomials can be constructed explicitly from the monomials $(z^j)_{j\geq 0}$ by Gram-Schmidt orthogonalisation, see Section 11.2 of O'Donnell (2014) for a detailed discussion. ⚠ Different conventions for the Hermite polynomials co-exist in the literature. The one we adopt in Definition 2, which is the most commonly employed in the machine learning literature, is known as the *probabilist*'s Hermite polynomial, in contrast to the *physicist*'s Hermite polynomials commonly found in the Physics literature.

Note that $(\mathrm{He}_j)_{j\geq 0}$ are not normalised. Indeed, as a particular case of property (b) we have:

$$||\mathrm{He}_j||_\gamma^2 = \mathbb{E}_{Z\sim\gamma}[|\mathrm{He}_j(Z)|^2] = j!$$

It is therefore convenient to define the normalised Hermite polynomials $(h_j)_{j\geq 0}$:

$$h_j(z) \coloneqq \frac{\mathrm{He}_j(z)}{\sqrt{j!}} \tag{A.9}$$

such that $\langle h_j, h_k\rangle_\gamma = \delta_{jk}$.

⚠ Note that the decomposition of $f \in L^2(\mathbb{R},\gamma)$ in the orthonormal basis $(h_j)_{j\geq 0}$ reads $f(z) = \sum_{j\geq 0}\mu_j h_j(z)$. It is important to keep in mind that $\hat{f}_j = \langle f, \mathrm{He}_j\rangle_\gamma = \sqrt{j!}\,\langle f, h_j\rangle_\gamma$. This is often a source of confusion. From the completeness relation in eq. (A.8), it is easy to show that:

**Lemma 2** (Plancharel Formula)**.** Let $f, g \in L^2(\mathbb{R},\gamma)$. Then:

$$\langle f, g\rangle_\gamma = \sum_{j=1}^{d} \hat{f}_j \hat{g}_j \tag{A.10}$$

## A.2 Multi-variate Hermite polynomials

The discussion above can be generalised to higher dimensions. Let $\gamma_d$ denote the multi-variate Gaussian pdf:

$$\gamma_d(\boldsymbol{z}) = \frac{1}{(2\pi)^{d/2}} e^{-\frac{1}{2}\|\boldsymbol{z}\|_2^2} \tag{A.11}$$

And as before define:

$$L^2(\mathbb{R}^d,\gamma_d) \coloneqq \left\{ f : \mathbb{R}^d \to \mathbb{R} : \mathbb{E}_{\boldsymbol{Z}\sim\gamma_d}\left[|f(\boldsymbol{Z})|^2\right] = \int_{\mathbb{R}^d} \gamma_d(\mathrm{d}\boldsymbol{z})|f(\boldsymbol{z})|^2 < \infty \right\} \tag{A.12}$$

The straightforward construction of an orthonormal polynomial basis for $L^2(\mathbb{R}^d,\gamma_d)$ is given by taking products of (normalised) Hermite polynomials $h_j$. For instance, for any multi-index $\boldsymbol{\alpha} \in \mathbb{N}^d$, we can define:

$$h_{\boldsymbol{\alpha}}(\boldsymbol{z}) = \prod_{i=1}^{d} h_{\alpha_i}(z_i) \tag{A.13}$$

It is easy to see that $(h_{\boldsymbol{\alpha}}(\boldsymbol{z}))_{\boldsymbol{\alpha}\in\mathbb{N}^d}$ is an orthonormal family, i.e. for any $\boldsymbol{\alpha},\boldsymbol{\beta} \in \mathbb{N}^d$:

$$\langle h_{\boldsymbol{\alpha}}, h_{\boldsymbol{\beta}}\rangle_{\gamma_d} = \mathbb{E}_{\boldsymbol{Z}\sim\gamma_d}[h_{\boldsymbol{\alpha}}(\boldsymbol{Z})h_{\boldsymbol{\beta}}(\boldsymbol{Z})] \overset{(a)}{=} \prod_{j=1}^{d} \mathbb{E}_{Z_j\sim\gamma}[h_{\alpha_j}(Z_j)h_{\beta_j}(Z_j)] = \prod_{j=1}^{d} \delta_{\alpha_j\beta_j} \tag{A.14}$$

where in (a) we used the fact that $Z_j \sim \gamma$ are i.i.d. to bring the product outside the expectation. Therefore, the family $(h_{\boldsymbol{\alpha}}(\boldsymbol{z}))_{\boldsymbol{\alpha}\in\mathbb{N}^d}$ form an orthonormal basis of $L^2(\mathbb{R}^d,\gamma_d)$, i.e. any $f \in L^2(\mathbb{R}^d,\gamma_d)$ can be written as:

$$f(\boldsymbol{z}) = \sum_{\boldsymbol{\alpha}\in\mathbb{N}^d} \hat{f}_{\boldsymbol{\alpha}} h_{\boldsymbol{\alpha}}(\boldsymbol{z}), \qquad \hat{f}_{\boldsymbol{\alpha}} \coloneqq \langle f, h_{\boldsymbol{\alpha}}\rangle_{\gamma_d} \tag{A.15}$$

**Example 4.** As a concrete illustration, let's look at $d = 2$. Letting $\boldsymbol{\alpha} = (\alpha_1, \alpha_2) \in \mathbb{N}^2$, we have the expansion of $f \in L^2(\mathbb{R}^2, \gamma_2)$:

$$\begin{aligned} f(z_1, z_2) &= \sum_{(\alpha_1,\alpha_2)\in\mathbb{N}^2} \hat{f}_{\alpha_1,\alpha_2} h_{\alpha_1,\alpha_2}(z_1, z_2) = \sum_{\alpha_1\geq 0}\sum_{\alpha_2\geq 0} \hat{f}_{\alpha_1\alpha_2} h_{\alpha_1}(z_1) h_{\alpha_2}(z_2) \\ &= \hat{f}_{00} + \sum_{\alpha_1\geq 1} \hat{f}_{\alpha_1 0} h_{\alpha_1}(z_1) + \sum_{\alpha_2\geq 1} \hat{f}_{0\alpha_2} h_{\alpha_2}(z_2) + \sum_{\alpha_1\geq 1}\sum_{\alpha_2\geq 1} \hat{f}_{\alpha_1\alpha_2} h_{\alpha_1}(z_1) h_{\alpha_2} h_{\alpha_2}(z_2) \\ &= \hat{f}_{00} + \hat{f}_{01} z_1 + \hat{f}_{10} z_2 + \hat{f}_{11} z_1 z_2 + \hat{f}_{20} \frac{z_1^2 - 1}{\sqrt{2}} + \hat{f}_{02} \frac{z_2^2 - 1}{\sqrt{2}} + \cdots \end{aligned}$$

The definition in eq. (A.13) is not the unique way to lift $h_j$ into an orthonormal basis of $L^2(\mathbb{R}^d, \gamma_d)$. Indeed, there are as many ways of constructing this extension as there are bases of $\mathbb{R}^d$. To see this, let $(\boldsymbol{u}_i)_{i\in[d]}$ denote an orthonormal basis of $\mathbb{R}^d$. By construction, the matrix $\boldsymbol{U} \in \mathbb{R}^{d\times d}$ obtained by concatenating $(\boldsymbol{u}_j)_{i\in[d]}$ row-wise defines an orthogonal matrix $\boldsymbol{U}^\top \boldsymbol{U} = \boldsymbol{I}_d$, and for any $\boldsymbol{z} \in \mathbb{R}^d$ we can write:

$$\boldsymbol{z} = \sum_{i=1}^{d} \langle \boldsymbol{u}_i, \boldsymbol{z}\rangle \boldsymbol{u}_i. \tag{A.16}$$

With this in mind, we can define a general lifting of $(h_j)_{j\geq 0}$:

**Definition 3** (Multi-variate Hermite polynomials)**.** Let $\boldsymbol{U} \in \mathcal{O}(d)$ denote an orthogonal matrix. For a multi-index $\boldsymbol{\alpha} \in \mathbb{N}^d$, we define the (normalised) multi-variate Hermite polynomial $h_{\boldsymbol{\alpha}}(U) \in L^2(\mathbb{R}^d, \gamma_d)$ with respect to $\boldsymbol{U}$ as:

$$h_{\boldsymbol{\alpha}}(\boldsymbol{U})(\boldsymbol{z}) = \prod_{i=1}^{d} h_{\alpha_i}\left(\langle \boldsymbol{u}_i, \boldsymbol{z}\rangle\right) \tag{A.17}$$

It satisfies the following useful properties:

(a) **Degree:** $h_{\boldsymbol{\alpha}}(\boldsymbol{U})(\boldsymbol{z})$ is a polynomial of degree $|\boldsymbol{\alpha}| = \sum_{i=1}^{d} \alpha_i$.

(b) **Orthogonality:** For any $\boldsymbol{\alpha}, \boldsymbol{\beta} \in \mathbb{N}^d$, we have:

$$\langle h_{\boldsymbol{\alpha}}(\boldsymbol{U}), h_{\boldsymbol{\beta}}(\boldsymbol{U})\rangle_{\gamma_d} = \delta_{\boldsymbol{\alpha}\boldsymbol{\beta}} \coloneqq \prod_{i=1}^{d} \delta_{\alpha_i\beta_j} \tag{A.18}$$

(c) **Completness:** Any $f \in L^2(\mathbb{R}^d, \gamma)$ can be decomposed :

$$f(\boldsymbol{z}) = \sum_{\boldsymbol{\alpha}\in\mathbb{N}^d} \hat{f}_j h_{\boldsymbol{\alpha}}(\boldsymbol{U})(\boldsymbol{z}), \qquad \hat{f}_{\boldsymbol{\alpha}} \coloneqq \langle f, h_{\boldsymbol{\alpha}}(\boldsymbol{U})\rangle_{\gamma_d} \tag{A.19}$$

It is easy to see that when $\boldsymbol{U} = \boldsymbol{I}_d$, i.e. $\boldsymbol{u}_i = \boldsymbol{e}_i$ is the canonical basis, this definition reduces to eq. (A.13).

# B Useful Matrix identities

Let $\boldsymbol{U} \in \mathbb{R}^{n\times d}$ and $\boldsymbol{V} \in \mathbb{R}^{d\times n}$ be two retangular matrices. We have the following useful identities:

- The traces of the resolvent and co-resolvent are related as:

$$\operatorname{Tr}(\boldsymbol{U}\boldsymbol{V} - z\boldsymbol{I}_n)^{-1} = \operatorname{Tr}(\boldsymbol{V}\boldsymbol{U} - z\boldsymbol{I}_d)^{-1} - \frac{n-d}{z} \tag{B.1}$$

  Taking the derivative with respect to $z$ on both sides, this also implies:

$$\operatorname{Tr}(\boldsymbol{U}\boldsymbol{V} - z\boldsymbol{I}_n)^{-2} = \operatorname{Tr}(\boldsymbol{V}\boldsymbol{U} - z\boldsymbol{I}_d)^{-2} - \frac{n-d}{z^2} \tag{B.2}$$

- Push-through identity:

$$(\boldsymbol{U}\boldsymbol{V} - z\boldsymbol{I}_n)^{-1}\boldsymbol{U} = \boldsymbol{U}(\boldsymbol{V}\boldsymbol{U} - z\boldsymbol{I}_d)^{-1} \tag{B.3}$$

- Block inversion formula:

$$\begin{bmatrix} \boldsymbol{A} & \boldsymbol{B} \\ \boldsymbol{C} & \boldsymbol{D} \end{bmatrix}^{-1} = \begin{bmatrix} (\boldsymbol{A} - \boldsymbol{B}\boldsymbol{D}^{-1}\boldsymbol{C})^{-1} & -(\boldsymbol{A} - \boldsymbol{B}\boldsymbol{D}^{-1}\boldsymbol{C})^{-1}\boldsymbol{B}\boldsymbol{D}^{-1} \\ -\boldsymbol{D}^{-1}\boldsymbol{C}(\boldsymbol{A} - \boldsymbol{B}\boldsymbol{D}^{-1}\boldsymbol{C})^{-1} & \boldsymbol{D}^{-1} + \boldsymbol{D}^{-1}\boldsymbol{C}(\boldsymbol{A} - \boldsymbol{B}\boldsymbol{D}^{-1}\boldsymbol{C})^{-1}\boldsymbol{B}\boldsymbol{D}^{-1} \end{bmatrix} \tag{B.4}$$

  where $\boldsymbol{A}^{n\times n}$, $\boldsymbol{B} \in \mathbb{R}^{n\times m}$, $\boldsymbol{C} \in \mathbb{R}^{m\times n}$ and $\boldsymbol{D} \in \mathbb{R}^{m\times m}$.

# C Random Matrix Theory

In this Appendix, we recall some useful notions and results from random matrix theory. We start by defining the following useful notions:

**Definition 4** (Resolvent)**.** Let $\boldsymbol{A} \in \mathbb{R}^{n\times n}$ denote a symmetric matrix with eigenvalues $\operatorname{spec}(\boldsymbol{A}) = \{\lambda_1, \dots, \lambda_n\}$. The *resolvent* of $\boldsymbol{A}$ is defined as:

$$\boldsymbol{R}(z; \boldsymbol{A}) = (\boldsymbol{A} - z\boldsymbol{I}_n)^{-1} \in \mathbb{R}^{n\times n}, \qquad z \in \mathbb{C} - \operatorname{spec}(\boldsymbol{A}). \tag{C.1}$$

**Definition 5** (Empirical spectral measure)**.** Let $\boldsymbol{A} \in \mathbb{R}^{n\times n}$ denote a symmetric matrix with eigenvalues $\operatorname{spec}(\boldsymbol{A}) = \{\lambda_1, \dots, \lambda_n\}$. We define its empirical measure:

$$\hat{\mu}_n(\lambda; \boldsymbol{A}) = \frac{1}{n}\sum_{i=1}^{n} \delta(\lambda - \lambda_i) \tag{C.2}$$

Note this is the normalised counting measure of $\operatorname{spec}(\boldsymbol{A})$. Moreover, note that by construction it is normalised $\int_{\mathbb{R}} \hat{\mu}_n(\mathrm{d}\lambda) = 1$, hence it is a probability measure.

**Definition 6** (Stieltjes transform)**.** Let $\mu$ denote a finite real measure with support $\operatorname{supp}(\mu) \subset \mathbb{R}$. We define its *Stieltjes transform*:

$$s(z; \mu) = \int_{\mathbb{R}} \frac{\mu(\mathrm{d}t)}{t - z}, \qquad z \in \mathbb{C} - \operatorname{supp}(\mu). \tag{C.3}$$

Note that if $\mu = \hat{\mu}_n(\cdot; \boldsymbol{A})$ is the empirical spectral measure associated to a real symmetric matrix $\boldsymbol{A} \in \mathbb{R}^{n\times n}$, we will some times denote $s(z; \hat{\mu}_n) = s_n(z; \boldsymbol{A})$ and refer to $\hat{s}_n$ as the "Stieltjes transform of the matrix $\boldsymbol{A}$", which can also be written as:

$$s_n(z; \boldsymbol{A}) = \frac{1}{n}\operatorname{Tr}\boldsymbol{R}(z; \boldsymbol{A}) = \frac{1}{n}\operatorname{Tr}\left(\boldsymbol{A} - z\boldsymbol{I}_n\right)^{-1}, \qquad z \in \mathbb{C} - \operatorname{spec}(\boldsymbol{A}), \tag{C.4}$$

where $\boldsymbol{R}(z; \boldsymbol{A})$ is the resolvent of $\boldsymbol{A}$. The Stieltjes transform satisfy the following useful properties:

- $$|s(z;\mu)| \leq \frac{1}{\operatorname{dist}(z, \operatorname{supp}(\mu))} \tag{C.5}$$

  In particular, if $\operatorname{supp}(\mu)$ is bounded, then:

$$\lim_{x \to \pm\infty} s(x;\mu) = 0 \tag{C.6}$$

- $$\operatorname{Im}(z) \operatorname{Im}(s(z;\mu)) \geq 0 \tag{C.7}$$

- The restriction of $s(z;\mu)$ to the real axis $\mathbb{R} - \operatorname{supp}(\mu)$ is an increasing function in all the connected components of $\operatorname{supp}(\mu)$.
- If $\mu$ admits a density $f$ at $x$:

$$f(x) = \frac{1}{\pi} \lim_{\varepsilon \to 0^+} \operatorname{Im}[s(x + i\varepsilon;\mu)] \tag{C.8}$$

**Definition 7** (Degrees-of-freedom)**.** Let $\boldsymbol{A} \in \mathbb{R}^{n \times n}$ denote a real symmetric matrix with $\operatorname{spec}(\boldsymbol{A}) = \{\lambda_1, \ldots, \lambda_n\}$. We define the degrees-of-freedom:

$$\operatorname{df}_\alpha(\lambda; \boldsymbol{A}) = \operatorname{Tr}\{\boldsymbol{A}^\alpha (\boldsymbol{A} + \lambda \boldsymbol{I}_n)^{-\alpha}\} = \sum_{i=1}^{n} \left( \frac{\lambda_i}{\lambda_i + \lambda} \right)^\alpha, \qquad \alpha \in \{1, 2\} \tag{C.9}$$

The degrees-of-freedom $\operatorname{df}_\alpha(\lambda; \boldsymbol{A})$ provide a notion of an effective dimension for the matrix $\boldsymbol{A}$, as justified by the following properties:

- $\operatorname{df}_\alpha(\lambda; \boldsymbol{A})$ are strictly decreasing functions of $\lambda$ with $\operatorname{df}_\alpha(0; \boldsymbol{A}) = n$.
- The following bound hold:

$$0 \leq \operatorname{df}_2(\lambda; \boldsymbol{A}) \leq \operatorname{df}_1(\lambda; \boldsymbol{A}) \leq n \tag{C.10}$$

- $\operatorname{df}_\alpha$ can be related to the restriction of the Stieltjes transform $s_n(z; \boldsymbol{A}) = {}^1\!/_n \operatorname{Tr}(\boldsymbol{A} - zI_n)^{-1}$ on the negative real axis $z = -\lambda \in \mathbb{R}_+ - \operatorname{spec}(\boldsymbol{A})$ as follows:

$$\frac{1}{p} \operatorname{df}_1(\lambda; \boldsymbol{A}) = 1 - \lambda s_n(-\lambda) \tag{C.11}$$

$$\tag{C.12}$$

**Definition 8** (Deterministic equivalent)**.** Let $\boldsymbol{M} \in \mathbb{R}^{n \times n}$ denote a symmetric random matrix. We say $\bar{M} \in \mathbb{R}^{n \times n}$ is a deterministic equivalent for $\boldsymbol{M}$ if for sequences of deterministic matrices $\boldsymbol{A} \in \mathbb{R}^{n \times n}$ with $||\boldsymbol{A}||_{op} = 1$ and unit vectors $\boldsymbol{u}, \boldsymbol{v} \in \mathbb{S}^{n-1}$ we have:

$$\frac{1}{n} \operatorname{Tr} \boldsymbol{A}(\boldsymbol{M} - \bar{\boldsymbol{M}}) \to 0, \qquad \boldsymbol{u}^\top (\boldsymbol{M} - \bar{\boldsymbol{M}}) \boldsymbol{v} \to 0, \tag{C.13}$$

as $n \to \infty$, where convergence can be almost surely or in probability. Often, we will denote $\boldsymbol{M} \sim \bar{\boldsymbol{M}}$ to say that $\bar{\boldsymbol{M}}$ is a deterministic equivalent for the random matrix $\boldsymbol{M}$.

## C.1 Wishart matrices

We now review some classical RMT results for Wishart matrices.

**Theorem 2** (Silverstein and Bai (1995))**.** Let $\boldsymbol{\Sigma}$ denote a symmetric positive semi-definite matrix with bounded operator norm, and define $\boldsymbol{X} = \boldsymbol{Z}\boldsymbol{\Sigma}^{1/2} \in \mathbb{R}^{n\times d}$ with $\boldsymbol{Z} \in \mathbb{R}^{n\times d}$ a matrix with i.i.d. sub-Gaussian entries with zero mean and unit variance. In the proportional asymptotic limit where $n, d \to \infty$ at fixed ratio $c = {}^{d}/_{n}$, the following deterministic equivalents hold:

$$\left({}^{1}/_{n}\boldsymbol{X}^\top\boldsymbol{X} - z\boldsymbol{I}_d\right)^{-1} \sim -\frac{1}{z}\left(\boldsymbol{I}_d + \tilde{s}_d(z)\boldsymbol{\Sigma}\right)^{-1} \tag{C.14}$$

$$\left({}^{1}/_{n}\boldsymbol{X}\boldsymbol{X}^\top - z\boldsymbol{I}_n\right)^{-1} \sim \tilde{s}_d(z)\boldsymbol{I}_n \tag{C.15}$$

where $(z, \tilde{s}_d(z))$ are the unique solution of the following self-consistent equation:

$$\frac{1}{\tilde{s}_d} + z = \frac{1}{n}\operatorname{Tr}\boldsymbol{\Sigma}(\boldsymbol{I}_d + \tilde{s}_d\boldsymbol{\Sigma})^{-1} \tag{C.16}$$

In particular, if $\boldsymbol{\Sigma}$ admits an asymptotic spectral measure $\nu$ as $d \to \infty$, then:

$$\hat{\mu}_d(\cdot; {}^{1}/_{n}\boldsymbol{X}^\top\boldsymbol{X}) \to \mu, \qquad \hat{\mu}_n(\cdot; {}^{1}/_{n}\boldsymbol{X}\boldsymbol{X}^\top) \to \tilde{\mu} \tag{C.17}$$

with associated Stieltjes transform $s(z) = s(z;\mu)$ and $\tilde{s} = s(z;\tilde{s})$ satisfying:

$$s(z) = \frac{1}{c}\tilde{s}(z) + \frac{1-c}{cz}, \qquad \frac{1}{\tilde{s}(z)} + z = c\int \nu(\mathrm{d}t)\frac{t}{1+t\tilde{s}(z)} \tag{C.18}$$

*Remark* 10. In machine learning, we will often be interested in the asymptotic equivalent of quantities of the type $\operatorname{Tr}\left(\hat{\boldsymbol{\Sigma}}_n + \lambda\boldsymbol{I}_d\right)^{-1}$ and $\operatorname{Tr}\hat{\boldsymbol{\Sigma}}_n(\hat{\boldsymbol{\Sigma}}_n + \lambda\boldsymbol{I}_d)^{-1}$, where $\lambda \geq 0$ and $\hat{\boldsymbol{\Sigma}}_n \coloneqq {}^{1}/_{n}\boldsymbol{X}^\top\boldsymbol{X}$ is the empirical covariance matrix. We can translate the result from Equation (C.16) to this case:

$$\operatorname{Tr}\hat{\boldsymbol{\Sigma}}_n(\hat{\boldsymbol{\Sigma}}_n + \lambda\boldsymbol{I}_d)^{-1} \sim \operatorname{Tr}\boldsymbol{\Sigma}(\boldsymbol{\Sigma} + \kappa(\lambda)\boldsymbol{I}_d)^{-1} \tag{C.19}$$

where we have evaluated eq. (C.16) at $z = -\lambda$ and defined:

$$\kappa(\lambda) = \frac{1}{\tilde{s}(-\lambda)}. \tag{C.20}$$

Therefore, $\kappa : \mathbb{R}_+ \to \mathbb{R}_+$ is an increasing function satisfying the following self-consistent equation:

$$1 - \frac{\lambda}{\kappa} = \frac{1}{n}\operatorname{Tr}\boldsymbol{\Sigma}(\boldsymbol{\Sigma} + \kappa\boldsymbol{I}_d)^{-1} = \frac{1}{n}\mathrm{df}_1(\kappa;\boldsymbol{\Sigma}) \tag{C.21}$$

**Corollary 1** (Other equivalents)**.** Other useful deterministic equivalents can be obtained from Theorem 2 by differentiation, see Bach (2024) for a derivation. Here, we list two that will be used later.

- $$\operatorname{Tr}\hat{\boldsymbol{\Sigma}}_n(\hat{\boldsymbol{\Sigma}}_n - z\boldsymbol{I}_d)^{-2} \sim \frac{n\operatorname{Tr}\boldsymbol{\Sigma}\left(\boldsymbol{\Sigma} + \frac{1}{\tilde{s}_d(z)}\boldsymbol{I}_d\right)^{-2}}{n - \operatorname{Tr}\boldsymbol{\Sigma}^2\left(\boldsymbol{\Sigma} + \frac{1}{\tilde{s}_d(z)}\boldsymbol{I}_d\right)^{-2}} \tag{C.22}$$

- $$\operatorname{Tr}\hat{\boldsymbol{\Sigma}}_n^2(\hat{\boldsymbol{\Sigma}}_n - z\boldsymbol{I}_d)^{-2} \sim \operatorname{Tr}\boldsymbol{\Sigma}^2\left(\boldsymbol{\Sigma} + \frac{1}{\tilde{s}_d(z)}\boldsymbol{I}_d\right)^{-2} + \frac{1}{\tilde{s}_d(z)^2}\frac{\left(\operatorname{Tr}\boldsymbol{\Sigma}\left(\boldsymbol{\Sigma} + \frac{1}{\tilde{s}_d(z)}\boldsymbol{I}_d\right)^{-2}\right)^2}{n - \operatorname{Tr}\boldsymbol{\Sigma}^2\left(\boldsymbol{\Sigma} + \frac{1}{\tilde{s}_d(z)}\boldsymbol{I}_d\right)^{-2}} \tag{C.23}$$

$$\operatorname{Tr}\left(\hat{\boldsymbol{\Sigma}}_n - z\boldsymbol{I}_d\right)^{-2} \sim \frac{1}{z}\frac{n\operatorname{Tr}\boldsymbol{\Sigma}\left(\boldsymbol{\Sigma} + \frac{1}{\tilde{s}_d(z)}\boldsymbol{I}_d\right)^{-2}}{n - \operatorname{Tr}\boldsymbol{\Sigma}^2\left(\boldsymbol{\Sigma} + \frac{1}{\tilde{s}_d(z)}\boldsymbol{I}_d\right)^{-2}} + \frac{1}{z^2\tilde{s}_d(z)}\operatorname{Tr}\left(\boldsymbol{\Sigma} + \frac{1}{\tilde{s}(z)}\boldsymbol{I}_d\right)^{-1} \tag{C.24}$$

**Corollary 2** (Isotropic Wishart)**.** In the isotropic case $\boldsymbol{\Sigma} = \boldsymbol{I}_d$, we have $\nu = \delta_1$ and eq. (C.18) simplifies to:

$$\frac{1}{\tilde{s}} + z = \frac{c}{1+\tilde{s}} \tag{C.25}$$

which admits two explicit solutions, depending on the branch of the complex square-root function:

$$\tilde{s}_{\pm}(z) = \frac{c-1-z \pm \sqrt{(1-c+z)^2 - 4z}}{2z} \tag{C.26}$$

However, only $\tilde{s}_+$ satisfies the property $\operatorname{Im}(z)\operatorname{Im}(\tilde{s}) \geq 0$ in eq. (C.7). Thefore, we have the following asymptotic Stieltjes transforms for the isotropic Wishart matrices:

$$s(z; {}^1\!/_n \boldsymbol{X}^\top \boldsymbol{X}) = \frac{1-c-z+\sqrt{(1-c+z)^2-4z}}{2cz} \tag{C.27}$$

$$\tilde{s}(z; {}^1\!/_n \boldsymbol{X}\boldsymbol{X}^\top) = \frac{c-1-z+\sqrt{(1-c+z)^2-4z}}{2z} \tag{C.28}$$

Using the Stieltjes inversion formula in eq. (C.8), we can also obtain explicit expressions for the asymptotic spectral distribution:

$$\mu(x) = \left(1 - \frac{1}{c}\right)_+ \delta(0) + \frac{\sqrt{(c_+ - x)(x - c_-)}}{2\pi c x} \mathbb{I}_{[c_-, c_+]}(x) \tag{C.29}$$

$$\tag{C.30}$$

where $c_\pm = (1 \pm \sqrt{c})^2$. Note that since $\boldsymbol{X}^\top\boldsymbol{X}$ has the same non-zero eigenvalues as $\boldsymbol{X}\boldsymbol{X}^\top$, the only difference between $\mu$ and $\tilde{\mu}$ is on the number of zero eigenvalues:

$$\mu(x; {}^1\!/_n \boldsymbol{X}\boldsymbol{X}^\top) = \mu(x; {}^1\!/_n \boldsymbol{X}^\top\boldsymbol{X}) + (1-c)\delta(x). \tag{C.31}$$

Finally, note that in the isotropic case we have:

$$\mathrm{df}_1(\lambda; \boldsymbol{I}_d) = \frac{d}{1+\lambda} \tag{C.32}$$

and therefore:

$$\kappa(\lambda) = \frac{1}{\tilde{s}(-\lambda)} = \frac{1}{2}\left(\lambda + c - 1 + \sqrt{(1-c-\lambda)^2 + 4\lambda}\right) \tag{C.33}$$

In particular, we also have:

$$\mathrm{df}_1(\lambda; \hat{\boldsymbol{\Sigma}}) \sim \mathrm{df}_1(\kappa(\lambda); \boldsymbol{I}_d) = \frac{2}{1+c+\lambda+\sqrt{(1-c-\lambda)^2+4\lambda}} \tag{C.34}$$

Finally, note we can use eq. (C.23) to get:

$$\mathrm{df}_2(\lambda; \hat{\boldsymbol{\Sigma}}) \sim \mathrm{df}_2(\kappa(\lambda); \boldsymbol{I}_d) + \frac{d^2}{(1+\kappa(\lambda))^4} \frac{\kappa(\lambda)^2}{n - \mathrm{df}_2(\kappa(\lambda); \boldsymbol{I}_d)} \tag{C.35}$$

$$= \frac{d}{(1+\kappa)^2}\left(1 + \frac{c\kappa(\lambda)^2}{(1+\kappa(\lambda))^2 - c}\right) \tag{C.36}$$

# D Generalised approximate message passing

## D.1 Derivation from Belief Propagation

In this appendix we discuss the derivation of the Generalised Approximate Message Passing (GAMP) algorithm from belief propagation. Our derivation closely follow the ones appearing in (Zdeborová and Krzakala, 2016). To simplify the exposition, we restrict the derivation to the following particular case of the model discussed in Section 3:

- Separable prior distribution:

$$\varphi(\boldsymbol{\theta}) = \prod_{k=1}^{p} \varphi(\theta_k). \tag{D.1}$$

  For a discussion of the non-separable case, we refer the reader to Berthier et al. (2020); Gerbelot and Berthier (2023).

- Isotropic covariates:

$$\boldsymbol{v}_i \to \boldsymbol{x}_i \sim \mathcal{N}(\mathbf{0}_p, {}^1\!/_p \boldsymbol{I}_p), \quad i.i.d. \tag{D.2}$$

  For notational convenience, we switch from $\boldsymbol{v}_i \to \boldsymbol{x}_i$. This is to avoid confusion with the letter $V$ which is commonly used for the pre-activation variance in AMP. The non-isotropic case follows from the isotropic one by adapting the prior denoising function, see Clarté et al. (2023) for a discussion.

Under these simplifications, the posterior measure for generalised linear estimation reads:

$$P(\boldsymbol{\theta}|\boldsymbol{X}, \boldsymbol{y}) = \frac{1}{Z_d(\boldsymbol{X}, \boldsymbol{y})} \prod_{i=1}^{n} \psi(y_i|\langle \boldsymbol{\theta}, \boldsymbol{x}_i \rangle) \prod_{k=1}^{p} \varphi(\theta_k) \tag{D.3}$$

where $\boldsymbol{X} \in \mathbb{R}^{n \times p}$ is the matrix with rows $\boldsymbol{x}_i \in \mathbb{R}^p$ and $\boldsymbol{y} \in \mathbb{R}^n$ the vector with entries $y_i$.

⚠ Recall that $\varphi$ and $\psi$ are not required to be probability densities, just positive functions.

**Factor graph** — For the derivation, it will be convenient to represent the posterior distribution eq. (D.3) in a *factor graph*. This is a bi-partite graph with two types of nodes: variable nodes (◯) and factor (■) nodes. Every variable node represent a random variable in the problem. For the posterior distribution eq. (D.3), the variable nodes are $\theta_k$, which we will always index with indices $k, l, m \in [k]$. Factor nodes represent non-negative functions of these random variables, and for the posterior distribution we have the prior $\varphi(\theta_k)$ and the likelihood $\psi(y_i|\langle v_i, \boldsymbol{\theta} \rangle)$. We will always denote factor nodes with indices $i, j \in [n]$. Edges are placed between variables and factors whenever they are functionally dependent. Edges can only connect variables to factors, never factors to factors or variables to variables. We adopt the notation from graph theory, and the set of edges by $E$, and $\partial i = \{k \in [p] : (ik) \in E\}$, $\partial k = \{i \in [n] : (ik) \in E\}$.

The factor graph for the posterior eq. (D.3) is shown in fig. 11. Note that every prior factor connect to a single variable, while likelihood factors connect to all variables - in physics jargon, we say the model is *fully-connected*.

### D.1.1 Belief Propagation

Belief propagation, also known as the *sum-product message passing* algorithm is an algorithm introduced by Pearl (1982) for performing Bayesian inference in graphical models. Its goal is to estimate the

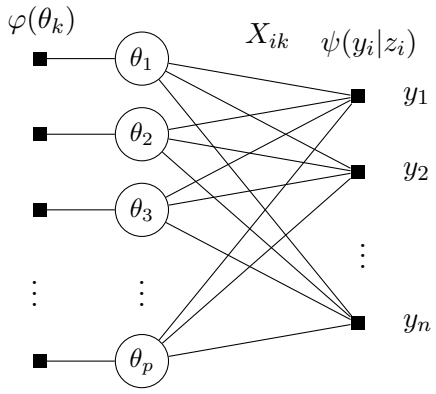


Figure 11: Factor graph for generalised linear estimation

marginals of probability distributions defined on graphs through an iterative message passing scheme. For the generalised linear estimation posterior eq. (D.3), the marginals are given by:

$$P(\theta_k|\boldsymbol{X},\boldsymbol{y}) \propto \int \prod_{l\neq k} \mathrm{d}\boldsymbol{\theta} P(\boldsymbol{\theta}|\boldsymbol{X},\boldsymbol{y}) = \varphi(\theta_k) \int \left(\prod_{l\neq k} \mathrm{d}\theta_l \varphi(\theta_l)\right) \prod_{i\in[n]} \psi(y_i|\langle \boldsymbol{\theta}, \boldsymbol{x}_i\rangle) \tag{D.4}$$

And BP seeks to estimate it by the procedure given in Algorithm 2.

*Remark* 11. A few remarks are in order.

- BP is composed of two types of messages: the variable-to-factor $m_{k\to i}$ and factor-to-variable $\hat{m}_{i\to k}$ beliefs. Factor-to-variable messages $m_{i\to k}(\theta_k)$ express the current belief of the factor $i$ over variable $k$ state. It aggregates all the variable-to-factor messages $m_{l\to i}(\theta_l)$ from the neighbouring variables excluding $k$ ($l \in \partial i \backslash k = [p]\backslash k$) and marginalise over them. Variable-to-factor messages $m_{i\to k}$ the beliefs of all factors in the neighbourhood of $k$ except $i$ - and therefore express the belief about variable $k$ state when factor $i$ is excluded.

- At every step, the algorithm requires going through all the factors and nodes of the factor graph. This requires **TODO:** number of operations per step. Different scheduling methods on what order to update the messages are possible.

- Note that an implicit assumption in BP is that the beliefs $m_{k\to i}$ are statistically independent. This is true for instance when the factor graph is a tree. Indeed, on tree-like[14] factor graphs, BP provably converges to the true marginals eq. (D.4) in a single forward-backward pass with an optimal scheduling. Although the factor graph of fully-connected models such as the GLM fig. 11 are not trees, as we will see later BP can still be exact. Indeed, this will be the case for mean-field models where the interaction between variables are $O(1/d)$, and therefore although present loops are subleading at large $d$.

⚠ Under the BP assumptions, $\theta_k \sim m_{k\to i}$ and $\theta_l \sim m_{l\to i}$ are independent random variables for $l \neq k$. This should not be confused with the components $\theta_k$ of posterior samples $\boldsymbol{\theta} \sim P(\boldsymbol{\theta}|\mathcal{D})$, which are not independent.

[14] Tree-like graphs are graphs for which the smallest loops are of size $O(\log d)$, and therefore they locally look like trees.

**Algorithm 2:** BP

**Input:** Data $\boldsymbol{V} \in \mathbb{R}^{n\times p}$, $\boldsymbol{y} \in \mathbb{R}^n$, likelihood $\psi$, prior $\varphi$.

**for** $t \leq T$ **do**

For every node $k \in [d]$ and factor $i \in [n]$:

$m^{t+1}_{k\to i}(\theta_k) = \frac{\varphi(\theta_k)\prod_{j\neq i}\hat{m}^t_{j\to k}(\theta_k)}{\int_{\mathbb{R}} \mathrm{d}\theta \varphi(\theta)\prod_{j\neq i}\hat{m}^t_{j\to k}(\theta)}$; /* *Update variable-to-factor belief* */

$\hat{m}^t_{i\to k}(\theta_k) = \frac{\int_{\mathbb{R}}\left(\prod_{l\neq k}\mathrm{d}\theta_l m^t_{l\to i}(\theta_l)\right)\psi(y_i|\langle \boldsymbol{x}_i,\boldsymbol{\theta}\rangle)}{\int \mathrm{d}\boldsymbol{\theta}\prod_{l\neq k} m^t_{l\to i}(\theta_l)\psi(y_i|\langle \boldsymbol{x}_i,\boldsymbol{\theta}\rangle)}$; /* *Update factor-to-variable belief*/

**end for**

**Return:** Marginals $m_k(\theta_k) = \frac{\prod_{i\in[n]}\hat{m}^T_{i\to k}(\theta_k)}{\int \mathrm{d}\theta\prod_{i\in[n]}\hat{m}^T_{i\to k}(\theta)}$

### D.1.2 Reduced Belief Propagation

When the posterior is a discrete probability distribution over a vocabulary $V$, the BP messages $(m_{i\to k}, \hat{m}_{k\to i})$ will be vectors of dimension $|V|$. However, when the posterior is a continuous probability density, as in the GLM case eq. (D.3), the messages are themselves probability densities. Therefore, running BP in practice for continuous probability densities requires binning the distribution, introducing new hyperparameters to tune and numerical errors.

In this case, it is common practice to assume the messages belong to a class of parametric probability density and reduce BP to an iterative algorithm on the parameters of this density. The simplest case is *Gaussian Belief Propagation*, in which case we assume the messages take the shape of a Gaussian density, see for example (Shental et al., 2008). In general, an ansatz introduces an inductive bias on the algorithm which will be reflected on the shape of the final BP marginals. This is a consequence of the fact that the model is fully-connected and with weakly interacting edges, implying that the distribution of the pre-activations is asymptotically Gaussian thanks to the central limit theorem. Gaussianity of the pre-activations will imply the posterior marginals take a particularly simple shape of a Gaussian convolution with the prior and likelihoods.

To see this, we start by looking on the factor-to-variable beliefs:

$$\hat{m}^t_{i\to k}(\theta_k) \propto \int_{\mathbb{R}}\left(\prod_{l\neq k}\mathrm{d}\theta_l m^t_{l\to i}(\theta_l)\right)\psi(y_i|\langle \boldsymbol{x}_i,\boldsymbol{\theta}\rangle) \tag{D.5}$$

Our goal is to understand how this distribution looks like for the GLM posterior. This expression takes the form of an expectation over $\theta_l \sim m_{l\to i}$, which by assumption are independent conditionally on $\theta_k$.

The first step is to study the distribution of the pre-activations $\boldsymbol{z} = \boldsymbol{X}\boldsymbol{\theta} \in \mathbb{R}^n$ (a.k.a. *local-fields* in the physics jargon). Note we can decompose it as:

$$z_i = \sum_{l=1}^{p} X_{il}\theta_l = \sum_{l\neq k} X_{il}\theta_l + X_{ik}\theta_k \tag{D.6}$$

This expression takes the form of an expectation over the variable-to-factor belief $m_{l\to i}$ for $l \in [p]\backslash k$. Recall that by assumption $\theta_l \sim m_{l\to i}$ are independent. Denoting by $\boldsymbol{z} = \boldsymbol{X}\boldsymbol{\theta} \in \mathbb{R}^n$ the pre-activations (a.k.a. *local-fields* in the physics jargon), we can decompose:

$$z_i = \sum_{l=1}^{p} X_{il}\theta_l = \sum_{l\neq k} X_{il}\theta_l + X_{ik}\theta_k \tag{D.7}$$

Note that the first term is a sum of $p-1$ independent random variables (recall $X_{il}$ are i.i.d. and by assumption $\theta_\ell \sim m_{l\to i}$ are independent). Its mean and variance under the $m_{l\to i}$ expectation reads:

$$\omega_{i\to k} = \mathbb{E}_{\theta_l \sim m_{l\to i}}\left[\sum_{l\neq k} X_{il}\theta_l\right] = \sum_{l\neq k} X_{il}\hat{\theta}_{l\to i}, \tag{D.8}$$

$$V_{i\to k} = \mathrm{Var}_{\theta_l \sim m_{l\to i}}\left[X_{il}\theta_l\right] = \sum_{l\neq k} V_{li}^2 \hat{c}_{l\to i}. \tag{D.9}$$

Where $\hat{\theta}_{k\to i}$ and $\hat{c}_{k\to i}$ are the mean and variances of the beliefs $m_{k\to i}$:

$$\hat{\theta}^t_{k\to i} = \int_{\mathbb{R}} \mathrm{d}\theta\; m^t_{k\to i}(\theta)\theta, \qquad \hat{c}^t_{k\to i}(\theta_k) = \int_{\mathbb{R}} \mathrm{d}\theta\; m_{k\to i}(\theta)^t\left[(\theta-\hat{\theta}_{k\to i})^2\right] \tag{D.10}$$

Since $X_{il}\sim\mathcal{N}(0,1/p)$, the mean is a $O(1)$ quantity, while the variance is $O(1/p)$. Therefore, the central limit theorem holds, and we have that:

$$z_i = z_{i\to k} + X_{ik}\theta_k, \qquad z_{i\to k}\sim\mathcal{N}(\omega_{i\to k}, V_{i\to k}) \tag{D.11}$$

Therefore, we can re-write the expectation in eq. (D.5) as:

$$\hat{m}^t_{i\to k}(\theta_k) \propto \int \frac{\mathrm{d}z}{\sqrt{2\pi V_{i\to k}}}\; e^{-\frac{1}{2V_{i\to k}}(z-\omega_{i\to k}-X_{ik}\theta_k)^2}\psi(y_i|z+X_{ik}\theta_k) \tag{D.12}$$

Note that this reduces the high-dimensional integral over $p-1$ variables to a single Gaussian integral. The attentive reader will notice that eq. (D.5) start to resemble a quantity we met in our replica computation for the GLM: the likelihood effective partition function $Z_y$ in eq. (4.44).

We can further simplify these equations by noting that since $X_{ik}\sim\mathcal{N}(0,1/p)$, the term $X_{ik}\theta_k = O(1/\sqrt{p})$. Therefore, we can expand eq. (D.12):

$$\hat{m}^t_{i\to k}(\theta_k) \propto \int \frac{\mathrm{d}z}{\sqrt{2\pi V_{i\to k}}}\; e^{-\frac{1}{2V_{i\to k}}(z-\omega_{i\to k})^2}\psi(y_i|z)\left[1-\frac{X_{ik}^2\theta_k^2}{2V_{i\to k}}+\frac{z-\omega_{i\to k}}{V_{i\to k}}X_{ik}\theta_k+\frac{(z-\omega_{i\to k})^2}{2V_{i\to k}}X_{ik}^2\theta_k^2+o(1/p)\right]$$

Defining the so-called likelihood *effective partition function* and *denoiser functions*:

$$Z_y(y,\omega,v) = \mathbb{E}_{z\sim\mathcal{N}(\omega,v)}[\psi(y|z)] = \int \frac{\mathrm{d}z}{\sqrt{2\pi v}} e^{-\frac{1}{2}(z-\omega)^2}\psi(y|z) \tag{D.13}$$

$$f_y(y,\omega,v) = \partial_\omega \log Z(y,\omega,v) = \frac{\int \frac{\mathrm{d}z}{\sqrt{2\pi v}} e^{-\frac{1}{2}(z-\omega)^2}\psi(y|z)\frac{z-\omega}{v}}{\int \frac{\mathrm{d}z}{\sqrt{2\pi v}} e^{-\frac{1}{2}(z-\omega)^2}\psi(y|z)} \tag{D.14}$$

and noting the following identity:

$$\partial_\omega f(y,\omega,v) = \frac{\partial_\omega^2 Z_y}{Z_y} - \left(\frac{\partial_\omega Z_y}{Z_y}\right)^2 = \frac{\int \frac{\mathrm{d}z}{\sqrt{2\pi v}} e^{-\frac{1}{2}(z-\omega)^2}\psi(y|z)\left(\frac{z-\omega}{v}\right)^2}{\int \frac{\mathrm{d}z}{\sqrt{2\pi v}} e^{-\frac{1}{2}(z-\omega)^2}\psi(y|z)} - f_y(y,\omega,v)^2 - \frac{1}{v} \tag{D.15}$$

we can write re-write eq. (D.13) as:

$$\begin{aligned}\hat{m}^t_{i\to k}(\theta_k) &\propto \mathcal{Z}_y(y,\omega_{i\to k},V_{i\to k})\left[1-\frac{X_{ik}^2\theta_k^2}{2V_{i\to k}} + f_y(y,\omega_{i\to k},V_{i\to k})X_{ik}\theta_k\right.\\ &\left.+\frac{1}{2}\left(\partial_\omega f_y(y,\omega_{i\to k},V_{i\to k}) + f_y(y,\omega_{i\to k},V_{i\to k})^2 + \frac{1}{V_{i\to k}}\right)V_{ik}^2\theta_k^2 + o(1/p)\right]\\ &= \mathcal{Z}_y(y,\omega_{i\to k},V_{i\to k})e^{\frac{1}{2}\partial_\omega f_y(y_i,\omega_{i\to k},V_{i\to k})V_{ik}^2\theta_k^2 + f_y(y_i,\omega_{i\to k},V_{i\to k})X_{ik}\theta_k} + o(1/p)\end{aligned} \tag{D.16}$$

where in the last equality we have re-exponentiated the expression. Finally, defining the following auxiliary variables:

$$A_{i\to k} = -\partial_\omega f_y(y_i, \omega_{i\to k}, V_{i\to k}) X_{ik}^2, \qquad b_{i\to k} = f_y(y_i, \omega_{i\to k}, V_{i\to k}) X_{ik} \tag{D.17}$$

and putting back the normalisation, we conclude that the factor-to-node beliefs are asymptotically given by a Gaussian density with sufficient statistics $(b_{i\to k}, A_{i\to k})$:

$$\hat{m}_{i\to k}(\theta_k) = \sqrt{\frac{A_{i\to k}}{2\pi}} e^{-\frac{1}{2}A_{i\to k}\theta_k^2 + b_{i\to k}\theta_k} \tag{D.18}$$

Gaussianity of the factor-to-variable beliefs $\hat{m}_{i\to k}$ directly imply Gaussianity of $m_{k\to i}$. Indeed, inserting the above in the update equations for the variable-to-factor beliefs give:

$$m_{k\to i}(\theta_k) = \frac{\varphi(\theta_k) e^{-\frac{1}{2}\sum_{j\neq i} A_{i\to k}\theta_k^2 + \sum_{j\neq i} b_{i\to k}\theta_k}}{\int_{\mathbb{R}} \mathrm{d}\theta\ \varphi(\theta) e^{-\frac{1}{2}\sum_{j\neq i} A_{i\to k}\theta^2 + \sum_{j\neq i} b_{i\to k}\theta}} \tag{D.19}$$

To close the loop, we note that our starting point, the varible-to-factor belief mean and variances defined in eq. (D.10) are related to the above by:

$$\hat{\theta}_{k\to i} = f_\theta\left(\sum_{j\neq i} b_{k\to j}, \sum_{j\neq i} A_{k\to i}\right), \qquad \hat{c}_{k\to i} = \partial_b f_\theta\left(\sum_{j\neq i} b_{k\to i}, \sum_{j\neq i} A_{k\to i}\right) \tag{D.20}$$

where we have defined the prior denoising function:

$$f_\theta(b, A) = \frac{\int_{\mathbb{R}} \mathrm{d}\theta \varphi(\theta)\theta e^{-\frac{1}{2}A\theta^2 + b\theta}}{\int_{\mathbb{R}} \mathrm{d}\theta \varphi(\theta) e^{-\frac{1}{2}A\theta^2 + b\theta}} \tag{D.21}$$

*Remark* 12. A few remarks about reduced BP are in order.

- All approximations we made are exact to order $o(1/p)$.
- For simplicity, we work with the first and second moments $(b_{i\to k}, A_{i\to k})$. Alternatively, we could have parametrised the factor-to-variable beliefs by the means and variances:

$$r_{i\to k} = \frac{b_{i\to k}}{A_{i\to k}}, \qquad \Sigma_{i\to k} = \frac{1}{A_{i\to k}} \tag{D.22}$$

  These are the variables used in some of the GAMP papers, e.g. (Zdeborová and Krzakala, 2016).
- We have introduced different variables, so it is good to keep in mind the meaning of each of them: $(\hat{\theta}_{k\to i}, \hat{c}_{k\to i})$ are the mean and variance of the (Gaussian) variable-to-factor beliefs; $(\omega_{i\to k}, V_{i\to k})$ are the mean and variances of the local fields $z_{i\to k} = \sum_{l\neq k} X_{il}$; $(b_{i\to k}, A_{i\to k})$ are sufficient statistics to the (Gaussian) factor-to-variable beliefs.
- In the context of fully-connected mean-field models where rBP is exact, it is also known as *reduced Belief Propagation* (rBP).

The Gaussian Belief Propagation algorithm is summarised in Algorithm 3

**Algorithm 3:** rBP

**Input:** Data $\boldsymbol{V} \in \mathbb{R}^{n\times p}$, $\boldsymbol{y} \in \mathbb{R}^n$, likelihood $\psi$, prior $\varphi$ (defining $f_\theta, f_y$).

Initialise $\hat{\theta}^0_{k\to i}$, $\hat{c}^0_{k\to i}$.

**for** $t \le T$ **do**

For every node $k \in [d]$ and factor $i \in [n]$:

*/* Update pre-activation mean and variances */*

$$V^t_{i\to k} = \sum_{l\neq k} X^2_{il}\hat{c}^t_{l\to i}; \quad \omega^t_{i\to k} = \sum_{l\neq k} X_{il}\hat{\theta}^t_{l\to i};$$

*/* Define auxiliary variables $A_{k\to i}, b_{k\to i}$ */*

$$b^t_{i\to k} = f_y(y_i, \omega^t_{i\to k}, V^t_{i\to k})X_{ik}; \quad A^t_{i\to k} = -\partial_\omega f_y(y_i, \omega_{i\to k}, V_{i\to k})X^2_{ik};$$

*/* Update marginal mean and variance beliefs */*

$$\hat{\theta}^{t+1}_{k\to i} = f_\theta\left(\sum_{j\neq i} b^t_{j\to k}, \sum_{j\neq i} A^t_{j\to k}\right); \quad \hat{c}^{t+1}_{j\to k} = \partial_b f_\theta\left(\sum_{j\neq i} b^t_{j\to k}, \sum_{j\neq i} A^t_{j\to k}\right);$$

**end for**

**Return:** */* Estimated marginal mean and variance */*

$$\hat{\theta}_k = f_\theta\left(\sum_{i=1}^n b^T_{i\to k}, \sum_{i=1}^n A^T_{i\to k}\right), \qquad \hat{c}_k = \partial_b f_\theta\left(\sum_{i=1}^n b^T_{i\to k}, \sum_{i=1}^n A^T_{i\to k}\right)$$

### D.1.3 From rBP to GAMP

Reduced BP 3 and GAMP 1 are pretty similar algorithms. A quick comparison reveals that the main different is on that GAMP depends only on the variable and factors nodes, while rBP has a dependence on the edges. Quite surprisingly, we don't loose anything by going from rBP to GAMP, as the approximation we will introduce in this section will only cost us a $o(1/p)$ factor, and therefore is asymptotically exact at the same order in which Gaussian GP is an exact approximation of BP. Although this might seem a small difference, it dropping the dependence of the edges has massive impact over the running time, reducing the number of operations to $O(np)$.

The key idea is to realise that the dependence of rBP on the factor/variable nodes is weak. For instance, consider the pre-activation variance:

$$V^t_{k\to i} = \sum_{l\neq k} X^2_{il}\hat{c}^t_{l\to i} = \sum_{l=1}^d X^2_{il}\hat{c}^t_{l\to i} - X^2_{ik}\hat{c}^t_{k\to i} \tag{D.23}$$

Note that the first term in the sum is independent of the variable node $k$ - all the dependence is in the second term. However, since $X_{ik} \sim \mathcal{N}(0, 1/p)$, this term is subleading in $p$: $V^2_{ik}\hat{c}^t_{k\to i} = O(1/p)$. The idea is to propagate this argument by keeping only leading order terms, which are $O(1/\sqrt{p})$. To implement this, we start by defining the following variable-independent messages:

$$\begin{cases} \omega^t_i & := \sum_{l=1}^d X_{il}\hat{\theta}^t_l \\ V^t_i & := \sum_{l=1}^d X^2_{il}\hat{c}^t_l \end{cases}, \qquad \begin{cases} b^t_k = \sum_{i=1}^n b_{i\to k} = \sum_{i=1}^n X_{ik} f_y(y_i, \omega^t_i, V^t_i) \\ A^t_k = \sum_{i=1}^n b_{i\to k} = -\sum_{i=1}^n X^2_{ik}\partial_\omega f_y(y_i, \omega^t_i, V^t_i) \end{cases}, \qquad \begin{cases} \hat{\theta}^{t+1}_k & = f_\theta(b_k, A_k) \\ \hat{c}^{t+1}_k & := \partial_b f_\theta(b_k, A_k) \end{cases} \tag{D.24}$$

⚠ Note that we define the factor independent quantities $(b_k, A_k)$ with the sum. This should not be confused with the messages $(b_{k\to i}, A_{k\to i})$ which are not summed over.

The first step is to note that the variables involving a $X_{il}^2 = O(1/p)$ factor can be directly simplified:

$$V_{k\to i}^t = V_i^t + O(1/p), \qquad \sum_{j\neq i} A_{j\to k}^t = A_k^t + O(1/p). \tag{D.25}$$

This is not the case for the other variables. Take for instance:

$$\hat{\theta}_{k\to i}^{t+1} = f_\theta\left(\sum_{j\neq i} b_{j\to k}^t, \sum_{j\neq i} A_{j\to k}^t\right) = f_\theta\left(b_k^t - b_{i\to k}^t, A_k^t\right) + O(1/p) \tag{D.26}$$

But since $b_{i\to k}^t = X_{ik} f_y(y_i, \omega_{i\to k}^t, V_{i\to k}^t) = O(1/\sqrt{p})$, we can expand on the first argument of $f_\theta$ to get:

$$\begin{aligned}\hat{\theta}_{k\to i}^{t+1} &= f_\theta\left(b_k^t, A_k^t\right) + \partial_b f_\theta\left(b_k^t, A_k^t\right) b_{k\to i} + O(1/p)\\ &= \hat{\theta}_k^{t+1} - \hat{c}_k^{t+1} b_{k\to i}^t + O(1/p)\end{aligned} \tag{D.27}$$

Note that by definition, this also implies that:

$$\hat{c}_{k\to i}^{t+1} = \mathrm{Var}(\hat{\theta}_{k\to i}^t) = \hat{c}_k^t + O(1/p) \tag{D.28}$$

since any correction will be quadratic on $b_{k\to i}^t$, hence $O(1/p)$. Note, however, that the equations are note closed, since they depend on edges through $b_{k\to i}^t$, which itself depends on $\omega_{k\to i}$. To close them, we need to further simplify their dependence on the edges. Looking at $\omega_{i\to k}^t$

$$\begin{aligned}\omega_{i\to k}^t &= \sum_{l\neq k} X_{il}\hat{\theta}_{l\to i} = \sum_{l\neq k}^{p} X_{il}\left(\hat{\theta}_l^{t+1} - \hat{c}_l^{t+1} b_{l\to i}^t + O(1/p)\right)\\ &= \omega_i^t - \sum_{k=1}^{p} X_{ik}\hat{c}_k^t b_{k\to i}^{t-1} + O(1/p)\end{aligned} \tag{D.29}$$

Now noting that:

$$\begin{aligned}b_{i\to k}^t &= X_{ik} f_y(y_i, \omega_{k\to i}^t, V_{k\to i}^t) = X_{ik} f_y(y_i, \omega_k^t - X_{ik}\hat{\theta}_k^t, V_k^t) + o(1/p)\\ &= X_{ik} f_y(y_i, \omega_i^t, V_i^t) - X_{ik}^2 \partial_\omega f_y(y_i, \omega_i^t, V_i^t)\hat{\theta}_k^t + o(1/p)\end{aligned} \tag{D.30}$$

we have:

$$\begin{aligned}\omega_{i\to k}^t &= \omega_i^t + f_y(y_i, \omega_i^{t-1}, V_i^{t-1}) \sum_{k=1}^{p} X_{ik}^2 \hat{c}_k^t + O(1/p)\\ &= \omega_i^t - V_i^t f_y(y_i, \omega_i^{t-1}, V_i^{t-1})\end{aligned} \tag{D.31}$$

which allow us to write all the updates using only variable and factors. The only missing update is $b_k^t$, which can be obtained by summing over eq. (D.30):

$$b_k^t = \sum_{i=1}^{n} b_{k\to k} = \sum_{i=1}^{n} X_{ik} f_y(y_i, \omega_i^t, V_i^t) - \hat{\theta}_k^t \sum_{i=1}^{n} X_{ik}^2 \partial_\omega f_y(y_i, \omega_i^t, V_i^t) \tag{D.32}$$

$$= \sum_{i=1}^{n} X_{ik} f_y(y_i, \omega_i^t, V_i^t) + A_i^t \hat{\theta}_k^t \tag{D.33}$$

Putting this together, we recover Algorithm 1 in the main. For convenience, we also write it in components here in Algorithm 4.

*Remark* 13. Two remarks about GAMP:

**Algorithm 4:** GAMP

**Input:** Data $\boldsymbol{X} \in \mathbb{R}^{n\times p}$, $\boldsymbol{y} \in \mathbb{R}^n$, likelihood $\psi$, prior $\varphi$ (defining $f_\theta, f_y$).

Initialise $\hat{\theta}_k^0$, $\hat{c}_k^0$, $g_i^{-1} = 0$.

**for** $t \leq T$ **do**

For every node $k \in [d]$ and factor $i \in [n]$:

*/* Update pre-activation mean and variances */*

$$V_{i\to k}^t = \sum_{k=1}^p X_{il}^2 \hat{c}_k^t; \quad \omega_i^t = \sum_{k=1}^p X_{ik}\hat{\theta}_k^t - V_i^t g_i^{t-1};$$

*/* Define likelihood denoisers */*

$$g_i^t = f_y(y_i, \omega_i^t, V_i^t); \quad \partial g_i = \partial_\omega f_y(y_i, \omega_i^t, V_i^t);$$

*/* Update $A_k, b_k$ */*

$$A_k^t = -\sum_{i=1}^n X_{ik}^2 \partial g_i; \quad b_k^t = \sum_{i=1}^n X_{ik} g_i^t + A_i^t \hat{\theta}_k^t;$$

*/* Update mean and variance*/*

$$\hat{\theta}_k^{t+1} = f_\theta\left(b_i^t, A_i^t\right); \quad \hat{c}_k^{t+1} = \partial_b f_\theta\left(b_k^t, A_k^t\right);$$

**end for**

**Return:** GAMP mean and variance $\hat{\theta}_k^T, \hat{c}_k^T$

- The term $V_i g_i^{t-1}$ is known as the *Onsager term*, and plays a fundamental in ensuring that the AMP marginals remain Gaussian at every iteration, despite the non-Gaussian correlations introduced when repeatedly iterating over the covariates matrix $X_{ik}$. The fact that at iteration $t$ it depends on the previous iterate $t-1$ is also crucial for the stability of the algorithm, and was a source of confusion in the early literature in the context of the TAP equations, see (Zdeborová and Krzakala, 2016) for a discussion.
- Since $X_{ki} \sim \mathcal{N}(0, 1/p)$, we have $\mathrm{Var}(X_{ki}^2) = O(1/p^2)$, which means that $X_{ik}^2$ concentrate to leading order in $O(1/p)$. Therefore, we can further take $X_{ki}^2 \to 1/p$ in all terms of GAMP 4 where $X_{ki}^2$ appears.

## D.2 State Evolution

In this section, we discuss the derivation of the state evolution equations. As in Appendix D.1, to simplify the exposition we will focus in the isotropic case $\boldsymbol{x} \sim \mathcal{N}(\boldsymbol{0}, 1/p \boldsymbol{I}_p)$, and will assume that data was independently drawn generated from:

$$y_i \sim P_\star(y_i|\langle \boldsymbol{\beta}_\star, \boldsymbol{x}\rangle), \qquad \boldsymbol{\beta}_\star \sim \prod_{k=1}^p P_\beta(\beta_k) \tag{D.34}$$

As before, the discussion can be generalised to the full GCM model, with non-separable signal distribution $P_{\boldsymbol{\beta}}$ (see (Berthier et al., 2020; Gerbelot and Berthier, 2023)) and non-isotropic covariates $(\boldsymbol{u}, \boldsymbol{v})$ (see (Clarté et al., 2023)).

The goal of state evolution is to derive a closed set of equations tracking the performance of the GAMP algorithm 4.

### D.2.1 Derivation of the state evolution equations

State evolution can be derived both from GAMP 4 or rBP 3 - since these are asymptotically equivalent algorithms, there is no difference. Here we opt to derive it from rBP since it is neater. The starting point is to note that the pre-activation of the data likelihood $\boldsymbol{\nu} = \boldsymbol{X}\boldsymbol{\beta}_\star$ is a Gaussian variable with zero mean and covariance $\mathbb{E}[\boldsymbol{\nu}\boldsymbol{\nu}^\top] = \rho \boldsymbol{I}_n$ with $\rho = \mathbb{E}[1/p||\boldsymbol{\beta}_\star||^2]$. The rBP estimator $\hat{\theta}_{k\to i}^t$ correlates with the target pre-activations $\nu$ through the covariates $\boldsymbol{X}_{ik}$, and our go is to track these correlations.

First, as we already noted in remark 13, to leading order $X_{ik}^2 = 1/p + o(1/p)$. Therefore, to leading order the variables $V_i$ and $A_k$ are independent of the covariates:

$$V_{i\to k}^t = V_i^t + o(1/p) = \frac{1}{p}\sum_{k=1}^{p} \hat{c}_k^t + o(1/p) = \frac{\mathbf{1}^\top \hat{\boldsymbol{c}}^t}{p} + o(1/p) \tag{D.35}$$

$$\sum_{j\neq i} A_{j\to k}^t = A_i^t + o(1/p) = -\frac{1}{p}\sum_{i=1}^{n} \partial g_i^t + o(1/p) = -\frac{\mathbf{1}_n^\top \boldsymbol{\partial g}^t}{p} + o(1/p) \tag{D.36}$$

This means all the dependence on $X_{ik}$ (and hence correlation with $\boldsymbol{\nu}$) is through $\omega_{i\to k}^t$ and $b_{i\to k}^t$. Let's start with the first.

At every iteration $t < T$, $\hat{\theta}_{l\to i}^t$ are independent of $X_{ik}$, and since $X_{ki} \sim \mathcal{N}(0, 1/p)$ the variables $\omega_{i\to k}^t = \sum_{l\neq k} X_{il}\hat{\theta}_{l\to i}$ are jointly Gaussian with mean $\mathbb{E}[\omega_{i\to k}] = 0$ and covariance given by:

$$\begin{aligned}\mathbb{E}[\omega_{i\to k}\omega_{j\to l}] &= \delta_{ij}\delta_{kl}\frac{1}{p}\sum_{m\neq k}\left(\hat{\theta}_{m\to i}^t\right)^2 = \delta_{ij}\delta_{kl}\frac{1}{p}\sum_{m}\left(\hat{\theta}_{m}^t\right)^2 + o(1/p)\\ &= \delta_{ij}\delta_{kl}\frac{1}{p}||\boldsymbol{\theta}^t||_2^2 + o(1/p)\end{aligned} \tag{D.37}$$

where $\boldsymbol{\theta}^t \in \mathbb{R}^p$ is the vector with components $\theta_k^t$. Moreover, $\omega_{i\to k}^t$ are correlated to the target pre-activation. Decomposing:

$$\nu_j = \sum_{l=1}^{p} X_{jk}\beta_{\star,l} = \sum_{l\neq k}^{p} X_{jl}\beta_{\star,l} + X_{jk}\beta_{\star,k} = \nu_{j\to k} + X_{jk}\beta_{\star,k} \tag{D.38}$$

We have:

$$\mathbb{E}[\omega_{i\to k}^t \nu_{j\to l}] = \delta_{ij}\delta_{kl}\frac{1}{p}\sum_{l\neq k}\hat{\theta}_{l\to i}\beta_{\star,l} = \delta_{ij}\delta_{kl}\frac{1}{p}\langle\hat{\boldsymbol{\theta}}^t, \boldsymbol{\beta}_\star\rangle + o(1/p) \tag{D.39}$$

⚠ Note that $\omega_{i\to k}^t = \sum_{l\neq k} X_{il}\hat{\theta}_l^t$ are independent of $X_{ik}$ but $\nu_j$ does. Therefore, this decomposition is not necessary at this step, since the additional term is zero under the expectation. However, this decomposition will be important for the following steps.

Therefore, defining:

$$q^t = \frac{||\hat{\boldsymbol{\theta}}^t||_2^2}{p}, \qquad m = \frac{\langle\hat{\boldsymbol{\theta}}^t, \boldsymbol{\beta}_\star\rangle}{p} \tag{D.40}$$

We conclude that to leading order:

$$(\omega_{i\to k}^t, \nu_{i\to k}) \sim \mathcal{N}\left(\mathbf{0}_2, \begin{bmatrix}\rho & m^t\\ m^t & q^t\end{bmatrix}\right) \tag{D.41}$$

independently over $i \in [n]$ and $k \in [p]$. Now let's look at $b$, which is a trickier term:

$$\sum_{j\neq i} b_{j\to k} = \sum_{j\neq i} X_{jk} f_y(y_j, \omega_{j\to k}, V_{j\to k}) \tag{D.42}$$

The variance $V_{j\to k}$ concentrates. Moreover, the variable $\omega_{j\to k}^t = \sum_{l\neq k} X_{il}\hat{\theta}_l^t$ is independent of $X_{jk}$. However, $y_j \sim P_\star(y|\langle\boldsymbol{\beta}_\star, \boldsymbol{x}_j\rangle)$ is correlated with $X_{jk}$ since it involves a sum over all $k \in [n]$. We need

to account for this dependence. For that, let's denote $y_i = f_\star(\nu_i)$ with $f_\star$ a stochastic function. Then, we can then write:

$$\begin{aligned}\sum_{j\neq i} b_{j\to k} &= \sum_{j\neq i} X_{jk} f_y\left(f_\star\left(\nu_{j\to k} + X_{jk}\beta_{\star,k}\right), \omega_{j\to k}, V_{j\to k}\right) \\ &= \sum_{j\neq i} X_{jk} f_y\left(f_\star\left(\nu_{j\to k}\right), \omega_{j\to k}, V_{j\to k}\right) + \beta_{\star,k}\sum_{j\neq i} X_{jk}^2 \partial_\nu f_y\left(f_\star\left(\nu_{j\to k}\right), \omega_{j\to k}, V_{j\to k}\right) \end{aligned} \tag{D.43}$$

Let's look at both terms in this sum separately. With our splitting, $f_y(f_\star(\nu_{j\to k}), \omega^t_{j\to k}, V^t_{j\to k})$ are independent of $X_{jk}$. Moreover, they are also independent over $j$. Therefore, again, the CLT holds, and this first term will be an asymptotically Gaussian random variable with zero mean and variance given by:

$$\hat{q} = \alpha \mathbb{E}_{(\nu,\omega)}[f_y(f_\star(\nu), \omega, v)^2] \tag{D.44}$$

where $\alpha = {}^n/_d$, $v = {}^1/_p\langle \mathbf{1}_p, \hat{\boldsymbol{c}}^t\rangle$ and the expectation is over eq. (D.41). To leading order, the second concentrates. Defining:

$$\hat{m} = \alpha \mathbb{E}_{(\nu,\omega)}[\partial_\nu f_y(f_\star(\nu), \omega, v)]. \tag{D.45}$$

We can write:

$$b^t_k = \sqrt{\hat{q}^t}\xi + \hat{m}^t \beta_{\star,k} \tag{D.46}$$

Moreover, note that we can also relate $A^t_i$ in eq. (D.36) to the above:

$$\hat{v} \coloneqq A^t_k = -\alpha \mathbb{E}_{(\nu,\omega^t)}[\partial_\omega f_y(f_\star(\nu), \omega^t, v^t)] \tag{D.47}$$

To close the equations, it remains to relate $(\hat{q}, \hat{m})$ to $(q, m, v)$. This can be achieved by using the definition of the updates:

$$\hat{\theta}^{t+1}_k = f_\theta\left(b^t_i, A^t_i\right), \qquad \hat{c}^{t+1}_k = \partial_b f_\theta\left(b^t_k, A^t_k\right) \tag{D.48}$$

Which give us:

$$q^{t+1} = \mathbb{E}_{\xi,\beta_\star}\left[f_\theta(\sqrt{\hat{q}^t}\xi + \hat{m}\beta_\star, \hat{q}^t)^2\right], \tag{D.49}$$

$$m^{t+1} = \mathbb{E}_{\xi,\beta_\star}\left[f_\theta(\sqrt{\hat{q}^t}\xi + \hat{m}^t\beta_\star, \hat{q}^t)\beta_\star\right] \tag{D.50}$$

$$v^{t+1} = \mathbb{E}_{\xi,\beta_\star}\left[\partial_b f_\theta(\sqrt{\hat{q}^t}\xi + \hat{m}^t\beta_\star, \hat{q}^t)\right] \tag{D.51}$$

Putting together, we have the following state evolution equations:

$$\begin{cases} \hat{v}^t = -\alpha \mathbb{E}_{(\nu,\omega^t)}[\partial_\omega f_y(f_\star(\nu), \omega^t, v^t)] \\ \hat{q}^t = \alpha \mathbb{E}_{(\nu,\omega^t)}[f_y(f_\star(\nu), \omega^t, v^t)^2] \\ \hat{m}^t = \alpha \mathbb{E}_{(\nu,\omega^t)}[\partial_\nu f_y(f_\star(\nu), \omega^t, v^t)] \end{cases}, \quad \begin{cases} v^{t+1} &= \mathbb{E}_{\xi,\beta_\star}\left[\partial_b f_\theta(\sqrt{\hat{q}^t}\xi + \hat{m}^t\beta_\star, \hat{v}^t)\right] \\ q^{t+1} &= \mathbb{E}_{\xi,\beta_\star}\left[f_\theta(\sqrt{\hat{q}^t}\xi + \hat{m}\beta_\star, \hat{v}^t)^2\right], \\ m^{t+1} &= \mathbb{E}_{\xi,\beta_\star}\left[f_\theta(\sqrt{\hat{q}^t}\xi + \hat{m}^t\beta_\star, \hat{v}^t)\beta_\star\right] \end{cases} \tag{D.52}$$

# E Massaging the self-consistent equations

## E.1 A simplified expression for ridge regression on the GCM

We are interested in finding a closed-form expression for the asymptotic excess risk of ridge regression for the Gaussian covariate model:

$$r - \sigma^2 = 1 + q - 2m \tag{E.1}$$

where $q, m$ solve the following self-consistent equations:

$$\begin{cases} \hat{v} = \frac{\alpha}{1+v} \\ \hat{q} = \alpha \frac{1+\sigma^2+q-2m}{(1+v)^2} \\ \hat{m} = \sqrt{\gamma}\frac{\alpha}{1+v} \end{cases} \qquad \begin{cases} v = \frac{1}{p}\operatorname{Tr}\boldsymbol{\Omega}(\lambda \boldsymbol{I}_p + \hat{v}\boldsymbol{\Omega})^{-1} \\ q = \frac{1}{p}\operatorname{Tr}\left(\hat{q}\boldsymbol{\Omega} + \hat{m}^2 \boldsymbol{\Phi}\boldsymbol{\beta}_\star\boldsymbol{\beta}_\star^\top\boldsymbol{\Phi}^\top\right)\boldsymbol{\Omega}(\lambda \boldsymbol{I}_p + \hat{v}\boldsymbol{\Omega})^{-2} \\ m = \sqrt{\gamma}\frac{\hat{m}}{p}\operatorname{Tr}\boldsymbol{\Phi}\boldsymbol{\beta}_\star\boldsymbol{\beta}_\star^\top\boldsymbol{\Phi}^\top(\lambda \boldsymbol{I}_p + \hat{v}\boldsymbol{\Omega})^{-1} \end{cases} \tag{E.2}$$

To alleviate the notation, we define a short-hand for the normalised trace of a matrix $\boldsymbol{A} \in \mathbb{R}^{n\times n}$:

$$\operatorname{tr}\boldsymbol{A} \coloneqq \frac{1}{n}\operatorname{Tr}\boldsymbol{A}. \tag{E.3}$$

and for the resolvent of $\boldsymbol{\Omega} \in \mathbb{R}^{p\times p}$ evaluated at the negative real axis $z = -t$:

$$\boldsymbol{R}(t;\boldsymbol{\Omega}) = (\boldsymbol{\Omega} + t\boldsymbol{I}_p)^{-1}, \quad t \in \mathbb{R}_+ \tag{E.4}$$

When it is clear from the context, we will omit the dependency on $\boldsymbol{\Omega}$ and denote $\boldsymbol{R}(t;\boldsymbol{\Omega})\boldsymbol{R}(t)$. First, note that many of these variables are redundant. For instance, we can solve for:

$$v = \frac{\alpha}{\hat{v}} - 1 \tag{E.5}$$

Inserting this in the equation for $v$, we get a closed equation for $\hat{v}$:

$$\frac{\alpha}{\hat{v}} - 1 = \operatorname{tr}\boldsymbol{\Omega}(\lambda \boldsymbol{I}_p + \hat{v}\boldsymbol{\Omega})^{-1} = \frac{1}{\hat{v}}\operatorname{tr}\boldsymbol{\Omega}\left(\frac{\lambda}{\hat{v}}\boldsymbol{I}_p + \boldsymbol{\Omega}\right)^{-1} = \frac{1}{\hat{v}}\mathrm{df}_1(\lambda/\hat{v}) \tag{E.6}$$

where we defined the (normalised) degrees of freedom (c.f. Definition 7):

$$\tilde{\mathrm{df}}_\alpha(\lambda;\boldsymbol{\Omega}) = \frac{1}{p}\mathrm{df}_\alpha(\nu;\boldsymbol{\Omega}) = \frac{1}{p}\operatorname{Tr}\boldsymbol{\Omega}^\alpha(\lambda \boldsymbol{I}_p + \boldsymbol{\Omega})^{-\alpha} \tag{E.7}$$

This suggests that it makes sense to work with the following variable $\nu = \lambda/\hat{v} \in \mathbb{R}_+$, so that eq. (E.6) read:

$$\alpha\nu - \lambda = \nu\tilde{\mathrm{df}}_1(\nu) \qquad \Leftrightarrow \qquad \alpha - \frac{\lambda}{\nu} = \tilde{\mathrm{df}}_1(\nu) \tag{E.8}$$

Similarly, we can trivially rewrite $\hat{m}$ as a function of $\hat{v}$:

$$\hat{m} = \sqrt{\gamma}\hat{v} \tag{E.9}$$

Which allow us to write $m$ entirely as a function of $\hat{v}$

$$\begin{aligned} m &= \gamma\hat{v}\operatorname{tr}\boldsymbol{\Phi}\boldsymbol{\beta}_\star\boldsymbol{\beta}_\star^\top\boldsymbol{\Phi}^\top(\lambda \boldsymbol{I}_p + \hat{v}\boldsymbol{\Omega})^{-1} \\ &= \gamma\operatorname{tr}\boldsymbol{\Phi}\boldsymbol{\beta}_\star\boldsymbol{\beta}_\star^\top\boldsymbol{\Phi}^\top(\lambda/\hat{v}\boldsymbol{I}_p + \boldsymbol{\Omega})^{-1} \\ &= \gamma\operatorname{tr}\boldsymbol{\Phi}\boldsymbol{\beta}_\star\boldsymbol{\beta}_\star^\top\boldsymbol{\Phi}^\top\boldsymbol{R}(\nu) \end{aligned} \tag{E.10}$$

We now note that:

$$\hat{q} = \frac{\hat{v}^2}{\alpha}r \tag{E.11}$$

Inserting in the equation for $q$:

$$\begin{aligned} q &= \hat{v}^2\operatorname{tr}\left(\frac{r}{\alpha}\boldsymbol{\Omega} + \gamma\boldsymbol{\Phi}\boldsymbol{\beta}_\star\boldsymbol{\beta}_\star^\top\boldsymbol{\Phi}^\top\right)\boldsymbol{\Omega}(\lambda \boldsymbol{I}_p + \hat{v}\boldsymbol{\Omega})^{-2} \\ &= \operatorname{tr}\left(\frac{r}{\alpha}\boldsymbol{\Omega} + \gamma\boldsymbol{\Phi}\boldsymbol{\beta}_\star\boldsymbol{\beta}_\star^\top\boldsymbol{\Phi}^\top\right)\boldsymbol{\Omega}(\lambda/\hat{v}\boldsymbol{I}_p + \boldsymbol{\Omega})^{-2} \qquad \text{(E.12)} \\ &= \operatorname{tr}\left(\frac{r}{\alpha}\boldsymbol{\Omega} + \gamma\boldsymbol{\Phi}\boldsymbol{\beta}_\star\boldsymbol{\beta}_\star^\top\boldsymbol{\Phi}^\top\right)\boldsymbol{\Omega}\boldsymbol{R}(\nu)^2 \qquad \text{(E.13)} \end{aligned}$$

which we can also write as:

$$q = \frac{r}{\alpha}\tilde{\mathrm{df}}_2(\nu) + \frac{\gamma}{p}\langle \boldsymbol{\Phi}\boldsymbol{\beta}_\star, \boldsymbol{\Omega}(\boldsymbol{\Omega} + \lambda \boldsymbol{I}_p)^{-2}\boldsymbol{\Phi}\boldsymbol{\beta}_\star\rangle \tag{E.14}$$

$$\tag{E.15}$$

This allow us to derive a self-consistent equation for the risk:

$$r - \sigma^2 \coloneqq \rho + q - 2m = \rho + \frac{r}{\alpha}\tilde{\mathrm{df}}_2(\nu) + B - 2m \tag{E.16}$$

with:

$$\begin{aligned} B &\coloneqq \frac{\gamma}{p}\langle \boldsymbol{\Phi}\boldsymbol{\beta}_\star, \boldsymbol{\Omega}(\boldsymbol{\Omega} + \nu \boldsymbol{I}_p)^{-2}\boldsymbol{\Phi}\boldsymbol{\beta}_\star\rangle \\ &\overset{(a)}{=} m - \frac{\gamma\nu}{p}\langle \boldsymbol{\Phi}\boldsymbol{\beta}_\star, (\boldsymbol{\Omega} + \nu \boldsymbol{I}_p)^{-2}\boldsymbol{\Phi}\boldsymbol{\beta}_\star\rangle \\ m &= \frac{\gamma}{p}\langle \boldsymbol{\Phi}\boldsymbol{\beta}_\star, (\boldsymbol{\Omega} + \nu \boldsymbol{I}_p)^{-1}\boldsymbol{\Phi}\boldsymbol{\beta}_\star\rangle \end{aligned} \tag{E.17}$$

where in (a) we used the following identity:

$$\boldsymbol{\Omega}(\nu \boldsymbol{I}_p + \boldsymbol{\Omega})^{-1} = (\boldsymbol{\Omega} + \nu \boldsymbol{I}_p - \nu \boldsymbol{I}_p)(\nu \boldsymbol{I}_p + \boldsymbol{\Omega})^{-1} = \boldsymbol{I}_p - \nu(\nu \boldsymbol{I}_p + \boldsymbol{\Omega})^{-1} \tag{E.18}$$

This can be solved to yield an expression for the excess risk:

$$\begin{aligned} r - \sigma^2 &= \frac{\alpha}{\alpha - \tilde{\mathrm{df}}_2(\nu)}\left(\rho + B - 2m\right) + \sigma^2 \frac{\tilde{\mathrm{df}}_2(\nu)}{\alpha - \tilde{\mathrm{df}}_2(\nu)} \\ &= \frac{\alpha}{\alpha - \tilde{\mathrm{df}}_2(\nu)}\left(\rho - \gamma/p\langle \boldsymbol{\Phi}\boldsymbol{\beta}_\star, (\boldsymbol{I}_p + \nu \boldsymbol{R}(\nu))\boldsymbol{R}(\nu)\boldsymbol{\Phi}\boldsymbol{\beta}_\star\rangle\right) + \sigma^2 \frac{\tilde{\mathrm{df}}_2(\nu)}{\alpha - \tilde{\mathrm{df}}_2(\nu)} \end{aligned} \tag{E.19}$$

and allow us to identify the bias and variance decomposition of the excess risk:

$$\mathcal{B}(\alpha, \gamma, \lambda) = \frac{\alpha}{\alpha - \tilde{\mathrm{df}}_2(\nu)}\left(\rho - \gamma/p\langle \boldsymbol{\Phi}\boldsymbol{\beta}_\star, (\boldsymbol{I}_p + \nu \boldsymbol{R}(\nu))\boldsymbol{R}(\nu)\boldsymbol{\Phi}\boldsymbol{\beta}_\star\rangle\right) \tag{E.20}$$

$$\mathcal{V}(\alpha, \gamma, \lambda) = \sigma^2 \frac{\tilde{\mathrm{df}}_2(\nu)}{\alpha - \tilde{\mathrm{df}}_2(\nu)} \tag{E.21}$$

with $\nu$ the solution of:

$$\alpha - \frac{\lambda}{\nu} = \tilde{\mathrm{df}}_1(\nu) \tag{E.22}$$

Finally, note that this expression can be written in terms of an integral. Consider the spectral decomposition of $\boldsymbol{\Omega}$:

$$\boldsymbol{\Omega} = \sum_{j=1}^{p} \sigma_j \boldsymbol{u}_j \boldsymbol{u}_j^\top \tag{E.23}$$

We can define:

$$\eta_j = \langle \boldsymbol{u}_j, \boldsymbol{\Phi}\boldsymbol{\beta}_\star\rangle \tag{E.24}$$

Then, defining the joint density:

$$\mu(\eta, \sigma) = \sum_{j=1}^{p} \delta(\eta - \eta_j)\delta(\sigma_j - \sigma) \tag{E.25}$$

Equation (E.19) can be rewritten as:

$$\mathcal{B}(\alpha,\gamma,\lambda) = \frac{\alpha}{\alpha - \mathbb{E}_{\sigma\sim\mu}\left[\frac{\sigma^2}{(\sigma+\nu)^2}\right]}\left(\rho - \gamma\mathbb{E}_{(\eta,\sigma)\sim\mu}\left[\frac{\eta^2}{\sigma+\nu}\left(\frac{\nu}{\sigma+\nu}+1\right)\right]\right) \tag{E.26}$$

$$\mathcal{V}(\alpha,\gamma,\lambda) = \sigma^2\frac{\mathbb{E}_{\sigma\sim\mu}\left[\frac{\sigma^2}{(\sigma+\nu)^2}\right]}{\alpha - \mathbb{E}_{\sigma\sim\mu}\left[\frac{\sigma^2}{(\sigma+\nu)^2}\right]} \tag{E.27}$$

with $\nu$ the solution of:

$$\alpha - \frac{\lambda}{\nu} = \mathbb{E}\left[\frac{\sigma}{\sigma+\nu}\right] \tag{E.28}$$

## E.2 Well-specified ridge regression

In the well specified case, we have $\gamma = 1$ ($p = d$) and $\boldsymbol{\Phi} = \boldsymbol{\Omega}$. Note the variance term remains the same, while the bias term now reads:

$$\mathcal{B}(\alpha,\gamma,\lambda) = \frac{\alpha}{\alpha - \tilde{\mathrm{df}}_2(\nu)}\left(\rho - {}^1\!/_p\langle\boldsymbol{\beta}_\star, \boldsymbol{\Omega}^2(\boldsymbol{I}_p + \nu\boldsymbol{R}(\nu))\boldsymbol{R}(\nu)\boldsymbol{\beta}_\star\rangle\right) \tag{E.29}$$

Using eq. (E.18) again:

$$\begin{aligned}\boldsymbol{\Omega}^2\boldsymbol{R}(\nu) &= \boldsymbol{\Omega} - \nu\boldsymbol{\Omega}\boldsymbol{R}(\nu)\\ \boldsymbol{\Omega}^2\boldsymbol{R}(\nu)^2 &= \boldsymbol{\Omega}\boldsymbol{R}(\nu) - \nu\boldsymbol{\Omega}\boldsymbol{R}(\nu)^2\end{aligned} \tag{E.30}$$

This allow us to simplify the inner product term:

$${}^1\!/_p\langle\boldsymbol{\beta}_\star, \boldsymbol{\Omega}^2(\boldsymbol{I}_p + \nu\boldsymbol{R}(\nu))\boldsymbol{R}(\nu)\boldsymbol{\beta}_\star\rangle = \rho - {}^{\nu^2}\!/_p\langle\boldsymbol{\beta}_\star, \boldsymbol{\Omega}\boldsymbol{R}(\nu)^2\boldsymbol{\beta}_\star\rangle \tag{E.31}$$

Therefore, the bias term simplifies, and is given by:

$$\mathcal{B}(\alpha,\lambda) = \frac{\alpha\nu^2}{\alpha - \tilde{\mathrm{df}}_2(\nu)}\frac{1}{p}\langle\boldsymbol{\beta}_\star, \boldsymbol{\Omega}\left(\nu\boldsymbol{I}_p + \boldsymbol{\Omega}\right)^{-2}\boldsymbol{\beta}_\star\rangle \tag{E.32}$$

**Isotropic covariance** — Consider the isotropic case where $\boldsymbol{\Omega} = \boldsymbol{I}_d$. We then have:

$$\alpha - \frac{\lambda}{\nu} = \frac{1}{1+\nu} \tag{E.33}$$

Which admits an explicit solution:

$$\nu_\star(\alpha,\lambda) = \frac{1 - \alpha + \lambda + \sqrt{(1-\alpha+\lambda)^2 + 4\alpha\lambda}}{2\alpha} \tag{E.34}$$

and the bias and variance terms can be simplified to:

$$\begin{aligned}\mathcal{B}(\alpha,\lambda) &= \frac{\alpha\nu_\star^2}{\alpha - \tilde{\mathrm{df}}_2(\nu_\star)}\frac{\rho}{(1+\nu_\star)^2} = \frac{\alpha\rho\nu_\star^2}{\alpha(1+\nu_\star)^2 - 1}\\ \mathcal{V}(\alpha,\lambda) &= \sigma^2\frac{\tilde{\mathrm{df}}_2(\nu_\star)}{\alpha - \tilde{\mathrm{df}}_2(\nu_\star)} = \frac{\sigma^2}{\alpha(1+\nu_\star)^2 - 1}\end{aligned} \tag{E.35}$$

Note in particular that at interpolation $\lambda = 0^+$, we have:

$$\nu_\star(0,\alpha) = \begin{cases}{}^1\!/_\alpha - 1 & \text{for } 0 \le \alpha < 1\\ 0 & \text{for } \alpha \ge 1\end{cases} \tag{E.36}$$

and therefore:

$$\mathcal{B}(\alpha,0)=\begin{cases}\frac{1-\alpha}{\alpha^2} & \text{for } 0\leq\alpha<1\\ 0 & \text{for } \alpha\geq 1\end{cases}, \qquad \mathcal{V}(\alpha,0)=\begin{cases}\frac{\sigma^2\alpha}{1-\alpha} & \text{for } 0\leq\alpha<1\\ \infty & \text{for } \alpha=1\\ \frac{\sigma^2}{\alpha-1} & \text{for } \alpha>1\end{cases} \tag{E.37}$$

In particular, note that we have $\nu_\star(\lambda,1)\sim\sqrt{\lambda}$ as $\lambda\to 0^+$, therefore $\mathcal{V}(\lambda,1)\sim\frac{1}{\sqrt{\lambda}}$ as $\lambda\to 0^+$.

### E.3 Random features ridge regression

We now consider the particular example of ridge regression on the random features model, where:

$$\boldsymbol{\Phi}=b_1\frac{\boldsymbol{W}_0}{\sqrt{d}}, \qquad \boldsymbol{\Omega}=b_1^2\frac{\boldsymbol{W}_0\boldsymbol{W}_0^\top}{d}+b_\star^2\boldsymbol{I}_p=b_1^2\frac{\boldsymbol{W}_0\boldsymbol{W}_0^\top}{d}+b_\star^2\boldsymbol{I}_p \tag{E.38}$$

In particular, note that we have:

$$\boldsymbol{\Omega}=\boldsymbol{\Phi}\boldsymbol{\Phi}^\top+b_\star^2\boldsymbol{I}_p \tag{E.39}$$

which means that $\boldsymbol{\Phi}\boldsymbol{\Phi}^\top$ and $\boldsymbol{\Omega}$ are jointly diagonalisable. We start the discussion by simplifying the bias term. Consider the inner product term in eq. (E.20):

$$^1\!/_p\langle\boldsymbol{\beta}_\star,\boldsymbol{\Phi}^\top(\boldsymbol{I}_p+\nu\boldsymbol{R}(\nu;\boldsymbol{\Omega}))\boldsymbol{R}(\nu;\boldsymbol{\Omega})\boldsymbol{\Phi}\boldsymbol{\beta}_\star\rangle \tag{E.40}$$

Recalling the following Woodbury identities for a rectangular matrix $\boldsymbol{U}\in\mathbb{R}^{p\times d}$:

$$\boldsymbol{U}^\top(\lambda\boldsymbol{I}_p+\boldsymbol{U}\boldsymbol{U}^\top)^{-1}\boldsymbol{U}=\boldsymbol{I}_d-\lambda(\lambda\boldsymbol{I}_d+\boldsymbol{U}^\top\boldsymbol{U})^{-1} \tag{E.41}$$

$$\boldsymbol{U}^\top(\lambda\boldsymbol{I}_p+\boldsymbol{U}\boldsymbol{U}^\top)^{-2}\boldsymbol{U}=(\lambda\boldsymbol{I}_d+\boldsymbol{U}^\top\boldsymbol{U})^{-1}-\lambda(\lambda\boldsymbol{I}_d+\boldsymbol{U}^\top\boldsymbol{U})^{-2} \tag{E.42}$$

Defining the shorthand $\tilde{\nu}=\nu+b_\star^2$ to lighten the notation and using the above, we can write:

$$\begin{aligned}\boldsymbol{\Phi}^\top\boldsymbol{R}(\tilde{\nu};\boldsymbol{\Phi}\boldsymbol{\Phi}^\top)\boldsymbol{\Phi}&=\boldsymbol{I}_d-\tilde{\nu}(\tilde{\nu}\boldsymbol{I}_d+\boldsymbol{\Phi}^\top\boldsymbol{\Phi})^{-1}\\ &=\boldsymbol{I}_d-\tilde{\nu}\boldsymbol{R}(\tilde{\nu};\boldsymbol{\Phi}^\top\boldsymbol{\Phi})\end{aligned} \tag{E.43}$$

$$\begin{aligned}\boldsymbol{\Phi}^\top\boldsymbol{R}(\tilde{\nu};\boldsymbol{\Phi}\boldsymbol{\Phi}^\top)^2\boldsymbol{\Phi}&=(\tilde{\nu}\boldsymbol{I}_d+\boldsymbol{\Phi}^\top\boldsymbol{\Phi})^{-1}-\tilde{\nu}(\tilde{\nu}\boldsymbol{I}_d+\boldsymbol{\Phi}^\top\boldsymbol{\Phi})^{-2}\\ &=\boldsymbol{R}(\tilde{\nu};\boldsymbol{\Phi}^\top\boldsymbol{\Phi})-\tilde{\nu}\boldsymbol{R}(\tilde{\nu};\boldsymbol{\Phi}^\top\boldsymbol{\Phi})^2\end{aligned} \tag{E.44}$$

With this, we can write the inner product term:

$$\begin{aligned}\langle\boldsymbol{\beta}_\star,\boldsymbol{\Phi}^\top(\boldsymbol{I}_p+\nu\boldsymbol{R}(\nu;\boldsymbol{\Omega}))\boldsymbol{R}(\nu;\boldsymbol{\Omega})\boldsymbol{\Phi}\boldsymbol{\beta}_\star\rangle&=||\boldsymbol{\beta}_\star||_2^2-(\tilde{\nu}-\nu)\langle\boldsymbol{\beta}_\star,\boldsymbol{R}(\tilde{\nu};\boldsymbol{\Phi}^\top\boldsymbol{\Phi})\boldsymbol{\beta}_\star\rangle+\nu\tilde{\nu}\langle\boldsymbol{\beta}_\star,\boldsymbol{R}(\tilde{\nu};\boldsymbol{\Phi}^\top\boldsymbol{\Phi})^2\boldsymbol{\beta}_\star\rangle\\ &=d\rho-\langle\boldsymbol{\beta}_\star,\left(b_\star^2+\nu(\nu+b_\star^2)\boldsymbol{R}(\tilde{\nu};\boldsymbol{\Phi}^\top\boldsymbol{\Phi})\right)\boldsymbol{R}(\tilde{\nu};\boldsymbol{\Phi}^\top\boldsymbol{\Phi})\boldsymbol{\beta}_\star\rangle\end{aligned} \tag{E.45}$$

Therefore, the bias term can be simplified to:

$$\mathcal{B}(\alpha,\gamma,\lambda)=\frac{\alpha\gamma\,^1\!/_p\langle\boldsymbol{\beta}_\star,\left(b_\star^2+\nu(\nu+b_\star^2)\boldsymbol{R}(\nu+b_\star^2;\boldsymbol{\Phi}^\top\boldsymbol{\Phi})\right)\boldsymbol{R}(\nu+b_\star^2;\boldsymbol{\Phi}^\top\boldsymbol{\Phi})\boldsymbol{\beta}_\star\rangle}{\alpha-\tilde{\mathrm{df}}_2(\nu+b_\star^2;\boldsymbol{\Phi}\boldsymbol{\Phi}^\top)} \tag{E.46}$$

If moreover we assume the target weights are random: $\boldsymbol{\beta}_\star\sim\mathcal{N}(\boldsymbol{0},\boldsymbol{I}_d)$, we have concentration of the quadratic forms on the trace, giving us:

$$\mathcal{B}(\alpha,\gamma,\lambda)=\frac{\alpha\,^1\!/_d\,\mathrm{Tr}\left\{\left(b_\star^2+\nu(\nu+b_\star^2)\boldsymbol{R}(\nu+b_\star^2;\boldsymbol{\Phi}^\top\boldsymbol{\Phi})\right)\boldsymbol{R}(\nu+b_\star^2;\boldsymbol{\Phi}^\top\boldsymbol{\Phi})\right\}}{\alpha-\tilde{\mathrm{df}}_2(\nu+b_\star^2;\boldsymbol{\Phi}\boldsymbol{\Phi}^\top)} \tag{E.47}$$

Note that the trace in the enumerator in over a $\mathbb{R}^{d\times d}$ matrix, while the one in the denominator is over $\mathbb{R}^{p\times p}$.

## References


Ehsan Abbasi, Fariborz Salehi, and Babak Hassibi. Universality in learning from linear measurements. *Advances in Neural Information Processing Systems*, 32, 2019.

Fabián Aguirre-López, Silvio Franz, and Mauro Pastore. Random features and polynomial rules. *arXiv preprint arXiv:2402.10164*, 2024.

Benjamin Aubin, Antoine Maillard, Florent Krzakala, Nicolas Macris, Lenka Zdeborová, et al. The committee machine: Computational to statistical gaps in learning a two-layers neural network. *Advances in Neural Information Processing Systems*, 31, 2018.

Benjamin Aubin, Bruno Loureiro, Antoine Maillard, Florent Krzakala, and Lenka Zdeborová. The spiked matrix model with generative priors. *Advances in Neural Information Processing Systems*, 32, 2019.

Benjamin Aubin, Bruno Loureiro, Antoine Baker, Florent Krzakala, and Lenka Zdeborová. Exact asymptotics for phase retrieval and compressed sensing with random generative priors. In *Mathematical and Scientific Machine Learning*, pages 55–73. PMLR, 2020.

Francis Bach. High-dimensional analysis of double descent for linear regression with random projections. *SIAM Journal on Mathematics of Data Science*, 6(1):26–50, 2024.

Maria-Florina Balcan, Avrim Blum, and Santosh Vempala. Kernels as features: On kernels, margins, and low-dimensional mappings. *Machine Learning*, 65(1):79–94, 2006.

Afonso S Bandeira, Ahmed El Alaoui, Samuel Hopkins, Tselil Schramm, Alexander S Wein, and Ilias Zadik. The franz-parisi criterion and computational trade-offs in high dimensional statistics. *Advances in Neural Information Processing Systems*, 35:33831–33844, 2022.

Jean Barbier and Dmitry Panchenko. Strong replica symmetry in high-dimensional optimal bayesian inference. *Communications in mathematical physics*, 393(3):1199–1239, 2022.

Jean Barbier, Florent Krzakala, Nicolas Macris, Léo Miolane, and Lenka Zdeborová. Optimal errors and phase transitions in high-dimensional generalized linear models. *Proceedings of the National Academy of Sciences*, 116(12):5451–5460, 2019.

Mohsen Bayati and Andrea Montanari. The dynamics of message passing on dense graphs, with applications to compressed sensing. *IEEE Transactions on Information Theory*, 57(2):764–785, 2011.

Mikhail Belkin, Daniel Hsu, Siyuan Ma, and Soumik Mandal. Reconciling modern machine-learning practice and the classical bias–variance trade-off. *Proceedings of the National Academy of Sciences*, 116(32):15849–15854, 2019.

Gérard Ben Arous and Sandrine Péché. Universality of local eigenvalue statistics for some sample covariance matrices. *Communications on Pure and Applied Mathematics: A Journal Issued by the Courant Institute of Mathematical Sciences*, 58(10):1316–1357, 2005.

Raphael Berthier, Andrea Montanari, and Phan-Minh Nguyen. State evolution for approximate message passing with non-separable functions. *Information and Inference: A Journal of the IMA*, 9(1): 33–79, 2020.

Hans A Bethe. Statistical theory of superlattices. *Proceedings of the Royal Society of London. Series A-Mathematical and Physical Sciences*, 150(871):552–575, 1935.

Erwin Bolthausen. An iterative construction of solutions of the tap equations for the sherrington–kirkpatrick model. *Communications in Mathematical Physics*, 325(1):333–366, 2014.

David Bosch, Ashkan Panahi, and Babak Hassibi. Precise asymptotic analysis of deep random feature models. In *The Thirty Sixth Annual Conference on Learning Theory*, pages 4132–4179. PMLR, 2023a.

David Bosch, Ashkan Panahi, Ayca Ozcelikkale, and Devdatt Dubhashi. Random features model with general convex regularization: A fine grained analysis with precise asymptotic learning curves. In *International Conference on Artificial Intelligence and Statistics*, pages 11371–11414. PMLR, 2023b.

S.P. Boyd and L. Vandenberghe. *Convex Optimization*. Number pt. 1 in Berichte über verteilte messysteme. Cambridge University Press, 2004. ISBN 9780521833783. URL https://books.google.fr/books?id=mYm0bLd3fcoC.

Abdulkadir Canatar, Blake Bordelon, and Cengiz Pehlevan. Spectral bias and task-model alignment explain generalization in kernel regression and infinitely wide neural networks. *Nature communications*, 12(1):2914, 2021.

Andrea Caponnetto and Ernesto De Vito. Optimal rates for the regularized least-squares algorithm. *Foundations of Computational Mathematics*, 7:331–368, 2007.

Michael Celentano, Andrea Montanari, and Yuchen Wu. The estimation error of general first order methods. In *Conference on Learning Theory*, pages 1078–1141. PMLR, 2020.

Chen Cheng and Andrea Montanari. Dimension free ridge regression. *arXiv preprint arXiv:2210.08571*, 2022.

Tin Sum Cheng, Aurelien Lucchi, Anastasis Kratsios, and David Belius. Characterizing overfitting in kernel ridgeless regression through the eigenspectrum. *arXiv preprint arXiv:2402.01297*, 2024.

Lucas Clarté, Bruno Loureiro, Florent Krzakala, and Lenka Zdeborová. On double-descent in uncertainty quantification in overparametrized models. In *International Conference on Artificial Intelligence and Statistics*, pages 7089–7125. PMLR, 2023.

P. Curie. *Propriétés magnétiques des corps à diverses températures*. Dissertations. Gauthier-Villars et fils, 1895. URL https://books.google.fr/books?id=QhMywOm_yNsC.

Alex Damian, Loucas Pillaud-Vivien, Jason D Lee, and Joan Bruna. The computational complexity of learning gaussian single-index models. *arXiv preprint arXiv:2403.05529*, 2024.

Yatin Dandi, Ludovic Stephan, Florent Krzakala, Bruno Loureiro, and Lenka Zdeborová. Universality laws for gaussian mixtures in generalized linear models. *Advances in Neural Information Processing Systems*, 36, 2024.

Lee H. Dicker. Ridge regression and asymptotic minimax estimation over spheres of growing dimension. *Bernoulli*, 22(1):1 – 37, 2016. doi: 10.3150/14-BEJ609. URL https://doi.org/10.3150/14-BEJ609.

Rainer Dietrich, Manfred Opper, and Haim Sompolinsky. Statistical mechanics of support vector networks. *Phys. Rev. Lett.*, 82:2975–2978, Apr 1999. doi: 10.1103/PhysRevLett.82.2975. URL https://link.aps.org/doi/10.1103/PhysRevLett.82.2975.

Edgar Dobriban and Stefan Wager. High-dimensional asymptotics of prediction: Ridge regression and classification. *The Annals of Statistics*, 46(1):247–279, 2018.

David Donoho and Jared Tanner. Observed universality of phase transitions in high-dimensional geometry, with implications for modern data analysis and signal processing. *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences*, 367(1906):4273–4293, 2009.

David L Donoho, Arian Maleki, and Andrea Montanari. Message-passing algorithms for compressed sensing. *Proceedings of the National Academy of Sciences*, 106(45):18914–18919, 2009.

Sofiia Dubova, Yue M Lu, Benjamin McKenna, and Horng-Tzer Yau. Universality for the global spectrum of random inner-product kernel matrices in the polynomial regime. *arXiv preprint arXiv:2310.18280*, 2023.

Samuel Frederick Edwards and Phil W Anderson. Theory of spin glasses. *Journal of Physics F: Metal Physics*, 5(5):965, 1975.

László Erdős, Horng-Tzer Yau, and Jun Yin. Bulk universality for generalized wigner matrices. *Probability Theory and Related Fields*, 154(1):341–407, 2012.

Zhou Fan and Zhichao Wang. Spectra of the conjugate kernel and neural tangent kernel for linear-width neural networks. *Advances in neural information processing systems*, 33:7710–7721, 2020.

Silvio Franz, Giorgio Parisi, Maxime Sevelev, Pierfrancesco Urbani, and Francesco Zamponi. Universality of the sat-unsat (jamming) threshold in non-convex continuous constraint satisfaction problems. *SciPost Physics*, 2(3):019, 2017.

Marylou Gabrié, Andre Manoel, Clément Luneau, Nicolas Macris, Florent Krzakala, Lenka Zdeborová, et al. Entropy and mutual information in models of deep neural networks. *Advances in neural information processing systems*, 31, 2018.

David Gamarnik. The overlap gap property: A topological barrier to optimizing over random structures. *Proceedings of the National Academy of Sciences*, 118(41):e2108492118, 2021.

Elizabeth Gardner and Bernard Derrida. Three unfinished works on the optimal storage capacity of networks. *Journal of Physics A: Mathematical and General*, 22(12):1983, 1989.

Federica Gerace, Bruno Loureiro, Florent Krzakala, Marc Mézard, and Lenka Zdeborová. Generalisation error in learning with random features and the hidden manifold model. In *International Conference on Machine Learning*, pages 3452–3462. PMLR, 2020.

Federica Gerace, Florent Krzakala, Bruno Loureiro, Ludovic Stephan, and Lenka Zdeborová. Gaussian universality of perceptrons with random labels. *Physical Review E*, 109(3):034305, 2024.

Cédric Gerbelot and Raphaël Berthier. Graph-based approximate message passing iterations. *Information and Inference: A Journal of the IMA*, 12(4):2562–2628, 2023.

Sebastian Goldt, Marc Mézard, Florent Krzakala, and Lenka Zdeborová. Modeling the influence of data structure on learning in neural networks: The hidden manifold model. *Physical Review X*, 10 (4):041044, 2020.

Sebastian Goldt, Bruno Loureiro, Galen Reeves, Florent Krzakala, Marc Mézard, and Lenka Zdeborová. The gaussian equivalence of generative models for learning with shallow neural networks. In *Mathematical and Scientific Machine Learning*, pages 426–471. PMLR, 2022.

Trevor Hastie, Andrea Montanari, Saharon Rosset, and Ryan J. Tibshirani. Surprises in high-dimensional ridgeless least squares interpolation. *The Annals of Statistics*, 50(2):949 – 986, 2022. doi: 10.1214/21-AOS2133. URL https://doi.org/10.1214/21-AOS2133.

J.A. Hertz. *Introduction To The Theory Of Neural Computation*. CRC Press, 1991. ISBN 9780429979293. URL https://books.google.fr/books?id=NwpQDwAAQBAJ.

Hong Hu and Yue M Lu. Universality laws for high-dimensional learning with random features. *IEEE Transactions on Information Theory*, 69(3):1932–1964, 2022.

Hong Hu, Yue M Lu, and Theodor Misiakiewicz. Asymptotics of random feature regression beyond the linear scaling regime. *arXiv preprint arXiv:2403.08160*, 2024.

Yukito Iba. The nishimori line and bayesian statistics. *Journal of Physics A: Mathematical and General*, 32(21):3875, 1999.

Ernst Ising. Beitrag zur theorie des ferromagnetismus. *Zeitschrift für Physik*, 31(1):253–258, 1925. doi: 10.1007/BF02980577. URL https://doi.org/10.1007/BF02980577.

Kurt Johansson Johansson. Universality of the local spacing distribution in certain ensembles of hermitian wigner matrices. *Communications in Mathematical Physics*, 215:683–705, 2001.

Yoshiyuki Kabashima. A cdma multiuser detection algorithm on the basis of belief propagation. *Journal of Physics A: Mathematical and General*, 36(43):11111, 2003a.

Yoshiyuki Kabashima. Propagating beliefs in spin-glass models. *Journal of the Physical Society of Japan*, 72(7):1645–1649, 2003b.

Yoshiyuki Kabashima and David Saad. The belief in tap. In M. Kearns, S. Solla, and D. Cohn, editors, *Advances in Neural Information Processing Systems*, volume 11. MIT Press, 1998a. URL https://proceedings.neurips.cc/paper_files/paper/1998/file/7949e456002b28988d38185bd30e77fd-Paper.pdf.

Yoshiyuki Kabashima and David Saad. Belief propagation vs. tap for decoding corrupted messages. *Europhysics Letters*, 44(5):668, 1998b.

Noureddine El Karoui. The spectrum of kernel random matrices. *The Annals of Statistics*, 38(1):1 – 50, 2010. doi: 10.1214/08-AOS648. URL https://doi.org/10.1214/08-AOS648.

Noureddine El Karoui. Asymptotic behavior of unregularized and ridge-regularized high-dimensional robust regression estimators: rigorous results. *arXiv preprint arXiv:1311.2445*, 2013.

Satish Babu Korada and Andrea Montanari. Applications of the lindeberg principle in communications and statistical learning. *IEEE transactions on information theory*, 57(4):2440–2450, 2011.

Anders Krogh and John Hertz. A simple weight decay can improve generalization. *Advances in neural information processing systems*, 4, 1991.

Anders Krogh and John A Hertz. Generalization in a linear perceptron in the presence of noise. *Journal of Physics A: Mathematical and General*, 25(5):1135, 1992.

Florent Krzakala, Marc Mézard, François Sausset, YF Sun, and Lenka Zdeborová. Statistical-physics-based reconstruction in compressed sensing. *Physical Review X*, 2(2):021005, 2012a.

Florent Krzakala, Marc Mézard, Francois Sausset, Yifan Sun, and Lenka Zdeborová. Probabilistic reconstruction in compressed sensing: algorithms, phase diagrams, and threshold achieving matrices. *Journal of Statistical Mechanics: Theory and Experiment*, 2012(08):P08009, 2012b.

Gen Li and Yuting Wei. A non-asymptotic distributional theory of approximate message passing for sparse and robust regression. *arXiv preprint arXiv:2401.03923*, 2024.

Gen Li, Wei Fan, and Yuting Wei. Approximate message passing from random initialization with applications to z 2 synchronization. *Proceedings of the National Academy of Sciences*, 120(31): e2302930120, 2023.

Cosme Louart, Zhenyu Liao, and Romain Couillet. A random matrix approach to neural networks. *The Annals of Applied Probability*, 28(2):1190–1248, 2018.

Bruno Loureiro, Cedric Gerbelot, Hugo Cui, Sebastian Goldt, Florent Krzakala, Marc Mezard, and Lenka Zdeborová. Learning curves of generic features maps for realistic datasets with a teacher-student model. *Advances in Neural Information Processing Systems*, 34:18137–18151, 2021.

Yue M Lu and Horng-Tzer Yau. An equivalence principle for the spectrum of random inner-product kernel matrices with polynomial scalings. *arXiv preprint arXiv:2205.06308*, 2022.

Antoine Maillard, Bruno Loureiro, Florent Krzakala, and Lenka Zdeborová. Phase retrieval in high dimensions: Statistical and computational phase transitions. *Advances in Neural Information Processing Systems*, 33:11071–11082, 2020.

Andre Manoel, Florent Krzakala, Marc Mézard, and Lenka Zdeborová. Multi-layer generalized linear estimation. In *2017 IEEE International Symposium on Information Theory (ISIT)*, pages 2098–2102. IEEE, 2017.

Enzo Marinari, Giorgio Parisi, and Felix Ritort. Replica field theory for deterministic models. ii. a non-random spin glass with glassy behaviour. *Journal of Physics A: Mathematical and General*, 27 (23):7647, 1994.

Song Mei and Andrea Montanari. The generalization error of random features regression: Precise asymptotics and the double descent curve. *Communications on Pure and Applied Mathematics*, 75 (4):667–766, 2022.

Marc Mézard, Giorgio Parisi, and Miguel Angel Virasoro. *Spin glass theory and beyond: An Introduction to the Replica Method and Its Applications*, volume 9. World Scientific Publishing Company, 1987.

Marc Mézard, Giorgio Parisi, and Riccardo Zecchina. Analytic and algorithmic solution of random satisfiability problems. *Science*, 297(5582):812–815, 2002.

Theodor Misiakiewicz and Andrea Montanari. Six lectures on linearized neural networks. *arXiv preprint arXiv:2308.13431*, 2023.

Theodor Misiakiewicz and Basil Saeed. A non-asymptotic theory of kernel ridge regression: deterministic equivalents, test error, and gcv estimator. *arXiv preprint arXiv:2403.08938*, 2024.

Rémi Monasson and Riccardo Zecchina. Learning and generalization theories of large committee-machines. *Modern Physics Letters B*, 9(30):1887–1897, 1995.

Marco Mondelli and Andrea Montanari. Fundamental limits of weak recovery with applications to phase retrieval. In *Conference On Learning Theory*, pages 1445–1450. PMLR, 2018.

Andrea Montanari and Phan-Minh Nguyen. Universality of the elastic net error. In *2017 IEEE International Symposium on Information Theory (ISIT)*, pages 2338–2342. IEEE, 2017.

Andrea Montanari and Basil N Saeed. Universality of empirical risk minimization. In *Conference on Learning Theory*, pages 4310–4312. PMLR, 2022.

Hidetoshi Nishimori. Exact results and critical properties of the ising model with competing interactions. *Journal of Physics C: Solid State Physics*, 13(21):4071, 1980.

Ryan O'Donnell. *Analysis of Boolean Functions*. Cambridge University Press, 2014.

M. Opper and R. Urbanczik. Universal learning curves of support vector machines. *Phys. Rev. Lett.*, 86:4410–4413, May 2001. doi: 10.1103/PhysRevLett.86.4410. URL https://link.aps.org/doi/10.1103/PhysRevLett.86.4410.

Ashkan Panahi and Babak Hassibi. A universal analysis of large-scale regularized least squares solutions. *Advances in Neural Information Processing Systems*, 30, 2017.

Giorgio Parisi and Marc Potters. Mean-field equations for spin models with orthogonal interaction matrices. *Journal of Physics A: Mathematical and General*, 28(18):5267, 1995.

Giorgio Parisi and Tommaso Rizzo. Universality and deviations in disordered systems. *Phys. Rev. B*, 81:094201, Mar 2010. doi: 10.1103/PhysRevB.81.094201. URL https://link.aps.org/doi/10.1103/PhysRevB.81.094201.

Judea Pearl. Reverend bayes on inference engines: a distributed hierarchical approach. In *Proceedings of the Second AAAI Conference on Artificial Intelligence*, AAAI'82, pages 133–136. AAAI Press, 1982.

Jeffrey Pennington and Pratik Worah. Nonlinear random matrix theory for deep learning. *Advances in neural information processing systems*, 30, 2017.

Luca Pesce, Bruno Loureiro, Florent Krzakala, and Lenka Zdeborová. Subspace clustering in high-dimensions: Phase transitions & statistical-to-computational gap. *Advances in Neural Information Processing Systems*, 35:27087–27099, 2022.

Luca Pesce, Florent Krzakala, Bruno Loureiro, and Ludovic Stephan. Are gaussian data all you need? the extents and limits of universality in high-dimensional generalized linear estimation. In *International Conference on Machine Learning*, pages 27680–27708. PMLR, 2023.

Ali Rahimi and Benjamin Recht. Random features for large-scale kernel machines. In J. Platt, D. Koller, Y. Singer, and S. Roweis, editors, *Advances in Neural Information Processing Systems*, volume 20. Curran Associates, Inc., 2007. URL https://proceedings.neurips.cc/paper_files/paper/2007/file/013a006f03dbc5392effeb8f18fda755-Paper.pdf.

Ali Rahimi and Benjamin Recht. Weighted sums of random kitchen sinks: Replacing minimization with randomization in learning. In D. Koller, D. Schuurmans, Y. Bengio, and L. Bottou, editors, *Advances in Neural Information Processing Systems*, volume 21. Curran Associates, Inc., 2008. URL https://proceedings.neurips.cc/paper_files/paper/2008/file/0efe32849d230d7f53049ddc4a4b0c60-Paper.pdf.

Sundeep Rangan. Generalized approximate message passing for estimation with random linear mixing. In *2011 IEEE International Symposium on Information Theory Proceedings*, pages 2168–2172. IEEE, 2011.

Maria Refinetti, Sebastian Goldt, Florent Krzakala, and Lenka Zdeborová. Classifying high-dimensional gaussian mixtures: Where kernel methods fail and neural networks succeed. In *International Conference on Machine Learning*, pages 8936–8947. PMLR, 2021.

Cynthia Rush and Ramji Venkataramanan. Finite sample analysis of approximate message passing algorithms. *IEEE Transactions on Information Theory*, 64(11):7264–7286, 2018.

Shriram Sarvotham, Dror Baron, and Richard G Baraniuk. Compressed sensing reconstruction via belief propagation. *preprint*, 14, 2006.

Dominik Schröder, Hugo Cui, Daniil Dmitriev, and Bruno Loureiro. Deterministic equivalent and error universality of deep random features learning. In *International Conference on Machine Learning*, pages 30285–30320. PMLR, 2023.

Dominik Schröder, Daniil Dmitriev, Hugo Cui, and Bruno Loureiro. Asymptotics of learning with deep structured (random) features. *arXiv preprint arXiv:2402.13999*, 2024.

Henry Schwarze and John Hertz. Generalization in a large committee machine. *Europhysics Letters*, 20(4):375, 1992.

Ori Shental, Paul H Siegel, Jack K Wolf, Danny Bickson, and Danny Dolev. Gaussian belief propagation solver for systems of linear equations. In *2008 IEEE international symposium on information theory*, pages 1863–1867. IEEE, 2008.

Jack W Silverstein and Zhi Dong Bai. On the empirical distribution of eigenvalues of a class of large dimensional random matrices. *Journal of Multivariate analysis*, 54(2):175–192, 1995.

Terence Tao and Van Vu. Random matrices: Universality of local eigenvalue statistics. *Acta Mathematica*, 206(1):127 – 204, 2011. doi: 10.1007/s11511-011-0061-3. URL https://doi.org/10.1007/s11511-011-0061-3.

David J Thouless, Philip W Anderson, and Robert G Palmer. Solution of'solvable model of a spin glass'. *Philosophical Magazine*, 35(3):593–601, 1977.

Christos Thrampoulidis, Samet Oymak, and Babak Hassibi. Regularized linear regression: A precise analysis of the estimation error. In Peter Grunwald, Elad Hazan, and Satyen Kale, editors, *Proceedings of The 28th Conference on Learning Theory*, volume 40 of *Proceedings of Machine Learning Research*, pages 1683–1709, Paris, France, 03–06 Jul 2015. PMLR. URL https://proceedings.mlr.press/v40/Thrampoulidis15.html.

Umberto M Tomasini, Antonio Sclocchi, and Matthieu Wyart. Failure and success of the spectral bias prediction for Laplace kernel ridge regression: the case of low-dimensional data. In Kamalika Chaudhuri, Stefanie Jegelka, Le Song, Csaba Szepesvari, Gang Niu, and Sivan Sabato, editors, *Proceedings of the 39th International Conference on Machine Learning*, volume 162 of *Proceedings of Machine Learning Research*, pages 21548–21583. PMLR, 17–23 Jul 2022.

Emanuele Troiani, Yatin Dandi, Leonardo Defilippis, Lenka Zdeborová, Bruno Loureiro, and Florent Krzakala. Fundamental limits of weak learnability in high-dimensional multi-index models. *arXiv preprint arXiv:2405.15480*, 2024.

Pierre Weiss. L'hypothèse du champ moléculaire et la propriété ferromagnétique. *J. Phys. Theor. Appl.*, 6(1):661–690, 1907.

Denny Wu and Ji Xu. On the optimal weighted $\ell_2$ regularization in overparameterized linear regression. *Advances in Neural Information Processing Systems*, 33:10112–10123, 2020.

Lechao Xiao, Hong Hu, Theodor Misiakiewicz, Yue M Lu, and Jeffrey Pennington. Precise learning curves and higher-order scaling limits for dot product kernel regression. *Journal of Statistical Mechanics: Theory and Experiment*, 2023(11):114005, 2023.

Yuansheng Xiong, Chulan Kwon, and Jong-Hoon Oh. The storage capacity of a fully-connected committee machine. *Advances in Neural Information Processing Systems*, 10, 1997.

Ilias Zadik, Min Jae Song, Alexander S Wein, and Joan Bruna. Lattice-based methods surpass sum-of-squares in clustering. In *Conference on Learning Theory*, pages 1247–1248. PMLR, 2022.

Lenka Zdeborová and Florent Krzakala. Statistical physics of inference: Thresholds and algorithms. *Advances in Physics*, 65(5):453–552, 2016.

Tong Zhang. Learning bounds for kernel regression using effective data dimensionality. *Neural computation*, 17(9):2077–2098, 2005.