# Spin models and ML

## 1 Intro

Bla

## 2 Nishimori identity also known as Bayes rule

The main contribution of this section is the introduction of a very convenient notation that will be used throughout these notes.

**Definition 2.1**

*Let $(\boldsymbol{X}, \boldsymbol{Y})$ be a couple of random variables on a polish space. We denote $\langle\cdot\rangle$ the expectation with respect to $\mathbb{P}(\boldsymbol{X} = \cdot \,|\, \boldsymbol{Y})$ and $\mathbb{E}$ the expectation with respect to $(\boldsymbol{X}, \boldsymbol{Y})$.*

*A replica denoted by $\boldsymbol{x}$ is a random variable sampled (given $\boldsymbol{Y}$) from the distribution $\mathbb{P}(\boldsymbol{X} = \cdot \,|\, \boldsymbol{Y})$, independently of every other random variables. For $k \geq 1$, we denote $k$ replicas as $\boldsymbol{x}^{(1)}, \dots, \boldsymbol{x}^{(k)}$ being $k$ i.i.d. samples (given $\boldsymbol{Y}$) from the distribution $\mathbb{P}(\boldsymbol{X} = \cdot \,|\, \boldsymbol{Y})$, independently of every other random variables.*

Formally, we defined random variables $(\boldsymbol{X}, \boldsymbol{Y}, \boldsymbol{x}, \boldsymbol{x}^{(1)}, \dots, \boldsymbol{x}^{(k)})$ with distribution $P_{(X,Y)} \otimes P_{(X|Y)} \otimes \dots \otimes P_{(X|Y)}$ and a notation $\langle\cdot\rangle$ such that for all continuous bounded function $f$:

$$\langle f(\boldsymbol{X}, \boldsymbol{Y}, \boldsymbol{x}, \boldsymbol{x}^{(1)}, \dots, \boldsymbol{x}^{(k)})\rangle \stackrel{\text{def}}{=} \mathbb{E}\left[f(\boldsymbol{X}, \boldsymbol{Y}, \boldsymbol{x}, \boldsymbol{x}^{(1)}, \dots, \boldsymbol{x}^{(k)}) \,|\, \boldsymbol{Y}\right].$$

The basic properties of conditional expectation gives:

$$\begin{aligned} \langle \boldsymbol{X} \rangle &= \mathbb{E}\left[\boldsymbol{X} \,|\, \boldsymbol{Y}\right] \\ \langle f(\boldsymbol{X}) g(\boldsymbol{Y}) \rangle &= \langle f(\boldsymbol{X}) \rangle g(\boldsymbol{Y}) \\ \mathbb{E}\langle \boldsymbol{X} \rangle &= \mathbb{E}\left[\boldsymbol{X}\right]. \end{aligned}$$

Moreover, the independence assumption ensures that

$$\langle f(\boldsymbol{X}) g(\boldsymbol{x}) \rangle = \langle f(\boldsymbol{X}) \rangle \langle g(\boldsymbol{x}) \rangle.$$

Thanks to Bayes rule, we have $P_{(X,Y)} = P_Y \otimes P_{(X|Y)}$ so that $(\boldsymbol{X}, \boldsymbol{Y})$ has the same law as $(\boldsymbol{x}, \boldsymbol{Y})$ and

$$\langle \boldsymbol{x} \rangle = \mathbb{E}\left[\boldsymbol{x} \,|\, \boldsymbol{Y}\right] = \langle \boldsymbol{X} \rangle$$

This argument extends easily and $(\boldsymbol{Y}, \boldsymbol{x}^{(1)}, \dots, \boldsymbol{x}^{(k)})$ has the same law as $(\boldsymbol{Y}, \boldsymbol{x}^{(1)}, \dots, \boldsymbol{x}^{(k-1)}, \boldsymbol{X})$, so that we get the following convenient reformulation:

**Proposition 2.1 (*Nishimori identity*)**

*With the notations introduced in Definition 2.1, we have for all continuous bounded function $f$*

$$\left\langle f(\boldsymbol{Y}, \boldsymbol{x}^{(1)}, \dots, \boldsymbol{x}^{(k)}) \right\rangle = \left\langle f(\boldsymbol{Y}, \boldsymbol{x}^{(1)}, \dots, \boldsymbol{x}^{(k-1)}, \boldsymbol{X}) \right\rangle.$$

We will now use the Nishimori identity in the particular case of two random vectors $\boldsymbol{X}$ and $\boldsymbol{Y}$ that live respectively in $\mathbb{R}^n$ and $\mathbb{R}^m$. In this setting, we have useful relations:

$$\langle \boldsymbol{X}^T \boldsymbol{x} \rangle = \langle \boldsymbol{X} \rangle^T \langle \boldsymbol{x} \rangle = \| \langle \boldsymbol{x} \rangle \|^2,$$

where the first equality follows from the independence assumption and the second one from Nishimori identity. Moreover, we have

$$\mathbb{E}\left[ \boldsymbol{X}^T \langle \boldsymbol{x} \rangle \right] = \mathbb{E}\left[ \langle \boldsymbol{X}^T \boldsymbol{x} \rangle \right] = \mathbb{E}\left[ \| \langle \boldsymbol{x} \rangle \|^2 \right],$$

so that we obtain the classical result:

**Proposition 2.2**

*$\langle \boldsymbol{x} \rangle$ is the orthognal projection of $\boldsymbol{X}$ on $L^2(\boldsymbol{Y})$:*

$$\mathbb{E}\left[ (\boldsymbol{X} - \langle \boldsymbol{x} \rangle)^T \langle \boldsymbol{x} \rangle \right] = 0.$$

As a direct consequence, we get another classical result about the problem of estimating $\boldsymbol{X}$ when observing $\boldsymbol{Y}$ where the performance of an estimator $\widehat{\theta}$ (i.e. a measurable function of the observations $\boldsymbol{Y}$) is given by its Mean Squared Error $\mathrm{MSE}(\widehat{\theta}) = \mathbb{E}\|\boldsymbol{X} - \widehat{\theta}(\boldsymbol{Y})\|^2$.

**Definition 2.2**

*For a signal $\boldsymbol{X}$ and observations $\boldsymbol{Y}$, the Minimum Mean Squared Error is defined by*

$$\mathrm{MMSE} \overset{\text{def}}{=} \min_{\widehat{\theta}} \mathrm{MSE}(\widehat{\theta}) = \mathbb{E}\left[ \| \boldsymbol{X} - \langle \boldsymbol{x} \rangle \|^2 \right] = \mathbb{E}\left[ \|\boldsymbol{X}\|^2 \right] - \mathbb{E}\left[ \langle \boldsymbol{x}^T \boldsymbol{X} \rangle \right],$$

*where the minimum is taken over all measurable function $\widehat{\theta}$ of the observations $\boldsymbol{Y}$. The optimal estimator (in term of Mean Squared Error) is the posterior mean of the signal $\boldsymbol{X}$ given $\boldsymbol{Y}$.*

**Proof.** Using $\langle \boldsymbol{X}^T \widehat{\theta}(\boldsymbol{Y}) \rangle = \langle \boldsymbol{X} \rangle^T \widehat{\theta}(\boldsymbol{Y})$, we have

$$\mathbb{E}\left[ \left\| \boldsymbol{X} - \widehat{\theta}(\boldsymbol{Y}) \right\|^2 \right] = \mathbb{E}\left[ \|\boldsymbol{X}\|^2 \right] - 2\mathbb{E}\left[ \langle \boldsymbol{X} \rangle^T \widehat{\theta}(\boldsymbol{Y}) \right] + \mathbb{E}\left[ \left\| \widehat{\theta}(\boldsymbol{Y}) \right\|^2 \right]$$

and

$$\mathbb{E}\left[ \left\| \langle \boldsymbol{X} \rangle - \widehat{\theta}(\boldsymbol{Y}) \right\|^2 \right] = \mathbb{E}\left[ \| \langle \boldsymbol{X} \rangle \|^2 \right] - 2\mathbb{E}\left[ \langle \boldsymbol{X} \rangle^T \widehat{\theta}(\boldsymbol{Y}) \right] + \mathbb{E}\left[ \left\| \widehat{\theta}(\boldsymbol{Y}) \right\|^2 \right] \geq 0,$$

we deduce

$$\begin{aligned}
\mathbb{E}\left[ \left\| \boldsymbol{X} - \widehat{\theta}(\boldsymbol{Y}) \right\|^2 \right] &\geq \mathbb{E}\left[ \|\boldsymbol{X}\|^2 \right] - \mathbb{E}\left[ \| \langle \boldsymbol{X} \rangle \|^2 \right] \\
&= \mathbb{E}\left[ \|\boldsymbol{X}\|^2 \right] - 2\mathbb{E}\left[ \langle \boldsymbol{X} \rangle^T \langle \boldsymbol{X} \rangle \right] + \mathbb{E}\left[ \| \langle \boldsymbol{X} \rangle \|^2 \right] \\
&= \mathbb{E}\left[ \| \boldsymbol{X} - \langle \boldsymbol{X} \rangle \|^2 \right] = \mathbb{E}\left[ \| \boldsymbol{X} - \langle \boldsymbol{x} \rangle \|^2 \right],
\end{aligned}$$

and the last equality follows from Proposition 2.2. □

**Example 2.1.** *In some particular cases, explicit computations can be done. The additive Gaussian scalar channel with Gaussian prior is a very simple case. We define*

$$Y = \sqrt{\lambda} X + Z \,, \tag{2.1}$$

*where $Z \sim \mathcal{N}(0,1)$ and $X$ is sampled from a distribution $P_0 = \mathcal{N}(0,1)$ over $\mathbb{R}$, independently of $Z$. We can compute the densities as $p(x) = \mathcal{N}(x|0,1)$, $p(y|x) = \mathcal{N}(y|\sqrt{\lambda}x, 1)$ so that $p(y) = \mathcal{N}(y|0, 1+\lambda)$ and the law of the replica is also Gaussian with $p(x|y) = \mathcal{N}\left(x|\frac{\sqrt{\lambda}}{1+\lambda}y, \frac{1}{1+\lambda}\right)$. In particular, we have*

$$\langle x \rangle = \frac{\sqrt{\lambda}}{1+\lambda} Y = \frac{\lambda}{1+\lambda} X + \frac{\sqrt{\lambda}}{1+\lambda} Z, \text{ and, } \mathbb{E}\left[\langle Xx \rangle\right] = \frac{\lambda}{1+\lambda}.$$

*In particular, we have* $\mathrm{MMSE}(\lambda) = \frac{1}{1+\lambda}$.

# 3 Bayesian inference with Gaussian additive noise

We start with the following model:

$$\boldsymbol{Y} = \sqrt{\lambda}\, \boldsymbol{X} + \boldsymbol{Z} \,, \tag{3.1}$$

where the signal $\boldsymbol{X}$ is sampled according to some probability distribution $P_X$ over $\mathbb{R}^n$, and where the noise $\boldsymbol{Z} = (Z_1, \dots, Z_n) \overset{\text{i.i.d.}}{\sim} \mathcal{N}(0,1)$ is independent from $\boldsymbol{X}$. The parameter $\lambda \geq 0$ plays the role of a signal-to-noise ratio. We assume that $P_X$ admits a finite second moment: $\mathbb{E}\|\boldsymbol{X}\|^2 < \infty$. No other assumption is made on $P_X$. In particular, the fact that the noise is i.i.d. in the dimension $n$ will be crucial in the computation below but no such assumption is made on the signal which is allowed to have any complicated structure component-wise.

Given the observation channel (3.1), the goal of the statistician is to estimate $\boldsymbol{X}$ given the observations $\boldsymbol{Y}$. We assume to be in the "Bayes-optimal" setting, where the statistician knows all the parameters of the inference model, that is the prior distribution $P_X$ and the signal-to-noise ratio $\lambda$. As will be clear below, the Mean Squared Error (MSE) is the natural measure of performance when dealing with Gaussian noise.

Following Definition 2.2, we see that a natural object to study is the posterior distribution of $\boldsymbol{X}$. We have for the joint distribution:

$$dP_{(X,Y)}(\boldsymbol{x}, \boldsymbol{y}) \propto dP_X(\boldsymbol{x}) e^{\frac{\|\boldsymbol{y} - \sqrt{\lambda}\boldsymbol{x}\|^2}{2}} d\boldsymbol{y}$$

By Bayes rule, the posterior distribution of $\boldsymbol{X}$ given $\boldsymbol{Y}$ is

$$dP(\boldsymbol{x} \,|\, \boldsymbol{Y}) = \frac{1}{\mathcal{Z}(\lambda, \boldsymbol{Y})} e^{H_{\lambda,\boldsymbol{Y}}(\boldsymbol{x})} dP_X(\boldsymbol{x}) \,, \tag{3.2}$$

where

$$H_{\lambda,\boldsymbol{Y}}(\boldsymbol{x}) = \sqrt{\lambda}\, \boldsymbol{x}^\intercal \boldsymbol{Y} - \frac{\lambda}{2}\|\boldsymbol{x}\|^2 = \sqrt{\lambda}\, \boldsymbol{x}^\intercal \boldsymbol{Z} + \lambda\, \boldsymbol{x}^\intercal \boldsymbol{X} - \frac{\lambda}{2}\|\boldsymbol{x}\|^2 \,.$$

**Definition 3.1**

*$H_{\lambda,\boldsymbol{Y}}$ is called the Hamiltonian*[1] *and the normalizing constant*

$$\mathcal{Z}(\lambda, \boldsymbol{Y}) = \int dP_X(\boldsymbol{x}) e^{H_{\lambda,\boldsymbol{Y}}(\boldsymbol{x})}$$

*is called the partition function.*

Expectations with respect the posterior distribution (3.2) will be denoted by the Gibbs brackets $\langle \cdot \rangle_\lambda$:

$$\Big\langle f(\boldsymbol{x}) \Big\rangle_\lambda = \mathbb{E}\Big[ f(\boldsymbol{X}) | \boldsymbol{Y} \Big] = \frac{1}{\mathcal{Z}(\lambda, \boldsymbol{Y})} \int dP_X(\boldsymbol{x}) f(\boldsymbol{x}) e^{H_{\lambda,\boldsymbol{Y}}(\boldsymbol{x})} \,,$$

for any measurable function $f$ such that $f(\boldsymbol{X})$ is integrable.

**Definition 3.2**

*$F(\lambda) = \mathbb{E} \log \mathcal{Z}(\lambda, \boldsymbol{Y})$ is called the free energy*[2]*. It is related to the mutual information between $\boldsymbol{X}$ and $\boldsymbol{Y}$ by*

$$F(\lambda) = \frac{\lambda}{2} \mathbb{E} \|\boldsymbol{X}\|^2 - I(\boldsymbol{X}; \boldsymbol{Y}) \,. \tag{3.3}$$

**Proof.** The mutual information $I(\boldsymbol{X}; \boldsymbol{Y})$ is defined as the Kullback-Leibler divergence between $P_{(X,Y)}$, the joint distribution of $(\boldsymbol{X}, \boldsymbol{Y})$ and $P_X \otimes P_Y$ the product of the marginal distributions of $\boldsymbol{X}$ and $\boldsymbol{Y}$. $P_{(X,Y)}$ is absolutely continuous with respect to $P_X \otimes P_Y$ with Radon-Nikodym derivative:

$$\frac{dP_{(X,Y)}}{dP_X \otimes P_Y}(\boldsymbol{X}, \boldsymbol{Y}) = \frac{\exp\left(-\frac{1}{2}\|\boldsymbol{Y} - \sqrt{\lambda}\boldsymbol{X}\|^2\right)}{\int \exp\left(-\frac{1}{2}\|\boldsymbol{Y} - \sqrt{\lambda}\boldsymbol{x}\|^2\right) dP_X(\boldsymbol{x})} \,.$$

Therefore

$$\begin{aligned}
I(\boldsymbol{X}; \boldsymbol{Y}) &= \mathbb{E} \log \left( \frac{dP_{(X,Y)}}{dP_X \otimes P_Y}(\boldsymbol{X}, \boldsymbol{Y}) \right) \\
&= -\mathbb{E} \log \int dP_X(\boldsymbol{x}) \exp\left( \sqrt{\lambda} \boldsymbol{x}^\mathsf{T} \boldsymbol{Y} - \sqrt{\lambda} \boldsymbol{X}^\mathsf{T} \boldsymbol{Y} - \frac{\lambda}{2}\|\boldsymbol{x}\|^2 + \frac{\lambda}{2}\|\boldsymbol{X}\|^2 \right) \\
&= -F(\lambda) + \sqrt{\lambda} \mathbb{E} \boldsymbol{X}^T \boldsymbol{Y} - \frac{\lambda}{2} \mathbb{E} \|\boldsymbol{X}\|^2 \\
&= -F(\lambda) + \frac{\lambda}{2} \mathbb{E} \|\boldsymbol{X}\|^2 \,,
\end{aligned}$$

since $\sqrt{\lambda} \boldsymbol{X}^T \boldsymbol{Y} = \lambda \|\boldsymbol{X}\|^2 + \sqrt{\lambda} \boldsymbol{X}^T \boldsymbol{Z}$. □

We state now two basic properties of the MMSE. A more detailed analysis can be found in [**?**, **?**].

**Proposition 3.1**

*$\lambda \mapsto \mathrm{MMSE}(\lambda)$ is non-increasing over $\mathbb{R}_{\geq 0}$. Moreover*

- $\mathrm{MMSE}(0) = \mathbb{E}\|\boldsymbol{X} - \mathbb{E}[\boldsymbol{X}]\|^2$,
- $\mathrm{MMSE}(\lambda) \xrightarrow[\lambda \to +\infty]{} 0$.

[1] According to the physics convention, this should be minus the Hamiltonian, since a physical system tries to minimize its energy. However, we chose here to remove it for simplicity.

[2] This is in fact minus the free energy, but we chose to remove the minus sign for simplicity.

**Proof.** Let $0 < \lambda_2 < \lambda_1$. Define $\Delta_1 = \lambda_1^{-1}$, $\Delta_2 = \lambda_2^{-1}$ and

$$\begin{cases} \boldsymbol{Y}_1 = \boldsymbol{X} + \sqrt{\Delta_1}\boldsymbol{Z}_1 \\ \boldsymbol{Y}_2 = \boldsymbol{X} + \sqrt{\Delta_1}\boldsymbol{Z}_1 + \sqrt{\Delta_2 - \Delta_1}\boldsymbol{Z}_2\,, \end{cases}$$

where $\boldsymbol{X} \sim P_X$ is independent from $\boldsymbol{Z}_1, \boldsymbol{Z}_2 \overset{\text{i.i.d.}}{\sim} \mathcal{N}(0, \mathrm{Id}_n)$. Now, by independence between $(\boldsymbol{X}, \boldsymbol{Y}_1)$ and $\boldsymbol{Z}_2$ we have

$$\begin{aligned} \mathrm{MMSE}(\lambda_1) &= \mathbb{E}\,\|\boldsymbol{X} - \mathbb{E}[\boldsymbol{X}|\boldsymbol{Y}_1]\|^2 = \mathbb{E}\,\|\boldsymbol{X} - \mathbb{E}[\boldsymbol{X}|\boldsymbol{Y}_1, \boldsymbol{Z}_2]\|^2 = \mathbb{E}\,\|\boldsymbol{X} - \mathbb{E}[\boldsymbol{X}|\boldsymbol{Y}_1, \boldsymbol{Y}_2]\|^2 \\ &\le \mathbb{E}\,\|\boldsymbol{X} - \mathbb{E}[\boldsymbol{X}|\boldsymbol{Y}_2]\|^2 = \mathrm{MMSE}(\lambda_2)\,. \end{aligned}$$

Next, notice that

$$\mathrm{MMSE}(\lambda_1) = \mathbb{E}\,\|\boldsymbol{X} - \mathbb{E}[\boldsymbol{X}|\boldsymbol{Y}_1]\|^2 \le \mathbb{E}\,\|\boldsymbol{X} - \mathbb{E}[\boldsymbol{X}]\|^2 = \mathrm{MMSE}(0)\,. \tag{3.4}$$

This shows that the MMSE is non-increasing on $\mathbb{R}_{\ge 0}$. The first point is obvious while the second follows from:

$$0 \le \mathrm{MMSE}(\lambda) = \mathbb{E}\|\boldsymbol{X} - \mathbb{E}[\boldsymbol{X}|\boldsymbol{Y}]\|^2 \le \mathbb{E}\|\boldsymbol{X} - \frac{1}{\sqrt{\lambda}}\boldsymbol{Y}\|^2 = \frac{n}{\lambda} \xrightarrow[\lambda\to+\infty]{} 0\,.$$

□

**Proposition 3.2**

*$\lambda \mapsto \mathrm{MMSE}(\lambda)$ is continuous over $\mathbb{R}_{\ge 0}$.*

The proof of Proposition 3.2 can be found in Appendix 8.1.

We present now the very useful "I-MMSE" relation from [**?**]. This relation was previously known (under a different formulation) as "de Brujin identity" see [**?**, Equation 2.12].

**Proposition 3.3**

*For all $\lambda \ge 0$,*

$$\frac{\partial}{\partial\lambda} I(\boldsymbol{X}; \boldsymbol{Y}) = \frac{1}{2}\mathrm{MMSE}(\lambda) \qquad \text{and} \qquad F'(\lambda) = \frac{1}{2}\mathbb{E}\langle \boldsymbol{x}^\mathsf{T}\boldsymbol{X}\rangle_\lambda = \frac{1}{2}\Big(\mathbb{E}\|\boldsymbol{X}\|^2 - \mathrm{MMSE}(\lambda)\Big)\,. \tag{3.5}$$

*$F$ thus is a convex, differentiable, non-decreasing, and $\frac{1}{2}\mathbb{E}\|\boldsymbol{X}\|^2$-Lipschitz function over $\mathbb{R}_{\ge 0}$. If $P_X$ is not a Dirac mass, then $F$ is strictly convex.*

TBC...

Proposition 3.3 is proved in Appendix **??**. Proposition 3.3 reduces the computation of the MMSE to the computation of the free energy. This will be particularly useful because the free energy $F$ is much easier to handle than the MMSE.

We end this section with the simplest model of the form (3.1), namely the additive Gaussian scalar channel:

$$Y = \sqrt{\lambda} X + Z\,, \tag{3.6}$$

where $Z \sim \mathcal{N}(0,1)$ and $X$ is sampled from a distribution $P_0$ over $\mathbb{R}$, independently of $Z$. The corresponding free energy and the MMSE are respectively

$$\psi_{P_0}(\lambda) = \mathbb{E} \log \int dP_0(x) e^{\sqrt{\lambda} Y x - \lambda x^2/2} \quad \text{and} \quad \mathrm{MMSE}_{P_0}(\lambda) = \mathbb{E}\Big[\Big(X - \mathbb{E}[X|Y]\Big)^2\Big]\,. \tag{3.7}$$

The study of this simple inference channel will be very useful in the following, because we will see that the inference problems that we are going to study enjoy asymptotically a "decoupling principle" that reduces them to scalar channels like (3.6).

Let us compute the mutual information and the MMSE for particular choices of prior distributions:

**Example 3.1** (Gaussian prior: $P_0 = \mathcal{N}(0,1)$)**.** *In that case* $\mathbb{E}[X|Y]$ *is simply the orthogonal projection of* $X$ *on* $Y$*:*

$$\mathbb{E}[X|Y] = \frac{\mathbb{E}[XY]}{\mathbb{E}[Y^2]} Y = \frac{\sqrt{\lambda}}{1+\lambda} Y.$$

*One deduces* $\mathrm{MMSE}_{P_0}(\lambda) = \frac{1}{1+\lambda}$*. Using* (3.5)*, we get* $I(X;Y) = \frac{1}{2}\log(1+\lambda)$ *and* $\psi_{P_0}(\lambda) = \frac{1}{2}\Big(\lambda - \log(1+\lambda)\Big)$*.*

**Remark 3.1** (Worst-case prior)**.** *Let* $P_0$ *be a probability distribution on* $\mathbb{R}$ *with unit second moment* $\mathbb{E}_{P_0}[X^2] = 1$*. By considering the estimator* $\widehat{x} = \frac{\sqrt{\lambda}}{1+\lambda} Y$*, one obtain* $\mathrm{MMSE}_{P_0}(\lambda) \leq \frac{1}{1+\lambda}$*. We conclude:*

$$\sup_{P_0} \mathrm{MMSE}_{P_0}(\lambda) = \frac{1}{1+\lambda} \qquad \textit{and} \qquad \inf_{P_0} \psi_{P_0}(\lambda) = \frac{1}{2}(\lambda - \log(1+\lambda)),$$

*where the supremum and infimum are both over the probability distributions that have unit second moment. The standard normal distribution* $P_0 = \mathcal{N}(0,1)$ *achieves both extrema.*

**Example 3.2** (Rademacher prior: $P_0 = \frac{1}{2}\delta_{+1} + \frac{1}{2}\delta_{-1}$)**.** *We compute* $\psi_{P_0}(\lambda) = \mathbb{E}\log\cosh(\sqrt{\lambda}Z + \lambda) - \frac{\lambda}{2}$ *and* $I(X;Y) = \lambda - \mathbb{E}\log\cosh(\sqrt{\lambda}Z + \lambda)$*. The I-MMSE relation gives*

$$\begin{aligned}
\frac{1}{2}\mathrm{MMSE}(\lambda) = \frac{\partial}{\partial\lambda} I(X;Y) &= 1 - \mathbb{E}\Big[\Big(\frac{1}{2\sqrt{\lambda}} Z + 1\Big)\tanh\Big(\sqrt{\lambda}Z + \lambda\Big)\Big] \\
&= 1 - \mathbb{E}\tanh(\sqrt{\lambda}Z + \lambda) - \frac{1}{2}\mathbb{E}\tanh'(\sqrt{\lambda}Z + \lambda) \\
&= \frac{1}{2} - \mathbb{E}\tanh(\sqrt{\lambda}Z + \lambda) + \frac{1}{2}\mathbb{E}\tanh^2(\sqrt{\lambda}Z + \lambda)
\end{aligned}$$

*where we used Gaussian integration by parts. Since by the Nishimori property* $\mathbb{E}\langle xX\rangle_\lambda = \mathbb{E}\langle x\rangle_\lambda^2$*, one has* $\mathbb{E}\tanh(\sqrt{\lambda}Z + \lambda) = \mathbb{E}\tanh^2(\sqrt{\lambda}Z + \lambda)$ *and therefore* $\mathrm{MMSE}(\lambda) = 1 - \mathbb{E}\tanh(\sqrt{\lambda}Z + \lambda)$*.*

# 4 Bayes-optimal inference

We introduce in this chapter some general properties of Bayes-optimal inference, that will be used repeatedly in the sequel. Let us first define what we mean by *Bayes-optimal inference*.

We consider a statistical problem where we would like to recover a signal vector $\boldsymbol{X} \in \mathbb{R}^n$ from some observations $\boldsymbol{Y} \in \mathbb{R}^m$. We assume that $(\boldsymbol{X}, \boldsymbol{Y})$ is drawn from some probability distribution $\mu$ over $\mathbb{R}^n \times \mathbb{R}^m$. Given a performance criterion, a Bayes-optimal estimator (or simply Bayes estimator) is an estimator of $\boldsymbol{X}$ given $\boldsymbol{Y}$ that achieves the best performance for this criterion. For instance if we measure the performance of an estimator $\widehat{\boldsymbol{x}}$ by its mean square error $\mathrm{MSE}(\widehat{\boldsymbol{x}}) = \mathbb{E}\|\boldsymbol{X} - \widehat{\boldsymbol{x}}(\boldsymbol{Y})\|^2$, then the Bayes-optimal estimator is simply the posterior mean $\widehat{\boldsymbol{x}}^{\mathrm{Bayes}}(\boldsymbol{Y}) = \mathbb{E}[\boldsymbol{X}|\boldsymbol{Y}]$.

The goal of this chapter is to present some general properties of Bayes-optimal estimators. In Section 5 we introduce what we will call (according to the statistical physics terminology) the "Nishimori identity" which is nothing more than a rewriting of Bayes rule. In Section 6 we will study the links between various natural performance metrics for estimators and show that they are in some sense equivalent. In Sections 3 we analyse the special case where $\boldsymbol{Y} = \sqrt{\lambda}\boldsymbol{X} + \boldsymbol{Z}$, where $\lambda \geq 0$ and $\boldsymbol{Z}$ is some Gaussian noise. This is the starting point of the study of the "spiked" matrix and tensor models. Finally we consider in Section 7 a simple example to illustrate the tools of this chapter.

# 5 The Nishimori identity

In order to analyze Bayes-optimal estimators, we will need to examine the posterior distribution of $\boldsymbol{X}$ given $\boldsymbol{Y}$. To do so we will often consider i.i.d. samples $\boldsymbol{x}^{(1)}, \dots, \boldsymbol{x}^{(k)}$ from the posterior distribution $P(\cdot \,|\, \boldsymbol{Y})$, independently of everything else. Such samples are called replicas. The (obvious) identity below (which is simply Bayes rule) is named after the works of Nishimori [**?**, **?**] on "gauge-symmetric" spin glasses. It states that the planted solution $\boldsymbol{X}$ behaves like a replica.

**Proposition 5.1 (*Nishimori identity*)**

*Let $(\boldsymbol{X}, \boldsymbol{Y})$ be a couple of random variables on a polish space. Let $k \geq 1$ and let $\boldsymbol{x}^{(1)}, \dots, \boldsymbol{x}^{(k)}$ be $k$ i.i.d. samples (given $\boldsymbol{Y}$) from the distribution $\mathbb{P}(\boldsymbol{X} = \cdot \,|\, \boldsymbol{Y})$, independently of every other random variables. Let us denote $\langle \cdot \rangle$ the expectation with respect to $\mathbb{P}(\boldsymbol{X} = \cdot \,|\, \boldsymbol{Y})$ and $\mathbb{E}$ the expectation with respect to $(\boldsymbol{X}, \boldsymbol{Y})$. Then, for all continuous bounded function $f$*

$$\mathbb{E}\Big\langle f(\boldsymbol{Y}, \boldsymbol{x}^{(1)}, \dots, \boldsymbol{x}^{(k)})\Big\rangle = \mathbb{E}\Big\langle f(\boldsymbol{Y}, \boldsymbol{x}^{(1)}, \dots, \boldsymbol{x}^{(k-1)}, \boldsymbol{X})\Big\rangle\,.$$

**Proof.** It is equivalent to sample the couple $(\boldsymbol{X}, \boldsymbol{Y})$ according to its joint distribution or to sample first $\boldsymbol{Y}$ according to its marginal distribution and then to sample $\boldsymbol{X}$ conditionally to $\boldsymbol{Y}$ from its conditional distribution $\mathbb{P}(\boldsymbol{X} = \cdot \,|\, \boldsymbol{Y})$. Thus the $(k+1)$-tuple $(\boldsymbol{Y}, \boldsymbol{x}^{(1)}, \dots, \boldsymbol{x}^{(k)})$ is equal in law to $(\boldsymbol{Y}, \boldsymbol{x}^{(1)}, \dots, \boldsymbol{x}^{(k-1)}, \boldsymbol{X})$. □

# 6 Performance measure and optimal estimators

We consider two random vectors $\boldsymbol{X}$ and $\boldsymbol{Y}$ that live respectively in $\mathbb{R}^n$ and $\mathbb{R}^m$. We assume (for simplicity) that $\|\boldsymbol{X}\| = 1$ almost surely. As explained above, given the observations $\boldsymbol{Y}$, our goal is to estimate $\boldsymbol{X}$ with an estimator $\widehat{\boldsymbol{x}}(\boldsymbol{Y})$. In order to evaluate the performance of such an estimator, what criterion should we take?

The probably most natural way to characterize the performance of $\widehat{\boldsymbol{x}}$ is by its mean-squared error:

$$\mathrm{MSE}(\widehat{\boldsymbol{x}}) = \mathbb{E}\|\boldsymbol{X} - \widehat{\boldsymbol{x}}(\boldsymbol{Y})\|^2.$$

By Pythagorean theorem, we know that the optimal estimator which respect to this metric is the posterior mean $\widehat{\boldsymbol{x}}(\boldsymbol{Y}) = \mathbb{E}[\boldsymbol{X}|\boldsymbol{Y}]$ which achieves the minimal mean square error:

$$\mathrm{MMSE}(\boldsymbol{X}|\boldsymbol{Y}) \stackrel{\text{def}}{=} \mathbb{E}\|\boldsymbol{X} - \mathbb{E}[\boldsymbol{X}|\boldsymbol{Y}]\|^2. \tag{6.1}$$

However, the MSE is not always an appropriate criterion. Indeed in many cases it is only possible to recover $\boldsymbol{X}$ up to its sign: think for instance of the Spiked Wigner Model $\boldsymbol{Y} = \boldsymbol{X}\boldsymbol{X}^\mathsf{T} + \text{Noise}$ with $\boldsymbol{X} \sim \mathrm{Unif}(\mathbb{S}^{n-1})$. In such case, $\mathbb{E}[\boldsymbol{X}|\boldsymbol{Y}] = 0$: the best estimator in term of MSE does not even depend on the observations $\boldsymbol{Y}$!

For this kind of problems one should rather consider the correlation (also known as cosine similarity) *in absolute value* between $\widehat{\boldsymbol{x}}$ and the signal $\boldsymbol{X}$:

$$\sup_{\|\widehat{\boldsymbol{x}}\|=1} \mathbb{E}\Big[|(\widehat{\boldsymbol{x}}(\boldsymbol{Y}); \boldsymbol{X})|\Big] \qquad \text{or} \qquad \sup_{\|\widehat{\boldsymbol{x}}\|=1} \mathbb{E}\Big[(\widehat{\boldsymbol{x}}(\boldsymbol{Y}); \boldsymbol{X})^2\Big], \tag{6.2}$$

where $(\cdot;\cdot)$ denotes the Euclidean inner product and where the suprema are taken over all estimators $\widehat{\boldsymbol{x}} : \mathbb{R}^m \to \mathbb{S}^{n-1}$.

Let us introduce some notations. We will use the Gibbs Bracket $\langle\cdot\rangle$ to write expectations with respect to the posterior distribution of $\boldsymbol{X}$ given $\boldsymbol{Y}$:

$$\langle f(\boldsymbol{x})\rangle = \mathbb{E}[f(\boldsymbol{X})|\boldsymbol{Y}],$$

for all measurable function $f$ such that $f(\boldsymbol{X})$ is integrable. In particular, we will be interested by the $n \times n$ positive semi-definite (random) matrix:

$$\boldsymbol{M} \stackrel{\text{def}}{=} \langle \boldsymbol{x}\boldsymbol{x}^\mathsf{T}\rangle = \mathbb{E}[\boldsymbol{X}\boldsymbol{X}^\mathsf{T}|\boldsymbol{Y}]. \tag{6.3}$$

$\boldsymbol{M}$ is the Bayes-optimal estimator (in terms of mean square error) for estimating the matrix $\boldsymbol{X}\boldsymbol{X}^\mathsf{T}$:

$$\mathrm{MMSE}(\boldsymbol{X}\boldsymbol{X}^\mathsf{T}|\boldsymbol{Y}) = \mathbb{E}\|\boldsymbol{X}\boldsymbol{X}^\mathsf{T} - \mathbb{E}[\boldsymbol{X}\boldsymbol{X}^\mathsf{T}|\boldsymbol{Y}]\|^2. \tag{6.4}$$

An easy computation gives $\mathrm{MMSE}(\boldsymbol{X}\boldsymbol{X}^\mathsf{T}|\boldsymbol{Y}) = 1 - \mathbb{E}[\mathrm{Tr}(\boldsymbol{M}^2)]$. The matrix $\boldsymbol{M}$ is also related to the second quantity of (6.2) through its largest eigenvalue $\lambda_{\max}(\boldsymbol{M})$:

**Lemma 6.1**

$$\sup_{\|\widehat{\boldsymbol{x}}\|=1} \mathbb{E}\Big[(\widehat{\boldsymbol{x}}(\boldsymbol{Y}); \boldsymbol{X})^2\Big] = \mathbb{E}\Big[\lambda_{\max}(\boldsymbol{M})\Big]$$

*and the optimal estimator for this metric is a unit eigenvector of $\boldsymbol{M}$ associated to its largest eigenvalue $\lambda_{\max}(\boldsymbol{M})$.*

**Proof.** Let $\widehat{\boldsymbol{x}}$ be an estimator of $\boldsymbol{X}$. By the Nishimori identity (Proposition 5.1), we have

$$\mathbb{E}[(\widehat{\boldsymbol{x}}(\boldsymbol{Y});\boldsymbol{X})^2] = \mathbb{E}[\widehat{\boldsymbol{x}}(\boldsymbol{Y})^\intercal \boldsymbol{X}\boldsymbol{X}^\intercal \widehat{\boldsymbol{x}}(\boldsymbol{Y})] = \mathbb{E}[\langle \widehat{\boldsymbol{x}}(\boldsymbol{Y})^\intercal \boldsymbol{x}\boldsymbol{x}^\intercal \widehat{\boldsymbol{x}}(\boldsymbol{Y})\rangle] = \mathbb{E}[\widehat{\boldsymbol{x}}(\boldsymbol{Y})^\intercal \boldsymbol{M} \widehat{\boldsymbol{x}}(\boldsymbol{Y})],$$

the lemma follows. □

Lemma 6.1 tells us that the top unit eigenvector $\widehat{\boldsymbol{v}}$ of $\boldsymbol{M}$ maximizes $\mathbb{E}\left[(\widehat{\boldsymbol{x}}(\boldsymbol{Y});\boldsymbol{X})^2\right]$. In the following, we will show that under a simple condition (that will hold for the models we consider in this manuscript), the estimator $\widehat{\boldsymbol{v}}$ is "asymptotically optimal" (in the limit of large dimension) for the two metrics (6.2) and $\lambda_{\max}\widehat{\boldsymbol{v}}\widehat{\boldsymbol{v}}^\intercal$ is optimal for the estimation of $\boldsymbol{X}\boldsymbol{X}^\intercal$ in terms of mean square error.

To introduce this condition and the asymptotic limit, we need to consider to a sequence of inference problems. We assume that for all $n \geq 1$ we have two random vectors $\boldsymbol{X}_{[n]}$ and $\boldsymbol{Y}_{[n]}$ respectively in $\mathbb{S}^{n-1}$ and $\mathbb{R}^{m_n}$, for some sequence $(m_n)_{n\geq 1}$. Our goal is again to estimate $\boldsymbol{X}_{[n]}$ from the observation of $\boldsymbol{Y}_{[n]}$ when $n$ is very large: we would like for instance to compute the limits of (6.2) and (6.4) as $n \to \infty$. Moreover, we would like to know which estimators are "asymptotically optimal", i.e. whose performance reach in the $n \to \infty$ limit the optimal one. In the following, in order to simplify the notations, we will write $\boldsymbol{X}$ and $\boldsymbol{Y}$ instead of $\boldsymbol{X}_{[n]}$ and $\boldsymbol{Y}_{[n]}$.

**Proposition 6.1**

*Let us denote by $G_n$ the posterior distribution of $\boldsymbol{X}$ given $\boldsymbol{Y}$. Notice that $G_n$ is a random probability distributions on $\mathbb{S}^{n-1}$. Assume that there exists $q \in [0,1]$ such that for $\boldsymbol{x}^{(1)}, \boldsymbol{x}^{(2)} \overset{\text{i.i.d.}}{\sim} G_n$ we have*

$$\left|\left(\boldsymbol{x}^{(1)};\boldsymbol{x}^{(2)}\right)\right| \xrightarrow[n\to\infty]{\text{(d)}} q. \tag{6.5}$$

*Then $\mathrm{Tr}(\boldsymbol{M}^2) \xrightarrow[n\to\infty]{\text{(d)}} q^2$ and*

$$\lambda_{\max}(\boldsymbol{M}) \xrightarrow[n\to\infty]{\text{(d)}} q.$$

**Proof.** Let us compute $\mathrm{Tr}(\boldsymbol{M}^2) = \mathrm{Tr}(\langle \boldsymbol{x}\boldsymbol{x}^\intercal\rangle\langle \boldsymbol{x}\boldsymbol{x}^\intercal\rangle) = \langle (\boldsymbol{x}^{(1)};\boldsymbol{x}^{(2)})^2\rangle \xrightarrow[n\to\infty]{} q^2$, by assumption. If $q = 0$, then the result is obvious since $\lambda_{\max}(\boldsymbol{M})^2 \leq \mathrm{Tr}(\boldsymbol{M}^2)$.

Notice that $\mathrm{Tr}(\boldsymbol{M}^3) \leq \lambda_{\max}(\boldsymbol{M})\mathrm{Tr}(\boldsymbol{M}^2)$, so it suffices to show that

$$\mathrm{Tr}(\boldsymbol{M}^3) = \left\langle (\boldsymbol{x}^{(1)};\boldsymbol{x}^{(2)})(\boldsymbol{x}^{(2)};\boldsymbol{x}^{(3)})(\boldsymbol{x}^{(3)};\boldsymbol{x}^{(1)})\right\rangle \xrightarrow[n\to\infty]{} q^3,$$

for $\boldsymbol{x}^{(1)}, \boldsymbol{x}^{(2)}, \boldsymbol{x}^{(3)} \overset{\text{i.i.d.}}{\sim} G_n$. This follows from Lemma 6.2 below. □

**Lemma 6.2**

*Under the assumptions of Proposition 6.1, we have for $\boldsymbol{x}^{(1)}, \boldsymbol{x}^{(2)}, \boldsymbol{x}^{(3)} \overset{\text{i.i.d.}}{\sim} G_n$,*

$$(\boldsymbol{x}^{(1)};\boldsymbol{x}^{(2)})(\boldsymbol{x}^{(2)};\boldsymbol{x}^{(3)})(\boldsymbol{x}^{(3)};\boldsymbol{x}^{(1)}) \xrightarrow[n\to\infty]{\text{(d)}} q^3.$$

Lemma 6.2 will be proved in Appendix **??**. From Proposition 6.1 we deduce the main result of this section:

**Proposition 6.2**

*Let $\widehat{\boldsymbol{v}}$ be a leading unit eigenvector of $\boldsymbol{M}$ (which is defined by (6.3)). Under the assumptions of Proposition 6.1, we have*

$$|(\widehat{\boldsymbol{v}};\boldsymbol{X})| \xrightarrow[n\to\infty]{\text{(d)}} \sqrt{q}. \tag{6.6}$$

*Further* $\lim_{n\to\infty} \mathrm{MMSE}(\boldsymbol{X}\boldsymbol{X}^\mathsf{T}|\boldsymbol{Y}) = 1-q^2$,

$$\lim_{n\to\infty} \sup_{\|\widehat{\boldsymbol{x}}\|=1} \mathbb{E}\Big[|(\widehat{\boldsymbol{x}}(\boldsymbol{Y});\boldsymbol{X})|\Big] = \sqrt{q}, \qquad \textit{and} \qquad \lim_{n\to\infty} \sup_{\|\widehat{\boldsymbol{x}}\|=1} \mathbb{E}\Big[(\widehat{\boldsymbol{x}}(\boldsymbol{Y});\boldsymbol{X})^2\Big] = q. \tag{6.7}$$

**Proof.** Let us abbreviate $\lambda_{\max} \overset{\text{def}}{=} \lambda_{\max}(\boldsymbol{M})$. By Lemma 6.1 and Proposition 6.1 we have

$$\mathbb{E}\big[(\widehat{\boldsymbol{v}};\boldsymbol{X})^2\big] = \mathbb{E}[\lambda_{\max}] \xrightarrow[n\to\infty]{} q. \tag{6.8}$$

Hence if $q=0$, the Proposition follows easily. Assume now that $q>0$. Using Pythagorean Theorem and the Nishimori identity (Proposition 5.1) we get

$$\begin{aligned}
\mathbb{E}\|\boldsymbol{X}^{\otimes 4} - \lambda_{\max}^2 \widehat{\boldsymbol{v}}^{\otimes 4}\|^2 &\geq \mathbb{E}\|\boldsymbol{X}^{\otimes 4} - \langle \boldsymbol{x}^{\otimes 4}\rangle\|^2 \\
&= 1 + \mathbb{E}\big[(\langle \boldsymbol{x}^{\otimes 4}\rangle;\langle \boldsymbol{x}^{\otimes 4}\rangle)\big] - 2\mathbb{E}\big[(\boldsymbol{X}^{\otimes 4};\langle \boldsymbol{x}^{\otimes 4}\rangle)\big] \\
&= 1 - \mathbb{E}\Big[\Big\langle (\boldsymbol{x}^{(1)};\boldsymbol{x}^{(2)})^4 \Big\rangle\Big] \xrightarrow[n\to\infty]{} 1-q^4,
\end{aligned}$$

where the last limit follows from the assumption (6.5). Since

$$\mathbb{E}\|\boldsymbol{X}^{\otimes 4} - \lambda_{\max}^2 \widehat{\boldsymbol{v}}^{\otimes 4}\|^2 = 1 + \mathbb{E}[\lambda_{\max}^4] - 2\mathbb{E}[\lambda_{\max}^2 (\boldsymbol{X};\widehat{\boldsymbol{v}})^4],$$

using Proposition 6.1, we deduce (recall that we assumed $q>0$) that $\limsup_{n\to\infty} \mathbb{E}\big[(\widehat{\boldsymbol{v}};\boldsymbol{X})^4\big] \leq q^2$. Together with (6.8) this gives that $|(\widehat{\boldsymbol{v}};\boldsymbol{X})| \xrightarrow[n\to\infty]{} \sqrt{q}$.

The next point is a consequence of Proposition 6.1 because $\mathrm{MMSE}(\boldsymbol{X}\boldsymbol{X}^\mathsf{T}|\boldsymbol{Y}) = 1-\mathbb{E}[\mathrm{Tr}(\boldsymbol{M}^2)]$. To prove (6.7) simply notice that

$$\mathbb{E}\big[|(\widehat{\boldsymbol{v}};\boldsymbol{X})|\big]^2 \leq \sup_{\|\widehat{\boldsymbol{x}}\|=1} \mathbb{E}\big[|(\widehat{\boldsymbol{x}}(\boldsymbol{Y});\boldsymbol{X})|\big]^2 \leq \sup_{\|\widehat{\boldsymbol{x}}\|=1} \mathbb{E}\big[(\widehat{\boldsymbol{x}}(\boldsymbol{Y});\boldsymbol{X})^2\big] = \mathbb{E}[\lambda_{\max}],$$

which proves (6.7) using (6.6) and Proposition 6.1. □

From Proposition 6.2, we deduce that the estimator $\widehat{\boldsymbol{A}} \overset{\text{def}}{=} \lambda_{\max}(\boldsymbol{M})\,\widehat{\boldsymbol{v}}\widehat{\boldsymbol{v}}^\mathsf{T}$ achieves asymptotically the minimal mean square error for the estimation of $\boldsymbol{X}\boldsymbol{X}^\mathsf{T}$:

$$\lim_{n\to\infty} \mathbb{E}\|\boldsymbol{X}\boldsymbol{X}^\mathsf{T} - \widehat{\boldsymbol{A}}\|^2 = 1-q^2.$$

**Remark 6.1.** *For simplicity we assumed in this section that $\|\boldsymbol{X}\|^2 = 1$ almost surely. However we will need to work in the next chapters under a slightly weaker condition, namely $\|\boldsymbol{X}\|^2 \xrightarrow[n\to\infty]{} 1$. It is not difficult to modify the proofs of this section to see that Lemma 6.1, Proposition 6.1, Lemma 6.2 and Proposition 6.2 still hold, provided that $\|\boldsymbol{X}\|^2 \xrightarrow[n\to\infty]{} 1$ for the Wasserstein distance of order 4 (i.e. $\|\boldsymbol{X}\|^2 \xrightarrow[n\to\infty]{} 1$ in distribution and $\mathbb{E}\|\boldsymbol{X}\|^8 \xrightarrow[n\to\infty]{} 1$).*

# 7 A warm-up: the "needle in a haystack" problem

In order to illustrate the results seen in the previous sections, we study now a very simple inference model. Let $(e_1, \dots, e_{2^n})$ be the canonical basis of $\mathbb{R}^{2^n}$. Let $\sigma_0 \sim \mathrm{Unif}(\{1, \dots, 2^n\})$ and define $\boldsymbol{X} = e_{\sigma_0}$ (i.e. $\boldsymbol{X}$ is chosen uniformly over the canonical basis of $\mathbb{R}^{2^n}$). Suppose here that we observe:

$$\boldsymbol{Y} = \sqrt{\lambda n}\boldsymbol{X} + \boldsymbol{Z}\,,$$

where $\boldsymbol{Z} = (Z_1, \dots, Z_{2^n}) \overset{\text{i.i.d.}}{\sim} \mathcal{N}(0,1)$, independently from $\sigma_0$. The goal here is to estimate $\boldsymbol{X}$ or equivalently to find $\sigma_0$. The posterior distribution reads:

$$\begin{aligned}\mathbb{P}(\sigma_0 = \sigma|\boldsymbol{Y}) = \mathbb{P}(\boldsymbol{X} = e_\sigma|\boldsymbol{Y}) &= \frac{1}{\mathcal{Z}_n(\lambda)} 2^{-n} \exp\left(\sqrt{\lambda n} e_\sigma^\intercal \boldsymbol{Y} - \frac{\lambda n}{2}\|e_\sigma\|^2\right) \\ &= \frac{1}{\mathcal{Z}_n(\lambda)} 2^{-n} \exp\left(\sqrt{\lambda n} Z_\sigma + \lambda n \mathbb{1}(\sigma = \sigma_0) - \frac{\lambda n}{2}\right),\end{aligned}$$

where $\mathcal{Z}_n(\lambda)$ is the partition function

$$\mathcal{Z}_n(\lambda) = \frac{1}{2^n}\sum_{\sigma=1}^{2^n} \exp\left(\sqrt{\lambda n} Z_\sigma + \lambda n \mathbb{1}(\sigma = \sigma_0) - \frac{\lambda n}{2}\right).$$

We will be interested in computing the free energy $F_n(\lambda) = \frac{1}{n}\mathbb{E}\log \mathcal{Z}_n(\lambda)$ in order to deduce then the minimal mean squared error using the I-MMSE relation (3.5) presented in the previous section.

Although its simplicity, this model is interesting for many reasons. First, it is one of the simplest statistical model for which one observes a phase transition. Second it is the "planted" analog of the random energy model (REM) introduced in statistical physics by Derrida [**?**, **?**], for which the free energy reads $\frac{1}{n}\mathbb{E}\log\sum_\sigma \frac{1}{2^n}\exp\left(\sqrt{\lambda n} Z_\sigma\right)$. Third, as we will see in Section **??**, this model correspond to the "large order limit" of a rank-one tensor estimation model.

We start by computing the limiting free energy:

**Theorem 7.1**

$$\lim_{n\to\infty} F_n(\lambda) = \begin{cases} 0 & \text{if} \quad \lambda \le 2\log 2\,, \\ \frac{\lambda}{2} - \log(2) & \text{if} \quad \lambda \ge 2\log 2\,. \end{cases}$$

**Proof.** Using Jensen's inequality

$$\begin{aligned} F_n(\lambda) &\le \frac{1}{n}\mathbb{E}\log\mathbb{E}\left[\mathcal{Z}_n(\lambda)|\sigma_0, Z_{\sigma_0}\right] = \frac{1}{n}\mathbb{E}\log\left(1 - \frac{1}{2^n} + e^{\sqrt{\lambda n}Z_{\sigma_0} + \frac{\lambda n}{2} - \log(2)n}\right) \\ &\le \frac{1}{n}\mathbb{E}\log\left(1 + e^{\frac{\lambda n}{2} - \log(2)n}\right) + \sqrt{\frac{\lambda}{n}} \xrightarrow[n\to\infty]{} \begin{cases} 0 & \text{if} \quad \lambda \le 2\log(2)\,, \\ \frac{\lambda}{2} - \log(2) & \text{if} \quad \lambda \ge 2\log(2)\,. \end{cases}\end{aligned}$$

$F_n$ is non-negative since $F_n(0) = 0$ and $F_n$ is non-decreasing. We have therefore $F_n(\lambda) \xrightarrow[n\to\infty]{} 0$ for all $\lambda \in [0, 2\log(2)]$. We have also, by only considering the term $\sigma = \sigma_0$:

$$F_n(\lambda) \ge \frac{1}{n}\mathbb{E}\log\left(\frac{e^{\sqrt{\lambda n}Z_{\sigma_0} + \frac{\lambda n}{2}}}{2^n}\right) = \frac{\lambda}{2} - \log(2)\,.$$

We obtain therefore that $F_n(\lambda) \xrightarrow[n\to\infty]{} \frac{\lambda}{2} - \log(2)$ for $\lambda \geq 2\log(2)$. □

Using the I-MMSE relation (3.5), we deduce the limit of the minimum mean squared error $\mathrm{MMSE}_n(\lambda) = \min_{\widehat{\theta}} \mathbb{E}\|\boldsymbol{X} - \widehat{\theta}(\boldsymbol{Y})\|^2$:

$$\mathrm{MMSE}_n(\lambda) = \mathbb{E}\|\boldsymbol{X}\|^2 - 2F_n'(\lambda) = 1 - 2F_n'(\lambda)\,.$$

$F_n$ is a convex function of $\lambda$, thus (see Proposition **??**) its derivative converges to the derivative of its limit at each $\lambda$ at which the limit is differentiable, i.e. for all $\lambda \in (0, +\infty) \setminus \{2\log(2)\}$. We obtain therefore that for all $\lambda > 0$,

- if $\lambda < 2\log(2)$, then $\mathrm{MMSE}_n(\lambda) \xrightarrow[n\to\infty]{} 1$: one can not recover $\boldsymbol{X}$ better than a random guess.
- if $\lambda > 2\log(2)$, then $\mathrm{MMSE}_n(\lambda) \xrightarrow[n\to\infty]{} 0$: one can recover $\boldsymbol{X}$ perfectly.

Of course, the result we obtain here is (almost) trivial since the maximum likelihood estimator

$$\widehat{\sigma}(\boldsymbol{Y}) = \operatorname*{arg\,max}_{1\leq\sigma\leq 2^n} Y_\sigma$$

of $\sigma_0$ is easy to analyze. Indeed, $\max_\sigma Z_\sigma \simeq \sqrt{2\log(2)n}$ with high probability so that the maximum likelihood estimator recovers perfectly the signal for $\lambda > 2\log(2)$ with high probability.

# 8 Appendix

## 8.1 Proof of Proposition 3.2

We start by proving that MMSE is continuous at $\lambda = 0$. Let $\lambda \geq 0$ and consider $\boldsymbol{Y}, \boldsymbol{X}, \boldsymbol{Z}$ as given by (3.1). By dominated convergence one has almost surely that

$$\mathbb{E}[\boldsymbol{X}|\boldsymbol{Y}] = \frac{\int dP_X(\boldsymbol{x})\boldsymbol{x}e^{-\frac{1}{2}\|\sqrt{\lambda}\boldsymbol{x}-\boldsymbol{Y}\|^2}}{\int dP_X(\boldsymbol{x})e^{-\frac{1}{2}\|\sqrt{\lambda}\boldsymbol{x}-\boldsymbol{Y}\|^2}} \xrightarrow[\lambda\to 0]{} \mathbb{E}[\boldsymbol{X}]\,.$$

Then by Fatou's Lemma we get

$$\liminf_{\lambda\to 0} \mathrm{MMSE}(\lambda) \geq \mathbb{E}\Big[\liminf_{\lambda\to 0} \Big\|\boldsymbol{X} - \mathbb{E}[\boldsymbol{X}|\boldsymbol{Y}]\Big\|^2\Big] = \mathbb{E}\Big\|\boldsymbol{X} - \mathbb{E}[\boldsymbol{X}]\Big\|^2\,.$$

Combining this with the bound $\mathrm{MMSE}(\lambda) \leq \mathbb{E}\|\boldsymbol{X} - \mathbb{E}[\boldsymbol{X}]\|^2$ gives $\mathrm{MMSE}(\lambda) \xrightarrow[\lambda\to 0]{} \mathbb{E}\|\boldsymbol{X} - \mathbb{E}[\boldsymbol{X}]\|^2$. This proves that the MMSE is continuous at $\lambda = 0$.

Let us now prove that the MMSE is continuous on $\mathbb{R}^*_{\geq 0}$. We need here a technical lemma:

**Lemma 8.1**

*For all $\lambda > 0$, $p \geq 1$*

$$\mathbb{E}\|\boldsymbol{X} - \langle \boldsymbol{x}\rangle_\lambda\|^{2p} \leq \frac{2^p(2p!)}{\lambda^p p!} n^{p+1}\,.$$

**Proof.** We reproduce here the proof from [**?**], Proposition 5. We start with the equality

$$\sqrt{\lambda}\,(\boldsymbol{X} - \langle \boldsymbol{x} \rangle_\lambda) = \sqrt{\lambda}\boldsymbol{X} - \mathbb{E}[\sqrt{\lambda}\boldsymbol{X}|\boldsymbol{Y}] = \boldsymbol{Y} - \boldsymbol{Z} - \mathbb{E}[\boldsymbol{Y} - \boldsymbol{Z}|\boldsymbol{Y}] = \mathbb{E}[\boldsymbol{Z}|\boldsymbol{Y}] - \boldsymbol{Z}\,.$$

We have therefore

$$\mathbb{E}\|\boldsymbol{X} - \langle \boldsymbol{x} \rangle_\lambda\|^{2p} = \frac{1}{\lambda^p}\mathbb{E}\|\mathbb{E}[\boldsymbol{Z}|\boldsymbol{Y}] - \boldsymbol{Z}\|^{2p} \leq \frac{2^{2p-1}}{\lambda^p}\mathbb{E}\big[\|\mathbb{E}[\boldsymbol{Z}|\boldsymbol{Y}]\|^{2p} + \|\boldsymbol{Z}\|^{2p}\big] \leq \frac{2^{2p}}{\lambda^p}\mathbb{E}\|\boldsymbol{Z}\|^{2p}\,.$$

It remains to bound

$$\mathbb{E}\|\boldsymbol{Z}\|^{2p} \leq n^p \mathbb{E}\left[\sum_{i=1}^n Z_i^{2p}\right] = n^{p+1}\frac{(2p)!}{2^p p!}\,.$$

□

Let $\lambda_0 > 0$. The family of random variables $\left(\|\boldsymbol{X} - \langle \boldsymbol{x} \rangle_\lambda\|^2\right)_{\lambda \geq \lambda_0}$ is bounded in $L^2$ by Lemma 8.1 and is therefore uniformly integrable. The function $\lambda \mapsto \|\boldsymbol{X} - \langle \boldsymbol{x} \rangle_\lambda\|^2$ is continuous on $[\lambda_0, +\infty)$, the uniform integrability ensures then that MMSE : $\lambda \mapsto \mathbb{E}\|\boldsymbol{X} - \langle \boldsymbol{x} \rangle_\lambda\|^2$ is continuous over $[\lambda_0, +\infty)$. This is valid for all $\lambda_0 > 0$: we conclude that MMSE is continuous over $(0, +\infty)$.