A statistical physics perspective on statistical inference

InformedAI Summer School 2026

Bruno Loureiro

Département d'Informatique, École Normale Supérieure (ENS) – PSL, CNRS, France

June 15-18, 2026

1 Basic notions

Given samples from a target distribution $Z \sim p_\star$, *statistical inference* is the subject concerned with the question of how to estimate properties of $p_\star$ from the observed data Z. This can be estimating the full distribution or some of its statistics, e.g. moments.

Our focus in this lectures will be on *parametric statistical inference*, when the target distribution is a parametric function $p_\star = p_{\theta_\star}$, which reduces the problem to estimating the *ground truth* parameters $\theta_\star$. Most of distributions studied in basic probability courses are parametric, e.g. the normal distribution $\mathcal{N}(\mu, \sigma^2)$ for which $\theta = (\mu, \sigma^2) \in \mathbb{R} \times (0, \infty)$, the Bernouilli distribution $\mathrm{Bernouilli}(p)$ for which $\theta = p$, the Poisson distribution $\mathrm{Pois}(\lambda)$ for which $\theta = \lambda \in (0, \infty)$, etc.

An important class of parametric statistical inference tasks are *inverse problems* in signal processing. These are problems in which a signal is sent through a potentially noisy channel, and the goal of the statistician is to reconstruct or detect the signal given the observations. In this case, the ground truth $\theta_\star$ plays the role of the signal, while the distribution $p_{\theta_\star}$ parametrises the noise or transmission channel.

Example 1.1. A few examples of inverse problems are:

- **Denoising:** In scalar denoising, $Y = \theta_\star + \sigma Z$ with $\mathbb{E}[Z] = 0$ and $\mathbb{E}[\varepsilon^2] = 1$. A variation of this problem is vector denoising, where $\boldsymbol{Y} \in \mathbb{R}^n$ and:

$$\boldsymbol{Y} = \boldsymbol{\theta}_\star + \sigma \boldsymbol{Z} \tag{1.1}$$

  with the noise ε_j often assumed i.i.d. from a distribution with zero mean and unit variance. An important particular case is when the noise is i.i.d. Gaussian $Z_i \sim \mathcal{N}(0, \sigma^2)$, and hence $p_{\boldsymbol{\theta}_\star}(y) = \mathcal{N}(\boldsymbol{\theta}_\star, \sigma^2 \boldsymbol{I}_n)$. Denoising will be one of the guiding examples in these notes.

- **Linear estimation:** In linear estimation, the observation is a set of pairs $Y = \{(\boldsymbol{x}_i, y_i) \in \mathbb{R}^d \times \mathbb{R} : i \in [n]\}$, where $y_i = \langle \boldsymbol{\theta}_\star, \boldsymbol{x}_i \rangle + \boldsymbol{\varepsilon}_i$ with $\boldsymbol{x}_i \sim p_x$ and i.i.d. noise. This is also often denoted in matrix form as:

$$\boldsymbol{y} = \boldsymbol{X}\boldsymbol{\theta}_\star + \boldsymbol{\varepsilon} \tag{1.2}$$

  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 $\boldsymbol{y}, \boldsymbol{\varepsilon} \in \mathbb{R}^d$.

- **Generalised linear estimation:** In linear estimation, the observations also come in pairs $Y_i = (\boldsymbol{x}_i, y_i) \in \mathbb{R}^d \times \mathbb{R}$, but now the noise is parametrised by a more general likelihood $P_{y|x}$:

$$y_i \sim P_{y|x}(\cdot|\langle \boldsymbol{\theta}_\star, \boldsymbol{x}_i \rangle), \quad \text{i.i.d.} \tag{1.3}$$

  Note that linear estimation is a particular case of the above. But generalised linear models also contain a richer class of problems where the relationship between signal and observation can be non-linear, e.g. $y_i = \sigma(\langle \boldsymbol{\theta}_\star, \boldsymbol{x}_i \rangle) + \varepsilon_i$, also known as a *single-index model*.

- **Spiked matrix factorisation:** In spiked matrix factorisation, the observation is a matrix $\boldsymbol{Y} \in \mathbb{R}^{n \times n}$:

$$\boldsymbol{Y} = \boldsymbol{v}_\star \boldsymbol{v}_\star^\top + \boldsymbol{Z} \tag{1.4}$$

  where the signal $\boldsymbol{\theta}_\star \in \mathbb{R}^d$ is known as the spike, and the noise is a random matrix $\boldsymbol{Z} \in \mathbb{R}^{n \times n}$. When $\boldsymbol{W} \sim \text{GOE}(n)$, this is also known as the *spiked Wigner model*. A close variation of this problem considers a rectangular observation matrix $\boldsymbol{Y} \in \mathbb{R}^{n \times d}$:

$$\boldsymbol{Y} = \boldsymbol{u}_\star \boldsymbol{v}_\star^\top + \boldsymbol{Z} \tag{1.5}$$

  When $W_{ij} \sim \mathcal{N}(0, 1)$, this is known as the *spiked Wishart model*.

In all these inverse problems, the signal $\theta_\star$ is typically structured, and the goal is to understand how structure impacts the statistical complexity of estimation. Typical modelling assumptions in the classical literature are sparsity,[1] or quasi-sparsity, e.g. belonging to a ℓ_p ball (Donoho and Johnstone, 1994). A stronger assumption, more common in the high-dimensional inference literature, is to make the signal probabilistic.

Assumption 1.2 (Random ground truth)**.** We assume the ground truth parameters are drawn at random from a prior distribution $\theta_\star \sim \pi$.

Note that without specifying π, assumption 1.2 is not really restrictive, as one can always recover a deterministic setting by taking $\pi = \delta_{\theta_\star}$. However, as it will become clearer in the following, the interest of assumption 1.2 is when π is a structured distribution, and therefore assumption 1.2 states that $\theta_\star$ is a "typical" realisation of this distribution. This should be contrasted to a minimax approach for this problem, where $\theta_\star$ can be taken adversarially from the constraining set of interest. Sometimes, this probabilistic setting is refereed to as the *typical-case approach* in opposition to the *worst-case*, minimax approach.[2]

Definition 1.3 (Posterior distribution)**.** Consider a parametric inference problem with $Y \sim p_{\theta_\star}(z)$, where $\theta_\star \sim \pi$. The *posterior distribution* of θ given the observation Z is given by:

$$p(\theta|Y) = \frac{p_\theta(Y)\pi(\theta)}{p(Y)} \tag{1.6}$$

In order to stress the relationship between the posterior $p(\theta|Z)$ and the likelihood $p_\theta(y)$, it is common to employ a Bayesian notation for the likelihood $p_\theta(Y) \equiv p(Y|\theta)$.

The posterior distribution enjoys of some interesting statistical properties.

[1] A vector $\boldsymbol{\theta} \in \mathbb{R}^d$ is said to be s-sparse if $s < d$ components are non-zero.

[2] See for instance Donoho, 2000 for an interesting discussion on the relationship between these approaches.

Proposition 1.4 (Nishimori identity)**.** *Consider a statistical inference problem with observations $Y \sim p(\cdot|\theta_\star)$ for $\theta_\star \sim \pi$. Then, for any measurable bounded function of the observations F, we have:*

$$\begin{aligned}&\mathbb{E}_{\theta_\star\sim\pi,\, Y\sim p(\cdot|\theta_\star)}\left[\mathbb{E}\left[F(Y,\theta_\star,\theta^{(1)},\cdots,\theta^{(k)})\mid Y\right]\right]\\&\quad=\mathbb{E}_{\theta_\star\sim\pi,\, Y\sim p(\cdot|\theta_\star)}\left[\mathbb{E}\left[F(Y,\theta^{(0)},\theta^{(1)},\cdots,\theta^{(k)})\mid Y\right]\right].\end{aligned}\tag{1.7}$$

where $\theta^{(0)},\theta^{(1)},\cdots,\theta^{(k)}$ are independent samples from the posterior distribution in eq. (1.6). In other words, under the expectation over the data generating process a sample from the posterior is equivalent to a sample from the prior. This can be equivalently stated as:

$$\left(Y,\theta_\star,\theta^{(1)},\cdots,\theta^{(k)}\right)\overset{(d)}{=}\left(Y,\theta^{(0)},\theta^{(1)},\cdots,\theta^{(k)}\right)\tag{1.8}$$

Remark 1.5. Note that proposition 1.4 does not mean that $\theta_\star$ and the replicas are independent unconditionally. They are generally correlated through the shared data Y. The statement is that, given Y, the planted signal $\theta_\star$ is distributed exactly as a posterior sample. This is precisely the Bayes-optimal, well-specified condition.

1.1 Estimators

As the name suggests, an estimator $\hat{\theta}(Y)$ is a data-dependent estimate of the ground truth parameters $\theta_\star$. Estimators are commonly denoted in statistics with a hat $\hat{\bullet}$. Formally, an estimator is defined as any measurable function of the data Y. The three most popular estimators in statistics are:

- **Maximum likelihood:** The maximum likelihood estimator (MLE) is obtained by maximising the (log) likelihood:

$$\hat{\theta}_{\text{mle}}(Y)=\underset{\hat{\theta}(Y)\in\text{supp}(\pi)}{\arg\max}\ \log p_\theta(Y)\tag{1.9}$$

- **Maximum a posteriori:** The maximum a posteriori estimator (MAP) is obtained by maximising the (log) posterior:

$$\begin{aligned}\hat{\theta}_{\text{map}}(Y)&=\underset{\hat{\theta}(Z)}{\arg\max}\log p(\theta|Y)\\&=\underset{\hat{\theta}(Y)}{\arg\max}\left\{\log p_\theta(Y)+\log\pi(\theta)\right\}\end{aligned}\tag{1.10}$$

- **Posterior mean:** The posterior mean estimator, sometimes also refereed as the Bayes estimator, is given by:

$$\hat{\theta}_{\text{Bayes}}(Z)=\mathbb{E}[\theta|Y]\tag{1.11}$$

  where the expectation is over the posterior distribution eq. (1.6). Note that differently from the MLE and MAP, this is not a point-estimator, but rather an average.

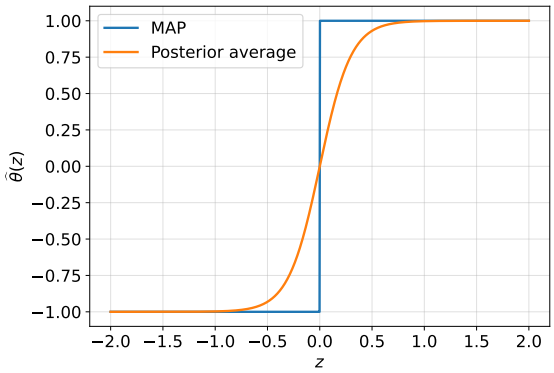

Figure 1: Comparison of MAP and Bayes estimator for scalar Gaussian denoising with a Radamacher prior $\pi = \frac{1}{2}(\delta_{+1} + \delta_{-1})$ with noise level $\sigma^2 = 0.3$

Example 1.6 (Scalar Gaussian denoising)**.** Consider Gaussian denoising problem where $Y = \theta_\star + \sigma Z$ with $Z \sim \mathcal{N}(0, 1)$. As discussed above, we have $p_{\theta_\star} = \mathcal{N}(\theta_\star, \sigma^2)$, and therefore:

$$\hat{\theta}_{\text{mle}}(Y) = \underset{\hat{\theta}(Y) \in \text{supp}(\pi)}{\arg\max} \left\{ -\frac{1}{2\sigma^2}(Y - \theta)^2 \right\} = Y \tag{1.12}$$

which is the observation itself. Note that, differently from the MAP and the Bayes estimator, this is independent of the prior distribution over $\theta_\star$. Now let's consider two examples of priors.

- **Gaussian signal**: Let $\theta_\star \sim \mathcal{N}(\mu, \tau^2)$. Then, the posterior is itself a Gaussian distribution:

$$p(\theta|Z) = \mathcal{N}(m(Y), v), \qquad m(y) = \frac{\tau^2}{\sigma^2 + \tau^2} y + \frac{\sigma^2}{\sigma^2 + \tau^2}\mu, \qquad v = \left(\frac{1}{\tau^2} + \frac{1}{\sigma^2} \right)^{-1} \tag{1.13}$$

  Therefore, since the posterior is unimodal, in this simple example the MAP and Bayes estimators are equivalent: $\hat{\theta}_{\text{map}}(Y) = \hat{\theta}_{\text{Bayes}}(Y) = m(Y)$. Differently from the MLE, this is a weighted average over the observation and the prior, with the weights determined by the relative variance of both distributions. This is sometimes called a *shrinkage estimator*.

- **Radamacher signal**: In general, the posterior can be multi-modal, implying that the MAP and Bayes estimators are different. As a concrete example, consider a mixture prior:

$$\pi = \rho\delta_{+1} + (1 - \rho)\delta_{-1} \tag{1.14}$$

  with $\rho \in [0, 1]$. Then, the posterior is itself a mixture:

$$p(\theta|Y = y) = \varphi(y)\delta_{+1} + (1 - \varphi(y))\,\delta_{-1} \tag{1.15}$$

  where:

$$\varphi(y) = \text{sigmoid}\left(\log \frac{\rho}{1 - \rho} + \frac{2y}{\sigma^2} \right) \tag{1.16}$$

  In this case, the MAP is different from the Bayes estimator:

$$\hat{\theta}_{\text{map}}(Z) = \begin{cases} +1 & \text{if } \varphi(Y) > \frac{1}{2} \\ -1 & \text{if } \varphi(Y) < \frac{1}{2} \end{cases}, \qquad \hat{\theta}_{\text{Bayes}}(Y) = 2\varphi(Y) - 1 \tag{1.17}$$

Therefore, the MAP is implementing a hard thresholding function, while the Bayes estimator implements a soft-thresholding. To see this more clearly, we can rewrite:

$$\hat{\theta}_{\mathrm{map}}(z) = \begin{cases} +1 & \text{if } y > \frac{\sigma^2}{2}\log\frac{1-\rho}{\rho} \\ -1 & \text{if } y < \frac{\sigma^2}{2}\log\frac{1-\rho}{\rho} \end{cases}, \qquad \hat{\theta}_{\mathrm{Bayes}}(Y) = \tanh\left(\frac{y}{\sigma^2} + \frac{1}{2}\log\frac{\rho}{1-\rho}\right) \tag{1.18}$$

See fig. 1 for an illustration.

1.2 Performance and optimality

Given an estimator $\hat{\theta}(Y)$, we are typically interested in assessing its performance. For that, we introduce a loss function $\ell(\theta_\star, \hat{\theta})$ which allows to quantify how good is the estimation with respect to the ground truth. The Bayes risk associated with the loss ℓ is defined as the minimal expected loss achieved by any estimator:

$$R_\star = \inf_{\hat{\theta}} \mathbb{E}_{\theta_\star \sim \pi, Y \sim p(\cdot|\theta_\star)} \left[\ell(\hat{\theta}(Y), \theta_\star)\right] \tag{1.19}$$

where the infimum is taken over all measurable functions of the data. An important example, which will be our main performance measure in the following, is given by the mean-squared error $\ell(\hat{\theta}, \theta_\star) = ||\hat{\theta} - \theta_\star||_2^2$.[3] In this context, the Bayes risk is also known as the *minimum mean-squared error* (MMSE), and is achieved by the Bayes estimator.

Lemma 1.7. *The minimum mean-squared error is achieved by the expectation over the Bayes posterior:*

$$\mathrm{MMSE} = \inf_{\hat{\theta}} \mathbb{E}_{\theta_\star \sim \pi, Y \sim p(\cdot|\theta_\star)} \left[||\hat{\theta} - \theta_\star||_2^2\right] = \mathbb{E}_{\theta_\star \sim \pi, Y \sim p(\cdot|\theta_\star)} \left[||\mathbb{E}[\theta|Y] - \theta_\star||_2^2\right] \tag{1.20}$$

This justifies our nomenclature of $\hat{\theta}_{\mathrm{Bayes}}(Y) = \mathbb{E}[\theta|Y]$ as the Bayes estimator. This shows that in order to understand the fundamental limitations in estimation, we must understand the posterior distribution.

2 Denoising

We now study in detail one of the simplest inference problems: denoising.

2.1 Scalar denoising

To warm up, let's start with the scalar case. Assume we observe:

$$Y = \theta + \sigma Z \tag{2.1}$$

where $Z \sim \mathcal{N}(0, 1)$ and $\theta \sim \pi$. The posterior distribution is given by:

$$p(\mathrm{d}\theta|Y = y) = \frac{1}{Z_\sigma(y)} e^{-\frac{(y-\theta)^2}{2\sigma^2}} \pi(\mathrm{d}\theta) \tag{2.2}$$

where the partition function is:

$$Z_\sigma(y) = \int \pi(\mathrm{d}\theta) e^{-\frac{(y-\theta)^2}{2\sigma^2}} \tag{2.3}$$

which is simply a convolution of the prior with the Gaussian likelihood.

[3] We implicitly assumed $\hat{\theta}$ is a vector here. But a similar notion can be defined for matrices by taking the Frobenius norm.

Bayes estimator — The Bayes estimator is given by:

$$\hat{\theta}_{\text{Bayes}}(Y) = \mathbb{E}[\theta|Y] = \frac{\int \pi(\mathrm{d}\theta) e^{-\frac{(y-\theta)^2}{2\sigma^2}} \theta}{\int \pi(\mathrm{d}\theta) e^{-\frac{(y-\theta)^2}{2\sigma^2}}} \tag{2.4}$$

which as we discussed before is optimal in mean-squared error, i.e. it achieves the MMSE:

$$\text{MMSE}(\sigma^2) = \mathbb{E}[(\theta - \hat{\theta}_{\text{Bayes}}(Y))^2] \tag{2.5}$$

Free energy — We now discuss the relationship between statistical physics and information theory. One of the key quantities in statistical physics is the free energy:

$$F(\sigma^2) = -\mathbb{E}[\log Z_\sigma(Y)] \tag{2.6}$$

where the expectation is taken over Y. This is closely related to an important quantity in information theory, the mutual information.

Lemma 2.1. *Up to an additive constant, the mutual information is equal to the free energy:*

$$I(\theta; \theta + \sigma Z) = -\frac{1}{2} + F(\sigma^2) \tag{2.7}$$

Proof. This follows from a simple calculation. By definition,

$$I(X;Y) = D_{\text{KL}}(P_{X,Y} \,||\, P_X \otimes P_Y). \tag{2.8}$$

Applying this to $X = \theta$ and Y in our problem:

$$\begin{aligned} I(\theta;Y) &= \mathbb{E}\left[\log \frac{p(\theta, Y)}{p(\theta)p(Y)}\right] \overset{(a)}{=} \mathbb{E}\left[\log \frac{p(Y|\theta)}{p(Y)}\right] \\ &= \mathbb{E}\left[\log p(Y|\theta)\right] - \mathbb{E}[\log p(Y)]. \end{aligned} \tag{2.9}$$

where in (a) we used $p(\theta, Y) = p(Y|\theta)p(\theta)$. Since:

$$p(Y) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(y-\theta)^2}{2\sigma^2}} = \frac{1}{\sqrt{2\pi\sigma^2}} Z_\sigma(Y) \tag{2.10}$$

We have:

$$\mathbb{E}[\log p(Y)] = -\tfrac{1}{2}\log 2\pi\sigma^2 - F(\sigma^2) \tag{2.11}$$

On the other hand, since $p(Y|\theta) = \mathcal{N}(\theta, \sigma^2)$

$$\begin{aligned} \mathbb{E}[\log p(Y|\theta)] &= -\frac{1}{2}\log 2\pi\sigma^2 - \mathbb{E}\left[\frac{(Y-\theta)^2}{2\sigma^2}\right] \\ &= -\frac{1}{2}\log 2\pi\sigma^2 - \frac{1}{2} \end{aligned} \tag{2.12}$$

Putting together yields the desired result:

$$I(\theta;Y) = -\frac{1}{2} + F(\sigma^2) \tag{2.13}$$

□

The mutual information quantifies how much information two random processes share. Therefore, $I(\theta; Y)$ is essentially a measure on how much information the observation Y carries about the signal θ. Surprisingly, the mutual information can be related to smallest mean-squared error that can be achieved by any estimator, i.e. the MMSE in eq. (2.5).

Theorem 2.2 (I-MMSE theorem, Guo et al., 2004)**.** *The MMSE is proportional to the derivative of the free energy with respect to noise variance.*

$$\frac{d}{d\sigma^2} I(\theta; \theta + \sigma Z) = -\frac{1}{2\sigma^4} \mathrm{MMSE}(\sigma^2) \tag{2.14}$$

Theorem 2.2 follows from a direct computation. The tricky part is that the mutual information depends on the noise level both through the expectation over $Y = \theta + \sigma Z$ and the log-partition function $\log Z_\sigma(Y)$. Therefore, it will be convenient to first prove first two auxiliary lemmas, which are useful results in information theory.

Lemma 2.3 (De Bruijn's identity)**.** *Let $X \in \mathbb{R}^n$ and $Z \sim \mathcal{N}(0, I_d)$ independent of X. Then:*

$$\frac{d}{dt} h(X + \sqrt{t} Z) = \frac{1}{2} J(X + \sqrt{t} Z) \tag{2.15}$$

where $h(Y) = -\mathbb{E}[\log p_Y(Y)]$ is the differential entropy and $J(Y) = \mathbb{E}[||\nabla \log p_Y(Y)||_2^2]$ is the Fisher information of the random vector $Y \sim p_Y$.

Proof. Let $Y_t = X + \sqrt{t} Z$ and denote by $p_t(y) = \mathbb{E}[\varphi_t(y - X)]$ the law of Y_t, where φ_t is the density of $\sqrt{t} Z \sim \mathcal{N}(0, t I_n)$. A standard property of ϕ_t is that it satisfies the heat equation (check this!):

$$\partial_t \varphi_t(z) = \nabla_z^2 \varphi_t(z) \tag{2.16}$$

By differentiation under the expectation, this implies that p_t itself satisfies a heat equation:

$$\partial_t p_t(y) = \partial_t \mathbb{E}[\varphi_t(y - X)] = \mathbb{E}[\partial_t \varphi_t(y - X)] = \mathbb{E}[\nabla_z^2 \varphi_t(y - X)] = \nabla_y^2 p_t(y)(y) \tag{2.17}$$

Therefore:

$$\partial_t h(Y_t) = -\partial_t \int \mathrm{d}y \; p_t(y) \log p_t(y) = -\int \mathrm{d}y \; \partial_t p_t(y) \log p_t(y) - \int \mathrm{d}y \; \partial_t p_t \tag{2.18}$$

However, since p_t is a density, we have that $\int \mathrm{d}y \; p_t(y) = 1$ and therefore:

$$\int \mathrm{d}y \; \partial_t p_t(y) = 0 \tag{2.19}$$

Therefore, using the heat-equation and integration by parts:

$$\begin{aligned}
\partial_t h(Y_t) &= -\int \mathrm{d}y \; \nabla_y^2 p_t(y) \log p_t(y) = \int \mathrm{d}y \; \langle \nabla_y p_t(y), \nabla_y \log p_y(y) \rangle \\
&= \int \mathrm{d}y \; p_y(y) \left\langle \frac{\nabla p_t(y)}{p_t(t)}, \nabla_y \log p_y(y) \right\rangle \\
&= \mathbb{E}\left[||\nabla_y \log p_y(Y)]||_2^2 \right]
\end{aligned}$$

which is the desired result. □

Lemma 2.4 (Tweedie's formula). *Consider the process $Y = X + \sqrt{t}Z$ with $X \sim \pi$ independently from $Z \sim \mathcal{N}(0, I_n)$. Then:*

$$\mathbb{E}[X|Y = X + \sqrt{t}Z] = Y + t\nabla \log p(Y) \tag{2.20}$$

Proof. The result follows from direct computation:

$$\nabla_y \log p(y) = \frac{1}{p_y(y)} \frac{\nabla_y e^{-\frac{||y-X||_2^2}{2t}}}{(2\pi t)^{n/2}} = -\frac{y - X}{t} \tag{2.21}$$

□

We can now move to the proof of Theorem 2.2.

Proof. Note that, up to a normalisation, the mutual information is precisely the differential entropy of Y:

$$I(\theta; Y) = -\frac{1}{2} - \mathbb{E}[\log Z_\sigma(Y)] = -\frac{1}{2} - \frac{1}{2}\log 2\pi\sigma^2 + h(Y) \tag{2.22}$$

Therefore, by Lemma 2.3:

$$\frac{\mathrm{d}}{\mathrm{d}\sigma^2} I(\theta; Y) = -\frac{1}{2\sigma^2} + \frac{1}{2}\mathbb{E}[(\partial_y \log p(Y))^2] \tag{2.23}$$

But by lemma 2.4:

$$\partial_y \log p(y) = \frac{\mathbb{E}[\theta|Y] - Y}{\sigma^2} \tag{2.24}$$

Therefore:

$$\begin{aligned}
\mathbb{E}[(\partial_y \log p(Y))^2] &= \frac{1}{\sigma^4}\mathbb{E}[(\mathbb{E}[\theta|Y] - Y)^2] = \frac{1}{\sigma^4}\mathbb{E}[(\mathbb{E}[\theta|Y] - \theta - \sigma Z)^2] \\
&= \frac{1}{\sigma^4}\mathbb{E}[(\mathbb{E}[\theta|Y] - \theta)^2] - \frac{2}{\sigma^3}(\mathbb{E}[\theta|Y] - \theta)Z] + \frac{1}{\sigma^2} \\
&\overset{(a)}{=} \frac{1}{\sigma^2} - \frac{1}{\sigma^4}\mathrm{MMSE}
\end{aligned} \tag{2.25}$$

Where in (a) we used that:

$$\begin{aligned}
\mathbb{E}[(\hat{\theta}(Y) - \theta)Z] &= \sigma^{-1}\mathbb{E}[(\hat{\theta}(Y) - \theta)(Y - \theta)] \\
&= -\sigma^{-1}\mathbb{E}[(\hat{\theta}(Y) - \theta)\theta] \\
&= -\sigma^{-1}\mathbb{E}[(\hat{\theta}(Y) - \theta)^2].
\end{aligned} \tag{2.26}$$

Putting together yields the desired result:

$$\begin{aligned}
\frac{\mathrm{d}}{\mathrm{d}\sigma^2} I(\theta; Y) &= -\frac{1}{2\sigma^2} + \frac{1}{2}\mathbb{E}[(\partial_y \log p(Y))^2] = -\frac{1}{2\sigma^2} - \frac{1}{2\sigma^4}\mathrm{MMSE} \\
&= -\frac{1}{2\sigma^4}\mathrm{MMSE}
\end{aligned} \tag{2.27}$$

□

2.2 Sparse vector denoising

We now move to the case of a vector denoising problem:

$$Y = \theta + \sigma_n Z \tag{2.28}$$

with $\theta \sim \pi$ and $Z \sim \mathcal{N}(0, I_n)$ independently. Note that if π factorises:

$$\pi(\theta) = \prod_{i=1}^{n} \pi_i(\theta_i) \tag{2.29}$$

then the posterior itself is a factorised distribution, implying that the problem simply decouples into n independent scalar estimation problems. Therefore, the interesting case to consider is when the entries of the signal are correlated.

As a concrete example, we will consider a widely studied case in signal processing: a s-sparse signal, for which only a subset $S \subset [n]$ of size $|S| = s$ the coordinates of θ are non-zero.

In particular, let's consider the extreme case where one coordinate $K \in [n]$, taken uniformly at random, is non-zero. This type of problem is known as *needle in the haystack*, since denoising requires finding a single non-zero coordinate among the $n - 1$ noise variables.

Remark 2.5. Note that π is not a factorised distribution, since coordinates are correlated. In particular, choosing the non-zero coordinate automatically forces all the remaining coordinates to be exactly zero.

More formally, we define a random coordinate $K \sim \mathrm{Unif}([n])$ and let $\theta | K = k \sim \delta_{e_k}$. Since $Y | K = k = \mathcal{N}(e_k, \sigma^2 I_n)$, the posterior distribution is given by:

$$\mathbb{P}(\theta = e_k | Y) \propto \underbrace{\frac{e^{-\frac{1}{2\sigma^2}||Y - e_k||_2^2}}{(2\pi\sigma^2)^{n/2}}}_{p(Y|\theta=e_k)} \underbrace{\frac{1}{n}}_{\pi(\theta=e_k)} = \frac{e^{\frac{1}{\sigma^2}Y_k}}{\sum_{j=1}^{n} e^{\frac{1}{\sigma^2}Y_j}} = \mathrm{softmax}\left(\frac{Y}{\sigma^2}\right)_k \tag{2.30}$$

Remark 2.6 (Support of the posterior)**.** Note that although the prior is constant, it does constraint the support of the posterior distribution. In other words, knowledge of the fact that the underlying signal is 1-sparse means that the posterior is effectively on a single coordinate $K \in [n]$.

The question we would like to understand is how small σ_n needs to be for us to be able to tell apart the signal $\theta = e_k$ from the noise? Since the posterior mean is optimal in MMSE, studying it should teach us about the fundamental limitations of recovery. But before, let's intuitively understand what is the trade-off between signal and noise in this problem.

Scaling: Estimation can be seen as a competition between the entropy from the prior and the information carried by the signal in the likelihood. Indeed, without any observation on the model, the best one can do is to take a random draw from the prior distribution, which is the uniform distribution over the coordinates. This can be measured by the entropy:

$$H(\pi) = \log n \tag{2.31}$$

How much information does the observation $Y = e_K + \sigma Z$ carry about the signal? To answer this question, we must separate the information carried by the signal from the information carried by the noise. The distribution of the noise part is $\sigma Z \sim \mathcal{N}(0, \sigma^2 I_n)$. The amount of

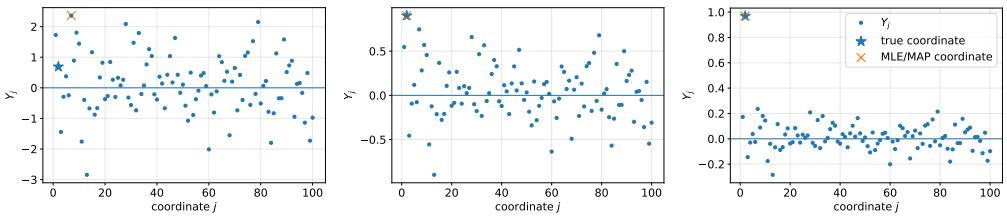

Figure 2: Y_j as a function of the coordinate $j \in [n]$ for $n = 100$ and three different noise-levels: (**Left**) Impossible recovery $\sigma_n \gg \frac{1}{2\log n}$, (**Center**) Critical $\sigma_n \approx \frac{1}{2\log n}$, (**Right**) Easy recovery $\sigma_n \ll \frac{1}{2\log n}$.

information carried by the signal can be measured by the relative entropy between the noise and the observation:

$$D_{\mathrm{KL}}(\mathcal{N}(e_K, \sigma^2 I_n)||\mathcal{N}(0, \sigma^2 I_n)) = \frac{||e_K||_2^2}{2\sigma^2} = \frac{1}{2\sigma^2} \tag{2.32}$$

Hence, we expect that $\sigma_n^2 \asymp \frac{1}{\log n}$ for the likelihood to remove a $O(1)$ fraction of entropy of the original uncertainty from the prior. From this, we should expect:

$$\begin{cases} \sigma_n \gg \frac{1}{\log n} & \text{impossible to identify } K \text{ (not enough information)} \\ \sigma_n = O\left(\frac{1}{\log n}\right) & \text{critical regime} \\ \sigma_n \ll \frac{1}{\log n} & \text{easy to identify } K \end{cases} \tag{2.33}$$

2.2.1 Maximum likelihood and maximum a posteriori estimators

Note that since the prior $\pi(\theta = e_k) = \frac{1}{n}$ is constant, the MAP and MLE estimators are equivalent for the 1-sparse denoising problem. They are explicitly given by:

$$\hat{\theta}_{\mathrm{mle}}(Y) = e_{\hat{K}}, \qquad \hat{K}(Y) = \arg\max_{k \in [n]} Y_k \tag{2.34}$$

This is very intuitive: we the MLE/MAP simply selects the largest value in the observations. Assume that the true coordinate is given by $k_\star \in [n]$. In order for the MLE/MAP to correlate with the signal we must have:

$$1 + \sigma_n Z_1 > \max_{j \neq k_\star} \sigma_n Z_j \tag{2.35}$$

By a classical result in extremal value statistics for Gaussian variables, for all $\varepsilon > 0$ with probability $1 - o(1)$ when $n \to \infty$ we have:

$$\max_{j \neq k_\star} Z_j \in \sqrt{2\log n}[1 - \varepsilon, 1 + \varepsilon] \tag{2.36}$$

Since the left-hand side is $O_{\mathbb{P}}(1)$, this implies that the:

$$\lim_{n \to \infty} \mathbb{P}(\hat{\theta}_{\mathrm{mle}}(Y) = e_{k_\star}) = \begin{cases} 0 & \text{for } \sigma_n^2 > \frac{1}{2\log n} \\ 1 & \text{for } \sigma_n^2 < \frac{1}{2\log n} \end{cases} \tag{2.37}$$

Which precisely agrees with the scale previously discussed !

2.2.2 Bayes optimal estimator

We have shown that the recovery threshold for MLE/MAP is $\sigma_c^2 = \frac{1}{2\log n}$. However, is this optimal? In other words, is there an estimator which recovers the signal above σ_c? To answer this question, we must characterise the MMSE, or equivalently (from theorem 2.2) the asymptotic free energy of the posterior distribution in the limit $n \to \infty$. Motivated by the scaling discussion above, let:

$$n = 2^d, \qquad \sigma_d^2 = \frac{\Delta}{d} \tag{2.38}$$

such that the critical scaling $\sigma_c^2 = O_d(d)$, and define the *free energy density*:

$$f_d(\Delta) = -\frac{1}{d}\mathbb{E}\Big[\log \underbrace{\sum_{j=1}^{2^d} e^{-\frac{||Y-\theta_j||_2^2}{2\sigma^2}}\pi(e_j)}_{Z_d(Y)}\Big] = \log 2 + \frac{2^d}{2d} + \frac{1}{\Delta} - \frac{1}{d}\mathbb{E}\Big[\log \underbrace{\sum_{j=1}^{2^d} e^{\frac{d}{\Delta}Y_j}}_{\tilde{Z}_d(Y)}\Big] \tag{2.39}$$

which, as discussed in the previous section is closely related to the *mutual information density*:

$$i_d(\Delta) = \log 2 + \frac{1}{\Delta} - \frac{1}{d}\mathbb{E}[\log \tilde{Z}_d(Y)] \tag{2.40}$$

Our goal is to understand its asymptotic value when $d \to \infty$. While it is possible to compute the asymptotic free energy density using standard tools from probability, this will be a nice opportunity to introduce a tool from statistical physics known as the replica method (Mézard et al., 1988).

Replica trick: The replica method addresses the challenge in computing the limiting free energy in eq. (2.39), which consists of dealing with the expectation of a logarithm. The key idea is to use the following identity:

$$\log x = \lim_{s \to 0^+} \partial_s x^s \tag{2.41}$$

also known as the *replica trick*. Applying this to eq. (2.39) reduces the problem to computing moments of the partition function:

$$\mathbb{E}[\log \tilde{Z}_d(Y)] = \lim_{s \to 0^+} \frac{1}{d}\mathbb{E}[\partial_s \tilde{Z}_d(Y)^s] \tag{2.42}$$

Now, we can explicitly write:

$$\tilde{Z}_d(Y)^r = \sum_{j_1,\cdots,j_r=1}^{2^d} e^{\frac{d}{\Delta}\sum_{a=1}^{r} Y_{j_a}} \tag{2.43}$$

Average over randomness: By symmetry, we can assume without loss of generality that the ground truth signal is in the first coordinate $\theta = e_1$. Noting that:

$$Y_j = \langle e_j, Y\rangle = \langle e_j, e_1\rangle + \sqrt{\frac{\Delta}{d}}\langle Z, e_j\rangle \tag{2.44}$$

we can take the average:

$$\mathbb{E}\left[e^{\frac{d}{\Delta}\sum_{a=1}^{r}\langle Y,e_{j_a}\rangle}\right] = e^{\frac{d}{\Delta}\sum_{a=1}^{r}\langle e_1,e_{j_a}\rangle}\mathbb{E}\left[e^{\sqrt{\frac{d}{\Delta}}\sum_{a=1}^{r}\langle e_{j_a},Z\rangle}\right] = e^{\frac{d}{\Delta}\sum_{a=1}^{r}\langle e_1,e_{j_a}\rangle+\frac{d}{2\Delta}\sum_{a,b=1}^{r}\langle e_{j_a},e_{j_b}\rangle} \tag{2.45}$$

This motivate us to introduce the following *summary statistics*:

$$m_a = \langle e_1, e_{j_a}\rangle, \qquad q_{ab} = \langle e_{j_a}, e_{j_b}\rangle \tag{2.46}$$

which measure the correlation between a replica and the signal and two independent replicas, respectively. Putting together, we can write:

$$\mathbb{E}\left[Z_d(Y)^r\right] = \sum_{j_1,\cdots,j_r=1}^{2^d} e^{\frac{d}{\Delta}\sum_{a=1}^{r} m_a+\frac{d}{2\Delta}\sum_{a,b=1}^{r} q_{ab}} \tag{2.47}$$

Applying Laplace's method: Note that the exponent depends only on the matrices $m \in \{0,1\}^r$ and $q_{ab} \in \{0,1\}^{r\times r}$. In order to derive a free energy potential, we can rewrite the sum over configurations $\{e_{j_a}\}$ in terms of a sum over the summary statistics:

$$\begin{aligned}\mathbb{E}\left[\tilde{Z}_d(Y)^r\right] = &\sum_{m\in\{0,1\}^r}\sum_{q\in\{0,1\}^{r\times r}} \exp\left\{\frac{d}{\Delta}\sum_{a=1}^{r} m_a + \frac{d}{2\Delta}\sum_{a,b=1}^{r} Q_{ab}\right\} \\ &\times \underbrace{\sum_{j_1,\cdots,j_r=1}^{2^d}\prod_{a=1}^{r}\mathbf{1}(m_a = \langle e_1, e_{j_a}\rangle)\prod_{1\le a<b\le r}\mathbf{1}(q_{ab} = \langle e_{j_a}, e_{j_b}\rangle)}_{\text{\# configurations with a given } m_a,\, q_{ab}}.\end{aligned} \tag{2.48}$$

Defining:

$$\Sigma_d(m,q) = \frac{1}{d}\log\sum_{j_1,\cdots,j_r=1}^{2^d}\prod_{a=1}^{r}\mathbf{1}(m_a = \langle e_1, e_{j_a}\rangle)\prod_{1\le a<b\le r}\mathbf{1}(q_{ab} = \langle e_{j_a}, e_{j_b}\rangle). \tag{2.49}$$

We can rewrite:

$$\mathbb{E}\left[\tilde{Z}_d(Y)^r\right] = \sum_{m\in\{0,1\}^r}\sum_{q\in\{0,1\}^{r\times r}} e^{d\Phi(m,q)} \tag{2.50}$$

where the *free energy potential* is given by:

$$\Phi_r(m,q) = \frac{1}{\Delta}\sum_{a=1}^{r} m_a + \frac{1}{2\Delta}\sum_{a,b=1}^{r} q_{ab} + \Sigma_d(m,q) \tag{2.51}$$

In the limit $d\to\infty$, Laplace's method give that the sum above concentrates at $m_\star, q_\star$ which extremise Φ_r:

$$\lim_{d\to\infty}\frac{1}{d}\log\mathbb{E}[\tilde{Z}_d(Y)^r] = \sup_{\substack{m\in\{0,1\}^r\\ q\in\{0,1\}^{r\times r}}}\Phi_r(m,q) \tag{2.52}$$

Therefore, we have reduced the problem from the space $\{0,1\}^{2^d}$ to the space of summary statistics $m\in\{0,1\}^r, q\in\{0,1\}^{r\times r}$.

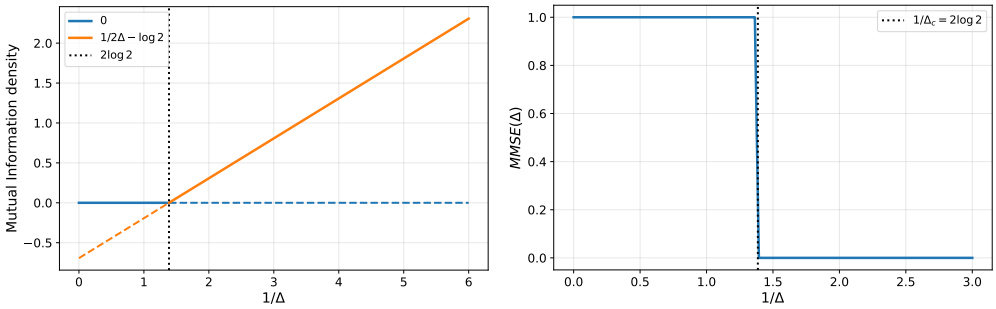

Figure 3: (**Left**) Asymptotic value of the mutual information density as a function of the snr for 1-sparse vector denoising. (**Right**) Minimum mean-squared error as a function of the snr.

Replica symmetry: Despite this simplification, solving the optimisation problem above is a daunting task. To make progress, we restrict our search to the following *replica symmetric* ansatz:

$$\begin{cases} m_a = m & a \in [r] \\ q_{ab} = q & 1 \leq a < b \leq r \end{cases} \tag{2.53}$$

In particular, by the Nishimori property we must also have $q = m \in \{0, 1\}$, with $m = 0$ denoting no recovery and $m = 1$ denote recovering. This considerable simplifies the free energy potential. In particular, we have:

$$\sum_{a=1}^{r} m_a = rm, \qquad \sum_{a,b=1}^{r} q_{ab} = r(1-q) + r^2 q \tag{2.54}$$

As for the entropy potential, we have that $\Sigma(m = 1) = \log 1 = 0$, as there is only a single way all replicas can align with the signal ($j_a = 1$ for all $a \in [r]$). On the other hand, for $m = 0$ we must have all replicas on distinct coordinates. The number of such configurations is:

$$(n-1)(n-2)\cdots(n-r) \underset{n\to\infty}{\asymp} n^r \tag{2.55}$$

Therefore:

$$\Sigma(m=0) = \lim_{d\to\infty} \frac{1}{d} \log (n-1)(n-2)\cdots(n-r) = r \log 2 \tag{2.56}$$

Therefore, we have:

$$\Phi_r(m) = \begin{cases} r(\frac{1}{2\Delta} + \log 2) & m = 0 \\ \frac{r}{\Delta} + \frac{r^2}{2\Delta} & m = 1 \end{cases} \tag{2.57}$$

Zero replica limit: Putting together allow us to take the $r \to 0^+$ limit explicitly:

$$\Phi(m) := \lim_{r\to 0^+} \partial_r \Phi_r(m) = \begin{cases} \frac{1}{2\Delta} + \log 2 & m = 0 \\ \frac{1}{\Delta} & m = 1 \end{cases} \tag{2.58}$$

and therefore:

$$\lim_{d\to\infty} \frac{1}{d} \mathbb{E}[\log Z_d(Y)] = \sup_{m\in\{0,1\}} \Phi(m) = \max\left\{ \frac{1}{2\Delta} + \log 2, \frac{1}{\Delta} \right\}$$

$$= \begin{cases} \log 2 + \frac{1}{2\Delta} & \text{if } \Delta > \frac{1}{2\log 2} \\ \frac{1}{\Delta} & \text{if } \Delta < \frac{1}{2\log 2} \end{cases} \tag{2.59}$$

with:

$$m_\star = \begin{cases} 0 & \text{if } \Delta > \frac{1}{2\log 2} \\ 1 & \text{if } \Delta < \frac{1}{2\log 2} \end{cases} \tag{2.60}$$

This implies that the asymptotic mutual information is given by:

$$i(\Delta) = \lim_{d\to\infty} i_d(\Delta) = \begin{cases} 0 & \text{if } \Delta > \frac{1}{2\log 2} \\ \frac{1}{2\Delta} - \log 2 & \text{if } \Delta < \frac{1}{2\log 2} \end{cases} \tag{2.61}$$

which gives a critical recovery threshold of $\Delta_c = \frac{1}{2\log 2}$. Recalling that:

$$\Delta = d\sigma^2 = \frac{\log n}{\log 2}\sigma^2 \tag{2.62}$$

We retrieve exactly the same threshold for signal recovery as the MLE/MAP in section 2.2.1!

Minimum mean-squared error: Finally, the I-MMSE theorem 2.2 allow us to compute the asymptotic MMSE, given by:

$$\mathrm{MMSE}(\Delta) = 1 - m_\star = \begin{cases} 1 & \text{if } \Delta > \frac{1}{2\log 2} \\ 0 & \text{if } \Delta < \frac{1}{2\log 2} \end{cases} \tag{2.63}$$

Note that to get this result, one needs to account for the difference of normalisation between the derivation of theorem 2.2 and eq. (2.39).

References

Donoho, David L and Iain M Johnstone (1994). "Minimax risk over l_q-balls for l_q-error". In: *Probability theory and related fields* 99.2, pp. 277–303.

Donoho, David Leigh (2000). *Counting bits with Kolmogorov and Shannon*. Department of Statistics, Stanford University.

Guo, Dongning, Shlomo Shamai, and Sergio Verdú (2004). "Mutual information and MMSE in Gaussian channels". In: *International Symposium onInformation Theory, 2004. ISIT 2004. Proceedings*. IEEE, pp. 349–349.

Mézard, Marc, Giorgio Parisi, Miguel Angel Virasoro, and David J Thouless (1988). *Spin glass theory and beyond*.