STAT260 Mean Field Asymptotics in Statistical Learning Lecture 5 - 02/03/2021

Lecture 5: $\mathbb{Z}_2$ synchronization and the free energy approach

Lecturer: Song Mei *Scriber: Omer Ronen* *Proof reader: Tae Joo Ahn*

1 Asymptotic risk in $\mathbb{Z}_2$ synchronization

1.1 Model setup

Let $\boldsymbol{\theta} = (\theta_1, \theta_2, \ldots, \theta_n) \in \Theta = \{\pm 1\}^n$ be the signal vector with $\theta_i \sim_{iid} \mathrm{Unif}(\mathbb{Z}_2 = \{\pm 1\})$. We observe $\boldsymbol{Y} \in \mathbb{R}^{n\times n}$ which is a noisy version of the signal $\boldsymbol{\theta}\boldsymbol{\theta}^\mathsf{T}$:

$$\boldsymbol{Y} = \frac{\lambda}{n}\boldsymbol{\theta}\boldsymbol{\theta}^\mathsf{T} + \boldsymbol{W} \in \mathbb{R}^{n\times n}. \tag{1}$$

The noise matrix follows $\boldsymbol{W} \sim \mathrm{GOE}(n)$, that is, we have $W_{ij} \sim_{iid} N(0, 1/n)$ for $1 \le i \le j \le n$ with the restriction that $W_{ij} = W_{ji}$, and $W_{ii} \sim_{iid} N(0, 2/n)$ for $1 \le i \le n$.

Our goal is to estimate $\boldsymbol{\theta}$ given its noisy observation $\boldsymbol{Y}$. For an estimator $\widehat{\boldsymbol{\Theta}} : \mathbb{R}^{n\times n} \to \mathbb{R}^{n\times n}$, we denote its expected risk by

$$R(\widehat{\boldsymbol{\Theta}}(\boldsymbol{Y})) = \mathbb{E}_{\boldsymbol{\theta}\sim\mathrm{Unif}(\mathbb{Z}_2^n), \boldsymbol{Y}\sim\mathbb{P}(\boldsymbol{Y}|\boldsymbol{\theta})}[\|\widehat{\boldsymbol{\Theta}}(\boldsymbol{Y}) - \boldsymbol{\theta}\boldsymbol{\theta}^\mathsf{T}\|_F^2]/n^2.$$

Our goal is to calculate the expected risk for certain interesting estimators.

1.2 Stochastic block model

The stochastic block model is a random graph model, that produces graphs containing communities. This model has a strong connection to the $\mathbb{Z}_2$ synchronization problem.

Suppose we have people coming from two groups (more generally, k groups):

- People from the same group form an edge with probability p independently ($p = a/n$)
- People from different groups form an edge with probability q independently ($q = b/n$)

We assume that $p > q$ and is scaled by n to obtain the desired property that the expected number of friends (edges) does not scale with the number of people in the population (n). We observe the adjacency matrix $\boldsymbol{A}$ and try to infer the two groups. Formally: let $\boldsymbol{\theta} \sim \mathrm{Unif}(\{\pm 1\}^n)$, we generate a graph $\mathcal{G} = (V, E)$, $V = \{1, 2, \ldots, n\}$. We define the two groups

$$V_+ = \{i \in V : \theta_i = +1\},\ V_- = \{i \in V : \theta_i = -1\}.$$

The adjacency matrix gives

$$A_{ij} \sim_{iid} \begin{cases} \mathrm{Ber}(p), & i, j \text{ are in the same group}, \\ \mathrm{Ber}(q), & i, j \text{ are not in the same group}. \end{cases}$$

We take $A_{ii} = 0$. Taking conditional expectation, we have that

$$\mathbb{E}[\boldsymbol{A}|\boldsymbol{\theta}] = \frac{1}{2}(p+q)\mathbf{1}\mathbf{1}^\mathsf{T} + \underbrace{\frac{1}{2}(p+q)\boldsymbol{\theta}\boldsymbol{\theta}^\mathsf{T}}_{\text{signal}} - p\mathbf{I}.$$

We define the de-noised version of $\boldsymbol{A}$ to be

$$\boldsymbol{Y} = \boldsymbol{A} - \frac{1}{2}(p+q)\mathbf{1}\mathbf{1}^\mathsf{T} + p\mathbf{I},$$

and we get that

$$\mathbb{E}[\boldsymbol{Y}|\boldsymbol{\theta}] = \frac{1}{2}(p-q)\boldsymbol{\theta}\boldsymbol{\theta}^{\mathsf{T}} = \frac{a-b}{2n}\boldsymbol{\theta}\boldsymbol{\theta}^{\mathsf{T}}.$$

To define our noise matrix $\boldsymbol{W}$, we center $\boldsymbol{Y}$ to have mean zero

$$\boldsymbol{W} = \boldsymbol{Y} - \frac{a-b}{2n}\boldsymbol{\theta}\boldsymbol{\theta}^{\mathsf{T}}. \tag{2}$$

Calculating the variance and neglecting terms on the order of n^{-2} we get

$$\mathrm{Var}(W_{ij}) = \frac{p(1-p)}{2} + \frac{q(1-q)}{2} \approx \frac{a+b}{2}.$$

Re-arranging (2) we get

$$\boldsymbol{Y} = \frac{a-b}{2}\boldsymbol{\theta}\boldsymbol{\theta}^{\mathsf{T}} + \boldsymbol{W}. \tag{3}$$

This form is very similar in structure to Eq. (1). Rescaling $\boldsymbol{W}$ to have similar variance values (with difference in distribution), we obtain the effective signal-to-noise parameter $\lambda = (a-b)/\sqrt{2(a+b)}$.

1.3 Estimators in $\mathbb{Z}_2$ synchronization

In the last lecture we have derived a few estimators for the parameter vector $\boldsymbol{\theta}$:

- MLE:

$$\hat{\boldsymbol{\theta}}_{\mathrm{ML}} = \arg\max_{\boldsymbol{\sigma}\in\{\pm 1\}^n} \langle \boldsymbol{\sigma}, \boldsymbol{Y}\boldsymbol{\sigma}\rangle$$

  Unfortunately it is NP-hard to compute this estimator.

- Spectral estimator:

$$\hat{\boldsymbol{\theta}}_{\mathrm{sp}}(\boldsymbol{Y}) = \arg\max_{\boldsymbol{\sigma}\in\mathbb{S}^{n-1}(\sqrt{n})} \langle \boldsymbol{\sigma}, \boldsymbol{Y}\boldsymbol{\sigma}\rangle = \sqrt{n} \quad \underbrace{\boldsymbol{v}_{\max}(\boldsymbol{Y})}_{\text{leading eigenvector}} .$$

  The spectral estimator is derived by relaxing the constraint set of the MLE ($\{\pm 1\}^n \subset \mathbb{S}^{n-1}(\sqrt{n})$), so that it can be computed efficiently. The matrix estimator gives

$$\widehat{\boldsymbol{\Theta}}_{\mathrm{sp}}(\boldsymbol{Y}) = \hat{\boldsymbol{\theta}}_{\mathrm{sp}}\hat{\boldsymbol{\theta}}_{\mathrm{sp}}^{T} \in \mathbb{R}^{n\times n}.$$

- SDP estimator:

$$\begin{aligned} \widehat{\boldsymbol{\Theta}}_{\mathrm{SDP}}(\boldsymbol{Y}) = \quad & \arg\max_{\boldsymbol{X}} \quad \langle \boldsymbol{Y}, \boldsymbol{X}\rangle \\ & \text{s.t.} \quad \boldsymbol{X} \succeq 0 \\ & \qquad\;\; X_{ii} = 1. \end{aligned}$$

  With the constraint that rank(X)=1, this estimator is equivalent to the MLE. So the SDP estimator is also a relaxation of the MLE.

- Bayes estimator:

$$\widehat{\boldsymbol{\Theta}}_{\mathrm{Bayes}}(\boldsymbol{Y}) = \mathbb{E}[\boldsymbol{\theta}\boldsymbol{\theta}^T|\boldsymbol{Y}] = \sum_{\boldsymbol{\sigma}\in\{\pm 1\}^n} \boldsymbol{\sigma}\boldsymbol{\sigma}^{\mathsf{T}}\mathbb{P}(\boldsymbol{\sigma}|\boldsymbol{Y}),$$

  where $\mathbb{P}(\boldsymbol{\sigma}|\boldsymbol{Y}) \propto \exp(\lambda\langle\boldsymbol{\sigma}, \boldsymbol{Y}\boldsymbol{\sigma}\rangle)/2)$.

1.4 Expected risk

Here we plot the expected risk $\mathbb{E}\|\widehat{\mathbf{\Theta}}(\boldsymbol{Y}) - \boldsymbol{\theta}\boldsymbol{\theta}^{\mathsf{T}}\|_F^2/n^2$ as a function of the signal to noise ratio parameter λ, for a few different estimators $\widehat{\mathbf{\Theta}}$.

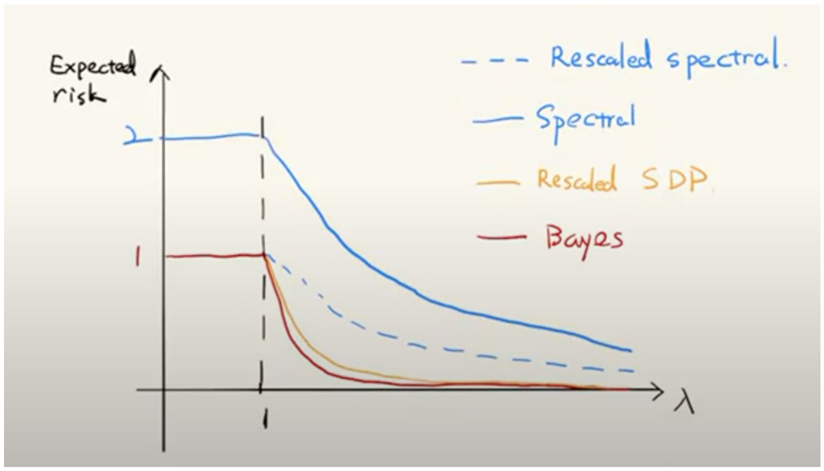

Figure 1: The asymptotic expected risk of the estimators ($n \to \infty$)

- As $\lambda \to \infty$ we can see that all estimators are consistent.
- For this problem the value $\lambda = 1$ is called "the information-theoretical threshold". For λ below this threshold, even the Bayes estimator is no better than the 0 estimator. There is a phase transition at $\lambda = 1$.
- The Bayes estimator can be computed efficiently in this model. That means, there's no statistical-computational gap in this model (in a few other models, this is not the case).

1.5 Asymptotic formula

Proposition 1. *Let $\hat{\boldsymbol{\theta}}_{\mathrm{sp}}$ be the spectral estimator. Then, we have almost surely*

$$\lim_{n\to\infty} \langle \hat{\boldsymbol{\theta}}_{\mathrm{sp}}, \boldsymbol{\theta} \rangle^2/n^2 = \begin{cases} 0 & \text{for } \lambda \le 1, \\ 1 - \frac{1}{\lambda^2} & \text{for } \lambda > 1. \end{cases}$$

This is known as **BBP phase transition**. *Furthermore, we have*

$$\lim_{n\to\infty} \|\hat{\boldsymbol{\theta}}_{\mathrm{sp}}\hat{\boldsymbol{\theta}}_{\mathrm{sp}}^{\mathsf{T}} - \boldsymbol{\theta}\boldsymbol{\theta}^{\mathsf{T}}\|_F^2/n = \begin{cases} 2 & \text{for } \lambda \le 1, \\ \frac{2}{\lambda^2} & \text{for } \lambda > 1. \end{cases}$$

To prove this we first show concentration and then calculate the limit. We will show this deviation in later lectures.

Proposition 2. *Let $\widehat{\mathbf{\Theta}}_{\mathrm{Bayes}}$ be the Bayes estimator. Then, we have almost surely*

$$\lim_{n\to\infty} \langle \boldsymbol{\theta}, \widehat{\mathbf{\Theta}}_{\mathrm{Bayes}}\boldsymbol{\theta} \rangle^2/n^2 = \begin{cases} 0 & \text{for } \lambda \le 1, \\ q_\star(\lambda)^2 & \text{for } \lambda > 1. \end{cases}$$

$$\lim_{n\to\infty} \|\widehat{\mathbf{\Theta}}_{\mathrm{Bayes}} - \boldsymbol{\theta}\boldsymbol{\theta}^{\mathsf{T}}\|_F^2/n^2 = \begin{cases} 1 & \text{for } \lambda \le 1, \\ 1 - q_\star(\lambda)^2 & \text{for } \lambda > 1. \end{cases}$$

Here $q_\star(\lambda)^2$ is the unique non-negative solution of

$$q = \mathbb{E}_{G\sim\mathcal{N}(0,1)}[\tanh(\lambda^2 q + \lambda\sqrt{q}G)^2].$$

2 Derivation of asymptotic risk

We now calculate the asymptotic formulas using the free energy approach. Our quantity of interest is

$$\lim_{n\to\infty} \mathbb{E}[\|\widehat{\boldsymbol{\Theta}}_{\text{Bayes}} - \boldsymbol{\theta}\boldsymbol{\theta}^{\mathsf{T}}\|_F^2]/n^2 = 1 - 2a_\star + c_\star,$$

with

$$\begin{aligned} a_\star &= \lim_{n\to\infty} \mathbb{E}[\langle \boldsymbol{\theta}, \widehat{\boldsymbol{\Theta}}_{\text{Bayes}}\boldsymbol{\theta}\rangle]/n^2, \\ c_\star &= \lim_{n\to\infty} \mathbb{E}[\|\widehat{\boldsymbol{\Theta}}_{\text{Bayes}}\|_F^2]/n^2. \end{aligned}$$

Remark 3. *General recipe for calculating $m_\star = \lim_{n\to\infty}\mathbb{E}[X_n]$ for some X_n*

1. *Find:*
   - *A configuration space Ω and a reference measure ν_0,*
   - *An observable $M : \Omega \to \mathbb{R}$,*
   - *A perturbed Hamiltonian $H_\lambda : \Omega \to \mathbb{R}, H_\lambda(\boldsymbol{\sigma}) = H_0(\boldsymbol{\sigma}) + \lambda M(\boldsymbol{\sigma})$,*
   - *so that $\mathbb{P}_{\beta,\lambda} \propto \exp(-\beta H_\lambda(\boldsymbol{\sigma}))$,*

   such that $\mathbb{E}[X_n] = \langle M\rangle_{\beta,\lambda}$ for some β, λ.
2. *Calculate the free energy density analytically*

$$f(\beta,\lambda) = \lim_{n\to\infty} -\frac{1}{n\beta}\mathbb{E}\Big[\log\int_\Omega \exp\{-\beta H_\lambda(\boldsymbol{\sigma})\}\nu_0(\mathrm{d}\boldsymbol{\sigma})\Big].$$

3. *$m_\star = \lim_{n\to\infty}\mathbb{E}[X_n] = \partial_\lambda f(\beta,\lambda)$.*

We now apply the general recipe to calculate $a_\star$ and $c_\star$.

2.1 Calculation of $a_\star$

$$\begin{aligned} a_\star &= \lim_{n\to\infty} \mathbb{E}[\langle \boldsymbol{\theta}, \widehat{\boldsymbol{\Theta}}_{\text{Bayes}}\boldsymbol{\theta}\rangle]/n^2 \\ &= \lim_{n\to\infty} \mathbb{E}[\langle \boldsymbol{\theta}, \sum_{\boldsymbol{\sigma}\in\{\pm1\}^n} \boldsymbol{\sigma}\boldsymbol{\sigma}^{\mathsf{T}}\mathbb{P}(\boldsymbol{\sigma}|\boldsymbol{Y})\boldsymbol{\theta}\rangle]/n^2 \\ &= \lim_{n\to\infty} \mathbb{E}[\sum_{\boldsymbol{\sigma}\in\{\pm1\}^n} (\langle \boldsymbol{\sigma}, \boldsymbol{\theta}\rangle^2/n^2)\mathbb{P}(\boldsymbol{\sigma}|\boldsymbol{Y})]. \end{aligned}$$

Now we arrive at an expression that looks like an ensemble average for the measure

$$\begin{aligned} \mathbb{P}(\boldsymbol{\sigma}|\boldsymbol{Y}) &\propto \exp\{\lambda\langle\boldsymbol{\sigma}, \boldsymbol{Y}\boldsymbol{\sigma}\rangle/2\} \\ &= \exp\Big\{-\lambda[-\frac{1}{2}\langle\boldsymbol{\sigma}, \boldsymbol{W}\boldsymbol{\sigma}\rangle - \frac{\lambda}{2n}\langle\boldsymbol{\theta},\boldsymbol{\sigma}\rangle^2]\Big\}. \end{aligned}$$

Suppose we define

- $\Omega = \{\pm1\}^n, \nu_0 = \text{Unif}$,

- $M(\boldsymbol{\sigma}) = -\langle \boldsymbol{\sigma}, \boldsymbol{\theta} \rangle^2 / 2n$,
- $H_\lambda(\boldsymbol{\sigma}) = -\langle \boldsymbol{\sigma}, \boldsymbol{W}\boldsymbol{\sigma} \rangle / 2 + \lambda M(\boldsymbol{\sigma})$,
- $\mathbb{P}_{\beta,\lambda}(\boldsymbol{\sigma}) \propto \exp(-\beta H_\lambda(\boldsymbol{\sigma}))$,
- $\mathbb{P}(\boldsymbol{\sigma}|\boldsymbol{Y}) = \mathbb{P}_{\beta,\lambda}(\boldsymbol{\sigma})|_{\beta=\lambda}$.

Then we have

$$a_\star = \lim_{n\to\infty} -2\mathbb{E}[\langle M \rangle_{\lambda,\lambda}]/n.$$

We will show in later lectures that, using the replica method, the free energy density gives:

$$f(\beta, \lambda) = \max_{b,q} f_{\rm mf}(b, q, \beta, \lambda).$$

for

$$f_{\rm mf}(b, q, \beta, \lambda) = -\frac{1}{4}\beta(1-q)^2 + \frac{1}{2}\lambda b^2 - \frac{1}{\beta}\mathbb{E}_{G\sim\mathcal{N}(0,1)}\Big[\log 2\cosh\Big(\beta(\lambda b + \sqrt{q}G)\Big)\Big].$$

The optimizers of the variational problem above solves the following self-consistent equations

$$b_\star = \mathbb{E}\Big[\tanh\big(\beta(\lambda b_\star + \sqrt{q_\star}G)\big)\Big], \tag{4}$$

$$q_\star = \mathbb{E}\Big[\tanh\big(\beta(\lambda b_\star + \sqrt{q_\star}G)\big)^2\Big]. \tag{5}$$

To compute $m_\star$, we take partial derivatives

$$m_\star = \partial_\lambda f(\beta, \lambda) \overset{(1)}{=} \frac{1}{2}b^2 - \frac{1}{\beta}\mathbb{E}\Big[\tanh\beta\big(\lambda b + \sqrt{q}G\big)\beta b\Big]|_{b=b_\star, q=q_\star} \overset{(2)}{=} \frac{1}{2}b_\star^2 - b_\star^2 = -\frac{1}{2}b_\star^2$$

and obtain our result

$$a_\star = -2m_\star(\beta,\lambda)|_{\beta=\lambda} = b_\star^2(\lambda,\lambda).$$

Here (1) comes from Implicit differentiation: $\partial_\lambda f_\star(\lambda) = \partial_\lambda[\max_q f(q,\lambda)] = \partial_\lambda f(q,\lambda)|_{q=q_\star}$ where $q_\star = \arg\max_q f(q,\lambda)$, and (2) comes from directly plugging in Eq. (4).

2.2 Calculation of $\mathbf{c}_\star$

We have

$$\begin{aligned}
\mathbb{E}[\|\widehat{\boldsymbol{\Theta}}_{\rm Bayes}\|_F^2]/n^2 &= \Big\langle \sum_{\boldsymbol{\sigma}\in\Theta} \boldsymbol{\sigma}\boldsymbol{\sigma}^\mathsf{T} P(\boldsymbol{\sigma}|\boldsymbol{Y}), \sum_{\boldsymbol{\sigma}\in\Theta} \boldsymbol{\sigma}\boldsymbol{\sigma}^\mathsf{T} P(\boldsymbol{\sigma}|\boldsymbol{Y}) \Big\rangle / n^2 \\
&= \sum_{\boldsymbol{\sigma}_1\in\Theta, \boldsymbol{\sigma}_2\in\Theta} \big\langle \boldsymbol{\sigma}_1\boldsymbol{\sigma}_1^\mathsf{T}, \boldsymbol{\sigma}_2\boldsymbol{\sigma}_2^\mathsf{T} \big\rangle P(\boldsymbol{\sigma}_1|\boldsymbol{Y}) P(\boldsymbol{\sigma}_2|\boldsymbol{Y})/n^2 \\
&= \sum_{\boldsymbol{\sigma}_1\in\Theta, \boldsymbol{\sigma}_2\in\Theta} \big\langle \boldsymbol{\sigma}_1, \boldsymbol{\sigma}_2 \big\rangle^2 P(\boldsymbol{\sigma}_1|\boldsymbol{Y}) P(\boldsymbol{\sigma}_2|\boldsymbol{Y})/n^2 \\
&= \sum_{(\boldsymbol{\sigma}_1, \boldsymbol{\sigma}_2)\in\Theta\times\Theta} \big\langle \boldsymbol{\sigma}_1, \boldsymbol{\sigma}_2 \big\rangle^2 \mu(\boldsymbol{\sigma}_1, \boldsymbol{\sigma}_2|\boldsymbol{Y})/n^2.
\end{aligned}$$

Here $\mu(\boldsymbol{\sigma}_1, \boldsymbol{\sigma}_2|\boldsymbol{Y}) = P(\boldsymbol{\sigma}_1|\boldsymbol{Y})P(\boldsymbol{\sigma}_2|\boldsymbol{Y}) \propto \exp(\lambda\langle \boldsymbol{\sigma}_1, \boldsymbol{Y}\boldsymbol{\sigma}_1\rangle/2 + \lambda\langle \boldsymbol{\sigma}_2, \boldsymbol{Y}\boldsymbol{\sigma}_2\rangle/2)$. We proceed to define the model:

- $\Omega = \Theta \times \Theta, \nu_0 = \mathrm{Unif}(\Theta) \times \mathrm{Unif}(\Theta)$,
- $M(\boldsymbol{\sigma}) = \langle \boldsymbol{\sigma}_1, \boldsymbol{\sigma}_2 \rangle^2 / n$,

- $H_{\lambda,h}(\boldsymbol{\sigma}) = -\frac{1}{2}\langle \boldsymbol{\sigma}_1, \boldsymbol{W}\boldsymbol{\sigma}_1\rangle - \frac{\lambda}{2n}\langle \boldsymbol{\theta}, \boldsymbol{\sigma}_1\rangle^2 - \frac{1}{2}\langle \boldsymbol{\sigma}_2, \boldsymbol{W}\boldsymbol{\sigma}_2\rangle - \frac{\lambda}{2n}\langle \boldsymbol{\theta}, \boldsymbol{\sigma}_2\rangle^2 + hM(\boldsymbol{\sigma})$,
- $\mathbb{P}_{\beta,\lambda,h}(\boldsymbol{\sigma}) \propto \exp(-\beta H_{\lambda,h}(\boldsymbol{\sigma}))$,

and we get a representation of $b_\star$ as an ensemble average of gibbs distribution

$$c_\star = \lim_{n\to\infty} \mathbb{E}\big[\langle M\rangle_{\beta,\lambda,h}\big]\Big|_{\beta=\lambda, h=0}.$$

2.3 Calculation of $m_\star = \lim_{n\to\infty} \mathbb{E}\langle \hat{\theta}_{\rm sp}(\boldsymbol{Y}), \theta\rangle^2/n^2$

The spectral estimator was defined to be

$$\hat{\theta}_{\rm sp}(\boldsymbol{Y}) = \sup_{\boldsymbol{\sigma}\in\mathbb{S}^{n-1}(\sqrt{n})} \langle \boldsymbol{\sigma}, \boldsymbol{Y}\boldsymbol{\sigma}\rangle.$$

We define

- $\Omega = \mathbb{S}^{n-1}$, ν_0 is the uniform distribution on $\mathbb{S}^{n-1}$,
- $M(\boldsymbol{\sigma}) = \langle \boldsymbol{\sigma}, \boldsymbol{\theta}\rangle^2/n$,
- $H_\lambda(\boldsymbol{\sigma}) = -\langle \boldsymbol{\sigma}, \boldsymbol{Y}\boldsymbol{\sigma}\rangle/2$,
- $\mathbb{P}_{\beta,\lambda,h}(\boldsymbol{\sigma}) \propto \exp\big(-\beta[H_\lambda(\boldsymbol{\sigma}) + hM(\boldsymbol{\sigma})]\big)$.

Then we have

$$\begin{aligned}
m_\star &= \lim_{n\to\infty} \mathbb{E}[M(\hat{\theta}_{\rm sp})]/n \\
&= \lim_{n\to\infty} \mathbb{E}[M(\arg\min_{\boldsymbol{\sigma}\in\Omega} H_\lambda(\boldsymbol{\sigma}))]/n \\
&\overset{(4)}{=} \lim_{n\to\infty}\lim_{\beta\to\infty} \mathbb{E}[\langle M\rangle_{\beta,\lambda,h}]/n|_{h=0} \\
&\overset{(5)}{=} \lim_{\beta\to\infty}\lim_{n\to\infty} \mathbb{E}[\langle M\rangle_{\beta,\lambda,h}]/n|_{h=0} \\
&= \lim_{\beta\to\infty}\lim_{n\to\infty} \mathbb{E}[\partial_h F(\beta,\lambda,h)]/n|_{h=0} \\
&= \lim_{\beta\to\infty} \partial_h f(\beta,\lambda,h)|_{h=0}.
\end{aligned}$$

Here (4) is from the low temperature limit of Gibbs measure (which concentrates on $\arg\min_{\boldsymbol{\sigma}} H(\boldsymbol{\sigma})$), and (5) we assume this change of limit is valid without justification here.

2.4 An exercise

Suppose $\boldsymbol{\beta}_0 \sim N(0, \boldsymbol{\sigma}_0^2 I_d), \boldsymbol{y} = \boldsymbol{X}\boldsymbol{\beta}_0 + \boldsymbol{\varepsilon}, X_{ij} \sim_{iid} N(0, 1/d)$ and $\varepsilon_i \sim_{iid} N(0, \boldsymbol{\sigma}^2)$. Denote the ridge estimator

$$\hat{\boldsymbol{\beta}} = \arg\min_{\boldsymbol{\beta}} \|\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta}\|_2^2 + \lambda\|\boldsymbol{\beta}\|_2^2 = (\boldsymbol{X}^\mathsf{T}\boldsymbol{X} + \lambda\mathbf{I})^{-1}\boldsymbol{X}^\mathsf{T}\boldsymbol{y}.$$

1. Please figure out:
    - A configuration space Ω, ν_0,
    - An observable $M : \Omega \to \mathbb{R}$,
    - A perturbed Hamiltonian $H_\lambda : \Omega \to \mathbb{R}$,

such that $\langle \hat{\boldsymbol{\beta}}, \boldsymbol{\beta}_0 \rangle / n = \lim_{\beta \to 0} \partial_\lambda F(\beta, \lambda)/n$ where $F(\beta, \lambda)$ is the free energy associated with Hamiltonian H_λ at temperature β.

2. Please repeat this for $\|\hat{\beta}\|_2^2/n$.

3. Hopefully, your $H_\lambda(\boldsymbol{\sigma})$ is of quadratic form and ν_0 is the Lebesgue measure. In this case,

$$\int_\Omega \exp\{-\beta H_\lambda(\boldsymbol{\sigma})\} \nu_0(\mathrm{d}\boldsymbol{\sigma})$$

is a Gaussian integration, and can be written explicitly. Please simplify $\mathbb{E}_{X,\beta}[F(\beta, \lambda)]$ as much as possible.