STAT260 Mean Field Asymptotics in Statistical Learning

Replica methods

Lecturer: Song Mei *Assignment 2 - Due on 3/10/2021*

Homework submissions are expected to be in pdf format produced by LaTeX. For programming exercises, you can use your favorite programming language, but matlab or Python are recommended. Grading of programming exercises will be based on the figures produced by the simulations, but you need also submit the source code in the original format (e.g. '.m' or '.py').

Exercise 1: Large deviation of the overlap matrix

Let $\boldsymbol{x}_0 = (x_{0,1}, \dots, x_{0,d})^\mathsf{T} \in \mathbb{R}^d$ with $\frac{1}{d}\sum_{i=1}^d \delta_{x_{0,i}} \to \mathbb{P}_0$ and $\mathbb{E}_{X\sim\mathbb{P}_0}[X^2] = 1$ (this will be satisfied with high probability if $x_{0,i} \sim_{iid} \mathbb{P}_0$). Let $\boldsymbol{x}_a \sim_{iid} \mathrm{Unif}([-1,1]^d)$ for $1 \le a \le k$. Denote $\boldsymbol{X} = (\boldsymbol{x}_a)_{0\le a\le k}$, and denote

$$\widehat{\boldsymbol{\mu}}(\boldsymbol{X}) = \Big(\langle \boldsymbol{x}_0, \boldsymbol{x}_a\rangle/d\Big)_{1\le a\le k} \in \mathbb{R}^k, \qquad \widehat{\boldsymbol{Q}}(\boldsymbol{X}) = \Big(\langle \boldsymbol{x}_a, \boldsymbol{x}_b\rangle/d\Big)_{1\le a,b\le k} \in \mathbb{R}^{k\times k}.$$

We further define (assuming the limit exists)

$$\mathrm{ent}(k, \boldsymbol{\mu}, \boldsymbol{Q}) \equiv \lim_{\varepsilon\to 0}\lim_{d\to\infty} \frac{1}{d}\log \mathbb{P}\Big(\|\widehat{\boldsymbol{\mu}}(\boldsymbol{X}) - \boldsymbol{\mu}\|_2^2 \le \varepsilon, \|\widehat{\boldsymbol{Q}}(\boldsymbol{X}) - \boldsymbol{Q}\|_F^2 \le \varepsilon\Big).$$

Question 1: Please use the field theoretic approach to calculate $\mathrm{Ent}(k, \boldsymbol{\mu}, \boldsymbol{Q})$.
Hint: You can start with the following heuristics (for $\boldsymbol{Q} = (q_{ab})_{0\le a,b\le k}$)

$$\mathbb{P}(\boldsymbol{Q}(\boldsymbol{X}) \approx \boldsymbol{Q}) \doteq \frac{1}{2^{dk}} \int_{[-1,1]^{dk}} \prod_{1\le a\le k} \delta(\langle \boldsymbol{x}_0, \boldsymbol{x}_a\rangle - d\cdot\mu_a) \prod_{1\le a\le b\le k} \delta(\langle \boldsymbol{x}_a, \boldsymbol{x}_b\rangle - d\cdot q_{ab}) \prod_{1\le a\le k} \mathrm{d}\boldsymbol{x}_a,$$

where $\int_{[-1,1]^d} f(\boldsymbol{x}_a)\mathrm{d}\boldsymbol{x}_a$ is integration with respect to Lebesgue measure on $[-1,1]^d$.

Question 2: Take

$$\boldsymbol{\mu} = \begin{bmatrix} \mu \\ \cdots \\ \mu \end{bmatrix}, \qquad \boldsymbol{Q} = \begin{bmatrix} q_1 & \cdots & q_0 \\ \cdots & \cdots & \cdots \\ q_0 & \cdots & q_1 \end{bmatrix}.$$

That is $\mu_a = \mu$, $q_{aa} = q_1$ for $1 \le a \le k$, and $Q_{ab} = q_0$ for $1 \le a \ne b \le k$.

Please simplify $\overline{\mathrm{ent}}(k, \mu, q_0, q_1) \equiv \mathrm{ent}(k, \boldsymbol{\mu}, \boldsymbol{Q})$.

Question 3: Give a simple expression for

$$T(\mu, q_0, q_1) \equiv \lim_{k\to 0} \frac{1}{k}\overline{\mathrm{ent}}(k, \mu, q_0, q_1).$$

Hint: Use the Gaussian identity formula and the replica trick $\lim_{k\to 0} k^{-1}\log \mathbb{E}[Z^k] = \mathbb{E}[\log Z]$.

Exercise 2: Replica formula for Box-LASSO

Let $\boldsymbol{x}_0 \in \mathbb{R}^d$ with $\frac{1}{d}\sum_{i=1}^d \delta_{x_{0,i}} \to \mathrm{Unif}(\{1,-1\})$ (this will be satisfied with high probability if $x_{0,i} \sim_{iid} \mathrm{Unif}(\{1,-1\})$). Let $\boldsymbol{A} \in \mathbb{R}^{n\times d}$ with $A_{ij} \sim_{iid} \mathcal{N}(0, 1/n)$. Let $\boldsymbol{\varepsilon} \in \mathbb{R}^n$ with $\varepsilon_i \sim_{iid} \mathcal{N}(0, \sigma^2)$. Assume all these random variables are independent. Let $\boldsymbol{y} = \boldsymbol{A}\boldsymbol{x}_0 + \boldsymbol{\varepsilon}$. Denote the Box-LASSO estimator by

$$\hat{\boldsymbol{x}} \equiv \arg\min_{\boldsymbol{x}\in[-1,1]^d} \frac{1}{2n}\|\boldsymbol{y} - \boldsymbol{A}\boldsymbol{x}\|_2^2,$$

and the training error by

$$L_n \equiv \min_{\boldsymbol{x}\in[-1,1]^d} \frac{1}{2n}\|\boldsymbol{y}-\boldsymbol{A}\boldsymbol{x}\|_2^2.$$

Question 1: Calculate $\lim_{n\to\infty}\mathbb{E}[L_n]$ using the replica method. Many results in class can be reused here. You will also use the results in Exercise 1. The answer is expected to be represented as some variational formula which involves integration with respect to some Gaussians.

Question 2: For $\psi:\mathbb{R}^2\to\mathbb{R}$, give a prediction for the expected observable

$$\lim_{d\to\infty, n/d\to\delta} \frac{1}{d}\sum_{i=1}^{d}\mathbb{E}[\psi(\hat{x}_i, x_{0,i})].$$

Especially, derive the asymptotics for (here we choose not to round the estimator $\hat{\boldsymbol{x}}$)

$$\lim_{d\to\infty, n/d\to\delta} \mathbb{E}[\|\hat{\boldsymbol{x}}-\boldsymbol{x}_0\|_2^2]/d.$$

Exercise 3: Simulating the risk of Box-LASSO (Programming exercise)

Please do the following simulation: for each $d = 50, 200, 800$, and take $n = 2d$, and each $\sigma^2 \in [0, 8]$ with grid size 0.2, obtain $n_s = 10$ **independent** samples of instances $(\boldsymbol{A}^{(i)}, \boldsymbol{x}_0^{(i)}, \boldsymbol{y}^{(i)})$, $i \in [n_s]$. Compute the Box-LASSO estimator $\hat{\boldsymbol{x}}$ as in Exercise 2. Denote

$$R = \frac{1}{d}\|\hat{\boldsymbol{x}}-\boldsymbol{x}_0\|_2^2.$$

Make an error bar plot of $(\hat{\mathbb{E}}[R], \widehat{\mathrm{std}}[R])$ versus σ^2 (σ^2 will be the x-axis), with $\hat{\mathbb{E}}[R]$ the mean variable and $\widehat{\mathrm{std}}[R]$ the error variable. Please plot the curve for $d = 50, 200, 800$ on the same figure.

Exercise 4: Prediction versus simulations (Extra credit)

Plot your prediction on the same figure (in Exercise 2). You earn the extra credit if the prediction and the simulation match.