STAT260 Mean Field Asymptotics in Statistical Learning Lecture 4 - 01/31/2021

Lecture 4: Statistical decision theory

Lecturer: Song Mei *Scriber: Alexander Tsigler* *Proof reader: Taejoo Ahn*

In this lecture we introduce the concepts of statistical decision theory and their relation to statistical physics. We demonstrate those relations using $\mathbb{Z}_2$ synchronization problem. We will show that the quantities of interest that arise in $\mathbb{Z}_2$ synchronization problem can be exactly translated into physical quantities in the Sherrington-Kirkpatrick spin glass model.

1 Statistical models, parameter space, likelihood function

In statistical decision theory an agent collects some (random) data and uses it to make the best action. We begin by introducing the model from which the data is generated, i.e. we give the following

Definition 1 (Statistical model)**.** *Statistical model is a family $\{\mathbb{P}_\theta : \theta \in \Theta\}$ of probability distributions on a common space $\mathcal{X}$ parametrized by $\theta \in \Theta$. Θ is called the parameter space of the statistical model.*

An observation X sampled from the distribution $\mathbb{P}_\theta$ gives the agent information about the unknown parameter θ. The quantity that measures how plausible it is for some value $\sigma \in \Theta$ to be the true parameter is given by the following

Definition 2 (Likelihood function)**.** *Suppose all measures in $\{\mathbb{P}_\theta : \theta \in \Theta\}$ have densities with respect to some reference measure ν. The likelihood of parameter $\boldsymbol{\sigma} \in \Theta$ given that $X = x$ is defined as*

$$L(\boldsymbol{\sigma}|x) = \frac{\mathrm{d}\mathbb{P}_{\boldsymbol{\sigma}}}{\mathrm{d}\nu}(x).$$

Remark 3. *In most cases of interest, either the set $\mathcal{X}$ is discrete or $\mathcal{X} \subseteq \mathbb{R}^d$ for some natural d. In the first case, the standard choice of ν is the counting measure, i.e. $L(\boldsymbol{\sigma}|x) = (\mathrm{d}\mathbb{P}_{\boldsymbol{\sigma}}/\mathrm{d}\nu)(x) = \mathbb{P}_{\boldsymbol{\sigma}}(X = x)$. In the second case, ν is usually the Lebesgue measure on $\mathbb{R}^d$.*

Remark 4. *The Hamiltonian of a physical system corresponds to negative log-likelihood of statistical model:*

$$H(\boldsymbol{\sigma}) = -\log L(\sigma|x), \quad \boldsymbol{\sigma} \in \Theta.$$

Example: $\mathbb{Z}_2$ syncronization. Let $\Theta = \{-1, 1\}^n$ be the parameter space. Our observation will be

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

where $\boldsymbol{W} \sim \mathrm{GOE}(n)$ (Gaussian orthogonal ensemble).

$$\begin{aligned} W_{i,j} &\sim_{i.i.d.} \mathcal{N}\Big(0, \frac{1}{n}\Big) && 1 \le i < j \le n, \\ W_{j,i} &= W_{i,j} && 1 \le i < j \le n, \\ W_{i,i} &\sim_{i.i.d.} \mathcal{N}\Big(0, \frac{2}{n}\Big) && 1 \le i \ \le n. \end{aligned}$$

The distribution of the random matrix $\boldsymbol{W}$ has the following density with respect to the Lebesgue measure on the space of real symmetric matrices:

$$p(\boldsymbol{W}) = \frac{1}{Z_n} \exp\{-n\|\boldsymbol{W}\|_F^2/4\}.$$

Therefore, we have for the likelihood

$$L(\sigma|\boldsymbol{Y}) = \frac{1}{Z_n}\exp\{-n\|\boldsymbol{Y} - \lambda\boldsymbol{\sigma}\boldsymbol{\sigma}^\mathsf{T}/n\|_F^2/4\},$$
$$\log L(\sigma|\boldsymbol{Y}) = -n\|\boldsymbol{Y} - \lambda\boldsymbol{\sigma}\boldsymbol{\sigma}^\mathsf{T}/n\|_F^2/4 - \text{const},$$

where the constant does not depend on σ. The scalar quantity λ is called **signal-to-noise ratio**. We assume that it is a known constant.

2 Loss function

Consider some space $\mathcal{A}$ that we will call **action space**. As we stated before, the goal of an agent is to choose "the best action" i.e. the element of the space $\mathcal{A}$ that minimizes some function $L : \mathcal{A}\times\Theta \to \mathbb{R}$. The function L is called **loss function**.

A rather common situation is when one wants to estimate the parameter θ, so $\mathcal{A} = \Theta$. A classical example is $\mathcal{A} = \Theta = \mathbb{R}$ and $L(a, \theta) = (a - \theta)^2$.

There are two possible choices of the action space and the corresponding loss function for the $\mathbb{Z}_2$ synchronization model:

- Vector square loss: $\mathcal{A} = [-1, 1]^n$ and $L(\boldsymbol{a}, \boldsymbol{\theta}) = \|\boldsymbol{a} - \boldsymbol{\theta}\|^2/n$.
- Matrix square loss: $\mathcal{A} = [-1, 1]^{n\times n}$ and $L(\boldsymbol{A}, \boldsymbol{\theta}) = \|\boldsymbol{A} - \boldsymbol{\theta}\boldsymbol{\theta}^\mathsf{T}\|_F^2/n^2$.

3 Statistical estimator

The goal of an agent is to minimize the loss. However, the loss depends on the unknown parameter θ, so the choice of agent's action can only depend on the observation X sampled from $\mathbb{P}_\theta$. This leads us to the following

Definition 5. *A function $\hat{\theta} : \mathcal{X} \to \mathcal{A}$ is called a statistical estimator.*

A classical example of statistical estimator for the case of parameter estimation (i.e. $\mathcal{A} = \Theta$) is the maximum likelihood estimator: $\hat{\theta}_{\text{ML}} = \arg\max_{\sigma\in\Theta} L(\sigma|X)$.

For $\mathbb{Z}_2$ synchronization model we can derive

$$\begin{aligned}
\hat{\boldsymbol{\theta}}_{\text{ML}} &= \arg\max_{\boldsymbol{\sigma}\in\Theta} L(\boldsymbol{\sigma}|\boldsymbol{Y}) \\
&= \arg\min_{\sigma\in\{-1,+1\}^n} \|\boldsymbol{Y} - \lambda\boldsymbol{\sigma}\boldsymbol{\sigma}^\mathsf{T}/n\|_F^2 \\
&= \arg\min_{\sigma\in\{-1,+1\}^n} \|\boldsymbol{Y}\|_F^2 - 2\lambda\langle\boldsymbol{\sigma}, \boldsymbol{Y}\boldsymbol{\sigma}\rangle/n + \lambda^2\|\boldsymbol{\sigma}\boldsymbol{\sigma}^\mathsf{T}\|_F^2/n^2 \\
&= \arg\max_{\sigma\in\{-1,+1\}^n} \langle\boldsymbol{\sigma}, \boldsymbol{Y}\boldsymbol{\sigma}\rangle.
\end{aligned}$$

4 Risk function

The notion of loss that we introduced depends the unknown parameter θ as well as the random observation X from $\mathbb{P}_\theta$, which makes it inconvenient to use loss to compare different estimators. The first step to overcome this difficulty is to integrate out the dependence on the random element X. This leads to the following

Definition 6 (Risk function)**.** *For a statistical estimator $\hat{\theta} : \mathcal{X} \to \mathcal{A}$ and a parameter $\theta \in \Theta$ the risk of that estimator is*

$$R(\hat{\theta}, \theta) = \mathbb{E}_{X\sim\mathbb{P}_\theta} L(\hat{\theta}(X), \theta) = \int_{\mathcal{X}} L(x, \theta)\mathbb{P}_\theta(\mathrm{d}x).$$

The function $R(\cdot, \cdot)$ is called the risk function.

5 Bayes optimality

Despite being non-random, our notion of risk still depends on the unknown θ, so it doesn't yet allow direct comparison of different estimators: one estimator can be better for some values of θ but worse for others. There are two most common approaches to fix this: Bayesian and minimax. We choose to take the Bayesian approach as its connections to statistical physics are more straightforward. Moreover, the minimax approach can be reformulated as Bayesian approach where one takes the worst possible prior for each estimator.

The Bayesian approach is to integrate over θ just as we integrated over X to eliminate the dependence on it. For a given measure Q on the space Θ we introduce the following

Definition 7 (Expected risk, Bayes risk, Bayes estimator)**.**

$$\begin{aligned} \textit{Expected risk is} \quad & R_B(\hat{\theta}, Q) = \int_\Theta R(\hat{\theta}, \theta) Q(\mathrm{d}\theta). \\ \textit{Bayes risk is} \quad & R_B(Q) = \inf_{\hat{\theta}: \mathcal{X} \to \mathcal{A}} R_B(\hat{\theta}, Q). \\ \textit{Bayes estimator is} \quad & \hat{\theta}_{\mathrm{Bayes}} = \arg\min_{\hat{\theta}: \mathcal{X} \to \mathcal{A}} R_B(\hat{\theta}, Q). \end{aligned}$$

The measure Q is often called the **prior distribution** of the parameter θ.

The main virtue of Bayesian approach is that for any $x \in \mathcal{X}$ the value of $\hat{\theta}_{\mathrm{Bayes}}(x)$ can be computed without any knowledge of $\hat{\theta}_{\mathrm{Bayes}}(y)$ for any other $y \in \mathcal{X}$. The formal statement of this result is given by the following

Theorem 8 (Bayes theorem)**.** *Bayes estimator minimizes the posterior expected value of a loss function:*

$$\hat{\theta}_{\mathrm{Bayes}}(x) = \arg\min_{a \in \mathcal{A}} \int_\Theta L(a, \boldsymbol{\sigma}) L(\boldsymbol{\sigma} | X) Q(\mathrm{d}\boldsymbol{\sigma}).$$

Proof. The full rigorous proof is left as an exercise to the reader. We only provide the main idea, which is to write the following

$$\begin{aligned} \hat{\theta}_{\mathrm{Bayes}} &= \arg\min_{\hat{\theta}} \int_\Theta \int_{\mathcal{X}} L(\hat{\theta}(x), \boldsymbol{\sigma}) \mathbb{P}_{\boldsymbol{\sigma}}(\mathrm{d}x) Q(\mathrm{d}\boldsymbol{\sigma}) \\ &= \arg\min_{\hat{\theta}} \int_\Theta \int_{\mathcal{X}} L(\hat{\theta}(x), \boldsymbol{\sigma}) L(\boldsymbol{\sigma} | x) \nu(\mathrm{d}x) Q(\mathrm{d}\boldsymbol{\sigma}) \\ &= \arg\min_{\hat{\theta}} \int_{\mathcal{X}} \Big(\int_\Theta L(\hat{\theta}(x), \boldsymbol{\sigma}) L(\boldsymbol{\sigma} | x) Q(\mathrm{d}\boldsymbol{\sigma}) \Big) \nu(\mathrm{d}x). \end{aligned}$$

The quantity inside the brackets only depends on the value of $\hat{\theta}$ in x, so to maximize the whole integral one can separately maximize $\int_\Theta L(\hat{\theta}(x), \boldsymbol{\sigma}) L(\boldsymbol{\sigma} | x) Q(\mathrm{d}\boldsymbol{\sigma})$ in $\hat{\theta}(x)$ for each x. □

Now let's see how Bayesian approach can be applied to $\mathbb{Z}_2$ synchronization. Consider the matrix squared loss $L(\boldsymbol{A}, \boldsymbol{\theta}) = \|\boldsymbol{A} - \boldsymbol{\sigma}\boldsymbol{\sigma}^\mathsf{T}\|_F^2 / n^2$, and choose Q to be the uniform distribution on $\Theta = \{-1, +1\}^n$. The Bayes estimator becomes

$$\hat{\theta}_{\mathrm{Bayes}}(\boldsymbol{Y}) = \arg\min_{\boldsymbol{A} \in [-1,1]^{n \times n}} \int_\Theta (\|\boldsymbol{A} - \boldsymbol{\sigma}\boldsymbol{\sigma}^\mathsf{T}\|_F^2 / n^2) L(\boldsymbol{\sigma} | \boldsymbol{Y}) Q(\mathrm{d}\boldsymbol{\sigma}).$$

To find the minimizer we can take the derivative in $\boldsymbol{A}$ and put it equal to zero. Plugging in $\nabla_{\boldsymbol{A}} \|\boldsymbol{A} - \boldsymbol{\sigma}\boldsymbol{\sigma}^\mathsf{T}\|_F^2 = 2(\boldsymbol{A} - \boldsymbol{\sigma}\boldsymbol{\sigma}^\mathsf{T})$ gives us

$$\hat{\theta}_{\mathrm{Bayes}}(\boldsymbol{Y}) = \int_\Theta \boldsymbol{\sigma}\boldsymbol{\sigma}^\mathsf{T} P(\boldsymbol{\sigma} | \boldsymbol{Y}) Q(\mathrm{d}\boldsymbol{\sigma}),$$

where

$$P(\boldsymbol{\sigma}|\boldsymbol{Y}) = \frac{L(\boldsymbol{\sigma}|\boldsymbol{Y})}{\int_\Theta L(\boldsymbol{\sigma}|\boldsymbol{Y})Q(\mathrm{d}\boldsymbol{\sigma})}.$$

In the following $\xi \propto \eta$ means that ξ/η does not depend on $\boldsymbol{\sigma}$. We write

$$\begin{aligned} L(\boldsymbol{\sigma}|\boldsymbol{Y}) &\propto \exp\{-n\|\boldsymbol{Y} - \lambda\boldsymbol{\sigma}\boldsymbol{\sigma}^\mathsf{T}/n\|_F^2/4\} \\ &= \exp\{-n\|\boldsymbol{Y}\|_F^2/4 - n\lambda^2\|\boldsymbol{\sigma}\boldsymbol{\sigma}^\mathsf{T}/n\|_F^2/4 + \lambda\langle\boldsymbol{\sigma}, \boldsymbol{Y}\boldsymbol{\sigma}\rangle/2\} \\ &\propto \exp\{\lambda\langle\boldsymbol{\sigma}, \boldsymbol{Y}\boldsymbol{\sigma}\rangle/2\}. \end{aligned}$$

To make the last transition we used the fact that $\boldsymbol{\sigma} \in \{-1, 1\}^n$, so $\|\boldsymbol{\sigma}\boldsymbol{\sigma}^\mathsf{T}/n\|_F^2 = 1$ — does not depend on $\boldsymbol{\sigma}$.

Finally, we see that the Bayes estimator is the average value of $\boldsymbol{\sigma}\boldsymbol{\sigma}^\mathsf{T}$ with respect to the measure on discrete cube $\{-1, 1\}^n$, whose density is proportional to $\exp\{\lambda\langle\boldsymbol{\sigma}, \boldsymbol{W}\boldsymbol{\sigma}\rangle/2 + \lambda\langle\boldsymbol{\sigma}, \boldsymbol{\theta}\rangle^2/(2n)\}$ (plug in $\boldsymbol{Y} = \boldsymbol{W} + \lambda\boldsymbol{\theta}\boldsymbol{\theta}^\mathsf{T}/n$). From this expression one can already see that the Bayesian estimator is exactly the average of the observable with respect to the Gibbs measure. We will further discuss the connections with statistical physics in the next section.

6 Connection of $\mathbb{Z}_2$ synchronization with Sherrington-Kirkpatrick spin glass

Recall the Sherrington-Kirkpatrick model that we defined previously:

- The configuration space is $\Omega = \{-1, +1\}^n$.
- The Hamiltonian is $H_{n,\lambda}(\boldsymbol{\sigma}) = -\langle\boldsymbol{\sigma}, \boldsymbol{W}\boldsymbol{\sigma}\rangle/2 - \lambda\langle\boldsymbol{\sigma}, \mathbf{1}\rangle^2/(2n)$, where $\boldsymbol{W} \sim \mathrm{GOE}$.
- The Gibbs measure is $P_{n,\beta,\lambda}(\boldsymbol{\sigma}) \propto \exp\{-\beta H_{n,\lambda}(\boldsymbol{\sigma})\}$.

We can immediately see the connections with $\mathbb{Z}_2$ synchronization:

- The configuration space of Sherrington-Kirkpatrick model is exactly the parameter space of $\mathbb{Z}_2$ synchronization model.
- If $\boldsymbol{\theta} = \mathbf{1}$, then the posterior distribution of the $\mathbb{Z}_2$ synchronization model is exactly the Gibbs distribution of Sherrington-Kirkpatrick model with inverse temperature $\beta = \lambda$. To extend this correspondence to general $\boldsymbol{\theta}$, one can generalize the definition of Sherrington-Kirkpatrick model by substituting vector $\mathbf{1}$ by an arbitrary vector $\boldsymbol{\theta}$ in the definition of $H_{n,\lambda}(\boldsymbol{\sigma})$.
- Under the same conditions as in the previous bullet point(i.e. $\beta = \lambda$ and $\boldsymbol{\theta}$ corresponds to $\mathbf{1}$), Bayes estimator of the $\mathbb{Z}_2$ synchronization model is exactly the average of the observable $\boldsymbol{\sigma}\boldsymbol{\sigma}^\mathsf{T}$ in Sherrington-Kirkpatrick model.
- The maximum likelihood estimator in the $\mathbb{Z}_2$ synchronization model is exactly the minimizer (the ground state) of the Hamiltonian in the Sherrington-Kirkpatrick model.

7 Key questions in $\mathbb{Z}_2$ synchronization

The main questions that we want to answer about the $\mathbb{Z}_2$ synchronization model are

- What is the asymptotic risk of $\hat{\theta}_{\mathrm{Bayes}}$, $\hat{\theta}_{\mathrm{ML}}$ and other statistical estimators of interest?
- How to efficiently compute those estimators?

Further we will see how connections with statistical physics help us answer these questions.