Lecture-18: Statistical decision theory

1 Sherrington-Kirkpatrick model

Definition 1.1. Consider a space of row vectors $\mathcal{X} \triangleq \mathbb{R}^{1\times d}$ for any finite $d \in \mathbb{N}$. The inner product $\langle\rangle : \mathcal{X}\times\mathcal{X} \to \mathbb{R}$ is defined as $\langle a, b\rangle \triangleq ab^T = \sum_{i=1}^d a_i b_i$ for all $a, b \in \mathcal{X}$. The outer product $\otimes : \mathcal{X}\times\mathcal{X} \to \mathbb{R}^{d\times d}$ is defined as

$$a \otimes b \triangleq a^T b = \sum_{i=1}^N \sum_{j=1}^N a_i b_j e_i^T e_j.$$

Definition 1.2. We define the Frobenius norm on space of real matrices as $\|\|_F : \mathbb{R}^{d\times d} \to \mathbb{R}_+$ for any real matrix $A \in \mathbb{R}^{d\times d}$ as $\|A\|_F \triangleq \sqrt{\sum_{i,j=1}^d a_{i,j}^2}$.

Remark 1. Let $A \in \mathbb{R}^{d\times d}$, then $(AA^T)_{ij} = \sum_{k=1}^d a_{i,k}a_{j,k}$ for all $i, j \in [d]$. It follows that $\mathrm{tr}(AA^T) = \sum_{i,k=1}^d a_{i,k}^2 = \|A\|_F^2$. We can define the Frobenius inner product $\langle\rangle_F : \mathbb{R}^{d\times d}\times\mathbb{R}^{d\times d} \to \mathbb{R}$ as $\langle A, B\rangle \triangleq \mathrm{tr}\, AB^T$ for any two matrices $A, B \in \mathbb{R}^{d\times d}$. It follows that

$$\|A - B\|_F^2 = \|A\|_F^2 + \|B\|_F^2 - \langle A, B\rangle_F - \langle B, A\rangle_F .$$

Definition 1.3. A random matrix $W : \Omega \to \mathbb{R}^{N\times N}$ is called a sample from **Gaussian orthogonal ensemble (N)** if (a) it is symmetric, i.e. $W_{i,j} = W_{j,i}$, and (b) all lower-diagonal matrix entries are independent zero-mean Gaussian random variables with the following variances

$$\mathbb{E}W_{i,j}^2 = \frac{1}{N}, \quad i \leqslant j, \qquad\qquad \mathbb{E}W_{i,i}^2 = \frac{2}{N}.$$

Lemma 1.4. *Consider a random matrix $W : \Omega \to \mathbb{R}^{N\times N}$ sampled from a Gaussian orthogonal ensemble. The density of W with respect to the Lebesgue measure on the space of real symmetric measures is*

$$p(W) \triangleq \frac{1}{Z_N} e^{-\frac{N}{4}\|W\|_F^2},$$

where the partition function $Z_N \triangleq \int_W e^{-\frac{N}{4}\|W\|_F^2} dW$.

Definition 1.5 (Sherrington-Kirkpatrick model). Consider a spin state in $\mathcal{Z} \triangleq \{-1, 1\}$ and state space of N interacting spin particles $\mathcal{X} \triangleq \mathcal{Z}^N$. We define a ferromagnetic model with N interacting spin particles with parametrized energy function $E_{\lambda,W} : \mathcal{X} \to \mathbb{R}$ in terms of parameter λ and random matrix W from Gaussian orthogonal ensemble (N), for spin configurations $\sigma \in \mathcal{X}$, as

$$E_{\lambda,W}(\sigma) \triangleq -\frac{1}{2}\sum_{i,j=1}^N W_{i,j}\sigma_i\sigma_j - \frac{\lambda}{2N}\sum_{i,j=1}^N \sigma_i\sigma_j.$$

Remark 2. We define row vector $1 \in \mathcal{Z}^{1\times N}$ where $1_i = 1$ for all $i \in [N]$. In terms of row vectors $\sigma, 1 \in \mathcal{Z}^{1\times N}$ we can write the energy as $E_{\lambda,W} = -\frac{1}{2}\langle\sigma W, \sigma\rangle - \frac{\lambda}{2N}\langle\sigma, 1\rangle^2$.

2 Statistical models

We show the relation between statistical decision theory and statistical physics by showing the one-to-one correspondence between the quantities of interest in the $\mathbb{Z}_2$ synchronization problem and the Sherrington-Kirkpatrick spin glass model.

Definition 2.1. We consider a **statistical model** to be a family $\mathcal{P}(\Theta, \mathcal{Z}) \triangleq \{P_\theta \in \mathcal{M}(\mathcal{Z}) : \theta \in \Theta\}$ of probability distributions on a common space $\mathcal{Z}$ parametrized by $\theta \in \Theta$, where Θ is the parameter space of the statistical model. We assume that observations $X : \Omega \to \mathcal{X} \triangleq \mathcal{Z}^N$ are *i.i.d.* with a common distribution P_θ for some parameter $\theta \in \Theta$.

Definition 2.2. Consider the case when all measures in the family $\mathcal{P}(\Theta, \mathcal{X})$ admit a density function with respect to a reference measure $\nu \in \mathcal{M}(\mathcal{Z})$. The **likelihood** of observation $\{X_n = z\}$ given a parameter $\theta \in \Theta$ is defined as

$$\mathcal{L}(\theta \mid z) \triangleq \frac{dP_\theta}{d\nu}(z).$$

Remark 3. We consider mostly the following two cases.

1. The set $\mathcal{Z}$ is discrete and the reference measure ν is the counting measure. In this case, $\mathcal{L}(\theta \mid z) = \frac{dP_\theta}{d\nu}(z) = P_\theta\{X_n = z\}$.
2. The set $\mathcal{Z} \subseteq \mathbb{R}^d$ for some $d \in \mathbb{N}$, and the reference measure is the Lebesgue measure on $\mathbb{R}^d$. In this case, $\mathcal{L}(\theta \mid z) = \frac{dP_\theta}{d\nu}(z) = p_\theta(z)$ is the density of P_θ evaluated at $z \in \mathcal{Z}$.

Remark 4. The energy of the physical system corresponds to the negative log-likelihood of the statistical model, i.e., $E(\theta) \triangleq -\ln \mathcal{L}(\theta \mid z)$ for all parameters $\theta \in \Theta$.

Example 2.3 ($\mathbb{Z}_2$ synchronization). Consider the parameter space $\Theta \triangleq \{-1, 1\}^N$ and samples W from Gaussian orthogonal ensemble-(N). For some fixed but unknown parameter $\theta_0 \in \Theta$, let the observations be samples of random vector $Y : \Omega \to \mathbb{R}^{N \times N}$ defined as

$$Y \triangleq \frac{\lambda}{N}\theta_0 \otimes \theta_0 + W.$$

From the density of W, we can write the likelihood of θ given an observation Y as

$$\mathcal{L}(\theta \mid Y) = \frac{1}{Z_N} e^{-\frac{N}{4}\left\|Y - \frac{\lambda}{N}\theta\otimes\theta\right\|_F^2}.$$

We can write the log-likelihood of unknown parameter θ in terms of observation Y as

$$\ell(\theta \mid Y) \triangleq \log \mathcal{L}(\theta \mid Y) = -\frac{N}{4}\left\|Y - \frac{\lambda}{N}\theta \otimes \theta\right\|_F^2 - \log Z_N.$$

We assume that λ is a known constant and is called the **signal-to-noise ratio**. We observe that $\log Z_N$ doesn't depend on parameter θ.

3 Statistical estimators

Definition 3.1. We denote the **prediction** or **action space** by $\mathcal{A}$, and the **loss function** by $L : \mathcal{A} \times \Theta \to \mathbb{R}$ that maps the loss of prediction from actual parameter to real number.

Remark 5. A common case is when prediction space $\mathcal{A} = \Theta = \mathbb{R}^d$ and $L(a, \theta) = \|a - \theta\|^2$ for all $a, \theta \in \mathbb{R}^d$.

Example 3.2 ($\mathbb{Z}_2$ synchronization). There are two common square loss models considered.

Vector square loss. The prediction space is taken as the convex hull of Θ so that

$$\mathcal{A} \triangleq \operatorname{cvx}(\Theta) \triangleq \{\lambda\theta + (1-\lambda)\theta' : \theta, \theta' \in \Theta, \lambda \in [0,1]\} = [-1,1]^N.$$

The loss function $L : \mathcal{A} \times \Theta \to \mathbb{R}$ is defined as $L(a,\theta) \triangleq \frac{1}{N}\|a-\theta\|^2$ for all $(a,\theta) \in \mathcal{A} \times \Theta$.

Matrix square loss. The prediction space is taken as the convex hull of $\Theta \otimes \Theta$ so that

$$\mathcal{A} \triangleq \operatorname{cvx}(\Theta \otimes \Theta) \triangleq \{\lambda\sigma + (1-\lambda)\sigma' : \sigma, \sigma' \in \Theta \otimes \Theta, \lambda \in [0,1]\} = [-1,1]^{N \times N}.$$

The loss function $L : \mathcal{A} \times \Theta \to \mathbb{R}$ is defined as $L(A,\theta) \triangleq \frac{1}{N^2}\|A-\theta\otimes\theta\|_F^2$ for all $(A,\theta) \in \mathcal{A} \times \Theta$.

Definition 3.3. A function $f : \mathcal{X} \to \mathcal{A}$ is called a statistical estimator.

3.1 Maximum likelihood estimator

Definition 3.4. A **maximum likelihood estimtor** $f_{\mathrm{ML}} : \mathcal{X} \to \mathcal{A} = \Theta$ is the one that maximizes the likelihood of a parameter $\theta \in \Theta$ for each observation $x \in \mathcal{X}$, i.e.,

$$f_{\mathrm{ML}}(x) \triangleq \arg\max \{\mathcal{L}(\theta \mid x) : \theta \in \Theta\}.$$

Remark 6. From the monotonicity of log function, it follows that the maximum likelihood estimator $f_{\mathrm{ML}}(x) = \arg\min\{-\ell(\theta \mid x) : \theta \in \Theta\}$ for all $x \in \mathcal{X}$.

Example 3.5 ($\mathbb{Z}_2$ synchronization). Recall that $\|Y\|_F^2$ doesn't change for any $\theta \in \Theta$. Further, we observe that $\theta \otimes \theta = \theta^T\theta$ is a real symmetric matrix and $\langle\theta,\theta\rangle = \operatorname{tr}\theta\theta^T = 1$ for all $\theta \in \Theta$. It follows that

$$\|\theta\otimes\theta\|_F^2 = \operatorname{tr}(\theta^T\theta)(\theta^T\theta) = \operatorname{tr}\theta^T\theta = 1.$$

The inner product $\left\langle Y, \theta^T\theta\right\rangle_F = \operatorname{tr}(Y\theta^T\theta) = \left\langle\theta, \theta Y^T\right\rangle$ and $\left\langle\theta^T\theta, Y\right\rangle_F = \operatorname{tr}(\theta^T\theta Y^T) = \left\langle\theta Y^T, \theta\right\rangle$. Thus, the maximum likelihood estimator for this problem is given by

$$\hat{\theta}_{\mathrm{ML}} = f_{\mathrm{ML}}(Y) = \arg\min\left\{\left\|Y - \frac{\lambda}{N}\theta\otimes\theta\right\|_F^2 : \theta \in \Theta\right\} = \arg\max\left\{\left\langle\theta, \theta Y^T\right\rangle : \theta \in \Theta\right\}.$$

3.2 Bayes estimator

Definition 3.6. Let $\mathcal{P}(\Theta, \mathcal{X})$ be a statistical model for observations $\mathcal{X}$ that admits density for each parameter. Then, the **risk function** $R_L : \mathcal{A}^{\mathcal{X}} \times \Theta \to \mathbb{R}$ for a loss function $L : \mathcal{A} \times \Theta \to \mathbb{R}$ is defined as $R_L(f,\theta) \triangleq \int_{\mathcal{X}} L(f(x),\theta)p_\theta(x)dx$, for any statistical estimator $f : \mathcal{X} \to \mathcal{A}$ and parameter $\theta \in \Theta$.

Definition 3.7. Let $Q \in \mathcal{M}(\Theta)$ be the **prior parameter distribution**, then the **expected risk** for a loss function $L : \mathcal{A} \times \Theta \to \mathbb{R}$ is defined as

$$R(Q,f,L) \triangleq \int_\Theta R_L(f,\theta)dQ(\theta).$$

The **Bayes risk function** $R_B : \mathcal{M}(\Theta) \times \mathbb{R}^{\mathcal{A}\times\Theta} \to \mathbb{R}$ is defined as $R_B(Q,L) \triangleq \inf\left\{R(Q,f,L) : f \in \mathbb{R}^{\mathcal{X}}\right\}$ for all prior distributions $Q \in \mathcal{M}(\Theta)$ and loss functions $L : \mathcal{A} \times \Theta \to \mathbb{R}$. The **Bayes estimator** $f_B : \mathcal{X} \to \mathcal{A}$ for a fixed prior distribution $Q \in \mathcal{M}(\Theta)$ and loss function $L : \mathcal{A} \times \Theta \to \mathbb{R}$ is defined as

$$f_B \triangleq \arg\min\left\{R(Q,f,L) : f \in \mathcal{A}^{\mathcal{X}}\right\}.$$

Proposition 3.8 (Bayes). *For a fixed prior distribution $Q \in \mathcal{M}(\Theta)$ and a loss function $L : \mathcal{A} \times \Theta \to \mathbb{R}$, the Bayes estimator minimizes the posterior expected value of this loss function, i.e.,*

$$f_B(x) \triangleq \arg\min \left\{ \int_\Theta L(a,\theta)\mathcal{L}(\theta \mid x) dQ(\theta) : a \in \mathcal{A} \right\}, \quad x \in \mathcal{X}. \tag{1}$$

Proof. For a fixed prior distribution Q and a loss function L, we can write the Bayes estimator as

$$f_B = \arg\min \left\{ \int_\Theta dQ(\theta) \int_\mathcal{X} L(f(x),\theta) p_\theta(x) dx : f \in \mathcal{A}^\mathcal{X} \right\}.$$

Recall that $\mathcal{L}(\theta \mid x) = p_\theta(x)$. Then, by interchanging the order of the integrals, we obtain

$$f_B = \arg\min \left\{ \int_\mathcal{X} dx \int_\Theta dQ(\theta) L(f(x),\theta)\mathcal{L}(\theta \mid x) : f \in \mathcal{A}^\mathcal{X} \right\}.$$

The result follows from the fact that the inner integral depends only on $a = f(x)$ for each $x \in \mathcal{X}$, and it is minimized for each x by the Bayes estimator defined in (**??**). □

Example 3.9 ($\mathbb{Z}_2$ synchronization). For $\mathcal{A} = \operatorname{cvx}(\Theta \otimes \Theta) = [-1,1]^{N\times N}$, the matrix squared loss function $L(A,\theta) = \|A - \theta \otimes \theta\|_F^2$ for all $(A,\theta) \in \mathcal{A} \times \Theta$, and uniform distribution $Q \in \mathcal{M}(\Theta)$, the Bayes estimator is

$$f_B(Y) = \arg\min \left\{ \int_\Theta \|A - \theta \otimes \theta\|_F^2 \, \mathcal{L}(\theta \mid Y) dQ(\theta) : A \in [-1,1]^{N\times N} \right\}.$$

Recall that $\nabla_A \|A - \theta \otimes \theta\|_F^2 = \nabla_A \operatorname{tr}(A - \theta^T\theta)(A^T - \theta^T\theta) = 2(A - \theta^T\theta)$. It follows that the Bayes estimator $f_B(Y)$ is the solution to the following equation

$$f_B(Y) \int_\Theta \mathcal{L}(\theta \mid Y) dQ(\theta) = \int_\Theta \theta^T\theta \mathcal{L}(\theta \mid Y) dQ(\theta).$$

Using the fact that Q is a uniform distribution, we can write the Bayes estimator as

$$f_B(Y) = \sum_{\theta\in\Theta} \theta^T\theta \frac{\mathcal{L}(\theta \mid Y)}{\sum_{\theta\in\Theta} \mathcal{L}(\theta \mid Y)}.$$

We further recall that the likelihood function is

$$\mathcal{L}(\theta \mid Y) \propto e^{-\frac{N}{4}\left\|Y - \frac{\lambda}{N}\theta^T\theta\right\|_F^2}.$$

Since $\left\|\theta^T\theta\right\|_F^2 = 1$ and $\|Y\|_F^2$ remains same for all $\theta \in \Theta$, we obtain

$$\mathcal{L}(\theta \mid Y) \propto e^{\frac{\lambda}{2}\left\langle \theta, \theta Y^T \right\rangle}.$$

Recall that W is real-symmetric, and so is $Y = \frac{\lambda}{N}\theta_0^T\theta_0 + W$. It follows that $Y = Y^T$ and

$$\frac{\lambda}{2}\left\langle \theta, \theta Y^T \right\rangle = \frac{\lambda}{2}\left\langle \theta, \theta W + \frac{\lambda}{N}\left\langle \theta, \theta_0 \right\rangle \theta_0 \right\rangle = \frac{\lambda}{2}\left\langle \theta, W \right\rangle + \frac{\lambda^2}{2N}\left\langle \theta, \theta_0 \right\rangle^2.$$

Defining $\nu_{\lambda,W} \triangleq \frac{\mathcal{L}(\theta|Y)}{\sum_{\theta\in\Theta}\mathcal{L}(\theta|Y)} \in \mathcal{M}(\Theta)$, we observe that $f_B = \sum_\theta \theta^T\theta \nu_{\lambda,W}(\theta)$.

3.3 Relation between $\mathbb{Z}_2$ synchronization and Sherrington-Kirkpatrick model

We observe that the configuration space in the Sherrington-Kirkpatrick model is identical to the parameter space in $\mathbb{Z}_2$ synchronization. If $\theta_0 = 1$, then $\mu_\lambda = \nu_{\lambda,W}$. In this case, $f_B = \left\langle \theta^T\theta \right\rangle$ and f_{ML} is the ground state configuration that minimizes energy $E_{\lambda,W}$.