Mean Field Asymptotics in Statistical Learning. Jan 31 th.

Lecture 4: Statistical decision theory, $\mathbb{Z}_2$ synchronization problem, and Spin glass models.

① Statistical models, parameter space, likelihood function.

Concept: A statististical model is family of distribution P. on a common space $\mathcal{X}$. parameterized by $\theta \in \Theta$.

$$P = \{ P_\theta : \theta \in \Theta \}.$$

Θ: parameter space. (Configuration space)

$X \sim \mathbb{P}_\theta$: Sample under measure $\mathbb{P}_\theta$.

Likelihood function: $L(\sigma \mid x) = \mathbb{P}_\sigma(x) = \mathbb{P}_\sigma(X = x)$ $\sigma \in \Theta$.

$H(\sigma) = -\log L(\sigma \mid x) : \Theta \to \mathbb{R}$. (Hamiltonian)

$\mathbb{Z}_2$ synchrozation:

Let $\theta \in \Theta_n = \{\pm 1\}^n$ parameter spae

Observation $Y = \underset{\text{scaler}}{\lambda} \underbrace{\frac{\theta\theta^T}{n}}_{\mathbb{R}^{n\times n}} + \underset{\text{matrax } \in \mathbb{R}^{n\times n}}{W} \in \mathbb{R}^{n\times n}$

$W \sim GOE(n)$, (Gaussian orthogonal ensemble).

$W_{ij} \sim_{i.i.d.} N(0, \frac{1}{n})$ $1 \le i < j \le n$, $\perp$ $W_{ii} \sim_{i.i.d.} N(0, \frac{2}{n})$

$W_{ji} = W_{ij}$ $1 \le i < j \le n$. $n \sum_{i=1}^{n} W_{ii}^2/4 + n \sum_{1\le i<j\le n} W_{ij}^2/2$

$P(W) = \frac{1}{Z_n} \exp\{-n\|W\|_F^2/4\}$. Z_n doesn't depend on W

$\lambda \ge 0$: signal-to-noise ratio. constant and known

$P_\sigma(Y) = \frac{1}{Z_n} \exp\{-n\|Y - \lambda\sigma\sigma^T/n\|_F^2/4\}$ $\sigma \in \Theta_n$.

Likelikehood function.

$L(\sigma \mid Y) = -\log P_\sigma(Y) = n\|Y - \lambda\sigma\sigma^T/n\|_F^2/4 - \text{const}.$

② Loss function.

Sometimes $\mathcal{A} = \Theta$

Concept: $L : \mathcal{A} \times \Theta \to \mathbb{R}$.

$(a, \theta) \mapsto \mathbb{R}$.

$\Theta = \mathcal{A} = \mathbb{R}$.

Example: $L(a, \theta) = (a - \theta)^2$.

$\mathbb{Z}_2$ sync:

a) Vector square loss.

$\mathcal{A} = [-1, 1]^n$: $L : \mathcal{A} \times \Theta \to \mathbb{R}$

$(a, \theta) \mapsto \| a - \theta \|_2^2 / n$.

b) Matrix square loss.

$A = [-1, 1]^{n \times n}$: $L : [-1, 1]^{n \times n} \times \{\pm 1\}^n \to \mathbb{R}$

$(A, \theta) \mapsto \frac{1}{n^2} \| A - \theta\theta^T \|_F^2$

③ Statistical estimator.

Concept: A statistical estimator $\hat{\theta} : \mathcal{X} \to \mathcal{A}$.

Example: Maximum likelihood. $\hat{\theta}_{ML}(x) = \underset{\sigma \in \Theta}{\arg\max}\ L(\sigma | x)$.

(Asymptotic efficient in low dimension).

$\mathbb{Z}_2$ sync

MLE

$\sigma \in \{\pm 1\}^n$

$$\begin{aligned}\hat{\theta}_{ML}(Y) &= \underset{\sigma \in \Theta_n}{\arg\max}\ L(\sigma | Y) \\ &= \underset{\sigma}{\arg\min}\ \| Y - \lambda \sigma\sigma^T / n \|_F^2 . \\ &= \underset{\sigma}{\arg\min}\ \| Y \|_F^2 - 2\lambda \langle \sigma, Y\sigma \rangle / n + \lambda^2 \frac{\| \sigma\sigma^T \|_F^2}{n^2} \\ &= \underset{\sigma \in \{\pm 1\}^n}{\arg\max}\ \langle \sigma, Y\sigma \rangle .\end{aligned}$$

$\| \sigma \|_2^4$

④ Risk function.

$L : A \times \Theta \to \mathbb{R}$

$\hat{\theta} : \mathcal{X} \to A$

Sample: $x \in \mathcal{X}$ $x \sim \mathbb{P}_\theta$. $\theta \in \Theta$.

Concept:

$X \sim \mathbb{P}_\theta$

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

Measure of the quality of a statistical estimator $\hat{\theta}$,

$\mathbb{Z}_2$ sync:

$$R(\hat{\theta}_{ML}, \theta) = \mathbb{E}_\theta \| \hat{\theta}_{ML}(Y) - \theta \|_2^2 / n .$$

Question: Given two estimator $\hat{\theta}_1$, $\hat{\theta}_2$, How to fairly compare?

Two approach: a) Bayes risk b) Minimax.

⑤ Bayes optimality.

Concept : Given a prior Q on the space Θ. (Reference measure)

Expected risk : $R_B(\hat{\theta}, Q) = \int_\Theta R(\hat{\theta}, \theta)\, Q(d\theta)$.

Bayes risk : $R_B(Q) = \inf_{\hat{\theta} : \mathcal{X} \to A} R_B(\hat{\theta}, Q) = R_B(\hat{\theta}_{Bayes}, Q)$.

Bayes estimator : $\hat{\theta}_{Bayes} = \operatorname{argmin}_{\hat{\theta} : \mathcal{X} \to A} R_B(\hat{\theta}, Q)$.

Thm (Bayes thm) Bayes estimator minizes the posterior expected value of a loss function.

$$\hat{\theta}_{Bayes}(x) = \operatorname{argmin}_a \int_\Theta L(a, \theta)\, \underbrace{p(\sigma|x)}_{\uparrow\ \frac{p(\sigma, x)}{p(x)}}\, d\sigma$$

Proof: $\hat{\theta}$ minimize $R_B(\hat{\theta}, \theta) = \int L(\hat{\theta}(x), \sigma)\, \mathbb{P}_\sigma(x)\, Q(\sigma)\, dx\, d\sigma$

iff $\forall x \in \mathcal{X}$, $\hat{\theta}(x) = a$ where "a" minimize] exercise

$$\int L(a, \sigma) \underbrace{\mathbb{P}_\sigma(x) Q(\sigma)}_{= p(x, \sigma) = p(\sigma|x) \times p(x)} d\sigma. \quad \square.$$

$\mathbb{Z}_2$ sync : $L(A, \theta) = \| A - \theta\theta^T \|_F^2 / n^2$.

$Q = \mathbb{P}_0$ uniform dist. $\{\pm 1\}^n$.

$$\hat{\Theta}_{Bayes}(\tilde{Y}) = \operatorname{argmin}_A \int_\Theta \left(\| A - \sigma\sigma^T \|_F^2 / n^2 \right) p(\sigma|\tilde{Y})\, d\sigma$$

$$= \int_\Theta \sigma\sigma^T p(\sigma|\tilde{Y})\, d\sigma$$

$$= \mathbb{E}[\theta\theta^T \mid Y = \cdot\,]\Big|_{\cdot = \tilde{Y}}$$

$$p(\sigma|\tilde{Y}) \equiv \frac{p(\sigma, \tilde{Y})}{p(\tilde{Y})} = \frac{\mathbb{P}_0(\sigma)\, \mathbb{P}_\sigma(\tilde{Y})}{\int \mathbb{P}_0(\sigma)\, \mathbb{P}_\sigma(\tilde{Y})\, d\sigma}$$

$\sigma \in \Theta$.

$$p(\sigma|\tilde{Y}) \propto \mathbb{P}_0(\sigma)\, \mathbb{P}_\sigma(\tilde{Y})$$

$$\propto \exp\{ -n \| \tilde{Y} - \lambda\sigma\sigma^T/n \|_F^2 / 4 \}\, \mathbb{P}_0(\sigma) \qquad \|\sigma\sigma^T\|_F^2 = \|\sigma\|_2^4 = n^2$$

$$= \exp\{ -n [\underbrace{\|\tilde{Y}\|_F^2}_{\text{Const}} - 2\langle \sigma, \tilde{Y}\sigma \rangle / n + \underline{\frac{\lambda^2 \|\sigma\sigma^T\|_F^2}{n^2}}] / 4 \}\, \underline{\mathbb{P}_0(\sigma)}$$

$$\propto \exp\{ \langle \sigma, \tilde{Y}\sigma \rangle / 2 \}\, \mathbb{P}_0(\sigma).$$

⑥. Connection $\mathbb{Z}$ sync with SK.

Gibbs dist. of SK.

$$\Omega = \{\pm 1\}^n.$$

$$\frac{\lambda}{2n}\langle \sigma, \theta\rangle^2$$

$$H_{n,\lambda}(\sigma) = -\frac{1}{2}\langle \sigma, W\sigma\rangle - \underline{\frac{\lambda}{2n}\langle \sigma, 1\rangle^2} \qquad \underline{+h\langle \sigma, \theta\rangle}$$

$$P_{n,\beta,\lambda}(\sigma) \propto \exp\{-\beta H_{n,\lambda}(\sigma)\}.$$

$$\partial_h \phi(\beta,\lambda,h)\big|_{\beta=\lambda}.$$

$$= \lim_{n\to\infty} \langle \theta, \hat\theta_{Bayes}\rangle/n$$

$$\langle \sigma \rangle_{\beta,\lambda} = \sum_{\sigma\in\Omega} \sigma\, P_{n,\beta,\lambda}(\sigma).$$

$$\partial_\lambda \phi(\beta,\lambda,h)\big|_{\beta=\lambda}$$

$$\overset{?}{=} \lim_{n\to\infty} \|\hat\theta_{Bayes}\|_2^2/n$$

Bayes posterior $\mathbb{Z}_2$ sync.

$$\Theta_n = \{\pm 1\}^n. \qquad Y = \frac{\lambda}{n}\theta\theta^T + W.$$

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

$$\hat\theta_{Bayes}(Y) = \sum_{\sigma\in\Theta} \sigma\, P(\sigma|Y).$$

$$\hat\theta_{ML}(Y) = \arg\max_{\sigma} P_\sigma(Y) = \arg\max_{\sigma} \langle \sigma, Y\sigma\rangle.$$

Connection: $P(\sigma|Y) \propto \exp\{-\lambda[-\langle \sigma, W\sigma\rangle/2 - \frac{1}{2n}\langle \sigma,\theta\rangle^2]\}$

$$= P_{n,\beta=\lambda,\lambda}(\sigma).$$

$$\langle \sigma\rangle_{\lambda,\lambda} = \hat\theta_{Bayes}(Y).$$

$$\hat\theta_{ML}(Y) = \arg\max_{\sigma} \langle \sigma, Y\sigma\rangle = \arg\min_{\sigma} H_{n,\lambda}(\sigma).$$

⑦ Key questions in $\mathbb{Z}_2$ sync.

- What is the asymptotic risk of $\hat\theta_{Bayes}$, $\hat\theta_{ML}$, $\hat\theta_{spectr}$... ?
- How to efficiently compute

$$\lim_{n\to\infty} \|\hat\theta_{Bayes} - \theta\|_2^2/n \longleftrightarrow \lim_{n\to\infty} \mathbb{E}_W[f(\langle \sigma\rangle_{\beta,\lambda})]$$

$$\lim_{n\to\infty} \mathbb{E}_W[\langle f(\sigma)\rangle_{\beta,\lambda}]$$

$$\mathbb{E}_{\substack{Y\sim\mathbb{P}_\theta \\ \theta\sim\mathbb{P}_0}}$$