Infotheory for Statistics and Learning

Lecture 8

- Selected recap
  - Basics statistical decision theory [PW, Chap. 28]
  - Variational representation of f-divergence [PW, Sect. 7.13]
- Statistical (lower) bounds [PW, Chap. 29]
  - Hammersley-Chapman-Robbins bound
  - Cramér-Rao bound
  - Fisher information

Framework of Statistical Decision Problem

Statistical experiment: Nature picks distribution with **parameter** θ from the set of probability distributions defined on a common probability space $(\mathcal{X}, \mathcal{F})$

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

- **Data** $X \sim P_\theta$ is observed
  - can be a random variable, vector, process etc. depending on $\mathcal{X}$

Estimator: We want to estimate $T(\theta)$ which is defined on $\mathcal{Y}$, which can be a θ itself, a relevant aspect or a function of θ.

- **Decision rule:** Compute $\hat{T} \in \hat{\mathcal{Y}}$ based on observed data X

$$\hat{T} : \mathcal{X} \to \hat{\mathcal{Y}}$$

  - randomized estimator $\hat{T} = \hat{T}(X, U)$, external RV U or $P_{\hat{T}|X}$

Choice of estimator depends on different factors including estimator properties, but mostly on the performance objective.

- **Loss function:**

$$l : \mathcal{Y} \times \hat{\mathcal{Y}} \to \mathbb{R}, \quad T \times \hat{T} \mapsto l(T, \hat{T})$$

  - example: $T(\theta) = \theta$ and $l(\theta, \hat{\theta}) = \|\theta - \hat{\theta}\|_2^2$

- **Risk** of estimator $\hat{T}$ at θ:

$$R_\theta = E_\theta[l(T, \hat{T})] = \int l(T(\theta), \hat{t}) P_{\hat{T}|X}(\hat{t}|x) P_\theta(x)\, d(x, \hat{t})$$

  - $P_{\hat{T}|X}(\hat{t}|x)$ denotes the likelihood of $\hat{t}$ after observing x
  - log-likelihood function $\log P_{\hat{T}|X}(\hat{t}|x)$ is sometimes numerically beneficial, e.g, when x denotes a vector of iid observations
  - converses correspond to lower bounds on the optimal loss/risk (achievable results/implementations are upper bounds)

Maximum Likelihood Estimator

- **Maximum Likelihood (ML) estimator.** Maximize the likelihood (fct) over parameter θ so that the observed data x is most likely
  - e.g. $T(\theta) = \theta$
  $$\hat{T}(x) = \arg\max_{\theta \in \Theta} P_\theta(x)$$
- **Gaussian Location Model** (Additive Gaussian Noise)
  - $\mathcal{P} = \{\mathcal{N}(\theta, \sigma^2) : \theta \in \mathbb{R}\}$
  - $X_i = \theta + Z_i$ with $Z_i \overset{iid}{\sim} \mathcal{N}(0, \sigma^2)$
  - likelihood (fct) after observing $x_1, \ldots, x_n$:
  $P_\theta(x_1^n) = \prod_{i=1}^n P_\theta(x_i) = \prod_{i=1}^n \frac{1}{\sqrt{2\pi\sigma^2}} \exp(-\frac{(x_i-\theta)^2}{2\sigma^2}) = \frac{1}{(\sqrt{2\pi\sigma^2})^n} \exp(-\frac{1}{2\sigma^2} \sum_{i=1}^n (x_i - \theta)^2)$
    - Note that $P_\theta(x_1^n)$ is maximized if we minimize $\sum_{i=1}^n (x_i - \theta)^2$ $0 = \frac{d}{d\theta} \sum_{i=1}^n (x_i - \theta)^2 = \sum_{i=1}^n -2(x_i - \theta)$ so that the minimizer is $\theta = \frac{1}{n} \sum_{i=1}^n x_i$

$\Rightarrow$ ML estimate $\hat{T}(x_1, \ldots, x_n) = \frac{1}{n} \sum_{i=1}^n x_i$

Fundamental limit – "Best estimator"

Performance is measured by the risk

$$R_\theta(\hat{\theta}) = E_\theta[l(\theta, \hat{\theta})]$$

Approaches to identify *a best* estimator

- **Naïve method:** Search for estimator $\hat{\theta}$ that is better than all other estimator θ' for all $\theta \in \Theta$, i.e. $R_\theta(\hat{\theta}) \leq R_\theta(\theta') \forall \theta' \forall \theta$.
  - often there does not exists one $\hat{\theta}$ that is uniformly the best

Standard approaches that reduce the candidate set

- **Method 1:** Limit the class of competitors of $\hat{\theta}$
  - e.g. restricting to unbiased estimators or invariant estimators
- **Method 2:** Bayes (Bayesian) approach - average analysis
- **Method 3:** Minimax approach - worst-case analysis

Bayes risk

Average risk analysis with **prior** probability distribution π on Θ

$$R_\pi(\hat{\theta}) = E_{\theta\sim\pi} R_\theta(\hat{\theta}) = E_{\theta,X}[l(\theta,\hat{\theta})]$$

- **Bayes risk:** Minimum average risk $R_\pi^* = \inf_{\hat{\theta}} R_\pi(\hat{\theta})$
- Limitation: Need to know/assume the prior distribution
  - Worst case Bayes risk: $R_B^* = \sup_\pi R_\pi^*$

Example:

- **MMSE:** Minimum mean square error $R_\pi^* = E[\|\theta - E[\theta|X]\|_2^2]$

Minimax risk

Worst-case risk analysis is based on **minimax risk**

$$R^* = \inf_{\hat{\theta}} \sup_{\theta \in \Theta} R_\theta(\hat{\theta})$$

Theorem (Minimax risk $\geq$ worst-case Bayes risk)

$$R^* \geq R_B^* = \sup_\pi R_\pi^* = \sup_\pi \inf_{\hat{\theta}} R_\pi(\hat{\theta})$$

Proof.
$\forall \hat{\theta}, \pi : \sup_{\theta \in \Theta} R_\theta(\hat{\theta}) \geq E_{\theta \sim \pi}[R_\theta(\hat{\theta})] = R_\pi(\hat{\theta})$, consider $\sup_\pi \inf_{\hat{\theta}}$ $\square$

- key idea also later for lower bounds on minimax risk: Consider Bayes risk with smart prior results in lower bound on R^*.
- result is *weak duality*, minimax theorem is *strong duality*

Variational representation of f-divergence

Legendre-Fenchel transform: Let $f : \mathcal{X} \to \bar{\mathbb{R}}$ be a function (not necessarily convex), then $f^* : \mathcal{X} \to \bar{\mathbb{R}}$ with

$$f^*(a) = \sup_{x \in \mathcal{X}} [\langle a, x \rangle - f(x)]$$

is the conjugate of f (aka Legendre-Fenchel conjugate).

- f^* is convex.
- If f is convex, then $(f^*)^* = f$ (biconjugation)

Similarly, the convex conjugate for any convex functional $\Psi(P)$ defined on the space of measures can be defined as

$$\Psi^*(g) = \sup_{P \in \mathcal{P}} \int g \mathrm{d}P - \Psi(P)$$

Biconjugation holds under certain conditions (e.g. domain of g is finite)

$$\Psi(P) = \sup_g \int g \mathrm{d}P - \Psi^*(P)$$

This can be applied to convex functional $P \mapsto D_f(P\|Q)$ which provides variational representation of f-divergence,[1] where f^* denotes the convex conjugate of f

$$D_f(P\|Q) = E_Q\left[f\left(\frac{P}{Q}\right)\right] = \sup_{g:\mathcal{X}\to\mathrm{dom}(f^*)} E_P\left[g(X)\right] - E_Q\left[f^*(g(X))\right]$$

where g is such that both expectations are finite.

[1] Generalization to infinite domains requires a technical partition argument, for more details see http://people.lids.mit.edu/yp/homepage/data/LN_fdiv.pdf

- Total variation: $f(x) = \frac{1}{2}|x-1|$ with convex conjugate $f^*(y) = \sup_x\{xy - \frac{1}{2}|x-1|\} = \begin{cases} +\infty & \text{if } |y| > \frac{1}{2} \\ y & \text{if } |y| \leq \frac{1}{2} \end{cases}$

$$TV(P,Q) = \sup_{g:|g|\leq\frac{1}{2}} E_P[g(X)] - E_Q[g(X)]$$

- Relative entropy (aka KL divergence), $f(x) = x\log x$ with $f^*(y) = \exp(y-1)$

$$D(P\|Q) = 1 + \sup_{g:\mathcal{X}\to\mathbb{R}} E_P[g(X)] - E_Q[\exp(g(X))]$$

  - Donsker-Varadhan representation (proof see [PW, Sect. 3.3]) $D(P\|Q) = \sup_{g:\mathcal{X}\to\mathbb{R}} E_P[g(X)] - \log E_Q[\exp(g(X))]$, which is stronger since RHS is tighter for any g due to $\log(1+t) \leq t$

- χ^2-divergence, $f(x) = (x-1)^2$ with $f^*(y) = y + \frac{1}{4}y^2$ (HW)

$$\chi^2(P,Q) = \sup_{g:\mathcal{X}\to\mathbb{R}} E_P\left[g(X)\right] - E_Q\left[g(X) + \tfrac{1}{4}g^2(X)\right],$$

  - with substitution $h(x) = \frac{1}{2}g(x) + 1$ we get

$$\chi^2(P,Q) = \sup_{h:\mathcal{X}\to\mathbb{R}} 2E_P\left[h(X)\right] - E_Q\left[h^2(X)\right] - 1,$$

Variational representations provide a systematic analytical approach to obtain lower bounds: $\chi^2(P,Q)$ representation restricted to affine functions $h(x) = ax + b$

$$\chi^2(P,Q) \geq \sup_{a,b\in\mathbb{R}} 2(aE_P\left[X\right] + b) - E_Q\left[(aX+b)^2\right] - 1$$

$$\overset{\text{(HW)}}{=} \frac{(E_P\left[X\right] - E_Q\left[X\right])^2}{\mathrm{Var}_Q[X]} \qquad (1)$$

Hammersley-Chapman-Robbins lower bound

Setup: Data $X \sim P_\theta$, parameter of interest $\theta \in \Theta$, estimator $\hat{\theta}(X)$ (possibly random), cost of prediction error $l(\theta, \hat{\theta}) = (\theta - \hat{\theta})^2$.

- Interested in lower bound on risk $R_\theta(\hat{\theta}) = E_\theta[(\theta - \hat{\theta})^2]$ of estimator $\hat{\theta}$ given the distribution of real parameter θ!

$$E_\theta[(\theta - \hat{\theta})^2] = E_\theta[(\theta - E_\theta[\hat{\theta}] + E_\theta[\hat{\theta}] - \hat{\theta})^2] = \ldots = E_\theta[(bias(\hat{\theta}))^2] + \mathrm{Var}_\theta[\hat{\theta}]$$

Theorem (Hammersley-Chapman-Robbins lower bound)

For the quadratic loss $l(\theta, \hat{\theta}) = (\theta - \hat{\theta})^2$, any estimator $\hat{\theta}(X)$ satisfies

$$R_\theta(\hat{\theta}) \geq \sup_{\theta' \neq \theta} \frac{(E_{\theta'}[\hat{\theta}] - E_\theta[\hat{\theta}])^2}{\chi^2(P_{\theta'}, P_\theta)} \qquad \forall \theta \in \Theta$$

Proof Hammersley-Chapman-Robbins lower bound

Approach: Utilize derived bound (1) on $\chi^2(P, Q)$. Identify distributions P and Q & data processing ineq. In more detail:

- In (1) set $Q = P_\theta$. For P, suppose X was produced by $P_{\theta'}$ with $\theta \neq \theta' \in \Theta$.
- Let $Q_{\hat{\theta}}$ and $P_{\hat{\theta}}$ denote the distributions on $\hat{\theta}$ generated by X distributed according to P_θ and $P_{\theta'}$ respectively.
  - Estimator $\hat{\theta}(X)$ acts a channel that transfers X into $\hat{\theta}$.

$$\chi^2(P_{\theta'}, P_\theta) \overset{\text{data proc.ineq.}}{\geq} \chi^2(P_{\hat{\theta}}, Q_{\hat{\theta}}) \overset{(1)}{\geq} \frac{(E_{\theta'}[\hat{\theta}] - E_\theta[\hat{\theta}])^2}{\mathrm{Var}_\theta[\hat{\theta}]}$$

- Swap LHS with denominator and use $R_\theta(\hat{\theta}) \geq \mathrm{Var}_\theta[\hat{\theta}]$.
- Bound holds for all $\theta' \in \Theta$ and $R_\theta(\hat{\theta})$ does not depend on θ', thus tighten bound by taking $\sup_{\theta' \neq \theta}$ provides desired result.

$\square$

Cramér-Rao lower bound

- Cramér-Rao lower bound can be derived from Hammersley-Chapman-Robbins lower bound
- Restricted to unbiased estimators, i.e., $E_\theta[\hat{\theta}(\theta)] = \theta$.
- Derivation requires regularity conditions to be satisfied

Theorem (Cramér-Rao lower bound)

$$Var_\theta[\hat{\theta}] \geq \frac{1}{I(\theta)}$$

with $I(\theta) = \int \frac{\left(\frac{\mathrm{d}P_\theta(x)}{\mathrm{d}\theta}\right)^2}{P_\theta(x)} \, dx$, *which is the* **Fisher information** *of the parametric family of densities* $\{P_\theta : \theta \in \Theta\}$ *at* θ *(if it exists).*

- Interpretation: The Fisher information is a measure of information in the data that is useful for the estimation task.

Proof Cramér-Rao lower bound

- HCR bound for unbiased estimators and $\theta' \to \theta$ becomes

$$\mathrm{Var}_\theta[\hat{\theta}] \overset{\text{HCR}}{\geq} \sup_{\theta' \neq \theta} \frac{(E_{\theta'}[\hat{\theta}] - E_\theta[\hat{\theta}])^2}{\chi^2(P_{\theta'}, P_\theta)} \geq \lim_{\theta' \to \theta} \frac{(\theta' - \theta)^2}{\chi^2(P_{\theta'}, P_\theta)} \quad \forall \theta \in \Theta.$$

- Taylor series expansion for $P_\theta - P_{\theta'}$ at θ' for θ close to θ':

$$P_\theta - P_{\theta'} = (\theta - \theta') \frac{d(P_\theta - P_{\theta'})}{d\theta} + o((\theta - \theta')^2) = (\theta - \theta') \frac{dP_\theta}{d\theta} + o((\theta - \theta')^2)$$

- With $\chi^2(P_{\theta'}, P_\theta) = \int \frac{(P_\theta - P_{\theta'})^2}{P_\theta} = (\theta' - \theta)^2 \int \frac{(\frac{dP_\theta}{d\theta} + \frac{o((\theta-\theta')^2)}{\theta - \theta'})^2}{P_\theta}$

$$\lim_{\theta' \to \theta} \frac{(\theta' - \theta)^2}{\chi^2(P_{\theta'}, P_\theta)} = \lim_{\theta' \to \theta} \frac{1}{\int \frac{(\frac{dP_\theta}{d\theta} + \frac{o((\theta-\theta')^2)}{\theta - \theta'})^2}{P_\theta}} = \frac{1}{\int \frac{(\frac{dP_\theta}{d\theta})^2}{P_\theta}}$$

□

Fisher information

$$I(\theta) = \int \left(\frac{\frac{\mathrm{d}P_\theta(x)}{\mathrm{d}\theta}}{P_\theta(x)} \right)^2 P_\theta(x)\,\mathrm{d}x = E_\theta\left[\left(\frac{\mathrm{d}\log P_\theta(x)}{\mathrm{d}\theta} \right)^2 \right]$$

- **Regularity condition (HW):** $I(\theta) = -E_\theta\left[\frac{\mathrm{d}^2 \log P_\theta}{\mathrm{d}\theta^2}\right]$ if P_θ is twice differentiable and we have

$$\int \frac{\mathrm{d}^2 P_\theta(x)}{\mathrm{d}\theta^2}\mathrm{d}x = \frac{\mathrm{d}^2}{\mathrm{d}\theta^2}\int P_\theta(x)\mathrm{d}x = 0.$$

- **Multiple samples (HW):** Let $X_1, ..., X_n \sim P_\theta$ iid, then

$$I_n(\theta) = nI(\theta)$$

holds where $I_n(\theta)$ and $I(\theta)$ denote the vector-valued and single-letter Fisher information.

Multivariate HCR/CR lower bounds

Consider multi-dimensional case with $\theta, \theta', \hat{\theta}$ and x defined on $\mathbb{R}^p$

- Multivariate version of HCR lower bound: $\forall \theta, \theta \in \Theta$

$$\chi^2(P_\theta', P_\theta) \geq \left(E_{\theta'}[\hat{\theta}] - E_\theta[\hat{\theta}]\right)^T cov_\theta[\hat{\theta}]^{-1} \left(E_{\theta'}[\hat{\theta}] - E_\theta[\hat{\theta}]\right)$$

  with $cov_\theta[\hat{\theta}] = E_\theta\left[(\hat{\theta} - E_\theta[\hat{\theta}])(\hat{\theta} - E_\theta[\hat{\theta}])^T\right] \in \mathbb{R}^{p \times p}$

- Multivariate CR lower bound
  - considering unbiased estimators $\hat{\theta}$, i.e. $E_\theta[\hat{\theta}] = \theta$

$$cov_\theta[\hat{\theta}] \succeq I(\theta)^{-1}$$

  with Fisher information matrix $I(\theta) = \int \frac{\nabla_\theta P_\theta(x)(\nabla_\theta P_\theta(x))^T}{P_\theta(x)} \mathrm{d}x$

  - $I(\theta) = -E_\theta\left[\frac{\partial^2 \log P_\theta}{\partial \theta_i \partial \theta_j}\right]$ if Hessian satisfies regularity condition

Bayesian Cramér-Rao lower bound

- **Bayesian approach:** Parameter $\theta \in \mathbb{R}$ with prior dist. π
- loss function $l(\theta, \hat{\theta}) = (\theta - \hat{\theta})^2$
- consider unbiased estimators $\hat{\theta}$, i.e. $E_\theta[\hat{\theta}] = \theta$

Theorem (Bayesian Cramér-Rao lower bound)

$$R_\pi^* = \inf_{\hat{\theta}} R_\pi(\hat{\theta}) = \inf_{\hat{\theta}} E_{\theta\sim\pi}[l(\theta, \hat{\theta})] \geq \frac{1}{E_{\theta\sim\pi}[I(\theta)] + I(\pi)}$$

with $I(\pi) = \int \frac{(\mathrm{d}\pi(\theta)/\mathrm{d}\theta)^2}{\pi(\theta)}\mathrm{d}\theta$ *Fisher information of the prior given that suitable regularity conditions hold such as (*)* $\int \frac{\partial^2}{\partial\theta^2}(P_\theta(X)\pi(\theta))\mathrm{d}\theta = \frac{\partial^2}{\partial\theta^2}\int(P_\theta(X)\pi(\theta))\mathrm{d}\theta = 0.$

- Result can be derived with previous arguments deriving first Bayesian HCR with clever choice of distribution in χ^2-term.

Classical proof for Bayesian CR lower bound

- Due to the regularity condition and integration by parts we have $\int(-\theta)\frac{\partial(P_\theta(x)\pi(\theta))}{\partial\theta}\mathrm{d}\theta = \int P_\theta(x)\pi(\theta)\mathrm{d}\theta$ and $\int \hat{\theta}(x)\frac{\partial}{\partial\theta}(P_\theta(x)\pi(\theta))\mathrm{d}\theta = 0$ so that

$$E_{\theta X}\left[(\hat{\theta}(X)-\theta)\frac{\partial \log(P_\theta(X)\pi(\theta))}{\partial\theta}\right]$$
$$= \int\int(\hat{\theta}(x)-\theta)\frac{\partial(P_\theta(x)\pi(\theta))}{\partial\theta}\frac{P_\theta(x)\pi(\theta)}{P_\theta(x)\pi(\theta)}\mathrm{d}\theta\mathrm{d}x = 1$$

- Using Cauchy-Schwarz inequality on (LHS)^2 and rearrange

$$1 = \left(E_{\theta X}\left[(\hat{\theta}(X)-\theta)\frac{\partial \log(P_\theta(X)\pi(\theta))}{\partial\theta}\right]\right)^2$$
$$\leq \underbrace{E_{\theta X}\left[(\hat{\theta}(X)-\theta)^2\right]}_{=R_\pi(\hat{\theta})} \quad \underbrace{E_{\theta X}\left[\left(\frac{\partial \log(P_\theta(X)\pi(\theta))}{\partial\theta}\right)^2\right]}_{\overset{(*)}{=}-E_{\theta X}\left[\frac{\partial^2}{\partial\theta^2}\log(P_\theta(X)\pi(\theta))\right]=E_\theta[I(\theta)]+I(\pi)} \qquad \square$$