Lecture 3: f-divergence, joint range, and statistical decision theory

Lecturer: Yanjun Han

April 5, 2021

Announcement

Reminder:

- students enrolled in letter grade are required to scribe some notes
- first lecture is lecture 6, next Wednesday (4/14)
- scribe group size: $1 - 3$
- template on course website
- sign-up link: `https://bit.ly/31quUIb`

Today's plan

Fundamentals of statistical closeness and statistical model

- total variation (TV) distance, Le Cam's first lemma
- f-divergences, sharp inequalities and joint range
- general setting of statistical decision theory

Total variation (TV) distance

Definition (TV distance)

For two probability distributions P and Q defined on the same set $\mathcal{X}$, the TV distance is defined to be

$$\|P - Q\|_{\mathrm{TV}} = \frac{1}{2} \sum_{x \in \mathcal{X}} |p(x) - q(x)| \quad \in [0, 1]$$

continuous: p, q are densities $\|p - q\|_{TV} = \frac{1}{2} \int |p(x) - q(x)| \, dx$

in general: $\|p - q\|_{TV} = \frac{1}{2} \int |dP - dQ|$

ν s.t. $P, Q \ll \nu$ $(\nu = \frac{P+Q}{2})$ $\quad \frac{1}{2} \int \left| \frac{dP}{d\nu} - \frac{dQ}{d\nu} \right| d\nu$

Equivalent formulations:

$$\|P - Q\|_{\mathrm{TV}} = \sup_{A \subseteq \mathcal{X}} |P(A) - Q(A)| = \sup_{0 \le f \le 1} |\mathbb{E}_P[f] - \mathbb{E}_Q[f]|$$

$$= 1 - \sum_{x \in \mathcal{X}} \min\{p(x), q(x)\} = \sum_{x \in \mathcal{X}} \max\{p(x), q(x)\} - 1.$$

Examples of TV computation

Example I: $\mathcal{X} = \{0, 1\}, P = \text{Bern}(p), Q = \text{Bern}(q)$

$$\|P - Q\|_{TV} = |p - q|$$

Example II: $\mathcal{X} = \mathbb{R}, P = \mathcal{N}(\mu_1, \sigma^2), Q = \mathcal{N}(\mu_2, \sigma^2)$

$$\|P - Q\|_{TV} = P(A) - Q(A), \quad A = \left\{x : \frac{p(x)}{q(x)} \geq 1\right\}$$

$$= 2\Phi\left(\left|\frac{\mu_1 - \mu_2}{2\sigma}\right|\right) - 1 \qquad A = \left\{x : x \geq \frac{\mu_1 + \mu_2}{2}\right\}.$$

Φ is CDF of $N(0,1)$.

Why TV distance?

Theorem (Le Cam's first lemma)

Let X follow either P or Q, and $\Psi(X) \in \{0, 1\}$ be a test. Then

$$\min_{\Psi}[P(\Psi(X) = 1) + Q(\Psi(X) = 0)] = 1 - \|P - Q\|_{\mathrm{TV}}$$

$H_0: X \sim P$
$H_1: X \sim Q$

$\psi(X) \in \{0,1\}$ → $\psi(x)=0 \Leftarrow H_0$ holds; $\psi(x)=1 \Leftarrow H_1$ holds

Type I error $= \mathbb{P}(\psi(x)=1 \mid H_0) = P(\psi(x)=1)$

Type II error $= \mathbb{P}(\psi(x)=0 \mid H_1) = Q(\psi(x)=0)$

Proof:

$\mathcal{X} = A \cup A^c$

if $x \in A$ $\psi(x) = 0$

$x \in A^c$ $\psi(x) = 1$

$P(\psi(X)=1) + Q(\psi(X)=0)$

$= P(A^c) + Q(A)$

$= 1 - (P(A) - Q(A))$

max. $= \|P-Q\|_{TV}$

General case: f-divergence

Definition (f-divergence)

Let $f : \mathbb{R}_+ \to \mathbb{R}$ be a convex function with $f(1) = 0$. The f-divergence is then defined as

$$D_f(P, Q) = \sum_{x \in \mathcal{X}} q(x) f\left(\frac{p(x)}{q(x)}\right)$$

$D_f(P,Q) \neq D_f(Q,P)$

$D_f(P,Q) = 0$ if $P = Q$

- $D_f(P, Q) \geq 0$: $D_f(P,Q) = \sum_x q(x) f\left(\frac{p(x)}{q(x)}\right) \geq f\left(\sum_x q(x) \cdot \frac{p(x)}{q(x)}\right) = f(1) = 0.$
- A general version of test error:

$$f(x) = a + bx + \frac{1}{2}\int |x - t| f''(dt). \quad \text{if } f \text{ convex}$$

$|x-t|$: generalization of TV when applied to $\frac{p(x)}{q(x)}$.

- Comparison with other divergences:
  - Bregman divergence: $D(P,Q) = F(P) - F(Q) - \langle \nabla F(Q), P - Q \rangle$, convex F.

    $\{\text{Bregman}\} \cap \{f\text{-}\} = \{\text{KL}\}.$
  - Integral probability metric (IPM):

$$D_{\mathcal{F}}(P,Q) = \sup_{f \in \mathcal{F}} |\mathbb{E}_P f - \mathbb{E}_Q f|$$

Data-processing inequality

Theorem (Data-processing inequality)

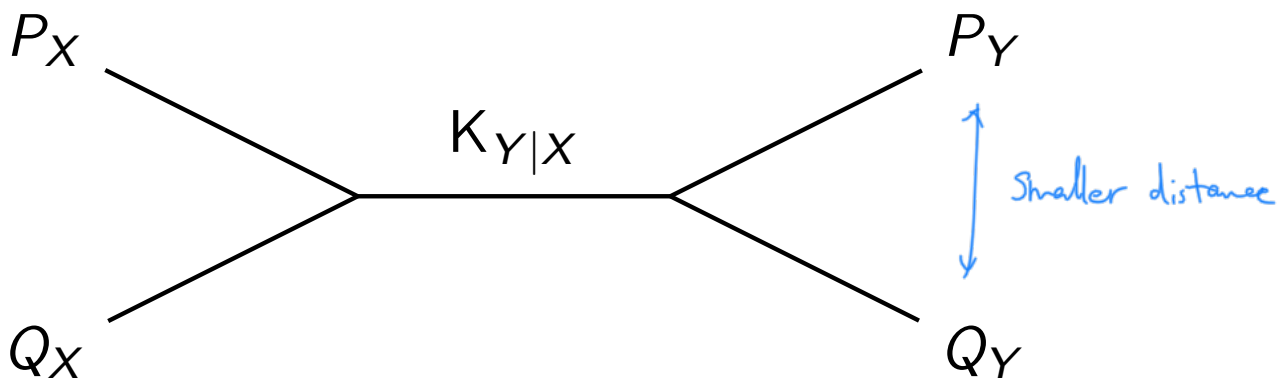
Let K be a transition kernel. Then

$$D_f(KP, KQ) \leq D_f(P, Q).$$

$D_f(KP, KQ) \leq c(K) \cdot D_f(P, Q)$

$c(K) \leq 1$.

Diagram:



Proof. $p(y) = \sum_x p(x,y) = \sum_x p(x) K(y|x)$.

$$D_f(P_X, Q_X) = \sum_x q(x) f\left(\frac{p(x)}{q(x)}\right) = \sum_x q(x) f\left(\frac{p(x,y)}{q(x,y)}\right) \quad \text{(any } y\text{)} = \sum_x \sum_y q(x,y) f\left(\frac{p(x,y)}{q(x,y)}\right)$$

$$= \sum_y q(y) \sum_x q(x|y) f\left(\frac{p(x,y)}{q(x,y)}\right) \geq \sum_y q(y) f\left(\sum_x q(x|y) \cdot \frac{p(x,y)}{q(x,y)}\right)$$

$$= \sum_y q(y) f\left(\sum_x \frac{p(x,y)}{q(y)}\right) = \sum_y q(y) f\left(\frac{p(y)}{q(y)}\right) = D_f(P_Y, Q_Y).$$

Widely-used examples of f-divergence

- TV distance: $f(x)=\frac{1}{2}|x-1|$

$$\|P-Q\|_{\mathrm{TV}}=\frac{1}{2}\sum_{x\in\mathcal{X}}|p(x)-q(x)|$$

$\|P^{\otimes n}-Q^{\otimes n}\|_{TV}\le n\cdot\|P-Q\|_{TV}$

typical: $\|P^{\otimes n}-Q^{\otimes n}\|_{TV}\le\sqrt{n}\cdot\|P-Q\|_{TV}$

- squared Hellinger distance: $f(x)=(\sqrt{x}-1)^2$

$$H^2(P,Q)=\sum_{x\in\mathcal{X}}\left(\sqrt{p(x)}-\sqrt{q(x)}\right)^2$$

$\|P-Q\|_{TV}\asymp O(\sqrt{D_{KL}(P\|Q)})$

$D_{KL}(P^{\otimes n}\|Q^{\otimes n})=nD_{KL}(P\|Q)$

$\Downarrow$

$\|P^{\otimes n}-Q^{\otimes n}\|_{TV}\le\sqrt{\frac{1}{2}D_{KL}(P^{\otimes n}\|Q^{\otimes n})}$

$=\sqrt{\frac{n}{2}D_{KL}(P\|Q)}$

$\approx\sqrt{n}\cdot\|P-Q\|_{TV}$.

- Kullback–Leibler (KL) divergence: $f(x)=x\log x$

$$D_{\mathrm{KL}}(P\|Q)=\sum_{x\in\mathcal{X}}p(x)\log\frac{p(x)}{q(x)}$$

- chi-squared divergence: $f(x)=(x-1)^2$

$$\chi^2(P,Q)=\sum_{x\in\mathcal{X}}\frac{(p(x)-q(x))^2}{q(x)}=\sum_{x\in\mathcal{X}}\frac{p(x)^2}{q(x)}-1$$

Hellinger distance

Advantages:

- $0 \leq H^2(P, Q) \leq 2$
- compatible with TV:

$$H^2(P, Q) \to 0 \iff \|P - Q\|_{\mathrm{TV}} \to 0$$
$$H^2(P, Q) \to 2 \iff \|P - Q\|_{\mathrm{TV}} \to 1$$

- tensorization property:

$$H^2\left(\prod_i P_i, \prod_i Q_i\right) = 2 - 2\prod_i \left(1 - \frac{H^2(P_i, Q_i)}{2}\right)$$

$1 - \frac{H^2(P,Q)}{2} = \sum_x \sqrt{p(x)q(x)}$

$1 - \frac{H^2(P_X \times P_Y, Q_X \times Q_Y)}{2} = \sum_x \sum_y \sqrt{p(x)p(y) \cdot q(x)q(y)} = \left(\sum_x \sqrt{p(x)q(x)}\right)\left(\sum_y \sqrt{p(y)q(y)}\right)$

$= \left(1 - \frac{H^2(P_X, Q_X)}{2}\right)\left(1 - \frac{H^2(P_Y, Q_Y)}{2}\right)$.

KL divergence

Advantages:

- tensorization property:

$$D_{\mathrm{KL}}\left(\prod_i P_i \| \prod_i Q_i\right) = \sum_i D_{\mathrm{KL}}(P_i \| Q_i)$$

- chain rule even for correlated random variables (HW1):

$$D_{\mathrm{KL}}(P_{XY} \| Q_{XY}) = D_{\mathrm{KL}}(P_X \| Q_X) + D_{\mathrm{KL}}(P_{Y|X} \| Q_{Y|X} \mid P_X)$$

$= \sum_x P_X(x) \cdot D_{\mathrm{KL}}(P_{Y|X=x} \| Q_{Y|X=x})$

- vast applications in many other fields, e.g. relationships to mutual information, and Donsker-Varadhan

$$D_{\mathrm{KL}}(P \| Q) = \sup_f \mathbb{E}_P[f] - \log \mathbb{E}_Q[e^f]$$

χ^2 divergence

Advantages:

- nice behavior under mixture distribution (HW1):

$$\chi^2(\mathbb{E}_\theta[P_\theta], Q) = \mathbb{E}_{\theta,\theta'}\left[\sum_{x\in\mathcal{X}} \frac{p_\theta(x)p_{\theta'}(x)}{q(x)}\right] - 1.$$

(e.g. $\frac{N(0,1)+N(1,1)}{2}$)

θ' an ind. copy of θ

- variational representation:

$$\chi^2(P, Q) = \sup_h \frac{(\mathbb{E}_P[h] - \mathbb{E}_Q[h])^2}{\mathrm{Var}_Q(h)}$$

- tensorization property:

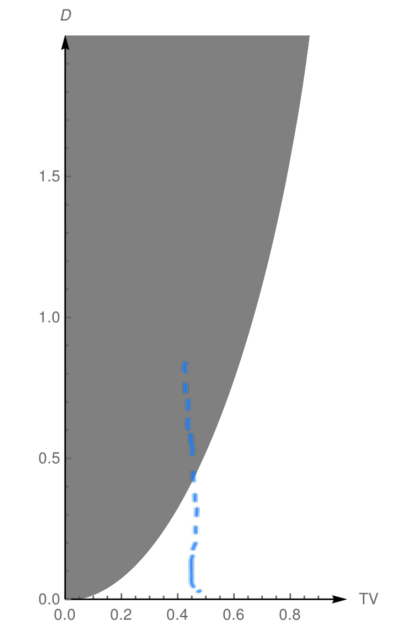
$$\chi^2\left(\prod_i P_i, \prod_i Q_i\right) = \prod_i (1 + \chi^2(P_i, Q_i)) - 1.$$

Joint range and inequalities

Question: given two f-divergences $D_f(P, Q)$ and $D_g(P, Q)$, what is the tight inequality between them?

Definition (Joint range)

$$\mathcal{R} = \{(D_f(P, Q), D_g(P, Q)) \in \mathbb{R}_+^2 : P, Q \text{ probability measures}\}$$



Joint range of (TV, KL)

Characterization of joint range

Notations:

- $\mathcal{R} = \{(D_f(P,Q), D_g(P,Q)) : P, Q \text{ probability measures}\}$
- $\mathcal{R}_k = \{(D_f(P,Q), D_g(P,Q)) : P, Q \text{ probability measures on } [k]\}$

$\mathcal{R}_k \subseteq \mathcal{R}$.

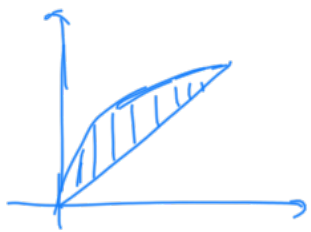
Theorem (Harremoës and Vajda, 2011)

$$\mathcal{R} = \text{conv}(\mathcal{R}_2) = \mathcal{R}_4.$$

convex hull

Implication:

- choose $P = (p, 1-p), Q = (q, 1-q)$
- vary $(p, q) \in [0,1]^2$, plot the joint range
- take the convex hull

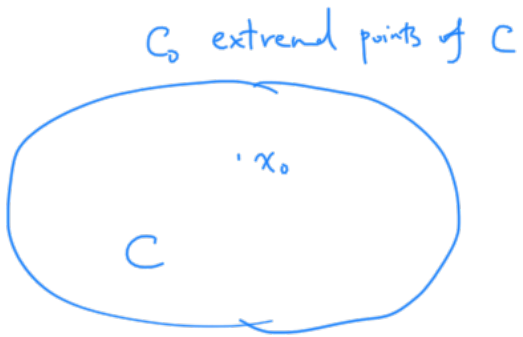


Proof technique: convex geometry

Theorem (Choquet, Bishop, and de Leeuw)

Let C be a metrizable convex compact subset of a locally convex topological vector space V. For any point $x_0 \in C$, there exists a probability measure μ supported on extremal points of C such that

$$x_0 = \int x \mu(dx).$$



$x \in C$ extremal point
$\Leftrightarrow$ x cannot be written as a convex combination of $y, z \in C$ with $y \neq x$ $z \neq x$.

Proof of $\mathcal{R} = \text{conv}(\mathcal{R}_2)$

Equivalent representation:

- $\mathcal{R} = \{(\mathbb{E}[f(X)], \mathbb{E}[g(X)]) : \mathbb{E}[X] \leq 1\}$
- $\mathcal{R}_k = \{(\mathbb{E}[f(X)], \mathbb{E}[g(X)]) : \mathbb{E}[X] \leq 1, X \text{ supported on } k \text{ points}\}$ (at most)

$X = \frac{p(x)}{q(x)}$ w.p. $q(x)$ $\quad \mathbb{E}[f(X)] = \sum_x q(x) f\left(\frac{p(x)}{q(x)}\right) = D_f(P, Q)$

$\mathbb{E}[X] = \sum_x q(x) \cdot \frac{p(x)}{q(x)} = 1$

Extremal points:

$C = \{$ probability distributions of X satisfying above constraints $\}$.

extremal points of $C = \begin{cases} \text{singleton}, \\ \text{two-point/binary distribution w. } \mathbb{E}[X]=1. \end{cases}$

$\mathbb{E}[f(X)]$ linear in distribution of X.

Proof of $\text{conv}(\mathcal{R}_2) = \mathcal{R}_4$

Theorem (Carathéodory)

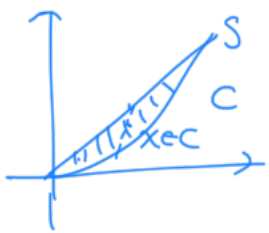
Let $S \subseteq \mathbb{R}^d$ and $C = \text{conv}(S)$. Then any $x \in C$ could be written as $x = \text{conv}(x_1, \cdots, x_{d+1})$ for some $x_1, \cdots, x_{d+1} \in S$.

Furthermore, if S is connected, then $d + 1$ could be replaced by d.

$d=2$

2-mixture of binary distributions

$\Longrightarrow$ distribution on 4 points



Some widely-used inequalities

- TV vs. Hellinger: tight

$$\frac{1}{2}H^2(P,Q) \le \|P-Q\|_{\mathrm{TV}} \le H(P,Q)\sqrt{1-\frac{H^2(P,Q)}{4}}$$

- TV vs. KL: not tight

$$\|P-Q\|_{\mathrm{TV}} \le \sqrt{\frac{1}{2}D_{\mathrm{KL}}(P\|Q)} \Leftarrow \text{Pinsker's inequality}$$

$$\|P-Q\|_{\mathrm{TV}} \le \sqrt{1-\exp(-D_{\mathrm{KL}}(P\|Q))}$$

- KL vs. χ^2: tight

$$D_{\mathrm{KL}}(P\|Q) \le \log(1+\chi^2(P,Q))$$

Statistical decision theory

$$\theta \in \Theta \xrightarrow{X \sim P_\theta} X \in \mathcal{X} \xrightarrow{a \sim \delta(\cdot \mid X)} a \in \mathcal{A}$$

- Θ: parameter space
- $\mathcal{X}$: observation space
- $\mathcal{A}$: action space
- loss function $L : \Theta \times \mathcal{A} \to \mathbb{R}_+$

Definition (Risk)

The risk of the decision rule δ under loss function L and the true parameter θ is defined as

$$R_{\theta,L}(\delta) = \mathbb{E}_{a \sim \delta(\cdot|X)} \mathbb{E}_{X \sim P_\theta}[L(\theta, a)]$$

Risk comparison: Bayes and minimax

For two decision rules δ_1, δ_2, typically $R_{\theta,L}(\delta_1) \geq R_{\theta,L}(\delta_2)$ for some θ, and $R_{\theta,L}(\delta_1) \leq R_{\theta,L}(\delta_2)$ for other θ

- minimax criterion: compare $\max_{\theta \in \Theta} R_{\theta,L}(\delta)$
- Bayes criterion: fix a prior π on θ, compare $\mathbb{E}_{\theta \sim \pi}[R_{\theta,L}(\delta)]$

Exercise

Bayes estimator is easy to find in principle:

$$T(x) \in \arg\min_{a \in \mathcal{A}} \mathbb{E}_{\theta \sim \pi(\cdot|x)}[L(\theta, a)]$$

Example I: linear regression

$$\theta \in \mathbb{R}^p \longrightarrow (x_1, y_1), \cdots, (x_n, y_n) \longrightarrow \widehat{\theta} \in \mathbb{R}^p$$
$$x_1, \cdots, x_n \sim P_X$$
$$y_i \mid x_i \sim \mathcal{N}(x_i^\top \theta, 1)$$

- estimation error: $L_1(\theta, \widehat{\theta}) = \|\theta - \widehat{\theta}\|_2^2$
- prediction error: $L_2(\theta, \widehat{\theta}) = \mathbb{E}_{(x,y)\sim P_\theta}[(y - x^\top \widehat{\theta})^2]$

Example II: density estimation

$$f \in \mathcal{F} \longrightarrow x_1, \cdots, x_n \sim f \longrightarrow T$$

- loss at a point: $L_1(f, T) = |T - f(0)|$
- global loss: $L_2(f, T) = \int |T(x) - f(x)| dx$
- functional estimation: $L_3(f, T) = |T - \|f\|_2|$

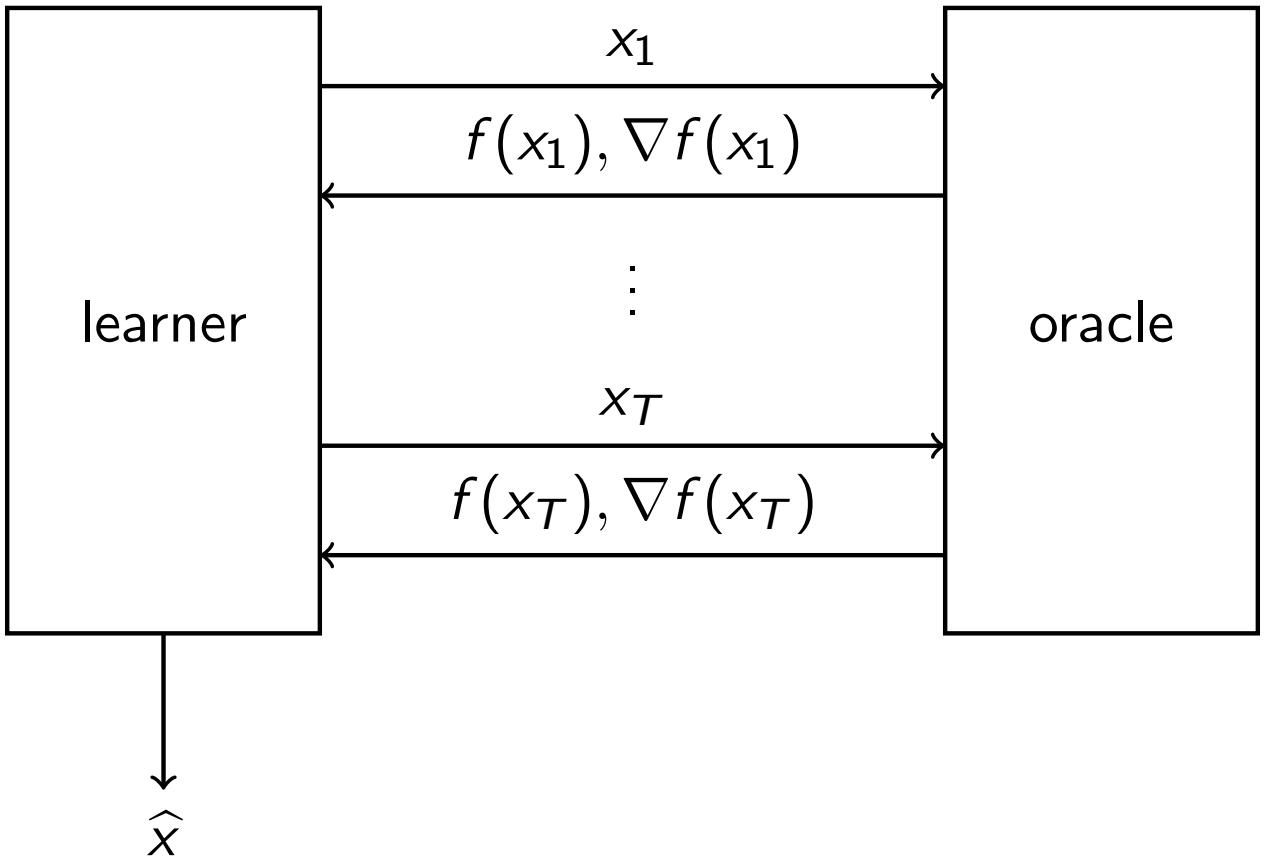
Example III: learning theory

$$P_{XY} \longrightarrow (x_1, y_1), \cdots, (x_n, y_n) \sim P_{XY} \longrightarrow f \in \mathcal{F}$$

- excess risk:

$$L(P_{XY}, f) = \mathbb{E}_{(x,y)\sim P_{XY}}[L_0(f(x), y)] - \min_{f^\star \in \mathcal{F}} \mathbb{E}_{(x,y)\sim P_{XY}}[L_0(f^\star(x), y)]$$

Example IV: optimization



- suboptimality gap:

$$L(f, \widehat{x}) = f(\widehat{x}) - \min_{x^\star} f(x^\star)$$

References

- Alexandre Tsybakov. "Introduction to nonparametric estimation." Springer-Verlag, 2009. (Chapter 2)
- Peter Harremoës and Igor Vajda. "On pairs of f-divergences and their joint range." *IEEE Transactions on Information Theory* 57.6 (2011): 3230–3235.
- Adityanand Guntuboyina, Sujayam Saha, and Geoffrey Schiebinger. "Sharp inequalities for f-divergences." *IEEE transactions on information theory* 60, no. 1 (2013): 104–121.

Next lecture: reduction between statistical models