Approximation and Uncertainty in Parameter Estimation

R. Kulhavý and F. Hrnčíř

Institute of Information Theory and Automation
Academy of Sciences of the Czech Republic

Outline

- Problem Statement
- Theoretical Optimum
- Approximate Scheme
- Key Properties
- Asymptotic Behaviour
- Illustrative Example
- Summary of Main Points
- Prospective Directions

Problem Statement

sequence of random variables

$$\boldsymbol{X} = (X_1, \ldots, X_k), \quad X_j \in \mathfrak{X} = \{1, 2, \ldots, |\mathfrak{X}|\}$$

independent and identically distributed

$$S_\theta(x), \quad \theta \in \mathfrak{T} = \{1, 2, \ldots, |\mathfrak{T}|\}$$

posterior distribution of $\boldsymbol{\Theta}$

$$P_{\boldsymbol{x}}(\theta) \propto P(\theta) \prod_{j=1}^{k} S_\theta(x_j)$$

Theoretical Optimum

empirical distribution

$$R_{\boldsymbol{x}}(a) \;=\; \frac{1}{k}\, N_{\boldsymbol{x}}(a)$$

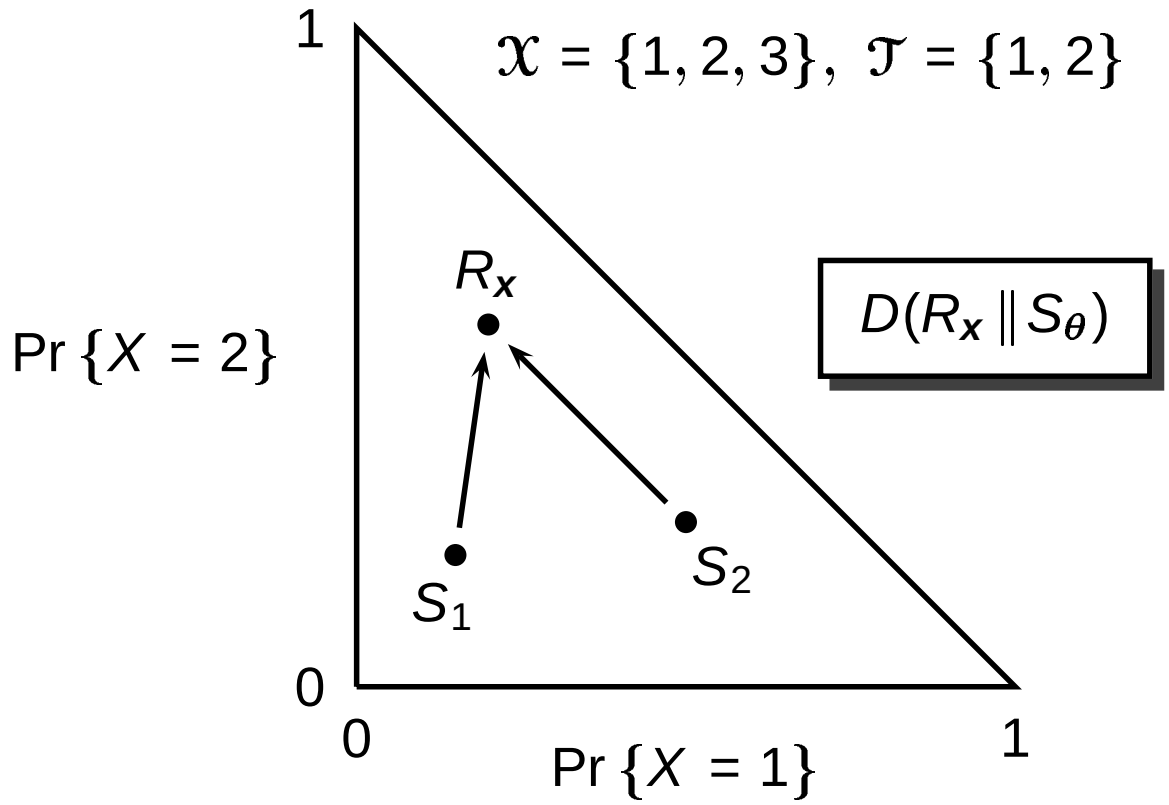
relative entropy

$$D(R \,\|\, S) \;=\; \sum_{a \in \mathcal{X}} R(a) \;\log\; \frac{R(a)}{S(a)}$$

posterior distribution

$$P_{\boldsymbol{x}}(\theta) \;\propto\; P(\theta) \;\exp\; \{-k\, D(R_{\boldsymbol{x}} \,\|\, S_{\theta})\}$$

Probability Simplex



Approximate Scheme

data statistic

$$T(\boldsymbol{X}) \;=\; \frac{1}{k}\sum_{j=1}^{k} h(X_j) \;=\; E_{R_{\boldsymbol{X}}}\, h(X) \;\geq\; \xi$$

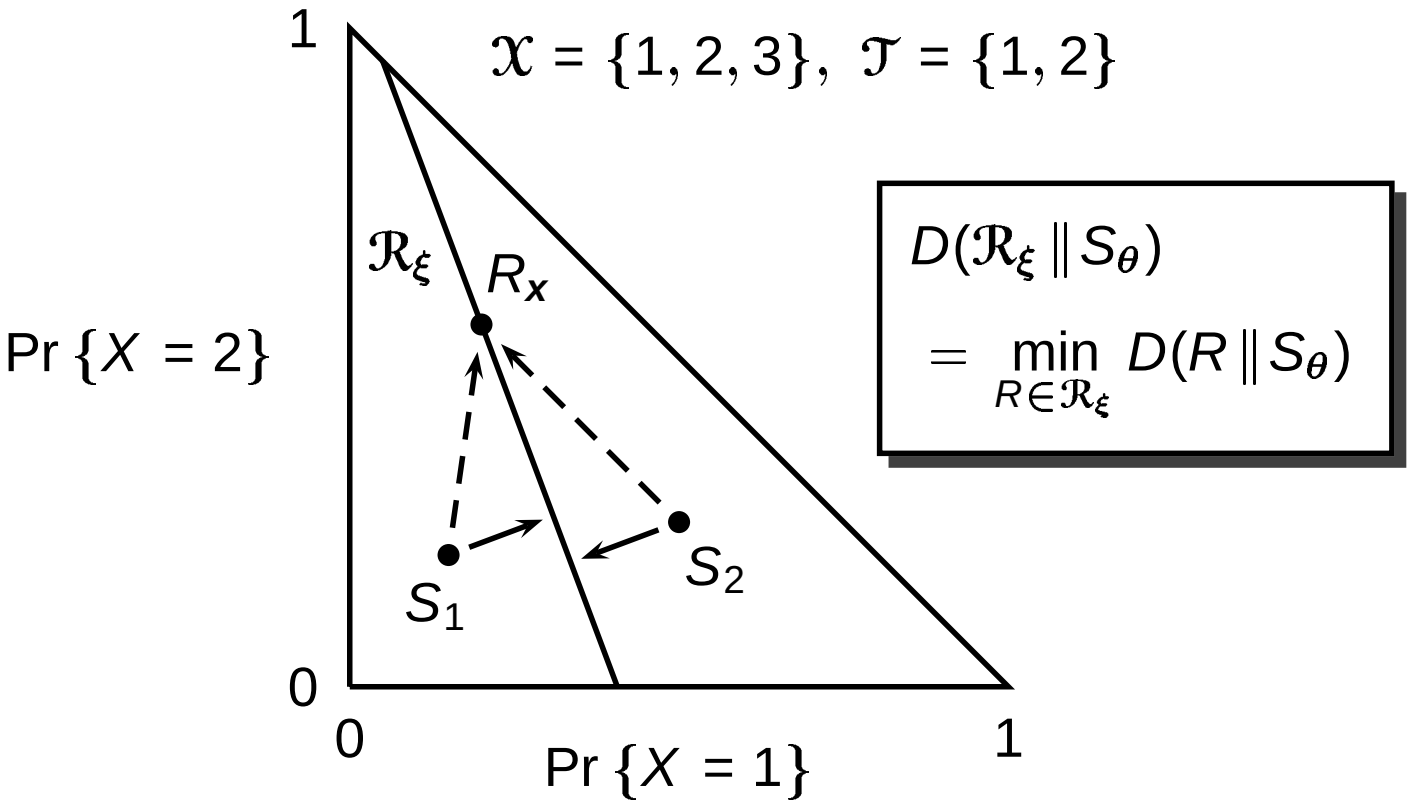
data compression makes $R_{\boldsymbol{X}}$ uncertain!

$$R_{\boldsymbol{X}} \;\in\; \mathcal{R}_{\xi} \;=\; \{R \in \mathcal{R} : E_R\, h(X) \geq \xi\}$$

approximate posterior

$$\hat{P}_{\xi}(\theta) \;\propto\; P(\theta)\;\exp\left\{-k \min_{R \in \mathcal{R}_{\xi}} D(R \,\|\, S_{\theta})\right\}$$

Probability Simplex II.



Key Properties

meaningful bounds

$$0 \leq D(\mathcal{R}_\xi \| S_\theta) \leq D(R_x \| S_\theta)$$

monotonicity

$$\mathcal{R}_\xi \subset \mathcal{R}'_\xi \Rightarrow D(\mathcal{R}'_\xi \| S_\theta) \leq D(\mathcal{R}_\xi \| S_\theta)$$

extreme cases

$$\mathcal{R}_\xi = \{R_x\} \Rightarrow D(\mathcal{R}_\xi \| S_\theta) = D(R_x \| S_\theta) \Rightarrow \hat{P}_\xi(\theta) = P_x(\theta)$$

$$\mathcal{R}_\xi = \mathcal{R} \Rightarrow D(\mathcal{R}_\xi \| S_\theta) = 0 \Rightarrow \hat{P}_\xi(\theta) = P(\theta)$$

Asymptotic Behaviour

point-estimation consistency

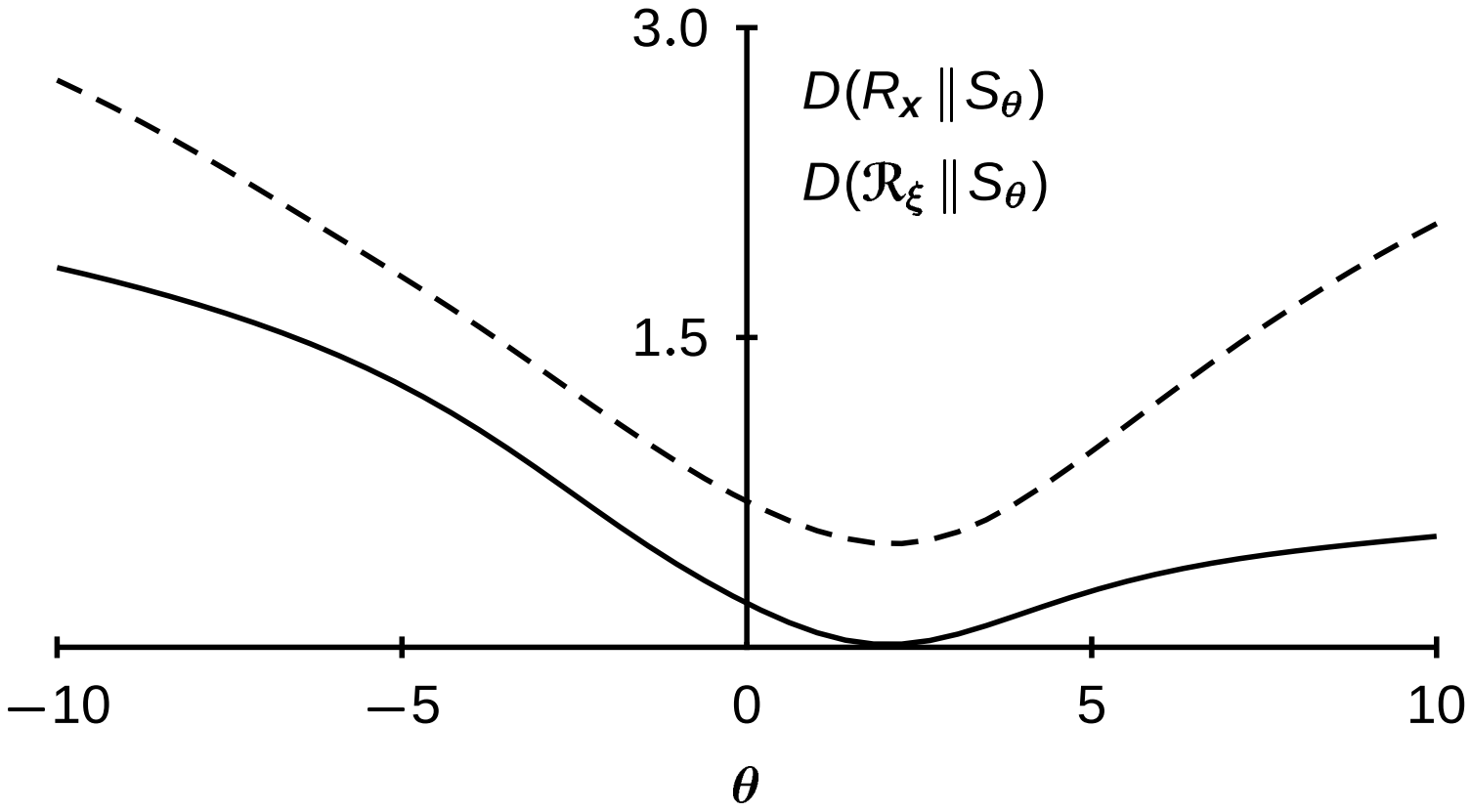
$$R_{\boldsymbol{x}} \to S_{\theta_0} \;\Rightarrow\; D(\mathcal{R}_{\xi} \,\|\, S_{\theta_0}) \to 0$$

$$P(\theta_0) > 0 \;\Rightarrow\; \lim_{k\to\infty} \hat{P}_{\xi}(\theta_0) > 0$$

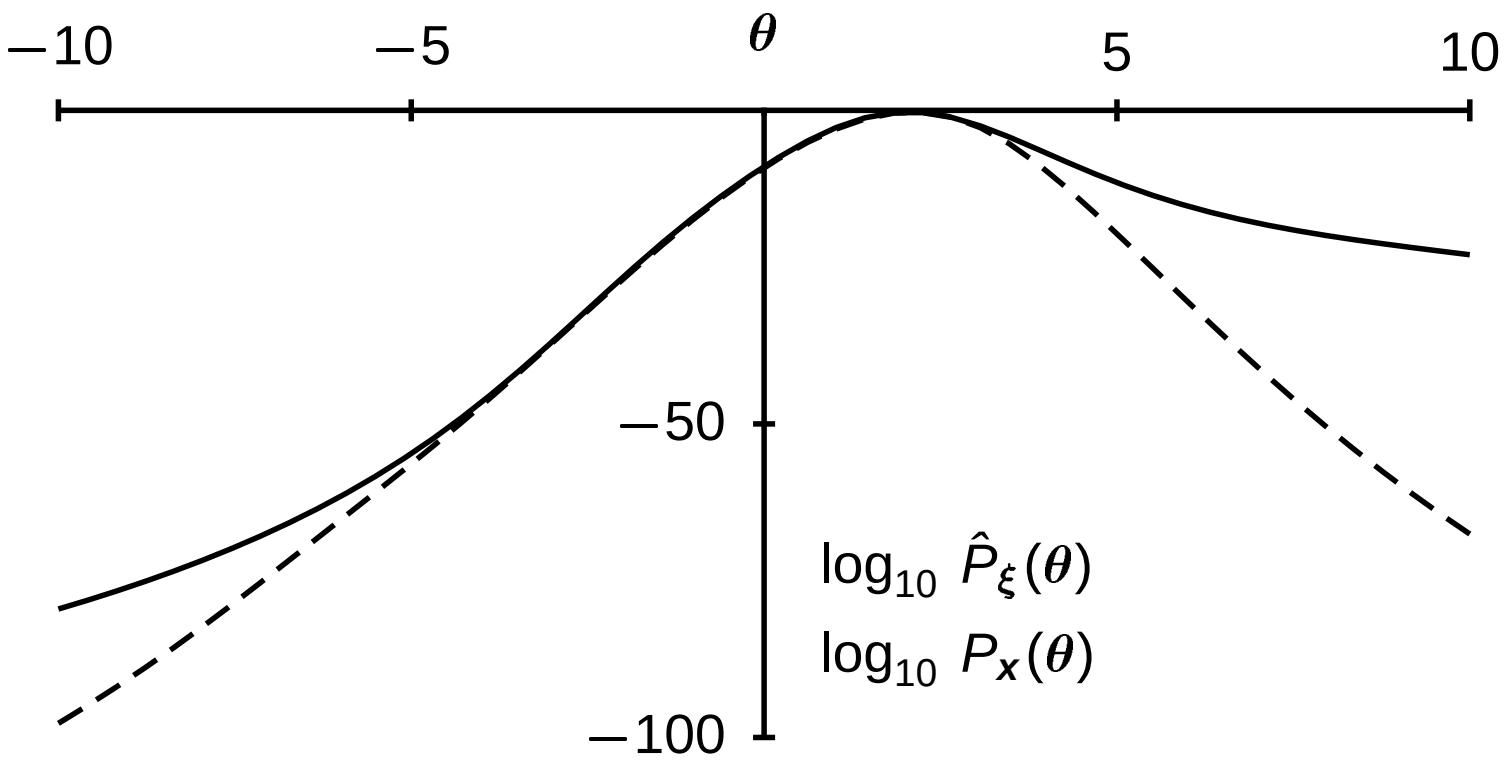
large deviation property (*Sanov's theorem*)

$$\lim_{k\to\infty} \frac{1}{k} \log \Pr\left\{\boldsymbol{x} \in \mathcal{X}^k : R_{\boldsymbol{x}} \in \mathcal{C}\right\} = -\min_{R\in\mathcal{C}} D(R \,\|\, S)$$

Example: Relative Entropy



Example: Posterior Distribution



Summary of Main Points

- the scheme beats largely the curse of dimensionality

  it copes effectively with compressed data

  *it **might** even cope with complexity of* $D(\mathcal{R}_\xi \,\|\, S_\theta)$

- the idea works in most cases of interest

  i.i.d. data, Markov chains, $\mathcal{T} \subset \mathbb{R}^M$, $\mathcal{X} \subset \mathbb{R}^N$

- the scheme is a **shift** of the underlying paradigm

  the answer is $D(\mathcal{R}_\xi \,\|\, S_\theta)$ *rather than* $P_{\boldsymbol{x}}(\theta)$

Prospective Directions

- analysis of mismodelling effects

  $P_\xi(\theta)$ *measures the* ***relative*** *degree of belief*

  $D(\mathcal{R}_\xi \,\|\, S_\theta)$ *measures the* ***actual distance***

- evaluation of uncertainty of point estimates

  $D(\mathcal{R}_\xi \,\|\, S_\theta)$ *admits any model class* $\{S_\theta\}$

- estimation with missing data

  $\mathcal{R}_\xi$ *becomes correspondingly larger*

Good vs Bad Models