Minimax Estimation of KL Divergence between Discrete Distributions
Yanjun Han, Jiantao Jiao and Tsachy Weissman
Information Systems Laboratory, Department of Electrical Engineering, Stanford University
Email: {yjhan, jiantao, tsachy}@stanford.edu

Problem formulation

Problem: given $m$ and $n$ i.i.d. samples from unknown distributions $P$ and $Q$, respectively, how to estimate the following Kullback–Leibler distance

$$D(P) = \sum_{i=1}^{n} p_i \ln \frac{p_i}{q_i}$$

where $S$ is the support size.

Fact: $P \approx Q$ only guarantees the finiteness of $D(P)Q$ but no estimator works in this case. Hence, the likelihood ratio is assumed to be bounded from above, i.e., the uncertainty set is

$$U_{e,k}(x) = \{(P, Q) : P, Q \in M, \frac{1}{n} \leq u(S), 1 \leq i \leq S\}$$

where $M$ denotes the set of discrete distributions with support size at most $S$.

Main results [1]

Optimal estimator: the minimax $L_2$ risk for estimating $D(P)Q$ over $(P, Q) \in U_{e,k}$ satisfies if $u(S) \geq \ln n^2$,

$$\inf_{\hat{D}} \sup_{(P, Q) \in U_{e,k}} \mathbb{E}(\hat{D} - D(P)Q)^2 \geq \left(\frac{s}{m} + \frac{\ln(S)}{m} \right)^2$$

It is necessary and sufficient to have $m \gg \frac{1}{S}$ and $n \gg \frac{1}{S^2}$ samples for a consistent estimation.

Plug-in approach: the performance of the plug-in approach is

$$\sup_{(P, Q) \in U_{e,k}} \mathbb{E}(\hat{D} - D(P)Q)^2 \leq \left(\frac{S}{m} + \frac{\ln(S)}{m} \right)^2$$

The plug-in approach requires $m \gg S$ and $n \gg S$ samples to be consistent.

Effective sample size enlargement: the performance of the optimal estimator with $(m, n)$ samples is essentially that of the plug-in approach with $(m \ln m, n \ln n)$ samples.

General recipe for constructing the optimal estimator

Consider the estimation of $G(\theta), \theta \in \Theta$, and an estimator $\hat{\theta}$ in $\Theta$. Denote by $\Theta'$ the set of all non-analytic points of $G(\cdot)$ in $\Theta'$. Localization: consider a statistical model $(P_{\theta}, Q_{\theta})$ and an estimator $\hat{\theta}$ in $\Theta'$. A localization of level $r \in [0, 1]$, or an $r$-localization, is a collection of sets $U(x) \subseteq \Theta$ for any $x \in \Theta'$, and $U_{\theta}(x) = U(\theta) \cap \Theta$. The complexity of estimating $R_{\text{entropy}} \leq \ln n/m$, and then essentially employ the plug-in approach $G(\hat{\theta})$ to estimate $G(\theta)$. We are in the “smooth” regime if $U(\hat{\theta}) \subset \Theta_{\theta}$, and then essentially employ the plug-in approach $G(\hat{\theta})$ to estimate $G(\theta)$. We are in the “non-smooth” regime if $U(\hat{\theta}) \subset \Theta_{\theta}$, and then use an unbiased estimator of another plug-in approach $G(\hat{\theta})$ which approximates $G(\theta)$ well on $U(\hat{\theta})$.

Estimator of $\rho \ln q$ by general recipe

Estimator construction:

- unbiased estimate of best polynomial approximation of order $n$

$$f(\hat{p}) = \frac{\ln(\frac{2\sqrt{n}}{1 + \sqrt{1 - \frac{1}{n}}})}{2}$$

- “smooth” regime:

$$\hat{p} \approx \hat{p}_1$$

- “non-smooth” regime I:

$$\hat{p} \approx \hat{p}_2$$

- “non-smooth” regime II:

$$\hat{p} \approx \hat{p}_3$$

References