Kullback-Leibler Divergence Estimation of Continuous Distributions

Fernando Pérez-Cruz

Princeton University

University Carlos III in Madrid

Workshop on Representations and Inference on Probability Distributions

8th December 2007

Overview

- Problem Set Up.
- One Dimensional Solution.
- Multi-Dimensional Solution.
- Some Examples.

Problem Set up

- The Kullback-Leibler divergence is given by:

$$D(P||Q) = \int_{\mathbb{R}^d} p(\mathbf{x}) \log \frac{p(\mathbf{x})}{q(\mathbf{x})} d\mathbf{x} \geq 0.$$

  This divergence is finite whenever P is absolutely continuous with respect to Q and it is only zero if $P = Q$.

- We want to estimate $D(P||Q)$ given:
  - n i.i.d. samples from $p(\mathbf{x})$: $\mathcal{X} = \{\mathbf{x}_i\}_{i=1}^n$.
  - m i.i.d. samples from $q(\mathbf{x})$: $\mathcal{X}' = \{\mathbf{x}'_j\}_{j=1}^m$.

- and we want to avoid estimating: $\widehat{p}(\mathbf{x})$ and $\widehat{q}(\mathbf{x})$.

One Dimensional Solution

- Our estimate:

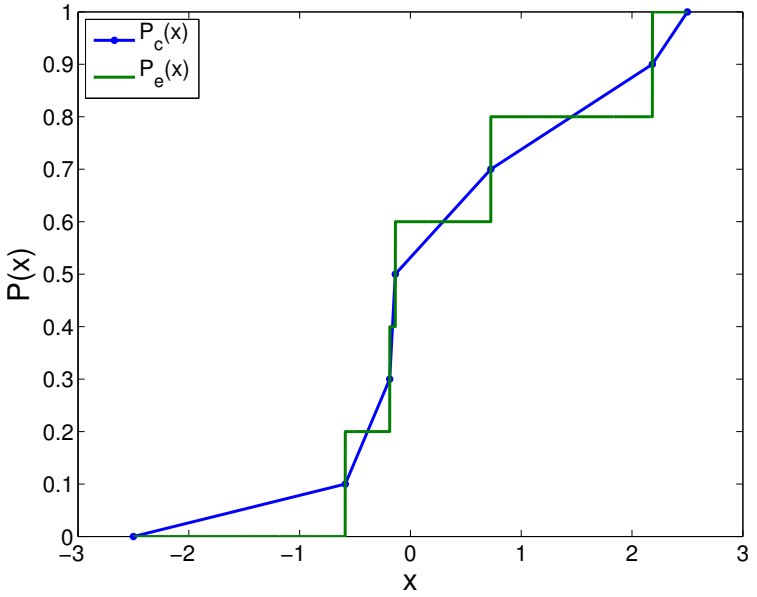
$$\widehat{D}(P||Q) = \frac{1}{n}\sum_{i=1}^{n} \log \frac{\delta P_c(x_i)}{\delta Q_c(x_i)} \quad \xrightarrow{a.s.} \quad D(P||Q) + 1$$

where $\delta P_c(x_i) = P_c(x_i) - P_c(x_i - \epsilon)$

for a sufficiently small ϵ.

$$P_e(x) = \frac{1}{n}\sum_{i=1}^{n} U(x - x_i)$$

$$P_c(x) = \begin{cases} 0, & x < x_0 \\ a_i x + b_i, & x_{i-1} \leq x < x_i \\ 1, & x_{n+1} \leq x \end{cases}$$



Key result in the proof

$$\frac{1}{n}\sum_{i=1}^{n}\log\frac{P(x_i)-P(x_{i-1})}{P_c(x_i)-P_c(x_{i-1})}=\frac{1}{n}\sum_{i=1}^{n}\log n\Delta P(x_i)$$

- x_i is distributed according to $p(x)$.
- $P(x_i)$ is distributed according to a uniform random variable.
- $z_i = n\Delta P(x_i)$ is the difference (waiting time) between two consecutive samples from a uniform distribution between 0 and n with one arrival per unit-time.

Multidimensional Solution

$$\widehat{D}_k(P||Q) = \frac{1}{n}\sum_{i=1}^{n} \log \frac{\widehat{p}_k(\mathbf{x}_i)}{\widehat{q}_k(\mathbf{x}_i)} = \frac{d}{n}\sum_{i=1}^{n} \log \frac{r_k(\mathbf{x}_i)}{s_k(\mathbf{x}_i)} + \log \frac{m}{n-1}$$

where

$$\widehat{p}_k(\mathbf{x}_i) = \frac{k\pi^{d/2}}{(n-1)\Gamma(d/2+1)r_k(\mathbf{x}_i)^d}$$

$$\widehat{q}_k(\mathbf{x}_i) = \frac{k\pi^{d/2}}{m\Gamma(d/2+1)s_k(\mathbf{x}_i)^d}$$

and $r_k(\mathbf{x}_i)$ and $s_k(\mathbf{x}_i)$ are, respectively, the Euclidean distances to the k^{th} nearest-neighbour of $\mathbf{x}_i$ in $\mathcal{X}\backslash\mathbf{x}_i$ and $\mathcal{X}'$.

Lemma I

- $p(\mathbf{x})/\widehat{p}_1(\mathbf{x}) \quad \xrightarrow{a.s.} \quad \mathcal{E}(1).$

- Proof:
  - Assume $p(\mathbf{x})$ is d-dimensional uniform distribution.
  - Define $\mathcal{S}_{\mathbf{x},R} = \{\mathbf{x}_i | \ \|\mathbf{x}_i - \mathbf{x}\|_2 \leq R\}$.
  - $\|\mathbf{x}_i - \mathbf{x}\|_2^d \sim U(0, R^d)$ for $\mathbf{x}_i \in \mathcal{S}_{\mathbf{x},R}$.
  - $r_1(\mathbf{x})^d = \min_{\mathbf{x}_j \in \mathcal{S}_{\mathbf{x},R}} (\|\mathbf{x}_j - \mathbf{x}\|_2^d)$ is an exponential random variable.
  - For non-uniform $p(\mathbf{x})$: $p(\mathbf{x}_{nn}) \rightarrow p(\mathbf{x})$, as $n \rightarrow \infty$.

Theorem

- $\widehat{D}_k(P||Q) \xrightarrow{a.s.} D(P||Q)$.

- Proof:

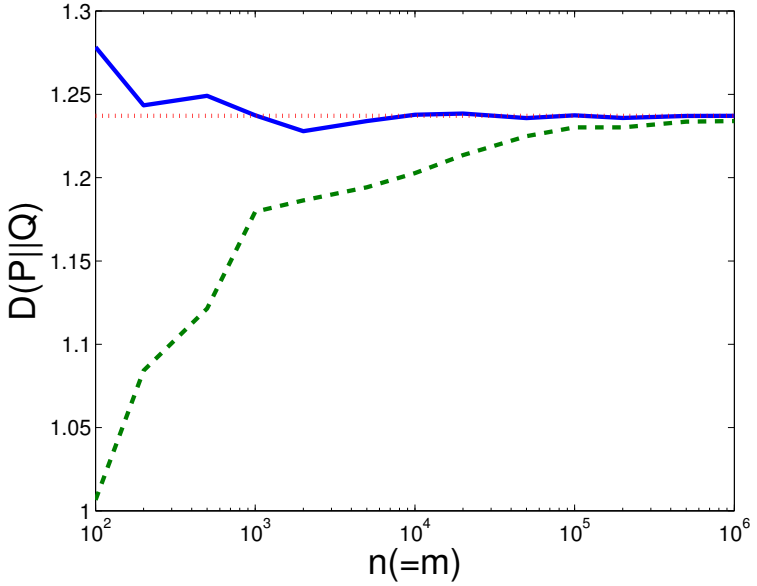
$$\widehat{D}_k(P||Q) = \frac{1}{n}\sum_{i=1}^{n} \log \frac{p(\mathbf{x}_i)}{q(\mathbf{x}_i)} - \frac{1}{n}\sum_{i=1}^{n} \log \frac{p(\mathbf{x}_i)}{\widehat{p}_k(\mathbf{x}_i)} + \frac{1}{n}\sum_{i=1}^{n} \log \frac{q(\mathbf{x}_i)}{\widehat{q}_k(\mathbf{x}_i)}$$

  The sum of random variables that converge in probability converges almost surely.

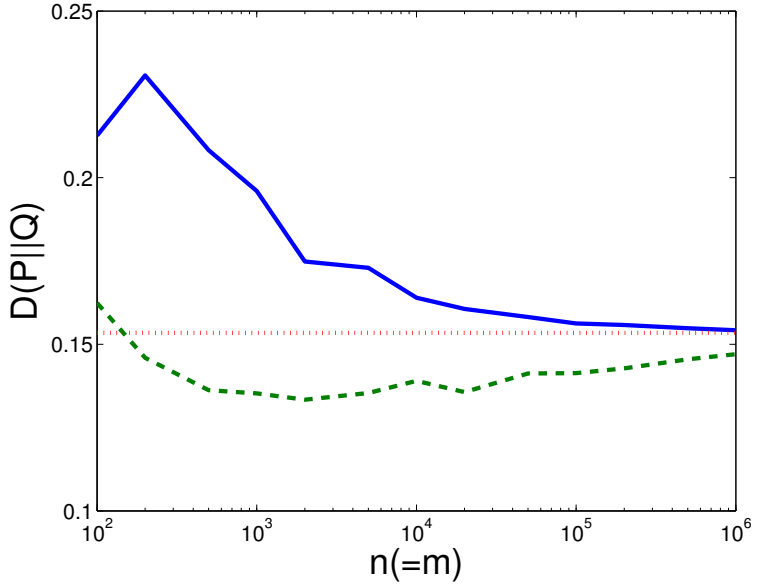
Example I

(a) $D(\mathcal{E}(1)||\mathcal{N}(3,4))$.

(b) $D(\mathcal{N}(0,2)||\mathcal{N}(0,1))$.



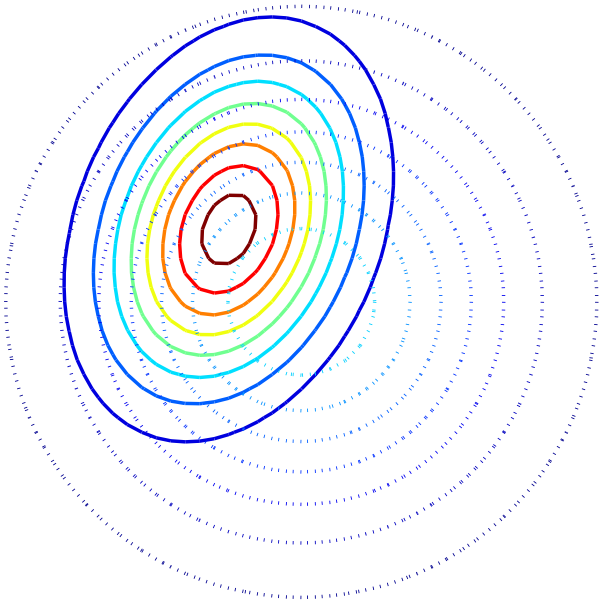
(a)



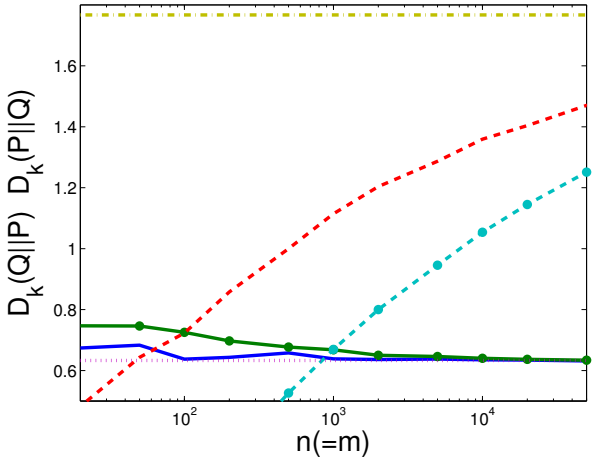
(b)

Example II

$$p(\mathbf{x}) = \mathcal{N}(\mathbf{0}, \mathbf{I}) \qquad q(\mathbf{x}) = \mathcal{N}\left(\begin{bmatrix} 0.5 \\ -0.5 \end{bmatrix}, \begin{bmatrix} 0.5 & 0.1 \\ 0.1 & 0.3 \end{bmatrix}\right)$$



(a)

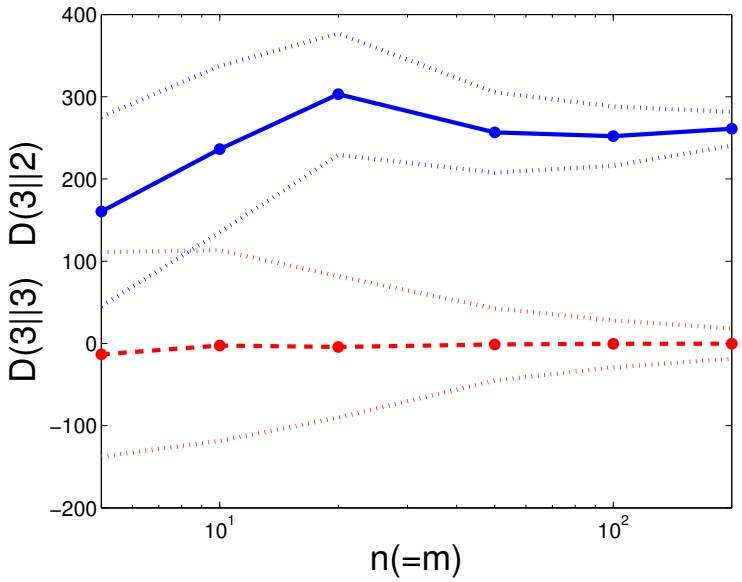


(b)

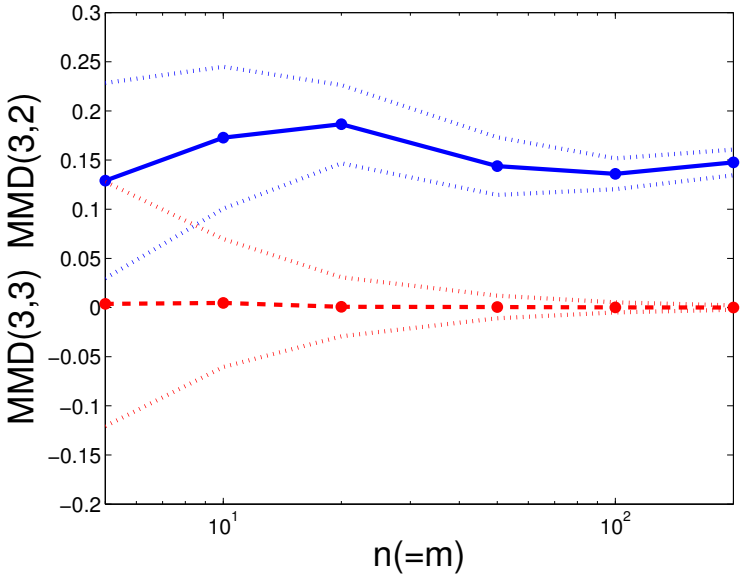
Example III: MNIST Data Set

(a) $\widehat{D}_1(P||Q)$.

(b) MMD test.



(a)



(b)

Annex

Theorem 1 *Let P and Q be absolutely continuous probability measures and assume its KL divergence is finite. Let $\mathcal{X} = \{x_i\}_{i=1}^n$ and $\mathcal{X}' = \{x_i'\}_{i=1}^m$ be i.i.d. samples sorted in increasing order, respectively, from P and Q, then*

$$\widehat{D}(P||Q) - 1 \xrightarrow{a.s.} D(P||Q)$$

We can rearrange $\widehat{D}(P||Q)$ as follows:

$$\widehat{D}(P||Q) = \frac{1}{n}\sum_{i=1}^{n} \log \frac{\delta P_c(x_i)/\epsilon}{\delta Q_c(x_i)/\epsilon} = \frac{1}{n}\sum_{i=1}^{n} \log \frac{\Delta P_c(x_i)/\Delta x_i}{\Delta Q_c(x'_{mi})/\Delta x'_{mi}}$$

where $\Delta P_c(x_i) = P_c(x_i) - P_c(x_{i-1})$, $\Delta x_i = x_i - x_{i-1}$ and $\Delta x'_{mi} = \min\{x'_j | x'_j \geq x_i\} - \max\{x'_j | x'_j < x_i\}$. The last equality holds because $P_c(x)$ and $Q_c(x)$ are piecewise linear approximations to their cdfs.

$$\widehat{D}(P||Q) = \frac{1}{n}\sum_{i=1}^{n} \log \frac{\frac{\Delta P(x_i)}{\Delta x_i}}{\frac{\Delta Q(x'_{mi})}{\Delta x'_{mi}}} - \frac{1}{n}\sum_{i=1}^{n} \log \frac{\Delta P(x_i)}{\Delta P_c(x_i)} + \frac{1}{n}\sum_{i=1}^{n} \log \frac{\Delta Q(x'_{mi})}{\Delta Q_c(x'_{mi})} =$$

$$= \widehat{D}_e(P||Q) - C_1(P) + C_2(P,Q)$$

The first term:

$$\widehat{D}_e(P||Q) = \frac{1}{n}\sum_{i=1}^{n} \log \frac{\Delta P(x_i)/\Delta x_i}{\Delta Q(x'_{mi})/\Delta x'_{mi}} \xrightarrow{a.s.} D(p||q),$$

because $\lim_{n\to\infty} \Delta P(x_i)/\Delta x_i = p(x_i)$ and $\lim_{n\to\infty} \Delta Q(x'_{mi})/\Delta x'_{mi} = q(x_i)$, due to $p(x)$ is absolutely continuous with respect to $q(x)$.

The second term:

$$C_1(P) = \frac{1}{n}\sum_{i=1}^{n} \log \frac{\Delta P(x_i)}{\Delta P_c(x_i)} = \frac{1}{n}\sum_{i=1}^{n} \log n\Delta P(x_i)$$

As x_i is distributed according to $p(x)$, $P(x_i)$ is distributed according to a uniform random variable between 0 and 1. $z_i = n\Delta P(x_i)$ is the difference (waiting time) between two consecutive samples from a uniform distribution between 0 and n with one arrival per unit-time, therefore it is distributed like an unit-mean exponential random variable. Consequently

$$C_1(P) = \frac{1}{n}\sum_{i=1}^{n} \log z_i \xrightarrow{a.s.} \int_0^\infty \log z e^{-z} dz = -0.577$$

and $C_1(P)$ is independent of $p(x)$.

The third term:

$$C_2(P,Q) = \frac{1}{n}\sum_{j=1}^{m} n\Delta P_e(x'_j)\log\frac{\Delta Q(x'_j)}{\Delta Q_c(x'_j)} =$$

$$= \frac{1}{m}\sum_{j=1}^{m}\frac{\frac{\Delta P_e(x'_j)}{\Delta x'_j}}{\frac{\Delta Q(x'_j)}{\Delta x'_j}} m\Delta Q(x'_j)\log m\Delta Q(x'_j)$$

$$\xrightarrow{a.s.} \int \frac{p_e(x)}{q(x)} z\log z e^{-z} q(x)dzdx = \int_0^\infty z\log z e^{-z}dz \int_{\mathbb{R}} p_e(x)dx = 0.423,$$

where $n\Delta P_e(x'_j)$ counts the number of samples from the set $\mathcal{X}$ between two consecutive samples from $\mathcal{X}'$. $\Delta Q(x'_j)/\Delta x'_j$ and $\Delta P_e(x'_j)/\Delta x'_j$ tend, respectively, to $q(x)$ and to $p_e(x)$. $p_e(x)$ is a density model, but it does not need to tend to $p(x)$ for $C_2(P,Q)$ to converge to 0.423.

The three terms, respectively, converge almost surely to $D(P||Q)$, -0.577 and 0.423, due to the strong law of large numbers. Integrating by parts one can easily show that $\int_0^\infty \log z(z-1)e^{-z} = 1$.

Note: we can readily understand that we are using a data-dependent histogram, in which we put one sample in each bin, as density estimate for $p(x)$ and $q(x)$, e.g. $\widehat{p}(x_i) = 1/n\Delta x_i$, to estimate the KL divergence. Data-dependent histograms converge to their true measures when two conditions are met. The first condition states that the number of bins must grow sublinearly with the number of samples and this condition is violated by our density estimate. Hence our KL divergence estimator converges almost surely, but it is based on non-convergent density estimates.