Akademie věd České republiky
Ústav teorie informace a automatizace

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

RESEARCH REPORT

Jitka Homolová, Miroslav Kárný

Evaluation of Kullback-Leibler Divergence

2349 May 2015

ÚTIA AVČR, P.O.Box 18, 182 08 Prague,
Czech Republic
Fax: (+420)286890378, http://www.utia.cas.cz, E-mail: utia@utia.cas.cz

Abstract

Kullback-Leibler divergence is a leading measure of similarity or dissimilarity of probability distributions. This technical paper collects its analytical and numerical expressions for the broad range of distributions.

Keywords

Kullback-Leibler divergence, cross-entropy, Bayesian decision making, Bayesian learning and approximation vspace2cm

1 Introduction

Kullback-Leibler divergence (Kullback and Leibler (1951), a.k.a. cross-entropy) has become a standard measure within Bayesian approximation (Bernardo, 1979) or the extension of partial knowledge via minimum-cross entropy principle (Shore and Johnson, 1980) and the fully probabilistic design of decision-making strategies (Kárný and Guy, 2006; Kárný and Kroupa, 2012). The multitude of other cases can be also found. For some common families of distributions, the Kullback-Leibler divergence KLD can be derived in a closed form. There are known analytical expressions of the KLD in some important cases or standard numerical procedures in others. They are, however, spread in various sources and we found it useful to collect them on a single place and partially to complement them. The current paper provides the results of this effort.

2 Kullback-Leibler Divergence and its Elementary Properties

The Kullback-Leibler divergence (Kullback and Leibler, 1951) is a non-symmetric measure of the dissimilarity between two probability distributions P and Q and it is denoted $\mathsf{D}_{KL}(P||Q)$. In probability and information theory, it can be interpreted as a measure of the information lost when Q is used to approximate P (Burnham and Anderson, 2002) or as a measure of the expected number of extra bits required to code samples from P when using a code based on Q, rather than using a code based on P. Typically, P represents observations, data measurements, or a precisely calculated theoretical distributions. On the other hand, Q then represents a theoretical model, description or approximation of P. KL divergence is a special case of a broader class of divergences called f-divergences (Csiszár, 1963).

2.1 Discrete probability distributions

For discrete probability distributions P and Q, the Kullback-Leibler divergence from P to Q is defined as

$$\mathsf{D}_{KL}(P||Q) = \sum_i \ln\left(\frac{P(i)}{Q(i)}\right) P(i), \tag{1}$$

where P and Q both sum to 1 and $Q(i) = 0$ implies $P(i) = 0$ for all i (i.e. it is absolutely continuous).

In other words, the KLD is the expectation of the logarithmic difference between the probabilities P and Q, where the expectation is taken using the probabilities P.

2.2 Continuous probability distributions

Let's assume that P and Q are probability measures over a set $\mathbf{\Theta}$, and P is absolutely continuous with respect to Q, then KLD is defined as

$$\mathsf{D}_{KL}(P||Q) = \int_{\mathbf{\Theta}} \ln\left(\frac{\mathrm{d}P}{\mathrm{d}Q}\right) \mathrm{d}P, \tag{2}$$

where $\frac{\mathrm{d}P}{\mathrm{d}Q}$ is the Radon-Nikodým derivative of P with respect to Q (Bishop, 2006), if the expression on the right-hand side exists. This relationship can be also written as follows:

$$\mathsf{D}_{KL}(P||Q) = \int_{\mathbf{\Theta}} \ln\left(\frac{\mathrm{d}P}{\mathrm{d}Q}\right) \frac{\mathrm{d}P}{\mathrm{d}Q} \mathrm{d}Q, \tag{3}$$

which can be interpreted as the entropy of P relative to Q.

Provided that θ is any measure of the continuous random variable Θ on the set $\mathbf{\Theta}$ for which $p = \frac{\mathrm{d}P}{\mathrm{d}\theta}$ and $q = \frac{\mathrm{d}Q}{\mathrm{d}\theta}$ exist, the KLD from P to Q is given as

$$\mathsf{D}_{KL}(P||Q) = \int_{\mathbf{\Theta}} p \ln \frac{p}{q} \mathrm{d}\theta. \tag{4}$$

It can be seen, that the discrete case of the KLD can be related to (4) when considering $\mathrm{d}\theta$ to be a counting measure.

2.3 Properties of KLD

This subsection summarizes key properties of the Kullback-Leibler divergence:

Non-negativity: the KLD is always non-negative, i.e.

$$\mathsf{D}_{KL}(P||Q) \geq 0; \tag{5}$$

This result is known as Gibbs' inequality in case of discrete probabilities.

Self similarity and self identification:

$$\begin{aligned} &\mathsf{D}_{KL}(P||P) = 0 \\ &\mathsf{D}_{KL}(P||Q) = 0 \Leftrightarrow P = Q \quad \text{(almost surely);} \end{aligned} \tag{6}$$

Asymmetry: the KLD is a non-symmetric functional, i.e.

$$\mathsf{D}_{KL}(P||Q) \neq \mathsf{D}_{KL}(Q||P); \tag{7}$$

Upper limit: the Kullback-Leibler distance is infinite if $p = \frac{\mathrm{d}P}{\mathrm{d}\theta} = 0$ and $q = \frac{\mathrm{d}Q}{\mathrm{d}\theta} > 0$ on a set of a positive $\mathrm{d}\theta$-volume;

Triangle inequality: the KLD does not obey the triangle inequality. It means that for the probability distributions P, Q and R holds

$$\mathsf{D}_{KL}(P||R) \not\leq \mathsf{D}_{KL}(P||Q) + \mathsf{D}_{KL}(Q||R); \tag{8}$$

Convex function: the KLD follows the Jensen's inequality, that relates the value of a convex function of an integral to the integral of the convex function. The KLD is convex in its first argument, but not necessarily in the second argument. If φ_P is a convex function, then

$$\varphi_P(\mathsf{D}_{KL}(P||R)) \leq \mathsf{D}_{KL}(\varphi_P(P)||Q), \tag{9}$$

where $\varphi_P(P)$ is point-wisely defined. Moreover, it can be easily shown (Kullback and Leibler, 1951) that the KLD is strictly convex in P, i.e. $\varphi_P(\mathsf{D}_{KL}(P||R)) < \mathsf{D}_{KL}(\varphi_P(P)||Q)$, because the KLD as the integral part over the support complement of the probability distribution P is zero, so it (almost surely) holds $\mathsf{supp}(P) \subseteq \mathsf{supp}(Q)$. This property can be seen from a key result about Bregman divergences, covering the generalized Kullback-Leibler divergence, saying the mean vector minimizes the expected Bregman divergence from the random vector (Banerjee et al., 2005), (Frigyik et al., 2008).

Parameter transformation: the KLD is invariant under parameter transformations. For example, if a transformation is made from variable θ to variable $\phi(\theta)$ and since $P(\theta)\,\mathrm{d}\theta = P(\phi)\,\mathrm{d}\phi$ and $Q(\theta)\,\mathrm{d}\theta = Q(\phi)\,\mathrm{d}\phi$, then the KLD may be rewritten:

$$\begin{aligned}\mathsf{D}_{KL}(P||Q) &= \int_{\mathbf{\Theta}} p(\theta)\ln\frac{p(\theta)}{q(\theta)}\,\mathrm{d}\theta = \int_{\mathbf{\Phi}} p(\phi)\ln\frac{p(\phi)\frac{\mathrm{d}\phi}{\mathrm{d}\theta}}{q(\phi)\frac{\mathrm{d}\phi}{\mathrm{d}\theta}}\,\mathrm{d}\phi = \\ &= \int_{\mathbf{\Phi}} p(\phi)\ln\frac{p(\phi)}{q(\phi)}\,\mathrm{d}\phi; \end{aligned} \tag{10}$$

Additivity: the KLD is additive for distributions of independent random variables. In case that P_1, P_2 are distributions of two independent variables, with the joint distribution $P(\theta_1,\theta_2) = P_1(\theta_1)P_2(\theta_2)$, and Q, Q_1, Q_2 likewise, then:

$$\mathsf{D}_{KL}(P||Q) = \mathsf{D}_{KL}(P_1||Q_1) + \mathsf{D}_{KL}(P_2||Q_2). \tag{11}$$

Existence of minimum: the minimum $P^* \in \mathbf{P}$ of the distribution $P \in \mathbf{P}$ exists and moreover, it is the unique one for the given Q:

$$P^* = \arg\min_{P\in\mathbf{P}} \mathsf{D}_{KL}(P||Q) \quad \text{is unique,} \tag{12}$$

wherever the set $\mathbf{P}$ is non-empty, convex and closed.

3 Analytical Expressions

The subsequent sections provide the main practical contribution of the paper by collecting cases in which the KLD can be evaluated in a closed form. Often, it is done using the form of the KLD in terms of expected values $\mathsf{E}(\bullet) = \int_{\mathbf{\Theta}} \bullet p(\theta)\,\mathrm{d}\theta$ or in terms of information entropy:

$$\mathsf{D}_{KL}(P||Q) = -\mathsf{E}(\ln q) + \mathsf{E}(\ln p) = \mathsf{K}(P,Q) - \mathsf{H}(P), \tag{13}$$

where $\mathsf{H}(P) = -\mathsf{E}(\ln p)$ is the information entropy of P and $\mathsf{K}(P,Q)$ is the Kerridge inaccuracy (Kerridge, 1961) of P and Q.

From the point of view of generality, the article only focuses on multivariate continuous random variable and its joint probability distribution.

3.1 Notation

This subsection summarizes the notation of parameters, variables or functions used in next sections. A continuous random variable is denoted Θ as before. It is a continuous random variable with values $\theta \in \mathbf{\Theta}$ that is the subset of finite dimensional real space. The random variable has the form of a vector (or a matrix) and its length (the number of entries) is denoted $\ell_{\{\Theta\}}$. Probability density functions of probabilities P and Q are denoted by the symbol $\mathsf{f}_S(\theta)$, where the index $S \in \mathbf{S}$ is a value of a finite dimensional statistic from the finite dimensional real space. It is referred as *Parameters* and determines the probability densities:

$$\frac{\mathrm{d}P}{\mathrm{d}\theta} = \mathsf{f}_{S_a}(\theta), \quad \frac{\mathrm{d}Q}{\mathrm{d}\theta} = \mathsf{f}_{S_b}(\theta), \quad S_a, S_b \in \mathbf{S}. \tag{14}$$

Using the definition equation (3) and the above mentioned densities (14), the Kullback-Leibler divergence can be referred as follows:

$$\mathsf{D}_{KL}(P||Q) = \int_{\mathbf{\Theta}} \ln\left(\frac{\mathrm{d}P}{\mathrm{d}Q}\right) \frac{\mathrm{d}P}{\mathrm{d}Q}\ \mathrm{d}Q = \int_{\mathbf{\Theta}} \mathsf{f}_{S_a} \ln \frac{\mathsf{f}_{S_a}}{\mathsf{f}_{S_b}}\ \mathrm{d}\theta = \mathsf{D}(\mathsf{f}_{S_a}||\mathsf{f}_{S_b}). \tag{15}$$

3.2 KLD on Exponential Family of Probability Densities

The exponential family of probability densities has an exceptional role because its members have non-trivial finite-dimensional sufficient statistics (Koopman, 1936) and consequently they admit non-trivial conjugated distributions. These probability densities have the following form (Barndorff-Nielsen, 1978)

$$\begin{aligned} \mathsf{f}_S(\theta) &= \frac{\exp\langle S, \mathsf{C}(\theta)\rangle}{\mathcal{I}(S)}, \\ \mathcal{I}(S) &= \int_{\boldsymbol{\theta}} \exp\langle S, \mathsf{C}(\theta)\rangle\ \mathrm{d}\theta, \end{aligned} \tag{16}$$

where $\langle\cdot,\cdot\rangle$ is a scalar product and the statistic S is a finite-dimensional real array compatible with the values of the mapping $\mathsf{C}(\theta)$ (with respect to the scalar product). The definition makes sense only if the normalization integral is limited from above, i.e. $\mathcal{I}(S) < \infty$.

Often, the probability density $\mathsf{f}_S(\theta)$ (16) serves as an approximation of another probability density, let's say $\mathsf{g}(\theta)$. Then the statistics S in (16) is chosen to minimise the Kullback-Leibler divergence $\mathsf{D}(\mathsf{g}||\mathsf{f}_S)$, (Bernardo, 1979).

Applying the basic KLD definition (4) on probability densities (16), we easily obtain

$$\begin{aligned} \mathsf{D}(\mathsf{g}||\mathsf{f}_S) &= \ln(\mathcal{I}(S)) - \langle S, \mathsf{E}_\mathsf{g}(\mathsf{C})\rangle + \int_{\boldsymbol{\theta}} \mathsf{g}(\theta)\ln(\mathsf{g}(\theta))\ \mathrm{d}\theta, \\ \mathsf{E}_\mathsf{g}(\mathsf{C}) &= \int_{\boldsymbol{\theta}} \mathsf{g}(\theta)\mathsf{C}(\theta)\ \mathrm{d}\theta. \end{aligned} \tag{17}$$

Thus, the KLD depends on the expected value $\mathsf{E}_\mathsf{g}(\mathsf{C})$ of the function $\mathsf{C}(\theta)$ with respect to the approximated probability density $\mathsf{g}(\theta)$. The necessary condition for extreme with respect to S is the zero value of the first derivative. Due to (17), it has the following form:

$$\begin{aligned} \frac{\mathrm{d}\mathsf{D}(\mathsf{g}||\mathsf{f}_S)}{\mathrm{d}S} &= \frac{\mathrm{d}\ln(\mathcal{I}(S))}{\mathrm{d}S} - \mathsf{E}_\mathsf{g}(\mathsf{C}) = \\ &= \mathsf{E}_{\mathsf{f}_S}(\mathsf{C}) - \mathsf{E}_\mathsf{g}(\mathsf{C}) = 0, \\ \mathsf{E}_{\mathsf{f}_S}(\mathsf{C}) &= \int_{\boldsymbol{\theta}} \mathsf{f}_S(\theta)\mathsf{C}(\theta)\ \mathrm{d}\theta. \end{aligned} \tag{18}$$

As seen from (18), the optimal approximation thus fits the moment $\mathsf{E}_{\mathsf{f}_S}(\mathsf{C})$ to the moment $\mathsf{E}_\mathsf{g}(\mathsf{C})$. The choice of the fitted moments is *not* arbitrary but, on the contrary, uniquely determined by the considered approximate probability density $\mathsf{f}_S(\theta)$.

It is uncomplicated to verify that the second derivative of $\mathsf{D}(\mathsf{g}||\mathsf{f}_S)$ with respect to S is the covariance of $\mathsf{C}(\theta)$ with respect to $\mathsf{f}_S(\theta)$ and thus it is positive semi-definite. It means that $\mathsf{D}(\mathsf{g}||\mathsf{f}_S)$ is (not-necessarily strictly) the convex function of S. Moreover, for all S for which probability density $\mathsf{f}_S(\theta)$ is non-degenerate, the second derivative is positive definite. In this generic case, the minimiser is the unique one.

4 KLD in Recursive Bayesian Estimation

Approximate recursive Bayesian estimation often relies on the probability density from the exponential family in the form that can be defined by the relation (16) (Kárný et al., 2006; Kárný, 2014). Essentially, it deals with probability density $\mathsf{g}(\theta)$ resulting from the Bayes rule (Peterka, 1981) and having the form

$$\begin{aligned} \mathsf{g}_{S_a}(\theta) &= \frac{\mathsf{m}(\theta)\exp\langle S_a, \mathsf{C}(\theta)\rangle}{\mathcal{I}_\mathsf{m}(S_a)}, \\ \mathcal{I}_\mathsf{m}(S_a) &= \int_{\boldsymbol{\theta}} \mathsf{m}(\theta)\exp\langle S_a, \mathsf{C}(\theta)\rangle\ \mathrm{d}\theta, \end{aligned} \tag{19}$$

where $\mathsf{m}(\theta) \geq 0$ is the likelihood function determined by the estimated parametric model and measured data inserted into it.

The necessary condition (18) for the optimal approximation of $\mathsf{g}_{S_a}(\theta)$ by the probability density $\mathsf{f}_{S_b}(\theta)$ (16) is determined by the statistics $S_b = \Delta + S_a$, with Δ solving the equation

$$\int_{\mathbf{\Theta}} \exp\langle S_a, \mathsf{C}(\theta)\rangle \, \mathsf{C}(\theta) \left[\frac{\mathcal{I}_{\mathsf{m}}(S_a)\exp\langle \Delta, \mathsf{C}(\theta)\rangle}{\mathcal{I}_{\mathsf{m}}(S_a+\Delta)} - \mathsf{m}(\theta)\right] \mathrm{d}\theta = 0, \tag{20}$$

whose solution generally requires repetitive integration (say by combination of Laplace and Monte Carlo methods (Robert E. Kass, 1995)) and a standard iterative solution of algebraic equations.

4.1 Analytical Example

Let's assume the f_{S_a} to be the exponential probability density with parameter $\lambda > 0$:

$$\mathsf{f}_{S_a}(\theta) = \lambda e^{-\lambda\theta}, \quad \theta > 0. \tag{21}$$

In terms of (16), it must hold:

$$S_a = -\lambda, \mathsf{C}(\theta) = \theta, \mathcal{I}(S_a) = \int_0^\infty \mathsf{e}^{-\lambda\theta} \, \mathrm{d}\theta = \frac{1}{\lambda}. \tag{22}$$

Applying (17), the KLD is then defined

$$\begin{aligned} \mathsf{D}(\mathsf{g}||\mathsf{f}_{S_a}) &= \ln\left(\frac{1}{\lambda}\right) + \lambda \mathsf{E}_{\mathsf{g}}(\mathsf{C}) + \int_{\boldsymbol{\theta}} \mathsf{g}(\theta)\ln(\mathsf{g}(\theta)) \, \mathrm{d}\theta, \\ \mathsf{E}_{\mathsf{g}}(\mathsf{C}) &= \int_{\boldsymbol{\theta}} \theta \mathsf{g}(\theta) \, \mathrm{d}\theta. \end{aligned} \tag{23}$$

Now, let's assume the probability distribution g to be also an exponential distribution with parameters S_b:

$$\mathsf{g} = \mathsf{f}_{S_b}(\theta) = \beta e^{-\beta\theta}, \quad \theta > 0. \tag{24}$$

Due to 23, the KLD can be calculated by

$$\mathsf{D}(\mathsf{f}_{S_b}||\mathsf{f}_{S_a}) = \ln\left(\frac{\beta}{\lambda}\right) + \frac{\lambda}{\beta} - 1. \tag{25}$$

5 Review of the Analytical Expressions

This section contains KLDs between two distributions from a same distribution family for which an analytical expression exists. It focuses on the frequently used distributions, mostly from the exponential family. For the clarity, the

main information about probability distribution, random variable, probability density functions, entropy and Kullback-Leibler divergence related to a specific distribution family are arranged into a table of the following form:

Name:	**Name of the probability distribution**
Variable:	Description of the random variable Θ.
Parameters:	The parameter S determining the probability density function.
Pdf:	Probability density function $\mathsf{f}_S(\theta)$.
Entropy:	Entropy of the probability distribution $\mathsf{H}(\mathsf{f}_S)$.
***KLD*:**	Kullback-Leibler divergence $\mathsf{D}(\mathsf{f}_{S_a}\|\|\mathsf{f}_{S_b})$.
***Remarks*:**	Typically, comments on proof or source where it comes from.

In formulas in text, the symbol $'$ denotes vector transposition. The symbol tr[$\cdot$] denotes matrix trace, i.e. the sum of the entries on the main matrix diagonal.

5.1 Gaussian (Normal) distribution

First, the KLD between two multivariate Gaussian (normal) distributions with and their corresponding means and nonsingular covariance matrices is treated.

Name:	**Gaussian Distribution**
Variable:	Θ – $\ell_{\{\Theta\}}$-dimensional column vector
Parameters:	S =[μ; R] =[mean; covariance matrix]= =[$\ell_{\{\Theta\}}$-dimensional column vector; $\ell_{\{\Theta\}} \times \ell_{\{\Theta\}}$ positive definite matrix ($R > 0$)].
Pdf:	$\mathsf{f}_S(\theta) = (2\pi)^{-\frac{\ell_{\{\Theta\}}}{2}} \|R\|^{-\frac{1}{2}} \exp\left[-\frac{1}{2}(\theta-\mu)' R^{-1} (\theta-\mu)\right]$
Entropy:	$\mathsf{H}(\mathsf{f}_S) = \frac{1}{2}\ell_{\{\Theta\}}(1+\ln(2\pi)) + \frac{1}{2}\ln\|R\|$
***KLD*:**	$\mathsf{D}(\mathsf{f}_{S_a}\|\|\mathsf{f}_{S_b}) = \frac{1}{2}\left[\mathrm{tr}\left[R_b^{-1}R_a\right] + (\mu_b-\mu_a)' R_b^{-1} (\mu_b-\mu_a)\right] - -\ell_{\{\Theta\}} + \ln\|R_b R_a^{-1}\|$
***Remarks*:**	See (Penny, 2001).

Some applications (Peterka, 1981) work with a matrix form of the Gaussian probability density functions. This form means that when arranging Θ into column vector (by stacking matrix columns) then the same arrangement of μ provides its mean and it has the covariance matrix $R \otimes Q$, where $\otimes$ denotes Kronecker product.

Name: **Gaussian Distribution - Matrix Form**

Variable: Θ – $(\ell_{\{T\}}, \ell_{\{R\}})$-dimensional matrix

Parameters: S= $[\mu; (R,T)]$=[mean; factors of the covariance matrix]=
= [$(\ell_{\{T\}}, \ell_{\{R\}})$-dimensional matrix;
Kronecker factors of covariance matrix, R>0, T>0].

Pdf: $\mathsf{f}_S(\theta) = |2\pi R|^{-\frac{\ell_{\{T\}}}{2}} |2\pi T|^{-\frac{\ell_{\{R\}}}{2}} \exp\left\{-\frac{\mathrm{tr}[R^{-1}(\theta-\mu)'T^{-1}(\theta-\mu)]}{2}\right\}$

Entropy: $\mathsf{H}(\mathsf{f}_S)$=$\frac{1}{2}\ell_{\{T\}}(1+\ln(2\pi)+\ln|R|)+$
$+\frac{1}{2}\ell_{\{R\}}(1+\ln(2\pi)+\ln|T|)$

KLD: $\mathsf{D}(\mathsf{f}_{S_a}||\mathsf{f}_{S_b})$= $\frac{1}{2}\left\{\ell_{\{R\}}\ln|T_bT_a^{-1}| + \ell_{\{T\}}\ln|R_bR_a^{-1}| - \right.$
$-\ell_{\{R\}}\ell_{\{T\}} + \mathrm{tr}\left[R_a\otimes T_aT_b^{-1}\otimes R_b^{-1}\right] +$
$\left.+\mathrm{tr}[R_b^{-1}(\mu_b-\mu_a)'T_b^{-1}(\mu_b-\mu_a)\right\}.$

Remarks: See Proposition 8.10 in (Kárný et al., 2006). The result is obtained by direct application of the results for the multivariate Gaussion distribution and the following identities (Graham, 1981):
$\mathrm{tr}[A\otimes B] = \mathrm{tr}[A]\mathrm{tr}[B]$,
$[A\otimes B]^{-1} = A^{-1}\otimes B^{-1}$,
$|A\otimes B| \quad = |A|^{\ell_{\{B\}}}|B|^{\ell_{\{A\}}}$.

5.2 Gaussian-Inverse-Gamma Distribution

Gaussian-inverse-gamma (or normal-inverse-gamma) distribution is conjugated to Gaussian regression model with single output (Berger, 1985; Kárný et al., 2006). It is one-dimensional version of Gauss-inverse-Wishart distribution (Zellner, 1976).

Multivariate output can always be treated as the collection of the single output cases (Zellner, 1976):

$$\begin{aligned}\mathsf{f}_S(\theta|\bullet) &= \prod_{i=1}^{n}\mathsf{f}_S(\theta_i|\theta_{i+1},\ldots,\theta_n,\bullet), \quad \text{where} \\ \theta &= [\theta_1,\ldots,\theta_i,\theta_{i+1},\ldots,\theta_n]'.\end{aligned} \tag{26}$$

Name: **Gaussian-Inverse-Gamma Distribution**

Variable: $\Theta = [\theta; r]$=[mean; variance]=
$= [\ell_{\{\Theta\}}$-dimensional column vector; positive variance]

Parameters: $S = [\hat{\theta}, \hat{r}; C; \nu]$=[maximum likelihood estimates of θ and r, $\hat{r} > 0$; covariance factor of θ, $C > 0$; degrees of freedom, $\nu > 0$]

Pdf: $\mathsf{f}_S(\theta) = \frac{r^{-\frac{1}{2}(\nu+\ell_{\{\theta\}}+2)}}{\mathcal{I}(S)} \exp\left\{-\frac{1}{2r}\left[(\theta-\hat{\theta})'C^{-1}(\theta-\hat{\theta})+\hat{r}\right]\right\}$

$\mathcal{I}(S) = \Gamma(\frac{\nu}{2})\hat{r}^{-\frac{\nu}{2}}\,|C|^{\frac{1}{2}}\,(2\nu)^{\frac{\nu}{2}}(2\pi)^{\frac{\ell_{\{\Theta\}}}{2}}$,

$\Gamma(\tau) = \int_0^\infty z^{\tau-1}\exp(-z)\,\mathrm{d}z$, $\ell_{\{\Theta\}} = \ell_{\{\hat{\Theta}\}}$.

KLD: $\mathsf{D}(\mathsf{f}_{S_a}||\mathsf{f}_{S_b}) = \ln\left(\frac{\Gamma(\frac{\nu_b}{2})}{\Gamma(\frac{\nu_a}{2})}\right) - \frac{1}{2}\ln\left|C_aC_b^{-1}\right| + \frac{\nu_b}{2}\ln\left(\frac{\hat{r}_a}{\hat{r}_b}\right) +$
$+\frac{1}{2}(\nu_a-\nu_b)\frac{\partial}{\partial(\frac{\nu_a}{2})}\ln\left(\Gamma(\frac{\nu_a}{2})\right) + \frac{1}{2}\mathrm{tr}\left[C_aC_b^{-1}\right] -$
$-\frac{\ell_{\{\theta\}}}{2} - \frac{\nu_a}{2} + \frac{\nu_a}{2\hat{r}_a}\left[\left(\hat{\theta}_a-\hat{\theta}_b\right)'C_b^{-1}\left(\hat{\theta}_a-\hat{\theta}_b\right)+\hat{r}_b\right]$.

Remarks: See (Penny, 2001).

5.3 Uniform Distribution

Uniform distribution is a prototype of probability density with Θ-dependent supports and it is out of the exponential family.

Name: **Uniform Distribution**

Variable: $\boldsymbol{\Theta}$ – $\ell_{\{\Theta\}}$-dimensional vector

Parameters: S– two $\ell_{\{\Theta\}}$-dimensional vectors describing minimum and maximum support values

Pdf: $\mathsf{f}_S(\Theta)=\frac{\chi_S(\Theta)}{|S|}$,
$\chi_S(\Theta)$ is indicator of S, $\chi_S(\Theta)$=1 if and only if $\Theta \in S$,
$|S|$ the denotes volume of the interval determined by S.

Entropy: $\mathsf{H}(\mathsf{f}_S) = \ln|S|$

KLD: $\mathsf{D}(\mathsf{f}_{S_a}||\mathsf{f}_{S_b}) = \begin{cases} \ln\left(\frac{|S_b|}{|S_a|}\right) & \text{if } S_a \subseteq S_b \\ \infty & \text{otherwise} \end{cases}$

Remarks: Resulting from the definitions of uniform pdf and KLD (4).

5.4 Dirichlet Distribution

Dirichlet distribution is conjugated probability distribution to multi-nominal probability density (Kárný et al., 2006).

Name:	**Dirichlet Distribution**
Variable:	$\boldsymbol{\Theta}$ – $\ell_{\{\Theta\}}$-dimensional positive vectors summing to one
Parameters:	S – $\ell_{\{\Theta\}}$-dimensional positive vector
Pdf:	$\mathsf{f}_S(\Theta) = \frac{\prod_{i=1}^{\ell_{\{\Theta\}}} \Theta_i^{S_i-1} \chi_{\boldsymbol{\Theta}}(\Theta)}{\mathcal{I}(S)}$
	$\mathcal{I}(S) = \frac{\prod_{i=1}^{\ell_{\{\Theta\}}} \Gamma(S_i)}{\Gamma(\nu)}$, $\nu = \sum_{i=1}^{\ell_{\{\Theta\}}} S_i$.
Entropy:	$\mathsf{H}(\mathsf{f}_S) = \log\left(\mathcal{I}(S)\right) + (\nu - \ell_{\{\Theta\}})\Psi(\nu) - \sum_{j=1}^{\ell_{\{\Theta\}}} (S_j - 1)\Psi(S_j)$,
	$\Psi(x) = \frac{\mathsf{d}}{\mathsf{d}x} \ln \Gamma(x)$.
KLD:	$\mathsf{D}(\mathsf{f}_{S_a} \| \mathsf{f}_{S_b}) = \sum_{d=1}^{\|\boldsymbol{d}\|} \left[(S_{d;a} - S_{d;b})\, \Psi(S_{d;a}) + \ln\left(\frac{\Gamma\left(S_{d;b}\right)}{\Gamma\left(S_{d;a}\right)}\right)\right] -$
	$-(\nu_a - \nu_b)\Psi(\nu_a) + \ln\left(\frac{\Gamma(\nu_a)}{\Gamma(\nu_b)}\right)$.
***Remarks*:**	See (T.W. Rauber and Berns, 2008).

6 Conclusions

In this paper, the definition of the Kullback-Leibler divergence is recalled for both cases of the probability distributions: discrete and continuous. The way how to find the analytical expression of the KLD is described and also demonstrated by one analytical example. The paper is focused on the frequently used distributions, mostly from the exponential family. The analytical expressions of the KLD between two distributions from the same family, for which an analytical expression exists, are collected and served by the table form.

Acknowledgements

This research has been supported by GAČR 13-13502S.

References

Banerjee, A., Merugu, S., Dhillon, I., Ghosh, J., 2005. Clustering with bregman divergences. J. of Machine Learning Research (JMLR) 6, 1705–1749.

Barndorff-Nielsen, O., 1978. Information and exponential families in statistical theory. Wiley, New York.

Berger, J., 1985. Statistical Decision Theory and Bayesian Analysis. Springer, New York.

Bernardo, J., 1979. Expected information as expected utility. The Annals of Statistics 7, 686–690.

Bishop, C., 2006. Pattern Recognition and Machine Learning. Springer.

Burnham, K., Anderson, D., 2002. Model Selection and Multi-Model Inference: A Practical Information-Theoretic Approach. Springer. 2nd edition.

Csiszár, I., 1963. Eine informationtheoretische ungleichung und ihre anwendung auf den beweis der ergodizitat von markoffschen ketten 8, 85108.

Frigyik, B., Srivastava, S., Gupta, M., 2008. Functional Bregman divergence and Bayesian estimation of distributions. Information Theory, IEEE Transactions on 54, 5130–5139.

Graham, A., 1981. Kronecker products and matrix calculus: with applications. Ellis Horwood series in mathematics and its applications, Horwood.

Kárný, M., 2014. Approximate Bayesian recursive estimation. Information Sciences 289, 100–111. DOI 10.1016/j.ins.2014.01.048.

Kárný, M., Böhm, J., Guy, T.V., Jirsa, L., Nagy, I., Nedoma, P., Tesař, L., 2006. Optimized Bayesian Dynamic Advising: Theory and Algorithms. Springer.

Kárný, M., Guy, T.V., 2006. Fully probabilistic control design. Systems & Control Letters 55, 259–265.

Kárný, M., Kroupa, T., 2012. Axiomatisation of fully probabilistic design. Information Sciences 186, 105–113.

Kerridge, D., 1961. Inaccuracy and inference. J. of the Royal Statistical Society B 23, 284–294.

Koopman, R., 1936. On distributions admitting a sufficient statistic. Trans. of American Mathematical Society 39, 399.

Kullback, S., Leibler, R., 1951. On information and sufficiency. Annals of Mathematical Statistics 22, 79–87.

Penny, W., 2001. KL-Divergences of Normal, Gamma, Dirichlet and Wishart Densities. Technical Report. Wellcome Department of Cognitive Neurology. London.

Peterka, V., 1981. Bayesian system identification, in: Eykhoff, P. (Ed.), Trends and Progress in System Identification. Pergamon Press, Oxford, pp. 239–304.

Robert E. Kass, A.E.R., 1995. Bayes factors. Journal of the American Statistical Association 90, 773–795.

Shore, J., Johnson, R., 1980. Axiomatic derivation of the principle of maximum entropy & the principle of minimum cross-entropy. IEEE Tran. on Inf. Th. 26, 26–37.

T.W. Rauber, T.B., Berns, K., 2008. Probabilistic distance measures of the dirichlet and beta distributions. Pattern Recognition 41 (2), 637–645.

Zellner, A., 1976. An Introduction to Bayesian Inference in Econometrics. J. Wiley, New York.