RENYI DIVERGENCE FOR UNIVARIATE DISTRIBUTIONS

Aljaž Ostrež

Faculty of Mathematics and Physics, Ljubljana

Institute Jozef Stefan, Ljubljana

July 21, 2020

Contents

1 Introduction

Finding an accurate method to detect errors is one of the basic problems in the control system area. Some of those methods are based on comparing two statistical samples.

At first, we will build the definition of divergence and list some examples. Then, we will focus on Renyi divergence. We will discuss methods that compare two statistical samples by using Renyi divergence: the first method is based on constructing probability density estimations from statistical samples, and the second method is based on constructing histograms.

2 Definition of divergence

As mentioned in the introduction, divergence is a measure for the difference between two statistical samples.

Definition 2.1 *Let S be space of all probability distributions with common support (i.e. all distributions in S have non-zero values on the common support).* ***Divergence*** *is the function $D(\cdot\|\cdot) : S \times S \to \mathbb{R}$, such that:*

1. $D(p\|q) \geq 0$ *for all* $p, q \in S$,
2. $D(p\|q) = 0 \Leftrightarrow p = q$.

Dual divergence D^* *is defined as* $D^*(p\|q) = D(q\|p)$.

Divergence is not necessarily symmetric and is not affected by triangle inequality, so it cannot be equated with a metric.

Several different divergences have been defined. Most of them have beneficial properties for our data analysis (e.g. one divergence is more sensitive to the mean value of the distributions, and the other divergence is more sensitive to the variance of the distributions).

In the next section our main focus will be Renyi divergence, however some other example of divergences are listed below (without proof that these are indeed divergences).

Example:

1. ***Kullback-Leibler divergence**:*

$$D_{KL}(p\|q) = \int_{\Omega} p(x) \cdot \log\left(\frac{p(x)}{q(x)}\right) dx, \tag{1}$$

where p and q are probability density functions with support Ω.

2. ***f-divergence**: This is a family of divergences generated by the function f, such that:*

   - *f is convex on $\mathbb{R}^+$,*
   - *$f(1) = 0$.*

   The elements of this family are:

$$D_f(p\|q) = \int_{\Omega} p(x) \cdot f\left(\frac{p(x)}{q(x)}\right) dx, \tag{2}$$

   where p and q are probability density functions with support Ω.

3. ***Hellinger distance**:*

$$H^2(p,q) = 2\int_{\Omega} \left(\sqrt{p(x)} - \sqrt{q(x)}\right)^2 dx, \tag{3}$$

   where p and q are probability density functions with support Ω.

Only formulas for continuous variables are given, since the formula for discrete variables are analogous.

3 Renyi divergence for continuous variables

When taking a closer look at **Renyi divergence**, we will only limit ourselves to continuous variables, since the definition of discrete is analogous.

Definition 3.1 *Let P and Q be probability distribution with support Ω, p and q probability density function of P and Q, and $\alpha > 0$, $\alpha \neq 1$. Then **Renyi divergence** is defined as:*

$$D_\alpha(P\|Q) = \frac{1}{\alpha - 1} \cdot \log \int_\Omega \Big(p(x)\Big)^\alpha \Big(q(x)\Big)^{1-\alpha} \, dx. \tag{4}$$

Let us prove that Renyi divergence is indeed a divergence.

Proof. Let S be a space of probability density functions. We must prove:

1. $D(p\|q) \geq 0$ for all $p, q \in S$,
2. $D(p\|q) = 0 \Leftrightarrow p = q$.

First, let us prove that Renyi divergence is always positive. Instead of $p(x)$ in $q(x)$ we write p in q. We must prove:

$$\frac{1}{\alpha - 1} \log \int_\Omega p^\alpha q^{1-\alpha} \, dx \geq 0. \tag{5}$$

We distinguish between cases:

- $\alpha > 1$: since the first factor in inequality (5) is positive for $\alpha > 1$, it is equivalently to prove that

  $$\int_\Omega p^\alpha q^{1-\alpha} \, dx \geq 1$$

  or

  $$\int_\Omega \Big(\frac{p}{q}\Big)^\alpha q \, dx \geq 1.$$

  Let's use Jensen's inequality for a convex function ϕ:

  $$\phi(\int f(x)dx) \leq \int (\phi \circ f)(x)dx,$$

where we select a function $\phi(t) = t^{\alpha}$:

$$\int_{\Omega} \left(\frac{p}{q}\right)^{\alpha} q \, dx \geq \left(\int_{\Omega} \frac{p}{q} q dx\right)^{\alpha} = \left(\int p dx\right)^{\alpha} = 1,$$

and where we take into account that $\int_{\Omega} p dx = 1$ by the probability density function definition.

- $0 < \alpha < 1$: since the first factor in inequality (5) is negative for $0 < \alpha < 1$, it is equivalently to prove that

$$0 < \int_{\Omega} p^{\alpha} q^{1-\alpha} \, dx \leq 1$$

or

$$0 < \int_{\Omega} \left(\frac{q}{p}\right)^{1-\alpha} p \, dx \leq 1.$$

Let's use Jensen's inequality for a concave function ϕ:

$$\phi(\int f(x) dx) \geq \int (\phi \circ f)(x) dx,$$

where we select a function $\phi(t) = t^{1-\alpha}$:

$$\int_{\Omega} \left(\frac{q}{p}\right)^{1-\alpha} p \, dx \leq \left(\int_{\Omega} \frac{q}{p} p \, dx\right)^{1-\alpha} = \int_{\Omega} q \, dx = 1,$$

and where we take into account that $\int_{\Omega} p dx = 1$ by the probability density function definition.

Let's prove second condition:

$$\frac{1}{\alpha - 1} \log \int_{\Omega} p^{\alpha} q^{1-\alpha} \, dx = 0 \Leftrightarrow p = q.$$

Let us first prove the implication from right to left ($\Leftarrow$):

$$\frac{1}{\alpha-1}\log\int_\Omega \left(\frac{p}{p}\right)^\alpha p\,dx = \frac{1}{\alpha-1}\log\int_\Omega p\,dx = \frac{1}{\alpha-1}\log 1 = 0.$$

In the other direction we will just do short deliberation. The equation on the left side of equivalence will be true when:

1. $\alpha = 0$, which contradicts the assumption that $\alpha > 0$,
2. $p = q$, because then $\log\int_\Omega p\,dx = \log 1 = 0$.

The last implication is proven superficial, since it could also happen that the left size of the equivalence in point 2 holds if $p \neq q$. We conclude that due to the properties of probability densities, this cannot happen. □

Renyi divergence is not defined in $\alpha = 1$, but we know its limit in this point:

Theorem 3.1 *Let $D_\alpha(P\|Q)$ be Renyi divergence of distributions P and Q. Then the following applies:*

$$\lim_{\alpha\to 1} D_\alpha(P\|Q) = \int_\Omega p(x)\cdot\log\left(\frac{p(x)}{q(x)}\right)dx, \tag{6}$$

*where the expression on the right is exactly **Kullback-Leibler divergence** of distributions P and Q, i.e.*

$$\lim_{\alpha\to 1} D_\alpha(P\|Q) = D_{KL}(P\|Q). \tag{7}$$

Let us prove Theorem 3.1:

Proof. Let's calculate the limit $D_\alpha(P\|Q)$ as α goes to 1.

$$\lim_{\alpha\to 1}\frac{\log\int p(x)^\alpha q(x)^{1-\alpha}\,dx}{\alpha-1} = \text{“}\frac{0}{0}\text{”},$$

so we can use L'Hôpital's rule:

$$\lim_{x\to a}\frac{f(x)}{g(x)} = \lim_{x\to a}\frac{f'(x)}{g'(x)}.$$

We separately calculate the derivatives of the numerator and the denominator. The denominator's derivative is: $(\alpha - 1)' = 1$. Let's calculate the derivative of the numerator:

$$\begin{aligned}\frac{d}{d\alpha}\Big(\log\int p(x)^\alpha q(x)^{1-\alpha}\,dx\Big) &= \frac{1}{\int p(x)^\alpha q(x)^{1-\alpha}\,dx}\quad \frac{d}{d\alpha}\int p(x)^\alpha q(x)^{1-\alpha}\,dx \overset{(*)}{=}\\ &\overset{(*)}{=} \underbrace{\frac{1}{\int p(x)^\alpha q(x)^{1-\alpha}\,dx}}_{\overset{\alpha\to 1}{\longrightarrow} 1}\quad \int \frac{\partial}{\partial\alpha}\Big(p(x)^\alpha q(x)^{1-\alpha}\,dx\Big) =\\ &= \int \Big(p(x)^\alpha \cdot \log p(x)\cdot q(x)^{1-\alpha} - p(x)^\alpha \cdot q(x)^{1-\alpha}\cdot \log q(x)\Big)dx =\\ &= \int p(x)^\alpha \cdot q(x)^{1-\alpha}\cdot \log\frac{p(x)}{q(x)}\;dx,\end{aligned}$$

where at $(*)$ we take into account that $F(\alpha) = p(x)^\alpha \cdot q(x)^{1-\alpha}$ is a continuous function and $\int p(x)^\alpha q(x)^{1-\alpha}\,dx$ goes to 1 as $\alpha \to < 1$ by the probability density definition. If we calculate derivatives quotient limit, we get:

$$\lim_{\alpha\to 1}\frac{\int p(x)^\alpha \cdot q(x)^{1-\alpha}\cdot \log\frac{p(x)}{q(x)}\;dx}{1} = \int p(x)\cdot\log\frac{p(x)}{q(x)}\;dx,$$

which is by definition exactly a Kullback-Leibler divergence. □

4 Renyi divergence of two histograms

Histograms are rough estimations of distributions. We can think of a histogram as a probability density function, so we can use the Renyi divergence for continuous variables on histograms too. But with some adjustments, we can avoid integration, which will accelerate the numerical calculation of the Renyi divergence. To do that, we must find a clever way to represent histograms.

We will use Python way to represent histograms (with one exception: we start counting from 1 rather than 0). So each histogram is a pair (x, y), where x and y are lists and length of x is one more than length of y. x represents edges of histogram bins, and y represents heights of histogram bins. To be able to understand the histogram as a probability density function, it must be normalized.

Let $z[i]$ represents the i-th element in the list z. First, we need to combine bins' edges so that we can compare two histograms by their bins. Let's take a union $x = x_1 \cup x_2$ and sort its elements in increasing order. Next, we cut bins in both histograms, so we get new heights: y_1', y_2'. Both of these lists have length n, one less than length of x. If $\exists i \in \{1, \ldots, n\}$, such that $y_1'[i]$ or $y_2'[i]$ is not defined (there is no bin in first or second histogram such that $x[i]$ lies in this bin) then $y_1' = 0$ or $y_2' = 0$. We get two new histograms (x, y_1') and (x, y_2'). This way we did not change histograms (areas under histograms are still equal to 1), we just broke their bins into multiple pieces.

Now we edit formula (4) to apply to histograms:

$$D_\alpha((x_1, y_1), (x_2, y_2)) = \frac{1}{1-\alpha} \cdot \log\Big(\sum_{i=1}^{n-1} \big(y_1'[i]\big)^\alpha \big(y_2'[i]\big)^\alpha \big(x[i+1] - x[i]\big) \Big), \tag{8}$$

where x is a sorted union of x_1 and x_2, and y_1', y_2' are heights after combining bins of histograms (x_1, y_1) and (x_2, y_2).

5 Numerical problems when calculating Renyi divergence

Certain difficulties occur in numerical calculations. Due to the underflow small numbers absolutely less than numerical resolution constant get rounded to 0. This leads to division with zero or log(0) problems.

First, we list some rules to operate with these problems:

- $\frac{0}{0} = 0$,
- $\frac{a}{0} = \infty$ for $a > 0$,
- $\log\infty = \infty$,
- $\log 0 = -\infty$.

There is another possibly even more complex way to bypass such problems, but one needs to be careful, because probability density functions must be changed a bit.

Let us take any probability densities p, q and expand them to $\mathbb{R}$ ($p(x) = 0,\ \forall x \notin \Omega$, same for q). Let ϵ be numerical resolution constant. Let m be such real number, that:

$$\forall x_0 < m, \forall \delta > 0 : \Big(p(x_0) < \epsilon \ \wedge \ q(x_0) < \epsilon\Big) \wedge \Big(p(m+\delta) \geq \epsilon \ \vee \ q(m+\delta) \geq \epsilon\Big),$$

i.e. m is a number such that probability densities are less than ϵ on $(-\infty, m)$ and m is a maximal such number. We know, that m exists because $\lim_{x\to-\infty} p(x) = 0$ for any probability density p (area under p is equal to 1). Let M be such real number, that:

$$\forall x_0 > M, \forall \delta > 0 : \Big(p(x_0) < \epsilon \ \wedge \ q(x_0) < \epsilon\Big) \wedge \Big(p(M-\delta) \geq \epsilon \ \vee \ q(M-\delta) \geq \epsilon\Big),$$

i.e. M is a number such that probability densities are less than ϵ on (M, ∞) and M is a minimal such number. We know, that M exists because $\lim_{x\to-\infty} p(x) = 0$ for any probability density p.

On $\mathbb{R}-[m,M]$ both probability densities are equal to 0 due to the underflow. Let's define new functions on interval $[m,M]$:

$$f(x) := \begin{cases} p(x) & , \ p(x) > \epsilon \\ \\ \epsilon & , \ p(x) \leq \epsilon \end{cases} \qquad \text{and} \qquad g(x) := \begin{cases} q(x) & , \ q(x) > \epsilon \\ \\ \epsilon & , \ q(x) \leq \epsilon \end{cases} .$$

Functions f and g differ from p and q, where p and q are 0 because of the underflow. We must check, if $\int_{[m,M]} f(x)dx \approx 1$ and $\int_{[m,M]} g(x)dx \approx 1$. This will be true, if the total length of subintervals on $[m,M]$, where $p,q < \epsilon$, are small enough. Now we make an approximation $D_\alpha(p\|q) \approx D_\alpha(f\|g)$, therefore eliminating the zero division problem. However, these values can be very large due to an ϵ division.

Remark: When we can round $\int_{[m,M]} f(x)dx$ to 1 depends on what kind of error is still satisfactory for us. For example, if total length of subintervals on $[m,M]$, where $p < \epsilon$, equals to 1000 and our numerical resolution constant is $\epsilon \approx 2.22\text{e}{-16}$, value of $\int_{[m,M]} f(x)dx$ will differ absolutely from 1 for less or equal to $2.22\text{e}{-13}$.

We can use both approaches for problem solving when calculating Renyi divergence of two histograms. The second approach is a lot simpler when operating with histograms. We simply have to substitute all 0 in y_2' with ϵ. If required, we can do the same thing with y_1' (when calculating Kullback-Leibler divergence).

6 Calculating Renyi divergence of two data samples

Let X and Y be sets of univariate data samples. We want to find the difference between them by using (Renyi) divergence. We need to follow the listed steps:

1. find probability density estimations p and q of samples X and Y,
2. calculate (Renyi) divergence $D_\alpha(p\|q)$.

In the first step we can estimate probability densities by using histograms or other probability density estimators such as Gaussian kernel density estimator. We only describe these two methods, however other estimation methods can apply.

Step 2 has already been resolved in the previous sections.

Histograms

We will not discuss, how we construct histograms from a data sample. After constructing histograms from X and Y, the only thing left to do is to use formula (8), that calculates Renyi divergence of given histograms.

Gaussian kernel density estimation

Firstly, we need to define what a kernel is.

Definition 6.1 *A **kernel** is a non-negative real-valued integrable function K, with additional properties:*

- $\int_{-\infty}^{\infty} K(x)\, dx = 1$ *and*
- $K(-x) = K(x), \quad \forall x \in \mathbb{R}$ *(symmetry).*

Finally, a kernel density estimation can be defined.

Definition 6.2 *A **kernel density estimation** is a non-parametric way to estimate the probability density function. Given kernel K and data-set $X = \{x_1, \ldots, x_n\}$, kernel density estimation of data-set X is defined as the function:*

$$f_h(x) = \frac{1}{n \cdot h} \sum_{i=1}^{n} K\Big(\frac{x - x_i}{h}\Big), \tag{9}$$

where h is a smoothing parameter called the bandwidth.

Gaussian kernel density estimator uses Gaussian kernel (Gaussian probability density function with the mean value 0 and the variance 1), defined as

$$K(x) = \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}x^2}. \tag{10}$$

Without listing the proof, we mention than Gaussian kernel density estimation is a probability density function. In step 1, we define p as Gaussian kernel density estimation of X and q as Gaussian kernel density estimation of Y. For step 2, we calculate Renyi divergence of p and q using formula (4).

We usually also limit integration area, when calculating Renyi divergence via formula (4). Let's define $m = \min\{\min(X), \min(Y)\}$ and $M = \max\{\max(X), \max(Y)\}$. Instead of integrating over the whole support area, we integrate over interval $[m, M]$, however we must be very careful not to cut to much area, which could distort our results. In order to do this, we must have an overview of the data-set.

References

[1] van Erven, T. in Harremoës, P., 2007. Renyi Divergence and Kullback-Leibler Divergence. Accessibility: https://arxiv.org/pdf/1206.2459.pdf