

Non-negative Monte Carlo estimation of f -divergences

Frank Nielsen*

Sony Computer Science Laboratories Inc.

Tokyo, Japan

6th January 2020

Abstract

We show how to guarantee non-negative Monte Carlo estimations of f -divergences by considering the corresponding *extended* f -divergences. We apply the method for estimating non-negatively the Kullback-Leibler divergence and the Jensen-Shannon divergence.

1 Problem with naive Monte Carlo estimations of f -divergences

Let (X, F, μ) be a probability space [5] with X denoting the sample space, F the σ -algebra, and μ a reference positive measure. The f -divergence [3, 6] between two probability measures P and Q both absolutely continuous with respect to μ for a convex generator $f : (0, \infty) \rightarrow \mathbb{R}$ strictly convex at 1 and satisfying $f(1) = 0$ is

$$I_f(P : Q) = I_f(p : q) = \int p(x) f\left(\frac{q(x)}{p(x)}\right) d\mu(x),$$

where $P = p d\mu$ and $Q = q d\mu$ (i.e., p and q are Radon-Nikodym derivatives with respect to μ). We use the following conventions:

$$0f\left(\frac{0}{0}\right) = 0, \quad f(0) = \lim_{u \rightarrow 0^+} f(u), \quad \forall a > 0, 0f\left(\frac{a}{0}\right) = \lim_{u \rightarrow 0^+} uf\left(\frac{a}{u}\right) = a \lim_{u \rightarrow \infty} \frac{f(u)}{u}.$$

When $f(u) = -\log u$, we retrieve the Kullback-Leibler divergence (KLD):

$$D_{\text{KL}}(p : q) = \int p(x) \log \frac{p(x)}{q(x)} d\mu(x).$$

The KLD is usually difficult to calculate in closed-form, say, for example, between statistical mixture models [7]. A common technique is to estimate the KLD using Monte Carlo sampling using a proposal distribution r :

$$\widehat{\text{KL}}_n(p : q) = \frac{1}{n} \sum_{i=1}^n \frac{p(x_i)}{r(x_i)} \log \frac{p(x_i)}{q(x_i)},$$

*E-mail: Frank.Nielsen@acm.org. <https://franknielsen.github.io/>

where $x_1, \dots, x_n \sim_{\text{iid}} r$. When r is chosen as p , the KLD can be estimated as

$$\widehat{\text{KL}}_n(p : q) = \frac{1}{n} \sum_{i=1}^n \log \frac{p(x_i)}{q(x_i)}. \quad (1)$$

Monte Carlo estimators are consistent under mild conditions: $\lim_{n \rightarrow \infty} \widehat{\text{KL}}_n(p : q) = \text{KL}(p : q)$.

In practice, one problem when implementing Eq. 1, is that we may end up potentially with $\widehat{\text{KL}}_n(p : q) < 0$. This may have disastrous consequences as algorithms implemented by programs consider non-negative divergences to execute a correct workflow. The potential negative value problem of Eq. 1 comes from the fact that $\sum_i p(x_i) \neq 1$ and $\sum_i q(x_i) \neq 1$.

2 Non-negative Monte Carlo estimates from extended f -divergences

A f -divergence is defined for a convex generator $f(u)$ with $f(u) = 1$, strictly convex at 1 (hence $f'(1)$ exists). The non-negativeness of f -divergences follow from the Jensen's inequality:

$$I_f(p : q) = \int p(x) f\left(\frac{q(x)}{p(x)}\right) d\mu(x) \geq f\left(\int p(x) \frac{q(x)}{p(x)} d\mu(x)\right) = f(1) = 0.$$

Two f -divergences coincide, i.e. $I_f(p : q) = I_g(p : q)$, iff there exists a real λ such that $g(u) = f(u) + \lambda(u - 1)$. In particular, we can choose the following equivalent generator

$$g(u) = f(u) - (u - 1)f'(1) = f(u) - f(1) - (u - 1)f'(1) =: B_f(u : 1), \quad (2)$$

where $B_f(a : b)$ is a *scalar Bregman divergence* [2]:

$$B_f(a : b) = f(a) - f(b) - (a - b)f'(b) \geq 0. \quad (3)$$

Bregman divergences are always non-negative and equal to zero iff $a = b$.

Thus we given an alternative proof of the Gibb's inequality of f -divergences as:

$$I_f(p : q) = I_g(p : q) = \int p(x) \left(f\left(\frac{q(x)}{p(x)}\right) - f'(1) \left(\frac{q(x)}{p(x)} - 1\right) \right) d\mu(x) \quad (4)$$

$$= \int p(x) \underbrace{B_f\left(\frac{q(x)}{p(x)} : 1\right)}_{\geq 0} d\mu(x) \geq 0. \quad (5)$$

One way to circumvent this negative Monte Carlo estimation problem is to consider the extended f -divergences:

Definition 1 (Extended f -divergence) *The extended f -divergence for a convex generator f , strictly convex at 1 and satisfying $f(1) = 0$ is defined by*

$$I_f^e(p : q) = \int p(x) \left(f\left(\frac{q(x)}{p(x)}\right) - f'(1) \left(\frac{q(x)}{p(x)} - 1\right) \right) d\mu(x).$$

Setting $a = \frac{q(x)}{p(x)}$ and $b = 1$ in Eq. 3, and using the fact that $f(1) = 0$, we get

$$f\left(\frac{q(x)}{p(x)}\right) - \left(\frac{q(x)}{p(x)} - 1\right) f'(1) \geq 0.$$

Therefore we define the *extended f -divergences* as

$$\boxed{I_f^e(p : q) = \int p(x) B_f\left(\frac{q(x)}{p(x)} : 1\right) d\mu(x) \geq 0.} \quad (6)$$

That is, the formula for the extended f -divergences is

$$I_f^e(p : q) = \int p(x) \left(f\left(\frac{q(x)}{p(x)}\right) - f'(1) \left(\frac{q(x)}{p(x)} - 1\right) \right) d\mu(x) \geq 0. \quad (7)$$

Then we estimate the extended f -divergence using importance sampling of the integral with respect to distribution r , using n variates $x_1, \dots, x_n \sim_{\text{iid}} p$ as:

$$\boxed{\hat{I}_{f,n}(p : q) = \frac{1}{n} \sum_{i=1}^n f\left(\frac{q(x_i)}{p(x_i)}\right) - f'(1) \left(\frac{q(x_i)}{p(x_i)} - 1\right) \geq 0.}$$

For example, for the KLD, we obtain the following Monte Carlo estimator:

$$\widehat{\text{KL}}_n(p : q) = \frac{1}{n} \sum_{i=1}^n \left(\log \frac{p(x_i)}{q(x_i)} + \frac{q(x_i)}{p(x_i)} - 1 \right) \geq 0, \quad (8)$$

since the extended KLD is

$$D_{\text{KL}}^e(p : q) = \int \left(p(x) \log \frac{p(x)}{q(x)} + q(x) - p(x) \right) d\mu(x).$$

Eq. 8 can be interpreted as a sum of scalar Itakura-Saito divergences since the Itakura-Saito divergence is scale-invariant: $\widehat{\text{KL}}_n(p : q) = \frac{1}{n} \sum_{i=1}^n D_{\text{IS}}(p(x_i) : q(x_i))$ with the scalar Itakura-Saito divergence

$$D_{\text{IS}}(a : b) = D_{\text{IS}}\left(\frac{a}{b} : 1\right) = \frac{a}{b} - \log \frac{a}{b} - 1 \geq 0,$$

a Bregman divergence obtained for the generator $f(u) = -\log u$.

Notice that the extended f -divergence is a f -divergence for the generator

$$f_e(u) = f(u) - f'(1)(u - 1).$$

We check that the generator f_e satisfies both $f(1) = 0$ and $f'(1) = 0$, and we have $I_f^e(p : q) = I_{f_e}(p : q)$. Thus $D_{\text{KL}}^e(p : q) = I_{f_{\text{KL}}}^e(p : q)$ with $f_{\text{KL}}^e(u) = -\log u + u - 1$.

Let us remark that we only need to have the scalar function strictly convex at 1 to ensure that $B_f\left(\frac{a}{b} : 1\right) \geq 0$. Indeed, we may use the definition of Bregman divergences extended to strictly convex functions but not necessarily smooth functions [4, 8]:

$$B_f(x : y) = \max_{g(y) \in \partial f(y)} \{f(x) - f(y) - (x - y)g(y)\},$$

where $\partial f(y)$ denotes the subderivative of f at y .

As a working example, consider the Jensen-Shannon divergence (bounded divergence which does not require matching supports of distributions):

$$\text{JS}[p : q] := \frac{1}{2} \text{KL} \left[p : \frac{p+q}{2} \right] + \frac{1}{2} \text{KL} \left[q : \frac{p+q}{2} \right]$$

- A first estimation consists in estimating the KLDs separately:

$$\widehat{\text{JS}}_{n_1, n_2}[p : q] := \frac{1}{2n_1} \sum_{i=1}^{n_1} \left(\log \left(\frac{2p(x_i)}{p(x_i) + q(x_i)} \right) + \frac{q(x_i) - p(x_i)}{2p(x_i)} \right) \quad (9)$$

$$\begin{aligned} &+ \frac{1}{2n_2} \sum_{i=1}^{n_2} \left(\log \left(\frac{2q(y_i)}{p(y_i) + q(y_i)} \right) + \frac{p(y_i) - q(y_i)}{2q(y_i)} \right) \\ &= -\frac{1}{2} + \frac{1}{2n_1} \sum_{i=1}^{n_1} \left(\log \left(\frac{2p(x_i)}{p(x_i) + q(x_i)} \right) + \frac{q(x_i)}{2p(x_i)} \right) \\ &+ \frac{1}{2n_2} \sum_{i=1}^{n_2} \left(\log \left(\frac{2q(y_i)}{p(y_i) + q(y_i)} \right) + \frac{p(y_i)}{2q(y_i)} \right), \end{aligned} \quad (10)$$

with $x_1, \dots, x_{n_1} \sim_{\text{iid}} p(x)$ and $y_1, \dots, y_{n_2} \sim_{\text{iid}} q(x)$.

- A second estimation consists in expressing it as a f -divergence for the generator:

$$f_{\text{JS}}(u) := -\frac{1+u}{2} \log \left(\frac{1+u}{2} \right) + \frac{u}{2} \log(u).$$

Indeed, we check that

$$\begin{aligned} I_{f_{\text{JS}}}(p : q) &:= \int p f \left(\frac{q}{p} \right) d\mu, \\ &= \frac{1}{2} \int \left((p+q) \log \frac{2p}{p+q} + q \log \frac{q}{p} \right) d\mu, \\ &= \frac{1}{2} \int p \log \frac{p}{m} d\mu + \frac{1}{2} \int q \log \frac{p}{m} \frac{q}{p} d\mu, \\ &= \frac{1}{2} \int p \log \frac{p}{m} d\mu + \frac{1}{2} \int q \log \frac{q}{m} d\mu, \\ &= \frac{1}{2} \text{KL} \left[p : \frac{p+q}{2} \right] + \frac{1}{2} \text{KL} \left[q : \frac{p+q}{2} \right], \\ &=: \text{JS}(p : q). \end{aligned}$$

Since we have $f'_{\text{JS}}(1) = 0$, we get the simplified non-negative f -divergence estimation formula:

$$\hat{I}_{f,n}(p : q) = \frac{1}{n} \sum_{i=1}^n f \left(\frac{q(x_i)}{p(x_i)} \right) \geq 0.$$

with $x_1, \dots, x_n \sim p$.

- A third method consists in using the extended KLD:

$$\text{JS}_e[p : q] := \frac{1}{2} \text{KL}_e \left[p : \frac{p+q}{2} \right] + \frac{1}{2} \text{KL}_e \left[q : \frac{p+q}{2} \right],$$

where $\text{KL}_e[p : q] = \text{KL}[p : q] + \int (q - p) d\mu$.

Then we estimate JS_e using stochastic Monte Carlo integration using the proposal distribution $r = m = \frac{p+q}{2}$. It follows that

$$\text{JS}_{e,n}[p : q] := \frac{1}{2n} \sum_{i=1}^n \left(\frac{2p(x_i)}{p(x_i) + q(x_i)} \log \frac{2p(x_i)}{p(x_i) + q(x_i)} + \frac{2q(x_i)}{p(x_i) + q(x_i)} \log \frac{2q(x_i)}{p(x_i) + q(x_i)} \right) \geq 0, \quad (11)$$

where $x_1, \dots, x_n \sim_{\text{iid}} \frac{p+q}{2}$ (sampled from the mixture distribution of p and q).

Note that if we define a divergence by $R_f(p : q) = \int p(x) B_f \left(1 : \frac{q(x)}{p(x)} \right) d\mu(x) \geq 0$ for f strictly convex everywhere (but we do not require $f'(1) = 0$ here), and expand the formula, we end up with the following inequality for the f -divergence:

$$I_f(p : q) \leq \int (q(x) - p(x)) f' \left(\frac{q(x)}{p(x)} \right) d\mu(x).$$

In particular, we find that $\text{KL}(p : q) \leq \int (p - q) \frac{p}{q} d\mu = \int \frac{p^2}{q} d\mu - 1$.

3 Conclusion

The f -divergence $I_f(p : q) = \int p(x) f \left(\frac{q(x)}{p(x)} \right) d\mu(x)$ is defined for a convex generator satisfying $f(1) = 0$ since it follows from Jensen inequality that $I_f(p : q) \geq f \left(\int p(x) \frac{q(x)}{p(x)} d\mu(x) \right) = f(1) = 0$. For densities, the generator f is equivalent to the family of generators $f_\lambda(u) = f(u) + \lambda(u - 1)$ where $\lambda \in \mathbb{R}$: $I_f(p : q) = I_{f_\lambda}(p : q)$. We showed that we can express the f -divergence as a scaled integral of a scalar Bregman divergence: $I_f(p : q) = \int p(x) B_f \left(\frac{q(x)}{p(x)} : 1 \right) d\mu(x)$ provided that $f'(1) = 0$. This can always be done by choosing the equivalent generator f_λ such that $f'_\lambda(1) = f'(1) + \lambda = 0$, i.e. $\lambda = -f'(1)$. It follows that in order to have the f -divergences satisfying the law of the indiscernibles, we need to have strict convexity of f at 1. Expressing the f -divergence using a Bregman divergence allows one to

1. calculate non-negative Monte Carlo estimates $\hat{I}_f(p : q) = \frac{1}{s} \sum_{i=1}^s \frac{p(x_i)}{r(x_i)} B_f \left(\frac{q(x_i)}{p(x_i)} : 1 \right) \geq 0$ where $x_1, \dots, x_s \sim_{\text{iid}} r$, a proposal distribution, and
2. extend the f -divergences to positive densities.

Furthermore, noticing that $I_{\lambda f}(p : q) = \lambda I_f(p : q)$ for $\lambda > 0$, we may enforce that $f''(1) = 1$, and obtain a *standard f -divergence* [1] which enjoys the property that $I_f(p_\theta(x) : p_{\theta+d\theta}(x)) = d\theta^\top I(\theta) d\theta$, where $I(\theta)$ denotes the Fisher information matrix of the parametric family $\{p_\theta\}_\theta$ of densities.

References

- [1] S. Amari. *Information Geometry and Its Applications*. Applied Mathematical Sciences. Springer Japan, 2016.
- [2] Lev M Bregman. The relaxation method of finding the common point of convex sets and its application to the solution of problems in convex programming. *USSR computational mathematics and mathematical physics*, 7(3):200–217, 1967.
- [3] Imre Csiszár. Information-type measures of difference of probability distributions and indirect observation. *studia scientiarum Mathematicarum Hungarica*, 2:229–318, 1967.
- [4] G. J. Gordon. *Approximate solutions to Markov decision processes*. PhD thesis, Department of Computer Science, Carnegie Mellon University, 1999.
- [5] Robert W Keener. *Theoretical statistics: Topics for a core course*. Springer, 2011.
- [6] Frank Nielsen and Richard Nock. On the chi square and higher-order chi distances for approximating f -divergences. *IEEE Signal Processing Letters*, 21(1):10–13, 2013.
- [7] Frank Nielsen and Ke Sun. Guaranteed bounds on the Kullback–Leibler divergence of univariate mixtures. *IEEE Signal Processing Letters*, 23(11):1543–1546, 2016.
- [8] Matus Telgarsky and Sanjoy Dasgupta. Agglomerative Bregman clustering. In *Proceedings of the 29th International Conference on International Conference on Machine Learning*, pages 1011–1018. Omnipress, 2012.