Applications of the f-divergence to variational inference

Sabin Roman [1] **Alexandra Peşte** [1]

Divergences are central object of study within information geometry (Amari, 2016). A widely studied class of divergences are the f-divergences (Csiszár, 2008). The f-divergence between two distributions p and q is defined as (Amari & Cichocki, 2010):

$$D_f[q:p] = \int q(z) f\left(\frac{p(z)}{q(z)}\right) dz \tag{1}$$

where we choose f convex with $f(1) = 0$. Specific examples of f-divergences have also found applications in machine learning, particularly in variational inference (Hoffman et al., 2013; Ranganath et al., 2014; Kingma & Welling, 2014; Rezende et al., 2014).

The most commonly used example of an f-divergence is the Kullback-Liebler (KL) divergence which, in the case of variational inference, is used in deriving the evidence bound:

$$\begin{aligned}\mathcal{L}(x;q) &= \log p(x) - KL[q(z|x)||p(z|x)] \\ &= E_q\left[\log \frac{p(x,z)}{q(z|x)}\right]\end{aligned} \tag{2}$$

where $p(x)$ is the likelihood we want to estimate through the bound, $p(x,z)$ is the joint likelihood of the data x with the latent variable z, $p(z|x)$ is the true posterior and $q(z|x)$ is the approximate posterior.

There have been several attempts at generalizing the above framework to bounds derived from other divergence functions. In particular, (Dieng et al., 2017) aims to perform variational inference employing a variational bound derived from the χ^2-divergence between the true and approximate posterior, while (Hernandez-Lobato et al., 2016) derives a bound from an α-divergence. In addition to these developments, variational auto-encoders (VAE) have in recent years become one of the main tools for performing simultaneous inference and generative aspects of Bayesian learning (Kingma & Welling, 2014). (Li & Turner, 2016) adapts VAE to be used with the Renyi divergence (which is closely related to an α-divergence). Here we aim to generalize the treatment to an arbitrary f-divergence and propose a framework to perform f-likelihood optimization. This provides a unified framework for variational inference and its use in conjuncture with auto-encoders.

To generalize the treatment of (Kingma & Welling, 2014), (Dieng et al., 2017) and (Hernandez-Lobato et al., 2016) for a general f-divergence we propose the following variational bound:

$$E_q\left[f\left(\frac{p(x,z)}{q(z|x)}\right)\right] \tag{3}$$

where f is convex and $f(1) = 0$.

Theorem 1. *If $p(x)$ is the likelihood and $q(z|x)$ is the approximate posterior, then:*

$$f(p(x)) \leq E_q\left[f\left(\frac{p(x,z)}{q(z|x)}\right)\right] \tag{4}$$

Furthermore, the following identity holds:

$$\begin{aligned}&E_q\left[f\left(\frac{p(x,z)}{q(z|x)}\right)\right] - E_q\left[f\left(\frac{p(z|x)}{q(z|x)}\right)\right] = f(p(x)) \\ &+ \sum_{n=2} \frac{p(x)^n f^{(n)}(p(x)) - f^{(n)}(1)}{n!} \int q(z|x)\left(\frac{p(z|x)}{q(z|x)} - 1\right)^n dz\end{aligned} \tag{5}$$

Proof. In equation (4) the inequality follows from Jensen's inequality. Let $y = \frac{p(z|x)}{q(z|x)}$ and $g(y) = f(p(x)y)$, then $g(1) = f(p(x))$ and the n-th order derivative is $g^{(n)}(y) = f^{(n)}(p(x)y)p(x)^n$. We Taylor expand $g(y)$ and $f(y)$ around 1, substitute in the left hand side of (5) and readily obtain the desired result. □

Equation (4) generalizes the evidence bound used in variational inference and (5) generalizes identity (2) for an arbitrary f-divergence. A common variant of the variational bound (3) is obtained by applying the $\log$, as done in (Hernandez-Lobato et al., 2016; Li & Turner, 2016; Dieng et al., 2017). For specific choices of f-divergences there is greater analytic tractability and lower and upper bounds can be derived for $\log p(x)$.

By minimizing the variational bounds in equation (3) we can perform both likelihood maximization and variational inference. This allows for the generalization and unification of the procedures used throughout the literature.

[1]DeepRiemann project, Romanian Institute of Science and Technology, Cluj-Napoca, Romania. Correspondence to: Sabin Roman <roman@rist.ro>, Alexandra Peşte <peste@rist.ro>.

Acknowledgements

We would like to thank Septimia Sârbu, Luigi Malagò and the Romanian Institute for Science and Technology (RIST).

References

Amari, Shun-ichi. *Information geometry and its applications*. Springer, 2016.

Amari, Shun-ichi and Cichocki, Andrzej. Information geometry of divergence functions. *Bulletin of the Polish Academy of Sciences: Technical Sciences*, 58(1):183–195, 2010.

Csiszár, Imre. Axiomatic characterizations of information measures. *Entropy*, 10(3):261–273, 2008.

Dieng, Adji Bousso, Tran, Dustin, Ranganath, Rajesh, Paisley, John, and Blei, David. Variational inference via χ upper bound minimization. In *Advances in Neural Information Processing Systems*, pp. 2729–2738, 2017.

Hernandez-Lobato, Jose, Li, Yingzhen, Rowland, Mark, Bui, Thang, Hernandez-Lobato, Daniel, and Turner, Richard. Black-box α-divergence minimization. In *International Conference on Machine Learning*, pp. 1511–1520, 2016.

Hoffman, Matthew D, Blei, David M, Wang, Chong, and Paisley, John. Stochastic variational inference. *The Journal of Machine Learning Research*, 14(1):1303–1347, 2013.

Kingma, Diederik P and Welling, Max. Auto-encoding variational bayes. In *Proceedings of the International Conference on Learning Representations (ICLR)*, 2014.

Li, Yingzhen and Turner, Richard E. Rényi divergence variational inference. In *Advances in Neural Information Processing Systems*, pp. 1073–1081, 2016.

Ranganath, Rajesh, Gerrish, Sean, and Blei, David. Black box variational inference. In *Artificial Intelligence and Statistics*, pp. 814–822, 2014.

Rezende, Danilo Jimenez, Mohamed, Shakir, and Wierstra, Daan. Stochastic backpropagation and approximate inference in deep generative models. In *International Conference on Machine Learning*, pp. 1278–1286, 2014.