f-Divergence Variational Inference

Skorik Sergey

MIPT, 2022

April 11, 2023

Motivation

Introduction

Let consider the **Bayesian inference**

$$p(z|x) = \frac{p(z,x)}{p(x)}$$

MCMC algorithms estimate the evidence $p(x) = \int p(z,x)dx$ via sampling. However, since sampling tends to be a slow and computationally intensive process, VI becomes a good alternative to perform Bayesian inference. Let's denote a family of approximate densities $\mathcal{Q}$. The VI problem is to find the member $q^*(z) \in \mathcal{Q}$ that minimizes the distance between the true posterior $D(q(z)||p(z|x))$. This distance called **divergence**.

Some examples of divergences:

1. Kullback-Leibler: $\int p(x)\frac{p(x)}{q(x)}dx$
2. Pearson χ^2: $\int \frac{(q(x)-p(x))^2}{p(x)}dx$

f-divergence

Definition 1 *The f-divergence from probability density functions $q(z)$ to $p(z)$ is defined as*

$$D_f(q(z)||p(z)) =: \int f\left(\frac{q(z)}{p(z)}\right)p(z)dz = \mathbb{E}_p\left[f\left(\frac{q(z)}{p(z)}\right)\right], \qquad (1)$$

where $f(\cdot)$ is a convex function with $f(1) = 0$.

Table: Divergences $D_f(q||p)$

Divergences	$f(t)$
KL divergecne	$t \log t$
General χ^n-divergence	$t^n - 1,\ n \in \mathbb{R}\backslash(0,1)$
Hellinger α-divergence $\mathcal{H}_\alpha$	$(t^\alpha - 1)/(\alpha - 1),\ \alpha \in \mathbb{R}^+\backslash\{1\}$

f-divergence

Definition 2 *Given a function* $f : (0,1) \to \mathbb{R}$, *the dual function* $f^* : (0,1) \to \mathbb{R}$ *is defined as*

$$f^*(t) = t \cdot f(1/t)$$

Properties:

1. $(f^*)^* = f$
2. if f is convex, f^* is also convex
3. if $f(1) = 0$, then $f^*(1) = 0$
4. With dual function f^* an identity between the forward and reverse f-divergences can be established

$$D_{f^*}(p||q) = \int \frac{p(z)}{q(z)} f\left(\frac{q(z)}{p(z)}\right) q(z)dz = D_f(q||p)$$

f-divergence

In order to facilitate the derivation of f-variational bound, especially when the latent variable model is involved, we introduce a *surrogate f-divergence* D_{f_λ} defined by the *generator function*

$$f_\lambda(\cdot) = f(\lambda \cdot) - f(\lambda) \tag{2}$$

where $\lambda \geqslant 0$ is a constant.

Proposition

Proposition 1: Given two probability distributions q and p, a convergent sequence $\lim_{n\to\infty} \lambda_n = 1$, $\lambda_n \geqslant 0$ and a convex function $f : (0, +\infty) \to \mathbb{R}$ such that $f(1) = 0$ and $f(\cdot)$ is uniformly continuous, the f-divergences between q and p satisfy

$$D_{f_{\lambda_n}}(q||p) \to D_f(q||p) \tag{3}$$

almost everywhere as $n \to \infty$

f-divergence

Proof

Proof:

$$\lim_{n\to\infty} D_{f_{\lambda_n}}(q||p) = \lim_{n\to\infty}\int p(z)\left[f\left(\lambda_n\frac{q(z)}{p(z)}\right) - f(\lambda_n)\right]dz =$$

$$= \lim_{n\to\infty}\int p(z)f\left(\lambda_n\frac{q(z)}{p(z)}\right)dz - \lim_{n\to\infty} f(\lambda_n)\int p(z)dz =$$

$$= \int \lim_{n\to\infty} p(z)f\left(\lambda_n\frac{q(z)}{p(z)}\right)dz = D_f(q||p)$$

Shifted homogeneity

We then introduce a class of f-functions equipped with a structural advantage in decomposition, which will be invoked later to derive the coordinate-wise VI algorithm under mean-field assumption.

Definition 3 A convex function f belongs to $\mathcal{F}_{\{0,1\}}$, if $f(1) = 0$ and for any $t, \tilde{t} \in \mathbb{R}$ we have

$$f(t\tilde{t}) = t^{\gamma} f(\tilde{t}) + f(t)\tilde{t}^{\eta}, \tag{4}$$

where $\gamma \in \mathbb{R}$ and $\eta \in \{0, 1\}$. Function f is type 0 shifted homogeneous or $f \in \mathcal{F}_0$ if $\eta = 0$, and type 1 shifted homogeneous or $f \in \mathcal{F}_1$ if $\eta = 1$.

Shifted homogeneity

Propositions

The duality property between F0 and F1 is stated in Proposition 2.
Proposition 2 Given $f_0 \in \mathcal{F}_0$ and $f_1 \in \mathcal{F}_1$, the dual functions $f_0^* \in \mathcal{F}_1$ and $f_1^* \in \mathcal{F}_0$.
When $f \in \mathcal{F}_{\{0,1\}}$, we can establish a more profound relationship, in contrast with Proposition 1, between f-divergence D_f and surrogate divergence D_{f_λ}
Proposition 3 When $f \in \mathcal{F}_{\{0,1\}}$ and $\lambda > 0$, an f-divergence D_f and its surrogate divergence D_{f_λ} satisfy

$$D_{f_\lambda}(q||p) = \lambda^\gamma D_f(q||p) \tag{5}$$

Shifted homogeneity

Proofs

Proof of Proposition 2. Let $f_0 \in \mathcal{F}_0$. Since

$$f^*(t\tilde{t}) = t\tilde{t} \cdot f\left(\frac{1}{t\tilde{t}}\right) = t\tilde{t}\left[\left(\frac{1}{t}\right)^{\gamma_0} \cdot f\left(\frac{1}{\tilde{t}}\right) + f\left(\frac{1}{t}\right)\right] =$$

$$= t^{1-\gamma_0} \cdot f_0^*(\tilde{t}) + f_0^*(t) \cdot \tilde{t}$$

by letting $\gamma = 1 - \gamma_0$ we can conclude that $f_0^* \in \mathcal{F}_1$. Case $f_1 \in \mathcal{F}_1$ proved analogous.

Proof of Proposition 3. We start this proof by substituting (1), (2) and (4) into the LHS of (5)

$$D_{f_\lambda}(q||p) = \mathbb{E}_p[f_\lambda(q/p)] = \mathbb{E}_p[f(\lambda q/p)] - f(\lambda) =$$

$$= \lambda^\gamma \mathbb{E}_p[f(q/p)] + f(\lambda)\mathbb{E}_p[(q/p)^\eta] - f(\lambda) = \lambda^\gamma D_f(q||p)$$

f-variational bounds

Given a convex function f such that $f(1) = 0$ and a set of i.i.d. samples $\mathcal{D} = \{x^{(n)}\}_{n=1}^{N}$, the generator function $f_{p(\mathcal{D})^{-1}}$ with $p(\mathcal{D}) > 0$ can induce a surrogate f-divergence.

$$D_{f_{p(\mathcal{D})^{-1}}}(q(z)||p(z|\mathcal{D})) = \frac{1}{p(\mathcal{D})}\mathbb{E}_{q(z)}\left[f^*\left(\frac{p(z,\mathcal{D})}{q(z)}\right)\right] - f\left(\frac{1}{p(\mathcal{D})}\right) \quad (6)$$

Multiplying both sides of (6) by $p(\mathcal{D})$ and with rearrangements, we have

$$\mathcal{L}_f(q,\mathcal{D}) = \mathbb{E}_{q(z)}\left[f^*\left(\frac{p(z,\mathcal{D})}{q(z)}\right)\right] = f^*(p(\mathcal{D})) + p(\mathcal{D}) \cdot D_{f_{p(\mathcal{D})^{-1}}}(q(z)||p(z|\mathcal{D})) \quad (7)$$

f-variational bounds

Theorem 1 Dual function of evidence $f^*(p(\mathcal{D}))$ is bounded above by f-variational bound $\mathcal{L}_f(q, \mathcal{D})$

$$\mathcal{L}_f(q, \mathcal{D}) = \mathbb{E}_{q(z)}\left[f^*\left(\frac{p(z, \mathcal{D})}{q(z)}\right)\right] \geqslant f^*(p(\mathcal{D})), \tag{8}$$

Examples:

1. KL-divergence: $f(t) = t \log t \Rightarrow f^*(t) = -\log t$ which is convex and decreasing.

$$\log p(\mathcal{D}) \geqslant \mathbb{E}_{q(z)}[\log p(z, \mathcal{D})] - \mathbb{E}_{q(z)}[\log q(z)] = ELBO$$

2. χ^2-divergence. $f(t) = t^{-1} - t \Rightarrow f^*(t) = t^2 - 1$, so

$$\mathbb{E}_{q(z)}\left[\left(\frac{p(z, \mathcal{D})}{q(z)}\right)^2 - 1\right] \geqslant p(x)^2 - 1$$

Importance-weighted VI

Corollary 1 When $1 \leqslant L_1 \leqslant L_2$, the importance-weighted f-variational bound $\mathcal{L}_f^{IW}(q, \mathcal{D}, L)$ and the f-variational bound $\mathcal{L}_f(q, \mathcal{D})$ satisfy

$$\mathcal{L}_f(q, \mathcal{D}) \geqslant \mathcal{L}_f^{IW}(q, \mathcal{D}, L_1) \geqslant \mathcal{L}_f^{IW}(q, \mathcal{D}, L_2) \xrightarrow{L \to \infty} f^*(p(\mathcal{D}))$$

where $\mathcal{L}_f^{IW}(q, \mathcal{D}, L)$ is defined as

$$\mathcal{L}_f^{IW}(q, \mathcal{D}, L) = \mathbb{E}_{z_{1:L} \sim q(z)} \left[f^* \left(\frac{1}{L} \sum_{l=1}^{L} \frac{p(z_l, \mathcal{D})}{q(z_l)} \right) \right]$$

, and $z_{1:L} = \{z_l\}_{l=1}^{L}$ are $L \in \mathbb{N}$ i.i.d. samples from $q(z)$.

Sandwich formula

After composing both sides of (8) with the inverse dual function $(f^*)^{-1}$, we have the following observations:

1. When the dual function f^* is increasing (or non-decreasing) on $\mathbb{R}^+$, the composition gives an evidence upper bound:

$$(f^*)^{-1} \circ \mathcal{L}_f(q, \mathcal{D}) \geqslant p(\mathcal{D})$$

2. When the dual function f^* is decreasing (or non-increasing) on $\mathbb{R}^+$, the composition gives an evidence lower bound:

$$(f^*)^{-1} \circ \mathcal{L}_f(q, \mathcal{D}) \leqslant p(\mathcal{D})$$

3. When the dual function f^* is non-monotonic on $\mathbb{R}^+$, the composition gives a local evidence bound by applying previous two observations on a monotonic interval f^*:

$$(f^*)^{-1} \circ \mathcal{L}_f(q, \mathcal{D}) \geqslant p(\mathcal{D})$$

Sandwich formula

Based on these observations, we can readily imply a sandwich formula for evidence $p(\mathcal{D})$, which is essential for accurate VI.

Corollary 2 Given convex functions f and g such that $f(1) = g(1) = 0$ on an interval where f^* is increasing and g^* is decreasing the evidence $p(\mathcal{D})$ satisfy

$$(g^*)^{-1} \circ \mathbb{E}_{q(z)} \left[g^* \left(\frac{p(z,\mathcal{D})}{q(z)} \right) \right] \leqslant p(\mathcal{D}) \leqslant (f^*)^{-1} \circ \mathbb{E}_{q(z)} \left[f^* \left(\frac{p(z,\mathcal{D})}{q(z)} \right) \right]. \tag{9}$$

Stochastic optimization

An intuitive approach to apply stochastic optimization is to compute the standard gradient of $\mathcal{L}_f(q, \mathcal{D})$ or $\mathcal{L}_f^{IW}(q, \mathcal{D})$ w.r.t. θ

$$\nabla_\theta \mathcal{L}_f(q_\theta, \mathcal{D}) = \mathbb{E}_{q_\theta(z)}\left[f'\left(\frac{q_\theta(z)}{p(z, \mathcal{D})}\right) \cdot \nabla_\theta \log q_\theta(z)\right], \tag{10}$$

where $f'(t)$ denotes $\partial f(t)/\partial t$.
An unbiased Monte Carlo (MC) estimator for (10) can be obtained by drawing $z_1, z_2, \ldots, z_K$ from $q_\theta(z)$ and

$$\nabla_\theta \hat{\mathcal{L}}_f(q_\theta, \mathcal{D}) = \frac{1}{K}\sum_{k=1}^{K}\left[f'\left(\frac{q_\theta(z_k)}{p(z_k, \mathcal{D})}\right) \cdot \nabla_\theta \log q_\theta(z_k)\right] \tag{11}$$

Stochastic optimization

An alternative to the score function gradient is the reparameterization gradient, which empirically has a lower estimation variance. Let $\varepsilon \sim \mathcal{N}(0, 1)$ and $z = g_\theta(\varepsilon) = \mu + \Sigma^{\frac{1}{2}}\varepsilon$.
The gradient of $\mathcal{L}_f(q, \mathcal{D})$ after reparameterization becomes

$$\nabla_\theta \mathcal{L}_f^{rep}(q_\theta, \mathcal{D}) = \nabla_\theta \mathbb{E}_{p(\varepsilon)} \left[f^* \left(\frac{p(g_\theta(\varepsilon), \mathcal{D})}{q_\theta(g_\theta(\varepsilon))} \right) \right]. \tag{12}$$

An unbiased MC estimator for (12) is

$$\nabla_\theta \hat{\mathcal{L}}_f^{rep}(q_\theta, \mathcal{D}) = \frac{1}{K} \sum_{k=1}^{K} \left[\nabla_\theta f^* \left(\frac{p(g_\theta(\varepsilon_k), \mathcal{D})}{q_\theta(g_\theta(\varepsilon_k))} \right) \right], \tag{13}$$

where $\varepsilon_1, \varepsilon_2, \ldots, \varepsilon_K$ are drawn from $p(\varepsilon)$

Synthetic example

Let $x = sin(z) + \mathcal{N}(0, 0.01)$, $z \sim U[0, \pi]$,
$p(z) = U[0, \pi] \Rightarrow p(x|z) = \mathcal{N}(sin(z), 0.01)$ and $q_\theta(z) = U\left[\frac{1-\theta}{2}\pi, \frac{1+\theta}{2}\pi\right]$.
$\theta_0 = 1.5$ is fixed.

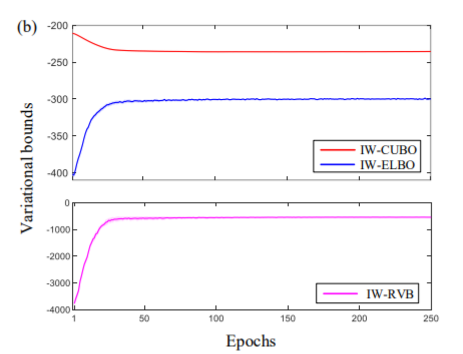

Figure: f-variational bounds on synthetic data.

Bayesian neural network

Table 2: Average test error and negative log likelihood.

Dataset	Test RMSE (lower is better)				Test negative log-likelihood (lower is better)			
	KL-VI	χ-VI	α-VI	f_{c1}-VI	KL-VI	χ-VI	α-VI	f_{c1}-VI
Airfoil	**2.16±.07**	2.36±.14	2.30±.08	2.34±.09	**2.17±.03**	2.27±.03	2.26±.02	2.29±.02
Aquatic	**1.12±.06**	1.20±.06	1.14±.07	1.14±.06	**1.54±.04**	1.60±.08	1.54±.07	1.54±.06
Boston	**2.76±.36**	2.99±.37	2.86±.36	2.87±.36	2.49±.08	2.54±.18	**2.48±.13**	2.49±.13
Building	1.38±.12	2.82±.51	1.83±.22	1.80±.21	6.62±.02	6.94±.13	6.79±.03	6.74±.04
CCPP	**4.05±.09**	4.14±.11	4.06±.08	4.33±.12	**2.82±.02**	2.84±.03	2.82±.02	2.95±.01
Concrete	5.40±.24	**3.32±.34**	5.32±.27	5.26±.21	3.10±.04	**2.61±.18**	3.09±.04	3.09±.03
Fish Toxicity	0.88±.04	0.90±.04	0.89±.04	0.88±.03	1.28±.04	1.27±.04	1.29±.04	1.29±.03
Protein	1.93±.19	2.45±.42	**1.87±.17**	1.97±.21	**2.00±.07**	2.01±.08	2.04±.08	2.21±.04
Real Estate	7.48±1.41	7.51±1.44	**7.46±1.42**	7.52±1.40	3.60±.30	3.70±.45	**3.59±.32**	3.62±.33
Stock	3.85±1.12	3.90±1.09	3.88±1.13	**3.82±1.11**	-1.09±.04	-1.09±.04	-1.09±.04	-1.09±.04
Wine	.642±.018	.640±.021	.638±.018	.643±.019	.966±.027	.965±.028	.964±.025	.975±.027
Yacht	**0.78±.12**	1.18±.18	0.99±.12	1.00±.18	**1.70±.02**	1.79±.03	1.82±.01	2.05±.01

Figure: BNN test results

Bayesian variational autoencoder

	KL-VI	χ-VI	α-VI	TV-VI	f_{c1}-VI	f_{c2}-VI
Caltech 101	**73.80±2.27**	73.84±2.16	74.95±2.76	74.32±2.26	74.87±2.56	74.85±2.94
Frey Face	160.85±.72	160.57±.95	161.06±1.16	161.11±1.00	**160.52±.88**	160.65±.87
MNIST	**59.06±.40**	62.13±.50	61.90±.69	62.44±.41	59.60±.25	59.53±.42
Omniglot	109.62±.20	110.57±.28	110.81±.32	110.21±.31	**107.13±.39**	108.29±.28

Figure: Average test reconstruction errors of f-VAEs