Bayesian Machine Learning Project, 2020 1–7

Alpha-Beta Divergence For Variational Inference

Hamidreza Behjoo HAMIDREZA.BEHJOO@SKOLTECH.COM
Jaspers Williamn Huanay JASPERS.HUANAY@SKOLTECH.RU

Abstract

In this report Alpha-Beta divergence is investigated and compared with respect to other well known divergences in literature. Using numerical experiments we showed that as dimension grows variance of estimation grows and we can not find reliable estimation expect for 1D or 2D based on Alpha-Beta divergence.

1. Introduction

The quality of the posterior approximation is a core question in variational inference. When using the KL-divergence averaging with respect to the approximate distribution, standard VI methods such as mean-field underestimate the true variance of the target distribution. In this scenario, such behavior is sometimes known as *mode seeking*. On the other end, by (approximately) averaging over the target distribution as in Expectation-Propagation, we might assign much mass to low-probability regions (*mass covering*). In an effort to smoothly interpolate between such behaviors, some recent contributions have exploited parameterized families of divergences such as the alpha-divergence ,and the Renyi-divergence. Another fundamental property of an approximation is its *robustness to outliers*.

We study here a variational objective to simultaneously trade off effects of mass-covering, spread and outlier robustness. This is done by developing a variational inference objective using the alpha-beta (AB) divergence, a family of divergence governed by two parameters and covering many of the divergences already used for VI as special cases.

2. Divergences Classes

In this section we follow flow of ideas toward the Alpha-Beta divergence.

2.1. Kullback-Leibler Divergence

$$D_{KL}(q||p) = \int q(\theta) \log \left(\frac{q(\theta)}{p(\theta)} \right) d\theta \tag{1}$$

It offers a relatively simple to optimize objective. However, because the KL-divergence considers the log-likelihood ratio p/q, it tends to penalize more the region where $q > p$ —i.e, for any given region over-estimating the true posterior is penalized more than underestimating it. The approximation derived tends to poorly cover regions of small probability in the target.

2.2. Rényi divergence

$$D_R^{\alpha}(p||q) = \frac{1}{\alpha - 1} \log \int p(\theta)^{\alpha} q(\theta)^{1-\alpha} d\theta \tag{2}$$

For this family, the meta-parameter α can be used to control the influence granted to likelihood ratio p/q on the objective in regions of over/under estimation. This flexibility has allowed for improvements on traditional VI on complex models, by fine-tuning the meta-parameter to the problem.

2.3. Gamma-Divergence

$$\begin{aligned} D_{\gamma}^{\beta}(p||q) = & \frac{1}{\beta(\beta+1)} \log \int p(\theta)^{\beta+1} d\theta \\ & + \frac{1}{\beta+1} \log \int q(\theta)^{\beta+1} d\theta - \frac{1}{\beta} \log \int p(\theta) q(\theta)^{\beta} d\theta. \end{aligned} \tag{3}$$

In this family, the parameter β controls how much importance is granted to elements of small probability. The upshot is that in the case the data is contaminated with *outliers* – here interpreted as data points contaminated with noise, which are assumed to be spurious and must not be covered by the model.

2.4. Alpha-Beta Divergence

$$\begin{aligned} D_{sAB}^{\alpha,\beta}(p||q) \equiv & \frac{1}{\beta(\alpha+\beta)} \log \int p(\theta)^{\alpha+\beta} d\theta \\ & + \frac{1}{\alpha(\alpha+\beta)} \log \int q(\theta)^{\alpha+\beta} d\theta \\ & - \frac{1}{\alpha\beta} \log \int p(\theta)^{\alpha} q(\theta)^{\beta} d\theta, \end{aligned} \tag{4}$$

for $(\alpha, \beta) \in \mathrm{R}^2$ such that $\alpha \neq 0$, $\beta \neq 0$ and $\alpha + \beta \neq 0$.

2.4.1. Special Cases

When $\alpha = 0$ and $\beta = 1$ the sAB-divergence reduces down to the Kullback-Leibler divergence. By symmetry, the reverse KL is obtained for $\alpha = 1$ and $\beta = 0$.
More generally, when $\alpha + \beta = 1$

$$D_{sAB}^{\alpha+\beta=1}(p||q) = \frac{1}{\alpha(\alpha-1)} \log \int p(\theta)^{\alpha} q(\theta)^{1-\alpha} d\theta, \tag{5}$$

and the sAB-divergence is proportional to the Rényi-divergence.

When $\alpha = 1$ and $\beta \in \mathrm{R}$, becomes

$$\begin{aligned} D_{sAB}^{\alpha=1,\beta}(p||q) = & \frac{1}{\beta(\beta+1)} \log \int p(\theta)^{\beta+1} d\theta \\ & + \frac{1}{\beta+1} \log \int q(\theta)^{\beta+1} d\theta - \frac{1}{\beta} \log \int p(\theta) q(\theta)^{\beta} d\theta. \end{aligned} \tag{6}$$

and the sAB-divergence is equivalent to Gamma-divergence.

By changing different value of (α, β) we can cover whole range of divergences as shown in Fig 1.

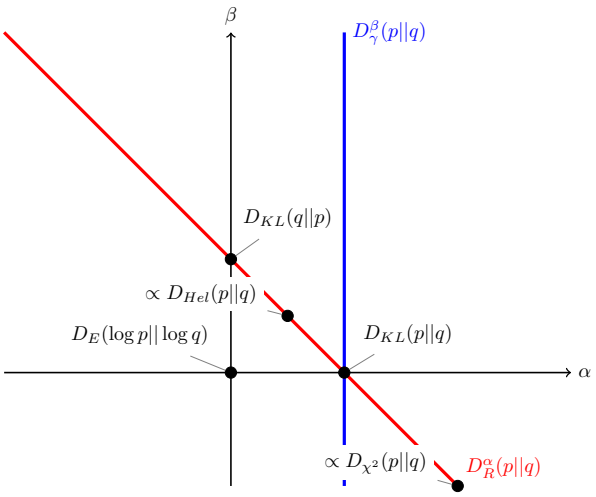

Figure 1: Mapping of the (α, β) space. The sAB-divergence reduces down to many known divergences but also interpolates smoothly in between them and cover a much broader spectrum than the Rényi or the gamma-divergence. For (α, β) equals $(0.5, 0.5)$ and $(2, -1)$ the sAB divergence is proportional to respectively the Hellinger and the Chi-square divergences.

2.4.2. Alpha-Beta Variational Objective

Alpha-Beta divergence can be written in term of expectation with respect to q as follows:

$$\begin{aligned} D_{sAB}^{\alpha,\beta}(q(\theta)||p(\theta|\mathbf{X})) = & \frac{1}{\alpha(\alpha+\beta)} \log \mathbb{E}_q \left[\frac{p(\theta, \mathbf{X})^{\alpha+\beta}}{q(\theta)} \right] \\ & + \frac{1}{\beta(\alpha+\beta)} \log \mathbb{E}_q \left[q(\theta)^{\alpha+\beta-1} \right] \\ & - \frac{1}{\alpha\beta} \log \mathbb{E}_q \left[\frac{p(\theta, \mathbf{X})^{\beta}}{q(\theta)^{1-\alpha}} \right] \end{aligned} \tag{7}$$

In usual setting of Varitional Inference we minimize the evidence lower bound (ELBO), but in (7) we directly optimize the divergence itself. Equation(7) has three main components,

- The first term ensure the objective satisfies the properties of a divergence. D_{sAB} is always positive and it is equal to 0 if and only if $p = q$.
- The second element and the weighting of the ratio $p(\theta, \mathbf{X})/q(\theta)$ in the third element by $q(\theta)^{\alpha+\beta-1}$ control the sensibility to outliers, by setting $\lambda = \alpha + \beta$ to small values below 2, one can achieve robustness to outliers whilst maintaining the efficiency of the objective.
- The scaling on the ratio $p(\theta, \mathbf{X})/q(\theta)$ by a power β in the last element favors the mass-covering property.

2.4.3. Optimization

We can not directly optimize (7) and we need to do some approximation to solve the problem. A simple Monte Carlo (MC) method equipped with reparametrization trick is deployed, which uses finite samples $\theta_k \sim q(\theta)$, $k = 1, \ldots, K$ to approximate $D_{sAB}^{\alpha,\beta} \approx \hat{D}_{sAB}^{\alpha,\beta,K}$.

$$\begin{aligned}
&\hat{D}_{sAB}^{\alpha,\beta,K}(q(.)||p(.|\mathbf{X})) \\
&= \frac{1}{\alpha(\alpha+\beta)} \log \frac{1}{K} \sum_{k=1}^{K} \frac{p(\theta_k, \mathbf{X})^{\alpha+\beta}}{q(\theta_k|\mathbf{X})} \\
&\quad + \frac{1}{\beta(\alpha+\beta)} \log \frac{1}{K} \sum_{k=1}^{K} q(\theta_k|\mathbf{X})^{\alpha+\beta-1} \\
&\quad - \frac{1}{\alpha\beta} \log \frac{1}{K} \sum_{k=1}^{K} \left[\frac{p(\theta_k, \mathbf{X})^{\beta}}{q(\theta_k|\mathbf{X})^{1-\alpha}} \right].
\end{aligned} \tag{8}$$

3. Numerical Experiments

In this section we do numerical experiments on synthetic data and UCI dataset. We also investigate the effect of outlier and how to mitigate them using Alpha-Beta divergence.

3.1. Bayesian Linear Regression

We fit a Baysein linear regression to a one dimensional dataset and consider different divergence to find the posterior.

$$y = w^{\top}\mathbf{X} + \epsilon \tag{9}$$

where $\mathbf{X}$ is selected uniformly at random between $[-1, 1]$ and we put a normal prior on $w \sim \mathcal{N}(0, 1)$. Also $\epsilon \sim \mathcal{N}(0, 0.1)$ is the noise. In Fig 2 the result of approximating posterior based on different divergences are shown. In Fig. 2b we see that approximating the posterior using KL divergence is not perfect and we have some discrepancy. We approximating KL divergence using Monte-Carlo samples

$$\hat{\mathrm{KL}}(q||p) = \frac{1}{n} \sum_{i=1}^{n} \log \frac{q(x_i)}{p(x_i)} \tag{10}$$

where $x_i \sim q(x)$. Monte Carlo estimators are consistent under mild conditions:

$$\lim_{n\to\infty} \hat{\mathrm{KL}}(q||p) = \mathrm{KL}(q||p)$$

In practice, one problem when implementing Eq. 10, is that we may end up potentially with $\hat{\mathrm{KL}}(q||p) < 0$. 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. 10 comes from the fact that $\sum_i p(x_i) \neq 1$ and $\sum_i q(x_i) \neq 1$.

In Fig. 2d for approximating the posterior based on Alpha-Beta divergence a grid search is employed to find the best value $(\alpha, \beta) = (0.75, 0.25)$. Note that Alpha-Beta divergence is very sensitive to the (α, β) selection and without careful selection of them we wont obtain a good fit for posterior.

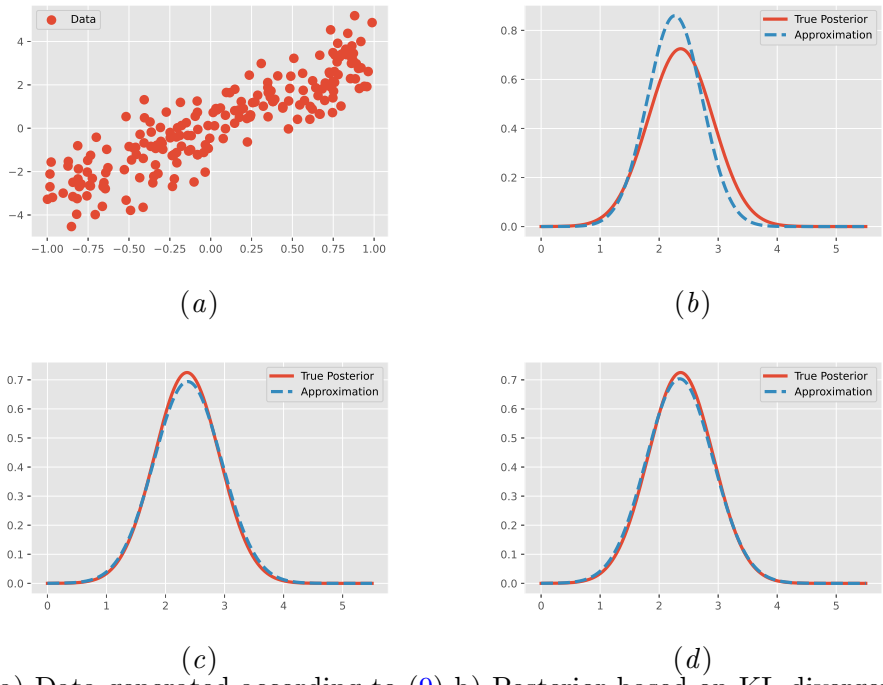

Figure 2: a) Data generated according to (9) b) Posterior based on KL divergence c) Posterior based on Rényi divergence d) Posterior based on Alpha-Beta divergence

3.1.1. Effect of outlier

In this example we investigate effect of outlier on different divergences. We contaminated 5% of data generated based on 9 and call the outlier. In Fig. 3 the mean of the predictive distributions for various values of (α, β) are displayed and results are summarized in Table 1. As expected, the network trained with Kl divergence is highly sensitive to outliers and thus has poor predictive abilities at test time Fig 3b. By careful selection of (α, β) we can mitigate outlier and find a good fit (Fig. 3d).

(α, β)	MAE	MSE
$(1, 0, 0.0)$ (KL)	0.68	0.60
$(0.7, 0.3)$ (Renyi)	0.52	0.50
$(\mathbf{2.2}, \mathbf{-0.3})$ **(sAB)**	**0.30**	**0.20**

Table 1: Average Mean Square Error and Mean Absolute Error over 40 regression experiments on the same toy dataset where the training data contain a 5% proportion of corrupted values.

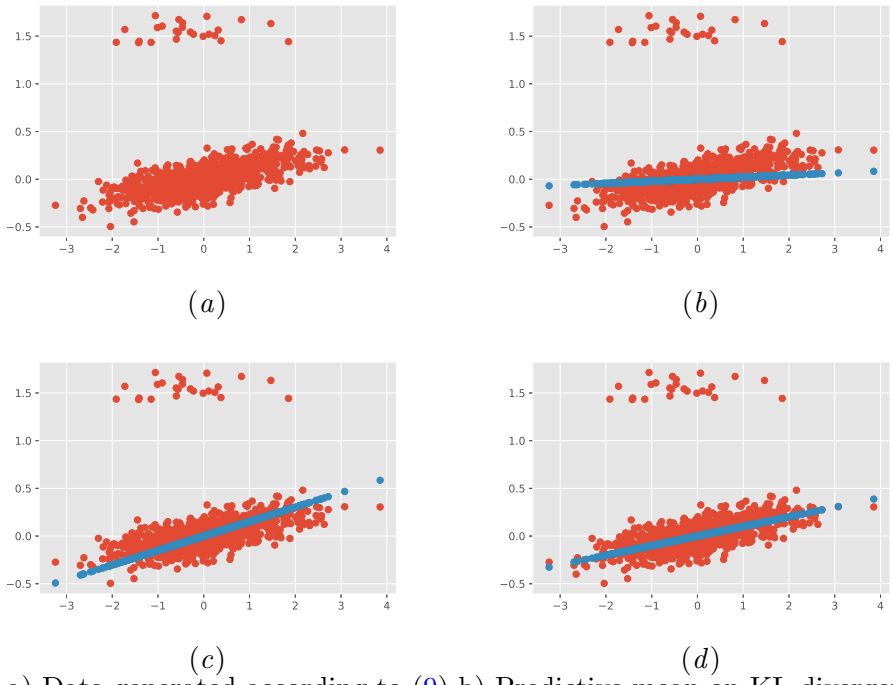

Figure 3: a) Data generated according to (9) b) Predictive mean on KL divergence c) Predictive mean based on Rényi divergence d) Predictive mean based on Alpha-Beta divergence

3.2. Boston Housing Prices

We conducted the same experiments described in the previous section for the Boston Housing dataset collected from the UCI dataset repository. We follow (Baptiste Silva, 2018) setup, meaning a Bayesian neural network with 50 hidden units, two layers and relu activations.

Below are show results for two groups, first to the data without any modification, and second with 10% outliers added.

(α, β)	MAE	MSE
$(1, 0, 0.0)$ (KL)	0.68	0.60
$(0.7, 0.3)$ (Renyi)	0.52	0.50
$(\mathbf{2.2}, \mathbf{-0.3})$ **(sAB)**	**0.30**	**0.20**

Table 2: Average Mean Square Error and Mean Absolute Error over 40 regression experiments on the same toy dataset where the training data contain a 5% proportion of corrupted values.

4. Conclusion

In one dimentional case Alpha-Beta divergence was successful to mitigate outliers and find a good fit for data. In higher dimension we have a high variance for estimation of the parameters and Alpha-Beta divergence is not a good choice to follow.

References