Alpha-Beta Divergence For Variational Inference

Hamidreza Behjoo
Jaspers Williamn Huanay

Skolkovo Institute of Science and Technology

October 24, 2020

Kullback-Leibler Divergence

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

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.

Rényi divergence

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

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.

Gamma-Divergence

$$D_{\gamma}^{\beta}(p||q) = \frac{1}{\beta(\beta+1)} \log \int p(\theta)^{\beta+1} d\theta + \frac{1}{\beta+1} \log \int q(\theta)^{\beta+1} d\theta - \frac{1}{\beta} \log \int p(\theta) q(\theta)^{\beta} d\theta.$$

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.

Alpha-Beta Divergence

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

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

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 $\beta = 0$.

Alpha-Beta Divergence

Special Cases

More generally, when $\alpha + \beta = 1$

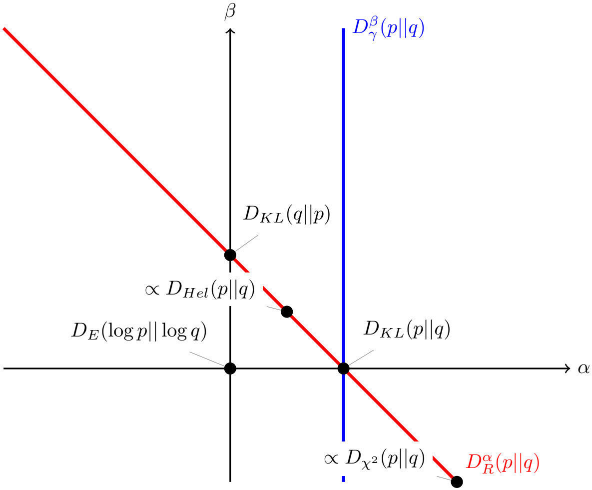
$$D_{sAB}^{\alpha+\beta=1}(p||q) = \frac{1}{\alpha(\alpha-1)} \log \int p(\theta)^{\alpha} q(\theta)^{1-\alpha} d\theta,$$

and the sAB-divergence is proportional to the Rényi-divergence.
When $\alpha = 1$ and $\beta \in \mathrm{R}$, becomes

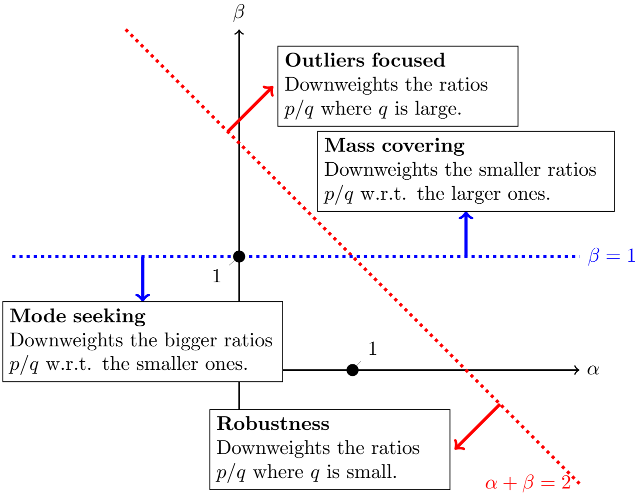
$$D_{sAB}^{\alpha=1,\beta}(p||q) = \frac{1}{\beta(\beta+1)} \log \int p(\theta)^{\beta+1} d\theta$$
$$+ \frac{1}{\beta+1} \log \int q(\theta)^{\beta+1} d\theta - \frac{1}{\beta} \log \int p(\theta) q(\theta)^{\beta} d\theta.$$

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

Alpha-Beta Divergence



Alpha-Beta Divergence



Alpha-Beta Divergence

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

Alpha-Beta Divergence

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 $D_{sAB}^{\alpha,\beta} \approx \hat{D}_{sAB}^{\alpha,\beta,K}$.

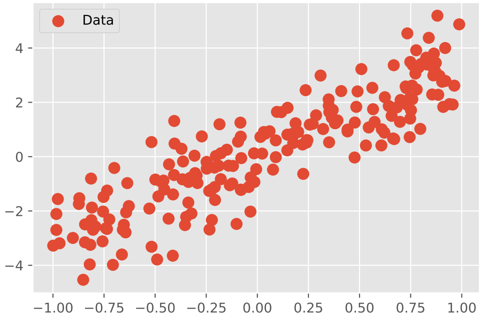
$$
\begin{aligned}
&\hat{D}_{sAB}^{\alpha,\beta,K}(q(.)||p(.|\mathbf{X})) \\
&= \frac{1}{\alpha(\alpha+\beta)} \log \frac{1}{K} \sum_{k=1}^{K} \frac{p(\theta_k, \mathbf{X})^{\alpha+\beta}}{q(\theta_k|\mathbf{X})} \\
&\quad + \frac{1}{\beta(\alpha+\beta)} \log \frac{1}{K} \sum_{k=1}^{K} q(\theta_k|\mathbf{X})^{\alpha+\beta-1}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

Numerical Experiment

Bayesian Linear Regression

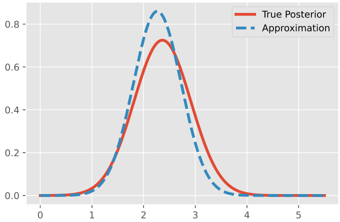
$$y = w^{\top}\mathbf{X} + \epsilon$$

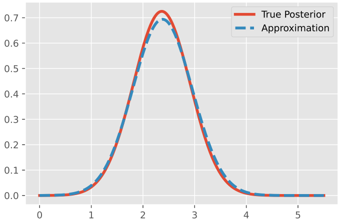
$$w \sim \mathcal{N}(0, 1)$$

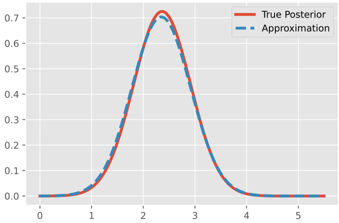


Numerical Experiment

Bayesian Linear Regression





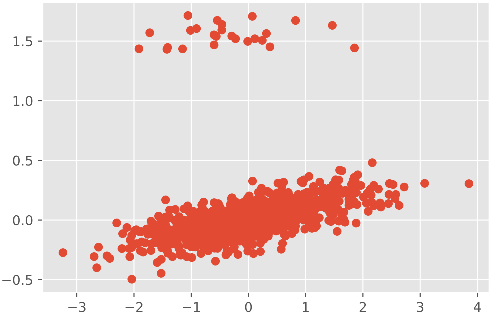


Numerical Experiment

Data contaminated with Outlier

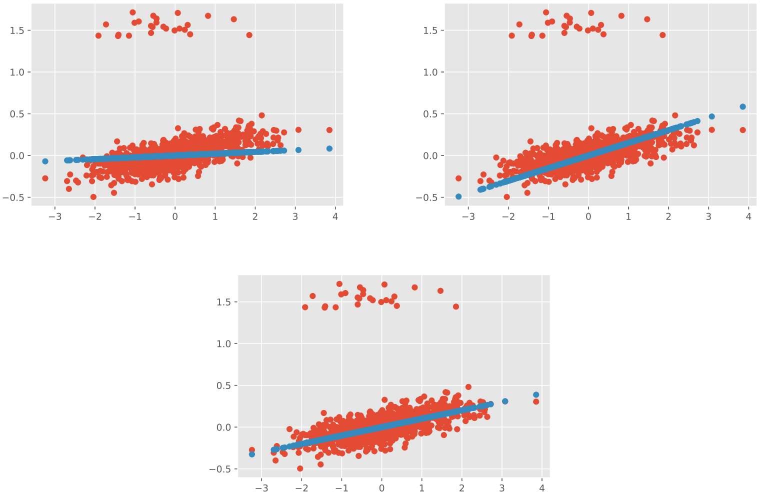
$$y = w^\top \mathbf{X} + \epsilon$$

$$w \sim \mathcal{N}(0, 1)$$



Numerical Experiments

Data contaminated with Outlier



Numerical Experiments

Data contaminated with Outlier

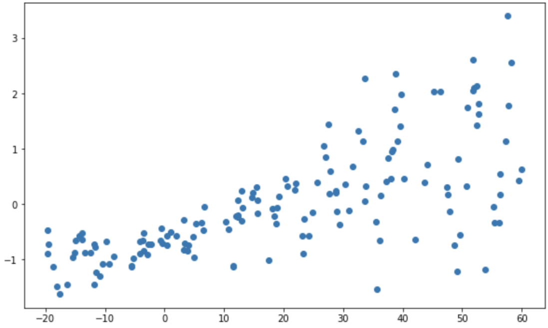
(α, β)	MAE	MSE
$(1, 0, 0.0)$ (KL)	0.68	0.60
$(0.7, 0.3)$ (Renyi)	0.52	0.50
$(\mathbf{2.2}, \mathbf{-0.3})$ **(sAB)**	**0.30**	**0.20**

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.

Numerical Experiments

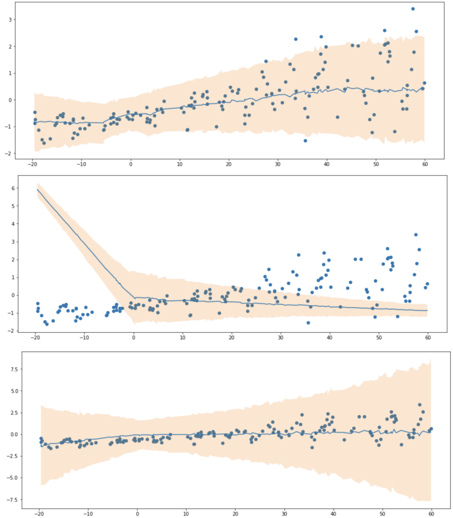
Data without Outlier

Sampled from normal distribution with noise dependent on X



Numerical Experiments

Data without Outlier, with noise dependent on X



Numerical Experiments

Data without Outlier, with noise dependent on X

(α, β)	MSE
$(1, 0, 0.0)$ (KL)	0.75
$(0.5, 0.3)$ (Renyi)	0.78
$\mathbf{(1.25, -0.3)}$ **(sAB)**	**1.06**

Table 2: AB divergence did not perform better than Kl or Renyi div, this might be due to high variance

Fin