A Remedy For Negative Monte Carlo Estimated Values Of KL-divergence

COMP4560 Final Presentation

Presenter: Jiaxu Liu[1,2]

[1]The Australian National University, Australia

[2]Shandong University, China

UID: u6920122

Supervisors: Tom Gedeon; Zhenyue Qin

Under Review – PRICAI 2021

Contribution

In sum, our contributions are three-fold:

1. We propose two novel KL estimators which guarantee the non-negativity, and bring forward new divergence based on these estimators, which are new measures for computing the difference between two distributions, one of them satisfies the axioms of semi-metrics, the other brings new insights on reducing the variance of estimators and can be analogized to which in the family of f-divergence.
2. Using D_{AKL} and D_{TKL} respectively, we derive new lower bounds of ELBO as new objective functions of VAEs.
3. We show via experimental results VAEs using our proposed measurement support more stable training and produce more realistic results.

Problem Statement

The KL-divergence:

$$D_{\mathrm{KL}}(Q\|P) = \int \log \frac{q(\mathrm{z})}{p(\mathrm{z})} q(\mathrm{z}) d\mathrm{z},$$

The KL-divergence can be intractable for some distributions due to the integral calculation. To address such intractability, it is common to estimate KL-divergence using MC sampling methods.
As a result, the estimated KL-divergence is:

$$\hat{D}_{\mathrm{KL}}(Q\|P) = \frac{1}{N} \sum_{i=1}^{N} \underbrace{\left(\log q(\mathrm{z}^{(i)}) - \log p(\mathrm{z}^{(i)})\right)}_{\text{Estimator (Naïve) } \hat{\theta}}.$$

Cannot Guarantee Non-negativity. Mislead optimization.

$\hat{\theta}_{\text{AKL}}$: Alleviation to High Variance

One intuitive approach is to calculate the absolute of difference of $(\log q(\mathrm{z}^{(i)}) - \log p(\mathrm{z}^{(i)}))$, *i.e.*

$$\left|\log q(\mathrm{z}^{(i)}) - \log p(\mathrm{z}^{(i)})\right|$$

This yields our first proposed estimator that ensure the non-negativity of MC-based estimation:

$$\hat{\theta}_{\text{Abs}} = \left|\hat{\theta}\right| = \left|\log \frac{q(\mathrm{z})}{p(\mathrm{z})}\right| \tag{1}$$

Pros: Lower variance
Cons: Biased

D_{AKL} Definition

We propose the Absolute-KL-divergence (D_{AKL}):

$$D_{\mathrm{AKL}}(Q\|P) = \int_{\Omega} \left| \log \frac{q(\mathrm{z})}{p(\mathrm{z})} \right| r(\mathrm{z}) d\mathrm{z} \tag{2}$$

where the Ω is the sample space. We use a new PDF $r(\mathrm{z})$ as the reference distribution, defined as:

$$r(\mathrm{z}) = \frac{p(\mathrm{z}) + q(\mathrm{z})}{2} \tag{3}$$

Reason: To avoid asymmetry. With this property, we can prove that D_{AKL} is a semi-metric.
Semi-metric: For a metric d, it holds: (1) non-negativity, *i.e.* $d(x; y) \geq 0$; (2) identity of indiscernibles, *i.e.* $d(x, y) = 0$ iff $x = y$; (3) symmetry, *i.e.* $d(x, y) = d(y, x)$.

$\hat{\theta}_{\mathrm{TKL}}$: Between Convex Function and Tangent Plane

Though $\hat{\theta}_{\mathrm{AKL}}$ has remarkably low bias, we still aim to construct a estimator that is both unbiased and has low variance.
Inspired by Bregman Divergence and the property of Control Variate in statistic, we can construct a regression estimator:

$$\begin{aligned}\hat{\theta}_{\mathrm{Tan}} &= -\log \nu(z) + \nu(z) - 1 \\ &= \frac{p(z)}{q(z)} - 1 - \log \frac{p(z)}{q(z)} \\ &= e^{-\hat{\theta}} - 1 + \hat{\theta},\end{aligned} \tag{4}$$

which is unbiased, guarantee the non-negativity, also exhibits low variance.
(Two pages of derivation, refer to thesis)

D_{TKL}: Tangent-KL-divergence

Similar to D_{AKL}, we propose the Tangent-KL-divergence:

$$D_{\mathrm{TKL}}(Q\|P) = \int_{\Omega} \left(\frac{p(\mathrm{z})}{q(\mathrm{z})} - 1 - \log \frac{p(\mathrm{z})}{q(\mathrm{z})} \right) q(\mathrm{z}) d\mathrm{z} \tag{5}$$

Cons: Unfortunately, the D_{TKL} is not symmetric regardless of the reference distribution of space Ω, even if we leverage the PDF $r(\mathrm{z})$ in Eq.(3), this metric is not symmetric.
Pros: Analogism to the family of f-divergence.

Analogize to f-divergence

Proposed general estimator for any f-divergence $D_f(Q\|P)$ derived by above method:

$$\begin{aligned}\hat{\theta}_f &= f\left(\nu(\mathrm{z})^{-1}\right) - f'(1)\left(\nu(\mathrm{z})^{-1} - 1\right) \\ &= f\left(\frac{q(\mathrm{z})}{p(\mathrm{z})}\right) - f'(1)\left(\frac{q(\mathrm{z})}{p(\mathrm{z})} - 1\right)\end{aligned} \tag{6}$$

An illustration in the case of D_{KL} where $f(t) = t\log t$, $f'(t=1) = \log t + 1 = 1$:

$$\hat{\theta}_{f_{\mathrm{KL}}} = \frac{q(\mathrm{z})}{p(\mathrm{z})}\log\left(\frac{q(\mathrm{z})}{p(\mathrm{z})}\right) - \frac{q(\mathrm{z})}{p(\mathrm{z})} + 1 \tag{7}$$

(Detailed derivation in thesis)

Toy Experiment

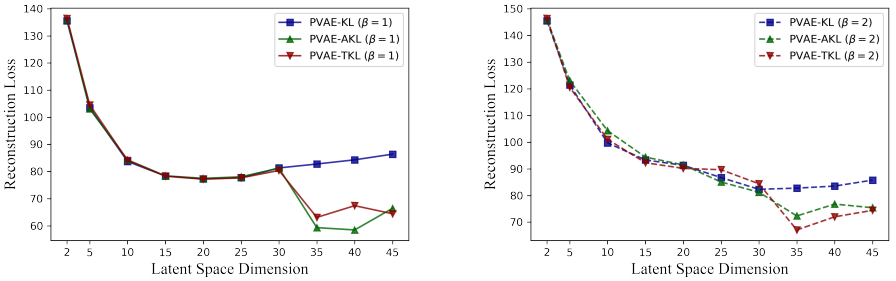
distribution family	probability density function	data distribution $p(\cdot;\theta)$	variational distribution $q(\cdot;\phi)$	Bias $B(\hat{\theta})$	Bias $B(\hat{\theta}_{\text{Abs}})$	Bias $B(\hat{\theta}_{\text{Tan}})$	Variance $Var(\hat{\theta})$	Variance $Var(\hat{\theta}_{\text{Abs}})$	Variance $Var(\hat{\theta}_{\text{Tan}})$
Beta	$f(\cdot;\alpha,\beta)$ $= \frac{x^{\alpha-1}(1-x)^{\beta-1}}{B(\alpha,\beta)}$	$Beta(0.5,0.5)$	$Beta(5,1)$	1.5e-4	1.4e-1	-1.2e-1	6.0e-1	3.6e-1	9.8e1
		$Beta(0.5,0.5)$	$Beta(1,3)$	-7.3e-5	1.7e-1	5.2e-3	5.7e-1	3.4e-1	1e2
		$Beta(0.5,0.5)$	$Beta(2,2)$	4.2e-4	2.1e-1	1e-2	5.9e-1	3.5e-1	6.2e1
		$Beta(0.5,0.5)$	$Beta(2,5)$	-1.2e-4	1.6e-1	-1.1e-1	6.4e-1	3.8e-1	5.4e1
Gaussian	$f(\cdot;\mu,\sigma)$ $= \frac{e^{-\frac{1}{2}(\frac{x-\mu}{\sigma})^2}}{\sigma\sqrt{2\pi}}$	$\mathcal{N}(0,1)$	$\mathcal{N}(0,0.2)$	-1.3e-4	1.1e-1	-3.1e-1	6.7e-1	4.5e-1	3.6e1
		$\mathcal{N}(0,1)$	$\mathcal{N}(0,5)$	-5.3e-3	6.1e-1	-5.6e-3	1.8e1	1.7e1	1.6e1
		$\mathcal{N}(0,1)$	$\mathcal{N}(0.1,1)$	6.2e-6	7.4e-2	7.7e-7	1e-1	6.0e-2	7.1e-3
		$\mathcal{N}(0,1)$	$\mathcal{N}(1,1)$	2.5e-4	3.9e-1	9.4e-5	1.0	6.6e-1	8.4e-1
Laplace	$f(\cdot;\mu,b)$ $= \frac{e^{-\frac{\lvert x-\mu\rvert}{b}}}{2b}$	$Laplace(0,1)$	$Laplace(0,2)$	3.3e-4	3.8e-1	1.4e-4	1.0	7.8e-1	5.7e-1
		$Laplace(0,1)$	$Laplace(0,4)$	-1.0e-3	5.5e-1	-1.1e-3	2.9	2.6	2.4
		$Laplace(0,1)$	$Laplace(0.1,1)$	-4.1e-6	9.2e-2	1.6e-7	9.8e-2	1.2e-2	7.9e-4
		$Laplace(0,1)$	$Laplace(-5,4)$	-1.5e-3	4.9e-1	-6.2e-5	3.8	3.3	3.2

Table: evaluation results of the bias and variance of different estimators under Beta, Gaussian and Laplace distributions with different parameter settings. The values marked in green are the best of the group and those in red are the worst.

Application: Lower Reconstruction Errors of Hyperbolic VAEs

$$\log p(\mathrm{x}) \geq \underbrace{\mathbb{E}_{\mathrm{z}\sim q_\phi(\mathrm{z}|\mathrm{x})}[\log p_\theta(\mathrm{x}|\mathrm{z})] - \beta D_{\star\mathrm{KL}}(q_\phi(\mathrm{z}|\mathrm{x})\|p(\mathrm{z}))}_{\text{our new lower bound of ELBO}}, \tag{8}$$

Averaged Test Reconstruction Error



(a) $\beta = 1$ (b) $\beta = 2$

Figure: Comparison of reconstruction loss of optimizing $\mathcal{P}^c$-VAEs with KL-, AKL- and TKL-divergence vs dimensionality of the latent space.

Application: Lower Reconstruction Errors of Hyperbolic VAEs

Different Curvatures of Hyperbolic Manifold

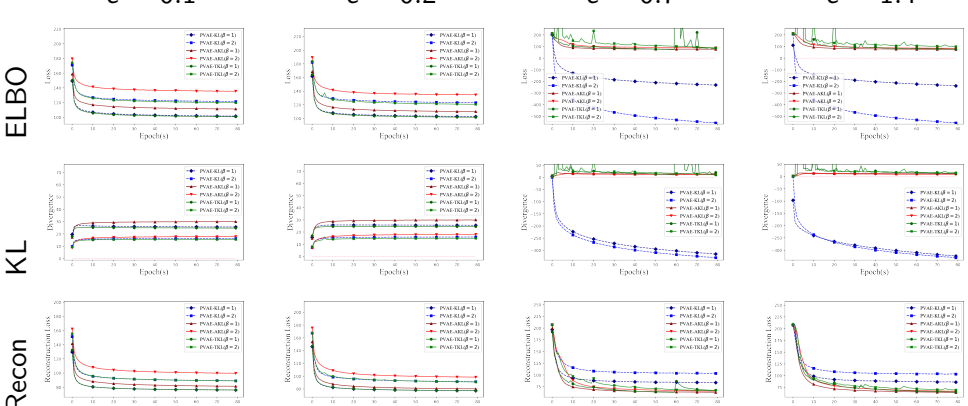


Figure: Comparison of ELBO, reconstruction loss and divergence of optimizing $\mathcal{P}^c$-VAEs with KL-divergence and AKL-divergence and TKL-divergence, vs. curvature $c = 0.1$, 0.2, 0.7, 1.4 respectively. The latent dimension is fixed at 40.

Application: Lower Reconstruction Errors of Hyperbolic VAEs

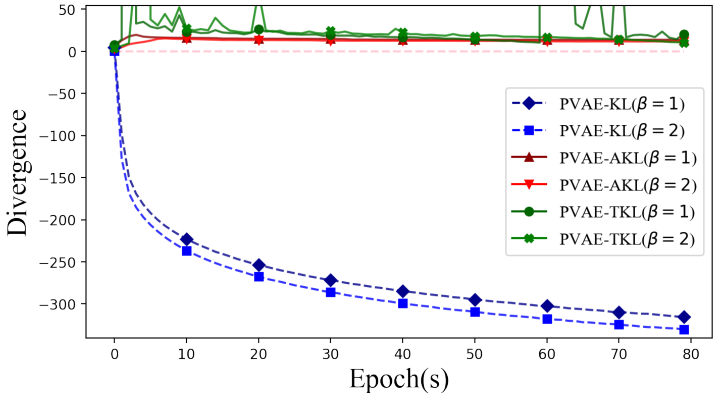


Figure: c=0.7; KL optimization

Application: Lower Reconstruction Errors of Hyperbolic VAEs

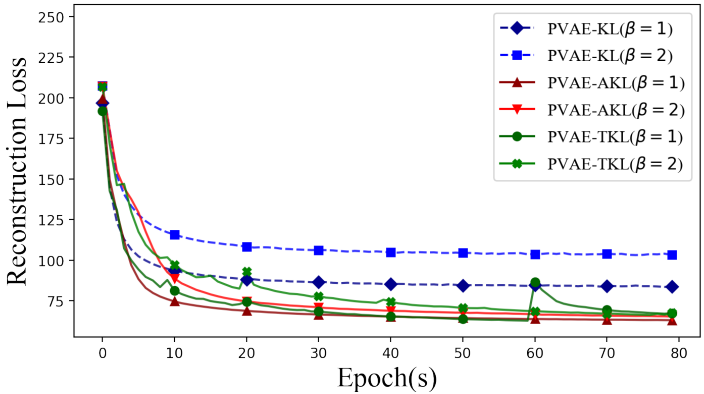


Figure: c=0.7; Reconstruction

Application: Realistic Human Appearance Generation

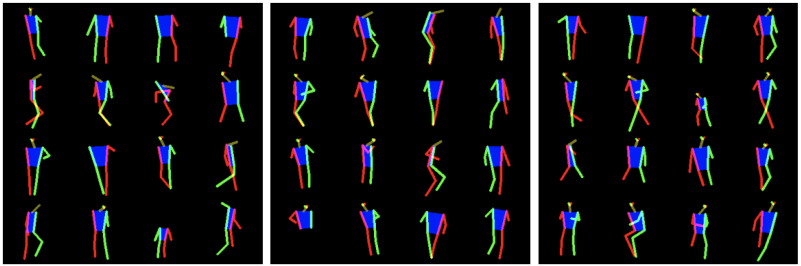


Figure: Examples of human geometrical information y in the Market-1501 dataset

Application: Realistic Human Appearance Generation

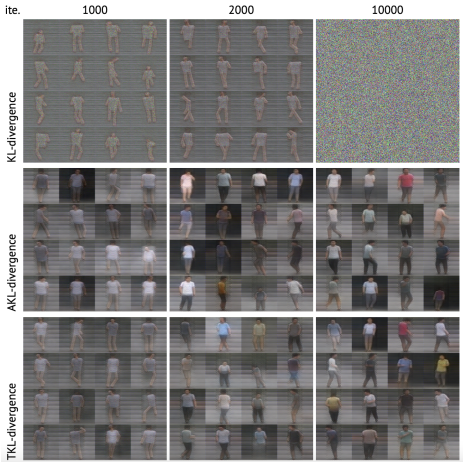


Figure: Comparison of optimizing vunets with KL- and AKL- and TKL-divergence on the market-1501 dataset. The *ite.* stands for iteration.

Thank You!

Thank you for listening