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

Zhenyue Qin[1], Jiaxu Liu[1*], Yang Liu[1,2], Saeed Anwar[1,2], Pan Ji[3], Tom Gedeon[1]
[1]Australian National University
[2]Data61 CSIRO
[3]OPPO US Research Center

***Abstract*—Kullback–Leibler divergence (KL-divergence) is widely used as an objective function in machine learning, in specific scenario, computing its analytical form is challenging; hence, Monte Carlo techniques are standard alternatives to estimate the KL-divergence. However, the unbiased estimation of Monte Carlo methods requires an unrealistically infinite available data samples. In contrast, standard small sample sizes can cause negatively estimate KL-divergence, whereas KL-divergence is non-negative by definition, such non-negativity of estimated KL-divergence results in severe noises and high variance. In this paper, we propose new estimators of KL-divergence, namely absolute and tangent estimator, we formulate new divergence using these estimators and propose Absolute-KL and Tangent-KL, abbreviated as AKL- and TKL-divergence, new quantities to measure the difference between two distributions, with a similar form of KL-divergence. The experiment demonstrated that the Monte Carlo estimations of our proposed divergence maintain non-negativity, facilitating more accurate approximations than with KL-divergence.**

***Keywords*—Divergence; Variational Inference; Monte Carlo;**

I. Introduction

Kullback-Leibler divergence, abbreviated as KL-divergence, is a widely-used measure for evaluating the difference between two distributions [1]. KL-divergence has many applications, examples including computing mutual information between two variables [2] and devising objective functions of neural networks [3]. To some pairs of distributions, the exact quantity of their KL-divergence is intractable [4], such as two Gaussian Mixture Models [5].

We denote KL-divergence as D_{KL}. When KL is hard to compute, Monte Carlo (MC) method is a common approach [6]. To briefly illustrate, we sample N results in order to approximate $D_{\mathrm{KL}}(Q\|P)$ from distribution Q as $\{\mathbf{z}_i\}_{i=1}^{N}$, then to compute $\frac{1}{N}\sum_{i=1}^{N}(\log q(\mathbf{z}_i) - \log p(\mathbf{z}_i))$ as the estimated D_{KL} [7]. However, such estimations can lead to severe noisy results. To illustrate: by definition, D_{KL} is non-negative [8], whereas MC estimation can violate the non-negativity of D_{KL} [9]. Furthermore, the negative estimated D_{KL} is also irrational since D_{KL} stands for the difference between two distributions, and it is unfounded to say a difference is negative.

To fix the negativity of MC estimation of D_{KL}, we can calculate the absolute value of difference $(\log q(\mathbf{z}_i) - \log p(\mathbf{z}_i))$ *i.e.*, to utilize $|\log q(\mathbf{z}_i) - \log p(\mathbf{z}_i)|$ instead of directly using the log difference. Consequently, MC estimation becomes non-negative due to the non-negativity of absolute. We also show in the paper that the new absolute form is a semi-metric, supporting its discriminability for two distributions. Another approach is to utilize the variance reduction technique to generate the *regression estimator* of KL-divergence used in Monte Carlo estimation. By choosing specific coefficient, we can guarantee the non-negativity of estimation. We give detailed derivation of these two novel estimators and the corresponding new divergences and lower-bound functions used in optimization. For convenience's sake, we denote the divergence derived by using the first and the second approach as D_{AKL} and D_{TKL} respectively.

Apart from theoretic advantages, in practice, we also demonstrate that one can replace KL-divergence within the objective function of variational autoencoders (VAEs) [10] with these two divergence, introducing new lower bounds of evidence for VAEs lower bound estimation (ELBO). We conduct experiments with two VAEs using different divergence, where it has been shown that the analytical forms of KL-divergence are hard to compute. Our experimental results indicate that MC sampling with new divergence is more stable compared to the vanilla model. Moreover, the generated results using D_{AKL} and D_{TKL} are more realistic than employing D_{KL}, supported by quantitative lower reconstruction error and qualitative more realistic looking of generated sampled images.

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. They 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.

*Equal contribution with Zhenyue Qin.

II. Problems of KL Estimation with Monte Carlo

The KL-divergence is a widely-used quantity that measures the difference between two distributions Q and P:

$$D_{\mathrm{KL}}(Q\|P) = \mathbb{E}_{\mathbf{z}\sim q(\mathbf{z})} \log \frac{q(\mathbf{z})}{p(\mathbf{z})}.$$

For discrete variables, KL-divergence is defined as:

$$D_{\mathrm{KL}}(Q\|P) = \sum_{\mathbf{z}} \log \frac{Q(\mathbf{z})}{P(\mathbf{z})} Q(\mathbf{z}),$$

where $Q(\mathbf{z})$ evaluates the probability of variable $\mathbf{z}$. Furthermore, for the sake of generalizability, we consider variables to be continuous, then mathematically, the KL-divergence is expressed as:

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

where $q(\mathbf{z})$ and $p(\mathbf{z})$ are the probability density functions (PDFs) of Q and P, respectively. Moreover, using Jensen's inequality, we can show that $D_{\mathrm{KL}}(Q\|P)$ is non-negative. Intuitively, the before-mentioned non-negativity of KL-divergence is plausible since it reflects the difference between two distributions, and negative differences are not sensible.

The KL-divergence defined in (II) 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:

$$\hat{D}_{\mathrm{KL}}(Q\|P) = \frac{1}{N} \sum_{i=1}^{N} \big(\log q(\mathbf{z}^{(i)}) - \log p(\mathbf{z}^{(i)}) \big).$$

Furthermore, when $N \to \infty$, the estimated KL-divergence by MC attains equality with the true KL-divergence, *i.e.*, $\hat{D}_{\mathrm{KL}} = D_{\mathrm{KL}}$.

In practice, the infinite sampling size is impossible. As a result, $\hat{D}_{\mathrm{KL}}(Q\|P)$ can be negative. However, by definition, the true KL-divergence $D_{\mathrm{KL}}(Q\|P)$ cannot be negative. Thus, MC sampling can cause strong noises to KL-divergence estimation, leading to failures of optimization tasks for minimizing KL-divergence. For example, to minimize $\hat{D}_{\mathrm{KL}}(Q\|P)$ via back-propagation on the parameters of distribution Q, if the estimation is negative, back-propagation will update the parameters to make $\hat{D}_{\mathrm{KL}}(Q\|P)$ even more negative, causing optimization going more astray.

III. Alleviation to MC Noises

Motivated by the negative MC estimation of KL-divergence using naïve estimator and the above concern mentioned regarding metrics, in this paper, we propose two estimators to tackle the non-negativity. We show our proposed estimators consistently produce non-negative MC values, regardless of samples sizes.

Furthermore, we exhibit the new divergence, one of them is by definition a semi-metric [11], supporting the rationality of using our estimators to measure the difference between two distributions. In [12], Abou-Moustafa *et al*. have shown that divergences of distributions have larger discriminability if the divergences meet the axioms of semi-metrics. Another proposed reformulated estimator can be generalised to the family of f-divergence inspired by the derivation of Bregman divergence [13] and the idea of control variates [14]. This estimator is unbiased and its variance empirically lower in most cases.

To start with, we will start by briefly reviewing the naïve estimator for KL-divergence. Then, for the two proposed estimators of KL-divergence, we will in detail demonstrate the motivation, derivation and proofs of their properties.

A. Naïve KL Estimator

The KL-divergence that measures the difference between two distributions Q and P. We utilize MC to estimtate the value of KL-divergence:

$$D_{\mathrm{KL}}(Q\|P) \approx \frac{1}{N} \sum_{i=1}^{N} \log \frac{q(\mathbf{z}^{(i)})}{p(\mathbf{z}^{(i)})}.$$

Let θ be the parameter that needs to be estimated and $\hat{\theta}$ is the estimator of θ. In this case, the naïve estimator $\hat{\theta}$ is defined as

$$\hat{\theta}(\mathbf{z}) = \log \frac{q(\mathbf{z})}{p(\mathbf{z})} \quad \text{or} \quad \hat{\theta}(\mathbf{z}) = \log q(\mathbf{z}) - \log p(\mathbf{z}). \tag{1}$$

B. Alleviation to High Variance

To fix the negative MC estimation of KL-divergence, one intuitive approach is to calculate the absolute of difference $(\log q(\mathbf{z}^{(i)}) - \log p(\mathbf{z}^{(i)}))$, *i.e.*, to utilize $\big|\log q(\mathbf{z}^{(i)}) - \log p(\mathbf{z}^{(i)})\big|$, instead of directly using the log difference. Consequently, MC estimation becomes non-negative due to the non-negativity of the absolute values. This yields our first proposed estimator that ensures the non-negativity of MC-based estimation:

$$\hat{\theta}_{\mathrm{Abs}} = \left|\hat{\theta}\right| = \left|\log \frac{q(\mathbf{z})}{p(\mathbf{z})}\right| \tag{2}$$

where $\hat{\theta}$ is the naïve estimator. Intuitively, the new estimator is better because each sample can tell how far apart q and p positively. Empirically, $\hat{\theta}_{\mathrm{Abs}}$ also has lower variance than $\hat{\theta}$. Nonetheless, is a biased estimator. Although the absolute seems similar to the original form of the log difference, it is questionable whether the new absolute form is an appropriate metric to measure the difference between two distributions, lacking theoretic grounds.

C. D_{AKL}: Absolute-KL-divergence

Since the new distribution metric with estimator $\hat{\theta}_{\mathrm{Abs}}$ is similar to the form of KL-divergence, yet with an additional *absolute* component, we name it as Absolute-KL-divergence, abbreviated as AKL-divergence (D_{AKL}). Formally, it is defined as follows:

$$D_{\mathrm{AKL}}(Q\|P) = \int_{\Omega} \left|\log \frac{q(\mathbf{z})}{p(\mathbf{z})}\right| r(\mathbf{z}) d\mathbf{z} \tag{3}$$

where the Ω is the sample space. To avoid asymmetry, we use a new PDF $r(\mathbf{z})$ as the reference distribution on the sample space instead of using $q(\mathbf{z})$. The new PDF $r(\mathbf{z})$ is defined as:

$$r(\mathbf{z}) = \frac{p(\mathbf{z}) + q(\mathbf{z})}{2} \tag{4}$$

Despite the similarity between D_{KL} and D_{AKL}, the two are distinct quantities for measuring the difference between two distributions. Therefore, one may have a concern about the plausibility of using D_{AKL} to evaluate distribution differences. In the following, we dispel such a concern by showing D_{AKL} satisfies all the three axioms of a semi-metric [11]. Furthermore, as Abou-Moustafa *et al*. have shown in [12], divergences for measuring the differences between two distributions can have greater discriminability if the divergences meet the semi-metric axioms [15].

D. D_{AKL} Is A Semi-Metric

The semi-metric axioms consist of four requirements [16], *i.e*., 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*., $d(x, y) = d(y, x)$. The requirements of non-negativity and symmetry are obvious for D_{AKL}.

In the following, we show D_{AKL} also meets condition: identity of indiscernibles.

Proposition 1. *AKL-Divergence satisfies identity of indiscernible. That is,*
$D_{AKL}(Q\|Q) = 0$ and if $D_{AKL}(Q\|P) = 0$, then $Q = P$.

Proof. It is obvious to observe $D_{\text{AKL}}(Q\|Q) = 0$. We focus demonstrating if $D_{\text{AKL}}(Q\|P) = 0$, then $Q = P$, by contradiction. Suppose there exists Q and P such that $Q \neq P$, yet $D_{\text{AKL}}(Q\|P) = 0$. Then, since $Q \neq P$, $\exists\ \mathbf{z}'$: $Q(\mathbf{z}') \neq P(\mathbf{z}')$, and thus $r(\mathbf{z}') \neq 0$ as well as $\left|\log \frac{q(\mathbf{z}')}{p(\mathbf{z}')}\right| \neq 0$. Since $\forall\ \mathbf{z}$: $r(\mathbf{z})\left|\log \frac{q(\mathbf{z})}{p(\mathbf{z})}\right| \geq 0$, we have $\exists!\ \mathbf{z}''$: $\left|\log \frac{q(\mathbf{z}'')}{p(\mathbf{z}'')}\right| = -\left|\log \frac{q(\mathbf{z}')}{p(\mathbf{z}')}\right|$. Therefore, $\left|\log \frac{q(\mathbf{z}')}{p(\mathbf{z}')}\right|$ cannot be canceled out, and thus D_{AKL} cannot be 0. □

E. Between Convex Function and Tangent Plane

Compared with $\hat{\theta}$, $\hat{\theta}_{\text{Abs}}$ is theoretic-grounded remedy for the high variance of naïve MC estimation since (1) it is a semi-metric and (2) its expectation is an f-divergence [9]. We further investigate how to construct an estimator that is unbiased while also has low variance. We are inspired by the Bregman divergence [13], where we define a function F for distribution q and p, such that the Bregman divergence is the difference between the value of F at p, and the value of the first-order Taylor expansion of F around q evaluated at p. This idea follows the same spirit with finding a tangent plane (using Taylor expansion on specific point) of a convex function F and looking at the difference between and measure the difference between F and the generated plane.

Let the unknown parameter of interest be μ, suppose we have a statistic $f(\mathbf{X})$ such that $\mu = \mathbb{E}[f(\mathbf{X})]$, letting $\hat{\mu} = \frac{1}{n}\sum_{i=1}^{n} f(\mathbf{X}_i)$, then $\hat{\mu}$ is an unbiased estimator of μ. Suppose we know another statistic $h(\mathbf{X})$ such that $\tau = \mathbb{E}[h(\mathbf{X})]$ is a known value, letting letting $\hat{\tau} = \frac{1}{n}\sum_{i=1}^{n} h(\mathbf{X}_i)$. Then for coefficient $\alpha \in \mathbb{R}$ we can estimate μ by:

$$\hat{\mu}_\alpha = \frac{1}{n}(f(\mathbf{X}_i) + \alpha h(\mathbf{X}_i)) - \alpha\tau = \hat{\mu} + \alpha(\hat{\tau} - \tau),$$

where $\hat{\mu}_\alpha$ is the ***regression estimator*** and the known mean $h(\mathbf{X})$ is the ***control variate*** [14], [17]–[19], it is a variance reduction technique used in Monte Carlo methods. To employ this technique, recall the value we want to estimate $\mathbb{E}_{q(\mathbf{z})} \log \frac{q(\mathbf{z})}{p(\mathbf{z})}$, let

$$\nu(\mathbf{z}) = \frac{p(\mathbf{z})}{q(\mathbf{z})}, \tag{5}$$

we have $\mathbb{E}[\nu(\mathbf{z})] = 1$ since the optimization target is to approximate p_θ with q_ϕ and they become identical, *i.e*., $\frac{q(\mathbf{z}^\star)}{p(\mathbf{z}^\star)} = 1$ when the optimization converge. According to the definition of *regression estimator*, let $\mu(\mathbf{z})$ be the *control variate*, as $\hat{\mu} = \hat{\theta} = \log \frac{q(\mathbf{z})}{p(\mathbf{z})} = \log \nu(\mathbf{z})^{-1} = -\log \nu(\mathbf{z})$ is an unbiased estimator, let the statistic $\hat{\mu} = -\log \nu(\mathbf{z})$, another statistic $\hat{\tau} = \nu(\mathbf{z})$, and $\tau = 1$, generating the regression estimator:

$$\hat{\mu}_\alpha = -\log \nu(\mathbf{z}) + \alpha(\nu(\mathbf{z}) - 1), \tag{6}$$

which is by definition also an unbiased estimator of KL-divergence. With respect to the coefficient α, although it is possible calculate the optimal value of α that generate the estimator with minimum variance by:

$$\alpha^\star = -\frac{Cov(\hat{\mu}, \hat{\tau})}{Var(\hat{\mu})},$$

in most of the scenario the p_θ is considered as not analytical so the whole expression of the estimator is hard to compute. However, according to the log inequality $\forall x > 0$: $f(x) = x - 1$ is the tangent plane to $f(x) = \log x$ with tangent point $x = 1$, such that $x - 1 \geq \log x$. Hence if $\alpha = 1$ in Eq.(6), $\hat{\mu}_\alpha = -\log \nu(\mathbf{z}) + \nu(\mathbf{z}) - 1 \geq 0$ always establish, yielding the definition of our second proposed estimator:

$$\begin{aligned} \hat{\theta}_{\text{Tan}} &= -\log \nu(\mathbf{z}) + \nu(\mathbf{z}) - 1 \\ &= \frac{p(\mathbf{z})}{q(\mathbf{z})} - 1 - \log \frac{p(\mathbf{z})}{q(\mathbf{z})} \\ &= e^{-\hat{\theta}} - 1 + \hat{\theta}, \end{aligned} \tag{7}$$

where $\hat{\theta}$ is the naïve estimator. We fix $\alpha = 1$ to guarantee the **non-negativity** of the estimator no matter what the value of $q(\mathbf{z}^{(i)})$ and $p(\mathbf{z}^{(i)})$ are.

F. D_{TKL}: Tangent-KL-divergence

Similar to D_{AKL}, we propose the Tangent-KL-divergence, abbreviated as TKL-divergence (D_{TKL}), where the estimator $\hat{\theta}$ is reformulated to $\hat{\theta}_{\text{Tan}}$ based on the idea of Bregman divergence.

$$D_{\text{TKL}}(Q\|P) = \int_\Omega \left(\frac{p(\mathbf{z})}{q(\mathbf{z})} - 1 - \log \frac{p(\mathbf{z})}{q(\mathbf{z})}\right) q(\mathbf{z}) d\mathbf{z} \tag{8}$$

Nonetheless, the D_{TKL} is not symmetric regardless of the reference distribution of space Ω. Even if we leverage the

PDF $r(\mathbf{z})$ in Eq.(4), this metric is not symmetric. Thus the D_{TKL} cannot be a semi-metric. However, such reformulation of the divergence still have one good property, which is the analogism to the family of f-divergence [20].

G. Connection to f-Divergence

The f-divergence is a function $D_f(Q\|P)$ that measures the difference between two probability distributions Q and P

$$D_f(Q\|P) = \int_{\Omega} f\left(\frac{q(\mathbf{z})}{p(\mathbf{z})}\right) p(\mathbf{z}) d\mathbf{z} \tag{9}$$

where the function f is a convex function such that $f'(1) = 0$, which could be interpreted as a weighted average of odds ratio given by Q and P. The KL-divergence is a special case of the f-divergence (where $f(t) = t\log t$), and some of the familiar divergences *e.g.* total variation ($f(t) = \frac{|t-1|}{2}$), Jensen-Shannon divergence ($f(t) = (t+1)\log\left(\frac{2}{t+1} + t\log t\right)$) also belong to the family of f-divergence.

Under the definition of $\nu(\mathbf{z})$ in Eq.(5), we have the conclusion that $\mathbb{E}[\nu(\mathbf{z})] = \mathbb{E}[\nu(\mathbf{z})^{-1}] = 1$. Thus, we always have an estimator of f-divergence according to the principle of control variation:

$$\hat{\theta}_f^{\star} = f\left(\nu(\mathbf{z})^{-1}\right) - c\left(\nu(\mathbf{z})^{-1} - 1\right) \tag{10}$$

Since f is convex function, we let $c = f'(1)$ to get the tangent plane at $\nu(x) = 1$ in order to guarantee the non-negativity of the estimator. This gives us a general estimator for any f-divergence $D_f(Q\|P)$:

$$\begin{aligned}\hat{\theta}_f &= f\left(\nu(\mathbf{z})^{-1}\right) - f'(1)\left(\nu(\mathbf{z})^{-1} - 1\right) \\ &= f\left(\frac{q(\mathbf{z})}{p(\mathbf{z})}\right) - f'(1)\left(\frac{q(\mathbf{z})}{p(\mathbf{z})} - 1\right)\end{aligned} \tag{11}$$

To give an illustration in the case of D_{KL} where $f(t) = t\log t$, $f'(t=1) = \log t + 1 = 1$, if we replace t with $\frac{q(\mathbf{z})}{p(\mathbf{z})}$, we have one possible estimator for the D_{KL}:

$$\hat{\theta}_{f_{\mathrm{KL}}} = \frac{q(\mathbf{z})}{p(\mathbf{z})}\log\left(\frac{q(\mathbf{z})}{p(\mathbf{z})}\right) - \frac{q(\mathbf{z})}{p(\mathbf{z})} + 1 \tag{12}$$

In this case, we assume the tangent point is at $\nu(\mathbf{z}) = 1$ since it the most straightforward and intuitive. However, we can leverage any possible value on the domain of $\nu(\mathbf{z})$ as the tangent point as long as it gives us a feasible tangent plane that avoids negativity.

IV. Experiment

A. Lower Reconstruction Errors of Hyperbolic VAEs

One application of our proposed estimators is to derive new evidence of lower bound estimation (ELBO) regarding variational autoencoders (VAEs). The new ELBO functions are more robust when using MC to estimate D_{AKL} and D_{TKL}.

Original VAEs and the subsequent proposed modifications consider latent representation spaces as Euclidean spaces. However, Euclidean spaces cannot support embedding data in a hierarchical fashion and has limitation in distance metric, which is not a good representation for many kinds of data structure, *e.g.* graph-like data. To facilitate embedding hierarchies in hyperbolic space, *Poincaré*-VAEs ($\mathcal{P}^c$-VAEs) have recently been introduced [21], which learns the projection from the latent space to hyperbolic ball using an encoder-decoder architecture. The $\mathcal{P}^c$-VAEs can be viewed as a generalisation of vanilla VAEs, *i.e.*, $\mathcal{P}^c$-VAE $\longrightarrow \mathcal{N}$-VAE when $c = 0$. Nevertheless, when switching from Euclidean spaces to hyperbolic spaces, the analytical form of D_{KL} becomes hard to compute. Thus, again MC techniques are utilized for estimating D_{KL}.

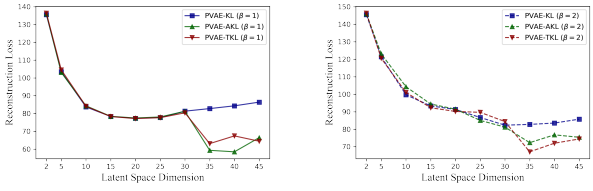
In our experiments, we optimize $\mathcal{P}^c$-VAEs with the new lower bounds (Eq.(13)) constructed by our proposed divergence. We investigate different latent dimensions and curvatures of poincaré spaces using the MNIST [22] dataset. We compare the quality of reconstructed images by comparing the reconstruction loss values, we compare the optimization efficiency and correctness by looking at the convergence of D_{KL}.

$$\log p(\mathbf{x}) \geq \underbrace{\mathbb{E}_{\mathbf{z}\sim q_\phi(\mathbf{z}|\mathbf{x})}[\log p_\theta(\mathbf{x}|\mathbf{z})] - \beta D_{\star\mathrm{KL}}(q_\phi(\mathbf{z}|\mathbf{x})\|p(\mathbf{z}))}_{\text{our new lower bound of ELBO}}, \tag{13}$$

which can be extended for Riemannian latent spaces by

$$\begin{aligned}\log p(\mathbf{x}) &= \log \int_{\mathcal{M}} p_\theta(\mathbf{x}|\mathbf{z}) p(\mathbf{z}) d\mathcal{M}(\mathbf{z}) \\ &= \log \int_{\mathcal{M}} \frac{p_\theta(\mathbf{x}|\mathbf{z})p(\mathbf{z})}{q_\phi(\mathbf{z}|\mathbf{x})} q_\phi(\mathbf{z}|\mathbf{x}) d\mathcal{M}(\mathbf{z}) \\ &\geq \int_{\mathcal{M}} \log \frac{p_\theta(\mathbf{x}|\mathbf{z})p(\mathbf{z})}{q_\phi(\mathbf{z}|\mathbf{x})} q_\phi(\mathbf{z}|\mathbf{x}) d\mathcal{M}(\mathbf{z}) \\ &= \mathbb{E}_{\mathbf{z}\sim q_\phi(\mathbf{z}|\mathbf{x})\mathcal{M}(\mathbf{z})}[\log p_\theta(\mathbf{x}|\mathbf{z}) - \log q_\phi(\mathbf{z}|\mathbf{x}) + \log p(\mathbf{z})] \\ &\geq \mathbb{E}_{\mathbf{z}\sim q_\phi(\mathbf{z}|\mathbf{x})\mathcal{M}(\mathbf{z})}[\log p_\theta(\mathbf{x}|\mathbf{z})] - \beta D_{\star\mathrm{KL}}(q_\theta(\mathbf{z}|\mathbf{x})\|p_\theta(\mathbf{z})).\end{aligned} \tag{14}$$

The $\mathcal{M}$ is a smooth manifold which is a set of point $\mathbf{z}$. The $D_{\star\mathrm{KL}}$ can be replaced by D_{KL}, D_{AKL} and D_{TKL}, for each of them, we evaluated the performance of two βs, *i.e.*, the trade off coefficient of the KL-divergence term against the reconstruction error term in ELBO on $\beta = 1$ and $\beta = 2$. More specifically, the β in Eq.(13) stand for the KKT [23], [24] multiplier for the Lagrangian of original ELBO function, which can be intepreted as a regularisation coefficient that constrains the capacity of the latent information channel $\mathbf{z}$ and puts implicit independence pressure on the learnt posterior [25], we change this paramter to adjust the degree of applied learning pressure during training so that we can test the ability of our proposed divergence on encouraging the latent representation learning in comparison with the vanilla KL-divergence.



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

Fig. 1. Comparison of reconstruction loss of optimizing $\mathcal{P}^c$-VAEs with KL-, AKL- and TKL-divergence vs dimensionality of the latent space.

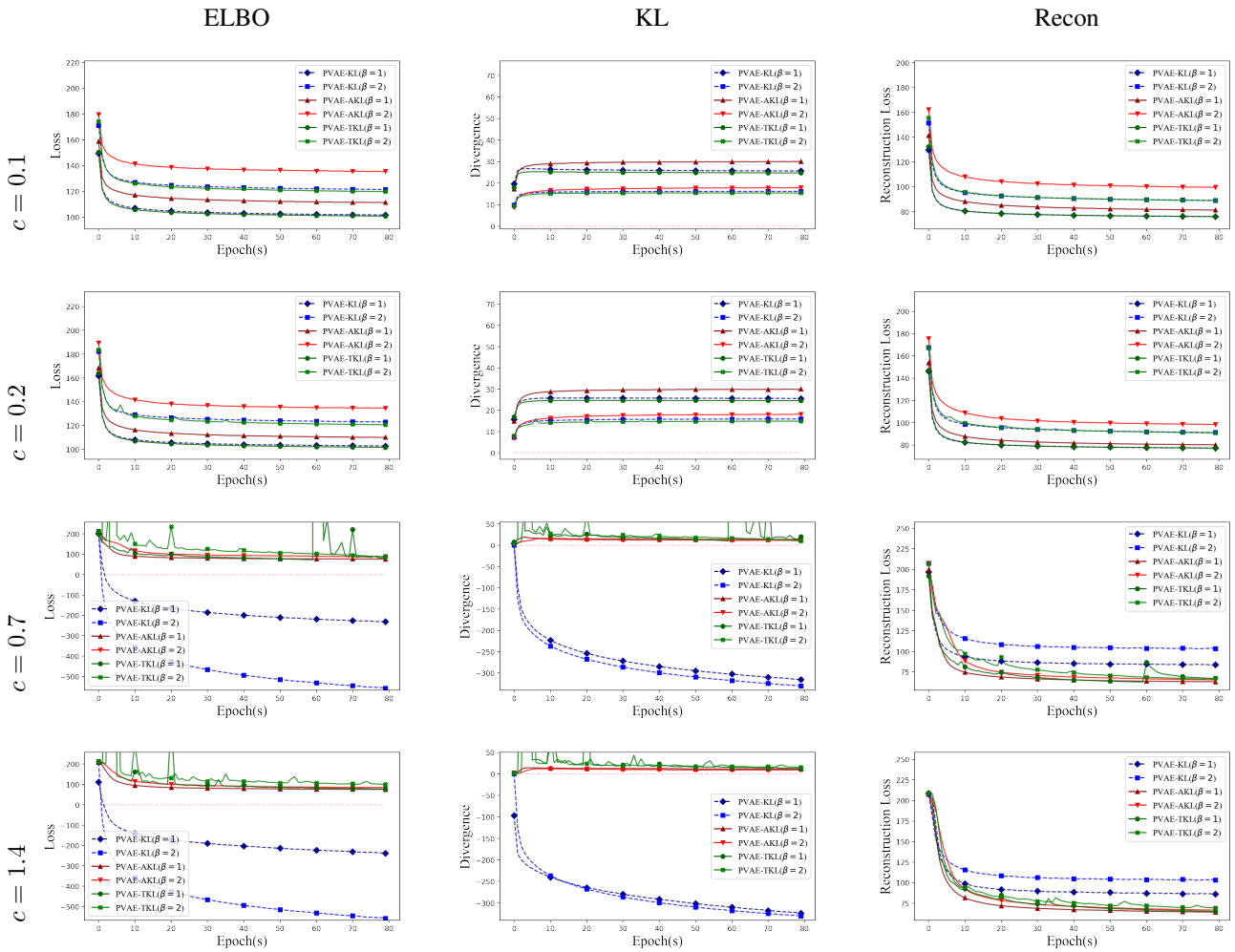


Fig. 2. Comparison of ELBO, reconstruction loss and divergence of optimizing $\mathcal{P}^c$-VAEs with KL-divergence and AKL-divergence and TKL-divergence, vs. curvature $c = 0.1, 0.2, 0.7, 1.4$ respectively. The latent dimension is fixed at 40.

1) Averaged Test Reconstruction Error: We conduct an ablation study on the usefulness of different divergence on higher dimensional latent space, to do so, we estimate the reconstruction loss for poíncare models achieved by different latent dimension where the curvature c of hyperbolic spaces is a fixed negative constant to get a wrapped manifold. As Fig.1(a) and Fig.1(b) illustrated, although using D_{AKL} and D_{TKL} as new lower bounds of ELBO seems unable to decrease reconstruction loss with lower dimensional latent space, as the latent spaces dimension become larger, the reconstruction loss of $\mathcal{P}^c$-VAE-KL start to rise slowly while those with D_{AKL} and D_{TKL} drop significantly on approximately latent dimension equal to 30, suggesting that optimizing the ELBO with D_{AKL} and D_{TKL} obtains increasing advantages over using D_{KL}.

2) Different Curvatures of Hyperbolic Manifold: We also investigated the performance as curvature c varies. We fix the latent dimension to 40. From Fig.2, we can observe that as curvature increases, MC estimation of D_{KL} reflects a downward trend toward negative approximated values, resulting in more noises and growing reconstruction loss. In contrast, greater curvature values lead to lower reconstruction loss for optimization with D_{AKL} and D_{TKL}.

The inaccuracy for KL estimation is specifically demonstrated by the negative KL estimation values, where we observe in the D_{KL} column of Fig.2 that the blue curve that represents the estimation D_{KL} by naïve estimator achieved negative value although the D_{KL} is by definition non-negative (the estimation below pink dashed line are considered as error), such results suggest serious problems with those estimators suffering from the uncertainty of non-negativity, especially for high-dimensional dependent variables and high curvature models.

With our proposed estimators, such miscalculation has improved, particularly in Fig.2 where $c = 0.7$ and $c = 1.4$, compared to $c = 0.1$ and $c = 0.2$, the red and dark-red curve indicating the D_{AKL} estimation, the green and dark-green curve indicating the D_{TKL} estimation are much better than the D_{KL} estimation (blue and dark-blue curve) by having a non-negative convergence in the divergence term and the overall ELBO loss term.

In average, with parameter $\beta = 1$, the model performs slightly better than those with $\beta = 2$, reflecting in a lower reconstruction loss, however, the divergence loss of $\beta = 2$ somehow outperforms those with $\beta = 1$, this is due to the loss of high frequency details when passing through a constrained latent bottleneck controlled by β, when $\beta > 1$, the it will put a

stronger constraint on the latent bottleneck than in the original VAE and thus force the model to learn better disentangled representation from input data, resulting in more informative $q_\phi(z|x)$ and thus approximate better to the ground truth.

V. Conclusion

Although estimating D_{KL} with MC methods is a commonly applied technique, such estimating can lead to unrealistic negative values of D_{KL}, whereas by definition, D_{KL} is non-negative. To fix such severe noises of MC, we propose Absolute-KL-divergence (D_{AKL}) and Tangent-KL-divergence (D_{TKL}) respectively, based on the absolute estimator (θ_{Abs}) and the tangent estimator (θ_{Tan}). The MC estimations of D_{AKL} and D_{TKL} are intrinsically non-negative, using MC for our proposed divergence is free from the previous unrealistic estimated results. We further show D_{AKL} satisfies the axioms of semi-metric. Therefore, D_{AKL} is a plausible quantity to measure the difference between two distributions. We also show the analogy of D_{TKL} toward the family of f-divergence. In practice, our derived divergence can help design VAEs. We replace KL-divergence with AKL- and TKL-divergence to derive a new lower bound of the traditional VAE ELBO. Via optimizing the new lower ELBO, two VAEs can have lower reconstruction loss values.

References

[1] A. Malinin and M. Gales, "Reverse kl-divergence training of prior networks: Improved uncertainty and adversarial robustness," in *Advances in Neural Information Processing Systems*, 2019, pp. 14 547–14 558.

[2] M. I. Belghazi, A. Baratin, S. Rajeshwar, S. Ozair, Y. Bengio, A. Courville, and D. Hjelm, "Mutual information neural estimation," in *International Conference on Machine Learning*, 2018, pp. 531–540.

[3] S. Nowozin, B. Cseke, and R. Tomioka, "f-gan: Training generative neural samplers using variational divergence minimization," in *Advances in neural information processing systems*, 2016, pp. 271–279.

[4] A. B. Dieng, D. Tran, R. Ranganath, J. Paisley, and D. Blei, "Variational inference via chi upper bound minimization," in *Advances in Neural Information Processing Systems*, 2017, pp. 2732–2741.

[5] J. Goldberger, S. Gordon, and H. Greenspan, "An efficient image similarity measure based on approximations of kl-divergence between two gaussian mixtures," in *ICCV*. IEEE, 2003, p. 487.

[6] J.-Y. Chen, J. R. Hershey, P. A. Olsen, and E. Yashchin, "Accelerated monte carlo for kullback-leibler divergence between gaussian mixture models," in *2008 IEEE International Conference on Acoustics, Speech and Signal Processing*. IEEE, 2008, pp. 4553–4556.

[7] J. R. Hershey and P. A. Olsen, "Approximating the kullback leibler divergence between gaussian mixture models," in *2007 IEEE International Conference on Acoustics, Speech and Signal Processing-ICASSP'07*, vol. 4. IEEE, 2007, pp. IV–317.

[8] D. J. MacKay and D. J. Mac Kay, *Information theory, inference and learning algorithms*. Cambridge university press, 2003.

[9] F. Nielsen, "Non-negative monte carlo estimation of f-divergences," 2020.

[10] D. P. Kingma and M. Welling, "Auto-encoding variational bayes," *arXiv preprint arXiv:1312.6114*, 2013.

[11] W. A. Wilson, "On semi-metric spaces," *American Journal of Mathematics*, vol. 53, no. 2, pp. 361–373, 1931.

[12] K. T. Abou-Moustafa and F. P. Ferrie, "A note on metric properties for some divergence measures: The gaussian case," in *Asian Conference on Machine Learning*, 2012, pp. 1–15.

[13] H. H. Bauschke and J. M. Borwein, "Joint and separate convexity of the bregman distance," in *Inherently Parallel Algorithms in Feasibility and Optimization and their Applications*, ser. Studies in Computational Mathematics, D. Butnariu, Y. Censor, and S. Reich, Eds. Elsevier, 2001, vol. 8, pp. 23–36.

[14] C. Lemieux, *Control Variates*. American Cancer Society, 2017, pp. 1–8.

[15] S. Ontañón, "An overview of distance and similarity functions for structured data," *Artificial Intelligence Review*, pp. 1–43, 2020.

[16] D. J. Galas, G. Dewey, J. Kunert-Graf, and N. A. Sakhanenko, "Expansion of the kullback-leibler divergence, and a new class of information metrics," *Axioms*, vol. 6, no. 2, p. 8, 2017.

[17] P. Glasserman, *Monte Carlo methods in financial engineering*. Springer Science & Business Media, 2013, vol. 53.

[18] Z. Botev and A. Ridder, "Variance reduction," *Wiley StatsRef: Statistics Reference Online*, pp. 1–6, 2014.

[19] A. B. Owen, *Monte Carlo theory, methods and examples*, 2013.

[20] A. Rényi *et al.*, "On measures of entropy and information," in *Proceedings of the Fourth Berkeley Symposium on Mathematical Statistics and Probability, Volume 1: Contributions to the Theory of Statistics*. The Regents of the University of California, 1961.

[21] E. Mathieu, C. Le Lan, C. J. Maddison, R. Tomioka, and Y. W. Teh, "Continuous hierarchical representations with poincaré variational auto-encoders," in *Advances in neural information processing systems*, 2019, pp. 12 565–12 576.

[22] Y. LeCun and C. Cortes, "MNIST handwritten digit database," 2010. [Online]. Available: http://yann.lecun.com/exdb/mnist/

[23] O. L. Mangasarian, *Nonlinear programming*. SIAM, 1994.

[24] G. Gordon and R. Tibshirani, "Karush-kuhn-tucker conditions," *Optimization*, vol. 10, no. 725/36, p. 725, 2012.

[25] I. Higgins, L. Matthey, A. Pal, C. Burgess, X. Glorot, M. Botvinick, S. Mohamed, and A. Lerchner, "beta-vae: Learning basic visual concepts with a constrained variational framework," 2016.