Implicit λ-Jeffreys Autoencoders: Taking the Best of Both Worlds

Aibek Alanov[1,2,*], Max Kochurov[1,3,*], Artem Sobolev[1], Daniil Yashkov[5], Dmitry Vetrov[2,3,4]

[1]Samsung AI Center in Moscow
[2]National Research University Higher School of Economics
[3]Skolkovo Institute of Science and Technology
[4]Joint Samsung-HSE lab
[5]FRC "Informatics and Management" of the Russian Academy of Sciences

November 1, 2019

Contents

Probability Distribution Divergences

Function $D(\cdot\|\cdot)$ is a divergence if

1. $D(p\|q) \geqslant 0 \quad \forall\, p, q$ distributions;
2. $D(p|q) = 0 \quad \Leftrightarrow \quad p = q$.

$p^*(x)$ - data distribution, $p_\theta(x)$ - model distribution.
Examples of divergences:

- Forward Kullback-Leibler (KL) divergence:

$$D_{\mathrm{KL}}(p^*(x)\|p_\theta(x)) = \mathbb{E}_{p^*(x)} \log \frac{p^*(x)}{p_\theta(x)}$$

- Reverse KL divergence:

$$D_{\mathrm{KL}}(p_\theta(x)\|p^*(x)) = \mathbb{E}_{p^*(x)} \log \frac{p^*(x)}{p_\theta(x)}$$

Probability Distribution Divergences

- Jensen-Shanon divergence:

$$\mathrm{JSD}(p^*(x)\|p_\theta(x)) = \frac{1}{2}D_{\mathrm{KL}}\left(p^*(x)\,\middle\|\,\frac{1}{2}(p^*(x)+p_\theta(x))\right) + \\ +\frac{1}{2}D_{\mathrm{KL}}\left(p_\theta(x)\,\middle\|\,\frac{1}{2}(p^*(x)+p_\theta(x))\right)$$

- λ-Jeffreys divergence:

$$\mathrm{J}_\lambda(p_\theta(x)\|p^*(x)) = \lambda D_{\mathrm{KL}}(p^*(x)\|p_\theta(x)) + \\ +(1-\lambda)D_{\mathrm{KL}}(p_\theta(x)\|p^*(x))$$

Generative Adversarial Networks (GANs)

GAN:

- generator $G_\theta(z)$, $z \sim p(z)$, $p_\theta(x) = \int \delta_{G_\theta(z)}(x) p(z) dz$;
- discriminator $D_\psi(x)$ classifies $p^*(x)$ vs $p_\theta(x)$.

Discriminator's objective:

$$\mathbb{E}_{p^*(x)} \log D_\psi(x) + \mathbb{E}_{p_\theta(x)} \log(1 - D_\psi(x)) \quad \rightarrow \quad \max_\psi$$

Generator's objective:

1. $-\mathbb{E}_{p_\theta(x)} \log(1 - D_\psi(x)) \quad \rightarrow \quad \max_\theta$
2. $\mathbb{E}_{p_\theta(x)} \log D_\psi(x) \quad \rightarrow \quad \max_\theta$
3. $\mathbb{E}_{p_\theta(x)} \log \dfrac{D_\psi(x)}{1 - D_\psi(x)} \quad \rightarrow \quad \max_\theta$

Generative Adversarial Networks (GANs)

Let $D_{\psi^*}(x) = \arg\max_{D} \left[\mathbb{E}_{p^*(x)} \log D_\psi(x) + \mathbb{E}_{p_\theta(x)} \log(1 - D_\psi(x))\right]$, then

1. $-\nabla_\theta \mathbb{E}_{p_\theta(x)} \log \frac{D_{\psi^*}(x)}{1 - D_{\psi^*}(x)} = \nabla_\theta D_{\mathrm{KL}}(p_\theta(x) \| p^*(x))$;
2. $\nabla_\theta \mathbb{E}_{p_\theta(x)} \log(1 - D_{\psi^*}(x)) = \nabla_\theta \mathrm{JSD}(p_\theta(x) \| p^*(x))$

It follows

$$\mathbb{E}_{p_\theta(x)} \log \frac{D_{\psi^*}(x)}{1 - D_{\psi^*}(x)} \to \max_\theta \quad \Leftrightarrow \quad D_{\mathrm{KL}}(p_\theta(x) \| p^*(x)) \to \min_\theta$$

$$-\mathbb{E}_{p_\theta(x)} \log(1 - D_{\psi^*}(x)) \to \max_\theta \quad \Leftrightarrow \quad \mathrm{JSD}(p_\theta(x) \| p^*(x)) \to \min_\theta$$

Variational Autoencoders (VAE)

VAE:

- generator $p_\theta(x|G_\theta(z)) = \mathcal{N}(x|G_\theta(z), \sigma I),\ z \sim p(z)$, $p_\theta(x) = \int p_\theta(x|G_\theta(z))p(z)dz$;
- encoder $q_\varphi(z|E_\varphi(x)) = \mathcal{N}(z|E^\mu_\varphi(x), E^\sigma_\varphi(x))$.

VAE's objective:

$$\theta^* = \arg\max_\theta \left[\max_\varphi \mathbb{E}_{p^*(x)}\mathbb{E}_{q_\varphi(z|x)} \log \frac{p_\theta(x|z)p(z)}{q_\varphi(z|x)}\right] =$$

$$= \arg\max_\theta \mathbb{E}_{p^*(x)} \log p_\theta(x) = \arg\max_\theta \left[-\mathbb{E}_{p^*(x)} \log \frac{p^*(x)}{p_\theta(x)}\right] =$$

$$= \arg\min_\theta D_{\mathrm{KL}}(p^*(x) \| p_\theta(x))$$

GAN and VAE objectives

GAN minimizes Reverse KL or JS divergence:

$$D_{\mathrm{KL}}(p_\theta(x) \| p^*(x)) \to \min_\theta \quad \text{or} \quad \mathrm{JSD}(p_\theta(x) \| p^*(x)) \to \min_\theta$$

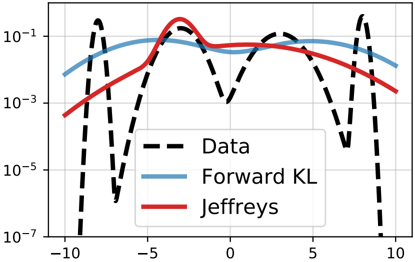
VAE minimizes Forward KL:

$$D_{\mathrm{KL}}(p^*(x) \| p_\theta(x)) \to \min_\theta$$

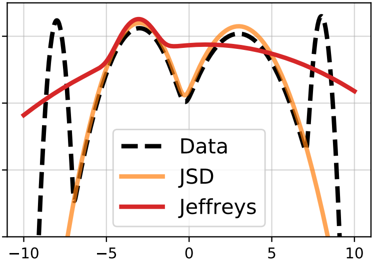
Divergence Properties: Toy Experiment

Toy example:

$$p^*(x) = 0.15\mathcal{N}(x|-8, 0.2^2) + 0.35\mathcal{N}(x|-3, 0.8^2)+$$
$$+0.3\mathcal{N}(x|3,1) + 0.2\mathcal{N}(x|8, 0.2^2),$$
$$p_\theta(x) = 0.5\mathcal{N}(x|\theta_1, \exp(\theta_2)) + 0.5\mathcal{N}(x|\theta_3, \exp(\theta_4))$$

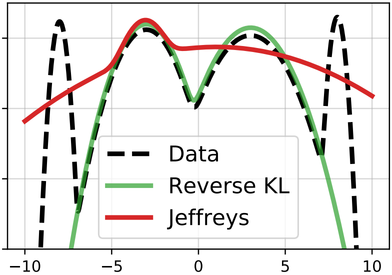


a) Jeffreys vs Forward KL

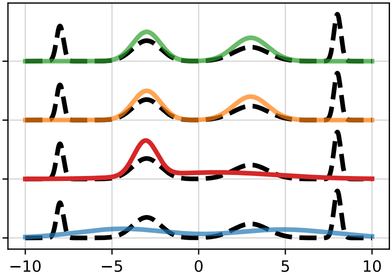


b) Jeffreys vs JSD

Divergence Properties: Toy Experiment



c) Jeffreys vs Reverse KL



d) Learned densities

Divergence Properties

Reverse KL and JS divergences lead to **mode-seeking** behaviour of $p_\theta(x)$:

- $p_\theta(x)$ captures some modes of $p^*(x)$, i.e. it can generate very realistic samples;
- $p_\theta(x)$ can ignore high value regions of $p^*(x)$.

Forward KL leads to **mass-covering** behaviour of $p_\theta(x)$:

- $p_\theta(x)$ captures all modes of $p^*(x)$;
- $p_\theta(x)$ covers low-probability regions of $p^*(x)$ as well.

Implicit λ-Jeffreys Autoencoder

We propose to minimize λ-Jeffreys divergence:

$$\mathrm{J}_\lambda(p_\theta(x)\|p^*(x)) = \lambda D_{\mathrm{KL}}(p^*(x)\|p_\theta(x)) + (1-\lambda) D_{\mathrm{KL}}(p_\theta(x)\|p^*(x))$$

We can balance between mode-seeking and mass-covering behaviours by adjusting the weight λ.

GAN part:

$$D_{\mathrm{KL}}(p_\theta(x)\|p^*(x)) \to \min_\theta \quad \Leftrightarrow \quad \mathbb{E}_{p_\theta(x)} \log \frac{D_{\psi^*}(x)}{1 - D_{\psi^*}(x)} \to \max_\theta$$

VAE part:

$$D_{\mathrm{KL}}(p^*(x)\|p_\theta(x)) \to \min_\theta \quad \Leftrightarrow$$
$$\Leftrightarrow \quad \mathbb{E}_{p^*(x)} \left[\mathbb{E}_{q_\varphi(z|x)} \log p_\theta(x|G_\theta(z)) - D_{\mathrm{KL}}(q_\varphi(z|x)\|p(z)) \right]$$

Implicit Conditional Likelihood

Standard choices for $p_\theta(x|G_\theta(z))$ are $\mathcal{N}(x|G_\theta(x), \sigma I)$ or $Laplace(x|G_\theta, \sigma I)$.

We propose a more general class of likelihoods - **symmetric likelihood** $r(x|y)$:

Definition

A density $r(\cdot|\cdot) : \mathcal{X} \times \mathcal{X} \to \mathbb{R}_+$ is a symmetric likelihood if

(*i*) $r(x = a|y = b) = r(x = b|y = a) \quad \forall\, a, b \in \mathcal{X}$;

(*ii*) $r(x = a|y = b)$ has a mode at $a = b$.

Examples: $\mathcal{N}(x|G_\theta(x), \sigma I)$ and $Laplace(x|G_\theta, \sigma I)$ are symmetric likelihoods.
Our model allows to train **implicit symmetric likelihoods**.

Implicit Conditional Likelihood

- Assume we are given implicit symmetric likelihood $r(y|x)$.
- We want to use it as $p_\theta(x|G_\theta(z))$, i.e. $p_\theta(x|G_\theta(z)) = r(x|G_\theta(z))$.
- Our aim is to compute $\nabla_\theta \mathbb{E}_{p^*(x)} \mathbb{E}_{q_\varphi(z|x)} \log r(x|G_\theta(z))$.

We introduce a discriminator $D_\tau(x, z, y)$ which classifies two types of triplets:

- real class: $(x, z, y) \sim p^*(x) q_\varphi(z|x) r(y|x)$;
- fake class: $(x, z, y) \sim p^*(x) q_\varphi(z|x) r'(y|G_\theta(z))$.

$$\begin{gathered}\mathbb{E}_{p^*(x) q_\varphi(z|x)} \left[\mathbb{E}_{r(y|x)} \log D_\tau(x, z, y) + \right. \\ \left. + \mathbb{E}_{r'(y|G_\theta(z))} \log(1 - D_\tau(x, z, y)) \right] \rightarrow \max_\tau \end{gathered} \tag{1}$$

Implicit Conditional Likelihood

Theorem

Let $D_{\tau^*}(x, z, y)$ be the optimal solution for the objective (1) and $r(y|x)$ and $r'(y|x)$ are symmetric likelihoods. Then $\nabla_\theta \mathbb{E}_{p^*(x)} \mathbb{E}_{q_\varphi(z|x)} \log \frac{D_{\tau^*}(x,z,G_\theta(z))}{1-D_{\tau^*}(x,z,G_\theta(z))} = \nabla_\theta \mathbb{E}_{p^*(x)} \mathbb{E}_{q_\varphi(z|x)} \log r(x|G_\theta(z))$.

We do not require an access to an analytic form of $r(y|G_\theta(z))$.

Choice of Symmetric Likelihood $r(y|x)$

It is an open question what is the best choice for the $r(y|G_\theta(z))$. Our expectations from $r(y|G_\theta(z))$:

- it should encourage realistic reconstructions;
- it should highly penalize for visually distorted images.

We chose as $r(y|x)$ a distribution over cyclic shifts in all directions of an image x. This distribution is symmetric with respect to all directions and has a mode in x, therefore it is the symmetric likelihood.

Although $r(y|x)$ is an explicit discrete distribution due to non-optimality of $D_\tau(x, z, y)$ the ratio $\log \frac{D_\tau(x,z,G_\theta(z))}{1-D_\tau(x,z,G_\theta(z))}$ sets *implicit likelihood* of reconstructions.

Implicit Encoder

The KL term $D_{\mathrm{KL}}(q_\varphi(z|x)\|p(z))$ from ELBO can be optimized adversarially using implicit $q_\varphi(z|x)$ defined by sampler $E_\varphi(x,\xi)$ where $\xi \sim \mathcal{N}(\cdot|0,I)$ [1].

We consider a disriminator $D_\zeta(x,z)$:

$$\mathbb{E}_{p^*(x)p(z)} \log D_\zeta(x,z) + \mathbb{E}_{p^*(x)q_\varphi(z|x)} \log(1 - D_\zeta(x,z)) \;\to\; \max_\zeta$$

Then

$$-\nabla_\varphi \mathbb{E}_{q_\varphi(z|x)} \log \frac{D_\zeta(x,z)}{1 - D_\zeta(x,z)} = \nabla_\varphi D_{\mathrm{KL}}(q_\varphi(z|x)\|p(z))$$

Final Objectives

$$
\begin{aligned}
\mathcal{L}_{\lambda\text{-IJAE}}(\theta, \varphi) =& (1-\lambda) D_{\mathrm{KL}}(p_\theta(x) \| p^*(x)) - \lambda \mathcal{L}_{\mathsf{ELBO}}(\theta, \varphi) = \\
=& -(1-\lambda)\mathbb{E}_{p_\theta(x)} \log \frac{D_{\psi^*}(x)}{1 - D_{\psi^*}(x)} - \\
& - \lambda \mathbb{E}_{p^*(x)} \mathbb{E}_{q_\varphi(z|x)} \left[\log \frac{D_{\tau^*}(x, z, G_\theta(z))}{1 - D_{\tau^*}(x, z, G_\theta(z))} + \right. \\
& \left. + \log \frac{D_{\zeta^*}(x, z)}{1 - D_{\zeta^*}(x, z)} \right] \quad \to \quad \min_{\theta, \varphi}
\end{aligned}
$$

Final Objectives

$$\mathcal{L}_G(\theta) = -(1-\lambda)\mathbb{E}_{p_\theta(x)} \log \frac{D_\psi(x)}{1 - D_\psi(x)} -$$

$$-\lambda \mathbb{E}_{p^*(x)} \mathbb{E}_{q_\varphi(z|x)} \log \frac{D_\tau(x, z, G_\theta(z))}{1 - D_\tau(x, z, G_\theta(z))} \rightarrow \min_\theta$$

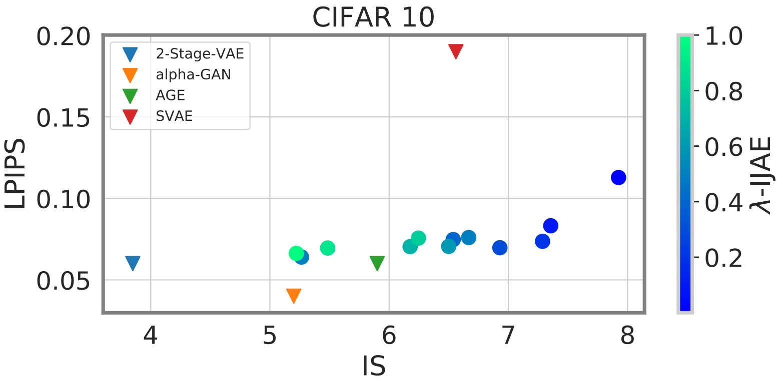
$$\mathcal{L}_E(\varphi) = -\lambda \mathbb{E}_{p^*(x)} \mathbb{E}_{q_\varphi(z|x)} \left[\log \frac{D_\tau(x, z, G_\theta(z))}{1 - D_\tau(x, z, G_\theta(z))} + \right.$$

$$\left. + \log \frac{D_\zeta(x, z)}{1 - D_\zeta(x, z)} \right] \rightarrow \min_\varphi$$

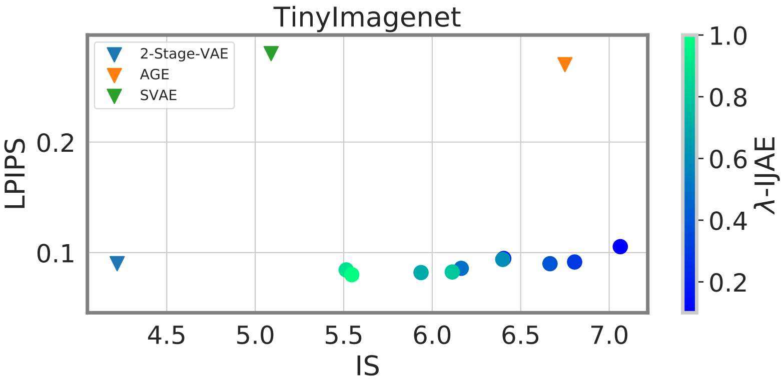
Experiment Results: Evaluation

- We evaluate our model on both generation and reconstruction tasks.
- The quality of the former is assessed using Inception Score (IS) and Fréchet Inception Distance (FID).
- The reconstruction quality is evaluated using LPIPS. It was show that LPIPS is a good metric which captures perceptual similarity between images.

Results on CIFAR-10



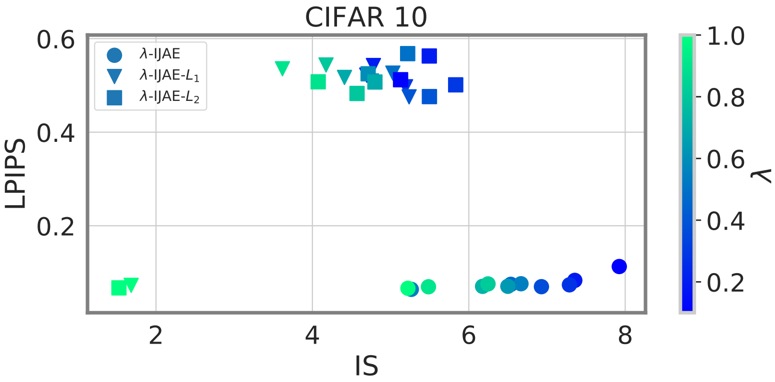
Results on TinyImageNet



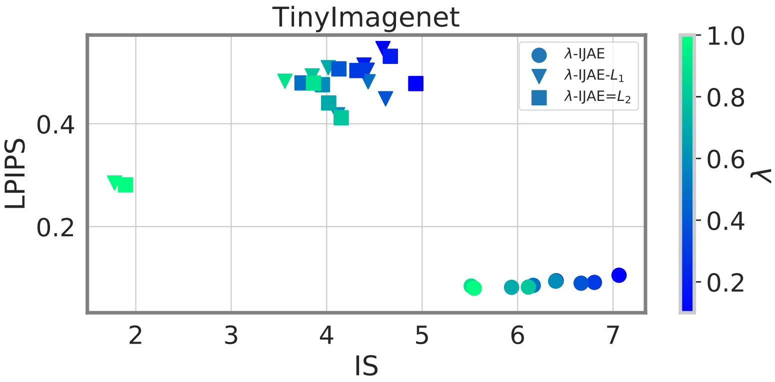
Results on CIFAR-10 and TinyImageNet

Method	Generation Quality FID ↓	IS ↑	Reconstruction Quality LPIPS ↓
CIFAR 10			
WAE (Tolstikhin et al., 2017)	87.7	4.18 ± 0.04	
ALI (Dumoulin et al., 2017))		5.34 ± 0.04	
ALICE (Li et al., 2017)		6.02 ± 0.03	
AS-VAE (Pu et al., 2017b)		6.3	
VAE (resnet)	150.3	3.45 ± 0.02	0.09 ± 0.03
2S-VAE (Dai & Wipf, 2019)	94.53	3.85 ± 0.03	0.06 ± 0.03
α-GAN (Rosca et al., 2017)	54.98	5.20 ± 0.08	0.04 ± 0.02
AGE (Ulyanov et al., 2018)	39.13	5.90 ± 0.04	0.06 ± 0.02
SVAE (Chen et al., 2018)	44.73	6.56 ± 0.07	0.19 ± 0.08
λ-IJAE ($\lambda = 0.3$)	**29.46**	$\mathbf{6.98} \pm \mathbf{0.09}$	0.07 ± 0.03
TinyImagenet			
AGE (Ulyanov et al., 2018)	39.51	6.75 ± 0.09	0.27 ± 0.09
SVAE (Chen et al., 2018)	79.50	5.09 ± 0.05	0.28 ± 0.08
2Se-VAE (Dai & Wipf, 2019)	72.90	4.22 ± 0.05	$\mathbf{0.09} \pm \mathbf{0.05}$
λ-IJAE ($\lambda = 0.3$)	**35.49**	$\mathbf{6.85} \pm \mathbf{0.06}$	$\mathbf{0.11} \pm \mathbf{0.04}$

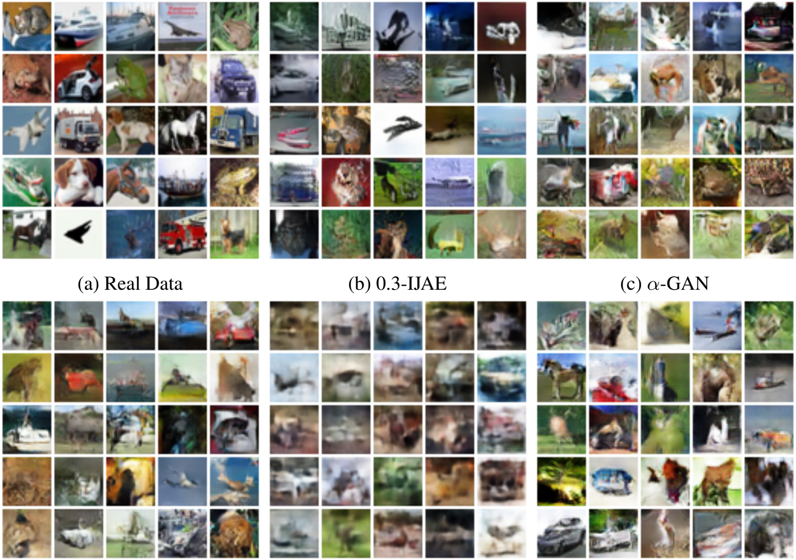
Ablation Study



Ablation Study



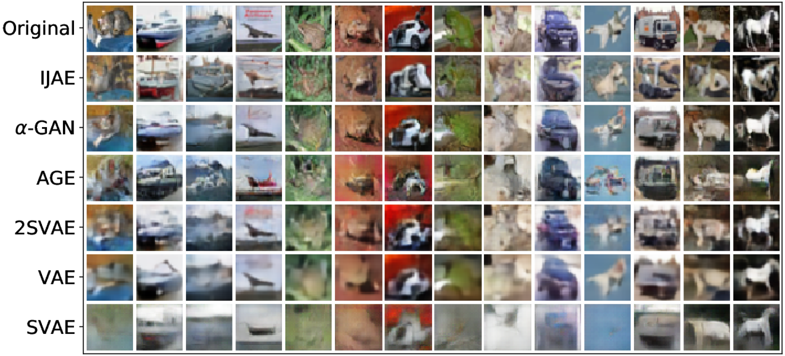
CIFAR10 Samples



(a) Real Data (b) 0.3-IJAE (c) α-GAN

(d) AGE (e) TwoStage-VAE (f) SVAE

CIFAR10 Reconstructions



Conclusion

- We propose a novel auto-encoding generative model
- We provide a theoretical analysis of our objective and show that it is equivalent to the λ-Jeffreys divergence.
- In experiments, we demonstrate that our model achieves the state-of-the-art balance between generation and reconstruction quality
- It confirms our assumption that the λ-Jeffreys divergence is the right choice for learning complex high-dimensional distributions in the case of the limited capacity of the model