Rényi Divergence Variational Inference

Arne Gebert, Wei-Chen Lee, Evan Neill, Rohan Saphal
University of Oxford Department of Computer Science
{arne.gebert, wei-chen.lee, evan.neill, rohan.saphal}@cs.ox.ac.uk

Introduction

Variational Inference (VI) is a means of approximating posterior probability densities $p(z|x)$ in a Latent Variable Model (LVM) wherein z is a latent variable representation of data x. This is achieved by minimizing a *divergence* measuring the dissimilarity between the parameterized proposed approximation $q_\psi(z)$ and the posterior $p(z|x)$. A typical and performant choice of divergence is the Kullback-Leibler (KL) divergence. The paper Rényi Divergence Variational Inference [1] introduces the parametrized family of **Rényi's α-divergences**:

$$D_\alpha(p(x)||q(x)) := \frac{1}{\alpha - 1} \log \mathbb{E}_p \left[\left(\frac{p(x)}{q(x)} \right)^{\alpha - 1} \right]$$

Rényi's α-divergences include multiple commonly used divergence functions as special cases, including the KL divergence for $\alpha \to 1$. Like in traditional variational inference this divergence can be used to derive an **optimizable bound** $\mathcal{L}_\alpha$ which can be approximated:

$$\mathcal{L}_\alpha = \frac{1}{1-\alpha} \log \mathbb{E}_q \left[\left(\frac{p_\theta(x,z)}{q_\psi(z|x)} \right)^{1-\alpha} \right] \approx \frac{1}{1-\alpha} \log \frac{1}{K} \sum_{k=1}^{K} \left(\frac{p_\theta(x,z_i)}{q_\psi(z_i|x)} \right)^{1-\alpha} = \mathcal{L}_{\alpha,K}$$

Applying Rényi α-divergences to VAEs

The **reparameterization trick** is applied to able to use $\hat{\mathcal{L}}_{\alpha,K}$ in the VAE framework:

$$\hat{\mathcal{L}}_{\alpha,K} = \frac{1}{1-\alpha} \log \frac{1}{K} \sum_{i=1}^{K} \left[\left(\frac{p(x, z_{i,\epsilon})}{q(z_{i,\epsilon}|x)} \right)^{1-\alpha} \right]$$

with $z_\epsilon = \mu + \Sigma\epsilon$ where $\epsilon \sim \mathcal{N}(0, \mathcal{I})$ and $\mathcal{I}$ is the identity matrix.
Special cases: $\hat{\mathcal{L}}_{\alpha,K}$ recovers the objective of the 'vanilla' VAE [2] with $\alpha \to 1$ and IWAE [3] with $\alpha = 0$.
The proposes altering the optimization of $\hat{\mathcal{L}}_{\alpha,K}$ with an algorithm termed VR-α algorithm which only needs to backpropagate one sample per K Monte-Carlo samples:

Algorithm 1: One step of the VR-α algorithm

1: sample $\epsilon_1, ..., \epsilon_K \sim \mathcal{N}(0, \mathcal{I})$
2: for $i = 1, ..., K$ compute: $\log \hat{w}_i := \log \frac{p_\theta(x, z_{i,\epsilon})}{q_\psi(z_{i,\epsilon}|x)}$
3: build Multinomial distribution $Mul(w)$ weighted according to $\log \hat{w}_i^{(1-\alpha)}$
4: sample one $\log \hat{w}_j^{(1-\alpha)} \sim Mul(w)$
5: backpropagate $\log \hat{w}_j^{(1-\alpha)}$

When $\alpha = -\infty$, the sample with largest unnormalised importance weight $\log \hat{w}_i$ is always chosen to backpropagate. In this case, the algorithm is termed **VR-max**.

Replication of VAE results

Implementation: We reimplement VAE, IWAE, and the new VR-α and VR-max in PyTorch. We then adapt the training procedure to reproduce the results with reduced computational resources:

Dataset	K	VAE	IWAE	VR-max	VR-0.5
Caltech 101 Silhouettes	5	-108.51	-107.61	**-107.00**	-107.09
		(-119.69)	**(-117.89)**	(-118.01)	
	50	-109.47	-106.99	**-106.09**	-106.51
		(-119.61)	(-117.21)	**(-117.10)**	
MNIST	5	-88.83	-87.33	**-87.20**	-87.61
		(-86.47)	**(-85.41)**	(-85.42)	
	50	-89.25	**-86.08**	-86.27	-86.65
		(-86.35)	**(-84.80)**	(-84.81)	

Table 1:Our test negative log likelihoods compared to [1]'s (in parentheses)

Analysis: We are generally able to reproduce the paper's results.

- Like in [1], VR-max and IWAE perform almost indistinguishably and better than VAE.
- Based on the tightness of the respective bounds VAE $<$ VR-0.5 $<$ IWAE. We show that this corresponds to the analogue result in relative performance.
- We were able to improve the NLL baseline on Silhouettes by almost 10% due to hyperparameter tuning.

Further investigations on VAEs

The authors show in a toy example how large positive and large negative α respectively forces the posterior approximation to mode-seek or mass-cover. We attempt to uncover this behaviour in a VAE experiment:

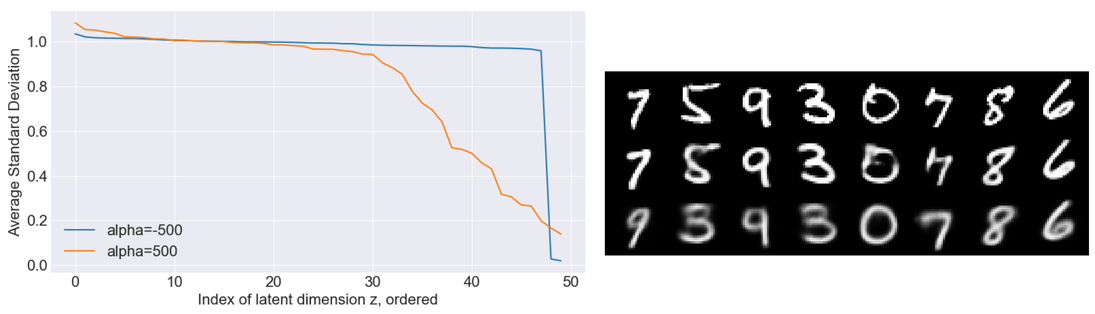

Figure 1: Left: mean standard deviations of latent dimension on two MNIST models. Right: their reconstructions (original sample top, $\alpha = -500$ middle, $\alpha = 500$ bottom)

Analysis: We successfully recover this behaviour.

- Standard deviations around 1 indicate maximizing the mass under the prior distribution $\mathcal{N}(\mathbf{0}, \mathcal{I})$, while standard deviations under 1 indicate preventing sampling from distal regions of those dimensions.
- Where there is a difference, it indicates there is a region of some dimension for each model where the $\alpha = 500$ model can't always create good reconstructions and won't risk expansion to that region, while $\alpha = -500$ sees that some reconstructions are possible and tries to cover it.

Tuning alpha: We further investigate the influence of α on performance:

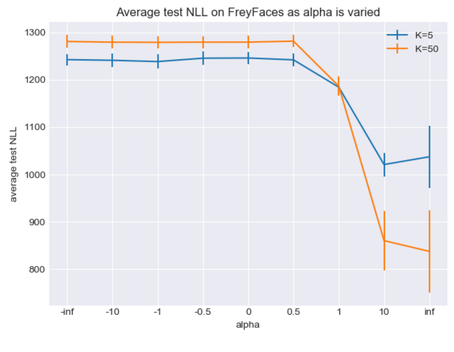

Figure 2: Performance of VR-α dependent on α and K

Analysis: We find that the VAE framework is largely invariant to the choice of α, as long as $\alpha \leq 0$. The choice of K is more impactful than the choice of α.

Applying Rényi α-divergences to BNNs

In addition to Monte Carlo approximation, a mini-batching technique, '**energy approximation**', is used to efficiently train the Bayesian neural network (BNN), defined by the following approximation (termed black box-α [4]) of the VR bound:

$$\mathcal{L}_{BB-\alpha} = \frac{1}{1-\alpha} \log \mathbb{E}_q \left[\left(\frac{p(w)(\prod_i^M p(x_i|w))^{N/M}}{q(w)} \right)^{1-\alpha} \right]$$

where M is the mini-batch size and N the size of the training set. This approximation effectively scales the 'average likelihood' of the mini-batch to the training size N.

Replication of BNN results

Analysis: Our replication of the BNN experiments produces largely consistent average test root-mean-squared-error (RMSE) and average test negative log likelihood (NLL) compared to the paper, except for smaller datasets such as Boston.

Dataset	$\alpha \to -\infty$	$\alpha = 0$	$\alpha = 0.5$	$\alpha = 1$ (VI)	$\alpha \to \infty$
power	4.12±0.04	4.17±0.04	4.09±0.03	**4.05±0.03**	4.08±0.03
	(4.08±0.03)	(4.10±0.04)	**(4.07±0.04)**	(4.07±0.04)	(4.08±0.04)
protein	4.56±0.01	4.52±0.03	4.56±0.03	**4.51±0.02**	4.55±0.02
	(4.57±0.05)	**(4.44±0.03)**	(4.51±0.03)	(4.45±0.02)	(4.45±0.01)
wine	0.63±0.01	0.63±0.01	0.64±0.01	0.63±0.01	**0.63±0.01**
	(0.64±0.01)	(0.64±0.01)	(0.64±0.01)	(0.63±0.01)	**(0.63±0.01)**
power	2.82±0.01	2.82±0.01	**2.81±0.01**	2.82±0.01	2.83±0.01
	(2.82±0.01)	(2.83±0.01)	**(2.82±0.01)**	(2.82±0.01)	(2.83±0.01)
protein	2.94±0.00	2.93±0.00	2.94±0.01	**2.93±0.00**	2.94±0.00
	(2.94±0.01)	**(2.91±0.00)**	(2.92±0.01)	(2.91±0.00)	(2.91±0.00)
wine	**0.95±0.01**	0.95±0.01	0.96±0.01	0.96±0.01	0.96±0.01
	(0.95±0.01)	(0.95±0.01)	**(0.95±0.01)**	(0.96±0.01)	(0.97±0.01)

Table 2: BNN regression replication: Average test RMSE (top) and NLL (bottom), ± standard error, on selected datasets. Lowest mean value for each dataset shown in bold. Original results shown in parentheses.

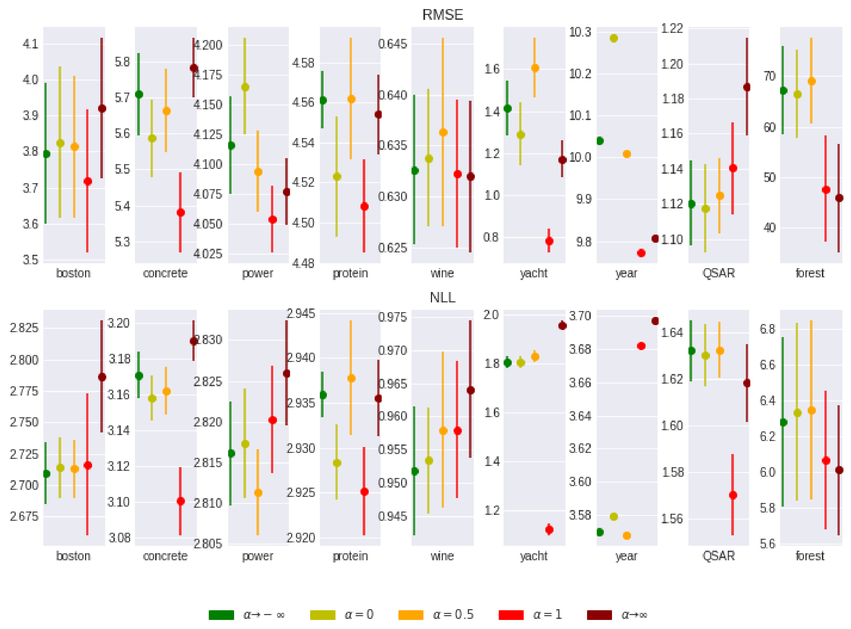

Figure 3: BNN replication results for 7 out of 10 datasets in the paper, plus two additional datasets QSAR and Forest

Further investigations on BNNs

The paper's main hypotheses are:

1. Models with high α values minimise RMSE while models with low α values minimise NLL.
2. Monte Carlo approximation biases the model towards $\alpha = 1$ (VI), so the effective optimal α is further from 1 than the true optimal α. The bias $\to 0$ as $K \to \infty$.

Analysis:

- No strong evidence to support the first hypothesis. Figure 3 shows that $\alpha = 1$ (VI) often performs well against other values of α for both RMSE and NLL, though there are clear exceptions.
- Some evidence to support the second. Figure 4 shows that when the true optimal α is 1, increasing K increases the performance of VI relative to other values of α. However, when the true optimal appears to be negative (bottom left), we do not observe strong bias in favour of negative α values when K is small.
- More extensive testing is required to establish relationships between α, performance measures (RMSE and NLL), experimental parameters (e.g. Monte Carlo sample size K and mini-batch size M), and dataset characteristics (e.g. uni-modal or multi-modal).

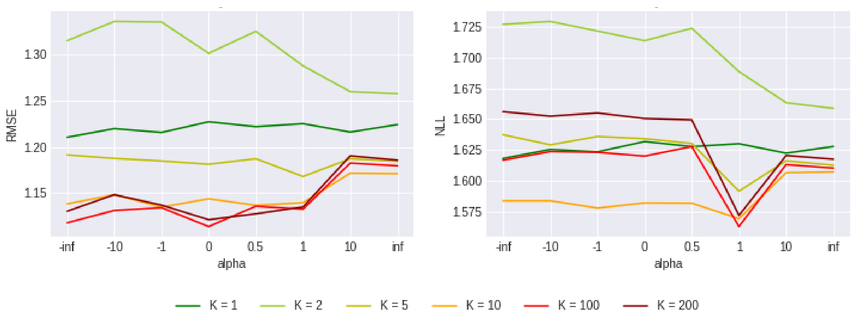

Figure 4: Impact of K on optimal α, for dataset QSAR

Conclusions

- **Replication of results:** We are able to generally replicate and validate the results of the paper with moderate computational resources.
- **Application to VAE:** IWAEs perform at least as well as VR-α without any trade-offs calling the practical usability into question.
- **Application to BNN:** We find that $\alpha = 1$ (VI) generally performs well against other values of α, though there are clear exceptions. Extensive testing is required to better understand the circumstances under which alternative values of α would outperform VI.

References

[1] Yingzhen Li and Richard E. Turner. Rényi divergence variational inference, 2016.
[2] Diederik P Kingma and Max Welling. Auto-encoding variational bayes, 2013.
[3] Yuri Burda, Roger Grosse, and Ruslan Salakhutdinov. Importance weighted autoencoders, 2015.
[4] José Miguel Hernández-Lobato, Yingzhen Li, Mark Rowland, Daniel Hernández-Lobato, Thang Bui, and Richard E. Turner. Black-box -divergence minimization, 2015.