A Tighter Monte Carlo Objective with Rényi α-divergence Measures

Stefan Webb
University of Oxford
stefanw@robots.ox.ac.uk

Yee Whye Teh
University of Oxford
y.w.teh@stats.ox.ac.uk

Abstract

Deep generative models (DGMs) offer the promise of improved unsupervised and semi-supervised learning, and could overcome current limitations of established deep learning methods such as by being robust to adversarial examples and being able to learn from smaller data sets. However, progress in learning is required before these methods become practical. To these ends, we propose to combine multi-sample, or rather Monte Carlo, objectives with the Rényi α-divergence measure to obtain an improved bound on the log-likelihood.

1 Introduction

The goal of unsupervised learning is often to learn a distribution over the input data. Models such as the variational autoencoder ([KW13]) and sigmoid belief network (SBN) assume that the data, $\mathbf{x}$ is generated by a number of independent latent causes, $\mathbf{h}$, and simultaneously learn the parameters to the model, $p(\mathbf{x}, \mathbf{h})$, and an approximation to the posterior, $q(\mathbf{h} \mid \mathbf{x})$, by optimizing a bound to the intractable log-likelihood of the inputs.

Learning methods for these models take as their starting point the standard variational inference bound (i.e. the ELBO),

$$\mathcal{L}(q\ ;\ \mathbf{x}) = \mathbb{E}_q\left[\ln\left(\frac{p_\phi(\mathbf{x}, \mathbf{h})}{q_\psi(\mathbf{h} \mid \mathbf{x})}\right)\right].$$

and develop methods for reducing the variance over the naive gradient of this bound based on the log-derivative trick (which is too great to be practical).

Further work has involved modifying the variational objective to improve learning. For instance, [BGS16] showed a tighter bound on the log-likelihood can be obtained by using multiple samples inside the expectation,

$$\mathcal{L}_K(q\ ;\ \mathbf{x}) = \mathbb{E}_q\left[\ln\left(\frac{1}{K}\sum_{k=1}^{K}\frac{p_\phi(\mathbf{x}, \mathbf{h}_k)}{q_\psi(\mathbf{h}_k \mid \mathbf{x})}\right)\right].$$

They prove that the bound approaches the log-likelihood in the limit as K goes to infinity. Intuitively, taking more samples inside the sum lessens the need for a single sample to explain the posterior well. [MR16] develop a control variate for this multi-sample objective with the log-derivative trick so that the technique can be used with discrete valued variables.

In other work, [LT16] investigated replacing the KL-divergence with Rényi's α-divergence measure to obtain the bound,

$$\mathcal{L}_\alpha(q\ ;\ \mathbf{x}) = \frac{1}{1-\alpha}\ln\mathbb{E}_q\left[\left(\frac{p_\phi(\mathbf{x}, \mathbf{h})}{q_\psi(\mathbf{h} \mid \mathbf{x})}\right)^{1-\alpha}\right].$$

Workshop on Bayesian Deep Learning, NIPS 2016, Barcelona, Spain.

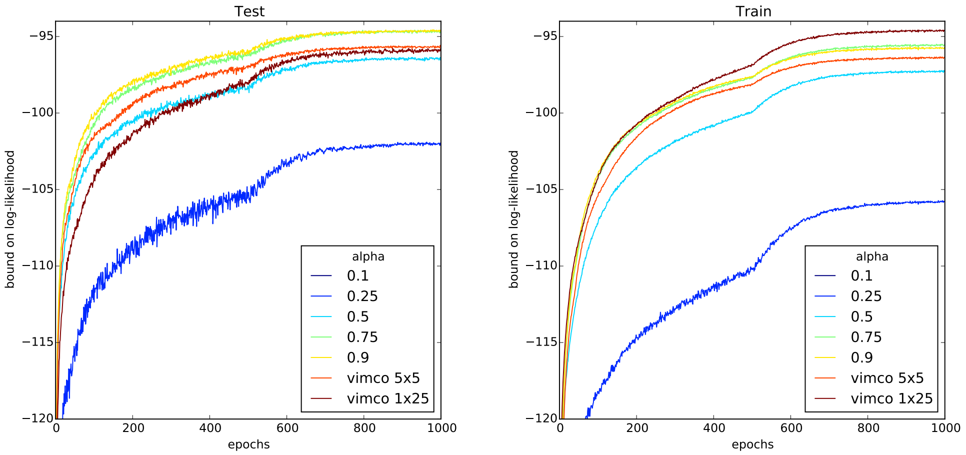

Figure 1: Varying the value of α for α-VIMCO and comparing to standard VIMCO.

They show that the ELBO is recovered in the limit $\alpha \to 1$, and as $\alpha \to 0$ the bound on the log-likelihood is tightened.

We propose to combine the two techniques by using the following multi-sample Rényi bound,

$$\mathcal{L}_{\alpha,K}(q\,;\,\mathbf{x}) = \frac{1}{1-\alpha}\ln \mathbb{E}_q\left[\left(\frac{1}{K}\sum_{k=1}^{K}\frac{p_\phi(\mathbf{x},\mathbf{h}_k)}{q_\psi(\mathbf{h}_k \mid \mathbf{x})}\right)^{1-\alpha}\right].$$

and conjecture that the resulting method is superior to either alone. We follow the approach of [MR16], termed VIMCO, and use the log-derivative trick with a control variate to form a Monte Carlo estimate of the gradient, which we call α-VIMCO. Our estimate has the same advantage of VIMCO in that it can be used with both discrete and continuous variables, unlike those based on the reparametrization trick, but has two ways of tightening the bound. In order to calculate the estimate of the gradient, samples $\{\mathbf{h}_{j,k}\}$ must be taken over a second dimension j, which is used to estimate an expectation over the sum.

2 Experiments

In preliminary experiments, α-VIMCO was compared to standard VIMCO for unsupervised learning on the MNIST data set. For α-VIMCO, we used five samples of size five, and for VIMCO both five samples of size five, and one sample of size twenty-five.

The generative model used for $p(\mathbf{x},\mathbf{h})$ was a sigmoid belief network (SBN). The latent space comprised a single hidden layer of 200 binary valued variables. The prior, $p(\mathbf{h})$, on the latent variables defined them to be independently Bernoulli distributed with probability 0.5. For the decoder $p(\mathbf{x} \mid \mathbf{h})$, the value of $\mathbf{h}$ fed through three densely connected deterministic layers of 200 variables each with ReLU activations and batch normalization, followed by a logistic activation to give the parameters to the distribution over the stochastic input $\mathbf{x}$ of 784 binary features. The proposal, $q(\mathbf{h} \mid \mathbf{x})$ used an analogous architecture with the edges reversed.

The bound was optimized with SGD and an Adam adaptive stepsize scheme ([KB14]). The gradient was normalized so that its sum did not exceed 5. Then the individual elements were clipped to $[-1, 1]$. Learning proceeded for 1000 epochs, with the Adam learning rate being annealed after 500 epochs by decreasing it by a percent every subsequent epoch. The bound estimate was reported both on the training and test sets every epoch. See Figure 1.

We see that for moderate values of α, that is, $\alpha = 0.9$ and $\alpha = 0.75$, α-VIMCO obtains a higher bound than VIMCO on the test set. For lower values of α it underperforms and this may be due to increased Monte Carlo noise or the more mass-covering (as opposed to mode-seeking) nature of the objective. Samples were drawn from all parameter settings after 1000 epochs—it was hard to tell any qualitative differences apart from the case $\alpha = 0.1$. Further results are available upon request.

3 Discussion

Further work is required to understand the relationship of the variance of the Monte Carlo estimate of the bound as α is varied. Also, further experiments are necessary to determine the effect of varying the sample size in both dimensions, and to evaluate deeper architectures with more stochastic layers on other data sets.

We hypothesize that the full benefits of α-VIMCO become apparent with deeper architectures and increased number of samples. It would also be interesting to compare the performance of the algorithm with $\alpha = 0.5$ to the method of [BSFB16], in which a Bhattacharyya distance is minimized (the case $\alpha = 0.5$ is a function of the Hellinger distance, which is closely related to the Bhattacharyya distance).

References

[BGS16] Yuri Burda, Roger Grosse, and Ruslan Salakhutdinov. Importance weighted autoencoders. *arXiv preprint arXiv:1509.00519v3 [cs.LG]*, 2016.

[BSFB16] Jorg Bornschein, Samira Shabanian, Asja Fischer, and Yoshua Bengio. Bidirectional helmholtz machines. *arXiv preprint arXiv:1506.03877v5 [cs.LG]*, 2016.

[KB14] Diederik Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.

[KW13] Diederik P Kingma and Max Welling. Auto-encoding variational bayes. *arXiv preprint arXiv:1312.6114v10 [stat.ML]*, 2013.

[LT16] Yingzhen Li and Richard E. Turner. Rényi divergence variational inference. *arXiv preprint arXiv:1602.02311v3 [stat.ML]*, 2016.

[MR16] Andriy Mnih and Danilo J. Rezende. Variational inference for Monte Carlo objectives. *arXiv preprint arXiv:1602.06725v2 [cs.LG]*, 2016.