# Black-box $\alpha$-divergence for Deep Generative Models

**Thang D. Bui**[*]
University of Cambridge
tdb40@cam.ac.uk

**Daniel Hernández-Lobato**[*]
Universidad Autónoma de Madrid
daniel.hernandez@uam.es

**José Miguel Hernández-Lobato**[*]
University of Cambridge
jmh233@cam.ac.uk

**Yingzhen Li**
University of Cambridge
yl494@cam.ac.uk

**Richard E. Turner**
University of Cambridge
ret26@cam.ac.uk

## Abstract

We propose using the black-box $\alpha$-divergence [1] as a flexible alternative to variational inference in deep generative models. By simply switching the objective function from the variational free-energy to the black-box $\alpha$-divergence objective we are able to learn better generative models, which is demonstrated by a considerable improvement of the test log-likelihood in several preliminary experiments.

## 1 Generative models and inference networks

We consider a probabilistic model for $N$ $D$-dimensional observations $\mathbf{x} = \{x_n\}_{n=1}^N$ and assume $K$-dimensional continuous latent variables $\mathbf{z} = \{z_n\}_{n=1}^N, z_n \in \mathbb{R}^K$ as follows,

$$p(\mathbf{z}) = \mathcal{N}(\mathbf{z}; \mathbf{0}, \mathbf{I}) \tag{1}$$

$$p(\mathbf{x}|\mathbf{z}, \theta) = \prod_n p(x_n|z_n, \theta) \tag{2}$$

where $p(x_n|z_n, \theta)$ is typically Gaussian $\mathcal{N}(x_n; f_\theta(z_n), \sigma_x^2\mathbf{I})$ if $x_n \in \mathbb{R}^D$, or Bernoulli $\mathcal{B}(x_n; \Phi(f_\theta(z_n)))$ with a sigmoidal link function $\Phi(\cdot)$ for binary data $x_n$, and $f_\theta$ is a deterministic mapping parameterised by $\theta$. Exact Bayesian inference in this model class, in general, is intractable, and we have to resort to approximate inference schemes. One popular deterministic approximate inference technique is variational inference (VI), which is the first step towards turning the aforementioned intractable inference problem into a tractable optimisation problem. By introducing a variational approximation $q_\gamma(\mathbf{z})$, a negative variational free-energy can be obtained, which is a lower bound to the log marginal likelihood of the observed data,

$$\mathcal{L}(\theta) = \log p(\mathbf{x}|\theta) \geq \int \mathrm{d}\mathbf{z} q_\gamma(\mathbf{z}) \log \frac{p(\mathbf{x}, \mathbf{z}|\theta)}{q_\gamma(\mathbf{z})} = \mathcal{F}_{\mathrm{VFE}}(\theta, \gamma), \tag{3}$$

where $\gamma$ are the variational parameters. For continuous latent variables, a typical choice of the approximate distribution is a diagonal Gaussian $q_\gamma(\mathbf{z}) = \prod_n q_\gamma(z_n)$, i.e. mean-field between latent variables[2]. The variational lower bound can be decomposed further,

$$\mathcal{F}(\theta, \gamma) = -\mathrm{KL}(q_\gamma(\mathbf{z})||p(\mathbf{z})) + \sum_{n=1}^N \int \mathrm{d}z_n q_\gamma(z_n) \log p(x_n|z_n). \tag{4}$$

In complex generative models, evaluating the expectations of the log likelihood terms, and hence, $\mathcal{F}$ and its gradients, is not tractable. However, they can be approximately computed using the *log-derivative* trick or Monte Carlo with the *reparameterisation* trick [2, 3]. Additionally, the standard

[2] and, in general, also between latent dimensions

VI formalism requires $N$ sets of parameters to be optimised. This can be avoided using a recognition model or inference network, $q_\gamma(z_n) = \mathcal{N}(z_n; \mu_\gamma(x_n), \sigma^2_\gamma(x_n))$, where $\gamma$ parameterises the mapping from the observed data $x_n$ to the mean and variance of $q$. Because the objective is a variational free energy, and the architecture involves a generative mapping (decoder) and a recognition mapping (encoder), this class of models is often referred to as the variational auto-encoder (VAE).

One simple, but powerful, extension of the VAE is the importance weighted auto-encoder (IWAE) [4]. This method is derived using importance sampling, but it returns for a tighter lower bound compared to the VAE free energy (for $K > 1$),

$$\mathcal{L}(\theta) = \log p(\mathbf{x}|\theta) \geq \sum_{n=1}^{N} \log \frac{1}{K} \sum_{k=1}^{K} \frac{p(x_n, z_{n,k}|\theta)}{q_\gamma(z_{n,k})} = \mathcal{F}_{\text{IWAE}}(\theta, \gamma), \tag{5}$$

where $\mathbf{z}_{n,k=1:K}$ are $K$ independent samples from $q_\gamma(z_n)$. The bound in eq. (5) saturates as $K \to \infty$.

## 2 Black-box $\alpha$-divergence

Black-box $\alpha$-divergence is a recently proposed framework for approximate inference which includes VI as a particular case [1]. This general scheme uses the factor tying scheme employed in Stochastic Expectation Propagation [5], but applies this directly to the Power-Expectation Propagation energy, yielding a closed form energy function that can be directly optimised without the need for message passing. Furthermore, black-box $\alpha$-divergence can be applied to a variety of probabilistic models and can be scaled to large-scale datasets.

The black-box $\alpha$-divergence objective for the generative model described above is

$$\mathcal{F}_{\text{BBAE}} = \sum_{n=1}^{N} \left( \log \mathcal{Z}_{p(z_n)} - \log \mathcal{Z}_{q(z_n)} - \frac{1}{\alpha} \log \mathrm{E}_q \left[ \left( \frac{p(x_n|z_n, \theta)}{f(z_n)} \right)^\alpha \right] \right) \tag{6}$$

where $p(z_n) = \exp(s(z_n)^\intercal \lambda_0 - \log \mathcal{Z}_{p(z_n)})$, $f(z_n) = \exp(s(z_n)^\intercal \lambda_n)$, $q(z_n) = \exp(s(z_n)^\intercal(\lambda_0 + \lambda_n) - \log \mathcal{Z}_{q(z_n)})$ are the prior, un-normalised factor approximation and posterior distribution corresponding to the $n$-th data-point, respectively. Similar to the VAE, we can employ a recognition model to parameterise the factor approximation. Namely, $f(z_n) \propto \mathcal{N}(z_n; \mu_\gamma(x_n), \sigma^2_\gamma(x_n))$. We will refer to this objective for the recognition model ($\theta$) and the generative model ($\gamma$) as the black-box $\alpha$ auto-encoder (BBAE).

In general, the expectations in the objective above can not be evaluated analytically. However, they can be approximated using Monte-Carlo with $K$ samples drawn from $q(z_n)$,

$$\mathcal{F}_{\text{BBAE}} \approx \sum_{n=1}^{N} \left( \log \mathcal{Z}_{p(z_n)} - \log \mathcal{Z}_{q(z_n)} - \frac{1}{\alpha} \log \frac{1}{K} \sum_{k=1}^{K} \left( \frac{p(x_n|z_{n,k}, \theta)}{f(z_{n,k})} \right)^\alpha \right) \tag{7}$$

### 2.1 Connections to VAE, IWAE, and the variational Rényi free energy

- As $\alpha \to 0$, the objective in eq. (6) becomes the VAE objective in eq. (4). See [1, 6] for further details.
- When $\alpha = 1$, the Monte-Carlo estimate of the BBAE objective in eq. (7) is identical to the IWAE objective in eq. (5). This identity is, perhaps, surprising given the difference between the motivation and approaches taken to arrive at the IWAE and the BBAE.
- The black-box $\alpha$ divergence objective is a special case of the variational Rényi (VR) bound [7], when the size of data minibatches in the stochastic VR objective is 1.

## 3 Preliminary experiments

The proposed training scheme is evaluated using a deep generative model on two datasets: binarised MNIST [8] and Omniglot [9]. The generative and recognition mappings are both neural networks, each with one deterministic hidden layer of 400 units. We train the model using the BBAE objective with different $\alpha$ values, and the VAE objective, during a total of 2000 epochs. The quality of the

inference and generative networks learnt is compared using the log-likelihood of test images, which is computed using importance sampling with 2000 samples drawn from the recognition model [3]. The results averaged over 5 trials are displayed in fig. 1. These results show a gain obtained by simply changing the training objective from VAE/IWAE to BBAE, demonstrated by a better log-likelihood on test data for $\alpha$ values bigger than 1, such as $\alpha = 2$.

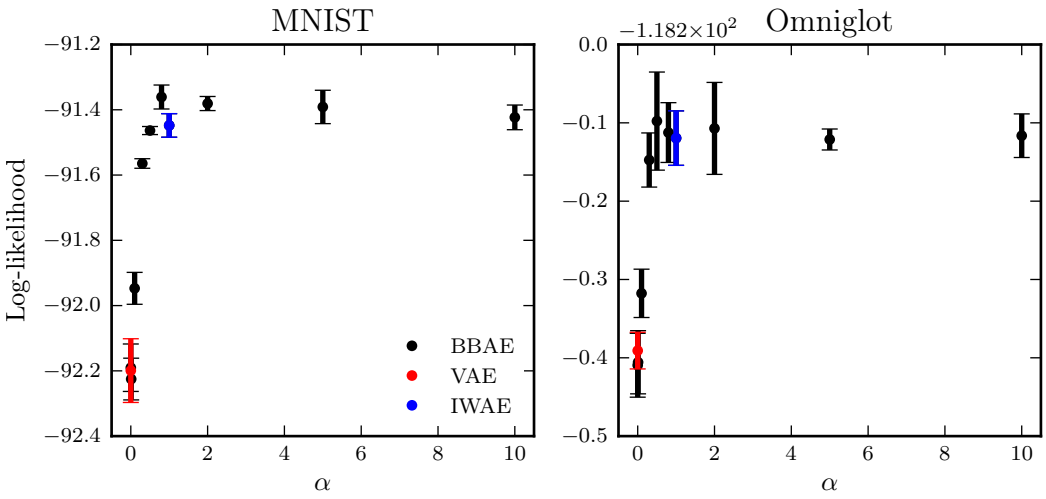


Figure 1: Average test log-likelihood in nats.

## 4 Conclusions and Future Work

We have described how Black-box $\alpha$-divergence can be used in the context of un-supervised learning to find a generative mapping (decoder) and a recognition mapping (encoder) to explain the observed data. Black-box $\alpha$-divergence is a very general method for approximate inference that is able to obtain, as a particular case, previous methods for un-supervised learning, including the VAE ($\alpha \to 0$) and the IWAE ($\alpha = 1$). Furthermore, by changing the value of $\alpha$ we are able to obtain generative and recognition models with better properties. In particular, the results show that setting $\alpha > 1$ provides better results in terms of the test log-likelihood than the VAE or the IWAE.

Regarding future work we will consider a deep analysis of the reasons for which $\alpha > 1$ gives better results on the datasets investigated. The is complex as it involves an interaction between the $\alpha$-divergence selected (i.e. setting of $\alpha$), and the Monte Carlo approximation. Moreover, we also plan to extend the variational framework to handle the uncertainty in the weights of the generative model ($\theta$). By doing so we expect to capture complicated patterns in the generative distribution and to obtain improved results in consequence.

## References


[1] J. M. Hernández-Lobato, Y. Li, M. Rowland, D. Hernández-Lobato, T. Bui, and R. E. Turner, "Black-box $\alpha$-divergence minimization," in *33rd International Conference on Machine Learning*, 2016.

[2] D. P. Kingma and M. Welling, "Auto-encoding variational Bayes," in *International Conference on Learning Representations*, 2014.

[3] D. J. Rezende, S. Mohamed, and D. Wierstra, "Stochastic backpropagation and approximate inference in deep generative models," in *31st International Conference on Machine Learning*, pp. 1278–1286, 2014.

[4] Y. Burda, R. Grosse, and R. Salakhutdinov, "Importance weighted autoencoders," *International Conference on Learning Representations*, 2016.

[5] Y. Li, J. M. Hernández-Lobato, and R. E. Turner, "Stochastic expectation propagation," in *Advances in Neural Information Processing Systems 29*, 2015.

[6] T. Minka, "Divergence measures and message passing," tech. rep., Microsft Research, 2005.

[7] Y. Li and R. E. Turner, "Rényi divergence variational inference," in *Advances in Neural Information Processing Systems 30*, 2016.

[8] H. Larochelle and I. Murray, "The neural autoregressive distribution estimator.," in *15th International Conference on Artificial Intelligence and Statistics*, 2011.

[9] B. M. Lake, R. R. Salakhutdinov, and J. Tenenbaum, "One-shot learning by inverting a compositional causal process," in *Advances in Neural Information Processing Systems 27*, pp. 2526–2534, 2013.