
Auto-Encoding Variational Bayes

Diederik P. Kingma

Machine Learning Group
Universiteit van Amsterdam
dpkingma@gmail.com

Max Welling

Machine Learning Group
Universiteit van Amsterdam
welling.max@gmail.com

Abstract

Can we efficiently learn the parameters of directed probabilistic models, in the presence of continuous latent variables with intractable posterior distributions, and in case of large datasets? We introduce a novel learning and approximate inference method that works efficiently, under some mild conditions, even in the on-line and intractable case. The method involves optimization of a stochastic objective function that can be straightforwardly optimized w.r.t. all parameters, using standard gradient-based optimization methods.

The method does not require the typically expensive sampling loops per data-point required for Monte Carlo EM, and all parameter updates correspond to optimization of the variational lower bound of the marginal likelihood, unlike the wake-sleep algorithm. These theoretical advantages are reflected in experimental results.

1 Introduction

How to efficiently learn the parameters of directed probabilistic models whose continuous latent variables have intractable posterior distributions? The variational approach to approximate Bayesian inference involves the introduction of an approximate posterior to the intractable posterior, used to maximize the variational lower bound of the marginal likelihood. Unfortunately, the common mean-field approach requires analytical solutions to expectations w.r.t. the approximate posterior, which are also intractable in the general case. We show how for continuous latent variables, a reparameterization of the expectation w.r.t. the approximate posterior yields a novel and practical estimator of the variational lower bound that can be differentiated and jointly optimized w.r.t. all parameters, i.e. both the variational parameters and regular parameters, using standard stochastic gradient ascent techniques.

The objective contains, in addition to regularization terms dictated by the variational bound, a noisy data reconstruction term, exposing a novel connection between auto-encoders and stochastic variational inference. In contrast to a typical objective for auto-encoders [BCV13], all parameters updates, including those of the noise distribution, correspond to optimization of the variational lower bound of the marginal likelihood. From the learned generative model it is straightforward to generate samples, without the typical requirement of running Markov chains. The probabilistic encoder can be used for fast approximate inference of latent variables, i.e. for recognition, representation or visualization purposes. Furthermore, the lower bound estimator can be used for unsupervised inference tasks such as denoising and inpainting.

2 Method

The strategy in the following section can be used to derive a lower bound estimator (a stochastic objective function) for a variety of directed graphical models with continuous latent variables. We will restrict ourselves here to the common case where we have an i.i.d. dataset with latent variables

per datapoint, and where we like to perform ML or MAP inference on the (global) parameters, and variational inference on the latent variables. It is, for example, straightforward to extend this scenario to the case where we also perform variational inference on the global parameters; that algorithm is put in the appendix, but experiments with that case are left to future work. Note that our method can be applied to online, non-stationary settings, e.g. streaming data, but here we assume a fixed dataset for simplicity.

2.1 Problem scenario

Let us consider some dataset $\mathbf{X} = \{\mathbf{x}^{(i)}\}_{i=1}^N$ consisting of N i.i.d. samples of some continuous or discrete variable \mathbf{x} . We assume that the data are generated by some random process, involving an unobserved continuous random variable \mathbf{z} . The process consists of two steps: (1) a value $\mathbf{z}^{(i)}$ is generated from some prior distribution $p_{\theta^*}(\mathbf{z})$; (2) a value $\mathbf{x}^{(i)}$ is generated from some conditional distribution $p_{\theta^*}(\mathbf{x}|\mathbf{z})$. We assume that the prior $p_{\theta^*}(\mathbf{z})$ and likelihood $p_{\theta^*}(\mathbf{x}|\mathbf{z})$ come from parametric families of distributions $p_{\theta}(\mathbf{z})$ and $p_{\theta}(\mathbf{x}|\mathbf{z})$, and that their PDFs are differentiable almost everywhere w.r.t. both θ and \mathbf{z} . Unfortunately, a lot of this process is hidden from our view: the true parameters θ^* as well as the values of the latent variables $\mathbf{z}^{(i)}$ are unknown to us.

Very importantly, we *do not* make the usual simplifying assumptions common in the literature. Conversely, we are here interested in a general algorithm that even works in the case of:

1. *Intractability*: the case where the integral of the marginal likelihood $p_{\theta}(\mathbf{x}) = \int p_{\theta}(\mathbf{z})p_{\theta}(\mathbf{x}|\mathbf{z}) d\mathbf{z}$ is intractable (so we cannot evaluate or differentiate the marginal likelihood), where the true posterior density $p_{\theta}(\mathbf{z}|\mathbf{x}) = p_{\theta}(\mathbf{x}|\mathbf{z})p_{\theta}(\mathbf{z})/p_{\theta}(\mathbf{x})$ is intractable (so the EM algorithm cannot be used), and where the required integrals for any reasonable mean-field Variational Bayes are also intractable. These intractabilities are quite common and already appear in case of moderately complicated likelihood functions $p_{\theta}(\mathbf{x}|\mathbf{z})$, e.g. a neural network with a nonlinear hidden layer.
2. *A large dataset*: we have so much data that batch optimization is too costly; we would like to make parameter updates using small minibatches or even single datapoints. Sampling-based solutions, e.g. Monte Carlo EM, would in general be too slow, since it involves a typically expensive sampling loop per datapoint.

We are interested in, and propose a solution to, three related problems in the above scenario:

1. Efficient approximate maximum likelihood (ML) or maximum a posteriori (MAP) estimation for the parameters θ . The parameters can be of interest themselves, e.g. if we are analyzing some natural process. They also allow us to mimic the hidden random process and generate artificial data that resembles the real data.
2. Efficient approximate posterior inference of the latent variable \mathbf{z} given an observed value \mathbf{x} for a choice of parameters θ . This is useful for coding or data representation tasks.
3. Efficient approximate marginal inference of the variable \mathbf{x} . This allows us to perform all kinds of inference tasks where a prior over \mathbf{x} is required. Common applications in computer vision include image denoising, inpainting and super-resolution.

For the purpose of solving the above problems, let us introduce the parametric variational approximation $q_{\phi}(\mathbf{z}|\mathbf{x})$: an approximation to the intractable true posterior $p_{\theta}(\mathbf{z}|\mathbf{x})$. Note that in contrast with the approximate posterior in mean-field variational inference, it is not necessarily factorial and its parameters are not computed from some closed-form expectation. Instead, its parameters ϕ are learned jointly with the parameters of the generative model.

From a coding theory perspective, the unobserved variables \mathbf{z} have an interpretation as a latent representation or *code*. In this paper we will therefore also refer to $q_{\phi}(\mathbf{z}|\mathbf{x})$ as a (*variational*) *encoder* or *recognition model*, since given a datapoint \mathbf{x} it produces a distribution (e.g. a Gaussian) over the possible values of the code \mathbf{z} from which the datapoint \mathbf{x} could have been generated. In a similar vein we will refer to $p_{\theta}(\mathbf{x}|\mathbf{z})$ as a (*generative*) *decoder*, since given a code \mathbf{z} it produces a distribution over the possible corresponding values of \mathbf{x} .

2.2 The variational bound

The marginal likelihood is composed of a sum over the marginal likelihoods of individual datapoints $\log p_{\theta}(\mathbf{x}^{(1)}, \dots, \mathbf{x}^{(N)}) = \sum_{i=1}^N \log p_{\theta}(\mathbf{x}^{(i)})$, which can each be rewritten as:

$$\log p_{\theta}(\mathbf{x}^{(i)}) = D_{KL}(q_{\phi}(\mathbf{z}|\mathbf{x}^{(i)}) || p_{\theta}(\mathbf{z}|\mathbf{x}^{(i)})) + \mathcal{L}(\boldsymbol{\theta}, \boldsymbol{\phi}; \mathbf{x}^{(i)}) \quad (1)$$

The first RHS term is the KL divergence of the approximate from the true posterior, which is non-negative. The second RHS term $\mathcal{L}(\boldsymbol{\theta}, \boldsymbol{\phi}; \mathbf{x}^{(i)})$ denotes the variational lower bound of the marginal likelihood of datapoint i :

$$\log p_{\theta}(\mathbf{x}^{(i)}) \geq \mathcal{L}(\boldsymbol{\theta}, \boldsymbol{\phi}; \mathbf{x}^{(i)}) = \int q_{\phi}(\mathbf{z}|\mathbf{x}) \left(\log p_{\theta}(\mathbf{x}^{(i)}|\mathbf{z}) + \log p_{\theta}(\mathbf{z}) - \log q_{\phi}(\mathbf{z}|\mathbf{x}) \right) d\mathbf{z} \quad (2)$$

Note that the bound equals the true marginal when the divergence of the approximate from true posterior distribution is zero.

The expectation on the RHS of eq. (2) can obviously be written as a sum of three separate expectations, of which the second and third component can sometimes be analytically solved, e.g. when both $p_{\theta}(\mathbf{x})$ and $q_{\phi}(\mathbf{z}|\mathbf{x})$ are Gaussian. For generality we will here assume that each of these expectations are intractable.

We would like to optimize the lower bound $\mathcal{L}(\boldsymbol{\theta}, \boldsymbol{\phi}; \mathbf{x}^{(i)})$ (eq. (2)) using stochastic gradients. Note that following these gradients would either decrease the KL divergence between the approximate and true posterior distributions, or increase the marginal likelihood, or both. A naïve attempt to compute a stochastic gradient would be to draw samples $\{\mathbf{z}^{(l)}\}_{l=1}^L$ from q_{ϕ} and then differentiate the following Monte Carlo estimate of the lower bound:

$$\mathcal{L}(\boldsymbol{\theta}, \boldsymbol{\phi}; \mathbf{x}^{(i)}) \simeq \frac{1}{L} \sum_{l=1}^L \left(\log p_{\theta}(\mathbf{x}^{(i)}|\mathbf{z}^{(l)}) + \log p_{\theta}(\mathbf{z}^{(l)}) - \log q_{\phi}(\mathbf{z}^{(l)}|\mathbf{x}^{(i)}) \right) \quad \text{where } \mathbf{z}^{(l)} \sim q_{\phi}(\mathbf{z}|\mathbf{x})$$

While the above expression is an unbiased estimator of the marginal likelihood (i.e. it will equal the lower bound in the limit $L \rightarrow \infty$), differentiating it w.r.t. the parameters $\boldsymbol{\phi}$ will not result in an unbiased gradient: the variational parameters $\boldsymbol{\phi}$ indirectly influence the estimate through the samples $\mathbf{z}^{(l)} \sim q_{\phi}(\mathbf{z}|\mathbf{x})$, and it is impossible to differentiate through this sampling process. Existing work on stochastic variational bayes provide workarounds [BJP12], but not a solution to this problem.

2.3 Our estimator of the lower bound

Under certain mild conditions outlined in section 2.4 for a chosen approximate posterior $q_{\phi}(\mathbf{z}|\mathbf{x})$ we can reparameterize its conditional samples $\tilde{\mathbf{z}} \sim q_{\phi}(\mathbf{z}|\mathbf{x})$ as

$$\tilde{\mathbf{z}} = g_{\phi}(\boldsymbol{\epsilon}, \mathbf{x}) \quad \text{with } \boldsymbol{\epsilon} \sim p(\boldsymbol{\epsilon}) \quad (3)$$

where we choose a prior $p(\boldsymbol{\epsilon})$ and a function $g_{\phi}(\boldsymbol{\epsilon}, \mathbf{x})$ such that the following holds:

$$\begin{aligned} \mathcal{L}(\boldsymbol{\theta}, \boldsymbol{\phi}; \mathbf{x}^{(i)}) &= \int q_{\phi}(\mathbf{z}|\mathbf{x}) \left(\log p_{\theta}(\mathbf{x}^{(i)}|\mathbf{z}) + \log p_{\theta}(\mathbf{z}) - \log q_{\phi}(\mathbf{z}|\mathbf{x}) \right) d\mathbf{z} \\ &= \int p(\boldsymbol{\epsilon}) \left(\log p_{\theta}(\mathbf{x}^{(i)}|\mathbf{z}) + \log p_{\theta}(\mathbf{z}) - \log q_{\phi}(\mathbf{z}|\mathbf{x}) \right) \Big|_{\mathbf{z}=g_{\phi}(\boldsymbol{\epsilon}, \mathbf{x}^{(i)})} d\boldsymbol{\epsilon} \end{aligned} \quad (4)$$

For notational conciseness we introduce a shorthand notation $f_{\theta, \phi}(\mathbf{x}, \mathbf{z})$ for the sum of three PDFs:

$$f_{\theta, \phi}(\mathbf{x}, \mathbf{z}) = \log p_{\theta}(\mathbf{x}|\mathbf{z}) + \log p_{\theta}(\mathbf{z}) - \log q_{\phi}(\mathbf{z}|\mathbf{x}) \quad (5)$$

Using eq. (4), the Monte Carlo estimate of the variational lower bound, given datapoint $\mathbf{x}^{(i)}$, is:

$$\mathcal{L}(\boldsymbol{\theta}, \boldsymbol{\phi}; \mathbf{x}^{(i)}) \simeq \frac{1}{L} \sum_{l=1}^L f_{\theta, \phi}(\mathbf{x}^{(i)}, g_{\phi}(\boldsymbol{\epsilon}^{(l)}, \mathbf{x}^{(i)})) \quad \text{where } \boldsymbol{\epsilon}^{(l)} \sim p(\boldsymbol{\epsilon}) \quad (6)$$

The estimator only depends on samples from $p(\boldsymbol{\epsilon})$ which are obviously not influenced by $\boldsymbol{\phi}$, therefore we can use it as an objective function that can be differentiated and jointly optimized w.r.t. both $\boldsymbol{\theta}$ and

Algorithm 1 Pseudocode for computing a stochastic gradient using our estimator. See section 2.3 for meaning of the functions $f_{\theta,\phi}$ and g_ϕ . The minibatch $\mathbf{X}^M = \{\mathbf{x}^{(i)}\}_{i=1}^M$ is a randomly drawn subset of the full dataset \mathbf{X} . We use settings $M = 100$ and $L = 1$ in experiments.

Require: θ, ϕ (Current value of parameters)

```

 $\mathbf{g} \leftarrow 0$ 
 $\mathbf{X}^M \leftarrow$  Random subset (minibatch) of  $M$  datapoints from dataset
for each  $\mathbf{x} \in \mathbf{X}^M$  do
    for  $l$  is 1 to  $L$  do
         $\epsilon \leftarrow$  Random draw from prior of  $p(\epsilon)$ 
         $\mathbf{g} \leftarrow \mathbf{g} + \nabla_{\theta,\phi} f_{\theta,\phi}(\mathbf{x}, g_\phi(\epsilon, \mathbf{x}))$ 
    end for
end for
return  $(N/(M \cdot L)) \cdot \mathbf{g}$ 
```

ϕ . Given multiple datapoints from the dataset \mathbf{X} , we can easily construct a minibatch-based version of the estimator: $\mathcal{L}(\theta, \phi; \mathbf{X}) \simeq \frac{N}{M} \sum_{i=1}^M \mathcal{L}(\theta, \phi; \mathbf{x}^{(i)})$ where the minibatch $\mathbf{X}^M = \{\mathbf{x}^{(i)}\}_{i=1}^M$ is a randomly drawn subset of the full dataset \mathbf{X} . In our experiments we found that the number of samples L per datapoint can be set to 1 as long as the minibatch size M was large enough, e.g. $M = 100$. Derivatives $\nabla_{\theta,\phi} \tilde{\mathcal{L}}(\theta; \mathbf{X}^M)$ can be taken, and the resulting gradients can be used in conjunction with stochastic optimization methods such as SGD or Adagrad [DHS10]. See algorithm 1 for a basic approach to compute the stochastic gradients.

A connection with auto-encoders becomes clear when looking at the objective function given at eq. (6). The variational approximation $q_\phi(\mathbf{z}|\mathbf{x}^{(i)})$ (the encoder) maps a datapoint $\mathbf{x}^{(i)}$ to a distribution over latent variables \mathbf{z} from which the datapoint could have been generated. The function $g_\phi(\cdot)$ is chosen such that it maps a datapoint $\mathbf{x}^{(i)}$ and a random noise vector $\epsilon^{(l)}$ to a sample from the approximate posterior for that datapoint: $\mathbf{z}^{(i,l)} = g_\phi(\epsilon^{(l)}, \mathbf{x}^{(i)})$ where $\mathbf{z}^{(i,l)} \sim q_\phi(\mathbf{z}|\mathbf{x}^{(i)})$. Subsequently, the sample $\mathbf{z}^{(i,l)}$ is then input to function $f_{\theta,\phi}(\cdot)$, which consists of three parts. The first part ($\log p_\theta(\mathbf{x}^{(i)}|\mathbf{z}^{(i,l)})$) can be interpreted as the *negative reconstruction error* in neural network parlance. The second and third part can be interpreted as regularization terms that make sure the code activations have high entropy due to the term $\log q_\phi(\mathbf{z}|\mathbf{x})$, while not being too far from the prior due to the term $\log p_\theta(\mathbf{z})$.

2.4 The deterministic parameterization trick

In order to solve the problem we invoked a reparameterization trick that is perhaps best known in literature for a different application , namely an efficient Gibbs sampling technique going under the name of *non-centered parameterization* [PRS07] or *ancillary augmentation* (AA) [YM11].

The essential parameterization trick is quite simple. Let $q_\phi(\mathbf{z}|\mathbf{x})$ be some conditional distribution parameterized by ϕ . It is then often possible to express the random variable \mathbf{z} given \mathbf{x} as a deterministic variable $\mathbf{z} = g_\phi(\epsilon, \mathbf{x})$, where ϵ is an auxiliary variable with independent marginal $p(\epsilon)$, and $g_\phi(\cdot)$ is some vector-valued function parameterized by ϕ .

This reparameterization is useful for our case since it can be used to rewrite an expectation w.r.t. $q_\phi(\mathbf{z}|\mathbf{x})$ such that the Monte Carlo estimate of the expectation is differentiable w.r.t. ϕ . A proof is as follows. Given the deterministic mapping $\mathbf{z} = g_\phi(\epsilon, \mathbf{x})$ we know that $q_\phi(\mathbf{z}|\mathbf{x}) \prod_i dz_i = p(\epsilon) \prod_i d\epsilon_i$. Therefore¹, $\int q_\phi(\mathbf{z}|\mathbf{x}) f(\mathbf{z}) d\mathbf{z} = \int p(\epsilon) f(g_\phi(\epsilon, \mathbf{x})) d\epsilon = \int p(\epsilon) f(g_\phi(\epsilon, \mathbf{x})) d\epsilon$. It follows that a differentiable estimator can be constructed: $\int q_\phi(\mathbf{z}|\mathbf{x}) f(\mathbf{z}) d\mathbf{z} \simeq \frac{1}{L} \sum_{l=1}^L f(g_\phi(\mathbf{x}, \epsilon^{(l)}))$ where $\epsilon^{(l)} \sim p(\epsilon)$. In section 2.3 we applied this trick to obtain a differentiable estimator of the variational lower bound.

Take, for example, the univariate Gaussian case: let z be distributed as $p(z|x) = \mathcal{N}(z, \sigma)$. The random variable z is partially explained by x , but there is some uncertainty left indicated by σ .

¹Note that for infinitesimals we use the notational convention $d\mathbf{z} = \prod_i dz_i$

In this case, a deterministic parameterization is $z = x + \sigma\epsilon$, where ϵ is an independent auxiliary variable $\epsilon \sim \mathcal{N}(0, 1)$. In this univariate Gaussian case, $\phi = \{\sigma\}$ and $g_\phi(\epsilon, y) = y + \sigma\epsilon$.

When can we do this, i.e., for which $q_\phi(\mathbf{z}|\mathbf{x})$ can we choose such a $g_\phi(\cdot)$ and $p(\epsilon)$? There are three basic approaches:

1. Tractable inverse CDF. In this case, let $\epsilon \sim \mathcal{U}(\mathbf{0}, \mathbf{I})$, and let $g_\phi(\epsilon, \mathbf{x})$ be the inverse CDF of $q_\phi(\mathbf{z}|\mathbf{x})$. Examples: Exponential, Cauchy, Logistic, Rayleigh, Pareto, Weibull, Reciprocal, Gompertz, Gumbel and Erlang distributions.
2. Analogous to the Gaussian example, for any "location-scale" family of distributions (with differentiable log-PDF) we can choose the standard distribution (with location = 0, scale = 1) as the auxiliary variable E , and let $g(\cdot) = \text{location} + \text{scale} \cdot \epsilon$. Examples: Laplace, Elliptical, Student's t, Logistic, Uniform, Triangular and Gaussian distributions.
3. Composition: It is often possible to express variables as functions of component variables with distributions that are reparameterizable using either of the above two approaches. Examples: Log-Normal (exponentiation of normally distributed variable), Gamma (a sum over exponentially distributed variables), Dirichlet (weighted sum of Gamma variates), Beta, Chi-Squared, and F distributions.

When all three approaches fail, good approximations to the inverse CDF exist requiring computations with time complexity comparable to the PDF (see e.g. [Dev86] for some methods).

3 Example

Here we'll give an example generative model and posterior approximation used in experiments.

Let the prior over the latent variables be the centered isotropic Gaussian $p_\theta(\mathbf{z}) = \mathcal{N}(\mathbf{z}; \mathbf{0}, \mathbf{I})$. Note that in this case, the prior lacks parameters. Let $p_\theta(\mathbf{x}|\mathbf{z})$ (the decoder) be a multivariate Bernoulli whose probabilities are computed from \mathbf{z} with a fully-connected neural network with a single hidden layer:

$$\log p_\theta(\mathbf{x}|\mathbf{z}) = \sum_{i=1}^D x_i \log y_i - (1 - x_i) \cdot \log(1 - y_i)$$

where $\mathbf{y} = \exp(\mathbf{W}_2 \tanh(\mathbf{W}_1 \mathbf{z} + \mathbf{b}_1) + \mathbf{b}_2)$ (7)

While there is much freedom in the choice of $q_\phi(\mathbf{z}|\mathbf{x})$ (the decoder), we'll for a moment assume a relatively simple case: let's assume that the true posterior $p_\theta(\mathbf{z}|\mathbf{x})$ takes on a approximate Gaussian form with an approximately diagonal covariance. In this case, we can let the variational approximate posterior be a multivariate Gaussian with a diagonal covariance structure²:

$$\log q_\phi(\mathbf{z}|\mathbf{x}) = \log \mathcal{N}(\mathbf{z}; \boldsymbol{\mu}, \boldsymbol{\sigma}^2 \mathbf{I}) (8)$$

where $\boldsymbol{\mu}$ and $\boldsymbol{\sigma}$ are yet unspecified functions of \mathbf{x} . We can sample from $q_\phi(\mathbf{z}|\mathbf{x})$ using $\tilde{\mathbf{z}} = h_\phi(\mathbf{x}, \epsilon) = \boldsymbol{\mu} + \boldsymbol{\sigma} \odot \epsilon$ where $\epsilon \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$. With \odot we signify an element-wise product. Therefore, given a minibatch \mathbf{X}^M of data, and using the $f_{\theta, \phi}(\cdot)$ abbreviation of eq. (5), our estimator of the lower bound is:

$$\mathcal{L}(\boldsymbol{\theta}, \boldsymbol{\phi}; \mathbf{x}^{(i)}) \simeq \frac{1}{L} \sum_{l=1}^L f_{\theta, \phi}(\mathbf{x}^{(i)}, \mathbf{z}^{(i, l)}) \Big|_{\mathbf{z}^{(i, l)} = \boldsymbol{\mu}^{(i)} + \boldsymbol{\sigma}^{(i)} \odot \epsilon^{(l)}} \quad \text{where} \quad \epsilon^{(l)} \sim \mathcal{N}(\mathbf{0}, \mathbf{I}) (9)$$

where $\boldsymbol{\mu}^{(i)}$ and $\boldsymbol{\sigma}^{(i)}$ denote the mean and s.d. of the approximation of the posterior $q_\phi(\mathbf{z}|\mathbf{x}^{(i)})$, which we didn't yet specify. Let the mean $\boldsymbol{\mu}^{(i)}$ and variance $\boldsymbol{\sigma}^{(i)}$ of the Gaussian encoding distribution be the following nonlinear function of \mathbf{x} , (a neural network):

$$\begin{aligned} \log q_\phi(\mathbf{z}|\mathbf{x}) &= \log \mathcal{N}(\mathbf{z}; \boldsymbol{\mu}, \boldsymbol{\sigma}^2 \mathbf{I}) \\ \text{where } \boldsymbol{\mu} &= \mathbf{W}_4 \mathbf{h} + \mathbf{b}_4, \quad \text{and} \quad \log \boldsymbol{\sigma}^2 = \mathbf{W}_5 \mathbf{h} + \mathbf{b}_5, \quad \text{and} \quad \mathbf{h} = \tanh(\mathbf{W}_3 \mathbf{x} + \mathbf{b}_3) \end{aligned} (10)$$

²Note that this is just a (simplifying) choice, and not a limitation our method.

Note that the generative (decoding) parameters are $\theta = \{\mathbf{W}_j, \mathbf{b}_j\}_{j=1}^2$ and the variational (encoding) parameters are $\phi = \{\mathbf{W}_j, \mathbf{b}_j\}_{j=3}^5$. These definitions for the encoder and decoder can be plugged in eq. 6, and the lower bound can subsequently be differentiated and optimized w.r.t. the parameters.

In this model both $p_\theta(\mathbf{z})$ and $q_\phi(\mathbf{z}|\mathbf{x})$ are Gaussian; in this special case, the second and third term of $f_{\theta,\phi}$ (eq. (5)) can be solved analytically. This results in an estimator with a lower variance than the generic estimator given in eq. (9). The resulting estimator is:

$$\mathcal{L}(\theta, \phi; \mathbf{x}^{(i)}) \simeq \frac{1}{2} \sum_{j=1}^J \left(1 + \log((\sigma_j^{(i)})^2) - (\mu_j^{(i)})^2 - (\sigma_j^{(i)})^2 \right) + \frac{1}{L} \sum_{l=1}^L \log p_\theta(\mathbf{x}^{(i)} | \mathbf{z}^{(i,l)}) \quad (11)$$

See the appendix for the derivation.

4 Related work

Perhaps the most relevant related method is the Wake-Sleep algorithm [HDFN95]. Like AEVB, the wake-sleep algorithm employs an encoder (called a recognition network) that approximates the true posterior. A well-known drawback of the wake-sleep algorithm is that it lacks a theoretically justified method for learning the parameters of the recognition network: its updates correspond to optimization of the divergence $KL(p||q)$ instead of the divergence $KL(q||p)$ dictated by the lower bound. A theoretical advantage of the wake-sleep algorithm is that it also applies to models with discrete latent variables. Wake-Sleep has the same computational complexity as AEVB per datapoint.

AEVB corresponds to the optimization of a type of auto-encoder, exposing a connection between generative models and auto-encoders. A connection between *linear* auto-encoders and a certain class of generative linear-Gaussian models has long been known. In [Row98] it was shown that PCA corresponds to the maximum-likelihood (ML) solution of a special case of the linear-Gaussian model with a prior $p(\mathbf{z}) = \mathcal{N}(0, \mathbf{I})$ and a conditional distribution $p(\mathbf{x}|\mathbf{z}) = \mathcal{N}(\mathbf{x}; \mathbf{W}\mathbf{z}, \epsilon\mathbf{I})$, specifically the case with infinitesimally small ϵ .

In relevant recent work on autoencoders [VLL⁺10] it was shown that the training criterion of unregularized autoencoders corresponds to maximization of a lower bound (see the infomax principle [Lin89]) of the mutual information between input X and latent representation Z . Maximizing (w.r.t. parameters) of the mutual information is equivalent to maximizing the conditional entropy, which is lower bounded by the expected loglikelihood of the data under the autoencoding model [VLL⁺10], i.e. the negative reconstruction error. However, it is well known that this reconstruction criterion is in itself not sufficient for learning useful representations [BCV13]. Regularization techniques have been proposed to make autoencoders learn useful representations, such as denoising, contractive and sparse autoencoder variants [BCV13]. Related are also encoder-decoder architectures such as the predictive sparse decomposition (PSD) [KRL08] from which we drew some inspiration. In contrast to our method, these methods fall under the umbrella of either unnormalized (or energy-based) models or sparse coding. Our objective function contains (hyper-parameter free) regularization terms dictated by the variational bound (eq. (5)).

5 Experiments

We trained generative models of images from the MNIST and Frey Face datasets³ and compared learning algorithms in terms of the variational lower bound, and the estimated marginal likelihood.

The generative model (encoder) and variational approximation (decoder) from section 3 were used, where the described encoder and decoder have an equal number of hidden units. We write *hidden units*

All parameters were updated according to the MAP criterion $\nabla_{\theta,\phi} \log p(\theta, \phi | \mathbf{X}) = \nabla_{\theta,\phi} \log p_\theta(\mathbf{X}) + \nabla_{\theta,\phi} \log p(\theta, \phi)$, with a prior $p(\theta, \phi) = \mathcal{N}(0, \mathbf{I})$. Optimization of this MAP objective is equivalent to plain likelihood maximization with the addition of a weight decay term.

³Available at <http://www.cs.nyu.edu/~roweis/data.html>

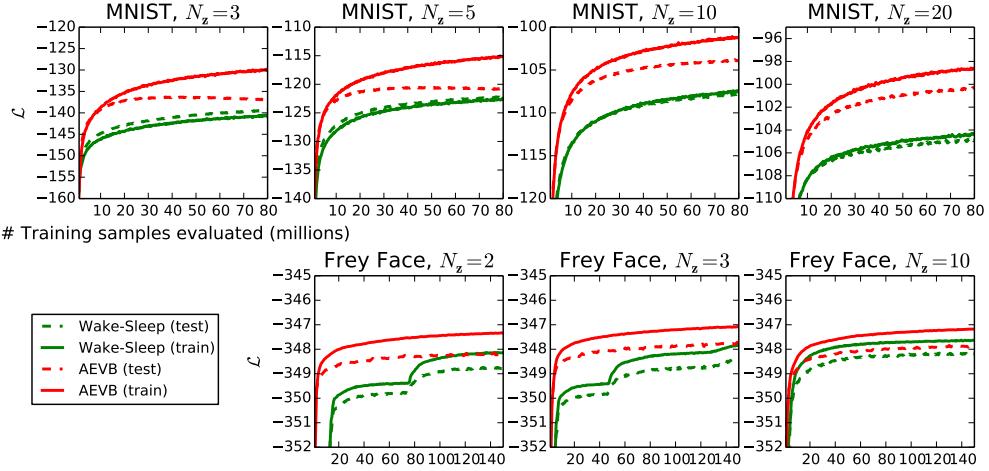


Figure 1: Comparison of our AEVB method to the wake-sleep algorithm, in terms of optimizing the lower bound, for different dimensionality of latent space (N_z). Our method converged considerably faster and reached a better solution in all experiments. Vertical axis: the estimated average variational lower bound per datapoint. The estimator variance was small (< 1) and omitted. Horizontal axis: amount of training points evaluated. Computation took around 20 minutes per million training samples with a dated quad-core Xeon CPU.

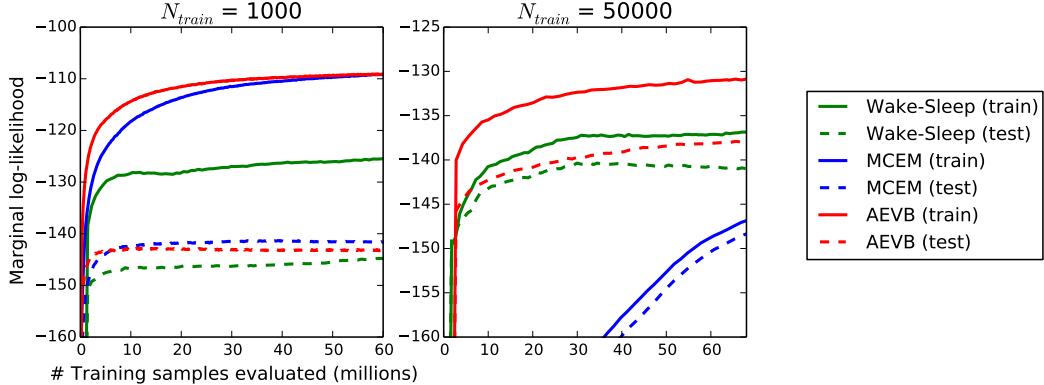
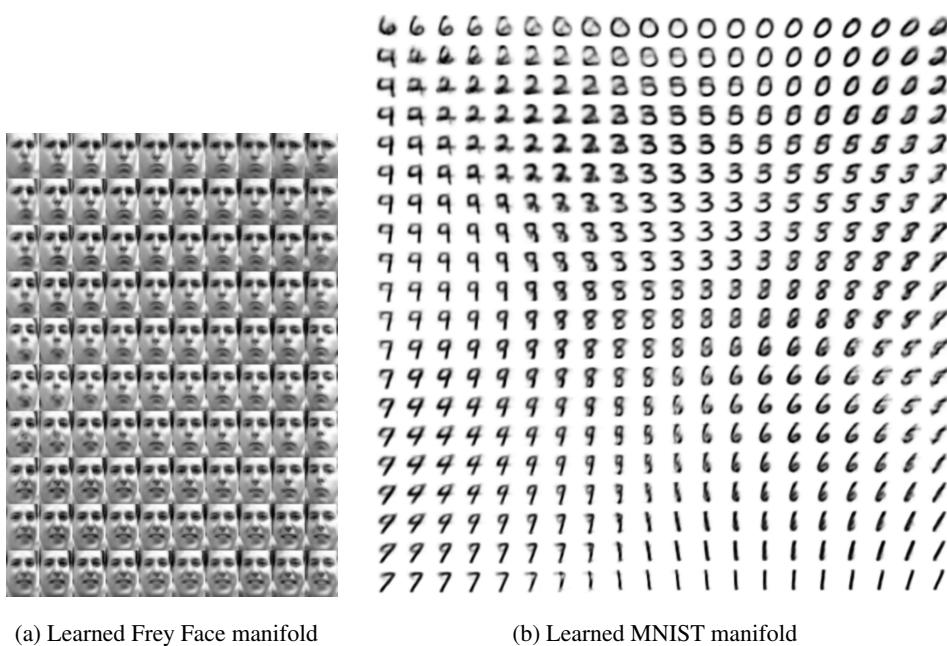


Figure 2: Comparison of AEVB to the wake-sleep algorithm and Monte Carlo EM, in terms of the estimated marginal likelihood, for a different number of training points. The Monte Carlo EM algorithm is (unlike AEVB and the wake-sleep method) asymptotically unbiased but cannot be applied online such that it becomes inefficient for large datasets (right figure).

The likelihood gradient was approximated by the gradient of the lower bound: $\nabla_{\theta, \phi} \log p_{\theta}(\mathbf{X}) \approx \nabla_{\theta, \phi} \mathcal{L}(\theta, \phi; \mathbf{X})$. We compared performance of AEVB to the wake-sleep algorithm [HDFN95]. We employed the same encoder (also called recognition network) for the wake-sleep algorithm and the variational auto-encoder. All parameters, both variational and generative, were initialized by random sampling from $\mathcal{N}(0, 0.01)$, and were jointly stochastically optimized using the MAP criterion. Stepsizes were adapted with Adagrad [DHS10]; the Adagrad global stepsize parameters were chosen from $\{0.01, 0.02, 0.1\}$ based on performance on the training set in the first few iterations. Minibatches of size $M = 100$ were used, with $L = 1$ samples per datapoint.

Likelihood lower bound We trained generative models (decoders) and corresponding encoders (a.k.a. recognition networks) having 500 hidden units in case of MNIST, and 200 hidden units in case of the Frey Face dataset (to prevent overfitting, since it is a considerably smaller dataset). Figure 1 shows the results when comparing the lower bounds.



(a) Learned Frey Face manifold

(b) Learned MNIST manifold

Figure 3: Visualisations of learned data manifold for generative models with two-dimensional latent space, learned with AEVB. Since the prior of the latent space is Gaussian, linearly spaced coordinates on the unit square were transformed through the inverse CDF of the Gaussian to produce values of the latent variables \mathbf{z} . For each of these values \mathbf{z} , we plotted the corresponding generative $p_{\theta}(\mathbf{x}|\mathbf{z})$ with the learned parameters θ .

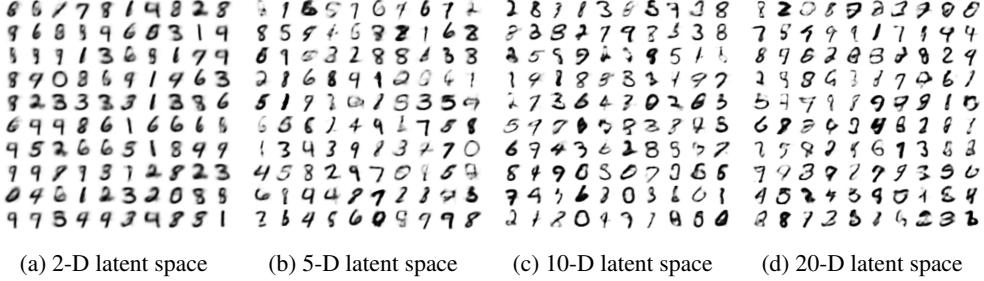


Figure 4: Random samples from learned generative models of MNIST for different dimensionalities of latent space.

Marginal likelihood For very low-dimensional latent space it is possible to estimate the marginal likelihood of the learned generative models using an MCMC estimator. More information about the marginal likelihood estimator is available in the appendix. For the encoder and decoder we again used neural networks, this time with 100 hidden units, and 3 latent variables; for higher dimensional latent space the estimates became unreliable. The AEVB and Wake-Sleep methods were compared to Monte Carlo EM (MCEM) with a Hybrid Monte Carlo (HMC) [DKPR87] sampler; details are in the appendix. We compared the convergence speed for the three algorithms, for a small and large training set size. Results are in figure 2.

6 Conclusion

We have introduced a novel online learning and approximate inference method for models with continuous latent variables, that works for the case where mean-field VB and EM methods are intractable. The proposed estimator can be straightforwardly differentiated and optimized w.r.t. all parameters, resulting in stochastic gradients that are easily plugged into existing stochastic gradient

optimization methods. The method learns an encoder, or variational approximation to the posterior, that can be used for fast approximate inference of the distribution of the latent variables. The theoretical advantages are reflected in experimental results.

References

- [BCV13] Yoshua Bengio, Aaron Courville, and Pascal Vincent. Representation learning: A review and new perspectives. 2013.
- [BJP12] David M Blei, Michael I Jordan, and John W Paisley. Variational bayesian inference with stochastic search. In *Proceedings of the 29th International Conference on Machine Learning (ICML-12)*, pages 1367–1374, 2012.
- [BK88] Hervé Bourlard and Yves Kamp. Auto-association by multilayer perceptrons and singular value decomposition. *Biological cybernetics*, 59(4-5):291–294, 1988.
- [Dev86] Luc Devroye. Sample-based non-uniform random variate generation. In *Proceedings of the 18th conference on Winter simulation*, pages 260–265. ACM, 1986.
- [DHS10] John Duchi, Elad Hazan, and Yoram Singer. Adaptive subgradient methods for online learning and stochastic optimization. *Journal of Machine Learning Research*, 12:2121–2159, 2010.
- [DKPR87] Simon Duane, Anthony D Kennedy, Brian J Pendleton, and Duncan Roweth. Hybrid monte carlo. *Physics letters B*, 195(2):216–222, 1987.
- [Gra11] Alex Graves. Practical variational inference for neural networks. In *Advances in Neural Information Processing Systems*, pages 2348–2356, 2011.
- [HDFN95] Geoffrey E Hinton, Peter Dayan, Brendan J Frey, and Radford M Neal. The “wake-sleep” algorithm for unsupervised neural networks. *SCIENCE-NEW YORK THEN WASHINGTON-*, pages 1158–1158, 1995.
- [KRL08] Koray Kavukcuoglu, Marc’Aurelio Ranzato, and Yann LeCun. Fast inference in sparse coding algorithms with applications to object recognition. Technical Report CBLL-TR-2008-12-01, Computational and Biological Learning Lab, Courant Institute, NYU, 2008.
- [Lin89] Ralph Linsker. *An application of the principle of maximum information preservation to linear systems*. Morgan Kaufmann Publishers Inc., 1989.
- [PRS07] Omiros Papaspiliopoulos, Gareth O Roberts, and Martin Sköld. A general framework for the parametrization of hierarchical models. *Statistical Science*, pages 59–73, 2007.
- [Row98] Sam Roweis. EM algorithms for PCA and SPCA. *Advances in neural information processing systems*, pages 626–632, 1998.
- [VLL⁺10] Pascal Vincent, Hugo Larochelle, Isabelle Lajoie, Yoshua Bengio, and Pierre-Antoine Manzagol. Stacked denoising autoencoders: Learning useful representations in a deep network with a local denoising criterion. *The Journal of Machine Learning Research*, 9999:3371–3408, 2010.
- [YM11] Yaming Yu and Xiao-Li Meng. To Center or Not to Center: That Is Not the Question—An Ancillarity-Sufficiency Interweaving Strategy (ASIS) for Boosting MCMC Efficiency. *Journal of Computational and Graphical Statistics*, 20(3):531–570, 2011.