University of Cambridge

MLSALT 4

Tom
Will Tebbutt, wct23, Darwin College
Paweł Budzianowski, pfb30, Clare Hall

1. Introduction

This paper concerns itself with the scenario in which we wish to find a point-estimate to the parameters θ of some parametric model in which we generate each observations $\mathbf{x}_i$ by first sampling a "local" latent variable $\mathbf{z}_i \sim p_\theta(\mathbf{z})$ and then sampling the associated observation $\mathbf{x}_i \sim p_\theta(\mathbf{x} \,|\, \mathbf{z})$. The conditional independence assumptions in this model are shown in the graphical model in figure fig. 1.

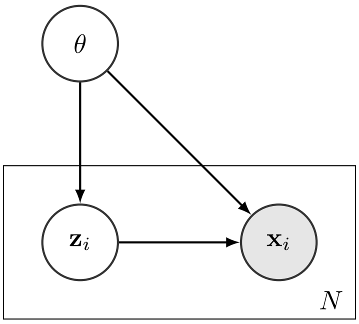

Figure 1: Directed graphical model representing the conditional independencies in the proposed problem scenario. Each i^{th} observation is conditionally independent given the model parameters θ.

The posterior $p_\theta(\mathbf{z}_i \,|\, \mathbf{x}_i)$ is intractable for a continuous latent space whenever either the prior $p_\theta(\mathbf{z}_i)$ or the likelihood $p_\theta(\mathbf{x}_i \,|\, \mathbf{z}_i)$ are non-Gaussian, meaning that approximate inference is required. To this end Autoencoding Variational Bayes makes two contributions in terms of methodology, introducing a differentiable stochastic estimator for the variational lower bound to the model evidence, using this to learn a recognition model to provide a fast method to compute an approximate posterior distribution over "local" latent variables given observations.

In this report we will first discuss the methodological contributions of the Autoencoding Variational Bayes paper, with care taken to treat each of the afformentioned methodological contributions separately. We then discuss their use of this framework to derive the Variational Autoencoder, a model exactly of the form described above with a Gaussian prior, multi-layer perceptron (MLP) parameterised likelihood and MLP-parameterised recognition model. We reproduce the key experiments from the original paper and conduct several more including investigating the sensitivity of the reported results to the network architecture and whether the Variational Autoencoder provides improved reconstructive performance over a traditional "vanilla" autoencoder with the same architecture.

2. Stochastic Variational Inference

The aim of Variational Inference is to provide a deterministic approximation to an intractable posterior distribution by finding parameters ϕ such that $\mathrm{D}_{KL}(q_\phi(\theta) \;||\; p(\theta \,|\, D))$ is minimised. This is achieved by noting that

$$\begin{aligned} \mathrm{D}_{KL}(q_\phi(\theta) \;||\; p(\theta \,|\, D)) &= \log p(D) + \mathbb{E}_{q_\phi(\theta)}[\log q_\phi(\theta) - \log p(\theta, D)] \\ &=: \log p(X) - \mathcal{L}(\phi; D). \end{aligned} \tag{1}$$

Noting that $\log p(D)$ is constant w.r.t. ϕ, we can now minimise the KL-divergence by maximising the evidence lower bound (ELBO) $\mathcal{L}$ (that this is indeed a lower bound follows from the non-negativity of the KL-divergence).

Aside from some notable exceptions (eg. [Titsias, 2009]) this quantity is not tractably point-wise evaluable. However, if $q_\phi(\theta)$ and $\log p(\theta, D)$ are point-wise evaluable, it can be approximated using Monte Carlo as

$$\mathcal{L}(\phi; D) \approx \frac{1}{L}\sum_{l=1}^{L} \log p(\theta_l, D) - \log q_\phi(\theta_l), \quad \theta_l \sim q_\phi(\theta) \tag{2}$$

This stochastic approximation to the ELBO is not differentiable w.r.t. ϕ as the distribution from which each θ_l is sampled itself depends upon ϕ, meaning that the gradient of the log likelihood cannot be exploited to perform inference. One of the primary contributions of the paper being reviewed is to provide a differentiable estimator for $\mathcal{L}$ that allows gradient information to be exploited. In particular it notes that if there exists a tractable reparameterisation of the random variable $\tilde{\theta} \sim q_\phi(\theta)$ such that

$$\tilde{\theta} = g_\phi(\epsilon), \quad \epsilon \sim p(\epsilon), \tag{3}$$

then we can approximate the gradient of the ELBO as

$$\mathcal{L}(\phi; D) = \mathbb{E}_{p(\epsilon)}\left[\log p(\theta, X) - q_\phi(\theta)\right] \approx \frac{1}{L}\sum_{l=1}^{L} \log p(\theta_l, X) - \log q_\phi(\theta_l) =: \tilde{\mathcal{L}}^1(\phi; X), \tag{4}$$

where $\theta_l = g_\phi(\epsilon_l)$ and $\epsilon_l \sim p(\epsilon)$. Thus the dependence of the sampled parameters θ on ϕ has been removed, yielding a differentiable estimator provided that both q_ϕ and $\log p(\theta, D)$ are themselves differentiable. Approxiate inference can now be performed by computing the gradient of $\tilde{\mathcal{L}}^1$ w.r.t. ϕ either by hand or using one's favourite reverse-mode automatic differentiation package (eg. Autograd [Maclaurin et al.,]) and performing gradient-based stochastic optimisation to maximise the elbo using, for example, AdaGrad [Duchi et al., 2011].

The authors also point out that one can re-express the elbo in the following manner

$$\mathcal{L}(\phi; D) = \mathbb{E}_{q_\phi(\theta)}\left[\log p(D \mid \theta)\right] - \mathrm{D}_{KL}\left(q_\phi(\theta) \;||\; p(\theta)\right). \tag{5}$$

This is useful as the KL-divergence between the variational approximation $q_\phi(\theta)$ and the prior over the parameters θ has a tractable closed-form expression in a number of useful cases. This leads to a second estimator for the elbo:

$$\tilde{\mathcal{L}}^2(\phi; D) := \frac{1}{L}\sum_{l=1}^{L} \log p(D \mid \theta_l) - \mathrm{D}_{KL}\left(q_\phi(\theta) \;||\; p(\theta)\right). \tag{6}$$

It seems probable that this estimator will in general have lower variance than $\tilde{\mathcal{L}}^1$.

So far Stochastic Variational Inference has been discussed only in a general parametric setting. The paper's other primary contribution is to use a differentiable recognition network to learn to parameterise the posterior distribution over latent variables z_i local to each observation x_i in a parametric model. In particular, they assume that given some global parameters θ, $\mathbf{z}_i \sim p_\theta(\mathbf{z})$ and $\mathbf{x}_i \sim p_\theta(\mathbf{x}_i \mid \mathbf{z}_i)$. In the general case the posterior distribution over each z_i will be intractable. Furthermore, the number of latent variables $\mathbf{z}_i$ increases as the number of observations increases, meaning that under the framework discussed above we would have to optimise the variational objective with respect to each of them independently. This is potentially computationally intensive and quite wasteful as it completely disregards the any information about the posterior distribution over the z_i provided by the similarities between inputs locations $\mathbf{x}_{\neq i}$and corresponding posteriors $\mathbf{z}_{\neq i}$. To rectify this the recognition model $q_\phi(\mathbf{z} \mid \mathbf{x})$ is introduced.

Given the recognition model and a point estimate for θ, the ELBO becomes

$$\begin{aligned}\mathcal{L}(\theta, \phi; D) &= \mathbb{E}_{q_\phi(\mathbf{z_1}),\ldots,q_\phi(\mathbf{z}_N)}\left[\log\prod_{i=1}^{N} p_\theta(\mathbf{x}_i, \mathbf{z}_i) - \log\prod_{i=1}^{N} q_\phi(\mathbf{z})_i\right] \\ &= \sum_{i=1}^{N} \mathbb{E}_{q_\phi(\mathbf{z}_i)}\left[\log p_\theta(\mathbf{x}_i, \mathbf{z}_i) - \log q_\phi(\mathbf{z}_i)\right]\end{aligned} \tag{7}$$

For this ELBO we can derive a similar result to $\tilde{\mathcal{L}}^1$, where we do no assume a closed-form solution for the KL divergence between distributions and include mini-batching to obtain an estimator for the ELBO for a mini-batch of observations

$$\tilde{\mathcal{L}}^A(\theta, \phi; D) \approx \frac{N}{LM}\sum_{i=1}^{M}\sum_{l=1}^{L} \log p_\theta(\mathbf{x}_i, \mathbf{z}_{i,l}) - \log q_\phi(\mathbf{z}_{i,l}), \quad \mathbf{z}_{i,l} = g_\phi(\mathbf{x}_i, \epsilon_{i,l}), \ \epsilon_{i,l} \sim p(\epsilon) \tag{8}$$

where the M observations in the mini-batch are drawn uniformly from the data set comprised of N obnservations and for each observation we draw L samples from the approximate posterior $q_\phi\left(\mathbf{z}_i \,|\, \mathbf{x}_i\right)$. Similarly, if $q_\phi\left(\mathbf{z} \,|\, \mathbf{x}\right)$ and $p_\theta\left(\mathbf{z}\right)$ are such that the KL-divergence between them has a tractable closed-form solution then we can use an approximate bound which we could reasonably expect to have a lower variance:

$$\begin{aligned}
\mathcal{L}\left(\theta, \phi; D\right) = & \, \mathbb{E}_{q_\phi(\mathbf{z}_1),\ldots,q_\phi(\mathbf{z}_N)}\left[\log \prod_{i=1}^{N} p_\theta\left(\mathbf{x}_i \,|\, \mathbf{z}_i\right)\right] - \mathrm{D}_{KL}\left(\prod_{i=1}^{N} q_\phi\left(\mathbf{z}_i \,|\, \mathbf{x}_i\right) \;||\; \prod_{i=1}^{N} p_\theta\left(\mathbf{z}_i\right)\right) \\
= & \sum_{i=1}^{N} \mathbb{E}_{q_\phi(\mathbf{z}_i \,|\, \mathbf{x}_i)}\left[\log p_\theta\left(\mathbf{x}_i \,|\, \mathbf{z}_i\right)\right] - \mathrm{D}_{KL}\left(q_\phi\left(\mathbf{z}_i \,|\, \mathbf{x}_i\right) \;||\; p_\theta\left(\mathbf{z}_i\right)\right) \\
\approx & \frac{N}{M}\sum_{i=1}^{M}\left[\frac{1}{L}\sum_{l=1}^{L} \log p_\theta\left(\mathbf{x}_i \,|\, \mathbf{z}_{i,l}\right) - \mathrm{D}_{KL}\left(q_\phi\left(\mathbf{z}_i \,|\, \mathbf{x}_i\right) \;||\; p_\theta\left(\mathbf{z}_i\right)\right)\right] =: \tilde{\mathcal{L}}^B\left(\theta, \phi; D\right), \qquad (9)
\end{aligned}$$

where $z_{i,l} = g_\phi\left(\mathbf{x}_i, \epsilon_{i,l}\right)$ and $\epsilon_{i,l} \sim p\left(\epsilon\right)$.

3. The Variational Autoencoder

The Variational Autoencoder exploits the methods described in the previous section to define a probabilistic model whose elbo is highly reminiscent of the objective optimised in a traditional autoencoder. In particular we define a generative model where we assume that the i^{th} observation was generated by first sampling a latent variable $\mathbf{z}_i \sim \mathcal{N}\left(0, I\right)$ and that the each observation vector real-valued observation $\mathbf{x}_i \sim \mathcal{N}\left(\mu_\theta\left(\mathbf{z}_i\right), \sigma_\theta^2\left(\mathbf{z}_i\right)\right)$ where

$$\mu_\theta\left(\mathbf{z}_i\right) = \mathbf{h}_i W_\mu^{(dec)} + \mathbf{b}_\mu^{(dec)}, \qquad (10)$$

$$\log \sigma_\theta^2\left(\mathbf{z}_i\right) = \mathbf{h}_i W_\sigma^{(dec)} + \mathbf{b}_\sigma^{(dec)}, \qquad (11)$$

$\mathbf{h}_i \in \mathbb{R}^{1 \times D_h}$ is the output at the final hidden layer of the "decoder MLP", $W_\mu^{(dec)}, W_\sigma^{(dec)} \in \mathbb{R}^{D_h \times D_z}$ are matrices mapping from the D_h hidden units to the D_z dimensional latent space. Similarly, $\mathbf{b}_\mu^{(dec)}, \mathbf{b}_\sigma^{(dec)} \in \mathbb{R}^{1 \times D_z}$ are row vector biases. Note that the variances are parameterised implicitely through their logs to ensure that they are correctly valued only on the positive reals.

If the output vectors are binary valued then $x_i \sim \text{Bernoulli}\left(f_\theta\left(\mathbf{z}_i\right)\right)$ where again given $\mathbf{h}_i$, the output at the final hidden layer of the decoder MLP for latent-space value $\mathbf{z}_i$

$$f_\theta\left(\mathbf{z}_i\right) = \left[1 + \exp\left(-\mathbf{h}_i W^{(dec)} - \mathbf{b}^{(dec)}\right)\right]^{-1}, \qquad (12)$$

where $W^{(dec)} \in \mathbb{R}^{D_h \times D_z}$ and $b^{(dec)} \in \mathbb{R}^{1 \times D_z}$.

The recognition model is given by

$$q_\phi\left(\mathbf{z} \,|\, \mathbf{x}\right) = \mathrm{N}\left(\mathbf{z} \,\middle|\, \mu_\phi\left(\mathbf{x}_i\right), \sigma_\phi^2\left(\mathbf{x}_i\right)\right), \qquad (13)$$

where the distributional parameters $\mu_\phi\left(\mathbf{x}_i\right)$ and $\sigma_\phi^2\left(\mathbf{x}_i\right)$ are again given by an MLP, which will be refered to as the "encoding MLP", whose input is $\mathbf{x}_i$.

In this case, there is a tractable closed form solution for the KL-divergence between the variational posterior $q_\phi\left(\mathbf{z} \,|\, \mathbf{x}\right)$ and the prior $p_\theta\left(\mathbf{z}\right)$ meaning that we can approximate the ELBO using equation 9. Additionally we have that

$$g_\phi\left(\mathbf{x}, \epsilon\right) = \mu_\phi\left(\mathbf{x}_i\right) + \sigma_\phi\left(\mathbf{x}_i\right) \odot \epsilon, \quad \epsilon \sim \mathcal{N}\left(0, I\right). \qquad (14)$$

The algorithm to perform inference in the Variational Autoencoder is provided in algorithm 1.

4. Reconstruction

On frey faces: used epochs 100 for latent space 2, 550 for 10 and 20

Tried mean and sampling for VAE, mean worked better

Algorithm 1 Procedure by which inference is performed in the Variational Autoencoder.

```
$\theta, \phi \leftarrow$ Initialisation.
while not converged in $\theta, \phi$ do
    Pick subset of size $\mathbf{x}_{1:M}$ from the full dataset uniformly at random.
    Compute $\mu_{1:M}, \log \sigma^2_{1:M}$ using the encoding MLP.
    For all $i \in \{1, ..., M\}$, sample $\epsilon_i \sim \mathcal{N}(0, I)$.
    For all $i \in \{1, ..., M\}$, $\mathbf{z}_i \leftarrow \mu_i + \sigma_i \odot \epsilon_i$.
    For all $i \in \{1, ..., M\}$, compute $\log p(\mathbf{x}_i \mid \mathbf{z}_i)$ using the decoder MLP and likelihood function.
    Compute $g \leftarrow \nabla_{\theta,\phi} \tilde{\mathcal{L}}^B(\theta, \phi; \mathbf{x}_{1:M}, \epsilon_{1:M})$
    Update $\theta, \phi$ using noisy gradient estimate $g$.
end while
```

5. Full variantional bayes

Tried FVB on random initialisation, however performed really badly: continuous: Lower bound: -53582.4758831, time: 0.483903589 Lower bound on validation set: -245029.915134 discrete: Lower bound: -57463.0461588, time: 4.11004209518 Lower bound on validation set: -5434312.97123

6. Extensions

An obvious extension to the paper to investigate is to simply change the form of the prior and the variational approximation in an attempt to induce a particular form of latent space. For example a particularly interesting set up would be to define a sparsity inducing prior that encourages each dimension of the latent space to be approximately valued on $\{0, 1\}$. An obvious choice would be a set of sparse Beta distributions (ie. ones in which the shape parameters $\alpha, \beta < 1$), but one could also use pairs of univariate Gaussians with means 0 and 1 and small variances.

Such a prior would be useful for two reasons - firstly it would allow one to provide a binary encoding for a data set by truncating the posterior approximation for any particular observation to be exactly vector binary valued allowing for a large amount of lossy compression. The posterior distribution over the parameters θ and latent values $\mathbf{z}_i$ also contains rotational symmetry which may affect the quality of the approximate inference if it attempts to place posterior mass over the entirety of this. Were a prior such as the one proposed used, this rotational symmetry would be destroyed and replaced with a "permutation symmetry", similar to that found in a finite mixture model.

7. Conclusion

References

[Duchi et al., 2011] Duchi, J., Hazan, E., and Singer, Y. (2011). Adaptive subgradient methods for online learning and stochastic optimization. *The Journal of Machine Learning Research*, 12:2121–2159.

[Maclaurin et al.,] Maclaurin, D., Duvenaud, D., and Adams, R. P. Autograd: Effortless gradients in numpy.

[Titsias, 2009] Titsias, M. K. (2009). Variational learning of inducing variables in sparse gaussian processes. In *International Conference on Artificial Intelligence and Statistics*, pages 567–574.