# Advanced Variational Inference (Part II)

Caterina De Bacco and Isabel Valera

## 1 Amortized Variational Inference

Let us assume that the joint distribution over the observed variables (the data) $X = \{x_n\}_{n=1}^N$ and the set of local latent variables $Z = \{z_n\}_{n=1}^N$ can be written as

$$p(X,Z) = \prod_{n=1}^{N} p(x_n, z_n). \tag{1}$$

For simplicity of notation and without loss of generality, in this lecture, we do not consider global latent variables $\beta$, as in the previous lecture.

Let us now assume that we rely on VI to approximate the posterior distribution $p(Z|X)$. To this end, we assume a variational distribution $q_\phi(Z)$ and maximize the evidence lower bound (ELBO),[1] which is given by $\mathscr{L}(x,\phi) = \mathbb{E}_{q_\phi(Z)}[\log p(X,Z)] - \mathbb{E}_{q_\phi(Z)}\left[\log q_\phi(Z)\right]$. Up to this point, we have assumed that the variational distribution over the local variables factorizes with respect to the number of observations $Z$, i.e.,

$$q(Z) = \prod_{n=1}^{N} q_{\phi_n}(z_n).$$

Thus, standard VI makes it necessary to optimize a variational set of parameters $\phi_n$ for each data point $x_n$, which is computationally expensive.

The basic idea behind *amortized* VI is to use a powerful predictor to predict the optimal local variational parameters $\phi_n$ based on the observed features of $x_n$, i.e., $\phi_n = f_\phi(x_n)$. Thus, we can now use an artificial neural network (NN) parameterized by $\phi$, with input $x_n$ and output $\phi_n$, to determine the variational family of distributions:

$$q_\phi(Z) = \prod_{n=1}^{N} q_\phi(z_n|x_n), \tag{2}$$

where we remark that the NN parameters $\phi$ are shared by all the observations $\{x_n\}_{n=1}^N$. In the case of amortized variational inference, the goal is now to find the NN parameters $\phi^*$ that maximize the ELBO:

$$\begin{aligned}\mathscr{L}(x,\phi) &= \mathbb{E}_{x\sim p_{\mathscr{D}}(x)}\left[\left(\mathbb{E}_{q_\phi(z|x)}[\log p(x|z] - \mathrm{KL}(q_\phi(z|x)\| p(z))\right)\right] \\ &= \sum_n \left(\mathbb{E}_{q_\phi(z_n|x_n)}[\log p(x_n|z_n)] - \mathrm{KL}(q_\phi(z_n|x_n)\| p(z_n))\right). \end{aligned} \tag{3}$$

As an example, we may assume a Normal variational distribution over the local variables, $q_\phi(z_n|x_n) = \mathscr{N}(z_n; \mu(x_n), \sigma(x_n))$, where $\mu(x_n)$ and $\Sigma(x_n)$ are two NNs, which may in turn share parameters and take as input the observation $x_n$ and output, respectively, the mean vector $\mu(x_n)$ and the covariance matrix $\Sigma(x_n)$ of $z_n$.

[1] Or equivalently, that minimise the Kullback-Leibler divergence from $q_\phi(Z)$ to $p(Z|X)$ [2].

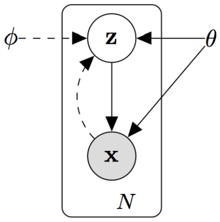

**Figure 1:** Graphical illustration of a VAE. Solid lines denote the generative model $p(z)p_\theta(x|z)$, dashed lines denote the variational approximation $q_\phi(z|x)$ to the intractable posterior $p_\theta(z|x)$. The variational parameters $\phi$ are learned jointly with the generative model parameters $\theta$.

**Obs. 1**: Note that in amortized VI, the number of (local) parameters does not increase with the number of data points in the training set, and training with very large datasets becomes feasible.

# 2 Variational Autoencoders (VAEs)

From Lecture 1, we recall that the objective of a latent variable model, a.k.a. generative model, is to fit the data distribution as accurately as possible by relying on latent variables, such that

$$p(X) = \int p(X,Z)\,dZ. \tag{4}$$

In the previous section, we have seen an example of how to make use of NNs in the context of Bayesian inference, to improve the flexibility as well as the data scalability of variational inference for a given generative model $p(X,Z)$, i.e., $q_\phi(Z|X) \approx p(Z|X)$. However, if the generative model is not flexible enough to accurately capture the underlying structure in the observed data, we will only have an accurate approximation for the posterior of a *poor* latent variable model.

For instance, images are a popular kind of data for which we might create generative models. Each "datapoint" (image) has thousands or millions of dimensions (pixels), and the generative model's job is to somehow capture the dependencies between pixels. In this context, it seems difficult to design a simple generative model (e.g., clustering model) that accurately captures all the dependencies between pixels. Instead, one may propose to use a deep generative model, which relies on NN architectures (e.g., convolutional neural networks) to capture the latent structure of complex high-dimensional data, such as images. As an example, one may propose the following generative model:

$$p_\theta(X) = \int p_\theta(X,Z)\,dZ = \int p_\theta(X|Z)p(Z)\,dZ., \tag{5}$$

where the likelihood model $p_\theta(X|Z)$ is paramaterized using a NN architecture with parameters $\theta$. Thus, the goal here is to find the paramaters of the NN $\theta$ that minimize the KL-divergence between $p_\theta(x)$ and the true data distribution $p_{\mathscr{D}}(x)$, i.e.,

$$\mathrm{KL}(p_{\mathscr{D}}(x)\|\, p_\theta(x)) = \mathbb{E}_{x\sim p_{\mathscr{D}}(x)}[\log p_{\mathscr{D}}(x)] - \mathbb{E}_{x\sim p_{\mathscr{D}}(x)}[\log p_\theta(x)], \tag{6}$$

or equivalently, paramaters $\theta$ that maximize the log-evidence $\mathbb{E}_{x\sim p_{\mathscr{D}}(x)}[\log p_\theta(x)]$, since the first term in the previous expression is independent of the parameters $\theta$.

Unfortunately, the log-evidence $\mathbb{E}_{x\sim p_{\mathscr{D}}(x)}[\log p_\theta(x)]$ is in general hard to evaluate, and thus, to optimize. Fortunately, its lower bound, i.e., the ELBO can be optimized. In more detail, we can now write

$$\mathbb{E}_{x\sim p_{\mathscr{D}}(x)}[\log p_\theta(x)] = \mathscr{L}(x,\theta,\phi) + \mathrm{KL}(q_\phi(z|x)\|\, p_\theta(z|x)) \geq \mathscr{L}(x,\theta,\phi), \tag{7}$$

where we make use of the fact that the $KL \geq 0$ and the ELBO is now given by

$$\begin{aligned}\mathscr{L}(x,\theta,\phi) &= \Big(\mathbb{E}_{q_\phi(z_n|x_n)}[\log p_\theta(x_n|z_n)] - \mathrm{KL}(q_\phi(z_n|x_n)\|\, p(z_n))\Big) && (8)\\ &= \sum_n \Big(\mathbb{E}_{q_\phi(z_n|x_n)}[\log p_\theta(x_n|z_n)] - \mathrm{KL}(q_\phi(z_n|x_n)\|\, p(z_n))\Big). && (9)\end{aligned}$$

In a *Variational autoencoder (VAE)*, which is graphically illustrated in Fig. 1, the goal is to find the set of parameters for both the generative model, $\theta$, and the variational distribution (a.k.a. recognition model), $\phi$, that maximize the ELBO in 9.

**Obs. 1**: We remark that here the goal is to maximize the log-evidence $\mathbb{E}_{x \sim p_{\mathscr{D}}(x)}[\log p_\theta(x)]$, which does not depend on the variational parameters $\phi$. Indeed, the variational distribution $q_\phi(z|x)$ is here introduced as a trick to get a lower bound of the log-evidence. Thus, it is important to notice that when learning the parameters of the generative model $\theta$, maximizing the log-evidence $\mathbb{E}_{x \sim p_{\mathscr{D}}(x)}[\log p_\theta(x)]$ is not equivalent to maximize the ELBO, since as shown in Eq. 7, the KL divergence between the approximation of the posterior and the true posterior, $\mathrm{KL}(q_\phi(z|x) \| p_\theta(z|x))$, depends on $\theta$.

**Obs. 2**: The VAE was originally proposed in [1] and its name arises from the fact that the joint training of the generative and recognition network resembles the structure of autoencoders, a class of unsupervised, deterministic models. Specifically, the first (likelihood) and second (KL) terms in 9 resemble, respectively, the reconstruction loss and the regularization term in a deterministic autoencoder.

**Obs. 3**: However, it can be proved theoretically that VAE may learn to ignore the latent variable $z$, i.e., the code, when the generative model is a universal approximator.

**For a complete overview of advances in VI, including amortized VI and VAE, please refer to [2].**

# References

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

[2] Cheng Zhang, Judith Bütepage, Hedvig Kjellström, and Stephan Mandt. Advances in variational inference. *IEEE transactions on pattern analysis and machine intelligence*, 41(8):2008–2026, 2018.