

is shared with generative models that optimize a log-likelihood, or equivalently, $D_{\text{KL}}(p_{\text{data}} \| p_{\text{model}})$, as argued by Theis *et al.* (2015) and by Huszar (2015). Another troubling issue with contemporary VAE models is that they tend to use only a small subset of the dimensions of \mathbf{z} , as if the encoder was not able to transform enough of the local directions in input space to a space where the marginal distribution matches the factorized prior.

The VAE framework is very straightforward to extend to a wide range of model architectures. This is a key advantage over Boltzmann machines, which require extremely careful model design to maintain tractability. VAEs work very well with a diverse family of differentiable operators. One particularly sophisticated VAE is the **deep recurrent attention writer** or DRAW model (Gregor *et al.*, 2015). DRAW uses a recurrent encoder and recurrent decoder combined with an attention mechanism. The generation process for the DRAW model consists of sequentially visiting different small image patches and drawing the values of the pixels at those points. VAEs can also be extended to generate sequences by defining variational RNNs (Chung *et al.*, 2015b) by using a recurrent encoder and decoder within the VAE framework. Generating a sample from a traditional RNN involves only non-deterministic operations at the output space. Variational RNNs also have random variability at the potentially more abstract level captured by the VAE latent variables.

The VAE framework has been extended to maximize not just the traditional variational lower bound, but instead the **importance weighted autoencoder** (Burda *et al.*, 2015) objective:

$$\mathcal{L}_k(\mathbf{x}, q) = \mathbb{E}_{\mathbf{z}^{(1)}, \dots, \mathbf{z}^{(k)} \sim q(\mathbf{z} | \mathbf{x})} \left[\log \frac{1}{k} \sum_{i=1}^k \frac{p_{\text{model}}(\mathbf{x}, \mathbf{z}^{(i)})}{q(\mathbf{z}^{(i)} | \mathbf{x})} \right]. \quad (20.79)$$

This new objective is equivalent to the traditional lower bound \mathcal{L} when $k = 1$. However, it may also be interpreted as forming an estimate of the true $\log p_{\text{model}}(\mathbf{x})$ using importance sampling of \mathbf{z} from proposal distribution $q(\mathbf{z} | \mathbf{x})$. The importance weighted autoencoder objective is also a lower bound on $\log p_{\text{model}}(\mathbf{x})$ and becomes tighter as k increases.

Variational autoencoders have some interesting connections to the MP-DBM and other approaches that involve back-propagation through the approximate inference graph (Goodfellow *et al.*, 2013b; Stoyanov *et al.*, 2011; Brakel *et al.*, 2013). These previous approaches required an inference procedure such as mean field fixed point equations to provide the computational graph. The variational autoencoder is defined for arbitrary computational graphs, which makes it applicable to a wider range of probabilistic model families because there is no need to restrict the choice