EQ2415 – Machine Learning and Data Science
HT22

Tutorial 9 A. Honoré, A. Ghosh

1 Approximate inference for graphical models

Suppose that $\mathbf{x} \in \mathbb{R}^d$ is an observable random variable and that $\mathbf{z} \in \mathbb{R}^n$ is a latent variable. We model the relation of the these two variables with a graphical model which allow us to calculate the joint distribution $p(\mathbf{x}, \mathbf{z})$. We call the evidence of the data, the log likelihood of the data $\ln p(\mathbf{x})$.

1.1 Variational inference

1.1.1 Evidence Lower Bound

Question 1 Using standard laws of propability, propose two ways to calculate the evidence and explain why they are not feasible in practice.

Solution: We can marginalize:

$$p(\mathbf{x}) = \int_{\mathbb{R}^n} p(\mathbf{x}, \mathbf{z}) d\mathbf{z} \tag{1}$$

but this is not feasible in general because intractable for complex models.

We can also use Bayes rule:

$$p(\mathbf{x}) = \frac{p(\mathbf{x}, \mathbf{z})}{p(\mathbf{z}|\mathbf{x})} \tag{2}$$

But we do not have access to the posterior $p(\mathbf{z}|\mathbf{x})$ ■

Question 2 Suppose that we have access to an approximate distribution $q_\phi(\mathbf{z}|\mathbf{x})$ of the true posterior $p(\mathbf{z}|\mathbf{x})$ where ϕ is a set of parameters. Show that $E_{q_\phi(\mathbf{z}|\mathbf{x})}[\ln \frac{p(\mathbf{x},\mathbf{z})}{q_\phi(\mathbf{z}|\mathbf{x})}]$ is a lower bound for $\ln p(\mathbf{x})$.

Recall:

- The KL-divergence between two probability distribution p and q:

$$D_{KL}(q||p) = E_q\left[\ln \frac{q}{p}\right] \geq 0 \tag{3}$$

Solution:

$$\begin{aligned}
\ln p(\mathbf{x}) &= \ln p(\mathbf{x}) \int q_\phi(\mathbf{z}|\mathbf{x}) d\mathbf{z} &(4)\\
&= E_{q_\phi(\mathbf{z}|\mathbf{x})}\left[\ln p(\mathbf{x})\right] &(5)\\
&= E_{q_\phi(\mathbf{z}|\mathbf{x})}\left[\ln \frac{p(\mathbf{x}, \mathbf{z})}{p(\mathbf{z}|\mathbf{x})}\right] &(6)\\
&= E_{q_\phi(\mathbf{z}|\mathbf{x})}\left[\ln \frac{p(\mathbf{x}, \mathbf{z}) q_\phi(\mathbf{z}|\mathbf{x})}{p(\mathbf{z}|\mathbf{x}) q_\phi(\mathbf{z}|\mathbf{x})}\right] &(7)\\
&= E_{q_\phi(\mathbf{z}|\mathbf{x})}\left[\ln \frac{p(\mathbf{x}, \mathbf{z})}{q_\phi(\mathbf{z}|\mathbf{x})}\right] + E_{q_\phi(\mathbf{z}|\mathbf{x})}\left[\ln \frac{q_\phi(\mathbf{z}|\mathbf{x})}{p(\mathbf{z}|\mathbf{x})}\right] &(8)\\
&= E_{q_\phi(\mathbf{z}|\mathbf{x})}\left[\ln \frac{p(\mathbf{x}, \mathbf{z})}{q_\phi(\mathbf{z}|\mathbf{x})}\right] + D_{KL}(q_\phi(\mathbf{z}|\mathbf{x})||p(\mathbf{z}|\mathbf{x})) &(9)\\
&\geq E_{q_\phi(\mathbf{z}|\mathbf{x})}\left[\ln \frac{p(\mathbf{x}, \mathbf{z})}{q_\phi(\mathbf{z}|\mathbf{x})}\right] &(10)
\end{aligned}$$

with equality between the ELBO and the evidence when $D_{KL} = 0$. ■

Question 3 Let us now introduce variational auto-encoders. Let us introduce parameters for the conditional distribution $p_\theta(\mathbf{x}|\mathbf{z})$. Show that maximizing the ELBO consists in maximizing a "reconstruction" cost: $E_{q_\phi(\mathbf{z}|\mathbf{x})}[\ln p_\theta(\mathbf{x}|\mathbf{z})]$ and minimizing a "prior matching" term: $D_{KL}(q_\phi(\mathbf{z}|\mathbf{x})||p(\mathbf{z}))$.

Solution:

$$E_{q_\phi(\mathbf{z}|\mathbf{x})}\left[\ln \frac{p_\theta(\mathbf{x},\mathbf{z})}{q_\phi(\mathbf{z}|\mathbf{x})}\right] = E_{q_\phi(\mathbf{z}|\mathbf{x})}\left[\ln \frac{p_\theta(\mathbf{x}|\mathbf{z})p(\mathbf{z})}{q_\phi(\mathbf{z}|\mathbf{x})}\right] \tag{11}$$

$$= E_{q_\phi(\mathbf{z}|\mathbf{x})}[\ln p_\theta(\mathbf{x}|\mathbf{z})] + E_{q_\phi(\mathbf{z}|\mathbf{x})}\left[\ln \frac{p(\mathbf{z})}{q_\phi(\mathbf{z}|\mathbf{x})}\right] \tag{12}$$

$$= E_{q_\phi(\mathbf{z}|\mathbf{x})}[\ln p_\theta(\mathbf{x}|\mathbf{z})] - D_{KL}(q_\phi(\mathbf{z}|\mathbf{x})||p(\mathbf{z})) \tag{13}$$

■

Question 4 In the context of an auto-encoder (See figure 1), what quantity can be interpreted as the encoder for a vector $\mathbf{x}$ and what quantity can be interpreted as a decoder of a vector $\mathbf{z}$?

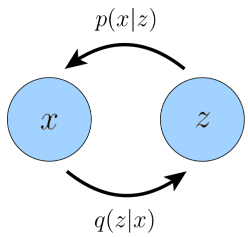

Figure 1: Graphical model for variational autoencoders.[1]

Solution: The encoder for a vector $\mathbf{x}$: $q_\phi(\mathbf{z}|\mathbf{x})$. The decoder for a latent vector $\mathbf{z}$: $p_\theta(\mathbf{x}|\mathbf{z})$ ■

Question 5 Let us consider a generalization of VAEs: Markovian Hierarchical VAEs (Markovian HVAE, Figure 2).

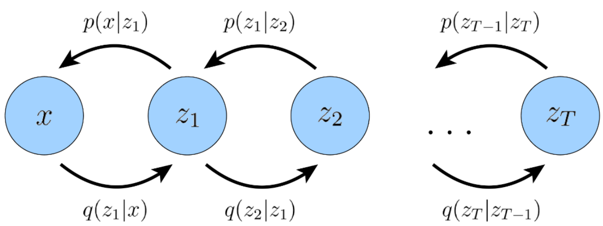

Figure 2: Markovian HVAE [1]

Question 5a Factorize the joint distribution

$$p_\theta(\mathbf{x}, \mathbf{z}_1, \ldots, \mathbf{z}_T), \tag{14}$$

in terms of the quantities on the edges of the graph in figure 2?

Solution:

$$p_\theta(\mathbf{x}, \mathbf{z}_1, \ldots, \mathbf{z}_T) = p(\mathbf{z}_T)p_\theta(\mathbf{x}|\mathbf{z}_1)\prod_{t=2}^{T} p_\theta(\mathbf{z}_{t-1}|\mathbf{z}_t). \tag{15}$$

■

Question 5b Similarly, factorize the posterior of the Markovian HVAE:

$$q_\phi(\mathbf{z}_1, \ldots, \mathbf{z}_T|\mathbf{x}). \tag{16}$$

Solution:

$$q_\phi(\mathbf{z}_1, \ldots, \mathbf{z}_T|\mathbf{x}) = q_\phi(\mathbf{z}_1|\mathbf{x}) \prod_{t=2}^{T} q_\phi(\mathbf{z}_t|\mathbf{z}_{t-1}) \tag{17}$$

■

1.1.2 Variational diffusion models (VDM)

The easiest way to think of a Variational Diffusion Model (VDM) is simply as a restricted Markovian Hierarchical Variational Autoencoder. The architecture, depicted on figure 3, is behind stable diffusion models.

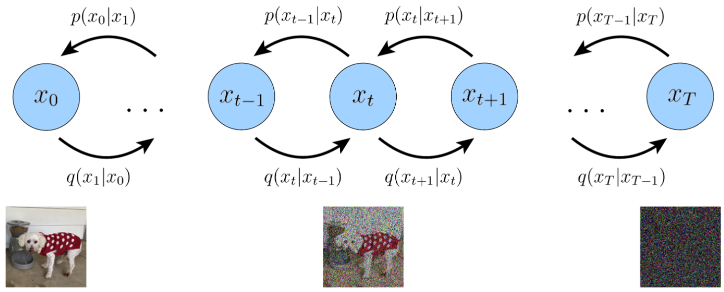

Figure 3: A visual representation of a Variational Diffusion Model; x_0 represents true data observations such as natural images, x_T represents pure Gaussian noise, and x_t is an intermediate noisy version of x_0. Each $q(x_t|x_{t-1})$ is modeled as a Gaussian distribution that uses the output of the previous state as its mean. [1]

2 Variational distributions

2.1 Factorized approximation

Consider a factorized variational distribution $q(\mathbf{z})$ of the form:

$$q(\mathbf{z}) = \prod_{i=1}^{M} q_i(\mathbf{z}_i) \tag{18}$$

Question 1: By using the technique of Lagrange multipliers, verify that the minimization of the Kullback-Leibler divergence $KL(p||q)$ with respect to one of the factors $q_j(\mathbf{z}_j)$, keeping all other factors fixed, leads to the solution:

$$q_j^\star(\mathbf{z}_j) = \int p(\mathbf{z}) \prod_{i \neq j} d\mathbf{z}_i = p(\mathbf{z}_j) \tag{19}$$

Solution: We will write the KL divergence and then minimize it with respect to a factor $q_j(\mathbf{z}_j)$.

We start by writing the KL divergence:

$$D_{KL}(p(\mathbf{z})||q(\mathbf{z})) = -\int p(\mathbf{z}) \left[\sum_{i=1}^{M} \ln q_i(\mathbf{z}_i)\right] d\mathbf{z} + cst \tag{20}$$

the cst term depends only on $p(\mathbf{z})$ and will be removed when deriving, next we isolate the term indexed with j.

$$
\begin{aligned}
D_{KL} &= -\int p(\mathbf{z}) \left[\ln q_j(\mathbf{z}_j) + \sum_{i \neq j}^{M} \ln q_i(\mathbf{z}_i) \right] d\mathbf{z} + cst \\
&= -\int p(\mathbf{z}) \ln q_j(\mathbf{z}_j) d\mathbf{z} + cst \\
&= -\int \left[\int p(\mathbf{z}) \prod_{i \neq j} d\mathbf{z}_i \right] \ln q_j(\mathbf{z}_j) d\mathbf{z}_j + cst \\
&= -\int p(\mathbf{z}_j) \ln q_j(\mathbf{z}_j) d\mathbf{z}_j + cst
\end{aligned} \tag{21}
$$

where the factors $q_i(\mathbf{z}_i)$ with $i \neq j$ are in the cst term, and we used the definition of the marginal $p(\mathbf{z}_j) = \int p(\mathbf{z}) \prod_{i \neq j}$ in the last step.

Before deriving, we need to construct an objective that enforces that the marginal factor $q_j(\mathbf{z}_j)$ integrates to 1, we use a Lagrangian multiplier:

$$L(q_j(\mathbf{z}_j), \lambda) = -\int p(\mathbf{z}_j) \ln q_j(\mathbf{z}_j) d\mathbf{z}_j + \lambda(\int q_j(\mathbf{z}_j) d\mathbf{z}_j - 1) \tag{22}$$

$$\text{Deriving wrt} q_j(\mathbf{z}_j) \text{ and setting to 0} \tag{23}$$

$$-\frac{p(\mathbf{z}_j)}{q_j^*(\mathbf{z}_j)} + \lambda = 0 \tag{24}$$

$$\lambda q_j(\mathbf{z}_j) = p(\mathbf{z}_j) \tag{25}$$

$$\text{Integrating both sides, we find } \lambda = 1, \text{ this gives} \tag{26}$$

$$q_j^*(\mathbf{z}_j) = p(\mathbf{z}_j) \tag{27}$$

References

[1] C. Luo, "Understanding Diffusion Models: A Unified Perspective," Aug. 2022.

■