

Generative Models - DRE7053

Lecture 2

NORA Summer School 2024

Rogelio A Mancisidor
Assistant Professor
Department of Data Science and Analytics
BI Norwegian Business School

June 11, 2024

1 Variational Autoencoder

- Assumptions and Background
- Amortized Inference
- The ELBO
- ELBO Closed Form Solution
- Reparameterization Trick
- Reparameterized Gradients vs Score Gradients

2 Should we optimize the ELBO?

- Bound on Mutual Information
- Posterior Collapse
- Yet another way to derive the ELBO

3 Semi-supervised Learning with VAEs

- Theoretical Background
- Generative and Inference Models
- Semi-supervised ELBO

1 Variational Autoencoder

- Assumptions and Background
- Amortized Inference
- The ELBO
- ELBO Closed Form Solution
- Reparameterization Trick
- Reparameterized Gradients vs Score Gradients

2 Should we optimize the ELBO?

- Bound on Mutual Information
- Posterior Collapse
- Yet another way to derive the ELBO

3 Semi-supervised Learning with VAEs

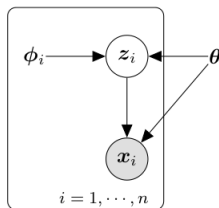
- Theoretical Background
- Generative and Inference Models
- Semi-supervised ELBO

Variational Autoencoder - I

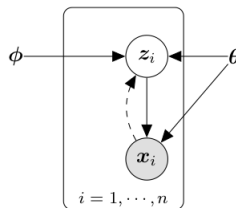
- The Variational Autoencoder (VAE) is an example of LVM where the posterior distribution is approximated using the variational inference principle
- Assume we observe $\mathbf{X} = \{\mathbf{x}_i\}_i^n$, and for each $\mathbf{x}_i \in \mathbb{R}^{d_x}$ we have one latent variable $\mathbf{z} \in \mathbb{R}^{d_z}$. Hence, $\mathbf{Z} = \{\mathbf{z}_i\}_i^n$.
- VAE assumes a *mean-field* factorization

$$q(\mathbf{Z}|\mathbf{X}; \phi) = \prod_i^n q_i(\mathbf{z}_i|\mathbf{x}_i; \phi), \quad (1)$$

Amortized Inference



(a)



(b)

- a) panel mean-field approximation, and b) panel VAE
- Note that ϕ does not depend on the i -th latent variable
- Amortized variational inference shares ϕ across all data points!

Variational Autoencoder - II

- Generative model
 - $p(\mathbf{z})p(\mathbf{x}|\mathbf{z})$
- Inference (recognition) model
 - $q(\mathbf{z}|\mathbf{x})$
- VAE assumes the following distributions:

$$\begin{aligned}p(\mathbf{z}) &\sim \mathcal{N}(\mathbf{0}, \mathbf{1}) \\p(\mathbf{x}|\mathbf{z}) &\sim f(\cdot) \\q(\mathbf{z}|\mathbf{x}) &\sim \mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma}),\end{aligned}$$

where $\boldsymbol{\Sigma}$ is a diagonal matrix with main diagonal $\boldsymbol{\sigma}^2 = (\sigma_1^2, \dots, \sigma_{d_z}^2)$

- $p(\mathbf{x}|\mathbf{z})$ can take different distributions depending on the data, e.g. Gaussian, Bernoulli, Laplace, etc.

Another way to derive the ELBO

- Note that $\text{ELBO} = \mathbb{E}_q[\log p(\mathbf{x}|\mathbf{z}) + \log p(\mathbf{z}) - \log q(\mathbf{z}|\mathbf{x})]$ is composed by

$$\int q(\mathbf{z}|\mathbf{x}) \log p(\mathbf{x}|\mathbf{z}) d\mathbf{z} = \int \mathcal{N}(\mathbf{z}|\mathbf{x}) \log \mathcal{N}(\mathbf{x}|\mathbf{z}) d\mathbf{z} \quad (2)$$

$$\int q(\mathbf{z}|\mathbf{x}) \log p(\mathbf{z}) d\mathbf{z} = \int \mathcal{N}(\mathbf{z}|\mathbf{x}) \log \mathcal{N}(\mathbf{z}) d\mathbf{z} \quad (3)$$

$$- \int q(\mathbf{z}|\mathbf{x}) \log q(\mathbf{z}|\mathbf{x}) d\mathbf{z} = - \int \mathcal{N}(\mathbf{z}|\mathbf{x}) \log \mathcal{N}(\mathbf{z}|\mathbf{x}) d\mathbf{z} \quad (4)$$

- According to Lemma 1 in [11] (page 48)

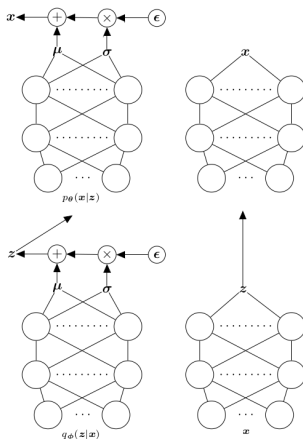
$$\int q(\mathbf{x}) \log p(\mathbf{z}) d\mathbf{x} = \sum_{j=1}^{d_x} -\frac{1}{2} \log(2\pi\sigma_{1,j}^2) - \frac{\sigma_{2,j}^2}{2\sigma_{1,j}^2} - \frac{(\mu_{2,j} - \mu_{1,j})^2}{2\sigma_{1,j}^2} \quad (5)$$

where $\sigma_{i,j}^2$ and $\mu_{i,j}$ are the j -th element of their respective $\boldsymbol{\mu}_1$ and $\boldsymbol{\sigma}_1^2$ parameters of $p(\mathbf{x})$ or $\boldsymbol{\mu}_2$ and $\boldsymbol{\sigma}_2^2$ of $q(\mathbf{x})$.

Closed form

Variational Autoencoder - III

- $q(\mathbf{z}|\mathbf{x})$ is often referred to as *probabilistic encoder*
- $p(\mathbf{x}|\mathbf{z})$ is often referred to as *probabilistic decoder*
- The reason is its similarity with autoencoders



Reparameterization Trick

- The VAE parameterize the distribution parameters with neural networks, i.e.,

$$p(\mathbf{x}|\mathbf{z}) \sim \mathcal{N}(\mathbf{x}|\mathbf{z}; \boldsymbol{\mu}_{\mathbf{x}|\mathbf{z}} = f_{\boldsymbol{\theta}}(\mathbf{z}), \boldsymbol{\sigma}_{\mathbf{x}|\mathbf{z}}^2 = f_{\boldsymbol{\theta}}(\mathbf{z})), \quad (6)$$

and

$$q(\mathbf{z}|\mathbf{x}) \sim \mathcal{N}(\mathbf{z}|\mathbf{x}; \boldsymbol{\mu}_{\mathbf{z}|\mathbf{x}} = f_{\boldsymbol{\phi}}(\mathbf{x}), \boldsymbol{\sigma}_{\mathbf{z}|\mathbf{x}}^2 = f_{\boldsymbol{\phi}}(\mathbf{x})), \quad (7)$$

where $f_{\boldsymbol{\phi}}(\mathbf{x})$ and $f_{\boldsymbol{\theta}}(\mathbf{z})$ are neural networks with trainable parameters $\boldsymbol{\phi}$ and $\boldsymbol{\theta}$

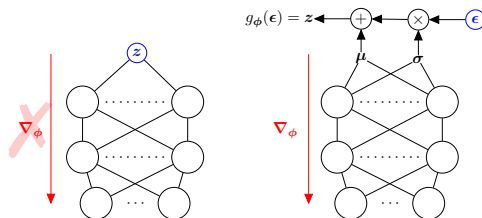
- Note that

$$\begin{aligned} \text{ELBO} &= \mathbb{E}_q \left[\log \frac{p_{\boldsymbol{\theta}}(\mathbf{x}, \mathbf{z})}{q_{\boldsymbol{\phi}}(\mathbf{z}|\mathbf{x})} \right] = \mathbb{E}_q [\log p_{\boldsymbol{\theta}}(\mathbf{x}|\mathbf{z})] - KL[q_{\boldsymbol{\phi}}(\mathbf{z}|\mathbf{x}) || p(\mathbf{z})] \\ &= \mathbb{E}_q [f_{\boldsymbol{\theta}}(\mathbf{z})] - KL[q_{\boldsymbol{\phi}}(\mathbf{z}|\mathbf{x}) || p(\mathbf{z})] \end{aligned}$$

- Remember, the closed form solution includes the parameters of $q(\mathbf{z}|\mathbf{x})$!

Reparameterization Trick - II

- To backpropagate $q_{\phi}(\mathbf{z}|\mathbf{x})$ we adopt the following architecture



- But the reparameterization trick is more than that...

Reparameterized Gradients

- Use an invertible function, e.g.

$$\mathbf{z} = g_{\phi}(\epsilon) = \boldsymbol{\mu} + \boldsymbol{\sigma}\epsilon, \quad (8)$$

where $\epsilon \sim N(0, 1)$.

- Use the *change of variable* result (see Lemma 2 in [11]) that says

$$\begin{aligned} \int q(\mathbf{z}|\mathbf{x})f(\mathbf{z})d\mathbf{z} &= \int p(\epsilon)f(\mathbf{z})d\epsilon \\ &= \int p(\epsilon)f(g_{\phi}(\epsilon))d\epsilon \\ \mathbb{E}_q[f(\mathbf{z})] &= \mathbb{E}_p[f(g_{\phi}(\epsilon))] \end{aligned} \quad (9)$$

- Therefore, the Monte Carlo estimate

$$\frac{1}{L} \sum_{l=1}^L \log p(\mathbf{x}_i | \mathbf{z}_i = \boldsymbol{\mu}_i + \boldsymbol{\sigma}_i \epsilon_i)$$

is an expectation over $p(\epsilon)$!

Variance of Reparameterized and Score Gradients

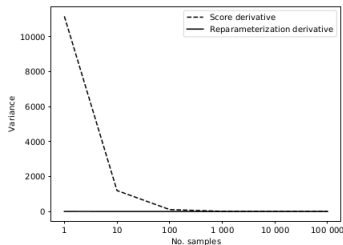
- Assume that $p(x) \sim \mathcal{N}(\theta, 1)$ and we want to minimize

$$\arg \min_{\theta} \mathbb{E}_p[x^2]. \quad (10)$$

- The score derivative is given by
- Use the reparameterization $x = \theta + \epsilon$ where $q(\epsilon) \sim \mathcal{N}(0, 1)$.
Therefore, $\mathbb{E}_p[x^2] = \mathbb{E}_q[(\theta + \epsilon)^2]$ and its derivative is

Simulation

- We simulate $N = [1, 10, 100, 1000, 10000, 100000]$ samples from $p(x) \sim \mathcal{N}(\theta, 1)$, where $\theta = 10$, and $q(\epsilon) \sim \mathcal{N}(0, 1)$ to estimate the variance of 100 Monte Carlo estimates of Equation ?? and ??.



- 1 Variational Autoencoder
 - Assumptions and Background
 - Amortized Inference
 - The ELBO
 - ELBO Closed Form Solution
 - Reparameterization Trick
 - Reparameterized Gradients vs Score Gradients
- 2 Should we optimize the ELBO?
 - Bound on Mutual Information
 - Posterior Collapse
 - Yet another way to derive the ELBO
- 3 Semi-supervised Learning with VAEs
 - Theoretical Background
 - Generative and Inference Models
 - Semi-supervised ELBO

Bound on Mutual Information

- Let's define the following distributions

$$q_{\phi}(\mathbf{x}, \mathbf{z}) = p(\mathbf{x})q_{\phi}(\mathbf{z}|\mathbf{x}) \quad (11)$$

$$q_{\phi}(\mathbf{z}) = \mathbb{E}_{p(\mathbf{x})}[q_{\phi}(\mathbf{z}|\mathbf{x})] \approx \frac{1}{N} \sum_n q(\mathbf{z}|\mathbf{x}_n) \quad (12)$$

$$q_{\phi}(\mathbf{x}, \mathbf{z}) = p_{\theta}(\mathbf{x}|\mathbf{z})q_{\phi}(\mathbf{z}) \quad (13)$$

- Note

$$l(\mathbf{z}, \mathbf{x}) = \mathbb{E}_{q_{\phi}(\mathbf{z}, \mathbf{x})} \left[\log \frac{q_{\phi}(\mathbf{z}, \mathbf{x})}{p(\mathbf{x})q_{\phi}(\mathbf{z})} \right]$$

- Meaning that

(14)

- What does that mean?

Posterior Collapse

- In practice, we maximize

$$\text{ELBO} = \mathbb{E}_{p(\mathbf{x})} [\mathbb{E}_q[\log p_\theta(\mathbf{x}|\mathbf{z})] - \text{KL}[q_\phi(\mathbf{z}|\mathbf{x})||p(\mathbf{z})]] \quad (15)$$

- We know that $\mathbb{E}_{p(\mathbf{x})} \text{KL}[q_\phi(\mathbf{z}|\mathbf{x})||p(\mathbf{z})] \geq I(\mathbf{z}, \mathbf{x})$
- Minimizing the average KL makes the posterior *collapse* into the prior, i.e.

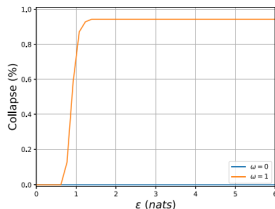
$$q_\phi(\mathbf{z}|\mathbf{x}) \approx p(\mathbf{z})$$

meaning \mathbf{z} is independent of \mathbf{x} !!!

- Ideally VAEs should embed as much information of \mathbf{x} into \mathbf{z}

Measuring Posterior Collapse

- We measure posterior collapse as the proportion of latent dimensions that are within ϵ KL divergence of the prior for at least 99% of the data sample.



Yet another way to derive the ELBO - I

- The expectation in equation 15 is taken with the empirical distribution of the data, i.e.

$$\text{ELBO} = \frac{1}{N} \sum_{n=1}^N \mathbb{E}_q[\log p_{\theta}(\mathbf{x}_n | \mathbf{z}_n)] - KL[q_{\phi}(\mathbf{z}_n | \mathbf{x}_n) || p(\mathbf{z}_n)] \quad (16)$$

- The term-by-term KL is minimized when $q_{\phi}(\mathbf{z}_n | \mathbf{x}_n) = p(\mathbf{z}_n)$ for all n
- Using the derivation in equation 14 we obtain

$$\frac{1}{N} \sum_{n=1}^N KL[q_{\phi}(\mathbf{z}_n | \mathbf{x}_n) || p(\mathbf{z}_n)] = I(\mathbf{x}, \mathbf{z}) + KL[q_{\phi}(\mathbf{z}_n) || p(\mathbf{z}_n)] \quad (17)$$

Yet another way to derive the ELBO - II

- Therefore

$$\begin{aligned} ELBO &= \frac{1}{N} \sum_{n=1}^N \log p(\mathbf{x}_n | \mathbf{z}_n) \rightarrow \textcircled{1} \\ &\quad - (\log N - \mathbb{E}_{q(\mathbf{z})} [\mathbb{H}(p(\mathbf{x} | \mathbf{z}))]) \rightarrow \textcircled{2} \\ &\quad - KL[q_\phi(\mathbf{z}) || p(\mathbf{z})] \rightarrow \textcircled{3} \end{aligned} \tag{18}$$

- ① average reconstruction
- ② mutual information
- ③ marginal (aggregated) KL divergence

- ① and ② are in tension with each other. Good reconstructions required \mathbf{z}_n to be specific to \mathbf{x}_n which corresponds to a low entropy.
- ② is bounded below and above

$$0 \leq \log N - \mathbb{E}_{q(\mathbf{z})}[\mathbb{H}(p(\mathbf{x}|\mathbf{z}))] \leq \log N$$

- The prior only appears in ③. We could choose the prior to be $q(\mathbf{z})$, so the divergence is 0.

- 1 Variational Autoencoder
 - Assumptions and Background
 - Amortized Inference
 - The ELBO
 - ELBO Closed Form Solution
 - Reparameterization Trick
 - Reparameterized Gradients vs Score Gradients
- 2 Should we optimize the ELBO?
 - Bound on Mutual Information
 - Posterior Collapse
 - Yet another way to derive the ELBO
- 3 Semi-supervised Learning with VAEs
 - Theoretical Background
 - Generative and Inference Models
 - Semi-supervised ELBO

Theoretical Background

- Semi-supervised learning considers the problem of classification when only a subsets of data has class labels
- We observe N pairs of labeled data

$$(\mathbf{X}, \mathbf{Y}) = \{(\mathbf{x}, y)_1, (\mathbf{x}, y)_2, \dots, (\mathbf{x}, y)_N\}$$

and M unlabeled observations

$$\mathbf{X} = \{\mathbf{x}_{n+1}, \mathbf{x}_{n+2}, \dots, \mathbf{x}_{N+M}\}$$

The M2 model

- We assume the following generative model

$$p(\mathbf{x}, y, \mathbf{z}) = p(\mathbf{z})p(y)p(\mathbf{x}|\mathbf{z}, y)$$

$$p(\mathbf{z}) \sim \mathcal{N}(\mathbf{0}, \mathbf{1})$$

$$p(y) \sim \text{Cat}(\boldsymbol{\pi}_0)$$

$$p(\mathbf{x}|\mathbf{z}, y) \sim f_{\boldsymbol{\theta}}(\boldsymbol{\theta})$$

- The inference model is factorized as $q(\mathbf{z}, y|\mathbf{x}) = q(\mathbf{z}|\mathbf{x}, y)q(y|\mathbf{x})$

$$q(\mathbf{z}|\mathbf{x}, y) \sim \mathcal{N}(\boldsymbol{\mu} = f_{\phi}(\mathbf{x}, y), \boldsymbol{\sigma}^2 = f_{\phi}(\mathbf{x}, y))$$

$$q(y|\mathbf{x}) \sim \text{Cat}(\boldsymbol{\pi} = \mathbf{f}_{\phi}(\mathbf{x}))$$

ELBO and Training Scheme

- Labeled data
- Unlabeled data