Auto-Encoding Variational Bayes

베이즈 스터디

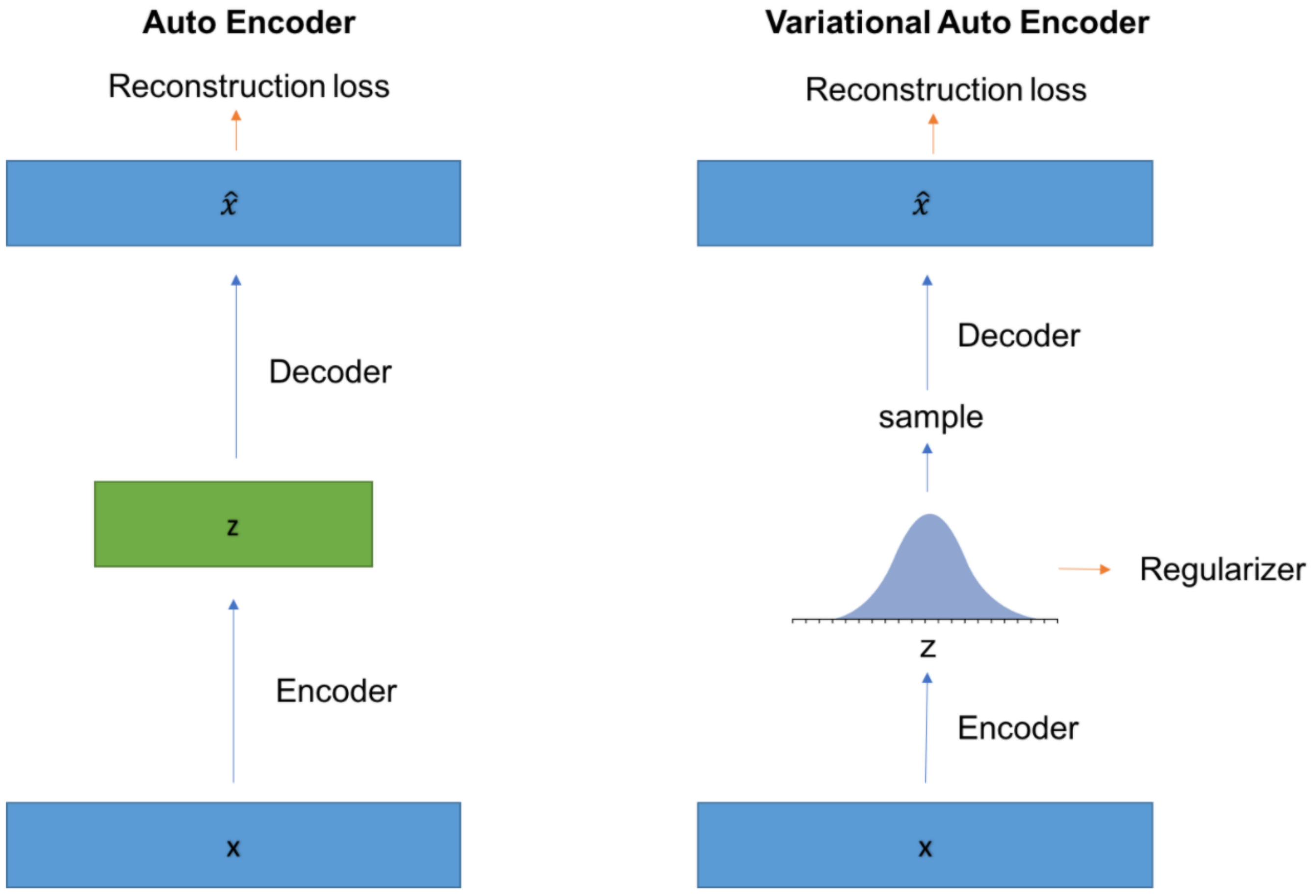
이현정

목차

1. Introduction
- VAE
- Variational Inference
2. Methods
- Scenario
- Variational Lower Bound
- Reparameterization tricK
- Stochastic Gradient Variational Bayes / AEVB Algorithm

Introduction - VAE

- Auto-encoder는 입력 데이터 x 자신을 다시 만들어내려는 neural network 모델로 VAE는 Auto-Encoder 특성을 물려 받음.
- 구조는 그림처럼 latent code z를 만드는 encoder와 x를 만드는 decoder가 맞붙어 있는 형태.
- Auto-encoder에서는 z가 단순히 계산 중간에 나오는 deterministic한 값일 뿐.
- 반면, VAE에서는 latent variable z가 continuous한 분포를 가지는 random variable이고, latent variable z의 분포는 training 과정에서 data로부터 학습됨.
- 즉, latent variable z는 평균과 표준편차로부터 결정되는 확률 분포를 갖는다



Introduction – VAE & AEVB

- VAE Encoder: encoder는 주어진 x로부터 z를 얻을 확률 p(z|x) (recognition model)
- VAE Decoder: decoder는 z로부터 x를 얻을 확률 p(x|z) (generative model)
- Encoder 부분인 p(z|x)는 Bayes 정리를 통해 우리가 다뤘던 posterior distribution에 해당함
- 많은 경우 이러한 posterior distribution을 다루기가 어렵다. (intractable)
- 그래서 이 논문에서는 Encoder 에 해당하는 posterior distribution 대신 계산할 수 있는 q(z|x)라는 분포를 대신 도입해 p(z|x)로 근사 시키는 방법을 사용함.
- 이런 방법을 *Variational Bayesian methods* 또는 *Variational Inference*라고 부르고, VAE의 'Variational'도 거기에서 온 것

- Auto-Encoding Variational Bayes는 q(z|x)를 inference하는 방식에 관한 논문이고, 파생되어서 나온 architecture가 VAE임.

- VAE의 loss fuction에 해당하는 부분의 수리적인 내용이 어떻게 도출 됐는지에 관한 논문이라 할 수 있음

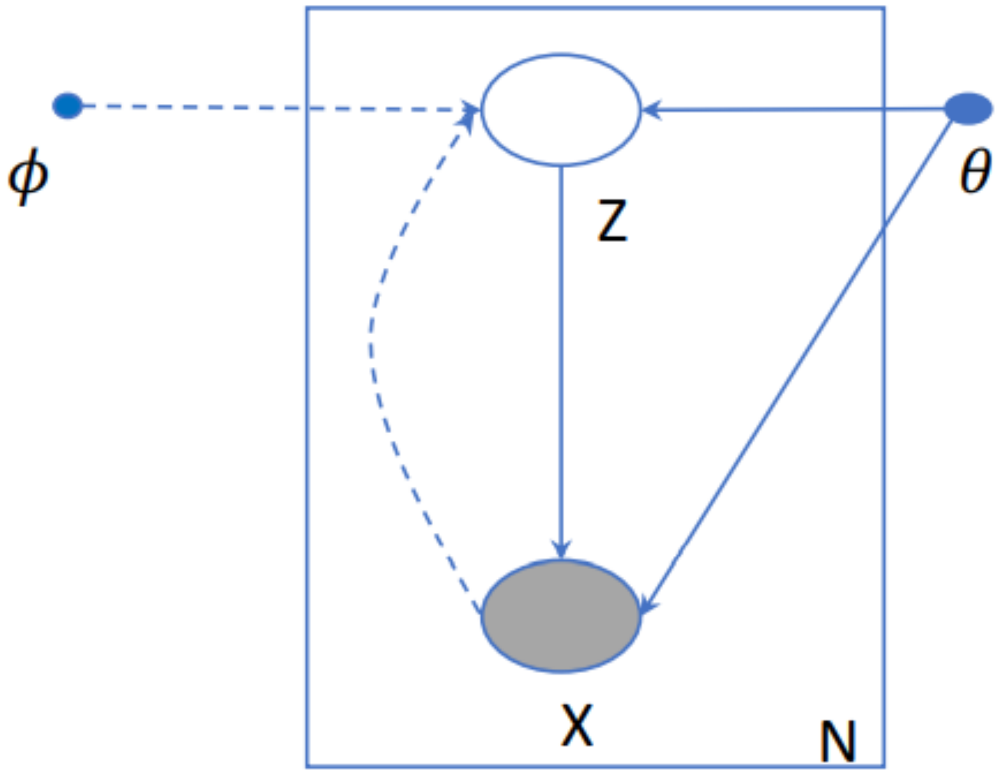
Introduction – Variational Inference

- 관측된 데이터 X가 주어졌을 때, 관측되지 않은 latent variable z의 분포는, simple한 distribution q를 가정하는데 이를 variational distribution이라 함
- 이때, q를 inference하는 것을 variational inference라 하는데 원래의 확률분포 p와 q의 dissimilarity를 d라 가정하고, 이 크기가 가장 작은 q를 찾는 과정을 뜻함.
- 가장 많이 사용하는 dissimilarity척도가 Kullback-Leibler divergence이다.

$$KL(P||Q) = E_P[log\frac{P}{Q}] = E_P[logP - logQ] = \sum_i P(i)log\frac{P(i)}{Q(i)}$$

Methods – Scenario

- probabilistic graphical model
- The process consists of two steps:
- (1) a value z(i) is generated from some prior distribution pθ*(z)
- (2) a value x(i) is generated from some conditional distribution pθ*(x|z).
- We assume that the prior pθ*(z) and likelihood pθ*(x|z) come from parametric families of distributions pθ(z) and pθ(x|z), and their PDFs are differentiable almost everywhere w.r.t both θ and z.



[AEVB algorithm]

Methods – Variational Lower Bound

- derive a lower bound estimator(a stochastic objective function)
- for a variety of directed graphical models with continuous latent variables
- variational lower bound: 두 분포 사이의 거리인 KL divergence를 최소화 시키기 위해 도입되는 개념
- qλ*(z|x)=argminλKL(qλ(z|x)||p(z|x)).

$$KL(q_\lambda(z|x)||p(z|x))$$

$$= E_{q_\lambda}[logq_\lambda(z|x) - logp(z|x)]$$

$$= E_{q_\lambda}[logq_\lambda(z|x) - log\frac{p(x,z)}{p(x)}]$$

$$= E_{q_\lambda}[logq_\lambda(z|x) - logp(x,z) + logp(x)]$$

$$= E_{q_\lambda}[logq_\lambda(z|x) - logp(x,z)] + logp(x)$$

$$= E_{q_\lambda}[logq_\lambda(z|x)] - E_{q_\lambda}[logp(x,z)] + logp(x)$$

Methods – Variational Lower Bound

$$logp(x) = E_{q_\lambda}[logp(x,z)] - E_{q_\lambda}[logq_\lambda(z|x)] + KL(q_\lambda(z|x)||p(z|x))$$

$$ELBO(\lambda) = E q_\lambda[logp(x,z)] - E q_\lambda[logq_\lambda(z|x)]$$

$$logp(x) = ELBO(\lambda) + KL(q_\lambda(z|x)||p(z|x))$$

$$\begin{aligned} &ELBO(\lambda) \\ &= E_{q_\lambda}[\log p(x,z)] - E_{q_\lambda}[\log q_\lambda(z|x)] \\ &= E_{q_\lambda}[\log p(x,z) - \log q_\lambda(z|x)] \\ &= E_{q_\lambda}[\log p(x|z)p(z) - \log q_\lambda(z|x)] \\ &= E_{q_\lambda}[\log p(x|z) + \log p(z) - \log q_\lambda(z|x)] \\ &= E_{q_\lambda}[\log p(x|z) - (\log q_\lambda(z|x) - \log p(z))] \\ &= E_{q_\lambda}[\log p(x|z)] - KL(q_\lambda(z|x)||p(z)) \end{aligned}$$

$$ELBO_i(\lambda) = E_{q_\lambda(z|x_i)}[\log p(x_i|z)] - KL(q_\lambda(z|x_i)||p(z))$$

- Kullback-Leibler divergence는 항상 0보다 같거나 크므로 ELBO 는 logp(x)가 될 수 있는 최솟값을 뜻한다. 여기서, ELBO는 evidence lower bound를 뜻함.
- logp(x)≥ELBO(λ)
- KL divergence를 최소화하는 것은 ELBO를 최대화 하는 것과 같고, 목적함수가 됨.

Methods – Variational Lower Bound

The marginal likelihood is composed of a sum over the marginal likelihoods of individual datapoints $\log p_{\boldsymbol{\theta}}(\mathbf{x}^{(1)}, \cdots, \mathbf{x}^{(N)}) = \sum_{i=1}^{N} \log p_{\boldsymbol{\theta}}(\mathbf{x}^{(i)})$, which can each be rewritten as:

$$\log p_{\boldsymbol{\theta}}(\mathbf{x}^{(i)}) = D_{KL}(q_{\boldsymbol{\phi}}(\mathbf{z}|\mathbf{x}^{(i)})||p_{\boldsymbol{\theta}}(\mathbf{z}|\mathbf{x}^{(i)})) + \mathcal{L}(\boldsymbol{\theta}, \boldsymbol{\phi}; \mathbf{x}^{(i)}) \tag{1}$$

The first RHS term is the KL divergence of the approximate from the true posterior. Since this KL-divergence is non-negative, the second RHS term $\mathcal{L}(\boldsymbol{\theta}, \boldsymbol{\phi}; \mathbf{x}^{(i)})$ is called the (variational) *lower bound* on the marginal likelihood of datapoint i, and can be written as:

$$\log p_{\boldsymbol{\theta}}(\mathbf{x}^{(i)}) \geq \mathcal{L}(\boldsymbol{\theta}, \boldsymbol{\phi}; \mathbf{x}^{(i)}) = \mathbb{E}_{q_{\boldsymbol{\phi}}(\mathbf{z}|\mathbf{x})}\left[-\log q_{\boldsymbol{\phi}}(\mathbf{z}|\mathbf{x}) + \log p_{\boldsymbol{\theta}}(\mathbf{x}, \mathbf{z})\right] \tag{2}$$

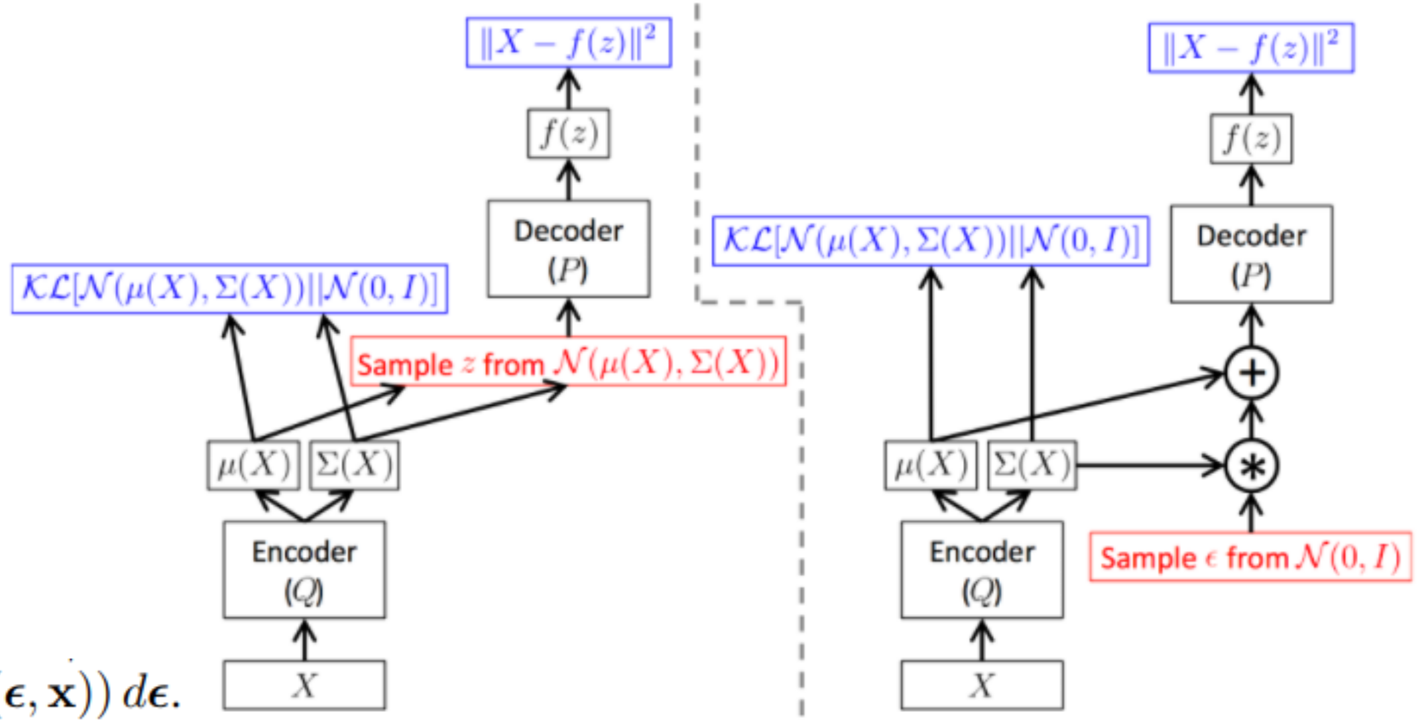
which can also be written as:

$$\mathcal{L}(\boldsymbol{\theta}, \boldsymbol{\phi}; \mathbf{x}^{(i)}) = -D_{KL}(q_{\boldsymbol{\phi}}(\mathbf{z}|\mathbf{x}^{(i)})||p_{\boldsymbol{\theta}}(\mathbf{z})) + \mathbb{E}_{q_{\boldsymbol{\phi}}(\mathbf{z}|\mathbf{x}^{(i)})}\left[\log p_{\boldsymbol{\theta}}(\mathbf{x}^{(i)}|\mathbf{z})\right] \tag{3}$$

Methods – Reparameterization trick

$$\mathbb{E}_{q_{\phi}(\mathbf{z}|\mathbf{x}^{(i)})}\left[\log p_{\boldsymbol{\theta}}(\mathbf{x}^{(i)}|\mathbf{z})\right]$$

- In order to solve our problem we invoked an alternative method for generating samples from qφ(z|x)
- z ~ qφ(z|x) be some conditional distribution. It is then often possible to express the random variable z as a deterministic variable z = gφ(ϵ, x) (z=μ+σ∗ϵ), where ϵ is an auxiliary variable with independent marginal p(ϵ), and gφ(.) is some vector-valued function parameterized by φ.



$$\int q_{\phi}(\mathbf{z}|\mathbf{x})f(\mathbf{z})\,d\mathbf{z} = \int p(\boldsymbol{\epsilon})f(\mathbf{z})\,d\boldsymbol{\epsilon} = \int p(\boldsymbol{\epsilon})f(g_{\phi}(\boldsymbol{\epsilon},\mathbf{x}))\,d\boldsymbol{\epsilon}.$$

Methods – Reparameterization trick

$$\int q_{\phi}(\mathbf{z}|\mathbf{x}) f(\mathbf{z})\, d\mathbf{z} = \int p(\boldsymbol{\epsilon}) f(\mathbf{z})\, d\boldsymbol{\epsilon} = \int p(\boldsymbol{\epsilon}) f(g_{\phi}(\boldsymbol{\epsilon}, \mathbf{x}))\, d\boldsymbol{\epsilon}.$$

$$\mathbb{E}_{q_{\phi}(\mathbf{z}|\mathbf{x}^{(i)})}\left[\log p_{\boldsymbol{\theta}}(\mathbf{x}^{(i)}|\mathbf{z})\right] \quad \mathbb{E}_{\mathcal{N}(z;\mu,\sigma^2)}[f(z)] = \mathbb{E}_{\mathcal{N}(\epsilon;0,1)}[f(\mu+\sigma\epsilon)] \simeq \frac{1}{L}\sum_{l=1}^{L} f(\mu+\sigma\epsilon^{(l)}) \text{ where } \epsilon^{(l)} \sim \mathcal{N}(0,1).$$

$$\nabla_{\phi}\mathbb{E}_{q_{\phi}(\mathbf{z})}[f(\mathbf{z})]$$

z = gφ(ϵ, x)

$$\begin{aligned}
\nabla_{\phi} L(\theta,\phi) &= \nabla_{\phi}\mathbf{E}_{x\sim p_{\phi}(x)}[f_{\theta}(x)] \\
&= \nabla_{\phi}\mathbf{E}_{\epsilon\sim p(\epsilon)}[f_{\theta}(g(\phi,\epsilon)] \\
&= \mathbf{E}_{\epsilon\sim p(\epsilon)}[\nabla_{\phi} f_{\theta}(g(\phi,\epsilon)] \\
&= \mathbf{E}_{\epsilon\sim p(\epsilon)}[f'_{\theta}(g(\phi,\epsilon))\nabla_{\phi} g(\phi,\epsilon)]
\end{aligned}$$

- Reparameterization trick을 이용하면 원래의 함수를 왼쪽과 같이 재표현이 가능해짐.
- Z가 Random variable일 때는 미분이 불가 했는데, deterministic한 변수로 재표현이 되어 미분이 가능해짐.(parameter φ 에 대하여)

Methods – Reparameterization trick

1. Tractable inverse CDF. In this case, let $\boldsymbol{\epsilon} \sim \mathcal{U}(\mathbf{0}, \mathbf{I})$, and let $g_{\boldsymbol{\phi}}(\boldsymbol{\epsilon}, \mathbf{x})$ be the inverse CDF of $q_{\boldsymbol{\phi}}(\mathbf{z}|\mathbf{x})$. Examples: Exponential, Cauchy, Logistic, Rayleigh, Pareto, Weibull, Reciprocal, Gompertz, Gumbel and Erlang distributions.
2. Analogous to the Gaussian example, for any "location-scale" family of distributions we can choose the standard distribution (with location $= 0$, scale $= 1$) as the auxiliary variable $\boldsymbol{\epsilon}$, and let $g(.) = \text{location} + \text{scale} \cdot \boldsymbol{\epsilon}$. Examples: Laplace, Elliptical, Student's t, Logistic, Uniform, Triangular and Gaussian distributions.
3. Composition: It is often possible to express random variables as different transformations of auxiliary variables. Examples: Log-Normal (exponentiation of normally distributed variable), Gamma (a sum over exponentially distributed variables), Dirichlet (weighted sum of Gamma variates), Beta, Chi-Squared, and F distributions.

Methods – SGVB estimator / AEVB Algorithm

- **SGVB Estimator A**

$$\mathcal{L}(\boldsymbol{\theta}, \boldsymbol{\phi}; \mathbf{x}^{(i)}) = \mathbb{E}_{q_{\boldsymbol{\phi}}(\mathbf{z}|\mathbf{x})} \left[-\log q_{\boldsymbol{\phi}}(\mathbf{z}|\mathbf{x}) + \log p_{\boldsymbol{\theta}}(\mathbf{x}, \mathbf{z}) \right]$$

$$\widetilde{L^A}(\theta, \phi; x^i) = \frac{1}{L} \sum_l logp_\theta(x^i, z^{i,l}) - logq_\phi(z^{i,l}|x^i)$$

$$where, g_\phi(\epsilon, x^i), \epsilon^l \sim p(\epsilon)$$

  - $logp_\theta(x^i, z^{i,l})$ is defined by probabilistic graphical model
  - $logq_\phi(z^{i,l}|x^i)$ can be determined $g_\phi(\epsilon, x^i), \epsilon^l \sim p(\epsilon)$

- **SGVB Estimator B**

$$ELBO_i(\lambda) = E_{q_\lambda(z|x_i)}[\log p(x_i|z)] - KL(q_\lambda(z|x_i)||p(z))$$

$$\widetilde{L^B}(\theta, \phi; x^i) = \frac{1}{L} \sum_l logp_\theta(x^i|z^{i,l}) - D_{KL}(q_\phi(z|x^i)||p_\theta(z))$$

$$where, g_\phi(\epsilon, x^i), \epsilon^l \sim p(\epsilon)$$

  - $D_{KL}(q_\phi(z|x^i)||p_\theta(z))$ is analytically evaluated (Gaussian Dist.)
  - $D_{KL}(q_\phi(z|x^i)||p_\theta(z))$ doesn't require sample z, only require parameter of approximate dist.

Methods – SGVB estimator / AEVB Algorithm

Algorithm 1 Minibatch version of the Auto-Encoding VB (AEVB) algorithm. Either of the two SGVB estimators in section 2.3 can be used. We use settings $M = 100$ and $L = 1$ in experiments.

$\boldsymbol{\theta}, \boldsymbol{\phi} \leftarrow$ Initialize parameters
repeat
  $\mathbf{X}^M \leftarrow$ Random minibatch of M datapoints (drawn from full dataset)
  $\boldsymbol{\epsilon} \leftarrow$ Random samples from noise distribution $p(\boldsymbol{\epsilon})$
  $\mathbf{g} \leftarrow \nabla_{\boldsymbol{\theta},\boldsymbol{\phi}} \widetilde{\mathcal{L}}^M(\boldsymbol{\theta}, \boldsymbol{\phi}; \mathbf{X}^M, \boldsymbol{\epsilon})$ (Gradients of minibatch estimator (8))
  $\boldsymbol{\theta}, \boldsymbol{\phi} \leftarrow$ Update parameters using gradients $\mathbf{g}$ (e.g. SGD or Adagrad [DHS10])
until convergence of parameters $(\boldsymbol{\theta}, \boldsymbol{\phi})$
return $\boldsymbol{\theta}, \boldsymbol{\phi}$

$$\mathcal{L}(\boldsymbol{\theta}, \phi; \mathbf{X}) \simeq \widetilde{\mathcal{L}}^M(\boldsymbol{\theta}, \phi; \mathbf{X}^M) = \frac{N}{M} \sum_{i=1}^{M} \widetilde{\mathcal{L}}(\boldsymbol{\theta}, \phi; \mathbf{x}^{(i)})$$