VI & BNN (1)

Stochastic Variational Inference (SVI) and Variational Autoencoder (VAE)

Keywords : Variational Inference, Scalable Variational Inference, ELBO, Probabilistic Deep Learning, Variational Autoencoder

Seunghan Lee

February 01, 2021

Contents

Paper List (+ references)

- Practical variational inference for neural networks (2011)
- Stochastic variational inference (2013)
- Auto-Encoding Variational Bayes (2013)
- Variational Inference : A review for statisticians (2017)
- Advances in variational inference (2018)
- Deep Bayes (https://deepbayes.ru/)

1. Variational Inference

1. Variational Inference

1-1. MCMC vs Variational Inference

Two main approaches to find the **(intractable) posterior** in Bayesian Inference!

(1) MCMC : sampling from the unnormalized posterior

- (Pros) Unbiased
- (Cons) High computational cost

(2) Variational Inference : Approximating target distn with a simpler distn

- (Pros) Faster & Scalable
- (Cons) Biased

1. Variational Inference

1-1. MCMC vs Variational Inference

Will be going to focus on **Variational Inference**

Before getting on…

- **KL divergence**

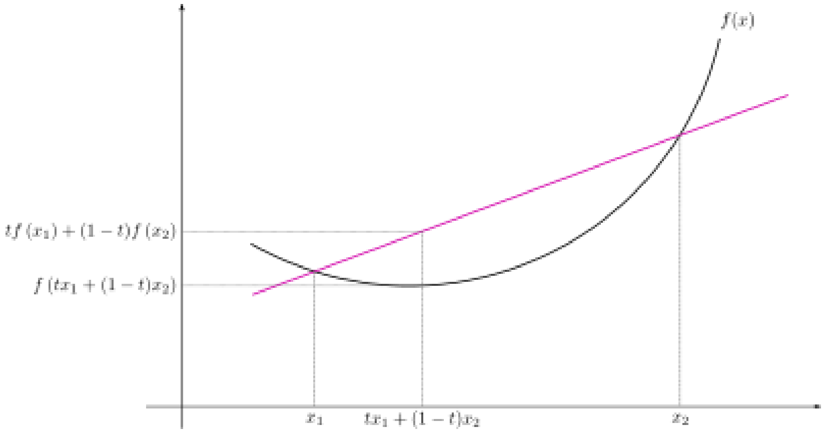
$$KL(q(\theta)\| p(\theta \mid x)) = \int q(\theta) \log \frac{q(\theta)}{p(\theta \mid x)} d\theta$$

- $KL(q\| p) \geq 0$
- $KL(q\| p) = 0 \Leftrightarrow q = p$
- $KL(q\| p) \neq KL(p\| q)$

- **Jensen's Inequality**

If $g(x)$ is a convex function on R_X and $E[g(X)]$ and $g(E[X])$ are finite,

$E[g(X)] \geq g(E[X])$



1. Variational Inference

1-2. Evidence Lower Bound (ELBO)

$$\log p(x) = \int q(\theta) \log p(x) d\theta = \int q(\theta) \log \frac{p(x,\theta)}{p(\theta \mid x)} d\theta =$$

$$= \int q(\theta) \log \frac{p(x,\theta) q(\theta)}{p(\theta \mid x) q(\theta)} d\theta =$$

$$= \int q(\theta) \log \frac{p(x,\theta)}{q(\theta)} d\theta + \int q(\theta) \log \frac{q(\theta)}{p(\theta \mid x)} d\theta =$$

$$= \mathcal{L}(q(\theta)) + KL(q(\theta) \| p(\theta \mid x))$$ (Non-negative)

1. Variational Inference

1-2. Evidence Lower Bound (ELBO)

$$\log p(x) = \int q(\theta) \log p(x) d\theta = \int q(\theta) \log \frac{p(x,\theta)}{p(\theta \mid x)} d\theta =$$

$$= \int q(\theta) \log \frac{p(x,\theta) q(\theta)}{p(\theta \mid x) q(\theta)} d\theta =$$

$$= \int q(\theta) \log \frac{p(x,\theta)}{q(\theta)} d\theta + \int q(\theta) \log \frac{q(\theta)}{p(\theta \mid x)} d\theta =$$

$$= \mathcal{L}(q(\theta)) + KL(q(\theta) \| p(\theta \mid x))$$ (Non-negative)

$$\begin{aligned} \log p(x) &= \log \int_z p(x,z) \\ &= \log \int_z p(x,z) \frac{q(z)}{q(z)} \\ &= \log \left(\mathbb{E}_q \left[\frac{p(x,z)}{q(z)} \right] \right) \\ &\geq \mathbb{E}_q [\log p(x,z)] - \mathbb{E}_q [\log q(z)] \end{aligned}$$

Jensen's Inequality

ELBO (Evidence Lower Bound)

Why called like that? Think of "Evidence" in Bayes rule!

$$\log p(x) \geq \mathcal{L}(q(\theta))$$

1-2. Evidence Lower Bound (ELBO)

$$\log p(x) = \int q(\theta)\log p(x)d\theta = \int q(\theta)\log\frac{p(x,\theta)}{p(\theta \mid x)}d\theta =$$

$$= \int q(\theta)\log\frac{p(x,\theta)q(\theta)}{p(\theta \mid x)q(\theta)}d\theta =$$

$$= \int q(\theta)\log\frac{p(x,\theta)}{q(\theta)}d\theta + \int q(\theta)\log\frac{q(\theta)}{p(\theta \mid x)}d\theta =$$

$$= \mathcal{L}(q(\theta)) + KL(q(\theta) \| p(\theta \mid x))$$

(Non-negative)

Minimizing KL-divergence

ELBO (Evidence Lower Bound)

Why called like that? Think of "Evidence" in Bayes rule!

$$\log p(x) \geq \mathcal{L}(q(\theta))$$

1. Variational Inference

1-2. Evidence Lower Bound (ELBO)

$$\log p(x) = \int q(\theta) \log p(x) d\theta = \int q(\theta) \log \frac{p(x,\theta)}{p(\theta \mid x)} d\theta =$$

$$= \int q(\theta) \log \frac{p(x,\theta) q(\theta)}{p(\theta \mid x) q(\theta)} d\theta =$$

$$= \int q(\theta) \log \frac{p(x,\theta)}{q(\theta)} d\theta + \int q(\theta) \log \frac{q(\theta)}{p(\theta \mid x)} d\theta =$$

$$= \mathcal{L}(q(\theta)) + KL(q(\theta) \| p(\theta \mid x))$$ (Non-negative)

Maximizing ELBO

ELBO (Evidence Lower Bound)

Why called like that? Think of “Evidence” in Bayes rule!

$$\log p(x) \geq \mathcal{L}(q(\theta))$$

1-2. Evidence Lower Bound (ELBO)

Interpretation of ELBO

Rewrite ELBO as below

$$\mathcal{L}(q(\theta)) = \int q(\theta) \log \frac{p(x,\theta)}{q(\theta)} d\theta = \int q(\theta) \log \frac{p(x \mid \theta) p(\theta)}{q(\theta)} d\theta =$$

$$= \int q(\theta) \log p(x \mid \theta) d\theta + \int q(\theta) \log \frac{p(\theta)}{q(\theta)} d\theta =$$

$$= \mathbb{E}_{q(\theta)} \log p(x \mid \theta) - KL(q(\theta) \| p(\theta))$$

1-2. Evidence Lower Bound (ELBO)

Interpretation of ELBO

Rewrite ELBO as below

$$\mathcal{L}(q(\theta)) = \int q(\theta) \log \frac{p(x,\theta)}{q(\theta)} d\theta = \int q(\theta) \log \frac{p(x \mid \theta) p(\theta)}{q(\theta)} d\theta =$$

$$= \int q(\theta) \log p(x \mid \theta) d\theta + \int q(\theta) \log \frac{p(\theta)}{q(\theta)} d\theta =$$

$$= \underline{\mathbb{E}_{q(\theta)} \log p(x \mid \theta)} - \underline{KL(q(\theta) \| p(\theta))}$$

Term 1) Encourage good fit!

Term 2) Regularize! Encourage posterior to be close to prior

1. Variational Inference

1-3. Mean Field Variational Inference (MFVI)

We have to optimize w.r.t ELBO, but how?

$$\mathcal{L}(q(\theta)) = \int q(\theta) \log \frac{p(x,\theta)}{q(\theta)} d\theta \rightarrow \max_{q(\theta)\in\mathcal{Q}}$$

Mean Field Assumption

(For simplicity, variational parameters are factorized as below, with an Independent Assumption)

$$q(\theta) = \prod_{j=1}^{m} q_j(\theta_j), \quad \theta = [\theta_1, \ldots, \theta_m]$$

Due to the assumption, its flexibility is limited!

1. Variational Inference

1-3. Mean Field Variational Inference (MFVI)

$$\mathcal{L}(q(\theta)) = \int q(\theta) \log \frac{p(x,\theta)}{q(\theta)} d\theta \rightarrow \max_{q(\theta) \in \mathcal{Q}}$$

$$\mathcal{L}(q(\theta)) = \int q(\theta) \log \frac{p(x,\theta)}{q(\theta)} d\theta \rightarrow \max_{q(\theta) = q_1(\theta_1) \cdot \ldots \cdot q_m(\theta_m)}$$

1-3. Mean Field Variational Inference (MFVI)

$$\mathcal{L}(q(\theta)) = \int q(\theta) \log \frac{p(x,\theta)}{q(\theta)} d\theta \rightarrow \max_{q(\theta) \in \mathcal{Q}}$$

$$\mathcal{L}(q(\theta)) = \int q(\theta) \log \frac{p(x,\theta)}{q(\theta)} d\theta \rightarrow \max_{q(\theta) = q_1(\theta_1) \cdot \ldots \cdot q_m(\theta_m)}$$

Solution (proof on the next page)

$$q_j(\theta_j) = r_j(\theta_j) = \frac{1}{Z_j} \exp\left(\mathbb{E}_{q_{i \neq j}} \log p(x,\theta)\right)$$

Coordinate-ascent method

CAVI (Coordinate Ascent Variational Inference)

1. Variational Inference

1-3. Mean Field Variational Inference (MFVI)

$$q_j\left(\theta_j\right) = r_j(\theta_j) = \frac{1}{Z_j}\exp\left(\mathbb{E}_{q_{i\neq j}} \log p(x,\theta)\right)$$

Proof)

$$\mathcal{L}(q(\theta)) = \mathbb{E}_{q(\theta)} \log p(x,\theta) - \mathbb{E}_{q(\theta)} \log q(\theta) =$$

$$= \mathbb{E}_{q(\theta)} \log p(x,\theta) - \sum_{k=1}^{m} \mathbb{E}_{q_k(\theta_k)} \log q_k(\theta_k) =$$

$$= \mathbb{E}_{q_j(\theta_j)} \left[\mathbb{E}_{q_{i\neq j}} \log p(x,\theta)\right] - \mathbb{E}_{q_j(\theta_j)} \log q_j(\theta_j) + Const =$$

$$= \left\{ r_j(\theta_j) = \frac{1}{Z_j}\exp\left(\mathbb{E}_{q_{i\neq j}} \log p(x,\theta)\right)\right\} =$$

$$= \mathbb{E}_{q_j(\theta_j)} \log \frac{r_j(\theta_j)}{q_j(\theta_j)} + Const = -KL\left(q_j(\theta_j) \| \, r_j(\theta_j)\right) + Const$$

1. Variational Inference

1-3. Mean Field Variational Inference (MFVI)

Algorithm

Initialize $q(\theta) = \prod_{j=1}^{m} q_j(\theta_j)$

Iterations:

- Update each factor $q_1, \ldots, q_m$:

$$q_j(\theta_j) = \frac{1}{Z_j} \exp\left(\mathbb{E}_{q_{i \neq j}} \log p(x, \theta)\right)$$

- Compute ELBO $\mathcal{L}(q(\theta))$

Repeat until convergence of ELBO

2. Stochastic Variational Inference

2. Stochastic Variational Inference

Stochastic Variational Inference (SVI)?

a stochastic optimization algorithm for mean-field variational inference,

that can **handle massive dataset (scalability)**

- Mean Field Variational Inference
- Stochastic Optimization

Journal of Machine Learning Research 14 (2013) 1303-1347 Submitted 6/12; Published 5/13

Stochastic Variational Inference

Matthew D. Hoffman MATHOFFM@ADOBE.COM
Adobe Research
Adobe Systems Incorporated
601 Townsend Street
San Francisco, CA 94103, USA

David M. Blei BLEI@CS.PRINCETON.EDU
Department of Computer Science
Princeton University
35 Olden Street
Princeton, NJ 08540, USA

Chong Wang CHONGW@CS.CMU.EDU
Machine Learning Department
Carnegie Mellon University
Gates Hillman Centers, 8110
5000 Forbes Avenue
Pittsburgh, PA 15213, USA

John Paisley JPAISLEY@BERKELEY.EDU
Computer Science Division
University of California
Berkeley, CA 94720-1776, USA

Editor: Tommi Jaakkola

Abstract

We develop stochastic variational inference, a scalable algorithm for approximating posterior distributions. We develop this technique for a large class of probabilistic models and we demonstrate

2. Stochastic Variational Inference

Mean Field Variational Inference

(This time, we will set hidden variables into 2 parts)

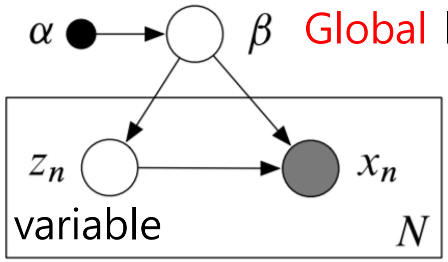


Figure 2: A graphical model with observations $x_{1:N}$, local hidden variables $z_{1:N}$ and global hidden variables β. The distribution of each observation x_n only depends on its corresponding local variable z_n and the global variables β. (Though not pictured, each hidden variable z_n, observation x_n, and global variable β may be a collection of multiple random variables.)

Factorize joint distribution

$$p(x, z, \beta \mid \alpha) = p(\beta \mid \alpha) \prod_{n=1}^{N} p(x_n, z_n \mid \beta)$$

Approximate using MFVI

$$p(z, \beta \mid x) = \frac{p(x, z, \beta)}{\int p(x, z, \beta) dz d\beta}$$

←

Approximating distn :

$$q(z, \beta) = q(\beta \mid \lambda) \prod_{n=1}^{N} \prod_{j=1}^{J} q(z_{nj} \mid \phi_{nj})$$

2. Stochastic Variational Inference

Mean Field Variational Inference

$$q(z, \beta) = q(\beta \mid \lambda) \prod_{n=1}^{N} \prod_{j=1}^{J} q(z_{nj} \mid \phi_{nj})$$

Set $q(\beta \mid \lambda)$ and $q\left(z_{nj} \mid \phi_{nj}\right)$ to be in the same exponential family

- $q(\beta \mid \lambda) = h(\beta) \exp\{\lambda^{\top} t(\beta) - a_g(\lambda)\}$
- $q\left(z_{nj} \mid \phi_{nj}\right) = h\left(z_{nj}\right) \exp\left\{\phi_{nj}^{\top} t\left(z_{nj}\right) - a_{\ell}\left(\phi_{nj}\right)\right\}$

$$\nabla_{\lambda} \mathcal{L} = \nabla_{\lambda}^{2} a_g(\lambda)(\mathbb{E}_q[\eta_g(x, z, \alpha)] - \lambda) \Rightarrow \boxed{\lambda = \mathbb{E}_q[\eta_g(x, z, \alpha)]}$$

$$\nabla_{\phi_{nj}} \mathcal{L} = \nabla_{\phi_{nj}}^{2} a_{\ell}(\phi_{nj})(\mathbb{E}_q[\eta_{\ell}(x_n, z_{n,-j}, \beta)] - \phi_{nj}) \Rightarrow \boxed{\phi_{nj} = \mathbb{E}_q[\eta_{\ell}(x_n, z_{n,-j}, \beta)]}$$

2. Stochastic Variational Inference

Initialize $\lambda^{(0)}$ randomly.
repeat
  for each local variational parameter ϕ_{nj} **do**
    Update ϕ_{nj}, $\phi_{nj}^{(t)} = \mathbb{E}_{q^{(t-1)}}[\eta_{\ell,j}(x_n, z_{n,-j}, \beta)]$.
  end for
  Update the global variational parameters, $\lambda^{(t)} = \mathbb{E}_{q^{(t)}}[\eta_g(z_{1:N}, x_{1:N})]$.
until the ELBO converges

Figure 3: Coordinate ascent mean-field variational inference.

1. Variational Inference

1-3. Mean Field Variational Inference (MFVI)

Algorithm

Initialize $q(\theta) = \prod_{j=1}^{m} q_j(\theta_j)$
Iterations:
- Update each factor $q_1, \ldots, q_m$:

$$q_j(\theta_j) = \frac{1}{Z_j} \exp\left(\mathbb{E}_{q_{i \neq j}} \log p(x, \theta)\right)$$

- Compute ELBO $\mathcal{L}(q(\theta))$

Repeat until convergence of ELBO

2. Stochastic Variational Inference

Initialize $\lambda^{(0)}$ randomly.
repeat
  for each local variational parameter ϕ_{nj} **do**
    Update ϕ_{nj}, $\phi_{nj}^{(t)} = \mathbb{E}_{q^{(t-1)}}[\eta_{\ell,j}(x_n, z_{n,-j}, \beta)]$.
  end for
  Update the global variational parameters, $\lambda^{(t)} = \mathbb{E}_{q^{(t)}}[\eta_g(z_{1:N}, x_{1:N})]$.
until the ELBO converges

Each data has its OWN local variational parameter

Figure 3: Coordinate ascent mean-field variational inference.

1. Variational Inference

1-3. Mean Field Variational Inference (MFVI)

Algorithm

Initialize $q(\theta) = \prod_{j=1}^{m} q_j(\theta_j)$
Iterations:
- Update each factor $q_1, \ldots, q_m$:

$$q_j(\theta_j) = \frac{1}{Z_j} \exp\left(\mathbb{E}_{q_{i \neq j}} \log p(x, \theta)\right)$$

- Compute ELBO $\mathcal{L}(q(\theta))$

Repeat until convergence of ELBO

INEFFICIENT for large data sets!

(should optimize the local variational params for each data, before re-estimating the global variational params)

SVI uses "**stochastic optimization**" to fit global variational parameters.

2. Stochastic Variational Inference

Algorithm

1: Initialize $\lambda^{(0)}$ randomly.
2: Set the step-size schedule ρ_t appropriately.
3: **repeat**
4: Sample a data point x_i uniformly from the data set.
5: Compute its local variational parameter,

$$\phi = \mathbb{E}_{\lambda^{(t-1)}}[\eta_g(x_i^{(N)}, z_i^{(N)})].$$

6: Compute intermediate global parameters as though x_i is replicated N times,

$$\hat{\lambda} = \mathbb{E}_{\phi}[\eta_g(x_i^{(N)}, z_i^{(N)})].$$

7: Update the current estimate of the global variational parameters,

$$\lambda^{(t)} = (1-\rho_t)\lambda^{(t-1)} + \rho_t\hat{\lambda}.$$

8: **until** forever

Figure 4: Stochastic variational inference.

Simple! Just think it as

(1) SGD + (2) VI

$$\begin{aligned} \lambda^{(t)} &= \lambda^{(t-1)} + \rho_t\left(\hat{\lambda}_t - \lambda^{(t-1)}\right) \\ &= (1-\rho_t)\lambda^{(t-1)} + \rho_t\hat{\lambda}_t. \end{aligned}$$

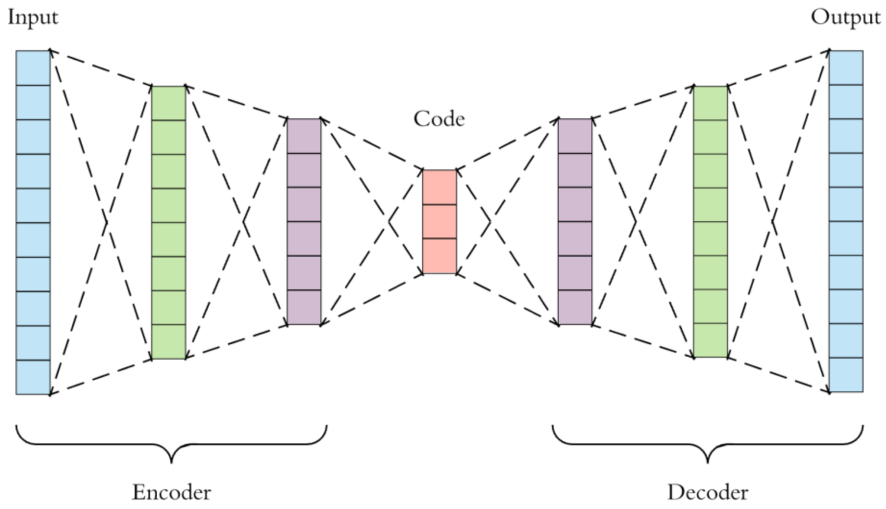
3. Variational Auto Encoder (VAE)

3. Variational Auto Encoder

3-1. Structure of VAE

Auto Encoder : "The aim of an **autoencoder** is to learn a **representation (encoding) for a set of data** "

What's the difference between AE & VAE?



Auto Encoder (AE)

3. Variational Auto Encoder

3-1. Structure of VAE

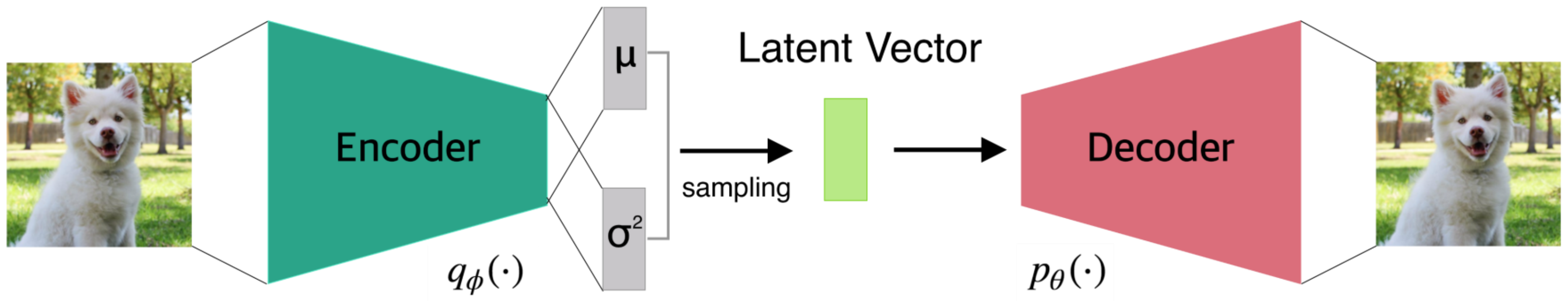
(1) Encoder

- inference network
- input : x, output : z
- $q(z \mid x, \phi)$

(2) Decoder

- generative network
- input : z, output : as closely as x
- $p(x \mid z, \theta)$

Variational Autoencoder



3. Variational Auto Encoder

3-1. Structure of VAE

Variational Inference : approximate with variational distribution!

$$p(z_i \mid x_i, \theta) \approx q(z_i \mid x_i, \phi) = \prod_{i=1}^{d} N(z_{ij} \mid \underline{\mu_j(x_i)}, \underline{\sigma_j^2(x_i)})$$

(use Neural Network! Flexible!)

Objective Function (ELBO)

- Minimize KL-divergence $q(Z \mid X, \phi) = \arg\min_{\phi} KL(q(Z \mid X, \phi) \| p(Z \mid X, \theta))$

= Maximize ELBO $\mathcal{L}(\phi, \theta) = \int q(Z \mid X, \phi) \log \frac{p(X|Z,\theta)p(Z)}{q(Z|X,\phi)} dZ \rightarrow \max_{\phi,\theta}$

3. Variational Auto Encoder

3-1. Structure of VAE

Update the parameters of **Encoder & Decoder**.

Due to flexible & complex model (Neural Network), it seems hard to solve…

But using some techniques (+ tricks), we can solve it!

Stochastic Optimization

1) Mini-batch
2) Monte Carlo Estimation

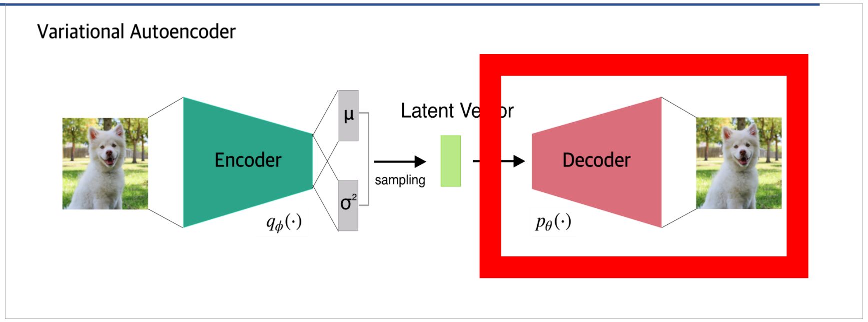
Tricks

1) Log-derivative trick

2) Reparameterization trick

3. Variational Auto Encoder

3-2. Update Decoder (parameter : θ)



Not that hard to solve!

(1) Mini-batch

$$\nabla_\theta \mathcal{L}(\phi,\theta) = \nabla_\theta \sum_{i=1}^{n} \int q(z_i \mid x_i, \phi) \log \frac{p(x_i \mid z_i, \theta)\, p(z_i)}{q(z_i \mid x_i, \phi)} dz_i$$

$$= \sum_{i=1}^{n} \int q(z_i \mid x_i, \phi)\, \nabla_\theta \log p(x_i \mid z_i, \theta)\, dz_i$$

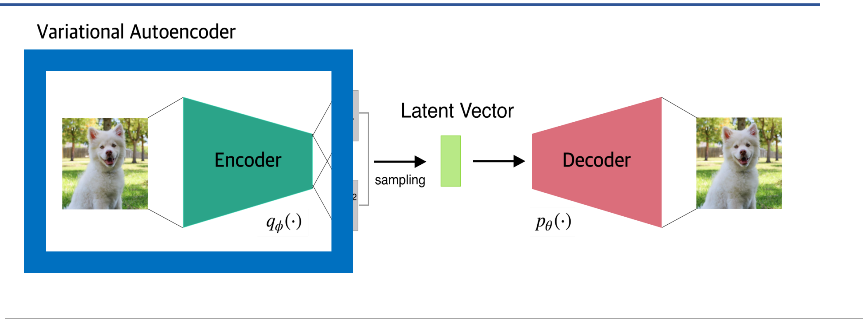
$$\approx n \int q(z_i \mid x_i, \phi)\, \nabla_\theta \log p(x_i \mid z_i, \theta)\, dz_i, \quad i \sim \mathcal{U}\{1, \ldots, n\}$$

(2) Monte-Carlo estimation

$$n \int q(z_i \mid x_i, \phi)\, \nabla_\theta \log p(x_i \mid z_i, \theta)\, dz_i \approx n \nabla_\theta \log p(x_i \mid z_i^*, \theta), \quad z_i^* \sim q(z_i \mid x_i, \phi)$$

3. Variational Auto Encoder

3-3. Update Encoder (parameter : ϕ)



A bit trickier than updating Encoder… ☹

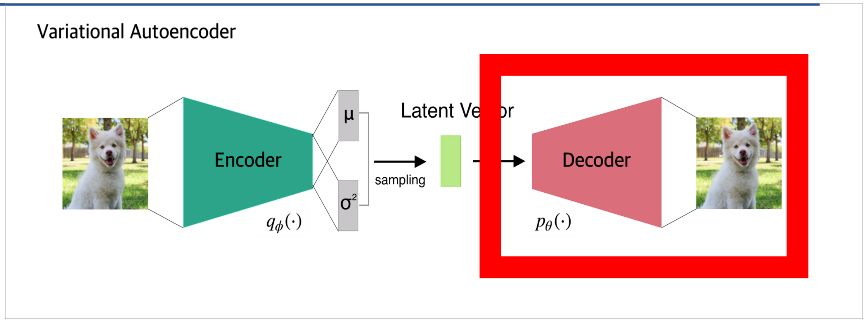
$$\nabla_\phi \mathcal{L}(\phi,\theta) = \nabla_\phi \sum_{i=1}^{n} \int q(z_i \mid x_i, \phi) \log \frac{p(x_i \mid z_i, \theta)\, p(z_i)}{q(z_i \mid x_i, \phi)} dz_i$$

$$\neq \sum_{i=1}^{n} \int q(z_i \mid x_i, \phi) \nabla_\phi \log p(x_i \mid z_i, \theta)\, dz_i$$

We need some tricks to solve this!

3. Variational Auto Encoder

3-3. Update Encoder (parameter : ϕ)



a. Log Derivate Trick

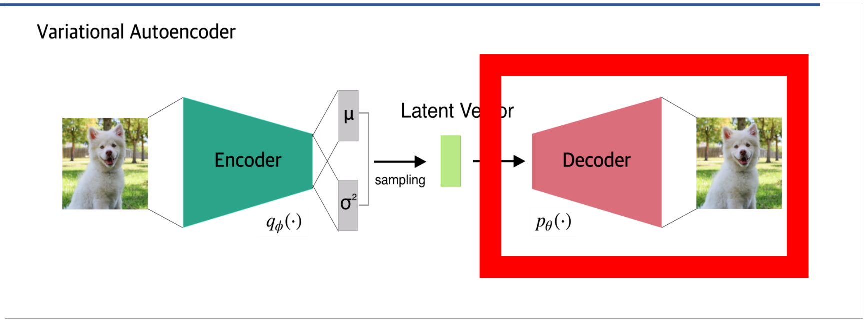
$$\frac{\partial}{\partial x} p(y \mid x) = p(y \mid x) \frac{\partial}{\partial x} \log p(y \mid x)$$

For question like solving $E_{y|x} h(x, y)$

$$\begin{aligned} \frac{\partial}{\partial x} \int p(y \mid x) h(x, y) dy &= \int \frac{\partial}{\partial x} (p(y \mid x) h(x, y)) dy \\ &= \int \left(h(x, y) \frac{\partial}{\partial x} p(y \mid x) + p(y \mid x) \frac{\partial}{\partial x} h(x, y) \right) dy \\ &= \int p(y \mid x) \frac{\partial}{\partial x} h(x, y) dy + \int h(x, y) \frac{\partial}{\partial x} p(y \mid x) dy \end{aligned}$$

3. Variational Auto Encoder

3-3. Update Encoder (parameter : ϕ)



a. Log Derivate Trick

$$\frac{\partial}{\partial x}\int p(y \mid x)h(x,y)dy = \int \frac{\partial}{\partial x}(p(y \mid x)h(x,y))dy$$
$$= \int \left(h(x,y)\frac{\partial}{\partial x}p(y \mid x) + p(y \mid x)\frac{\partial}{\partial x}h(x,y)\right)dy$$
$$= \int p(y \mid x)\frac{\partial}{\partial x}h(x,y)dy + \int h(x,y)\frac{\partial}{\partial x}p(y \mid x)dy$$

First term

- $\int p(y \mid x)\frac{\partial}{\partial x}h(x,y)dy$ - easy to solve (using Monte Carlo estimation)

Second term

- $\int h(x,y)\frac{\partial}{\partial x}p(y \mid x)dy$ - hard to solve - but can be solved using "Log-derivative trick"

3. Variational Auto Encoder

3-3. Update Encoder (parameter : ϕ)

a. Log Derivate Trick

Apply it to our derivative of ELBO!

$$\frac{\partial}{\partial \phi}\mathcal{L}(\phi, \theta) = \frac{\partial}{\partial \phi}\int q(Z \mid X, \phi) \log p(X \mid Z, \theta) dZ$$

$$\approx n \log p\left(x_i \mid z_i^*, \theta\right) \frac{\partial}{\partial \phi} \log q\left(z_i^* \mid x_i, \phi\right)$$

1. **Mini batching**
2. **Log-derivative Trick**

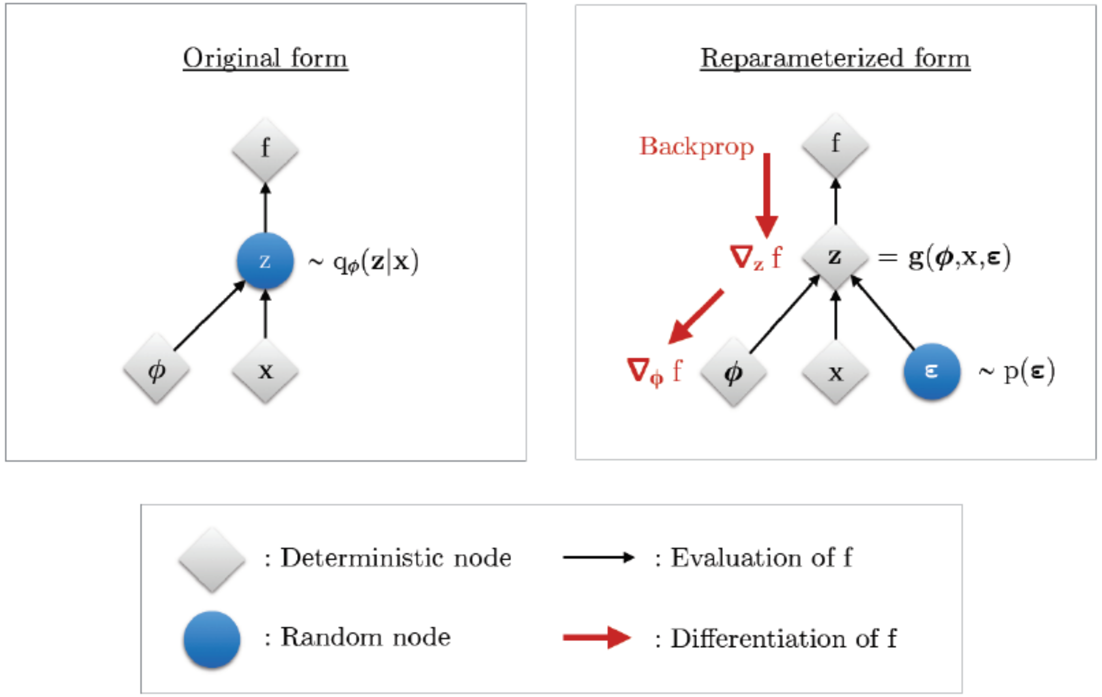
3. Variational Auto Encoder

3-3. Update Encoder (parameter : ϕ)

b. Reparameterization Trick

We can not backpropagate through "random" variable!

It should be deterministic!



$$\frac{\partial}{\partial x}\int p(y \mid x)h(x, y)dy = \frac{\partial}{\partial x}\int r(\epsilon)h(x, g(\epsilon, x))d\epsilon$$

$$\approx \frac{d}{dx}h(x, g(x, \hat{\epsilon}))$$

$$= \frac{\partial}{\partial x}h(x, g(x, \hat{\epsilon})) + \frac{\partial}{\partial g}h(x, g(x, \hat{\epsilon}))\frac{\partial}{\partial x}g(x, \hat{\epsilon}) \quad \text{where} \quad \hat{\epsilon} \sim r(\epsilon)$$

3. Variational Auto Encoder

3-3. Update Encoder (parameter : ϕ)

b. Reparameterization Trick

Apply it to our derivative of ELBO!

$$n\frac{\partial}{\partial\phi}\int q\left(z_i \mid x_i,\phi\right)\log p\left(x_i \mid z_i,\theta\right) = n\frac{\partial}{\partial\phi}\int r(\epsilon)\log p\left(x_i \mid g\left(\epsilon, x_i,\phi\right),\theta\right)d\epsilon$$

$$\approx n\frac{\partial}{\partial\phi}\log p\left(x_i \mid g\left(\hat{\epsilon}, x_i,\phi\right),\theta\right), \quad \text{where } \hat{\epsilon}\sim r(\epsilon)$$

(simple example)

$$q_\phi\left(\mathbf{z}_i|\mathbf{x}_i\right) = \mathrm{N}\left(\boldsymbol{\mu}_i, \boldsymbol{\sigma}_i^2\mathbf{I}\right)$$

➡ $\mathbf{z}_i = \boldsymbol{\mu}_i + \boldsymbol{\sigma}_i^2 \odot \boldsymbol{\epsilon}_i$ where $\boldsymbol{\epsilon}_i \sim \mathrm{N}\left(\mathbf{0},\mathbf{I}\right)$

Summary of VAE

- step 1) sample $i \sim \mathcal{U}\{1, \ldots, n\}$
- step 2) compute stochastic gradient of ELBO (w.r.t θ and ϕ)
  - Update θ (decoder parameter)

    - stoch. $\text{grad}_\theta\, \mathcal{L}(\phi, \theta) = n \frac{\partial}{\partial \theta} \log p\left(x_i \mid z_i^*, \theta\right)$

    where $z_i^* \sim q\left(z_i \mid x_i, \phi\right)$
  - Update ϕ (encoder parameter)

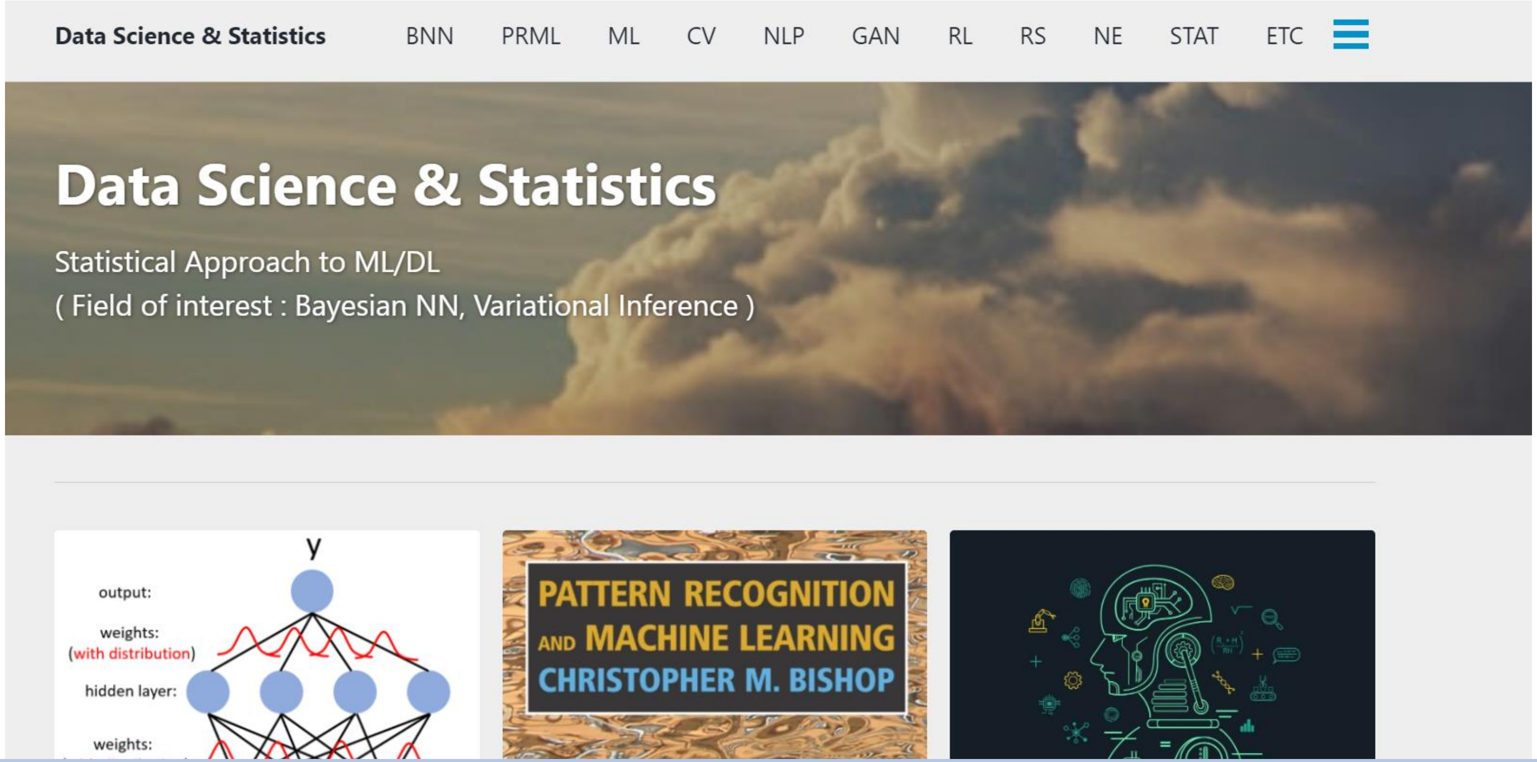
    - stoch.grad $_\phi\mathcal{L}(\phi, \theta) = n \frac{\partial}{\partial \phi} \log p\left(x_i \mid g\left(\hat{\epsilon}, x_i, \phi\right), \theta\right) - \frac{\partial}{\partial \phi} KL\left(q\left(z_i \mid x_i, \phi\right) \| p\left(z_i\right)\right)$

    where $\hat{\epsilon} \sim r(\epsilon)$
- Update until stopping criterion reaches

3. Variational Auto Encoder

3-4. Implementation using Pytorch

https://seunghan96.github.io/stat/gan/bnn/code-6.Variational-Auto-Encoder/

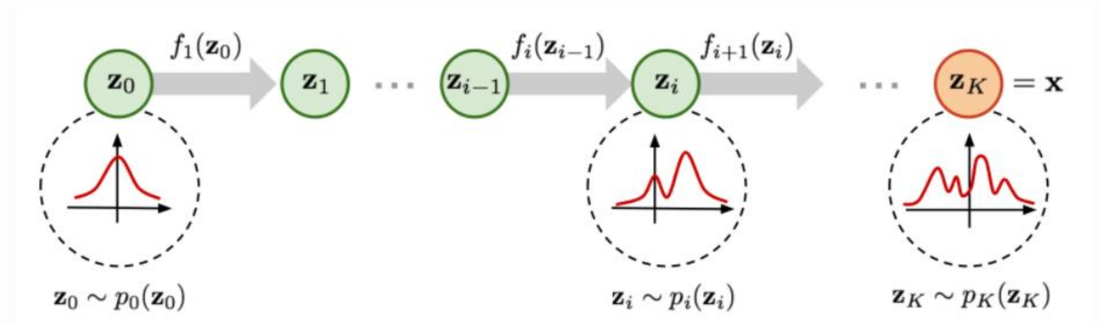


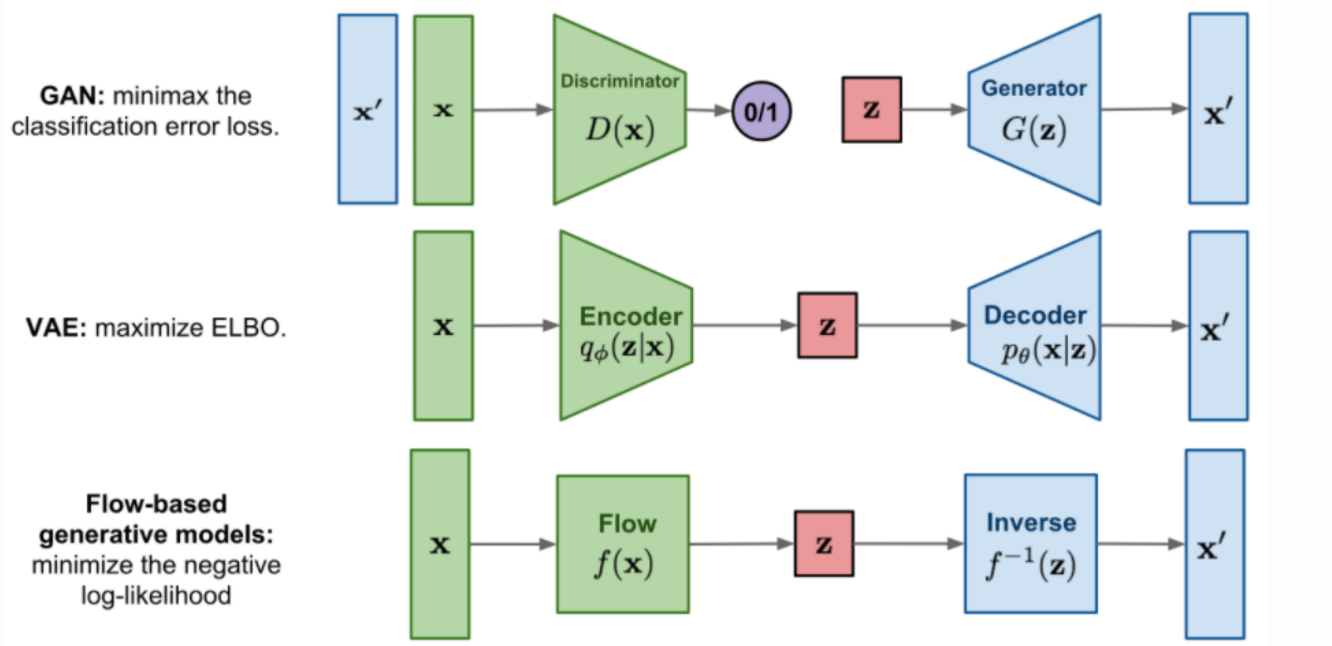
Summary

Have dealt with basic concepts to know before reading papers about

various **variational inference methods, Bayesian NN.**

Next Presentation : Normalizing Flow





Thank you !

Review of Papers regarding VI/BNN

(+ some Statistical Models / Machine Learning / Deep Learning)

can be found in my github blog below :)

https://seunghan96.github.io/

(for more about VI/BNN https://seunghan96.github.io/categories/bnn/)