Paper Review 6

Auto-Encoding Variational Bayes

0. Abstract

본 논문은 interactable posterior 분포를 따르는 연속적인 잠재 변수들이 존재하고 큰 데이터 셋이 있을 때, 방향성 확률 모델이 어떻게 효율적인 추론과 학습을 할 수 있을 지 의문을 제시한다. 연구진은 이에 대해 큰 데이터 셋을 다루고, 여러 미분 가능성이 까다로운 조건에서도 작동할 수 있는 확률적 변수 추론과 학습 알고리즘을 제안한다.

이 알고리즘은 두 가지 면에서 기여한다.

- variational lower bound의 reparameterization이 일반적인 확률적 기울기 방법론(standard stochastic gradient methods)을 사용하여 직접적으로 최적화 될 수 있는 lower bound estimator를 생성한다.
- 각 포인트가 연속형 잠재변수를 가지는 i.i.d 데이터셋에 대해 lower bound estimator를 사용해 approximated inference model을 계산이 불가능한 posterior에 fitting하여 posterior 추론이 효율적으로 가능하게 한다.

1. Introduction

우리는 어떻게 interactable posterior 분포와 연속적 잠재 변수들이 존재하는 방향성 확률 모델으로 효율적인 추론과 학습을 하도록 할까? VB(Variational Bayesian) 접근법은 다루기 힘든 posterior에 대한 근사치를 최적화하는 방법을 제시한다.

본 논문은 SGVB(Stochastic Gradient Variation Bayes) estiamator 를 활용하여 효율적인 approximate posterior inference를 수행한다. 또한, AEVB(Auto Encoding Variational Encoder)를 사용하는 알고리즘을 제안한다. AEVB 알고리즘에서 SGVB estimator로 recognition 모델을 최적화 할 수 있고, 이렇게 학습된 approximate posterior inference 모델은 recognition, denoising, representation, visualization에 활용될 수 있다. 또한, 신경망이 recognition 모델에 활용될 때 VAE(Variational Auto-Encoder)을 만들 수 있다.

2. Method

연속적 잠재변수를 가지는 다양한 모델에 대해 lower bound estimator를 도출하는 전략을 설명한다. 각 포인트가 연속형 잠재변수를 가지는 i.i.d 데이터셋에서 전역 변수/잠재 변수들

의 variational inference에 ML(Maximum likelihood)과 MAP(Maximum a posteriori) 추론을 수행한다.

2.1 Problem scenario

데이터 셋 $X = \{x_i\}_{i=1}^{N}$은 N개의 i.i.d 샘플로 구성된다. 관측되지 않은 연속적 랜덤 변수 z를 포함한 모든 데이터는 random process로 생성된다고 가정한다. 하지만 실제 parameter인 θ^*값과 잠재 변수인 $z^{(i)}$는 알려져 있지 않다.

이러한 문제를 해결하기 위해, 아래의 해결책을 제시한다.

1. θ의 efficient approximate ML / MAP 추정
2. θ의 선택에 따른 관측 값 x가 주어졌을 때, z의 efficient approximate posterior inference
3. 변수 x의 efficient approximate marginal inference

본 논문은 이를 위해 recognition model인 $q_\phi(z \mid x)$($p_\theta(z \mid x)$의 근사치)를 제안하며, recognition model의 매개변수 ϕ와 generative model의 매개변수 θ를 함께 학습하는 방법을 소개한다. 따라서 $q_\phi(z \mid x)$을 확률적 encoder를, $p_\theta(z \mid x)$은 확률적 decoder를 나타낸다.

2.2 The variational bound

Marginal likelihood는 다음과 같이 나타낼 수 있다.

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

우변의 첫 항은 true posterior 근사치의 KL divergence이고, 두 번째 항은 marginal likelihood의 lower bound이다. lower bound는 아래와 같이 다시 나타낼 수 있다.

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

이 lower bound를 미분하고 최적화해야한다.

2.3 The SGVB estimator and AEVB algorithm

lower bound의 practical estimator와 미분값을 얻기 위해, $q_\phi(z \mid x)$ 형태의 approximate posterior을 가정한다.

함수 f에 대해 Monte Carlo estimate를 하면 아래와 같이 나타낼 수 있다.

$$\mathbb{E}_{q_{\phi}(\mathbf{z}|\mathbf{x}^{(i)})}\left[f(\mathbf{z})\right] = \mathbb{E}_{p(\boldsymbol{\epsilon})}\left[f(g_{\boldsymbol{\phi}}(\boldsymbol{\epsilon}, \mathbf{x}^{(i)}))\right] \simeq \frac{1}{L}\sum_{l=1}^{L} f(g_{\boldsymbol{\phi}}(\boldsymbol{\epsilon}^{(l)}, \mathbf{x}^{(i)})) \quad \text{where} \quad \boldsymbol{\epsilon}^{(l)} \sim p(\boldsymbol{\epsilon})$$

이 estimator를 variational lower bound에 적용하면 SGVB estimator값을 얻을 수 있으며, 다음과 같다.

$$\widetilde{\mathcal{L}}^{A}(\boldsymbol{\theta}, \boldsymbol{\phi}; \mathbf{x}^{(i)}) = \frac{1}{L}\sum_{l=1}^{L} \log p_{\boldsymbol{\theta}}(\mathbf{x}^{(i)}, \mathbf{z}^{(i,l)}) - \log q_{\boldsymbol{\phi}}(\mathbf{z}^{(i,l)}|\mathbf{x}^{(i)})$$
$$\text{where} \quad \mathbf{z}^{(i,l)} = g_{\boldsymbol{\phi}}(\boldsymbol{\epsilon}^{(i,l)}, \mathbf{x}^{(i)}) \quad \text{and} \quad \boldsymbol{\epsilon}^{(l)} \sim p(\boldsymbol{\epsilon})$$

KL-divergence term은 $q_{\phi}(z \mid x)$가 $p_{\theta}(x)$와 가까워지도록 ϕ를 제어함으로써 얻을 수 있다. 이로써 얻을 수 있는 두 번째 SGVB estimator는 다음과 같다.

$$\widetilde{\mathcal{L}}^{B}(\boldsymbol{\theta}, \boldsymbol{\phi}; \mathbf{x}^{(i)}) = -D_{KL}(q_{\boldsymbol{\phi}}(\mathbf{z}|\mathbf{x}^{(i)})||p_{\boldsymbol{\theta}}(\mathbf{z})) + \frac{1}{L}\sum_{l=1}^{L}(\log p_{\boldsymbol{\theta}}(\mathbf{x}^{(i)}|\mathbf{z}^{(i,l)}))$$
$$\text{where} \quad \mathbf{z}^{(i,l)} = g_{\boldsymbol{\phi}}(\boldsymbol{\epsilon}^{(i,l)}, \mathbf{x}^{(i)}) \quad \text{and} \quad \boldsymbol{\epsilon}^{(l)} \sim p(\boldsymbol{\epsilon})$$

따라서 multiple datapoint에서 marginal likelihood lower bound의 estimator는 다음과 같이 나타낼 수 있다.

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

lower bound의 미분값은 stochastic optimization method를 통해 구할 수 있다. 이에 대한 알고리즘은 아래와 같다.

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 $\boldsymbol{\theta}, \boldsymbol{\phi}$

2.4 The reparameterization trick

$q_\phi(z \mid x)$의 generating samples를 얻기 위해, reparameterization trick을 사용한다. 이러한 reparmeterization은 $q_\phi(z \mid x)$을 미분 가능한 estimator로 만든다.

표준 정규 분포에서 $z = g_\phi(\epsilon, x)$일 때,

$$\int q_\phi(z|x) f(z)\, dx = \int p_\epsilon f(z)\, d\epsilon = \int p(\epsilon) f(z)\, d\epsilon = \int p(\epsilon) f(g_\phi(\epsilon, x))\, d\epsilon$$

로 나타낼 수 있으며, 이는 미분 가능하다.

다음 3가지 경우에서 reparameterization trick이 유효하다.

1. Tractable한 inverse CDF
2. $g(.) = location + scale \cdot \epsilon$형태로 표현 가능한 location-scale 분포
3. Composition

3. Example: Variational Auto-Encoder

확률적 인코더 $q_\phi(z \mid x)$에 신경망 네트워크를 사용하면 VAE를 만들 수 있다. $p_\theta(z \mid x)$를 다변량 가우시안/베르누이 분포를 따른다고 할 때, z를 MLP로 계산할 수 있다. 하지만, $p_\theta(z \mid x)$은 intractable하기 때문에 직접 계산할 수 없다. 이때 근사치인 $q_\phi(z \mid x)$이 다변량 가우시안 분포를 따른다고 한다면, 아래와 같이 나타낼 수 있다.

$$\log q_{\boldsymbol{\phi}}(\mathbf{z}|\mathbf{x}^{(i)}) = \log \mathcal{N}(\mathbf{z}; \boldsymbol{\mu}^{(i)}, \boldsymbol{\sigma}^{2(i)}\mathbf{I})$$

위 2.4에서와 같이, z의 sample 결과는 아래와 같이 나타낼 수 있다.

$$z^{(i,l)} = g_\phi\left(x^{(i)}, \epsilon^{(l)}\right) = \mu^{(i)} + \sigma^{(i)}\epsilon^{(l)} \quad \text{where} \quad \epsilon^{(l)} \sim \mathcal{N}(0, I)$$

이 모델에서는 $p_\theta(z)$와 $q_\phi(z \mid x)$가 모두 가우시안 분포를 따르기 때문에, 2.3과 같이 KL-divergence는 추정 없이 계산 및 미분된다. 그 결과, 각 datapoint x(i)에서의 estimator는 다음과 같이 나타낼 수 있다.

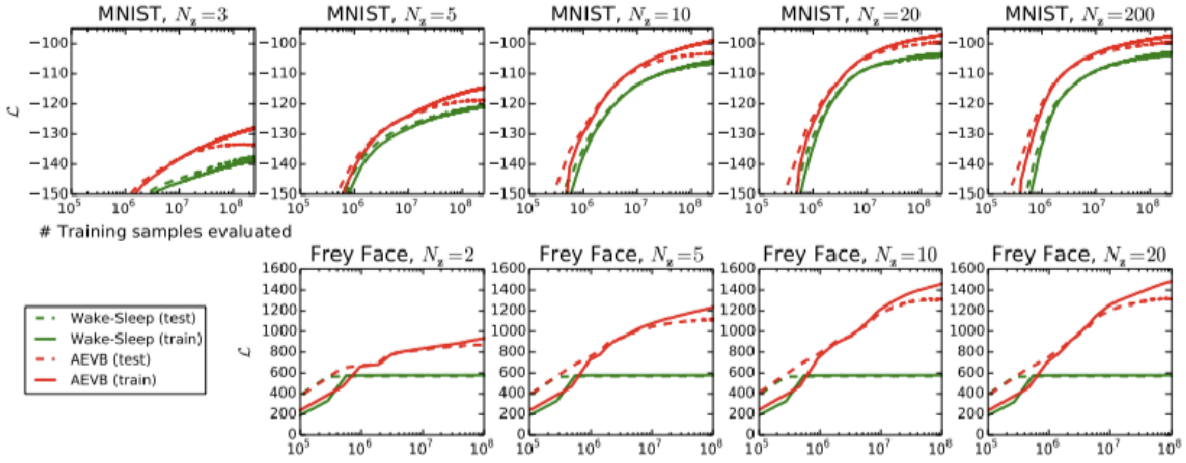
$$\mathcal{L}(\boldsymbol{\theta}, \boldsymbol{\phi}; \mathbf{x}^{(i)}) \simeq \frac{1}{2}\sum_{j=1}^{J}\left(1 + \log((\sigma_j^{(i)})^2) - (\mu_j^{(i)})^2 - (\sigma_j^{(i)})^2\right) + \frac{1}{L}\sum_{l=1}^{L}\log p_{\boldsymbol{\theta}}(\mathbf{x}^{(i)}|\mathbf{z}^{(i,l)})$$

$$\text{where} \quad \mathbf{z}^{(i,l)} = \boldsymbol{\mu}^{(i)} + \boldsymbol{\sigma}^{(i)} \odot \boldsymbol{\epsilon}^{(l)} \quad \text{and} \quad \boldsymbol{\epsilon}^{(l)} \sim \mathcal{N}(0, \mathbf{I})$$

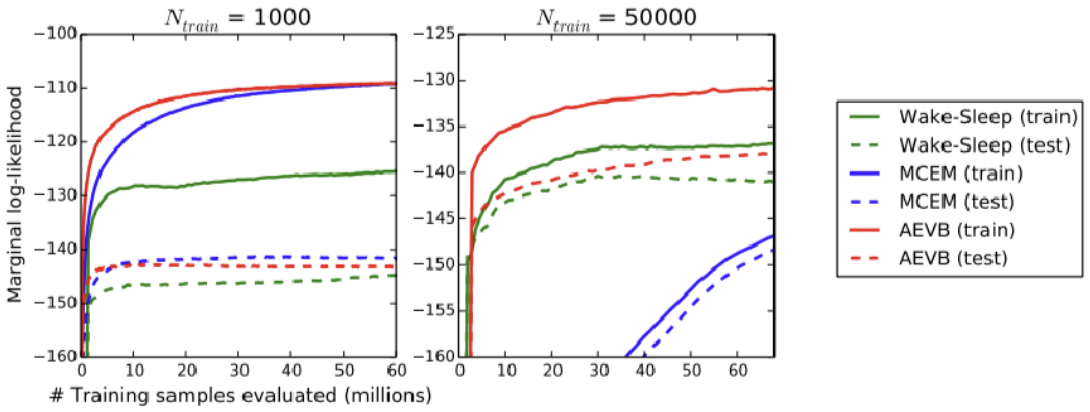
decoding term인 $\log p_\theta \left(x^{(i)} \mid z^{(i,l)}\right)$는 베르누이/가우시안 MLP이다.

4. Experiement

MNIST dataset과 Frey Face dataset을 사용하였으며, variational lower bound와 marginal likelihood의 추정값을 비교하였다.



AEVB algorithm과 Wake-Sleep algorithm을 비교하면, AEVB의 Likelihood lower bound가 더 높다는 것을 알 수 있다.



AEVB algorithm, Wake-Sleep algorithm, Monte Carlo EM을 비교하면, AEVB의 Marginal likelihood가 가장 높다.

5. Conclusion

연속적 잠재변수와 효율적인 근사 추론을 위해 SGVB(Stochastic Gradient VB) estimator와 AEVB(Auto-Encoding VB) algorithm을 제안하였다. 각 datapoint가 연속

적 잠재변수를 가지는 i.i.d 데이터셋에서, SGVB estimator는 variational lower bound를 최적화하고, 이를 활용하면 AEVB algorithm의 효율적인 추론 및 학습이 가능하다.

6. Future work

- encoder와 decoder에 사용되는 deep neural networks를 통해 계층적 생성 아키텍처를 AEVB와 함께 학습
- 시계열 모델
- global parameter에 SGVB 적용
- 복잡한 노이즈 분포 학습에 유용한 잠재변수를 가진 지도 학습 모델