서울대학교 통계학과 베이즈통계 연구실

베이즈 딥러닝 모형

January 29, 2023

이경원

목차

신경망 모형과 인공지능

- 최근의 인공지능 시스템은 대부분 신경망(neural network) 모형을 활용하고 있음

Figure 1: Tesla Autopilot

불확실성 추론

- 특정 값의 예측(prediction)을 넘어, 예측의 불확실성(uncertainty)를 함께 추론하는 문제가 중요해지고 있다.
- 불확실성의 추론이 가능한 신경망 모형인 베이즈 신경망 모형(Bayesian neural network model)에 대해 소개하고자 한다.

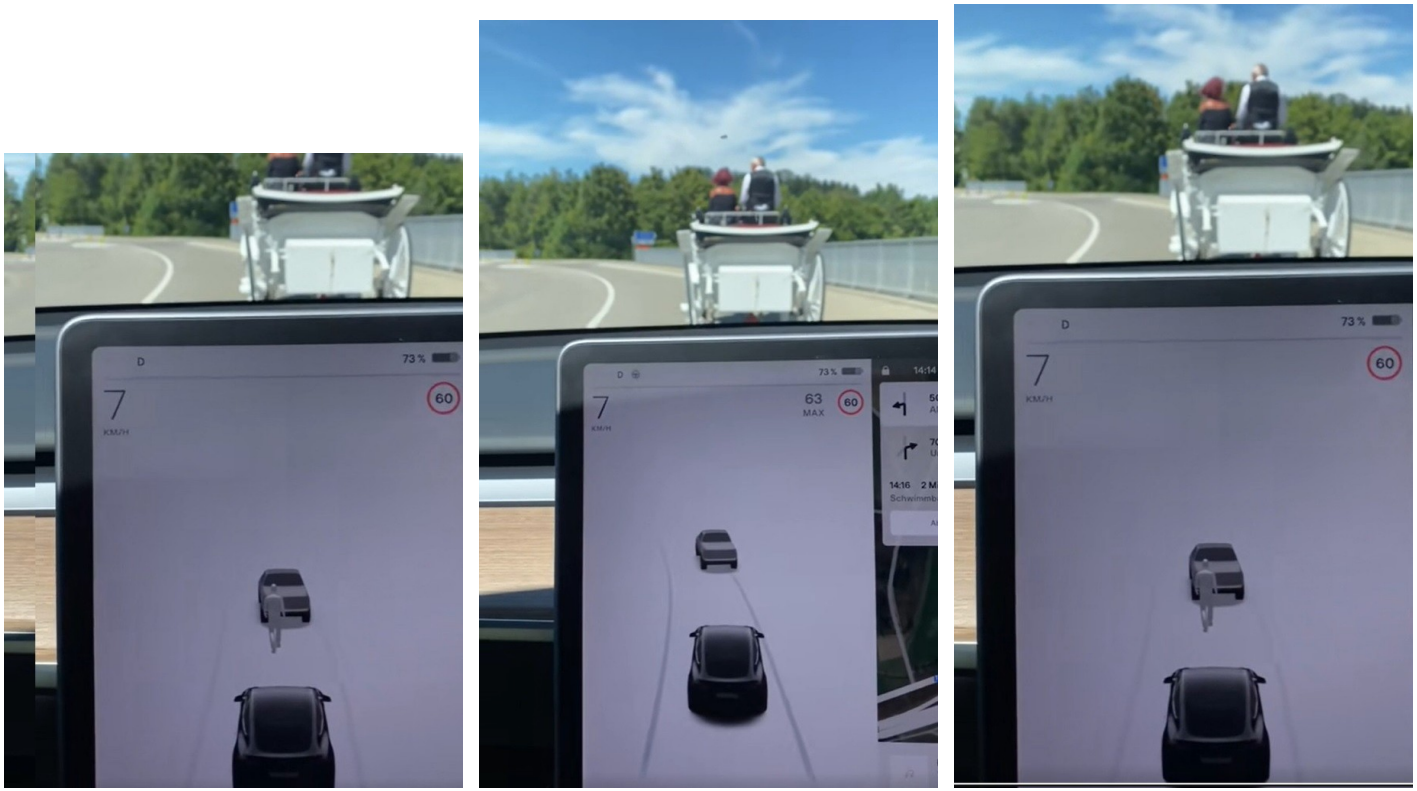


Figure 2: ./figs/ 마주한 테슬라 자율주행 시스템

목차

불확실성의 분해

- 예측의 불확실성은 크게 우연적(aleatoric) 및 인식적(epistemic) 불확실성이라고 하는 두 가지 불확실성으로 분해할 수 있다 (Hüllermeier & Waegeman, 2021).

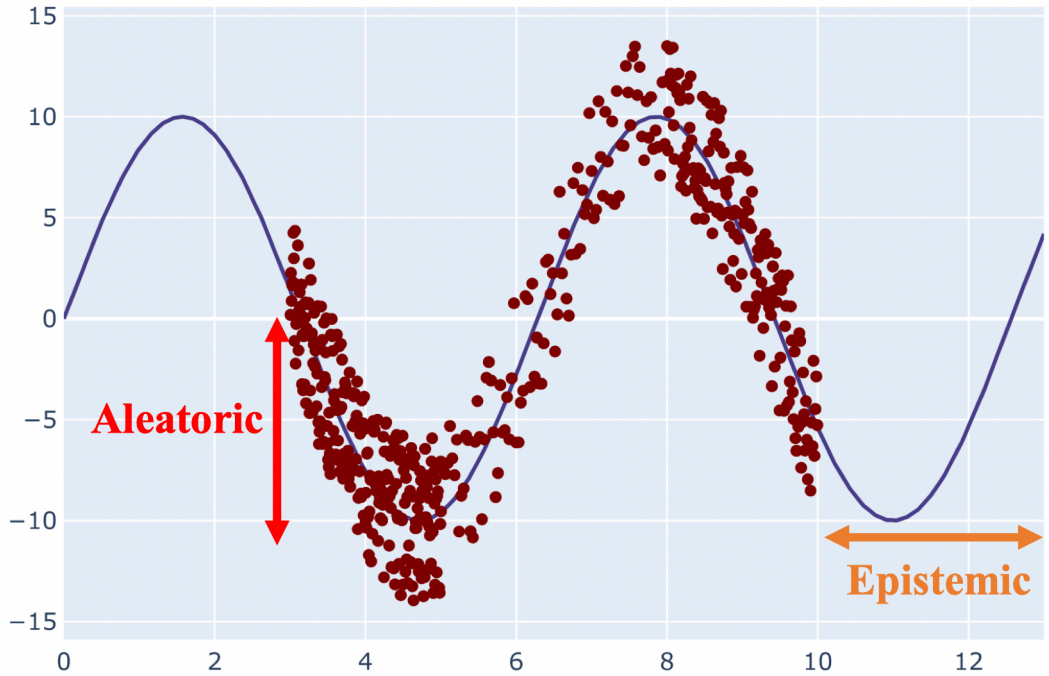


Fig. 1: A schematic view of main differences between aleatoric and epistemic uncertainties.

Figure 3: Fig. 1 in Abdar et al. (2021).

불확실성의 분해

우연적 불확실성

- 우연적 불확실성은 대게 자료에 내재된 확률 분포에서 발생한다 (Phan, 2019).
- 이러한 유형의 불확실성은 모형이 아니라 자료에 내재된 것이므로 줄일 수 없다 (irreducible).

인식적 불확실성

- 인식적 불확실성은 자료와 모형에 대한 지식의 부족에서 발생한다.
- 이러한 불확실성은 더 많은 정보를 수집하는 것으로 줄일 수 있다.

통계적 모형화

비모수 회귀모형

n개의 관측치 $\mathbb{D}_n = (X_i, y_i)_{i=1}^n \subset [0,1]^d \times \mathbb{R}$에 다음과 같은 회귀모형을 가정하자.

$$y_i = f_0(X_i) + \xi_i \ (i = 1, 2, \cdots, n),$$

여기서 $(\xi_i)_{i=1}^n$은 서로 독립이고 동일한 분포 $\mathcal{N}(0, \sigma^2)$를 따르는 분산 $\sigma^2 > 0$이 알려진 정규오차이며 f_0는 어떤 함수 공간 $\mathscr{F}$에 속하는 실제 회귀함수라 하자.

신경망 모형

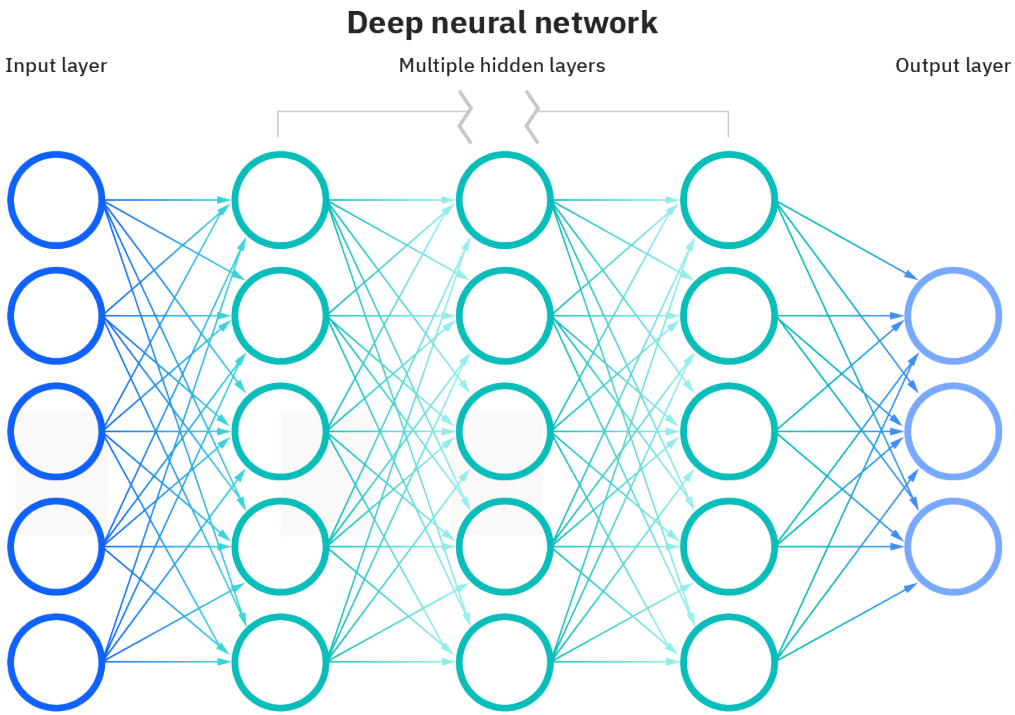


Figure 4: 간단한 형태의 신경망 모형, source: IBM.

신경망 모형

신경망 모형

신경망 모형은 선형변환(linear transformation)과 활성 함수(activation function)라는 비선형변환(non-linear transformation)을 조합하여 복잡한 구조를 표현하는 모형이다. 수학적으로는 다음과 같이 나타낼 수 있다.

$$\Phi(\Theta) = \Big\{ f_\theta(x) = (W^{(L+1)}(\cdot) + b^{(L+1)}) \circ \zeta \circ \cdots \circ \zeta \circ (W^{(1)}x + b^{(1)}) : \\ \theta = (W^{(1)}, b^{(1)}, \cdots, W^{(L+1)}, b^{(L+1)}) \in \Theta \Big\},$$

여기서 ζ는 활성함수, Θ는 가중치(weight), 편향(bias) 모수를 모아놓은 모수 공간이다.

불확실성의 분해

앞서 언급하였듯이, 예측의 불확실성(predictive uncertainty; PU)는 인식적 불확실성(epistemic uncertainty; EU)과 우연적 불확실성(aleatoric uncertainty; AU)으로 분해할 수 있다. 즉, 다음과 같이 나타낼 수 있음을 의미한다.

$$PU = EU + AU. \tag{1}$$

인식적 불확실성은 모형 공간에 대한 확률분포의 형태로, 우연적 불확실성은 자료에 대한 확률분포의 형태로 표현할 수 있다 (Abdar et al., 2021).

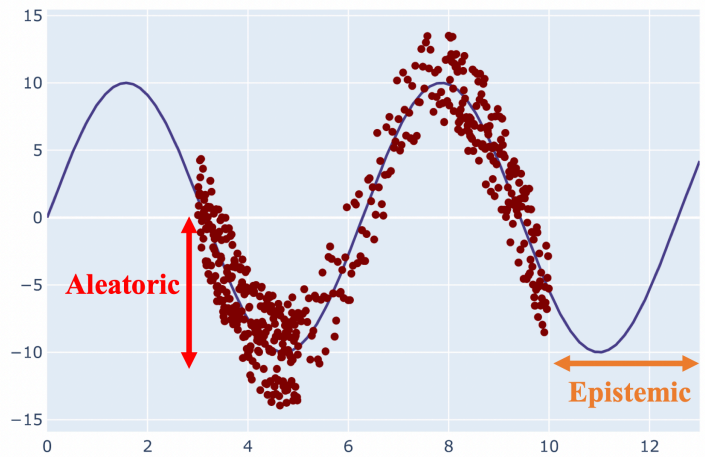


Fig. 1: A schematic view of main differences between aleatoric and epistemic uncertainties.

베이즈 신경망 모형

- 베이즈 신경망 모형은 신경망 모형의 불확실성을 신경망 모형의 모수 공간에 확률분포를 부여하는 것으로 표현한다.
- 이때, 자료를 관측하기 전 모수에 대한 확률분포 $\pi(\theta)$를 사전 분포(prior distribution), 자료를 관측한 뒤 모수에 대한 확률분포 $\pi(\theta|\mathbb{D}_n)$를 사후 분포(posterior distribution)이라 부른다.
- 베이즈 모형은 베이즈 정리를 바탕으로 자료로부터 모수에 대한 정보를 업데이트 한다.
- 사후 분포를 계산하고 나면, 다음과 같이 새로운 자료 X^*에서의 확률분포인 사후 예측 분포(posterior predictive distribution)를 계산할 수 있고, 이 분포를 통해 예측에 대한 불확실성을 표현할 수 있다.

$$p(y^*|X^*, \mathbb{D}_n) = \int p(y^*|X^*, \theta)\pi(\theta|\mathbb{D}_n)d\theta.$$

베이즈 신경망 모형

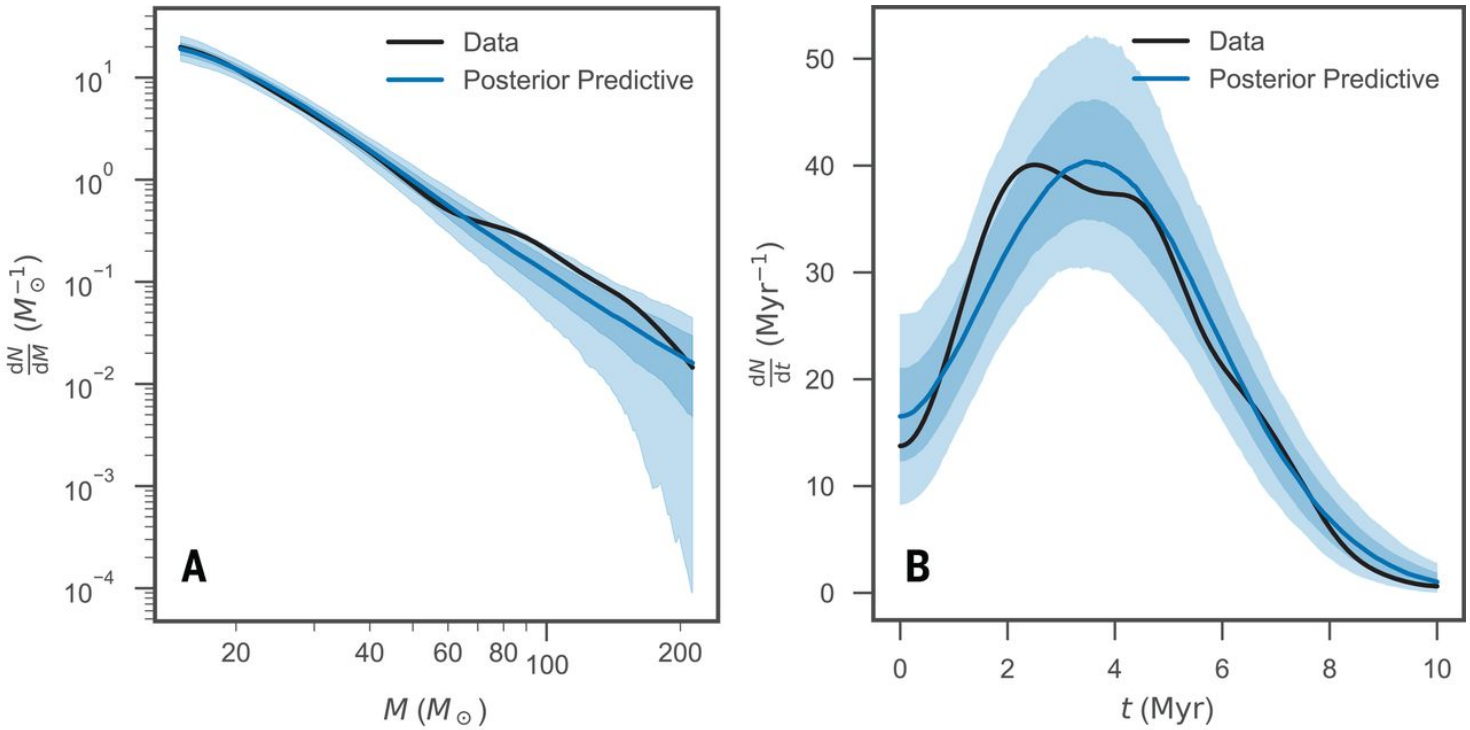


Figure 5: 사후 예측 분포의 예, figure from Farr and Mandel (2018)

베이즈 신경망 모형

- 베이즈 신경망 모형은 기존의 (빈도론적) 신경망 모형이 제공하지 못하는 인식적 불확실성에 대한 정보를 제공할 수 있어 기대를 받고 있다 (Abdar et al., 2021).
- 베이즈 신경망 모형은 과적합에 강건하며 훈련 자료의 크기가 작은 상황에서도 잘 작동한다는 것이 알려져 있다 (Kucukelbir, Tran, Ranganath, Gelman, & Blei, 2016).

베이즈 신경망 모형

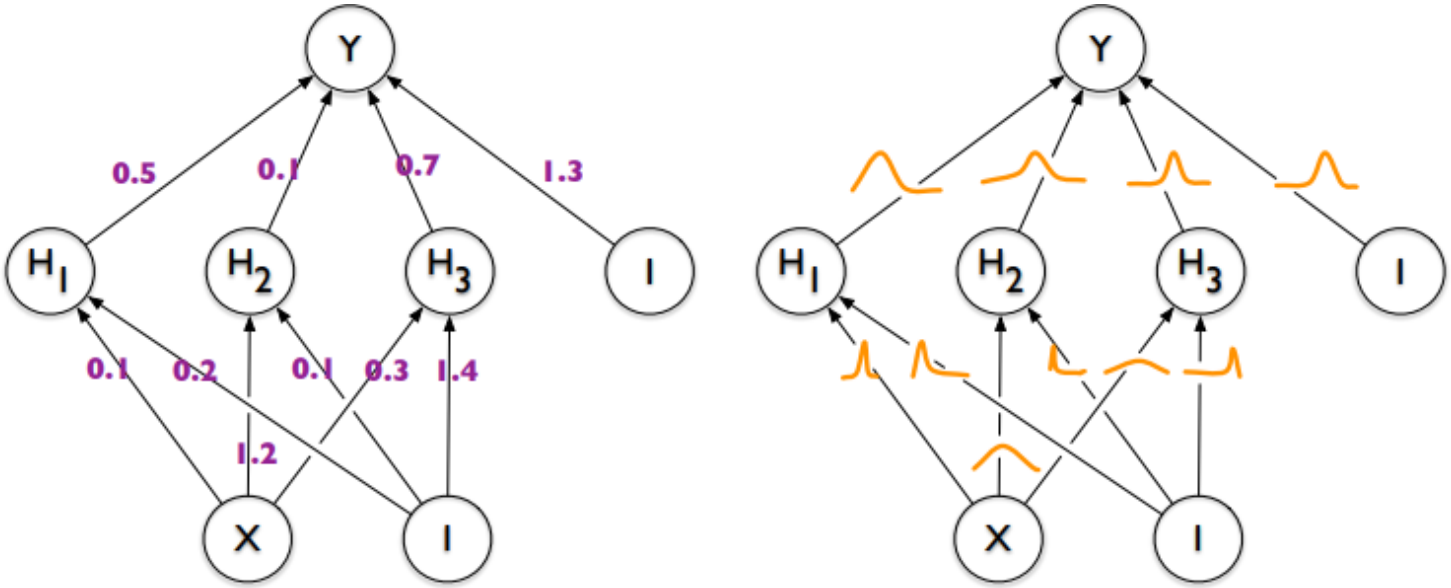


Figure 6: 기존의 신경망 모형과 베이즈 신경망 모형의 비교, figure from Blundell et al. (2015).

목차

베이즈 신경망 모형의 문제점

- 베이즈 신경망 모형 또한, 다른 베이즈 모형과 마찬가지로 계산과 관련된 여러 문제를 겪는다.
- 자료의 수가 많을 때 가능도를 반영한 사후 분포의 계산이 어렵다.
- 모수 공간의 차원이 높을 때, 사후 분포의 계산이 어려워지는데 신경망 모형은 모수공간의 차원이 매우 높아 이러한 문제가 두드러진다.
- 여기서는 베이즈 신경망 모형을 계산하기 위한 다양한 방법들을 소개한다.

변분 근사

- 사후 분포 $\pi(\theta|\mathbb{D}_n)$ 계산이 해석적으로 어려울 때, 이를 적당한 변분 분포 $q_\phi(\theta)$로 근사하는 방법에 대해 생각해볼 수 있다.
- 이 방법의 핵심은 사후 분포와 가장 가까운 변분 분포를 찾는 데 있다.
- 일반적으로는, 다음과 같이 쿨벡-라이블러 발산을 최소화하는 변분 모수 ϕ를 찾는다.

$$KL(q_\phi(\theta)||p(\omega|D_n)) = \int q_\phi(\theta) \log \frac{q_\phi(\theta)}{p(\omega|D_n)} d\omega. \tag{2}$$

변분 근사

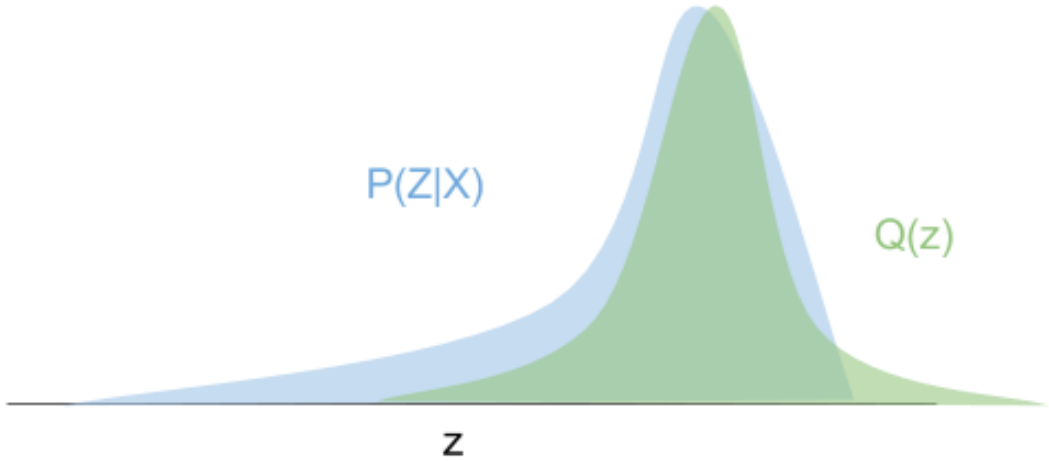


Figure 7: 확률분포 $P(Z|X)$를 근사하는 변분 분포 $Q(Z)$

사후 예측 분포 또한 다음과 같이 변분 분포를 사용하여 근사할 수 있다.

$$p(y^*|X^*, D_n) \approx \int p(y^*|X^*, \theta) q_\phi(\theta) d\theta =: q_\phi^*(y^*, X^*), \tag{3}$$

변분 근사

사후 분포의 형태를 모르는 상황에서도 최적의 변분 분포를 찾는 것은 가능한데, 다음과 같이 정의되는 증거의 하한(evidence lower bound; ELBO)를 최대화하는 변분 모수를 찾으면 된다.

$$\mathcal{L}_{VI}(\phi) := \int q_\phi(\theta) \log p(D_n|\theta) d\theta - KL(q_\phi(\theta)||p(\theta)), \tag{4}$$

이러한 과정을 **변분 추론(variational inference; VI)** 혹은 **변분 근사(variational approximation)**이라 부른다.

베이즈 신경망 모형과 변분 근사

- 베이즈 신경망 모형에서 변분 근사는 모수 공간의 변분 분포를 찾는 것으로 이루어진다.
- 변분 근사는 사후 분포를 계산하는 문제는 변분 모수를 찾는 최적화 (optimization) 문제로 환원하며, 확률적 경사 하강법(stochastic gradient descent)와 같은 방법을 사용하여 대용량 자료에도 적용이 가능하다.

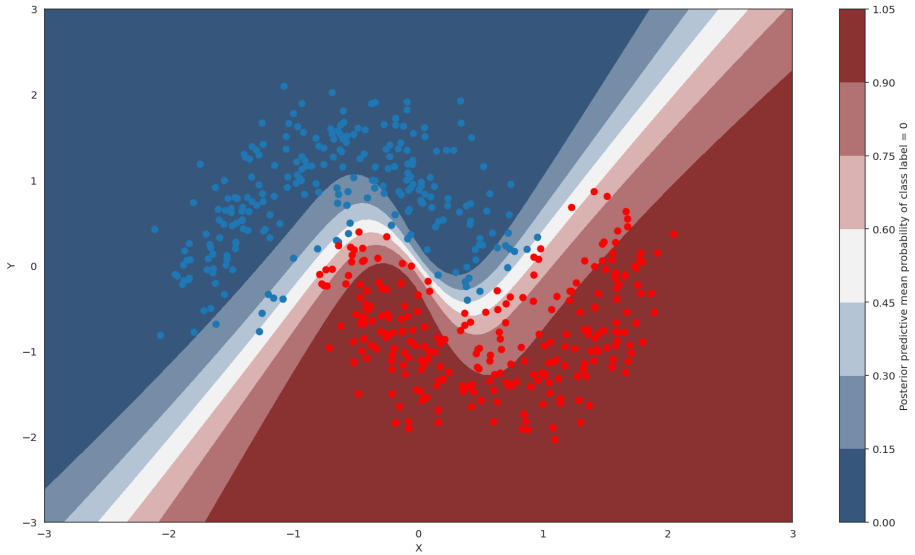


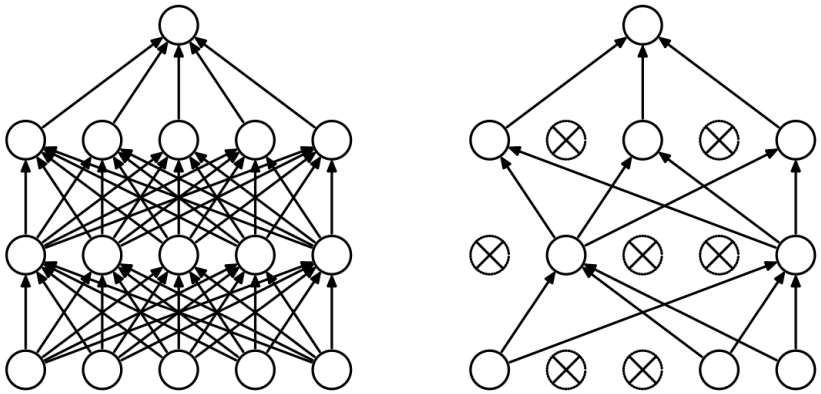
Figure 8: 변분 근사를 적용한 베이즈 분류 신경망의 예.

몬테 카를로 드롭아웃 - 배경

- 몬테 카를로(Monte Carlo) 방법은 복잡한 적분을 확률적으로 근사하는 방법으로 고차원 적분의 계산에서 주로 사용된다.
- 베이즈 추론에서는 적당한 과정을 통해 얻은 사후 표본에 몬테 카를로 방법을 적용하여 근사적으로 베이즈 추론을 실시한다 (Neal, 2012).
- 모수공간의 차원이 매우 높은 베이즈 신경망 모형에서 몬테 카를로 방법은 유용하게 사용된다.

드롭아웃

- 드롭아웃(dropout, Srivastava, Hinton, Krizhevsky, Sutskever, and Salakhutdinov (2014))은 학습 중 신경망 모형의 일부 모수를 비활성화하는 방법으로, 신경망 모형의 과적합 현상을 해결하기 위해 주로 사용된다.



(a) Standard Neural Net (b) After applying dropout.

Figure 9: Figure 1 in Srivastava et al. (2014), Dropout Neural Net Model. Left: A standard neural net. Right: An example of a thinned net produced by applying dropout to the network on the left.

몬테 카를로 드롭아웃

- Gal and Ghahramani (2016)에 의해 제안된 몬테 카를로 드롭아웃은 드롭아웃을 적당한 베이즈 모형의 근사로 해석하여 드롭아웃에서 얻어진 모수를 사후 표본으로 간주하여 몬테 카를로 근사를 적용한다.

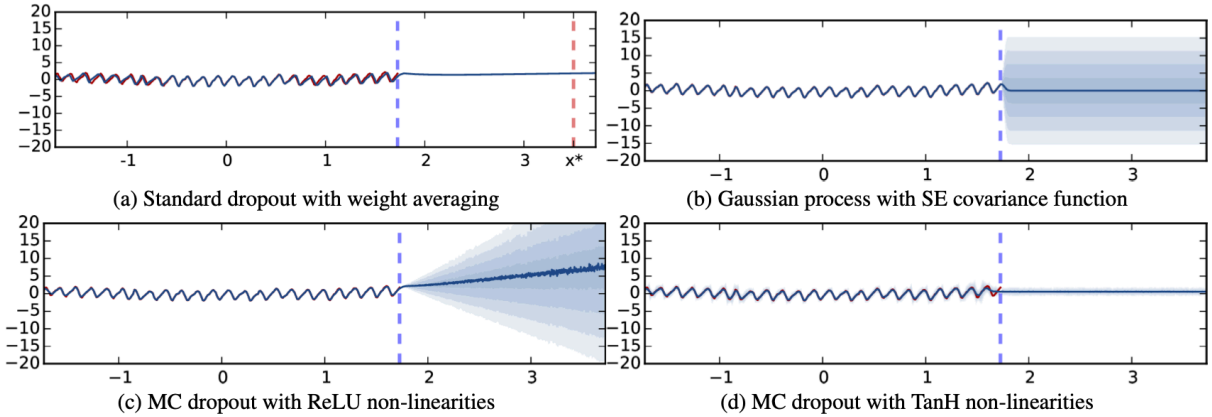


Figure 10: Figure 2 in Gal and Ghahramani (2016), Predictive mean and uncertainties on the Mauna Loa CO2 concentrations dataset, for various models.

몬테 카를로 드롭아웃

- Gal and Ghahramani (2016)은 몬테 카를로 드롭아웃 방법이 사후분포를 두 개의 점과 가중치로 근사하는 변분 추론임을 보였다.
- 몬테 카를로 드롭아웃은 기존 신경망 모형을 그대로 활용할 수 있고 계산이 빨라 복잡한 신경망 모형의 불확실성 추론에서 자주 사용된다.

Figure 11: MC dropout example for simple regression problem.

마르코프 체인 몬테 카를로

- 마르코프 체인 몬테 카를로(Markov chain Monte Carlo; MCMC) 방법은 마르코프 체인을 통해 근사적으로 사후 표본을 얻고, 몬테 카를로 방법을 통해 근사적으로 베이즈 추론을 실시하는 방법이다.
- MCMC 알고리듬은 마르코프 체인으로부터 얻어진 표본이 사후 분포를 정상 분포(stationary distribution)로 갖도록 구성한다.
- 구성 방법에 따라 여러 가지 알고리듬들이 제안되어 왔는데, 대표적으로 메트로폴리스-헤이스팅스 (MH) 알고리듬, 깁스 표집기(Gibbs sampler), 해밀토니안 몬테 카를로(Hamiltonian monte carlo; HMC, Duane, Kennedy, Pendleton, and Roweth (1987)) 등이 있다.

깁스 표집기

깁스 표집기는 사후 분포의 조건부 분포로부터 모수의 성분을 교대로 추출하여 표본을 얻는다.

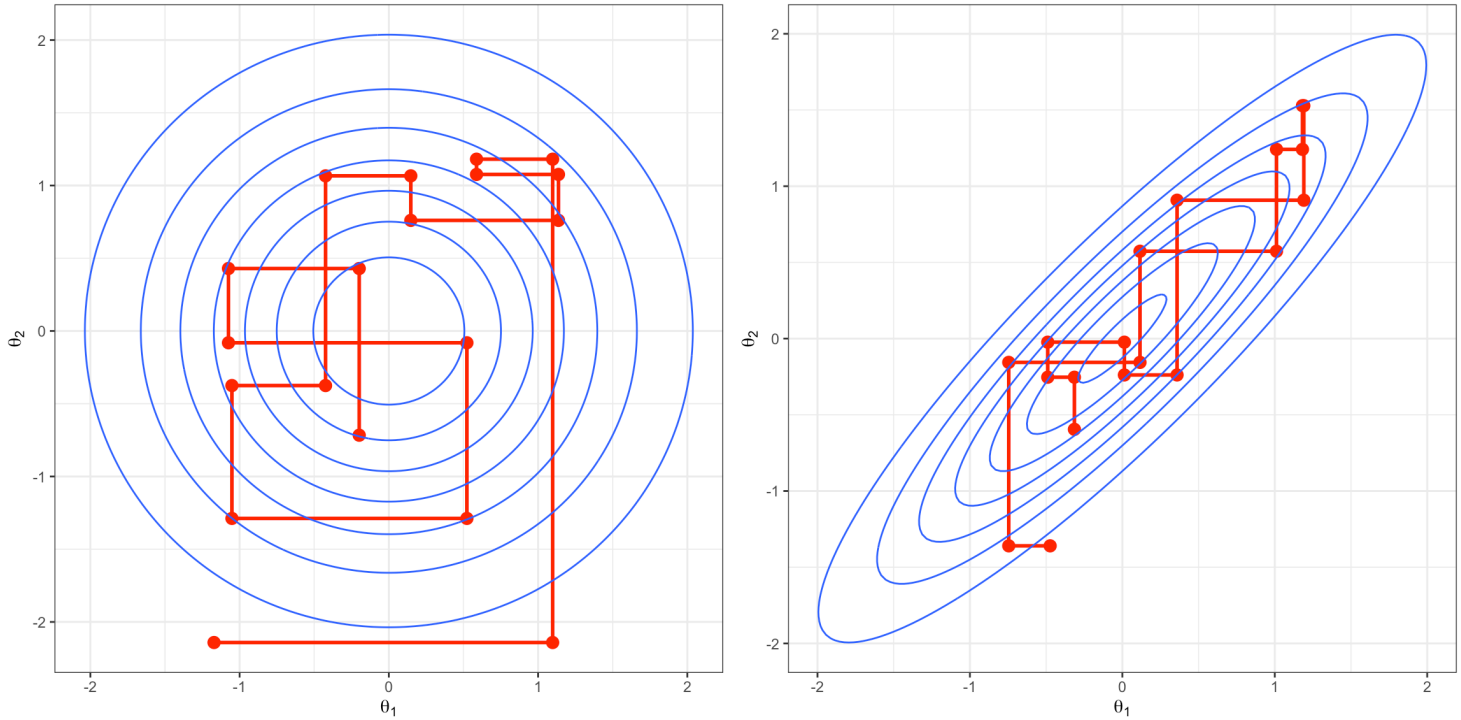


Figure 12: Gibbs sampler for bivariate normal distribution, from 이재용 and 이기재 (2022)

해밀토니안 몬테 카를로

해밀토니안 몬테 카를로는 해밀토니안 동역학(Hamiltonian dynamics)의 원리를 응용하여 표본을 얻는다.

Algorithm 1: Hamiltonian Monte Carlo

Input: Starting position $\theta^{(1)}$ and step size ϵ
for $t = 1, 2 \cdots$ **do**
  Resample momentum r
  $r^{(t)} \sim \mathcal{N}(0, M)$
  $(\theta_0, r_0) = (\theta^{(t)}, r^{(t)})$
  Simulate discretization of Hamiltonian dynamics in Eq. (4)*:*
  $r_0 \leftarrow r_0 - \frac{\epsilon}{2}\nabla U(\theta_0)$
  for $i = 1$ ***to*** m **do**
    $\theta_i \leftarrow \theta_{i-1} + \epsilon M^{-1} r_{i-1}$
    $r_i \leftarrow r_{i-1} - \epsilon \nabla U(\theta_i)$
  end
  $r_m \leftarrow r_m - \frac{\epsilon}{2}\nabla U(\theta_m)$
  $(\hat{\theta}, \hat{r}) = (\theta_m, r_m)$
  Metropolis-Hastings correction:
  $u \sim \text{Uniform}[0, 1]$
  $\rho = e^{H(\hat{\theta}, \hat{r}) - H(\theta^{(t)}, r^{(t)})}$
  if $u < \min(1, \rho)$, **then** $\theta^{(t+1)} = \hat{\theta}$
end

Figure 13: Hamiltonian Monte Carlo, from (Chen et al., 2014)

해밀토니안 몬테 카를로

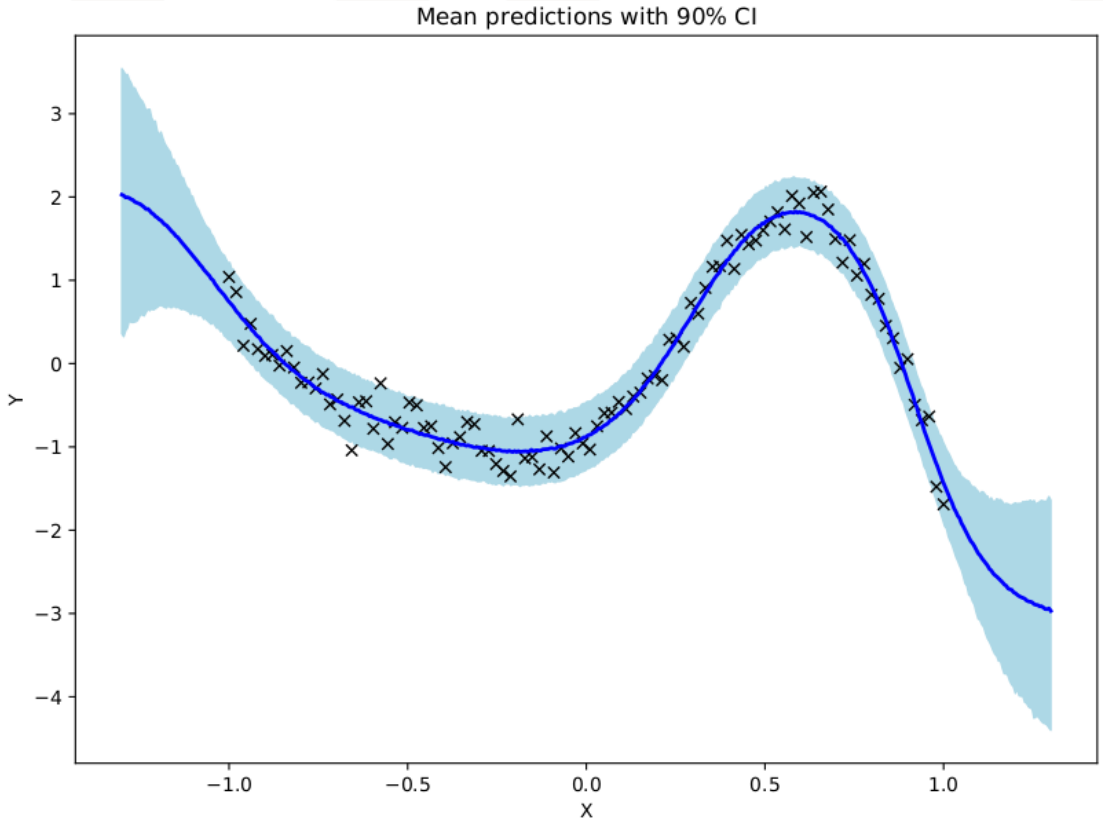


Figure 14: Figure in numpyro docs, applying NUTS to do inference on a simple Bayesian neural network with two hidden layers.

마르코프 체인 몬테 카를로 방법의 문제점

- MCMC 방법은 이론적으로 충분한 표본을 얻으면 사후 분포를 근사할 수 있음이 알려져 있으나 계산이 상당히 비싸다는 문제를 갖는다 (Neal, 2012).
- 이러한 문제는 고차원 모수공간에서 두드러지는데, 베이즈 신경망 모형과 같은 고차원 베이즈 모형을 학습시키는 데는 해밀토니안 몬테 카를로 방법이 잘 작동하는 것이 알려져있다.
- 해밀토니안 몬테 카를로 방법은 자료의 수가 클 때 계산에 많은 시간이 소요된다는 문제점을 갖는다.

확률적 경사 마르코프 체인 몬테 카를로

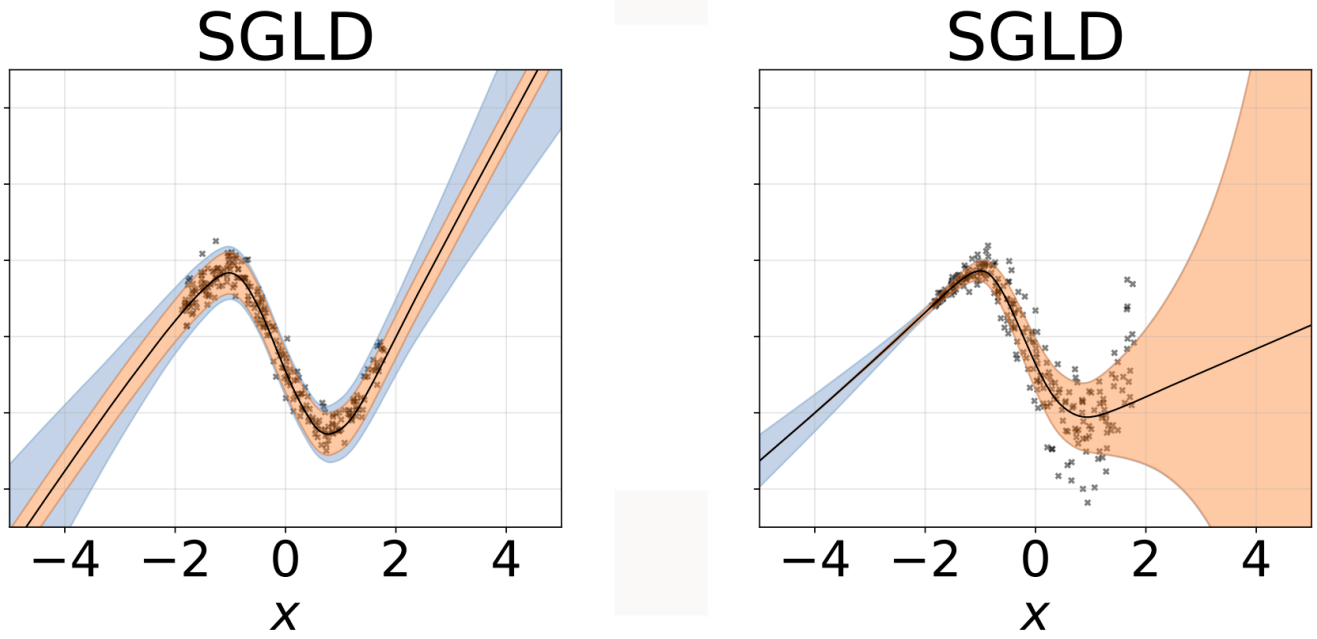
- Chen et al. (2014); Ding et al. (2014) 등에 의해 제안된 확률적 경사 마르코프 체인 몬테 카를로 (stochastic gradient MCMC; SG-MCMC)은 자료를 분할하여 마르코프 체인에서 표본을 얻는다.

Algorithm 2: Stochastic Gradient HMC

for $t = 1, 2 \cdots$ **do**
  optionally, resample momentum r *as*
  $r^{(t)} \sim \mathcal{N}(0, M)$
  $(\theta_0, r_0) = (\theta^{(t)}, r^{(t)})$
  simulate dynamics in Eq.(13)*:*
  for $i = 1$ ***to*** m **do**
    $\theta_i \leftarrow \theta_{i-1} + \epsilon_t M^{-1} r_{i-1}$
    $r_i \leftarrow r_{i-1} - \epsilon_t \nabla \tilde{U}(\theta_i) - \epsilon_t C M^{-1} r_{i-1}$
      $+\mathcal{N}(0, 2(C - \hat{B})\epsilon_t)$
  end
  $(\theta^{(t+1)}, r^{(t+1)}) = (\theta_m, r_m)$, no M-H step
end

Figure 15: Stochastic Gradient Hamiltonian Monte Carlo, from (Chen et al., 2014)

확률적 경사 마르코프 체인 몬테 카를로



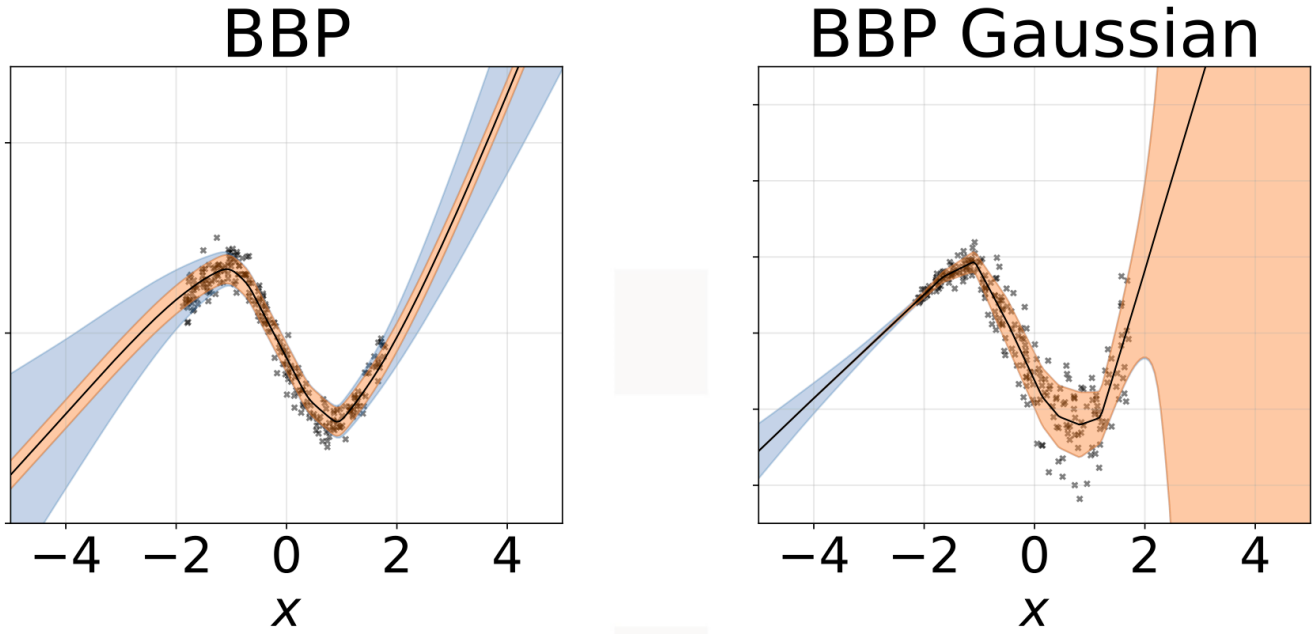
(a) Homoscedastic model

(b) Heteroscedastic model

Figure 16: BNN with stochastic gradient Langevin dynamics (SGLD, Welling and Teh (2011)), figure from github.

이 외에도 다음과 같은 알고리듬들이 제안되었다.

- Blundell et al. (2015)에 의해 제안된 베이즈 역전파(Bayes by Backprop) 알고리듬
- Maddox, Izmailov, Garipov, Vetrov, and Wilson (2019)에 의해 제안된 SWAG(stochastic weight averaging Gaussian) 알고리듬



(a) Homoscedastic model

(b) Heteroscedastic model

Figure 17: Bayes by Backprop, figure from github.

목차

베이즈 신경망 모형의 점근적 성질

- 신경망 모형이 복잡한 문제에서 잘 작동한다는 것이 알려져 있으나, 그 이론적 배경은 많이 밝혀지지 않았다.
- 여기서는 (베이즈) 신경망 모형이 이론적으로 어떤 성질을 갖는 지에 대해 간략히 소개한다.
- 점근적 성질이란, 자료의 수가 충분히 클 때 모형이 갖는 성질을 의미한다.
- 빈도론 모형의 점근적 성질은 일치성(consistency)으로 추정량이 실제 값에 가까워 지는 지, 얼마나 빠른 속도(speed)로 가까워지는 지에 대해 관심을 갖는다. 이때, 최적의 수렴 속도는 최소최대수렴속도(minimax convergence rate)을 고려한다.
- 베이즈 모형의 점근적 성질은 사후일치성(posterior consistency)으로 사후 분포가 실제 값으로 충분히 가까워 지는 지, 얼마나 빠른 속도로 가까워지는 지에 대해 관심을 갖는다.

관련 연구

	Approximation	Frequentist estimation	Bayesian estimation
Hölder space C^s			
Author	Yarotsky (2017)	Schmidt-Hieber (2020)	Polson and Ročková (2018)
Accuracy	$\tilde{O}(N^{-s/d})$	$\tilde{O}(n^{-s/(2s+d)})$	$\tilde{O}(n^{-s/(2s+d)})$
Besov space $B^s_{p,q}$			
Author	Suzuki (2019)	Suzuki (2019)	Our work (Lee & Lee, 2023)
Accuracy	$\tilde{O}(N^{-s/d})$	$\tilde{O}(n^{-s/(2s+d)})$	$\tilde{O}(n^{-s/(2s+d)})$

Table 1: 관련 연구들을 요약한 표이다. N은 신경망 모형의 복잡도를, n은 자료의 크기를, d는 자료의 차원을, s는 실제 함수의 평활도(smoothness) 모수를 의미한다. 모든 속도는 L^2 노음에 대해 계산된 것이며 $\tilde{O}$는 부드러운 big-O 표기로 로그항을 무시한(polylogarithmic) 표기이다.

베소프 공간

- 베소프 공간 $B^s_{p,q}$는 함수의 연속성과 미분가능성에 무관하게 평활도를 정의한 함수 공간으로, 횔더 공간(Hölder space)과 소볼레프 공간(Sobolev space)를 일반화한 공간이다.

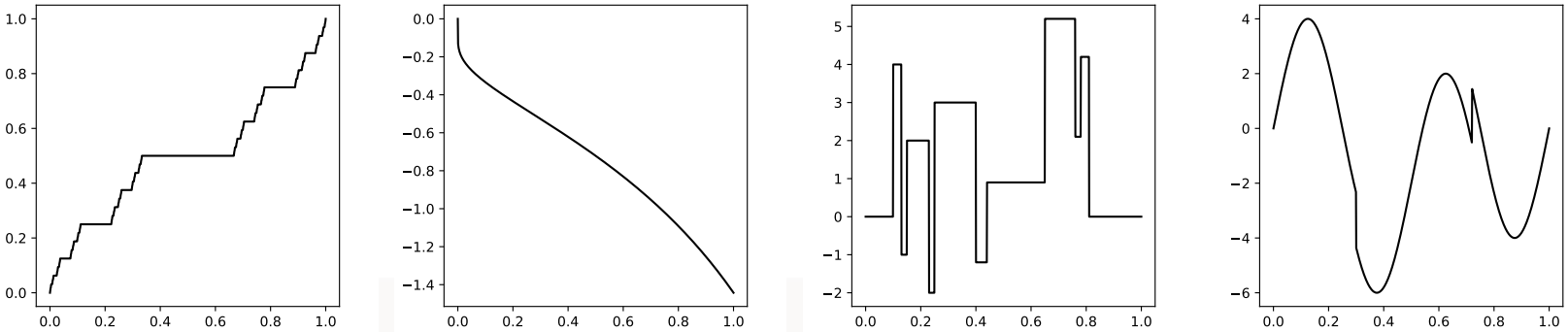


Figure 18: 베소프 공간에 속하는 다양한 함수들

- Donoho and Johnstone (1998)는 베소프 공간 $B^s_{p,q}(\Omega)$에서 최소최대수렴속도가 $n^{-\frac{s}{2s+d}}$ 임을 보였다.
- 해당 논문에서는 커널 회귀모형(kernel regression)과 같은 선형 추정량(linear estimator)의 수렴속도가 $n^{-\frac{s-(1/p-1/2)_+}{2s+1-2(1/p-1/2)_+}}$ 보다 크다는 것을 보였다.

Lee and Lee (2023)

가정

- 실제 회귀 함수 f_0는 베소프 공간 $B^s_{p,q}([0,1]^d)$ 에 속하는 균등 유계(uniformly bounded) 함수이다. 즉, $f_0 \in B^s_{p,q}([0,1]^d) \cap UB(F)$, $UB(F) = \{f : |f(x)| \leq F\}$.
- 평활도 모수 $p,\ s$와 자료의 차원 d에 대해 $d(1/p - 1/2)_+ < s$ 이 성립한다.
- 활성함수는 ReLU$(x) = \max\{0, x\}$이다.
- 설명변수의 주변확률분포 P_X는 유계인 밀도함수를 갖는다.

참고

- $d(1/p - 1/2)_+ \leq s < d/p$일 때, $f \in B^s_{p,q}(\Omega)$는 불연속함수일 수 있다.
- $d = 1$이고 $p < 2$이면 다음이 성립한다

$$n^{-\frac{s}{2s+1}} \ll n^{-\frac{s-(1/p-1/2)_+}{2s+1-2(1/p-1/2)_+}}.$$

즉, 신경망 모형이 선형 추정량보다 잘 작동한다.

베소프 공간에서 신경망 모형의 이론적 성질

- Suzuki (2019)와 Lee and Lee (2023)은 모두 다음과 같은 제약이 있는 신경망 모형 공간을 고려한다.

$$\Theta(L, W, S, B) := \Big\{ W^{(l)} \in \mathbb{R}^{p_{l-1} \times p_l}, b^{(l)} \in \mathbb{R}^{p_l},$$

$$p_l = W, l = 1, 2 \cdots, L,\ p_0 = d, p_{L+1} = 1,\ \|\theta\|_0 \leq S,\ \|\theta\|_\infty \leq B \Big\},$$

$\Phi(L, W, S, B) := \Phi(\Theta(L, W, S, B))$ and

$$\Phi = \bigcup_{L=1}^{\infty} \bigcup_{W=1}^{\infty} \bigcup_{S=0}^{\infty} \bigcup_{B=0}^{\infty} \Phi(L, W, S, B).$$

베소프 공간에서 신경망 모형의 이론적 성질

- Suzuki (2019)는 다음과 같은 모형 모수(model parameter)

$$N_n = \lceil n^{d/(2s+d)} \rceil,\ L_n = O(\log n),$$

$$W_n = O(N_n),\ S_n = O(N_n \log n),\ B_n = O(N_n^{\Xi})$$

for some constant $\Xi = \Xi(d, p, s) \geq 0$. 를 갖는 신경망 모형이 최대가능도추정에서 (거의) 최적의 수렴속도 $n^{-s/(2s+d)}(\log n)^{3/2}$를 가짐을 보였다.

- Lee and Lee (2023)는 이 결과를 베이즈 신경망 모형으로 확장하였다.

Sparsity: spike-and-slab prior

Lee and Lee (2023)의 첫 번째 정리는 다음과 같다.

- Assume prior distribution

$$\pi(\theta_j|\gamma_j, L, W, S, B) = \gamma_j \tilde{\pi}(\theta_j|L, W, S, B) + (1-\gamma_j)\delta_0(\theta_j),$$
$$\pi(\boldsymbol{\gamma}|L, W, S, B) = \frac{1}{\binom{T}{S}}, \tag{5}$$

$$\pi(L = L_n) = \pi(W = W_n) = \pi(S = S_n) = \pi(B = B_n) = 1, \tag{6}$$

  where $\tilde{\pi}(\theta_j|L, W, S, B) = U(\theta_j; [-B, B])$ and $T = |\Theta(L, W, S, B)|$.

- Then the posterior distribution concentrates at the true function with a rate $\epsilon_n = n^{-s/(2s+d)}(\log n)^{3/2}$.
- That is,

$$\Pi(f_\theta \in \Phi \cap UB(F) : \|f_\theta - f_0\|_n > M_n \epsilon_n | \mathbb{D}_n) \to 0$$

  in $P_{f_0}^{(n)}$-probability as $n \to \infty$ for any $M_n \to \infty$.

Sparsity: spike-and-slab prior

참고

- 이 정리는 베이즈 신경망 모형이 베소프 공간에서 빈도론 신경망 모형과 동일한 이론적 최적성을 가짐을 보여준다.
- Polson and Ročková (2018)은 횔더 공간에서 spike-and-slab 사전분포를 고려한 베이즈 신경망 모형이 이론적 최적성을 가짐을 보였다.
- $p = q = \infty$ 일때, 횔더 공간 $C^s(\Omega)$은 베소프 공간 $B^s_{p,q}(\Omega)$에 연속적으로 포함(continuously embed)된다. 즉, Polson and Ročková (2018)의 정리는 위의 정리의 특수한 경우로 해석할 수 있다.

Adaptivity: unknown smoothness

- We assumed knwon model parameters $N_n,\ W_n,\ S_n$ and B_n.
- These parameters depend on the smoothness parameters $p,\ q,\ s$.
- Polson and Ročková (2018) showed that the Bayesian neural network with spike-and-slab prior achive same optimal rate in Hölder space $C^s(\Omega)$.
- We extened the results to Besov space with following priors on model parameters:

$$\begin{aligned} \tilde{L}_n(H) &= \lceil H(\log n) \rceil \vee 1,\ \tilde{W}_n(H,N) = HN, \\ \tilde{S}_n(H,N) &= HN\tilde{L}_n(H),\ \tilde{B}_n(H,N) = N^H. \end{aligned} \tag{7}$$

where

$$N \overset{d}{\equiv} 1 \vee \lceil Z/(\log n)^2 \rceil, \quad \pi_Z(Z) = \frac{\lambda_N^Z}{Z!(e^{\lambda_N}-1)} \quad \text{for } Z = 1, 2, \cdots. \tag{8}$$

- We get the same posterior convergence rate $n^{-s/(2s+d)}(\log n)^{3/2}$.

Relaxation: shrinkage prior

- Unlike spike-and-slab prior, shrinkage priors are relatively straightforward to implement.
- The shrinkage prior avoids the posterior computation with varying dimensions, and enables feasible computation.
- We extened the results to shrinkage prior

$$\pi(\theta|L,W,S,B) = \prod_{j=1}^{T} g(\theta_j|L,W,S,B), \tag{9}$$

where $g(t) := g(t|L,W,S,B)$ is a symmetric density function and decreasing on $t > 0$.

Relaxation: shrinkage prior

1
$$a_n \leq \frac{\epsilon_n}{72 L_n (B_n \vee 1)^{L_n - 1} (W_n + 1)^{L_n}}$$
$$u_n = \int_{[-a_n, a_n]} g(t|L_n, W_n, S_n, B_n) dt \tag{10}$$

satisfies

$$\frac{S_n}{T_n} > 1 - u_n \geq \frac{S_n}{T_n} \eta_n, \tag{11}$$

where $\eta_n = \exp(-K n \epsilon_n^2 / S_n)$.

2 $g(t|L_n, W_n, S_n, B_n)$ is continuous on $[-B_n, B_n]$ and

$$-\log g(B_n|L_n, W_n, S_n, B_n) \lesssim (\log n)^2, \tag{12}$$

3
$$v_n = \int_{[-B_n, B_n]^c} g(t|L_n, W_n, S_n, B_n) dt = o\left(e^{-K_0 n \epsilon_n^2}\right). \tag{13}$$

for some $K,\ K_0 > 4$.

Relaxation: shrinkage prior

Remarks

- Note that
  1. The first condition (11) is a continuous relaxation for the spike part of (5).
  2. The second condition (12) means that the prior should have enough thick tail to sample true function.
  3. The last condition (13) restricts the thickness of the tail to prevent a function from being divergent.
- One can use the relaxed spike-and-slab prior which replace the spike prior (δ_0) to the continuous prior with enough mass around zero.

목차

요약

본 발표에서는 인공지능 분야에서의 불확실성 추론을 위한 베이즈 추론 방법들에 대해 소개하였다. 특히,

- Variational Inference
- Monte Carlo Dropout
- Markov Chain Monte Carlo

와 같은 베이즈 계산 방법들과 베이즈 신경망 모형의 이론적 성질들에 대해 소개하였다.

References

Abdar, M., Pourpanah, F., Hussain, S., Rezazadegan, D., Liu, L., Ghavamzadeh, M., ⋯ others (2021). A review of uncertainty quantification in deep learning: Techniques, applications and challenges. *Information Fusion*.

Blundell, C., Cornebise, J., Kavukcuoglu, K., & Wierstra, D. (2015). Weight uncertainty in neural network. In *International conference on machine learning* (pp. 1613–1622).

Chen, T., Fox, E., & Guestrin, C. (2014). Stochastic gradient hamiltonian monte carlo. In *International conference on machine learning* (pp. 1683–1691).

Ding, N., Fang, Y., Babbush, R., Chen, C., Skeel, R., & Neven, H. (2014). Bayesian sampling using stochastic gradient thermostats.

Donoho, D. L., & Johnstone, I. M. (1998). Minimax estimation via wavelet shrinkage. *The annals of Statistics*, *26*(3), 879–921.

Duane, S., Kennedy, A. D., Pendleton, B. J., & Roweth, D. (1987). Hybrid monte carlo. *Physics letters B*, *195*(2), 216–222.

Farr, W. M., & Mandel, I. (2018). Comment on "an excess of massive stars in the local 30 doradus starburst" . *Science*, *361*(6400).

Gal, Y., & Ghahramani, Z. (2016). Dropout as a bayesian approximation: Representing model uncertainty in deep learning. In *international conference on machine learning* (pp. 1050–1059).

References

Hüllermeier, E., & Waegeman, W. (2021). Aleatoric and epistemic uncertainty in machine learning: An introduction to concepts and methods. *Machine Learning*, *110*(3), 457–506.

Kucukelbir, A., Tran, D., Ranganath, R., Gelman, A., & Blei, D. M. (2016). Automatic differentiation variational inference. *arXiv preprint arXiv:1603.00788*.

Lee, K., & Lee, J. (2023). Asymptotic properties for bayesian neural network in besov space. *Advances in Neural Information Processing Systems*, *36*.

Maddox, W. J., Izmailov, P., Garipov, T., Vetrov, D. P., & Wilson, A. G. (2019). A simple baseline for bayesian uncertainty in deep learning. *Advances in Neural Information Processing Systems*, *32*.

Neal, R. M. (2012). *Bayesian learning for neural networks* (Vol. 118). Springer Science & Business Media.

Phan, B. T. (2019). *Bayesian deep learning and uncertainty in computer vision* (Unpublished master's thesis). University of Waterloo.

Polson, N. G., & Ročková, V. (2018). Posterior concentration for sparse deep learning. *Advances in Neural Information Processing Systems*, *31*.

Schmidt-Hieber, J. (2020). Nonparametric regression using deep neural networks with relu activation function. *The Annals of Statistics*, *48*(4), 1875–1897.

Srivastava, N., Hinton, G., Krizhevsky, A., Sutskever, I., & Salakhutdinov, R. (2014). Dropout: a simple way to prevent neural networks from overfitting. *The journal of machine learning research*, *15*(1), 1929–1958.

References

Suzuki, T. (2019). Adaptivity of deep relu network for learning in besov and mixed smooth besov spaces: optimal rate and curse of dimensionality.

Welling, M., & Teh, Y. W. (2011). Bayesian learning via stochastic gradient langevin dynamics. In *Proceedings of the 28th international conference on machine learning* (*icml-11*) (pp. 681–688).

Yarotsky, D. (2017). Error bounds for approximations with deep relu networks. *Neural Networks*, *94*, 103–114.

이재용, & 이기재. (2022). 베이즈데이터분석. 한국방송통신대학교 출판문화원.

Thank you!