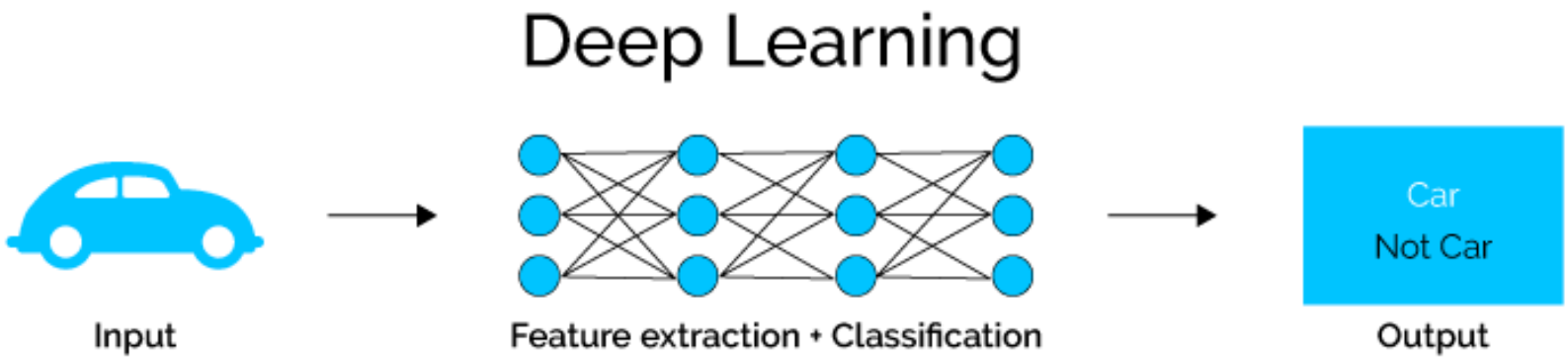
M1399.000400 / M3309.005200 딥러닝의 통계적 이해 (Deep Learning: Statistical Perspective) 2025 2학기 (Fall)

김지수 (Jisu KIM)

2025-09-02

심층학습은 여러 층으로 쌓인 구조로 되어 있는 모형을 학습합니다.

- 심층학습은 폭발적인 발전에 비해 통계적 이론 분석은 매우 더딘 편입니다.
- 심층학습에 관련된 통계 학습 이론을 배웁니다.



심층학습은 기계학습(Machine Learning)의 일종으로 볼 수 있습니다.

- 기계학습에는 여러 분야들이 있지만, 대표적으로 다음과 같습니다.
- 지도학습(Supervised Learning)
  - 입력변수를 이용하여 출력변수를 예측
  - 회귀(regression), 분류(classification) 등
- 비지도학습(Unsupervised Learning)
  - 입력변수 간의 복잡합 관계를 규명(출력변수가 없음)
  - 군집분석(clustering), 주성분분석(principal component analysis), 요인분석(factor analysis) 등
- 강화학습(reinforcement learning)
  - 행동에 따라 변화하는 환경에서 의사결정 방법을 학습
  - multi-armed bandit problem, markov decision process 등
- 심층학습은 대부분 지도학습(Supervised Learning)과 연관돼 있고, 비지도학습(Unsupervised Learning)에도 사용됩니다.

심층학습의 통계적 이론을 배웁니다.

- 자료(data)는 항상 랜덤한 잡음(noise)을 가지고 있습니다.
- 랜덤한 잡음(noise)이 기계학습 알고리즘에 어떠한 영향을 끼치는지, 알고리즘이 잘 작동하는지 알기 위해서는 통계적 분석이 필요합니다.
- 심층학습에서 통계 이론은 주로 다음의 분석을 해줍니다: 일치성 (consistency), 최적합(optimality) 등

지도학습(Supervised Learning)의 기본 모형

- 입력(Input) / 설명 변수(Covariate) : $x \in \mathbb{R}^d$
- 출력(Output) / 반응 변수(Response): $y \in \mathcal{Y}$
- 모형(Model) : $y \approx f(x)$, $f \in \mathcal{M}$
- 손실 함수(loss function): $l(y, a)$
- 목적: $f^0 = \arg\min_{f \in \mathcal{F}} \mathbb{E}_{(Y,X)} [l(Y, f(X))]$ 을 찾는 것을 목적으로 합니다, 이 때 $\mathcal{F}$는 $\mathcal{M}$과 다를 수 있습니다.
- 예측: $\hat{f}$가 추정한 함수이고, 새로운 자료 x가 들어오면, y를 $\hat{f}(x)$로 추정합니다.

고차원에서 일반화가법모형(Generalized Additive Models)을 사용하는 이유를 알아봅니다.

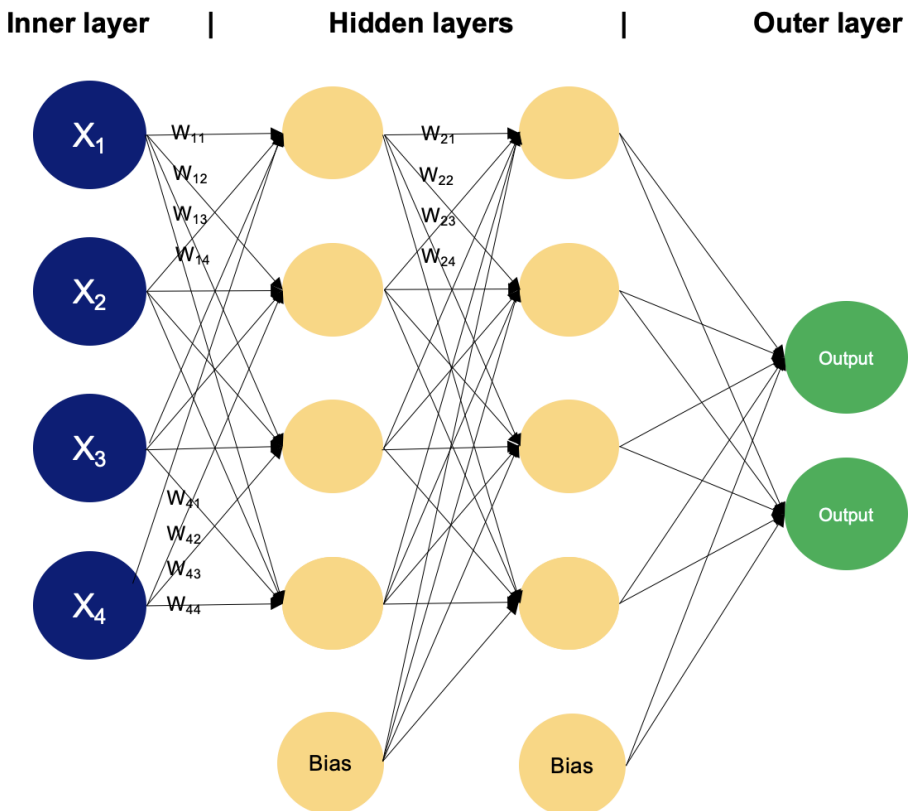
- 저차원에서는 함수를 추정하는 여러 방법들이 있습니다.
- 고차원에서는 차원의 저주(curse of dimensionality)로 인해 잘 작동하지 않습니다.
- 고차원에서 함수를 추정하는 여러 기법들을 알아봅니다: 일반화가법모형(Generalized Additive Models), 사영추적회귀(Projection Pursuit Regression)

Deep learning 의 구조에 대해 알아봅니다.

- Deep Learning Network 중 hidden layer 가 1개인 모형은 다음과 같이 나타낼 수 있습니다:

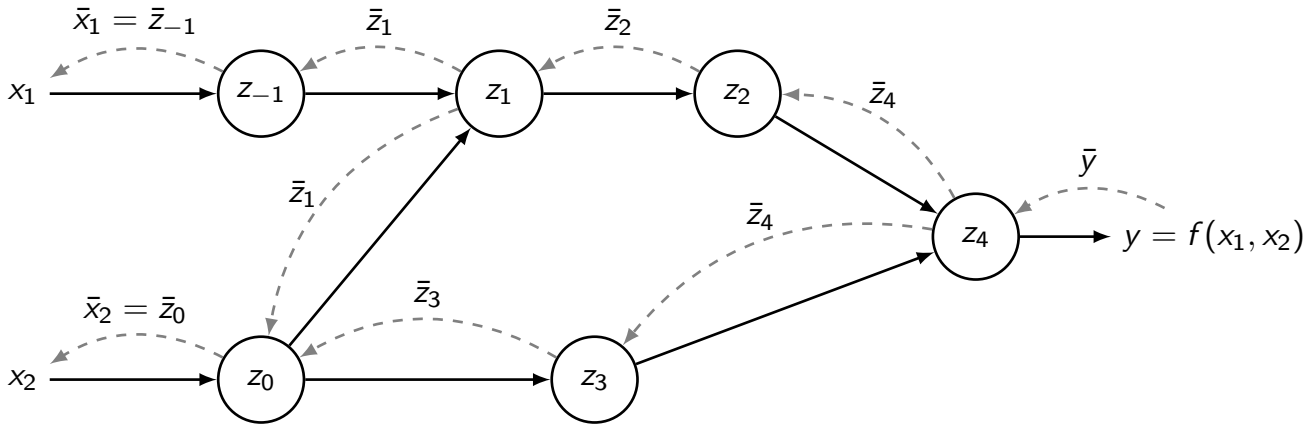
$$f_\theta(x) = g\left(\beta_0 + \sum_{j=1}^{M} \beta_j \sigma(b_j + w_j^\top x)\right).$$

- 더 일반적으로 hidden layer 가 1개 이상 있을 수 있는 모형을 고려할 수 있습니다: 심층학습의 가장 기본적이면서 일반적인 모형입니다.



Backpropagation의 원리에 대해 알아봅니다.

- 매개변수들을 학습할 때 backpropagation을 활용합니다.



Concentration inequality는 확률변수의 행태를 확률적으로 통제합니다.

- 약한 대수의 법칙(law of large number)은, $X_1, \ldots, X_n$이 iid 자료이고 모평균 $\mathbb{E}[X_1] = \mu < \infty$일 때, 표본평균이 모평균으로 확률수렴함을 뜻합니다:

$$\lim_{n\to\infty} P\left(\left|\frac{1}{n}\sum_{i=1}^{n} X_i - \mu\right| > \epsilon\right) = 0 \text{ for all } \epsilon > 0.$$

- 표본평균을 더 일반적인 함수로 확장시켰을 때 확률변수의 행태를 다음과 같이 확률적으로 부등호로 통제하는 것 (및 이에 필요한 기술들)을 Concentration inequality라 합니다:

$$P\left(\sup_{f\in\mathcal{F}} |f(X_1, \ldots, X_n) - \mathbb{E}\left[f(X_1, \ldots, X_n)\right]| > \epsilon\right) < \delta.$$

Concentration inequality는 확률변수의 행태를 확률적으로 통제합니다.

- 표본평균을 더 일반적인 함수로 확장시켰을 때 확률변수의 행태를 다음과 같이 확률적으로 부등호로 통제하는 것 (및 이에 필요한 기술들)을 Concentration inequality라 합니다:

$$P\left(\sup_{f\in\mathcal{F}}|f(X_1,\ldots,X_n)-\mathbb{E}\left[f(X_1,\ldots,X_n)\right]|>\epsilon\right)<\delta.$$

- 다음과 같은 concentration inequality를 배웁니다: Hoeffding's inequality, McDiarmid's inequality, Bernstein's inequality, Rademacher Complexity, VC dimension, uniform bound 등

학습된 함수와 목적함수의 차이를 분해할 수 있습니다.

- 지도학습(Supervised Learning)에서, $y \approx f_*(x)$인 참 회귀함수(true regression function) $f_* \in \mathcal{M}$을 목적함수(target function)로 놓습니다.
- $\mathcal{F}$에서 평균위험(Expected Risk)를 최소화하는 f^0를 평균위험최소함수(expected risk mimizing function)로 놓습니다, 즉

$$f^0 = \arg\min_{f \in \mathcal{F}} \mathbb{E}_{(Y,X)\sim P} \left[\ell(Y, f(X))\right].$$

- $\mathcal{F}$에서 경험위험(Empirical Risk)를 최소화하는 $\hat{f}$를 경험위험최소함수(empirical risk mimizing function)으로 놓습니다, 즉

$$\hat{f} = \arg\min_{f \in \mathcal{F}} \frac{1}{n} \sum_{i=1}^{n} \ell(y_i, f(x_i)).$$

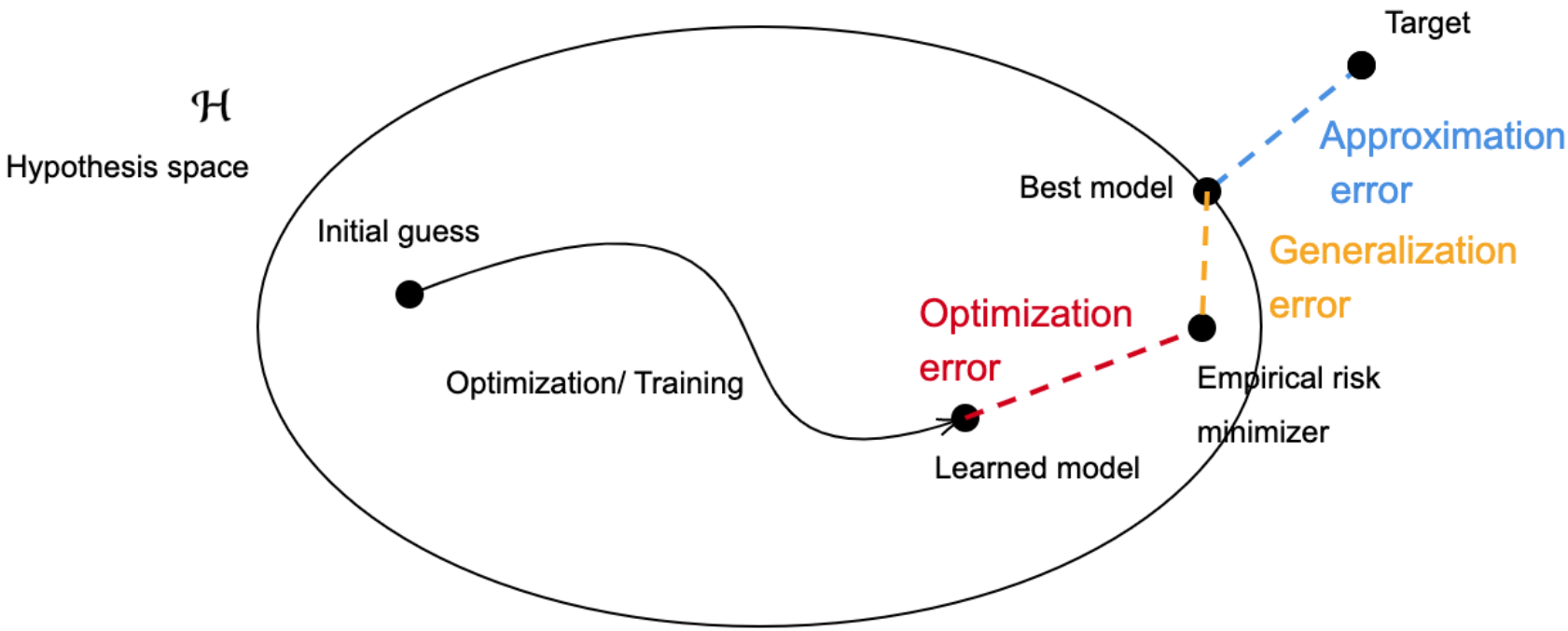
- $\hat{f}$를 최적화 기법(Optimization)으로 근사한 함수 $\tilde{f}$를 학습된 함수(learned function)로 놓습니다.

학습된 함수와 목적함수의 차이를 분해할 수 있습니다.

- 목적함수(target function) f_*와 학습된 함수(learned function) $\tilde{f}$의 차이를 아래와 같이 분해할 수 있습니다:

$$f_* - \hat{f} = \underbrace{f_* - f^0}_{\text{approximation error}} + \underbrace{f^0 - \hat{f}}_{\text{generalization error}} + \underbrace{\hat{f} - \tilde{f}}_{\text{optimization error}} .$$

- 여기선 approximation error $f_* - f^0$와 generalization error $f^0 - \hat{f}$를 다룹니다.



Universal Approximation Theorem

- $f_* - f^0$(approximation error)를 분석합니다.
- 모형 $\mathcal{M}$을 적절히 선택할 경우, f^0가 f_*에 임의로 가깝게 갈 수 있습니다: Universal Approximation Theorem

Approximation Error를 상한하는 방법을 배웁니다.

- $f_* - f^0$(approximation error)를 분석합니다.
- 모형 $\mathcal{M}$을 적절히 선택할 경우, approximation error를 m(심층학습의 폭)으로 상한할 수 있습니다.
- $$\left(f_* - f^0\right)^2 = O\left(\frac{1}{m}\right).$$

심층학습을 깊게 쌓을 때 Approximation Error에서의 이점을 알아봅니다.

- 심층학습에서 깊게 쌓을 때 Approximation error에서의 이점을 알아봅니다.
- 특정 함수는 $2K^2 + 4$ 개의 layer 와 $3K^2 + 6$ 개의 노드를 가진 f_*로 표현할 수 있으나, layer 가 K개 이하인 f^0로 근사할 때에는 노드를 2^K 개 사용하더라도 근사가 제대로 안 됩니다:

$$\int_0^1 \left| f_*(x) - f^0(x) \right| dx \geq \frac{1}{32}.$$

Generalization Error를 상한하는 방법을 배웁니다.

- $f^0 - \hat{f}$(generalization error)를 분석합니다.
- 학습에 사용하는 모형 $\mathcal{F}$의 Rademacher complexity 를 n(자료의 개수)으로 상한할 수 있습니다:

$$\mathrm{Rad}_n(\mathcal{F}) = O\left(\sqrt{\frac{1}{n}}\right).$$

- 이를 이용하여 generalization error $f^0 - \hat{f}$를 n(자료의 개수)로 상한할 수 있습니다.
-

$$\mathbb{E}\left[\left(f^0(x) - \hat{f}(x)\right)^2\right] = O\left(\sqrt{\frac{1}{n}}\right).$$

심층학습을 깊게 쌓을 때 Generalization Error의 이점을 알아봅니다.

- $f^0 - \hat{f}$(generalization error)를 분석합니다.
- 깊게 쌓을 경우, 학습에 사용하는 모형 $\mathcal{F}$의 Rademacher complexity를 n(자료의 개수)으로 상한할 수 있습니다:

$$\mathrm{Rad}_n(\mathcal{F}) = O\left(\sqrt{\frac{1}{n}}\right).$$

- 이를 이용하여 generalization error $f^0 - \hat{f}$를 n(자료의 개수)으로 상한할 수 있습니다.

$$\mathbb{E}\left[\left(f^0(x) - \hat{f}(x)\right)^2\right] = O\left(\sqrt{\frac{1}{n}}\right).$$

얕은 신경학습(shallow neural network)을 분석합니다.

- 얕은 신경학습망(shallow neural network)을 분석합니다: 모형 $\mathcal{F}$가 hidden neuron이 m개인 다음의 f_W들을 모아놓은 집합으로 놓습니다:

$$f_W(x) = \frac{\alpha}{m} \sum_{r=1}^{m} a_r \sigma(\omega_r^\top x).$$

- Neural Tangent Kernel: $\alpha = \sqrt{m}$.
- Mean Field: $\alpha = 1$.

Neural Tangent Kernel

- 모형 매개변수(model parameter) W에 대해 테일러 전개를 해서 분석합니다:

$$f_{W_{(k)}}(x) = f_{W_{(\mathbf{0})}}(x) + \left\langle \nabla f_{W_{(\mathbf{0})}}(x), W_{(k)} - W_{(0)} \right\rangle + o\left(\left\| W_{(k)} - W_{(0)} \right\|_F^2 \right).$$

- $K^m(x_i, x_j) := \langle \nabla f_{W_\mathbf{0}}(x_i), \nabla f_{W_\mathbf{0}}(x_j) \rangle$이 kernel regression의 kernel 같은 역할을 합니다.
- $K^\infty := \lim_{m \to \infty} K^m$라 할 때,

$$\mathbb{E}\left[(f_{W_k}(x) - y)^2 \right] \leq O\left(\sqrt{\frac{y^\top (K^\infty)^{-1} y}{m}} \right).$$

Mean Field

- Mean Field라는 이름은 수리물리(mathematical physics)의 mean field model에서 비롯됐습니다: 매개변수 $\theta_r := (a_r, \omega_r)$를 입자처럼 생각합니다.
- 입자가 랜덤하게 주어진다고 생각하고 다음의 수렴을 이용합니다: $\sigma_*(x;\theta_r) := a_r\sigma(\omega_r^\top x)$ 라 놓으면,

$$\frac{1}{m}\sum_{r=1}^{m}\sigma_*(x;\theta_r) \overset{m\to\infty}{\to} f(x;\rho) = \int \sigma_*(x;\theta)\rho(d\theta).$$

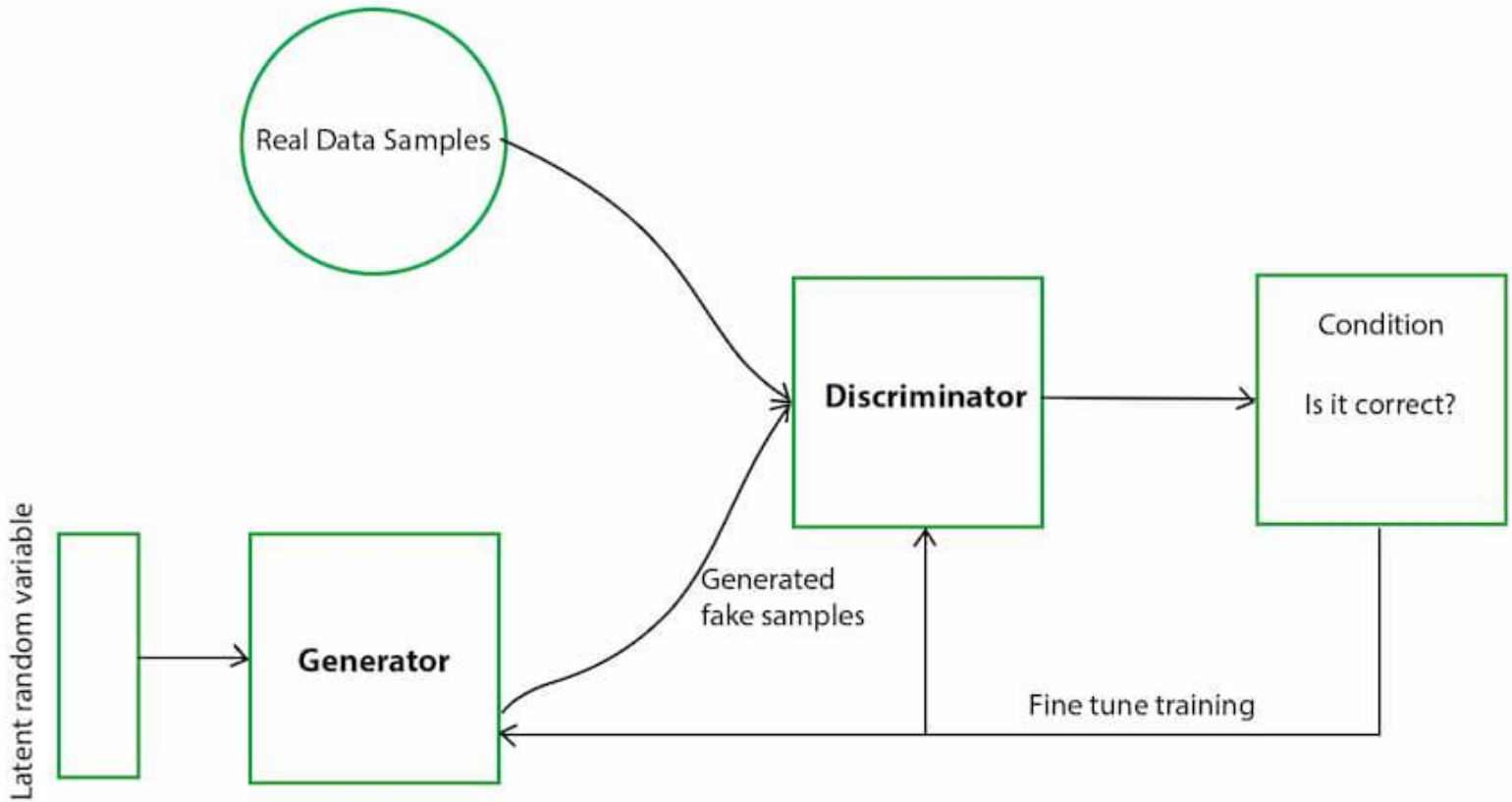
- $\mathcal{R}(f) := \mathbb{E}\left[(y - f(x))^2\right]$라고 할 때,

$$\left|\mathcal{R}(\hat{f}) - \mathcal{R}(\tilde{f})\right| = O\left(\sqrt{\frac{1}{m}}\right).$$

심층학습을 이용한 생성 모형에 관한 통계적 이론을 알아봅니다.

- 심층학습(deep learning)을 이용한 여러 생성 모형들이 있습니다: Generative Adversarial Network, Diffusion model, Large Language Model 등
- Generative Adversarial Network와 Diffusion model은 원본 자료 $\{x_1, \ldots, x_n\}$와 비슷한 자료 $\{y_1, \ldots, y_m\}$을 만들어내고, Large Language Model은 $\{x_1, f(x_1), \ldots, x_n, f(x_n)\}$으로 학습해서 x_{query}가 주어졌을 때 $f(x_{query})$를 생성합니다.
- 생성된 자료가 얼마나 좋은지에 대한 통계적인 이론을 알아봅니다.

Generative Adversarial Network



[3]

[3] https://www.geeksforgeeks.org/generative-adversarial-network-gan/

Diffusion Model

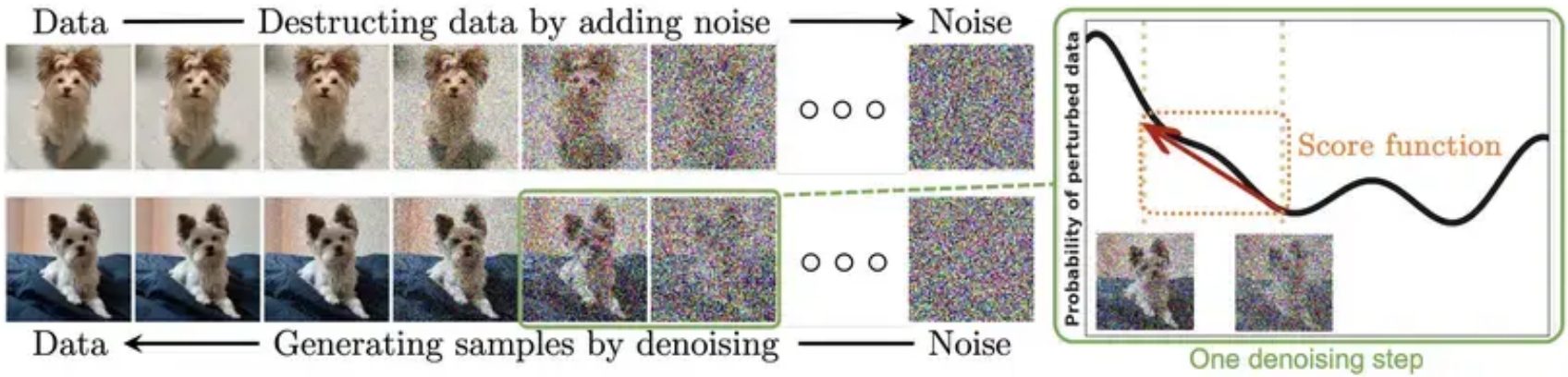


Fig. 2. Diffusion models smoothly perturb data by adding noise, then reverse this process to generate new data from noise. Each denoising step in the reverse process typically requires estimating the score function (see the illustrative figure on the right), which is a gradient pointing to the directions of data with higher likelihood and less noise.

4

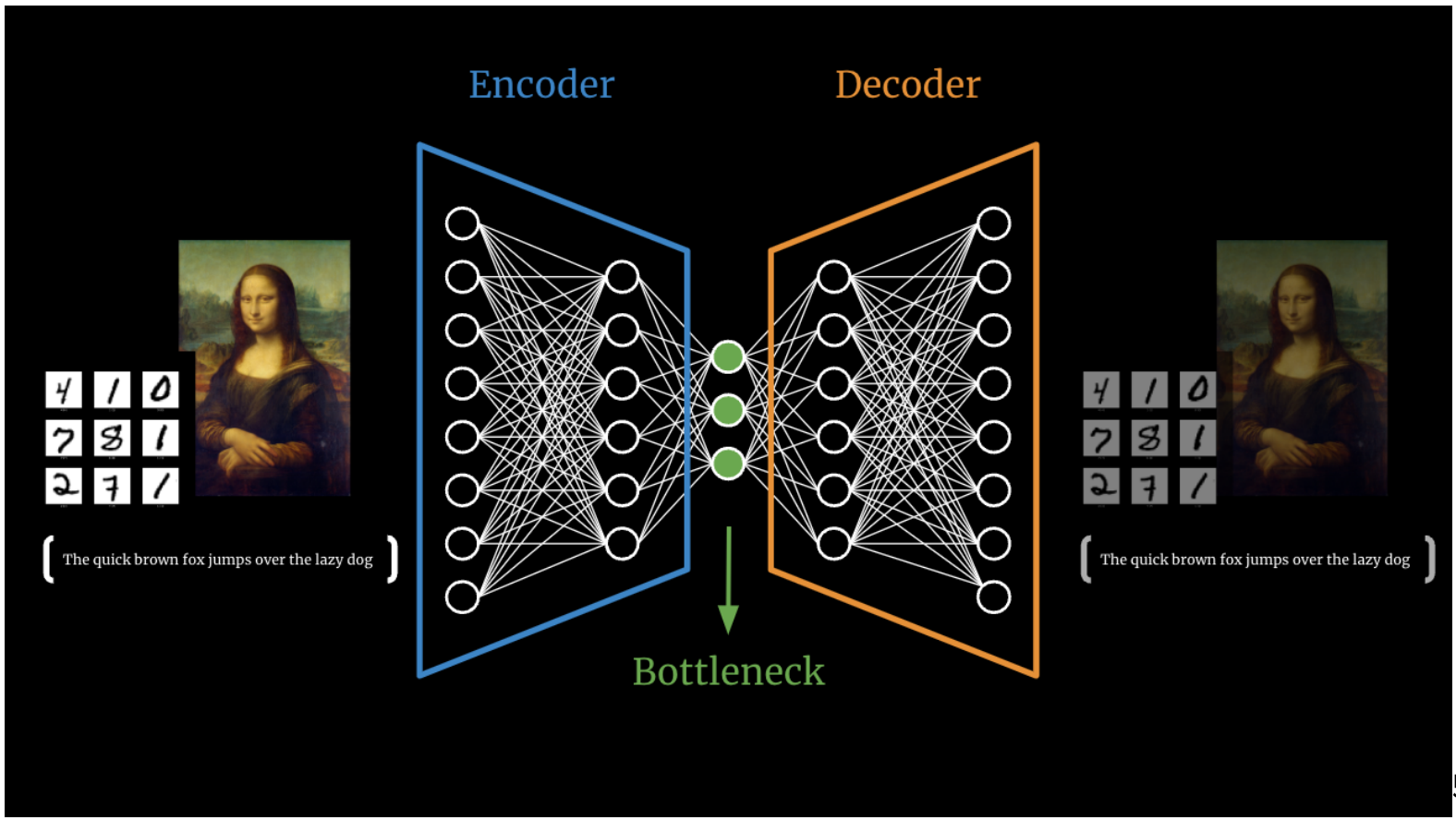
[4] https://www.superannotate.com/blog/diffusion-models

비지도학습(Unsupervised Learning)은 입력변수 간의 관계를 규명합니다.

- 비지도학습(Unsupervised Learning)은 입력변수 간의 복잡합 관계를 규명하며, 출력변수가 없습니다.
- 대표적으로 군집분석(clustering), 주성분분석(principal component analysis), 요인분석(factor analysis) 등이 있습니다.

심층학습을 이용한 비지도학습(Unsupervised Learning)의 통계적 이론을 알아봅니다.

- 심층학습을 활용한 비지도학습의 대표적인 예로 Autoencoder 가 있습니다.
- 관련된 통계적 이론을 알아봅니다.



[5]

[5] https://www.saberhq.com/blog/autoencoders

기하와 위상이 정규조건으로 어떻게 사용되는지 알아봅니다.

- 기하 및 위상 조건이 심층학습의 학습에 어떤 영향을 끼치는지 알아봅니다.

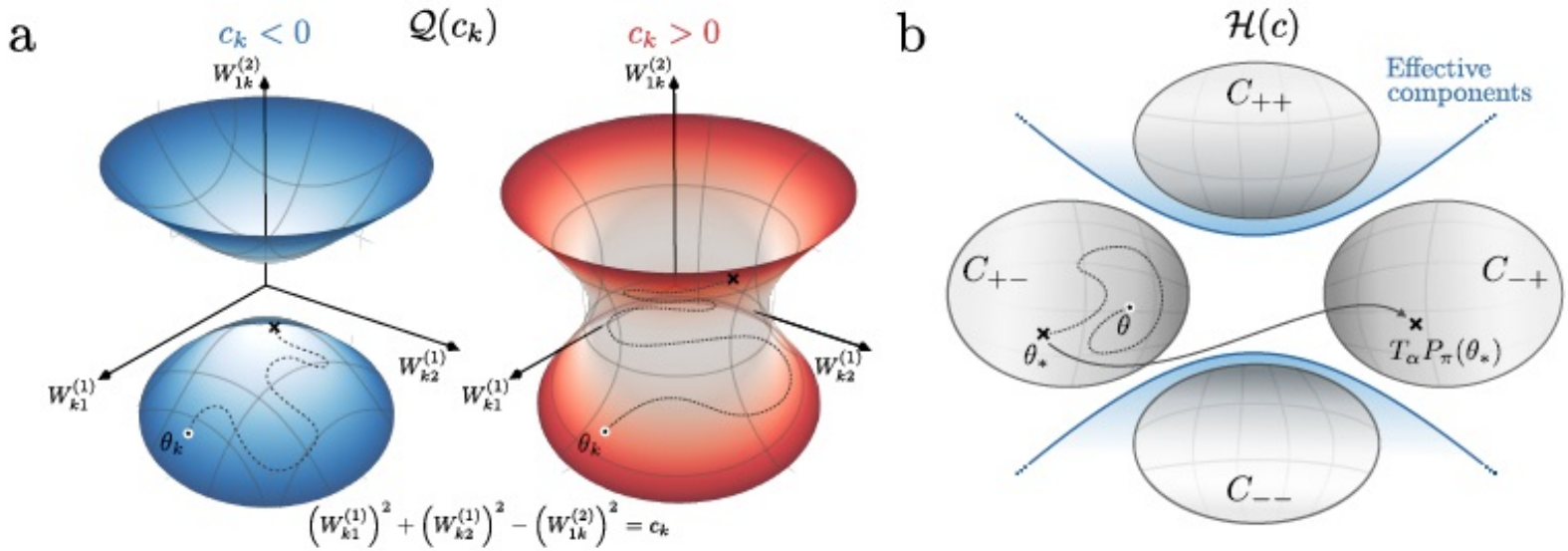


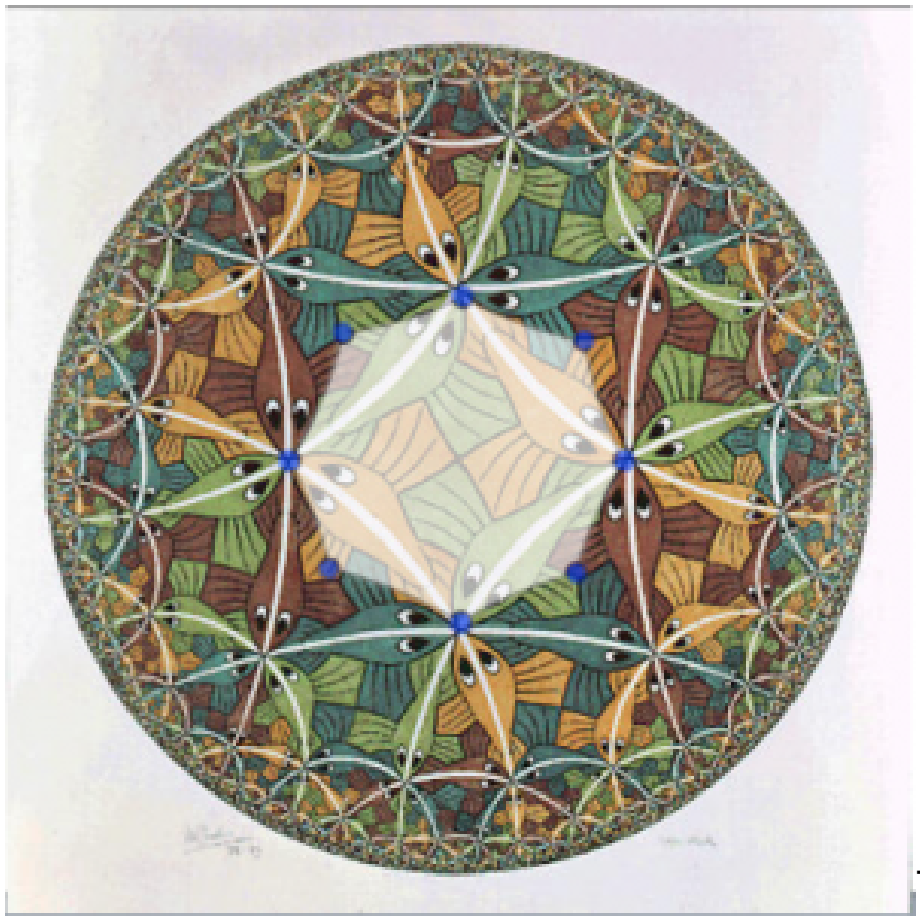
Figure 2: **a.** The invariant hyperquadric $\mathcal{Q}(c_k)$ of a neuron with two inputs ($d = 2$) and one output ($e = 1$) in the cases where $c_k < 0$ (left) and $c_k > 0$ (right). **b.** Depiction of the invariant set $\mathcal{H}(c)$ in the case where $l_- = 2$ so that there are $2^{l_-} = 4$ connected components. $C_{\pm\mp}$ denotes the connected component such that $s = (\pm 1, \mp 1)$. The blue lines separate the different effective components of $\mathcal{H}(c)$.

6

6

non-Euclidean space 가 deep learning에 어떻게 응용되는지 알아봅니다.

- non-Euclidean space 중 hyperbolic space 가 deep learning 에 어떻게 응용되는지 알아봅니다.

 [7]

[7] https://web.colby.edu/thegeometricviewpoint/2016/12/21/tessellations-of-the-hyperbolic-plane-and-m-c-escher/