

Review of Probability Theory

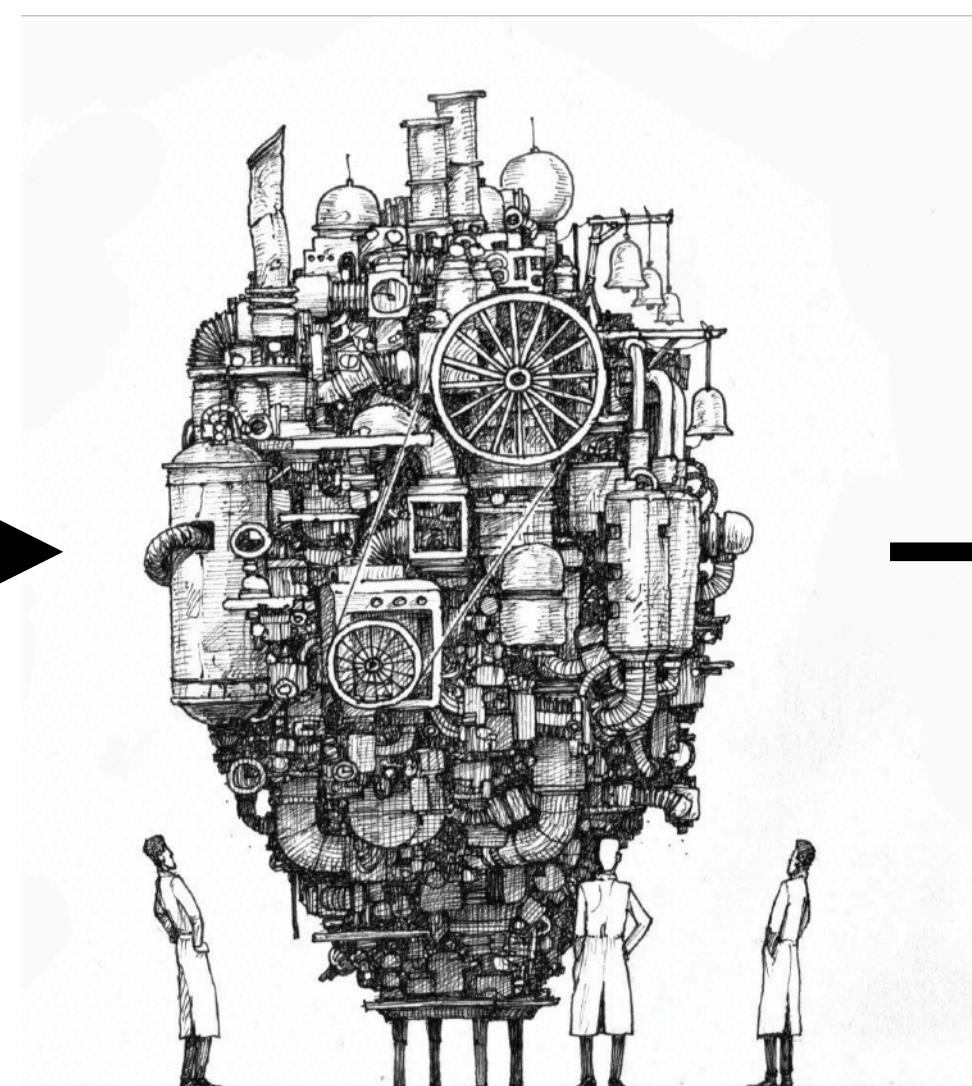
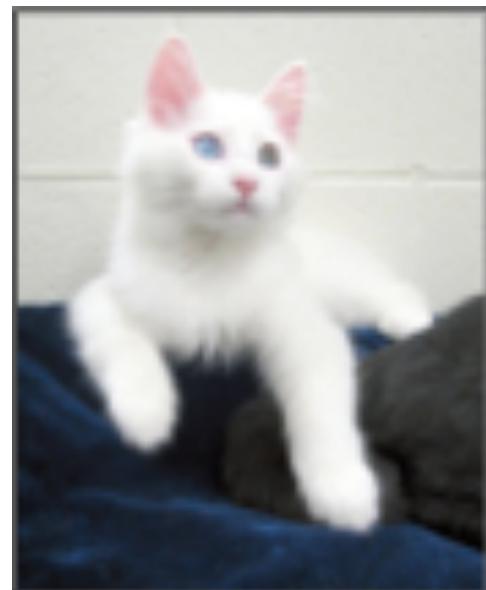
2020

Soochul Park

Random Variables & Probability Distributions

How to Represent Probability Distributions

$$p(y|x)$$



“Cat”

$$p(x|y)$$

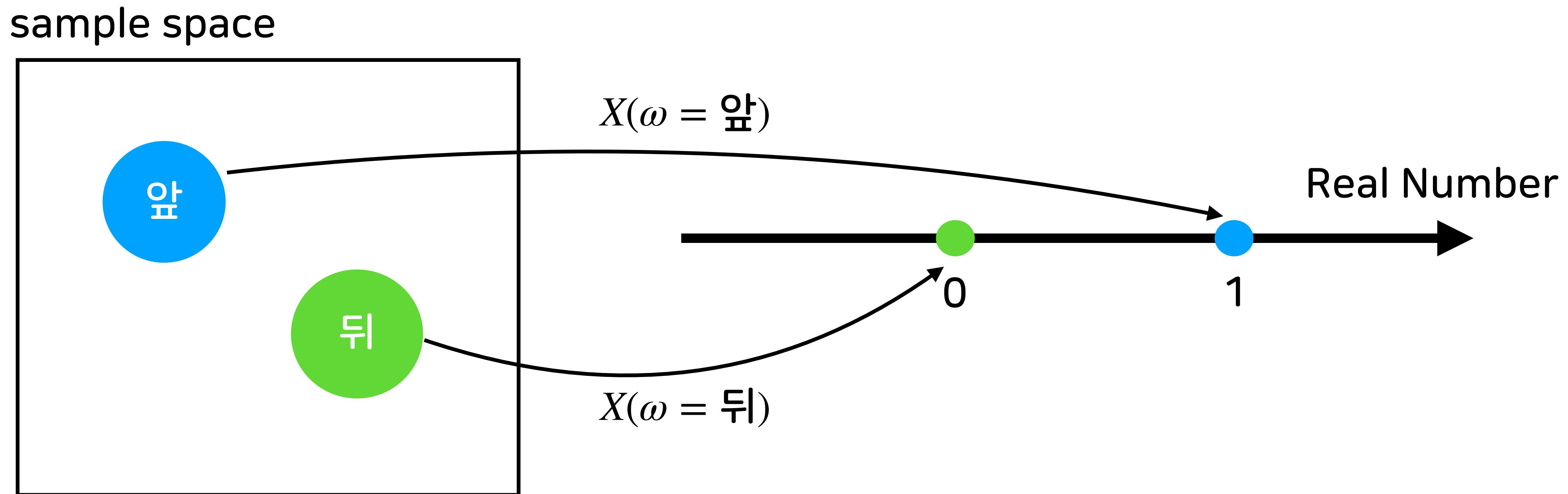


Discriminative Models

Generative Models

Random Variable

- 확률 변수 (Random Variable)는 무작위적으로 다른 값을 가질 수 있는 변수를 나타냅니다.
- 좀 더 엄밀하게는 sample space 내의 예측할 수 없는 각 사건들을 실수값에 대응시키는 함수로 생각할 수 있습니다.



Probability Distribution

- 확률 분포 (probability distribution)는 random variable이 가질 수 있는 값들의 가능성을 나타냅니다.
- Probability distribution은 discrete random variable에 대한 Probability Mass Function (PMF)와 continuous random variable에 대한 Probability Density Function(PDF) 두 종류가 있습니다.

Bernoulli-정의

- 베르누이 분포 (Bernoulli distribution)은 0 또는 1 두가지 값을 가지는 random variable의 확률 분포입니다.
- Random variable이 1 값을 가질 확률을 나타내는 μ 를 parameter로 가집니다.
- Bernoulli distribution의 PMF는 다음과 같은 식으로 표현됩니다.
- $$p(x|\mu) = Ber(x|\mu) = \begin{cases} \mu, & \text{if } x = 1 \\ 1 - \mu, & \text{if } x = 0 \end{cases}$$
- 또는 다음과 같이 표현 할 수도 있습니다.
- $$p(x|\mu) = Ber(x|\mu) = \mu^x(1 - \mu)^{(1-x)}$$

Bernoulli-최적화

- Dataset $X = \{x_1, x_2, \dots, x_N\}$ 이 주어진 경우, likelihood function을 다음과 같이 표현할 수 있습니다.

$$p(X|\mu) = \prod_{n=1}^N p(x_n|\mu) = \prod_{n=1}^N \mu^{x_n}(1-\mu)^{(1-x_n)}$$

- 빈도주의적 관점에서 위 식을 최대화 하는 parameter인 μ 를 구할 수 있습니다. 또는 단조 증가 함수인 log함수를 likelihood에 적용하여 최대화 할 수도 있습니다.

$$\log p(X|\mu) = \sum_{n=1}^N \log p(x_n|\mu) = \sum_{n=1}^N \{x_n \log \mu + (1-x_n) \log (1-\mu)\}$$

Bernoulli-최적화

- Data의 likelihood 또는 log likelihood를 최대화하는 최적화 방법을 Maximum Likelihood라고 합니다.
- Maximum Likelihood를 통해 얻은 파라메터 μ 의 값은 다음과 같습니다.
- $$\mu_{ML} = \frac{1}{N} \sum_{n=1}^N x_n$$
- 이는 전체 데이터 중 1값을 갖는 데이터의 비율과 같습니다.

Binomial-정의

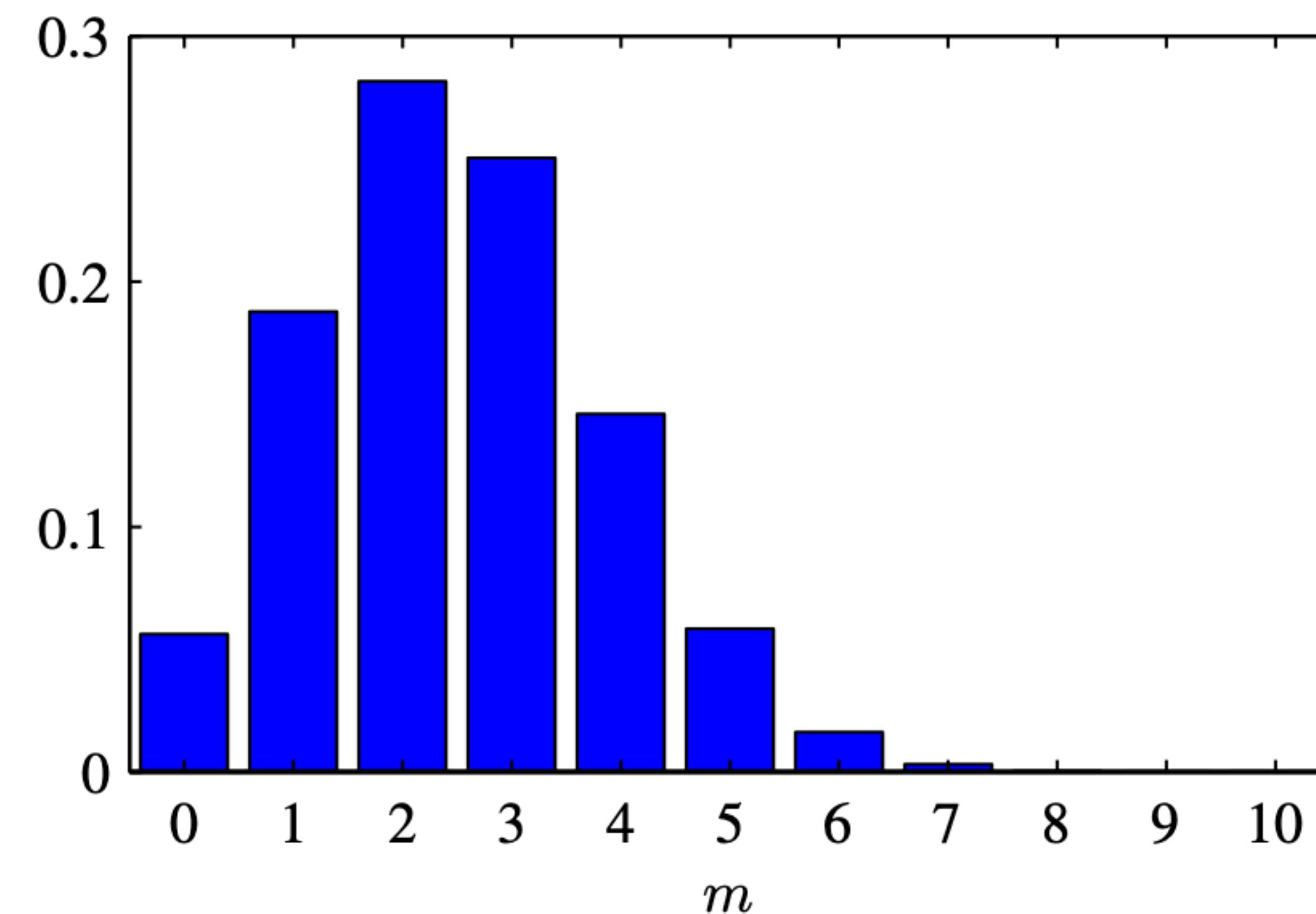
- Binomial distribution은 베르누이 시행(Bernoulli trials)을 N 번 독립적으로 했을 때 얻을 수 있는 1의 갯수에 대한 확률 분포입니다.
- 총 시행 횟수 N 과 1 값이 발생할 확률 μ 를 파라메터로 갖습니다.
- Binomial distribution의 PMF는 다음과 같이 정의할 수 있습니다.

$$Bin(m | N, \mu) = \binom{N}{m} \mu^m (1 - \mu)^{N-m}$$

$$\text{where } \binom{N}{m} \equiv \frac{N!}{(N - m)!m!}$$

Binomial

Figure 2.1 Histogram plot of the binomial distribution (2.9) as a function of m for $N = 10$ and $\mu = 0.25$.



Categorical-정의

- Categorical distribution은 K 개의 discrete 값을 가질 수 있는 random variable에 대한 probability distribution입니다.
- Parameter로 각 카테고리에 대한 확률 값을 나타내는 벡터 $\mu = (\mu_1, \mu_2, \dots, \mu_K)$ 를 갖습니다.
- Random variable이 갖는 값은 one-hot 벡터로 나타낼 수 있습니다. 예를 들어 6개의 카테고리가 있고 random variable이 3번째 카테고리 값을 갖는다면 다음과 같이 나타낼 수 있습니다.

$$\mathbf{x} = (0, 0, 1, 0, 0, 0)^T$$

- Categorical distribution의 PMF는 다음과 같이 표현할 수 있습니다.

$$p(\mathbf{x} | \mu) = \prod_{k=1}^K \mu_k^{x_k}$$

Categorical-최적화

- Dataset $\mathbf{X} = \{\mathbf{x}_1, \mathbf{x}_2, \dots, \mathbf{x}_N\}$ 이 주어진 경우, likelihood function을 다음과 같이 표현할 수 있습니다.

$$p(\mathbf{X} | \boldsymbol{\mu}) = \prod_{n=1}^N \prod_{k=1}^K \mu_k^{x_{nk}} = \prod_{k=1}^K \mu_k^{(\sum_n x_{nk})} = \prod_{k=1}^K \mu_k^{N_k}$$

- 빈도주의적 관점에서 위 식을 최대화 하는 파라메터인 $\boldsymbol{\mu}$ 를 구할 수 있습니다. 또는 단조 증가 함수인 log함수를 likelihood에 적용하여 최대화 할 수도 있습니다.

$$\log p(\mathbf{X} | \boldsymbol{\mu}) = \sum_{k=1}^K N_k \log \mu_k$$

Categorical-최적화

- 단, μ_k 값들의 합이 1이 되어야 하는 조건을 걸기 위해, Lagrange multiplier를 이용하여 다음 식을 최대화할 수 있습니다.



$$\log p(\mathbf{X} | \boldsymbol{\mu}) = \sum_{k=1}^K N_k \log \mu_k + \lambda \left(\sum_{k=1}^K \mu_k - 1 \right)$$

- 위 식을 최대화하는 μ_k 는 다음과 같습니다.



$$\mu_k = \frac{N_k}{N}$$

- 이는 전체 데이터 중 k번째 카테고리를 갖는 데이터의 비율과 같습니다.

Categorical-최적화

- 증명

$$\frac{\partial \log p(\mathbf{X} | \boldsymbol{\mu})}{\partial \mu_k} = \frac{N_k}{\mu_k} + \lambda, \frac{\partial \log p(\mathbf{X} | \boldsymbol{\mu})}{\partial \lambda} = \sum_{k=1}^K \mu_k - 1$$

- 위 두 편미분 값을 0으로 두면,

$$\frac{N_k}{\mu_k} + \lambda = 0, \sum_{k=1}^K \mu_k - 1 = 0$$

- 위 두식을 정리하면,

$$\lambda = -N, \mu_k = \frac{N_k}{N}$$

Multinomial-정의

- Multinomial distribution은 K 개의 다른 카테고리를 가질 수 있는 random variable에서 N 번 독립적으로 값을 얻었을 때, 각 카테고리가 N_k 번씩 선택될 확률에 대한 분포입니다.
- Parameter로 각 카테고리에 대한 확률 값 $\mu = (\mu_1, \mu_2, \dots, \mu_K)$ 과 시행 횟수 N 을 갖습니다.
- Multinomial distribution의 PMF는 다음과 같이 표현할 수 있습니다.

$$Mult(N_1, N_2, \dots, N_K | \mu, N) = \binom{N}{N_1 N_2 \cdots N_K} \prod_{k=1}^K \mu_k^{N_k}$$

$$\text{where } \binom{N}{N_1 N_2 \cdots N_K} \equiv \frac{N!}{N_1! N_2! \cdots N_K!}$$

Gaussian-정의

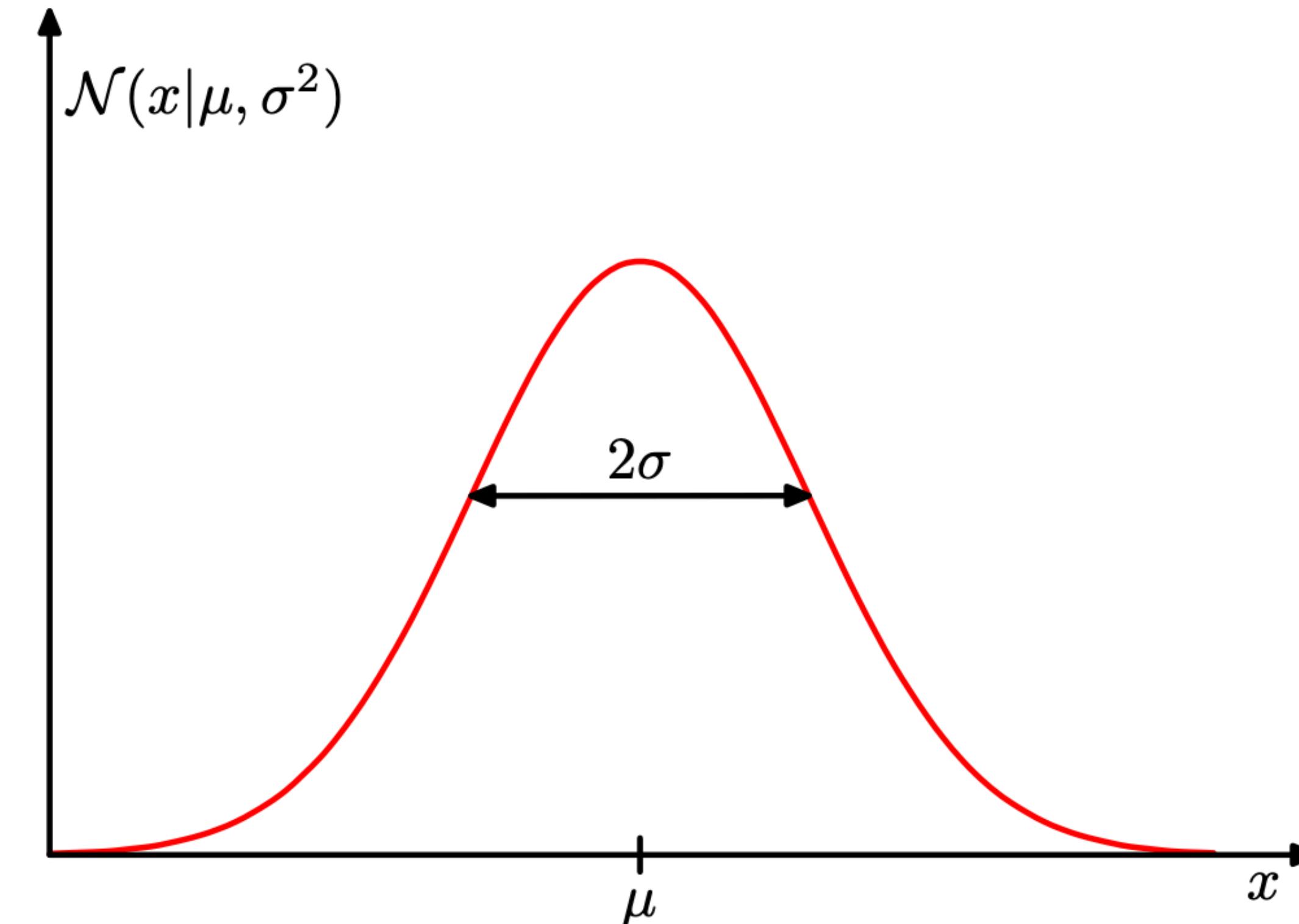
- Gaussian Distribution의 PDF(Probability Distribution Function)은 다음과 같습니다.

$$\mathcal{N}(x | \mu, \sigma^2) = \frac{1}{(2\pi\sigma^2)^{1/2}} \exp \left\{ -\frac{1}{2\sigma^2}(x - \mu)^2 \right\}$$

- Parameter로 mean값을 나타내는 μ 와, 분산을 나타내는 σ^2 을 갖습니다.

Gaussian

Figure 1.13 Plot of the univariate Gaussian showing the mean μ and the standard deviation σ .



Gaussian-최적화

- Gaussian Distribution에서 독립적으로 샘플링한 데이터셋 $X = \{x_1, \dots, x_N\}^T$ 에 대해 log likelihood function은 다음과 같이 쓸 수 있습니다.

$$\ln p(X | \mu, \sigma^2) = -\frac{N}{2} \ln 2\pi - \frac{N}{2} \ln \sigma^2 - \frac{1}{2\sigma^2} \sum_{n=1}^N (x_n - \mu)^2$$

- Frequentist 관점에서 위 식을 최대화하는 μ 와 Σ 는 다음과 같습니다.

$$\begin{aligned}\mu_{\text{ML}} &= \frac{1}{N} \sum_{n=1}^N x_n \\ \sigma_{\text{ML}}^2 &= \frac{1}{N} \sum_{n=1}^N (x_n - \mu_{\text{ML}})^2\end{aligned}$$

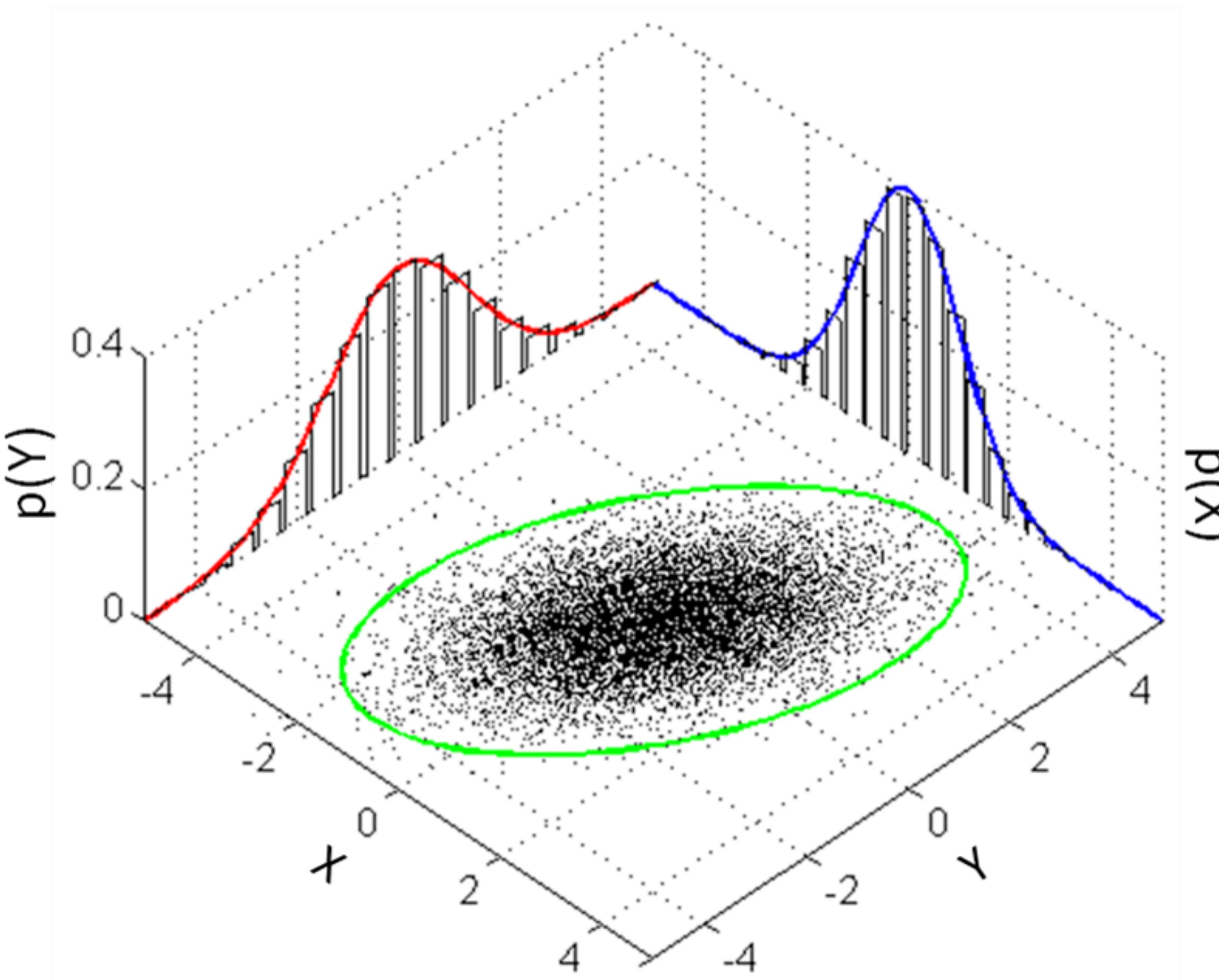
Multivariate Gaussian-정의

- 데이터를 나타내는 \mathbf{x} 가 D 차원일 벡터일 때, Multivariate Gaussian Distribution의 PDF(Probability Distribution Function)은 다음과 같습니다.

$$\mathcal{N}(\mathbf{x} | \boldsymbol{\mu}, \boldsymbol{\Sigma}) = \frac{1}{(2\pi)^{D/2}} \frac{1}{|\boldsymbol{\Sigma}|^{1/2}} \exp \left\{ -\frac{1}{2} (\mathbf{x} - \boldsymbol{\mu})^T \boldsymbol{\Sigma}^{-1} (\mathbf{x} - \boldsymbol{\mu}) \right\}$$

- Parameter로 mean값을 나타내는 $\boldsymbol{\mu}$ 와, covariance matrix인 $\boldsymbol{\Sigma}$ 을 갖습니다.

Bivariate Gaussian distribution



Multivariate Gaussian-최적화

- Multivariate Gaussian Distribution에서 독립적으로 샘플링한 데이터셋 $\mathbf{X} = \{\mathbf{x}_1, \dots, \mathbf{x}_N\}^T$ 에 대해 log likelihood function은 다음과 같이 쓸 수 있습니다.

$$\ln p(\mathbf{X} | \boldsymbol{\mu}, \boldsymbol{\Sigma}) = -\frac{ND}{2} \ln(2\pi) - \frac{N}{2} \ln |\boldsymbol{\Sigma}| - \frac{1}{2} \sum_{n=1}^N (\mathbf{x}_n - \boldsymbol{\mu})^T \boldsymbol{\Sigma}^{-1} (\mathbf{x}_n - \boldsymbol{\mu})$$

- Frequentist 관점에서 위 식을 최대화하는 $\boldsymbol{\mu}$ 와 $\boldsymbol{\Sigma}$ 는 다음과 같습니다.

$$\boldsymbol{\mu}_{\text{ML}} = \frac{1}{N} \sum_{n=1}^N \mathbf{x}_n$$
$$\boldsymbol{\Sigma}_{\text{ML}} = \frac{1}{N} \sum_{n=1}^N (\mathbf{x}_n - \boldsymbol{\mu}_{\text{ML}}) (\mathbf{x}_n - \boldsymbol{\mu}_{\text{ML}})^T$$

Probability Theory

Joint & Marginal Probability Distribution

$P(x = x, y = y)$	y_1	y_2	y_3	$P(x = x)$
x_1	3/20	5/20	4/20	12/20
x_2	2/20	3/20	3/20	8/20
$P(y = y)$	5/20	8/20	7/20	20/20

Joint prob. distribution

$$P(x = x, y = y)$$

Marginal probability distribution

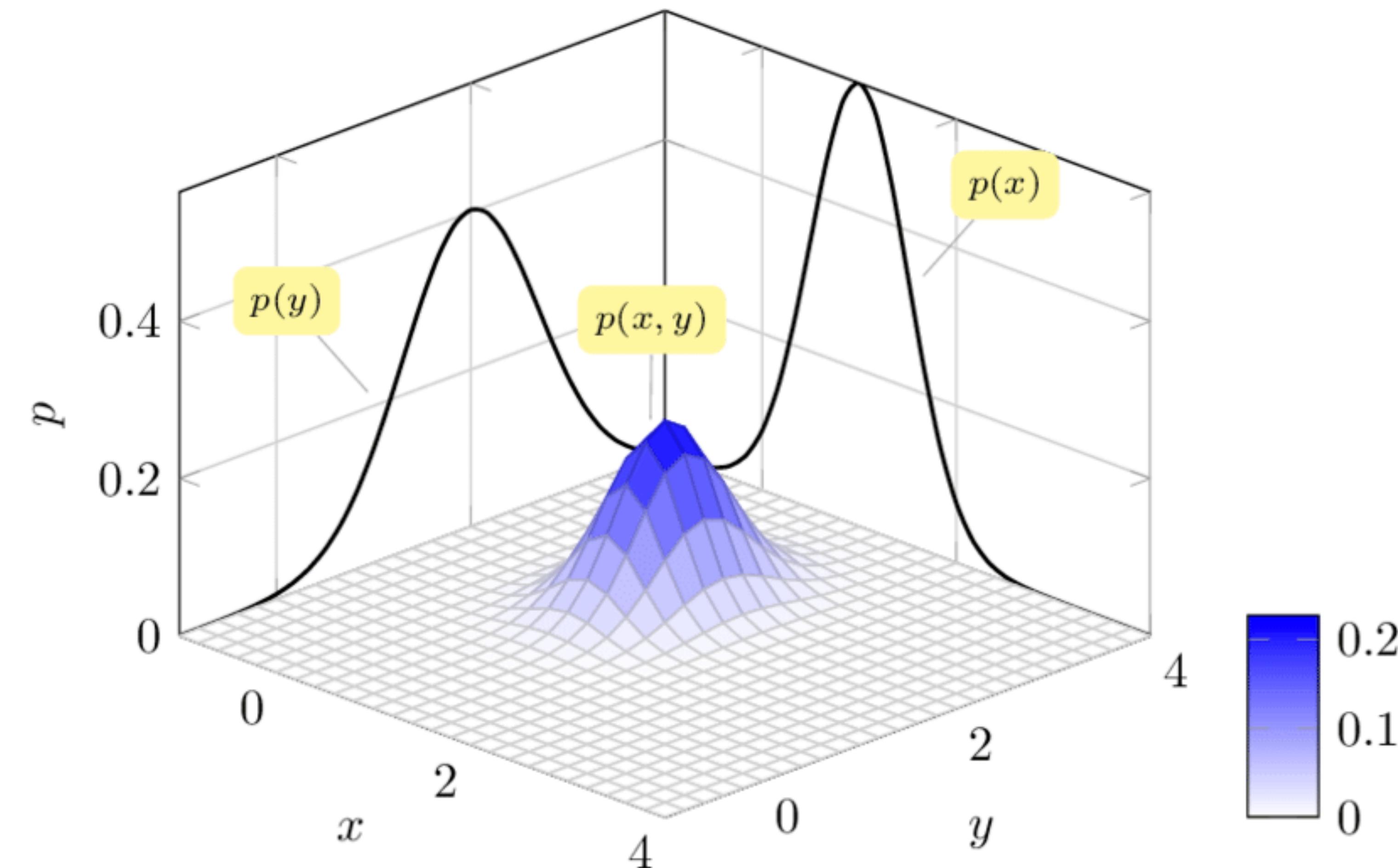
$$\forall x \in X, P(x = x) = \sum_y P(x = x, y = y)$$

Continuous case

$$p(x) = \int p(x, y) dy$$

$$\forall y \in Y, P(y = y) = \sum_x P(x = x, y = y)$$

Joint & Marginal Probability Distribution



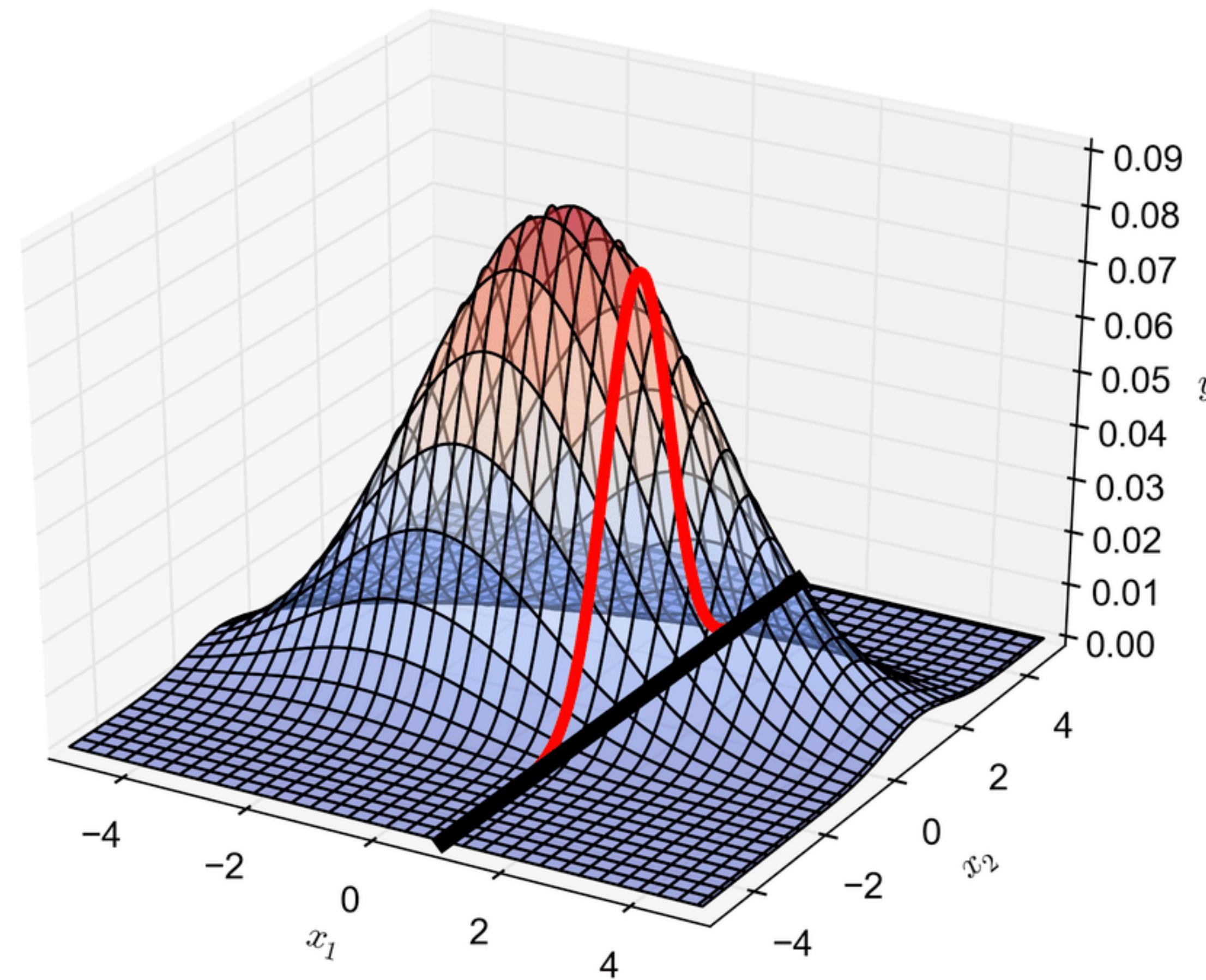
Conditional Probability Distribution

$P(x = x, y = y)$	y_1	y_2	y_3	$P(x = x)$
x_1	3/20	5/20	4/20	12/20
x_2	2/20	3/20	3/20	8/20
$P(y = y)$	5/20	8/20	7/20	20/20

Conditional probability

$$P(y = y|x = x) = \frac{P(y = y, x = x)}{P(x = x)}, \quad \text{when } P(x = x) > 0$$

Conditional Probability Distribution



Expectation

- Discrete probability distribution $p(x)$ 에 대한 function $f(x)$ 의 기댓값(expectation)은 다음과 같이 계산됩니다.



$$\mathbb{E}[f] = \sum_x p(x)f(x)$$

- 또는 $p(x)$ 가 continuous인 경우 다음과 같습니다.



$$\mathbb{E}[f] = \int p(x)f(x)dx$$

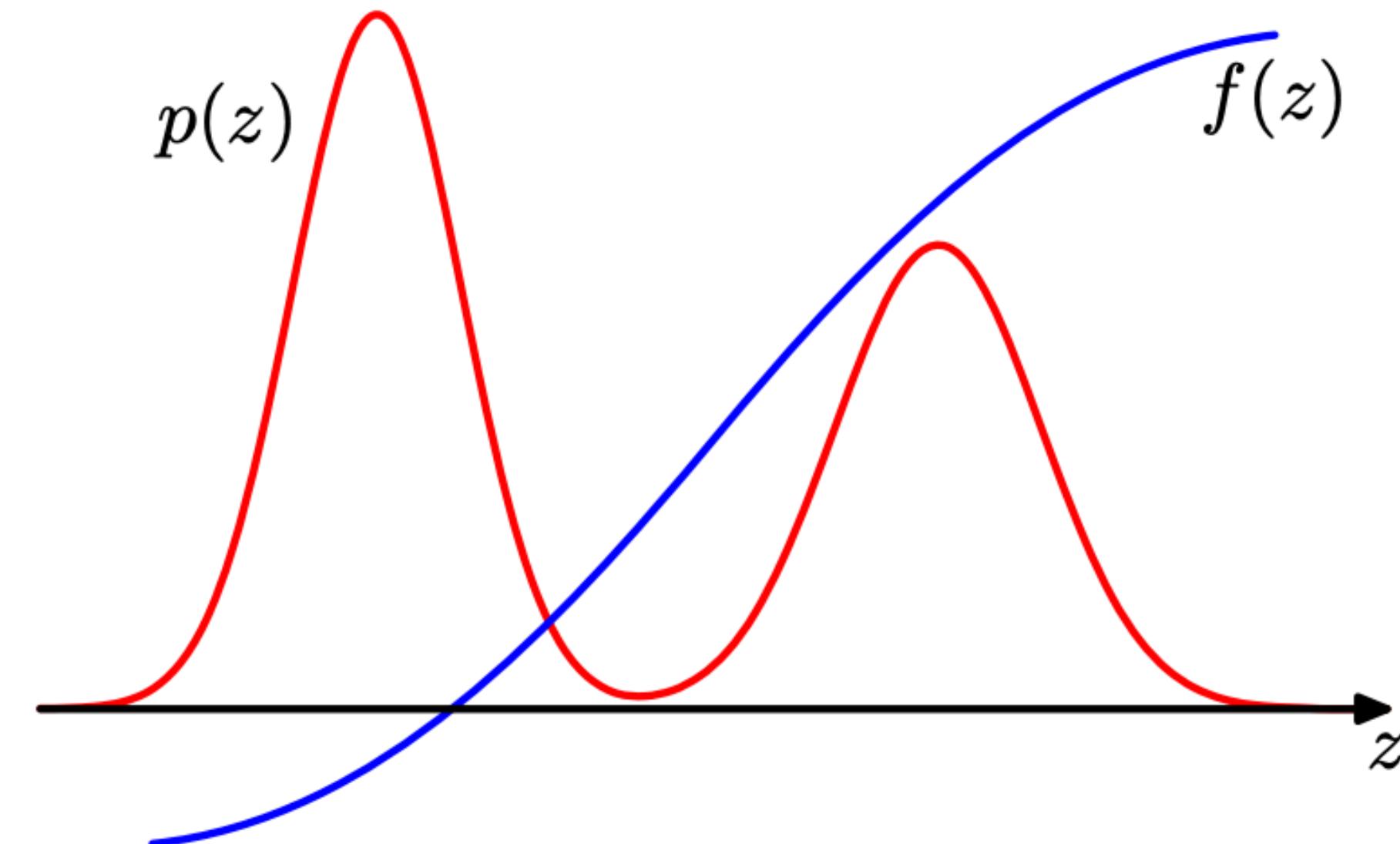
- Expectation은 다음과 같이 sampling을 통한 근사(approximation)으로 계산할 수도 있습니다.



$$\mathbb{E}[f] \simeq \frac{1}{N} \sum_{n=1}^N f(x_n)$$

Expectation

Figure 11.1 Schematic illustration of a function $f(z)$ whose expectation is to be evaluated with respect to a distribution $p(z)$.



Likelihood

- Likelihood function: **function of the parameter** of a statistical model given data
 - **Probability** [$P_\theta(x = x)$]: to describe plausibility of a future outcome, given the parameter
 - **Likelihood** [$\mathcal{L}(\theta; x) = P_\theta(x = x)$]: to describe plausibility of a parameter value
 - cf. continuous case: likelihood function is equivalent to density function

Maximum Log-Likelihood

- Approximate the expected log-likelihood

$$\mathbb{E}_{x \sim P_{\text{data}}} [\log P_{\theta}(x)]$$

- with the empirical log-likelihood

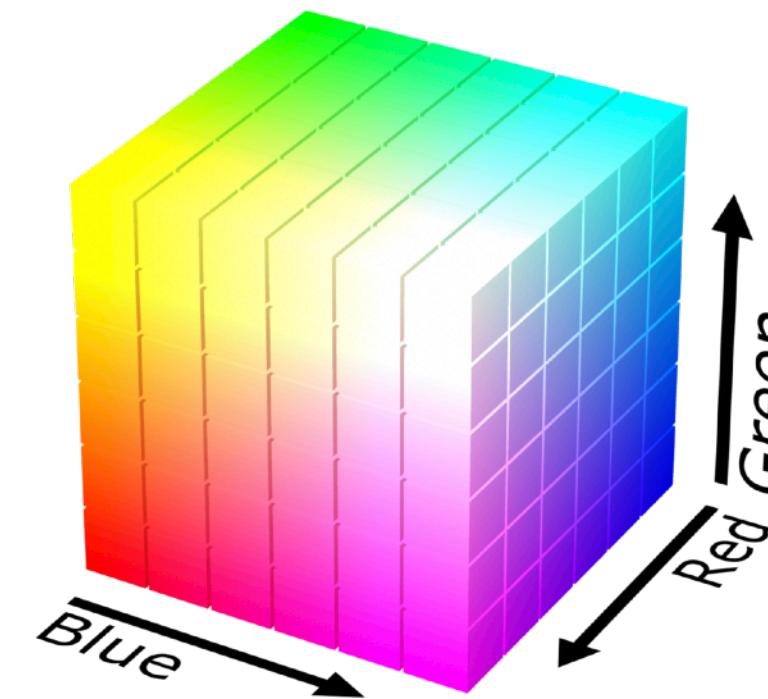
$$\mathbb{E}_{\mathcal{D}} [\log P_{\theta}(x)] = \frac{1}{|\mathcal{D}|} \sum_{x \in \mathcal{D}} \log P_{\theta}(x)$$

- Maximum (log) likelihood

$$\max_{P_{\theta}} \frac{1}{|\mathcal{D}|} \sum_{x \in \mathcal{D}} \log P_{\theta}(x)$$

Example of Joint Distributions

- Modeling a single pixel's color. Three discrete random variables:



- Possible values of all Channels : $\{0, \dots, 255\}$
- Sampling: $(r, g, b) \sim P(R, G, B)$
- How many parameters do we need to specify the joint distribution?
 - Sol) $256 \times 256 \times 256 - 1$

Example of Joint Distributions



- Suppose x_1, \dots, x_n are binary (Bernoulli) random variables
 - $\text{Val}(x_1) = \{0, 1\} = \{\text{Black}, \text{White}\}$
 - How many possible states? $\underbrace{2 \times 2 \times \dots \times 2}_{n \text{ times}} = 2^n$
- Sampling from $P(x_1, \dots, x_n)$ generates an image
- How many parameters? $2^n - 1$

Structure Through Independence

- If x_1, \dots, x_n are independent, then

$$p(x_1, \dots, x_n) = p(x_1)p(x_2) \cdots p(x_n)$$

- How many possible states? 2^n
- How many parameters to specify the joint distribution $p(x_1, \dots, x_n)$?
 - How many to specify the marginal distribution $p(x_1)$? 1
- 2^n entries can be described by just n numbers (if $|\text{Val}(x_i)| = 2$)!
- Independence assumption is too strong. Model not likely to be useful
 - Each pixel chosen independently



Chain Rule

- Chain rule or product rule of probability

$$P(x_1, \dots, x_n) = P(x_1) \prod_{i=2}^n P(x_i | x_1, \dots, x_{i-1})$$

- For example

$$P(a, b, c) = P(a | b, c)P(b, c)$$

$$P(b, c) = P(b | c)P(c)$$

$$P(a, b, c) = P(a | b, c)P(b | c)P(c)$$

$$P(a, b, c, d) = P(a | b, c, d)P(b | c, d)P(c | d)P(d)$$

$$P(a, b, c, d) = P(c | a, b, d)P(b | a, d)P(d | a)P(a)$$