Review of Probability Theory

2021
Multicampus

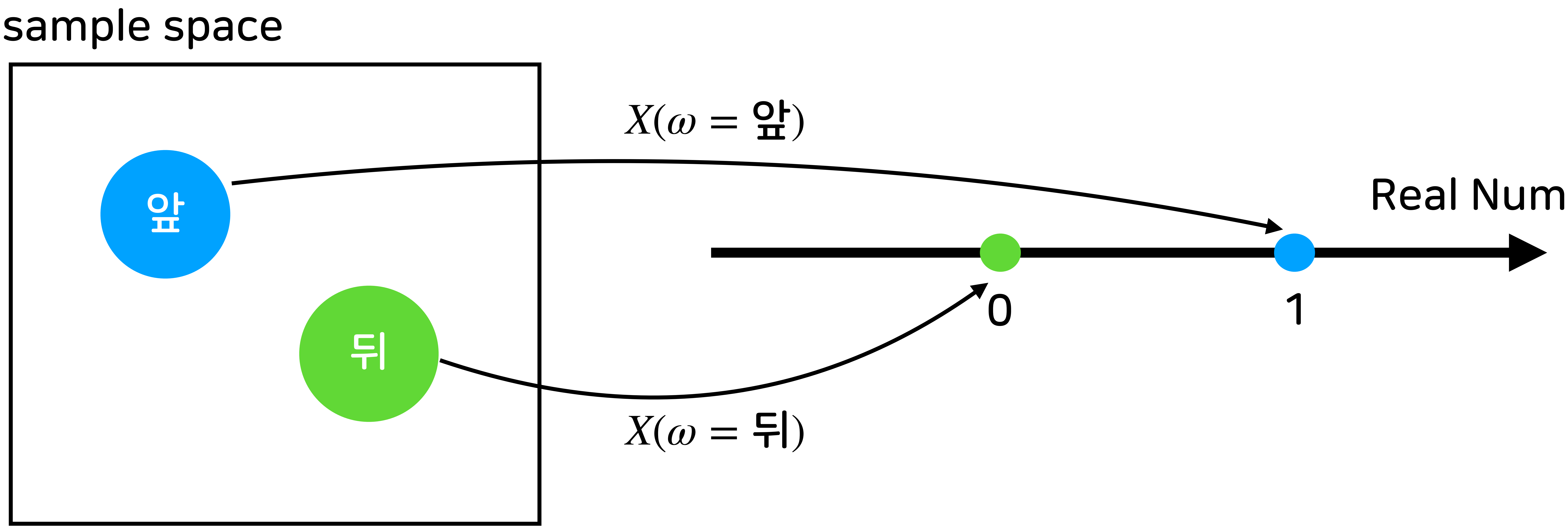
Il Gu Yi
ModuLabs, Research Scientist
Soochul Park
GaudioLab, Research Scientist

v1.2

Random Variables & Probability Distriubtions

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 μ를 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 $X = \{x_1, x_2, \ldots, x_N\}$이 주어진 경우, likelihood function을 다음과 같이 표현할 수 있습니다.
- $$p(X|\mu) = \prod_{n=1}^{N} p(x_n|\mu) = \prod_{n=1}^{N} \mu^{x_n}(1-\mu)^{(1-x_n)}$$
- Frequentist 관점에서 위 식을 최대화 하는 parameter인 μ를 구할 수 있습니다. 또는 단조 증가 함수인 $\log$함수를 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 μ의 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 N번 독립적으로 했을 때 얻을 수 있는 1의 갯수에 대한 확률 분포입니다.

- 총 시행 횟수 N과 1 값이 발생할 확률 μ를 파라메터로 갖습니다.

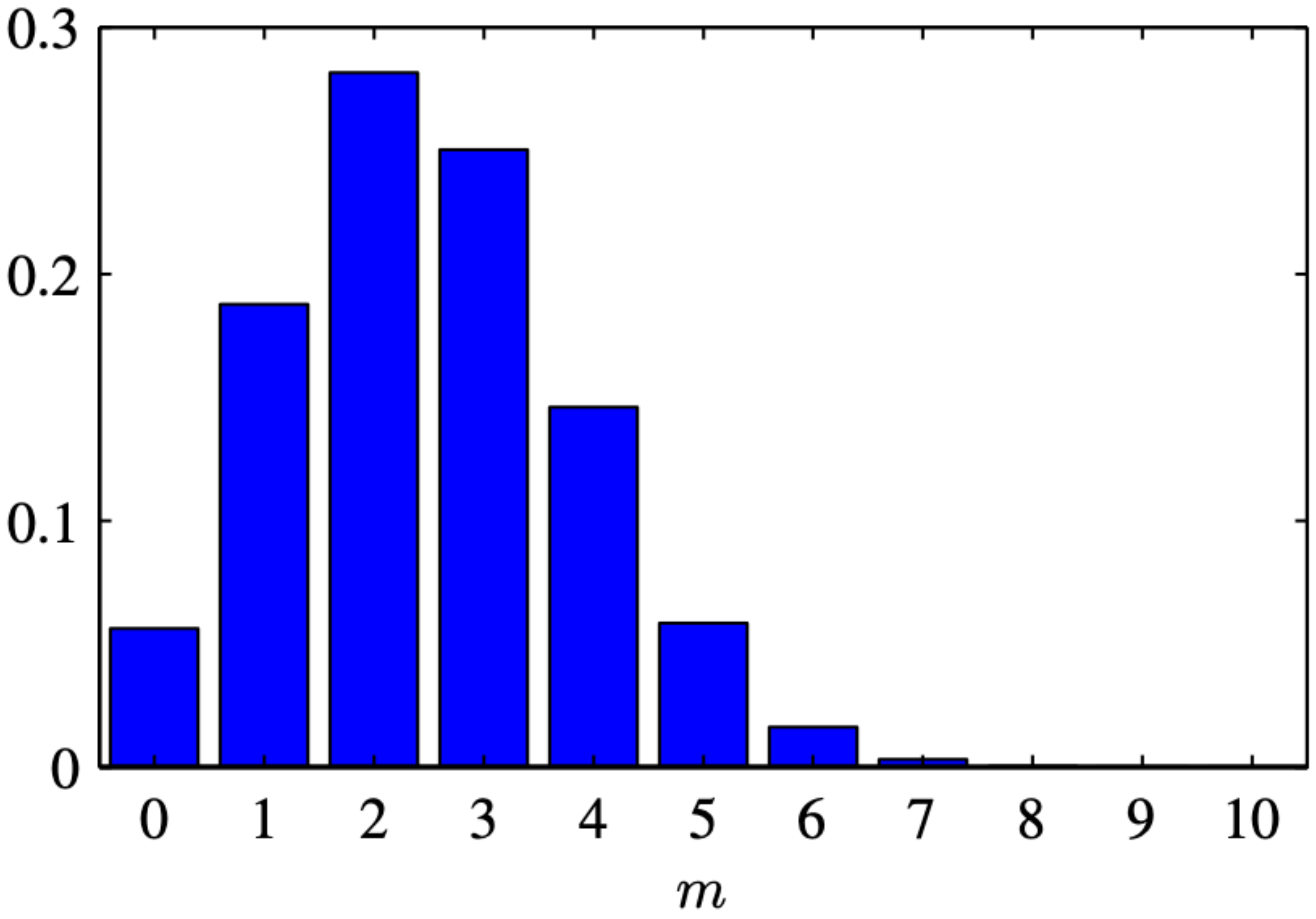
- Binomial distribution의 PMF는 다음과 같이 정의할 수 있습니다.

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

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

Binomial

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



Christopher M. Bishop, Pattern Recognition And Machine Learning 2006, p.70

Categorical-정의

- Categorical distribution은 K개의 discrete 값을 가질 수 있는 random variable에 대한 probability distribution입니다.

- Parameter로 각 카테고리에 대한 확률 값들을 나타내는 벡터 $\boldsymbol{\mu} = (\mu_1, \mu_2, \ldots, \mu_K)$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.

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

Categorical-최적화

- Dataset $\mathbf{X} = \{\mathbf{x_1}, \mathbf{x_2}, \ldots, \mathbf{x_N}\}$이 주어진 경우, likelihood function을 다음과 같이 표현할 수 있습니다.
- $$p(\mathbf{X}|\boldsymbol{\mu}) = \prod_{n=1}^{N}\prod_{k=1}^{K}\mu_k^{\mathbf{x}_{nk}} = \prod_{k=1}^{K}\mu_k^{\left(\sum_n \mathbf{x}_{nk}\right)} = \prod_{k=1}^{K}\mu_k^{N_k}$$
- Frequentist 관점에서 위 식을 최대화 하는 파라메터인 $\boldsymbol{\mu}$를 구할 수 있습니다. 또는 단조 증가 함수인 $\log$함수를 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_k값들의 합이 1이 되어야 하는 조건을 걸기 위해, Lagrange multiplier를 이용하여 다음 식을 최대화할 수 있습니다.
- $$\log p(\mathbf{X} | \boldsymbol{\mu}) = \sum_{k=1}^{K} N_k \log \mu_k + \lambda\Big(\sum_{k=1}^{K} \mu_k - 1 \Big)$$
- 위 식을 최대화하는 μ_k는 다음과 같습니다.
- $$\boxed{\mu_k = \frac{N_k}{N}}$$
- 이는 전체 데이터 중 k번째 카테고리를 갖는 데이터의 비율과 같습니다.

Categorical-최적화

- 증명

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

- 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 K개의 다른 카테고리를 가질 수 있는 random variable에서 N번 독립적으로 값을 얻었을 때, 각 카테고리가 N_k번씩 선택될 확률에 대한 분포입니다.

- Parameter로 각 카테고리에 대한 확률 값 $\boldsymbol{\mu} = (\mu_1, \mu_2, \ldots, \mu_K)$와 시행 횟수 N을 갖습니다.

- Multinomial distribution의 PMF는 다음과 같이 표현할 수 있습니다.

-

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

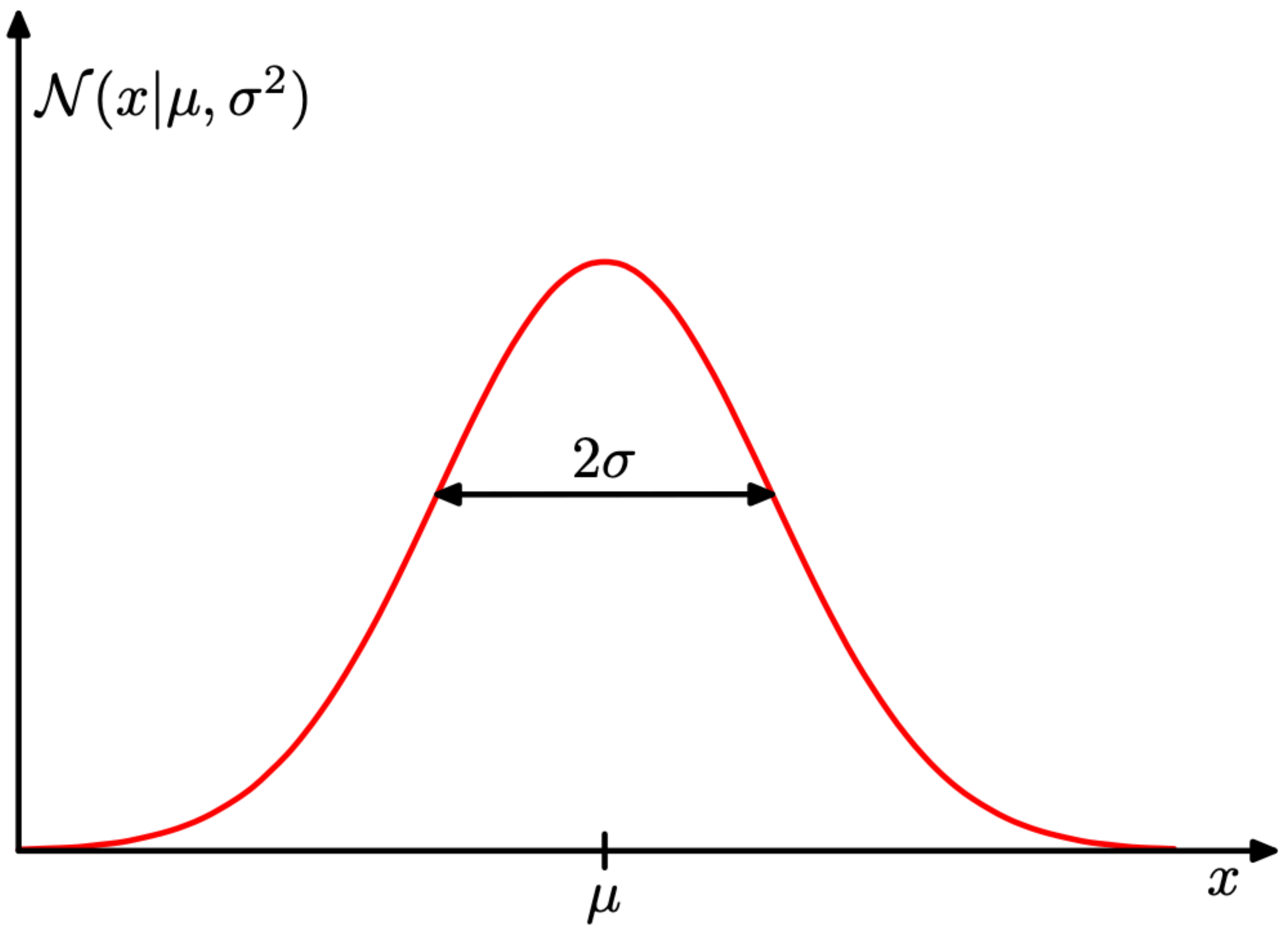
$$\text{where } \binom{N}{N_1 N_2 \cdots N_K} \equiv \frac{N!}{N_1! N_2! \cdots N_K!}$$

Gaussian-정의

- Gaussian Distribution의 PDF(Probability Distribution Function)은 다음과 같습니다.
- $$\mathcal{N}\left(x \mid \mu, \sigma^2\right) = \frac{1}{\left(2\pi\sigma^2\right)^{1/2}} \exp\left\{-\frac{1}{2\sigma^2}(x-\mu)^2\right\}$$
- Parameter로 mean값을 나타내는 μ와, 분산을 나타내는 σ^2을 갖습니다.

Gaussian

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



Christopher M. Bishop, Pattern Recognition And Machine Learning 2006, p.25

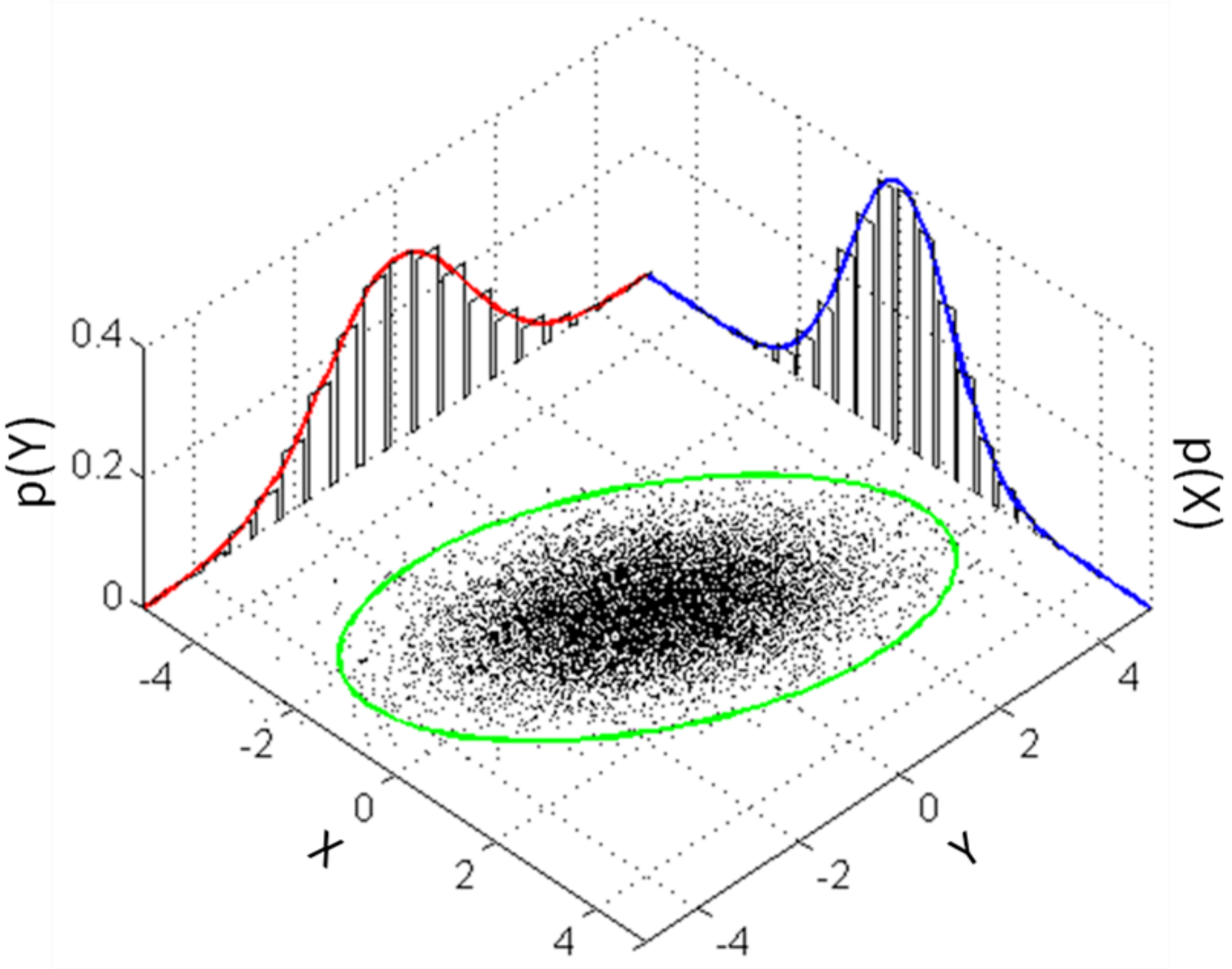
Gaussian-최적화

- Gaussian Distribution에서 독립적으로 샘플링한 데이터셋 $X = \{x_1, \dots, x_N\}^T$에 대해 log likelihood function은 다음과 같이 쓸 수 있습니다.
- $$\ln p(X \mid \mu, \sigma^2) = -\frac{N}{2}\ln 2\pi - \frac{N}{2}\ln \sigma^2 - \frac{1}{2\sigma^2}\sum_{n=1}^{N}(x_n - \mu)^2$$
- Frequentist 관점에서 위 식을 최대화하는 $\boldsymbol{\mu}$와 $\boldsymbol{\Sigma}$는 다음과 같습니다.
- $$\mu_{\mathrm{ML}} = \frac{1}{N}\sum_{n=1}^{N} x_n$$
$$\sigma^2_{\mathrm{ML}} = \frac{1}{N}\sum_{n=1}^{N}(x_n - \mu_{\mathrm{ML}})^2$$

Multivariate Gaussian-정의

- 데이터를 나타내는 $\mathbf{x}$ 가 D 차원일 벡터일 때, Multivariate Gaussian Distribution의 PDF(Probability Distribution Function)은 다음과 같습니다.
- $$\mathcal{N}(\mathbf{x}\,|\,\boldsymbol{\mu}, \boldsymbol{\Sigma}) = \frac{1}{(2\pi)^{D/2}} \frac{1}{|\boldsymbol{\Sigma}|^{1/2}} \exp\left\{ -\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu})^{\mathrm{T}} \boldsymbol{\Sigma}^{-1} (\mathbf{x}-\boldsymbol{\mu}) \right\}$$
- Parameter로 mean값을 나타내는 $\boldsymbol{\mu}$와, covariance matrix인 $\boldsymbol{\Sigma}$을 갖습니다.

Bivariate Gaussian distribution



en.wikipedia.org/wiki/Multivariate_normal_distribution

Multivariate Gaussian-최적화

- Multivariate Gaussian Distribution에서 독립적으로 샘플링한 데이터셋 $\mathbf{X} = \{\mathbf{x}_1, \ldots, \mathbf{x}_N\}^T$에 대해 log likelihood function은 다음과 같이 쓸 수 있습니다.
- $$\ln p(\mathbf{X} \mid \boldsymbol{\mu}, \boldsymbol{\Sigma}) = -\frac{ND}{2}\ln(2\pi) - \frac{N}{2}\ln|\boldsymbol{\Sigma}| - \frac{1}{2}\sum_{n=1}^{N}\left(\mathbf{x}_n - \boldsymbol{\mu}\right)^{\mathrm{T}}\boldsymbol{\Sigma}^{-1}\left(\mathbf{x}_n - \boldsymbol{\mu}\right)$$
- Frequentist 관점에서 위 식을 최대화하는 $\boldsymbol{\mu}$와 $\boldsymbol{\Sigma}$는 다음과 같습니다.
- $$\boldsymbol{\mu}_{\mathrm{ML}} = \frac{1}{N}\sum_{n=1}^{N}\mathbf{x}_n$$
  $$\boldsymbol{\Sigma}_{\mathrm{ML}} = \frac{1}{N}\sum_{n=1}^{N}\left(\mathbf{x}_n - \boldsymbol{\mu}_{\mathrm{ML}}\right)\left(\mathbf{x}_n - \boldsymbol{\mu}_{\mathrm{ML}}\right)^{\mathrm{T}}$$

Probability Theory

Joint & Marginal Probability Distribution

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

Joint prob. distribution

$$P(\mathrm{x}=x, \mathrm{y}=y)$$

Marginal probability distribution

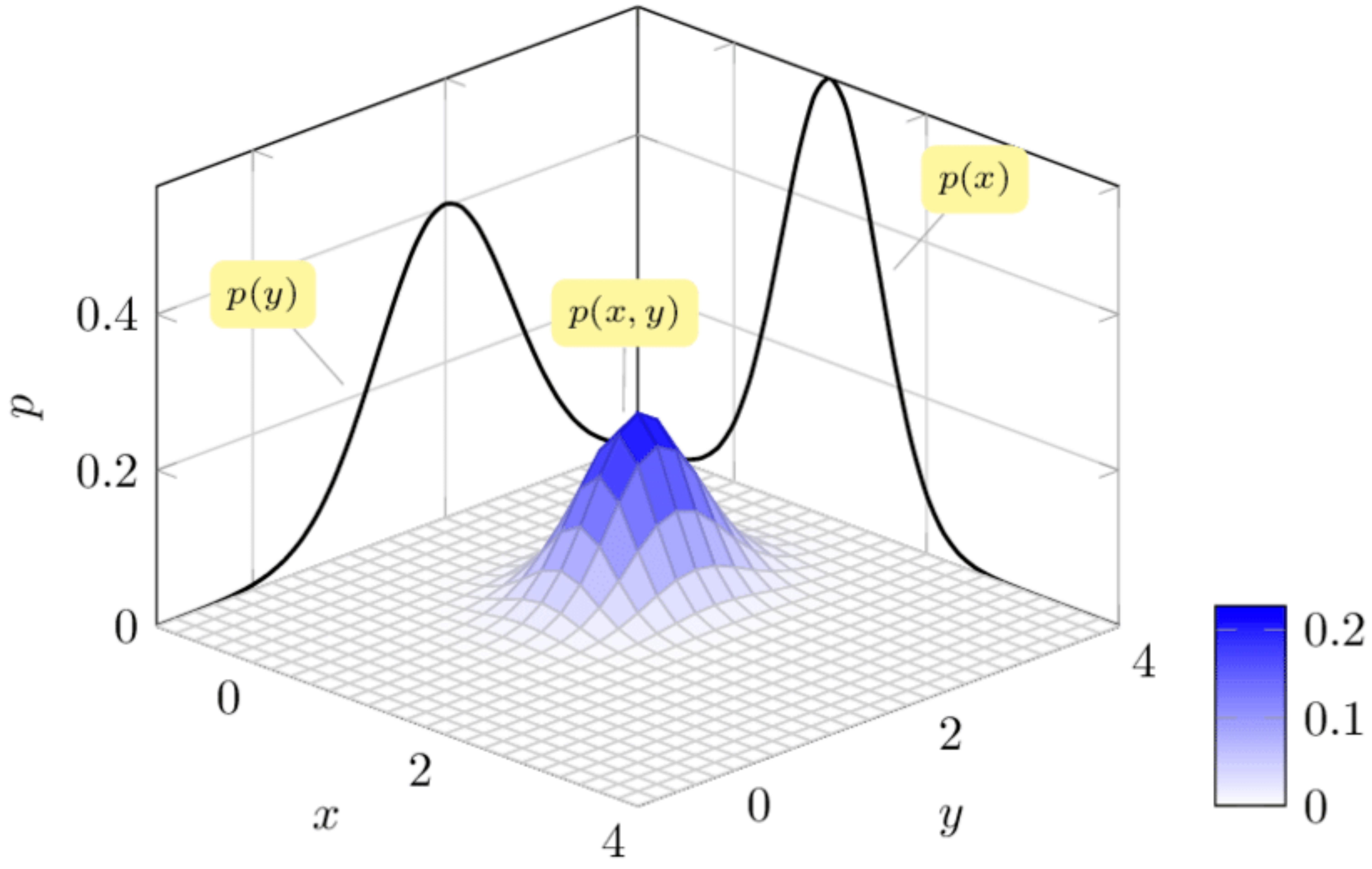
$$\forall x \in \mathrm{x}, P(\mathrm{x}=x) = \sum_y P(\mathrm{x}=x, \mathrm{y}=y)$$

$$\forall y \in \mathrm{y}, P(\mathrm{y}=y) = \sum_x P(\mathrm{x}=x, \mathrm{y}=y)$$

Continuous case

$$p(x) = \int p(x,y)\mathrm{d}y$$

Joint & Marginal Probability Distribution



https://www.researchgate.net/figure/Illustration-of-a-bivariate-Gaussian-distribution-The-marginal-and-joint-probability_fig1_320182941

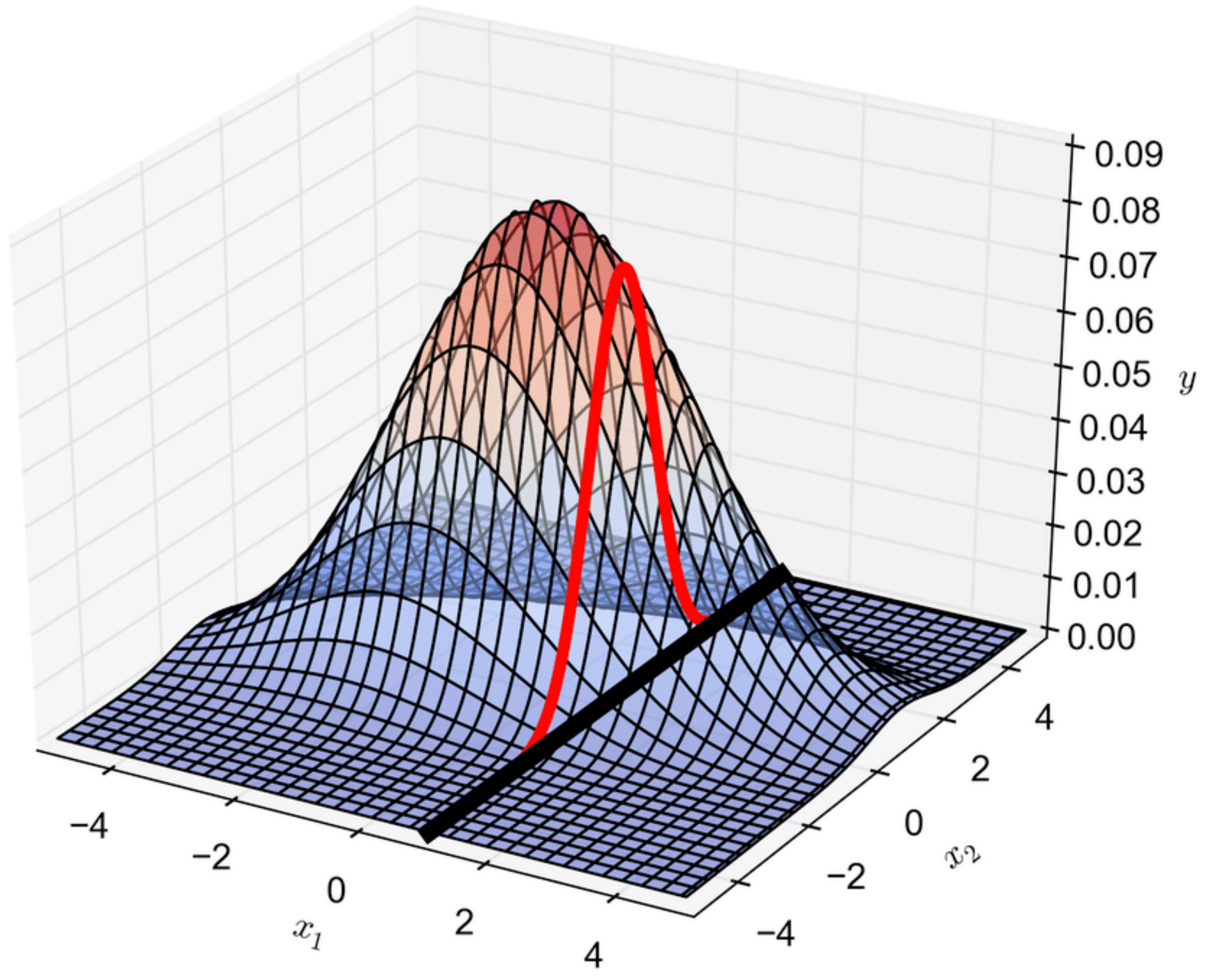
Conditional Probability Distribution

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

Conditional probability

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

Conditional Probability Distribution



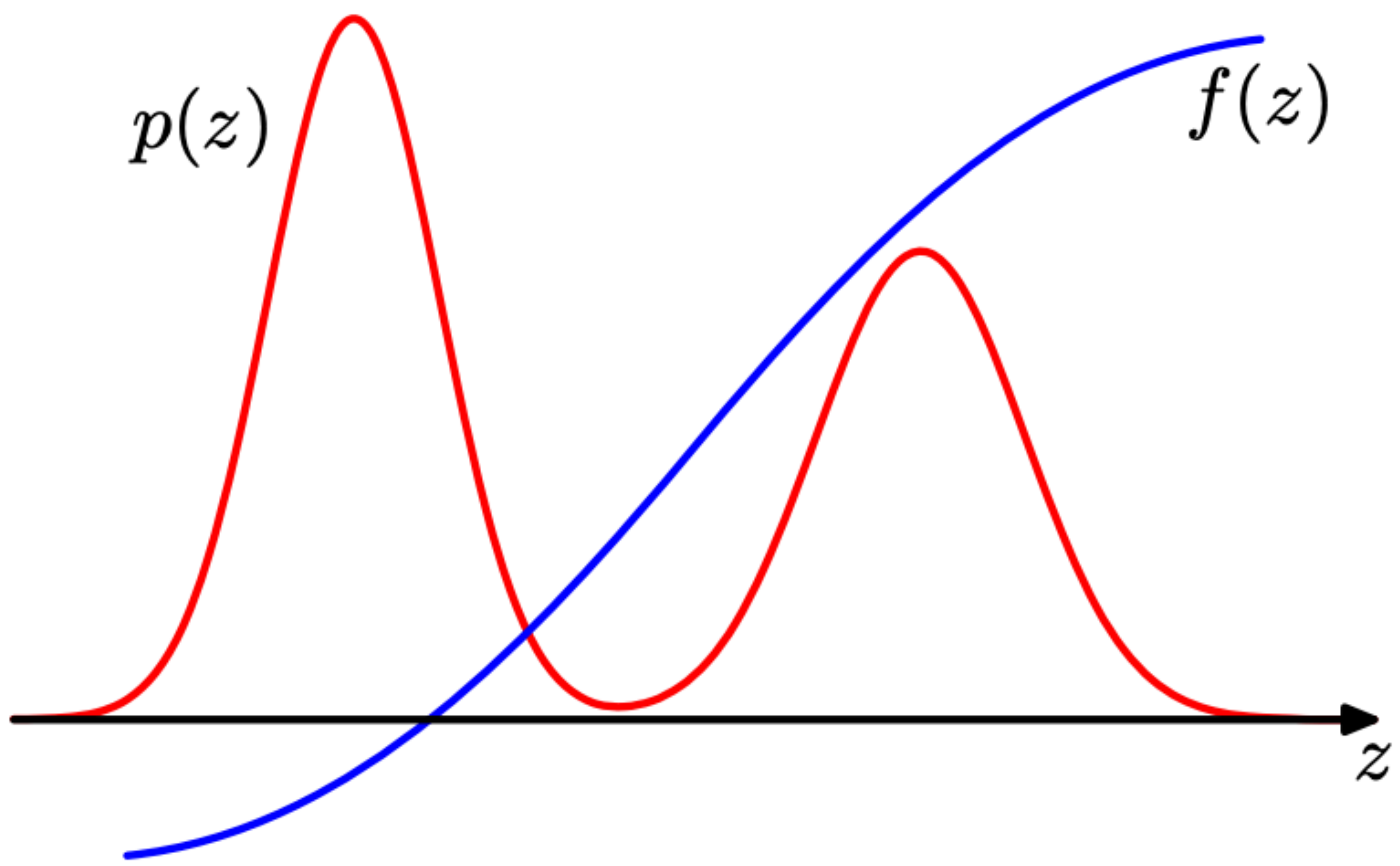
https://www.researchgate.net/figure/Linking-two-points-using-a-bivariate-Gaussian-distribution-enables-the-conditional-pdf_fig1_318429539

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.
- $$\mathbb{E}[f] = \int p(x)f(x)\mathrm{d}x$$
- Expectation은 다음과 같이 sampling을 통한 근사(approximation)으로 계산할 수도 있습니다.
- $$\mathbb{E}[f] \simeq \frac{1}{N}\sum_{n=1}^{N} f\left(x_n\right)$$

Expectation

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



Christopher M. Bishop, Pattern Recognition And Machine Learning 2006, p.524

KL-Divergence

Information Theory

- 다음과 같은 기준들에 의해 정보(information)을 수량화 합니다.
- 1. 자주 일어나는 사건은 낮은 정보량을 갖는다.

  2. 드물게 일어나는 사건은 높은 정보량을 갖는다.

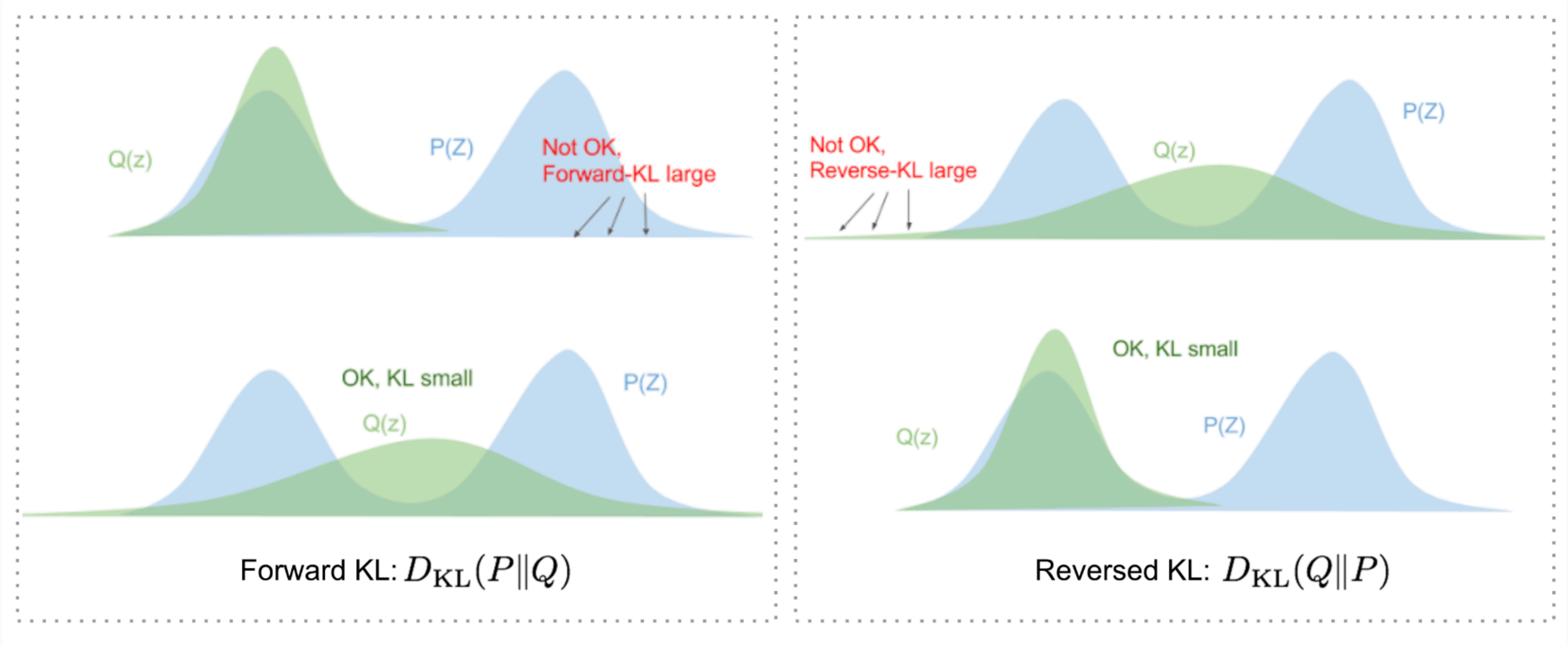
  3. 독립된 사건의 정보량은 각 사건의 정보량을 더하여 구한다.
- $$h(x) = -\log p(x)$$
- Random variable $\mathbf{x}$의 distribution이 $p(x)$일 때, $\mathbf{x}$의 엔트로피(entropy)는 다음과 같이 정보량의 기댓값으로 정의합니다.
- $$H[\mathbf{x}] = \mathbb{E}_{p(x)}[h(x)] = \begin{cases} \sum_x p(x)\{-\log p(x)\}, & \text{discrete} \\ \int p(x)\{-\log p(x)\}dx, & \text{continuous} \end{cases}$$

KL-Divergence

- KL-divergence는 두 분포의 차이를 재는 범함수(functional)입니다.
- 두 분포 P, Q에 대해서 다음과 같이 KL-divergence를 정의합니다.
- $$D_{\mathrm{KL}}(P \| Q) = \mathbb{E}_{x \sim P}\left[\log \frac{P(x)}{Q(x)}\right] = \begin{cases} \sum_x P(x) \log \frac{P(x)}{Q(x)}, \text{ discrete} \\ \int P(x) \log \frac{P(x)}{Q(x)} dx, \text{ continuous} \end{cases}$$
- KL-divergence는 symmetric하지 않으므로 수학적으로 distance의 개념이 아닙니다.
- $$D_{\mathrm{KL}}(P \| Q) \neq D_{\mathrm{KL}}(Q \| P)$$
- KL-divergence 값은 항상 0보다 같거나 크며, 두 distribution P, Q가 같을 때만 0이 됩니다.
- $$\mathbb{E}_{x \sim P}\left[-\log \frac{Q(x)}{P(x)}\right] \geq -\log\left(\mathbb{E}_{x \sim P}\left[\frac{Q(x)}{P(x)}\right]\right) = -\log\left(\sum_x P(x) \frac{Q(x)}{P(x)}\right) = 0$$

by Jensson's inequality

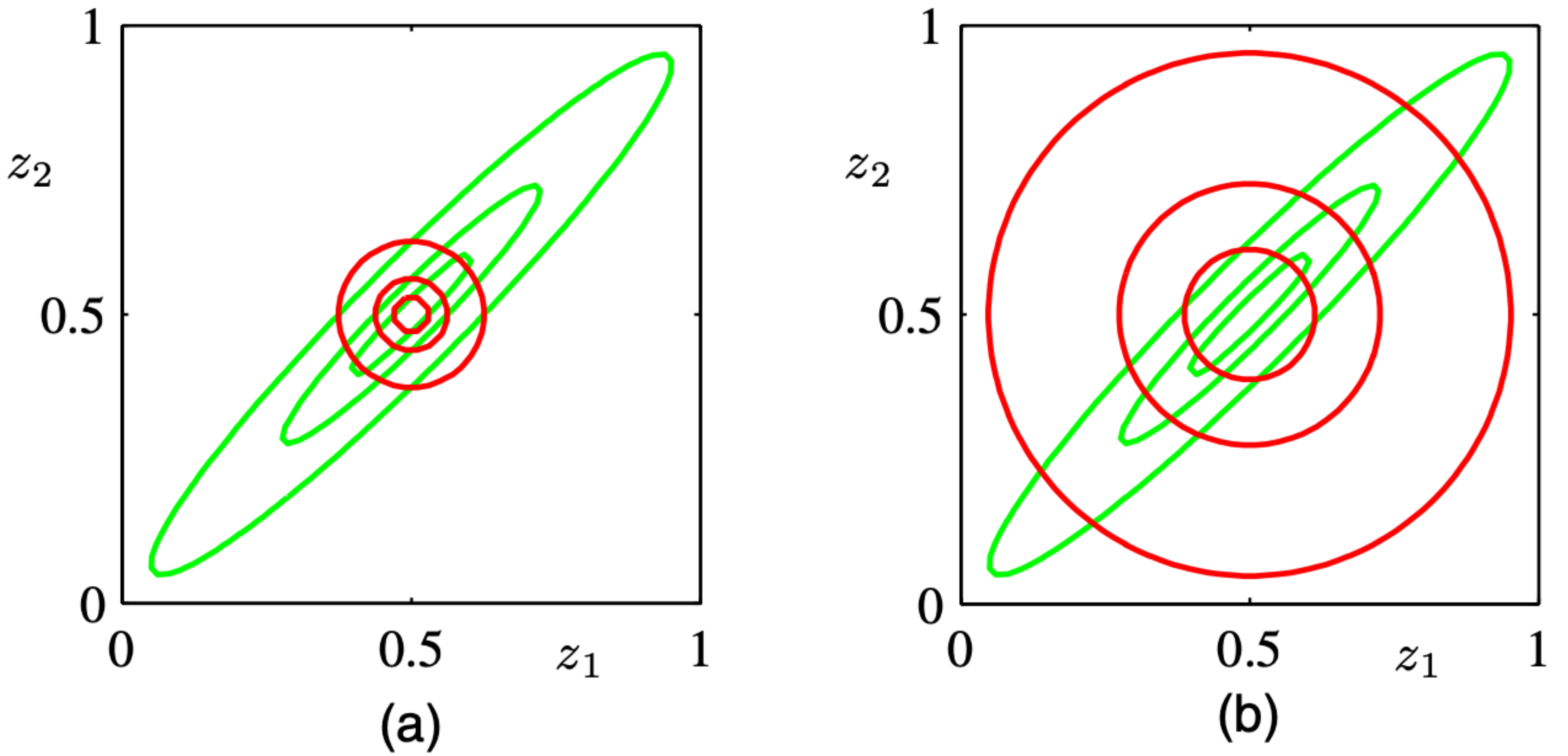
KL-Divergence



https://blog.evjang.com/2016/08/variational-bayes.html

KL-Divergence

Figure 10.2 Comparison of the two alternative forms for the Kullback-Leibler divergence. The green contours corresponding to 1, 2, and 3 standard deviations for a correlated Gaussian distribution $p(\mathbf{z})$ over two variables z_1 and z_2, and the red contours represent the corresponding levels for an approximating distribution $q(\mathbf{z})$ over the same variables given by the product of two independent univariate Gaussian distributions whose parameters are obtained by minimization of (a) the Kullback-Leibler divergence $\mathrm{KL}(q\|p)$, and (b) the reverse Kullback-Leibler divergence $\mathrm{KL}(p\|q)$.



Christopher M. Bishop, Pattern Recognition And Machine Learning 2006, p.468

KL-Divergence

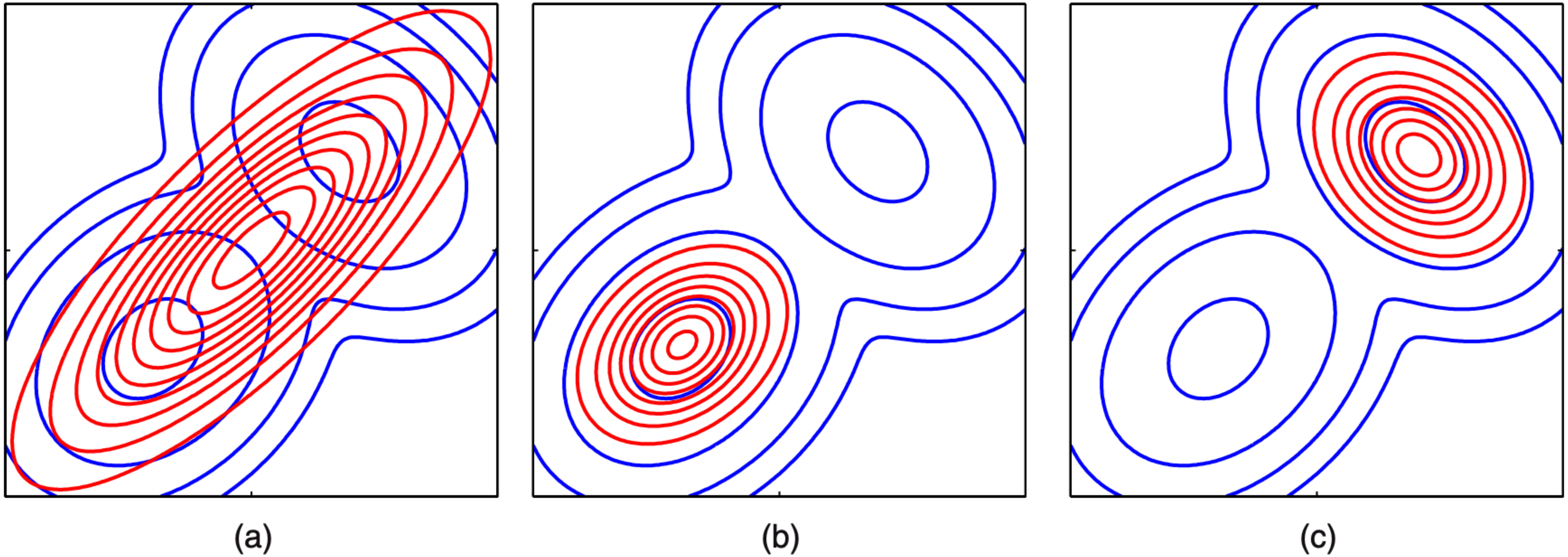


Figure 10.3 Another comparison of the two alternative forms for the Kullback-Leibler divergence. (a) The blue contours show a bimodal distribution $p(\mathbf{Z})$ given by a mixture of two Gaussians, and the red contours correspond to the single Gaussian distribution $q(\mathbf{Z})$ that best approximates $p(\mathbf{Z})$ in the sense of minimizing the Kullback-Leibler divergence $\mathrm{KL}(p\|q)$. (b) As in (a) but now the red contours correspond to a Gaussian distribution $q(\mathbf{Z})$ found by numerical minimization of the Kullback-Leibler divergence $\mathrm{KL}(q\|p)$. (c) As in (b) but showing a different local minimum of the Kullback-Leibler divergence.

Christopher M. Bishop, Pattern Recognition And Machine Learning 2006, p.469