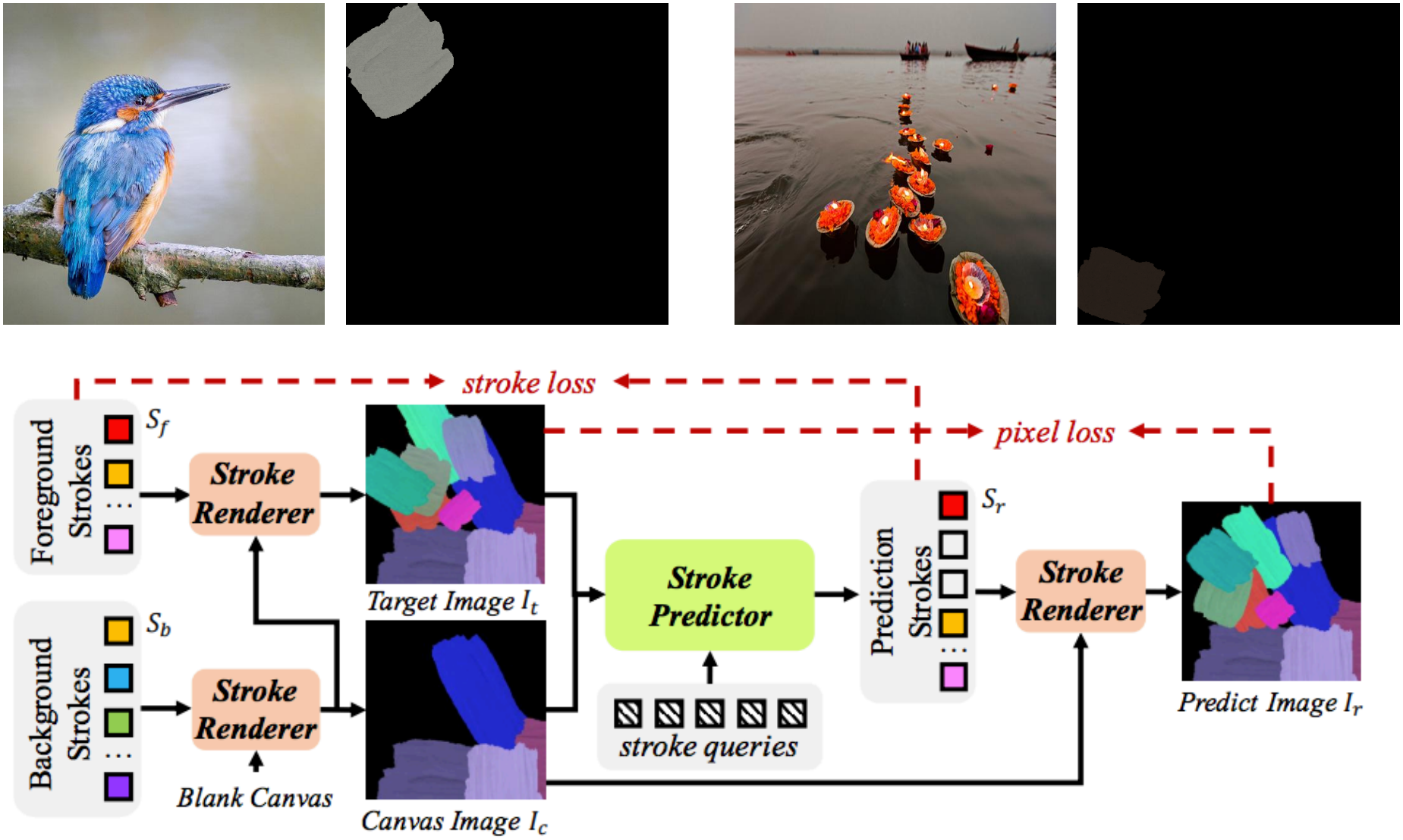
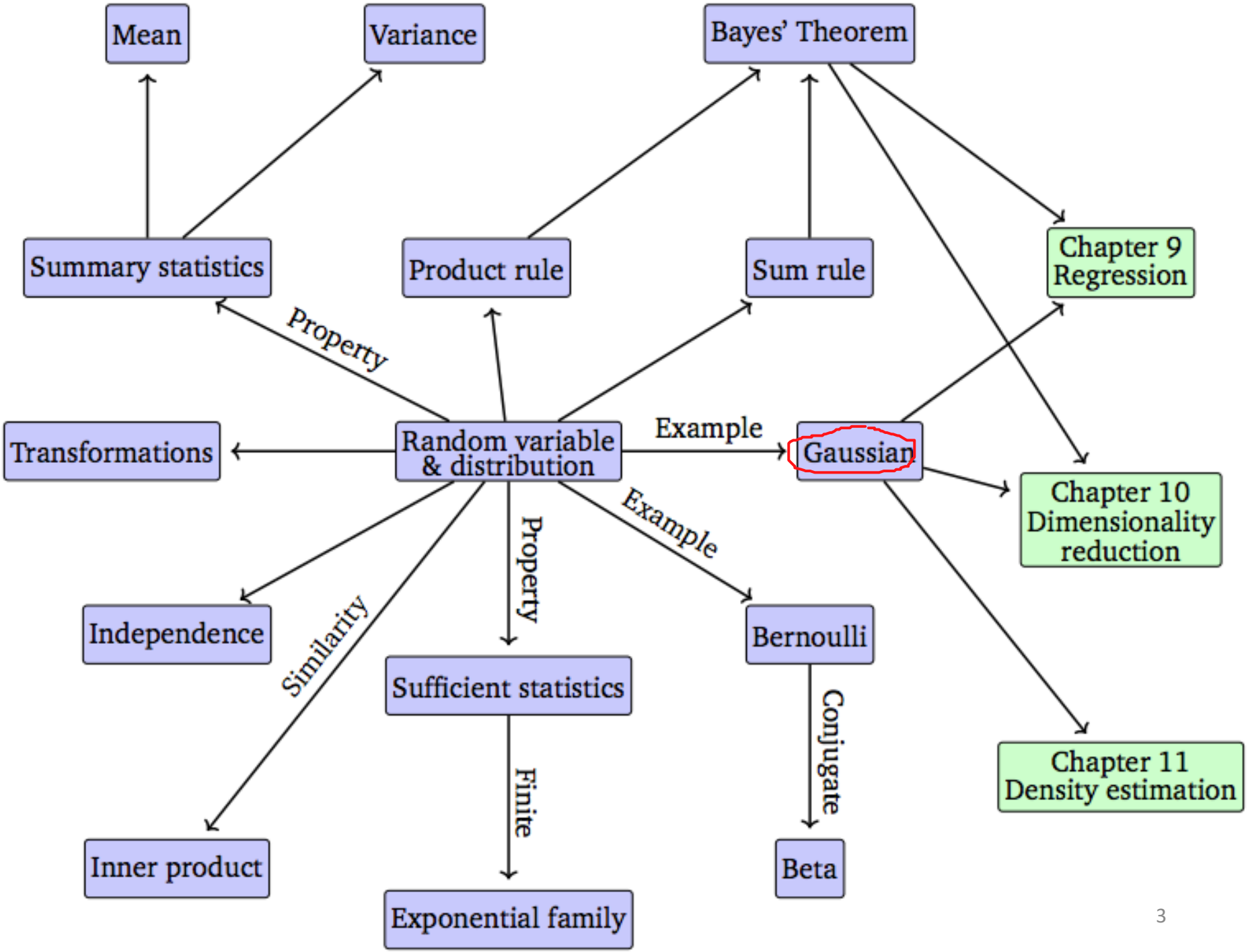
Probability and Distributions

Liang Zheng
Australian National University
liang.zheng@anu.edu.au

Paint Transformer: Feed Forward Neural Painting with Stroke Prediction, ICCV 2021




Mean
Variance
Bayes' Theorem
Summary statistics
Product rule
Sum rule
Chapter 9
Regression
Property
Transformations
Random variable
& distribution
Example
Gaussian
Chapter 10
Dimensionality
reduction
Property
Example
Independence
Similarity
Sufficient statistics
Bernoulli
Conjugate
Chapter 11
Density estimation
Finite
Inner product
Beta
Exponential family

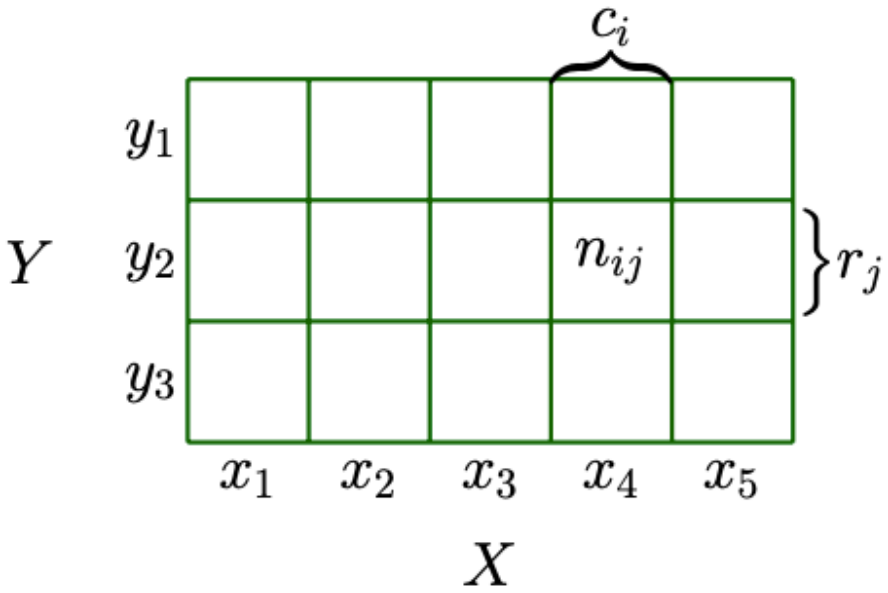
6.2.1 Discrete Probabilities

- When the target space is discrete, we can imagine the probability distribution of multiple random variables as filling out a (multidimensional) array of numbers.
- We define the joint probability as the entry of both values jointly

$$P(X = x_i, Y = y_j) = \frac{n_{ij}}{N}$$

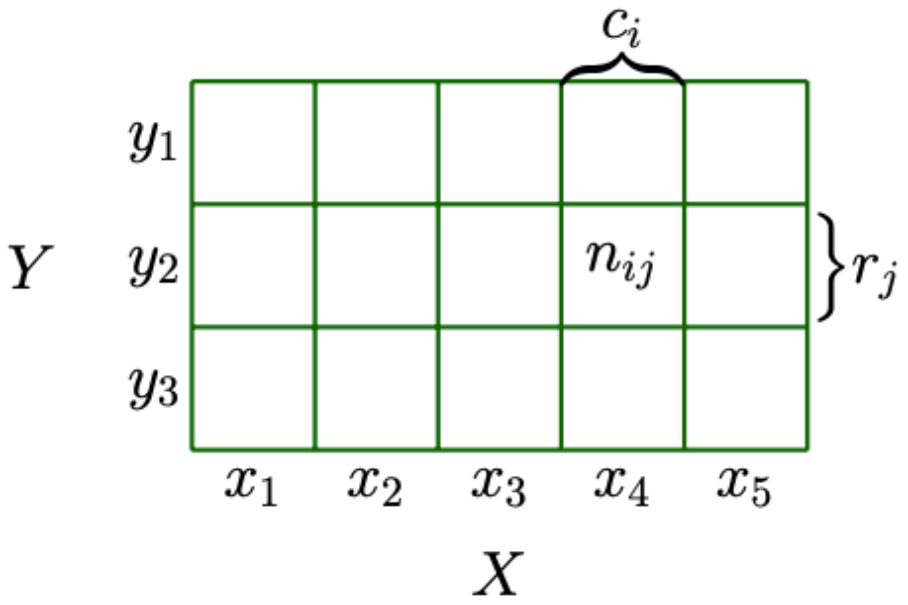
n_{ij} is the number of events with state x_i and y_j and N total number of events

- The probability that $X = x, Y = y$ is written as $p(x, y)$



6.2.1 Discrete Probabilities

- The marginal probability that $X = x$ irrespective of the value of Y is written as $p(x)$
- We write $X \sim p(x)$ to denote that the random variable X is distributed according to $p(x)$
- If we consider only the instances where $X = x$, then the fraction of instances (conditional probability) for which $Y = y$ is written as $p(y|x)$.



Example

Lebron James

≥ 30	13	6
< 30	4	17
	< 30	≥ 30

Anthony Davis

- X: AD scoring. Y: LBJ scoring.
- X has two possible states; Y has two possible states
- We use n_{ij} to denote the number of events with state $X = x$ and $Y = y$.
- Total number of events $N = 13 + 6 + 4 + 17 = 40$
- Value c_i is the event sum of the ith column, i.e., $c_i = \sum_{j=1}^{2} n_{ij}$
- r_j is the row sum, i.e., $r_j = \sum_{i=1}^{2} n_{ij}$
- The probability distribution of each random variable, the marginal probability, can be seen as the sum over a row or column

$$P(X = x_i) = \frac{c_i}{N} = \frac{\sum_{j=1}^{2} n_{ij}}{N}$$

and

$$P(Y = y_j) = \frac{r_j}{N} = \frac{\sum_{i=1}^{2} n_{ij}}{N}$$

$$P(LBJ\ scores\ at\ least\ 30\ pts) = \frac{13+6}{40}$$

Example

- For discrete random variables with a finite number of events, we assume that probabilities sum up to one, that is

$$\sum_{i=1}^{2} P(X = x_i) = 1 \text{ and } \sum_{j=1}^{2} P(Y = y_j) = 1$$

- The conditional probability is the fraction of a row or column in a particular cell. For example, the conditional probability of Y given X is

$$P(Y = y_j \mid X = x_i) = \frac{n_{ij}}{c_i}$$

- and the conditional probability of X given Y is

$$P(X = x_i \mid Y = y_j) = \frac{n_{ij}}{r_j}$$

r: row, c: col

Lebron James

	< 30	≥ 30
≥ 30	13	6
< 30	4	17

Anthony Davis

- $P(LBJ < 30 \mid AD \geq 30) = \frac{n_{12}}{c_2} = \frac{17}{6+17}$

6.2.2 Continuous Probabilities

- A function $f\colon \mathbb{R}^D \longrightarrow \mathbb{R}$ is called a probability density function (pdf) if
$$\forall \boldsymbol{x} \in \mathbb{R}^D : f(\boldsymbol{x}) \geq 0$$
- Its integral exists and
$$\int_{\mathbb{R}^D} f(\boldsymbol{x}) d\boldsymbol{x} = 1.$$
- Observe that the probability density function is any function f that is non-negative and integrates to one. We associate a random variable X with this function f by
$$P(a \leq X \leq b) = \int_a^b f(x) dx$$
where $a, b \in \mathbb{R}$; $x \in \mathbb{R}$ are outcomes of the continuous random variable X. This association is called the distribution of the random variable X.
- Note: the probability of a continuous random variable X taking a particular value $P(X = x)$ is zero. This is to specify an interval where $a = b$

6.2.2 Continuous Probabilities

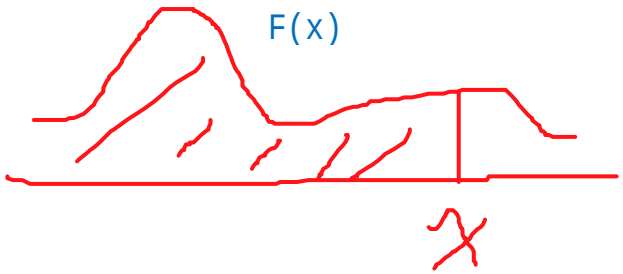
- A cumulative distribution function (cdf) of a multivariate real-valued random variable X with states $\boldsymbol{x} \in \mathbb{R}^D$ is given by

$$F_X(\boldsymbol{x}) = P(X_1 \leq x_1, \cdots, X_D \leq x_D).$$

where $X = [X_1, \cdots, X_D]^T$, $\boldsymbol{x} = [x_1, \cdots, x_D]^T$, and the right-hand side represents the probability that random variable X_i takes the value smaller than or equal to x_i.

- The cdf can be expressed also as the integral of the probability density function $f(\boldsymbol{x})$ so that

$$F_X(\boldsymbol{x}) = \int_{-\infty}^{x_1} \cdots \int_{-\infty}^{x_D} f(z_1, \cdots, z_D) dz_1 \cdots dz_D$$

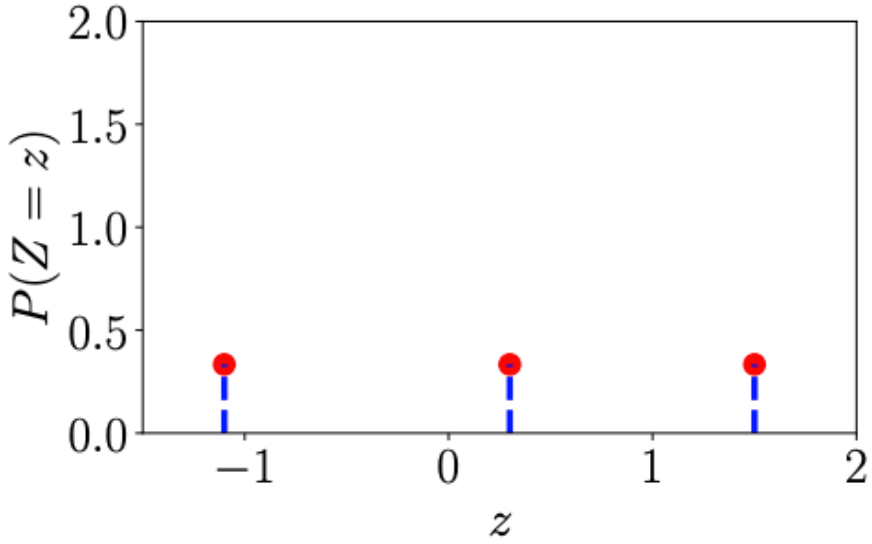


f(x)=dFX(x)/dx

6.2.3 Contrasting Discrete and Continuous Distributions

- Let Z be a discrete uniform random variable with three states $\{z = -1.1, z = 0.3, z = 1.5\}$. The probability mass function can be represented as a table of probability values:

z	-1.1	0.3	1.5
$P(Z = z)$	$\frac{1}{3}$	$\frac{1}{3}$	$\frac{1}{3}$



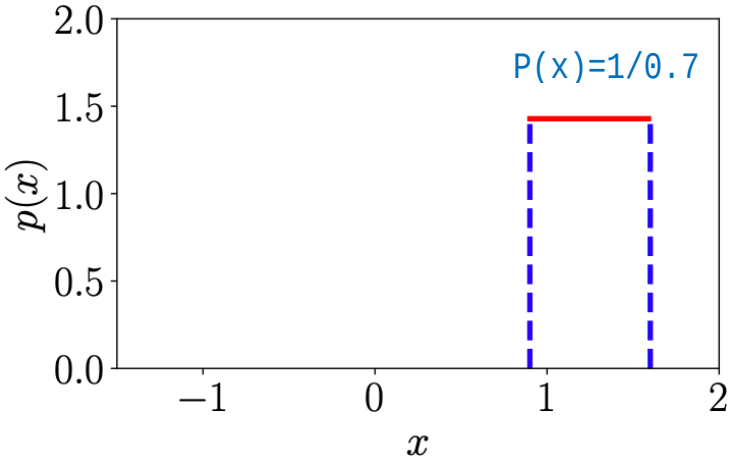
- States can be located on the x-axis, and the y-axis represents the probability of a particular state

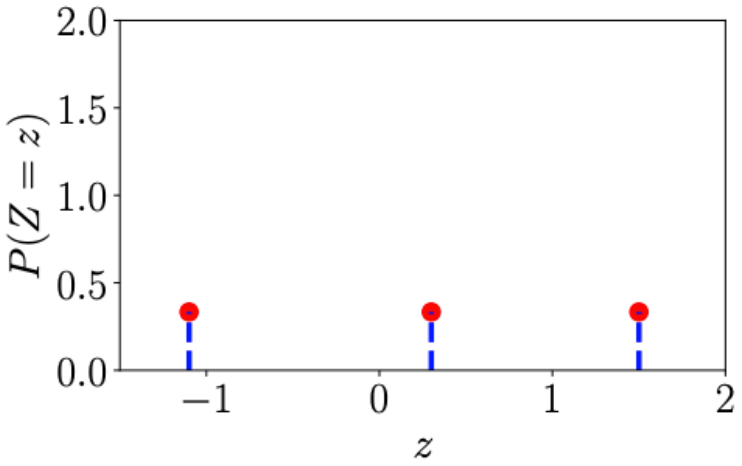
Uniform distribution (discrete): a finite number of values are equally likely to be observed; every one of n values has equal probability $1/n$

6.2.3 Contrasting Discrete and Continuous Distributions

- Let X be a continuous random variable taking values in range $0.9 \leq X \leq 1.6$
- Observe that the height of the density can be greater than 1. However, it needs to hold that

$$\int_{0.9}^{1.6} p(x)dx = 1$$





Uniform distribution (continuous): denoted as $U(a, b)$

$$f(x) = \begin{cases} \dfrac{1}{b-a} & \text{for } a \leq x \leq b \\ 0 & \text{otherwise} \end{cases}$$

6.3 Sum Rule, Product Rule, and Bayes' Theorem

- $p(\boldsymbol{x}, \boldsymbol{y})$ is the joint distribution of the two random variables $\boldsymbol{x}, \boldsymbol{y}$
- $p(\boldsymbol{x})$ and $p(\boldsymbol{y})$ are the marginal distributions
- $p(\boldsymbol{y}|\boldsymbol{x})$ is the conditional distribution of $\boldsymbol{y}$ given $\boldsymbol{x}$
- The sum rule states that

$$p(\boldsymbol{x}) = \begin{cases} \sum_{\boldsymbol{y} \in \mathcal{Y}} p(\boldsymbol{x}, \boldsymbol{y}) & \text{if } \boldsymbol{y} \text{ is discrete} \\ \int_{\mathcal{Y}} p(\boldsymbol{x}, \boldsymbol{y}) \mathrm{d}\boldsymbol{y} & \text{if } \boldsymbol{y} \text{ is continuous} \end{cases}$$

  where $\mathcal{Y}$ are the states of the target space of random variable Y.
- We sum out (or integrate out) the set of states $\boldsymbol{y}$ of the random Y
- The sum rule is also known as the marginalization property.
- If $\boldsymbol{x} = [x_1, \cdots, x_D]^{\mathrm{T}}$, we obtain the marginal

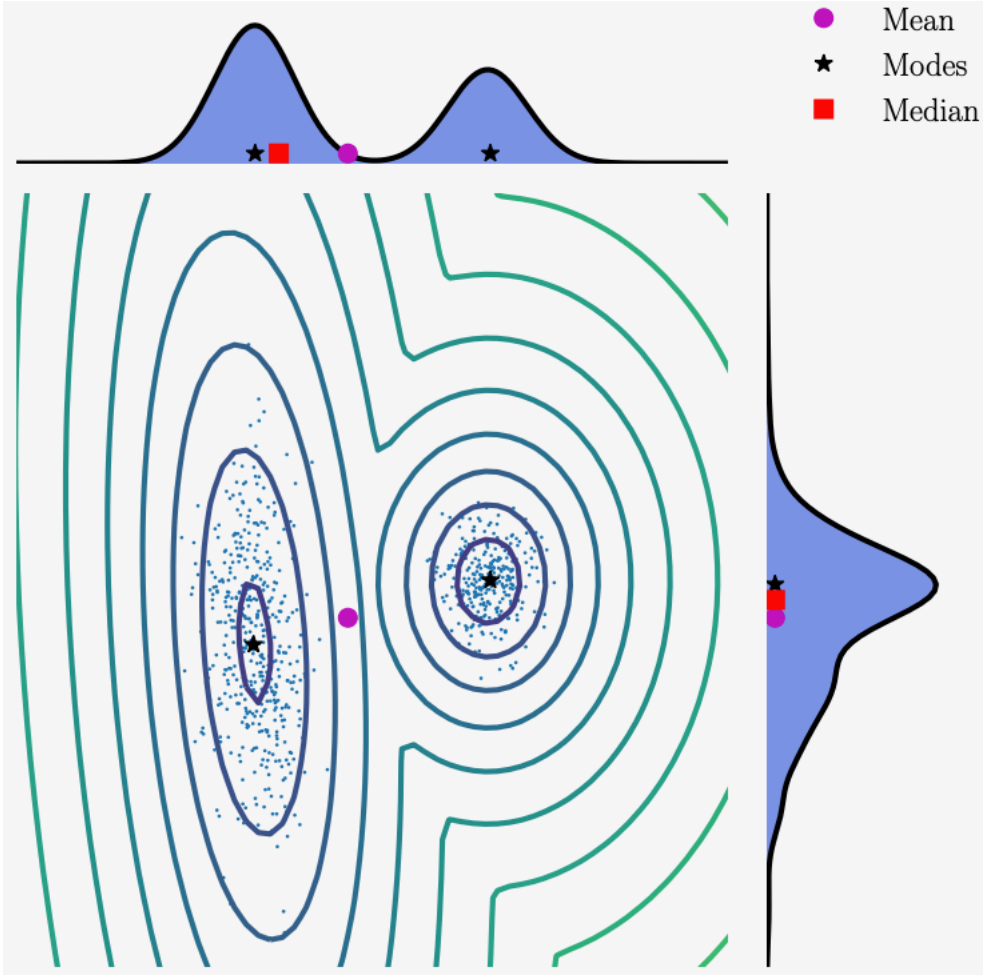
$$p(x_i) = \int p(x_1, \cdots, x_D) \mathrm{d}\boldsymbol{x}_{\setminus i}$$

by repeated application of the sum rule where we integrate/sum out all random variables except x_i, which is indicated by \i, which reads all "except i."

$$p(x) = 0.4\mathcal{N}\left(\boldsymbol{x} \mid \begin{bmatrix}10\\2\end{bmatrix}, \begin{bmatrix}1 & 0\\0 & 1\end{bmatrix}\right) + 0.6\mathcal{N}\left(\boldsymbol{x} \mid \begin{bmatrix}0\\0\end{bmatrix}, \begin{bmatrix}8.4 & 2.0\\2.0 & 1.7\end{bmatrix}\right)$$

- The distribution is bimodal (has two peaks)
- One of marginal distributions is unimodal (has one peak)

$p(x_1, x_2)$



6.3 Sum Rule, Product Rule, and Bayes' Theorem

- The product rule relates the joint distribution to the conditional distribution

$$p(\boldsymbol{x}, \boldsymbol{y}) = p(\boldsymbol{y} \mid \boldsymbol{x})p(\boldsymbol{x})$$

- Every joint distribution of two random variables can be factorized (written as a product) of two other distributions
- The product rule also implies

$$p(\boldsymbol{x}, \boldsymbol{y}) = p(\boldsymbol{x} \mid \boldsymbol{y})p(\boldsymbol{y})$$

6.3 Sum Rule, Product Rule, and Bayes' Theorem

- Let us assume we have some prior knowledge $p(\boldsymbol{x})$ about an unobserved random variable $\boldsymbol{x}$ and some relationship $p(\boldsymbol{y} \mid \boldsymbol{x})$ between $\boldsymbol{x}$ and a second random variable $\boldsymbol{y}$, which we can observe. If we observe $\boldsymbol{y}$, we can use Bayes' theorem (also *Bayes' rule* or *Bayes' law*) to draw some conclusions about $\boldsymbol{x}$ given the observed values of $\boldsymbol{y}$.

$$\underbrace{p(\boldsymbol{x} \mid \boldsymbol{y})}_{\text{posterior}} = \frac{\overbrace{p(\boldsymbol{y} \mid \boldsymbol{x})}^{\text{likelihood}} \overbrace{p(\boldsymbol{x})}^{\text{prior}}}{\underbrace{p(\boldsymbol{y})}_{\text{evidence}}}$$

is a direct consequence of the product rule, since

$$p(\boldsymbol{x}, \boldsymbol{y}) = p(\boldsymbol{x} \mid \boldsymbol{y}) p(\boldsymbol{y})$$

and

$$p(\boldsymbol{x}, \boldsymbol{y}) = p(\boldsymbol{y} \mid \boldsymbol{x}) p(\boldsymbol{x})$$

so that

$$p(\boldsymbol{x} \mid \boldsymbol{y}) p(\boldsymbol{y}) = p(\boldsymbol{y} \mid \boldsymbol{x}) p(\boldsymbol{x}) \Leftrightarrow p(\boldsymbol{x} \mid \boldsymbol{y}) = \frac{p(\boldsymbol{y} \mid \boldsymbol{x}) p(\boldsymbol{x})}{p(\boldsymbol{y})}$$

$$p(\boldsymbol{x} \mid \boldsymbol{y}) = \frac{p(\boldsymbol{y} \mid \boldsymbol{x})p(\boldsymbol{x})}{p(\boldsymbol{y})}$$

$$p(LBJ < 30 \mid AD \geq 30) = \frac{p(AD \geq 30 | LBJ < 30)p(LBJ < 30)}{p(AD \geq 30)} = \frac{\frac{17}{21} \cdot \frac{21}{40}}{\frac{23}{40}} = \frac{17}{23}$$

Lebron James

≥ 30	13	6
< 30	4	17
	< 30	≥ 30

Anthony Davis

6.3 Sum Rule, Product Rule, and Bayes' Theorem

likelihood

Linear regression: $p(y|\boldsymbol{x}, \boldsymbol{\theta})$

P(θ|x,y)

$$p(\boldsymbol{x} \mid \boldsymbol{y}) = \frac{\overbrace{p(\boldsymbol{y} \mid \boldsymbol{x})}^{\text{likelihood}} \overbrace{p(\boldsymbol{x})}^{\text{prior}}}{p(\boldsymbol{y})}$$

posterior

evidence

- $p(\boldsymbol{x})$ is the prior, which encapsulates our subjective prior knowledge of the unobserved (latent) variable $\boldsymbol{x}$ before observing any data
- $p(\boldsymbol{y} \mid \boldsymbol{x})$, the likelihood, describes how $\boldsymbol{x}$ and $\boldsymbol{y}$ are related
- $p(\boldsymbol{y} \mid \boldsymbol{x})$ is the probability of the data $\boldsymbol{y}$ if we were to know the latent variable $\boldsymbol{x}$
- We call $p(\boldsymbol{y} \mid \boldsymbol{x})$ either the "likelihood of $\boldsymbol{x}$ (given $\boldsymbol{y}$)" or the "probability of $\boldsymbol{y}$ given $\boldsymbol{x}$" ($\boldsymbol{y}$ is observed; $\boldsymbol{x}$ is latent)
- $p(\boldsymbol{x} \mid \boldsymbol{y})$, the posterior, is the quantity of interest in Bayesian statistics because it expresses exactly what we are interested in, i.e., what we know about $\boldsymbol{x}$ after having observed $\boldsymbol{y}$ (e.g., linear regression or Gaussian mixture models)

6.3 Sum Rule, Product Rule, and Bayes' Theorem

$$\underbrace{p(\boldsymbol{x} \mid \boldsymbol{y})}_{\text{posterior}} = \frac{\overbrace{p(\boldsymbol{y} \mid \boldsymbol{x})}^{\text{likelihood}} \overbrace{p(\boldsymbol{x})}^{\text{prior}}}{\underbrace{p(\boldsymbol{y})}_{\text{evidence}}}$$

- The quantity

$$p(\boldsymbol{y}) := \int p(\boldsymbol{y} \mid \boldsymbol{x}) p(\boldsymbol{x}) \, d\boldsymbol{x} = \mathbb{E}_X[p(\boldsymbol{y} \mid \boldsymbol{x})]$$

- is the *marginal likelihood/evidence*.
- The marginal likelihood integrates the numerator with respect to the latent variable $\boldsymbol{x}$

6.4 Summary Statistics and Independence
6.4.1 Means and Covariances

- The expected value of a function $g : \mathbb{R} \longrightarrow \mathbb{R}$ of a univariate continuous random variable $X \sim p(x)$ is given by

$$\mathbb{E}_X[g(x)] = \int_{\mathcal{X}} g(x)p(x)dx$$

- Correspondingly, the expected value of a function g of a discrete random variable $X \sim p(x)$ is given by

$$\mathbb{E}_X[g(x)] = \sum_{x \in \mathcal{X}} g(x)p(x)$$

where $\mathcal{X}$ is the set of possible outcomes (the target space) of the random variable X

- We consider multivariate random variables X as a finite vector of univariate random variables $[X_1, \cdots, X_n]^{\mathrm{T}}$. For multivariate random variables, we define the expected value element wise

$$\mathbb{E}_X[g(\boldsymbol{x})] = \begin{bmatrix} \mathbb{E}_{X_1}[g(x_1)] \\ \vdots \\ \mathbb{E}_{X_D}[g(x_D)] \end{bmatrix} \in \mathbb{R}^D$$

where the subscript $\mathbb{E}_{X_d}$ indicates that we are taking the expected value with respect to the dth element of the vector $\boldsymbol{x}$

6.4.1 Means and Covariances

- The mean of a random variable X with states $\boldsymbol{x} \in \mathbb{R}^D$ is an average and is defined as

$$\mathbb{E}_X[\boldsymbol{x}] = \begin{bmatrix} \mathbb{E}_{X_1}[x_1] \\ \vdots \\ \mathbb{E}_{X_D}[x_D] \end{bmatrix} \in \mathbb{R}^D$$

where

$$\mathbb{E}_{x_d}[x_d] := \begin{cases} \int_{\mathcal{X}} x_d p(x_d) \mathrm{d}x_d & \text{if } X \text{ is a continuous random variable} \\ \sum_{x_i \in \mathcal{X}} x_i p(x_d = x_i) & \text{if } X \text{ is a discrete random variable} \end{cases}$$

for $d = 1, \ldots, D$, where the subscript d indicates the corresponding dimension of $\boldsymbol{x}$. The integral and sum are over the states $\mathcal{X}$ of the target space of the random variable X.

6.4.1 Means and Covariances

- The expected value is a linear operator. For example, given a real-valued function $f(\boldsymbol{x}) = ag(\boldsymbol{x}) + bh(\boldsymbol{x})$ where $a, b \in \mathbb{R}$ and $\boldsymbol{x} \in \mathbb{R}^D$, we obtain

$$\mathbb{E}_X[f(\boldsymbol{x})] = \int f(\boldsymbol{x})p(\boldsymbol{x})\mathrm{d}\boldsymbol{x}$$

$$= \int [ag(\boldsymbol{x}) + bh(\boldsymbol{x})]p(\boldsymbol{x})\mathrm{d}\boldsymbol{x}$$

$$= a\int g(\boldsymbol{x})p(\boldsymbol{x})\mathrm{d}\boldsymbol{x} + b\int h(\boldsymbol{x})p(\boldsymbol{x})\mathrm{d}\boldsymbol{x}$$

$$= a\mathbb{E}_X[g(\boldsymbol{x})] + b\mathbb{E}_X[h(\boldsymbol{x})]$$

6.4.1 Means and Covariances

协方差矩阵描述随机变量之间的线性相关关系

- The covariance（协方差） between two univariate random variables $X, Y \in \mathbb{R}$ is given by the expected product of their deviations from their respective means, i.e.,

$$Cov_{X,Y}[x, y] := \mathbb{E}_{X,Y}[(x - \mathbb{E}_X[x])(y - \mathbb{E}_Y[y])]$$

- By using the linearity of expectations, It can be rewritten as the expected value of the product minus the product of the expected values, i.e.,

$$Cov[x, y] = \mathbb{E}[xy] - \mathbb{E}[x]\mathbb{E}[y]$$

- The covariance of a variable with itself $Cov[x, x]$ is called the variance and is denoted by $\mathbb{V}_X[x]$.
- The square root of the variance is called the standard deviation and is often denoted by $\sigma(x)$.

- If we consider two multivariate random variables X and Y with states $\boldsymbol{x} \in \mathbb{R}^D$ and $\boldsymbol{y} \in \mathbb{R}^E$ respectively, the covariance between X and Y is defined as

$$\mathrm{Cov}[\boldsymbol{x}, \boldsymbol{y}] = \mathbb{E}[\boldsymbol{x}\boldsymbol{y}^{\mathrm{T}}] - \mathbb{E}[\boldsymbol{x}]\mathbb{E}[\boldsymbol{y}]^{\mathrm{T}} = Cov[\boldsymbol{y}, \boldsymbol{x}]^{\mathrm{T}} \in \mathbb{R}^{D \times E}$$

6.4.1 Means and Covariances

- The variance of a random variable X with states $\boldsymbol{x} \in \mathbb{R}^D$ and a mean vector $\boldsymbol{\mu} \in \mathbb{R}^D$ is defined as

$$\mathbb{V}_X[\boldsymbol{x}] = \mathrm{Cov}_X[\boldsymbol{x}, \boldsymbol{x}]$$
$$= \mathbb{E}_X[(\boldsymbol{x} - \boldsymbol{\mu})(\boldsymbol{x} - \boldsymbol{\mu})^{\mathrm{T}}] = \mathbb{E}_X[\boldsymbol{x}\boldsymbol{x}^{\mathrm{T}}] - \mathbb{E}_X[\boldsymbol{x}]\mathbb{E}_X[\boldsymbol{x}]^{\mathrm{T}}$$

$$= \begin{bmatrix} Cov[x_1, x_1] & Cov[x_1, x_2] & \cdots & Cov[x_1, x_D] \\ Cov[x_2, x_1] & Cov[x_2, x_2] & \cdots & Cov[x_2, x_D] \\ \vdots & \vdots & \ddots & \vdots \\ Cov[x_D, x_1] & \cdots & \cdots & Cov[x_D, x_D] \end{bmatrix}$$

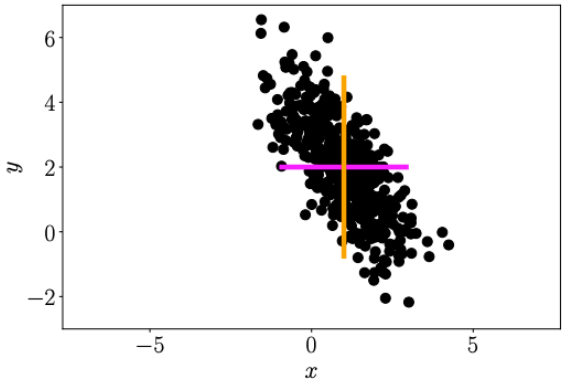
- The $D \times D$ matrix is called the covariance matrix of the multivariate random variable X. The covariance matrix is symmetric and positive definite and tells us something about the spread of the data.
- On its diagonal, the covariance matrix contains the variances of x_i

6.4.1 Means and Covariances

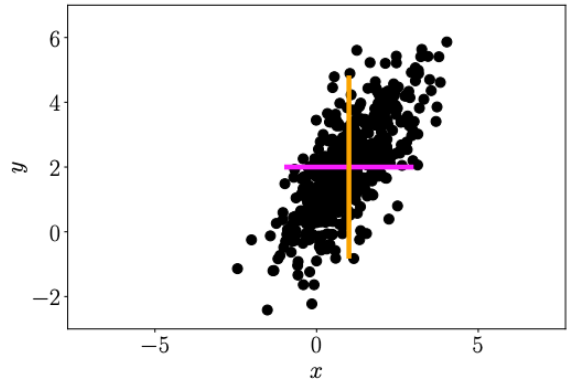
- The correlation between two random variables X, Y is given by

$$\text{corr}[x, y] = \frac{\text{Cov}[x, y]}{\sqrt{\mathbb{V}[x]\mathbb{V}[y]}} \in [-1, 1]$$

- The correlation is the covariance of standardized random variables, $x/\sigma(x)$. In other words, each random variable is divided by its standard deviation (the square root of the variance) in the correlation.
- The covariance (and correlation) indicate how two random variables are related;
- Positive correlation $\text{corr}[x, y]$ means that when x grows, then y is also expected to grow. Negative correlation means that as x increases, then y decreases



(a) x and y are negatively correlated.



(b) x and y are positively correlated.

Two-dimensional datasets with identical means and variances along each axis (colored lines) but with different covariances.

6.4.2 Empirical Means and Covariances

- In 6.4.1 we defined population mean and covariance, as it refers to the true statistics for the population 但是在ML中通常因无法统计全部数据而难以计算全局均值和协方差
- In machine learning, we have a finite dataset of size N
- The empirical mean vector is the arithmetic average of the observations for each variable, and it is defined as

$$\bar{\boldsymbol{x}} := \frac{1}{N}\sum_{n=1}^{N} \boldsymbol{x}_n$$

  Where $\boldsymbol{x}_n \in \mathbb{R}^D$.

- The empirical covariance matrix is a $D \times D$ matrix

$$\sum := \frac{1}{N}\sum_{n=1}^{N} (\boldsymbol{x}_n - \bar{\boldsymbol{x}})(\boldsymbol{x}_n - \bar{\boldsymbol{x}})^{\mathrm{T}}$$

- To compute the statistics for a particular dataset, we would use the realizations (observations) $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_N$ and use the two equations above.

Example - Computation of covariance matrix

- $X = [\boldsymbol{x}_1, \boldsymbol{x}_2] = \begin{bmatrix} 1 & 3 \\ 2 & 1 \\ 3 & 1 \end{bmatrix}$
- $\boldsymbol{\mu} = \begin{bmatrix} 2 \\ 1.5 \\ 2 \end{bmatrix}$;
- $X - \boldsymbol{\mu} = \begin{bmatrix} -1 & 1 \\ 0.5 & -0.5 \\ 1 & -1 \end{bmatrix}$
- $\sum = \frac{1}{2}(X - \boldsymbol{\mu})(X - \boldsymbol{\mu})^T = \frac{1}{2}\begin{bmatrix} 2 & -1 & -2 \\ -1 & 0.5 & 1 \\ -2 & 1 & 2 \end{bmatrix}$

the matrix is symmetric

negative number means the two vectors are moving to different direction (负相关), positive number means the same direction (正相关).

6.4.3 Three Expressions for the Variance

- The standard definition of variance is the expectation of the squared deviation of a random variable X from its expected value μ, i.e.,
$$\mathbb{V}_X[x] := \mathbb{E}_X[(x - \mu)^2]$$
- This is equivalent to the mean of a new random variable $Z := (X - \mu)^2$.
- We use a two-pass algorithm: one pass through the data to calculate μ, and then a second pass using this estimate $\hat{\mu}$ to calculate the variance.
- It can be converted to the so-called raw-score formula for variance:
$$\mathbb{V}_X[x] = \mathbb{E}_X[x^2] - (\mathbb{E}_X[x])^2$$
- The mean of the square minus the square of the mean. It can be calculated empirically in one pass
- A third way to understand the variance is that it is a sum of pairwise differences between all pairs of observations. Consider a sample $x_1, \ldots, x_N$ of realizations of random variable X, and we compute the squared difference between pairs of x_i and x_j. By expanding the square, we can show that the sum of N^2 pairwise differences is the empirical variance of the observations:
$$\frac{1}{N^2}\sum_{i,j=1}^{N}(x_i - x_j)^2 = \cdots = 2\left[\frac{1}{N}\sum_{i=1}^{N}{x_i}^2 - \left(\frac{1}{N}\sum_{i=1}^{N}x_i\right)^2\right] \text{(lab question)}$$

6.4.4 Sums and Transformations of Random Variables

- Consider two random variables X, Y with states $\boldsymbol{x}, \boldsymbol{y} \in \mathbb{R}^D$. Then:
- $\mathbb{E}[\boldsymbol{x}+\boldsymbol{y}] = \mathbb{E}[\boldsymbol{x}] + \mathbb{E}[\boldsymbol{y}]$
- $\mathbb{E}[\boldsymbol{x}-\boldsymbol{y}] = \mathbb{E}[\boldsymbol{x}] - \mathbb{E}[\boldsymbol{y}]$
- $\mathbb{V}[\boldsymbol{x}+\boldsymbol{y}] = \mathbb{V}[\boldsymbol{x}] + \mathbb{V}[\boldsymbol{y}] + \mathrm{Cov}[\boldsymbol{x},\boldsymbol{y}] + \mathrm{Cov}[\boldsymbol{y},\boldsymbol{x}]$ can you prove it?
- $\mathbb{V}[\boldsymbol{x}-\boldsymbol{y}] = \mathbb{V}[\boldsymbol{x}] + \mathbb{V}[\boldsymbol{y}] - \mathrm{Cov}[\boldsymbol{x},\boldsymbol{y}] - \mathrm{Cov}[\boldsymbol{y},\boldsymbol{x}]$
- Mean and (co)variance have useful properties when it comes to affine transformation of random variables. Consider a random variable X with mean $\boldsymbol{\mu}$ and covariance matrix $\boldsymbol{\Sigma}$ and an affine transformation $\boldsymbol{y} = \boldsymbol{A}\boldsymbol{x} + \boldsymbol{b}$ of $\boldsymbol{x}$. Then $\boldsymbol{y}$ is itself a random variable whose mean vector and covariance matrix are given by
- $\mathbb{E}_Y[\boldsymbol{y}] = \mathbb{E}_X[\boldsymbol{A}\boldsymbol{x}+\boldsymbol{b}] = \boldsymbol{A}\mathbb{E}_X[\boldsymbol{x}] + \boldsymbol{b} = \boldsymbol{A}\boldsymbol{\mu} + \boldsymbol{b}$
- $\mathbb{V}_Y[\boldsymbol{y}] = \mathbb{V}_X[\boldsymbol{A}\boldsymbol{x}+\boldsymbol{b}] = \mathbb{V}_X[\boldsymbol{A}\boldsymbol{x}] = \boldsymbol{A}\mathbb{V}_X[\boldsymbol{x}]\boldsymbol{A}^{\mathrm{T}} = \boldsymbol{A}\boldsymbol{\Sigma}\boldsymbol{A}^{\mathrm{T}}$ can you prove it?
- Furthermore,
- $\mathrm{Cov}[\boldsymbol{x},\boldsymbol{y}] = \mathbb{E}[\boldsymbol{x}(\boldsymbol{A}\boldsymbol{x}+\boldsymbol{b})^{\mathrm{T}}] - \mathbb{E}[\boldsymbol{x}]\mathbb{E}[\boldsymbol{A}\boldsymbol{x}+\boldsymbol{b}]^{\mathrm{T}}$
- $= \mathbb{E}[\boldsymbol{x}]\boldsymbol{b}^{\mathrm{T}} + \mathbb{E}[\boldsymbol{x}\boldsymbol{x}^{\mathrm{T}}]\boldsymbol{A}^{\mathrm{T}} - \boldsymbol{\mu}\boldsymbol{b}^{\mathrm{T}} - \boldsymbol{\mu}\boldsymbol{\mu}^{\mathrm{T}}\boldsymbol{A}^{\mathrm{T}}$
- $= \boldsymbol{\mu}\boldsymbol{b}^{\mathrm{T}} - \boldsymbol{\mu}\boldsymbol{b}^{\mathrm{T}} + \left(\mathbb{E}[\boldsymbol{x}\boldsymbol{x}^{\mathrm{T}}] - \boldsymbol{\mu}\boldsymbol{\mu}^{\mathrm{T}}\right)\boldsymbol{A}^{\mathrm{T}}$
- $= \boldsymbol{\Sigma}\boldsymbol{A}^{\mathrm{T}}$
- where $\Sigma = \mathbb{E}[\boldsymbol{x}\boldsymbol{x}^{\mathrm{T}}] - \boldsymbol{\mu}\boldsymbol{\mu}^{\mathrm{T}}$ is the covariance of X.

6.4.5 Statistical Independence

- Two random variables X, Y are statistically independent if and only if
$$p(\boldsymbol{x}, \boldsymbol{y}) = p(\boldsymbol{x})p(\boldsymbol{y})$$
- Intuitively, two random variables X and Y are independent if the value of $\boldsymbol{y}$ (once known) does not add any additional information about $\boldsymbol{x}$ (and vice versa). If X, Y are (statistically) independent, then
$$p(\boldsymbol{y} \mid \boldsymbol{x}) = p(\boldsymbol{y})$$
$$p(\boldsymbol{x} \mid \boldsymbol{y}) = p(\boldsymbol{x})$$
$$\mathrm{Cov}_{X,Y}[\boldsymbol{x}, \boldsymbol{y}] = \boldsymbol{0}$$
$$\mathbb{V}_{X,Y}[\boldsymbol{x} + \boldsymbol{y}] = \mathbb{V}_X[\boldsymbol{x}] + \mathbb{V}_Y[\boldsymbol{y}]$$
- The last point may not hold in converse, i.e., two random variables can have covariance zero but are not statistically independent. To understand why, recall that covariance measures only linear dependence. Therefore, random variables that are nonlinearly dependent could have covariance zero
- **Example**. Consider a random variable X with zero mean ($\mathbb{E}_X[x] = 0$) and also $\mathbb{E}_X[x^3] = 0$. Let $y = x^2$ (hence, Y is dependent on X) and consider the covariance between X and Y. But this gives
$$\mathrm{Cov}[x, y] = \mathbb{E}[xy] - \mathbb{E}[x]\mathbb{E}[y] = \mathbb{E}[x^3] = 0$$

6.5 Gaussian Distribution

- The Gaussian distribution is the most well-studied probability distribution for continuous-valued random variables.
- It is also referred to as the normal distribution.
- For a univariate random variable, the Gaussian distribution has a density that is given by

mean standard variance

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

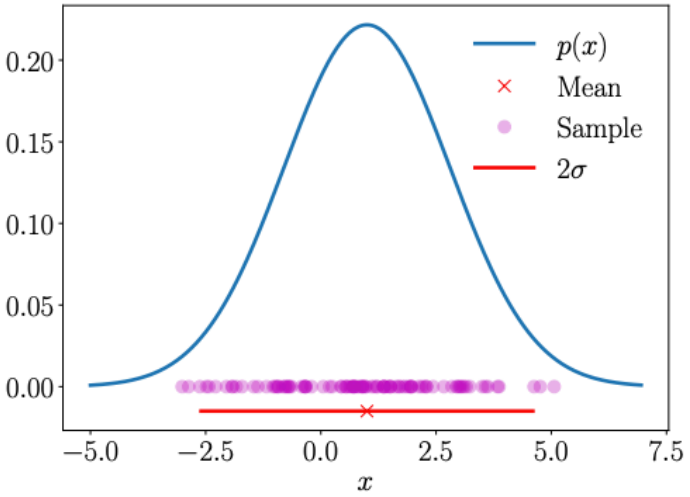
- The multivariate Gaussian distribution is fully characterized by a mean vector $\boldsymbol{\mu}$ and a covariance matrix $\boldsymbol{\Sigma}$ and defined as

$$p(\boldsymbol{x}|\boldsymbol{\mu},\boldsymbol{\Sigma}) = (2\pi)^{-\frac{D}{2}}|\boldsymbol{\Sigma}|^{-\frac{1}{2}}\exp(-\frac{1}{2}(\boldsymbol{x}-\boldsymbol{\mu})^T\boldsymbol{\Sigma}^{-1}(\boldsymbol{x}-\boldsymbol{\mu}))$$

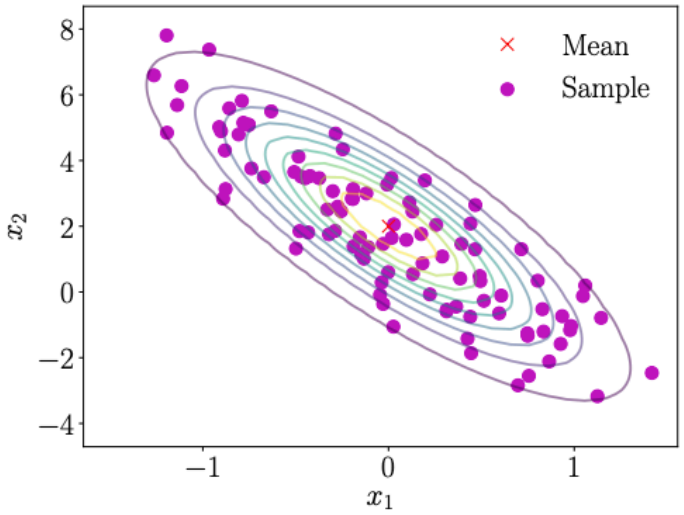
  where $\boldsymbol{x} \in \mathbb{R}^D$. We write $p(x) = \mathcal{N}(\boldsymbol{x}|\boldsymbol{\mu},\boldsymbol{\Sigma})$ or $X \sim \mathcal{N}(\boldsymbol{\mu},\boldsymbol{\Sigma})$.
- The special case of the Gaussian with zero mean and identity covariance, that is, $\boldsymbol{\mu} = \boldsymbol{0}$ and $\boldsymbol{\Sigma} = \boldsymbol{I}$, is referred to as the standard normal distribution.

6.5 Gaussian Distribution

- Figure below shows a univariate Gaussian and a bivariate Gaussian with corresponding samples.



(a) Univariate (one-dimensional) Gaussian; The red cross shows the mean and the red line shows the extent of the variance.



(b) Multivariate (two-dimensional) Gaussian, viewed from top. The red cross shows the mean and the colored lines show the contour lines of the density.

Spherical Gaussian

- General probability density function

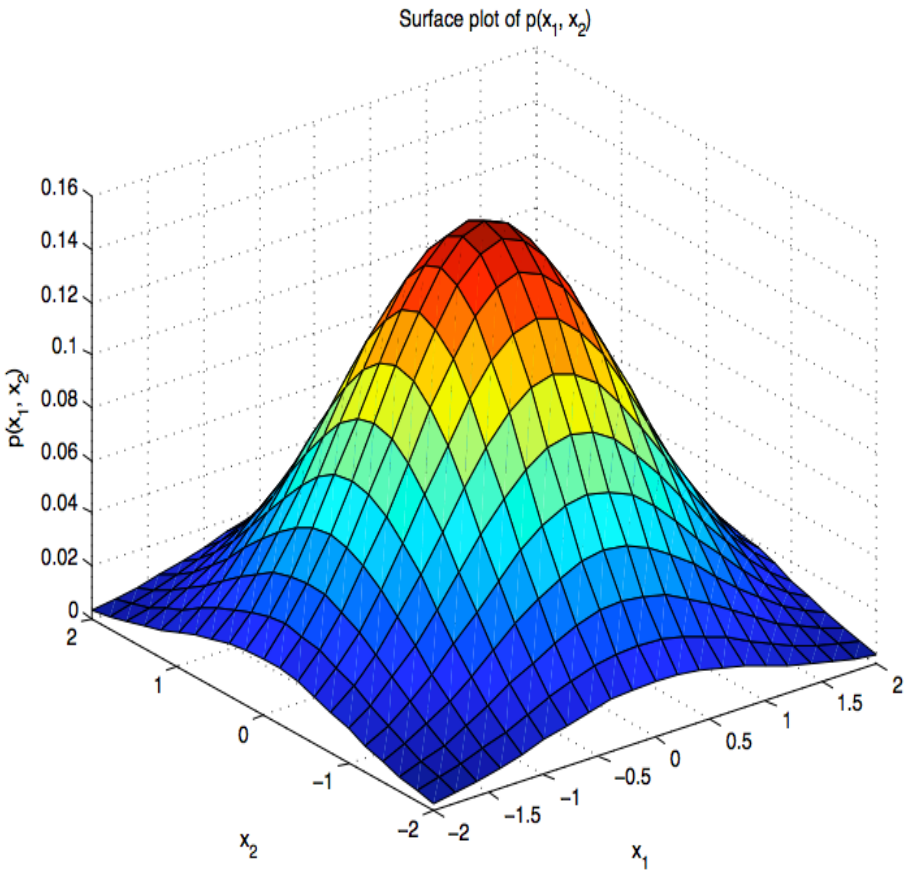
$$p(\boldsymbol{x}|\boldsymbol{\mu},\boldsymbol{\Sigma}) = (2\pi)^{-\frac{D}{2}}|\boldsymbol{\Sigma}|^{-\frac{1}{2}}\exp(-\frac{1}{2}(\boldsymbol{x}-\boldsymbol{\mu})^T\boldsymbol{\Sigma}^{-1}(\boldsymbol{x}-\boldsymbol{\mu}))$$

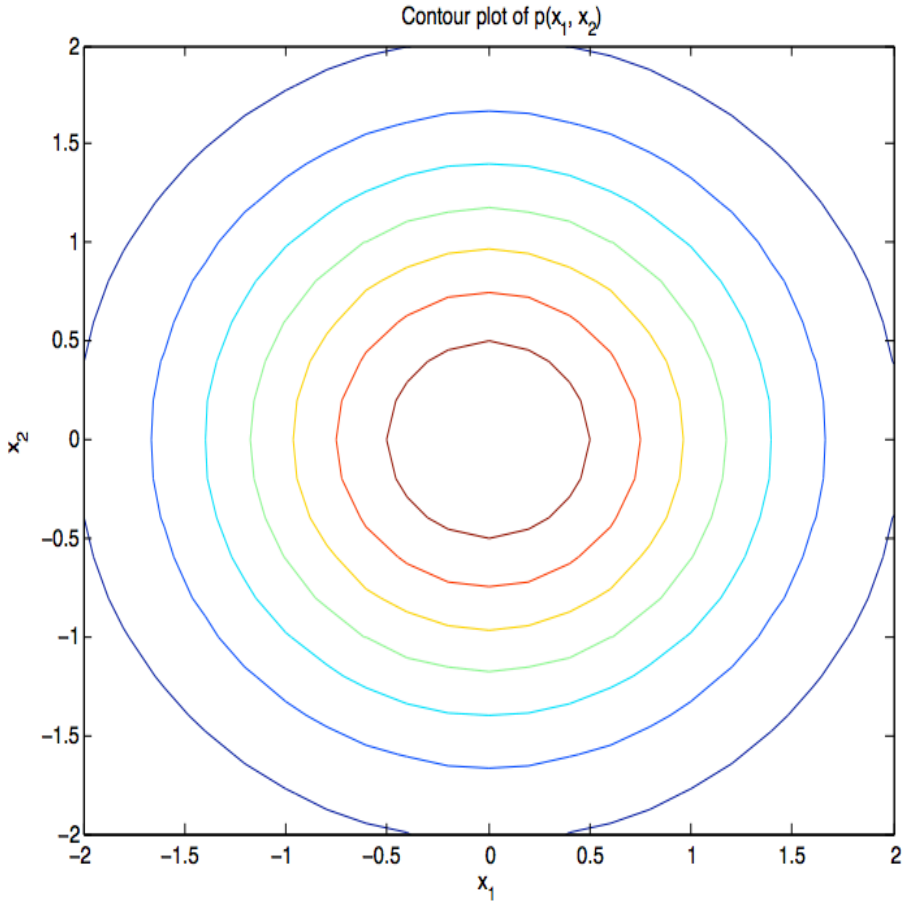
- Spherical Gaussian
- $p(\boldsymbol{x} \mid \boldsymbol{\mu}, \sigma^2) = (2\pi\sigma^2)^{-D/2} \exp\left\{-\frac{1}{2\sigma^2}\|\boldsymbol{x}-\boldsymbol{\mu}\|^2\right\}, \quad \boldsymbol{\mu} \in \mathbb{R}^D, \sigma \in \mathbb{R}.$

$$\sum = \begin{bmatrix} \sigma^2 & 0 & 0 \\ 0 & \sigma^2 & 0 \\ 0 & 0 & \sigma^2 \end{bmatrix}$$

diagonal covariance, equal variances

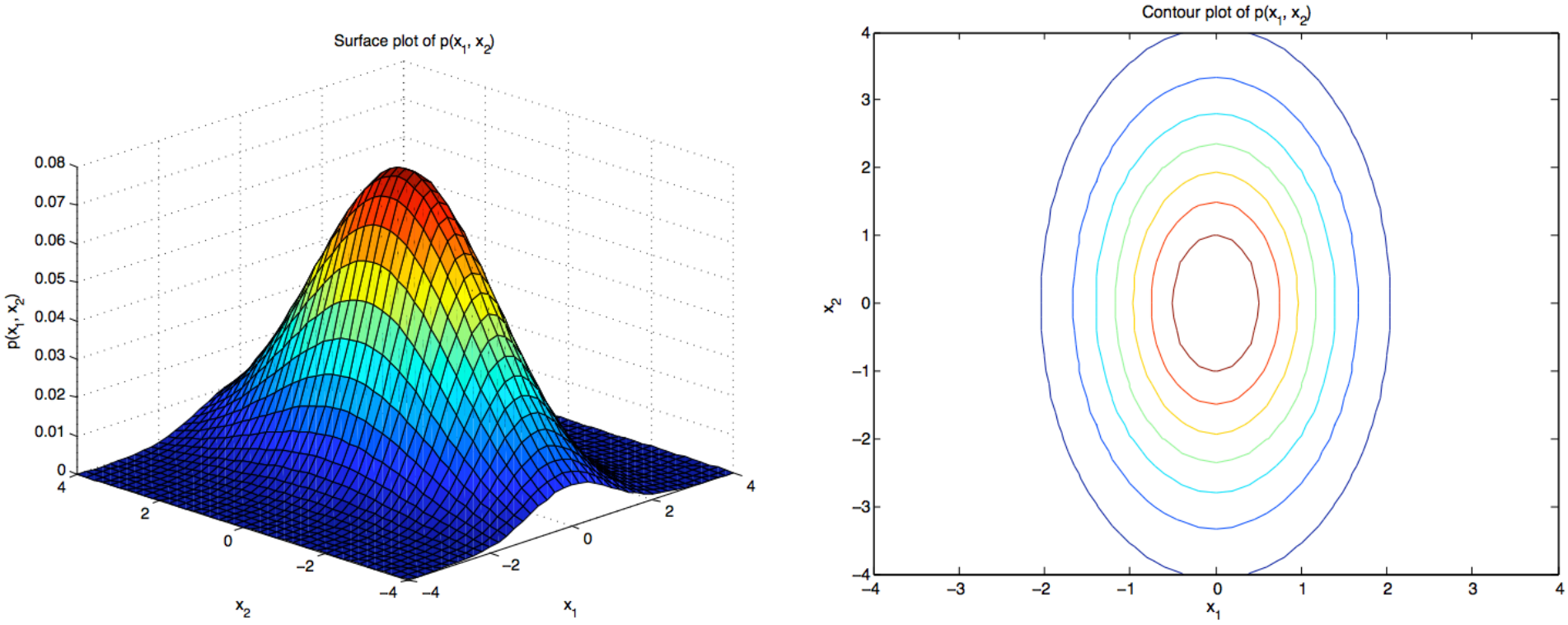
Spherical Gaussian





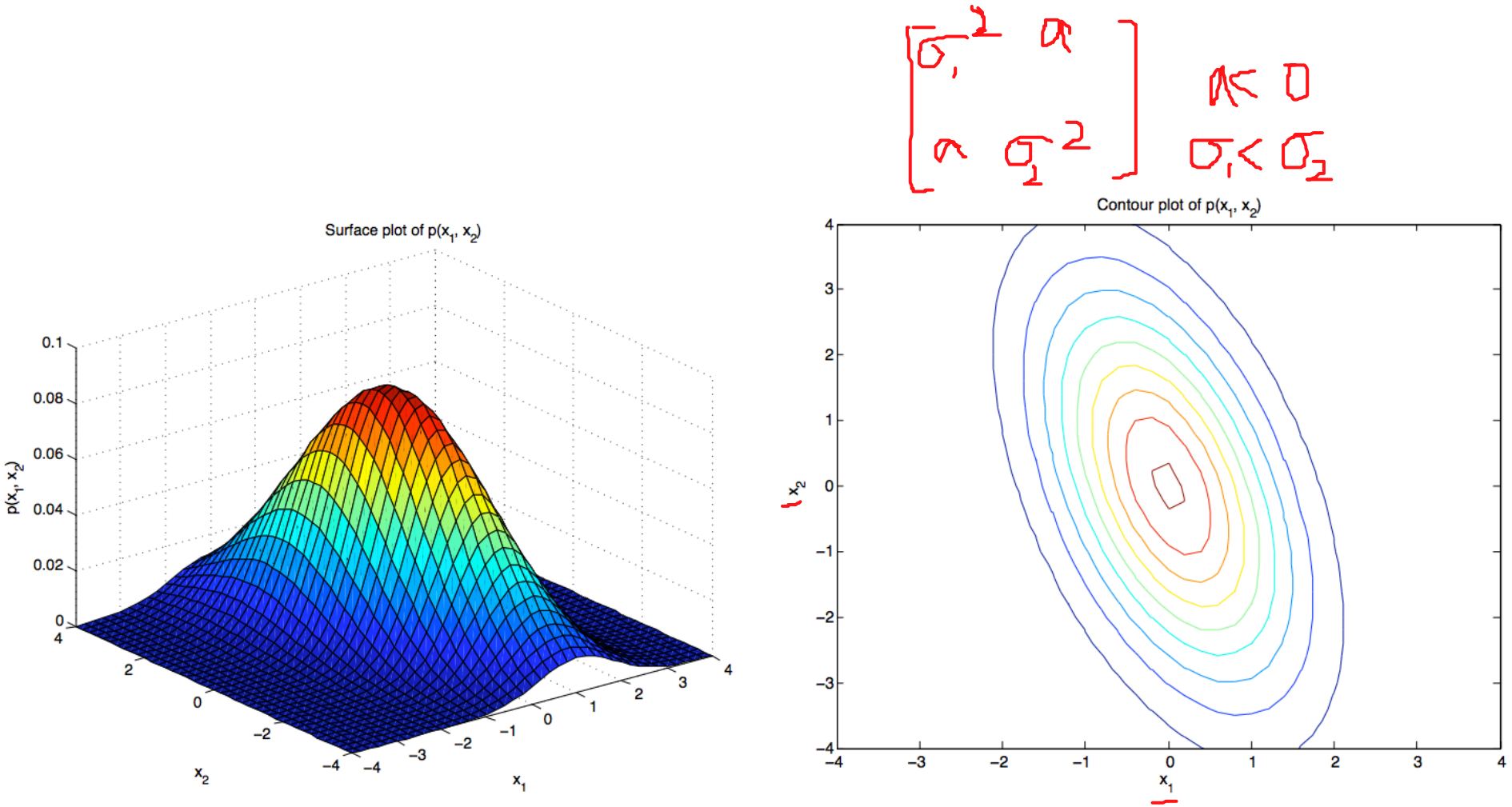
(a) Spherical Gaussian (diagonal covariance, equal variances)

Gaussian with diagonal covariance matrix (variance not equal for different x_i)



(b) Gaussian with diagonal covariance matrix

Gaussian with full covariance matrix



(c) Gaussian with full covariance matrix

6.5.1 Marginals and Conditionals of Gaussians are Gaussians

- Let X and Y be two multivariate random variables that may have different dimensions.
- We write the Gaussian distribution in terms of the concatenated states $[\boldsymbol{x}, \boldsymbol{y}]^{\mathrm{T}}$,

$$p(\boldsymbol{x}, \boldsymbol{y}) = \mathcal{N}\left(\begin{bmatrix}\boldsymbol{\mu}_x \\ \boldsymbol{\mu}_y\end{bmatrix}, \begin{bmatrix}\boldsymbol{\Sigma}_{xx} & \boldsymbol{\Sigma}_{xy} \\ \boldsymbol{\Sigma}_{yx} & \boldsymbol{\Sigma}_{yy}\end{bmatrix}\right)$$

where $\boldsymbol{\Sigma}_{xx} = \mathrm{Cov}[\boldsymbol{x}, \boldsymbol{x}]$ and $\boldsymbol{\Sigma}_{yy} = \mathrm{Cov}[\boldsymbol{y}, \boldsymbol{y}]$ are the marginal covariance matrices of $\boldsymbol{x}$ and $\boldsymbol{y}$, respectively, and $\boldsymbol{\Sigma}_{xy} = \mathrm{Cov}[\boldsymbol{x}, \boldsymbol{y}]$ is the cross-covariance matrix between $\boldsymbol{x}$ and $\boldsymbol{y}$.

- The conditional distribution $p(\boldsymbol{x} \mid \boldsymbol{y})$ is also Gaussian and given by

$$p(\boldsymbol{x} \mid \boldsymbol{y}) = \mathcal{N}\left(\boldsymbol{\mu}_{x|y}, \boldsymbol{\Sigma}_{x|y}\right)$$
$$\boldsymbol{\mu}_{x|y} = \boldsymbol{\mu}_x + \boldsymbol{\Sigma}_{xy}\boldsymbol{\Sigma}_{yy}^{-1}(\boldsymbol{y} - \boldsymbol{\mu}_y)$$
$$\boldsymbol{\Sigma}_{x|y} = \boldsymbol{\Sigma}_{xx} - \boldsymbol{\Sigma}_{xy}\boldsymbol{\Sigma}_{yy}^{-1}\boldsymbol{\Sigma}_{yx}$$

- The marginal distribution $p(\boldsymbol{x})$ of a joint Gaussian distribution $p(\boldsymbol{x}, \boldsymbol{y})$ is itself Gaussian and computed by applying the sum rule and given by

$$p(\boldsymbol{x}) = \int p(\boldsymbol{x}, \boldsymbol{y}) d\boldsymbol{y} = \mathcal{N}(\boldsymbol{x} \mid \boldsymbol{\mu}_x, \boldsymbol{\Sigma}_{xx})$$

- Consider the bivariate Gaussian distribution

$$p(x_1, x_2) = \mathcal{N}\left(\begin{bmatrix}0\\2\end{bmatrix}, \begin{bmatrix}0.3 & -1\\ -1 & 5\end{bmatrix}\right)$$

- We can compute the parameters of the univariate Gaussian, conditioned on $x_2 = -1$, to obtain the mean and variance respectively.

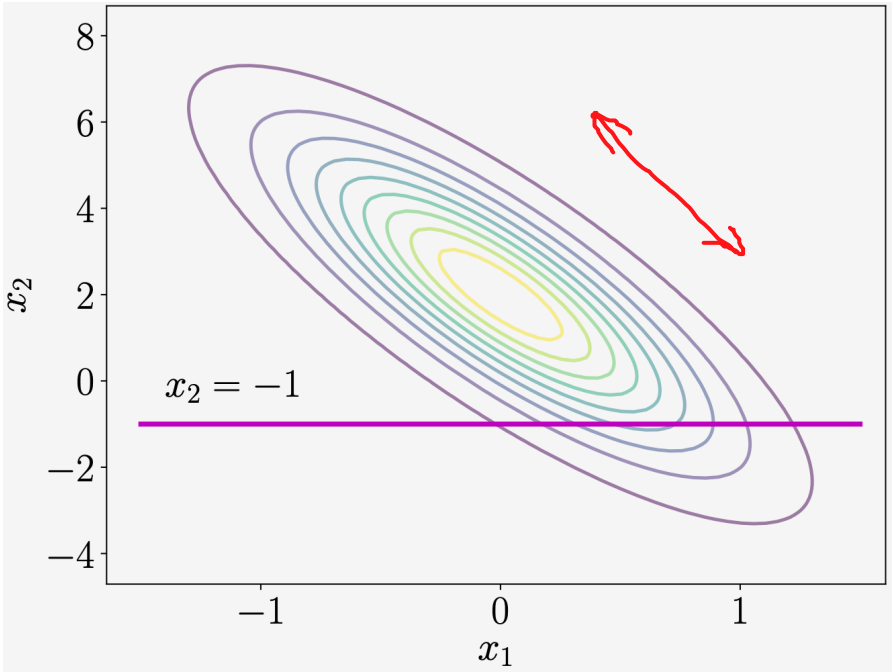
$$\boldsymbol{\mu}_{x|y} = \boldsymbol{\mu}_x + \boldsymbol{\Sigma}_{xy}\boldsymbol{\Sigma}_{yy}^{-1}(\boldsymbol{y} - \boldsymbol{\mu}_y)$$
$$\boldsymbol{\Sigma}_{x|y} = \boldsymbol{\Sigma}_{xx} - \boldsymbol{\Sigma}_{xy}\boldsymbol{\Sigma}_{yy}^{-1}\boldsymbol{\Sigma}_{yx}$$

- Numerically, this is

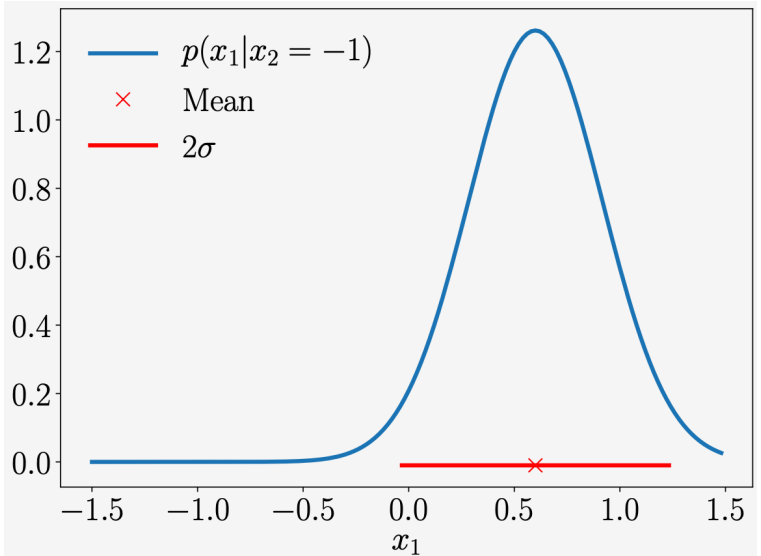
$$\mu_{x_1|x_2=-1} = 0 + (-1)(0.2)(-1-2) = 0.6$$
$$\sigma^2_{x_1|x_2=-1} = 0.3 - (-1)(0.2)(-1) = 0.1$$

- Therefore, the conditional Gaussian is given by

$$p(x_1 \mid x_2 = -1) = \mathcal{N}(0.6, 0.1)$$



$$\boldsymbol{\mu}_{x|y} = \boldsymbol{\mu}_x + \boldsymbol{\Sigma}_{xy}\boldsymbol{\Sigma}_{yy}^{-1}(\boldsymbol{y} - \boldsymbol{\mu}_y)$$
$$\boldsymbol{\Sigma}_{x|y} = \boldsymbol{\Sigma}_{xx} - \boldsymbol{\Sigma}_{xy}\boldsymbol{\Sigma}_{yy}^{-1}\boldsymbol{\Sigma}_{yx}$$



- Consider the bivariate Gaussian distribution

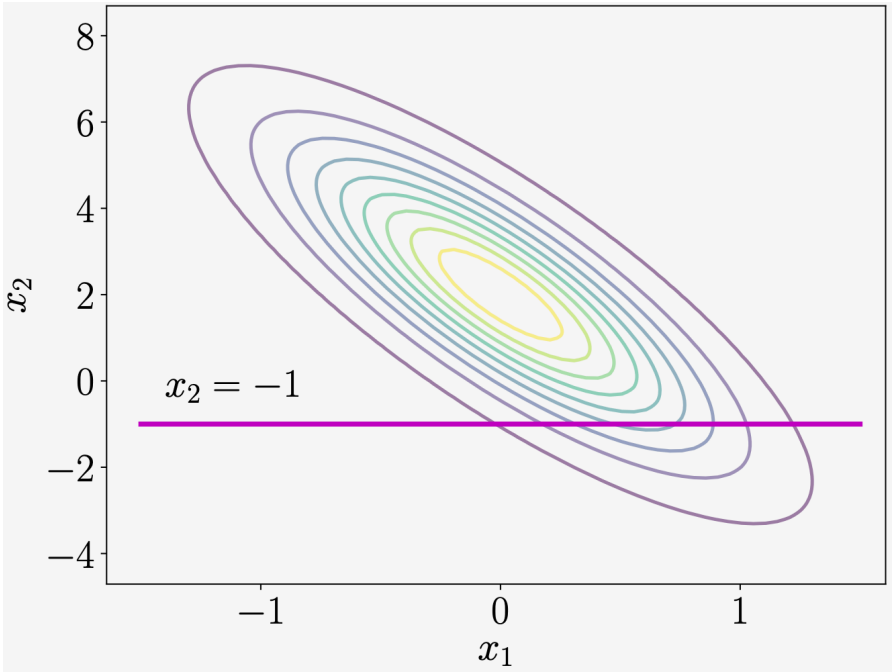
$$p(x_1, x_2) = \mathcal{N}\left(\begin{bmatrix}0\\2\end{bmatrix}, \begin{bmatrix}0.3 & -1\\-1 & 5\end{bmatrix}\right)$$

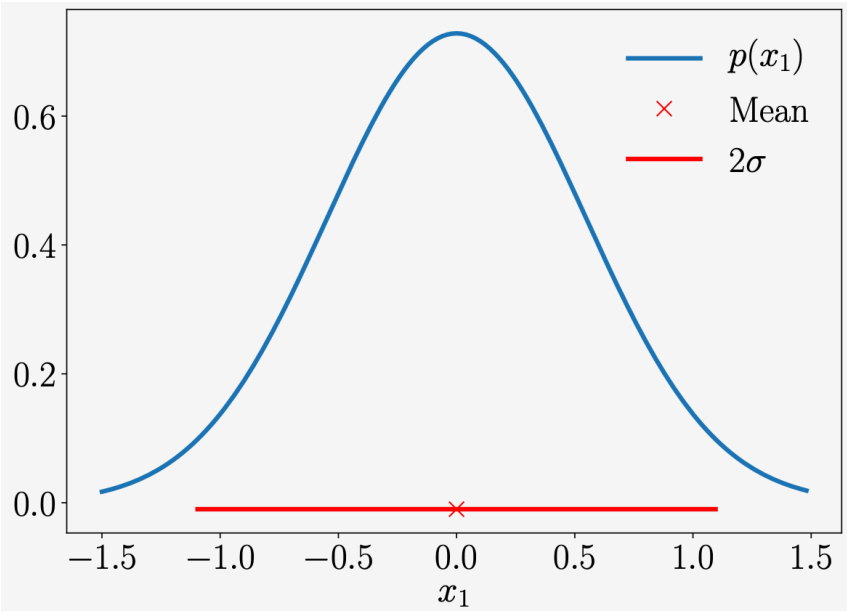
- The marginal distribution $p(x_1)$ can be obtained by

$$p(\boldsymbol{x}) = \int p(\boldsymbol{x}, \boldsymbol{y}) d\boldsymbol{y} = \mathcal{N}(\boldsymbol{x} \,|\, \boldsymbol{\mu}_x, \boldsymbol{\Sigma}_{xx})$$

which is essentially using the mean and variance of the random variable x_1. So we have,

$$p(x_1) = \mathcal{N}(0, 0.3)$$





Check your understanding

- Covariance (correlation) has two directions, i.e., negative and positive values have different meanings), while variance does not.

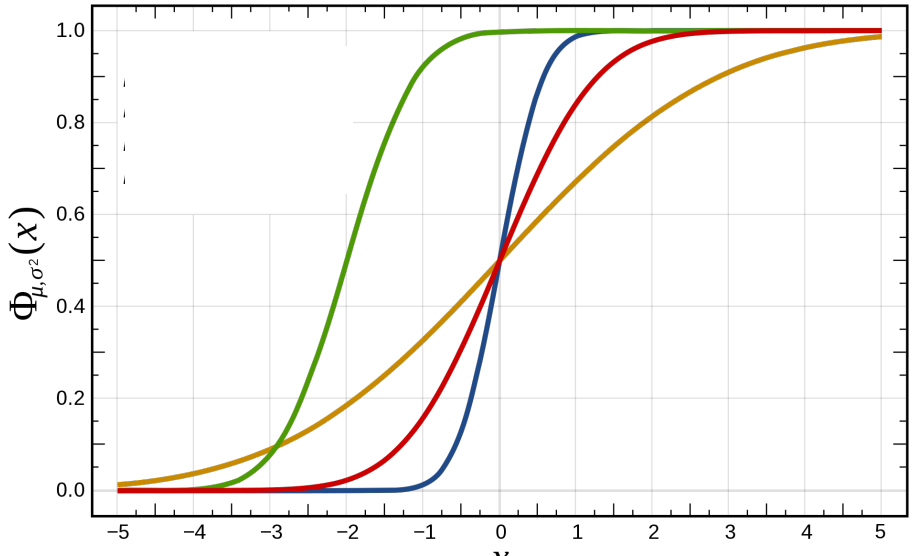
of univariate random variables

- When the covariance of two variables equals to the sum of their individual covariances, the two variables are independent.

non-linear dependent

- The empirical mean $\frac{1}{N}\sum_{n=1}^{N} \boldsymbol{x}_n$ approximates the expected value $\int_{\mathcal{X}} g(\boldsymbol{x})p(\boldsymbol{x})d\boldsymbol{x}$ of a random variable when N is large.

- In the figure (cumulative distribution function, cdf), which Gaussian has the largest variance?
  - Green? Blue? Red? Yellow?
- What is the mean of the red Gaussian?
- The cdf terminates at →1.
- The cdf starts from → 0.
- The cdf always increases.



https://en.wikipedia.org/wiki/Normal_distribution