

Probability and Distributions

Liang Zheng

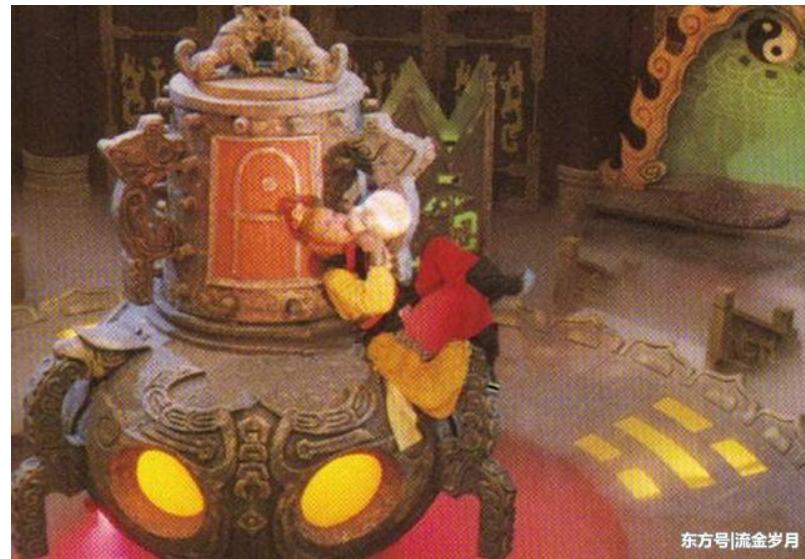
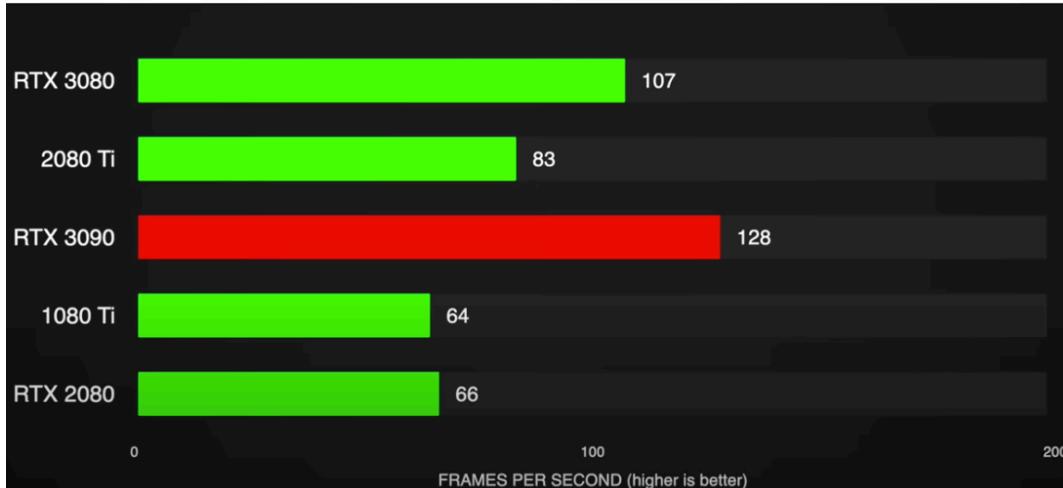
Australian National University

liang.zheng@anu.edu.au

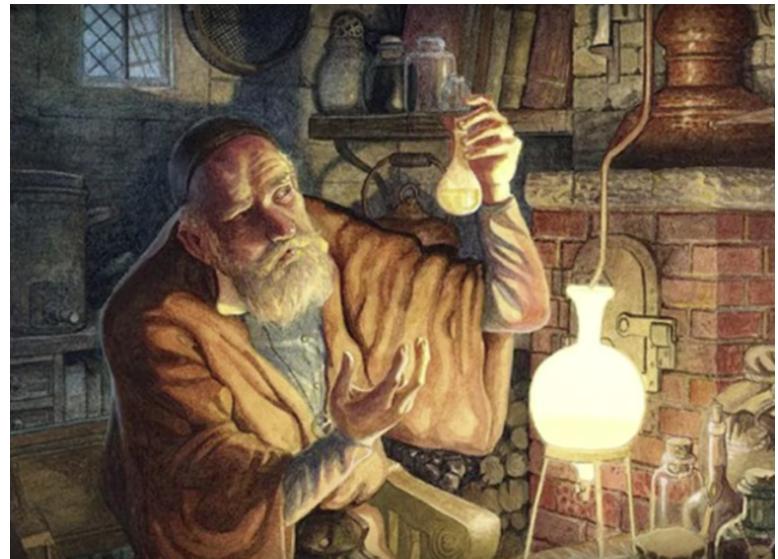
NVIDIA launches GeForce RTX 3090, 3080 and 3070 with Ampere

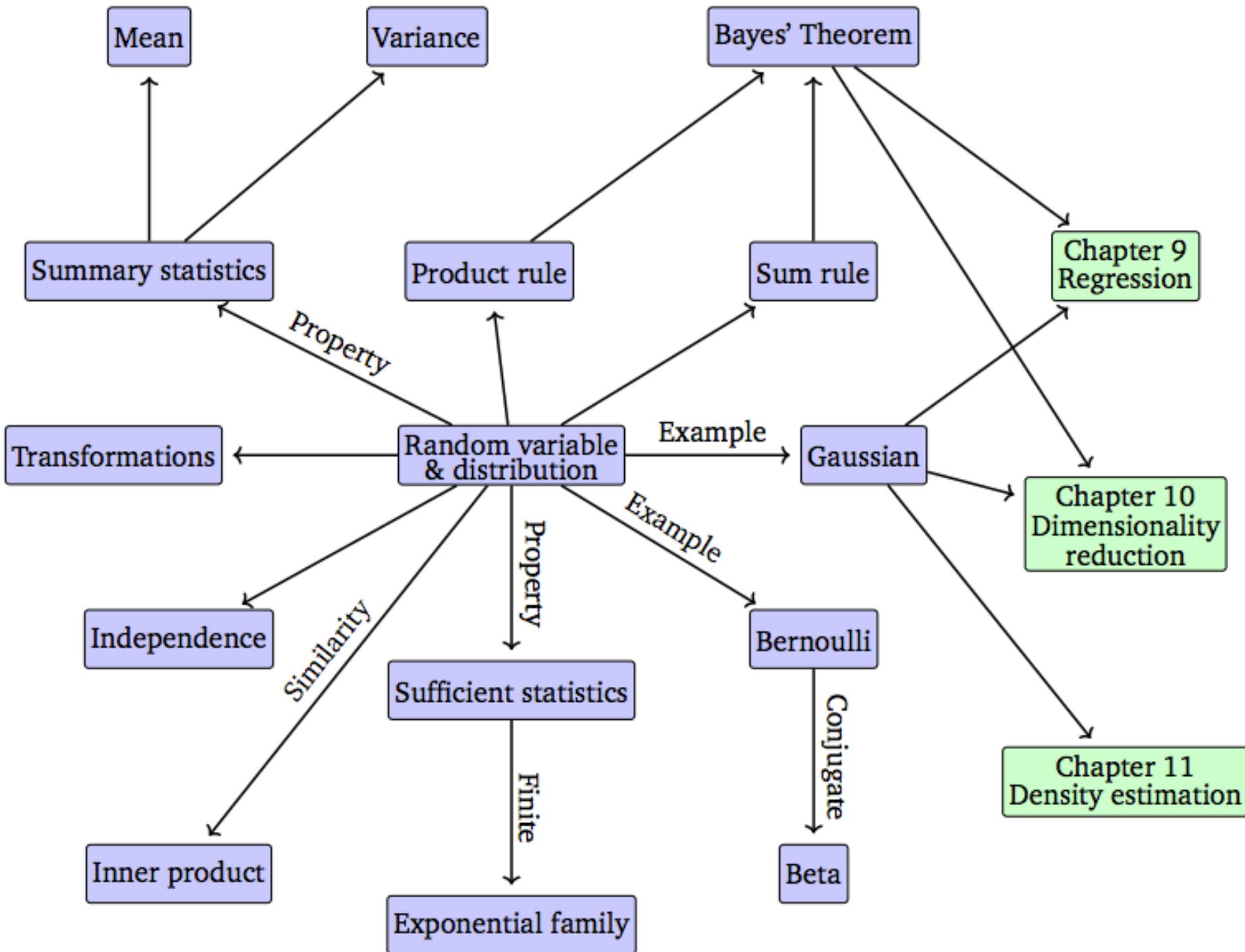
Death Stranding - 4K

Very High, TAA | Ryzen 9 3950X



东方号|流金岁月





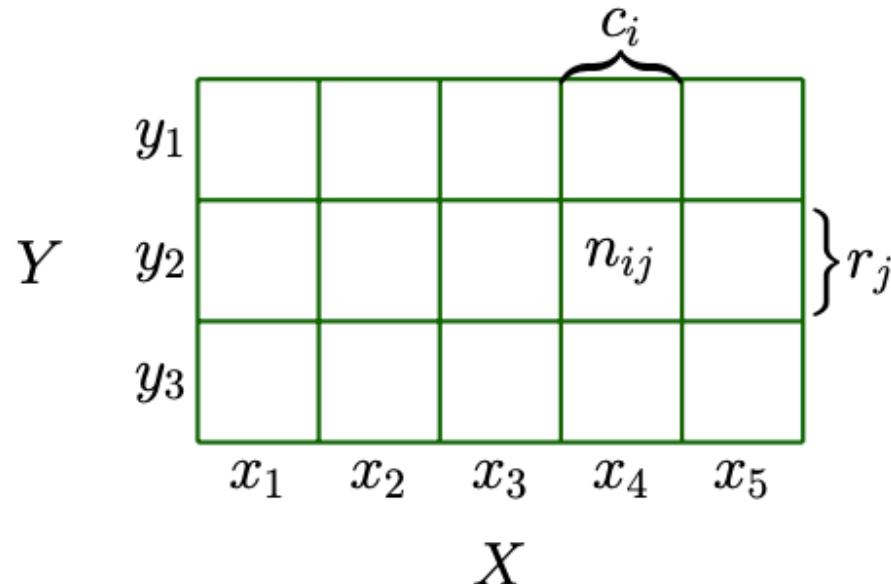
6.2.1 Discrete Probabilities

- 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

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

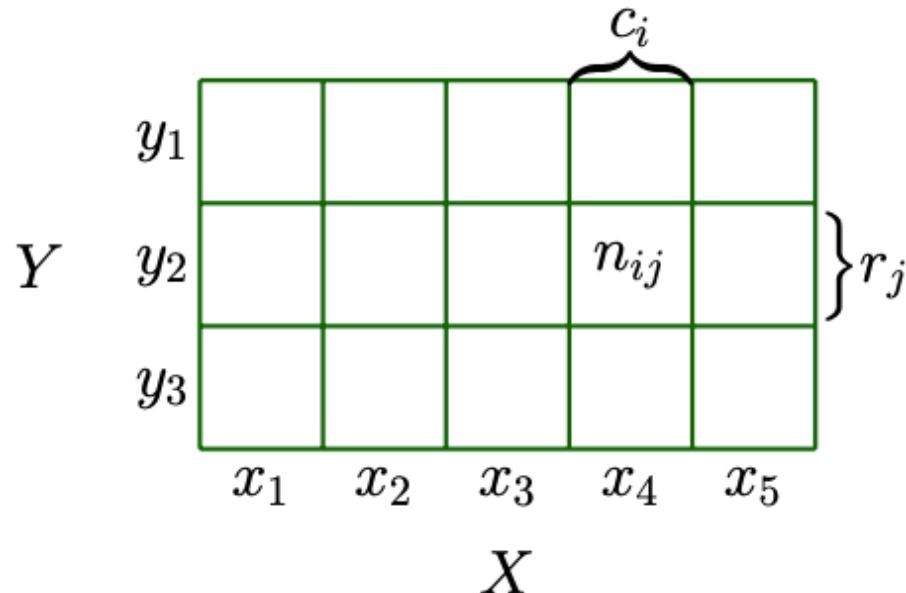
n_{ij} is the number of events with state x_i and y_j and N total number of events

- 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

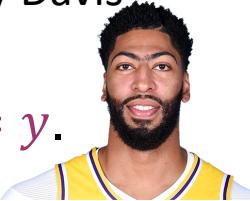


Lebron James

	13	6
	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 i th 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

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

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

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

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

- and the conditional probability of X given Y is

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



	13	6
≥ 30		
4		17
< 30		

Anthony Davis

Lebron James



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

6.2.2 Continuous Probabilities

- A function $f: \mathbb{R}^D \rightarrow \mathbb{R}$ is called a **probability density function (pdf)** if
$$\forall x \in \mathbb{R}^D: f(x) \geq 0$$
- Its integral exists and

$$\int_{\mathbb{R}^D} f(x) dx = 1.$$

- 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 $\mathbf{x} \in \mathbb{R}^D$ is given by

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

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

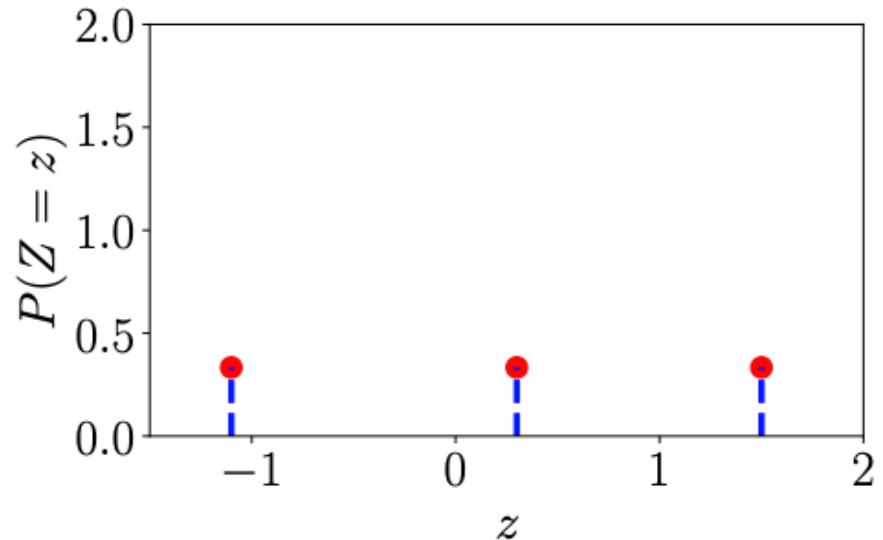
- The cdf can be expressed also as the integral of the probability density function $f(\mathbf{x})$ so that

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

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
$P(Z = z)$	$\frac{1}{3}$	$\frac{1}{3}$	$\frac{1}{3}$



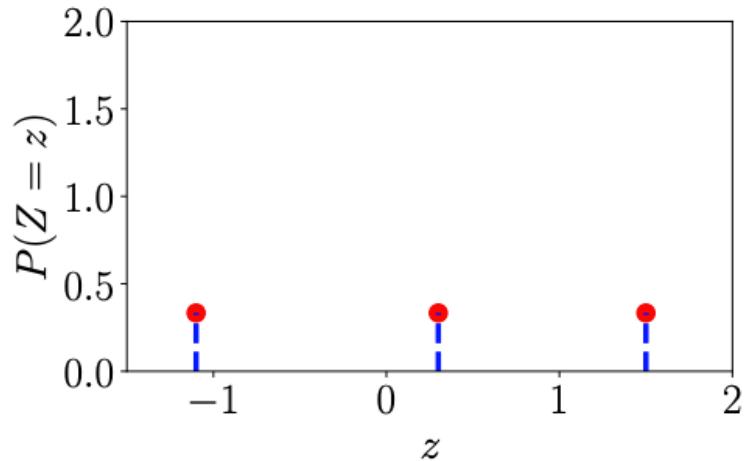
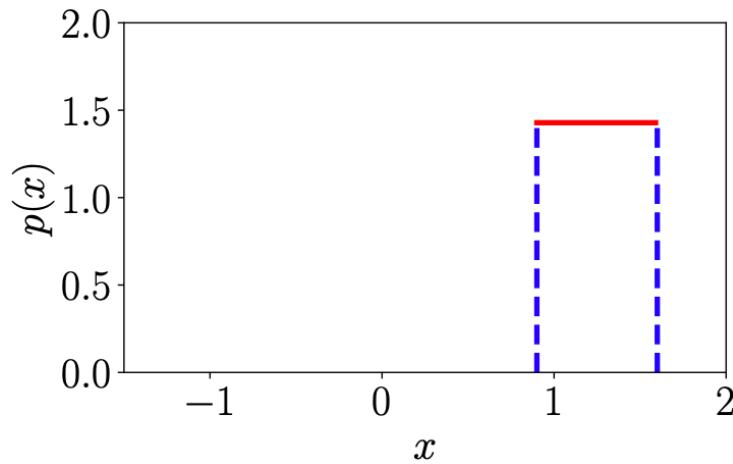
- States can be located on the x -axis, and the y -axis represents the probability of a particular state

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 $U(a, b)$

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

6.3 Sum Rule, Product Rule, and Bayes' Theorem

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

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

where \mathcal{Y} are the states of the target space of random variable \mathbf{Y} .

- We sum out (or integrate out) the set of states \mathbf{y} of the random \mathbf{Y}
- The sum rule is also known as the **marginalization property**.
- If $\mathbf{x} = [x_1, \dots, x_D]^T$, we obtain the marginal

$$p(x_i) = \int p(x_1, \dots, x_D) dx_{\setminus i}$$

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

6.3 Sum Rule, Product Rule, and Bayes' Theorem

- The **product rule** relates the joint distribution to the conditional distribution
$$p(\mathbf{x}, \mathbf{y}) = p(\mathbf{y} | \mathbf{x})p(\mathbf{x})$$
- Every joint distribution of two random variables can be factorized (written as a product) of two other distributions
- The product rule also implies

$$p(\mathbf{x}, \mathbf{y}) = p(\mathbf{x} | \mathbf{y})p(\mathbf{y})$$

6.3 Sum Rule, Product Rule, and Bayes' Theorem

- We are often interested in making inferences of unobserved (latent) random variables given that we have observed other random variables.
- Let us assume we have some prior knowledge $p(x)$ about an unobserved random variable x and some relationship $p(y | x)$ between x and a second random variable y , which we can observe. If we observe y , we can use **Bayes' theorem** (also *Bayes' rule* or *Bayes' law*) to draw some conclusions about x given the observed values of y .

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

is a direct consequence of the product rule, since

$$p(x, y) = p(x | y)p(y)$$

and

$$p(x, y) = p(y | x)p(x)$$

so that

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

6.3 Sum Rule, Product Rule, and Bayes' Theorem

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

- $p(x)$ is the **prior**, which encapsulates our subjective prior knowledge of the unobserved (latent) variable x before observing any data
- $p(y | x)$, the **likelihood**, describes how x and y are related
- $p(y | x)$ is the probability of the data y if we were to know the latent variable x
- We call $p(y | x)$ either the “likelihood of x (given y)” or the “probability of y given x ” (y is observed; x is latent)
- $p(x | y)$, 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 x after having observed y (e.g., linear regression or Gaussian mixture models)

6.3 Sum Rule, Product Rule, and Bayes' Theorem

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

- The quantity

$$p(y) := \int p(y | x)p(x) dx = \mathbb{E}_X[p(y | x)]$$

- is the *marginal likelihood/evidence*.
- The marginal likelihood integrates the numerator with respect to the latent variable x

$$p(x | y) = \frac{p(y | x)p(x)}{p(y)}$$

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

	≥ 30	13	6
Lebron James			
	< 30	4	17
		< 30	≥ 30
			Anthony Davis



6.4 Summary Statistics and Independence

6.4.1 Means and Covariances

- The **expected value** of a function $g : \mathbb{R} \rightarrow \mathbb{R}$ 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 $[X_1, \dots, X_n]^T$. For multivariate random variables, we define the expected value element wise

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

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

6.4.1 Means and Covariances

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

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

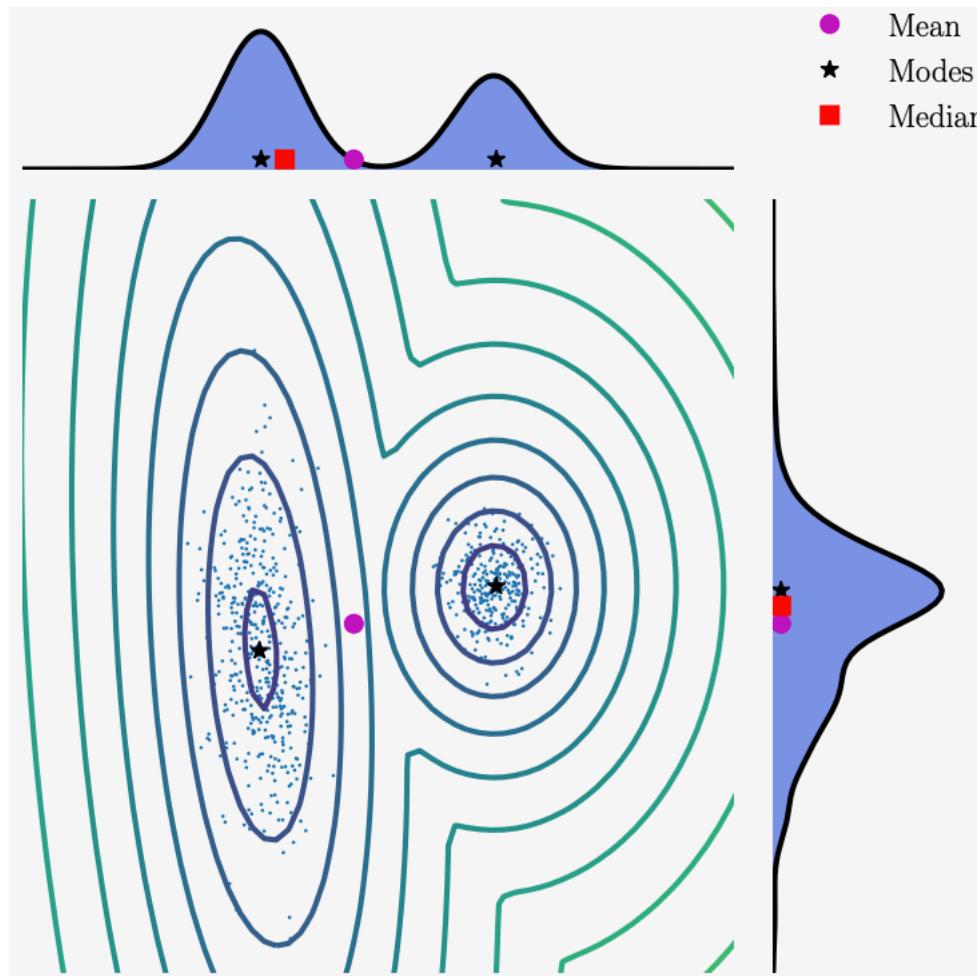
where

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

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

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

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



6.4.1 Means and Covariances

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

$$\begin{aligned}\mathbb{E}_X[f(\mathbf{x})] &= \int f(\mathbf{x})p(\mathbf{x})d\mathbf{x} \\ &= \int [ag(\mathbf{x}) + bh(\mathbf{x})]p(\mathbf{x})d\mathbf{x} \\ &= a \int g(\mathbf{x})p(\mathbf{x})d\mathbf{x} + b \int h(\mathbf{x})p(\mathbf{x})d\mathbf{x} \\ &= a\mathbb{E}_X[g(\mathbf{x})] + b\mathbb{E}_X[h(\mathbf{x})]\end{aligned}$$

6.4.1 Means and Covariances

- The covariance 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.,

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

- 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.,

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

- The covariance of a variable with itself $\text{Cov}[x, x]$ is called the variance and is denoted by $\text{V}_X[x]$.

- 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 $\mathbf{x} \in \mathbb{R}^D$ and $\mathbf{y} \in \mathbb{R}^E$ respectively, the covariance between X and Y is defined as

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

6.4.1 Means and Covariances

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

$$\begin{aligned}\mathbb{V}_X[\mathbf{x}] &= \text{Cov}_X[\mathbf{x}, \mathbf{x}] \\ &= \mathbb{E}_X[(\mathbf{x} - \boldsymbol{\mu})(\mathbf{x} - \boldsymbol{\mu})^T] = \mathbb{E}_X[\mathbf{x}\mathbf{x}^T] - \mathbb{E}_X[\mathbf{x}]\mathbb{E}_X[\mathbf{x}]^T\end{aligned}$$

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

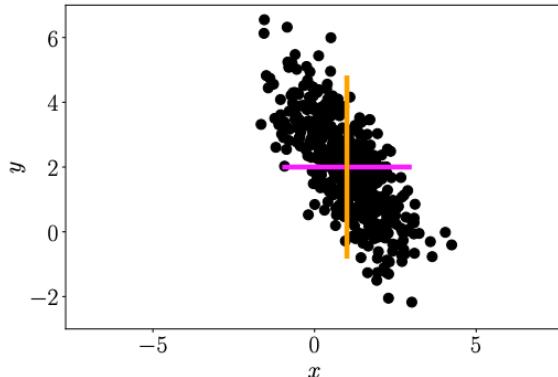
- The $D \times D$ matrix is called the **covariance matrix** of the multivariate random variable X . The covariance matrix is symmetric and positive definite and tells us something about the spread of the data.
- On its diagonal, the covariance matrix contains the variances of x_i

6.4.1 Means and Covariances

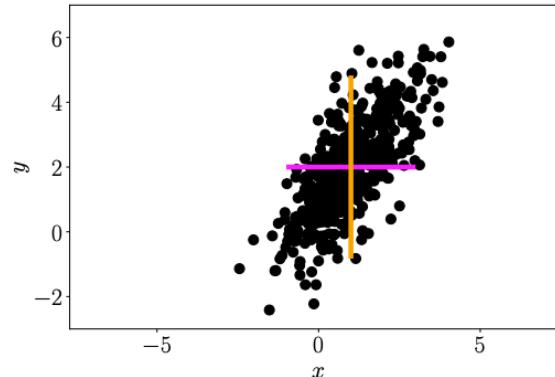
- The **correlation** between two random variables X, Y is given by

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

- 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
- 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

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

Where $x_n \in \mathbb{R}^D$.

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

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

- To compute the statistics for a particular dataset, we would use the realizations (observations) x_1, \dots, x_N and use the two equations above. Empirical covariance matrices are symmetric, positive semi-definite.

Example - Computation of covariant matrix

- $X = [x_1, x_2] = \begin{bmatrix} 1 & 3 \\ 2 & 1 \\ 3 & 1 \end{bmatrix} \begin{array}{l} x_1 \\ x_2 \\ x_3 \end{array}$
- $\mu = \begin{bmatrix} 2 \\ 1.5 \\ 2 \end{bmatrix};$
- $X - \mu = \begin{bmatrix} -1 & 1 \\ 0.5 & -0.5 \\ 1 & -1 \end{bmatrix} \begin{array}{l} x_1 \\ x_2 \\ x_3 \end{array}$
- $\Sigma = \frac{1}{2}(X - \mu)(X - \mu)^T = \frac{1}{2} \begin{bmatrix} 2 & -1 & -2 \\ -1 & 0.5 & 1 \\ -2 & 1 & 2 \end{bmatrix} \begin{array}{l} x_1 \\ x_2 \\ x_3 \end{array}$

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, \dots, 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:

$$\frac{1}{N^2} \sum_{i,j=1}^N (x_i - x_j)^2 = \dots = 2 \left[\frac{1}{N} \sum_{i=1}^N x_i^2 - \left(\frac{1}{N} \sum_{i=1}^N x_i \right)^2 \right] \text{(lab question)}$$

6.4.4 Sums and Transformations of Random Variables

- Consider two random variables X, Y with states $\mathbf{x}, \mathbf{y} \in \mathbb{R}^D$. Then:
 - $\mathbb{E}[\mathbf{x} + \mathbf{y}] = \mathbb{E}[\mathbf{x}] + \mathbb{E}[\mathbf{y}]$
 - $\mathbb{E}[\mathbf{x} - \mathbf{y}] = \mathbb{E}[\mathbf{x}] - \mathbb{E}[\mathbf{y}]$
 - $\mathbb{V}[\mathbf{x} + \mathbf{y}] = \mathbb{V}[\mathbf{x}] + \mathbb{V}[\mathbf{y}] + \text{Cov}[\mathbf{x}, \mathbf{y}] + \text{Cov}[\mathbf{y}, \mathbf{x}]$ can you prove it?
 - $\mathbb{V}[\mathbf{x} - \mathbf{y}] = \mathbb{V}[\mathbf{x}] + \mathbb{V}[\mathbf{y}] - \text{Cov}[\mathbf{x}, \mathbf{y}] - \text{Cov}[\mathbf{y}, \mathbf{x}]$
- 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 $\mathbf{y} = A\mathbf{x} + \mathbf{b}$ of \mathbf{x} . Then \mathbf{y} is itself a random variable whose mean vector and covariance matrix are given by
 - $\mathbb{E}_Y[\mathbf{y}] = \mathbb{E}_X[A\mathbf{x} + \mathbf{b}] = A\mathbb{E}_X[\mathbf{x}] + \mathbf{b} = A\boldsymbol{\mu} + \mathbf{b}$
 - $\mathbb{V}_Y[\mathbf{y}] = \mathbb{V}_X[A\mathbf{x} + \mathbf{b}] = \mathbb{V}_X[A\mathbf{x}] = A\mathbb{V}_X[\mathbf{x}]A^T = A\Sigma A^T$ can you prove it?
 - Furthermore,
 - $\text{Cov}[\mathbf{x}, \mathbf{y}] = \mathbb{E}[\mathbf{x}(A\mathbf{x} + \mathbf{b})^T] - \mathbb{E}[\mathbf{x}]\mathbb{E}[A\mathbf{x} + \mathbf{b}]^T$
 - $= \mathbb{E}[\mathbf{x}]\mathbf{b}^T + \mathbb{E}[\mathbf{x}\mathbf{x}^T]A^T - \boldsymbol{\mu}\mathbf{b}^T - \boldsymbol{\mu}\boldsymbol{\mu}^T A^T$
 - $= \boldsymbol{\mu}\mathbf{b}^T - \boldsymbol{\mu}\mathbf{b}^T + (\mathbb{E}[\mathbf{x}\mathbf{x}^T] - \boldsymbol{\mu}\boldsymbol{\mu}^T)A^T$
 - $= \boldsymbol{\Sigma}A^T$
 - where $\boldsymbol{\Sigma} = \mathbb{E}[\mathbf{x}\mathbf{x}^T] - \boldsymbol{\mu}\boldsymbol{\mu}^T$ is the covariance of X .

6.4.5 Statistical Independence

- Two random variables X, Y are **statistically independent** if and only if
$$p(x, y) = p(x)p(y)$$
- Intuitively, two random variables X and Y are independent if the value of y (once known) does not add any additional information about x (and vice versa). If X, Y are (statistically) independent, then

$$p(y | x) = p(y)$$

$$p(x | y) = p(x)$$

$$\text{Cov}_{X,Y}[x, y] = 0$$

$$\mathbb{V}_{X,Y}[x + y] = \mathbb{V}_X[x] + \mathbb{V}_Y[y]$$

- 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

$$\text{Cov}[x, y] = \mathbb{E}[xy] - \mathbb{E}[x]\mathbb{E}[y] = \mathbb{E}[x^3] = 0$$

6.4.5 Statistical Independence

- Two random variables X and Y are **conditionally independent** given Z if and only if

$$p(x, y|z) = p(x|z)p(y|z) \quad \text{for all } z \in \mathcal{Z} \quad (1)$$

where \mathcal{Z} is the set of states of random variable Z . We write $X \perp\!\!\!\perp Y|Z$ to denote that X is conditionally independent of Y given Z .

- By using the product rule of probability, we can expand the left-hand side of the equation above to obtain

$$p(x, y|z) = p(x|y, z)p(y|z) \quad (2)$$

- By comparing the right-hand side of (1) with (2), we see that $p(y|z)$ appears in both of them so that

$$p(x|y, z) = p(x|z)$$

- This equation provides an alternative definition of conditional independence, i.e., $X \perp\!\!\!\perp Y|Z$. This alternative presentation provides the interpretation “given that we know z , knowledge about y does not change our knowledge of x ”.

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

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

- The **multivariate Gaussian distribution** is fully characterized by a **mean vector** μ and a **covariance matrix** Σ and defined as

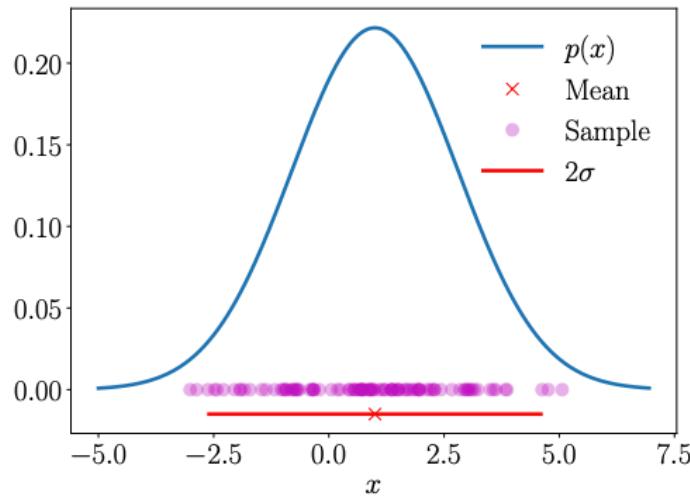
$$p(\mathbf{x}|\mu, \Sigma) = (2\pi)^{-\frac{D}{2}} |\Sigma|^{-\frac{1}{2}} \exp\left(-\frac{1}{2} (\mathbf{x} - \mu)^T \Sigma^{-1} (\mathbf{x} - \mu)\right)$$

where $\mathbf{x} \in \mathbb{R}^D$. We write $p(\mathbf{x}) = \mathcal{N}(\mathbf{x}|\mu, \Sigma)$ or $\mathbf{X} \sim \mathcal{N}(\mu, \Sigma)$.

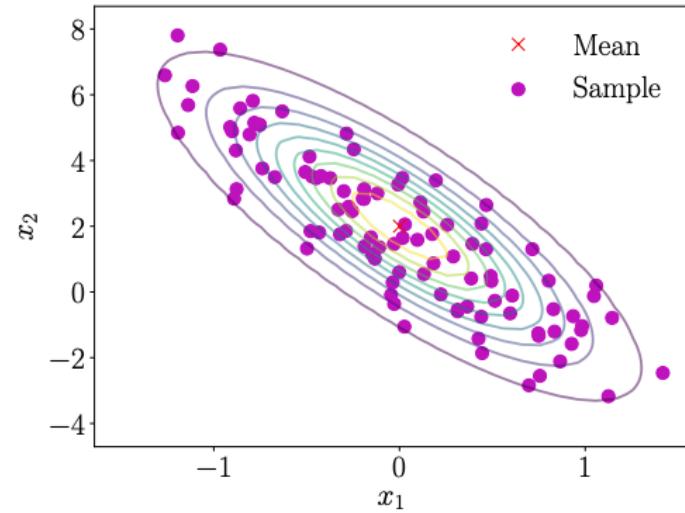
- The special case of the Gaussian with zero mean and identity covariance, that is, $\mu = \mathbf{0}$ and $\Sigma = \mathbf{I}$, is referred to as the **standard normal distribution**.

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

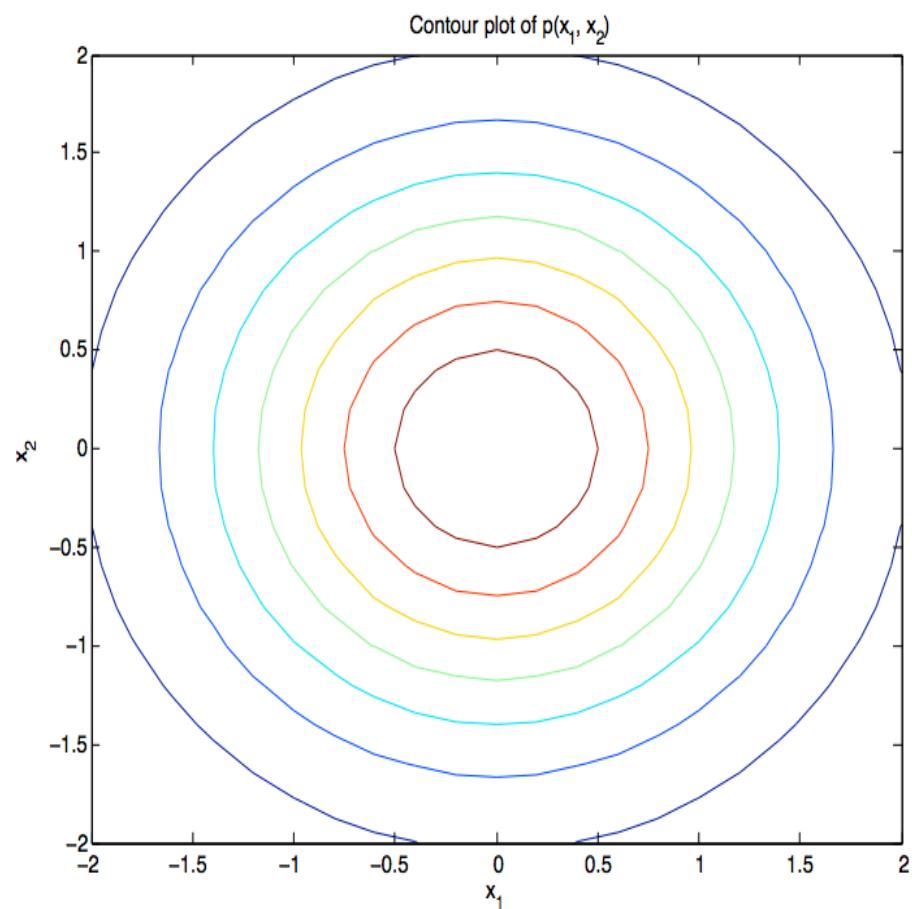
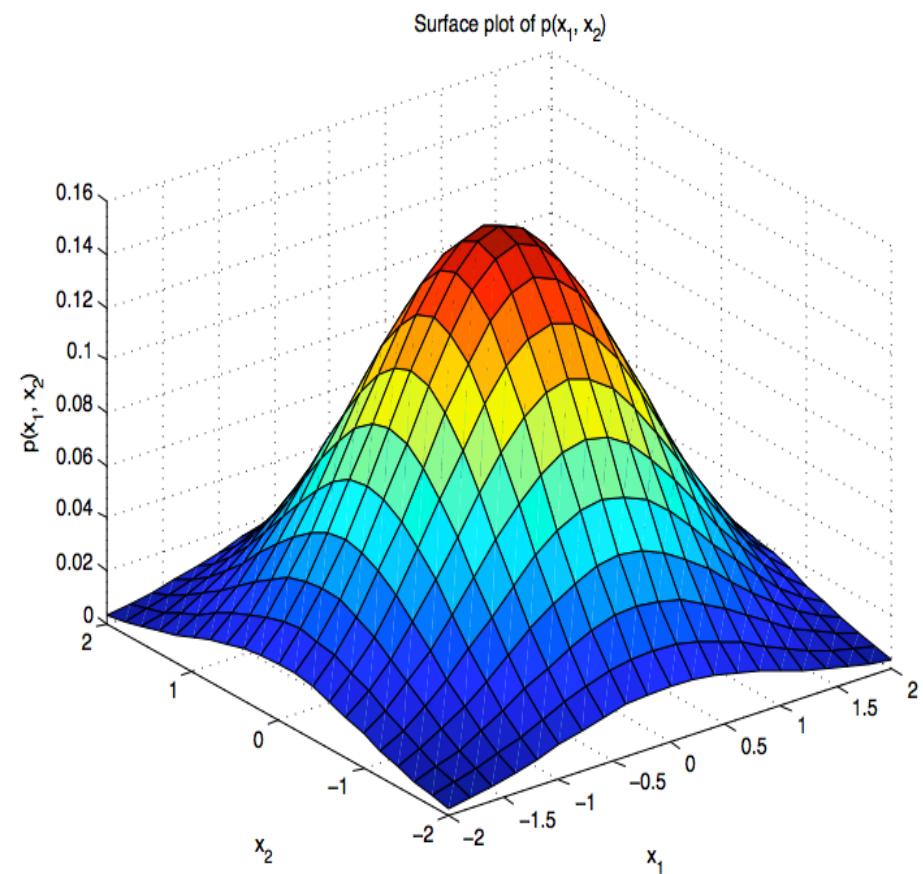
$$p(x|\mu, \Sigma) = (2\pi)^{-\frac{D}{2}} |\Sigma|^{-\frac{1}{2}} \exp\left(-\frac{1}{2}(x - \mu)^T \Sigma^{-1} (x - \mu)\right)$$

- Spherical Gaussian

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

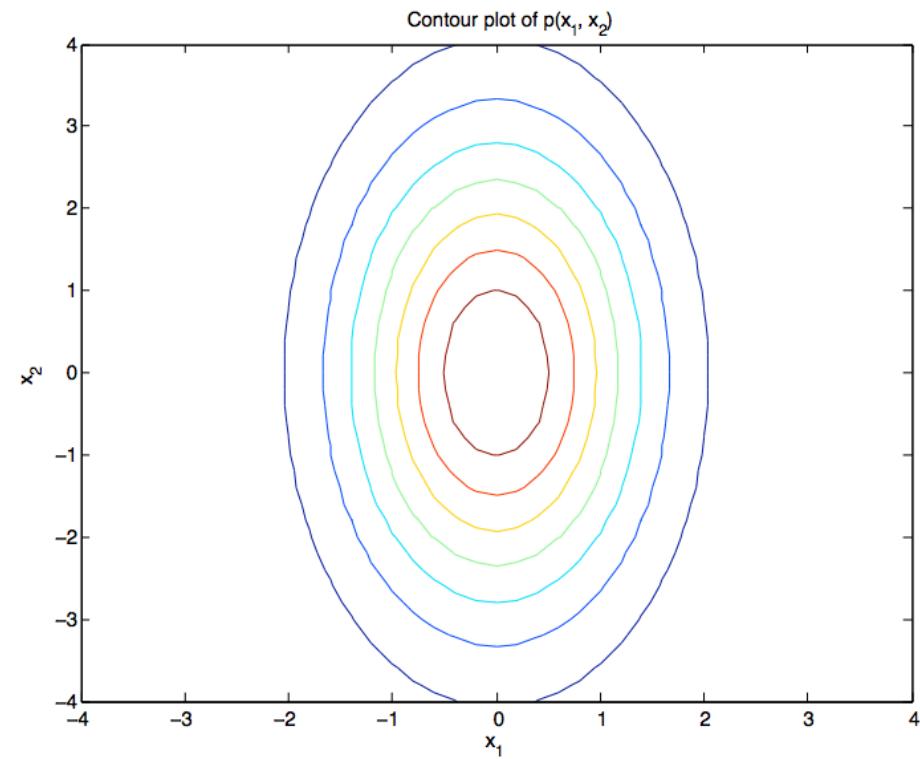
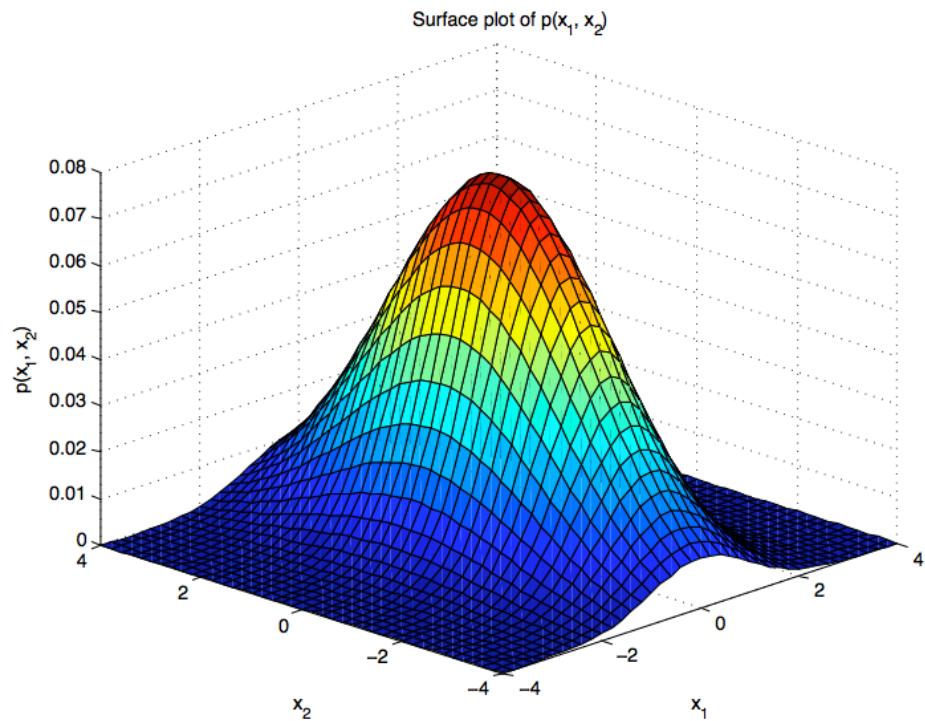
$$\Sigma = \begin{bmatrix} \sigma^2 & 0 & 0 \\ 0 & \sigma^2 & 0 \\ 0 & 0 & \sigma^2 \end{bmatrix} \quad \text{diagonal covariance, equal variances}$$

Spherical Gaussian



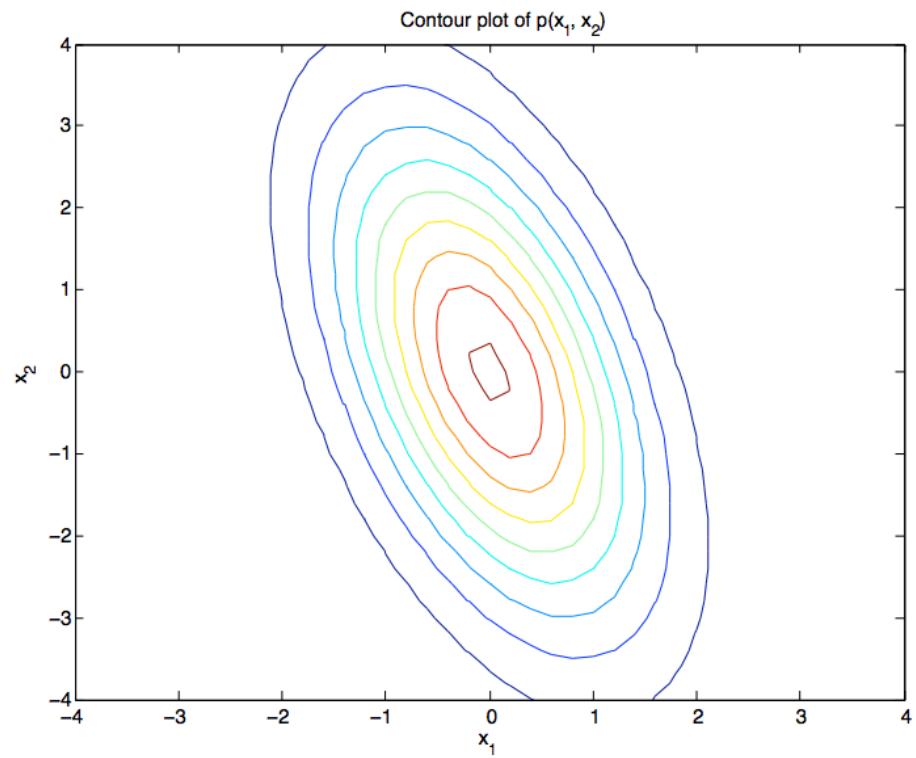
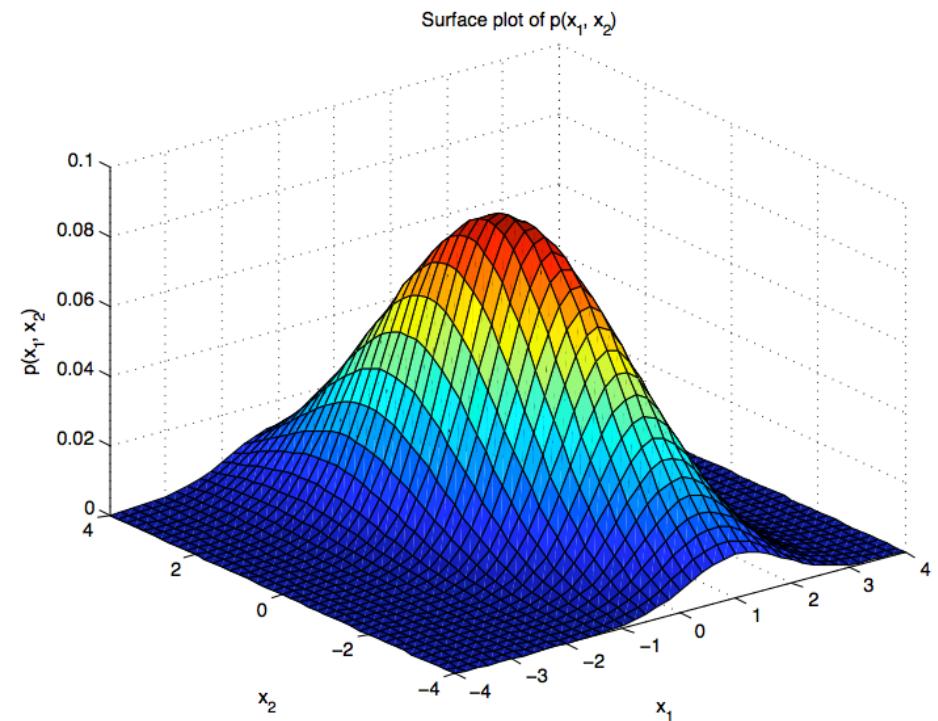
(a) Spherical Gaussian (diagonal covariance, equal variances)

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

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

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

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

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

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

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

- Consider the bivariate Gaussian distribution

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

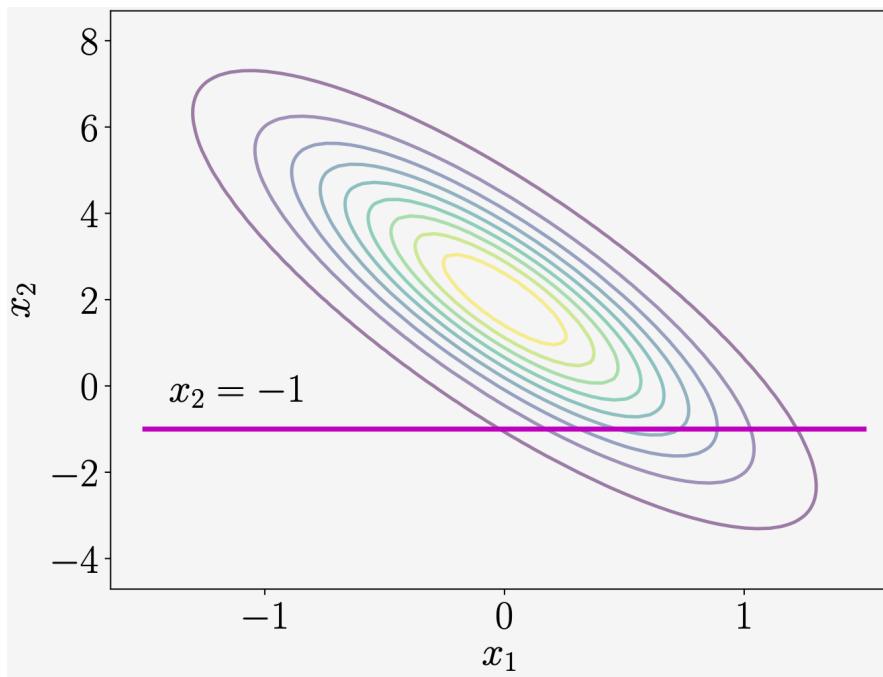
- We can compute the parameters of the univariate Gaussian, conditioned on $x_2 = -1$, to obtain the mean and variance respectively.

- Numerically, this is

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

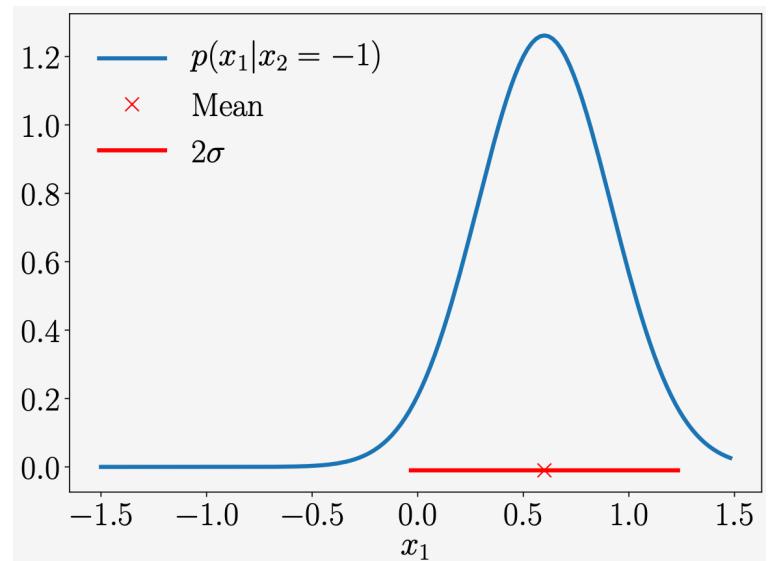
- Therefore, the conditional Gaussian is given by

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



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

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



- 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

$$p(x) = \int p(x, y) dy = \mathcal{N}(x | \mu_x, \Sigma_{xx})$$

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

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

