

Introduction to Linear Algebra

A d -dimensional (column) vector \mathbf{x} and its (row) transpose \mathbf{x}^t can be written as

$$\mathbf{x} = \begin{pmatrix} x_1 \\ x_2 \\ \vdots \\ x_d \end{pmatrix} \quad \text{and} \quad \mathbf{x}^t = (x_1 \ x_2 \ \dots \ x_d),$$

Matrix
representation
of size $n \times d$

$$\mathbf{M} = \begin{pmatrix} m_{11} & m_{12} & m_{13} & \dots & m_{1d} \\ m_{21} & m_{22} & m_{23} & \dots & m_{2d} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ m_{n1} & m_{n2} & m_{n3} & \dots & m_{nd} \end{pmatrix} \text{ and}$$

$$\mathbf{M}^t = \begin{pmatrix} m_{11} & m_{21} & \dots & m_{n1} \\ m_{12} & m_{22} & \dots & m_{n2} \\ m_{13} & m_{23} & \dots & m_{n3} \\ \vdots & \vdots & \ddots & \vdots \\ m_{1d} & m_{2d} & \dots & m_{nd} \end{pmatrix}.$$

Symmetric Matrix

A square $(d \times d)$ matrix is called symmetric if its entries obey $m_{ij} = m_{ji}$;

Diagonal Matrix

A general diagonal matrix (i.e., one having 0 for all off diagonal entries)

Matrix multiplication with a vector gives a transformed vector

We can multiply a vector by a matrix, $\mathbf{M}\mathbf{x} = \mathbf{y}$, i.e.,

$$\begin{pmatrix} m_{11} & m_{12} & \dots & m_{1d} \\ m_{21} & m_{22} & \dots & m_{2d} \\ \vdots & \vdots & \ddots & \vdots \\ m_{n1} & m_{n2} & \dots & m_{nd} \end{pmatrix} \begin{pmatrix} x_1 \\ x_2 \\ \vdots \\ x_d \end{pmatrix} = \begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{pmatrix},$$

where

$$y_j = \sum_{i=1}^d m_{ji} x_i.$$

Note that if \mathbf{M} is not square, the dimensionality of \mathbf{y} differs from that of \mathbf{x} .

$$\mathbf{x}^t \mathbf{y} = \sum_{i=1}^d x_i y_i = \mathbf{y}^t \mathbf{x}.$$

Inner Product

$$\|\mathbf{x}\| = \sqrt{\mathbf{x}^t \mathbf{x}};$$

Norm of a
vector

$$\cos \theta = \frac{\mathbf{x}^t \mathbf{y}}{\|\mathbf{x}\| \|\mathbf{y}\|}$$

Cosine of the angle
between 2 vectors

$$\|\mathbf{x}^t \mathbf{y}\| \leq \|\mathbf{x}\| \|\mathbf{y}\|$$

Cauchy Schwartz
Inequality

Linear Independence of vectors / Basis vectors

We say a set of vectors $\{\mathbf{x}_1, \mathbf{x}_2, \dots, \mathbf{x}_n\}$ is *linearly independent* if no vector in the set can be written as a linear combination of any of the others. Informally, a set of d linearly independent vectors spans an d -dimensional vector space, i.e., any vector in that space can be written as a linear combination of such spanning vectors.

Outer Product of 2 vectors

The outer product (sometimes called *matrix product*) of two column vectors yields a matrix

$$\mathbf{M} = \mathbf{xy}^t = \begin{pmatrix} x_1 \\ x_2 \\ \vdots \\ x_d \end{pmatrix} (y_1 \ y_2 \ \dots \ y_n) = \begin{pmatrix} x_1y_1 & x_1y_2 & \dots & x_1y_n \\ x_2y_1 & x_2y_2 & \dots & x_2y_n \\ \vdots & \vdots & \ddots & \vdots \\ x_dy_1 & x_dy_2 & \dots & x_dy_n \end{pmatrix}.$$

Rank of a matrix

- Number of independent rows / columns of a matrix
- Outer product between 2 vectors gives a matrix of rank 1.

Gradient of a scalar function

Suppose $f(\mathbf{x})$ is a scalar function of d variables x_i which we represent as the vector \mathbf{x} . Then the derivative or gradient of f with respect to this parameter vector is computed component by component, i.e.,

$$\frac{\partial f(\mathbf{x})}{\partial \mathbf{x}} = \begin{pmatrix} \frac{\partial f(\mathbf{x})}{\partial x_1} \\ \frac{\partial f(\mathbf{x})}{\partial x_2} \\ \vdots \\ \frac{\partial f(\mathbf{x})}{\partial x_d} \end{pmatrix}.$$

Gradient of a vector-valued function (Jacobian)

$$\mathbf{J}(\mathbf{x}) = \frac{\partial \mathbf{f}(\mathbf{x})}{\partial \mathbf{x}} = \begin{pmatrix} \frac{\partial f_1(\mathbf{x})}{\partial x_1} & \cdots & \frac{\partial f_1(\mathbf{x})}{\partial x_d} \\ \vdots & \ddots & \vdots \\ \frac{\partial f_n(\mathbf{x})}{\partial x_1} & \cdots & \frac{\partial f_n(\mathbf{x})}{\partial x_d} \end{pmatrix}.$$

Matrix / Vector Derivatives

$$\frac{\partial}{\partial \mathbf{x}} [\mathbf{M}\mathbf{x}] = \mathbf{M}$$

**Gradient of
Vector function**

$$\frac{\partial}{\partial \mathbf{x}} [\mathbf{y}^t \mathbf{x}] = \frac{\partial}{\partial \mathbf{x}} [\mathbf{x}^t \mathbf{y}] = \mathbf{y}$$

**Gradient of scalar
function**

$$\frac{\partial}{\partial \mathbf{x}} [\mathbf{x}^t \mathbf{M} \mathbf{x}] = [\mathbf{M} + \mathbf{M}^t] \mathbf{x}.$$

Eigen values and vectors

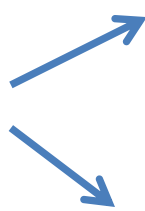
Given a $d \times d$ matrix M , a very important class of linear equations is of the form

$$M\mathbf{x} = \lambda\mathbf{x}, \quad (26)$$

which can be rewritten as

$$(M - \lambda I)\mathbf{x} = 0, \quad (27)$$

Diagonalization property

$$M = VDV^{-1}$$


Matrix, whose columns correspond to Eigenvectors of M

Diagonal matrix, corresponding to eigenvalues of M

Rank deficient square matrices

- A matrix **A** of rank r will have r independent rows / columns, and is said to be 'rank-deficient'.
- There will be r independent eigen vectors.....corresponding to r non- zero eigen values
- ' r ' eigen values are zero.

Matrix Inverse

The inverse of a $n \times d$ matrix \mathbf{M} , denoted \mathbf{M}^{-1} , is the $d \times n$ matrix such that

$$\mathbf{M}\mathbf{M}^{-1} = \mathbf{I}.$$

$$\mathbf{M}^{-1} = \frac{\text{Adj}[\mathbf{M}]}{|\mathbf{M}|}.$$

For square matrix **M**

If \mathbf{M}^{-1} does not exist — because the columns of \mathbf{M} are not linearly independent or \mathbf{M} is not square — one typically uses instead the *pseudoinverse* \mathbf{M}^\dagger , defined as

$$\mathbf{M}^\dagger = [\mathbf{M}^t \mathbf{M}]^{-1} \mathbf{M}^t,$$

Orthogonal Matrix

$\{q_1, q_2, \dots, q_M\}$ M vectors

$$q_i^T q_j = \begin{cases} 1 & i = j \\ 0 & i \neq j \end{cases}$$

 $Q^T Q = I$

columns of Q denote
 $\begin{bmatrix} q_1 & q_2 & \dots & q_M \end{bmatrix}$

Q is of size $M \times M$ and is said to be an 'orthogonal' matrix

Positive Definite Matrix

Positive definite square matrix **A**  $\mathbf{x}^T \mathbf{A} \mathbf{x} \geq 0$

\mathbf{x} denotes a vector of size $d \times 1$.

A denotes a vector of size $d \times d$.

Eigen values of positive definite matrix **A** are non-negative.

Real symmetric matrix

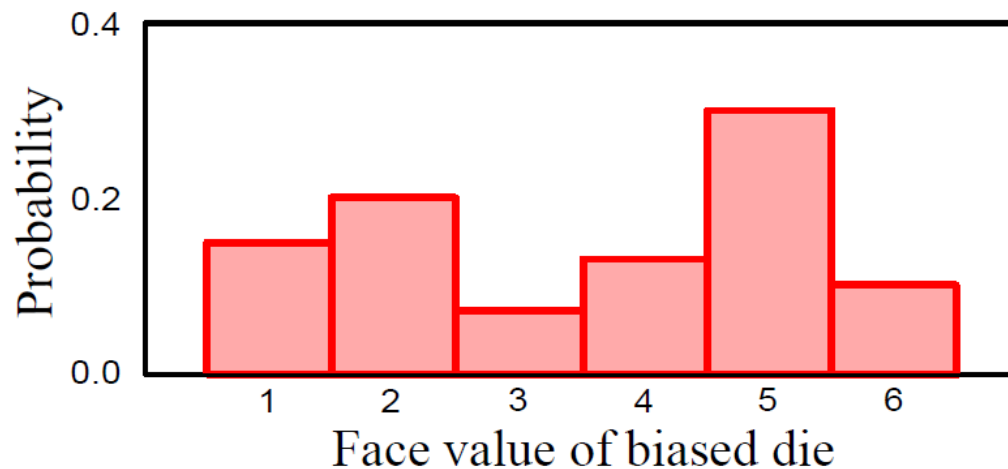
- Elements of a matrix are real numbers
- Matrix is symmetric
- Eigen values of a real symmetric matrix are real.

Probability Theory

Random variables

- A random variable x denotes a quantity that is uncertain
- May be result of experiment (flipping a coin) or a real world measurements (measuring temperature)
- If observe several instances of x we get different values
- Some values occur more than others and this information is captured by a probability distribution

Discrete Random Variables



$$P(x) \geq 0 \quad \text{and}$$
$$\sum_{x \in \mathcal{X}} P(x) = 1.$$

The *mean* or *expected value* or *average* of x is defined by

$$\mathcal{E}[x] = \mu = \sum_{x \in \mathcal{X}} xP(x) = \sum_{i=1}^m v_i p_i.$$

More generally, if $f(x)$ is any function of x , the expected value of f is defined by

$$\mathcal{E}[f(x)] = \sum_{x \in \mathcal{X}} f(x)P(x).$$

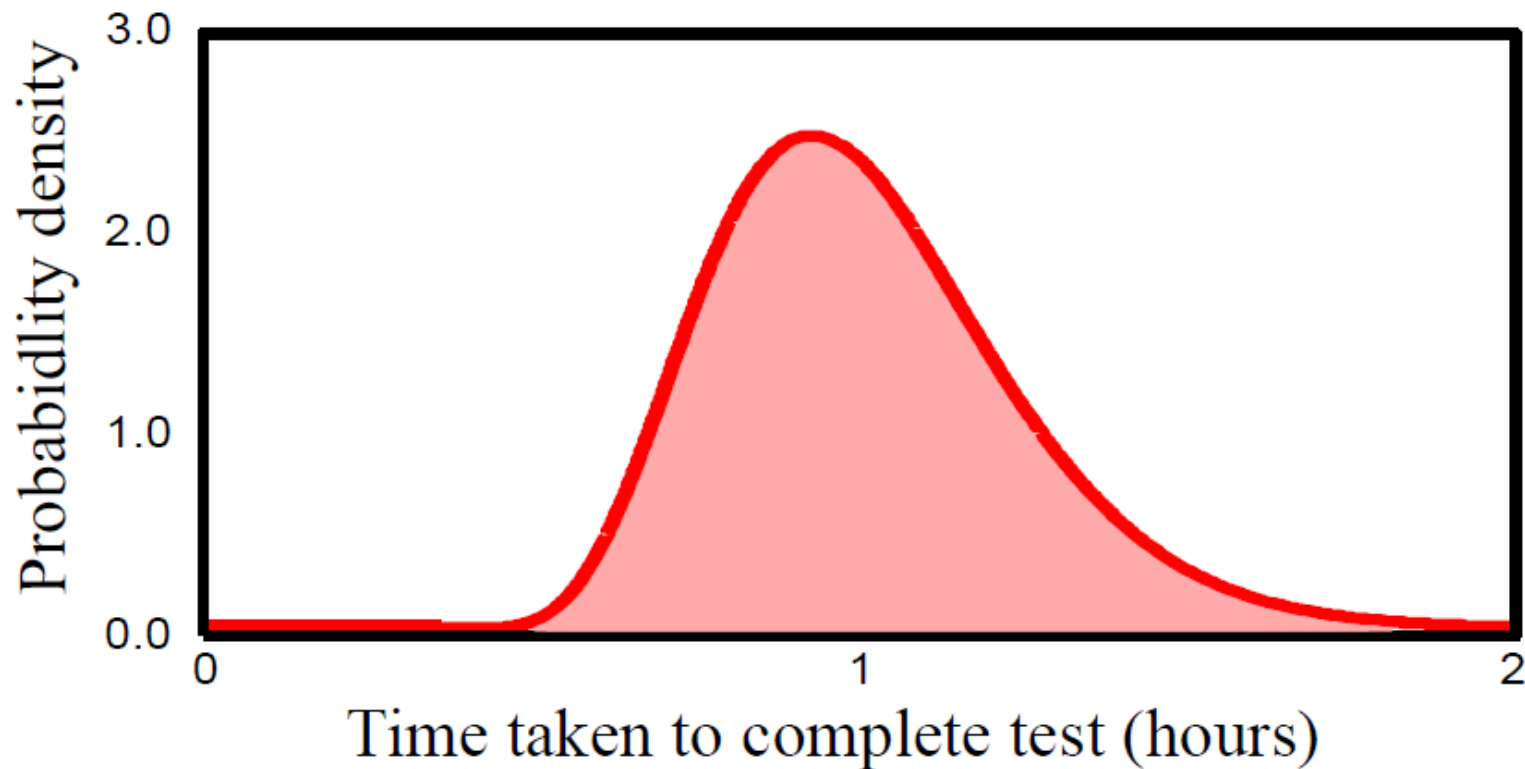
SECOND
MOMENT

$$\mathcal{E}[x^2] = \sum_{x \in \mathcal{X}} x^2 P(x)$$

VARIANCE

$$\text{Var}[x] = \mathcal{E}[(x - \mu)^2] = \sigma^2 = \sum_{x \in \mathcal{X}} (x - \mu)^2 P(x),$$

Continuous Random Variable



$$p(x) \geq 0 \quad \text{and} \\ \int_{-\infty}^{\infty} p(x) dx = 1.$$

Expectation/ Variance of continuous random variable

$$\mathcal{E}[f(x)] = \int_{-\infty}^{\infty} f(x)p(x) dx$$

$$\mu = \mathcal{E}[x] = \int_{-\infty}^{\infty} xp(x) dx$$

$$\text{Var}[x] = \sigma^2 = \mathcal{E}[(x - \mu)^2] = \int_{-\infty}^{\infty} (x - \mu)^2 p(x) dx,$$

Joint Probability

- Consider two random variables x and y
- If we observe multiple paired instances, then some combinations of outcomes are more likely than others
- This is captured in the joint probability distribution
- Written as $P(x,y)$
- Can read $P(x,y)$ as “probability of x and y ”

For 2 random variables x and y ,

$$P(x, y) \geq 0 \quad \text{and} \\ \sum_{x \in \mathcal{X}} \sum_{y \in \mathcal{Y}} P(x, y) = 1.$$

Marginalization property

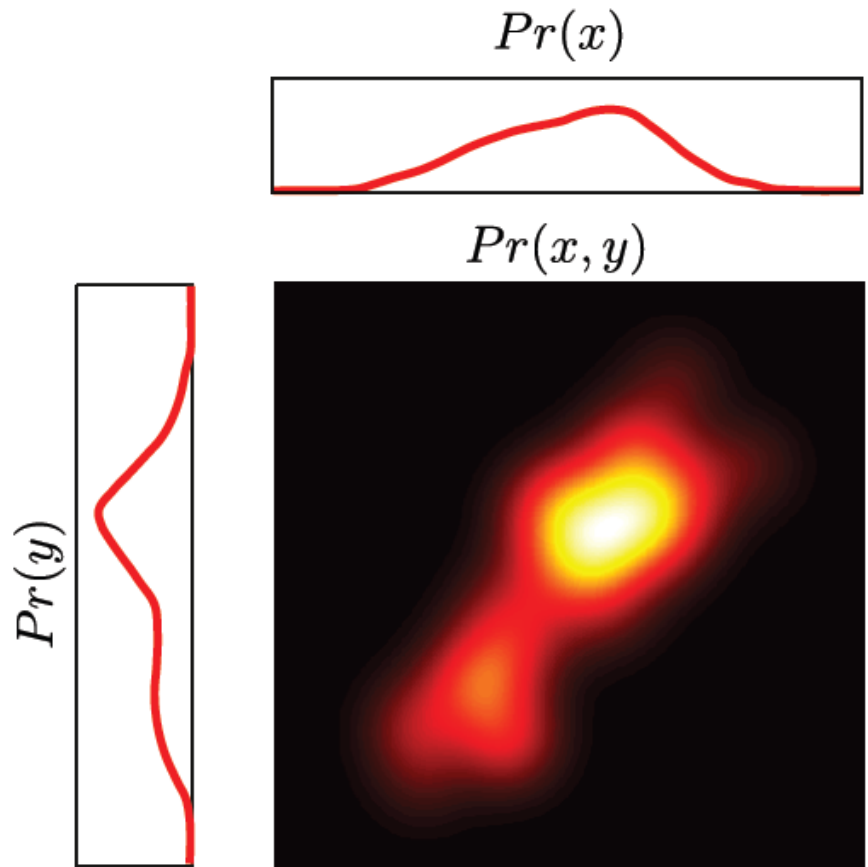
$$P_x(x) = \sum_{y \in \mathcal{Y}} P(x, y) \\ P_y(y) = \sum_{x \in \mathcal{X}} P(x, y)$$

Marginalization

We can recover probability distribution of any variable in a joint distribution by integrating (or summing) over the other variables

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

$$Pr(y) = \int Pr(x, y) dx$$

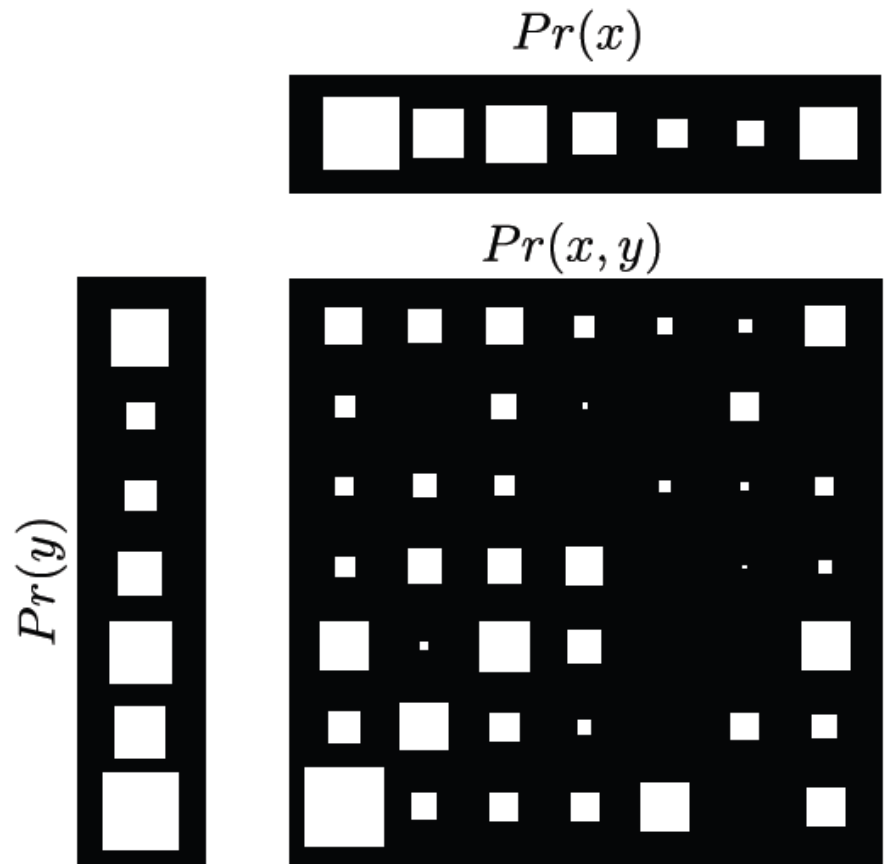


Marginalization

We can recover probability distribution of any variable in a joint distribution by integrating (or summing) over the other variables

$$Pr(x) = \sum_y Pr(x, y)$$

$$Pr(y) = \sum_x Pr(x, y)$$



Marginalization

We can recover probability distribution of any variable in a joint distribution by integrating (or summing) over the other variables

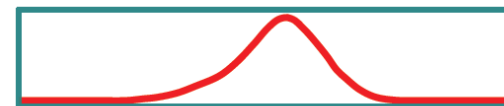
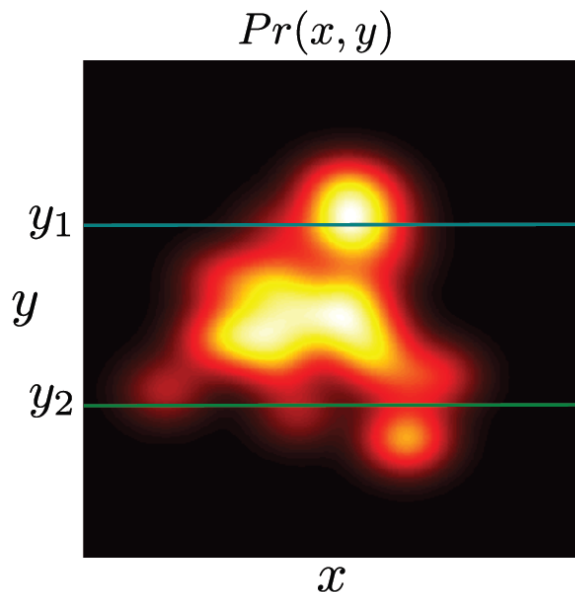
$$Pr(x) = \int Pr(x, y) dy$$

$$Pr(y) = \int Pr(x, y) dx$$

Works in higher dimensions as well – leaves joint distribution between whatever variables are left

Conditional Probability

- Conditional probability of x given that $y=y_1$ is relative propensity of variable x to take different outcomes given that y is fixed to be equal to y_1 .
- Written as $\Pr(x | y=y_1)$



$$\Pr(x | y = y_1)$$

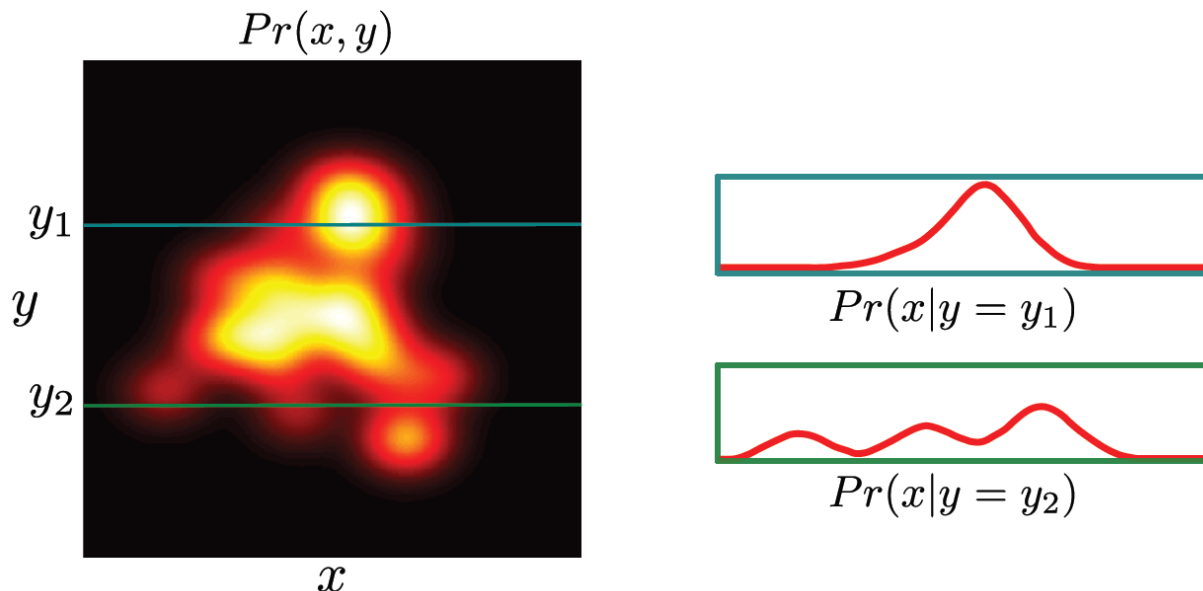


$$\Pr(x | y = y_2)$$

Conditional Probability

- Conditional probability can be extracted from joint probability
- Extract appropriate slice and normalize

$$Pr(x|y = y^*) = \frac{Pr(x, y = y^*)}{\int Pr(x, y = y^*)dx} = \frac{Pr(x, y = y^*)}{Pr(y = y^*)}$$



Conditional Probability

$$Pr(x|y = y^*) = \frac{Pr(x, y = y^*)}{\int Pr(x, y = y^*)dx} = \frac{Pr(x, y = y^*)}{Pr(y = y^*)}$$

- More usually written in compact form

$$Pr(x|y) = \frac{Pr(x, y)}{Pr(y)}$$

- Can be re-arranged to give

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

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

Conditional Probability

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

- This idea can be extended to more than two variables

$$\begin{aligned} Pr(w, x, y, z) &= Pr(w, x, y|z)Pr(z) \\ &= Pr(w, x|y, z)Pr(y|z)Pr(z) \\ &= Pr(w|x, y, z)Pr(x|y, z)Pr(y|z)Pr(z) \end{aligned}$$

Bayes' Rule

From before:

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

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

Combining:

$$Pr(y|x)Pr(x) = Pr(x|y)Pr(y)$$

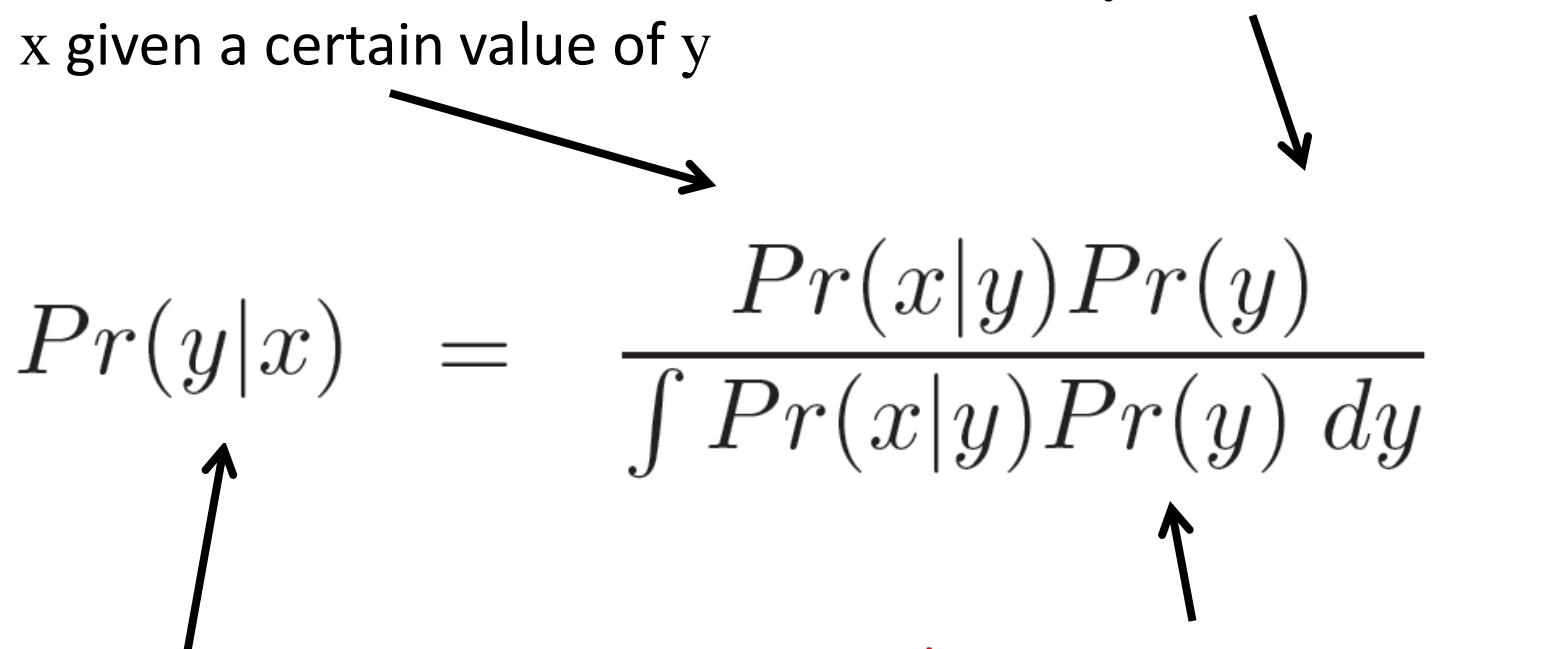
Re-arranging:

$$\begin{aligned} Pr(y|x) &= \frac{Pr(x|y)Pr(y)}{Pr(x)} \\ &= \frac{Pr(x|y)Pr(y)}{\int Pr(x, y) dy} \\ &= \frac{Pr(x|y)Pr(y)}{\int Pr(x|y)Pr(y) dy} \end{aligned}$$

Bayes' Rule Terminology

Likelihood – propensity for observing a certain value of x given a certain value of y

Prior – what we know about y before seeing x


$$Pr(y|x) = \frac{Pr(x|y)Pr(y)}{\int Pr(x|y)Pr(y) dy}$$

Posterior – what we know about y after seeing x

Evidence – a constant to ensure that the left hand side is a valid distribution

Independence

- If two variables x and y are independent then variable x tells us nothing about variable y (and vice-versa)

$$Pr(x|y) = Pr(x)$$

$$Pr(y|x) = Pr(y)$$

Independence

- When variables are independent, the joint factorizes into a product of the marginals:

$$\begin{aligned}Pr(x, y) &= Pr(x|y)Pr(y) \\ &= Pr(x)Pr(y)\end{aligned}$$

Discrete Random vectors

To extend these results from two variables x and y to d variables x_1, x_2, \dots, x_d , it is convenient to employ vector notation. The joint probability mass function $P(\mathbf{x})$ satisfies $P(\mathbf{x}) \geq 0$ and $\sum P(\mathbf{x}) = 1$ (Eq. 46), where the sum extends over all possible values for the vector \mathbf{x} . Note that $P(\mathbf{x})$ is a function of d variables, x_1, x_2, \dots, x_d ,

Continuous Random vectors

The probability density function $p(\mathbf{x})$ must satisfy

$$p(\mathbf{x}) \geq 0 \quad \text{and} \quad \int_{-\infty}^{\infty} p(\mathbf{x}) d\mathbf{x} = 1, \quad (78)$$

where the integral is understood to be a d -fold, multiple integral, and where $d\mathbf{x}$ is the element of d -dimensional volume $d\mathbf{x} = dx_1 dx_2 \cdots dx_d$. The corresponding moments for a general n -dimensional vector-valued function are

$$\mathcal{E}[\mathbf{f}(\mathbf{x})] = \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} \cdots \int_{-\infty}^{\infty} \mathbf{f}(\mathbf{x}) p(\mathbf{x}) dx_1 dx_2 \cdots dx_d = \int_{-\infty}^{\infty} \mathbf{f}(\mathbf{x}) p(\mathbf{x}) d\mathbf{x} \quad (79)$$

Continuous random vectors

$$\begin{aligned}\mu = \mathcal{E}[\mathbf{x}] &= \int_{-\infty}^{\infty} \mathbf{x} p(\mathbf{x}) d\mathbf{x} \\ \Sigma = \mathcal{E}[(\mathbf{x} - \mu)(\mathbf{x} - \mu)^t] &= \int_{-\infty}^{\infty} (\mathbf{x} - \mu)(\mathbf{x} - \mu)^t p(\mathbf{x}) d\mathbf{x}.\end{aligned}$$

Discrete random vectors

MEAN VECTOR

In particular, the d -dimensional *mean vector* μ is defined by

$$\mu = \mathcal{E}[\mathbf{x}] = \begin{bmatrix} \mathcal{E}[x_1] \\ \mathcal{E}[x_2] \\ \vdots \\ \mathcal{E}[x_d] \end{bmatrix} = \begin{bmatrix} \mu_1 \\ \mu_2 \\ \vdots \\ \mu_d \end{bmatrix} = \sum_{\mathbf{x}} \mathbf{x} P(\mathbf{x}).$$

COVARIANCE MATRIX

Similarly, the *covariance matrix* Σ is defined as the (square) matrix whose ij^{th} element σ_{ij} is the covariance of x_i and x_j :

$$\sigma_{ij} = \mathcal{E}[(x_i - \mu_i)(x_j - \mu_j)] \quad i, j = 1 \dots d, \quad (72)$$

$$\begin{aligned}
\Sigma &= \begin{bmatrix} \mathcal{E}[(x_1 - \mu_1)(x_1 - \mu_1)] & \mathcal{E}[(x_1 - \mu_1)(x_2 - \mu_2)] & \dots & \mathcal{E}[(x_1 - \mu_1)(x_d - \mu_d)] \\ \mathcal{E}[(x_2 - \mu_2)(x_1 - \mu_1)] & \mathcal{E}[(x_2 - \mu_2)(x_2 - \mu_2)] & \dots & \mathcal{E}[(x_2 - \mu_2)(x_d - \mu_d)] \\ \vdots & \vdots & \ddots & \vdots \\ \mathcal{E}[(x_d - \mu_d)(x_1 - \mu_1)] & \mathcal{E}[(x_d - \mu_d)(x_2 - \mu_2)] & \dots & \mathcal{E}[(x_d - \mu_d)(x_d - \mu_d)] \end{bmatrix} \\
&= \begin{bmatrix} \sigma_{11} & \sigma_{12} & \dots & \sigma_{1d} \\ \sigma_{21} & \sigma_{22} & \dots & \sigma_{2d} \\ \vdots & \vdots & \ddots & \vdots \\ \sigma_{d1} & \sigma_{d2} & \dots & \sigma_{dd} \end{bmatrix} = \begin{bmatrix} \sigma_1^2 & \sigma_{12} & \dots & \sigma_{1d} \\ \sigma_{21} & \sigma_2^2 & \dots & \sigma_{2d} \\ \vdots & \vdots & \ddots & \vdots \\ \sigma_{d1} & \sigma_{d2} & \dots & \sigma_d^2 \end{bmatrix}. \tag{73}
\end{aligned}$$

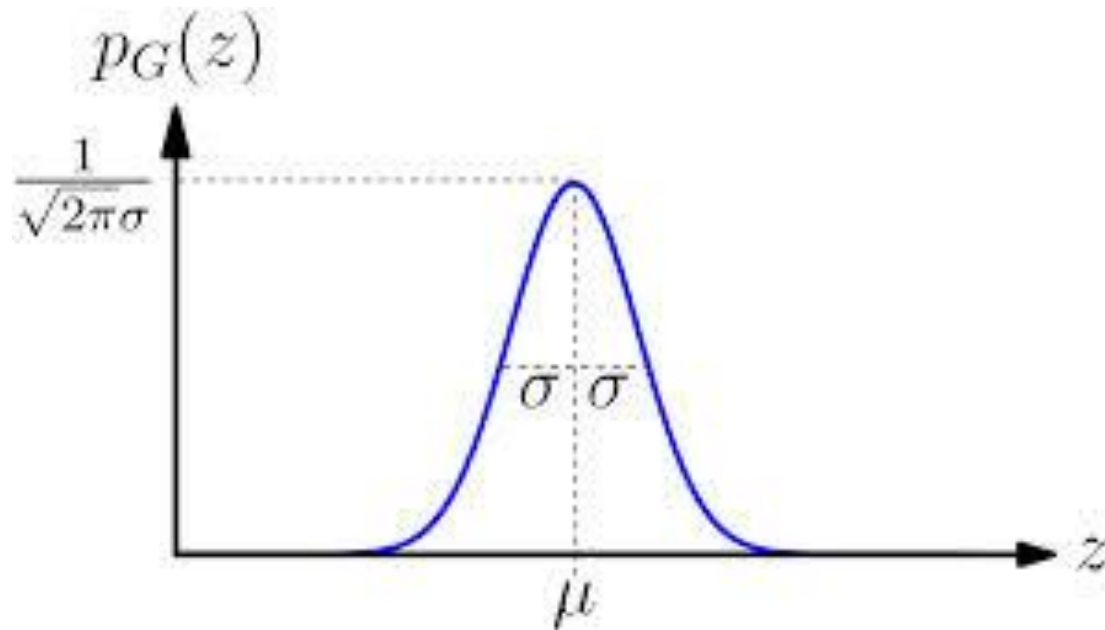
We can use the vector product $(\mathbf{x} - \boldsymbol{\mu})(\mathbf{x} - \boldsymbol{\mu})^t$, to write the covariance matrix as

$$\Sigma = \mathcal{E}[(\mathbf{x} - \boldsymbol{\mu})(\mathbf{x} - \boldsymbol{\mu})^t]. \tag{74}$$

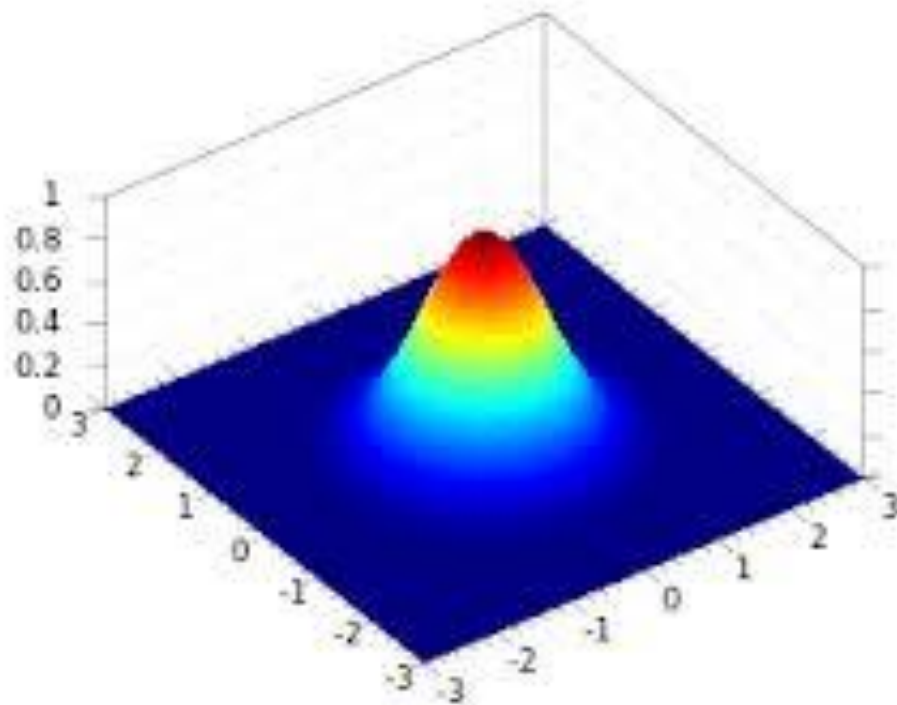
Properties of covariance matrix

Thus, the diagonal elements of Σ are just the variances of the individual elements of \mathbf{x} , which can never be negative; the off-diagonal elements are the covariances, which can be positive or negative. If the variables are statistically independent, the covariances are zero, and the covariance matrix is diagonal. The analog to the Cauchy-Schwarz inequality comes from recognizing that if \mathbf{w} is any d -dimensional vector, then the variance of $\mathbf{w}^t \mathbf{x}$ can never be negative. This leads to the requirement that the quadratic form $\mathbf{w}^t \Sigma \mathbf{w}$ never be negative. Matrices for which this is true are said to be *positive semi-definite*; thus, the covariance matrix Σ must be positive semi-definite. It can be shown that this is equivalent to the requirement that none of the eigenvalues of Σ can ever be negative.

1 dimensional Gaussian



2 dimensional Gaussian



Multi dimensional Gaussian

$$p(\mathbf{x}) = \frac{1}{(2\pi)^{d/2} |\Sigma|^{1/2}} e^{-\frac{1}{2}(\mathbf{x} - \mu)^t \Sigma^{-1} (\mathbf{x} - \mu)}.$$

$$\mu = \mathcal{E}[\mathbf{x}] = \int_{-\infty}^{\infty} \mathbf{x} p(\mathbf{x}) d\mathbf{x}$$

$$\Sigma = \mathcal{E}[(\mathbf{x} - \mu)(\mathbf{x} - \mu)^t] = \int_{-\infty}^{\infty} (\mathbf{x} - \mu)(\mathbf{x} - \mu)^t p(\mathbf{x}) d\mathbf{x},$$