

Machine Learning

CMPT 726

Mo Chen

SFU School of Computing Science

2022-10-04

Probability Review

Terminology

- **Sample space Ω :** Set of *all* possible outcomes of a random phenomenon
E.g.: Toss two coins - sample space is {HH, HT, TH, TT}
- **Event E :** A subset of the possible outcomes
E.g.: The event that the second coin turns out to be heads, i.e.: {HH, TH}
- **Probability (formally a “probability measure”) $\Pr(\cdot)$:** A function that assigns every possible event a number, representing the chance that the event happens
E.g.: If both coins are fair, $\Pr(\text{second coin turns out to be heads}) = \frac{1}{2}$
- **Random variables (RVs):** Variables whose values depend on the outcome of a random phenomenon
E.g.: $X_i = \begin{cases} 1 & \text{ith coin turns out to be heads} \\ 0 & \text{otherwise} \end{cases}$, or $Y = \sum_i X_i$ (the number of heads)

Terminology

- **Discrete random variables:** RVs that take on values from a discrete set
- **Continuous random variables:** RVs that take on values from a continuous range
- **Probability distribution:** a function that characterizes the probability of different realizations of RVs
 - Can be represented as cumulative distribution functions (cdfs), probability mass functions (pmfs) in the case of discrete RVs, or probability density functions (pdfs)
- **Support of distribution $\text{supp}(X)$:** the set of realizations of RVs where the pmf (in the case of discrete RVs) or pdf (in the case of continuous RVs) is non-zero

Discrete vs. Continuous RVs

Discrete random variables

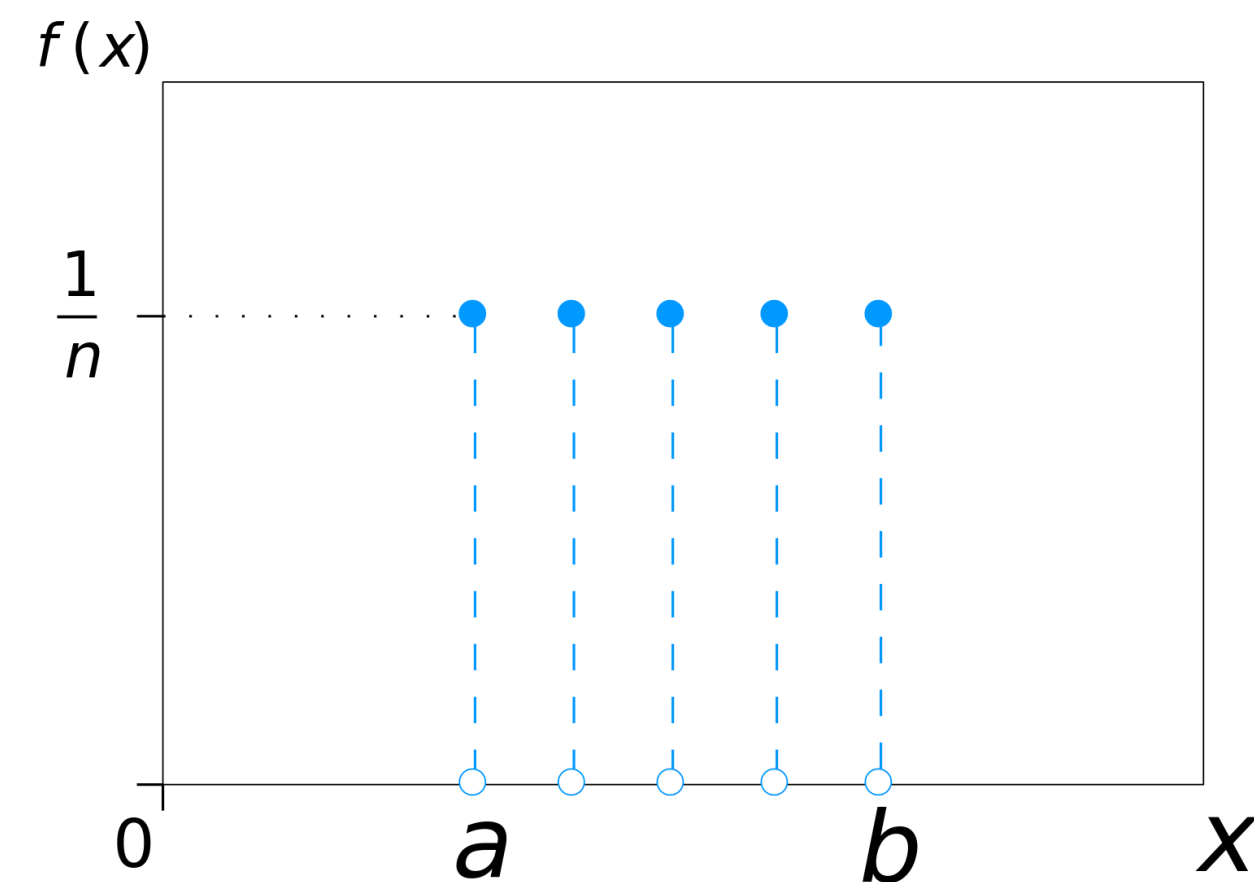
- Cumulative distribution functions (cdf): $F_X(x) = \Pr(X \leq x)$
- Probability mass functions (pmf): $p_X(x) = \Pr(X = x)$
- Examples: Bernoulli RVs, Categorical RVs

Continuous random variables

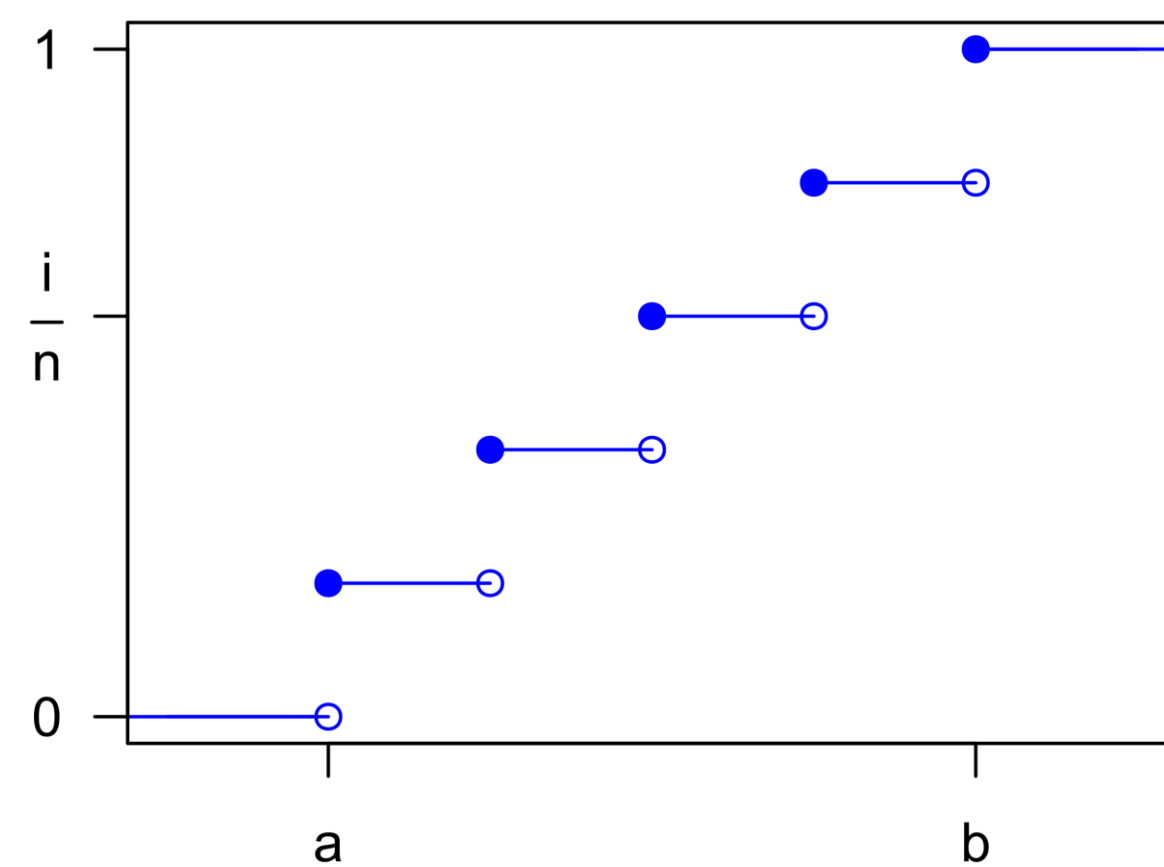
- Cumulative distribution functions (cdf): $F_X(x) = \Pr(X \leq x)$
- Probability density functions (pdf): $f_X(x) = \frac{d}{dx} F_X(x)$
- Examples: Uniform RVs, Normal RVs

Discrete vs. Continuous RVs

Discrete random variables



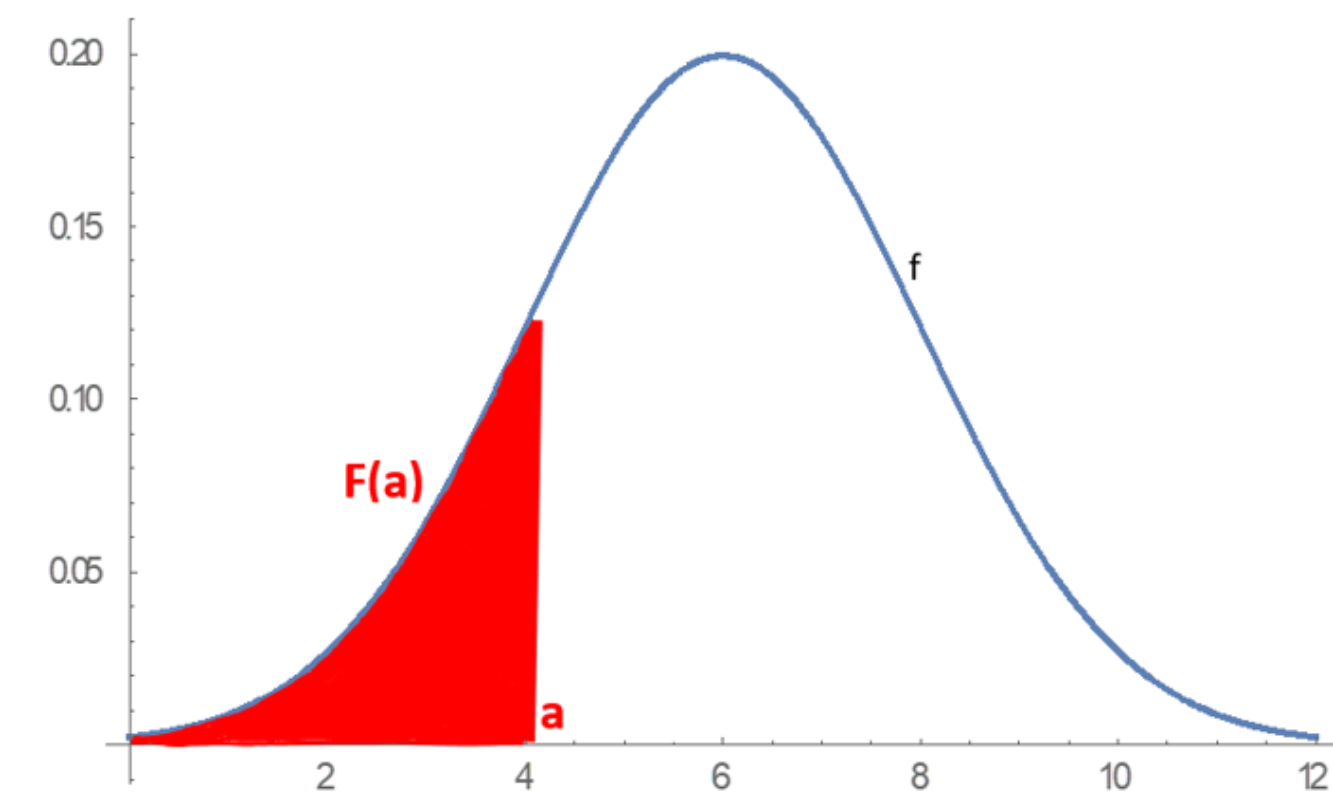
pmf



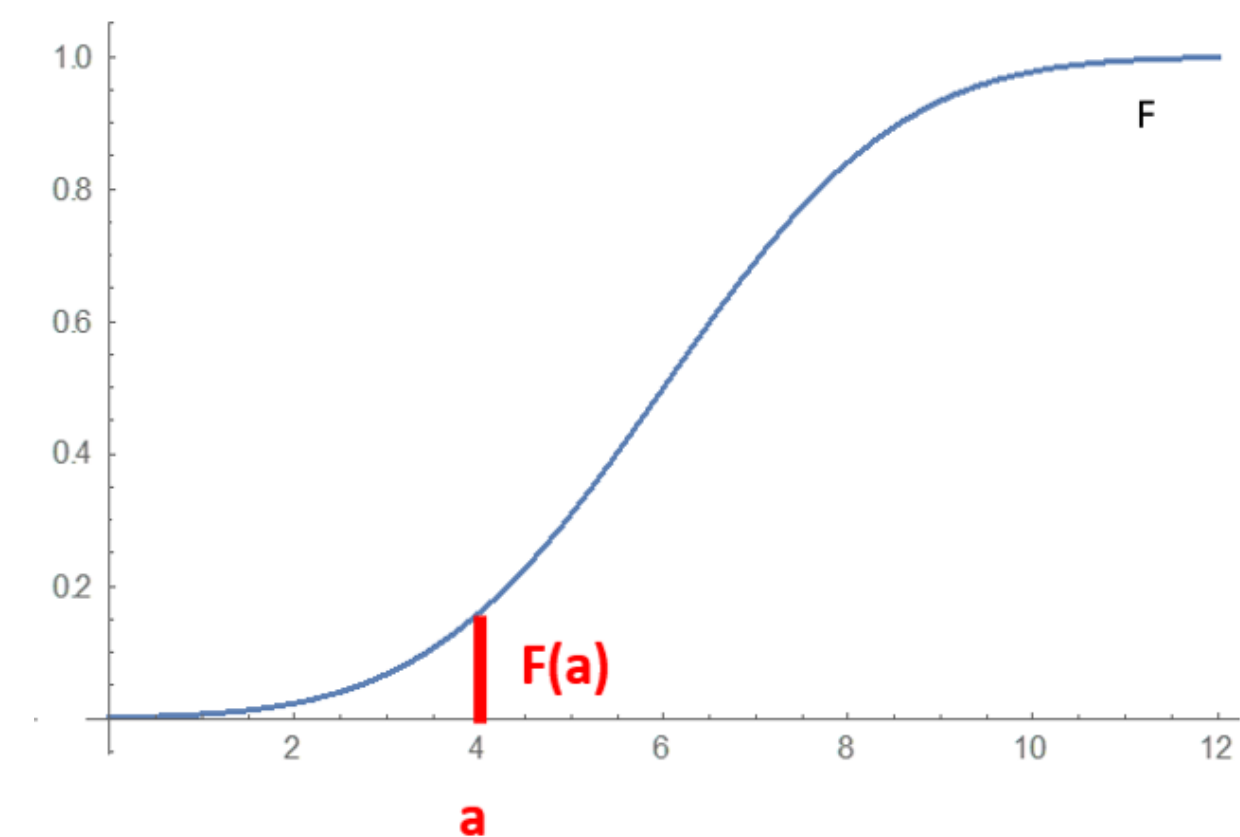
cdf

Credit: Wikipedia

Continuous random variables



pdf



cdf

Credit: Wikipedia

Discrete vs. Continuous RVs

Discrete random variables

$$p_X(x) = \Pr(X = x) \geq 0 \quad \forall x$$

$$p_X(x) \leq 1 \quad \forall x$$

$$\sum_{x \in \Omega} p_X(x) = 1$$

$$F_X(x) = \Pr(X \leq x) = \sum_{\tilde{x} \in \Omega: \tilde{x} \leq x} p_X(\tilde{x})$$

Continuous random variables

$$f_X(x) = \frac{d}{dx} F_X(x) \geq 0 \quad \forall x \quad \text{cdf is non-decreasing}$$

$$\Pr(X = x) = 0 \quad \forall x \implies f_X(x) \neq \Pr(X = x)$$

$f_X(x)$ may be larger than 1 (could be arbitrarily large)

$$\int_{-\infty}^{\infty} f_X(x) dx = 1$$

cdf could have arbitrarily high slope

$$F_X(x) = \Pr(X \leq x) = \int_{-\infty}^x f_X(s) ds$$

Common Discrete Distributions

Bernoulli distribution: $X \sim \text{Bernoulli}(p)$

$$p_X(x) = \Pr(X = x) = \begin{cases} p & x = 1 \\ 1 - p & x = 0 \end{cases}$$

More mathematically convenient form:

$$p_X(x) = \Pr(X = x) = p^x (1 - p)^{1-x}$$

Common Discrete Distributions

Categorical distribution:

$$p_X(x) = \Pr(X = x) = \begin{cases} p_1 & x = 1 \\ p_2 & x = 2 \\ \vdots & \vdots \\ p_k & x = k \end{cases}$$

More mathematically convenient form:

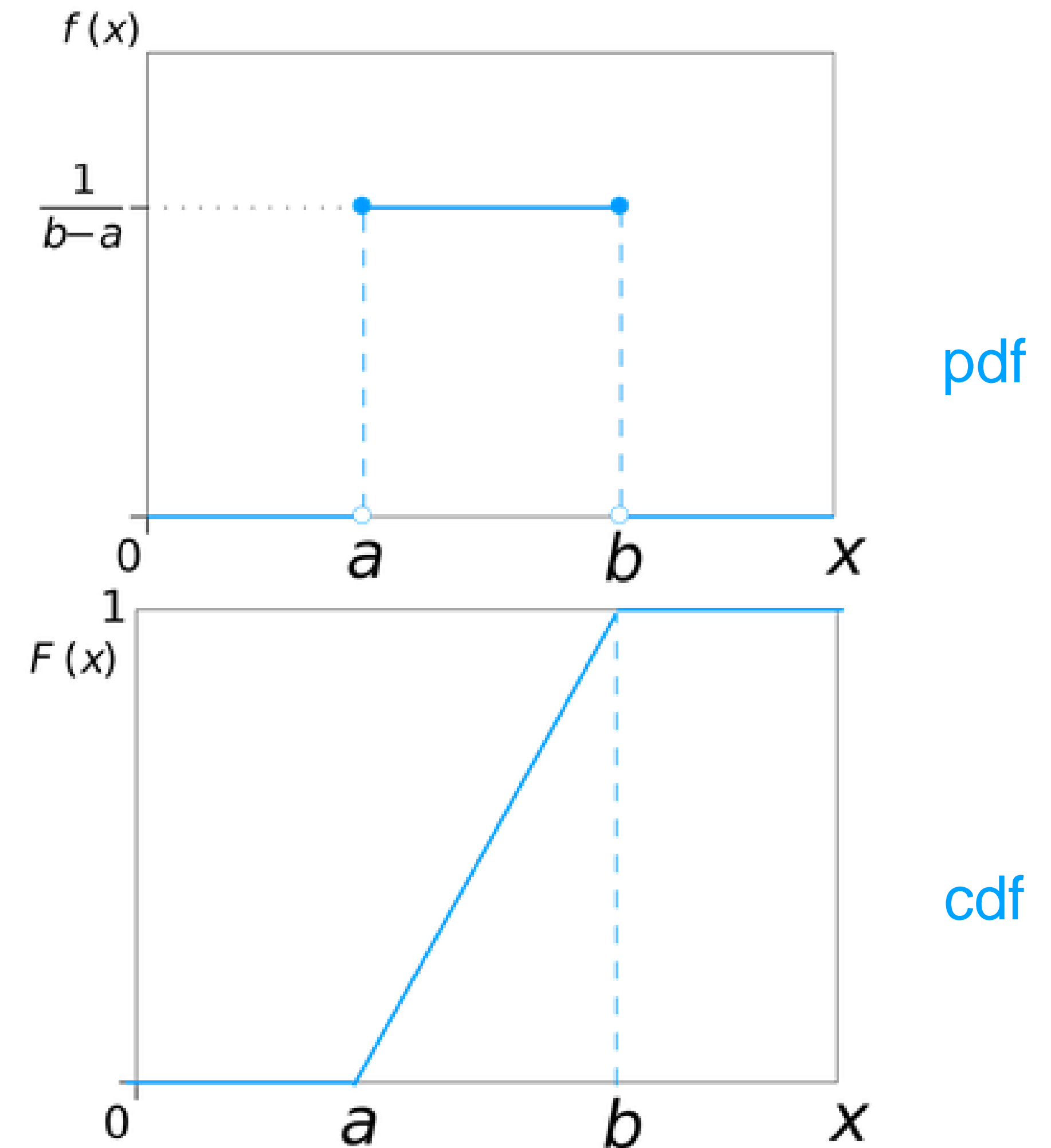
$$p_X(x) = \Pr(X = x) = \prod_{i=1}^k p_i^{[x=i]}, \text{ where } [x = i] = \begin{cases} 1 & x = i \\ 0 & x \neq i \end{cases}$$

Common Continuous Distributions

Uniform distribution: $X \sim \text{Uniform}(a, b)$

$$f_X(x) = \begin{cases} \frac{1}{b-a} & x \in [a, b] \\ 0 & x \notin [a, b] \end{cases}$$

$$\text{supp}(X) = [a, b]$$



Common Continuous Distributions

Normal distribution: $X \sim \mathcal{N}(\mu, \sigma^2)$

(In ML, more commonly referred to as the Gaussian distribution)

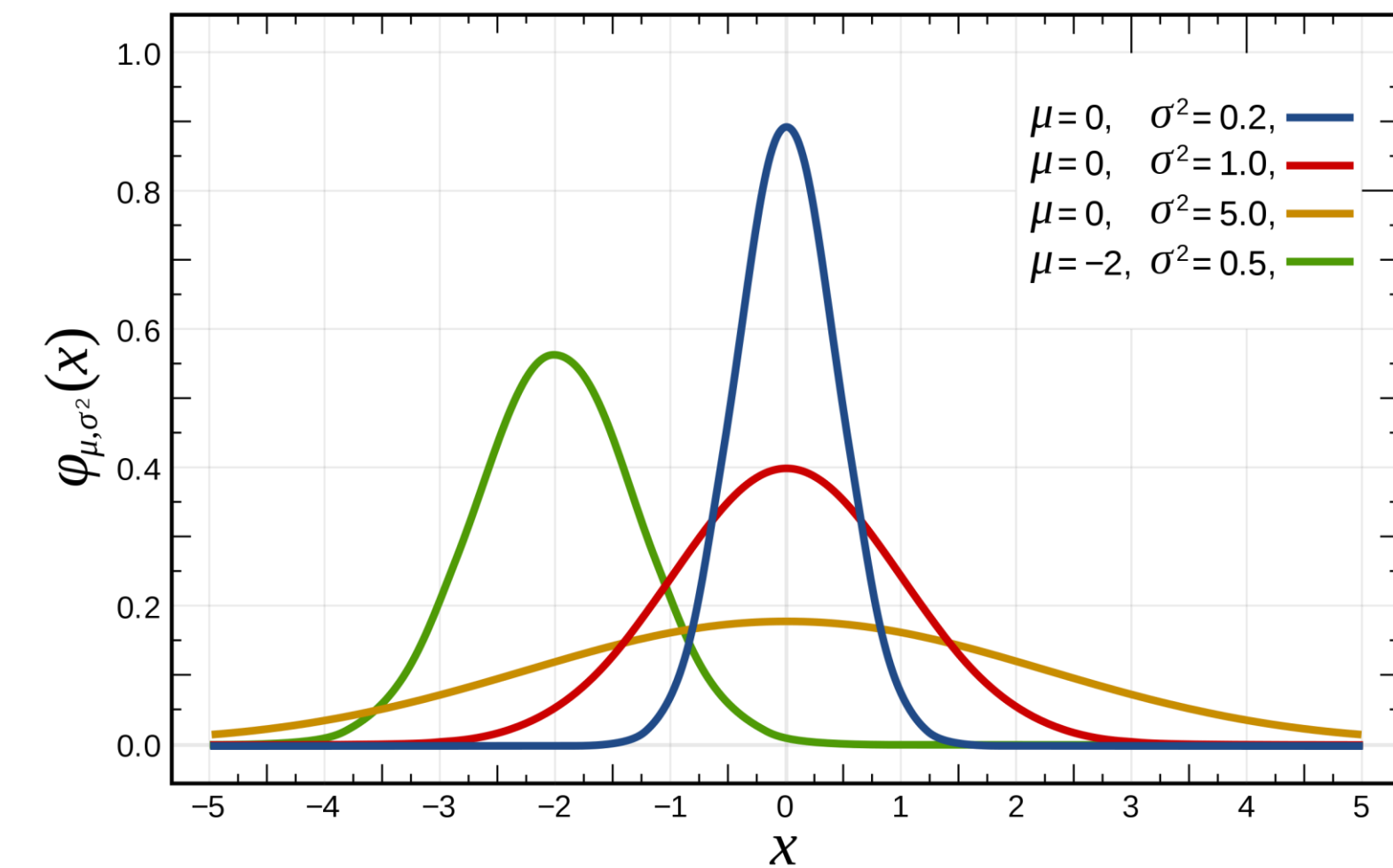
$$f_X(x) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$$

$$\text{supp}(X) = \mathbb{R}$$

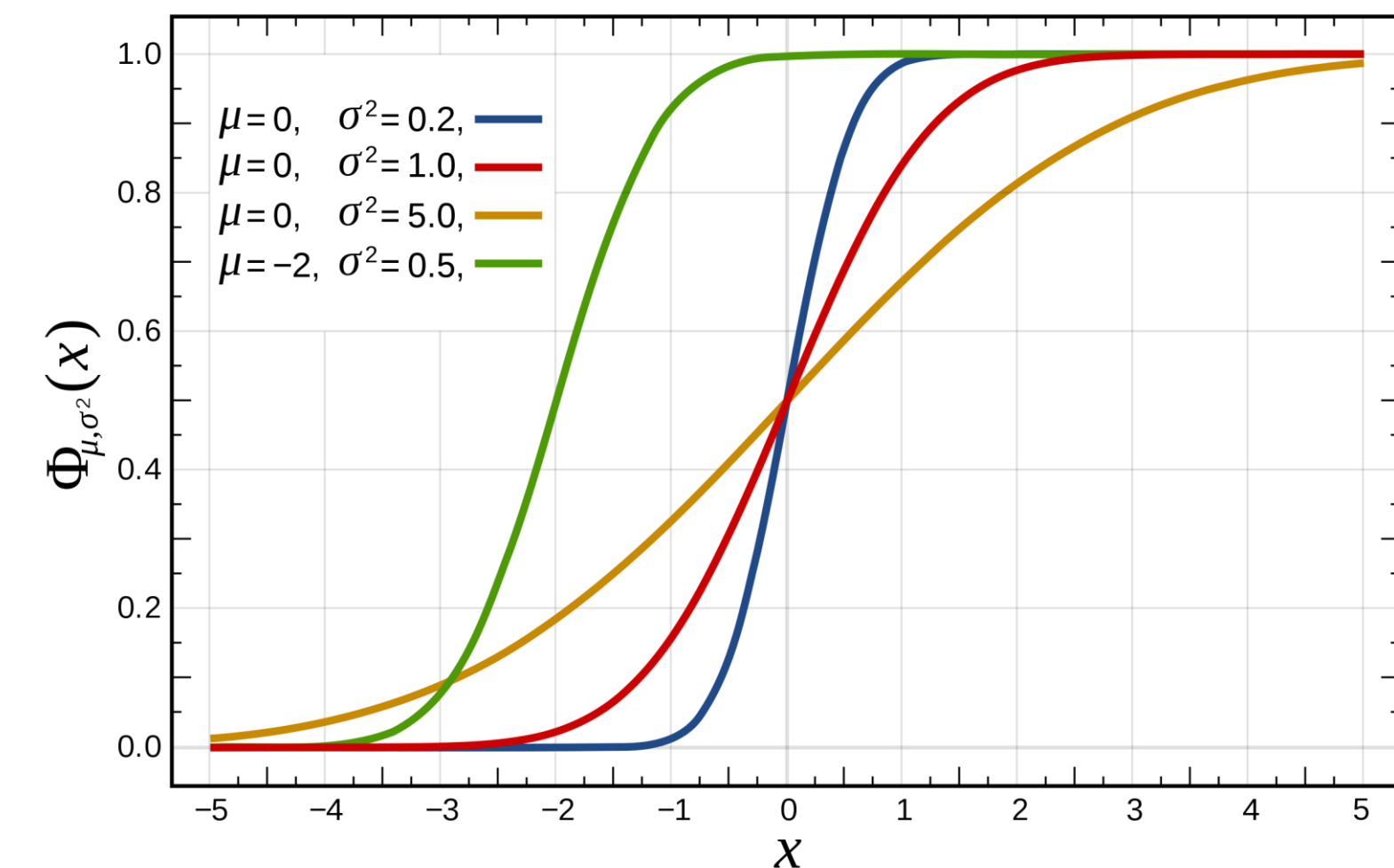
Standard normal distribution: $Z \sim \mathcal{N}(0,1)$

$$Z + \mu \sim \mathcal{N}(\mu, 1) \text{ and } \sigma Z \sim \mathcal{N}(0, \sigma^2) \\ \Rightarrow \mu + \sigma Z \sim \mathcal{N}(\mu, \sigma^2)$$

$$\text{Hence, } \frac{X-\mu}{\sigma} \sim \mathcal{N}(0,1)$$



pdf



cdf

Credit: Wikipedia

Multiple Random Variables

- What if we have multiple random variables, which may depend on one another? How do we represent the dependence between them?
 - E.g.: Tomorrow's temperature and snowfall
- Going forward, will use slightly different notation:
 - Will use capital letters, e.g.: X , to denote RVs and corresponding lowercase letters, e.g.: x , to denote a realized value of the RVs. Can therefore drop the subscripts in $p_X(x)$ and $F_X(x)$.
 - Will overload the notation $p(x)$ to mean the pmf $p_X(x)$ if X is discrete and the pdf $f_X(x)$ if X is continuous.
 - So, the cdf of X will be denoted as $F(x)$ and the pdf/pmf of X will be denoted as $p(x)$

Joint Probability Distributions

Two random variables:

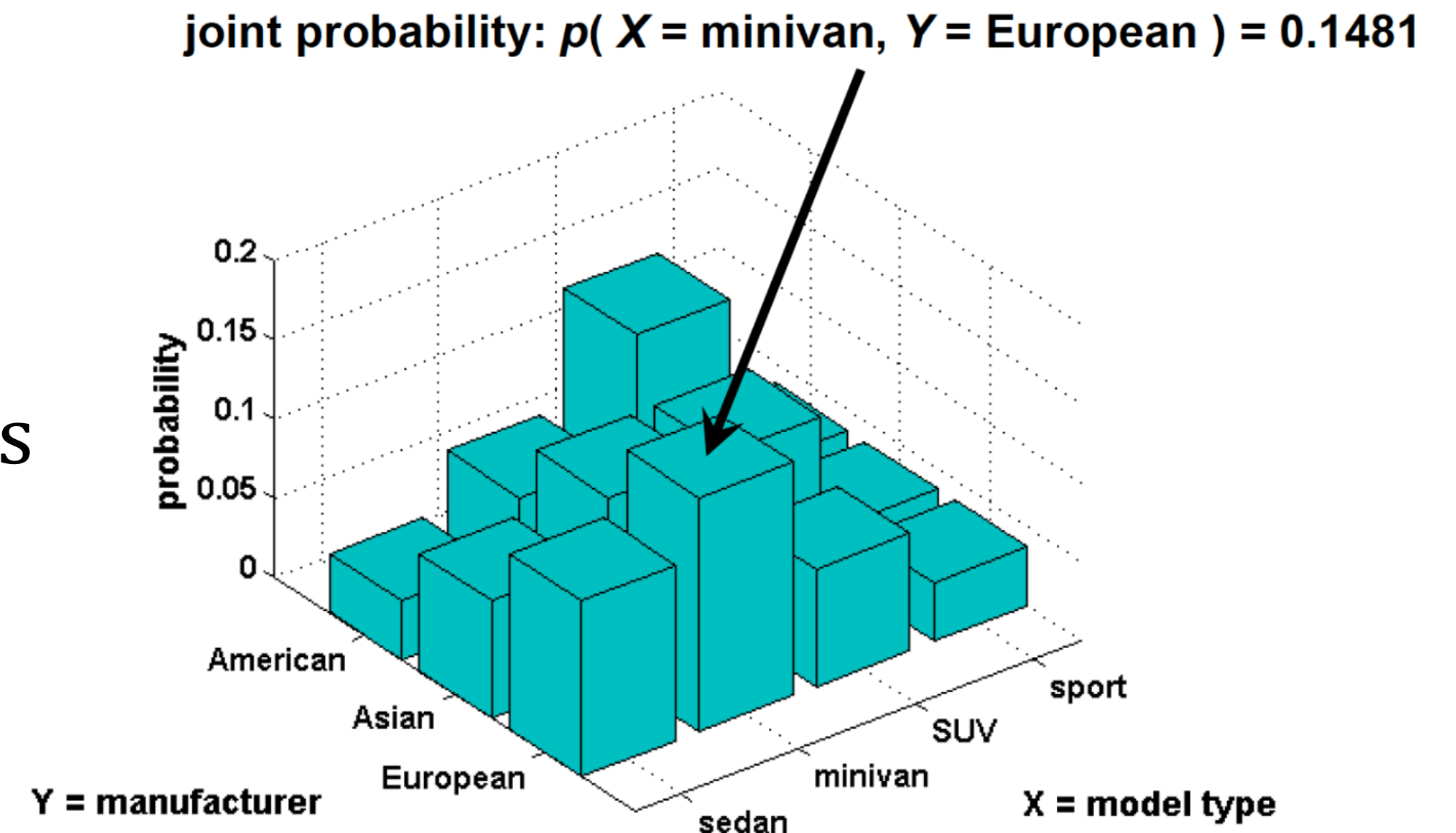
$$F(x, y) = \Pr(X \leq x \text{ and } Y \leq y)$$

$$p(x, y) = \begin{cases} \Pr(X = x \text{ and } Y = y) & X, Y \text{ are discrete} \\ \frac{\partial^2}{\partial x \partial y} F(x, y) & X, Y \text{ are continuous} \end{cases}$$

In general:

$$F(x_1, \dots, x_n) = \Pr(X_1 \leq x_1 \text{ and } \dots \text{ and } X_n \leq x_n)$$

$$p(x_1, \dots, x_n) = \begin{cases} \Pr(X_1 = x_1 \text{ and } \dots \text{ and } X_n = x_n) & X_1, \dots, X_n \text{ are discrete} \\ \frac{\partial^n}{\partial x_1 \dots \partial x_n} F(x_1, \dots, x_n) & X_1, \dots, X_n \text{ are continuous} \end{cases}$$



Credit: Jeff Howbert

Marginal Probability Distributions

Two random variables:

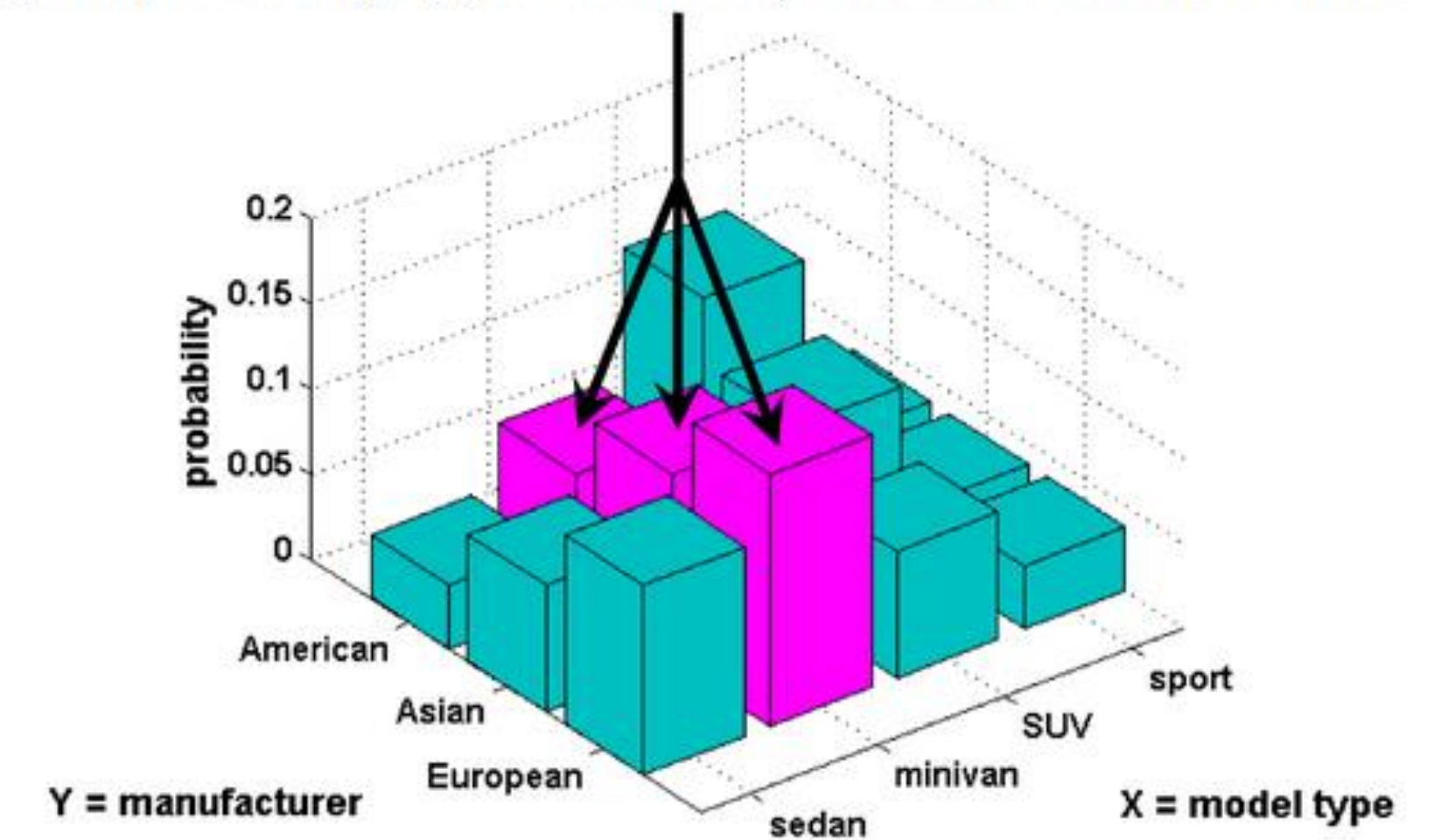
$$p(x) = \begin{cases} \sum_{y \in \Omega_Y} p(x, y) & X, Y \text{ are discrete} \\ \int_{-\infty}^{\infty} p(x, y) dy & X, Y \text{ are continuous} \end{cases}$$

In general:

$$p(x_1, \dots, x_m) = \begin{cases} \sum_{x_{m+1} \in \Omega_{X_{m+1}}} \dots \sum_{x_n \in \Omega_{X_n}} p(x_1, \dots, x_n) & X_1, \dots, X_n \text{ are discrete} \\ \int_{-\infty}^{\infty} \dots \int_{-\infty}^{\infty} p(x_1, \dots, x_n) dx_{m+1} \dots dx_n & X_1, \dots, X_n \text{ are continuous} \end{cases}$$

“Marginalizing out x_{m+1}, \dots, x_n ”

marginal probability: $p(X = \text{minivan}) = 0.0741 + 0.1111 + 0.1481 = 0.33$



Credit: Jeff Howbert

Conditional Probability Distributions

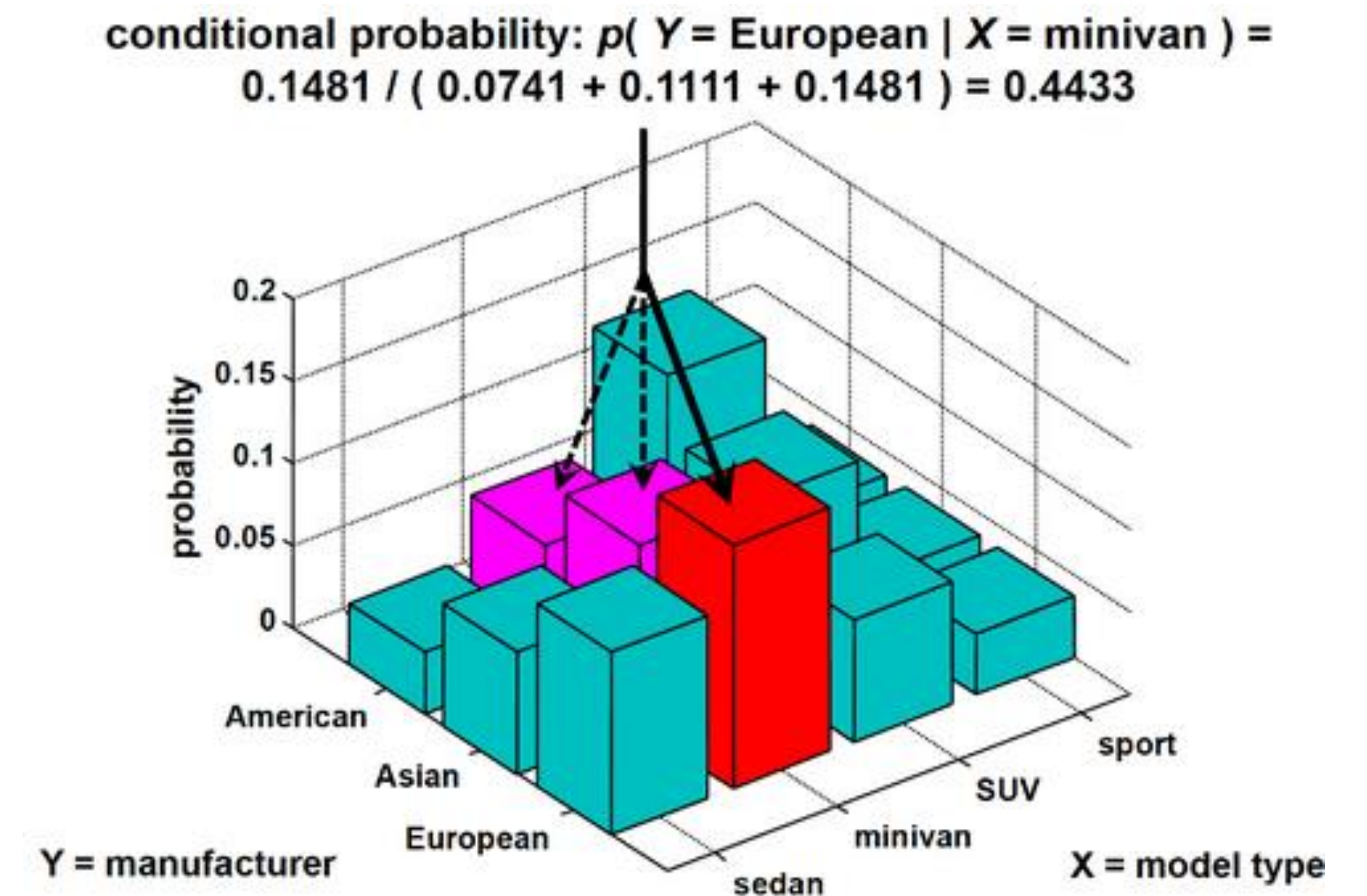
Two random variables:

$$p(y|x) = \frac{p(x, y)}{p(x)} = \begin{cases} \frac{p(x, y)}{\sum_{y \in \Omega_Y} p(x, y)} & X, Y \text{ are discrete} \\ \frac{p(x, y)}{\int_{-\infty}^{\infty} p(x, y) dy} & X, Y \text{ are continuous} \end{cases}$$

In general:

$$p(x_{m+1}, \dots, x_n | x_1, \dots, x_m) = \frac{p(x_1, \dots, x_n)}{p(x_1, \dots, x_m)}$$

$$= \begin{cases} \frac{p(x_1, \dots, x_n)}{\sum_{x_{m+1} \in \Omega_{X_{m+1}}} \dots \sum_{x_n \in \Omega_{X_n}} p(x_1, \dots, x_n)} & X_1, \dots, X_n \text{ are discrete} \\ \frac{p(x_1, \dots, x_n)}{\int_{-\infty}^{\infty} \dots \int_{-\infty}^{\infty} p(x_1, \dots, x_n) dx_{m+1} \dots dx_n} & X_1, \dots, X_n \text{ are continuous} \end{cases}$$



Credit: Jeff Howbert

Chain Rule of Probability

Two random variables:

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

In general:

$$p(x_1, \dots, x_n) = p(x_1)p(x_2|x_1)p(x_3|x_1, x_2) \cdots p(x_n|x_1, \dots, x_{n-1})$$

Chain Rule of Probability (Conditional Case)

Two random variables:

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

In general:

$$\begin{aligned} & p(x_1, \dots, x_n | z_1, \dots, z_l) \\ &= p(x_1 | z_1, \dots, z_l) p(x_2 | x_1, z_1, \dots, z_l) \cdots p(x_n | x_1, \dots, x_{n-1}, z_1, \dots, z_l) \end{aligned}$$

Independence

Two random variables X and Y are independent if:

$$p(y|x) = p(y) \quad \forall x \quad (\text{or equivalently, } p(x|y) = p(x) \quad \forall y)$$

Since $p(x, y) = p(x)p(y|x)$ in general, an equivalent definition is $p(x, y) = p(x)p(y)$

Random variables X_1, \dots, X_n are (mutually) independent if:

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

Conditional Independence

Two random variables X and Y are conditionally independent given $Z = z$ if:

$$p(y|x, z) = p(y|z) \quad \forall x \quad \text{or equivalently, } p(x|y, z) = p(x|z) \quad \forall y$$

Since $p(x, y|z) = p(x|z)p(y|x, z)$ in general, an equivalent definition is $p(x, y|z) = p(x|z)p(y|z)$

Random variables X_1, \dots, X_n are conditionally independent given $Z_1 = z_1, \dots, Z_l = z_l$ if:

$$p(x_1, \dots, x_n | z_1, \dots, z_l) = p(x_1 | z_1, \dots, z_l) \cdots p(x_n | z_1, \dots, z_l)$$

Bayes' Rule

An identity that relates $p(y|x)$ to $p(x|y)$:

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

Expanding $p(x)$ further is often useful:

$$p(y|x) = \frac{p(y)p(x|y)}{p(x)} = \frac{p(y)p(x|y)}{\int_{-\infty}^{\infty} p(x, y) dy} = \frac{p(y)p(x|y)}{\int_{-\infty}^{\infty} p(y)p(x|y) dy} \quad (\text{assuming continuous RVs})$$

True in the conditional case as well: (show this as an exercise)

$$p(y|x, z_1, \dots, z_l) = \frac{p(y|z_1, \dots, z_l)p(x|y, z_1, \dots, z_l)}{p(x|z_1, \dots, z_l)}$$

Expected Value

Two random variables:

$$E[f(X, Y)] = \begin{cases} \sum_{x \in \Omega_X} \sum_{y \in \Omega_Y} f(x, y) p(x, y) & X, Y \text{ discrete} \\ \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} f(x, y) p(x, y) dx dy & X, Y \text{ continuous} \end{cases}$$

In general:

$$E[f(X_1, \dots, X_n)] = \begin{cases} \sum_{x_1 \in \Omega_{X_1}} \dots \sum_{x_n \in \Omega_{X_n}} f(x_1, \dots, x_n) p(x_1, \dots, x_n) & X_i \text{ discrete} \\ \int_{-\infty}^{\infty} \dots \int_{-\infty}^{\infty} f(x_1, \dots, x_n) p(x_1, \dots, x_n) dx_1 \dots dx_n & X_i \text{ continuous} \end{cases}$$

(Technically this is the “law of the unconscious statistician” rather than the definition)

$f(\cdot)$ could be vector-valued, in which case $E[f(X_1, \dots, X_n)] = \begin{pmatrix} E[f_1(X_1, \dots, X_n)] \\ \vdots \\ E[f_m(X_1, \dots, X_n)] \end{pmatrix}$, where $E[f_i(X_1, \dots, X_n)]$ is the i th component of $f(\cdot)$.

Expected Value

Linearity of expectation:

$E[X + Y] = E[X] + E[Y]$ (always true, even if X and Y are dependent)

$$E[cX] = cE[X]$$

Not multiplicative unless independent:

In general, $E[XY] \neq E[X]E[Y]$

However, if X and Y are independent, $E[XY] = E[X]E[Y]$

Moments

Mean: $E[X]$

$$\begin{aligned}\text{Covariance: } \text{Cov}(X, Y) &= E[(X - E[X])(Y - E[Y])] \\ &= E[XY] - E[X]E[Y]\end{aligned}$$

Covariance is symmetric: $\text{Cov}(X, Y) = \text{Cov}(Y, X)$

$$\begin{aligned}\text{Variance: } \text{Var}(X) &:= \text{Cov}(X, X) = E[(X - E[X])^2] \\ &= E[X^2] - (E[X])^2\end{aligned}$$

Standard Deviation: $\sqrt{\text{Var}(X)}$

$$\text{Pearson's Correlation Coefficient: } \rho_{X,Y} = \frac{\text{Cov}(X,Y)}{\sqrt{\text{Var}(X)}\sqrt{\text{Var}(Y)}} \in [-1,1]$$

Zero Covariance vs. Independence

If X and Y are independent,

$$\begin{aligned}\text{Cov}(X, Y) &= E[XY] - E[X]E[Y] \\ &= E[X]E[Y] - E[X]E[Y] \\ &= 0\end{aligned}$$

However, if $\text{Cov}(X, Y) = 0$, X and Y are **not** necessarily independent.

Conditional Expectation

Two random variables X and Y conditioned on $Z = z$:

$$E[f(X, Y)|Z = z] = \begin{cases} \sum_{x \in \Omega_X} \sum_{y \in \Omega_Y} f(x, y)p(x, y|z) & X, Y \text{ discrete} \\ \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} f(x, y)p(x, y|z)dx dy & X, Y \text{ continuous} \end{cases}$$

In general:

$$E[f(X_1, \dots, X_n)|Z_1 = z_1, \dots, Z_l = z_l] \\ = \begin{cases} \sum_{x_1 \in \Omega_{X_1}} \dots \sum_{x_n \in \Omega_{X_n}} f(x_1, \dots, x_n)p(x_1, \dots, x_n|z_1, \dots, z_l) & X_i \text{ discrete} \\ \int_{-\infty}^{\infty} \dots \int_{-\infty}^{\infty} f(x_1, \dots, x_n)p(x_1, \dots, x_n|z_1, \dots, z_l)dx_1 \dots dx_n & X_i \text{ continuous} \end{cases}$$

Conditional Moments

Conditioning on one variable:

Conditional mean: $E[X|Z = z]$

Conditional variance: $\text{Var}(X|Z = z) = E[(X - E[X|Z = z])^2|Z = z]$

In general:

Conditional mean: $E[X|Z_1 = z_1, \dots, Z_l = z_l]$

Conditional variance:

$$\begin{aligned} &\text{Var}(X|Z_1 = z_1, \dots, Z_l = z_l) \\ &= E[(X - E[X|Z_1 = z_1, \dots, Z_l = z_l])^2|Z_1 = z_1, \dots, Z_l = z_l] \end{aligned}$$

Law of total expectation:

$$E_{Z_1, \dots, Z_l}[E_X[f(X, Z_1, \dots, Z_l)|Z_1, \dots, Z_l]] = E_{X, Z_1, \dots, Z_l}[f(X, Z_1, \dots, Z_l)]$$

Entropy

Entropy measures the amount of uncertainty in a discrete distribution.

For a **discrete** RV X , the entropy of X is defined as:

$$H(X) = E[-\log_b p(X)] = - \sum_{x \in \Omega_X} p(x) \log_b(p(x))$$

(When evaluating the above expression, $0 \log 0$ should be treated as if it evaluates to 0)

Typically the base b of the logarithm is e or 2. The units of entropy are known as “nats” if the base is e , and “bits” if the base is 2. When the base is not specified, in ML, typically the base is assumed to be e .

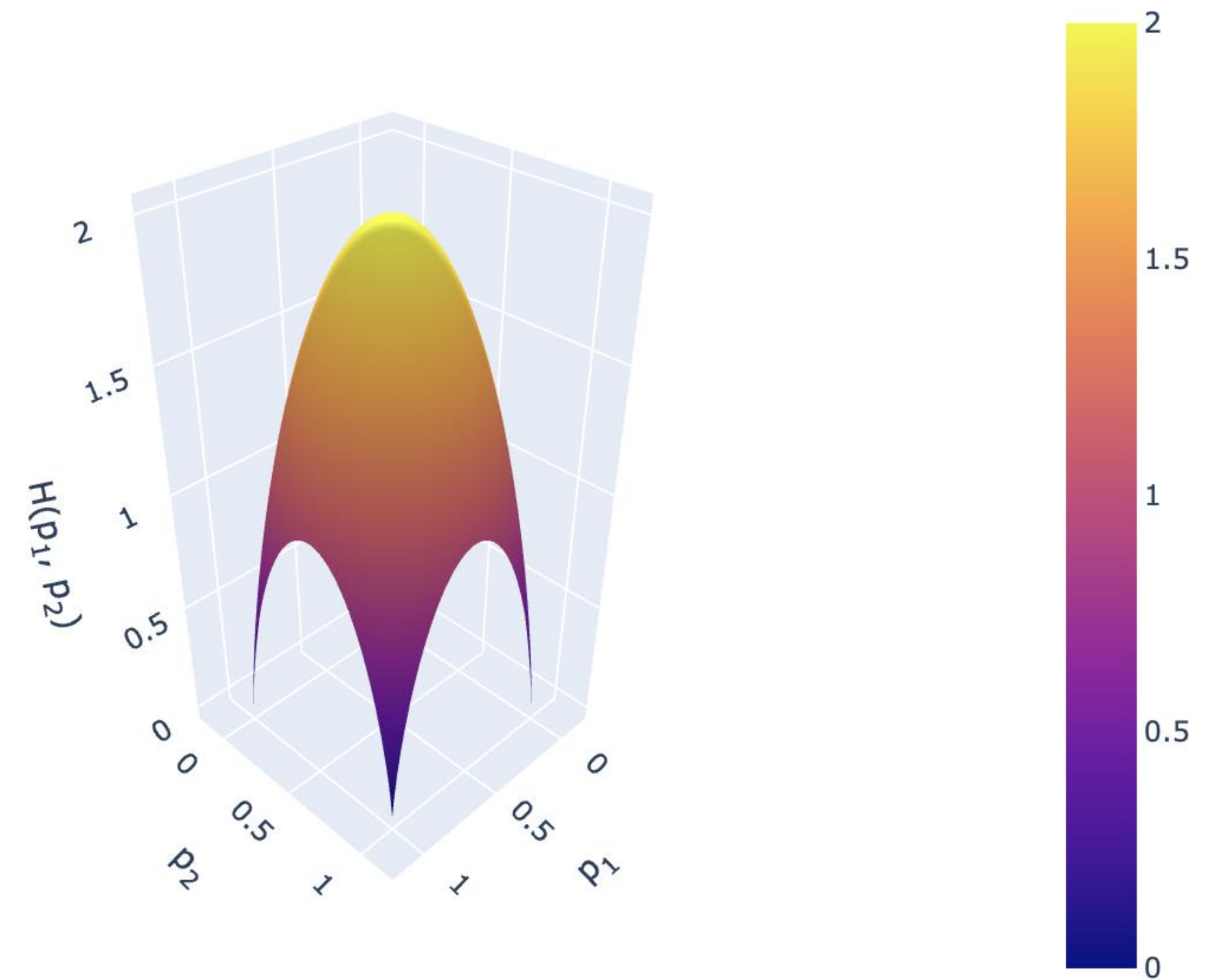
Entropy

Properties:

For any discrete RV X , $H(X) \geq 0$.

$H(X) = 0$ if and only if X is deterministic.

$H(X)$ is maximized when $p(x)$ is the same for all $x \in \Omega_X$ (i.e.: when the distribution is discrete uniform)



Credit: Ethan Weinberger

Joint Entropy

Joint entropy measures the total amount of uncertainty in a discrete joint distribution.

Two **discrete** RVs:

$$H(X, Y) = E[-\log_b p(X, Y)] = - \sum_{x \in \Omega_X} \sum_{y \in \Omega_Y} p(x, y) \log_b(p(x, y))$$

In general:

$$\begin{aligned} H(X_1, \dots, X_n) &= E[-\log_b p(X_1, \dots, X_n)] \\ &= - \sum_{x_1 \in \Omega_{X_1}} \dots \sum_{x_n \in \Omega_{X_n}} p(x_1, \dots, x_n) \log_b(p(x_1, \dots, x_n)) \end{aligned}$$

(When evaluating the above expressions, $0 \log 0$ should be treated as if it evaluates to 0)

Conditional Entropy

Two **discrete** RVs:

$$H(X|Y = y) = E_X[-\log_b p(X|y)|Y = y] = - \sum_{x \in \Omega_X} p(x|y) \log_b(p(x|y))$$

$$\begin{aligned} H(X|Y) &= E_Y[E_X[-\log_b p(X|Y)|Y]] \\ &= E_{X,Y}[-\log_b p(X|Y)] \end{aligned}$$

$$= - \sum_{x \in \Omega_X} \sum_{y \in \Omega_Y} p(x, y) \log_b(p(x|y))$$

$$= - \sum_{x \in \Omega_X} \sum_{y \in \Omega_Y} (p(x, y) \log_b(p(x, y)) - p(x, y) \log_b(p(y)))$$

(When evaluating the above expressions, $0 \log 0$ should be treated as if it evaluates to 0)

Conditional Entropy

In general:

$$H(X|Y_1 = y_1, \dots, Y_l = y_l) = E_X[-\log_b p(X|y_1, \dots, y_l)|Y_1 = y_1, \dots, Y_l = y_l]$$

$$= - \sum_{x \in \Omega_X} p(x|y_1, \dots, y_l) \log_b(p(x|y_1, \dots, y_l))$$

$$H(X|Y_1, \dots, Y_l) = E_{Y_1, \dots, Y_l}[E_X[-\log_b p(X|Y_1, \dots, Y_l)|Y_1, \dots, Y_l]]$$

$$= E_{X, Y_1, \dots, Y_l}[-\log_b p(X|Y_1, \dots, Y_l)]$$

$$= - \sum_{x \in \Omega_X} \sum_{y_1 \in \Omega_{Y_1}} \dots \sum_{y_l \in \Omega_{Y_l}} p(x, y_1, \dots, y_l) \log_b(p(x|y_1, \dots, y_l))$$

$$= - \sum_{x \in \Omega_X} \sum_{y_1 \in \Omega_{Y_1}} \dots \sum_{y_l \in \Omega_{Y_l}} (p(x, y_1, \dots, y_l) \log_b(p(x, y_1, \dots, y_l)) - p(x, y_1, \dots, y_l) \log_b(p(y_1, \dots, y_l)))$$

(When evaluating the above expressions, $0 \log 0$ should be treated as if it evaluates to 0)

Relative Entropy / KL Divergence

KL divergence stands for Kullback-Leibler divergence.

Recall definition of entropy:

$$H(X) = E[-\log_b p(X)] = - \sum_{x \in \Omega_X} p(x) \log_b(p(x))$$

Consider two distributions P and Q where the support of P is a subset of support of Q

$$D_{KL}(P \parallel Q) = - \sum_{x \in \Omega_X} p(x) \log_b \left(\frac{q(x)}{p(x)} \right)$$

When evaluating the above expressions, $0 \log 0$ should be treated as if it evaluates to 0)

Mutual Information

$$\begin{aligned} I(X; Y) &= H(X) - H(X|Y) \\ &= H(Y) - H(Y|X) \\ &= \sum_{x \in \Omega_X} \sum_{y \in \Omega_Y} p(x, y) \log_b \left(\frac{p(x, y)}{p(x)p(y)} \right) \end{aligned}$$

(When evaluating the above expressions, $0 \log 0$ should be treated as if it evaluates to 0)

Vector Notation

We can arrange multiple random variables X_1, \dots, X_n as a vector:

$$\vec{X} = \begin{pmatrix} X_1 \\ \vdots \\ X_n \end{pmatrix}$$

We can arrange the means of each RV into a vector as well, which can be represented as

$$E[\vec{X}] = \begin{pmatrix} E[X_1] \\ \vdots \\ E[X_n] \end{pmatrix} \quad \text{“Mean vector” or just the “mean”}$$

The covariances and variances can be arranged into a matrix:

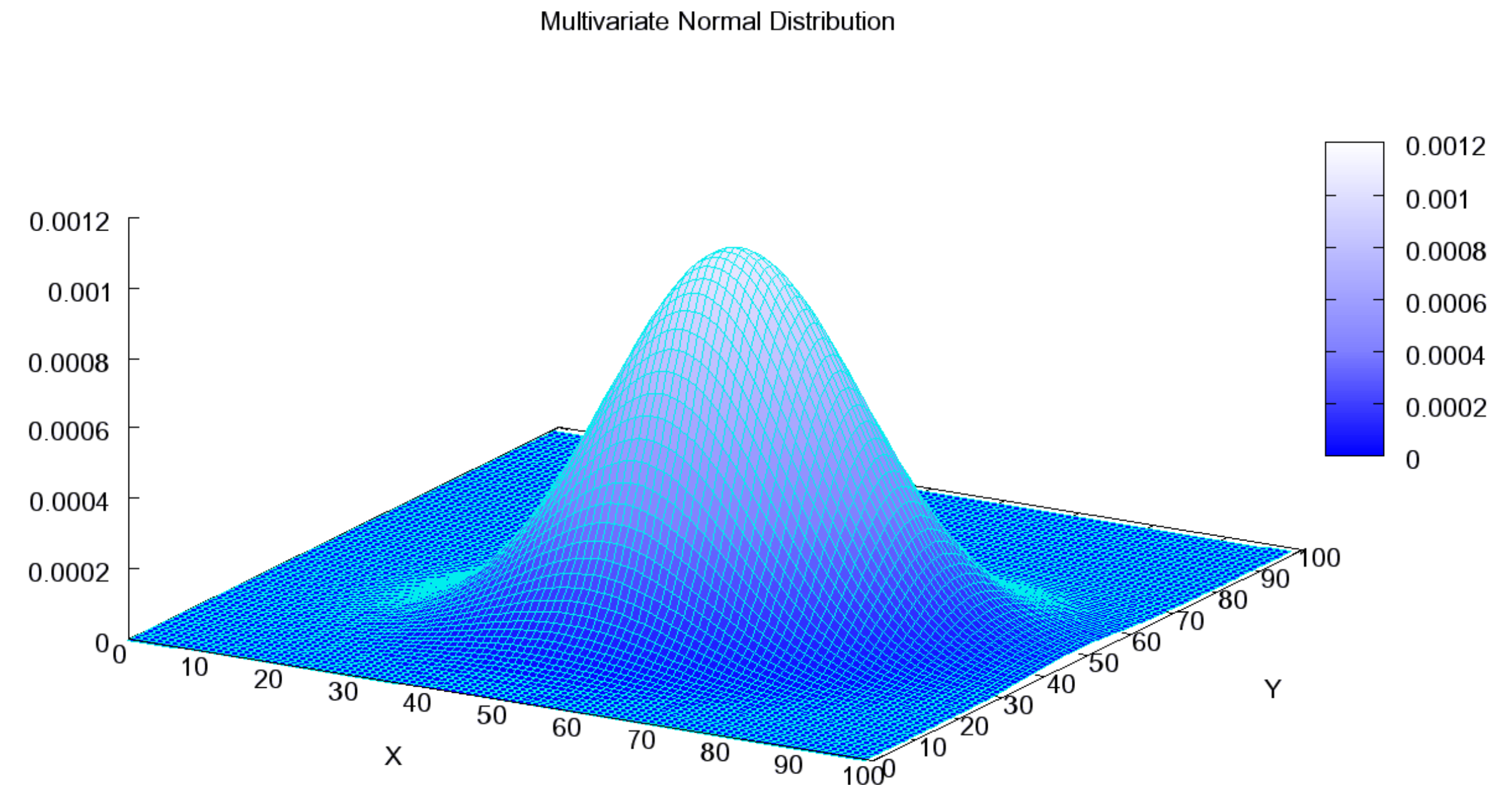
$$E[(\vec{X} - E[\vec{X}])(\vec{X} - E[\vec{X}])^\top] = E[\vec{X}\vec{X}^\top] - (E[\vec{X}]) (E[\vec{X}])^\top$$

“Covariance matrix” or just the “covariance”

Multivariate Normal Distribution

Generalization of the normal distribution to multiple random variables.

In ML, commonly referred to as a multivariate Gaussian distribution or simply Gaussian distribution.



Credit: Wikipedia

Multivariate Normal Distribution

Univariate normal:

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

Multivariate normal:

$\vec{X} = \begin{pmatrix} X_1 \\ \vdots \\ X_n \end{pmatrix} \sim \mathcal{N}(\vec{\mu}, \Sigma)$, where $\vec{\mu}$ denotes the mean vector and Σ denotes a

positive definite covariance matrix.

Quadratic form!

$$p(\vec{x}) = \frac{1}{\sqrt{(2\pi)^n \det(\Sigma)}} \exp\left(-\frac{1}{2} (\vec{x} - \vec{\mu})^\top \Sigma^{-1} (\vec{x} - \vec{\mu})\right)$$

Multivariate Normal Distribution

$$p(\vec{x}) = \frac{1}{\sqrt{(2\pi)^n \det(\Sigma)}} \exp\left(-\frac{1}{2}(\vec{x} - \vec{\mu})^\top \Sigma^{-1}(\vec{x} - \vec{\mu})\right)$$

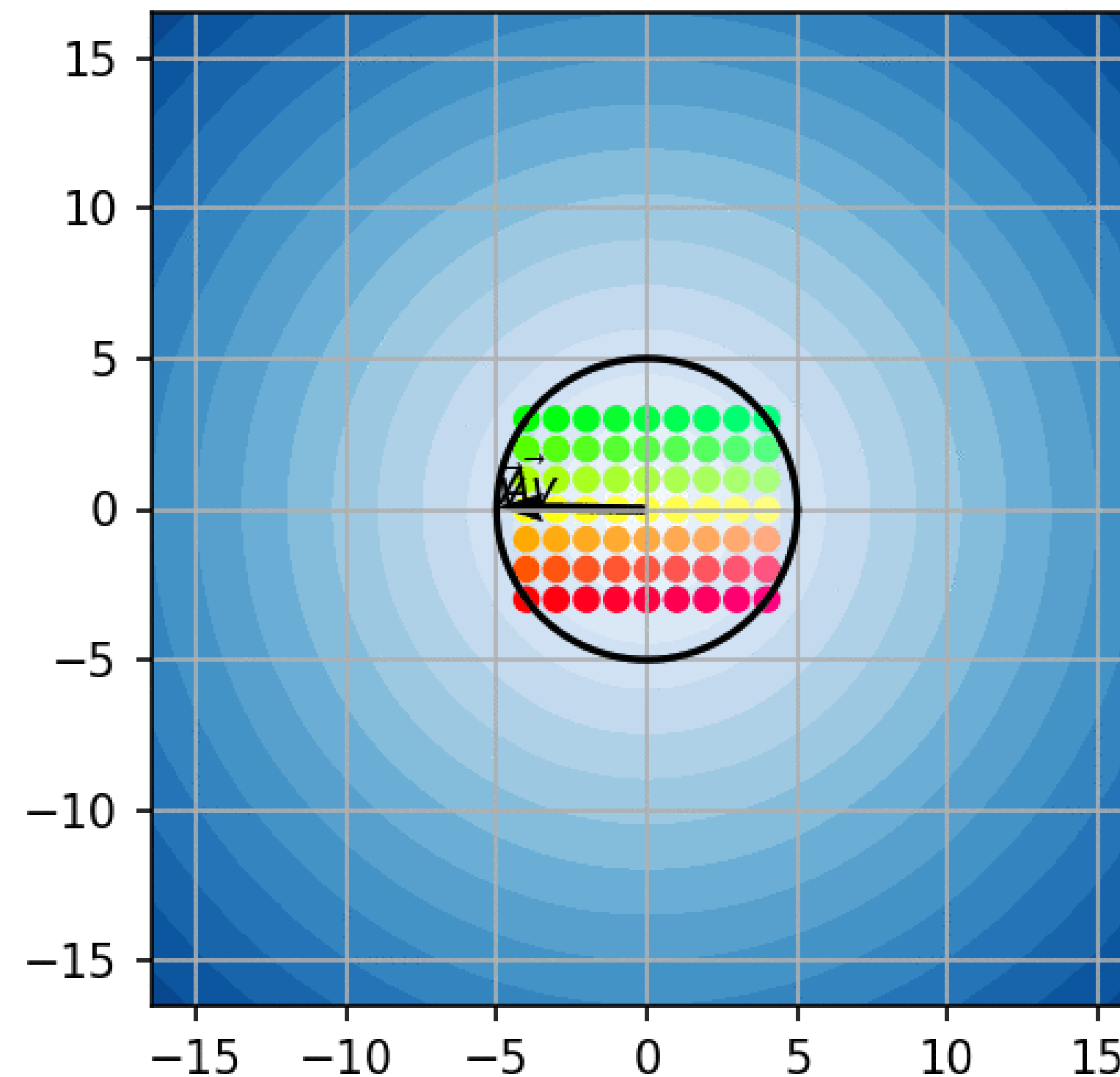
How does the quadratic form $(\vec{x} - \vec{\mu})^\top \Sigma^{-1}(\vec{x} - \vec{\mu})$ behave?

Recall: The right-singular vector of a matrix A with the largest singular value is the direction along which a unit vector becomes the longest after being transformed by A .

$$\vec{v}_{.1} = \arg \max_{\vec{x}: \|\vec{x}\|_2=1} \|A\vec{x}\|_2 = \arg \max_{\vec{x}: \|\vec{x}\|_2=1} \|A\vec{x}\|_2^2$$

$$\|A\vec{x}\|_2^2 = (A\vec{x})^\top (A\vec{x}) = \vec{x}^\top (A^\top A)\vec{x}$$

This is a quadratic form! The direction along which a vector grows the most is given by the first right-singular vector of A .



\vec{v} - the right singular vector
 $A\vec{v}$ - after matrix transformation

$$A = \begin{pmatrix} 1 & 2 \\ 0 & 1 \end{pmatrix}$$

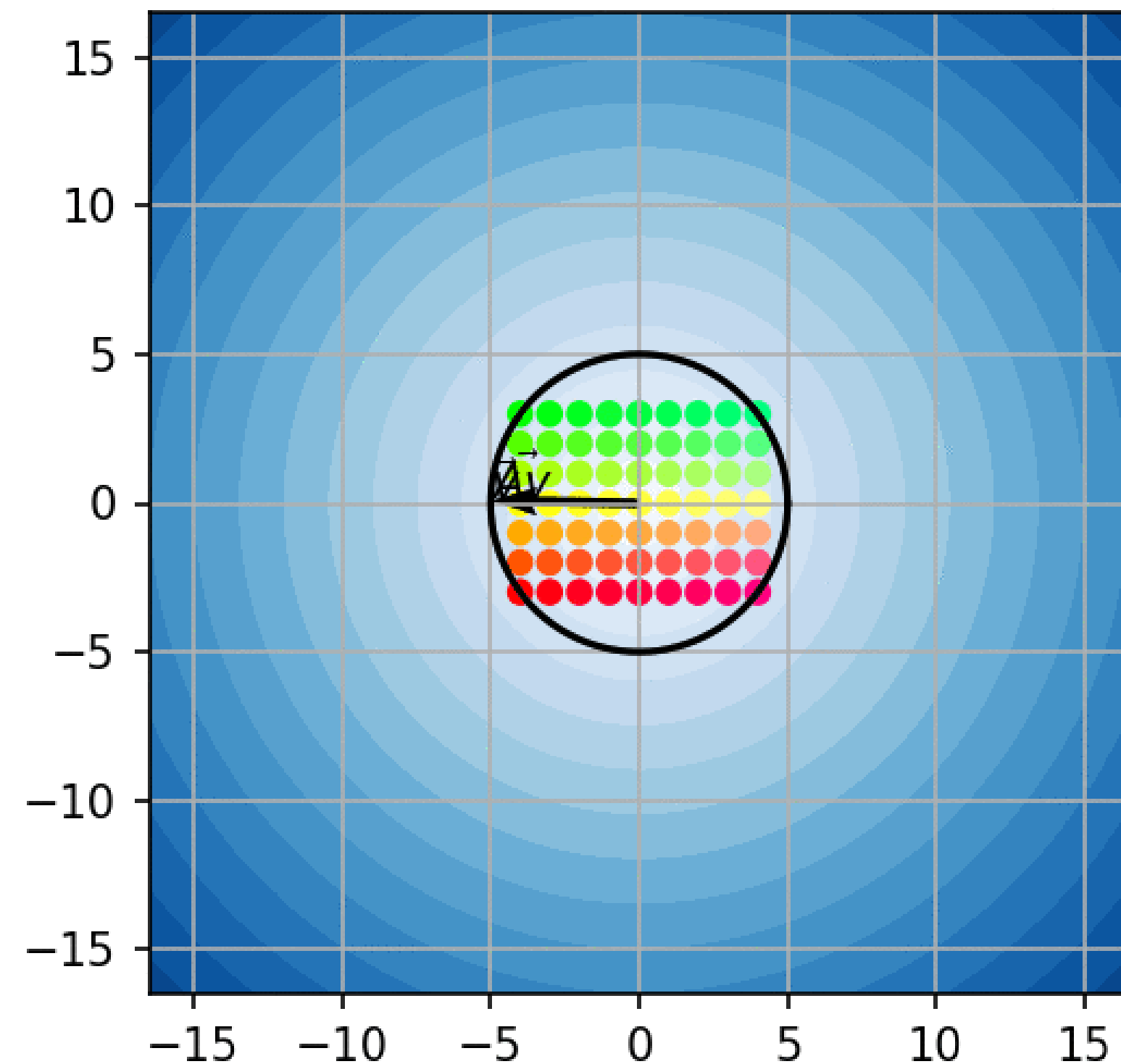
Multivariate Normal Distribution

$$p(\vec{x}) = \frac{1}{\sqrt{(2\pi)^n \det(\Sigma)}} \exp\left(-\frac{1}{2}(\vec{x} - \vec{\mu})^\top \Sigma^{-1}(\vec{x} - \vec{\mu})\right)$$

How does the quadratic form $(\vec{x} - \vec{\mu})^\top \Sigma^{-1}(\vec{x} - \vec{\mu})$ behave?

Problem: In the Gaussian density, we don't have separate matrices A^\top and A ; instead, we are only given the product $A^\top A =: \Sigma^{-1}$.

Recall: the right-singular vectors of A are the eigenvectors of $A^\top A$. So the direction along which a vector grows the most is given by the eigenvector of $A^\top A =: \Sigma^{-1}$ with the largest eigenvalue.



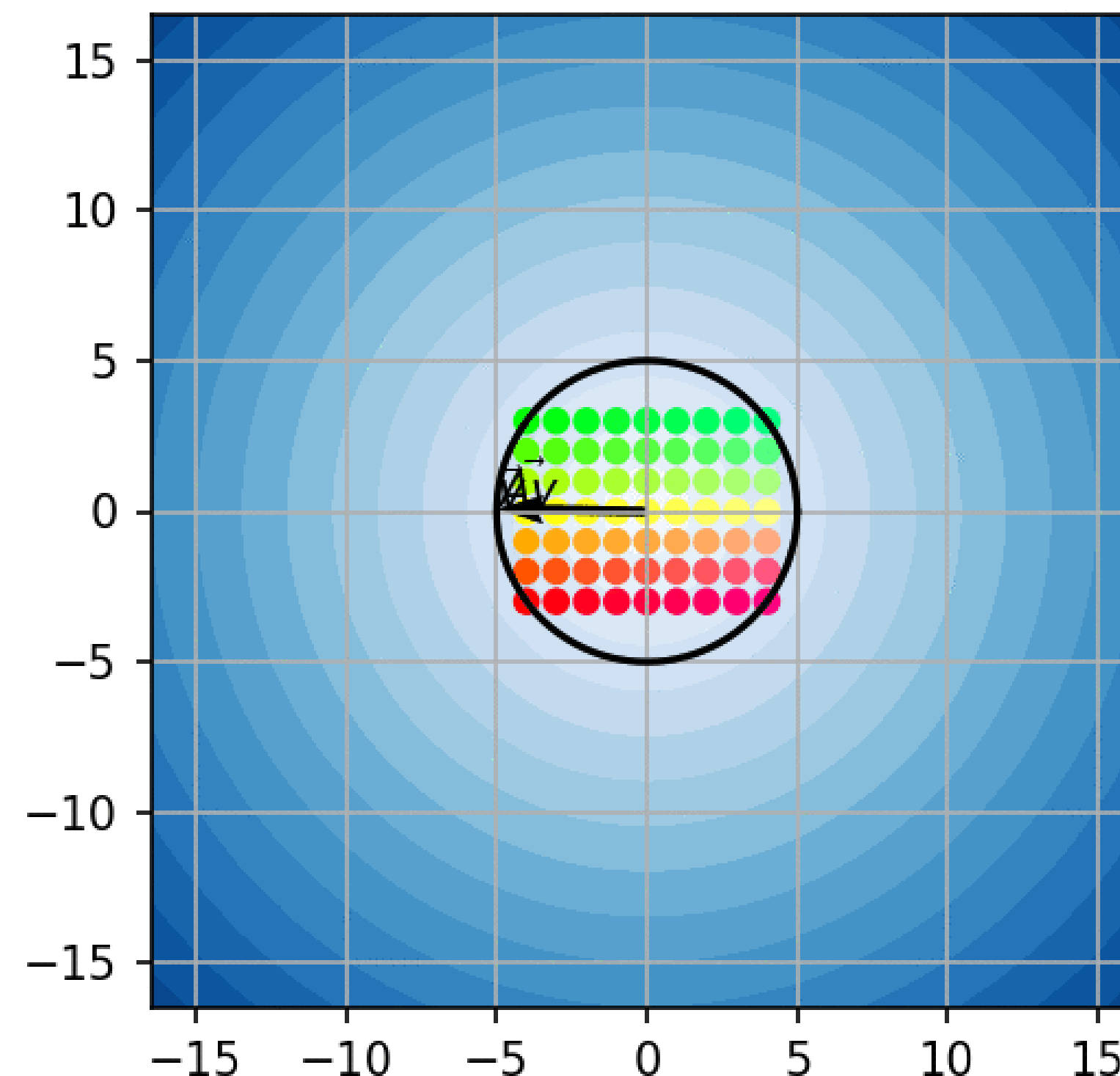
\vec{v} - the right singular vector
 $A\vec{v}$ - after matrix transformation

Multivariate Normal Distribution

$$p(\vec{x}) = \frac{1}{\sqrt{(2\pi)^n \det(\Sigma)}} \exp\left(-\frac{1}{2}(\vec{x} - \vec{\mu})^\top \Sigma^{-1}(\vec{x} - \vec{\mu})\right)$$

How does the quadratic form $(\vec{x} - \vec{\mu})^\top \Sigma^{-1}(\vec{x} - \vec{\mu})$ behave?

Because Σ is symmetric, recall that $\Sigma^{-1} = U\Lambda^{-1}U^\top$, where U denotes the eigenvector matrix of. Hence the eigenvector of Σ^{-1} with the largest eigenvalue is the eigenvector of Σ with the smallest eigenvalue.



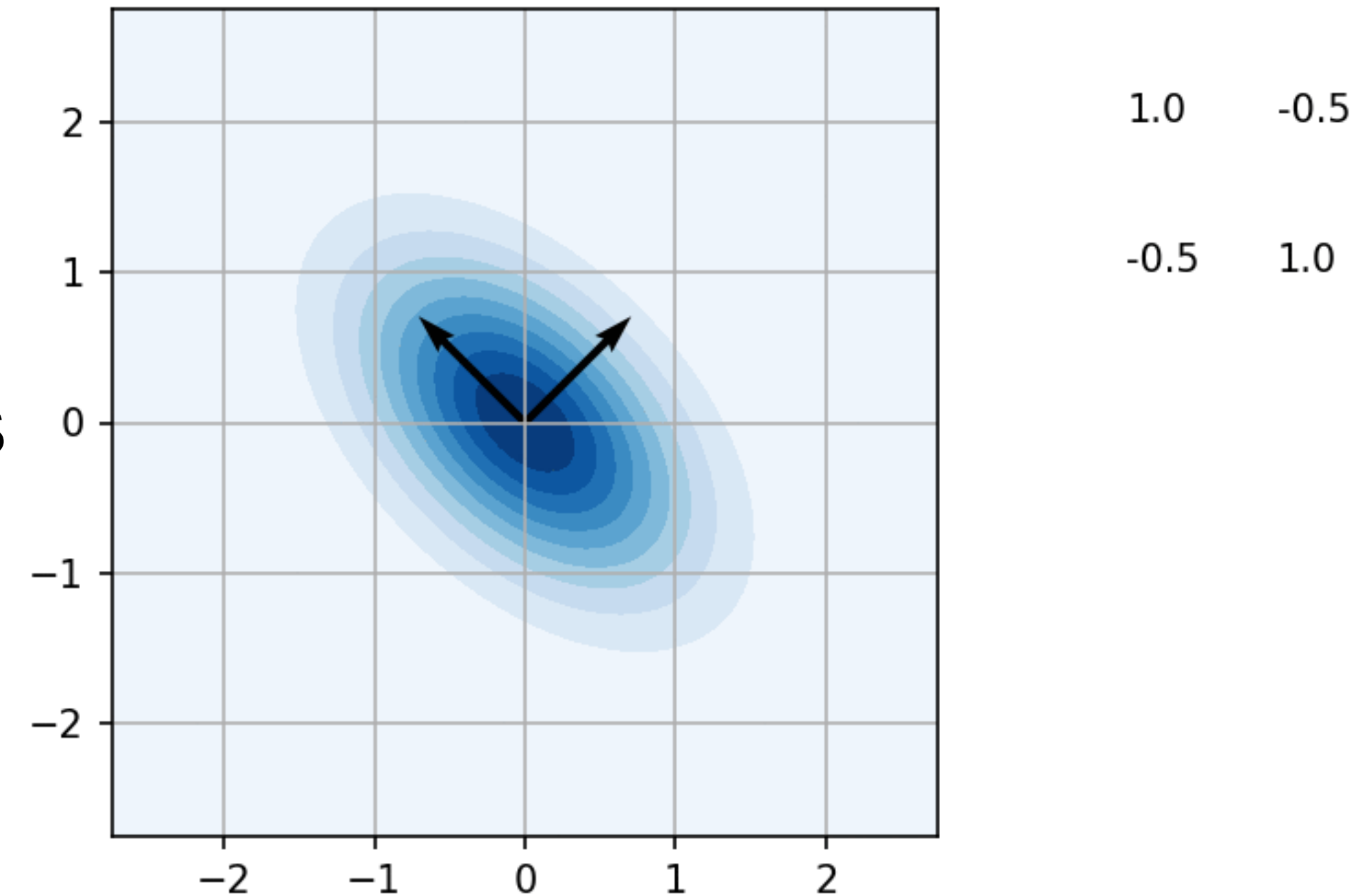
\vec{v} - the right singular vector
 $A\vec{v}$ - after matrix transformation

Multivariate Normal Distribution

$$p(\vec{x}) = \frac{1}{\sqrt{(2\pi)^n \det(\Sigma)}} \exp\left(-\frac{1}{2}(\vec{x} - \vec{\mu})^\top \Sigma^{-1}(\vec{x} - \vec{\mu})\right)$$

The black arrows denote the eigenvectors of Σ .

As shown, they correspond to the principal axes of the elliptical contours.

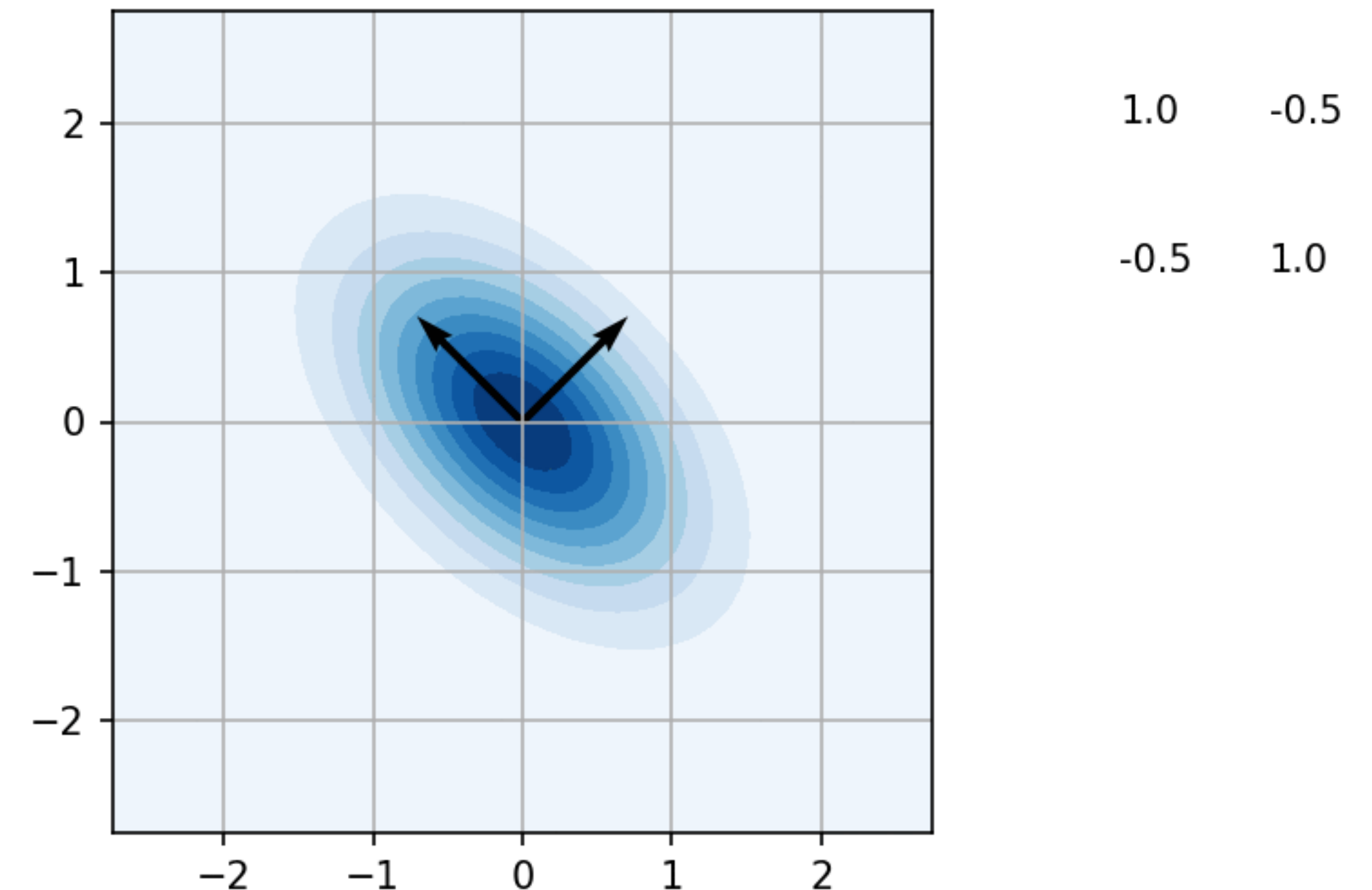


Multivariate Normal Distribution

$$p(\vec{x}) = \frac{1}{\sqrt{(2\pi)^n \det(\Sigma)}} \exp\left(-\frac{1}{2}(\vec{x} - \vec{\mu})^\top \Sigma^{-1}(\vec{x} - \vec{\mu})\right)$$

Now we visualize the Gaussian as the off-diagonal entries of the covariance matrix changes.

$$\Sigma = \begin{pmatrix} \text{Var}(X_1) & \text{Cov}(X_1, X_2) \\ \text{Cov}(X_1, X_2) & \text{Var}(X_2) \end{pmatrix} = \begin{pmatrix} 1 & \alpha \\ \alpha & 1 \end{pmatrix}$$

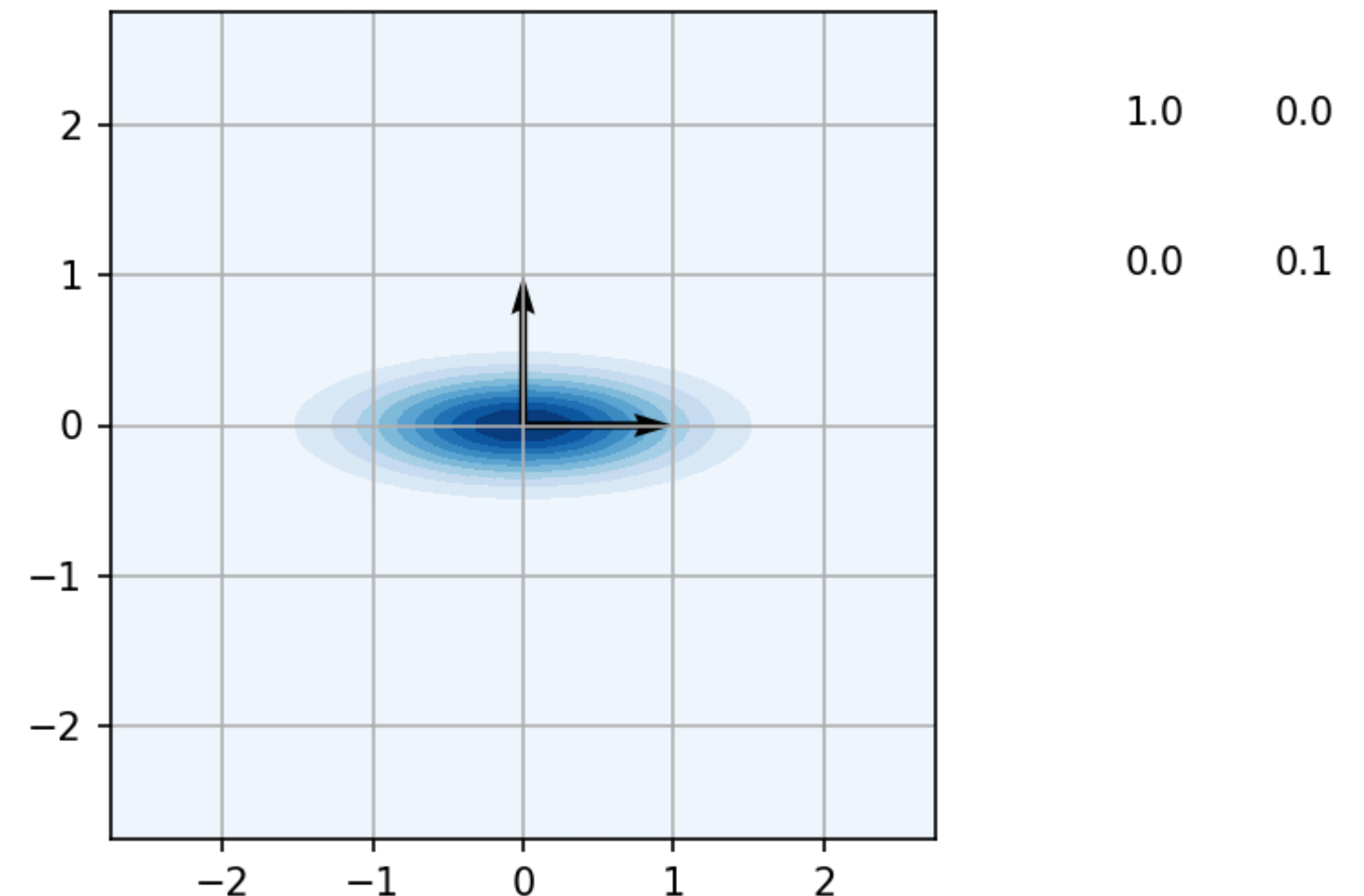


Multivariate Normal Distribution

$$p(\vec{x}) = \frac{1}{\sqrt{(2\pi)^n \det(\Sigma)}} \exp\left(-\frac{1}{2}(\vec{x} - \vec{\mu})^\top \Sigma^{-1}(\vec{x} - \vec{\mu})\right)$$

Now we visualize the Gaussian as one of the diagonal entries of the covariance matrix changes.

$$\Sigma = \begin{pmatrix} \text{Var}(X_1) & \text{Cov}(X_1, X_2) \\ \text{Cov}(X_1, X_2) & \text{Var}(X_2) \end{pmatrix} = \begin{pmatrix} 1 & 0 \\ 0 & \alpha \end{pmatrix}$$



Transformations of Multivariate Normals

If $\vec{x} \sim \mathcal{N}(\vec{\mu}, \Sigma)$, $\vec{x} + \vec{c} \sim \mathcal{N}(\vec{\mu} + \vec{c}, \Sigma)$

If $\vec{x} \sim \mathcal{N}(\vec{\mu}, \Sigma)$, $A\vec{x} \sim \mathcal{N}(A\vec{\mu}, A\Sigma A^\top)$

Special case: If $\vec{x} \sim \mathcal{N}(\vec{\mu}, \Sigma)$, $c\vec{x} \sim \mathcal{N}(c\vec{\mu}, c^2\Sigma)$

If $\vec{x} \perp \vec{y}$ (\vec{x} and \vec{y} are independent), $\vec{x} \sim \mathcal{N}(\vec{\mu}_X, \Sigma_X)$ and $\vec{y} \sim \mathcal{N}(\vec{\mu}_Y, \Sigma_Y)$, $\vec{x} + \vec{y} \sim \mathcal{N}(\vec{\mu}_X + \vec{\mu}_Y, \Sigma_X + \Sigma_Y)$

Standard multivariate normal: $\vec{z} \sim \mathcal{N}(\vec{0}, I)$

$\vec{z} + \vec{\mu} \sim \mathcal{N}(\vec{\mu}, I)$ and $\sigma\vec{z} \sim \mathcal{N}(\vec{0}, \sigma^2 I) \Rightarrow \vec{\mu} + \sigma\vec{z} \sim \mathcal{N}(\vec{\mu}, \sigma^2 I)$

Compare: Standard (univariate) normal: $Z \sim \mathcal{N}(0, 1)$

$Z + \mu \sim \mathcal{N}(\mu, 1)$ and $\sigma Z \sim \mathcal{N}(0, \sigma^2) \Rightarrow \mu + \sigma Z \sim \mathcal{N}(\mu, \sigma^2)$

“Isotropic Gaussian”

(Variance along every direction is the same)

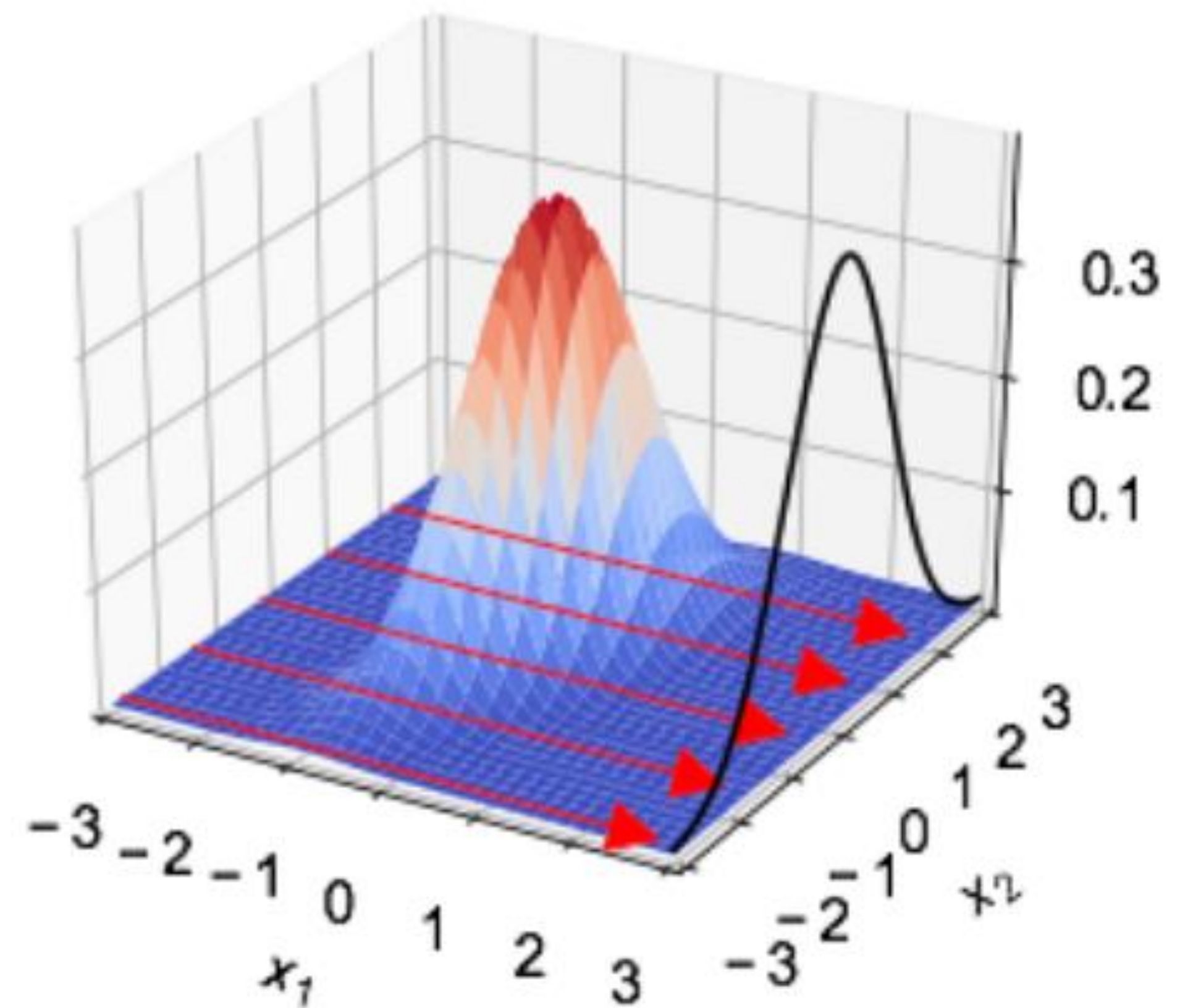
Marginalization of Multivariate Normals

Let $\vec{x} = \begin{pmatrix} \vec{x}_A \\ \vec{x}_B \end{pmatrix}$, where \vec{x}_A and \vec{x}_B correspond to a block of elements of \vec{x}

$$\text{If } \vec{x} = \begin{pmatrix} \vec{x}_A \\ \vec{x}_B \end{pmatrix} \sim \mathcal{N} \left(\begin{pmatrix} \vec{\mu}_A \\ \vec{\mu}_B \end{pmatrix}, \begin{pmatrix} \Sigma_{AA} & \Sigma_{AB} \\ \Sigma_{BA} & \Sigma_{BB} \end{pmatrix} \right),$$

$$\vec{x}_A \sim \mathcal{N}(\vec{\mu}_A, \Sigma_{AA})$$

$$\vec{x}_B \sim \mathcal{N}(\vec{\mu}_B, \Sigma_{BB})$$



Credit: Kris Hauser

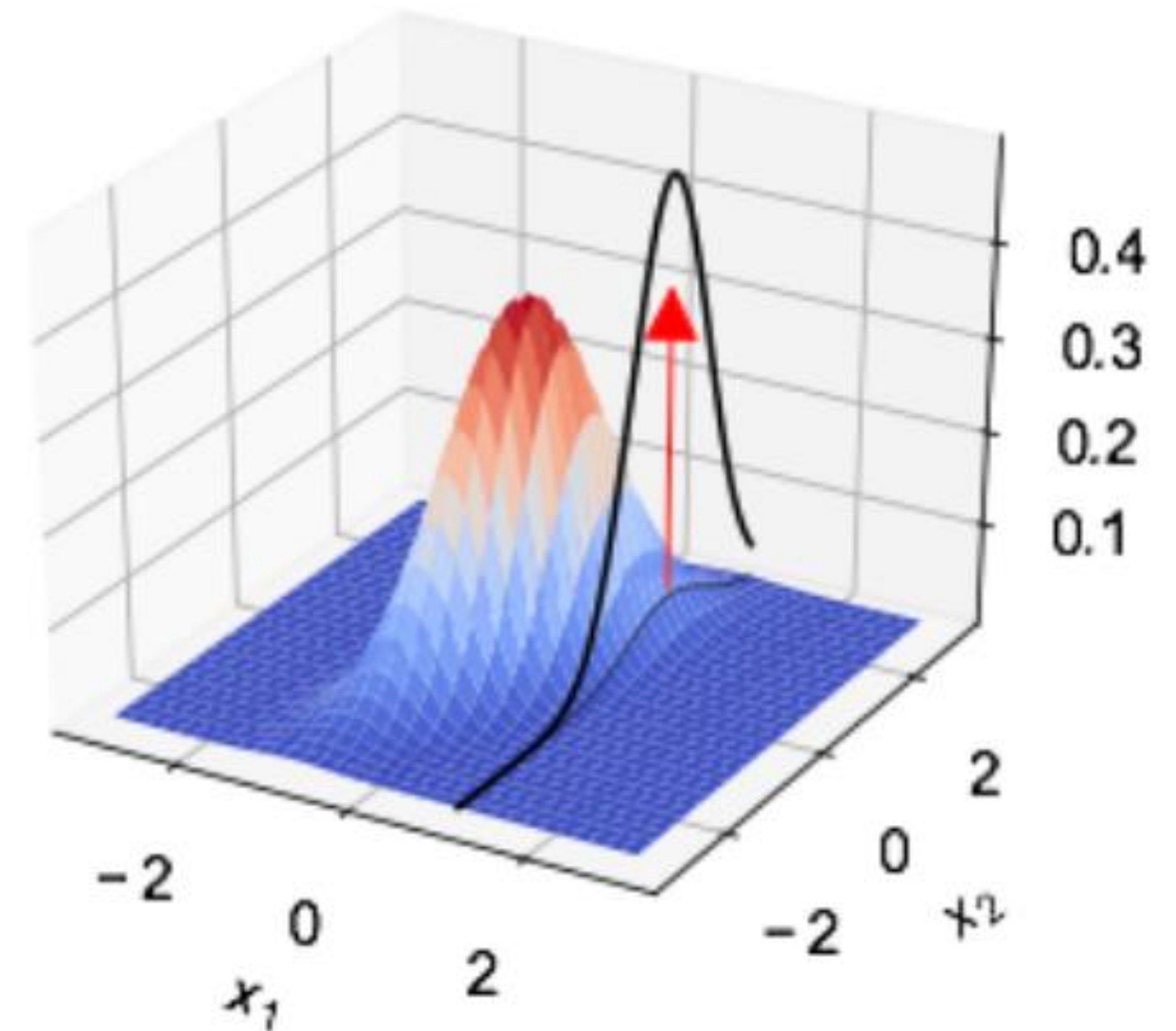
Conditioning of Multivariate Normals

Let $\vec{x} = \begin{pmatrix} \vec{x}_A \\ \vec{x}_B \end{pmatrix}$, where \vec{x}_A and \vec{x}_B correspond to a block of elements of \vec{x}

$$\text{If } \vec{x} = \begin{pmatrix} \vec{x}_A \\ \vec{x}_B \end{pmatrix} \sim \mathcal{N} \left(\begin{pmatrix} \vec{\mu}_A \\ \vec{\mu}_B \end{pmatrix}, \begin{pmatrix} \Sigma_{AA} & \Sigma_{AB} \\ \Sigma_{BA} & \Sigma_{BB} \end{pmatrix} \right),$$

$$\vec{x}_A | \vec{x}_B \sim \mathcal{N}(\vec{\mu}_A + \Sigma_{AB} \Sigma_{BB}^{-1} (\vec{x}_B - \vec{\mu}_B), \Sigma_{AA} - \Sigma_{AB} \Sigma_{BB}^{-1} \Sigma_{BA})$$

$$\vec{x}_B | \vec{x}_A \sim \mathcal{N}(\vec{\mu}_B + \Sigma_{BA} \Sigma_{AA}^{-1} (\vec{x}_A - \vec{\mu}_A), \Sigma_{BB} - \Sigma_{BA} \Sigma_{AA}^{-1} \Sigma_{AB})$$



Credit: Kris Hauser