# Quiz Practice

**Q1:** Which of the following facts about ridge regression is NOT true?

(A) Ridge regression is less prone to overfitting compared to ordinary least squares

(B) Ridge regression always has a unique optimal parameter vector

(C) Compared to ordinary least squares, ridge regression adds a regularizer

(D) Ridge regression uses more hyperparameters than ordinary least squares

(E) Ridge regression uses more parameters than ordinary least squares

(F) Ridge regression uses a strictly convex loss function

(G) All of the above are true

# Machine Learning

## CMPT 726

Mo Chen
SFU School of Computing Science
2021-10-18

# 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
  $E.g.$: $X_i = \begin{cases} 1 & i\text{th coin turns out to be heads} \\ 0 & \text{otherwise} \end{cases}$, 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** $\mathrm{supp}(X)$**:** 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**

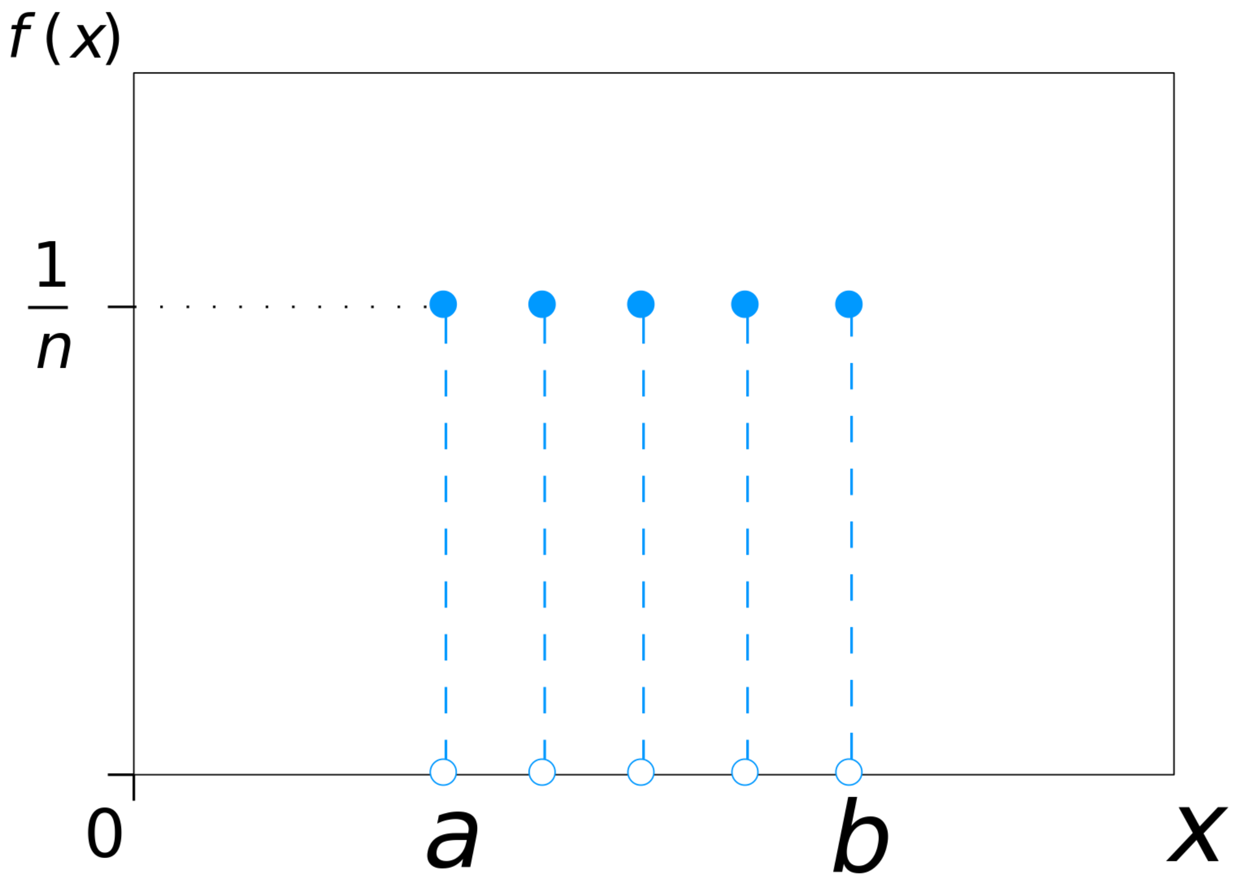
- Cumulative distribution functions (cdf): $F_X(x) = \Pr(X \leq x)$
- Probability mass functions (pmf): $p_X(x) = \Pr(X = x)$
- Examples: Bernoulli RVs, Categorical RVs

**Continuous random variables**

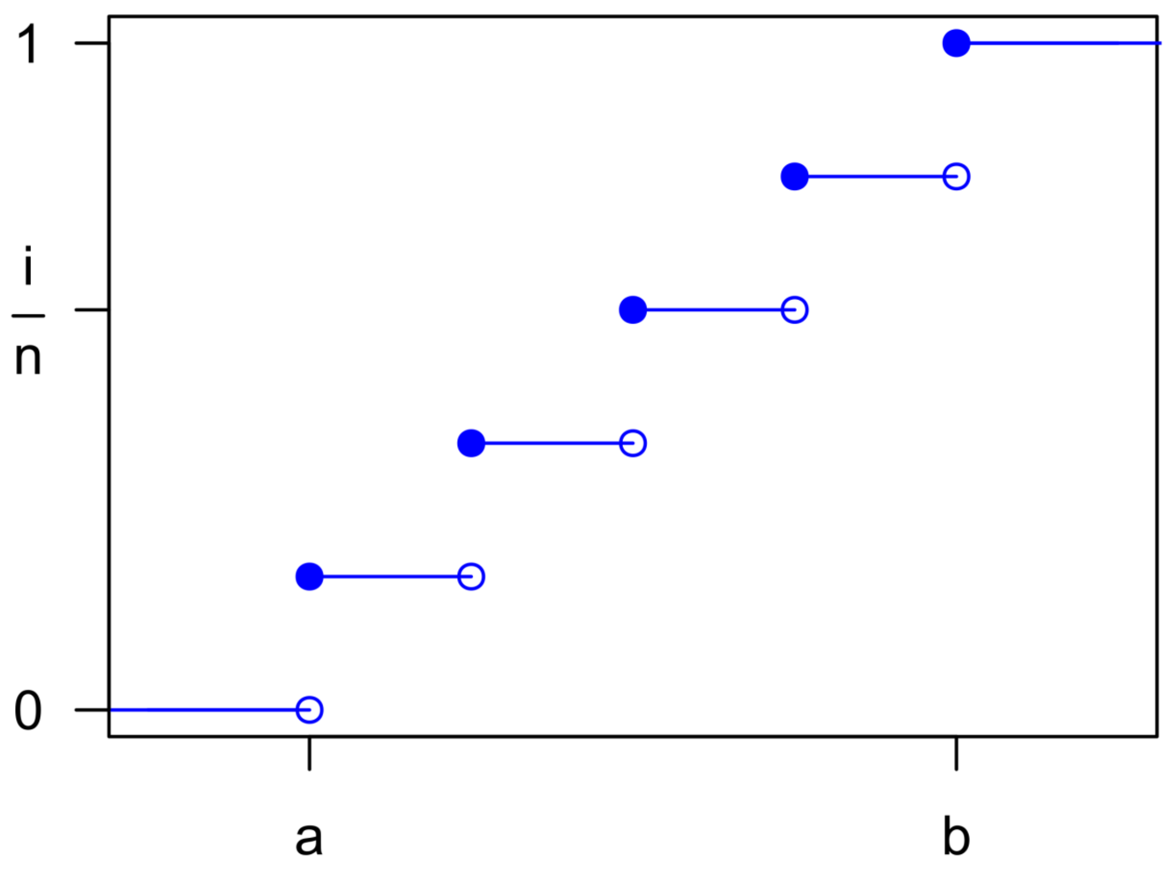
- Cumulative distribution functions (cdf): $F_X(x) = \Pr(X \leq x)$
- Probability density functions (pdf): $f_X(x) = \frac{d}{dx} F_X(x)$
- Examples: Uniform RVs, Normal RVs

# Discrete vs. Continuous RVs

**Discrete random variables**



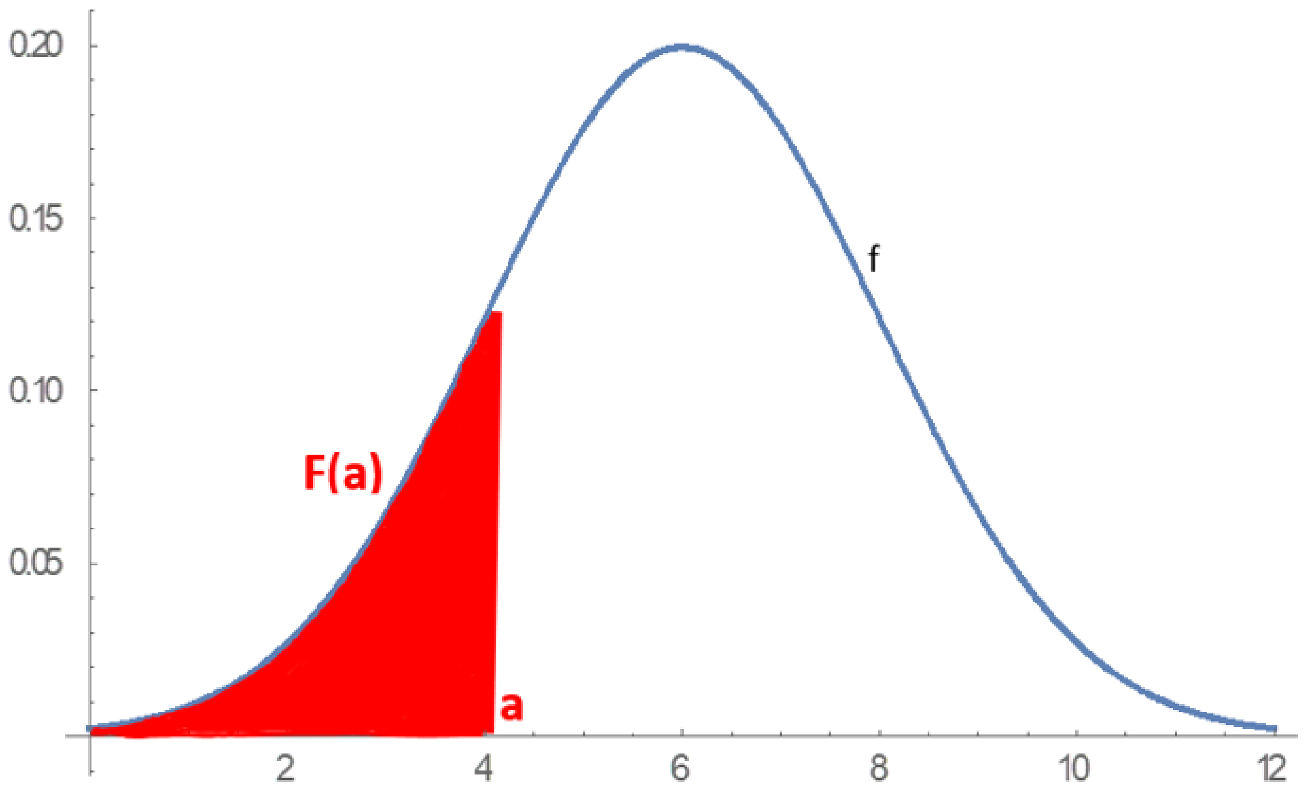

pmf



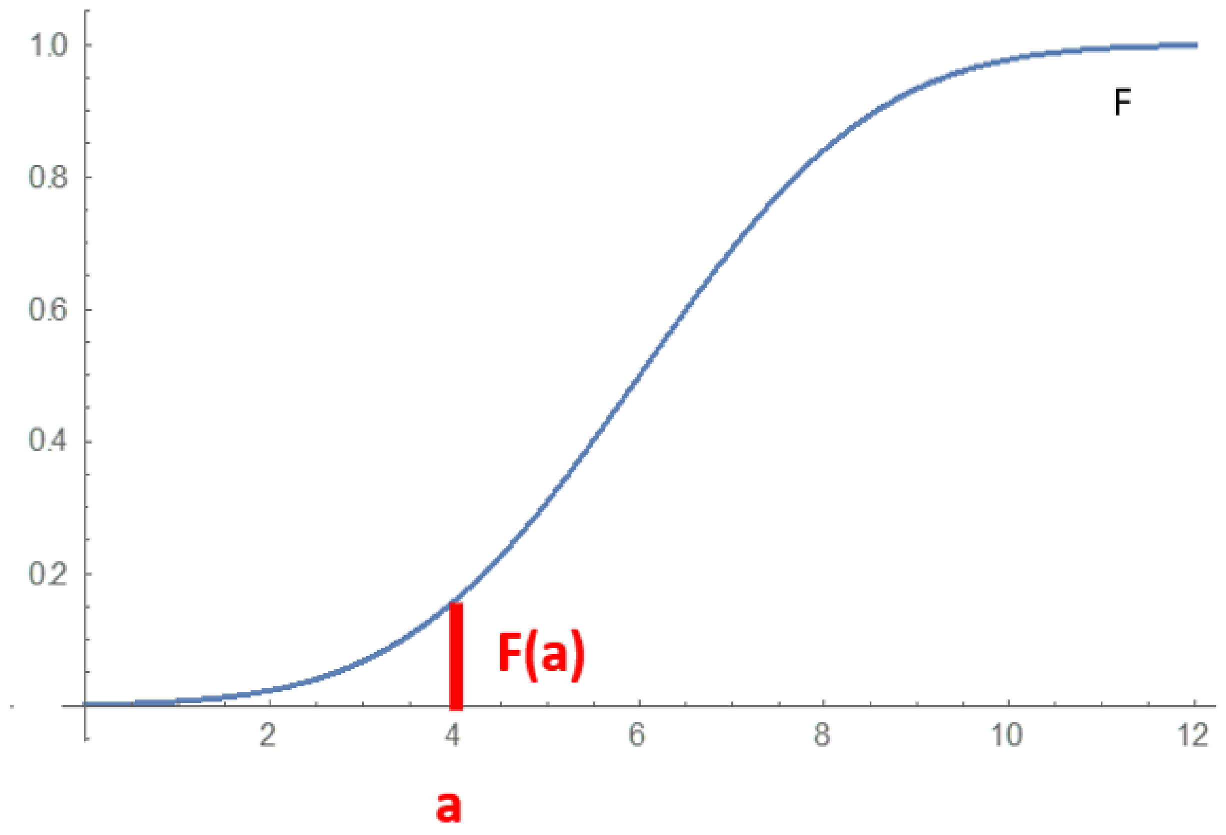

cdf

Credit: Wikipedia

**Continuous random variables**




pdf




cdf

Credit: Wikipedia

# Discrete vs. Continuous RVs

**Discrete random variables**

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

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

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

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

**Continuous random variables**

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

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

$f_X(x)$ 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:** $X \sim \mathrm{Bernoulli}(p)$

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

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:

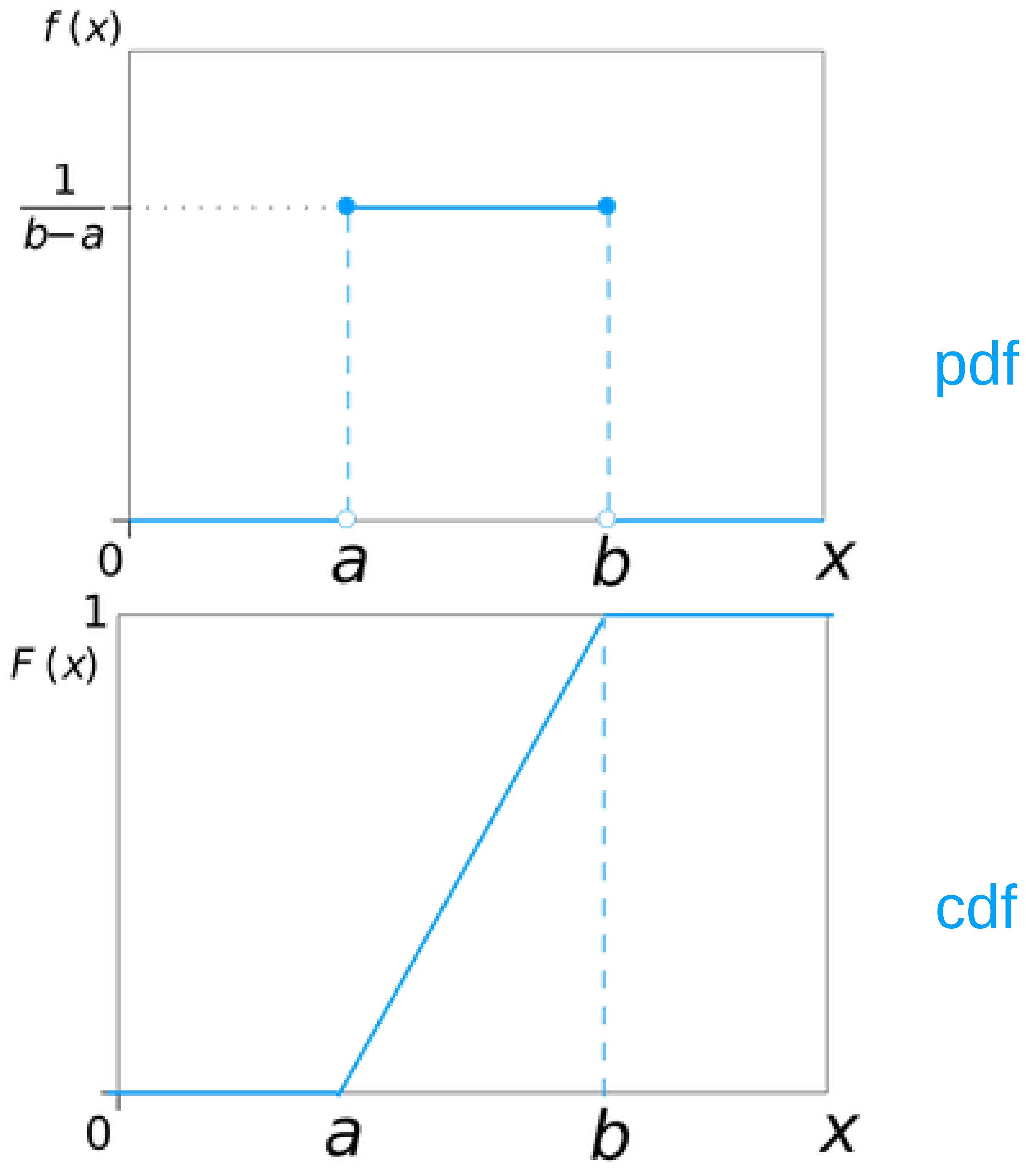
$$p_X(x) = \Pr(X = x) = \prod_{i=1}^{k} p_i^{[x=i]}, \text{where } [x = i] = \begin{cases} 1 & x = i \\ 0 & x \neq i \end{cases}$$

# Common Continuous Distributions

**Uniform distribution:** $\boldsymbol{X\sim\mathrm{Uniform}(a,b)}$

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

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




Credit: Wikipedia

# Common Continuous Distributions

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

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

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

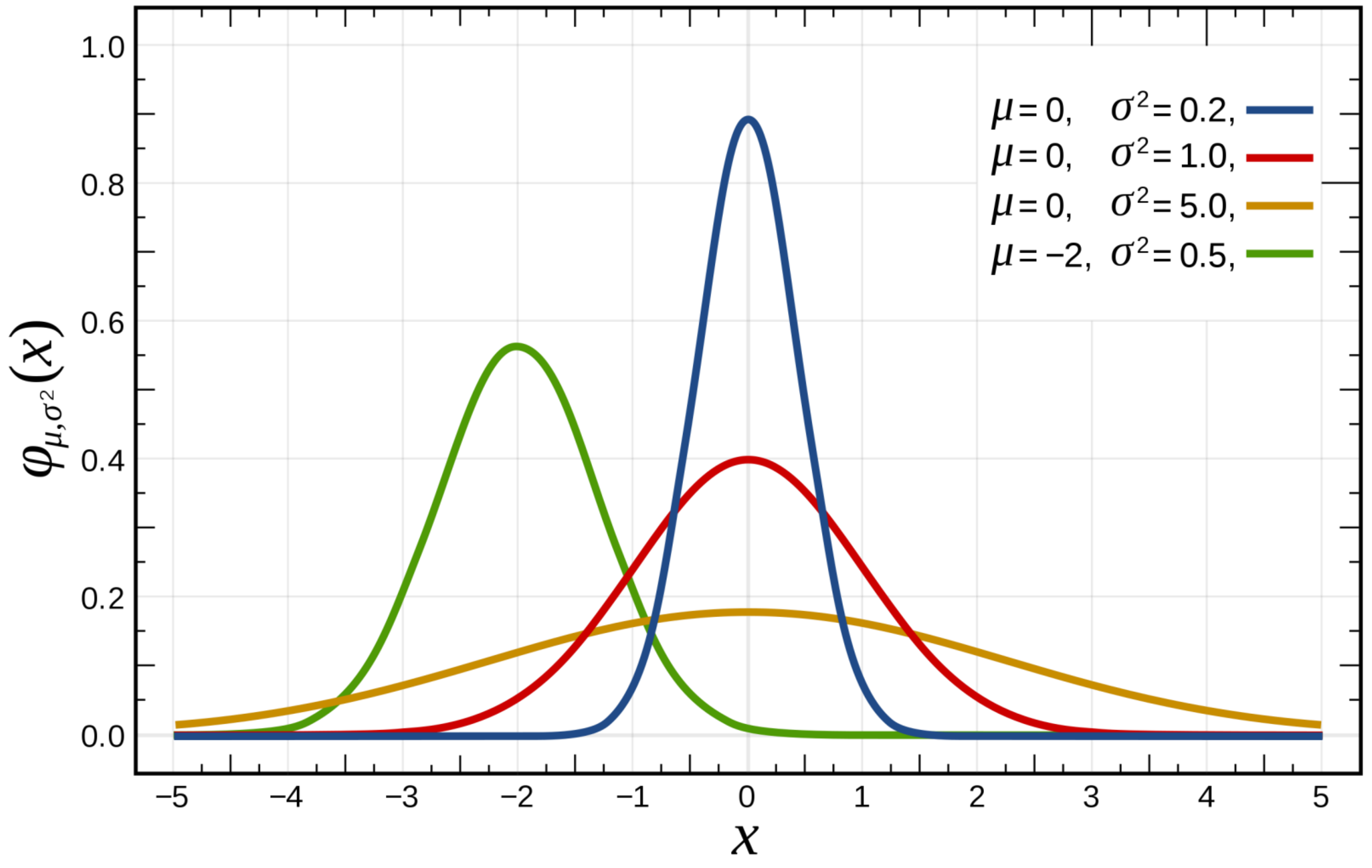
$$\mathrm{supp}(X) = \mathbb{R}$$

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

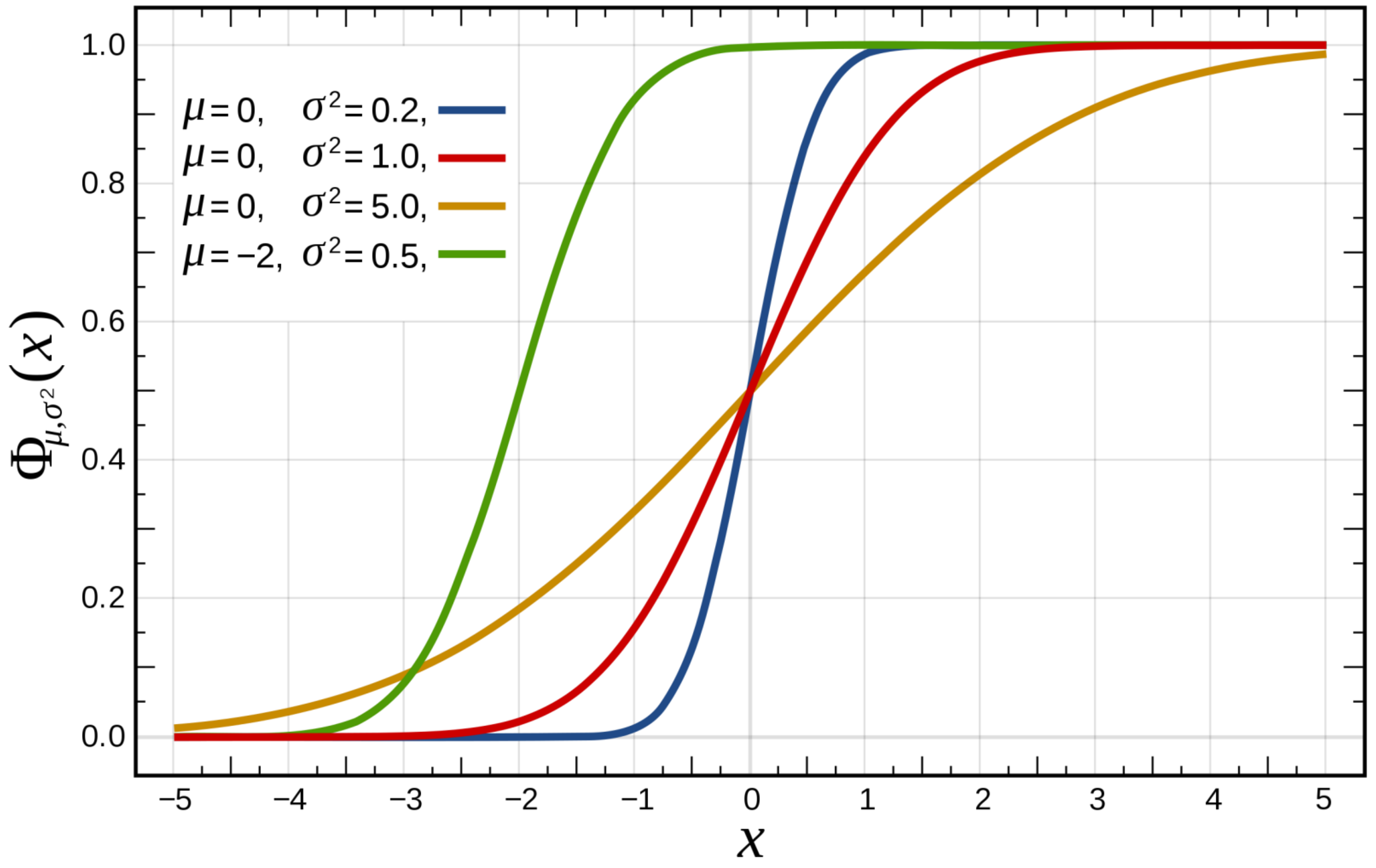
$Z + \mu \sim \mathcal{N}(\mu, 1)$ and $\sigma Z \sim \mathcal{N}(0, \sigma^2)$

$$\Longrightarrow \mu + \sigma Z \sim \mathcal{N}(\mu, \sigma^2)$$

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




pdf




cdf

Credit: Wikipedia

# Multiple Random Variables

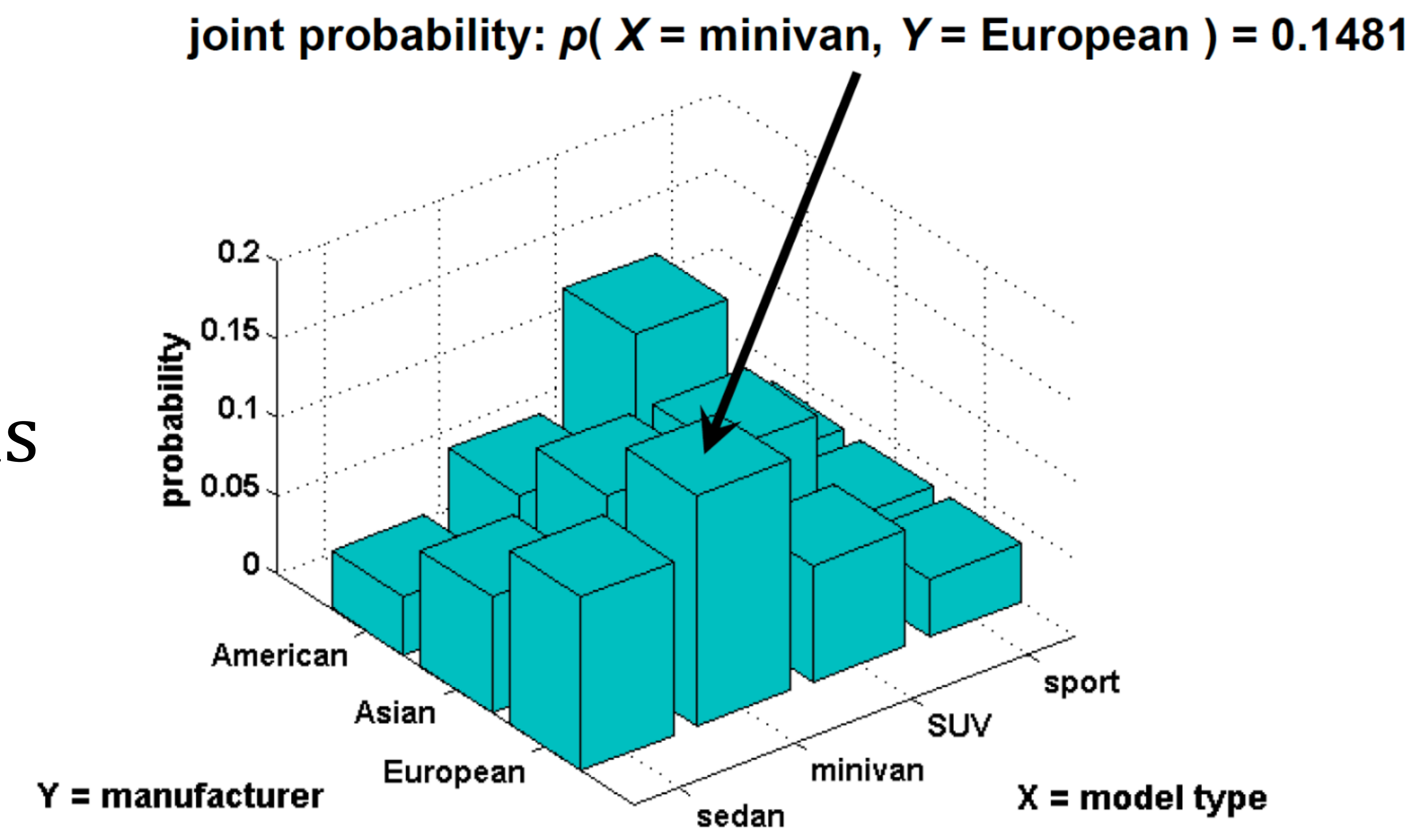
- What if we have multiple random variables, which may depend on one another? How do we represent the dependence between them?
    - E.g.: Tomorrow's temperature and snowfall
- Going forward, will use slightly different notation:
    - Will use capital letters, e.g.: $X$, to denote RVs and corresponding 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:

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

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




Credit: Jeff Howbert

In general:

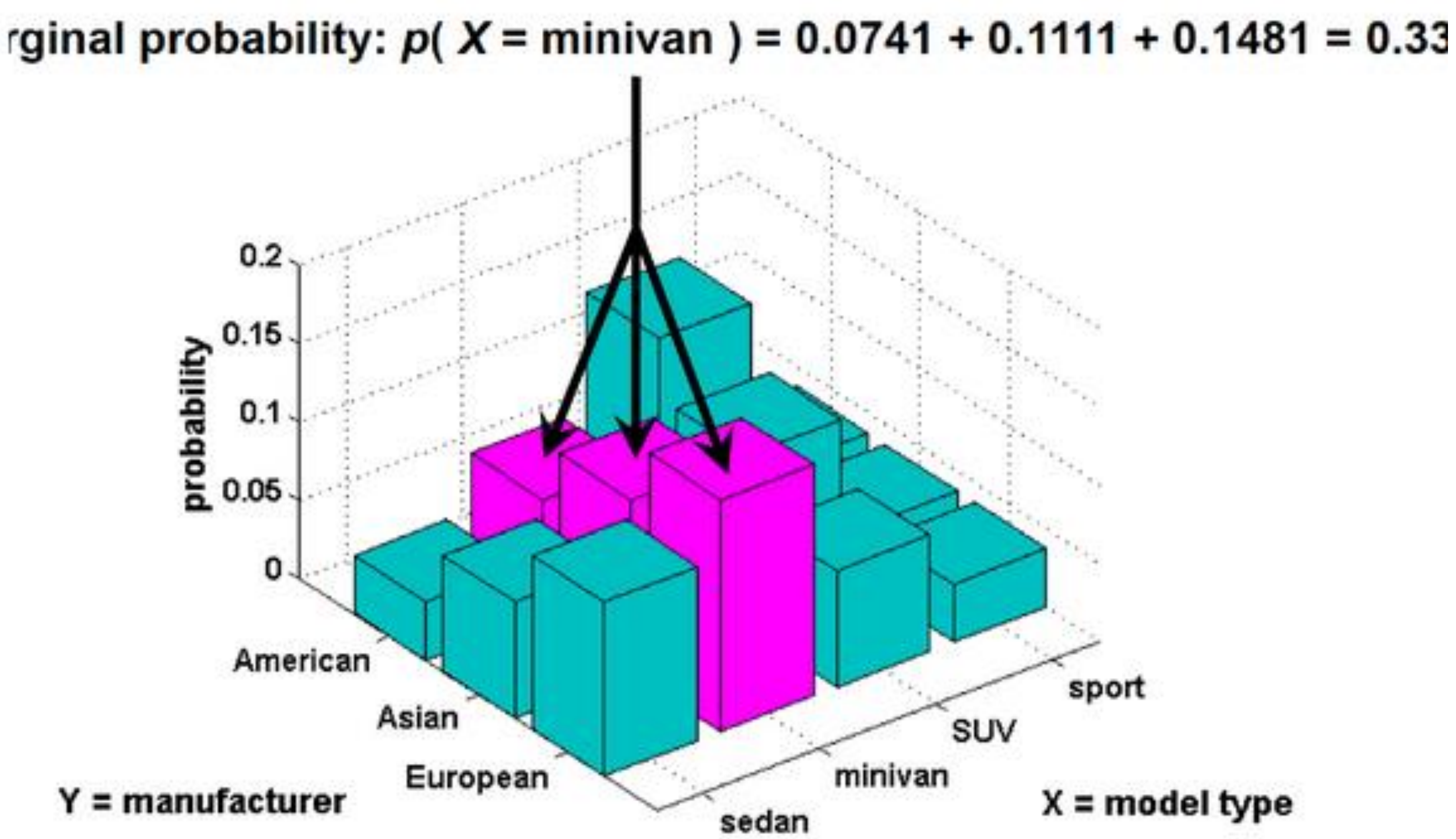
$$F(x_1, \ldots, x_n) = \Pr(X_1 \leq x_1 \text{ and} \cdots \text{and } X_n \leq x_n)$$

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

# Marginal Probability Distributions

Two random variables:

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




Credit: Jeff Howbert

In general:

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

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

# Conditional Probability Distributions

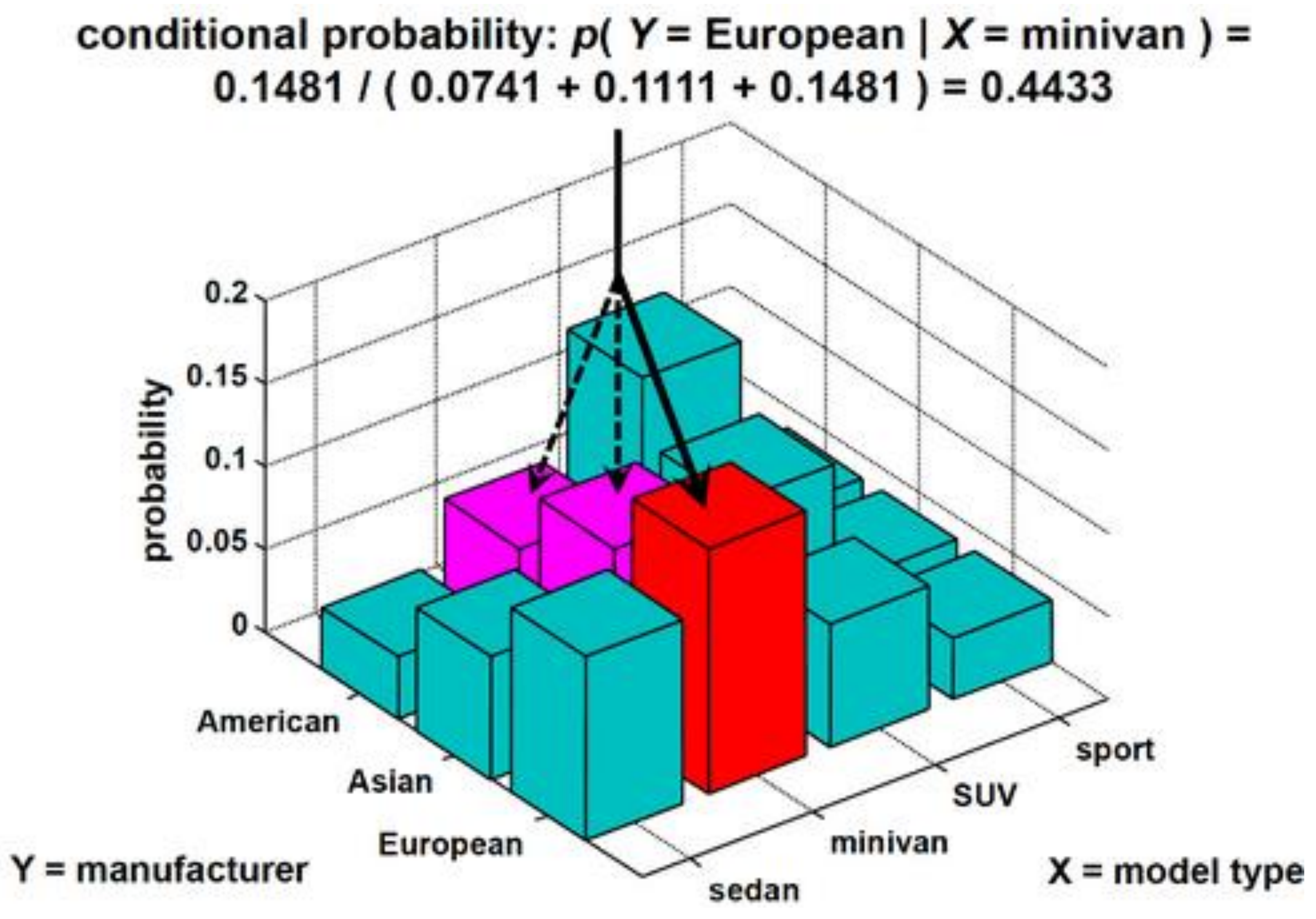
Two random variables:

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

In general:

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

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




Credit: Jeff Howbert

# Chain Rule of Probability

Two random variables:

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

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:

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

In general:

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

# Independence

Two random variables $X$ and $Y$ are independent if:

$p(y|x) = p(y)\ \forall x$ (or equivalently, $p(x|y) = p(x)\ \forall y$)

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

Random variables $X_1, \ldots, X_n$ are (mutually) independent if:
$p(x_1, \ldots, x_n) = p(x_1) \cdots p(x_n)$

# Conditional Independence

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

$p(y|x,z) = p(y|z) \forall x$ (or equivalently, $p(x|y,z) = p(x|z) \forall y$)

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 $X_1, \ldots, X_n$ are conditionally independent given $Z_1 = z_1, \ldots, Z_l = z_l$ if:
$p(x_1, \ldots, x_n | z_1, \ldots, z_l) = p(x_1|z_1, \ldots, z_l) \cdots p(x_n|z_1, \ldots, z_l)$

# 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:

$$p(y|x) = \frac{p(y)p(x|y)}{p(x)} = \frac{p(y)p(x|y)}{\int_{-\infty}^{\infty} p(x,y)dy} = \frac{p(y)p(x|y)}{\int_{-\infty}^{\infty} p(y)p(x|y)dy}$$ (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:

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

In general:

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

(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: $E[X]$

Covariance: $\mathrm{Cov}(X,Y) = E[(X - E[X])(Y - E[Y])]$
$= E[XY] - E[X]E[Y]$

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

Variance: $\mathrm{Var}(X) := \mathrm{Cov}(X,X) = E[(X - E[X])^2]$
$= E[X^2] - (E[X])^2$

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

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

# Zero Covariance vs. Independence

If $X$ and $Y$ are independent,

$$\mathrm{Cov}(X,Y) = E[XY] - E[X]E[Y]$$

$$= E[X]E[Y] - E[X]E[Y]$$
$$= 0$$

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

# Conditional Expectation

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

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

In general:

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

# Conditional Moments

Conditioning on one variable:

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

In general:

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

$$\mathrm{Var}(X|Z_1 = z_1, \dots, Z_l = z_l)$$
$$= E[(X - E[X|Z_1 = z_1, \dots, Z_l = z_l])^2|Z_1 = z_1, \dots, Z_l = z_l]$$

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:

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

(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 $e$.

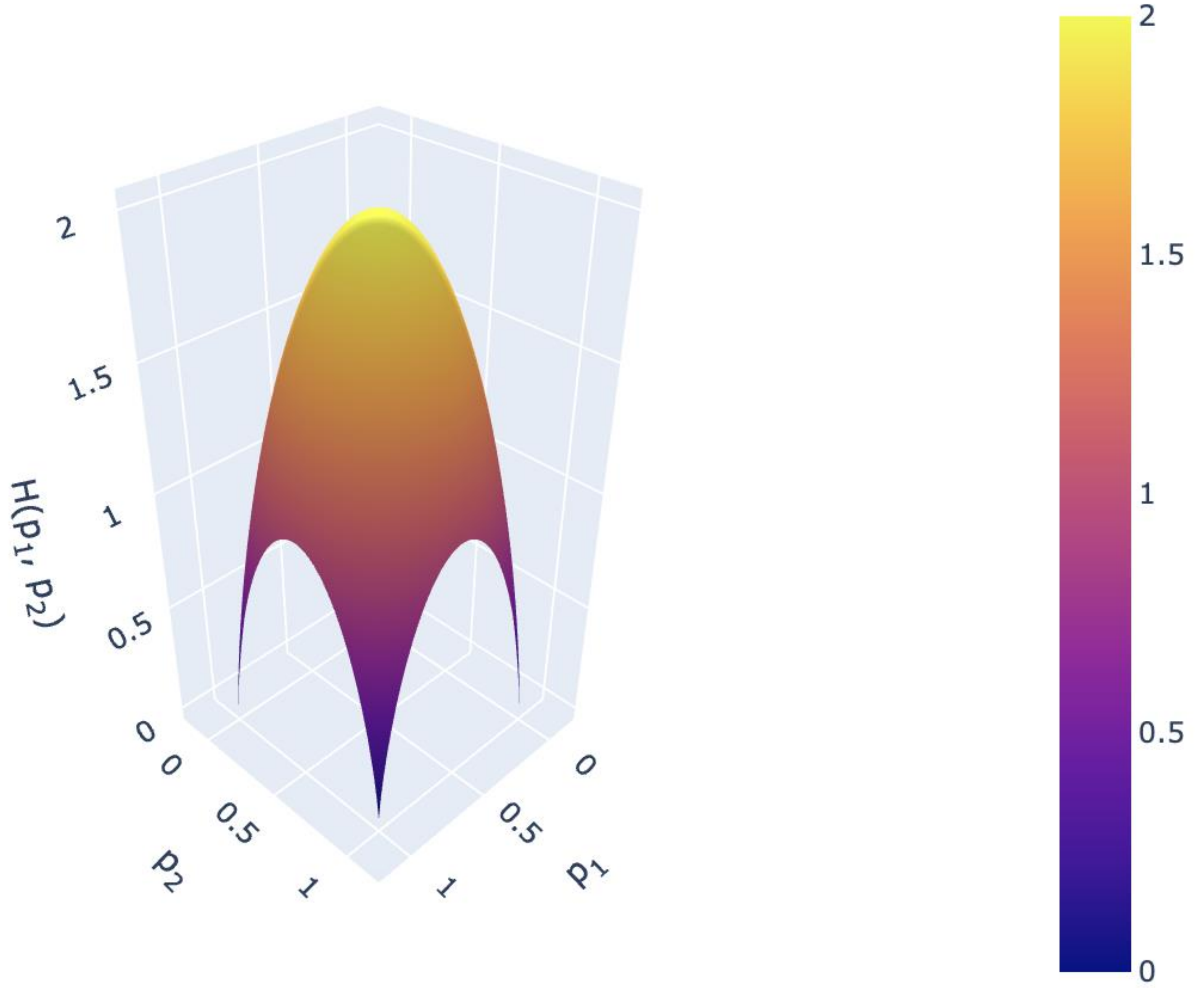
# Entropy

**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:

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

In general:

$$H(X_1,\dots,X_n) = E[-\log_b p(X_1,\dots,X_n)]$$
$$= -\sum_{x_1\in\Omega_{X_1}}\cdots\sum_{x_n\in\Omega_{X_n}} p(x_1,\dots,x_n)\log_b\big(p(x_1,\dots,x_n)\big)$$

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

# Conditional Entropy

Two **discrete** RVs:

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

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

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

# Conditional Entropy

In general:

$$H(X|Y_1 = y_1, \dots, Y_l = y_l) = E_X[-\log_b p(X|y_1, \dots, y_l)|Y_1 = y_1, \dots, Y_l = y_l]$$
$$= -\sum_{x \in \Omega_X} p(x|y_1, \dots, y_l) \log_b\big(p(x|y_1, \dots, y_l)\big)$$

$$H(X|Y_1, \dots, Y_l) = E_{Y_1,\dots,Y_l}[E_X[-\log_b p(X|Y_1, \dots, Y_l)|Y_1, \dots, Y_l]]$$
$$= E_{X,Y_1,\dots,Y_l}[-\log_b p(X|Y_1, \dots, Y_l)]$$
$$= -\sum_{x \in \Omega_X} \sum_{y_1 \in \Omega_{Y_1}} \cdots \sum_{y_l \in \Omega_{Y_l}} p(x, y_1, \dots, y_l) \log_b\big(p(x|y_1, \dots, y_l)\big)$$
$$= -\sum_{x \in \Omega_X} \sum_{y_1 \in \Omega_{Y_1}} \cdots \sum_{y_l \in \Omega_{Y_l}} \Big(p(x, y_1, \dots, y_l) \log_b\big(p(x, y_1, \dots, y_l)\big) - p(x, y_1, \dots, y_l) \log_b\big(p(y_1, \dots, y_l)\big)\Big)$$

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

# Mutual Information

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

(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 $X_1, \ldots, X_n$ 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

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

"Mean vector" or just the "mean"

The covariances and variances can be arranged into a matrix:

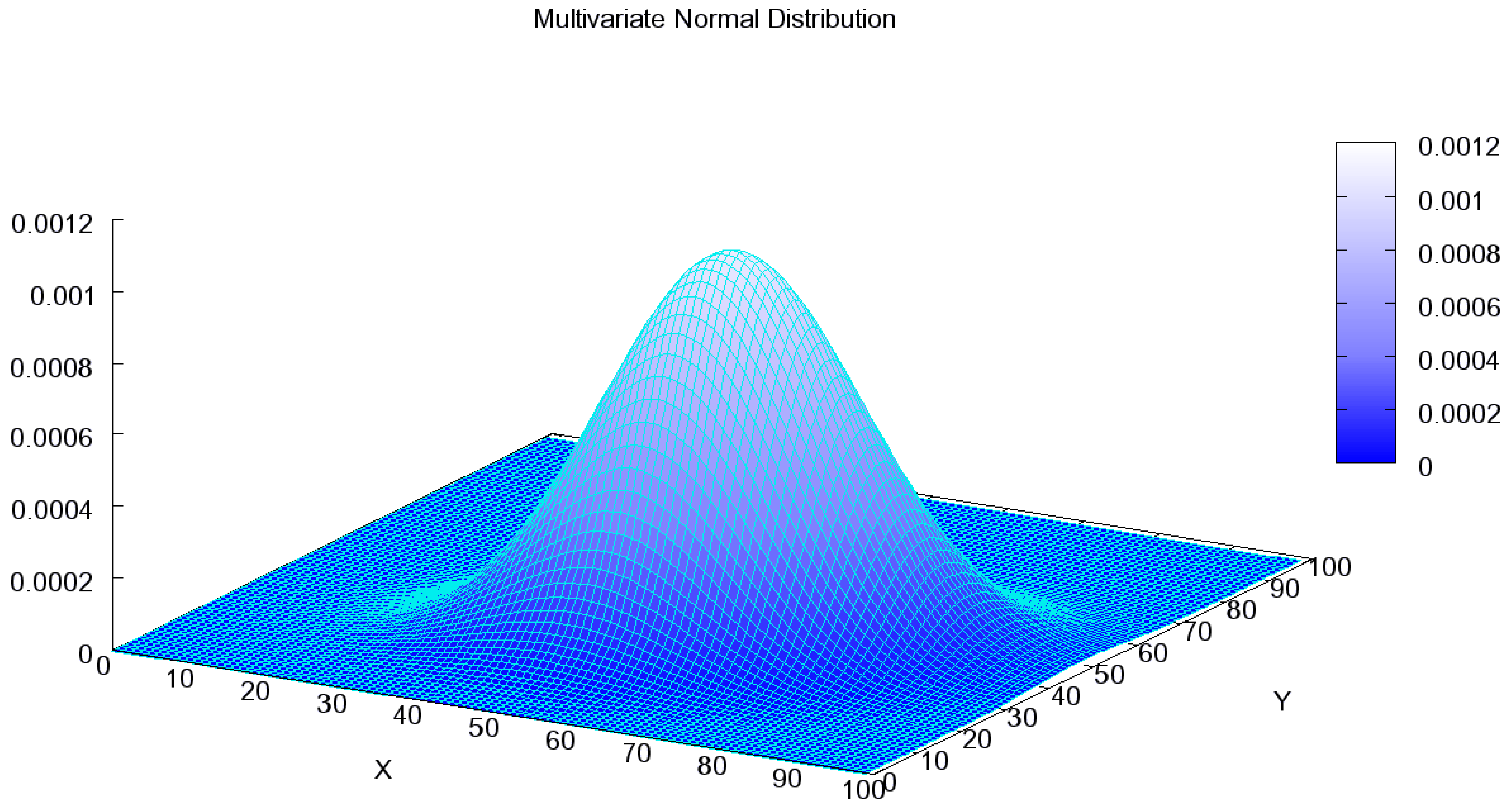
$$E\left[(\vec{X} - E[\vec{X}])(\vec{X} - E[\vec{X}])^\top\right] = E[\vec{X}\vec{X}^\top] - (E[\vec{X}])(E[\vec{X}])^\top$$

"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:

$$X \sim \mathcal{N}(\mu, \sigma)$$

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

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 $\Sigma$ 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

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

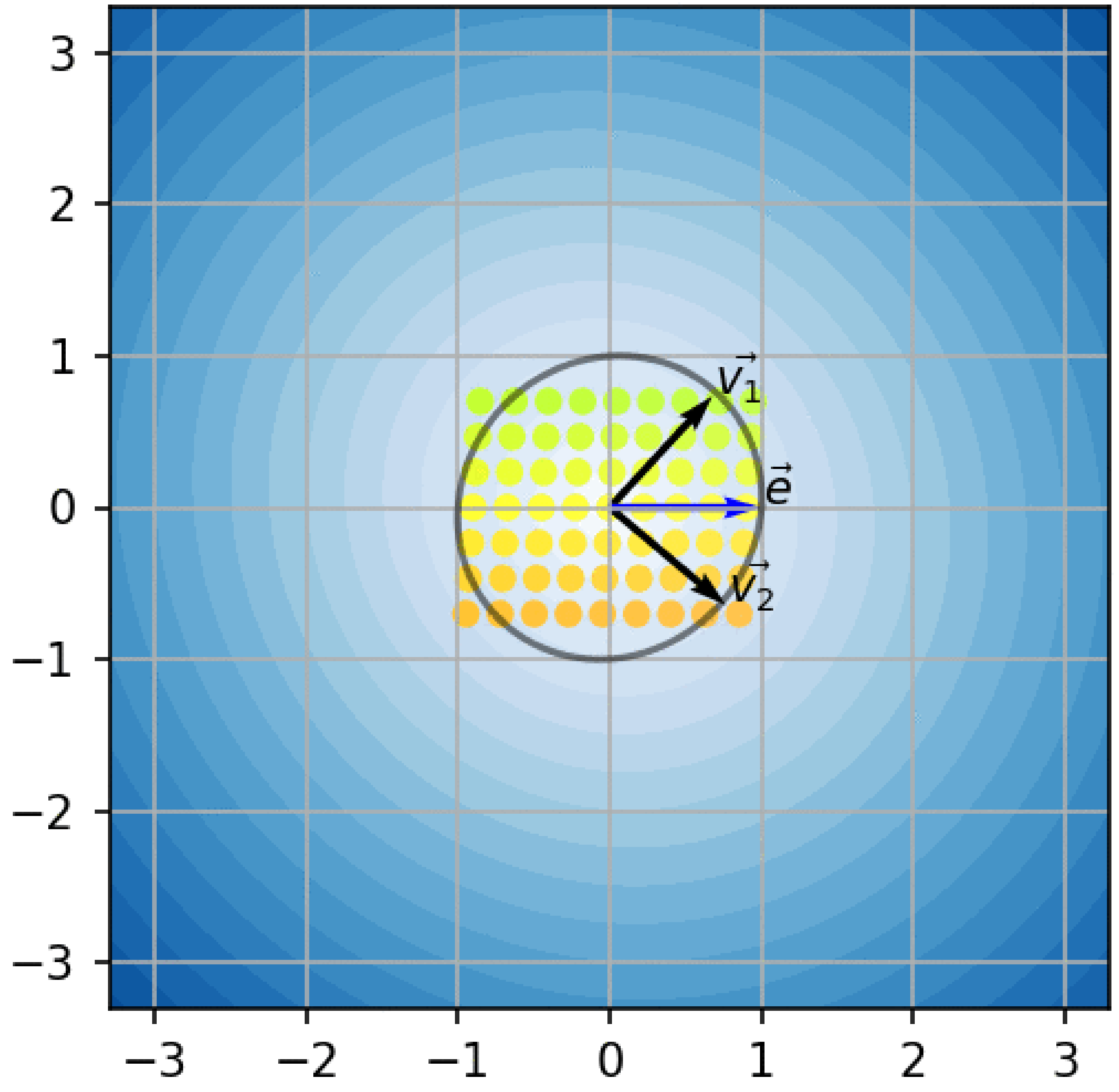
How does the quadratic form $(\vec{x}-\vec{\mu})^\top \Sigma^{-1}(\vec{x}-\vec{\mu})$ behave?

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 $A$.

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

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

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




$\vec{v_1}$ - right-singular vector

$\vec{v_2}$ - second right-singular vector

$\vec{e}$ - eigenvector

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

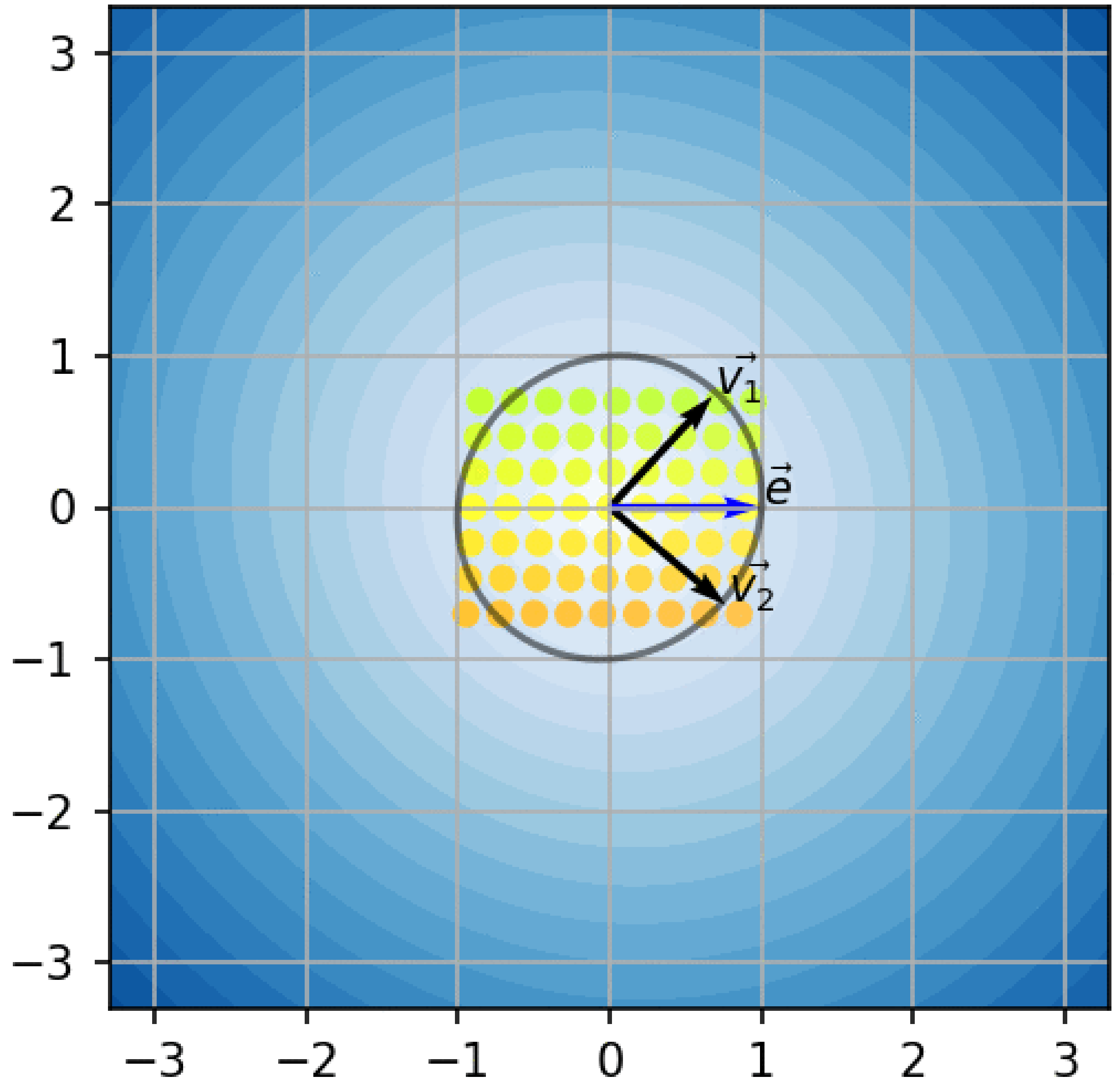
# Multivariate Normal Distribution

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

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.




$\vec{v}_1$ - right-singular vector

$\vec{v}_2$ - second right-singular vector

$\vec{e}$ - eigenvector

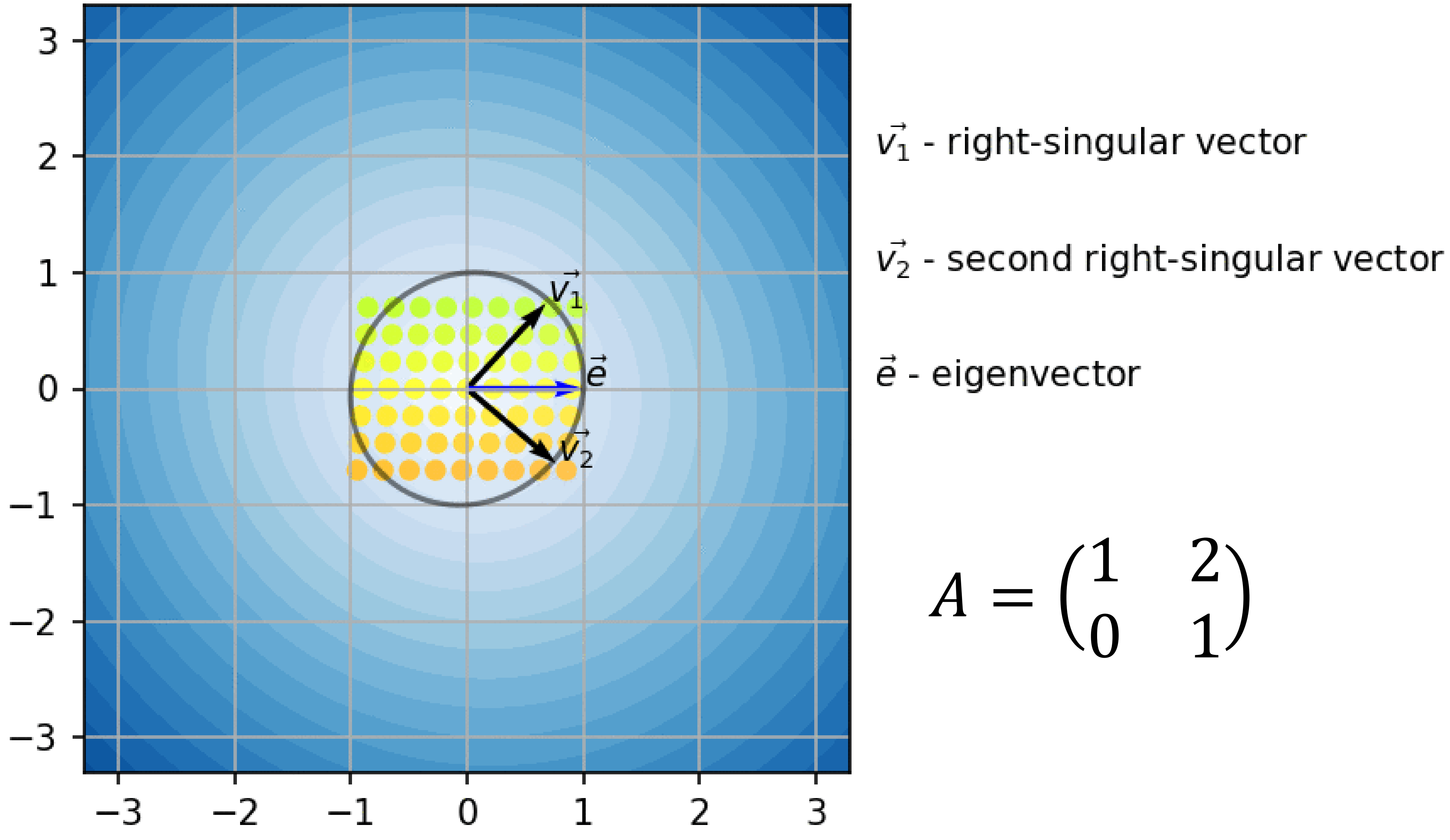
$$A = \begin{pmatrix} 1 & 2 \\ 0 & 1 \end{pmatrix}$$

# 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?

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



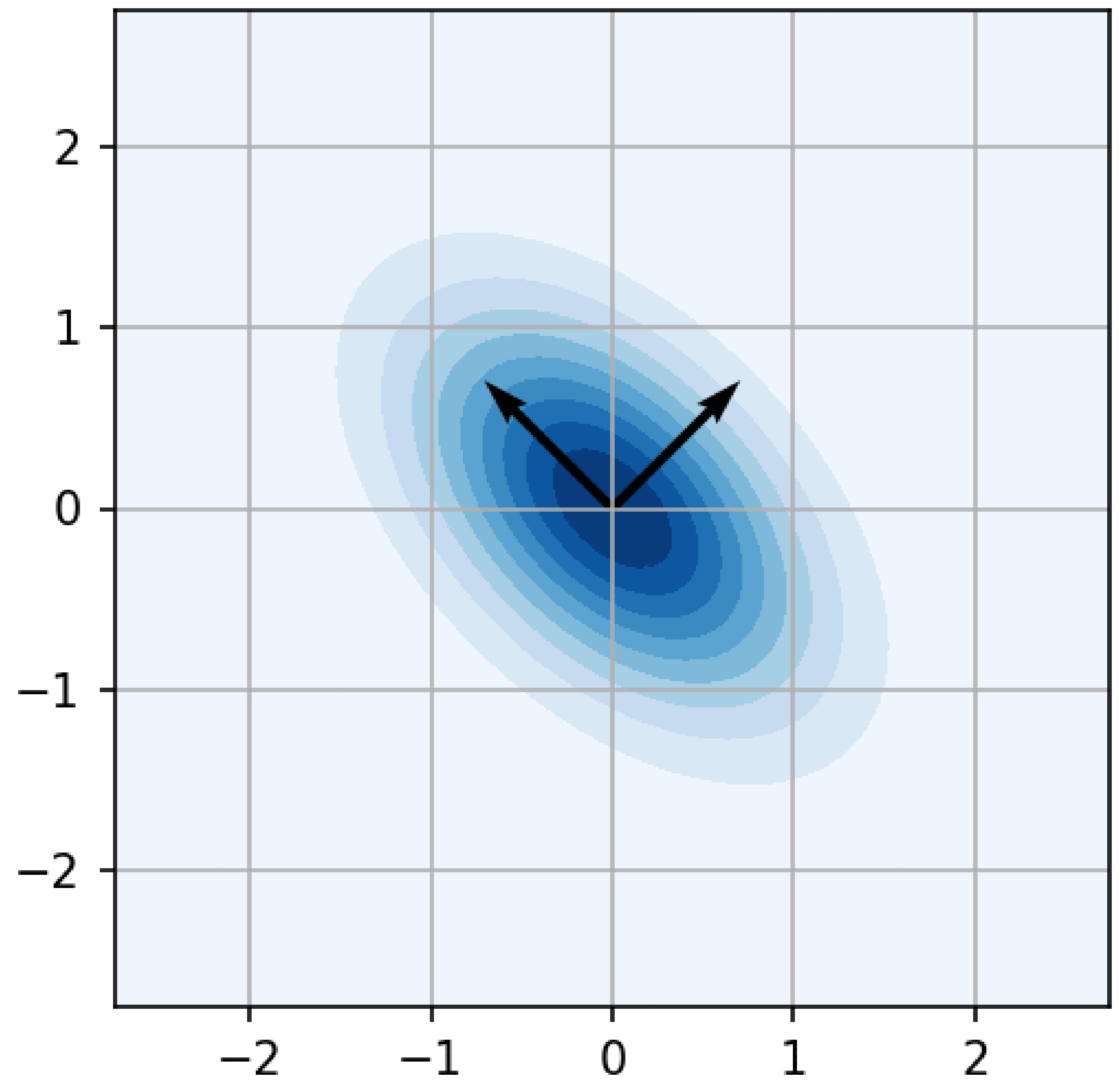

$$A = \begin{pmatrix} 1 & 2 \\ 0 & 1 \end{pmatrix}$$

# Multivariate Normal Distribution

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

The black arrows denote the eigenvectors of $\Sigma$.

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



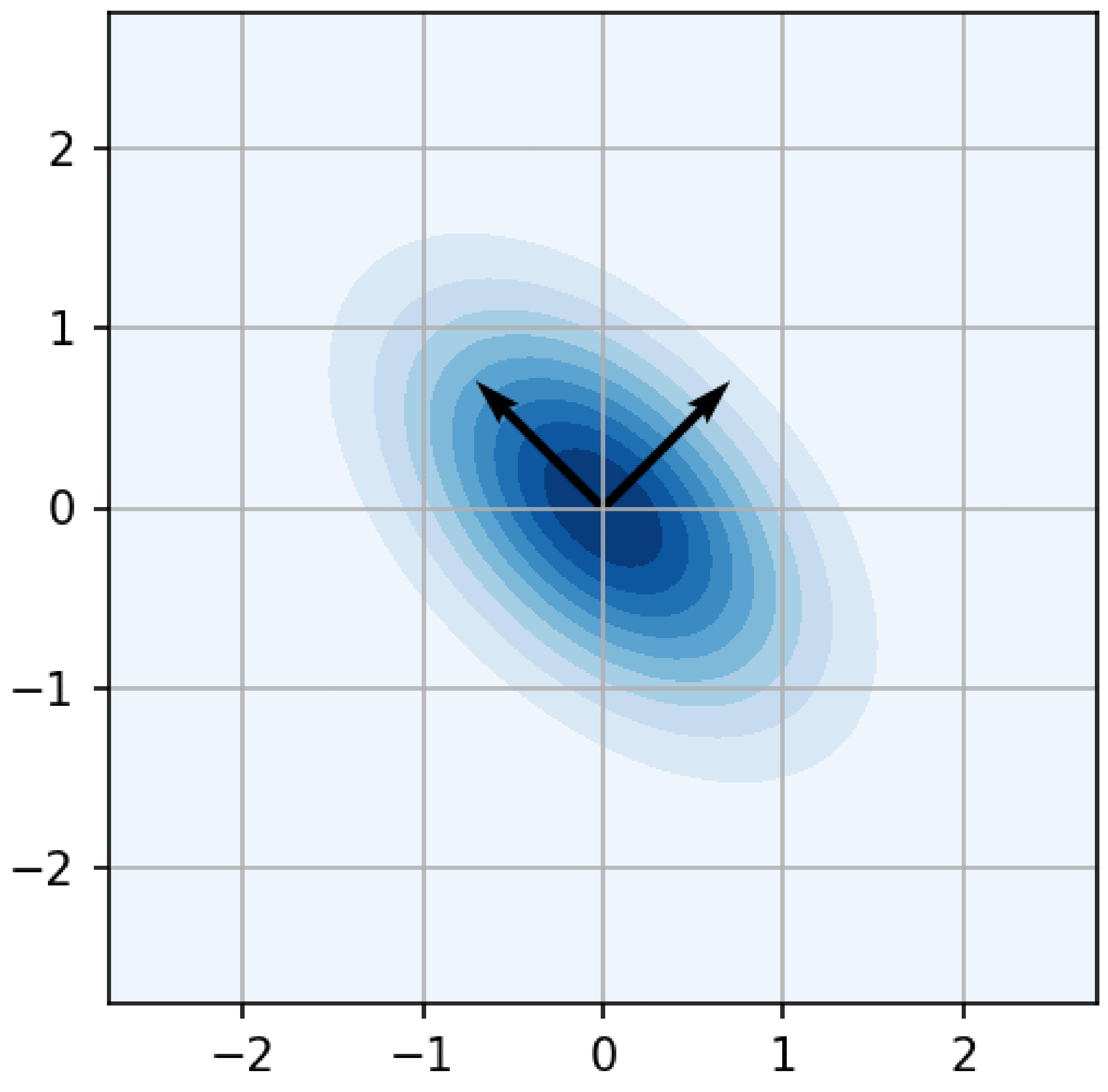

| | |
|---|---|
| 1.0 | -0.5 |
| -0.5 | 1.0 |

# 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.

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



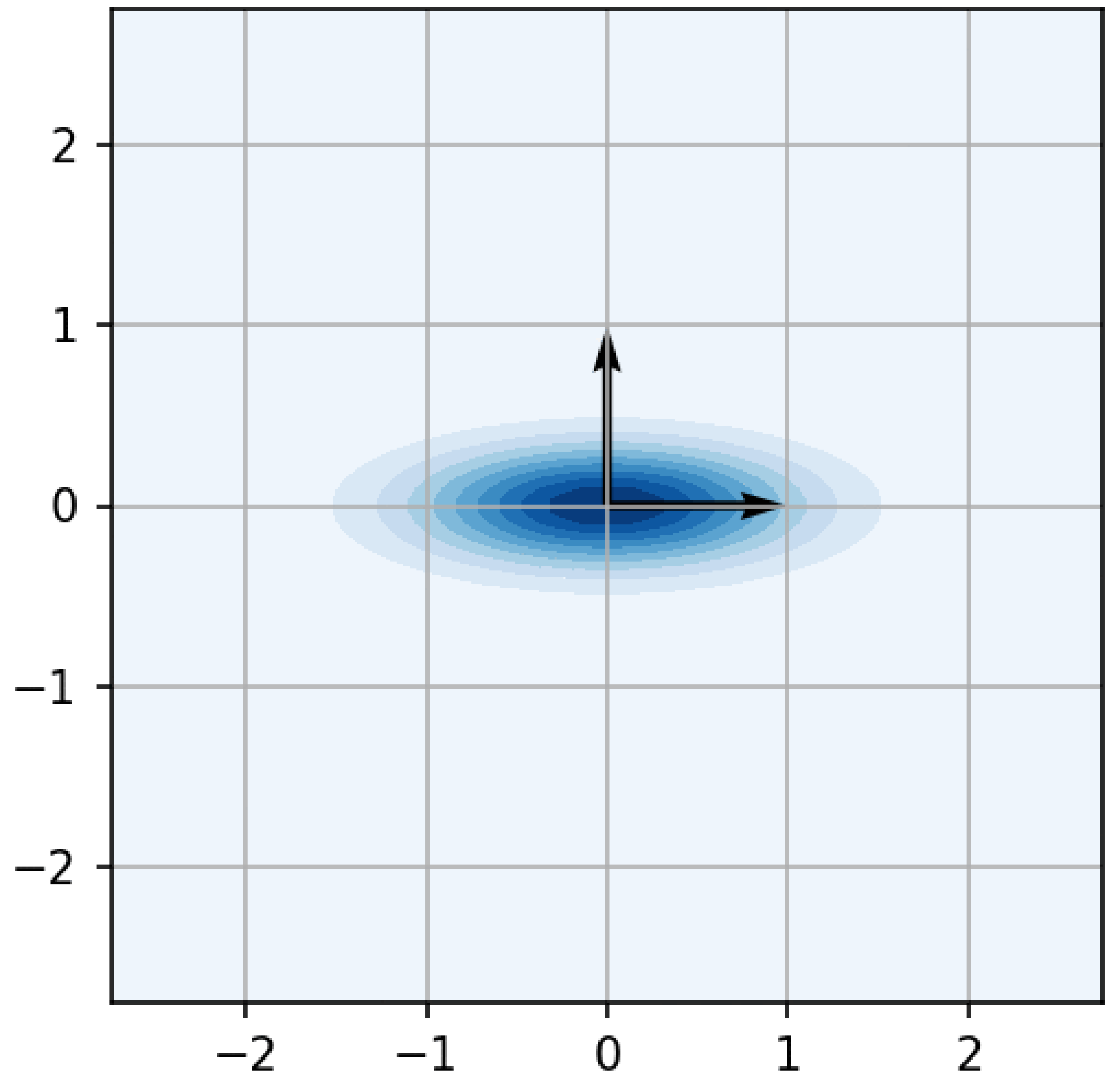

| | |
|---|---|
| 1.0 | -0.5 |
| -0.5 | 1.0 |

# 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.

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




| | |
|---|---|
| 1.0 | 0.0 |
| 0.0 | 0.1 |

# 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: $\vec{z} \sim \mathcal{N}(0, I)$

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

"Isotropic Gaussian"

(Variance along every direction is the same)

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

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

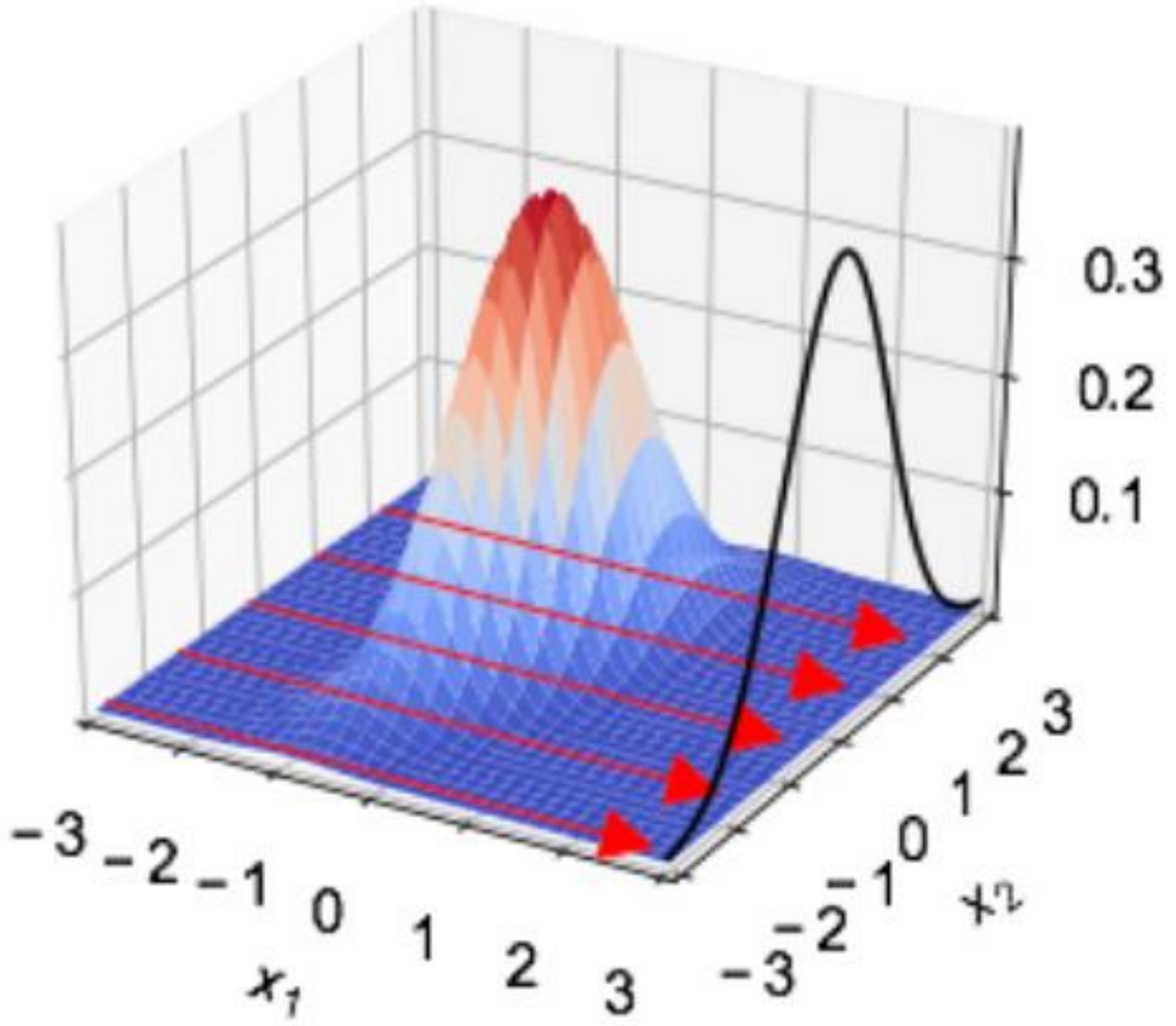
# Marginalization of Multivariate Normals

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 $\vec{x}$

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

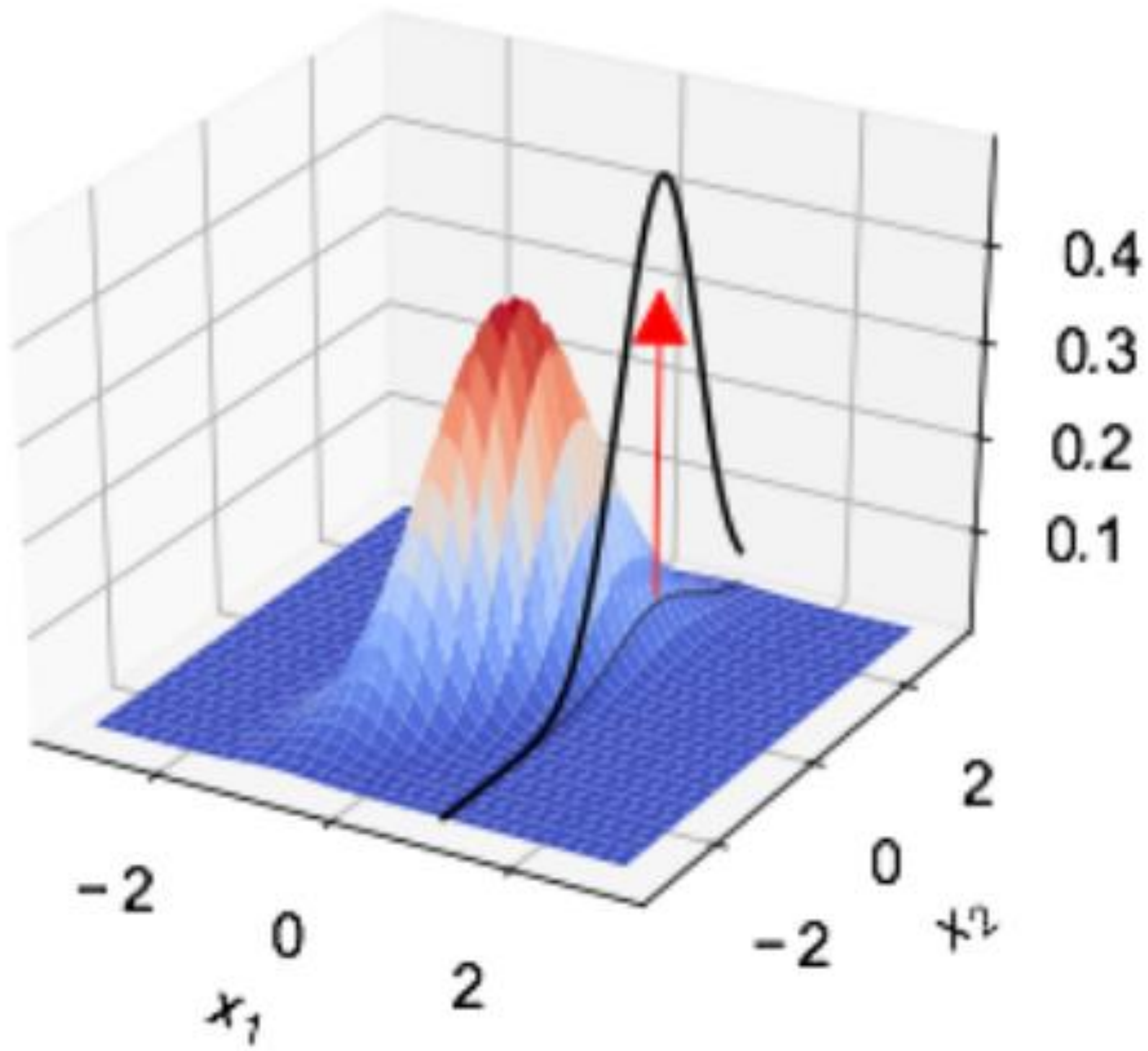
# Conditioning of Multivariate Normals

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 $\vec{x}$

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

# Quiz Practice

Which one of the following is not necessarily a Gaussian random variable?

(A) $X|Y = 1$, where $\begin{pmatrix} X \\ Y \end{pmatrix} \sim \mathcal{N}(\vec{0}, I)$

(B) $Y|X = 100$, where $\begin{pmatrix} X \\ Y \end{pmatrix} \sim \mathcal{N}(\vec{0}, I)$

(C) $\frac{1}{2}X - \frac{1}{3}Y$ ,where $\begin{pmatrix} X \\ Y \end{pmatrix} \sim \mathcal{N}(\vec{0}, I)$

(D) $-10X + 5$, where $\begin{pmatrix} X \\ Y \end{pmatrix} \sim \mathcal{N}(\vec{0}, I)$

(E) $Y - 10X$, where $X \sim \mathcal{N}(0,1)$, $Y \sim \mathcal{N}(0,1)$ and $X \perp Y$

(F) $X - Y$, where $X \sim \mathcal{N}(0,1)$, $Y \sim \mathcal{N}(0,1)$ and $\text{Cov}(X, Y) = -1$

(G) $X - Y$, where $X \sim \mathcal{N}(0,1)$, $Y \sim \mathcal{N}(0,1)$ and $\text{Cov}(X, Y) = -0.5$

(H) $-10X + 5$, where $X \sim \mathcal{N}(0,1)$, $Y \sim \mathcal{N}(0,1)$ and $\text{Cov}(X, Y) = -0.5$

(I) All are Gaussian random variables