Recap: Matrix Calculus & Basic Probability

EECE454 Intro. to Machine Learning Systems

Fall 2024

Last class

- Vectors & Matrices
- Multiplications
- Norms, Column / Row / Null space
- Eigendecomposition & SVD
- **Today.**
    - Gram-Schmidt
    - Matrix Calculus
    - Probability

Gram-Schmidt
(QR decomposition)

QR decomposition

- **Last class.** We reviewed SVD—a neat method to decompose any $\mathbf{A} \in \mathbb{R}^{m \times n}$ into $\mathbf{A} = \mathbf{U}\Sigma\mathbf{V}^\top$

QR decomposition

- **Last class.** We reviewed SVD—a neat method to decompose any $\mathbf{A} \in \mathbb{R}^{m \times n}$ into $\mathbf{A} = \mathbf{U}\boldsymbol{\Sigma}\mathbf{V}^\top$

- **Today.** A more compact decomposition, when $m \geq n$

$$\mathbf{A} = \mathbf{QR}$$

  - $\mathbf{Q} \in \mathbb{R}^{m \times m}$ is a unitary matrix (i.e., $\mathbf{Q}^\top = \mathbf{Q}^{-1}$)

  - $\mathbf{R} \in \mathbb{R}^{m \times n}$ is an upper triangular matrix

$$\mathbf{A} = \begin{bmatrix} | & \cdots & | \\ \mathbf{e}_1 & \cdots & \mathbf{e}_m \\ | & \cdots & | \end{bmatrix} \begin{bmatrix} r_{11} & r_{12} & \cdots & r_{1n} \\ 0 & r_{22} & \cdots & r_{2n} \\ & & \cdots & \\ 0 & 0 & \cdots & 0 \end{bmatrix}$$

Idea

$$\mathbf{A} = \begin{bmatrix} | & \cdots & | \\ \mathbf{e}_1 & \cdots & \mathbf{e}_m \\ | & \cdots & | \end{bmatrix} \begin{bmatrix} r_{11} & r_{12} & \cdots & r_{1n} \\ 0 & r_{22} & \cdots & r_{2n} \\ & & \cdots & \\ 0 & 0 & \cdots & 0 \end{bmatrix}$$

- Take a look at each column of $\mathbf{A}$:

$$\mathbf{a}_1 = \begin{bmatrix} | & \cdots & | \\ \mathbf{e}_1 & \cdots & \mathbf{e}_m \\ | & \cdots & | \end{bmatrix} \begin{bmatrix} r_{11} \\ 0 \\ 0 \\ \cdots \end{bmatrix}, \quad \mathbf{a}_2 = \begin{bmatrix} | & \cdots & | \\ \mathbf{e}_1 & \cdots & \mathbf{e}_m \\ | & \cdots & | \end{bmatrix} \begin{bmatrix} r_{12} \\ r_{22} \\ 0 \\ \cdots \end{bmatrix}, \quad \cdots$$

Idea

$$\mathbf{A} = \begin{bmatrix} | & \cdots & | \\ \mathbf{e}_1 & \cdots & \mathbf{e}_m \\ | & \cdots & | \end{bmatrix} \begin{bmatrix} r_{11} & r_{12} & \cdots & r_{1n} \\ 0 & r_{22} & \cdots & r_{2n} \\ & & \cdots & \\ 0 & 0 & \cdots & 0 \end{bmatrix}$$

- Take a look at each column of $\mathbf{A}$:

$$\mathbf{a}_1 = \begin{bmatrix} | & \cdots & | \\ \mathbf{e}_1 & \cdots & \mathbf{e}_m \\ | & \cdots & | \end{bmatrix} \begin{bmatrix} r_{11} \\ 0 \\ 0 \\ \cdots \end{bmatrix}, \quad \mathbf{a}_2 = \begin{bmatrix} | & \cdots & | \\ \mathbf{e}_1 & \cdots & \mathbf{e}_m \\ | & \cdots & | \end{bmatrix} \begin{bmatrix} r_{12} \\ r_{22} \\ 0 \\ \cdots \end{bmatrix}, \quad \cdots$$

$\Rightarrow$ $\mathbf{a}_1 = r_{11}\mathbf{e}_1$

$\mathbf{a}_2 = r_{12}\mathbf{e}_1 + r_{22}\mathbf{e}_2$

(...)

Procedure

$$\mathbf{a}_1 = r_{11}\mathbf{e}_1, \quad \mathbf{a}_2 = r_{12}\mathbf{e}_1 + r_{22}\mathbf{e}_2, \quad \cdots$$

- Now it is quite easy to see how it works — called Gram-Schmidt process

Procedure

$$\mathbf{a}_1 = r_{11}\mathbf{e}_1, \quad \mathbf{a}_2 = r_{12}\mathbf{e}_1 + r_{22}\mathbf{e}_2, \quad \cdots$$

- Now it is quite easy to see how it works — called Gram-Schmidt process

  - Make $\mathbf{e}_1$ by normalizing $\mathbf{a}_1$

$$\mathbf{e}_1 = \frac{\mathbf{a}_1}{\|\mathbf{a}_1\|_2}, \quad r_{11} = \|\mathbf{a}_1\|_2$$

Procedure

$$\mathbf{a}_1 = r_{11}\mathbf{e}_1, \quad \mathbf{a}_2 = r_{12}\mathbf{e}_1 + r_{22}\mathbf{e}_2, \quad \cdots$$

- Now it is quite easy to see how it works — called Gram-Schmidt process
  - Make $\mathbf{e}_1$ by normalizing $\mathbf{a}_1$

$$\mathbf{e}_1 = \frac{\mathbf{a}_1}{\|\mathbf{a}_1\|_2}, \quad r_{11} = \|\mathbf{a}_1\|_2$$

  - Make $\mathbf{e}_2$ by (1) subtracting the $\mathbf{a}_1$-direction, and (2) normalizing the remainder

$$r_{12} = \mathbf{a}_2^\top \mathbf{e}_1, \quad \mathbf{e}_2 = \frac{\mathbf{a}_2 - r_{12}\mathbf{e}_1}{\|\mathbf{a}_2 - r_{12}\mathbf{e}_1\|_2}, \quad r_{22} = \|\mathbf{a}_2 - r_{12}\mathbf{e}_1\|_2$$

Procedure

$$\mathbf{a}_1 = r_{11}\mathbf{e}_1, \quad \mathbf{a}_2 = r_{12}\mathbf{e}_1 + r_{22}\mathbf{e}_2, \quad \cdots$$

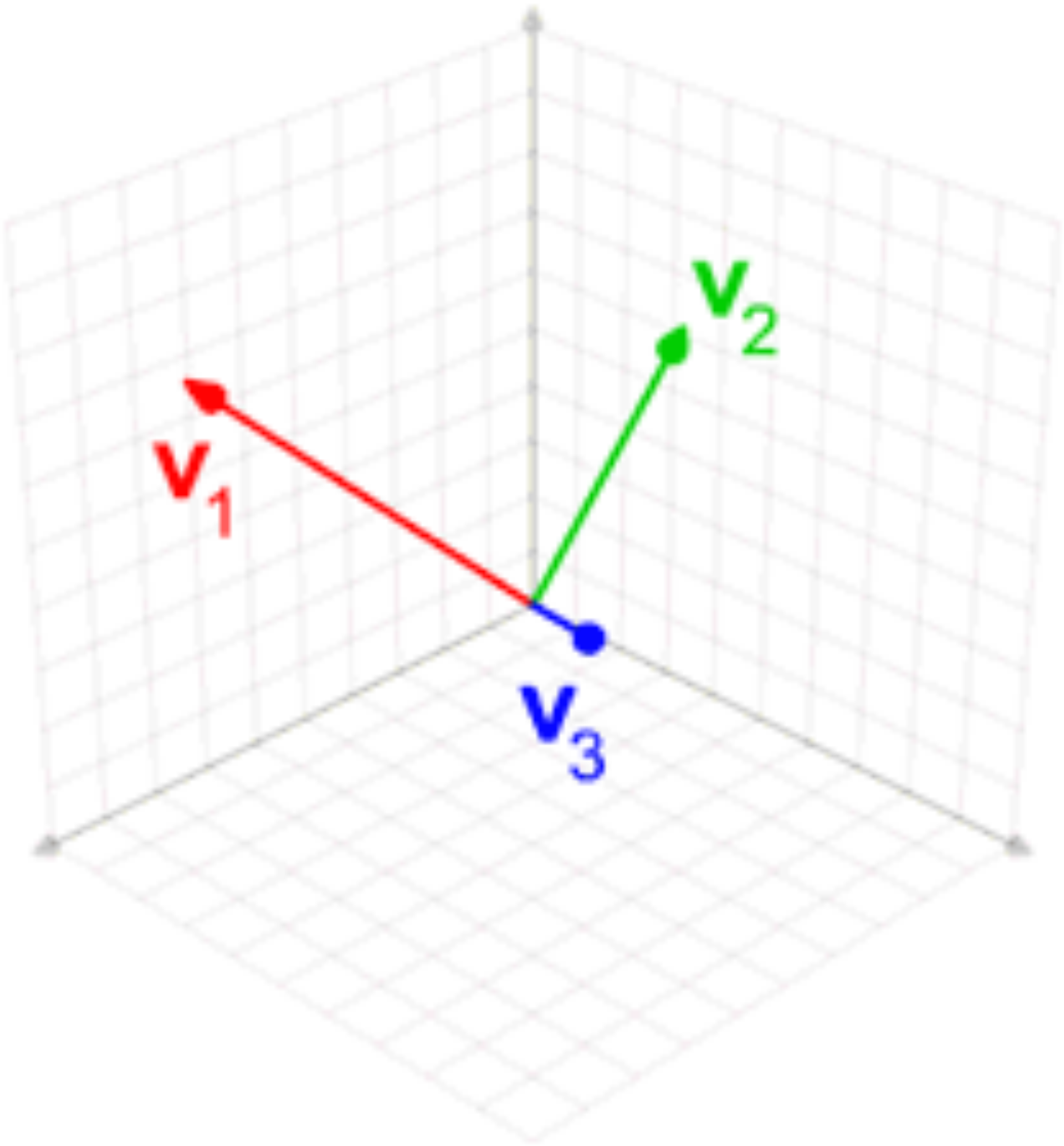
- Now it is quite easy to see how it works — called Gram-Schmidt process
  - Make $\mathbf{e}_1$ by normalizing $\mathbf{a}_1$

$$\mathbf{e}_1 = \frac{\mathbf{a}_1}{\|\mathbf{a}_1\|_2}, \quad r_{11} = \|\mathbf{a}_1\|_2$$

  - Make $\mathbf{e}_2$ by (1) subtracting the $\mathbf{a}_1$-direction, and (2) normalizing the remainder

$$r_{12} = \mathbf{a}_2^\top \mathbf{e}_1, \quad \mathbf{e}_2 = \frac{\mathbf{a}_2 - r_{12}\mathbf{e}_1}{\|\mathbf{a}_2 - r_{12}\mathbf{e}_1\|_2}, \quad r_{22} = \|\mathbf{a}_2 - r_{12}\mathbf{e}_1\|_2$$

  - Repeat!



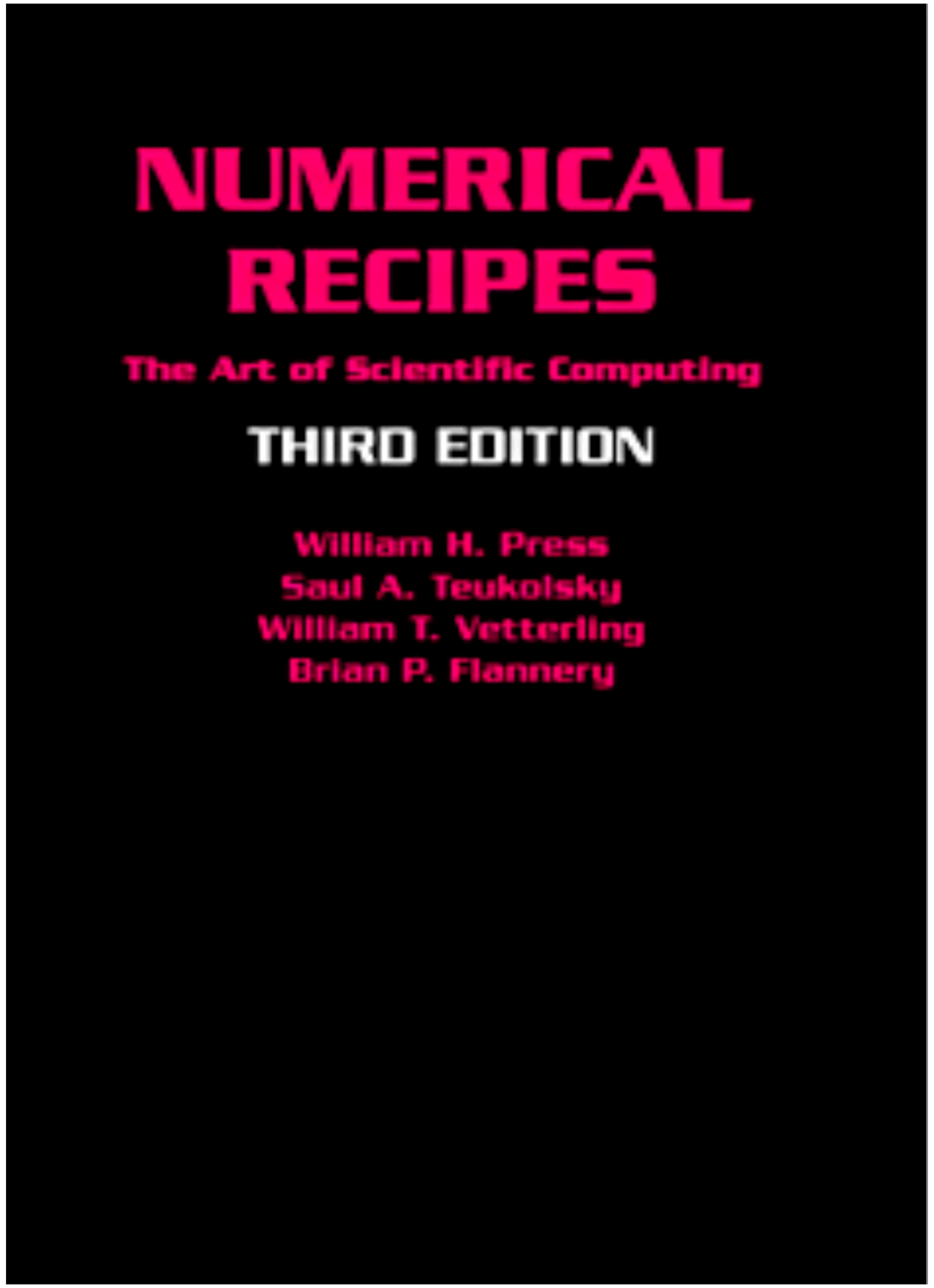
https://commons.wikimedia.org/wiki/File:Gram-Schmidt_orthonormalization_process.gif

Matrix decomposition

- There are plenty of these.
  - SVD, QR, Cholesky, LU, ...
- These tend to have different purposes.
  - People use QR for solving finding $\mathbf{x}$ such that $\mathbf{A}\mathbf{x} = \mathbf{y}$

Matrix decomposition

- There are plenty of these.
  - SVD, QR, Cholesky, LU, ...
- These tend to have different purposes.
  - People use QR for solving finding $\mathbf{x}$ such that $\mathbf{A}\mathbf{x} = \mathbf{y}$
  - Different strengths and weaknesses
    - Numerical stability of the algorithm dramatically differs! (Sec. 2 of “Numerical Recipes” is much recommended)



Matrix Calculus

Why matrix calculus?

- **Univariate calculus.** Finding an optimal scalar $w \in \mathbb{R}$ for a one-dimensional datum.

  - Example. Find a linear function $f(x) = wx$ that minimizes the loss function $\ell(\hat{y}, y) = (\hat{y} - y)^2$.

    Given a single datum (x_0, y_0), then we are solving

$$\min_{w\in\mathbb{R}} \mathscr{L}(w) \quad := \quad \min_{w\in\mathbb{R}} (y_0 - wx_0)^2$$

  - Question. How do we solve?

Why matrix calculus?

- **Univariate calculus.** Finding an optimal scalar $w \in \mathbb{R}$ for a one-dimensional datum.
  - Example. Find a linear function $f(x) = wx$ that minimizes the loss function $\ell(\hat{y}, y) = (\hat{y} - y)^2$.

    Given a single datum (x_0, y_0), then we are solving

$$\min_{w \in \mathbb{R}} \mathcal{L}(w) \quad := \quad \min_{w \in \mathbb{R}} (y_0 - wx_0)^2$$

  - Question. How do we solve?
  - Answer. Inspect the **critical points**, where $\dfrac{\partial \mathcal{L}(w)}{\partial w} = 0$

Why matrix calculus?

- **Multivariate case.** Use vector / matrix calculus to find optimal parameter.

  - Example. Find a linear model $f(\mathbf{x}) = \mathbf{W}\mathbf{x}, \quad \mathbf{W} \in \mathbb{R}^{m \times n}$
    that minimizes the squared ℓ_2 loss $\ell(\hat{\mathbf{y}}, \mathbf{y}) = \|\hat{\mathbf{y}} - \mathbf{y}\|_2^2$.

    Given a single datum $(\mathbf{x}_0, \mathbf{y}_0)$, we want to inspect the **critical point** where

$$\frac{\partial(\|\mathbf{y}_0 - \mathbf{W}\mathbf{x}_0\|_2^2)}{\partial \mathbf{W}} = \mathbf{0}$$

Why matrix calculus?

- **Multivariate case.** Use vector / matrix calculus to find optimal parameter.

  - Example. Find a linear model $f(\mathbf{x}) = \mathbf{W}\mathbf{x}, \quad \mathbf{W} \in \mathbb{R}^{m \times n}$
    that minimizes the squared ℓ_2 loss $\ell(\hat{\mathbf{y}}, \mathbf{y}) = \|\hat{\mathbf{y}} - \mathbf{y}\|_2^2$.

    Given a single datum $(\mathbf{x}_0, \mathbf{y}_0)$, we want to inspect the **critical point** where

$$\frac{\partial(\|\mathbf{y}_0 - \mathbf{W}\mathbf{x}_0\|_2^2)}{\partial \mathbf{W}} = \mathbf{0}$$

- Want to know. How to handle the **gradient w.r.t. matrices**

  - Note. Sometimes, we want to run iterative algorithms to find solutions (e.g., GD),
    —> this still requires evaluating gradients (more on next class)

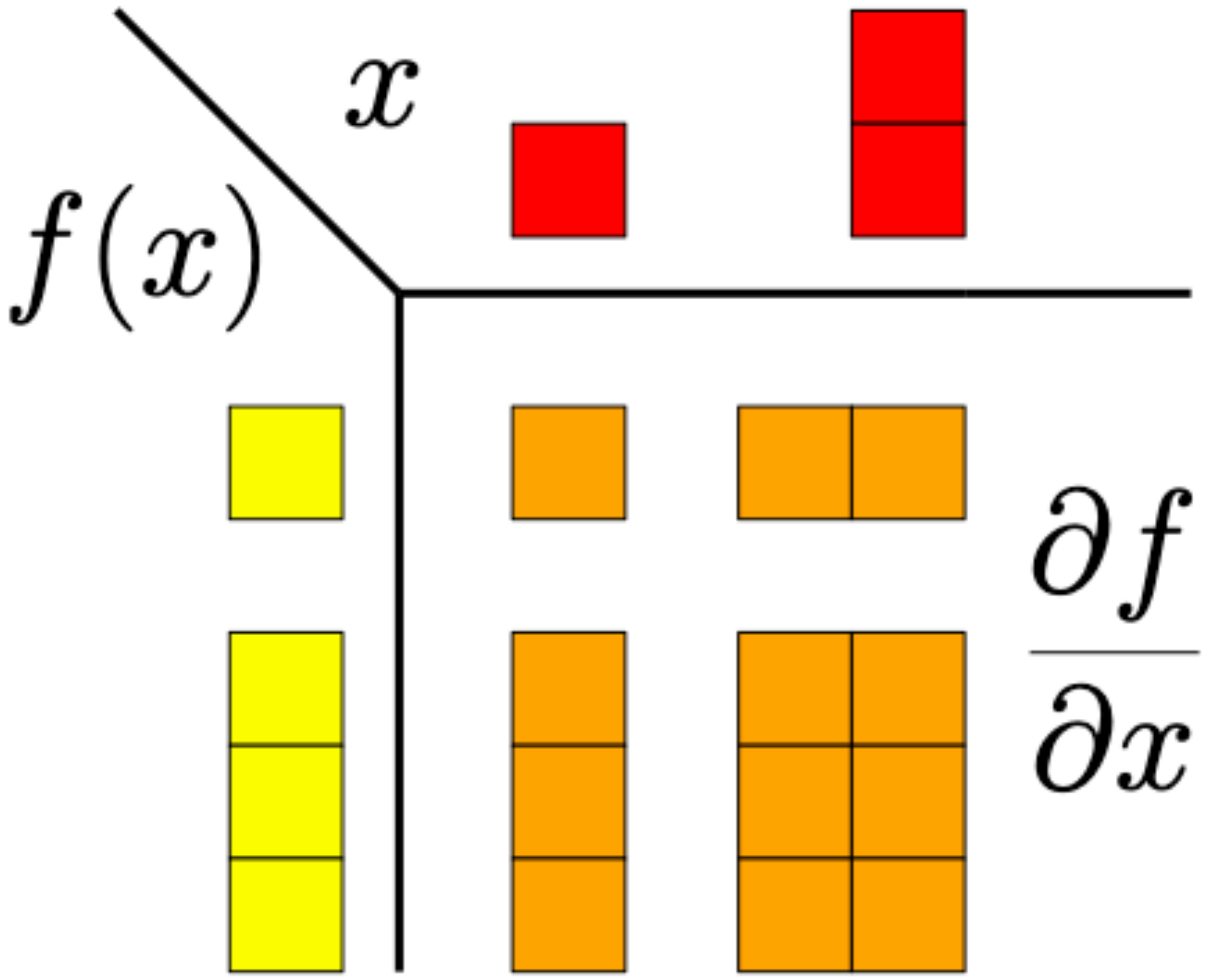
Gradients

- For a scalar variable x, differentiating a ...
  - Scalar function $y \in \mathbb{R}$: $$\frac{\partial y}{\partial x}$$
  - Vector function $\mathbf{y} \in \mathbb{R}^m$: $$\frac{\partial \mathbf{y}}{\partial x} = \begin{bmatrix} \frac{\partial y_1}{\partial x} & \cdots & \frac{\partial y_m}{\partial x} \end{bmatrix}^\top$$
  - Matrix function: $\mathbf{Y} \in \mathbb{R}^{m \times n}$: $$\frac{\partial \mathbf{Y}}{\partial x} = \begin{bmatrix} \frac{\partial y_{11}}{\partial x} & \cdots & \frac{\partial y_{1n}}{\partial x} \\ & \cdots & \\ \frac{\partial y_{m1}}{\partial x} & \cdots & \frac{\partial y_{mn}}{\partial x} \end{bmatrix}$$

Gradients

- For a vector variable $\mathbf{x} \in \mathbb{R}^n$, differentiating a ...
  - Scalar function $y \in \mathbb{R}$: $$\frac{\partial y}{\partial \mathbf{x}} = \begin{bmatrix} \frac{\partial y}{\partial x_1} & \cdots & \frac{\partial y}{\partial x_n} \end{bmatrix}$$
  - Vector function $\mathbf{y} \in \mathbb{R}^m$: $$\frac{\partial \mathbf{y}}{\partial \mathbf{x}} = \begin{bmatrix} \frac{\partial y_1}{\partial x_1} & \cdots & \frac{\partial y_1}{\partial x_n} \\ & \cdots & \\ \frac{\partial y_m}{\partial x_1} & \cdots & \frac{\partial y_m}{\partial x_n} \end{bmatrix}$$
  - Note. the direction!

Figure 5.2 Dimensionality of (partial) derivatives.



Gradients

- For a matrix variable $\mathbf{X} \in \mathbb{R}^{m \times n}$, differentiating a ...
  - Scalar function $y \in \mathbb{R}$:

$$\frac{\partial y}{\partial \mathbf{X}} = \begin{bmatrix} \frac{\partial y}{\partial x_{11}} & \cdots & \frac{\partial y}{\partial x_{m1}} \\ & \cdots & \\ \frac{\partial y}{\partial x_{1n}} & \cdots & \frac{\partial y}{\partial x_{mn}} \end{bmatrix}$$

  - Note. again, the direction!

Reference for self-study

- MML book Sec. 5
- https://en.wikipedia.org/wiki/Matrix_calculus

Condition	Expression	Numerator layout, i.e. by $\mathbf{y}$ and $\mathbf{x}^\mathrm{T}$	Denominator layout, i.e. by $\mathbf{y}^\mathrm{T}$ and $\mathbf{x}$
$\mathbf{a}$ is not a function of $\mathbf{x}$	$\frac{\partial \mathbf{a}}{\partial \mathbf{x}} =$	$\mathbf{0}$	
	$\frac{\partial \mathbf{x}}{\partial \mathbf{x}} =$	$\mathbf{I}$	
$\mathbf{A}$ is not a function of $\mathbf{x}$	$\frac{\partial \mathbf{A}\mathbf{x}}{\partial \mathbf{x}} =$	$\mathbf{A}$	$\mathbf{A}^\top$
$\mathbf{A}$ is not a function of $\mathbf{x}$	$\frac{\partial \mathbf{x}^\top \mathbf{A}}{\partial \mathbf{x}} =$	$\mathbf{A}^\top$	$\mathbf{A}$
a is not a function of $\mathbf{x}$, $\mathbf{u} = \mathbf{u}(\mathbf{x})$	$\frac{\partial a\mathbf{u}}{\partial \mathbf{x}} =$	$a\frac{\partial \mathbf{u}}{\partial \mathbf{x}}$	
$v = v(\mathbf{x})$, $\mathbf{a}$ is not a function of $\mathbf{x}$	$\frac{\partial v\mathbf{a}}{\partial \mathbf{x}} =$	$\mathbf{a}\frac{\partial v}{\partial \mathbf{x}}$	$\frac{\partial v}{\partial \mathbf{x}}\mathbf{a}^\top$
$v = v(\mathbf{x})$, $\mathbf{u} = \mathbf{u}(\mathbf{x})$	$\frac{\partial v\mathbf{u}}{\partial \mathbf{x}} =$	$v\frac{\partial \mathbf{u}}{\partial \mathbf{x}} + \mathbf{u}\frac{\partial v}{\partial \mathbf{x}}$	$v\frac{\partial \mathbf{u}}{\partial \mathbf{x}} + \frac{\partial v}{\partial \mathbf{x}}\mathbf{u}^\top$
$\mathbf{A}$ is not a function of $\mathbf{x}$, $\mathbf{u} = \mathbf{u}(\mathbf{x})$	$\frac{\partial \mathbf{A}\mathbf{u}}{\partial \mathbf{x}} =$	$\mathbf{A}\frac{\partial \mathbf{u}}{\partial \mathbf{x}}$	$\frac{\partial \mathbf{u}}{\partial \mathbf{x}}\mathbf{A}^\top$
$\mathbf{u} = \mathbf{u}(\mathbf{x})$, $\mathbf{v} = \mathbf{v}(\mathbf{x})$	$\frac{\partial (\mathbf{u} + \mathbf{v})}{\partial \mathbf{x}} =$	$\frac{\partial \mathbf{u}}{\partial \mathbf{x}} + \frac{\partial \mathbf{v}}{\partial \mathbf{x}}$	
$\mathbf{u} = \mathbf{u}(\mathbf{x})$	$\frac{\partial \mathbf{g}(\mathbf{u})}{\partial \mathbf{x}} =$	$\frac{\partial \mathbf{g}(\mathbf{u})}{\partial \mathbf{u}}\frac{\partial \mathbf{u}}{\partial \mathbf{x}}$	$\frac{\partial \mathbf{u}}{\partial \mathbf{x}}\frac{\partial \mathbf{g}(\mathbf{u})}{\partial \mathbf{u}}$
$\mathbf{u} = \mathbf{u}(\mathbf{x})$	$\frac{\partial \mathbf{f}(\mathbf{g}(\mathbf{u}))}{\partial \mathbf{x}} =$	$\frac{\partial \mathbf{f}(\mathbf{g})}{\partial \mathbf{g}}\frac{\partial \mathbf{g}(\mathbf{u})}{\partial \mathbf{u}}\frac{\partial \mathbf{u}}{\partial \mathbf{x}}$	$\frac{\partial \mathbf{u}}{\partial \mathbf{x}}\frac{\partial \mathbf{g}(\mathbf{u})}{\partial \mathbf{u}}\frac{\partial \mathbf{f}(\mathbf{g})}{\partial \mathbf{g}}$

Probability

Why probability?

- In ML, many things are **random**

Why probability?

- In ML, many things are **random**
  - The **data** is drawn randomly
    - Training data $Z_1, \cdots, Z_n \sim P$
    - Test data $Z_{\text{new}} \sim \tilde{P}$

Why probability?

- In ML, many things are **random**
  - The **data** is drawn randomly
    - Training data $Z_1, \cdots, Z_n \sim P$
    - Test data $Z_{\text{new}} \sim \tilde{P}$
  - **Components of learning algorithms** are randomly selected
    - Examples. Initial parameter (neural nets, k-means)
      SGD ordering
      Noise
    - Reason. Enable efficient computation (Monte Carlo)
      Random "likely contains every direction"

Probability

- Mathematical foundation due to Kolmogorov (1930s)
- The **probability space** $(\Omega, \mathcal{F}, P)$ is a triplet of
  - **Sample space** Ω
    - Set of all possible outcomes
  - **Event space** $\mathcal{F}$
    - Set of all events (set of outcomes)
  - **Probability measure** $P : \mathcal{F} \rightarrow [0,1]$
    - Chances assigned to each event

Probability

- Consider **rolling a die**:
  - **Sample space**
    - $\Omega = \{1,2,3,4,5,6\}$
  - **Event space**
    - $\mathscr{F} = \left\{\varnothing, \{1\}, \cdots, \{6\}, \{1,2\}, \cdots, \{5,6\}, \cdots, \{1,2,3,4,5,6\}\right\}$
  - **Probability measure** $P : \mathscr{F} \to [0,1]$ (or probability distribution)
    - $P(\varnothing) = 0, \quad P(\{1\}) = 1/6, \quad \cdots, \quad P(\{1,2,3,4,5,6\}) = 1$
      - Note. This should satisfy certain properties!

Probability Measure

- A **probability measure** is a function $P : \mathscr{F} \to [0,1]$ satisfying the following axioms.
  - $P(\Omega) = 1$
    - i.e., an outcome will happen, eventually.

Probability Measure

- A **probability measure** is a function $P : \mathcal{F} \to [0,1]$ satisfying the following axioms.
  - $P(\Omega) = 1$
    - i.e., an outcome will happen, eventually.
  - $P(A) \geq 0, \qquad \forall A \in \mathcal{F}$
    - i.e., there is no such thing as negative probability

Probability Measure

- A **probability measure** is a function $P : \mathscr{F} \to [0,1]$ satisfying the following axioms.
  - $P(\Omega) = 1$
    - i.e., an outcome will happen, eventually.
  - $P(A) \geq 0, \qquad \forall A \in \mathscr{F}$
    - i.e., there is no such thing as negative probability
  - $P(A \cup B) = P(A) + P(B), \qquad \text{whenever } A \cup B = \varnothing$
    - called "additivity"   <— should hold for any **countable** number of mutually exclusive events
    - Note (advanced). To generalize to arbitrary space, people use special math (σ-algebra …)

Random variable

Random variable

- We avoid dealing directly with the probability space (for a good reason)

  - A **random variable** is a real-valued function $X : \Omega \to \mathbb{R}$

Random variable

- We avoid dealing directly with the probability space (for a good reason)
  - A **random variable** is a real-valued function $X : \Omega \to \mathbb{R}$
    - Example. For coin tossing where $\Omega = \{H, T\}$, we may define the random variable

$$X(H) = 1, \quad X(T) = 0$$

      - Here, we can say that the probability of $X = 1$ under P is equal to $P(\{H\})$
        - Simply use the shorthand $P(X = 1)$

Cumulative Distribution Function (CDF)

- A **CDF** is defined as

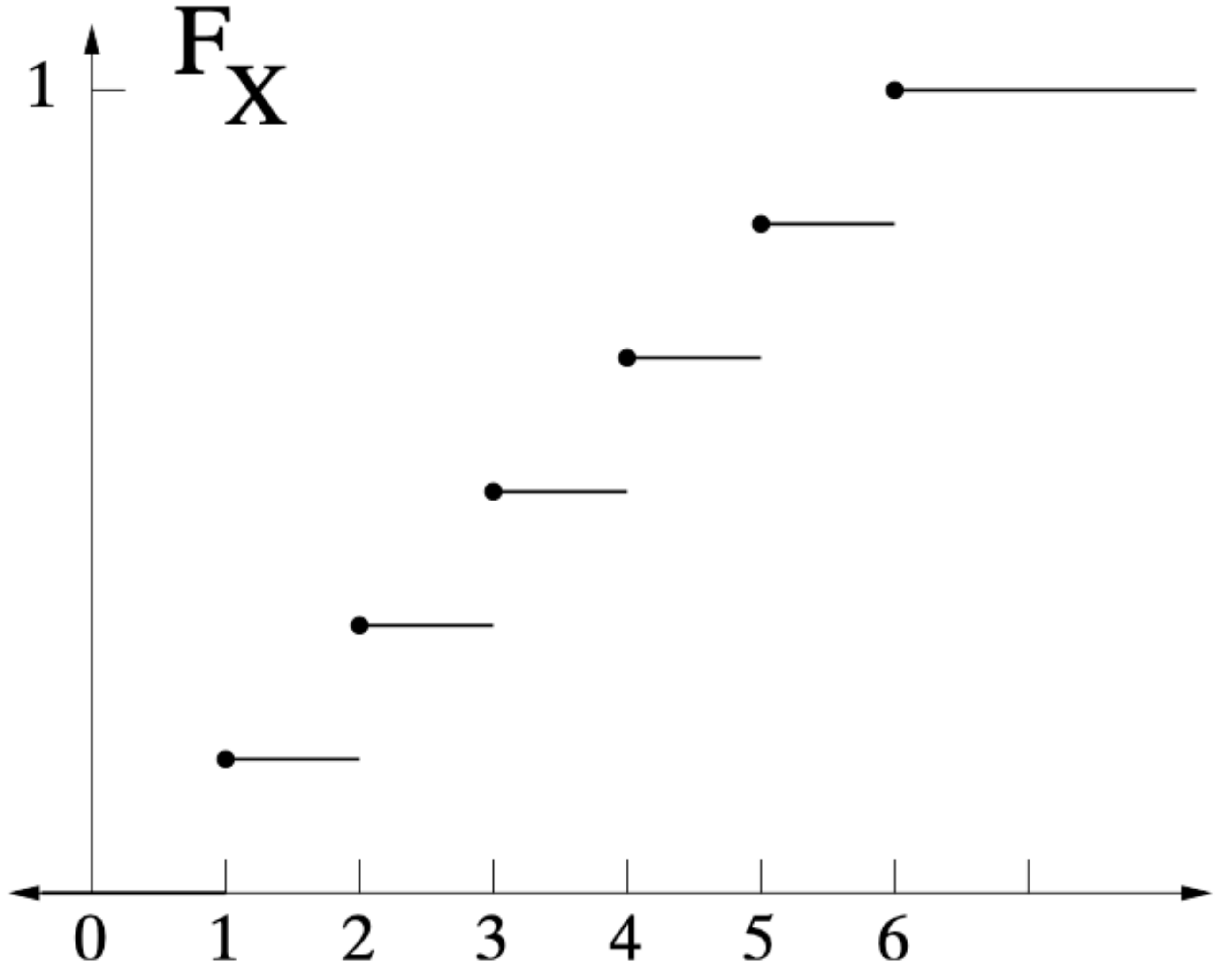
$$F_X(x) := P(X \leq x)$$

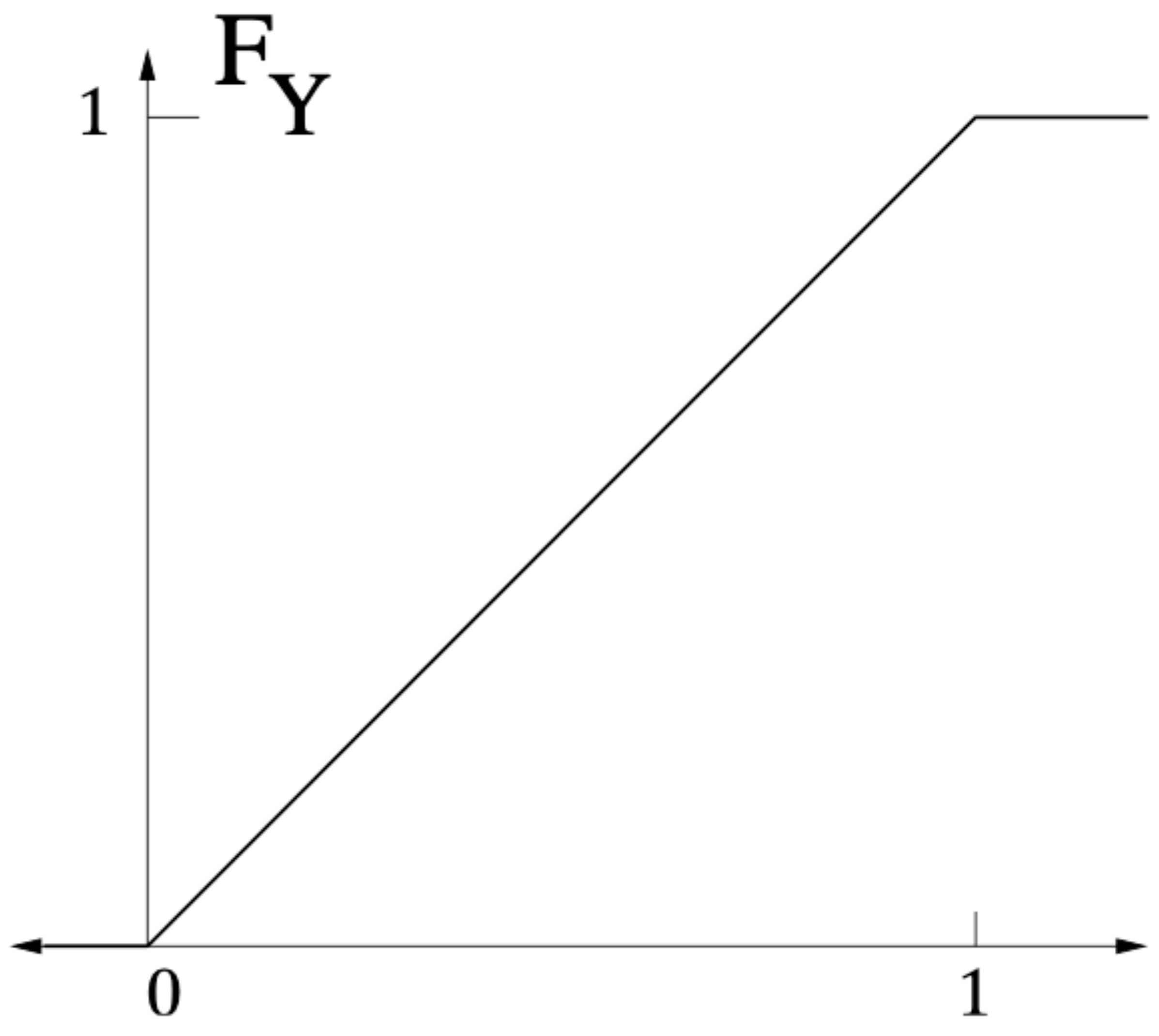
Cumulative Distribution Function (CDF)

- A **CDF** is defined as

$$F_X(x) := P(X \leq x)$$

- Properties.
  - $0 \leq F_X(x) \leq 1$
  - $F_X(-\infty) = 0$
  - $F_X(\infty) = 1$
  - If $x \leq y$, then $F_X(x) \leq F_X(y)$

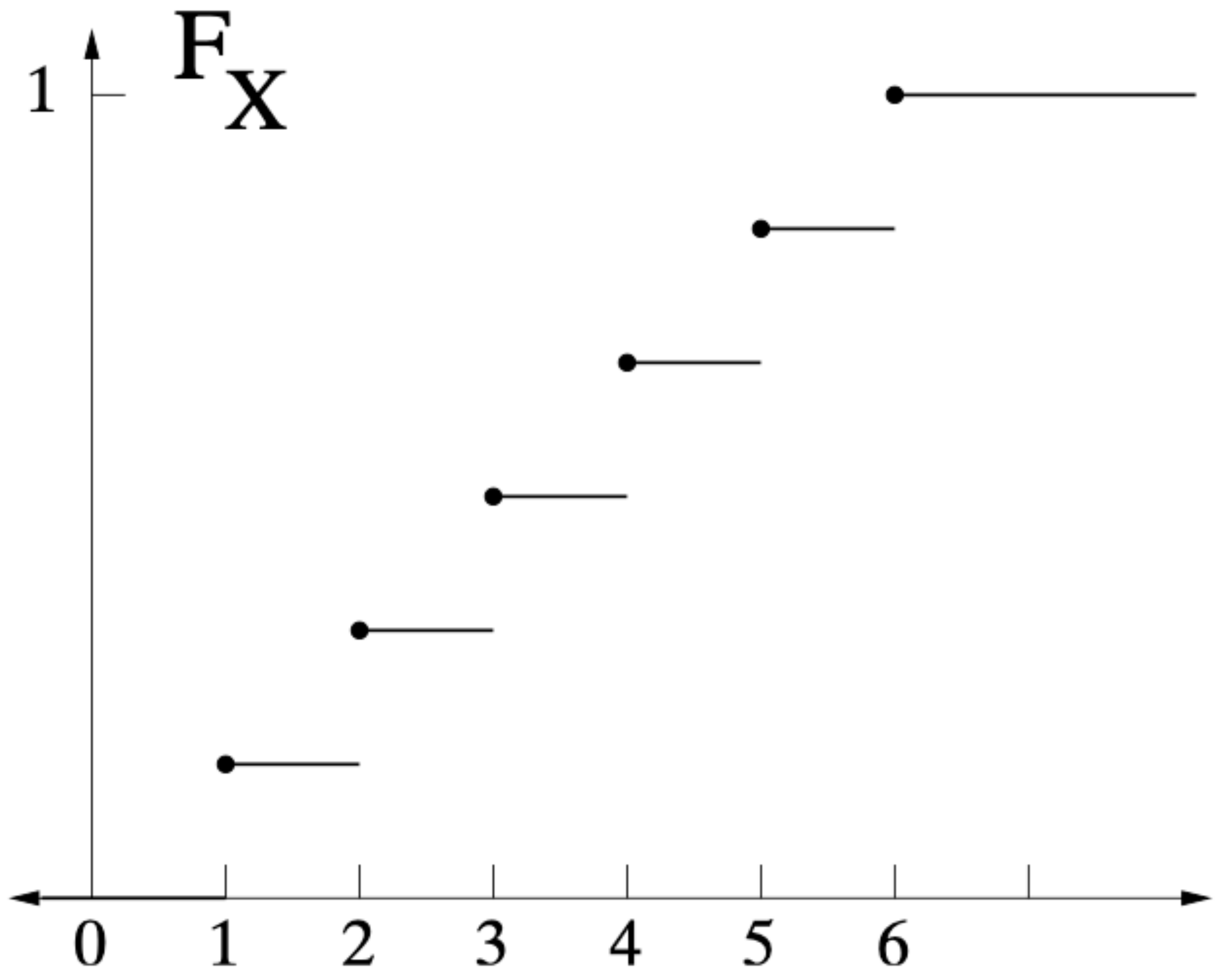




Probability Mass Function (PMF)

- For a <u>discrete random variable</u> X, the **PMF** is defined as

$$p_X(x) := P(X = x)$$

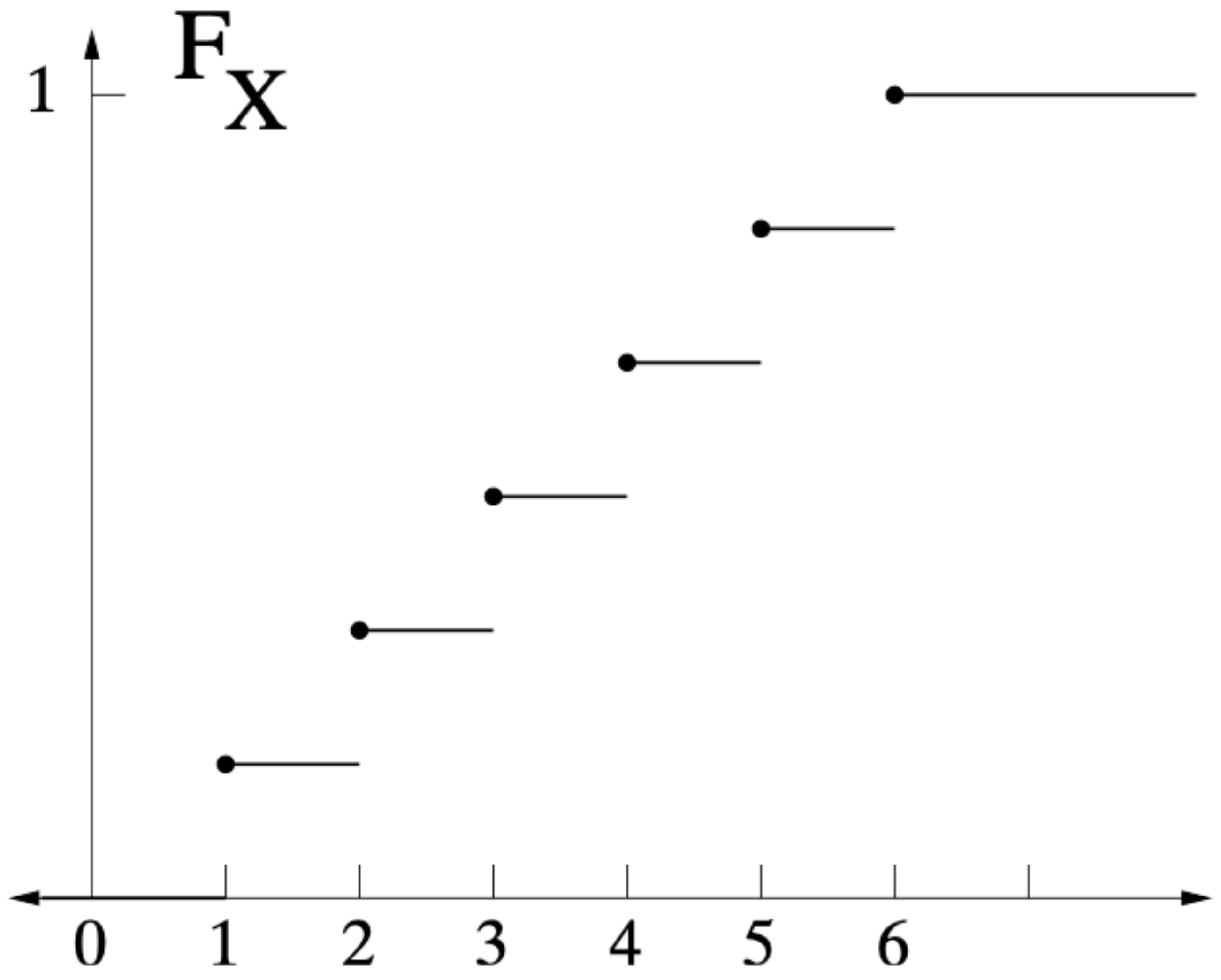


Probability Mass Function (PMF)

- For a <u>discrete random variable</u> X, the **PMF** is defined as

$$p_X(x) := P(X = x)$$

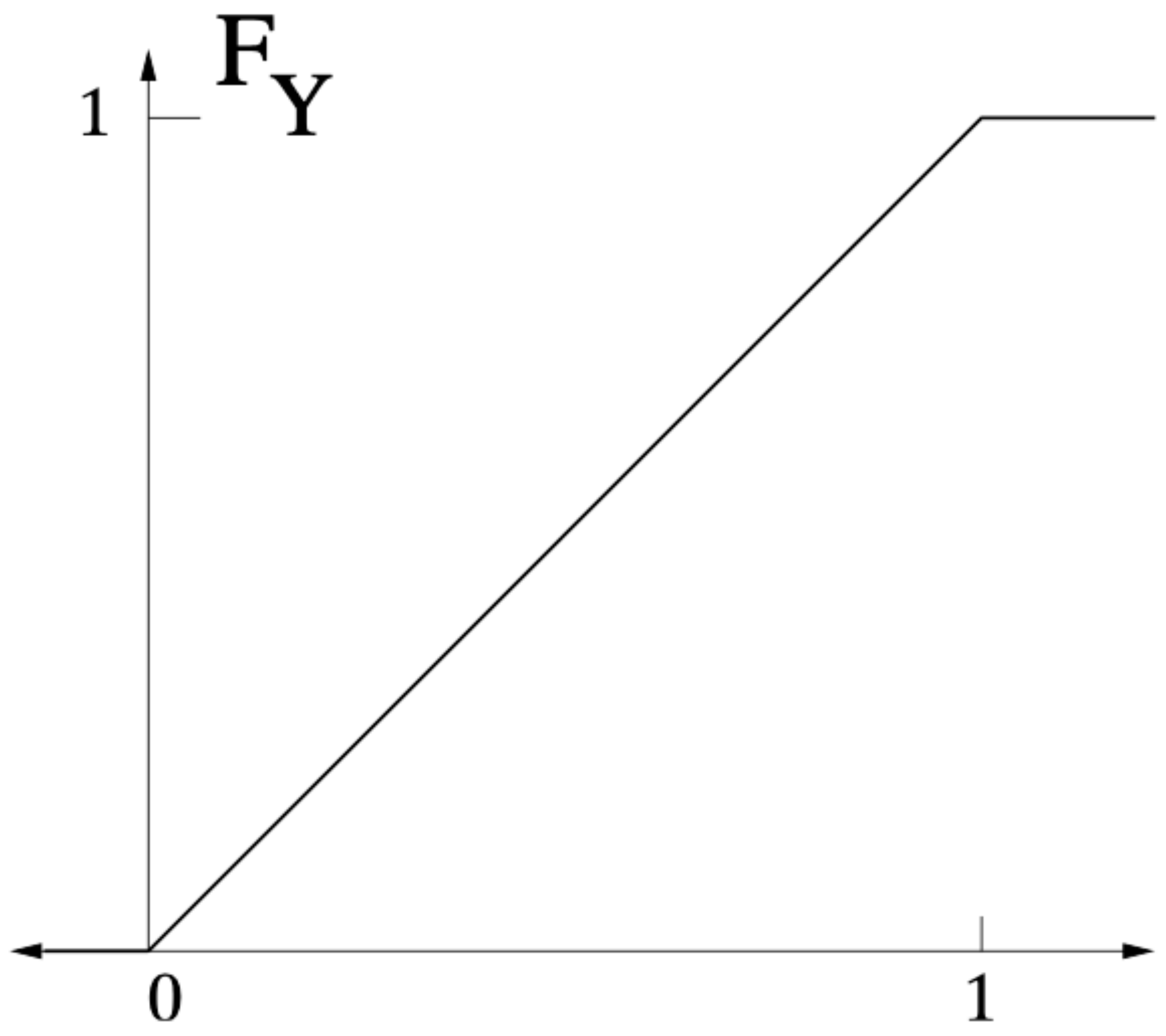
- <u>Properties</u>.
  - $0 \leq p_X(x) \leq 1$
  - $\sum_x p_X(x) = 1$
  - $\sum_{x \in A} p_X(x) = P(X \in A)$



Probability Density Function (PDF)

- For a <u>continuous random variable</u> X, the **PDF** is defined as

$$f_X(s) := \frac{\partial F_X(x)}{\partial x}(s)$$

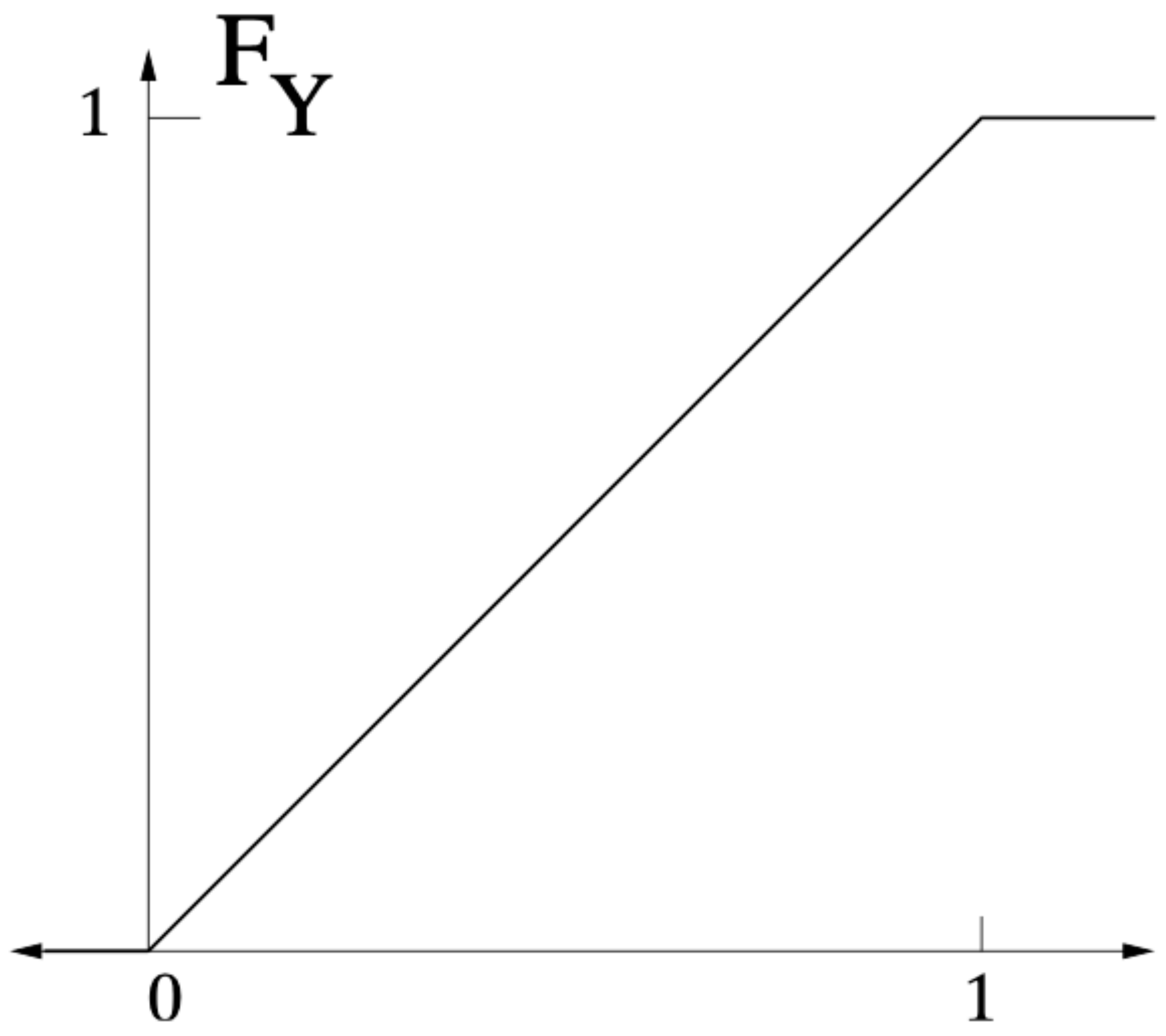


Probability Density Function (PDF)

- For a <u>continuous random variable</u> X, the **PDF** is defined as

$$f_X(s) := \frac{\partial F_X(x)}{\partial x}(s)$$

- <u>Properties</u>.
  - $0 \leq f_X(x)$
  - $\int_{\mathbb{R}} f_X(x)\, \mathrm{d}x = 1$
  - $\int_A f_X(x)\, \mathrm{d}x = P(X \in A)$



Probability Density Function (PDF)

- Note. PDF is not really the probability itself
  - Only gives you an estimate via

$$P(x \leq X \leq x + \mathrm{d}x) \approx p(x)\,\mathrm{d}x$$

used interchangeably with $f_X(x)$

  - Thus, it is okay to have $p(x) > 1$

Joint distribution

- Characterized by the **joint CDF**

$$F_{XY}(x, y) = P(X \leq x, Y \leq y)$$

Joint distribution

- Characterized by the **joint CDF**

$$F_{XY}(x, y) = P(X \leq x, Y \leq y)$$

- **Marginal CDF** can be recovered via

$$F_X(x) = \lim_{y \to \infty} F_{XY}(x, y), \qquad F_Y(y) = \lim_{x \to \infty} F_{XY}(x, y)$$

Joint distribution

- Characterized by the **joint CDF**

$$F_{XY}(x, y) = P(X \leq x, Y \leq y)$$

- **Marginal CDF** can be recovered via

$$F_X(x) = \lim_{y \to \infty} F_{XY}(x, y), \qquad F_Y(y) = \lim_{x \to \infty} F_{XY}(x, y)$$

- When discrete, we write the **joint PMF** as

$$p_{XY}(x, y) = P(X = x, Y = y)$$

where we have $p_X(x) = \sum_y p_{XY}(x, y)$

Conditional distribution

- **Conditional probability** of an event is given as

$$P(A \mid B) = \frac{P(A, B)}{P(B)}$$

both A and B happening; $P(A \cap B)$, precisely

Conditional distribution

- **Conditional probability** of an event is given as

$$P(A \mid B) = \frac{P(A, B)}{P(B)}$$

- Conditional PMF (discrete)

$$p_{Y|X}(y \mid x) = \frac{p_{XY}(x, y)}{p_X(x)}$$

- Conditional PDF (continuous)

$$f_{Y|X}(y \mid x) = \frac{f_{XY}(x, y)}{f(x)}$$

Basic arithmetics

- Product rule.

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

- Bayes' theorem.

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

Statistics of random variables

Expectation (1st order)

- For discrete random variables, the **expected value** is defined as a weighted sum

$$\mathbb{E}[g(X)] = \sum_x g(x) p_X(x)$$

- For continuous r.v.s, defined as

$$\mathbb{E}[g(X)] = \int_{\mathbb{R}} g(x) f_X(x) \, \mathrm{d}x$$

Expectation (1st order)

- For discrete random variables, the **expected value** is defined as a weighted sum

$$\mathbb{E}[g(X)] = \sum_x g(x)p_X(x)$$

- For continuous r.v.s, defined as

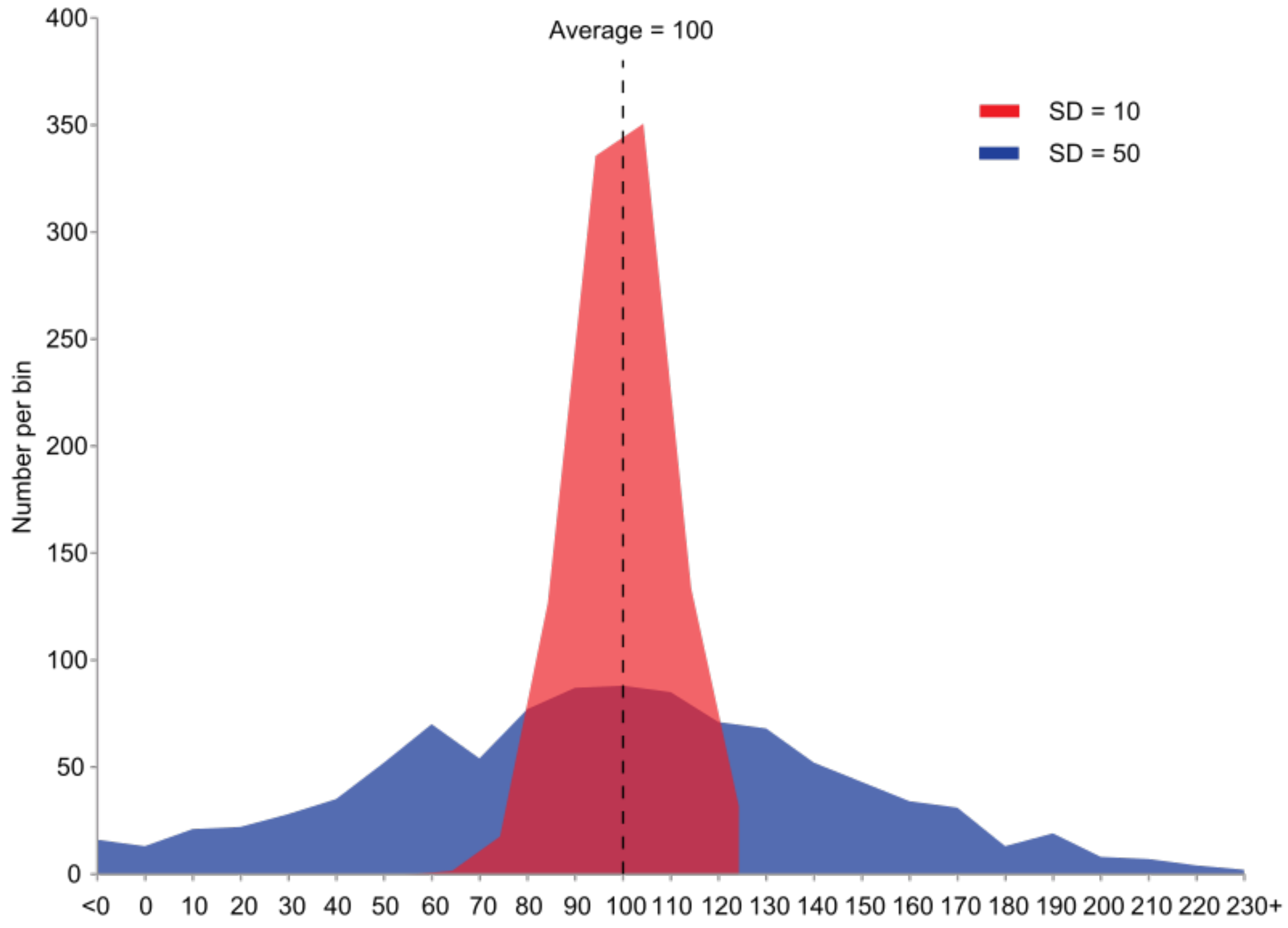
$$\mathbb{E}[g(X)] = \int_{\mathbb{R}} g(x)f_X(x)\,\mathrm{d}x$$

- Properties.
    - $\mathbb{E}[a] = a$, for a constant a
    - $\mathbb{E}[af(X) + bg(X)] = a\mathbb{E}[f(X)] + b\mathbb{E}[g(X)]$ (linearity)

Variance (2nd order)

- The **variance** is defined as

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



Variance (2nd order)

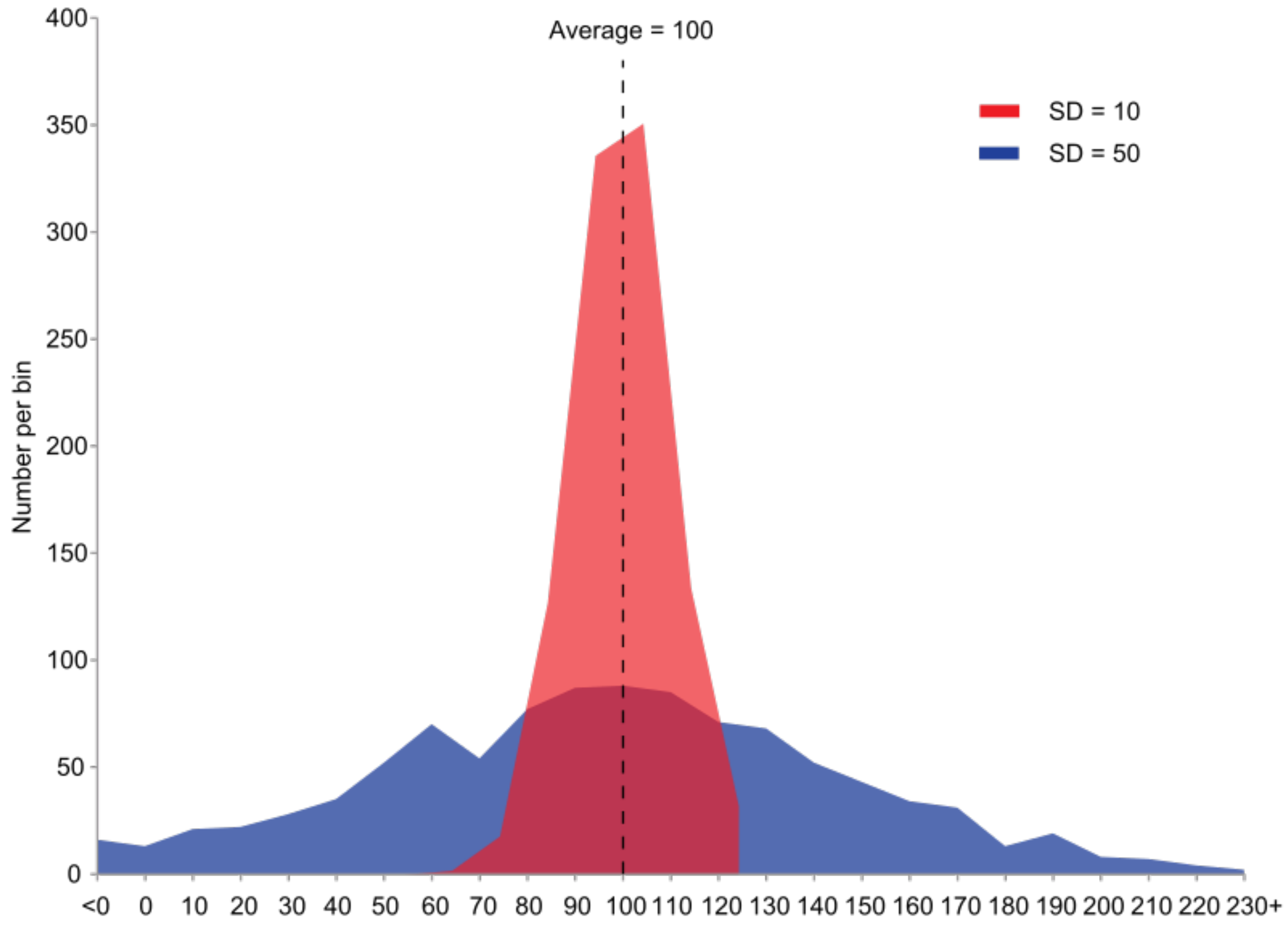
- The **variance** is defined as

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

  - Properties.
    - $\mathrm{Var}[a] = 0,$ for constant a
    - $\mathrm{Var}[af(X)] = a^2\mathrm{Var}[f(X)]$

- The **standard deviation** is defined as

$$\sigma_X = \sqrt{\mathrm{Var}(X)}$$



A fact

- **Question.** Suppose that we have a random variable X, with a known distribution $P(X)$.
    - What is our **best blind guess of X** when we want to minimize the expected squared error?

$$\min_{c \in \mathbb{R}} \mathbb{E}[(X - c)^2]$$

    - How much would the expected squared error be, for this estimate?

Another fact

- **Question.** What is our best guess, if we are no longer blind and can **utilize some observation Y** jointly distributed with X?

$$\min_f \mathbb{E}[(X - f(Y))^2]$$

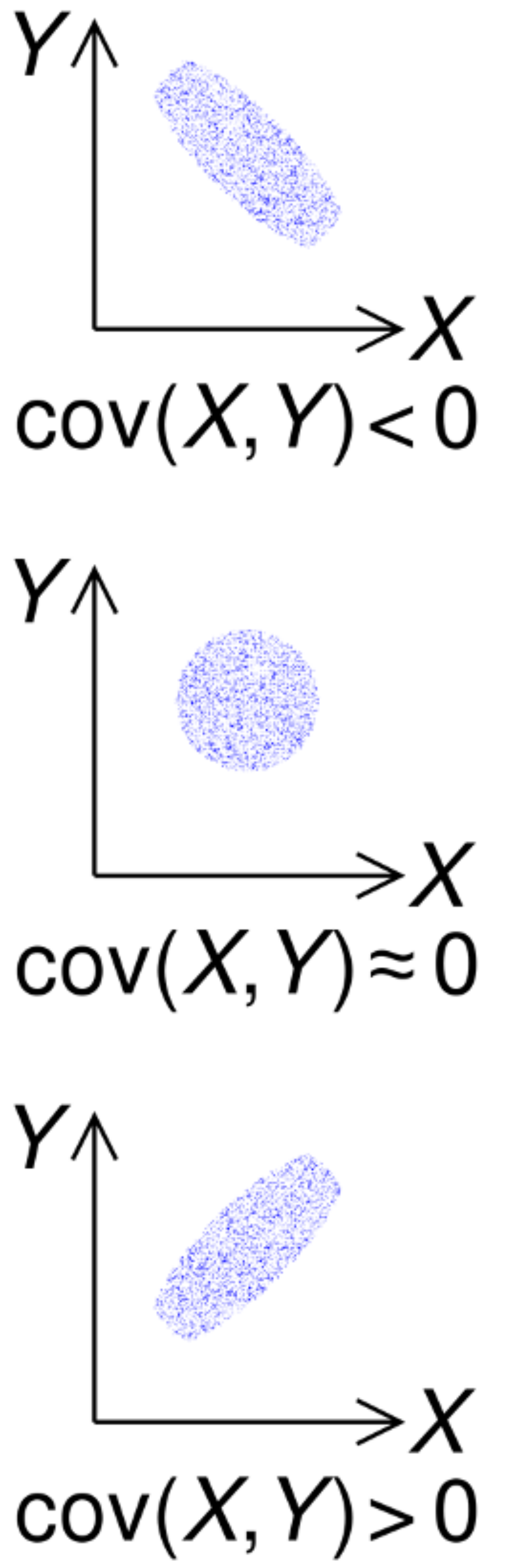
  - That is, $(X, Y) \sim p_{XY}$ and X is not known, Y is observed.

Covariance and Correlation

- **Covariance** measures the joint variability of two RVs.

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

  - Related to whether one variable is predictive of another



Covariance and Correlation

- **Covariance** measures the joint variability of two RVs.

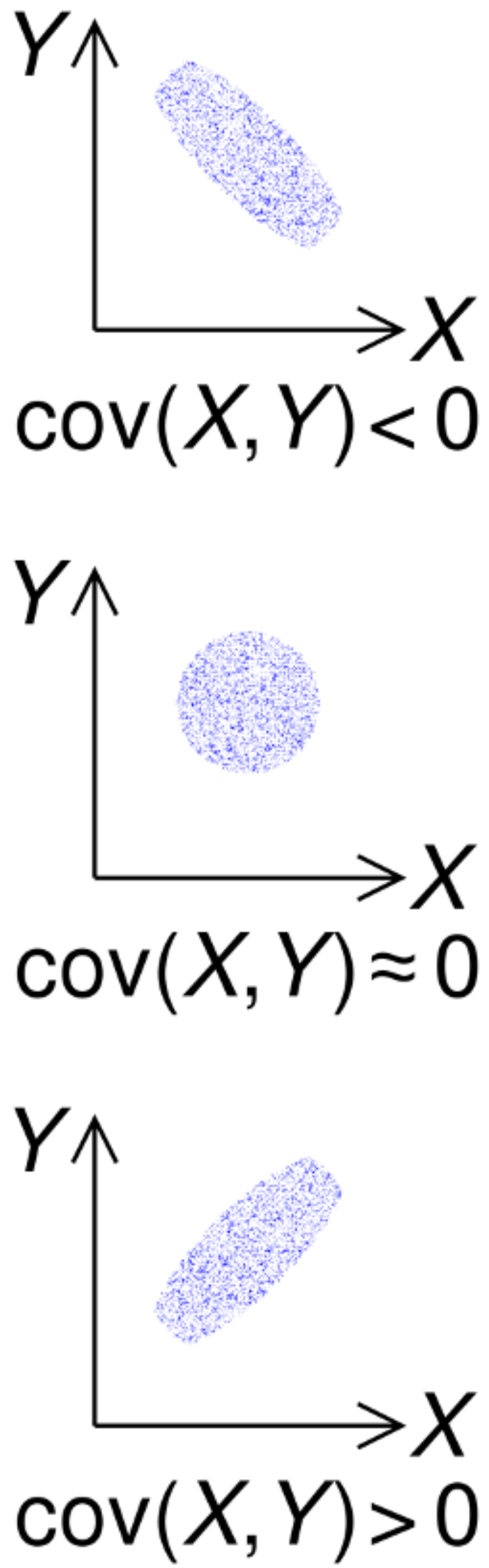
$$\mathrm{Cov}[X, Y] := \mathbb{E}[(X - \mathbb{E}[X])(Y - \mathbb{E}[Y])]$$

  - Related to whether one variable is predictive of another

- **(Pearson) Correlation** is defined as

$$\mathrm{corr}[X, Y] = \frac{\mathrm{Cov}[X, Y]}{\sigma_X \sigma_Y}$$

  - lies in [-1,+1]



Independence

Independence

- Two random variables X and Y are **independent** whenever

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

  - Properties. If independent...
    - $p(y \mid x) = p(y)$
    - $\mathrm{Var}[X + Y] = \mathrm{Var}[X] + \mathrm{Var}[Y]$
    - $\mathrm{Cov}[X, Y] = 0$

Conditional independence

- Random variables X and Y are **conditionally independent given Z** whenever

$$p(x, y \mid z) = p(x \mid z)p(y \mid z)$$

  - Denoted by $X \perp Y \mid Z$

Conditional independence

- Random variables X and Y are **conditionally independent given** Z whenever

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

  - Denoted by $X \perp Y | Z$

- **Theorem.** We have $X \perp Y | Z$, if and only if there exists two functions g, h such that

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

  - Neat tool to verify the conditional independence
    (no need to check whether each are valid probability functions)

Common probability distributions

Bernoulli (coin toss)

- A **Bernoulli random variable** $X \sim \mathrm{Bern}(p)$ is a binary random variable with

$$P(X = 1) = p, \quad P(X = 0) = 1 - p$$

  - $\mathbb{E}[X] = p$
  - $\mathrm{Var}[X] = p(1 - p)$

Binomial (many coin tosses)

- A **Binomial random variable** $X \sim \mathrm{Bin}(n, p)$ is a discrete r.v. with

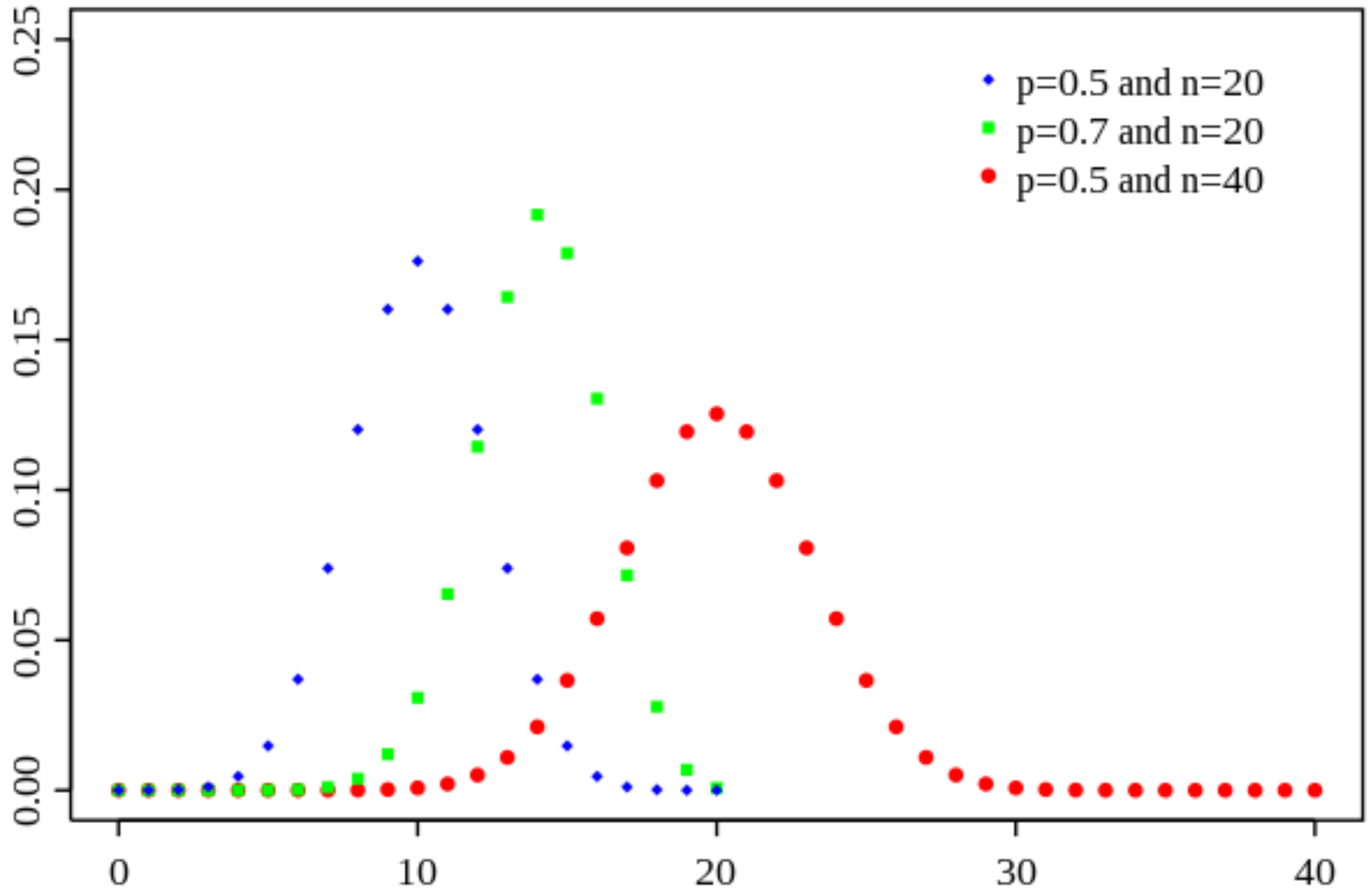
$$P(X = k) = \binom{n}{k} p^k (1-p)^{n-k}$$

  - Here, the shorthand is

$$\binom{n}{k} = \frac{n!}{k!(n-k)!}$$

  - $\mathbb{E}[X] = np$

  - $\mathrm{Var}[X] = np(1-p)$



Uniform

- Discrete. A **uniform random variable** $X \sim \text{Unif}(\{1,\ldots,k\})$ is a r.v. with

$$P(X = 1) = \cdots = P(X = k) = \frac{1}{k}$$

Uniform

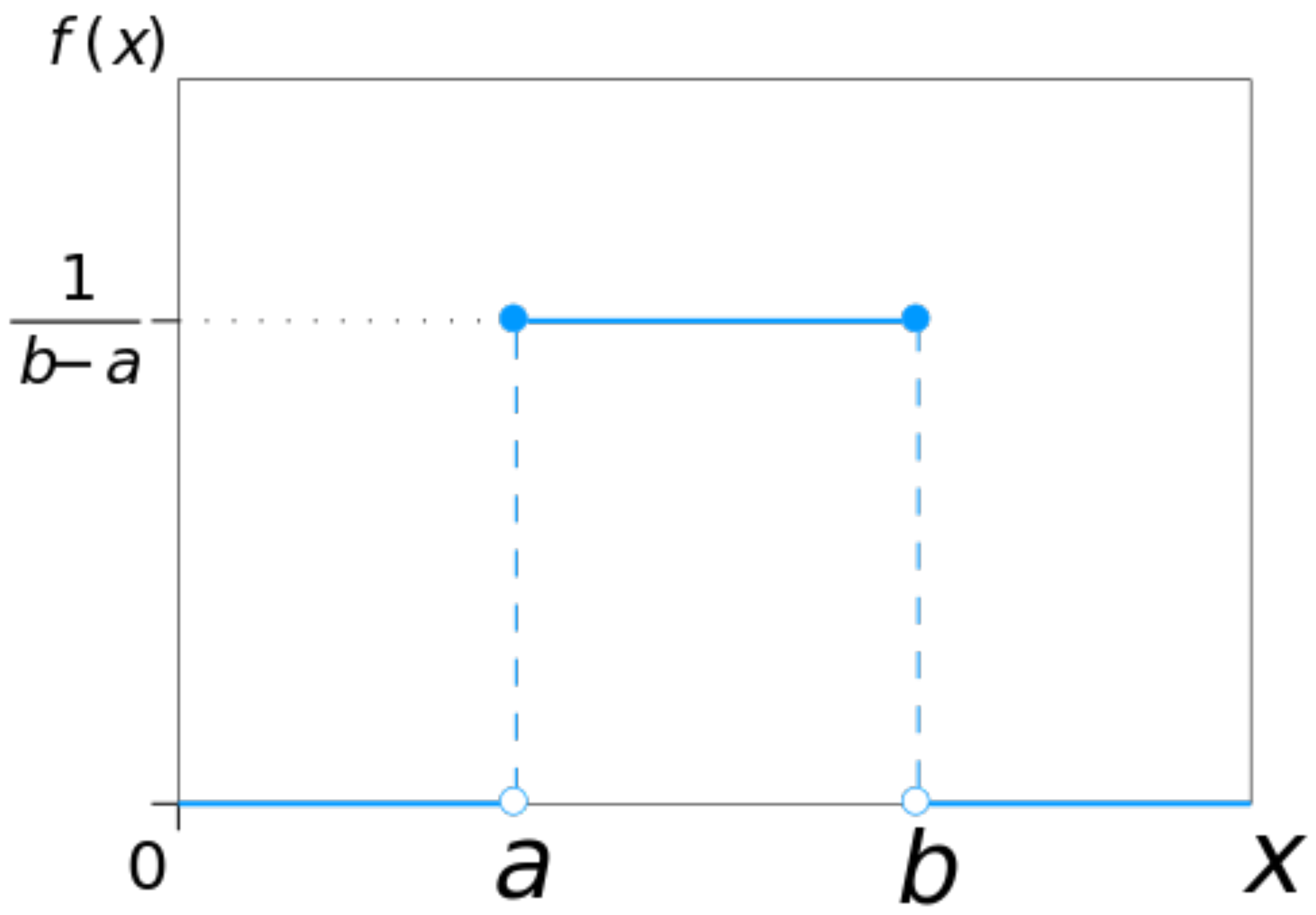
- Discrete. A **uniform random variable** $X \sim \mathrm{Unif}(\{1,\ldots,k\})$ is a r.v. with

$$P(X=1) = \cdots = P(X=k) = \frac{1}{k}$$

- Continuous. A **uniform random variable** $X \sim \mathrm{Unif}([a,b])$ is a r.v. with

$$f_X(x) = \frac{1}{b-a}\mathbf{1}\{x \in [a,b]\}$$

- $\mathbb{E}[X] = \dfrac{a+b}{2}$
- $\mathrm{Var}[X] = \dfrac{(b-a)^2}{12}$

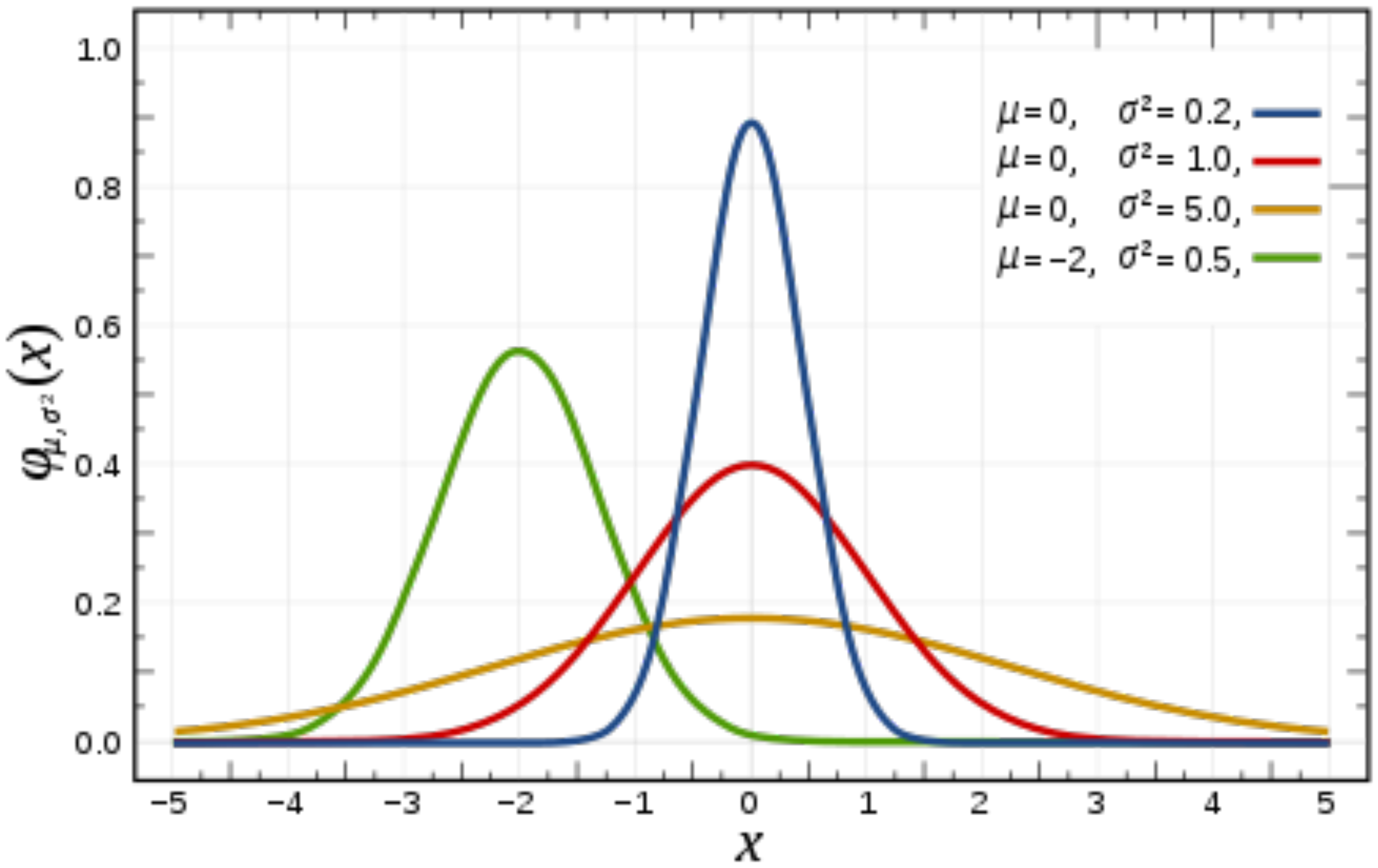


Gaussian (a.k.a. normal)

- A **Gaussian random variable** $X \sim \mathcal{N}(\mu, \sigma^2)$ is a continuous r.v. with

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

  - $\mathbb{E}[X] = \mu$
  - $\mathrm{Var}[X] = \sigma^2$
  - **Importance.** The central limit theorem (homework: review)

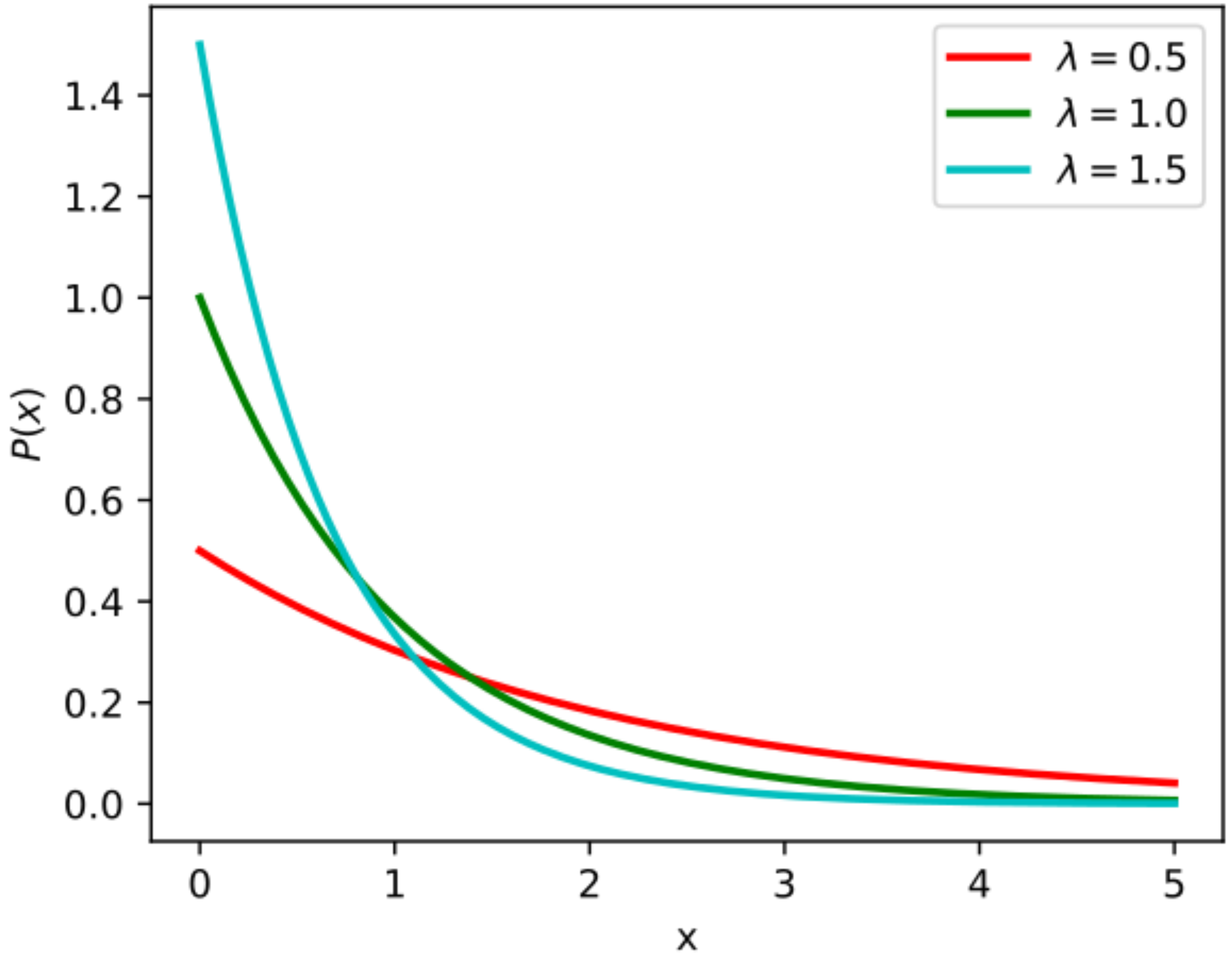


Exponential

- An **Exponential random variable** $X \sim \text{Exp}(\lambda)$ is a nonnegative continuous r.v. with

$$f_X(x) = \lambda \exp(-\lambda x)$$

  - $\mathbb{E}[X] = \dfrac{1}{\lambda}$
  - $\text{Var}[X] = \dfrac{1}{\lambda^2}$
  - Models an event that can either stop or continue at each infinitesimal time
    - Closely related with Poisson r.v. (not discussed today)

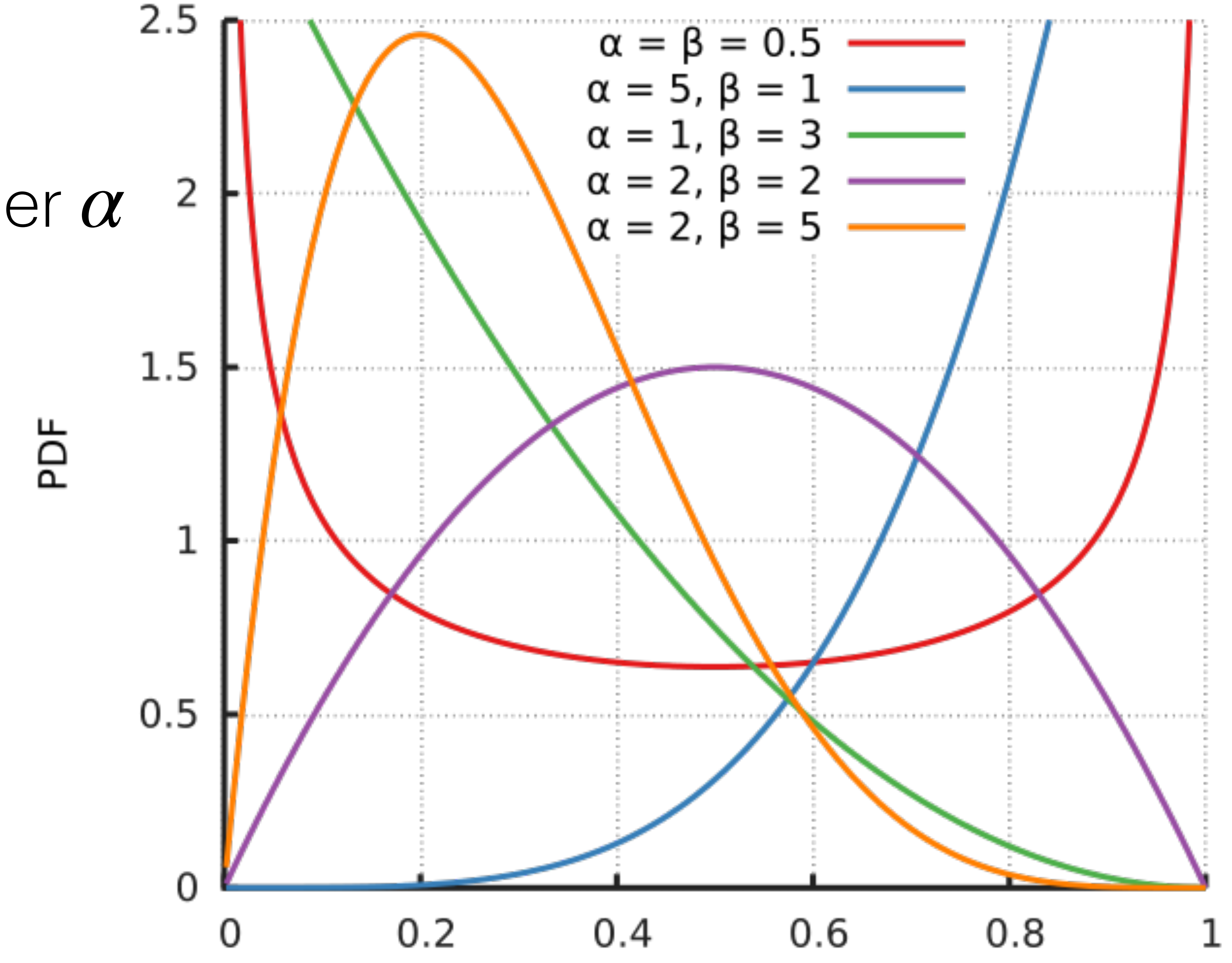


Beta

- A **Beta random variable** $X \sim \text{Beta}(\alpha, \beta)$ is a continuous r.v. with

$$f_X(x) = \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)} x^{\alpha-1}(1-x)^{\beta-1}, \qquad x \in [0,1]$$

  - Here, $\Gamma(\cdot)$ is the Gamma function
    - Complicated, but satisfies $\Gamma(\alpha) = (\alpha-1)!$ for integer α
  - $\mathbb{E}[X] = \dfrac{\alpha}{\alpha+\beta}$
  - $\text{Var}[X] = \dfrac{\alpha\beta}{(\alpha+\beta)^2(\alpha+\beta+1)}$
  - General version of uniform r.v.



Gamma

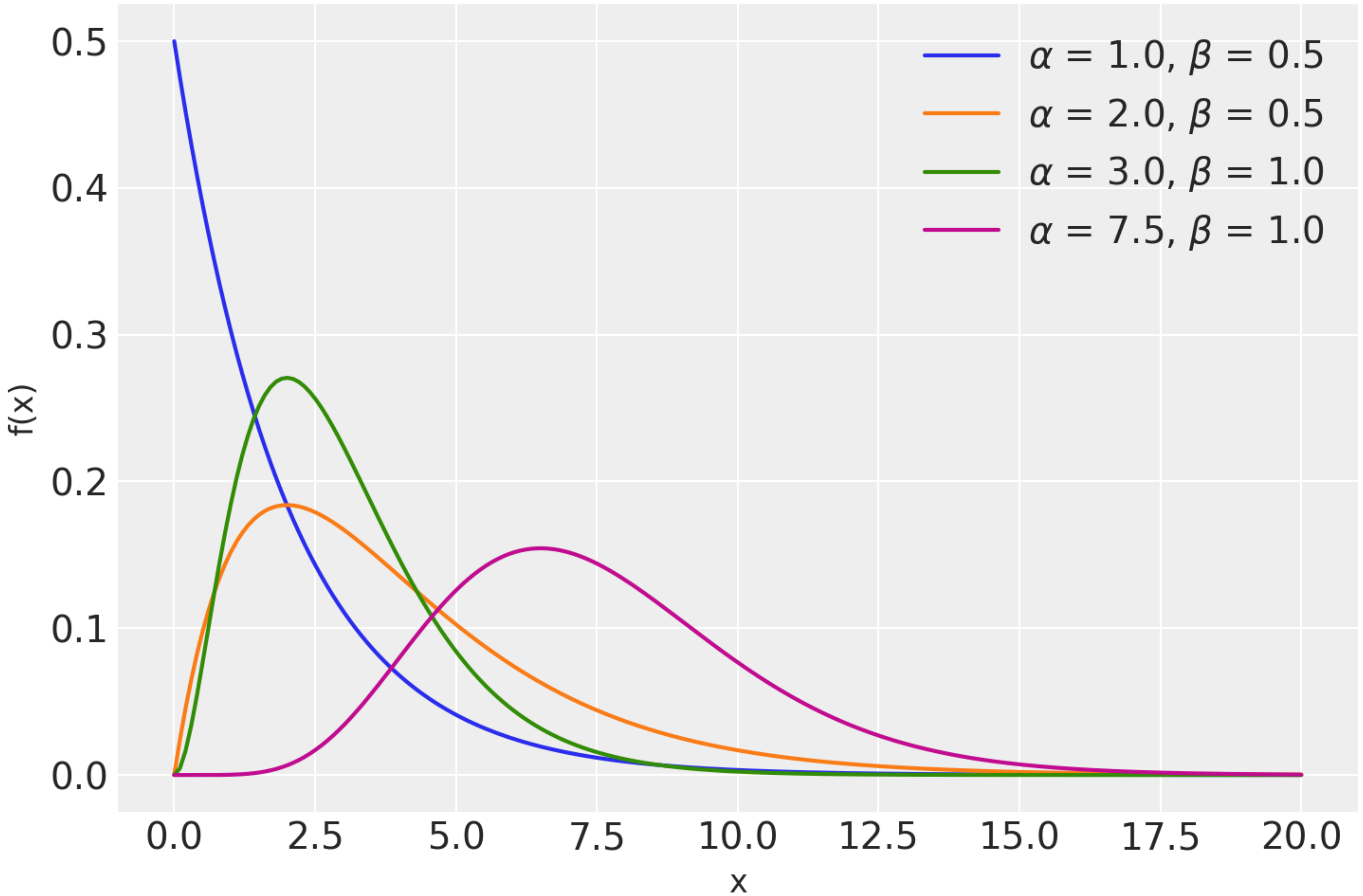
- A **Gamma random variable** $X \sim \text{Gamma}(\alpha, \beta)$ is a continuous r.v. with

$$f_X(x) = \frac{1}{\Gamma(a)} \beta^\alpha x^{\alpha-1} \exp(-\beta x)$$

- $\mathbb{E}[X] = \dfrac{\alpha}{\beta}$

- $\text{Var}[X] = \dfrac{\alpha}{\beta^2}$

- Generalizes the exponential distribution



Concentration inequalities

Concentration inequalities

- Gives more fine-grained information on the **tail behavior** of r.v.s

- Typically takes the form

$$P(X - \mathbb{E}[X] > t) \quad \leq \quad \text{small value}$$

Concentration inequalities

- Gives more fine-grained information on the **tail behavior** of r.v.s

- Typically takes the form

$$P(X - \mathbb{E}[X] > t) \quad \leq \quad \text{small value}$$

- Example. Two random variables

$$X \sim \mathcal{N}(0,1), \quad Y \sim \text{Unif}([-\sqrt{3}, \sqrt{3}])$$

have …

- Same mean and variance

- Very different tail probabilities

Standard inequalities

- **Markov.** For a nonnegative r.v. X, we have

$$P(X \geq a) \leq \frac{\mathbb{E}[X]}{a}, \qquad \forall a > 0$$

Standard inequalities

- **Markov.** For a nonnegative r.v. X, we have

$$P(X \geq a) \leq \frac{\mathbb{E}[X]}{a}, \qquad \forall a > 0$$

- **Chebyshev.** For a r.v. X with finite variance, we have

$$P(|X - \mathbf{E}[X]| \geq a) \leq \frac{\mathrm{Var}[X]}{a^2}, \qquad \forall a > 0$$

  - A simple application of Markov's inequality

Standard inequalities

- **Chernoff.** We have

$$P(X \geq a) \leq \mathbb{E}[\exp(t \cdot X)] \cdot \exp(-t \cdot a) \qquad \forall a \in \mathbb{R}, t > 0$$

  - Another simple application of Markov's inequality
  - Homework. Revisit moment-generating functions & cumulant-generating functions.
  - Note (advanced). Hoeffding's inequality
    McDiarmid's inequality
    Bernstein's inequality

Further readings

- Bruce Hajek, "Random Processes for Engineers"
  - https://hajek.ece.illinois.edu/ECE534Notes.html

Next up

- Finally some machine learning stuff!
    - Starting from linear models

Cheers