3. Recap: Matrix Calculus & Basic Probability

EECE454 Introduction to
Machine Learning Systems

2023 Fall, Jaeho Lee

New Ref. — Deep Learning

- Very cool book by Francois Fleuret: “The Little Book of Deep Learning” https://fleuret.org/francois/lbdl.html
  - Strongly recommended— Phone-sized PDFs!

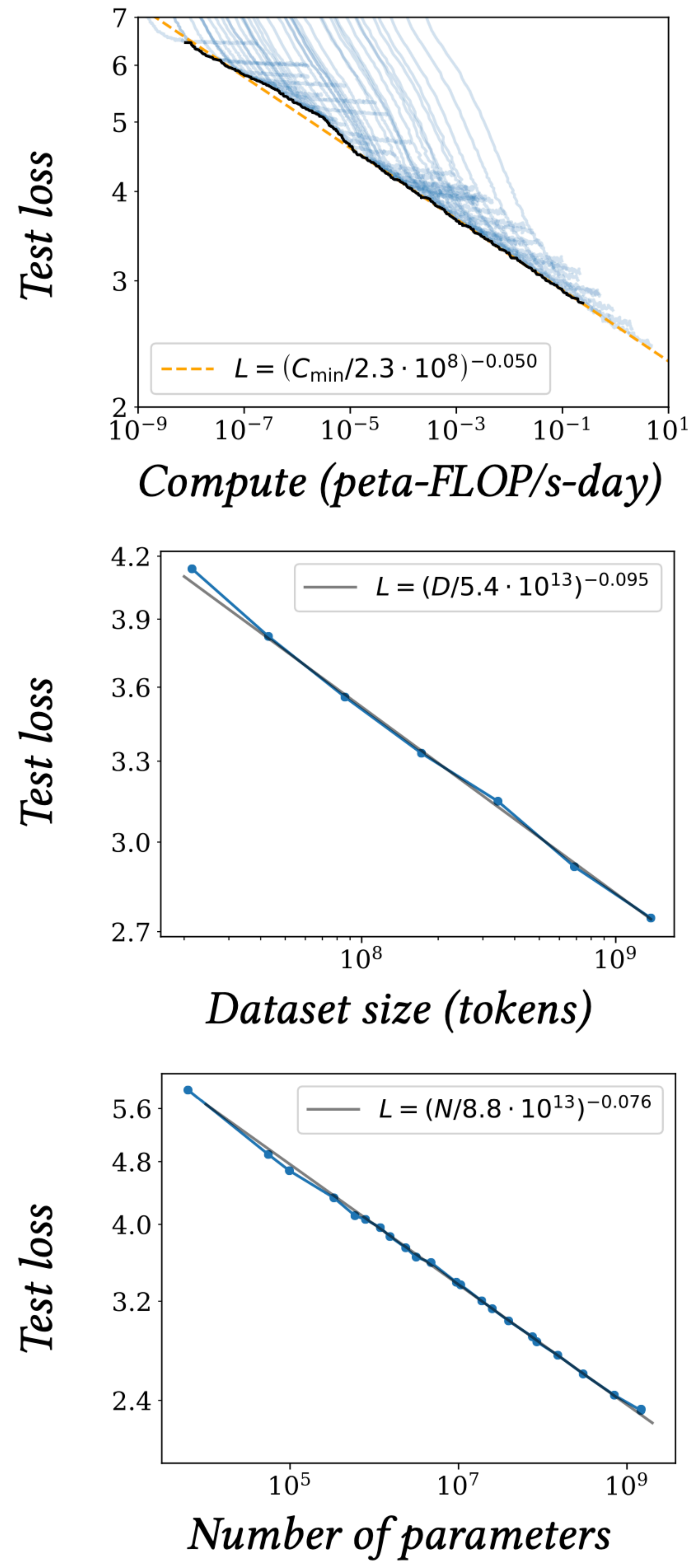


Figure 3.6: *Test loss of a language model vs. the amount of computation in petaflop/s-day, the dataset size in tokens, that is fragments of words, and the model size in parameters [Kaplan et al., 2020].*

52

Last Class

- Vectors, Matrices
- Multiplications (V-V, M-V, M-M)
- Vector norms # not covered matrix norms yet
- Column/Row/Null Space
- Eigenvalues, Eigenvectors
- Eigendecomposition, SVD
- **Today.** Gram-Schmidt, Matrix Calculus, Probability.

Gram-Schmidt (QR decomposition)

QR Decomposition

- Compact decomposition of matrix $\mathbf{A} \in \mathbb{R}^{m \times n}$ (with $m \geq n$)

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

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

$$\mathbf{A} = \mathbf{Q}\mathbf{R}$$

- This is identical to saying that

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

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

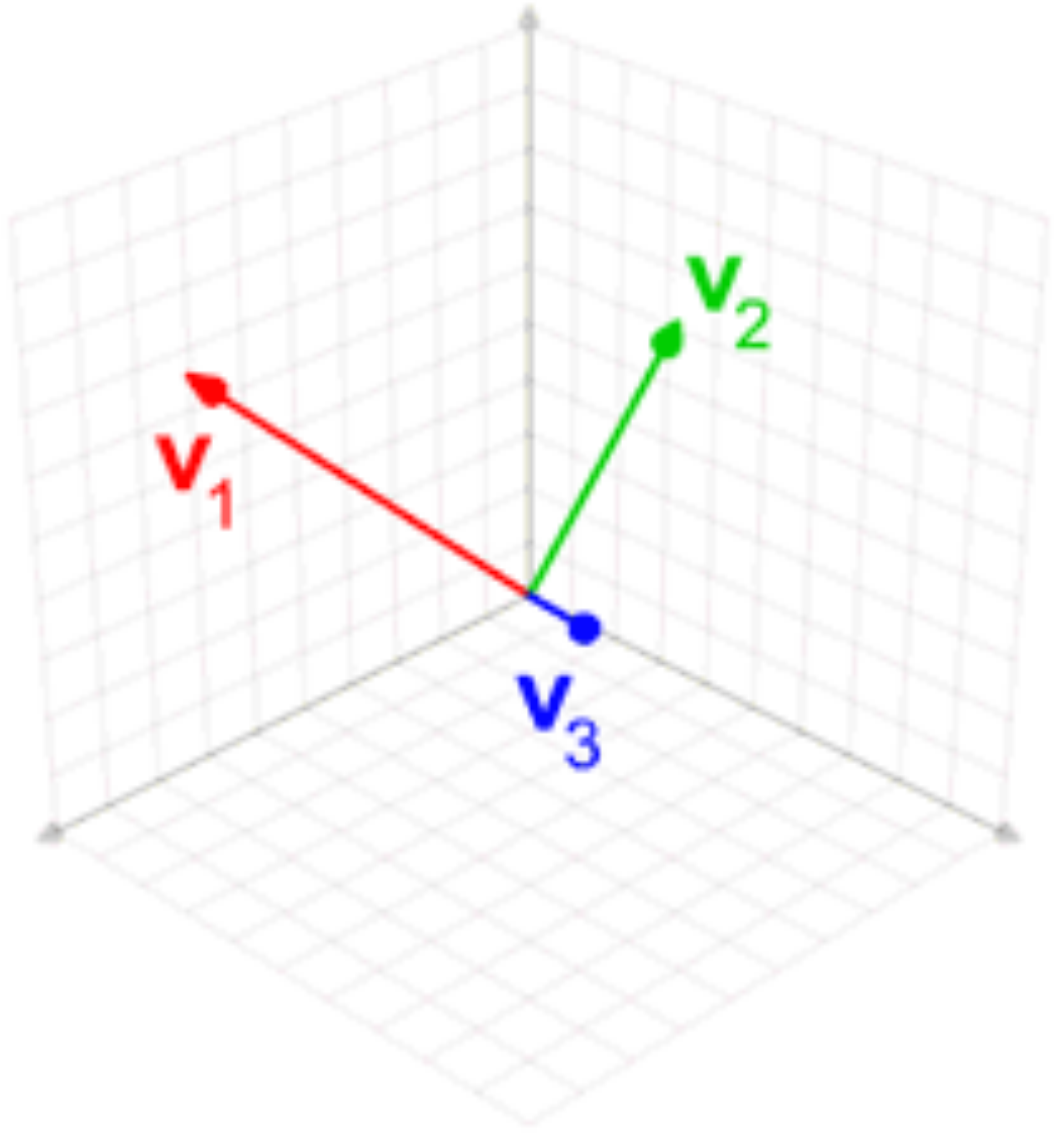
$$\mathbf{a}_2 = \mathbf{e}_1 r_{12} + \mathbf{e}_2 r_{22}$$

$$\cdots$$

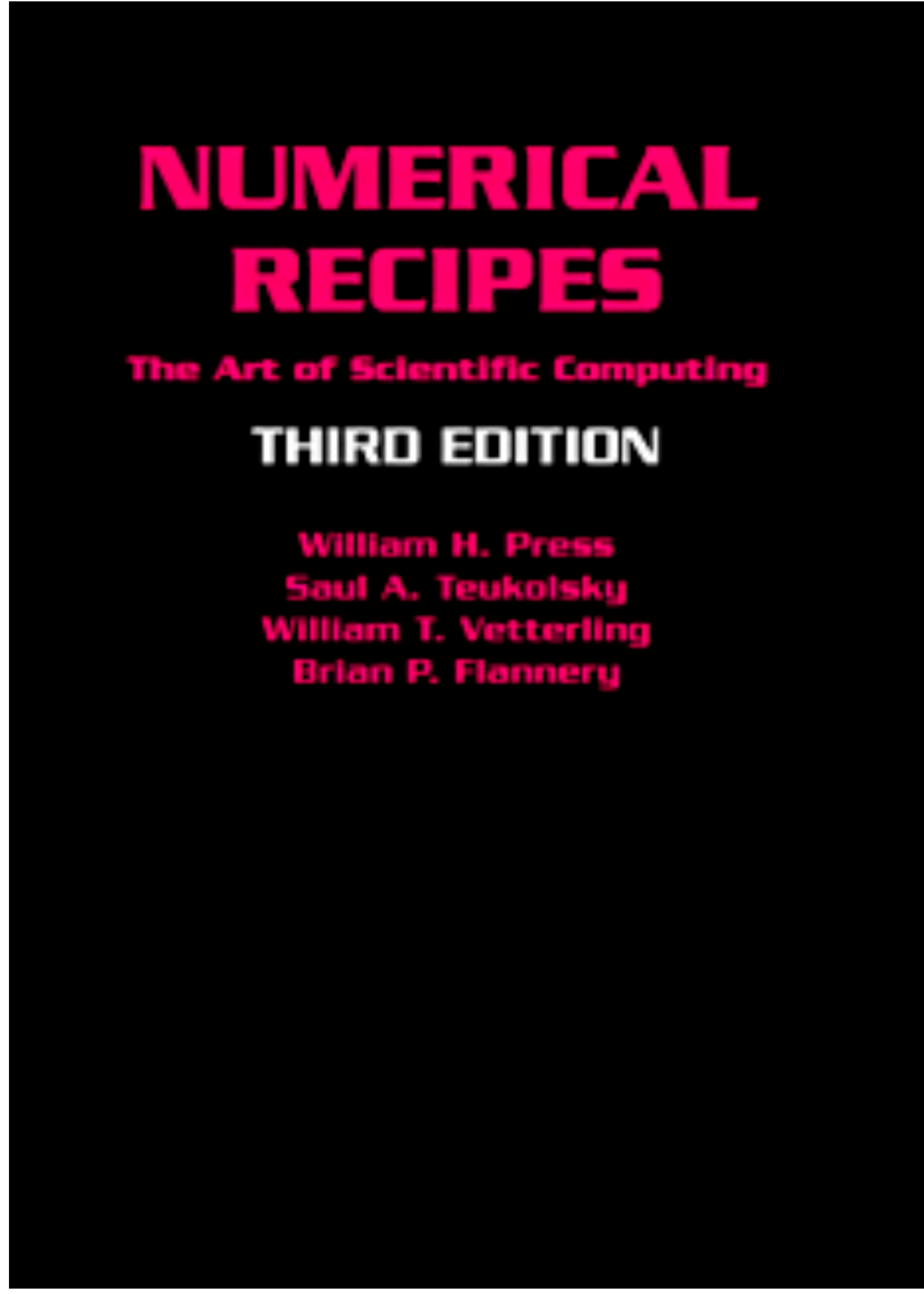
Procedure

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

- This can be done via Gram-Schmidt process
  - Make $\mathbf{e}_1$ by normlizing $\mathbf{a}_1$.
  - Make $\mathbf{e}_2$ by normalizing the remainder $\mathbf{a}_2 - \langle \mathbf{a}_2, \mathbf{e}_1 \rangle \cdot \mathbf{e}_1$
  - repeat ...



Matrix decompositions…



- There are many!
  - SVD, QR, Cholesky, LU, …
- These tend to have different purposes:
  - People use QR for solving $\mathbf{Ax} = \mathbf{y}$.
  - Different strengths / weaknesses (e.g., numerical stability)
    - See section 2 of "Numerical Recipes" for more info.

Matrix Calculus

Why Matrix Calculus?

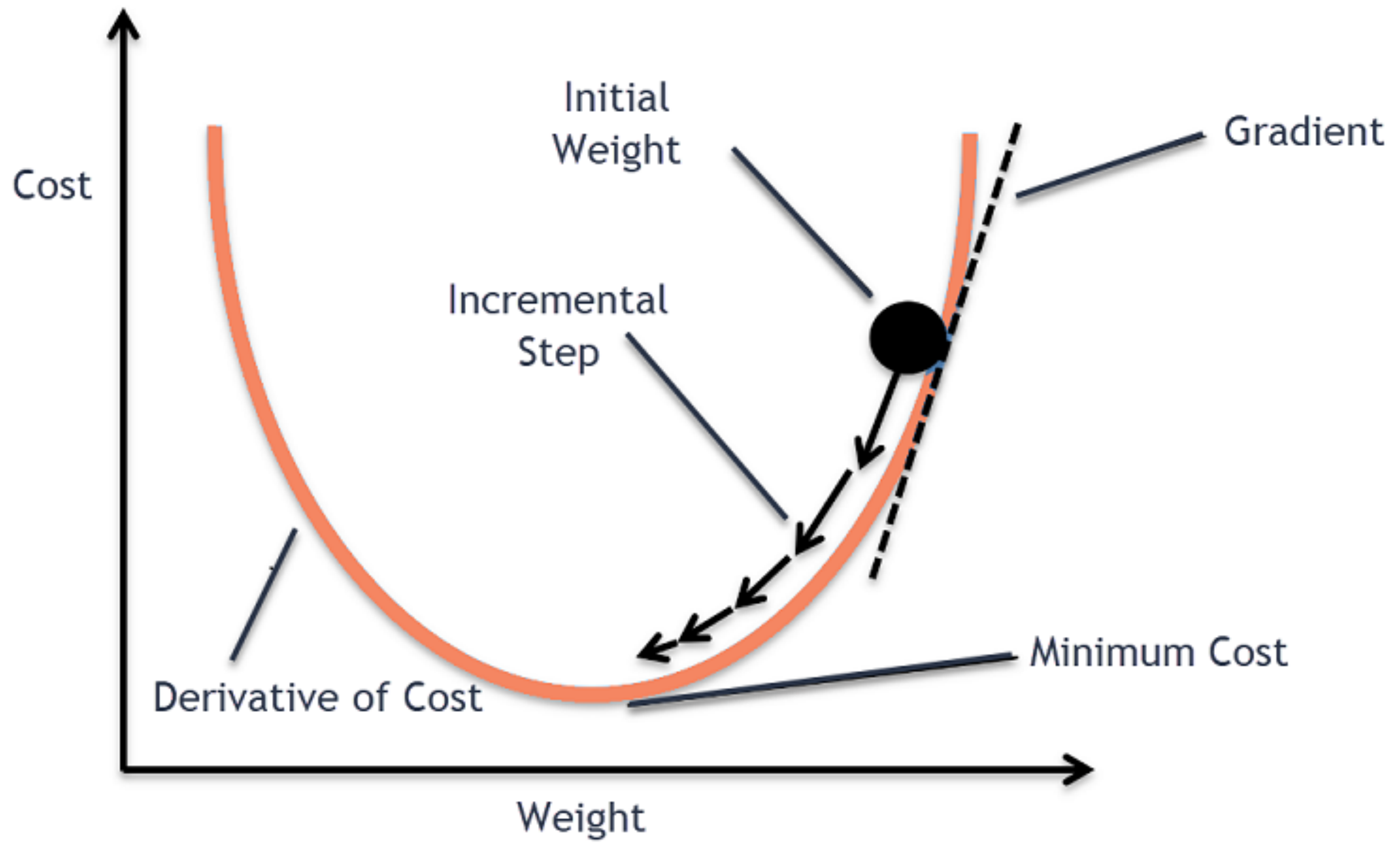
- Univariate Calculus, to find an optimal parameter.
- **Goal.** Find a good "model" $c \in \mathbb{R}$ for a single datum.

  That is, we want to minimize

$$(y_0 - cx_0)^2$$

  - How to solve?

    (either explicit solution or iterative method)



Why Matrix Calculus?

- Vector/Matrix Calculus, to find optimal parameters.
- **Goal.** Find a good "model" $\mathbf{W} \in \mathbb{R}^{m \times n}$ for high-dim data, with $\mathbf{x}_0 \in \mathbb{R}^n$, $\mathbf{y}_0 \in \mathbb{R}^m$. That is, we minimize

$$\|\mathbf{y}_0 - \mathbf{W}\mathbf{x}_0\|_2^2$$

  - How to solve?

(Later, we see even more complicated cases, where we use "gradient descent")

Gradients

- For a scalar variable x, differentiating a...

scalar function $y \in \mathbb{R}$: $$\frac{\partial y}{\partial x}$$

vector function $\mathbf{y} \in \mathbb{R}^m$: $$\begin{bmatrix} \frac{\partial y_1}{\partial x} & \cdots & \frac{\partial y_m}{\partial x} \end{bmatrix}^\top$$

matrix function $\mathbf{Y} \in \mathbb{R}^{m \times n}$: $$\begin{bmatrix} \frac{\partial y_{11}}{\partial x} & \cdots & \frac{\partial y_{1n}}{\partial x} \\ & \cdots & \\ \frac{\partial y_{m1}}{\partial x} & \cdots & \frac{\partial y_{mn}}{\partial x} \end{bmatrix}$$

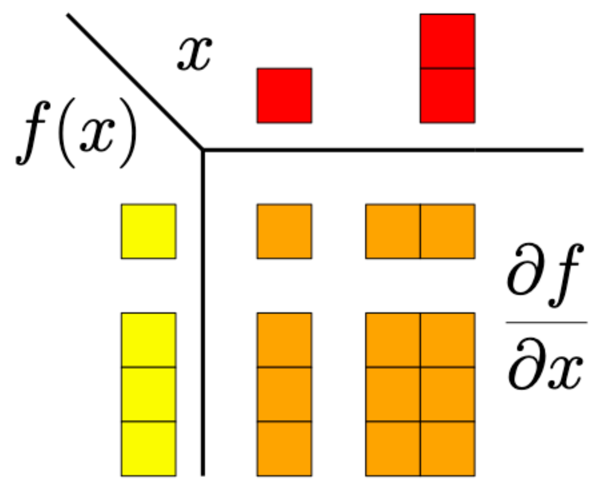
Gradients

- For a vector $\mathbf{x} \in \mathbb{R}^n$, differentiating a...

  scalar function $y \in \mathbb{R}$: $\begin{bmatrix} \dfrac{\partial y}{\partial x_1} & \cdots & \dfrac{\partial y}{\partial x_n} \end{bmatrix}$ (note: direction)

  vector function $\mathbf{y} \in \mathbb{R}^m$: $$\begin{bmatrix} \dfrac{\partial y_1}{\partial x_1} & \cdots & \dfrac{\partial y_1}{\partial x_n} \\ & \cdots & \\ \dfrac{\partial y_m}{\partial x_1} & \cdots & \dfrac{\partial y_m}{\partial x_n} \end{bmatrix}$$

Figure 5.2 Dimensionality of (partial) derivatives.



Gradients

- For a matrix $\mathbf{x} \in \mathbb{R}^{m \times n}$, differentiating…

scalar $y \in \mathbb{R}$:

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

(note: direction)

References for self-study

- MML book Section 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

Probability

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

Probability Space: Tossing a Die

- Consider tossing a die:
  - *Sample space*

    $$\Omega = \{1,2,3,4,5,6\}$$

  - Event space

    $$\mathcal{F} = \left\{ \emptyset, \{1\}, \cdots, \{6\}, \{1,2\}, \cdots, \{5,6\}, \cdots, \{1,2,3,4,5,6\} \right\}$$

  - Probability measure (or probability distribution)

    $$P(\emptyset) = 0, \quad P(\{1\}) = 1/6, \quad \cdots, \quad P(\{1,2,3,4,5,6\}) = 1$$

    (should satisfy certain properties!)

Probability Measure

- Roughly put, axiomatically defined by these properties:
  - $P(\Omega) = 1$
  - $P(A) \geq 0$ for any $A \in \mathcal{F}$
  - $P(A \cup B) = P(A) + P(B)$, whenever $A \cap B = \emptyset$
    - called "additivity," and we expect this to hold for any countable number of mutually exclusive events.

* to generalize to arbitrary space, people use special definitions like σ-algebra, σ-additivity, …

Random Variable

Random Variable

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

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

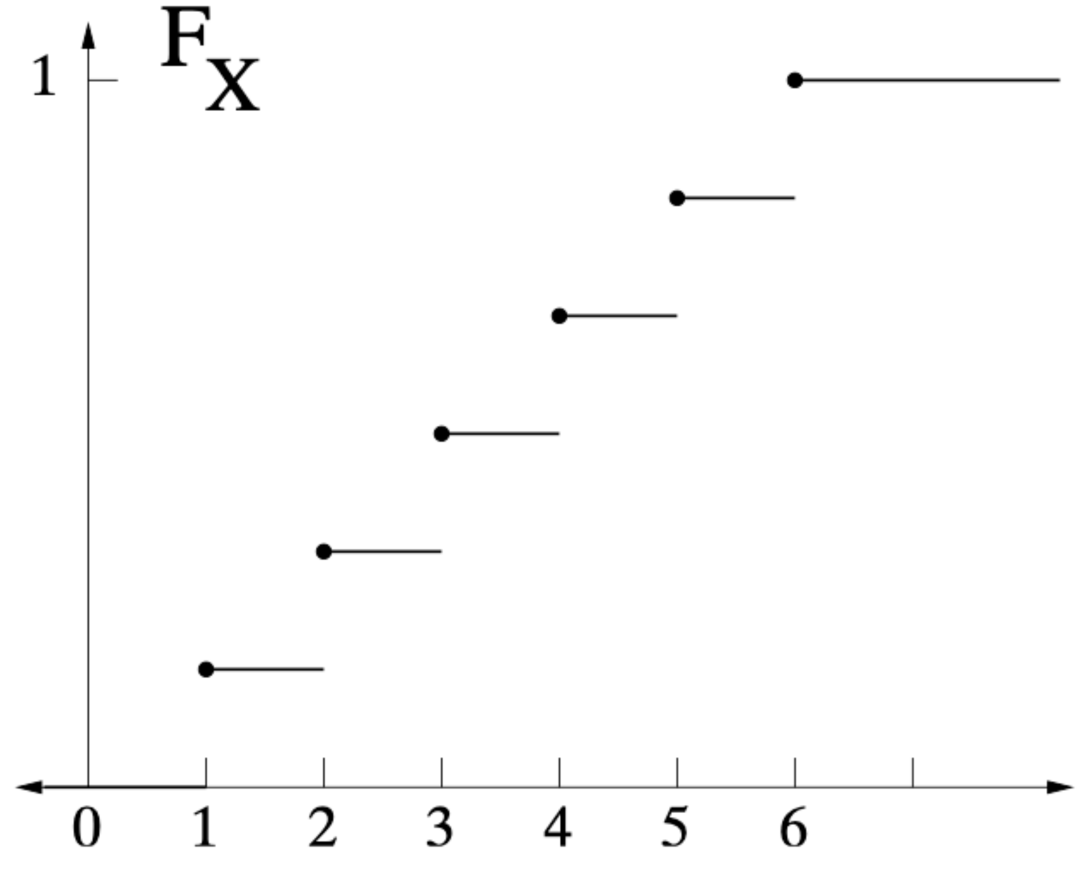
    - Here, we can say that "the probability of $X = 0$ under P" is equal to $P(\{H\})$.
      - We may use the shorthand $P(X = 0)$

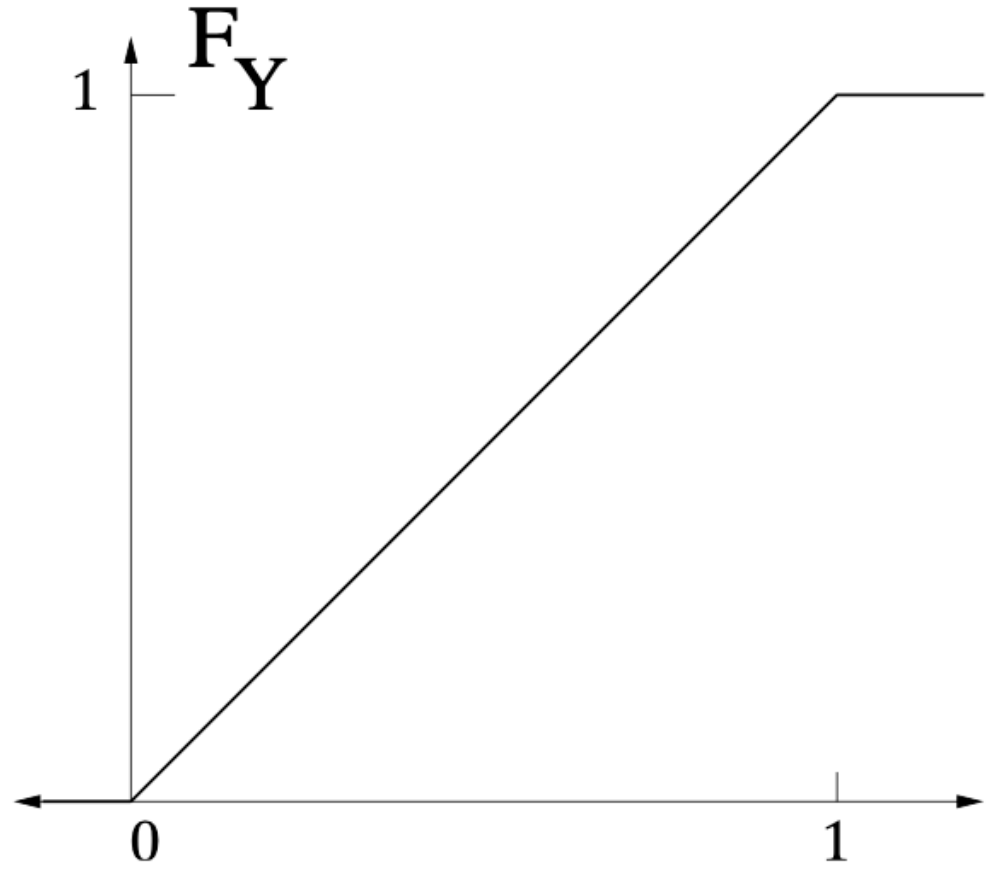
Cumulative Distribution Function (CDF)

- 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 $F_X(x) \leq F_X(y)$



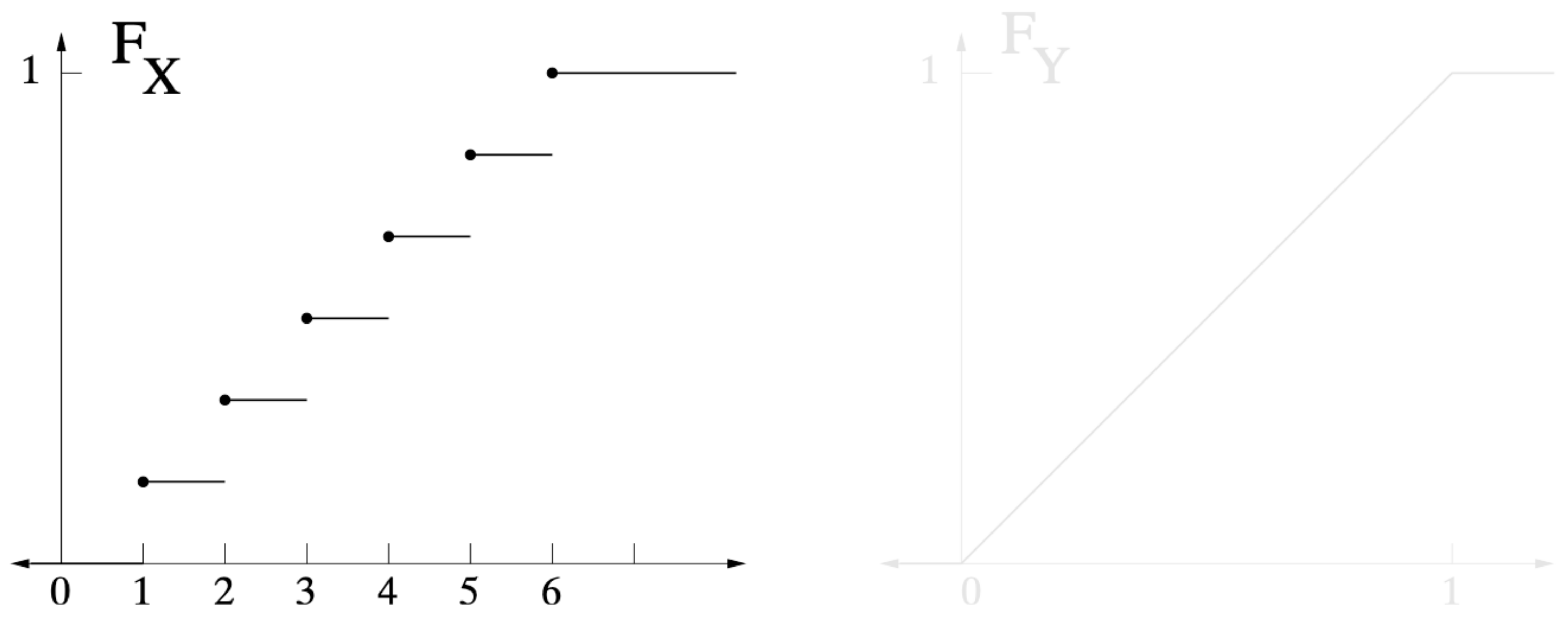


Probability Mass Function (PMF)

- Defined for discrete random variables

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

- *Properties.*
  - $0 \le p_X(x) \le 1$
  - $\sum_{x} p_X(x) = 1$
  - $\sum_{x \in A} p_X(x) = P(X \in A)$

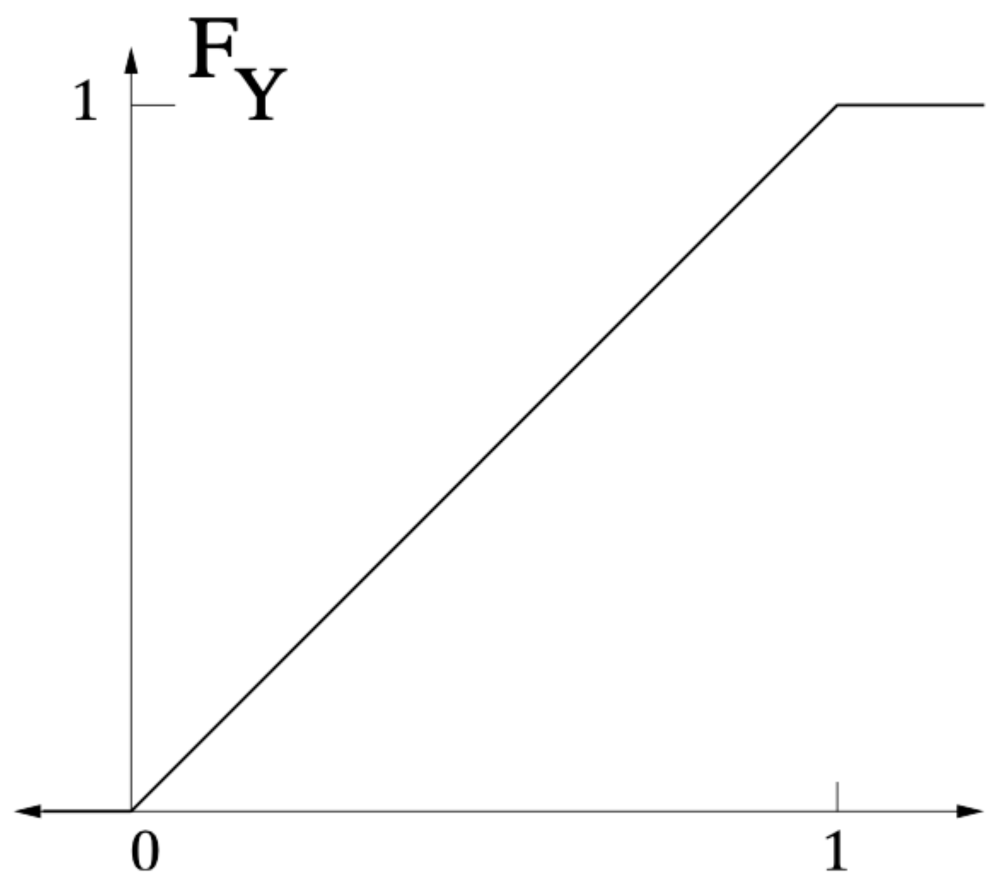


Probability Density Function (PDF)

- Defined for continuous random variables

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

- *Properties.*
  - $0 \le f_X(x)$
  - $\displaystyle\int_{\mathbb{R}} f_X(x)\,\mathrm{d}x = 1$
  - $\displaystyle\int_{A} f_X(x)\,\mathrm{d}x = P(X \in A)$



Probability Density Function (PDF)

- PDF is not really the "probability" itself, but gives you an estimate via:

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

used interchangeably with $f_X(x)$

(This is why $p(x) > 1$ is okay)

Joint distribution

- Defined by some 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 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

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

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

- 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

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

- Bayes' theorem

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

Statistics of RV

Expectation (1st order)

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

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

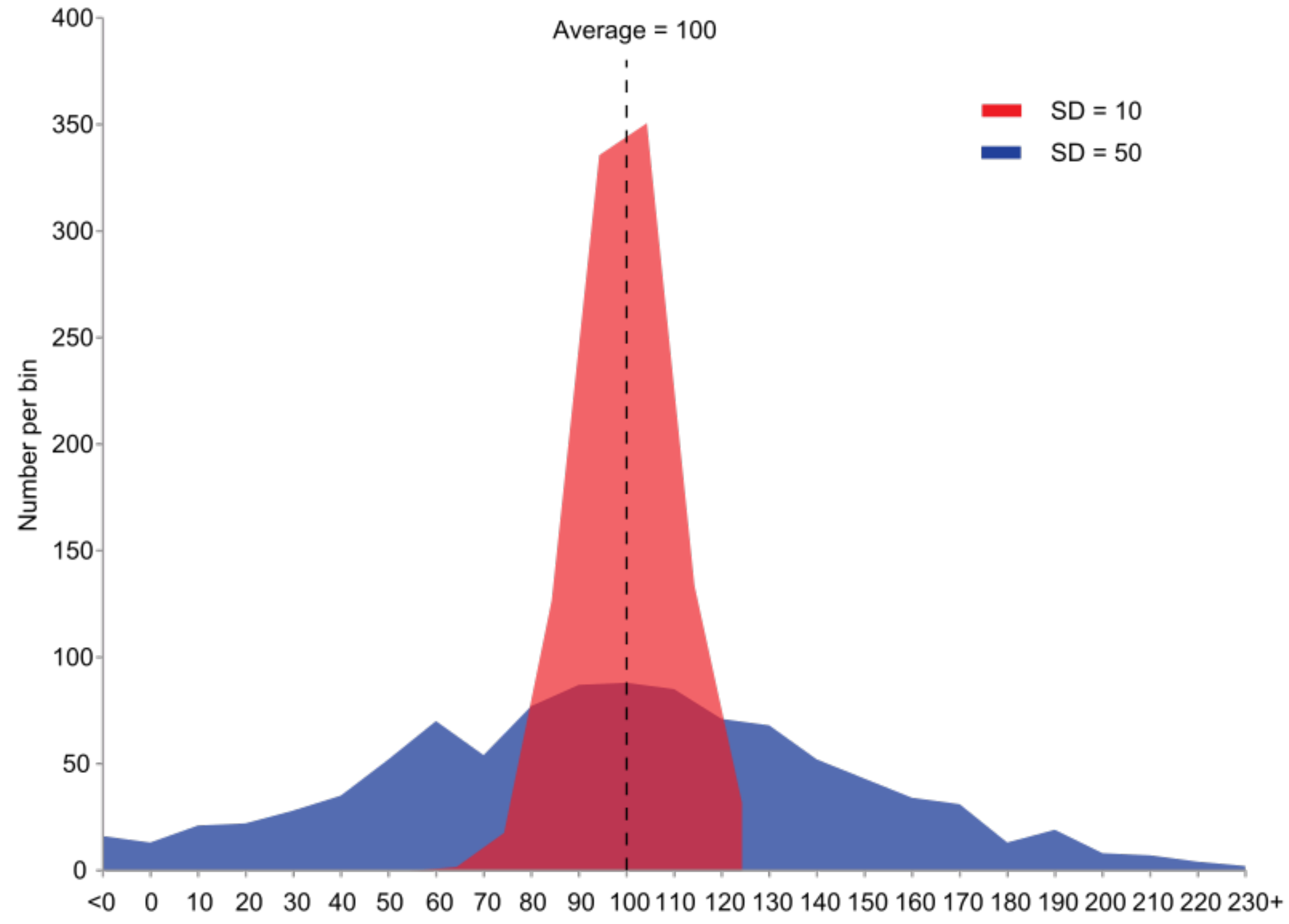
- *Properties.*
  - $\mathbb{E}[a] = a,$ for constant a.
  - $\mathbb{E}[af(X) + bg(X)] = a\mathbb{E}[f(X)] + b\mathbb{E}[g(X)]$ (linearity)

Variance (2nd order)

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

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

- Standard deviation.
- $\sigma_X = \sqrt{\mathrm{Var}(X)}$



Covariance & Correlation

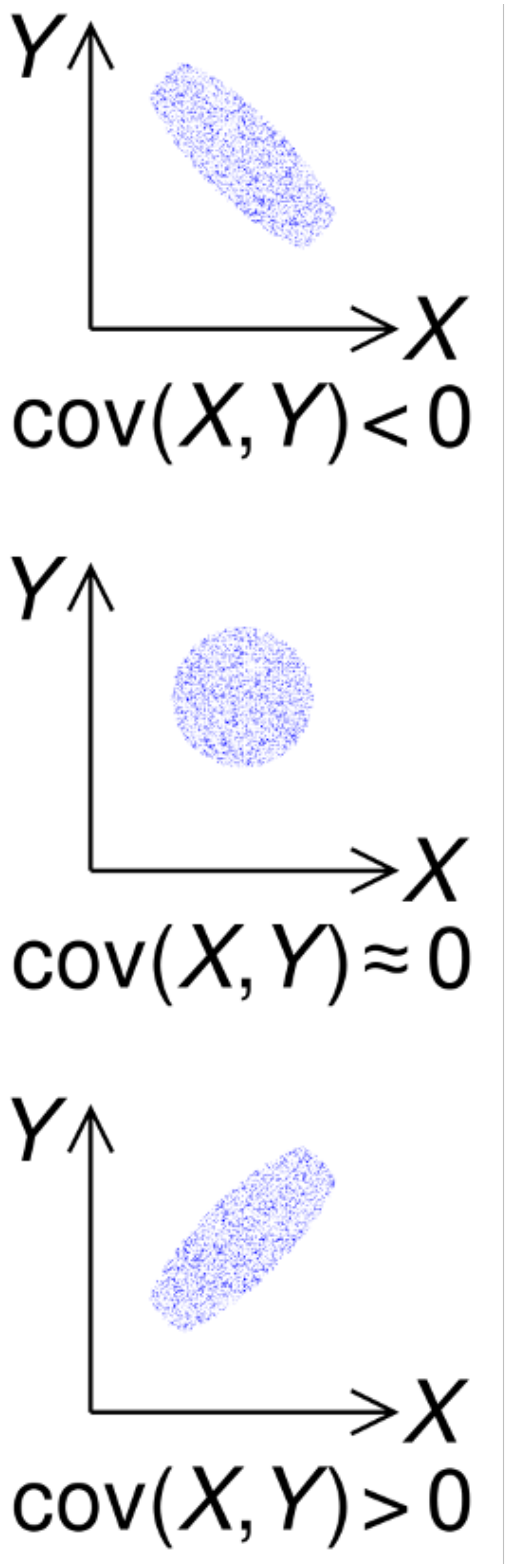
- Measures the joint variability of two RVs.

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

- **(Pearson) Correlation.**

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

(thus lies in $[-1, +1]$)



Independence

Independence

- X and Y are independent, whenever

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

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

Conditional Independence

- X and Y are conditionally independent given Z, whenever

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

(write $X \perp Y \mid Z$)

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

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

Common probability distributions

Bernoulli (a.k.a. coin toss)

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

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

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

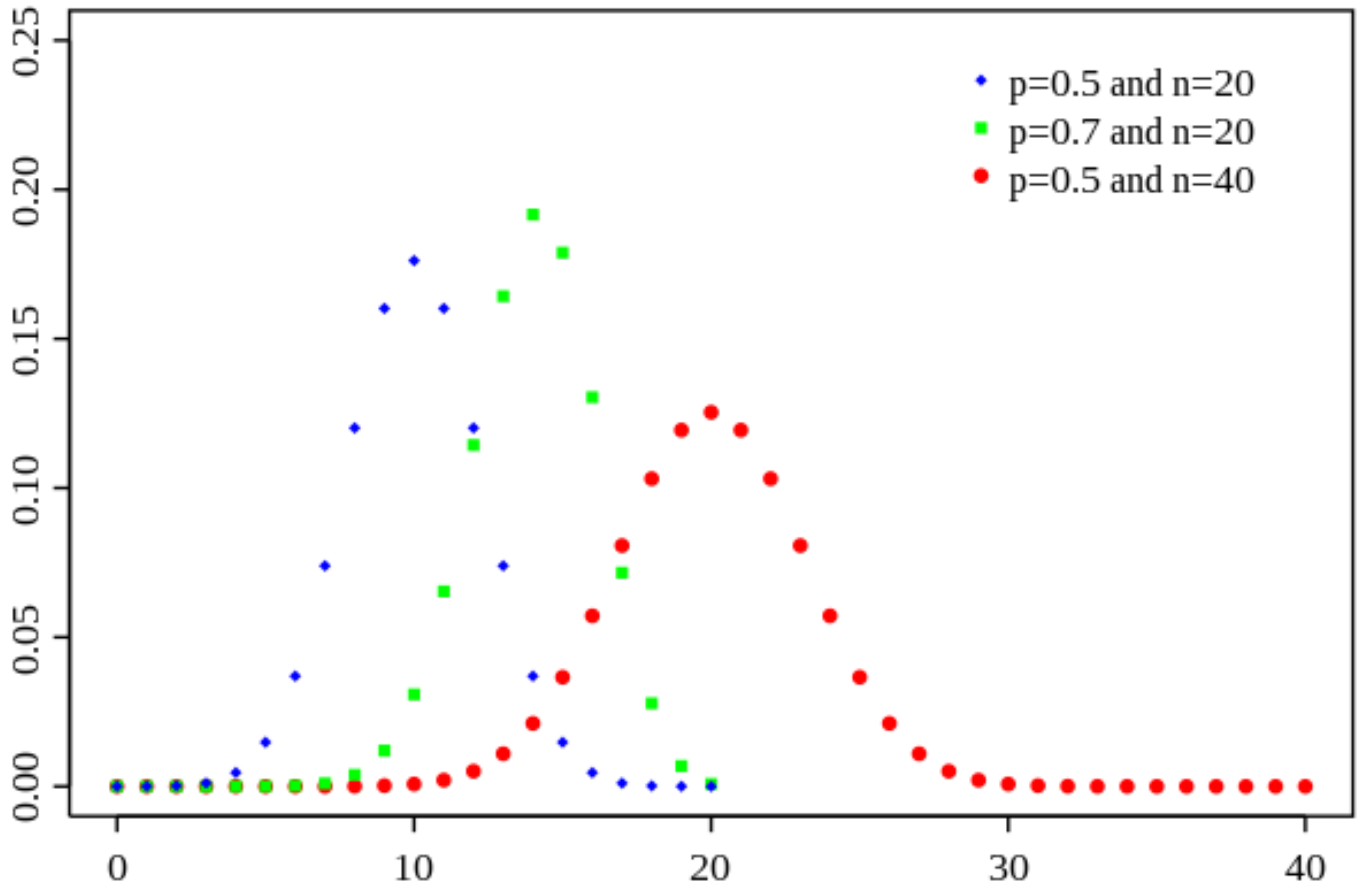
Binomial (a.k.a. many coins)

- $X \sim \mathrm{Bin}(n, p)$ is a discrete random variable with

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

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

(here, $\binom{n}{k} = \dfrac{n!}{k!(n-k)!}$)



Uniform

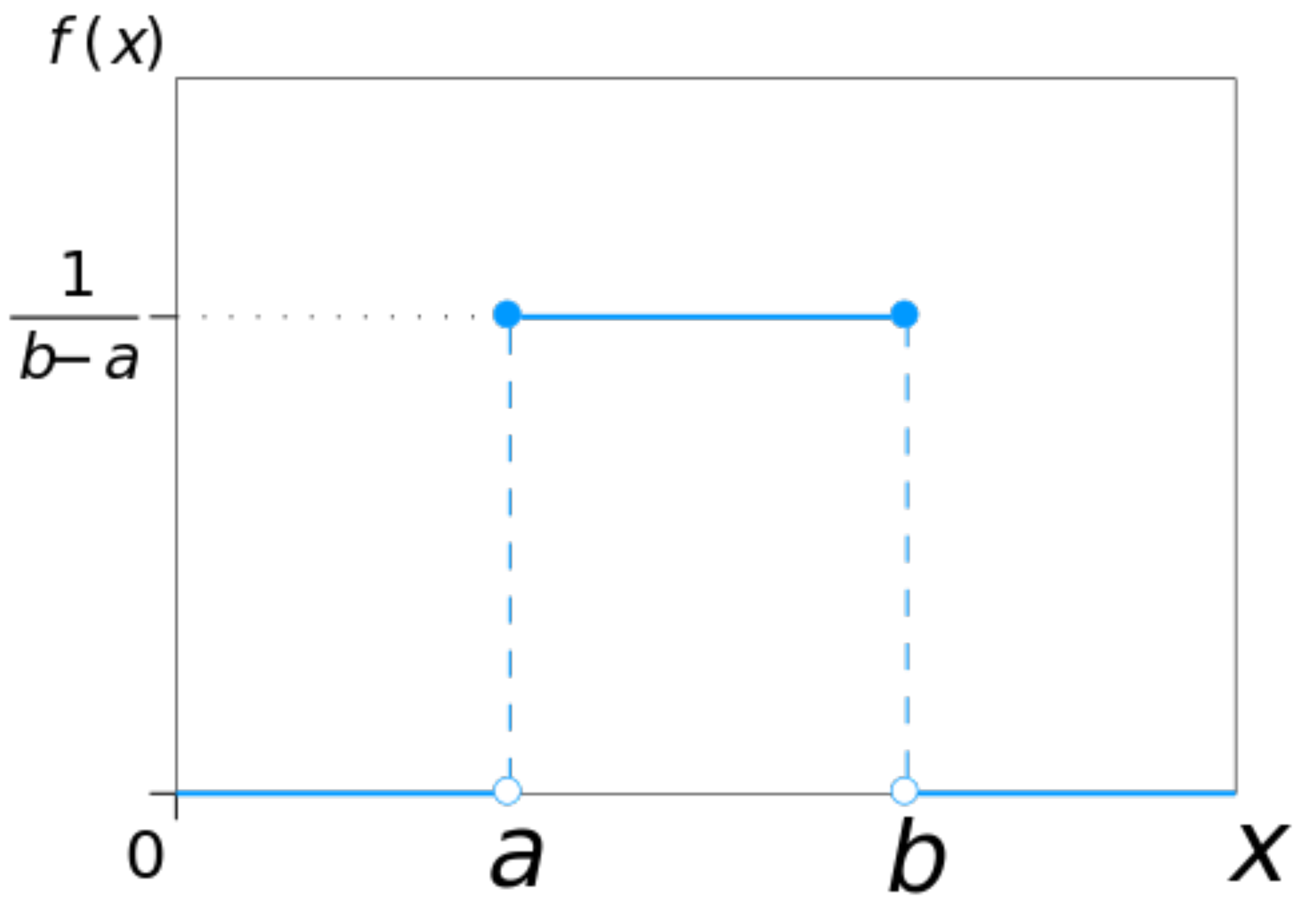
- **Discrete.** $X \sim \mathrm{Unif}(\{1,\ldots,k\})$ is a random variable with

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

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

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

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

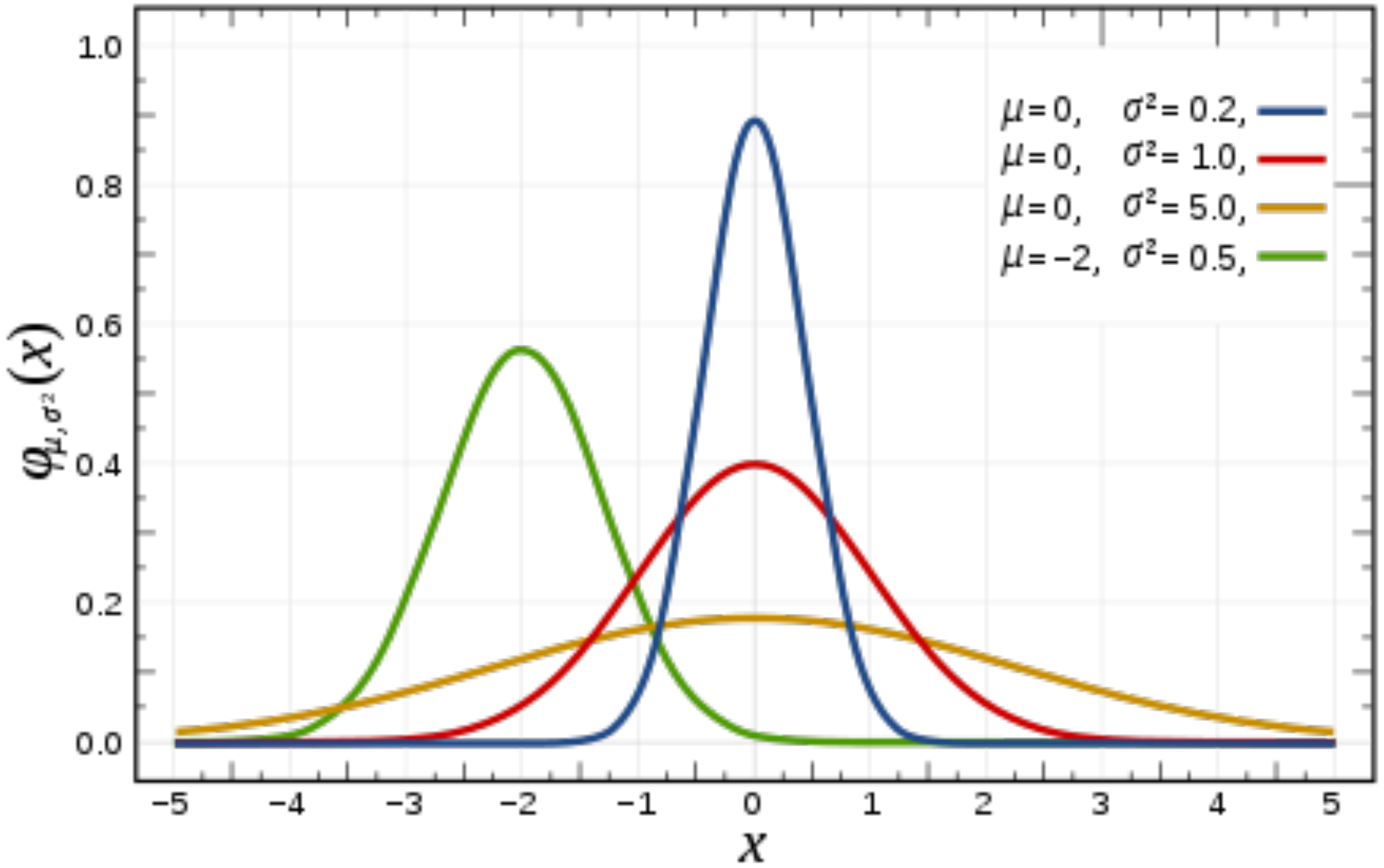


Gaussian (a.k.a. normal)

- $X \sim \mathcal{N}(\mu, \sigma^2)$ is a random variable with

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

- *Importance*. Central limit theorem
- $\mathbb{E}[X] = \mu$
- $\mathrm{Var}[X] = \sigma^2$

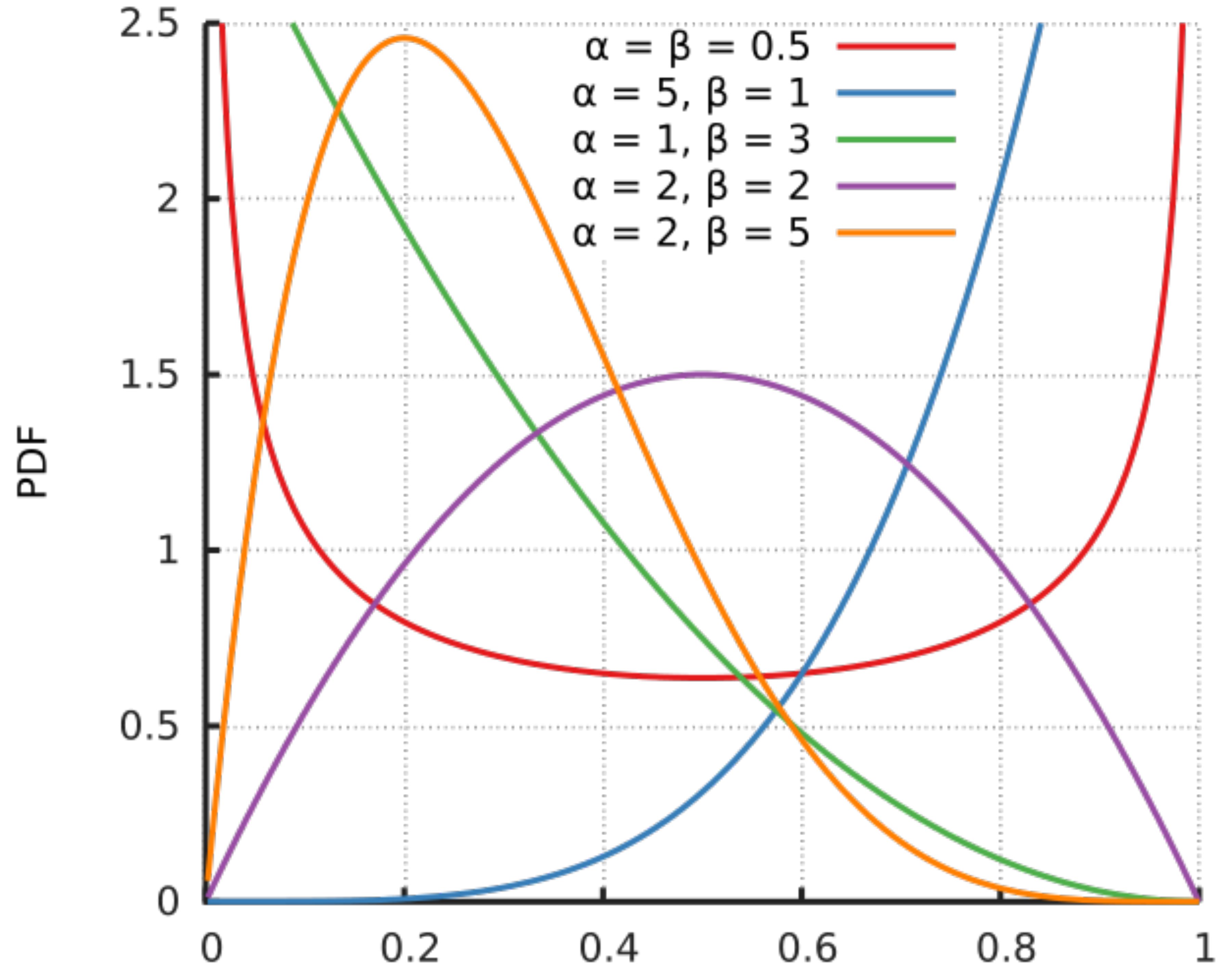


Beta

- $X \sim \text{Beta}(\alpha, \beta)$ is a continuous random variable 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 $\Gamma(\alpha) = (\alpha-1)!$ for integer α)
  - $\mathbb{E}[X] = \dfrac{\alpha}{\alpha+\beta}$
  - $\text{Var}[X] = \dfrac{\alpha\beta}{(\alpha+\beta)^2(\alpha+\beta+1)}$

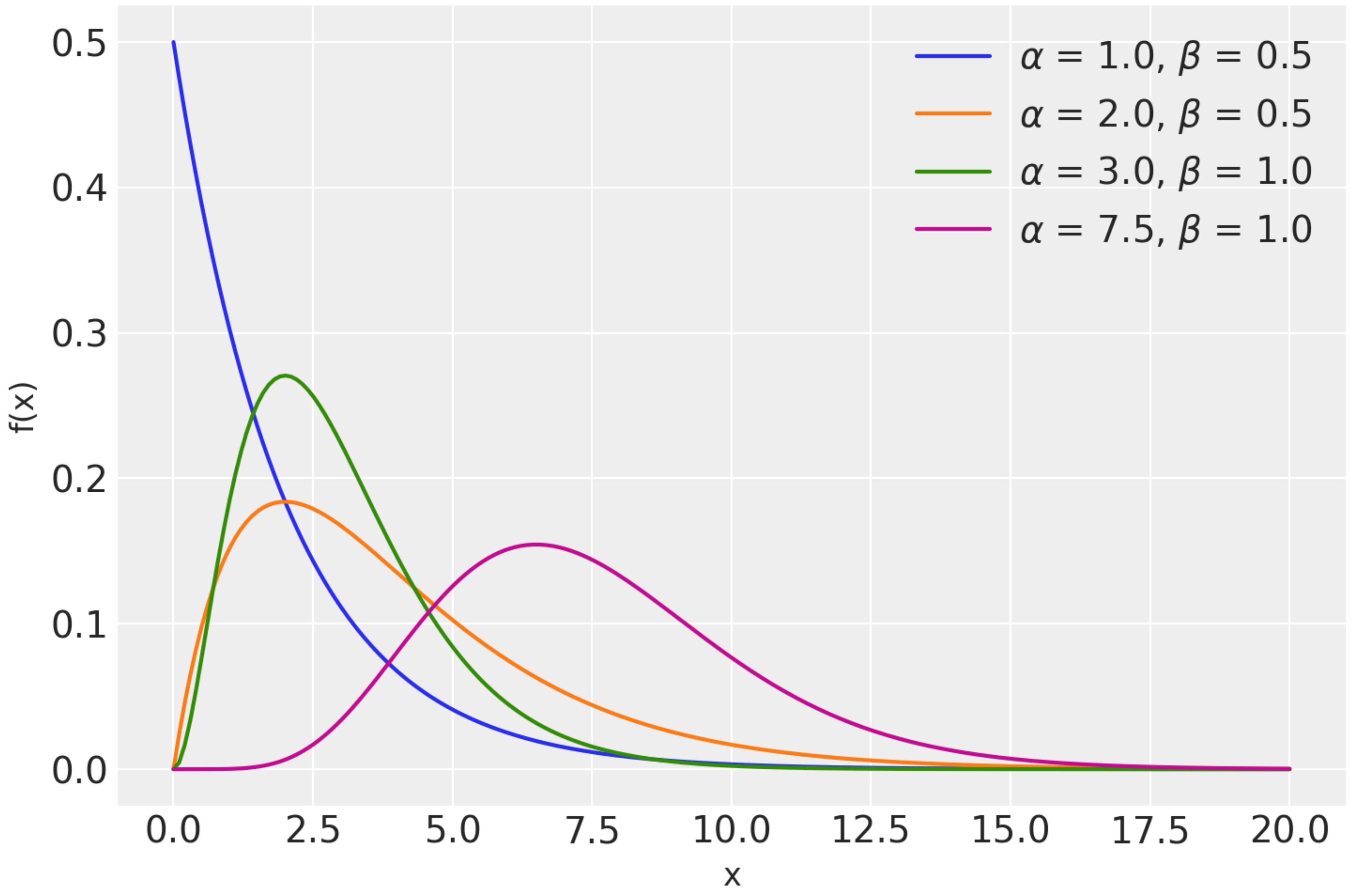


Gamma

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

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

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



Concentration Inequalities

Concentration inequalities

- Gives more fine-grained info. on the "tail behavior" of RVs.

- Typically takes the form:

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

  - *Example.* $X \sim \mathcal{N}(0,1)$ and $Y \sim \mathrm{Unif}([-\sqrt{3}, \sqrt{3}])$ has very different tails, while they have same mean and variances.

Standard Inequalities

- **Markov.** For a *nonnegative* RV X, we have

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

- **Chebyshev.** For a RV X, we have

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

Standard Inequalities

- **Chernoff.**

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

  - Revisit moment-generating functions,
    cumulant-generating functions, ...

Bounded RVs

Theorem 2.2.5 (Hoeffding's inequality, two-sided). *Let $X_1, \ldots, X_N$ be independent symmetric Bernoulli random variables, and $a = (a_1, \ldots, a_N) \in \mathbb{R}^N$. Then, for any $t > 0$, we have*

$$\mathbb{P}\left\{\left|\sum_{i=1}^{N} a_i X_i\right| \geq t\right\} \leq 2\exp\left(-\frac{t^2}{2\|a\|_2^2}\right).$$

Theorem 2.2.6 (Hoeffding's inequality for general bounded random variables). *Let $X_1, \ldots, X_N$ be independent random variables. Assume that $X_i \in [m_i, M_i]$ for every i. Then, for any $t > 0$, we have*

$$\mathbb{P}\left\{\sum_{i=1}^{N}(X_i - \mathbb{E}\,X_i) \geq t\right\} \leq \exp\left(-\frac{2t^2}{\sum_{i=1}^{N}(M_i - m_i)^2}\right).$$

Theorem 2.8.2 (Bernstein's inequality). *Let $X_1, \ldots, X_N$ be independent, mean zero, sub-exponential random variables, and $a = (a_1, \ldots, a_N) \in \mathbb{R}^N$. Then, for every $t \geq 0$, we have*

$$\mathbb{P}\left\{\left|\sum_{i=1}^{N} a_i X_i\right| \geq t\right\} \leq 2\exp\left[-c\min\left(\frac{t^2}{K^2\|a\|_2^2}, \frac{t}{K\|a\|_\infty}\right)\right]$$

where $K = \max_i \|X_i\|_{\psi_1}$.

Further Readings

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

Cheers

- *Next up.* Finally some machine learning.