COMP229: Introduction to Data Science
Lecture 8: Continuous distributions

Olga Anosova, O.Anosova@liverpool.ac.uk
Autumn 2023, Computer Science department
University of Liverpool, United Kingdom

Mid-term test

Mid-term test will take place on Canvas on the 7th November, between 9.00-17.00.

The test consists of 30 multiple choice questions, you have 60 minutes to attempt all questions. For each question, you are expected to select exactly one answer.

The test should be conducted alone, but you are allowed to refer to your notes, although be mindful that if you spend too long looking at your notes you will likely run out of time.

Before starting, make sure you won't be disturbed, and that you have a calculator, pen and paper available. Also make sure before starting that you have a good internet connection and, if doing the test on a laptop, that you are plugged in.

Recap: Random variables

- A **random variable** X in a probability space (Ω, E, P) is a function $X : \Omega \to \mathbb{R}$. A **discrete** random variable takes a finite number of values.
- The **probability mass function** is defined for a discrete variable X by the probabilities $p_i = P(X = v_i)$ with $\sum_{i=1}^{n} p_i = 1$.
- The **cumulative distribution function** of X is $F_X(v) = P(X < v)$.

Distribution function

If X takes all values from an interval, the probability of a single value is 0. Probabilities of $\{X < v\}$ are enough to express all other events, e.g. event $\{X \in [v_1, v_2)\} = \{X < v_2\} \backslash \{X < v_1\}$, hence cdf is enough to define all probabilities for a continuous function.

Properties of the distribution function:

1) $F(v_1) \leqslant F(v_2)$ for any $v_1 < v_2$,

2) $\lim_{v \to -\infty} F(v) = 0$, $\lim_{v \to +\infty} F(v) = 1$,

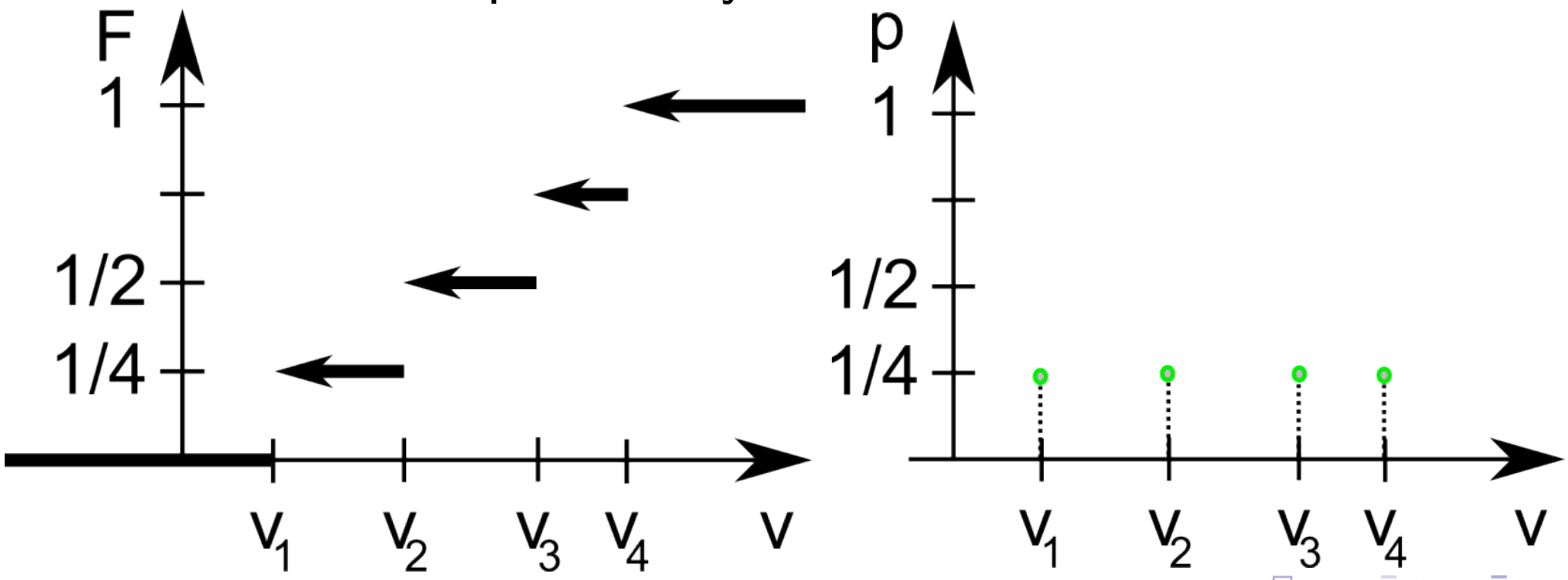
3) $P(v_1 \leqslant X < v_2) = F_X(v_2) - F_X(v_1)$.

Uniform discrete distribution

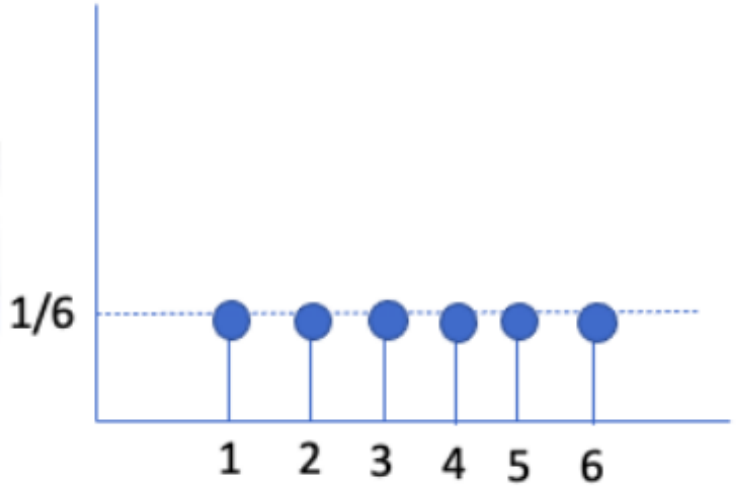
If X is discrete, the distribution is $F_X(v) = \sum_{v_i < v} p_i$.

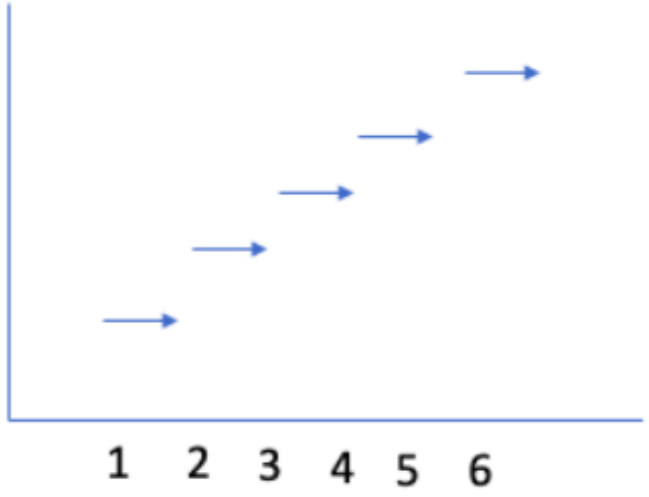
The **uniform discrete** variable X_n has n equally likely outcomes, each value v_i has the probability $p_i = \frac{1}{n}$, $i = 1, \ldots, n$.

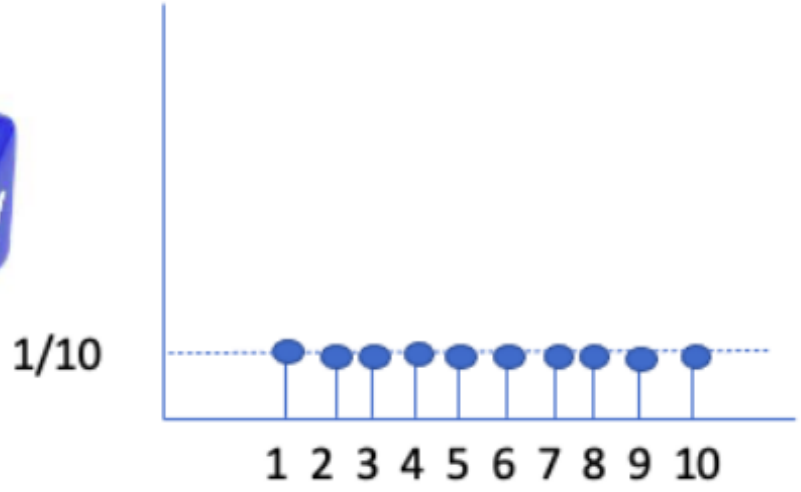
Distribution and probability mass function for $n = 4$:

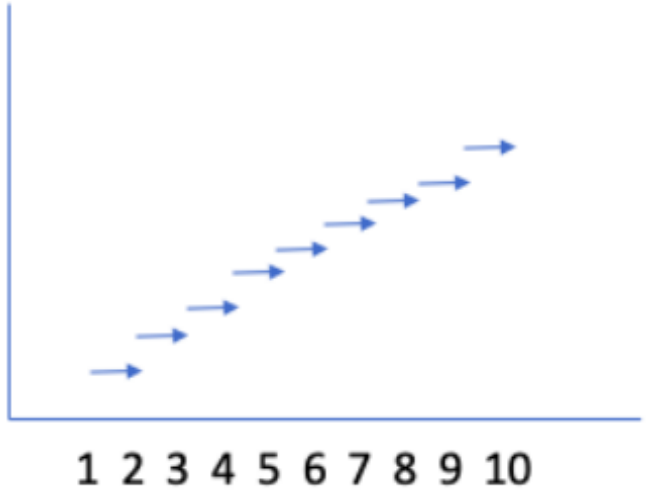


Discrete with $n \longrightarrow \infty$







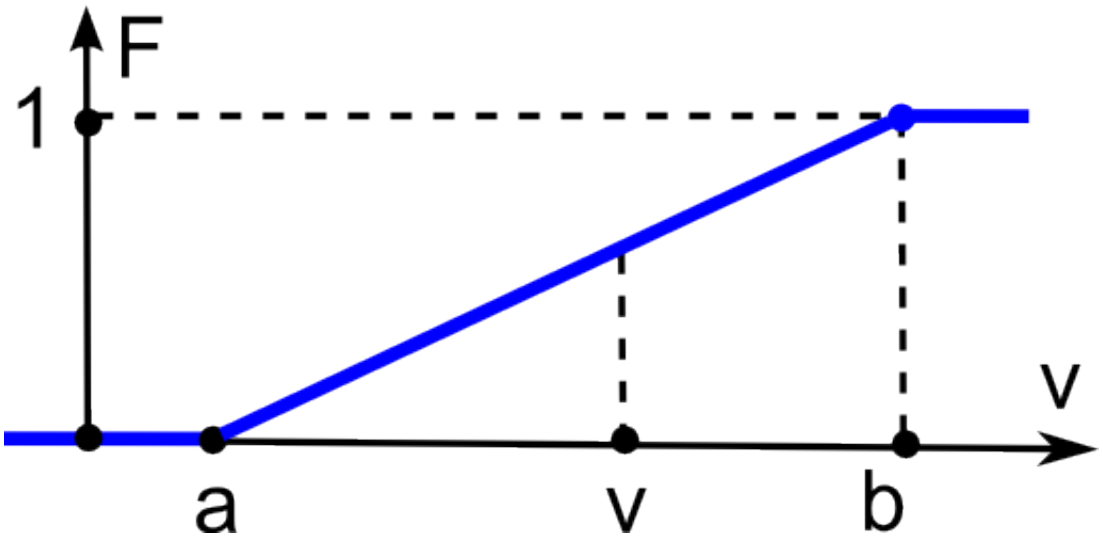


Images courtesy of Dr M.Bright

Uniform continuous variable

Definition 8.1. A random variable X is *uniform continuous* on an interval $[a, b]$ if its distribution function is

$$F_X(v) = \begin{cases} 0 & \text{for } v \leqslant a, \\ \dfrac{v-a}{b-a} & \text{for } v \in [a, b], \\ 1 & \text{for } v \geqslant b. \end{cases}$$



Uniform continuous on $[0, 2]$

Problem 8.2. Let X be the uniform continuous random variable on the interval $[0, 2] \subset \mathbb{R}$. Find the probabilities $P(X < 0)$, $P(X < 2)$, $P(X < 1.35)$, $P(X \leqslant 0)$.

Solution 8.2. $P(X < 0) = 0$, because X takes values within $[0, 1]$.

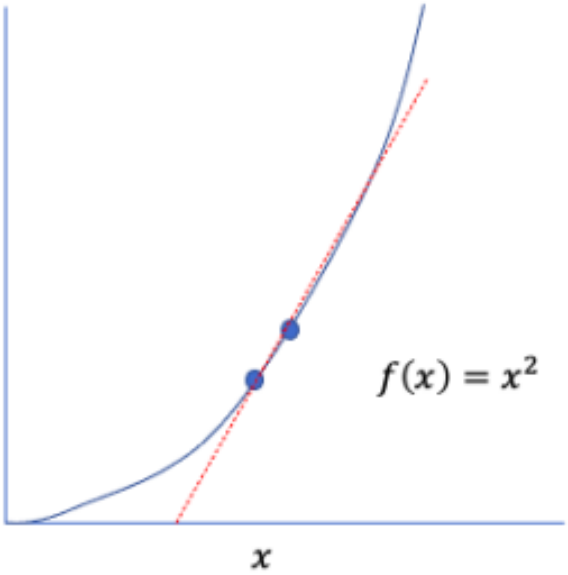
$P(X < 2) = F_X(2) = 1$, because $F_X(v) = \frac{1}{2}v$ for $v \in [0, 2]$, see the graph.

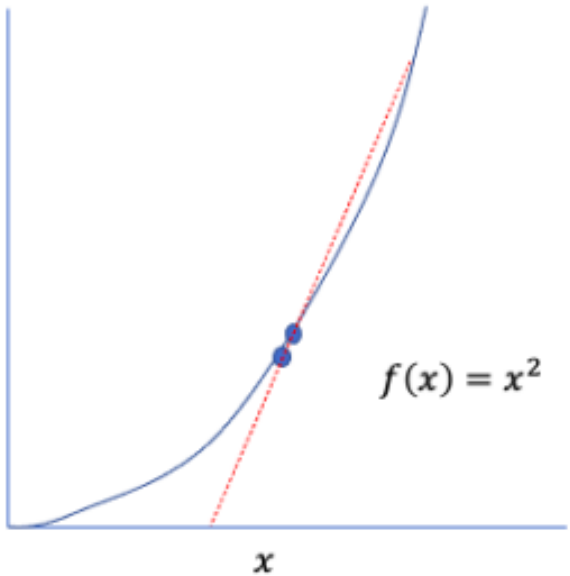
$P(X < 1.35) = F_X(1.35) = 0.5 * 1.35 = 0.675$

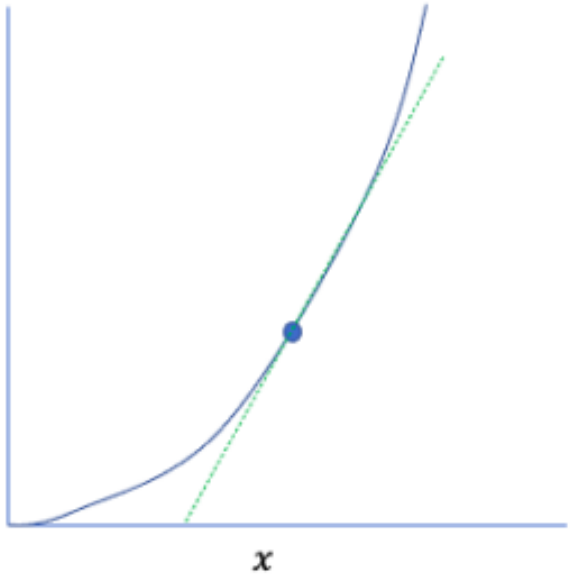
$P(X \leqslant 0) = 0$ as the limit of the probabilities

$P(X < \varepsilon) = F_X(\varepsilon) = \varepsilon$ for a parameter $\varepsilon \to 0$.

Informal calculus: Derivative

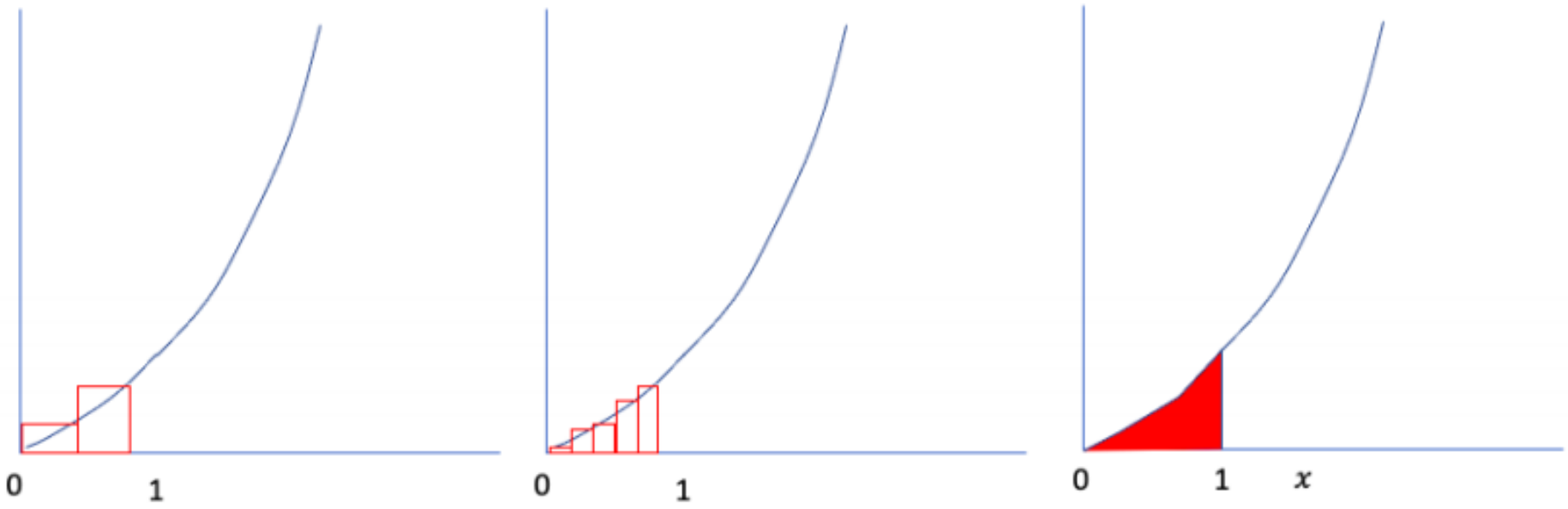






$$\text{Gradient (/) } = \frac{d}{dx} f(x) = f'(x)$$

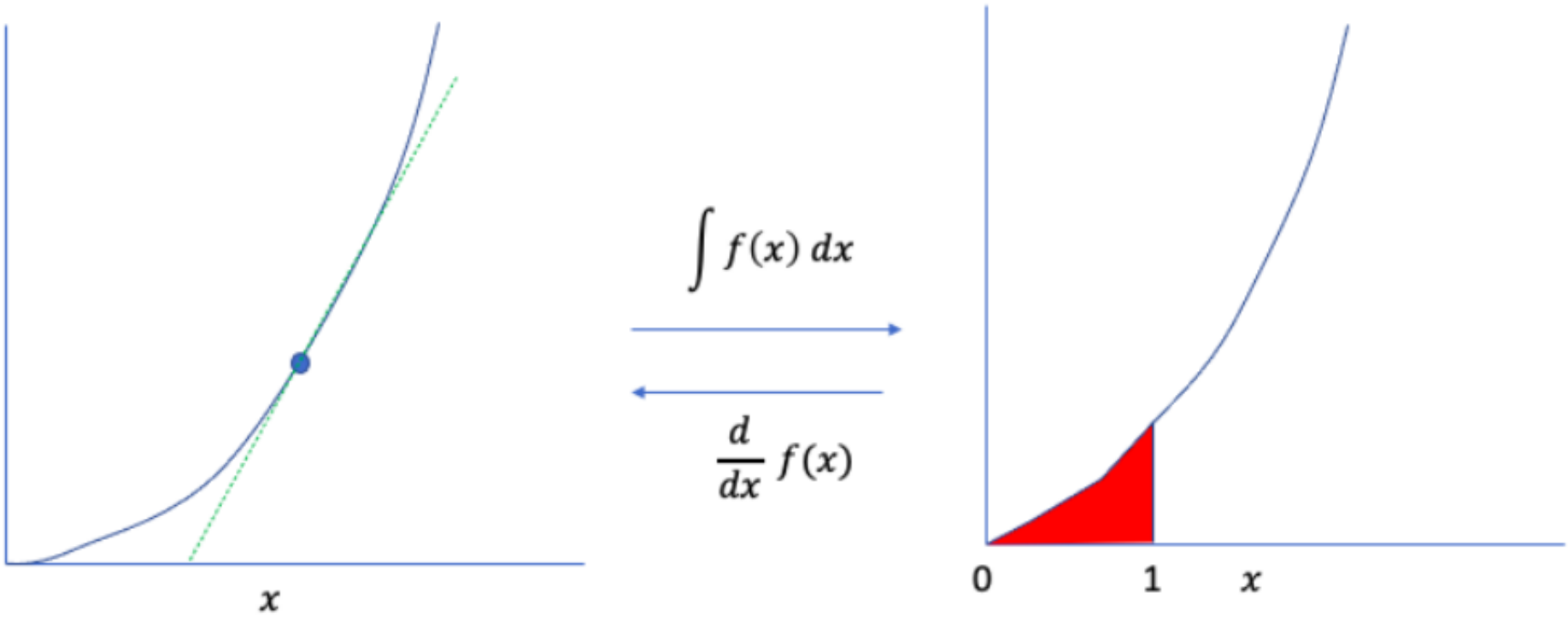
Informal calculus: Integral



$$\text{Area}\left(\ \right) = \int_0^1 f(x)dx$$

The Fundamental Theorem of calculus

Differentiation and integration are 'inverses' of each other

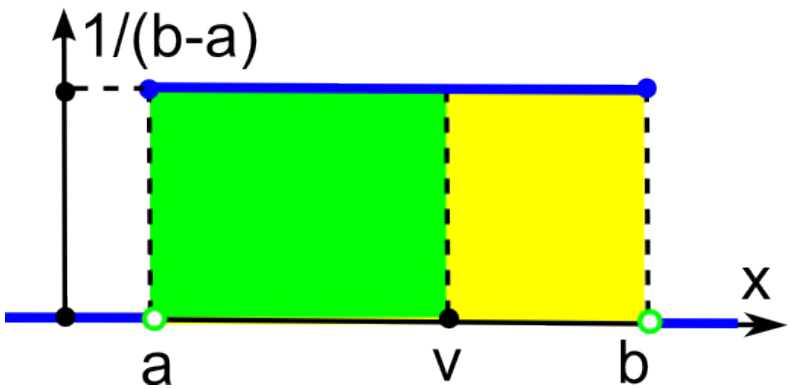


The density of a uniform variable

Definition 8.3.. The **Probability Density Function (PDF)** $f(x)$ of a continuous variable X is the derivative of $F_X(v)$ as a function of v.

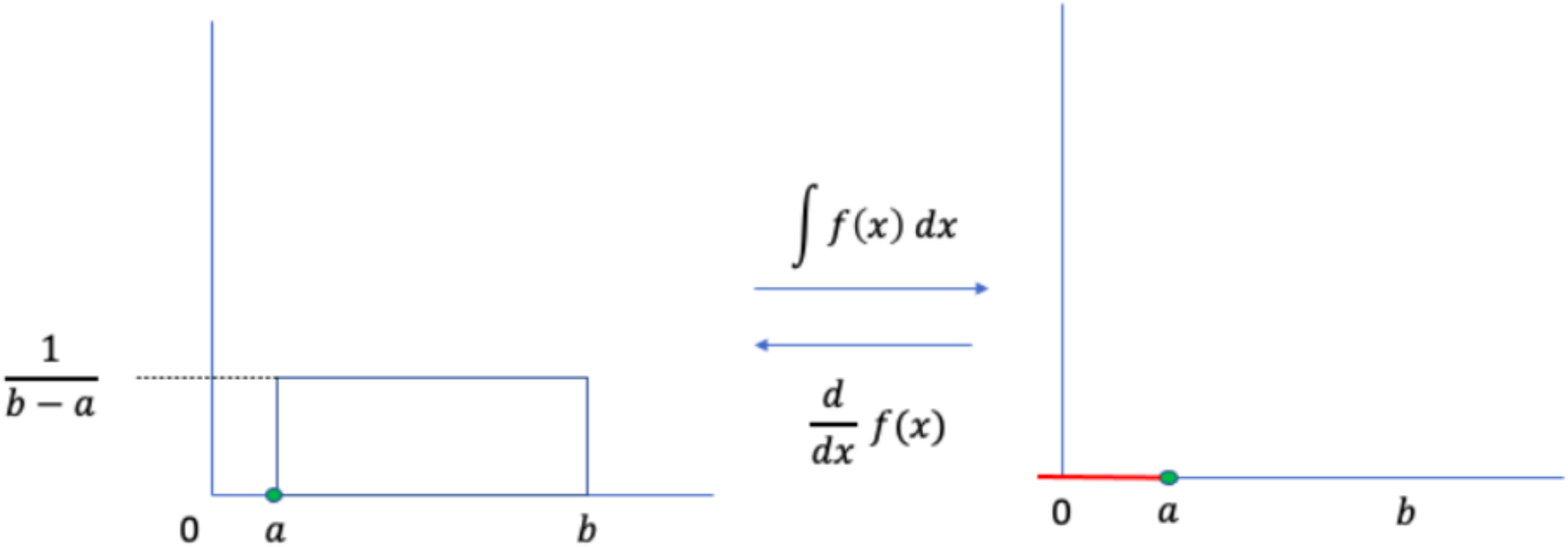
The uniform variable over $[a, b]$ has the *density*

$$f(x) = \begin{cases} \dfrac{1}{b-a} & \text{for } x \in [a, b], \\ 0 & \text{for } x \notin [a, b]. \end{cases}$$

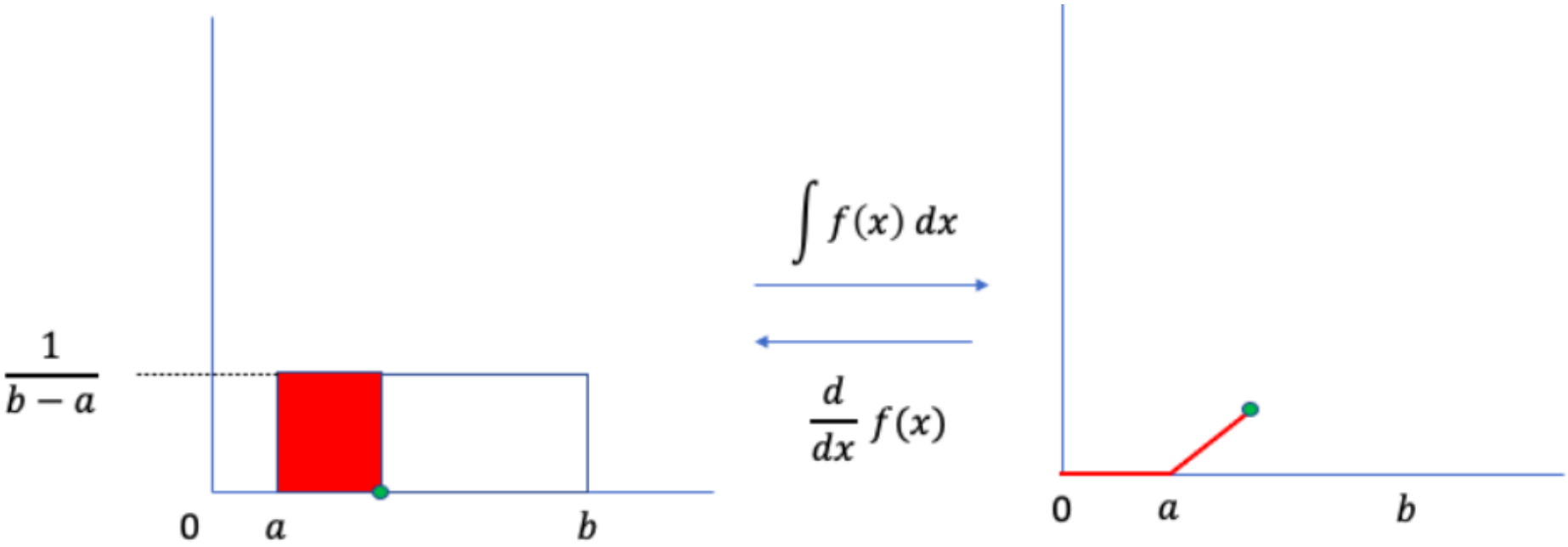


The rectangle with sides $b - a$ and $\dfrac{1}{b-a}$ has the area $\int_{-\infty}^{+\infty} f(x)dx = 1$.

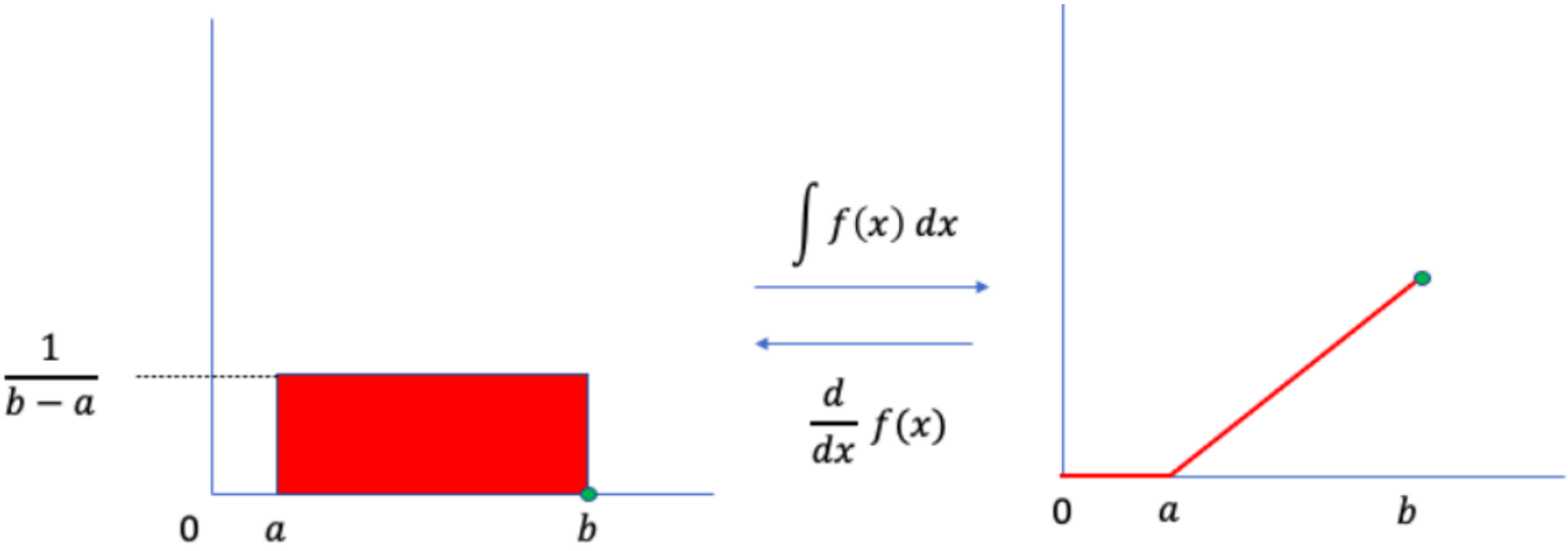
PDF vs CDF



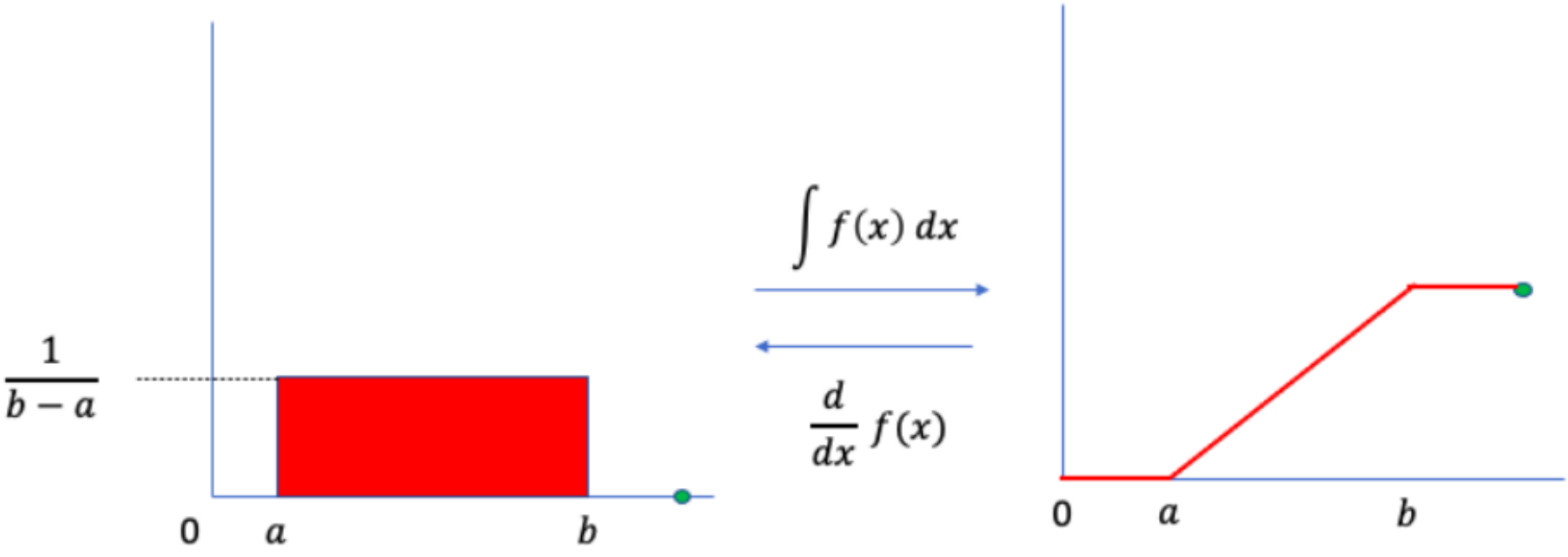
PDF vs CDF



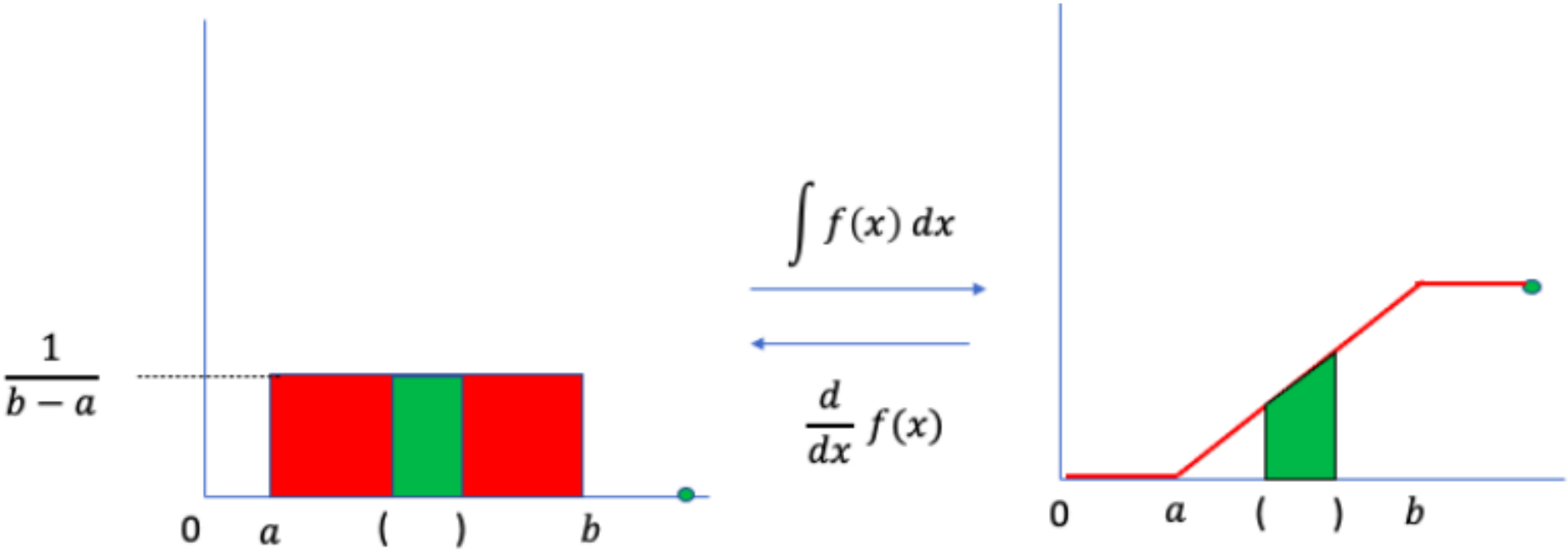
PDF vs CDF



PDF vs CDF



PDF vs CDF



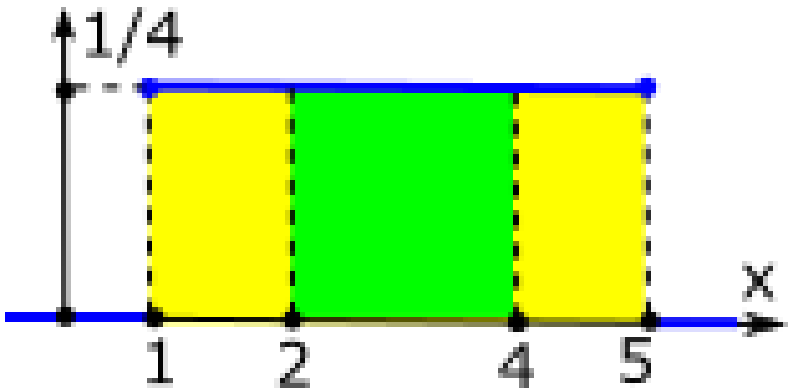
Example of a uniform variable

Problem 8.4. For a uniform variable X on [1,5] find $P(X \in [2,4))$ in two ways: from CDF and PDF.

Solution 8.4. For a uniform continuous X its distribution

$$F_X(v) = \begin{cases} 0, & v \leqslant 1, \\ \dfrac{v-1}{4}, & v \in [1,5], \\ 1, & v \geqslant 5 \end{cases} \quad \text{hence } P(X \in [2,4)) =$$

$$P(X<4) - P(X<2) = F_X(4) - F_X(2) = \tfrac{3}{4} - \tfrac{1}{4} = \tfrac{1}{2}.$$



$$f(x) = \begin{cases} \dfrac{1}{4} & \text{for } x \in [1,5], \\ 0 & \text{for } x \notin [a,b]. \end{cases}$$

The rectangle has the area $\int_2^4 \frac{1}{4} dx = \frac{1}{2}$.

A continuous random variable

Definition 8.5. A random variable X is called **continuous** if its distribution F_X is a continuous function. Often X is defined by a *probability density* $f(x) \geqslant 0$ such that

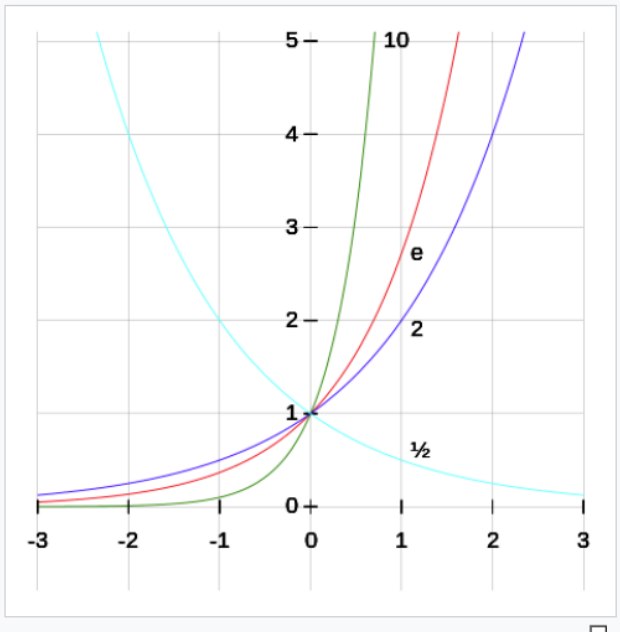
$$P(X < v) = F_x(v) = \int_{-\infty}^{v} f(x)dx.$$

$f(x)$ should quickly tend to 0 when $x \to \pm\infty$ and the area over $(-\infty, v]$ can be computed as $\int_{-\infty}^{v} f(x)dx$.

When $b \to +\infty$, the probability $P(X < v)$ tends to 1, so any probability density has $\int_{-\infty}^{+\infty} f(x)dx = 1$.

Exponential function

The real number e ($\approx 2.718\ldots$) occurs in many mathematical contexts. Like π, it is *transcendental* (i.e. not a solution of any polynomial with real coefficients).



Graphs of $y = b^x$ for various bases b: base 10, base e, base 2, base $\frac{1}{2}$. Each curve passes through the point $(0, 1)$ because any nonzero number raised to the power of 0 is 1. At $x = 1$, the value of y equals the base because any number raised to the power of 1 is the number itself.

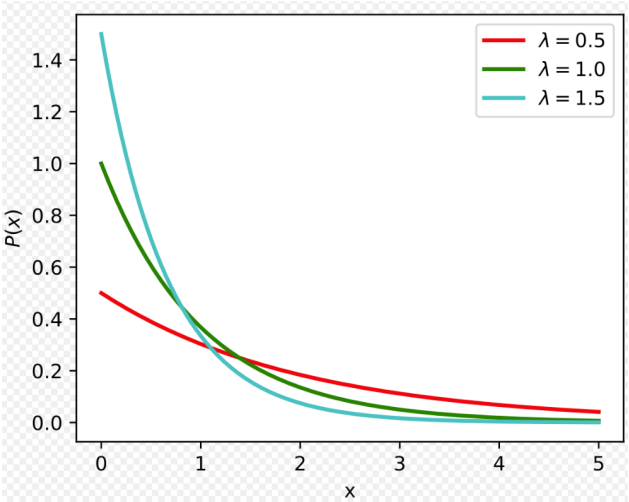
Images are taken from wiki.

Its most important property is that e is the unique positive real number such that

$$\frac{d}{dx}e^x = e^x.$$

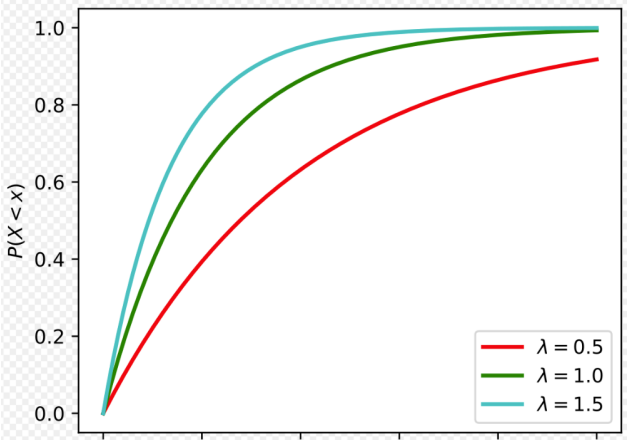
Exponential distribution

Definition 8.6. The **exponential distribution** describes time between events in a *Poisson process* when events occur continuously and independently at a constant average rate λ.



The pdf is $f(x;\lambda) =$

$$\begin{cases} \lambda e^{-\lambda x} & x \geqslant 0, \\ 0 & x < 0. \end{cases}$$



The cdf is $F(x;\lambda) =$

$$\begin{cases} 1 - e^{-\lambda x} & x \geqslant 0, \\ 0 & x < 0. \end{cases}$$

Higher dimensional variables

Definition 8.8. For any domain $D \subset \mathbb{R}^n$ with a volume $\mathrm{vol}(D)$, the *uniform continuous* random variable X on D has the density $f_X(p) = \dfrac{1}{\mathrm{vol}(D)}$ for any point $p \in D$ and $f_X(p) = 0$ for $p \in \mathbb{R}^n - D$.

This concept generalises Definition 8.1:
for $n = 1$ any interval domain $D = [a, b] \subset \mathbb{R}$ with the length ('volume') $\mathrm{vol}(D) = b - a$.
For any $A \subset \mathbb{R}^n$, the probability $P(X \in A)$ is $\mathrm{vol}(A \cap D)/\mathrm{vol}(D)$.

Curse of dimensionality

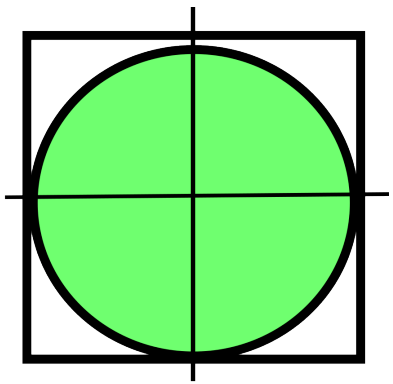
Higher dimensional analogues of objects from $\mathbb{R}^2$ or $\mathbb{R}^3$ may have rather counter intuitive properties.

Problem 8.9. Uniformly choose a random point q in the cube $[-1,1]^n$. What's the probability that q has a distance more than 1 from the origin $\vec{0} \in \mathbb{R}^n$?

For $n = 1$, any point $q \in [-1,1]$ has a distance $|q| \leqslant 1$, so the probability is 0. How about $n = 2$?

A 2D disk vs a 1D circle

Solution 8.9. The complementary event means that q is within a ball B^n with radius 1 and centre at the origin $\vec{0} \in \mathbb{R}^n$. For $n = 2$, this ball is called a *disk* and has the area π. A *circle* is the 1D boundary of the 2D disk, hence has *length* 2π and no area.



Since the square $[-1, 1]^2$ has the area $2^2 = 4$, the required probability is $\frac{4-\pi}{4} = 1 - \frac{\pi}{4} \approx 21.5\%$ for $n = 2$.

For $n = 3$ we need the volume of a 3D ball.

The volume of a ball in $\mathbb{R}^n$

The volume of the n-dimensional ball of radius 1 is $V_n = \frac{\pi^{n/2}}{\Gamma(\frac{n}{2}+1)}$, where $\Gamma(m) = (m-1)!$ and $\Gamma(m+\frac{1}{2}) = (m-\frac{1}{2})(m-\frac{3}{2})\ldots\frac{1}{2}\sqrt{\pi}$.

Then $V_2 = \frac{\pi}{1!}$,

$$V_3 = \frac{\pi^{3/2}}{\Gamma(\frac{5}{2})} = \frac{\pi^{3/2}}{\frac{3}{2}\frac{1}{2}\sqrt{\pi}} = \frac{4}{3}\pi.$$

The probability for a point $q \in [-1,1]^3 \subset \mathbb{R}^3$ to be outside the 3D ball is $P_3 = 1 - \frac{V_3}{2^3} = 1 - \frac{\pi}{6} \approx 47.6\%$, nearly a half!

Is the ball shrinking?

When $n \to +\infty$, the volume ratio $\frac{V_n}{2^n} = \frac{\pi^{n/2}}{2^n \Gamma\left(\frac{n}{2}+1\right)}$ quickly converges to 0, hence

the probability to miss the unit ball $P_n = 1 - \frac{V_n}{2^n}$ tends to 1,

e.g. $P_4 = 1 - \frac{\pi^2}{32} \approx 69\%$, $P_6 = 1 - \frac{\pi^3}{3!2^6} \approx 92\%$,

$P_8 = 1 - \frac{\pi^4}{4!2^8} \approx 98.4\%$.

In high-dimensional $\mathbb{R}^n$ a uniform random choice of n coordinates within $[-1, 1]$ almost always produces a point near corners of the cube away from the center $\vec{0}$.

Keep this in mind when performing multidimensional grid search!

Time to revise and ask questions

- The density (pdf) of a continuous random variable is the derivative of the cdf.
- The uniform continuous random variable is defined by a constant density function over a bounded domain such as an interval in $\mathbb{R}$.
- The curse of dimensionality doesn't allow us to use our low dimensional intuition in $\mathbb{R}^2$ or $\mathbb{R}^3$ to similarly analyse high dimensional data.

Problem 8.10.(Thick skin problem). What is the radius r of the ball whose volume is only 50% of the unit ball volume in $\mathbb{R}^n$?

IF YOU CHOOSE AN ANSWER TO THIS QUESTION AT RANDOM, WHAT IS THE CHANCE YOU WILL BE CORRECT?

A) 25%
B) 50%
C) 0%
D) 25%