Mathematical Primer

Sanjay Arora
AI Center of Excellence
Red Hat

Ulrich Drepper
AI Center of Excellence
Red Hat

Why do I need mathematics?

To understand existing literature and techniques

To reason about why an idea has the potential to work or not work, to make educated guesses about what can go wrong or what is needed for a technique to work

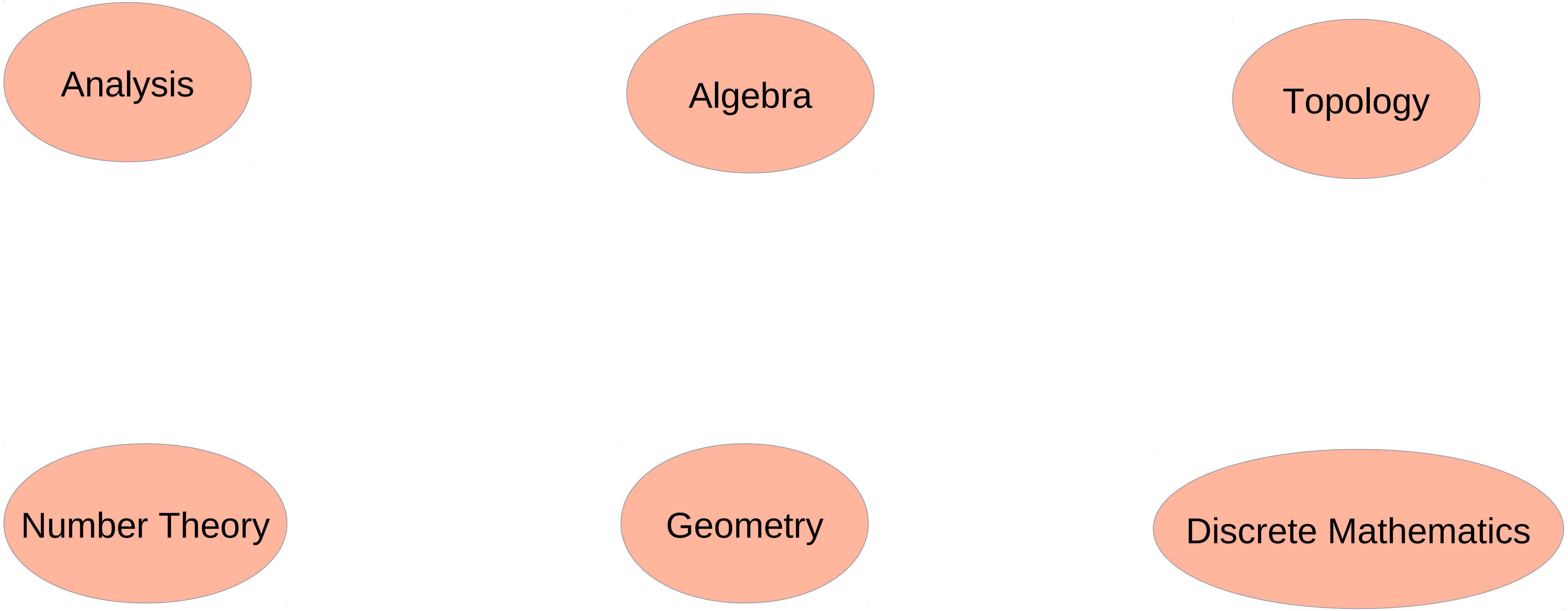
Why do I need mathematics?

You can go far without mathematics if the goal is to apply well-known techniques

Example:
Most kaggle competitors are NOT using advanced mathematics

They are creatively using existing techniques

The Mathematical Landscape



Red Hat

Analysis

- Real and complex analysis
- Ordinary and partial differential equations. Dynamical systems.
- Functional analysis
- Probability Theory
- Stochastic Calculus
- ...

Algebra

- Groups
- Rings
- Ideals
- Vector spaces
- Fields
- ...

Topology

- Topological spaces

Number Theory

Geometry

- Euclidean geometry
- Elliptic geometry
- Hyperbolic geometry
- ...

Discrete Mathematics

- Set Theory
- Combinatorics
- Graph Theory

Red Hat

Differential Geometry = Analysis + Topology / Geometry

Algebraic Topology = Algebra + Topology

Algebraic Geometry = Algebra + Geometry

Analytic Number Theory = Analysis + Number Theory

Algebraic Number Theory = Algebra + Number Theory

Red Hat

Information Theory

Signal Processing

Morse Theory

Concentration Inequalities

Category Theory

Commutative Algebra

Representation Theory

Symplectic Geometry

Homological Algebra

Knot Theory

Game Theory

Optimization

General Principle

Start with an abstract set with objects

Impose additional structure on set

Define mappings from one set to another that preserve some structure

Study properties of the sets and mappings

General Principle

Start with an abstract set with objects

Set of Vectors

Impose additional structure on set

Define mappings from one set to another that preserve some structure

Study properties of the sets and mappings

Red Hat

General Principle

Start with an abstract set with objects
Set of Vectors

Impose additional structure on set
Addition of vectors, multiplication by scalars/numbers

Define mappings from one set to another that preserve some structure

Study properties of the sets and mappings

General Principle

Start with an abstract set with objects

Set of Vectors

Impose additional structure on set

Addition of vectors, multiplication by scalars/numbers

Define mappings from one set to another that preserve some structure

Linear mappings from one vector space to another

Study properties of the sets and mappings

General Principle

Start with an abstract set with objects

Set of Vectors

Impose additional structure on set

Addition of vectors, multiplication by scalars/numbers

Define mappings from one set to another that preserve some structure

Linear mappings from one vector space to another

Study properties of the sets and mappings

Linear Algebra

Red Hat

Do I need all this?

No!

- Machine Learning uses a small subset of mathematics
- What do I need?
  - Probability – language to reason about uncertainty
  - Statistics – Probability + rules of thumb
  - Linear Algebra – mainly basic properties of matrices (linear mappings)
  - Some basic mathematical knowledge about functions

Red Hat

Do I need all this?

No!

- Applying machine learning models often requires even less mathematics.
- Danger:
  - Not understanding mathematics **can** lead one into subtle pitfalls.
- More theoretical investigations do require substantial mathematical knowledge as well as the ability to learn new mathematics.

Red Hat

Should I still learn mathematics?

Yes!

- Mathematics will help you reason about ideas in a deeper way
- It will help you understand current machine learning approaches
- It will make you creative where you can invent your own approaches or tweak existing ones in well-defined ways

Goals for this session

- To establish a baseline of terminology and concepts needed for upcoming sessions.
- To introduce you to a sampling of topics that you can explore in more detail.

Linear Algebra: Vectors and Matrices

What is Linear Algebra?

$$\text{Branch of mathematics that deals with :}$$

$$\text{vectors:} \quad \begin{bmatrix} 1 \\ -4 \\ 3 \end{bmatrix}$$

$$\text{and matrices :} \begin{bmatrix} 1 & 4 \\ 3 & 2 \\ -1 & -5 \end{bmatrix}$$

What is Linear Algebra?

More precisely :

Study of vector spaces and linear mappings (homomorphisms) between vector spaces

Ignore this definition for now

What is a vector?

$$\mathbf{v} = \begin{bmatrix} 3 \end{bmatrix}$$ → Point on 1-dimensional real line

$$\mathbf{v} = \begin{bmatrix} 3 \\ -1 \end{bmatrix}$$ → Point in 2-dimensional plane

$$\mathbf{v} = \begin{bmatrix} 3 \\ -1 \\ 5 \end{bmatrix}$$ → Point in 3-dimensional real space

$$\vdots$$

$$\mathbf{v} = \begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_n \end{bmatrix}$$ → Point in n-dimensional real space

What is a vector?

$$\mathbf{v} = \begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_n \end{bmatrix} \longrightarrow \text{Coordinates of point in n-dimensional space}$$

Note:

This simple expression represents a major leap in our ability. We can’t visualize higher dimensions but we can work in them through algebra.

What is a vector?

$$\mathbf{v} = \begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_n \end{bmatrix} \longrightarrow \text{Vector space} = \text{set of all such vectors}$$

Addition:

$$\begin{bmatrix} a_1 \\ a_2 \end{bmatrix} + \begin{bmatrix} b_1 \\ b_2 \end{bmatrix} = \begin{bmatrix} a_1 + b_1 \\ a_2 + b_2 \end{bmatrix}$$

Multiplication:

$$k * \begin{bmatrix} a_1 \\ a_2 \end{bmatrix} = \begin{bmatrix} k * a_1 \\ k * a_2 \end{bmatrix}$$

* There is a precise technical definition

What is a vector?

$$\begin{bmatrix} -1 \\ 3 \\ 0 \end{bmatrix} = -1 * \begin{bmatrix} 1 \\ 0 \\ 0 \end{bmatrix} + 3 * \begin{bmatrix} 0 \\ 1 \\ 0 \end{bmatrix} + 0 * \begin{bmatrix} 0 \\ 0 \\ 1 \end{bmatrix}$$

Special vectors called **basis vectors**

Same as unit vectors from elementary geometry: $\hat{i}, \hat{j}, \hat{k}$

Red Hat

What is a vector?

$$\begin{bmatrix} x_1 \\ x_2 \\ x_3 \end{bmatrix} = x_1 * \begin{bmatrix} 1 \\ 0 \\ 0 \end{bmatrix} + x_2 * \begin{bmatrix} 0 \\ 1 \\ 0 \end{bmatrix} + x_3 * \begin{bmatrix} 0 \\ 0 \\ 1 \end{bmatrix}$$

Can write **any** vector in terms of basis vectors

What is a vector?

$$\begin{bmatrix} -1 \\ 3 \\ 0 \end{bmatrix} = -2 * \begin{bmatrix} 1 \\ -1 \\ 0 \end{bmatrix} + 1 * \begin{bmatrix} 1 \\ 1 \\ 0 \end{bmatrix} + 0 * \begin{bmatrix} 0 \\ 0 \\ 1 \end{bmatrix}$$

Different basis!

What is a vector?

Basis vectors are fundamental building blocks of other vectors

Every vector space has a basis

There can be an infinite number of bases

Number of basis elements in a basis is fixed $\equiv$ **Dimension of vector space**

* $\mathbb{Z}_2$ does have only one unique basis

Inner/Dot Products

Can impose more structure on vectors

$$\langle \mathbf{v}, \mathbf{w} \rangle \longrightarrow \text{Real number (for these lectures)}$$

with properties:

$$\langle \mathbf{v}, \mathbf{w} \rangle = \langle \mathbf{w}, \mathbf{v} \rangle$$

$$\langle \mathbf{u} + \mathbf{v}, \mathbf{w} \rangle = \langle \mathbf{u}, \mathbf{w} \rangle + \langle \mathbf{v}, \mathbf{w} \rangle$$

$$\langle \mathbf{v}, \mathbf{v} \rangle \geq 0 \text{ with equality} \iff \mathbf{v} = 0$$

Usual dot product satisfies all these properties

Red Hat

Inner/Dot Products

Can impose more structure on vectors

$\langle \mathbf{v}, \mathbf{w} \rangle \longrightarrow$ Real number (for these lectures)

Intuition: Overlap/Similarity between two vectors

Usual Dot Product

$$\langle \mathbf{v}, \mathbf{w} \rangle = \mathbf{v}.\mathbf{w} = v_1 w_1 + \ldots + v_n w_n$$

Norms/Lengths

Can use inner products to define a **norm**

$$||\mathbf{v}|| \equiv \sqrt{\langle \mathbf{v}, \mathbf{v} \rangle}$$

Can use norm to define **distance**

$$d(\mathbf{v}, \mathbf{w}) \equiv ||\mathbf{v} - \mathbf{w}||$$

Example

$$\langle \mathbf{v}, \mathbf{w} \rangle \longrightarrow \mathbf{v}.\mathbf{w} = v_1 w_1 + \ldots + v_n w_n$$

$$||\mathbf{v}|| \equiv \sqrt{\langle \mathbf{v}, \mathbf{v} \rangle} \longrightarrow ||\mathbf{v}|| = \sqrt{v_1^2 + \ldots + v_n^2}$$

$$d(\mathbf{v}, \mathbf{w}) \equiv ||\mathbf{v} - \mathbf{w}|| \longrightarrow d(\mathbf{v}, \mathbf{w}) = \sqrt{(v_1 - w_1)^2 + \ldots + (v_n - w_n)^2}$$

Red Hat

Other Norms

Notion of **length** of a vector

$$||\mathbf{v}||_1 \equiv \Sigma_{i=1}^{n} |v_i|$$

ℓ_1 norm

$$||\mathbf{v}||_2 \equiv (\Sigma_{i=1}^{n} v_i^2)^{\frac{1}{2}}$$

ℓ_2 norm

$$||\mathbf{v}||_\infty \equiv \max_{i \in [1,n]} |v_i|$$

ℓ_∞ norm

$$\mathbf{v} = \begin{pmatrix} -1 \\ 3 \\ 0 \end{pmatrix}$$

ℓ_1 norm : $|-1| + |3| + |0| = 4$

ℓ_2 norm : $((-1)^2 + 3^2 + 0^2)^{\frac{1}{2}} = \sqrt{10}$

ℓ_∞ norm : $max(|-1|, |3|, |-0|) = 3$

What is a matrix?

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

Table of Numbers?

No!!!

Red Hat

What is a matrix?

Recall the definition of a function or mapping:

A function from set A to set B is a rule that assigns to **each** element of A **exactly one** element of B

Notation: $f : A \to B$

What is a matrix?

An **n (rows) x m (columns)** matrix is a **linear** mapping from $\mathbb{R}^m$ to $\mathbb{R}^n$

$$\mathbb{R}^1 = \text{real line}$$

$$\mathbb{R}^2 = \text{usual Cartesian x-y plane}$$

$$\mathbb{R}^3 = \text{usual Cartesian x-y-z volume}$$

$$\mathbb{R}^4 = \text{set of tuples (x,y,z,w)}$$

$$\underbrace{\begin{pmatrix} -2 \\ -1 \end{pmatrix}}_{\text{vector in } \mathbb{R}^2} = \underbrace{\begin{pmatrix} 1 & 2 & -1 \\ 2 & 1 & 1 \end{pmatrix}}_{\text{2 x 3 matrix}} \underbrace{\begin{pmatrix} -1 \\ 0 \\ 1 \end{pmatrix}}_{\text{vector in } \mathbb{R}^3}$$

Red Hat

What is a matrix?

Linear mapping

$$f(\mathbf{v} + \mathbf{w}) = f(\mathbf{v}) + f(\mathbf{w})$$

$$f(k\mathbf{w}) = kf(\mathbf{w})$$

What is a matrix?

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

$$\begin{pmatrix} 1 \\ 0 \end{pmatrix} \xrightarrow{A} \begin{pmatrix} 1 \\ 2 \end{pmatrix}$$

$$\begin{pmatrix} 0 \\ 1 \end{pmatrix} \xrightarrow{A} \begin{pmatrix} 2 \\ 1 \end{pmatrix}$$

These define mappings for all other vectors

$$\begin{pmatrix} a \\ b \end{pmatrix} = a \begin{pmatrix} 1 \\ 0 \end{pmatrix} + b \begin{pmatrix} 0 \\ 1 \end{pmatrix} \xrightarrow{A} a \begin{pmatrix} 1 \\ 2 \end{pmatrix} + b \begin{pmatrix} 2 \\ 1 \end{pmatrix}$$

What is a matrix?

Trace: $tr(A) = \Sigma A_{ii}$ Sum of diagonal entries

Transpose: $(A^T)_{ij} = A_{ji}$

$$\begin{cases} A = \begin{pmatrix} 1 & 2 \\ 2 & 1 \\ 3 & 0 \end{pmatrix} \\ A^T = \begin{pmatrix} 1 & 2 & 3 \\ 2 & 1 & 0 \end{pmatrix} \end{cases}$$

Symmetric: $A^T = A \iff A_{ij} = A_{ji}$

$$\left\{ A = \begin{pmatrix} 1 & 2 \\ 2 & 1 \end{pmatrix} \right.$$

Red Hat

Eigenvectors and Eigenvalues

Special Vectors: Matrix only scales vectors by constant value

$$\begin{pmatrix} 1 \\ 1 \end{pmatrix} \xrightarrow{A} \underbrace{3}_{\text{eigenvalue}} \begin{pmatrix} 1 \\ 1 \end{pmatrix}$$

$$\begin{pmatrix} 1 \\ -1 \end{pmatrix} \xrightarrow{A} \underbrace{-1}_{\text{eigenvalue}} \begin{pmatrix} 1 \\ -1 \end{pmatrix}$$

Red Hat

Diagonalizing a Matrix

Eigenvectors $\begin{pmatrix} 1 \\ 1 \end{pmatrix}, \begin{pmatrix} 1 \\ -1 \end{pmatrix}$

Eigenvectors form a basis

$$\begin{pmatrix} a \\ b \end{pmatrix} = a\begin{pmatrix} 1 \\ 0 \end{pmatrix} + b\begin{pmatrix} 0 \\ 1 \end{pmatrix} \quad \longleftrightarrow \quad A = \begin{pmatrix} 1 & 2 \\ 2 & 1 \end{pmatrix}$$

$$\begin{pmatrix} a \\ b \end{pmatrix} = \frac{a+b}{2}\begin{pmatrix} 1 \\ 1 \end{pmatrix} + \frac{a-b}{2}\begin{pmatrix} 1 \\ -1 \end{pmatrix} \quad \longleftrightarrow \quad \boxed{A = \begin{pmatrix} 3 & 0 \\ 0 & -1 \end{pmatrix}}$$

Diagonal

Red Hat

We like symmetric matrices

Result 1: If A is symmetric and real, its eigenvalues are real.

Proof:

$$A\mathbf{v} = \lambda\mathbf{v}$$

Multiply by: $\mathbf{v}^{*T}$

$$\mathbf{v}^{*T}A\mathbf{v} = \lambda\mathbf{v}^{*T}\mathbf{v}$$

Take transpose and complex conjugate:

$$\mathbf{v}^{*T}A^{*T}\mathbf{v} = \lambda^{*}\mathbf{v}^{*T}\mathbf{v}$$

Subtract:

= 0 since A real, symmetric

$$\mathbf{v}^{*T}(A^{*T} - A)\mathbf{v} = (\lambda^{*} - \lambda)\mathbf{v}^{*T}\mathbf{v}$$

Red Hat

We like symmetric matrices

Result 2: If A is symmetric and real, eigenvectors belonging to unequal eigenvalues are orthogonal

Proof:

$$A\mathbf{v}_1 = \lambda_1 \mathbf{v}_1 \qquad A\mathbf{v}_2 = \lambda_2 \mathbf{v}_2 \qquad \lambda_1 \neq \lambda_2$$

$$\mathbf{v}_2^T A\mathbf{v}_1 = \lambda_1 \mathbf{v}_2^T \mathbf{v}_1 \qquad \mathbf{v}_1^T A\mathbf{v}_2 = \lambda_2 \mathbf{v}_1^T \mathbf{v}_2$$

$$\mathbf{v}_1^T A^T \mathbf{v}_2 = \lambda_1 \mathbf{v}_1^T \mathbf{v}_2$$

Subtract: $$\mathbf{v}_1^T \underbrace{(A^T - A)}_{\text{Symmetric means 0}} \mathbf{v}_2 = (\lambda_1 - \lambda_2)\mathbf{v}_1^T \mathbf{v}_2$$

$$\lambda_1 \neq \lambda_2 \implies \mathbf{v}_1^T \mathbf{v}_2 = 0$$ **Eigenvectors orthogonal**

Red Hat

Summary

Punch-lines:

- Linear algebra = study of set of vectors and maps (matrices) between them.

- Every matrix represents a mapping or a function.

- There are certain “special” vectors that are only scaled by the mapping represented by the matrix (as long as it's a square matrix. These vectors are called eigenvectors. The scale factors are called eigenvalues.

- By representing every generic vector in terms of eigenvectors, the form of the matrix simplifies a lot: it becomes diagonal (non-diagonal entries are zero).

Why do we need all this for machine learning?

- Features can be represented as real-valued vectors in high-dimensional spaces.

- Want to use transformations/mappings/matrices in these spaces to make data more amenable to our algorithms.

Fourier Transforms

Feature Engineering

Doing a mathematical transformation on input features can:

Make patterns more obvious

Make it easier for machine-learning algorithms to run on data

Red Hat

Feature Engineering

Transforms are one way to engineer features

Fourier transforms

Laplace transforms

Wavelet transforms

Basis of a Vector Space

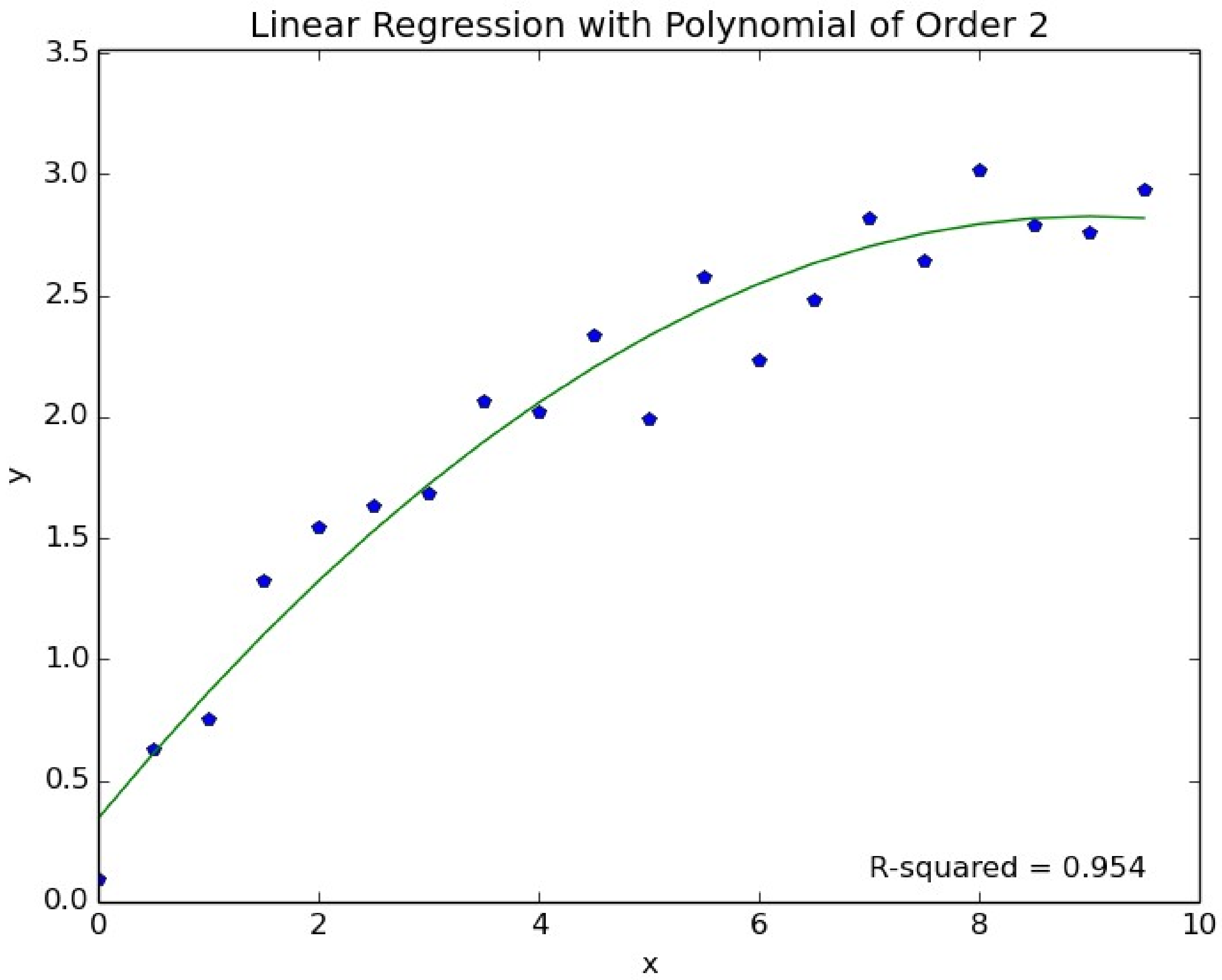
Recall definition of basis:

$$\mathbf{v} = \begin{pmatrix} -1 \\ 3 \\ 0 \end{pmatrix} = \text{-1} \begin{pmatrix} 1 \\ 0 \\ 0 \end{pmatrix} +3 \begin{pmatrix} 0 \\ 1 \\ 0 \end{pmatrix} 0 \begin{pmatrix} 0 \\ 0 \\ 1 \end{pmatrix}$$

Can decompose any vector as sum of these “special” vectors

$$\textbf{Dimension} = \text{number of basis elements}$$

Infinite-dimensional basis



Function (blue points) at regularly spaced points

Red Hat

Infinite-dimensional basis

Suppose we have a function (blue points) at regularly spaced points

x	y
x1	y1
x2	y2
x3	y3
...	...

Summarize the function in a table

$$\mathbf{v} = \begin{pmatrix} y_1 \\ y_2 \\ y_3 \\ \vdots \\ y_n \end{pmatrix}$$

Have n samples of function

Infinite-dimensional basis

Suppose we have a function (blue points) at regularly spaced points

$$\mathbf{v} = \begin{pmatrix} y_1 \\ y_2 \\ y_3 \\ \vdots \\ y_n \end{pmatrix} = \mathrm{y}_1 \begin{pmatrix} 1 \\ 0 \\ 0 \\ \vdots \\ 0 \end{pmatrix} + \mathrm{y}_2 \begin{pmatrix} 0 \\ 1 \\ 0 \\ \vdots \\ 0 \end{pmatrix} + \mathrm{y}_3 \begin{pmatrix} 0 \\ 0 \\ 1 \\ \vdots \\ 0 \end{pmatrix} + \ldots + \mathrm{y}_n \begin{pmatrix} 0 \\ 0 \\ 0 \\ \vdots \\ 1 \end{pmatrix}$$

One possible basis

$$\xrightarrow{\text{describe function exactly}}$$

$$n \to \infty$$

$$(0\,0\ldots0\,1\;0\ldots0)$$

$\downarrow$

1 in position i

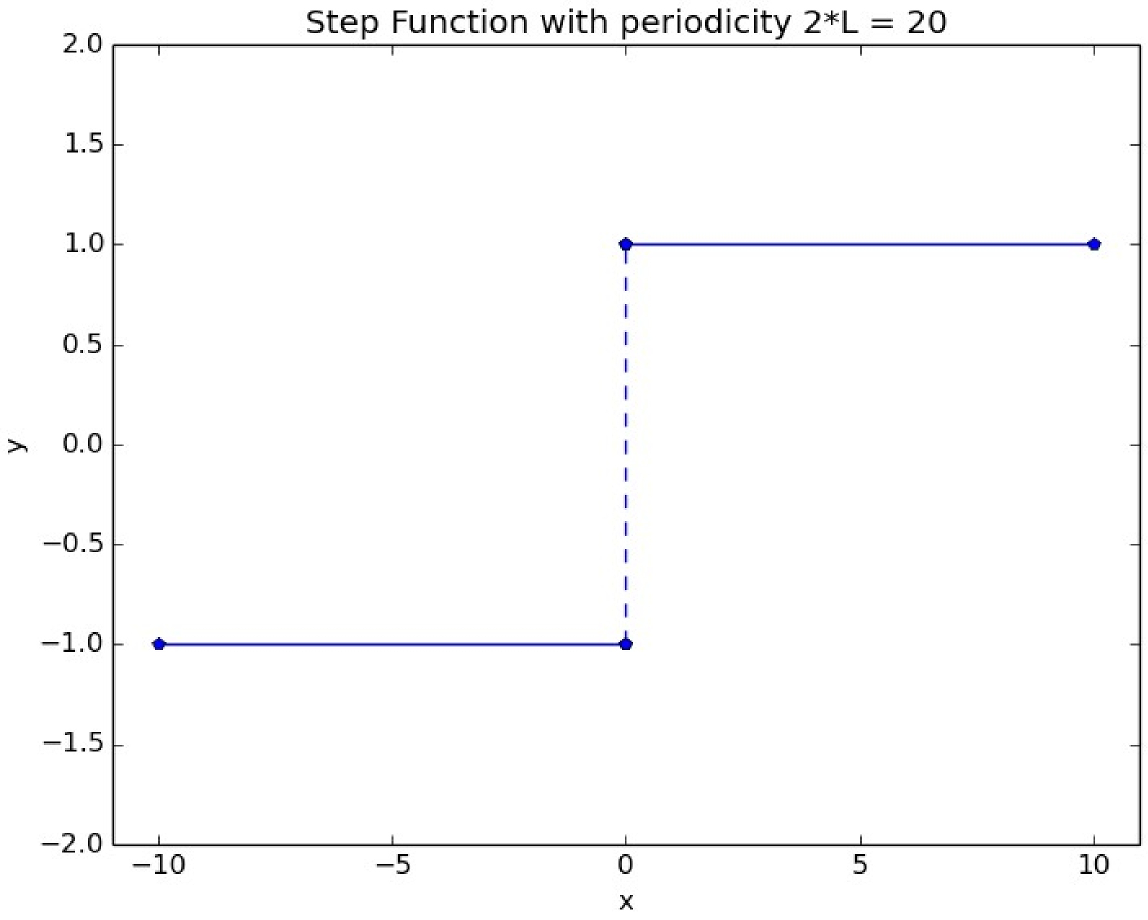
Any other basis?

$$
\begin{aligned}
f(x) = a_0 + a_1 \cos x + a_2 \cos 2x + a_3 \cos 3x + \dots \\
+ b_1 \sin x + b_2 \sin 2x + b_3 \sin 3x + \dots
\end{aligned}
$$

Expansion of a function in this **trigonometric basis** is called a **Fourier series**

Fourier Series

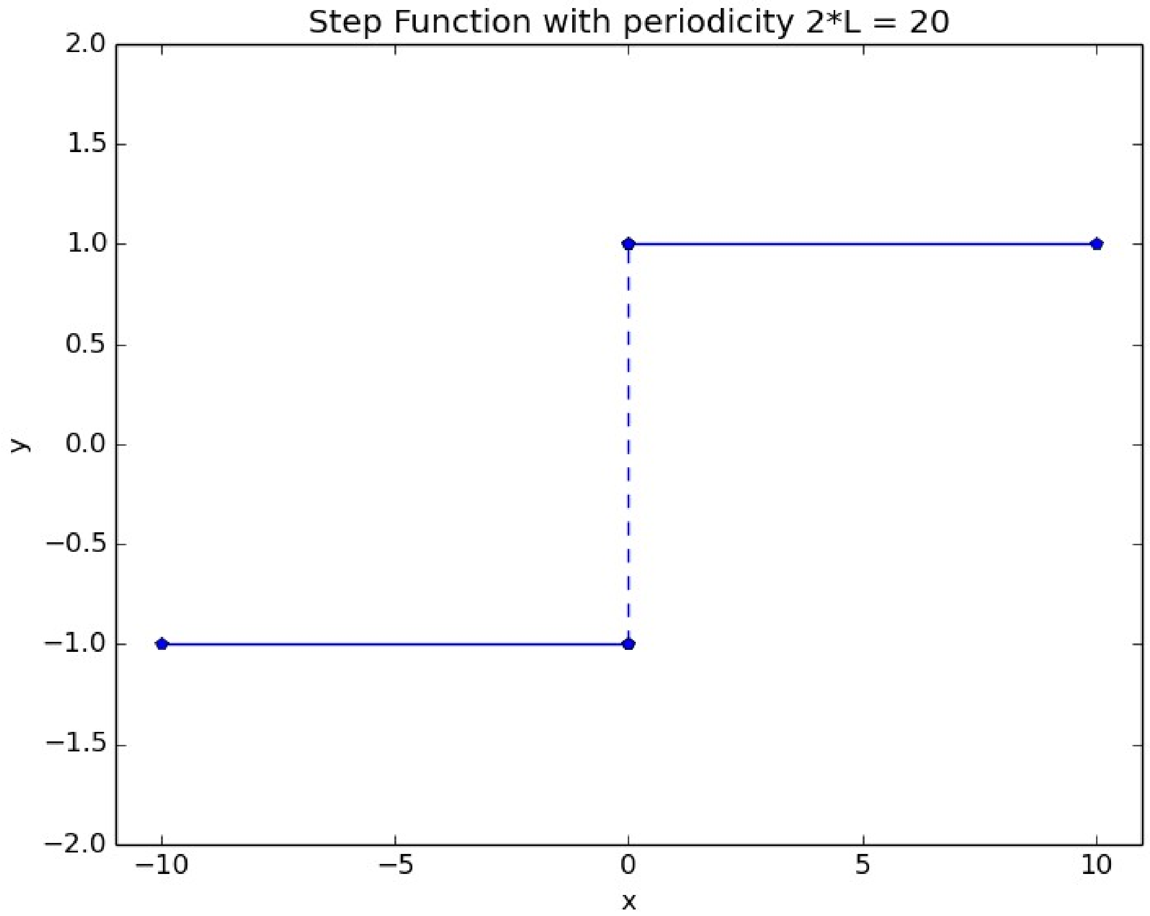
Consider a periodic function: $f(x)$



Repeats in both directions

Fourier Series

Consider a periodic function: $f(x)$



$$f(x) = a_0 + a_1 \cos\frac{\pi x}{L} + a_2 \cos\frac{2\pi x}{L} + \ldots b_1 \sin\frac{\pi x}{L} + b_2 \sin\frac{2\pi x}{L} + \ldots$$

Fourier Series

$$f(x) = a_0 + a_1 \cos\frac{\pi x}{L} + a_2 \cos\frac{2\pi x}{L} + \ldots b_1 \sin\frac{\pi x}{L} + b_2 \sin\frac{2\pi x}{L} + \ldots$$

$$\cos\frac{m\pi x}{L}, \sin\frac{m\pi x}{L} \text{ periodic on interval [-L, L]}$$

$\underline{x \to x + 2L}$:

$$\cos\frac{m\pi x}{L} \to \cos\left(\frac{m\pi x}{L} + 2m\pi\right) = \cos\frac{m\pi x}{L}$$

$$\sin\frac{m\pi x}{L} \to \sin\left(\frac{m\pi x}{L} + 2m\pi\right) = \sin\frac{m\pi x}{L}$$

Red Hat

Fourier Series

$$f(x) = \underbrace{a_0 + a_1 \cos\frac{\pi x}{L} + a_2 \cos\frac{2\pi x}{L} + \ldots b_1 \sin\frac{\pi x}{L} + b_2 \sin\frac{2\pi x}{L} + \ldots}$$

Periodic function – heading in the right direction

$\frac{m\pi x}{L}$: higher m $\rightarrow$ higher frequency

a_m, b_m : Strength of frequency m

Interpretation: Replace original function f(x) by frequency content

Red Hat

Fourier Series

Think of $\cos\dfrac{m\pi x}{L}, \sin\dfrac{m\pi x}{L}$ as basis elements

with inner/dot product defined as:

$$\left\langle \cos\frac{m\pi x}{L}, \cos\frac{n\pi x}{L} \right\rangle \equiv \int_{-L}^{L} \cos\frac{m\pi x}{L} \cos\frac{n\pi x}{L} dx$$

$$\left\langle \sin\frac{m\pi x}{L}, \sin\frac{n\pi x}{L} \right\rangle \equiv \int_{-L}^{L} \sin\frac{m\pi x}{L} \sin\frac{n\pi x}{L} dx$$

$$\left\langle \cos\frac{m\pi x}{L}, \sin\frac{n\pi x}{L} \right\rangle \equiv \int_{-L}^{L} \cos\frac{m\pi x}{L} \sin\frac{n\pi x}{L} dx$$

Red Hat

Fourier Series

$$\cos\frac{m\pi x}{L}, \sin\frac{m\pi x}{L} \text{ form an } \underline{\text{orthogonal basis!!}}$$

$$\boxed{\left\langle \cos\frac{m\pi x}{L}, \sin\frac{n\pi x}{L} \right\rangle \equiv \int_{-L}^{L} \cos\frac{m\pi x}{L}\sin\frac{n\pi x}{L}dx = 0\,, \forall m, n}$$

$$\boxed{\left\langle \cos\frac{m\pi x}{L}, \cos\frac{n\pi x}{L} \right\rangle \equiv \int_{-L}^{L} \cos\frac{m\pi x}{L}\cos\frac{n\pi x}{L}dx = \delta_{m,n}L\,, \forall m, n}$$

$$\boxed{\left\langle \sin\frac{m\pi x}{L}, \sin\frac{n\pi x}{L} \right\rangle \equiv \int_{-L}^{L} \sin\frac{m\pi x}{L}\sin\frac{n\pi x}{L}dx = \delta_{m,n}L\,, \forall m, n}$$

= 1 when m = n, 0 otherwise

Red Hat

Fourier Series

$$f(x) = a_0 + a_1 \cos\frac{\pi x}{L} + a_2 \cos\frac{2\pi x}{L} + \ldots b_1 \sin\frac{\pi x}{L} + b_2 \sin\frac{2\pi x}{L} + \ldots$$

Easy to calculate a_m, b_n

$$\int_{-L}^{L} f(x)dx = \int_{-L}^{L} a_0 + a_1 \cos\frac{\pi x}{L} + a_2 \cos\frac{2\pi x}{L} + \ldots b_1 \sin\frac{\pi x}{L} + b_2 \sin\frac{2\pi x}{L} + \ldots dx$$

$$= a_0 2L$$

$$\boxed{a_0 = \frac{1}{2L}\int_{-L}^{L} f(x)dx}$$

Red Hat

Fourier Series

$$f(x) = a_0 + a_1 \cos\frac{\pi x}{L} + a_2 \cos\frac{2\pi x}{L} + \ldots b_1 \sin\frac{\pi x}{L} + b_2 \sin\frac{2\pi x}{L} + \ldots$$

$$\begin{aligned} f(x)\cos\frac{m\pi x}{L} &= a_0 \cos\frac{m\pi x}{L} + a_1 \cos\frac{\pi x}{L}\cos\frac{m\pi x}{L} \\ &+ a_2 \cos\frac{2\pi x}{L}\cos\frac{m\pi x}{L} + \ldots b_1 \sin\frac{\pi x}{L}\cos\frac{m\pi x}{L} \\ &+ b_2 \sin\frac{2\pi x}{L}\cos m\frac{\pi x}{L} + \ldots \end{aligned}$$

Red Hat

Fourier Series

$$\int_{-L}^{L} f(x)\cos\frac{m\pi x}{L}dx = \int_{-L}^{L}\Big(a_0\cos\frac{m\pi x}{L} + a_1\cos\frac{\pi x}{L}\cos\frac{m\pi x}{L}$$

$$+a_2\cos\frac{2\pi x}{L}\cos\frac{m\pi x}{L} + \ldots b_1\sin\frac{\pi x}{L}\cos\frac{m\pi x}{L}$$

$$+ b_2\sin\frac{2\pi x}{L}\cos m\frac{\pi x}{L} + \ldots)$$

$$= a_m L$$

Red Hat

Fourier Series

$$a_0 = \frac{1}{2L}\int_{-L}^{L} f(x)dx$$

$$a_m = \frac{1}{L}\int_{-L}^{L} f(x)\cos\frac{m\pi x}{L}dx$$

$$b_m = \frac{1}{L}\int_{-L}^{L} f(x)\sin\frac{m\pi x}{L}dx$$

These measure the frequency content of the original function

Fourier Series

$$a_0 = \frac{1}{2L} \int_{-L}^{L} f(x) dx$$

0 if f odd $\longleftarrow$ $$a_m = \frac{1}{L} \int_{-L}^{L} f(x) \cos \frac{m\pi x}{L} dx$$

0 if f even $\longleftarrow$ $$b_m = \frac{1}{L} \int_{-L}^{L} f(x) \sin \frac{m\pi x}{L} dx$$

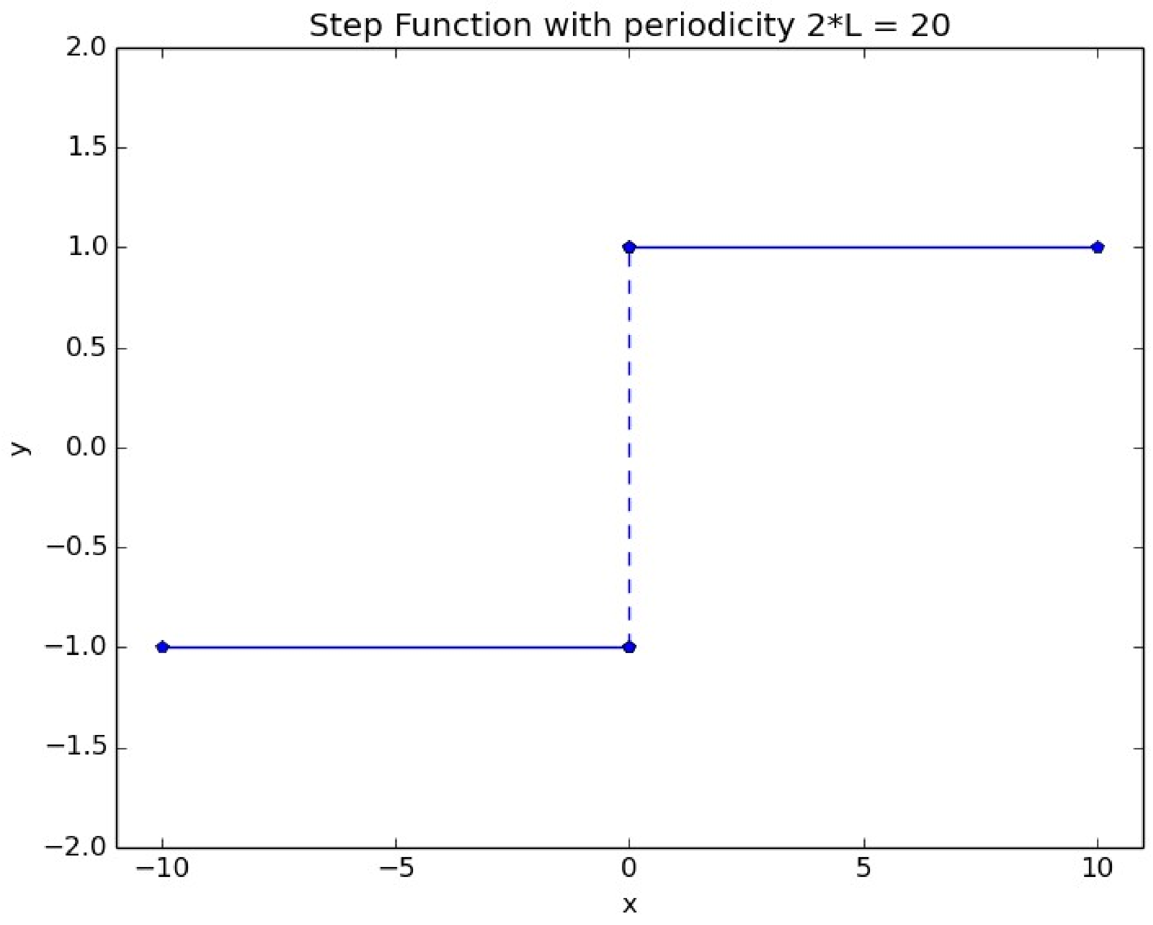
These measure the <u>frequency</u> content of the original function

Harmonic Analysis

General name of area of mathematics dealing with expansion of in functional bases

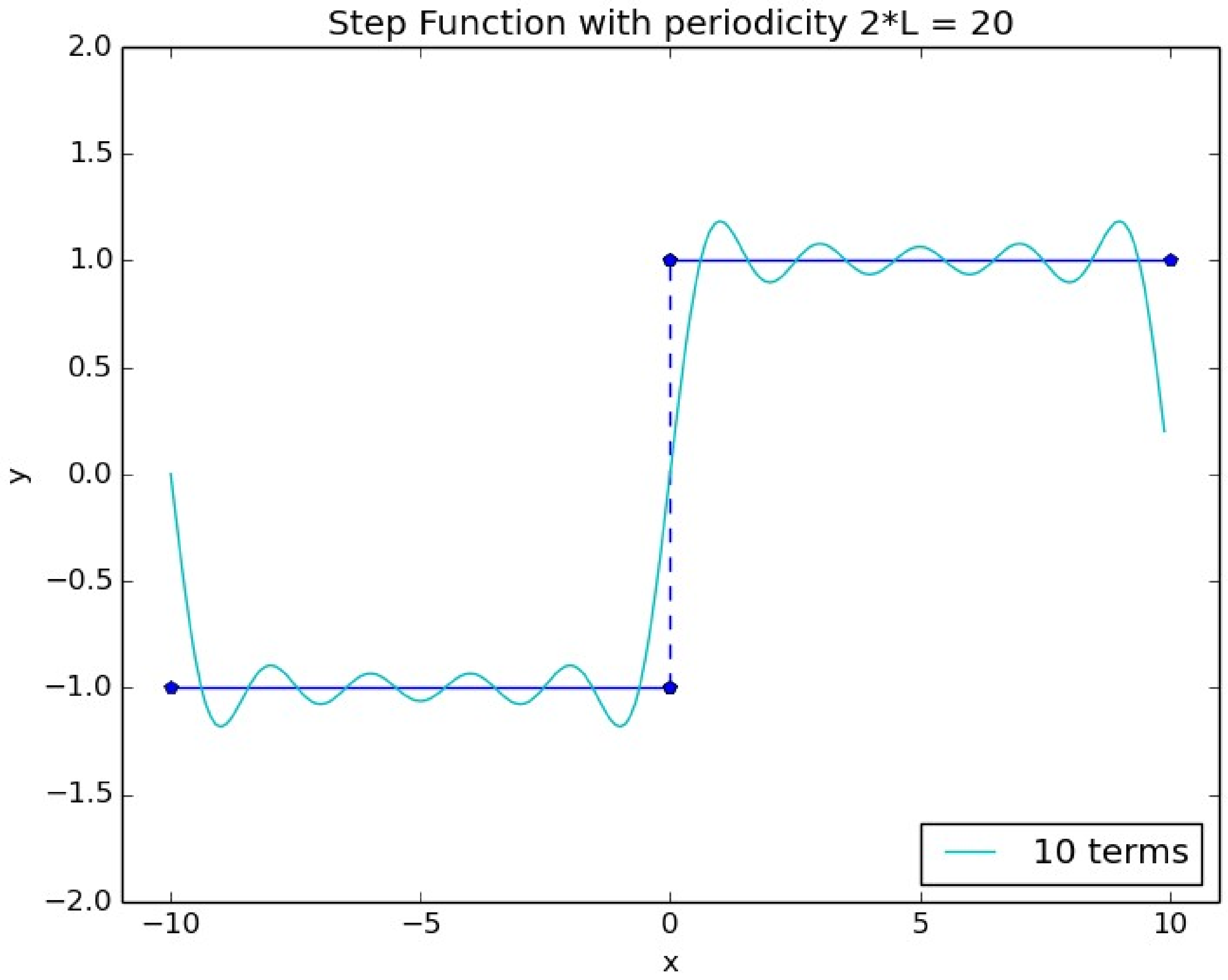
Look at books by Elias Stein, Princeton University

Example



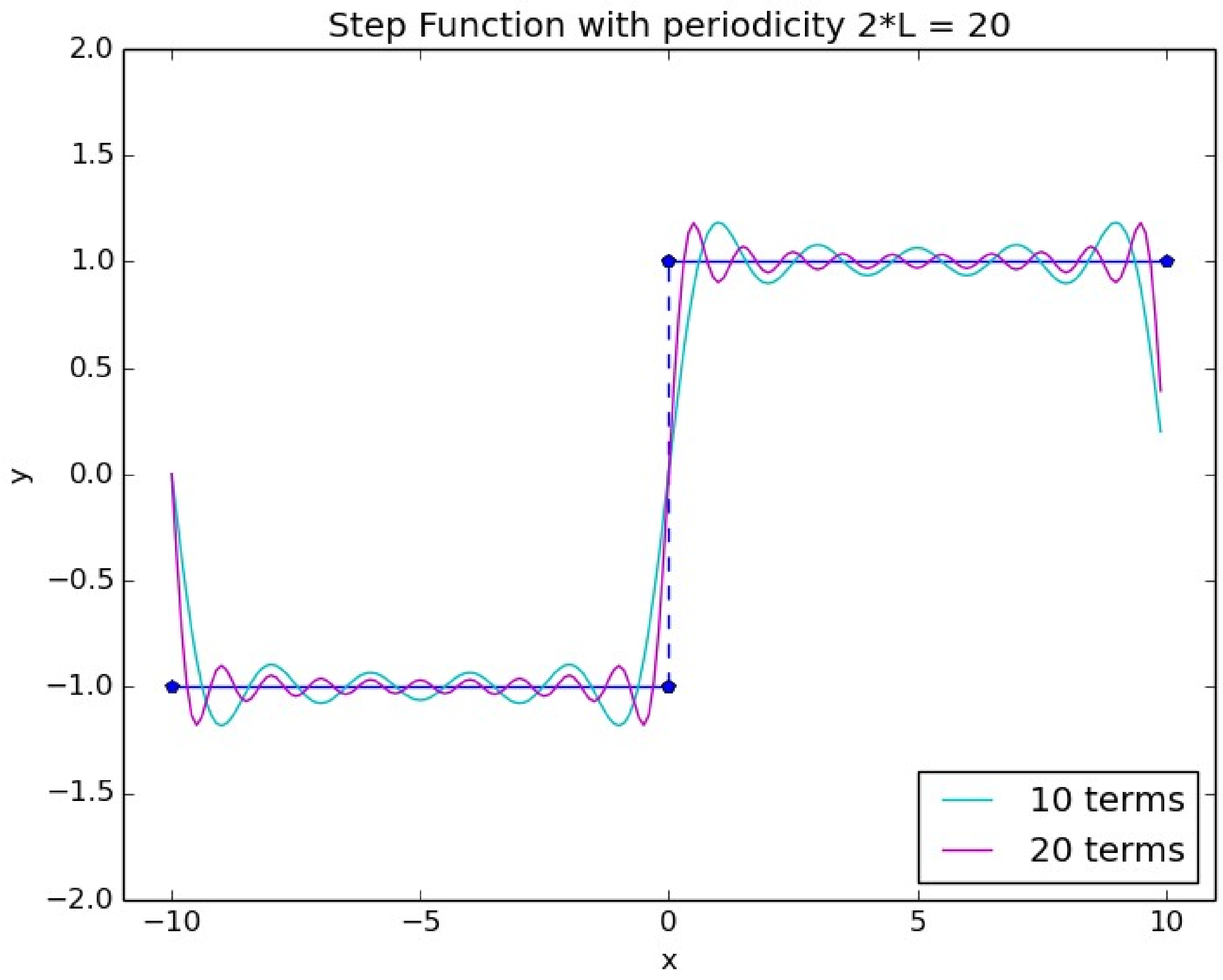
Red Hat

Example



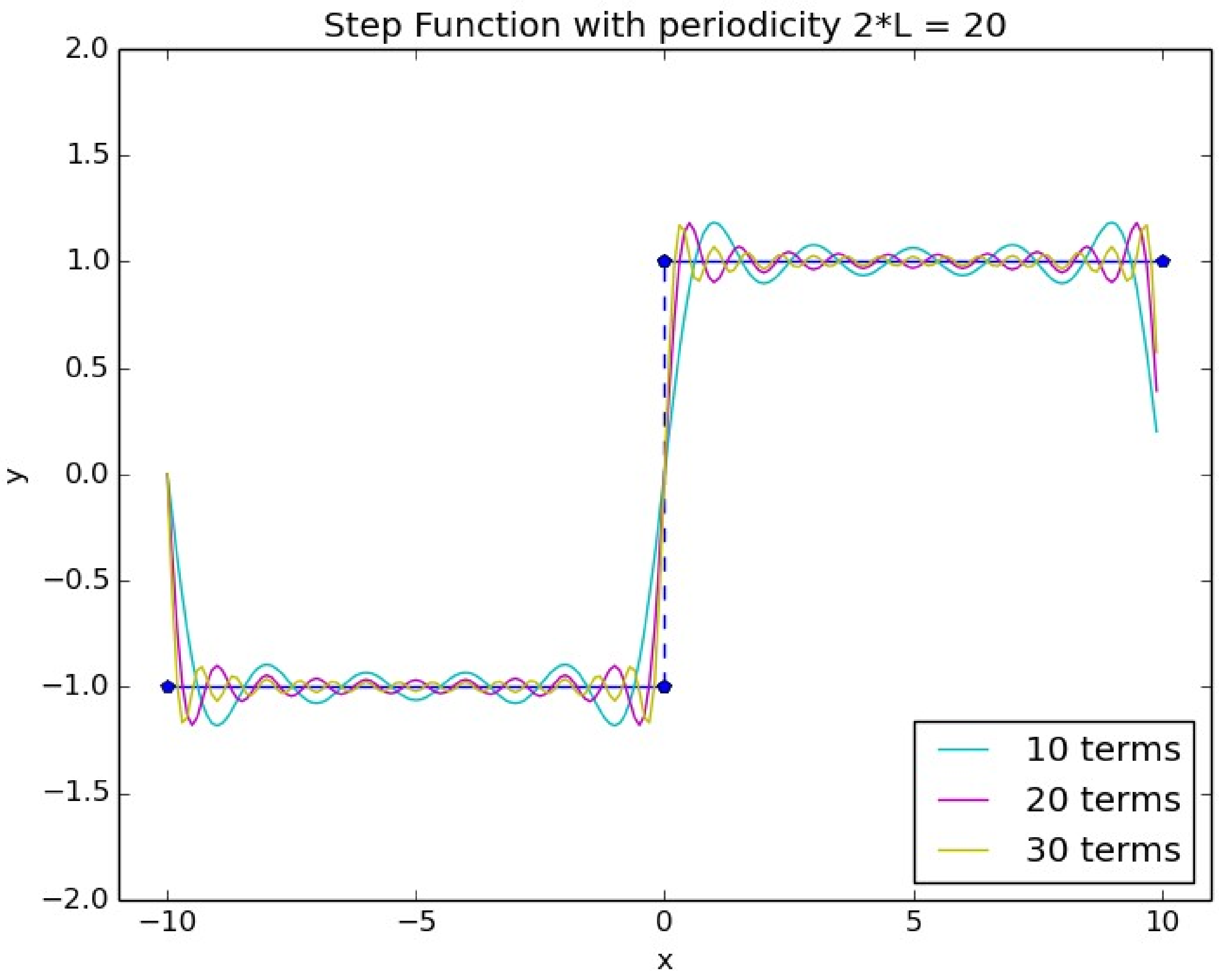
Step Function with periodicity 2*L = 20
2.0
1.5
1.0
0.5
0.0
−0.5
−1.0
−1.5
−2.0
y
−10
−5
0
5
10
x
10 terms

Example



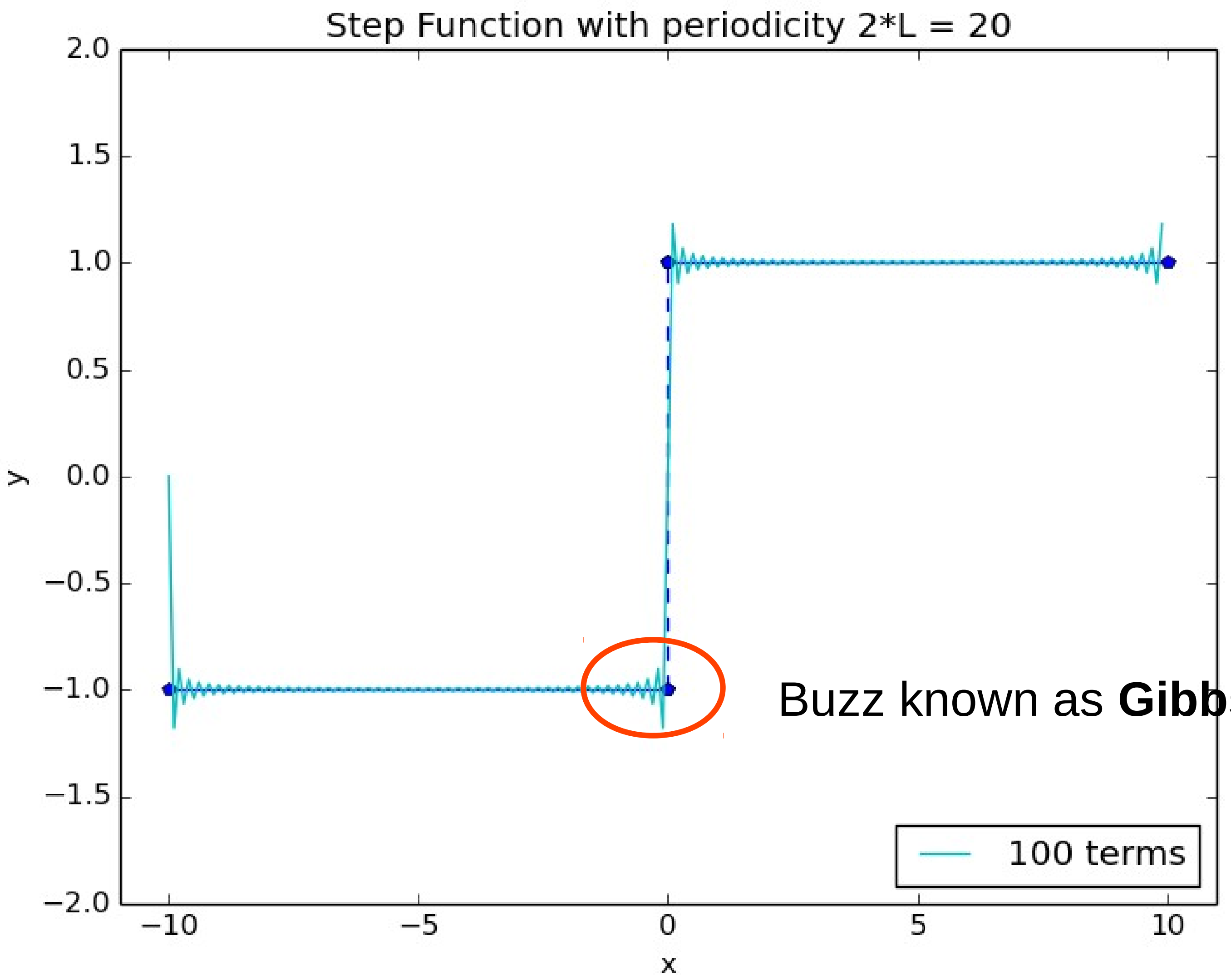
Step Function with periodicity 2*L = 20
2.0
1.5
1.0
0.5
0.0
−0.5
−1.0
−1.5
−2.0
y
−10
−5
0
5
10
x
10 terms
20 terms

Example



Red Hat

Example



Buzz known as **Gibbs phenomenon**

Fourier "Series" for Non-Periodic Functions

Define **Fourier Transform**:

$$\hat{f}(\omega) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} f(x) e^{i\omega x} dx$$

with **inverse Fourier transform**:

$$f(x) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} \hat{f}(\omega) e^{-i\omega x} d\omega$$

Fourier "Series" for Non-Periodic Functions

$$f(x) = a_0 + a_1 \cos\frac{\pi x}{L} + a_2 \cos\frac{2\pi x}{L} + \ldots b_1 \sin\frac{\pi x}{L} + b_2 \sin\frac{2\pi x}{L} + \ldots$$

Basis of trigonometric functions with different freq.

Original Function

Weights for different freq.

$$f(x) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} \hat{f}(\omega) e^{-i\omega x} d\omega$$

$$e^{i\omega x} = \cos\omega x + i \sin\omega x$$

Red Hat

Fourier "Series" for Non-Periodic Functions

$$f(x) = a_0 + a_1 \cos\frac{\pi x}{L} + a_2 \cos\frac{2\pi x}{L} + \ldots b_1 \sin\frac{\pi x}{L} + b_2 \sin\frac{2\pi x}{L} + \ldots$$

Describe f(x) by: $[a_0, a_1, \ldots, b_1, \ldots]$

$$f(x) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} \hat{f}(\omega) e^{-i\omega x} d\omega$$

Describe f(x) by: $\hat{f}(\omega)$

Aside on Dirac Delta

$$f(x) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} \hat{f}(\omega) e^{-i\omega x} d\omega \longleftarrow \hat{f}(\omega) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} f(x) e^{i\omega x} dx$$

$$f(x) = \frac{1}{\sqrt{2\pi}} \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} f(x') e^{i\omega x'} e^{-i\omega x} dx' d\omega$$

$$f(x) = \int_{-\infty}^{\infty} f(x') \underbrace{\left(\int_{-\infty}^{\infty} \frac{1}{2\pi} e^{i\omega(x'-x)} d\omega \right)}_{\equiv \delta(x'-x)} dx'$$

Very useful for computing complex distributions among other things

What's the distribution of $X^2 + Y^2$ if X, Y Gaussian$(0, \sigma^2)$

Please speak to me if you would like to know more

Numerical Fourier Transform

We relied on knowing analytic form of f(x)

to calculate a_0, a_m, b_m

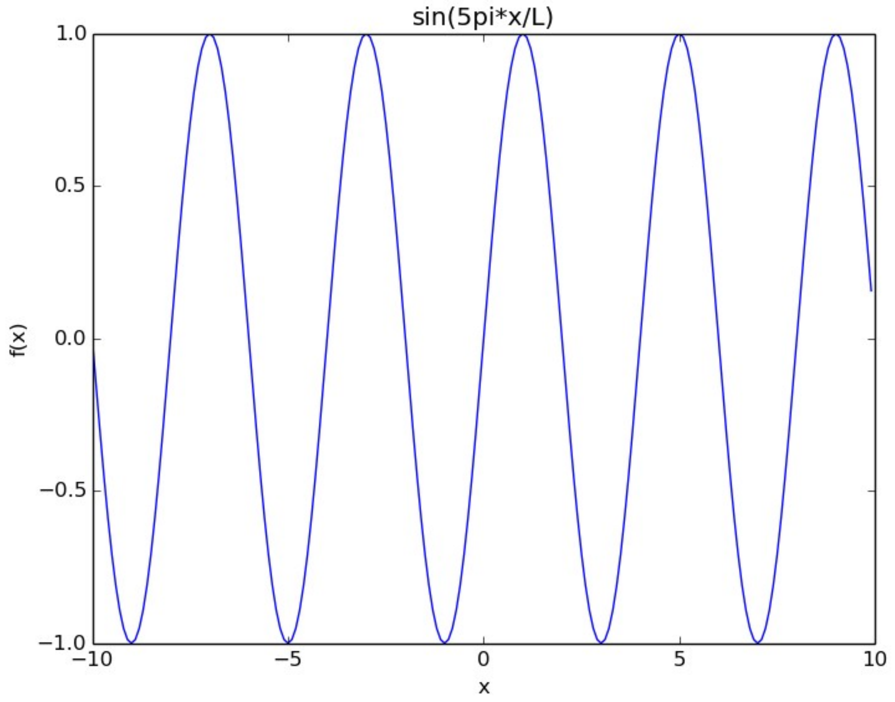
OR

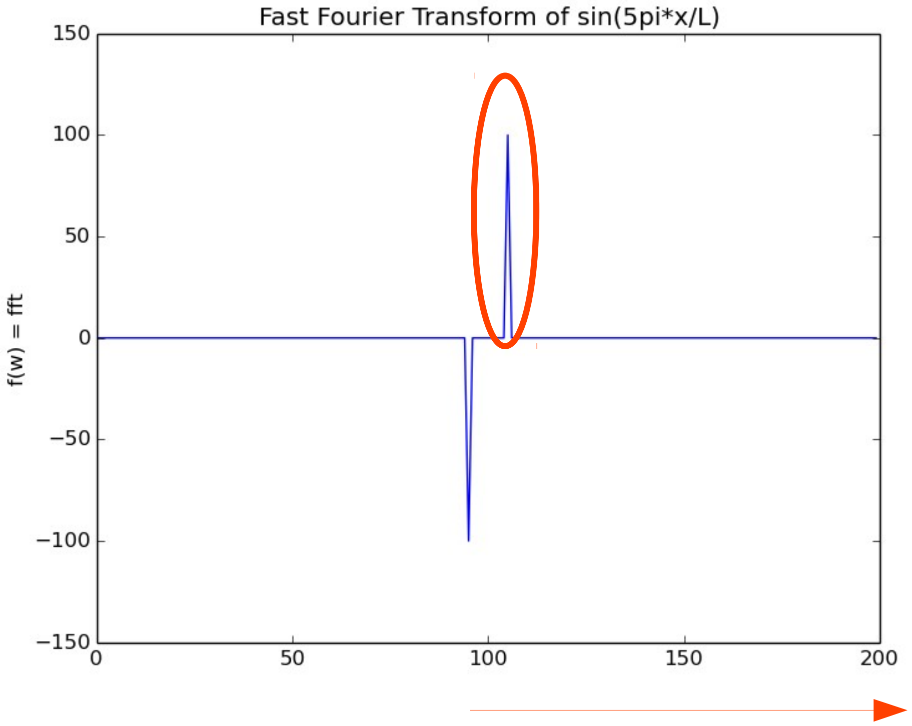
to calculate $\hat{f}(\omega)$

What if only have access to sampling of function?

$$\begin{bmatrix} x_1 & y_1 \\ x_2 & y_2 \\ \vdots & \vdots \end{bmatrix}$$

Fast Fourier Transform

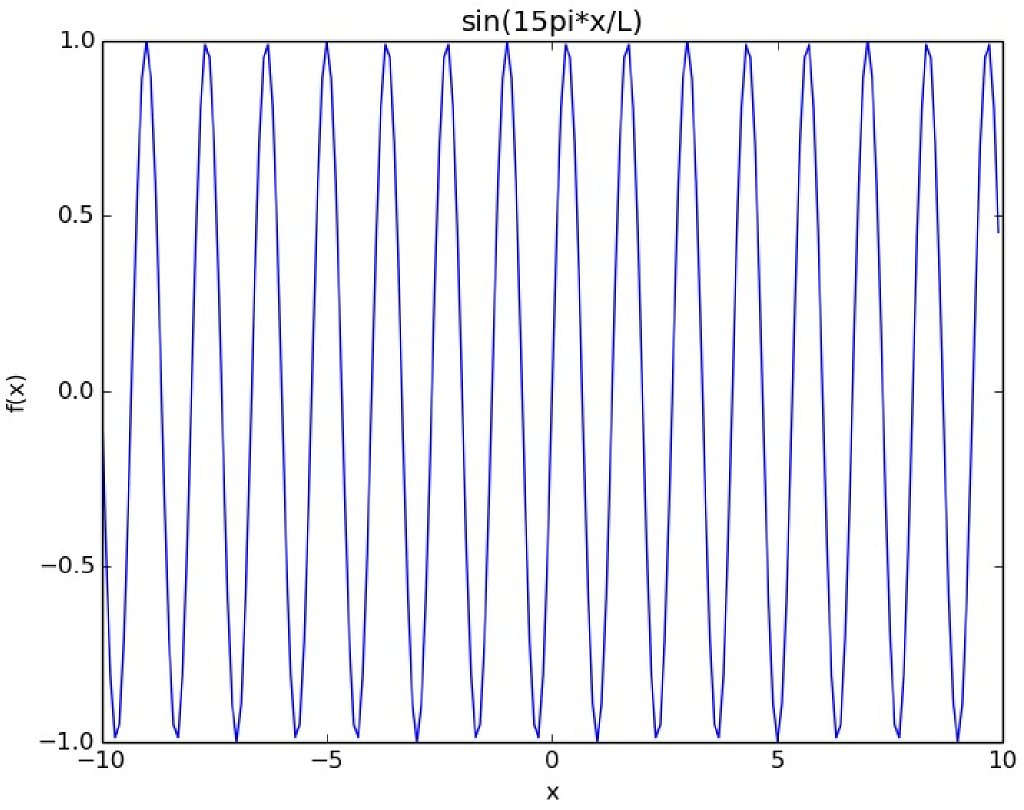


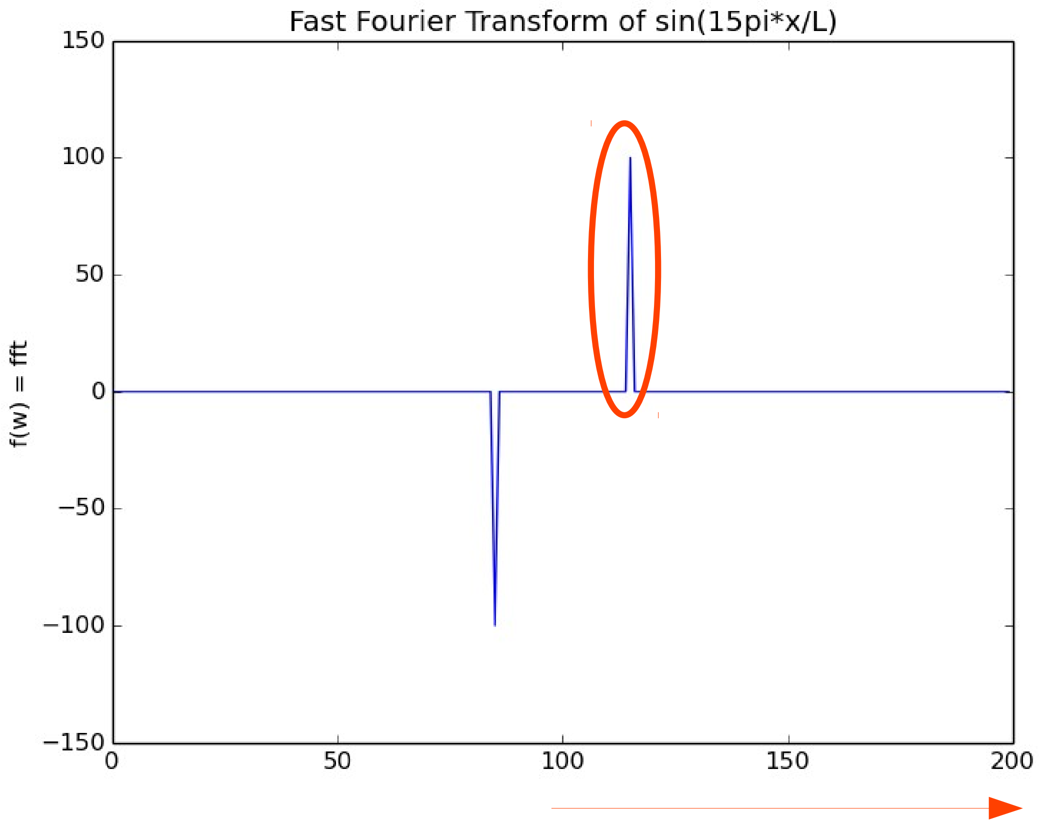


$$\sin \frac{\pi x}{L}$$

Frequency = 0

Fast Fourier Transform



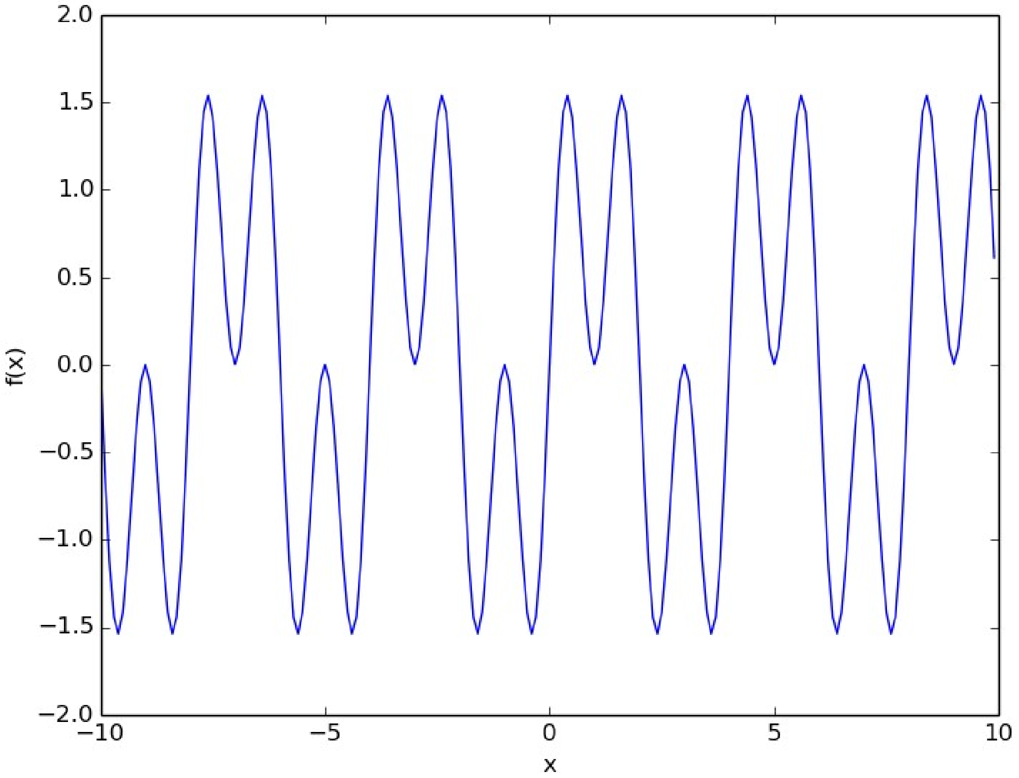


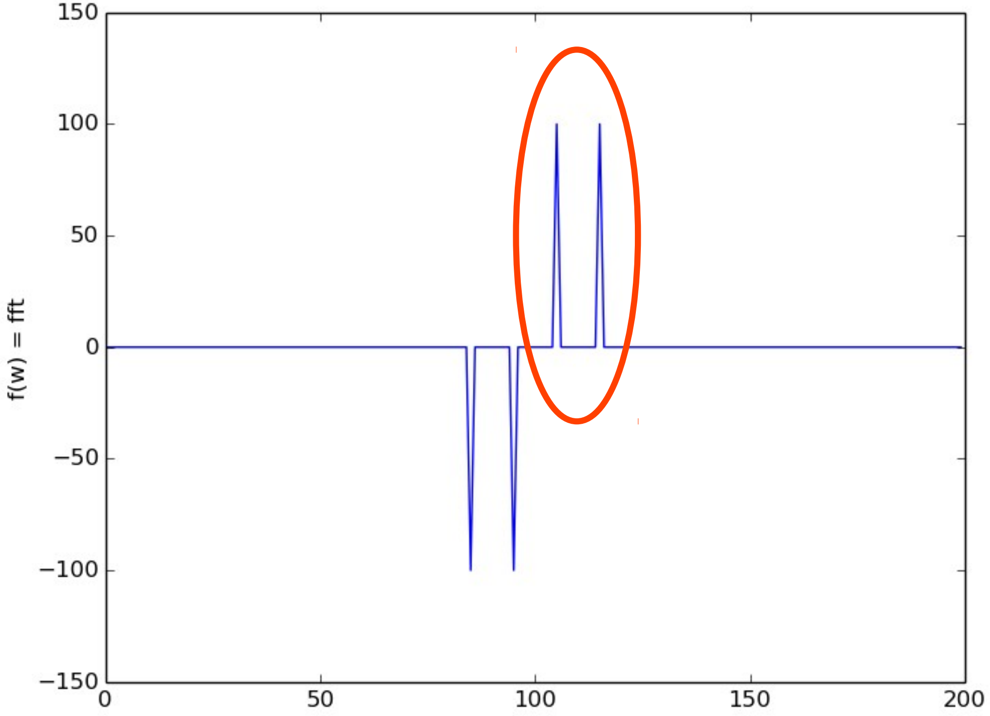
Frequency

$$\sin \frac{15\pi x}{L}$$

Red Hat

Fast Fourier Transform





Frequency

$$\sin \frac{5\pi x}{L} + \sin \frac{15\pi x}{L}$$

Probability

Probability

Solving problems with incomplete knowledge

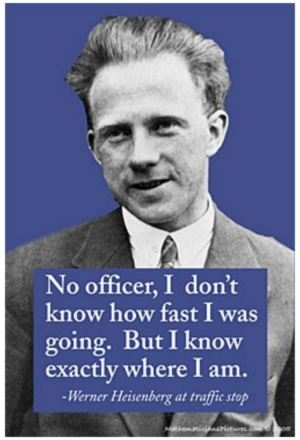
Examples

- **Coin Tosses** - Each toss starts with different positions and velocities → ignoring initial conditions gives illusion of “random” results

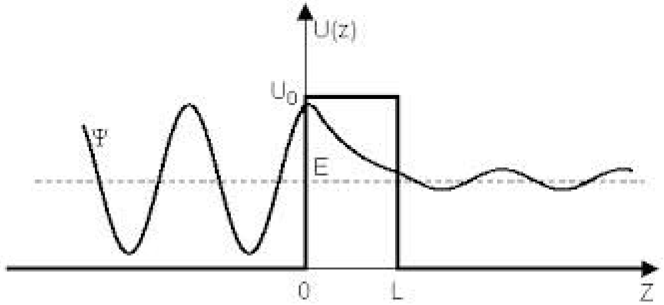
- **Gas molecules in a balloon** – $\sim 10^{23}$ molecules → Each molecule has 6 associated numbers (3 positions, 3 velocities) → Need $\sim 10^{12}$ TB to store state AT EACH INSTANT IN TIME → Average over molecules and work with macroscopic quantities like Temperature, Pressure, Volume

- **Stock Market** – extremely complex system

- **Quantum Mechanics** – Nature actually is fundamentally probabilistic! No hidden variables that can make it deterministic, even in principle.



$$\Delta x \Delta p \geq \frac{\hbar}{2}$$



Non-zero probability particle ends up on other side of barrier

Red Hat

Why Probability?

- Modeling, data science, analytics, data mining (many other buzz words) require appreciating the fragility of “insights” drawn from data.

- Probability and Statistics play a central and crucial role in measuring our degree of confidence.

- We are trying to model systems that might not even follow well-defined “laws” or “rules”. Alternatively, systems might be so complex that it might not be possible to infer the exact rules.

Red Hat

Why Probability? - Abstract View

Have system with N (independent) underlying inputs:

$$x_1, x_2, \ldots, x_N$$

Experimenter measures output value y. Presumably, there's a function:

$$y = f(x_1, x_2, \ldots, x_N)$$

Goal: Find this function

Why Probability? - Abstract View

N large and can reasonably measure only a few inputs, say x_1, x_2, x_3

$$y = g(x_1, x_2, x_3)$$

g approximation to **f**

Instead of:

$$y = f(x_1, x_2, \ldots, x_N)$$

Why Probability? - Abstract View

Instead of:

x1	x2	x3	Value
1	2	0	10
4	3	2	5

One output for each fixed input

Red Hat

Why Probability? - Abstract View

Instead of:

x1	x2	x3	Value
1	2	0	10
4	3	2	5

One output for each fixed input

Get:

x1	x2	x3	Value
1	2	0	10 (5/10) 4 (1/10) 12 (4/10)
4	3	2	5 (2/10) 6 (6/10) 4 (2/10)

Multiple outputs for each fixed input!

Red Hat

Why Probability? - Abstract View

Incomplete knowledge/Ignorance leads to probabilities!

This ignorance can be self-imposed:

Give up **unnecessary detail** and simplify system at the cost of introducing probabilities

Basic Rules

$$0 \leq p \leq 1$$

No chance

Certain

Basic Rules

Discrete

$$\Sigma p_i = 1$$

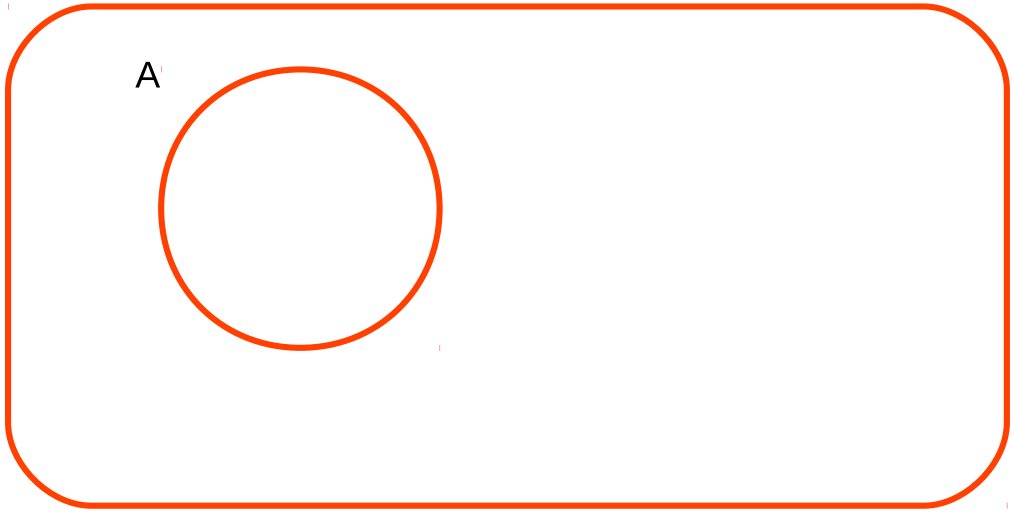
Continuous

$$\int p(x)dx = 1$$

Total Probability = 1
(something has to happen)

Red Hat

Basic Rules



Sample Space, S

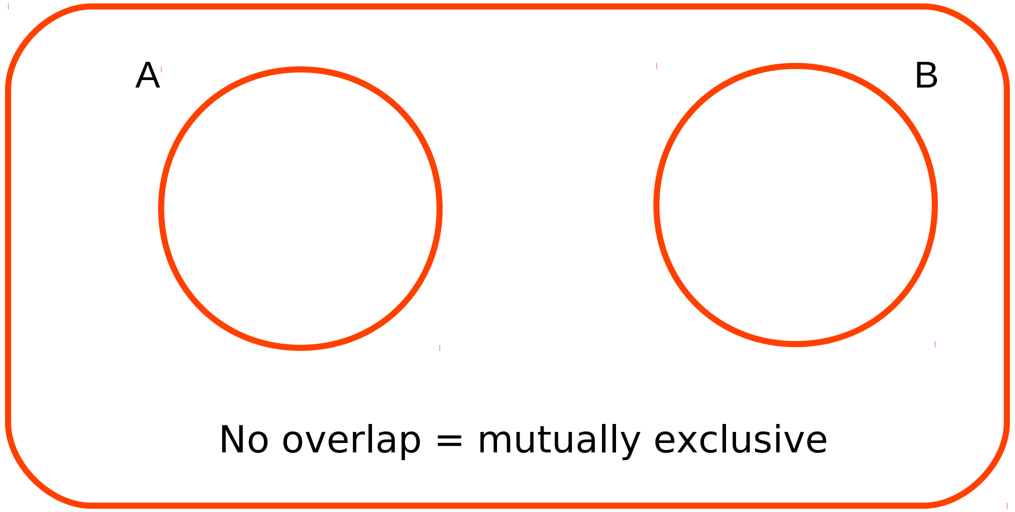
Event: A

$$p(A) = \text{Area of A}$$

$$p(S) = 1 \implies p(A) \leq 1$$

Red Hat

Basic Rules

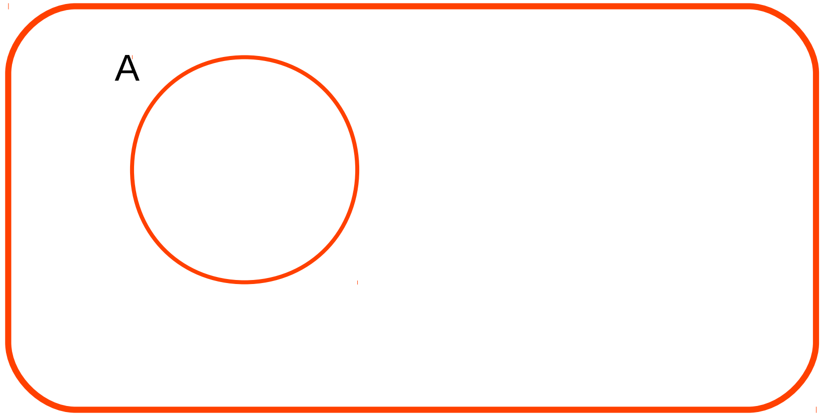


Sample Space, S

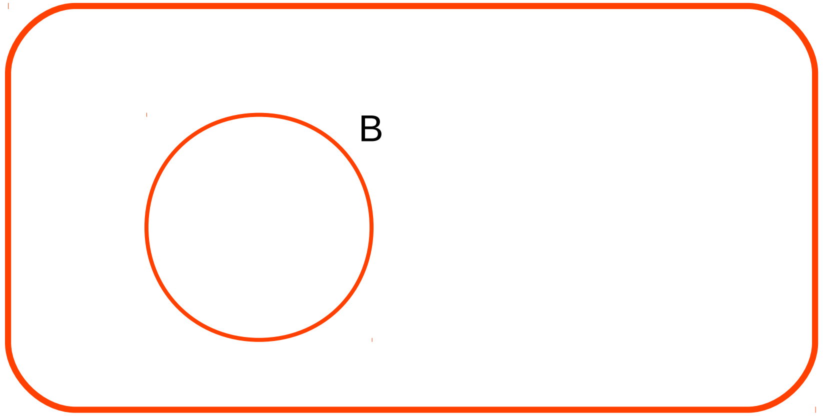
Events: A, B

$$\text{p(A or B)} = \text{p(A)} + \text{p(B)}$$

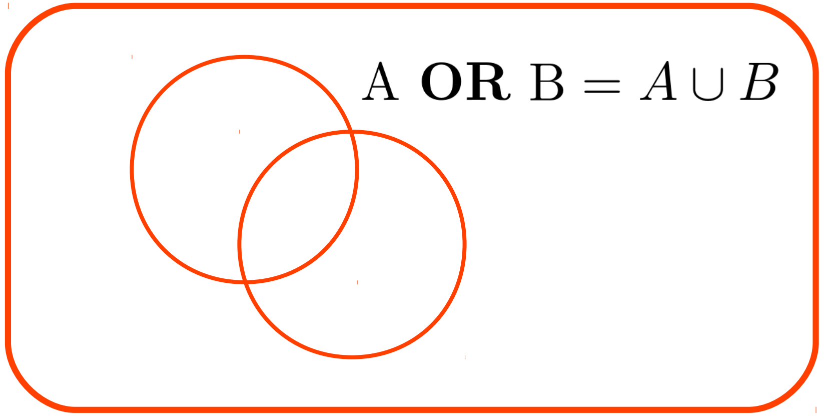
if A, B mutually exclusive



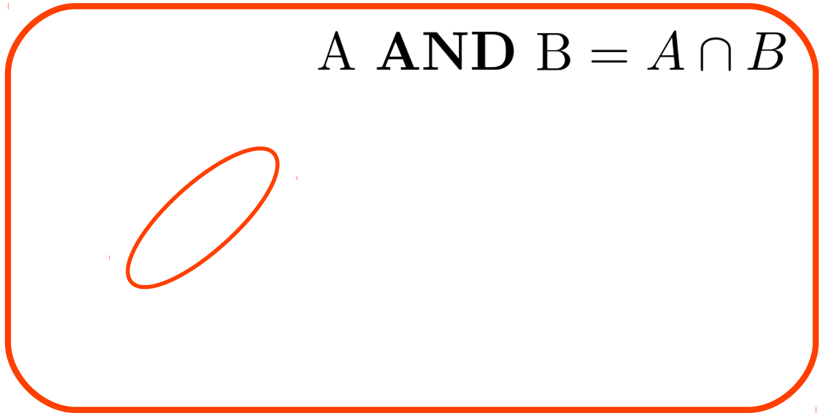
+



$$\mathrm{p(A) + p(B) = p(A\ and\ B) + p(A\ or\ B)}$$



+



Basic Rules

$$\mathrm{p(A\ and\ B) = p(A)p(B)}$$

if A, B independent

Independence = A cannot influence B and vice-versa

(usually an assumption based on knowledge
of field of application i.e. domain)

Basic Rules

If p(A) is known, then p(not A) = 1 - p(A)

Example: F1 or Gas Guzzler

Not open

Not open

Not open

Hint: you should want this one

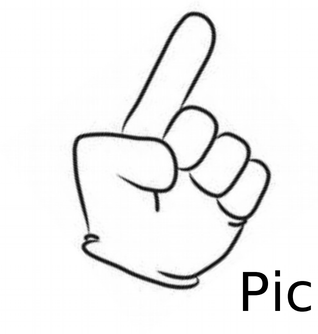
Red Hat

Step 1: You pick a door. But don't open!

Not open

Not open

Not open



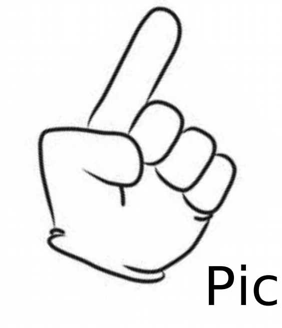
Picked this door

Step 2: Host opens an undesirable door

Not open

Open

Not open



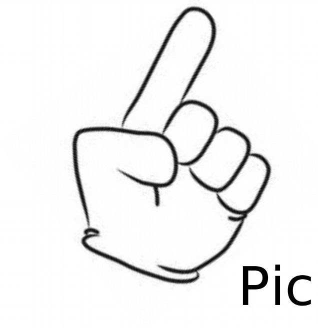
Picked this door

Step 3: You have option to switch to unopened door or stay with current door. Then, the doors are opened. Do you switch?

Not open

Open

Not open



Picked this door

Step 3: You have option to switch to unopened door or stay with current door. Then, the doors are opened. Do you switch?

Not open

Open

Not open

Intuitive answer:

Two doors – one with prize and one without – 50% probability each – doesn't matter if switch or not switch

Possibility 1: You picked undesirable door
Probability = 2/3

Not open

Open

Not open

Switch and win 2/3 of the time

Possibility 2: You picked winning door
Probability = 1/3

Not open

Open

Not open

Switch and lose 1/3 of the time

Picked this door

You should switch each time!

There’s a much higher probability you picked the wrong door

Host removes the other wrong door

The only remaining door is the right one

Takeaways:
Probability can be tricky

Our intuition can be very wrong – need deliberate analysis

Understanding a solution is much easier than coming up with one

Example 2: Effective Medical Test

- Invent a new test to detect a rare disease
  - p(disease) = 0.005 = 0.5% (rare!)

- The test returns one of two results:
  - “-” = didn't detect disease
  - “+” = detected disease

- Properties of a good test

Person	Test = +	Test = -
Has disease	Probability High (~1)	Probability Low (~0)
Doesn't have disease	Probability Low (~0)	Probability High (~1)

Red Hat

Example 2: Effective Medical Test

- Want to know

$$p(disease \mid +)$$

If test returns +, what's the probability you have the disease?

where:

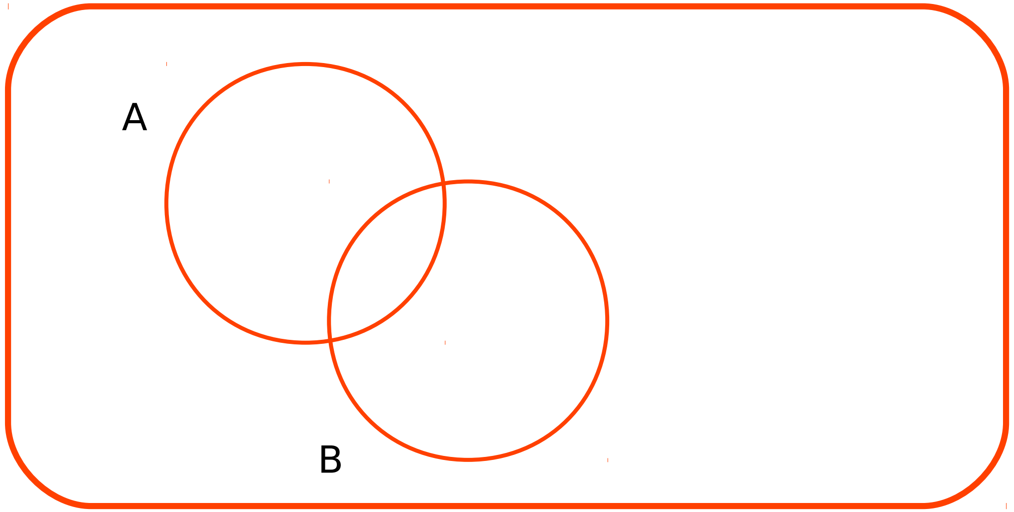
disease = patient has disease

no disease = patient doesn’t have disease

+ = test returns positive result

- = test returns negative result

Aside: Conditional Probability



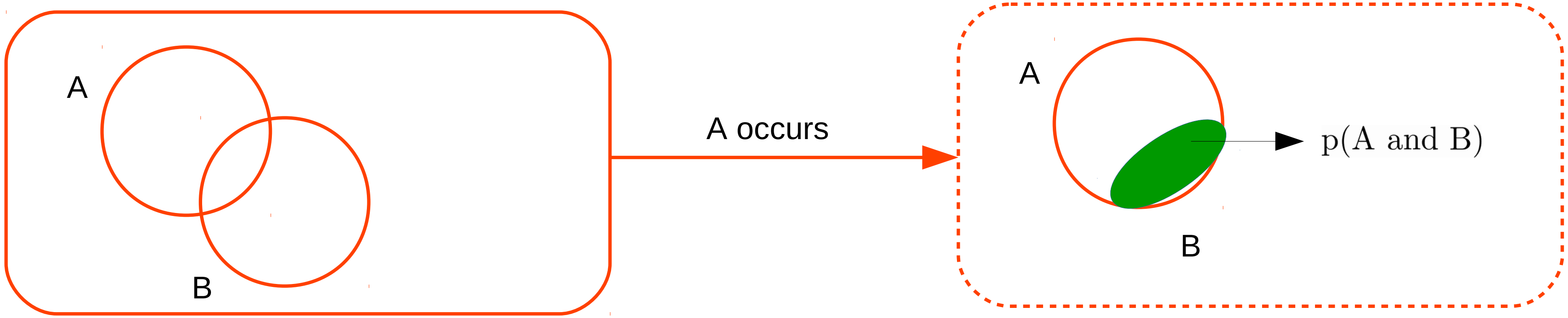
In most real-life cases, we already have some information about the system.

Example: What is the probability that facebook stock will go above $40 given that SP500 is above 1500?

What is the probability of getting heads on a coin toss given that the first 3 tosses gave heads too?

Red Hat

Aside: Conditional Probability



- Suppose, we know event A has happened (SP500 going above 1500).
- What is the probability that event B (facebook crossing 40) will occur?
- Answer: proportion of B's area within A

Read as:
“probability of B given A”

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

Green area

Area under A

- Invent a new test to detect a rare disease
  - p(disease) = 0.005 = 0.5% (rare!)
  - p(no disease) = 1 – p(disease) = 0.995

- Suppose the test is very accurate:
  - p(+|disease) = 0.99
  - P(-|no disease) = 0.98

Person	Test = +	Test = -
Has disease	99%	1%
Doesn't have disease	2%	98%

Person	Test = +	Test = -
Has disease	p(+|disease) = 99%	P(-|disease) = 1%
Doesn't have disease	p(+|no disease) = 2%	P(-|no disease) = 98%

What is the probability of having disease given test is +?

$$p(disease \mid +)$$

Bayes' Theorem

Remember:

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

Let's flip A and B around:

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

$p(A\, and\, B) = p(B\, and\, A)$ means:

$$\boxed{p(B \mid A)p(A) = p(A \mid B)p(B)}$$

Nothing deep here BUT
Very useful relationship that lets us "invert" probabilities!!

Red Hat

Bayes' Theorem

$$p(B \mid A)p(A) = p(A \mid B)p(B)$$

$$\text{Replace: } A \to \theta \text{ (model parameters)}$$

$$B \to \mathcal{D} \text{ (observed data)}$$

$$\boxed{p(\theta \mid \mathcal{D}) = \frac{p(\mathcal{D} \mid \theta)p(\theta)}{p(\mathcal{D})}}$$

Red Hat

Bayes' Theorem

$$p(B \mid A)p(A) = p(A \mid B)p(B)$$

$$\text{Replace: } A \to \theta \text{ (model parameters)}$$

$$B \to \mathcal{D} \text{ (observed data)}$$

$$p(\theta \mid \mathcal{D}) = \frac{p(\mathcal{D} \mid \theta)\boxed{p(\theta)}}{p(\mathcal{D})}$$

Prior $=$ What's your intuition about parameter values?

Bayes' Theorem

$$p(B \mid A)p(A) = p(A \mid B)p(B)$$

$$\text{Replace: } A \to \theta \text{ (model parameters)}$$

$$B \to \mathcal{D} \text{ (observed data)}$$

$$p(\theta \mid \mathcal{D}) = \frac{p(\mathcal{D} \mid \theta)\boxed{p(\theta)}}{p(\mathcal{D})}$$

Example: If think θ between 1 and 4
but have no more information,
use uniform distribution

Bayes' Theorem

$$p(B \mid A)p(A) = p(A \mid B)p(B)$$

$$\text{Replace: } A \to \theta \text{ (model parameters)}$$

$$B \to \mathcal{D} \text{ (observed data)}$$

$$p(\theta \mid \mathcal{D}) = \frac{\boxed{p(\mathcal{D} \mid \theta)}\, p(\theta)}{p(\mathcal{D})}$$

Likelihood = Given model (the function p) what's the total "probability" of observing the data you do observe?

Red Hat

Bayes' Theorem

$$p(B \mid A)p(A) = p(A \mid B)p(B)$$

Replace: $A \rightarrow \theta$ (model parameters)

$B \rightarrow \mathcal{D}$ (observed data)

$$p(\theta \mid \mathcal{D}) = \frac{\boxed{p(\mathcal{D} \mid \theta)}\, p(\theta)}{p(\mathcal{D})}$$

Example: N independent coin tosses with probability θ of getting heads

Observe m heads

$$\textbf{Likelihood} = \binom{N}{m} \theta^m (1-\theta)^{N-m}$$

Bayes' Theorem

$$p(B \mid A)p(A) = p(A \mid B)p(B)$$

Replace: $A \to \theta$ (model parameters)

$B \to \mathcal{D}$ (observed data)

$$p(\theta \mid \mathcal{D}) = \frac{p(\mathcal{D} \mid \theta)p(\theta)}{\boxed{p(\mathcal{D})}}$$

Normalization = To ensure left-hand side is a probability (sums to 1)

Red Hat

Bayes' Theorem

$$p(B \mid A)p(A) = p(A \mid B)p(B)$$

$$\text{Replace: } A \to \theta \text{ (model parameters)}$$

$$B \to \mathcal{D} \text{ (observed data)}$$

$$\boxed{p(\theta \mid \mathcal{D})} = \frac{p(\mathcal{D} \mid \theta)p(\theta)}{p(\mathcal{D})}$$

Posterior = Probability distribution of model parameters after observing data

Red Hat

Bayes' Theorem

$$p(\theta \mid \mathcal{D}) = \frac{p(\mathcal{D} \mid \theta)p(\theta)}{p(\mathcal{D})}$$

<u>Read as</u>:

Have some initial guess about distribution for θ

Model with parameters θ predicts probability to observe gathered data

$$p(\mathcal{D} \mid \theta)$$

Bayes' theorem lets us update θ to get a new distribution

$$p(\theta \mid \mathcal{D})$$

REPEAT with new data

Red Hat

Bayes' Theorem

$$p(\theta \mid \mathcal{D}) = \frac{p(\mathcal{D} \mid \theta)p(\theta)}{p(\mathcal{D})}$$

Repeat with new data, $\mathcal{D}'$ but use p$(\theta|\mathcal{D})$ in place of $p(\theta)$

$$p(\theta \mid \mathcal{D}', \mathcal{D}) = \frac{p(\mathcal{D}' \mid \theta, \mathcal{D})p(\theta \mid \mathcal{D})}{p(\mathcal{D}')}$$

Bayes’ Theorem

- What if two events, B and C, that are mutually exclusive (p(B and C)=0), always occur in conjunction with A (p(B) + p(C) = 1).
- Then,

$$\boxed{p(A) = p(A \mid B)p(B) + p(A \mid C)p(C)}$$

Mutually Exclusive: B happens or C happens but not both
Cover the space: Either B or C happens but not none

Red Hat

Effective Test?

Test returned positive – does patient have disease?

99% = very accurate test

$$p(disease \mid +) = \frac{p(+ \mid disease)p(disease)}{p(+)}$$

0.5% = very rare disease

No idea?

Effective Test?

Remember:

$$p(A) = p(A \mid B)p(B) + p(A \mid C)p(C)$$

Replace some letters:

$$p(+) = p(+ \mid \underbrace{disease}_{99\%})p(\underbrace{disease}_{0.5\%}) + p(+ \mid \underbrace{no\,disease}_{2\%})p(\underbrace{no\,disease}_{99.5\%})$$

Red Hat

Effective Test?

$$p(disease \mid +) = \frac{p(+ \mid disease)p(disease)}{p(+ \mid disease)p(disease) + p(+ \mid no\, disease)p(no\, disease)}$$

$$p(disease \mid +) = \frac{99\% * 0.5\%}{99\% * 0.5\% + 2\% * 99.5\%}$$

$$p(disease \mid +) = 19.9\%!!!!$$

Thought test was very good but there's ~80% chance you don't have the disease even if test positive.

Red Hat

Effective Test?

What if disease not rare but affects 50% of the population

$$p(disease \mid +) = \frac{99\% * 50\%}{99\% * 50\% + 2\% * 50\%} = \boxed{98\%!}$$

Good test

Effective Test: Concrete Numbers

- Population = 1000 people
- 5 have disease, 995 don't have disease
- p(+|disease) = 99% → 4.95 people return +
- p(+|no disease) = 2% → 19.9 people return +
- 24.85 people return +
- p(disease|+) = 4.95 / 24.85 = 19.9%

Shortest Introduction to Statistics Ever

- Combination of rigorous results from probability theory and rules-of-thumb from specific examples.

- Rules-of-thumb are assumed to hold in general. If we encounter a data set where they don't, we come up with new rules.

- Coin – 10 tosses gives 7 heads – fair or not?

- Estimate average height of 400 million Americans from 1000 people.

Random Variables and Probability Distributions

A variable that stores the result of an experiment

Example:

Coin Toss: X = 0 (heads), X = 1(tails)

Height Measurement: X = 178 cm, X = 165 cm

PDFs and Likelihood

- Probability distribution functions (p.d.f.):
  - Table of numbers if outcomes discrete:

3-sided Dice Outcome	Probability
1	1/10
2	5/10
3	4/10

PDFs and Likelihood

- Probability distribution functions (p.d.f.):
  - Table of numbers if outcomes discrete:

3-sided Dice Outcome	Probability
1	1/10
2	5/10
3	4/10

  - Continuous outcomes?

$p(\vec{x}) \longrightarrow \vec{x}$ describes continuous outcomes

Red Hat

PDFs and Likelihood

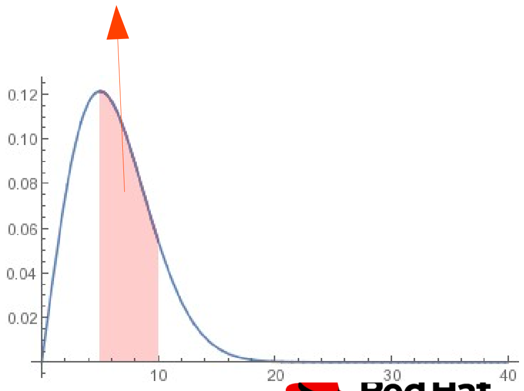
- Probability distribution functions (p.d.f.):
  - Continuous outcomes? Cannot tabulate since uncountably infinite possibilities

$$p(\vec{x}) \longrightarrow \vec{x} \quad \text{describes continuous outcomes}$$

  - 1-d version

$$p(x) \longrightarrow \text{Single outcome}$$

Red area = prob. observation between 5 and 10



Intepretation

$$p(x)dx = \text{Probability outcome between } x \text{ and } x + dx$$

$$p(\vec{x})\vec{dx} = \text{Probability outcome between } \vec{x} \text{ and } \vec{x} + \vec{dx}$$

PDFs and Likelihood

- Total probability = 1 →

$$\int p(x)dx = 1$$

$$\text{Probability outcome between } x \text{ and } x + dx$$

$$\int p(\vec{x})d\vec{x} = 1$$

$$\text{Probability outcome between } \vec{x} \text{ and } \vec{x} + d\vec{x}$$

= integral/sum over all outcomes

PDFs and Likelihood

- Parameters: Generally the p.d.f depends on not just the outcome but some other parameters

- Example: Gaussian/normal distributions have a mean and variance parameter (real numbers) that can be tuned externally (you pick what they are).

- This is often denoted as:

$$p(\vec{x}; \vec{\theta})$$

or

$$p_{\vec{\theta}}(\vec{x})$$

Red Hat

PDFs and Likelihood

- Another way of thinking about this:

$$\text{For each } \vec{\theta}, \text{ we get a different p.d.f. } p_{\vec{\theta}}(\vec{x})$$

- $\vec{\theta}$ describes a **family** of p.d.f. or informally "distributions"

- Often goal in statistics is to infer the $\vec{\theta}$ value that the data follows

Short Philosophical Rant

- Probability is not just a tool for calculating but a way of thinking.

- Probability abhors strong statements and opinions unless there’s overwhelming evidence.

- So strong (and stupid) statements like the ones below (specially a disease I have noticed in the software world) will raise red flags in a probabilistic/mathematical mind.

  This is the best coffee/(food of your choice)
  (What the speaker is trying to say: “Out of all the coffees I have tried, I like this one the most)

  All people from Country X are Y

  Republicans are X or Democrats are Y (sorry I am American)

Red Hat

Short Philosophical Rant

- Mathematically and logically, blanket statements are easy to disprove. I need just one counterexample. Generally one counterexample is indicative of many more and point to a gap in one’s experience.

- In daily life, we speak casually and one shouldn’t be precise with every sentence obviously.

- But probability and mathematics encourages and trains one to only make very strong statements if the opposite is obviously false (no evidence or provably false). General tip: strong statements spoken with utmost confidence are usually wrong, specially about matters in real life.

Red Hat

Short Philosophical Rant

… Neils Bohr divided true statements into two classes: the trivial ones and those of genius. Specifically, he regarded a true statement as trivial when the opposite statement is obviously false, and a true statement as genius when the opposite statement is just as non-obvious as the original, so that the question of truth of the opposite statement is interesting and worth studying.

- V.I. Arnold
Mathematical Understanding of Nature

Short Philosophical Rant

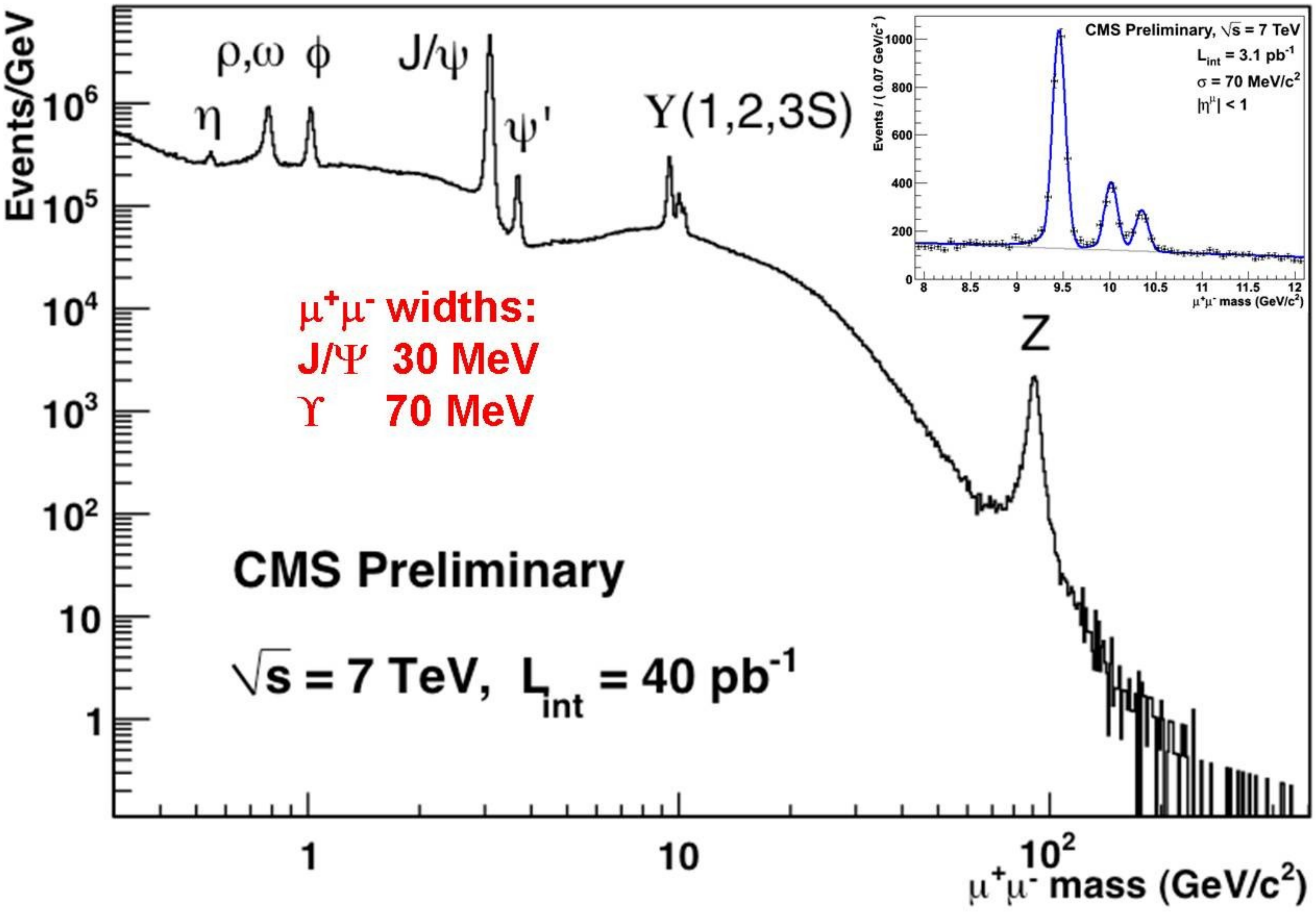
The first principle is that you must not fool yourself – and you are the easier person to fool

- R.P. Feynman

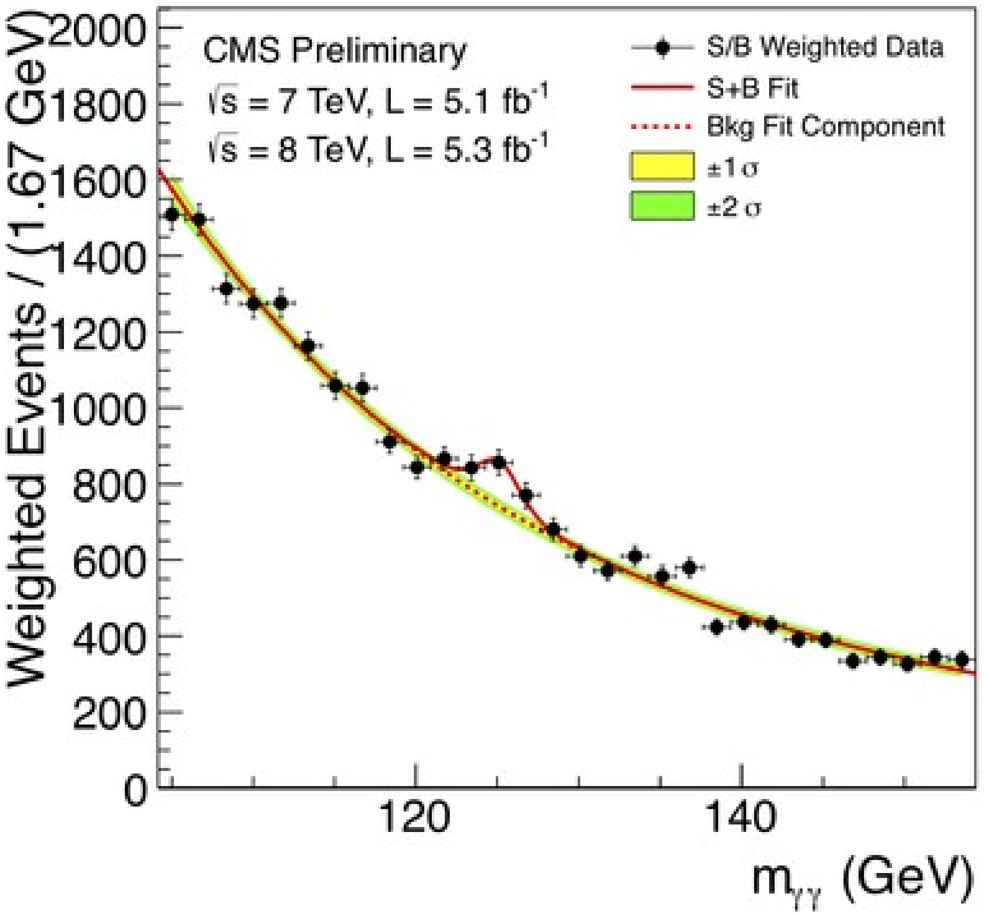
Cargo Cult Science (very highly recommended)

[http://calteches.library.caltech.edu/51/2/CargoCult.htm]

Examples

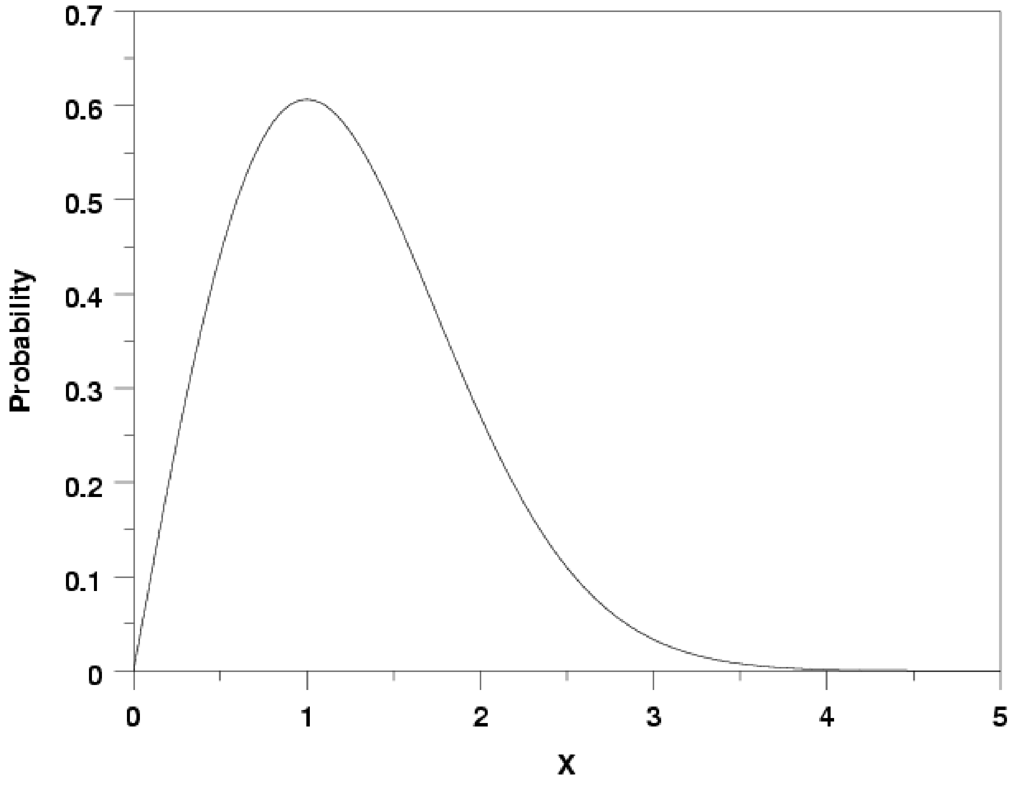

Events/GeV
ρ,ω
ϕ
J/ψ
η
ψ'
Υ(1,2,3S)
Z
μ⁺μ⁻ widths:
J/Ψ 30 MeV
Υ 70 MeV
CMS Preliminary
√s = 7 TeV, L_int = 40 pb⁻¹
μ⁺μ⁻ mass (GeV/c²)
CMS Preliminary, √s = 7 TeV
L_int = 3.1 pb⁻¹
σ = 70 MeV/c²
|η^μ| < 1
Events / (0.07 GeV/c²)
μ⁺μ⁻ mass (GeV/c²)

Examples



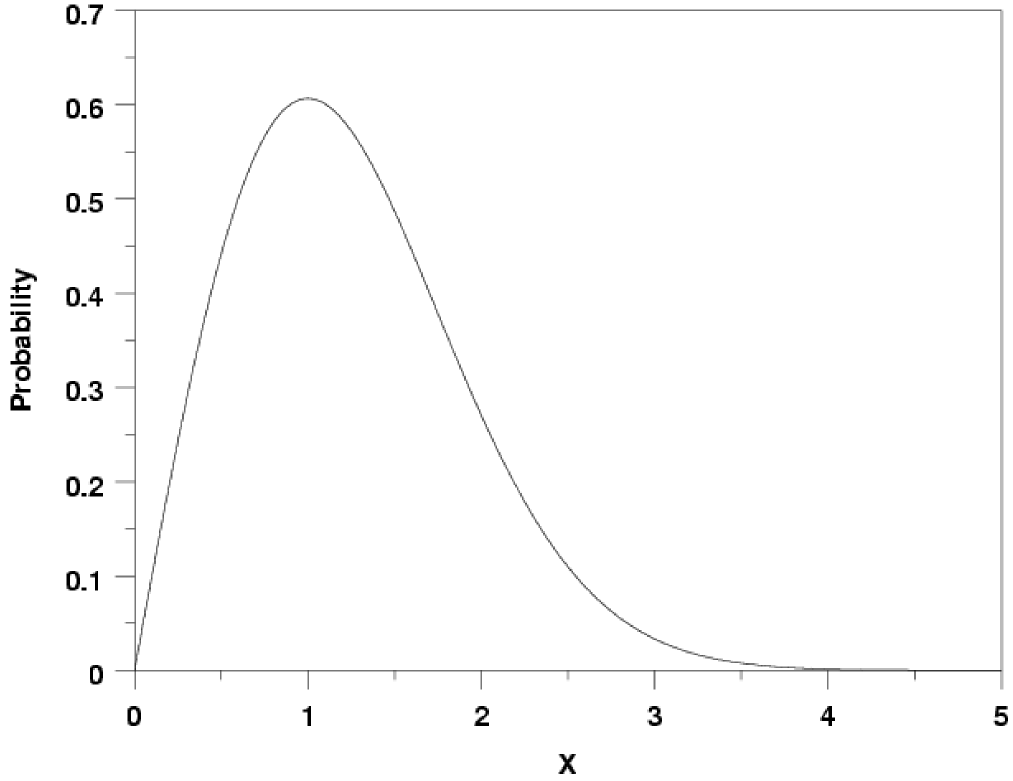
Higgs →2 photons

Summarizing Probability Distributions



Too much detail – can we *approximately* summarize this shape by a few numbers?

Summarizing Probability Distributions

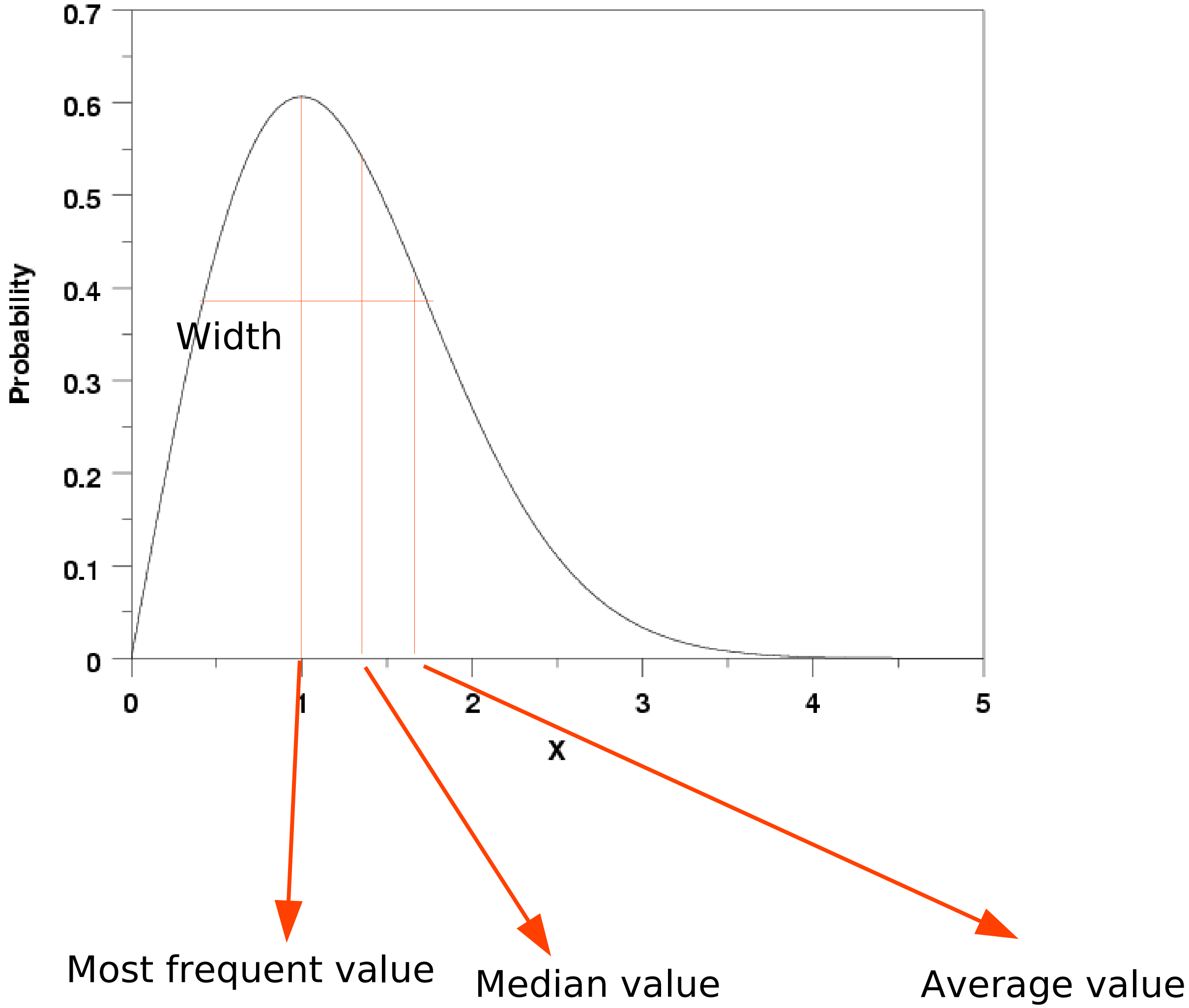


$$\mu = \text{Average/Mean/Expected Value}$$

$$\sigma = \text{Standard Deviation/Root-mean-squared value} = \text{"width"}$$

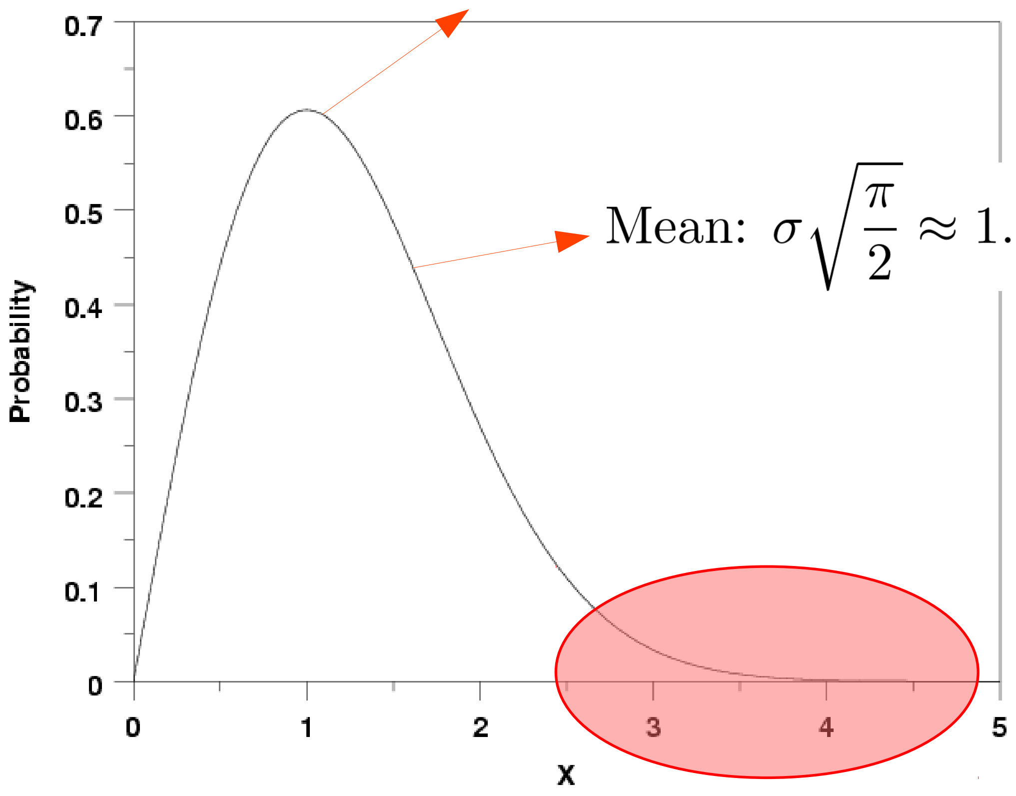
Higher moments: characterize effects of tails

Summarizing Probability Distributions



Summarizing Probability Distributions

Most frequent value: σ



Mean: $\sigma\sqrt{\frac{\pi}{2}} \approx 1.25\sigma$

Asymmetric Tail:
Shifts mean to the right of the peak

Summarizing Probability Distributions: Width or Standard Deviation

Variance

$$Var(X) = Average[(X - \mu)^2]$$

Standard Deviation

$$sd(X) = \sqrt{Var(X)}$$

The next few slides give a taste of some statistical topics.

They are not comprehensive at all.

Red Hat

Measurement of Parameters

- Repeat experiment N times and get results $x_1, x_2, \ldots, x_N$
- These values follow some distribution that has an average μ and a standard deviation σ. **We don't know μ and σ**!
- Can we get an **estimate of μ and σ** from our measurements?

Red Hat

Measuring Averages

- Estimate of the mean:

$$\boxed{\mu_{estimate} = \frac{x_1 + \ldots + x_N}{N}}$$

- How close is this to the actual but unknown μ

$$\boxed{Average[(\mu_{estimate} - \mu)^2] = \frac{\sigma^2}{N}}$$

Estimate from data

Actual mean but unknown

More measurements send this to 0

Red Hat

Measuring Standard Deviation

Variance

$$Var(X) = Average[(X - \mu)^2]$$

Standard Deviation

$$sd(X) = \sqrt{Var(X)}$$

Red Hat

Measuring Standard Deviation

Estimate of Variance

$$\sigma^2_{estimate} = \frac{(x_1 - \mu_{estimate})^2 + \ldots + (x_N - \mu_{estimate})^2}{N}$$

Biased – gives values that are too low!

$$\boxed{\sigma^2_{estimate} = \frac{(x_1 - \mu_{estimate})^2 + \ldots + (x_N - \mu_{estimate})^2}{N - 1}}$$

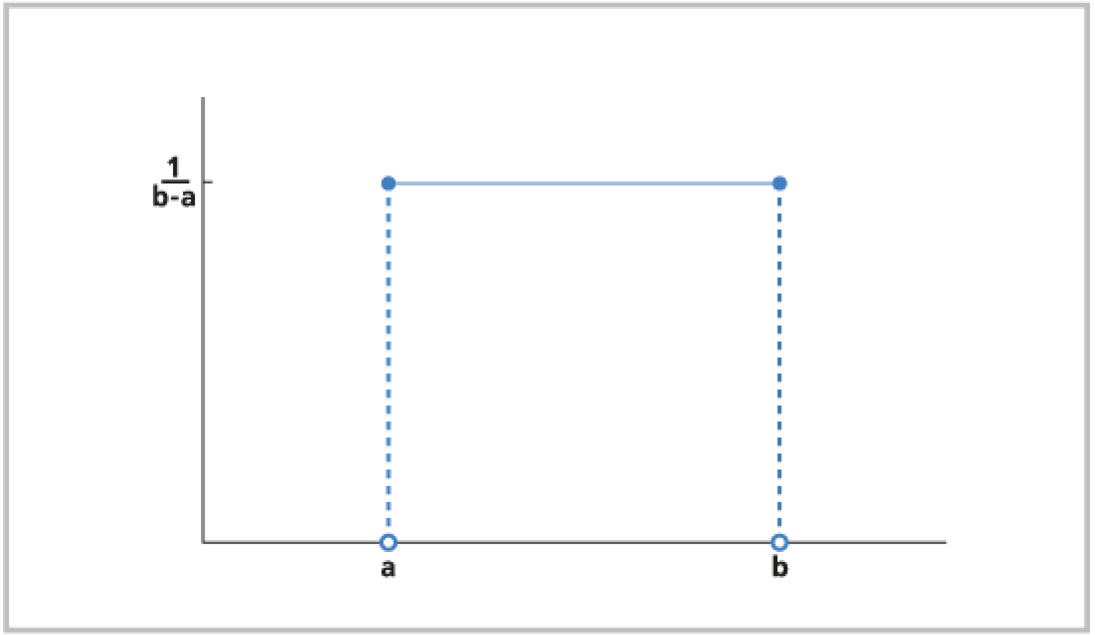
Unbiased Variance!

$sd_{estimate} = \sqrt{\sigma^2_{estimate}}$ biased though - bias negligible for $N > 10$

Bias drops off as $\frac{1}{N}$ so large N has small bias

Red Hat

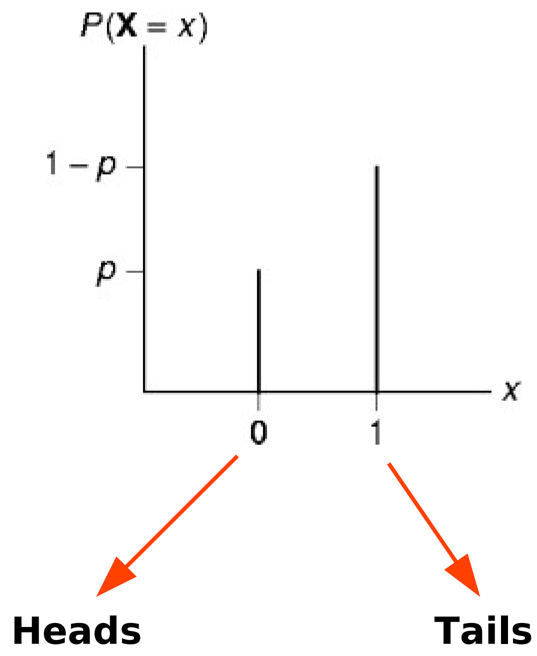
Examples of Distributions: Uniform



$$\mu = \frac{a+b}{2}$$

$$\sigma^2 = \frac{(b-a)^2}{12}$$

Examples of Distributions: Bernoulli



Model any two-state system

Red Hat

Examples of Distributions: Binomial

Given n occurrences in a two-state system(0/1),

what's the probability of observing m occurences of 1

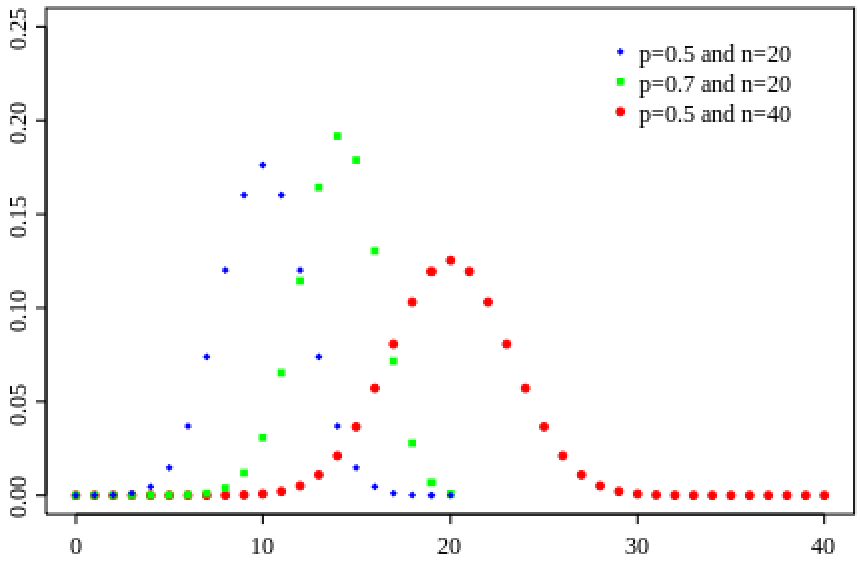
Total number of possibilities:
0011
0110
0101
1010
...

$$\binom{n}{m} p^m (1-p)^{n-m} \qquad m = 0, 1, \ldots, n$$

m occurrences of type 1

N-m occurrences of type 0

Examples of Distributions: Binomial

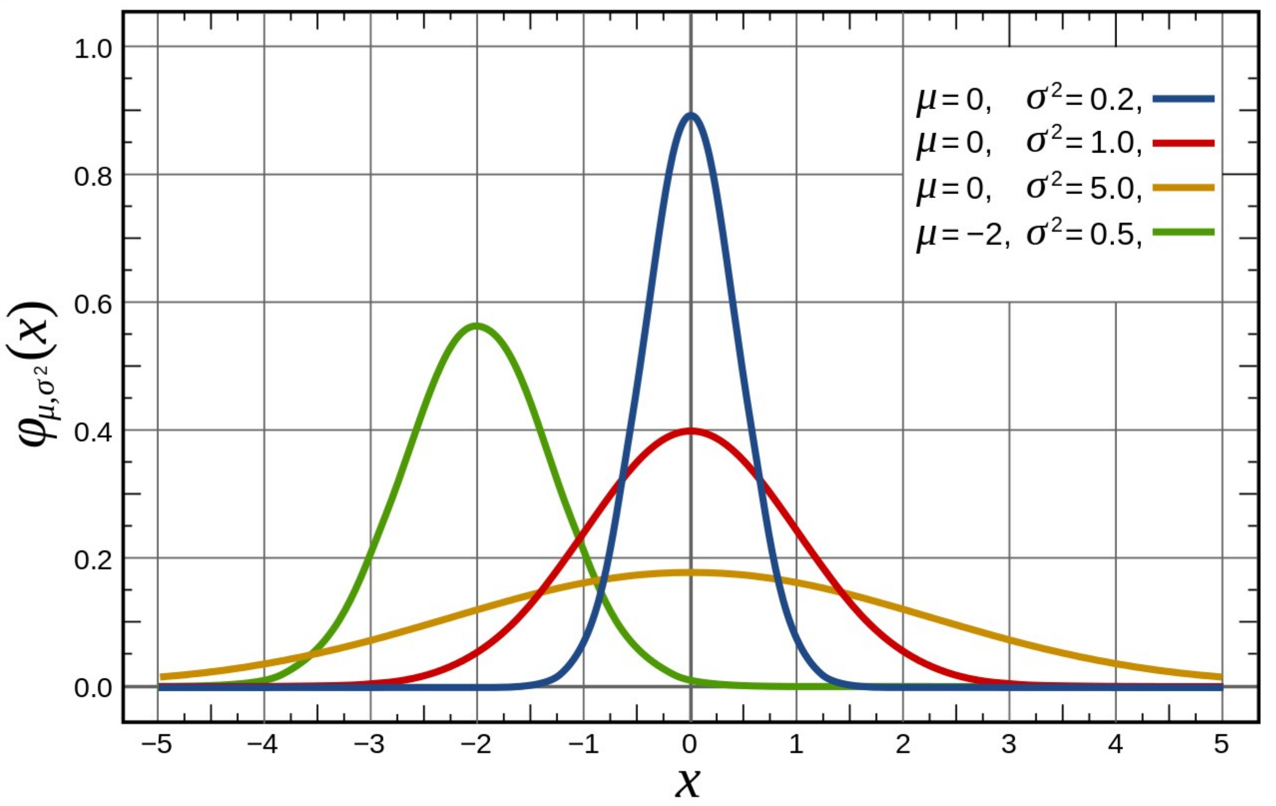


$$\binom{n}{m} p^m (1-p)^{n-m} \qquad m = 0, 1, \ldots, n$$

→ becomes Gaussian for large n

$$\mu = Np \qquad \sigma = \sqrt{Np(1-p)}$$

Examples of Distributions: Gaussian

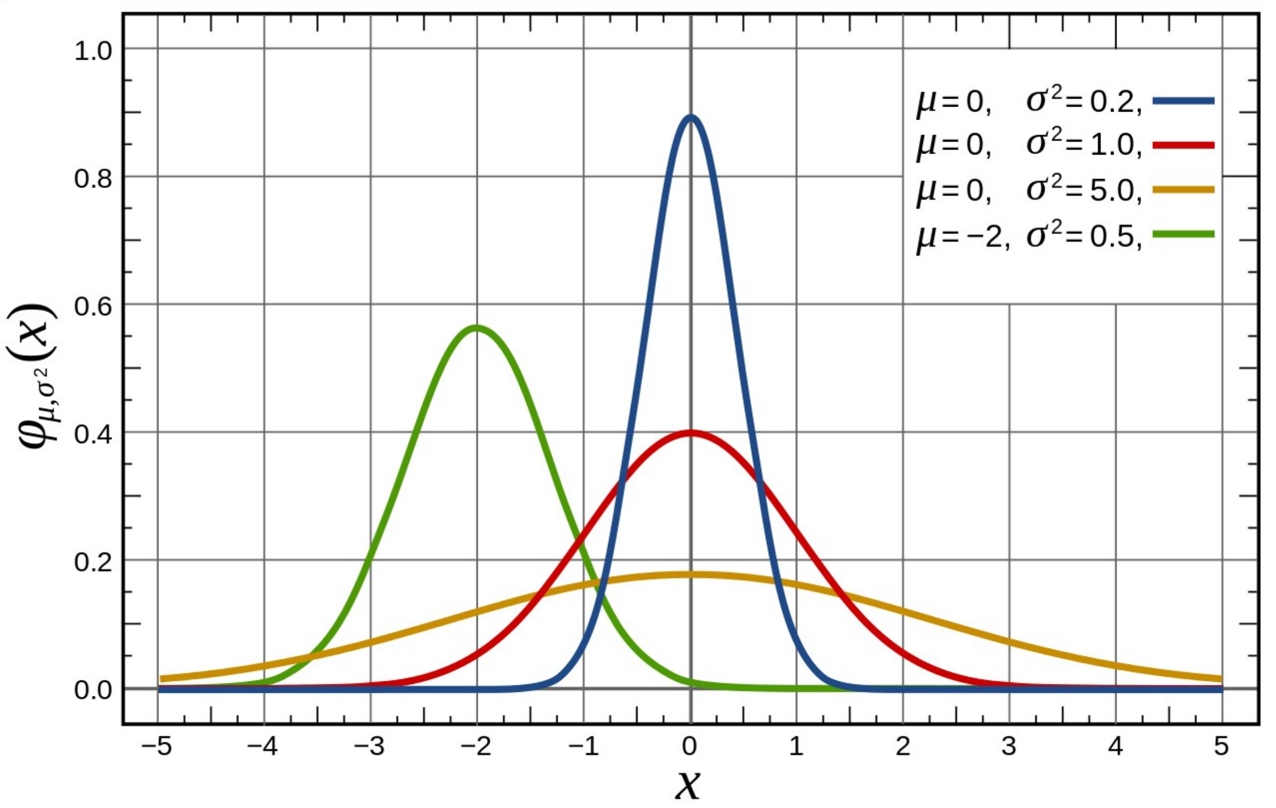


$$\frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$$

$$\mu = \text{ mean(center) of distribution}$$

$$\sigma = \text{ standard deviation(width) of distribution}$$

Examples of Distributions: Gaussian

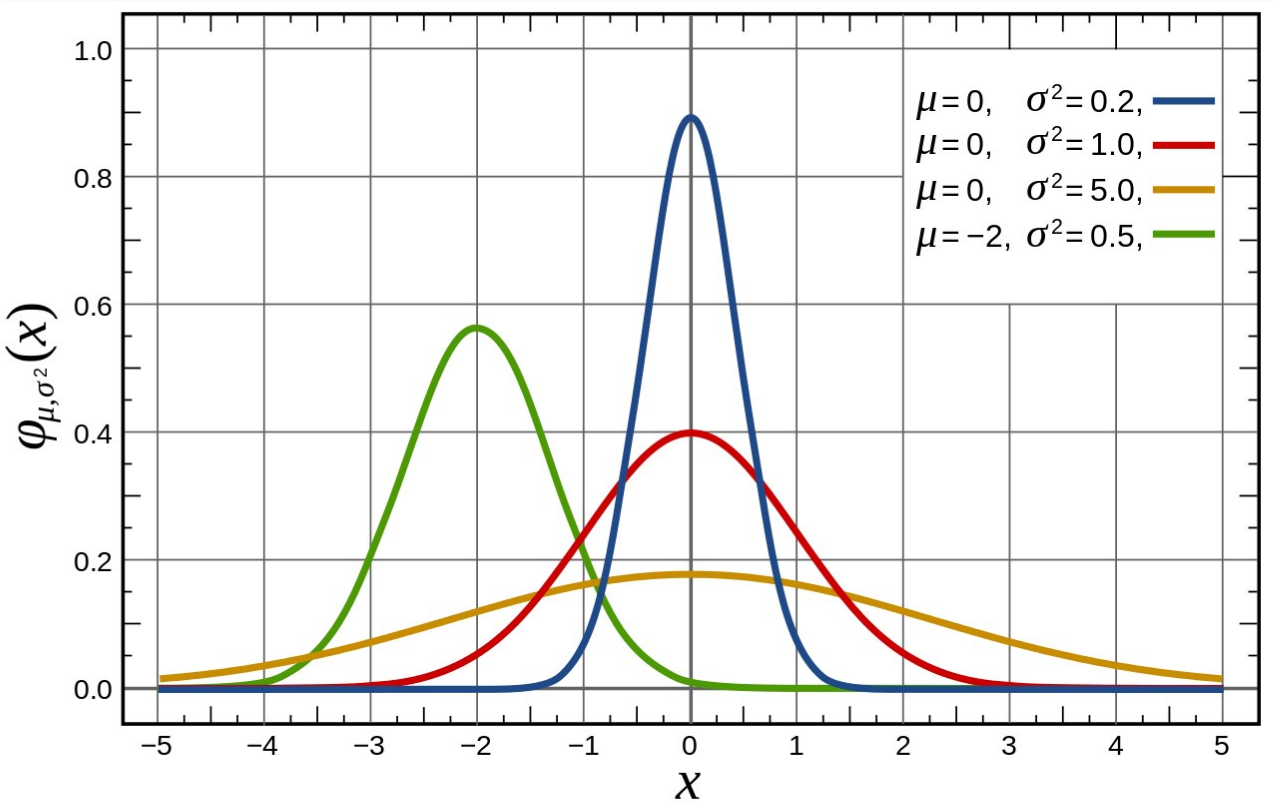


Advantages:

Mean, standard deviation independent parameters

Mathematically easy to work with

Examples of Distributions: Gaussian



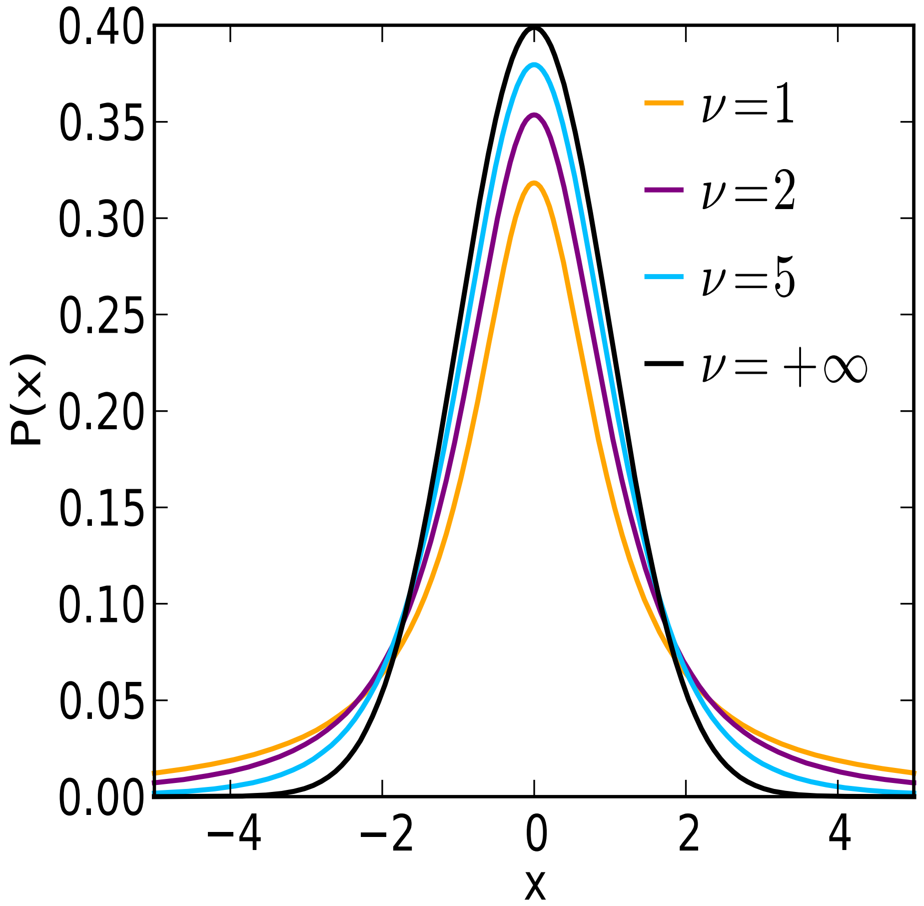
Disadvantage:

Measured quantity might have tails that deviate significantly from Gaussian tails which leads to over/under-estimating probability of tail events

Examples of Distributions: Student-t

$$\frac{\Gamma(\frac{\nu+1}{2})}{\sqrt{\nu\pi}\,\Gamma(\frac{\nu}{2})}\left(1+\frac{x^2}{\nu}\right)^{-\frac{\nu+1}{2}}$$

$$\nu = \text{degree of freedom}$$

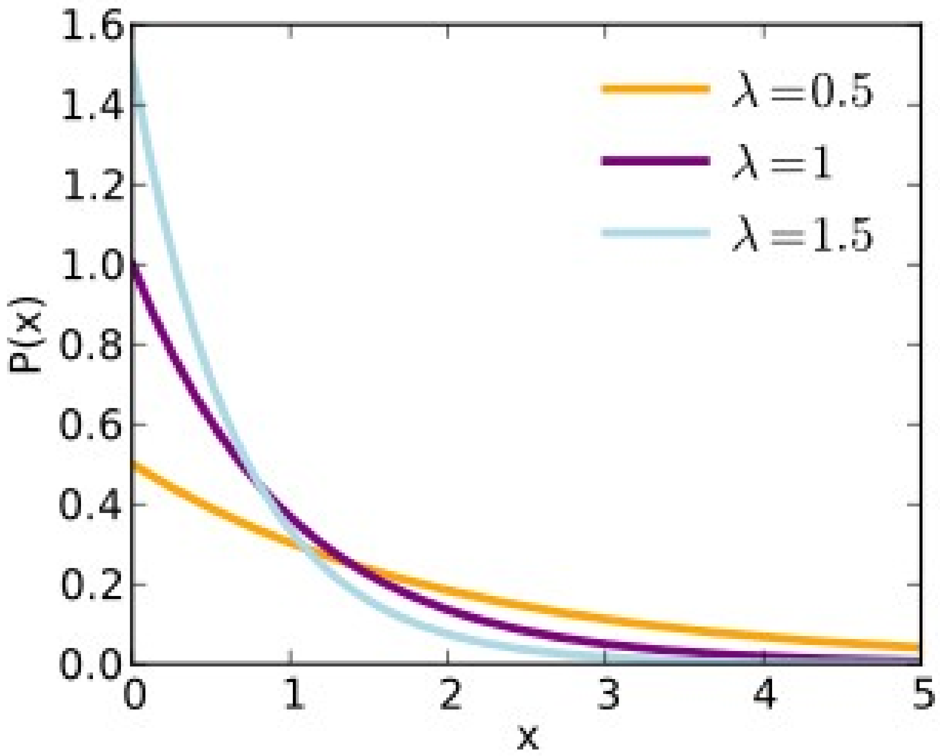


Normal Distribution not correct for small sample size

$\lim_{\nu\to\infty}$ of T-Distribution is Gaussian Distribution

Red Hat

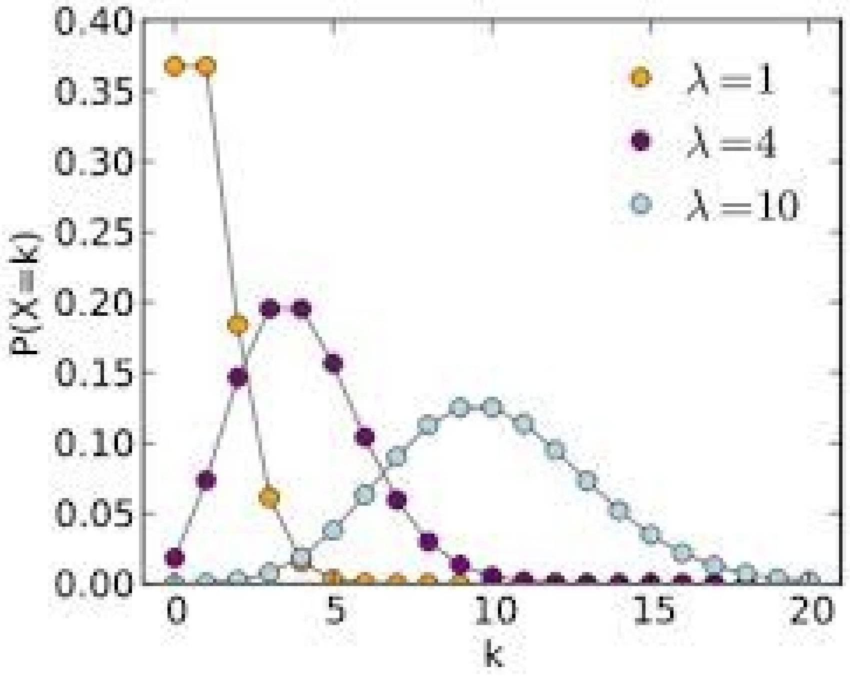
Examples of Distributions: Exponential



$$\lambda e^{-\lambda t}$$

- Have special incidents ("failures") and want to **model the lifetime** (some caveats here).
- Simplest version assumes probability of failure is equal through lifetime – easily generalized to varying probability of failure.

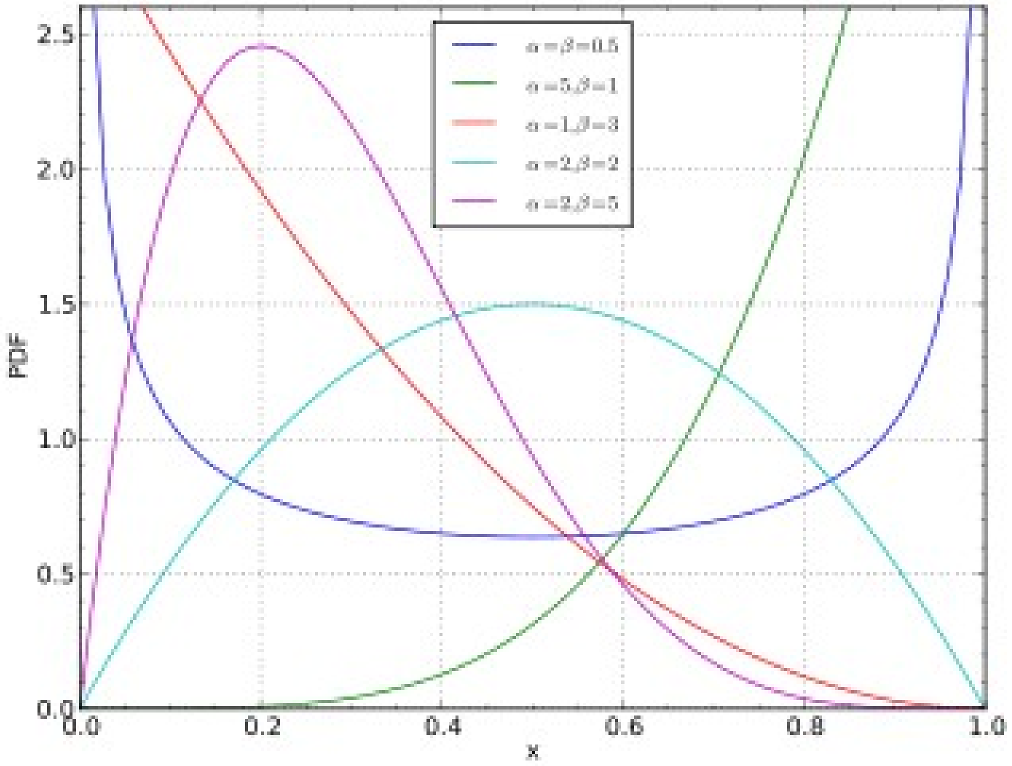
Examples of Distributions: Poisson



$$\frac{e^{-\lambda}\lambda^{n}}{n!}$$

- Counting **number of events** (e.g. failures) in a certain time period.
- Underlying assumption – lifetimes of failing parts follows **exponential distribution**.

Examples of Distributions: Beta



$$\frac{x^{\alpha-1}(1-x)^{\beta-1}}{B(\alpha,\beta)}$$

Normalization constant

Examples of Distributions: Beta

Recall Bayes:

$$p(\theta \mid \mathcal{D}) = \frac{p(\mathcal{D} \mid \theta)p(\theta)}{p(\mathcal{D})}$$

$$\text{Prior, } p(\theta) = Beta(\alpha, \beta) = \frac{\theta^{\alpha-1}(1-\theta)^{\beta-1}}{B(\alpha, \beta)}$$

<u>Experiment: See m 1s and n 0s</u>

Examples of Distributions: Beta

Recall Bayes:

$$p(\theta \mid \mathcal{D}) = \frac{p(\mathcal{D} \mid \theta)p(\theta)}{p(\mathcal{D})}$$

$$\text{Prior, } p(\theta) = Beta(\alpha, \beta) = \frac{\theta^{\alpha-1}(1-\theta)^{\beta-1}}{B(\alpha, \beta)}$$

<u>Experiment: See m 1s and n 0s</u>

$$\text{Likelihood, } p(\mathcal{D} \mid \theta) = \binom{n+m}{m}\theta^{m}(1-\theta)^{n}$$

Examples of Distributions: Beta

Recall Bayes:

$$p(\theta \mid \mathcal{D}) = \frac{p(\mathcal{D} \mid \theta)p(\theta)}{p(\mathcal{D})}$$

$$\text{Prior, } p(\theta) = Beta(\alpha, \beta) = \frac{\theta^{\alpha-1}(1-\theta)^{\beta-1}}{B(\alpha, \beta)}$$

$$\text{Likelihood, } p(\mathcal{D} \mid \theta) = \binom{n+m}{m}\theta^{m}(1-\theta)^{n}$$

$$\text{Posterior, } p(\theta \mid \mathcal{D}) \propto \text{Prior} * \text{Likelihood} \propto \binom{n+m}{m}\theta^{m+\alpha-1}(1-\theta)^{n+\beta-1}$$

Examples of Distributions: Beta

Recall Bayes:

$$p(\theta \mid \mathcal{D}) = \frac{p(\mathcal{D} \mid \theta)p(\theta)}{p(\mathcal{D})}$$

$$\text{Prior, } p(\theta) = Beta(\alpha, \beta) = \frac{\theta^{\alpha-1}(1-\theta)^{\beta-1}}{B(\alpha, \beta)}$$

$$\text{Likelihood, } p(\mathcal{D} \mid \theta) = \binom{n+m}{m}\theta^{m}(1-\theta)^{n}$$

Beta distribution again!!!! → Use as prior for next round of data

$$\text{Posterior, } p(\theta \mid \mathcal{D}) \propto \text{Prior} * \text{Likelihood} \propto \binom{n+m}{m}\theta^{m+\alpha-1}(1-\theta)^{n+\beta-1}$$

One more thing: Central Limit Theorem

Suppose we want to measure the average height of everyone in Brno

Pick N people randomly

Measure person i's height: X_i

Can think of X_i as just a number

BUT it is actually a sample from the underlying distribution/p.d.f. of heights

One more thing: Central Limit Theorem

$$\text{Average height: } \hat{\mu} = \frac{X_1 + X_2 + \ldots + X_N}{N}$$

Question: how close is this to the actual mean of the population?!

Question: If I picked a different sample of N people,
how different would the answer be?

Proof

$$\text{Answer:} \qquad \hat{\mu} \to \mathcal{N}(\mu, \frac{\sigma^2}{N})$$

$$\mu = \mathbb{E}(X_i) \qquad \sigma^2 = \mathrm{Var}(X_i)$$

Technical assumptions: μ, σ^2 finite

X_i are independent, identically distributed (can be relaxed)

Mathematical Optimization

What is Optimization?

Mathematical and numerical techniques to find the maximum or minimum of a function

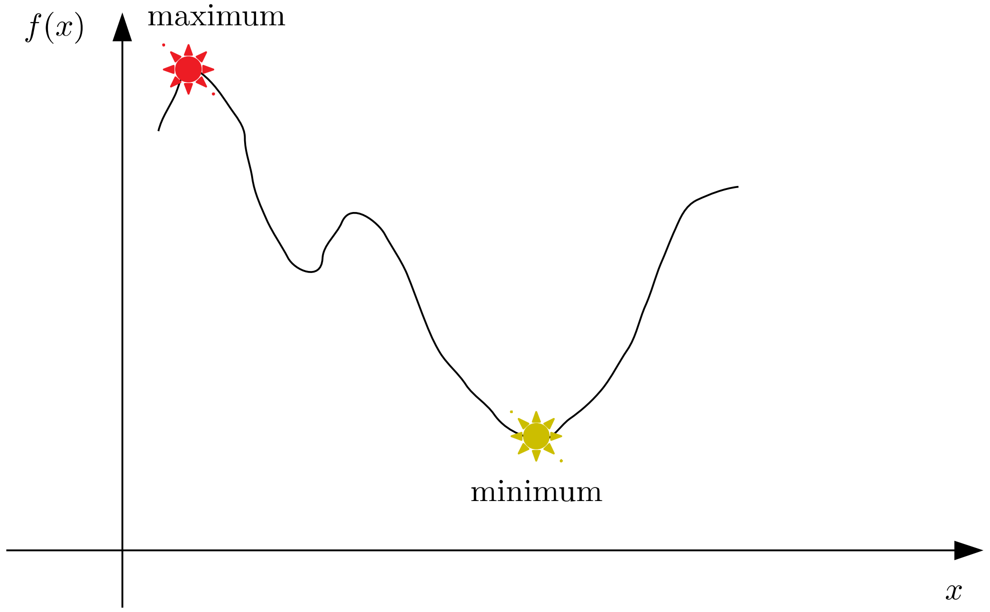
Will see this again on days 4/5

Will only focus on gradient descent - a tiny piece of the general landscape of optimization strategies

Red Hat

What is Optimization?

Mathematical and numerical techniques to find the maximum or minimum of a function



Terminology

Global maximum: the maximum value the function takes across its domain

Local maximum: the maximum value of the function in a small neighborhood around an x (input value)

Similar definitions for minima

Red Hat

Terminology

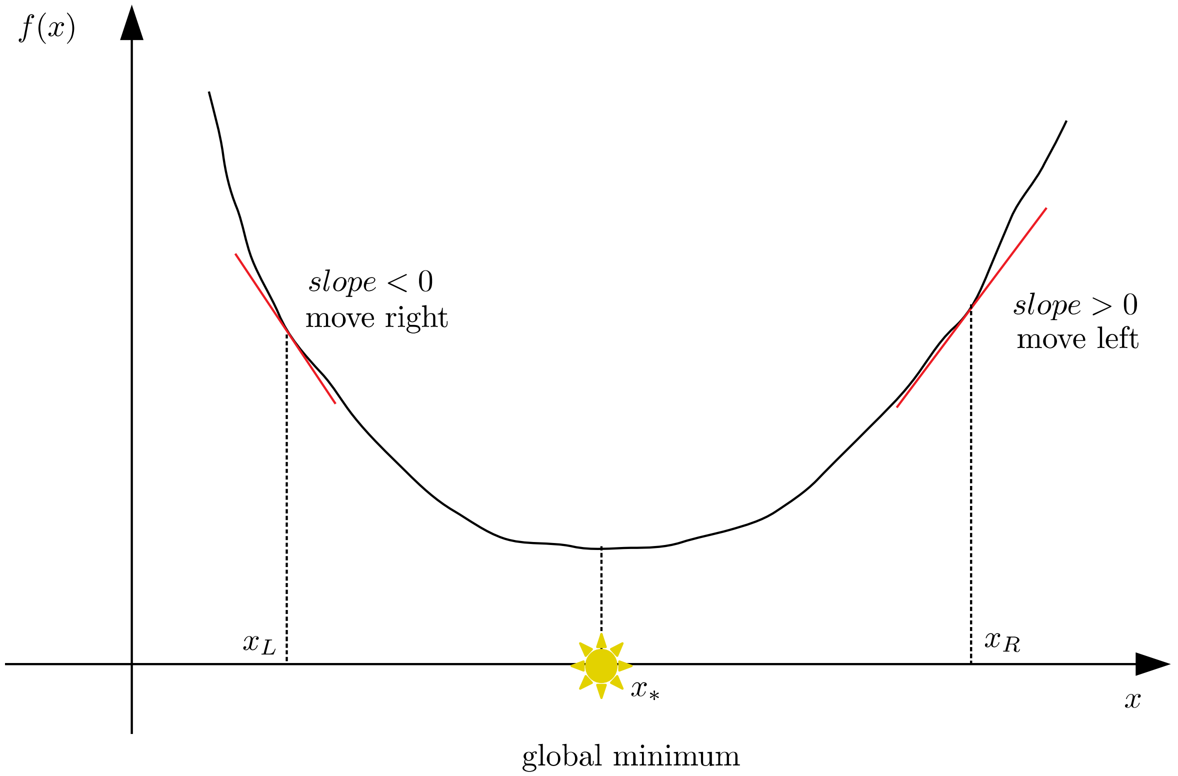
Training: Finding weights that minimize the loss function on the test set (“out of sample”) for a neural network is an optimization problem

Hyperparameter tuning: Finding the number of layers, the activation function, the number of nodes in each layer, and other parameters is also an optimization problem

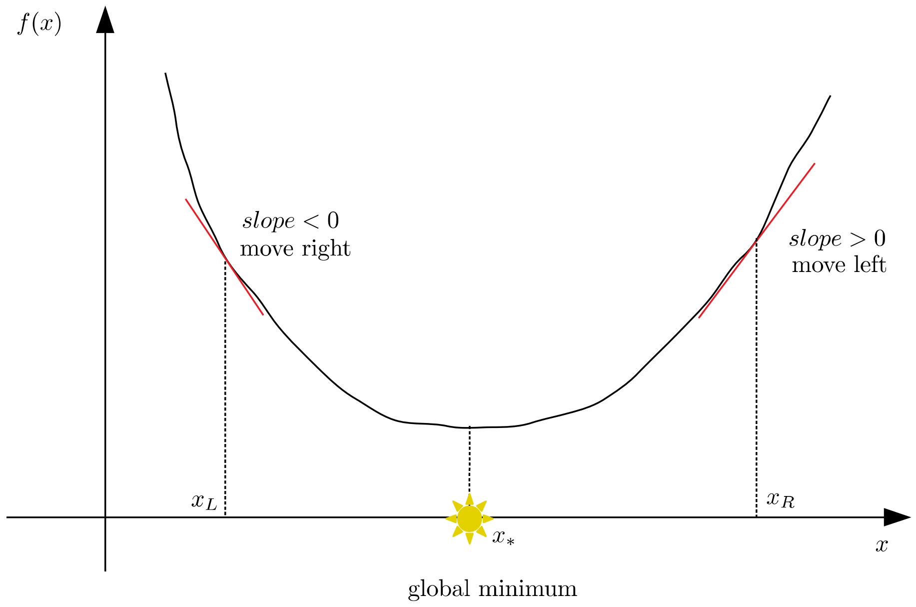
Obstacles

- f might be computationally expensive to evaluate.
- f might be discrete so no way to evaluate derivatives which can guide search for minima
- Even if f is continuous and well-behaved, evaluating higher derivatives (second, third etc.) is very expensive.
- f might have very complex structure with multiple (possibly infinite) local minima
- f might be very high-dimensional i.e. it has a large number of inputs and hence we are searching for the minima in a high-dimensional space with many more directions to explore.

Gradient Descent



Gradient Descent



Iterative method

Start at some reasonable point and keep updating

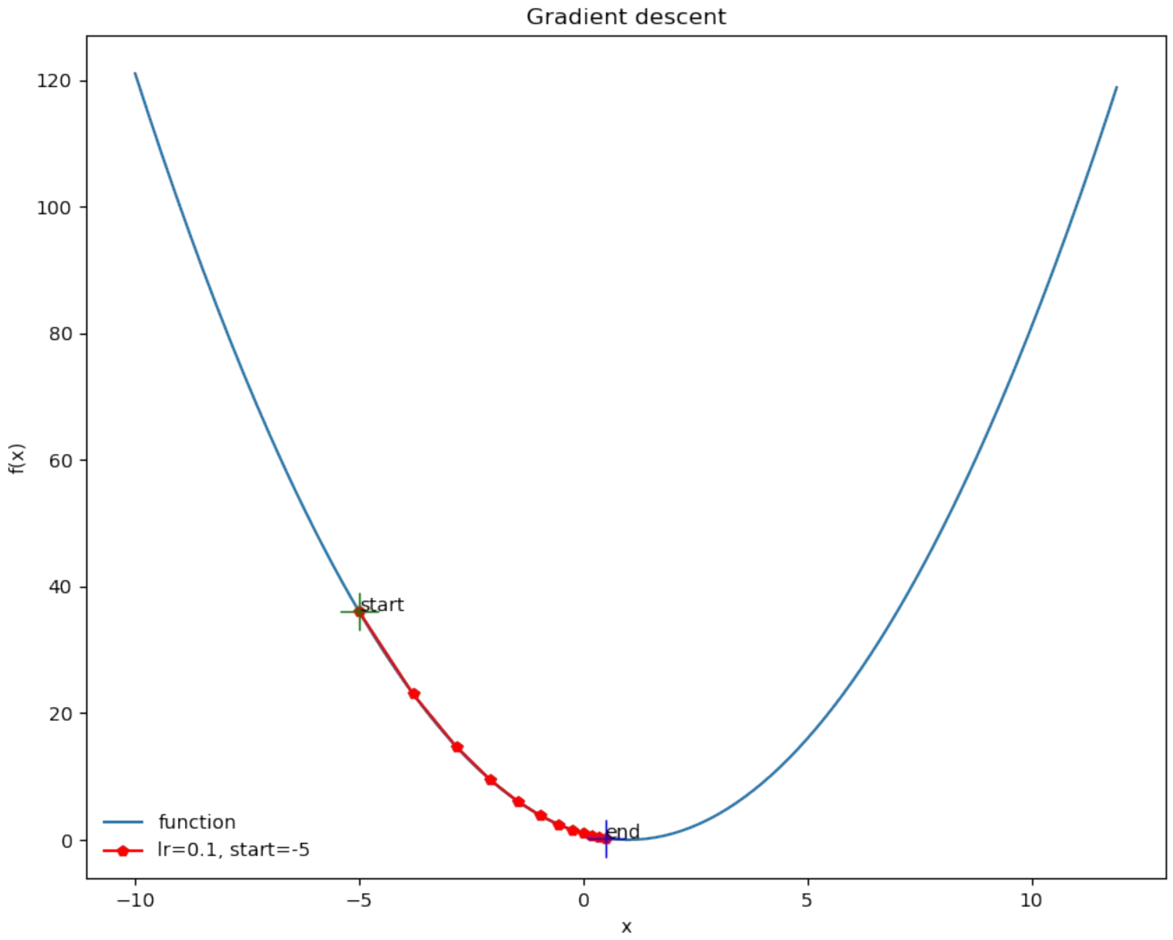
$$x^{(t+1)} = x^{(t)} + \text{update}$$

$$x^{(t+1)} = x^{(t)} - \eta \frac{df}{dx}(x^{(t)})$$

$$\eta = \text{learning rate}$$

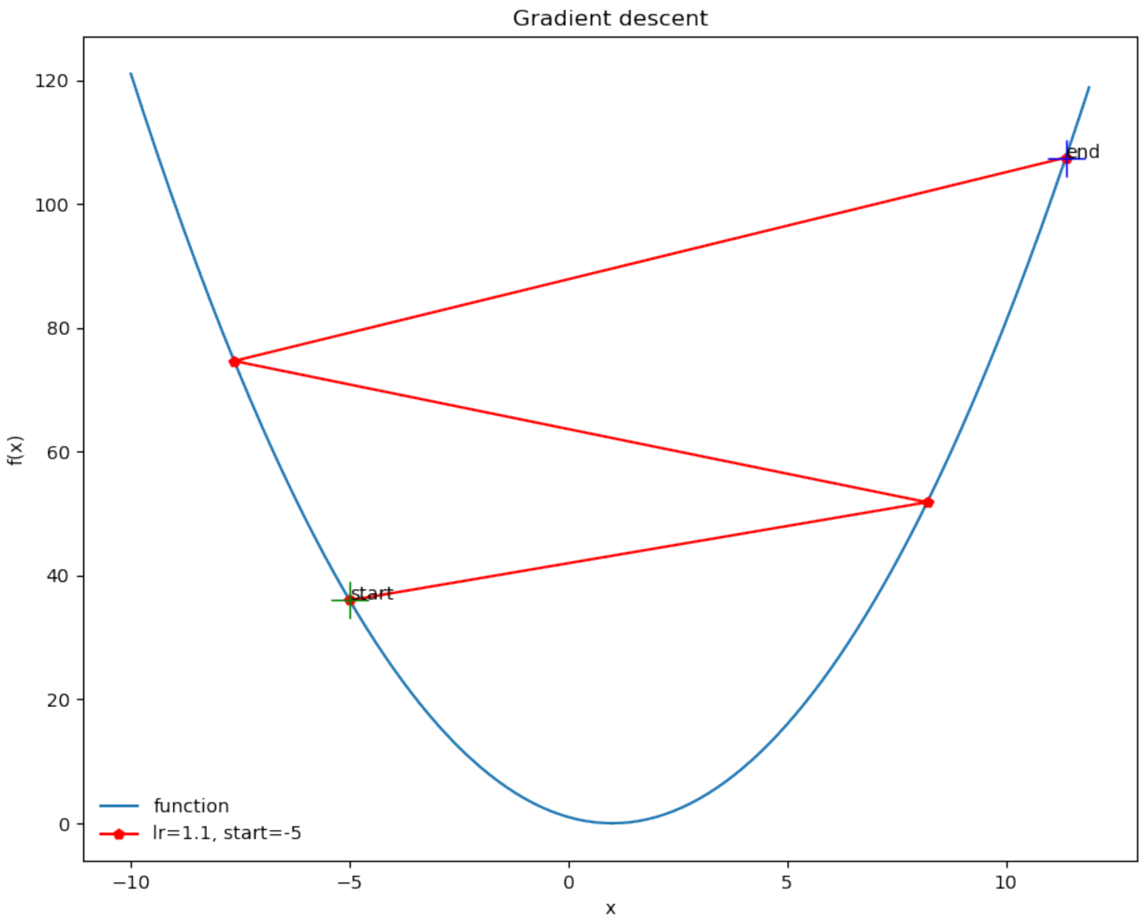
Red Hat

Intuition

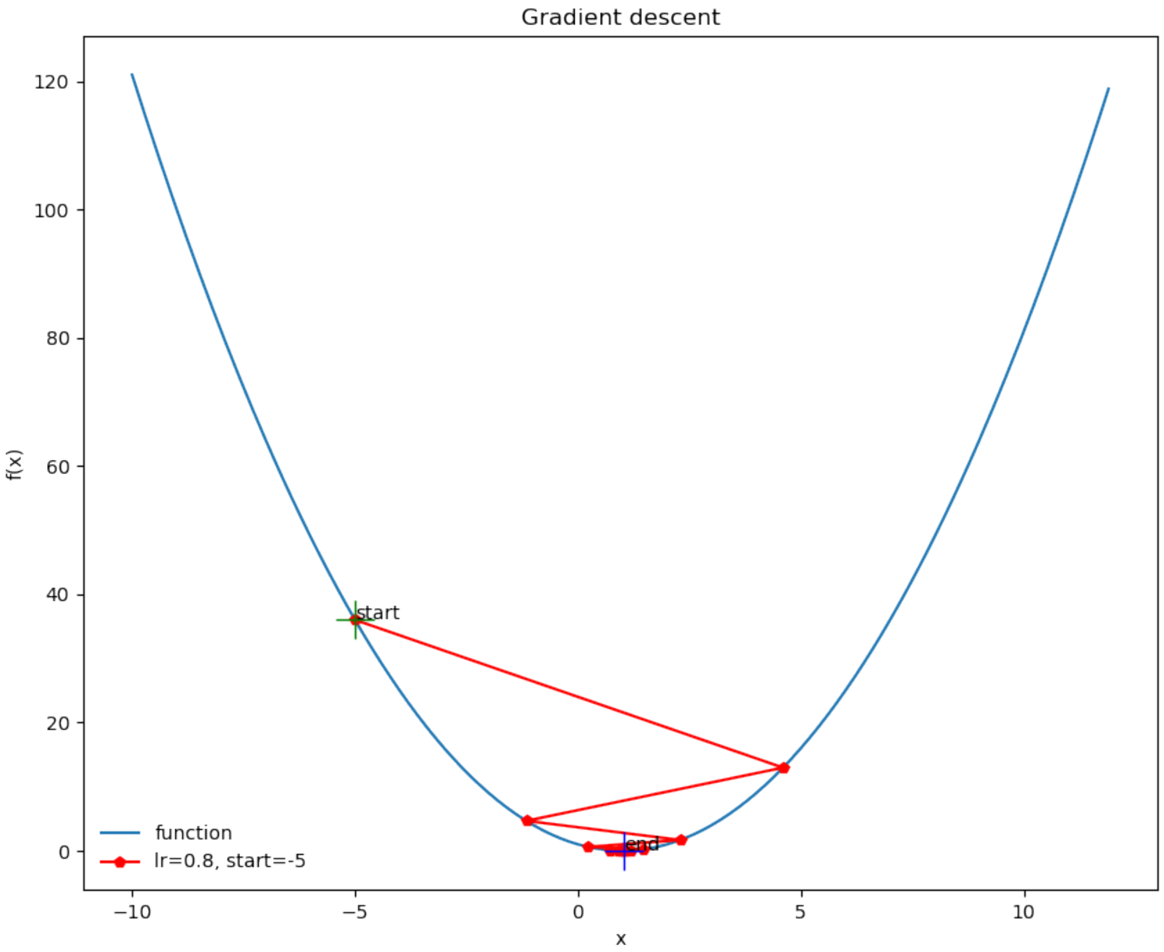


Gradient descent
120
100
80
60
40
20
0
f(x)
start
end
function
lr=0.1, start=-5
−10
−5
0
5
10
x

Intuition

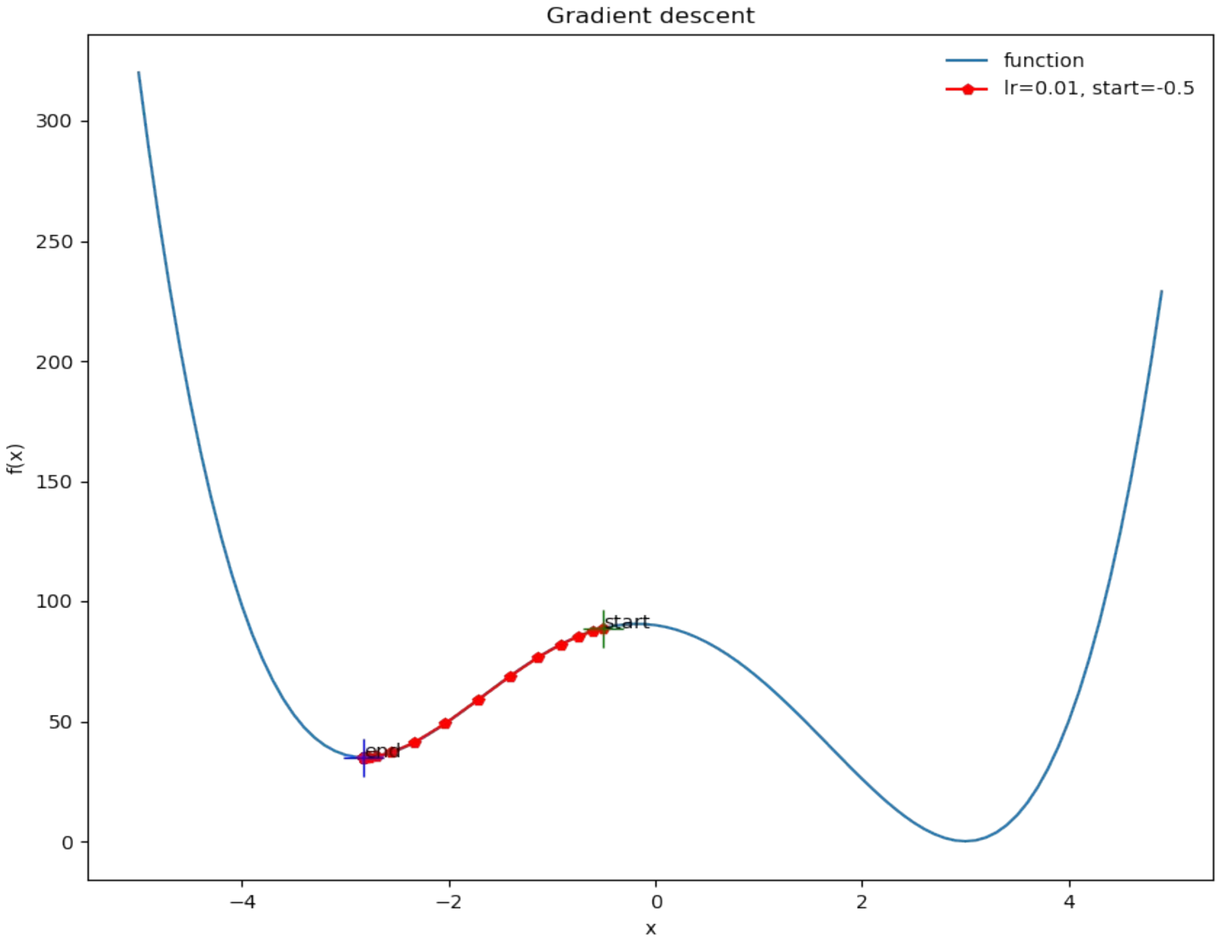

Gradient descent
120
100
80
60
40
20
0
f(x)
−10
−5
0
5
10
x
end
start
function
lr=1.1, start=-5

Intuition

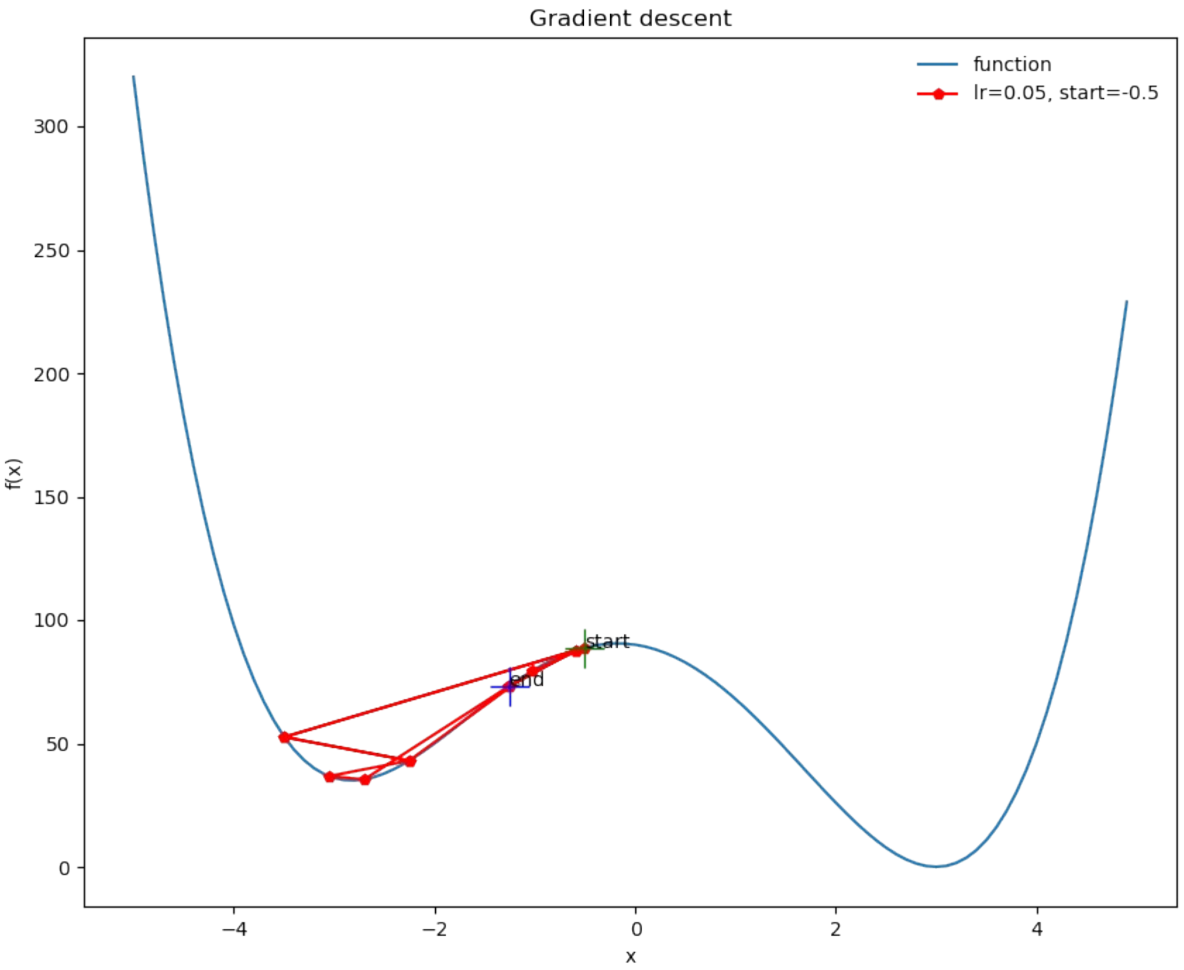


Red Hat

Intuition

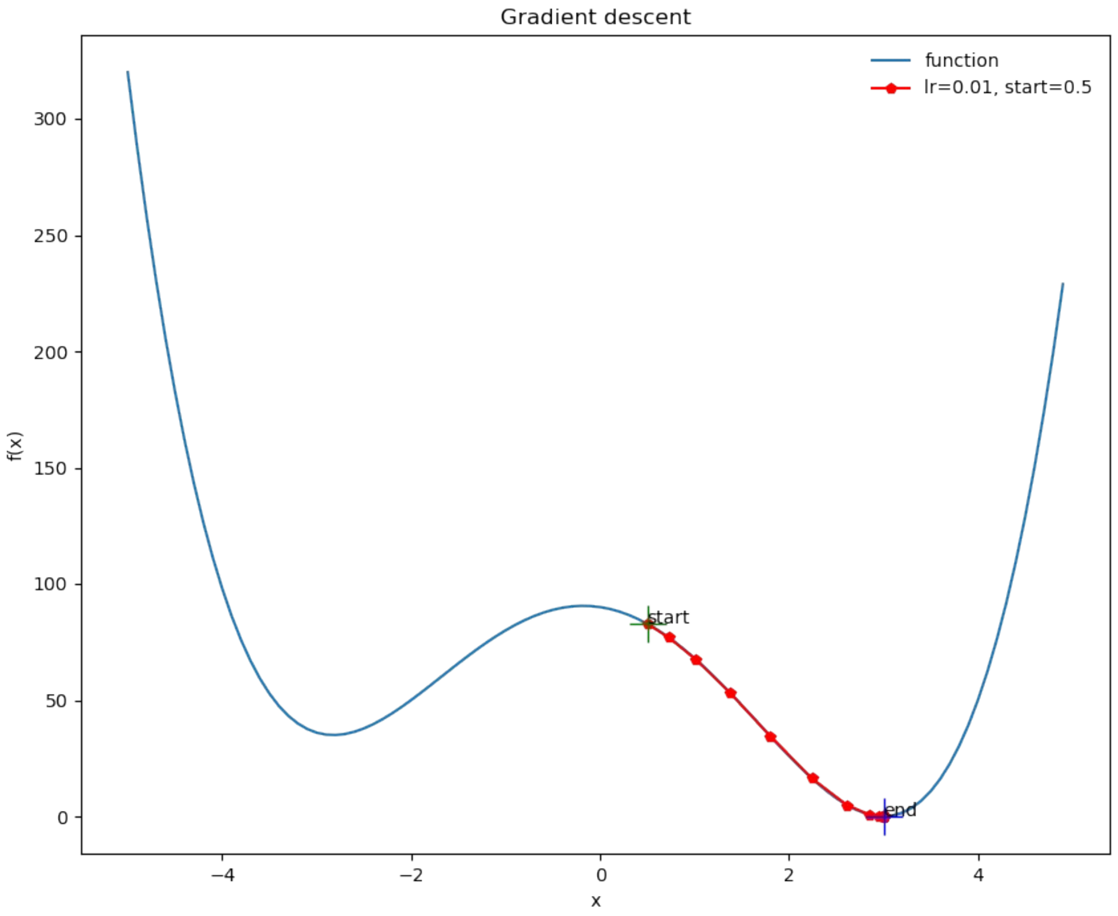

Gradient descent
function
lr=0.01, start=-0.5
300
250
200
150
100
50
0
f(x)
start
end
−4
−2
0
2
4
x

Intuition



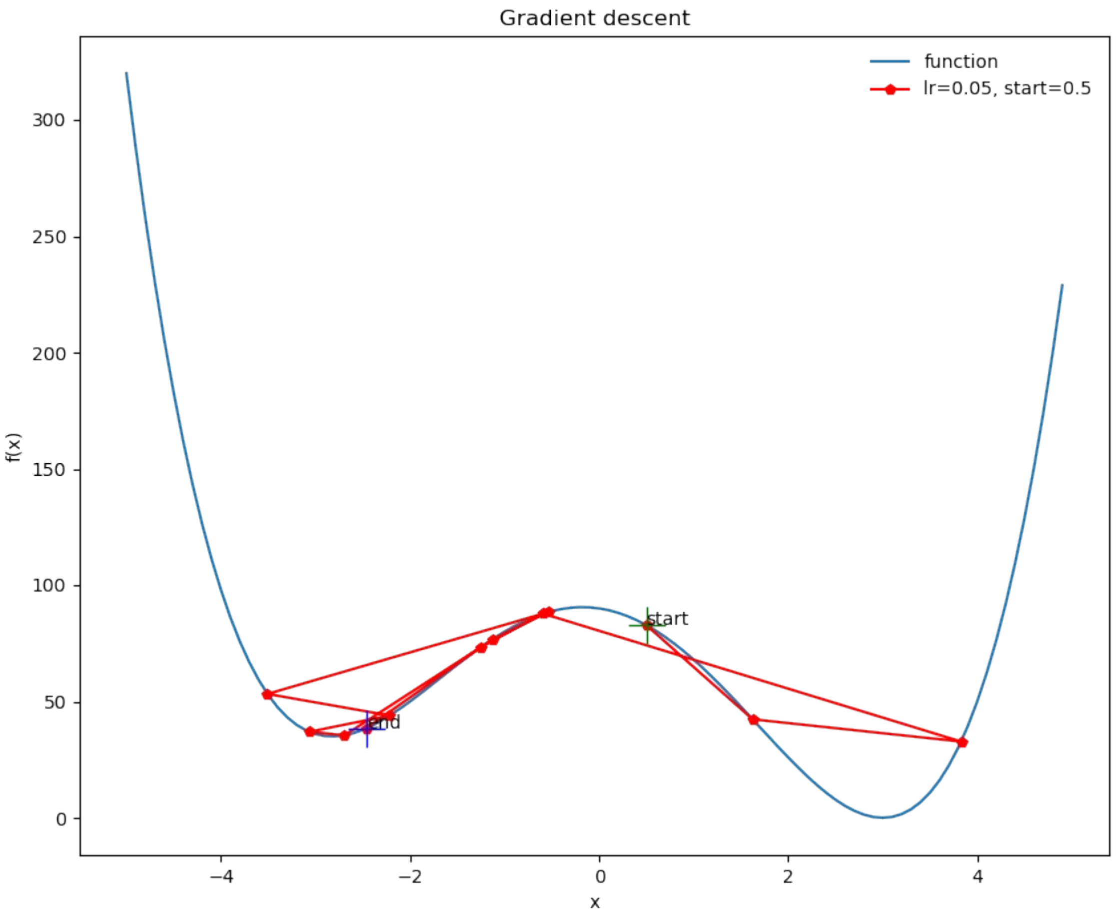
Gradient descent
function
lr=0.05, start=-0.5
300
250
200
150
100
50
0
f(x)
start
end
−4
−2
0
2
4
x

Intuition



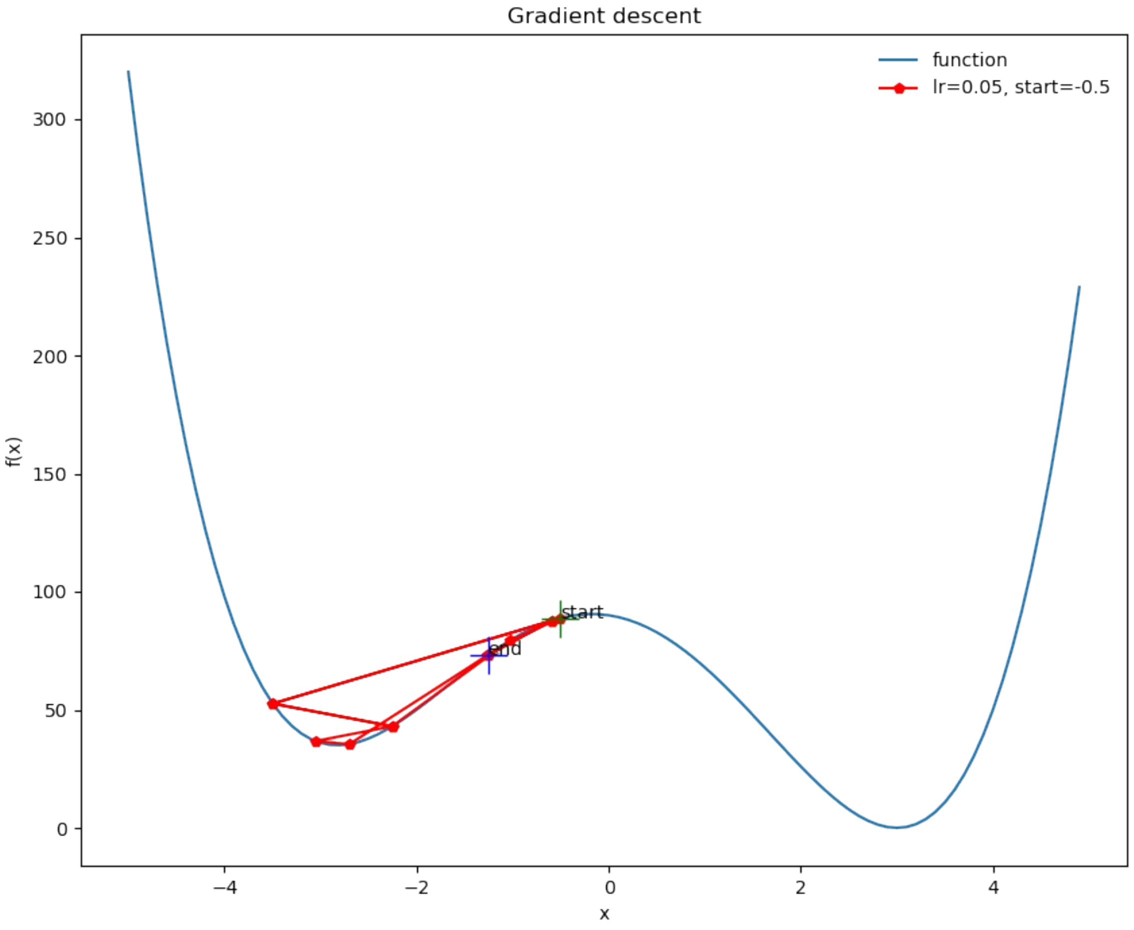
Gradient descent
function
lr=0.01, start=0.5
start
end
300
250
200
150
100
50
0
f(x)
−4
−2
0
2
4
x

Intuition



Red Hat

Intuition



Red Hat

Intuition

- Convergence very sensitive to learning rate

- Learning rate too small:
  - Will converge to **local** minimum
  - Might take a long time

- Learning rate too large:
  - Will bounce around or even escape valley one started in

- If function decreases gradually, will take a long time to converge.

- Obvious question:
  - Can learning rate be **adaptive** i.e. change in response to location?

Convex Convergence

Every smooth function can be locally approximated as a quadratic (convex) function

$$\mathrm{f(x)} = \underbrace{f(x_{min}) + \overset{0}{f'(x_{min})}(x - x_{min}) + \frac{1}{2}f''(x_{min})(x - x_{min})^2 + \mathcal{O}((x - x_{min})^3)}_{a+b(x-c)^2}$$

Taylor expansion around local minimum

Function of one variable

Convex Convergence

Every smooth function can be locally approximated as a quadratic (convex) function

$$f(x,y) = f(x_{min}, y_{min}) + \overbrace{\frac{\partial f}{\partial x}}^{0}(x - x_{min}) + \overbrace{\frac{\partial f}{\partial y}}^{0}(y - y_{min}) + \frac{1}{2}\frac{\partial^2 f}{\partial x^2}(x - x_{min})^2 + \frac{1}{2}\frac{\partial^2 f}{\partial y^2}(y - y_{min})^2 + \frac{\partial^2 f}{\partial x \partial y}(x - x_{min})(y - y_{min}) + \mathcal{O}(\text{cubic})$$

$$f(x,y) = a + b(x - x_{min})^2 + d(y - y_{min})^2 + g(x - x_{min})(y - y_{min})$$

Taylor expansion around local minimum

All derivatives evaluated at (x_{min}, y_{min})

Convex Convergence

$$f(x,y) = a + b(x - x_{min})^2 + d(y - y_{min})^2 + g(x - x_{min})(y - y_{min})$$

$$\text{Change of coordinates: } x' = x - x_{min}, y' = y - y_{min}$$

$$f(x,y) = a + \begin{bmatrix} x' & y' \end{bmatrix} \begin{bmatrix} b & g\epsilon \\ g(1-\epsilon) & d \end{bmatrix} \begin{bmatrix} x' \\ y' \end{bmatrix}$$

$$\epsilon = \frac{1}{2} \rightarrow \text{diagonalize symmetric matrix} \rightarrow f(x'', y'') = a + \alpha x''^2 + \beta y''^2$$

Convex Convergence

$$f(x) = a + \frac{b}{2}(x - c)^2$$

$$\text{Start point: } x^{(0)}$$

$$\text{Update: } x^{(t)} = x^{(t-1)} - \eta \frac{df}{dx}(x^{(t-1)})$$

$$\boxed{x^{(t)} = x^{(t-1)} - \eta b(x^{(t-1)} - c)}$$

Red Hat

Convex Convergence

$$x^{(t)} = x^{(t-1)} - \eta b(x^{(t-1)} - c)$$

$$\text{Minimum at: } x = c$$

$$\text{Consider Error: } \epsilon_t \equiv |x^{(t)} - c|$$

$$\underbrace{|x^{(t)} - c|}_{\epsilon_t} = |x^{(t-1)} - \eta b(x^{(t-1)} - c) - c| = \underbrace{|(x^{(t-1)} - c)|}_{\epsilon_{t-1}} |1 - \eta b|$$

$$\boxed{\epsilon_t = \epsilon_{t-1}|1 - \eta b|}$$

Red Hat

Convex Convergence

$$\epsilon_t = \epsilon_{t-1} |1 - \eta b|$$

$$\boxed{\epsilon_t = \epsilon_0 |1 - \eta b|^t}$$

Distance from minimum
at time t (ideally close to 0)

Distance from minimum
at time 0

Convex Convergence

$$\boxed{\epsilon_t = \epsilon_0 |1 - \eta b|^t} \qquad 0 < 1 - \eta b < 1$$

$$\text{Suppose we want } \frac{\epsilon_t}{\epsilon_0} = \delta, \text{ a small number}$$

$$t = \frac{\log \delta}{\log |1 - \eta b|} \approx \frac{-\log \delta}{\eta b} \qquad \eta b \text{ small}$$

Red Hat

Thank you

Red Hat is the world's leading provider of

enterprise open source software solutions.

Award-winning support, training, and

consulting services make

Red Hat a trusted adviser to the Fortune 500.

linkedin.com/company/red-hat

youtube.com/user/
RedHatVideos

facebook.com/redhatinc

twitter.com/RedHat