3: Linear independence

- linearity
- linear (in)dependence
- bases

Functions

- the notation $f : \mathbb{R}^n \to \mathbb{R}$ conveys the following information:
  - f is a **function**
  - f maps real n-vectors to real scalars
- if x is a vector in $\mathbb{R}^n$, then we can write $y = f(x)$
  - x is the **argument**
  - the dimension of y is determined by $f : \mathbb{R}^n \to \mathbb{R}^m$
- more generally we have $f : X \to Y$ where X and Y are sets
  - X is the **domain**
  - Y is the **co-domain**

Inner product as a function

- consider $f : \mathbb{R}^n \to \mathbb{R}$ where

$$f(x) = w_1x_1 + w_2x_2 + \cdots + w_nx_n$$

  for a given vector w

- f computes the **inner product** of its argument x and the fixed vector w

- let the argument to f be $\alpha x + \beta y$ where α, β are scalars and x, y are n-vectors:

$$\begin{aligned} f(\alpha x + \beta y) &= w^T(\alpha x + \beta y) \\ &= w^T(\alpha x) + w^T(\beta y) \\ &= \alpha(w^T x) + \beta(w^T y) \\ &= \alpha f(x) + \beta f(y) \end{aligned}$$

- the property $f(\alpha x + \beta y) = \alpha f(x) + \beta f(y)$ is called **superposition**

- functions that satisfy superposition are called **linear**

Linear functions

- superposition extends to an arbitrary number of scalar-vector additions

- let $x_1, \ldots x_n$ be n-vectors and $\alpha_1, \ldots, \alpha_n$ be scalars, then

$$f(\alpha_1 x_1 + \cdots + \alpha_n x_n) = \alpha_1 f(x_1) + \cdots + \alpha_n f(x_n)$$

- the definition of superposition is often stated in two parts

  1. **Homogeneity:** $f(\alpha x) = \alpha f(x)$
  2. **Additivity:** $f(x + y) = f(x) + f(y)$

Inner product representation

- consider the vector $y \in \mathbb{R}^n$, then we have

$$y = e_1 y_1 + \cdots + e_n y_n$$

- **all** linear functions can be represented as inner products, i.e.,

there exists a vector a such that $f(x) = a^T x$

- using the above observation, it follows that

$$\begin{aligned} f(x) &= f(e_1 x_1 + \cdots + e_n x_n) \\ &= x_1 f(e_1) + \cdots + x_n f(e_n) \end{aligned}$$

and so $a_i = f(e_i)$

Affine functions

- a function is **affine** it can be expressed as a linear function plus a constant, i.e.,

$$f(x) = w^T x + b$$

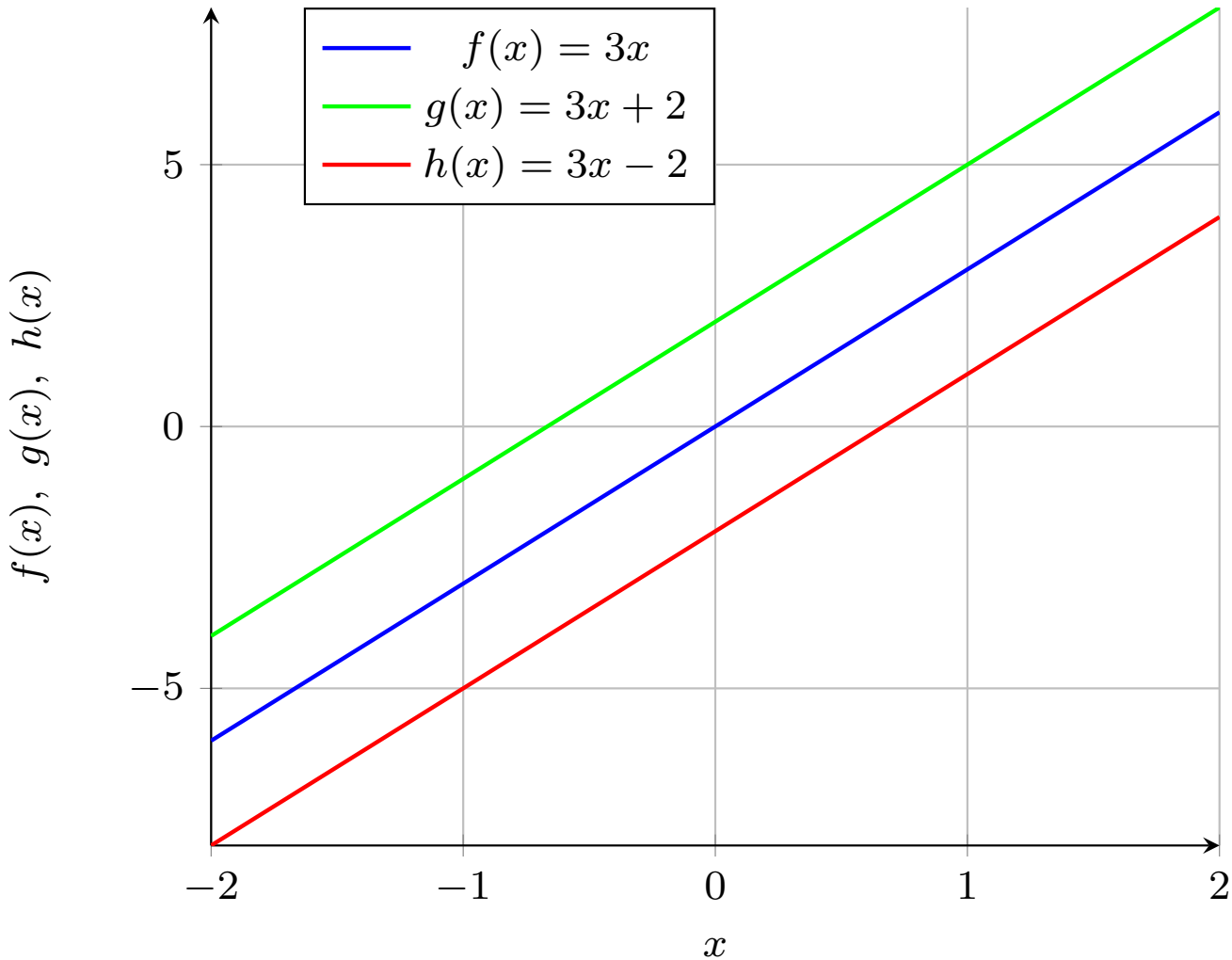
- this equivalent to a function satisfying the **restricted superposition** property

$$f(\alpha x + \beta y) = \alpha f(x) + \beta f(y) \quad \text{and} \quad \alpha + \beta = 1$$

- to see this, consider $x, y \in \mathbb{R}^n$ and scalars α, β,

$$\begin{aligned} f(\alpha x + \beta y) &= w^T(\alpha x + \beta y) + b \\ &= \alpha w^T x + \beta w^T y + (\alpha + \beta) b \quad \texttt{// uses } \alpha + \beta = 1 \\ &= \alpha(w^T x + b) + \beta(w^T y + b) \\ &= \alpha f(x) + \beta f(y) \end{aligned}$$

Linear and affine functions

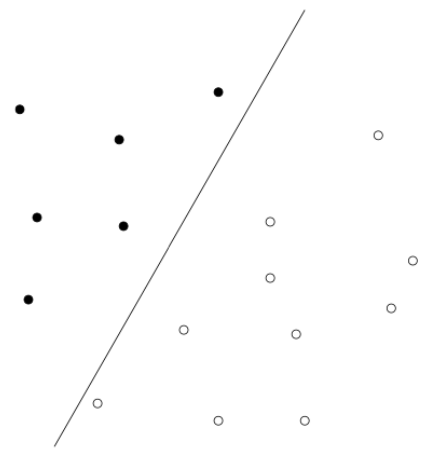


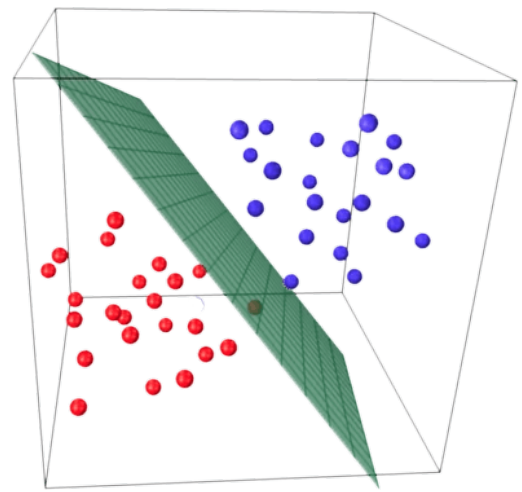
Affine functions in action

Data classification: linear classifiers

given two sets of vectors $\{x_1, \ldots, x_n\} \subset \mathbb{R}^n$ and $\{y_1, \ldots, y_m\} \subset \mathbb{R}^n$, find an affine function $f : \mathbb{R}^n \to \mathbb{R}$ such that

$$f(x_i) > 0, \quad \text{for } i = 1, \ldots, n \quad \text{and} \quad f(y_i) < 0, \quad \text{for } i = 1, \ldots, m$$





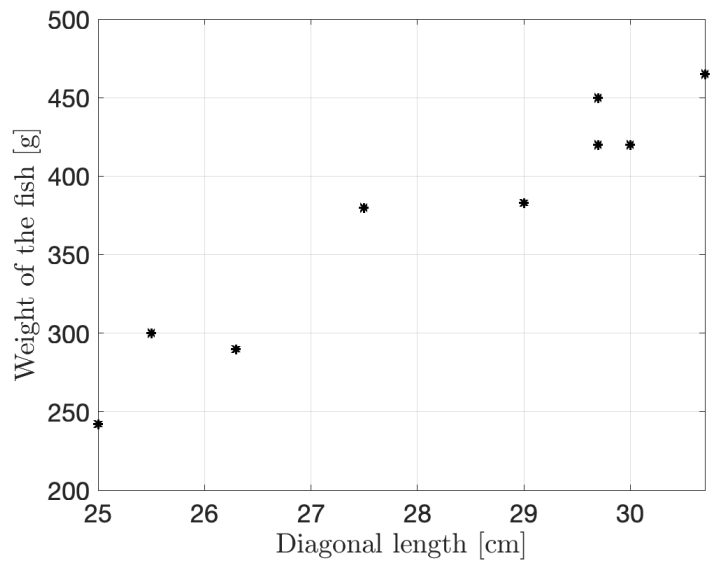
- the line $\{x \mid f(x) = 0\}$ is the classification boundary
- "linear" classifiers use affine functions $f(x) = a^T x + b$

Data fitting: Linear regression

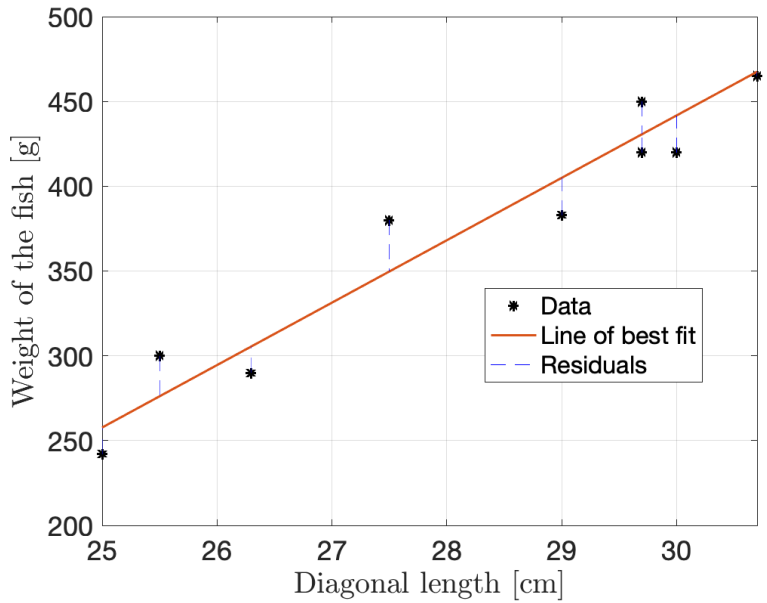
given some data, a candidate model is proposed as $\hat{y} = x^T\theta + v$, where

- $x \in \mathbb{R}^n$ is the **feature vector**
- entries x_i are called **regressors** or **independent variables**
- $\theta \in \mathbb{R}^n$, $v \in \mathbb{R}$ are the **weights** and **offset** respectively
- $\hat{y}$ is the prediction, i.e., approximates some true value y

diagonal width [cm]	weight [g]
25	242
26.3	290
25.5	300
29	383
32	530
29.7	450
29.7	420
30	420
27.5	380
30.7	465



solving the regression problem provides estimates $(\theta, v) = (36.8, -662.3)$



General interpretation

$$\hat{y} = \theta_1 x_1 + \theta_2 x_2 + \ldots + \theta_n x_n + v$$

- if $|\theta_i|$ is large then the model $\hat{y}$ depends heavily on feature x_i
- if $\theta_i > 0$, then $\hat{y}$ increases by θ_i when x_i increases by 1 (remaining θ_j constant)

Multiple regressors

a model that predicts the value of a house in Sacramento, based upon its area (measured in sq. feet) and no. of bedrooms

House	x_1 (area)	x_2 (beds)	y (price)
1	0.846	1	115.00
2	1.324	2	234.50
3	1.150	3	198.00
4	3.037	4	528.00
5	3.984	5	572.50

$$\hat{y} = \theta_1 x_1 + \theta_2 x_2 + v$$

- we want to choose θ, v to make $y - \hat{y}$ small
- can be formulated as a least squares problem (later in the course)
- for now we will take

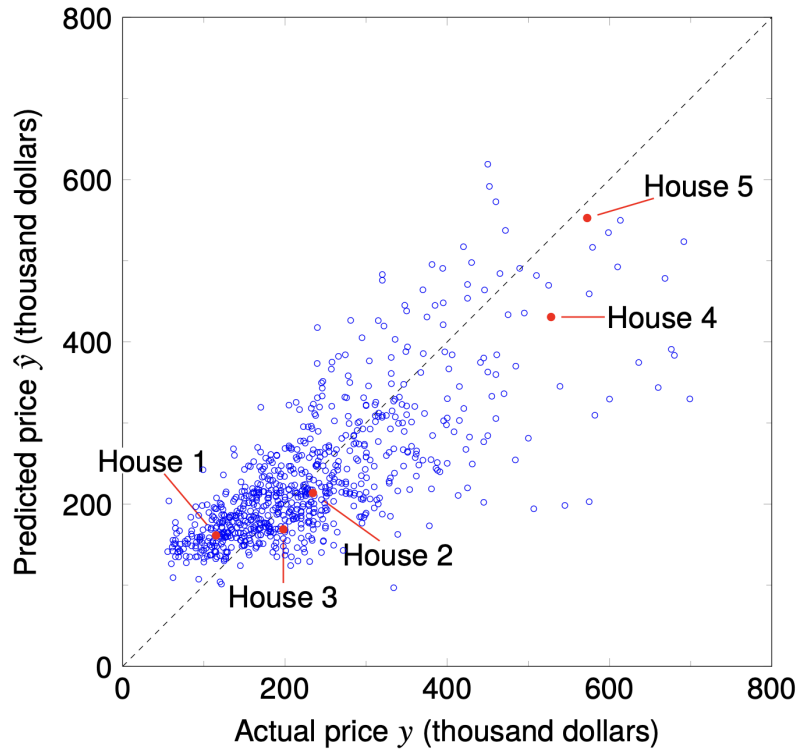
$$\theta = \begin{pmatrix} 148.73 \\ -18.85 \end{pmatrix}, \quad v = 54.4$$

[example adapted and image from Boyd & Vandenberghe, p.39]

Assessing the model

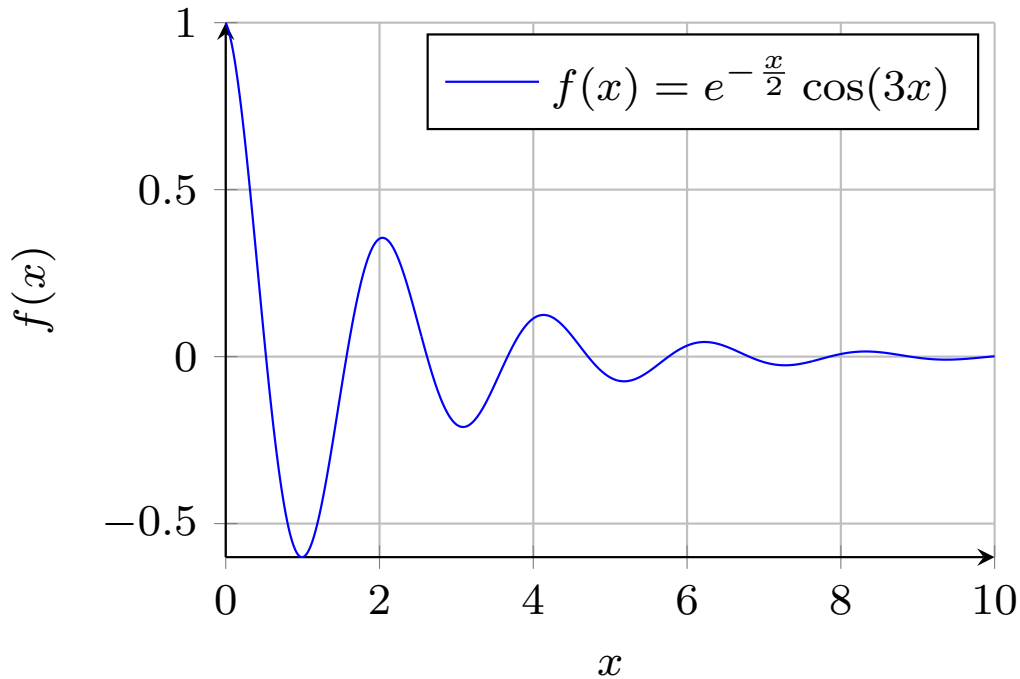
assume that $\theta = \begin{pmatrix} 148.73 \\ -18.85 \end{pmatrix}$, $v = 54.4$ is optimal for our choice of regression algorithm, we can visualize how well it does:

House	x_1 (area)	x_2 (beds)	y (price)	$\hat{y}$ (prediction)
1	0.846	1	115.00	161.37
2	1.324	2	234.50	213.61
3	1.150	3	198.00	168.88
4	3.037	4	528.00	430.67
5	3.984	5	572.50	552.66



Nonlinear functions

any function f that is not linear is said to be **nonlinear**



- the "real world" in nonlinear
- locally, linear functions often do a good job approximating a nonlinear function

Taylor approximation

consider a scalar differentiable function f, the **Taylor series expansion** of f around z is

$$f(x) = f(z) + f'(z)(x-z) + \frac{f''(z)}{2!}(x-z)^2 + \frac{f'''(z)}{3!}(x-z)^3 + \dots$$

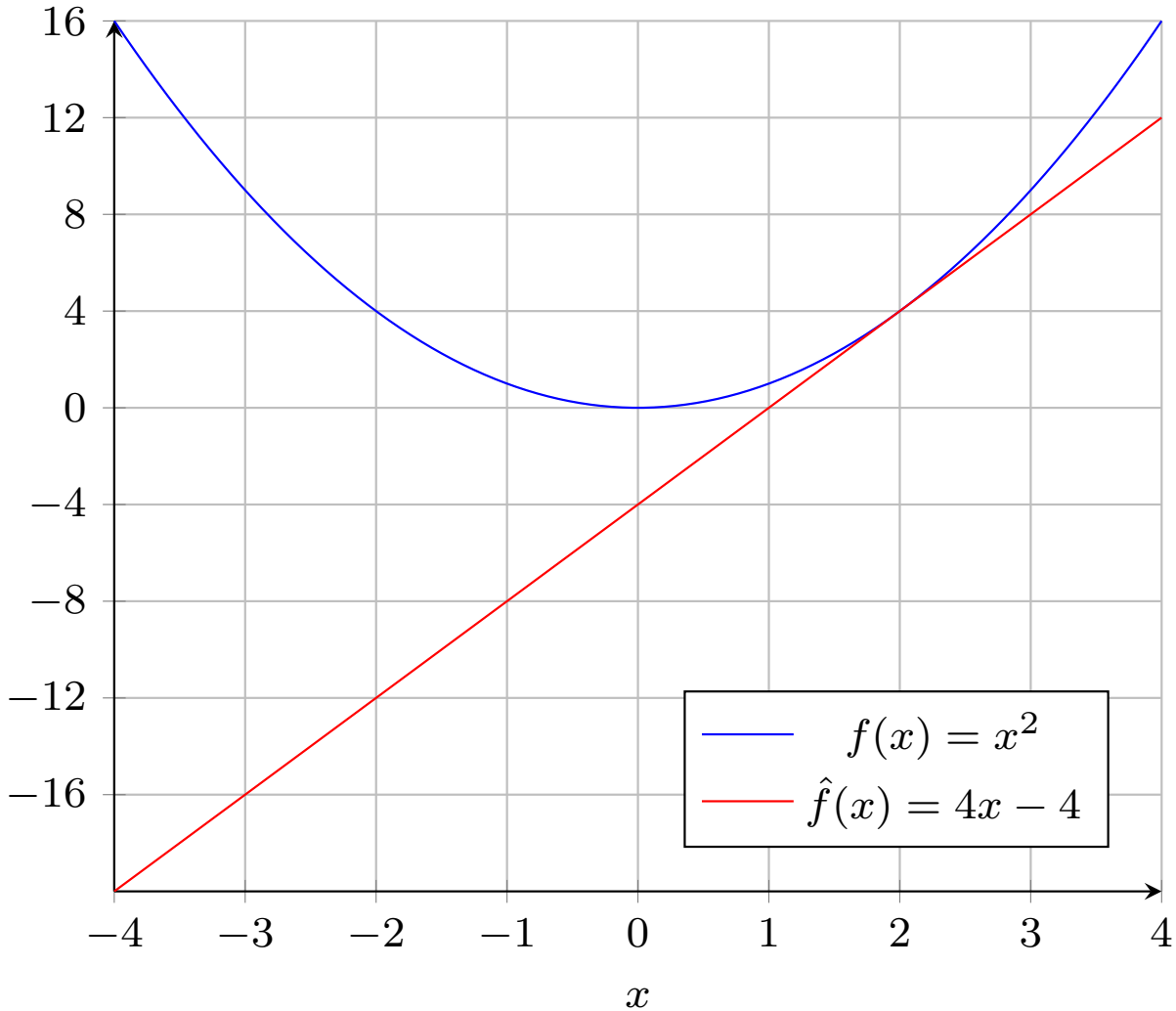
- $f' = \frac{\mathrm{d}f}{\mathrm{d}x}(z)$
- for x sufficiently close to z, $\frac{1}{n!}(x-z)^n$ higher order terms contribute very little
- the **first-order Taylor approximation** of a f around z is

$$\hat{f}(x) = f(z) + f'(z)(x-z)$$

- $\hat{f}$ is referred to the "linearization" of f

Example

linearization of $f(x) = x^2$ at $z = 2$



Multi-variable Taylor approximation

given a differentiable function $f : \mathbb{R}^n \to n$, its first-order Taylor approximation at z is

$$\hat{f}(x) = f(z) + \frac{\partial f}{\partial x_1}(z)(x_1 - z_1) + \ldots + \frac{\partial f}{\partial x_n}(z)(x_n - z_n)$$

- more compactly written using the gradient vector ∇f:

$$\hat{f}(x) = f(z) + \nabla f(z)^T (x - z)$$

  where

$$\nabla f(z) = \begin{pmatrix} \frac{\partial f}{\partial x_1}(z) \\ \vdots \\ \frac{\partial f}{\partial x_n}(z) \end{pmatrix}$$

- $\hat{f}$ is **affine** in x

Linear combinations

consider the set of vectors $v_1, v_2, \ldots, v_k \in \mathbb{R}^n$

- a **linear combination** of the set of vectors is

$$\lambda_1 v_1 + \lambda_2 v_2 + \ldots + \lambda_k v_k, \quad \text{with } \lambda_i \in \mathbb{R}$$

Linear combinations

consider the set of vectors $v_1, v_2, \ldots, v_k \in \mathbb{R}^n$

- a **linear combination** of the set of vectors is

$$\lambda_1 v_1 + \lambda_2 v_2 + \ldots + \lambda_k v_k, \quad \text{with } \lambda_i \in \mathbb{R}$$

- the **span** of the set is

$$\mathbf{span}(v_1, \ldots, v_k) = \{\lambda_1 v_1 + \lambda_2 v_2 + \ldots + \lambda_k v_k \mid \lambda_i \in \mathbb{R}\}$$

  i.e., the set of **all linear combinations** of the vectors in the set

Linear combinations

consider the set of vectors $v_1, v_2, \ldots, v_k \in \mathbb{R}^n$

- a **linear combination** of the set of vectors is

$$\lambda_1 v_1 + \lambda_2 v_2 + \ldots + \lambda_k v_k, \quad \text{with } \lambda_i \in \mathbb{R}$$

- the **span** of the set is

$$\mathbf{span}(v_1, \ldots, v_k) = \{\lambda_1 v_1 + \lambda_2 v_2 + \ldots + \lambda_k v_k \mid \lambda_i \in \mathbb{R}\}$$

  i.e., the set of **all linear combinations** of the vectors in the set

- $\mathbf{span}(v_1, \ldots, v_k)$ is a **subspace** of $\mathbb{R}^n$

Linear independence

the set of vectors $v_1, v_2, \ldots, v_k \in \mathbb{R}^n$ is **linearly independent** if

$$\lambda_1 v_1 + \lambda_2 v_2 + \ldots + \lambda_k v_k = 0 \quad \Longrightarrow \quad \lambda_1 = \lambda_2 = \ldots = \lambda_k = 0$$

sets that are not linearly independent are **linearly dependent**

Linear independence

the set of vectors $v_1, v_2, \ldots, v_k \in \mathbb{R}^n$ is **linearly independent** if

$$\lambda_1 v_1 + \lambda_2 v_2 + \ldots + \lambda_k v_k = 0 \quad \Longrightarrow \quad \lambda_1 = \lambda_2 = \ldots = \lambda_k = 0$$

sets that are not linearly independent are **linearly dependent**

Equivalent conditions

- no vector v_i can be expressed as a linear combination of the remaining vectors

$$v_i \neq \lambda_1 v_1 + \ldots + \lambda_{i-1} v_{i-1} + \lambda_{i+1} v_{i+1} + \ldots + \lambda_k v_k$$

- $v_i \notin \textbf{span}(v_1, \ldots, v_{i-1}, v_{+1}, \ldots, v_k)$

- coefficients are uniquely determined, i.e., if

$$\alpha_1 v_1 + \alpha_2 v_2 + \ldots + \alpha_k v_k = \beta_1 v_1 + \beta_2 v_2 + \ldots + \beta_k v_k$$

  then $\alpha_i = \beta_i$

Examples

1. let $e_1 = \begin{pmatrix} 1 \\ 0 \end{pmatrix}$ and $e_2 = \begin{pmatrix} 0 \\ 1 \end{pmatrix}$, then $\mathbf{span}(e_1, e_2) = \mathbb{R}^2$

2. $\left\{ \begin{pmatrix} 1 \\ 0 \end{pmatrix}, \begin{pmatrix} 3 \\ 0 \end{pmatrix} \right\}$

3. $\left\{ \begin{pmatrix} -2 \\ 0 \end{pmatrix}, \begin{pmatrix} 0 \\ 3 \end{pmatrix}, \begin{pmatrix} 4 \\ 4 \end{pmatrix} \right\}$

Increasing span theorem

Independence-dimension inequality

If the set of vectors $v_1, \ldots, v_k \in \mathbb{R}^n$ are linearly independent, then $k \leq n$.

Increasing span theorem

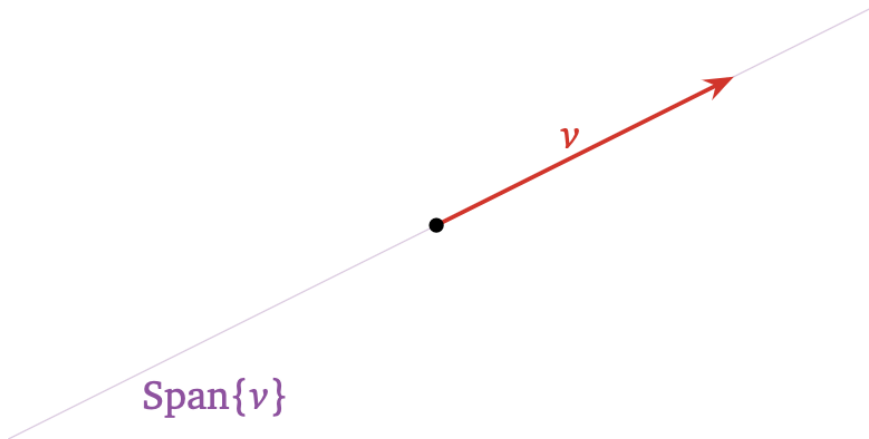
Independence-dimension inequality

If the set of vectors $v_1, \ldots, v_k \in \mathbb{R}^n$ are linearly independent, then $k \leq n$.

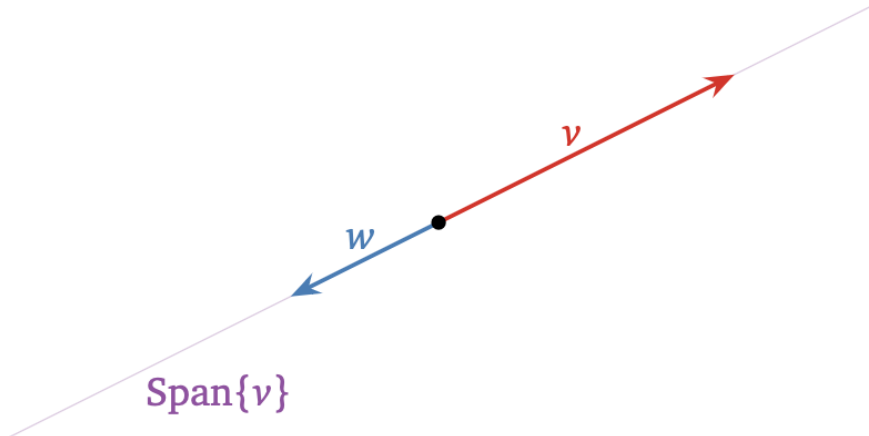
Consequence

If you build a set (of vectors) by adding one vector at a time, and if the span got bigger every time you added a vector, then your set is linearly independent.

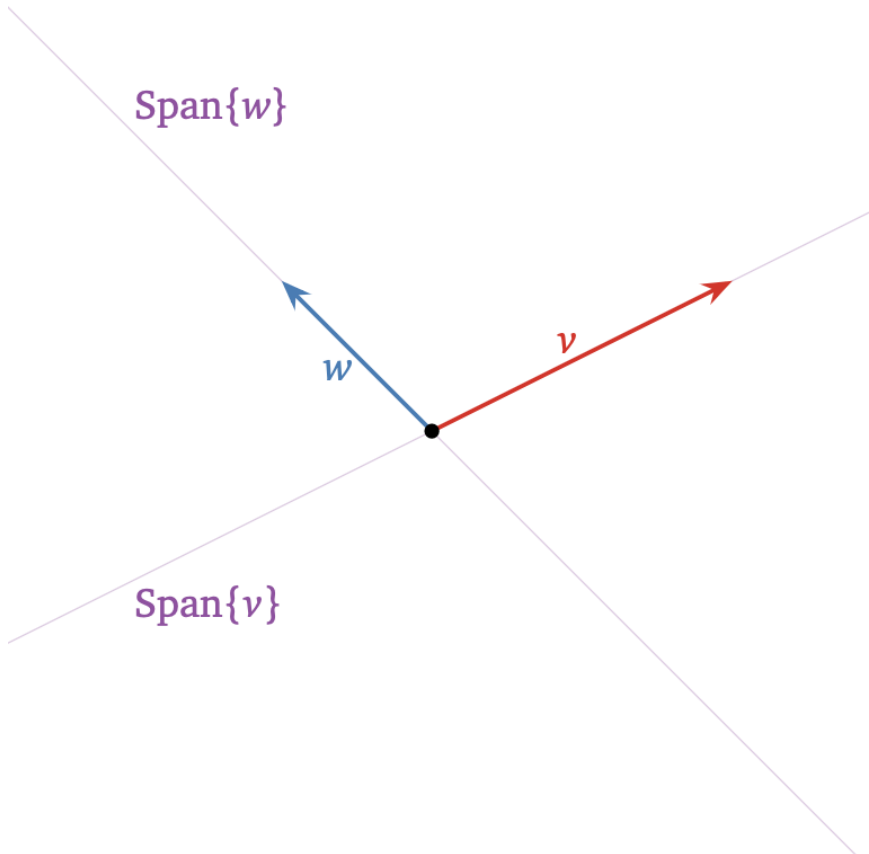
Graphical interpretation in $\mathbb{R}^2$



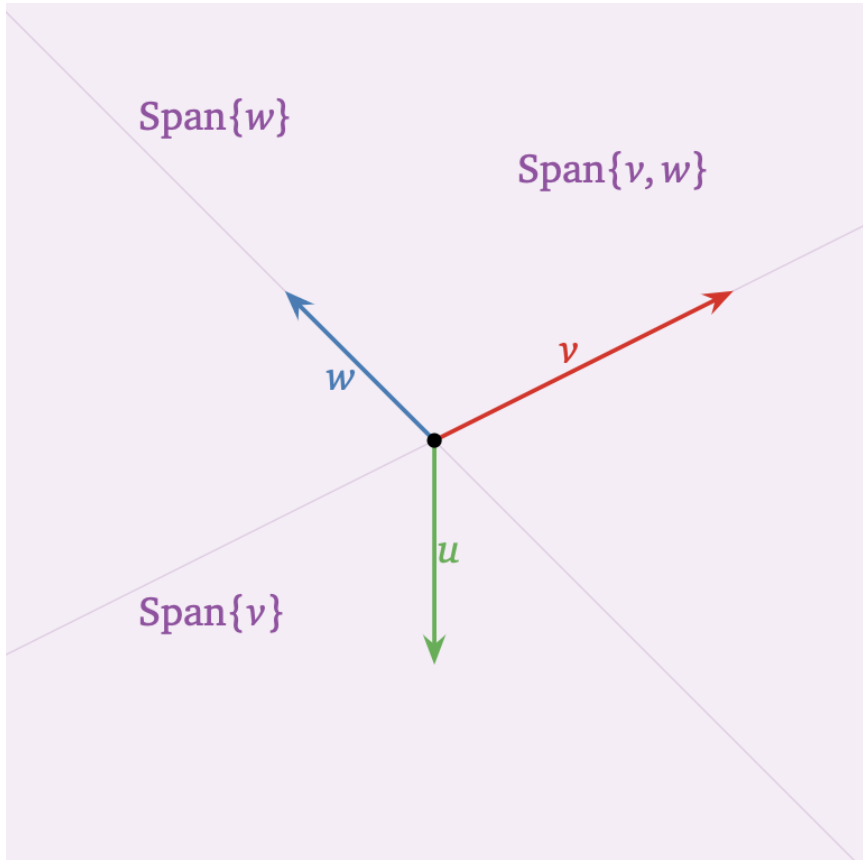
Graphical interpretation in $\mathbb{R}^2$



Graphical interpretation in $\mathbb{R}^2$



Graphical interpretation in $\mathbb{R}^2$



Basis and dimension

Basis

a set of vectors $v_1, \ldots, v_k$ is a **basis** for a subspace S if

1. $\mathbf{span}(v_1, \ldots, v_k) = S$, and
2. $v_1, \ldots, v_k$ are linearly independent

Basis and dimension

Basis

a set of vectors $v_1, \ldots, v_k$ is a **basis** for a subspace S if

1. $\mathbf{span}(v_1, \ldots, v_k) = S$, and
2. $v_1, \ldots, v_k$ are linearly independent

Dimension

the dimension of S, $\mathbf{dim}(S)$, is the number of elements in the basis

Basis and dimension

Basis

a set of vectors $v_1, \ldots, v_k$ is a **basis** for a subspace S if

1. $\mathbf{span}(v_1, \ldots, v_k) = S$, and
2. $v_1, \ldots, v_k$ are linearly independent

Dimension

the dimension of S, $\mathbf{dim}(S)$, is the number of elements in the basis

Notes

- equivalently, every element $x \in S$ can be uniquely expressed as

$$x = \lambda_1 v_1 + \lambda_2 v_2 + \ldots + \lambda_k v_k$$

- a basis is **not** unique, but the number of elements in the basis is always the same

Basis Theorem

Let V be a subspace of dimension m. Then,

- Any m linearly independent vectors in V form a basis for V.
- Any m vectors that span V form a basis for V.

Basis Theorem

Let V be a subspace of dimension m. Then,

- Any m linearly independent vectors in V form a basis for V.
- Any m vectors that span V form a basis for V.

How to apply:

- if you know $\textbf{dim}(V) = m$ then you only need to check one of the two conditions
- if you don't know the dimension, then check both