Math for Machine Learning

Week 1.2: Subspaces, Bases, and Orthogonality

By: Samuel Deng

Logistics and Announcements

- PROJECT OUT (DUE: MONDAY). ✓
- ☆ PS0 OUT (DUE: THURS, tmrrw). ✓
- PS1 OUT (DUE: next THURS). ✓
- 4 6 LATE DAYS! (from 4).
- SAM OUT OF TOWN: WEEK 4.

Lesson Overview

Regression. Fill in gaps from last time: invertibility and Pythagorean theorem.

Subspaces. Subsets of $\mathscr{S} \subseteq \mathbb{R}^n$ where we "stay inside" when performing linear combinations of vectors.

Bases. A "language" to describe all vectors in a subspace.

Orthogonality. Orthonormal bases are "good" bases to work with.

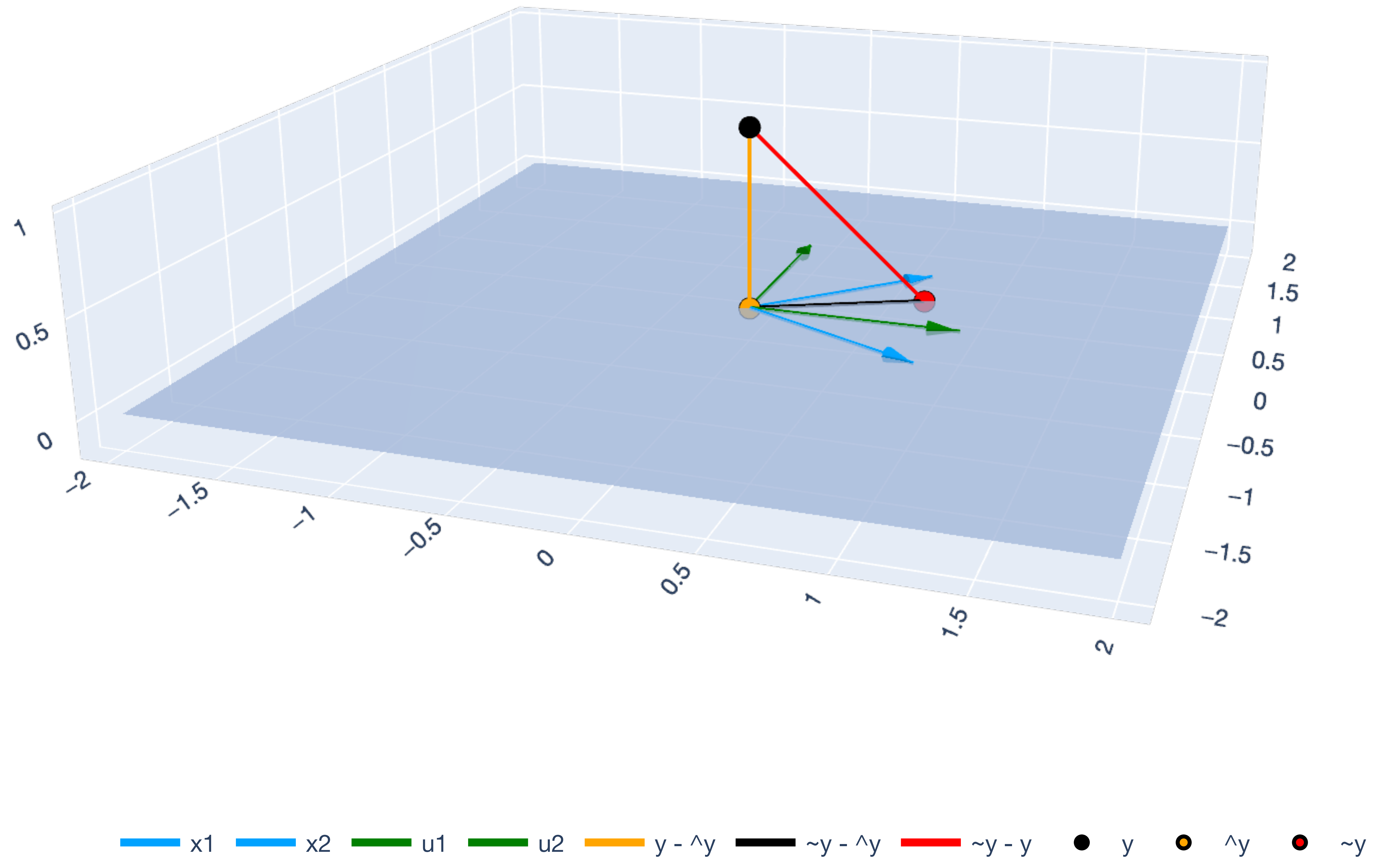
Projection. Formal definition of projection and the relationship between projection and least squares.

Least squares with orthonormal bases. If we have an orthonormal basis for $\text{span}(\text{col}(\mathbf{X}))$, least squares becomes much simpler.

* ORTHOGONALITY BASES.

Lesson Overview

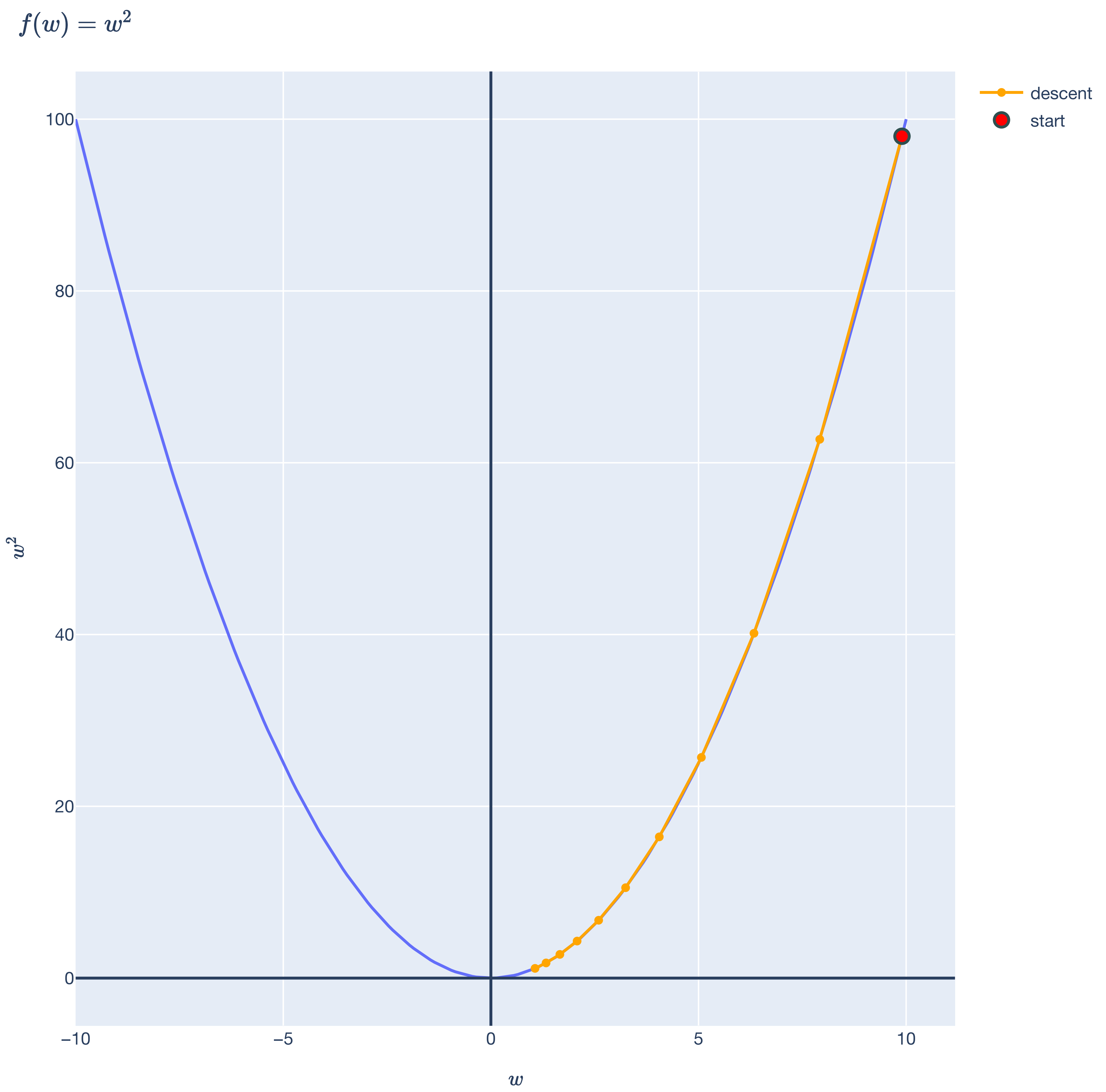
Big Picture: Least Squares

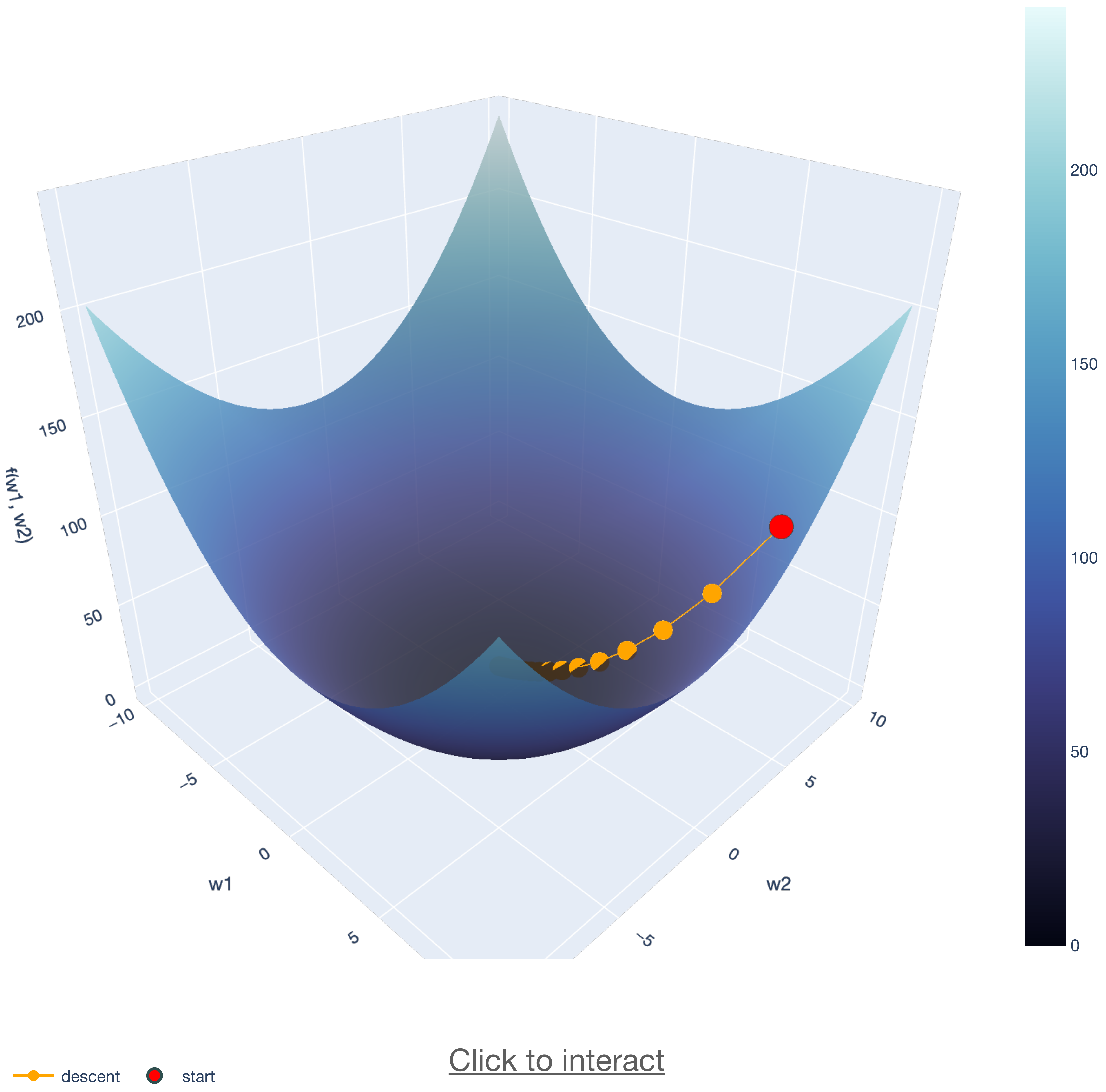


Lesson Overview

Big Picture: Gradient Descent

$SSR = err(w) = \|Xw - y\|^2$





Click to interact

Least Squares

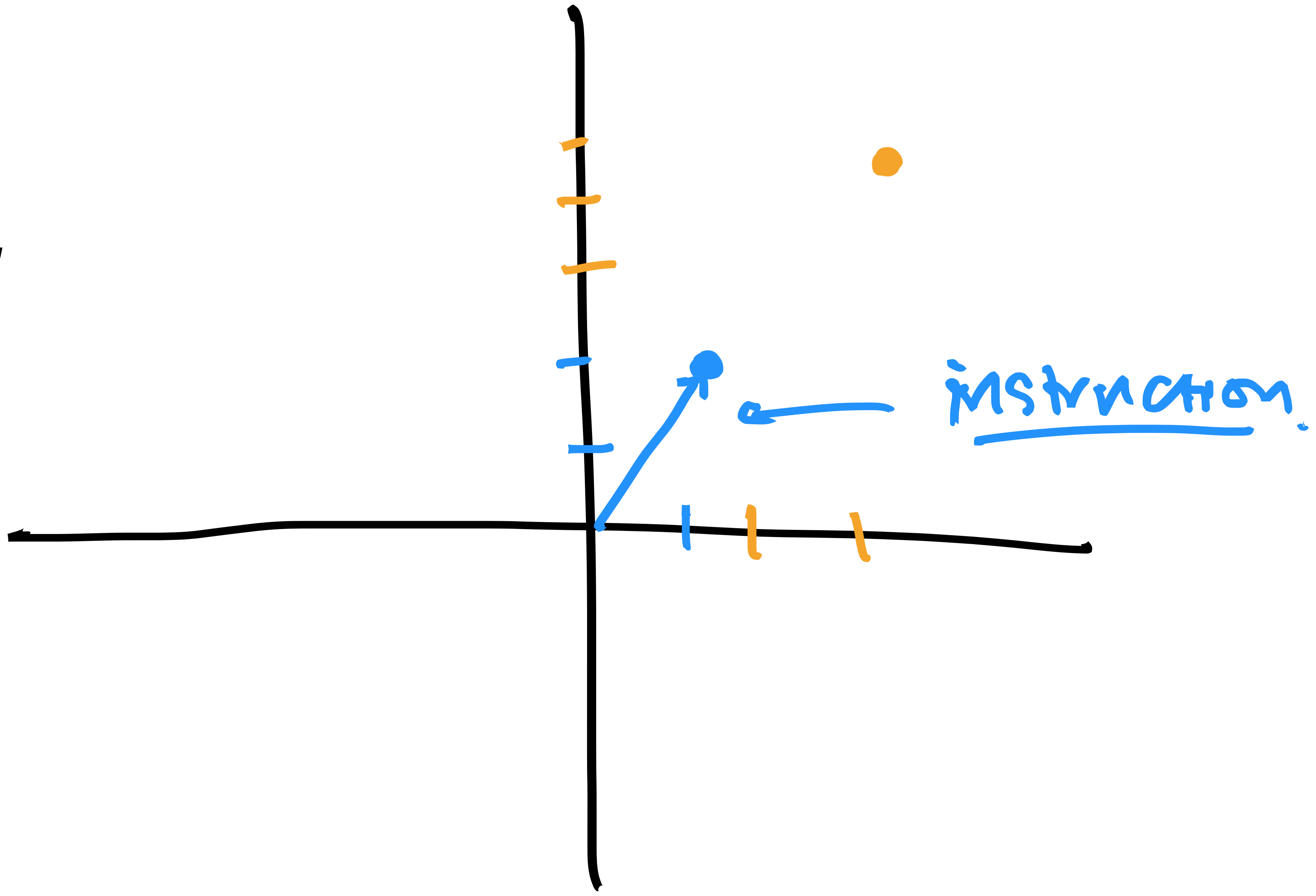
A Quick Review

Vectors

Review from linear algebra

$\mathbb{R}^2$ $v = (1, 2)$ New York
$u = (3, 5)$ Boston.

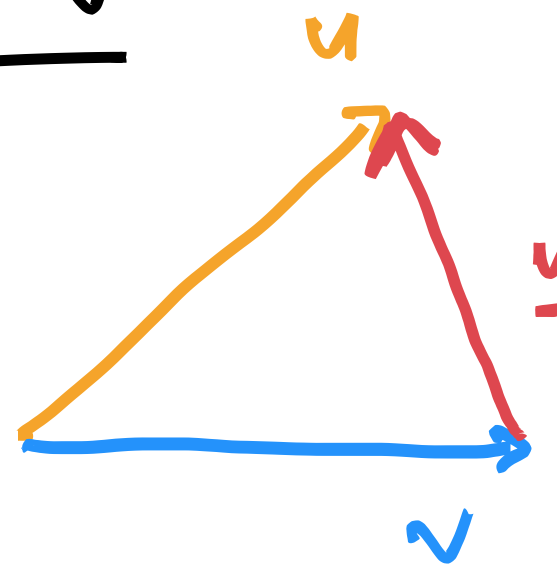
Vectors can interchangeably thought of as *points:*



or “*arrows*”:

Subtraction: $u - v$

points.



$u - v$ → distance: $\|u - v\|$.

⊛ To get a direction from 2 points → $\boxed{\text{SUBTRACT}}$

Regression

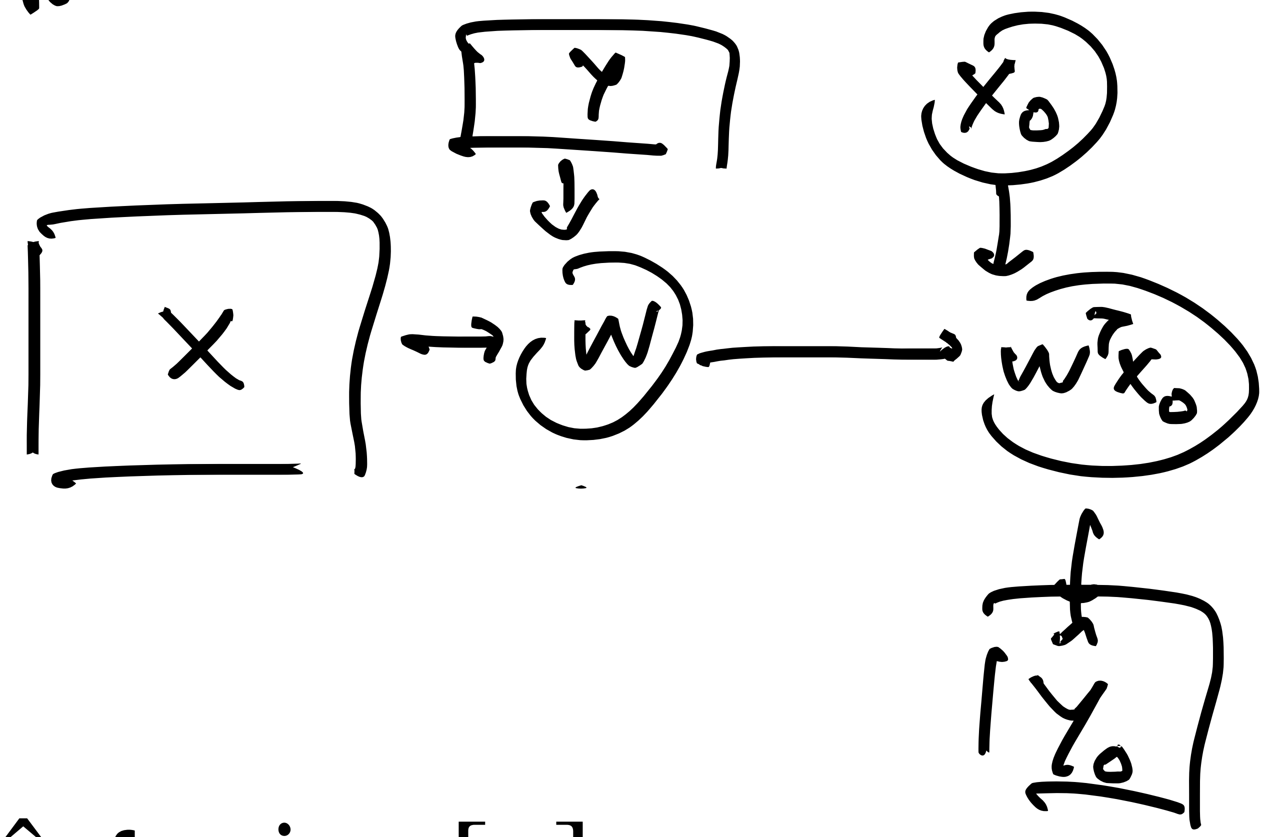
Setup

n datapoints

d features/measurements

Observed: Matrix of *training samples* $\mathbf{X} \in \mathbb{R}^{n \times d}$ and vector of *training labels* $\mathbf{y} \in \mathbb{R}^{\cancel{d}\, n}$.

$$\mathbf{X} = \begin{bmatrix} \uparrow & & \uparrow \\ \mathbf{x}_1 & \cdots & \mathbf{x}_d \\ \downarrow & & \downarrow \end{bmatrix} = \begin{bmatrix} \leftarrow & \mathbf{x}_1^\top & \rightarrow \\ & \vdots & \\ \leftarrow & \mathbf{x}_n^\top & \rightarrow \end{bmatrix}. \Big\} \mathbb{R}^{n \times d}$$



Unknown: *Weight vector* $\mathbf{w} \in \mathbb{R}^d$ with weights $w_1, \ldots, w_d$.

Goal: For each $i \in [n]$, we predict: $\hat{y}_i = \mathbf{w}^\top \mathbf{x}_i = w_1 x_{i1} + \ldots + w_d x_{id} \in \mathbb{R}$.

Choose a weight vector that "fits the training data": $\mathbf{w} \in \mathbb{R}^d$ such that $y_i \approx \hat{y}_i$ for $i \in [n]$, or:

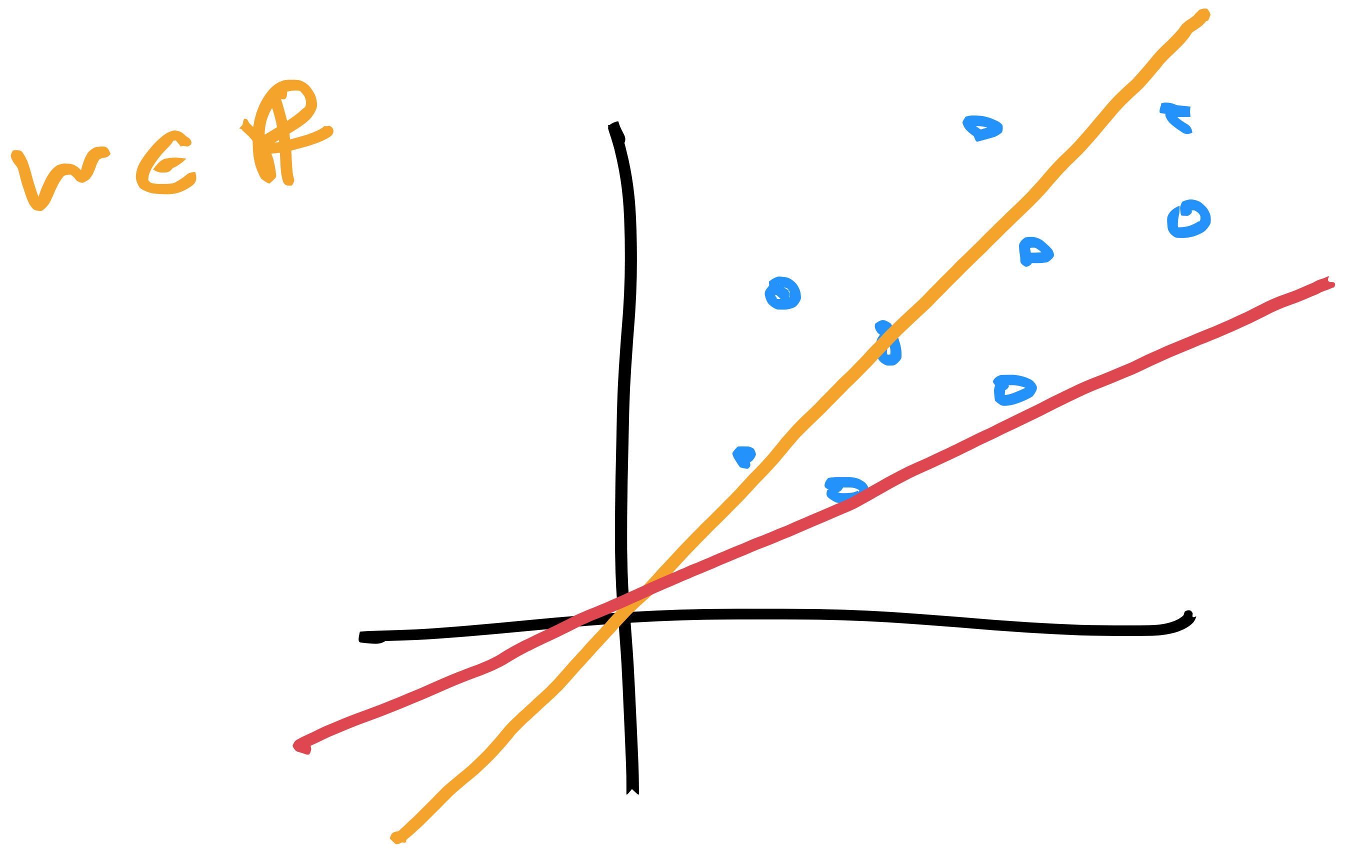
Test: $\boxed{x_0 \quad w^\top x_0 = \hat{y}_0}$

$$\boxed{\mathbf{X}\mathbf{w} = \hat{\mathbf{y}} \approx \mathbf{y}.}$$

$Xw = \hat{y} \approx y$

Regression

A note on intercepts



Goal: For each $i \in [n]$, we predict: $\hat{y}_i = \mathbf{w}^\top \mathbf{x}_i = w_1 x_{i1} + \ldots + w_d x_{id} \in \mathbb{R}$.

This "homogeneous" equation doesn't account for intercepts!

What if we want: $\hat{y}_i = \mathbf{w}^\top \mathbf{x}_i = \underbrace{w_1 x_{i1} + \ldots + w_d x_{id}} + w_0$?

$y = \underset{\text{slope}}{w} x + \underset{\text{y-intercept.}}{w_0}$

Regression

A note on intercepts

Goal: For each $i \in [n]$, we predict: $\hat{y}_i = \mathbf{w}^\top \mathbf{x}_i = w_1 x_{i1} + \ldots + w_d x_{id} \in \mathbb{R}$.

This "homogeneous" equation doesn't account for intercepts!

What if we want: $\hat{y}_i = \mathbf{w}^\top \mathbf{x}_i = w_1 x_{i1} + \ldots + w_d x_{id} + w_0$?

Solution: We modify add a "dummy" 1 to each example:

$$\mathbf{x}_i^\top = \begin{bmatrix} x_{i1} & \ldots & x_{id} & 1 \end{bmatrix}.$$

Same as transforming the data matrix $\mathbf{X} \in \mathbb{R}^{n \times d}$ into $\mathbf{X}' \in \mathbb{R}^{n \times (d+1)}$:

$$\mathbf{X} = \begin{bmatrix} \uparrow & & \uparrow \\ \mathbf{x}_1 & \ldots & \mathbf{x}_d \\ \downarrow & & \downarrow \end{bmatrix} \Longrightarrow \mathbf{X}' = \begin{bmatrix} \uparrow & & \uparrow & 1 \\ \mathbf{x}_1 & \ldots & \mathbf{x}_d & \vdots \\ \downarrow & & \downarrow & 1 \end{bmatrix}$$

$\in \mathbb{R}^{n \times (d+1)}$

Regression

A note on intercepts

Goal: For each $i \in [n]$, we predict: $\hat{y}_i = \mathbf{w}^\top \mathbf{x}_i = w_1 x_{i1} + \ldots + w_d x_{id} \in \mathbb{R}$.

This "homogeneous" equation doesn't account for intercepts!

What if we want: $\hat{y}_i = \mathbf{w}^\top \mathbf{x}_i = w_1 x_{i1} + \ldots + w_d x_{id} + w_0$?

Solution: We modify add a "dummy" 1 to each example:

$$\mathbf{x}_i^\top = \begin{bmatrix} x_{i1} & \ldots & x_{id} & 1 \end{bmatrix}.$$

Same as transforming the data matrix $\mathbf{X} \in \mathbb{R}^{n \times d}$ into $\mathbf{X}' \in \mathbb{R}^{n \times (d+1)}$:

$$\mathbf{X} = \begin{bmatrix} \uparrow & & \uparrow \\ \mathbf{x}_1 & \ldots & \mathbf{x}_d \\ \downarrow & & \downarrow \end{bmatrix} \implies \mathbf{X}' = \begin{bmatrix} \uparrow & & \uparrow & 1 \\ \mathbf{x}_1 & \ldots & \mathbf{x}_d & \vdots \\ \downarrow & & \downarrow & 1 \end{bmatrix}$$

Choose a weight vector that fits $\mathbf{X}'$: $\mathbf{w} \in \mathbb{R}^{d+1}$ such that $y_i \approx \hat{y}_i$ for $i \in [n]$, or:

$w = (w_1, \ldots, w_d, w_0)$

$\mathbf{X}'\mathbf{w} = \hat{\mathbf{y}} \approx \mathbf{y}$. The last $(d+1)$ entry of $\mathbf{w}$ is the intercept, w_0.

Regression

A note on intercepts

Goal: For each $i \in [n]$, we predict: $\hat{y}_i = \mathbf{w}^\top \mathbf{x}_i = w_1 x_{i1} + \ldots + w_d x_{id} + w_0$?

Solution: We modify add a "dummy" 1 to each example:

$$\mathbf{x}_i^\top = \begin{bmatrix} x_{i1} & \ldots & x_{id} & 1 \end{bmatrix}.$$

Same as transforming the data matrix $\mathbf{X} \in \mathbb{R}^{n \times d}$ into $\mathbf{X}' \in \mathbb{R}^{n \times (d+1)}$:

$$\mathbf{X} = \begin{bmatrix} \uparrow & & \uparrow \\ \mathbf{x}_1 & \ldots & \mathbf{x}_d \\ \downarrow & & \downarrow \end{bmatrix} \implies \mathbf{X}' = \begin{bmatrix} \uparrow & & \uparrow & 1 \\ \mathbf{x}_1 & \ldots & \mathbf{x}_d & \vdots \\ \downarrow & & \downarrow & 1 \end{bmatrix}$$

Choose a weight vector that fits $\mathbf{X}'$: $\mathbf{w} \in \mathbb{R}^{d+1}$ such that $y_i \approx \hat{y}_i$ for $i \in [n]$, or:

$\mathbf{X}'\mathbf{w} = \hat{\mathbf{y}} \approx \mathbf{y}$. The last $(d+1)$ entry of $\mathbf{w}$ is the intercept, w_0.

We can always do this WLOG, so we'll focus on the "homogeneous" case.

Least Squares

Summary

$\boxed{\hat{y} = X\hat{w}} \approx y$

sum of squared residuals.

$\hat{y} - y$ $\quad$ $y - \tilde{y}$ $\quad$ $\hat{y} - \tilde{y}$

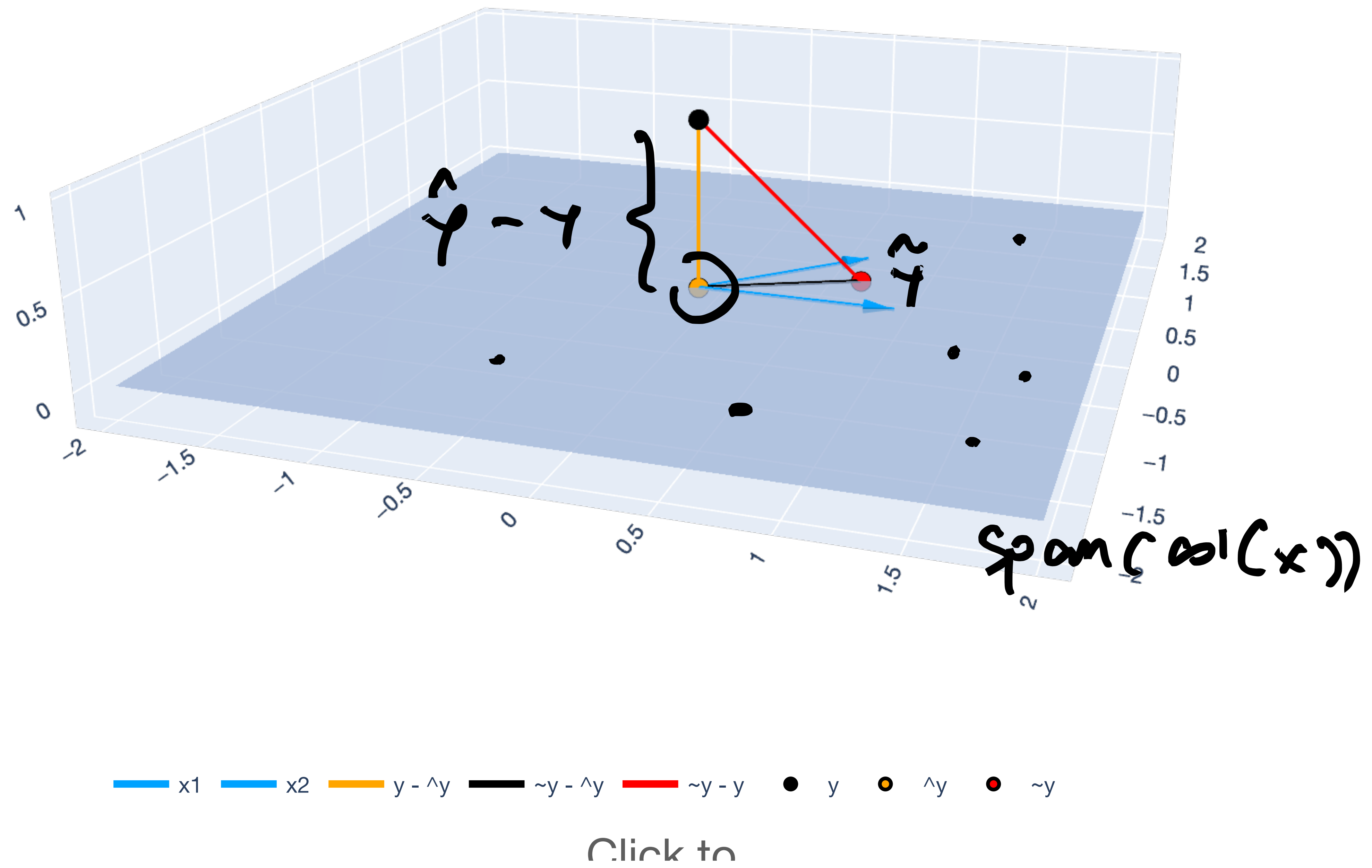
Use the principle of *least squares* to find the $\hat{\mathbf{w}} \in \mathbb{R}^d$ that minimizes

$$\|\hat{\mathbf{y}} - \mathbf{y}\|^2 = \|\mathbf{X}\mathbf{w} - \mathbf{y}\|^2.$$

Using geometric intuition: $\hat{\mathbf{y}}$ *is the vector for which* $\hat{\mathbf{y}} - \mathbf{y}$ *is perpendicular to* $\mathrm{span}(\mathrm{col}(\mathbf{X}))$.

By Pythagorean Theorem, any other vector $\tilde{\mathbf{y}} \in \mathrm{span}(\mathrm{col}(\mathbf{X}))$ gives a larger error:

$$\|\underline{\hat{\mathbf{y}} - \mathbf{y}}\|^2 \leq \|\tilde{\mathbf{y}} - \mathbf{y}\|^2.$$



Because $\hat{\mathbf{y}} - \mathbf{y}$ is perpendicular to $\mathrm{span}(\mathrm{col}(\mathbf{X}))$, we obtain the *normal equations:*

$(\hat{y} - y)^T X$ $\quad$ ($\hat{y} = X\hat{w}$) $\rightarrow$

$$\boxed{\mathbf{X}^\top\mathbf{X}\hat{\mathbf{w}} = \mathbf{X}^\top\mathbf{y}.}$$

If $n \geq d$ and $\mathrm{rank}(\mathbf{X}) = d$, then $\mathbf{X}^\top\mathbf{X}$ *is invertible*, and

$$* \quad \boxed{\hat{\mathbf{w}} = (\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{X}^\top\mathbf{y}.}$$

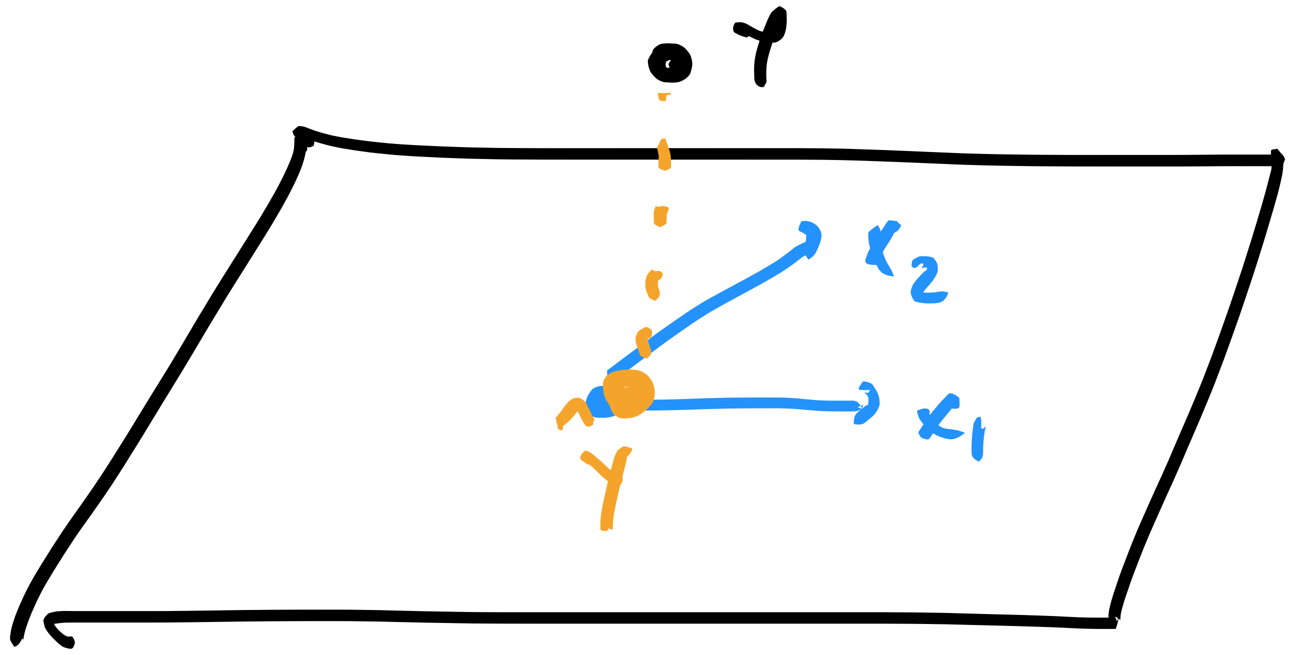
$X\hat{w} = \hat{y}$

$X\tilde{w} = \tilde{y}$

$$\begin{bmatrix} | & & | \\ x_1 & \cdots & x_d \\ | & & | \end{bmatrix} \begin{bmatrix} w_1 \\ \vdots \\ w_d \end{bmatrix}$$

Least Squares

Summary



$\hat{y} = X\hat{w}$

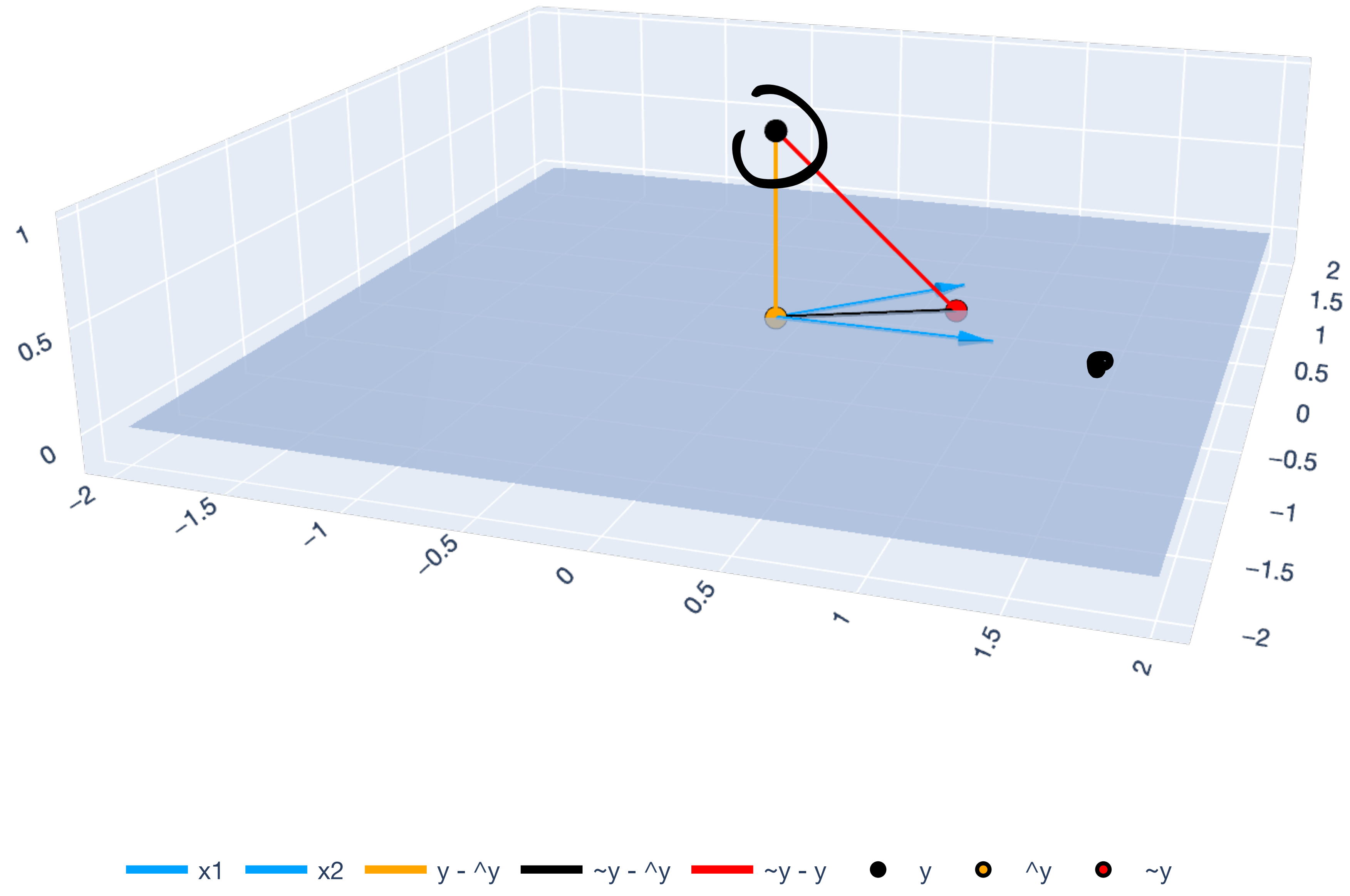
$\boxed{y \neq X\hat{w}}$

$X\hat{w} \approx y$

Use the principle of *least squares* to find the $\hat{\mathbf{w}} \in \mathbb{R}^d$ that minimizes

$$\|\hat{\mathbf{y}} - \mathbf{y}\|^2 = \|\mathbf{X}\mathbf{w} - \mathbf{y}\|^2.$$

Using geometric intuition: $\hat{\mathbf{y}}$ *is the vector for which* $\hat{\mathbf{y}} - \mathbf{y}$ *is perpendicular to* $\mathrm{span}(\mathrm{col}(\mathbf{X}))$.



By Pythagorean Theorem, any other vector $\tilde{\mathbf{y}} \in \mathrm{span}(\mathrm{col}(\mathbf{X}))$ gives a larger error:

$$\|\hat{\mathbf{y}} - \mathbf{y}\|^2 \leq \|\tilde{\mathbf{y}} - \mathbf{y}\|^2.$$

Because $\hat{\mathbf{y}} - \mathbf{y}$ is perpendicular to $\mathrm{span}(\mathrm{col}(\mathbf{X}))$, we obtain the *normal equations:*

$$\mathbf{X}^\top\mathbf{X}\hat{\mathbf{w}} = \mathbf{X}^\top\mathbf{y}.$$

If $n \geq d$ and $\mathrm{rank}(\mathbf{X}) = d$, then $\mathbf{X}^\top\mathbf{X}$ *is invertible*, and

$$\hat{\mathbf{w}} = (\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{X}^\top\mathbf{y}.$$

$\boxed{X\hat{w} = y}$ $d \times n$ n

$\hat{y} \approx y$ $\|\hat{y} - y\|^2 \rightarrow X^T X\hat{w} = \boxed{X^T y} \in \mathbb{R}^d$

Least Squares

First missing item: invertibility of $\mathbf{X}^\top \mathbf{X}$

$\boxed{\text{If } n \geq d \text{ and } \mathrm{rank}(\mathbf{X}) = d\text{, then } \mathbf{X}^\top \mathbf{X} \textit{ is invertible.}}$

"If there are no redundant features, then we can invert the normal equations"

Subspaces

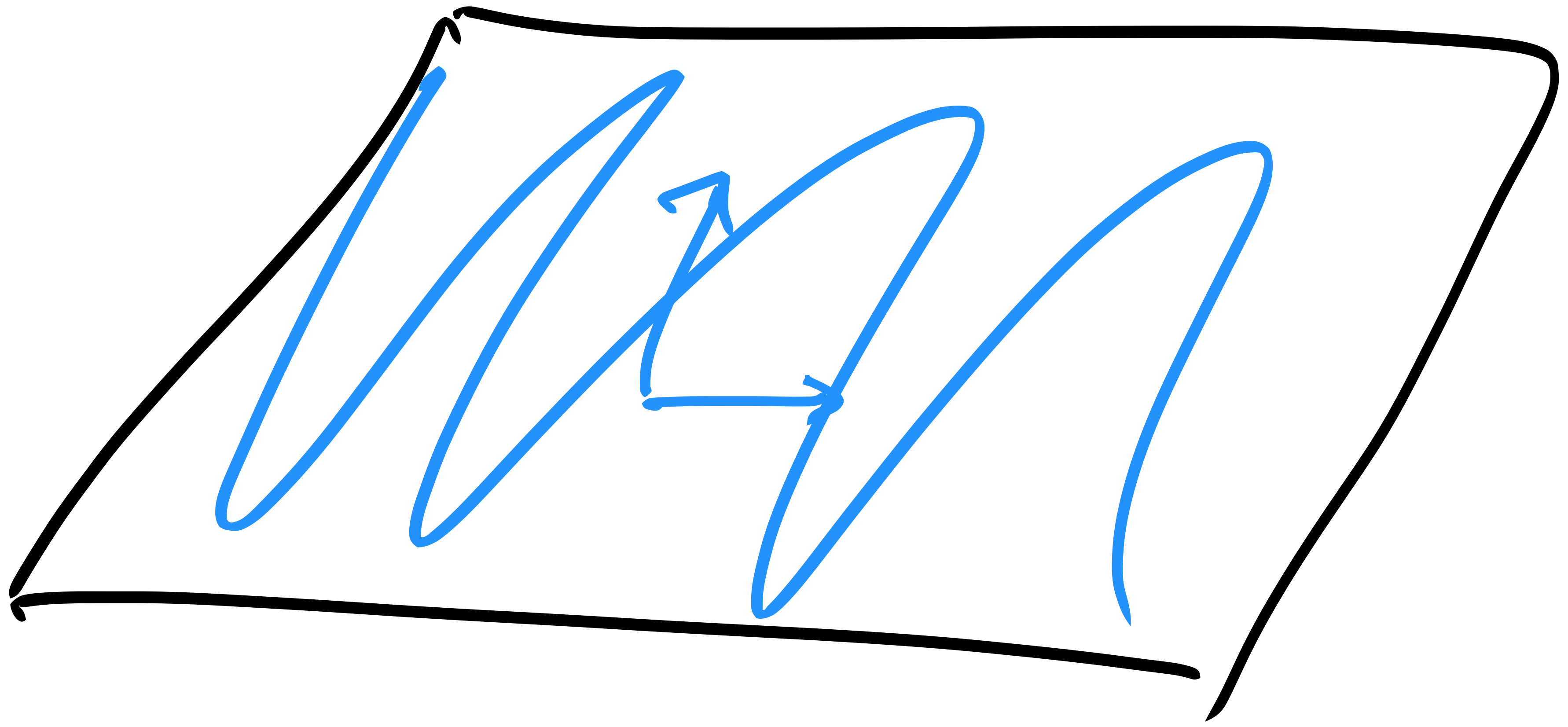
Subspaces

Idea

$\mathbb{R}^n$ = Euclidean space (n dimensions)

$S \subseteq \mathbb{R}^n$

A ***subspace*** is a set of vectors that "stays within" the set under all linear combinations of the vectors.



Subspaces

Definition

A ***subspace*** $\mathcal{S} \subseteq \mathbb{R}^n$ is a subset of vectors that satisfies the property: if $\mathbf{v}, \mathbf{w} \in \mathcal{S}$, then $\alpha\mathbf{v} + \beta\mathbf{w} \in \mathcal{S}$ for any $\alpha, \beta \in \mathbb{R}$.

$v - v = \vec{0}$

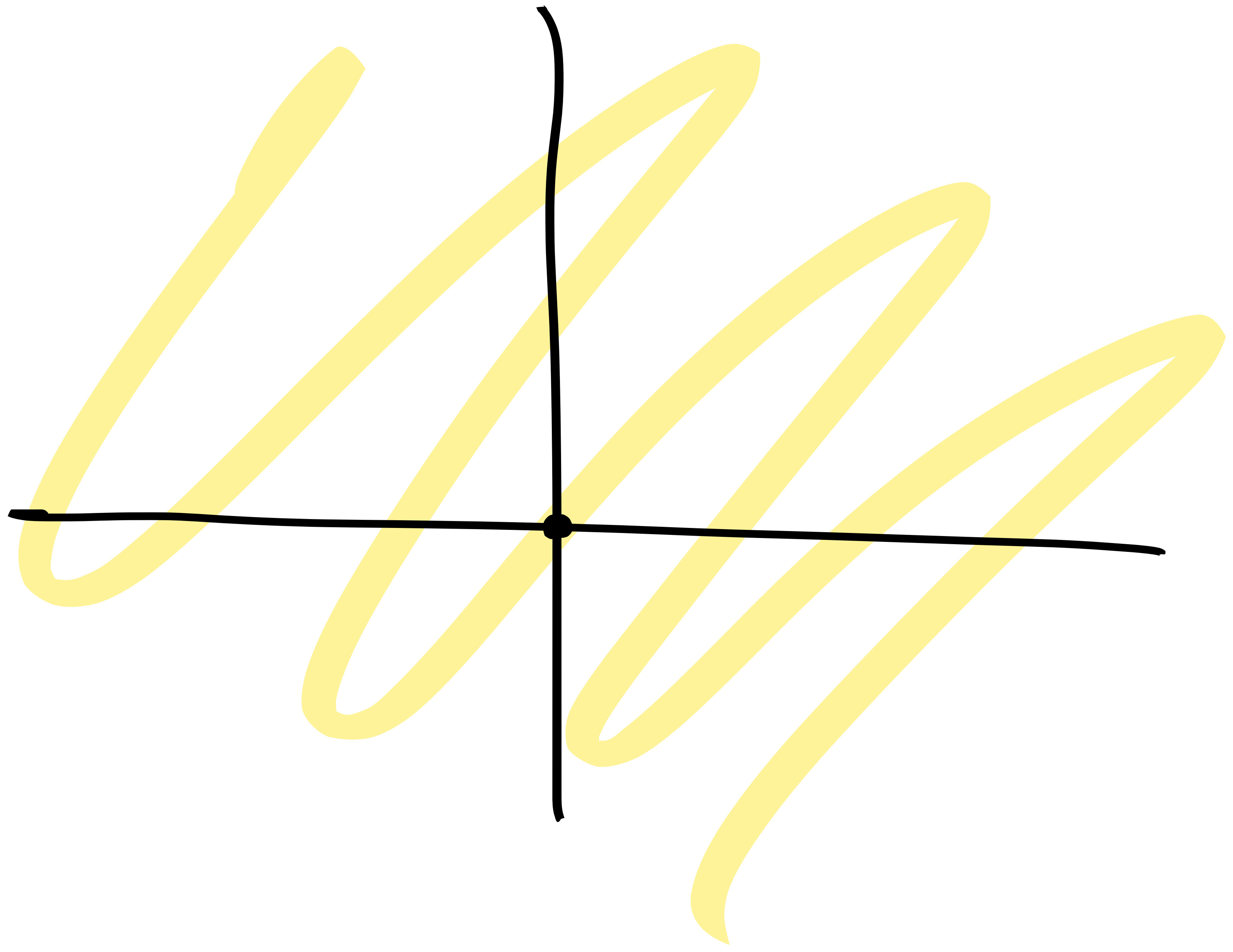
Any subspace $\mathcal{S}$ contains the zero vector: $\mathbf{0} \in \mathcal{S}$.

Subspaces

Examples

$\mathbb{R}^2 \subseteq \mathbb{R}^2$

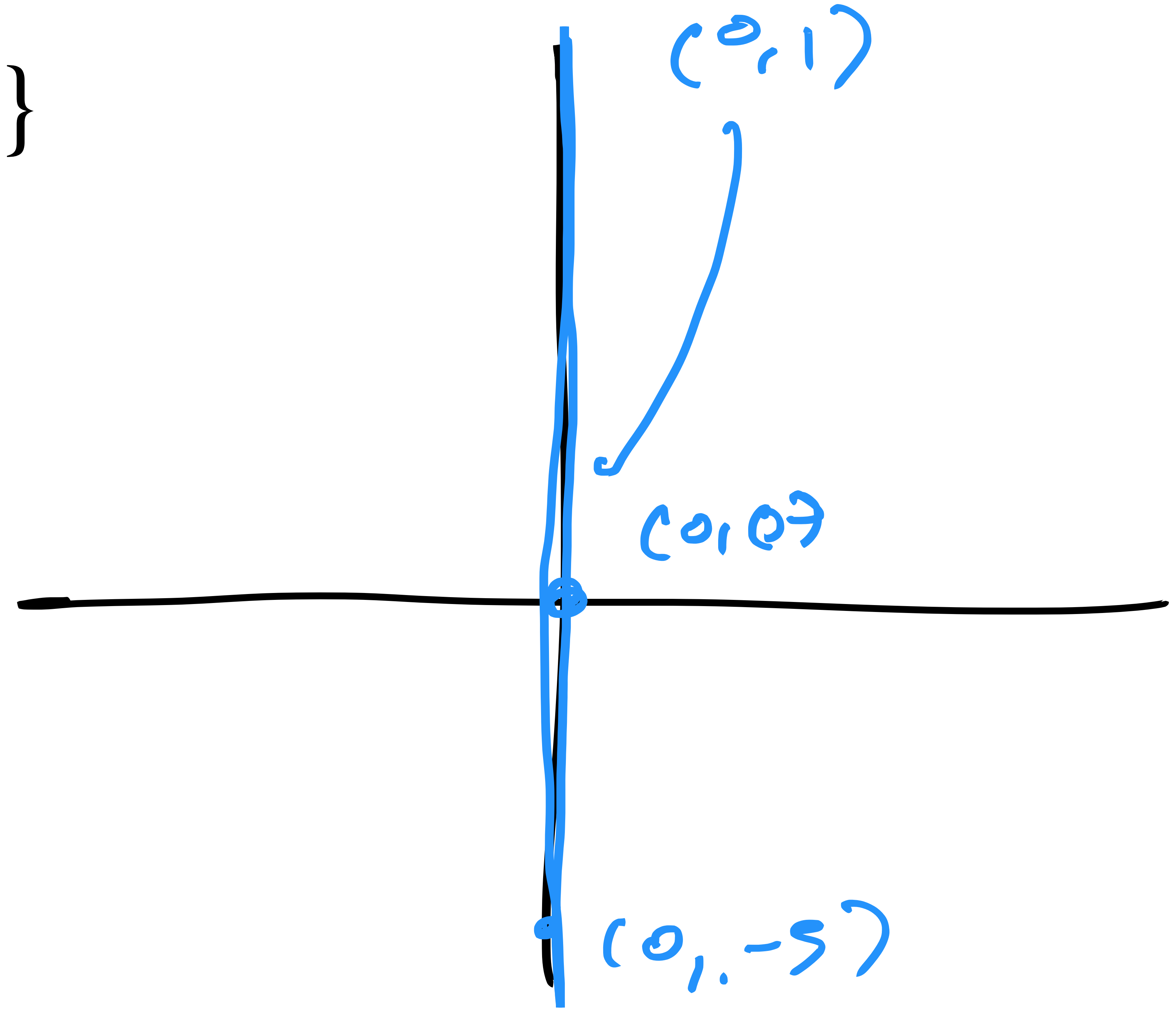
Example: $\mathcal{S}_0 := \mathbb{R}^2$



Subspaces

Examples

Example: $\mathcal{S}_1 := \{\mathbf{v} \in \mathbb{R}^2 : v_1 = 0\}$

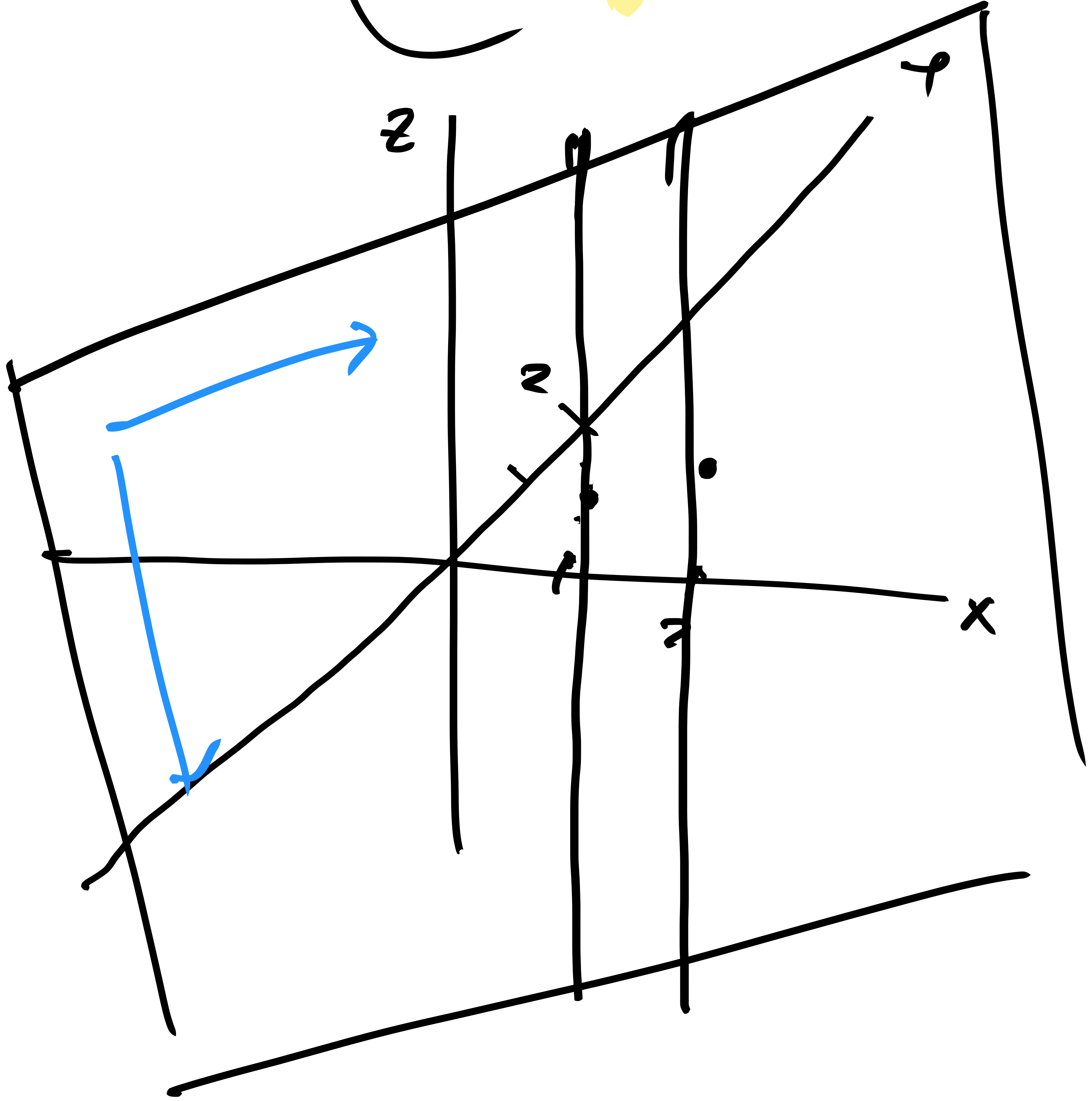


Subspaces

Examples

Example: $\mathcal{S}_2 := \{\mathbf{v} \in \mathbb{R}^3 : v_1 = v_2\}$

(1, 1, 2)
(1, 1, 3)
(2, 2, 2)



Span

Review

For a collection of vectors $\mathbf{a}_1, \ldots, \mathbf{a}_d \in \mathbb{R}^n$, the ***span*** is the set of vectors we can attain through linear combinations of $\mathbf{a}_1, \ldots, \mathbf{a}_d$.

$$\boxed{\text{span}(\mathbf{a}_1, \ldots, \mathbf{a}_d) = \left\{ \mathbf{y} \in \mathbb{R}^n : \mathbf{y} = \sum_{i=1}^{d} \alpha_i \mathbf{a}_i, \alpha_i \in \mathbb{R} \right\}}.$$

Recall that this is equivalent to all the $\mathbf{y} \in \mathbb{R}^{n \times d}$ we obtain from matrix vector multiplication!

$$\mathbf{y} = \mathbf{A}\alpha, \text{ i.e. } \begin{bmatrix} y_1 \\ \vdots \\ y_n \end{bmatrix} = \begin{bmatrix} \uparrow & \uparrow & \uparrow \\ \mathbf{a}_1 & \ldots & \mathbf{a}_d \\ \downarrow & \downarrow & \downarrow \end{bmatrix} \begin{bmatrix} \alpha_1 \\ \vdots \\ \alpha_d \end{bmatrix} = \alpha_1 \vec{a}_1 + \ldots + \alpha_d \vec{a}_d$$

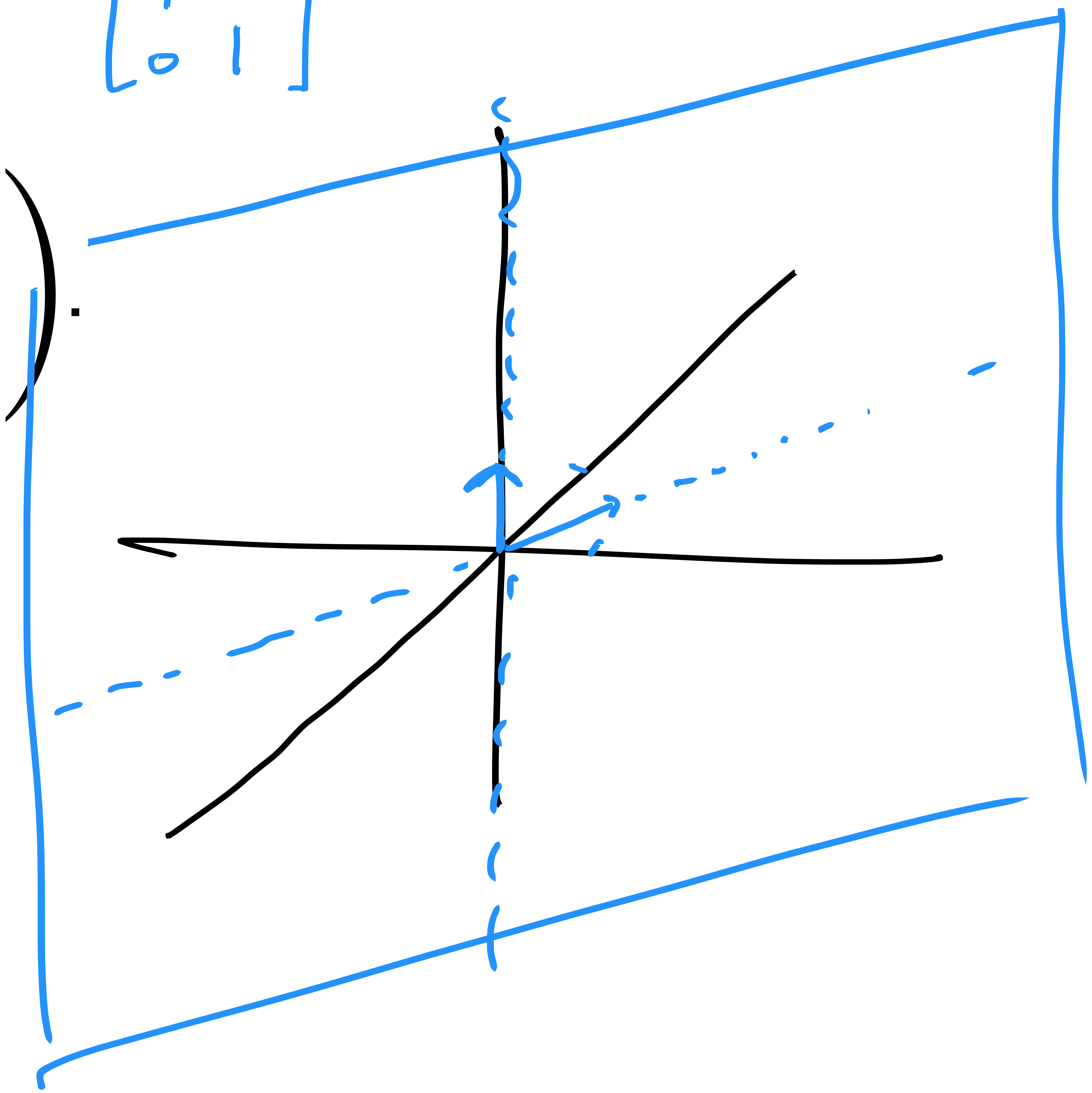
Linear combinations

Subspaces

Examples

$$\overset{3\times 2}{\begin{bmatrix}1&0\\1&0\\0&1\end{bmatrix}}$$

Example: $\mathscr{S}_3 := \operatorname{span}\left(\begin{bmatrix}1\\1\\0\end{bmatrix}, \begin{bmatrix}0\\0\\1\end{bmatrix}\right).$

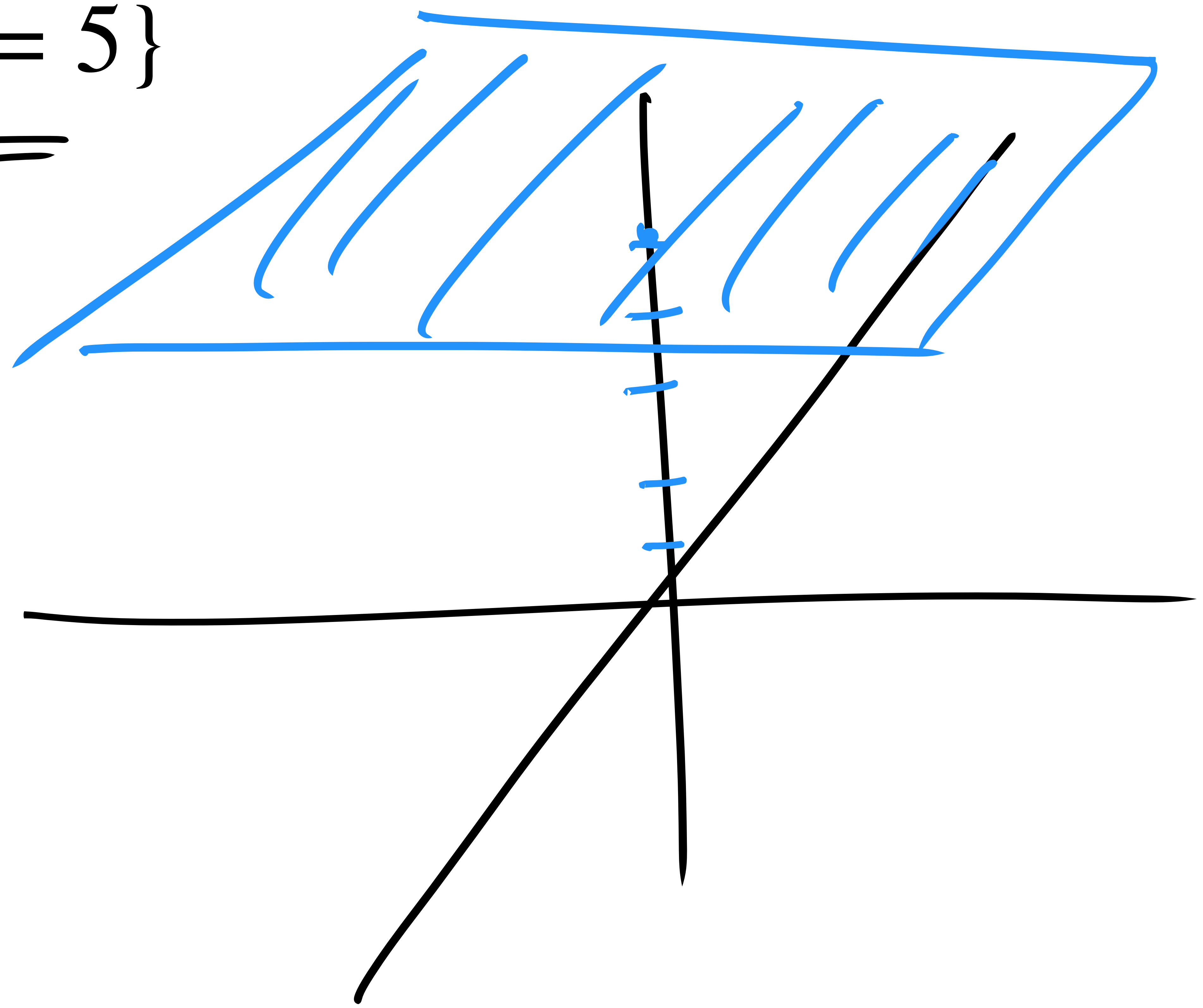


Subspaces

Examples

(Non)Example: $\mathcal{S}_4 := \{\mathbf{v} \in \mathbb{R}^3 : v_3 = 5\}$

(0, 0, 0)



Subspaces

Specific example: $\underbrace{\mathrm{span}(\mathrm{col}(\mathbf{X}))}_{\text{columnspace}}$

$$\begin{bmatrix} | & & | \\ x_1 & \cdots & x_d \\ | & & | \end{bmatrix}$$

Let $\mathbf{X} \in \mathbb{R}^{n \times d}$. The columns are $\mathbf{x}_1, \ldots, \mathbf{x}_d \in \mathbb{R}^n$.

$$\mathrm{span}(\mathrm{col}(\mathbf{X})) = \{\mathbf{y} \in \mathbb{R}^n : \mathbf{y} = w_1\mathbf{x}_1 + \ldots + w_d\mathbf{x}_d\}$$

* MATRIX VECTOR MULTIPLICATION

$\downarrow$

$\{y \in \mathbb{R}^n : y = Xw\}$.

Bases & Dimension

Basis

Idea

FINITE

For a subspace $\mathcal{S}$, a ***basis*** is a *minimal* set of vectors that can "linearly describe" *any* vector in $\mathcal{S}$. A "language" for vectors in $\mathcal{S}$.

Basis

Linear Independence and Span

equivalent definition of linear dependence

$$\vec{a}_j = \alpha_1 \vec{a}_1 + \ldots + \alpha_{j-1}\vec{a}_{j-1} + \alpha_{j+1}\vec{a}_{j+1} + \ldots + \alpha_d \vec{a}_d$$

Recall the following two notions.

A collection of vectors $\mathbf{a}_1, \ldots, \mathbf{a}_d \in \mathbb{R}^n$ is ***linearly independent*** if $\alpha_1 \mathbf{a}_1 + \ldots + \alpha_d \mathbf{a}_d = \mathbf{0}$ if and only if $\alpha_i = 0$ for all $i \in [d]$.

For a collection of vectors $\mathbf{a}_1, \ldots, \mathbf{a}_d \in \mathbb{R}^n$, the ***span*** is the set of vectors we can attain through linear combinations of $\mathbf{a}_1, \ldots, \mathbf{a}_d$:

$$\mathrm{span}(\mathbf{a}_1, \ldots, \mathbf{a}_d) = \left\{ \mathbf{y} \in \mathbb{R}^n : \mathbf{y} = \sum_{i=1}^{d} \alpha_i \mathbf{a}_i, \alpha_i \in \mathbb{R} \right\}.$$

Basis

Definition

For a subspace $\mathscr{S} \subseteq \mathbb{R}^n$, a set of vectors $\mathbf{a}_1, \ldots, \mathbf{a}_d \in \mathscr{S}$ is a ***basis*** for $\mathscr{S}$ if:

$$\underbrace{\mathscr{S} = \operatorname{span}(\mathbf{a}_1, \ldots, \mathbf{a}_d)} \text{ and } \underbrace{\mathbf{a}_1, \ldots, \mathbf{a}_d \text{ are linearly independent.}}$$

Bases are not unique — there are infinitely many bases for any subspace.

However, all bases have the same number of elements. → dimension.

Basis

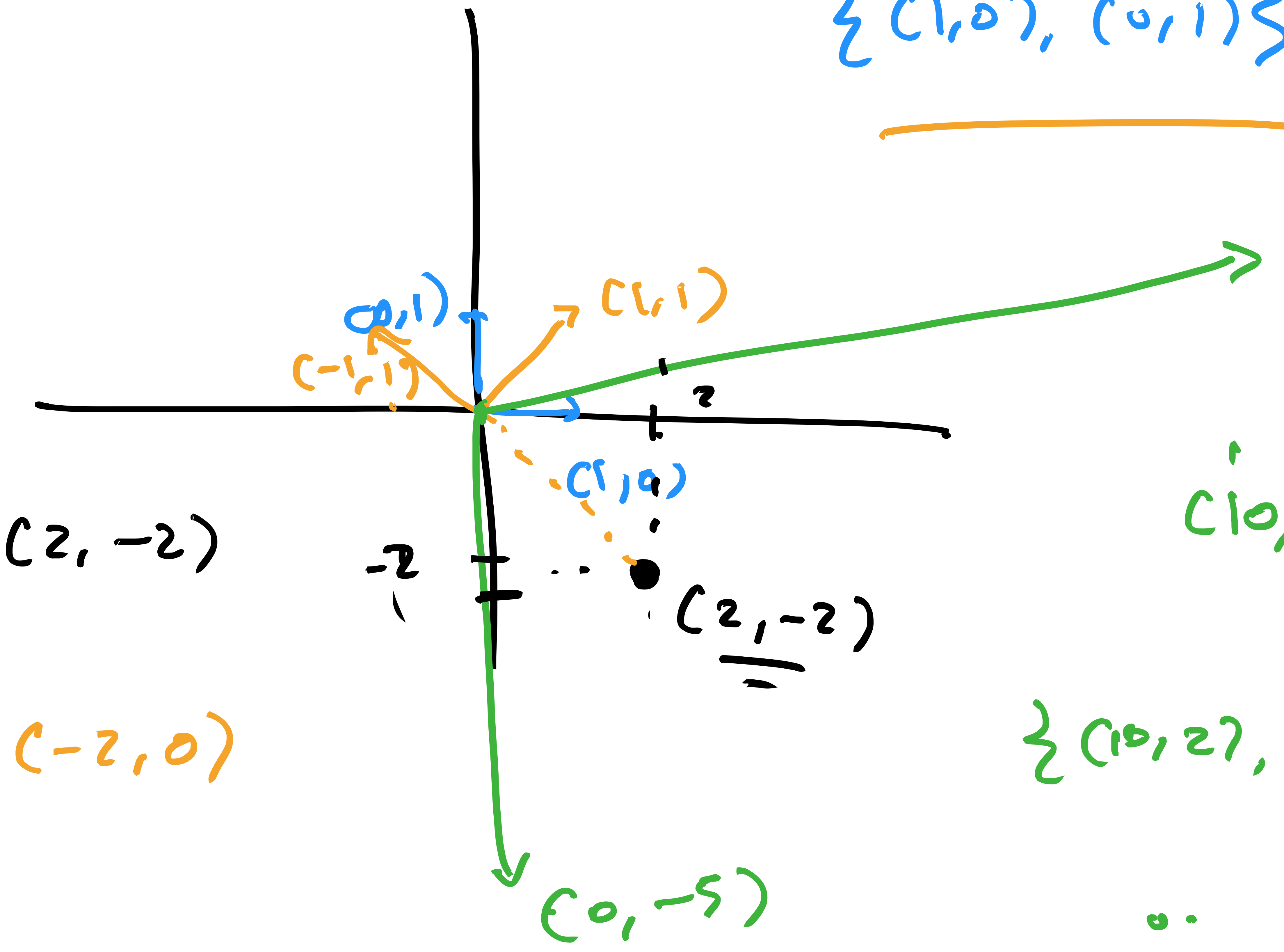
Examples

Example: $\mathscr{S}_0 := \mathbb{R}^2$

$\{$

$$\begin{bmatrix} 2 \\ -2 \end{bmatrix} = 2\begin{bmatrix} 1 \\ 0 \end{bmatrix} - 2\begin{bmatrix} 0 \\ 1 \end{bmatrix} \to (2,-2)$$

$$\begin{bmatrix} 2 \\ -2 \end{bmatrix} = -2\begin{bmatrix} -1 \\ 1 \end{bmatrix} + 0\begin{bmatrix} 1 \\ 1 \end{bmatrix} \to (-2,0)$$



$\{(1,0),\ (0,1)\}$

$(10,2)$

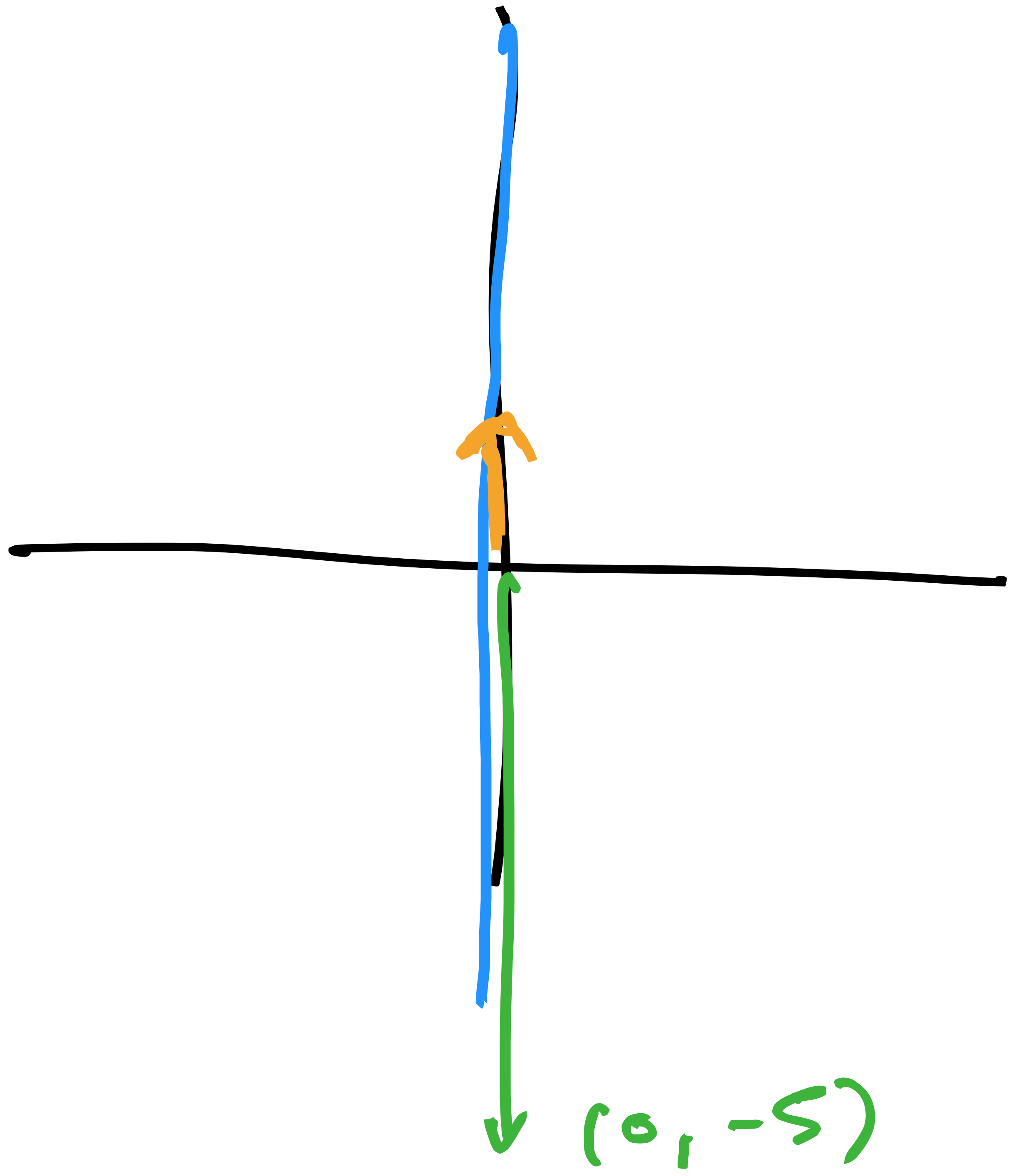
$\{(10,2),\ (0,-5)\}$

Basis

Examples

$\mathrm{span}(v_1, v_2) = \{ y \in \mathbb{R}^n : \alpha_1 v_1 + \alpha_2 v_2 = y \}$

Example: $\mathcal{S}_1 := \{\mathbf{v} \in \mathbb{R}^2 : v_1 = 0\}$



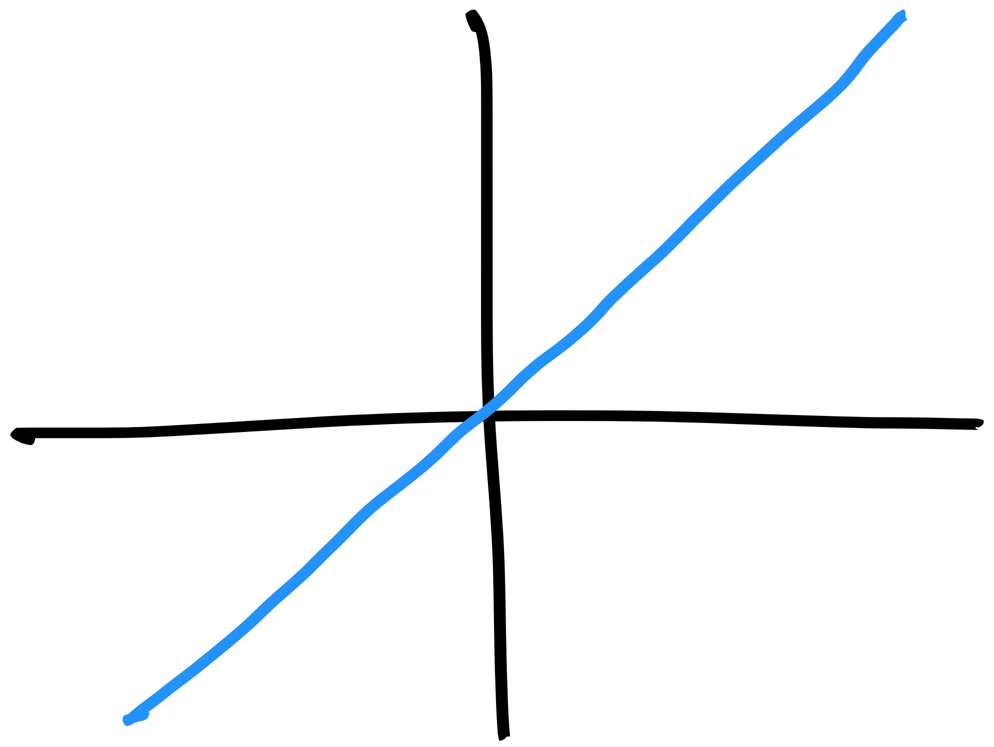
Basis

Examples

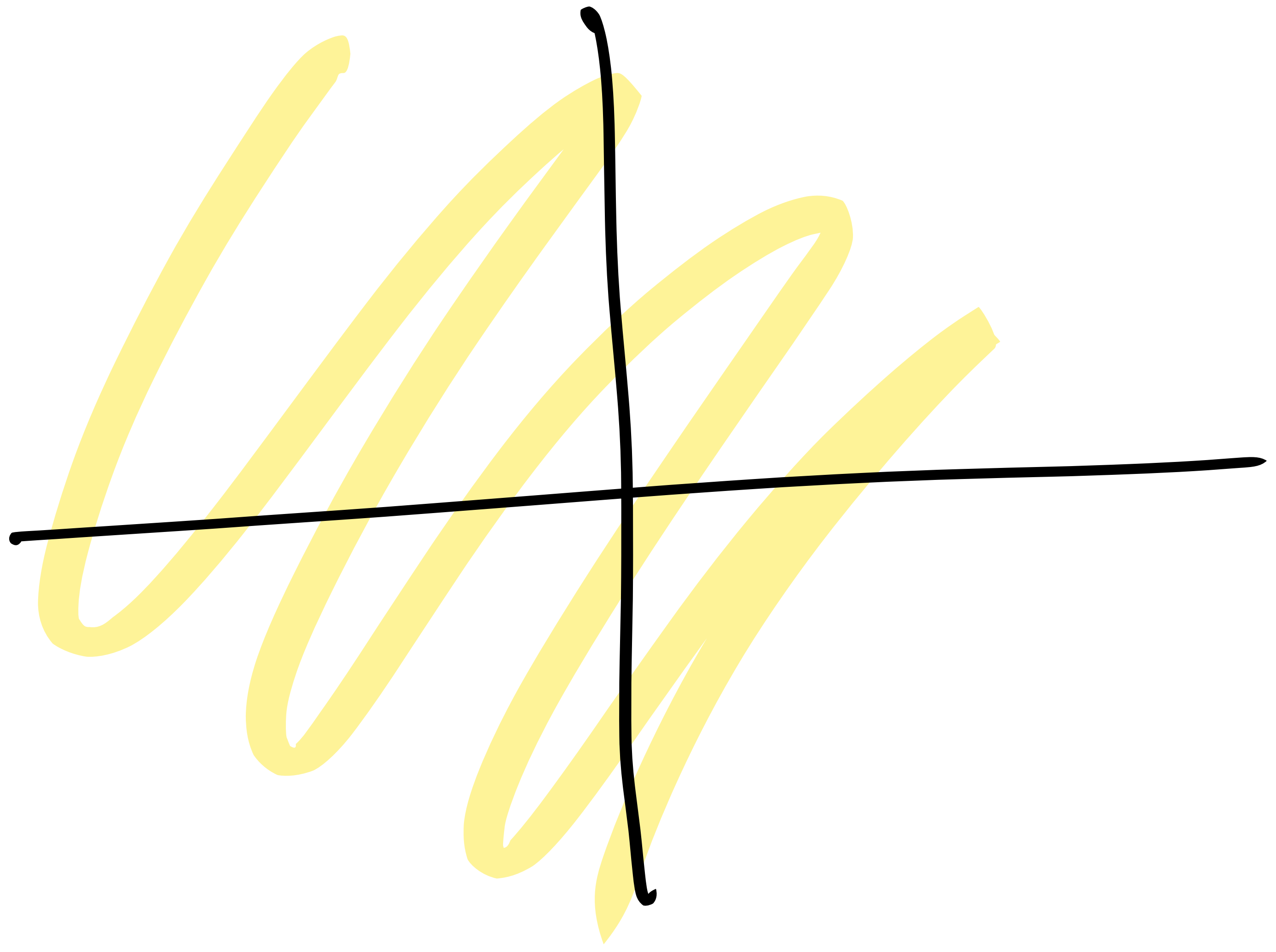
Example: $\mathcal{S}_2 := \{\mathbf{v} \in \mathbb{R}^3 : v_1 = v_2\}$

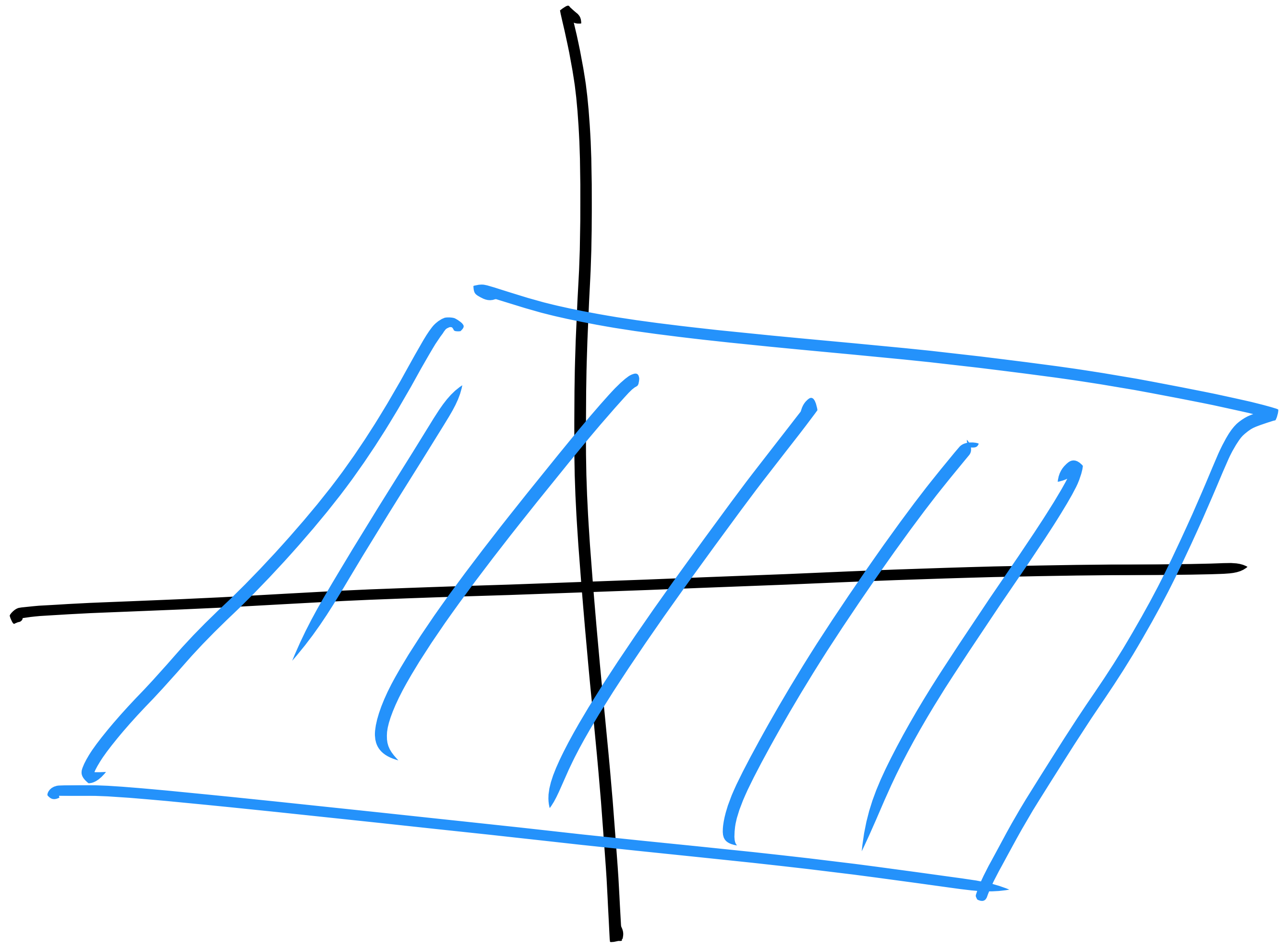
Dimension of a Subspace

Definition



The ***dimension*** of a subspace is the size of any of its bases. For a subspace $\mathscr{S}$, write this as $\dim(\mathscr{S})$.





Matrices & Subspaces

Every matrix comes with four subspaces

$$\begin{bmatrix} & X & \end{bmatrix}$$

Let $\mathbf{X} \in \mathbb{R}^{n \times d}$ be a matrix.

span(col(X))

Its ***columnspace*** is $\mathrm{col}(\mathbf{X}) = \{\mathbf{y} \in \mathbb{R}^{n} : \mathbf{y} = \mathbf{X}\mathbf{w},$ for any $\mathbf{w} \in \mathbb{R}^d\}$.

linear combination of columns
||
Any matrix-vector product.

Its ***nullspace/kernel*** is $\ker(\mathbf{X}) := \{\mathbf{w} \in \mathbb{R}^{d} : \mathbf{X}\mathbf{w} = \mathbf{0}\}$.

Its rowspace is $\mathrm{col}(\mathbf{X}^\top) = \{\mathbf{y} \in \mathbb{R}^d : \mathbf{y} = \mathbf{X}^\top \mathbf{v},$ for any $\mathbf{v} \in \mathbb{R}^n\}$.

Its *left nullspace* is $\ker(\mathbf{X}^\top) := \{\mathbf{v} \in \mathbb{R}^n : \mathbf{X}^\top \mathbf{v} = \mathbf{0}\}$.

Rank-nullity theorem: $\underbrace{n = \dim(\mathrm{col}(\mathbf{X})) + \dim(\ker(\mathbf{X}))}$.

Matrices & Subspaces

Columnspace of a matrix

Let $\mathbf{X} \in \mathbb{R}^{n \times d}$ be a matrix, with columns $\mathbf{x}_1, \ldots, \mathbf{x}_d \in \mathbb{R}^n$.

We can think of its columnspace as:

$$\begin{aligned} \mathrm{col}(\mathbf{X}) := & \{\mathbf{y} \in \mathbb{R}^n : \mathbf{y} = \mathbf{X}\mathbf{w}, \text{ for any } \mathbf{w} \in \mathbb{R}^d\} \quad \text{Matrix-vector mult.} \\ = & \{\mathbf{y} \in \mathbb{R}^n : \mathbf{y} = w_1\mathbf{x}_1 + \ldots + w_d\mathbf{x}_d, \text{ for any } w_i \in \mathbb{R}\} \quad \text{linear combo.} \\ = & \underbrace{\mathrm{span}(\mathbf{x}_1, \ldots, \mathbf{x}_d)} \end{aligned}$$

This is a subspace that "comes with" any matrix.

Matrices & Subspaces

Rank of a matrix

Let $\mathbf{X} \in \mathbb{R}^{n \times d}$ be a matrix, with columns $\mathbf{x}_1, \ldots, \mathbf{x}_d \in \mathbb{R}^n$.

The ***rank*** of $\mathbf{X}$ is the number of linearly independent columns (which is the same as the number of linearly independent rows).

It is always the case that: $\mathrm{rank}(\mathbf{X}) \leq \min\{n, d\}$. If $\mathrm{rank}(\mathbf{X}) = \min\{n, d\}$, then we say $\mathbf{X}$ is *full rank.*

Matrices & Subspaces

Rank & Invertibility

Let $\mathbf{X} \in \mathbb{R}^{d \times d}$ be a square matrix.

It is always the case that: $\mathrm{rank}(\mathbf{X}) \leq d$. If $\mathrm{rank}(\mathbf{X}) = d$, then we say $\mathbf{X}$ is *full rank.*

Basic fact from linear algebra:

$\mathbf{X}$ *is invertible if and only if it is full rank.*

Matrices & Subspaces

Dimension of the columnspace

Let $\mathbf{X} \in \mathbb{R}^{n \times d}$ be a matrix, with columns $\mathbf{x}_1, \ldots, \mathbf{x}_d \in \mathbb{R}^n$.

$$\mathrm{col}(\mathbf{X}) = \mathrm{span}(\mathbf{x}_1, \ldots, \mathbf{x}_d)$$

$\mathrm{rank}(\mathbf{X})$ = how many of $\mathbf{x}_1, \ldots, \mathbf{x}_d$ are linearly independent

So, if $\mathrm{rank}(\mathbf{X}) = d$, then $\mathbf{x}_1, \ldots, \mathbf{x}_d$ form a *basis for the columnspace*!

Least Squares

First missing item: invertibility of $\mathbf{X}^\top\mathbf{X}$

$$\boxed{X^\top X w = X^\top y}$$

If $n \geq d$ and $\text{rank}(\mathbf{X}) = d$, then $\mathbf{X}^\top\mathbf{X}$ *is invertible.*

"If there are no redundant features, then we can invert the normal equations"

Least Squares

First missing item: invertibility of $\mathbf{X}^\top \mathbf{X}$

Theorem (Invertibility of $\mathbf{X}^\top \mathbf{X}$). Let $\mathbf{X} \in \mathbb{R}^{n \times d}$ be a matrix, with columns $\mathbf{x}_1, \ldots, \mathbf{x}_d \in \mathbb{R}^n$. If $n \geq d$ and $\operatorname{rank}(\mathbf{X}) = d$, then $\mathbf{X}^\top \mathbf{X}$ *is invertible.* $\longrightarrow \mathbb{R}^{d \times d}$

Proof. To show that $\mathbf{X}^\top \mathbf{X}$ is invertible, show $\operatorname{rank}(\mathbf{X}^\top \mathbf{X}) = d$.

Least Squares

First missing item: invertibility of $\mathbf{X}^\top\mathbf{X}$

$w_1 \vec{x}_1 + \ldots + w_d \vec{x}_d = \vec{0}$

$\Updownarrow$

$w_1, \ldots, w_d = 0.$

Theorem (Invertibility of $\mathbf{X}^\top\mathbf{X}$). Let $\mathbf{X} \in \mathbb{R}^{n \times d}$ be a matrix, with columns $\mathbf{x}_1, \ldots, \mathbf{x}_d \in \mathbb{R}^n$. If $n \geq d$ and $\operatorname{rank}(\mathbf{X}) = d$, then $\mathbf{X}^\top\mathbf{X}$ *is invertible.*

Proof. To show that $\mathbf{X}^\top\mathbf{X}$ is invertible, show $\mathbf{X}^\top\mathbf{X}$ has d linearly independent columns.

$$\boxed{\mathbf{X}^\top\mathbf{X}\mathbf{w} = \mathbf{0} \iff \mathbf{w} = \mathbf{0}.}$$

$$\begin{bmatrix} X^T X \end{bmatrix} \begin{bmatrix} w_1 \\ \vdots \\ w_d \end{bmatrix} = \vec{0}$$

Least Squares

First missing item: invertibility of $\mathbf{X}^\top\mathbf{X}$

Theorem (Invertibility of $\mathbf{X}^\top\mathbf{X}$). Let $\mathbf{X} \in \mathbb{R}^{n\times d}$ be a matrix, with columns $\mathbf{x}_1, \ldots, \mathbf{x}_d \in \mathbb{R}^n$. If $n \geq d$ and $\underbrace{\operatorname{rank}(\mathbf{X}) = d}_{X \text{ has } d \text{ lin indep}}$, then $\mathbf{X}^\top\mathbf{X}$ *is invertible.*

Proof. To show that $\mathbf{X}^\top\mathbf{X}$ is invertible, show $\mathbf{X}^\top\mathbf{X}$ has d linearly independent columns.

If ... then:

$$\mathbf{X}^\top\mathbf{X}\mathbf{w} = \mathbf{0} \implies \mathbf{w} = \mathbf{0}.$$

Suppose $\mathbf{X}^\top\mathbf{X}\mathbf{w} = \mathbf{0}$. Let $\mathbf{w} \in \mathbb{R}^d$ be any vector.

Least Squares

First missing item: invertibility of $\mathbf{X}^\top \mathbf{X}$

Theorem (Invertibility of $\mathbf{X}^\top \mathbf{X}$). Let $\mathbf{X} \in \mathbb{R}^{n \times d}$ be a matrix, with columns $\mathbf{x}_1, \ldots, \mathbf{x}_d \in \mathbb{R}^n$. If $n \geq d$ and $\mathrm{rank}(\mathbf{X}) = d$, then $\mathbf{X}^\top \mathbf{X}$ *is invertible.*

Proof. To show that $\mathbf{X}^\top \mathbf{X}$ is invertible, show $\mathbf{X}^\top \mathbf{X}$ has d linearly independent columns.

$$\mathbf{X}^\top \mathbf{X} \mathbf{w} = \mathbf{0} \implies \mathbf{w} = \mathbf{0}.$$

Suppose $\mathbf{X}^\top \mathbf{X} \mathbf{w} = \mathbf{0}$. Let $\mathbf{w} \in \mathbb{R}^d$ be any vector. Take a dot product of both sides with $\mathbf{w}$;

$$\mathbf{w}^\top \mathbf{X}^\top \mathbf{X} \mathbf{w} = \mathbf{w}^\top \mathbf{0} = 0.$$

Least Squares

First missing item: invertibility of $\mathbf{X}^\top\mathbf{X}$

Theorem (Invertibility of $\mathbf{X}^\top\mathbf{X}$). Let $\mathbf{X} \in \mathbb{R}^{n\times d}$ be a matrix, with columns $\mathbf{x}_1, \ldots, \mathbf{x}_d \in \mathbb{R}^n$. If $n \geq d$ and $\operatorname{rank}(\mathbf{X}) = d$, then $\mathbf{X}^\top\mathbf{X}$ *is invertible.*

Proof. To show that $\mathbf{X}^\top\mathbf{X}$ is invertible, show $\mathbf{X}^\top\mathbf{X}$ has d linearly independent columns.

$$\mathbf{X}^\top\mathbf{X}\mathbf{w} = \mathbf{0} \implies \mathbf{w} = \mathbf{0}.$$

Suppose $\mathbf{X}^\top\mathbf{X}\mathbf{w} = \mathbf{0}$. Let $\mathbf{w} \in \mathbb{R}^d$ be any vector. Take a dot product of both sides with $\mathbf{w}$:

$$\boxed{(AB)^T = B^T A^T.}$$

$$\|Xw\| = \sqrt{(Xw)^T Xw}$$

$$\|Xw\| = 0 \Downarrow \boxed{Xw = 0}$$

$$\mathbf{w}^\top\mathbf{X}^\top\mathbf{X}\mathbf{w} = \|\mathbf{X}\mathbf{w}\|^2 = 0.$$

$$(w^T X^T) Xw = (\underbrace{Xw}_{\mathbb{R}^n})^T \underbrace{Xw}_{\mathbb{R}^n} = \|Xw\|^2$$

Least Squares

First missing item: invertibility of $\mathbf{X}^\top \mathbf{X}$

Linear Independence:

$w_1 \vec{x}_1 + \dots + w_d \vec{x}_d = \vec{0}$

$\Downarrow$

$w_1, \dots, w_d = 0$

Theorem (Invertibility of $\mathbf{X}^\top \mathbf{X}$). Let $\mathbf{X} \in \mathbb{R}^{n \times d}$ be a matrix, with columns $\mathbf{x}_1, \dots, \mathbf{x}_d \in \mathbb{R}^n$. If $n \geq d$ and $\mathrm{rank}(\mathbf{X}) = d$, then $\mathbf{X}^\top \mathbf{X}$ *is invertible.*

Proof. To show that $\mathbf{X}^\top \mathbf{X}$ is invertible, show $\mathbf{X}^\top \mathbf{X}$ has d linearly independent columns.

$$\mathbf{X}^\top \mathbf{X} \mathbf{w} = \mathbf{0} \implies \underline{\underline{\mathbf{w} = \mathbf{0}}}.$$

Suppose $\mathbf{X}^\top \mathbf{X} \mathbf{w} = \mathbf{0}$. Let $\mathbf{w} \in \mathbb{R}^d$ be any vector. Take a dot product of both sides with $\mathbf{w}$:

$\|v\| = 0 \rightarrow \vec{v} = \vec{0}$

$$\underline{\|\mathbf{X}\mathbf{w}\|^2} \implies \mathbf{X}\mathbf{w} = \mathbf{0}\,.$$

$\Downarrow$

Least Squares

First missing item: invertibility of $\mathbf{X}^\top\mathbf{X}$

$\text{rank}(X) \leq \min\{n, d\}$

Theorem (Invertibility of $\mathbf{X}^\top\mathbf{X}$). Let $\mathbf{X} \in \mathbb{R}^{n\times d}$ be a matrix, with columns $\mathbf{x}_1, \ldots, \mathbf{x}_d \in \mathbb{R}^n$. If $n \geq d$ and $\text{rank}(\mathbf{X}) = d$, then $\mathbf{X}^\top\mathbf{X}$ *is invertible.*

Proof. To show that $\mathbf{X}^\top\mathbf{X}$ is invertible, show $\mathbf{X}^\top\mathbf{X}$ has d linearly independent columns.

$$\mathbf{X}^\top\mathbf{X}\mathbf{w} = \mathbf{0} \implies \mathbf{w} = \mathbf{0}.$$

Suppose $\mathbf{X}^\top\mathbf{X}\mathbf{w} = \mathbf{0}$. Let $\mathbf{w} \in \mathbb{R}^d$ be any vector. Take a dot product of both sides with $\mathbf{w}$:

$$\|\mathbf{X}\mathbf{w}\|^2 \implies \mathbf{X}\mathbf{w} = \mathbf{0}\,.$$

$Xw = \vec{0} \Rightarrow \vec{w} = \vec{0}$

But $\text{rank}(\mathbf{X}) = d$, so $\mathbf{X}$ has d linearly independent columns. Therefore, $\mathbf{w} = \mathbf{0}$.

$$Xw = \begin{bmatrix} | & & | \\ x_1 & \cdots & x_d \\ | & & | \end{bmatrix} \begin{bmatrix} w_1 \\ \vdots \\ w_d \end{bmatrix} = \vec{0}.$$

Least Squares

First missing item: invertibility of $\mathbf{X}^\top \mathbf{X}$

Theorem (Invertibility of $\mathbf{X}^\top \mathbf{X}$). Let $\mathbf{X} \in \mathbb{R}^{n \times d}$ be a matrix, with columns $\mathbf{x}_1, \ldots, \mathbf{x}_d \in \mathbb{R}^n$. If $n \geq d$ and $\mathrm{rank}(\mathbf{X}) = d$, then $\mathbf{X}^\top \mathbf{X}$ *is invertible.*

Least Squares
Summary

Use the principle of *least squares* to find the $\hat{\mathbf{w}} \in \mathbb{R}^d$ that minimizes

$$\|\hat{\mathbf{y}} - \mathbf{y}\|^2 = \|\mathbf{X}\mathbf{w} - \mathbf{y}\|^2.$$

Using geometric intuition: $\hat{\mathbf{y}}$ is the vector for which $\hat{\mathbf{y}} - \mathbf{y}$ is perpendicular to $\operatorname{span}(\operatorname{col}(\mathbf{X}))$.

By Pythagorean Theorem, any other vector $\tilde{\mathbf{y}} \in \operatorname{span}(\operatorname{col}(\mathbf{X}))$ gives a larger error:

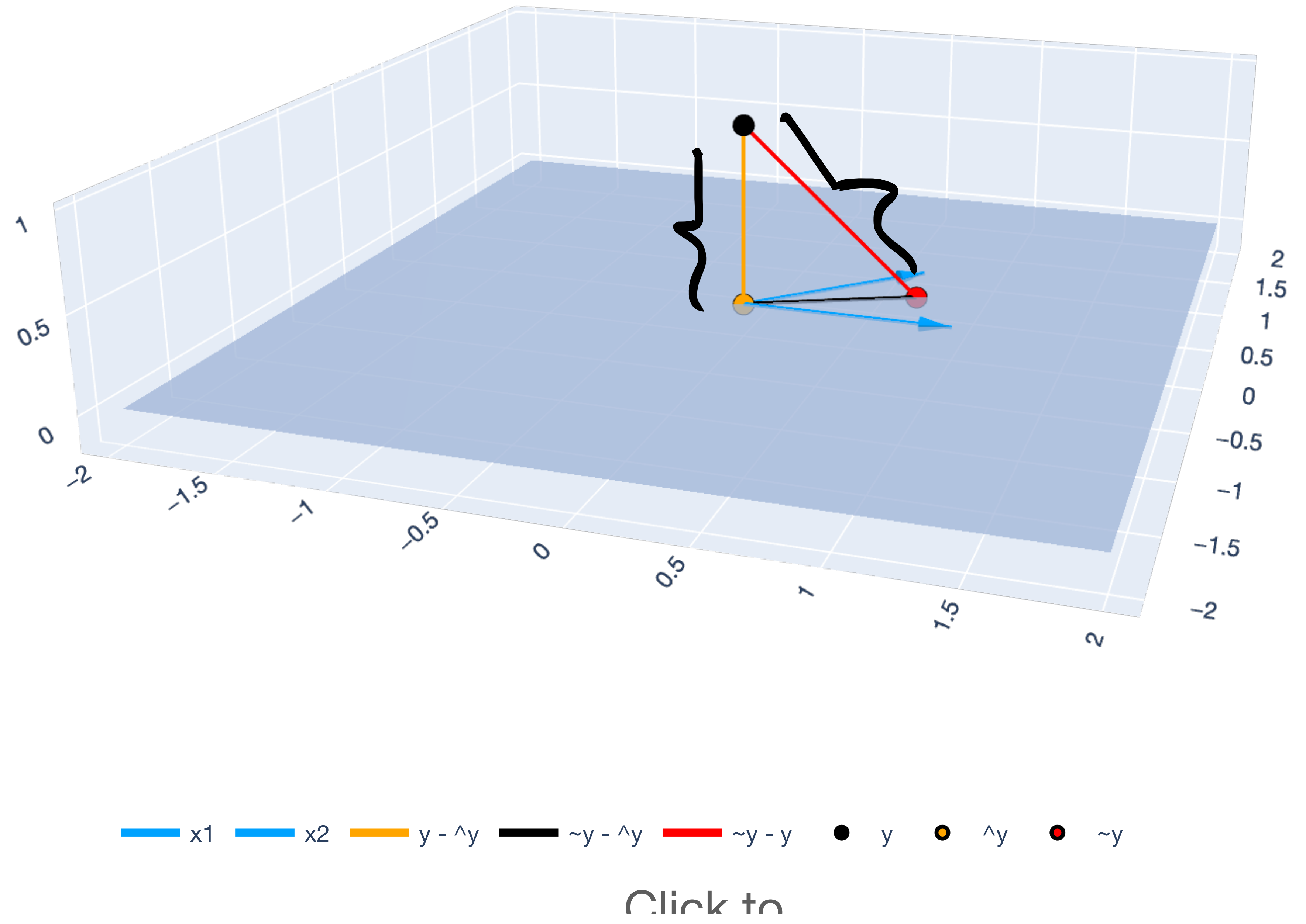
$$\|\underline{\hat{\mathbf{y}} - \mathbf{y}}\|^2 \leq \|\underline{\tilde{\mathbf{y}} - \mathbf{y}}\|^2.$$

Because $\hat{\mathbf{y}} - \mathbf{y}$ is perpendicular, we obtain the *normal equations:*

$$\boxed{\mathbf{X}^\top \mathbf{X} \hat{\mathbf{w}} = \mathbf{X}^\top \mathbf{y}.}$$

If $n \geq d$ and $\operatorname{rank}(\mathbf{X}) = d$, then $\mathbf{X}^\top\mathbf{X}$ *is invertible*, and

$$\boxed{\hat{\mathbf{w}} = (\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{X}^\top \mathbf{y}.}$$



Least Squares

Second missing item: Pythagorean Theorem

By Pythagorean Theorem, any other vector $\tilde{\mathbf{y}} \in \mathrm{span}(\mathrm{col}(\mathbf{X}))$ gives a larger error:

$$\boxed{\|\hat{\mathbf{y}} - \mathbf{y}\|^2 \leq \|\tilde{\mathbf{y}} - \mathbf{y}\|^2.}$$

"The vector closest to $\mathbf{y}$ in the subspace is perpendicular."

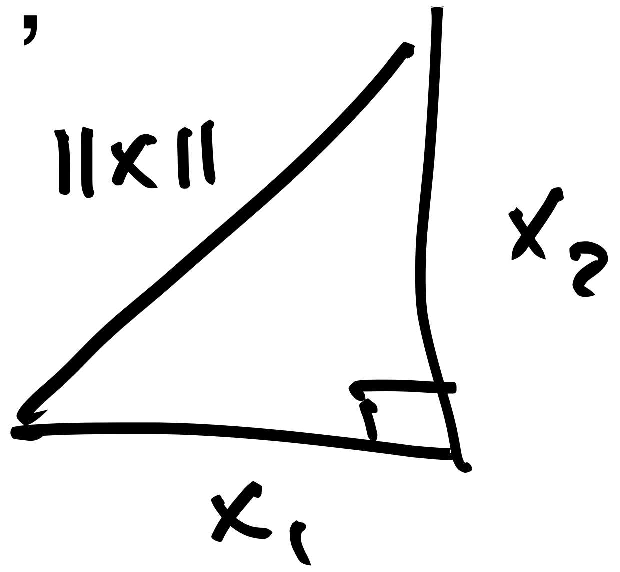
Orthogonality

Definition and Orthonormal Bases

Norms and Inner Products

Euclidean Norm

Recall the notion of “length” from $\mathbb{R}^2$. For a vector $\mathbf{x} = (x_1, x_2) \in \mathbb{R}^2$,



$$\|\mathbf{x}\|_2 := \sqrt{x_1^2 + x_2^2}.$$

Generalizing this, for $\mathbf{x} \in \mathbb{R}^n$, the ***Euclidean norm (ℓ_2-norm)*** is:

$$\|\mathbf{x}\|_2 := \sqrt{x_1^2 + \ldots + x_n^2} = \sqrt{\mathbf{x}^\top \mathbf{x}}.$$

$$\|\mathbf{x}\|_2^2 = \mathbf{x}^\top \mathbf{x}.$$

In this course, dropping the “2” and just writing $\|\mathbf{x}\|$ denotes the Euclidean norm.

Orthogonality

Definition

$$v_1 w_1 + \dots + v_d w_d$$

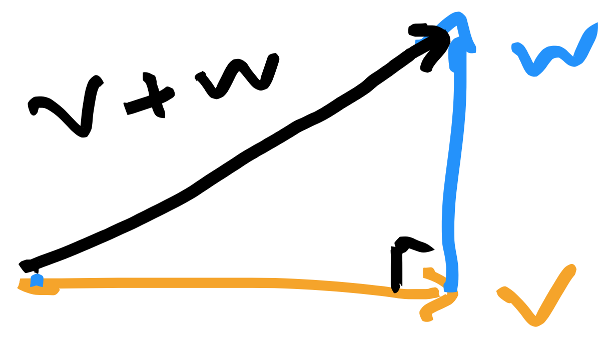
Two vectors $\mathbf{v}, \mathbf{w} \in \mathbb{R}^n$ are ***orthogonal*** if $\langle \mathbf{v}, \mathbf{w} \rangle = \mathbf{v}^\top \mathbf{w} = 0$. In $\mathbb{R}^2$ and $\mathbb{R}^3$, this corresponds to our geometric notion of "perpendicular."

A set of vectors is ***orthogonal*** if every pair of distinct vectors in the set is orthogonal.

$v_1 \quad v_2 \quad v_3$

Orthogonality

Pythagorean Theorem



<u>Theorem (Pythagorean Theorem).</u> If vectors $\mathbf{v}, \mathbf{w} \in \mathbb{R}^n$ are orthogonal, then

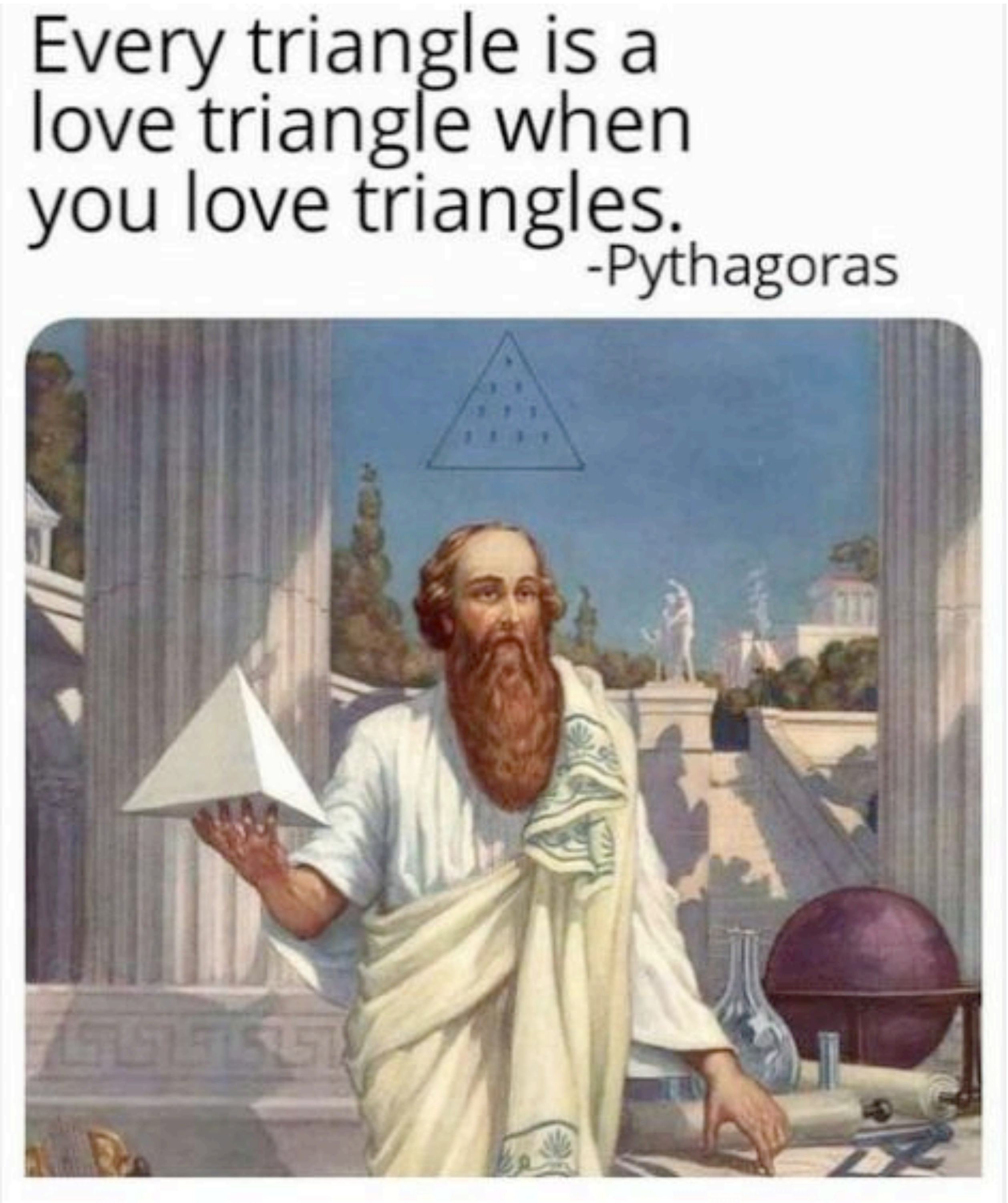
$$\|\mathbf{v} + \mathbf{w}\|^2 = \|\mathbf{v}\|^2 + \|\mathbf{w}\|^2.$$

Orthogonality

Pythagorean Theorem

Theorem (Pythagorean Theorem). If vectors $\mathbf{v}, \mathbf{w} \in \mathbb{R}^n$ are orthogonal, then

$$\|\mathbf{v} + \mathbf{w}\|^2 = \|\mathbf{v}\|^2 + \|\mathbf{w}\|^2.$$



Orthogonality

Pythagorean Theorem

Theorem (Pythagorean Theorem). If vectors $\mathbf{v}, \mathbf{w} \in \mathbb{R}^n$ are orthogonal, then

$$||\mathbf{v} + \mathbf{w}||^2 = ||\mathbf{v}||^2 + ||\mathbf{w}||^2.$$

Proof. Let $\mathbf{v}, \mathbf{w} \in \mathbb{R}^n$ be orthogonal vectors. Expand the square $||\mathbf{v} + \mathbf{w}||^2$.

Orthogonality

Pythagorean Theorem

Theorem (Pythagorean Theorem). If vectors $\mathbf{v}, \mathbf{w} \in \mathbb{R}^n$ are orthogonal, then

$$\|\mathbf{v} + \mathbf{w}\|^2 = \|\mathbf{v}\|^2 + \|\mathbf{w}\|^2.$$

inner product of v+w w/ itself.

Proof. Let $\mathbf{v}, \mathbf{w} \in \mathbb{R}^n$ be orthogonal vectors. Expand the square $\|\mathbf{v} + \mathbf{w}\|^2$.

$$\|\mathbf{v} + \mathbf{w}\|^2 = \langle \mathbf{v} + \mathbf{w}, \mathbf{v} + \mathbf{w} \rangle$$

Orthogonality

Pythagorean Theorem

Theorem (Pythagorean Theorem). If vectors $\mathbf{v}, \mathbf{w} \in \mathbb{R}^n$ are orthogonal, then

$$\|\mathbf{v} + \mathbf{w}\|^2 = \|\mathbf{v}\|^2 + \|\mathbf{w}\|^2.$$

Proof. Let $\mathbf{v}, \mathbf{w} \in \mathbb{R}^n$ be orthogonal vectors. Expand the square $\|\mathbf{v} + \mathbf{w}\|^2$.

$(a+b)(a+b)$

$$\begin{aligned} \|\mathbf{v} + \mathbf{w}\|^2 &= \langle \mathbf{v} + \mathbf{w}, \mathbf{v} + \mathbf{w} \rangle \\ &= \langle \mathbf{v}, \mathbf{v} \rangle + \langle \mathbf{v}, \mathbf{w} \rangle + \langle \mathbf{w}, \mathbf{v} \rangle + \langle \mathbf{w}, \mathbf{w} \rangle \end{aligned}$$

Linearity.

$\langle a+b, c \rangle$
$= \langle a, c \rangle + \langle b, c \rangle$.

Orthogonality

Pythagorean Theorem

<u>**Theorem (Pythagorean Theorem).**</u> If vectors $\mathbf{v}, \mathbf{w} \in \mathbb{R}^n$ are orthogonal, then

$$\|\mathbf{v} + \mathbf{w}\|^2 = \|\mathbf{v}\|^2 + \|\mathbf{w}\|^2.$$

Proof. Let $\mathbf{v}, \mathbf{w} \in \mathbb{R}^n$ be orthogonal vectors. Expand the square $\|\mathbf{v} + \mathbf{w}\|^2$.

$$\begin{aligned}
\|\mathbf{v} + \mathbf{w}\|^2 &= \langle \mathbf{v} + \mathbf{w}, \mathbf{v} + \mathbf{w} \rangle \\
&= \langle \mathbf{v}, \mathbf{v} \rangle + \langle \underline{\mathbf{v}}, \underline{\mathbf{w}} \rangle + \langle \underline{\mathbf{w}}, \underline{\mathbf{v}} \rangle + \langle \mathbf{w}, \mathbf{w} \rangle \qquad \text{linearity} \\
&= \langle \mathbf{v}, \mathbf{v} \rangle + \underline{2\langle \mathbf{v}, \mathbf{w} \rangle} + \langle \mathbf{w}, \mathbf{w} \rangle \qquad \text{symmetry}
\end{aligned}$$

Orthogonality

Pythagorean Theorem

$\langle v, w \rangle = 0$

Theorem (Pythagorean Theorem). If vectors $\mathbf{v}, \mathbf{w} \in \mathbb{R}^n$ are orthogonal, then

$$\|\mathbf{v} + \mathbf{w}\|^2 = \|\mathbf{v}\|^2 + \|\mathbf{w}\|^2.$$

Proof. Let $\mathbf{v}, \mathbf{w} \in \mathbb{R}^n$ be orthogonal vectors. Expand the square $\|\mathbf{v} + \mathbf{w}\|^2$.

$$\begin{aligned}
\|\mathbf{v} + \mathbf{w}\|^2 &= \langle \mathbf{v} + \mathbf{w}, \mathbf{v} + \mathbf{w} \rangle \\
&= \langle \mathbf{v}, \mathbf{v} \rangle + \langle \mathbf{v}, \mathbf{w} \rangle + \langle \mathbf{w}, \mathbf{v} \rangle + \langle \mathbf{w}, \mathbf{w} \rangle \\
&= \langle \mathbf{v}, \mathbf{v} \rangle + 2\langle \mathbf{v}, \mathbf{w} \rangle + \langle \mathbf{w}, \mathbf{w} \rangle \\
&= \|\mathbf{v}\|^2 + \|\mathbf{w}\|^2
\end{aligned}$$

∎

Orthogonality

Pythagorean Theorem

Theorem (Pythagorean Theorem). If vectors $\mathbf{v}, \mathbf{w} \in \mathbb{R}^n$ are orthogonal, then

$$\|\mathbf{v} + \mathbf{w}\|^2 = \|\mathbf{v}\|^2 + \|\mathbf{w}\|^2.$$

Proof. Let $\mathbf{v}, \mathbf{w} \in \mathbb{R}^n$ be orthogonal vectors. Expand the square $\|\mathbf{v} + \mathbf{w}\|^2$.

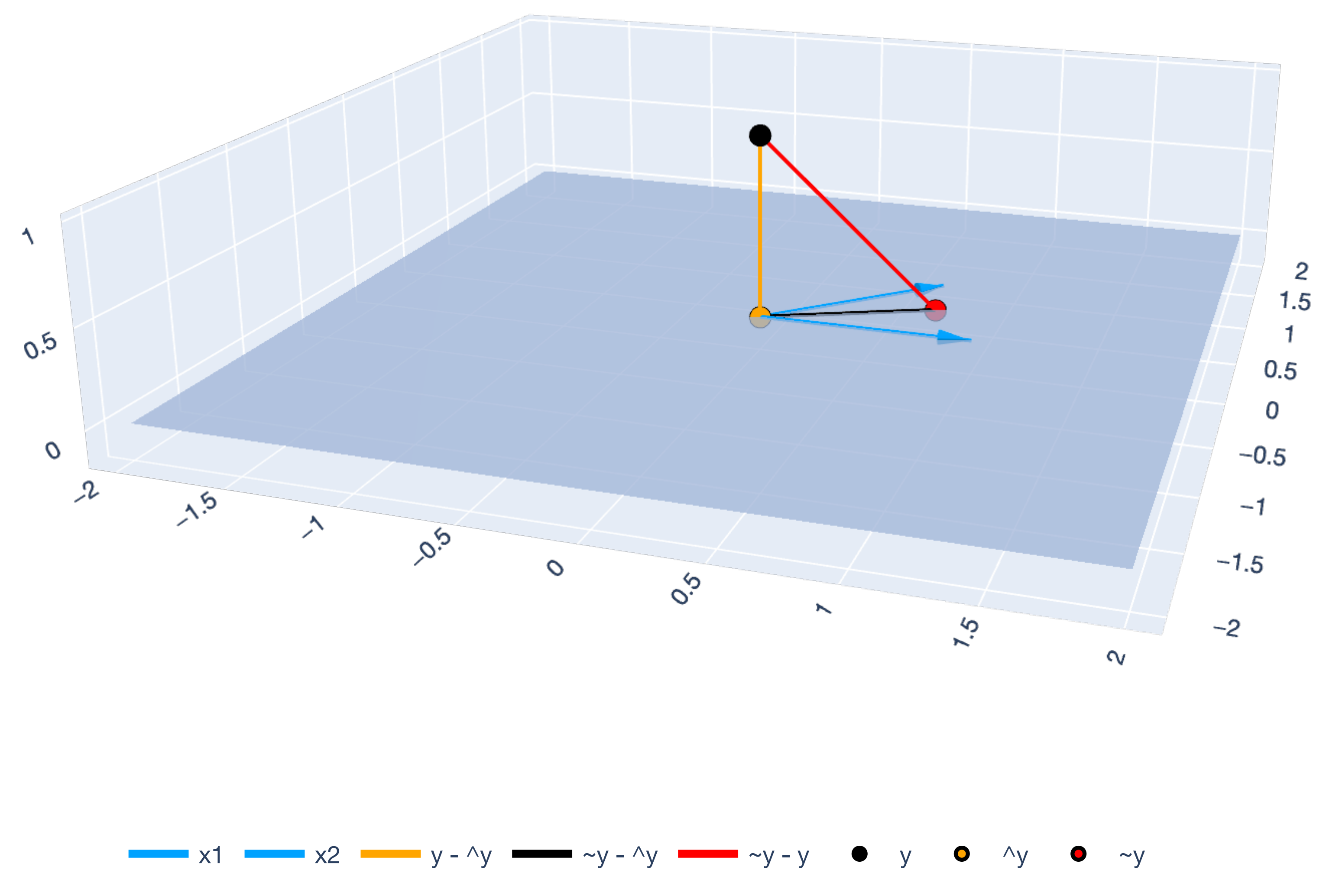
$$\begin{aligned}
\|\mathbf{v} + \mathbf{w}\|^2 &= (\mathbf{v} + \mathbf{w})^\top(\mathbf{v} + \mathbf{w}) \\
&= \mathbf{v}^\top\mathbf{v} + \mathbf{v}^\top\mathbf{w} + \mathbf{w}^\top\mathbf{v} + \mathbf{w}^\top\mathbf{w} \\
&= \mathbf{v}^\top\mathbf{v} + 2\mathbf{v}^\top\mathbf{w} + \mathbf{w}^\top\mathbf{w} \\
&= \|\mathbf{v}\|^2 + \|\mathbf{w}\|^2
\end{aligned}$$

Least Squares

Second missing item: Pythagorean Theorem

By Pythagorean Theorem, any other vector $\tilde{\mathbf{y}} \in \mathrm{span}(\mathrm{col}(\mathbf{X}))$ gives a larger error:

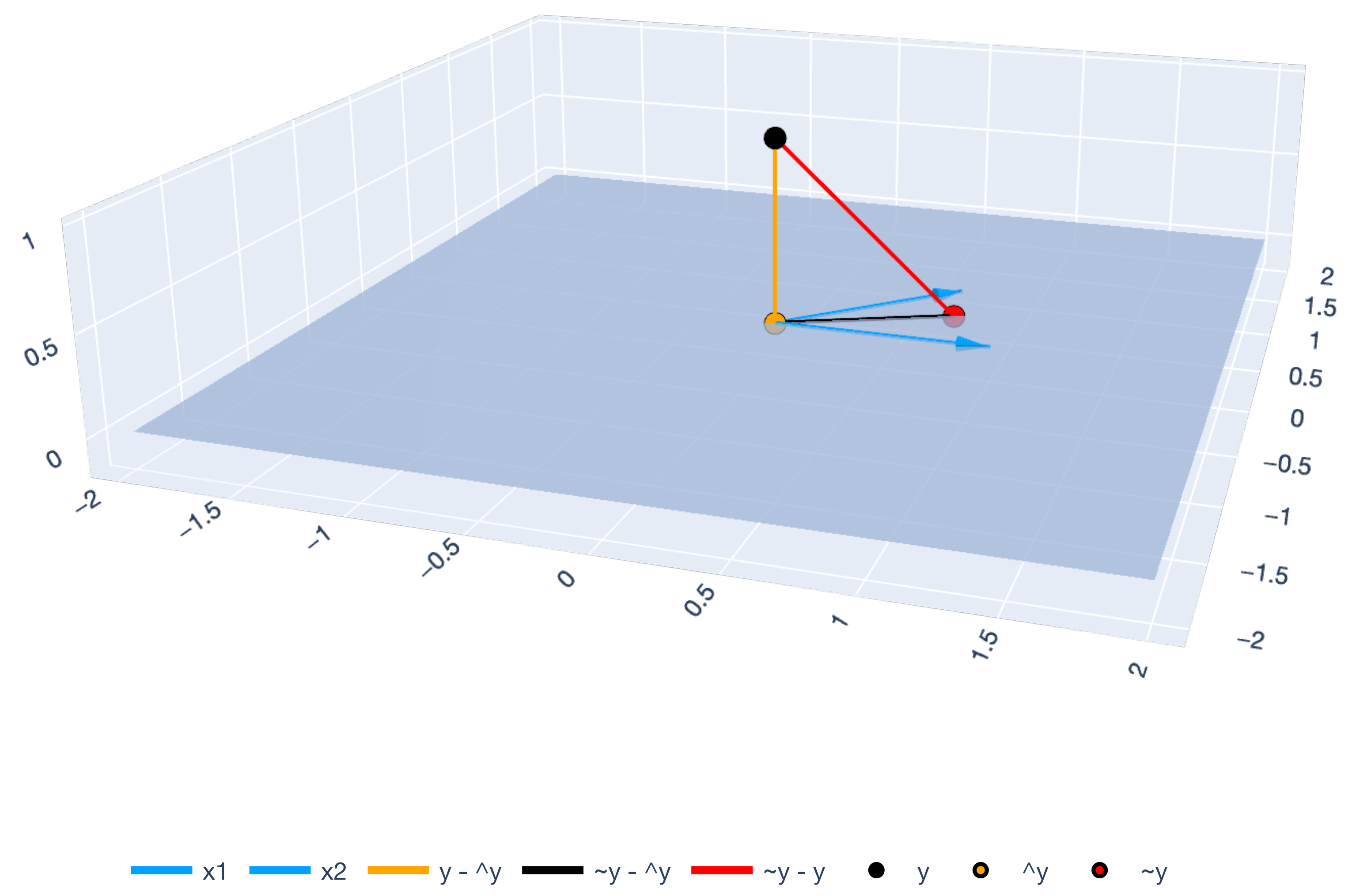
$$\boxed{\|\hat{\mathbf{y}} - \mathbf{y}\|^2 \leq \|\tilde{\mathbf{y}} - \mathbf{y}\|^2.}$$



Least Squares

Second missing item: Pythagorean Theorem

Theorem (Projection minimizes distance). Let $\boxed{\hat{\mathbf{y}}} \in \mathrm{span}(\mathrm{col}(\mathbf{X}))$ be the vector where $\hat{\mathbf{y}} - \mathbf{y}$ is orthogonal to any vector in $\mathrm{span}(\mathrm{col}(\mathbf{X}))$ and let $\boxed{\tilde{\mathbf{y}}} \in \mathrm{span}(\mathrm{col}(\mathbf{X}))$ be any other vector. Then $\|\hat{\mathbf{y}} - \mathbf{y}\|^2 \leq \|\tilde{\mathbf{y}} - \mathbf{y}\|^2$.



Click to

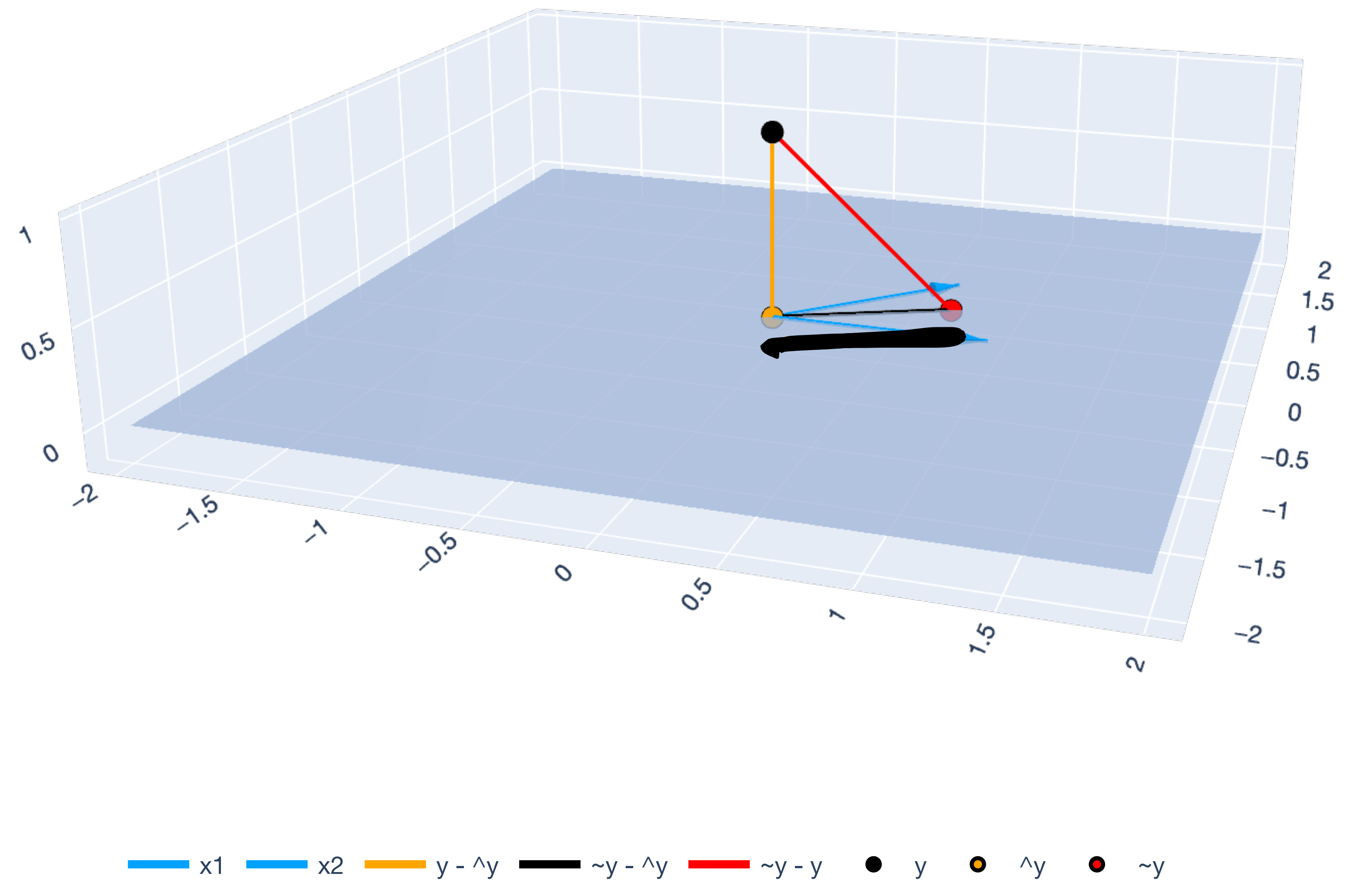
Least Squares

Second missing item: Pythagorean Theorem

Theorem (Projection minimizes distance). Let $\hat{\mathbf{y}} \in \text{span}(\text{col}(\mathbf{X}))$ be the vector where $\hat{\mathbf{y}} - \mathbf{y}$ is orthogonal to any vector in $\text{span}(\text{col}(\mathbf{X}))$ and let $\tilde{\mathbf{y}} \in \text{span}(\text{col}(\mathbf{X}))$ be any other vector. Then $||\hat{\mathbf{y}} - \mathbf{y}||^2 \leq ||\tilde{\mathbf{y}} - \mathbf{y}||^2$.

Proof. Because $\hat{\mathbf{y}} \in \text{span}(\text{col}(\mathbf{X}))$ and $\tilde{\mathbf{y}} \in \text{span}(\text{col}(\mathbf{X}))$ and $\text{span}(\text{col}(\mathbf{X}))$ is a subspace, $\tilde{\mathbf{y}} - \hat{\mathbf{y}} \in \text{span}(\text{col}(\mathbf{X}))$.

$\tilde{y} - \hat{y}$



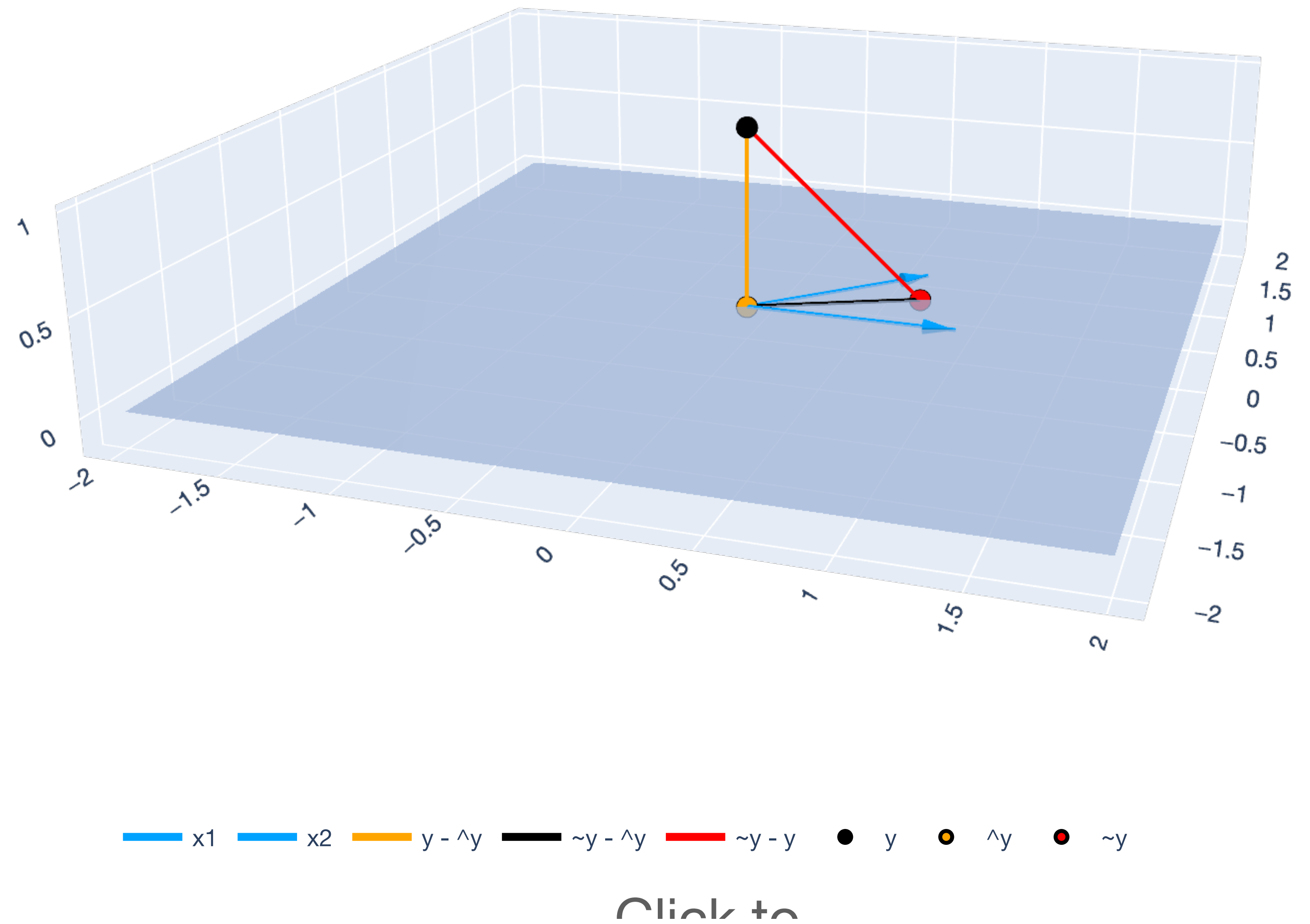
Least Squares

Second missing item: Pythagorean Theorem

Theorem (Projection minimizes distance). Let $\hat{\mathbf{y}} \in \text{span}(\text{col}(\mathbf{X}))$ be the vector where $\hat{\mathbf{y}} - \mathbf{y}$ is orthogonal to any vector in $\text{span}(\text{col}(\mathbf{X}))$ and let $\tilde{\mathbf{y}} \in \text{span}(\text{col}(\mathbf{X}))$ be any other vector. Then $\|\hat{\mathbf{y}} - \mathbf{y}\|^2 \leq \|\tilde{\mathbf{y}} - \mathbf{y}\|^2$.

Proof. Because $\hat{\mathbf{y}} \in \text{span}(\text{col}(\mathbf{X}))$ and $\tilde{\mathbf{y}} \in \text{span}(\text{col}(\mathbf{X}))$ and $\text{span}(\text{col}(\mathbf{X}))$ is a subspace, $\tilde{\mathbf{y}} - \hat{\mathbf{y}} \in \text{span}(\text{col}(\mathbf{X}))$.

The vector $\hat{\mathbf{y}} - \mathbf{y}$ is orthogonal to any vector in $\text{span}(\text{col}(\mathbf{X}))$, so $\hat{\mathbf{y}} - \mathbf{y}$ is orthogonal to $\tilde{\mathbf{y}} - \hat{\mathbf{y}}$.



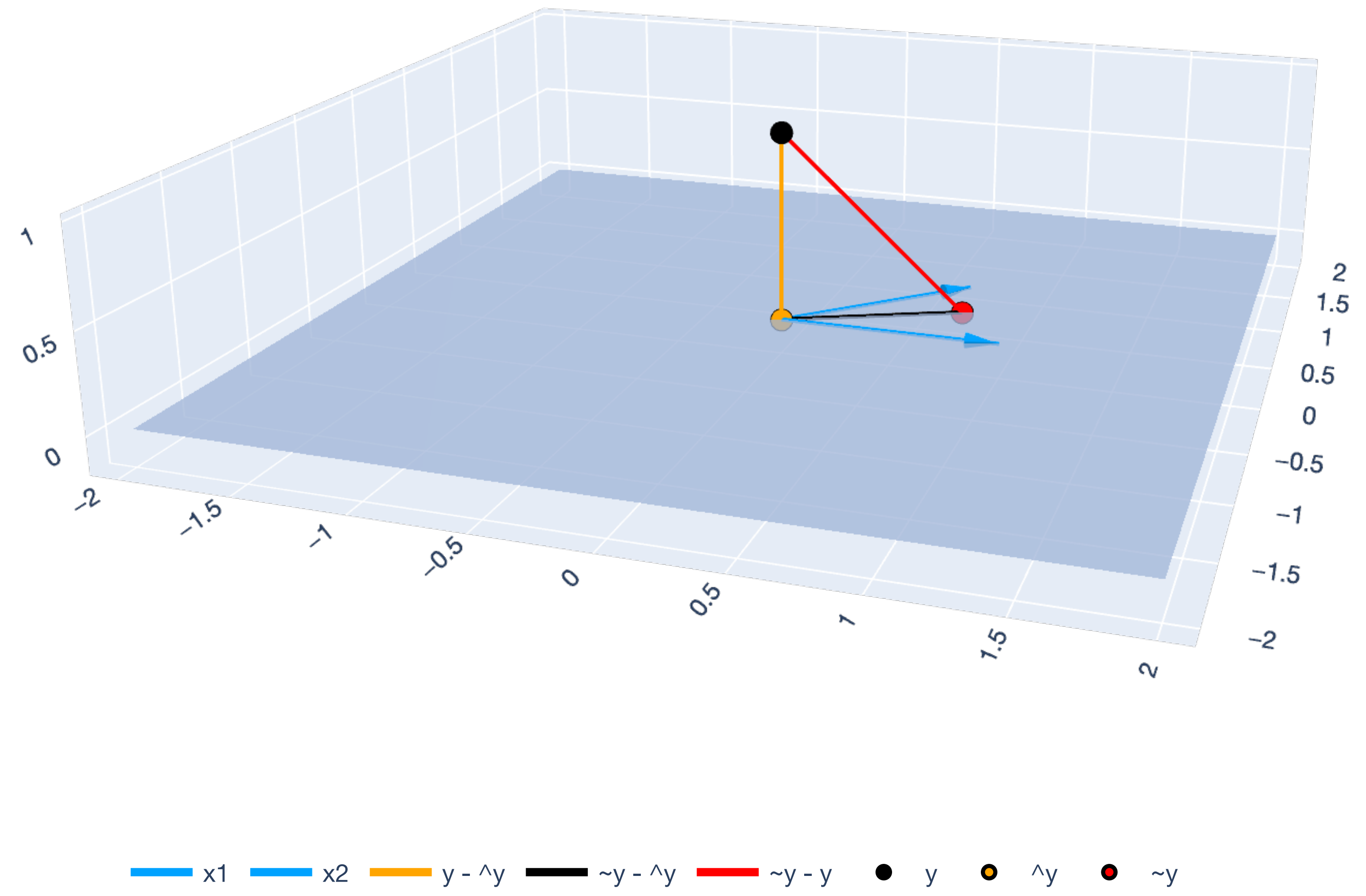
Least Squares

Second missing item: Pythagorean Theorem

Theorem (Projection minimizes distance). Let $\hat{\mathbf{y}} \in \mathrm{span}(\mathrm{col}(\mathbf{X}))$ be the vector where $\hat{\mathbf{y}} - \mathbf{y}$ is orthogonal to any vector in $\mathrm{span}(\mathrm{col}(\mathbf{X}))$ and let$\tilde{\mathbf{y}} \in \mathrm{span}(\mathrm{col}(\mathbf{X}))$ be any other vector. Then $\|\hat{\mathbf{y}} - \mathbf{y}\|^2 \leq \|\tilde{\mathbf{y}} - \mathbf{y}\|^2$.

Proof. Because $\hat{\mathbf{y}} \in \mathrm{span}(\mathrm{col}(\mathbf{X}))$ and $\tilde{\mathbf{y}} \in \mathrm{span}(\mathrm{col}(\mathbf{X}))$ and $\mathrm{span}(\mathrm{col}(\mathbf{X}))$ is a subspace, $\tilde{\mathbf{y}} - \hat{\mathbf{y}} \in \mathrm{span}(\mathrm{col}(\mathbf{X}))$.

The vector $\hat{\mathbf{y}} - \mathbf{y}$ is orthogonal to any vector in $\mathrm{span}(\mathrm{col}(\mathbf{X}))$, so $\hat{\mathbf{y}} - \mathbf{y}$ is orthogonal to $\tilde{\mathbf{y}} - \hat{\mathbf{y}}$.



By the Pythagorean Theorem: $\|u\|^2 + \|v\|^2 = \|u+v\|^2$

$$\|\hat{\mathbf{y}} - \mathbf{y}\|^2 + \|\tilde{\mathbf{y}} - \hat{\mathbf{y}}\|^2 = \|\underline{\hat{\mathbf{y}} - \mathbf{y} + \tilde{\mathbf{y}} - \hat{\mathbf{y}}}\|^2$$

$u = \hat{y} - y$

$v = \tilde{y} - \hat{y}$

Least Squares

Second missing item: Pythagorean Theorem

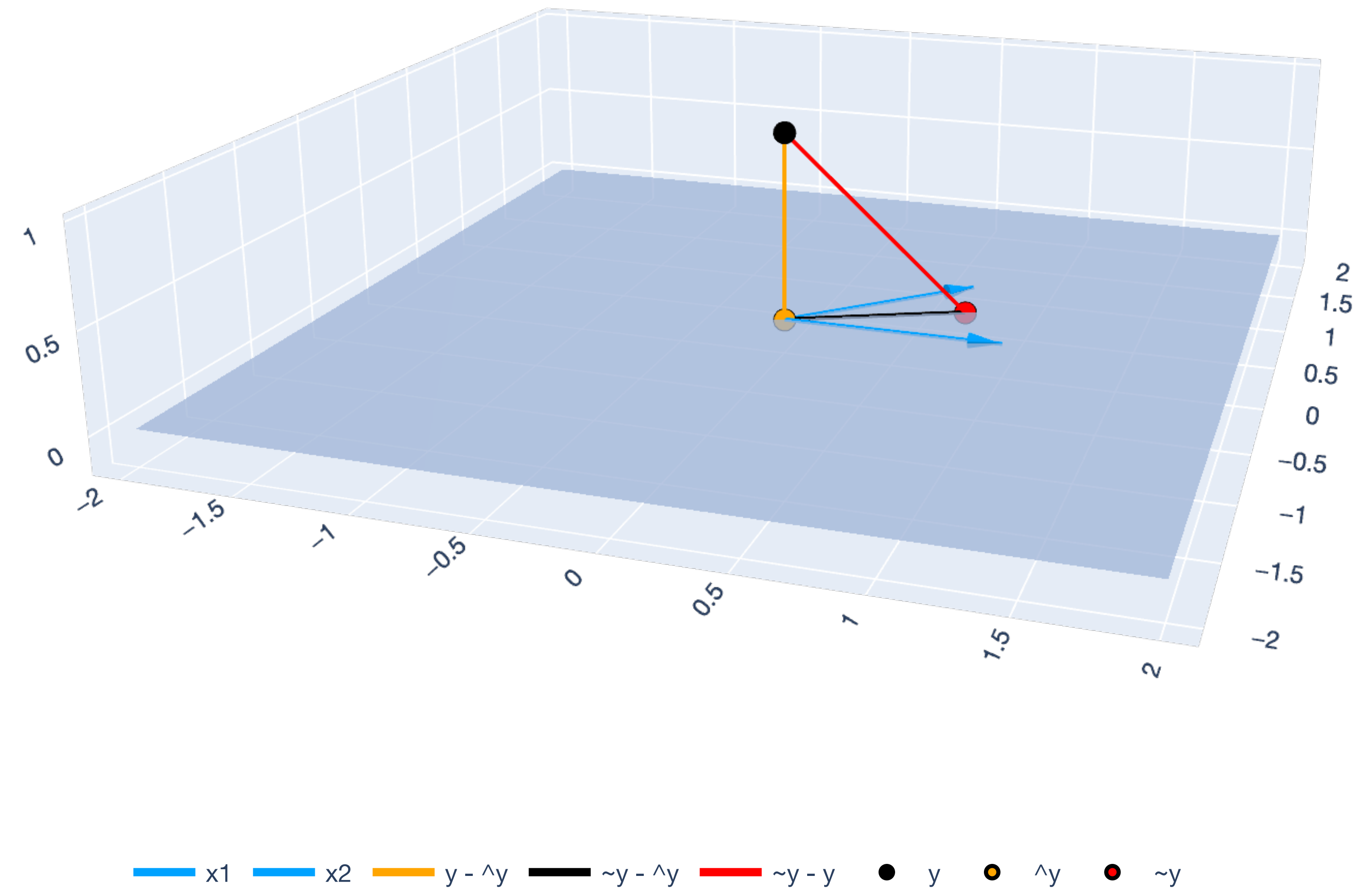
Theorem (Projection minimizes distance). Let $\hat{\mathbf{y}} \in \mathrm{span}(\mathrm{col}(\mathbf{X}))$ be the vector where $\hat{\mathbf{y}} - \mathbf{y}$ is orthogonal to any vector in $\mathrm{span}(\mathrm{col}(\mathbf{X}))$ and let $\tilde{\mathbf{y}} \in \mathrm{span}(\mathrm{col}(\mathbf{X}))$ be any other vector. Then $\|\hat{\mathbf{y}} - \mathbf{y}\|^2 \leq \|\tilde{\mathbf{y}} - \mathbf{y}\|^2$.

Proof. Because $\hat{\mathbf{y}} \in \mathrm{span}(\mathrm{col}(\mathbf{X}))$ and $\tilde{\mathbf{y}} \in \mathrm{span}(\mathrm{col}(\mathbf{X}))$ and $\mathrm{span}(\mathrm{col}(\mathbf{X}))$ is a subspace, $\tilde{\mathbf{y}} - \hat{\mathbf{y}} \in \mathrm{span}(\mathrm{col}(\mathbf{X}))$.

The vector $\hat{\mathbf{y}} - \mathbf{y}$ is orthogonal to any vector in $\mathrm{span}(\mathrm{col}(\mathbf{X}))$, so $\hat{\mathbf{y}} - \mathbf{y}$ is orthogonal to $\tilde{\mathbf{y}} - \hat{\mathbf{y}}$.

By the Pythagorean Theorem:

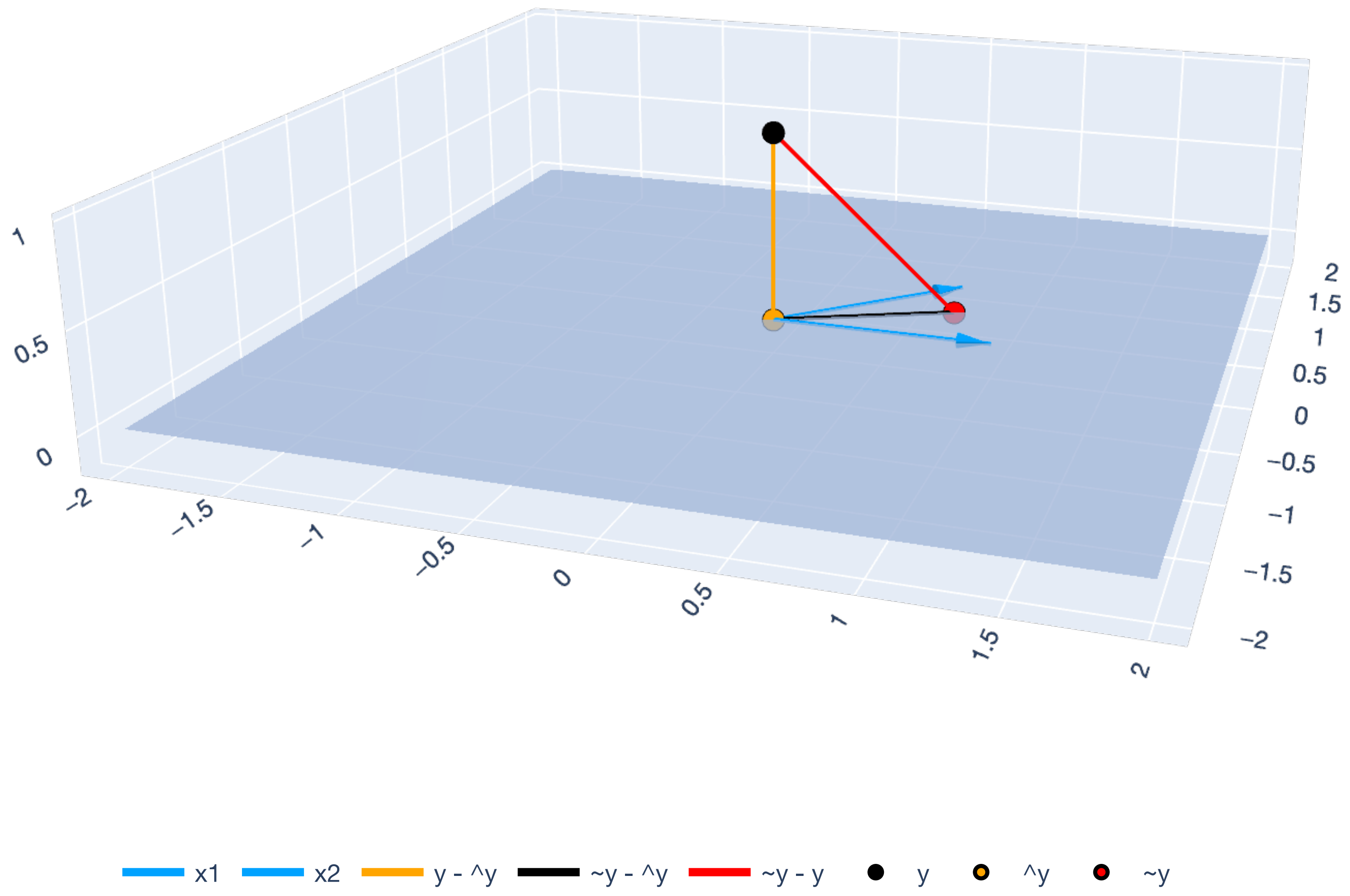
$$\|\hat{\mathbf{y}} - \mathbf{y}\|^2 + \|\tilde{\mathbf{y}} - \hat{\mathbf{y}}\|^2 = \|\hat{\mathbf{y}} - \mathbf{y} + \tilde{\mathbf{y}} - \hat{\mathbf{y}}\|^2 = \|\tilde{\mathbf{y}} - \mathbf{y}\|^2$$



Least Squares

Second missing item: Pythagorean Theorem

Theorem (Projection minimizes distance). Let $\hat{\mathbf{y}} \in \mathrm{span}(\mathrm{col}(\mathbf{X}))$ be the vector where $\hat{\mathbf{y}} - \mathbf{y}$ is orthogonal to any vector in $\mathrm{span}(\mathrm{col}(\mathbf{X}))$ and let $\tilde{\mathbf{y}} \in \mathrm{span}(\mathrm{col}(\mathbf{X}))$ be any other vector. Then $\|\hat{\mathbf{y}} - \mathbf{y}\|^2 \leq \|\tilde{\mathbf{y}} - \mathbf{y}\|^2$.



Proof. Because $\hat{\mathbf{y}} \in \mathrm{span}(\mathrm{col}(\mathbf{X}))$ and $\tilde{\mathbf{y}} \in \mathrm{span}(\mathrm{col}(\mathbf{X}))$ and $\mathrm{span}(\mathrm{col}(\mathbf{X}))$ is a subspace, $\tilde{\mathbf{y}} - \hat{\mathbf{y}} \in \mathrm{span}(\mathrm{col}(\mathbf{X}))$.

The vector $\hat{\mathbf{y}} - \mathbf{y}$ is orthogonal to any vector in $\mathrm{span}(\mathrm{col}(\mathbf{X}))$, so $\hat{\mathbf{y}} - \mathbf{y}$ is orthogonal to $\tilde{\mathbf{y}} - \hat{\mathbf{y}}$.

By the Pythagorean Theorem:

$$\|\hat{\mathbf{y}} - \mathbf{y}\|^2 + \|\tilde{\mathbf{y}} - \hat{\mathbf{y}}\|^2 = \|\hat{\mathbf{y}} - \mathbf{y} + \tilde{\mathbf{y}} - \hat{\mathbf{y}}\|^2 = \|\tilde{\mathbf{y}} - \mathbf{y}\|^2$$

But because norms are always nonnegative,

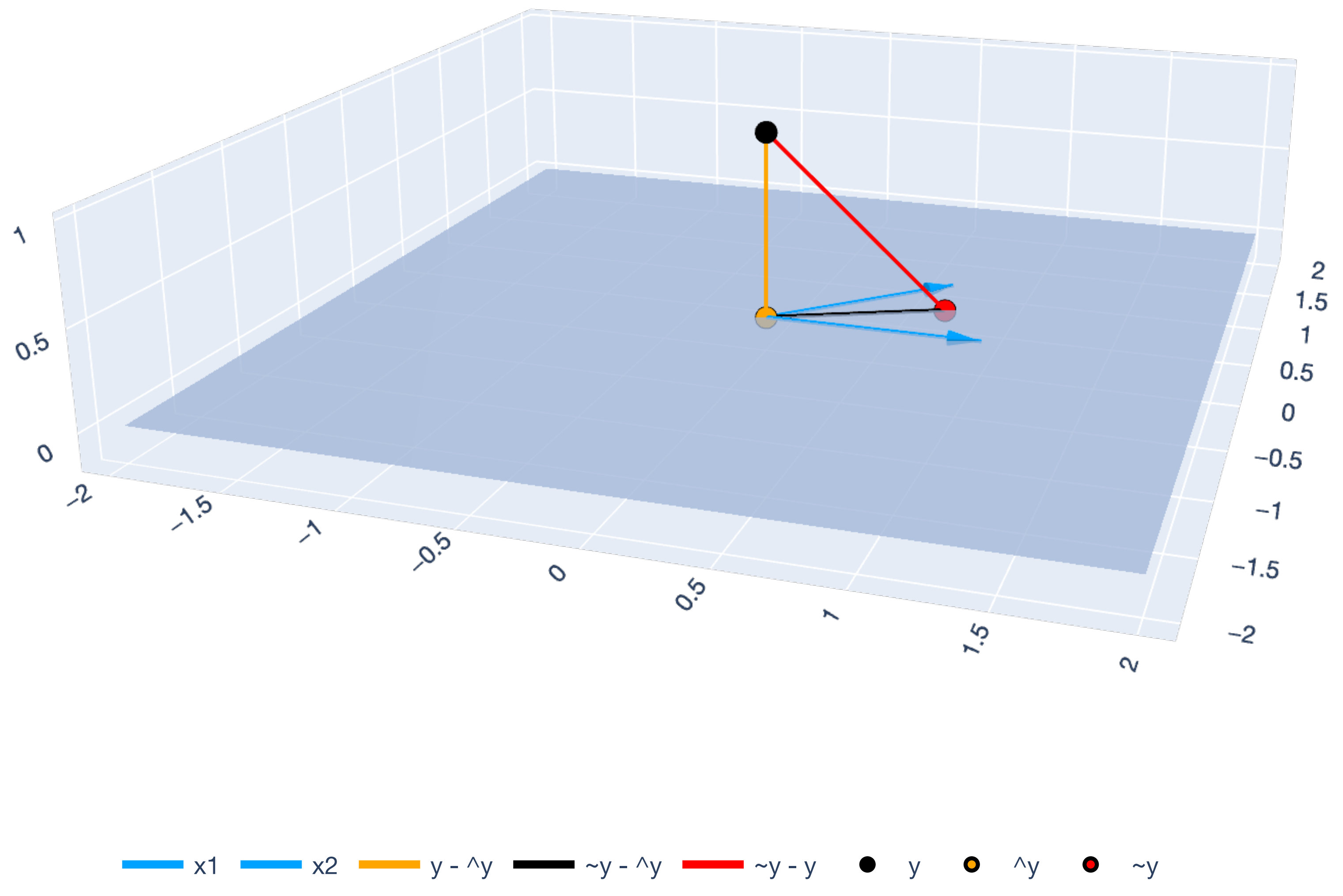
$$\|\hat{y} - y\|^2 + \underbrace{\|\tilde{y} - \hat{y}\|^2}_{\geq 0} = \|\tilde{y} - y\|^2$$

$$\|\hat{\mathbf{y}} - \mathbf{y}\|^2 \leq \|\tilde{\mathbf{y}} - \mathbf{y}\|^2.$$

Least Squares

Second missing item: Pythagorean Theorem

Theorem (Projection minimizes distance). Let $\hat{\mathbf{y}} \in \mathrm{span}(\mathrm{col}(\mathbf{X}))$ be the vector where $\hat{\mathbf{y}} - \mathbf{y}$ is orthogonal to any vector in $\mathrm{span}(\mathrm{col}(\mathbf{X}))$ and let $\tilde{\mathbf{y}} \in \mathrm{span}(\mathrm{col}(\mathbf{X}))$ be any other vector. Then $||\hat{\mathbf{y}} - \mathbf{y}||^2 \leq ||\tilde{\mathbf{y}} - \mathbf{y}||^2$.



Click to

Least Squares

Summary

Use the principle of *least squares* to find the $\hat{\mathbf{w}} \in \mathbb{R}^d$ that minimizes

$$\|\hat{\mathbf{y}} - \mathbf{y}\|^2 = \|\mathbf{X}\mathbf{w} - \mathbf{y}\|^2.$$

Using geometric intuition: $\hat{\mathbf{y}}$ is the vector for which $\hat{\mathbf{y}} - \mathbf{y}$ is perpendicular to $\mathrm{span}(\mathrm{col}(\mathbf{X}))$.

By Pythagorean Theorem, any other vector $\tilde{\mathbf{y}} \in \mathrm{span}(\mathrm{col}(\mathbf{X}))$ gives a larger error:

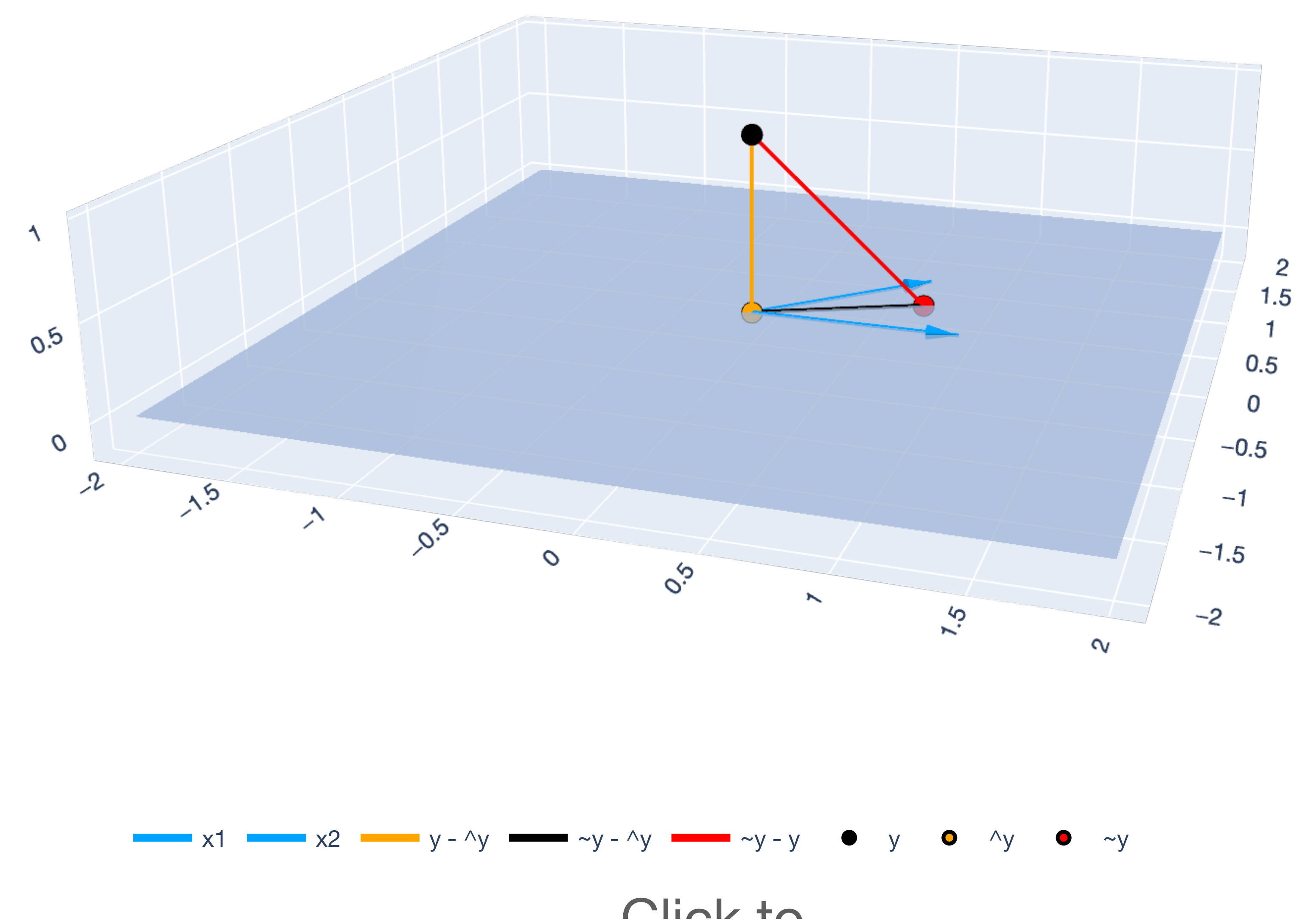
$$\|\hat{\mathbf{y}} - \mathbf{y}\|^2 \leq \|\tilde{\mathbf{y}} - \mathbf{y}\|^2.$$

Because $\hat{\mathbf{y}} - \mathbf{y}$ is perpendicular, we obtain the *normal equations:*

$$\mathbf{X}^\top \mathbf{X} \hat{\mathbf{w}} = \mathbf{X}^\top \mathbf{y}.$$

If $n \geq d$ and $\mathrm{rank}(\mathbf{X}) = d$, then $\mathbf{X}^\top \mathbf{X}$ *is invertible*, and

$$\boxed{\hat{\mathbf{w}} = (\mathbf{X}^\top \mathbf{X})^{-1} \mathbf{X}^\top \mathbf{y}.}$$



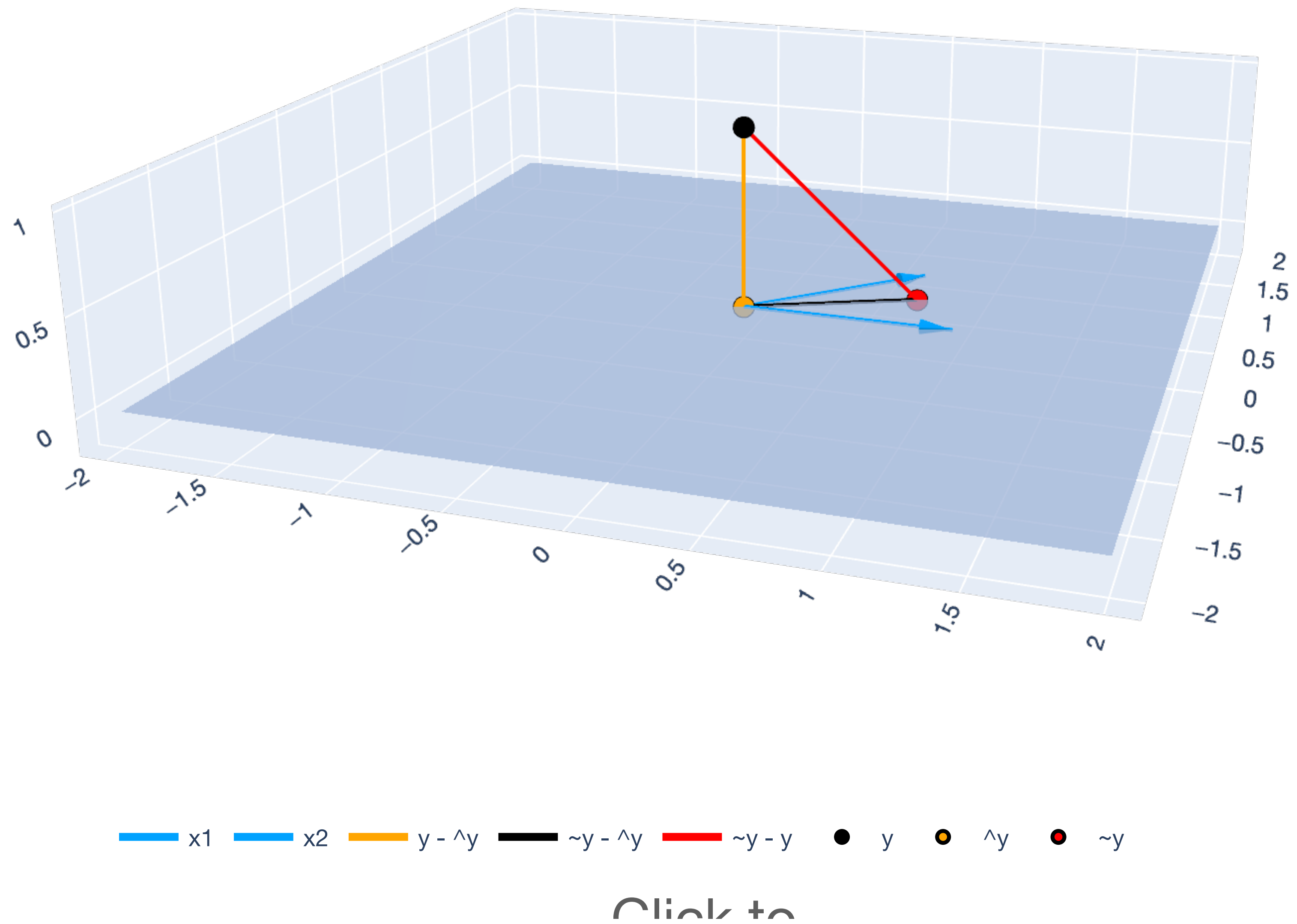
Least Squares
Summary

Goal: Find the $\hat{\mathbf{w}} \in \mathbb{R}^d$ that minimizes

$$||\mathbf{X}\mathbf{w} - \mathbf{y}||^2.$$

<u>**Theorem (OLS).**</u> If $n \geq d$ and $\operatorname{rank}(\mathbf{X}) = d$, then:

$$\hat{\mathbf{w}} = (\mathbf{X}^\top \mathbf{X})^{-1} \mathbf{X}^\top \mathbf{y}.$$



Least Squares

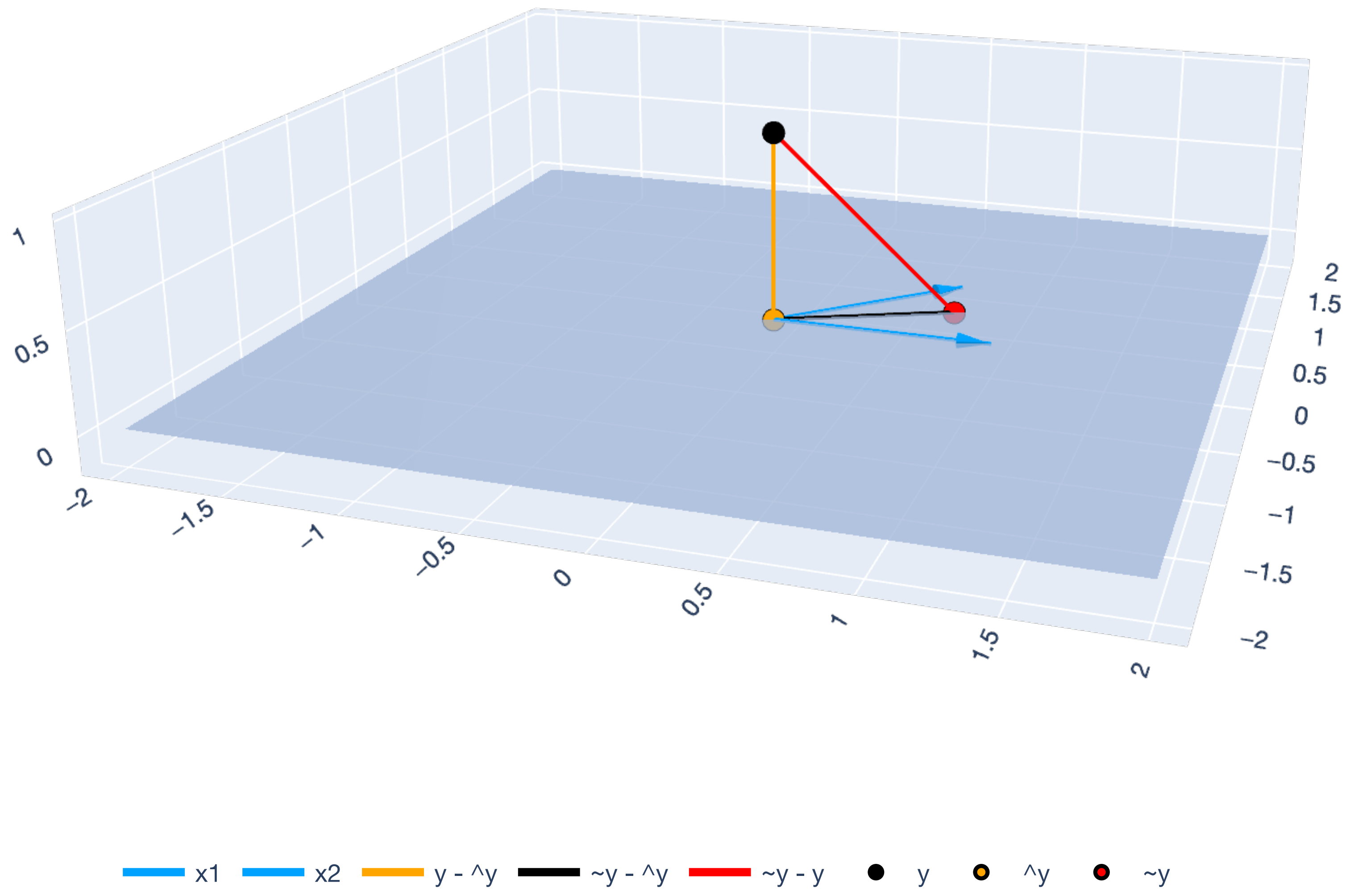
Summary

$\boxed{X\hat{w} \approx y}$

$\|X\hat{w} - y\| \leq \|Xw - y\|$

Goal: Find the $\hat{\mathbf{w}} \in \mathbb{R}^d$ that minimizes

$$\|\mathbf{X}\mathbf{w} - \mathbf{y}\|^2.$$



Theorem (OLS). If $n \geq d$ and $\text{rank}(\mathbf{X}) = d$, then:

$$\boxed{\hat{\mathbf{w}} = (\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{X}^\top\mathbf{y}.}$$

To get predictions $\hat{\mathbf{y}} \in \mathbb{R}^n$:

$$\hat{\mathbf{y}} = \mathbf{X}\hat{\mathbf{w}} = \mathbf{X}(\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{X}^\top\mathbf{y}.$$

$$\underbrace{\begin{bmatrix} - & x_1^T & - \\ & \vdots & \\ - & x_n^T & - \end{bmatrix}}_{X} \begin{bmatrix} | \\ w \\ | \end{bmatrix} = \begin{bmatrix} \hat{y}_1 \\ \vdots \\ \hat{y}_n \end{bmatrix}$$

$x_0 \in \mathbb{R}^d$

$x_0^T w = \hat{y}_0$

Least Squares
Summary

Goal: Find the $\hat{\mathbf{w}} \in \mathbb{R}^d$ that minimizes

$$\|\mathbf{X}\mathbf{w} - \mathbf{y}\|^2.$$

<u>Theorem (OLS).</u> If $n \geq d$ and $\mathrm{rank}(\mathbf{X}) = d$, then:

$$\boxed{\hat{\mathbf{w}} = (\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{X}^\top\mathbf{y}.}$$

To get predictions $\hat{\mathbf{y}} \in \mathbb{R}^n$:

$$\boxed{\hat{\mathbf{y}} = \mathbf{X}\hat{\mathbf{w}} = \underbrace{\mathbf{X}(\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{X}^\top}\mathbf{y}.}$$

Least Squares

Summary

Train Examples

X

Training Data

$\vec{y} \in \mathbb{R}^n$

$\hat{w}$

inverse

To get predictions $\hat{\mathbf{y}} \in \mathbb{R}^n$:

$\mathbb{R}^{d \times d}$

$O(d^3)$

$$\boxed{\hat{\mathbf{y}} = \mathbf{X}\hat{\mathbf{w}} = \boxed{\mathbf{X}(\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{X}^\top\mathbf{y}}}.$$

Test Examples: $\underbrace{z_1, \ldots, z_n \in \mathbb{R}^d}$

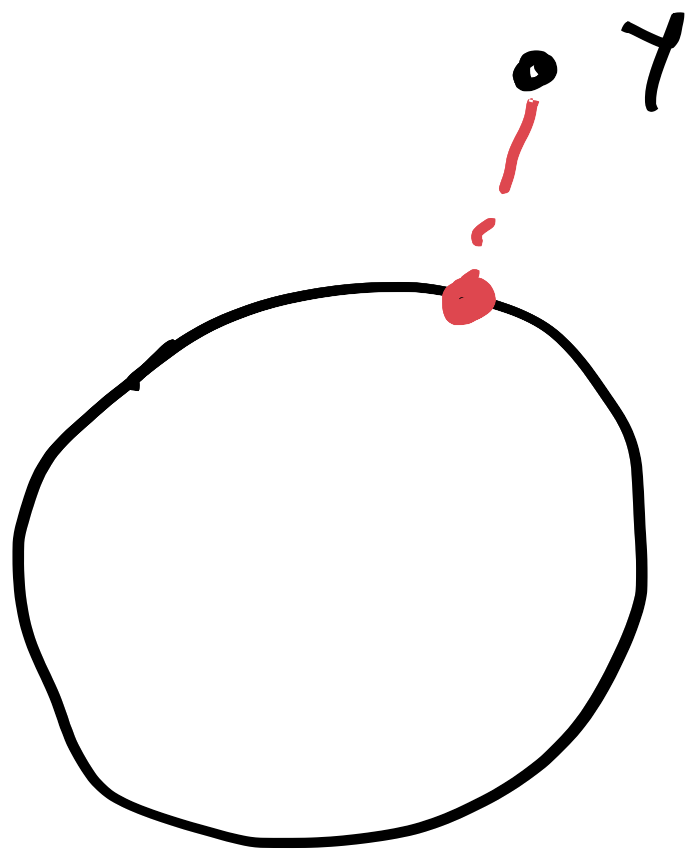
$$\boxed{Z\vec{w}}$$

Orthogonality

Projections

Projection

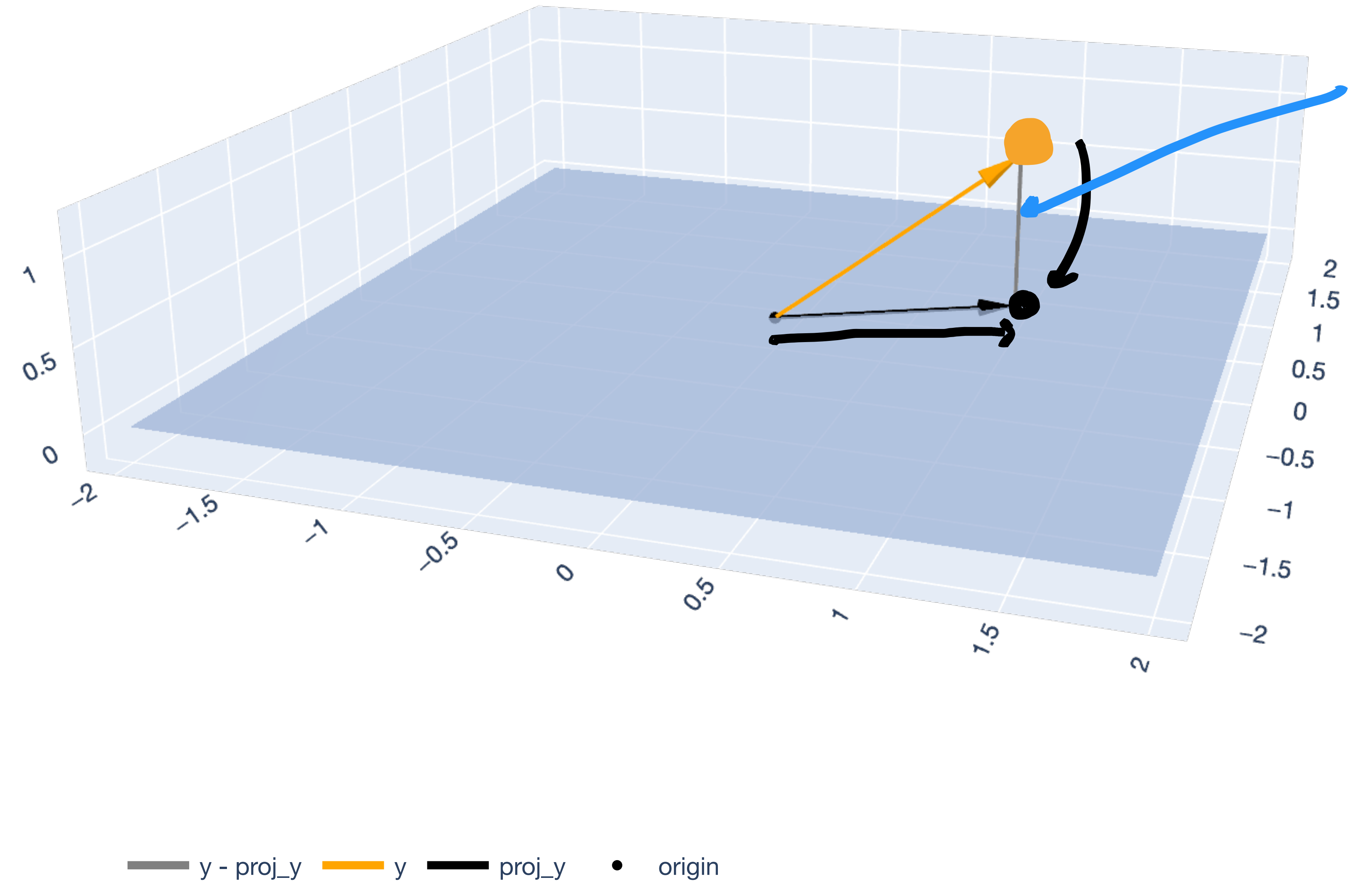
Idea: A vector's "shadow" on another set



For an arbitrary set $S \subseteq \mathbb{R}^n$, the ***projection*** of a vector $\mathbf{y} \in \mathbb{R}^n$ onto the set S is the closest vector $\hat{\mathbf{y}}$ in S to $\mathbf{y}$.

Denote this vector $\Pi_S(\mathbf{y}) := \hat{\mathbf{y}}$.

Projects $\vec{y}$ onto S.

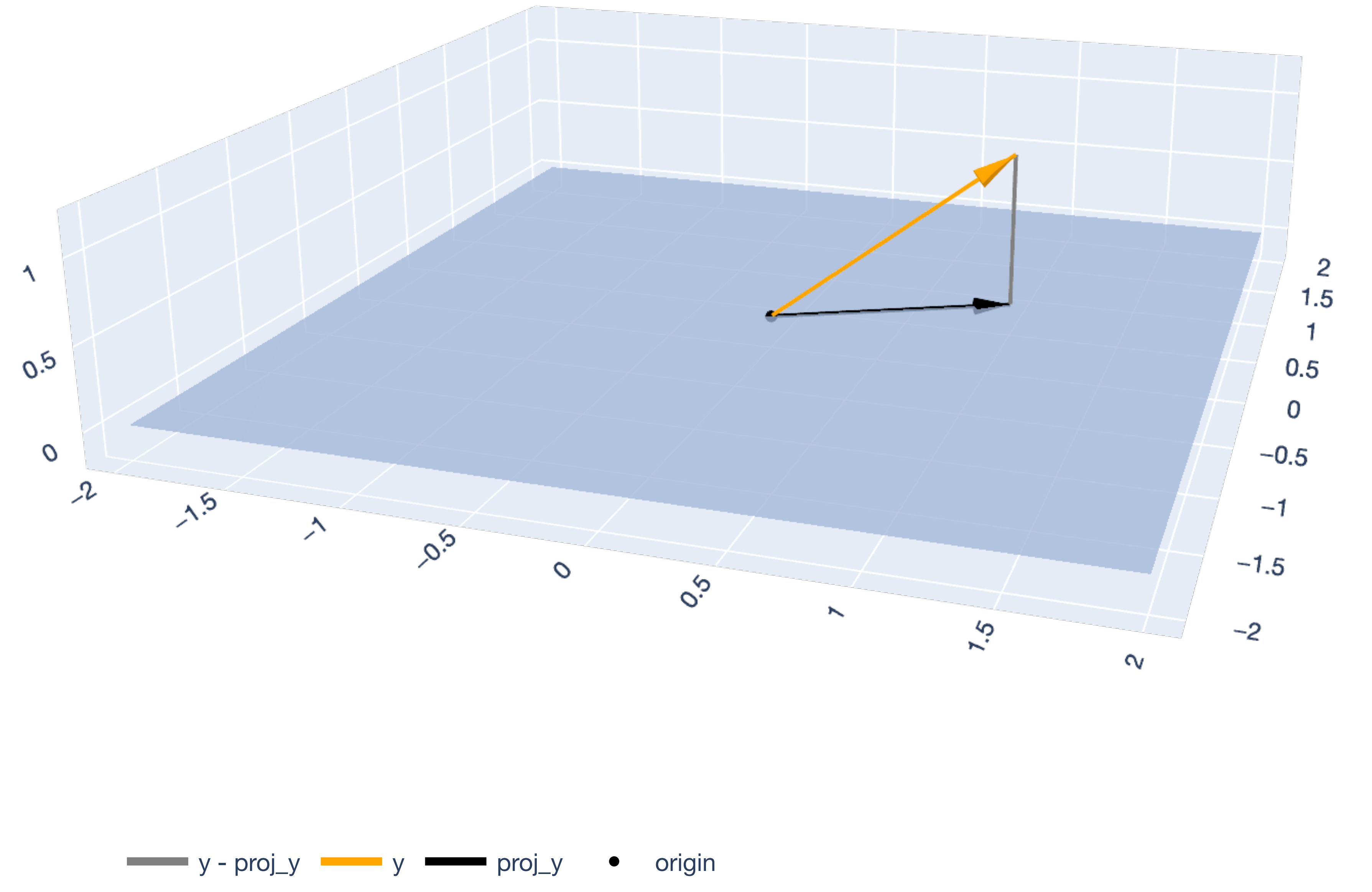


Projection

Projection of a vector onto an arbitrary set

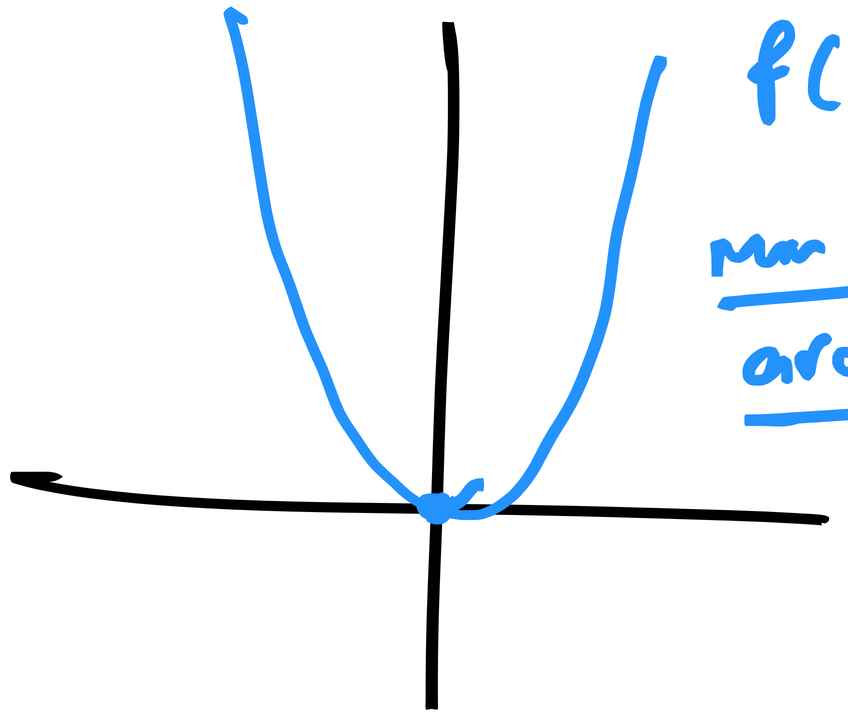
For an arbitrary set $S \subseteq \mathbb{R}^n$, the ***projection*** of a vector $\mathbf{y} \in \mathbb{R}^n$ onto the set S is the closest vector $\hat{\mathbf{y}}$ in S to $\mathbf{y}$.

Denote this vector $\Pi_S(\mathbf{y}) := \hat{\mathbf{y}}$.



Projection

Projection of a vector onto an arbitrary set

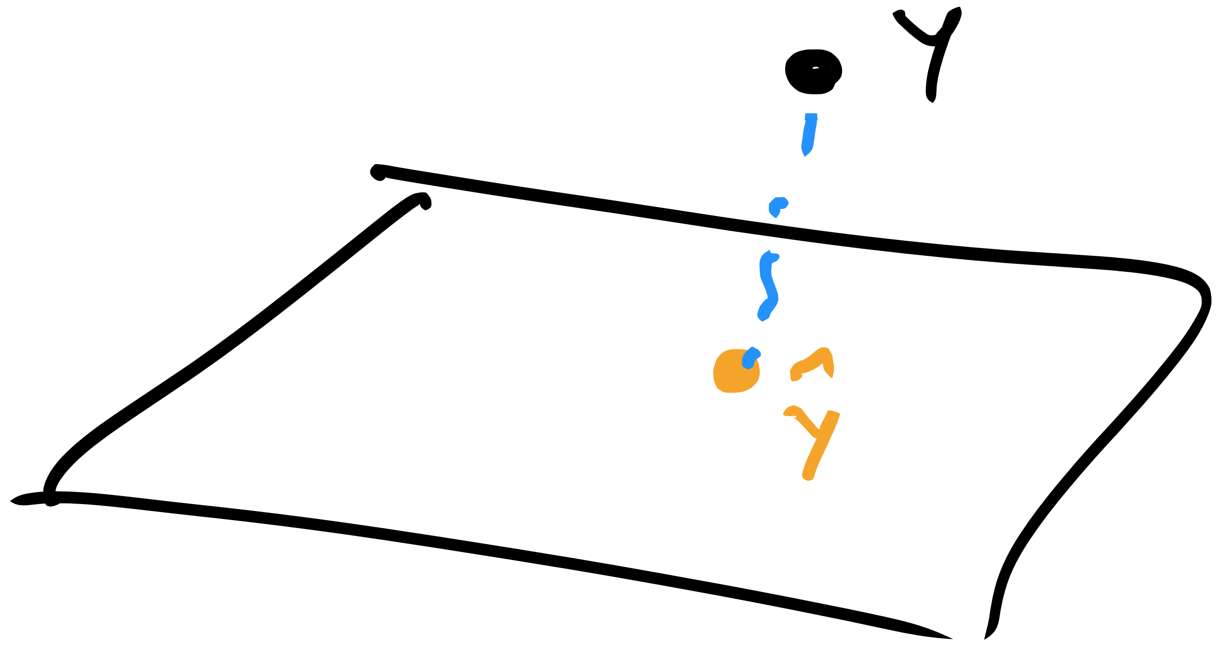


$f(x) = x^2$

min value : 0

argmin : 0

For an arbitrary set $S \subseteq \mathbb{R}^n$, the ***projection*** of a vector $\mathbf{y} \in \mathbb{R}^n$ onto the set S is the closest vector $\hat{\mathbf{y}}$ in S to $\mathbf{y}$.



Denote this vector $\Pi_S(\mathbf{y}) := \hat{\mathbf{y}}$.

"Closest" in a Euclidean ("least squares") distance sense:

$$\boxed{\Pi_S(\mathbf{y}) = \arg\min_{\hat{\mathbf{y}} \in S} \underbrace{\|\hat{\mathbf{y}} - \mathbf{y}\|}_{f(\hat{y})} = \|\hat{\mathbf{y}} - \mathbf{y}\|^2.}$$

Π_S

For a function $f(\hat{y})$, argmin, the input that gave the minimum.

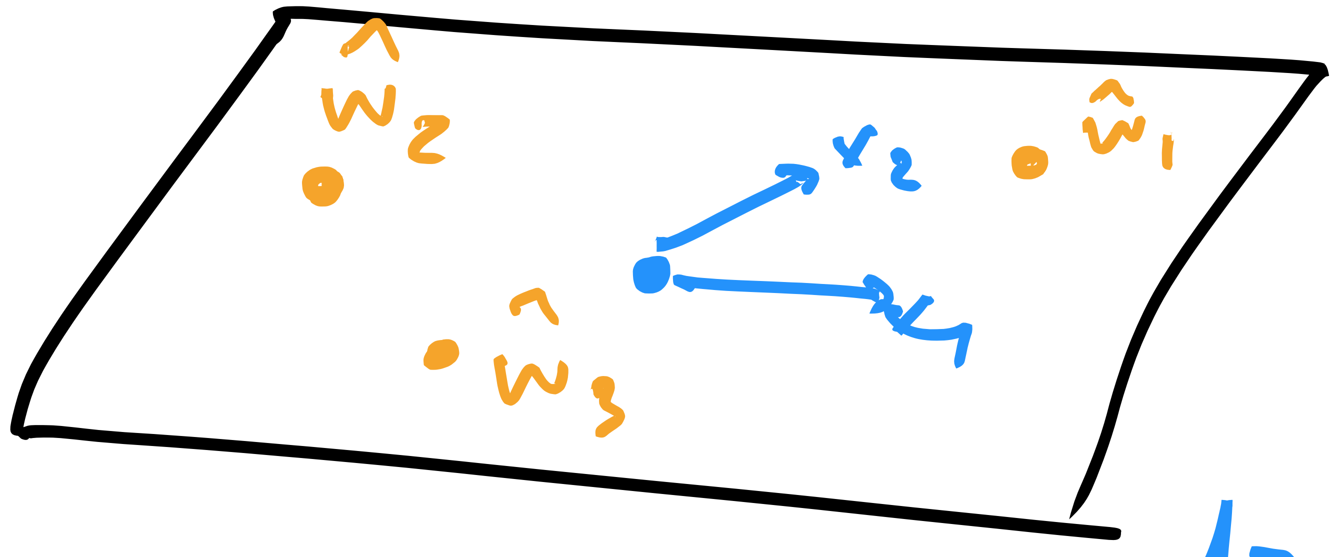
Projection

Projection of a vector onto a subspace

$$\begin{bmatrix} | & & | \\ x_1 & \cdots & x_d \\ | & & | \end{bmatrix} = X.$$

Let $\mathscr{X} \subseteq \mathbb{R}^n$ be a *subspace*, with the basis $\mathbf{x}_1, \ldots, \mathbf{x}_d \in \mathbb{R}^n$. Let $\mathbf{X} \in \mathbb{R}^{n \times d}$ be the matrix with $\mathbf{x}_1, \ldots, \mathbf{x}_d$ as its columns. *Any* point $\hat{\mathbf{y}} \in \mathscr{X}$ is a linear combination:

$$\boxed{\begin{aligned} \hat{\mathbf{y}} &= w_1\mathbf{x}_1 + \ldots + w_d\mathbf{x}_d \\ &= \mathbf{X}\mathbf{w} \end{aligned}}$$



$d = 2$

$\{x_1, x_2\}$

The projection of $\mathbf{y}$ onto $\mathscr{X}$ is:

$$\boxed{\Pi_{\mathscr{X}}(\mathbf{y}) = \arg\min_{\hat{\mathbf{y}} \in \mathscr{X}} \|\hat{\mathbf{y}} - \mathbf{y}\|^2}$$

Projection

Projection of a vector onto a subspace

Let $\mathscr{X} \subseteq \mathbb{R}^n$ be a *subspace*, with the basis $\mathbf{x}_1, \ldots, \mathbf{x}_d \in \mathbb{R}^n$. Let $\mathbf{X} \in \mathbb{R}^{n \times d}$ be the matrix with $\mathbf{x}_1, \ldots, \mathbf{x}_d$ as its columns. *Any* point $\hat{\mathbf{y}} \in \mathscr{X}$ is a linear combination:

$$\begin{aligned} \hat{\mathbf{y}} &= w_1 \mathbf{x}_1 + \ldots + w_d \mathbf{x}_d \\ &= \mathbf{X}\mathbf{w} \end{aligned}$$

This is equivalent to finding:

$$\hat{\mathbf{w}} = \arg\min_{\hat{\mathbf{w}} \in \mathbb{R}^d} \|\mathbf{X}\hat{\mathbf{w}} - \mathbf{y}\|^2$$

$\hat{y}$

$\hat{y} \in \mathscr{X}$ $\quad$ $X\hat{w}$

Least Squares as Projection

Projection Matrix

$d = 20$

$\vec{x} = (\text{weight, Height, Assists, Rebounds, ...})$

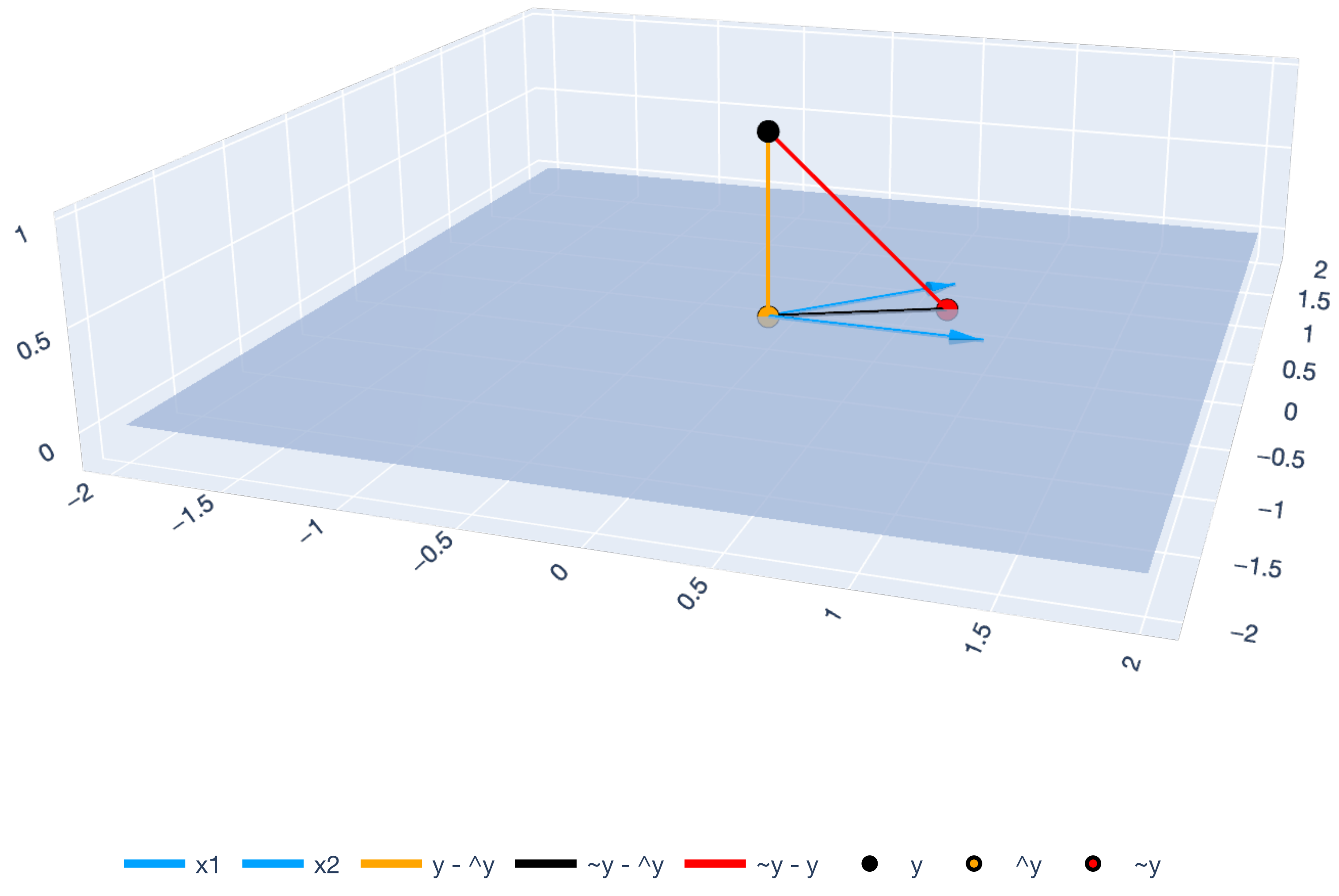
$n = 200$

$$\boxed{\hat{\mathbf{w}} = \underset{\hat{\mathbf{w}} \in \mathbb{R}^d}{\arg\min} \; \|\mathbf{X}\hat{\mathbf{w}} - \mathbf{y}\|^2}$$

This is just least squares! By what we've learned...

$$\boxed{\hat{\mathbf{w}} = (\mathbf{X}^\top \mathbf{X})^{-1}\mathbf{X}^\top \mathbf{y}}$$ OLS.

$$\Pi_{\mathscr{X}}(\mathbf{y}) = \underbrace{\hat{\mathbf{y}}}_{\text{prediction}} = \mathbf{X}(\mathbf{X}^\top \mathbf{X})^{-1}\mathbf{X}^\top \mathbf{y}$$



$x_1 \in \mathbb{R}^{200}$ ← weight of all 200.

$x_2 \in \mathbb{R}^{200}$ ← Height of all 200.

$\vdots$

$x_d \in \mathbb{R}^{200}$

Least Squares as Projection

Projection Matrix

$$\hat{\mathbf{w}} = \arg\min_{\hat{\mathbf{w}} \in \mathbb{R}^d} \|\mathbf{X}\hat{\mathbf{w}} - \mathbf{y}\|^2$$

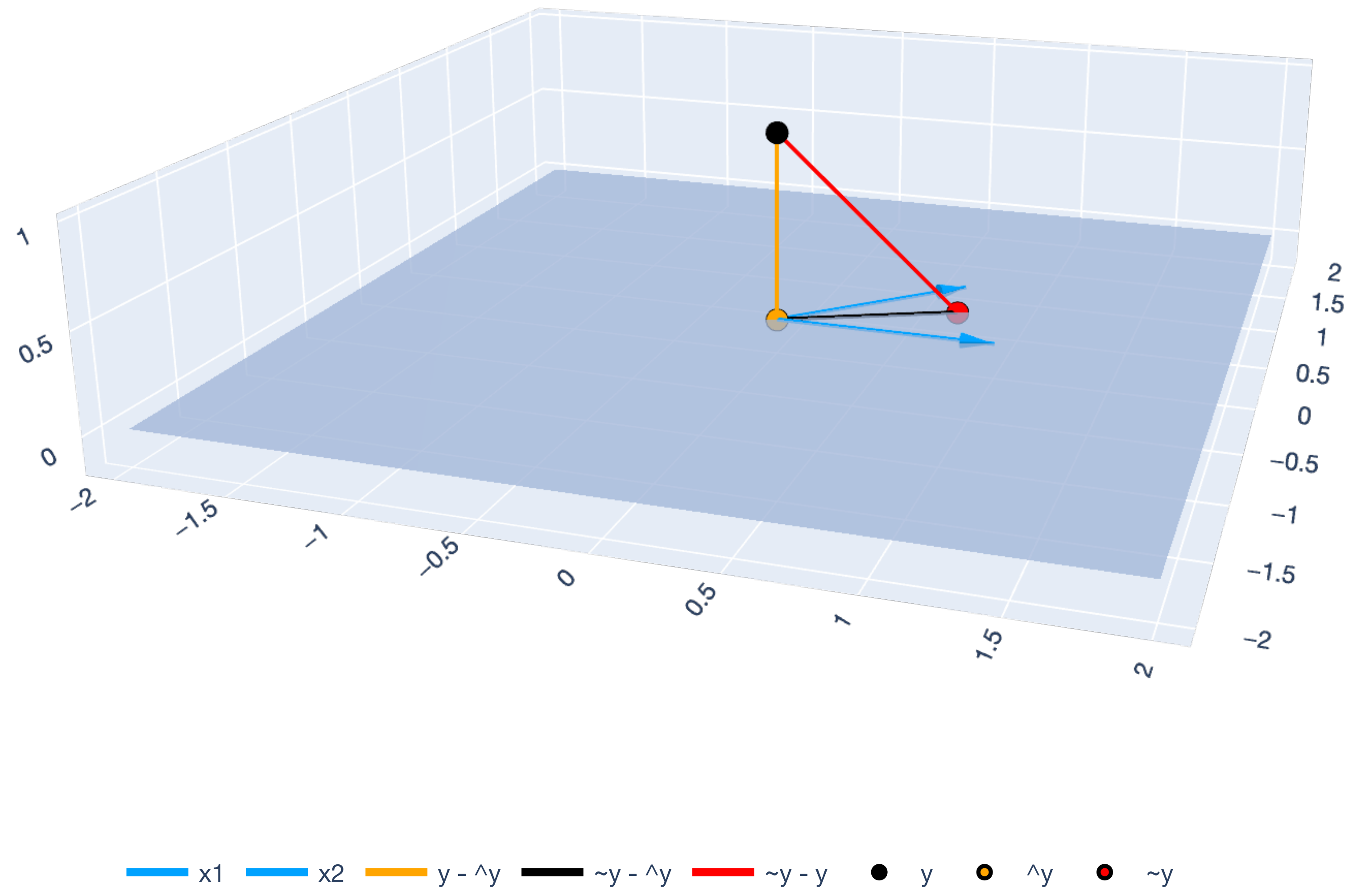
This is just least squares! By what we've learned…

$$\hat{\mathbf{w}} = (\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{X}^\top\mathbf{y}$$

$\in \mathbb{R}^{n \times n}$

$$\Pi_{\mathcal{X}}(\mathbf{y}) = \hat{\mathbf{y}} = \boxed{\mathbf{X}(\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{X}^\top}\mathbf{y}$$

Let $P_{\mathbf{X}} := \boxed{\mathbf{X}(\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{X}^\top \in \mathbb{R}^{n \times n}}$ be the ***projection matrix*** for $\mathrm{span}(\mathrm{col}(\mathbf{X}))$.



Linearity

Review from linear algebra

Linearity is the central property in linear algebra. Cooking is linear.

Bacon, egg, cheese (on roll)	*Bacon, egg, cheese (on bagel)*	*Lox sandwich*
1 egg	1 egg	0 egg
1 slice of cheese	1 slice of cheese	0 slice of cheese
1 slice bacon	1 slice bacon	0 slice bacon
1 Kaiser roll	0 Kaiser roll	0 Kaiser roll
0 cream cheese	0 cream cheese	1 cream cheese
0 slices of lox	0 slices of lox	2 slices of lox
0 bagel	1 bagel	1 bagel

Linearity

Review from linear algebra

Linearity is the central property in linear algebra. A function ("transformation") $T : \mathbb{R}^d \to \mathbb{R}^n$ is ***linear*** if T satisfies these two properties for any two vectors $\mathbf{a}, \mathbf{b} \in \mathbb{R}^d$:

$$T(\mathbf{a} + \mathbf{b}) = T(\mathbf{a}) + T(\mathbf{b})$$

$T(c\mathbf{a}) = cT(\mathbf{a})$ for any $c \in \mathbb{R}$.

Linearity

Review from linear algebra

Example. Consider the function $T : \mathbb{R}^3 \to \mathbb{R}$, defined by:

$$T(\mathbf{x}) = 2x_1 + 3x_3.$$

Linearity

Review from linear algebra

Matrices also play by these rules. Let $\mathbf{X} \in \mathbb{R}^{n\times d}$ be a matrix and let $\mathbf{w}, \mathbf{v} \in \mathbb{R}^{d}$ be vectors.

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

$$\mathbf{X}(c\mathbf{w}) = c(\mathbf{X}\mathbf{w}) \text{ for any } c \in \mathbb{R}.$$

Linearity

Review from linear algebra

Theorem (Equivalence of linear transformations and matrices).

Any linear transformation $T : \mathbb{R}^d \to \mathbb{R}^n$ has a corresponding matrix $\mathbf{A}_T \in \mathbb{R}^{n \times d}$ such that:

$$T(\mathbf{x}) = \mathbf{A}_T \mathbf{x}.$$

Any matrix $\mathbf{A} \in \mathbb{R}^{n \times d}$ has a corresponding linear transformation $T_{\mathbf{A}} : \mathbb{R}^d \to \mathbb{R}^n$ such that:

$$T_{\mathbf{A}}(\mathbf{x}) = \mathbf{A}\mathbf{x}.$$

Linearity

Review from linear algebra

$$T(\mathbf{x}) = \mathbf{A}_T\mathbf{x} \text{ and } T_{\mathbf{A}}(\mathbf{x}) = \mathbf{A}\mathbf{x}$$

This means that *matrix-vector multiplication is the same as applying a linear transformation*. So one way of thinking of a matrix is an "action" applied to vectors.

Least Squares as Projection

Projection Matrix

Let $\mathscr{X} \subseteq \mathbb{R}^d$ be a *subspace* with basis $\mathbf{x}_1, \ldots, \mathbf{x}_d \in \mathbb{R}^n$. If $\mathbf{x}_1, \ldots, \mathbf{x}_d$ are linearly independent, making up the matrix $\mathbf{X} \in \mathbb{R}^{n \times d}$,

$$\boxed{P_{\mathbf{X}} := \mathbf{X}(\mathbf{X}^\top \mathbf{X})^{-1} \mathbf{X}^\top \in \mathbb{R}^{n \times n}}$$

is the ***projection matrix*** onto $\mathscr{X}$. To project a vector $\mathbf{y} \in \mathbb{R}^n$ onto $\mathscr{X}$, compute:

$$\Pi_{\mathscr{X}}(\mathbf{y}) = \hat{\mathbf{y}} = P_{\mathbf{X}}\mathbf{y} = \mathbf{X}(\mathbf{X}^\top \mathbf{X})^{-1} \mathbf{X}\mathbf{y}.$$

Least Squares

Orthonormal Bases and Projection

Norms and Inner Products

Unit Vectors

A vector $\mathbf{v} \in \mathbb{R}^d$ is a ***unit vector*** if $\|\mathbf{v}\| = 1$.

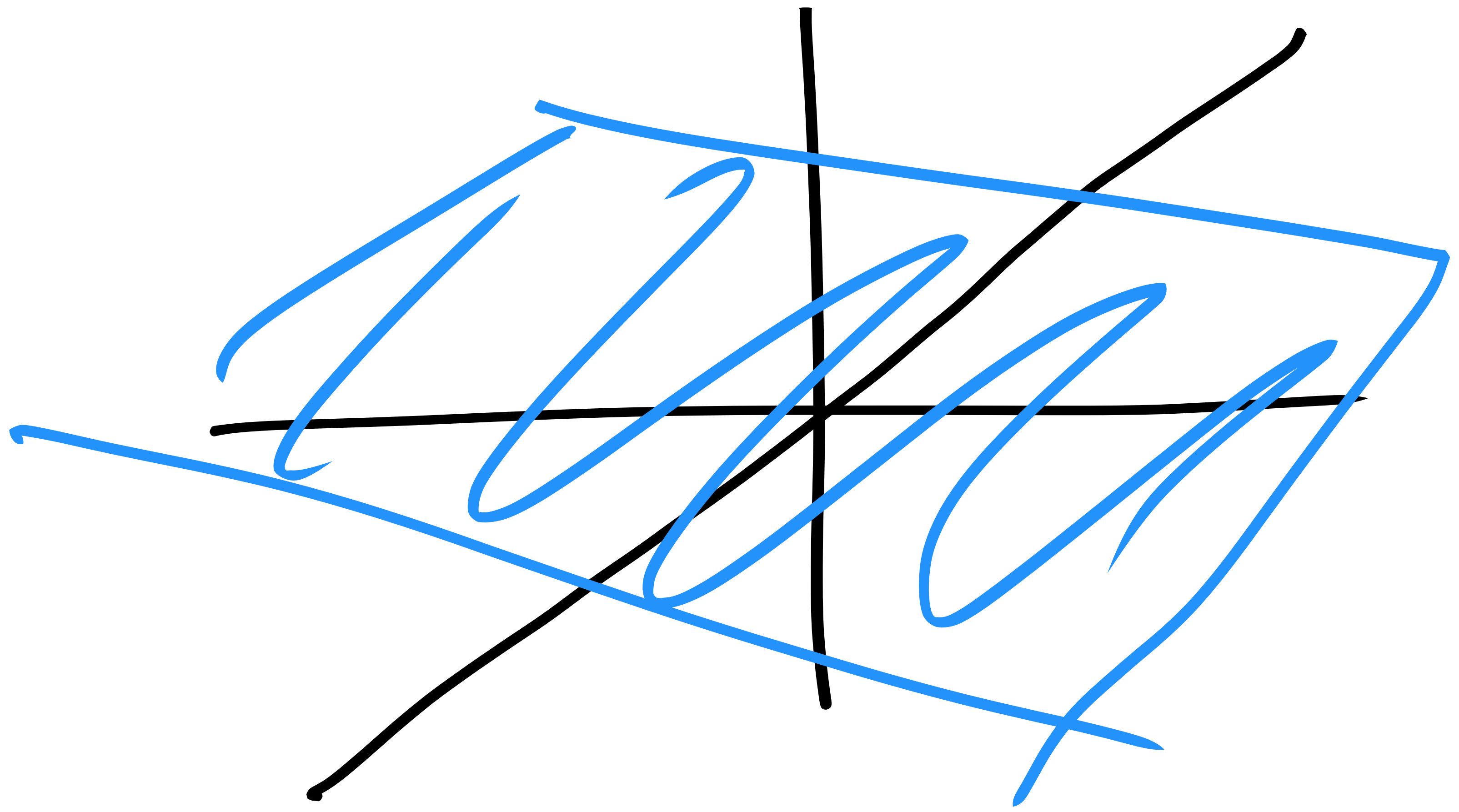
We can convert any vector into a unit vector by dividing itself by its norm:

$$\boxed{\frac{\mathbf{v}}{\|\mathbf{v}\|}}$$

$\left\| \frac{v}{\|v\|} \right\| = \frac{1}{\|v\|} \|v\| = 1.$

Orthonormal Basis

"Good" Bases



How should we represent a subspace?

Take, for example, the subspace $\mathcal{S} = \{\mathbf{v} \in \mathbb{R}^3 : v_3 = 0\}$.

Orthonormal Basis

“Good” Bases

$$\mathcal{S} = \{\mathbf{v} \in \mathbb{R}^3 : v_3 = 0\}$$

Attempt 1: Use the span of a set of vectors: $\text{span}\left(\underline{\begin{bmatrix}1\\0\\0\end{bmatrix}, \begin{bmatrix}0\\1\\0\end{bmatrix}, \begin{bmatrix}1\\1\\0\end{bmatrix}}\right).$

Orthonormal Basis

“Good” Bases

dim(S) = 2

$$\mathcal{S} = \{\mathbf{v} \in \mathbb{R}^3 : v_3 = 0\}$$

Attempt 1: Use the span of a set of vectors: $\operatorname{span}\left(\begin{bmatrix}2\\1\\0\end{bmatrix}, \begin{bmatrix}0\\1\\0\end{bmatrix}, \cancel{\begin{bmatrix}2\\3\\0\end{bmatrix}}\right).$

Attempt 2: Use the span of a set of linearly independent vectors (a basis):

$$\operatorname{span}\left(\begin{bmatrix}2\\1\\0\end{bmatrix}, \begin{bmatrix}0\\1\\0\end{bmatrix}\right).$$

Orthonormal Basis

"Good" Bases

$$\mathcal{S} = \{\mathbf{v} \in \mathbb{R}^3 : v_3 = 0\}$$

Attempt 1: Use the span of a set of vectors: $\operatorname{span}\left(\begin{bmatrix}2\\1\\0\end{bmatrix}, \begin{bmatrix}0\\1\\0\end{bmatrix}, \begin{bmatrix}2\\3\\0\end{bmatrix}\right).$

Attempt 2: Use the span of a set of linearly independent vectors (a basis):

$$\operatorname{span}\left(\begin{bmatrix}2\\1\\0\end{bmatrix}, \begin{bmatrix}0\\1\\0\end{bmatrix}\right).$$

Attempt 3: Use the span of an orthonormal set of vectors (an ***orthonormal basis***):

$$\operatorname{span}\left(\begin{bmatrix}1\\0\\0\end{bmatrix}, \begin{bmatrix}0\\1\\0\end{bmatrix}\right).$$

$1^2 + 0^2 + 0^2 = 1$

$\begin{bmatrix}1 & 0 & 0\end{bmatrix}^T \begin{bmatrix}0\\1\\0\end{bmatrix} = 0.$

Orthonormal Basis

"Good" Bases

$$\mathcal{S} = \{\mathbf{v} \in \mathbb{R}^3 : v_3 = 0\}$$

$$\text{span}\left(\begin{bmatrix}2\\1\\0\end{bmatrix}, \begin{bmatrix}0\\1\\0\end{bmatrix}, \begin{bmatrix}2\\3\\0\end{bmatrix}\right) \qquad \text{span}\left(\begin{bmatrix}2\\1\\0\end{bmatrix}, \begin{bmatrix}0\\1\\0\end{bmatrix}\right) \qquad \text{span}\left(\begin{bmatrix}1\\0\\0\end{bmatrix}, \begin{bmatrix}0\\1\\0\end{bmatrix}\right)$$

Bad (Redundant)

Bad (Not unit length)

Best :)

Orthonormal Basis

Definition

A set of vectors $\boxed{\mathbf{u}_1, \ldots, \mathbf{u}_n} \in \mathcal{S}$ is an ***orthonormal basis*** for the subspace $\mathcal{S}$ if they are a basis for $\mathcal{S}$ and, additionally:

$$\boxed{\langle \mathbf{u}_i, \mathbf{u}_j \rangle = 0 \text{ for } i \neq j.}$$

For any pair.

$$\boxed{\|\mathbf{u}_i\| = 1 \text{ for } i \in [n].}$$

Unit length.

Orthonormal Basis

Orthogonal Matrices

$$U = \begin{bmatrix} | & & | \\ u_1 & \cdots & u_d \\ | & & | \end{bmatrix} \quad d \times d$$

A square matrix $\mathbf{U} \in \mathbb{R}^{d \times d}$ is an ***orthogonal matrix*** if its columns $\mathbf{u}_1, \ldots, \mathbf{u}_d \in \mathbb{R}^d$ are orthogonal unit vectors:

$$\langle \mathbf{u}_i, \mathbf{u}_j \rangle = 0 \text{ for } i \neq j.$$

$$||\mathbf{u}_i|| = 1 \text{ for } i \in [d].$$

These form an orthonormal basis for $\text{span}(\text{col}(\mathbf{U}))$. subspace

Its rows are also orthogonal.

Orthonormal Basis

Orthogonal Matrices

A matrix $\mathbf{U} \in \mathbb{R}^{n \times d}$ is an ***semi-orthogonal matrix*** if its columns $\mathbf{u}_1, \ldots, \mathbf{u}_d \in \mathbb{R}^n$ are orthogonal unit vectors:

$$\langle \mathbf{u}_i, \mathbf{u}_j \rangle = 0 \text{ for } i \neq j.$$

$$||\mathbf{u}_i|| = 1 \text{ for } i \in [d].$$

These form an orthonormal basis for $\text{span}(\text{col}(\mathbf{U}))$.

Orthonormal Basis

Properties of Orthogonal Matrices

$$\begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix} = I.$$

$$\boxed{U^{-1} = U^T}$$

Let a square matrix $\mathbf{U} \in \mathbb{R}^{d\times d}$ be an ***orthogonal matrix.*** Then:

$\mathbf{U}^T$ is its ~~own~~ inverse: $\boxed{\mathbf{U}^\top\mathbf{U} = \mathbf{U}\mathbf{U}^\top = \mathbf{I}.}$

U

U is length-preserving: $\|\mathbf{U}\mathbf{v}\| = \|\mathbf{v}\|$.

$$\begin{bmatrix} - u_1 - \\ \\ \\ \end{bmatrix} \begin{bmatrix} | & | & & | \\ u_1 & u_2 & \cdots & u_d \\ | & | & & | \end{bmatrix}$$

Orthonormal Basis

Properties of Orthogonal Matrices

$n \neq d$

→ columns are an orthonormal basis for span(col(U)).

Let matrix $\mathbf{U} \in \boxed{\mathbb{R}^{n\times d}}$ be an ***semi-orthogonal matrix.*** Then:

$\mathbf{U}^{T}$ **is its ~~own~~ U left inverse:** $\boxed{\mathbf{U}^\top \mathbf{U} = \mathbf{I}}$.

$UU^T \neq I$

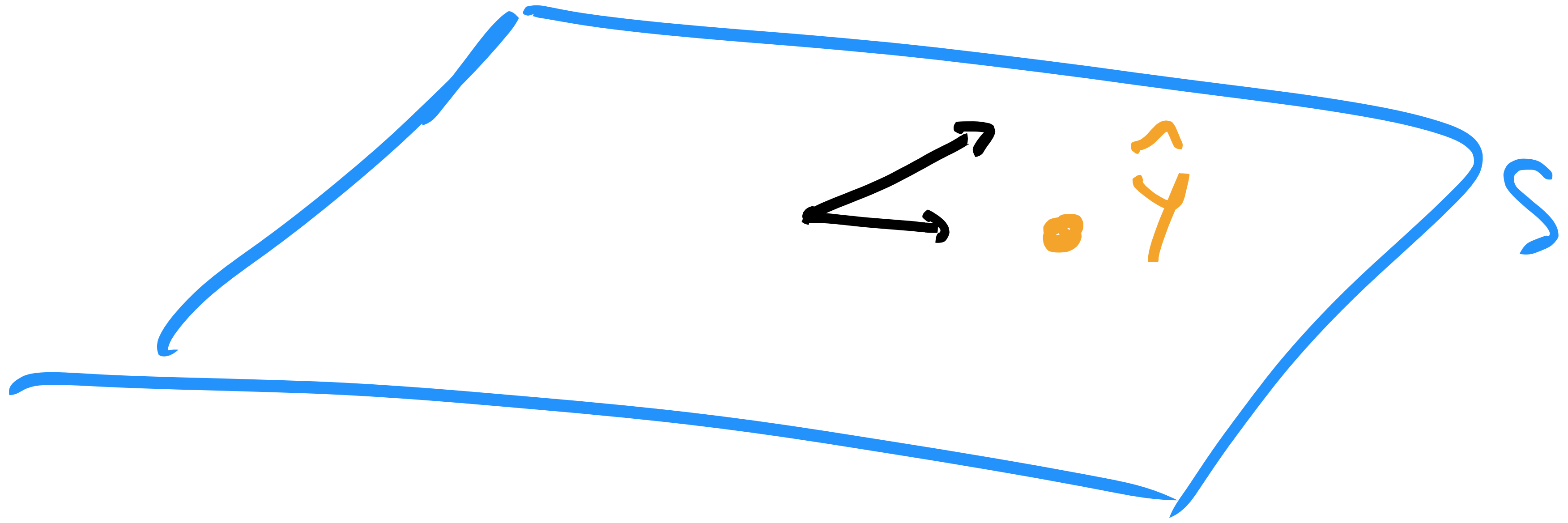
U is length-preserving: $||\mathbf{U}\mathbf{v}|| = ||\mathbf{v}||$.

Orthogonal Bases in Least Squares

What if we had an orthogonal basis?

A basis is just a “language” for representing vectors in a subspace. For example, consider the subspace $\mathcal{S} = \{\mathbf{v} \in \mathbb{R}^3 : v_3 = 0\}$ and the vector

$$\hat{\mathbf{y}} = \begin{bmatrix} 1 \\ 1 \\ 0 \end{bmatrix}$$



<u>Basis 1:</u> $\left\{ \begin{bmatrix} 2 \\ 1 \\ 0 \end{bmatrix}, \begin{bmatrix} 0 \\ 1 \\ 0 \end{bmatrix} \right\}$

Orthogonal Bases in Least Squares

What if we had an orthogonal basis?

A basis is just a "language" for representing vectors in a subspace. For example, consider the subspace $\boxed{\mathcal{S} = \{\mathbf{v} \in \mathbb{R}^3 : v_3 = 0\}}$ and the vector

$$\hat{\mathbf{y}} = \begin{bmatrix} 1 \\ 1 \\ 0 \end{bmatrix}$$

Basis 2: $\underline{\left\{ \begin{bmatrix} 1 \\ 0 \\ 0 \end{bmatrix}, \begin{bmatrix} 0 \\ 1 \\ 0 \end{bmatrix} \right\}}$ $\}$

$$\hat{y} = w_1 \begin{bmatrix} 1 \\ 0 \\ 0 \end{bmatrix} + w_2 \begin{bmatrix} 0 \\ 1 \\ 0 \end{bmatrix} = \begin{bmatrix} 1 & 0 \\ 0 & 1 \\ 0 & 0 \end{bmatrix} \begin{bmatrix} w_1 \\ w_2 \end{bmatrix}$$

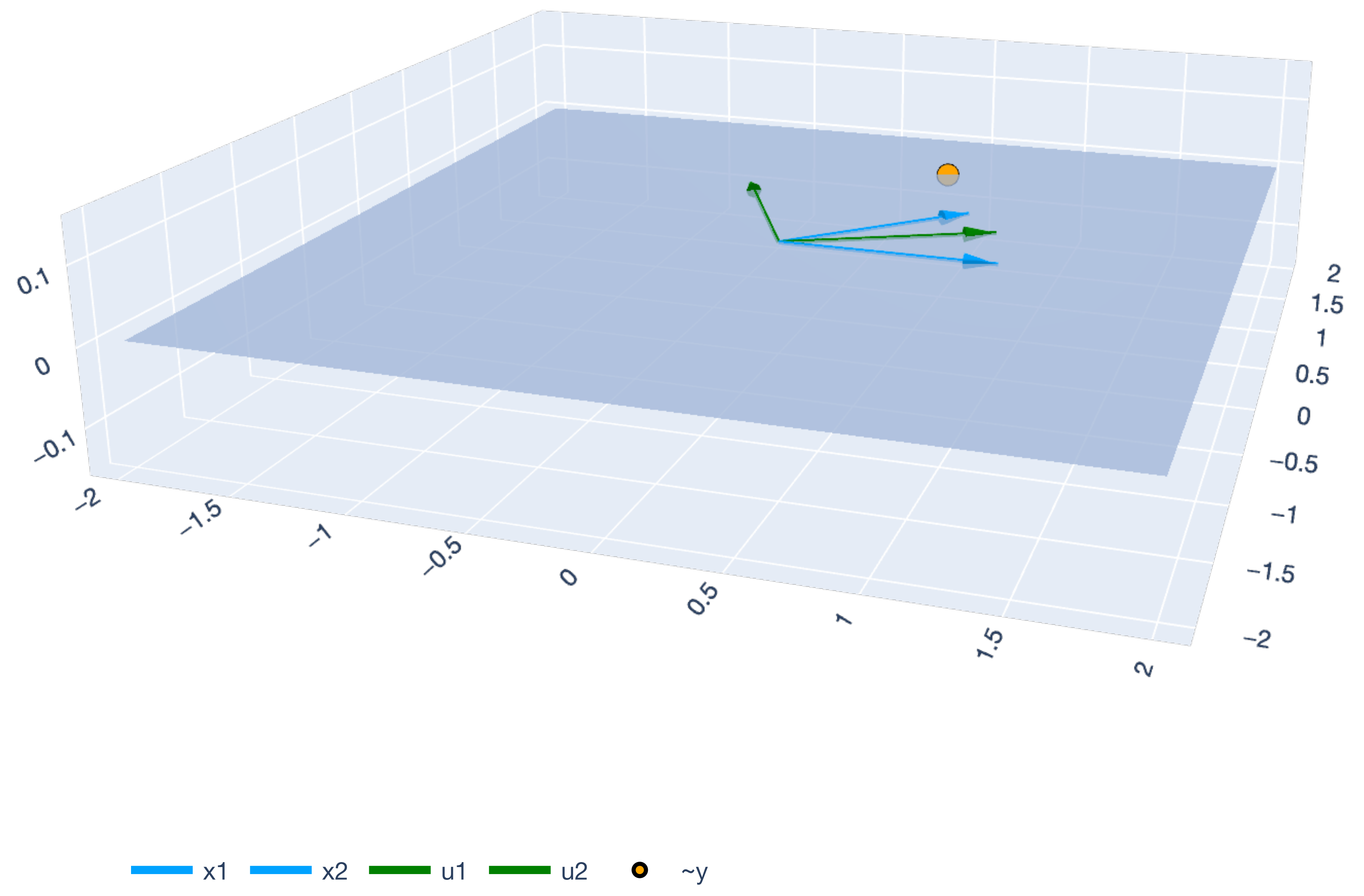
$$\boxed{\begin{bmatrix} 1 & 0 \\ 0 & 1 \\ 0 & 0 \end{bmatrix} \begin{bmatrix} w_1 \\ w_2 \end{bmatrix} = \hat{y}}$$

Orthogonal Bases in Least Squares

What if we had an orthogonal basis?

Every subspace $\mathscr{X} \subseteq \mathbb{R}^n$ has many choices of bases.

Some are better than others.



Orthogonal Bases in Least Squares

What if we had an orthogonal basis?

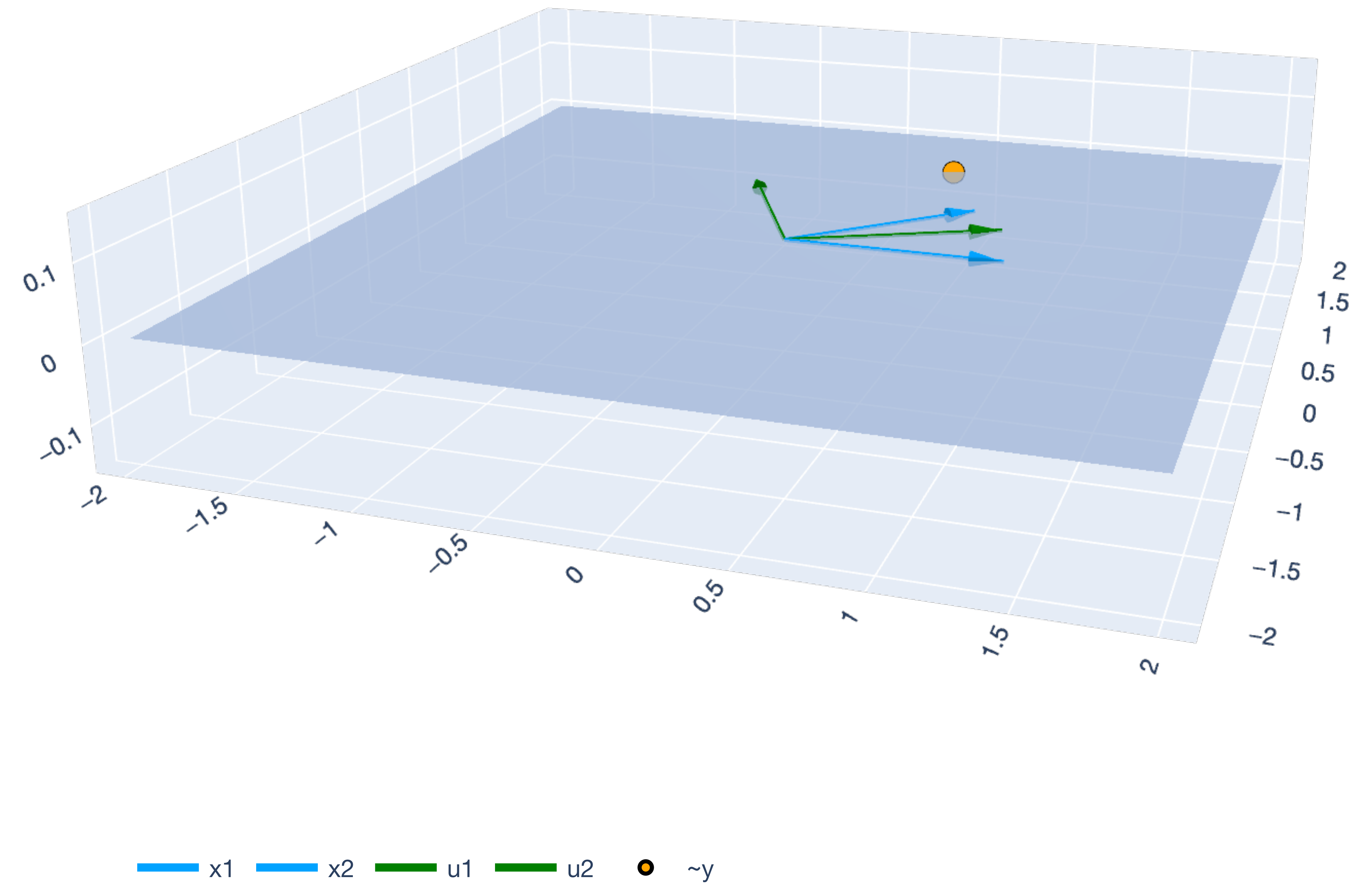
Let $\mathscr{X} \subseteq \mathbb{R}^n$ be a subspace, with $\dim(\mathscr{X}) = d$.

One basis: $\mathbf{x}_1, \ldots, \mathbf{x}_d \in \mathbb{R}^n$, with matrix $\mathbf{X} \in \mathbb{R}^{n \times d}$.

$= \begin{bmatrix} | & & | \\ x_1 & \cdots & x_d \\ | & & | \end{bmatrix} = X$

Another basis: $\mathbf{u}_1, \ldots, \mathbf{u}_d \in \mathbb{R}^n$, with matrix $\mathbf{U} \in \mathbb{R}^{n \times d}$.

$\begin{bmatrix} | & & | \\ u_1 & \cdots & u_d \\ | & & | \end{bmatrix} = U.$



$$\boxed{\operatorname{span}(\operatorname{col}(X)) = \operatorname{span}(\operatorname{col}(U)) = \mathscr{X}}$$

Orthogonal Bases in Least Squares

What if we had an orthogonal basis?

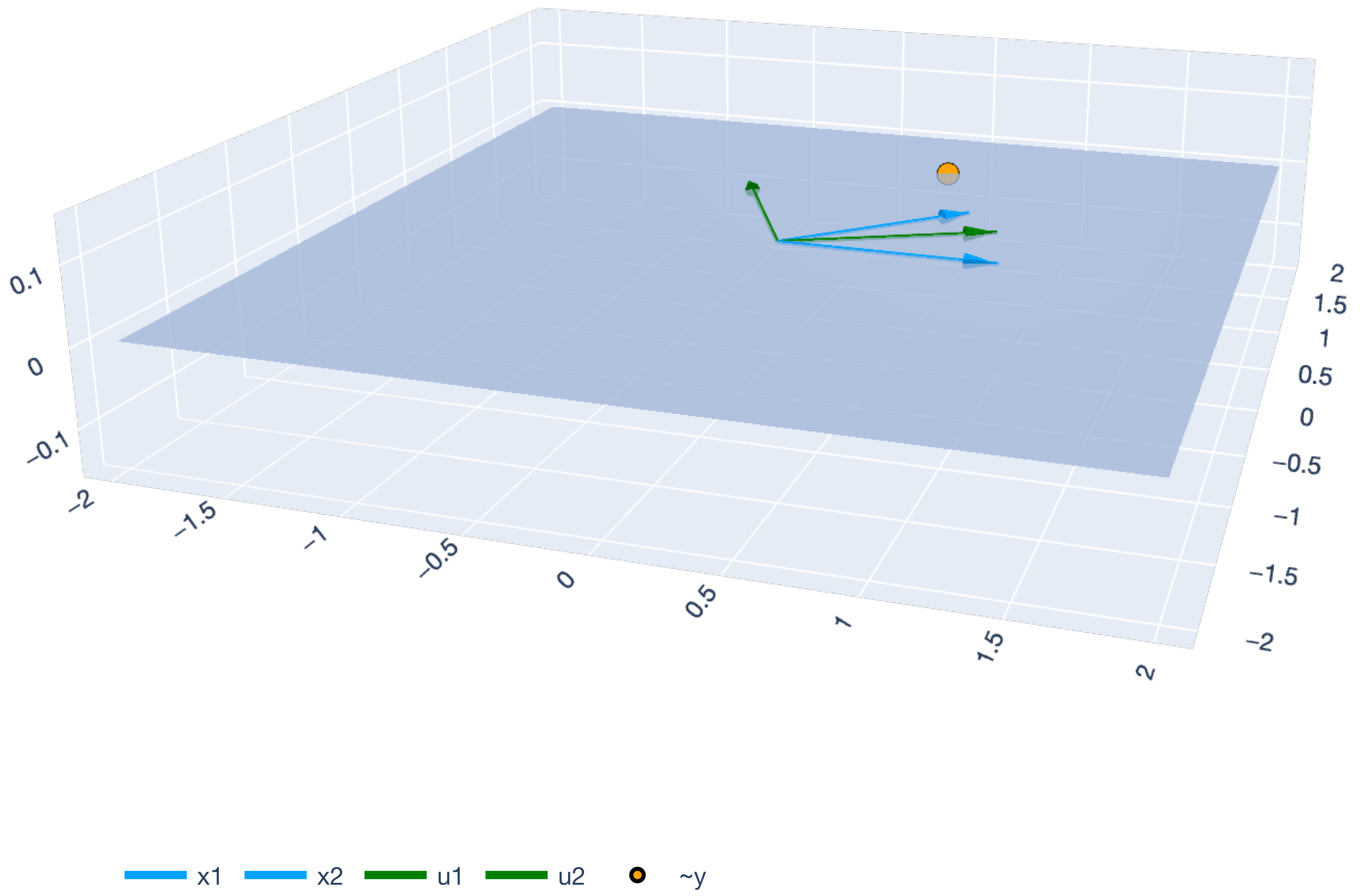
Let $\mathscr{X} \subseteq \mathbb{R}^n$ be a subspace, with $\dim(\mathscr{X}) = d$.

One basis: $\mathbf{x}_1, \ldots, \mathbf{x}_d \in \mathbb{R}^n$, with matrix $\mathbf{X} \in \mathbb{R}^{n \times d}$.

Another basis: $\mathbf{u}_1, \ldots, \mathbf{u}_d \in \mathbb{R}^n$, with matrix $\mathbf{U} \in \mathbb{R}^{n \times d}$.

Then,

$$\boxed{\mathscr{X} = \text{span}(\text{col}(\mathbf{U})) = \text{span}(\text{col}(\mathbf{X})).}$$



Orthogonal Bases in Least Squares

What if we had an orthogonal basis?

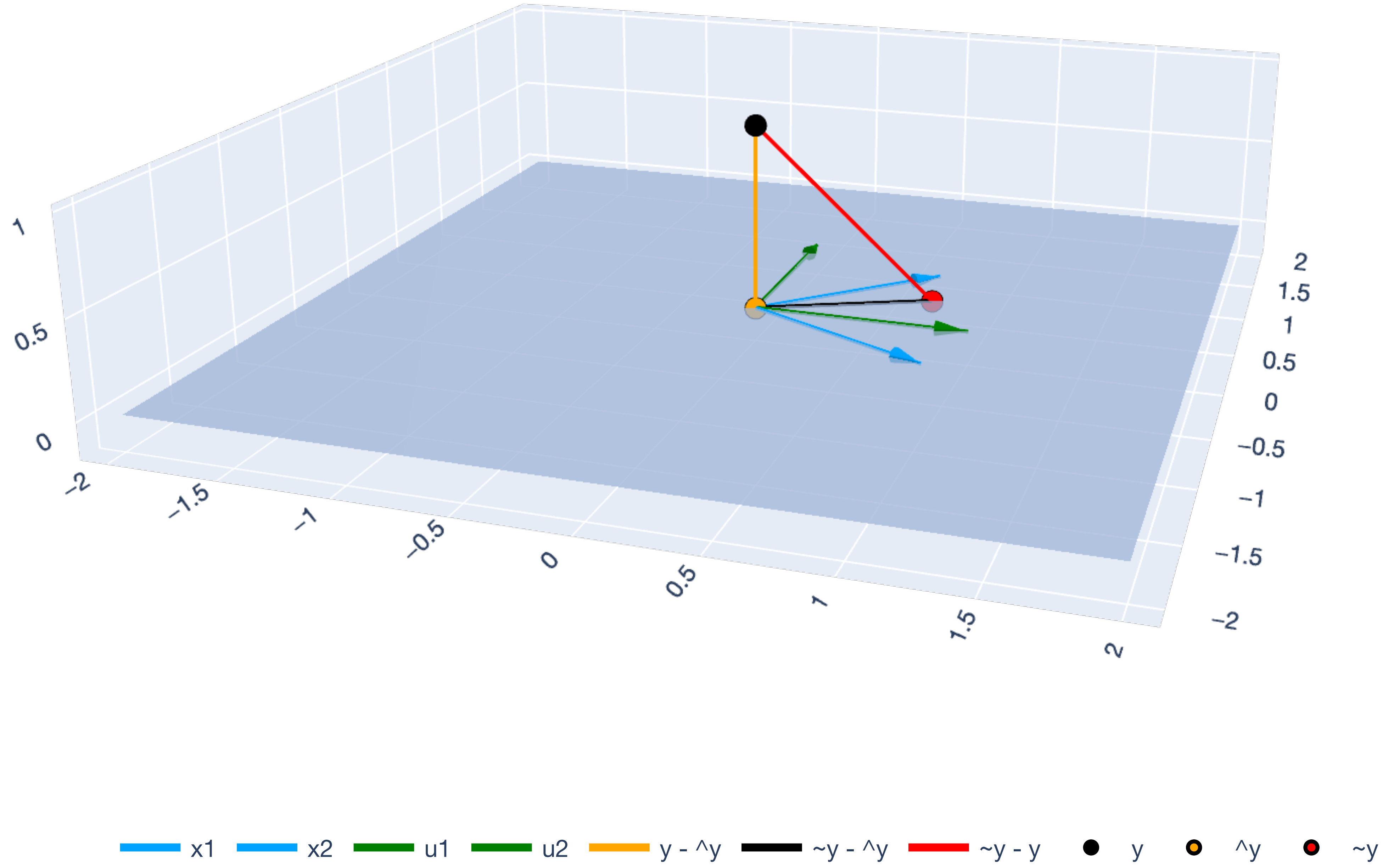
Let $\mathscr{X} \subseteq \mathbb{R}^n$ be a subspace, with $\dim(\mathscr{X}) = d$.

$$\mathscr{X} = \text{span}(\text{col}(\mathbf{U})) = \text{span}(\text{col}(\mathbf{X})).$$

Therefore, for any $\hat{\mathbf{y}} \in \mathscr{X}$, we can write:

$$\boxed{\hat{\mathbf{y}} = \mathbf{X}\hat{\mathbf{w}} = \mathbf{U}\hat{\mathbf{w}}_{onb}.}$$

Both $\hat{\mathbf{w}}, \hat{\mathbf{w}}_{onb} \in \mathbb{R}^d$ are valid ways to "represent" $\hat{\mathbf{y}}$.



$\hat{w} \qquad \hat{w}_{onb}$

Orthogonal Bases in Least Squares

What if we had an orthogonal basis?

How do we find $\hat{\mathbf{w}}_{onb} \in \mathbb{R}^d$ in $\hat{\mathbf{y}} = \mathbf{U}\hat{\mathbf{w}}_{onb}$? Least squares!

$\hat{Y} = X\hat{w}$

$$\boxed{\hat{\mathbf{w}}_{onb} = \arg\min_{\hat{\mathbf{w}}_{onb} \in \mathbb{R}^d} \|\mathbf{y} - \underline{\mathbf{U}\hat{\mathbf{w}}_{onb}}\|^2}$$

OLS

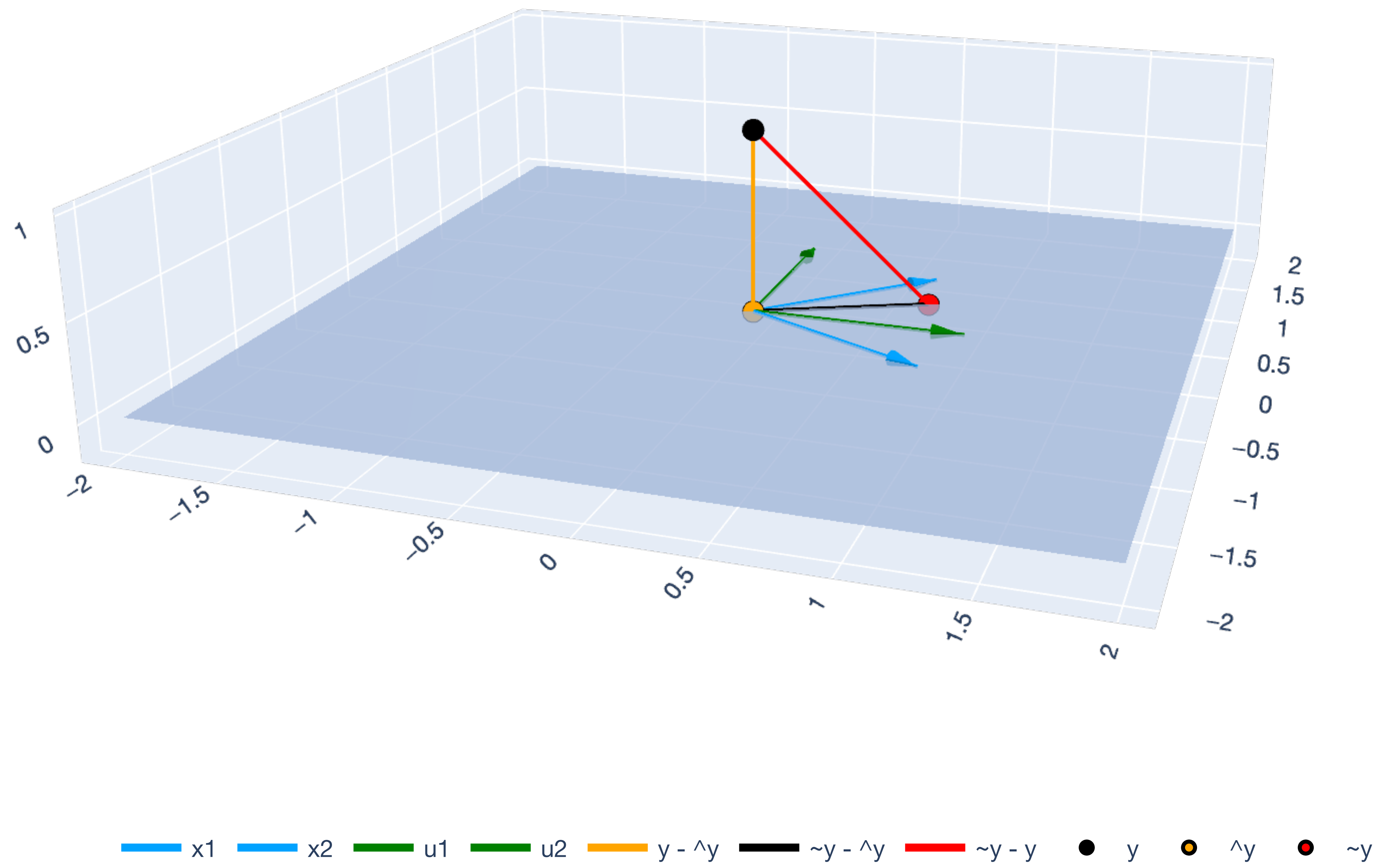
The columns of $\mathbf{U}$ give an ONB for $\mathcal{X}$…

$$\boxed{\hat{\mathbf{w}}_{onb} = \cancel{(\mathbf{U}^\top\mathbf{U})^{-1}}\mathbf{U}^\top\mathbf{y}}$$

U is semi-orthogonal

$U \in \mathbb{R}^{n \times d}$

$U^T U = I$.



Orthogonal Bases in Least Squares

What if we had an orthogonal basis?

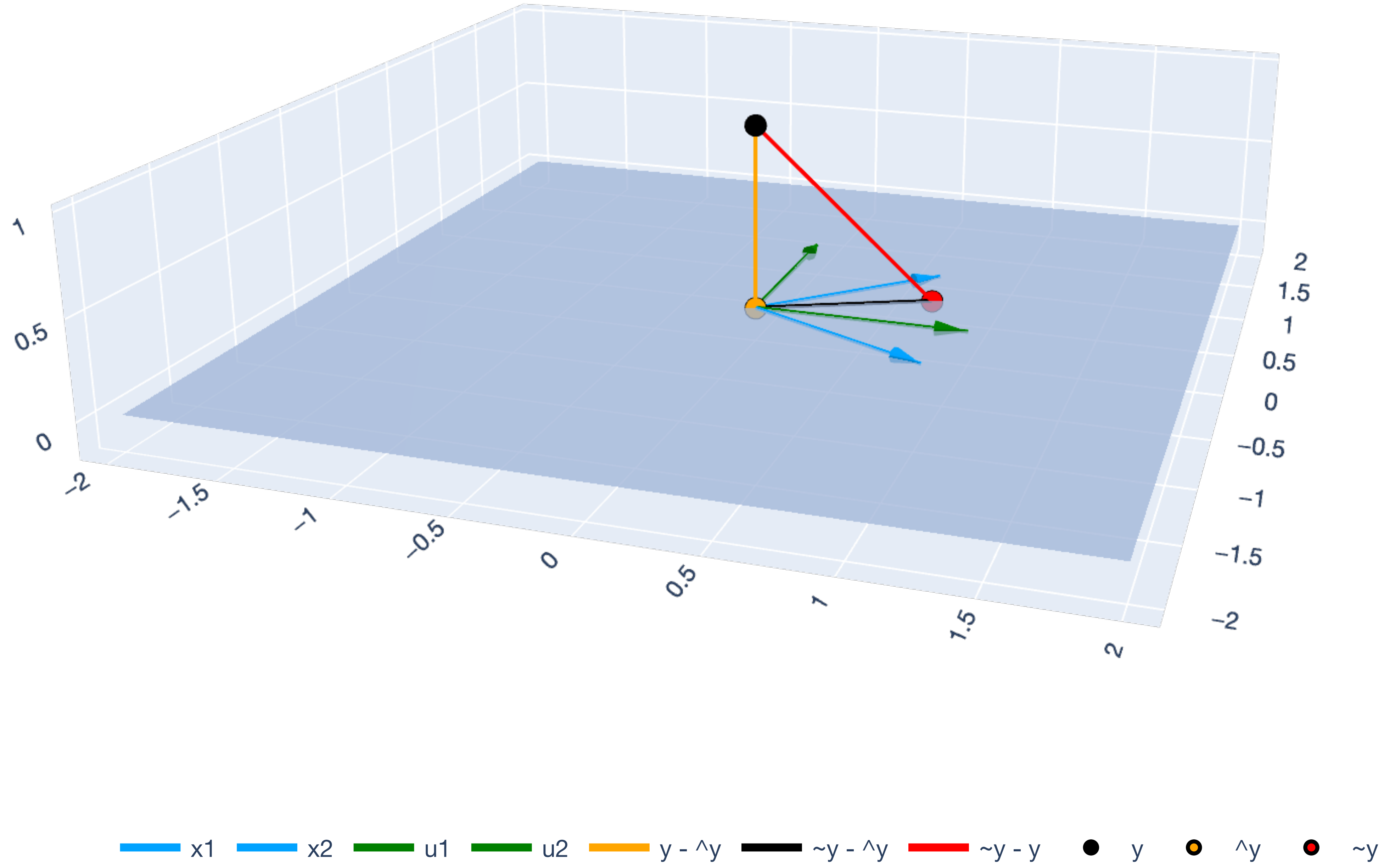
How do we find $\hat{\mathbf{w}}_{onb} \in \mathbb{R}^d$ in $\hat{\mathbf{y}} = \mathbf{U}\hat{\mathbf{w}}_{onb}$? Least squares!

$$\hat{\mathbf{w}}_{onb} = \underset{\hat{\mathbf{w}}_{onb} \in \mathbb{R}^d}{\arg\min} \; \|\mathbf{y} - \mathbf{U}\hat{\mathbf{w}}_{onb}\|^2$$

The columns of $\mathbf{U}$ give an ONB for $\mathscr{X}$…

$$\hat{\mathbf{w}}_{onb} = (\mathbf{U}^\top\mathbf{U})^{-1}\mathbf{U}^\top\mathbf{y}$$

$$= \mathbf{U}^\top\mathbf{y}$$



Orthonormal Basis

Why do we like an orthogonal basis?

Let $\mathscr{X}$ be a subspace. Let $\Pi_{\mathscr{X}}(\mathbf{y}) = \arg\min_{\hat{\mathbf{y}} \in \mathscr{X}} \|\hat{\mathbf{y}} - \mathbf{y}\|^2$ be the projection of $\mathbf{y}$ onto $\mathscr{X}$.

For an arbitrary matrix $\mathbf{X} \in \mathbb{R}^{n \times d}$ with $\text{span}(\text{col}(\mathbf{X})) = \mathscr{X}$,

$$\boxed{\hat{\mathbf{w}} = (\mathbf{X}^\top \mathbf{X})^{-1} \mathbf{X}^\top \mathbf{y} \text{ and } \hat{\mathbf{y}} = \mathbf{X}(\mathbf{X}^\top \mathbf{X})^{-1} \mathbf{X}^\top \mathbf{y}.}$$

OLS

For a *semi-orthogonal matrix* $\mathbf{U} \in \mathbb{R}^{n \times d}$ with $\text{span}(\text{col}(\mathbf{U})) = \mathscr{X}$,

$$\boxed{\hat{\mathbf{w}}_{onb} = \mathbf{U}^\top \mathbf{y} \text{ and } \hat{\mathbf{y}} = \mathbf{U}\mathbf{U}^\top \mathbf{y}.}$$

Much simpler — no inverse operations!

Orthonormal Basis

Why do we like an orthogonal basis?

OLS
$\Updownarrow$
Projection
$\Updownarrow$
OLS II (orthogonal)

Theorem (Projection with orthogonal matrices). Let $\mathcal{X} \subseteq \mathbb{R}^n$ be a subspace and let $\mathbf{u}_1, \ldots, \mathbf{u}_d \in \mathbb{R}^n$ be an orthonormal basis for $\mathcal{X}$, with semi-orthogonal matrix $\mathbf{U} \in \mathbb{R}^{n \times d}$. For any $\mathbf{y} \in \mathbb{R}^n$, the ***projection*** of $\mathbf{y}$ onto $\mathcal{X}$, i.e.

$$\Pi_{\mathcal{X}}(\mathbf{y}) = \arg\min_{\hat{\mathbf{y}} \in \mathcal{X}} \|\hat{\mathbf{y}} - \mathbf{y}\|^2$$

is given by

$$\Pi_{\mathcal{X}}(\mathbf{y}) = \mathbf{U}\mathbf{U}^\top \mathbf{y}.$$

Recap

Lesson Overview

Regression. Fill in gaps from last time: invertibility and Pythagorean theorem.

Subspaces. Subsets of $\mathcal{S} \subseteq \mathbb{R}^n$ where we "stay inside" when performing linear combinations of vectors.

Bases. A "language" to describe all vectors in a subspace.

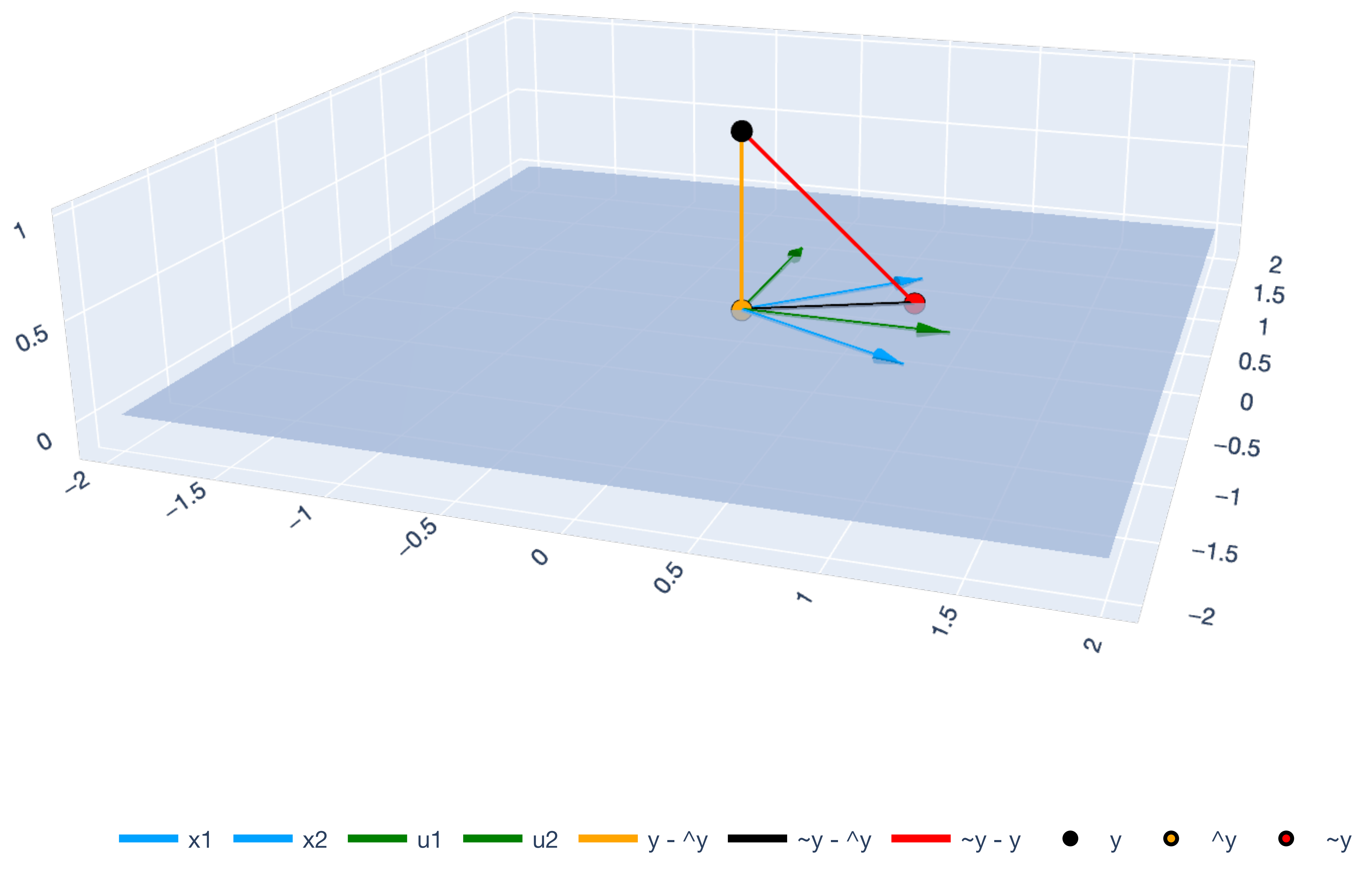
Orthogonality. Orthonormal bases are "good" bases to work with.

Projection. Formal definition of projection and the relationship between projection and least squares.

Least squares with orthonormal bases. If we have an orthonormal basis for $\text{span}(\text{col}(\mathbf{X}))$, least squares becomes much simpler.

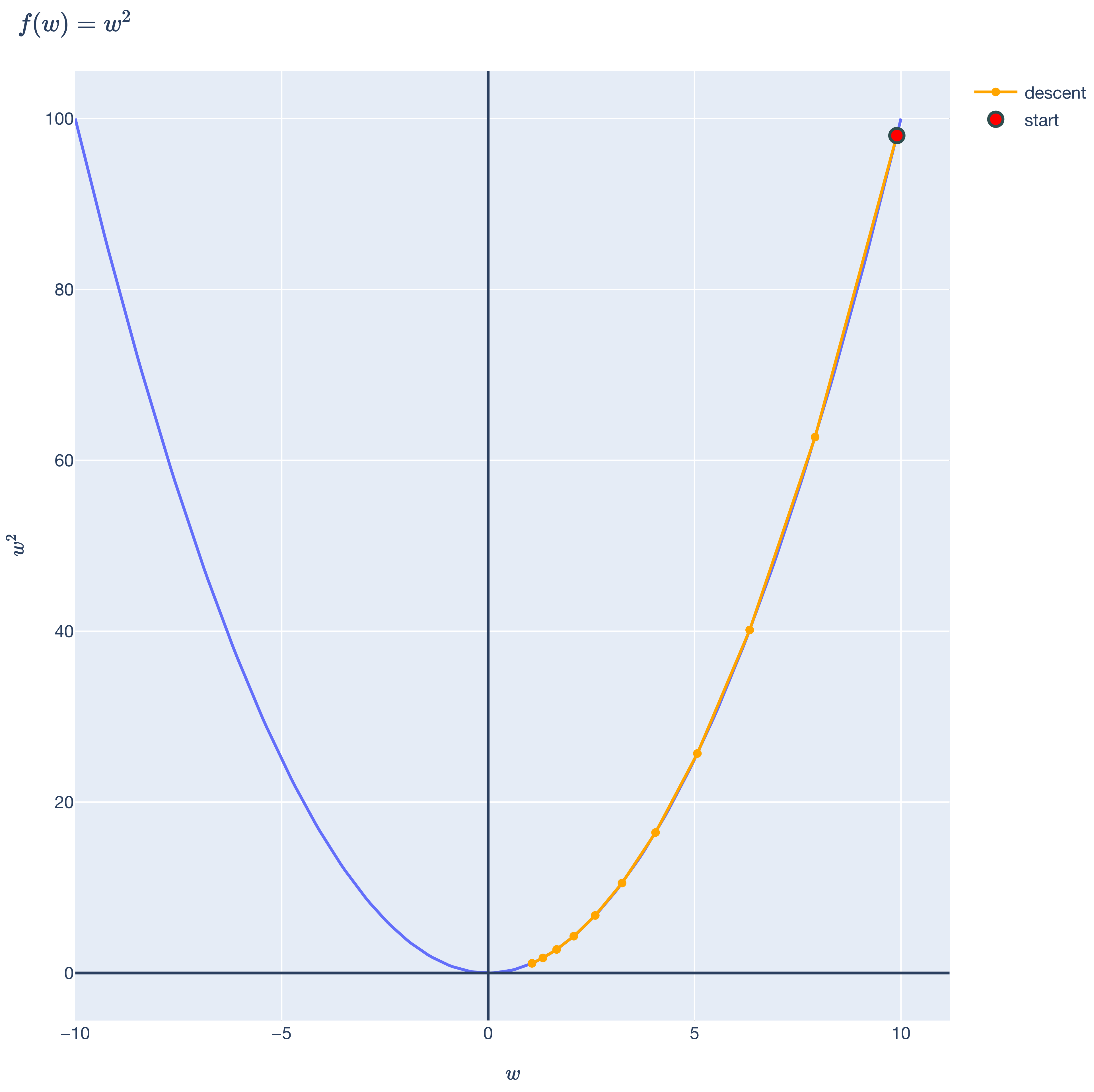
Lesson Overview

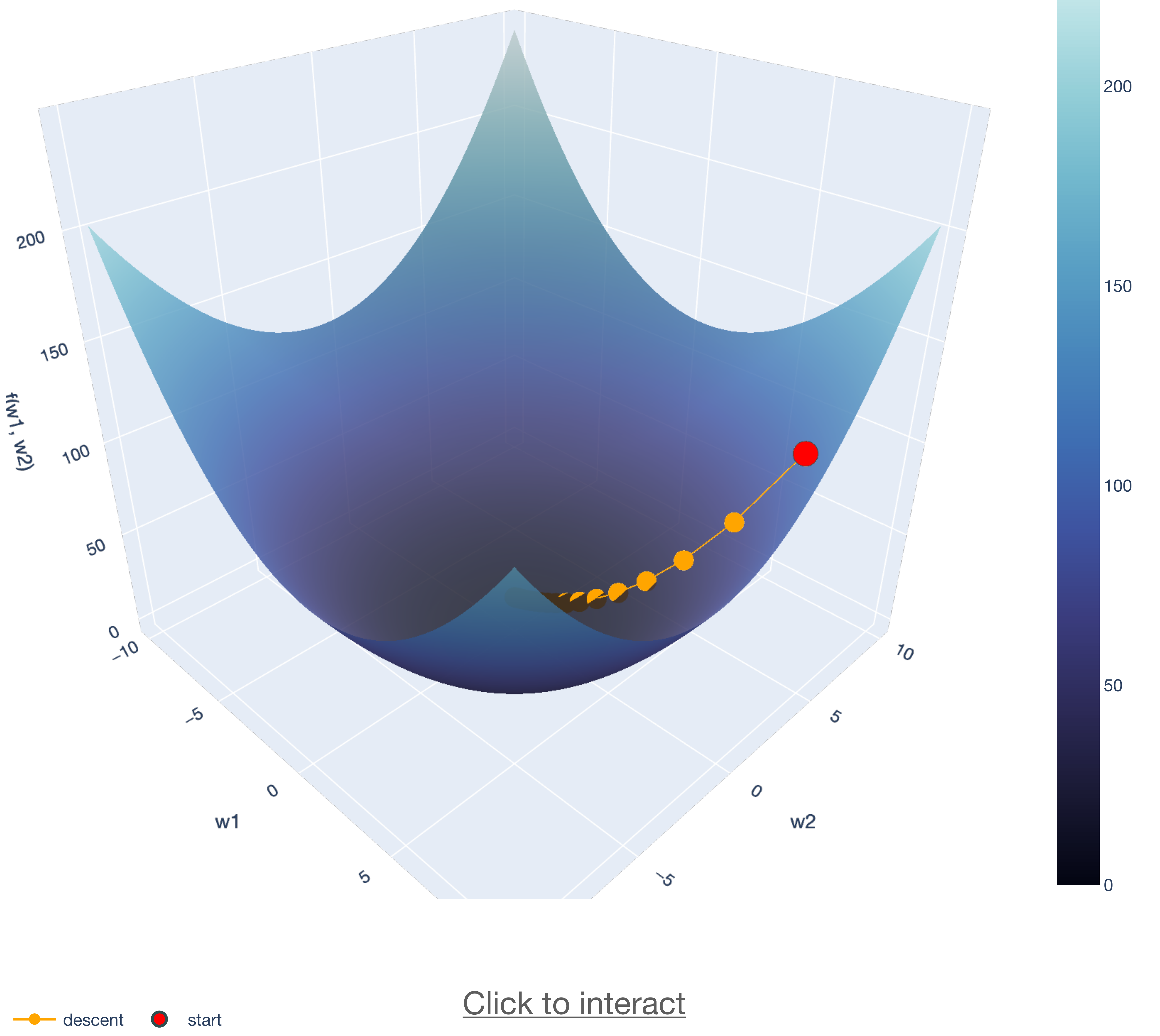
Big Picture: Least Squares



Lesson Overview

Big Picture: Gradient Descent





Click to interact

References

Mathematics for Machine Learning. Marc Pieter Deisenroth, A. Aldo Faisal, Cheng Soon Ong.

Vector Calculus, Linear Algebra, and Differential Forms: A Unified Approach. John H. Hubbard and Barbara Burke Hubbard.

Computational Linear Algebra Lecture Notes: Orthogonality. Daniel Hsu.

Mathematical Foundations for Machine Learning. Rebecca Willett.

Elements of Statistical Learning: Data Mining, Inference, and Prediction. Trevor Hastie, Robert Tibshirani, and Jerome Friedman.