Math for Machine Learning

Week 2.1: Singular Value Decomposition

By: Samuel Deng

Logistics & Announcements

Lesson Overview

Orthogonal complement and properties of projection. We go over several useful properties of the projection operation.

Derivation of the singular value decomposition (SVD). We derive the SVD from the "best-fitting subspace" problem using all the properties of projection.

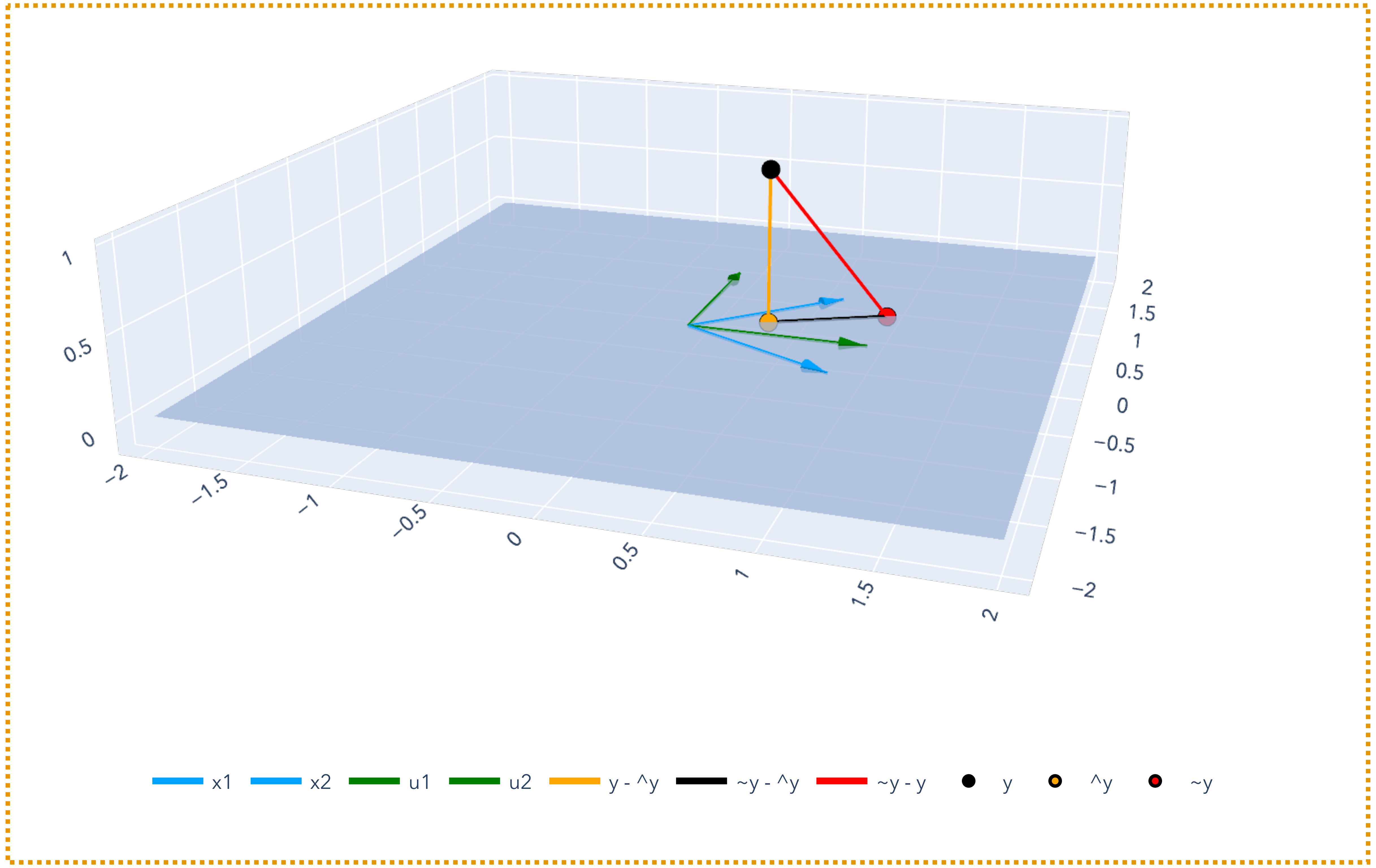
SVD Definition. We go over the definition of SVD and the geometric intuition as the factorization of a data matrix.

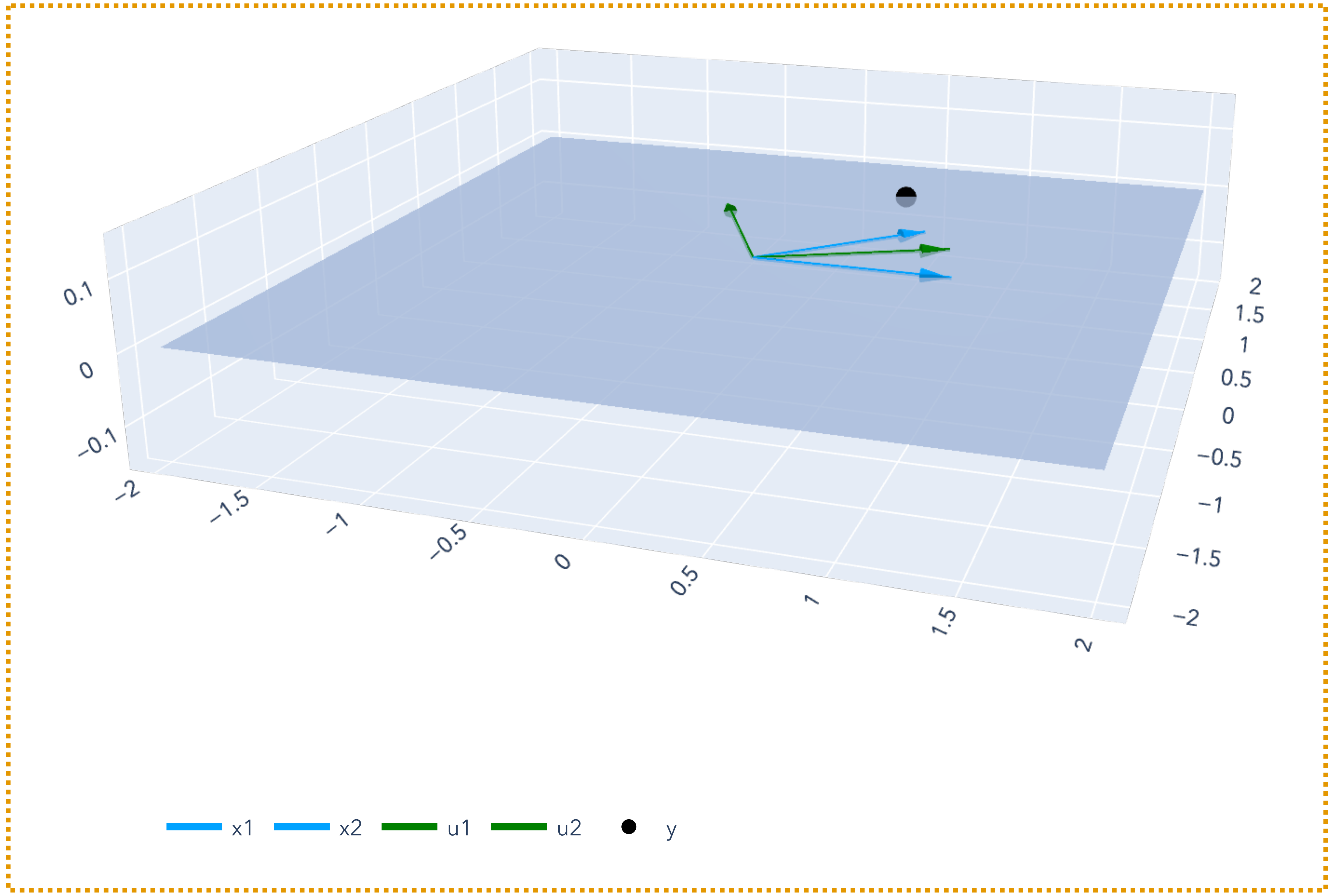
Application of SVD: rank-k approximation. We state and give an example of rank-k approximation, a common data compression technique using SVD.

Pseudoinverse. We unify our OLS solution from the perspective of SVD and the notion of the pseudoinverse, a generalization of inverses to rectangular matrices.

Lesson Overview

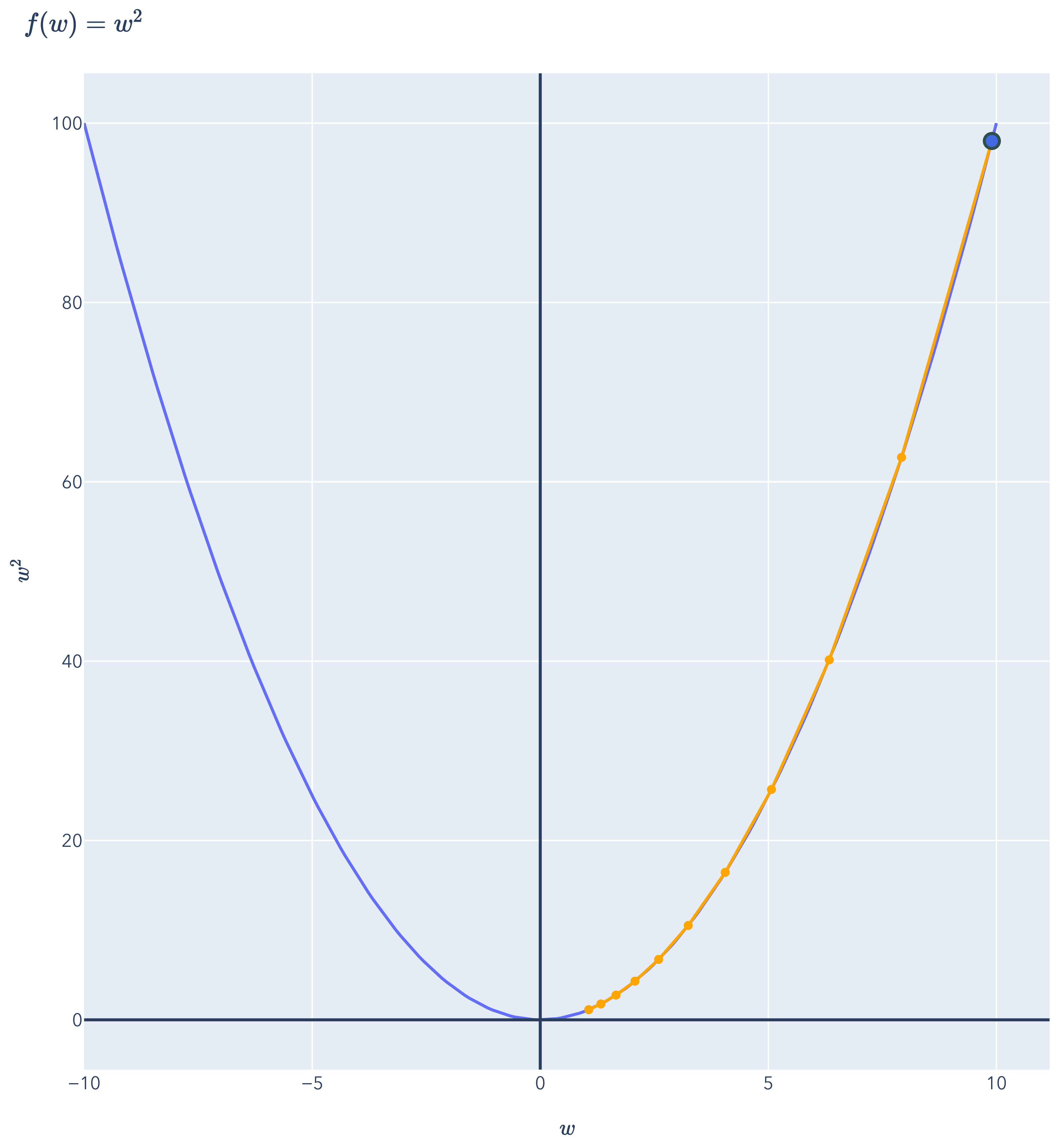
Big Picture: Least Squares

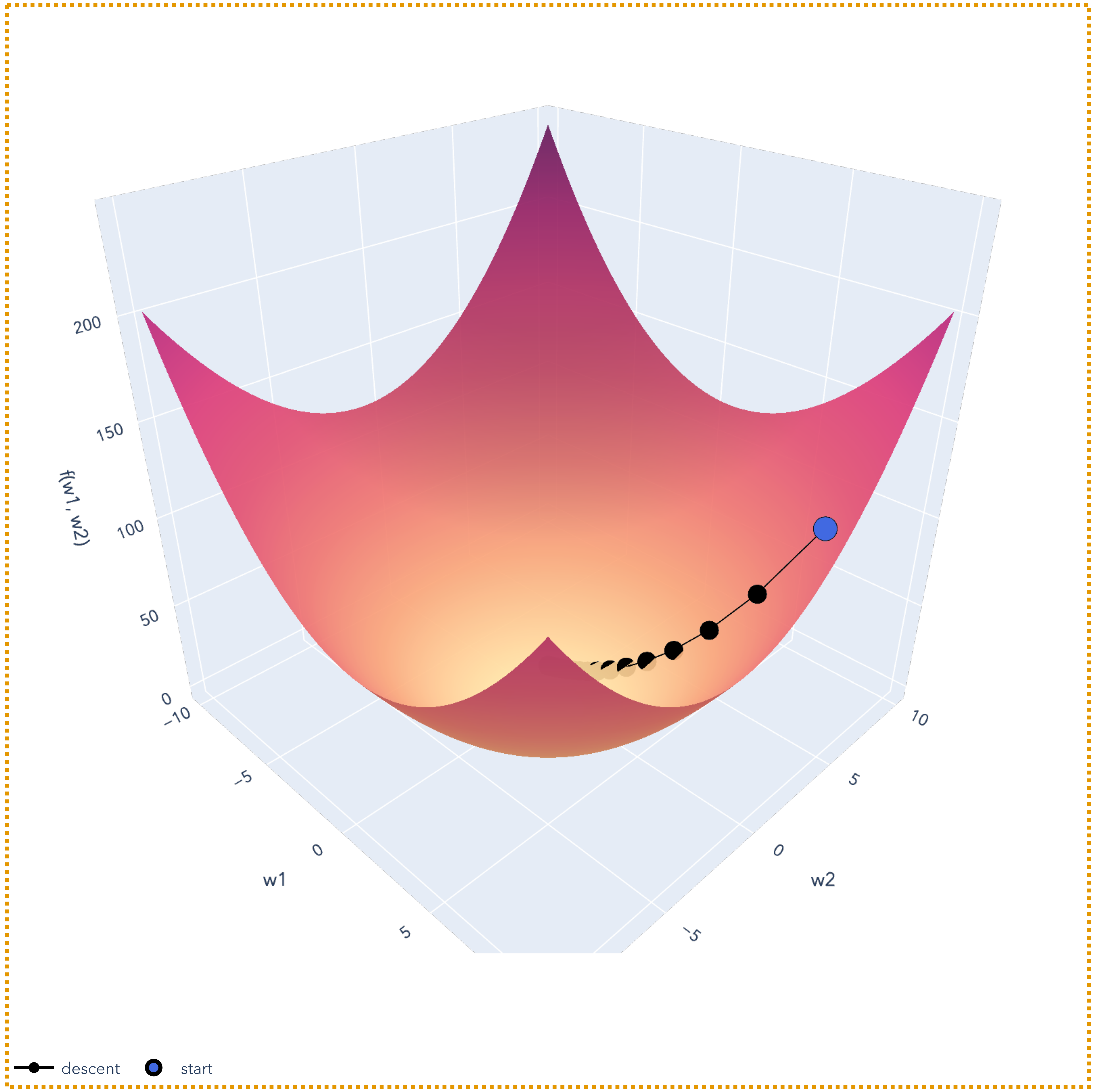




Lesson Overview

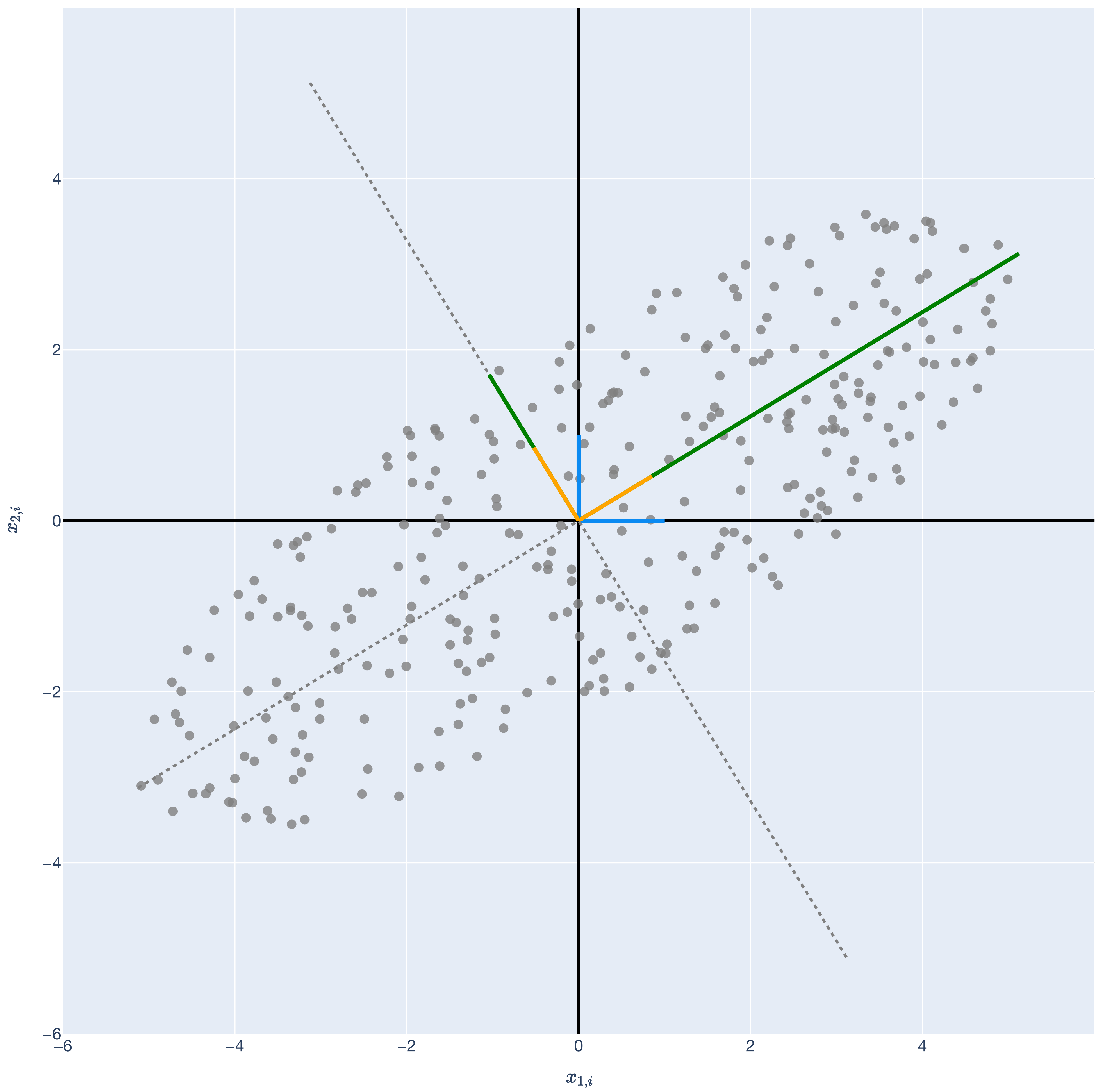
Big Picture: Gradient Descent

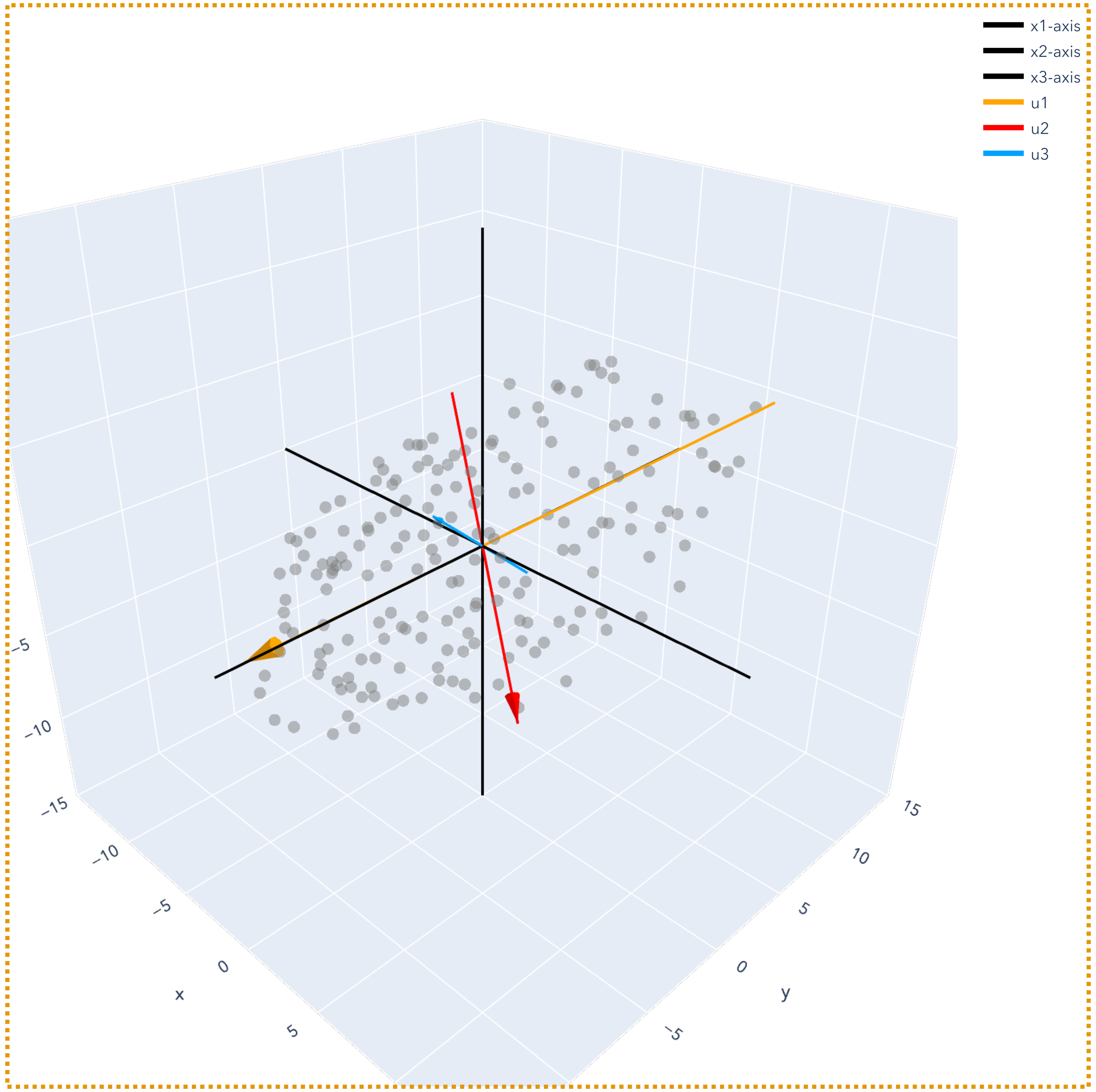




Lesson Overview

Big Picture: Singular Value Decomposition (SVD)





Least Squares

A Quick Review

Regression

Setup (Example View)

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

$$\mathbf{X} = \begin{bmatrix} \leftarrow & \mathbf{x}_1^\top & \rightarrow \\ & \vdots & \\ \leftarrow & \mathbf{x}_n^\top & \rightarrow \end{bmatrix} \quad \mathbf{y} = \begin{bmatrix} y_1 \\ \vdots \\ y_n \end{bmatrix}, \text{ where } \mathbf{x}_1, \ldots, \mathbf{x}_n \in \mathbb{R}^d.$$

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

<u>Goal:</u> 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:

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

Regression

Setup (Feature View)

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

$$\mathbf{X} = \begin{bmatrix} \uparrow & & \uparrow \\ \mathbf{x}_1 & \cdots & \mathbf{x}_d \\ \downarrow & & \downarrow \end{bmatrix} \quad \mathbf{y} = \begin{bmatrix} y_1 \\ \vdots \\ y_n \end{bmatrix}, \text{ where } \mathbf{x}_1, \ldots, \mathbf{x}_d \in \mathbb{R}^n.$$

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

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:

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

Regression

Setup

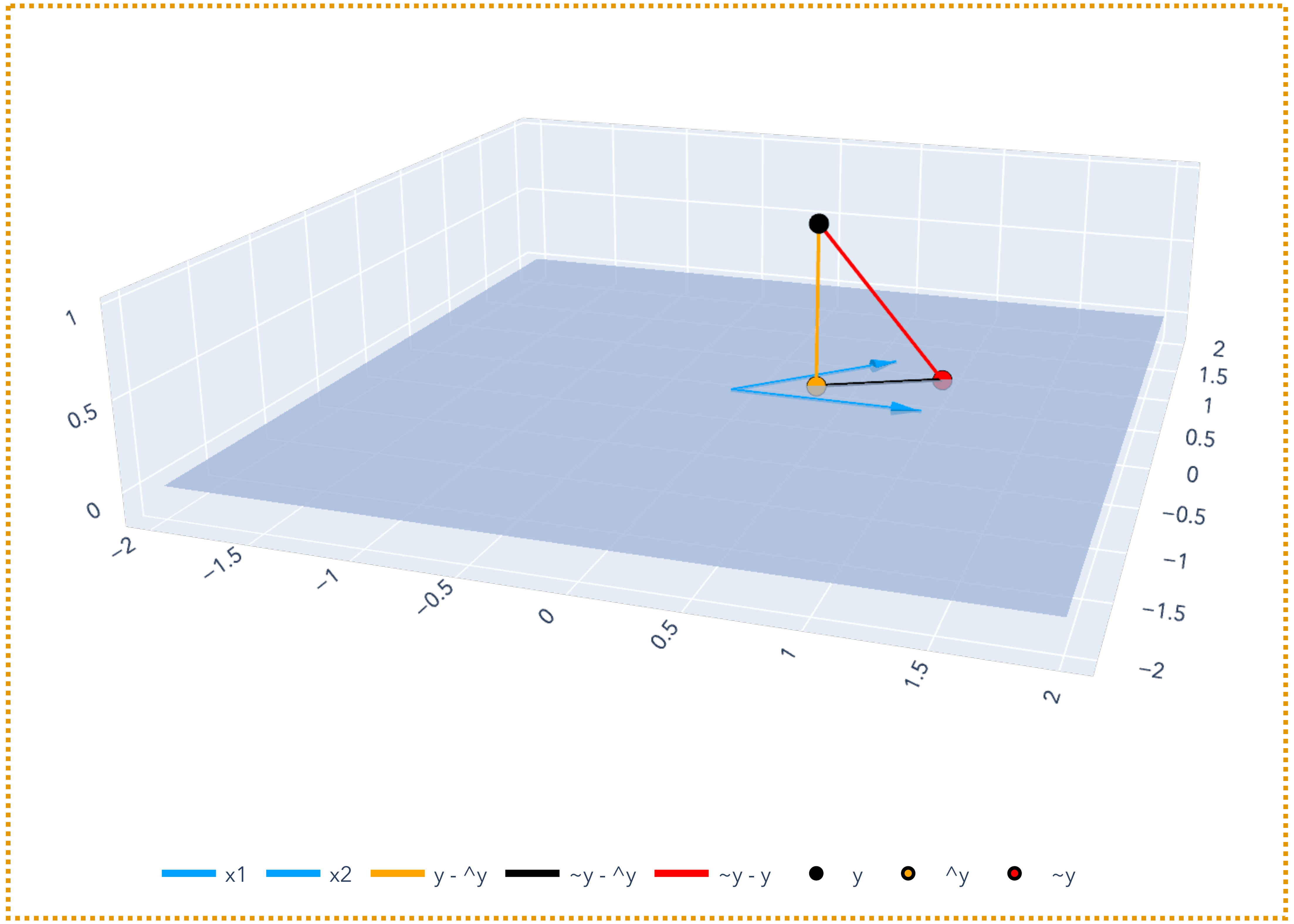
To find $\hat{\mathbf{w}}$, we follow the *principle of least squares.*

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

This gives the predictions $\hat{\mathbf{y}} \in \mathbb{R}^n$ that are close in a least squares sense:

$$\hat{\mathbf{y}} = \mathbf{X}\hat{\mathbf{w}} \text{ such that } \|\hat{\mathbf{y}} - \mathbf{y}\|^2 \leq \|\tilde{\mathbf{y}} - \mathbf{y}\|^2$$

(for $\tilde{\mathbf{y}} = \mathbf{X}\mathbf{w}$ from any other $\mathbf{w} \in \mathbb{R}^d$).



Least Squares

OLS Theorem

<u>Theorem (Ordinary Least Squares).</u> Let $\mathbf{X} \in \mathbb{R}^{n \times d}$ and $\mathbf{y} \in \mathbb{R}^n$. Let $\hat{\mathbf{w}} \in \mathbb{R}^d$ be the least squares minimizer:

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

If $n \geq d$ and $\operatorname{rank}(\mathbf{X}) = d$, then:

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

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

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

Least Squares

OLS with Orthogonal Basis

<u>Theorem (OLS with orthogonal basis).</u> 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 $\mathbf{U} \in \mathbb{R}^{n \times d}$. Let $\mathbf{y} \in \mathbb{R}^n$ and let $\hat{\mathbf{w}} \in \mathbb{R}^d$ be the least squares minimizer:

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

which is solved by:

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

Additionally, the projection $\hat{\mathbf{y}} \in \mathbb{R}^n$ is given by $\Pi_{\mathcal{X}}(\mathbf{y}) = \arg\min_{\hat{\mathbf{y}} \in \mathcal{X}} \|\hat{\mathbf{y}} - \mathbf{y}\|^2$:

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

Least Squares

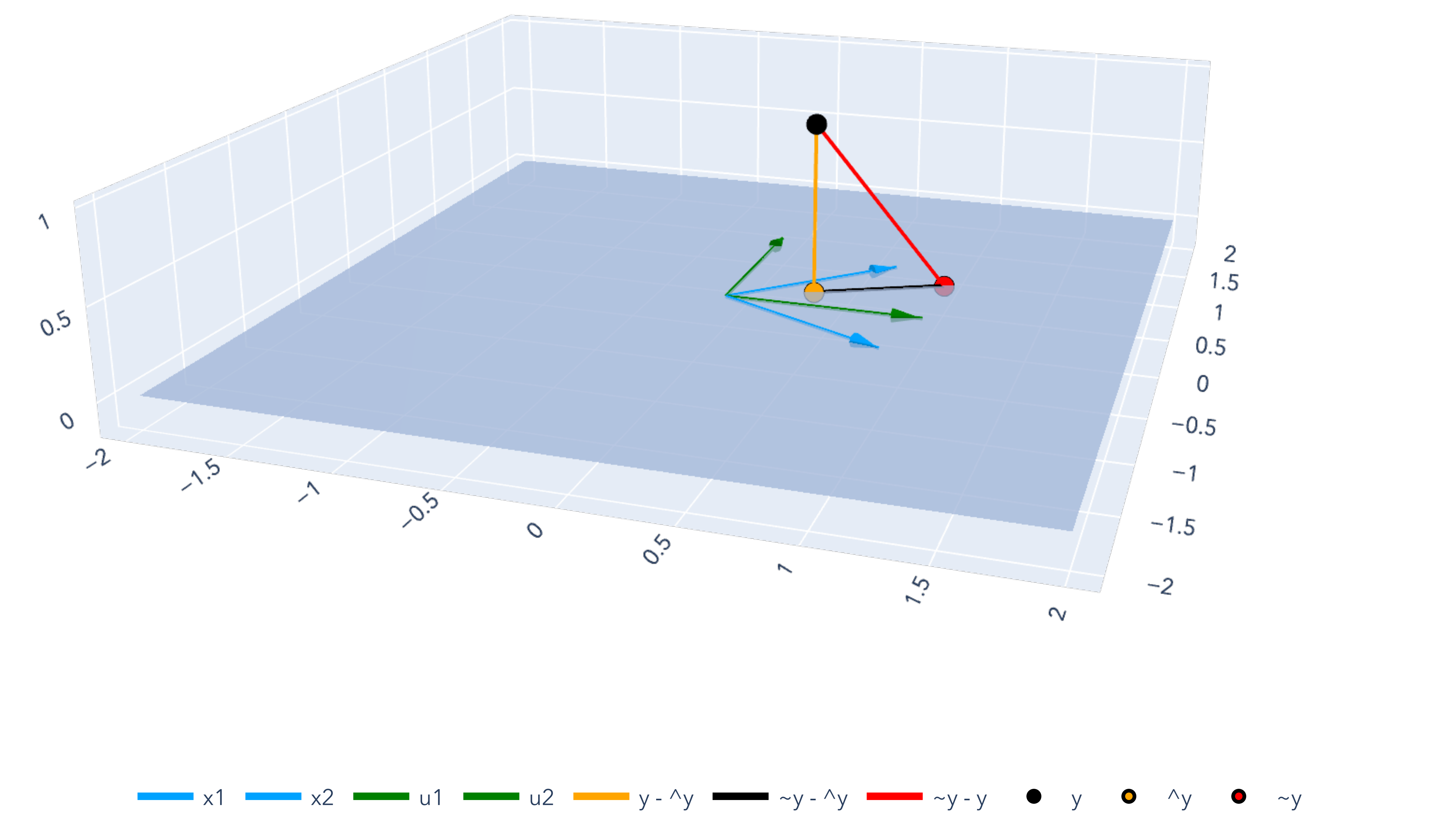
OLS with Orthogonal Basis

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

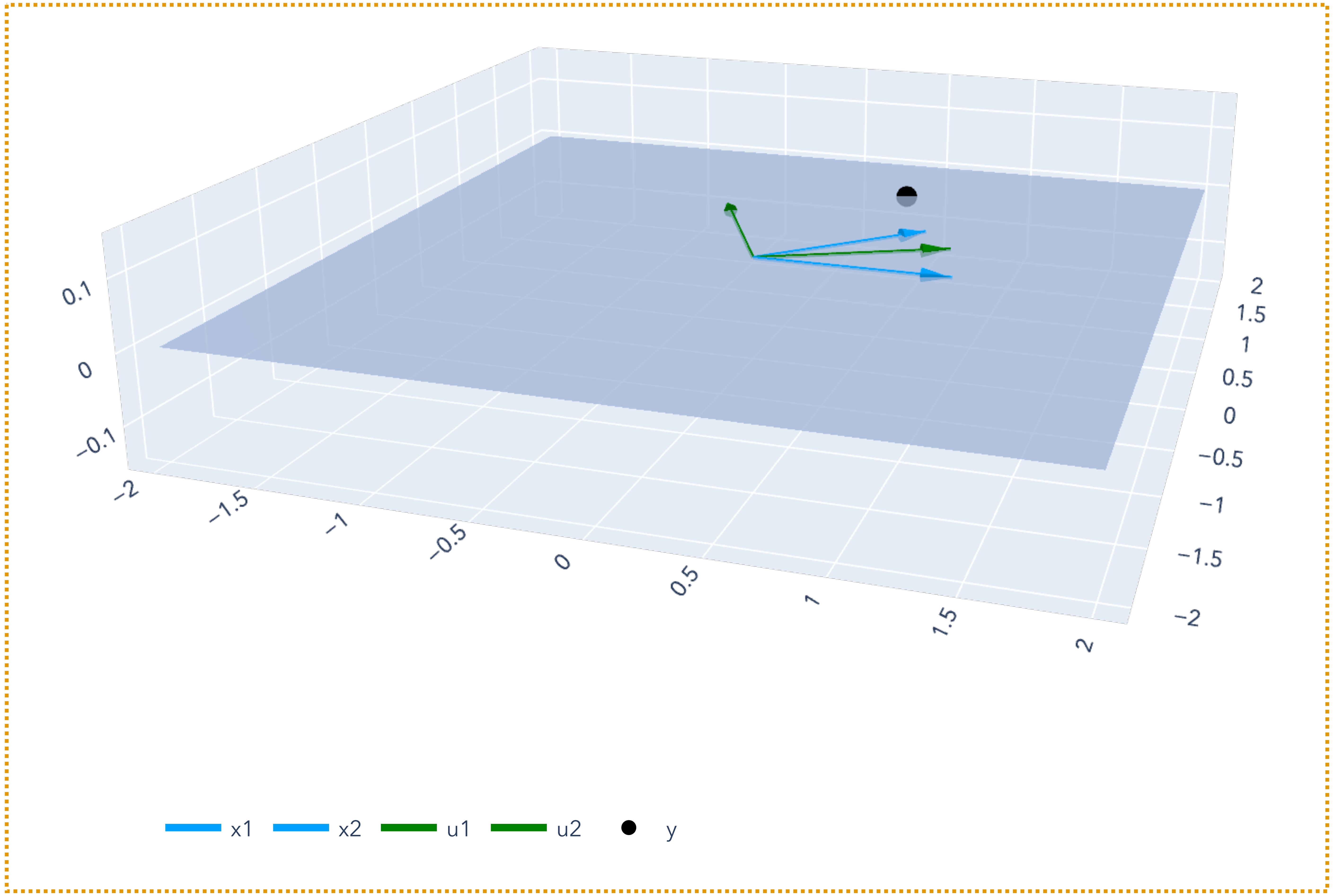
$$\hat{\mathbf{w}}_{onb} = \mathbf{U}^\top\mathbf{y}$$

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

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



How to find a good orthogonal basis?



Properties of Projections

Projection Matrices and Orthogonal Complement

Projection

Projection of a vector onto a subspace

For a subspace $\mathscr{X} \subseteq \mathbb{R}^n$, the projection of a vector $\mathbf{y} \in \mathbb{R}^n$ onto $\mathscr{X}$ is the closest vector $\hat{\mathbf{y}}$ in $\mathscr{X}$ to $\mathbf{y}$, in a Euclidean distance sense:

$$\hat{\mathbf{y}} = \underset{\hat{\mathbf{y}} \in \mathscr{X}}{\arg\min} \ \|\hat{\mathbf{y}} - \mathbf{y}\| = \|\hat{\mathbf{y}} - \mathbf{y}\|^2.$$

Let $\mathscr{X} = \mathrm{CS}(\mathbf{X})$. *Any* point $\hat{\mathbf{y}} \in \mathscr{X}$ is a linear combination $\hat{\mathbf{y}} = \mathbf{X}\hat{\mathbf{w}}$, with:

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

Least Squares as Projection

Projection Matrix

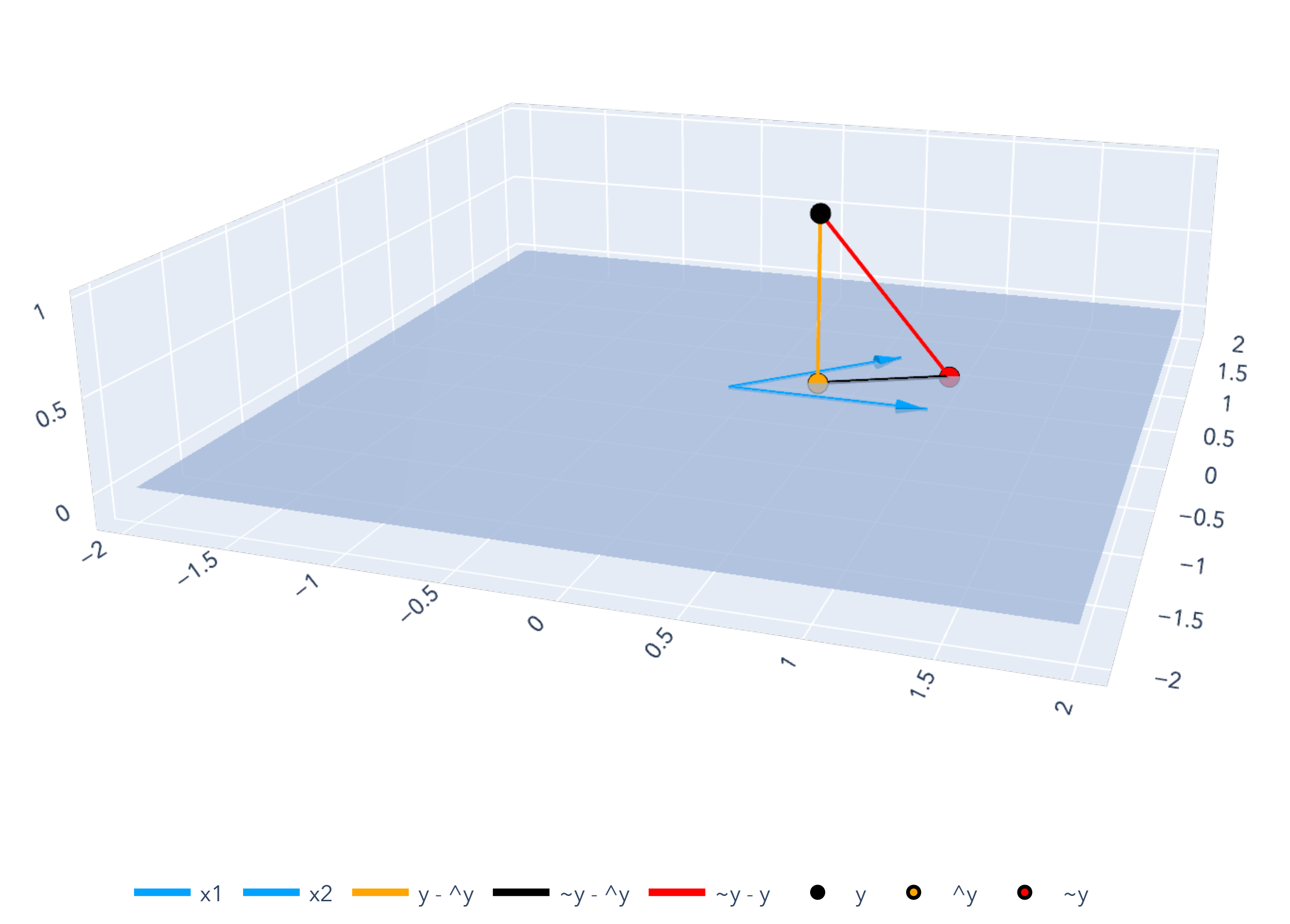
$$\hat{\mathbf{w}} = \arg\min_{\hat{\mathbf{w}} \in \mathbb{R}^d} \|\mathbf{X}\hat{\mathbf{w}} - \mathbf{y}\|^2$$

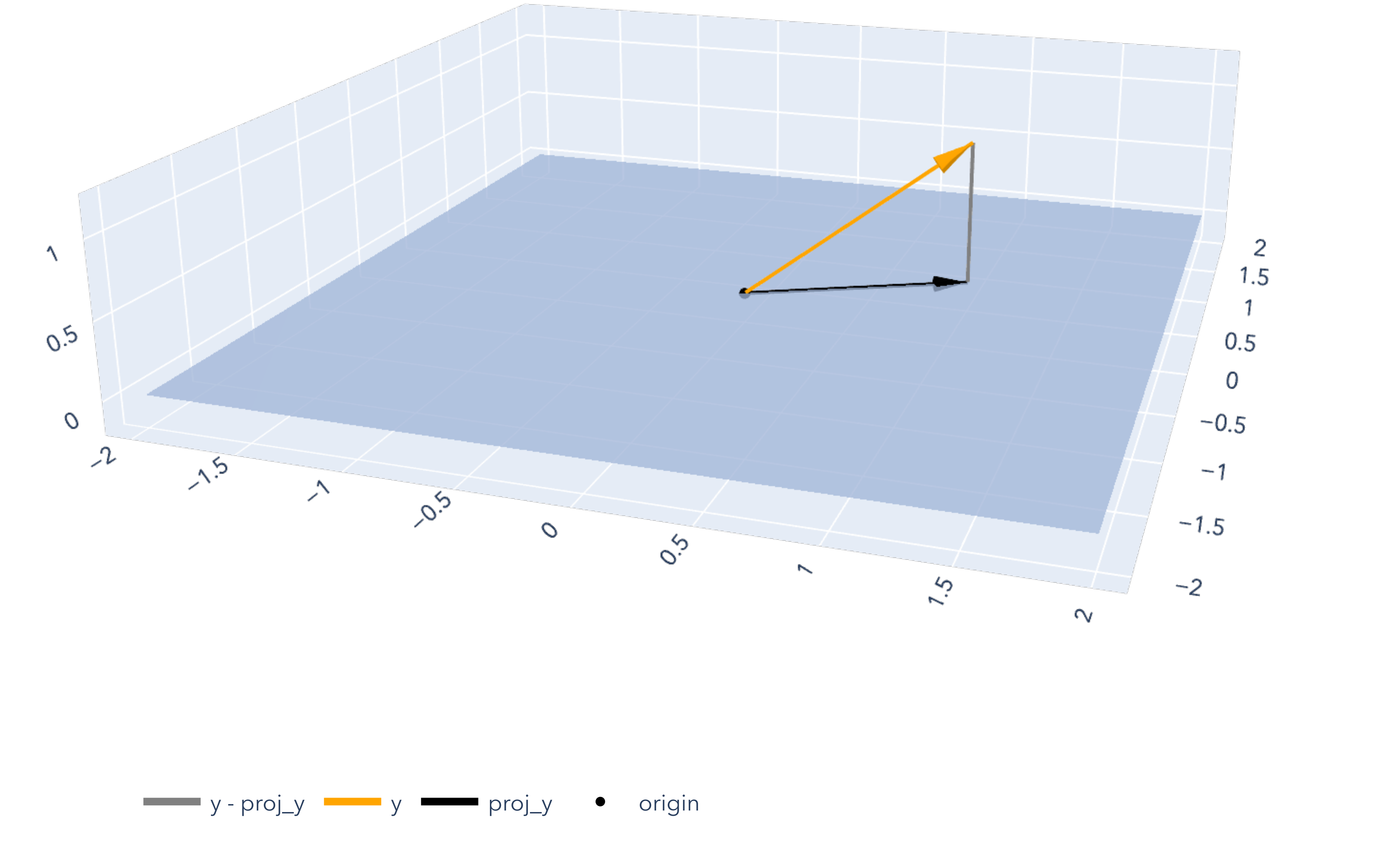
This is just least squares! By what we've learned…

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

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

The projection matrix is: $P_{\mathcal{X}} = \mathbf{X}(\mathbf{X}^\top \mathbf{X})^{-1} \mathbf{X}^\top \in \mathbb{R}^{n \times n}$.





Least Squares as Projection

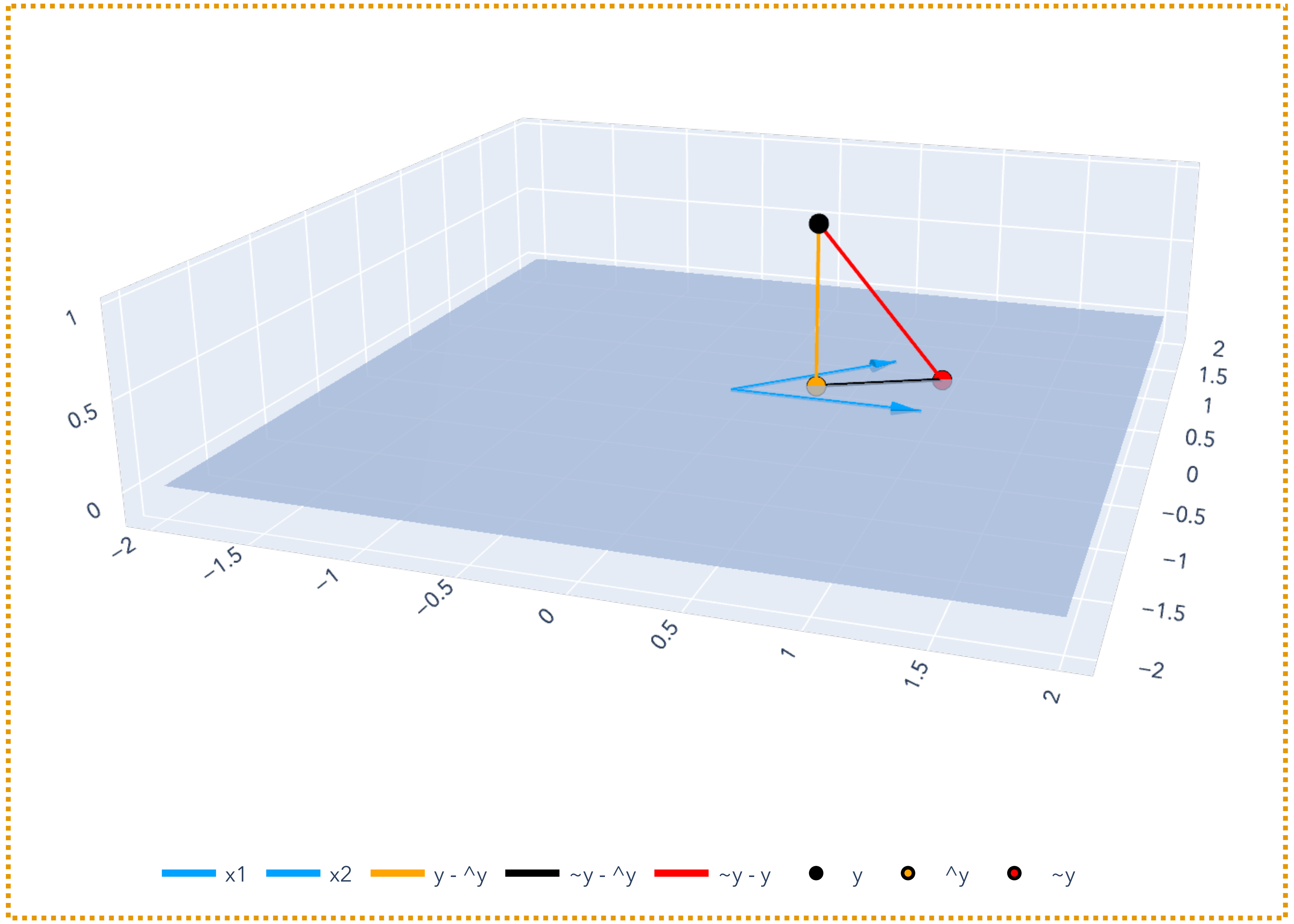
Projection Matrix

Any matrix $\mathbf{X} \in \mathbb{R}^{n \times d}$ has a subspace $\mathcal{X} = \mathrm{CS}(\mathbf{X})$.

If the columns $\mathbf{x}_1, \ldots, \mathbf{x}_d$ are *linearly independent,* then:

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

where $P_{\mathcal{X}} \in \mathbb{R}^{n \times n}$ is a projection matrix.



What else can we say about projections?

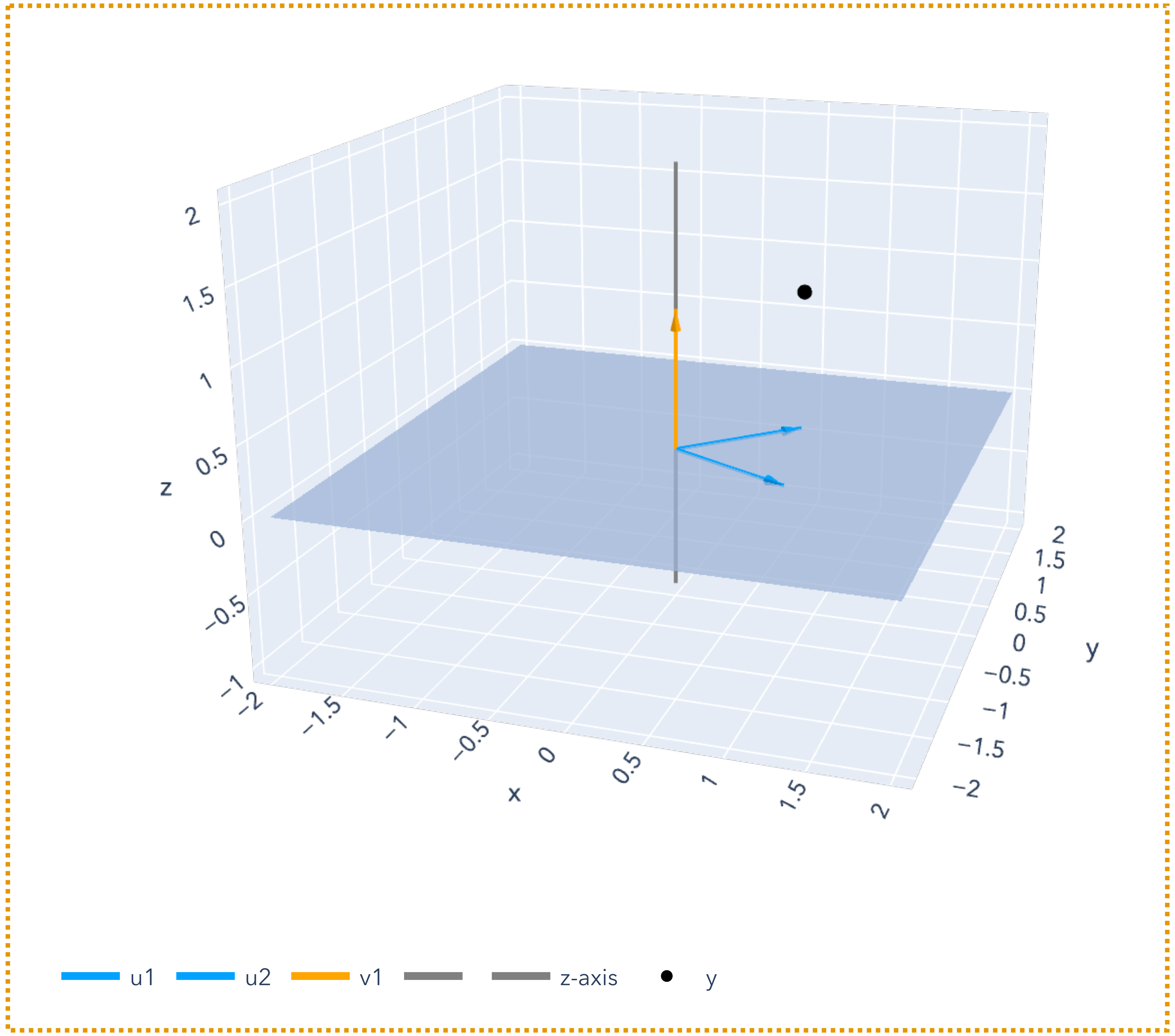
Orthogonal Complement

Intuition

Any subspace $A \subseteq \mathbb{R}^n$ has an orthogonal complement $A^{\perp}$.

All vectors in A are orthogonal to all the vectors in $A^{\perp}$, and vice versa.

Any vector $\mathbf{y} \in \mathbb{R}^n$ can be constructed by adding a vector from A to a vector from $A^{\perp}$.

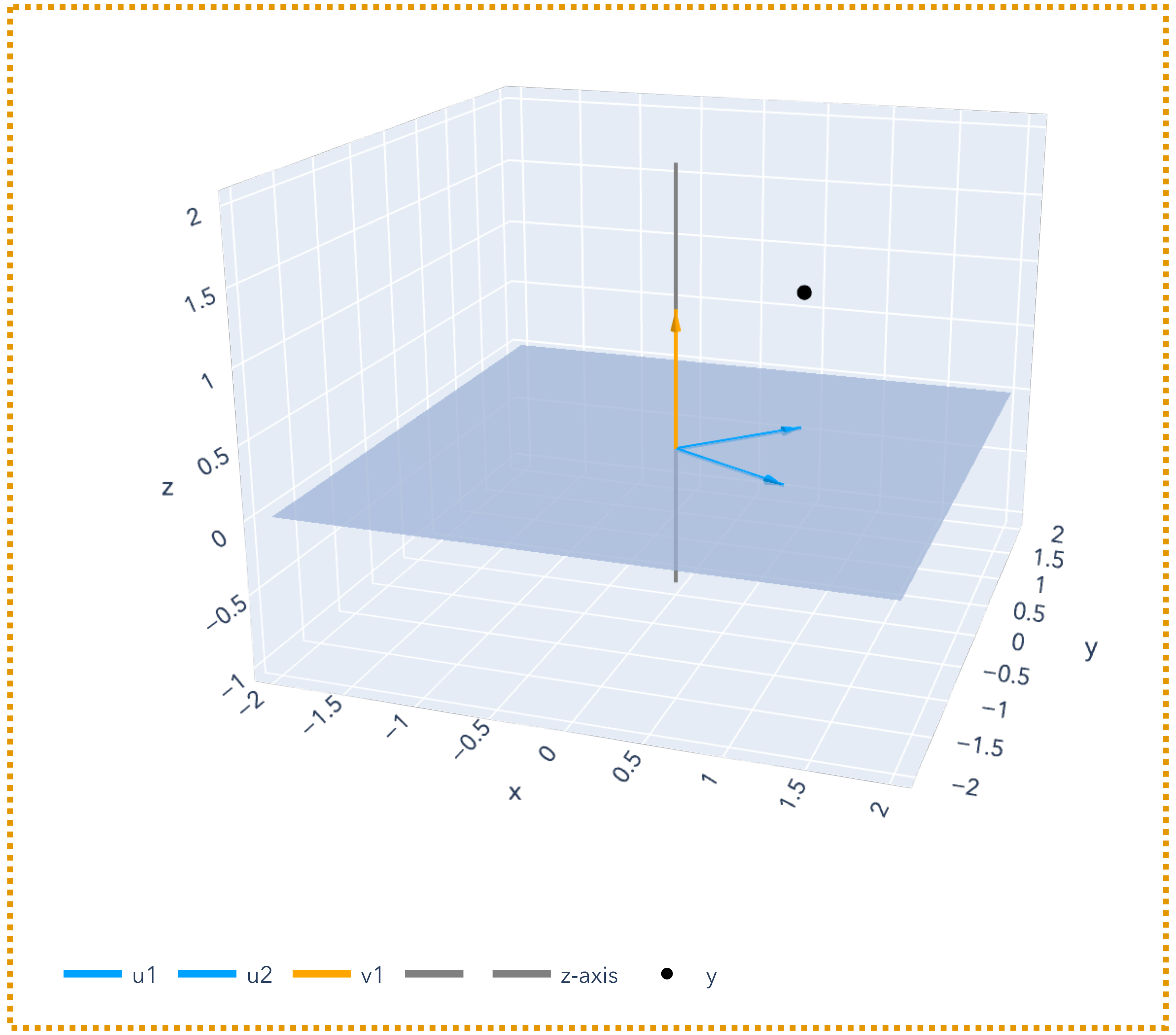


Orthogonal Complement

Definition

Let $A \subseteq \mathbb{R}^n$ be a subspace. The orthogonal complement of A, written $A^\perp$, is the set of vectors

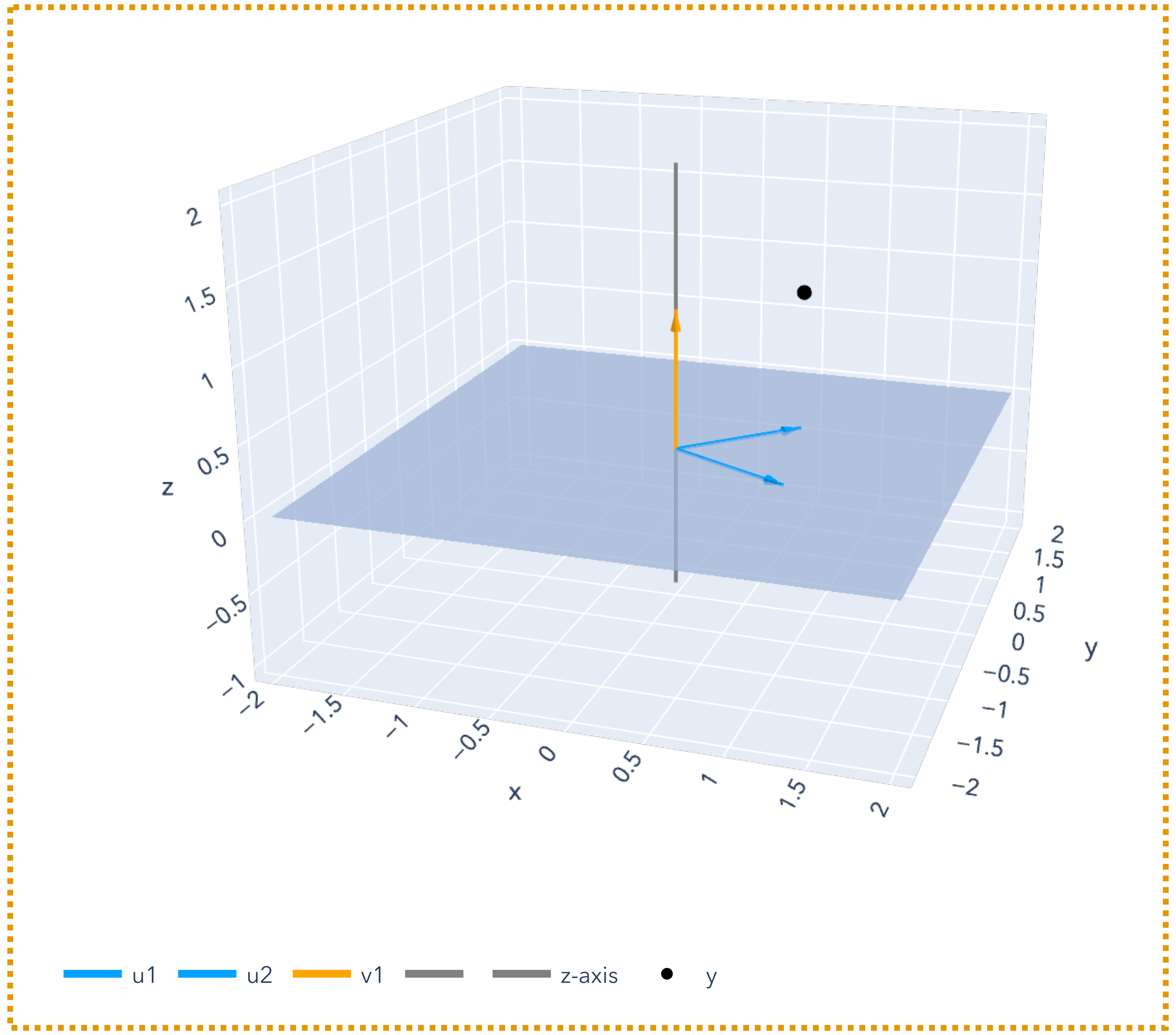
$$A^\perp := \{\mathbf{v} \in \mathbb{R}^n : \langle \mathbf{v}, \mathbf{u} \rangle = 0 \text{ for all } \mathbf{u} \in A\}.$$



Orthogonal Complement

Dimension

For any subspace $A \subseteq \mathbb{R}^n$ with $\dim(A) = d$, orthogonal complement $A^\perp$ has $\dim(A^\perp) = n - d$.



Orthogonal Complement

Orthogonal Complement and Matrices

Let $\mathbf{a}_1, \ldots, \mathbf{a}_d \in \mathbb{R}^n$ be a basis for the subspace $A \subseteq \mathbb{R}^n$.

Let $\mathbf{b}_1, \ldots, \mathbf{b}_{n-d}$ be a basis for the orthogonal complement, $A^\perp$.

Let $\mathbf{A} \in \mathbb{R}^{n\times d}$ have columns $\mathbf{a}_1, \ldots, \mathbf{a}_d$. Let $\mathbf{B} \in \mathbb{R}^{n\times(n-d)}$ have columns $\mathbf{b}_1, \ldots, \mathbf{b}_{n-d}$. Then:

$$\mathbf{A}^\top\mathbf{B} = \mathbf{0} \text{ and } \mathbf{B}^\top\mathbf{A} = \mathbf{0}.$$

We can break down any vector $\mathbf{x} \in \mathbb{R}^n$ into two projections:

$$\mathbf{x} = P_{\mathbf{A}}\mathbf{x} + P_{\mathbf{B}}\mathbf{x}.$$

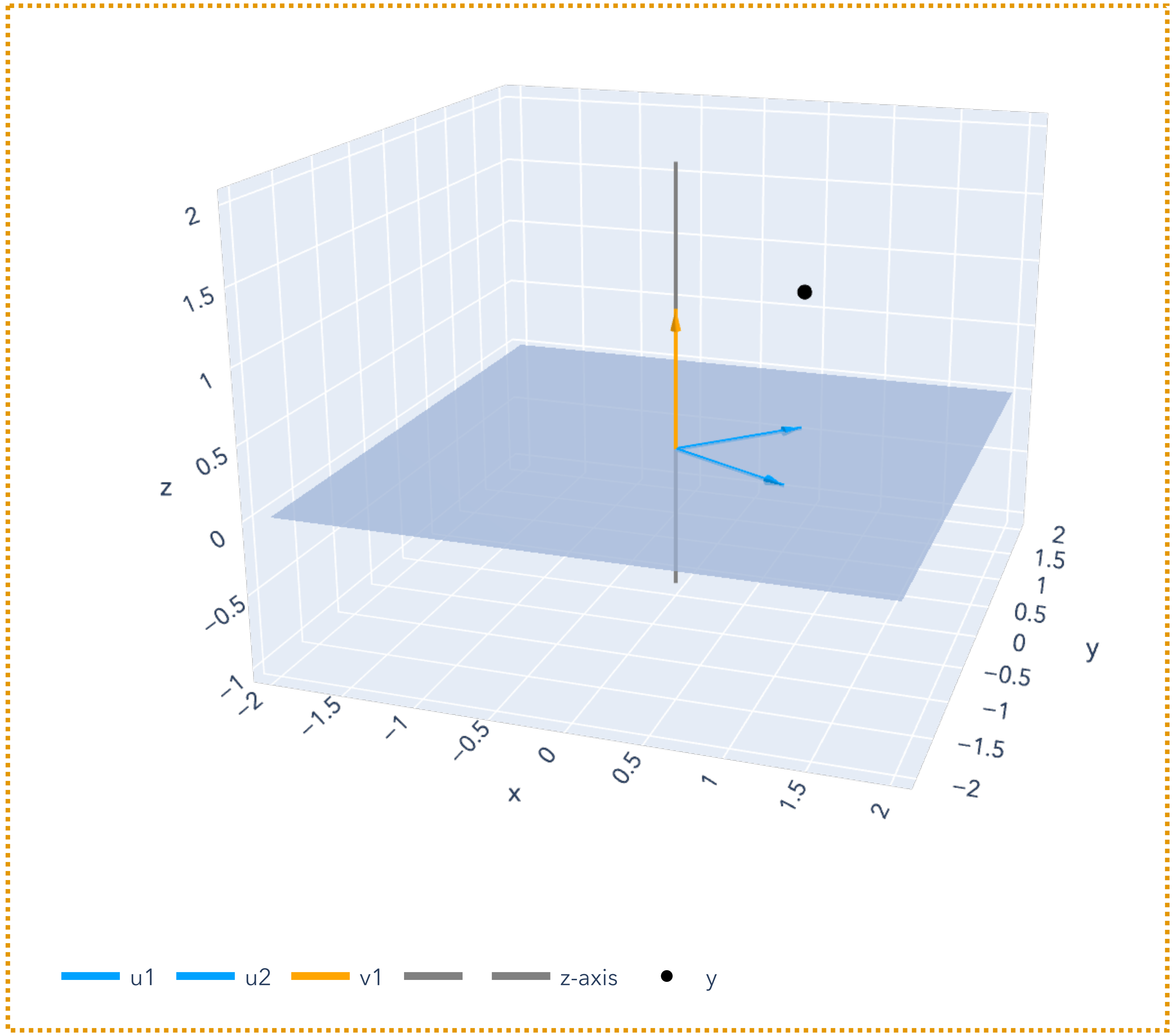
Orthogonal Complement

Orthogonal Complement and Projections

We can break down any vector $\mathbf{x} \in \mathbb{R}^n$ into two projections:

$$\mathbf{x} = P_{\mathbf{A}}\mathbf{x} + P_{\mathbf{B}}\mathbf{x}.$$

Additionally, $\mathbf{I} = P_{\mathbf{A}} + P_{\mathbf{B}}$.



Projection Matrices

Properties

$\mathbf{A} \in \mathbb{R}^{n \times d}$ has columnspace $\mathrm{CS}(\mathbf{A})$; $\mathbf{B} \in \mathbb{R}^{n \times (n-d)}$ has columns $\mathbf{b}_1, \ldots, \mathbf{b}_{n-d}$, a basis for $\mathrm{CS}(\mathbf{A})^\perp$.

<u>Prop (Orthogonal Decomposition).</u> For any vector $\mathbf{x} \in \mathbb{R}^n$, $\mathbf{x} = P_\mathbf{A}\mathbf{x} + P_\mathbf{B}\mathbf{x}$.

<u>Prop (Projection and Orthogonal Complement Matrices).</u> $P_\mathbf{A} + P_\mathbf{B} = \mathbf{I}$.

<u>Prop (Projecting twice doesn't do anything).</u> $P_\mathbf{A} = P_\mathbf{A} P_\mathbf{A} = P_\mathbf{A}^2$.

<u>Prop (Projections are symmetric).</u> $P_\mathbf{A} = P_\mathbf{A}^\top$.

<u>Prop (1D projection formula).</u> For the 1D subspace associated with $\mathbf{a} \in \mathbb{R}^n$: $P_\mathbf{a} = \dfrac{\mathbf{a}\mathbf{a}^\top}{\mathbf{a}^\top\mathbf{a}}$.

Singular Value Decomposition

1D Intuition and Derivation

Singular Value Decomposition (SVD)

1D Picture

<u>Observed:</u> Matrix of *training samples* $\mathbf{X} \in \mathbb{R}^{n \times d}$ (forget about *training labels* $\mathbf{y} \in \mathbb{R}^n$ for now).

$$\mathbf{X} = \begin{bmatrix} \uparrow & & \uparrow \\ \mathbf{x}_1 & \cdots & \mathbf{x}_d \\ \downarrow & & \downarrow \end{bmatrix}, \text{ where } \mathbf{x}_1, \ldots, \mathbf{x}_d \in \mathbb{R}^n.$$

<u>Goal:</u> Find the best one-dimensional subspace $\mathscr{U} \subseteq \mathbb{R}^n$ that fits the points.

A one-dimensional subspace is determined by a single vector $\mathbf{u} \in \mathbb{R}^n$:

$$\mathscr{U} = \{c\mathbf{u} : c \in \mathbb{R}\}.$$

Singular Value Decomposition (SVD)

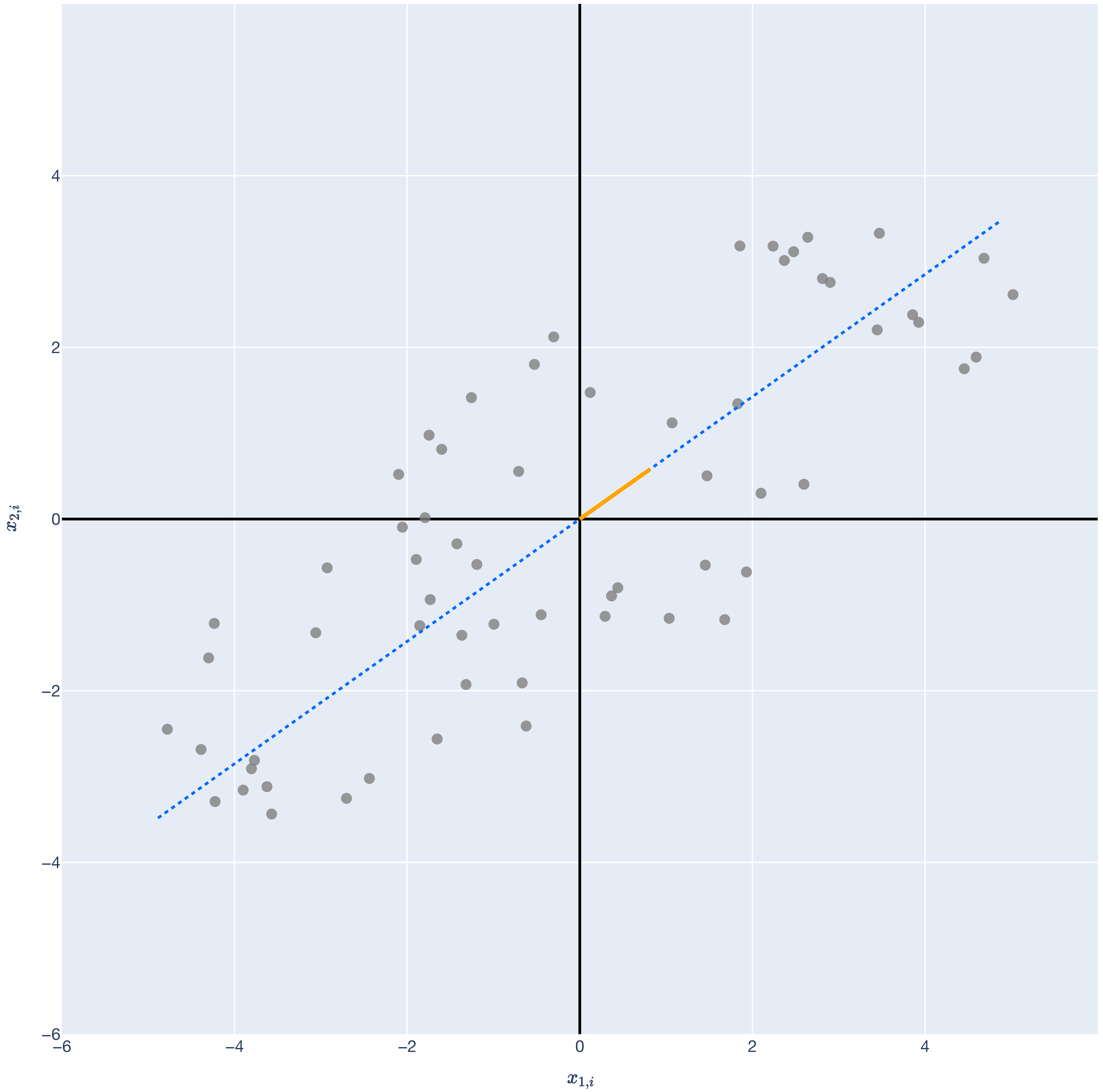
1D Picture

Observe data $\mathbf{x}_1, \ldots, \mathbf{x}_d \in \mathbb{R}^n$.

<u>**Goal:**</u> Find the best one-dimensional subspace $\mathcal{U} \subseteq \mathbb{R}^n$ that fits the points.

How? Find $\mathbf{u} \in \mathbb{R}^n$ that minimizes the sum of squared projection distances:

$$\arg\min_{\mathbf{u} \in \mathbb{R}^n} \sum_{i=1}^{d} \|\mathbf{x}_i - \Pi_{\mathbf{u}}(\mathbf{x}_i)\|^2.$$



Comparison with OLS

1D Pictures

OLS: Find best linear combination $\hat{\mathbf{w}} \in \mathbb{R}^d$ of $\mathbf{x}_1, \ldots, \mathbf{x}_d$ such that

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

Important: there is no $\mathbf{y}$ in our BFS problem!

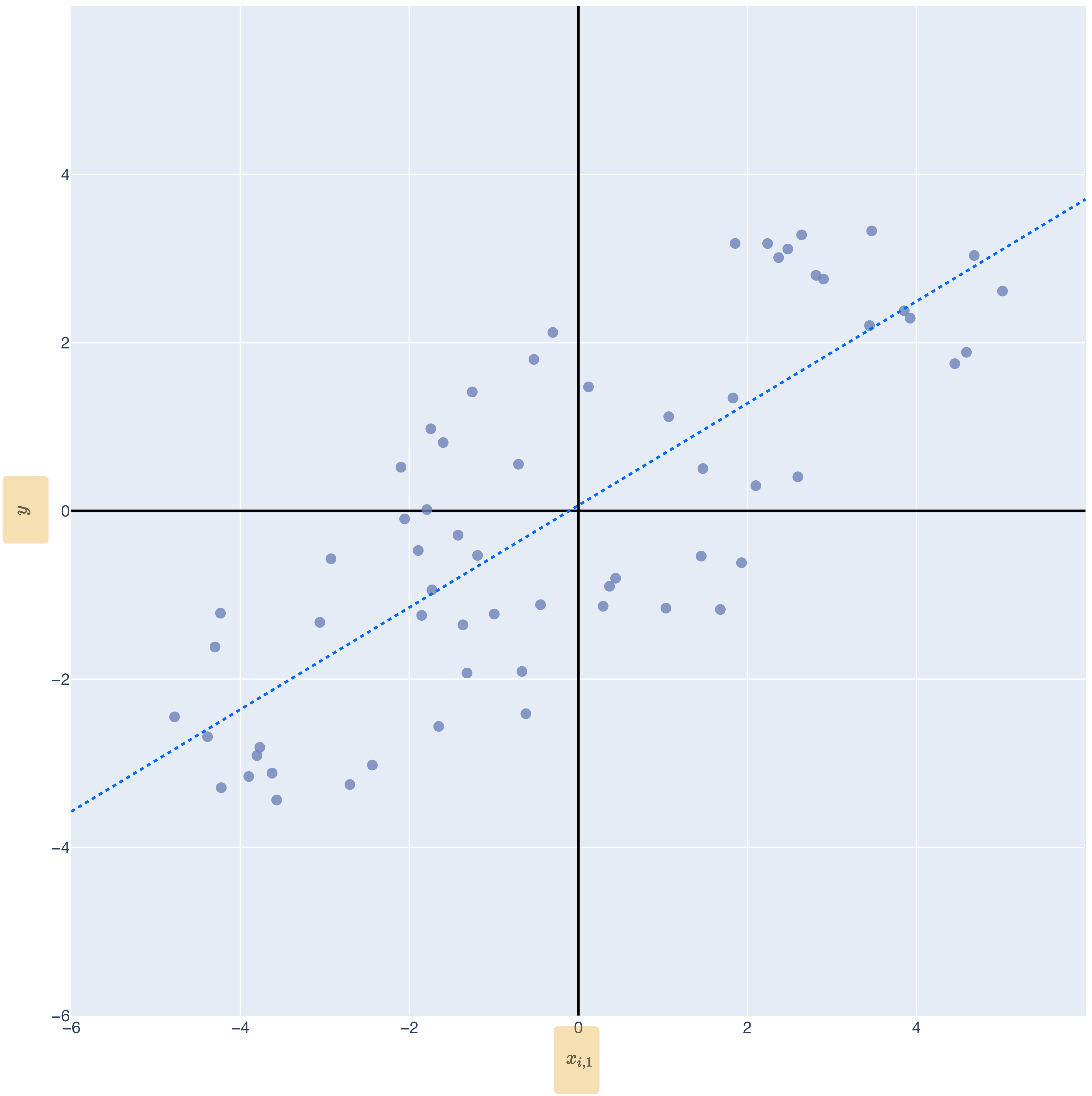
BFS: Find one-dimensional subspace determined by $\mathbf{u} \in \mathbb{R}^n$ such that

$$\arg\min_{\mathbf{u} \in \mathbb{R}^n} \sum_{i=1}^{d} \|\mathbf{x}_i - \Pi_{\mathbf{u}}(\mathbf{x}_i)\|^2$$

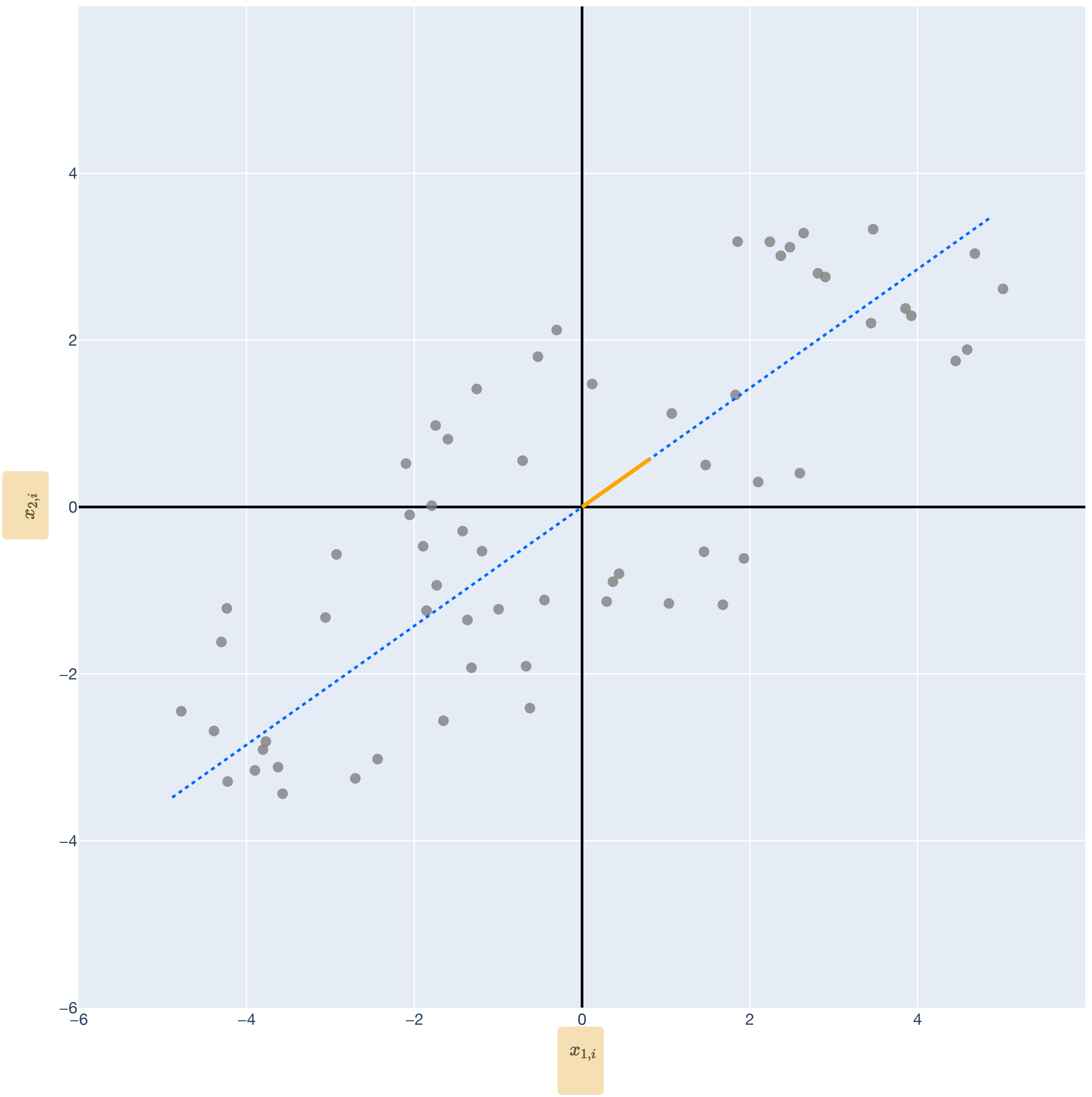
Comparison with OLS

1D Pictures

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



$$\underset{\mathbf{u} \in \mathbb{R}^n}{\arg\min} \sum_{i=1}^{d} \|\mathbf{x}_i - \Pi_{\mathbf{u}}(\mathbf{x}_i)\|^2$$



Best-fitting 1D Subspace

Step 1: Expand out squared projection distance

Find $\mathbf{u} \in \mathbb{R}^n$ that minimizes the sum of squared projection distances:

$$\arg\min_{\mathbf{u}\in\mathbb{R}^n} \sum_{i=1}^{d} \|\mathbf{x}_i - \Pi_{\mathbf{u}}(\mathbf{x}_i)\|^2 = \sum_{i=1}^{d} \boxed{\|\mathbf{x}_i - P_{\mathbf{u}}\mathbf{x}_i\|^2}.$$

1D projection

Orthogonal comp. to $\mathbf{u}$ subspace!

$$\|\mathbf{x}_i - P_{\mathbf{u}}\mathbf{x}_i\|^2 = \left\| \mathbf{x}_i - \left(\frac{\mathbf{u}\mathbf{u}^\top}{\mathbf{u}^\top\mathbf{u}}\right)\mathbf{x}_i \right\|^2 = \left\| \left(\mathbf{I} - \frac{\mathbf{u}\mathbf{u}^\top}{\mathbf{u}^\top\mathbf{u}}\right)\mathbf{x}_i \right\|^2 = \mathbf{x}_i^\top \left(\mathbf{I} - \frac{\mathbf{u}\mathbf{u}^\top}{\mathbf{u}^\top\mathbf{u}}\right)^\top \left(\mathbf{I} - \frac{\mathbf{u}\mathbf{u}^\top}{\mathbf{u}^\top\mathbf{u}}\right)\mathbf{x}_i$$

$$= \mathbf{x}_i^\top \left(\mathbf{I} - \frac{\mathbf{u}\mathbf{u}^\top}{\mathbf{u}^\top\mathbf{u}}\right)^2 \mathbf{x}_i = \boxed{\mathbf{x}_i^\top \left(\mathbf{I} - \frac{\mathbf{u}\mathbf{u}^\top}{\mathbf{u}^\top\mathbf{u}}\right)\mathbf{x}_i}$$

Projections are symmetric

Projecting twice doesn't do anything

Best-fitting 1D Subspace

Step 2: Simplify minimization problem into maximization

Find $\mathbf{u} \in \mathbb{R}^n$ that minimizes the sum of squared projection distances:

$$\underset{\mathbf{u}\in\mathbb{R}^n}{\arg\min} \sum_{i=1}^{d} \|\mathbf{x}_i - \Pi_{\mathbf{u}}(\mathbf{x}_i)\|^2 = \sum_{i=1}^{d} \|\mathbf{x}_i - P_{\mathbf{u}}\mathbf{x}_i\|^2 = \sum_{i=1}^{d} \boxed{\mathbf{x}_i^\top \left(\mathbf{I} - \frac{\mathbf{u}\mathbf{u}^\top}{\mathbf{u}^\top\mathbf{u}}\right)\mathbf{x}_i.}$$

$$= \sum_{i=1}^{d} \mathbf{x}_i^\top \mathbf{x}_i - \mathbf{x}_i^\top \left(\frac{\mathbf{u}\mathbf{u}^\top}{\mathbf{u}^\top\mathbf{u}}\right)\mathbf{x}_i$$

$$\mathbf{u} = \underset{\mathbf{u}\in\mathbb{R}^n}{\arg\min} \sum_{i=1}^{d} \mathbf{x}_i^\top \mathbf{x}_i - \mathbf{x}_i^\top \left(\frac{\mathbf{u}\mathbf{u}^\top}{\mathbf{u}^\top\mathbf{u}}\right)\mathbf{x}_i \iff \boxed{\underset{\mathbf{u}\in\mathbb{R}^n}{\arg\max} \sum_{i=1}^{d} \mathbf{x}_i^\top \left(\frac{\mathbf{u}\mathbf{u}^\top}{\mathbf{u}^\top\mathbf{u}}\right)\mathbf{x}_i}$$

Best-fitting 1D Subspace

Step 3: Derive "operator norm" from matrix outer products

Find $\mathbf{u} \in \mathbb{R}^n$ that minimizes the sum of squared projection distances:

$$\underset{\mathbf{u}\in\mathbb{R}^n}{\arg\min} \sum_{i=1}^{d} \|\mathbf{x}_i - \Pi_{\mathbf{u}}(\mathbf{x}_i)\|^2 = \sum_{i=1}^{d} \|\mathbf{x}_i - P_{\mathbf{u}}\mathbf{x}_i\|^2 = \sum_{i=1}^{d} \mathbf{x}_i^\top \left(\mathbf{I} - \frac{\mathbf{u}\mathbf{u}^\top}{\mathbf{u}^\top\mathbf{u}}\right)\mathbf{x}_i.$$

$$\iff \boxed{\underset{\mathbf{u}\in\mathbb{R}^n}{\arg\max} \sum_{i=1}^{d} \mathbf{x}_i^\top \left(\frac{\mathbf{u}\mathbf{u}^\top}{\mathbf{u}^\top\mathbf{u}}\right)\mathbf{x}_i}$$

$$= \boxed{\underset{\mathbf{u}\in\mathbb{R}^n}{\arg\max} \frac{\mathbf{u}^\top\mathbf{X}\mathbf{X}^\top\mathbf{u}}{\mathbf{u}^\top\mathbf{u}}}$$

squared operator norm of $\mathbf{X}$, i.e. $\|\mathbf{X}\|_{op}^2$

Singular Value Decomposition (SVD)

1D Picture

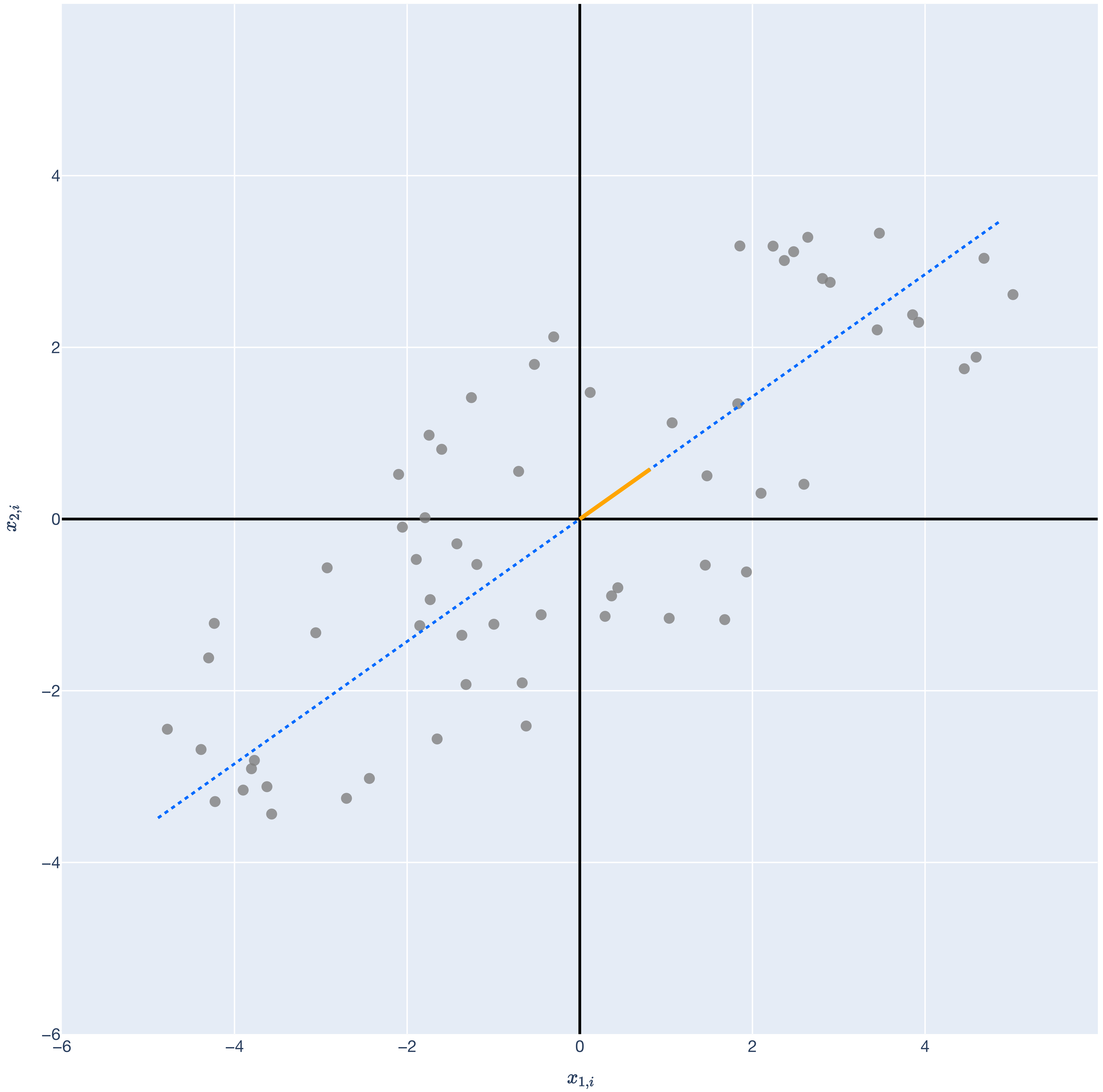
Observe data $\mathbf{x}_1, \ldots, \mathbf{x}_d \in \mathbb{R}^n$.

<u>Goal:</u> Find the best one-dimensional subspace $\mathscr{U} \subseteq \mathbb{R}^n$ that fits the points.

How? Find $\mathbf{u} \in \mathbb{R}^n$ that minimizes the sum of squared projection distances:

$$\arg\min_{\mathbf{u}\in\mathbb{R}^n} \sum_{i=1}^{d} \|\mathbf{x}_i - \Pi_{\mathbf{u}}(\mathbf{x}_i)\|^2 = \boxed{\arg\max_{\mathbf{u}\in\mathbb{R}^n} \frac{\mathbf{u}^\top \mathbf{X}\mathbf{X}^\top \mathbf{u}}{\mathbf{u}^\top \mathbf{u}}}.$$

$\mathbf{u} \in \mathbb{R}^n$ is the 1st <u>left singular vector</u> with 1st (squared) <u>singular value</u> $\sigma_1^2 = \dfrac{\mathbf{u}^\top \mathbf{X}\mathbf{X}^\top \mathbf{u}}{\mathbf{u}^\top \mathbf{u}}$



Singular Value Decomposition

Definition of Full SVD and Compact SVD

Singular Value Decomposition (SVD)

Building up the SVD

Observe data $\mathbf{x}_1, \ldots, \mathbf{x}_d \in \mathbb{R}^n$. Consider the following procedure…

For $t = 1,2,\ldots, n$:

1. Find $\mathbf{u}_1 \in \mathbb{R}^n$, the best one-dimensional subspace fit to $\mathbf{x}_1, \ldots, \mathbf{x}_d$.

   Let $\mathbf{x}_i^{(1)} = \mathbf{x}_i - \Pi_{\mathbf{u}_1}(\mathbf{x}_i)$.

2. Find $\mathbf{u}_2 \in \mathbb{R}^n$, the best one-dimensional subspace fit to $\mathbf{x}_1^{(1)}, \ldots, \mathbf{x}_d^{(1)}$.

   Let $\mathbf{x}_i^{(2)} = \mathbf{x}_i^{(1)} - \Pi_{\mathbf{u}_2}(\mathbf{x}_i) = \mathbf{x}_i - \Pi_{\mathbf{u}_1}(\mathbf{x}_i) - \Pi_{\mathbf{u}_2}(\mathbf{x}_i)$.

3. Find $\mathbf{u}_3 \in \mathbb{R}^n$, the best one-dimensional subspace fit to $\mathbf{x}_1^{(2)}, \ldots, \mathbf{x}_d^{(2)}\ldots$

Singular Value Decomposition (SVD)

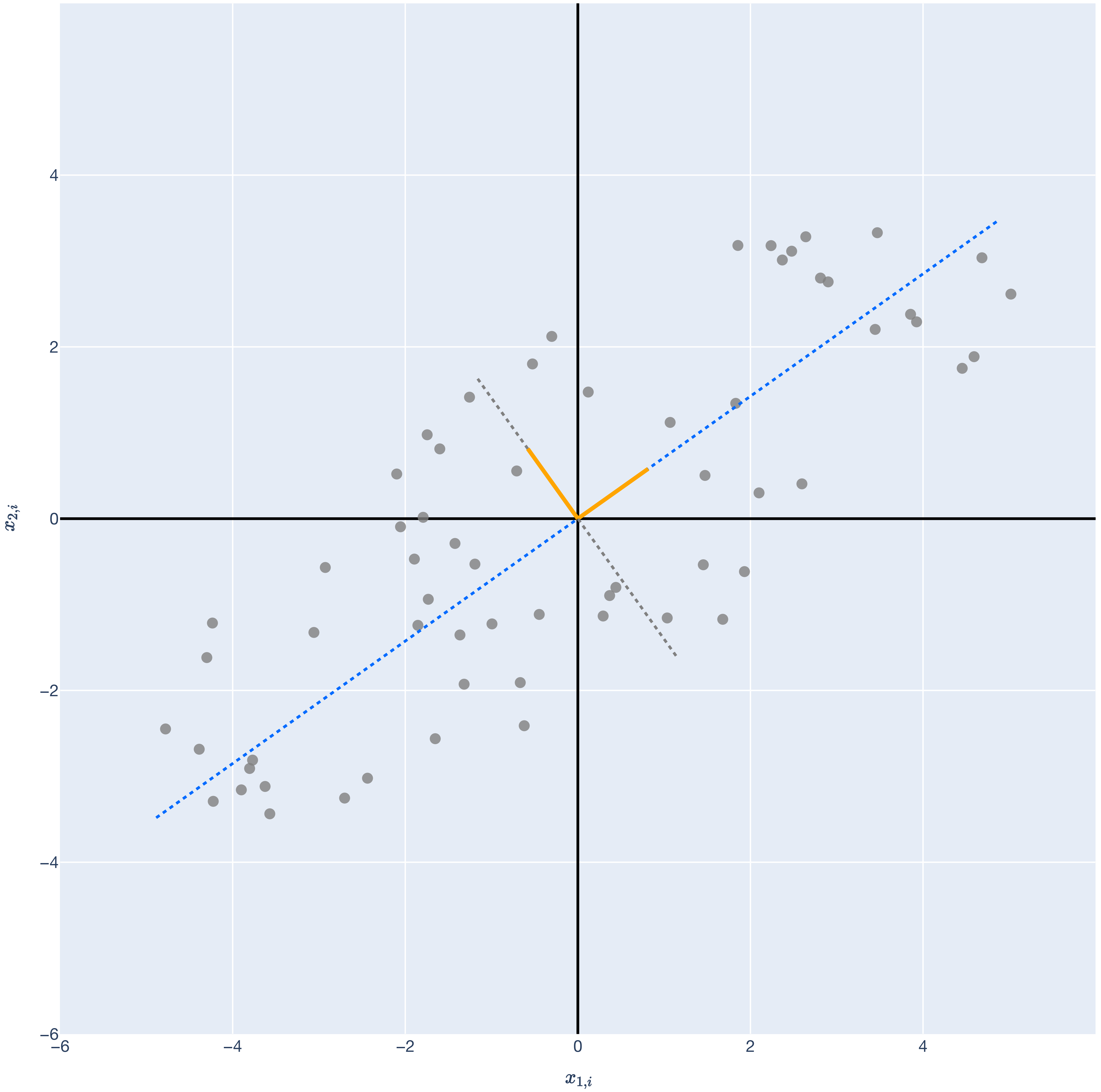
Building up the SVD

Observe data $\mathbf{x}_1, \ldots, \mathbf{x}_d \in \mathbb{R}^2$.

1. Find $\mathbf{u}_1 \in \mathbb{R}^2$, the best one-dimensional subspace fit to $\mathbf{x}_1, \ldots, \mathbf{x}_d$.

$$\text{Let } \mathbf{x}_i^{(1)} = \mathbf{x}_i - \Pi_{\mathbf{u}_1}(\mathbf{x}_i).$$

2. Find $\mathbf{u}_2 \in \mathbb{R}^n$, the best one-dimensional subspace fit to $\mathbf{x}_1^{(1)}, \ldots, \mathbf{x}_d^{(1)}$.



Singular Value Decomposition (SVD)

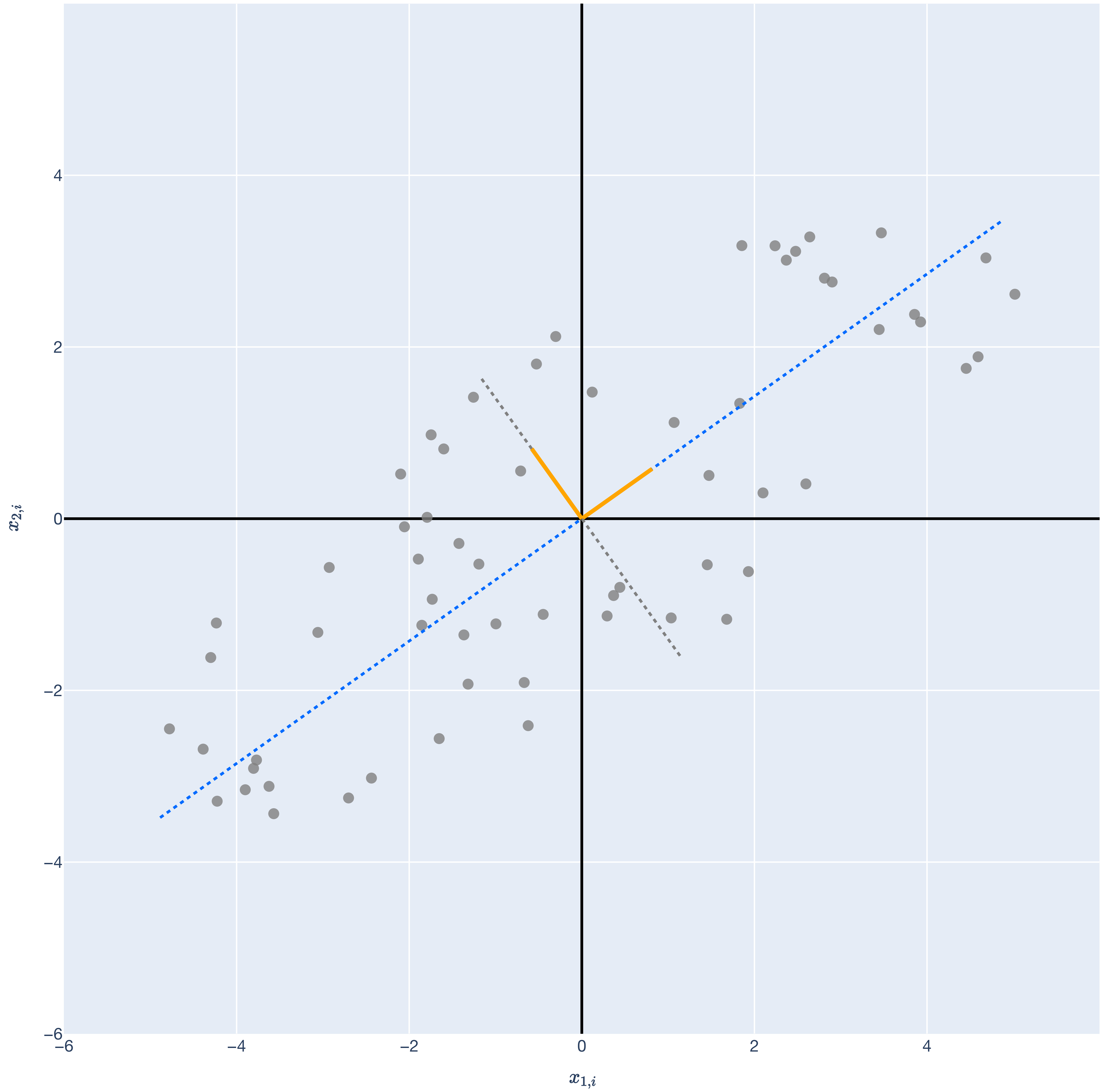
Building up the SVD

$\mathbf{u}_t \in \mathbb{R}^n$is the best one-dimensional subspace fit to:

$$\mathbf{x}_i - \sum_{k=1}^{t-1} \Pi_{\mathbf{u}_k}(\mathbf{x}_i).$$

These are the n left singular vectors of $\mathbf{X} \in \mathbb{R}^{n \times d}$.

Orthogonal, by construction (left singular vector $\mathbf{u}_k$ is in the orthogonal complement of $\mathbf{u}_1, \ldots, \mathbf{u}_{k-1}$).



Singular Value Decomposition (SVD)

Definition of the Full SVD

Consider any matrix $\mathbf{X} \in \mathbb{R}^{n \times d}$. The full singular value decomposition (SVD) is

$$\underbrace{\mathbf{X}}_{n \times d} = \underbrace{\mathbf{U}}_{n \times n} \underbrace{\mathbf{\Sigma}}_{n \times d} \underbrace{\mathbf{V}^\top}_{d \times d}.$$

The columns of $\mathbf{U} \in \mathbb{R}^{n \times n}$ are the left singular vectors and $\mathbf{U}$ is orthogonal: $\mathbf{U}^\top \mathbf{U} = \mathbf{U}\mathbf{U}^\top = \mathbf{I}$.

The columns of $\mathbf{V} \in \mathbb{R}^{d \times d}$ are the right singular vectors and $\mathbf{V}$ is orthogonal: $\mathbf{V}^\top \mathbf{V} = \mathbf{V}\mathbf{V}^\top = \mathbf{I}$.

$\mathbf{\Sigma} \in \mathbb{R}^{n \times d}$ is a diagonal matrix with singular values $\sigma_1 \geq \sigma_2 \geq \ldots \geq \sigma_d \geq 0$ on the diagonal.

The rank of $\mathbf{X}$ is equal to the number of $\sigma_i > 0$.

Singular Value Decomposition (SVD)

Shape of the $\mathbf{\Sigma}$ Matrix

$\mathbf{\Sigma} \in \mathbb{R}^{n \times d}$ is a diagonal matrix with singular values $\sigma_1 \geq \sigma_2 \geq \ldots \geq \sigma_{\min\{n,d\}} \geq 0$ on the diagonal.

$$\underbrace{\mathbf{\Sigma} = \begin{bmatrix} \sigma_1 & 0 & \ldots & 0 \\ 0 & \sigma_2 & \ldots & 0 \\ 0 & 0 & \ddots & \vdots \\ 0 & 0 & \ldots & \sigma_d \end{bmatrix}}_{n=d} \text{ or } \underbrace{\mathbf{\Sigma} = \begin{bmatrix} \sigma_1 & 0 & \ldots & 0 \\ 0 & \sigma_2 & \ldots & 0 \\ 0 & 0 & \ddots & \vdots \\ 0 & 0 & \ldots & \sigma_d \\ 0 & 0 & \ldots & 0 \\ 0 & 0 & \ldots & 0 \\ \vdots & \vdots & \vdots & \vdots \end{bmatrix}}_{n>d} \text{ or } \underbrace{\mathbf{\Sigma} = \begin{bmatrix} \sigma_1 & 0 & \ldots & 0 & 0 & 0 & \ldots \\ 0 & \sigma_2 & \ldots & 0 & 0 & 0 & \ldots \\ 0 & 0 & \ddots & \vdots & \vdots & \vdots & \ldots \\ 0 & 0 & \ldots & \sigma_n & 0 & 0 & \ldots \end{bmatrix}}_{d>n}$$

Interpreting the SVD

Example in $\mathbb{R}^2$

Let $\mathbf{x}_1, \ldots, \mathbf{x}_{212} \in \mathbb{R}^2$. The SVD is given by:

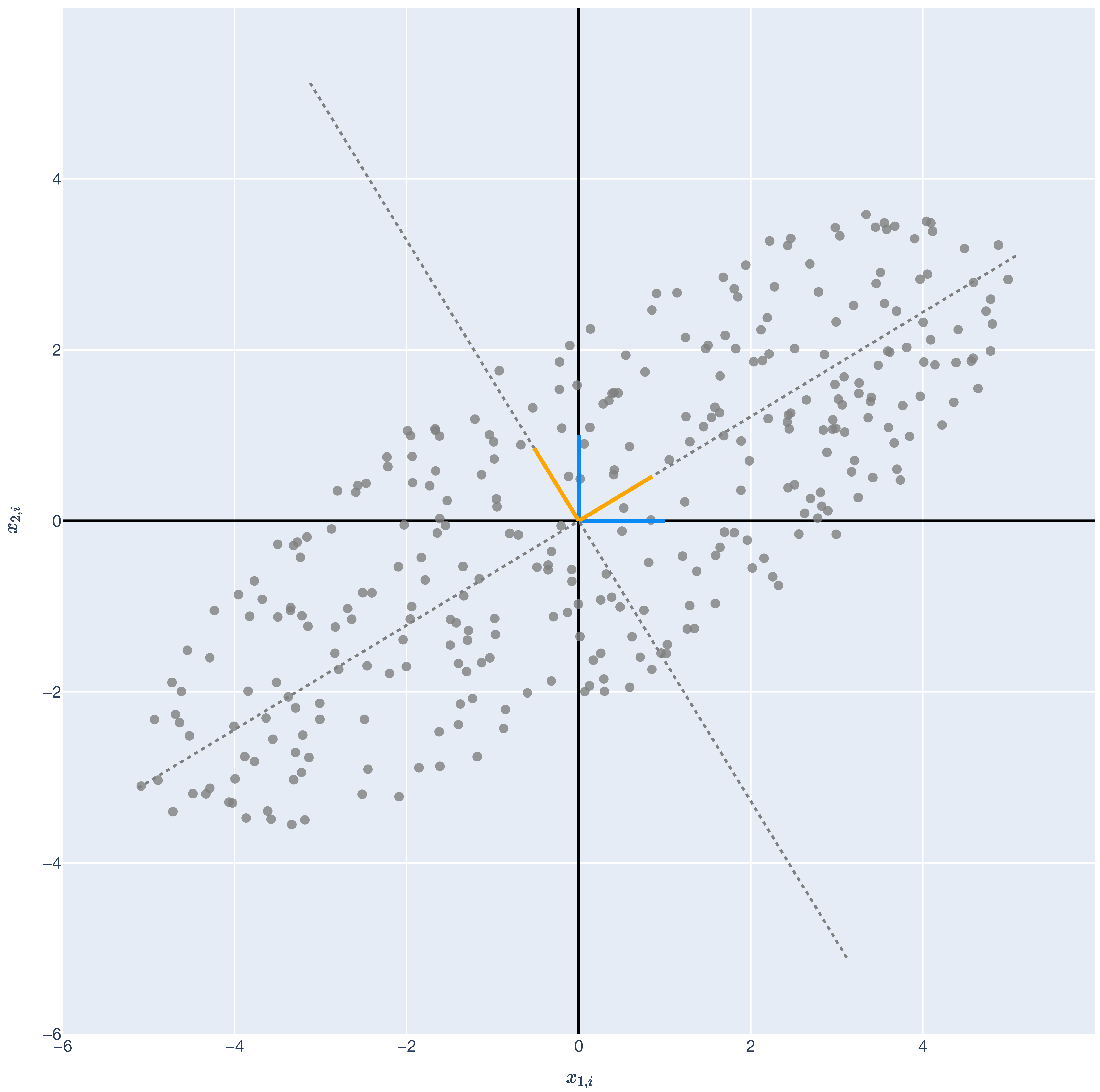
$$\underbrace{\mathbf{X}}_{2\times 212} = \underbrace{\mathbf{U}}_{2\times 2} \underbrace{\mathbf{\Sigma}}_{2\times 212} \underbrace{\mathbf{V}^\top}_{212\times 212}$$

Left Singular Vectors

Interpreting the $\mathbf{U}$ matrix

$$\underbrace{\mathbf{X}}_{2\times 212} = \underbrace{\mathbf{U}}_{2\times 2} \underbrace{\mathbf{\Sigma}}_{2\times 212} \underbrace{\mathbf{V}^{\top}}_{212\times 212}$$

The columns $\mathbf{u}_1, \mathbf{u}_2 \in \mathbb{R}^2$ of $\mathbf{U}$ are an orthonormal basis for $\mathrm{CS}(\mathbf{X})$.



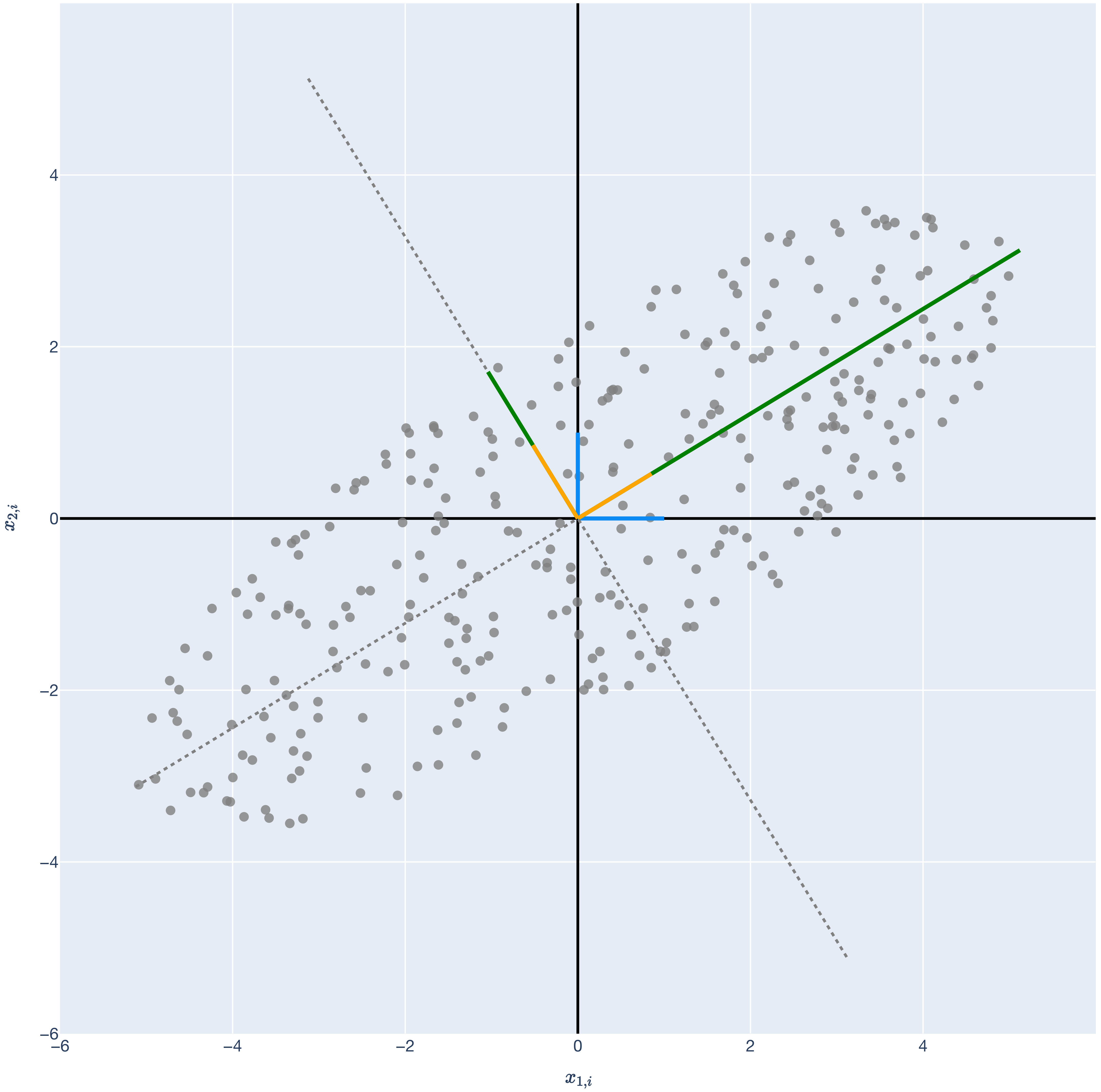
Singular Values

Interpreting the $\mathbf{\Sigma}$ matrix

$$\underbrace{\mathbf{X}}_{2\times212} = \underbrace{\mathbf{U}}_{2\times2} \underbrace{\mathbf{\Sigma}}_{2\times212} \underbrace{\mathbf{V}^\top}_{212\times212}$$

The singular values $\sigma_1, \sigma_2 > 0$ represent how to scale $\mathbf{u}_1$ and $\mathbf{u}_2$ to "fit" all the data.

They represent the relative "strength" of $\mathbf{u}_1$ and $\mathbf{u}_2$ in explaining the data.



Right Singular Vectors

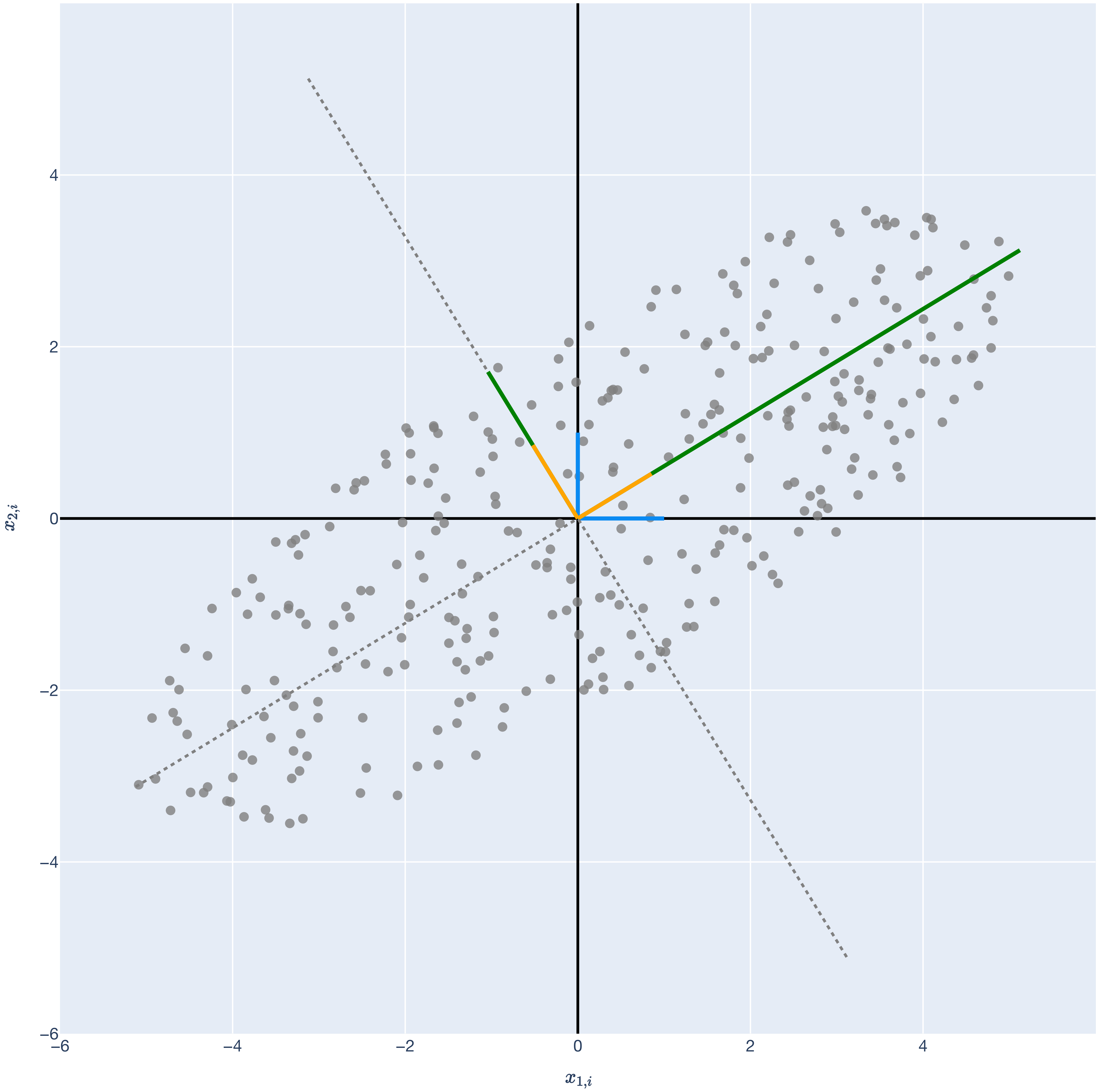
Interpreting the $\mathbf{V}$ matrix

$$\underbrace{\mathbf{X}}_{2\times 212} = \underbrace{\mathbf{U}}_{2\times 2} \underbrace{\mathbf{\Sigma}}_{2\times 212} \underbrace{\mathbf{V}^\top}_{212\times 212}$$

The rows of $\mathbf{V}^\top$ give the coordinates for each point under the basis $\sigma_1\mathbf{u}_1, \sigma_2\mathbf{u}_2$.

Specifically, for $j \in [d]$,

$$\mathbf{x}_j = v_{1j}\sigma_1\mathbf{u}_1 + v_{2j}\sigma_2\mathbf{u}_2.$$



Right Singular Vectors

Interpreting the $\mathbf{V}$ matrix

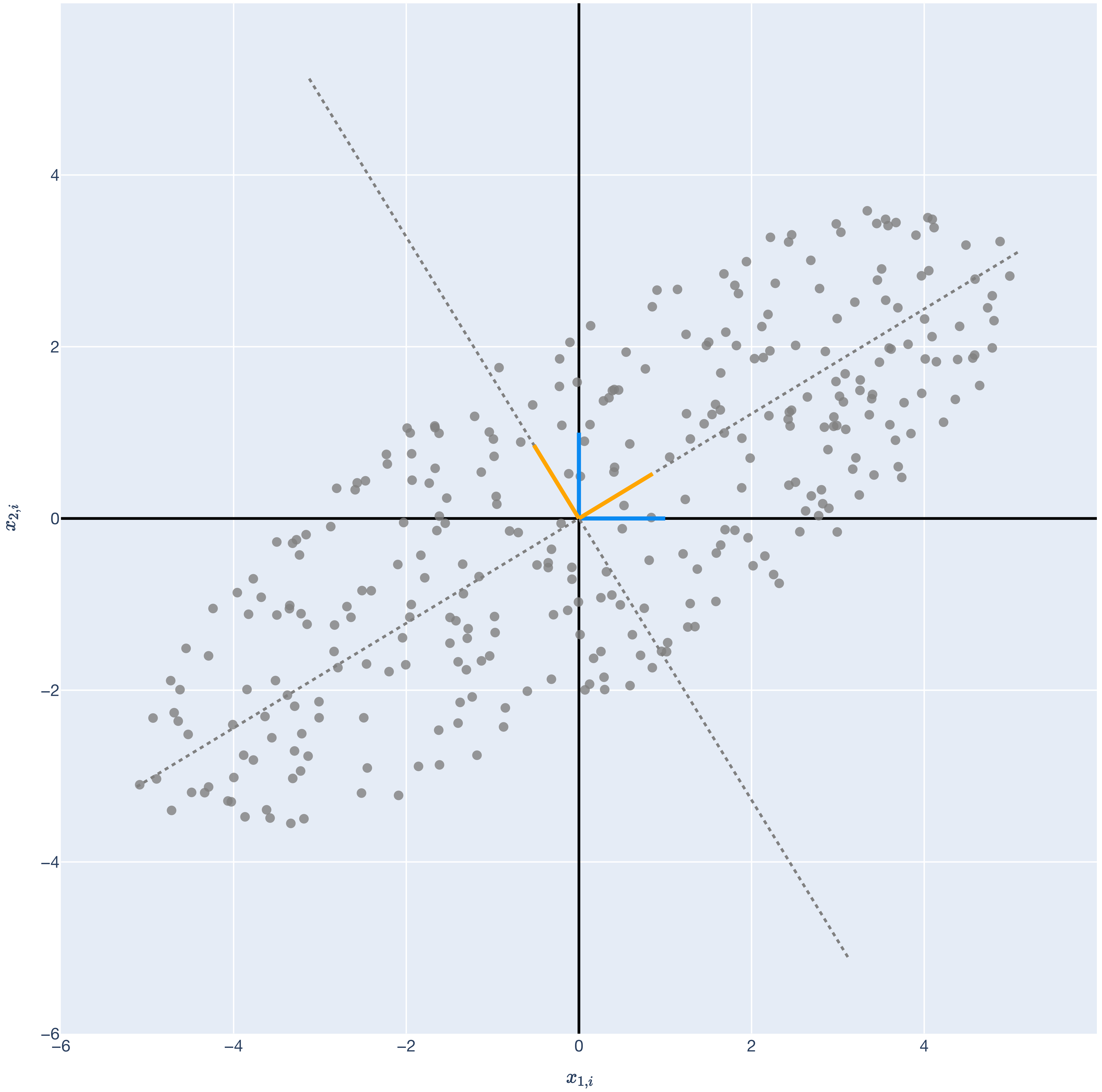
Specifically, for $j \in [d]$,

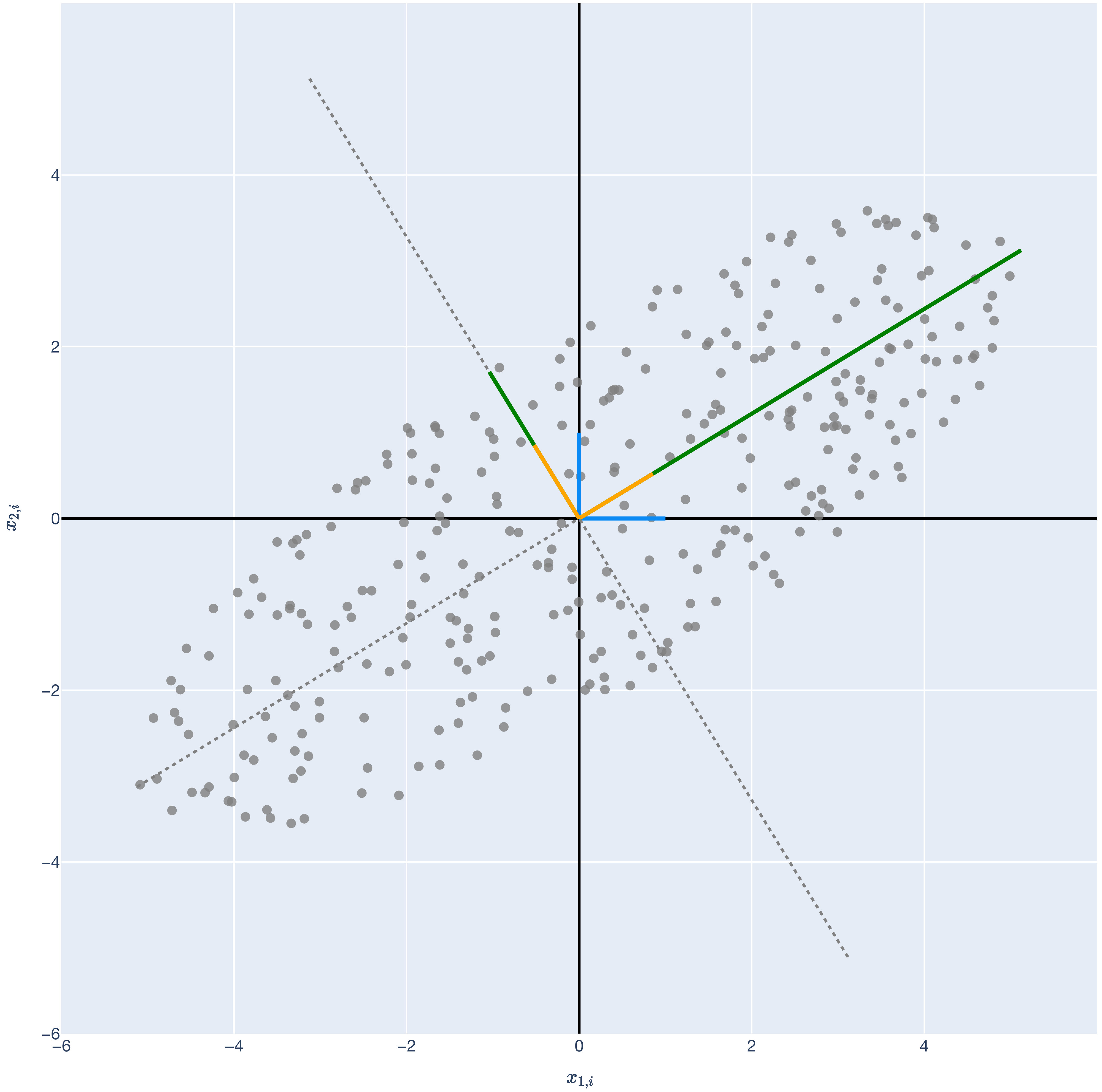
$$\mathbf{x}_j = v_{1j}\sigma_1\mathbf{u}_1 + v_{2j}\sigma_2\mathbf{u}_2.$$

$$\begin{bmatrix} \uparrow & \uparrow & \uparrow & \uparrow & \uparrow \\ \mathbf{x}_1 & \mathbf{x}_2 & \mathbf{x}_3 & \dots & \mathbf{x}_{212} \\ \downarrow & \downarrow & \downarrow & \downarrow & \downarrow \end{bmatrix} = \begin{bmatrix} \uparrow & \uparrow \\ \mathbf{u}_1 & \mathbf{u}_2 \\ \downarrow & \downarrow \end{bmatrix} \begin{bmatrix} \sigma_1 & 0 & 0 & \dots & 0 \\ 0 & \sigma_2 & 0 & \dots & 0 \end{bmatrix} \begin{bmatrix} \leftarrow & \mathbf{v}_1^\top & \rightarrow \\ \leftarrow & \mathbf{v}_2^\top & \rightarrow \\ \vdots & \vdots & \vdots \\ \leftarrow & \mathbf{v}_{212}^\top & \rightarrow \end{bmatrix}$$

Interpretation of the SVD

Full Interpretation of the SVD





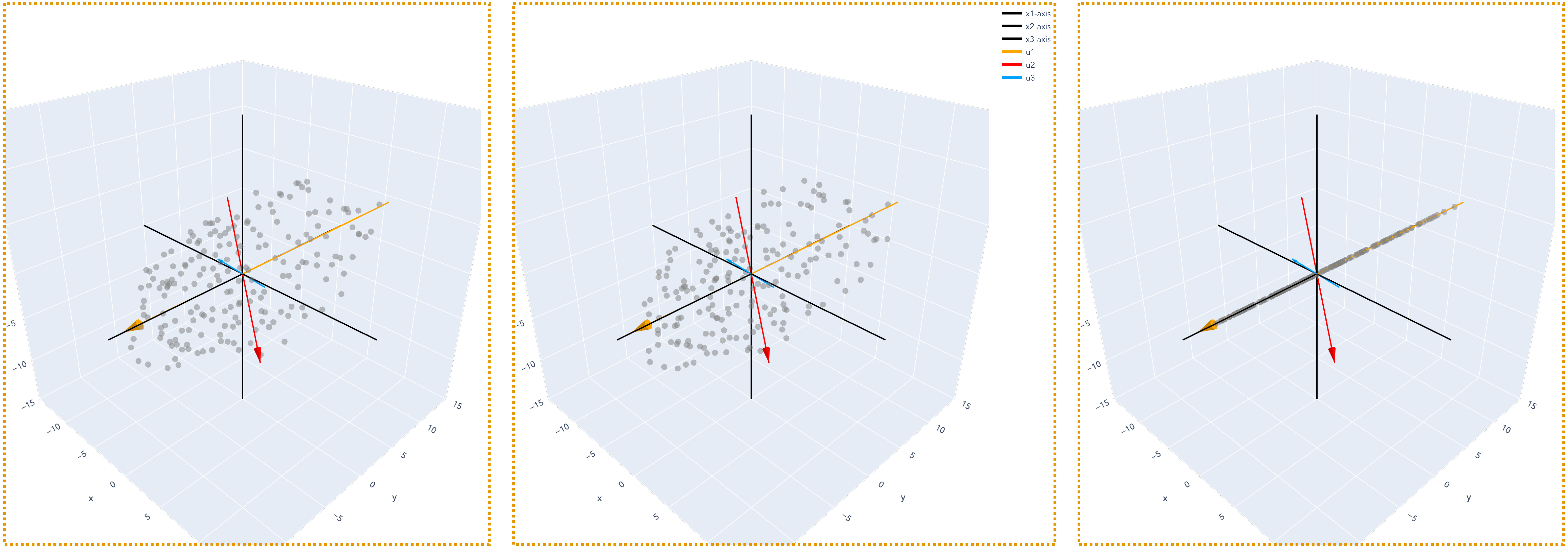
Singular Value Decomposition (SVD)

Example of SVD

$$\mathbf{X} = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 5 & 0 \\ 0 & 0 & 10 \end{bmatrix}$$

Singular Value Decomposition (SVD)

Example in $\mathbb{R}^3$



Singular Value Decomposition (SVD)

Definition of the Compact SVD

$\mathbf{X} \in \mathbb{R}^{n \times d}$ with rank $r \leq \min\{n, d\}$ has compact singular value decomposition (SVD):

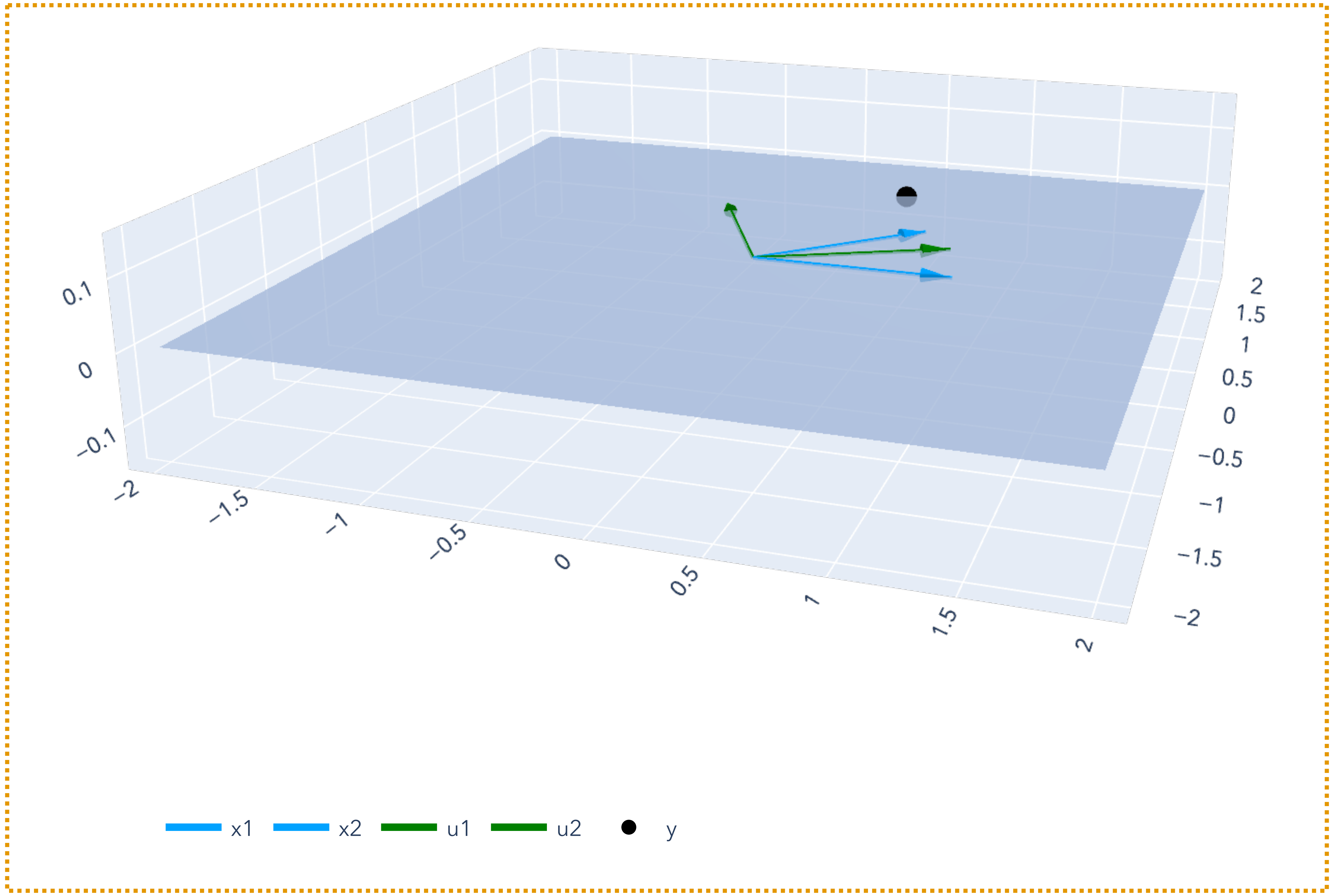
$$\underbrace{\mathbf{X}}_{n \times d} = \underbrace{\mathbf{U}}_{n \times r} \underbrace{\mathbf{\Sigma}}_{r \times r} \underbrace{\mathbf{V}^\top}_{r \times d}.$$

Columns of $\mathbf{U} \in \mathbb{R}^{n \times r}$ are the left singular vectors and $\mathbf{U}^\top \mathbf{U} = \mathbf{I}$, o.n.b. for $\mathrm{CS}(\mathbf{X})$.

Columns of $\mathbf{V} \in \mathbb{R}^{r \times d}$ are the right singular vectors and $\mathbf{V}^\top \mathbf{V} = \mathbf{I}$, o.n.b. for $\mathrm{CS}(\mathbf{X}^\top)$.

$\mathbf{\Sigma} \in \mathbb{R}^{r \times r}$ is a square diagonal matrix with singular values $\sigma_1 \geq \sigma_2 \geq \ldots \geq \sigma_r > 0$ on diagonal.

How to find a good orthogonal basis?



Least Squares

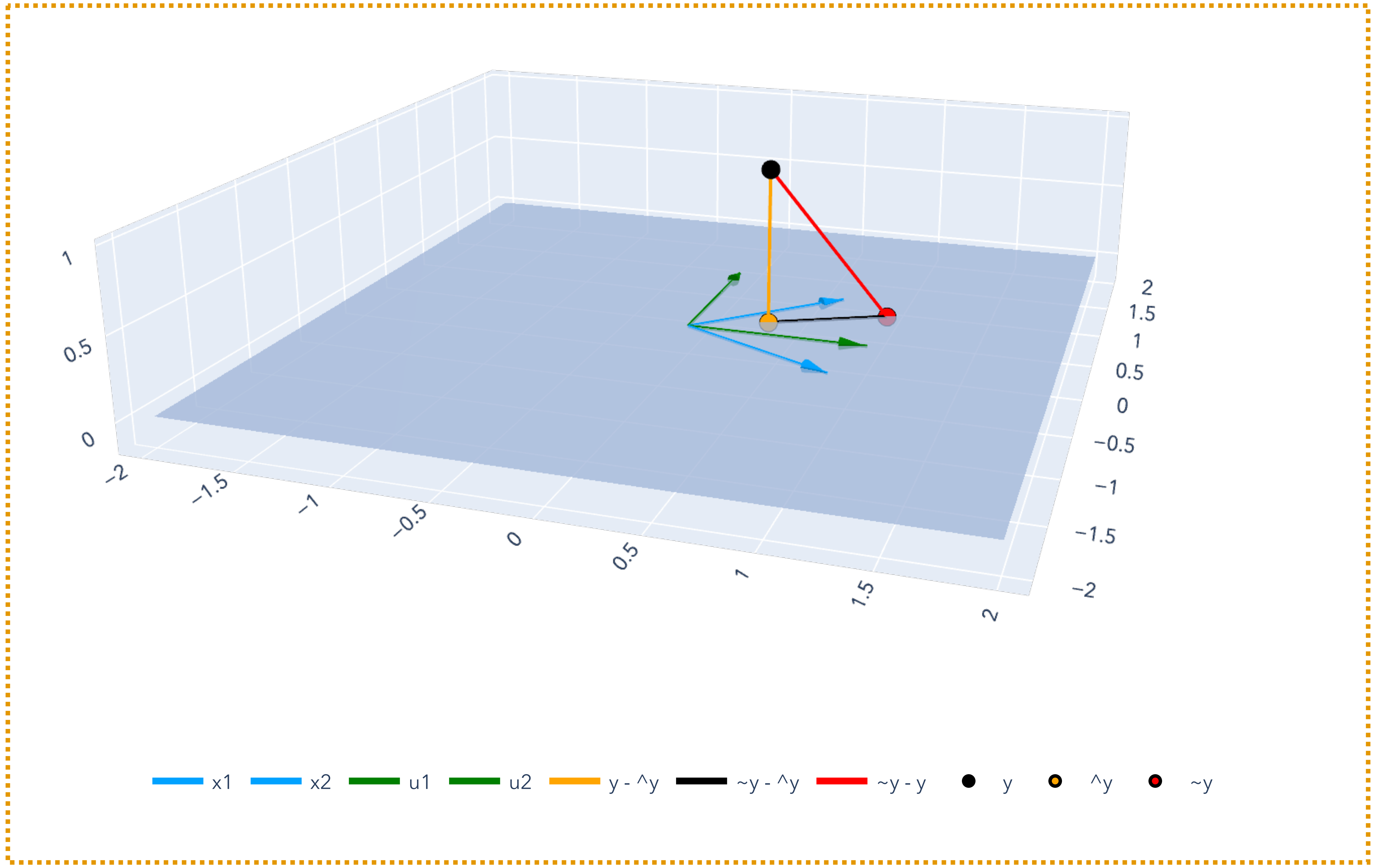
OLS with Orthogonal Basis

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

$$\hat{\mathbf{w}}_{onb} = \mathbf{U}^\top\mathbf{y}$$

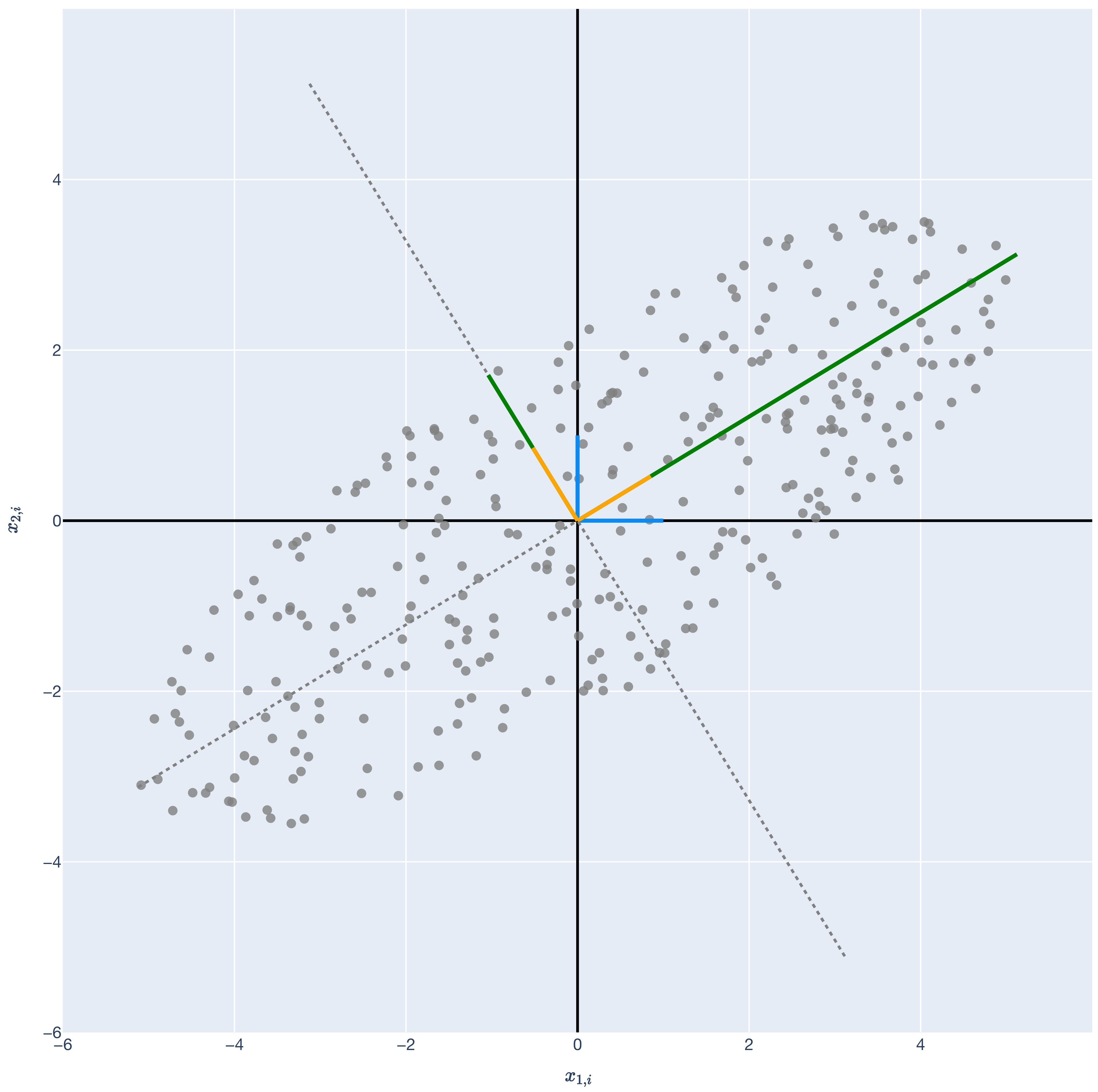
$$\hat{\mathbf{y}} = \Pi_{\mathscr{X}}(\mathbf{y}) = \mathbf{X}(\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{X}^\top\mathbf{y}$$

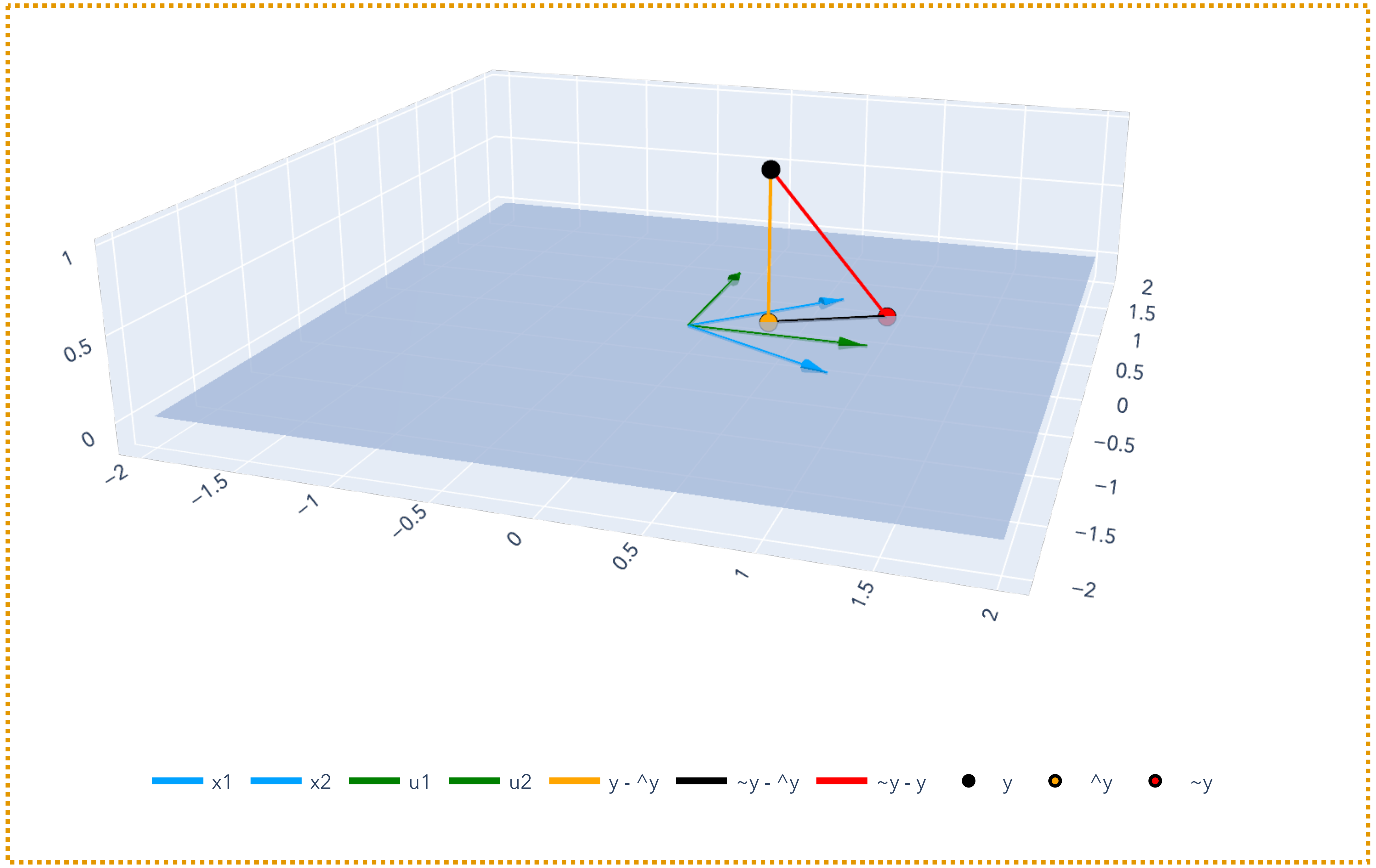
$$\hat{\mathbf{y}} = \Pi_{\mathscr{X}}(\mathbf{y}) = \mathbf{U}\mathbf{U}^\top\mathbf{y}$$



Least Squares

OLS with Orthogonal Basis





$$\hat{\mathbf{w}}_{onb} = \mathbf{U}^\top \mathbf{y}$$

$$\hat{\mathbf{y}} = \Pi_{\mathscr{X}}(\mathbf{y}) = \mathbf{U}\mathbf{U}^\top \mathbf{y}$$

Singular Value Decomposition

Application: Low-rank Approximation

Rank-*k* Approximation

Idea

In many applications, it is useful to *approximate* a matrix.

The *rank* of a matrix represents how many linearly independent columns (or rows) make up a matrix (i.e. how much "novel information" the matrix contains).

We might approximate a matrix $\mathbf{X}$ with $r = \operatorname{rank}(\mathbf{X})$ by asking:

> *What's the closest rank-k matrix (with $k \ll r$) to $\mathbf{X}$?*

One notion of "close" for matrices is the Frobenius norm: $\|\mathbf{X}\|_F := \sqrt{\sum_{i=1}^{n} \sum_{j=1}^{d} X_{ij}^2}$.

Rank-k Approximation

Theorem

Theorem (Rank-k Approximation). Let $\mathbf{X} \in \mathbb{R}^{n \times d}$. If $\mathbf{X} = \mathbf{U\Sigma V}^\top$ is the compact SVD of $\mathbf{X}$ with $\mathbf{U}_k \in \mathbb{R}^{n \times k}$, $\mathbf{\Sigma}_k \in \mathbb{R}^{k \times k}$, and $\mathbf{V}_k \in \mathbb{R}^{d \times k}$ as truncated matrices of $\mathbf{U}$, $\mathbf{\Sigma}$, and $\mathbf{V}$, respectively, then

$$\hat{\mathbf{X}}_k = \mathbf{U}_k \mathbf{\Sigma}_k \mathbf{V}_k^\top \text{ and } \|\mathbf{X} - \hat{\mathbf{X}}_k\|^2 = \sum_{i=k+1}^{r} \sigma_i^2.$$

Then, $\hat{\mathbf{X}}_k \in \mathbb{R}^{n \times d}$ is the rank-k approximation of $\mathbf{X}$ in Frobenius norm:

$$\hat{\mathbf{X}}_k = \arg\min_{\hat{\mathbf{X}} \in \mathbb{R}^{n \times d}} \|\mathbf{X} - \hat{\mathbf{X}}\|_F, \text{ such that } \mathrm{rank}(\hat{\mathbf{X}}) = k.$$

Rank-*k* Approximation

Outer Product Interpretation

The (compact) SVD of a matrix can also be written as a sum of rank-1 matrices.

$$\mathbf{X} = \underbrace{\sigma_1 \mathbf{u}_1 \mathbf{v}_1^\top}_{n \times d} + \sigma_2 \mathbf{u}_2 \mathbf{v}_2^\top + \ldots + \sigma_r \mathbf{u}_r \mathbf{v}_r^\top .$$

In this way, the rank-k approximation $\hat{\mathbf{X}}_k$ can be written as truncating this sum at k:

$$\hat{\mathbf{X}}_k = \sigma_1 \mathbf{u}_1 \mathbf{v}_1^\top + \ldots + \sigma_k \mathbf{u}_k \mathbf{v}_k^\top .$$

Rank-k Approximation

Example

Consider the 4×4 matrix:

$$\mathbf{X} = \begin{bmatrix} 100 & 0 & 0 & 0 \\ 0 & 90 & 0 & 0 \\ 0 & 0 & 8 & 0 \\ 0 & 0 & 0 & 2 \end{bmatrix}$$

Rank-*k* Approximation

Application in Image Processing


```
print(X)
print("Shape: {}".format(X.shape))
```

✓ 0.0s

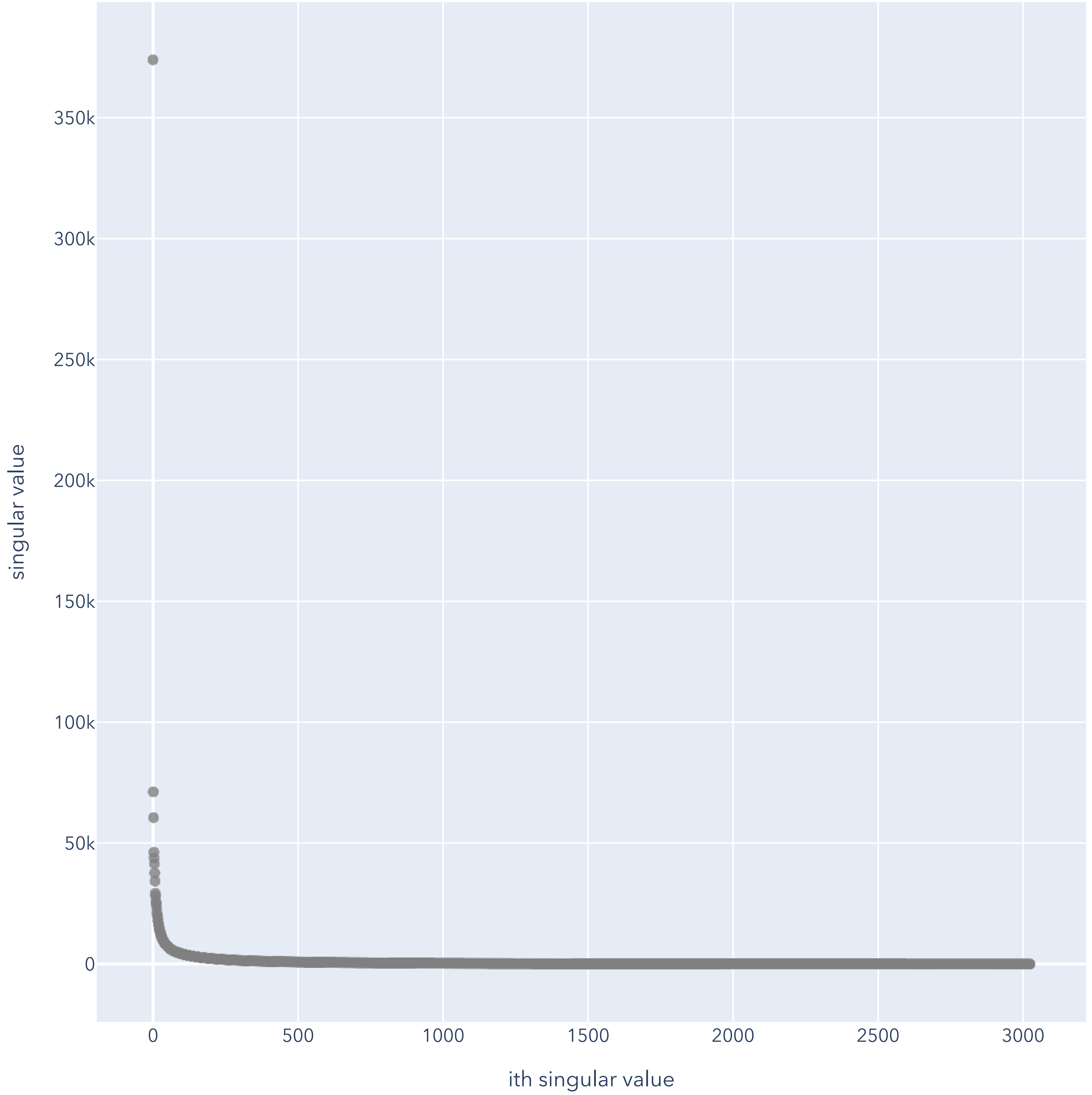
```
[[37 39 38 ... 32 31 29]
 [40 43 41 ... 32 30 27]
 [41 45 44 ... 32 30 27]
 ...
 [50 51 54 ... 57 58 58]
 [50 53 56 ... 57 58 60]
 [50 53 55 ... 58 60 63]]
Shape: (3024, 4032)
```

```
# Take an SVD
U, S, Vt = np.linalg.svd(X, full_matrices=False)
```

✓ 16.5s

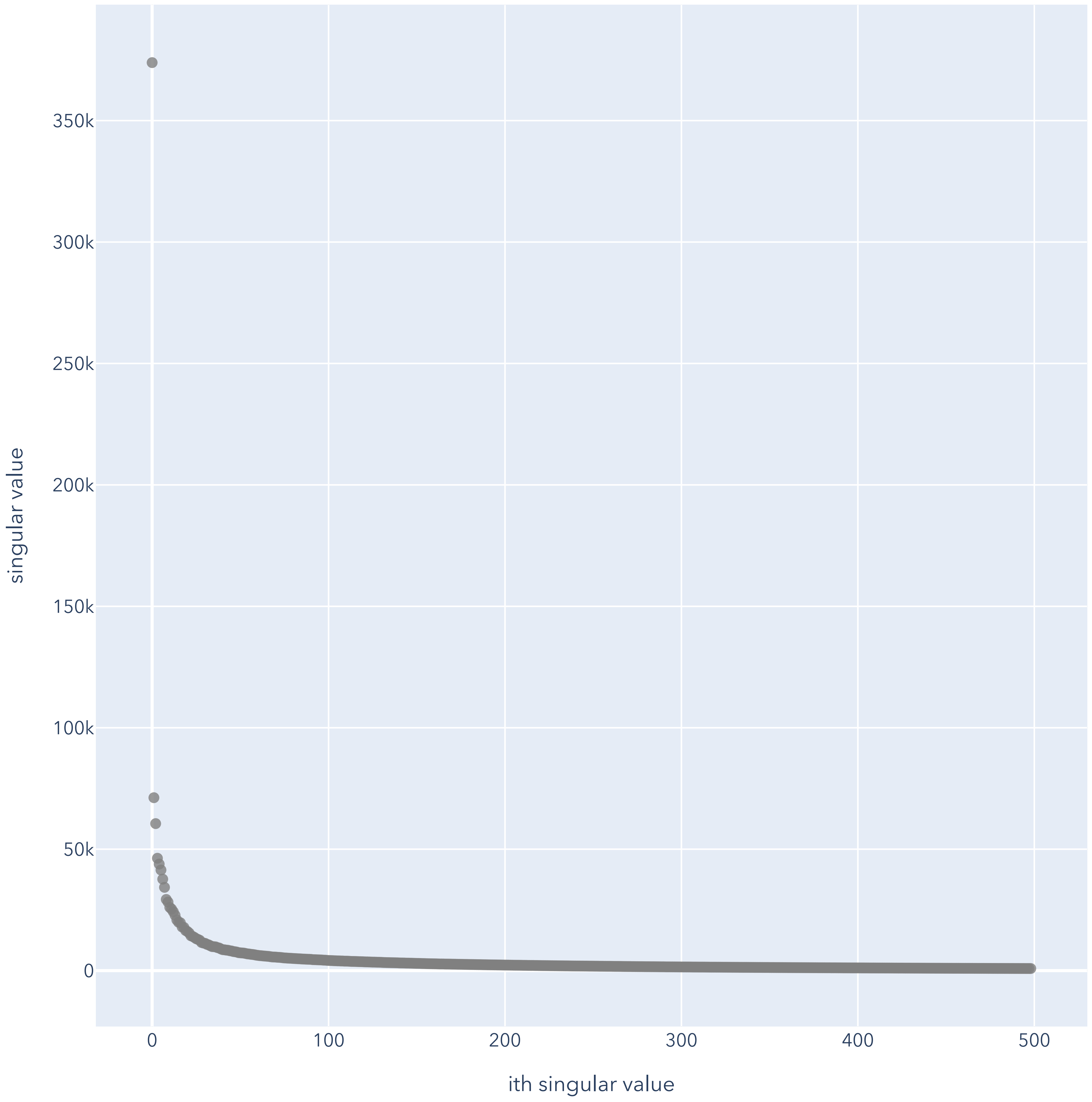
Rank-*k* Approximation

Application in Image Processing



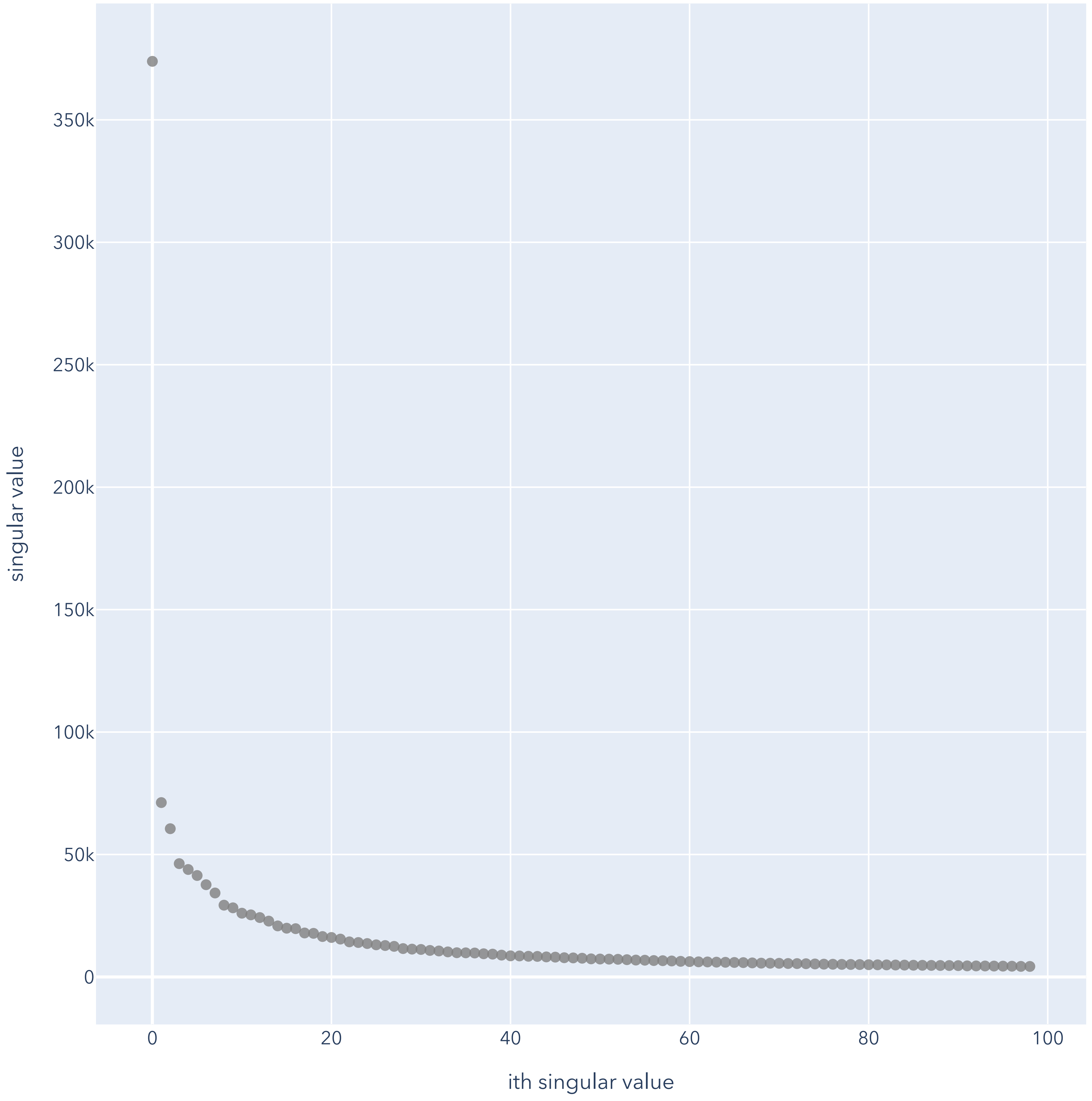
Rank-*k* Approximation

Application in Image Processing (k = *500*)



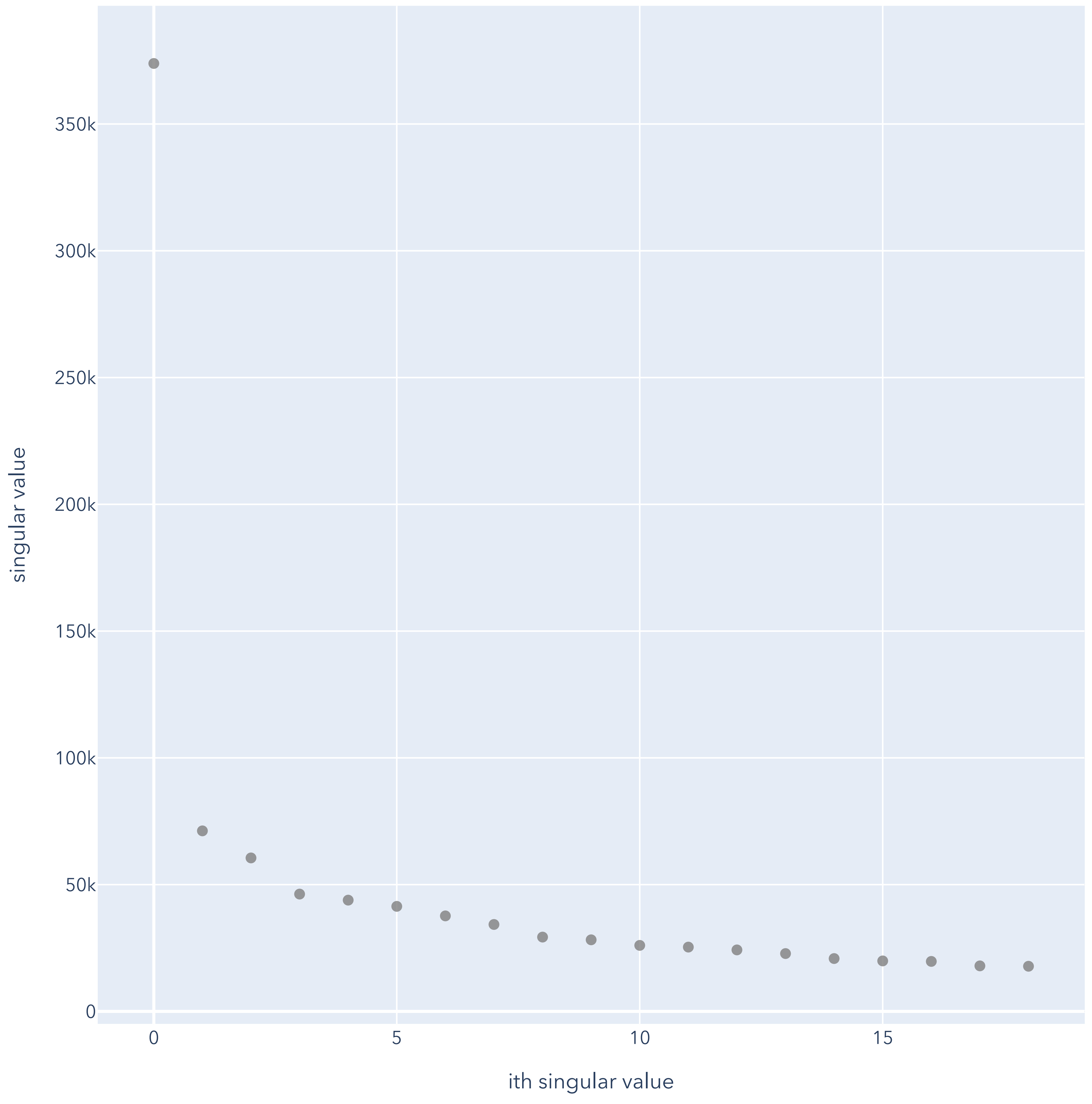
Rank-*k* Approximation

Application in Image Processing (*k* = *100*)



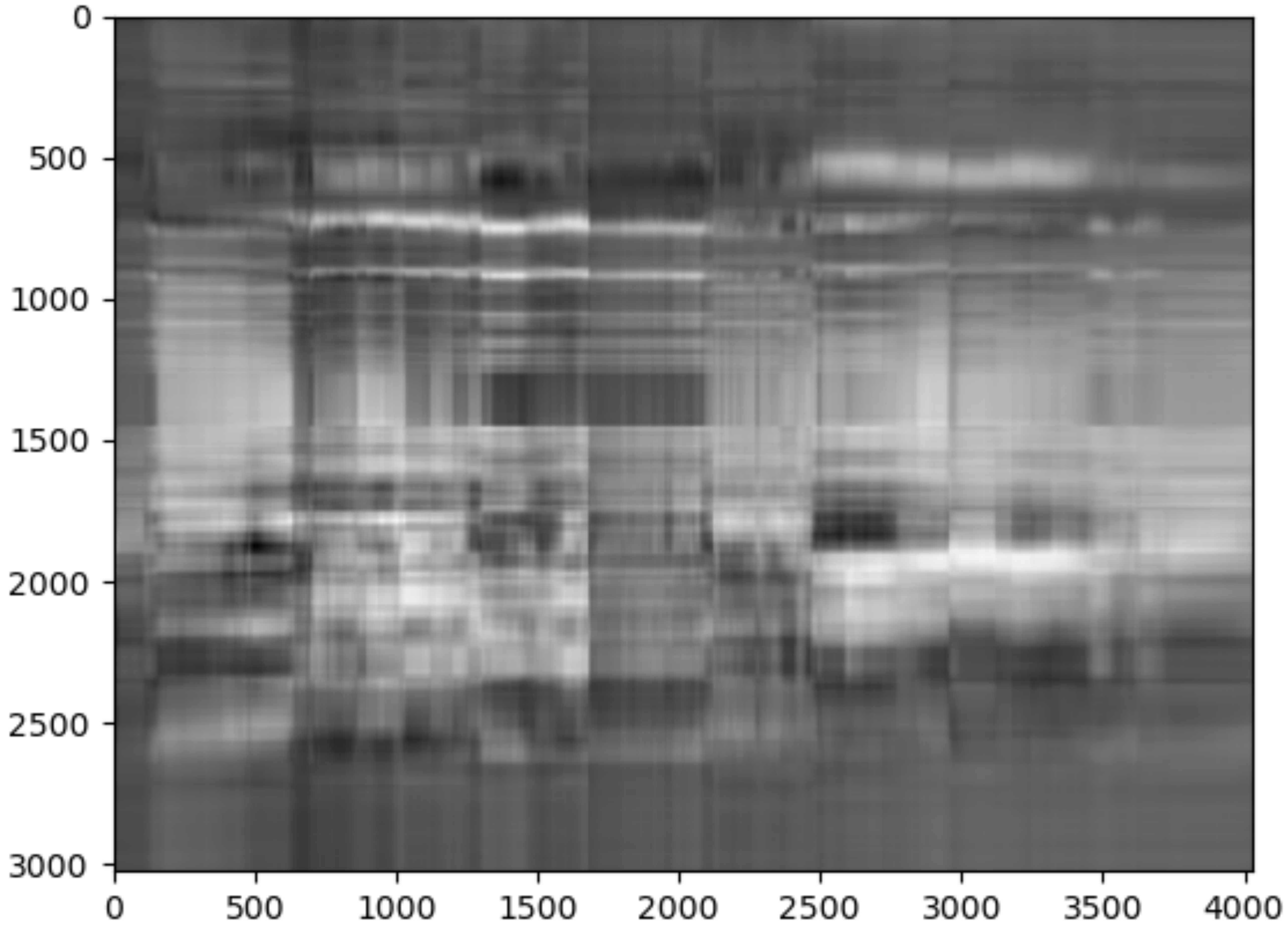
Rank-*k* Approximation

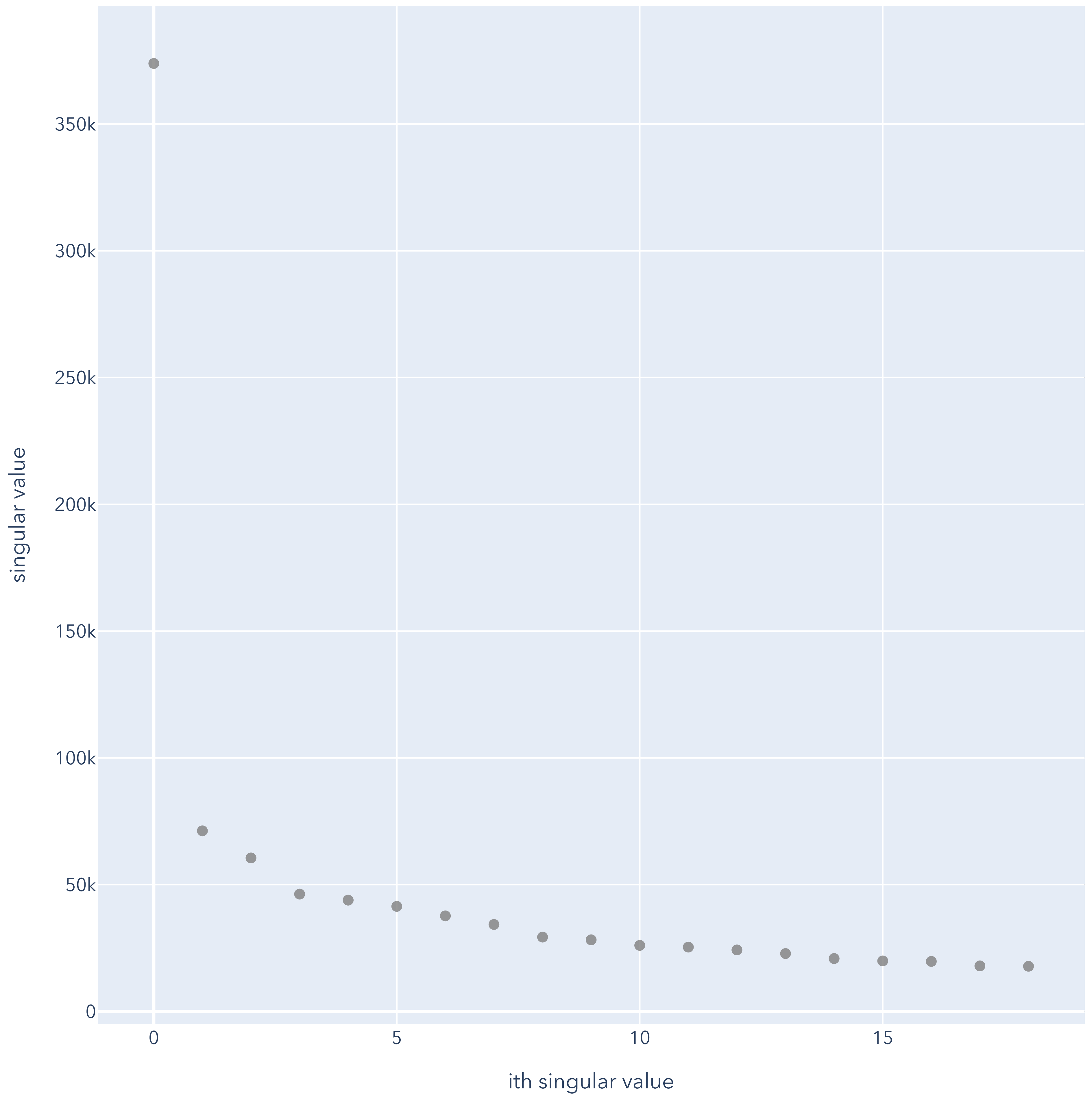
Application in Image Processing ($k = 20$)



Rank-*k* Approximation

Application in Image Processing (*k* = 5)





Least Squares

SVD and the Pseudoinverse

Regression

Setup (Example View)

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

$$\mathbf{X} = \begin{bmatrix} \leftarrow & \mathbf{x}_1^\top & \rightarrow \\ & \vdots & \\ \leftarrow & \mathbf{x}_n^\top & \rightarrow \end{bmatrix} \quad \mathbf{y} = \begin{bmatrix} y_1 \\ \vdots \\ y_n \end{bmatrix}, \text{ where } \mathbf{x}_1, \ldots, \mathbf{x}_n \in \mathbb{R}^d.$$

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

<u>Goal:</u> 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:

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

Regression

Setup (Feature View)

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

$$\mathbf{X} = \begin{bmatrix} \uparrow & & \uparrow \\ \mathbf{x}_1 & \cdots & \mathbf{x}_d \\ \downarrow & & \downarrow \end{bmatrix} \quad \mathbf{y} = \begin{bmatrix} y_1 \\ \vdots \\ y_n \end{bmatrix}, \text{ where } \mathbf{x}_1, \ldots, \mathbf{x}_d \in \mathbb{R}^n.$$

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

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:

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

Least Squares

OLS Theorem

Theorem (Ordinary Least Squares). Let $\mathbf{X} \in \mathbb{R}^{n\times d}$ and $\mathbf{y} \in \mathbb{R}^n$. Let $\hat{\mathbf{w}} \in \mathbb{R}^d$ be the least squares minimizer:

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

If $n \geq d$ and $\mathrm{rank}(\mathbf{X}) = d$, then:

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

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

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

Least Squares: SVD Perspective

Plugging in the SVD

By the full SVD, we can represent $\mathbf{X} = \mathbf{U}\mathbf{\Sigma}\mathbf{V}^\top$. How can we interpret the least squares solution now that we know the SVD?

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

Least Squares: SVD Perspective

Plugging in the SVD

By the full SVD, we can represent $\mathbf{X} = \mathbf{U}\boldsymbol{\Sigma}\mathbf{V}^\top$. How can we interpret the least squares solution now that we know the SVD?

$$\hat{\mathbf{w}} = (\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{X}^\top\mathbf{y} \; = (\mathbf{V}\boldsymbol{\Sigma}^\top\mathbf{U}^\top\mathbf{U}\boldsymbol{\Sigma}\mathbf{V}^\top)^{-1}(\mathbf{V}\boldsymbol{\Sigma}\mathbf{U}^\top)\mathbf{y} \text{ because } \mathbf{X}^\top = \mathbf{V}\boldsymbol{\Sigma}\mathbf{U}^\top$$

$$= (\mathbf{V}\boldsymbol{\Sigma}^\top\boldsymbol{\Sigma}\mathbf{V}^\top)^{-1}\mathbf{V}\boldsymbol{\Sigma}^\top\mathbf{U}^\top\mathbf{y} \text{ because } \mathbf{U}^\top\mathbf{U} = \mathbf{I}$$

$$= (\boldsymbol{\Sigma}^\top\boldsymbol{\Sigma}\mathbf{V}^\top)^{-1}\mathbf{V}^\top\mathbf{V}\boldsymbol{\Sigma}^\top\mathbf{U}^\top\mathbf{y} \text{ because } (\mathbf{A}\mathbf{B})^{-1} = \mathbf{B}^{-1}\mathbf{A}^{-1}$$

$$= (\boldsymbol{\Sigma}^\top\boldsymbol{\Sigma}\mathbf{V})^{-1}\boldsymbol{\Sigma}^\top\mathbf{U}^\top\mathbf{y} \text{ because } \mathbf{V}^\top\mathbf{V} = \mathbf{I}$$

$$= \mathbf{V}^\top(\boldsymbol{\Sigma}^\top\boldsymbol{\Sigma})^{-1}\boldsymbol{\Sigma}^\top\mathbf{U}^\top\mathbf{y} \text{ because } (\mathbf{A}\mathbf{B})^{-1} = \mathbf{B}^{-1}\mathbf{A}^{-1}$$

Pseudoinverse
Idea

Therefore, we derived:

$$\hat{\mathbf{w}} = \mathbf{V}(\boldsymbol{\Sigma}^\top \boldsymbol{\Sigma})^{-1}\boldsymbol{\Sigma}^\top \mathrm{U}^\top \mathbf{y} \quad \text{(when } n \geq d \text{ and } \mathrm{rank}(\mathbf{X}) = d\text{).}$$

Taking a closer look at the matrix $(\boldsymbol{\Sigma}^\top \boldsymbol{\Sigma})^{-1}\boldsymbol{\Sigma}^\top \in \mathbb{R}^{d \times n}$, we have:

$$(\boldsymbol{\Sigma}^\top \boldsymbol{\Sigma})^{-1}\boldsymbol{\Sigma}^\top \boldsymbol{\Sigma} = \mathbf{I}_{d \times d}.$$

In this way, $(\boldsymbol{\Sigma}^\top \boldsymbol{\Sigma})^{-1}\boldsymbol{\Sigma}^\top$ acts “like an inverse” to $\boldsymbol{\Sigma}$, though $\boldsymbol{\Sigma}$ may not be square.

Pseudoinverse

Definition

Let $\mathbf{X} \in \mathbb{R}^{n\times d}$ be a matrix, and let $\mathbf{X} = \mathbf{U}\mathbf{\Sigma}\mathbf{V}^\top$ be its full SVD.

If $n \geq d$, the matrix $\mathbf{\Sigma}^+ := (\mathbf{\Sigma}^\top\mathbf{\Sigma})^{-1}\mathbf{\Sigma}^\top \in \mathbb{R}^{d\times n}$ is the pseudoinverse of the matrix $\mathbf{\Sigma}$.

If $d > n$, the matrix $\mathbf{\Sigma}^+ := \mathbf{\Sigma}^\top(\mathbf{\Sigma}\mathbf{\Sigma}^\top)^{-1}$ is the pseudoinverse.

More generally, the matrix $\mathbf{X} \in \mathbb{R}^{n\times d}$ with full SVD $\mathbf{X} = \mathbf{U}\mathbf{\Sigma}\mathbf{V}^\top$ has the pseudoinverse:

$$\mathbf{X}^+ := \mathbf{V}\mathbf{\Sigma}^+\mathbf{U}^\top.$$

Note: If using the notation of the compact SVD, this is written differently (see PS2).

Pseudoinverse

Main Property

<u>Prop (Pseudoinverse as left/right inverse).</u> For any matrix $\mathbf{A} \in \mathbb{R}^{n \times d}$ with full SVD $\mathbf{A} = \mathbf{U}\boldsymbol{\Sigma}\mathbf{V}^\top$ and $\operatorname{rank}(\mathbf{A}) = \min\{n, d\}$, the pseudo inverse

$$\mathbf{A}^+ = \mathbf{V}\boldsymbol{\Sigma}^+\mathbf{U}^\top$$

has the following properties:

If $n = d$, then $\mathbf{A}^+$ is the *inverse*: $\mathbf{A}^+ = \mathbf{A}^{-1}$ and $\mathbf{A}^+\mathbf{A} = \mathbf{A}\mathbf{A}^+ = \mathbf{I}$.

If $n > d$, then $\mathbf{A}^+$ is a *left inverse*: $\mathbf{A}^+\mathbf{A} = \mathbf{I}_{d \times d}$.

If $d > n$, then $\mathbf{A}^+$ is a *right inverse*: $\mathbf{A}\mathbf{A}^+ = \mathbf{I}_{n \times n}$.

Pseudoinverse

Shape of $\mathbf{\Sigma}^+$

$\mathbf{\Sigma} \in \mathbb{R}^{n \times d}$ is a diagonal matrix with singular values $\sigma_1 \geq \sigma_2 \geq \ldots \geq \sigma_r \geq 0$, with $r \leq \min\{n, d\}$.

$$\underbrace{\mathbf{\Sigma} = \begin{bmatrix} \sigma_1 & 0 & \ldots & 0 \\ 0 & \sigma_2 & \ldots & 0 \\ 0 & 0 & \ddots & \vdots \\ 0 & 0 & \ldots & \sigma_d \end{bmatrix}}_{n=d} \text{ or } \underbrace{\mathbf{\Sigma} = \begin{bmatrix} \sigma_1 & 0 & \ldots & 0 \\ 0 & \sigma_2 & \ldots & 0 \\ 0 & 0 & \ddots & \vdots \\ 0 & 0 & \ldots & \sigma_d \\ 0 & 0 & \ldots & 0 \\ 0 & 0 & \ldots & 0 \\ \vdots & \vdots & \vdots & \vdots \end{bmatrix}}_{n>d} \text{ or } \underbrace{\mathbf{\Sigma} = \begin{bmatrix} \sigma_1 & 0 & \ldots & 0 & 0 & 0 & \ldots \\ 0 & \sigma_2 & \ldots & 0 & 0 & 0 & \ldots \\ 0 & 0 & \ddots & \vdots & \vdots & \vdots & \ldots \\ 0 & 0 & \ldots & \sigma_n & 0 & 0 & \ldots \end{bmatrix}}_{d>n}$$

Pseudoinverse

Shape of $\boldsymbol{\Sigma}^+$

$\boldsymbol{\Sigma} \in \mathbb{R}^{n \times d}$ is a diagonal matrix with singular values $\sigma_1 \geq \sigma_2 \geq \ldots \geq \sigma_r \geq 0$, with $r \leq \min\{n, d\}$.

$$\underbrace{\boldsymbol{\Sigma}^+ = \begin{bmatrix} 1/\sigma_1 & 0 & \ldots & 0 \\ 0 & 1/\sigma_2 & \ldots & 0 \\ 0 & 0 & \ddots & \vdots \\ 0 & 0 & \ldots & 1/\sigma_d \end{bmatrix}}_{n=d} \text{ or } \underbrace{\boldsymbol{\Sigma}^+ = \begin{bmatrix} 1/\sigma_1 & 0 & \ldots & 0 & 0 & 0 & \ldots \\ 0 & 1/\sigma_2 & \ldots & 0 & 0 & 0 & \ldots \\ 0 & 0 & \ddots & \vdots & \vdots & \vdots & \ldots \\ 0 & 0 & \ldots & 1/\sigma_d & 0 & 0 & \ldots \end{bmatrix}}_{n>d} \text{ or } \underbrace{\boldsymbol{\Sigma}^+ = \begin{bmatrix} 1/\sigma_1 & 0 & \ldots & 0 \\ 0 & 1/\sigma_2 & \ldots & 0 \\ 0 & 0 & \ddots & \vdots \\ 0 & 0 & \ldots & 1/\sigma_n \\ 0 & 0 & \ldots & 0 \\ 0 & 0 & \ldots & 0 \\ \vdots & \vdots & \vdots & \vdots \end{bmatrix}}_{d>n}$$

Least Squares: SVD Perspective

Using the pseudoinverse

Let $\mathbf{X} \in \mathbb{R}^{n \times d}$ and $\mathbf{y} \in \mathbb{R}^n$. Let $\hat{\mathbf{w}} \in \mathbb{R}^d$ be the least squares minimizer:

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

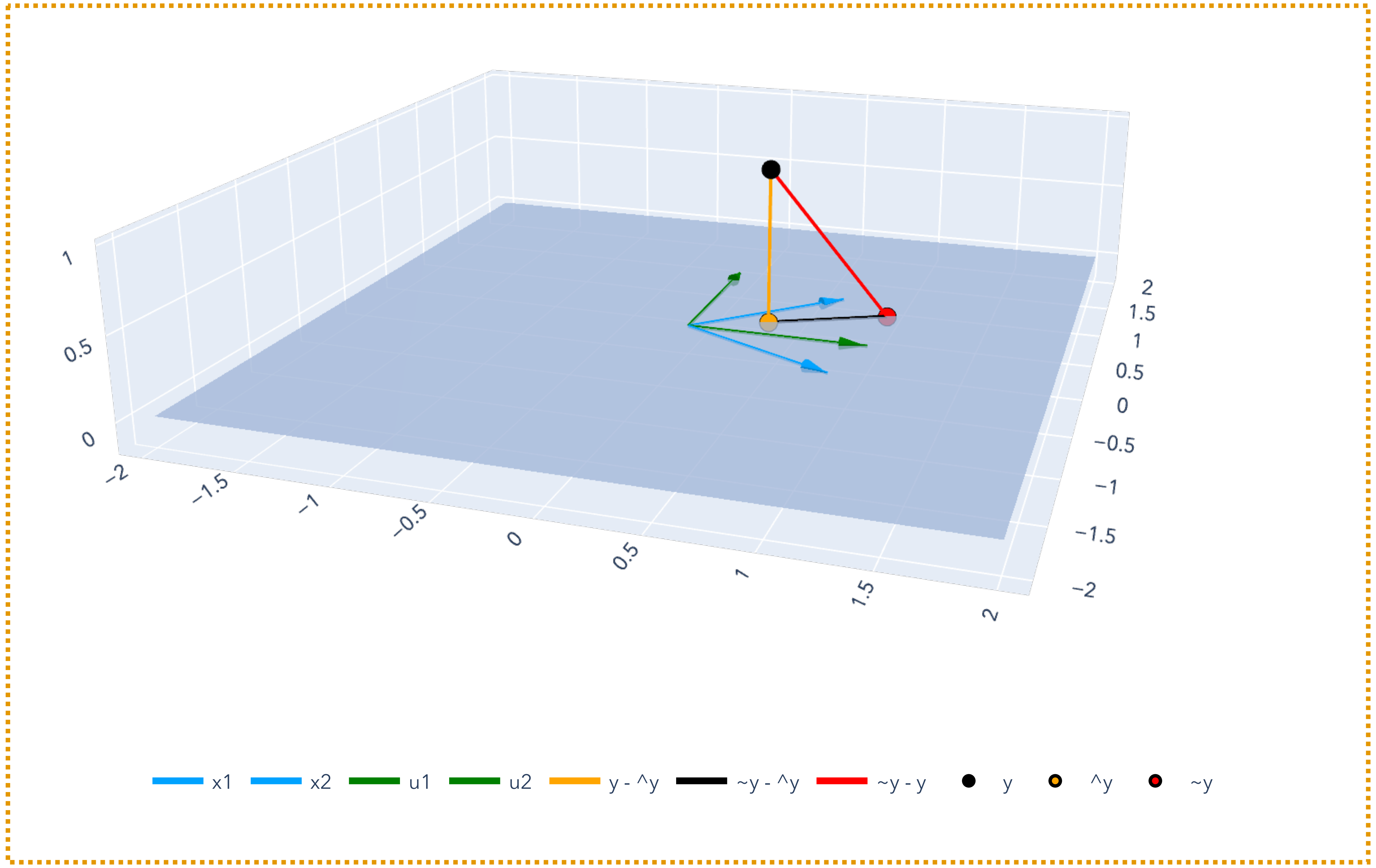
<u>Theorem (Ordinary Least Squares).</u>

If $n \geq d$ and $\operatorname{rank}(\mathbf{X}) = d$, then:

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

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

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



Least Squares: SVD Perspective

Using the pseudoinverse

Let $\mathbf{X} \in \mathbb{R}^{n \times d}$ and $\mathbf{y} \in \mathbb{R}^n$. Let $\hat{\mathbf{w}} \in \mathbb{R}^d$ be the least squares minimizer:

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

If $n = d$ and $\mathrm{rank}(\mathbf{X}) = d$, then we are just solving the system $\mathbf{X}\mathbf{w} = \mathbf{y}$, and:

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

We solved this by the principle of least squares because, when $n > d$, we don't have an inverse. We are solving for an *approximation*:

$$\mathbf{X}\mathbf{w} \approx \mathbf{y}.$$

Least Squares: SVD Perspective

Using the pseudoinverse

We solved this by the principle of least squares because, when $n > d$, we don't have an inverse. We are solving for an *approximation*:

$$\mathbf{X}\mathbf{w} \approx \mathbf{y}.$$

We don't have an inverse – but now we have a *pseudoinverse*:

$$\mathbf{X}^{+}\mathbf{X}\mathbf{w} \approx \mathbf{X}^{+}\mathbf{y} \implies \hat{\mathbf{w}} = \mathbf{X}^{+}\mathbf{y} = \mathbf{V}\boldsymbol{\Sigma}^{+}\mathbf{U}^{\top}\mathbf{y}.$$

Least Squares: SVD Perspective

Main Theorem (with pseudoinverse)

Let $\mathbf{X} \in \mathbb{R}^{n \times d}$ and $\mathbf{y} \in \mathbb{R}^n$. Let $\hat{\mathbf{w}} \in \mathbb{R}^d$ be the least squares minimizer:

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

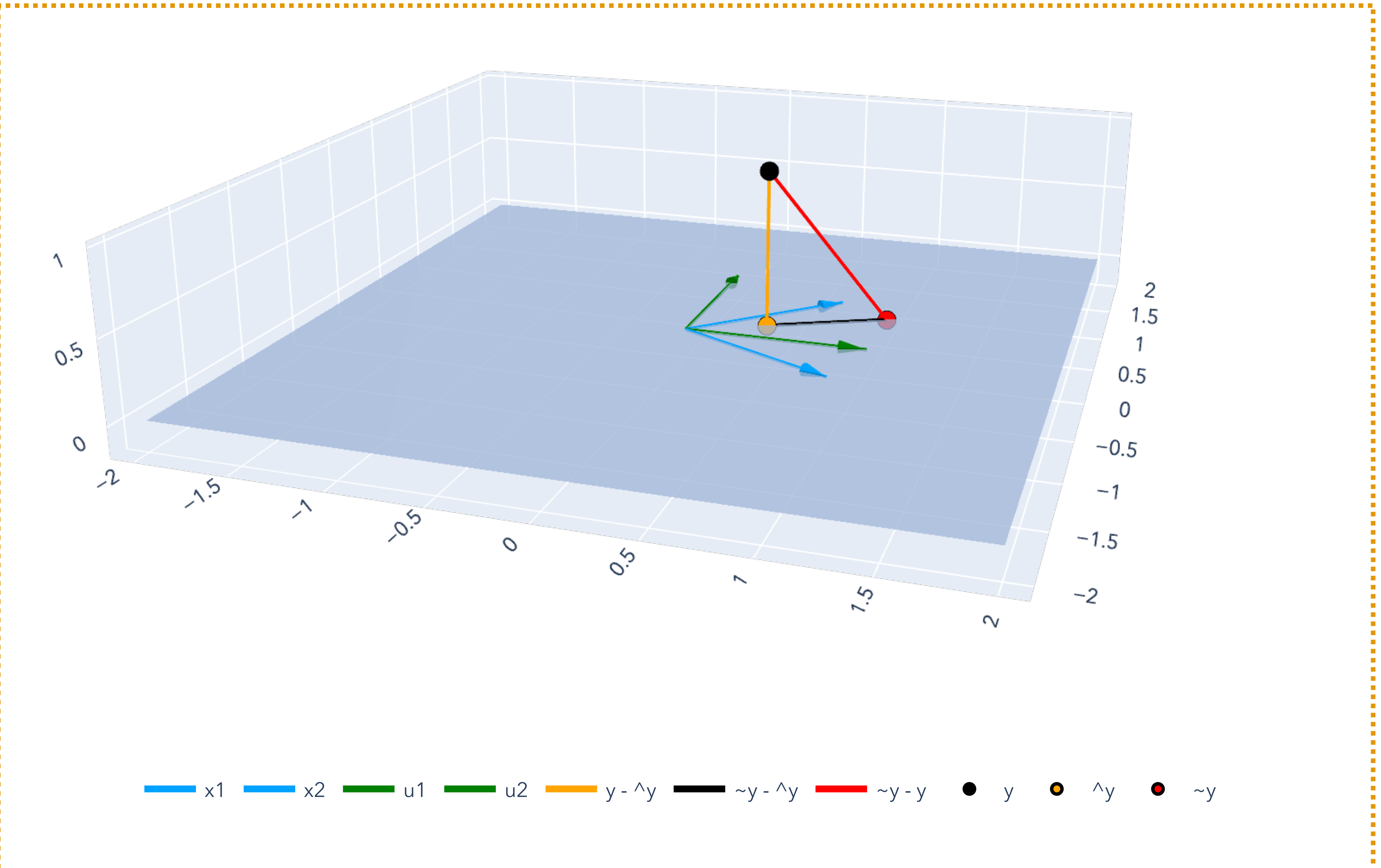
<u>Theorem (OLS with pseudoinverse).</u>

If $n \geq d$ and $\operatorname{rank}(\mathbf{X}) = d$, then:

$$\hat{\mathbf{w}} = \mathbf{X}^+\mathbf{y} = \mathbf{V}\boldsymbol{\Sigma}^+\mathbf{U}^\top\mathbf{y}.$$

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

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



Least Squares with $d \geq n$

Review: Systems of Linear Equations

So far, we've considered the case where $\mathbf{X} \in \mathbb{R}^{n \times d}$, $n \geq d$, and $\text{rank}(\mathbf{X}) = d$.

In general, our goal is to solve the system of linear equations:

$$\mathbf{X}\mathbf{w} = \mathbf{y}.$$

We know that there are three scenarios, if $\mathbf{X}$ is full rank (i.e., $\text{rank}(\mathbf{X}) = \min\{n, d\}$)…

If $n = d$, then number of equations = number of unknowns. *One unique solution:* $\hat{\mathbf{w}} = \mathbf{X}^{-1}\mathbf{y}$.

If $n > d$, then number of equations > number of unknowns. *One unique (approximate) solution:* $\hat{\mathbf{w}} = \mathbf{X}^{+}\mathbf{y}$.

If $d > n$, then number of unknowns > number of equations. *Infinitely many solutions!*

Systems of Linear Equations

Example: no solutions

In general, our goal is to solve the system of linear equations:

$$\mathbf{X}\mathbf{w} = \mathbf{y}.$$

Consider the system:

$$\begin{bmatrix} 1 & -1 \\ -1 & 1 \end{bmatrix} \begin{bmatrix} w_1 \\ w_2 \end{bmatrix} = \begin{bmatrix} 2 \\ 2 \end{bmatrix}$$

Systems of Linear Equations

Example: one unique solution, $n = d$

In general, our goal is to solve the system of linear equations:

$$\mathbf{X}\mathbf{w} = \mathbf{y}.$$

Consider the system:

$$\begin{bmatrix} 2 & 1 \\ 2 & -1 \end{bmatrix} \begin{bmatrix} w_1 \\ w_2 \end{bmatrix} = \begin{bmatrix} 3 \\ 3 \end{bmatrix}$$

Systems of Linear Equations

Example: one unique solution, $n > d$

In general, our goal is to solve the system of linear equations:

$$\mathbf{X}\mathbf{w} = \mathbf{y}.$$

Consider the system:

$$\begin{bmatrix} 2 & 1 \\ 2 & -1 \\ 4 & -2 \end{bmatrix} \begin{bmatrix} w_1 \\ w_2 \end{bmatrix} = \begin{bmatrix} 3 \\ 3 \\ 3 \end{bmatrix}$$

Systems of Linear Equations

Example: infinitely many solutions, $d > n$

In general, our goal is to solve the system of linear equations:

$$\mathbf{X}\mathbf{w} = \mathbf{y}.$$

Consider the system:

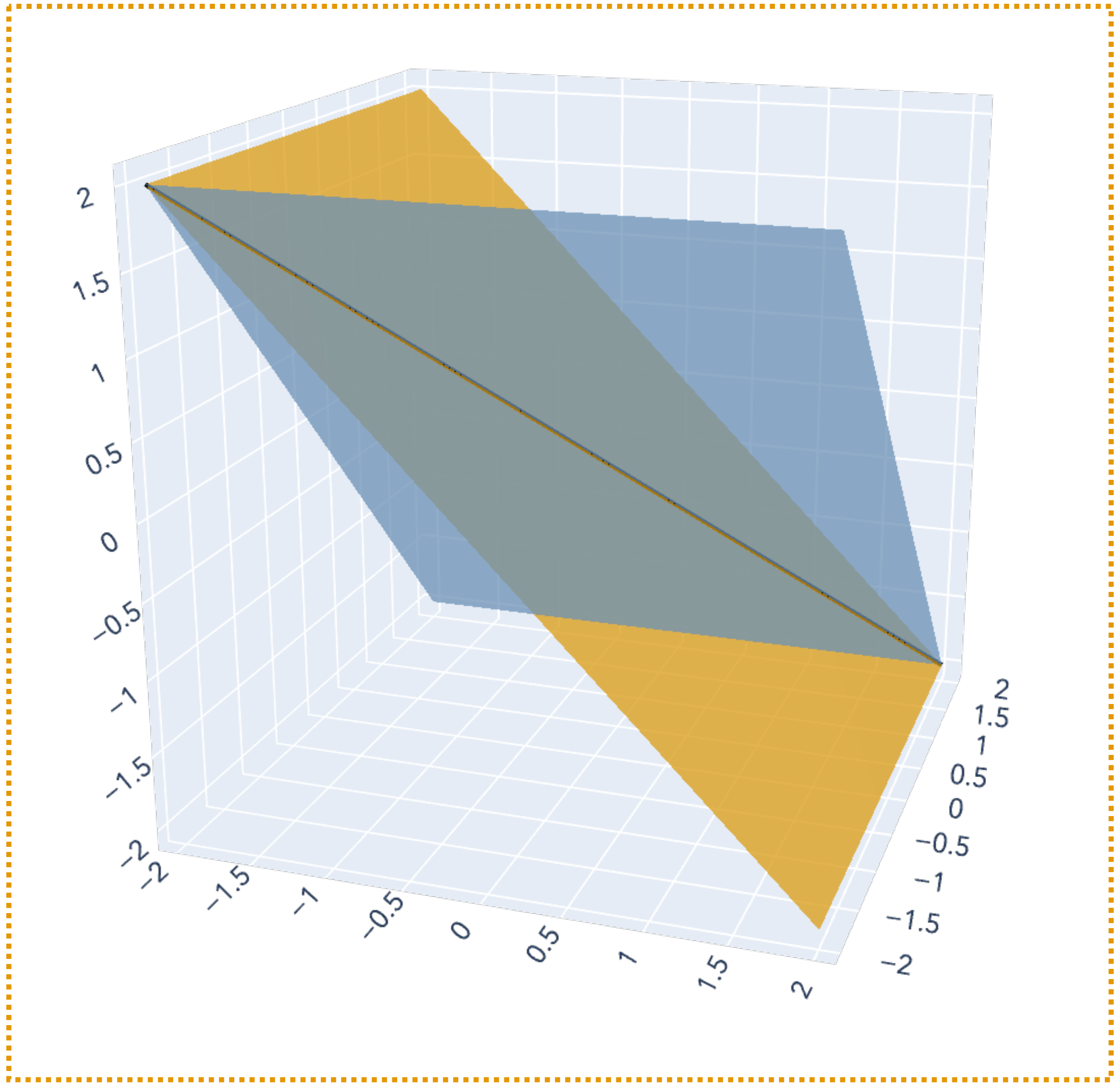
$$\begin{bmatrix} 2 & 1 & 1 \\ 2 & -1 & 0 \end{bmatrix} \begin{bmatrix} w_1 \\ w_2 \\ w_3 \end{bmatrix} = \begin{bmatrix} 3 \\ 3 \end{bmatrix}$$

Least Squares with $d > n$

Review: Systems of Linear Equations

When the number of equations < number of unknowns…

Example. $d = 3, n = 2$



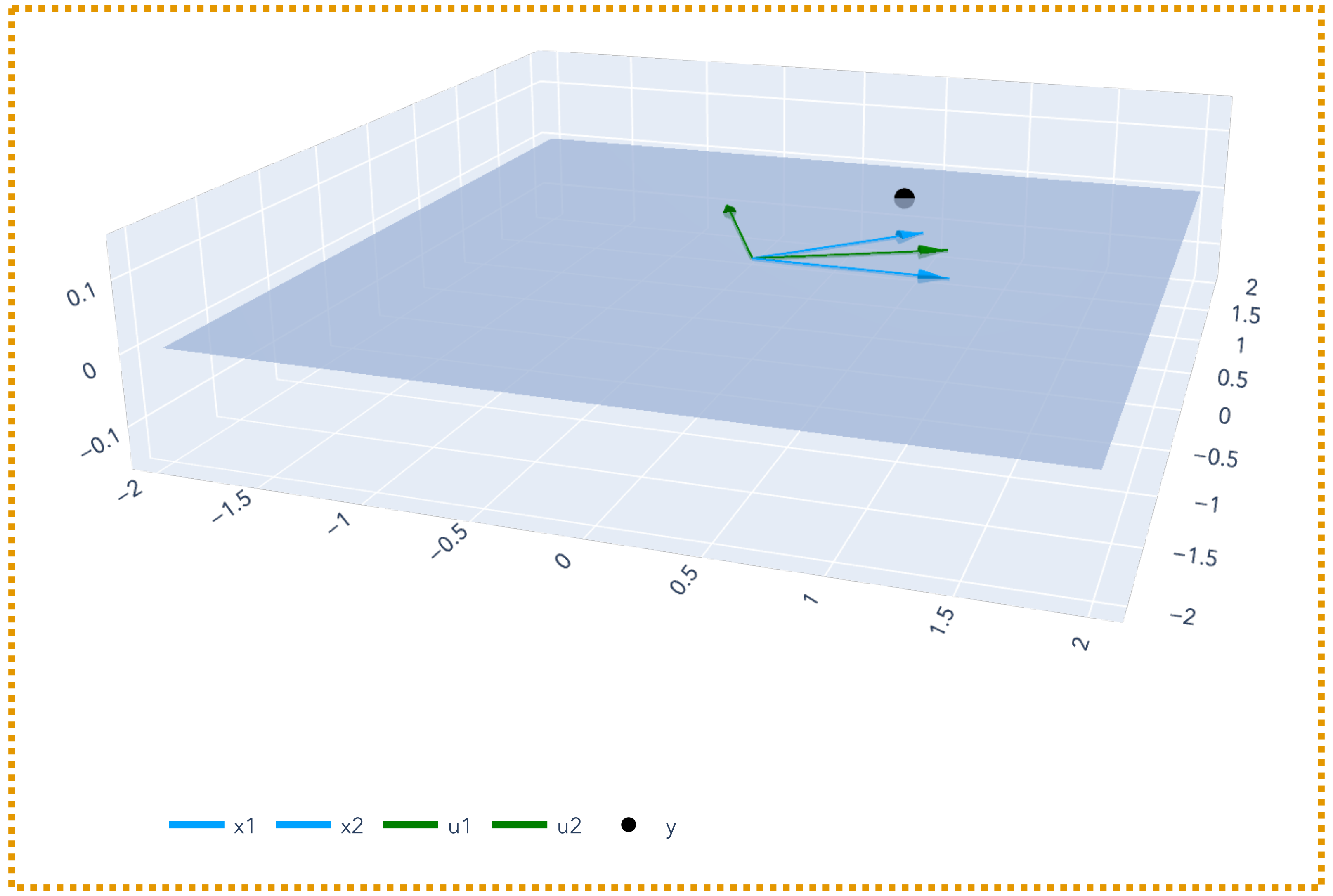
Least Squares with $d > n$

Problem Statement

Let $\mathbf{X} \in \mathbb{R}^{n \times d}$, let $d > n$, and let $\mathrm{rank}(\mathbf{X}) = n$. We want to solve the system of linear equations:

$$\mathbf{X}\mathbf{w} = \mathbf{y}.$$

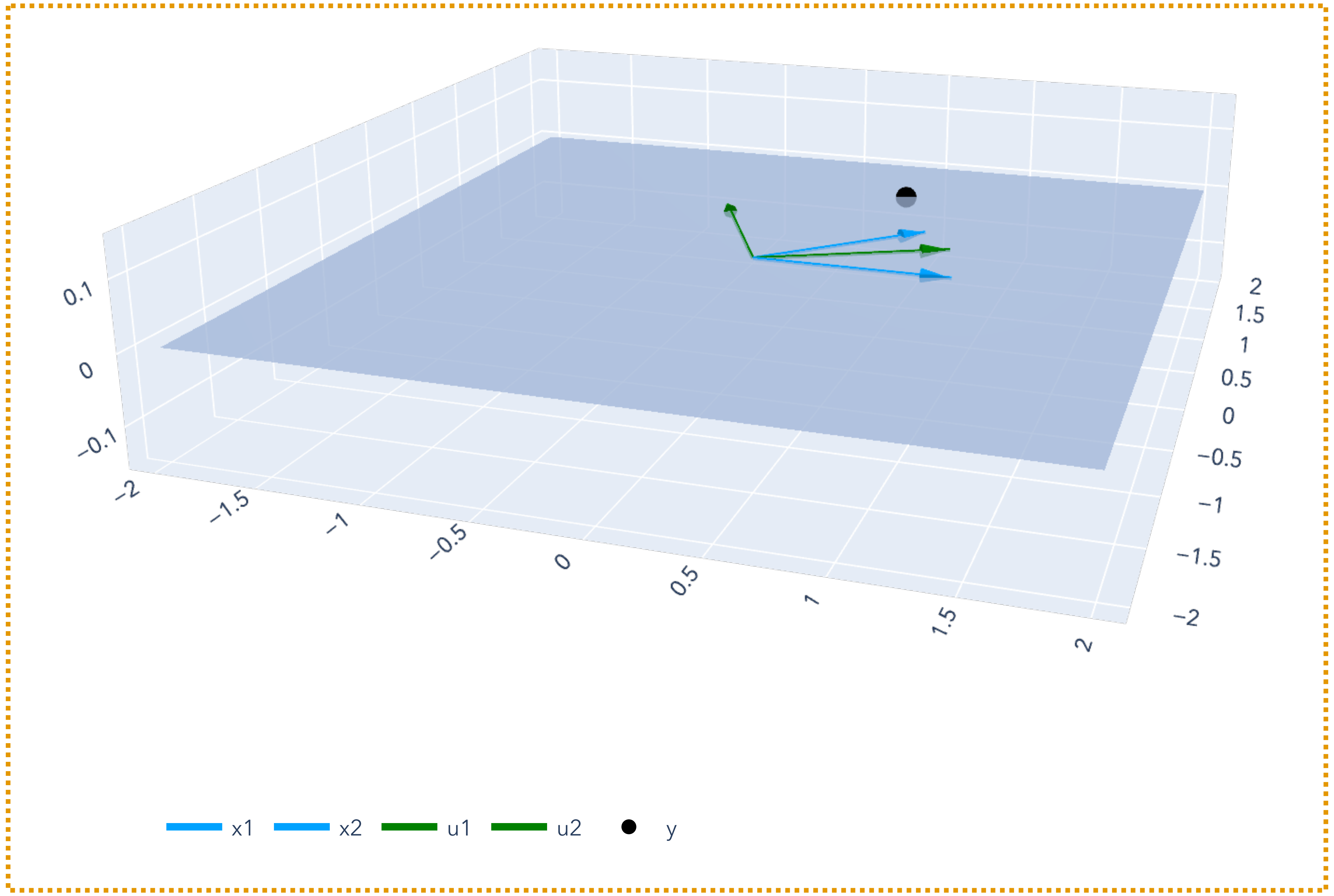
Because $\mathrm{rank}(\mathbf{X}) = n$, infinitely many *exact* solutions exist. Which to choose?



Least Squares with $d > n$

Using the Pseudoinverse

There are now infinitely many $\hat{\mathbf{w}} \in \mathbb{R}^d$ such that $\mathbf{X}\hat{\mathbf{w}} = \mathbf{y}$. Which $\hat{\mathbf{w}}$ to pick?



Pseudoinverse

Main Property

<u>Prop (Pseudoinverse as left/right inverse).</u> For any matrix $\mathbf{A} \in \mathbb{R}^{n \times d}$ with full SVD $\mathbf{A} = \mathbf{U}\boldsymbol{\Sigma}\mathbf{V}^\top$ and $\operatorname{rank}(\mathbf{A}) = \min\{n, d\}$, the pseudo inverse

$$\mathbf{A}^+ = \mathbf{V}\boldsymbol{\Sigma}^+\mathbf{U}^\top$$

has the following properties:

If $n = d$, then $\mathbf{A}^+$ is the *inverse*: $\mathbf{A}^+ = \mathbf{A}^{-1}$ and $\mathbf{A}^+\mathbf{A} = \mathbf{A}\mathbf{A}^+ = \mathbf{I}$.

If $n > d$, then $\mathbf{A}^+$ is a *left inverse*: $\mathbf{A}^+\mathbf{A} = \mathbf{I}_{d \times d}$.

If $d > n$, then $\mathbf{A}^+$ is a *right inverse*: $\mathbf{A}\mathbf{A}^+ = \mathbf{I}_{n \times n}$.

Least Squares with $d > n$

Using the Pseudoinverse

Let $\mathbf{X} \in \mathbb{R}^{n \times d}$ have the full SVD $\mathbf{X} = \mathbf{U}\mathbf{\Sigma}\mathbf{V}^\top$.

Choose $\hat{\mathbf{w}} = \mathbf{X}^+\mathbf{y} = \mathbf{V}\mathbf{\Sigma}^+\mathbf{U}^\top\mathbf{y}$ to use the pseudoinverse.

Least Squares with $d > n$

Using the Pseudoinverse

Let $\mathbf{X} \in \mathbb{R}^{n \times d}$ have the full SVD $\mathbf{X} = \mathbf{U}\mathbf{\Sigma}\mathbf{V}^\top$.

Choose $\hat{\mathbf{w}} = \mathbf{X}^+\mathbf{y} = \mathbf{V}\mathbf{\Sigma}^+\mathbf{U}^\top\mathbf{y}$ to use the pseudoinverse.

Then, $\hat{\mathbf{w}} \in \mathbb{R}^d$ is a solution:

$$\mathbf{X}\hat{\mathbf{w}} = \mathbf{X}\mathbf{X}^+\mathbf{y} = \mathbf{I}_{n \times n}\mathbf{y} = \mathbf{y},$$

where $\mathbf{X}^+ \in \mathbb{R}^{d \times n}$ is a right inverse by the previous property.

Least Squares with $d > n$

Theorem: Minimum norm solution

<u>Theorem (Minimum norm least squares solution).</u> Let $\mathbf{X} \in \mathbb{R}^{n \times d}$, let $d > n$, and let $\mathrm{rank}(\mathbf{X}) = n$. Then, $\hat{\mathbf{w}} = \mathbf{X}^{+}\mathbf{y} = \mathbf{V}\mathbf{\Sigma}^{+}\mathbf{U}^{\top}\mathbf{y}$ is the exact solution $\mathbf{X}\hat{\mathbf{w}} = \mathbf{y}$ with smallest Euclidean norm:

$$\|\mathbf{w}\|^2 \geq \|\hat{\mathbf{w}}\|^2 \text{ for all } \mathbf{w} \in \mathbb{R}^d \text{ such that } \mathbf{X}\mathbf{w} = \mathbf{y}.$$

Least Squares with $d > n$

Theorem: Minimum norm solution

<u>Theorem (Minimum norm least squares solution).</u> Let $\mathbf{X} \in \mathbb{R}^{n \times d}$, let $d > n$, and let $\operatorname{rank}(\mathbf{X}) = n$. Then, $\hat{\mathbf{w}} = \mathbf{X}^+\mathbf{y} = \mathbf{V}\boldsymbol{\Sigma}^+\mathbf{U}^\top\mathbf{y}$ is the exact solution $\mathbf{X}\hat{\mathbf{w}} = \mathbf{y}$ with smallest Euclidean norm:

$$\|\mathbf{w}\|^2 \geq \|\hat{\mathbf{w}}\|^2 \text{ for all } \mathbf{w} \in \mathbb{R}^d \text{ such that } \mathbf{X}\mathbf{w} = \mathbf{y}.$$

Proof. Consider any arbitrary $\mathbf{w} \in \mathbb{R}^d$ such that $\mathbf{X}\mathbf{w} = \mathbf{y}$.

$$\|\mathbf{w}\|^2 = \|(\mathbf{w} - \hat{\mathbf{w}}) + \hat{\mathbf{w}}\|^2 = \|\mathbf{w} - \hat{\mathbf{w}}\|^2 - 2(\mathbf{w} - \hat{\mathbf{w}})^\top\hat{\mathbf{w}} + \|\hat{\mathbf{w}}\|^2$$

$$(\mathbf{w} - \hat{\mathbf{w}})^\top\hat{\mathbf{w}} = (\mathbf{w} - \hat{\mathbf{w}})^\top\underbrace{\mathbf{X}^\top(\mathbf{X}\mathbf{X}^\top)^{-1}}_{\mathbf{X}^+ \text{ if } d > n}\mathbf{y} = \underbrace{(\mathbf{X}\mathbf{w} - \mathbf{X}\hat{\mathbf{w}})^\top(\mathbf{X}\mathbf{X}^\top)^{-1}\mathbf{y} = 0}_{\text{because both } \mathbf{w} \text{ and } \hat{\mathbf{w}} \text{ are exact solutions!}}$$

Therefore: $\|\mathbf{w}\|^2 = \|\mathbf{w} - \hat{\mathbf{w}}\|^2 + \|\hat{\mathbf{w}}\|^2 \implies \|\mathbf{w}\|^2 \geq \|\hat{\mathbf{w}}\|^2$.

Least Squares: SVD Perspective

Unified Picture

We want to solve $\mathbf{X}\mathbf{w} = \mathbf{y}$.

If $n = d$ and $\text{rank}(\mathbf{X}) = d\ldots$

We can solve exactly.

Choose

$$\hat{\mathbf{w}} = \mathbf{X}^{-1}\mathbf{y},$$

which is an exact solution.

If $n > d$ and $\text{rank}(\mathbf{X}) = d\ldots$

We approximate by least squares:

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

Choose

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

the best approximate solution:

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

If $n < d$ and $\text{rank}(\mathbf{X}) = n\ldots$

We can solve exactly, but there are infinitely many solutions.

Choose

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

the minimum norm (exact) solution:

$$\|\hat{\mathbf{w}}\|^2 \leq \|\mathbf{w}\|^2.$$

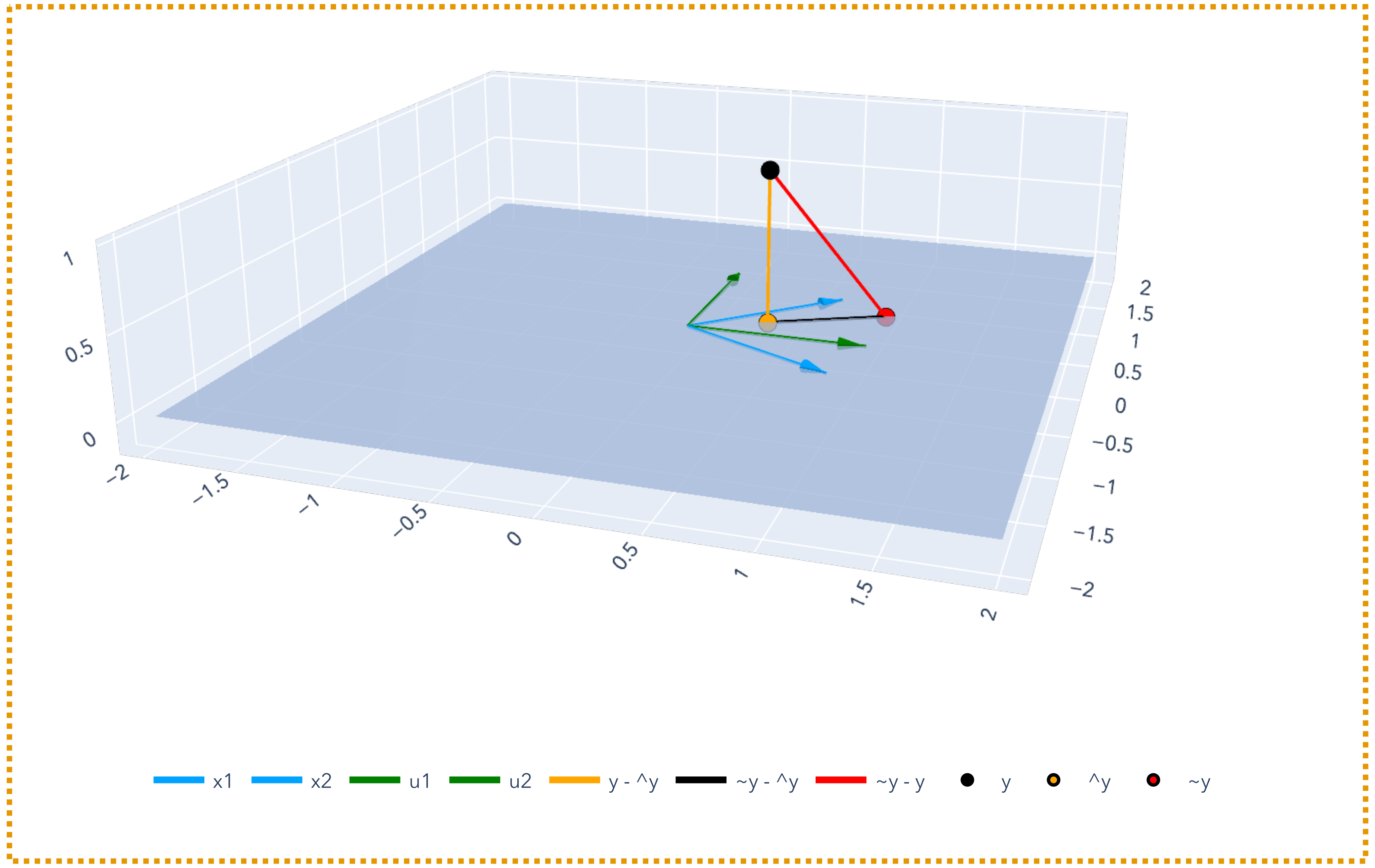
Least Squares: SVD Perspective

Unified Picture

We want to solve $\mathbf{X}\mathbf{w} = \mathbf{y}$.

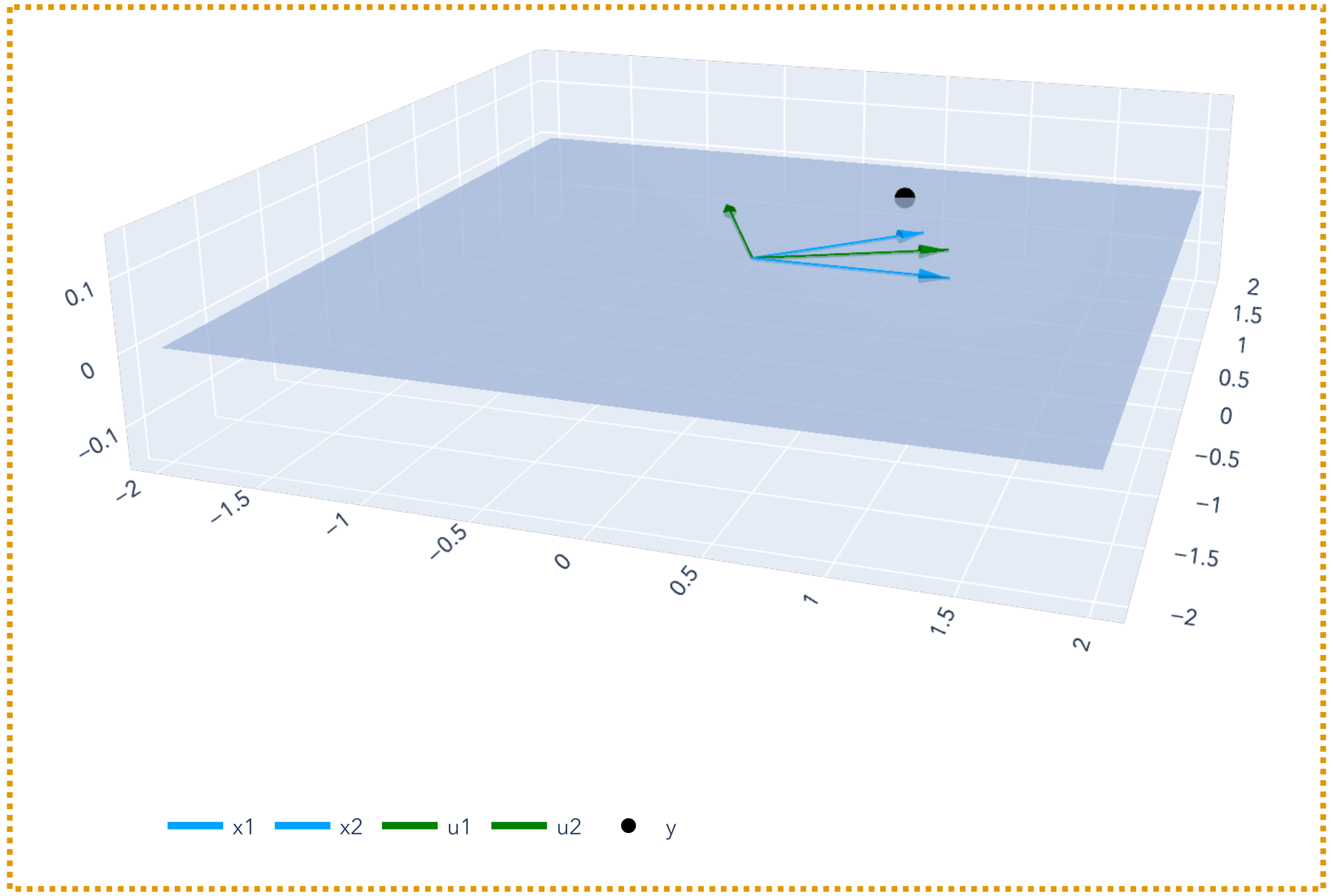
If $n > d$ and $\mathrm{rank}(\mathbf{X}) = d\ldots$

We approximate by least squares.



If $n < d$ and $\mathrm{rank}(\mathbf{X}) = n\ldots$

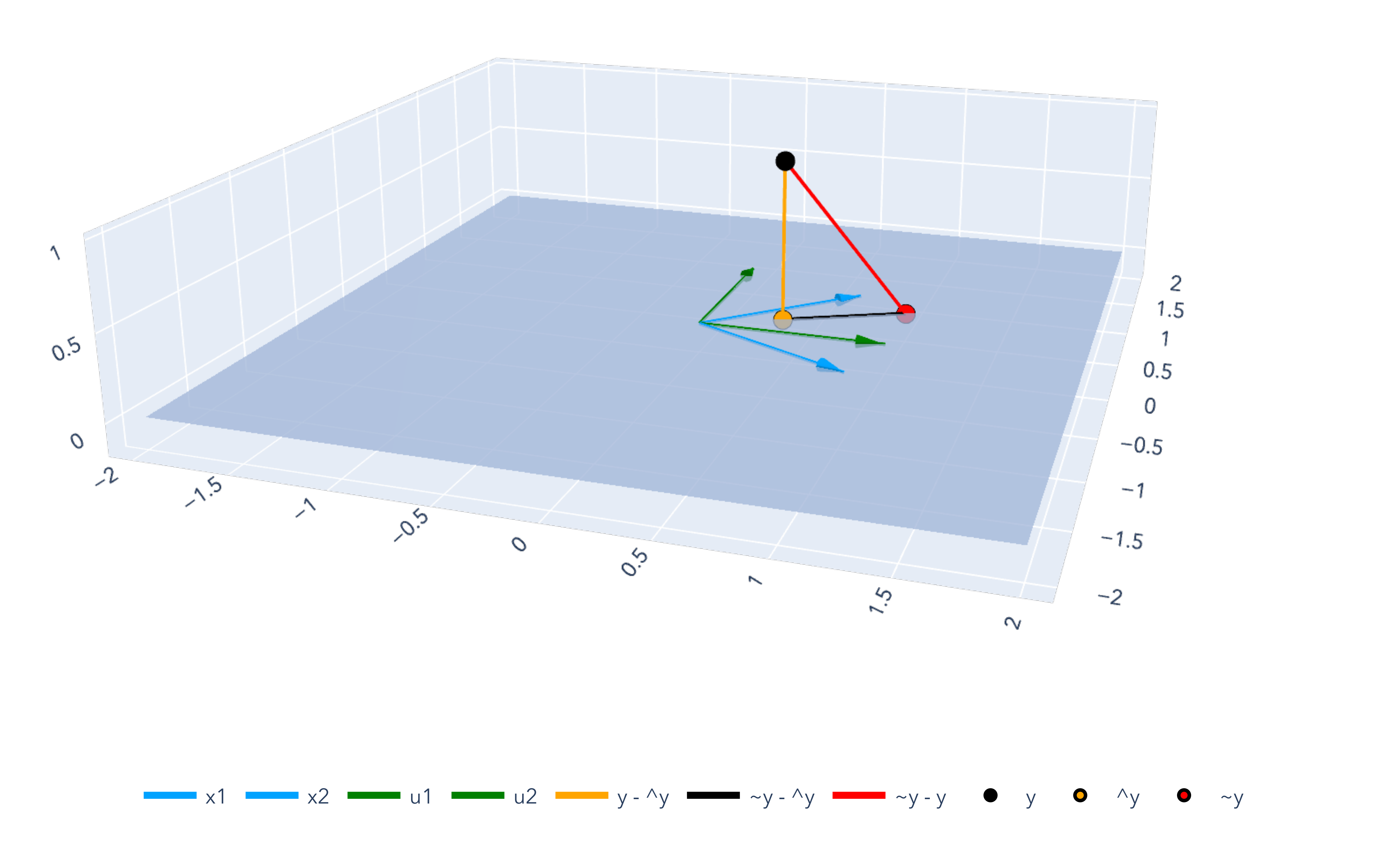
We can solve exactly, but there are infinitely many solutions.

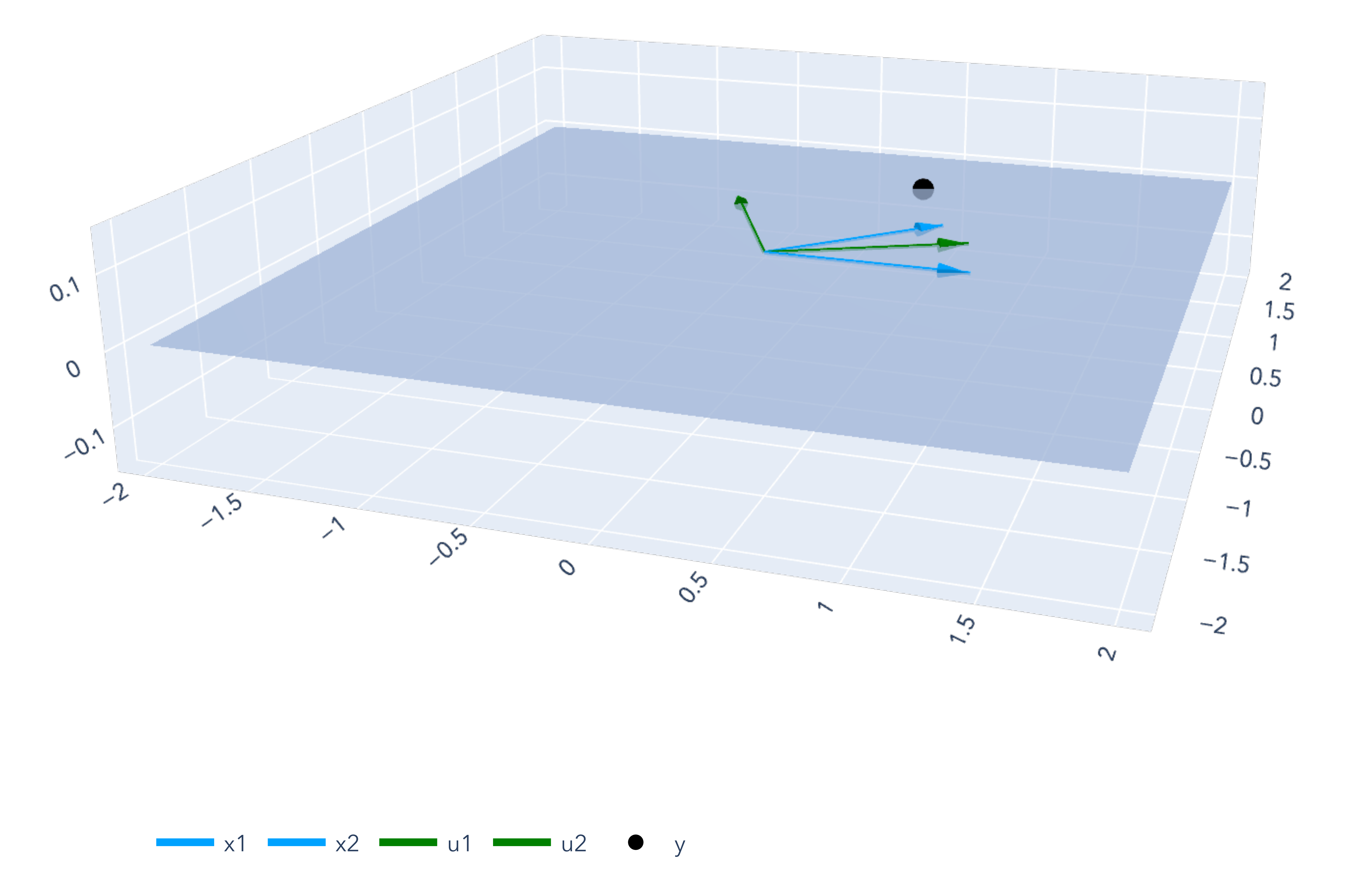


Recap

Lesson Overview

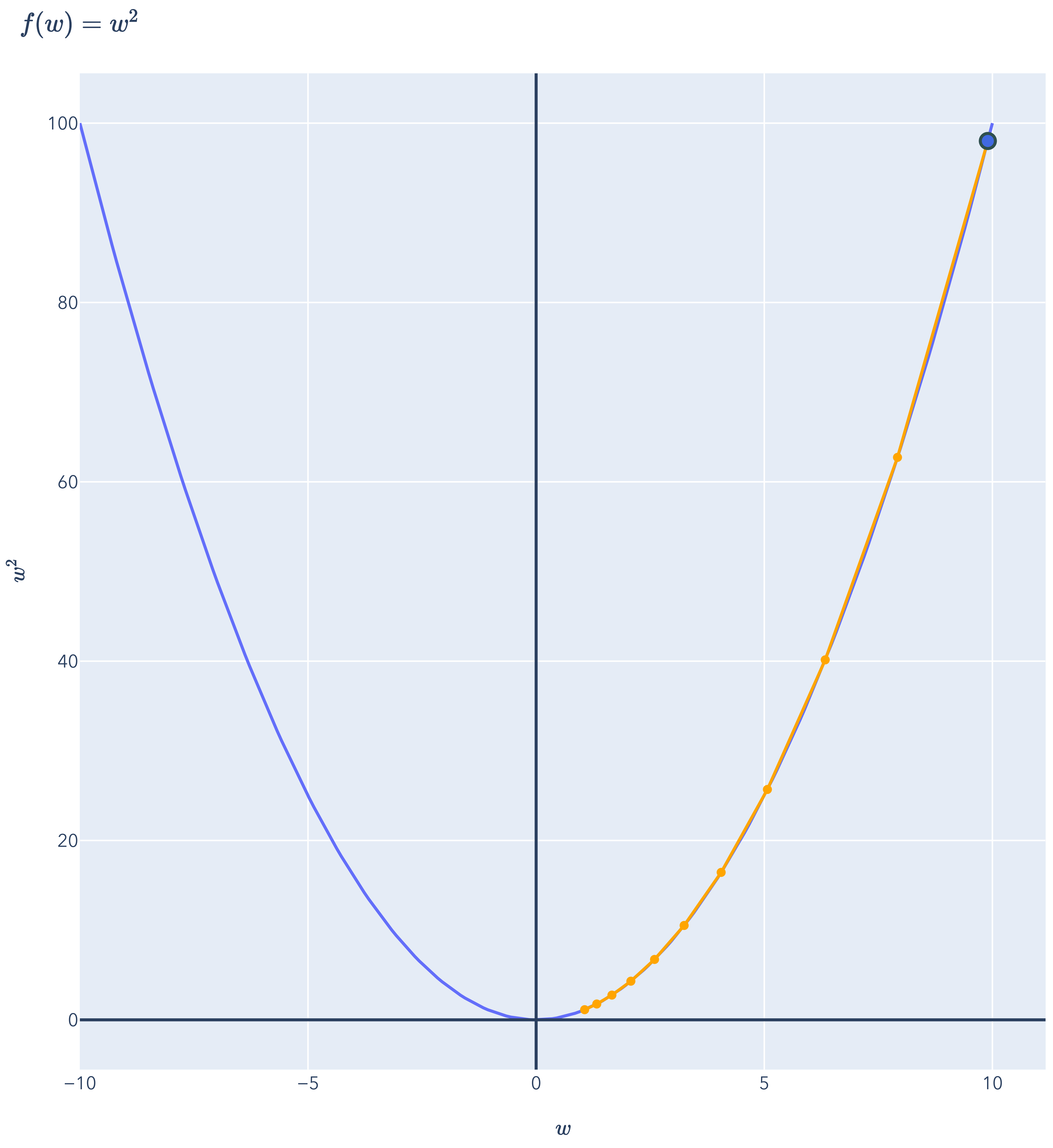
Big Picture: Least Squares

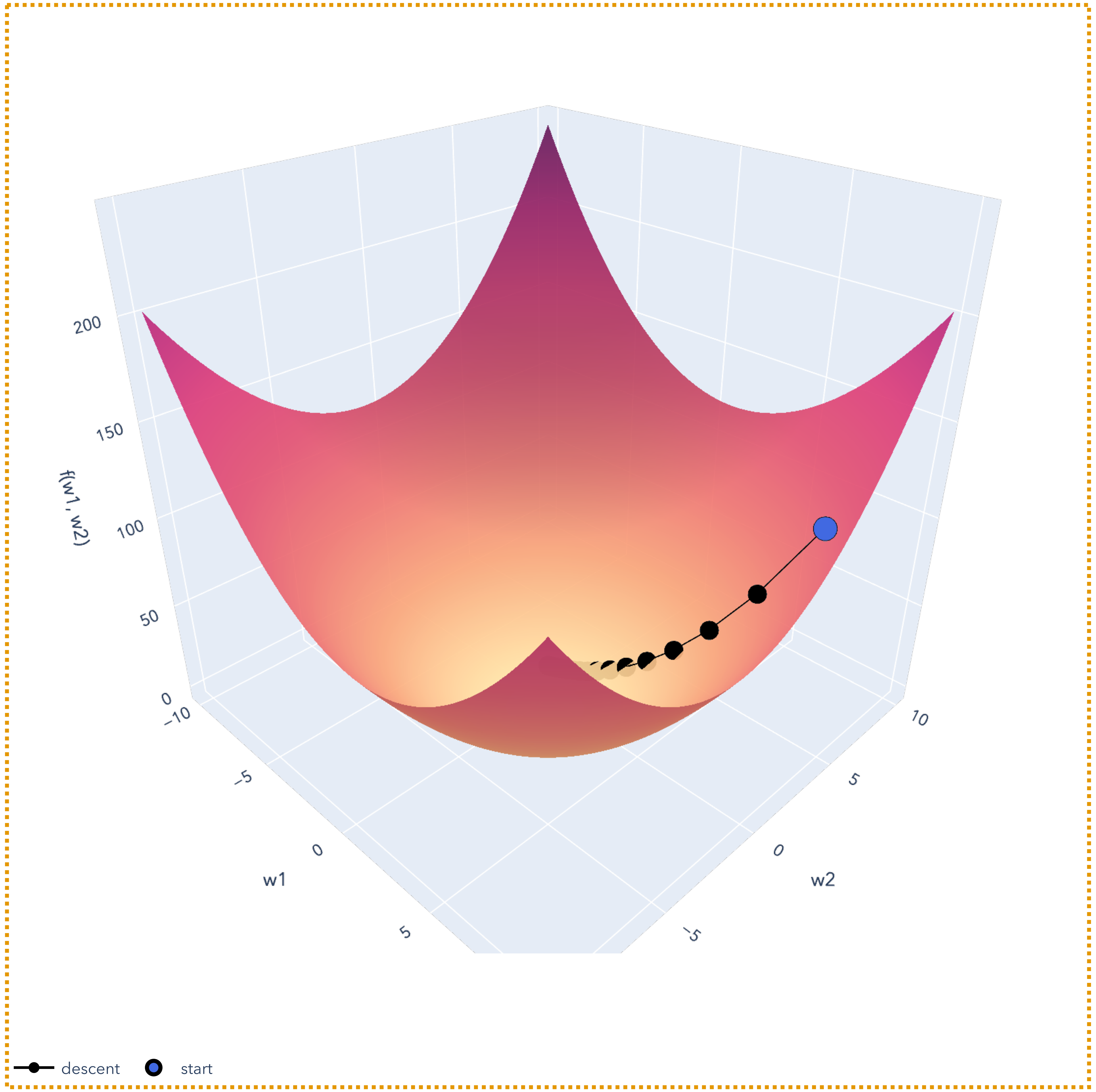




Lesson Overview

Big Picture: Gradient Descent





Lesson Overview

Big Picture: Singular Value Decomposition (SVD)