Math for Machine Learning

Week 2.1: Singular Value Decomposition

By: Samuel Deng

Logistics & Announcements

- DUE 11:59 PM tonight (Mon. Jul 8): Project 1st Reflection.
  → Pick a Paper, follow the instructions under "Project"
- DUE 11:59 PM (Thurs.): PS1.
- PS2 Released Tues. (~ Noon)

6 Late Days.

BREAKS: 10 minutes / 5 minutes each.

PollEv.com/samdeng.

Lesson Overview

MATRICES AS LINEAR TRANSFORMATIONS

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

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

$n \geq d$

$\hat{w} = (X^TX)^{-1}X^Ty$

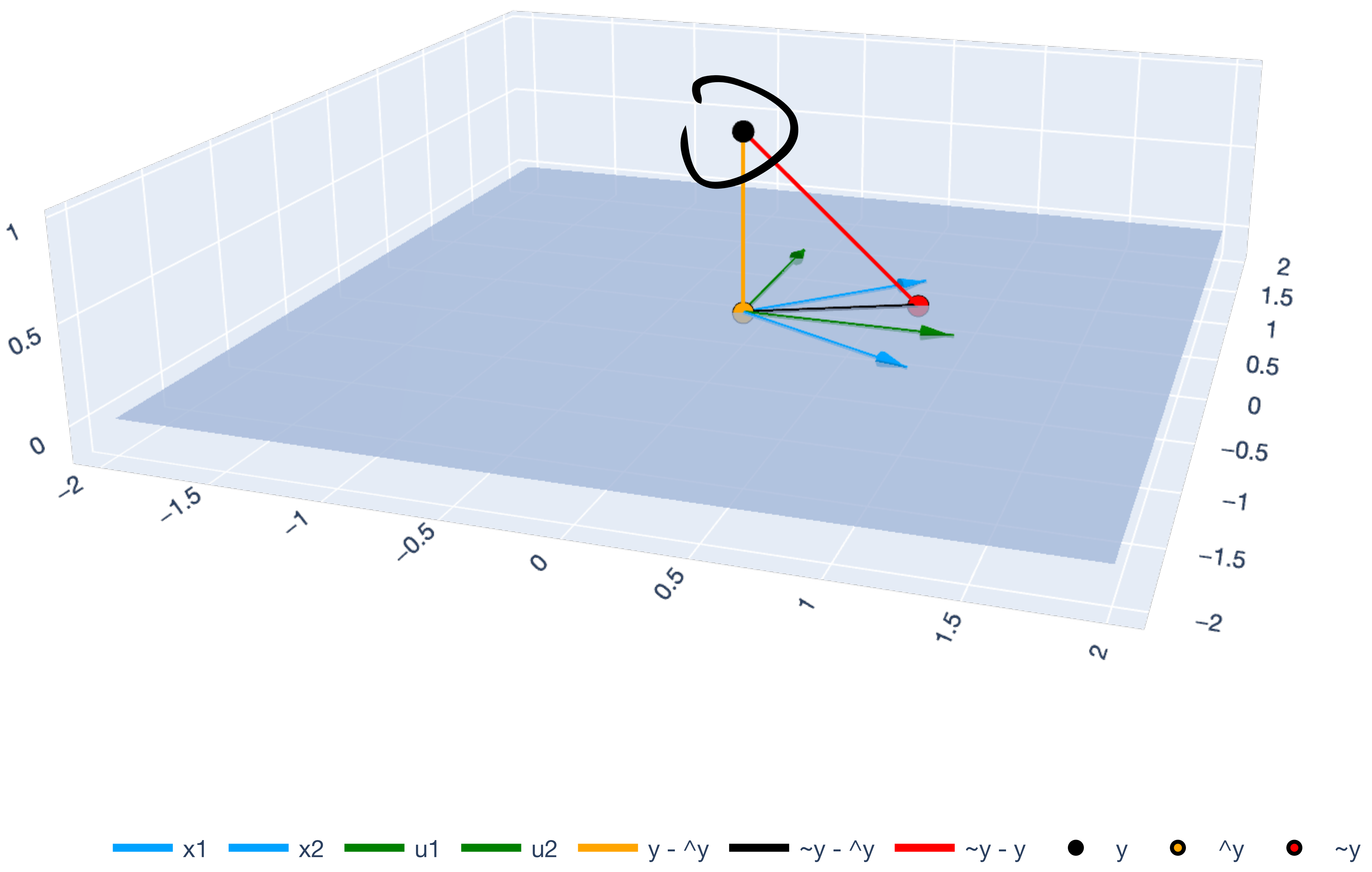
$\boxed{X\hat{w} \approx y}$

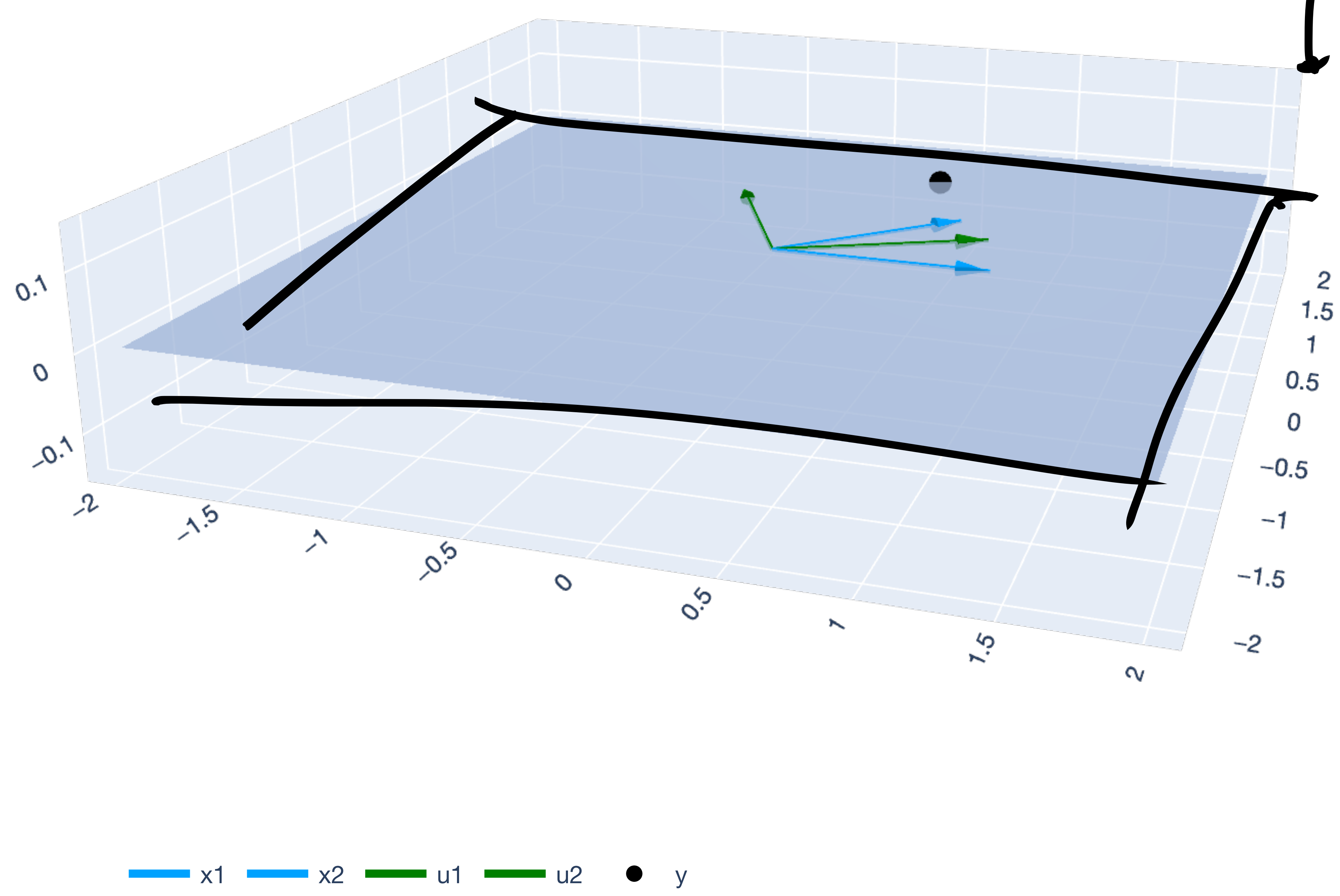
$\boxed{Xw = y}$

?

$\boxed{d \geq n}$

$\|\hat{w}\| \leq \|w\|$ $\forall w \in \mathbb{R}^d$



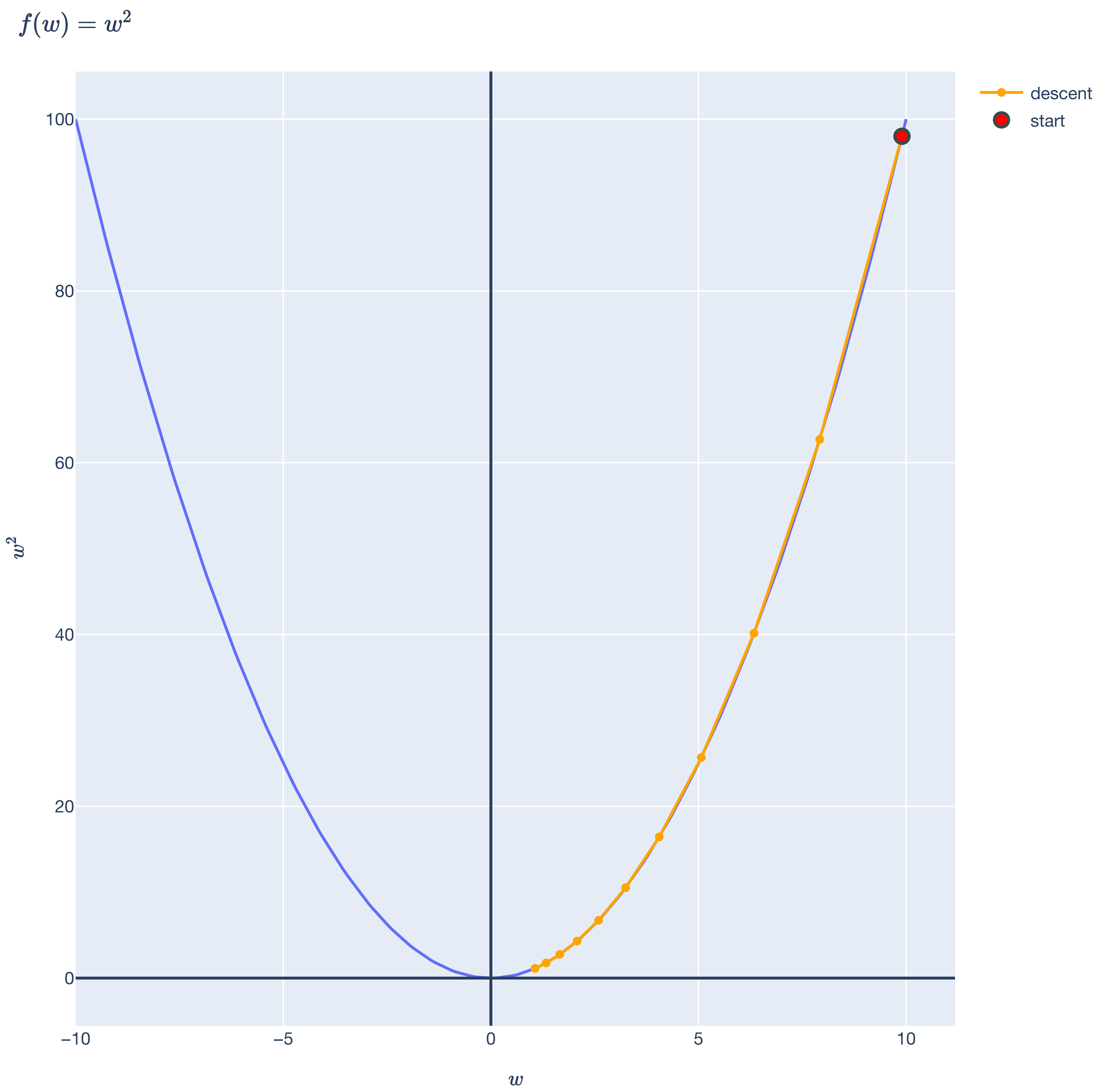


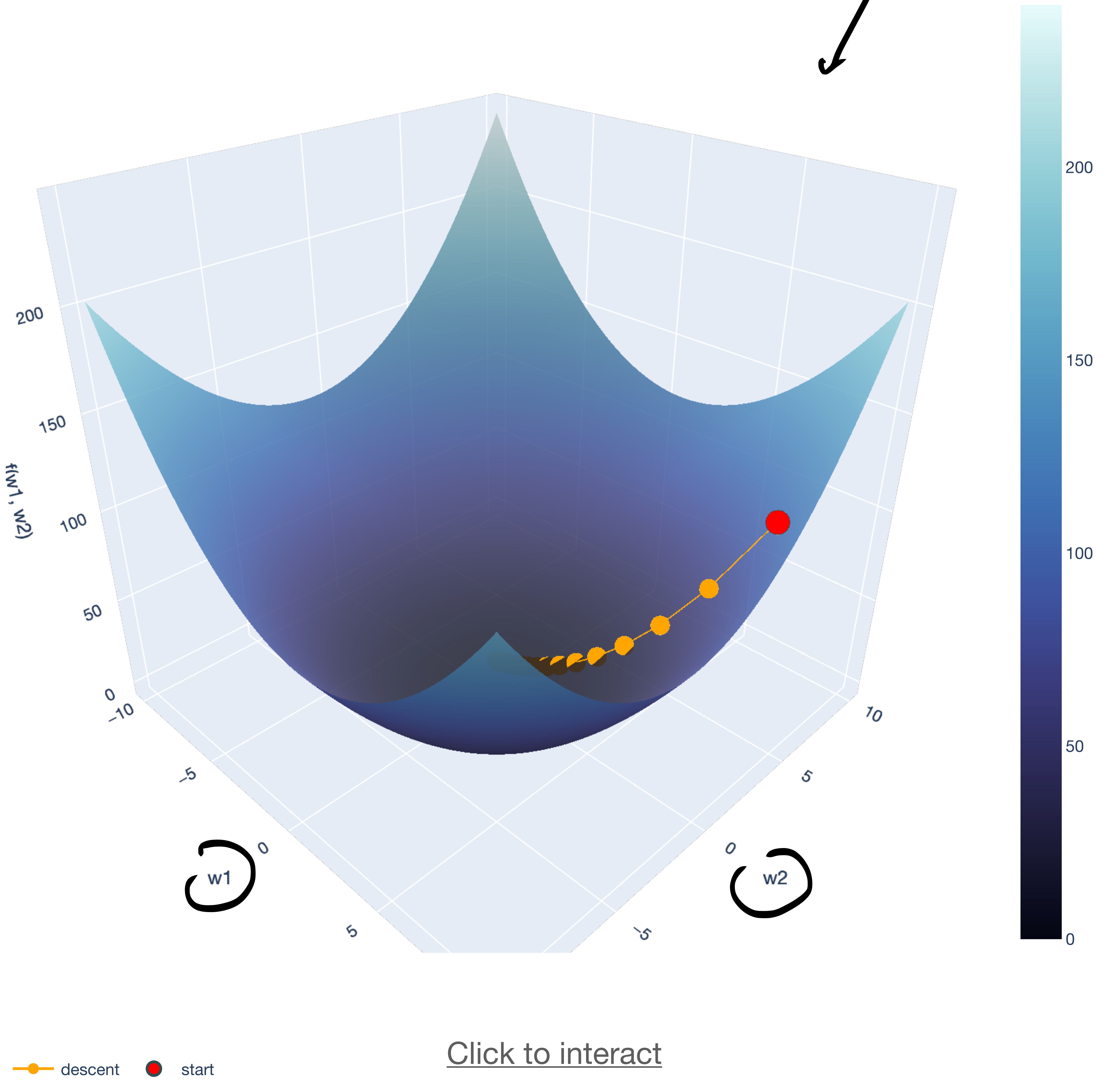
Lesson Overview

Big Picture: Gradient Descent

$w \in \mathbb{R}^2$

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

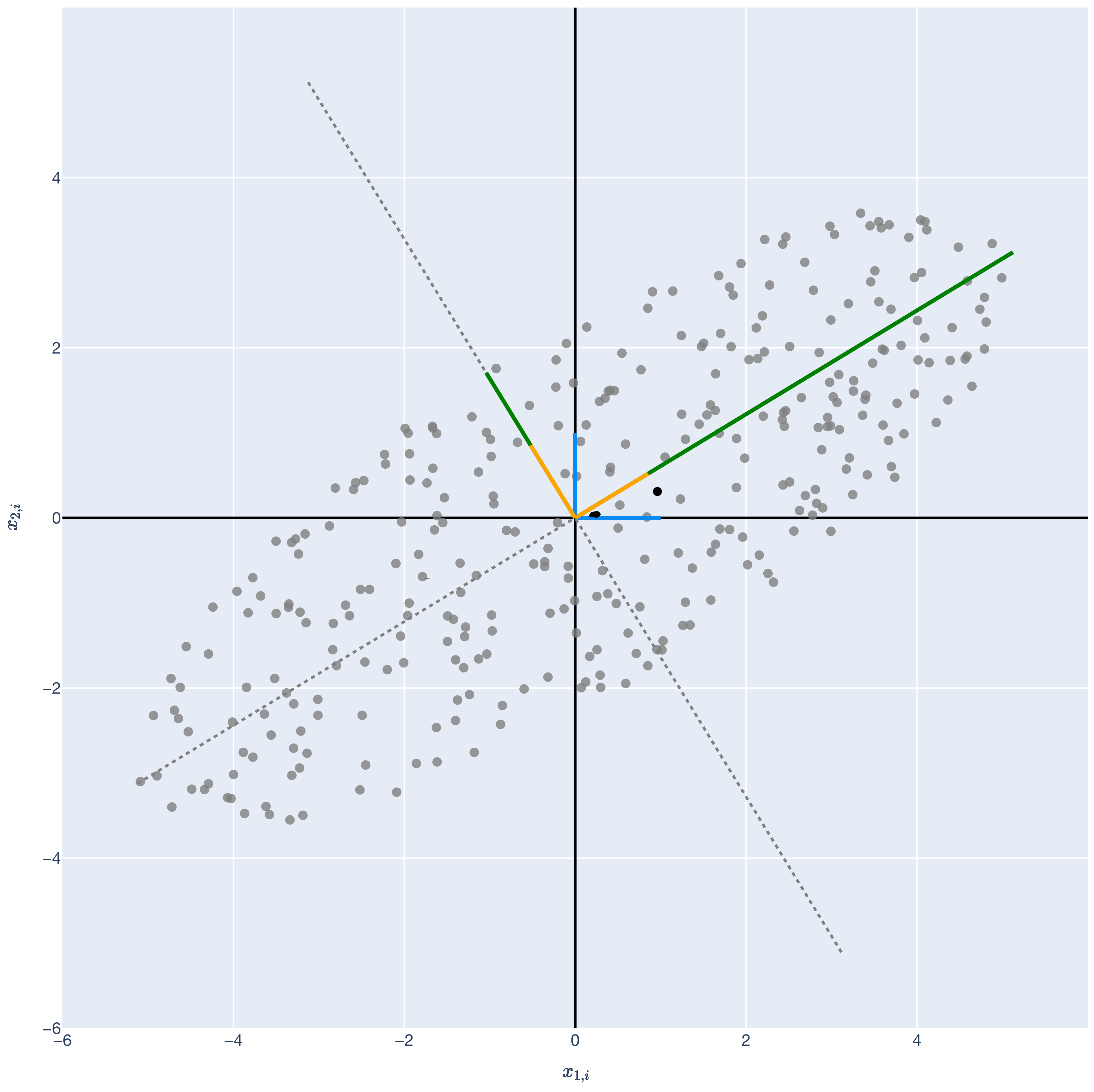


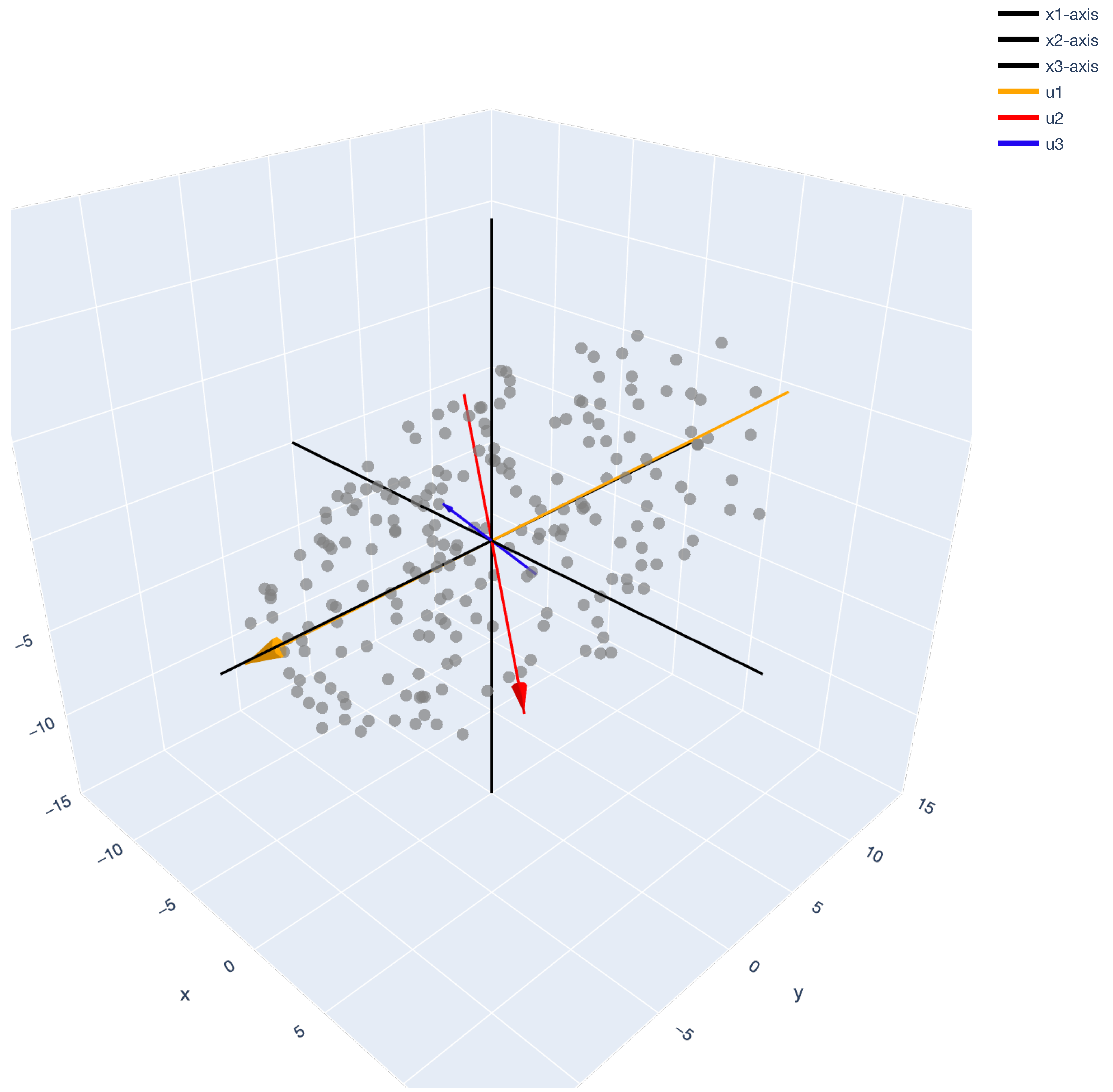


Click to interact

Lesson Overview

Big Picture: Singular Value Decomposition (SVD)





Least Squares
A Quick Review

Regression
Setup

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

$$\mathbf{X} = \begin{bmatrix} \uparrow & & \uparrow \\ \mathbf{x}_1 & \dots & \mathbf{x}_d \\ \downarrow & & \downarrow \end{bmatrix} = \begin{bmatrix} \leftarrow & \mathbf{x}_1^\top & \rightarrow \\ & \vdots & \\ \leftarrow & \mathbf{x}_n^\top & \rightarrow \end{bmatrix}.$$

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

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:

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

Regression
Setup

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": $\hat{\mathbf{w}} \in \mathbb{R}^d$ such that $y_i \approx \hat{y}_i$ for $i \in [n]$, or:

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

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

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

Regression

Setup

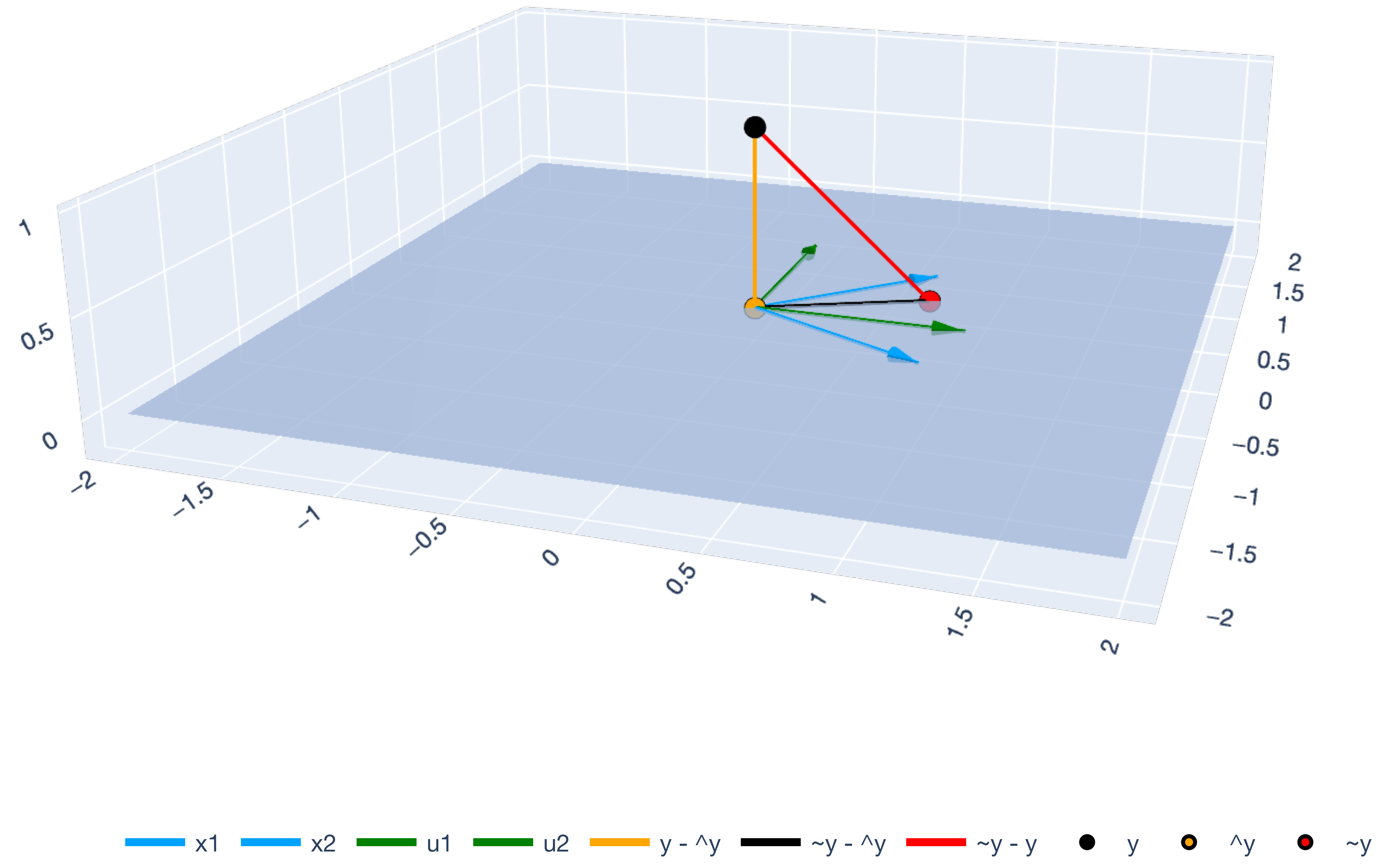
To find $\hat{\mathbf{w}}$, we follow the *principle of least squares.*

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

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:

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

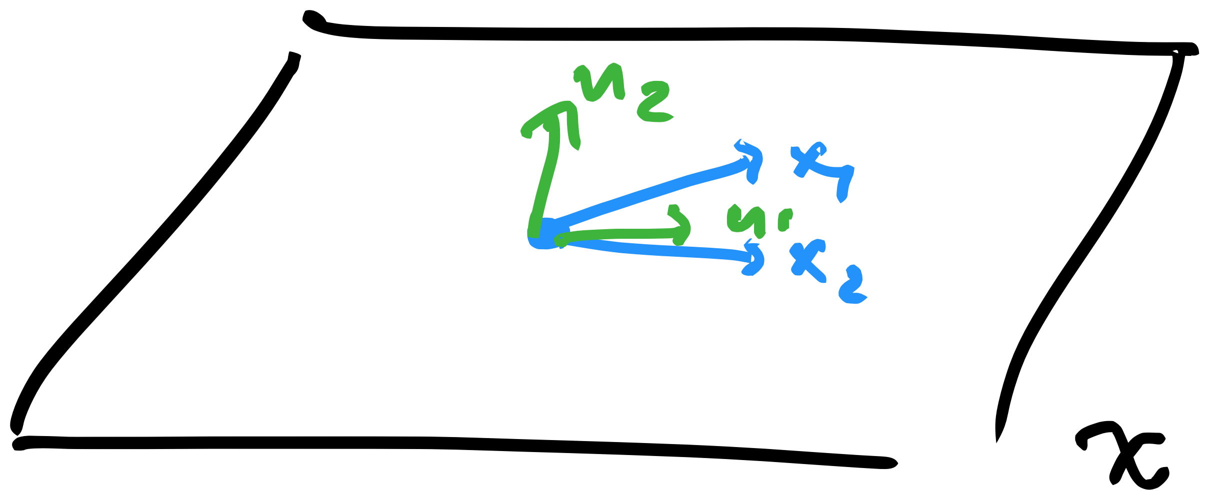
If $\underline{n \geq d}$ and $\underline{\mathrm{rank}(\mathbf{X}) = d}$, then:

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

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

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

Least Squares
OLS with Orthogonal Basis



$\mathcal{X} = \text{span}(\text{col}(X))$

<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:

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

which is solved by:

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

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

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

Least Squares
OLS with Orthogonal Basis

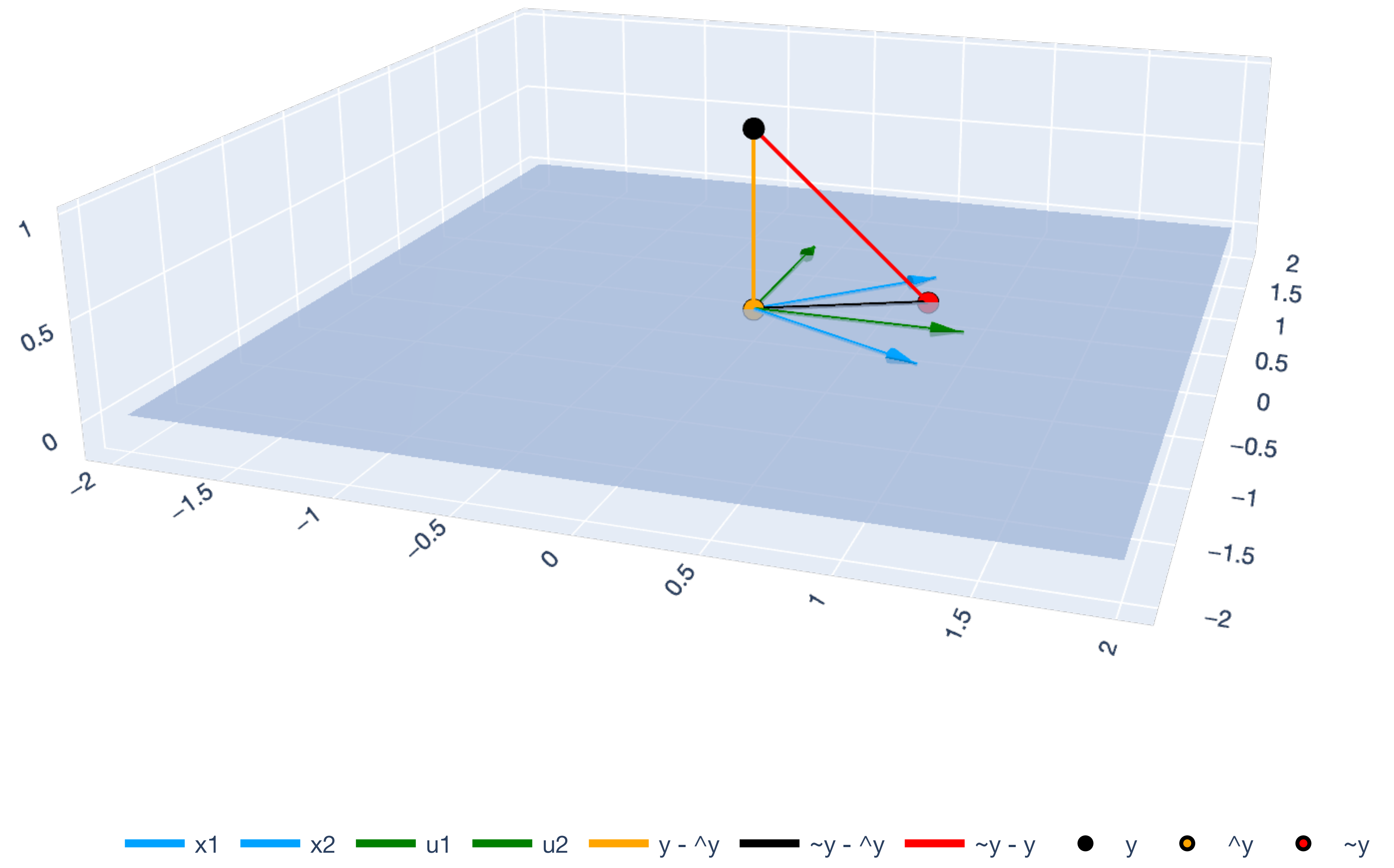
if we did have
$u_1, \ldots, u_d$

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

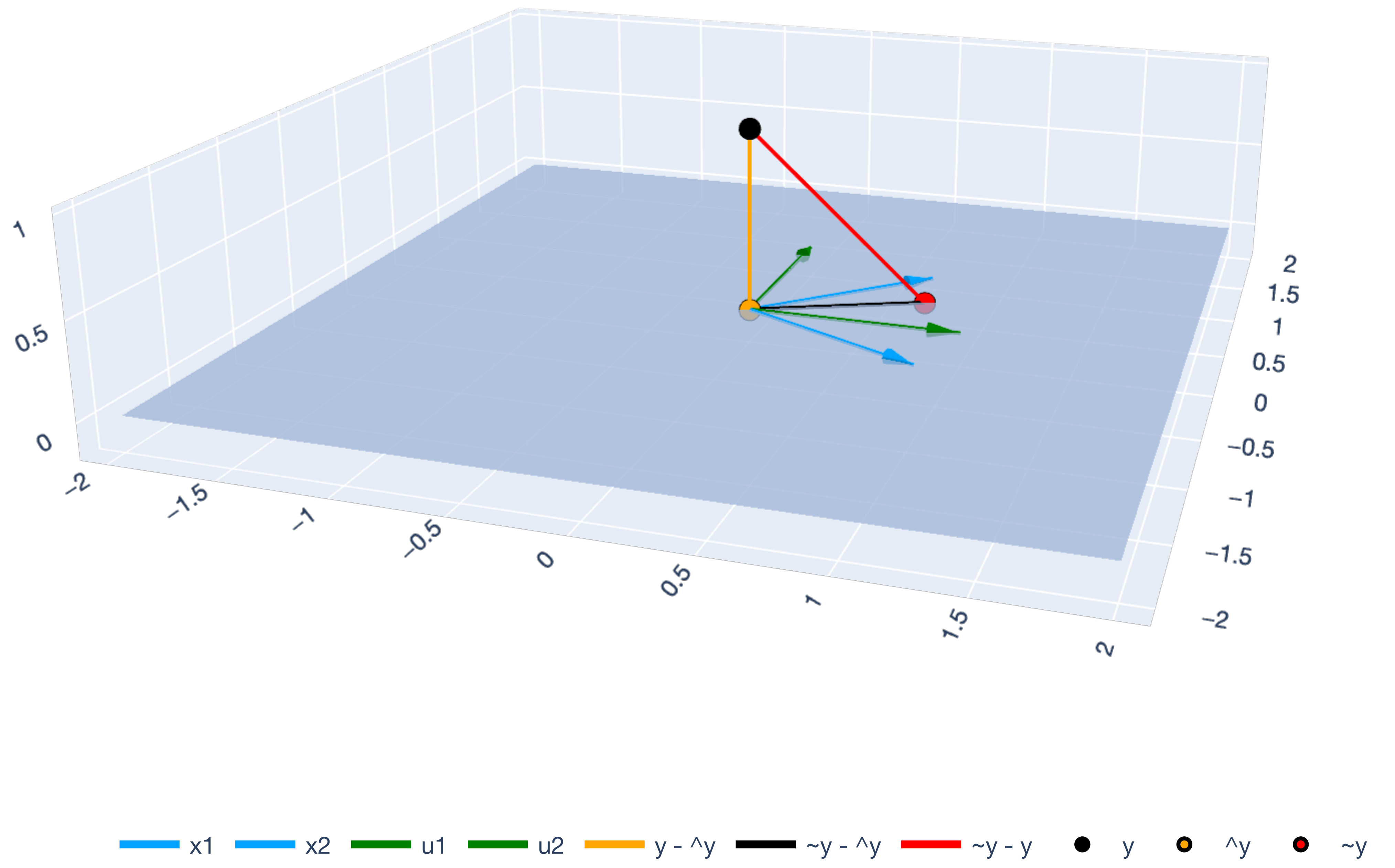
$$\hat{\mathbf{y}} = \Pi_{\mathscr{X}}(\mathbf{y}) = \mathbf{X}(\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{X}^\top\mathbf{y}$$

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

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



How to find a good orthogonal basis?

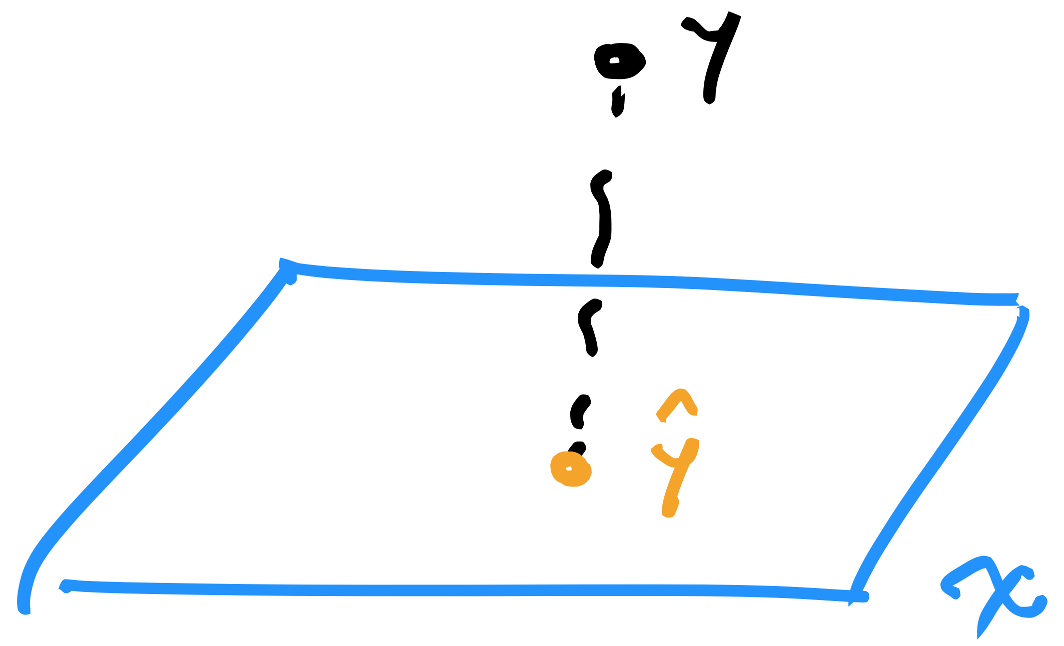


Properties of Projections

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 the set $\mathscr{X}$ is the closest vector $\hat{\mathbf{y}}$ in $\mathscr{X}$ to $\mathbf{y}$, in a Euclidean distance sense:

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

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

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

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…

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

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

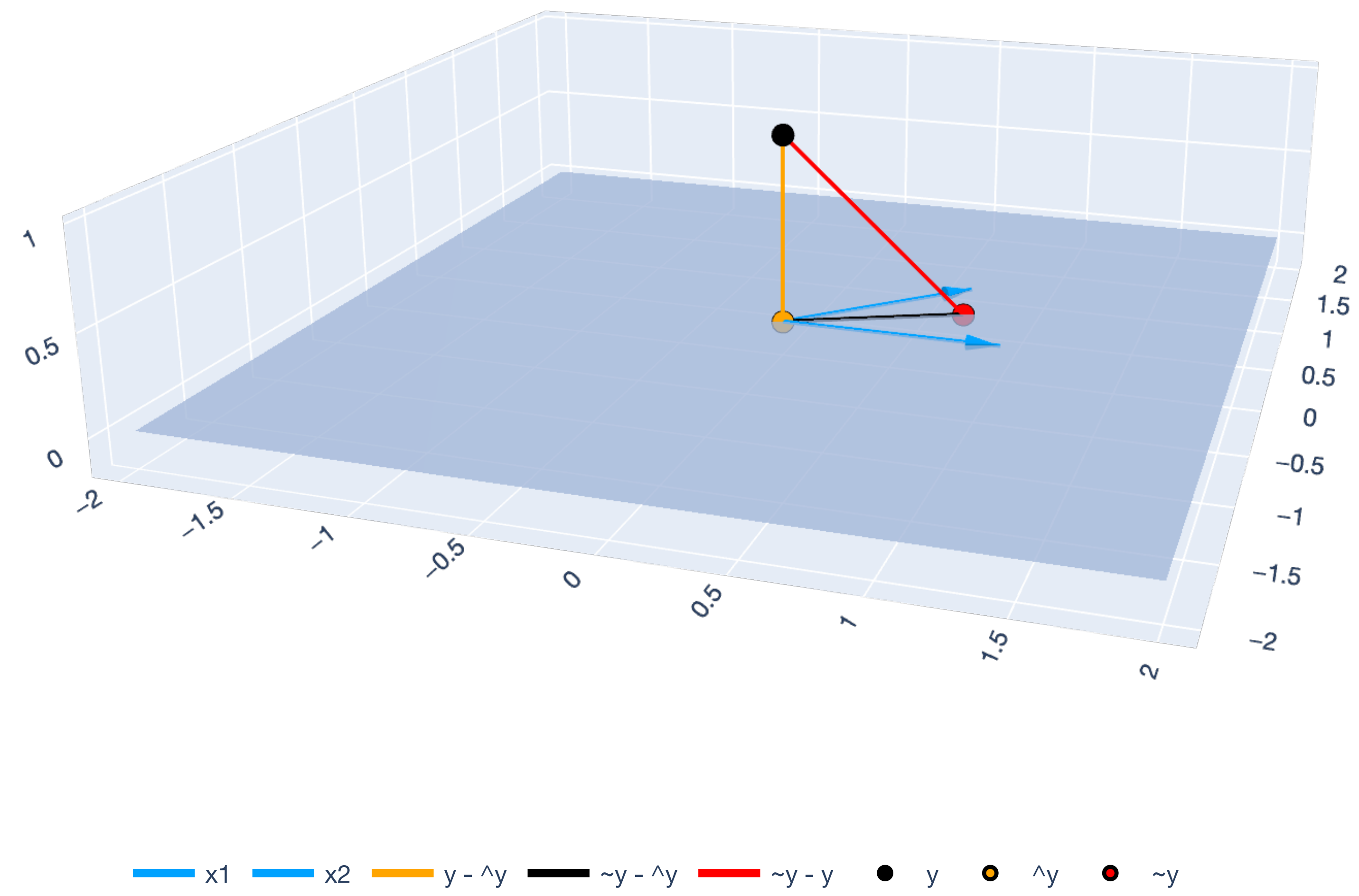
$$\boxed{\hat{\mathbf{y}} = \underbrace{\mathbf{X}}_{n\times d}(\underbrace{\mathbf{X}^\top}_{d\times n}\underbrace{\mathbf{X}}_{n\times d})^{-1}\underbrace{\mathbf{X}^\top}_{d\times n}\mathbf{y}}$$

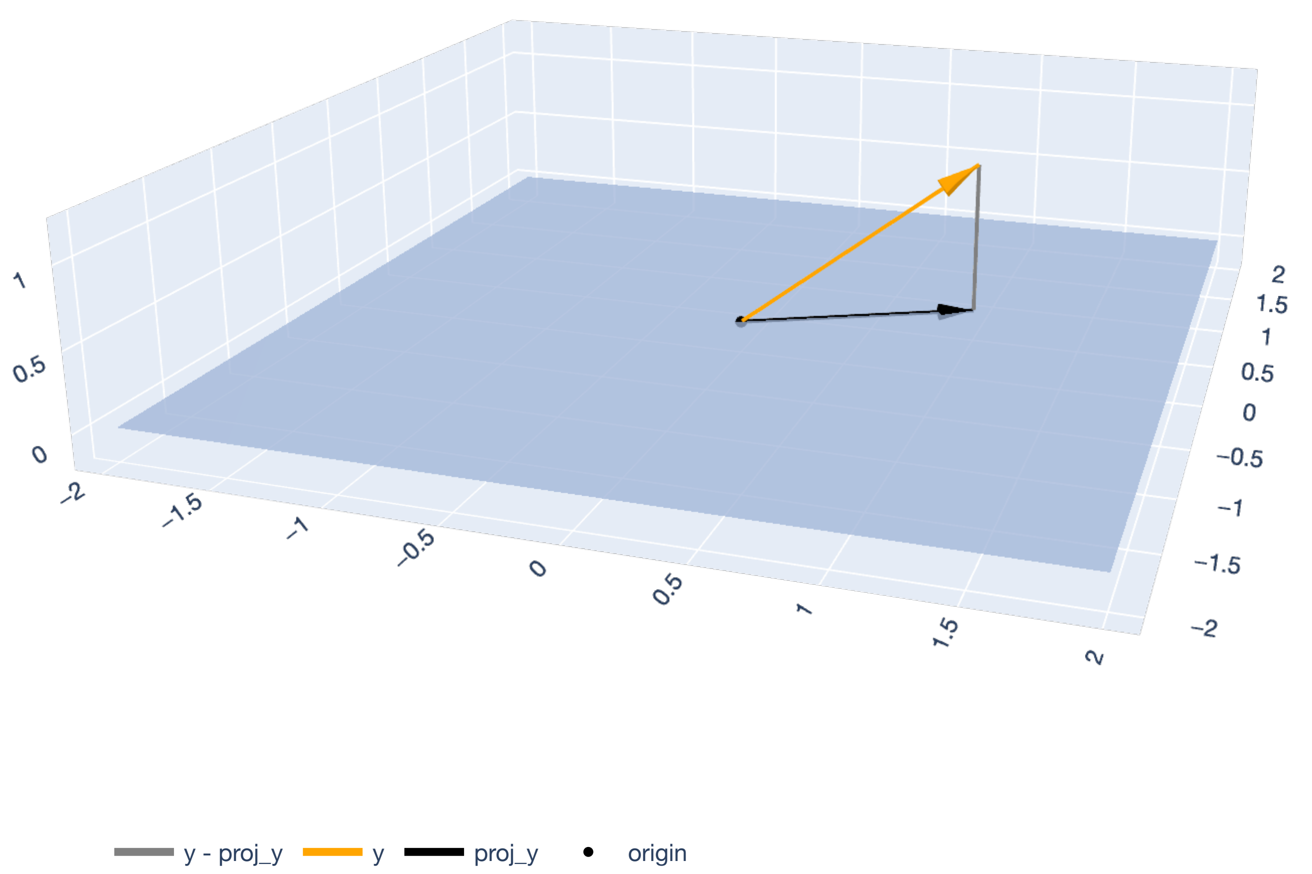
The ***projection matrix*** is:

$\mathcal{X}$

$$P_{\mathcal{X}} = \underbrace{\mathbf{X}(\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{X}^\top}_{n\times n} \in \mathbb{R}^{n\times n}.$$

$$\boxed{P_{\mathcal{X}}y = \Pi_{\mathcal{X}}(y)}$$





Least Squares as Projection

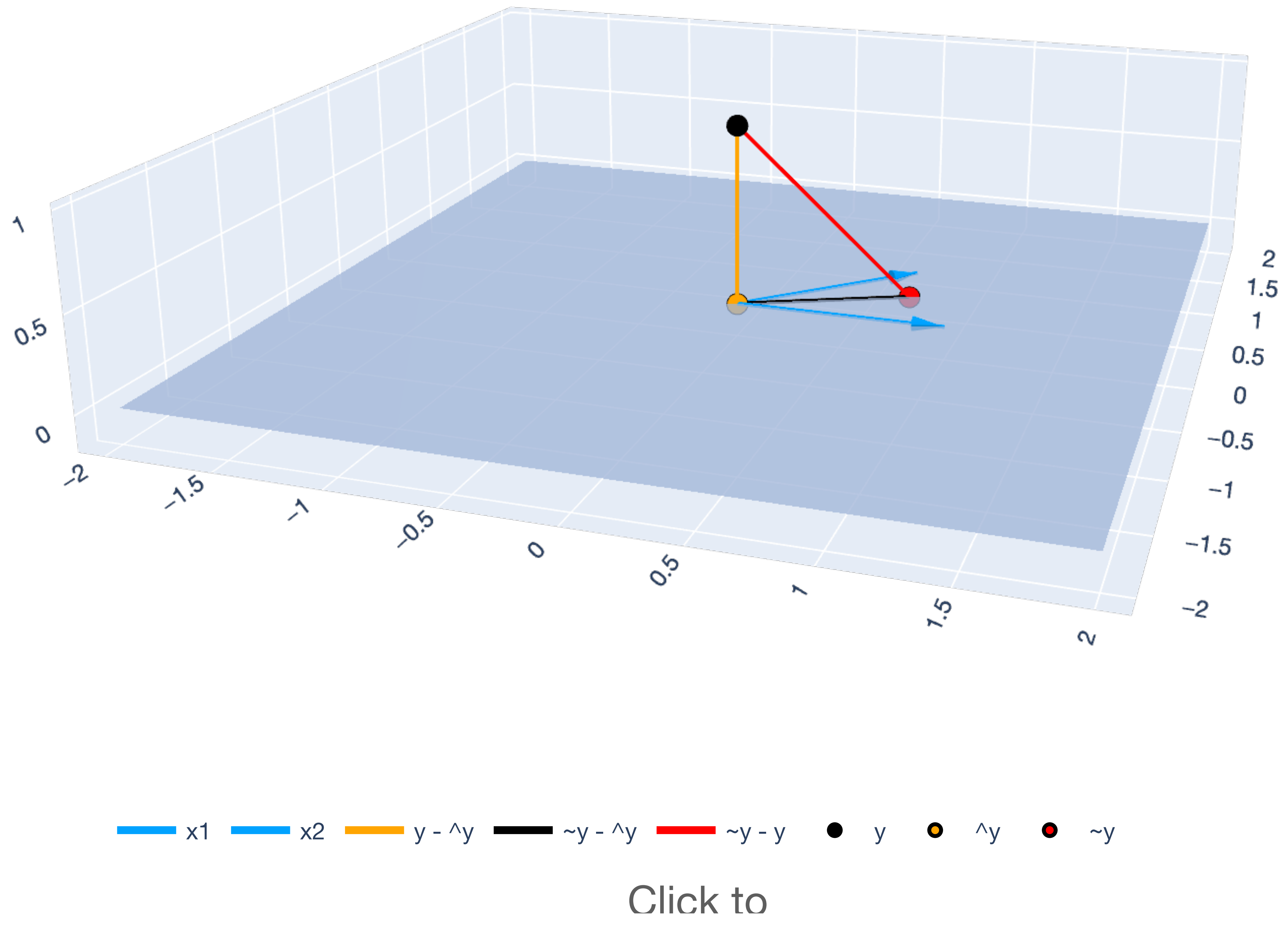
Projection Matrix

Any matrix $\mathbf{X} \in \mathbb{R}^{n \times d}$ has a subspace $\mathscr{X} = \text{span}(\text{col}(\mathbf{X}))$.

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

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

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



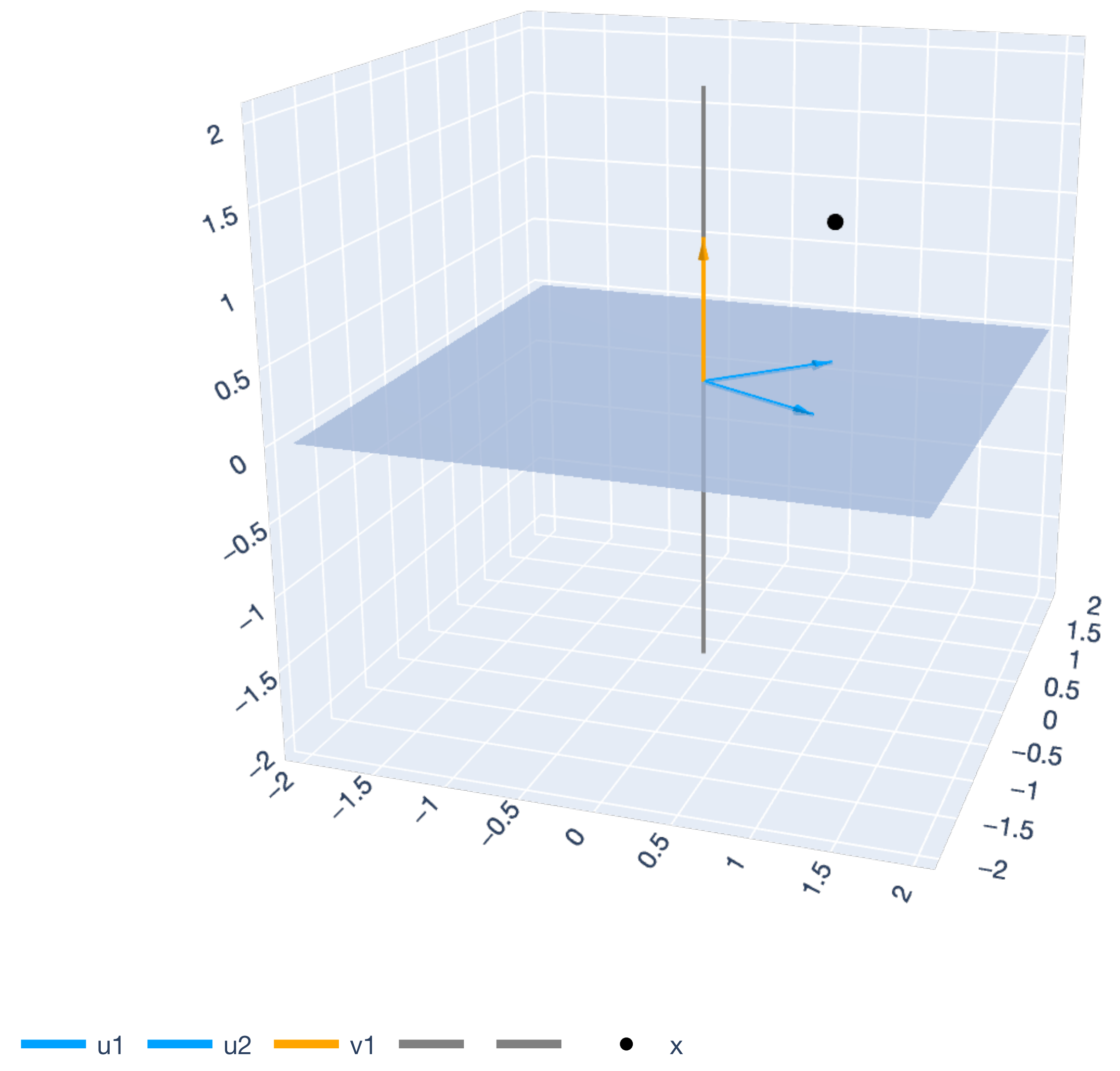
Orthogonal Complement

Intuition

"A- Perp"

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 $A^\perp$.

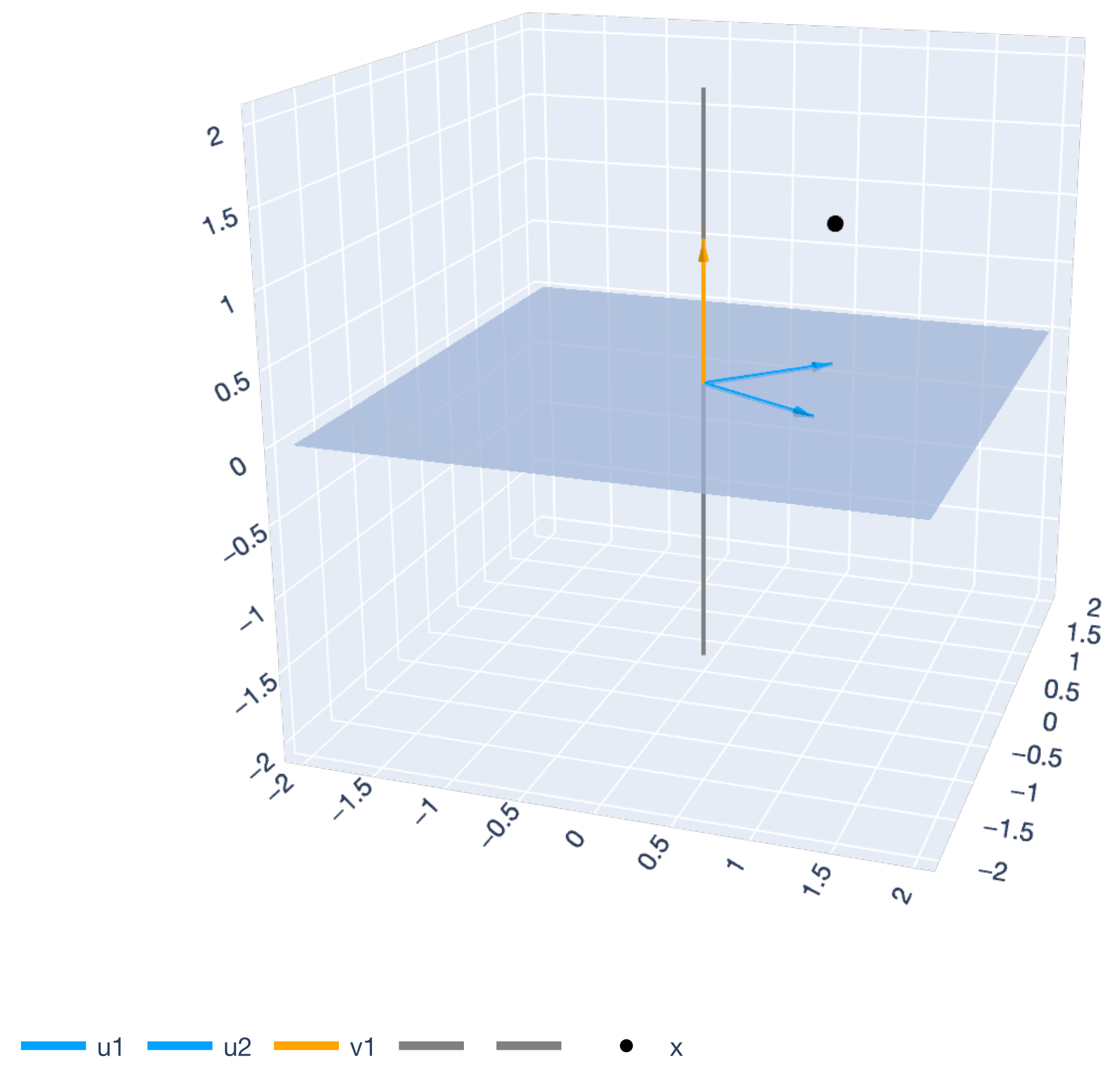


Orthogonal Complement

Definition

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

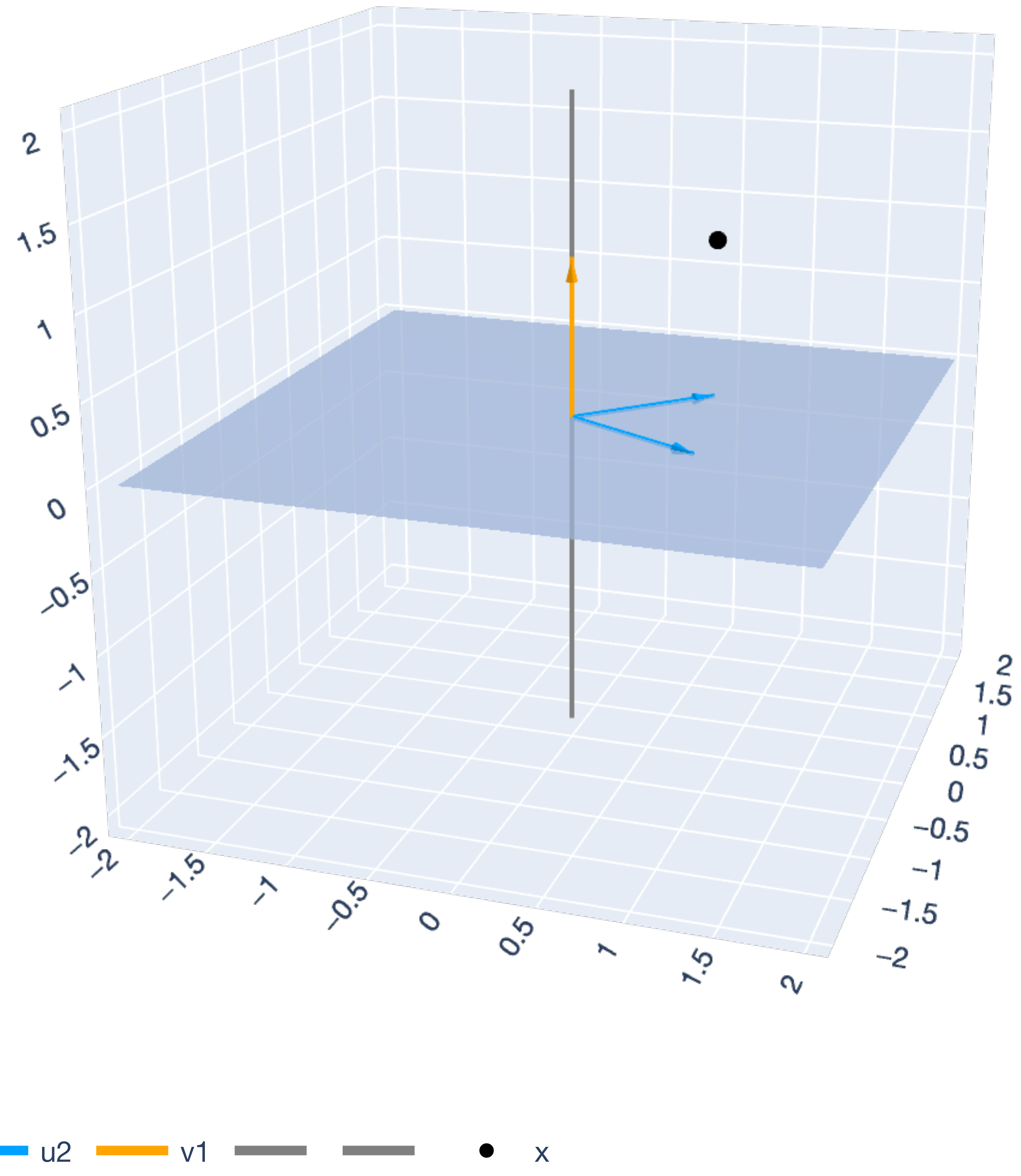
$$A^\perp := \{\mathbf{v} \in \mathbb{R}^n : \langle \mathbf{v}, \mathbf{u} \rangle = 0 \text{ for all } \mathbf{u} \in A\}.$$



Orthogonal Complement
Dimension

$A^{\perp} \subseteq \mathbb{R}^n$

For any subspace $A \subseteq \mathbb{R}^n$ with $\dim(A) = d$, the 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:

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

$$A = \begin{bmatrix} | & & | \\ a_1 & \cdots & a_d \\ | & & | \end{bmatrix} \Rightarrow A^T = \begin{bmatrix} - & a_1 & - \\ & \vdots & \\ - & a_d & - \end{bmatrix} \overset{B}{\begin{bmatrix} | & & | \\ b_1 & \cdots & b_{n-d} \\ | & & | \end{bmatrix}} = \begin{bmatrix} 0 \end{bmatrix}$$

Orthogonal Complement

Orthogonal Complement and Projections

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 $\mathbf{b}_1, \ldots, \mathbf{b}_{n-d}$, a basis for the orthogonal complement of $\mathrm{span}(\mathrm{col}(\mathbf{A}))$. 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:

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

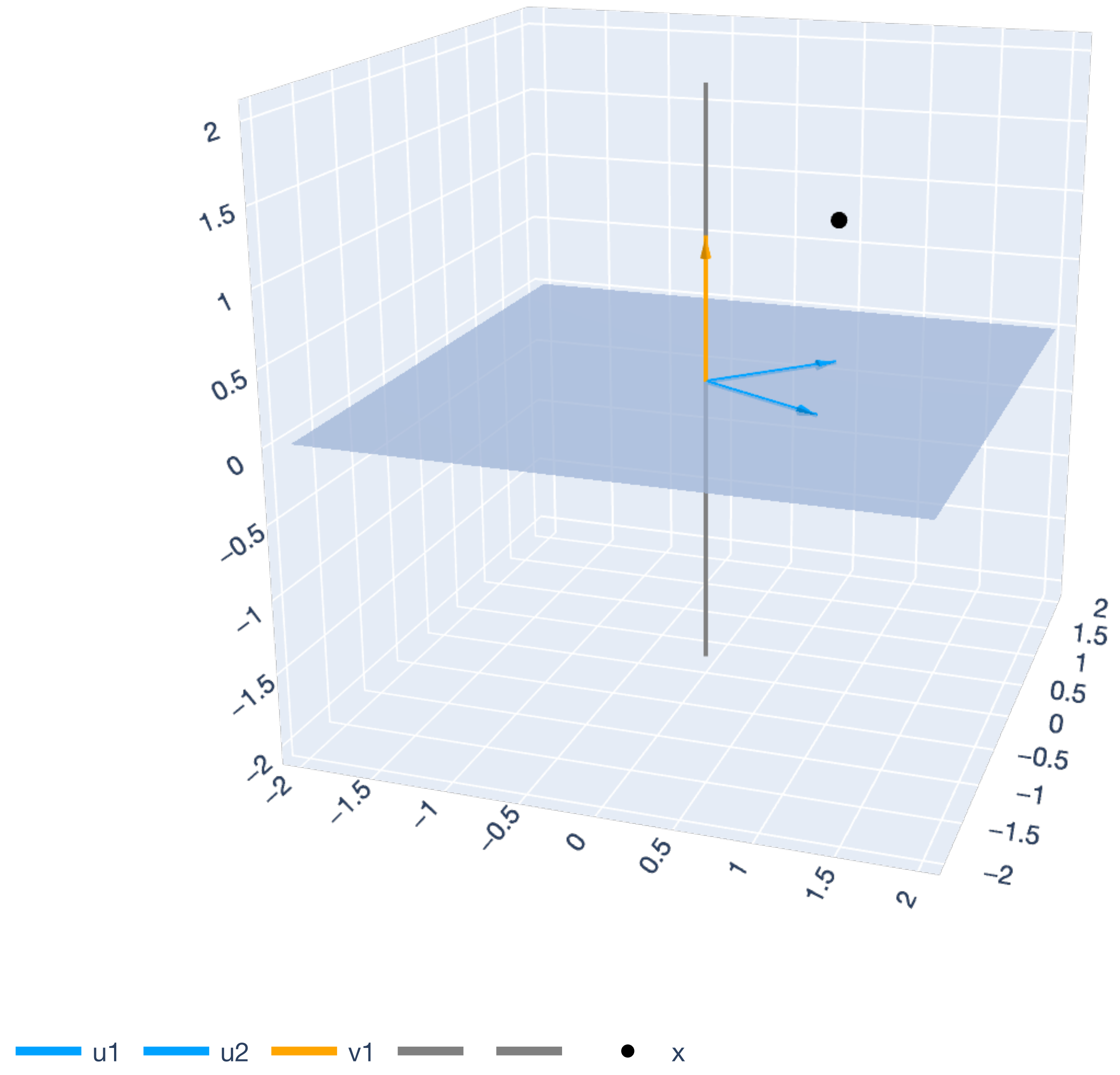
Orthogonal Complement

Orthogonal Complement and Projections

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

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

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



Projection Matrices

Properties

Let $\mathbf{A} \in \mathbb{R}^{n \times d}$ be a matrix and let $\mathbf{B} \in \mathbb{R}^{n \times (n-d)}$ have columns $\mathbf{b}_1, \ldots, \mathbf{b}_{n-d}$, a basis for the orthogonal complement of $\mathrm{span}(\mathrm{col}(\mathbf{A}))$.

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

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

Projection Matrices

Properties

Let $\mathbf{A} \in \mathbb{R}^{n \times d}$ be a matrix and let $\mathbf{B} \in \mathbb{R}^{n \times (n-d)}$ have columns $\mathbf{b}_1, \ldots, \mathbf{b}_{n-d}$, a basis for the orthogonal complement of $\mathrm{span}(\mathrm{col}(\mathbf{A}))$.

<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> $P_{\mathbf{A}} + P_{\mathbf{B}} = \mathbf{I}.$

Projection Matrices

Properties

Let $\mathbf{A} \in \mathbb{R}^{n \times d}$ be a matrix and let $\mathbf{B} \in \mathbb{R}^{n \times (n-d)}$ have columns $\mathbf{b}_1, \ldots, \mathbf{b}_{n-d}$, a basis for the orthogonal complement of $\text{span}(\text{col}(\mathbf{A}))$.

<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> $P_{\mathbf{A}} = P_{\mathbf{A}}P_{\mathbf{A}} = P_{\mathbf{A}}^2$.

Projection Matrices
Properties

Let $\mathbf{A} \in \mathbb{R}^{n \times d}$ be a matrix and let $\mathbf{B} \in \mathbb{R}^{n \times (n-d)}$ have columns $\mathbf{b}_1, \ldots, \mathbf{b}_{n-d}$, a basis for the orthogonal complement of $\text{span}(\text{col}(\mathbf{A}))$.

Prop (Orthogonal Decomposition). For any vector $\mathbf{x} \in \mathbb{R}^n$,

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

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

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

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

$P_X = X(X^T X)^{-1} X^T$

$P_X^T = X(X^T X)^{-1} X^T$

Projection Matrices

Properties

Let $\mathbf{A} \in \mathbb{R}^{n\times d}$ be a matrix and let $\mathbf{B} \in \mathbb{R}^{n\times(n-d)}$ have columns $\mathbf{b}_1, \ldots, \mathbf{b}_{n-d}$, a basis for the orthogonal complement of $\text{span}(\text{col}(\mathbf{A}))$.

Prop (Orthogonal Decomposition). For any vector $\mathbf{x} \in \mathbb{R}^n$,

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

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

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

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

$P_a \vec{x} = \dfrac{a a^T \vec{x}}{a^T a}$

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

Prop (1D projection formula). For the one-dimensional subspace associated to the vector $\mathbf{a} \in \mathbb{R}^n$, the projection matrix is:

$$\boxed{P_{\mathbf{a}} = \frac{\mathbf{a}\mathbf{a}^\top}{\mathbf{a}^\top\mathbf{a}}.}$$

$X(X^TX)^{-1}X^T$

$\Rightarrow a(a^Ta)^{-1}a^T = \dfrac{aa^T}{a^Ta}$

$$\boxed{X = U \Sigma V^T}$$

Singular Value Decomposition

1D Intuition and Derivation

Singular Value Decomposition (SVD)

1D Picture

Observe data $\mathbf{x}_1, \ldots, \mathbf{x}_d \in \mathbb{R}^n$, with matrix of *training samples* $\mathbf{X} \in \mathbb{R}^{n\times d}$:

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

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

Singular Value Decomposition (SVD)

1D Picture

u

Observe data $\mathbf{x}_1, \ldots, \mathbf{x}_d \in \mathbb{R}^n$, with matrix of *training samples* $\mathbf{X} \in \mathbb{R}^{n \times d}$:

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

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

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

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

Goal: 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

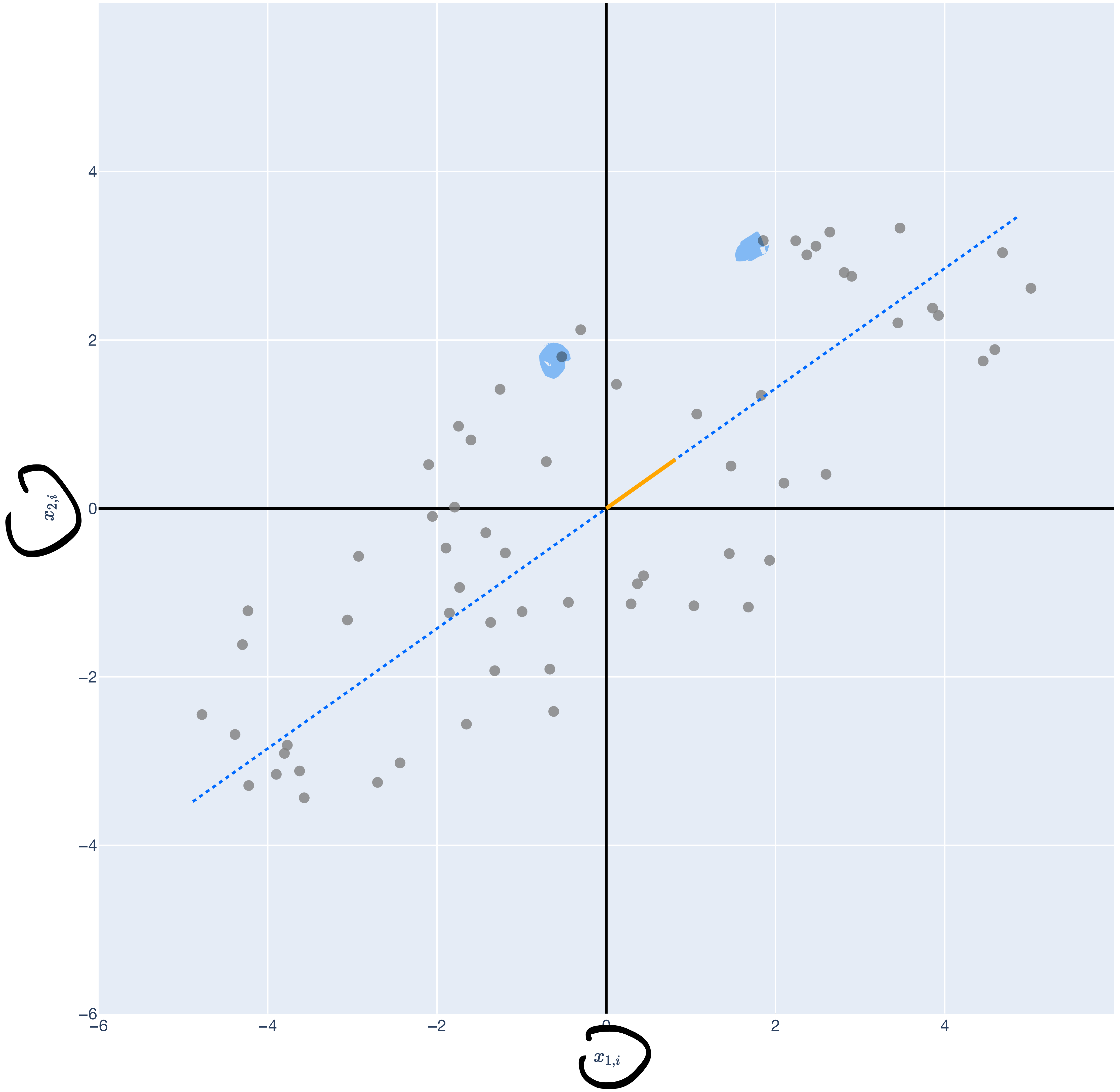
COLUMNS

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

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

$$X \in \mathbb{R}^{2 \times 50}$$

$$\begin{bmatrix} | & | & & | & | \\ x_1 & x_2 & \cdots & x_{49} & x_{50} \\ | & | & & | & | \end{bmatrix}$$



Singular Value Decomposition (SVD)

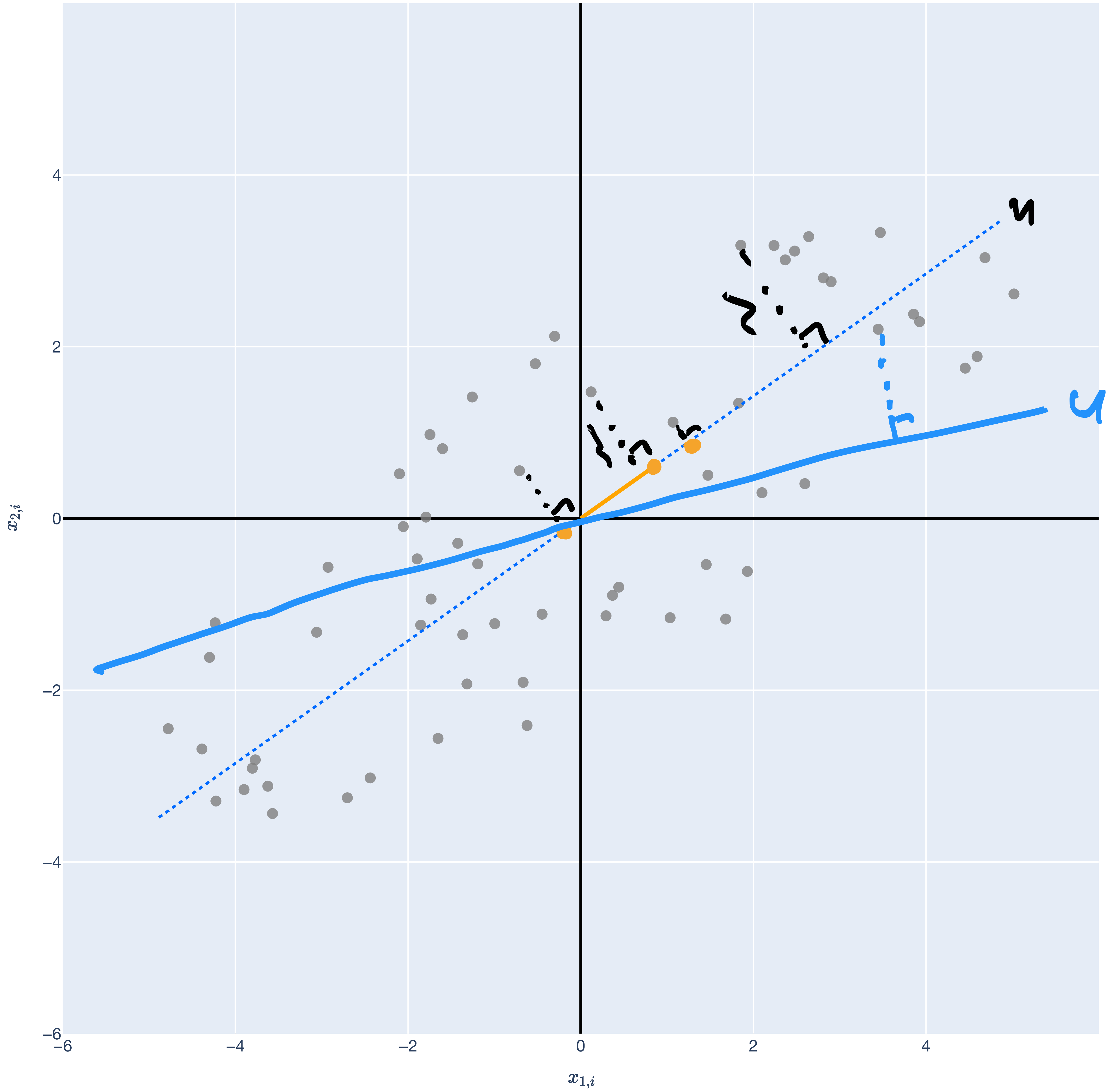
1D Picture

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

Goal: 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:

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



Comparison with OLS

1D Pictures

OLS: Observe data $\mathbf{x}_1, \ldots, \mathbf{x}_n \in \mathbb{R}^d$.

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

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

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

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

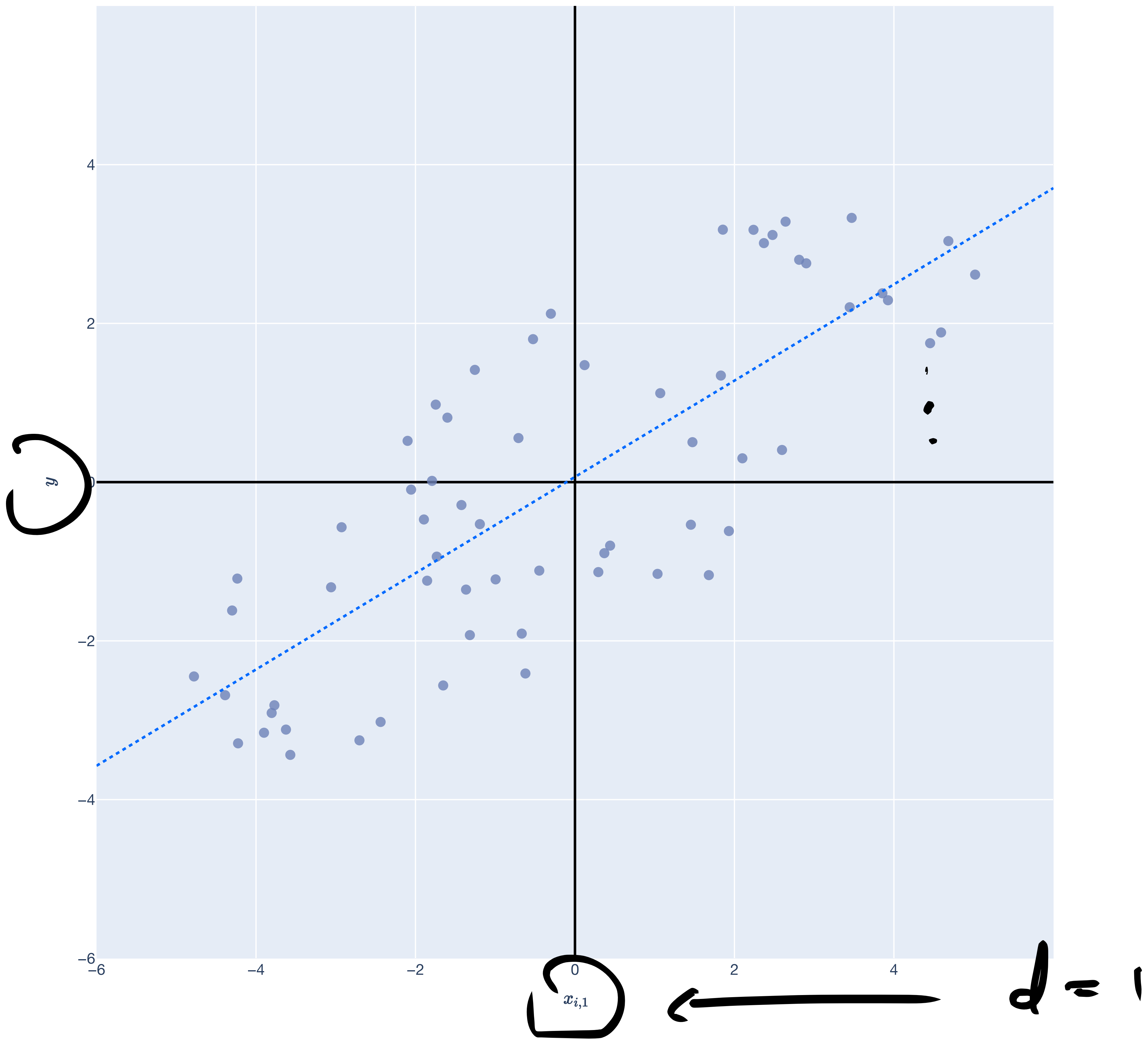
$$\underset{\mathbf{u} \in \mathbb{R}^n}{\arg\min} \ \sum_{i=1}^{d} \|\mathbf{x}_i - \Pi_{\mathbf{u}}(\mathbf{x}_i)\|^2$$

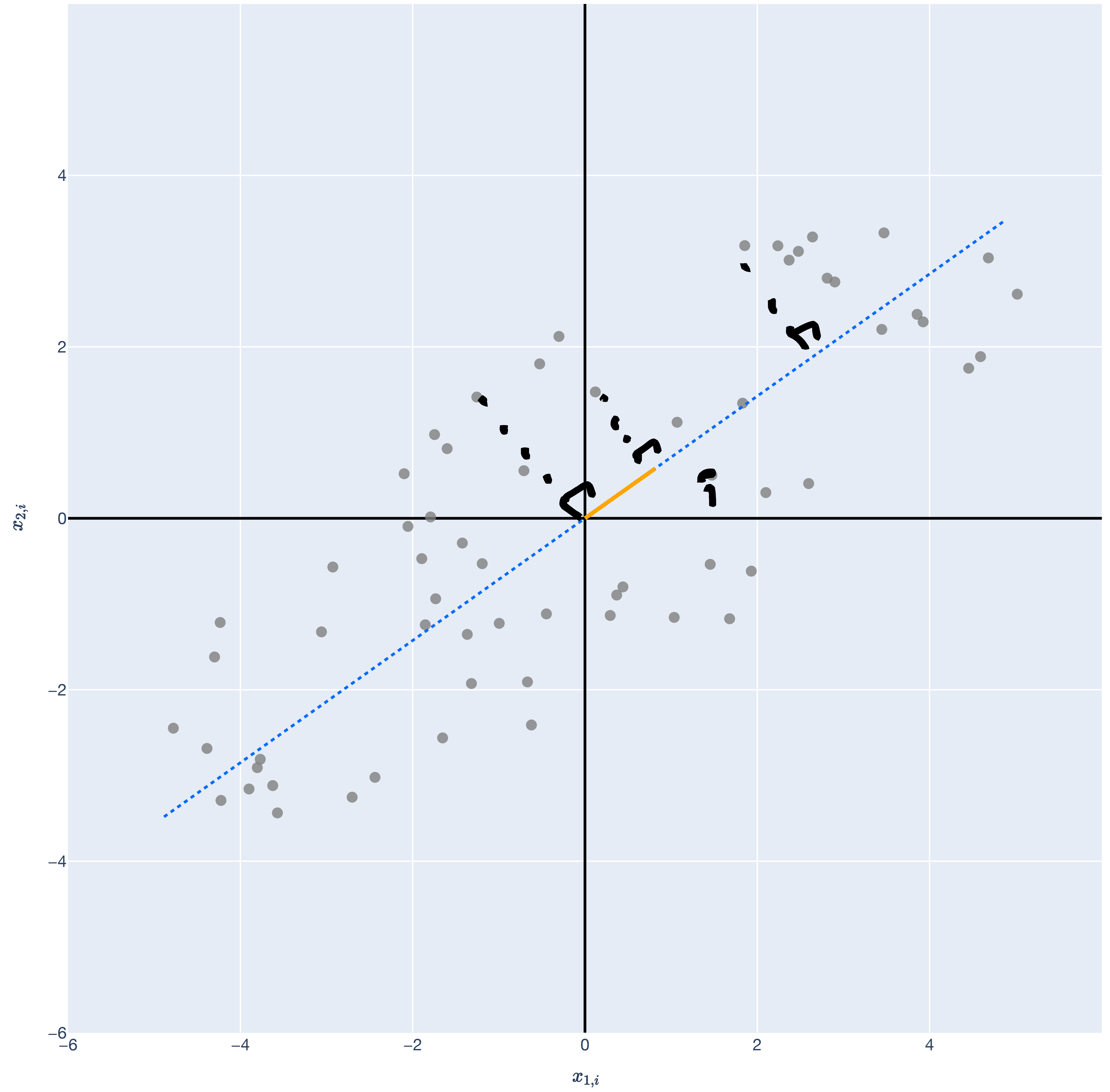
Comparison with OLS

1D Pictures

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

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





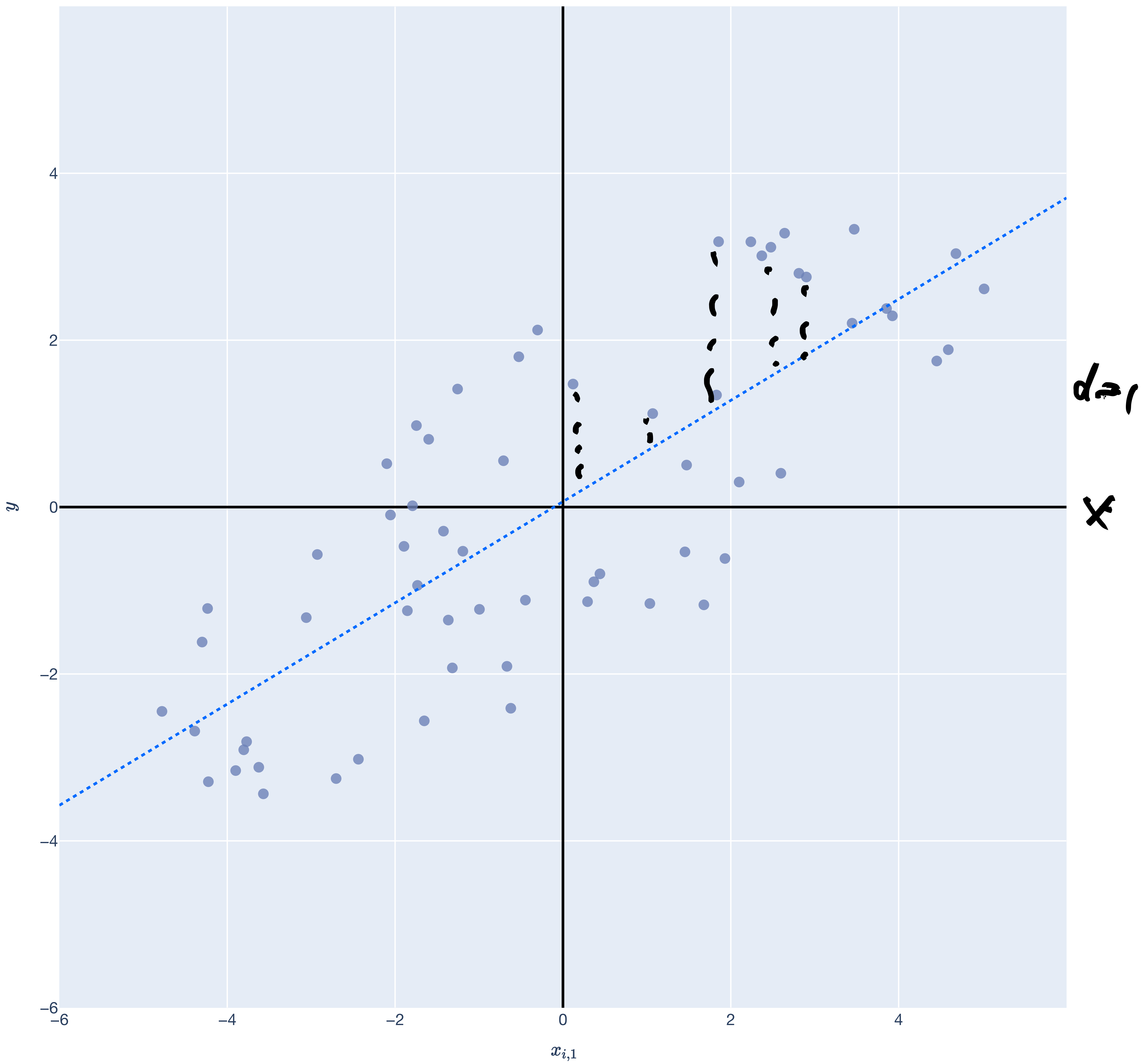
Comparison with OLS

1D Pictures

OLS: Observe data $\mathbf{x}_1, \ldots, \mathbf{x}_n \in \mathbb{R}^d$.

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

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



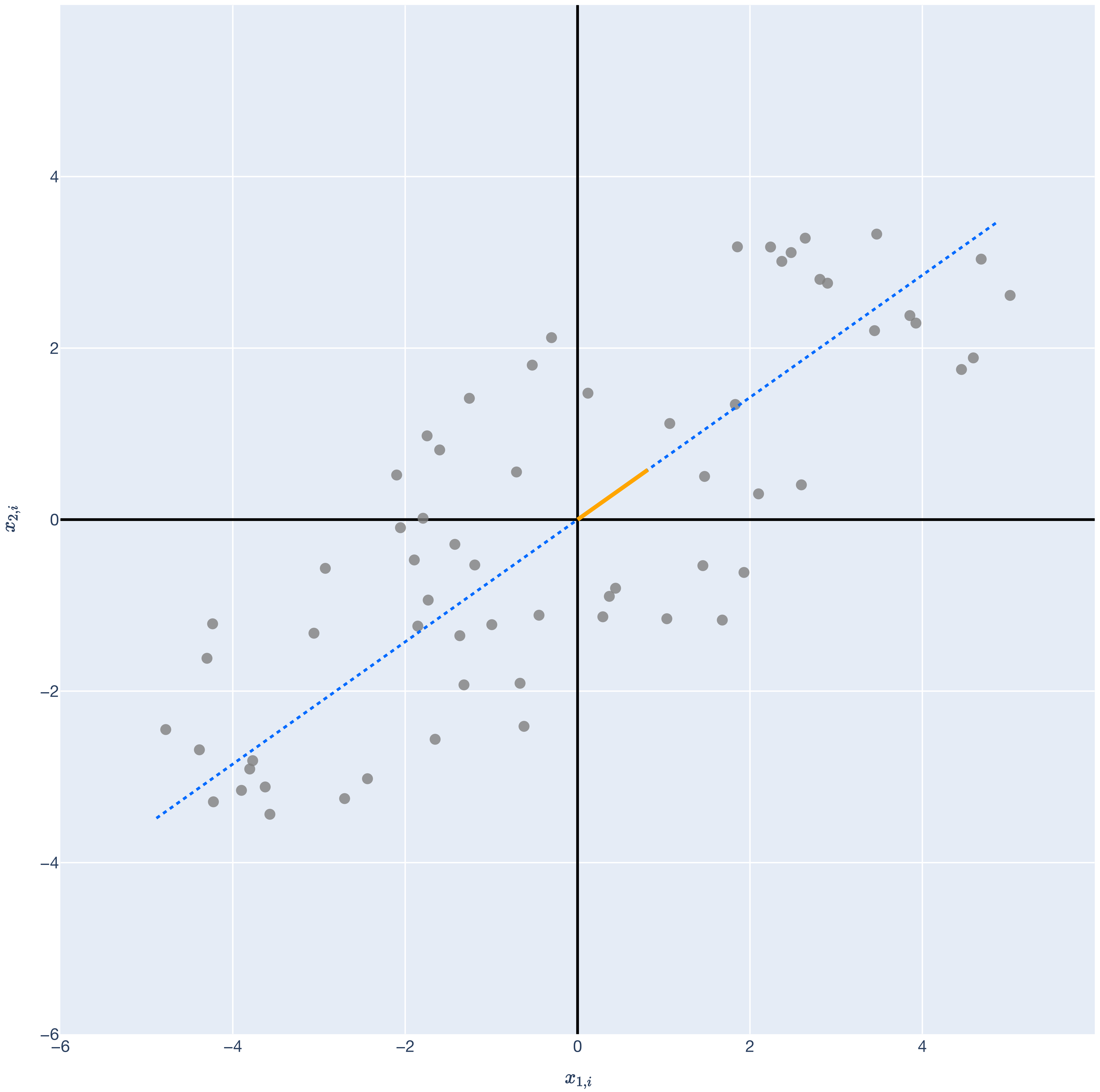
Comparison with OLS

1D Pictures

<u>BFS</u>: Observe data $\mathbf{x}_1, \ldots, \mathbf{x}_d \in \mathbb{R}^n$.

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

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



Singular Value Decomposition (SVD)
Deriving 1D SVD

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

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

Singular Value Decomposition (SVD)

Deriving 1D SVD

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

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

Singular Value Decomposition (SVD)

Deriving 1D SVD

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

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

What's $\|\mathbf{x}_i - P_{\mathbf{u}}\mathbf{x}_i\|^2$*?*

Singular Value Decomposition (SVD)

Deriving 1D SVD

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

$\frac{uu^T}{u^Tu} \in \mathbb{R}^n$

$\begin{bmatrix} | \\ u \\ | \end{bmatrix} \begin{bmatrix} - & u & - \end{bmatrix}$

$n \times 1 \quad 1 \times n \rightarrow \underline{n \times n}$

Consider any $i \in [d]$. Then,

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

(Prop: 1D projection formula)

Singular Value Decomposition (SVD)

Deriving 1D SVD

$$\boxed{P_u + P_{u^\perp} = I}$$

$$I - \frac{uu^T}{u^Tu} = \underline{\underline{I - P_u}}$$

Consider any $i \in [d]$. Then,

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

(Prop: 1D projection formula)

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

(Prop: Projection and Orthogonal Complement Matrices)

Singular Value Decomposition (SVD)

Deriving 1D SVD

Consider any $i \in [d]$. Then,

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

(Prop: 1D projection formula)

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

(Prop: Projection and Orthogonal Complement Matrices)

$\|Ax\|^2 = (Ax)^T Ax = x^T A^T A x$

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

Singular Value Decomposition (SVD)

$(AB)^T = B^T A^T$

$(A^{-1})^T = (A^T)^{-1}$

Deriving 1D SVD

Consider any $i \in [d]$. Then,

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

(Prop: 1D projection formula)

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

(Prop: Projection and Orthogonal Complement Matrices)

$$I - \frac{uu^T}{u^Tu} = \boxed{P_{u^\perp}}$$

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

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

$P = X(X^TX)^{-1}X^T$

$P^T = (\underbrace{X(X^TX)^{-1}}_{A}\underbrace{X^T}_{B})^T$

$= (X^T)^T(X(X^TX)^{-1})^T$

$= X((X^TX)^{-1})^TX^T$

$= X((X^TX)^T)^{-1}X^T = X(X^TX)^{-1}X^T$ ∎

$P^T = P$

(Prop: Projections are symmetric)

Singular Value Decomposition (SVD)

Deriving 1D SVD

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

- u spans a one dim. subspace.
- $P_{u^\perp}$ projects onto a $n-1$ dim. subspace

Consider any $i \in [d]$. Then,

$$\|\mathbf{x}_i - P_{\mathbf{u}}\mathbf{x}_i\|^2 = \left\| \mathbf{x}_i - \left(\frac{\mathbf{u}\mathbf{u}^\top}{\mathbf{u}^\top\mathbf{u}}\right)\mathbf{x}_i \right\|^2$$ (Prop: 1D projection formula)

$$= \left\| \left(\mathbf{I} - \frac{\mathbf{u}\mathbf{u}^\top}{\mathbf{u}^\top\mathbf{u}}\right)\mathbf{x}_i \right\|^2$$ (Prop: Projection and Orthogonal Complement Matrices)

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

$P^2 = P$

$$= \mathbf{x}_i^\top \left(\mathbf{I} - \frac{\mathbf{u}\mathbf{u}^\top}{\mathbf{u}^\top\mathbf{u}}\right)^2 \mathbf{x}_i$$ (Prop: Projections are symmetric)

$$= \mathbf{x}_i^\top \underbrace{\left(\mathbf{I} - \frac{\mathbf{u}\mathbf{u}^\top}{\mathbf{u}^\top\mathbf{u}}\right)}_{P_{u^\perp}} \mathbf{x}_i \leftarrow \boxed{x_i^\top P_{u^\perp} x_i}$$ (Prop: Projecting twice doesn't do anything)

Singular Value Decomposition (SVD)

Deriving 1D SVD

Therefore, for any $i \in [d]$,

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

Singular Value Decomposition (SVD)

Deriving 1D SVD

Therefore, for any $i \in [d]$,

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

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

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

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

Singular Value Decomposition (SVD)

Deriving 1D SVD

Therefore, for any $i \in [d]$,

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

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

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

Singular Value Decomposition (SVD)

Deriving 1D SVD

Therefore, for any $i \in [d]$,

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

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

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

$$\Longleftrightarrow \quad \arg\max_{\mathbf{u}\in\mathbb{R}^n} \sum_{i=1}^{d} \mathbf{x}_i^\top \underbrace{\left(\frac{\mathbf{u}\mathbf{u}^\top}{\mathbf{u}^\top\mathbf{u}} \right)}_{P_u} \mathbf{x}_i$$

$$\sum_{i=1}^{d} - x_i^T \left(\frac{uu^T}{u^Tu} \right) x_i$$

$$= - \sum_{i=1}^{d} x_i^T \left(\frac{uu^T}{u^Tu} \right) x_i$$

$$\sum_{i=1}^{d} x_i^T P_u x_i$$

Singular Value Decomposition (SVD)

Deriving 1D SVD

Therefore, for any $i \in [d]$,

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

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

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

$$\iff \arg\max_{\mathbf{u}\in\mathbb{R}^n} \sum_{i=1}^{d} \mathbf{x}_i^\top \left(\frac{\mathbf{u}\mathbf{u}^\top}{\mathbf{u}^\top\mathbf{u}} \right) \mathbf{x}_i = \arg\max_{\mathbf{u}\in\mathbb{R}^n} \frac{\boxed{\mathbf{u}^\top\mathbf{X}\mathbf{X}^\top\mathbf{u}}}{\mathbf{u}^\top\mathbf{u}}$$ (AKA: squared *operator norm* of $\mathbf{X}$, i.e. $\|\mathbf{X}\|_{op}^2$).

Singular Value Decomposition (SVD)

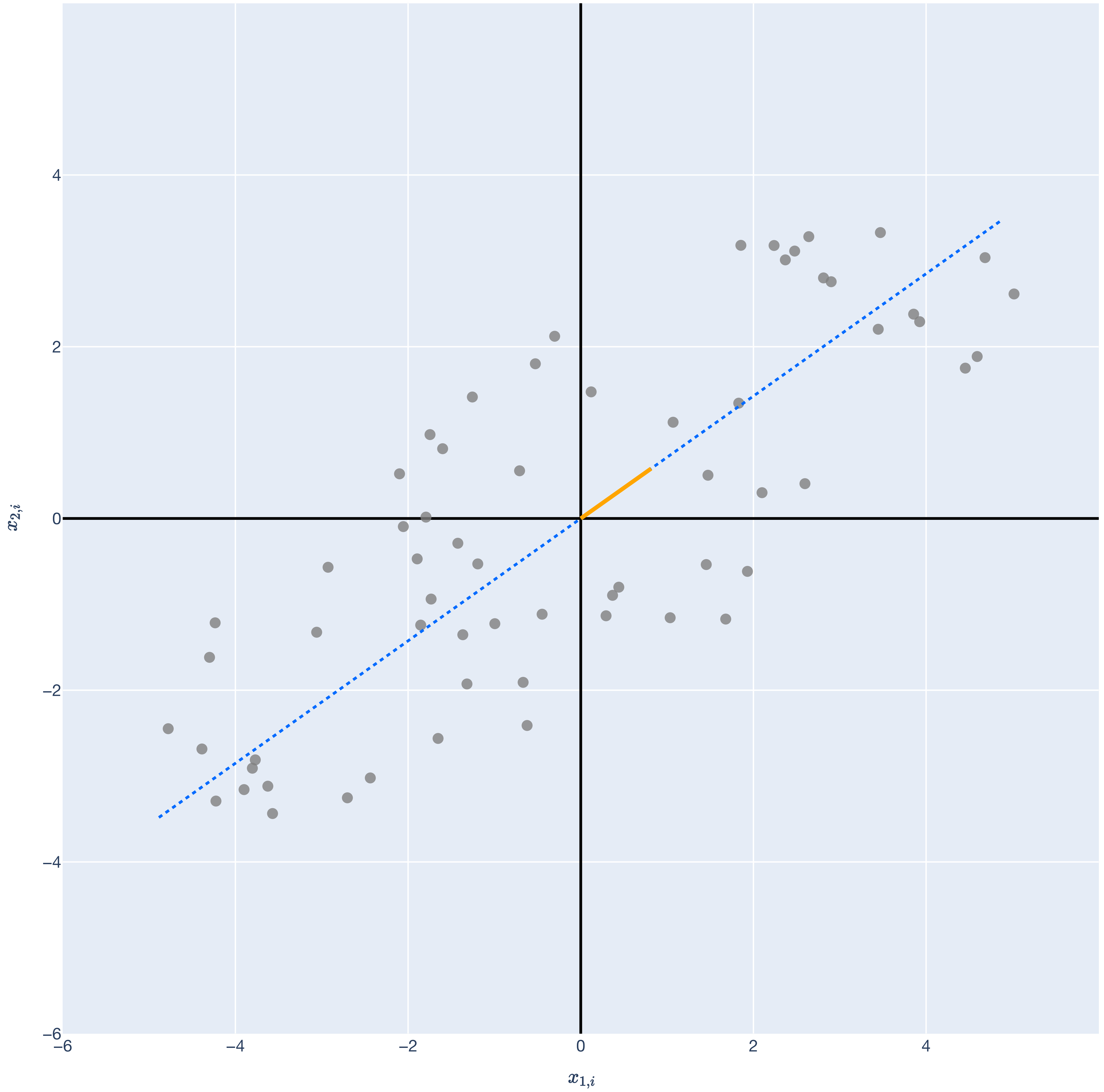
1D Picture

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:

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



Singular Value Decomposition (SVD)

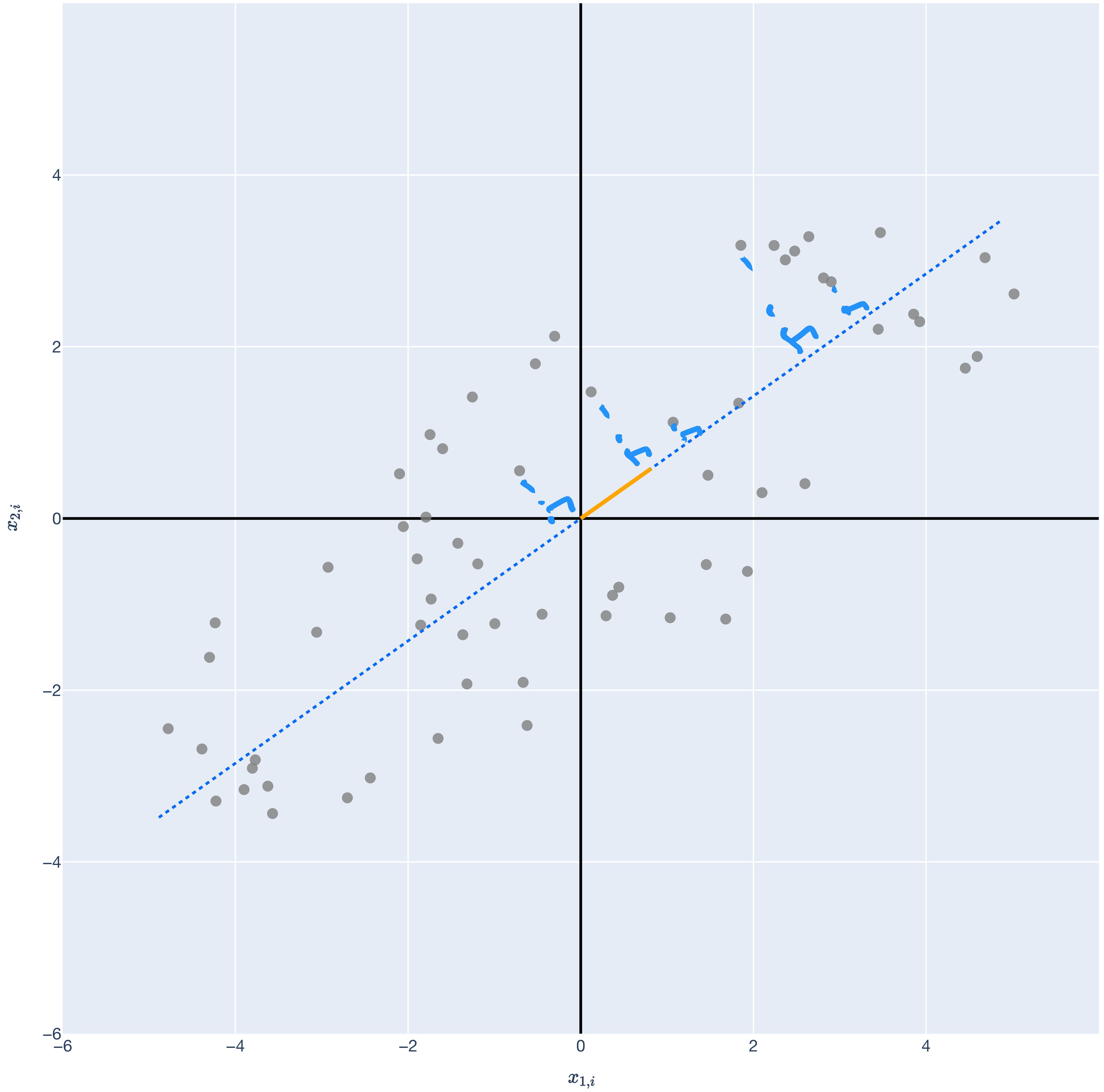
1D Picture

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:

$$\arg\max_{\mathbf{u} \in \mathbb{R}^n} \frac{\mathbf{u}^\top \mathbf{X} \mathbf{X}^\top \mathbf{u}}{\mathbf{u}^\top \mathbf{u}}.$$



$P = X(X^T X)^{-1} X^T$

Singular Value Decomposition (SVD)

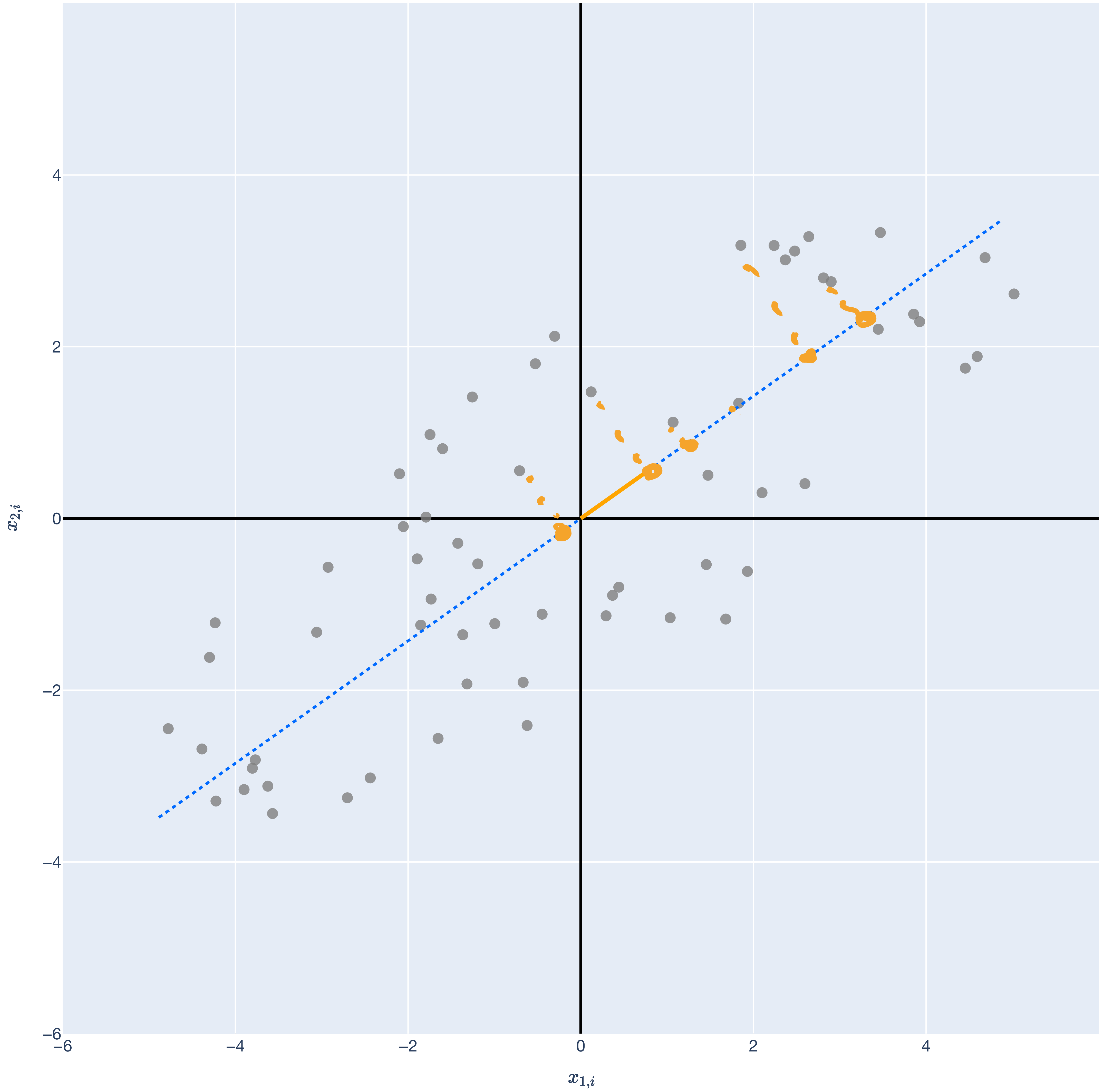
1D Picture

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

$$\boxed{\arg\max_{\mathbf{u} \in \mathbb{R}^n} \frac{\mathbf{u}^\top \mathbf{X}\mathbf{X}^\top \mathbf{u}}{\mathbf{u}^\top \mathbf{u}}.}$$

The vector $\mathbf{u} \in \mathbb{R}^n$ that achieves this maximum is the *1st* ***left singular vector***.

The value $\dfrac{\mathbf{u}^\top \mathbf{X}\mathbf{X}^\top \mathbf{u}}{\mathbf{u}^\top \mathbf{u}}$ is σ_1^2, the *squared 1st* ***singular value*** of $\mathbf{X}$.



Singular Value Decomposition

Definition of Full SVD and Compact SVD

Singular Value Decomposition (SVD)

Building up the SVD

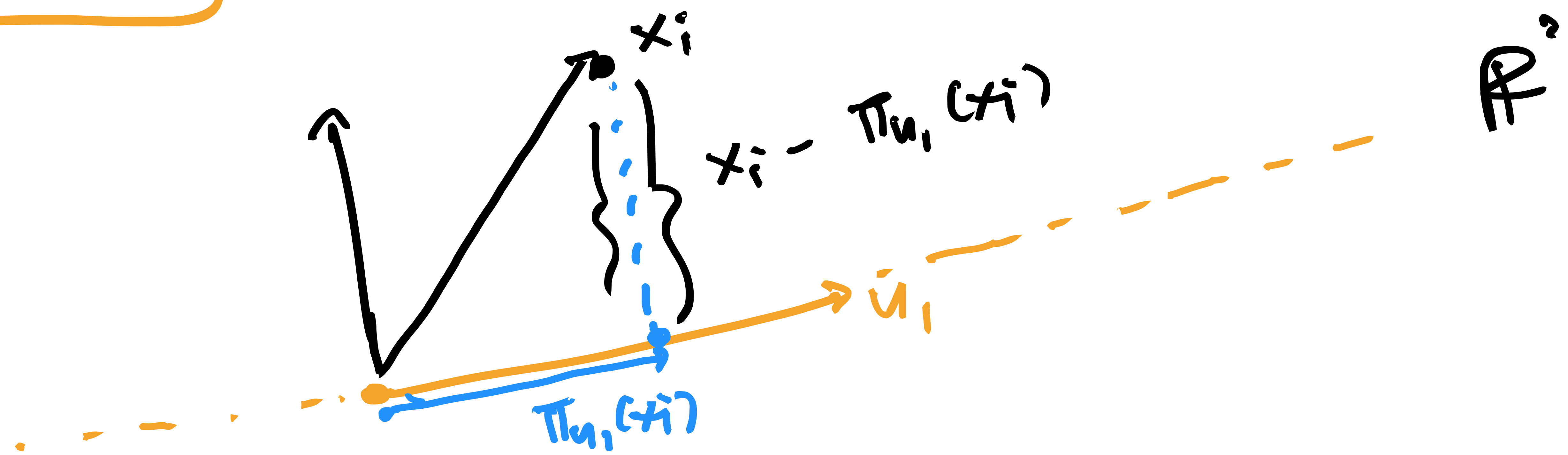
Observe data $\mathbf{x}_1, \ldots, \mathbf{x}_d \in \mathbb{R}^n$. Consider the following procedure…

For $t = 1,2,\ldots,n$…

$$\underset{u}{\operatorname{argmax}} \frac{u^T x x^T u}{u^T u}$$

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

- Let $\mathbf{x}_i^{(1)} = \mathbf{x}_i - \Pi_{\mathbf{u}_1}(\mathbf{x}_i)$. ← Part of each x_i "unexplained" by $\vec{u}$.



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 $\mathbf{x}_1, \ldots, \mathbf{x}_d$.

   Apply to modified $x_1^{(1)}, \ldots, x_d^{(1)}$

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

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

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)

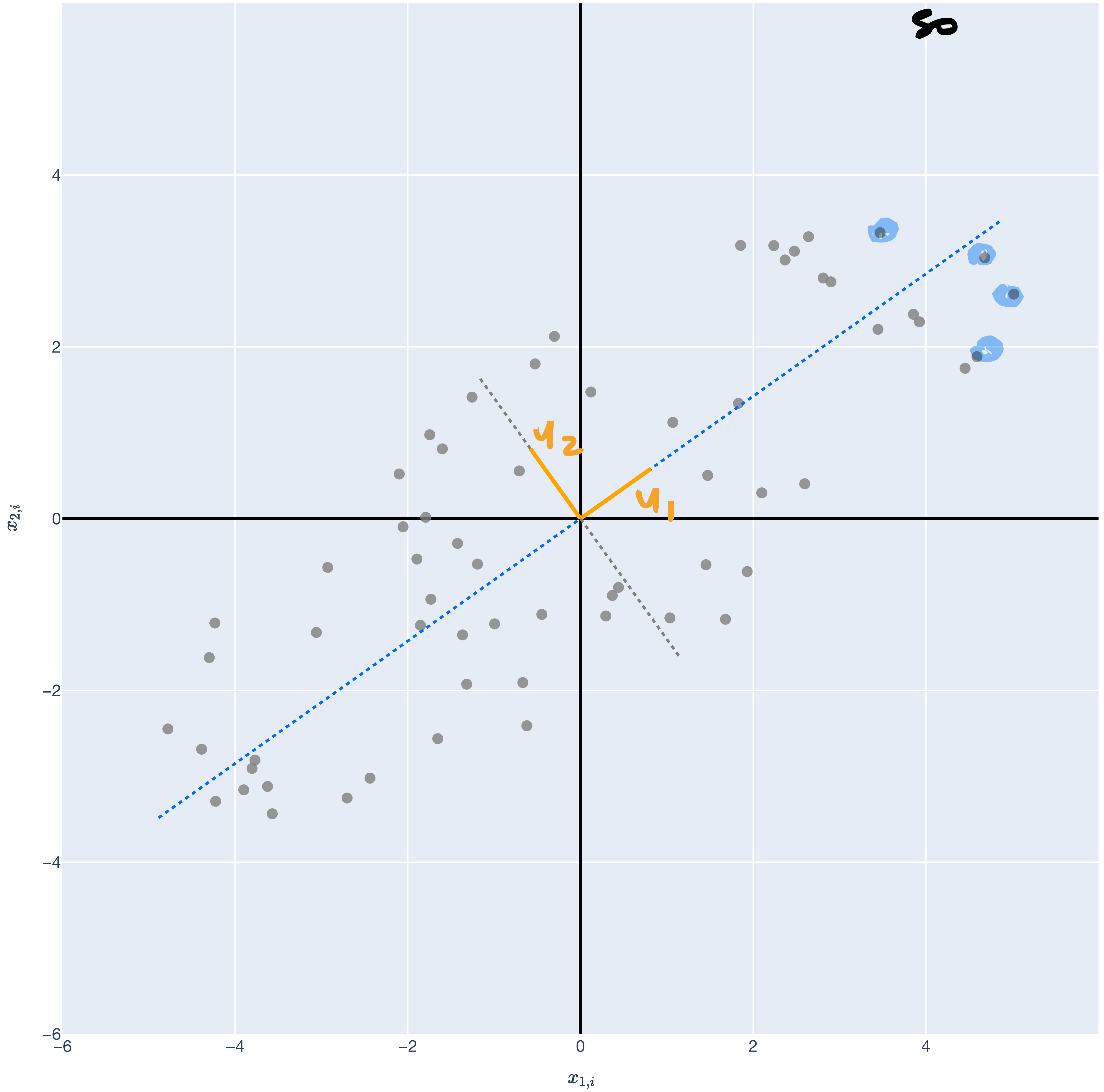
Building up the SVD

$n = 2$
$d = 50$

$X =_2 \{ [\ \underbrace{\quad\quad\quad}_{50}\ ]$

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

Then, $\mathbf{u}_1$ and $\mathbf{u}_2$ look like…



Singular Value Decomposition (SVD)

Building up the SVD

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

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

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

u_2

$u_1, \dots, u_n$

$= u_n$

The n left singular vectors are 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

$\sigma_1, \dots, \sigma_d \rightarrow$ values $\dfrac{u^T X X^T u}{u^T u}$

Consider any matrix $\mathbf{X} \in \mathbb{R}^{n \times d}$. By the ***full singular value decomposition (SVD)***, there exist matrices $\mathbf{U}, \Sigma, \mathbf{V}$ such that

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

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

Because $u_1, \dots, u_n$ are orthonormal

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

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

$$V = \begin{bmatrix} | & & | \\ v_1 & \cdots & v_d \\ | & & | \end{bmatrix}$$

$\mathbf{\Sigma} \in \mathbb{R}^{n \times d}$ is a diagonal matrix with ***singular values*** $\sigma_1 \geq \sigma_2 \geq \ldots \geq \sigma_d \geq 0$ on the diagonal. The rank is equal to the number of $\sigma_i > 0$.

$$\begin{bmatrix} \sigma_1 \cdots \sigma_d & 0 \\ 0 & \end{bmatrix}$$

$$\boxed{\text{rank}(X) = \#\text{ singular values } \sigma_i > 0}$$

Singular Value Decomposition (SVD)

Shape of the Σ Matrix

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

$d \times d$

$n \times n$

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

Interpreting the SVD

Example in $\mathbb{R}^2$

$$X = \begin{bmatrix} | & | & & | \\ x_1 & x_2 & \cdots & x_{212} \\ | & | & & | \end{bmatrix} \Big\} \begin{array}{l} n = 2 \\ d = 212. \end{array}$$

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

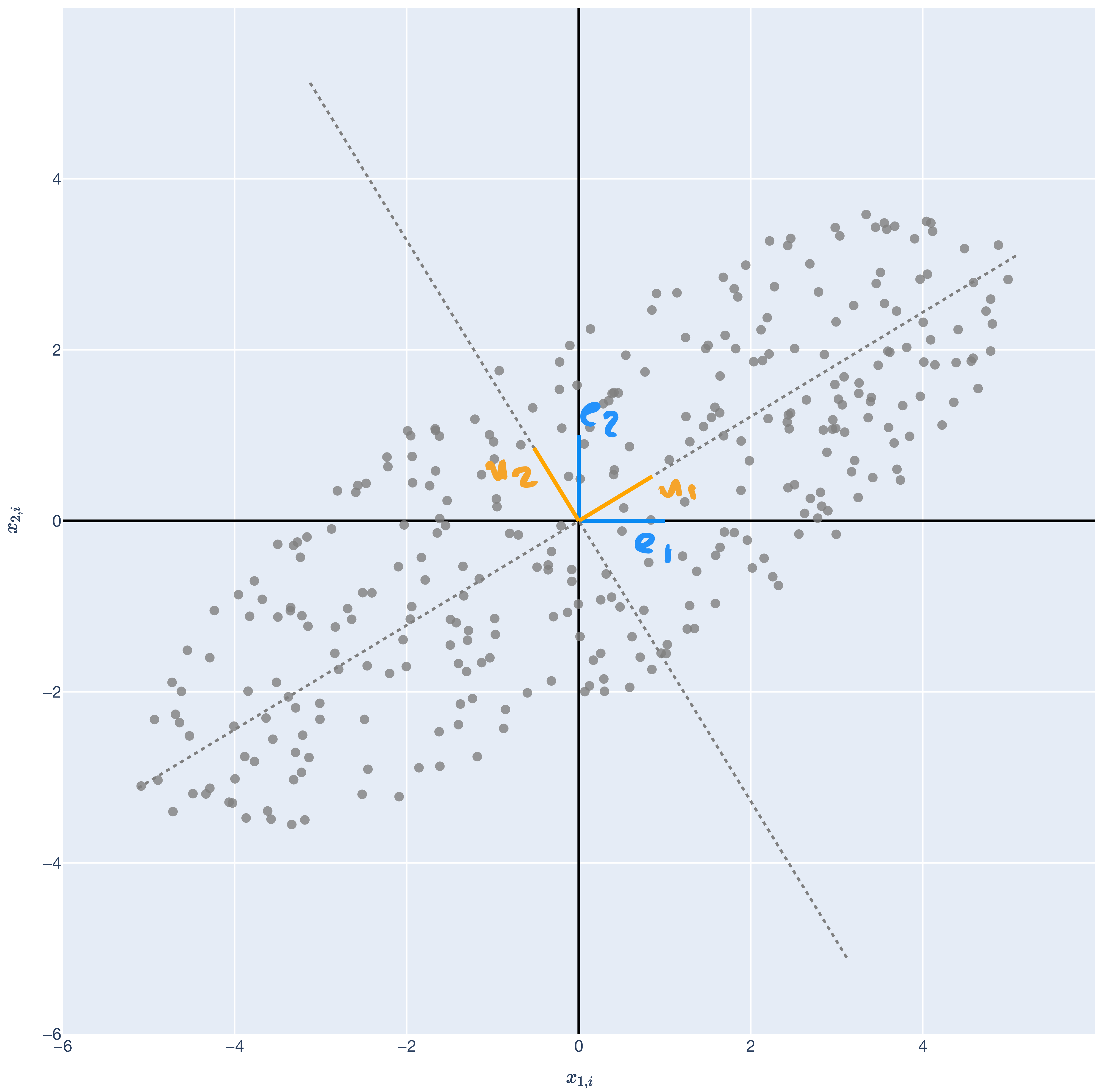
Interpreting the $\mathbf{U}$ matrix

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

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

$$\begin{bmatrix} | & | \\ u_1 & u_2 \\ | & | \end{bmatrix}$$

2×2.



Singular Values

Interpreting the Σ matrix

$$\begin{bmatrix} a & 0 \\ 0 & b \end{bmatrix}$$

$$U\Sigma = \begin{bmatrix} | & | \\ u_1 & u_2 \\ | & | \end{bmatrix}\begin{bmatrix} \sigma_1 & 0 & \vdots & 0 & 0 \\ 0 & \sigma_2 & \vdots & & \end{bmatrix}$$

2×212

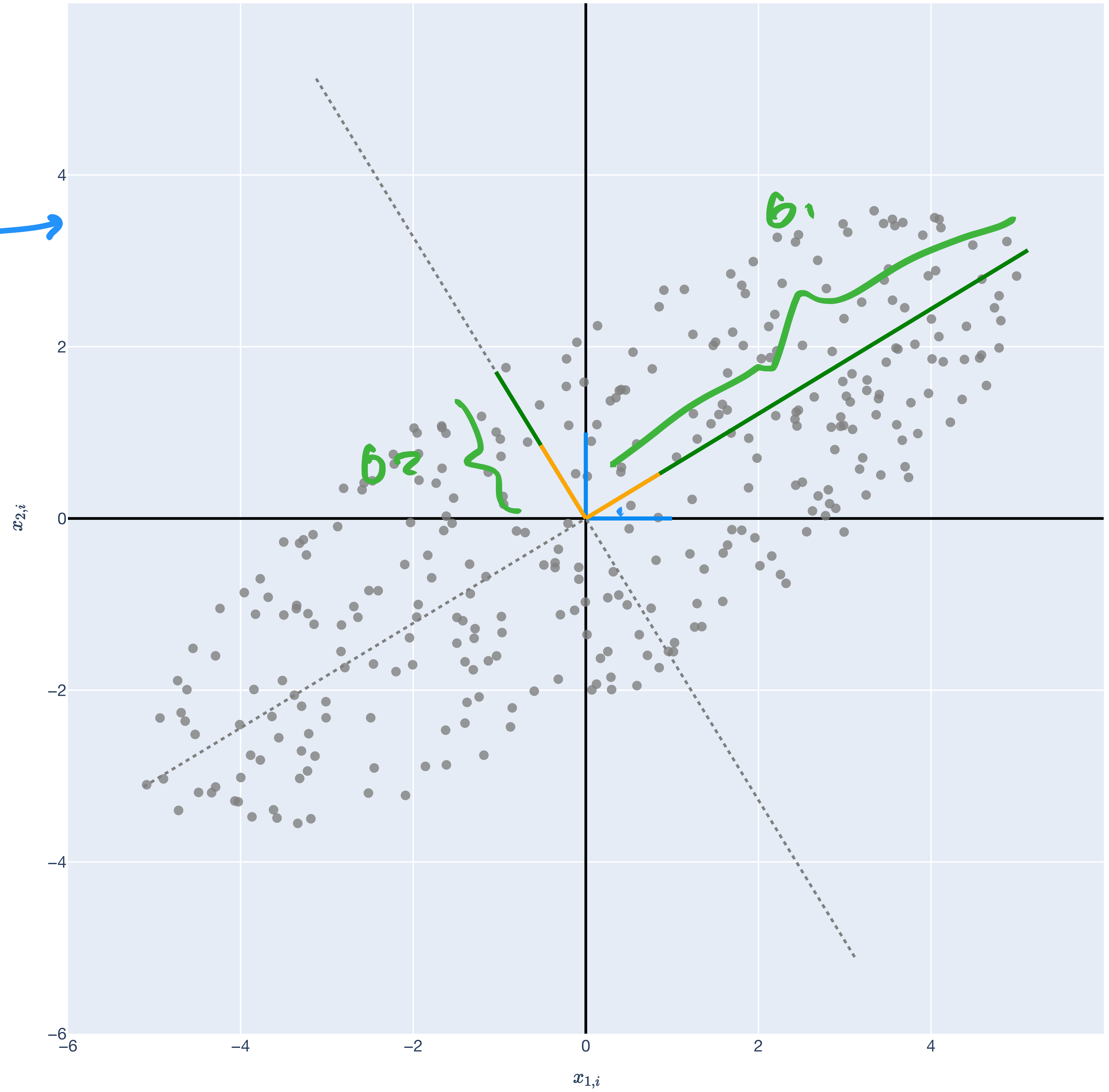
$$= \begin{bmatrix} | & | & \vdots & & & \\ \sigma_1 u_1 & \sigma_2 u_2 & \vdots & 0 & 0 & 0 \\ | & | & \vdots & & & \end{bmatrix}$$

$$\hat{w} = U^\top y$$

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

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

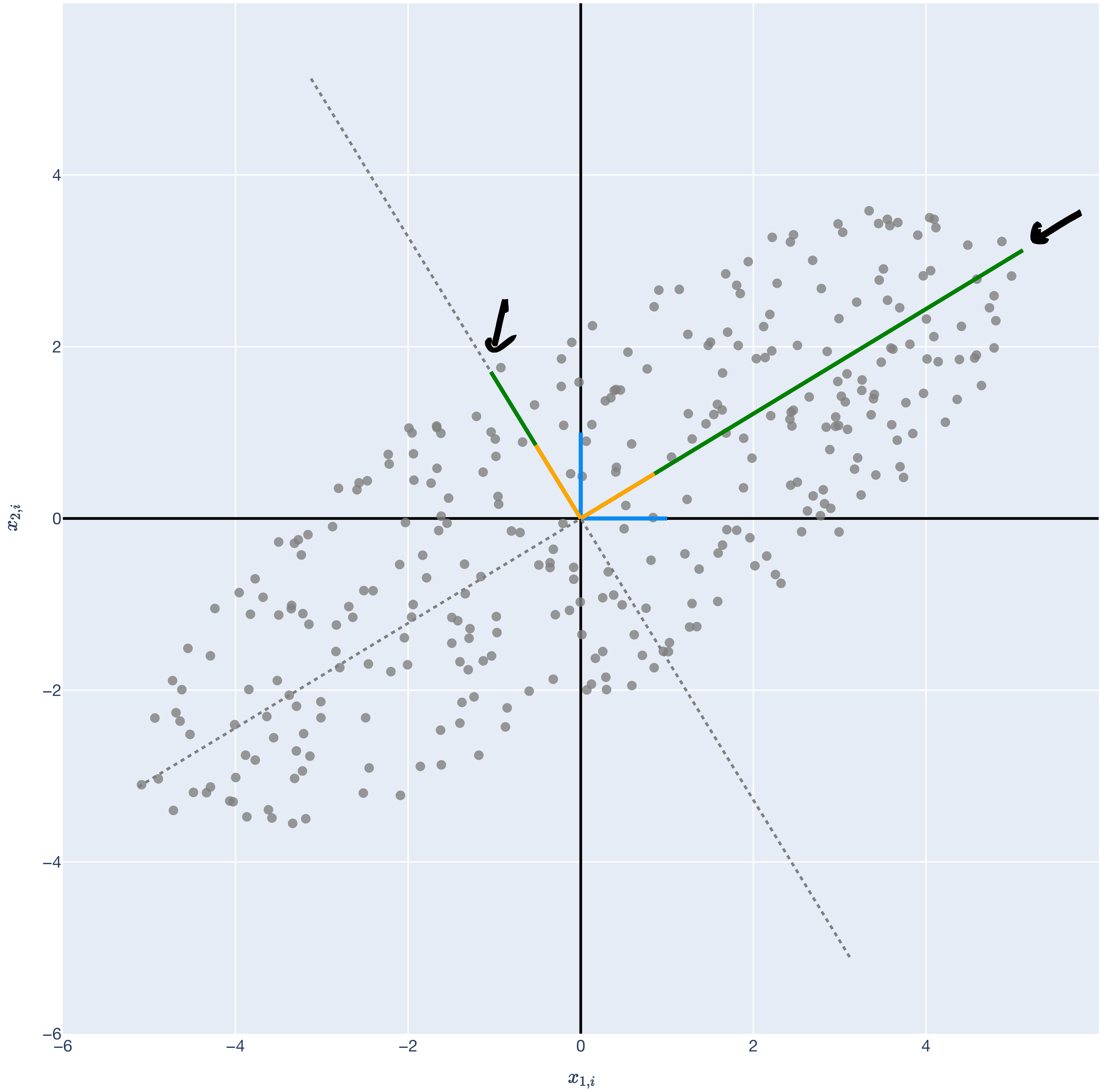
$$V = \begin{array}{c} \quad j \\ \begin{bmatrix} v_{1j} \\ v_{2j} \\ \cdots \end{bmatrix} \end{array} \quad \text{doesn't matter.}$$

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

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

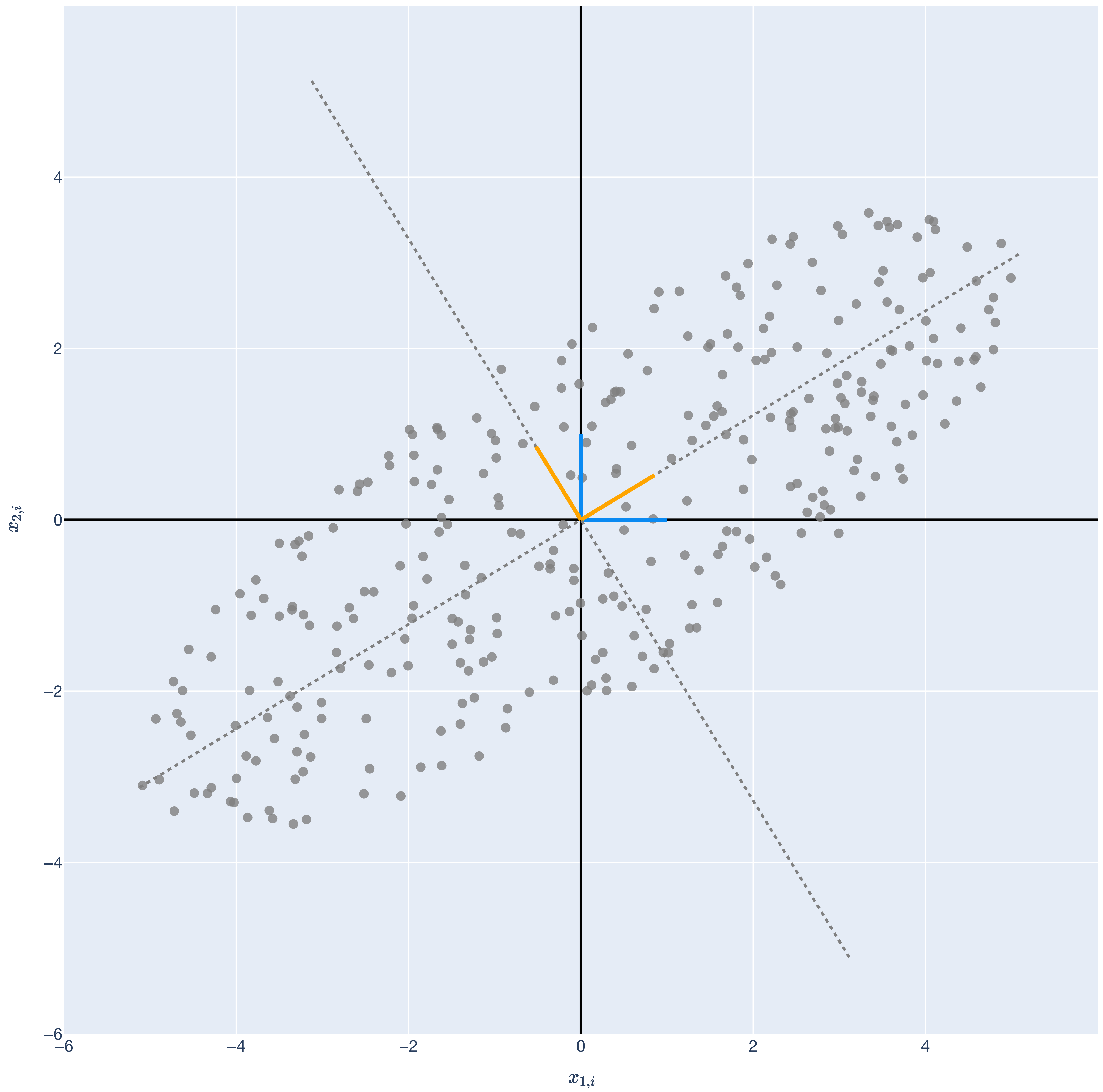
Specifically, for $j \in [d]$,

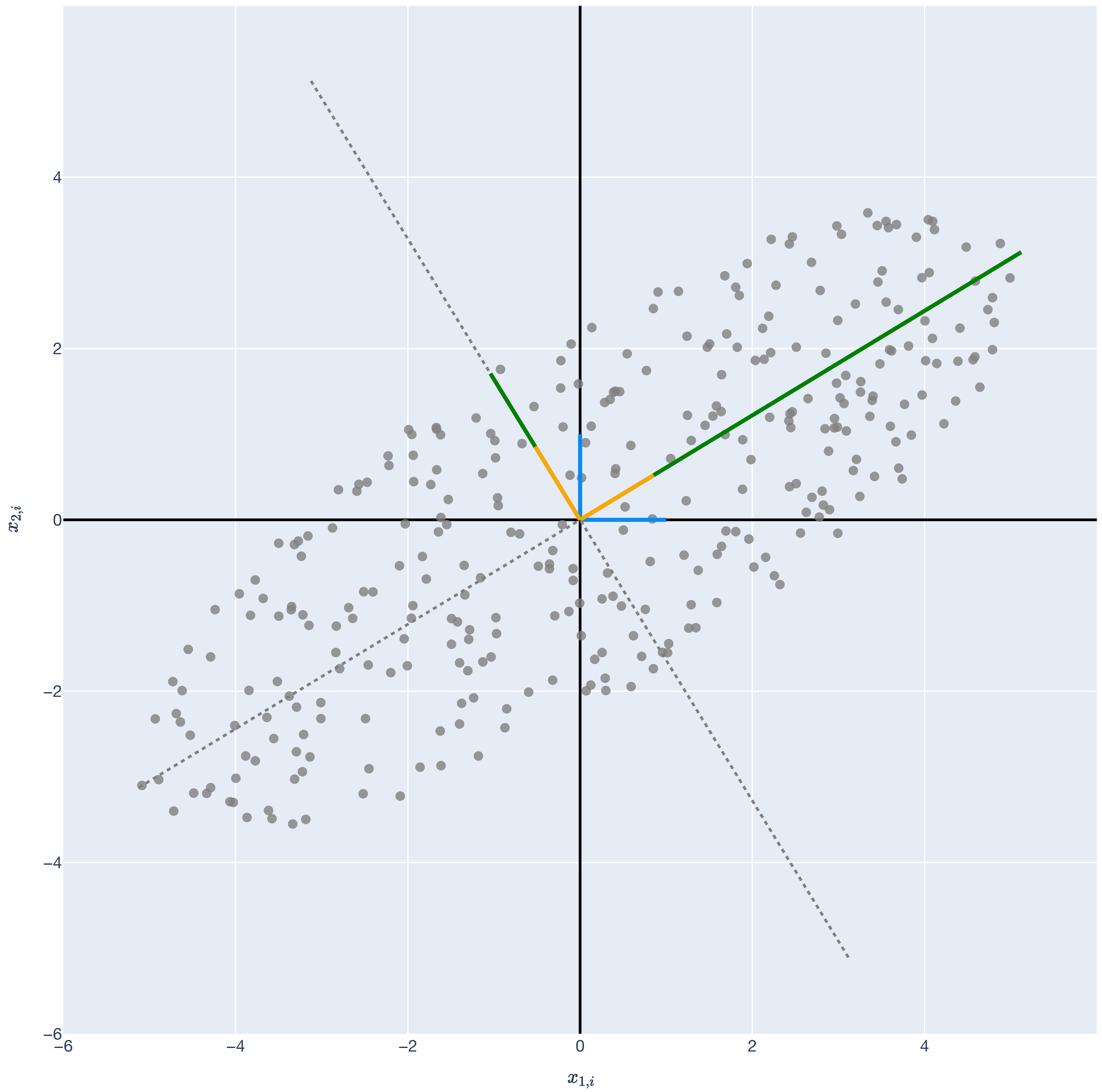
$$\boxed{\mathbf{x}_j} = v_{1j}\sigma_1\mathbf{u}_1 + v_{2j}\sigma_2\mathbf{u}_2.$$



Interpretation of the SVD

Full Interpretation of the SVD





Singular Value Decomposition (SVD)

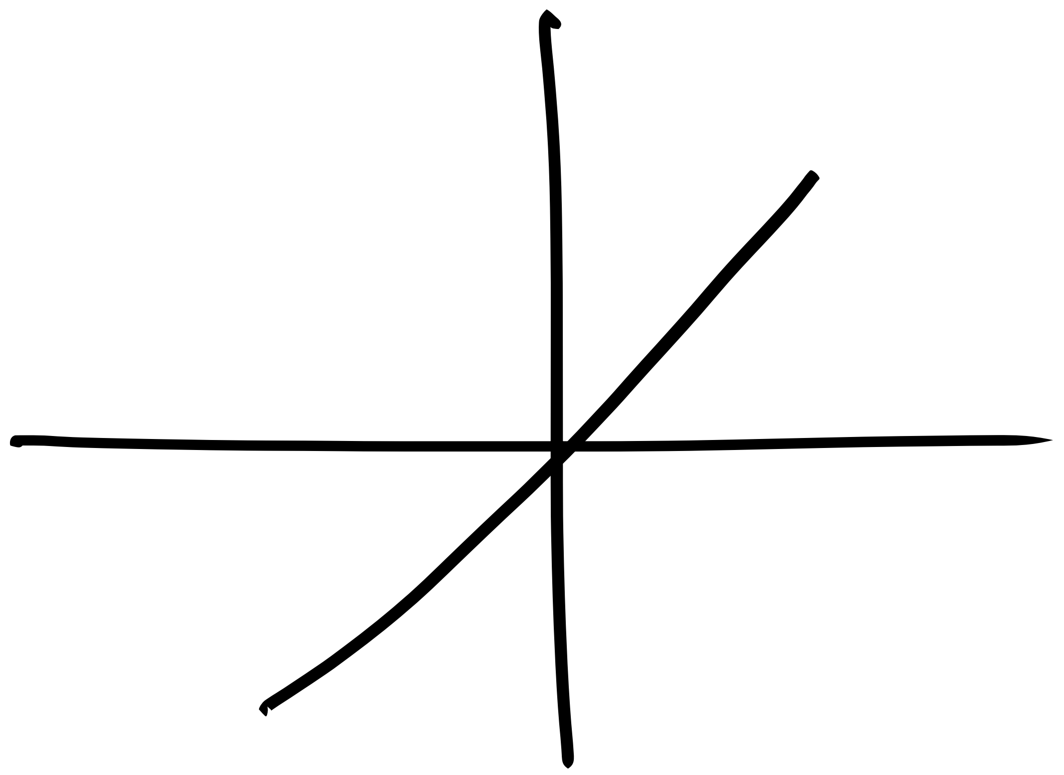
Example of SVD

U, S, Vt = np.linalg.svd (X)

3×3

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

$n = 3$
$d = 3$



$X = U \Sigma V^T$

$$U = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix}$$

$$\Sigma = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 5 & 0 \\ 0 & 0 & 1 \end{bmatrix}$$

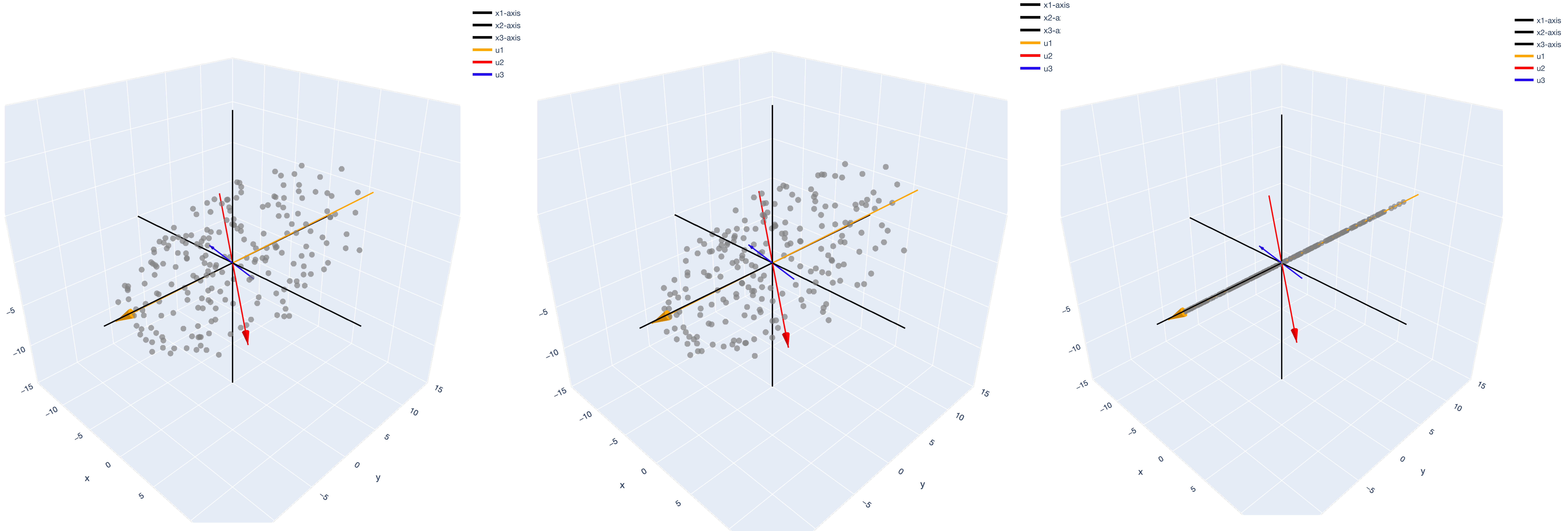
$$V^T = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix}$$

$$X = U \Sigma V^T = \underbrace{\begin{array}{ccc} u_1 & u_2 & u_3 \\ \end{array}\begin{bmatrix} 0 & 0 & 1 \\ 0 & 1 & 0 \\ 1 & 0 & 0 \end{bmatrix}}_{U} \underbrace{\begin{bmatrix} 10 & 0 & 0 \\ 0 & 5 & 0 \\ 0 & 0 & 1 \end{bmatrix}}_{\Sigma} \underbrace{\begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix}}_{V^T}$$

Singular Value Decomposition (SVD)

Example in $\mathbb{R}^3$

$$X = \begin{bmatrix} | & & | \\ x_1 & \cdots & x_{212} \\ | & & | \end{bmatrix} \quad \begin{matrix} n=3 \\ d=212 \end{matrix}$$



Singular Value Decomposition (SVD)

Definition of the Compact SVD

np.linalg.svd(full = False)

Consider any matrix $\mathbf{X} \in \mathbb{R}^{n \times d}$ with rank $r \leq \min\{n, d\}$. By the ***compact singular value decomposition (SVD)***, there exist matrices $\mathbf{U}, \Sigma, \mathbf{V}$ such that

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

diagonal .

The columns of $\mathbf{U} \in \mathbb{R}^{n \times r}$ are the ***left singular vectors*** and $\mathbf{U}^\top \mathbf{U} = \mathbf{I}$. They form an orthonormal basis for $\mathrm{span}(\mathrm{col}(\mathbf{X}))$, the columnspace of $\mathbf{X}$.

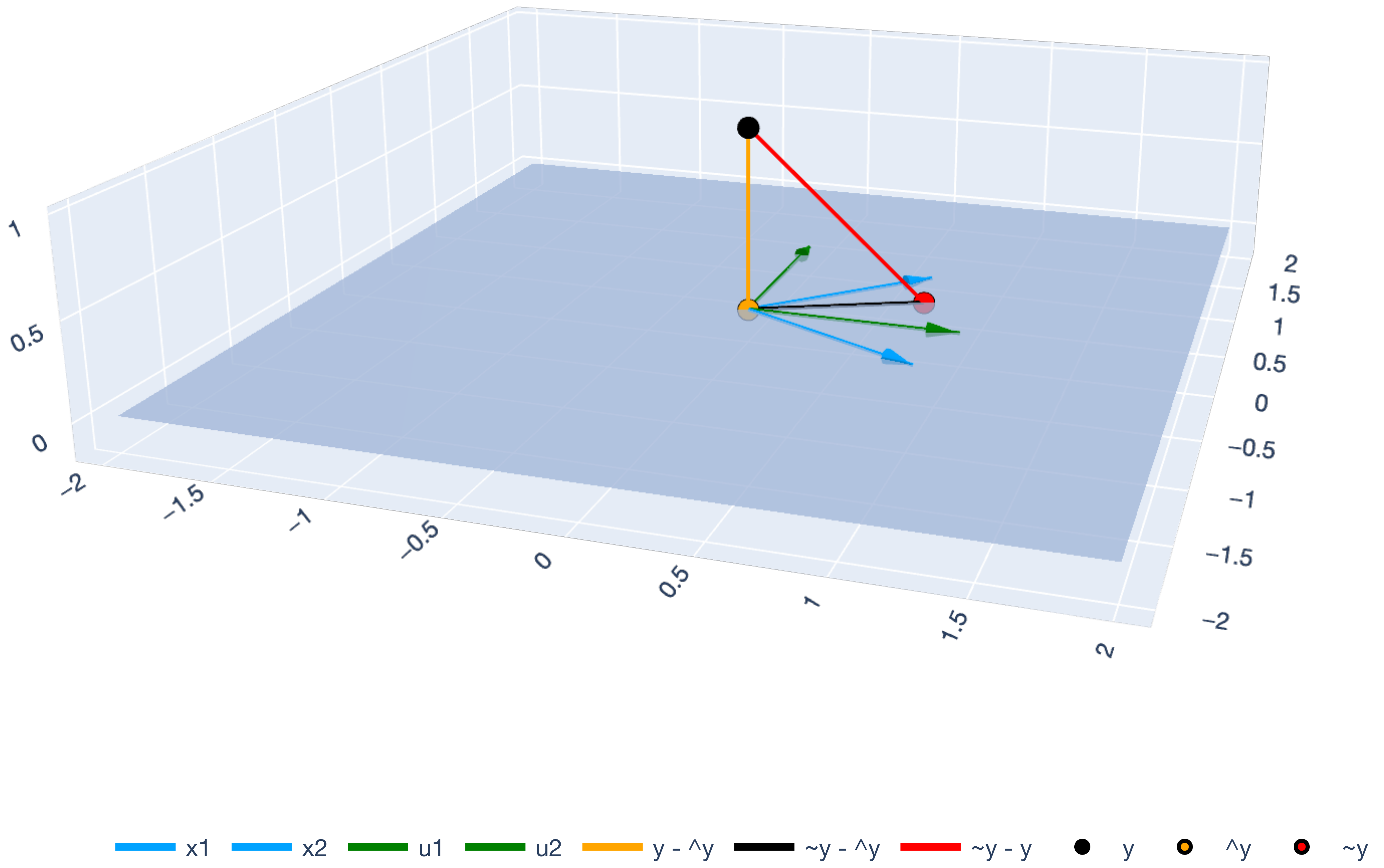
The columns of $\mathbf{V} \in \mathbb{R}^{r \times d}$ are the ***right singular vectors*** and $\mathbf{V}^\top \mathbf{V} = \mathbf{I}$. They form an orthonormal basis for $\mathrm{span}(\mathrm{row}(\mathbf{X}))$, the rowspace of $\mathbf{X}$.

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

How to find a ~~good~~ Best orthogonal basis?

numpy

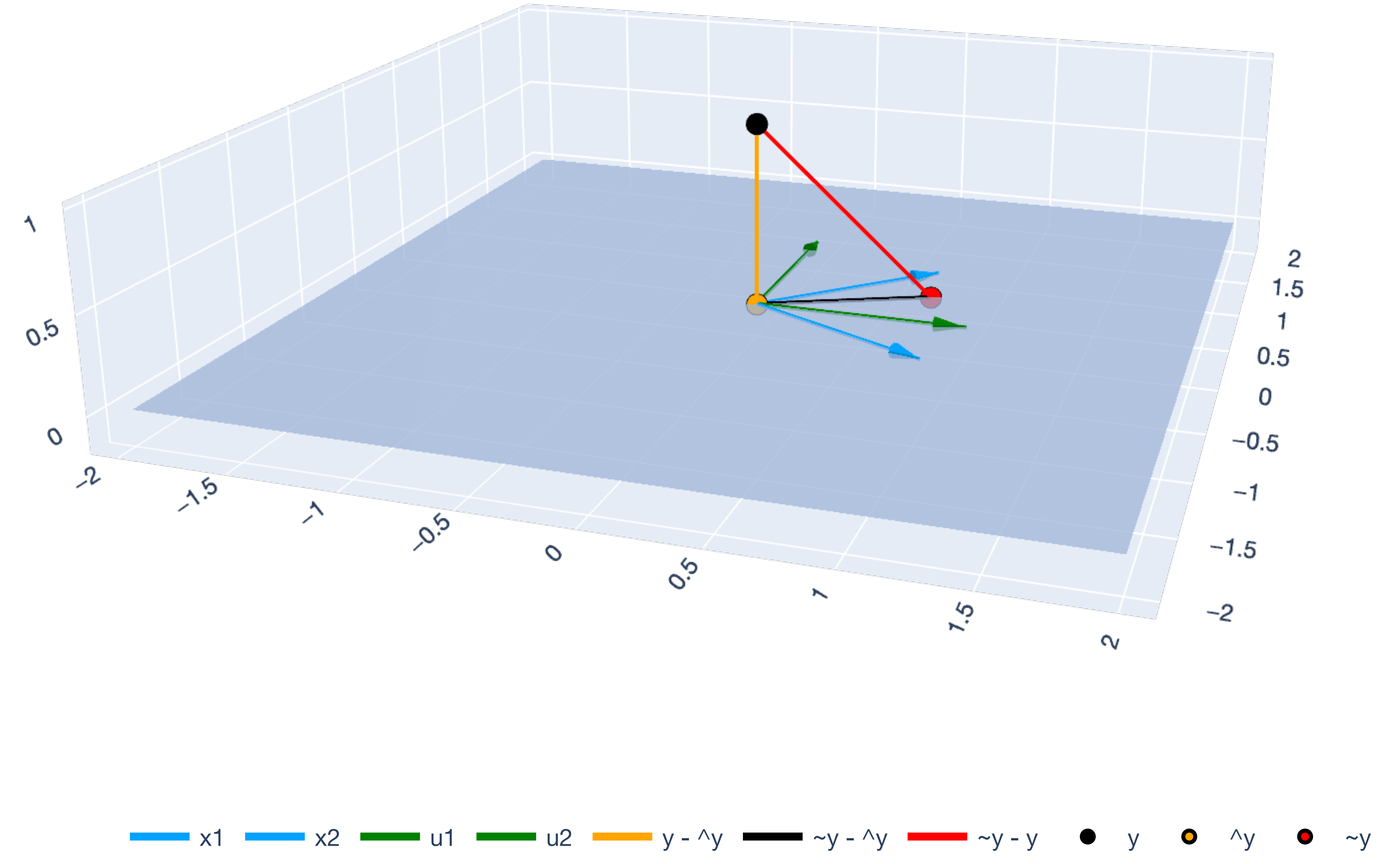
np.linalg.svd(x)



Least Squares
OLS with Orthogonal Basis

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

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



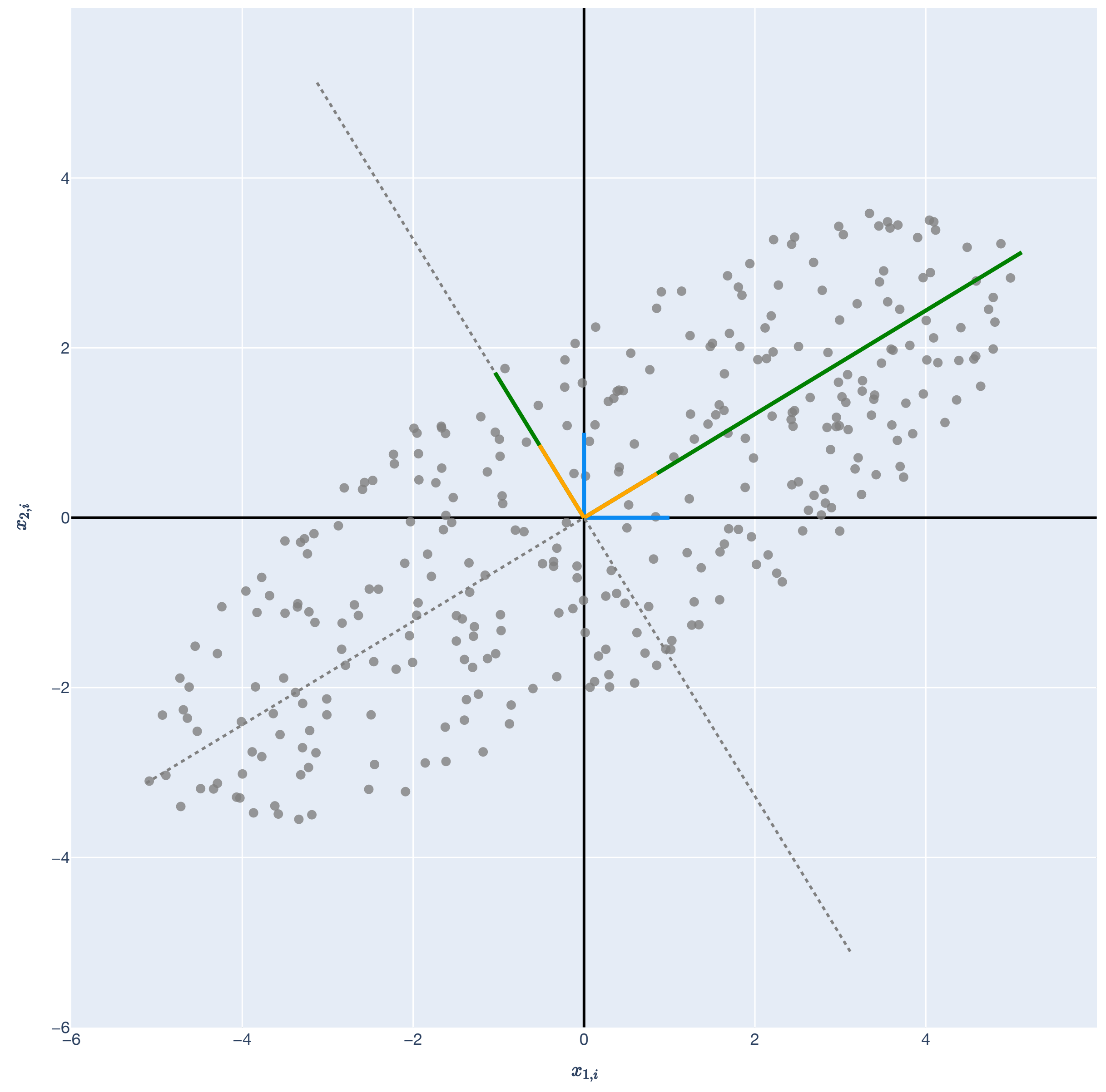
Least Squares

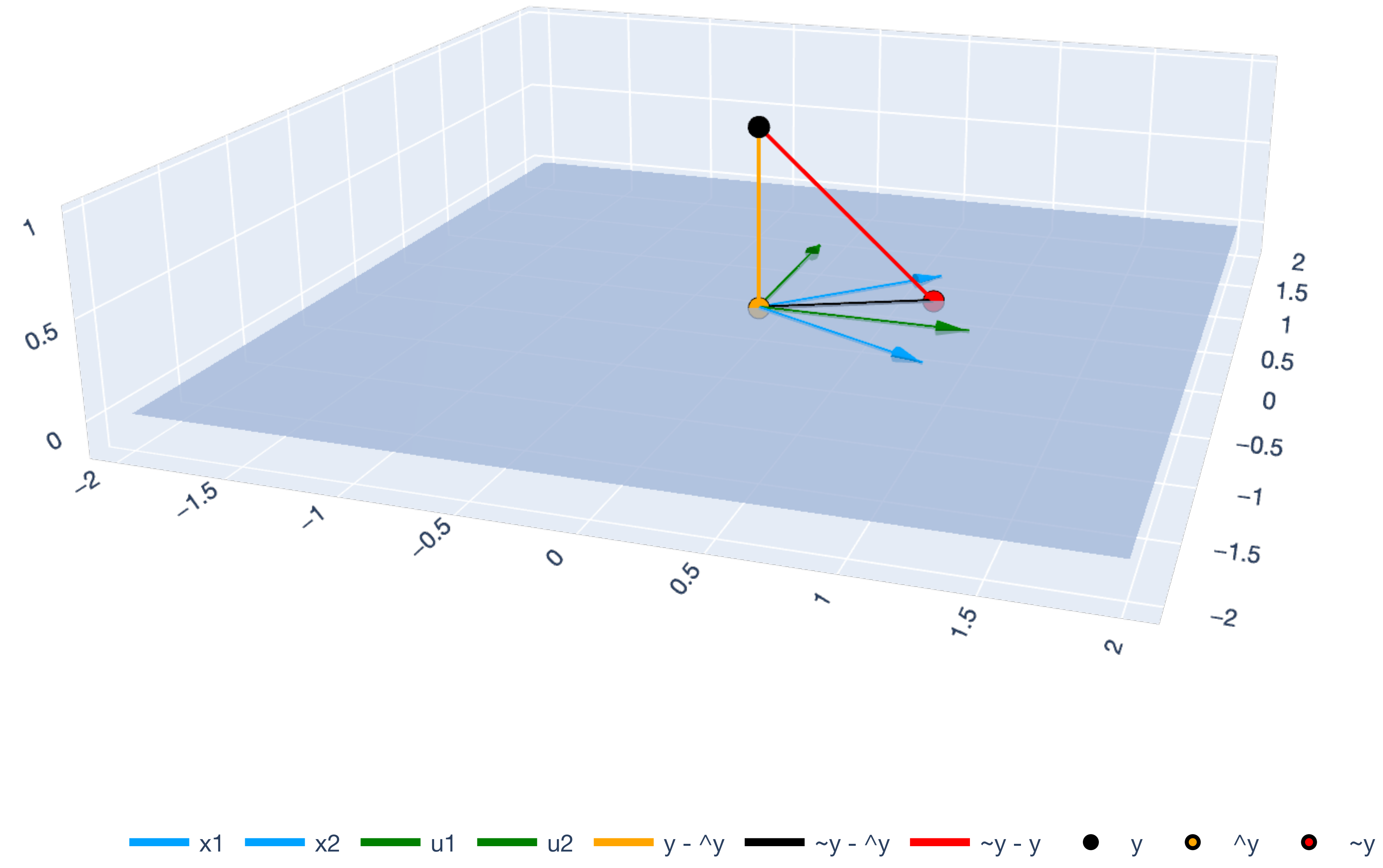
OLS with Orthogonal Basis

<u>Prop (OLS using the ONB from Compact SVD).</u>
Let $\mathbf{X} \in \mathbb{R}^{n \times d}$ and let $\mathscr{X} = \mathrm{span}(\mathrm{col}(\mathbf{X}))$ be a subspace, with dimension $\dim(\mathscr{X}) = \mathrm{rank}(\mathbf{X}) = r$.

Then, if the compact SVD of $\boxed{\mathbf{X} = \mathbf{U}\mathbf{\Sigma}\mathbf{V}^\top}$, then the columns $\mathbf{u}_1, \ldots, \mathbf{u}_r \in \mathbb{R}^n$ of $\mathbf{U}$ are an ONB for $\mathscr{X}$ and, hence, for any $\mathbf{y} \in \mathbb{R}^n$, the projection of $\mathbf{y}$ onto $\mathscr{X}$ is given by:

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





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 $r = \text{rank}(\mathbf{X})$ by asking:

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

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 $r = \mathrm{rank}(\mathbf{X})$ 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:***

$$\boxed{\|\mathbf{X}\|_F := \sqrt{\sum_{i=1}^{n} \sum_{j=1}^{d} X_{ij}^2}.}$$

Rank-*k* Approximation

Statement

Eckhart - Young

Theorem (Rank-k Approximation). Let $\mathbf{X} \in \mathbb{R}^{n\times d}$. Let $\hat{\mathbf{X}}_k \in \mathbb{R}^{n\times d}$ be the rank-k approximation of $\mathbf{X}$ in Frobenius norm:

closest

$$\hat{\mathbf{X}}_k = \underset{\hat{\mathbf{X}}\in\mathbb{R}^{n\times d}}{\arg\min}\ \underbrace{\|\mathbf{X} - \hat{\mathbf{X}}_k\|_F}_{},$$

such that $\mathrm{rank}(\hat{\mathbf{X}}) = k$.

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

$$V = \begin{bmatrix} | & & | \\ u_1 & \cdots & u_r \\ | & & | \end{bmatrix}_{n\times r}$$

Then, if $\boxed{\mathbf{X} = \mathbf{U}\boldsymbol{\Sigma}\mathbf{V}^\top}$ is the compact SVD of $\mathbf{X}$ with $\mathbf{U}_k \in \mathbb{R}^{n\times k}$, $\underbrace{\boldsymbol{\Sigma} \in \mathbb{R}^{k\times k}}_{}$, and $\mathbf{V} \in \mathbb{R}^{d\times k}$ are truncated matrices of $\mathbf{U}$, $\boldsymbol{\Sigma}$, and $\mathbf{V}$, respectively, then

$$\Sigma = \begin{bmatrix} r\times r \end{bmatrix}$$

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

Rank-*k* Approximation

Outer Product Interpretation

NEXT TIME

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:

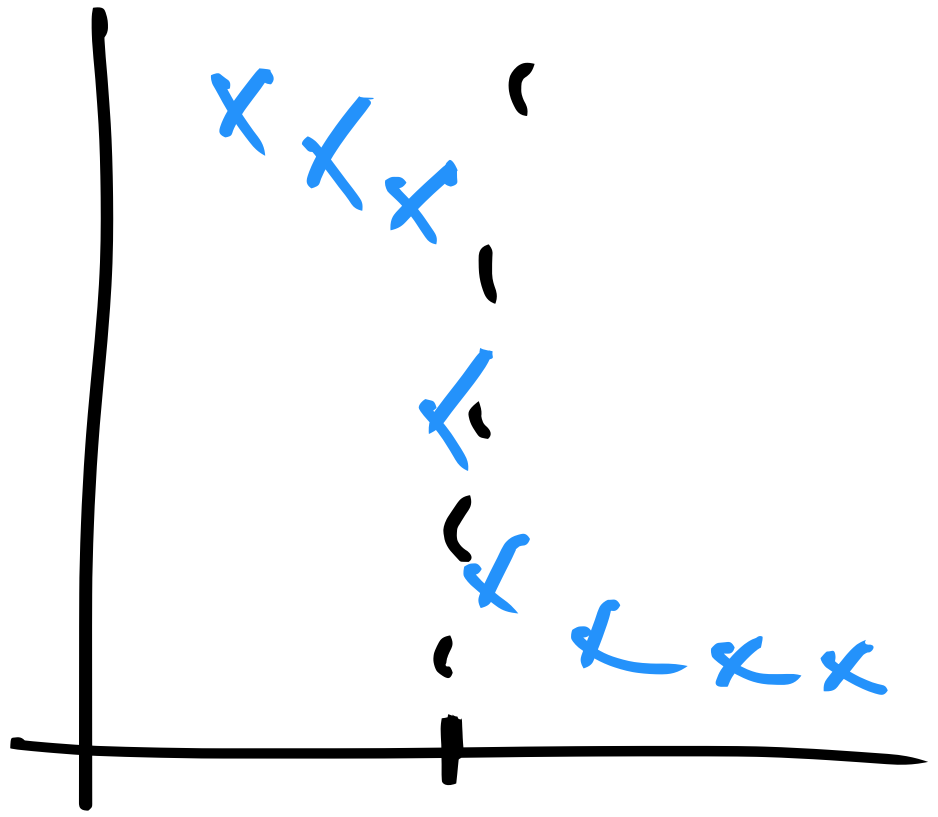
$$\hat{\mathbf{X}}_k = \sigma_1 \mathbf{u}_1 \mathbf{v}_1^\top + \ldots + \sigma_k \mathbf{u}_k \mathbf{v}_k^\top .$$

Rank-*k* Approximation

Example

$$\boxed{k=2}:\quad \hat{X} = \underbrace{\begin{bmatrix} 1 & 0 \\ 0 & 1 \\ 0 & 0 \\ 0 & 0 \end{bmatrix} \begin{bmatrix} 100 & 0 \\ 0 & 90 \end{bmatrix} \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \end{bmatrix}}_{n \times d}$$

Consider the 4×4 matrix



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

$$X = \underbrace{\begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix}}_{V} \begin{bmatrix} 100 & 0 & 0 & 0 \\ 0 & 90 & 0 & 0 \\ 0 & 0 & 8 & 0 \\ 0 & 0 & 0 & 2 \end{bmatrix} \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix}$$

Rank-*k* Approximation

Application in Image Processing

$n = 3000$
$d = 4000$

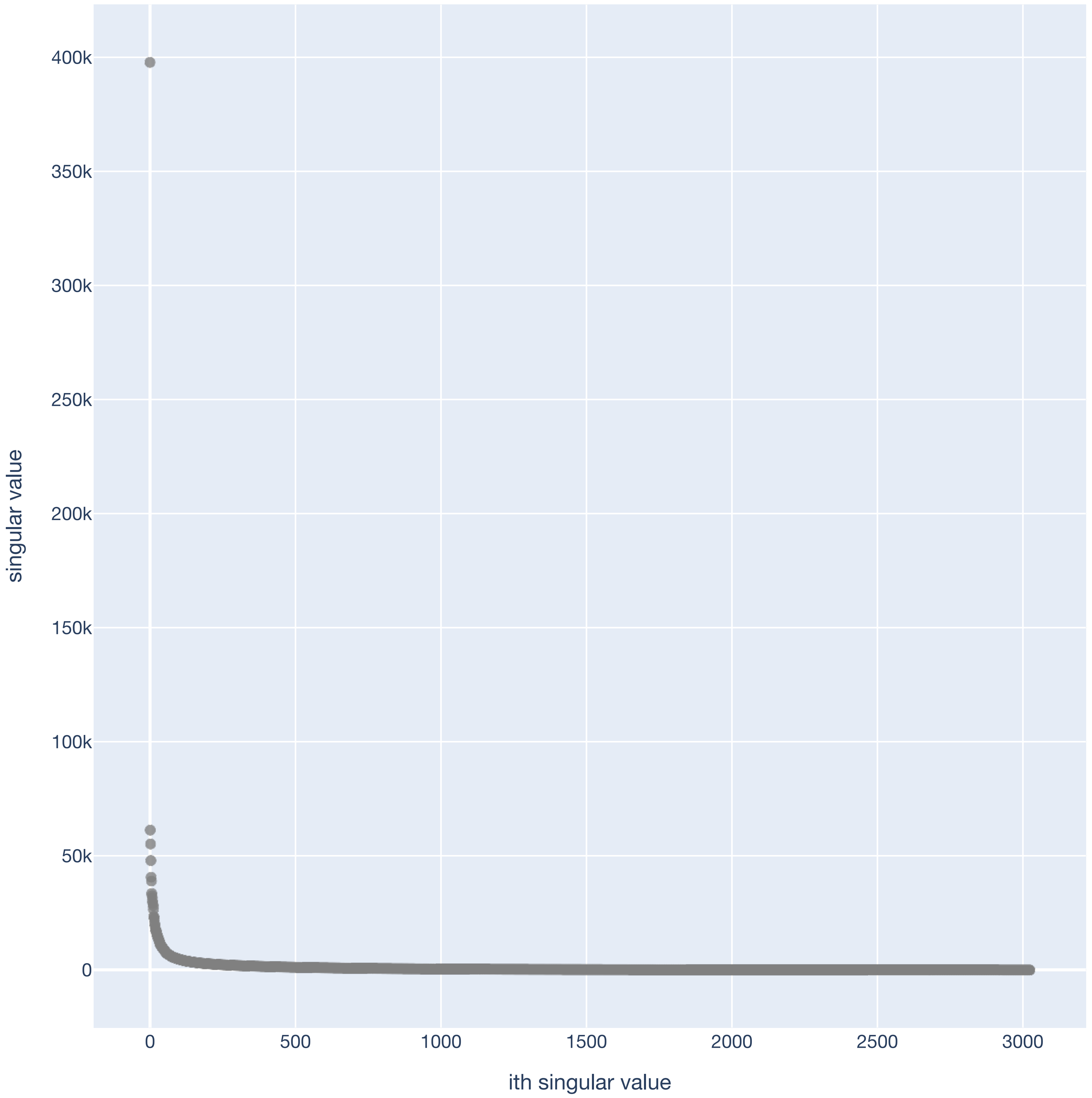

```
print(X)
print("Shape: {}".format(X.shape))
```

✓ 0.0s

```
[[ 78  78  78 ... 124 122 129]
 [ 82  81  79 ... 124 121 126]
 [ 81  80  78 ... 120 123 127]
 ...
 [ 40  42  40 ... 116  99 118]
 [ 40  41  40 ... 111 114 119]
 [ 41  39  40 ... 120 122  96]]
Shape: (3024, 4032)
```

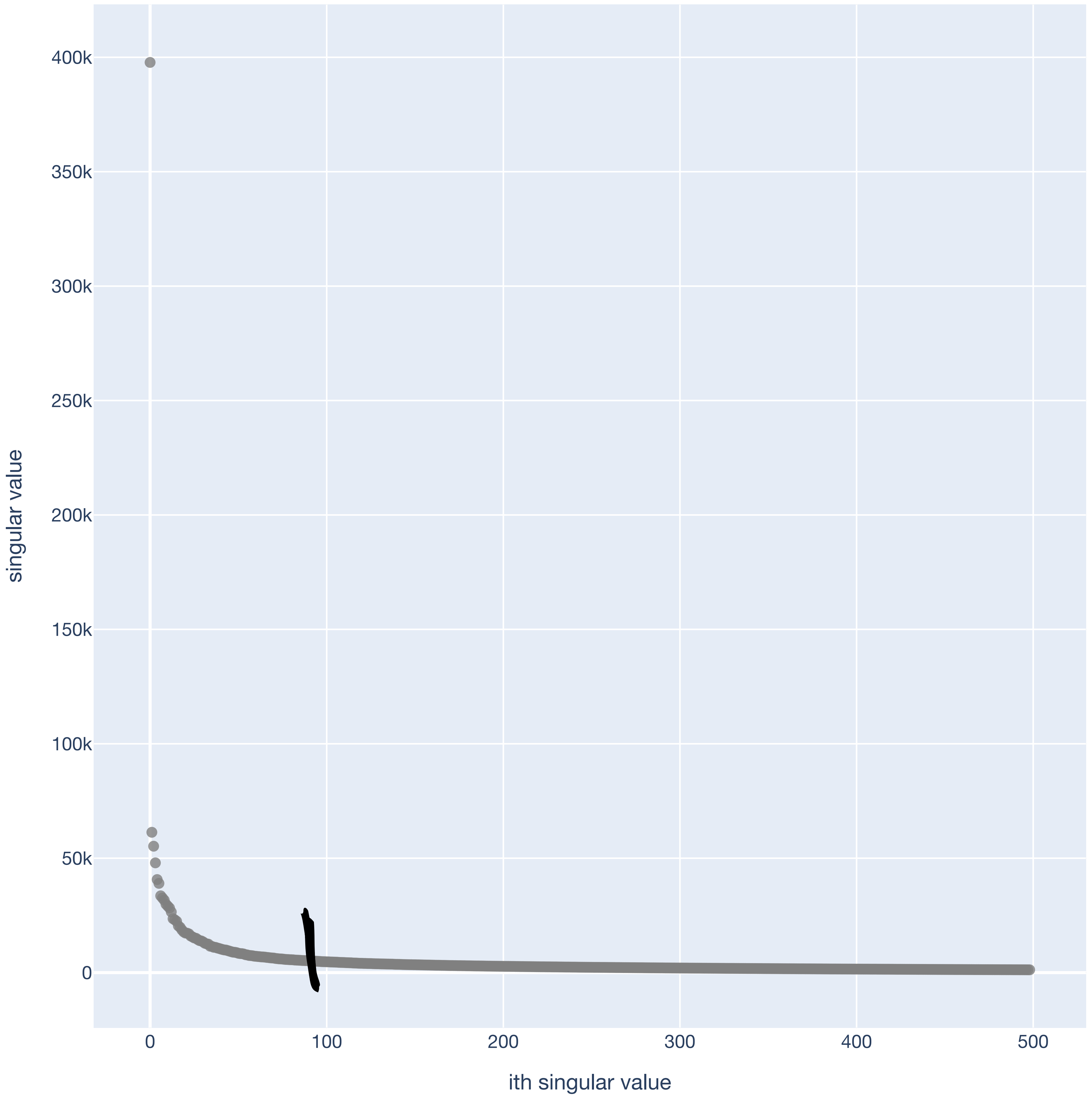
Rank-*k* Approximation

Application in Image Processing

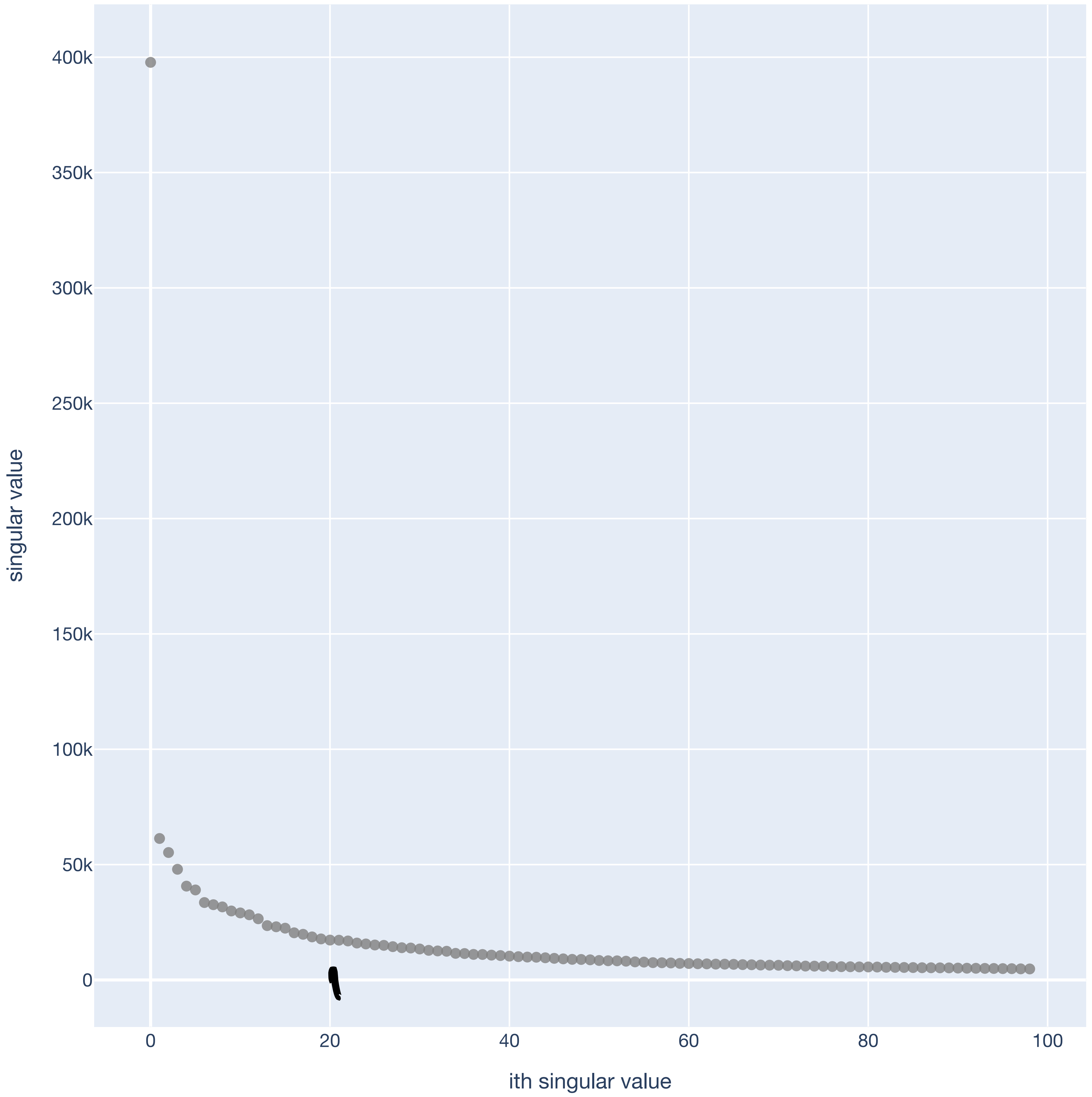


Rank-*k* Approximation

Application in Image Processing (*k* = *500*)



Rank-*k* Approximation
Application in Image Processing (*k* = *100*)



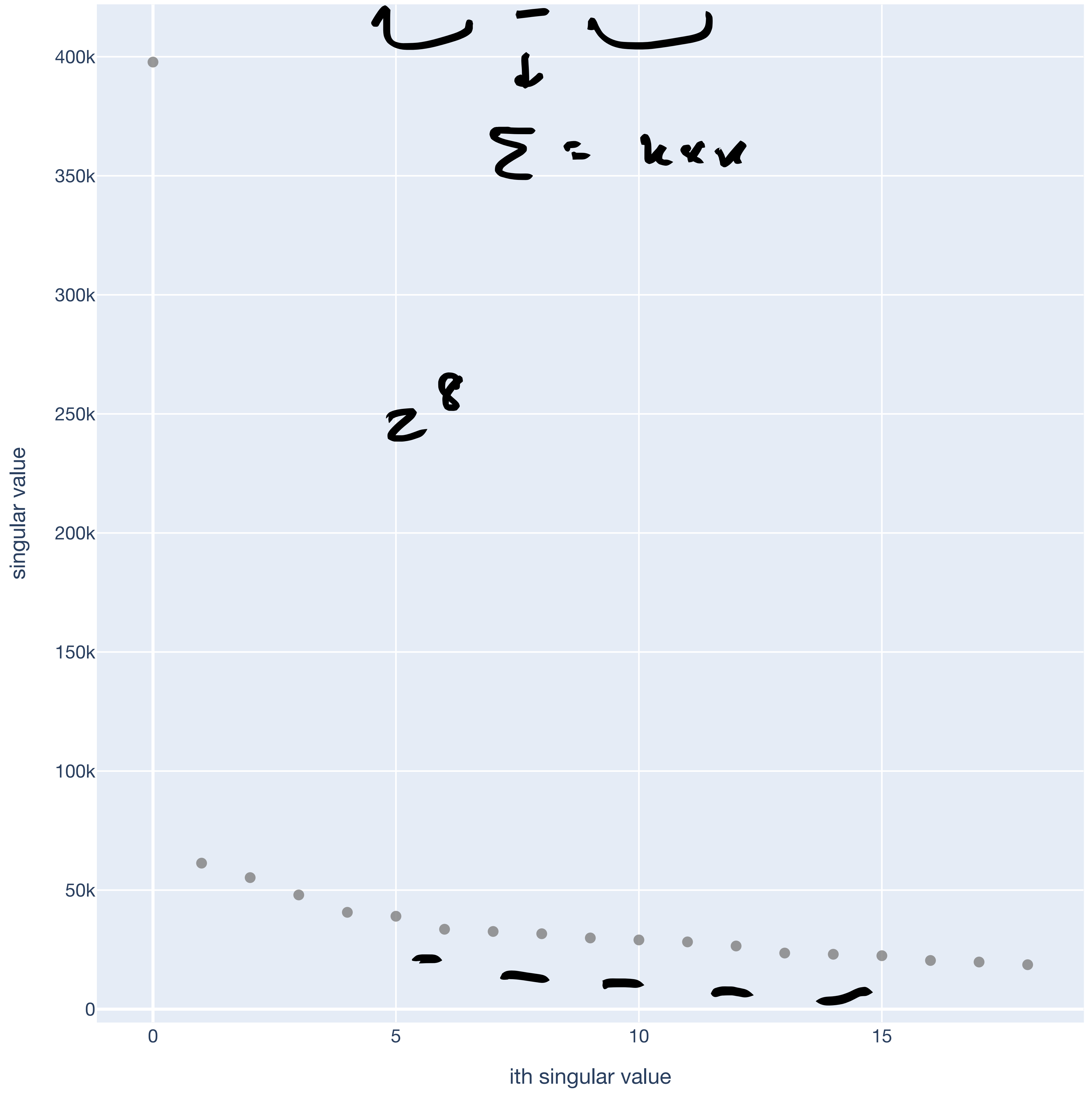
Rank-*k* Approximation

Application in Image Processing (*k* = *20*)

$X \in \mathbb{R}^{3000 \times 4000}$

np.linalg.svd(X)

$X = \underbrace{U}_{n \times r} \underbrace{\Sigma}_{r \times r} \underbrace{V^T}_{r \times d}$ $\qquad k << r$



$\Sigma = k \times k$

2^8

Least Squares

SVD and the Pseudoinverse

Regression

Setup

$$Xw = y$$
$d\times d$ $\quad d \quad d$

$$w = X^{-1}y$$

① $X \in \mathbb{R}^{d\times d}$
② $\mathrm{rank}(X) = d$

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

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

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:

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

Regression

Setup

<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": $\hat{\mathbf{w}} \in \mathbb{R}^d$ such that $y_i \approx \hat{y}_i$ for $i \in [n]$, or:

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

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

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

Least Squares
Main 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 $\operatorname{rank}(\mathbf{X}) = d$, 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

Plugging in the SVD

① $(AB)^T = B^T A^T$

② $(A^{-1})^T = (A^T)^{-1}$

$n \times d$

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

$(AB)^T = B^T A^T$

$(ABC)^T = C^T B^T A^T$

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?

$X^T = (U\Sigma V^T)^T = V\Sigma^T U^T = V\Sigma U^T$

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

$$= (\mathbf{V}\boldsymbol{\Sigma}\mathbf{U}^\top\mathbf{U}\boldsymbol{\Sigma}\mathbf{V}^\top)^{-1}\mathbf{V}\boldsymbol{\Sigma}\mathbf{U}^\top\mathbf{y} \qquad (\mathbf{X}^\top = \mathbf{V}\boldsymbol{\Sigma}\mathbf{U}^\top)$$

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?

$$
\begin{aligned}
\hat{\mathbf{w}} &= (\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{X}^\top\mathbf{y} \\
&= (\mathbf{V}\boldsymbol{\Sigma}\underbrace{\mathbf{U}^\top\mathbf{U}}\boldsymbol{\Sigma}\mathbf{V}^\top)^{-1}\mathbf{V}\boldsymbol{\Sigma}\mathbf{U}^\top\mathbf{y} \qquad && (\mathbf{X}^\top = \mathbf{V}\boldsymbol{\Sigma}\mathbf{U}^\top) \\
&= (\mathbf{V}\boldsymbol{\Sigma}^\top\boldsymbol{\Sigma}\mathbf{V}^\top)^{-1}\mathbf{V}\boldsymbol{\Sigma}\mathbf{U}^\top\mathbf{y} && (\underbrace{\mathbf{U}^\top\mathbf{U} = \mathbf{I}})
\end{aligned}
$$

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?

$$
\begin{aligned}
\hat{\mathbf{w}} &= (\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{X}^\top\mathbf{y} & \\
&= (\mathbf{V}\boldsymbol{\Sigma}\mathbf{U}^\top\mathbf{U}\boldsymbol{\Sigma}\mathbf{V}^\top)^{-1}\mathbf{V}\boldsymbol{\Sigma}\mathbf{U}^\top\mathbf{y} & (\mathbf{X}^\top = \mathbf{V}\boldsymbol{\Sigma}\mathbf{U}^\top) \\
&= (\underbrace{\mathbf{V}\boldsymbol{\Sigma}^\top\boldsymbol{\Sigma}}_{A}\underbrace{\mathbf{V}^\top}_{B})^{-1}\mathbf{V}\boldsymbol{\Sigma}\mathbf{U}^\top\mathbf{y} & (\mathbf{U}^\top\mathbf{U} = \mathbf{I}) \\
&= (\underbrace{\mathbf{V}^\top}_{B})^{-1}(\underbrace{\mathbf{V}\boldsymbol{\Sigma}^\top\boldsymbol{\Sigma}}_{A})^{-1}\mathbf{V}\boldsymbol{\Sigma}^\top\mathbf{U}^\top\mathbf{y} & (\underline{(\mathbf{AB})^{-1} = \mathbf{B}^{-1}\mathbf{A}^{-1}})
\end{aligned}
$$

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?

$$
\begin{aligned}
\hat{\mathbf{w}} &= (\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{X}^\top\mathbf{y} \\
&= (\mathbf{V}\mathbf{\Sigma}\mathbf{U}^\top\mathbf{U}\mathbf{\Sigma}\mathbf{V}^\top)^{-1}\mathbf{V}\mathbf{\Sigma}\mathbf{U}^\top\mathbf{y} && (\mathbf{X}^\top = \mathbf{V}\mathbf{\Sigma}\mathbf{U}^\top) \\
&= (\mathbf{V}\mathbf{\Sigma}^\top\mathbf{\Sigma}\mathbf{V}^\top)^{-1}\mathbf{V}\mathbf{\Sigma}\mathbf{U}^\top\mathbf{y} && (\mathbf{U}^\top\mathbf{U} = \mathbf{I}) \\
&= (\mathbf{V}^\top)^{-1}(\mathbf{V}\mathbf{\Sigma}^\top\mathbf{\Sigma})^{-1}\mathbf{V}\mathbf{\Sigma}^\top\mathbf{U}^\top\mathbf{y} && ((\mathbf{A}\mathbf{B})^{-1} = \mathbf{B}^{-1}\mathbf{A}^{-1}) \\
&= \mathbf{V}(\mathbf{\Sigma}^\top\mathbf{\Sigma})^{-1}\mathbf{V}^\top\mathbf{V}\mathbf{\Sigma}^\top\mathbf{U}^\top\mathbf{y} && (\underline{\mathbf{V}^{-1} = \mathbf{V}^\top})
\end{aligned}
$$

$V^T V = V V^T = I$.

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?

$$
\begin{aligned}
\hat{\mathbf{w}} &= (\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{X}^\top\mathbf{y} \\
&= (\mathbf{V}\boldsymbol{\Sigma}\mathbf{U}^\top\mathbf{U}\boldsymbol{\Sigma}\mathbf{V}^\top)^{-1}\mathbf{V}\boldsymbol{\Sigma}\mathbf{U}^\top\mathbf{y} && (\mathbf{X}^\top = \mathbf{V}\boldsymbol{\Sigma}\mathbf{U}^\top) \\
&= (\mathbf{V}\boldsymbol{\Sigma}^\top\boldsymbol{\Sigma}\mathbf{V}^\top)^{-1}\mathbf{V}\boldsymbol{\Sigma}\mathbf{U}^\top\mathbf{y} && (\mathbf{U}^\top\mathbf{U} = \mathbf{I}) \\
&= (\mathbf{V}^\top)^{-1}(\mathbf{V}\boldsymbol{\Sigma}^\top\boldsymbol{\Sigma})^{-1}\mathbf{V}\boldsymbol{\Sigma}^\top\mathbf{U}^\top\mathbf{y} && ((\mathbf{A}\mathbf{B})^{-1} = \mathbf{B}^{-1}\mathbf{A}^{-1}) \\
&= \mathbf{V}(\boldsymbol{\Sigma}^\top\boldsymbol{\Sigma})^{-1}\mathbf{V}^\top\mathbf{V}\boldsymbol{\Sigma}^\top\mathbf{U}^\top\mathbf{y} && (\mathbf{V}^{-1} = \mathbf{V}^\top) \\
&= \mathbf{V}(\boldsymbol{\Sigma}^\top\boldsymbol{\Sigma})^{-1}\boldsymbol{\Sigma}^\top\mathbf{U}^\top\mathbf{y} && (\mathbf{V}^\top\mathbf{V} = \mathbf{I})
\end{aligned}
$$

Least Squares: SVD Perspective

Plugging in the SVD

$n \times d$

$(U\Sigma V^T)^{-1} = (V^T)^{-1}\Sigma^{-1}U^{-1}$

$= V\Sigma^{-1}U^T$

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?

$$
\begin{aligned}
\hat{\mathbf{w}} &= (\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{X}^\top\mathbf{y} & \\
&= (\mathbf{V}\mathbf{\Sigma}\mathbf{U}^\top\mathbf{U}\mathbf{\Sigma}\mathbf{V}^\top)^{-1}\mathbf{V}\mathbf{\Sigma}\mathbf{U}^\top\mathbf{y} & (\mathbf{X}^\top = \mathbf{V}\mathbf{\Sigma}\mathbf{U}^\top) \\
&= (\mathbf{V}\mathbf{\Sigma}^\top\mathbf{\Sigma}\mathbf{V}^\top)^{-1}\mathbf{V}\mathbf{\Sigma}\mathbf{U}^\top\mathbf{y} & (\mathbf{U}^\top\mathbf{U} = \mathbf{I}) \\
&= (\mathbf{V}^\top)^{-1}(\mathbf{V}\mathbf{\Sigma}^\top\mathbf{\Sigma})^{-1}\mathbf{V}\mathbf{\Sigma}^\top\mathbf{U}^\top\mathbf{y} & ((\mathbf{A}\mathbf{B})^{-1} = \mathbf{B}^{-1}\mathbf{A}^{-1}) \\
&= \mathbf{V}(\mathbf{\Sigma}^\top\mathbf{\Sigma})^{-1}\mathbf{V}^\top\mathbf{V}\mathbf{\Sigma}^\top\mathbf{U}^\top\mathbf{y} & (\mathbf{V}^{-1} = \mathbf{V}^\top) \\
&= \mathbf{V}\underbrace{(\mathbf{\Sigma}^\top\mathbf{\Sigma})^{-1}\mathbf{\Sigma}^\top}_{\Sigma^+}\mathbf{U}^\top\mathbf{y} & (\mathbf{V}^\top\mathbf{V} = \mathbf{I})
\end{aligned}
$$

$\Sigma^+ \rightarrow V\Sigma^+U^T y$

Pseudoinverse

Idea

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

$$\begin{bmatrix} \sigma_1 & & & 0 \\ & \sigma_2 & & \\ & & \ddots & \\ 0 & & & \sigma_d \end{bmatrix}$$

$\Sigma^T \Sigma \in \mathbb{R}^{d \times d}$

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

Therefore, we derived:

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

$n \times d$

$\rightarrow$ d pos. singular values $\sigma_1, \ldots, \sigma_d > 0$.

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

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

$\Sigma^T \Sigma \in \mathbb{R}^{d \times d}$

$(\Sigma^T \Sigma)^{-1} (\Sigma^T \Sigma) = I$

In this way, $\underbrace{\boxed{(\mathbf{\Sigma}^\top\mathbf{\Sigma})^{-1}\mathbf{\Sigma}^\top}}_{d \times n}$ acts "like an inverse" to $\mathbf{\Sigma}$, though $\underbrace{\mathbf{\Sigma}}_{n \times d}$ may not be square.

Pseudoinverse

Definition

$\Sigma \to \Sigma^+ = (\Sigma^T\Sigma)^{-1}\Sigma^T$

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

Σ^+

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

$X = U\Sigma V^T \Rightarrow X^{-1} = (U\Sigma V^T)^{-1}$

$= (V^T)^{-1}\Sigma^+ U^{-1}$

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 ***(Moore-Penrose) pseudoinverse***: $\mathbf{X}^+ := \mathbf{V}\mathbf{\Sigma}^+\mathbf{U}^\top$.

$= V\Sigma^+ U^T$

Note: If using the notation of the compact SVD, this is written differently.

Pseudoinverse

Main Property

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

$$\boxed{\mathbf{A}^+ = \mathbf{V\Sigma}^+\mathbf{U}^\top} = \mathbf{V}(\mathbf{\Sigma}^\top\mathbf{\Sigma})^{-1}\mathbf{\Sigma}^\top\mathbf{U}^\top \quad n>d$$

$$\mathbf{V}(\mathbf{\Sigma}^\top\mathbf{\Sigma})^{-1}\mathbf{\Sigma}^\top\cancel{\mathbf{U}^\top} \cancel{U} \Sigma V^\top$$

$$= V\cancel{(\Sigma^\top\Sigma)^{-1}}\cancel{\Sigma^\top\Sigma} V^\top$$

$$= VV^\top = I.$$

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 Σ^+

What does $\boldsymbol{\Sigma}^+ = (\boldsymbol{\Sigma}^\top \boldsymbol{\Sigma})^{-1} \boldsymbol{\Sigma}^\top$ look like?

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

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

Pseudoinverse

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

What does $\mathbf{\Sigma}^+ = (\mathbf{\Sigma}^\top \mathbf{\Sigma})^{-1}\mathbf{\Sigma}^\top$ look like?

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

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

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:

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

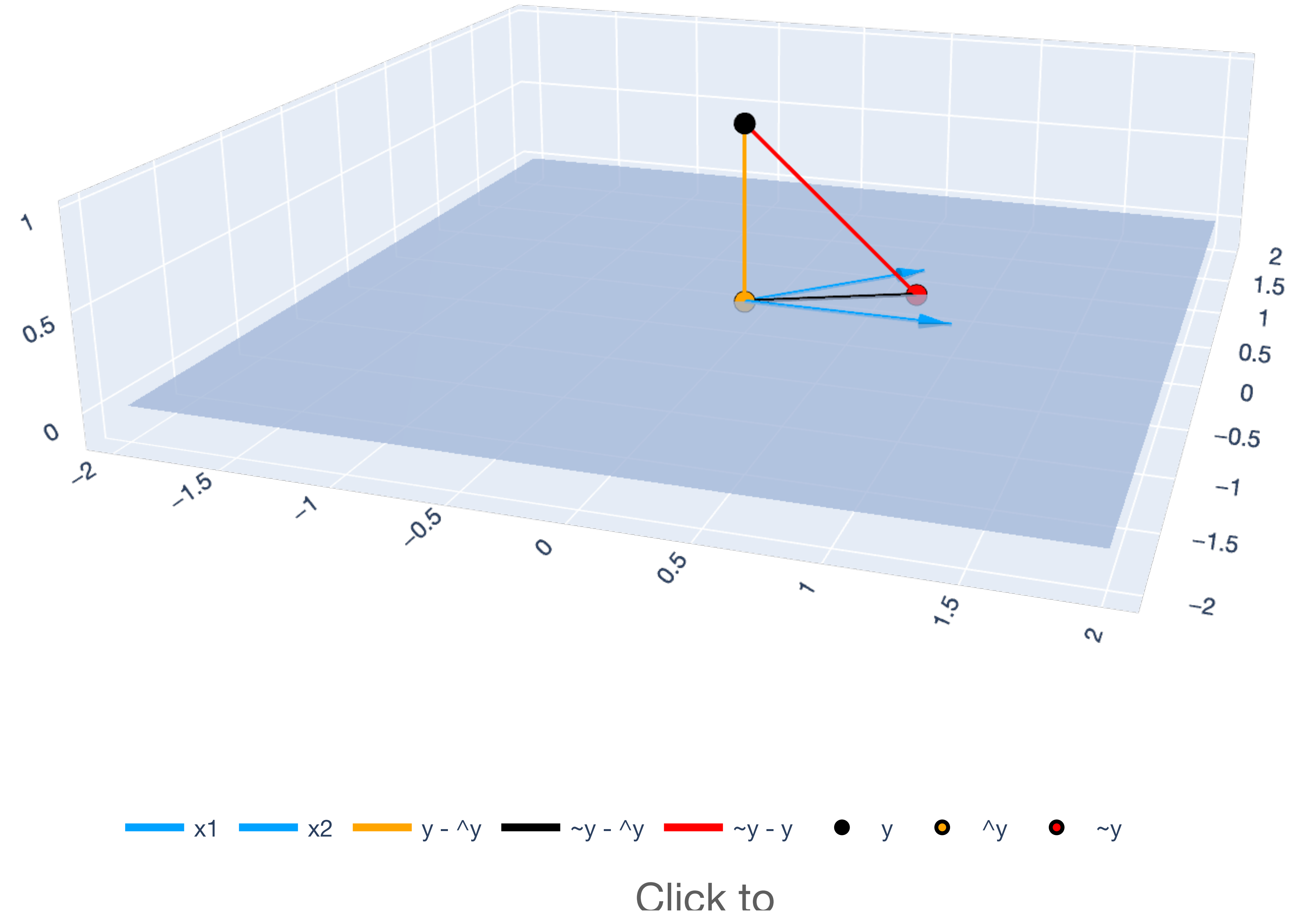
<u>**Theorem (Ordinary Least Squares).**</u>

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

$n \geq d \quad \text{rank}(X) = d$

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

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

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*:

$$\boxed{\mathbf{X}\mathbf{w} \approx \mathbf{y}.} \longrightarrow \min \|Xw - y\|^2 = 0.$$

$X^{-1}y$

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*:

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

Least Squares: SVD Perspective

Main Theorem (with pseudoinverse)

$Xw = y$

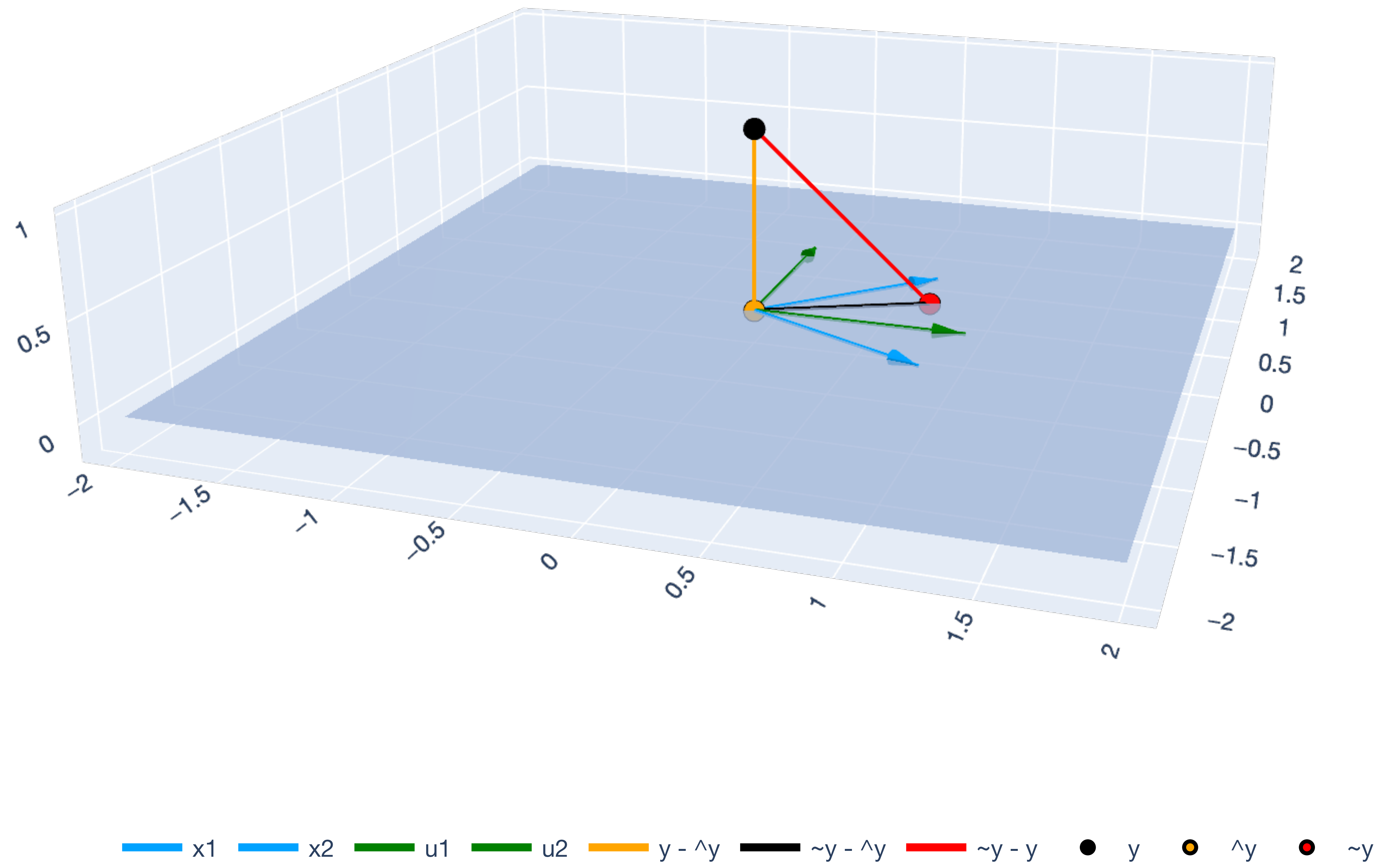
$\boxed{\hat{w} = X^{-1} y}$

$Xw \approx y$

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

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 $\mathrm{rank}(\mathbf{X}) = d$, then:

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

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

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

$d > n$?

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:

$$\boxed{\mathbf{X}\mathbf{w} = \mathbf{y}.} \rightarrow \begin{bmatrix} \text{—} \; x_1 \; \text{—} \\ \vdots \\ \text{—} \; x_n \; \text{—} \end{bmatrix} \begin{bmatrix} w_1 \\ \vdots \\ w_d \end{bmatrix}$$

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 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:

$$n=3 \left\{ \overbrace{\begin{bmatrix} 2 & 1 \\ 2 & -1 \\ 4 & -2 \end{bmatrix}}^{d=2} \right. \begin{bmatrix} w_1 \\ w_2 \end{bmatrix} = \begin{bmatrix} 3 \\ 3 \\ 3 \end{bmatrix}$$

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

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:

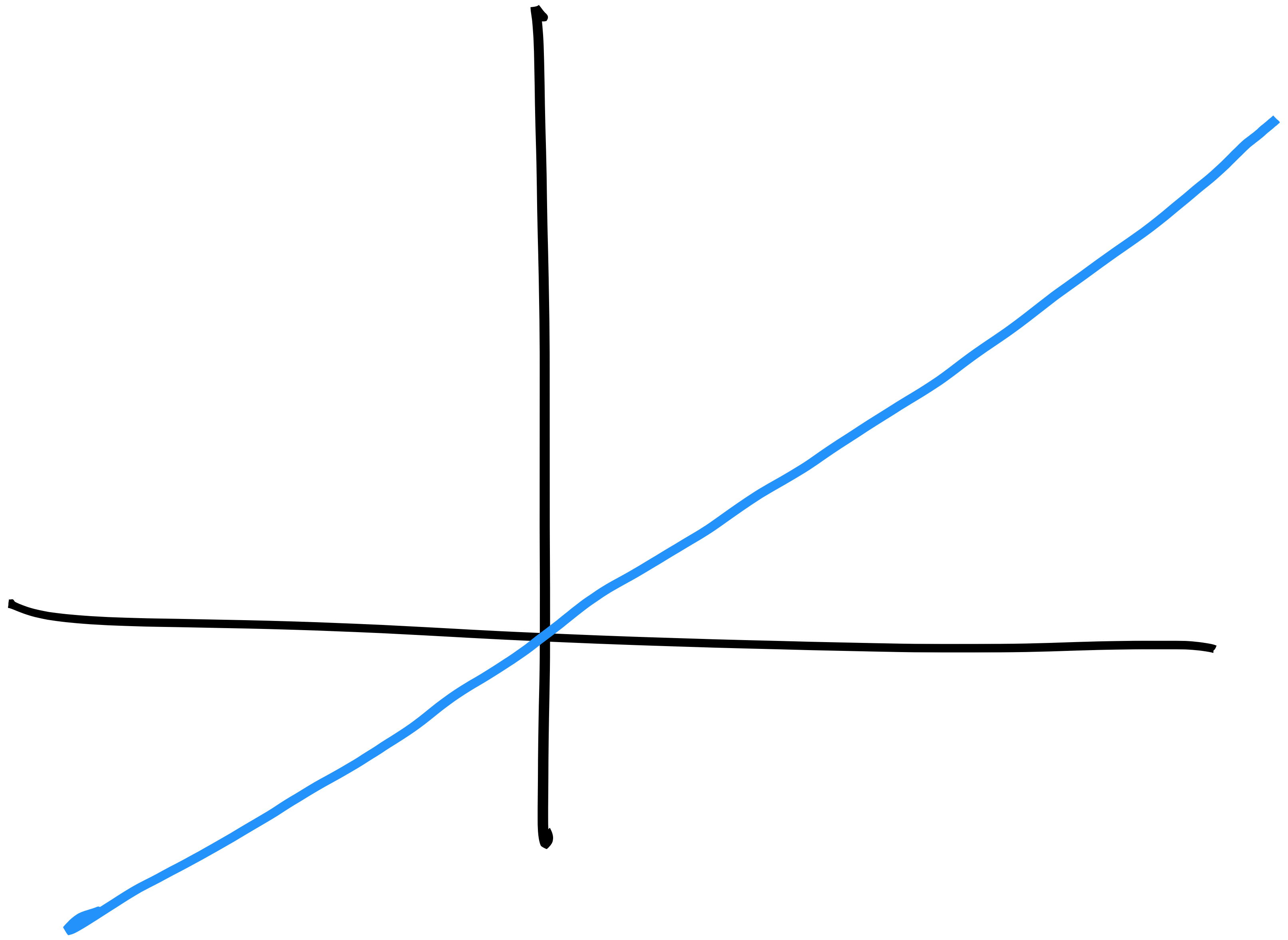
$$\begin{bmatrix} 2 & 1 & 1 \\ 2 & -1 & 0 \end{bmatrix} \begin{bmatrix} w_1 \\ w_2 \\ w_3 \end{bmatrix} = \begin{bmatrix} 3 \\ 3 \end{bmatrix}$$

Least Squares with $d \geq n$

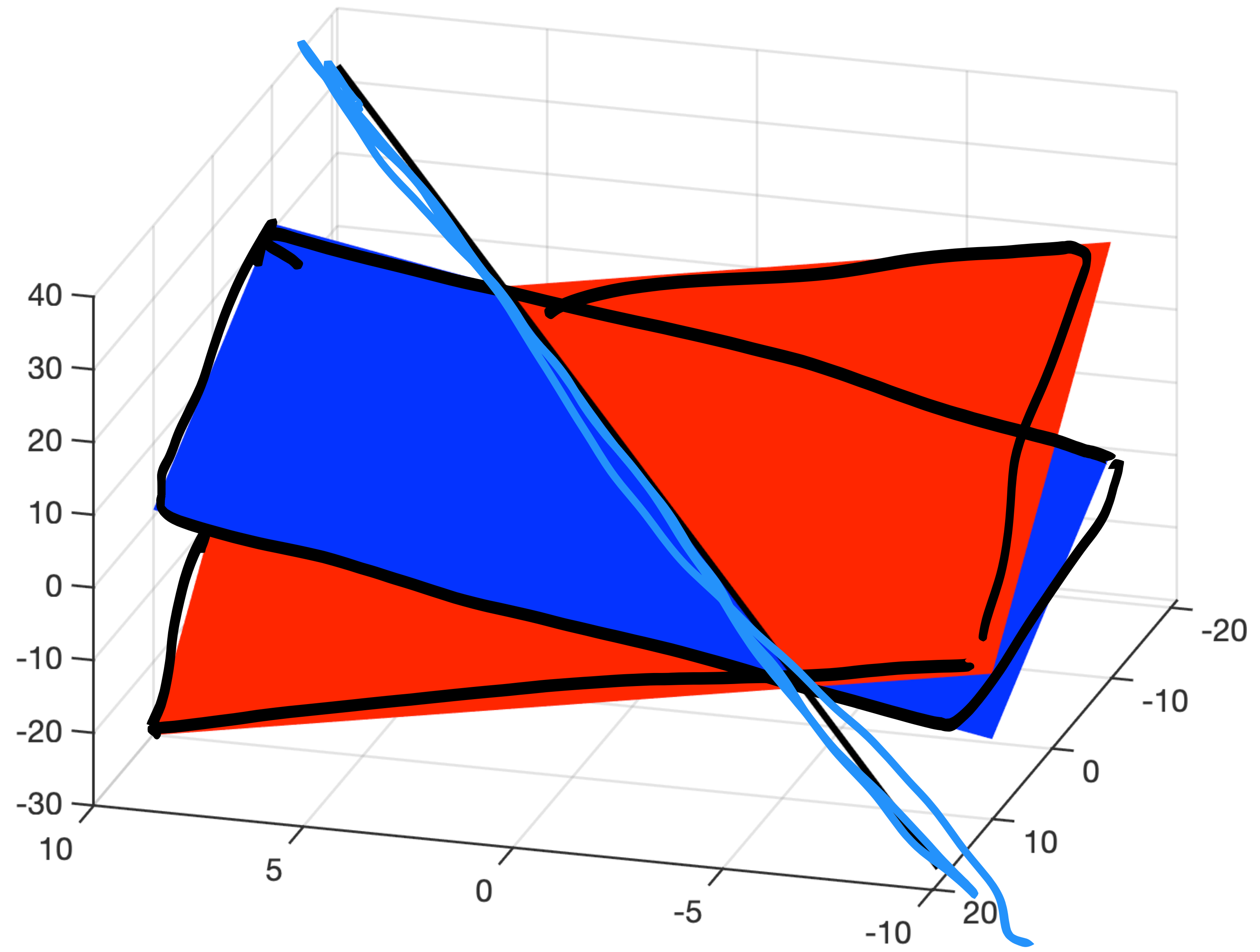
Review: Systems of Linear Equations

When the number of equations < number of unknowns…

$n = 2, \mathbb{R}^2$



$n = 3, \mathbb{R}^3$

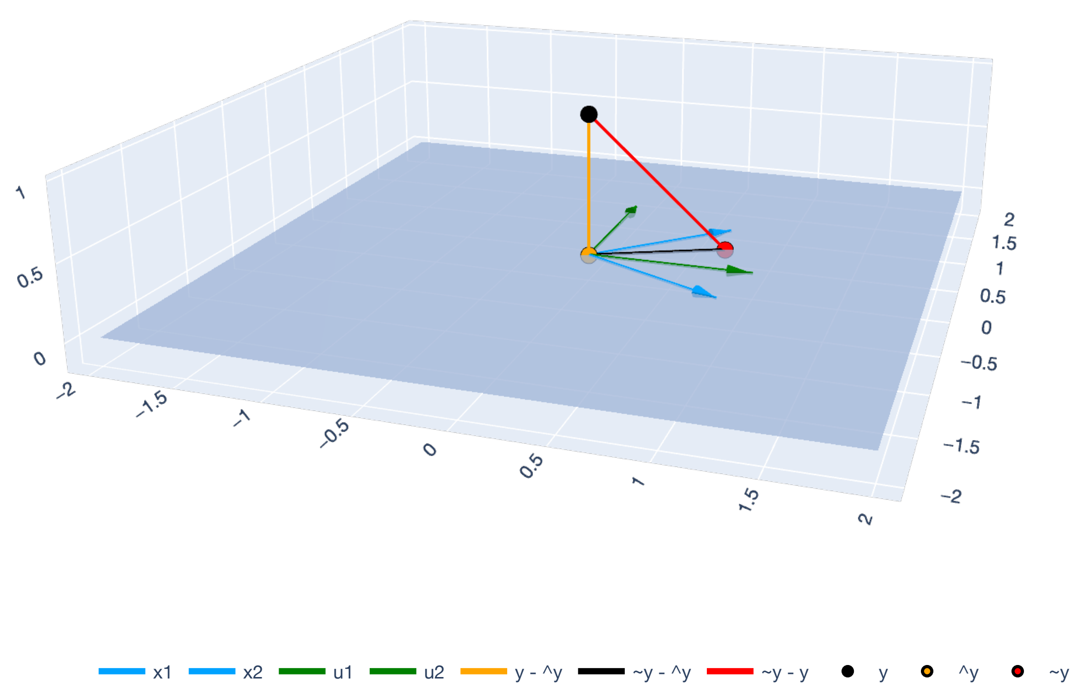


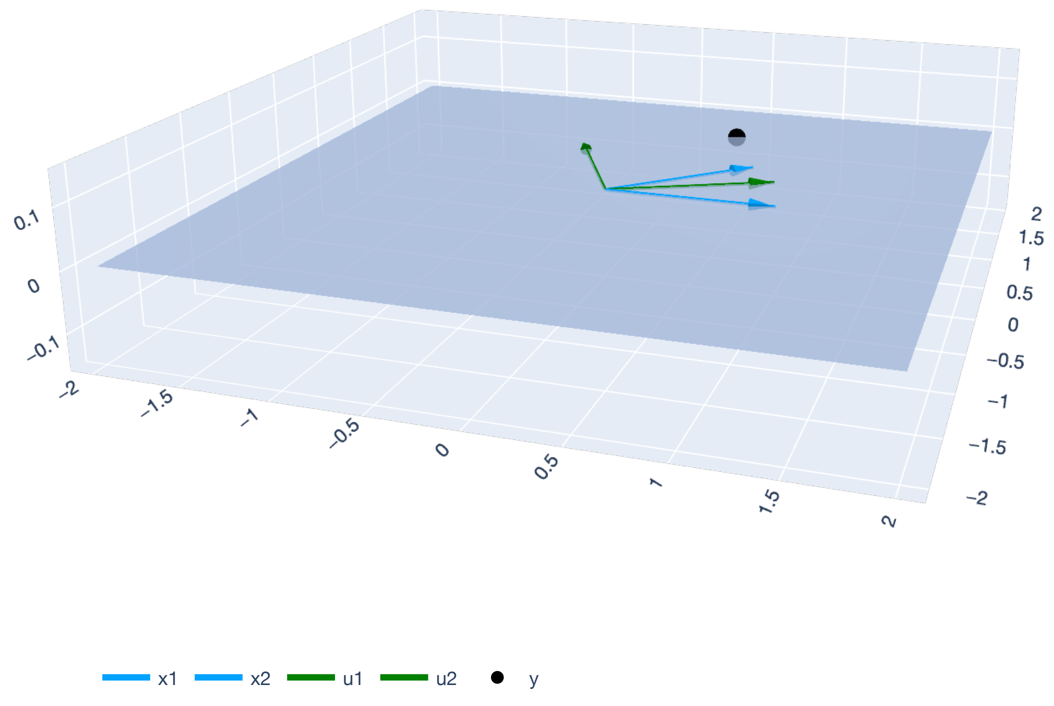
Least Squares with $d \geq n$
Problem Statement

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

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

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



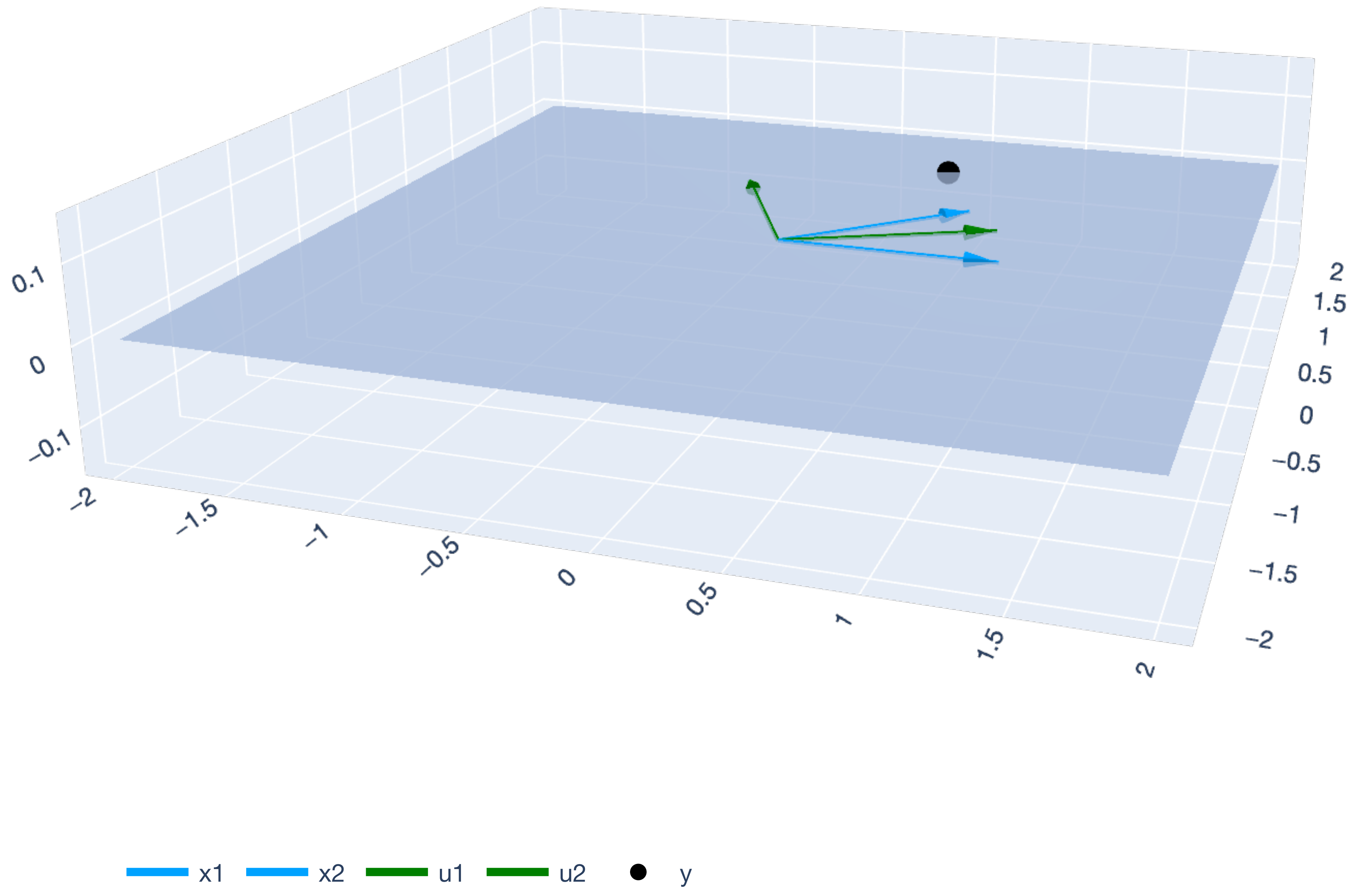


Least Squares with $d \geq n$

Using the Pseudoinverse

$\boxed{\hat{w} = X^{+} y}$

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 $\mathrm{rank}(\mathbf{A}) = \min\{n, d\}$, the pseudo inverse

$$\mathbf{A}^+ = \mathbf{V}\boldsymbol{\Sigma}^+\mathbf{U}^\top = \mathbf{V}(\boldsymbol{\Sigma}^\top\boldsymbol{\Sigma})^{-1}\boldsymbol{\Sigma}^\top\mathbf{U}^\top$$

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 $d \geq n$

Using the Pseudoinverse

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

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

Least Squares with $d \geq n$

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 $d \geq n$
Theorem: Minimum norm solution

Theorem (Minimum norm least squares solution). Let $\mathbf{X} \in \mathbb{R}^{n \times d}$, let $d \geq n$, 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:

$$\boxed{\|\mathbf{w}\|_2^2 \geq \|\hat{\mathbf{w}}\|_2^2 \text{ for all } \mathbf{w} \in \mathbb{R}^d.}$$

Least Squares with $d \geq n$

Theorem: Minimum norm solution

Theorem (Minimum norm least squares solution). Let $\mathbf{X} \in \mathbb{R}^{n \times d}$, let $d \geq n$, 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:

$$\|\mathbf{w}\|_2^2 \geq \|\hat{\mathbf{w}}\|_2^2 \text{ for all } \mathbf{w} \in \mathbb{R}^d.$$

Proof. Consider any arbitrary $\mathbf{w} \in \mathbb{R}^d$.

Least Squares with $d \geq n$

Theorem: Minimum norm solution

Theorem (Minimum norm least squares solution). Let $\mathbf{X} \in \mathbb{R}^{n \times d}$, let $d \geq n$, 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:

$$\|\mathbf{w}\|_2^2 \geq \|\hat{\mathbf{w}}\|_2^2 \text{ for all } \mathbf{w} \in \mathbb{R}^d.$$

Proof. Consider any arbitrary $\mathbf{w} \in \mathbb{R}^d$. We can write $\mathbf{w}$'s Euclidean norm as:

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

Least Squares with $d \geq n$

Theorem: Minimum norm solution

<u>**Theorem (Minimum norm least squares solution).**</u> Let $\mathbf{X} \in \mathbb{R}^{n \times d}$, let $d \geq n$, and let $\mathrm{rank}(\mathbf{X}) = n$. 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:

$$\|\mathbf{w}\|_2^2 \geq \|\hat{\mathbf{w}}\|_2^2 \text{ for all } \mathbf{w} \in \mathbb{R}^d.$$

Proof. Consider any arbitrary $\mathbf{w} \in \mathbb{R}^d$. We can write $\mathbf{w}$'s Euclidean norm as:

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

Consider the term $(\mathbf{w} - \hat{\mathbf{w}})^{\top}\hat{\mathbf{w}}$:

$$(\mathbf{w} - \hat{\mathbf{w}})^{\top}\hat{\mathbf{w}} = (\mathbf{w} - \hat{\mathbf{w}})^{\top}\underbrace{\mathbf{X}^{\top}(\mathbf{X}\mathbf{X}^{\top})^{-1}}_{\mathbf{X}^{+} \text{ if } d>n}\mathbf{y}$$

Least Squares with $d \geq n$

Theorem: Minimum norm solution

Theorem (Minimum norm least squares solution). Let $\mathbf{X} \in \mathbb{R}^{n \times d}$, let $d \geq n$, and let $\text{rank}(\mathbf{X}) = n$. 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:

$$\|\mathbf{w}\|_2^2 \geq \|\hat{\mathbf{w}}\|_2^2 \text{ for all } \mathbf{w} \in \mathbb{R}^d.$$

Proof. Consider any arbitrary $\mathbf{w} \in \mathbb{R}^d$. We can write $\mathbf{w}$'s Euclidean norm as:

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

Consider the term $(\mathbf{w} - \hat{\mathbf{w}})^\top\hat{\mathbf{w}}$:

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

Least Squares with $d \geq n$

Theorem: Minimum norm solution

Theorem (Minimum norm least squares solution). Let $\mathbf{X} \in \mathbb{R}^{n \times d}$, let $d \geq n$, 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 (i.e., $\mathbf{X}\hat{\mathbf{w}} = \mathbf{y}$) with smallest Euclidean norm:

$$\|\mathbf{w}\|_2^2 \geq \|\hat{\mathbf{w}}\|_2^2 \text{ for all } \mathbf{w} \in \mathbb{R}^d.$$

Proof. Consider any solution $\mathbf{w} \in \mathbb{R}^d$, such that $\mathbf{X}\mathbf{w} = \mathbf{y}$. We can write $\mathbf{w}$'s Euclidean norm as:

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

Consider the term $(\mathbf{w} - \hat{\mathbf{w}})^{\top}\hat{\mathbf{w}}$:

$$(\mathbf{w} - \hat{\mathbf{w}})^{\top}\hat{\mathbf{w}} = (\mathbf{w} - \hat{\mathbf{w}})^{\top}\mathbf{X}^{\top}(\mathbf{X}\mathbf{X}^{\top})^{-1}\mathbf{y} = (\mathbf{X}\mathbf{w} - \mathbf{X}\hat{\mathbf{w}})^{\top}(\mathbf{X}\mathbf{X}^{\top})^{-1}\mathbf{y} = 0,$$

because $\mathbf{w}$ and $\hat{\mathbf{w}}$ are both exact solutions.

Least Squares with $d \geq n$

Theorem: Minimum norm solution

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

$$\|\mathbf{w}\|_2^2 \geq \|\hat{\mathbf{w}}\|_2^2 \text{ for all } \mathbf{w} \in \mathbb{R}^d.$$

Proof. Consider any solution $\mathbf{w} \in \mathbb{R}^d$, such that $\mathbf{X}\mathbf{w} = \mathbf{y}$. We can write $\mathbf{w}$'s Euclidean norm as:

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

Consider the term $(\mathbf{w} - \hat{\mathbf{w}})^{\top}\hat{\mathbf{w}}$:

$$(\mathbf{w} - \hat{\mathbf{w}})^{\top}\hat{\mathbf{w}} = (\mathbf{w} - \hat{\mathbf{w}})^{\top}\mathbf{X}^{\top}(\mathbf{X}\mathbf{X}^{\top})^{-1}\mathbf{y} = (\mathbf{X}\mathbf{w} - \mathbf{X}\hat{\mathbf{w}})^{\top}(\mathbf{X}\mathbf{X}^{\top})^{-1}\mathbf{y} = 0,$$

because $\mathbf{w}$ and $\hat{\mathbf{w}}$ are both 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 $\operatorname{rank}(\mathbf{X}) = d$...

We can solve exactly.

Choose

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

which is an exact solution.

If $n > d$ and $\operatorname{rank}(\mathbf{X}) = d$...

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 $\operatorname{rank}(\mathbf{X}) = n$...

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

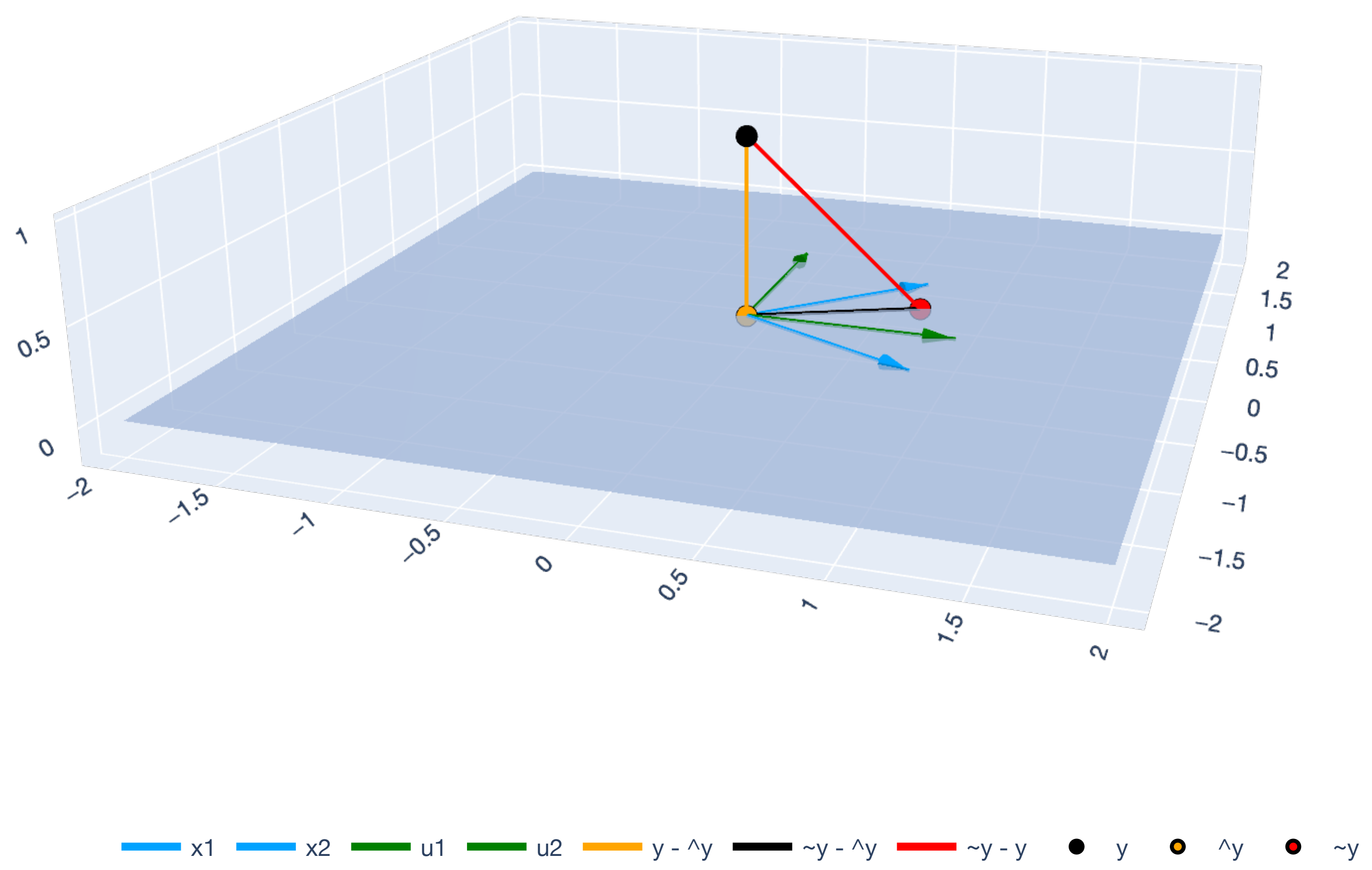
Unified Picture

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

If $n > d$ and $\operatorname{rank}(\mathbf{X}) = d$…

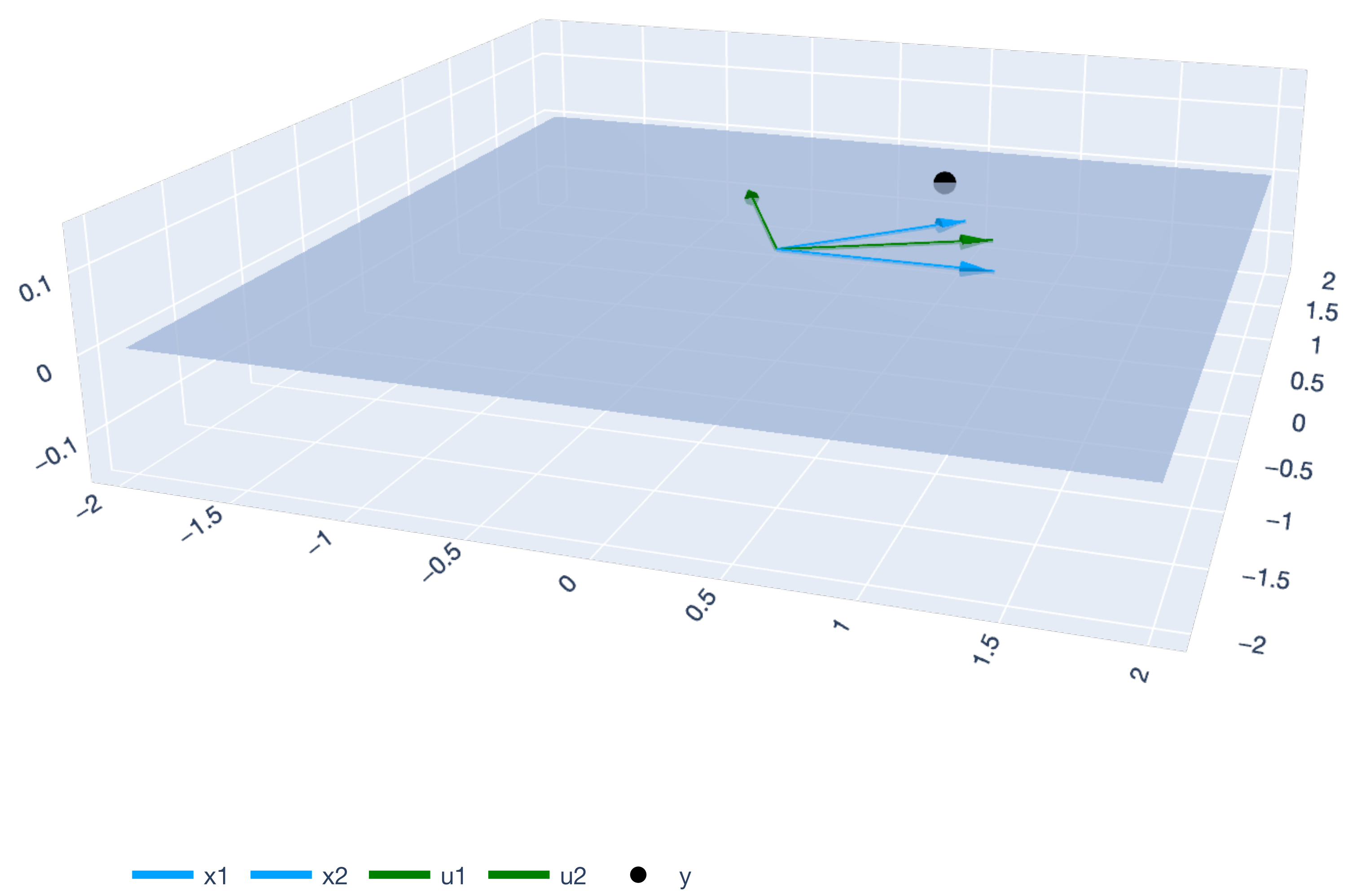
We approximate by least squares:

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



If $n < d$ and $\operatorname{rank}(\mathbf{X}) = n$…

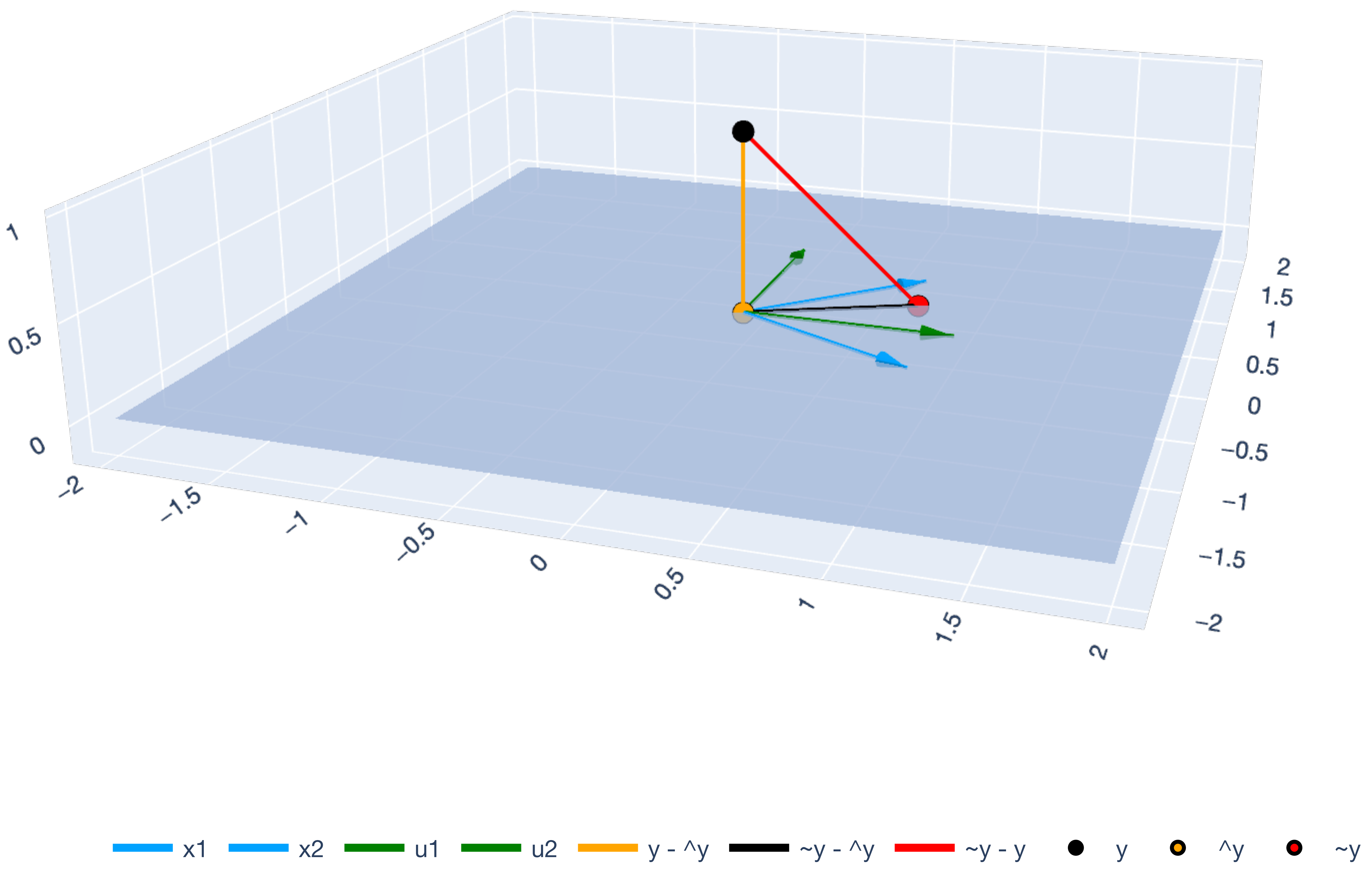
We can solve exactly, but there are infinitely many solutions.

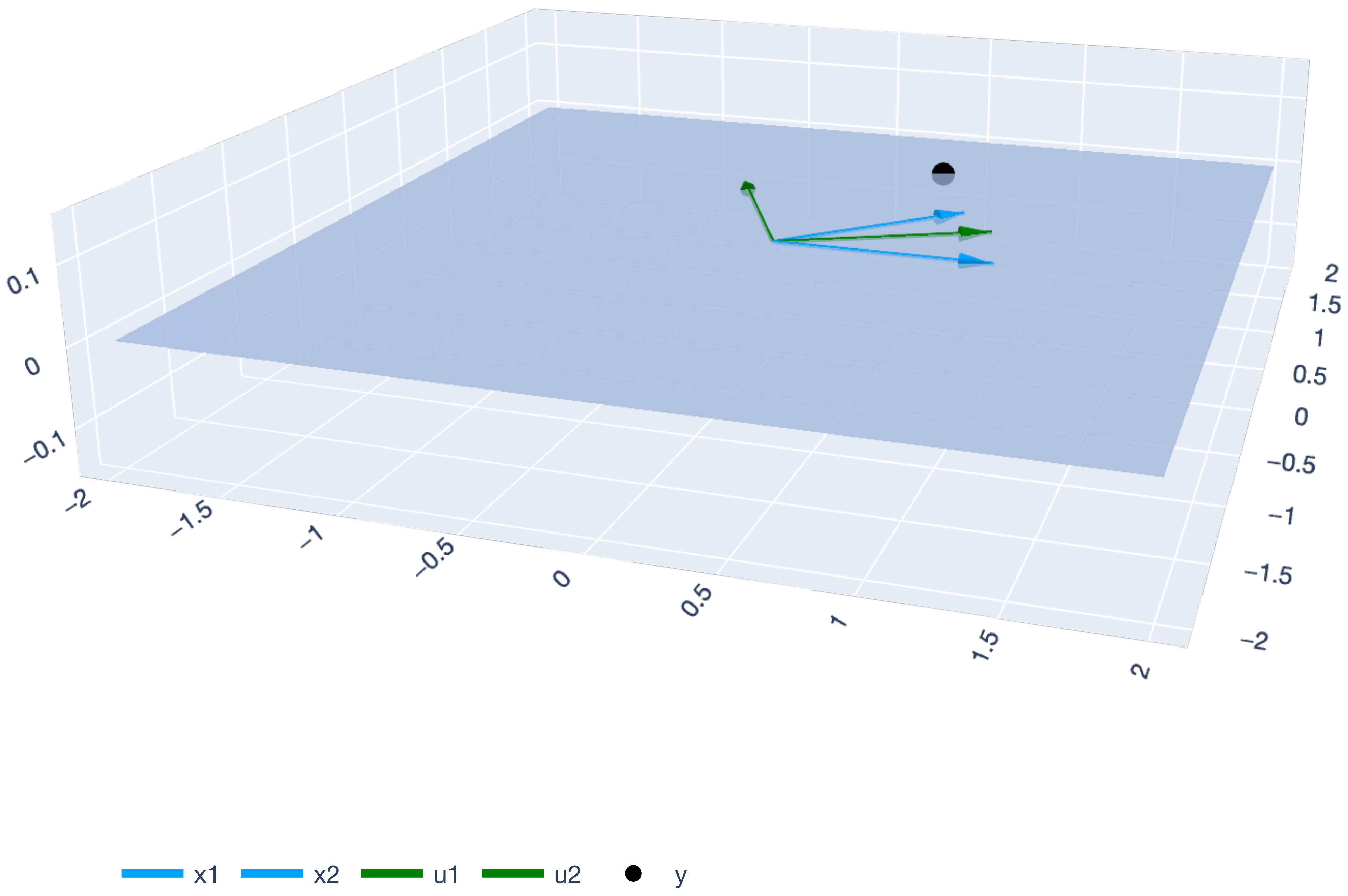


Recap

Lesson Overview

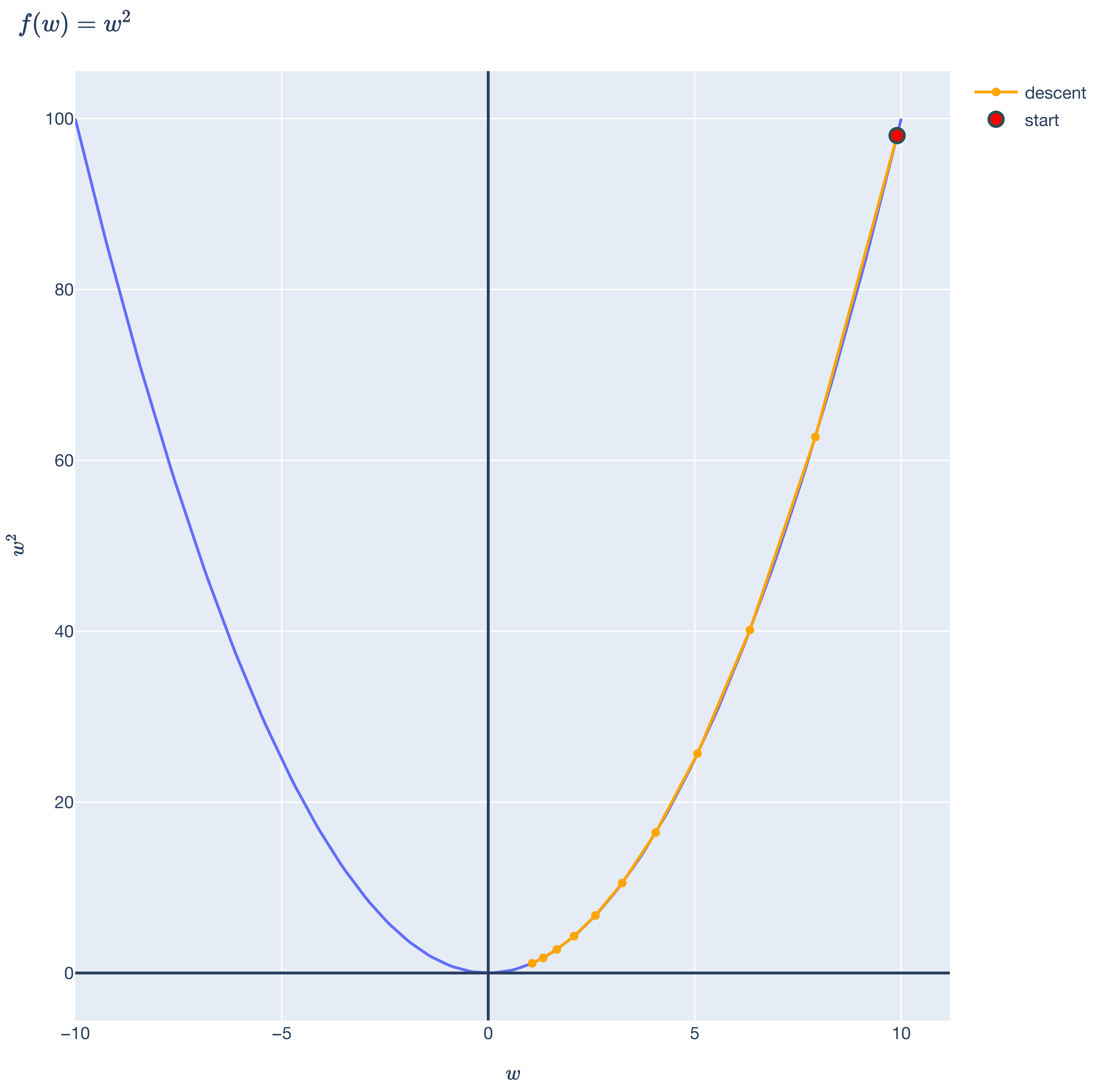
Big Picture: Least Squares

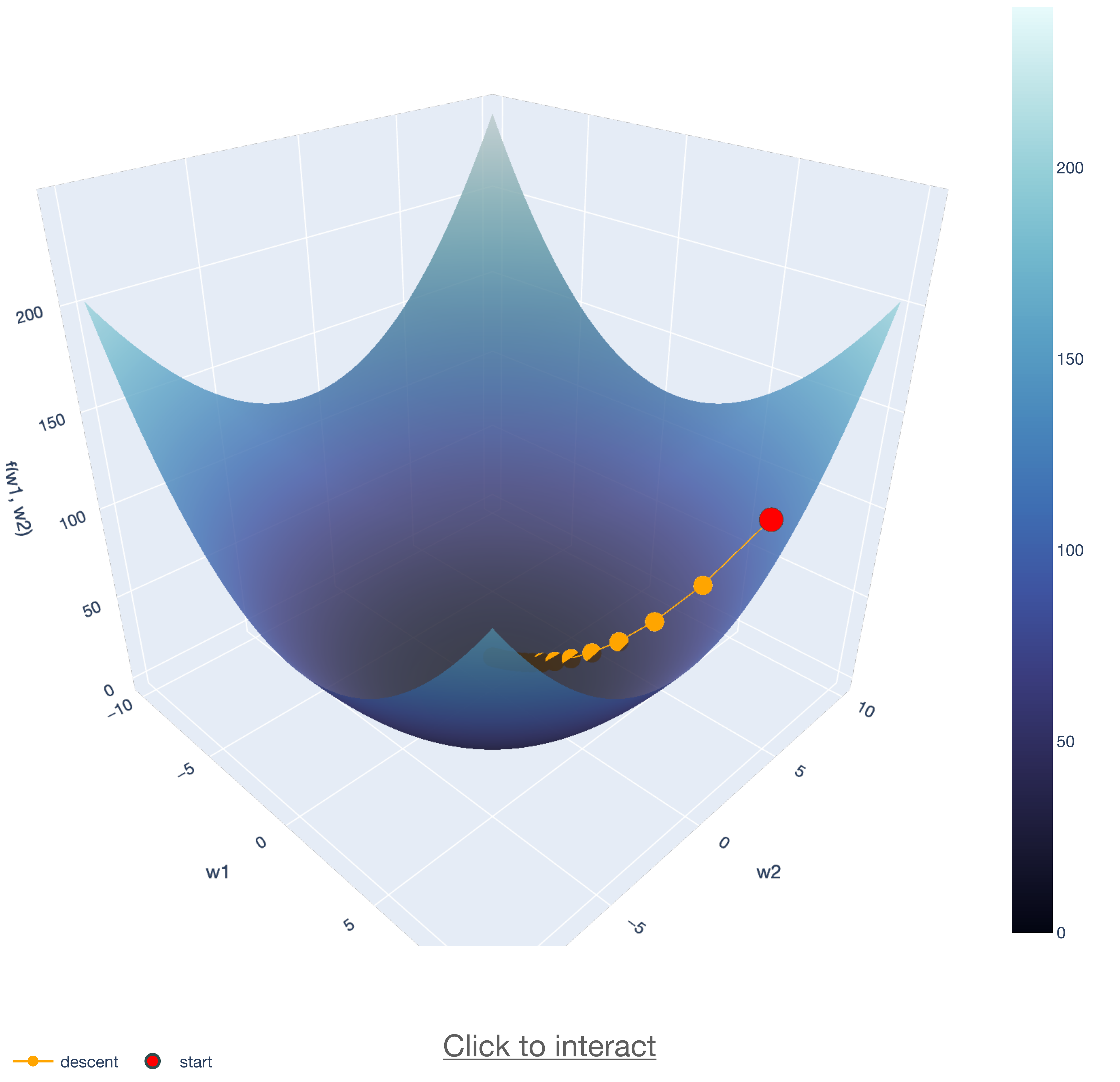




Lesson Overview

Big Picture: Gradient Descent

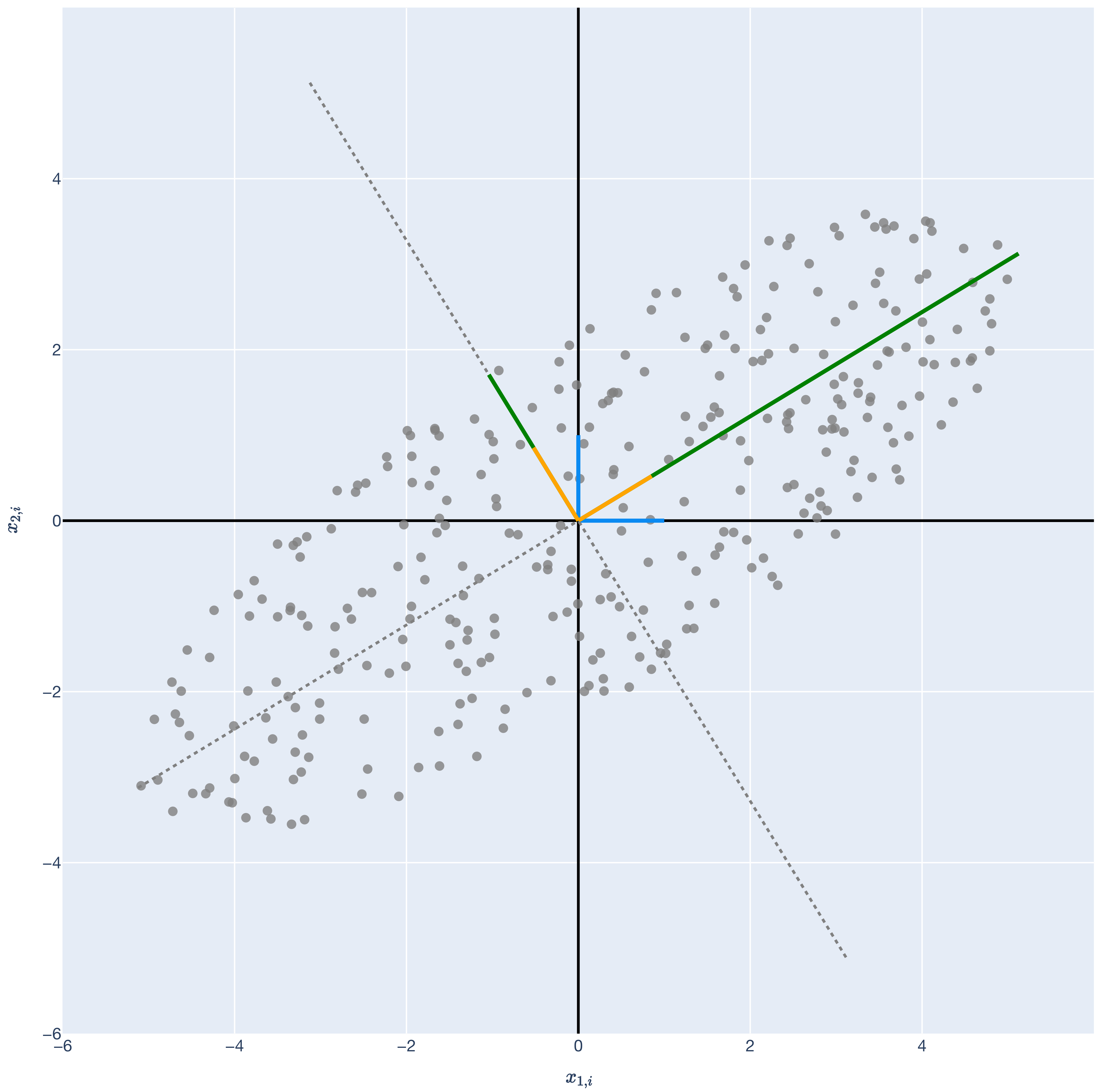


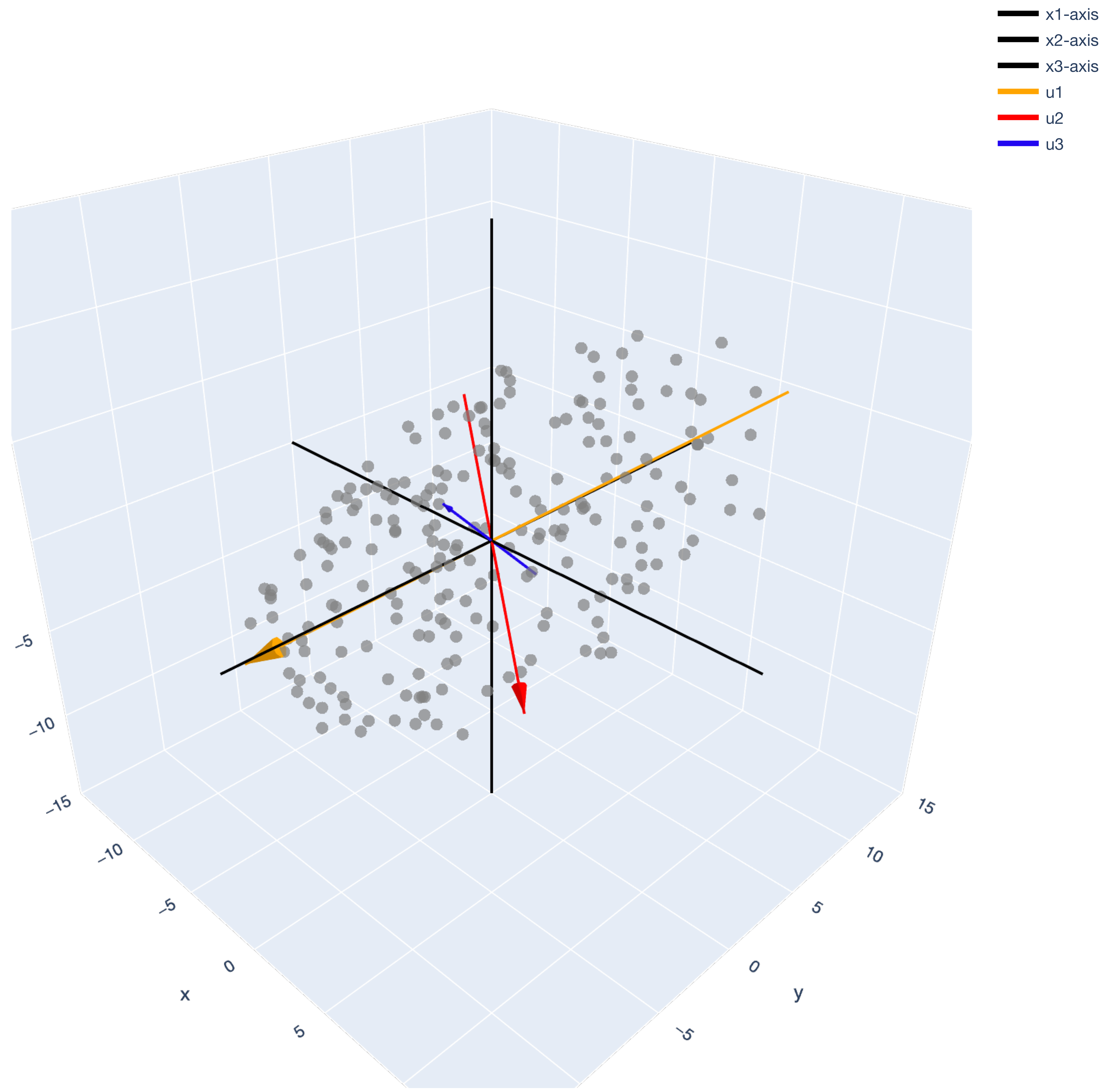


Click to interact

Lesson Overview

Big Picture: Singular Value Decomposition (SVD)





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: Singular Value Decomposition. 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.