

Math for Machine Learning

Week 2.1: Singular Value Decomposition

By: Samuel Deng

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.

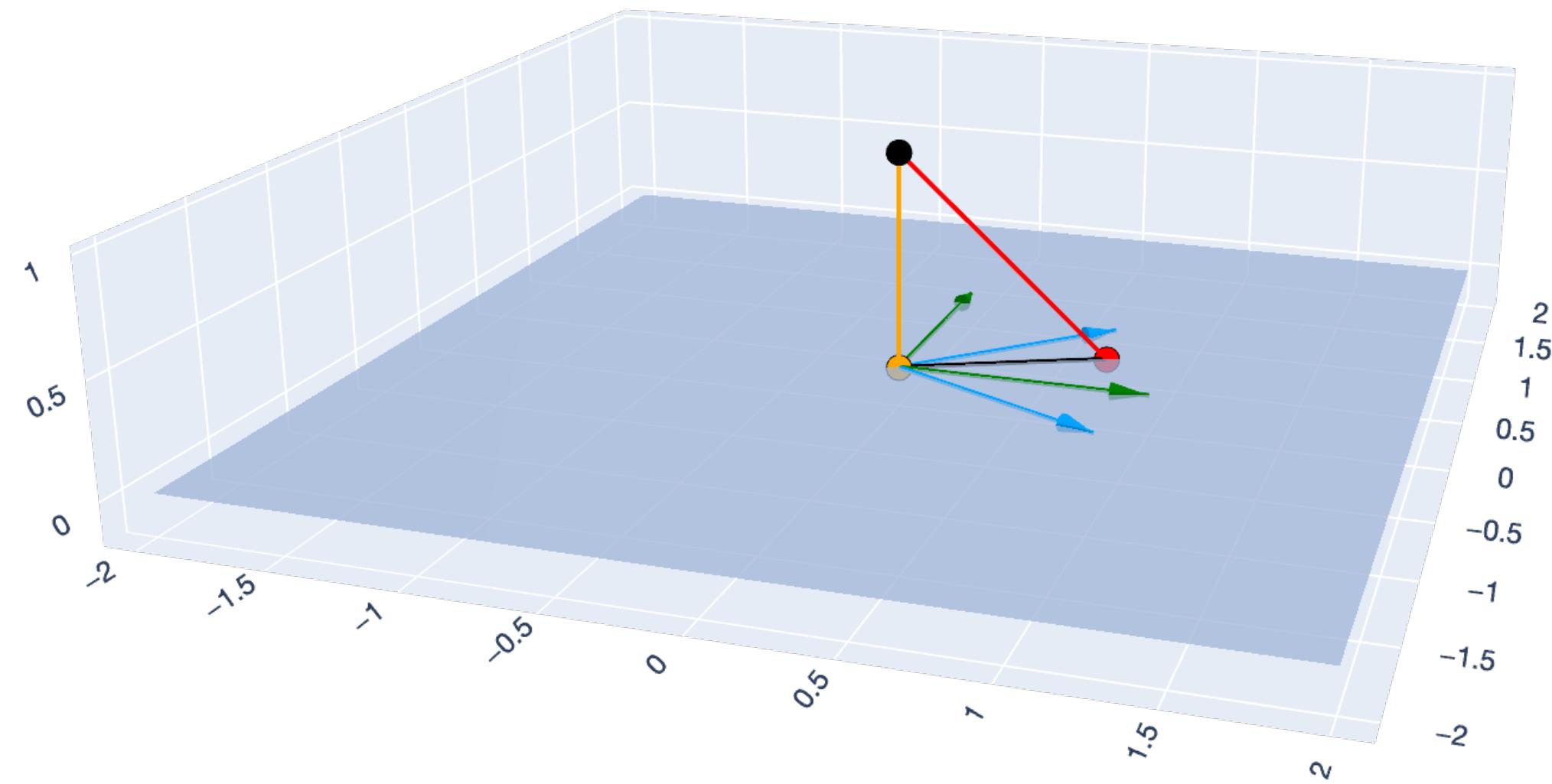
SVD Definition. We go over the definition of SVD and the geometric intuition as the factorization of a data matrix.

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

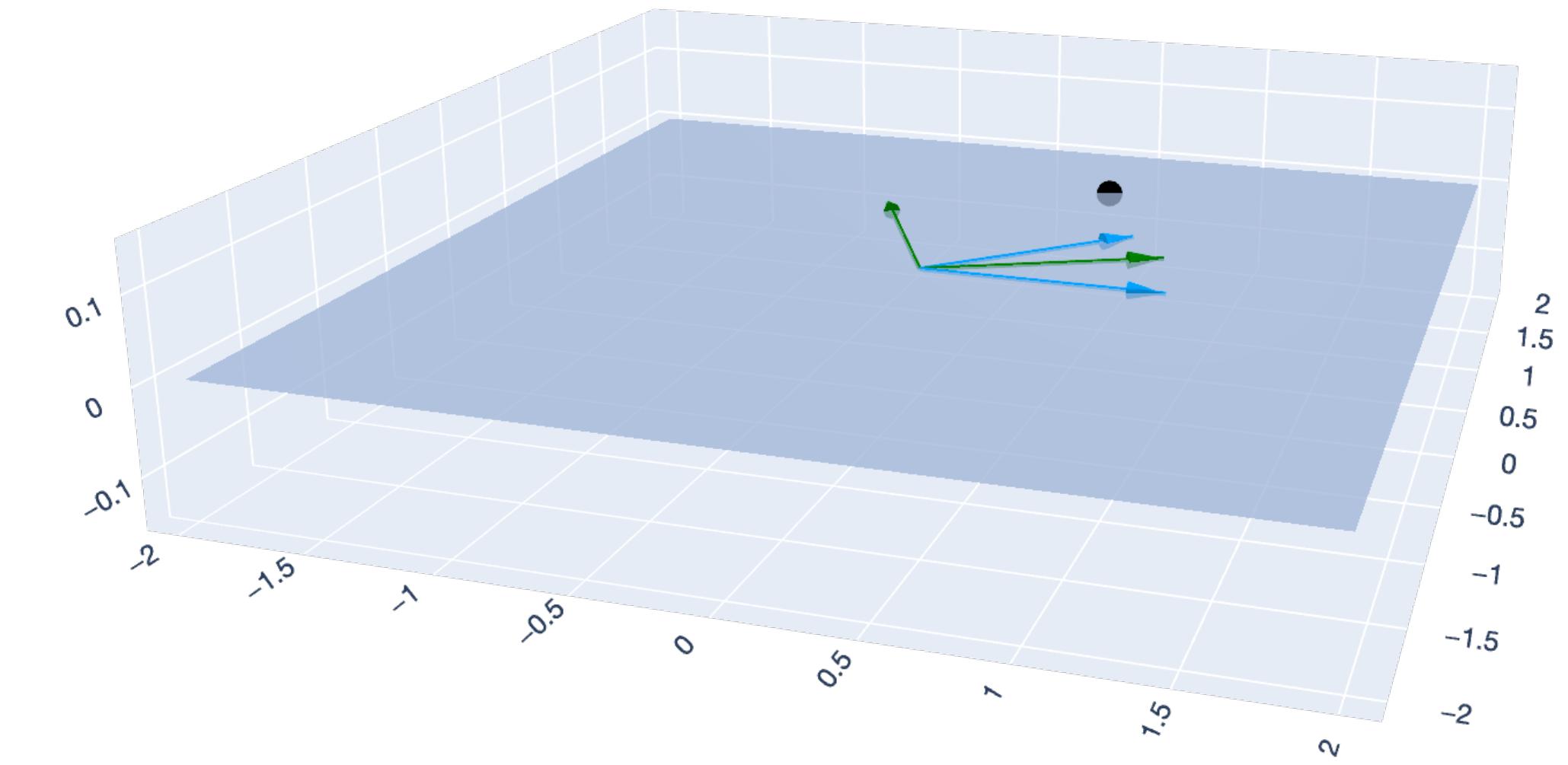
Pseudoinverse. We unify our OLS solution from the perspective of SVD and the notion of the [**pseudoinverse**](#), a generalization of inverses to rectangular matrices.

Lesson Overview

Big Picture: Least Squares



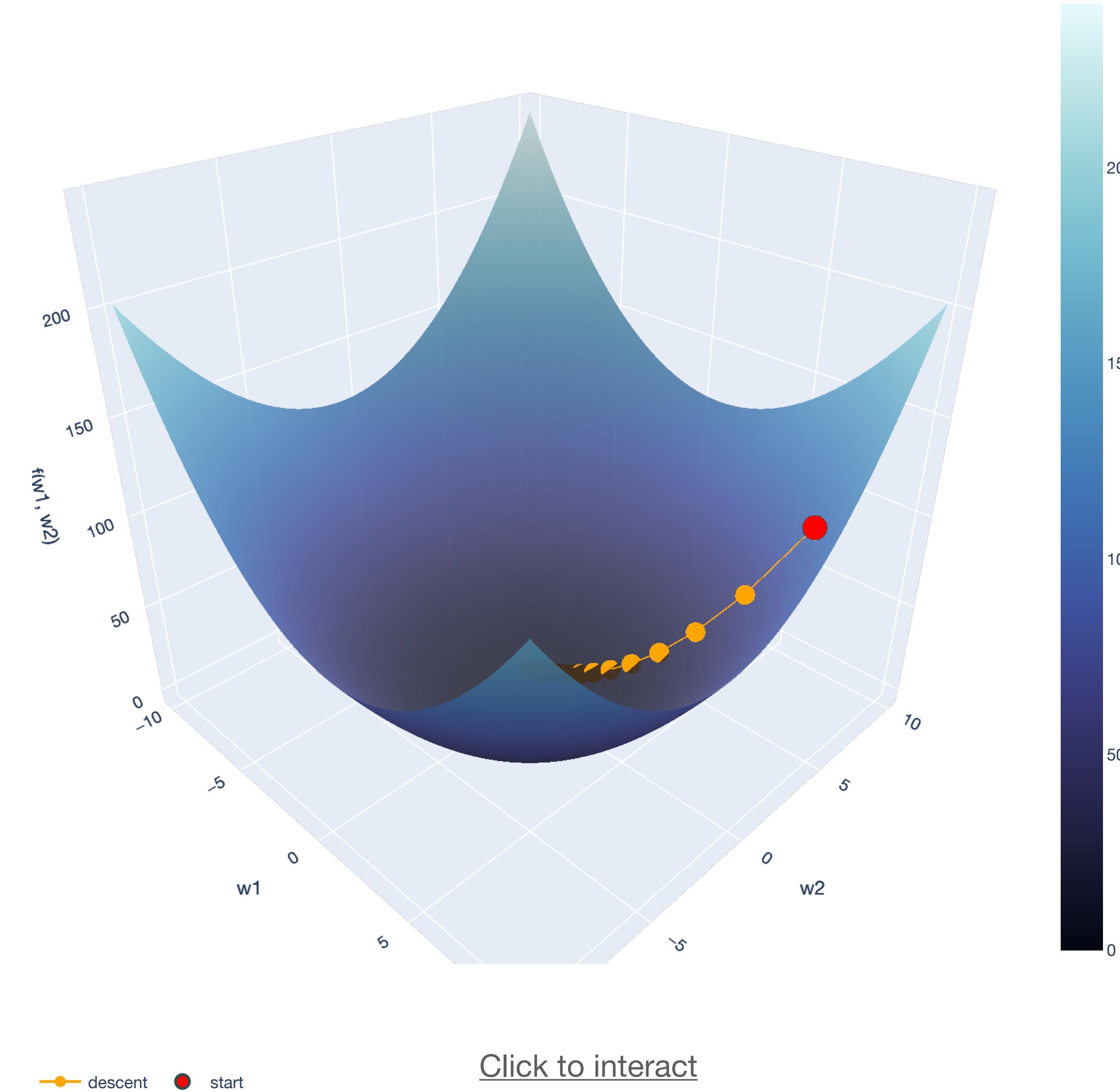
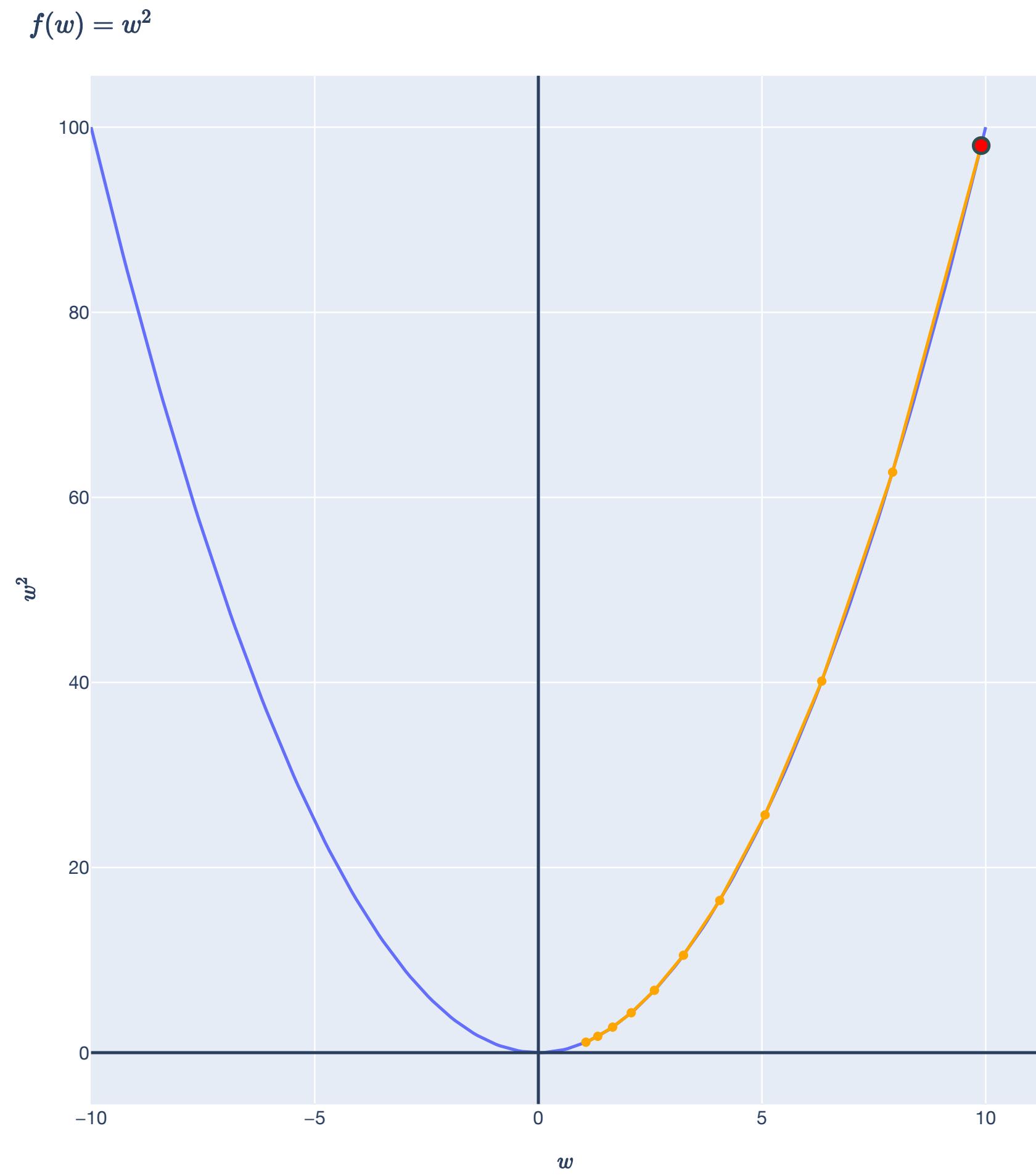
— x_1 — x_2 — u_1 — u_2 — $y - \hat{y}$ — $\hat{y} - y$ ● y ○ \hat{y} ● $\sim y$



— x_1 — x_2 — u_1 — u_2 ● y

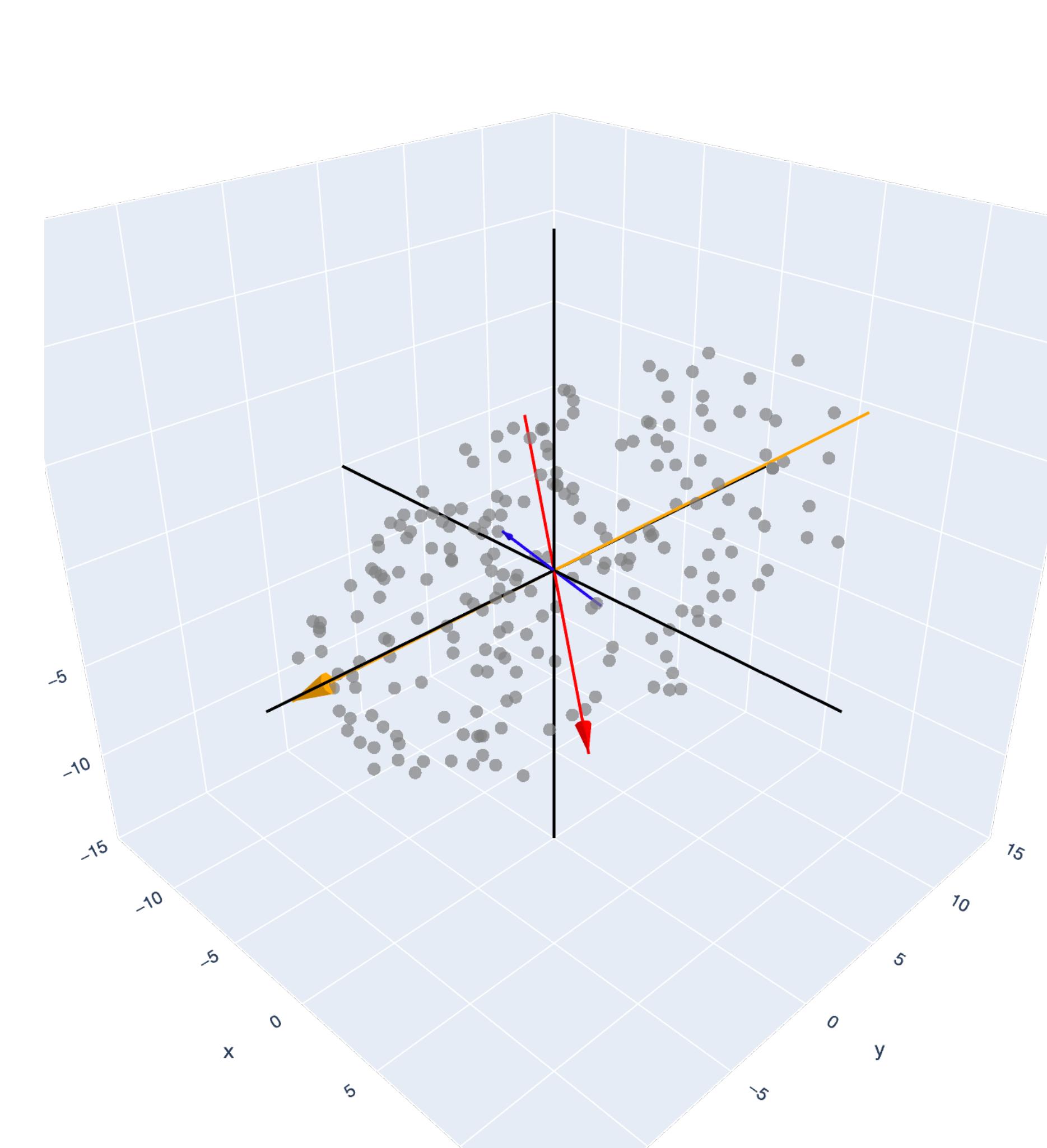
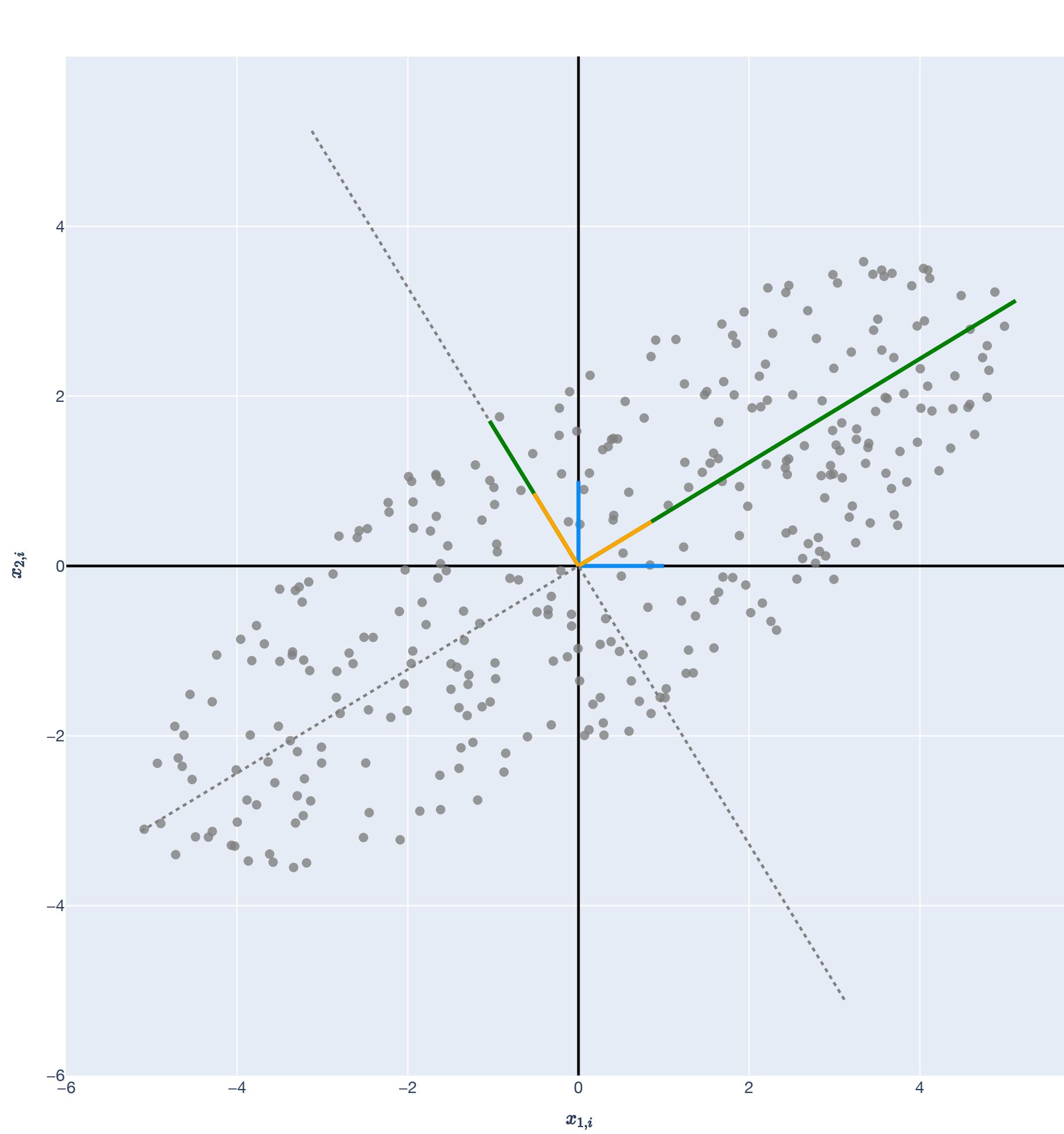
Lesson Overview

Big Picture: Gradient Descent



Lesson Overview

Big Picture: Singular Value Decomposition (SVD)



- x1-axis
- x2-axis
- x3-axis
- u_1
- u_2
- u_3

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* $\mathbf{y} \in \mathbb{R}^d$.

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

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

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

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

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

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*.

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

Regression Setup

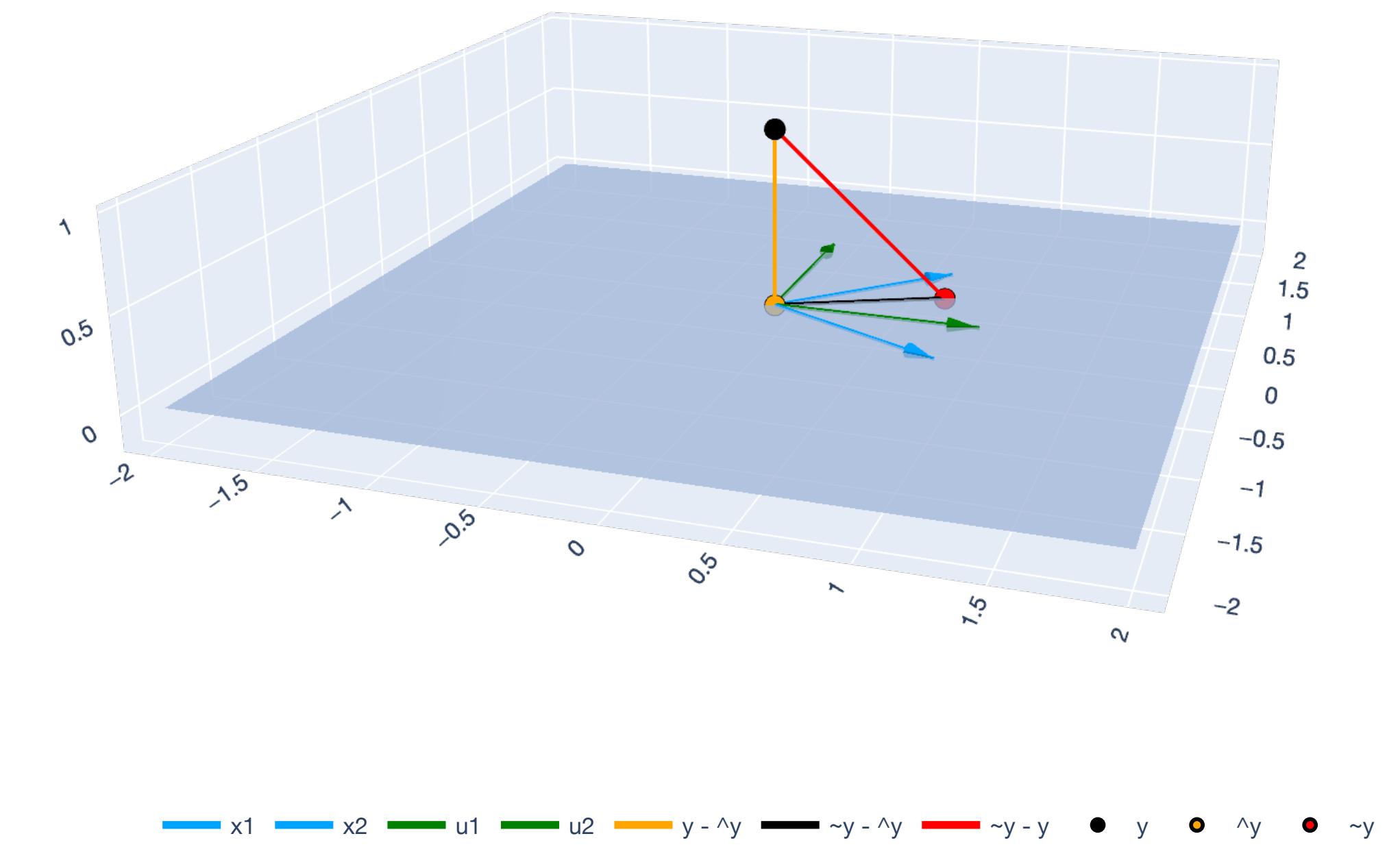
To find $\hat{\mathbf{w}}$, we follow the *principle of least squares*.

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

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

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 $\text{rank}(\mathbf{X}) = d$, 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

OLS with Orthogonal Basis

Theorem (OLS with orthogonal basis). Let $\mathcal{X} \subseteq \mathbb{R}^n$ be a subspace and let $\mathbf{u}_1, \dots, \mathbf{u}_d \in \mathbb{R}^n$ 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:

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

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

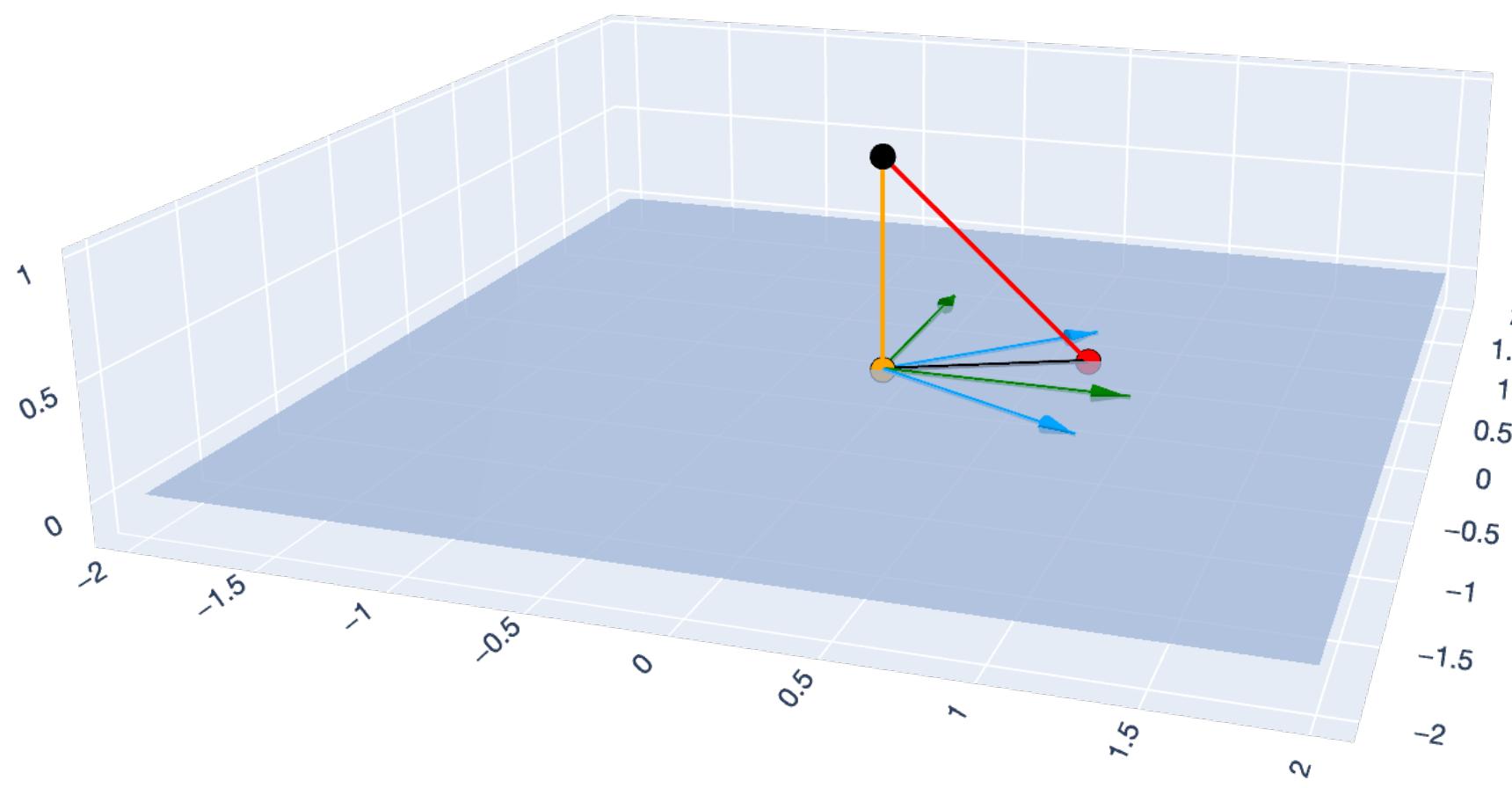
OLS with Orthogonal Basis

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

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

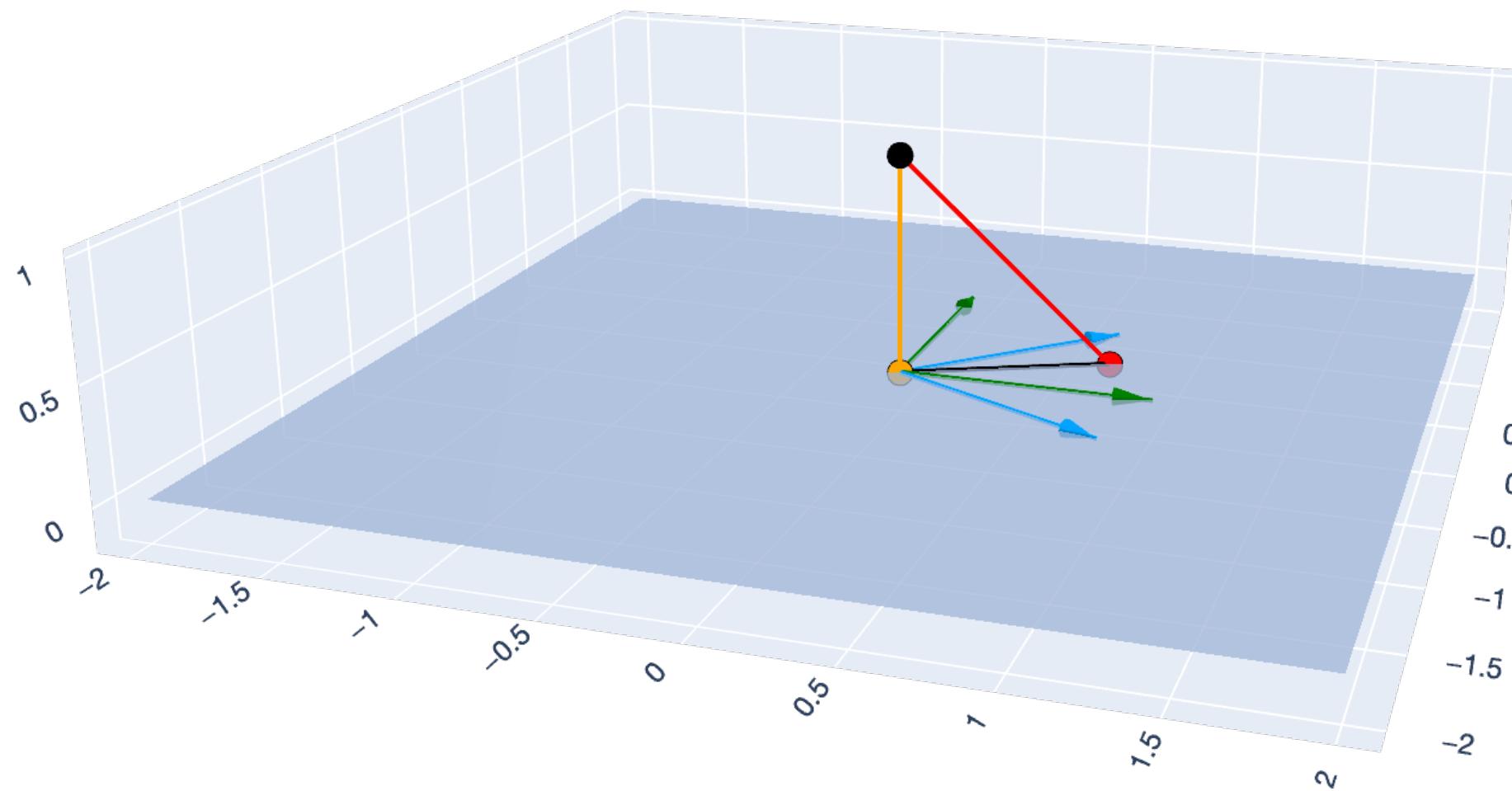
$$\hat{\mathbf{w}}_{onb} = \mathbf{U}^\top \mathbf{y}$$

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



— x_1 — x_2 — \mathbf{u}_1 — \mathbf{u}_2 — $\mathbf{y} - \hat{\mathbf{y}}$ — $\sim \mathbf{y} - \hat{\mathbf{y}}$ ● \mathbf{y} ● $\hat{\mathbf{y}}$ ● $\sim \mathbf{y}$

How to find a good orthogonal basis?



— x_1 — x_2 — u_1 — u_2 — $y - \hat{y}$ — $\sim y - \hat{y}$ — $\sim y - y$ ● y ● \hat{y} ● $\sim y$

Properties of Projections

Projection Matrices and Orthogonal Complement

Projection

Projection of a vector onto a subspace

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

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

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

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

Least Squares as Projection Projection Matrix

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

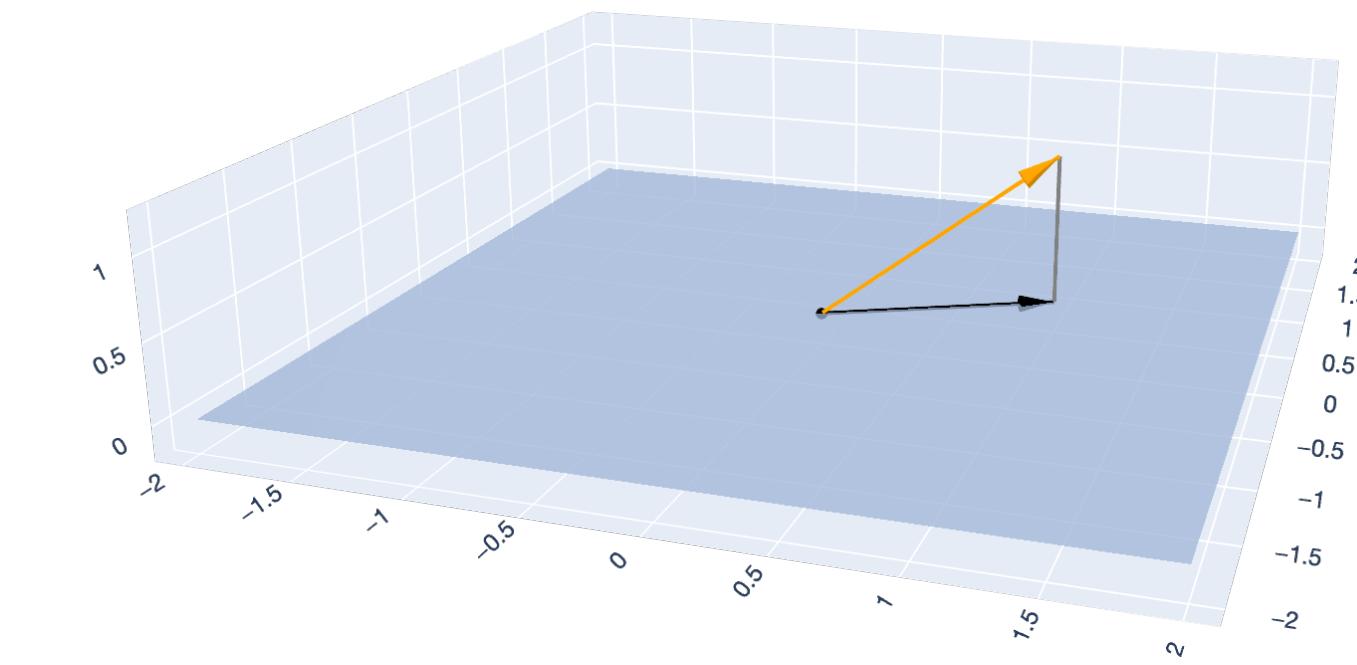
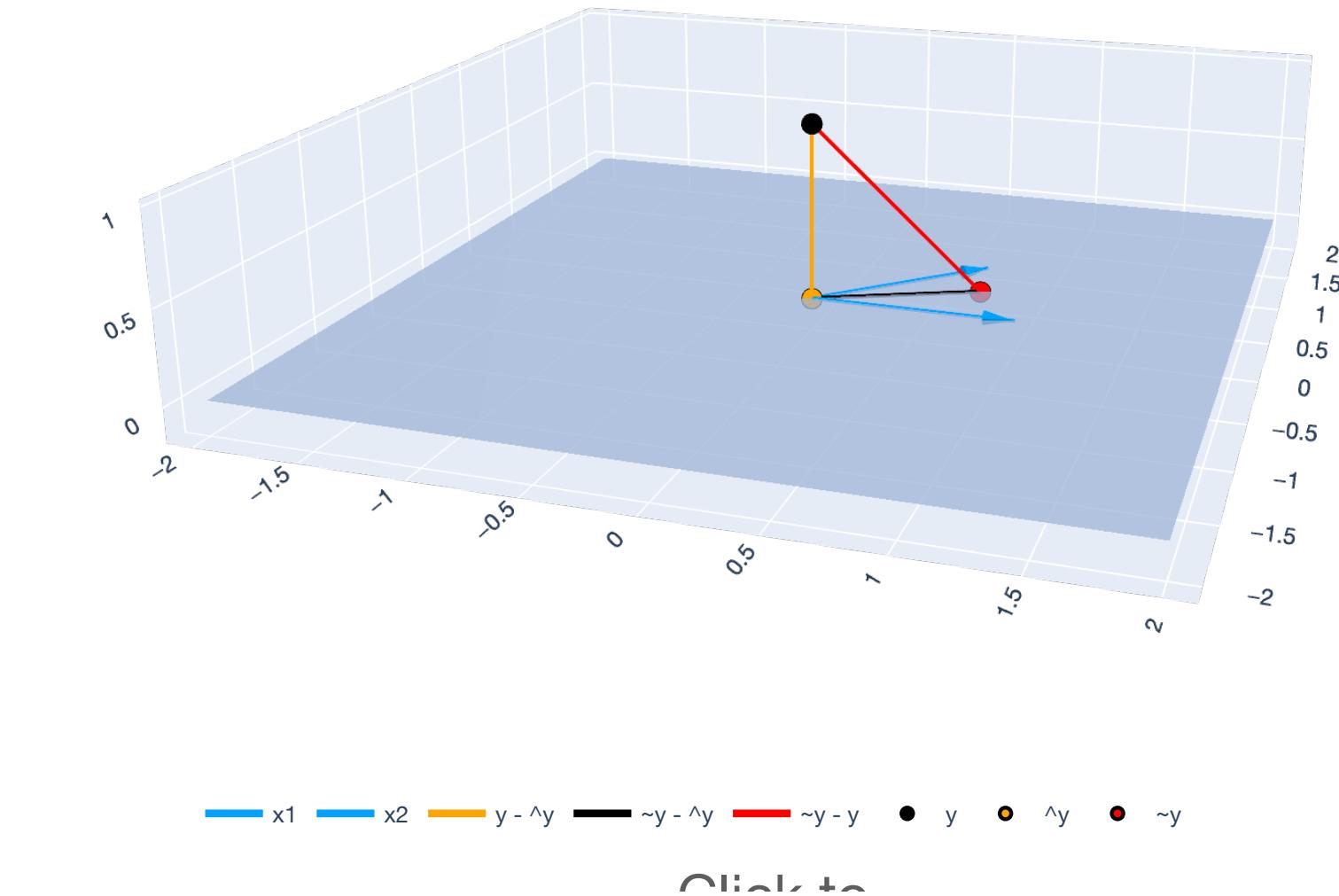
This is just least squares! By what we've learned...

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

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

The ***projection matrix*** is:

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



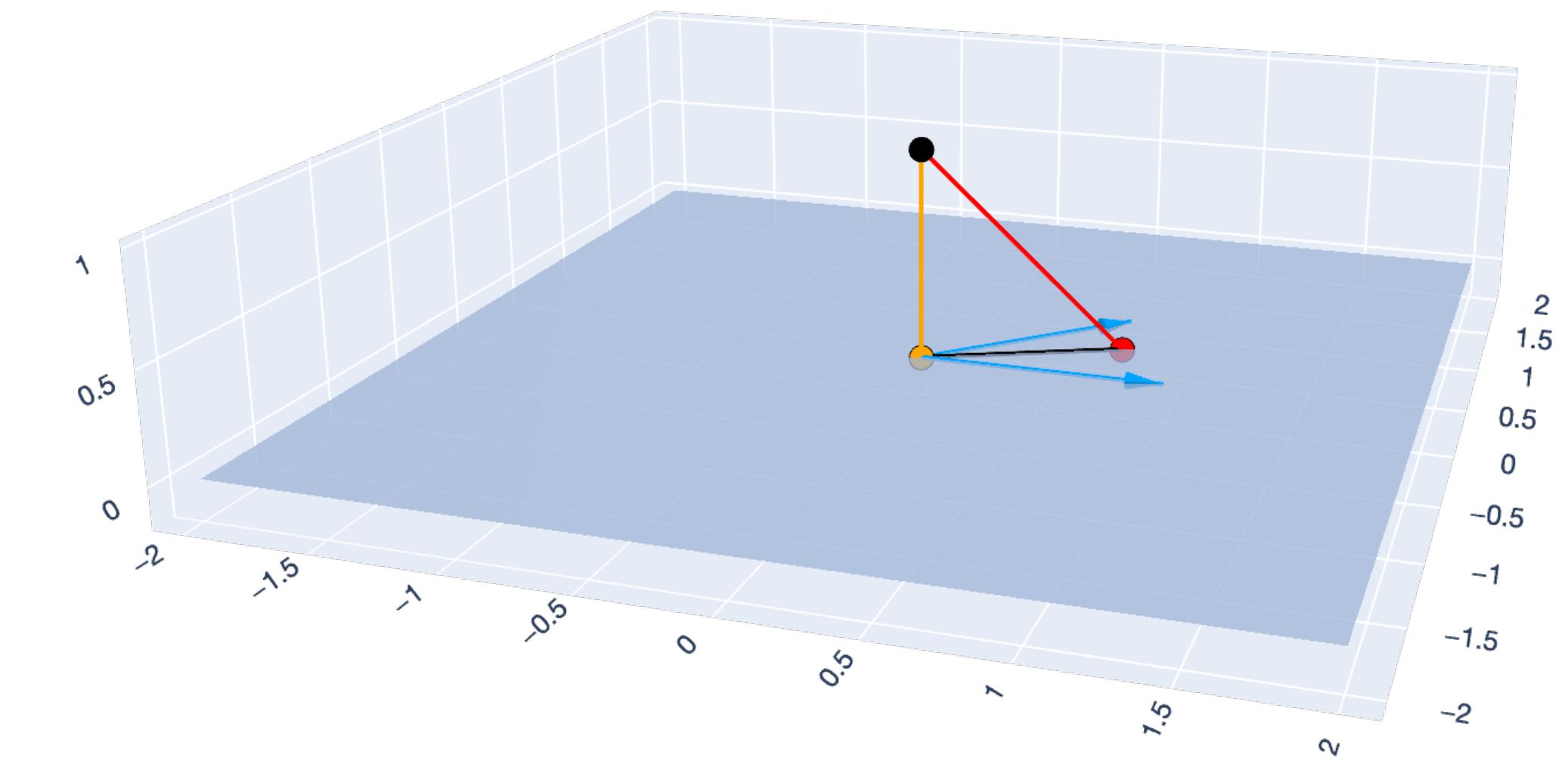
Least Squares as Projection Projection Matrix

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

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

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

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



— x1 — x2 — y - ~y — ~y - ~y — ~y - y • y • ~y • ~y

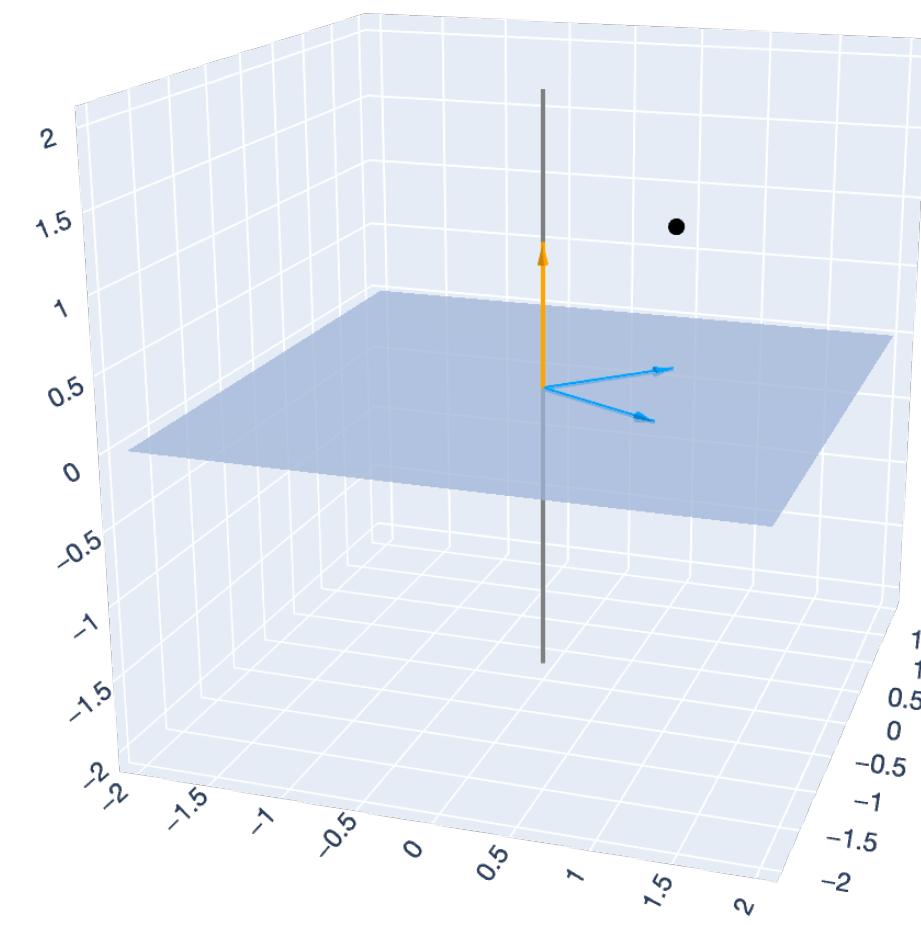
Click to

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



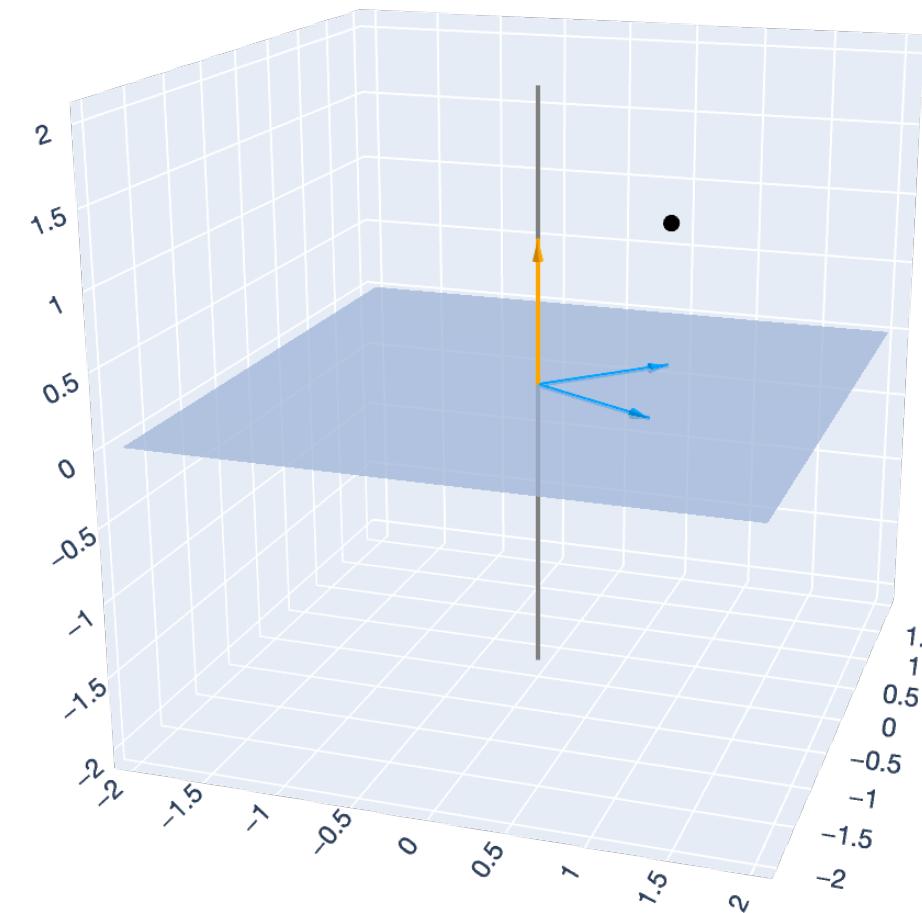
— u1 — u2 — v1 — — • x

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

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

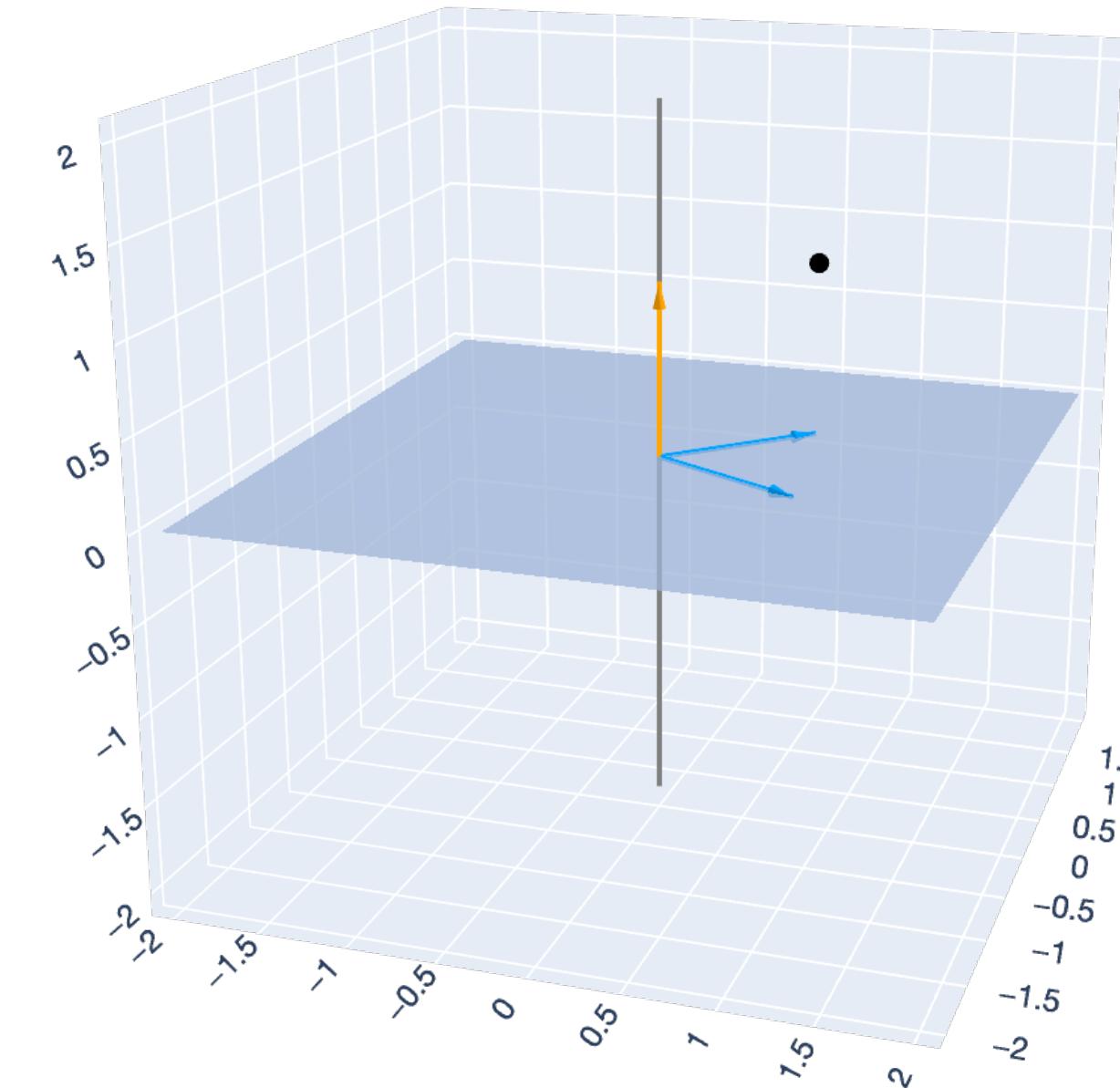


— u1 — u2 — v1 — — x

Orthogonal Complement

Dimension

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



— u1 — u2 — v1 — — x

Orthogonal Complement

Orthogonal Complement and Matrices

Let $\mathbf{a}_1, \dots, \mathbf{a}_d \in \mathbb{R}^n$ be a basis for the subspace $A \subseteq \mathbb{R}^n$. Let $\mathbf{b}_1, \dots, \mathbf{b}_{n-d}$ be a basis for the orthogonal complement, A^\perp .

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

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

Orthogonal Complement

Orthogonal Complement and Projections

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

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

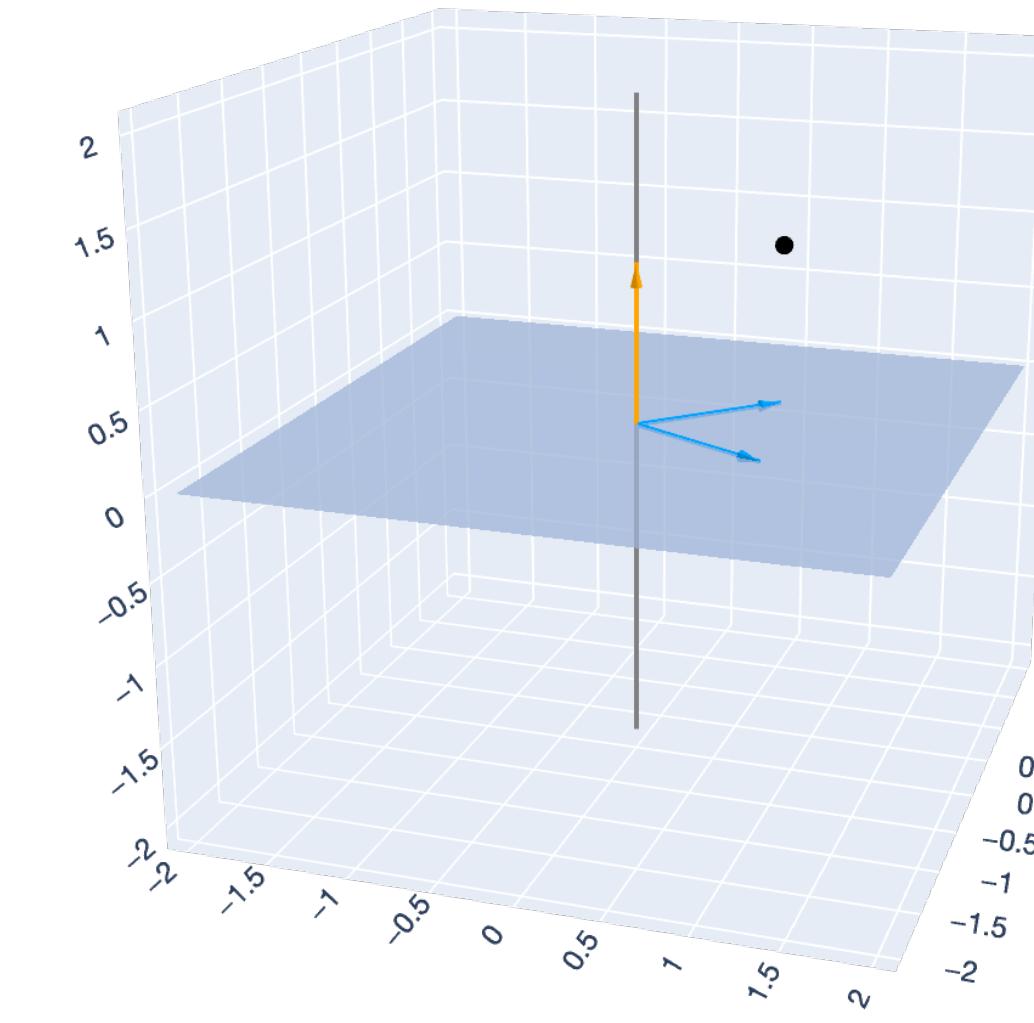
Orthogonal Complement

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



— u1 — u2 — v1 — — • 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 $\mathbf{b}_1, \dots, \mathbf{b}_{n-d}$, a basis for the orthogonal complement of $\text{span}(\text{col}(\mathbf{A}))$.

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

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

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 $\mathbf{b}_1, \dots, \mathbf{b}_{n-d}$, 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). $P_{\mathbf{A}} + P_{\mathbf{B}} = \mathbf{I}$.

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 $\mathbf{b}_1, \dots, \mathbf{b}_{n-d}$, 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). $P_{\mathbf{A}} = P_{\mathbf{A}}P_{\mathbf{A}} = P_{\mathbf{A}}^2$.

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 $\mathbf{b}_1, \dots, \mathbf{b}_{n-d}$, 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). $P_{\mathbf{A}} = P_{\mathbf{A}}^\top$.

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 $\mathbf{b}_1, \dots, \mathbf{b}_{n-d}$, 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). $P_{\mathbf{A}} = P_{\mathbf{A}}^\top$.

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

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

Singular Value Decomposition

1D Intuition and Derivation

Singular Value Decomposition (SVD)

1D Picture

Observe data $\mathbf{x}_1, \dots, \mathbf{x}_d \in \mathbb{R}^n$, 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 $\mathcal{U} \subseteq \mathbb{R}^n$ that fits the points.

Singular Value Decomposition (SVD)

1D Picture

Observe data $\mathbf{x}_1, \dots, \mathbf{x}_d \in \mathbb{R}^n$, 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 $\mathcal{U} \subseteq \mathbb{R}^n$ that fits the points.

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

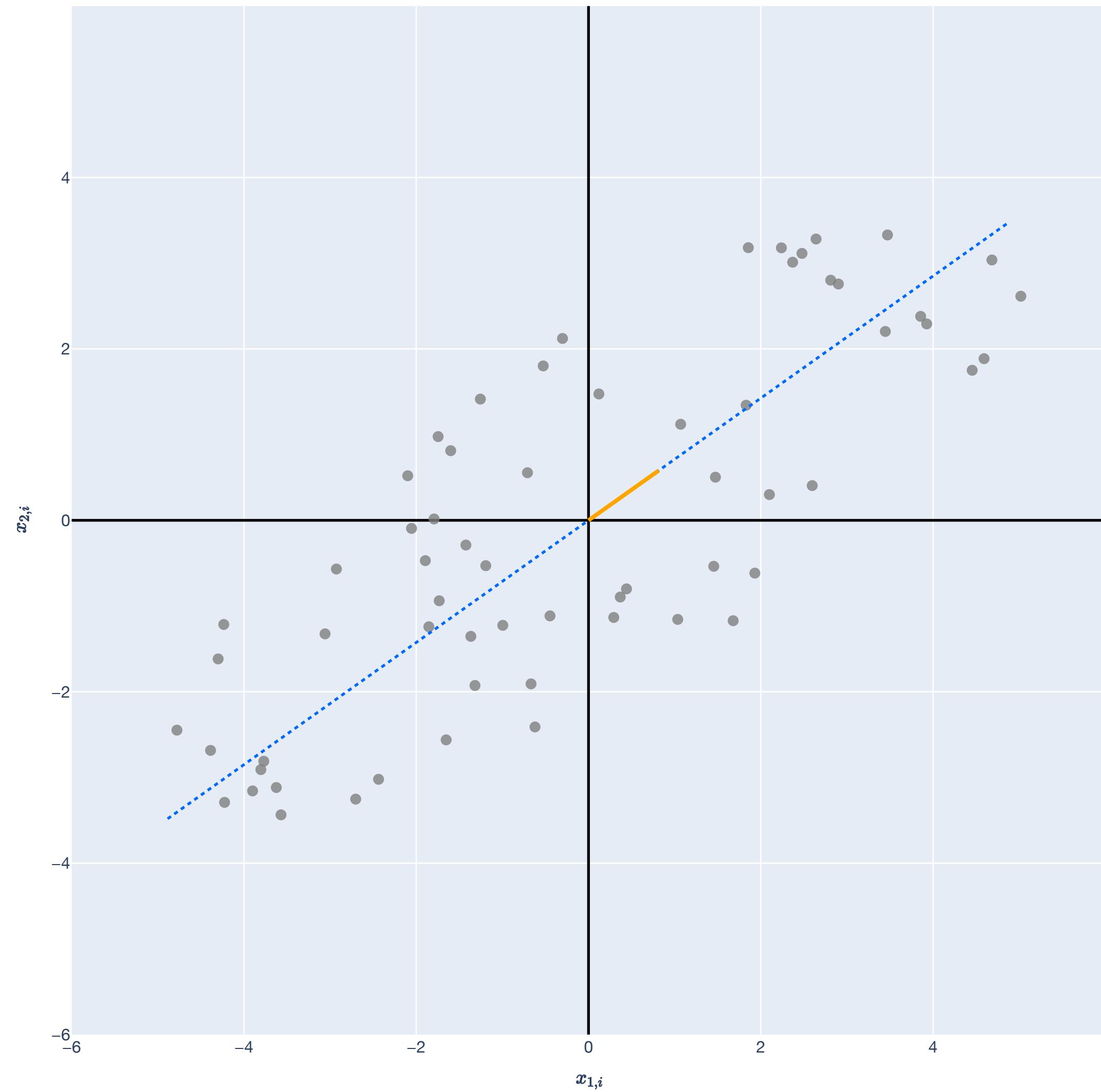
$$\mathcal{U} = \{c\mathbf{u} : c \in \mathbb{R}\}.$$

Singular Value Decomposition (SVD)

1D Picture

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

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



Singular Value Decomposition (SVD)

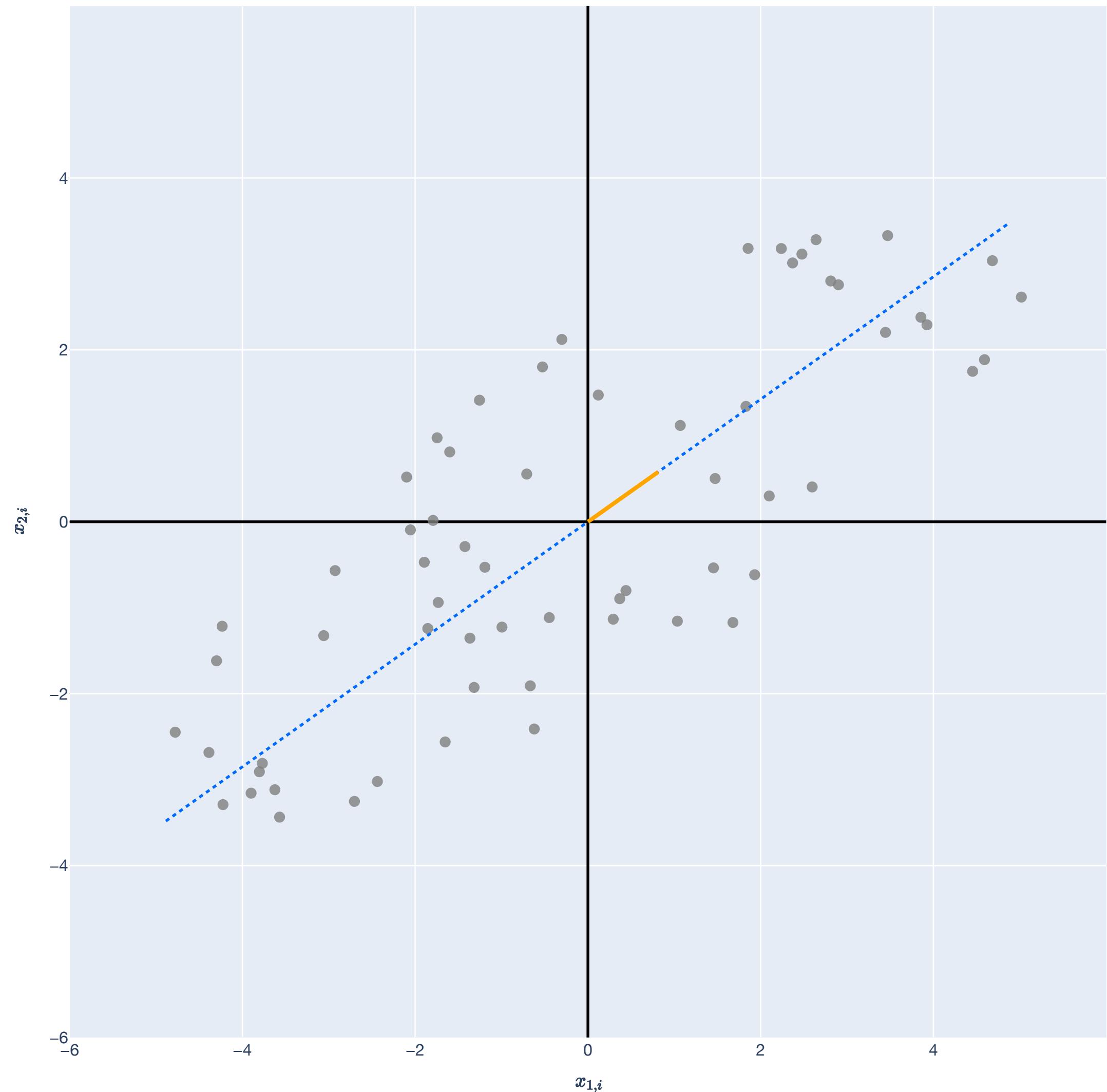
1D Picture

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

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

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

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

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

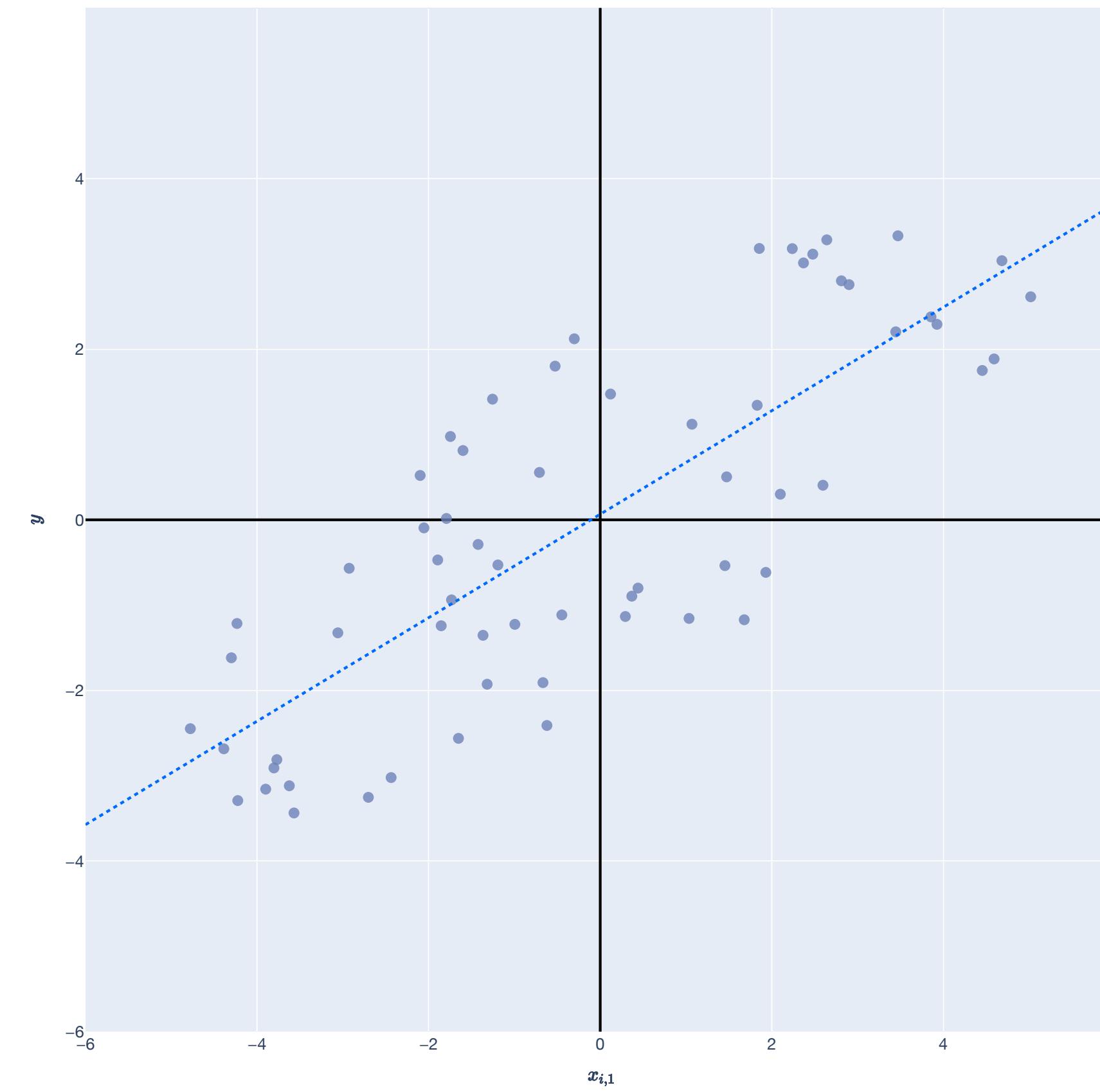
Goal: Find one-dimensional subspace determined by $\mathbf{u} \in \mathbb{R}^n$ such that

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

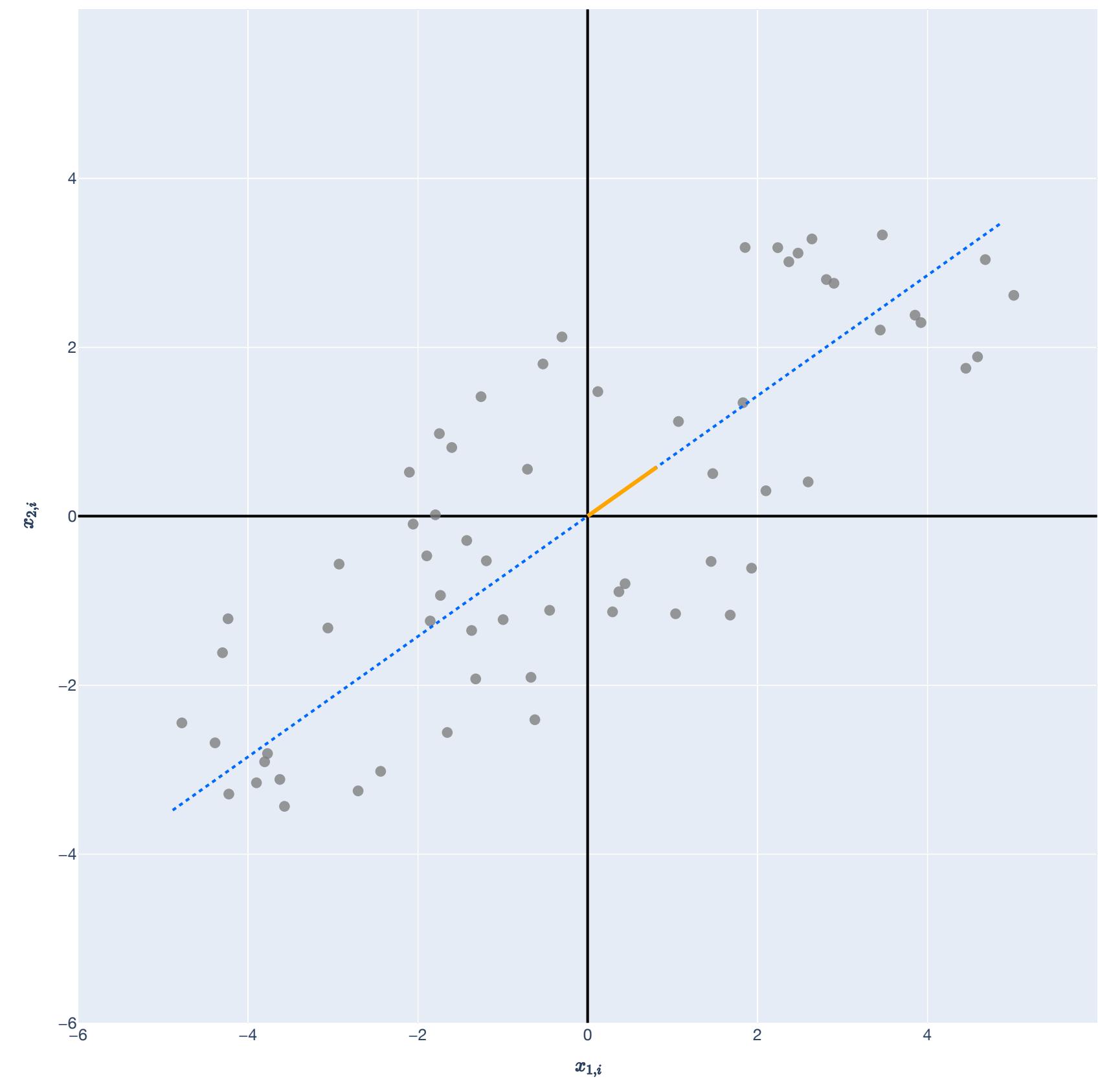
Comparison with OLS

1D Pictures

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



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



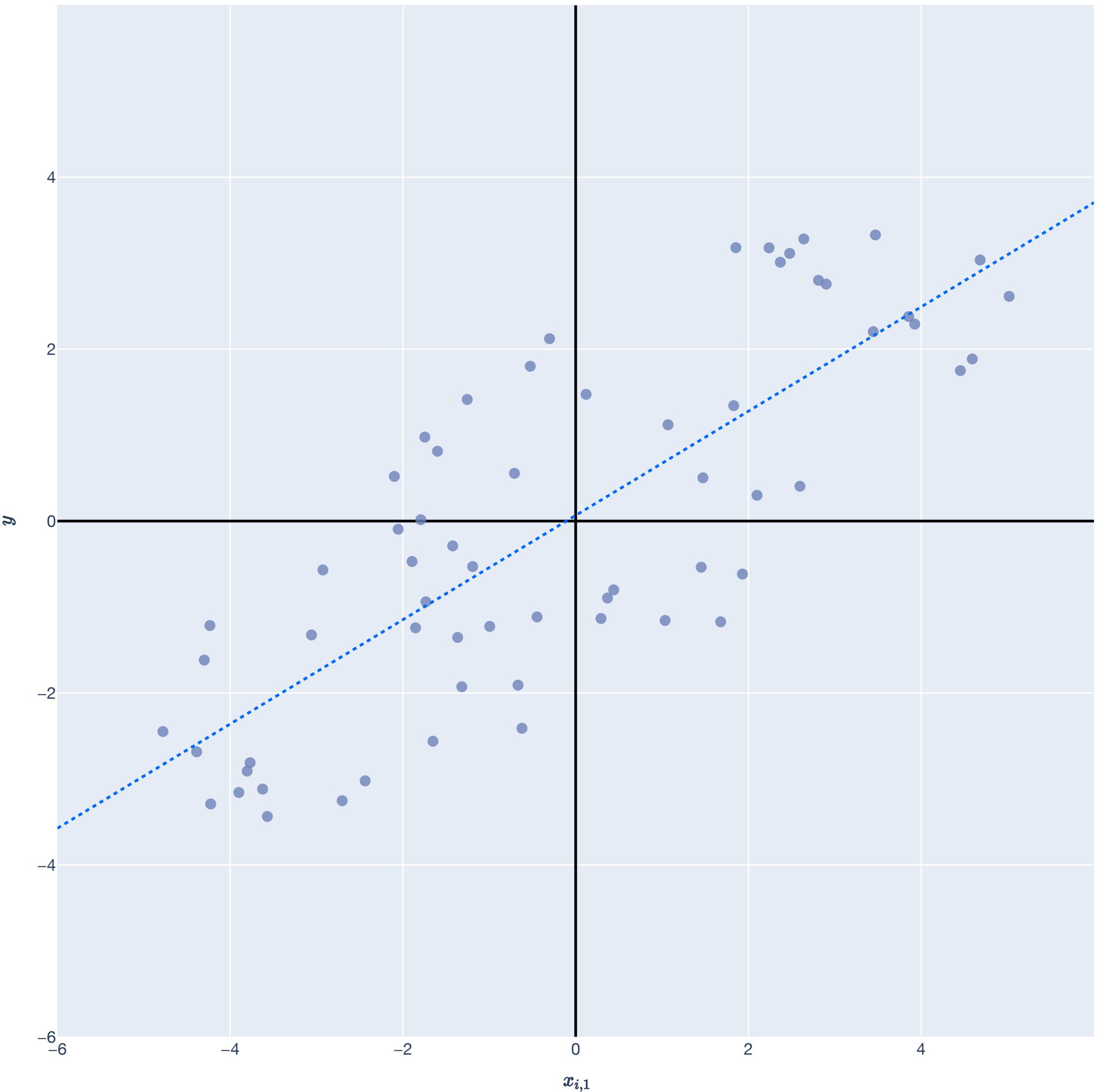
Comparison with OLS

1D Pictures

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

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

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



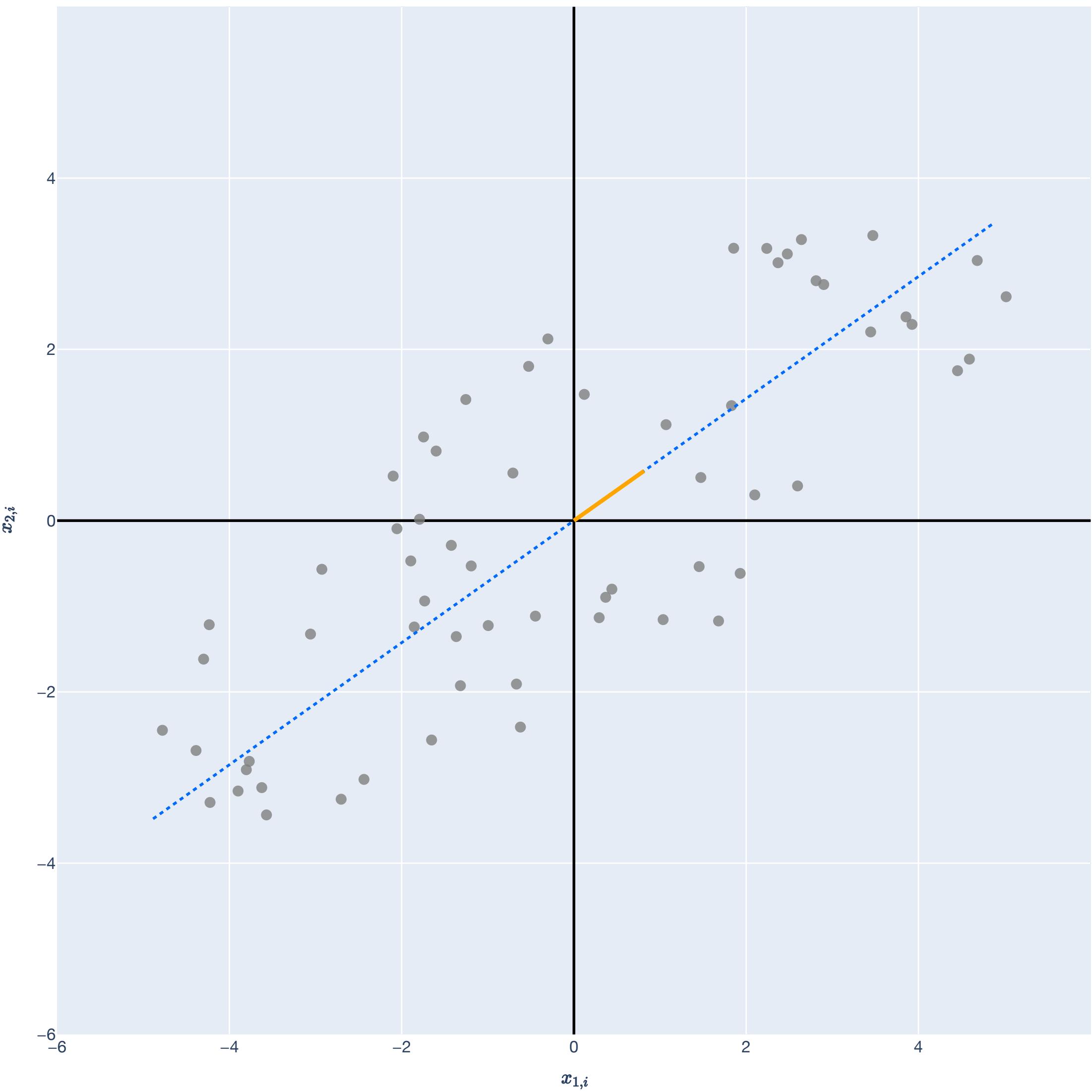
Comparison with OLS

1D Pictures

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

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

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



Singular Value Decomposition (SVD)

Deriving 1D SVD

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

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

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

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

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

Singular Value Decomposition (SVD)

Deriving 1D SVD

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

$$\begin{aligned}\|\mathbf{x}_i - P_{\mathbf{u}}\mathbf{x}_i\|^2 &= \left\| \mathbf{x}_i - \left(\frac{\mathbf{u}\mathbf{u}^\top}{\mathbf{u}^\top\mathbf{u}} \right) \mathbf{x}_i \right\|^2 && \text{(Prop: 1D projection formula)} \\ &= \left\| \left(\mathbf{I} - \frac{\mathbf{u}\mathbf{u}^\top}{\mathbf{u}^\top\mathbf{u}} \right) \mathbf{x}_i \right\|^2 && \text{(Prop: Projection and Orthogonal Complement Matrices)}\end{aligned}$$

Singular Value Decomposition (SVD)

Deriving 1D SVD

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

$$\begin{aligned}\|\mathbf{x}_i - P_{\mathbf{u}}\mathbf{x}_i\|^2 &= \left\| \mathbf{x}_i - \left(\frac{\mathbf{u}\mathbf{u}^T}{\mathbf{u}^T\mathbf{u}} \right) \mathbf{x}_i \right\|^2 && \text{(Prop: 1D projection formula)} \\ &= \left\| \left(\mathbf{I} - \frac{\mathbf{u}\mathbf{u}^T}{\mathbf{u}^T\mathbf{u}} \right) \mathbf{x}_i \right\|^2 && \text{(Prop: Projection and Orthogonal Complement Matrices)} \\ &= \mathbf{x}_i^T \left(\mathbf{I} - \frac{\mathbf{u}\mathbf{u}^T}{\mathbf{u}^T\mathbf{u}} \right)^T \left(\mathbf{I} - \frac{\mathbf{u}\mathbf{u}^T}{\mathbf{u}^T\mathbf{u}} \right) \mathbf{x}_i\end{aligned}$$

Singular Value Decomposition (SVD)

Deriving 1D SVD

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

$$\begin{aligned}\|\mathbf{x}_i - P_{\mathbf{u}}\mathbf{x}_i\|^2 &= \left\| \mathbf{x}_i - \left(\frac{\mathbf{u}\mathbf{u}^T}{\mathbf{u}^T\mathbf{u}} \right) \mathbf{x}_i \right\|^2 && \text{(Prop: 1D projection formula)} \\ &= \left\| \left(\mathbf{I} - \frac{\mathbf{u}\mathbf{u}^T}{\mathbf{u}^T\mathbf{u}} \right) \mathbf{x}_i \right\|^2 && \text{(Prop: Projection and Orthogonal Complement Matrices)} \\ &= \mathbf{x}_i^T \left(\mathbf{I} - \frac{\mathbf{u}\mathbf{u}^T}{\mathbf{u}^T\mathbf{u}} \right)^T \left(\mathbf{I} - \frac{\mathbf{u}\mathbf{u}^T}{\mathbf{u}^T\mathbf{u}} \right) \mathbf{x}_i \\ &= \mathbf{x}_i^T \left(\mathbf{I} - \frac{\mathbf{u}\mathbf{u}^T}{\mathbf{u}^T\mathbf{u}} \right)^2 \mathbf{x}_i && \text{(Prop: Projections are symmetric)}\end{aligned}$$

Singular Value Decomposition (SVD)

Deriving 1D SVD

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

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

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

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

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

$$= \mathbf{x}_i^T \left(\mathbf{I} - \frac{\mathbf{u} \mathbf{u}^T}{\mathbf{u}^T \mathbf{u}} \right) \mathbf{x}_i \quad (\text{Prop: Projecting twice doesn't do anything})$$

Singular Value Decomposition (SVD)

Deriving 1D SVD

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

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

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:

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

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:

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

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:

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

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:

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

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

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

Singular Value Decomposition (SVD)

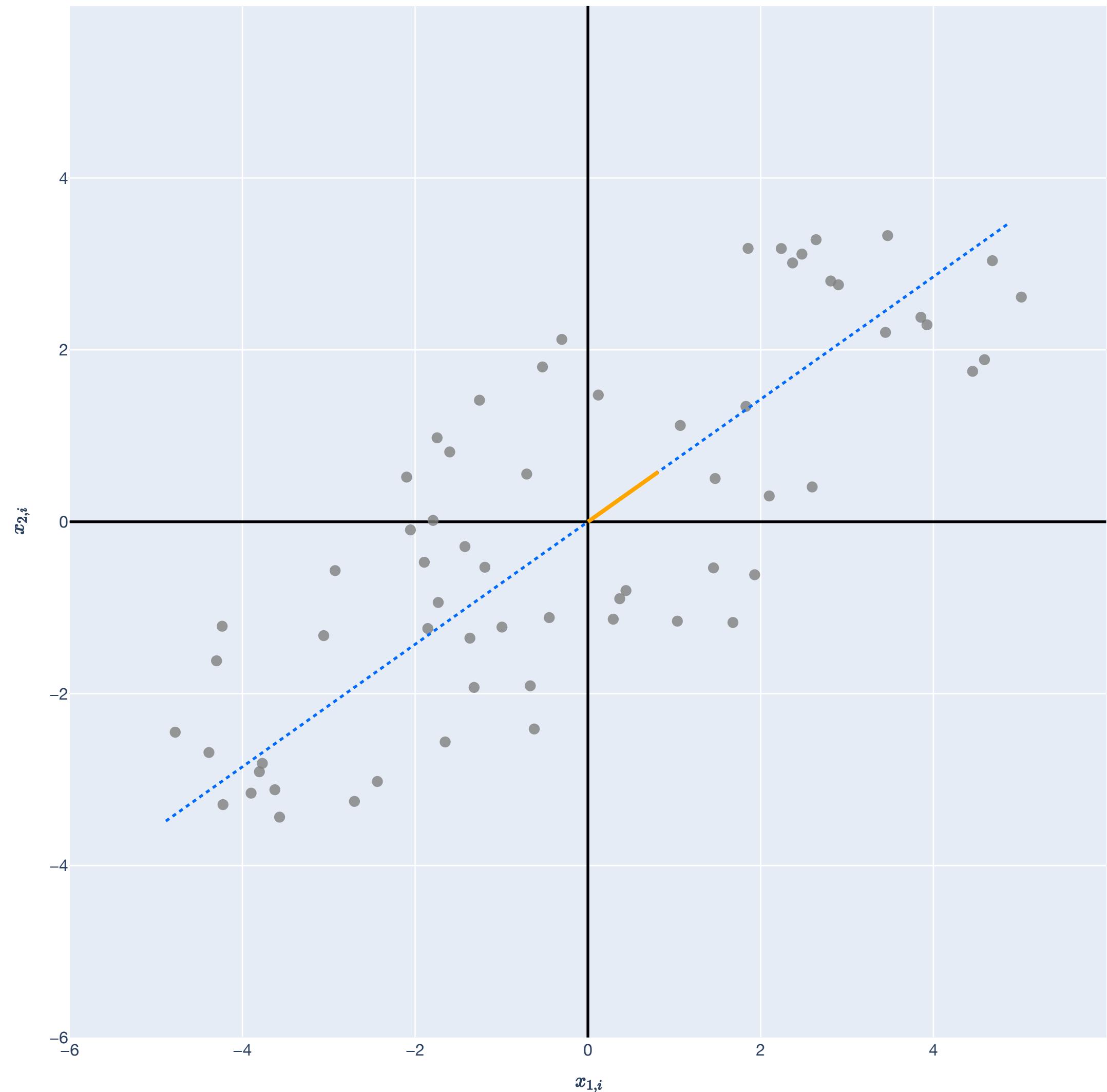
1D Picture

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

Goal: 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:

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



Singular Value Decomposition (SVD)

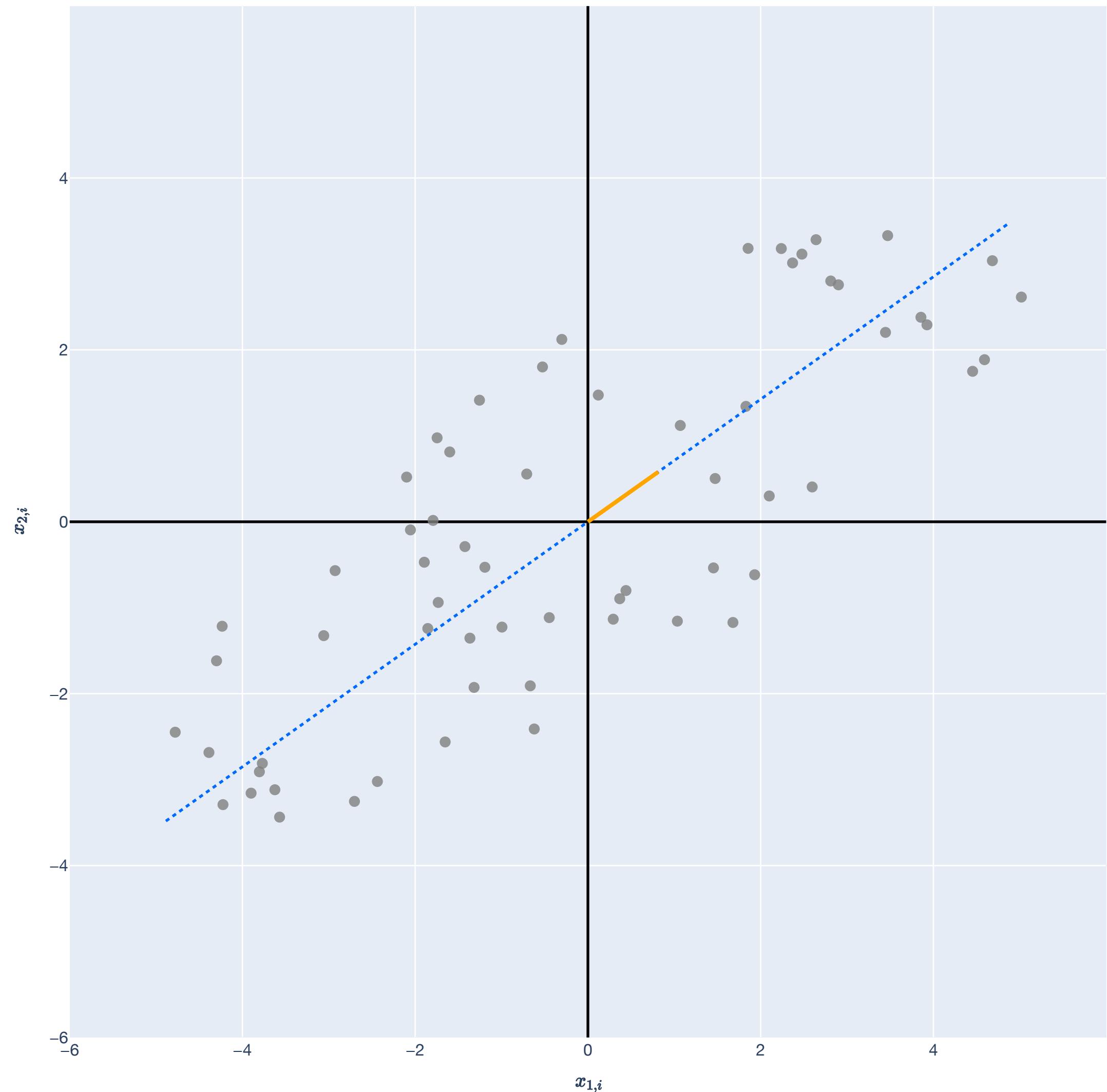
1D Picture

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

Goal: 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:

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



Singular Value Decomposition (SVD)

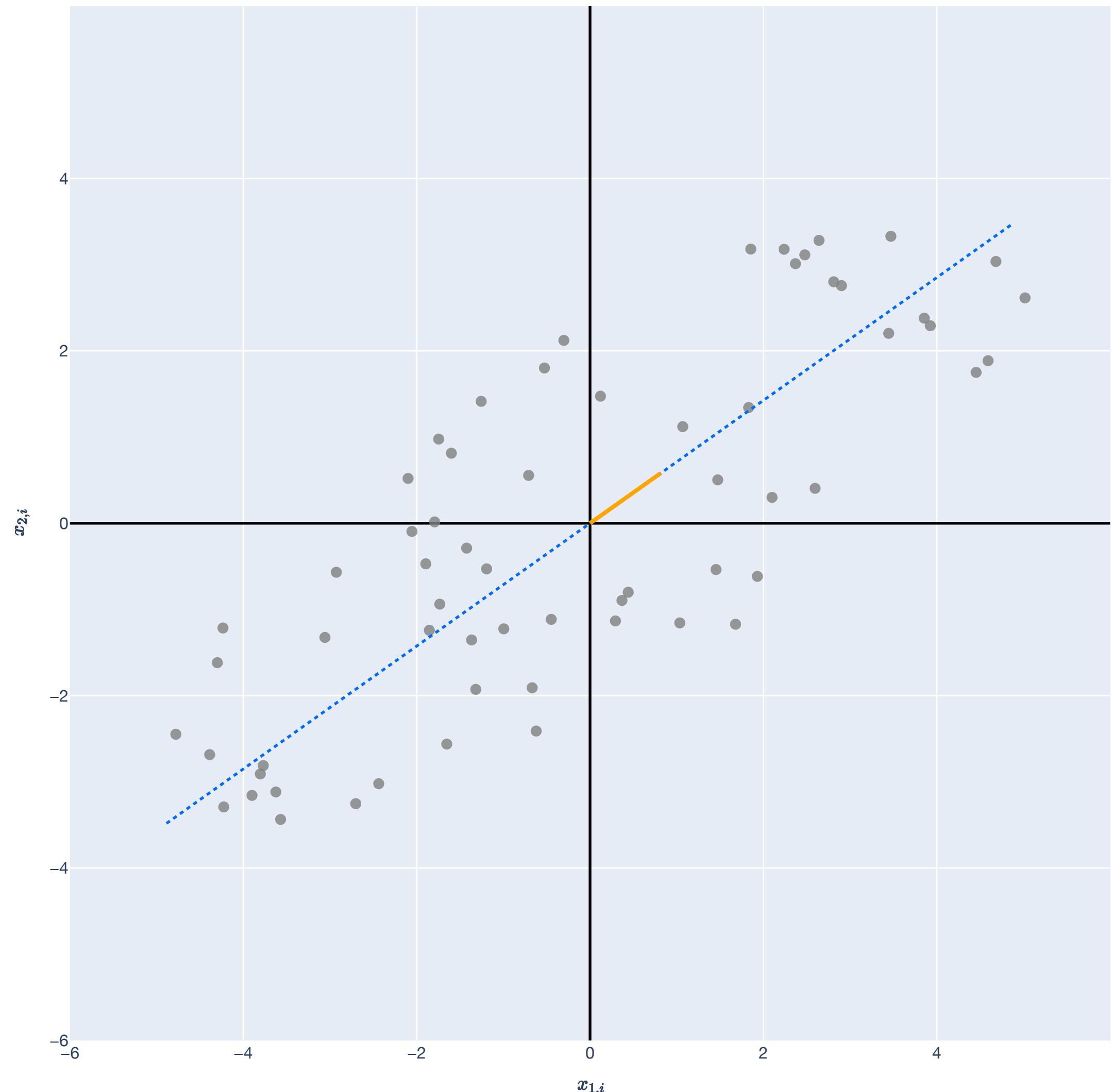
1D Picture

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

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

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

The value $\frac{\mathbf{u}^\top \mathbf{X} \mathbf{X}^\top \mathbf{u}}{\mathbf{u}^\top \mathbf{u}}$ 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 $\mathbf{x}_1, \dots, \mathbf{x}_d \in \mathbb{R}^n$. Consider the following procedure...

For $t = 1, 2, \dots, n$...

1. Find $\mathbf{u}_1 \in \mathbb{R}^n$, the best one-dimensional subspace fit to $\mathbf{x}_1, \dots, \mathbf{x}_d$.
 - Let $\mathbf{x}_i^{(1)} = \mathbf{x}_i - \Pi_{\mathbf{u}_1}(\mathbf{x}_i)$.

Singular Value Decomposition (SVD)

Building up the SVD

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

For $t = 1, 2, \dots, n$...

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

For $t = 1, 2, \dots, n$...

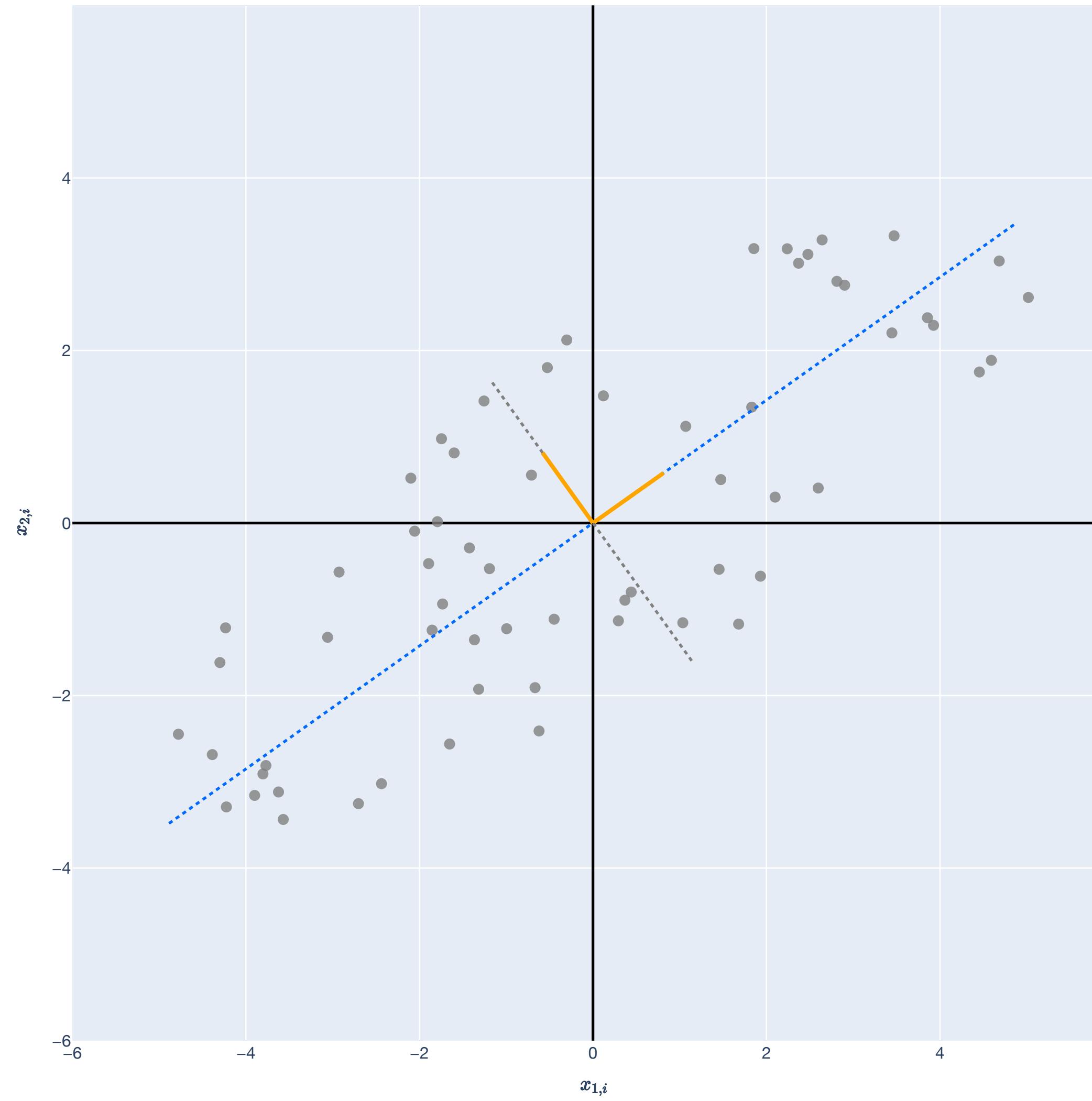
1. Find $\mathbf{u}_1 \in \mathbb{R}^n$, the best one-dimensional subspace fit to $\mathbf{x}_1, \dots, \mathbf{x}_d$.
 - Let $\mathbf{x}_i^{(1)} = \mathbf{x}_i - \Pi_{\mathbf{u}_1}(\mathbf{x}_i)$.
2. Find $\mathbf{u}_2 \in \mathbb{R}^n$, the best one-dimensional subspace fit to $\mathbf{x}_1^{(1)},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

Singular Value Decomposition (SVD)

Building up the SVD

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

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:

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

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

The n left singular vectors are orthogonal, by construction (left singular vector \mathbf{u}_k is in the orthogonal complement of $\mathbf{u}_1, \dots, \mathbf{u}_{k-1}$).

Singular Value Decomposition (SVD)

Definition of the Full SVD

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

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

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

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

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

Singular Value Decomposition (SVD)

Shape of the Σ Matrix

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

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

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

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

Interpreting the SVD

Example in \mathbb{R}^2

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

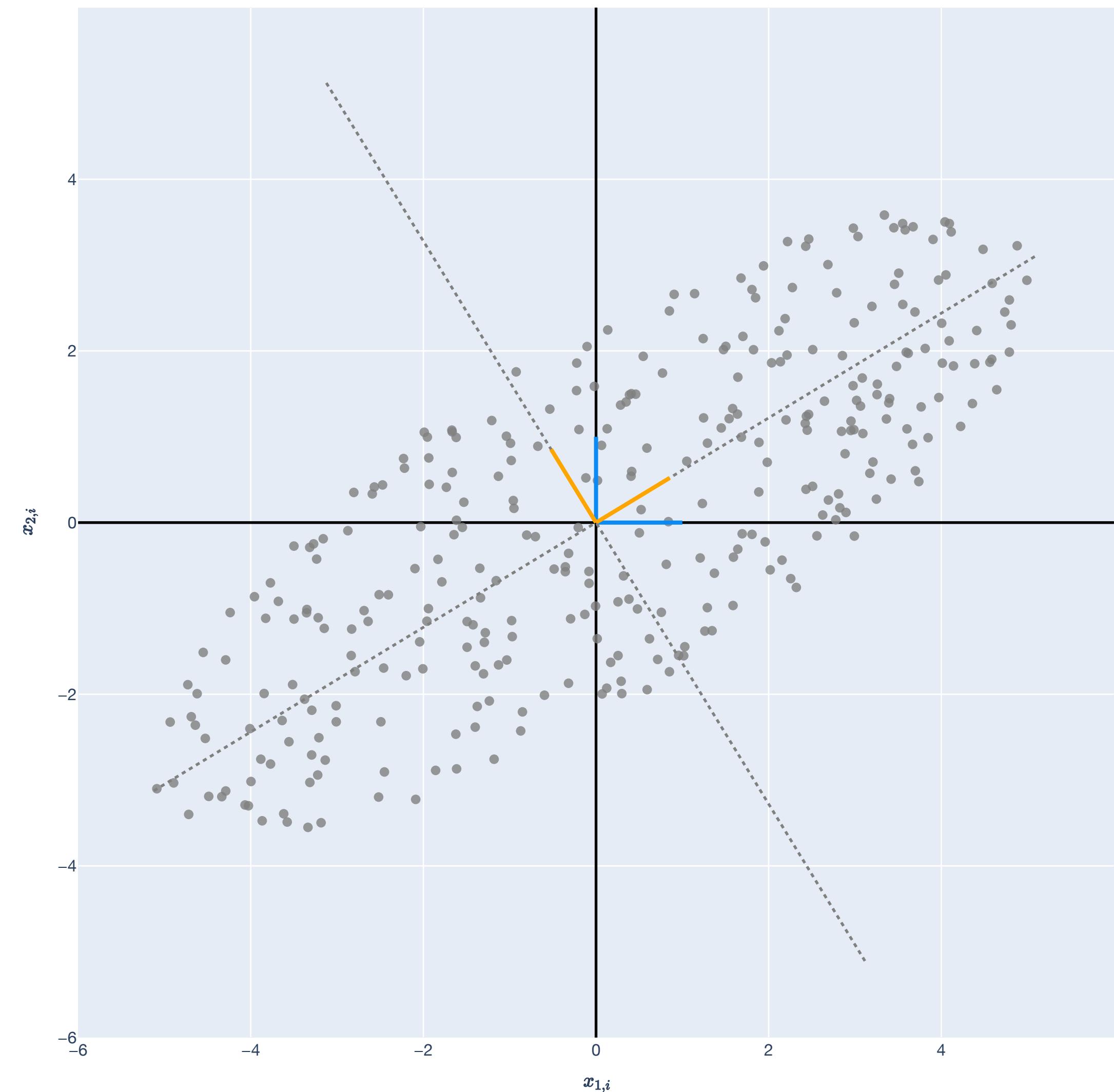
$$\underbrace{\mathbf{X}}_{2 \times 12} = \underbrace{\mathbf{U}}_{2 \times 2} \underbrace{\Sigma}_{2 \times 12} \underbrace{\mathbf{V}^\top}_{12 \times 12}.$$

Left Singular Vectors

Interpreting the \mathbf{U} matrix

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

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



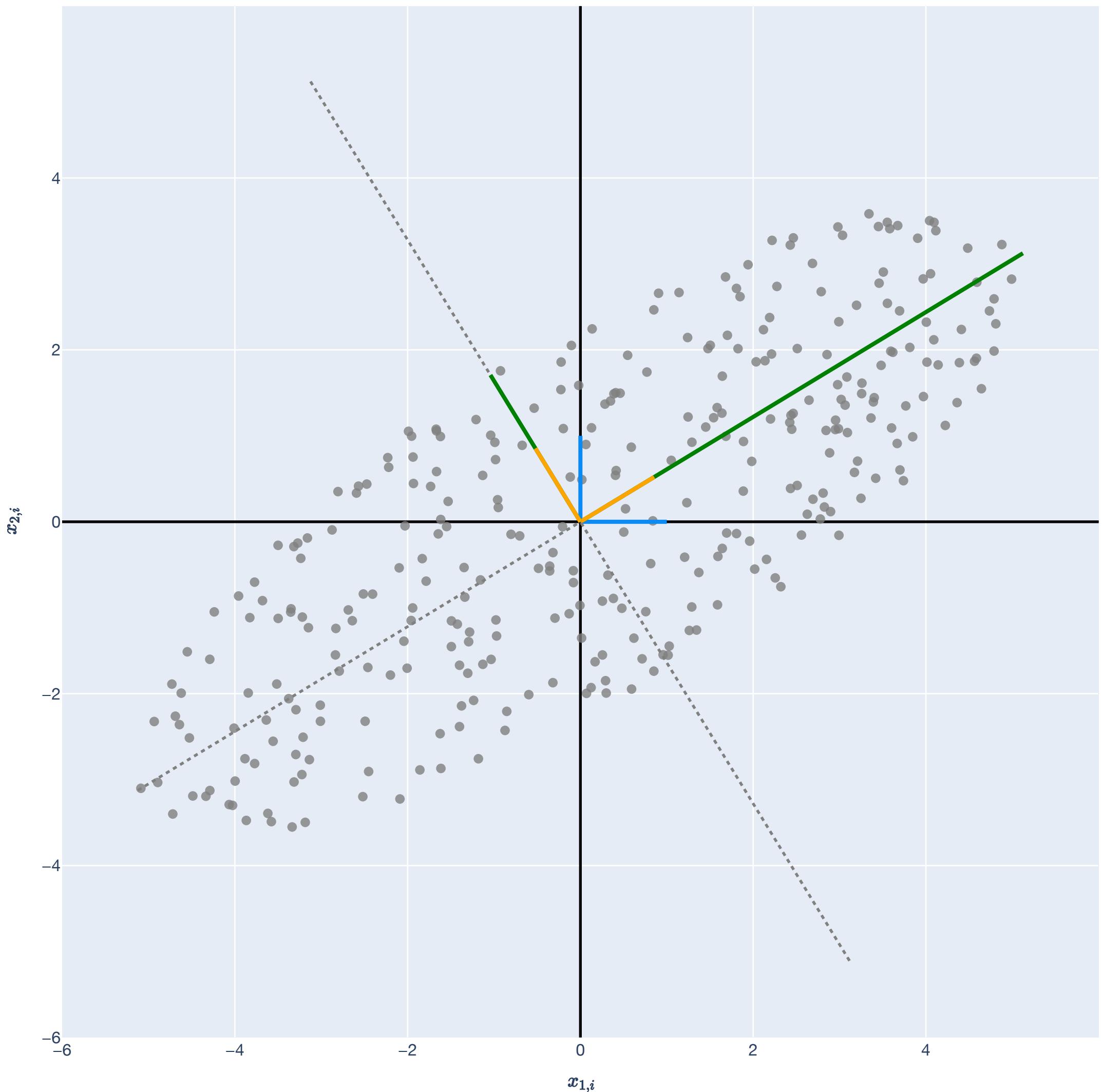
Singular Values

Interpreting the Σ matrix

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

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

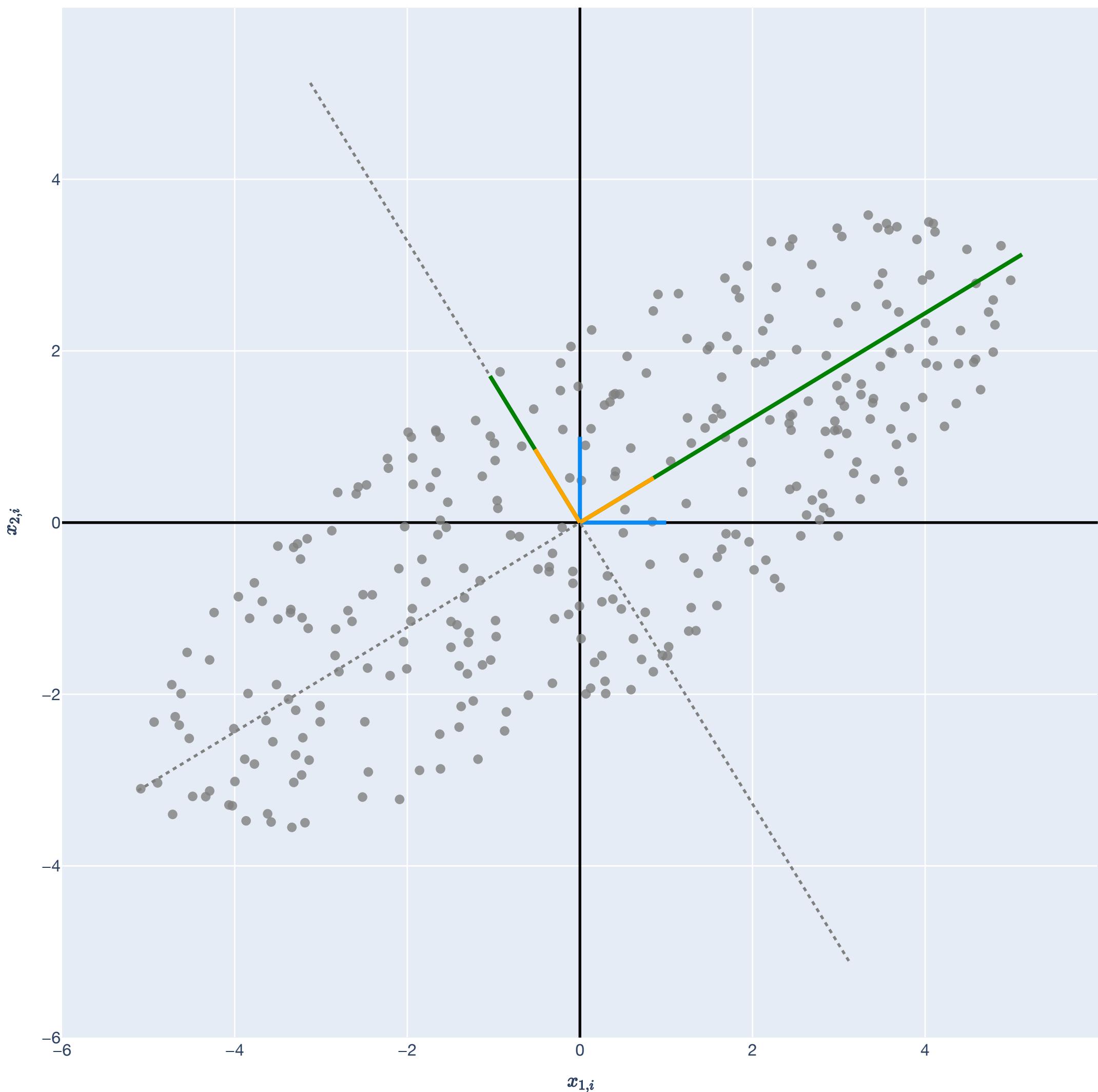
Interpreting the \mathbf{V} matrix

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

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

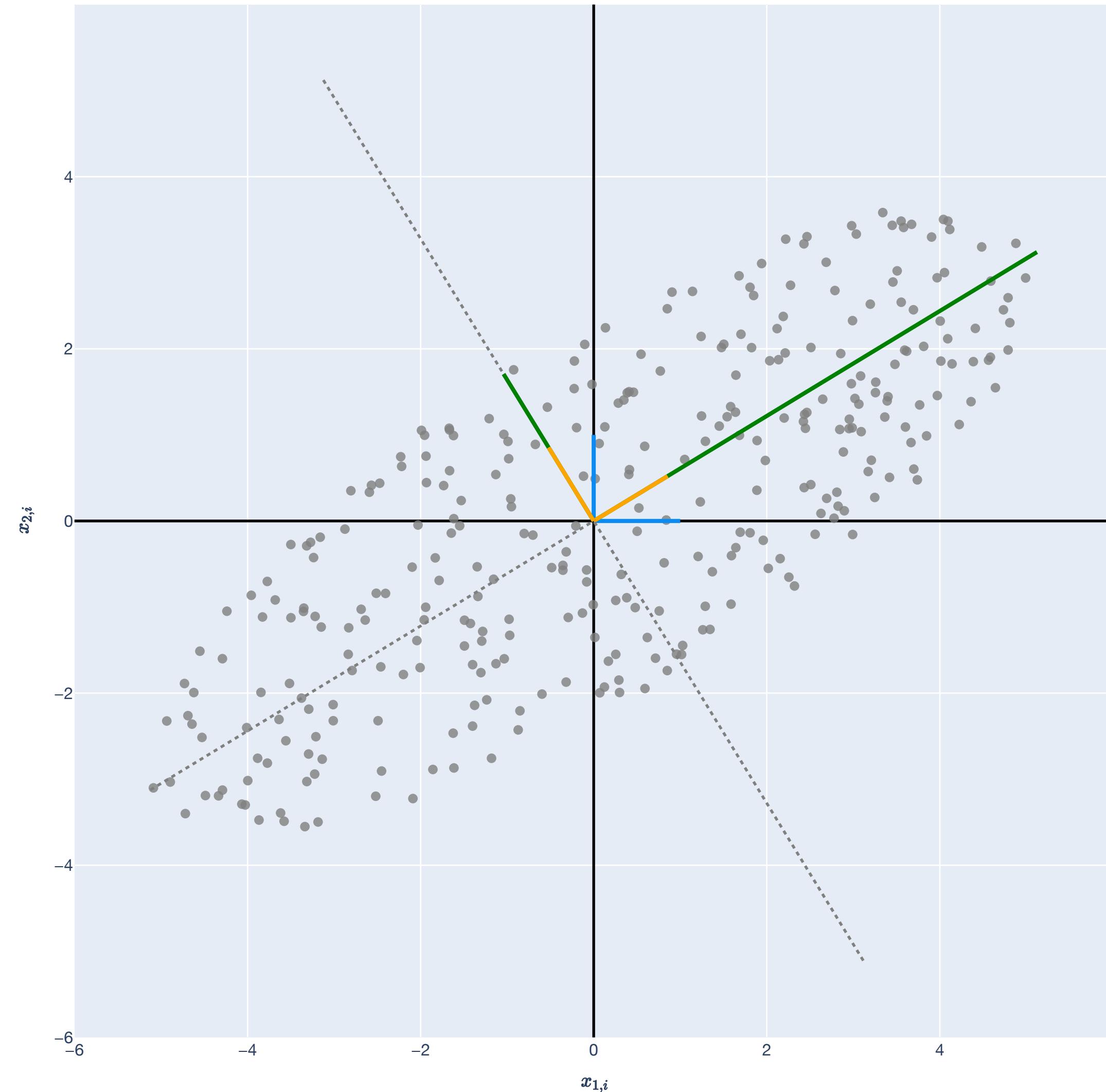
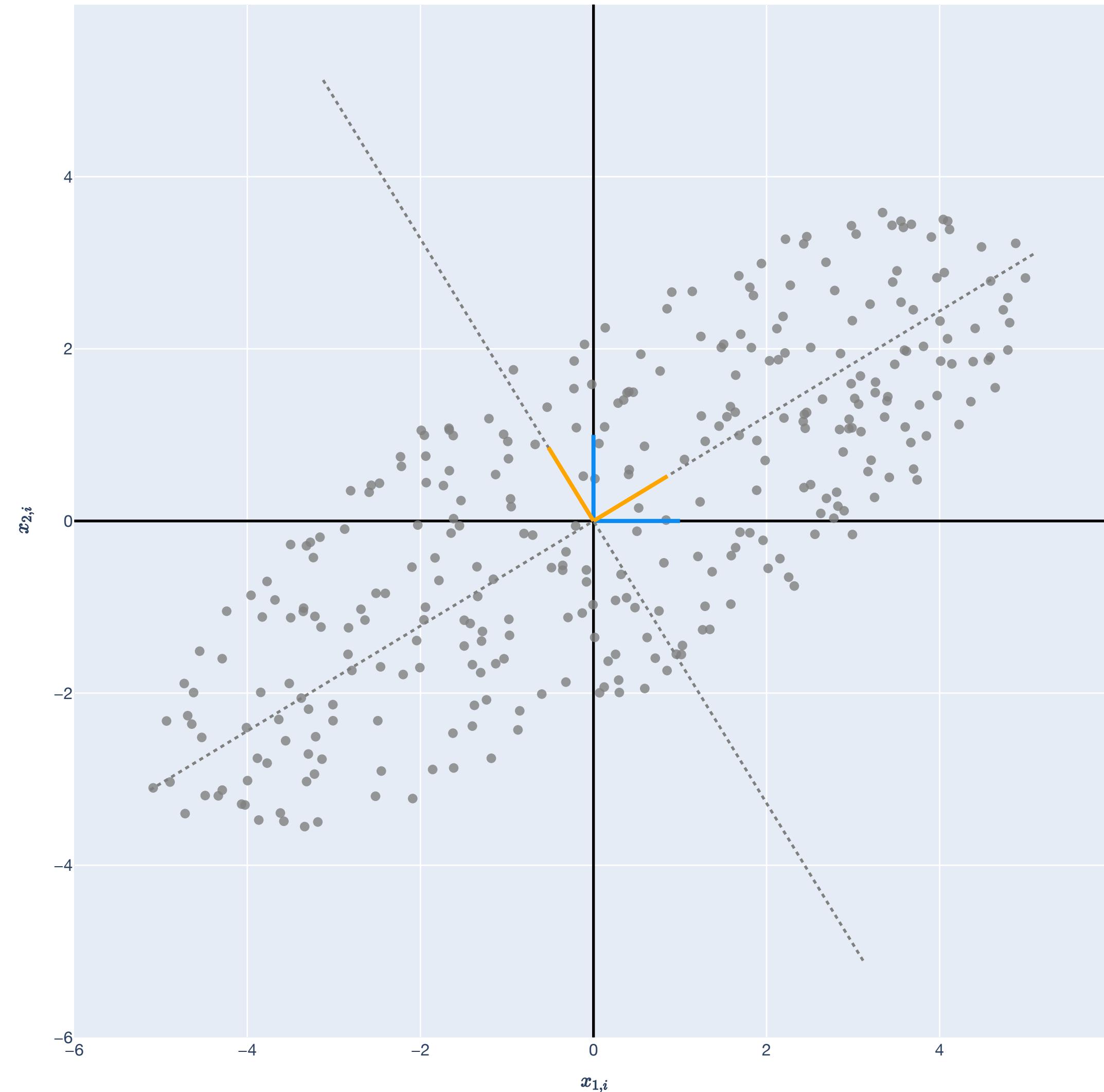
Specifically, for $j \in [d]$,

$$\mathbf{x}_j = \sigma_1 \mathbf{u}_1 \mathbf{v}_{j1} + \sigma_2 \mathbf{u}_2 \mathbf{v}_{j2}.$$



Interpretation of the SVD

Full Interpretation of the SVD



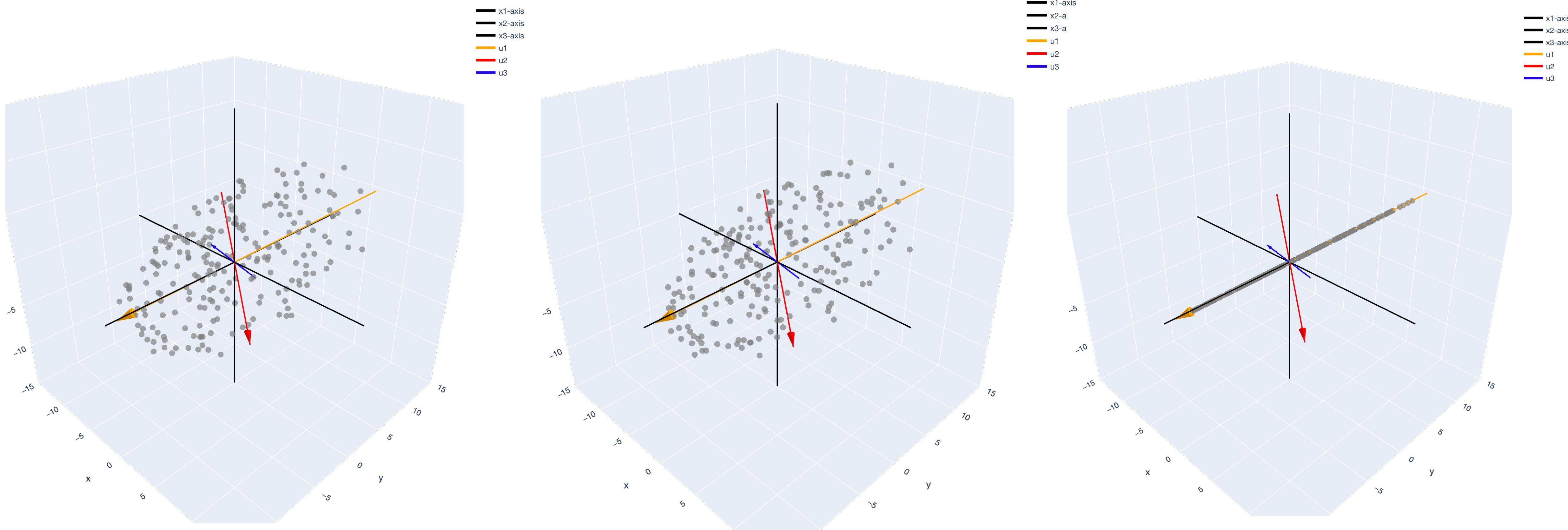
Singular Value Decomposition (SVD)

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

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

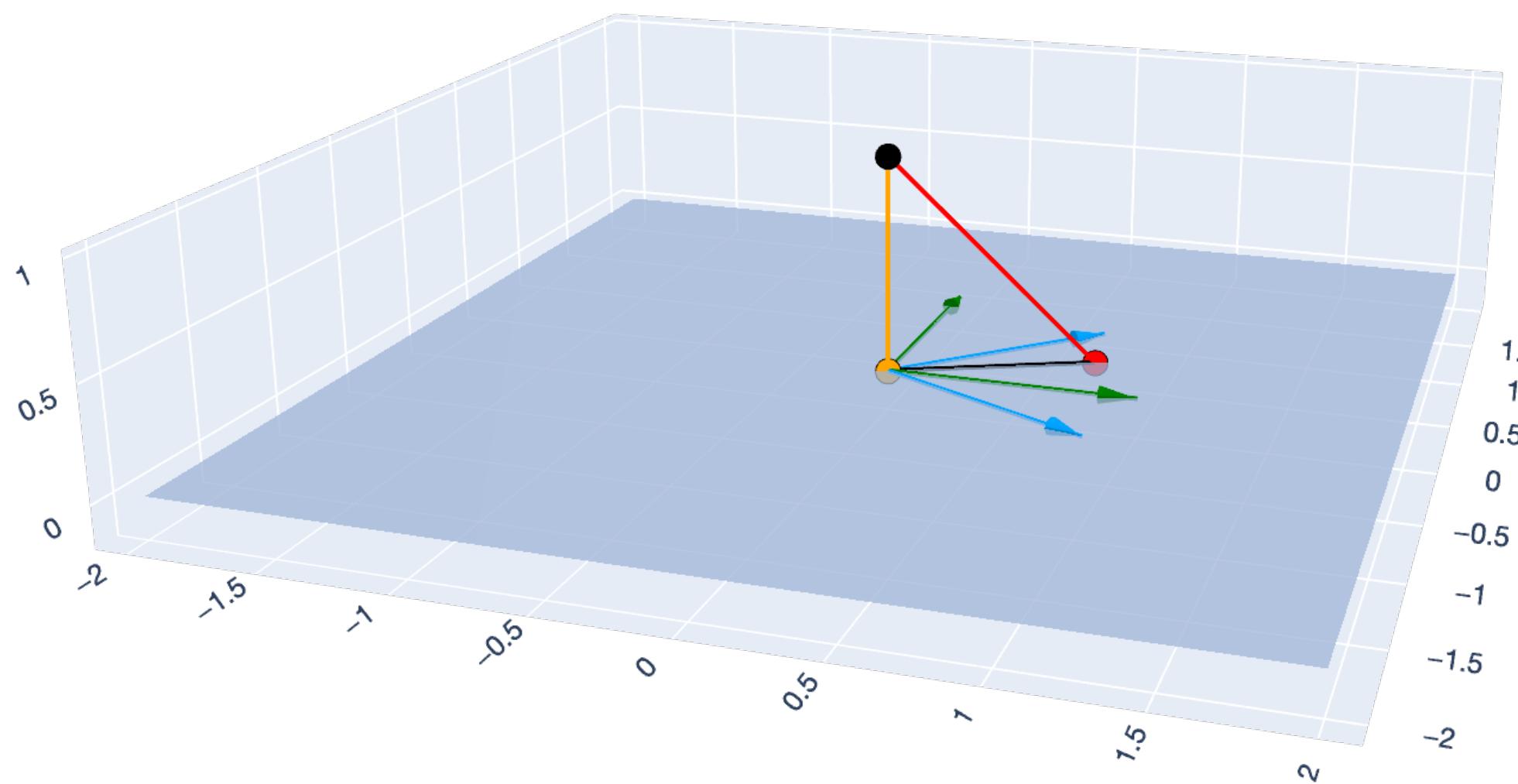
$$\underbrace{\mathbf{X}}_{n \times d} = \underbrace{\mathbf{U}}_{n \times r} \underbrace{\Sigma}_{r \times r} \underbrace{\mathbf{V}^T}_{r \times d}.$$

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

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

$\Sigma \in \mathbb{R}^{r \times r}$ is a diagonal matrix with [singular values](#) $\sigma_1 \geq \sigma_2 \geq \dots \geq \sigma_r > 0$ on the diagonal.

How to find a good orthogonal basis?



— x_1 — x_2 — u_1 — u_2 — $y - \hat{y}$ — $\sim y - \hat{y}$ ● y ○ \hat{y} ● $\sim y$

Least Squares

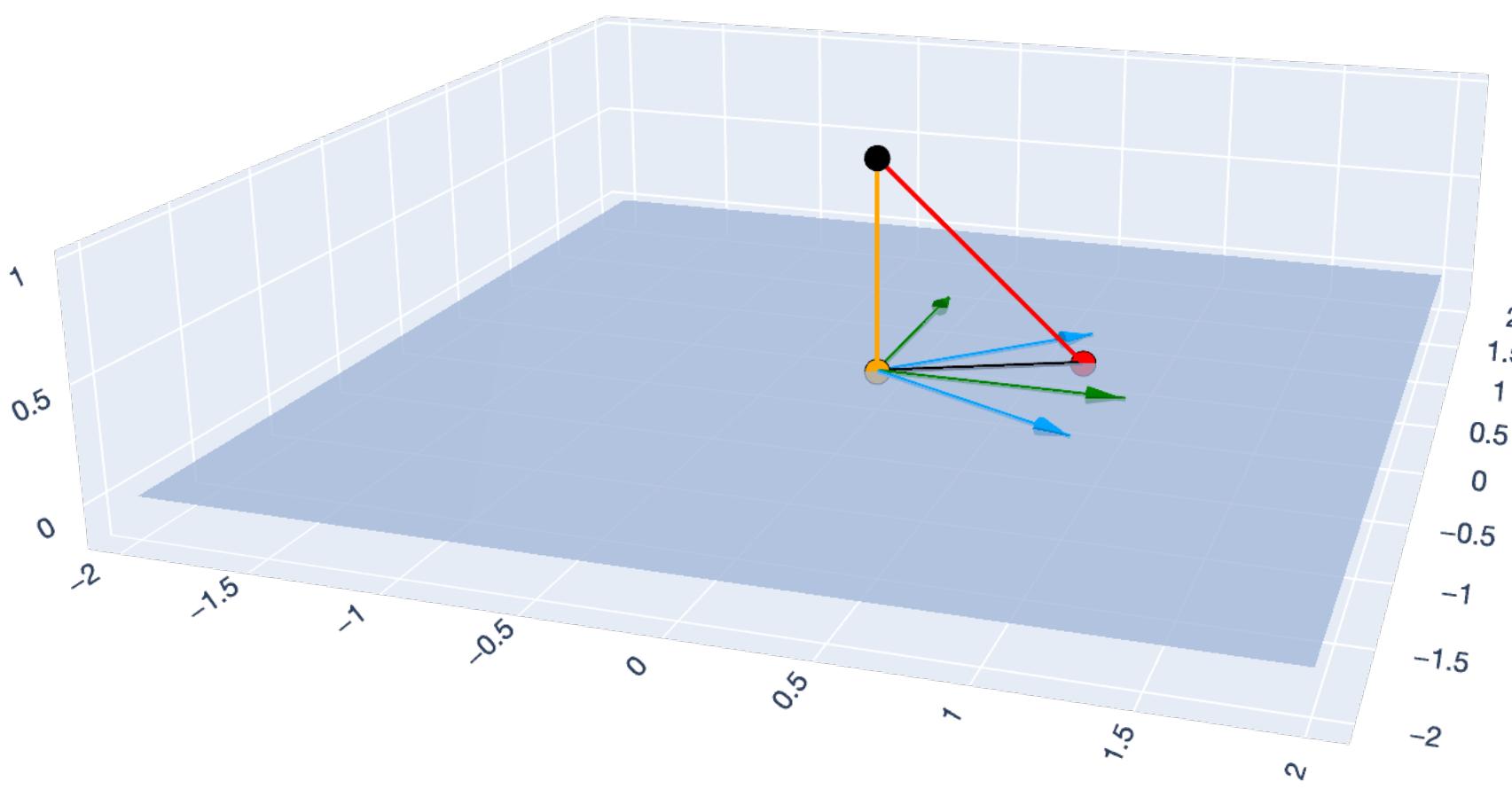
OLS with Orthogonal Basis

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

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

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

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



Least Squares

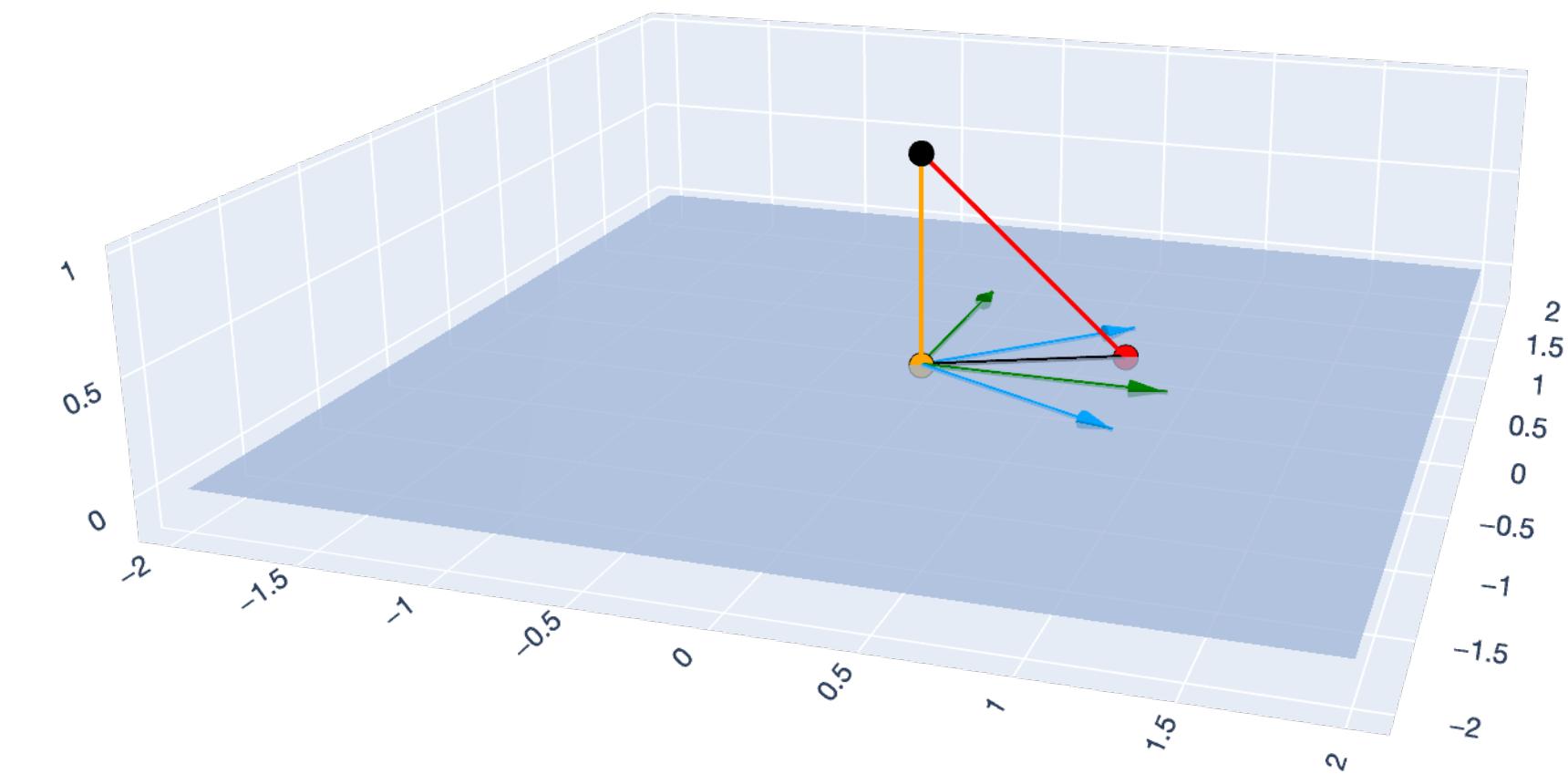
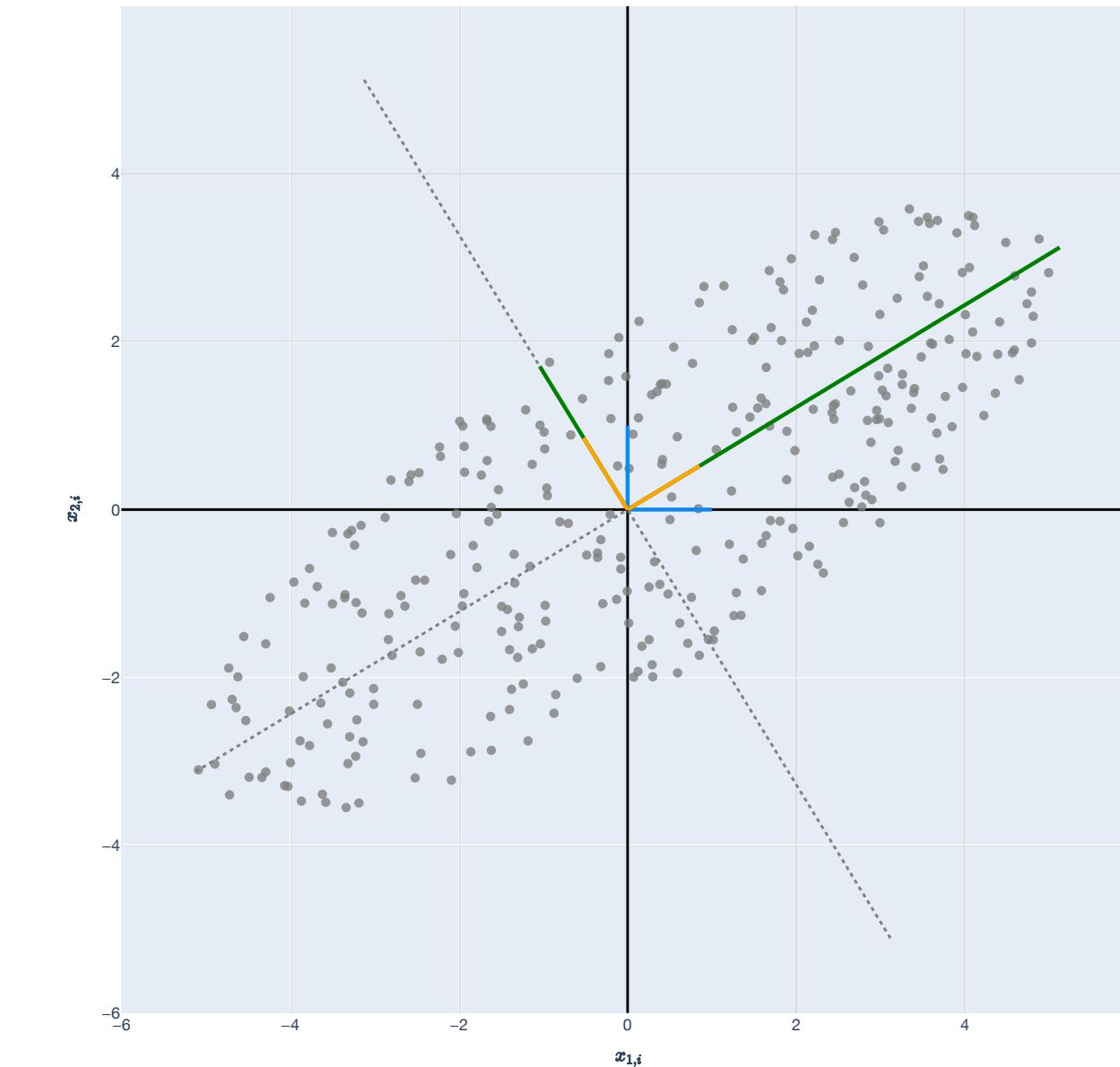
OLS with Orthogonal Basis

Prop (OLS using the ONB from Compact SVD).

Let $\mathbf{X} \in \mathbb{R}^{n \times d}$ and let $\mathcal{X} = \text{span}(\text{col}(\mathbf{X}))$ be a subspace, with dimension $\dim(\mathcal{X}) = \text{rank}(\mathbf{X}) = r$.

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

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



Legend:
— x1 — x2 — u1 — u2 — y - ^y — y - ~y — ~y - y
● y ● ^y ● ~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:

What's the closest rank- k matrix (with $k \ll r$) 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 $r = \text{rank}(\mathbf{X})$ 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 Statement

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:

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

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

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

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

Rank- k Approximation

Outer Product Interpretation

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

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

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

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

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

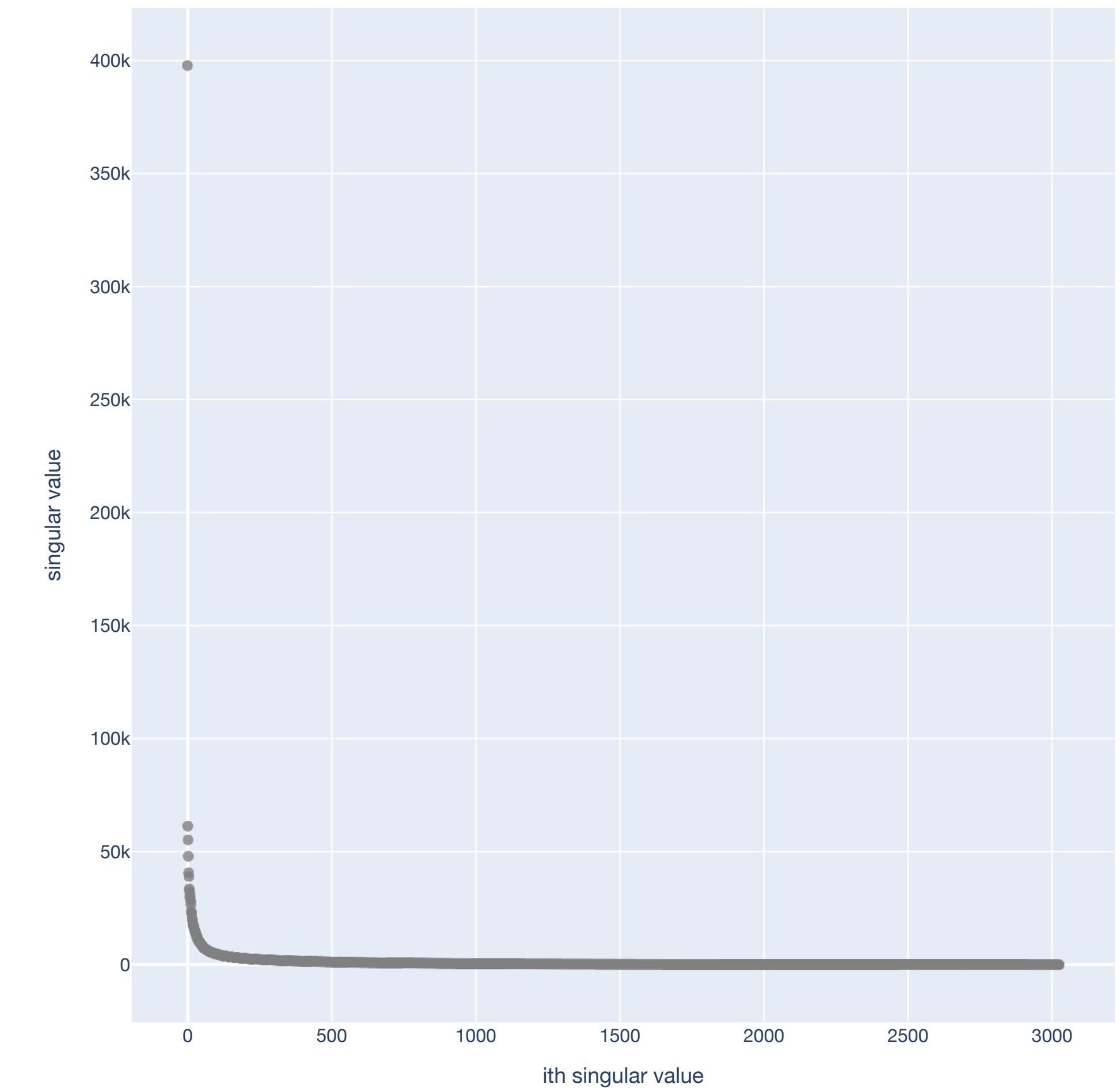
Application in Image Processing



```
print(X)
print("Shape: {}".format(X.shape))
✓ 0.0s
[[ 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)
```

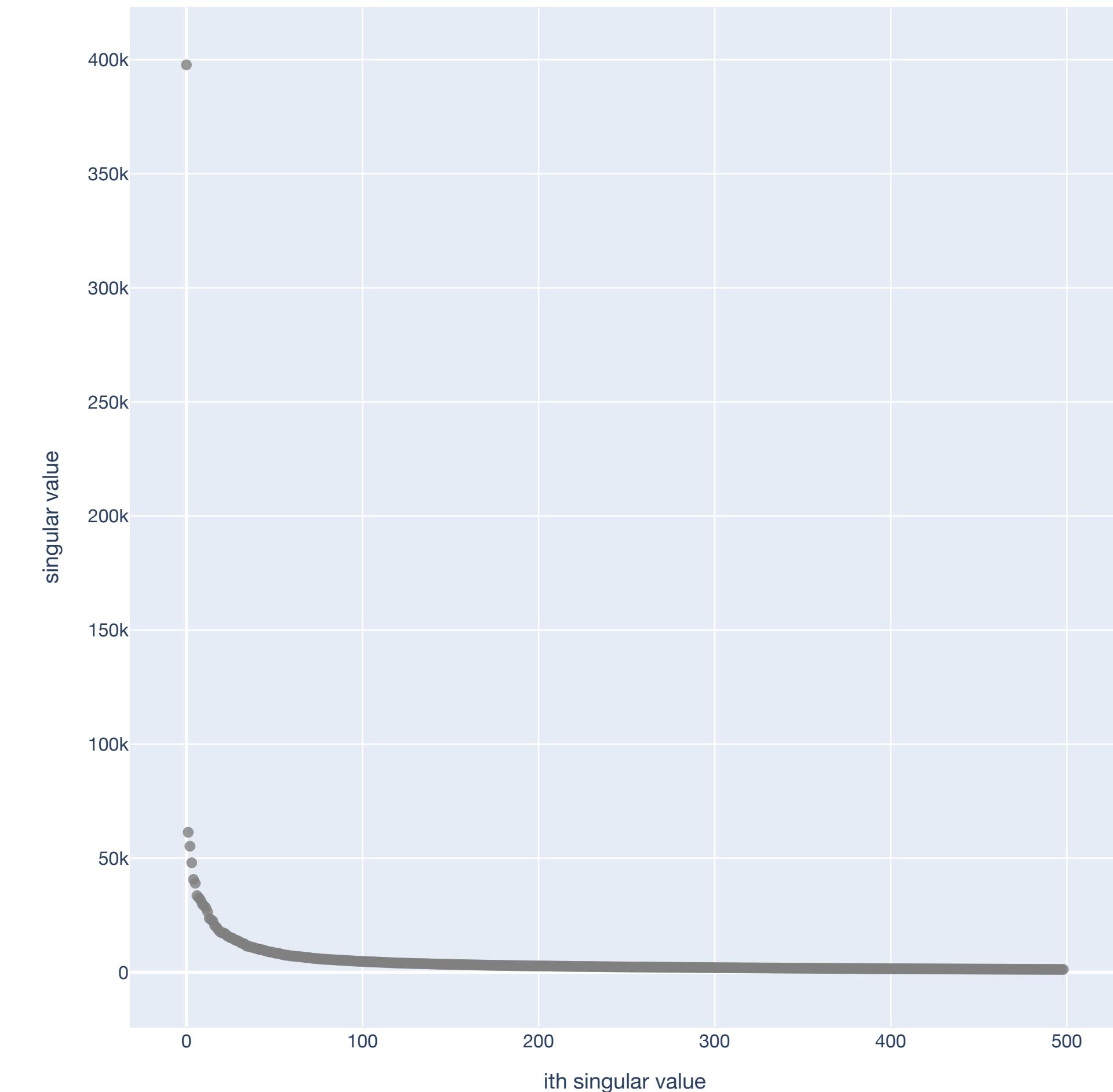
Rank- k Approximation

Application in Image Processing



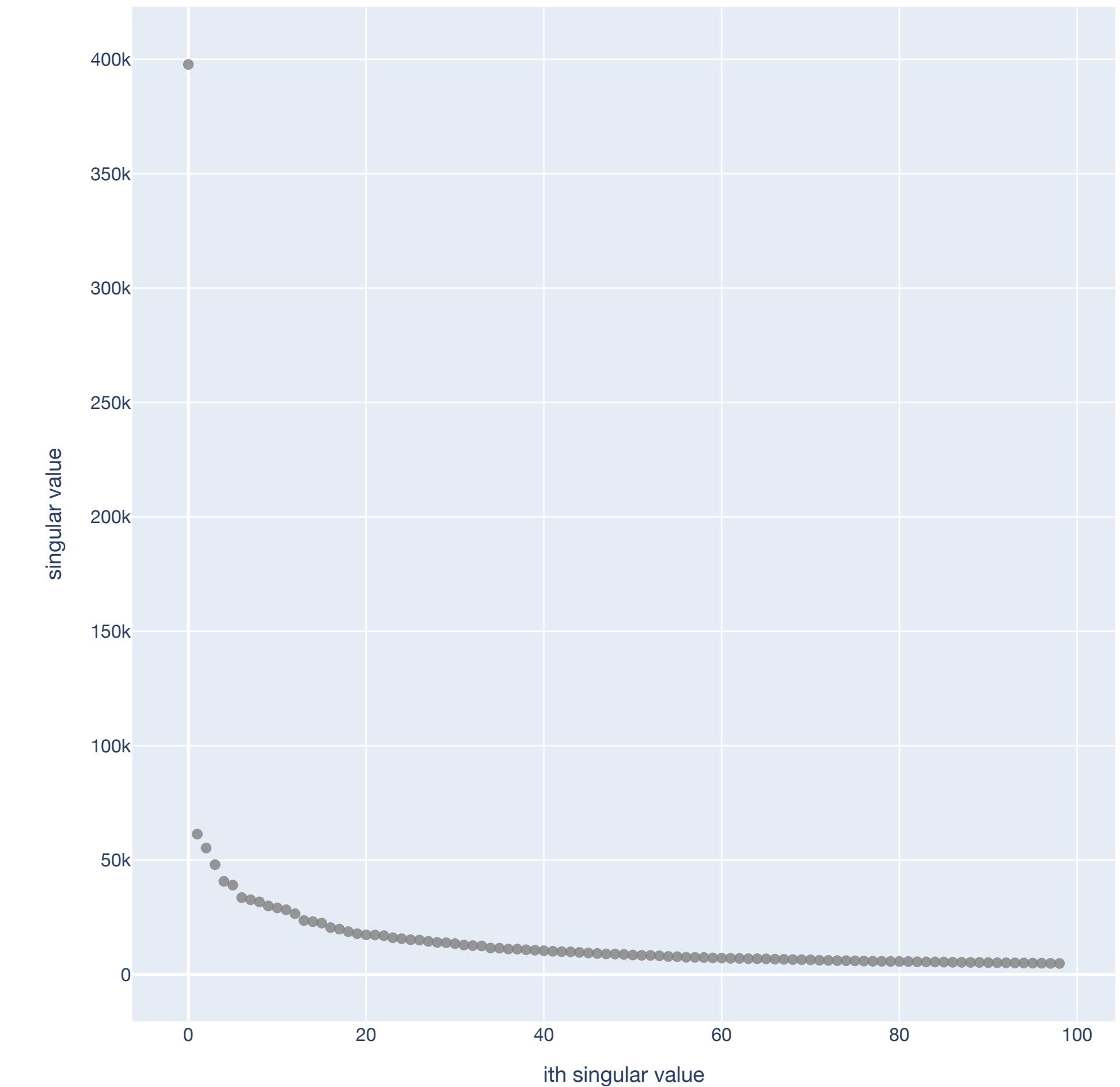
Rank- k Approximation

Application in Image Processing ($k = 500$)



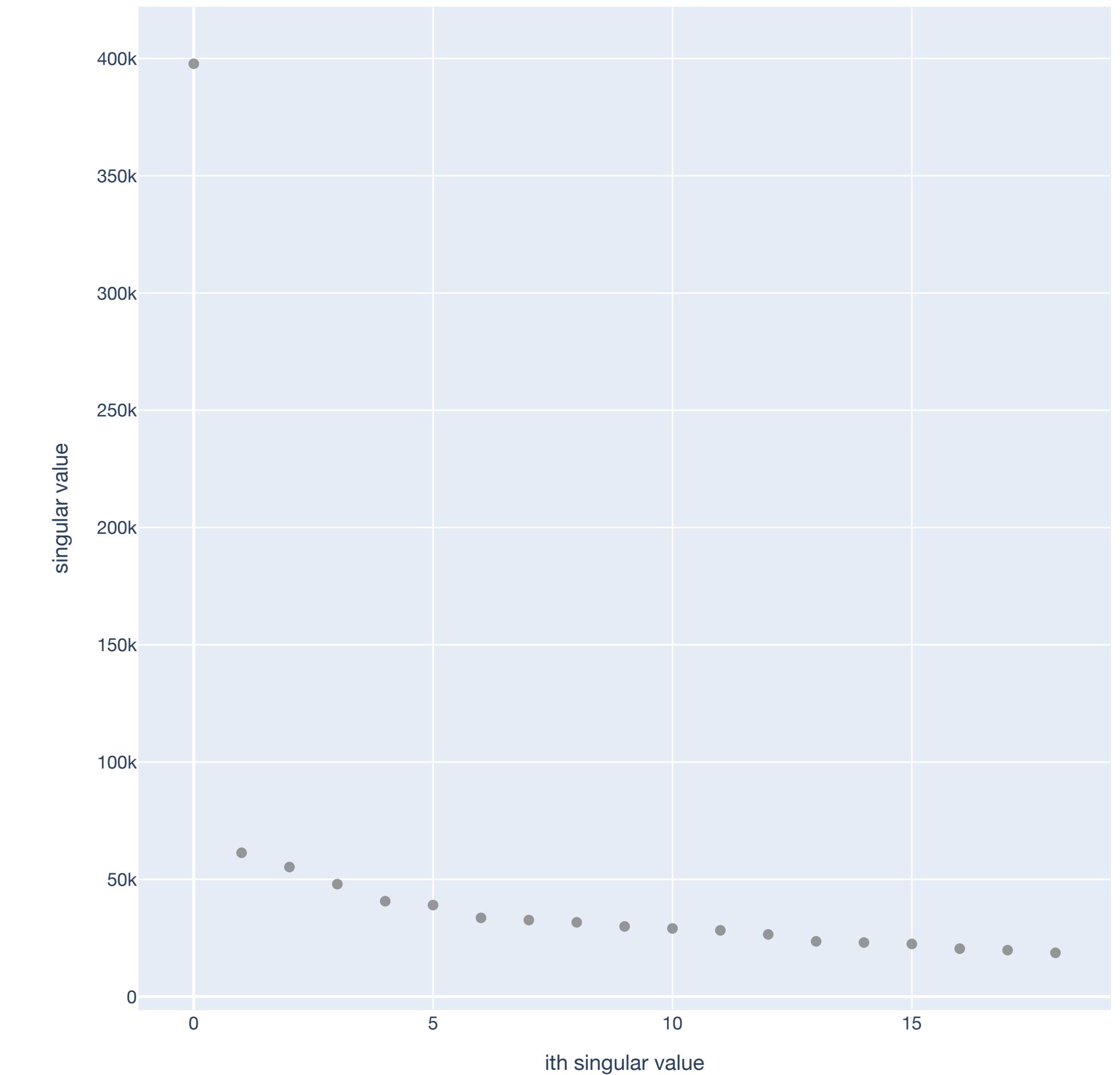
Rank- k Approximation

Application in Image Processing ($k = 100$)



Rank- k Approximation

Application in Image Processing ($k = 20$)



Least Squares

SVD and the Pseudoinverse

Regression Setup

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

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

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

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

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

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

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 $\text{rank}(\mathbf{X}) = d$, 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

Plugging in the SVD

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

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

Least Squares: SVD Perspective

Plugging in the SVD

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

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

Least Squares: SVD Perspective

Plugging in the SVD

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

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

Least Squares: SVD Perspective

Plugging in the SVD

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

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

Least Squares: SVD Perspective

Plugging in the SVD

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

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

Least Squares: SVD Perspective

Plugging in the SVD

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

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

Pseudoinverse

Idea

Therefore, we derived:

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

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

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

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

Pseudoinverse

Definition

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

If $n \geq d$, the matrix $(\Sigma^\top\Sigma)^{-1}\Sigma^\top \in \mathbb{R}^{d \times n}$ is the [**\(Moore-Penrose\) pseudoinverse**](#) of the matrix Σ , denoted $\Sigma^+ := (\Sigma^\top\Sigma)^{-1}\Sigma^\top$.

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

More generally, the matrix $\mathbf{X} \in \mathbb{R}^{n \times d}$ with full SVD $\mathbf{X} = \mathbf{U}\Sigma\mathbf{V}^\top$ has the [**\(Moore-Penrose\) pseudoinverse**](#): $\mathbf{X}^+ := \mathbf{V}\Sigma^+\mathbf{U}^\top$.

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 $\mathbf{A} = \mathbf{U}\Sigma\mathbf{V}^\top$ and $\text{rank}(\mathbf{A}) = \min\{n, d\}$, the pseudo inverse

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

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*: $\mathbf{A}\mathbf{A}^+ = \mathbf{I}_{n \times n}$.

Pseudoinverse

Shape of Σ^+

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

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

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

Pseudoinverse

Shape of Σ^+

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

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

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

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:

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

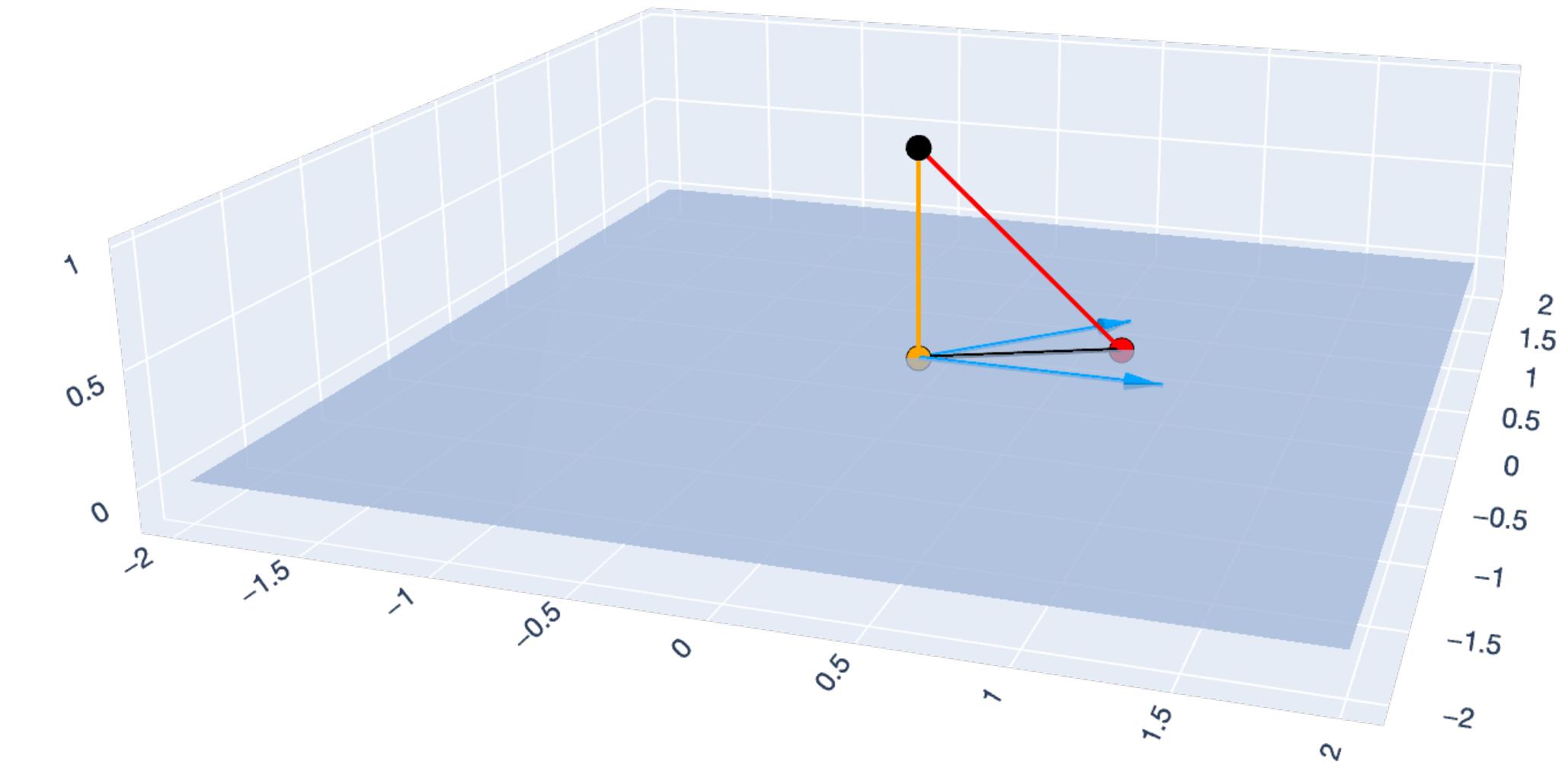
Theorem (Ordinary Least Squares).

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

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:

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

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*:

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

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:

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

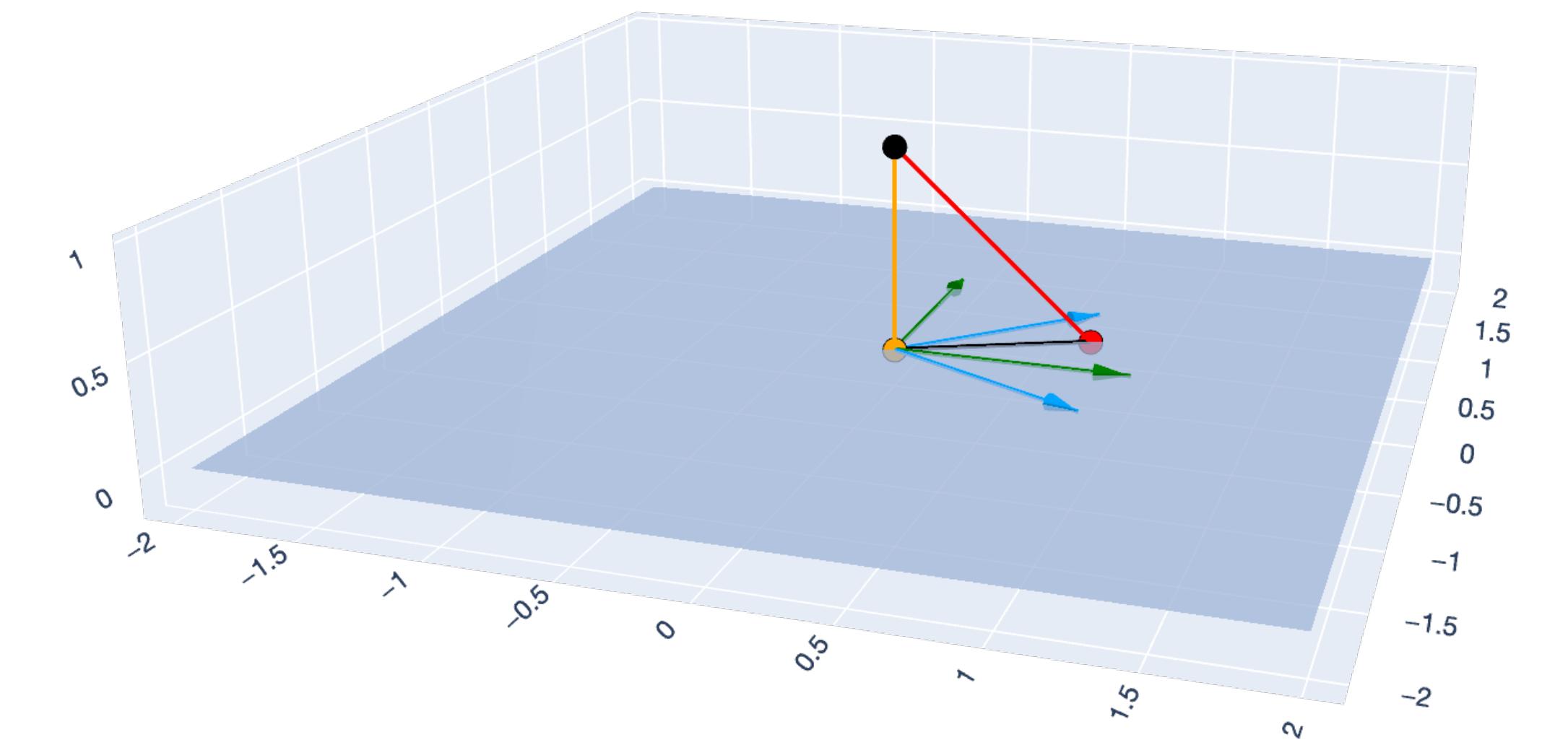
Theorem (OLS with pseudoinverse).

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

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

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

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



x1 x2 u1 u2 y - $\hat{\mathbf{y}}$ $\mathbf{y} - \hat{\mathbf{y}}$ $\mathbf{y} - \tilde{\mathbf{y}}$ \mathbf{y} $\hat{\mathbf{y}}$ $\tilde{\mathbf{y}}$

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 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:

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

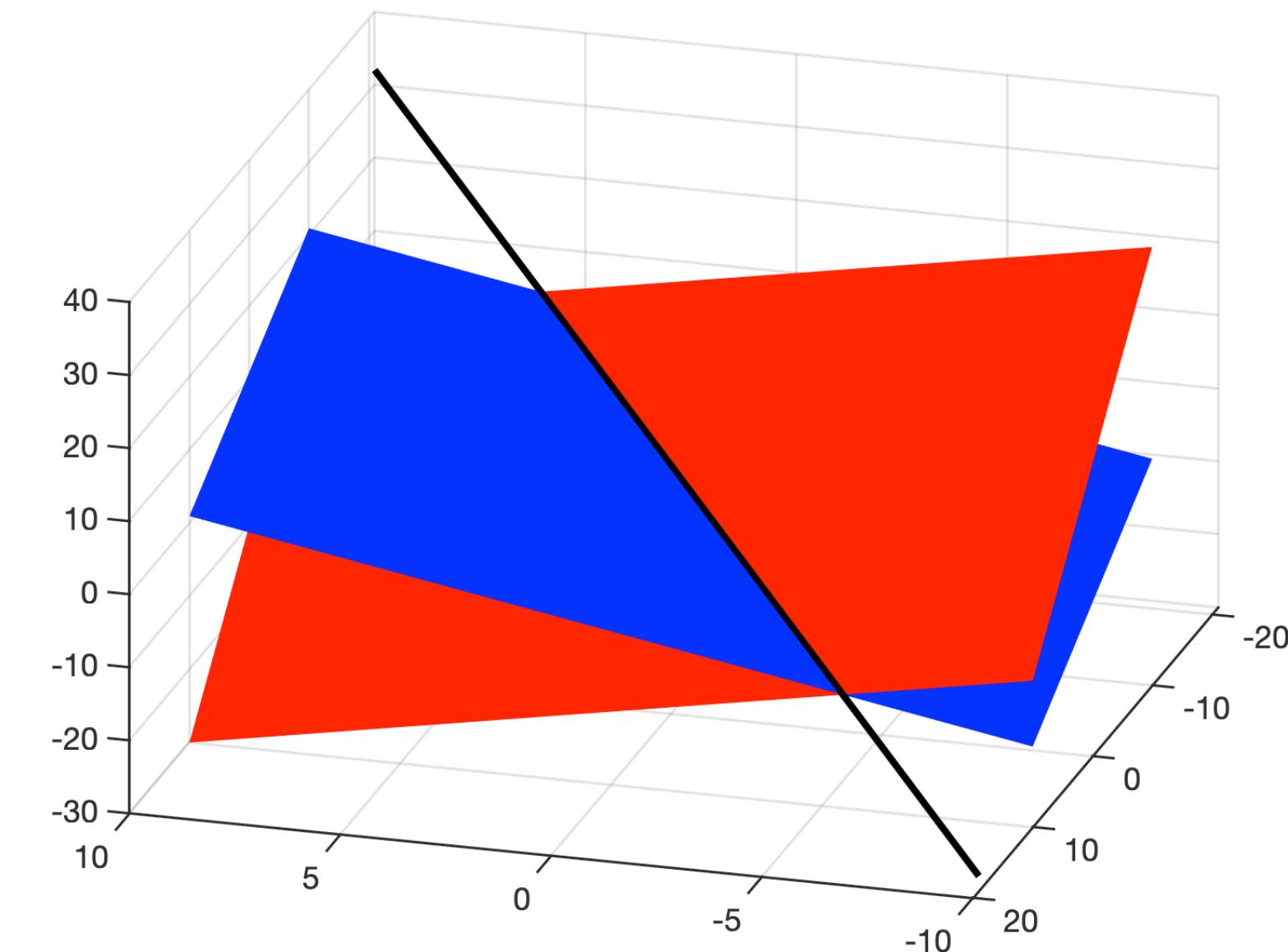
Least Squares with $d \geq n$

Review: Systems of Linear Equations

When the number of equations < number of unknowns...

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

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



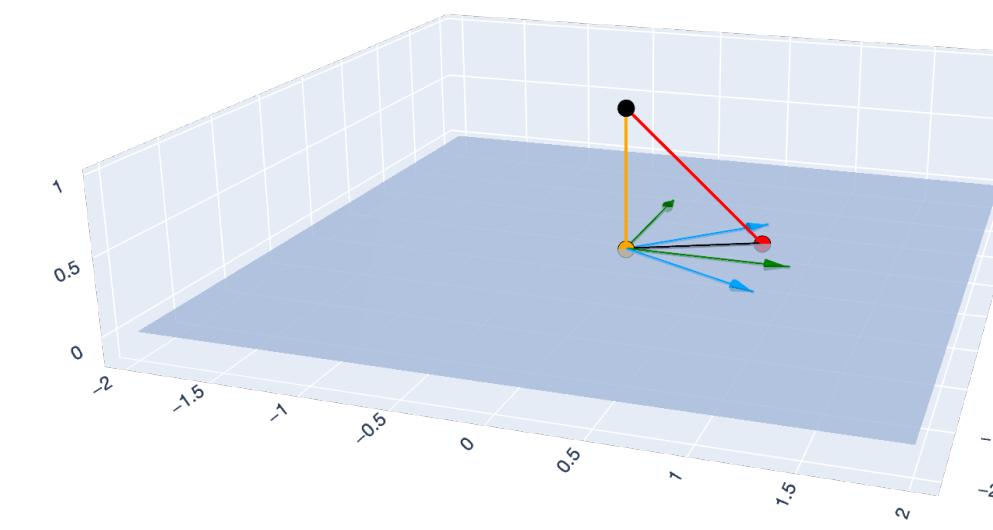
Least Squares with $d \geq n$

Problem Statement

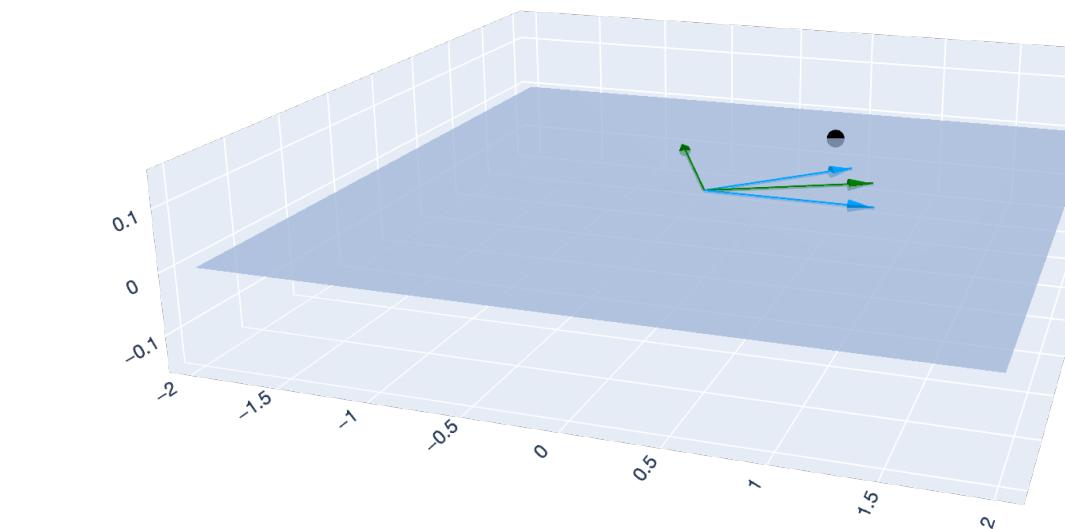
Let $\mathbf{X} \in \mathbb{R}^{n \times d}$, let $d \geq n$, and let $\text{rank}(\mathbf{X}) = n$. We want to solve the system of linear equations:

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

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



x1 x2 u1 u2 y - \hat{y} $\sim y - \hat{y}$ $\sim y - y$ \bullet y \circ \hat{y} \bullet $\sim y$

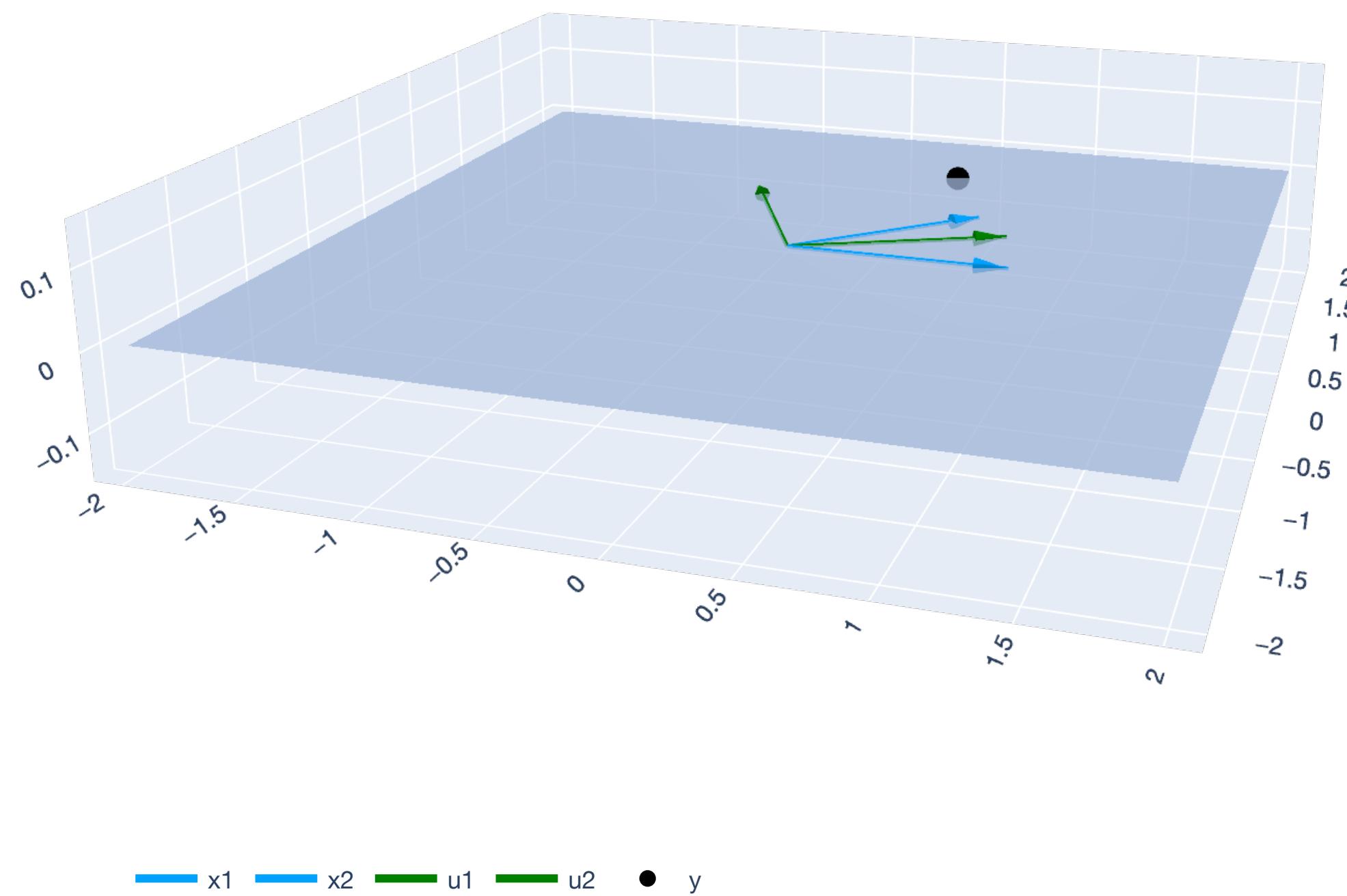


x1 x2 u1 u2 \bullet y

Least Squares with $d \geq n$

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

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

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

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

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

Choose $\hat{\mathbf{w}} = \mathbf{X}^+\mathbf{y} = \mathbf{V}\Sigma^+\mathbf{U}^\top\mathbf{y}$ 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 $\text{rank}(\mathbf{X}) = n$. 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:

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

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

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:

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

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

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

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 $\text{rank}(\mathbf{X}) = n$. Then, $\hat{\mathbf{w}} = \mathbf{X}^+ \mathbf{y} = \mathbf{V} \boldsymbol{\Sigma}^+ \mathbf{U}^\top \mathbf{y}$ is the exact solution (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

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 (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 $\text{rank}(\mathbf{X}) = d\dots$

We can solve exactly.

Choose

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

which is an exact solution.

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

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 $\text{rank}(\mathbf{X}) = n\dots$

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

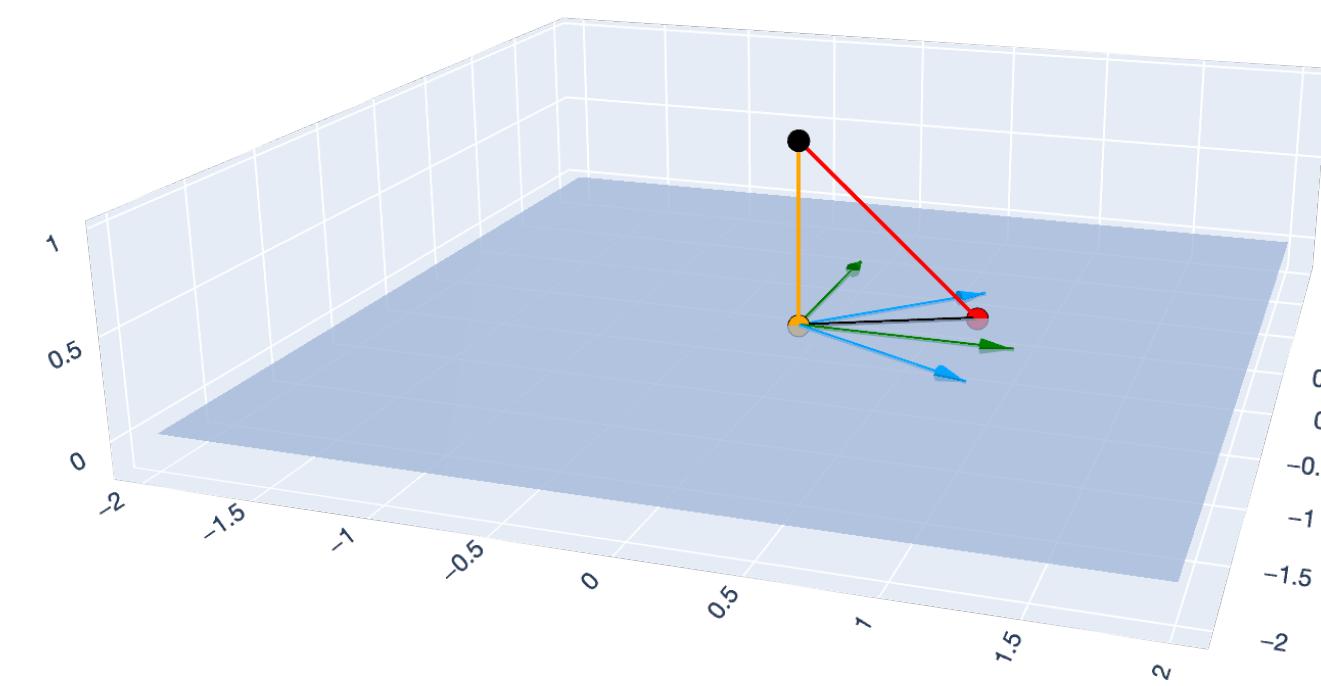
If $n > d$ and $\text{rank}(\mathbf{X}) = d$...

We approximate by least squares:

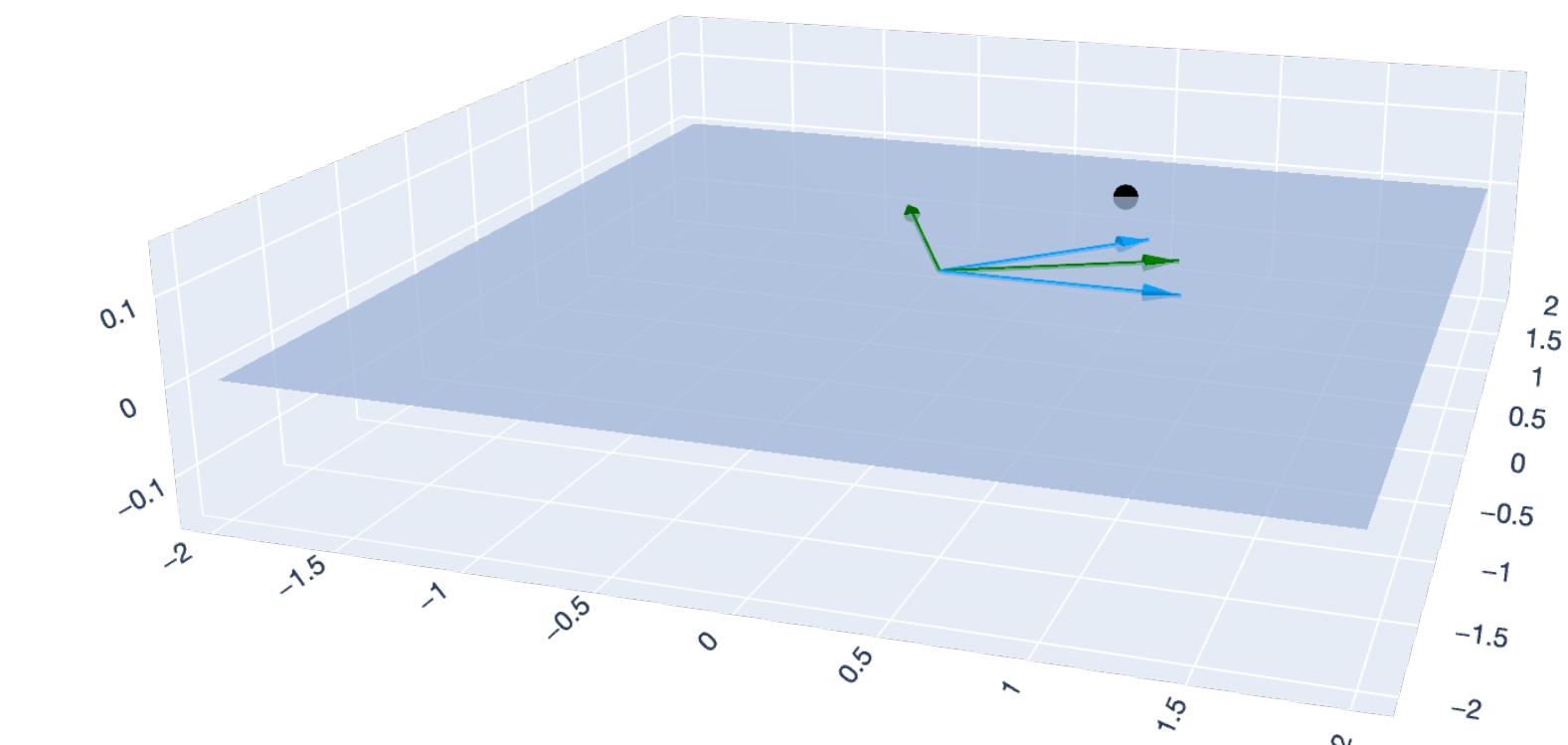
$$\hat{\mathbf{w}} = \arg \min_{\mathbf{w} \in \mathbb{R}^d} \|\mathbf{X}\mathbf{w} - \mathbf{y}\|^2.$$

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

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



x1 x2 u1 u2 y - \hat{y} - $y - \hat{y}$ $\sim y - \hat{y}$ $\sim y$ \hat{y} $\sim y$

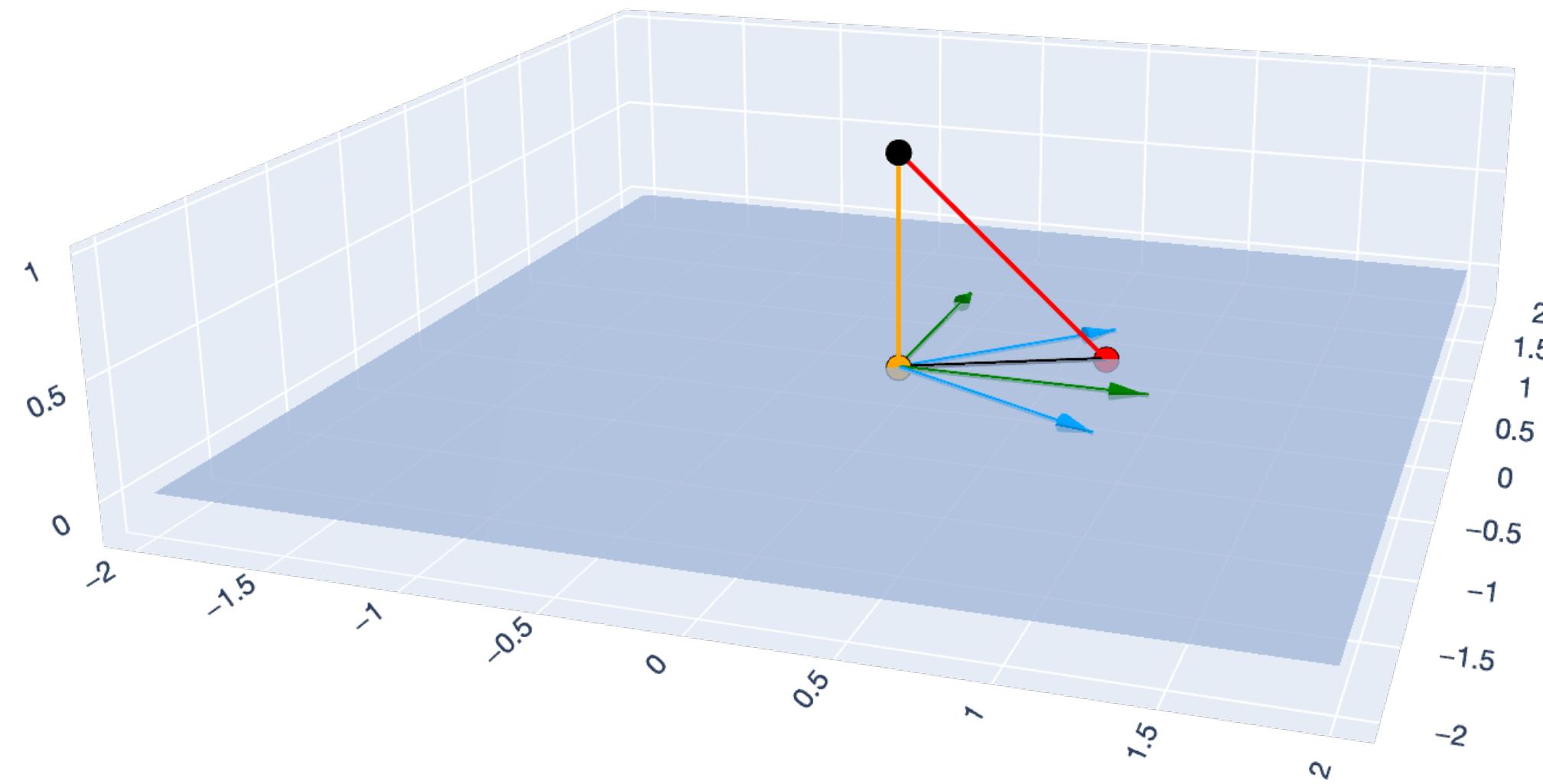


x1 x2 u1 u2 $\sim y$

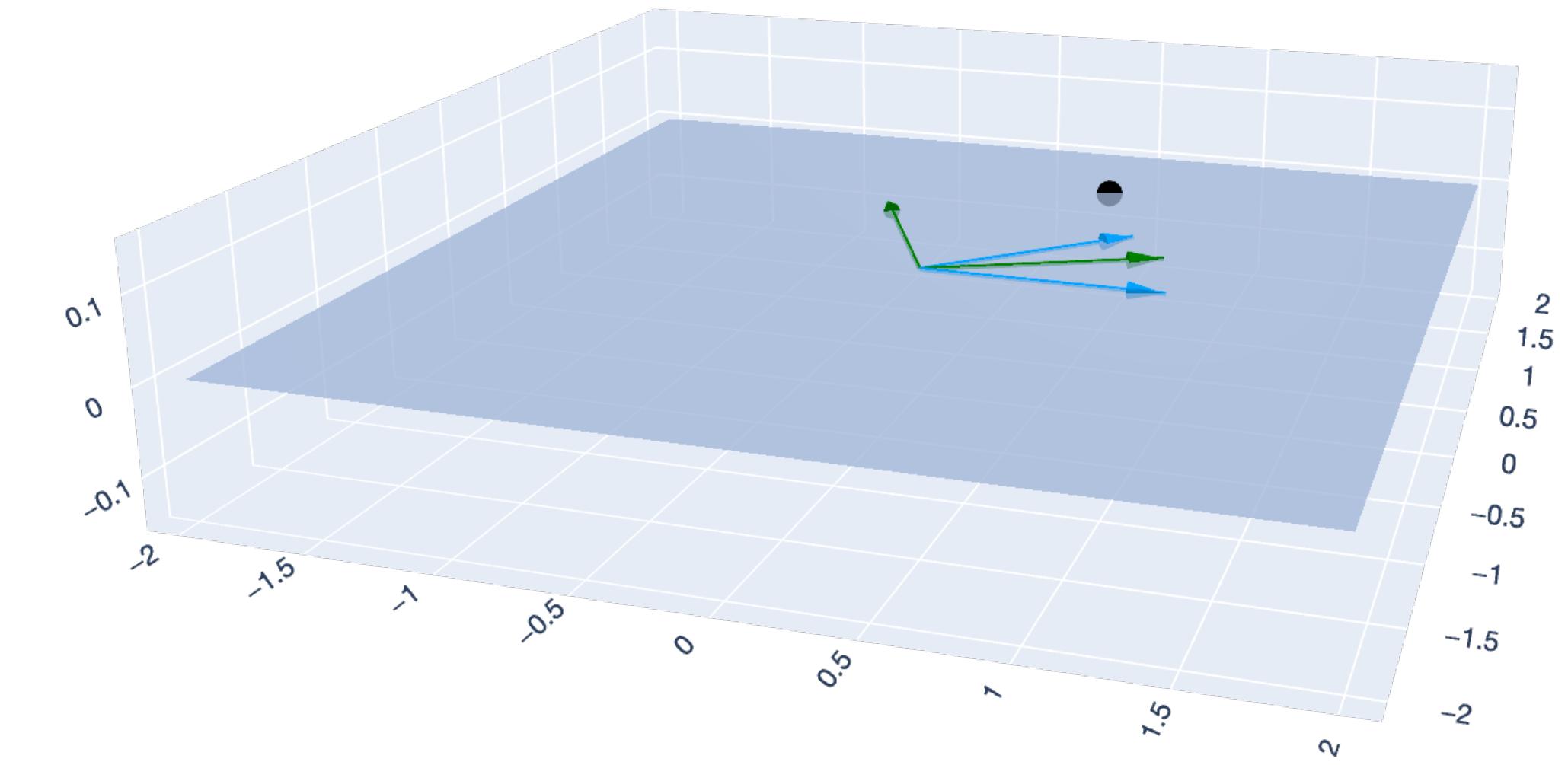
Recap

Lesson Overview

Big Picture: Least Squares



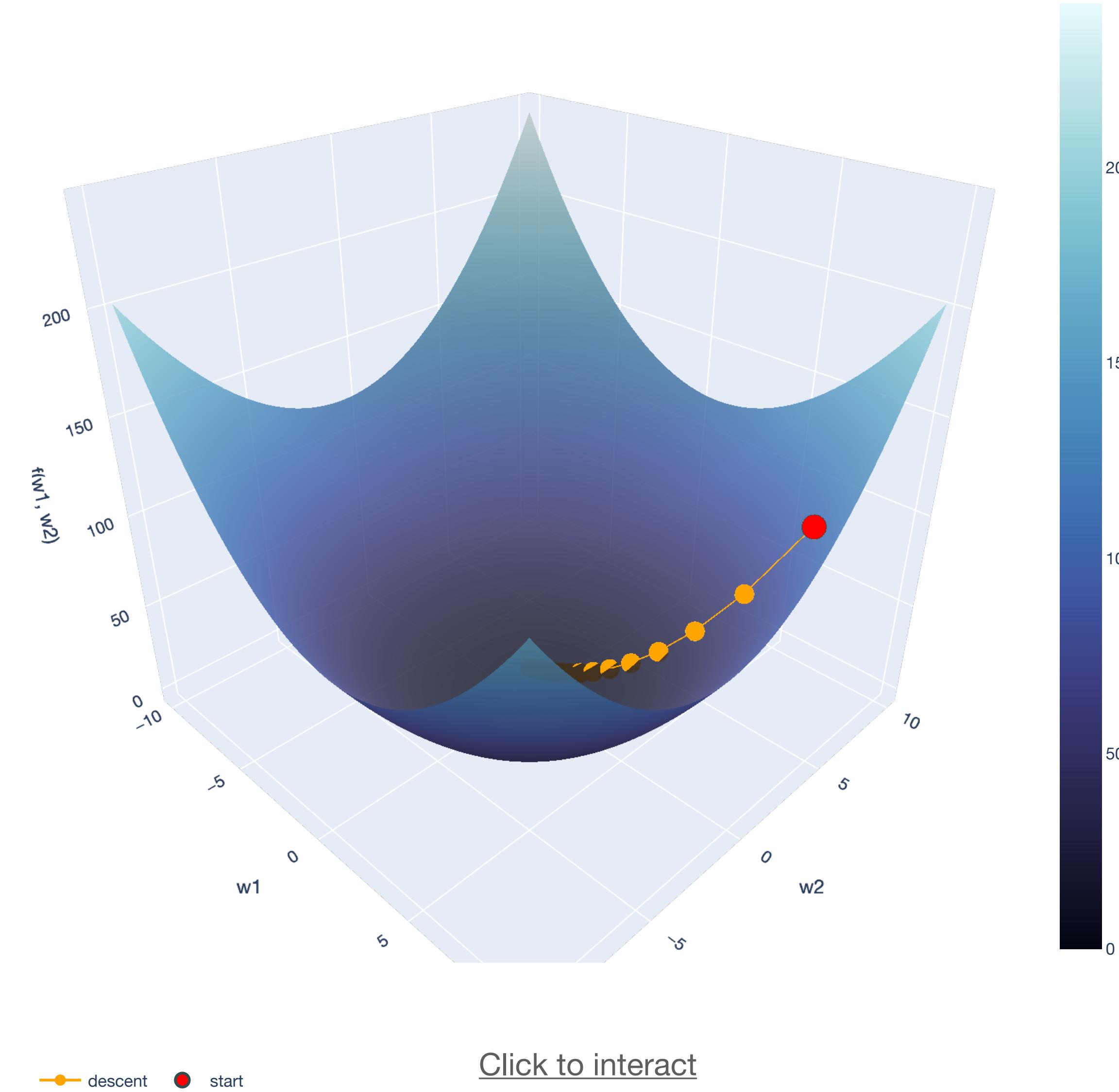
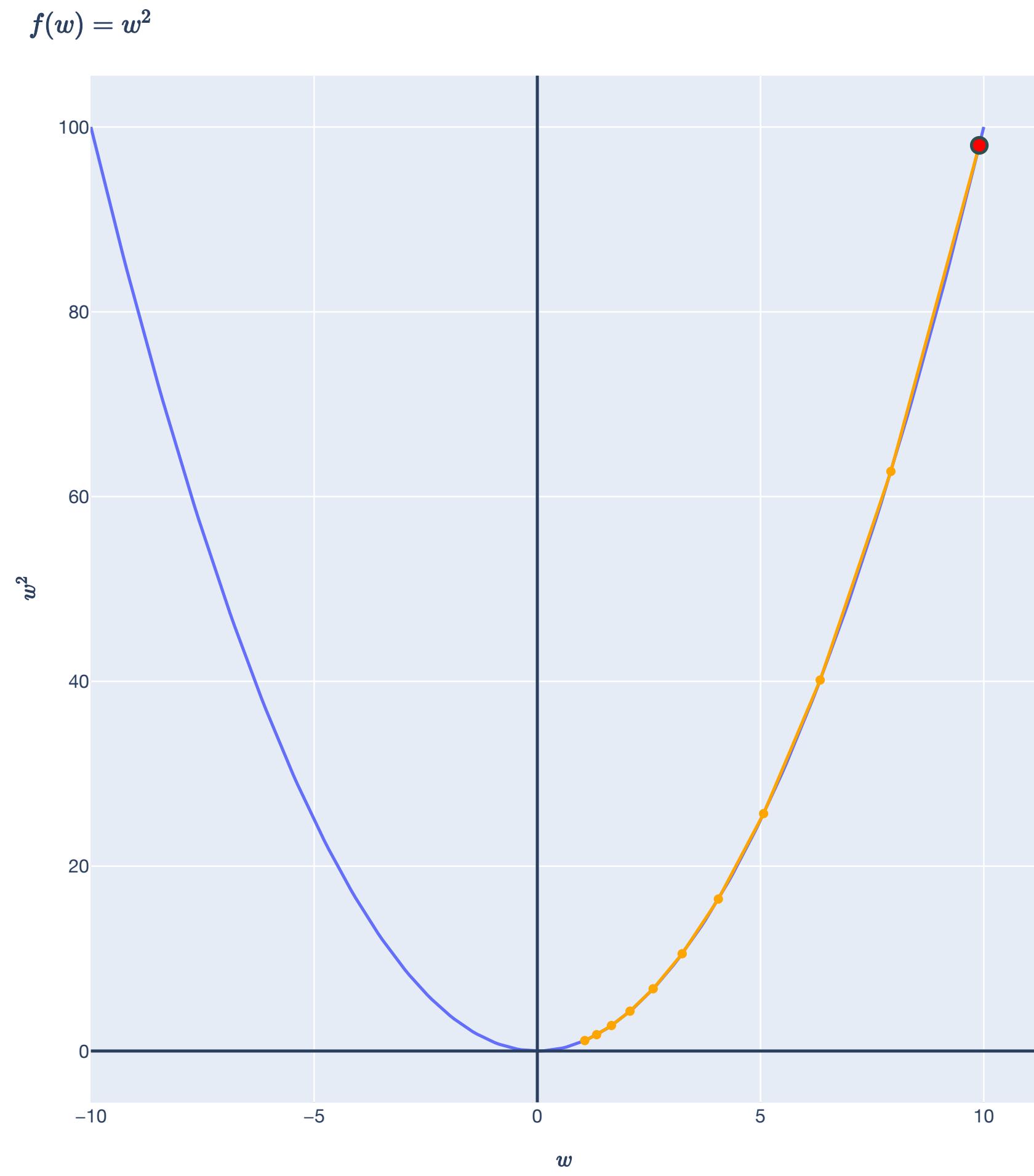
— x_1 — x_2 — u_1 — u_2 — $y - \hat{y}$ — $\hat{y} - y$ ● y ○ \hat{y} ● $\sim y$



— x_1 — x_2 — u_1 — u_2 ● y

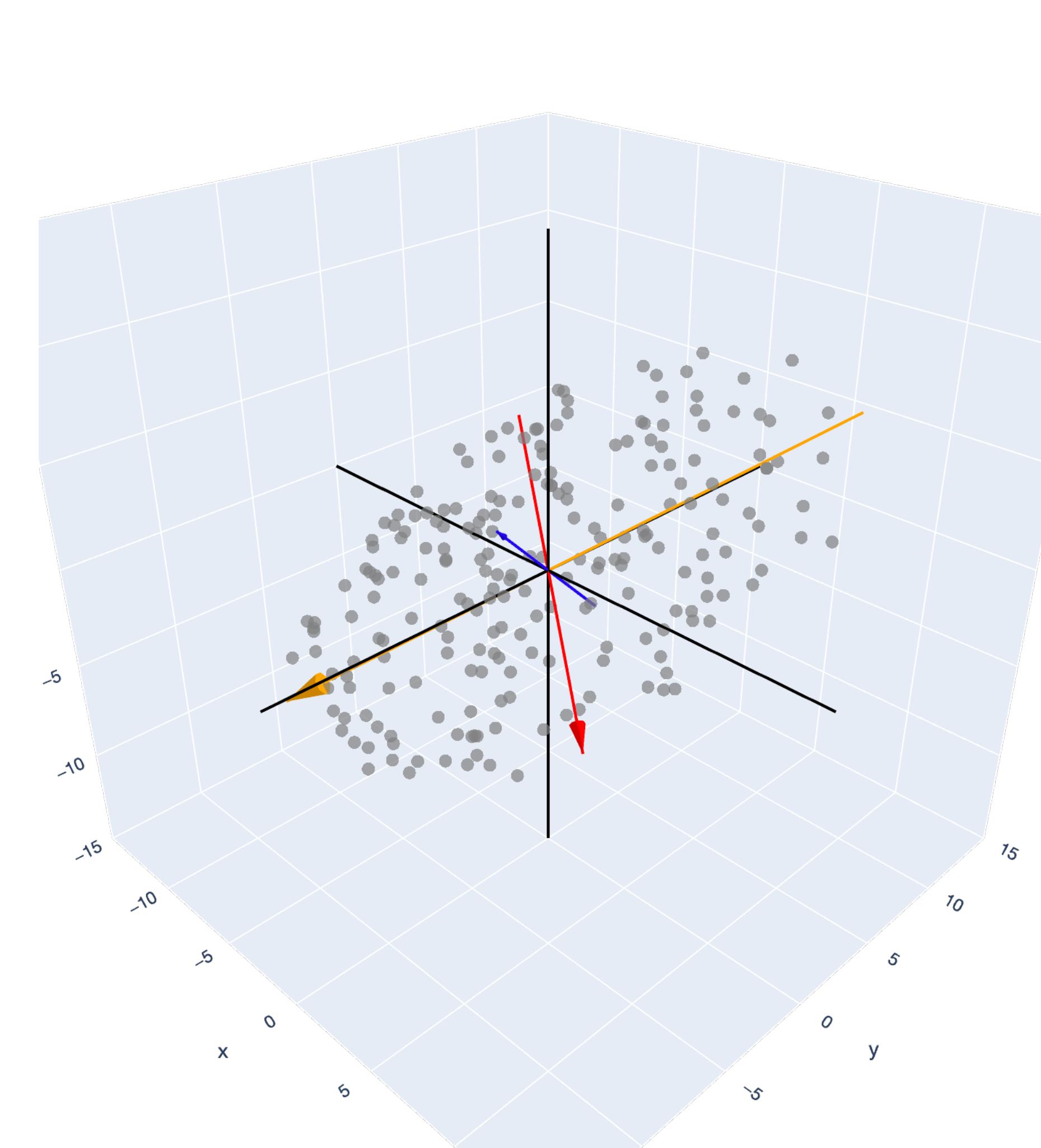
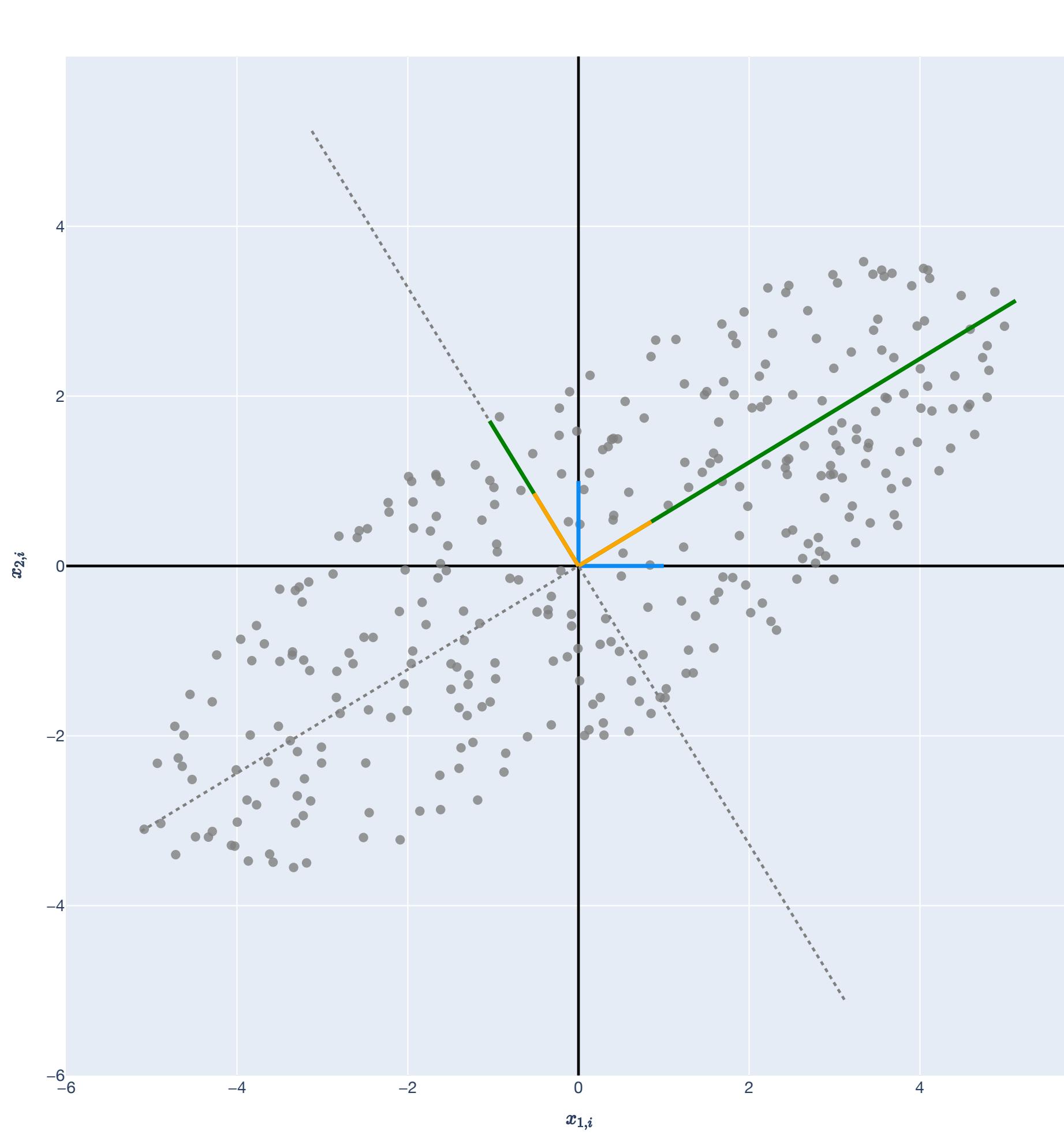
Lesson Overview

Big Picture: Gradient Descent



Lesson Overview

Big Picture: Singular Value Decomposition (SVD)



- x1-axis
- x2-axis
- x3-axis
- u1
- u2
- u3

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.