Math for Machine Learning

Week 2.2: Eigendecomposition and PSD Matrices

By: Samuel Deng

Logistics & Announcements

Lesson Overview

Linear dynamical systems example. Motivation for eigendecomposition as a way to make repeated matrix multiplication easier.

Eigendecomposition. Definition of eigenvectors, eigenvalues.

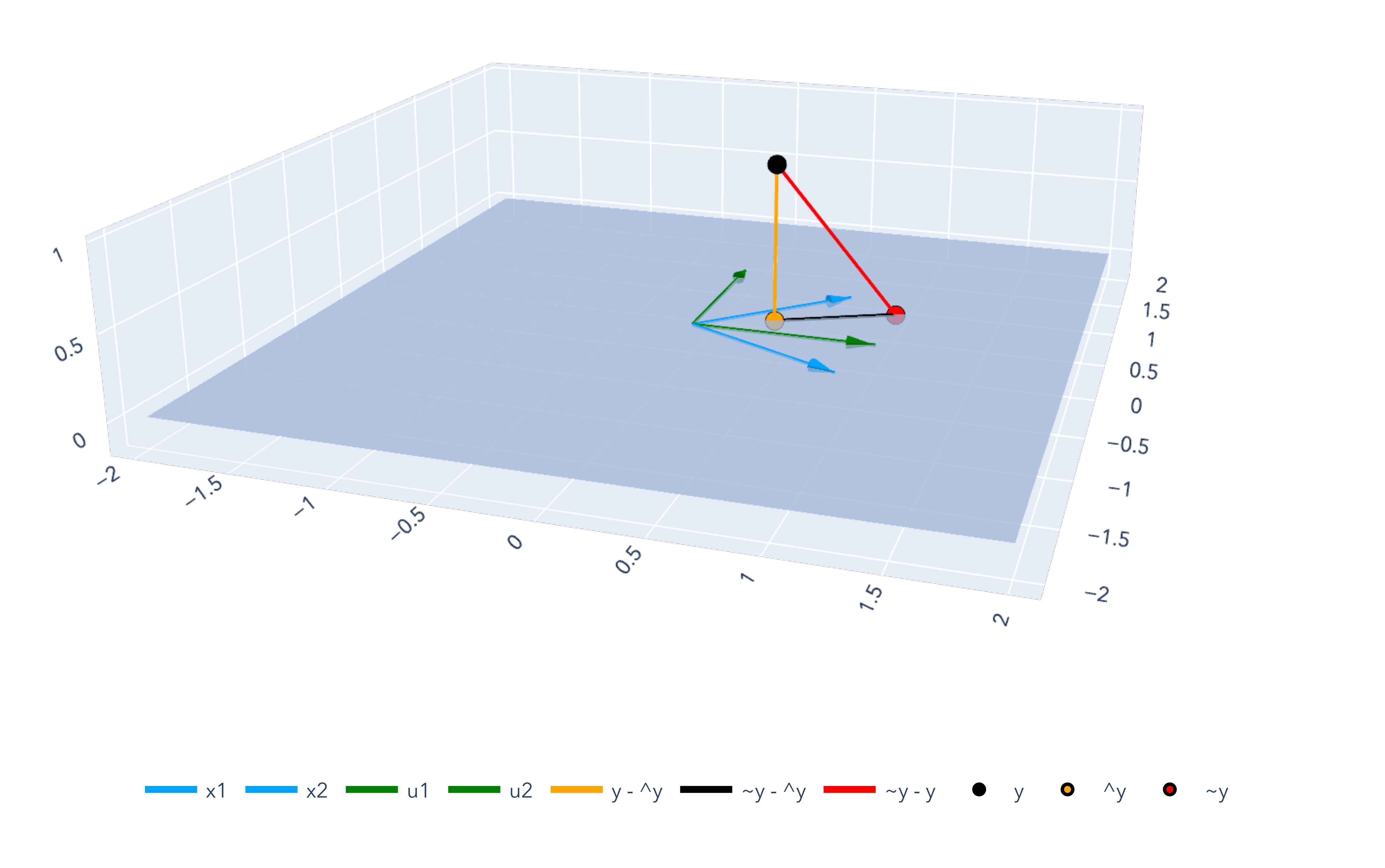
Eigendecomposition and SVD. The eigendecomposition drops out of the SVD.

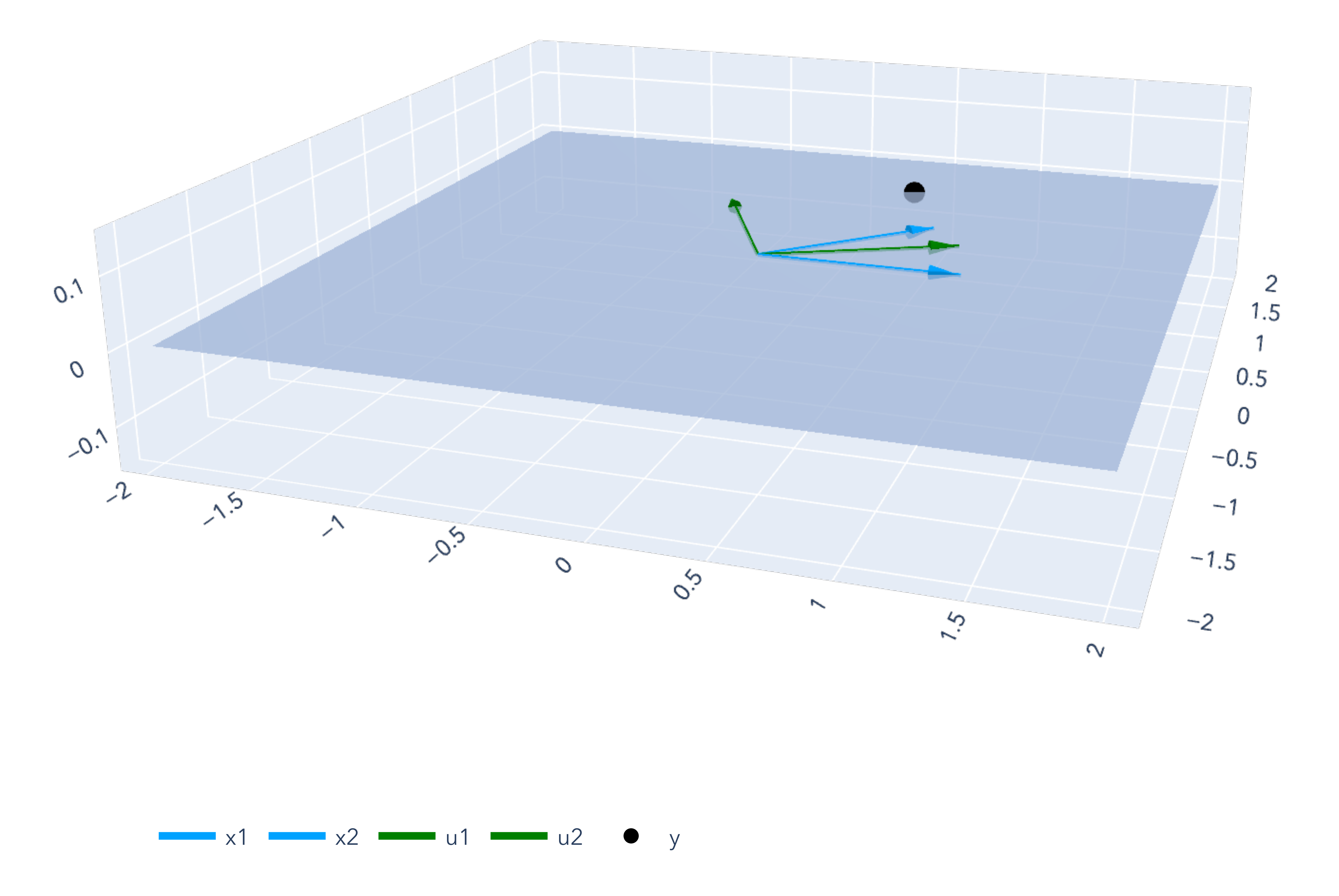
Spectral Theorem. Symmetric matrices are always diagonalizable.

Positive semidefinite matrices/positive definite matrices. Definition and some visual examples through the corresponding quadratic forms.

Lesson Overview

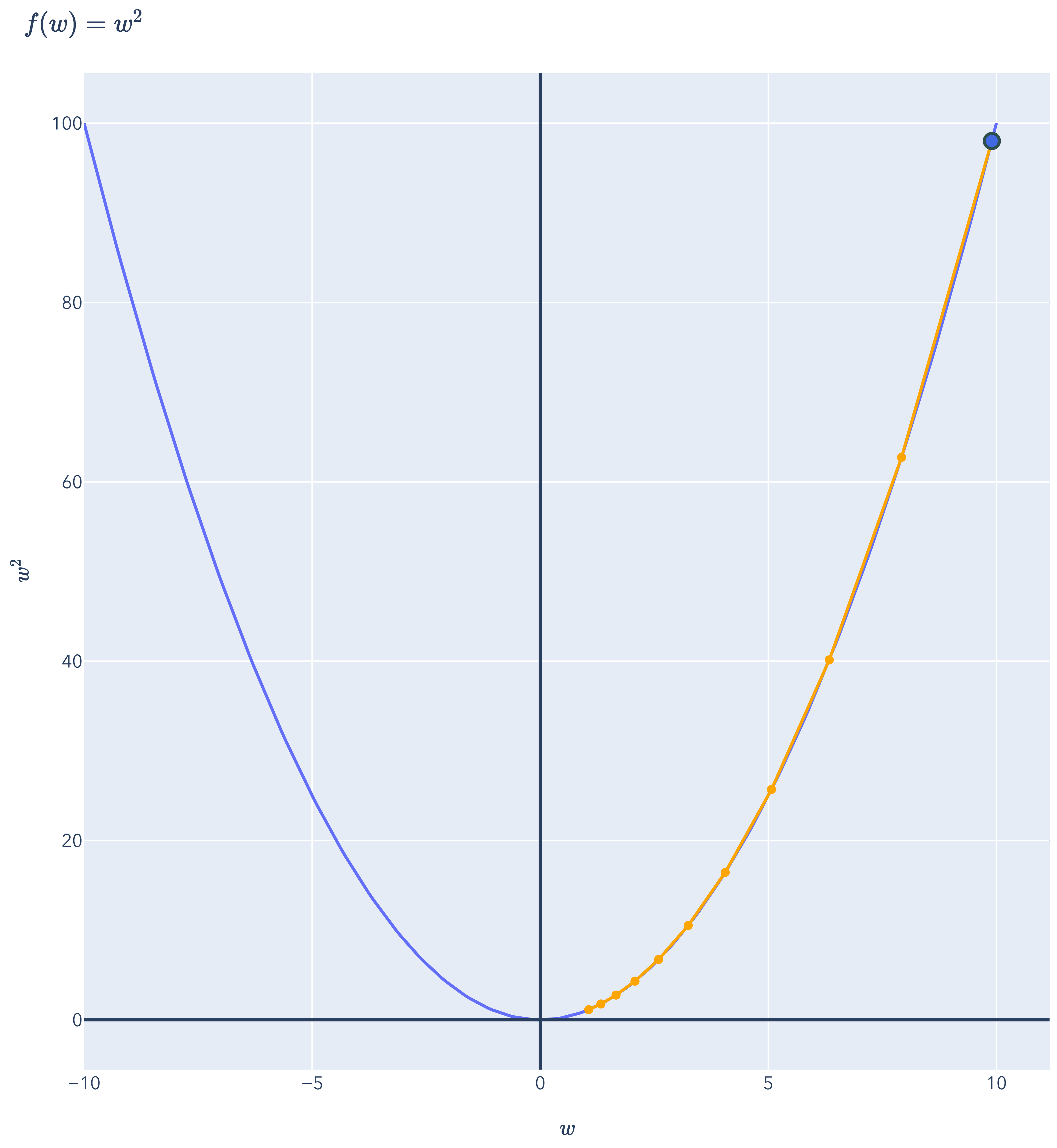
Big Picture: Least Squares

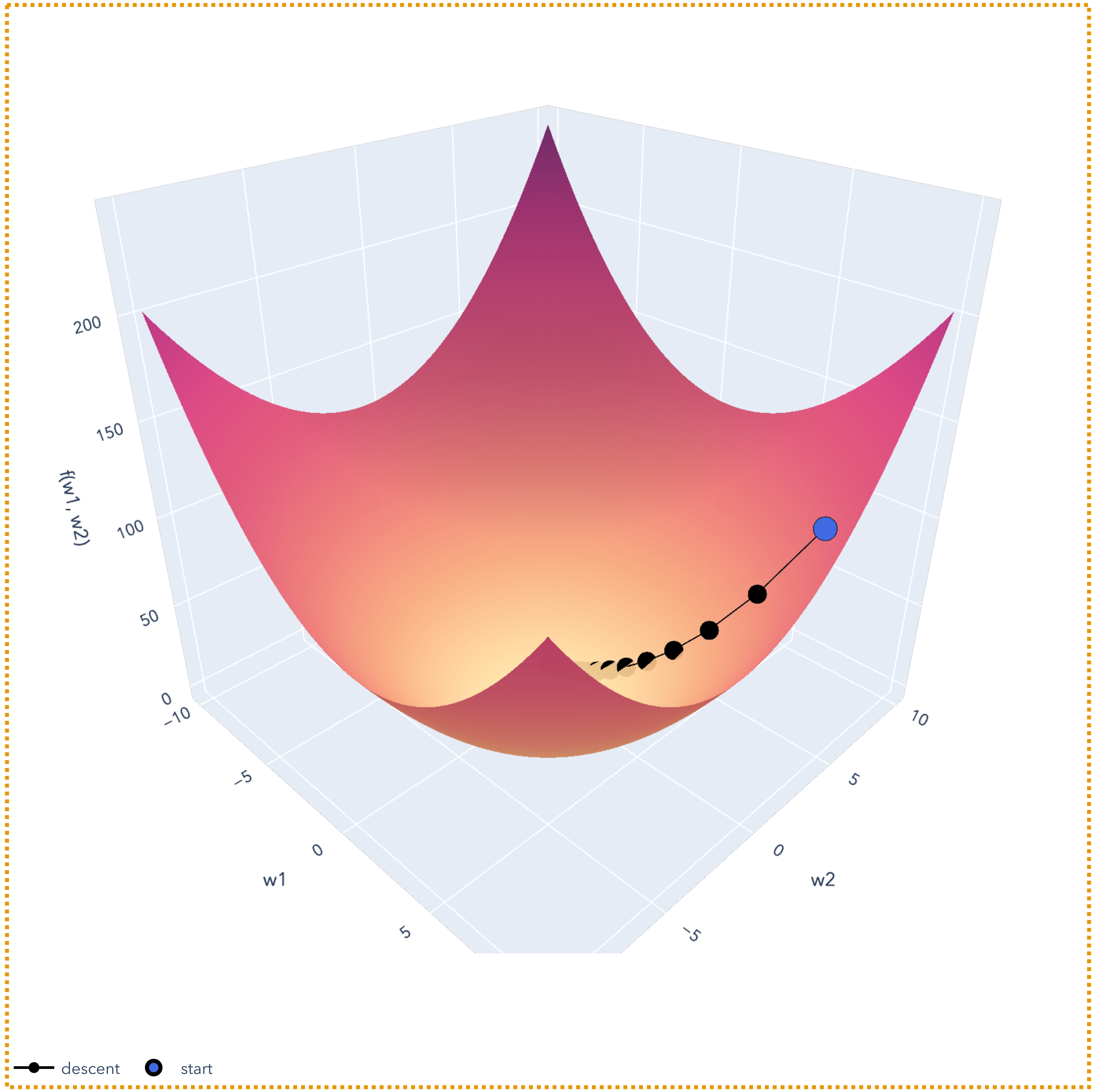




Lesson Overview

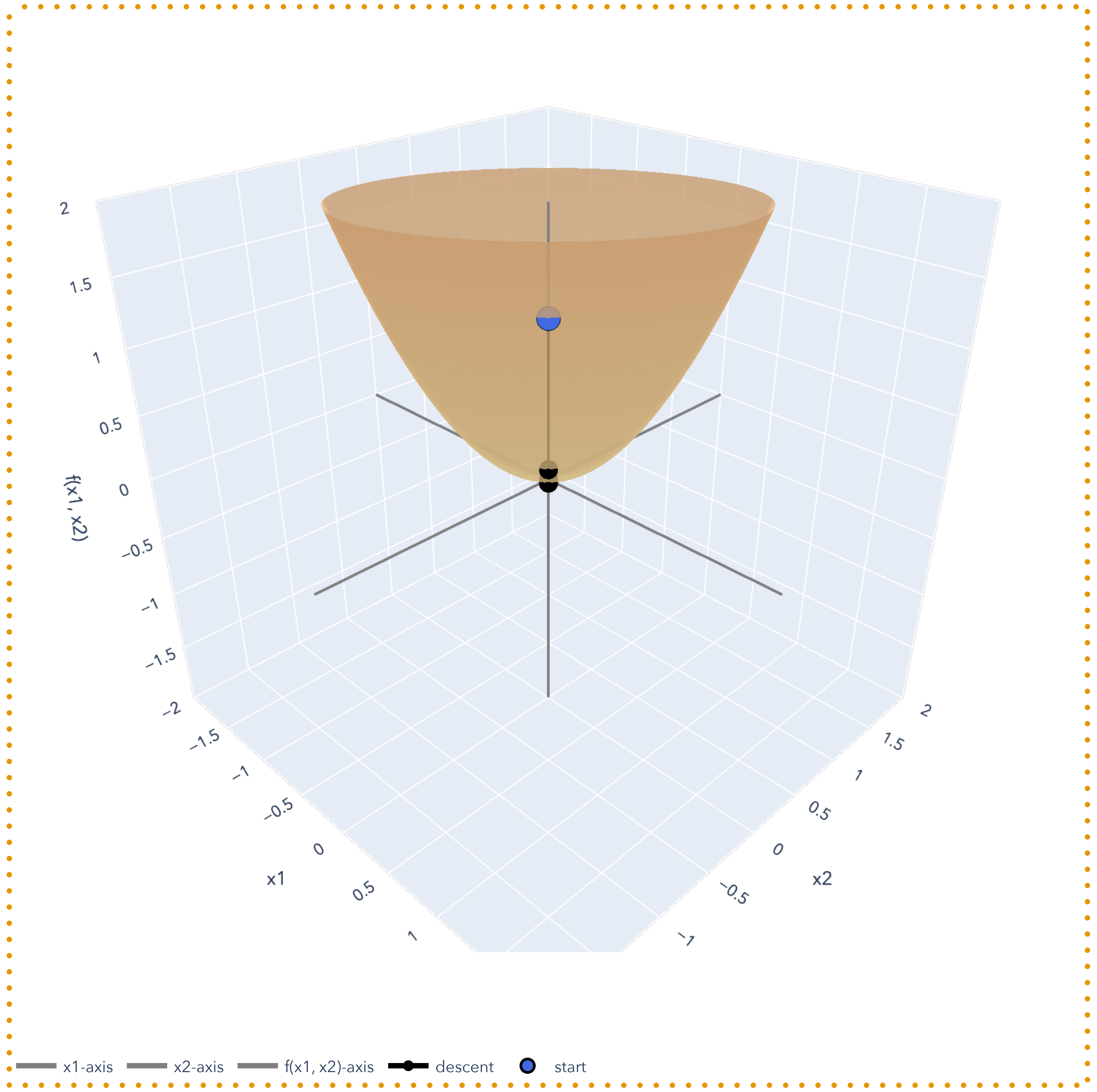
Big Picture: Gradient Descent

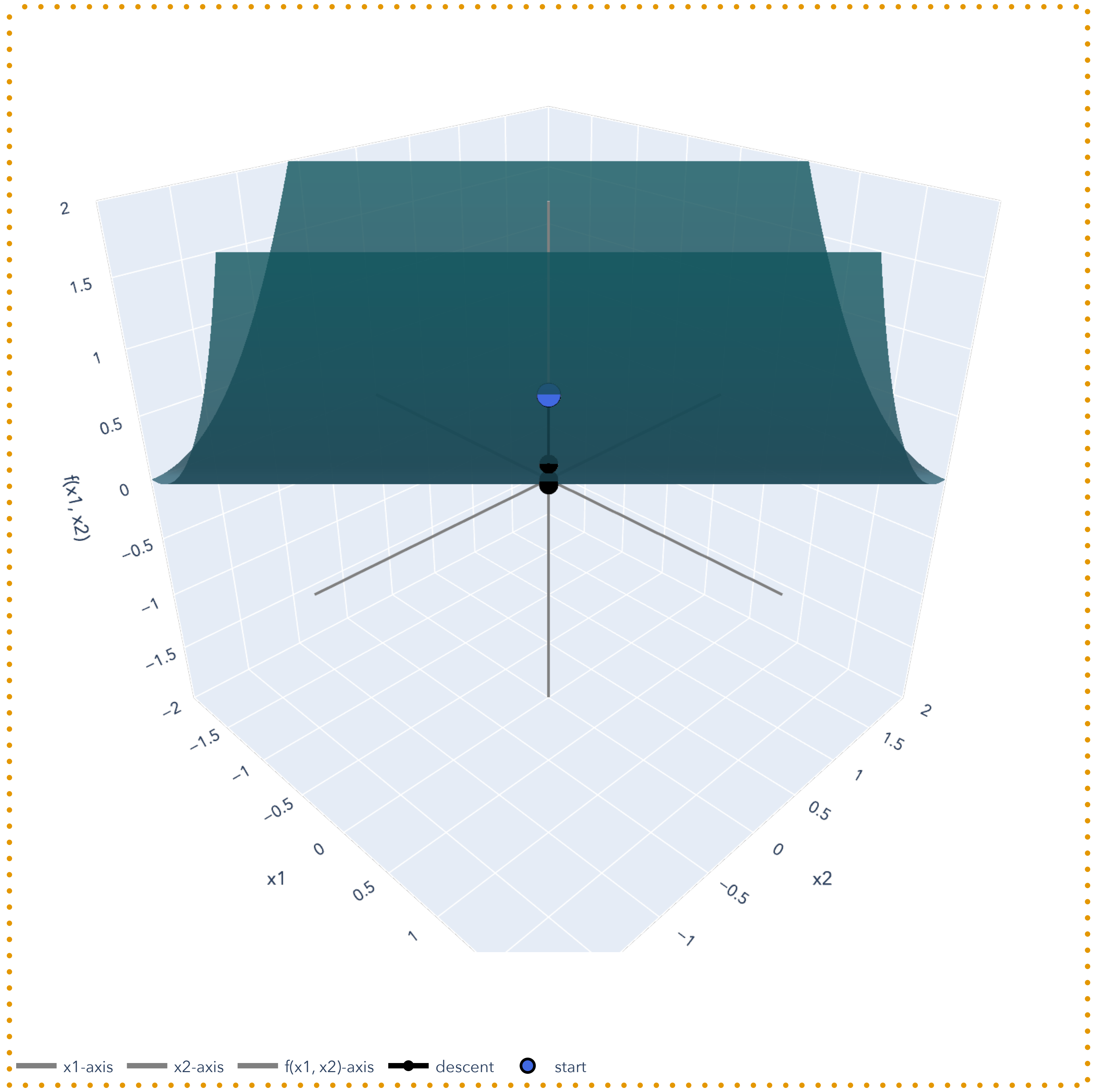


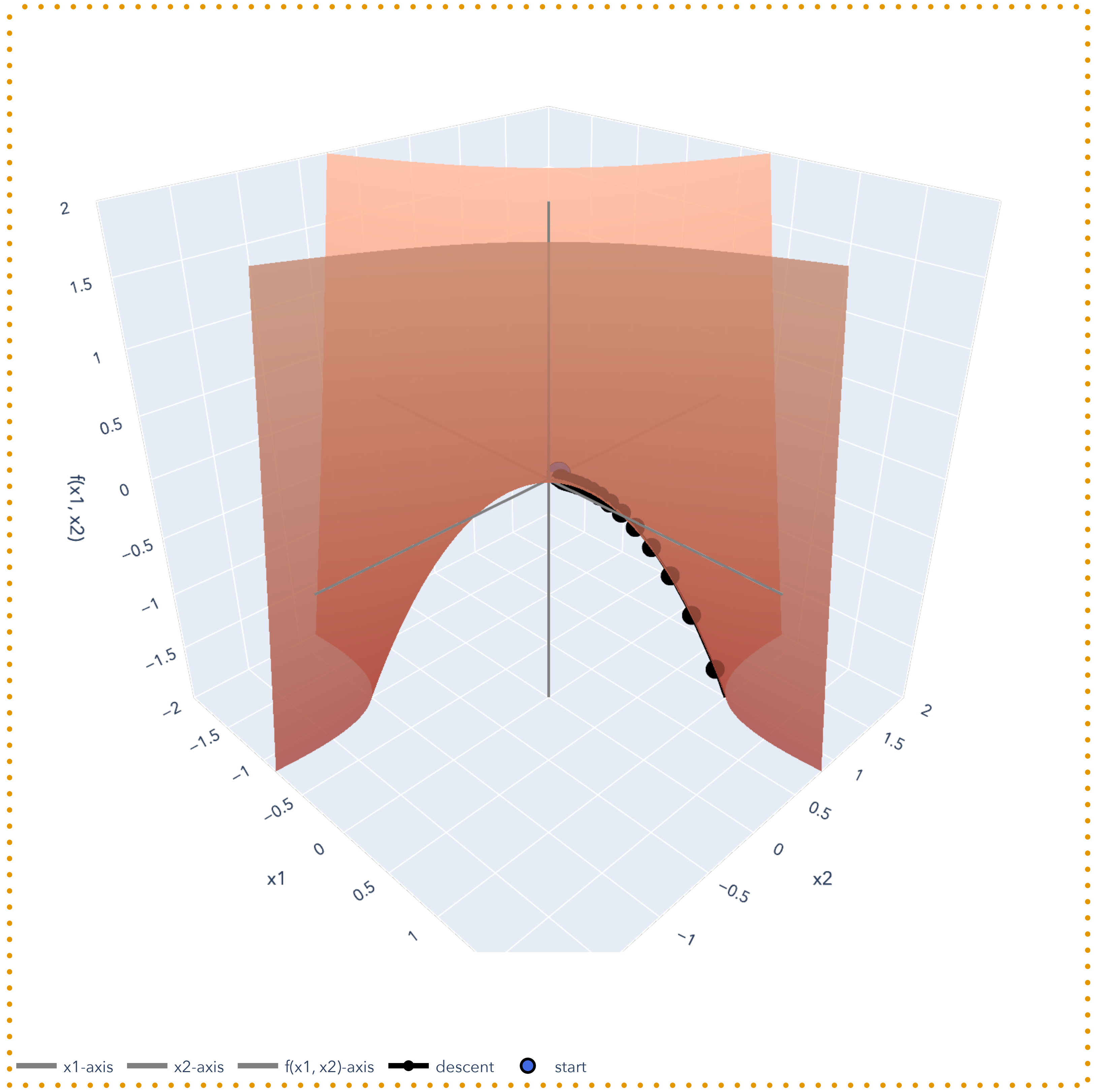


Lesson Overview

Big Picture: Gradient Descent







Least Squares

A Quick Review

Regression

Setup (Example View)

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

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

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 (Feature View)

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

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

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

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

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

Regression

Setup

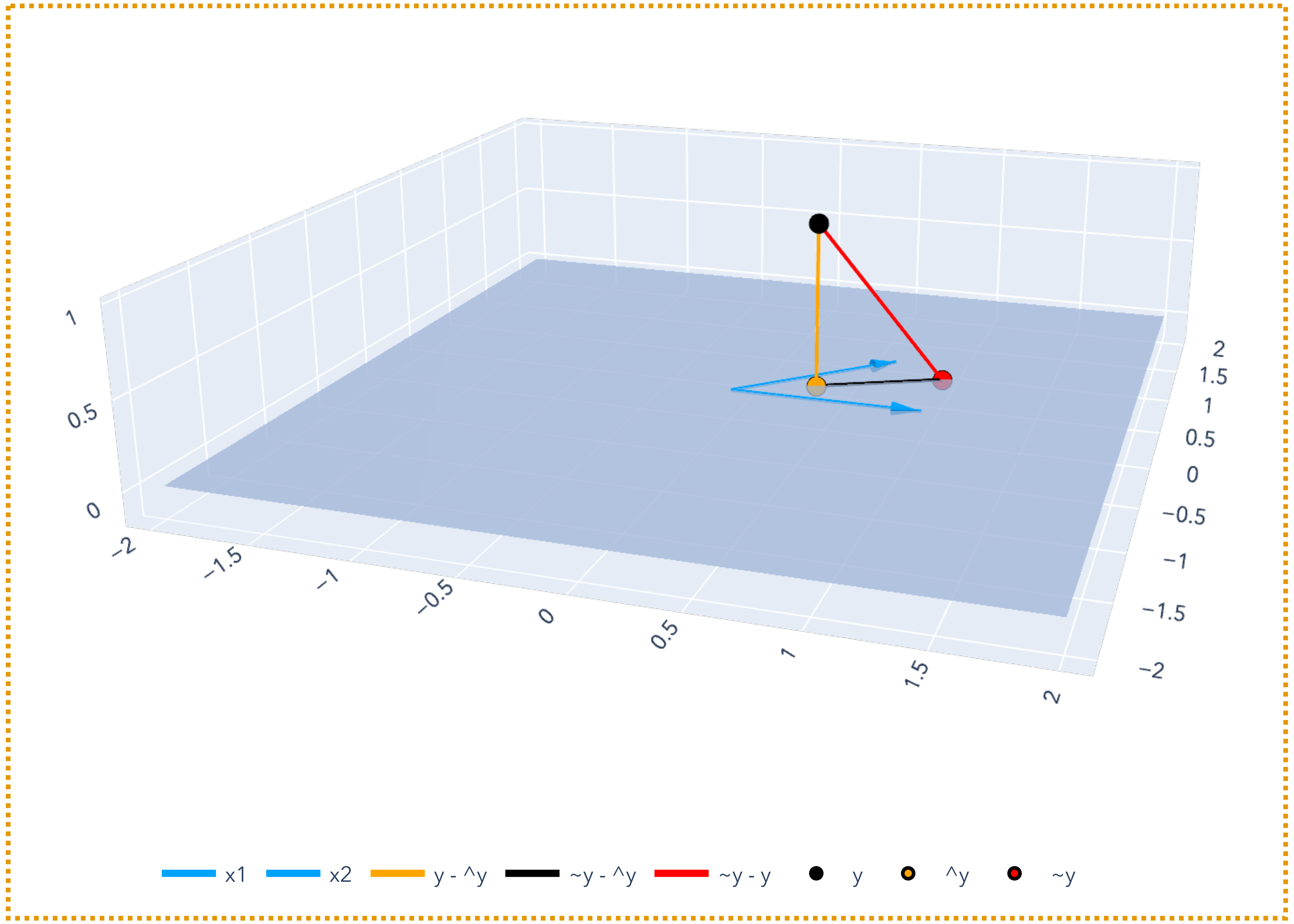
To find $\hat{\mathbf{w}}$, we follow the *principle of least squares*.

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

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

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

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



Singular Value Decomposition (SVD)

Matrix Decompositions

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

$\mathbf{U} \in \mathbb{R}^{n\times n}$ is orthogonal, i.e. $\mathbf{U}^\top\mathbf{U} = \mathbf{U}\mathbf{U}^\top = \mathbf{I}$.

$\mathbf{V} \in \mathbb{R}^{d\times d}$ is orthogonal, i.e.$\mathbf{V}^\top\mathbf{V} = \mathbf{V}\mathbf{V}^\top = \mathbf{I}$.

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

Pseudoinverse

Definition

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

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

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

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

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

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

Least Squares with Pseudoinverse

Unified Picture

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

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

We can solve exactly.

Choose

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

which is an exact solution.

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

We approximate by least squares:

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

Choose

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

the best approximate solution:

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

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

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

Choose

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

the minimum norm (exact) solution:

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

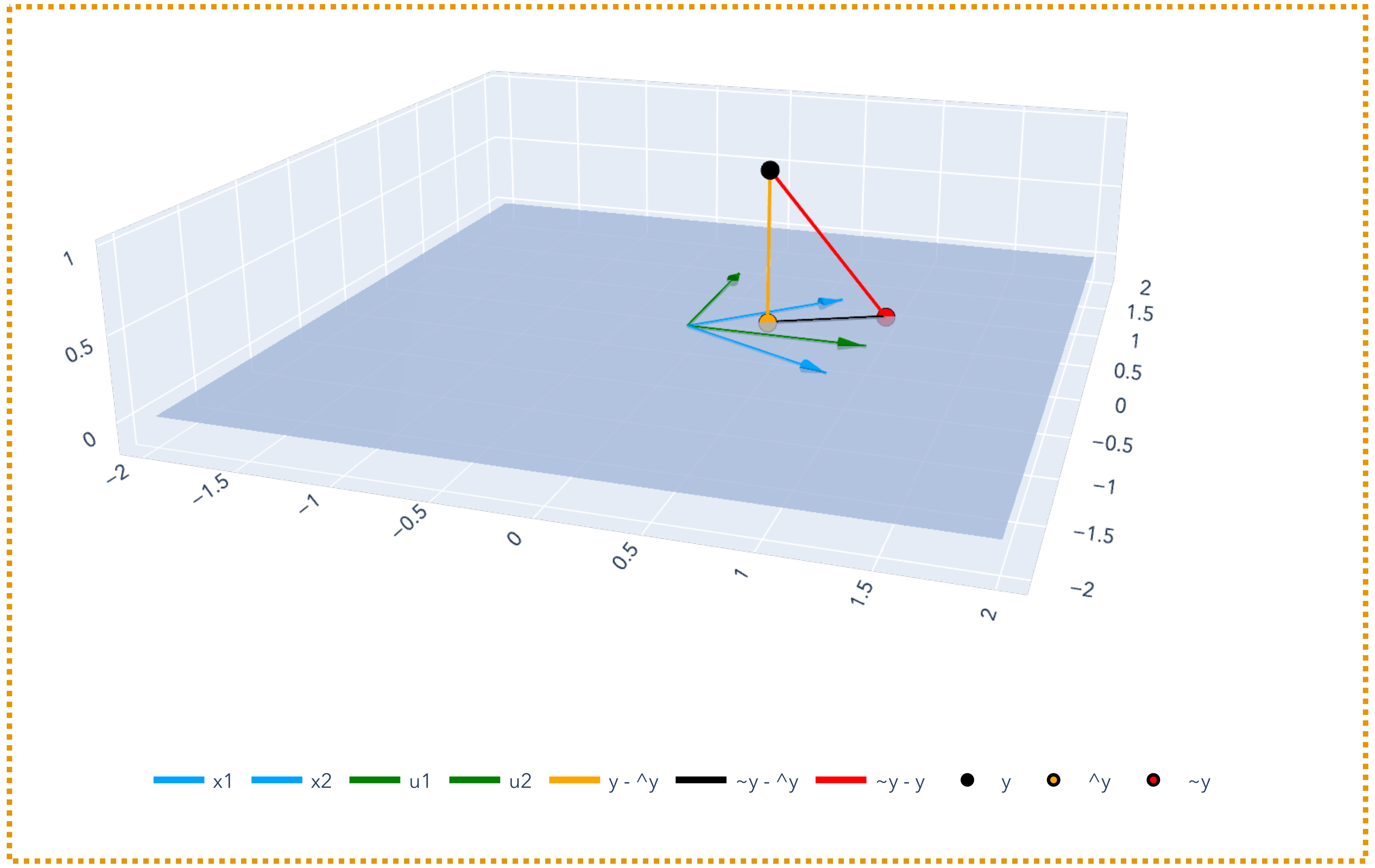
Least Squares with Pseudoinverse

Unified Picture

We want to solve $\mathbf{X}\mathbf{w} = \mathbf{y}$. Choose $\mathbf{w} = \mathbf{X}^{+}\mathbf{y}$!

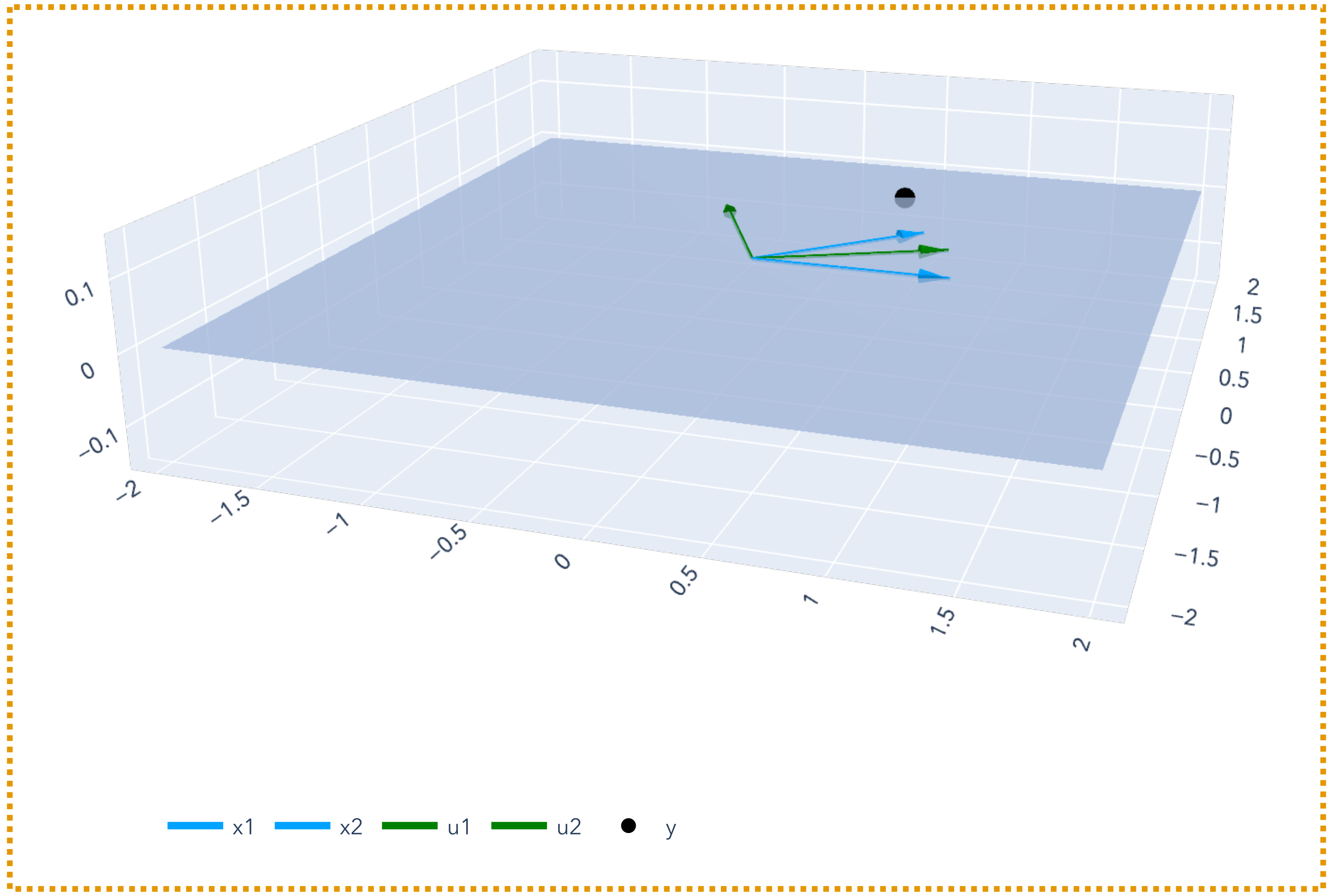
If $n > d$ and $\text{rank}(\mathbf{X}) = d\ldots$

We approximate by least squares.



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

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



What other matrix decompositions are out there?

Eigendecomposition

Motivation: Linear Dynamical System

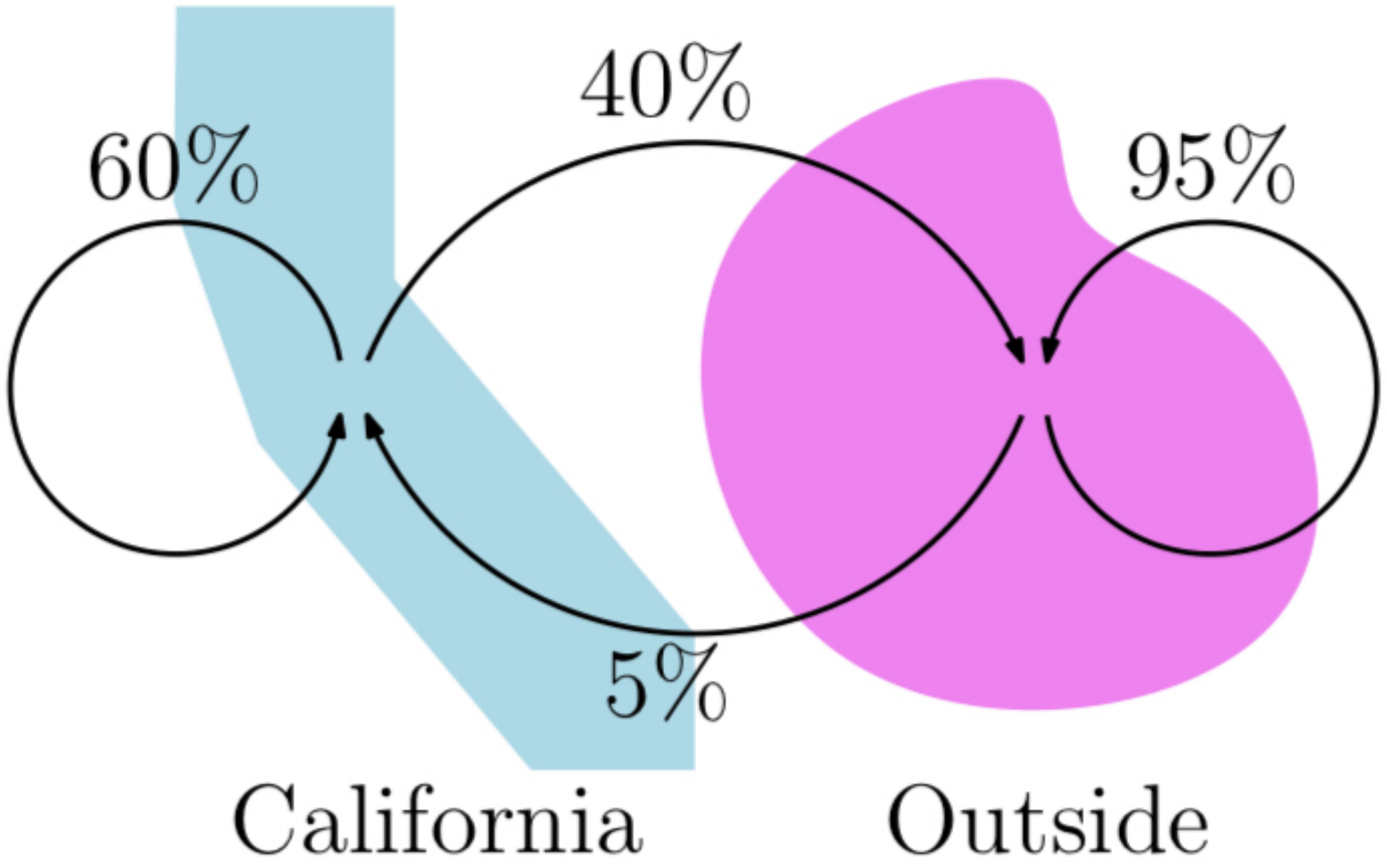
Population Change

Example of a linear dynamical system

x_{in} := people in California (at start of year)

x_{out} := people outside of California (at start of year)

$$\#\text{ inside at end of year} = 0.6x_{in} + 0.05x_{out}$$
$$\#\text{ outside at end of year} = 0.4x_{in} + 0.95x_{out}$$



Example and graphic from Daniel Hsu's course: *Computational Linear Algebra* (Fall 2022)

Population Change

Modeling with a transition matrix

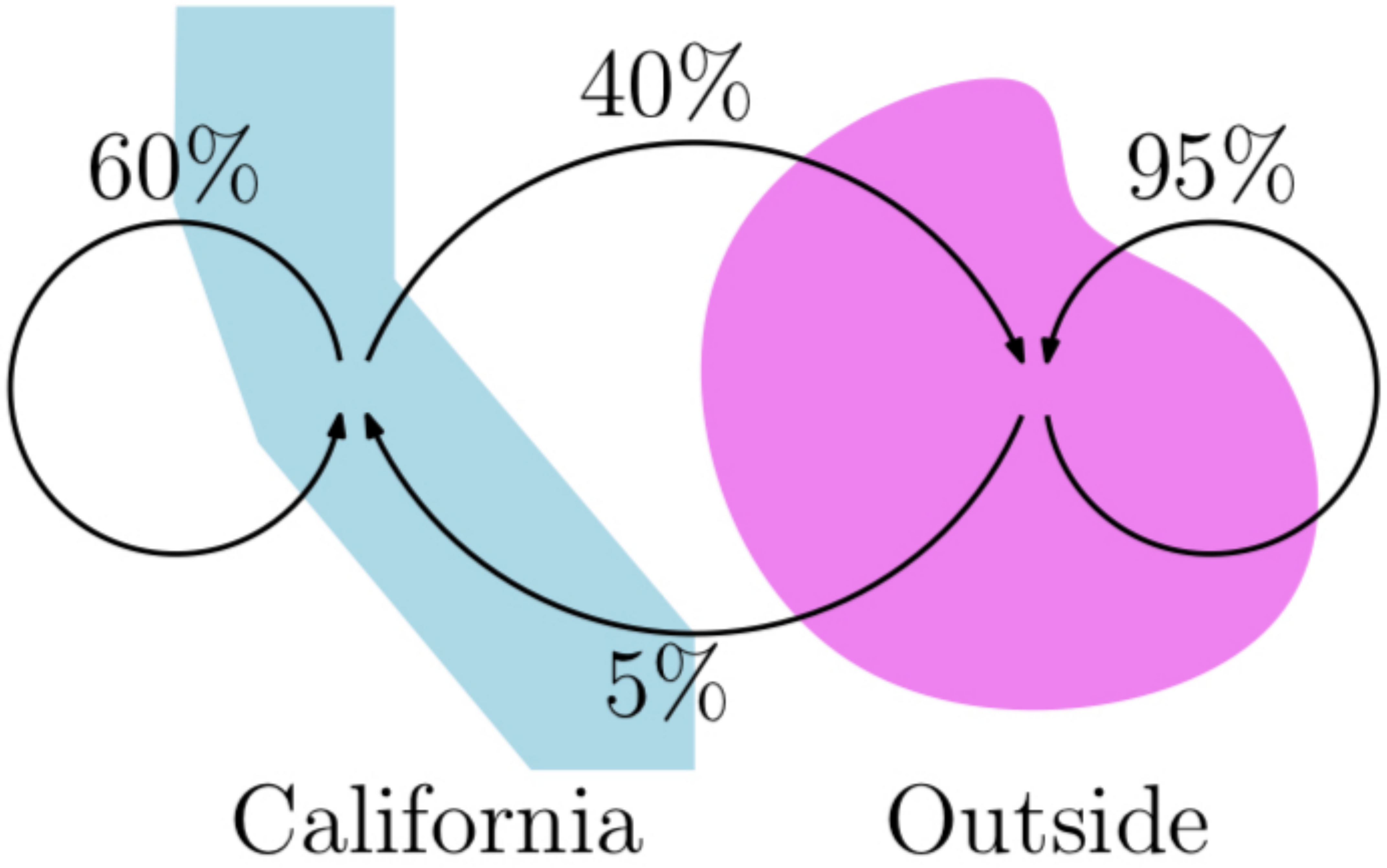
$$\#\text{ inside at end of year} = 0.6x_{in} + 0.05x_{out}$$
$$\#\text{ outside at end of year} = 0.4x_{in} + 0.95x_{out}$$

Model this with a *transition matrix*:

$$\mathbf{A} = \begin{bmatrix} in \to in & out \to in \\ in \to out & out \to out \end{bmatrix} = \begin{bmatrix} 0.6 & 0.05 \\ 0.4 & 0.95 \end{bmatrix}$$

and a system of linear equations:

$$\mathbf{Ax} = \begin{bmatrix} \text{in} \to \text{in} & \text{out} \to \text{in} \\ \text{in} \to \text{out} & \text{out} \to \text{out} \end{bmatrix} \begin{bmatrix} x_{in} \\ x_{out} \end{bmatrix} = \begin{bmatrix} 0.6 & 0.05 \\ 0.4 & 0.95 \end{bmatrix} \begin{bmatrix} x_{in} \\ x_{out} \end{bmatrix}$$



Example and graphic from Daniel Hsu's course: *Computational Linear Algebra* (Fall 2022)

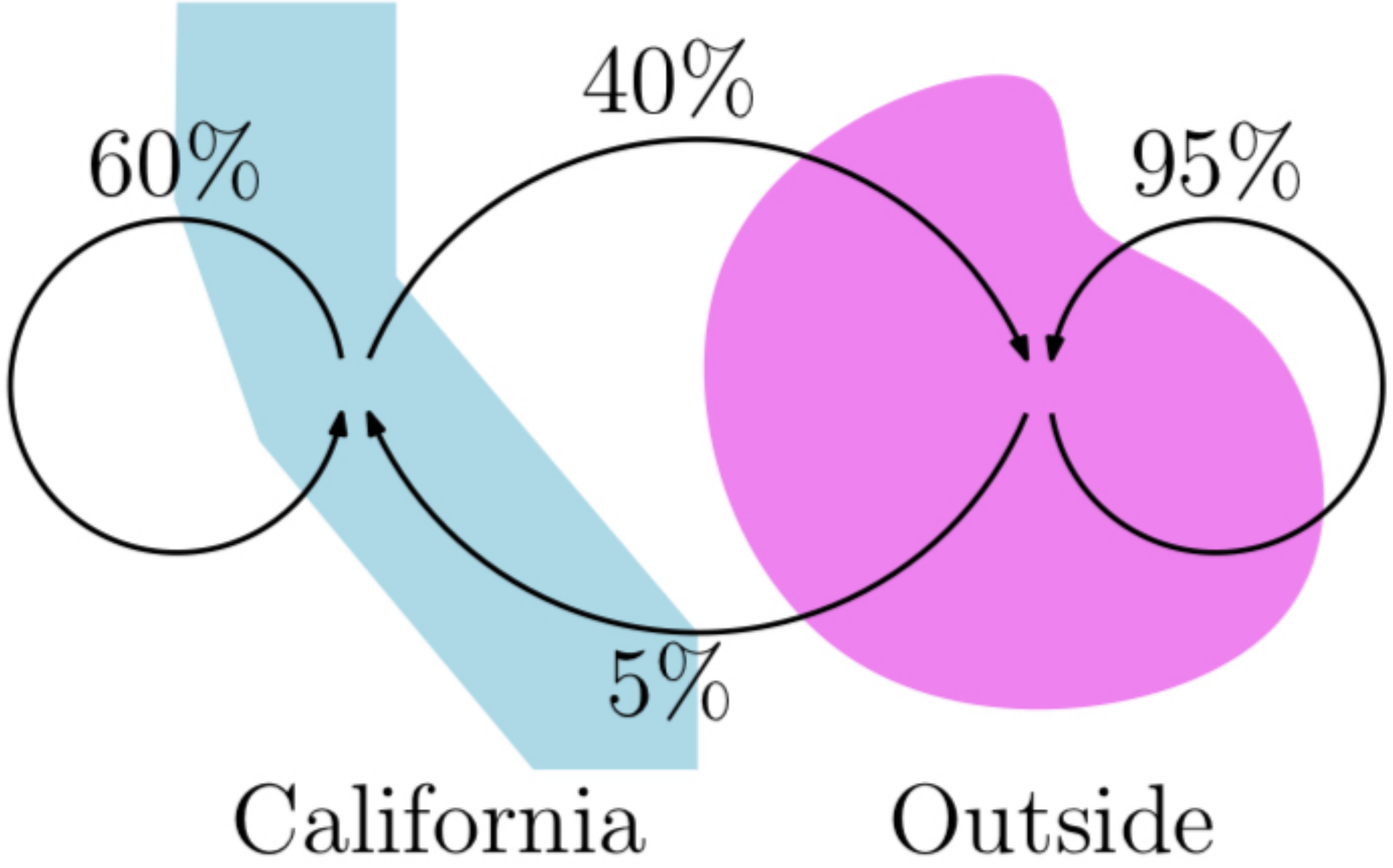
Population Change

Modeling with a transition matrix

$$\mathbf{Ax} = \begin{bmatrix} \text{in} \rightarrow \text{in} & \text{out} \rightarrow \text{in} \\ \text{in} \rightarrow \text{out} & \text{out} \rightarrow \text{out} \end{bmatrix} \begin{bmatrix} x_{in} \\ x_{out} \end{bmatrix} = \begin{bmatrix} 0.6 & 0.05 \\ 0.4 & 0.95 \end{bmatrix} \begin{bmatrix} x_{in} \\ x_{out} \end{bmatrix}.$$

$\mathbf{Ax} \in \mathbb{R}^2$ is people inside and outside of CA after one year, from the initial populations in $\mathbf{x} \in \mathbb{R}^2$.

How to find the number of people inside/outside of California after t years have passed?



Example and graphic from Daniel Hsu's course: *Computational Linear Algebra* (Fall 2022)

Population Change

Modeling with a transition matrix

$$\mathbf{A}\mathbf{x} = \begin{bmatrix} \text{in} \to \text{in} & \text{out} \to \text{in} \\ \text{in} \to \text{out} & \text{out} \to \text{out} \end{bmatrix} \begin{bmatrix} x_{in} \\ x_{out} \end{bmatrix} = \begin{bmatrix} 0.6 & 0.05 \\ 0.4 & 0.95 \end{bmatrix} \begin{bmatrix} x_{in} \\ x_{out} \end{bmatrix}.$$

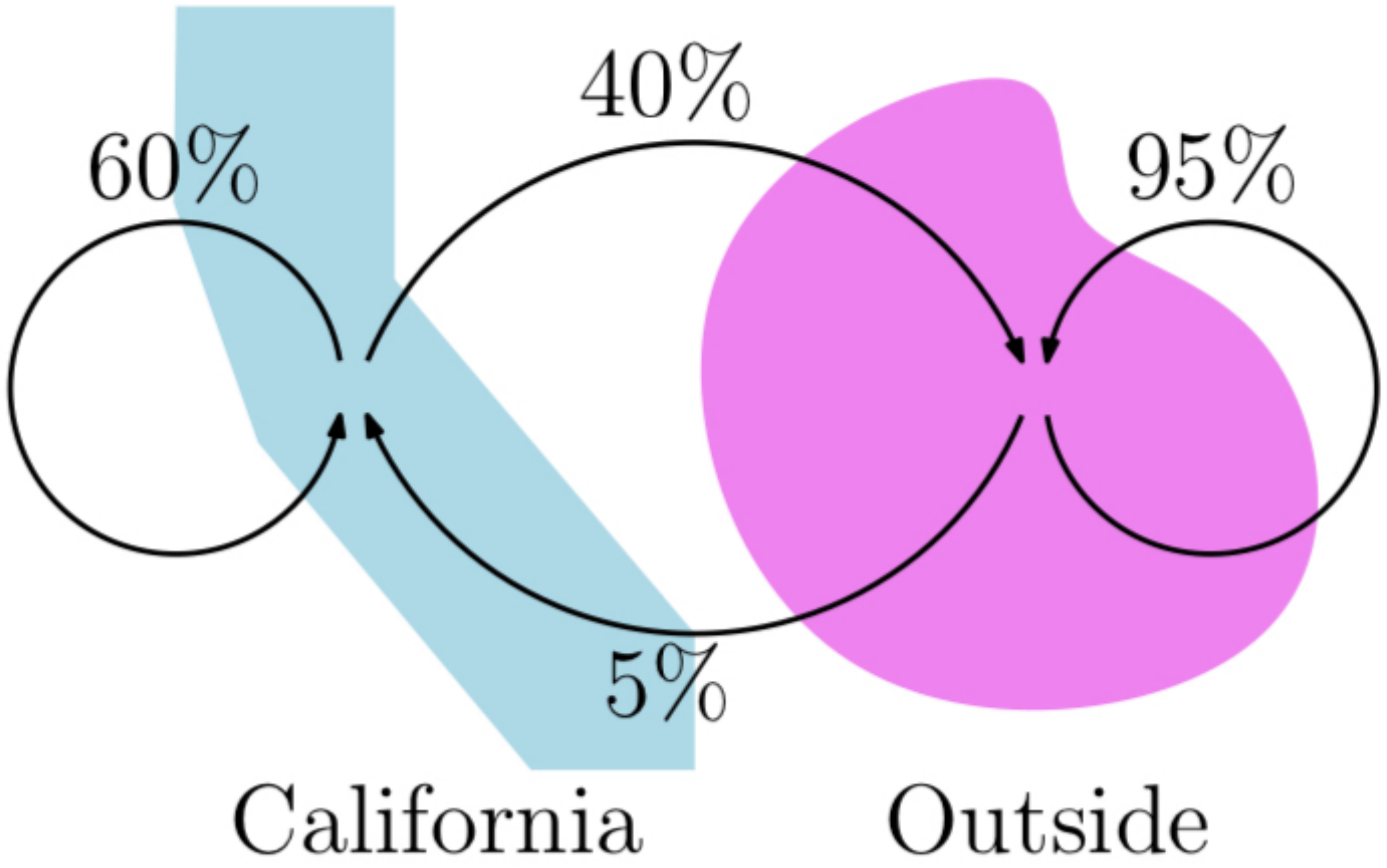
$\mathbf{A}\mathbf{x}^{(0)} \in \mathbb{R}^2$ is people inside and outside of CA after one year, from the initial populations in $\mathbf{x}^{(0)} \in \mathbb{R}^2$.

after one year: $\mathbf{x}^{(1)} = \mathbf{A}\mathbf{x}^{(0)}$

after two years: $\mathbf{x}^{(2)} = \mathbf{A}\mathbf{x}^{(1)} = \mathbf{A}\mathbf{A}\mathbf{x}^{(0)} = \mathbf{A}^2\mathbf{x}^{(0)}$

$$\vdots$$

after t years: $\mathbf{x}^{(t)} = \underbrace{\mathbf{A}\mathbf{A}\ldots\mathbf{A}}_{t \text{ products}} \mathbf{x}^{(0)} = \mathbf{A}^t\mathbf{x}^{(0)}$



Example and graphic from Daniel Hsu's course: *Computational Linear Algebra* (Fall 2022)

Population Change

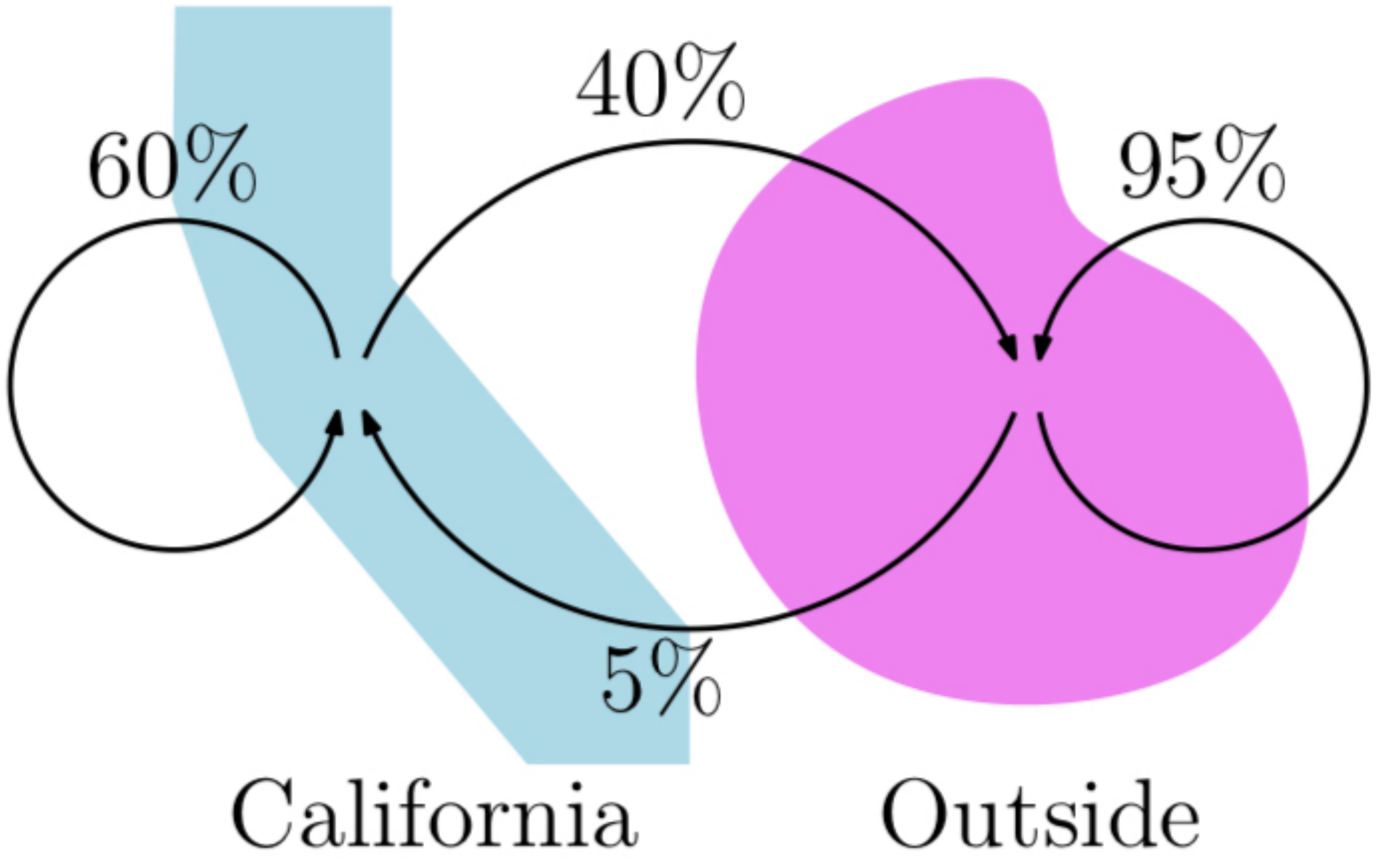
Modeling with a transition matrix

$$\mathbf{Ax} = \begin{bmatrix} \text{in} \to \text{in} & \text{out} \to \text{in} \\ \text{in} \to \text{out} & \text{out} \to \text{out} \end{bmatrix} \begin{bmatrix} x_{in} \\ x_{out} \end{bmatrix} = \begin{bmatrix} 0.6 & 0.05 \\ 0.4 & 0.95 \end{bmatrix} \begin{bmatrix} x_{in} \\ x_{out} \end{bmatrix}.$$

Let initial populations be $\mathbf{x}^{(0)} = \begin{bmatrix} 40 \\ 300 \end{bmatrix}$

What are the populations inside and outside of CA after t years?

$$\mathbf{x}^{(t)} = \mathbf{A}^t \mathbf{x}^{(0)} = \begin{bmatrix} 0.6 & 0.05 \\ 0.4 & 0.95 \end{bmatrix}^t \begin{bmatrix} 40 \\ 300 \end{bmatrix}$$



Example and graphic from Daniel Hsu's course: *Computational Linear Algebra* (Fall 2022)

Population Change

Annoying computation 😖

What are the populations inside and outside of CA after t years?

$$\mathbf{x}^{(t)} = \mathbf{A}^t \mathbf{x}^{(0)} = \begin{bmatrix} 0.6 & 0.05 \\ 0.4 & 0.95 \end{bmatrix}^t \begin{bmatrix} 40 \\ 300 \end{bmatrix}$$

Try calculating this…

$$\begin{bmatrix} 0.6 & 0.05 \\ 0.4 & 0.95 \end{bmatrix} \cdots \begin{bmatrix} 0.6 & 0.05 \\ 0.4 & 0.95 \end{bmatrix} \begin{bmatrix} 0.6 & 0.05 \\ 0.4 & 0.95 \end{bmatrix} \begin{bmatrix} 40 \\ 300 \end{bmatrix}$$ 😖

Population Change

Easy computation 😃

I hand you $\mathbf{u} = (1,8)$ and $\mathbf{v} = (-1,1)$. These two vectors have the properties:

$$\mathbf{Au} = \begin{bmatrix} 0.6 & 0.05 \\ 0.4 & 0.95 \end{bmatrix} \begin{bmatrix} 1 \\ 8 \end{bmatrix} = \begin{bmatrix} 1 \\ 8 \end{bmatrix}$$

$$\mathbf{Av} = \begin{bmatrix} 0.6 & 0.05 \\ 0.4 & 0.95 \end{bmatrix} \begin{bmatrix} -1 \\ 1 \end{bmatrix} = \frac{11}{20} \begin{bmatrix} -1 \\ 1 \end{bmatrix}$$

$$\mathbf{A}^t\mathbf{u} = \begin{bmatrix} 0.6 & 0.05 \\ 0.4 & 0.95 \end{bmatrix}^t \begin{bmatrix} 1 \\ 8 \end{bmatrix} = (1)^t \begin{bmatrix} 1 \\ 8 \end{bmatrix} = \begin{bmatrix} 1 \\ 8 \end{bmatrix}$$ 😃

$$\mathbf{A}^t\mathbf{v} = \begin{bmatrix} 0.6 & 0.05 \\ 0.4 & 0.95 \end{bmatrix}^t \begin{bmatrix} -1 \\ 1 \end{bmatrix} = \left(\frac{11}{20}\right)^t \begin{bmatrix} -1 \\ 1 \end{bmatrix}$$ 😃

Population Change

Easy computation 😃

I hand you $\mathbf{u} = (1,8)$ and $\mathbf{v} = (-1,1)$. These two vectors have the properties:

$$\mathbf{Au} = \begin{bmatrix} 0.6 & 0.05 \\ 0.4 & 0.95 \end{bmatrix} \begin{bmatrix} 1 \\ 8 \end{bmatrix} = \begin{bmatrix} 1 \\ 8 \end{bmatrix}$$

$$\mathbf{Av} = \begin{bmatrix} 0.6 & 0.05 \\ 0.4 & 0.95 \end{bmatrix} \begin{bmatrix} -1 \\ 1 \end{bmatrix} = \frac{11}{20} \begin{bmatrix} -1 \\ 1 \end{bmatrix}$$

$$\mathbf{A}^t\mathbf{u} = \begin{bmatrix} 0.6 & 0.05 \\ 0.4 & 0.95 \end{bmatrix}^t \begin{bmatrix} 1 \\ 8 \end{bmatrix} = (1)^t \begin{bmatrix} 1 \\ 8 \end{bmatrix} = \begin{bmatrix} 1 \\ 8 \end{bmatrix} \implies \mathbf{A}^t\mathbf{u} = \mathbf{u}$$

$$\mathbf{A}^t\mathbf{v} = \begin{bmatrix} 0.6 & 0.05 \\ 0.4 & 0.95 \end{bmatrix}^t \begin{bmatrix} -1 \\ 1 \end{bmatrix} = \left(\frac{11}{20}\right)^t \begin{bmatrix} -1 \\ 1 \end{bmatrix} \implies \mathbf{A}^t\mathbf{v} = \left(\frac{11}{20}\right)^t \mathbf{v}$$

Population Change

Using $\mathbf{u}$ and $\mathbf{v}$ for initial population

For $\mathbf{u} = (1,8)$ and $\mathbf{v} = (-1,1)$,

$$\mathbf{A}^t\mathbf{u} = \mathbf{u}$$

$$\mathbf{A}^t\mathbf{v} = \left(\frac{11}{20}\right)^t \mathbf{v}$$

Notice that $\mathbf{u}, \mathbf{v}$ are a basis for $\mathbb{R}^2$. Then, if we rewrite $\mathbf{x}^{(0)}$ as a linear combination of $\mathbf{u}$ and $\mathbf{v}$, i.e.

$$\mathbf{x}^{(0)} = a\mathbf{u} + b\mathbf{v},$$

we can obtain $\mathbf{x}^{(t)}$ with the following computation:

$$\mathbf{x}^{(t)} = \mathbf{A}^t\mathbf{x}^{(0)} = \mathbf{A}^t(a\mathbf{u} + b\mathbf{v}) = a\mathbf{A}^t\mathbf{u} + b\mathbf{A}^t\mathbf{v} = a\mathbf{u} + b(11/20)^t\mathbf{v}.$$

Population Change

Using $\mathbf{u}$ and $\mathbf{v}$ for initial population

For $\mathbf{u} = (1,8)$ and $\mathbf{v} = (-1,1)$, and $\mathbf{x}^{(0)}$ written as $a\mathbf{u} + b\mathbf{v}$:

$$\mathbf{x}^{(t)} = \mathbf{A}^t\mathbf{x}^{(0)} = \mathbf{A}^t(a\mathbf{u} + b\mathbf{v}) = a\mathbf{A}^t\mathbf{u} + b\mathbf{A}^t\mathbf{v} = a\mathbf{u} + b(11/20)^t\mathbf{v}.$$

In matrix form:

$$\mathbf{x}^{(0)} = \begin{bmatrix} \uparrow & \uparrow \\ \mathbf{u} & \mathbf{v} \\ \downarrow & \downarrow \end{bmatrix} \begin{bmatrix} a \\ b \end{bmatrix} = \mathbf{V} \begin{bmatrix} a \\ b \end{bmatrix}$$

$$\mathbf{x}^{(t)} = \begin{bmatrix} \uparrow & \uparrow \\ \mathbf{u} & \mathbf{v} \\ \downarrow & \downarrow \end{bmatrix} \begin{bmatrix} 1 & 0 \\ 0 & (11/20)^t \end{bmatrix} \begin{bmatrix} a \\ b \end{bmatrix} = \mathbf{V} \begin{bmatrix} 1 & 0 \\ 0 & (11/20)^t \end{bmatrix} \begin{bmatrix} a \\ b \end{bmatrix}$$

Population Change

Using $\mathbf{u}$ and $\mathbf{v}$ for initial population

For $\mathbf{u} = (1,8)$ and $\mathbf{v} = (-1,1)$, and $\mathbf{x}^{(0)}$ written as $a\mathbf{u} + b\mathbf{v}$:

$$\mathbf{x}^{(t)} = \mathbf{A}^t\mathbf{x}^{(0)} = \mathbf{A}^t(a\mathbf{u} + b\mathbf{v}) = a\mathbf{A}^t\mathbf{u} + b\mathbf{A}^t\mathbf{v} = a\mathbf{u} + b(11/20)^t\mathbf{v}.$$

In matrix form:

$$\mathbf{x}^{(0)} = \begin{bmatrix} \uparrow & \uparrow \\ \mathbf{u} & \mathbf{v} \\ \downarrow & \downarrow \end{bmatrix} \begin{bmatrix} a \\ b \end{bmatrix} = \mathbf{V} \begin{bmatrix} a \\ b \end{bmatrix} \iff \mathbf{V}^{-1}\mathbf{x}^{(0)} = \begin{bmatrix} a \\ b \end{bmatrix}$$

$$\mathbf{x}^{(t)} = \begin{bmatrix} \uparrow & \uparrow \\ \mathbf{u} & \mathbf{v} \\ \downarrow & \downarrow \end{bmatrix} \begin{bmatrix} 1 & 0 \\ 0 & (11/20)^t \end{bmatrix} \begin{bmatrix} a \\ b \end{bmatrix} = \mathbf{V} \begin{bmatrix} 1 & 0 \\ 0 & (11/20)^t \end{bmatrix} \begin{bmatrix} a \\ b \end{bmatrix}$$

Population Change

Using $\mathbf{u}$ and $\mathbf{v}$ for initial population

For $\mathbf{u} = (1,8)$ and $\mathbf{v} = (-1,1)$, and $\mathbf{x}^{(0)}$ written as $a\mathbf{u} + b\mathbf{v}$:

$$\mathbf{x}^{(t)} = \mathbf{A}^t\mathbf{x}^{(0)} = \mathbf{A}^t(a\mathbf{u} + b\mathbf{v}) = a\mathbf{A}^t\mathbf{u} + b\mathbf{A}^t\mathbf{v} = a\mathbf{u} + b(11/20)^t\mathbf{v}.$$

In matrix form:

$$\mathbf{x}^{(0)} = \begin{bmatrix} \uparrow & \uparrow \\ \mathbf{u} & \mathbf{v} \\ \downarrow & \downarrow \end{bmatrix} \begin{bmatrix} a \\ b \end{bmatrix} = \mathbf{V}\begin{bmatrix} a \\ b \end{bmatrix} \Longleftrightarrow \mathbf{V}^{-1}\mathbf{x}^{(0)} = \begin{bmatrix} a \\ b \end{bmatrix}$$

$$\mathbf{x}^{(t)} = \begin{bmatrix} \uparrow & \uparrow \\ \mathbf{u} & \mathbf{v} \\ \downarrow & \downarrow \end{bmatrix} \begin{bmatrix} 1 & 0 \\ 0 & (11/20)^t \end{bmatrix} \begin{bmatrix} a \\ b \end{bmatrix} = \mathbf{V}\begin{bmatrix} 1 & 0 \\ 0 & (11/20)^t \end{bmatrix} \begin{bmatrix} a \\ b \end{bmatrix}$$

Population Change

Using $\mathbf{u}$ and $\mathbf{v}$ for initial population

For $\mathbf{u} = (1{,}8)$ and $\mathbf{v} = (-1{,}1)$, and $\mathbf{x}^{(0)}$ written as $a\mathbf{u} + b\mathbf{v}$:

$$\mathbf{x}^{(t)} = \mathbf{A}^t\mathbf{x}^{(0)} = \mathbf{A}^t(a\mathbf{u} + b\mathbf{v}) = a\mathbf{A}^t\mathbf{u} + b\mathbf{A}^t\mathbf{v} = a\mathbf{u} + b(11/20)^t\mathbf{v}.$$

In matrix form:

$$\mathbf{x}^{(0)} = \begin{bmatrix} \uparrow & \uparrow \\ \mathbf{u} & \mathbf{v} \\ \downarrow & \downarrow \end{bmatrix} \begin{bmatrix} a \\ b \end{bmatrix} = \mathbf{V} \begin{bmatrix} a \\ b \end{bmatrix} \Longleftrightarrow \mathbf{V}^{-1}\mathbf{x}^{(0)} = \begin{bmatrix} a \\ b \end{bmatrix}$$

$$\mathbf{x}^{(t)} = \begin{bmatrix} \uparrow & \uparrow \\ \mathbf{u} & \mathbf{v} \\ \downarrow & \downarrow \end{bmatrix} \begin{bmatrix} 1 & 0 \\ 0 & (11/20)^t \end{bmatrix} \begin{bmatrix} a \\ b \end{bmatrix} = \mathbf{V} \begin{bmatrix} 1 & 0 \\ 0 & (11/20)^t \end{bmatrix} \begin{bmatrix} a \\ b \end{bmatrix} \Longleftrightarrow \mathbf{V} \begin{bmatrix} 1 & 0 \\ 0 & (11/20)^t \end{bmatrix} \mathbf{V}^{-1}\mathbf{x}^{(0)}$$

Population Change

Using $\mathbf{u}$ and $\mathbf{v}$ for initial population

For $\mathbf{u} = (1,8)$ and $\mathbf{v} = (-1,1)$, and $\mathbf{x}^{(0)}$ written as $a\mathbf{u} + b\mathbf{v}$:

$$\mathbf{x}^{(t)} = \mathbf{A}^t\mathbf{x}^{(0)} = \mathbf{A}^t(a\mathbf{u} + b\mathbf{v}) = a\mathbf{A}^t\mathbf{u} + b\mathbf{A}^t\mathbf{v} = a\mathbf{u} + b(11/20)^t\mathbf{v}.$$

In matrix form:

$$\mathbf{x}^{(0)} = \begin{bmatrix} \uparrow & \uparrow \\ \mathbf{u} & \mathbf{v} \\ \downarrow & \downarrow \end{bmatrix} \begin{bmatrix} a \\ b \end{bmatrix} = \mathbf{V} \begin{bmatrix} a \\ b \end{bmatrix} \Longleftrightarrow \mathbf{V}^{-1}\mathbf{x}^{(0)} = \begin{bmatrix} a \\ b \end{bmatrix}$$

$$\mathbf{x}^{(t)} = \begin{bmatrix} \uparrow & \uparrow \\ \mathbf{u} & \mathbf{v} \\ \downarrow & \downarrow \end{bmatrix} \begin{bmatrix} 1 & 0 \\ 0 & (11/20)^t \end{bmatrix} \begin{bmatrix} a \\ b \end{bmatrix} = \mathbf{V} \begin{bmatrix} 1 & 0 \\ 0 & (11/20)^t \end{bmatrix} \begin{bmatrix} a \\ b \end{bmatrix} \Longleftrightarrow \mathbf{V} \begin{bmatrix} 1 & 0 \\ 0 & (11/20)^t \end{bmatrix} \mathbf{V}^{-1}\mathbf{x}^{(0)}$$

Population Change

Using $\mathbf{u}$ and $\mathbf{v}$ for initial population

For $\mathbf{u} = (1,8)$ and $\mathbf{v} = (-1,1)$:

$$\mathbf{x}^{(t)} = \mathbf{V} \begin{bmatrix} 1 & 0 \\ 0 & (11/20)^t \end{bmatrix} \mathbf{V}^{-1} \mathbf{x}^{(0)}$$

where

$$\mathbf{V} = \begin{bmatrix} \uparrow & \uparrow \\ \mathbf{u} & \mathbf{v} \\ \downarrow & \downarrow \end{bmatrix}.$$

Population Change

Comparison of hard and easy computation

$$\mathbf{x}^{(t)} = \mathbf{A}^t \mathbf{x}^{(0)}$$

For initial populations $\mathbf{x}^{(0)} = (40,\ 300)$, the population after t years is:

$$\mathbf{x}^{(t)} = \begin{bmatrix} 0.6 & 0.05 \\ 0.4 & 0.95 \end{bmatrix}^t \begin{bmatrix} 40 \\ 300 \end{bmatrix}.$$

😖

$$\mathbf{x}^{(t)} = \mathbf{V} \begin{bmatrix} 1 & 0 \\ 0 & (11/20)^t \end{bmatrix} \mathbf{V}^{-1} \mathbf{x}^{(0)}$$

For initial populations $\mathbf{x}^{(0)} = (40,\ 300)$, the population after t years is:

$$\mathbf{x}^{(t)} = \begin{bmatrix} 1 & -1 \\ 8 & 1 \end{bmatrix} \begin{bmatrix} 1 & 0 \\ 0 & (11/20)^t \end{bmatrix} \begin{bmatrix} 1/9 & 1/9 \\ -8/9 & 1/9 \end{bmatrix} \begin{bmatrix} 40 \\ 300 \end{bmatrix}.$$

😃

Diagonal Matrices

Why we like diagonal matrices

Multiplying diagonal matrices with themselves many times is easy:

$$\begin{bmatrix} 1 & 0 \\ 0 & (11/20)^t \end{bmatrix} = \begin{bmatrix} 1 & 0 \\ 0 & (11/20) \end{bmatrix}^t.$$

Diagonal Matrices

Why we like diagonal matrices

Multiplying diagonal matrices with themselves many times is easy:

$$\begin{bmatrix} 1 & 0 \\ 0 & (11/20)^t \end{bmatrix} = \begin{bmatrix} 1 & 0 \\ 0 & (11/20) \end{bmatrix}^t.$$

But this matrix depended on a basis of vectors that we got out of nowhere:

$$\mathbf{u} = (1,8) \text{ and } \mathbf{v} = (-1,1).$$

In what cases (and how) can we obtain such nice bases?

Eigendecomposition

Intuition and Definition

Eigenvectors and eigenvalues

Intuition

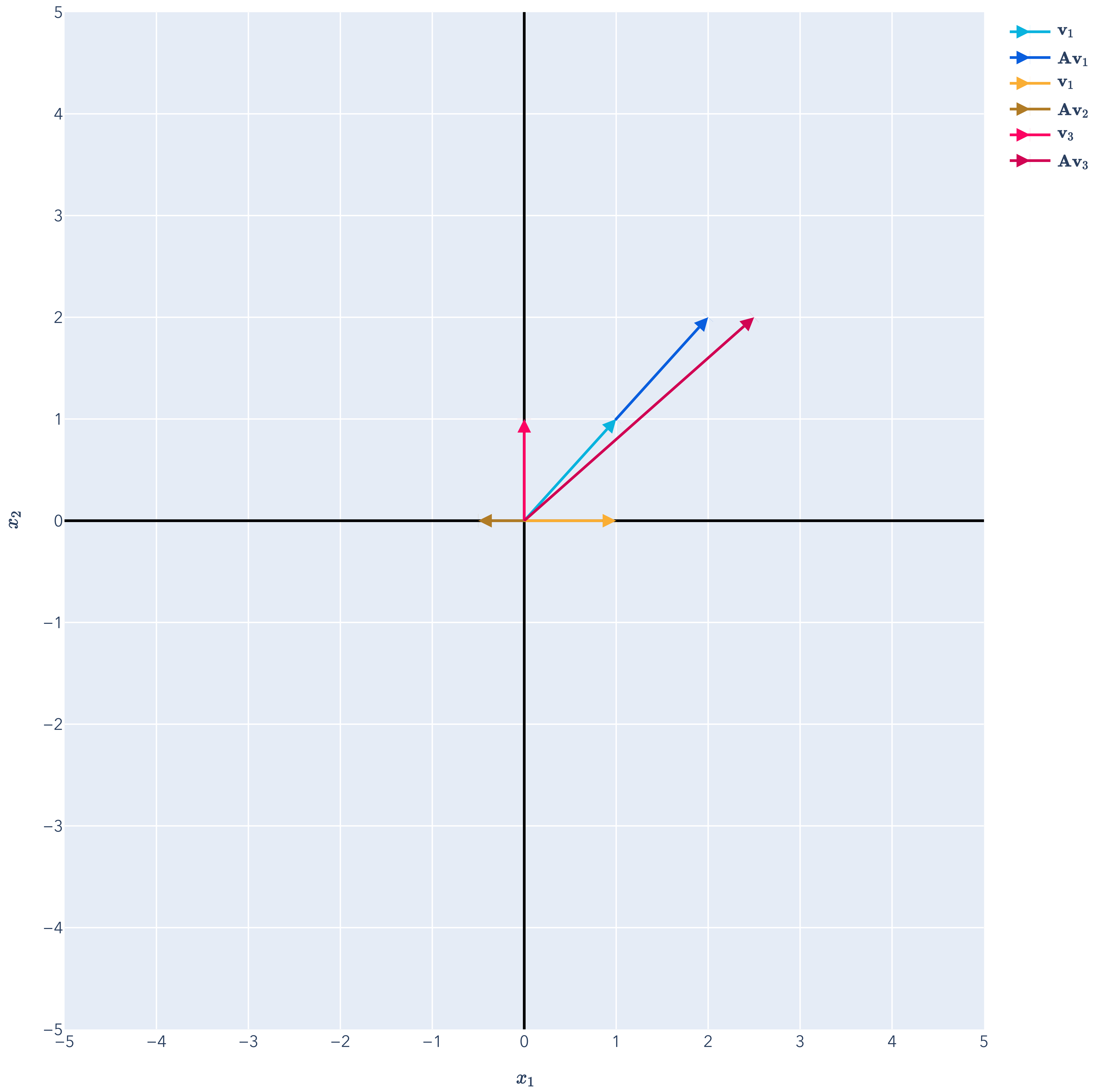
Let $\mathbf{A} \in \mathbb{R}^{d \times d}$ be a *square* matrix.

This represents a linear transformation from $\mathbb{R}^d$ to $\mathbb{R}^d$.

Eigenvectors are the vectors that just get scaled by $\mathbf{A}$.

Eigenvalues are how much $\mathbf{A}$ scales each eigenvector.

These only make sense for square matrices!



Eigenvectors and eigenvalues

Definition

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

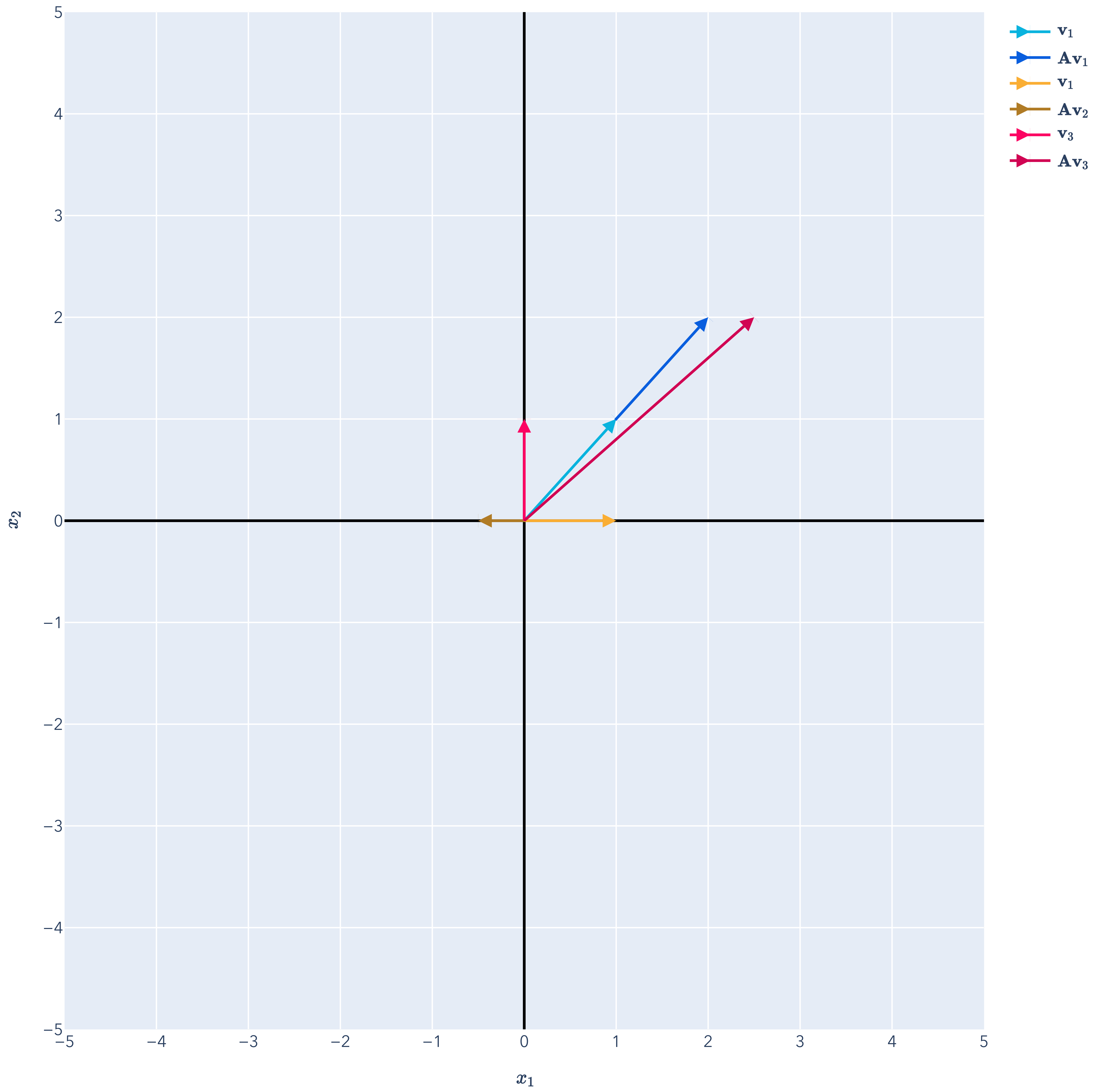
This represents a linear transformation from $\mathbb{R}^d$ to $\mathbb{R}^d$.

Eigenvectors are the nonzero vectors $\mathbf{v} \in \mathbb{R}^d$ such that:

$$\mathbf{A}\mathbf{v} = \lambda \mathbf{v}\,.$$

The scalar $\lambda \in \mathbb{R}$ is the eigenvalue associated with the eigenvector $\mathbf{v}$.

These only make sense for square matrices!



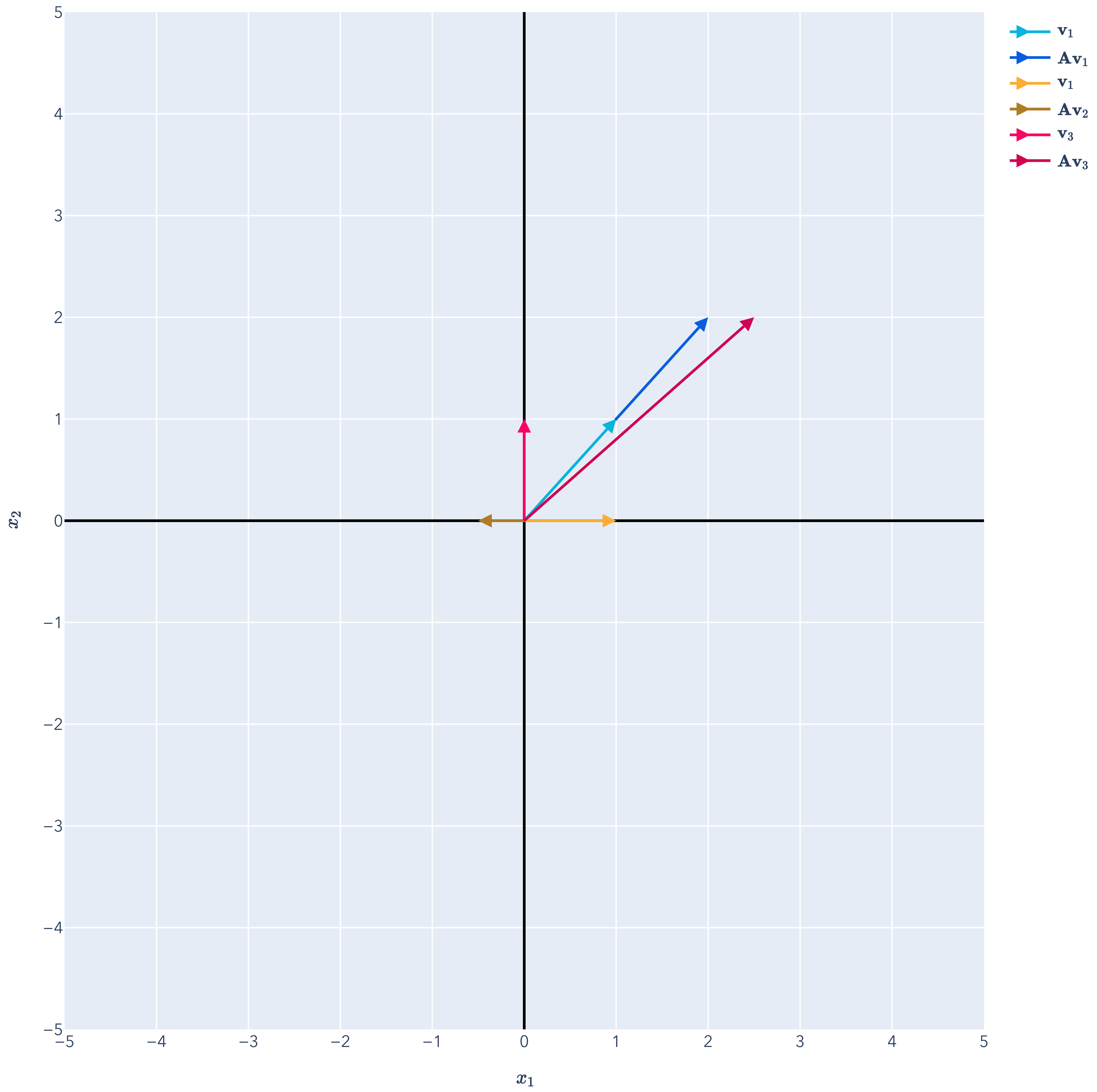
Eigenvectors and eigenvalues

Example

Consider the matrix $\mathbf{A} \in \mathbb{R}^{2\times 2}$ given by

$$\mathbf{A} = \begin{bmatrix} -1/2 & 5/2 \\ 0 & 2 \end{bmatrix}.$$

What happens to the vector $\mathbf{v}_1 = (1,1)$?



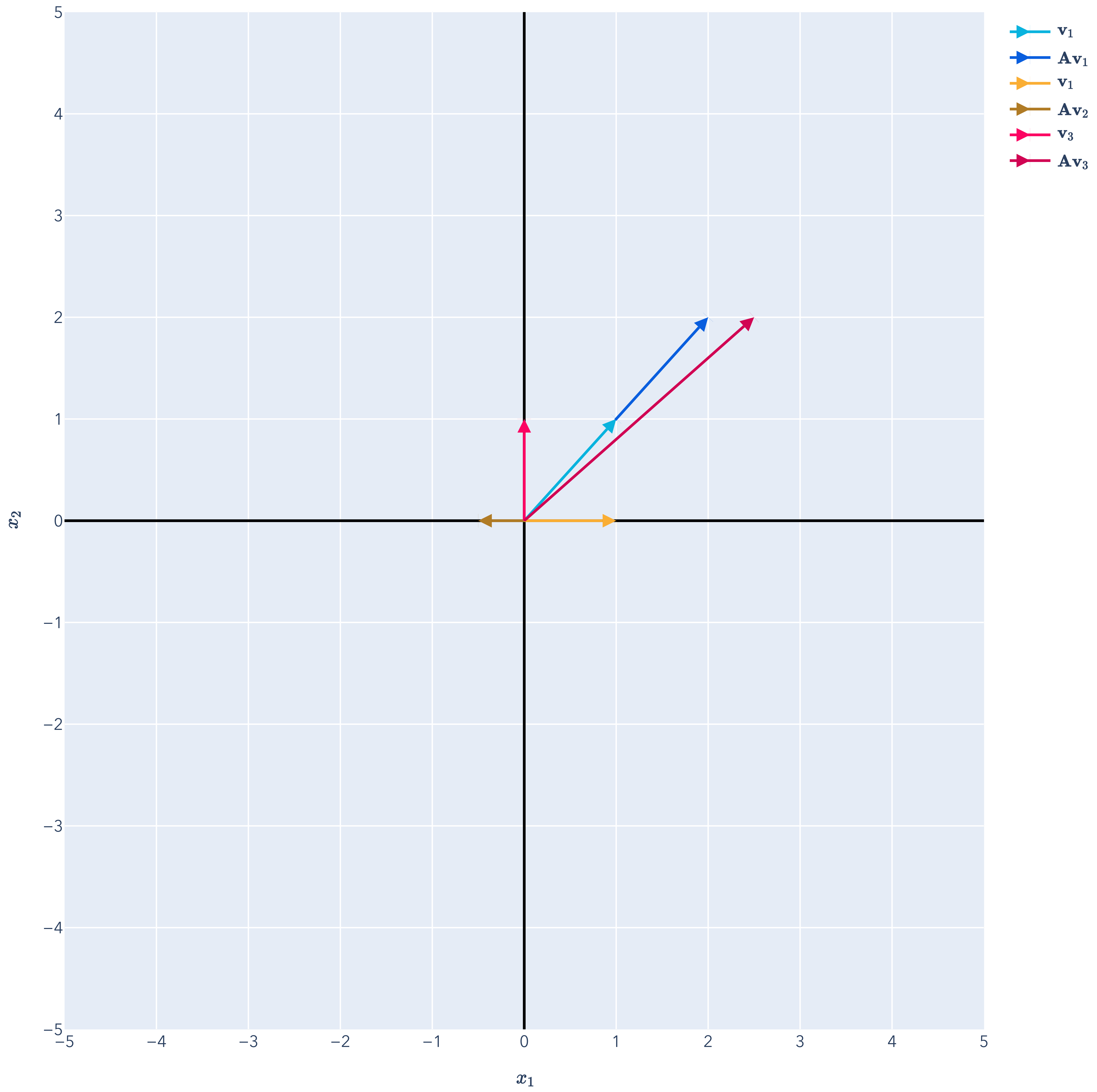
Eigenvectors and eigenvalues

Example

Consider the matrix $\mathbf{A} \in \mathbb{R}^{2\times 2}$ given by

$$\mathbf{A} = \begin{bmatrix} -1/2 & 5/2 \\ 0 & 2 \end{bmatrix}.$$

What happens to the vector $\mathbf{v}_2 = (1,0)$?



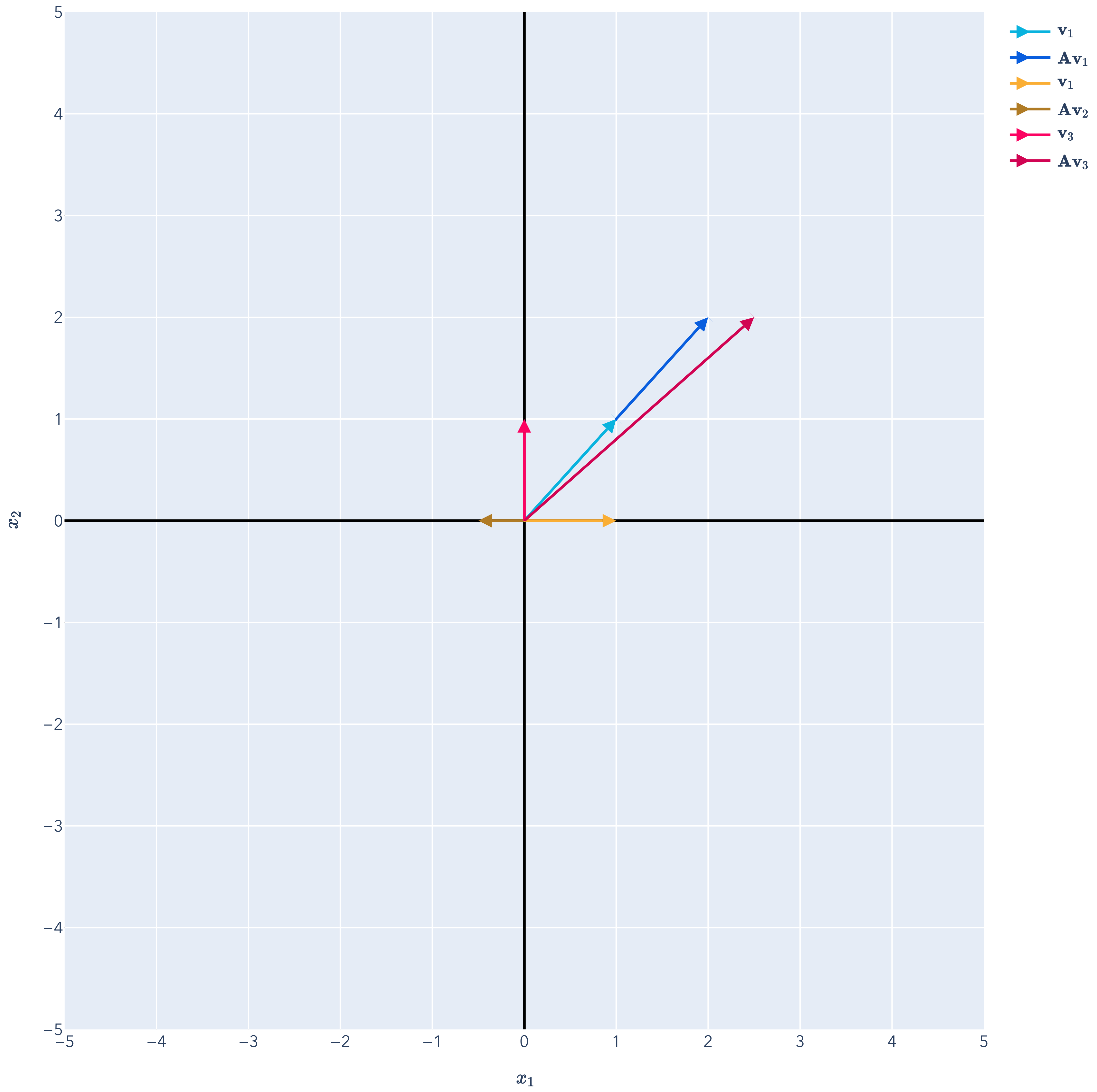
Eigenvectors and eigenvalues

Example

Consider the matrix $\mathbf{A} \in \mathbb{R}^{2\times 2}$ given by

$$\mathbf{A} = \begin{bmatrix} -1/2 & 5/2 \\ 0 & 2 \end{bmatrix}.$$

What happens to the vector $\mathbf{v}_3 = (0,1)$?



Eigenvectors and eigenvalues

Example

$$\mathbf{A} = \begin{bmatrix} -1/2 & 5/2 \\ 0 & 2 \end{bmatrix}.$$

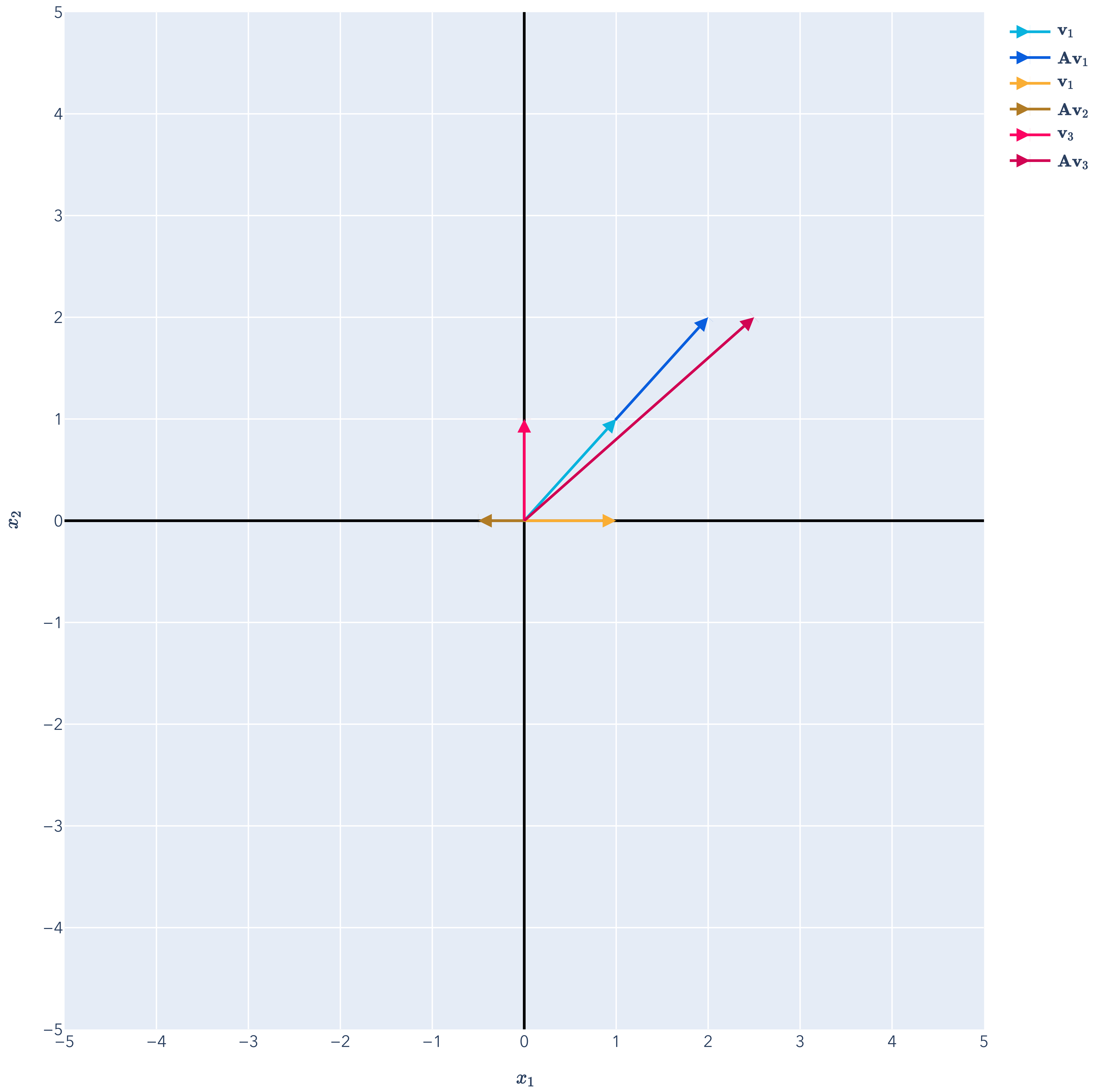
Eigenvectors (with eigenvalues $\lambda_1 = 2$ and $\lambda_2 = -1/2$):

$$\begin{bmatrix} -1/2 & 5/2 \\ 0 & 2 \end{bmatrix} \begin{bmatrix} 1 \\ 1 \end{bmatrix} = \begin{bmatrix} 2 \\ 2 \end{bmatrix} = 2 \begin{bmatrix} 1 \\ 1 \end{bmatrix}$$

$$\begin{bmatrix} -1/2 & 5/2 \\ 0 & 2 \end{bmatrix} \begin{bmatrix} 1 \\ 0 \end{bmatrix} = \begin{bmatrix} -1/2 \\ 0 \end{bmatrix} = -\frac{1}{2} \begin{bmatrix} 1 \\ 0 \end{bmatrix}$$

Not an eigenvector:

$$\begin{bmatrix} -1/2 & 5/2 \\ 0 & 2 \end{bmatrix} \begin{bmatrix} 0 \\ 1 \end{bmatrix} = \begin{bmatrix} 5/2 \\ 2 \end{bmatrix}$$



Eigenvectors and eigenvalues

Example

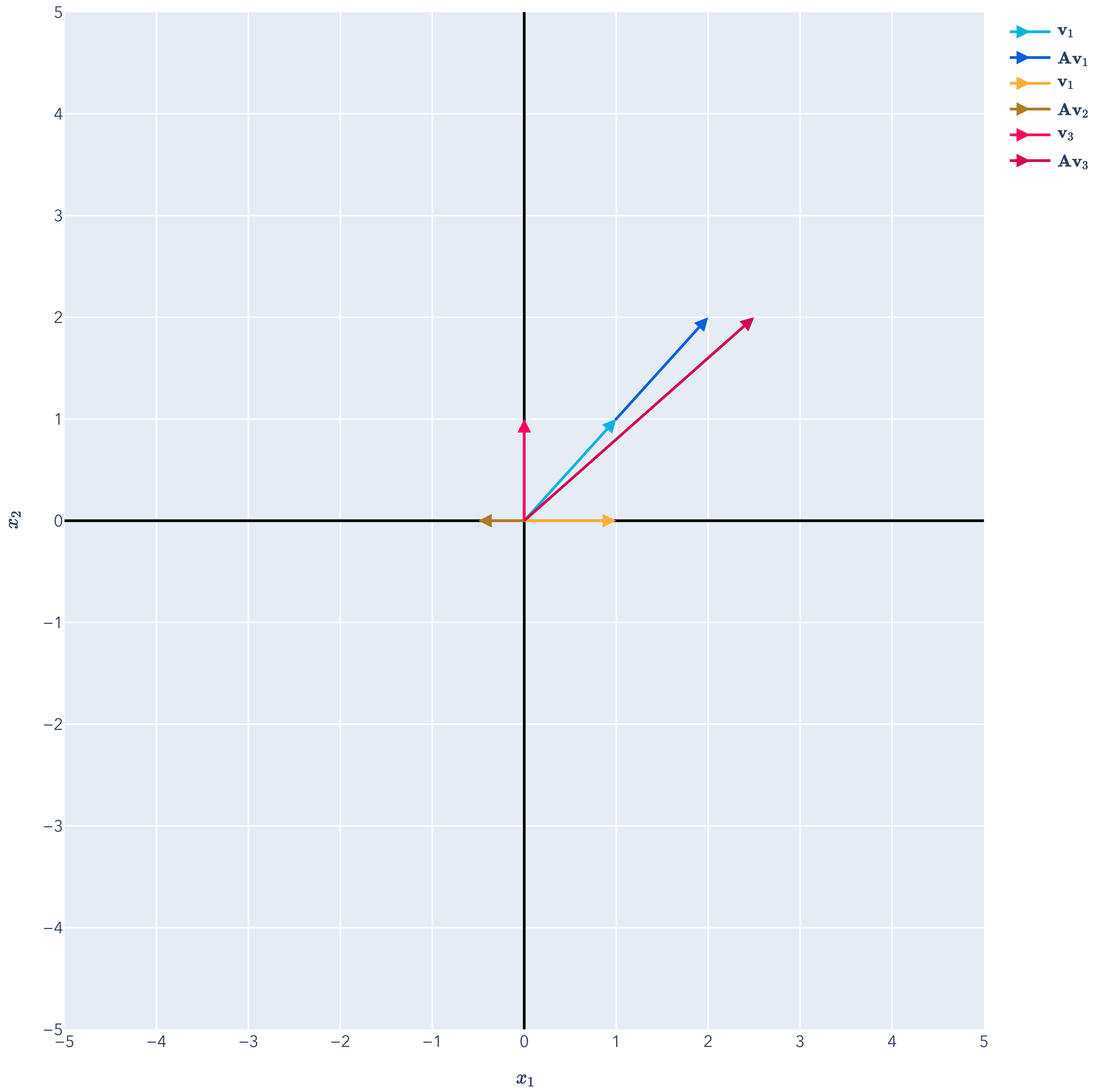
$$\mathbf{A} = \begin{bmatrix} -1/2 & 5/2 \\ 0 & 2 \end{bmatrix}$$

$\mathbf{v}_1 = (1,1)$ and $\mathbf{v}_2 = (1,0)$ form a basis for $\mathbb{R}^2$.

So any $\mathbf{x} \in \mathbb{R}^2$ can be written as: $\mathbf{x} = a\mathbf{v}_1 + b\mathbf{v}_2$.

$$\mathbf{x} = \begin{bmatrix} \uparrow & \uparrow \\ \mathbf{v}_1 & \mathbf{v}_2 \\ \downarrow & \downarrow \end{bmatrix} \begin{bmatrix} a \\ b \end{bmatrix}$$

$$\mathbf{A}^t\mathbf{x} = \mathbf{A}^t(a\mathbf{v}_1 + b\mathbf{v}_2) = a\mathbf{A}^t\mathbf{v}_1 + b\mathbf{A}^t\mathbf{v}_2 = a2^t\mathbf{v}_1 + b\left(-\frac{1}{2}\right)^t\mathbf{v}_2$$



Eigenvectors and eigenvalues

Example

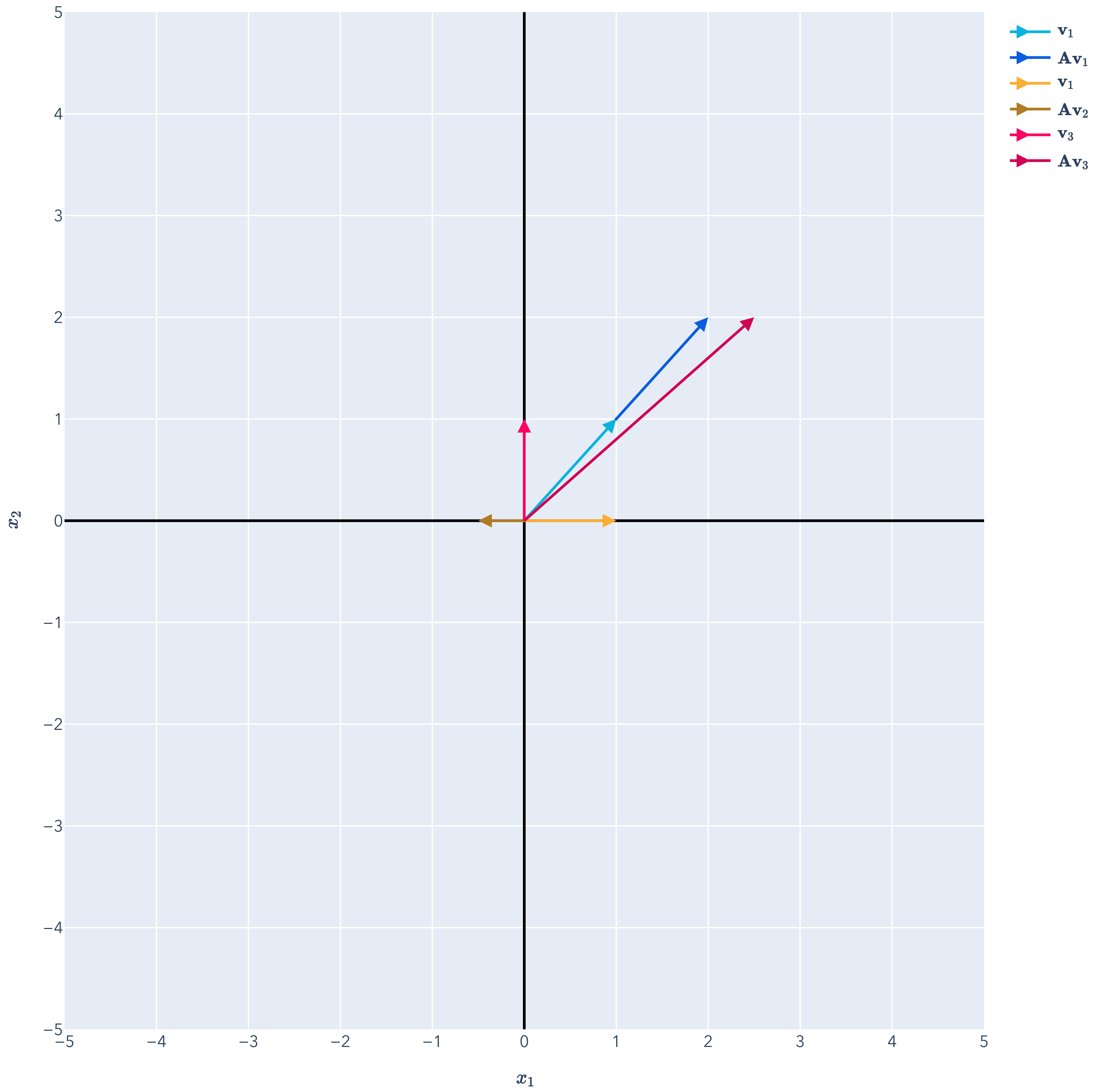
$\mathbf{v}_1 = (1,1)$ and $\mathbf{v}_2 = (1,0)$ form a basis for $\mathbb{R}^2$.

So any $\mathbf{x} \in \mathbb{R}^2$ can be written as: $\mathbf{x} = a\mathbf{v}_1 + b\mathbf{v}_2$.

$$\mathbf{x} = \begin{bmatrix} \uparrow & \uparrow \\ \mathbf{v}_1 & \mathbf{v}_2 \\ \downarrow & \downarrow \end{bmatrix} \begin{bmatrix} a \\ b \end{bmatrix} \implies \begin{bmatrix} a \\ b \end{bmatrix} = \mathbf{V}^{-1}\mathbf{x}$$

$$\mathbf{A}^t\mathbf{x} = \mathbf{A}^t(a\mathbf{v}_1 + b\mathbf{v}_2) = a\mathbf{A}^t\mathbf{v}_1 + b\mathbf{A}^t\mathbf{v}_2 = a2^t\mathbf{v}_1 + b\left(-\frac{1}{2}\right)^t \mathbf{v}_2$$

$$\implies \mathbf{A}^t\mathbf{x} = \mathbf{V}\begin{bmatrix} 2^t & 0 \\ 0 & (-1/2)^t \end{bmatrix}\mathbf{V}^{-1}\mathbf{x}$$



Eigenvectors and eigenvalues

Example

Repeated multiplication:

$$\mathbf{A}^t\mathbf{x} = \mathbf{A}^t(a\mathbf{v}_1 + b\mathbf{v}_2) = a\mathbf{A}^t\mathbf{v}_1 + b\mathbf{A}^t\mathbf{v}_2 = a2^t\mathbf{v}_1 + b\left(-\frac{1}{2}\right)^t\mathbf{v}_2 \implies \mathbf{A}^t\mathbf{x} = \mathbf{V}\begin{bmatrix}2^t & 0 \\ 0 & (-1/2)^t\end{bmatrix}\mathbf{V}^{-1}\mathbf{x}$$

Single multiplication:

$$\mathbf{A}\mathbf{x} = \mathbf{V}\begin{bmatrix}2 & 0 \\ 0 & -1/2\end{bmatrix}\mathbf{V}^{-1}\mathbf{x}$$

$\mathbf{A} = \mathbf{V}\mathbf{\Lambda}\mathbf{V}^{-1}$, where $\mathbf{\Lambda} \in \mathbb{R}^{2\times2}$ is diagonal.

Eigendecomposition

Definition

Prop (Eigendecomposition of a diagonalizable matrix). Let $\mathbf{A} \in \mathbb{R}^{d \times d}$ have d linearly independent eigenvectors, satisfying $\mathbf{A}\mathbf{v}_i = \lambda \mathbf{v}_i$ for $i \in [d]$. Then, $\mathbf{A}$ has the eigendecomposition:

$$\mathbf{A} = \mathbf{V}\boldsymbol{\Lambda}\mathbf{V}^{-1} = \begin{bmatrix} \uparrow & \dots & \uparrow \\ \mathbf{v}_1 & \dots & \mathbf{v}_d \\ \downarrow & \dots & \downarrow \end{bmatrix} \begin{bmatrix} \lambda_1 & 0 & \dots & 0 \\ 0 & \lambda_2 & \dots & 0 \\ 0 & 0 & \ddots & 0 \\ 0 & 0 & \dots & \lambda_d \end{bmatrix} \begin{bmatrix} \uparrow & \dots & \uparrow \\ \mathbf{v}_1 & \dots & \mathbf{v}_d \\ \downarrow & \dots & \downarrow \end{bmatrix}^{-1},$$

where $\boldsymbol{\Lambda} \in \mathbb{R}^{d \times d}$ and $\mathbf{V} \in \mathbb{R}^{d \times d}$.

A matrix with an eigendecomposition is called diagonalizable.

Eigendecomposition

Example

$\mathbf{A} = \begin{bmatrix} -1/2 & 5/2 \\ 0 & 2 \end{bmatrix}$ has the eigenvectors $\mathbf{v}_1 = (1,1)$ and $\mathbf{v}_2 = (1,0)$ because

$$\mathbf{A}\mathbf{v}_1 = 2\mathbf{v}_1 \text{ and } \mathbf{A}\mathbf{v}_2 = -\frac{1}{2}\mathbf{v}_2.$$

$\mathbf{v}_1$ and $\mathbf{v}_2$ are *linearly independent*, so $\mathbf{A}$ is *diagonalizable* with *eigendecomposition:*

$$\mathbf{A} = \mathbf{V}\boldsymbol{\Lambda}\mathbf{V}^{-1}$$

$$\begin{bmatrix} -1/2 & 5/2 \\ 0 & 2 \end{bmatrix} = \begin{bmatrix} 1 & 1 \\ 1 & 0 \end{bmatrix} \begin{bmatrix} 2 & 0 \\ 0 & -1/2 \end{bmatrix} \begin{bmatrix} 0 & 1 \\ 1 & -1 \end{bmatrix}$$

Question: But when do (square) matrices have a basis of eigenvectors?

Eigendecomposition

Connection with SVD

Connection with SVD

Eigendecomposition from SVD

Eigendecomposition only applies to *square* matrices $\mathbf{A} \in \mathbb{R}^{d \times d}$:

$$\mathbf{A} = \mathbf{V}\mathbf{\Lambda}\mathbf{V}^{-1}\,.$$

The SVD applies to *any* matrix $\mathbf{X} \in \mathbb{R}^{n \times d}$:

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

Connection with SVD

Eigendecomposition from SVD

The SVD applies to *any* matrix $\mathbf{X} \in \mathbb{R}^{n \times d}$:

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

Consider the square matrix $\mathbf{A} = \mathbf{X}^\top\mathbf{X} \in \mathbb{R}^{d \times d}$. By the SVD:

$$\begin{aligned} \mathbf{A} &= \mathbf{X}^\top\mathbf{X} \\ &= \mathbf{V}\boldsymbol{\Sigma}^\top\mathbf{U}^\top\mathbf{U}\boldsymbol{\Sigma}\mathbf{V}^\top \\ &= \mathbf{V}\boldsymbol{\Sigma}^\top\boldsymbol{\Sigma}\mathbf{V}^\top \end{aligned}$$

Connection with SVD

Eigendecomposition from SVD

The SVD applies to *any* matrix $\mathbf{X} \in \mathbb{R}^{n \times d}$:

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

Consider the square matrix $\mathbf{A} = \mathbf{X}^\top\mathbf{X} \in \mathbb{R}^{d \times d}$. By the SVD:

$$\mathbf{A} = \underbrace{\mathbf{V}}_{d \times d} \; \underbrace{\boldsymbol{\Sigma}^\top\boldsymbol{\Sigma}}_{d \times d} \; \underbrace{\mathbf{V}^\top}_{d \times d}$$

The *eigendecomposition* of $\mathbf{A}$ is:

$$\mathbf{A} = \underbrace{\mathbf{V}}_{d \times d} \; \underbrace{\boldsymbol{\Lambda}}_{d \times d} \; \underbrace{\mathbf{V}^{-1}}_{d \times d}$$

Connection with SVD

Eigendecomposition from SVD

Theorem (SVD and Eigendecomposition). Let $\mathbf{X} \in \mathbb{R}^{n \times d}$ be a matrix with $\operatorname{rank}(\mathbf{X}) = r$ and $\mathbf{A} = \mathbf{X}^\top \mathbf{X} \in \mathbb{R}^{d \times d}$. Let the SVD of $\mathbf{X} = \mathbf{U}\boldsymbol{\Sigma}\mathbf{V}^\top$ have nonzero singular values

Note: this isn't the original matrix!

$$\sigma_1 \geq \sigma_2 \geq \ldots \geq \sigma_r > 0,$$

and let $\mathbf{v}_1, \ldots, \mathbf{v}_d$ be the columns of $\mathbf{V} \in \mathbb{R}^{d \times d}$. Then, each $\mathbf{v}_i$ is an eigenvector for $\mathbf{A}$ with corresponding eigenvalue $\lambda_i = \sigma_i^2$, and the eigendecomposition of $\mathbf{A}$ is:

$$\mathbf{A} = \mathbf{V}\boldsymbol{\Lambda}\mathbf{V}^\top,$$

where $\boldsymbol{\Lambda} \in \mathbb{R}^{d \times d}$ is the diagonal matrix with entries $\lambda_i = \sigma_i^2$ for $i \in [d]$.

Connection with SVD

Eigendecomposition from SVD

Therefore, for *any* matrix $\mathbf{X} \in \mathbb{R}^{n \times d}$, if $\mathbf{A} = \mathbf{X}^\top \mathbf{X}$ we know that we have d linearly independent eigenvectors – this is a case when $\mathbf{A}$ is diagonalizable!

Moreover, the eigendecomposition looks like:

$$\mathbf{X}^\top \mathbf{X} = \mathbf{A} = \mathbf{V} \boldsymbol{\Lambda} \mathbf{V}^\top$$

where $\mathbf{X} = \mathbf{U} \boldsymbol{\Sigma} \mathbf{V}^\top$ is the SVD of $\mathbf{X}$.

Positive Semidefinite Matrices

Definition and Connections

Positive Semidefinite (PSD) Matrices

First definition

Square matrix $\mathbf{A} \in \mathbb{R}^{d \times d}$ is positive semidefinite (PSD) if there exists a matrix $\mathbf{X} \in \mathbb{R}^{n \times d}$ s.t.

$$\mathbf{A} = \mathbf{X}^\top \mathbf{X}.$$

Note: If you've seen PSD matrices before, this isn't the usual first definition (but it's equivalent).

Positive Semidefinite (PSD) Matrices

Symmetry of PSD Matrices

Square matrix $\mathbf{A} \in \mathbb{R}^{d \times d}$ is positive semidefinite (PSD) if there exists a matrix $\mathbf{X} \in \mathbb{R}^{n \times d}$ s.t.

$$\mathbf{A} = \mathbf{X}^\top \mathbf{X}.$$

Prop (Symmetry of PSD matrices). Any PSD matrix is symmetric. If $\mathbf{A} \in \mathbb{R}^{d \times d}$ is PSD, then

$$\mathbf{A} = \mathbf{A}^\top.$$

Positive Semidefinite (PSD) Matrices

Example

$\mathbf{A} = \begin{bmatrix} 5/2 & 3/2 \\ 3/2 & 5/2 \end{bmatrix}$ is positive semidefinite.

Its "square root" is the matrix

$$\mathbf{X} = \begin{bmatrix} \frac{2}{\sqrt{2}} & \frac{2}{\sqrt{2}} \\ \frac{1}{\sqrt{2}} & -\frac{1}{\sqrt{2}} \\ 0 & 0 \end{bmatrix} \text{ because } \mathbf{X}^\top \mathbf{X} = \begin{bmatrix} \frac{2}{\sqrt{2}} & \frac{1}{\sqrt{2}} & 0 \\ \frac{2}{\sqrt{2}} & -\frac{1}{\sqrt{2}} & 0 \end{bmatrix} \begin{bmatrix} \frac{2}{\sqrt{2}} & \frac{2}{\sqrt{2}} \\ \frac{1}{\sqrt{2}} & -\frac{1}{\sqrt{2}} \\ 0 & 0 \end{bmatrix} = \begin{bmatrix} 5/2 & 3/2 \\ 3/2 & 5/2 \end{bmatrix} = \mathbf{A}$$

PSD Matrices and Eigendecomposition

Connection to eigenvalues

By Theorem (SVD and Eigendecomposition), if $\mathbf{A}$ is PSD with $\mathbf{A} = \mathbf{X}^\top \mathbf{X}$ and $\mathbf{X} = \mathbf{U}\boldsymbol{\Sigma}\mathbf{V}^\top$ then

$$\mathbf{A} = \mathbf{V}\boldsymbol{\Lambda}\mathbf{V}^\top,$$

with orthonormal eigenvectors $\mathbf{v}_1, \ldots, \mathbf{v}_d$ and nonnegative eigenvalues $\lambda_1 = \sigma_1^2, \ldots, \lambda_d = \sigma_d^2$.

The reverse direction is also true!

PSD Matrices and Eigendecomposition

Second definition

A square matrix $\mathbf{A} \in \mathbb{R}^{d \times d}$ is positive semidefinite (PSD) if $\mathbf{A}$ has d eigenvectors forming an orthonormal basis for $\mathbb{R}^d$ with corresponding *nonnegative* eigenvalues $\lambda_1, \ldots, \lambda_d \geq 0$.

Positive Semidefinite (PSD) Matrices

Example

$$\mathbf{A} = \begin{bmatrix} 5/2 & 3/2 \\ 3/2 & 5/2 \end{bmatrix} \text{ is positive semidefinite.}$$

It has the eigenvectors $\mathbf{v}_1 = \left(1/\sqrt{2}, 1/\sqrt{2}\right)$ and $\mathbf{v}_2 = \left(1/\sqrt{2}, -1/\sqrt{2}\right)$:

$$\mathbf{A}\mathbf{v}_1 = \begin{bmatrix} 5/2 & 3/2 \\ 3/2 & 5/2 \end{bmatrix} \begin{bmatrix} 1/\sqrt{2} \\ 1/\sqrt{2} \end{bmatrix} = \begin{bmatrix} 4/\sqrt{2} \\ 4/\sqrt{2} \end{bmatrix} = 4 \begin{bmatrix} 1/\sqrt{2} \\ 1/\sqrt{2} \end{bmatrix} \implies \lambda_1 = 4$$

$$\mathbf{A}\mathbf{v}_2 = \begin{bmatrix} 5/2 & 3/2 \\ 3/2 & 5/2 \end{bmatrix} \begin{bmatrix} 1/\sqrt{2} \\ -1/\sqrt{2} \end{bmatrix} = \begin{bmatrix} 1/\sqrt{2} \\ -1/\sqrt{2} \end{bmatrix} \implies \lambda_1 = 1$$

The eigenvectors are orthonormal and $\lambda_1, \lambda_2 \geq 0$, so $\mathbf{A} = \mathbf{V}\boldsymbol{\Lambda}\mathbf{V}^\top$ is positive semidefinite.

Positive Semidefinite (PSD) Matrices

Third definition

A square matrix $\mathbf{A} \in \mathbb{R}^{d \times d}$ is positive semidefinite (PSD) if, for any $\mathbf{x} \in \mathbb{R}^d$,

$$\mathbf{x}^\top \mathbf{A} \mathbf{x} \geq 0.$$

This is often taken as the definition of PSD (but it is equivalent to the other two definitions).

Positive Semidefinite (PSD) Matrices

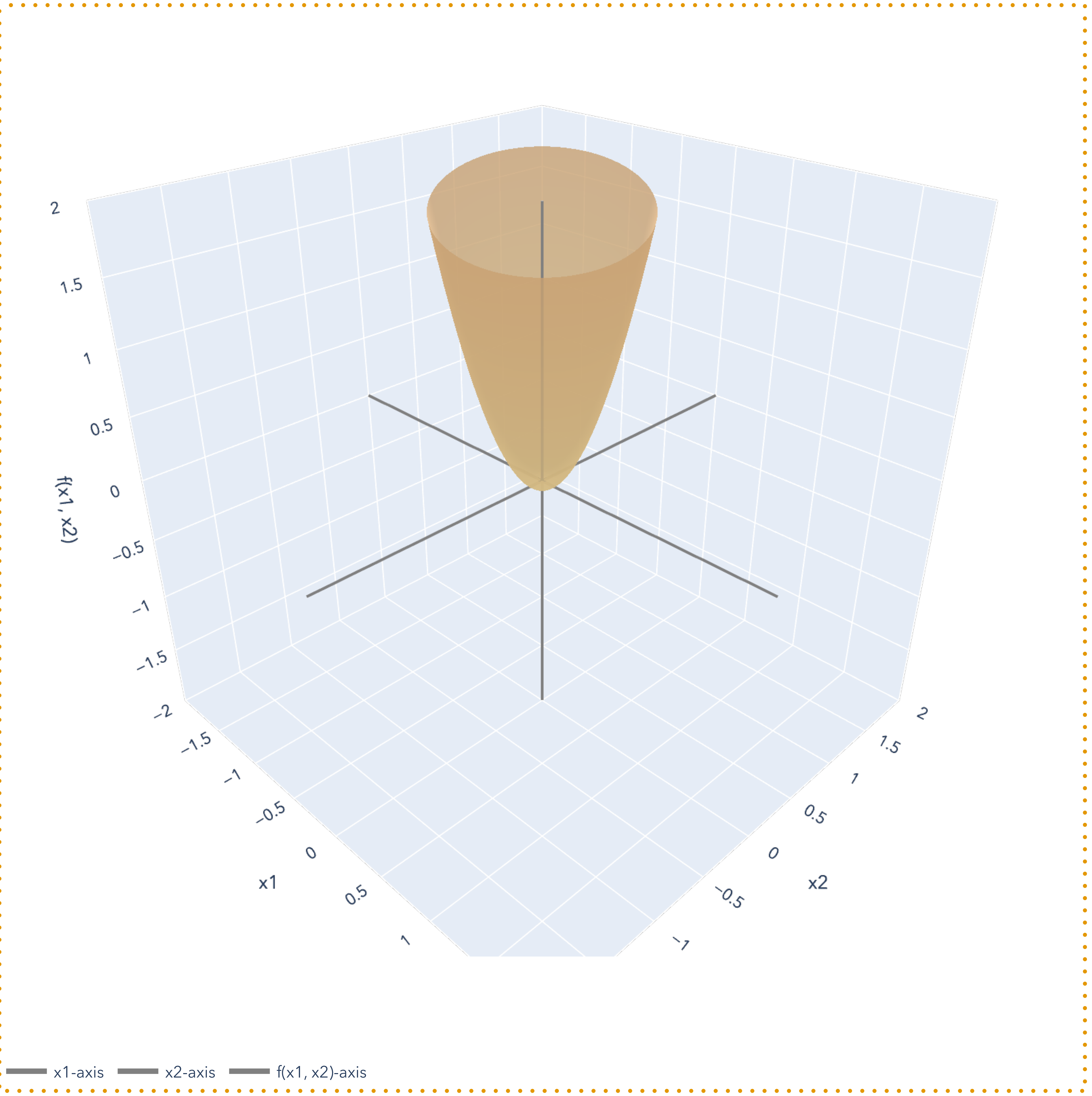
Example

$$\mathbf{A} = \begin{bmatrix} 5/2 & 3/2 \\ 3/2 & 5/2 \end{bmatrix} \text{ is positive semidefinite.}$$

Consider any vector $\mathbf{x} = (x_1, x_2) \in \mathbb{R}^d$.

$$\mathbf{x}^\top \mathbf{A}\mathbf{x} = [x_1 \quad x_2] \begin{bmatrix} 5/2 & 3/2 \\ 3/2 & 5/2 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \end{bmatrix} = [x_1 \quad x_2] \begin{bmatrix} (5/2)x_1 + (3/2)x_2 \\ (3/2)x_1 + (5/2)x_2 \end{bmatrix}$$

$$\mathbf{x}^\top \mathbf{A}\mathbf{x} = (5/2)x_1^2 + 3x_1x_2 + (5/2)x_2^2$$



Positive Semidefinite (PSD) Matrices

All definitions

A square matrix $\mathbf{A} \in \mathbb{R}^{d \times d}$ is positive semidefinite (PSD) if…

there exists $\mathbf{X} \in \mathbb{R}^{n \times d}$ such that $\mathbf{A} = \mathbf{X}^\top \mathbf{X}$.

$$\updownarrow$$

all eigenvalues of $\mathbf{A}$ are nonnegative: $\lambda_1 \geq 0, \ldots, \lambda_d \geq 0$.

$$\updownarrow$$

$\mathbf{x}^\top \mathbf{A} \mathbf{x} \geq 0$ for any $\mathbf{x} \in \mathbb{R}^d$.

Positive Definite (PD) Matrices

All definitions

A square matrix $\mathbf{A} \in \mathbb{R}^{d \times d}$ is positive definite (PD) if…

there exists *an invertible matrix* $\mathbf{X} \in \mathbb{R}^{d \times d}$ such that $\mathbf{A} = \mathbf{X}^\top \mathbf{X}$.

$\updownarrow$

all eigenvalues of $\mathbf{A}$ are positive: $\lambda_1 > 0, \ldots, \lambda_d > 0$.

$\updownarrow$

$\mathbf{x}^\top \mathbf{A} \mathbf{x} > 0$ for any $\mathbf{x} \in \mathbb{R}^d$.

Spectral Theorem

Statement

Question: But when does a square matrix $\mathbf{A} \in \mathbb{R}^{d \times d}$ have a basis of eigenvectors (and, hence, is diagonalizable)?

Answer: When $\mathbf{A}$ is positive semidefinite!

But even more generally…

Spectral Theorem

Statement

Theorem (Spectral Theorem). Let $\mathbf{A} \in \mathbb{R}^{d \times d}$ be a square, *symmetric* matrix (i.e. $\mathbf{A}^\top = \mathbf{A}$). Then, $\mathbf{A}$ is diagonalizable.

That is, $\mathbf{A}$ has an orthonormal basis of d eigenvectors $\mathbf{v}_1, \ldots, \mathbf{v}_d$ in the columns of a matrix $\mathbf{V} \in \mathbb{R}^{d \times d}$, associated eigenvalues $\lambda_1, \ldots, \lambda_d$ in diagonal matrix $\mathbf{\Sigma} \in \mathbb{R}^{d \times d}$ and eigendecomposition

$$\mathbf{A} = \mathbf{V}\mathbf{\Lambda}\mathbf{V}^\top .$$

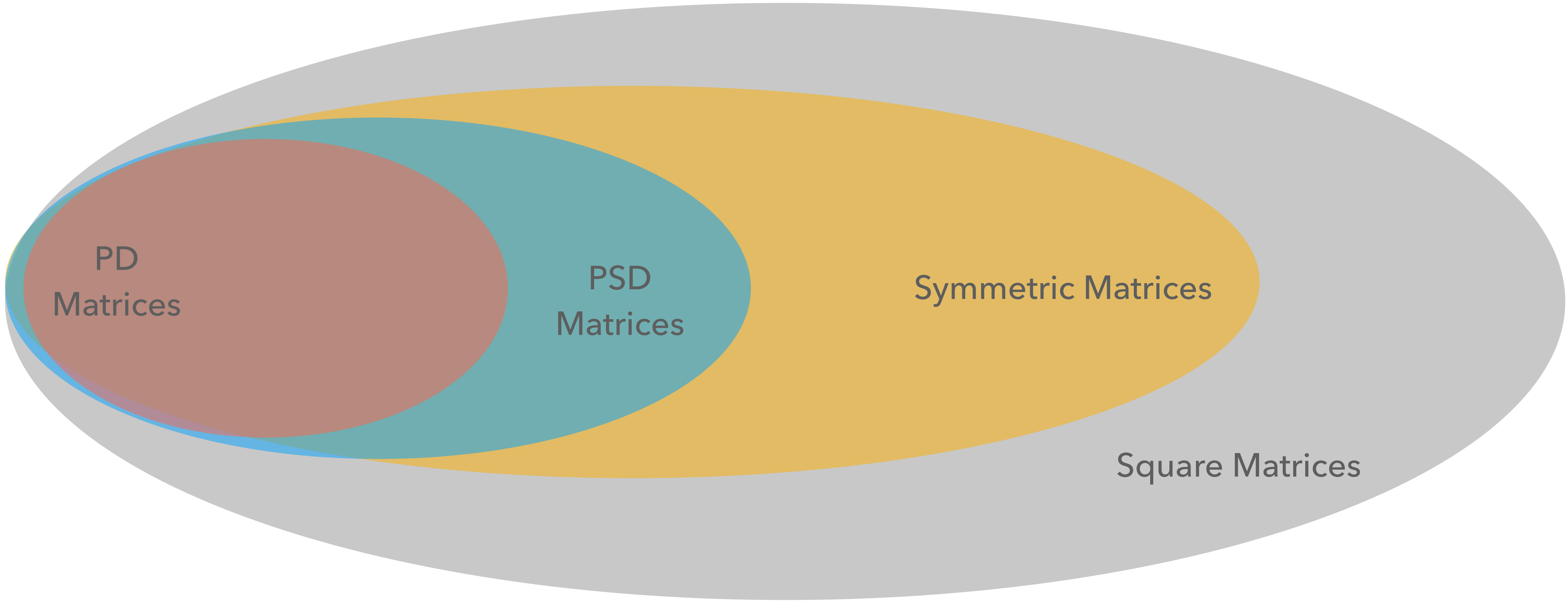
But, in this generality, λ_i can be negative!

Spectral Theorem

Statement

<u>Theorem (Spectral Theorem).</u> Let $\mathbf{A} \in \mathbb{R}^{d \times d}$ be a square, *symmetric* matrix (i.e. $\mathbf{A}^\top = \mathbf{A}$). Then, $\mathbf{A}$ is diagonalizable.

But, in this generality, λ_i can be negative!



Principal Components Analysis

Application of Eigendecomposition

Principal Components Analysis

Example: “Eigenfaces” and facial recognition

<u>Observed:</u> Matrix of *training samples* $\mathbf{X} \in \mathbb{R}^{n \times d}$ (no labels $\mathbf{y}$).

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

Each row is a “flattened” image vector. Typically, pixels are in $[0,\ 255]$ for grayscale images.

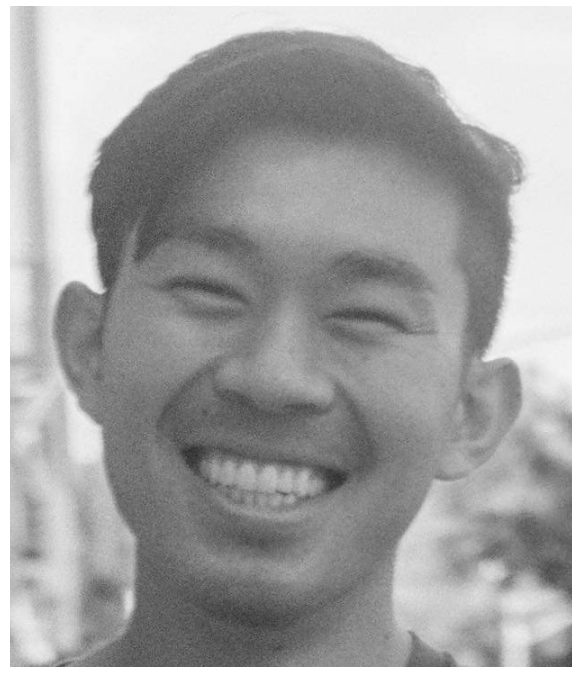
Images are very high-dimensional: $d =$ width in pixels $\times$ height in pixels.

Example: a 1080×1080 image has $d = 1080 \times 1080 = 1{,}166{,}400$.

Principal Components Analysis

Example: "Eigenfaces" and facial recognition

Consider a dataset of 1,000 grayscale face images $\mathbf{x}_1, \ldots, \mathbf{x}_{1000} \in \mathbb{R}^{1080 \times 1080} \ldots$

e.g. $\mathbf{x}_1 =$ 

Naive facial recognition: Get a new face, linear search over $1{,}000$ faces for the "closest" face (perhaps in Euclidean norm $\|\mathbf{x} - \mathbf{x}_i\|$).

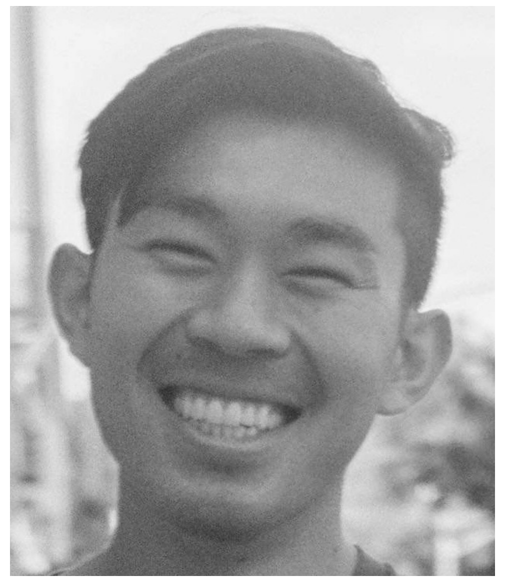
Storage: 1166400 integers $\times\, 1000$ images $\approx$ 1 GB.

Principal Components Analysis

Example: "Eigenfaces" and facial recognition

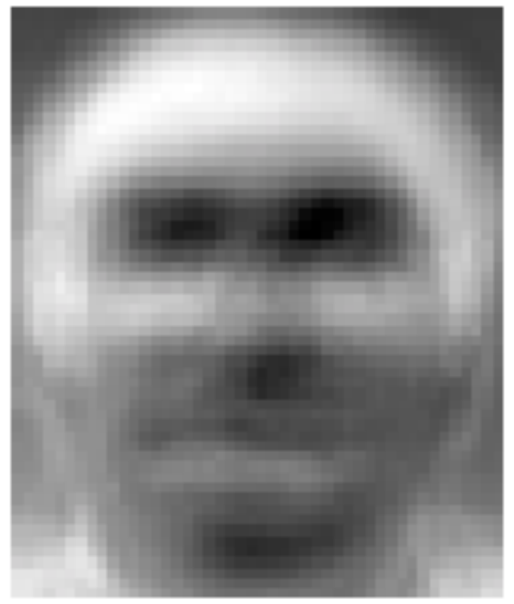
Suppose we can find a "basis" of representative faces: $\mathbf{v}_1, \ldots, \mathbf{v}_k$ where $k \ll n$.

Then, we can represent any face as a linear combination of the basis faces!

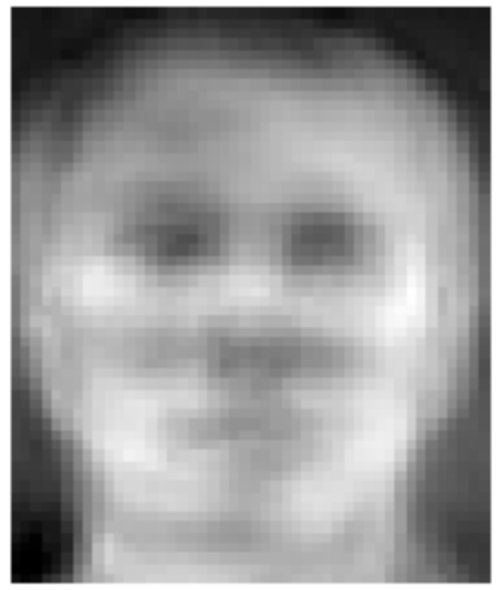


$=\ 0.45$

$+\ 0.21$



$+\ 0.12$



$+\ 0.05$

$+\ldots$

Principal Components Analysis

Example: “Eigenfaces” and facial recognition

Basis of eigenfaces: $\mathbf{v}_1, \ldots, \mathbf{v}_k$ where $k \ll n$ for subspace $\mathscr{V}$ with $\dim(\mathscr{V}) = k$.

Improved facial recognition:

Store the *projection* of n faces $\Pi_{\mathscr{V}}(\mathbf{x}_i)$ for each $\mathbf{x}_i$ in our dataset of faces.

Given a new face $\mathbf{x}_0$, project the face onto the eigenface subsapce $\mathscr{V}$ to get $\Pi_{\mathscr{V}}(\mathbf{x}_0)$.

Compare $\Pi_{\mathscr{V}}(\mathbf{x}_0)$ to each projected face in dataset in Euclidean norm $\|\Pi(\mathbf{x}_0) - \Pi(\mathbf{x}_i)\|$.

Principal Components Analysis

Example: "Eigenfaces" and facial recognition

What is this basis?

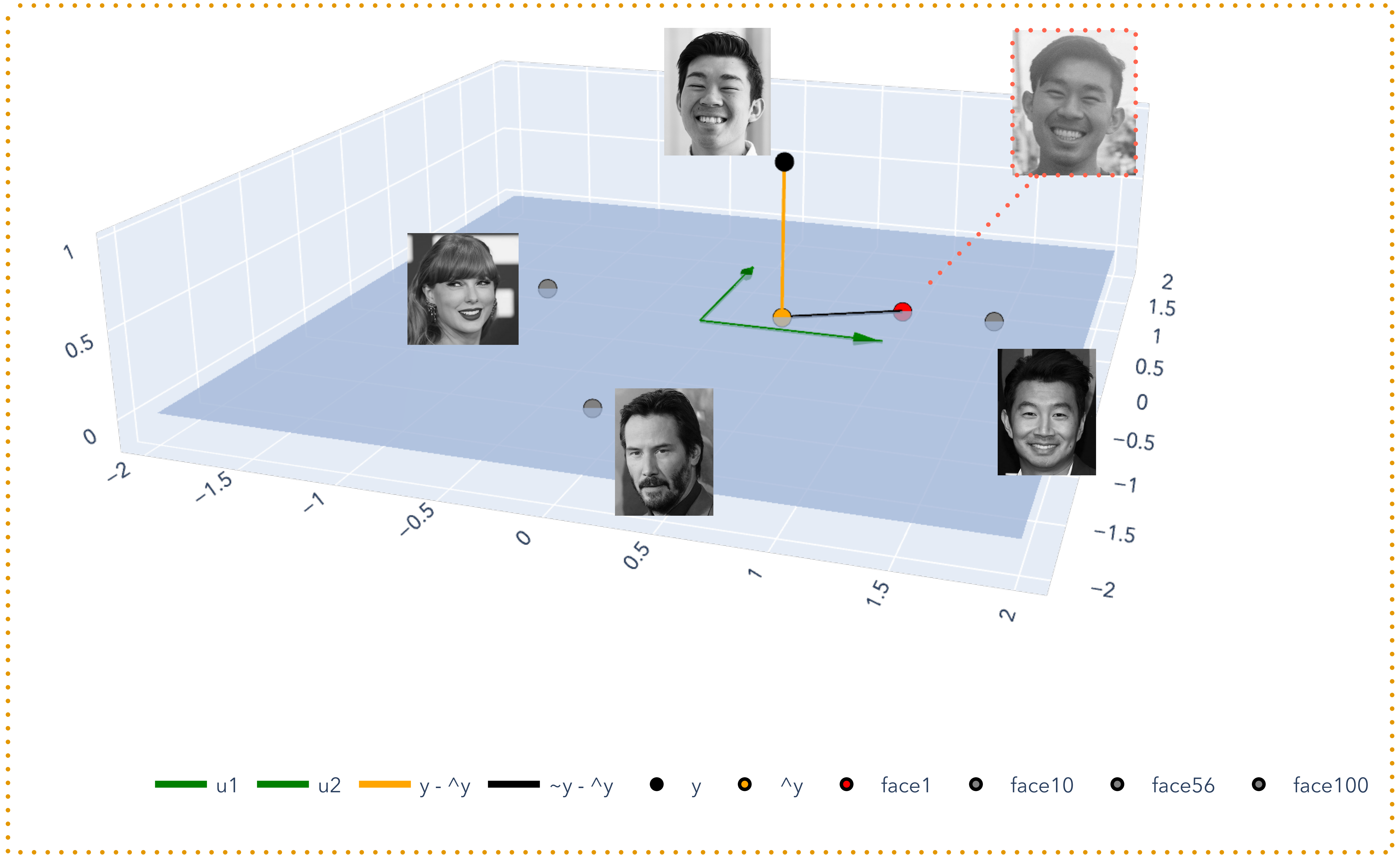
Basis of eigenfaces: $\mathbf{v}_1, \ldots, \mathbf{v}_k$ where $k \ll n$ for subspace $\mathscr{V}$ with $\dim(\mathscr{V}) = k$.

Improved facial recognition:

Store the *projection* of n faces $\Pi_{\mathscr{V}}(\mathbf{x}_i)$ for each $\mathbf{x}_i$ in our dataset of faces.

Given a new face $\mathbf{x}_0$, project the face onto the eigenface subsapce $\mathscr{V}$ to get $\Pi_{\mathscr{V}}(\mathbf{x}_0)$.

Compare $\Pi_{\mathscr{V}}(\mathbf{x}_0)$ to each projected face in dataset in Euclidean norm $||\Pi(\mathbf{x}_0) - \Pi(\mathbf{x}_i)||$.



Principal Components Analysis

Example: PCA in 2D

<u>Observed:</u> Matrix of *training points* $\mathbf{X} \in \mathbb{R}^{n \times 2}$, with columns $\mathbf{x}_1$ and $\mathbf{x}_2$.

$$\mathbf{X} = \begin{bmatrix} x_{11} & x_{12} \\ x_{21} & x_{22} \\ \vdots & \vdots \\ x_{n1} & x_{n2} \end{bmatrix}.$$

Want to find the directions that most explain the "variance" of the data.

Principal Components Analysis

Example: PCA in 2D

<u>Observed:</u> Matrix of *training points* $\mathbf{X} \in \mathbb{R}^{n \times 2}$, with columns $\mathbf{x}_1$ and $\mathbf{x}_2$.

$$\mathbf{X} = \begin{bmatrix} x_{11} & x_{12} \\ x_{21} & x_{22} \\ \vdots & \vdots \\ x_{n1} & x_{n2} \end{bmatrix}.$$

Want to find the directions that most explain the "variance" of the data.

The matrix $\mathbf{C} = \mathbf{X}^\top \mathbf{X} \in \mathbb{R}^{2 \times 2}$ is the (unnormalized) covariance matrix of the data.

Principal Components Analysis

Example: PCA in 2D

<u>Observed:</u> Matrix of *training points* $\mathbf{X} \in \mathbb{R}^{n \times 2}$, with columns $\mathbf{x}_1$ and $\mathbf{x}_2$.

$$\mathbf{X} = \begin{bmatrix} x_{11} & x_{12} \\ x_{21} & x_{22} \\ \vdots & \vdots \\ x_{n1} & x_{n2} \end{bmatrix}.$$

The matrix $\mathbf{C} = \mathbf{X}^\top \mathbf{X} \in \mathbb{R}^{2 \times 2}$ is the (unnormalized) covariance matrix of the data.

$$\mathbf{C} = \begin{bmatrix} \mathbf{x}_1^\top \mathbf{x}_1 & \mathbf{x}_1^\top \mathbf{x}_2 \\ \mathbf{x}_1^\top \mathbf{x}_2 & \mathbf{x}_2^\top \mathbf{x}_2 \end{bmatrix} = \begin{bmatrix} \|\mathbf{x}_1\|^2 & \mathbf{x}_1^\top \mathbf{x}_2 \\ \mathbf{x}_1^\top \mathbf{x}_2 & \|\mathbf{x}_2\|^2 \end{bmatrix}$$

Principal Components Analysis

Example: PCA in 2D

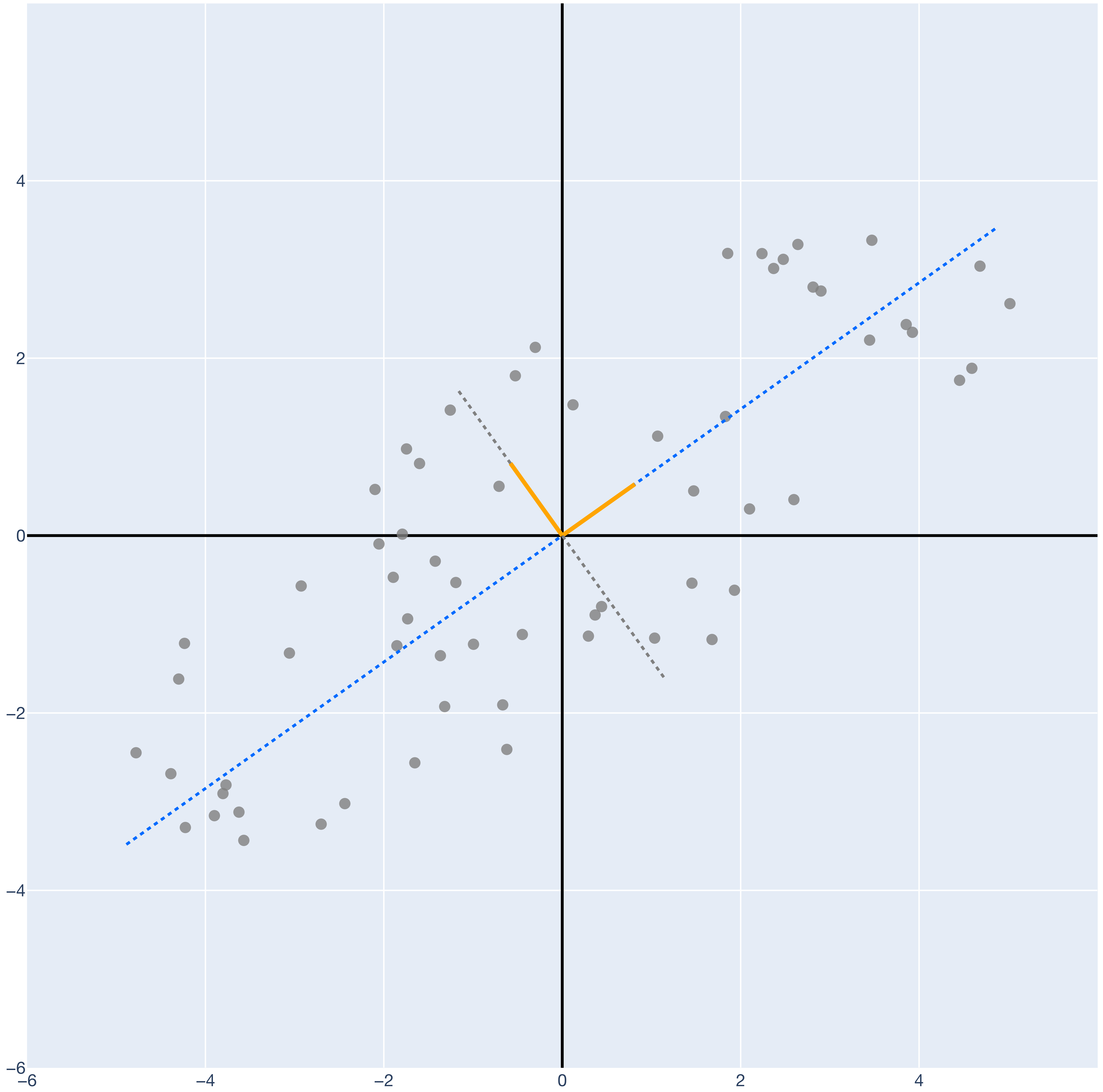
<u>Observed:</u> Matrix of *training points* $\mathbf{X} \in \mathbb{R}^{n \times 2}$, with columns $\mathbf{x}_1$ and $\mathbf{x}_2$.

$$\mathbf{X} = \begin{bmatrix} x_{11} & x_{12} \\ x_{21} & x_{22} \\ \vdots & \vdots \\ x_{n1} & x_{n2} \end{bmatrix}.$$

The matrix $\mathbf{C} = \mathbf{X}^\top \mathbf{X} \in \mathbb{R}^{2 \times 2}$ is the *covariance matrix* of the data.

$$\mathbf{C} = \begin{bmatrix} \mathbf{x}_1^\top \mathbf{x}_1 & \mathbf{x}_1^\top \mathbf{x}_2 \\ \mathbf{x}_1^\top \mathbf{x}_2 & \mathbf{x}_2^\top \mathbf{x}_2 \end{bmatrix} = \begin{bmatrix} \|\mathbf{x}_1\|^2 & \mathbf{x}_1^\top \mathbf{x}_2 \\ \mathbf{x}_1^\top \mathbf{x}_2 & \|\mathbf{x}_2\|^2 \end{bmatrix}$$

PCA: Find the ordered set of vectors $\mathbf{v}_1, \ldots, \mathbf{v}_d \in \mathbb{R}^d$ that explain the most variance to least variance in the data.



Derivation of PCA

Eigendecomposition and PCA

PCA = Eigendecomposition (SVD) of the covariance matrix!

Consider a (column-centered) dataset $\mathbf{X} \in \mathbb{R}^{n \times d}$ and construct its covariance matrix $\mathbf{C} = \mathbf{X}^\top \mathbf{X} \in \mathbb{R}^{d \times d}$. By definition, $\mathbf{C}$ is positive semidefinite.

Therefore, it is diagonalizable with eigendecomposition:

$$\mathbf{C} = \mathbf{X}^\top \mathbf{X} = \mathbf{V} \boldsymbol{\Lambda} \mathbf{V}^\top, \text{ with eigenvectors } \mathbf{v}_1, \ldots, \mathbf{v}_d.$$

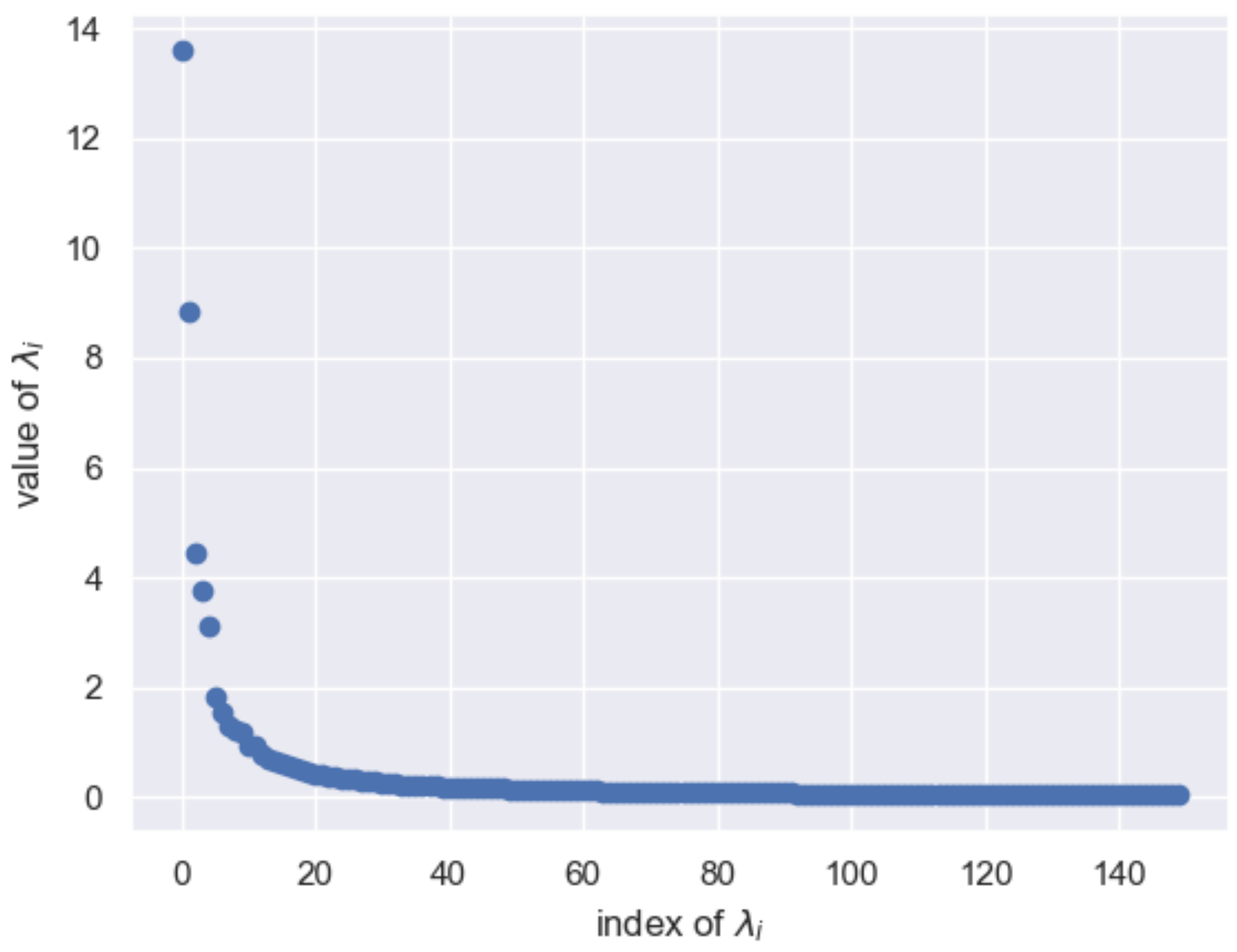
With eigenvectors ordered $\lambda_1 \geq \lambda_2 \geq \ldots \geq \lambda_d \geq 0$, choose a cutoff point $k \ll d$, and keep eigenvectors $\mathbf{v}_1, \ldots, \mathbf{v}_k$.

The eigenvectors $\mathbf{v}_1, \ldots, \mathbf{v}_k$ give an orthonormal basis for a k-dimensional subspace.

Derivation of PCA

Eigendecomposition and PCA

…with eigenvectors ordered $\lambda_1 \geq \lambda_2 \geq \ldots \geq \lambda_d \geq 0$, choose a cutoff point $k \ll d$, and keep eigenvectors $\mathbf{v}_1, \ldots, \mathbf{v}_k$.



Derivation of PCA

Eigendecomposition and PCA

PCA = Eigendecomposition (SVD) of the covariance matrix!

Consider a (column-centered) dataset $\mathbf{X} \in \mathbb{R}^{n \times d}$ and construct its covariance matrix $\mathbf{C} = \mathbf{X}^\top \mathbf{X} \in \mathbb{R}^{d \times d}$. By definition, $\mathbf{C}$ is positive semidefinite.

Therefore, it is diagonalizable with eigendecomposition:

$$\mathbf{C} = \mathbf{X}^\top \mathbf{X} = \mathbf{V} \boldsymbol{\Lambda} \mathbf{V}^\top.$$

(Could have also just taken the right singular vectors of $\mathbf{X} = \mathbf{U} \boldsymbol{\Sigma} \mathbf{V}^\top$ if we have efficient algorithm to find the SVD – true in practice).

Least Squares

Interpretation of Eigenvalues

Regression

Setup (Feature View)

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

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

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

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

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

Regression

Setup (Example View)

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

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

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

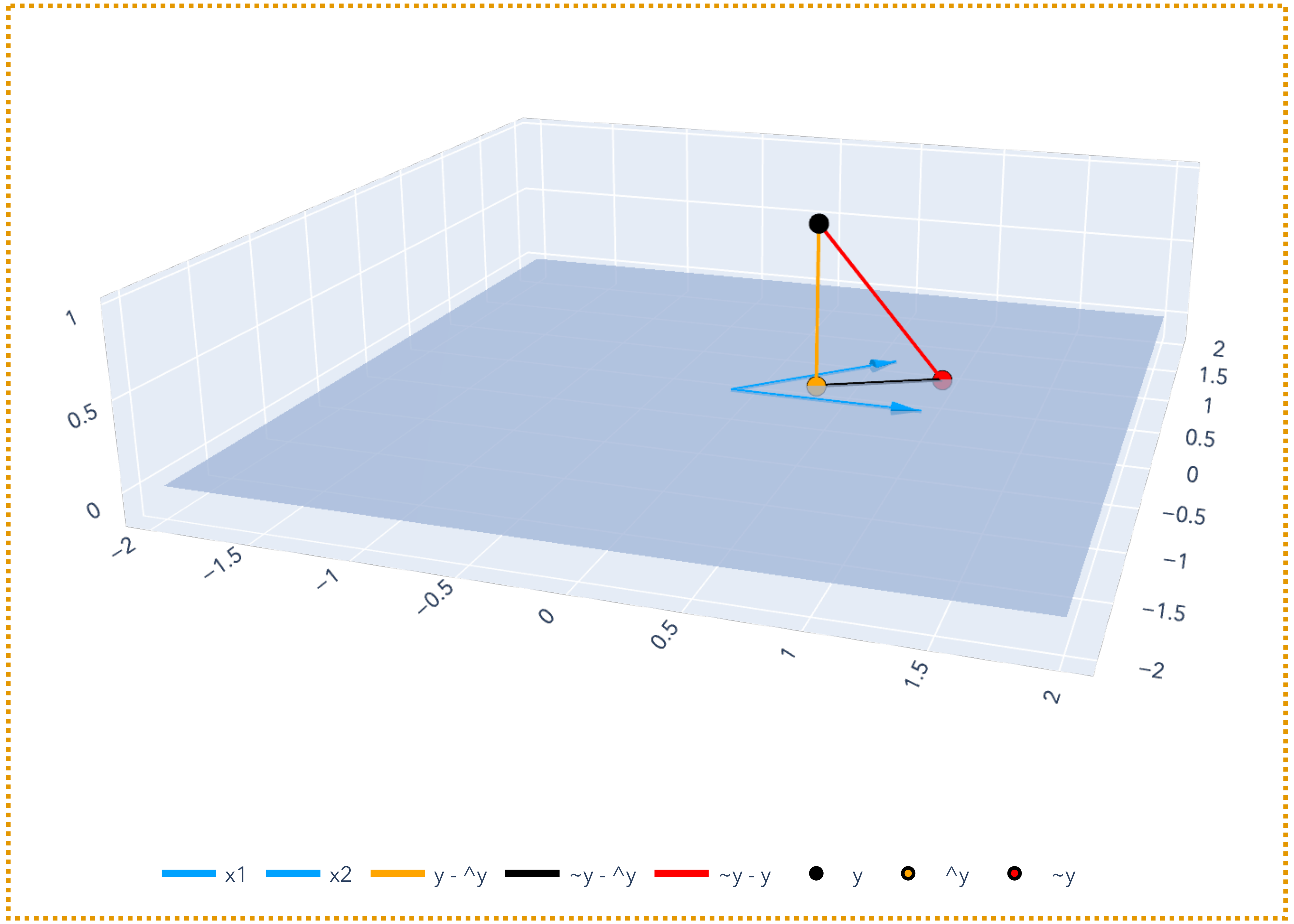
To find $\hat{\mathbf{w}}$, we follow the *principle of least squares*.

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

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



Error in Regression

Error using least squares model

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:

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

But $\hat{\mathbf{y}}$ might not be a perfect fit to $\mathbf{y}$!

Model this using a *true weight vector* $\mathbf{w}^* \in \mathbb{R}^d$ and an *error term* $\epsilon = (\epsilon_1, \ldots, \epsilon_n) \in \mathbb{R}^n$.

$$y_i = \mathbf{x}_i^\top \mathbf{w}^* + \epsilon_i \text{ for all } i \in [n] \text{ (here } \mathbf{x}_i \text{ are rows)}$$

$$\mathbf{y} = \mathbf{X}\mathbf{w}^* + \epsilon$$

Error in Regression

Error using least squares model

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:

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

But $\hat{\mathbf{y}}$ might not be a perfect fit to $\mathbf{y}$!

Model this using a *true weight vector* $\mathbf{w}^* \in \mathbb{R}^d$ and an *error term* $\epsilon = (\epsilon_1, \ldots, \epsilon_n) \in \mathbb{R}^n$.

$$y_i = \mathbf{x}_i^\top \mathbf{w}^* + \epsilon_i \text{ for all } i \in [n] \text{ (here } \mathbf{x}_i \text{ are rows)}$$

$$\mathbf{y} = \mathbf{X}\mathbf{w}^* + \epsilon$$

Error in Regression

Error using least squares model

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:

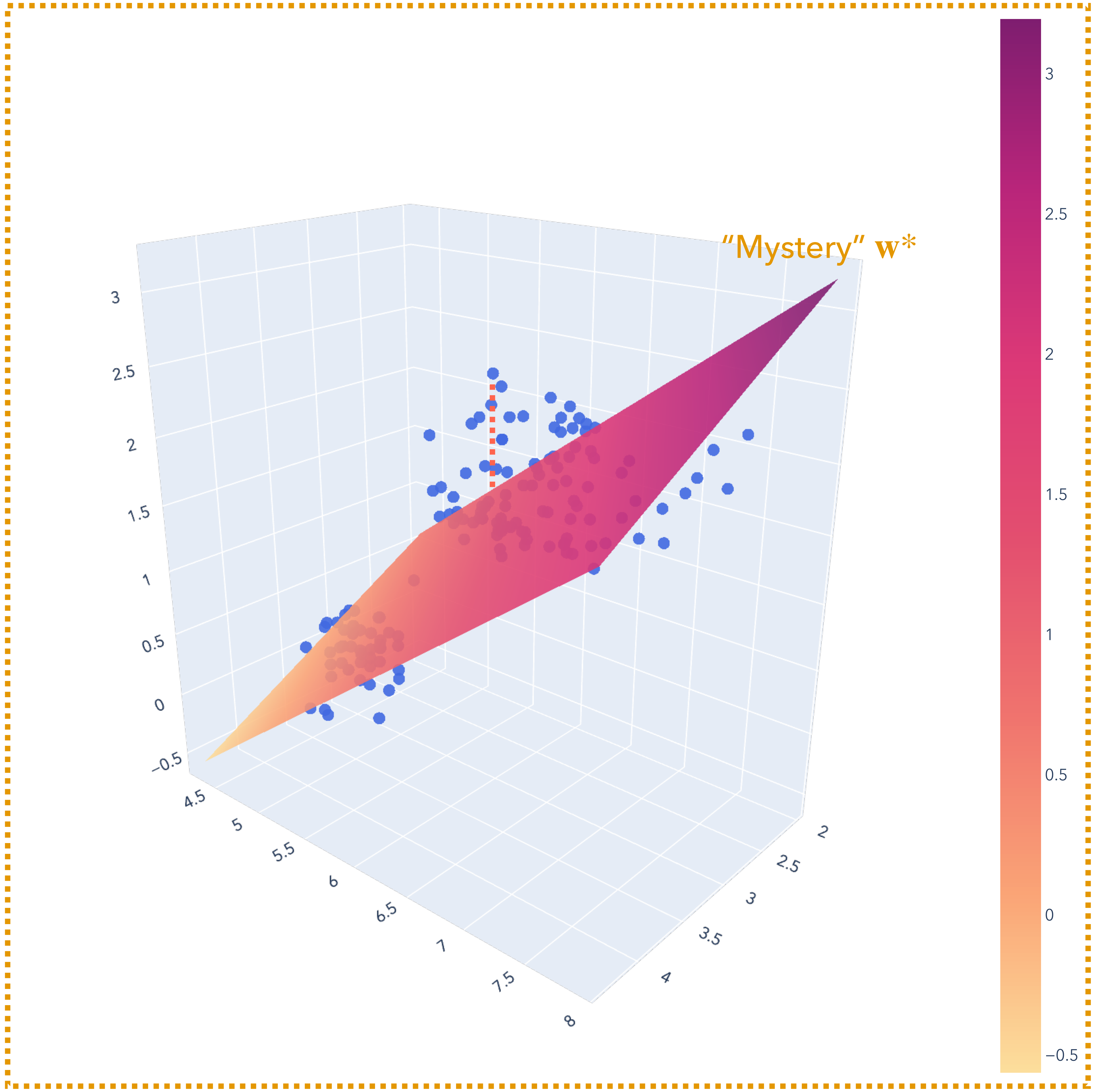
$$\mathbf{X}\hat{\mathbf{w}} = \hat{\mathbf{y}} \approx \mathbf{y}\,.$$

But $\hat{\mathbf{y}}$ might not be a perfect fit to $\mathbf{y}$!

Model this using a *true weight vector* $\mathbf{w}^* \in \mathbb{R}^d$ and an *error term* $\epsilon = (\epsilon_1, \ldots, \epsilon_n) \in \mathbb{R}^n$.

$$y_i = \mathbf{x}_i^\top \mathbf{w}^* + \epsilon_i \text{ for all } i \in [n] \text{ (here } \mathbf{x}_i \text{ are rows)}$$

$$\mathbf{y} = \mathbf{X}\mathbf{w}^* + \epsilon$$



Error in Regression

Error using least squares model

In our model of the world, true labels are given by: $\mathbf{y} = \mathbf{X}\mathbf{w}^* + \epsilon$.

What happens when we use the least squares weights $\hat{\mathbf{w}} = (\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{X}^\top\mathbf{y}$?

$$
\begin{aligned}
\hat{\mathbf{w}} &= (\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{X}^\top\mathbf{y} \\
&= (\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{X}^\top(\mathbf{X}\mathbf{w}^* + \epsilon) \\
&= (\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{X}^\top\mathbf{X}\mathbf{w}^* + (\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{X}^\top\epsilon \\
&= \mathbf{w}^* + (\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{X}^\top\epsilon
\end{aligned}
$$

Error in Regression

Error using least squares model

In our model of the world, true labels are given by: $\mathbf{y} = \mathbf{X}\mathbf{w}^* + \epsilon$.

What happens when we use the least squares weights $\hat{\mathbf{w}} = (\mathbf{X}^\top \mathbf{X})^{-1}\mathbf{X}^\top \mathbf{y}$?

$$
\begin{aligned}
\hat{\mathbf{w}} &= (\mathbf{X}^\top \mathbf{X})^{-1}\mathbf{X}^\top \mathbf{y} \\
&= (\mathbf{X}^\top \mathbf{X})^{-1}\mathbf{X}^\top (\mathbf{X}\mathbf{w}^* + \epsilon) \\
&= (\mathbf{X}^\top \mathbf{X})^{-1}\mathbf{X}^\top \mathbf{X}\mathbf{w}^* + (\mathbf{X}^\top \mathbf{X})^{-1}\mathbf{X}^\top \epsilon \\
&= \mathbf{w}^* + (\mathbf{X}^\top \mathbf{X})^{-1}\mathbf{X}^\top \epsilon
\end{aligned}
$$

When $\epsilon = 0$ ($\mathbf{y}$ is linearly related to $\mathbf{X}$), this is perfect: $\hat{\mathbf{w}} = \mathbf{w}^*$!

Error in Regression

Error using least squares model

In our model of the world, true labels are given by: $\mathbf{y} = \mathbf{X}\mathbf{w}^* + \epsilon$.

What happens when we use the least squares weights $\hat{\mathbf{w}} = (\mathbf{X}^\top \mathbf{X})^{-1}\mathbf{X}^\top \mathbf{y}$?

$$\begin{aligned}
\hat{\mathbf{w}} &= (\mathbf{X}^\top \mathbf{X})^{-1}\mathbf{X}^\top \mathbf{y} \\
&= (\mathbf{X}^\top \mathbf{X})^{-1}\mathbf{X}^\top (\mathbf{X}\mathbf{w}^* + \epsilon) \\
&= (\mathbf{X}^\top \mathbf{X})^{-1}\mathbf{X}^\top \mathbf{X}\mathbf{w}^* + (\mathbf{X}^\top \mathbf{X})^{-1}\mathbf{X}^\top \epsilon \\
&= \mathbf{w}^* + (\mathbf{X}^\top \mathbf{X})^{-1}\mathbf{X}^\top \epsilon
\end{aligned}$$

When $\epsilon \neq 0$, we have the difference of $\hat{\mathbf{w}} - \mathbf{w}^* = (\mathbf{X}^\top \mathbf{X})^{-1}\mathbf{X}^\top \epsilon$.

Error in Regression

Eigendecomposition perspective

Weight vector's difference from true $\mathbf{w}^*$: $\hat{\mathbf{w}} - \mathbf{w}^* = (\mathbf{X}^\top \mathbf{X})^{-1} \mathbf{X}^\top \epsilon$.

We know that $\mathbf{X}^\top \mathbf{X}$ (the *covariance matrix*) is PSD, so it is diagonalizable:

$$\mathbf{X}^\top \mathbf{X} = \mathbf{V} \boldsymbol{\Lambda} \mathbf{V}^\top \implies (\mathbf{X}^\top \mathbf{X})^{-1} = \mathbf{V}^\top \boldsymbol{\Lambda}^{-1} \mathbf{V}\,.$$

The inverse of the diagonal matrix $\boldsymbol{\Lambda}^{-1}$:

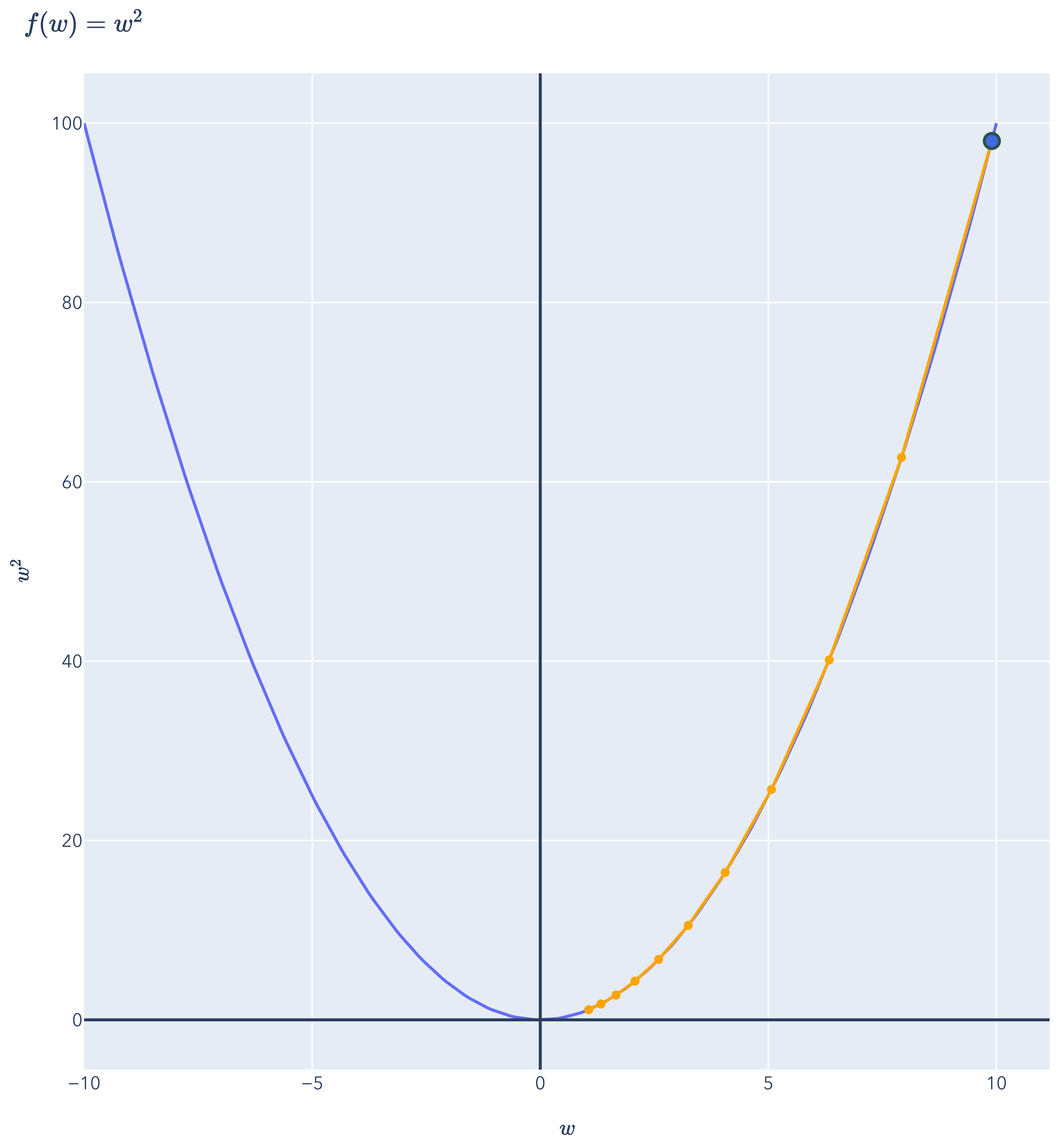
$$\boldsymbol{\Lambda}^{-1} = \begin{bmatrix} 1/\lambda_1 & \dots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \dots & 1/\lambda_d \end{bmatrix}$$, so if λ_i is small, the entries of $\hat{\mathbf{w}} - \mathbf{w}^*$ blow up!

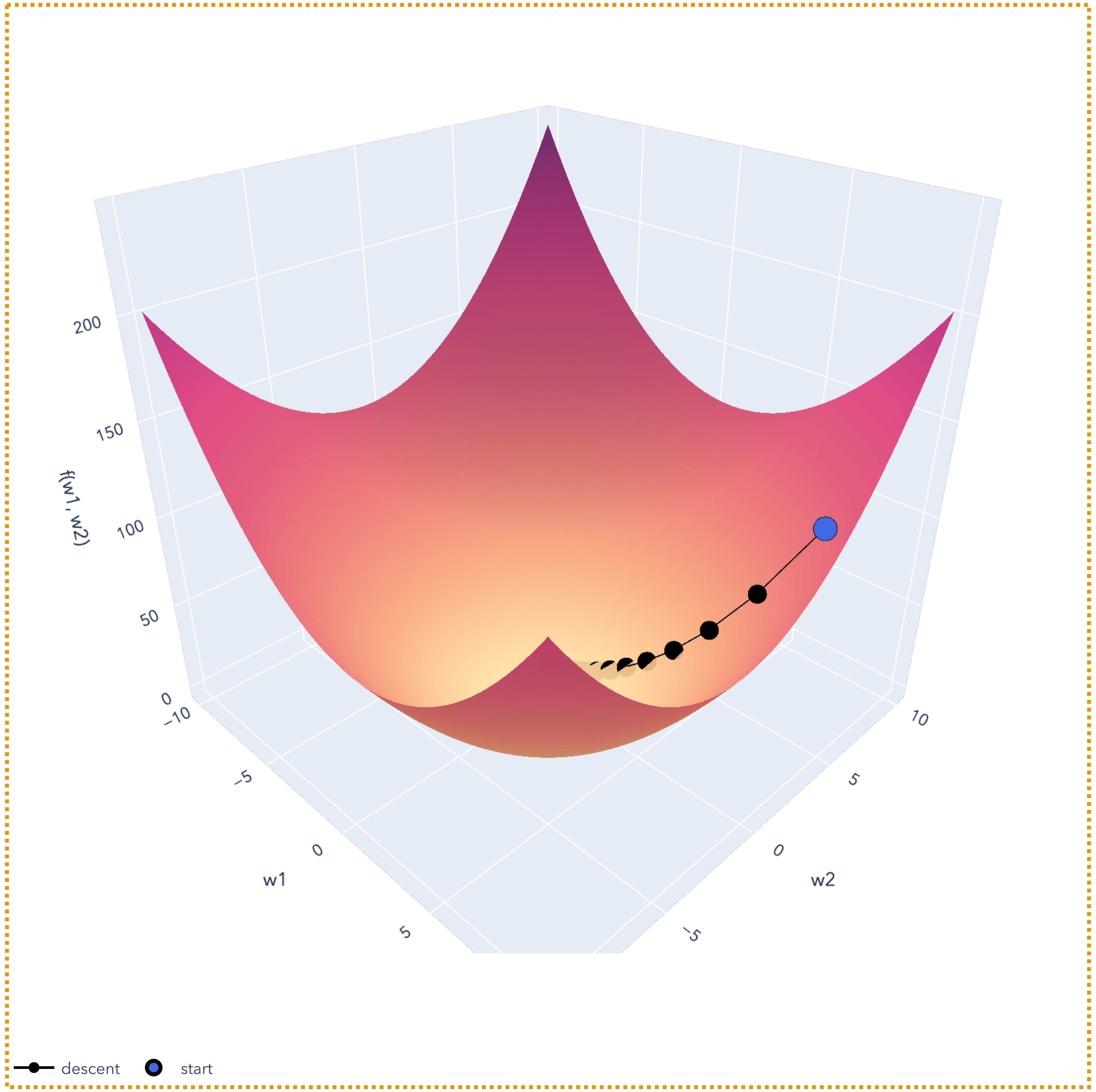
Gradient Descent

Positive Semidefinite Matrices and Convexity

Lesson Overview

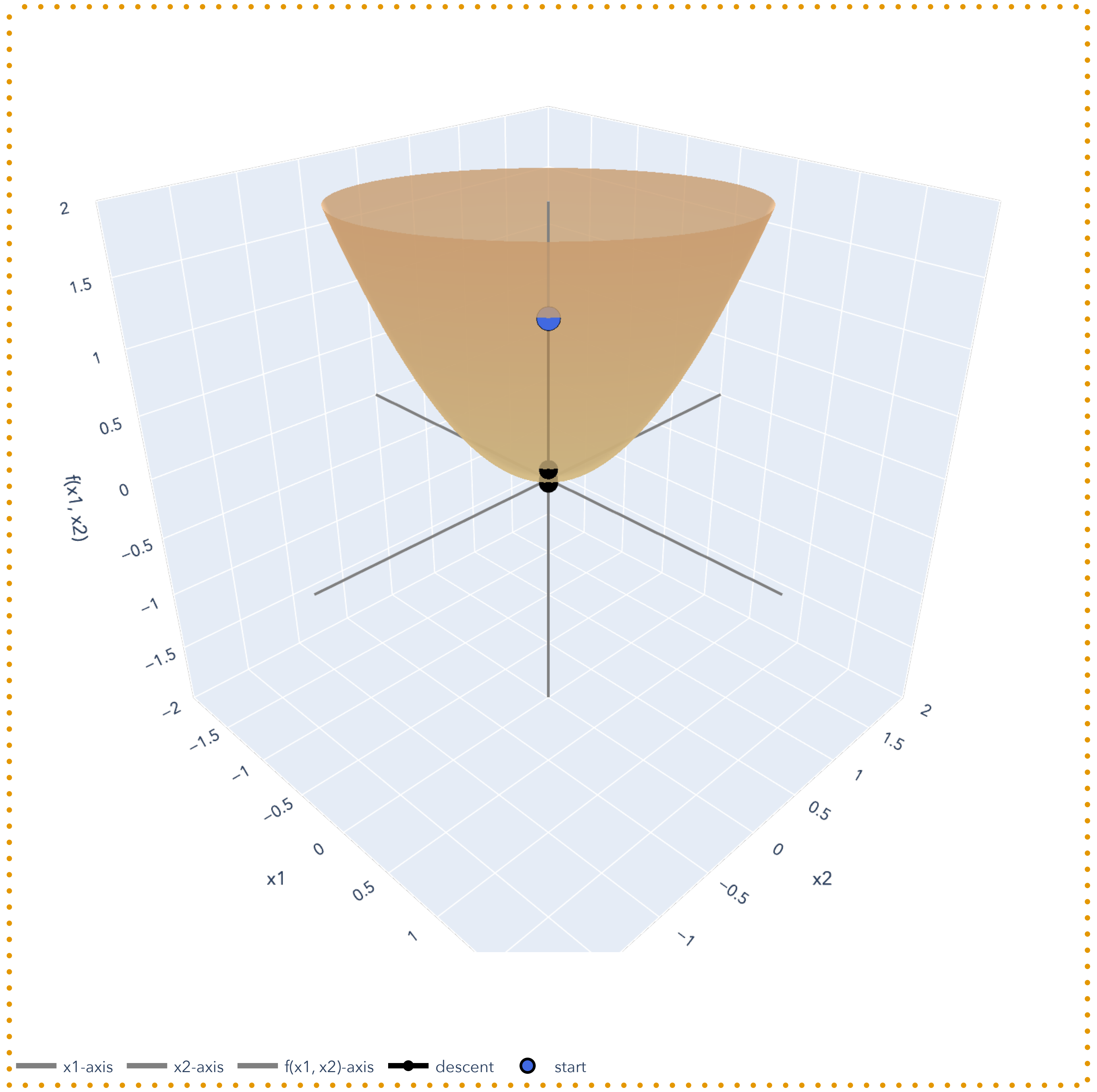
Big Picture: Gradient Descent

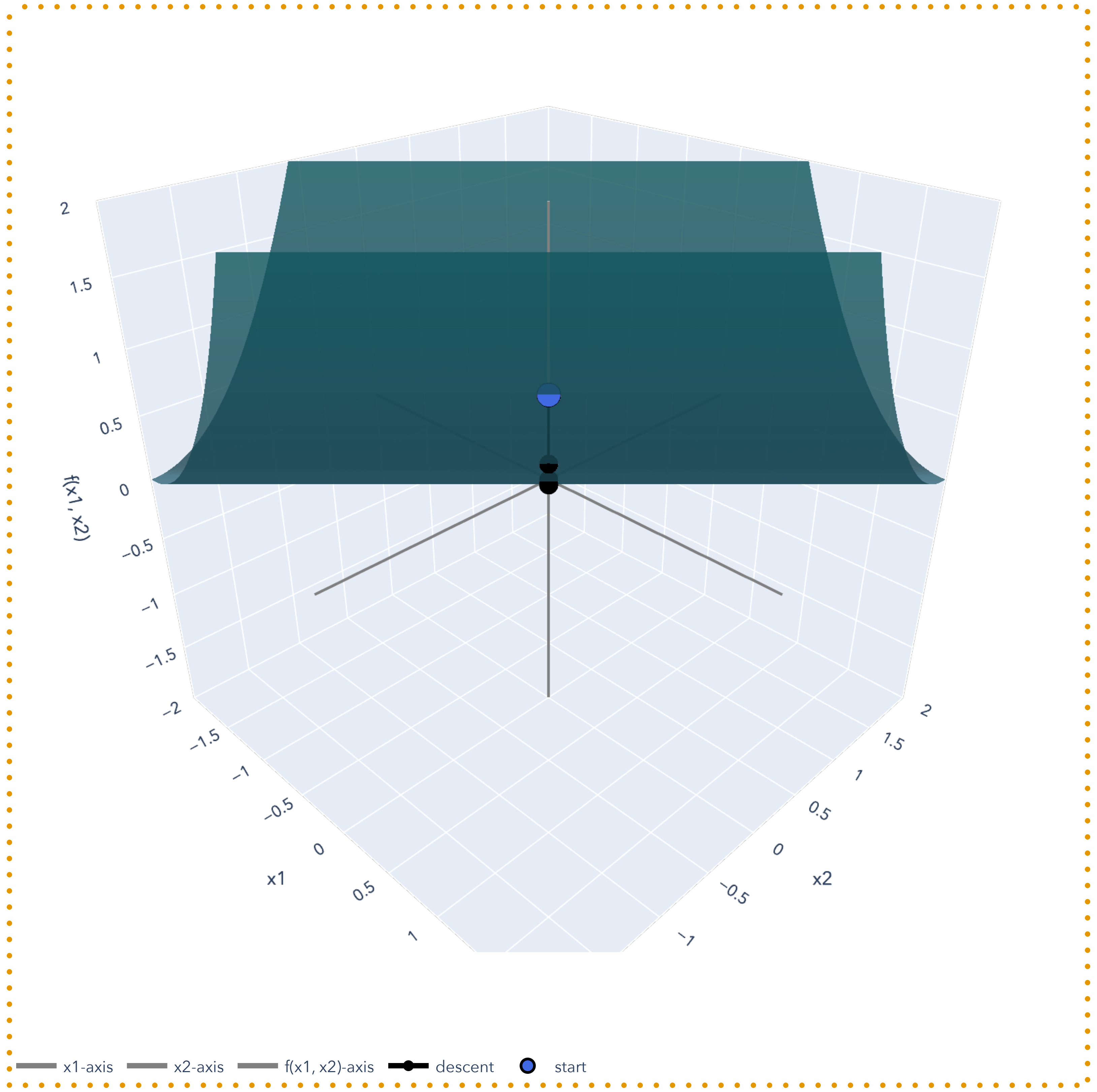


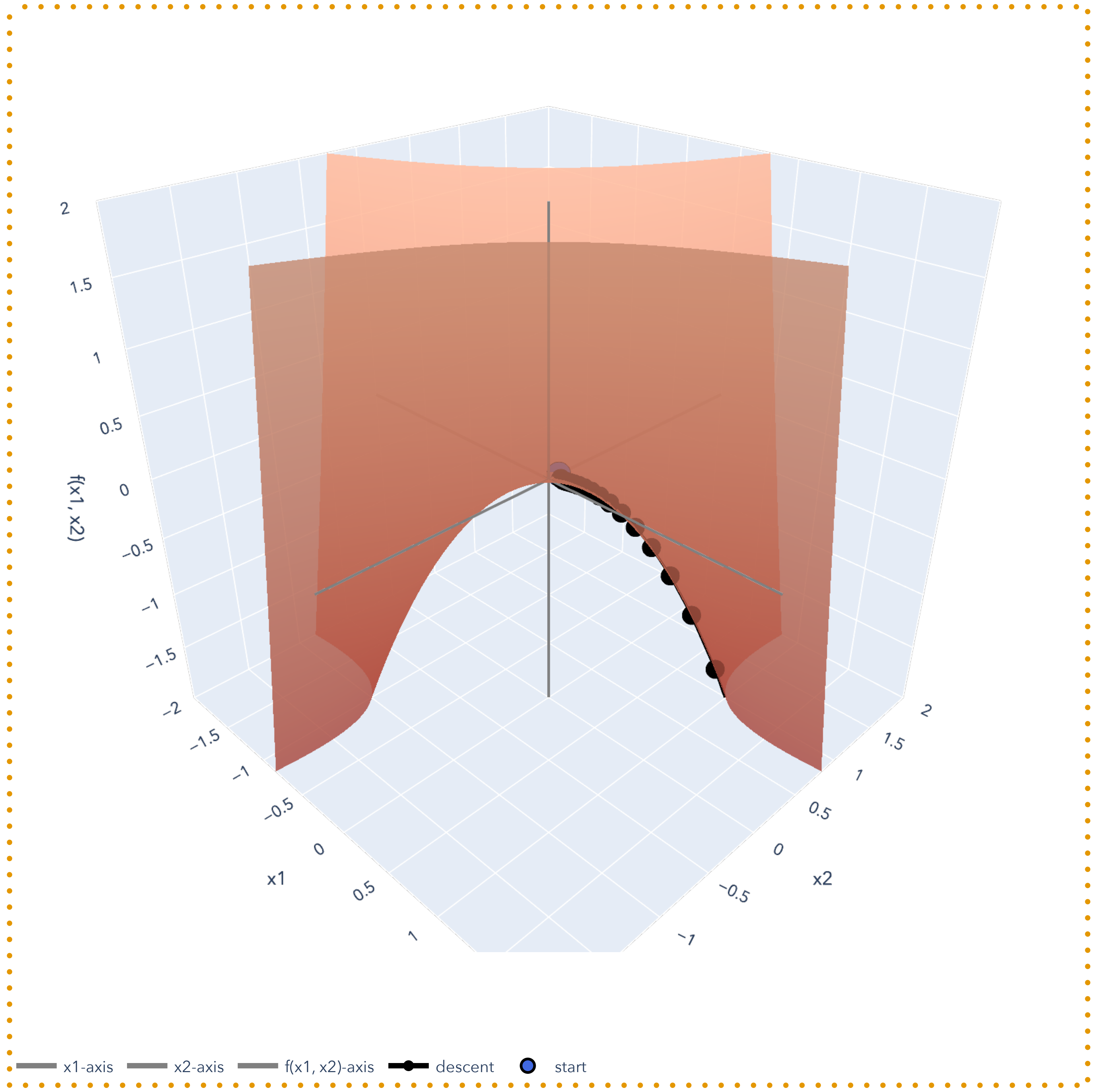


Lesson Overview

Big Picture: Gradient Descent







Quadratic Forms

2D Example

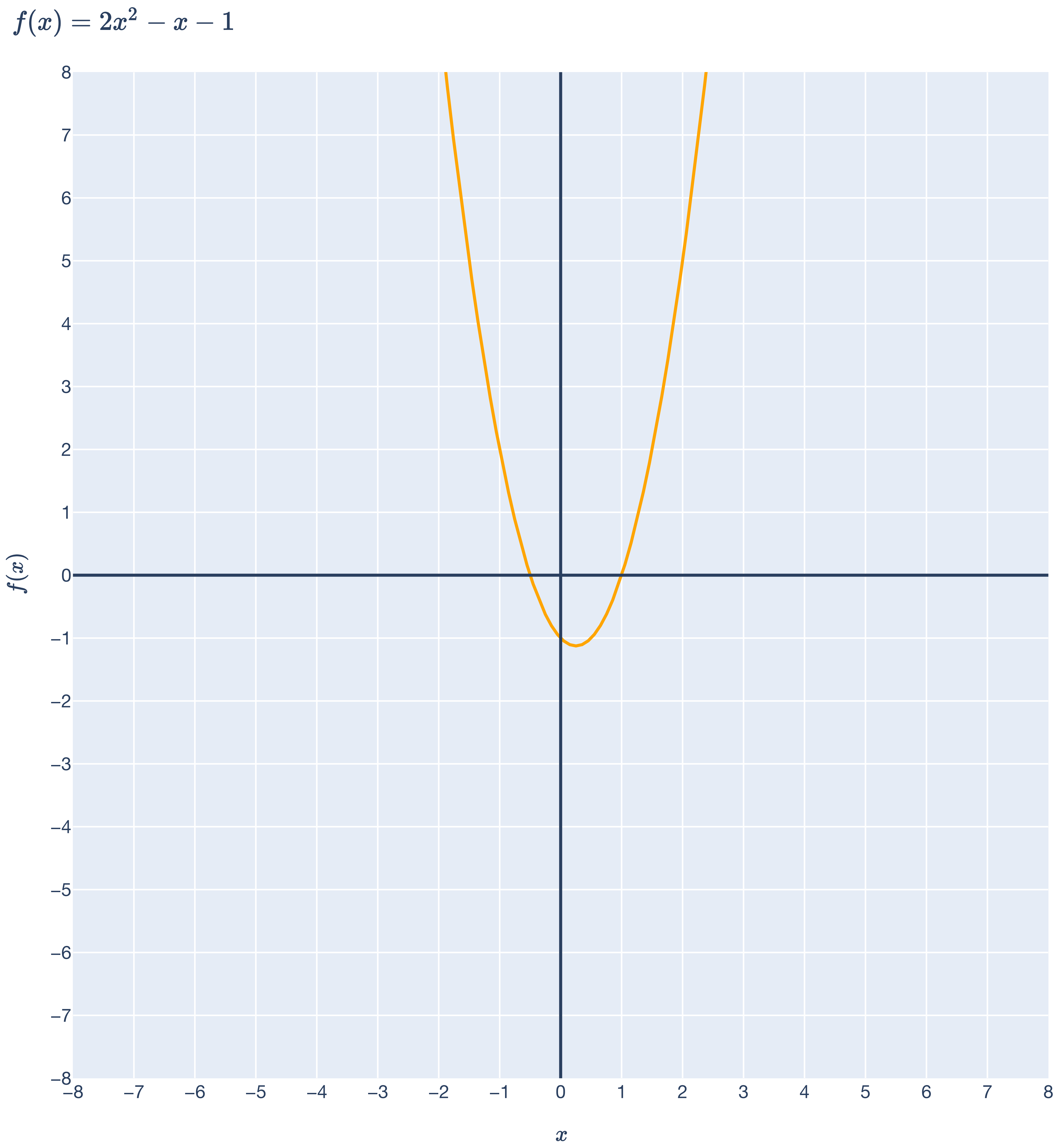
A *quadratic function* $f : \mathbb{R} \to \mathbb{R}$ has the form

$$f(x) = ax^2 + bx + c,$$

where $a, b, c \in \mathbb{R}$.

<u>Example:</u> $f(x) = 2x^2 - x - 1$

We will be concerned about finding *minima* of quadratic functions.



Quadratic Forms

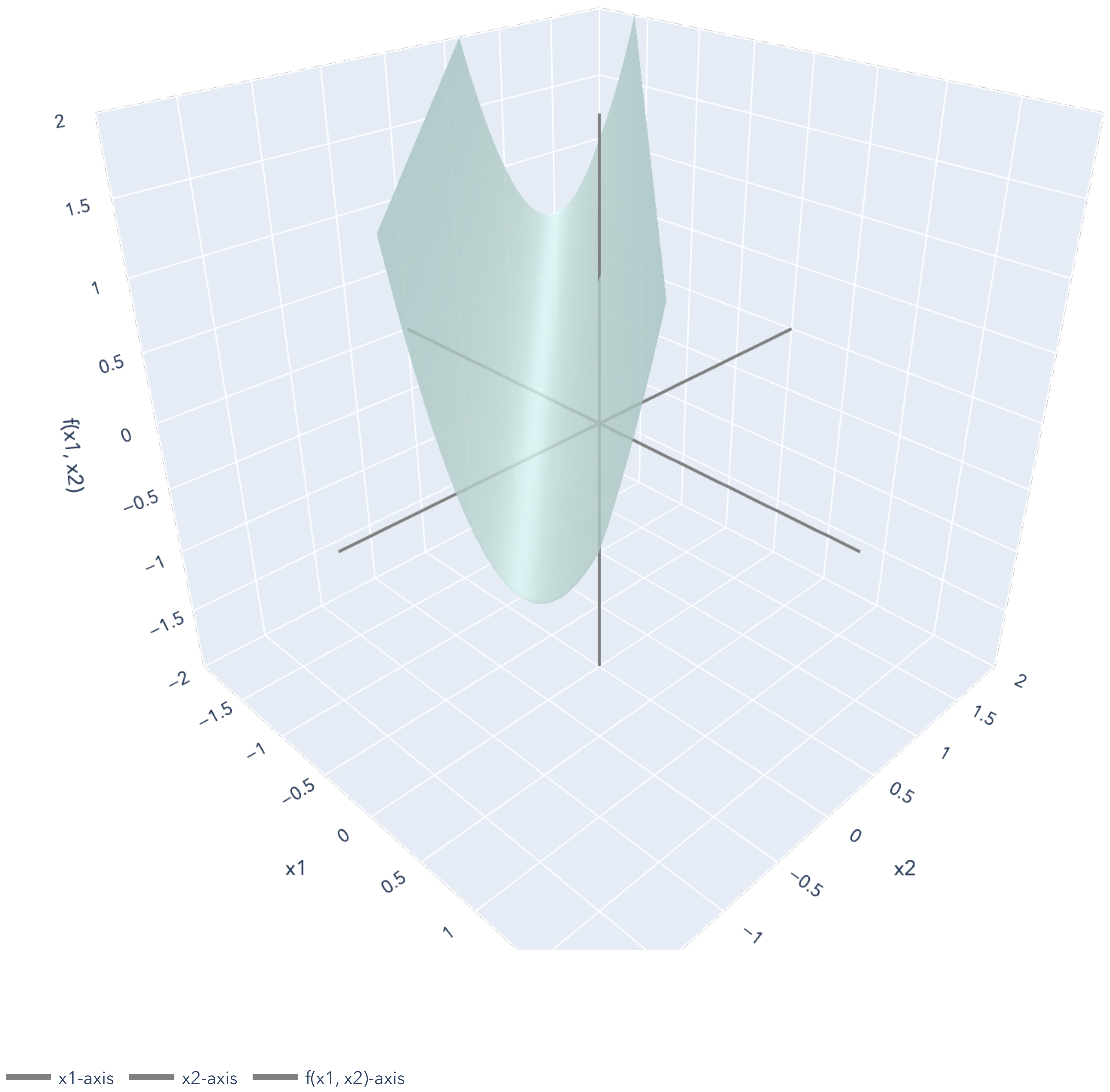
3D Example

In $d = 2$, a *quadratic function* $f : \mathbb{R}^2 \to \mathbb{R}$ has form:

$$f(x) = ax^2 + 2bxy + cy^2 + dx + ey + f,$$

where $a, b, c, d, e, f \in \mathbb{R}$ are all constants.

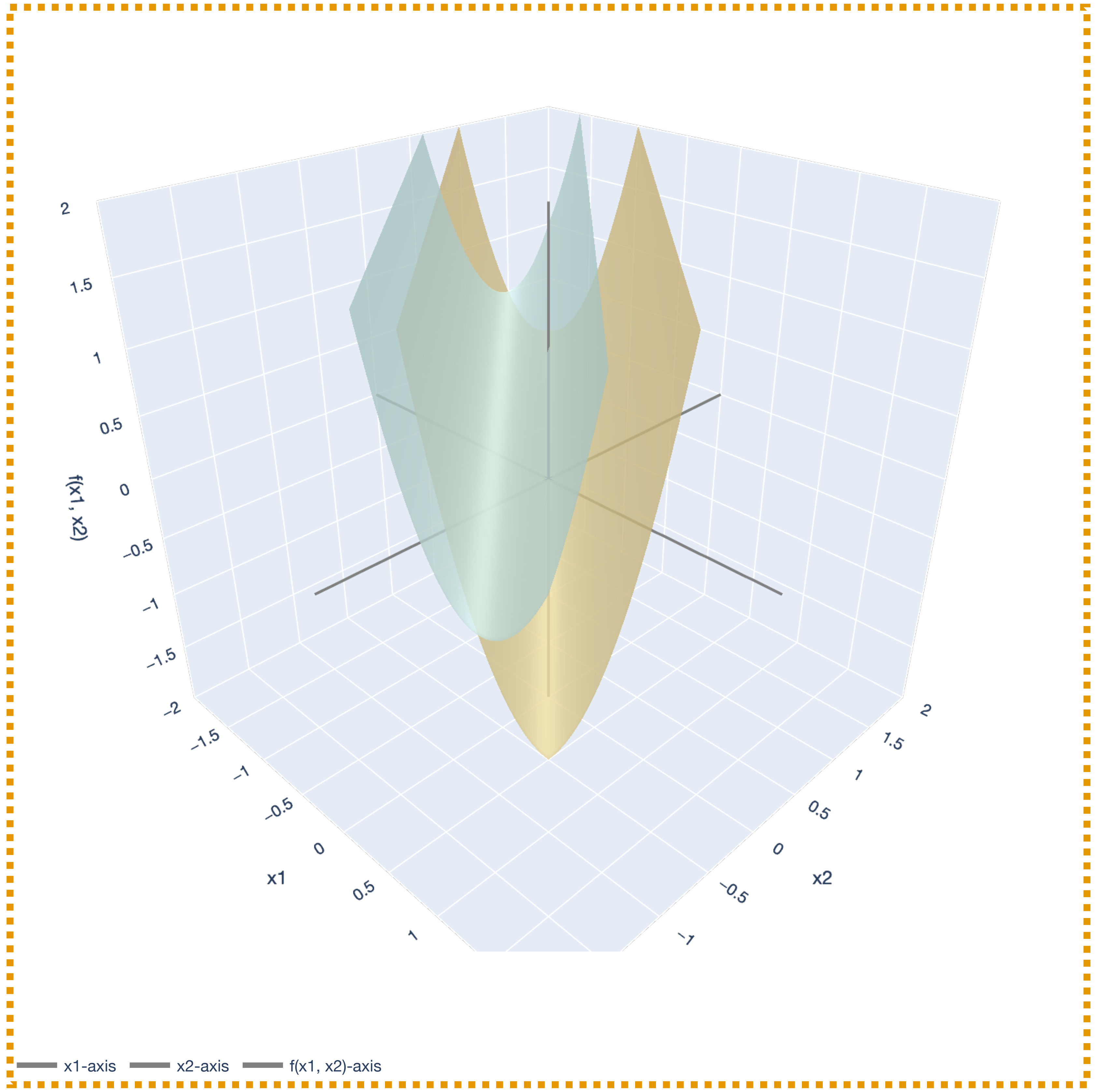
<u>Example:</u> $f(x) = 2x^2 + 4xy + 2y^2 + 2x + 2y + 1$



Quadratic Forms

3D Example

$f(x) = 2x^2 + 4xy + 2y^2 + 2x + 2y + 1$ vs. $f(x) = 2x^2 + 4xy + 2y^2$



Quadratic Forms

3D Example

In 3D, a *quadratic function* $f : \mathbb{R}^2 \to \mathbb{R}$ has the form

$$f(x) = \underbrace{ax^2 + 2bxy + cy^2}_{\text{quadratic}} + \underbrace{dx + ey}_{\text{linear}} + \underbrace{f}_{\text{constant}} .$$

Let's only examine the quadratic part!

$$f(x) = ax^2 + 2bxy + cy^2.$$

Quadratic Forms

Relationship with matrices and eigenvalues

A function $f : \mathbb{R}^2 \to \mathbb{R}$ is a quadratic form if it is a polynomial with terms of all degree two:

$$f(x) = ax^2 + 2bxy + cy^2 .$$

We can rewrite this in matrix form:

$$f(x, y) = \begin{bmatrix} x & y \end{bmatrix} \begin{bmatrix} a & b \\ b & c \end{bmatrix} \begin{bmatrix} x \\ y \end{bmatrix}$$

$$f(\mathbf{x}) = \mathbf{x}^\top \mathbf{A} \mathbf{x}$$

Quadratic Forms

Relationship with matrices and eigenvalues

Consider a quadratic form:

$$f(x, y) = \begin{bmatrix} x & y \end{bmatrix} \begin{bmatrix} a & b \\ b & c \end{bmatrix} \begin{bmatrix} x \\ y \end{bmatrix}$$

$$f(\mathbf{x}) = \mathbf{x}^\top \mathbf{A} \mathbf{x}$$

The matrix $\mathbf{A} \in \mathbb{R}^{2\times 2}$ is always symmetric, so it is diagonalizable!

$$\mathbf{A} = \mathbf{V}\boldsymbol{\Lambda}\mathbf{V}^\top, \text{ where } \boldsymbol{\Lambda} \in \mathbb{R}^{d\times d} \text{ is diagonal.}$$

Quadratic Forms

Relationship with matrices and eigenvalues

The matrix $\mathbf{A} \in \mathbb{R}^{2\times 2}$ is always symmetric, so it is diagonalizable!

$$\mathbf{A} = \mathbf{V}\mathbf{\Lambda}\mathbf{V}^\top, \text{ where } \mathbf{\Lambda} \in \mathbb{R}^{d\times d} \text{ is diagonal.}$$

$$\implies f(\mathbf{x}) = \mathbf{x}^\top\mathbf{A}\mathbf{x} = \mathbf{x}^\top\mathbf{V}\mathbf{\Lambda}\mathbf{V}^\top\mathbf{x}$$

$$\implies \bar{\mathbf{x}}^\top\mathbf{\Lambda}\bar{\mathbf{x}}, \text{ where } \bar{\mathbf{x}} = \mathbf{V}^\top\mathbf{x}.$$

$$\mathbf{\Lambda} = \begin{bmatrix} \lambda_1 & 0 \\ 0 & \lambda_2 \end{bmatrix}$$

Quadratic Forms

Relationship with matrices and eigenvalues

$\mathbf{A} = \mathbf{V}\boldsymbol{\Lambda}\mathbf{V}^\top$, where $\boldsymbol{\Lambda} \in \mathbb{R}^{d \times d}$ is diagonal.

$$\boldsymbol{\Lambda} = \begin{bmatrix} \lambda_1 & 0 \\ 0 & \lambda_2 \end{bmatrix}$$

There are three possibilities:

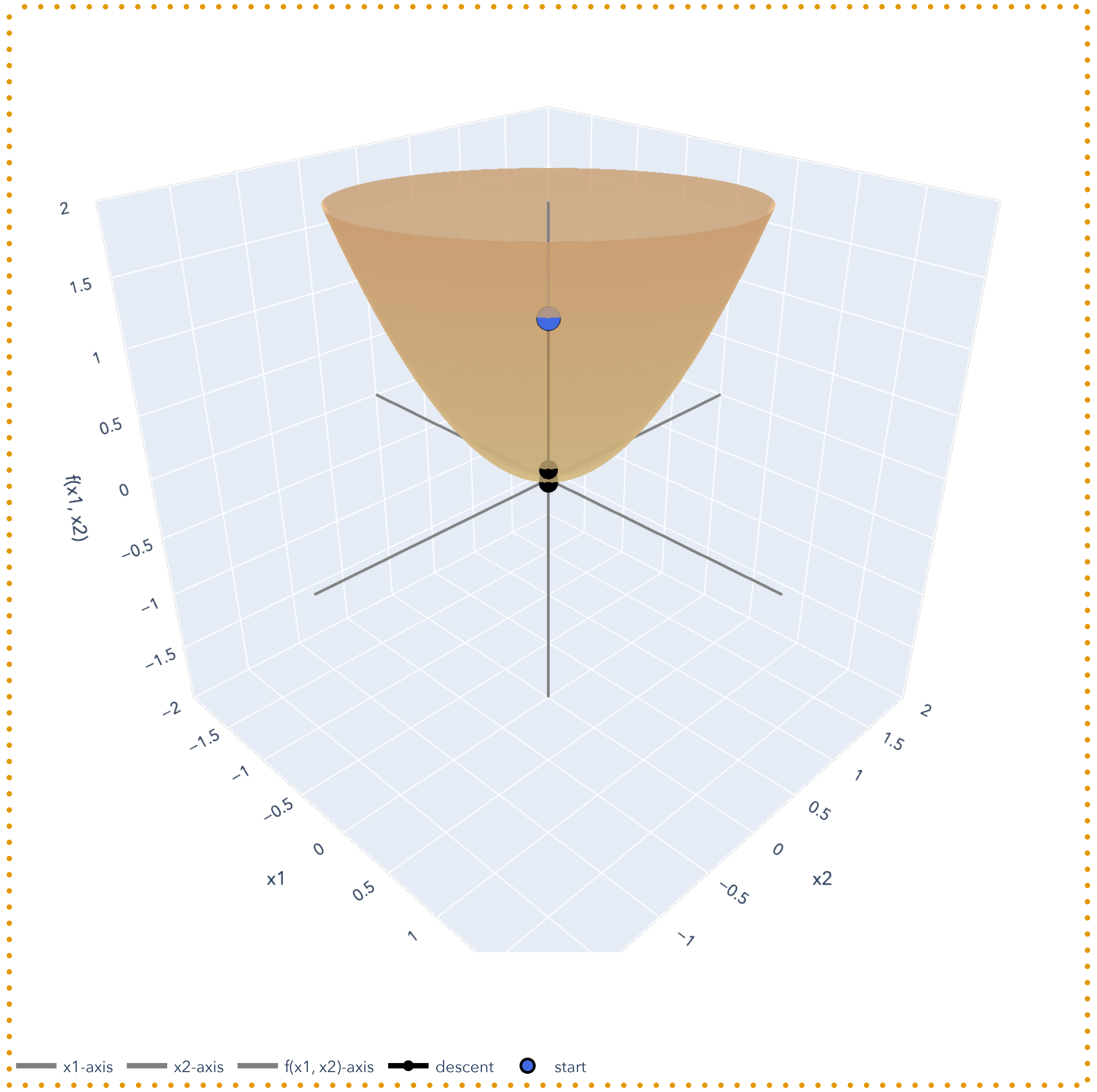
1. λ_1 and λ_2 are *both* positive (*positive definite*).

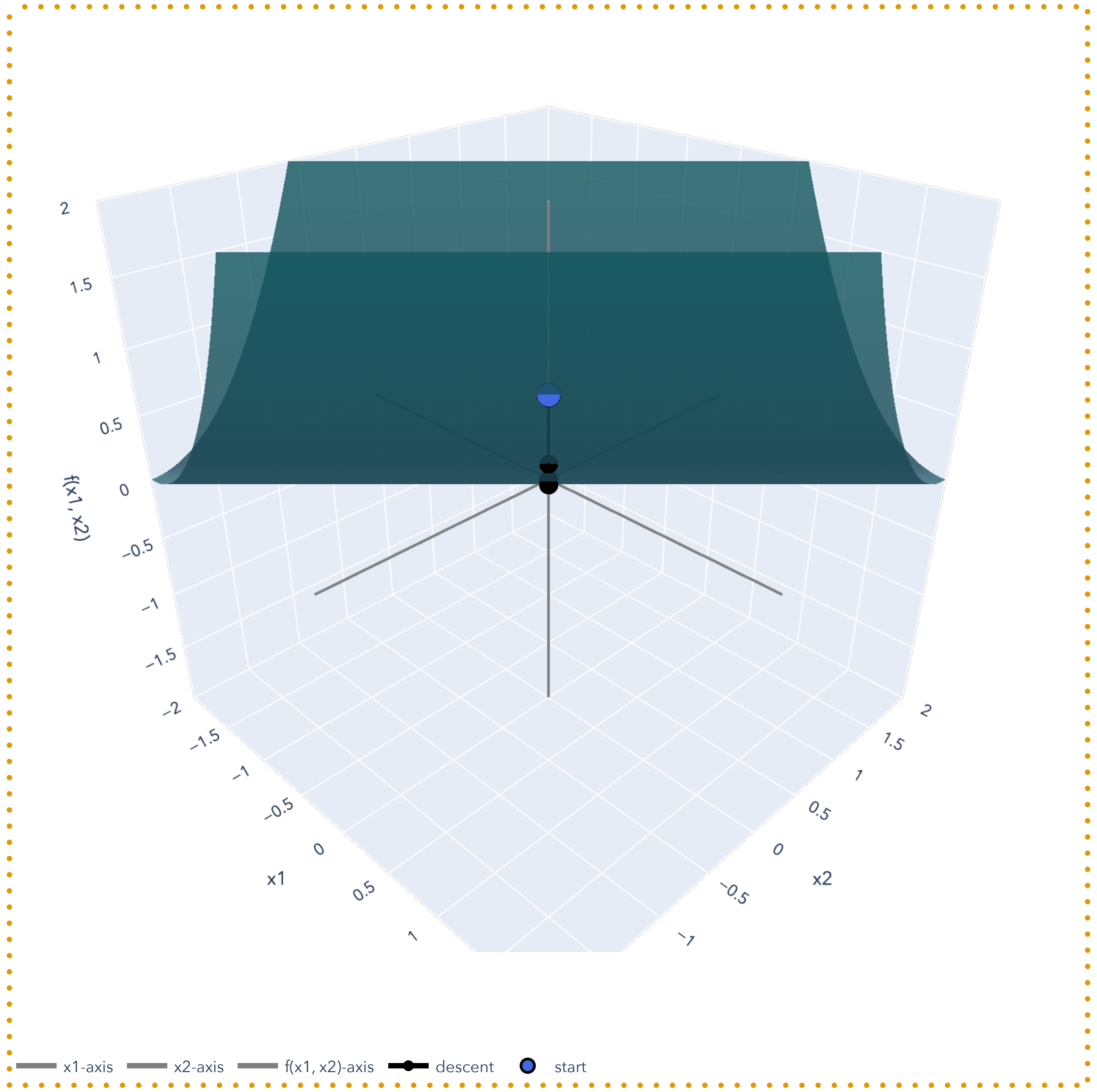
2. λ_1 or λ_2 is zero, and the other is positive (*positive semidefinite*).

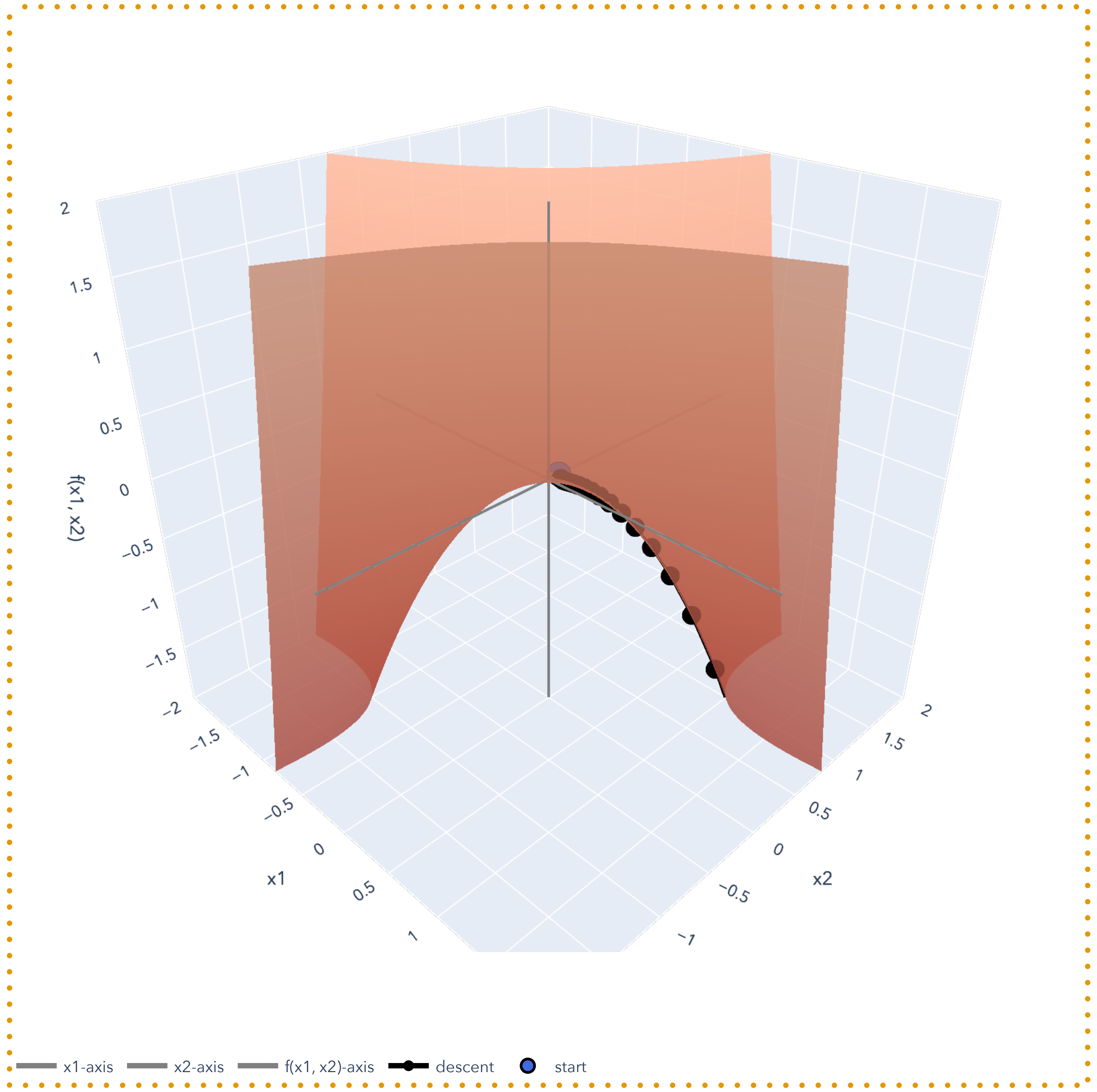
3. λ_1 or λ_2 is negative (*indefinite*).

Lesson Overview

Big Picture: Gradient Descent







Quadratic Forms

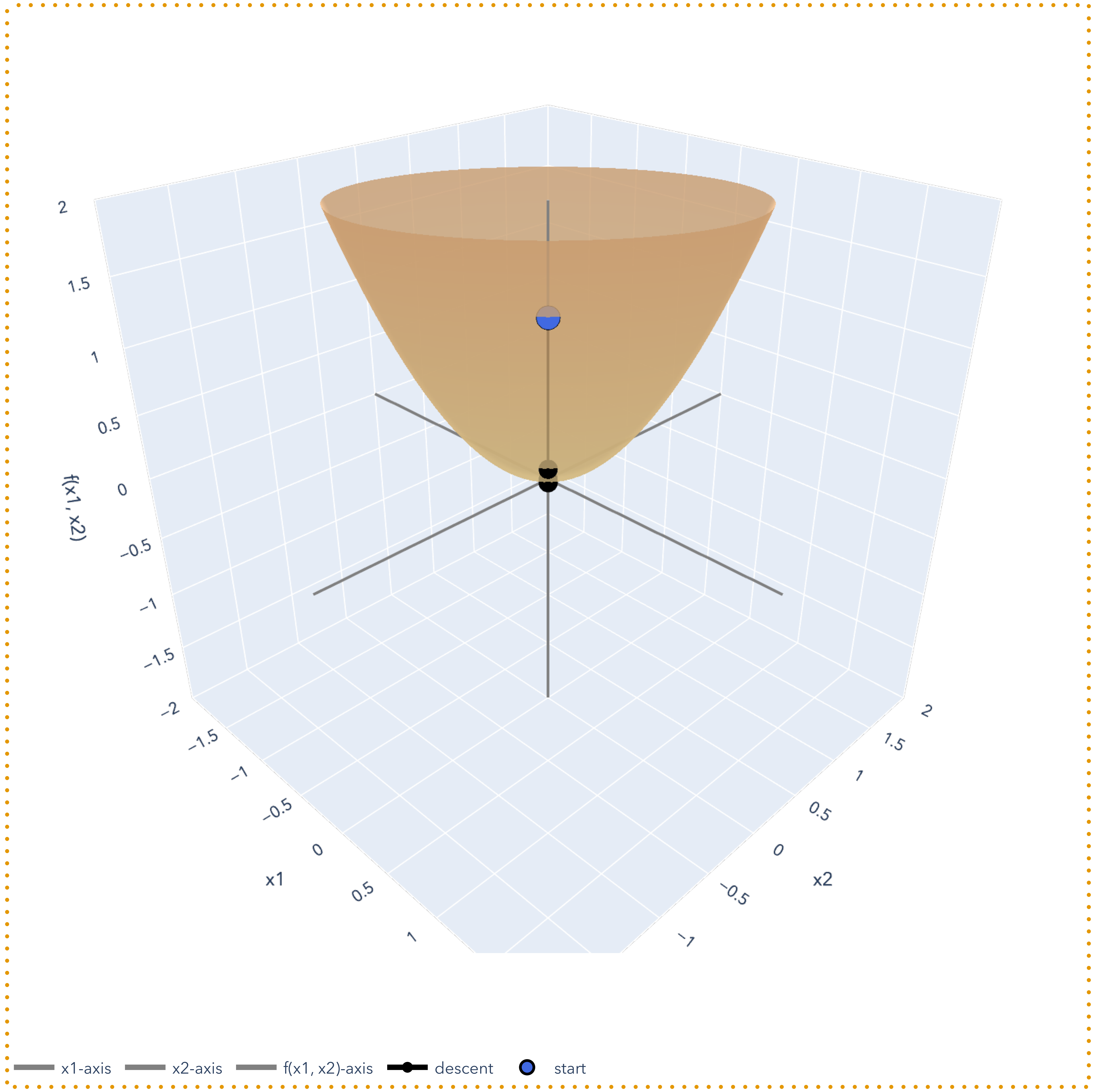
Example: positive definite

$$f(x, y) = \begin{bmatrix} x & y \end{bmatrix} \begin{bmatrix} 2 & -1 \\ -1 & 2 \end{bmatrix} \begin{bmatrix} x \\ y \end{bmatrix}$$

Eigendecomposition:

$$\begin{bmatrix} 2 & -1 \\ -1 & 2 \end{bmatrix} = \begin{bmatrix} -1/\sqrt{2} & 1/\sqrt{2} \\ 1/\sqrt{2} & 1/\sqrt{2} \end{bmatrix} \begin{bmatrix} 3 & 0 \\ 0 & 1 \end{bmatrix} \begin{bmatrix} -1/\sqrt{2} & 1/\sqrt{2} \\ 1/\sqrt{2} & 1/\sqrt{2} \end{bmatrix}$$

so $\mathbf{\Lambda} = \begin{bmatrix} 3 & 0 \\ 0 & 1 \end{bmatrix}$.



Quadratic Forms

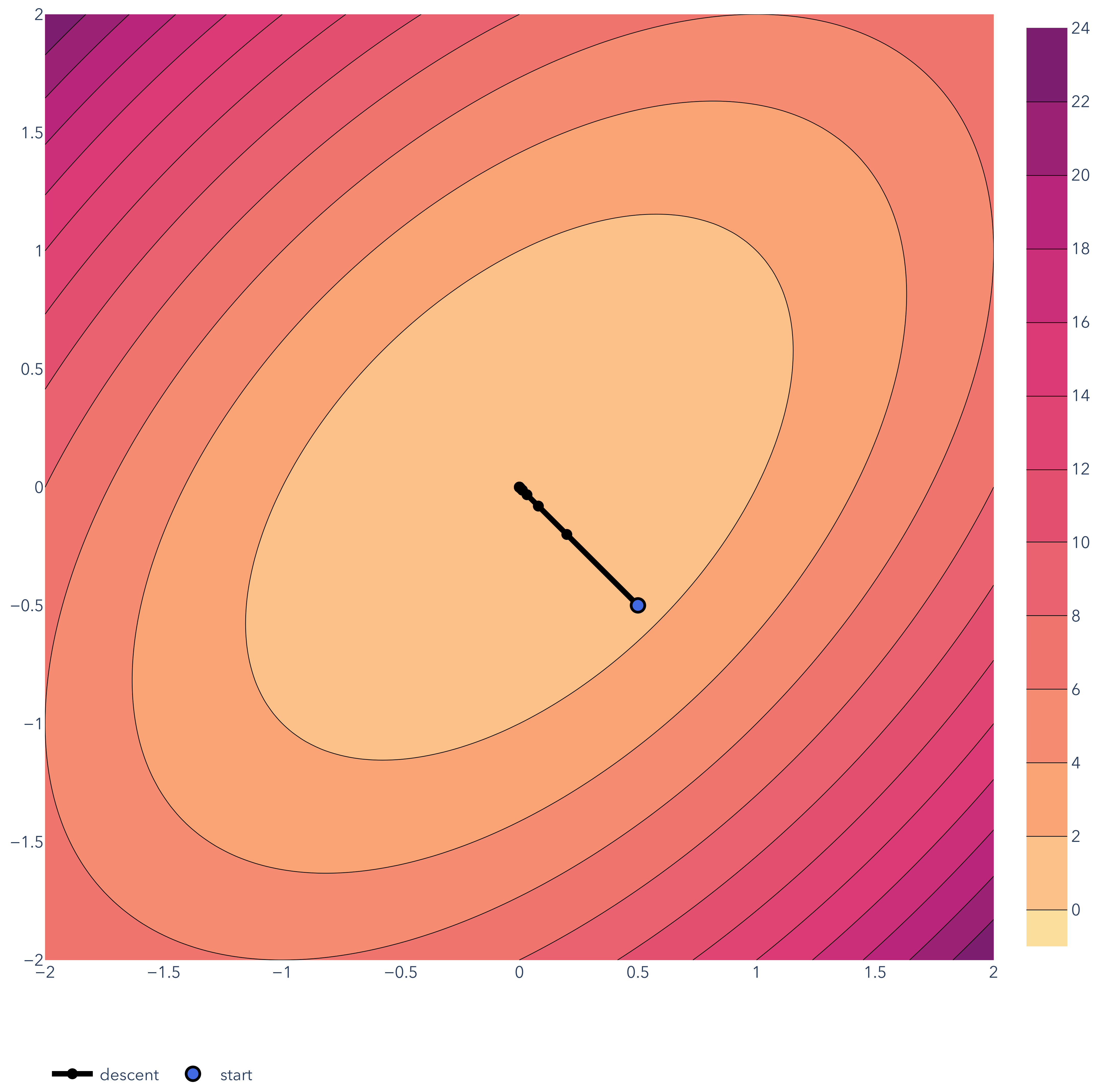
Example: positive definite

$$f(x, y) = \begin{bmatrix} x & y \end{bmatrix} \begin{bmatrix} 2 & -1 \\ -1 & 2 \end{bmatrix} \begin{bmatrix} x \\ y \end{bmatrix}$$

Eigendecomposition:

$$\begin{bmatrix} 2 & -1 \\ -1 & 2 \end{bmatrix} = \begin{bmatrix} -1/\sqrt{2} & 1/\sqrt{2} \\ 1/\sqrt{2} & 1/\sqrt{2} \end{bmatrix} \begin{bmatrix} 3 & 0 \\ 0 & 1 \end{bmatrix} \begin{bmatrix} -1/\sqrt{2} & 1/\sqrt{2} \\ 1/\sqrt{2} & 1/\sqrt{2} \end{bmatrix}$$

so $\mathbf{\Lambda} = \begin{bmatrix} 3 & 0 \\ 0 & 1 \end{bmatrix}$.



Quadratic Forms

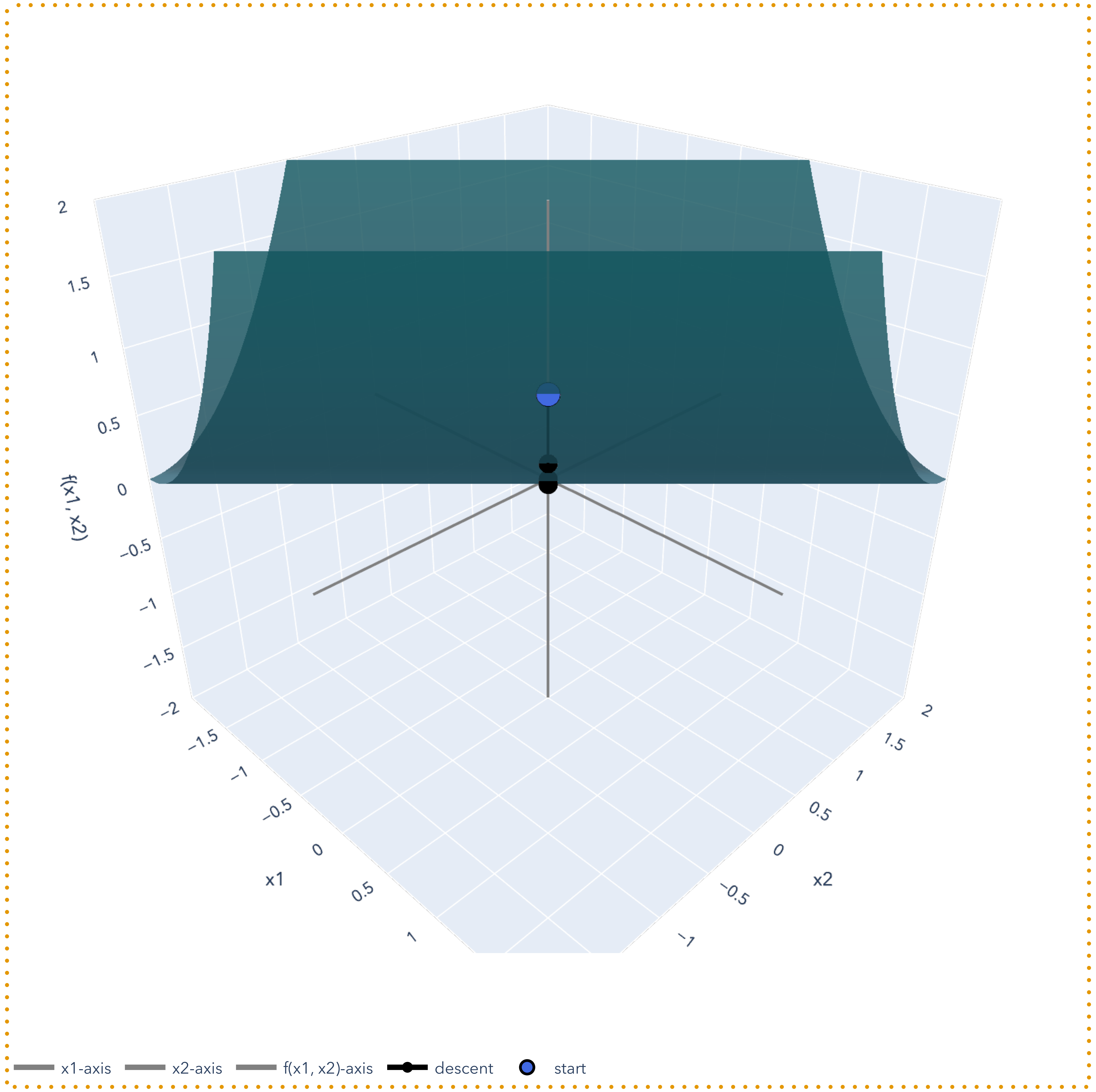
Example: positive semidefinite

$$f(x, y) = \begin{bmatrix} x & y \end{bmatrix} \begin{bmatrix} 1 & -1 \\ -1 & 1 \end{bmatrix} \begin{bmatrix} x \\ y \end{bmatrix}$$

Eigendecomposition:

$$\begin{bmatrix} 1 & -1 \\ -1 & 1 \end{bmatrix} = \begin{bmatrix} -1/\sqrt{2} & 1/\sqrt{2} \\ 1/\sqrt{2} & 1/\sqrt{2} \end{bmatrix} \begin{bmatrix} 2 & 0 \\ 0 & 0 \end{bmatrix} \begin{bmatrix} -1/\sqrt{2} & 1/\sqrt{2} \\ 1/\sqrt{2} & 1/\sqrt{2} \end{bmatrix}$$

so $\mathbf{\Lambda} = \begin{bmatrix} 2 & 0 \\ 0 & 0 \end{bmatrix}$.



Quadratic Forms

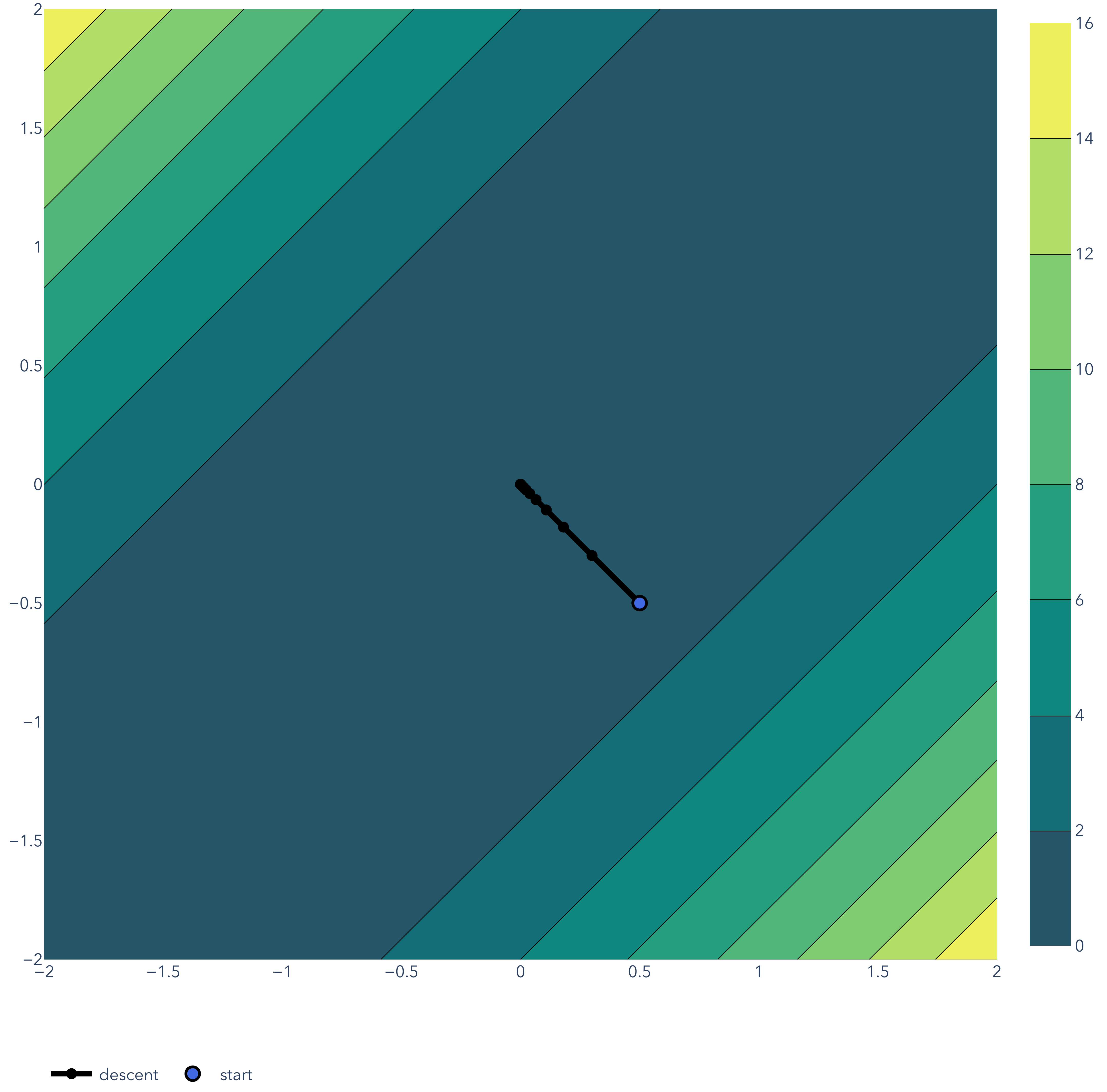
Example: positive semidefinite

$$f(x,y) = \begin{bmatrix} x & y \end{bmatrix} \begin{bmatrix} 1 & -1 \\ -1 & 1 \end{bmatrix} \begin{bmatrix} x \\ y \end{bmatrix}$$

Eigendecomposition:

$$\begin{bmatrix} 1 & -1 \\ -1 & 1 \end{bmatrix} = \begin{bmatrix} -1/\sqrt{2} & 1/\sqrt{2} \\ 1/\sqrt{2} & 1/\sqrt{2} \end{bmatrix} \begin{bmatrix} 2 & 0 \\ 0 & 0 \end{bmatrix} \begin{bmatrix} -1/\sqrt{2} & 1/\sqrt{2} \\ 1/\sqrt{2} & 1/\sqrt{2} \end{bmatrix}$$

$$\text{so } \mathbf{\Lambda} = \begin{bmatrix} 2 & 0 \\ 0 & 0 \end{bmatrix}.$$



Quadratic Forms

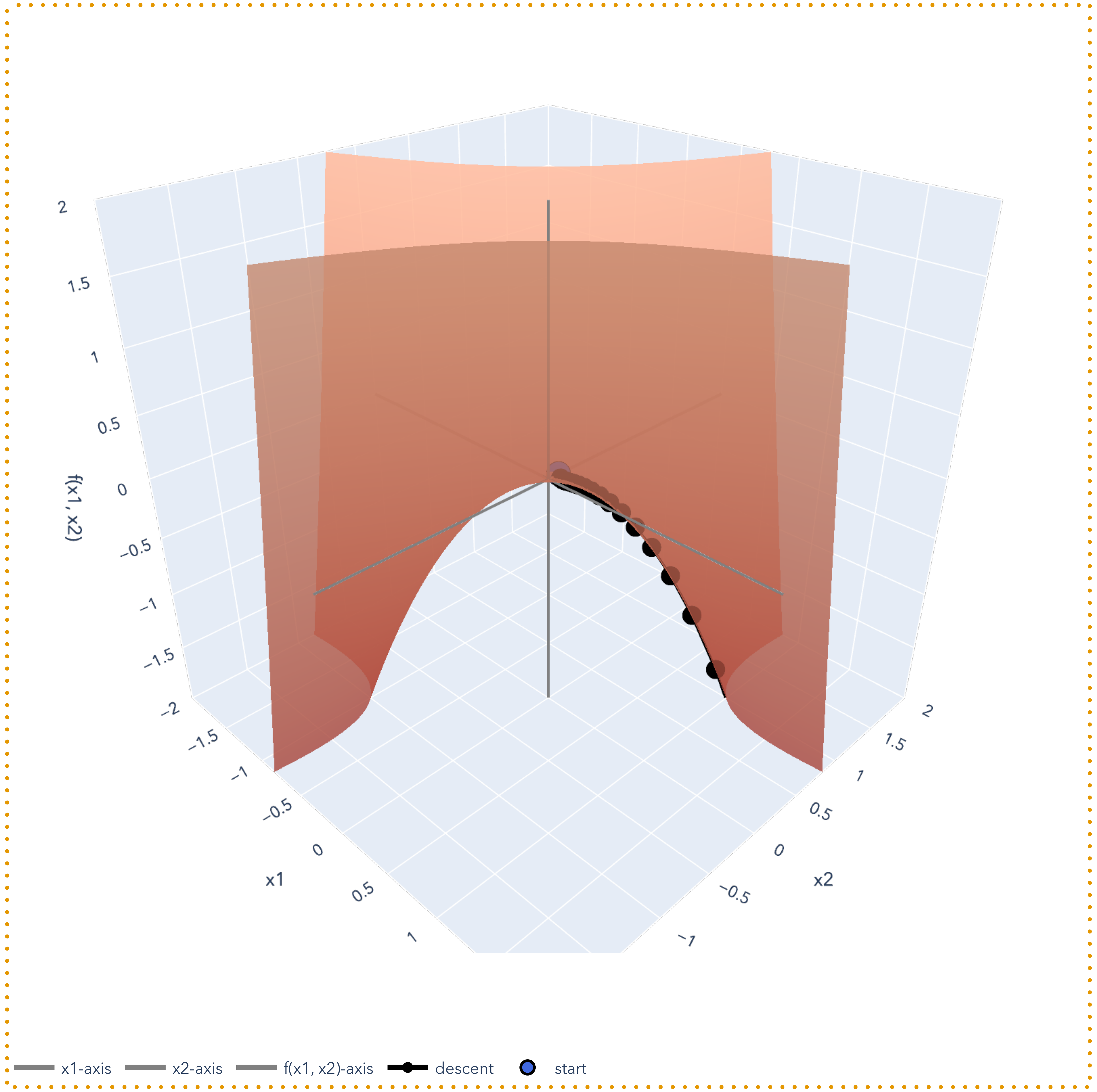
Example: indefinite

$$f(x,y) = \begin{bmatrix} x & y \end{bmatrix} \begin{bmatrix} 1 & -2 \\ -2 & 1 \end{bmatrix} \begin{bmatrix} x \\ y \end{bmatrix}$$

Eigendecomposition:

$$\begin{bmatrix} 1 & -2 \\ -2 & 1 \end{bmatrix} = \begin{bmatrix} -1/\sqrt{2} & 1/\sqrt{2} \\ 1/\sqrt{2} & 1/\sqrt{2} \end{bmatrix} \begin{bmatrix} 3 & 0 \\ 0 & -1 \end{bmatrix} \begin{bmatrix} -1/\sqrt{2} & 1/\sqrt{2} \\ 1/\sqrt{2} & 1/\sqrt{2} \end{bmatrix}$$

so $\mathbf{\Lambda} = \begin{bmatrix} 3 & 0 \\ 0 & -1 \end{bmatrix}$.



Quadratic Forms

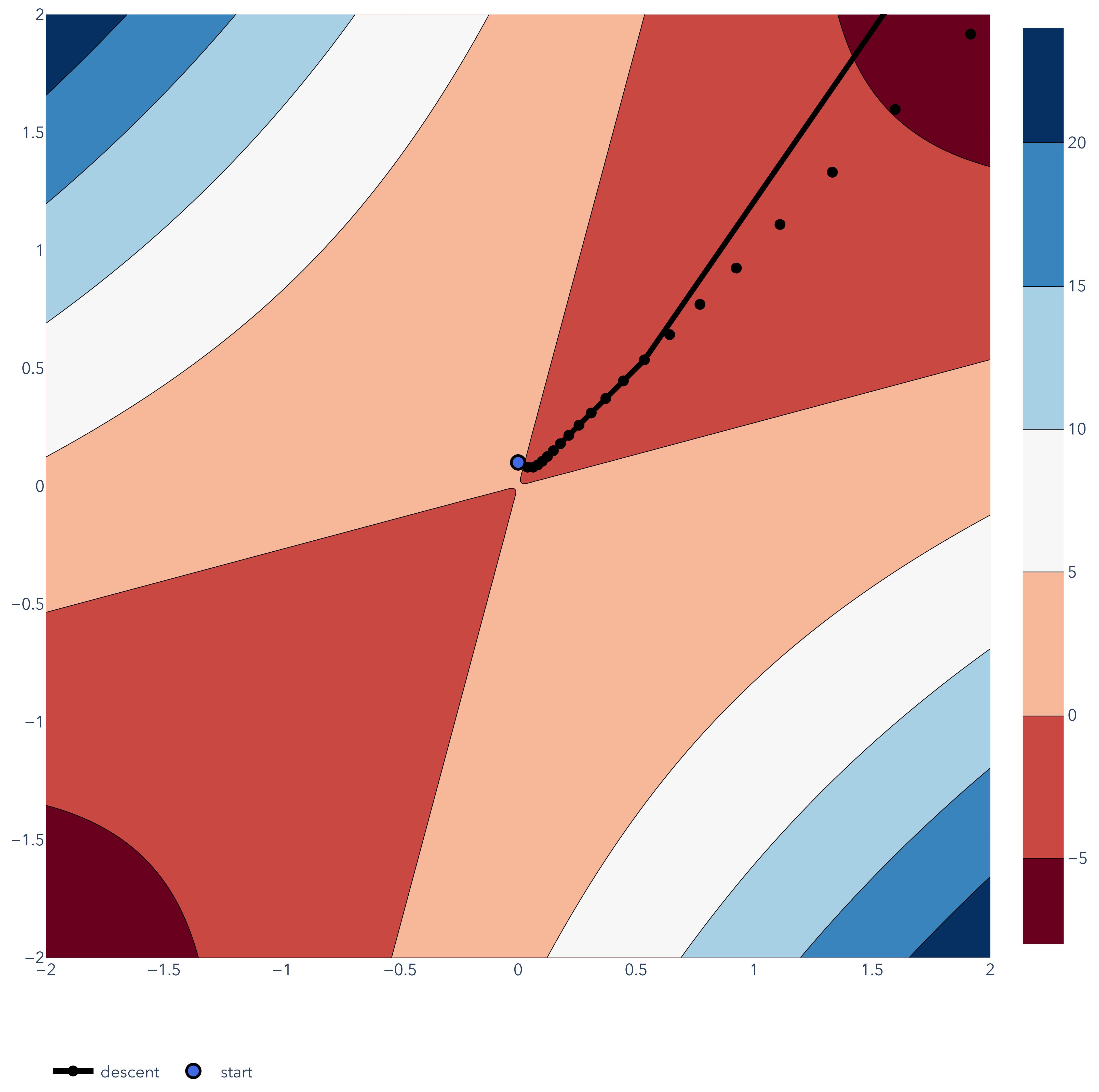
Example: indefinite

$$f(x, y) = \begin{bmatrix} x & y \end{bmatrix} \begin{bmatrix} 1 & -2 \\ -2 & 1 \end{bmatrix} \begin{bmatrix} x \\ y \end{bmatrix}$$

Eigendecomposition:

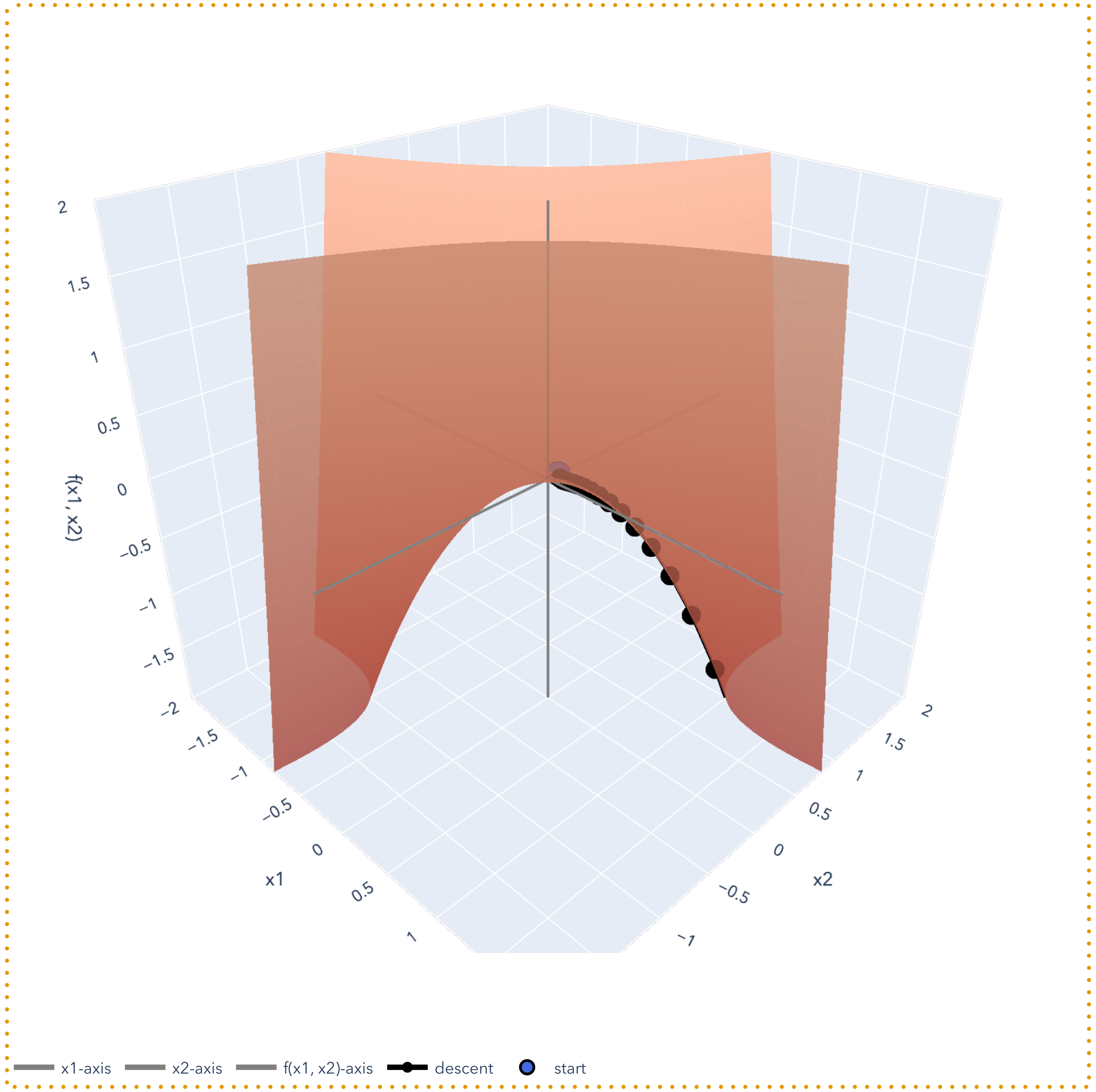
$$\begin{bmatrix} 1 & -2 \\ -2 & 1 \end{bmatrix} = \begin{bmatrix} -1/\sqrt{2} & 1/\sqrt{2} \\ 1/\sqrt{2} & 1/\sqrt{2} \end{bmatrix} \begin{bmatrix} 3 & 0 \\ 0 & -1 \end{bmatrix} \begin{bmatrix} -1/\sqrt{2} & 1/\sqrt{2} \\ 1/\sqrt{2} & 1/\sqrt{2} \end{bmatrix}$$

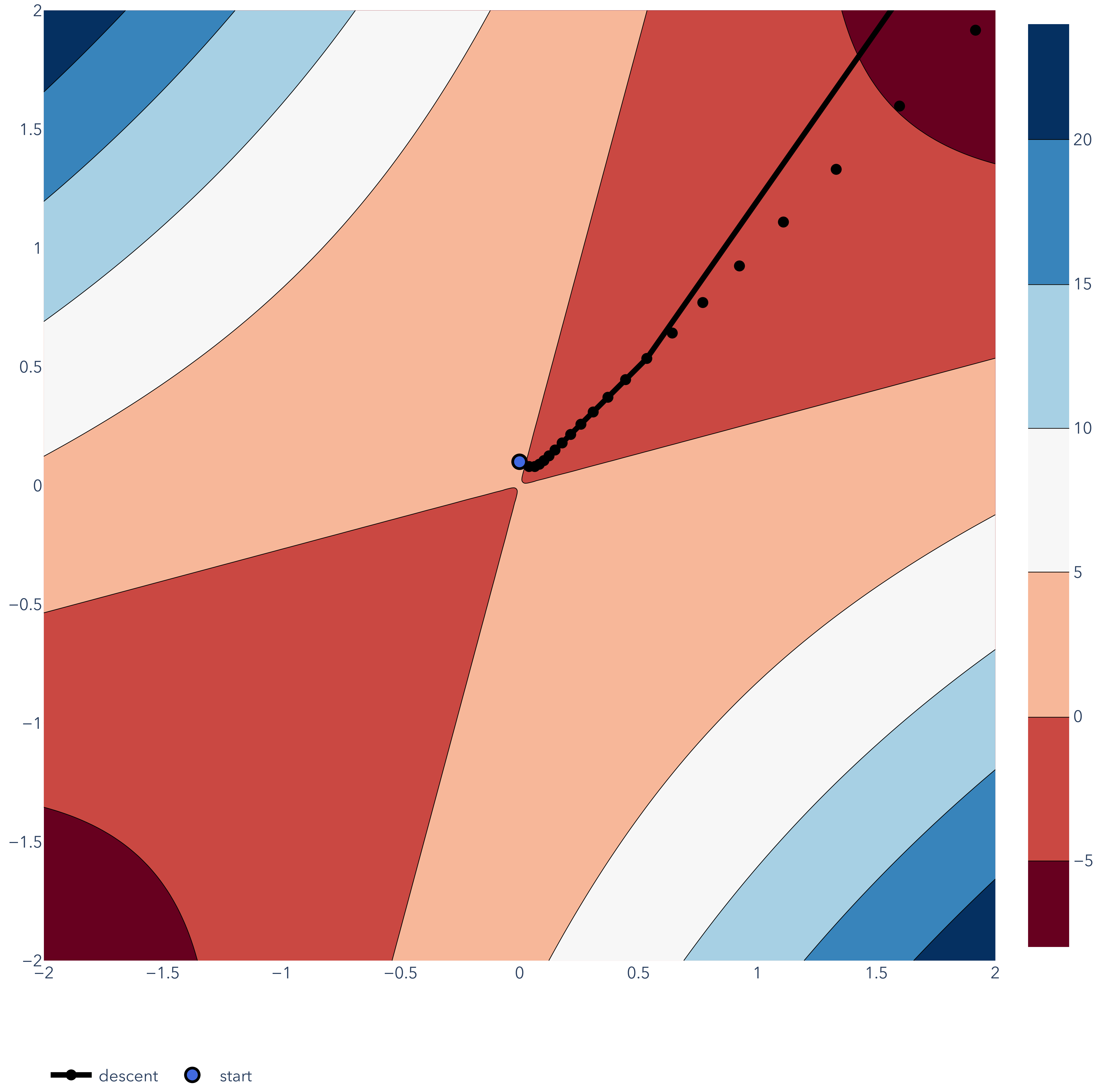
$$\text{so } \mathbf{\Lambda} = \begin{bmatrix} 3 & 0 \\ 0 & -1 \end{bmatrix}.$$



Quadratic Forms

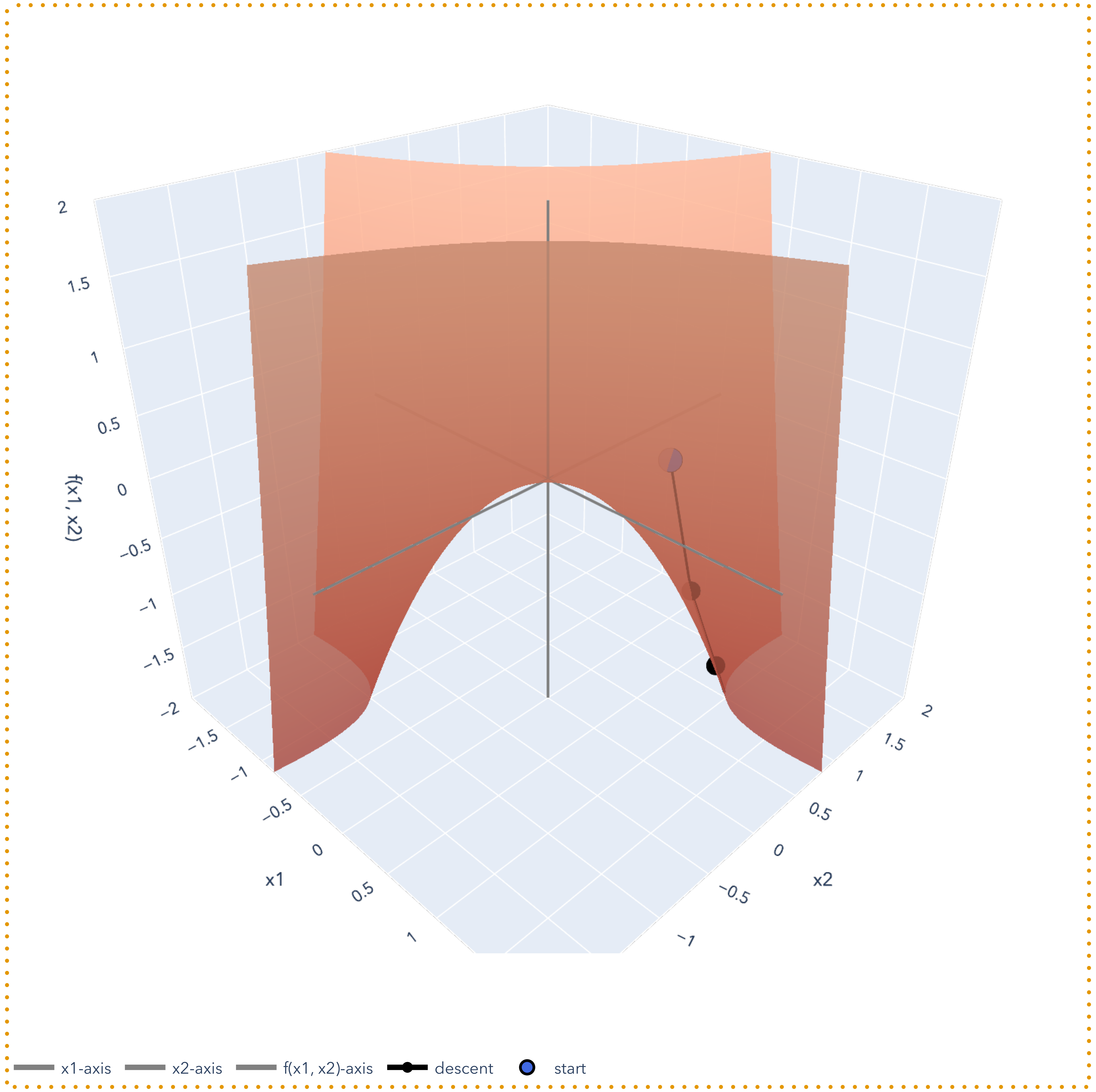
Example: indefinite

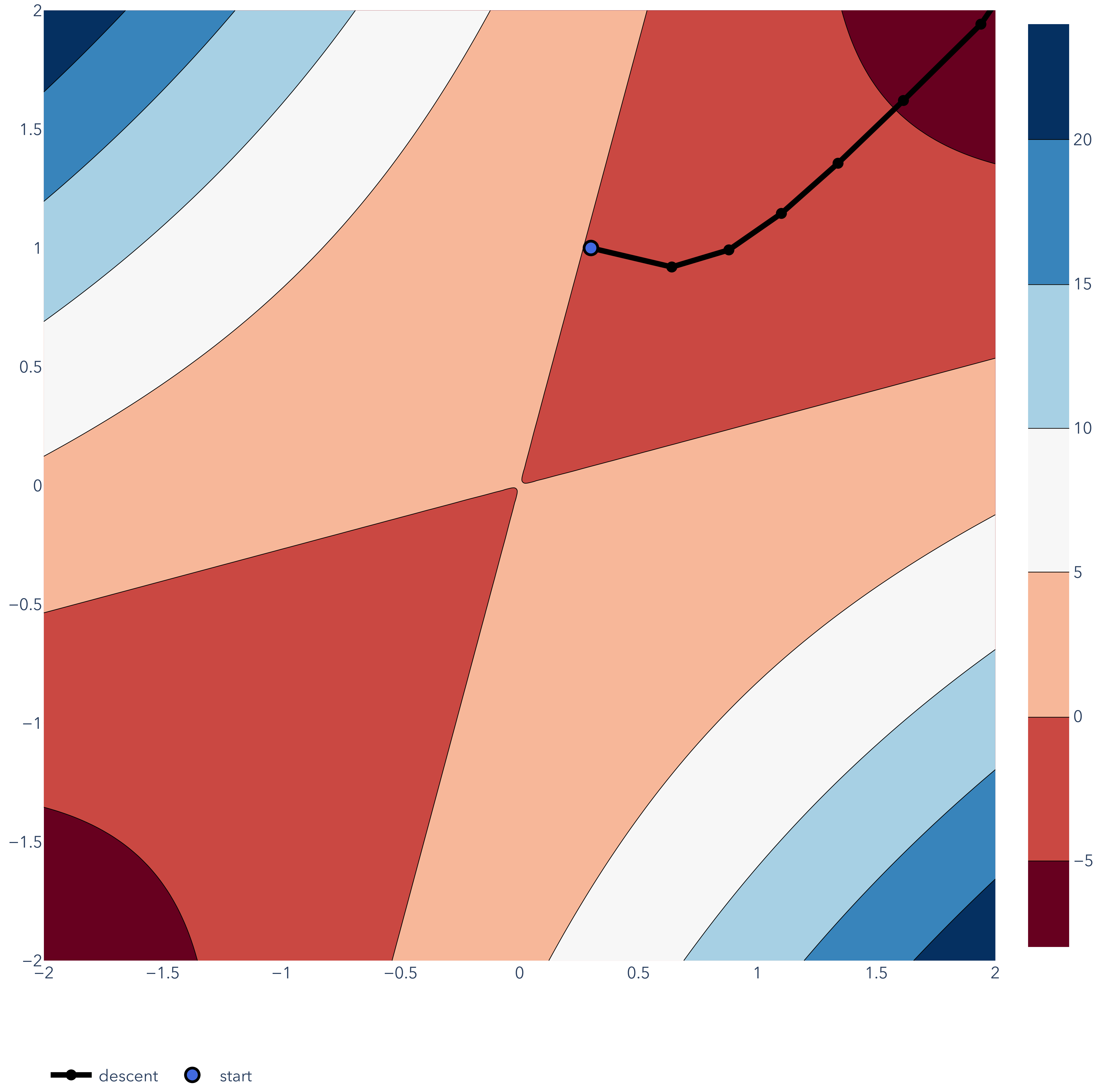




Quadratic Forms

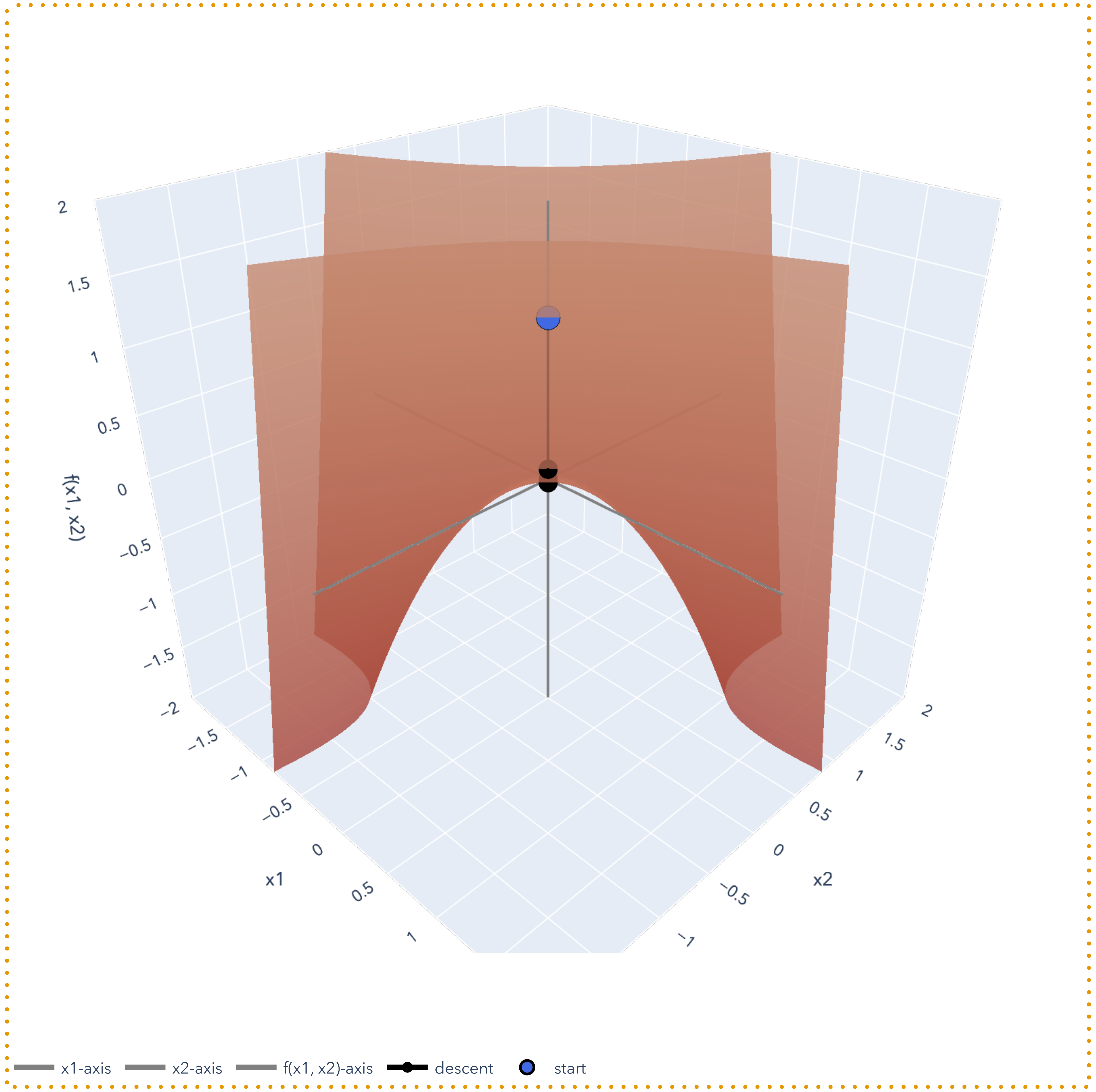
Example: indefinite

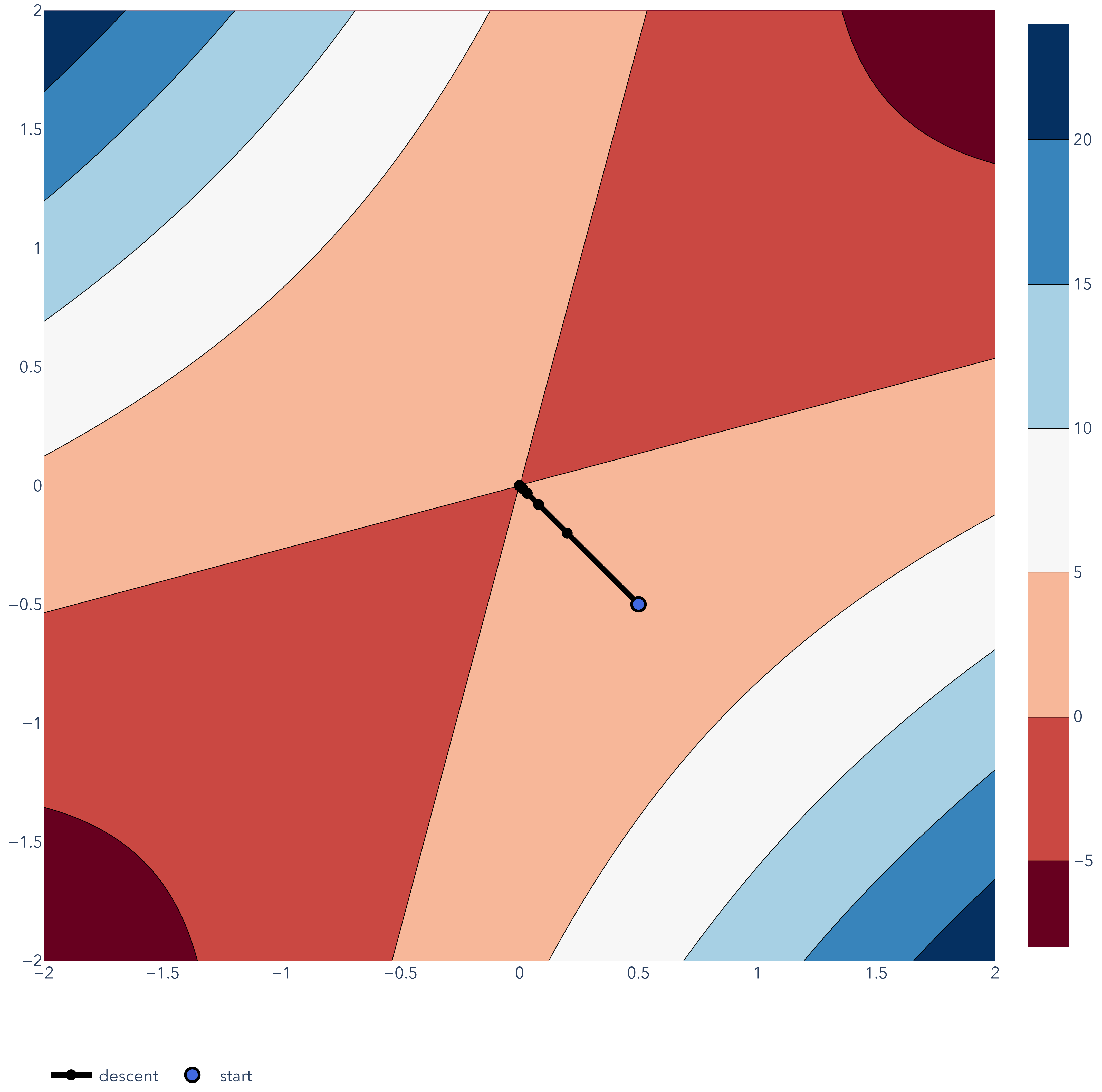




Quadratic Forms

Example: indefinite





Least Squares

Example of quadratic form

Consider the sum of squared residuals error function for least squares…

$$f(\mathbf{w}) = \|\mathbf{X}\mathbf{w} - \mathbf{y}\|^2$$

$$(\mathbf{X}\mathbf{w} - \mathbf{y})^\top(\mathbf{X}\mathbf{w} - \mathbf{y}) = \mathbf{w}^\top(\mathbf{X}^\top\mathbf{X})\mathbf{w} - 2\mathbf{w}^\top(\mathbf{X}^\top\mathbf{y}) + \mathbf{y}^\top\mathbf{y}.$$

The quadratic form $\mathbf{w}^\top(\mathbf{X}^\top\mathbf{X})\mathbf{w}$ is positive semidefinite!

Least Squares

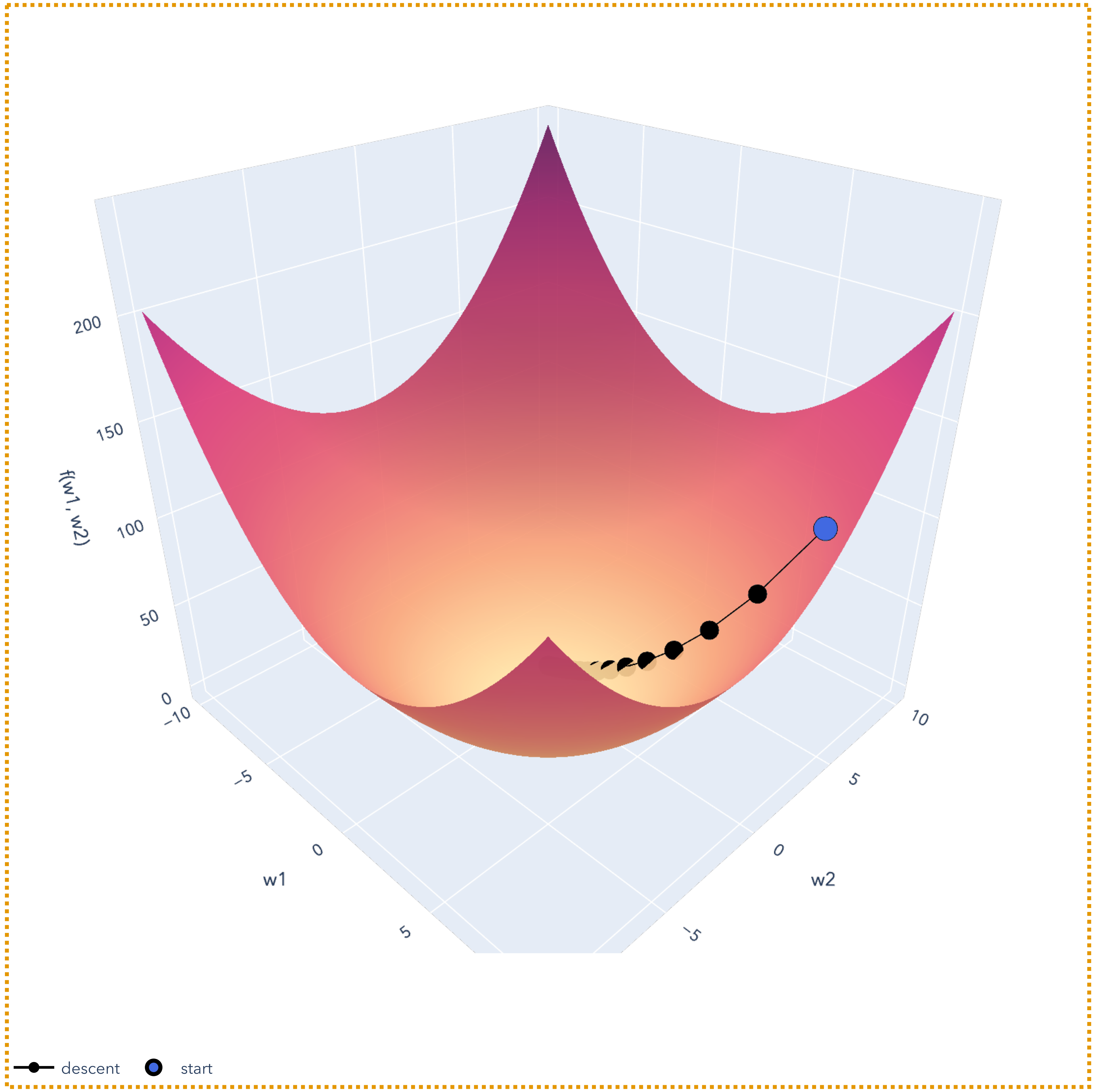
Example of quadratic form

Consider the sum of squared residuals error function for least squares…

$$f(\mathbf{w}) = \|\mathbf{X}\mathbf{w} - \mathbf{y}\|^2$$

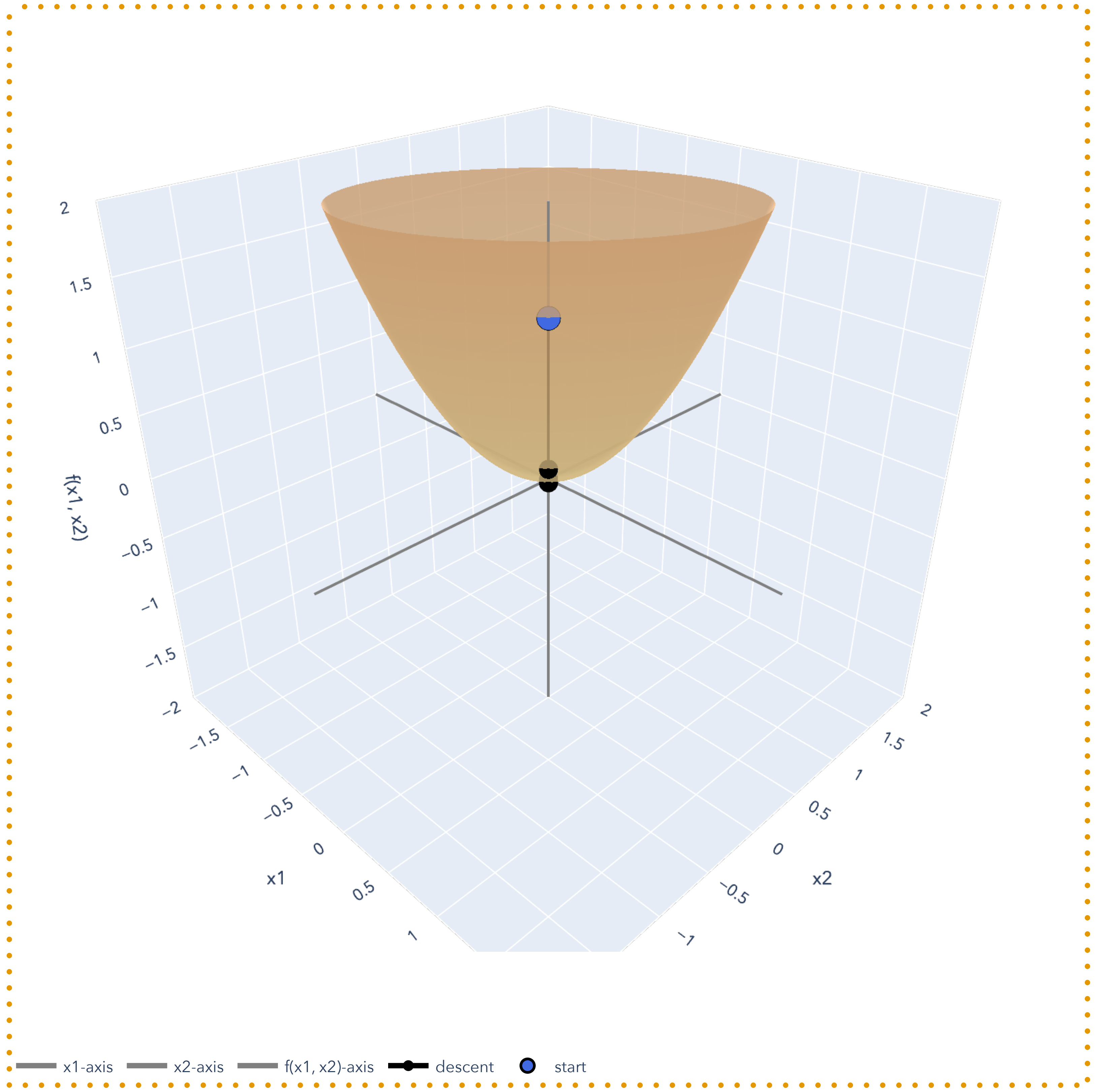
$$(\mathbf{X}\mathbf{w} - \mathbf{y})^\top(\mathbf{X}\mathbf{w} - \mathbf{y}) = \mathbf{w}^\top(\mathbf{X}^\top\mathbf{X})\mathbf{w} - 2\mathbf{w}^\top(\mathbf{X}^\top\mathbf{y}) + \mathbf{y}^\top\mathbf{y}$$

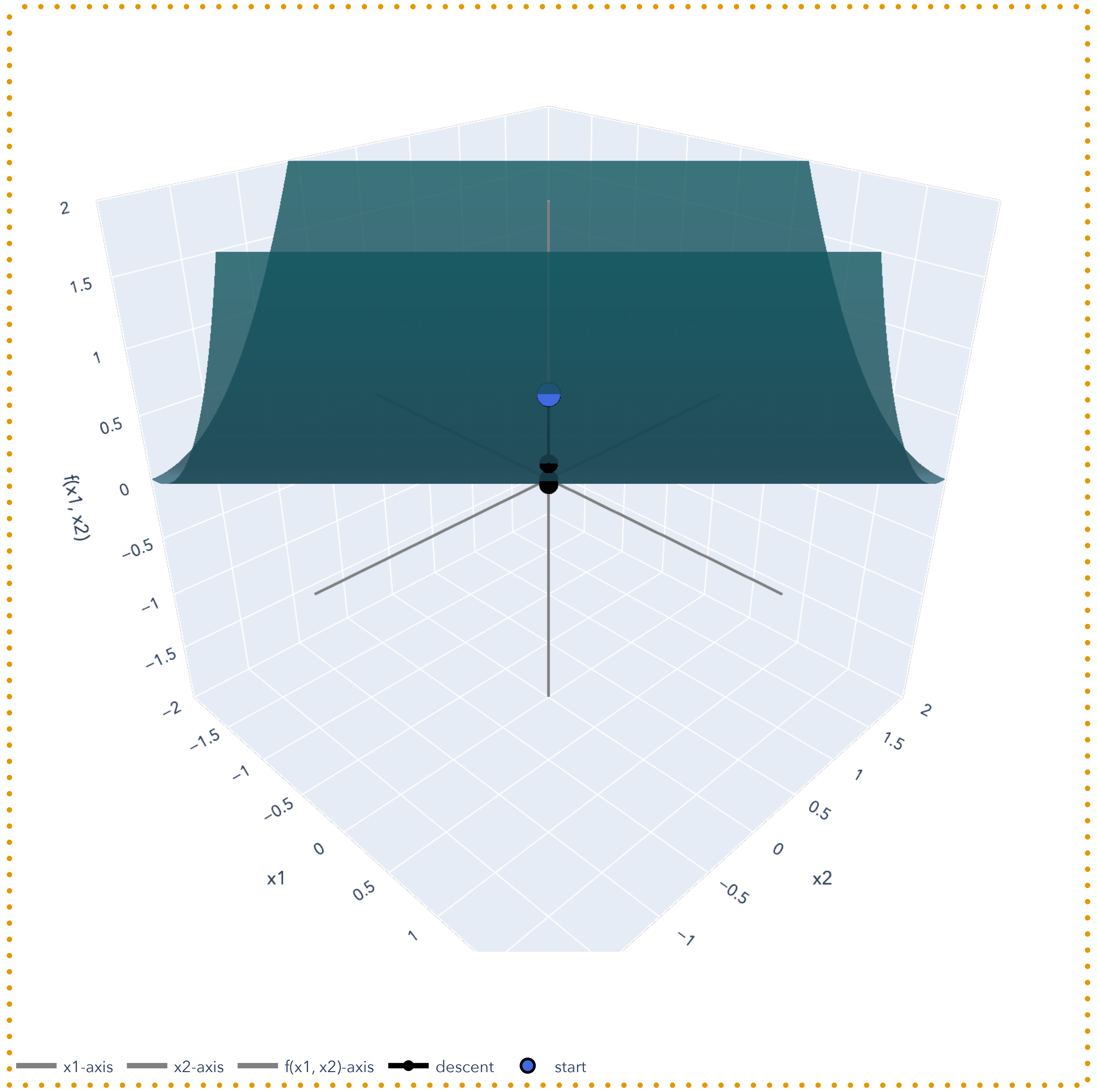
The quadratic form $\mathbf{w}^\top(\mathbf{X}^\top\mathbf{X})\mathbf{w}$ is positive semidefinite!

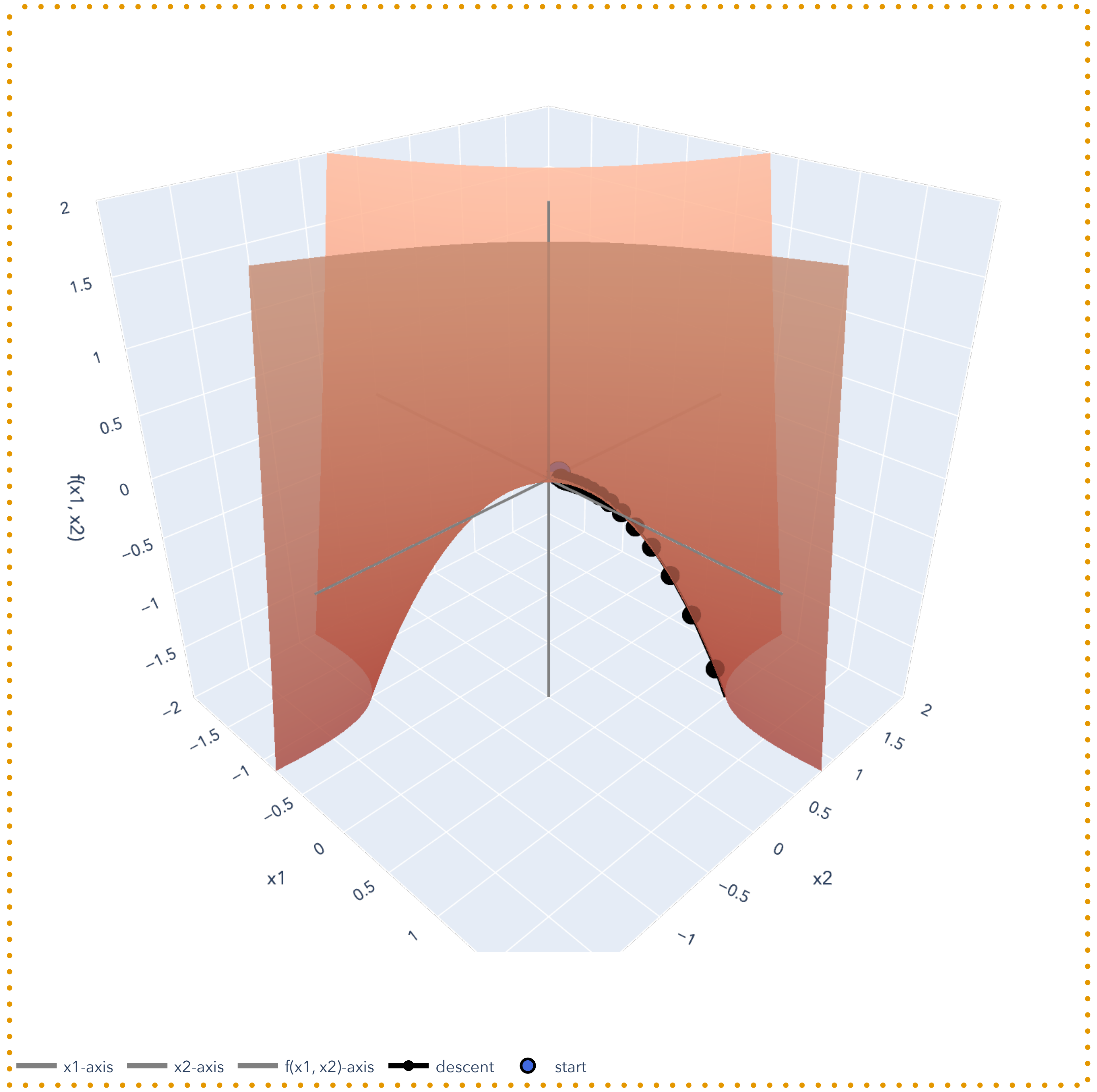


Gradient Descent

Preview







$$\mathbf{\Lambda} = \begin{bmatrix} 3 & 0 \\ 0 & 1 \end{bmatrix}$$

$$\mathbf{\Lambda} = \begin{bmatrix} 2 & 0 \\ 0 & 0 \end{bmatrix}$$

$$\mathbf{\Lambda} = \begin{bmatrix} 3 & 0 \\ 0 & -1 \end{bmatrix}$$

Recap

Lesson Overview

Linear dynamical systems example. Motivation for eigendecomposition as a way to make repeated matrix multiplication easier.

Eigendecomposition. Definition of eigenvectors, eigenvalues.

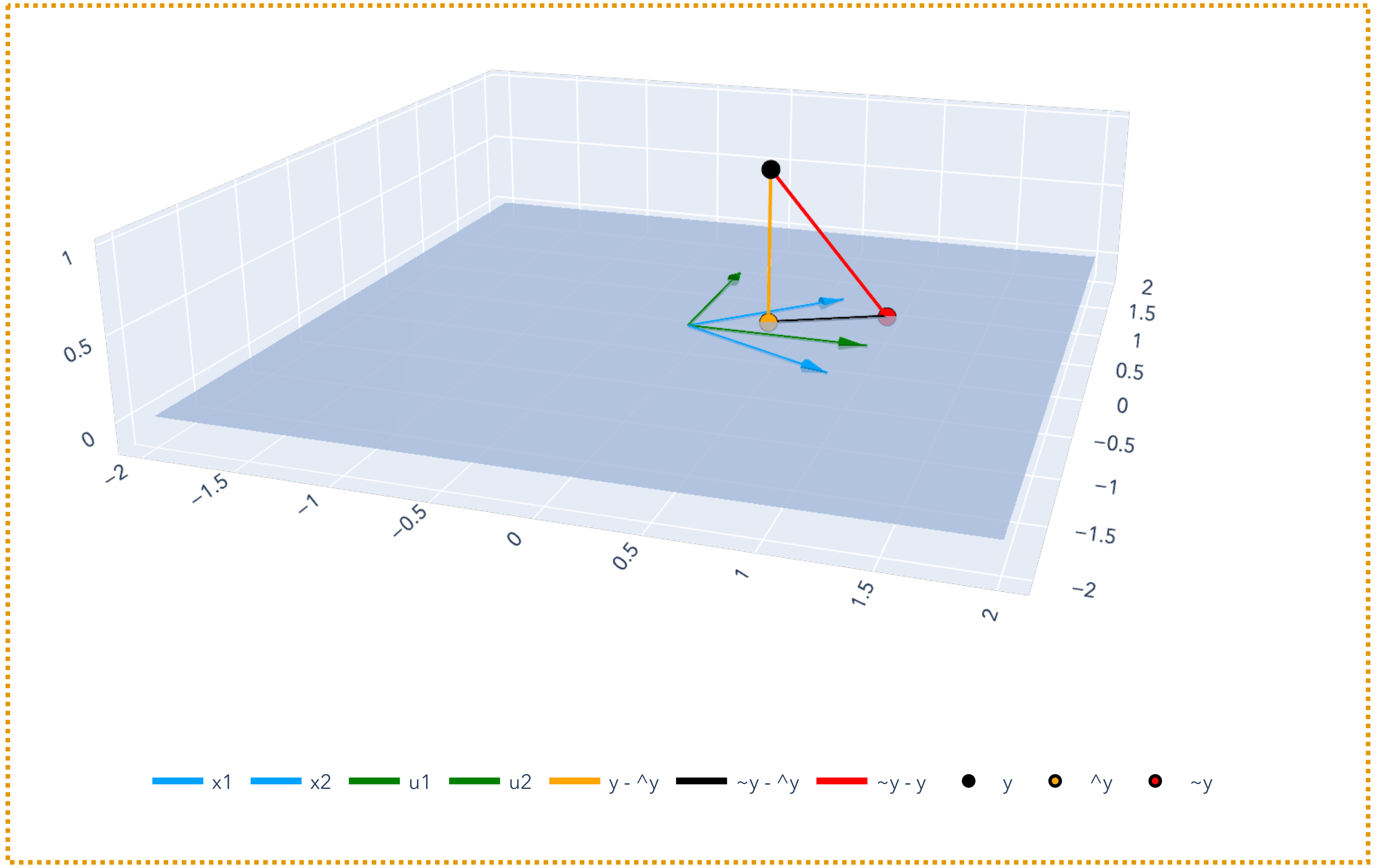
Eigendecomposition and SVD. The eigendecomposition drops out of the SVD.

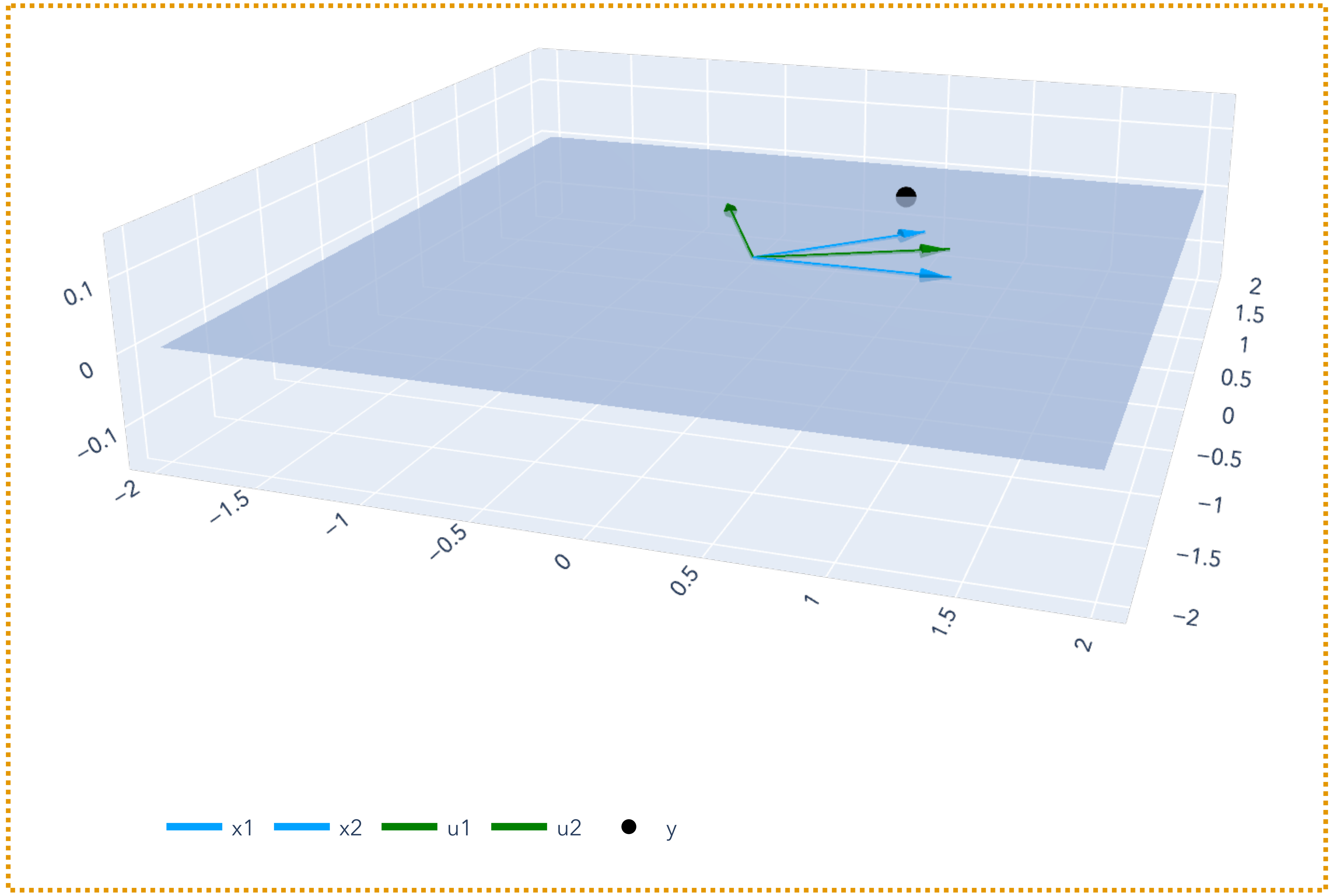
Spectral Theorem. Symmetric matrices are always diagonalizable.

Positive semidefinite matrices/positive definite matrices. Definition and some visual examples through the corresponding quadratic forms.

Lesson Overview

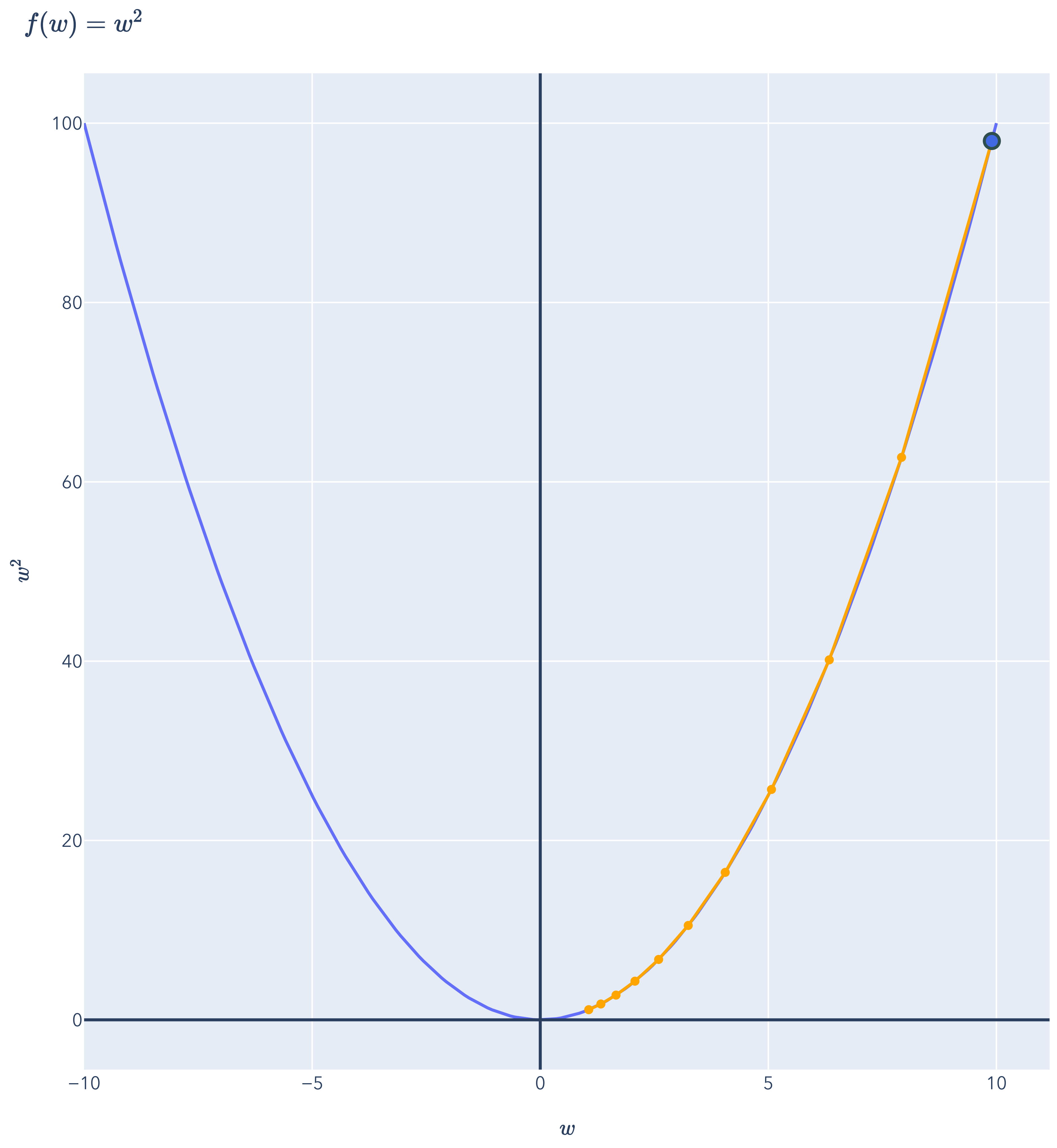
Big Picture: Least Squares

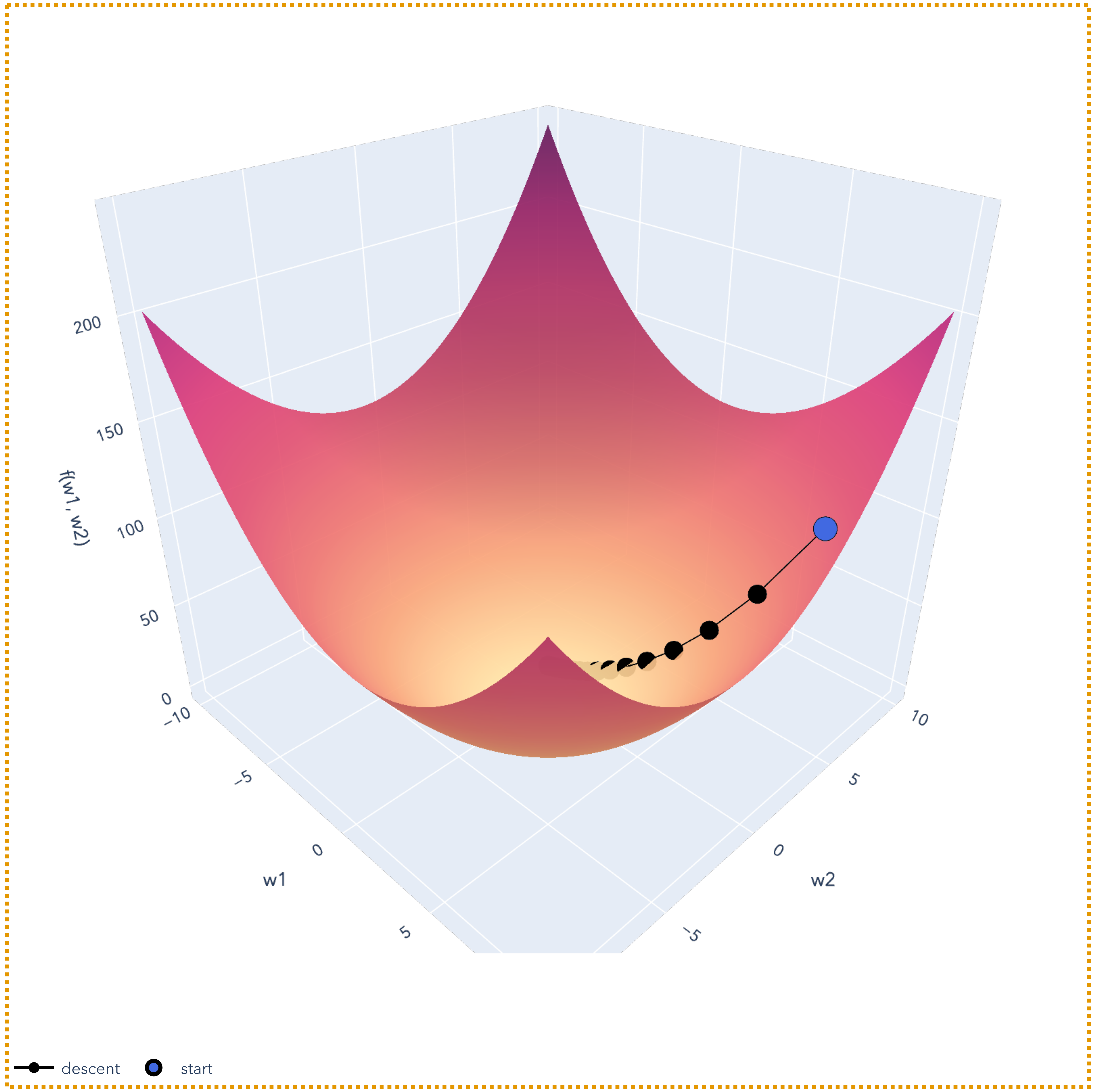




Lesson Overview

Big Picture: Gradient Descent





Lesson Overview

Big Picture: Gradient Descent