Math for Machine Learning

Week 2.2: Eigendecomposition and PSD Matrices

By: Samuel Deng

Logistics & Announcements

- HW 2: Due: July 18, next Thurs.
- HW 1: Due: July 11, tomorrow.
- OFFICE HOURS 3PM – 5PM (ZOOM).
- BREAKS: 8 minutes. = 16 min. ~~total~~.

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.

PCA

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

$n > d$

$Xw \approx y$

$\downarrow$

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

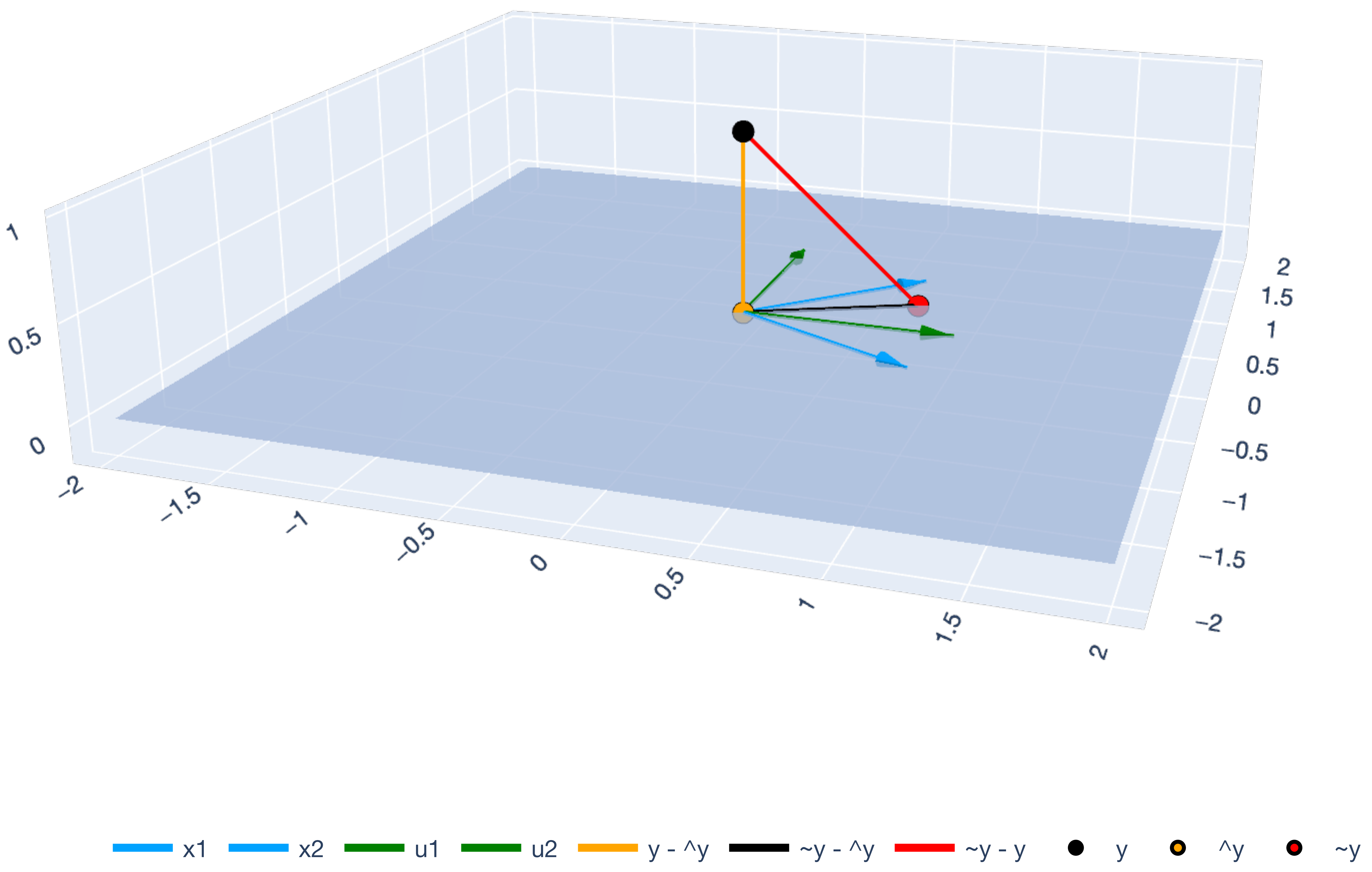
$\boxed{\hat{w} = X^{+} y}$.

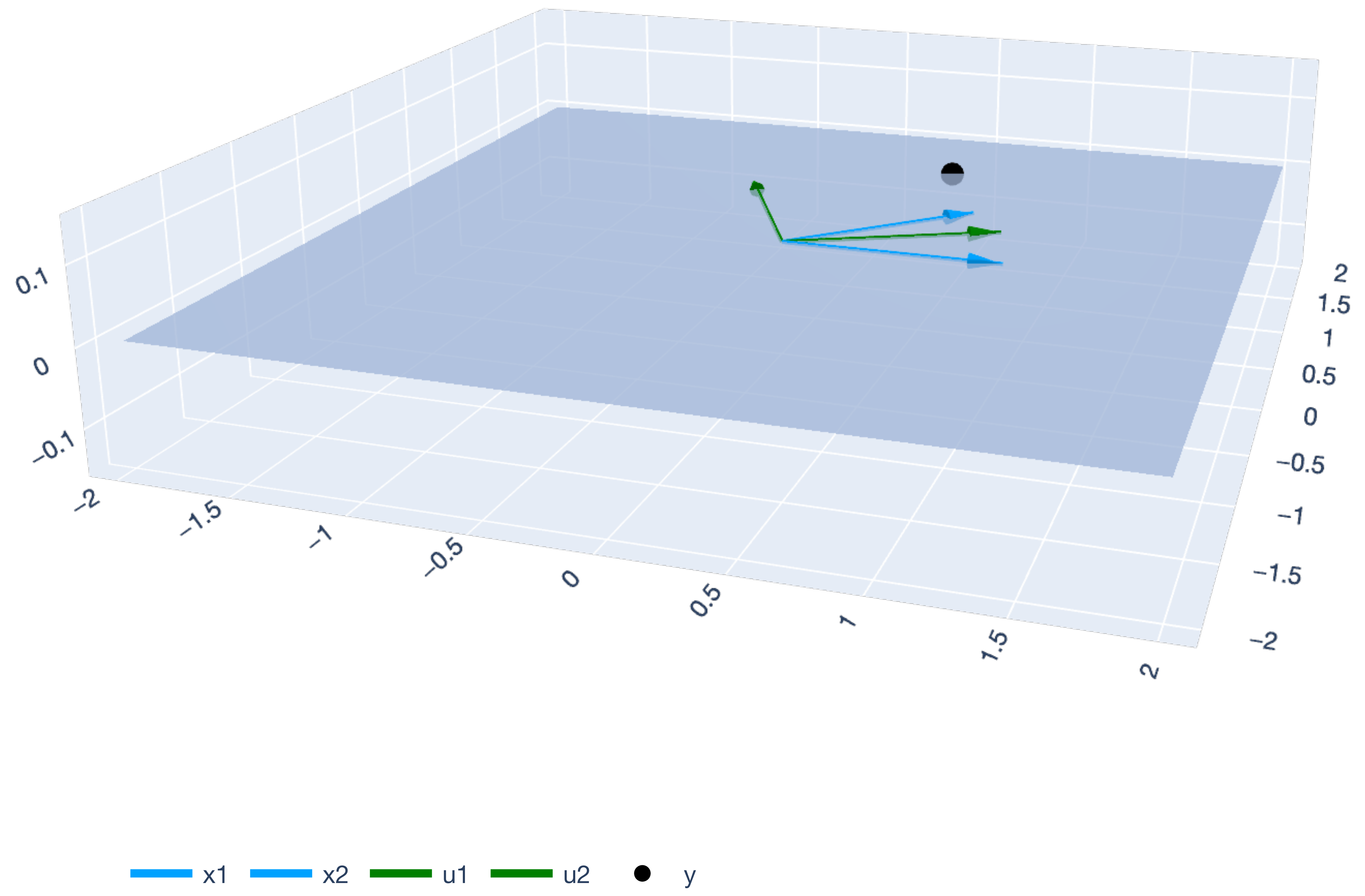
$d > n$

$\boxed{Xw = y}$

$\downarrow$

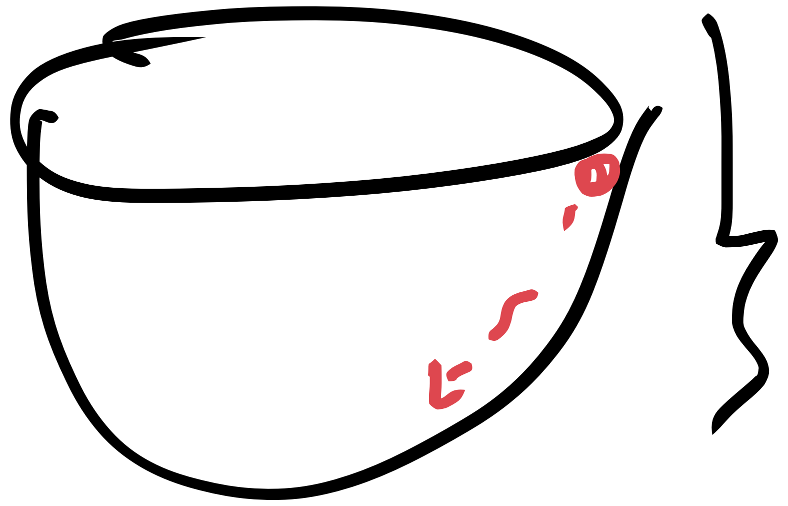
more unknowns than equations.



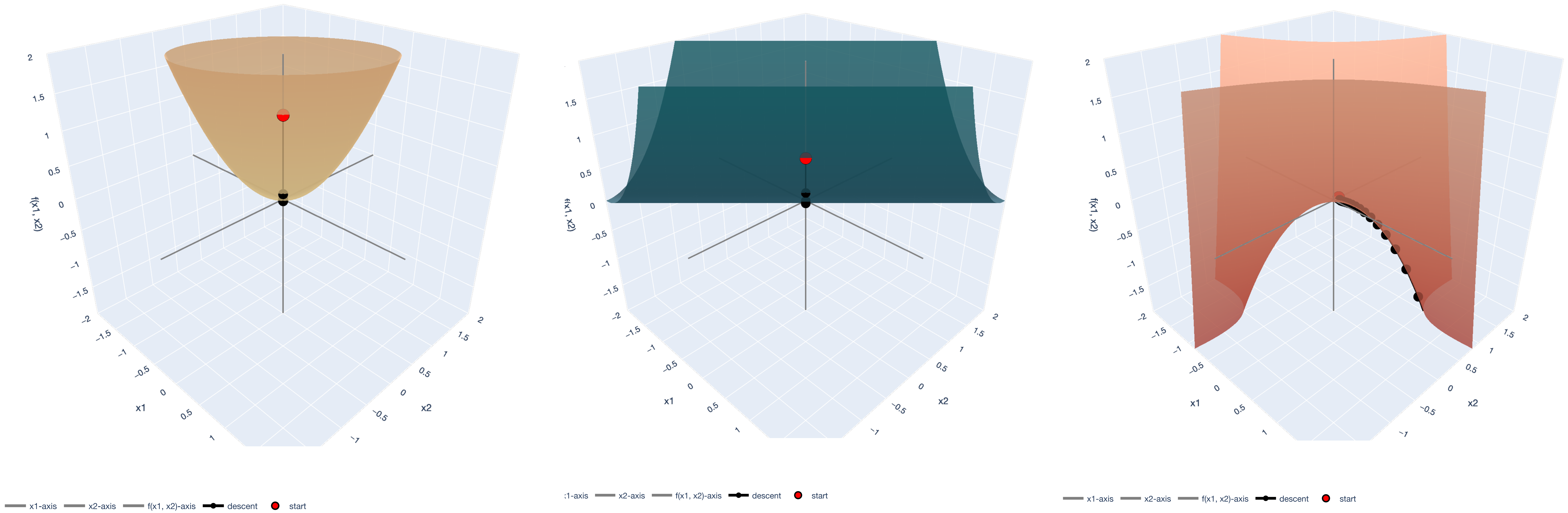


Lesson Overview

Big Picture: Gradient Descent



$\|Xw - y\|_2^2 = f(w)$



$\boxed{x^T A x}$

Least Squares

A Quick Review

Regression

Setup

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

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

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

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

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

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

Regression

Setup

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

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

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

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

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

SVD and Pseudoinverse

Review

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.

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

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

left Inverse.

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

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

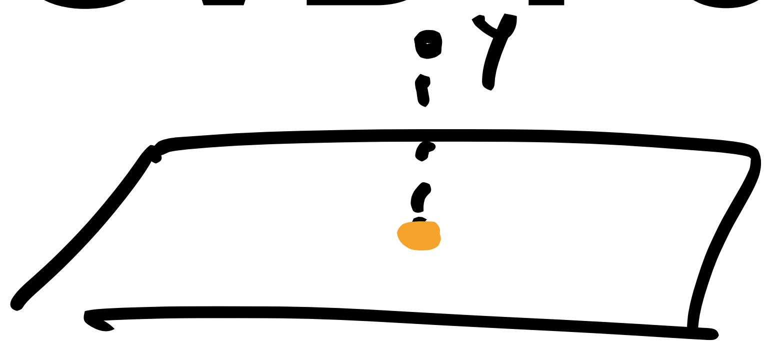
Right - inverse.

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

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

Least Squares: SVD Perspective

Unified Picture



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

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

We can solve exactly.

Choose

X^+

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

which is an exact solution.

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

We approximate by least squares:

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

Choose

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

the best approximate solution:

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

$\neq 0$

more unknowns > equations

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

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

Choose

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

the minimum norm solution:

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

Least Squares: SVD Perspective

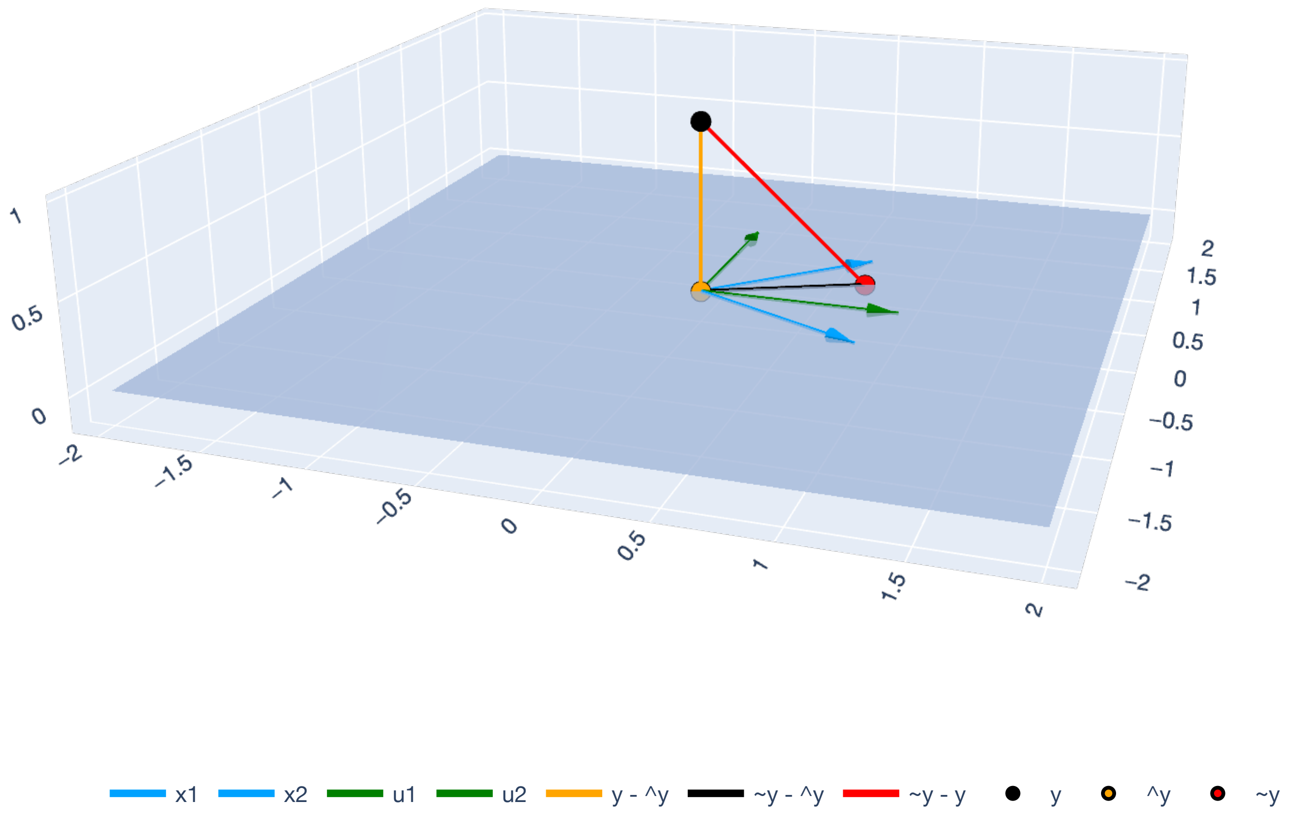
Unified Picture

We want to solve $\mathbf{Xw} = \mathbf{y}$. $\boxed{\text{Use } \hat{\mathbf{w}} = \mathbf{X}^{+}\mathbf{y}!}$

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

We approximate by least squares:

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

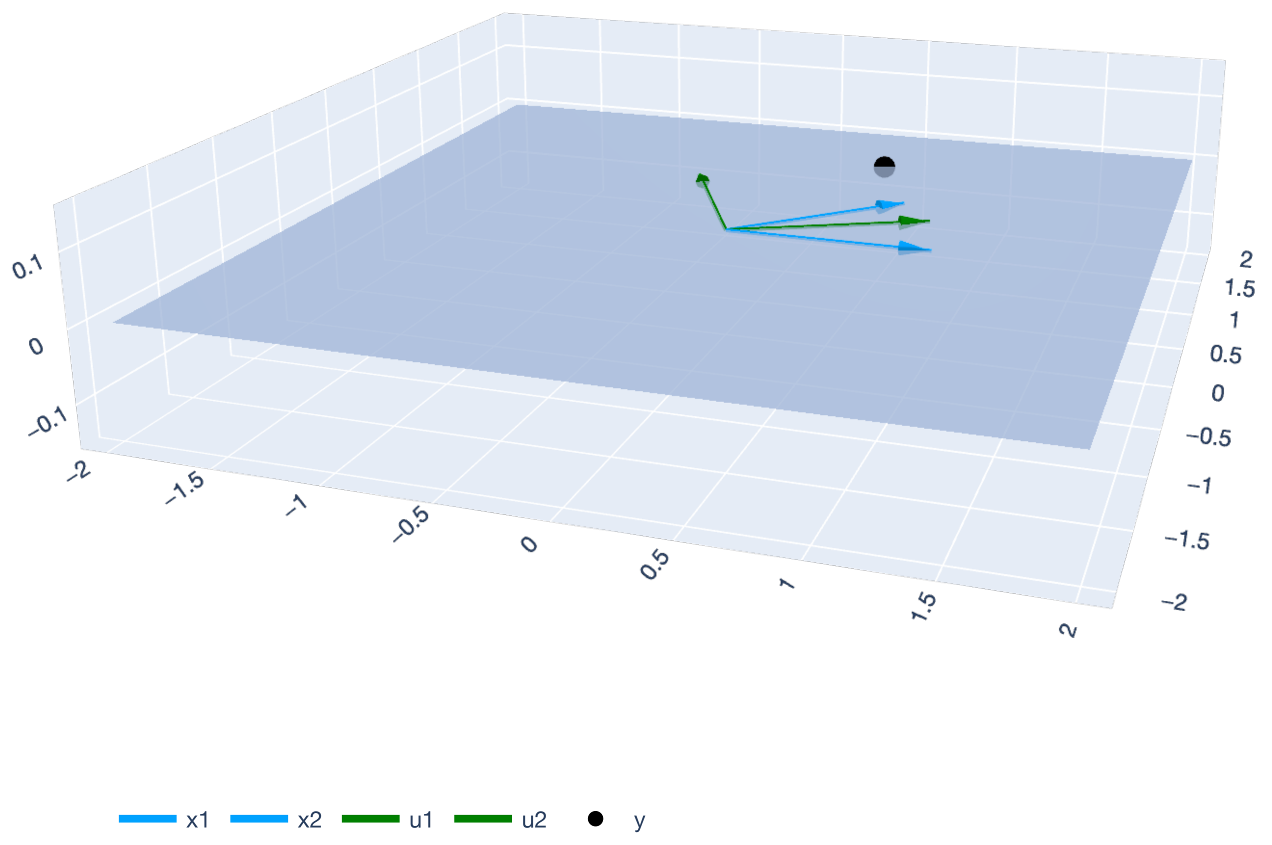


If $n < d$ and $\text{rank}(\mathbf{X}) = n$...

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

$y \in \text{col}(X)$

$n = 3$
$d = 2$



Singular Value Decomposition (SVD)

Matrix Decompositions

* IT APPLIES TO ANY MATRIX.

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

$\mathbf{U}$ is orthogonal, i.e. $\mathbf{U}^\top\mathbf{U} = \mathbf{U}\mathbf{U}^\top = \mathbf{I}$.

$\mathbf{V}$ 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. $\mathrm{rank}(\mathbf{X})$ is equal to the number of $\sigma_i > 0$.

$r = 5 \quad \sigma_5 \neq 0$

What other matrix decompositions are out there?

Eigendecomposition

Motivation: Linear Dynamical System

Population Change

Example of a linear dynamical system

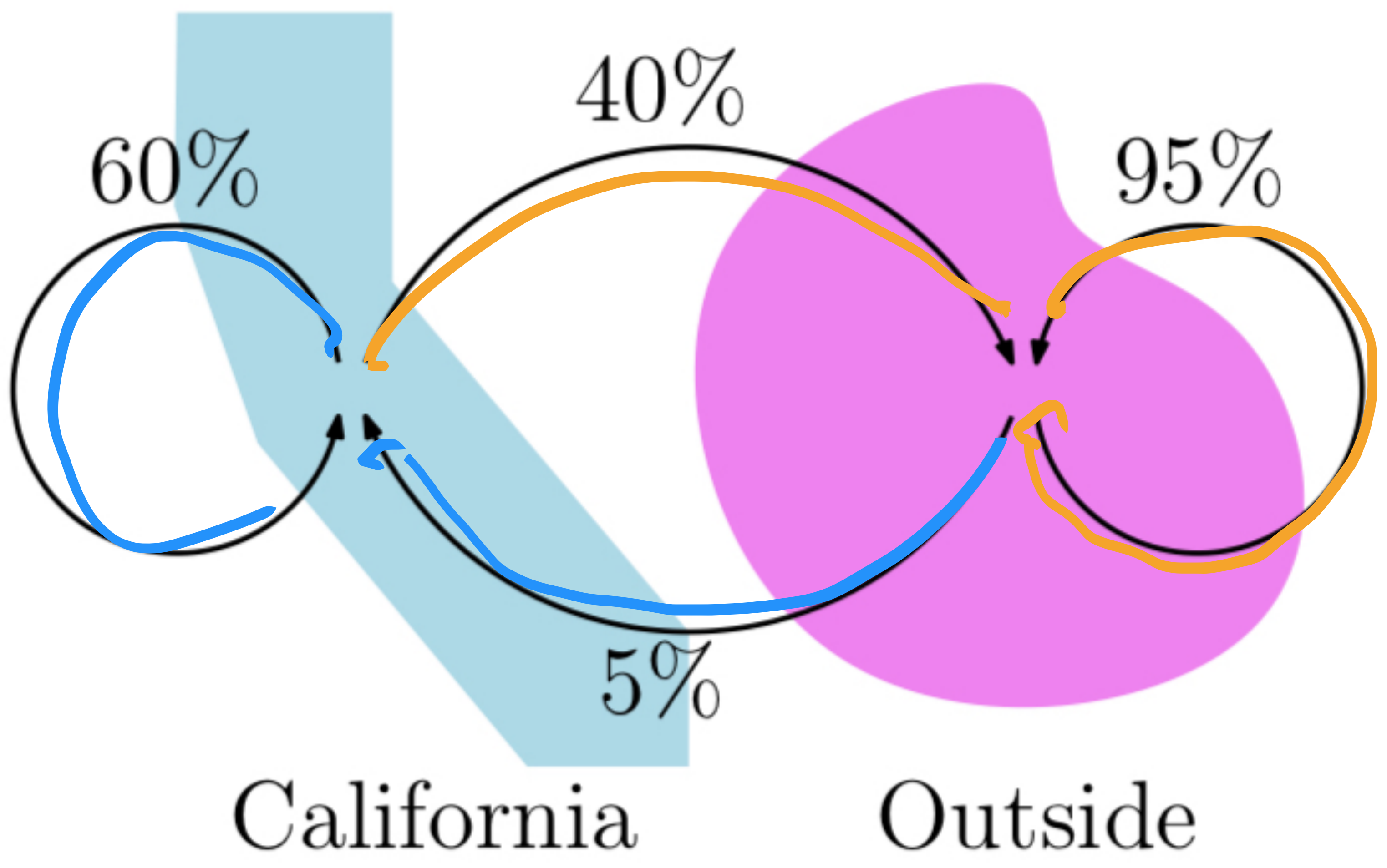
Consider the following example.

Suppose that

- of those who start a year in California, 60% stay in California and 40% move out of California by the end of the year.
- of those who start a year outside California, 95% stay out and 5% move to California by the end of the year.

If we know how many people are in California x_{in} and how many people are outside of California x_{out}, then we can find the number of people inside and outside of California at the end of the year:

$$\begin{aligned} \#\text{ inside} &= 0.6x_{in} + 0.05x_{out} \\ \#\text{ outside} &= 0.4x_{in} + 0.95x_{out} \end{aligned}$$



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

Population Change

Modeling with a transition matrix

Consider the following example.

Suppose that

- of those who start a year in California, 60% stay in California and 40% move out of California by the end of the year.
- of those who start a year outside California, 95% stay out and 5% move to California by the end of the year.

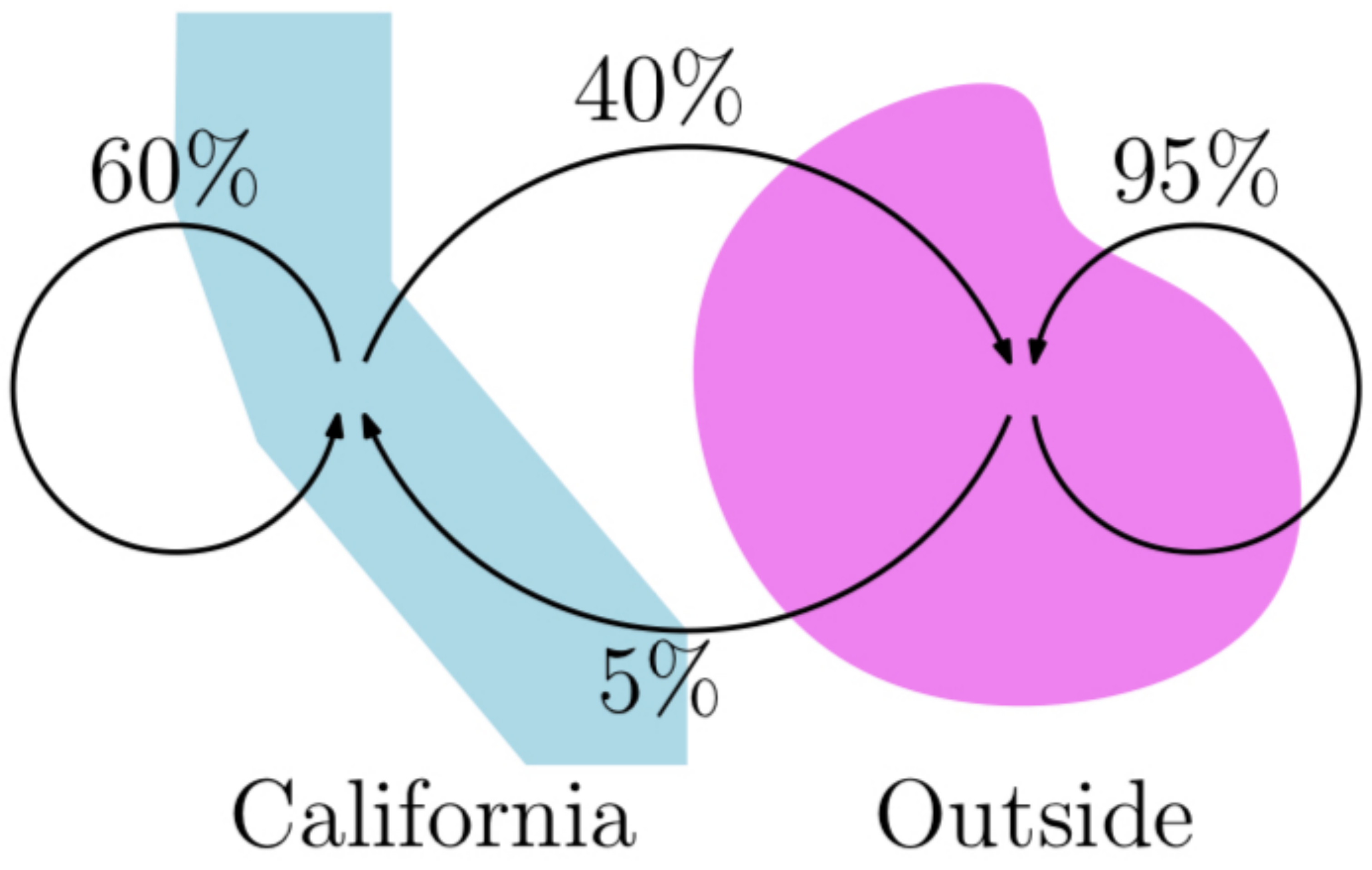
We can model this with a *transition matrix*

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

and a system of linear equations:

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

↑ # of people in / out at the end of year.



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

Population Change

Modeling with a transition matrix

Consider the transition matrix

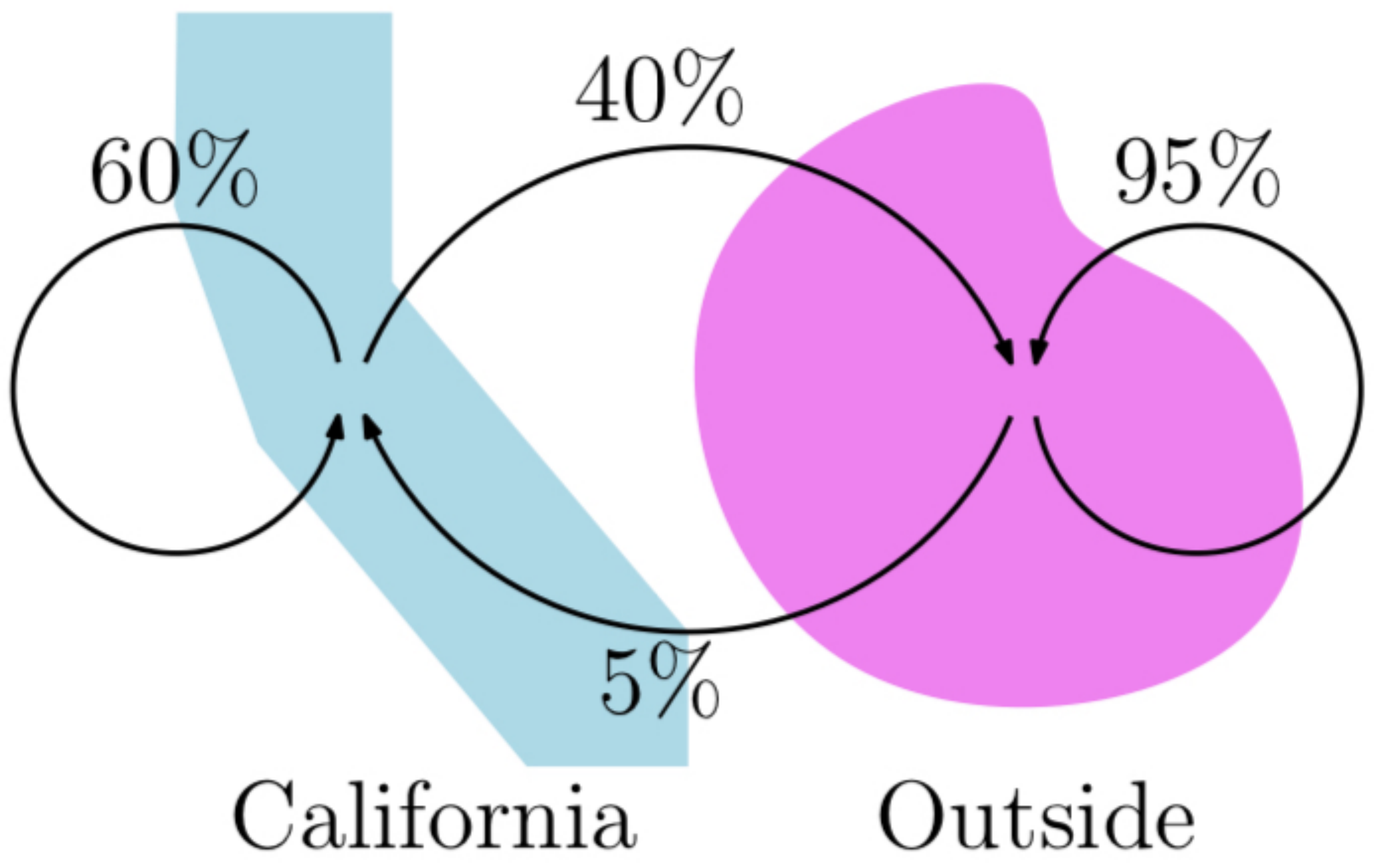
$$\mathbf{A} = \begin{bmatrix} \textit{in} \to \textit{in} & \textit{out} \to \textit{in} \\ \textit{in} \to \textit{out} & \textit{out} \to \textit{out} \end{bmatrix} = \boxed{\begin{bmatrix} 0.6 & 0.05 \\ 0.4 & 0.95 \end{bmatrix}}$$

with a corresponding system of linear equations:

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

The vector $\mathbf{Ax} \in \mathbb{R}^2$ gives the number of people inside and outside of California after a year has passed, 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)

Ax ← after 1 year.

$A(Ax)$ ← after 2 years A^2x. } $A^t x$

Population Change

Modeling with a transition matrix

Consider the transition matrix

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

with a corresponding system of linear equations:

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

The vector $\mathbf{Ax}^{(0)} \in \mathbb{R}^2$ gives the number of people inside and outside of California after a year has passed, from the initial populations in $\mathbf{x}^{(0)} \in \mathbb{R}^2$.

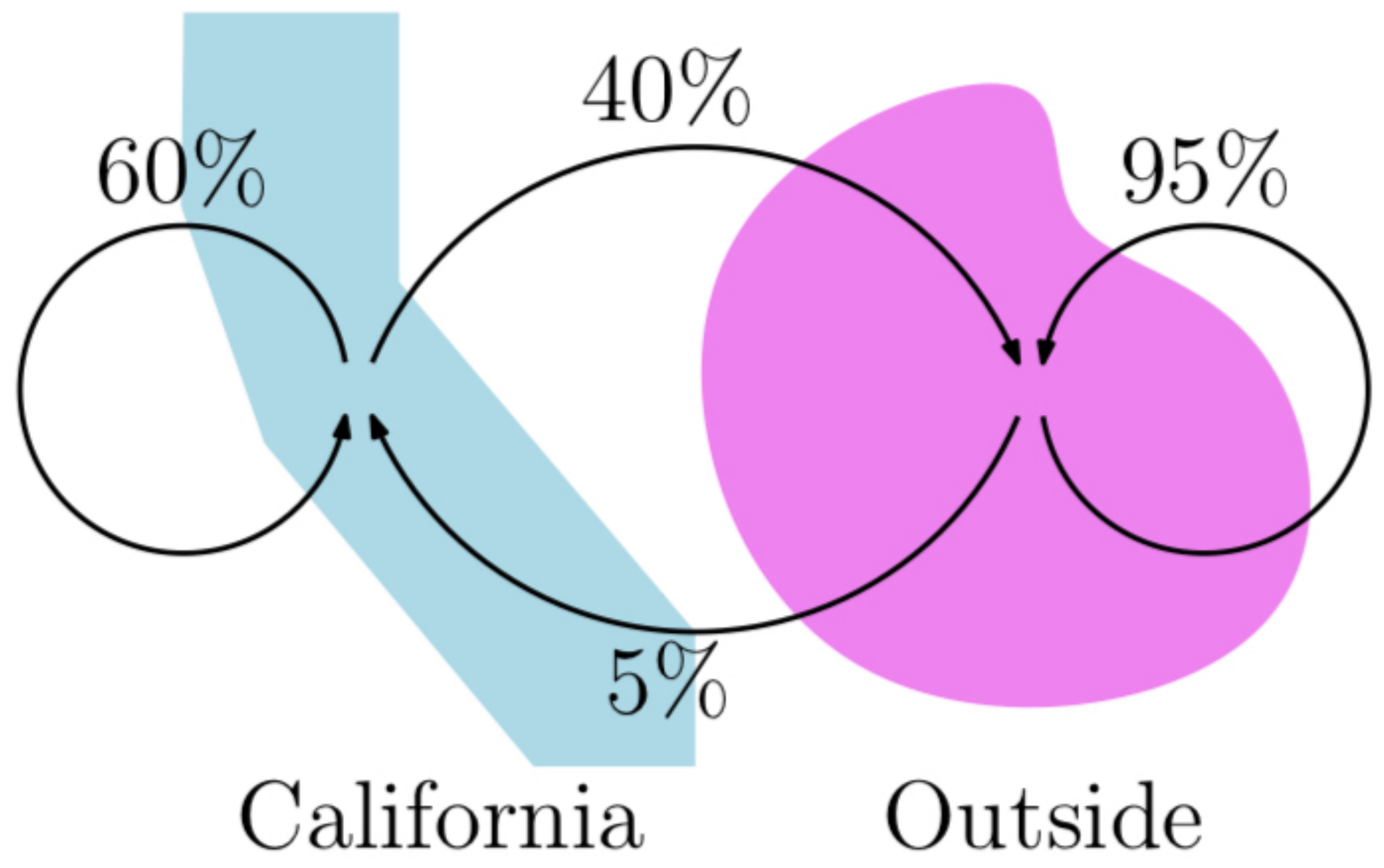
How to find the number of people inside/outside of California after t years have passed?

$$\mathbf{x}^{(1)} = \mathbf{Ax}^{(0)}$$

$$\mathbf{x}^{(2)} = \mathbf{Ax}^{(1)} = \mathbf{AAx}^{(0)} = \mathbf{A}^2\mathbf{x}^{(0)}$$

$$\vdots$$

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



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

Population Change

Modeling with a transition matrix

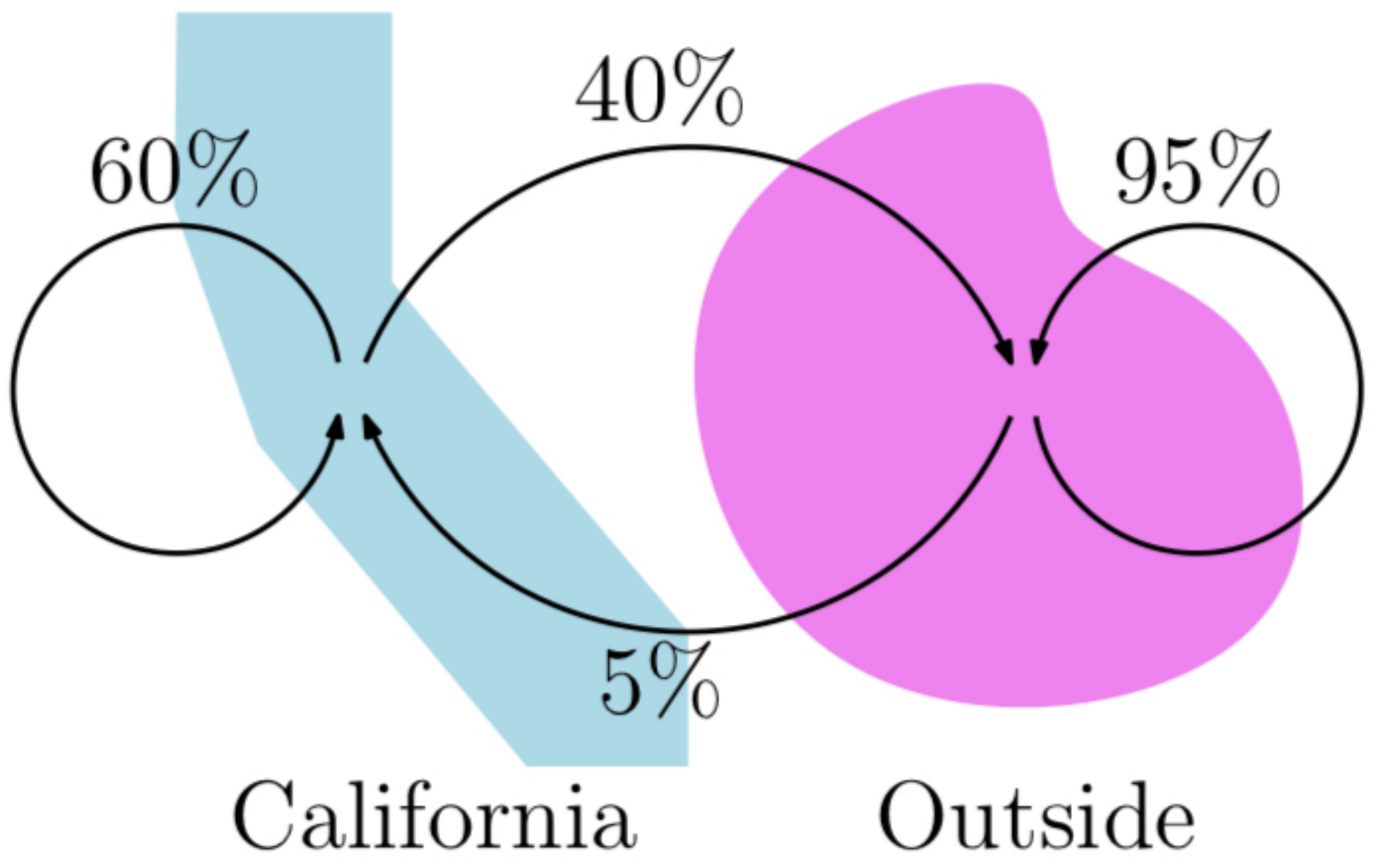
$$\mathbf{A}\mathbf{x} = \begin{bmatrix} \text{in} \to \text{in} & \text{out} \to \text{in} \\ \text{in} \to \text{out} & \text{out} \to \text{out} \end{bmatrix} \begin{bmatrix} x_{in} \\ x_{out} \end{bmatrix} = \begin{bmatrix} 0.6 & 0.05 \\ 0.4 & 0.95 \end{bmatrix} \begin{bmatrix} x_{in} \\ x_{out}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

Concretely, suppose there are 300 million outside of California and 40 million inside of California at the start of a year. Then,

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

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

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



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…

$$\underbrace{\begin{bmatrix} 0.6 & 0.05 \\ 0.4 & 0.95 \end{bmatrix} \cdots \begin{bmatrix} 0.6 & 0.05 \\ 0.4 & 0.95 \end{bmatrix} \begin{bmatrix} 0.6 & 0.05 \\ 0.4 & 0.95 \end{bmatrix}}_{t \text{ times.}} \begin{bmatrix} 40 \\ 300 \end{bmatrix}$$

Population Change

Easy computation 😃

Magically

Assume I gave you a couple of vectors, $\boxed{\mathbf{u} = (1,8) \text{ and } \mathbf{v} = (-1,1).}$ These two vectors have the properties:

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

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

Population Change

Easy computation 😃

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

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

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

Now, the repeated multiplication looks like:

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

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

Population Change

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

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

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

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

Now, the repeated multiplication looks like:

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

$$\mathbf{A}^t\mathbf{v} = \begin{bmatrix} 0.6 & 0.05 \\ 0.4 & 0.95 \end{bmatrix}^t \begin{bmatrix} -1 \\ 1 \end{bmatrix} = \left(\frac{11}{20}\right)^t \begin{bmatrix} -1 \\ 1 \end{bmatrix} \implies \boxed{\mathbf{A}^t\mathbf{v} = \left(\frac{11}{20}\right)^t \mathbf{v}} \; ②$$

Population Change

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

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

$$\boxed{\begin{aligned} &① \quad \mathbf{A}^t\mathbf{u} = \mathbf{u} \\ &② \quad \mathbf{A}^t\mathbf{v} = \left(\frac{11}{20}\right)^t \mathbf{v} \end{aligned}}$$

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.

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

scalars

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

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

any starting

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:

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

In matrix form:

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

Population Change

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

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

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

where

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

Writing $\mathbf{x}^{(0)}$ in matrix form as well, we have:

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

Population Change

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

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

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

where

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

Writing $\mathbf{x}^{(0)}$ in matrix form as well, we have:

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

Because $\mathbf{u}$ and $\mathbf{v}$ are linearly independent, $\mathbf{V} \in \mathbb{R}^{2\times 2}$ has $\mathrm{rank}(\mathbf{V}) = 2$, so we can invert:

$$\begin{bmatrix} a \\ b \end{bmatrix} = \mathbf{V}^{-1}\mathbf{x}^{(0)}.$$

$$V^{-1} = \begin{bmatrix} | & | \\ u & v \\ | & | \end{bmatrix}^{-1}$$

Population Change

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

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

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

where

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

Writing $\mathbf{x}^{(0)}$ in matrix form as well, we have:

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

Because $\mathbf{u}$ and $\mathbf{v}$ are linearly independent, $\mathbf{V} \in \mathbb{R}^{2\times 2}$ has $\operatorname{rank}(\mathbf{V}) = 2$, so we can invert:

$$\begin{bmatrix} a \\ b \end{bmatrix} = \mathbf{V}^{-1}\mathbf{x}^{(0)}.$$

Therefore,

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

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

$A^t \begin{bmatrix}1\\8\end{bmatrix} = \begin{bmatrix}1\\8\end{bmatrix}$

Hard computation:

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

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

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

😖

Easy computation:

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

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

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

😃

Diagonal Matrices

Why we like diagonal matrices

Multiplying diagonal matrices with themselves many times is easy:

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

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:

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

depended on A

(transition matrix)

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

Eigendecomposition

Intuition and Definition

Eigenvectors and eigenvalues

Intuition

$T_A : \mathbb{R}^d \to \mathbb{R}^d$

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

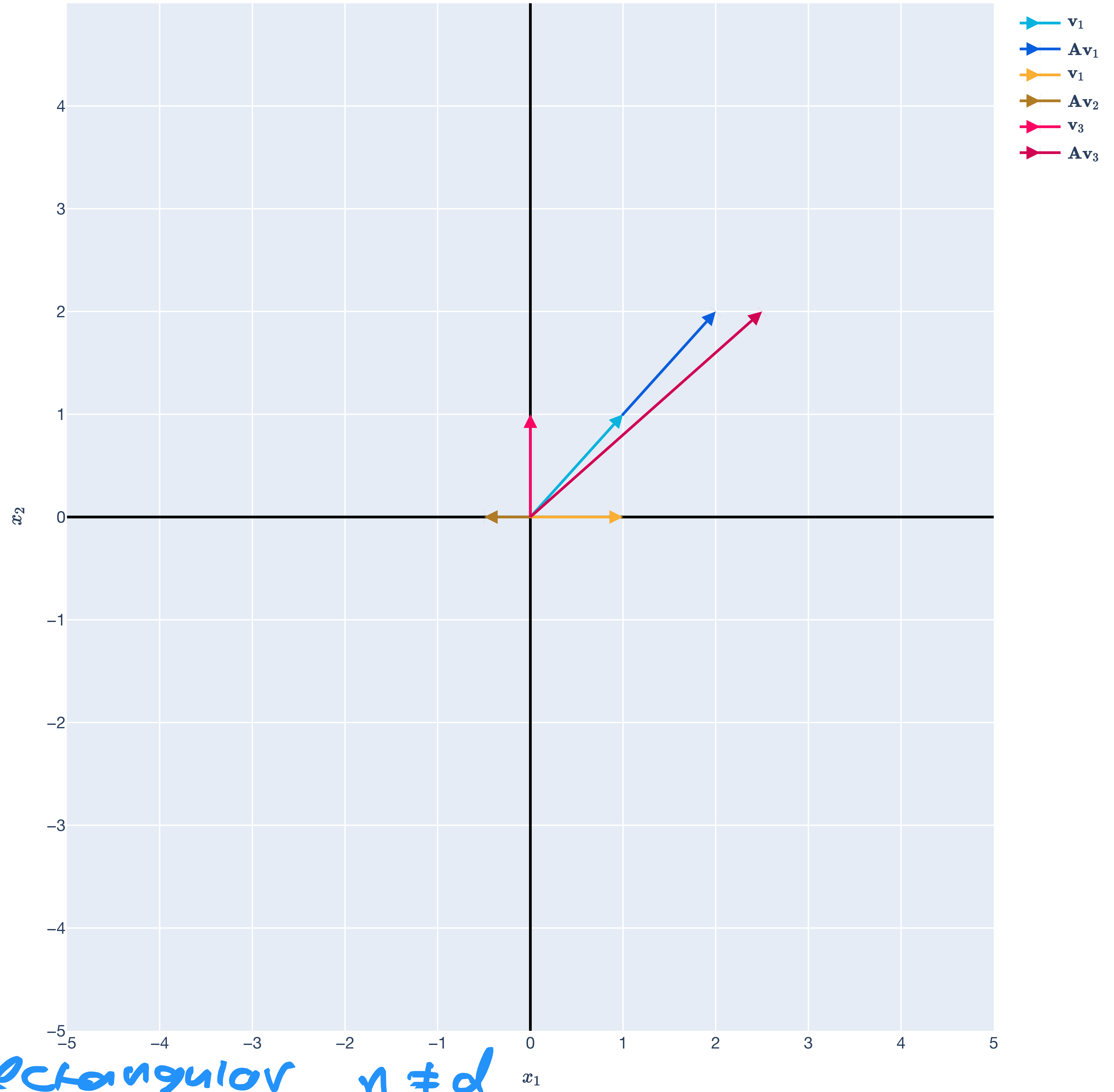
This represents a linear transformation from $\mathbb{R}^d$ to $\mathbb{R}^d$.

Eigenvectors are the vectors in $\mathbb{R}^d$ that just get scaled by $\mathbf{A}$.

Eigenvalues are how much each eigenvector gets scaled.

Eigenvectors/eigenvalues are properties of square matrices!

doesn't make sense to ask for rectangular $n \neq d$.



Eigenvectors and eigenvalues

Definition

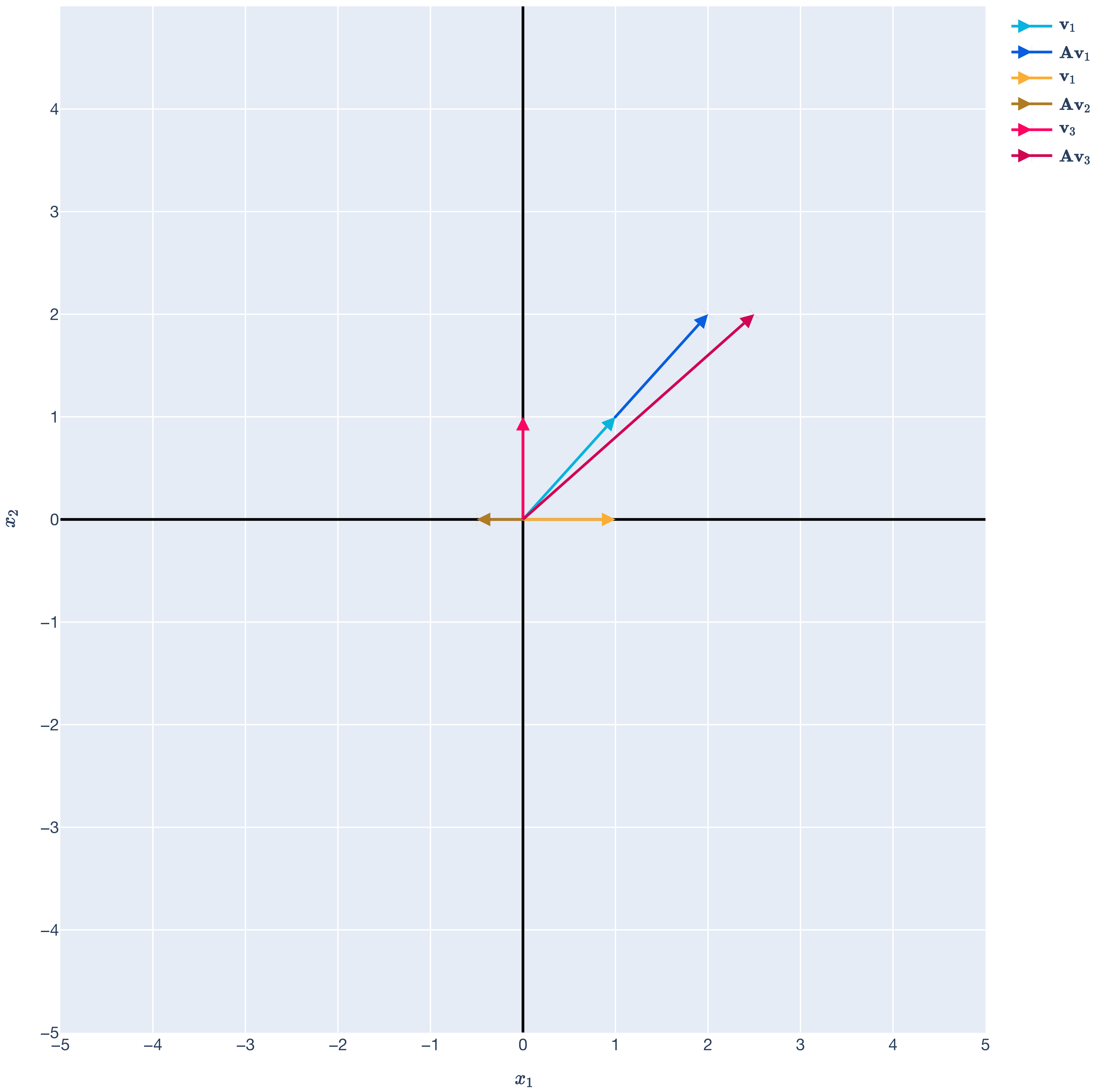
Let $\mathbf{A} \in \mathbb{R}^{d \times d}$ be a *square* matrix.

A nonzero vector $\mathbf{v} \in \mathbb{R}^d$ is an ***eigenvector*** if there exists a scalar $\lambda \in \mathbb{R}$ such that

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

The scalar λ is the ***eigenvalue*** associated with the eigenvector $\mathbf{v}$.

Eigenvectors/eigenvalues are properties of square matrices!



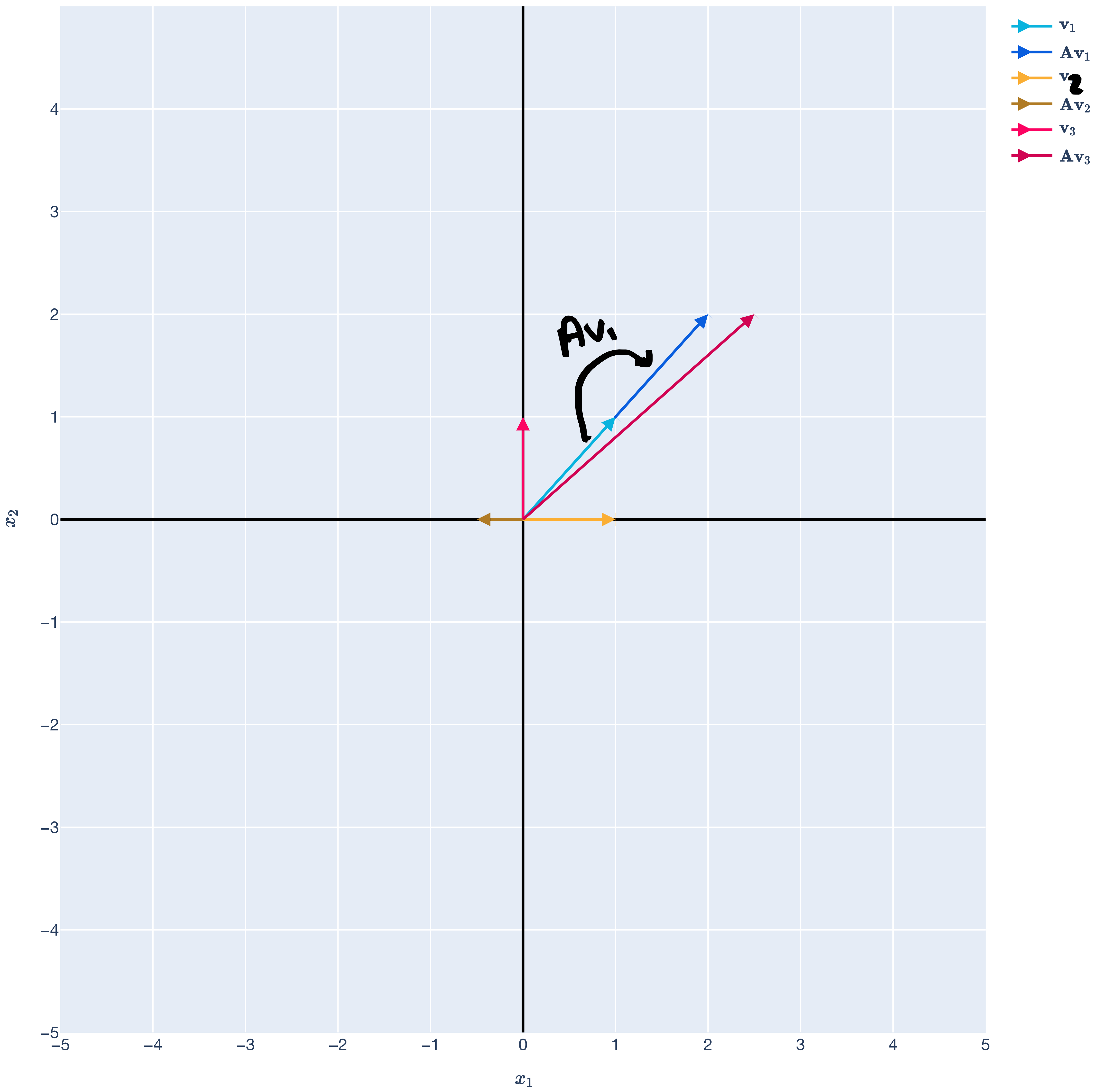
Eigenvectors and eigenvalues

Example

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

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

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



Eigenvectors and eigenvalues

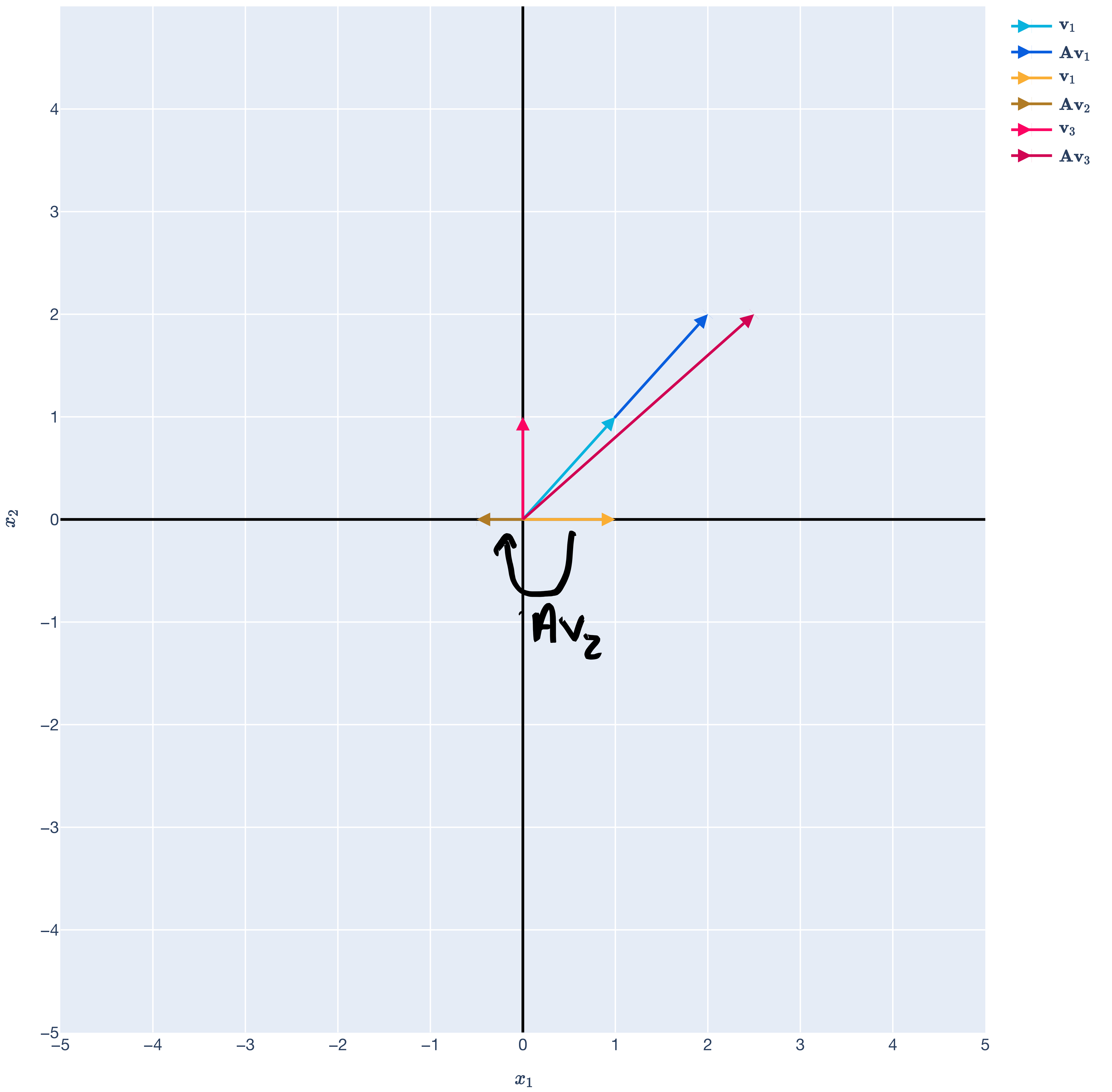
Example

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 $\mathbf{v}_2 = (1,0)$?

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



Eigenvectors and eigenvalues

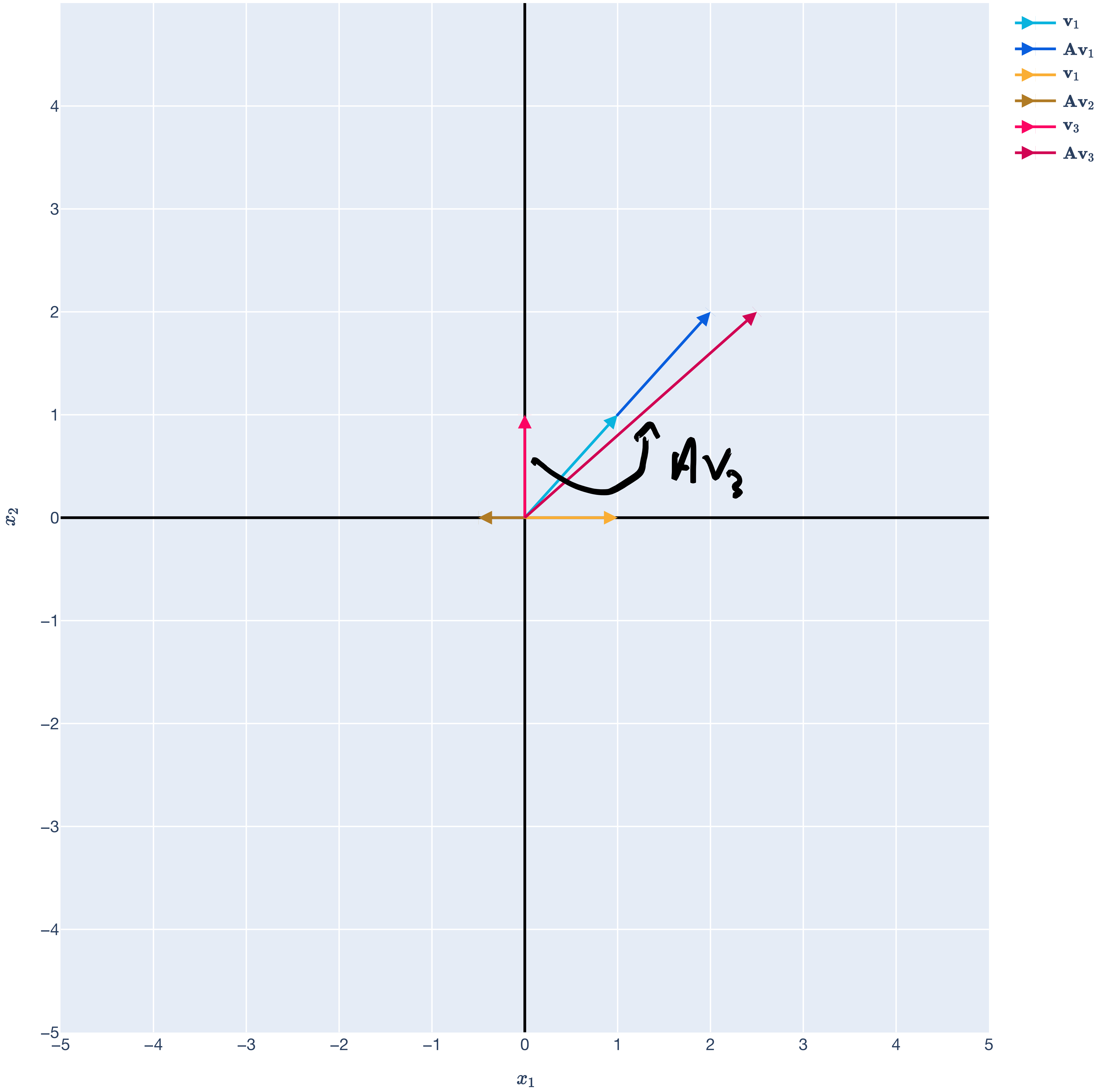
Example

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 $\mathbf{v}_3 = (0,1)$?

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



Eigenvectors and eigenvalues

Example

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

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

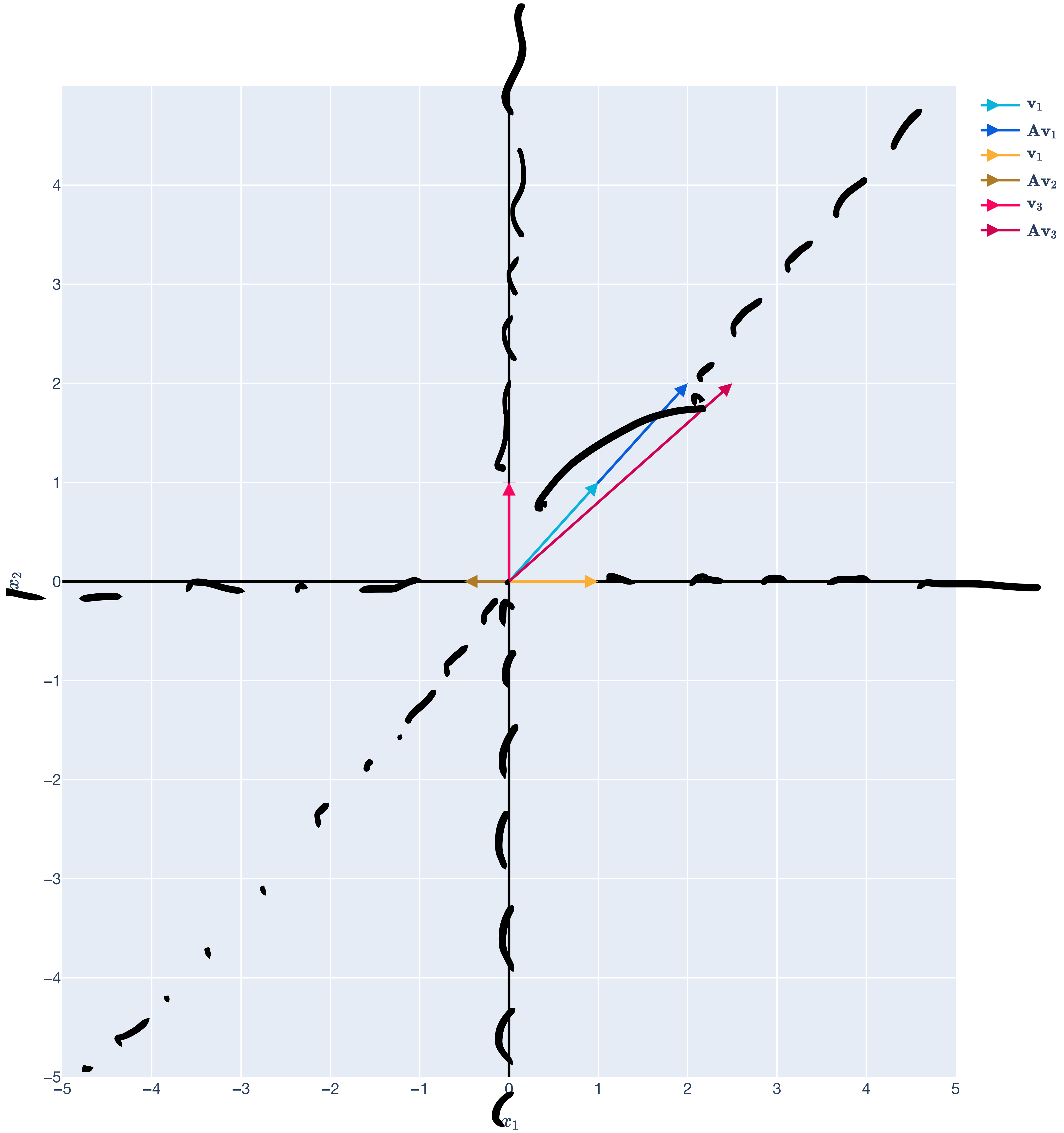
Eigenvectors (with eigenvalues $\lambda_1 = 2$ and $\lambda_2 = -1/2$):

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

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

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

3Blue1Brown

Example

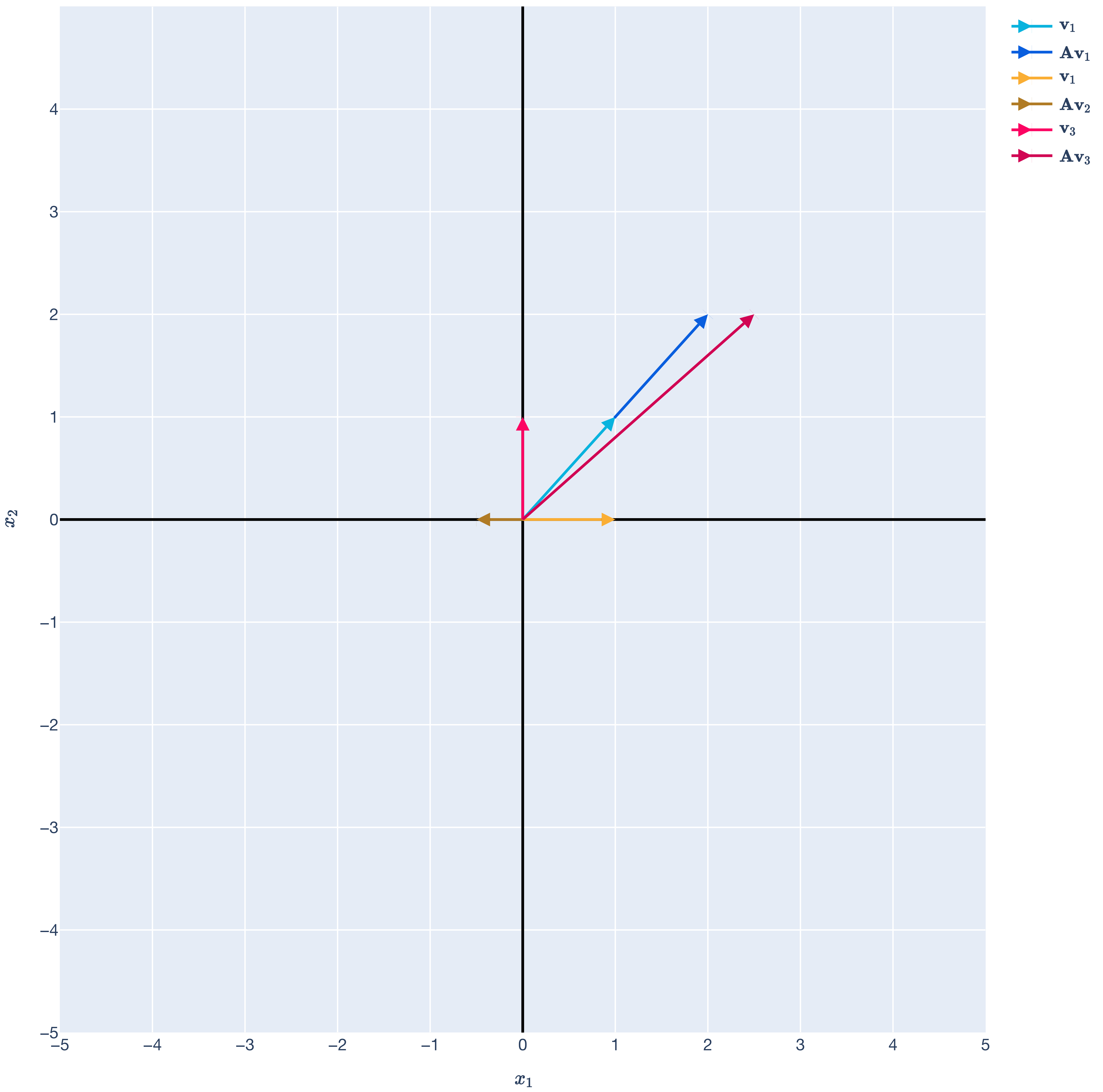
$$\mathbf{A} = \begin{bmatrix} -1/2 & 5/2 \\ 0 & 2 \end{bmatrix}$$

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

We can write any $\mathbf{x} \in \mathbb{R}^2$ in terms of $\mathbf{v}_1$ and $\mathbf{v}_2$:

$$\mathbf{x} = a\mathbf{v}_1 + b\mathbf{v}_2.$$

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



Eigenvectors and eigenvalues

Example

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

$\mathbf{v}_1 = (1,1)$ and $\mathbf{v}_2 = (1,0)$ are linearly independent *eigenvectors* — they form a basis for $\mathbb{R}^2$. Their *eigenvalues* are $\lambda_1 = 2$ and $\lambda_2 = -1/2$.

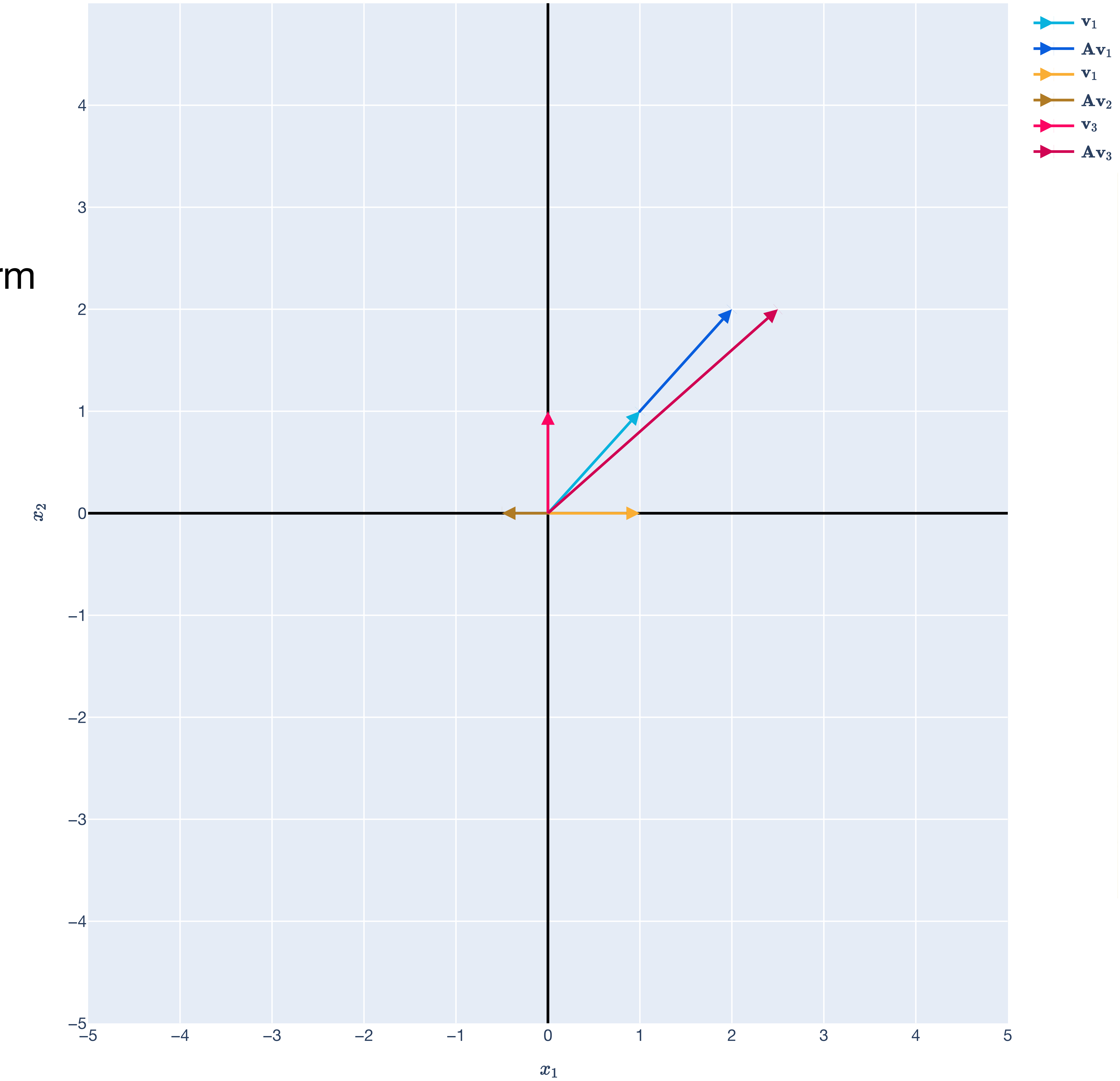
We can write any $\mathbf{x} \in \mathbb{R}^2$ in terms of $\mathbf{v}_1$ and $\mathbf{v}_2$:

$$\mathbf{x} = a\mathbf{v}_1 + b\mathbf{v}_2.$$

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

Repeated multiplication:

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



Eigenvectors and eigenvalues

Example

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

$\mathbf{v}_1 = (1,1)$ and $\mathbf{v}_2 = (1,0)$ are linearly independent *eigenvectors* — they form a basis for $\mathbb{R}^2$. Their *eigenvalues* are $\lambda_1 = 2$ and $\lambda_2 = -1/2$.

We can write any $\mathbf{x} \in \mathbb{R}^2$ in terms of $\mathbf{v}_1$ and $\mathbf{v}_2$:

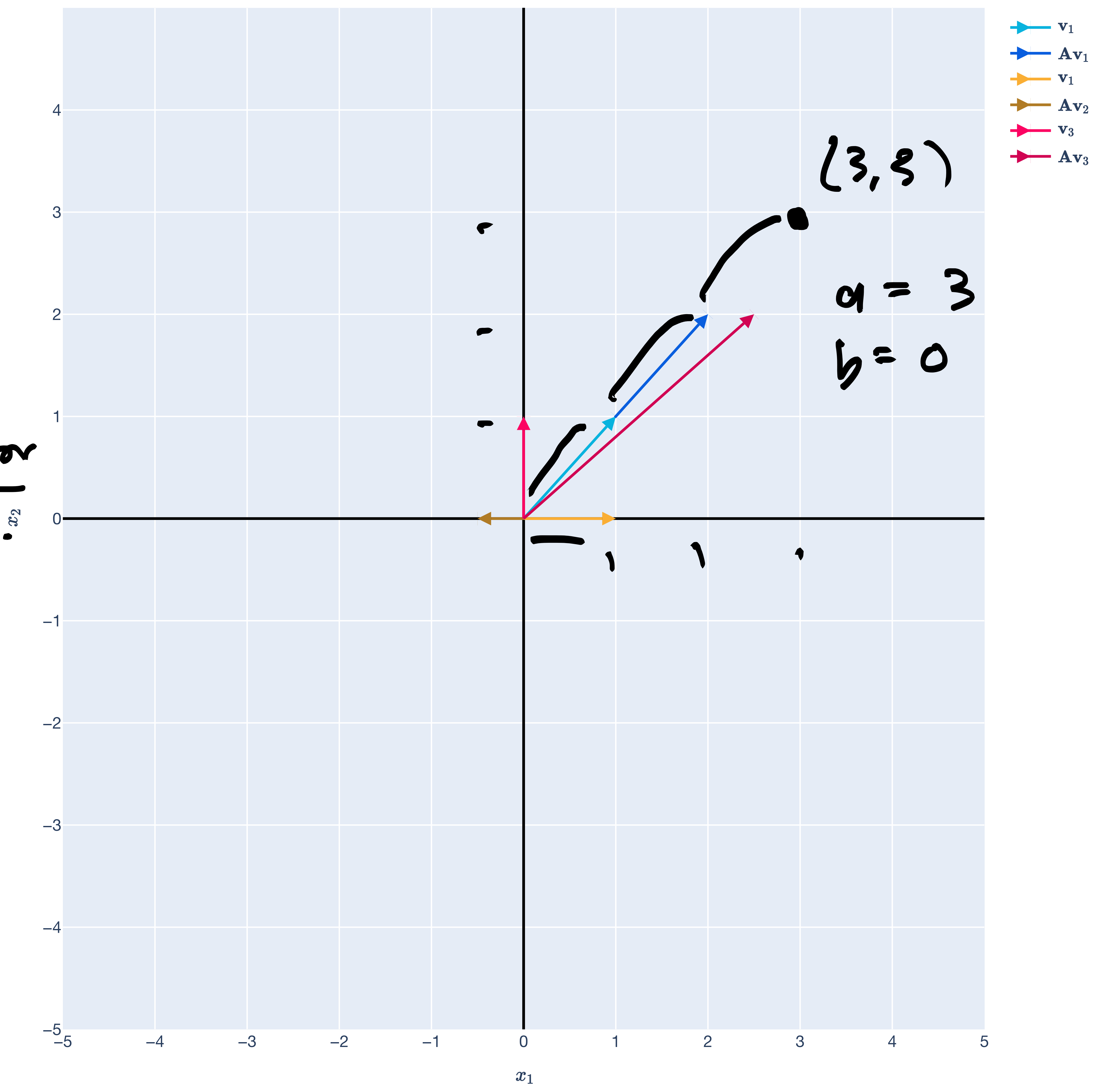
$$\mathbf{x} = a\mathbf{v}_1 + b\mathbf{v}_2.$$

coordinates of x in the eigenvector basis.

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

Repeated multiplication:

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



Eigenvectors and eigenvalues

Example

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

$\mathbf{v}_1 = (1,1)$ and $\mathbf{v}_2 = (1,0)$ are linearly independent *eigenvectors* — they form a basis for $\mathbb{R}^2$. Their *eigenvalues* are $\lambda_1 = 2$ and $\lambda_2 = -1/2$.

We can write any $\mathbf{x} \in \mathbb{R}^2$ in terms of $\mathbf{v}_1$ and $\mathbf{v}_2$:

$$\mathbf{x} = a\mathbf{v}_1 + b\mathbf{v}_2.$$

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

Repeated multiplication:

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

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:

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

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

Eigendecomposition

Definition

$\boxed{\begin{bmatrix} a \\ b \end{bmatrix} = V^{-1}x}$

$V\begin{bmatrix} a \\ b \end{bmatrix} = x$

$\begin{bmatrix} | & | \\ v_1 & v_2 \\ | & | \end{bmatrix}\begin{bmatrix} a \\ b \end{bmatrix} = x$

Prop (Eigendecomposition of a diagonalizable matrix). Let $\mathbf{A} \in \mathbb{R}^{d \times d}$ be a matrix with d linearly independent eigenvectors

$$\mathbf{A}\mathbf{v}_1 = \lambda_1 \mathbf{v}_1$$
$$\vdots$$
$$\mathbf{A}\mathbf{v}_d = \lambda_d \mathbf{v}_d$$

$V^{-1}x \to \lambda$

scales.

Then, $\mathbf{A}$ has the *eigendecomposition*:

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

Such a matrix is said to be *diagonalizable.*

$\to$ Eigendecomposition exists.

Eigendecomposition

Example

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

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

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

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

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

Eigendecomposition

Example

$\det(A - \lambda I) = 0$ X

numpy: $\boxed{\text{np.linalg.eig}}$

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

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

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

$$\mathbf{A} = \mathbf{Q}\mathbf{\Lambda}\mathbf{Q}^{-1}$$

$$\begin{bmatrix} -1/2 & 5/2 \\ 0 & 2 \end{bmatrix} = \underset{Q}{\begin{bmatrix} 1 & 1 \\ 1 & 0 \end{bmatrix}} \underset{\Lambda}{\begin{bmatrix} 2 & 0 \\ 0 & -1/2 \end{bmatrix}} \underset{Q^{-1}}{\begin{bmatrix} 0 & 1 \\ 1 & -1 \end{bmatrix}}$$

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

Eigendecomposition

Connection with SVD

Connection with SVD

Eigendecomposition from SVD

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

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

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

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

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

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:

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

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:

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

$$A = \underbrace{X^\top X}$$

Connection with SVD

Eigendecomposition from SVD

PS 2: for Proof. $n \times n$ $\quad n \times d,\ d \times d$

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 singular values

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

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

right singular vecs

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 $\boxed{\lambda_i = \sigma_i^2}$, 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, if $\boxed{\mathbf{A} = \mathbf{X}^\top\mathbf{X}}$ (for *any* matrix $\boxed{\mathbf{X} \in \mathbb{R}^{n\times d}}$), we know that we have d linearly independent eigenvectors — this is a case where $\mathbf{A}$ is diagonalizable!

Moreover, the diagonalization looks like:

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

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

$$\begin{bmatrix} \lambda_1 & & & & \\ & \ddots & & & \\ & & \lambda_r & & \\ & & & 0 & \\ & & & & 0 \end{bmatrix}$$

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

$$\boxed{A = X^T X}$$

Positive Semidefinite Matrices

Definition and Connections

Positive Semidefinite (PSD) Matrices

First definition

A square matrix $\mathbf{A} \in \mathbb{R}^{d \times d}$ is ***positive semidefinite (PSD)*** if there exists a matrix $\mathbf{X} \in \mathbb{R}^{n \times d}$ such that:

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

Note: If you've seen PSD matrices before, this isn't the usual definition (but it's equivalent, as we'll see in a bit).

Positive Semidefinite (PSD) Matrices

Symmetry of PSD Matrices

A square matrix $\mathbf{A} \in \mathbb{R}^{d \times d}$ is ***positive semidefinite (PSD)*** if there exists a matrix $\mathbf{X} \in \mathbb{R}^{n \times d}$ such that:

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

$$A^T = (X^T X)^T = X^T X = A.$$

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

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

Positive Semidefinite (PSD) Matrices

Example

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

Positive Semidefinite (PSD) Matrices

Example

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

Its “square root” is the matrix

$$\mathbf{X} = \begin{bmatrix} \frac{2}{\sqrt{2}} & \frac{2}{\sqrt{2}} \\ \frac{1}{\sqrt{2}} & -\frac{1}{\sqrt{2}} \\ 0 & 0 \end{bmatrix}.$$

To verify:

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

PSD Matrices and Eigendecomposition

Connection to eigenvalues

$X = V\Sigma V^T$

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

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

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 $\lambda_1, \ldots, \lambda_d \geq 0$.

d nonnegative eigenvalues.

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 $\mathbf{v}_1 = \left(\frac{1}{\sqrt{2}}, \frac{1}{\sqrt{2}}\right)$ and $\mathbf{v}_2 = \left(\frac{1}{\sqrt{2}}, -\frac{1}{\sqrt{2}}\right)$:

$$\left.\begin{array}{l} \mathbf{A}\mathbf{v}_1 = \begin{bmatrix} 5/2 & 3/2 \\ 3/2 & 5/2 \end{bmatrix} \begin{bmatrix} 1/\sqrt{2} \\ 1/\sqrt{2} \end{bmatrix} = \begin{bmatrix} 4/\sqrt{2} \\ 4/\sqrt{2} \end{bmatrix} = 4 \begin{bmatrix} 1/\sqrt{2} \\ 1/\sqrt{2} \end{bmatrix} \implies \lambda_1 = 4 \\ \mathbf{A}\mathbf{v}_2 = \begin{bmatrix} 5/2 & 3/2 \\ 3/2 & 5/2 \end{bmatrix} \begin{bmatrix} 1/\sqrt{2} \\ -1/\sqrt{2} \end{bmatrix} = \begin{bmatrix} 1/\sqrt{2} \\ -1/\sqrt{2} \end{bmatrix} \implies \lambda_1 = 1 \end{array}\right\} \quad 4, 1 \geq 0.$$

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

Positive Semidefinite (PSD) Matrices

Third definition

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

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

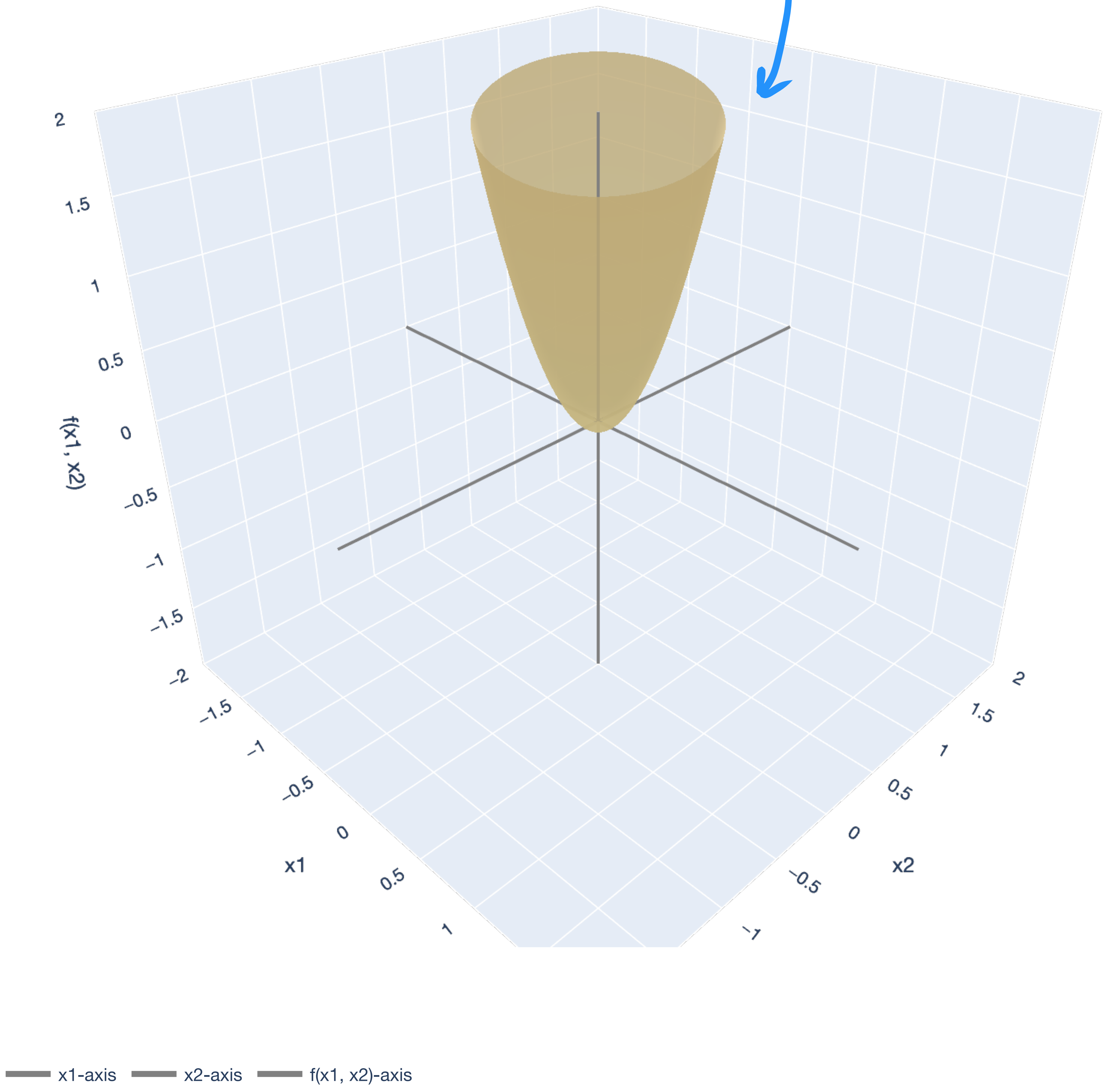
This is often taken as the definition of PSD (but it is equivalent to the other two definitions in previous slides).

$$\underbrace{x^\top}_{\mathbb{R}^d} A \underbrace{x}_{\mathbb{R}^d} \in \mathbb{R}.$$

Positive Semidefinite (PSD) Matrices

Example

$f(x_1, x_2) = x^\top A x$

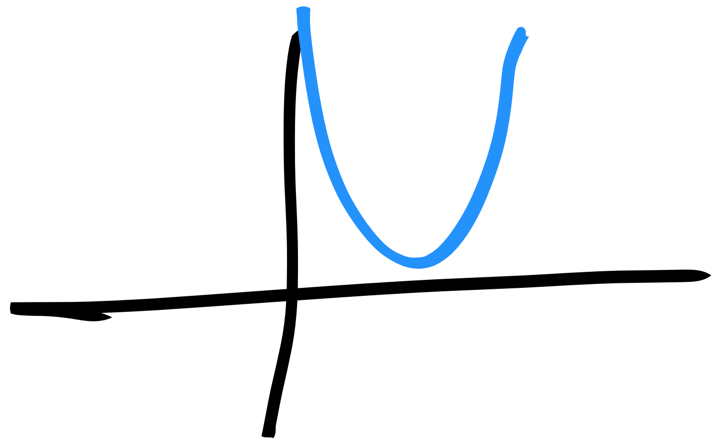


$$\mathbf{A} = \begin{bmatrix} 5/2 & 3/2 \\ 3/2 & 5/2 \end{bmatrix}$$ is positive semidefinite.

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

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

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



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…

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

$$\updownarrow$$

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

$\lambda_1 = 4$

$\lambda_2 = 1.$

$$\updownarrow$$

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

stricter

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)?*

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

$A^T = A$

$A = X^T X \rightarrow$ SVD $\rightarrow$ $v_1, \ldots, v_d$ is a basis

But even more generally…

Spectral Theorem

Statement

$A^T = A$.

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: $\mathbf{A}$ has an orthonormal basis of d eigenvectors and an eigendecomposition

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

$$\boxed{X = U \Sigma V^T}$$ SVD works for any $n \times d$.

But, in this generality, λ_i can be negative!

$\mathbb{R}^{100} \longrightarrow \mathbb{R}^3$ or $\mathbb{R}^2$.

Principal Components Analysis

Application of Eigendecomposition

Principal Components Analysis ≡ Eigendecomposition

Example: "Eigenfaces" and facial recognition

pixels

1000

1000

Observed: Matrix of *training images* $\mathbf{X} \in \mathbb{R}^{n \times d}$:

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

Each row is a "flattened" image vector. Typically, each pixel is in $[0, 255]$ for grayscale images.

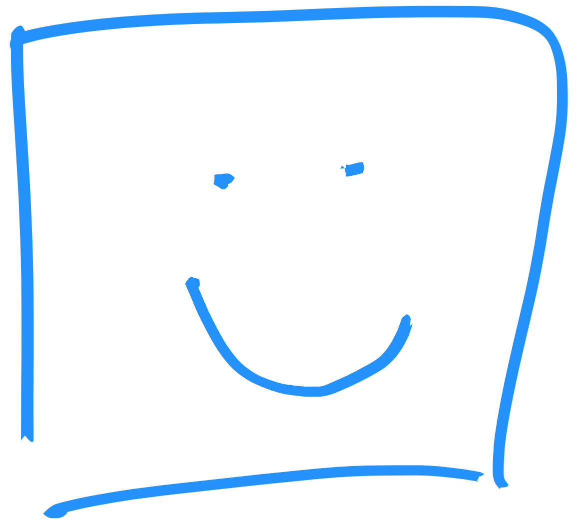
Images are very high-dimensional: $d =$ width in pixels $\times$ height in pixels (e.g. $d = 1080 \times 1080 = 1{,}166{,}400$).

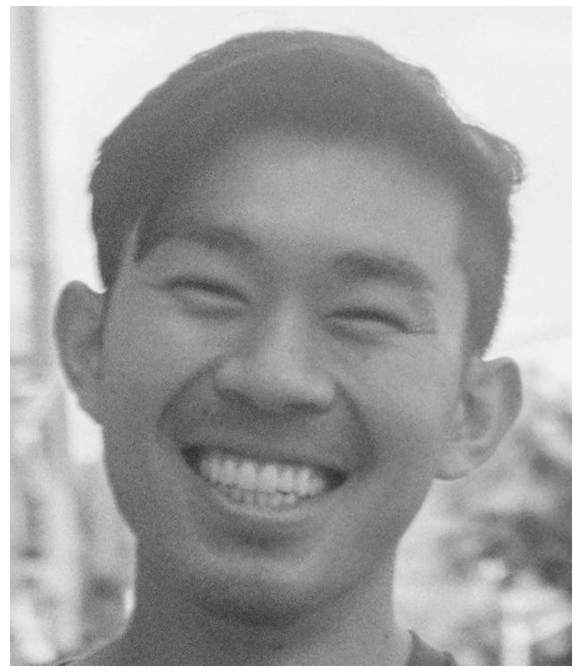
Principal Components Analysis

Example: “Eigenfaces” and facial recognition

1,000, 600

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

$x_0 =$ 

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 $\underbrace{||\mathbf{x}_0 - \mathbf{x}_i||}$).

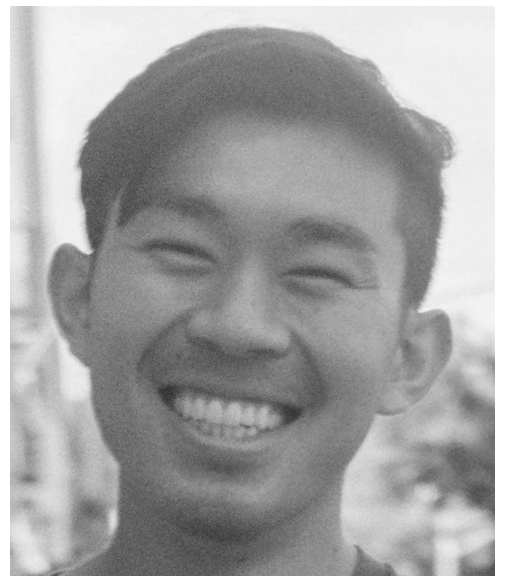
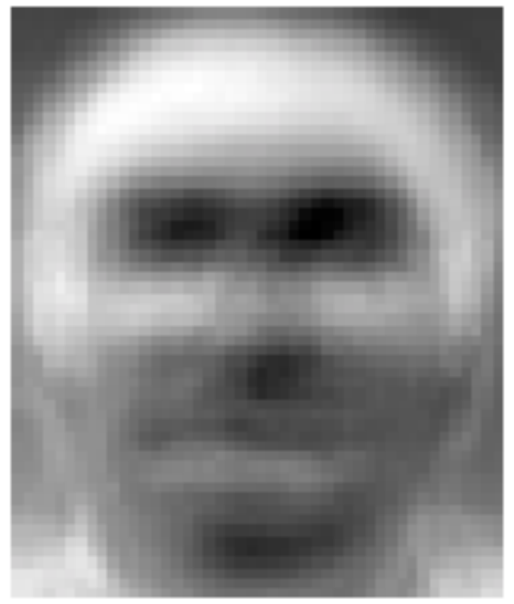
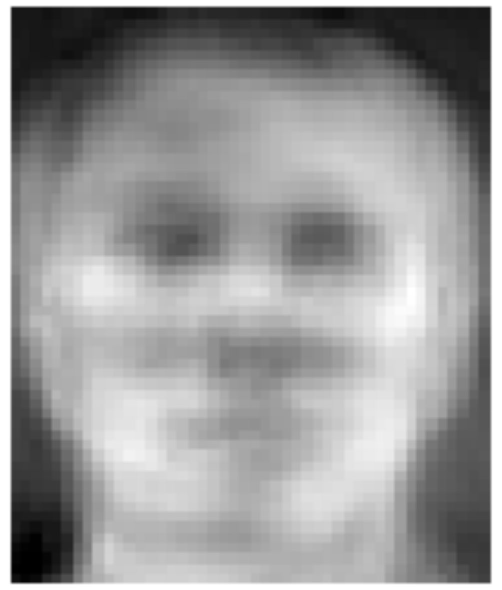
Storage: $\underbrace{\text{1166400 integers}}_{\text{Byte.}} \times \underbrace{1000 \text{ images}} \approx \underline{1 \text{ GB}}$.

Principal Components Analysis

Example: "Eigenfaces" and facial recognition

Suppose we can find a "basis" of representative faces: $\underbrace{\mathbf{v}_1, \ldots, \mathbf{v}_k}$ where $k \ll n$. ← 1000

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

 $= \underbrace{0.45}_{\hat{w}_1}$  $+ \underbrace{0.21}_{\hat{w}_2}$  $+ \underbrace{0.12}_{\hat{w}_3}$  $+ \; 0.05$  $+ \ldots$

v_1 $\quad v_2 \quad v_3$

Improved facial recognition: Store k "eigenfaces." Given a new face $\mathbf{x}_0$, project the face onto the subspace spanned by the eigenfaces to get $\Pi(\mathbf{x}_0)$. Compare $\Pi(\mathbf{x}_0)$ to each face's projection in the database in Euclidean norm $\|\Pi(\mathbf{x}_0) - \Pi(\mathbf{x}_i)\|$.

$\min \| X_0 - V\hat{w} \|^2$

$\hat{w} \to (V\hat{w}) \leftarrow \boxed{\hat{w} \in \mathbb{R}^k}$

Principal Components Analysis

Example: PCA in 2D

<u>**Observed:**</u> Matrix of *training points* $\mathbf{X} \in \mathbb{R}^{n \times 2}$:

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

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

Principal Components Analysis

Example: PCA in 2D

Observed: Matrix of *training points* $\mathbf{X} \in \mathbb{R}^{n\times 2}$:

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

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

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

Principal Components Analysis

Example: PCA in 2D

<u>Observed:</u> Matrix of *training points* $\mathbf{X} \in \mathbb{R}^{n\times 2}$:

positive semidefinite

$$\mathbf{X} = \begin{bmatrix} x_{11} & x_{12} \\ x_{21} & x_{22} \\ \vdots & \vdots \\ x_{n1} & x_{n2} \end{bmatrix} = \begin{bmatrix} \uparrow & \uparrow \\ \mathbf{x}_1 & \mathbf{x}_2 \\ \downarrow & \downarrow \end{bmatrix}$$

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

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

Principal Components Analysis

Example: PCA in 2D

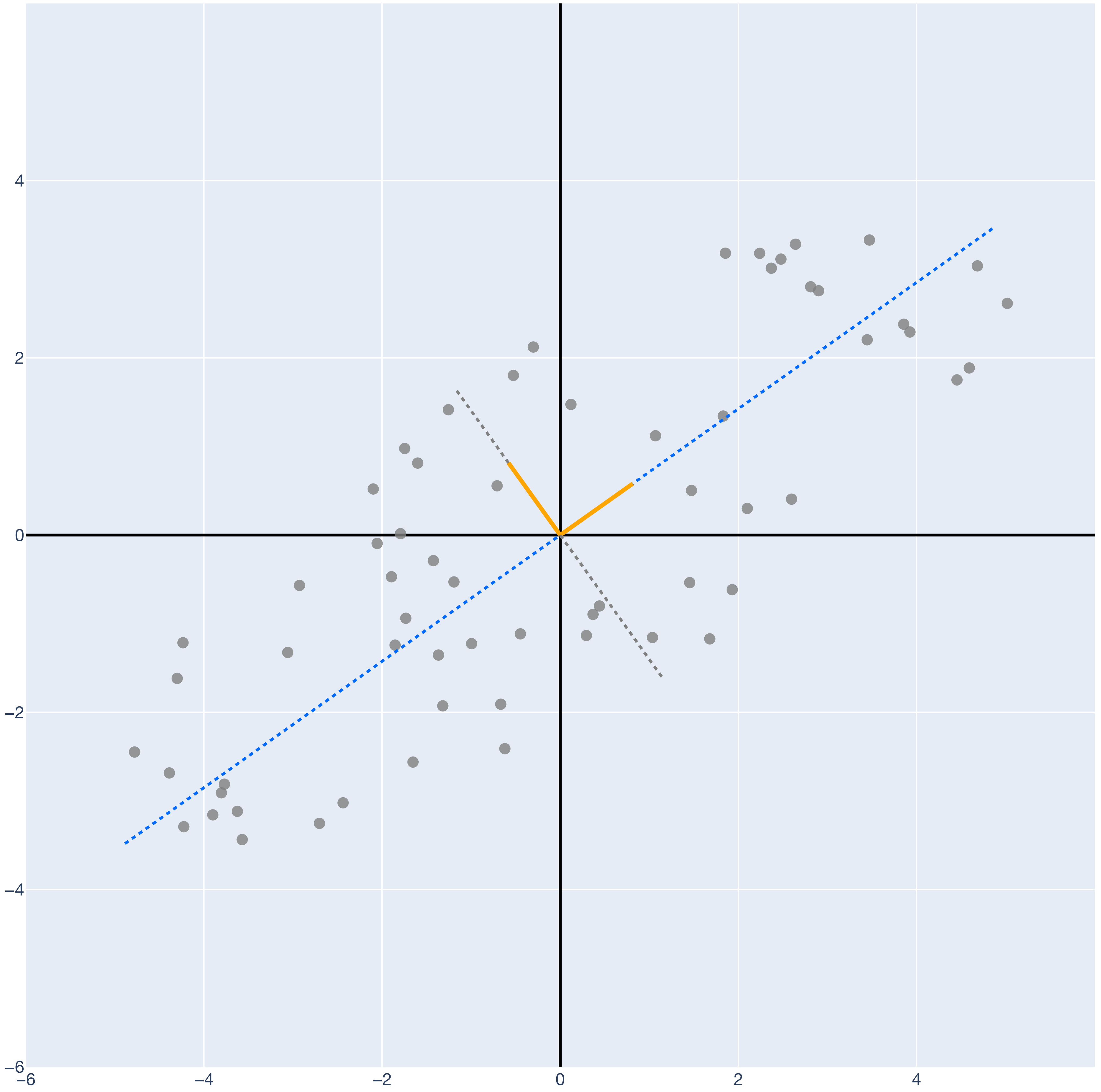
<u>**Observed:**</u> Matrix of *training points* $\mathbf{X} \in \mathbb{R}^{n \times 2}$:

$$\mathbf{X} = \begin{bmatrix} x_{11} & x_{12} \\ x_{21} & x_{22} \\ \vdots & \vdots \\ x_{n1} & x_{n2} \end{bmatrix} = \begin{bmatrix} \uparrow & \uparrow \\ \mathbf{x}_1 & \mathbf{x}_2 \\ \downarrow & \downarrow \end{bmatrix}$$

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. x_k



Derivation of PCA

$C = X^T X$ is symmetric

$\Rightarrow$ spectral thm $\Rightarrow$ Eigendecomposition

Eigendecomposition and PCA

$\boxed{\textit{PCA = Eigendecomposition of the covariance matrix!}}$

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:

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

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

PCA = Eigendecomposition of the covariance matrix!

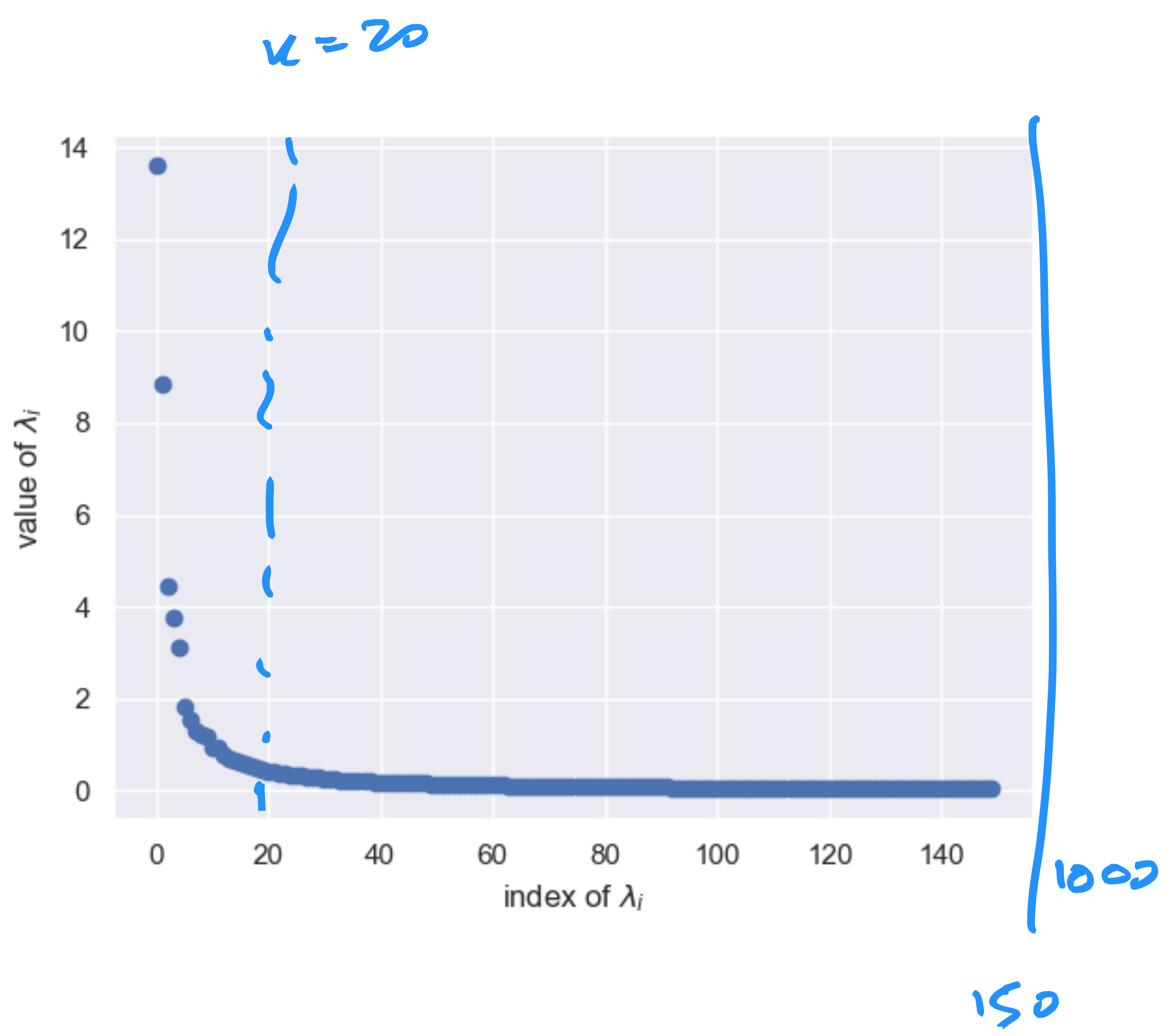
Consider a (column-centered) dataset $\mathbf{X} \in \mathbb{R}^{n \times d}$ and construct its covariance matrix $\mathbf{C} = \mathbf{X}^\top \mathbf{X} \in \mathbb{R}^{d \times d}$. By definition, $\mathbf{C}$ is positive semidefinite.

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

PCA = Eigendecomposition 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:

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

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

Least Squares

Interpretation of Eigenvalues

Regression
Setup

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

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

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

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

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

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

Regression
Setup

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

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

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

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

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

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:

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

$Xw^* = Y.$

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* $\epsilon = (\epsilon_1, \ldots, \epsilon_n) \in \mathbb{R}^n$.

$\varepsilon_i \sim D.$

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

$$\mathbf{y} = \mathbf{X}\mathbf{w}^* + \vec{\epsilon} \in \mathbb{R}^n$$

Error in Regression

Error using least squares model

OLS.

True labels: $\boxed{\mathbf{y} = \mathbf{X}\mathbf{w}^* + \epsilon}$.

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

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

Error in Regression

Error using least squares model

True labels: $\mathbf{y} = \mathbf{X}\mathbf{w}^* + \underbrace{\epsilon}_{=0}$.

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

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

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

True labels: $\mathbf{y} = \mathbf{X}\mathbf{w}^* + \epsilon$.

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

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

Spectral Thrm

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

$$\hat{w} - w^* = (x^T x)^{-1} x^T \varepsilon .$$

Error in Regression

Eigendecomposition perspective

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

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

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

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

λ_i is small $\to$ $1/\lambda_i$ is big.

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

Gradient Descent

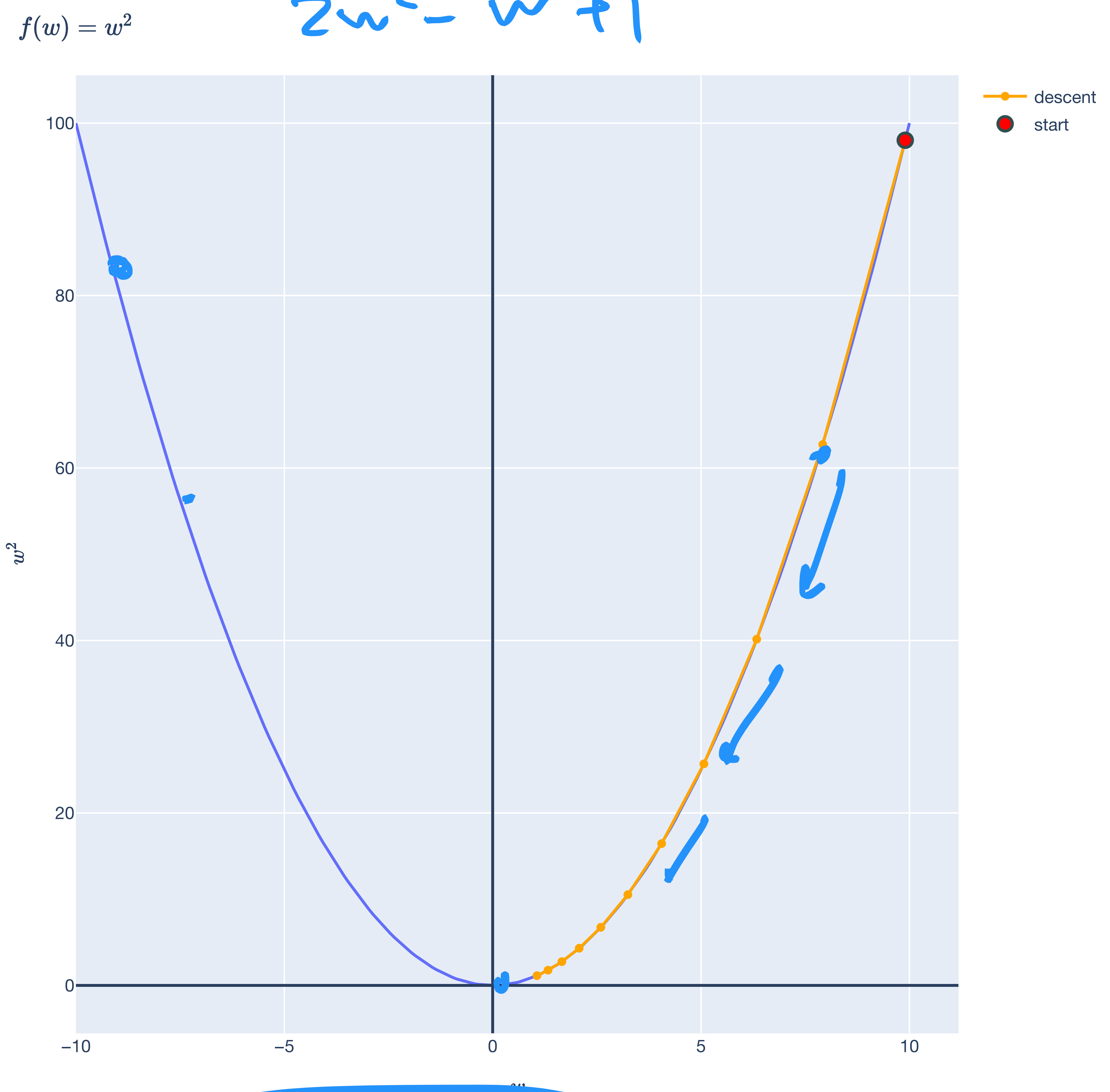
Positive Semidefinite Matrices and Convexity

$\boxed{x^T A x}$

Lesson Overview

Big Picture: Gradient Descent

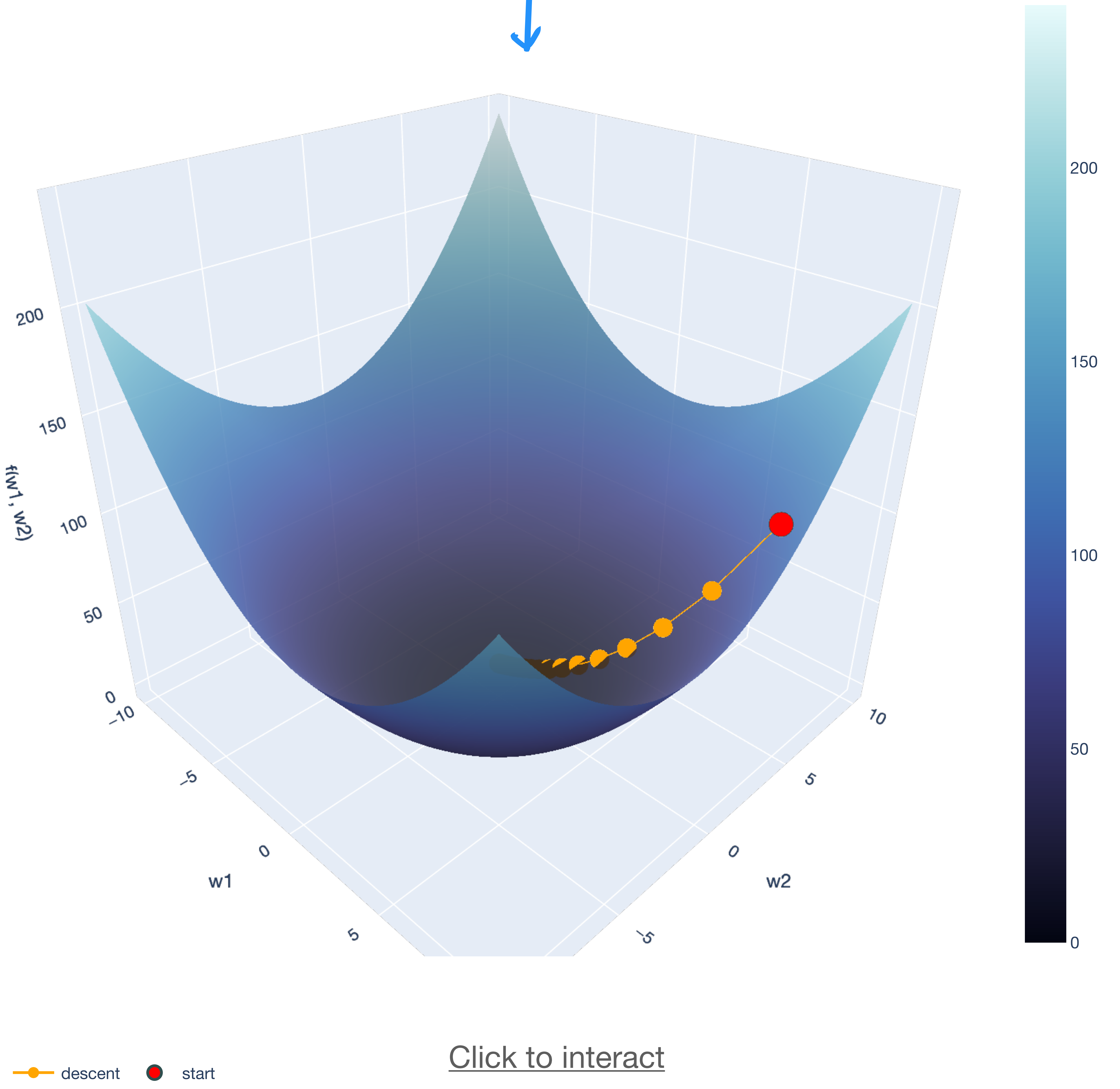
$2w^2 - w + 1$



$\boxed{f(w) = w^2}$

$f(w_1, w_2) = \| Xw - y \|^2$

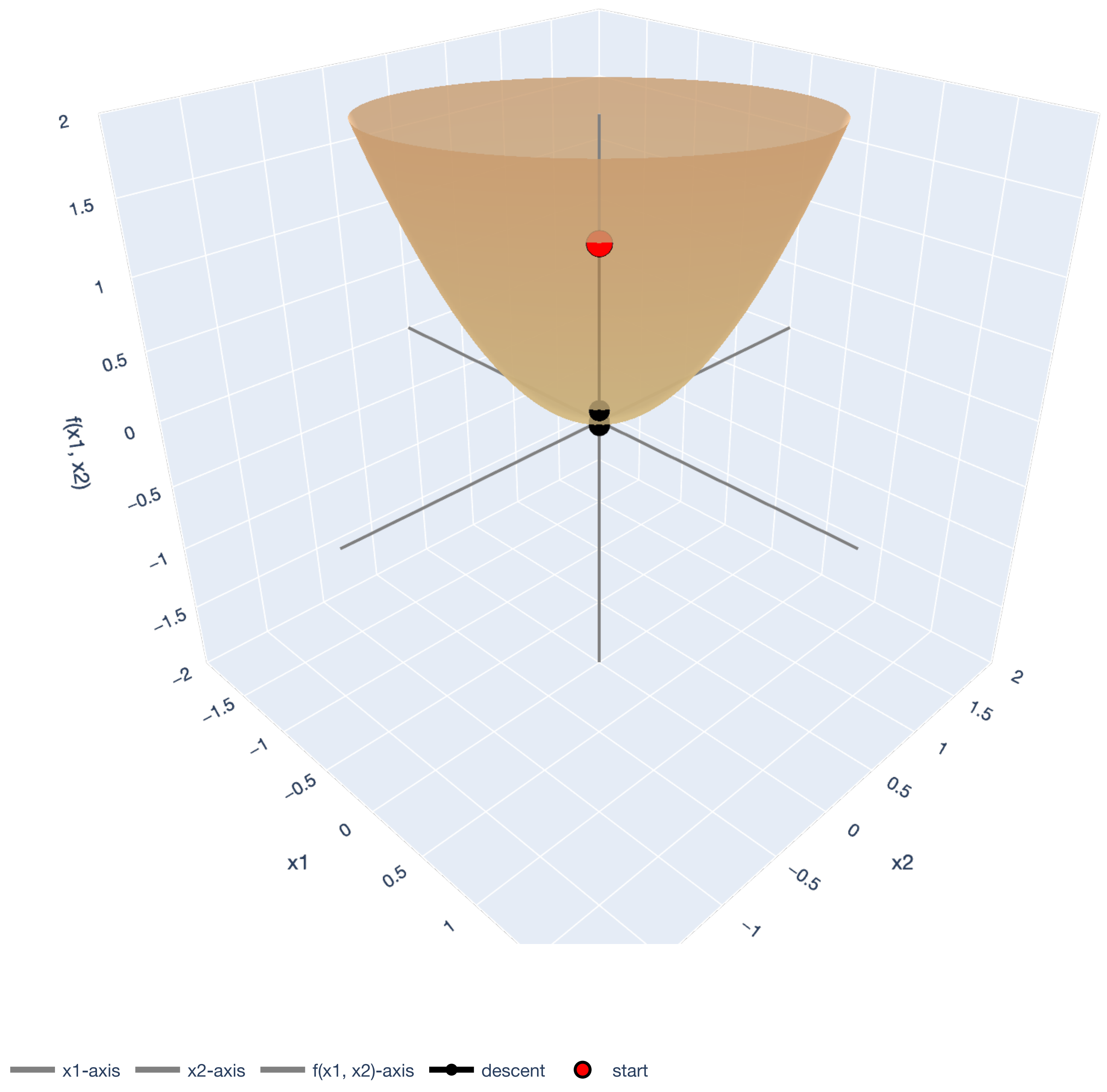
fixed X
fixed y.

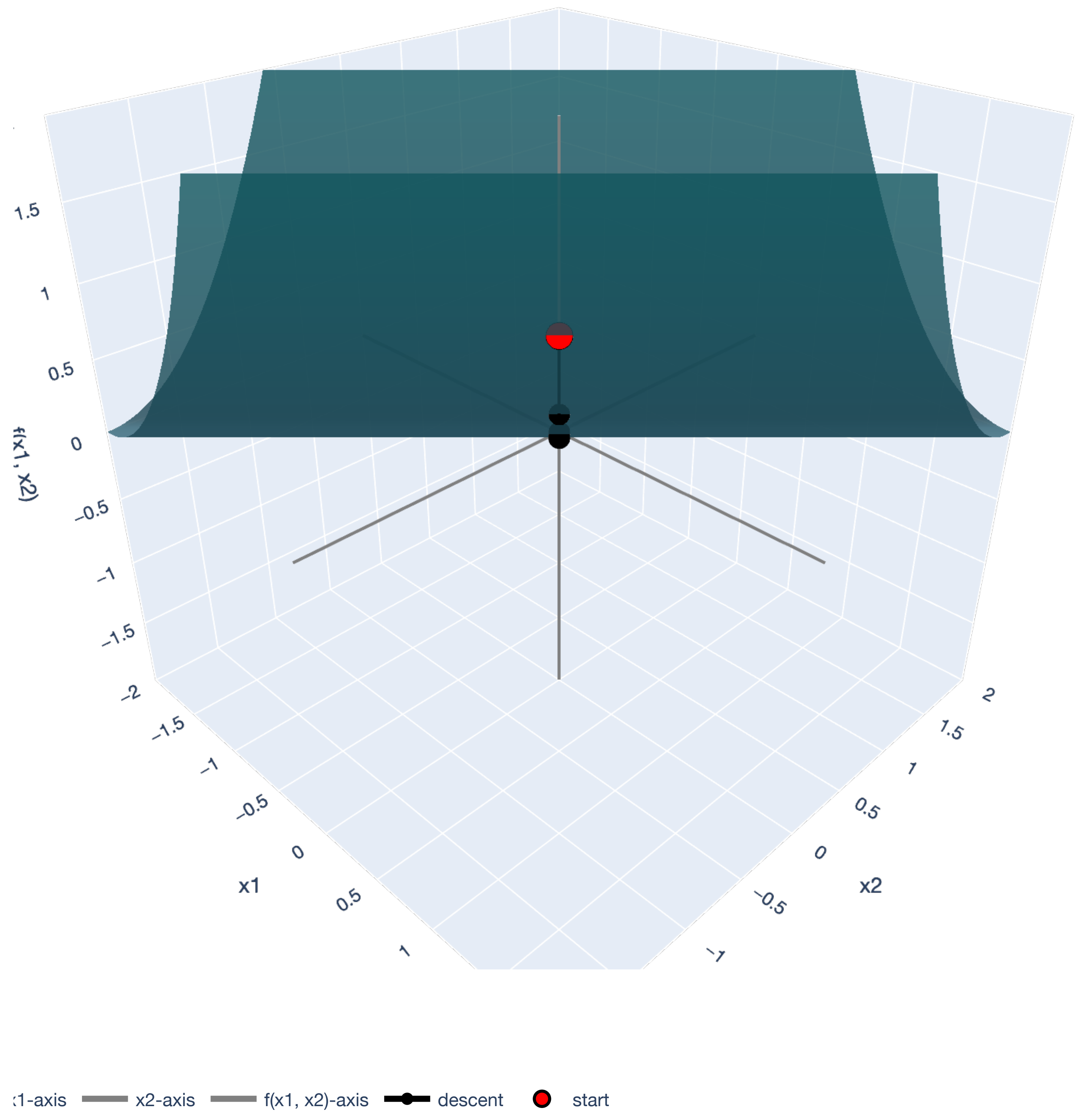


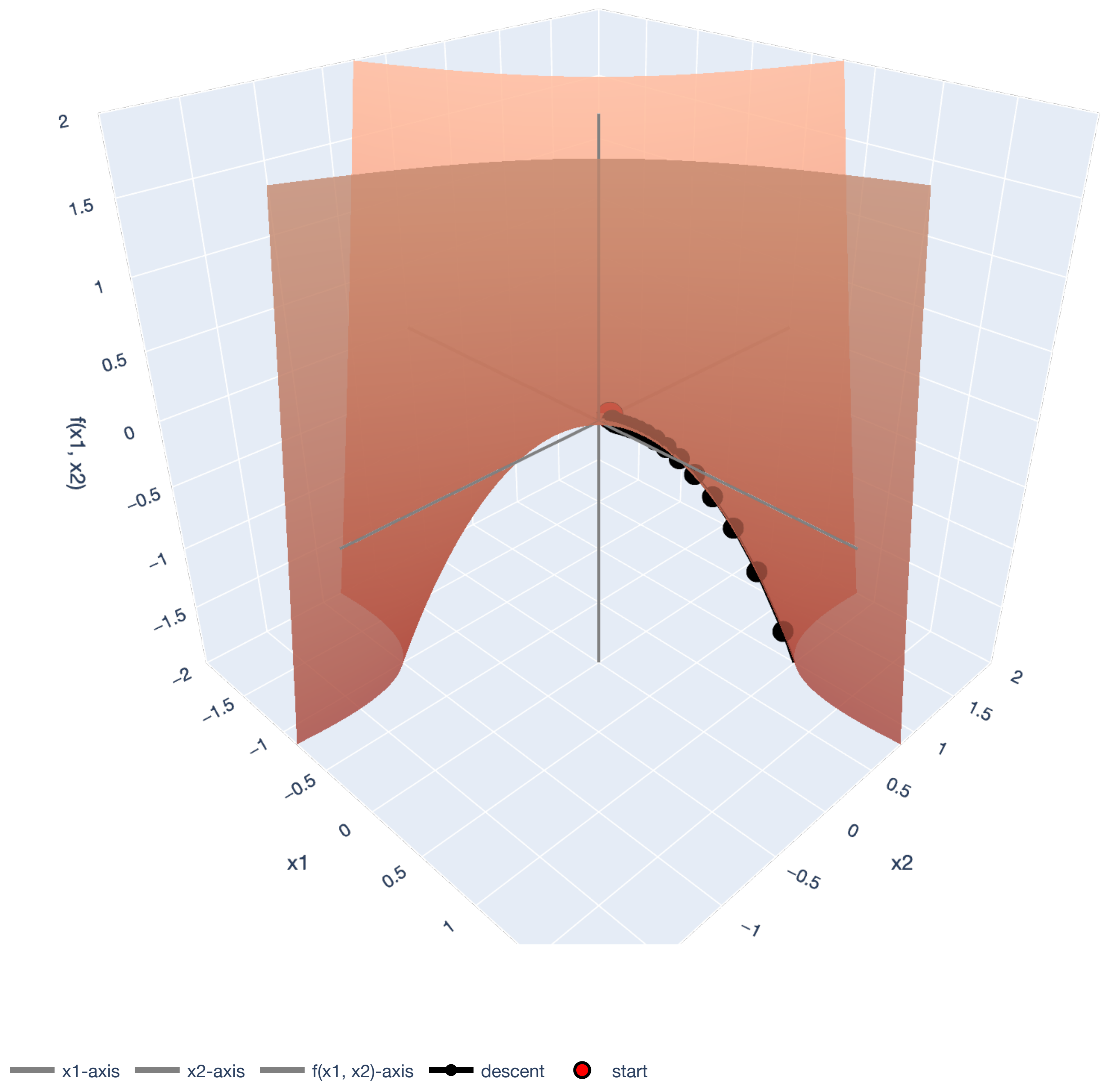
Click to interact

Lesson Overview

Big Picture: Gradient Descent







Quadratic Forms

2D Example

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

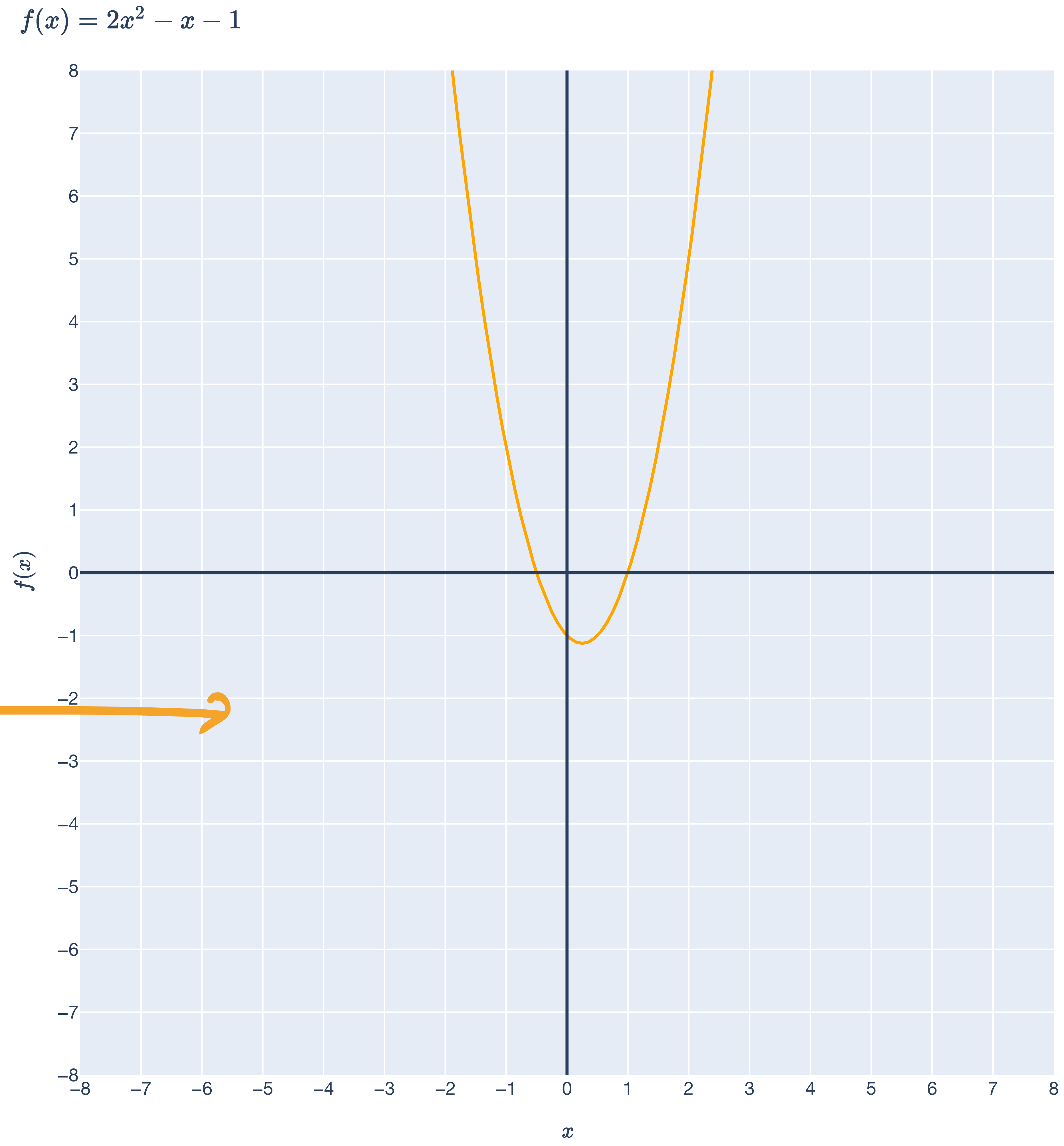
$$\boxed{f(x) = ax^2 + bx + c,}$$

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

$\neq 0$

Example: $f(x) = 2x^2 - x - 1$

$(2x - b')^2 + c'$



Quadratic Forms

2D Example

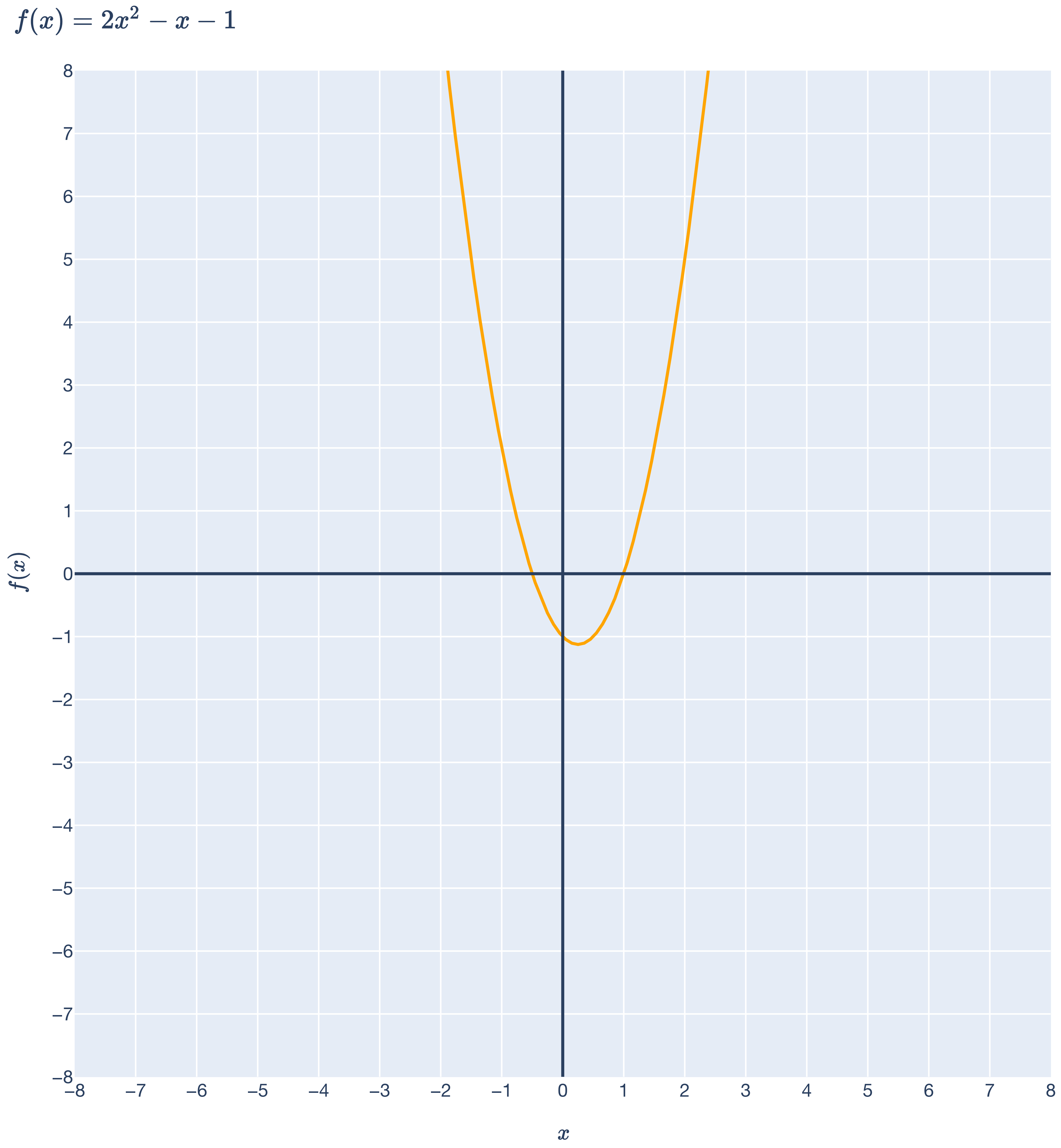
A *quadratic function* $f : \mathbb{R} \to \mathbb{R}$ has the form

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

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

Example: $f(x) = 2x^2 - x - 1$

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



Quadratic Forms

3D Example

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

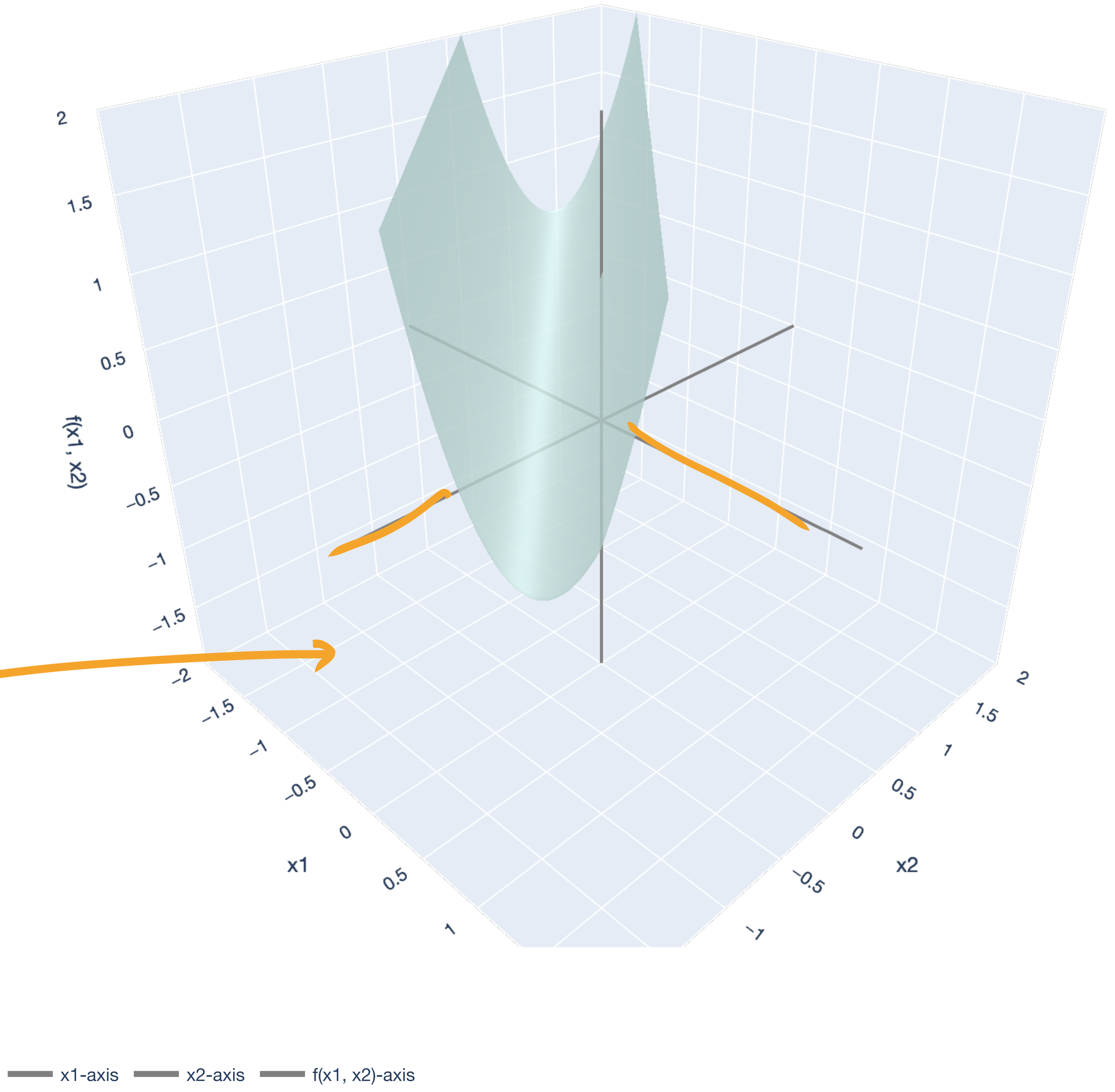
$$f(x) = \underbrace{ax^2 + 2bxy + cy^2 + dx + ey + f}, $$

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

$\neq 0$ $\neq 0$ $\neq 0$

Example:

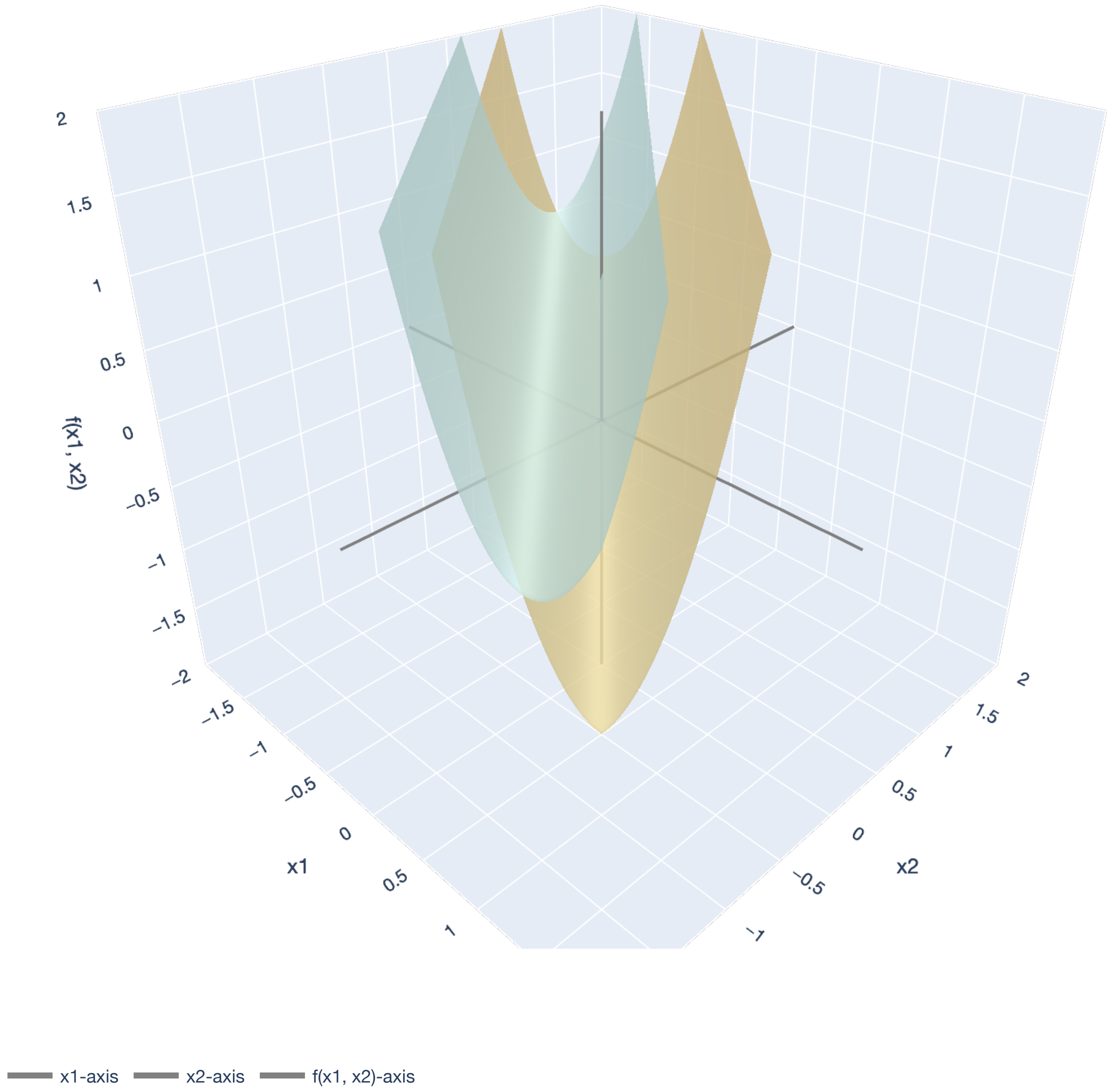
$$f(x) = \underbrace{2x^2 + 4xy + 2y^2 + 2x + 2y + 1}$$



Quadratic Forms

3D Example

$f(x) = \boxed{2x^2 + 4xy + 2y^2} + 2x + 2y + 1$ 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

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

Let's only examine the quadratic part!

$$\underbrace{f(x) = ax^2 + 2bxy + cy^2}_{\text{Quadratic Form}}.$$

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:

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

PSD

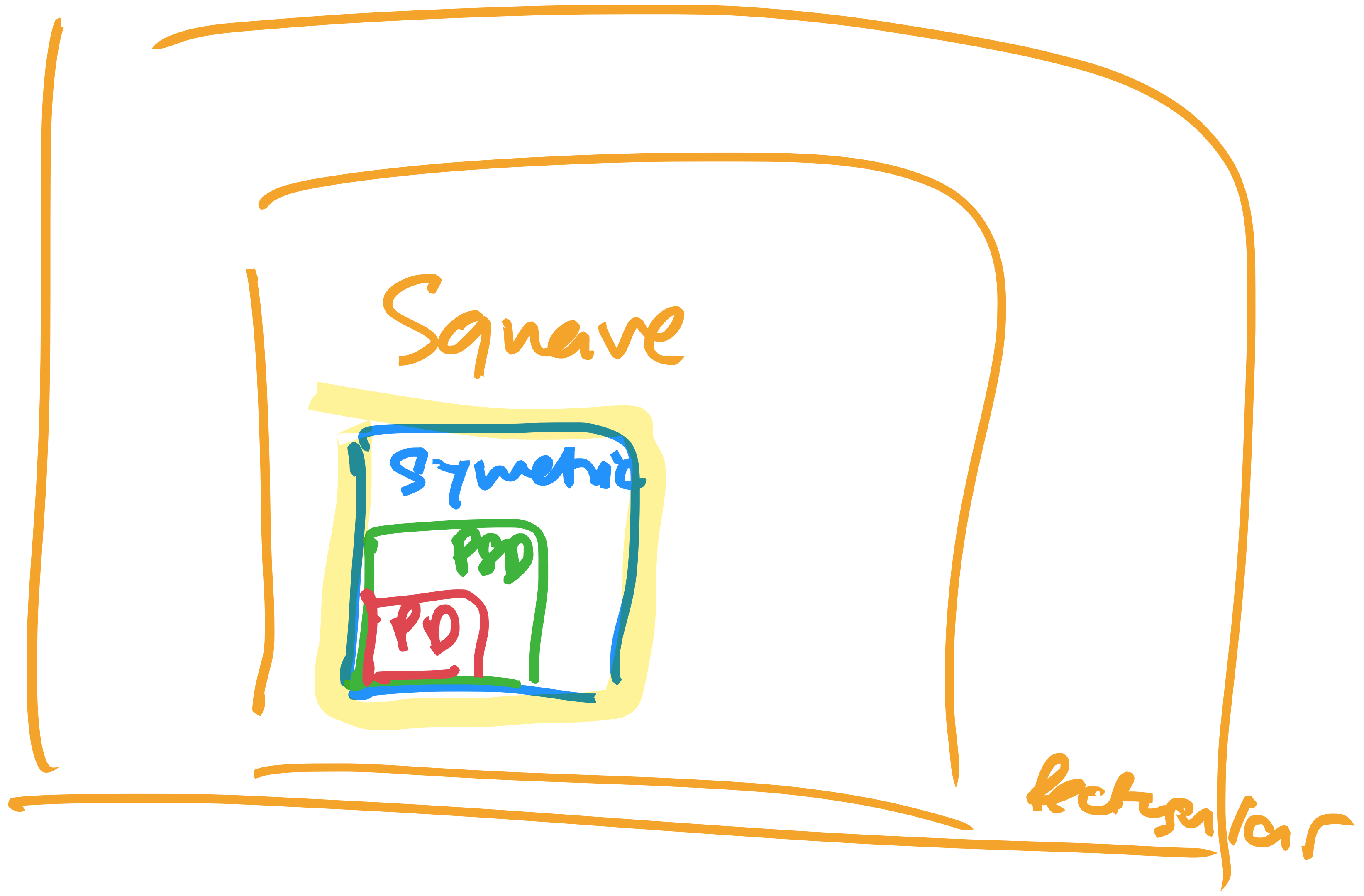
We can rewrite this in matrix form:

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

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

Symmetric $\xrightarrow{\text{spectral Thm}}$ Diagonalizable

Eigenvalues / Eigenvectors



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!

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

Quadratic Forms

Relationship with matrices and eigenvalues

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

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

$$\implies \underline{f(\mathbf{x}) = \mathbf{x}^\top\mathbf{A}\mathbf{x}} = \mathbf{x}^\top\mathbf{Q}\boldsymbol{\Lambda}\mathbf{Q}^\top\mathbf{x}$$

$$\implies \boxed{\bar{\mathbf{x}}^\top\boldsymbol{\Lambda}\bar{\mathbf{x}},} \text{ where } \underbrace{\bar{\mathbf{x}} = \mathbf{Q}^\top\mathbf{x}}.$$

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

$$\boxed{\bar{x}^T \Lambda \bar{x}}$$

$x = M_1 v_1 + M_2 v_2$

$x = \begin{bmatrix} | & | \\ v_1 & v_2 \\ | & | \end{bmatrix} \begin{bmatrix} M_1 \\ M_2 \end{bmatrix}$

$Q^T = Q^{-1}$

$\bar{x} = Q^T x$

Quadratic Forms

Relationship with matrices and eigenvalues

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

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

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*). $\lambda_1, \lambda_2 \geq 0.$
3. λ_1 or λ_2 is negative (*indefinite*).

Quadratic Forms

Example: positive definite

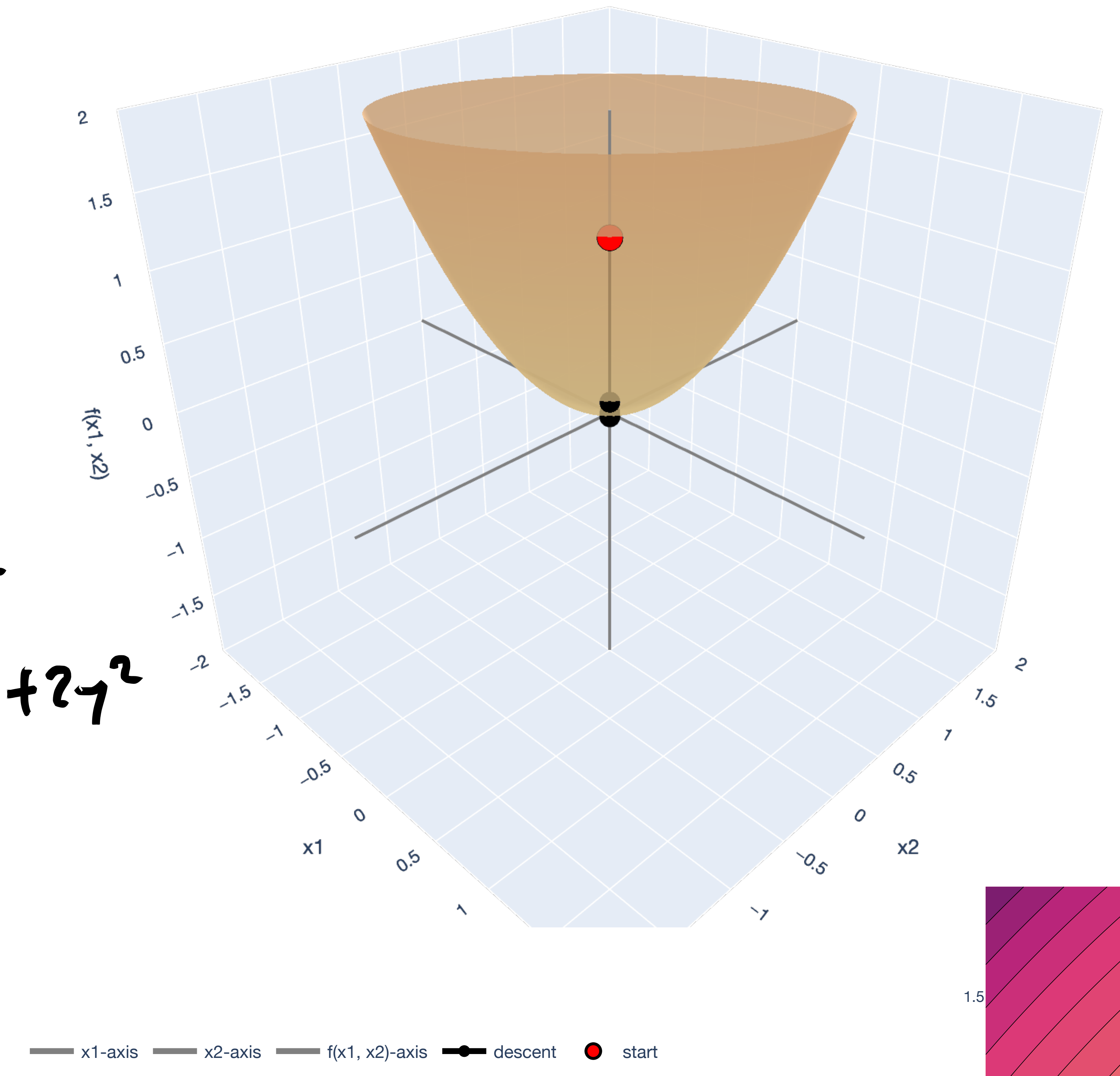
$$\begin{bmatrix} x & y \end{bmatrix} \begin{bmatrix} 2 & -1 \\ -1 & 2 \end{bmatrix} \begin{bmatrix} x \\ y \end{bmatrix} = \begin{bmatrix} x & y \end{bmatrix} \begin{bmatrix} 2x - y \\ -x + 2y \end{bmatrix}$$

$$= 2x^2 - xy - xy + 2y^2$$

$$= 2x^2 - 2xy + 2y^2$$

Example:

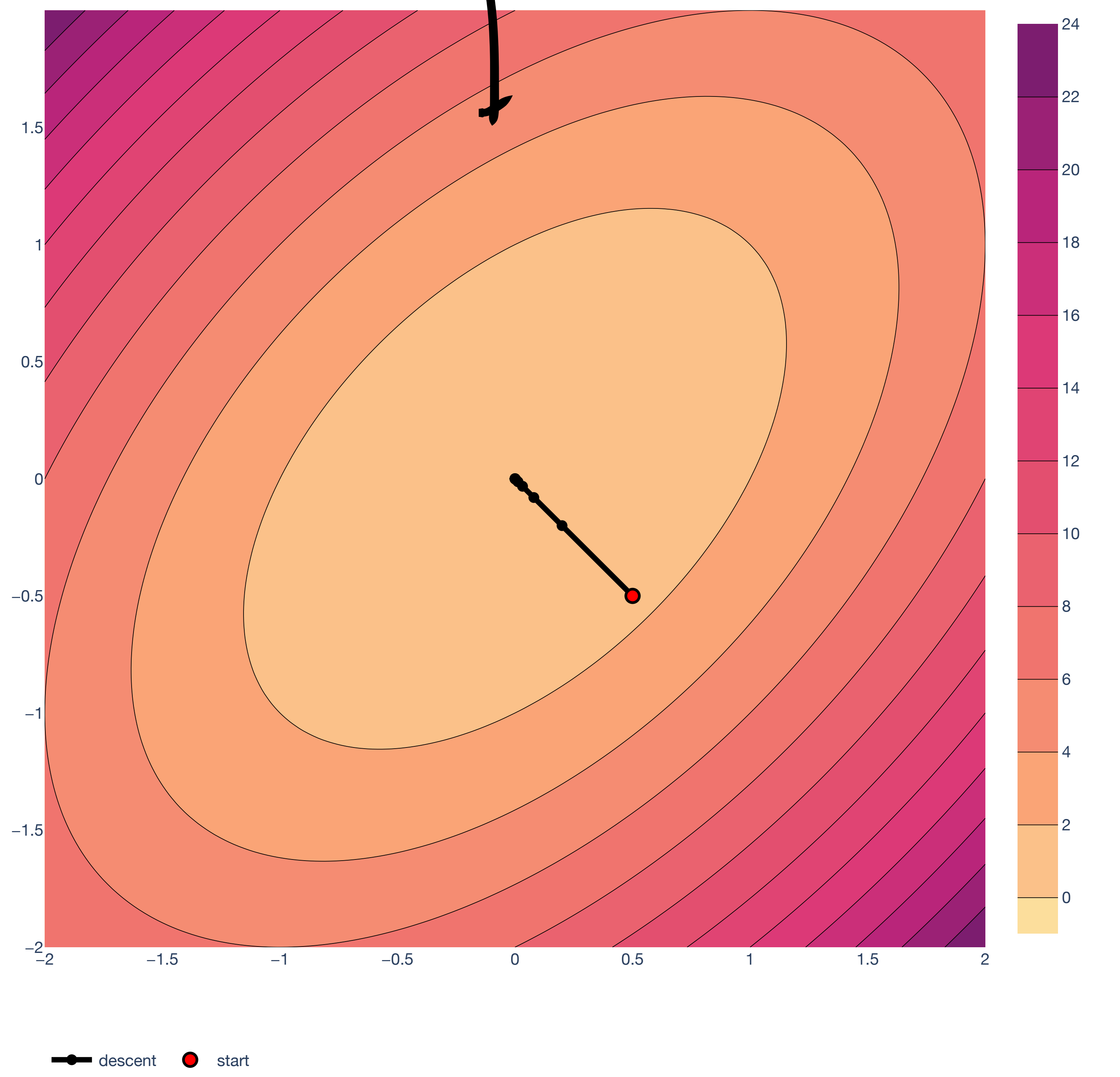
$$f(x, y) = \begin{bmatrix} x & y \end{bmatrix} \begin{bmatrix} 2 & -1 \\ -1 & 2 \end{bmatrix} \begin{bmatrix} x \\ y \end{bmatrix}$$



Eigendecomposition:

$$\begin{bmatrix} 2 & -1 \\ -1 & 2 \end{bmatrix} = \underbrace{\begin{bmatrix} -1/\sqrt{2} & 1/\sqrt{2} \\ 1/\sqrt{2} & 1/\sqrt{2} \end{bmatrix}}_{Q} \underbrace{\begin{bmatrix} 3 & 0 \\ 0 & 1 \end{bmatrix}}_{\lambda} \underbrace{\begin{bmatrix} -1/\sqrt{2} & 1/\sqrt{2} \\ 1/\sqrt{2} & 1/\sqrt{2} \end{bmatrix}}_{Q^T}$$

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



Quadratic Forms

Example: positive semidefinite

Example:

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

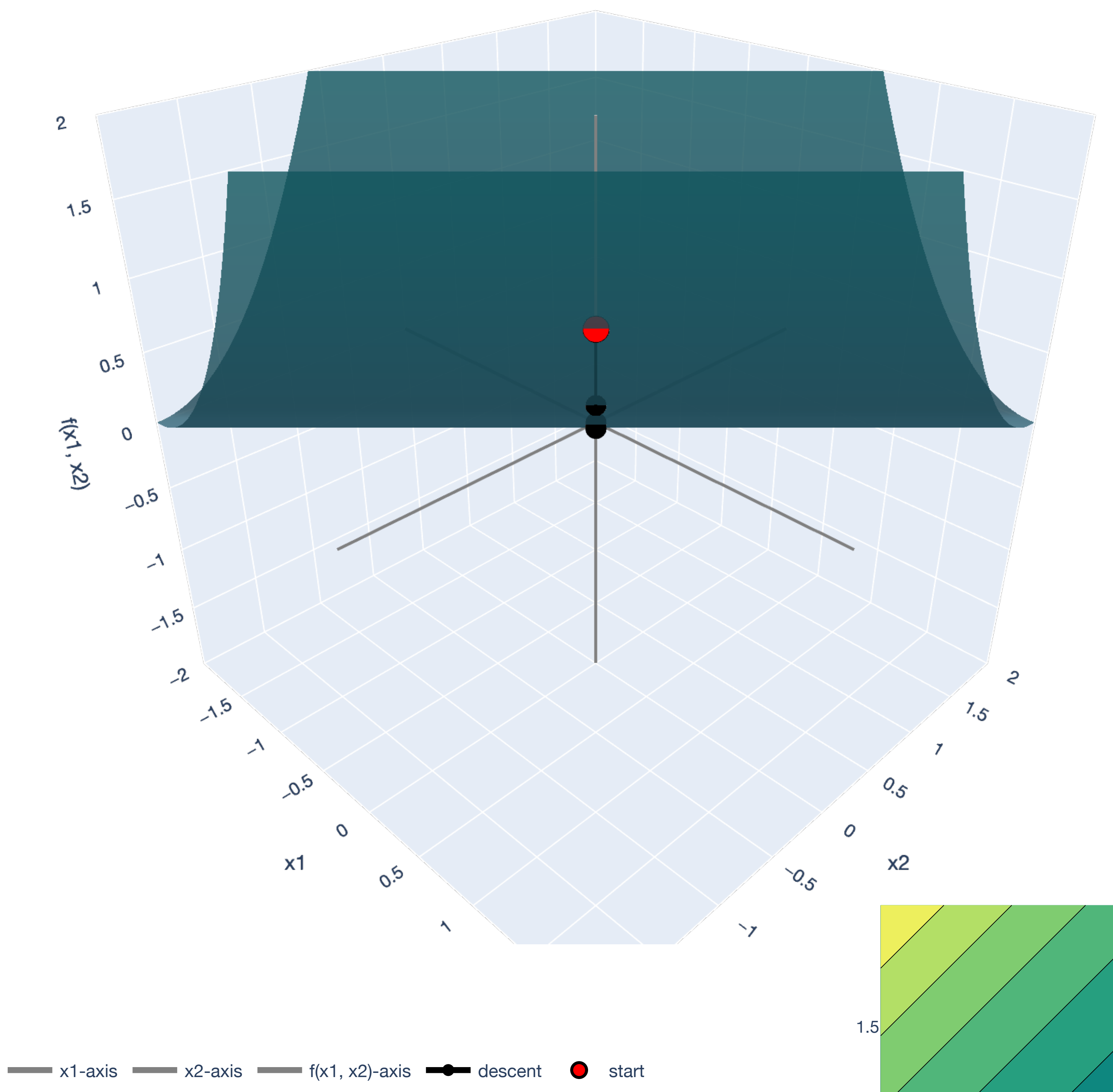
Eigendecomposition:

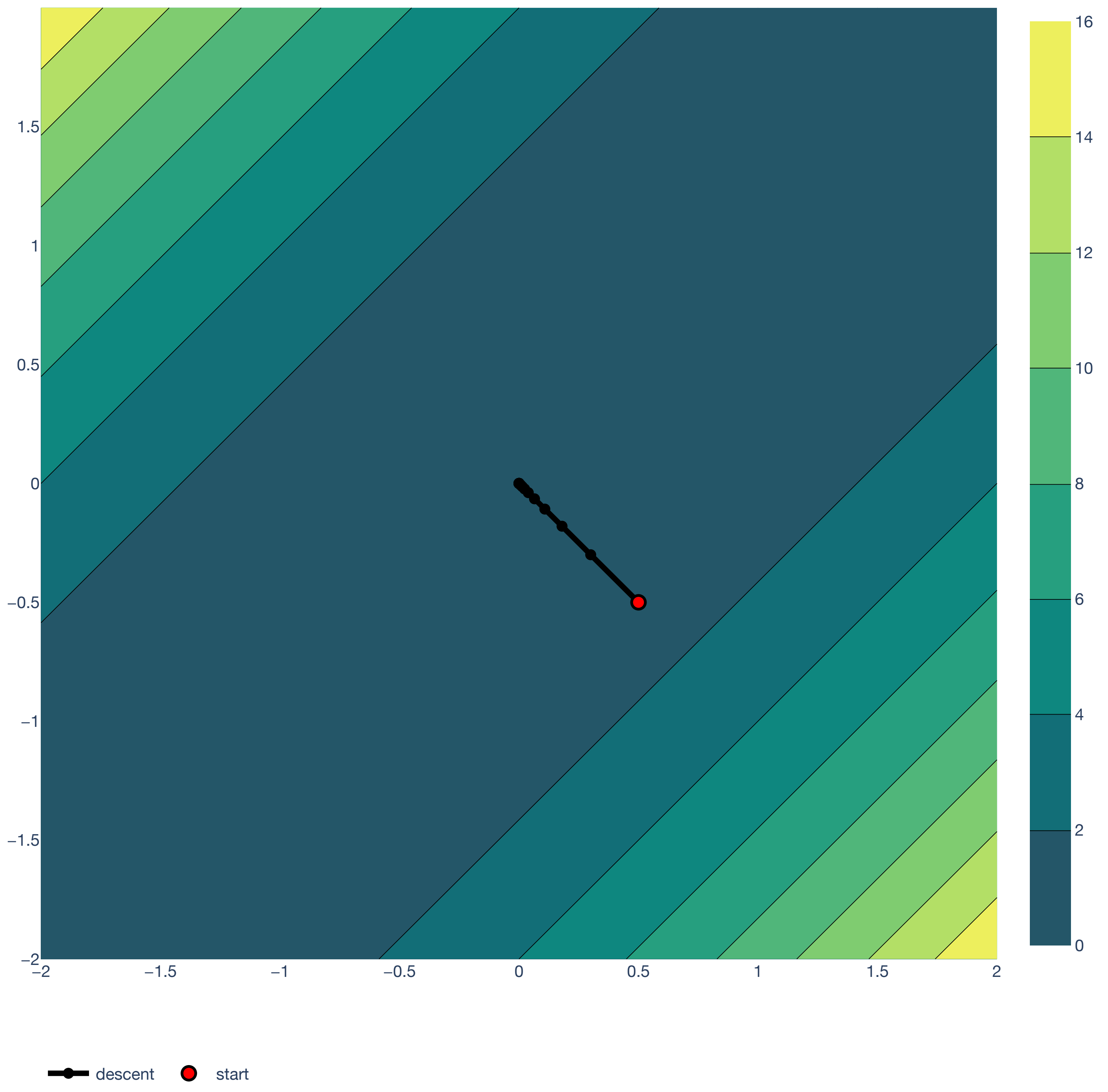
$$\begin{bmatrix} 1 & -1 \\ -1 & 1 \end{bmatrix} = \begin{bmatrix} -1/\sqrt{2} & 1/\sqrt{2} \\ 1/\sqrt{2} & 1/\sqrt{2} \end{bmatrix} \begin{bmatrix} 2 & 0 \\ 0 & 0 \end{bmatrix} \begin{bmatrix} -1/\sqrt{2} & 1/\sqrt{2} \\ 1/\sqrt{2} & 1/\sqrt{2} \end{bmatrix}$$

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

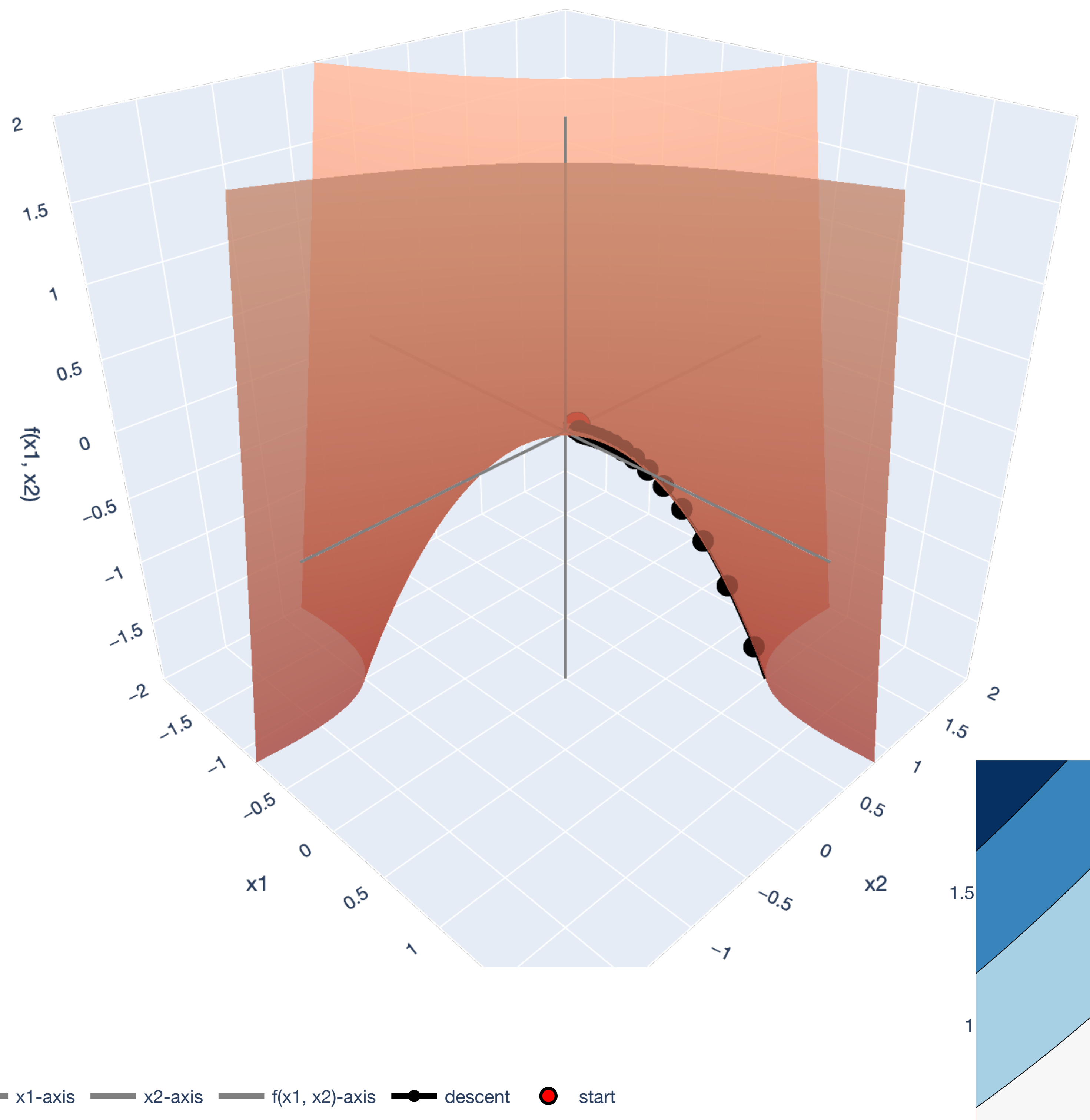
$\lambda_1 = 2$

$\lambda_2 = 0$





Quadratic Forms
Example: indefinite



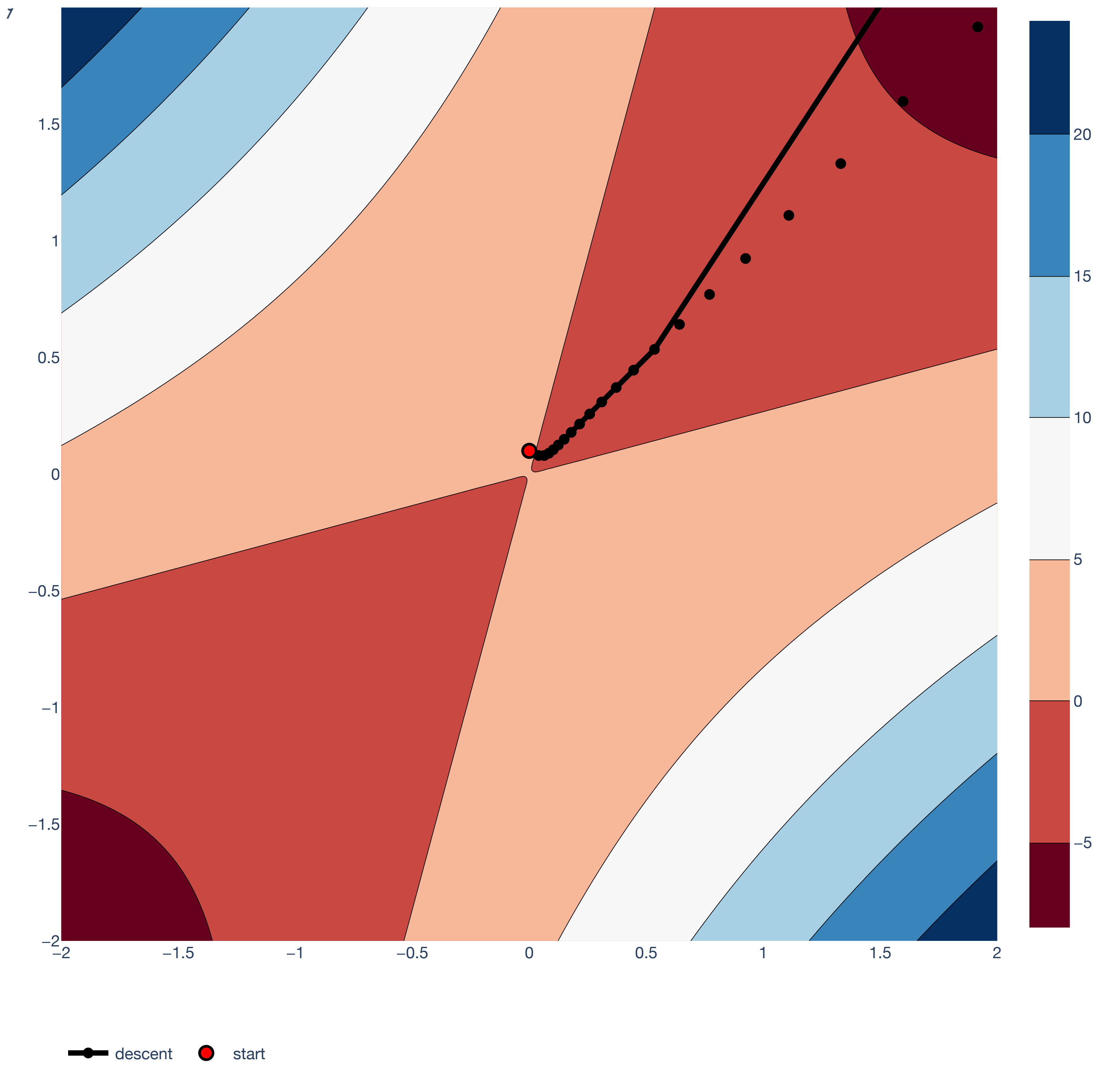
<u>Example:</u>

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

Eigendecomposition:

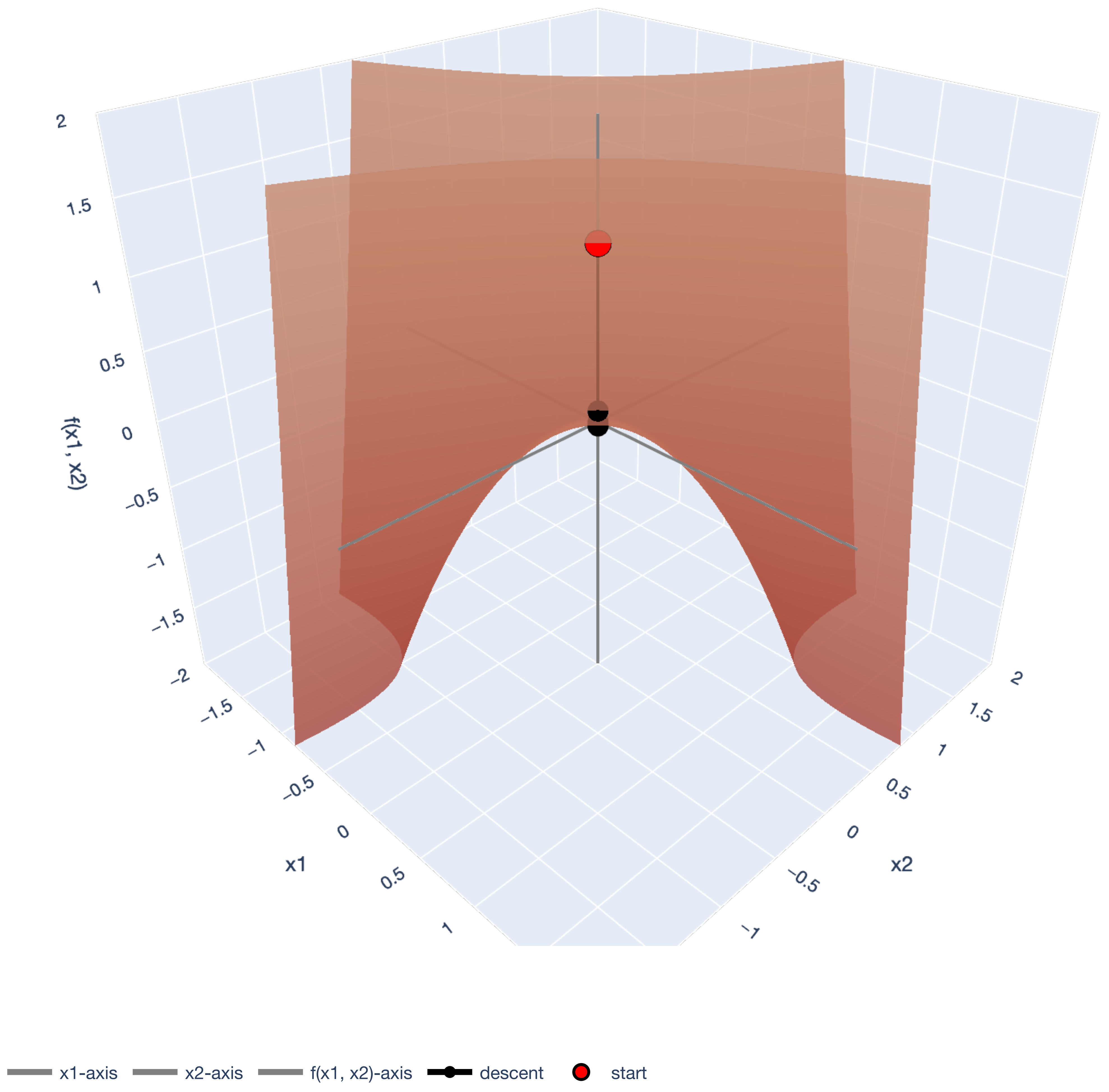
$$\begin{bmatrix} 1 & -2 \\ -2 & 1 \end{bmatrix} = \begin{bmatrix} -1/\sqrt{2} & 1/\sqrt{2} \\ 1/\sqrt{2} & 1/\sqrt{2} \end{bmatrix} \begin{bmatrix} 3 & 0 \\ 0 & -1 \end{bmatrix} \begin{bmatrix} -1/\sqrt{2} & 1/\sqrt{2} \\ 1/\sqrt{2} & 1/\sqrt{2} \end{bmatrix}$$

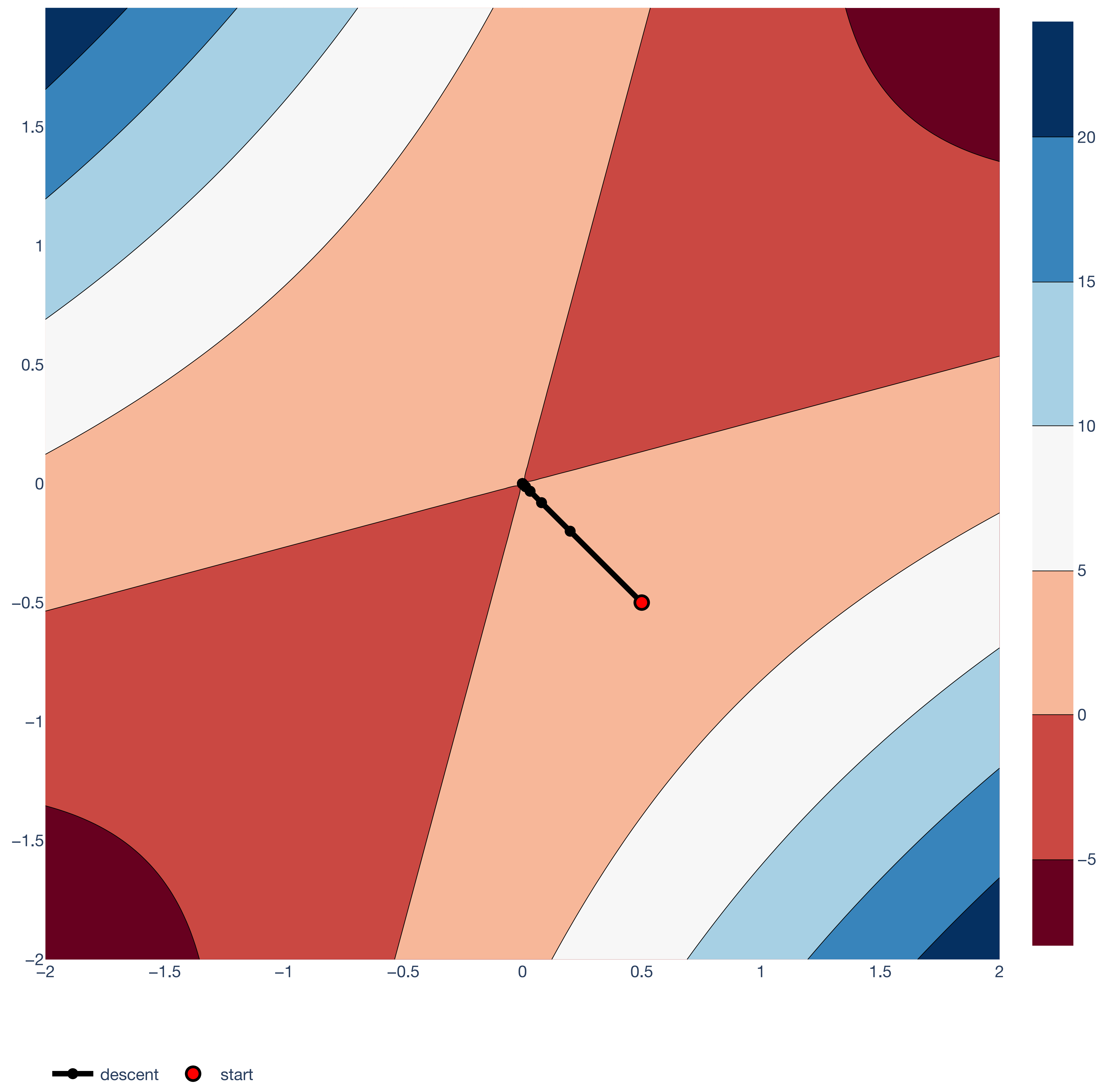
so $\boldsymbol{\Lambda} = \begin{bmatrix} 3 & 0 \\ 0 & -1 \end{bmatrix}$.



Quadratic Forms

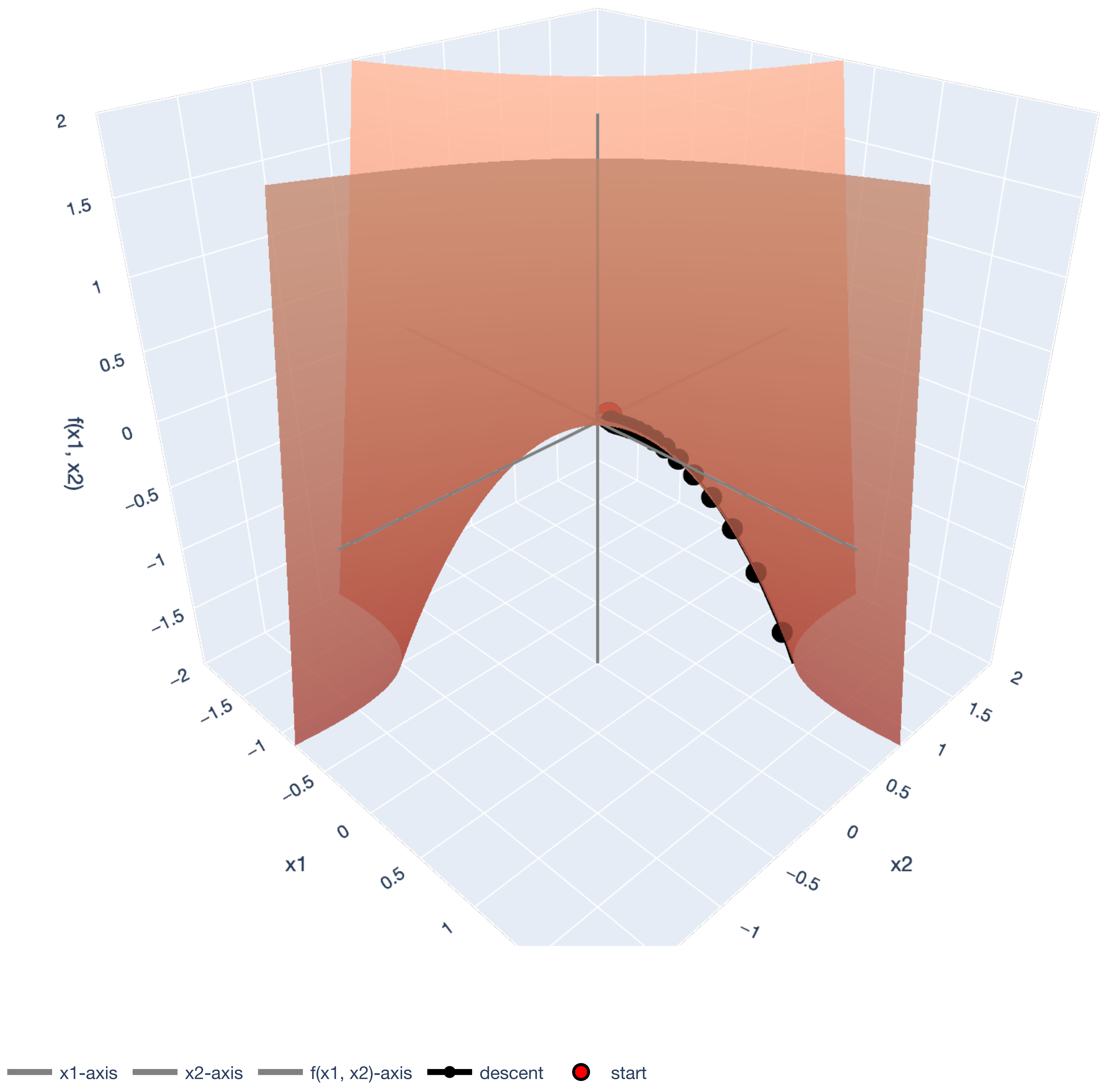
Example: indefinite

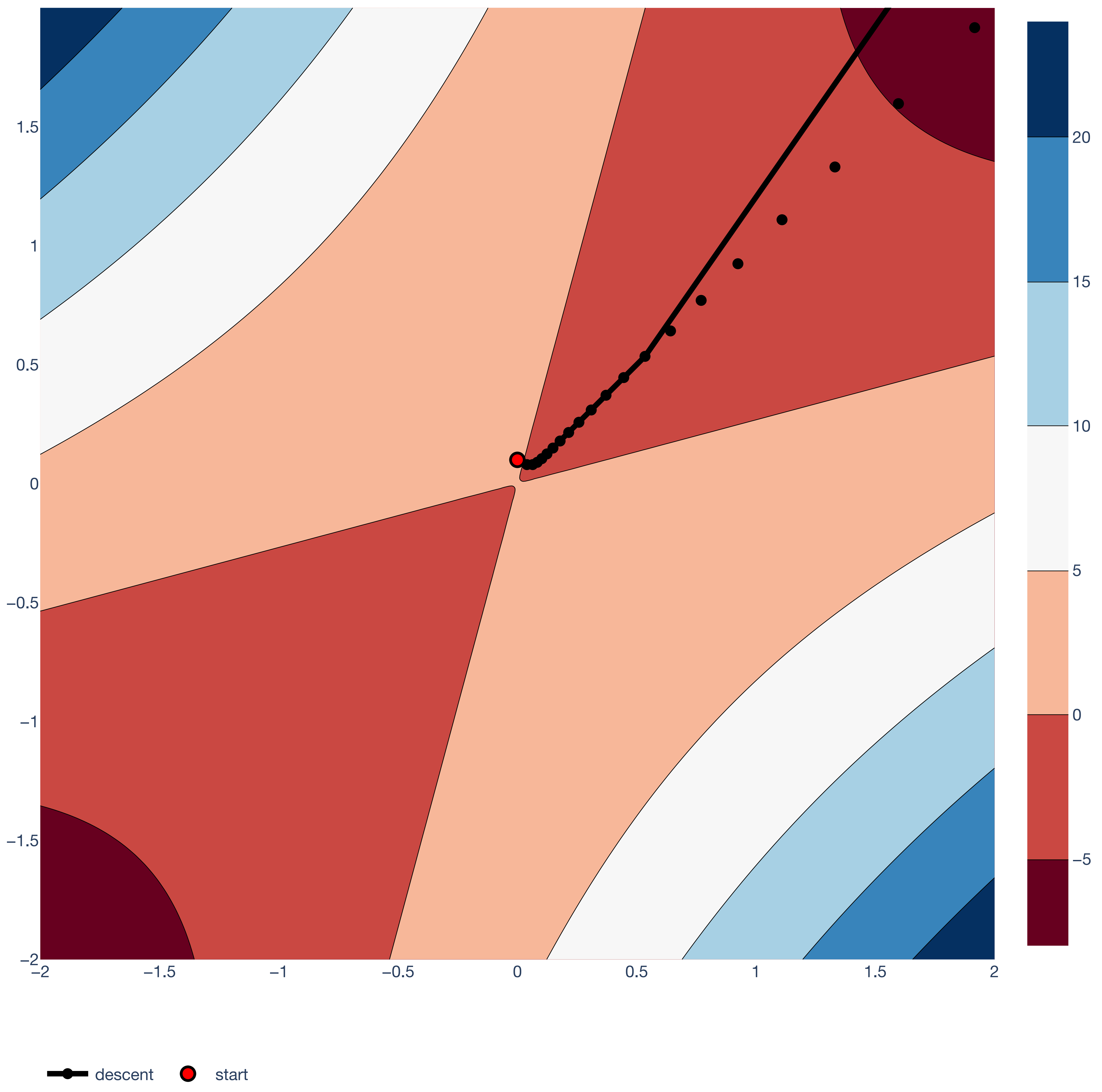




Quadratic Forms

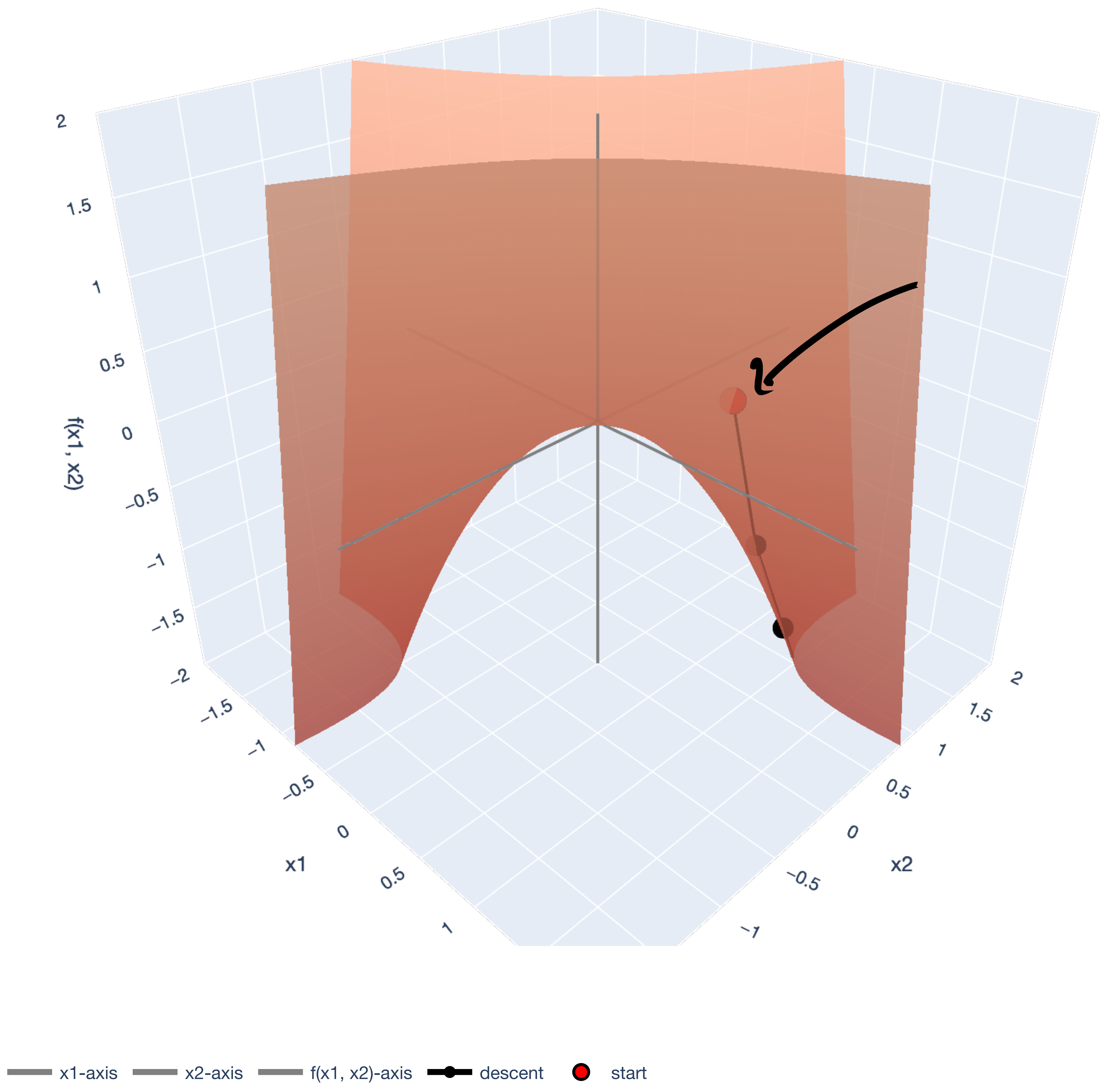
Example: indefinite

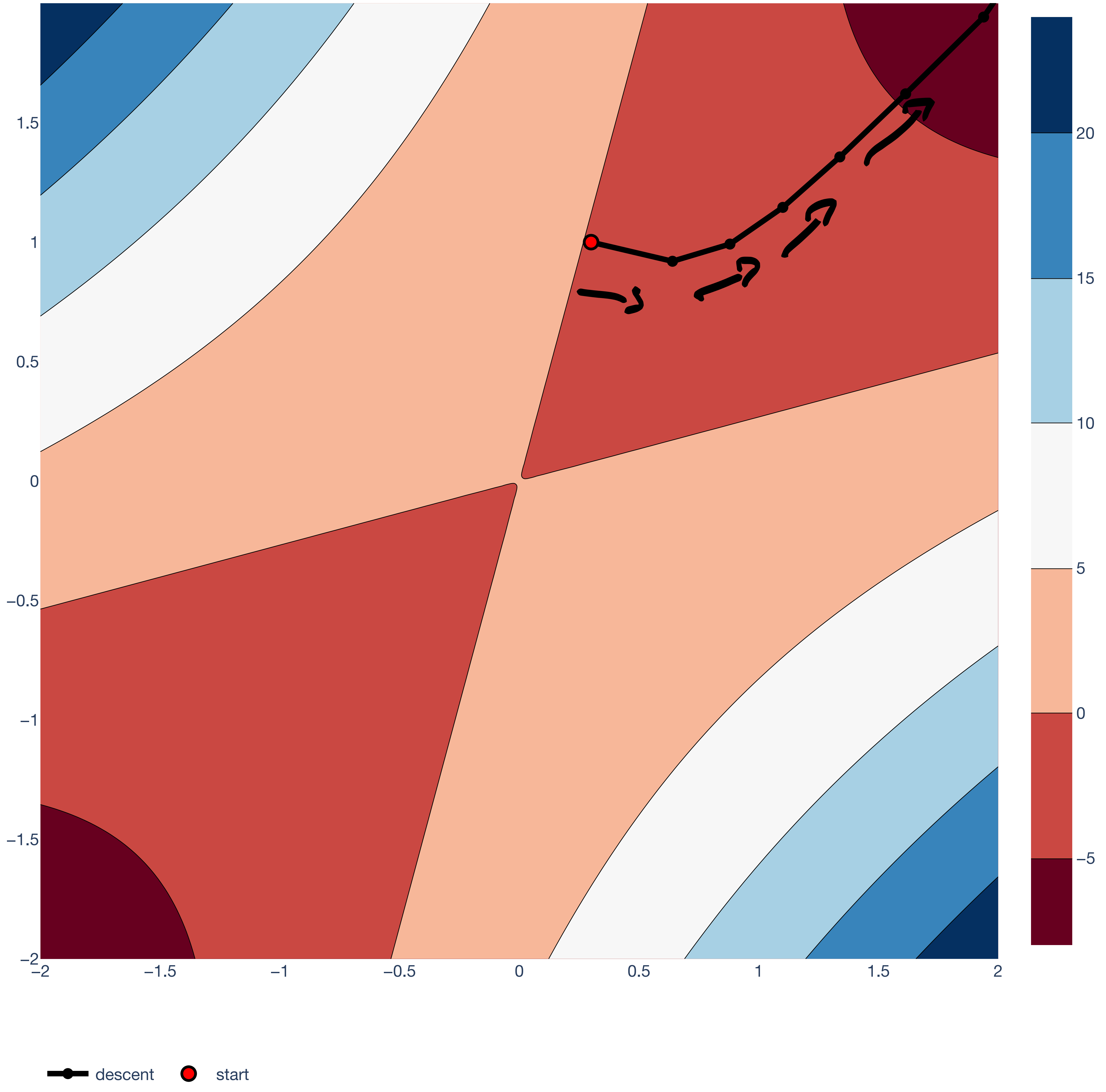




Quadratic Forms

Example: indefinite





Least Squares

Example of quadratic form

Consider the familiar function we've been thinking about:

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

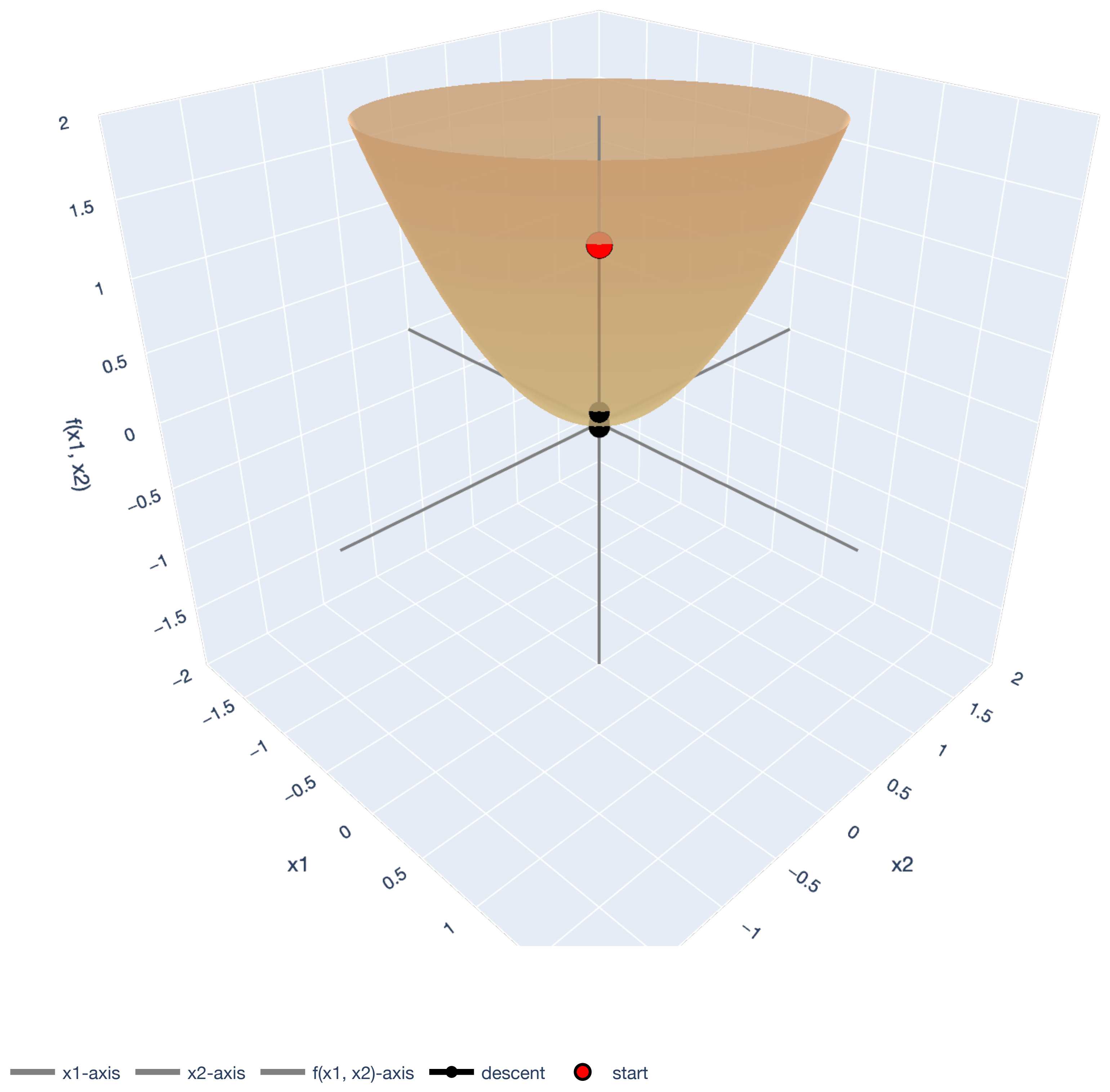
$$(\mathbf{X}\mathbf{w} - \mathbf{y})^\top(\mathbf{X}\mathbf{w} - \mathbf{y}) = \mathbf{w}^\top(\mathbf{X}^\top\mathbf{X})\mathbf{w} - 2\mathbf{w}^\top(\mathbf{X}^\top\mathbf{y}) + \mathbf{y}^\top\mathbf{y}.$$

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

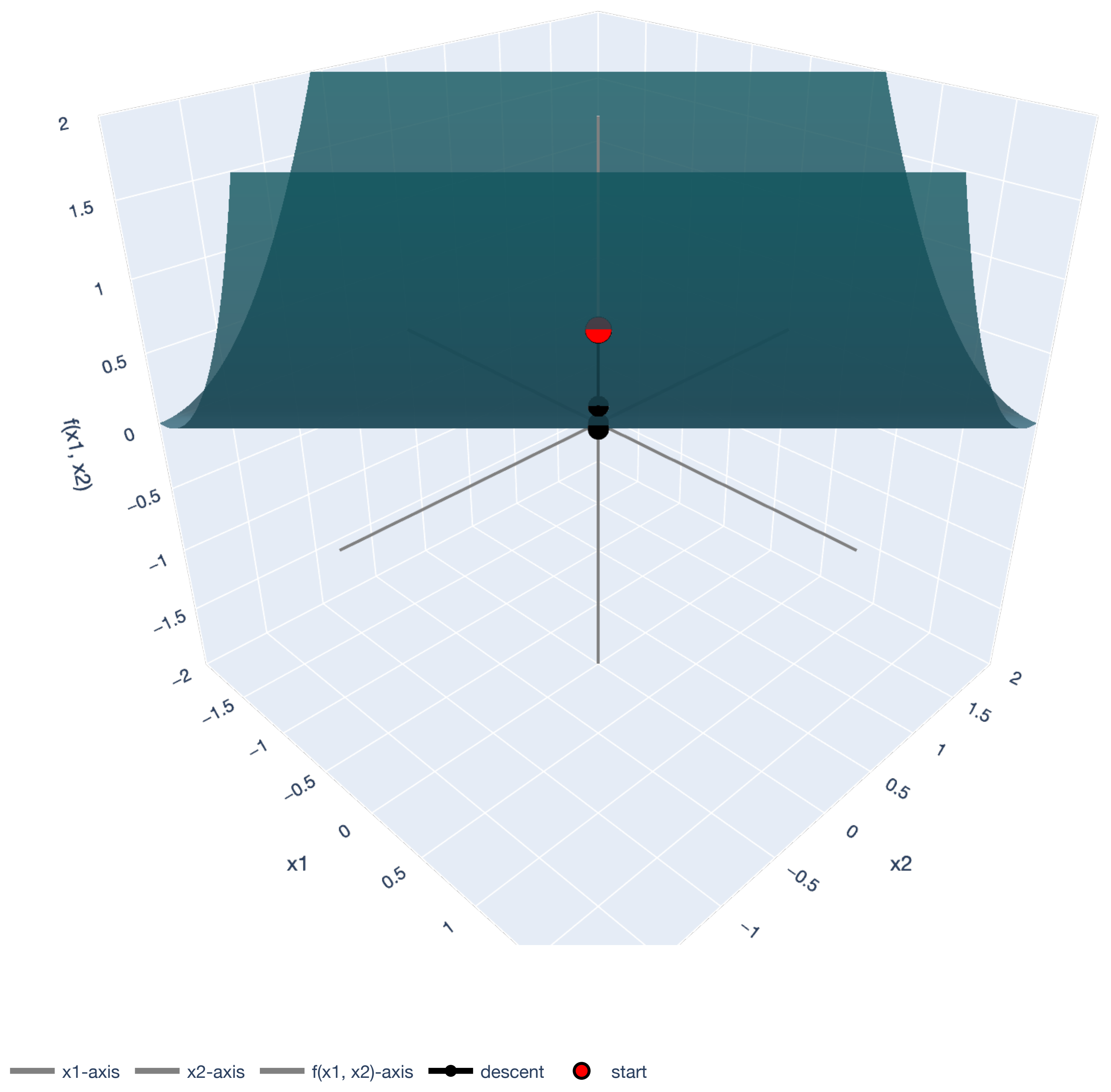
$A = X^T X$ is PSD.

Gradient Descent

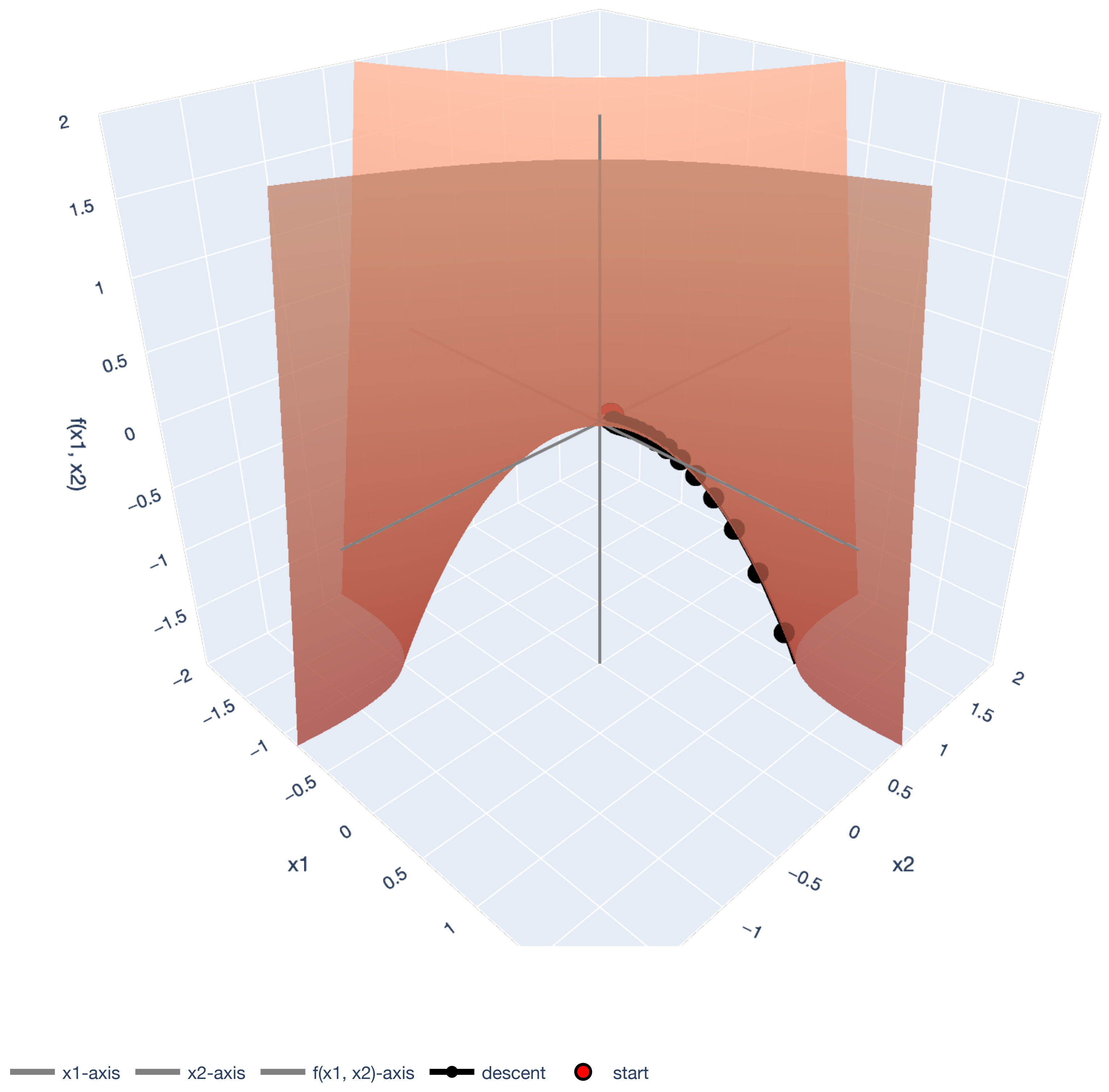
Preview



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



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



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

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.

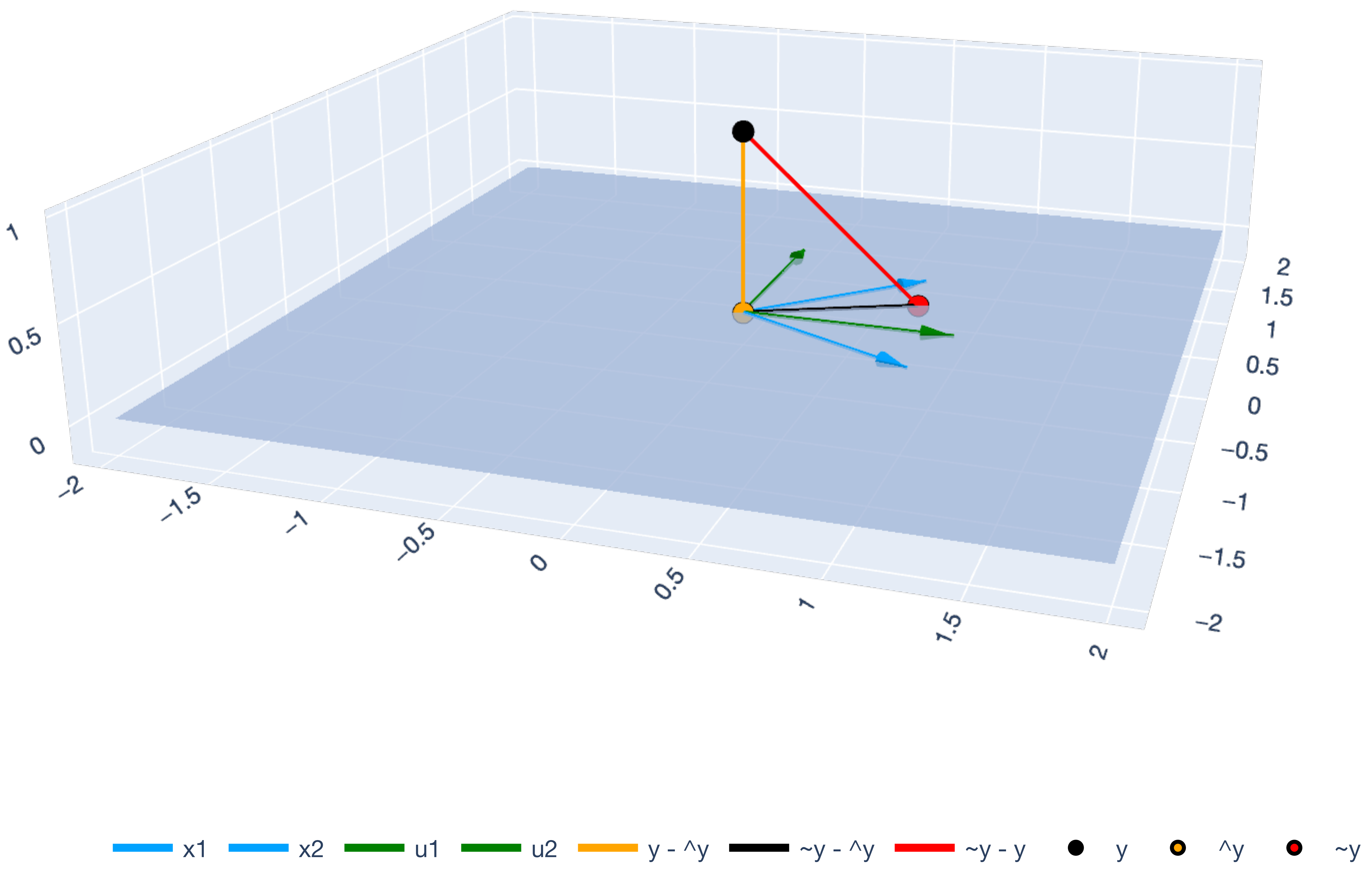
Spectral Theorem. Symmetric matrices are always diagonalizable.

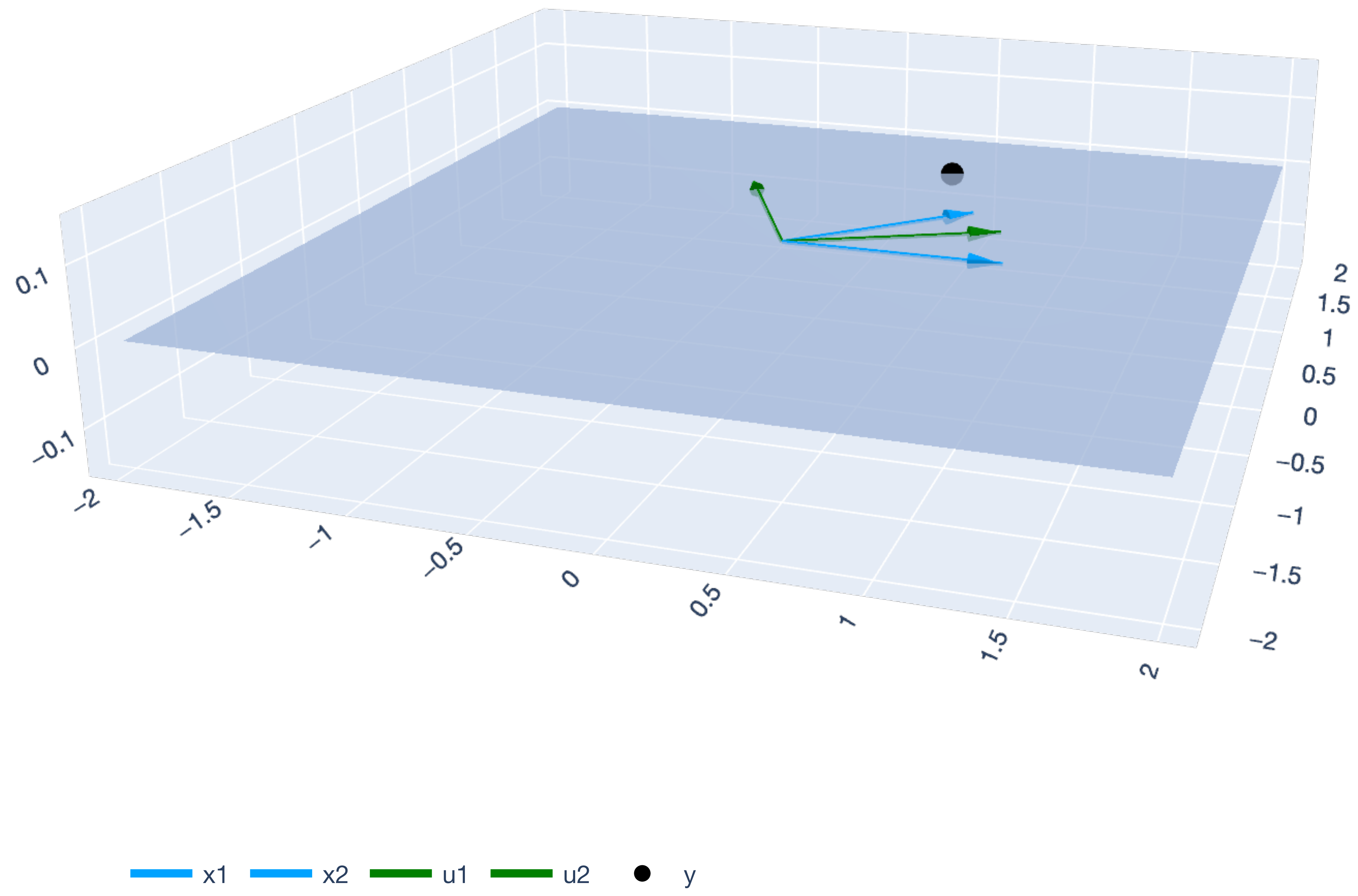
Positive semidefinite matrices/positive definite matrices. Definition and some visual examples through the corresponding quadratic forms.

Lesson Overview

Big Picture: Least Squares

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

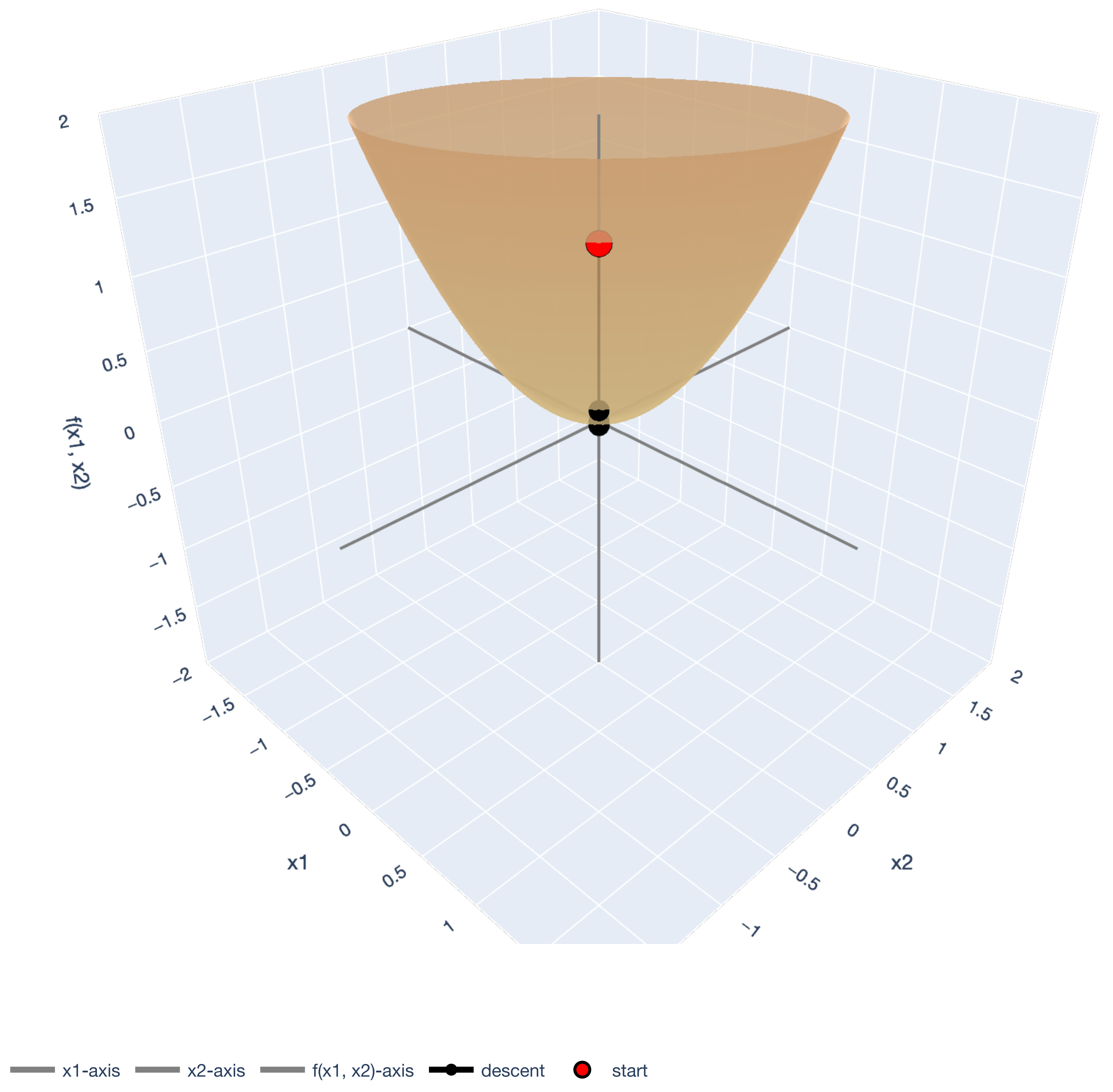


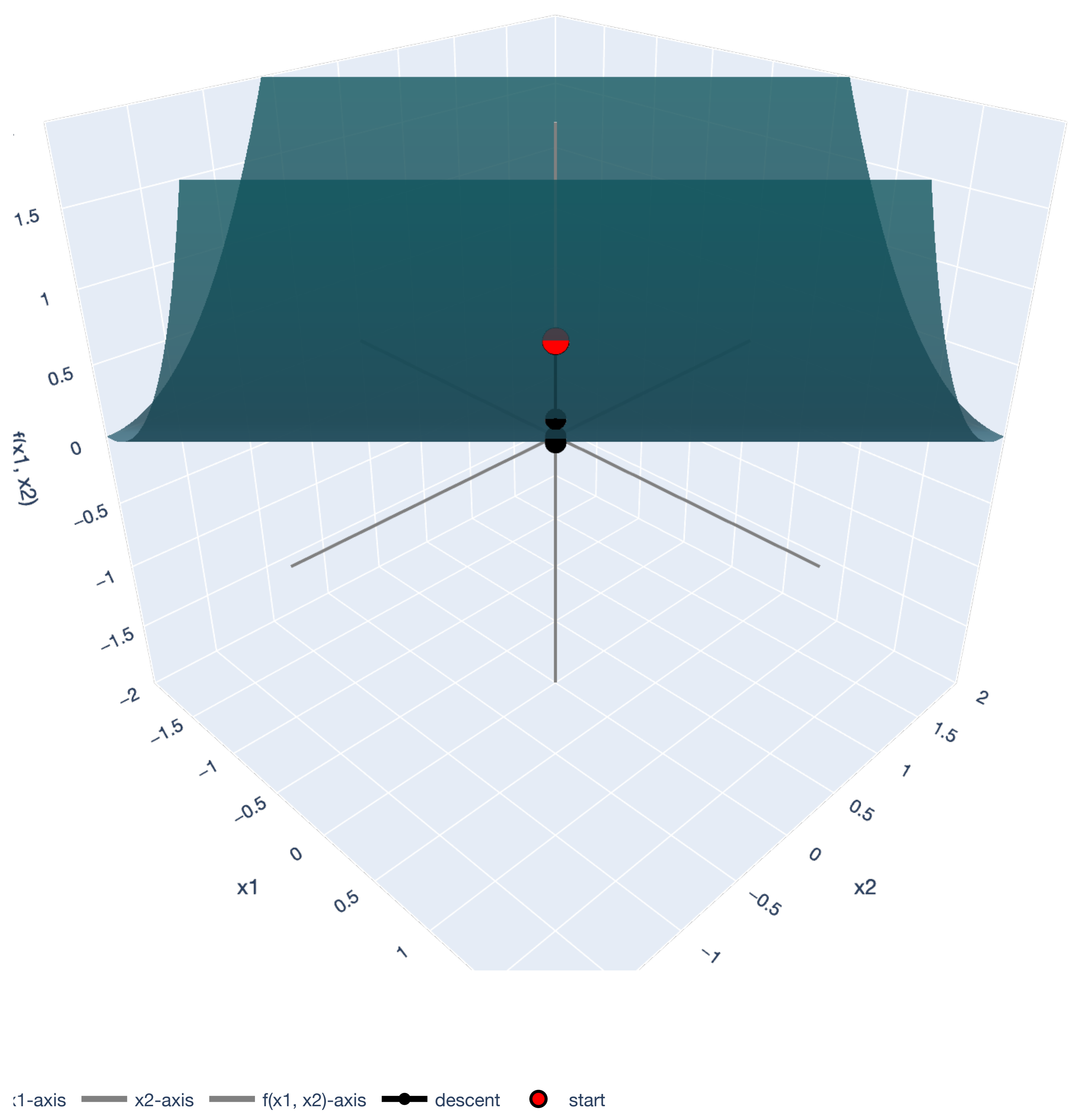


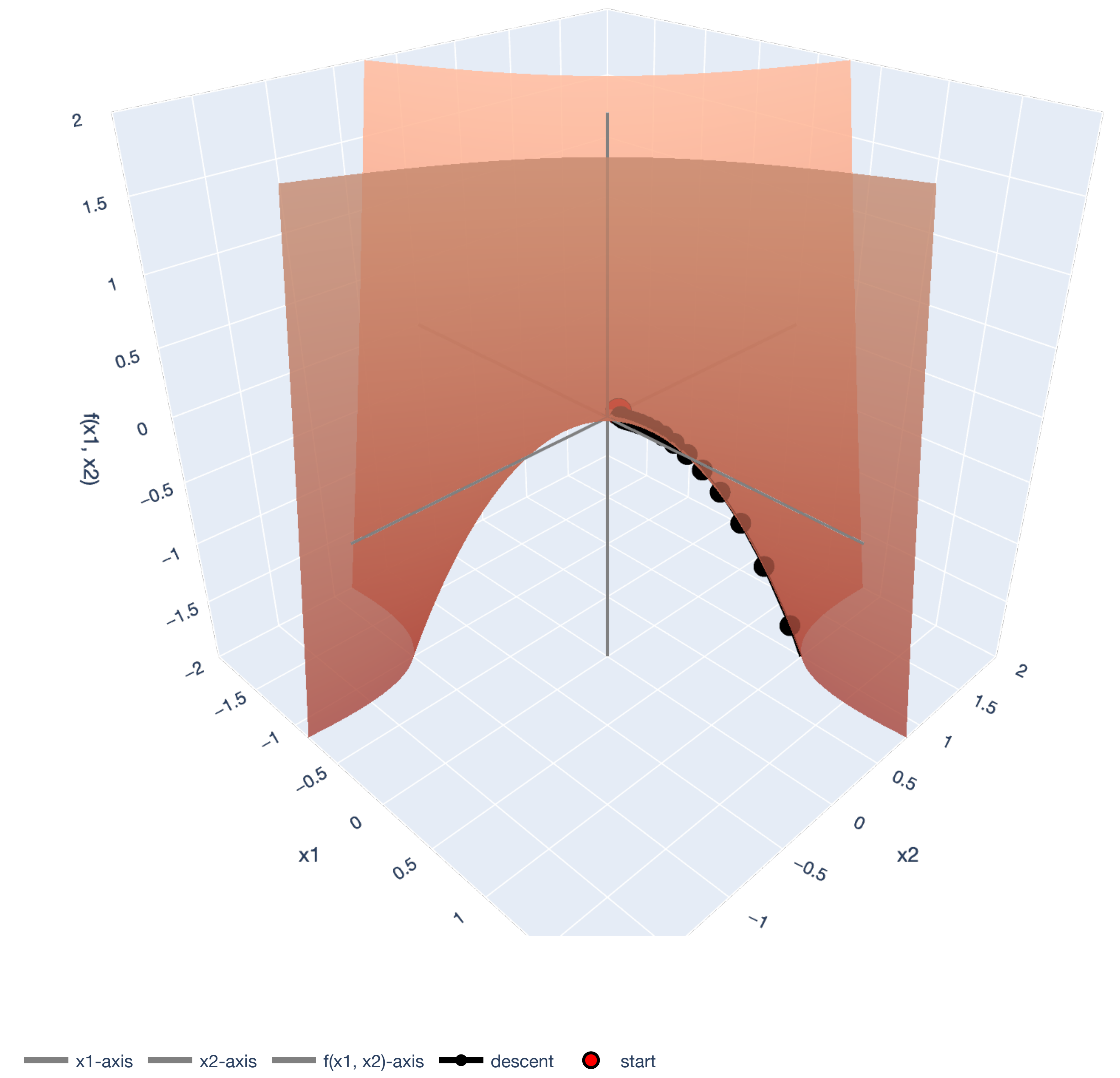
Lesson Overview

Big Picture: Gradient Descent

QUADRATIC FUNCTIONS







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: Eigenvalues and eigenvectors. Daniel Hsu.