Linear Classifiers: Expressiveness

Machine Learning

Lecture outline

- Linear models: Introduction

- What functions do linear classifiers express?

Where are we?

- Linear models: Introduction
- What functions do linear classifiers express?
  - Conjunctions and disjunctions
  - m-of-n functions
  - Not all functions are linearly separable
  - Feature space transformations
  - Exercises

Which Boolean functions can linear classifiers represent?

- Linear classifiers are an expressive hypothesis class

- Many Boolean functions are linearly separable
  - Not all though
  - Recall: In comparison, decision trees can represent any Boolean function

Conjunctions and disjunctions

$y = x_1 \land x_2 \land x_3$ is equivalent to "$y = 1$ whenever $x_1 + x_2 + x_3 \geq 3$"

x_1	x_2	x_3	y
0	0	0	0
0	0	1	0
0	1	0	0
0	1	1	0
1	0	0	0
1	0	1	0
1	1	0	0
1	1	1	1

Conjunctions and disjunctions

$y = x_1 \wedge x_2 \wedge x_3$ is equivalent to "$y = 1$ whenever $x_1 + x_2 + x_3 \geq 3$"

x_1	x_2	x_3	y	$x_1 + x_2 + x_3 - 3$	sign
0	0	0	0	-3	0
0	0	1	0	-2	0
0	1	0	0	-2	0
0	1	1	0	-1	0
1	0	0	0	-2	0
1	0	1	0	-1	0
1	1	0	0	-1	0
1	1	1	1	0	1

Conjunctions and disjunctions

$y = x_1 \wedge x_2 \wedge x_3$ is equivalent to "$y = 1$ whenever $x_1 + x_2 + x_3 \geq 3$"

x_1	x_2	x_3	y	$x_1 + x_2 + x_3 - 3$	sign
0	0	0	0	-3	0
0	0	1	0	-2	0
0	1	0	0	-2	0
0	1	1	0	-1	0
1	0	0	0	-2	0
1	0	1	0	-1	0
1	1	0	0	-1	0
1	1	1	1	0	1

Negations are okay too.
In general, use $1 - x$ in the linear threshold unit if x is negated

$$y = x_1 \wedge x_2 \wedge \neg x_3$$

corresponds to

$$x_1 + x_2 + (1 - x_3) \geq 3$$

Conjunctions and disjunctions

$y = x_1 \wedge x_2 \wedge x_3$ is equivalent to "$y = 1$ whenever $x_1 + x_2 + x_3 \geq 3$"

x_1	x_2	x_3	y	$x_1 + x_2 + x_3 - 3$	sign
0	0	0	0	-3	0
0	0	1	0	-2	0
0	1	0	0	-2	0
0	1	1	0	-1	0
1	0	0	0	-2	0
1	0	1	0	-1	0
1	1	0	0	-1	0
1	1	1	1	0	1

Negations are okay too.
In general, use $1 - x$ in the linear threshold unit if x is negated

$$y = x_1 \wedge x_2 \wedge \neg x_3$$

corresponds to

$$x_1 + x_2 + (1 - x_3) \geq 3$$

Exercise: What would the linear threshold function be if the conjunctions here were replaced with disjunctions?

Conjunctions and disjunctions

$y = x_1 \wedge x_2 \wedge x_3$ is equivalent to "$y = 1$ whenever $x_1 + x_2 + x_3 \geq 3$"

x_1	x_2	x_3	y	$x_1 + x_2 + x_3 - 3$	sign
0	0	0	0	-3	0
0	0	1	0	-2	0
0	1	0	0	-2	0
0	1	1	0	-1	0
1	0	0	0	-2	0
1	0	1	0	-1	0
1	1	0	0	-1	0
1	1	1	1	0	1

Negations are okay too.
In general, use $1 - x$ in the linear threshold unit if x is negated

$y = x_1 \wedge x_2 \wedge \neg x_3$

corresponds to

$x_1 + x_2 + (1 - x_3) \geq 3$

Questions?

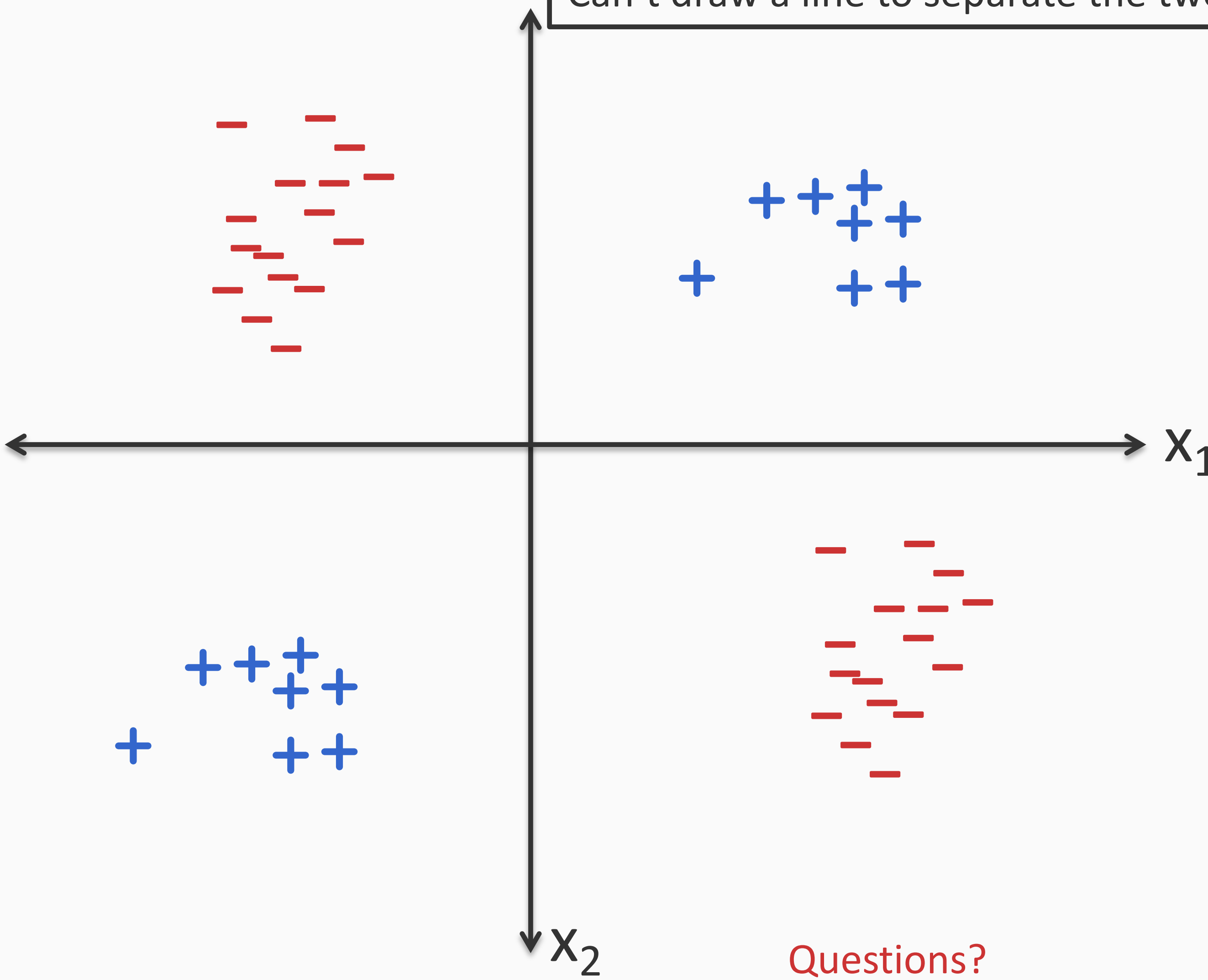
Exercise: What would the linear threshold function be if the conjunctions here were replaced with disjunctions?

Not all functions are linearly separable

Parity is not linearly separable

(The XOR function)

Can’t draw a line to separate the two classes



Questions?

Not all functions are linearly separable

- XOR is not linear
  - $y = x \text{ XOR } y = (x \land \neg y) \lor (\neg x \land y)$
  - *Parity* cannot be represented as a linear classifier
    - f(**x**) = 1 if the number of 1's is even

- Many non-trivial Boolean functions
  - Example: $y = (x_1 \land x_2) \lor (x_3 \land \neg x_4)$
  - The function is not linear in the four variables

Even these functions can be *made* linear

These points are not separable in 1-dimension by a line

What is a one-dimensional line, by the way?

Even these functions can be *made* linear

These points are not separable in 1-dimension by a line

What is a one-dimensional line, by the way?

The trick: Change the representation

The blown up feature space

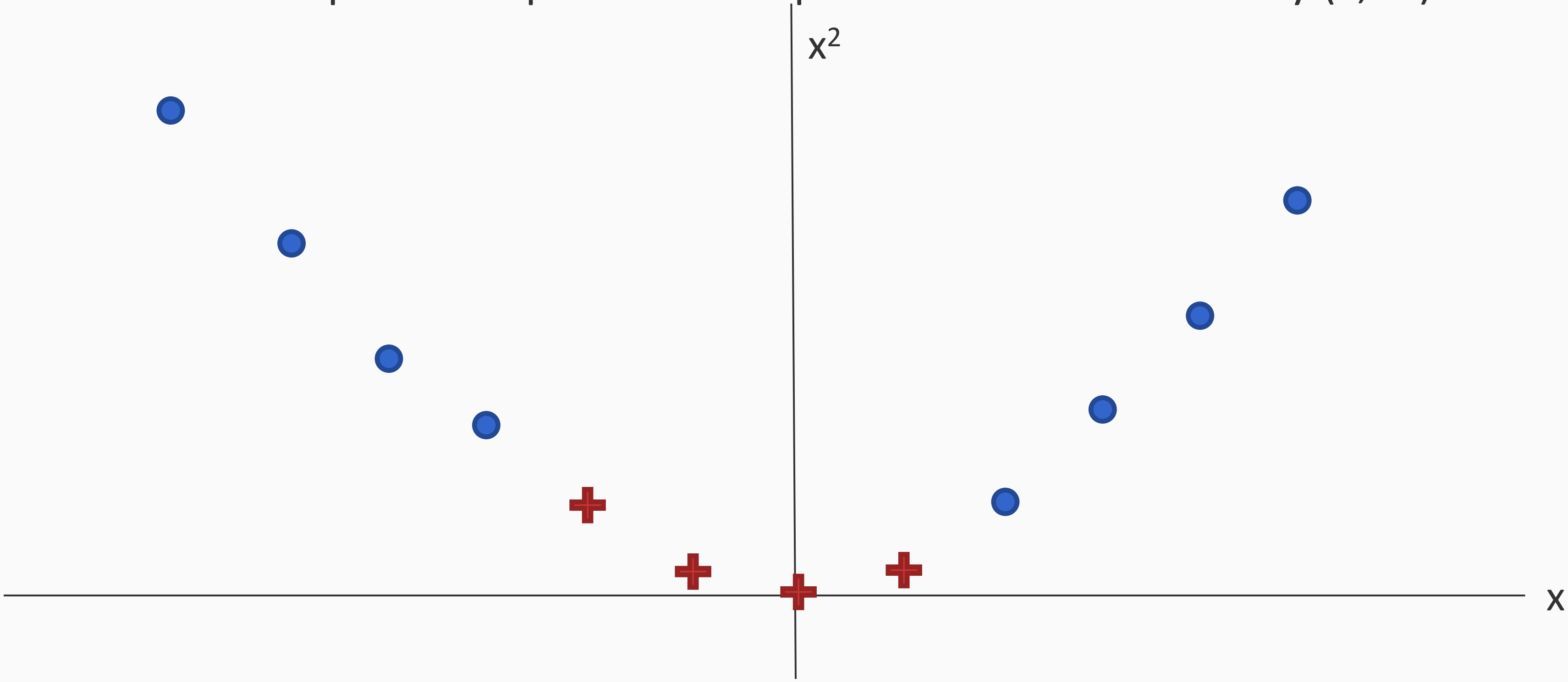
The trick: Use feature *conjunctions*

Transform points: Represent each point x in 2 dimensions by (x, x^2)

The blown up feature space

The trick: Use feature *conjunctions*

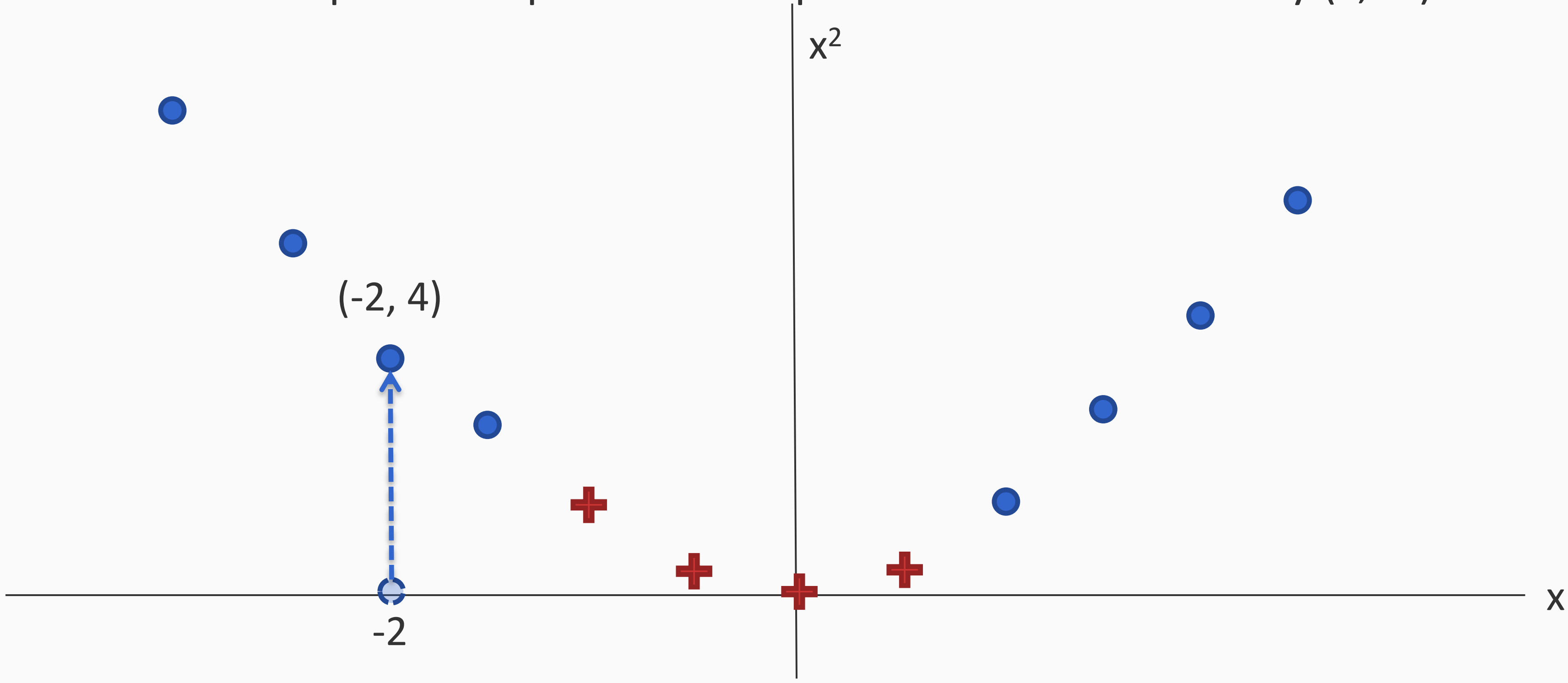
Transform points: Represent each point x in 2 dimensions by (x, x^2)



The blown up feature space

The trick: Use feature *conjunctions*

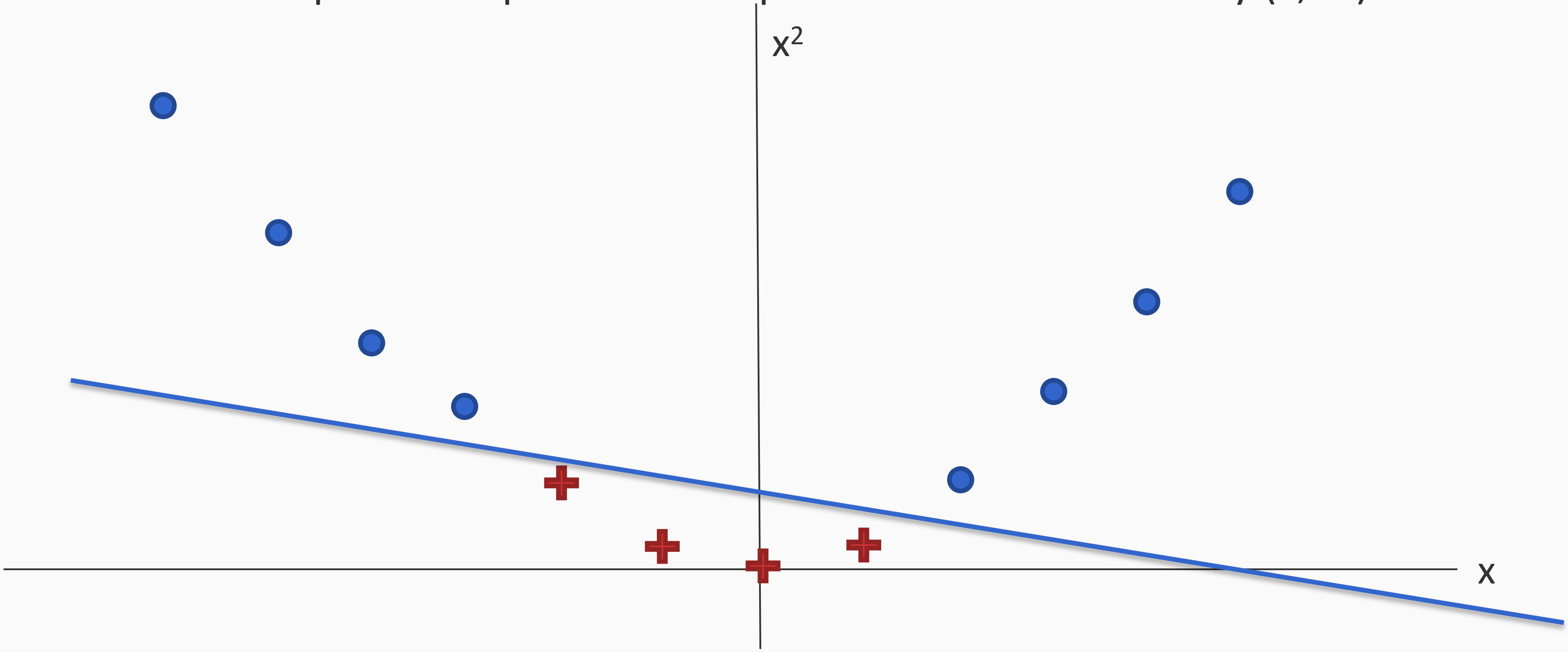
Transform points: Represent each point x in 2 dimensions by (x, x^2)



The blown up feature space

The trick: Use feature *conjunctions*

Transform points: Represent each point x in 2 dimensions by (x, x^2)



Now the data is linearly separable in this space!

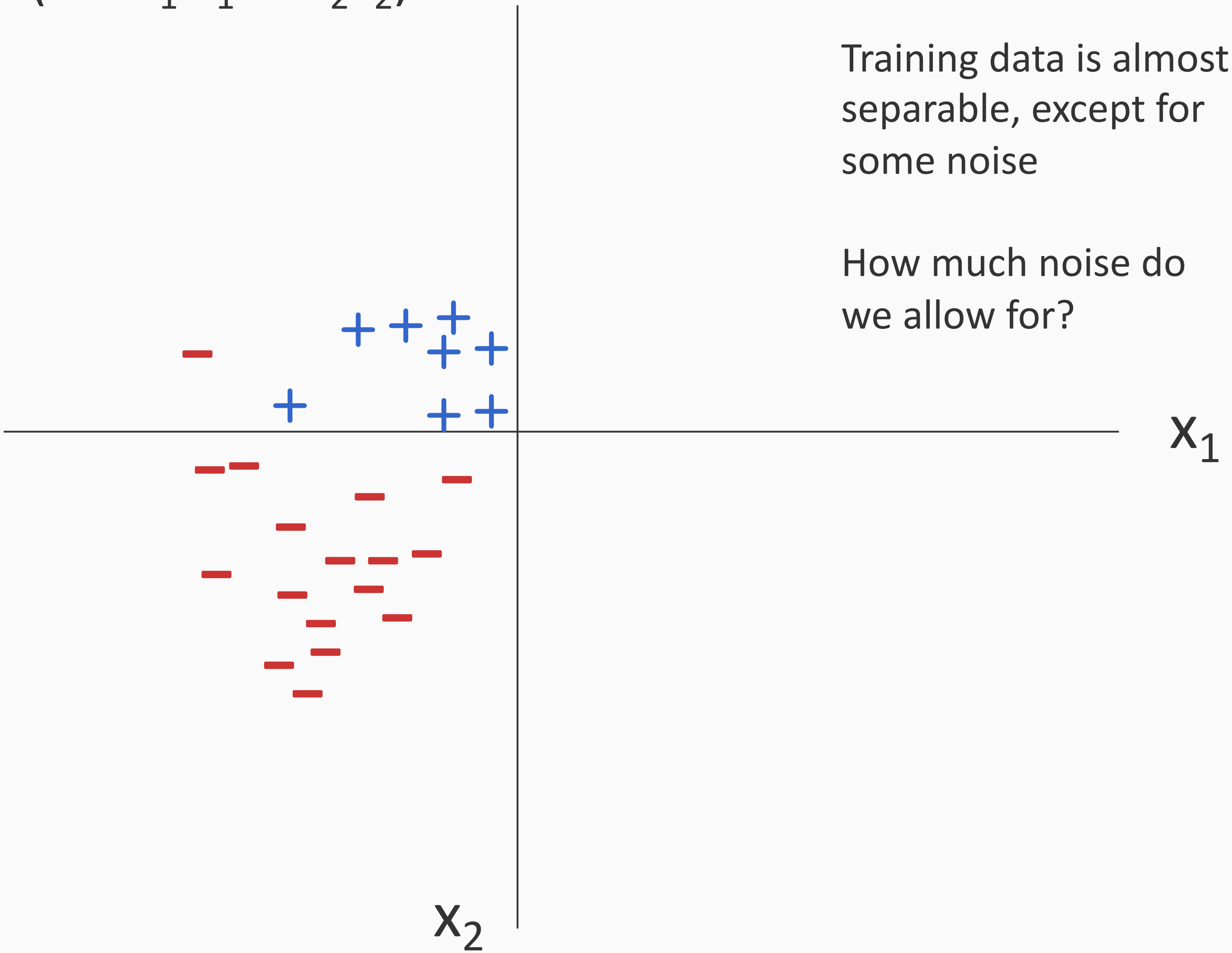
Exercise

How would you use the feature transformation idea to make XOR in two dimensions linearly separable in a new space?

To answer this question, you need to think about a function that maps examples from two dimensional space to a higher dimensional space.

Almost linearly separable data

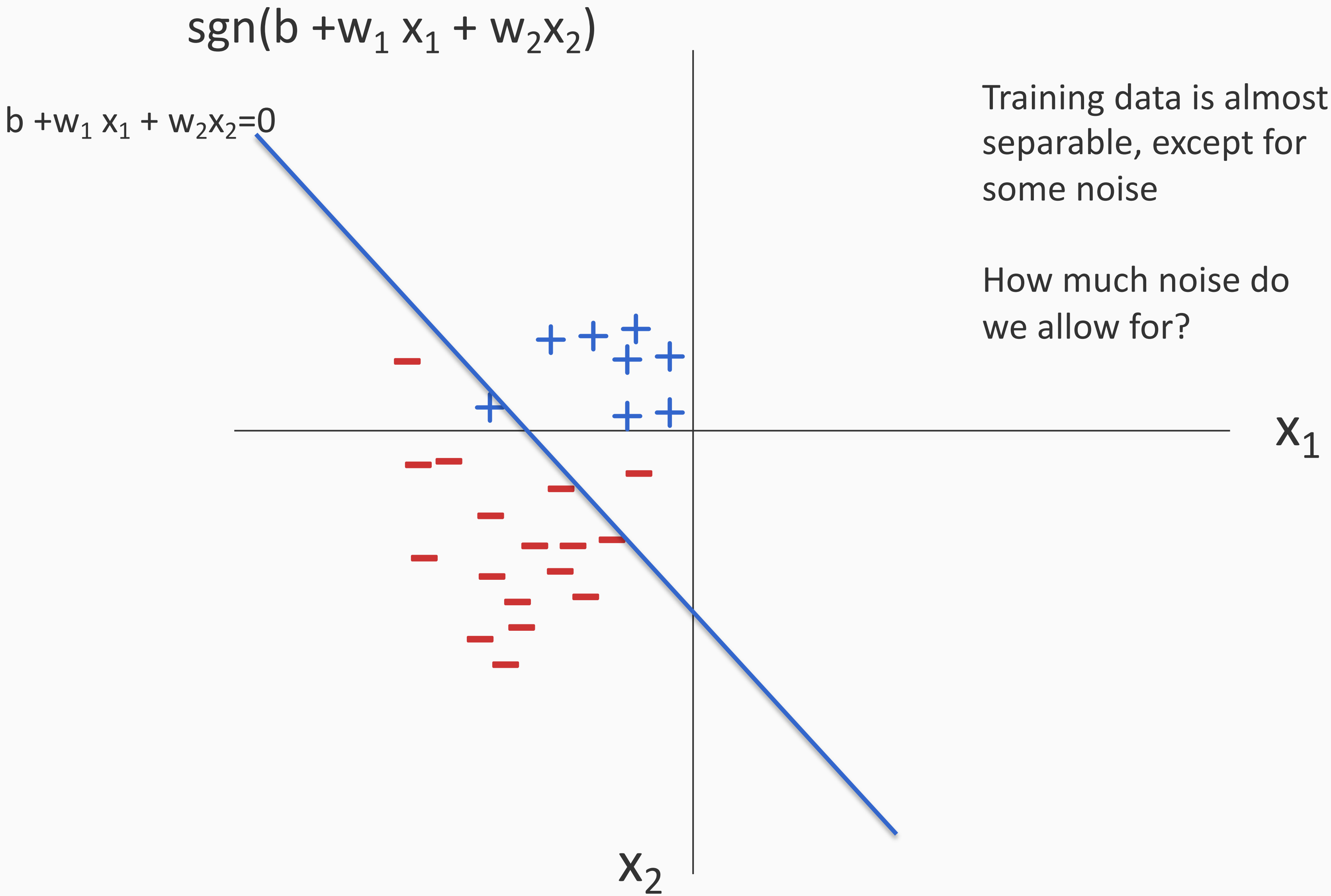
$\mathrm{sgn}(b + w_1 x_1 + w_2 x_2)$



Training data is almost separable, except for some noise

How much noise do we allow for?

Almost linearly separable data



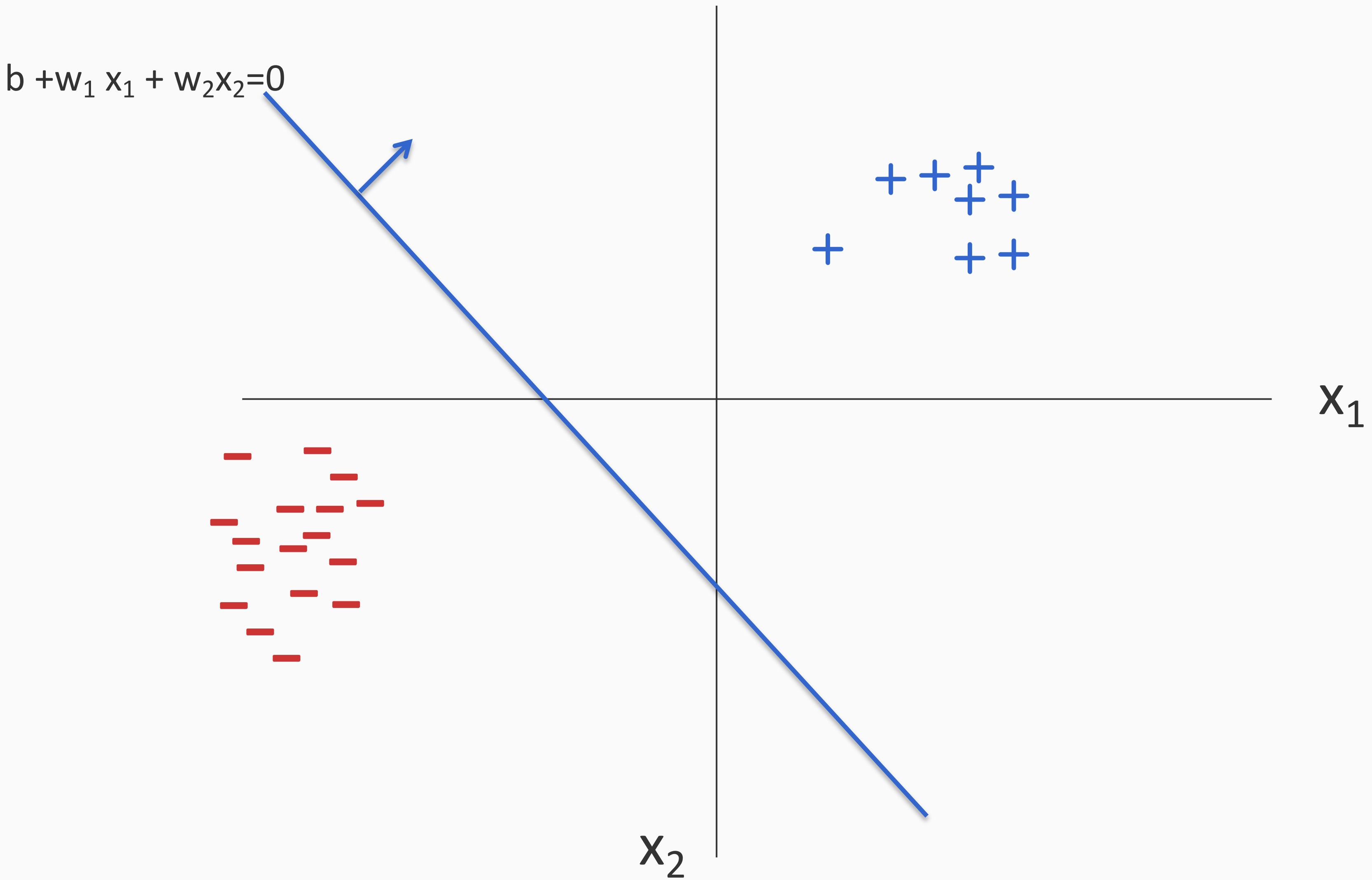
Training data is almost separable, except for some noise

How much noise do we allow for?

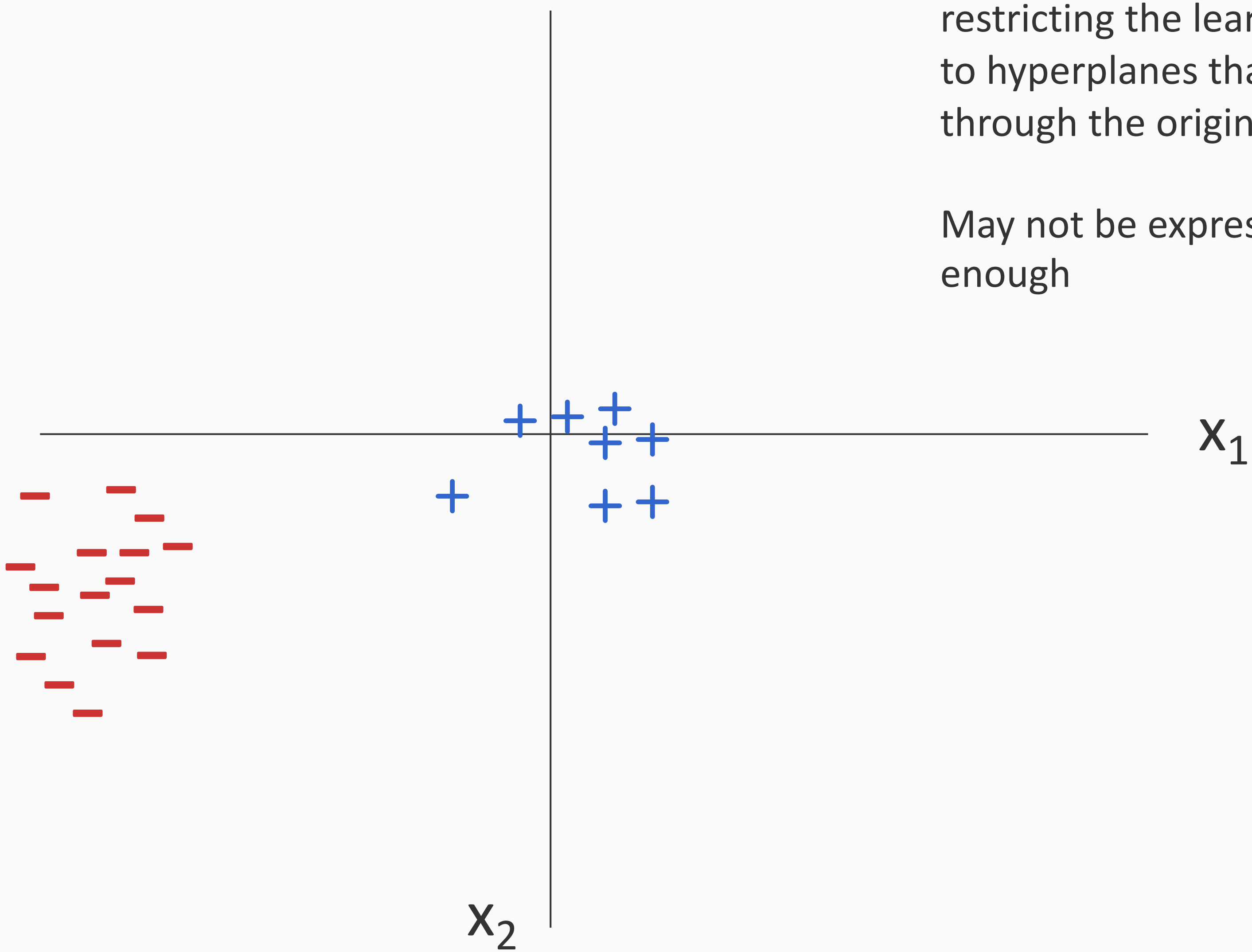
Linear classifiers: An expressive hypothesis class

- Many functions are linear
- Often a good guess for a hypothesis space
- Some functions are not linear
  - The XOR function
  - Non-trivial Boolean functions
- But there are ways of making them linear in a higher dimensional feature space

Why is the bias term needed?



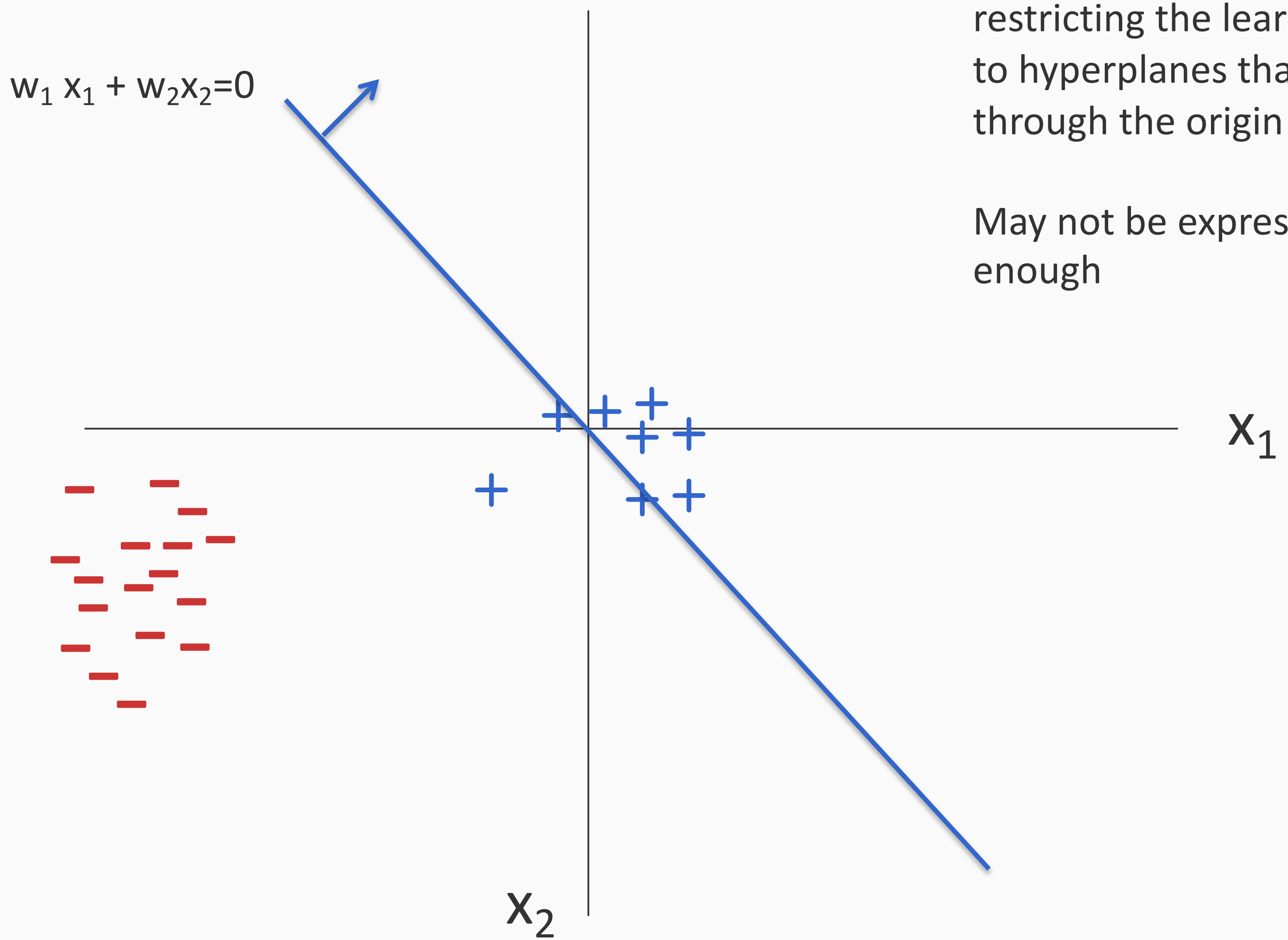
Why is the bias term needed?



If b is zero, then we are restricting the learner only to hyperplanes that go through the origin

May not be expressive enough

Why is the bias term needed?



If b is zero, then we are restricting the learner only to hyperplanes that go through the origin

May not be expressive enough

Exercises

1. Represent the simple disjunction as a linear classifier.

2. How would you apply the feature space expansion idea for the XOR function?