6.036/6.862: Introduction to Machine Learning

Lecture: starts Tuesdays 9:35am (Boston time zone)

Course website: introml.odl.mit.edu

Who's talking? Prof. Tamara Broderick

Questions? discourse.odl.mit.edu ("Lecture 6" category)

Materials: Will all be available at course website

Last Time(s)

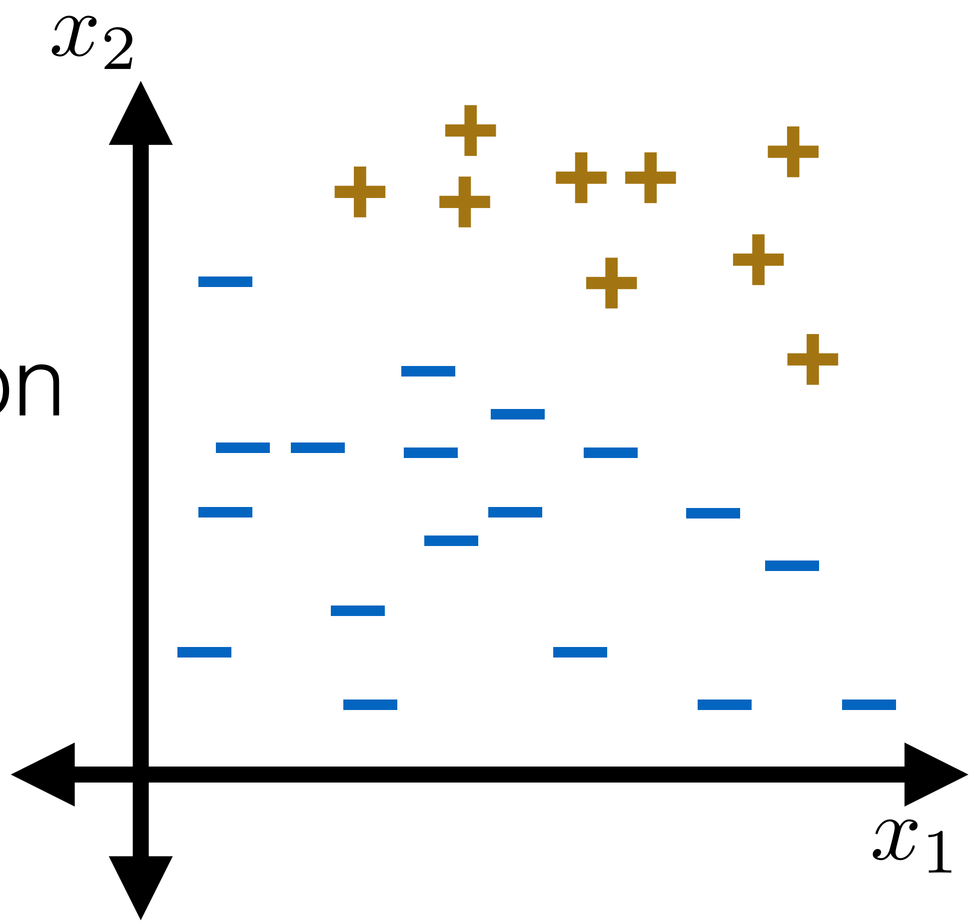
I. Linear classification
II. Linear regression
III. Choosing features

Today's Plan

I. Step-function features
II. Neural nets: hypothesis class
III. Neural nets: learning

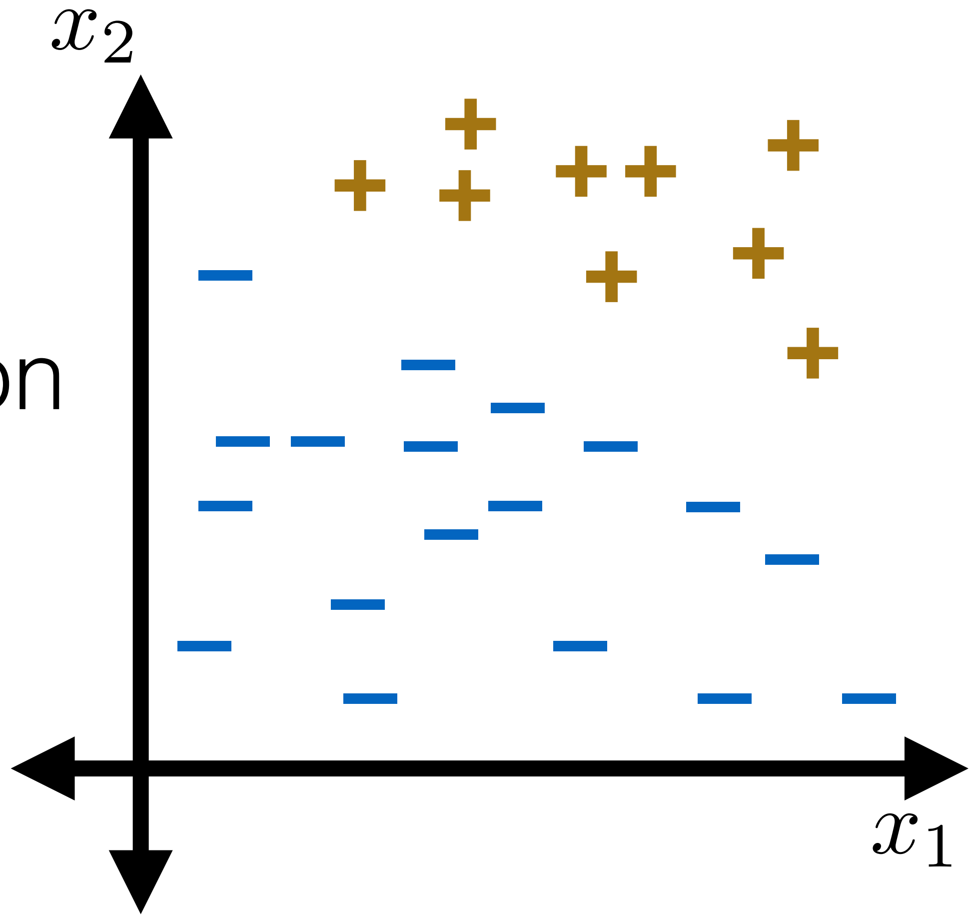
Recall

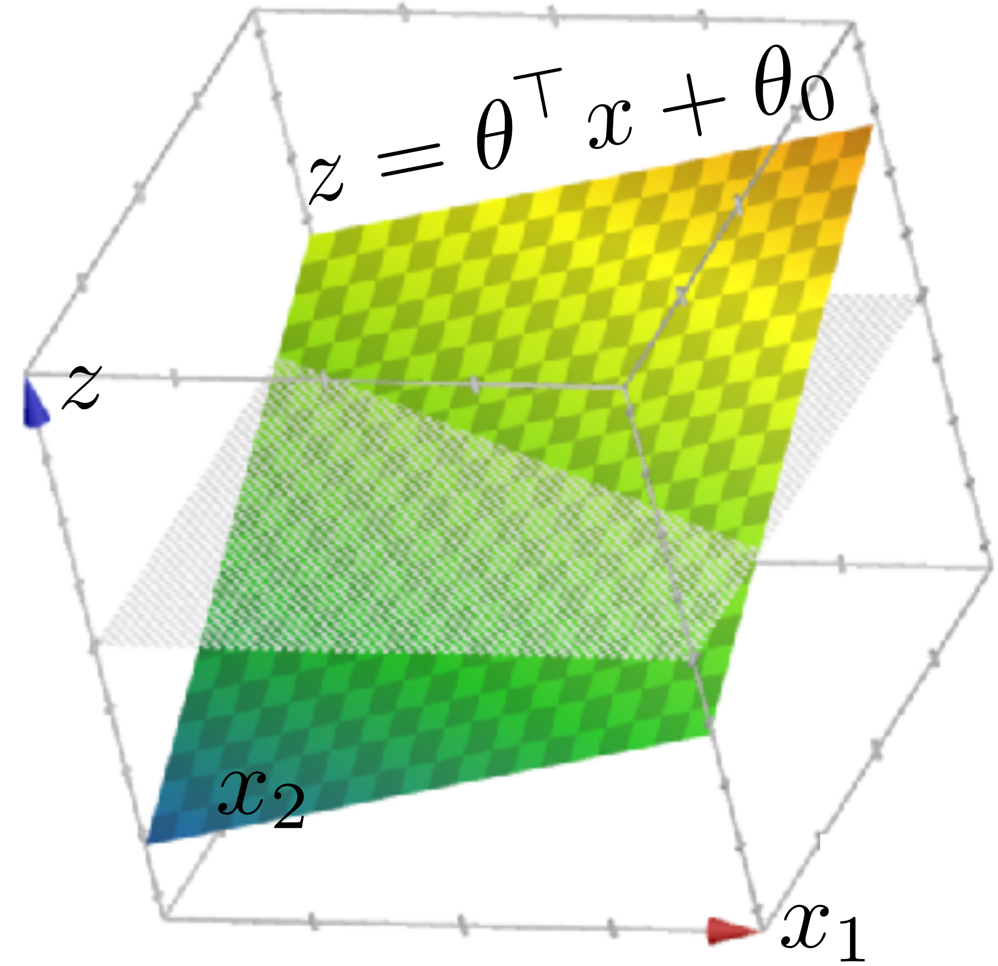
- Linear classification with default features:



Recall

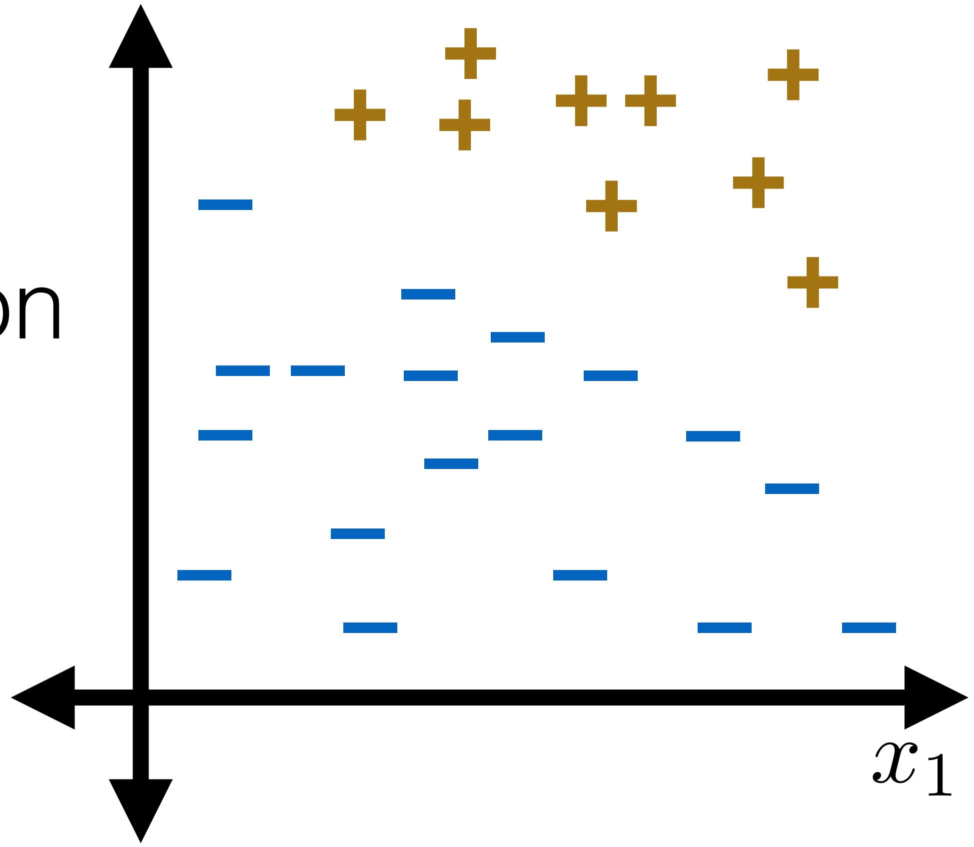
- Linear classification with default features:

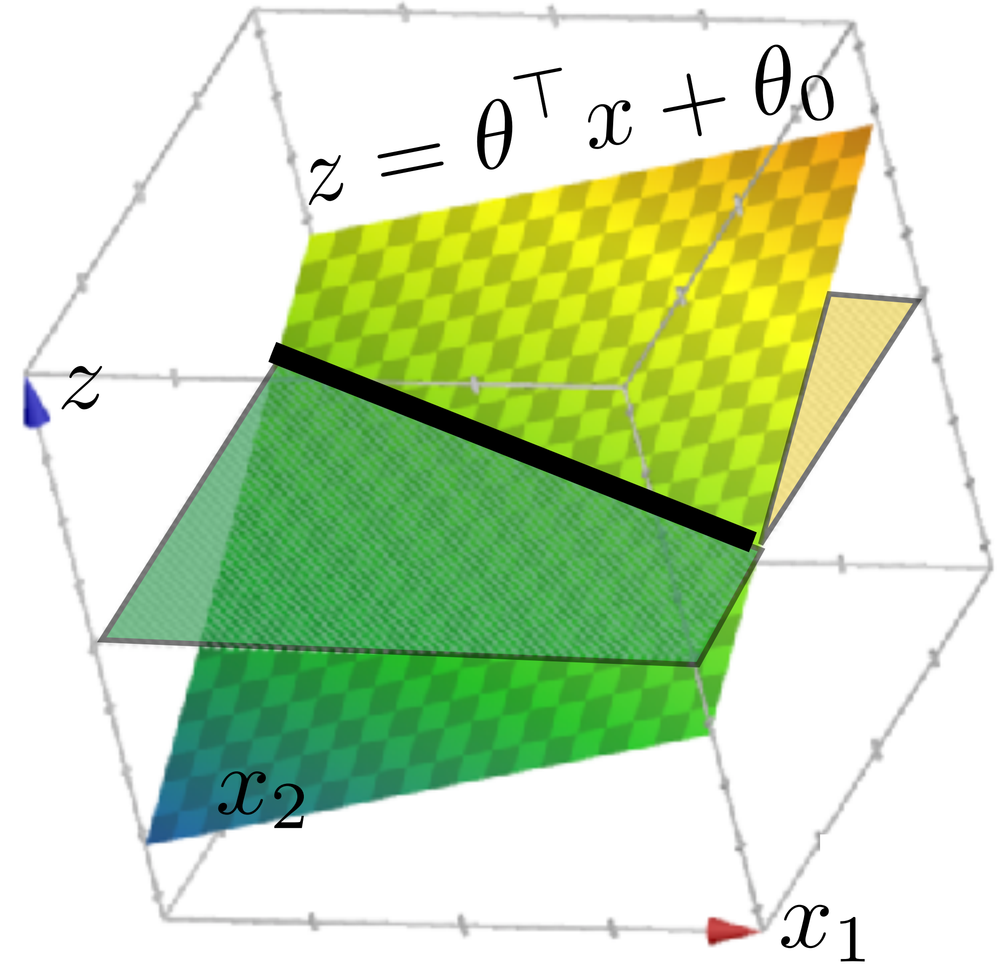




Recall

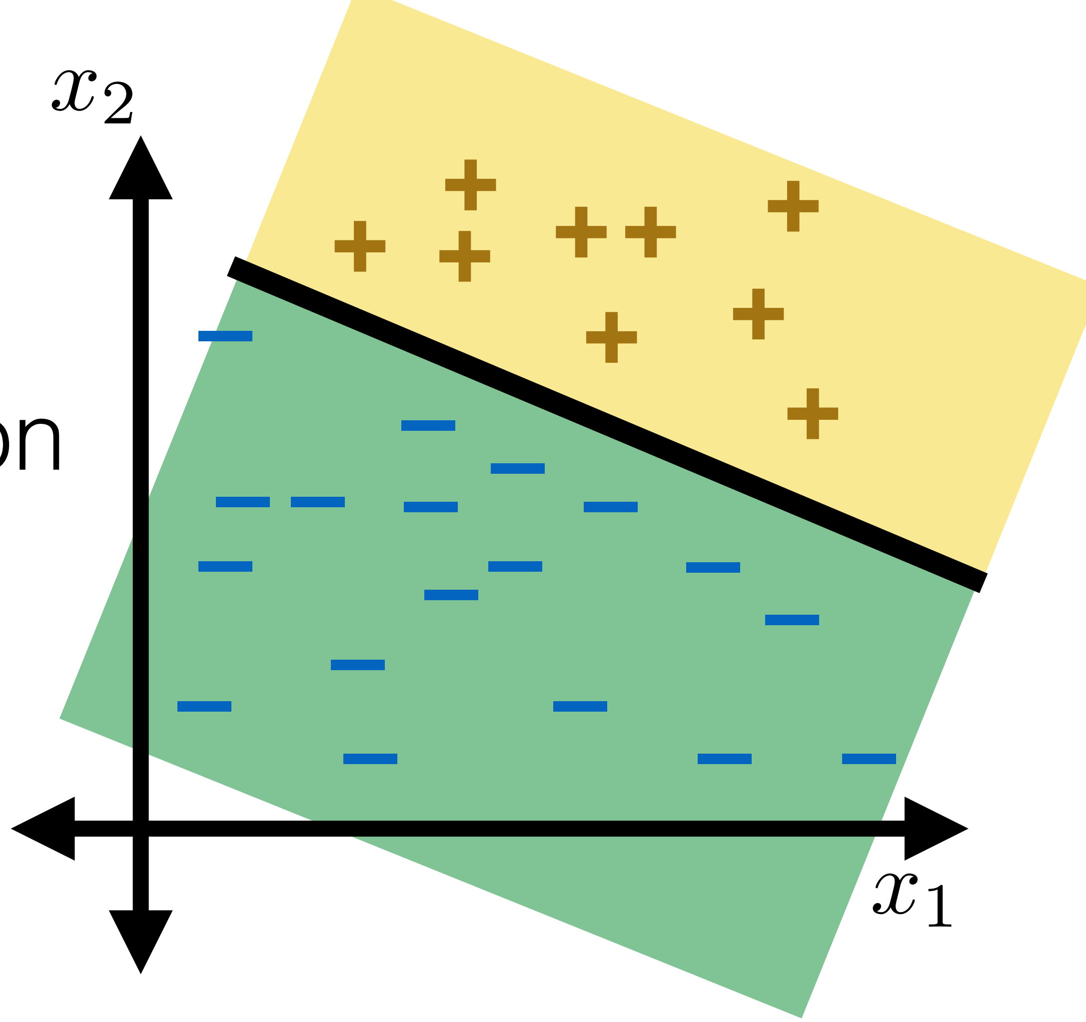
- Linear classification with default features:

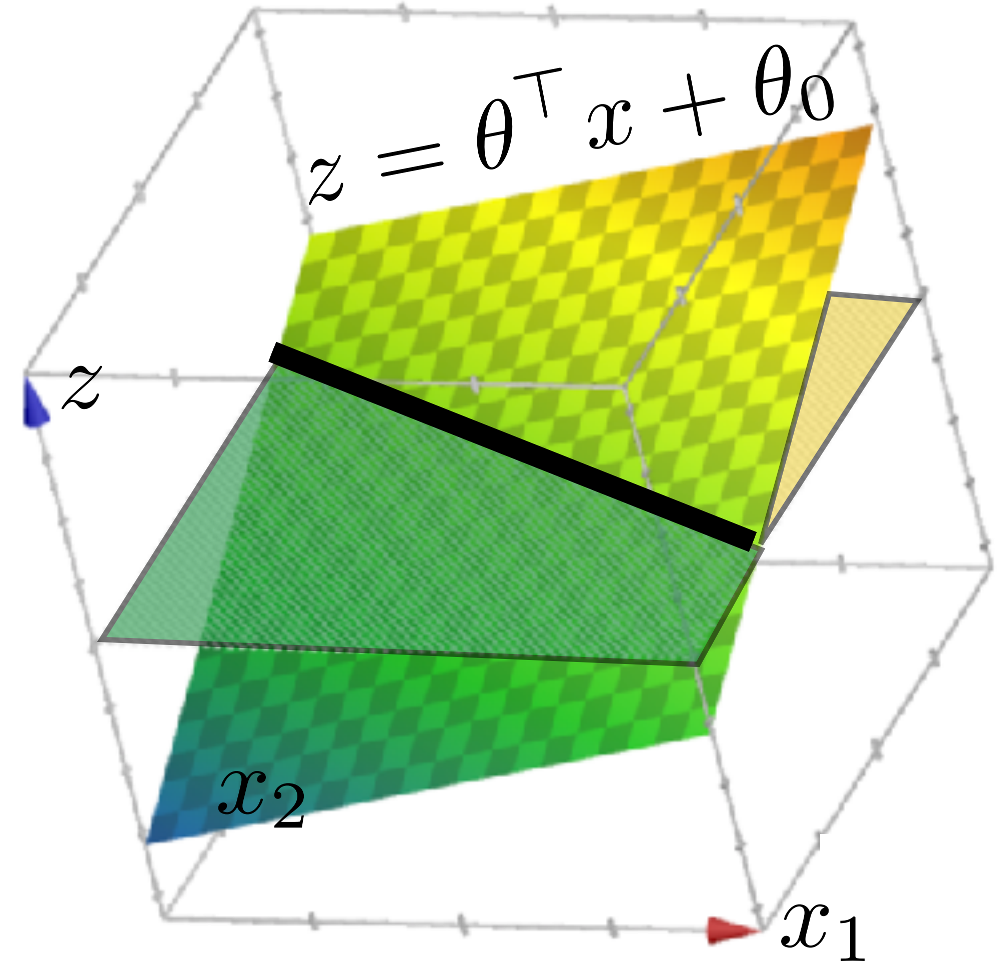




Recall

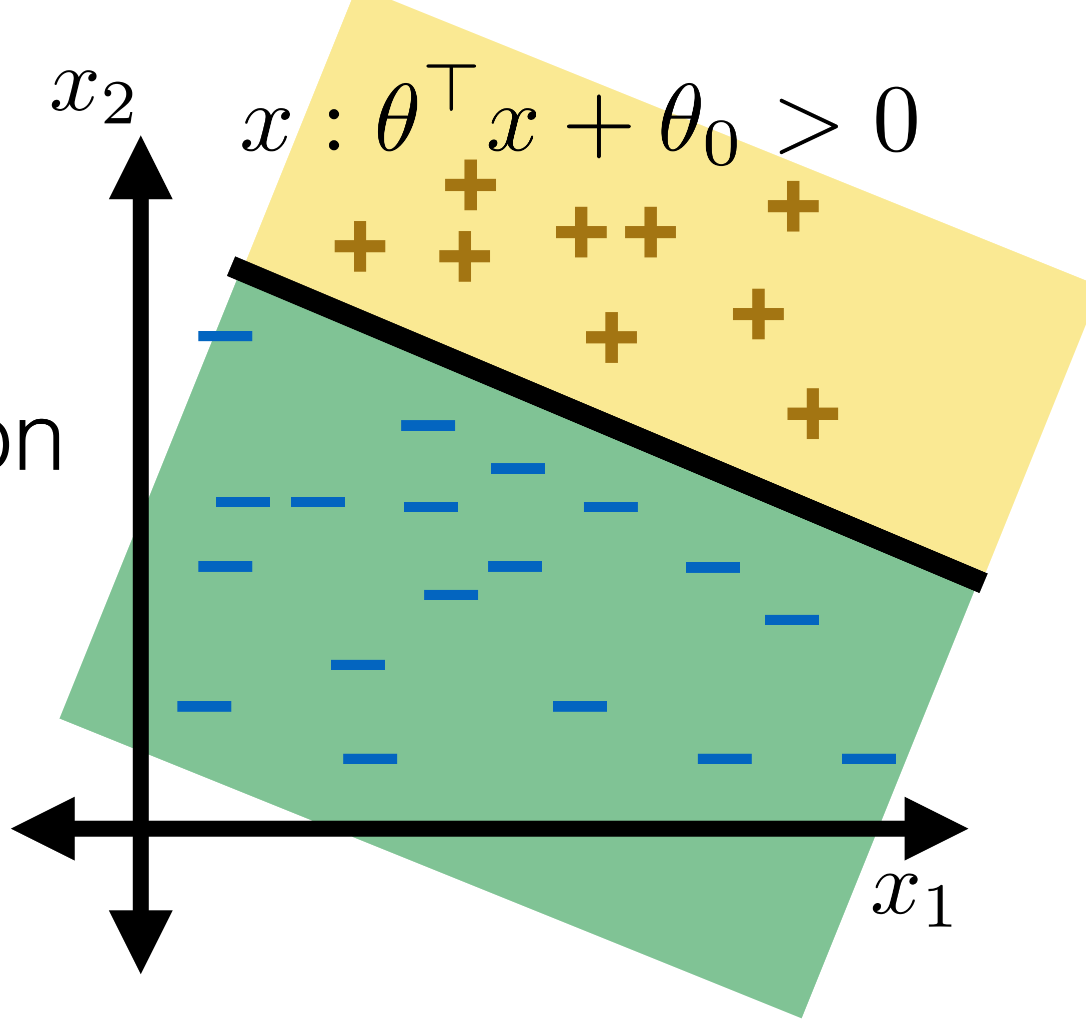
- Linear classification with default features:

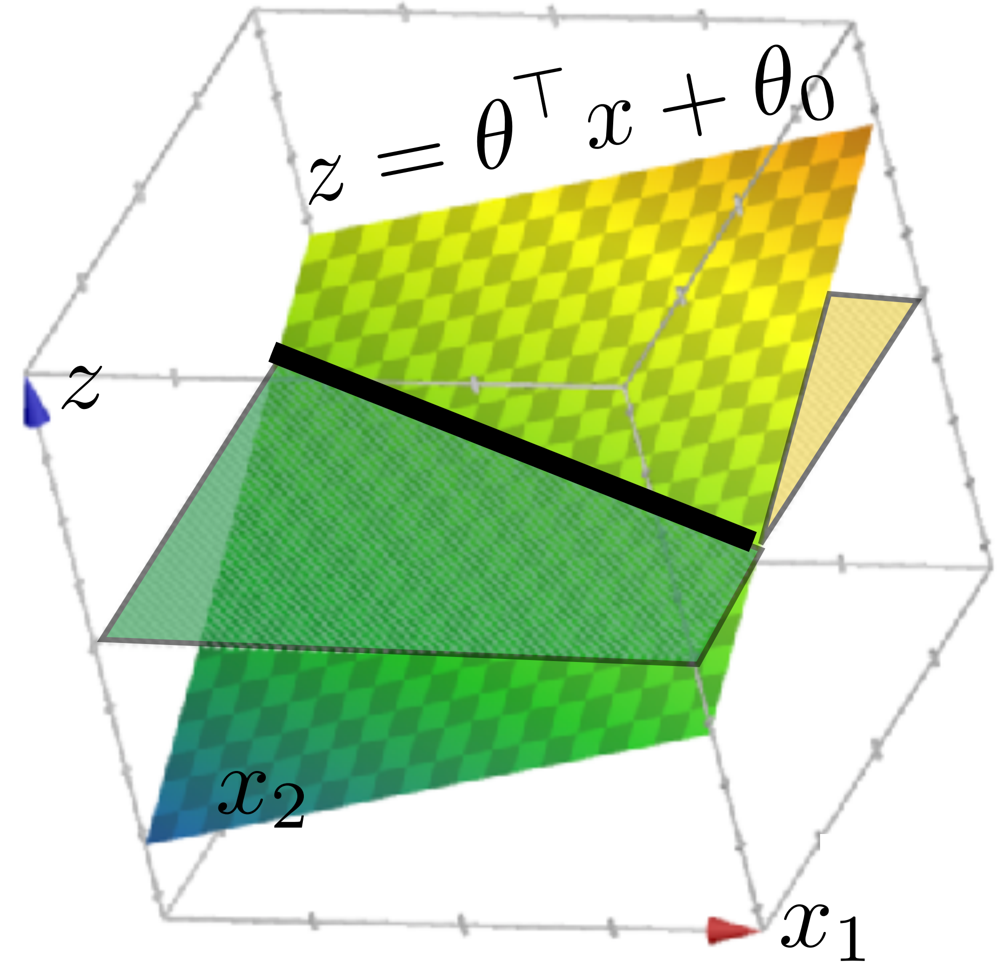




Recall

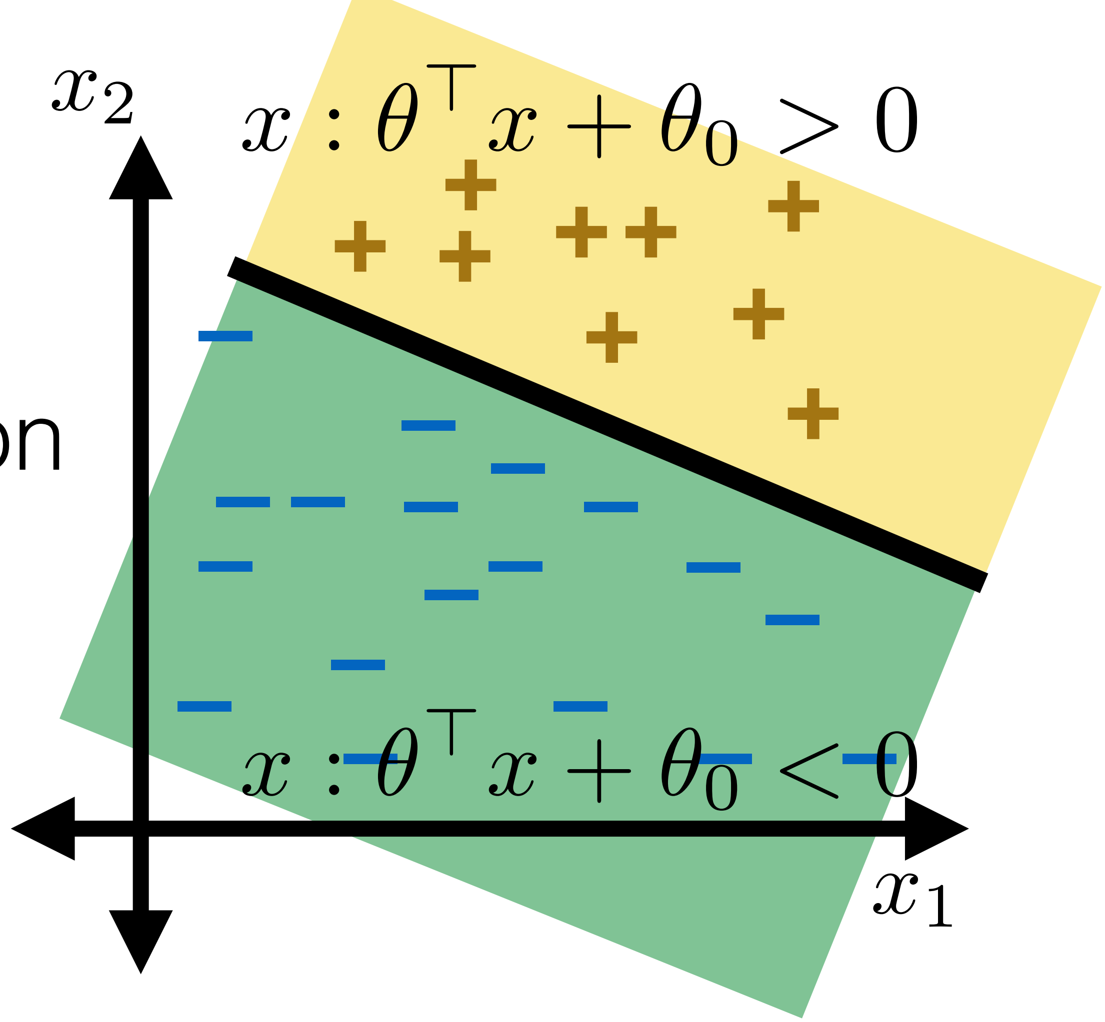
- Linear classification with default features:

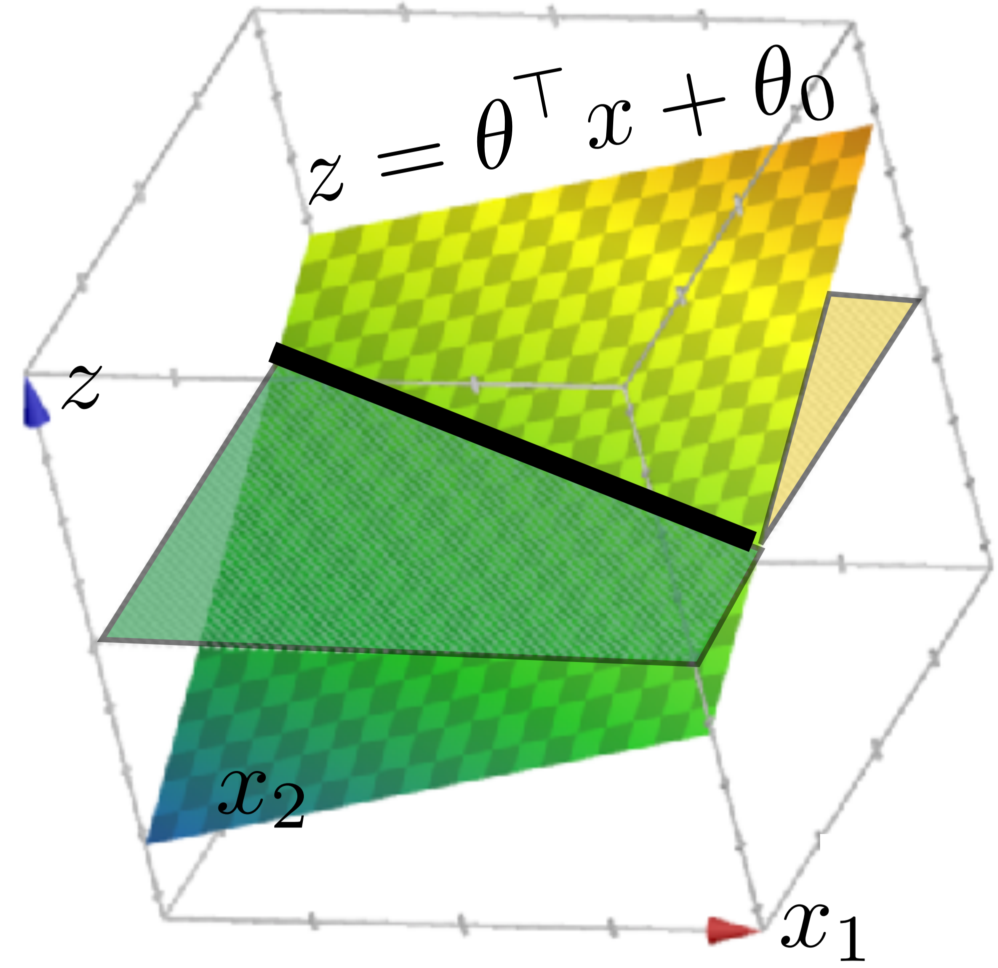




Recall

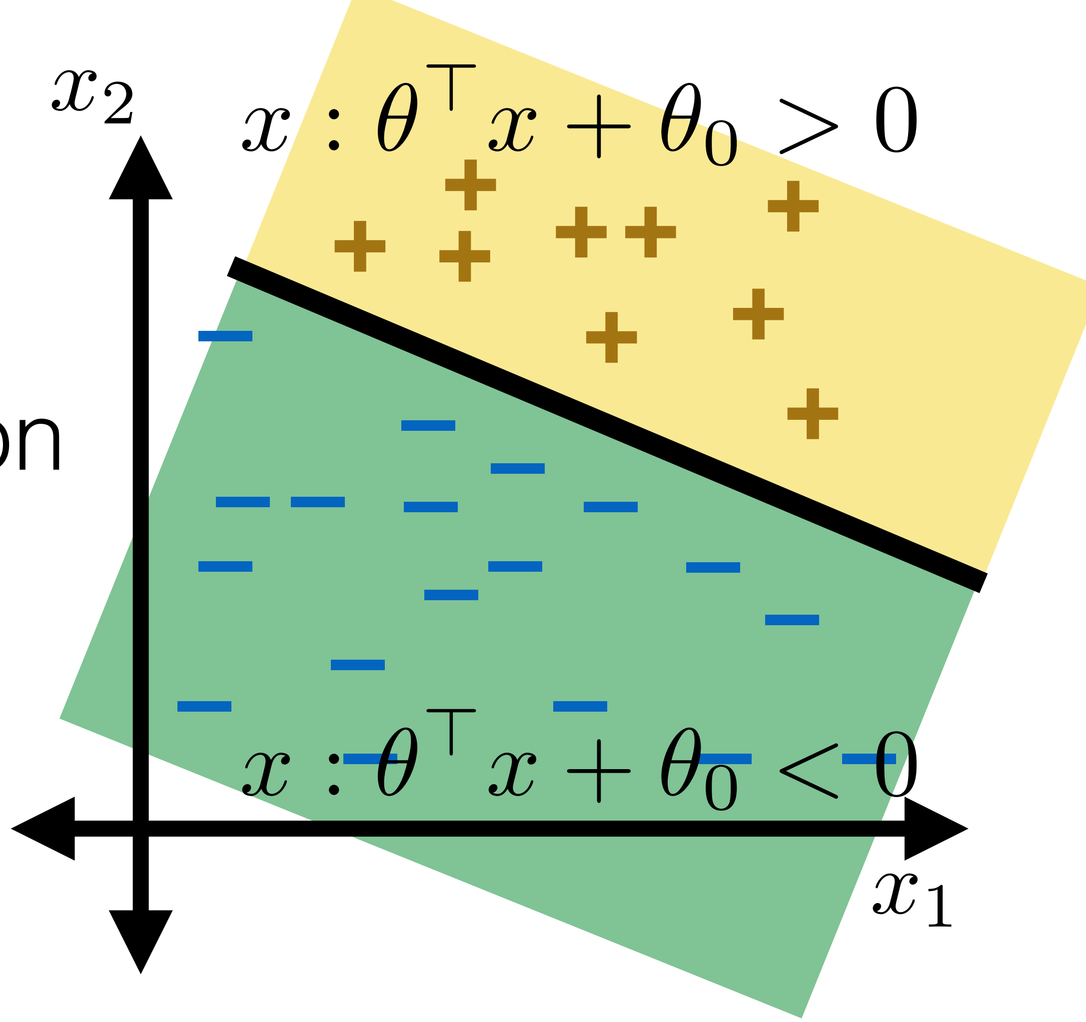
- Linear classification with default features:

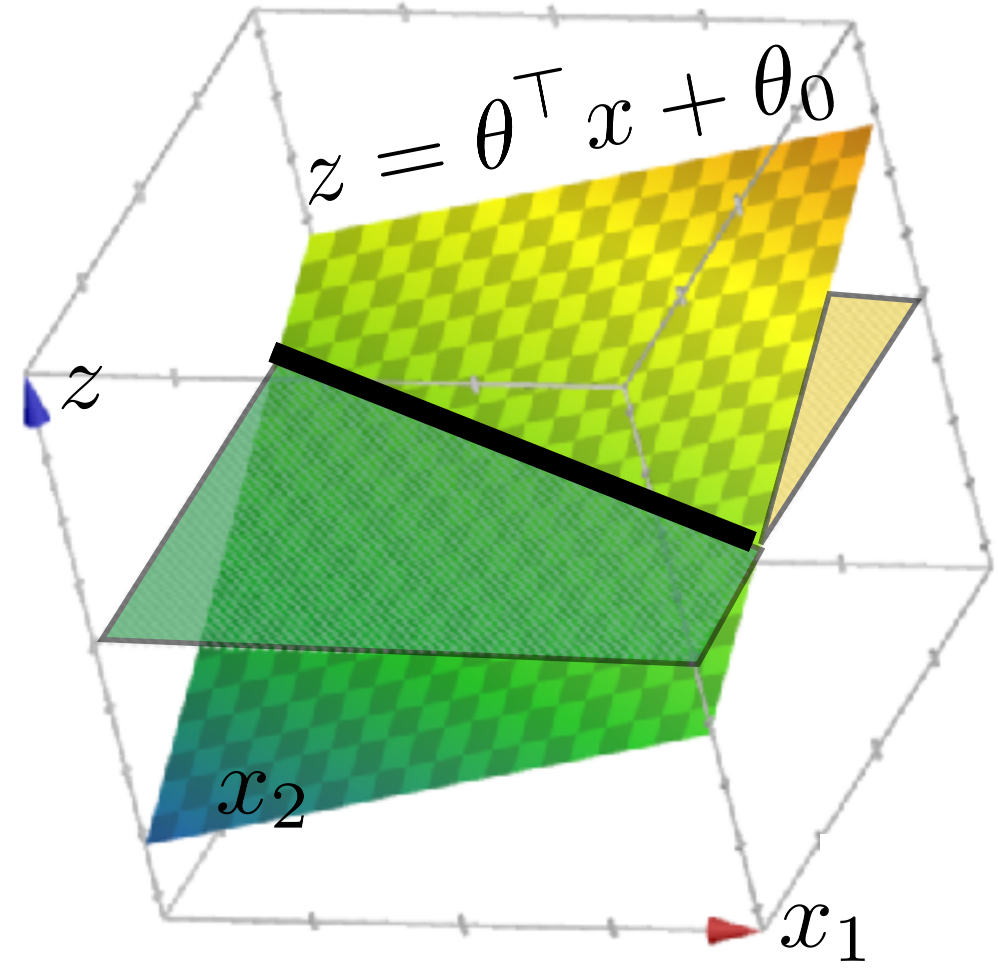




Recall

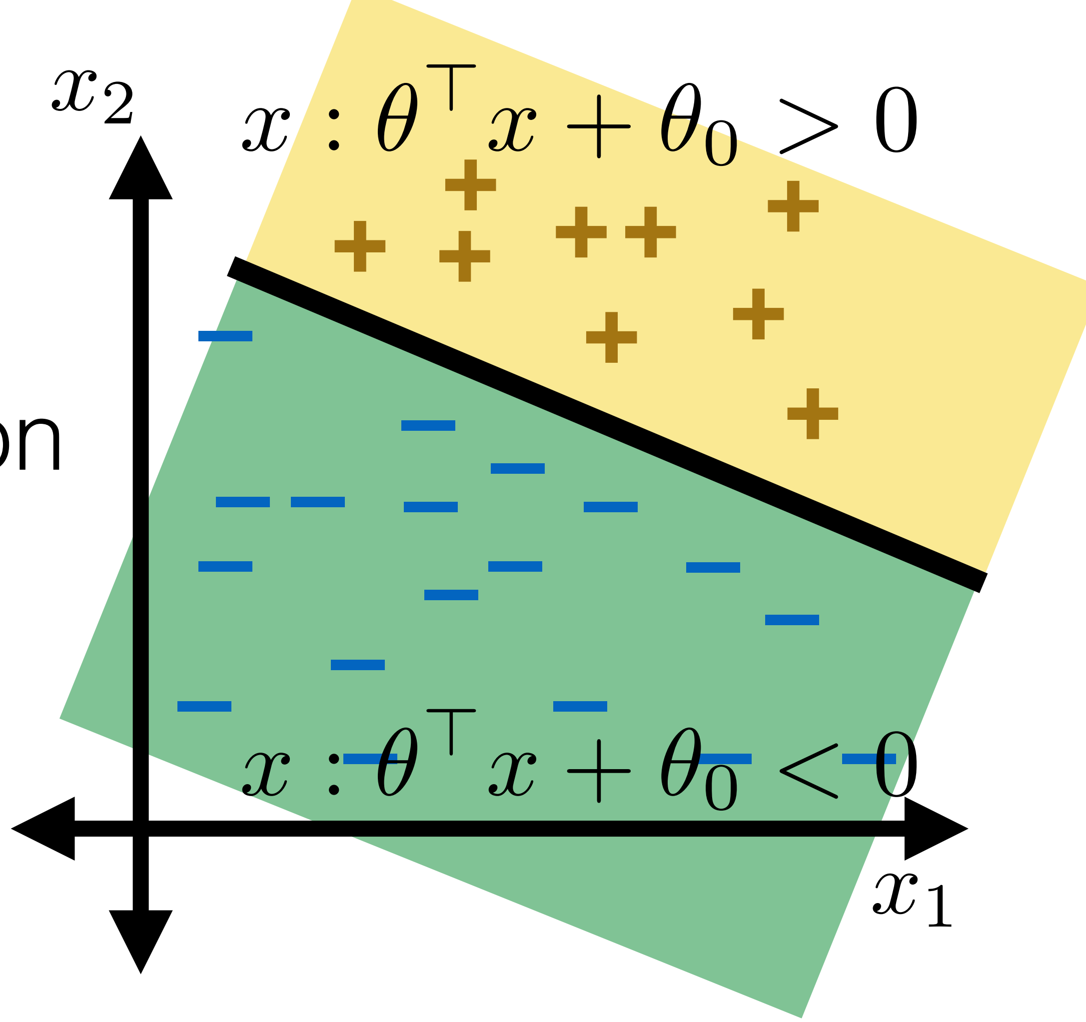
- Linear classification with default features:

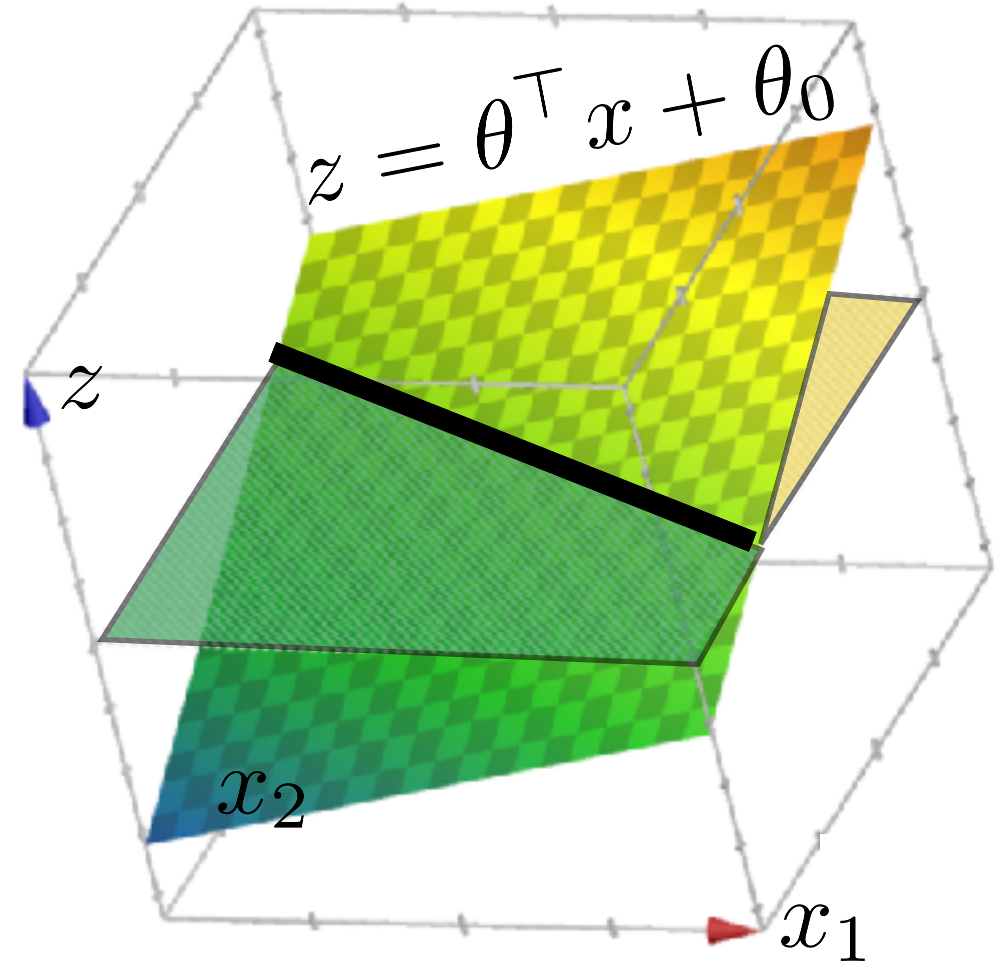




- Linear classification with polynomial features:

Recall





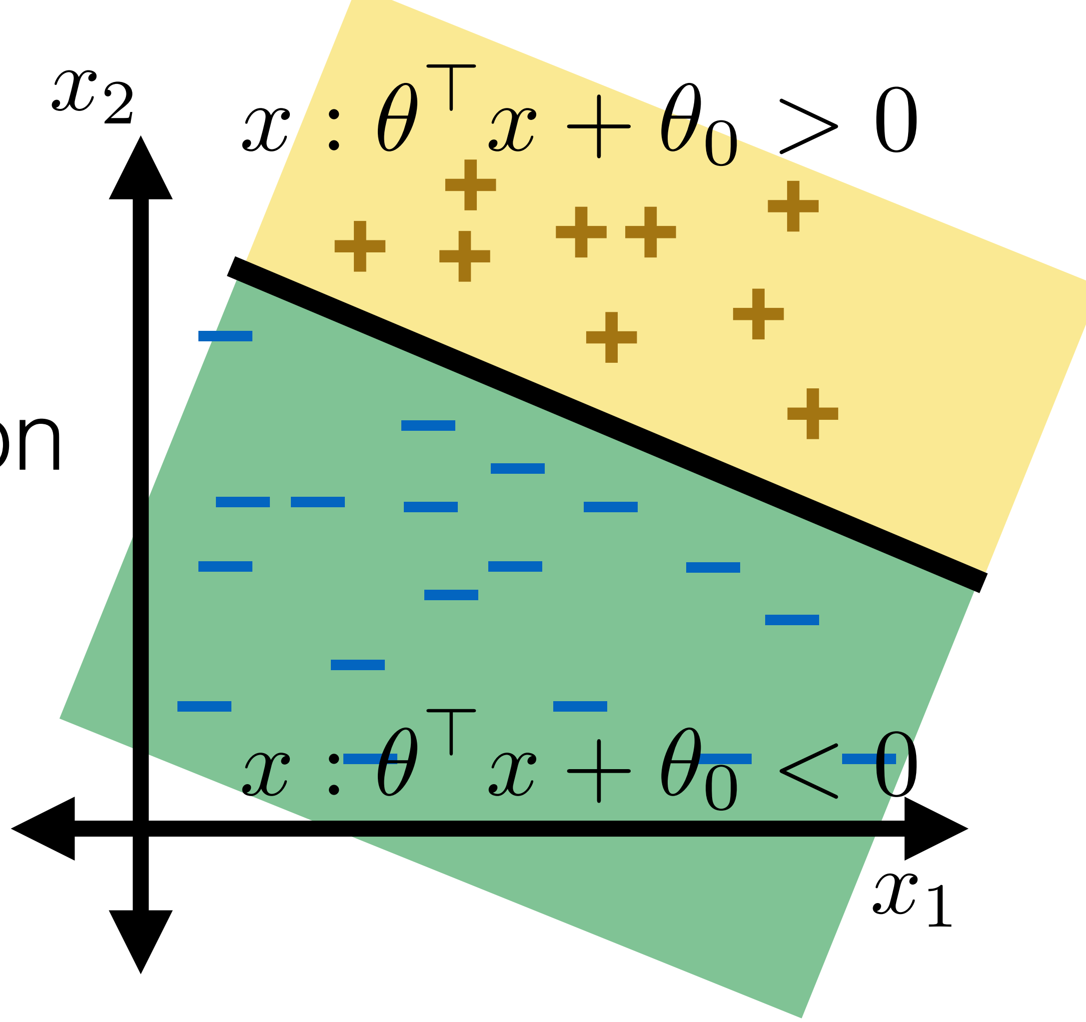
- Linear classification with default features:

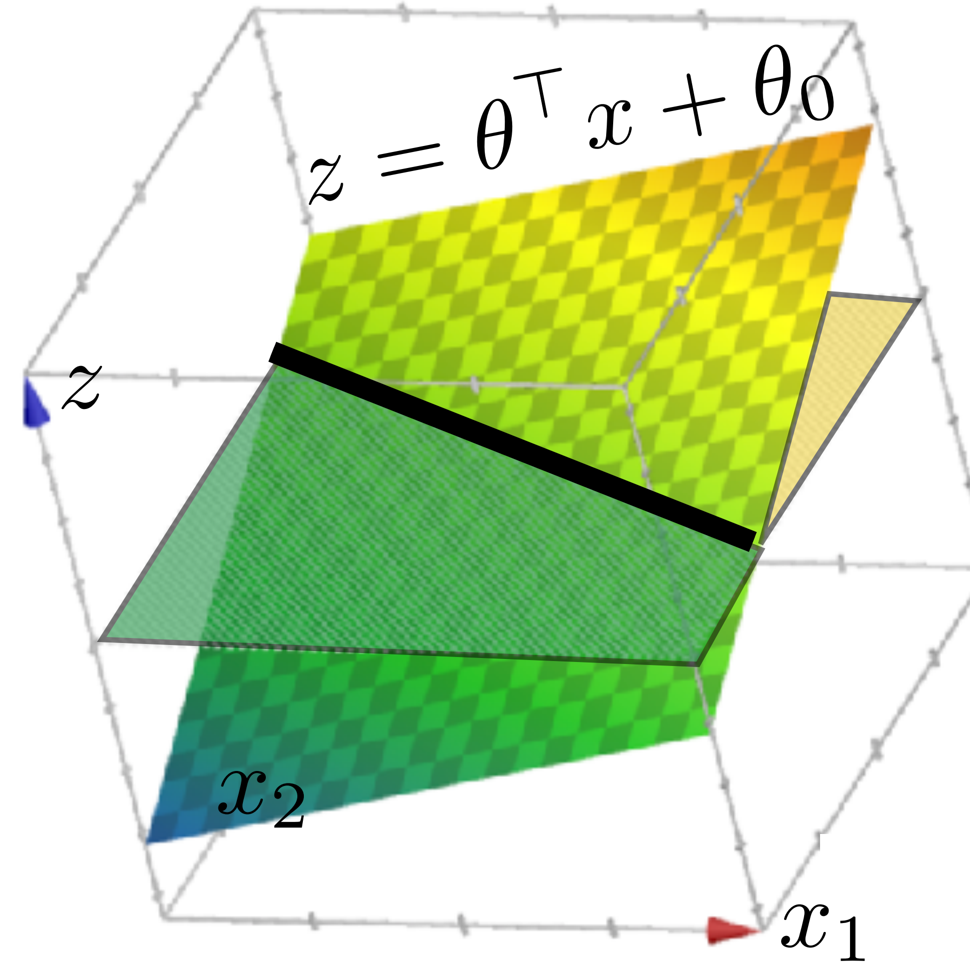
- Linear classification with polynomial features:

$\phi(x)$
$= [x_1, x_2, x_1^2,$
$x_1 x_2, x_2^2]^\top$

Recall

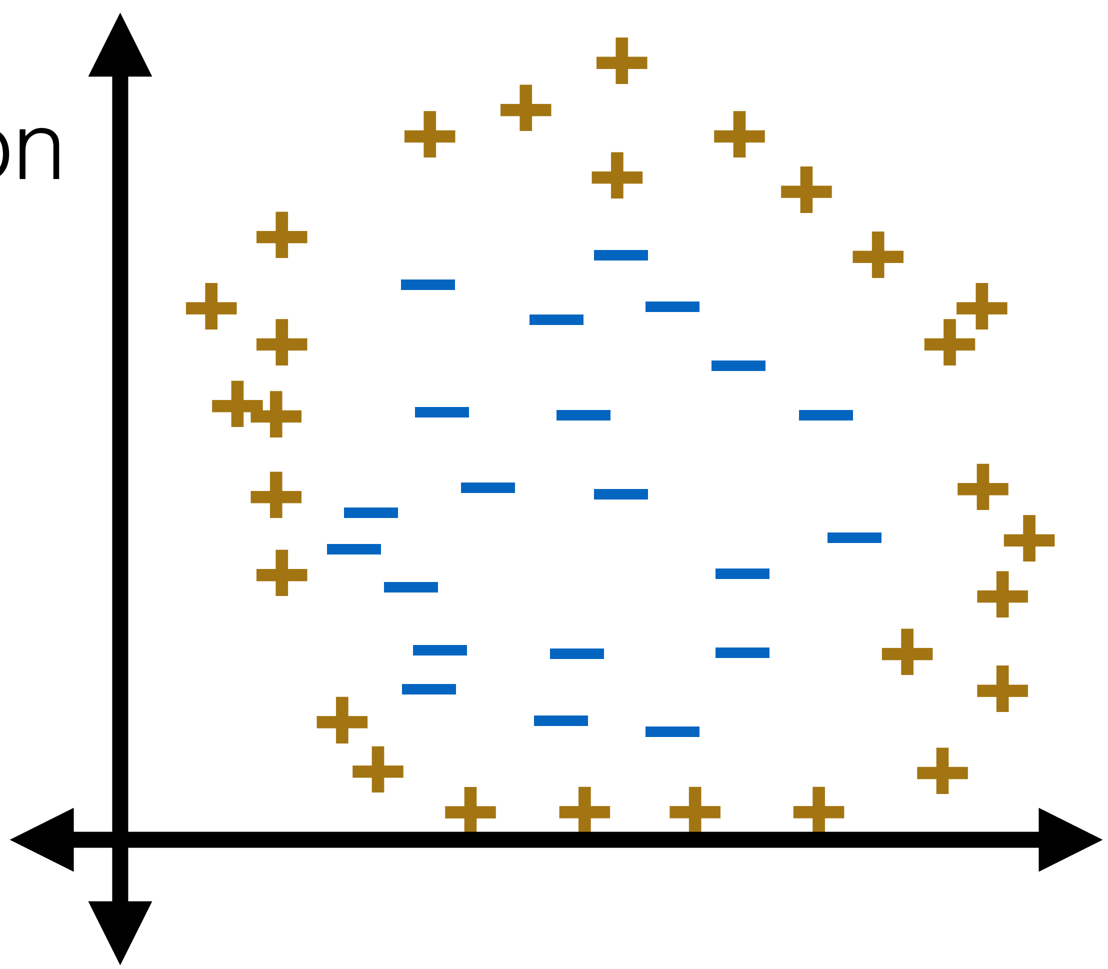
- Linear classification with default features:





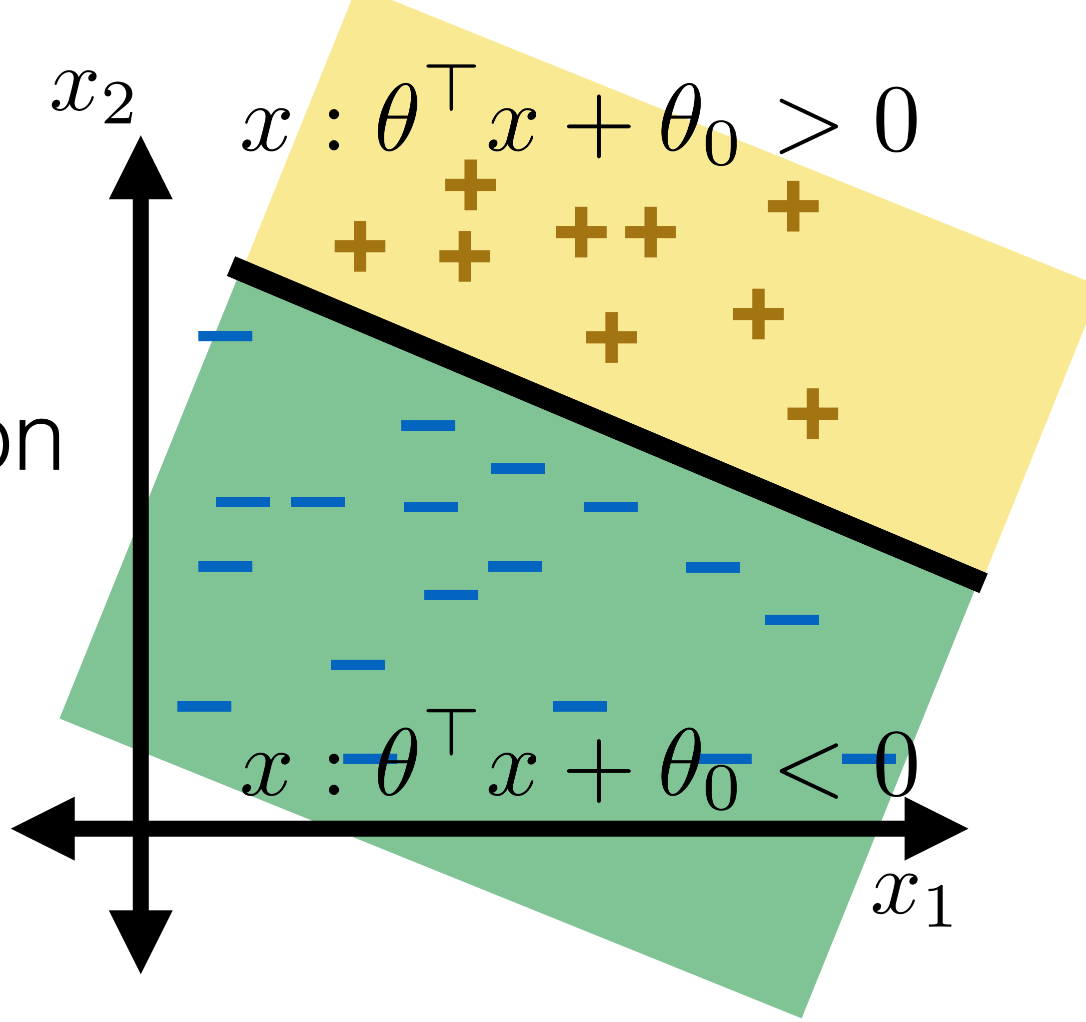
- Linear classification with polynomial features:

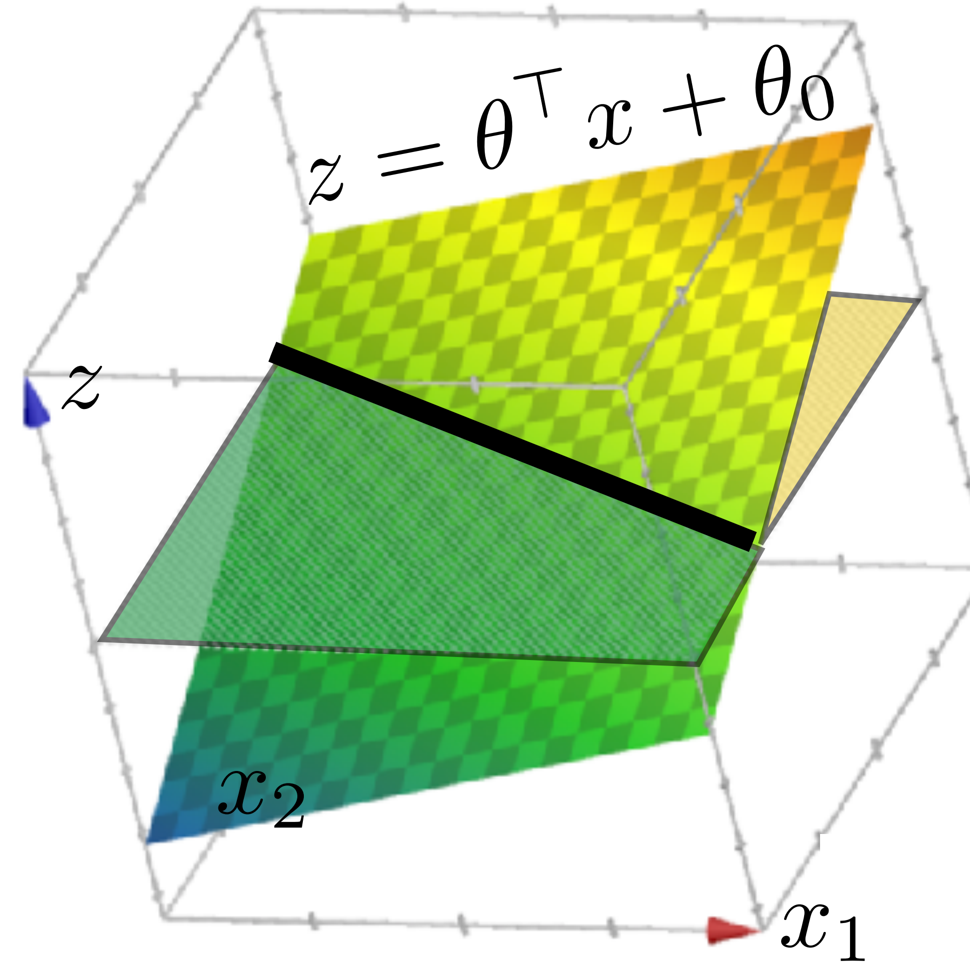
$\phi(x)$
$= [x_1, x_2, x_1^2,$
$x_1 x_2, x_2^2]^\top$



Recall

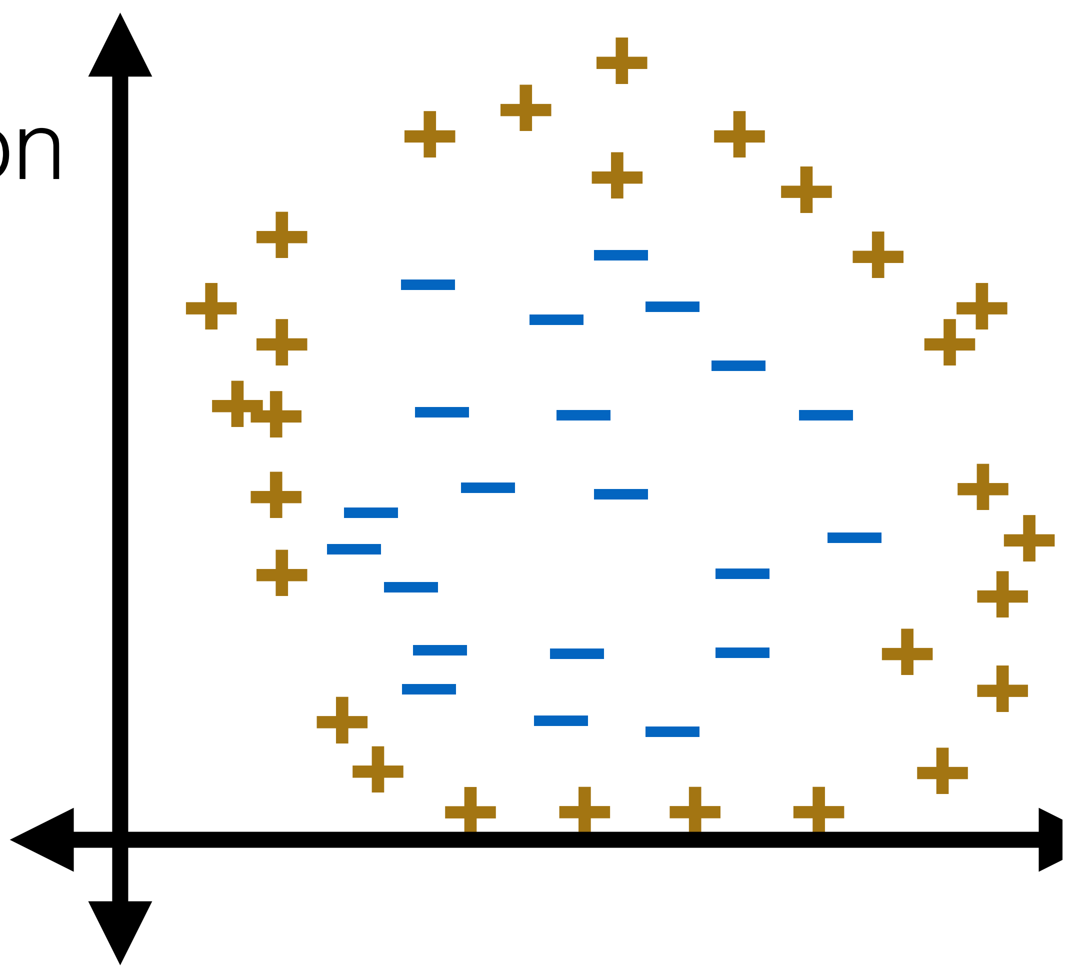
- Linear classification with default features:

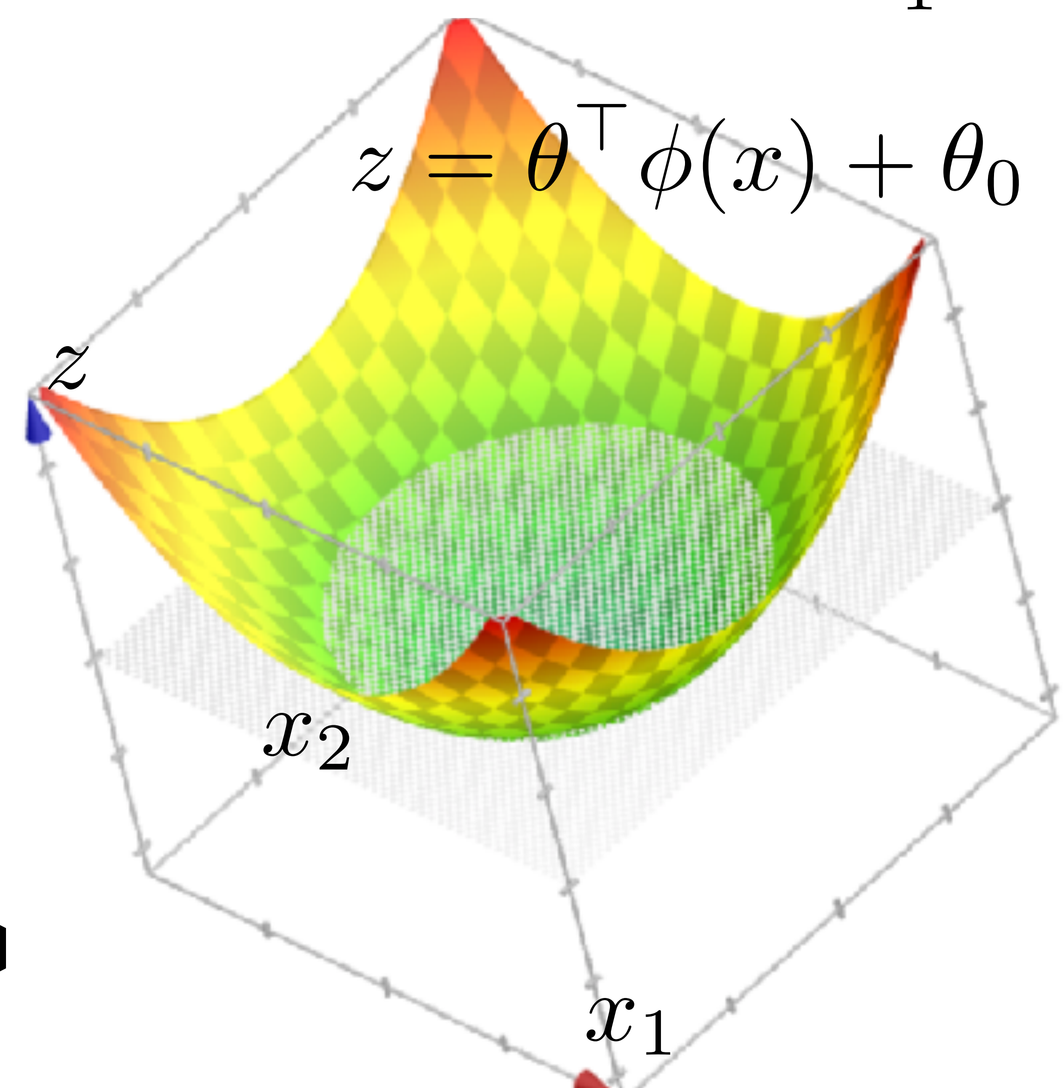




- Linear classification with polynomial features:

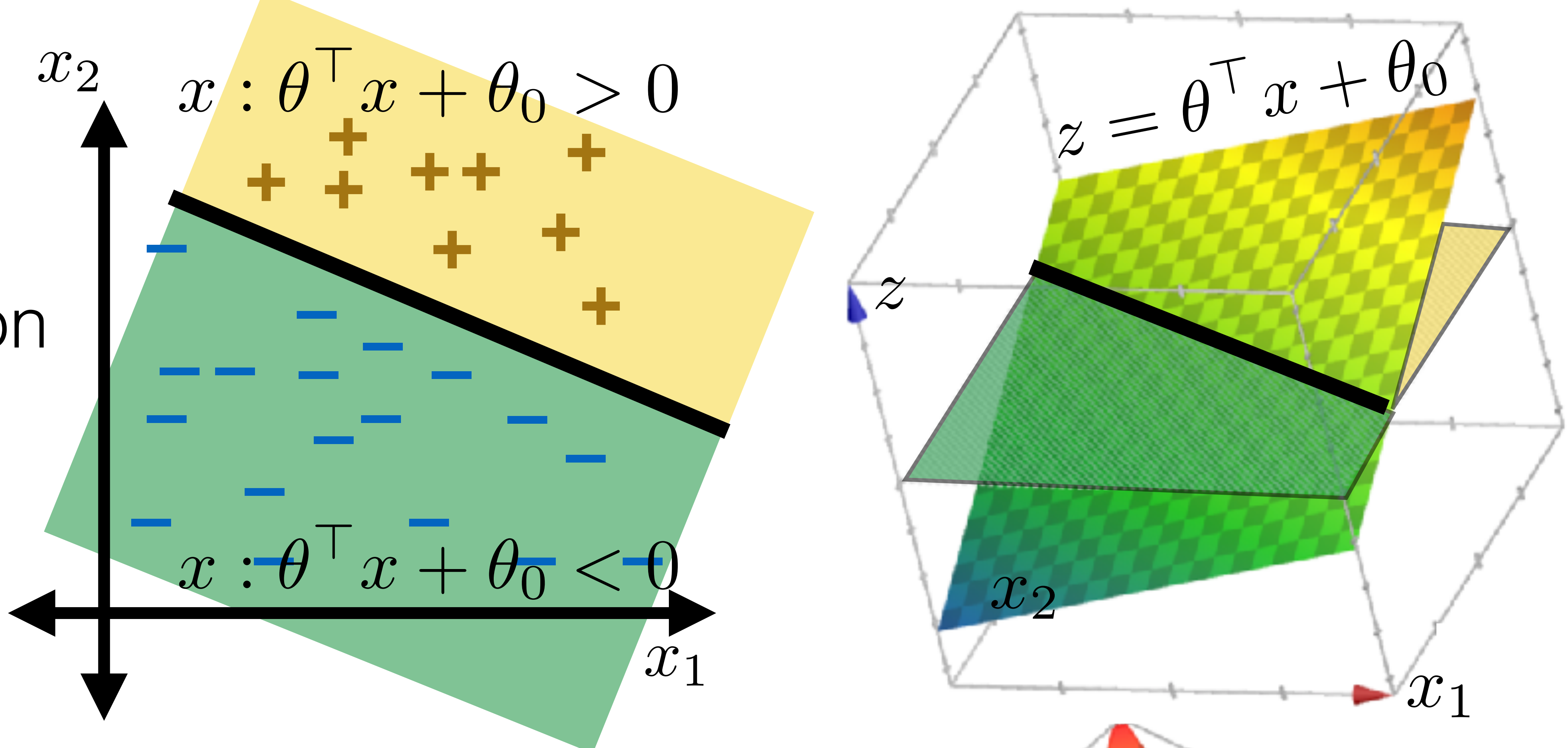
$\phi(x)$
$= [x_1, x_2, x_1^2,$
$x_1 x_2, x_2^2]^\top$





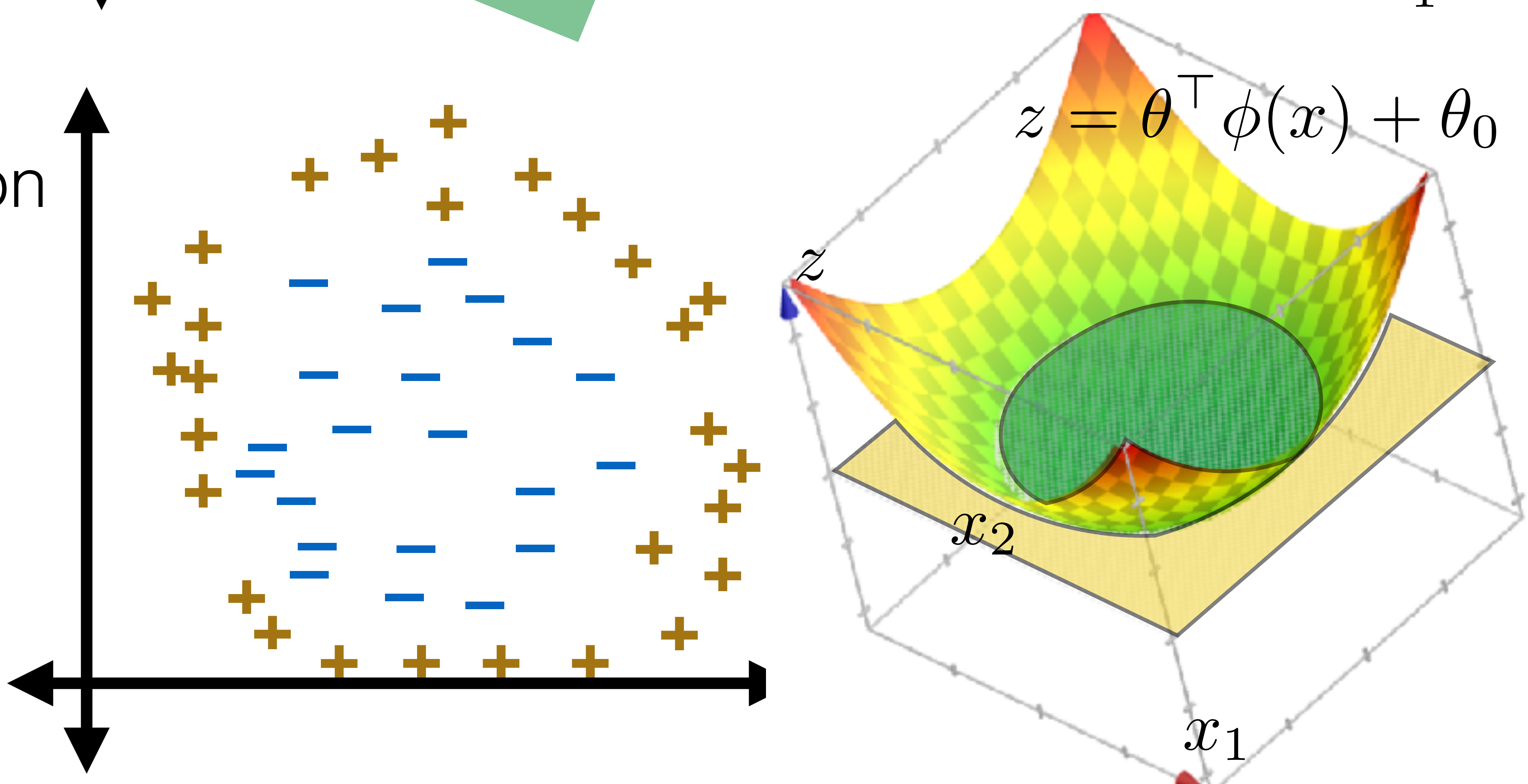
Recall

- Linear classification with default features:



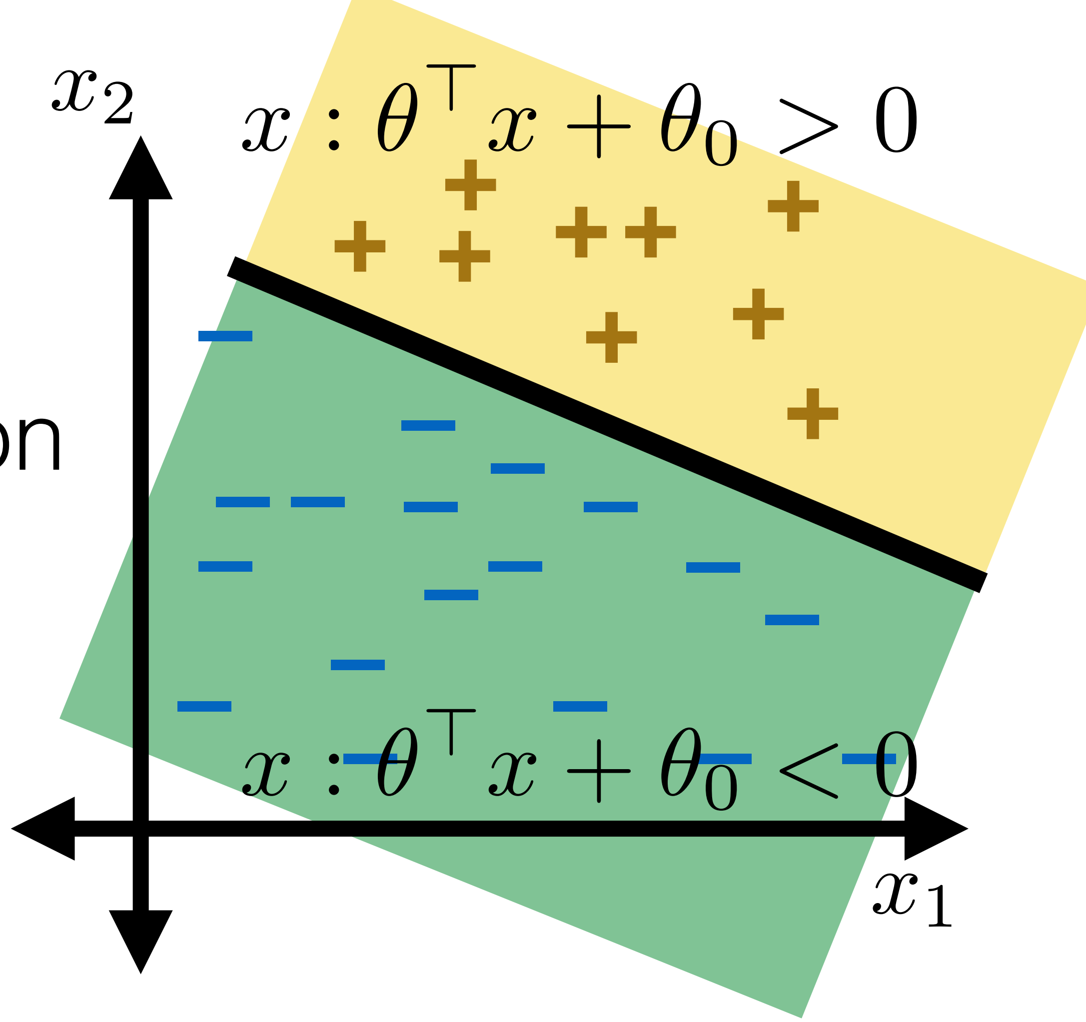
- Linear classification with polynomial features:

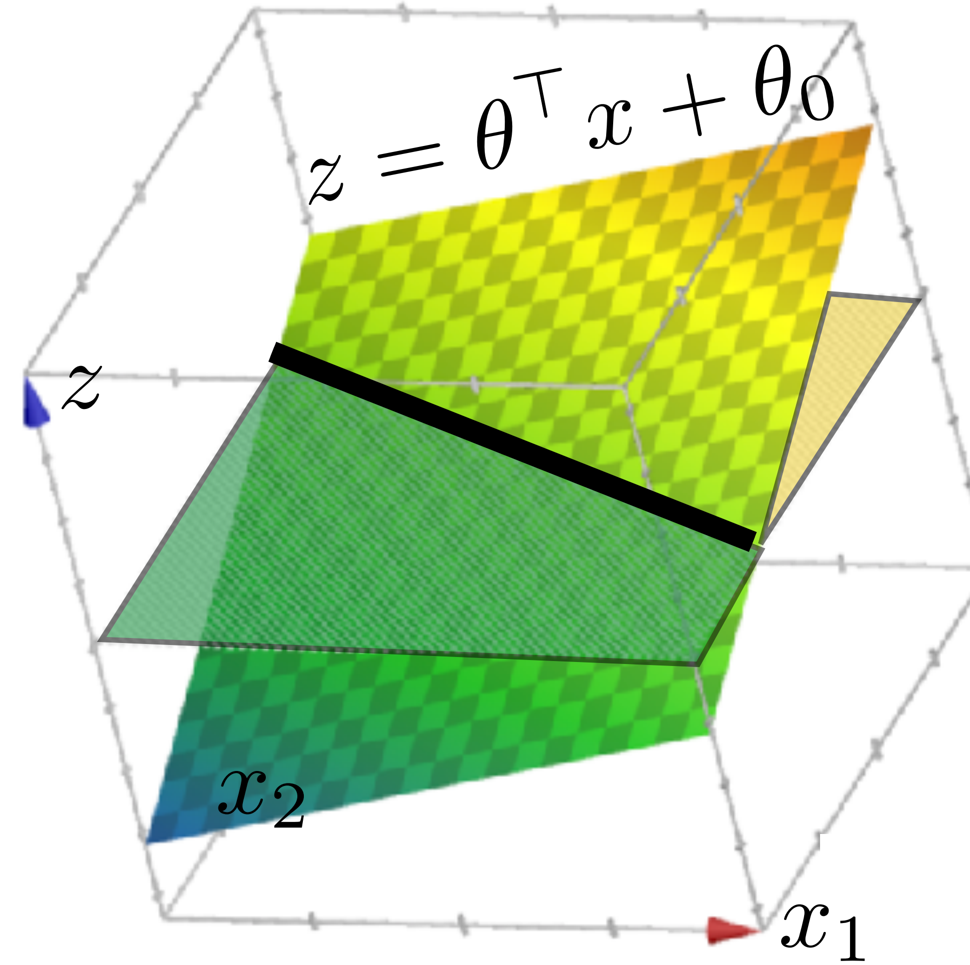
$\phi(x)$
$= [x_1, x_2, x_1^2,$
$x_1 x_2, x_2^2]^\top$



Recall

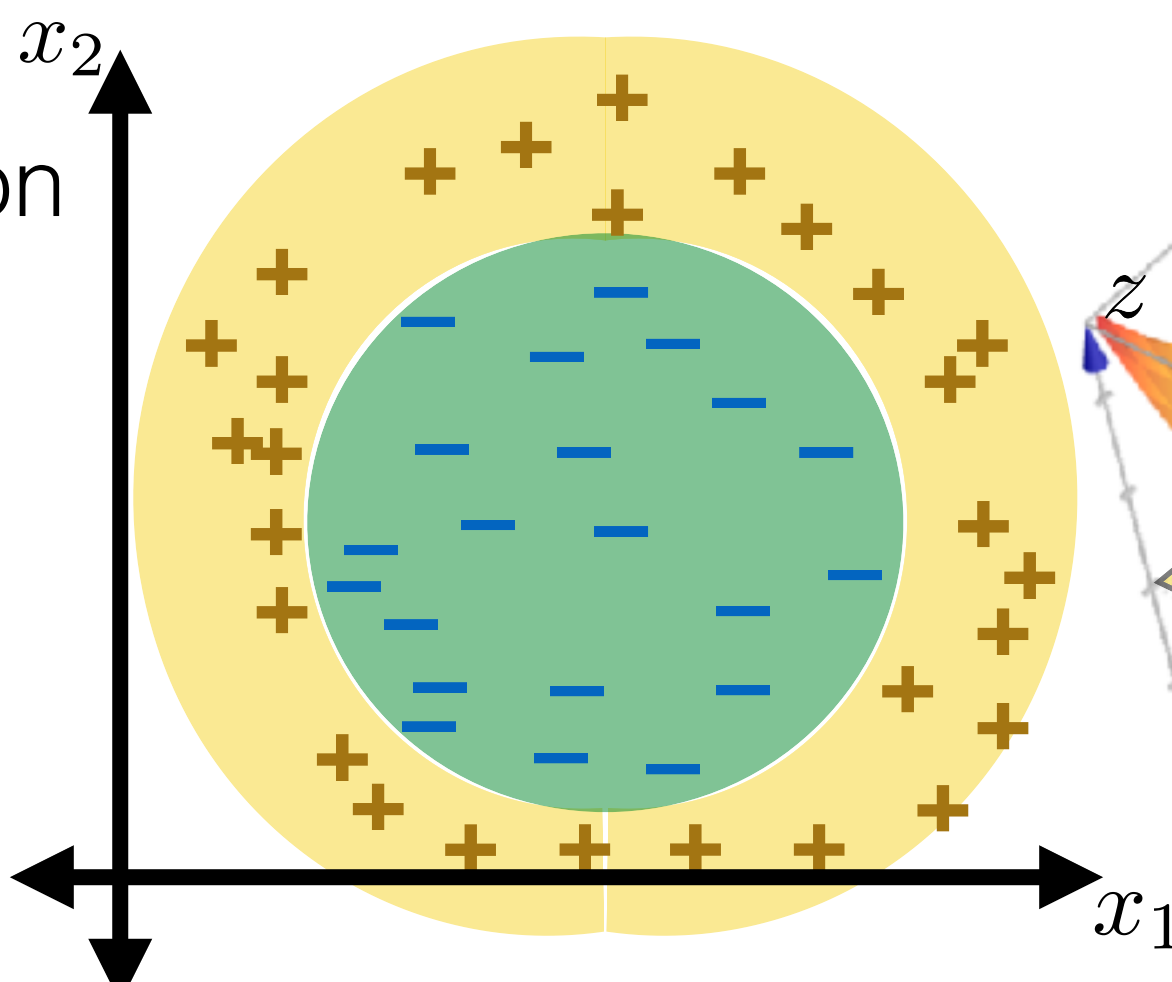
- Linear classification with default features:

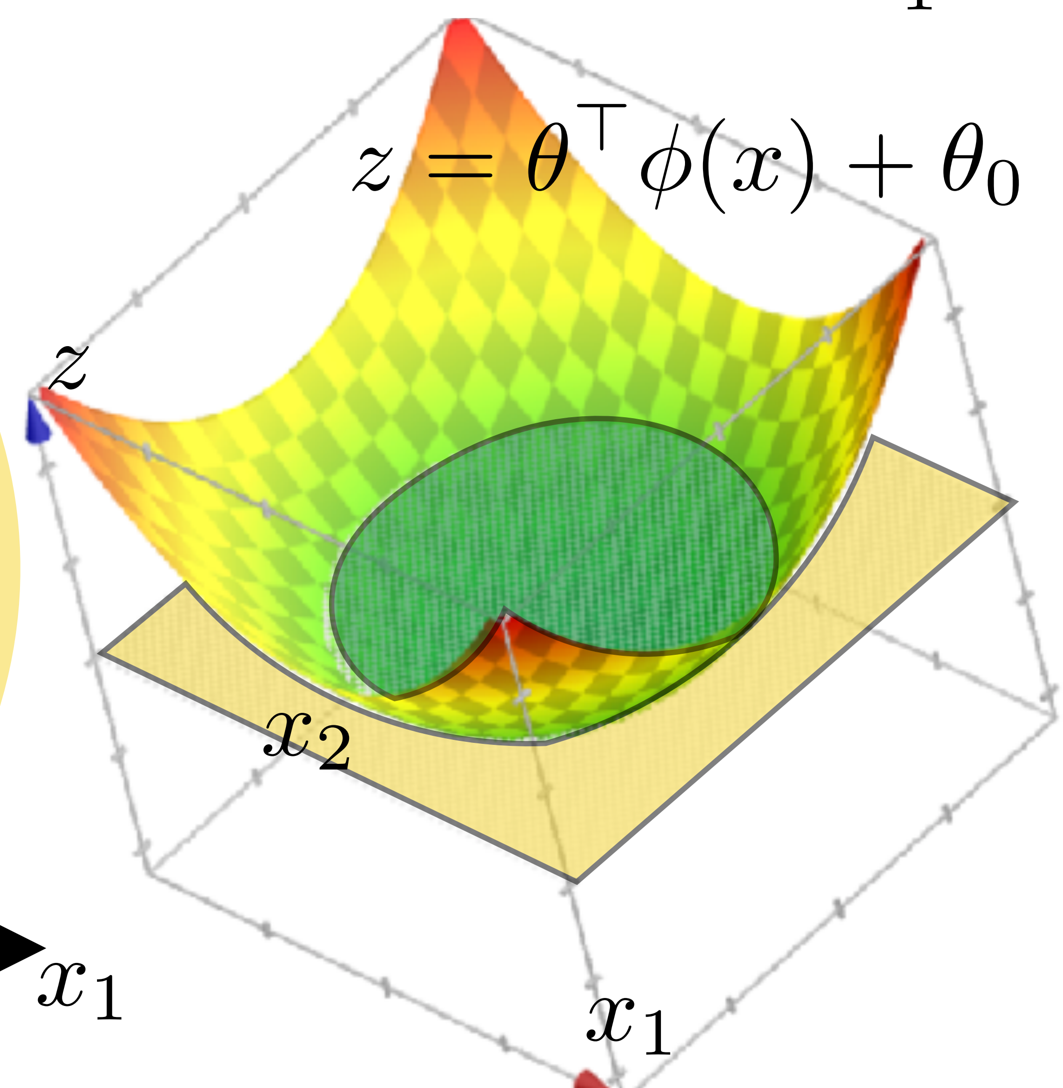




- Linear classification with polynomial features:

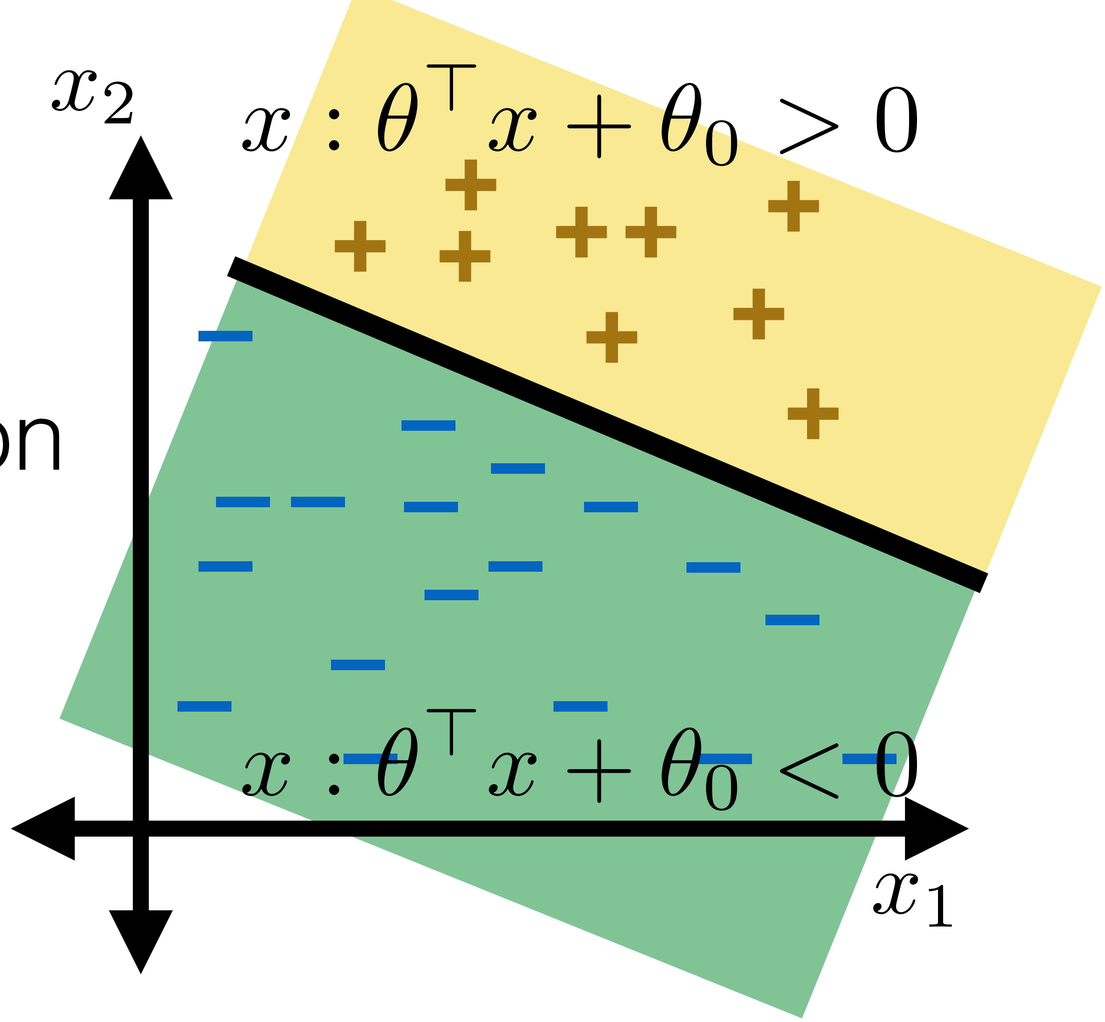
$\phi(x)$
$= [x_1, x_2, x_1^2,$
$x_1 x_2, x_2^2]^\top$

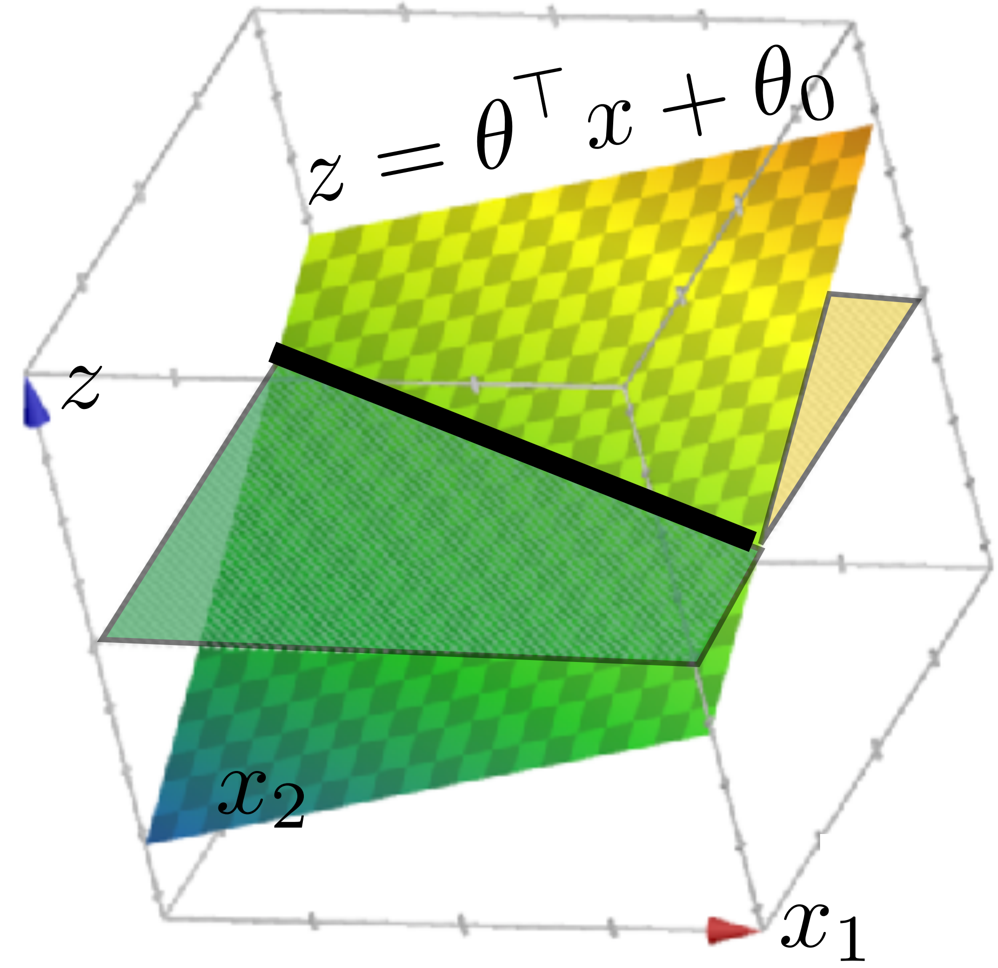




Recall

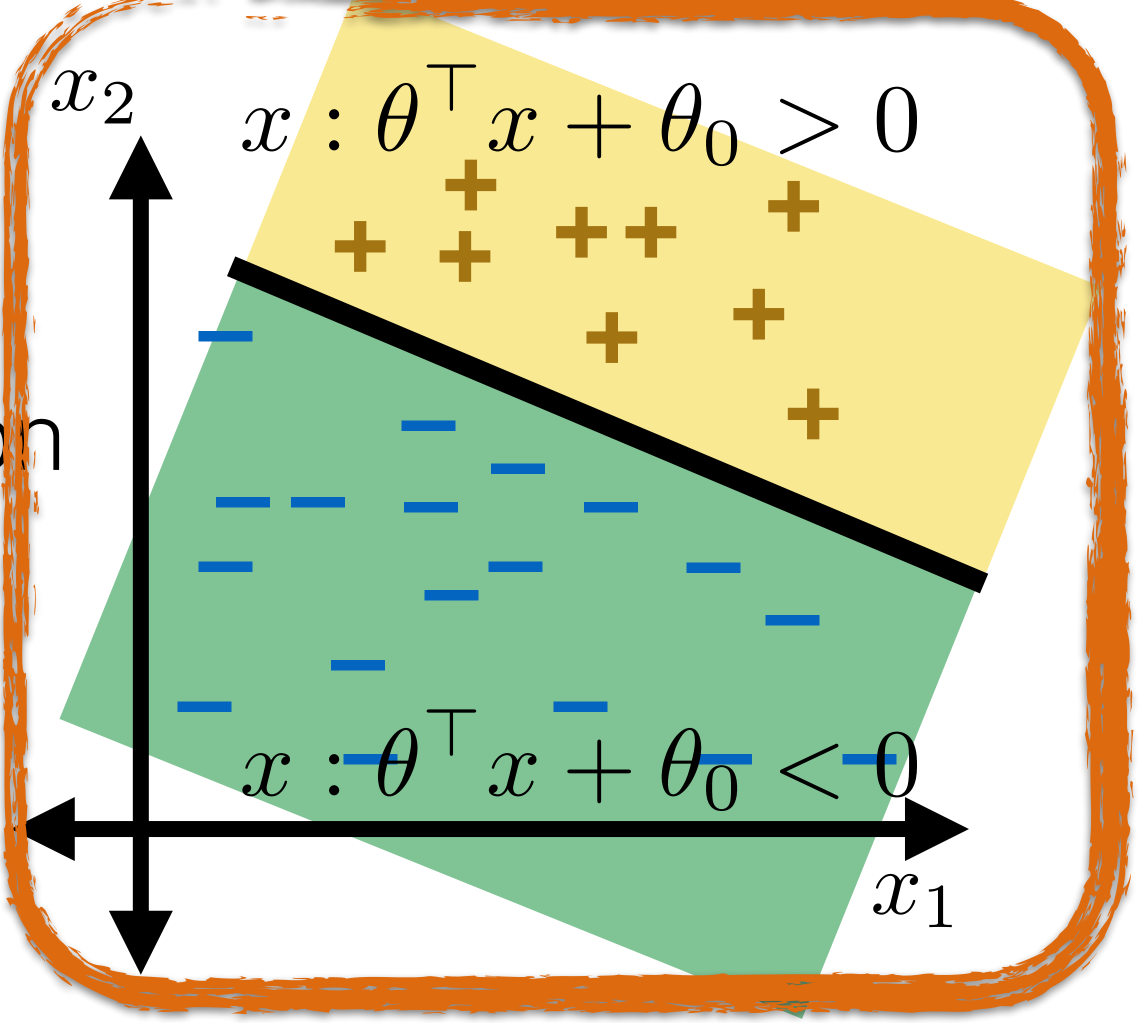
- Linear classification with default features:

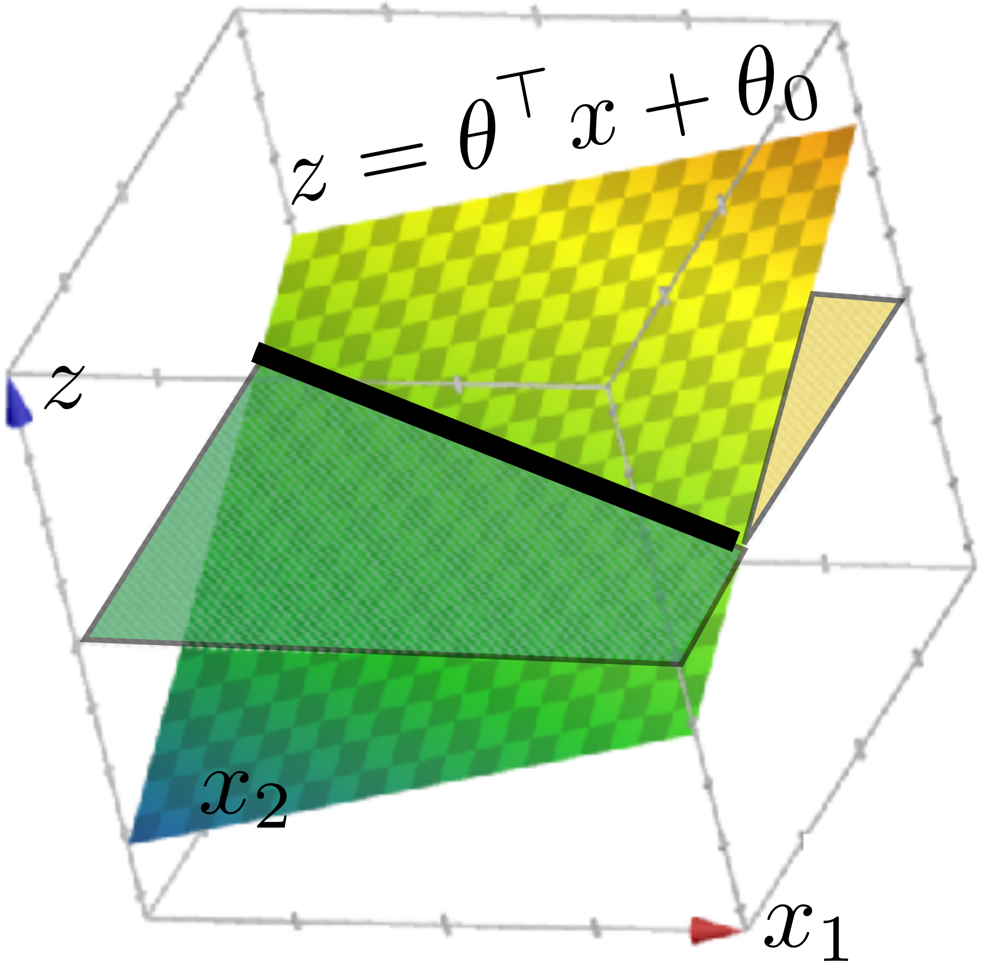




Recall

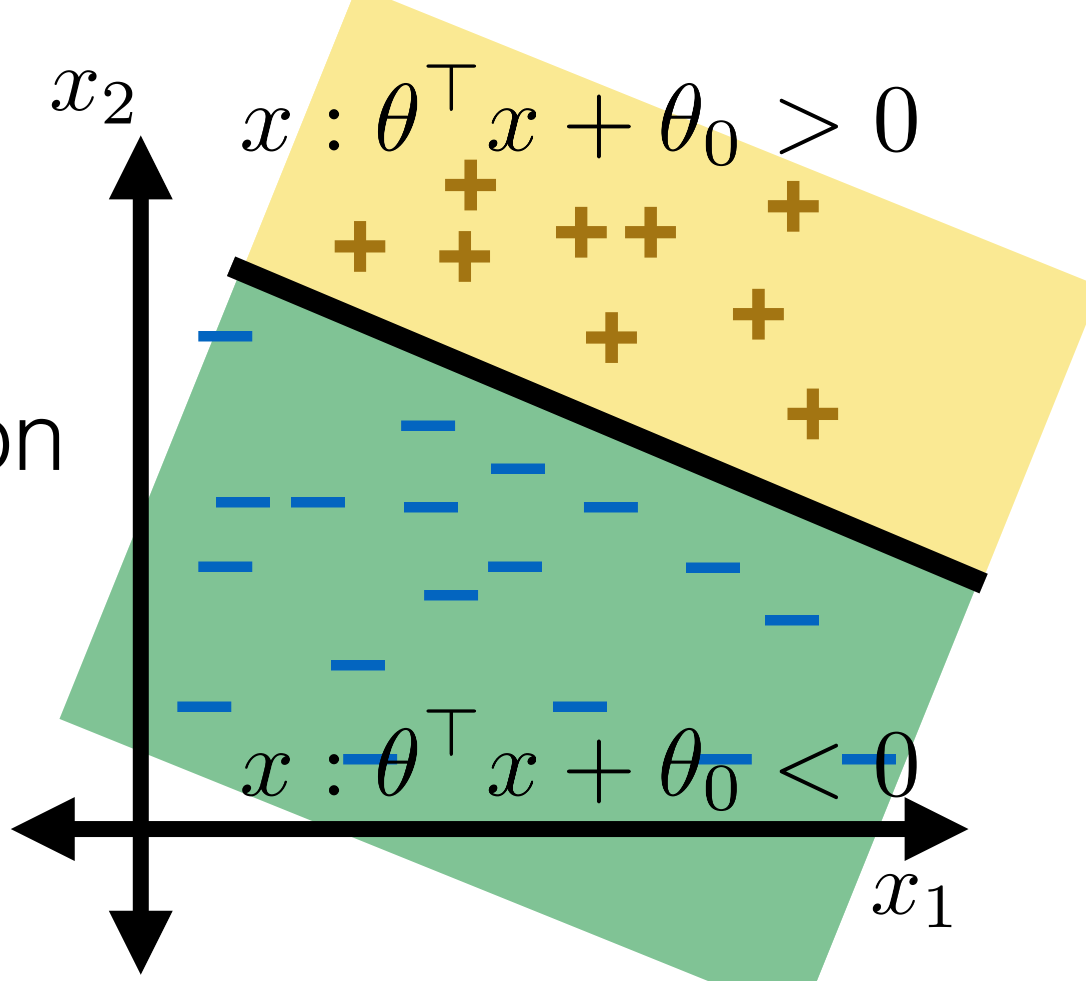
- Linear classification with default features:

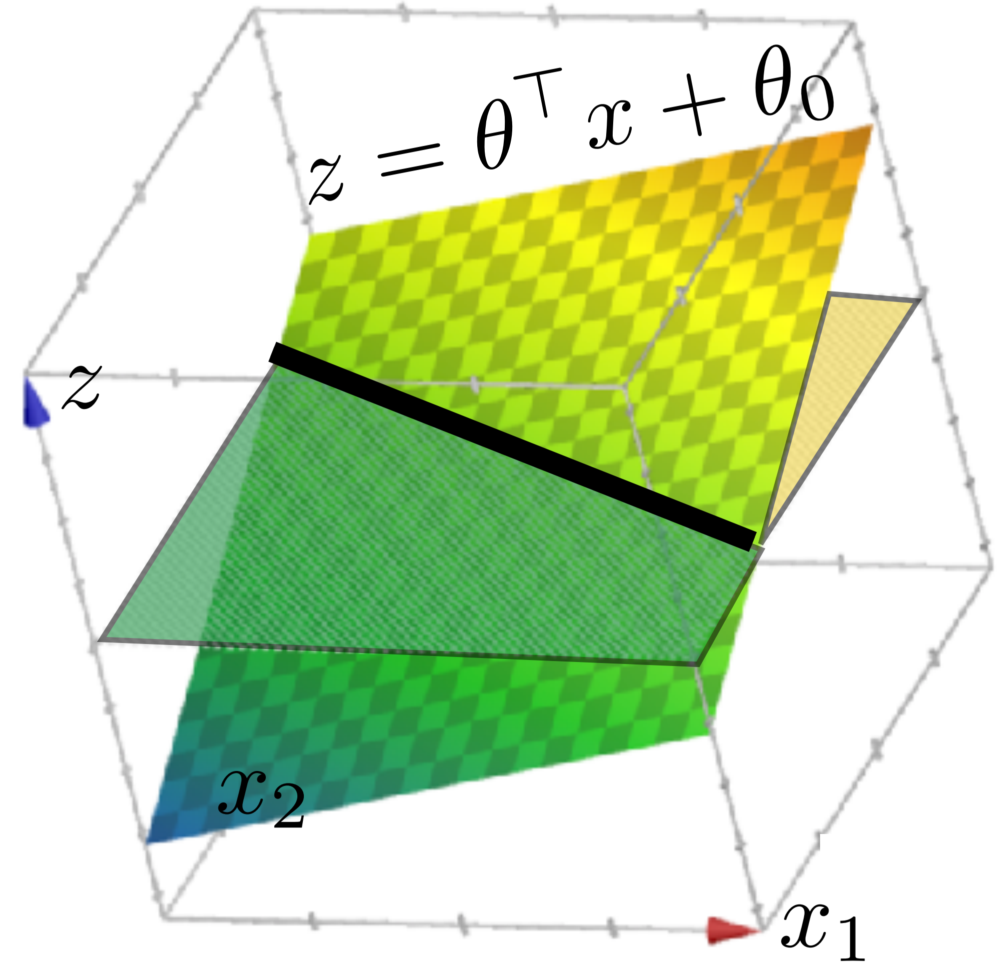


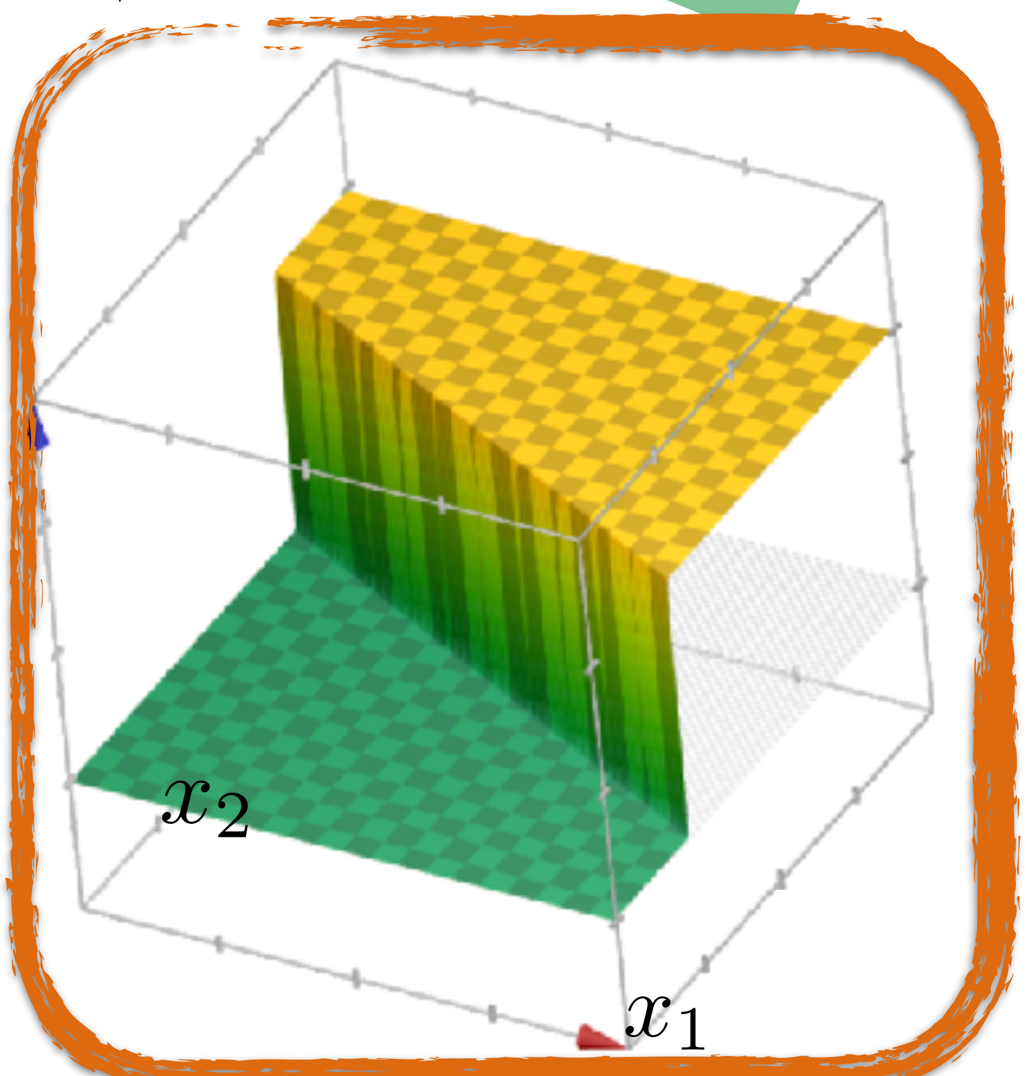


Recall

- Linear classification with default features:

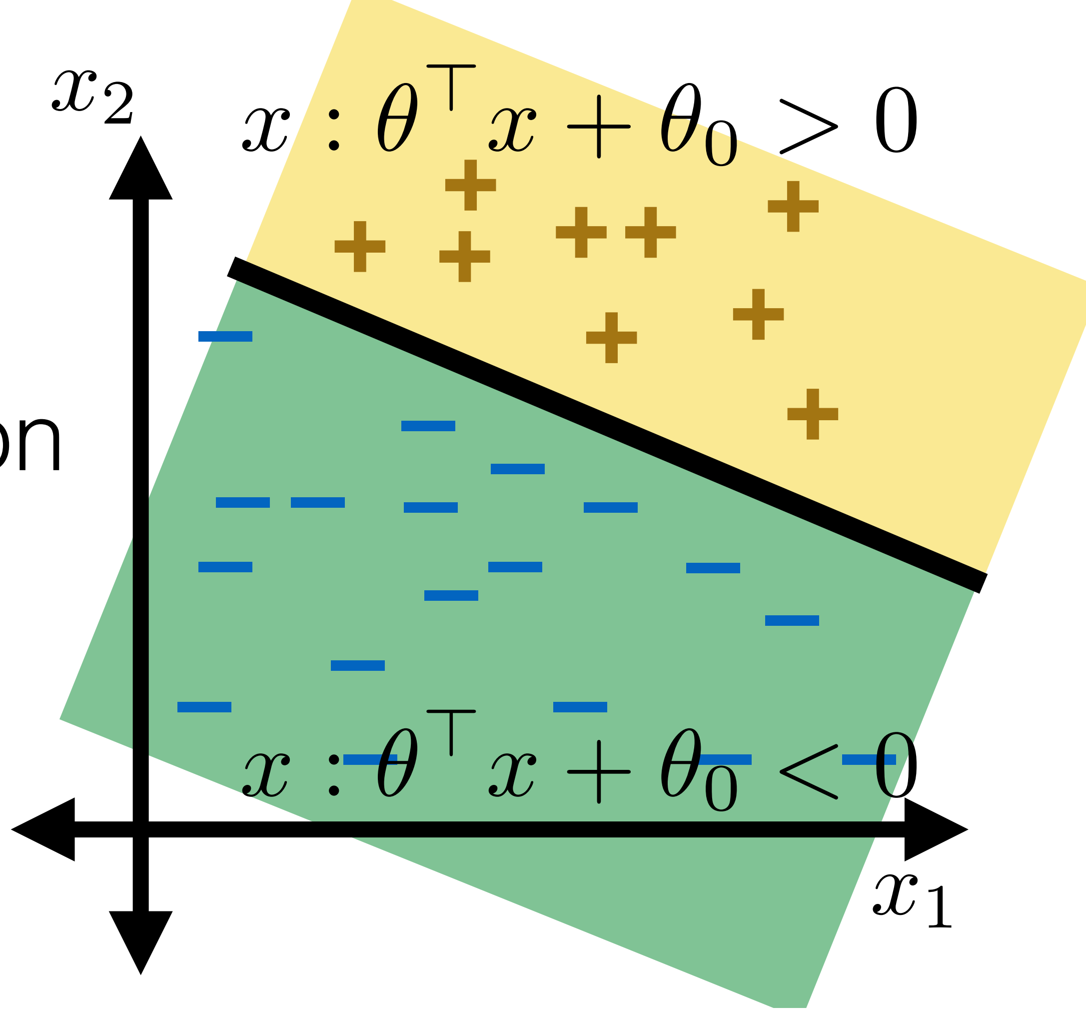


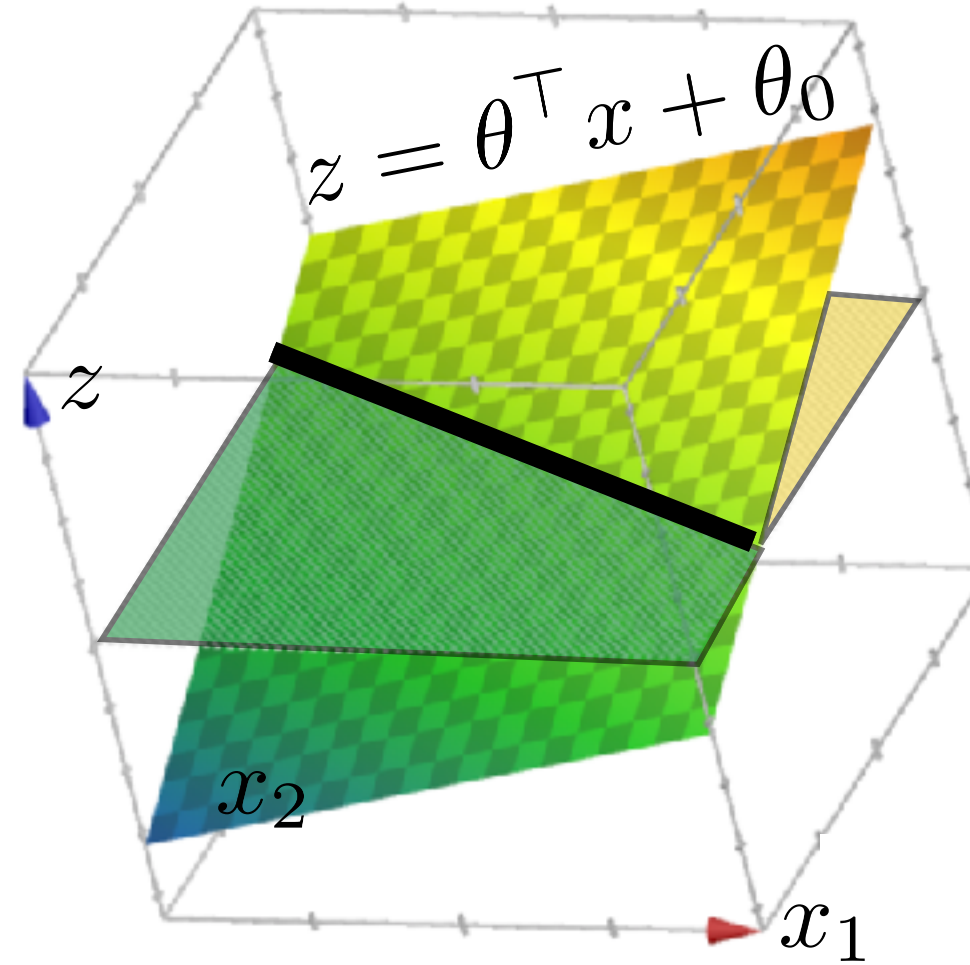


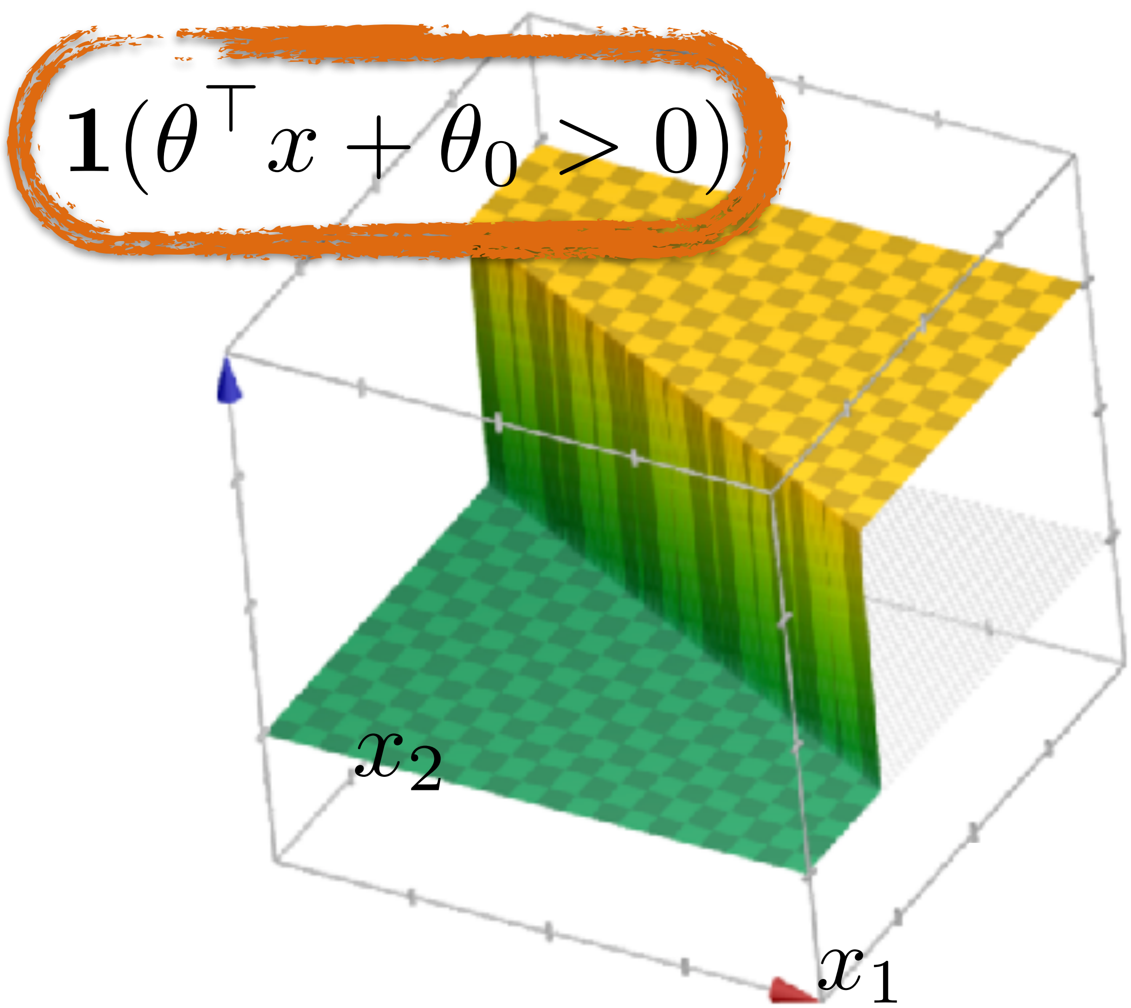


Recall

- Linear classification with default features:

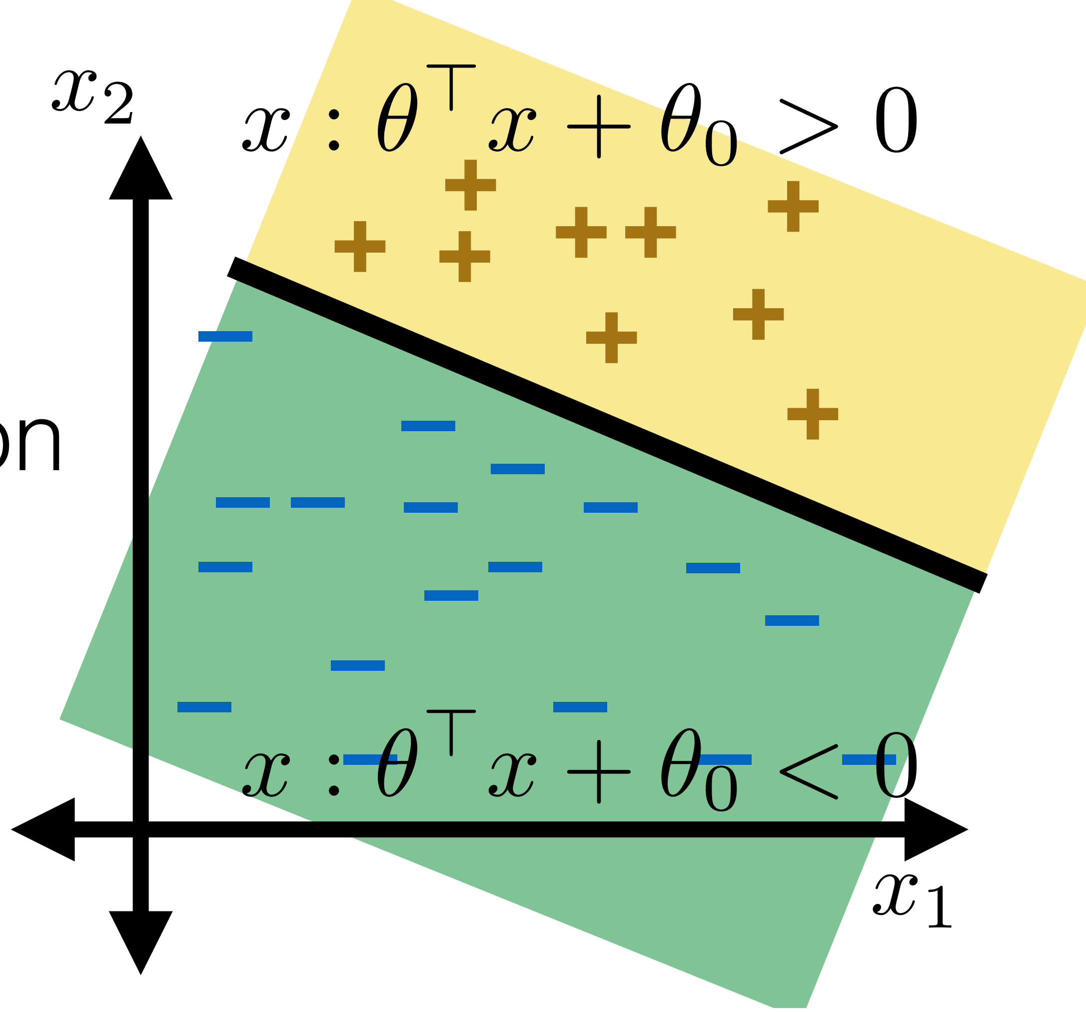


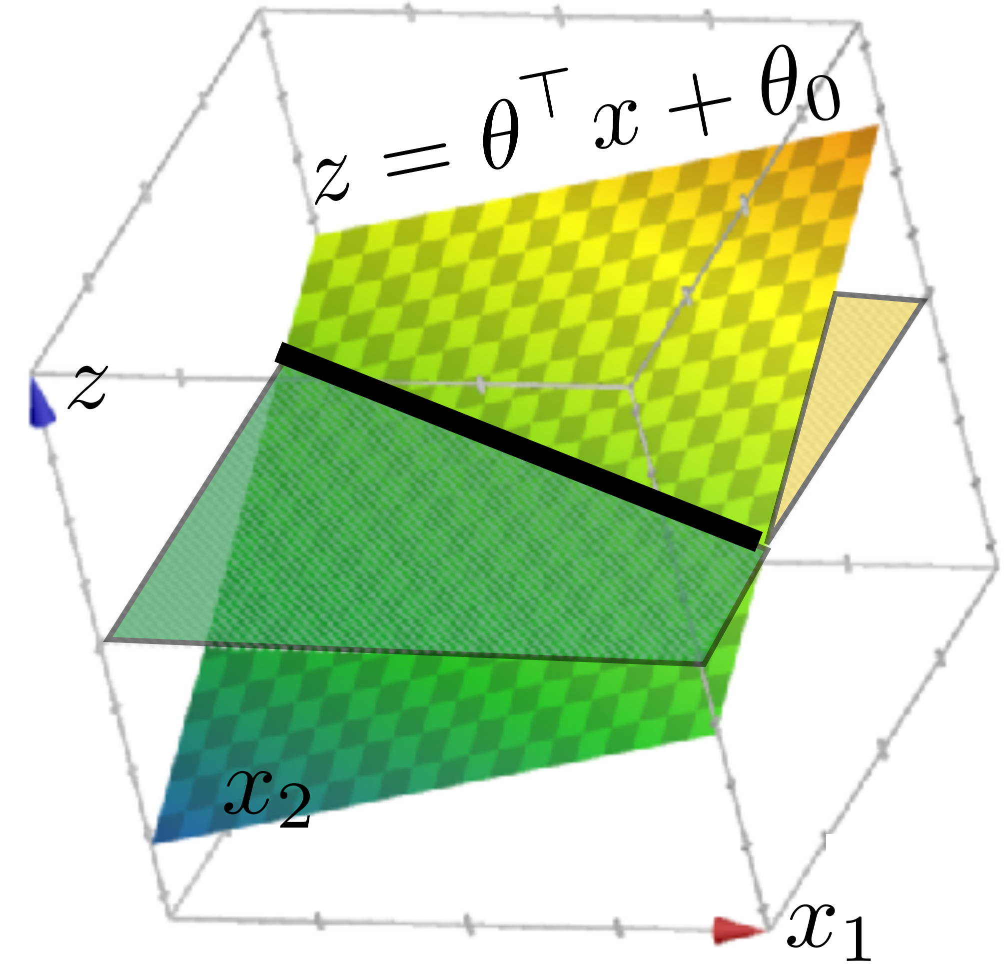


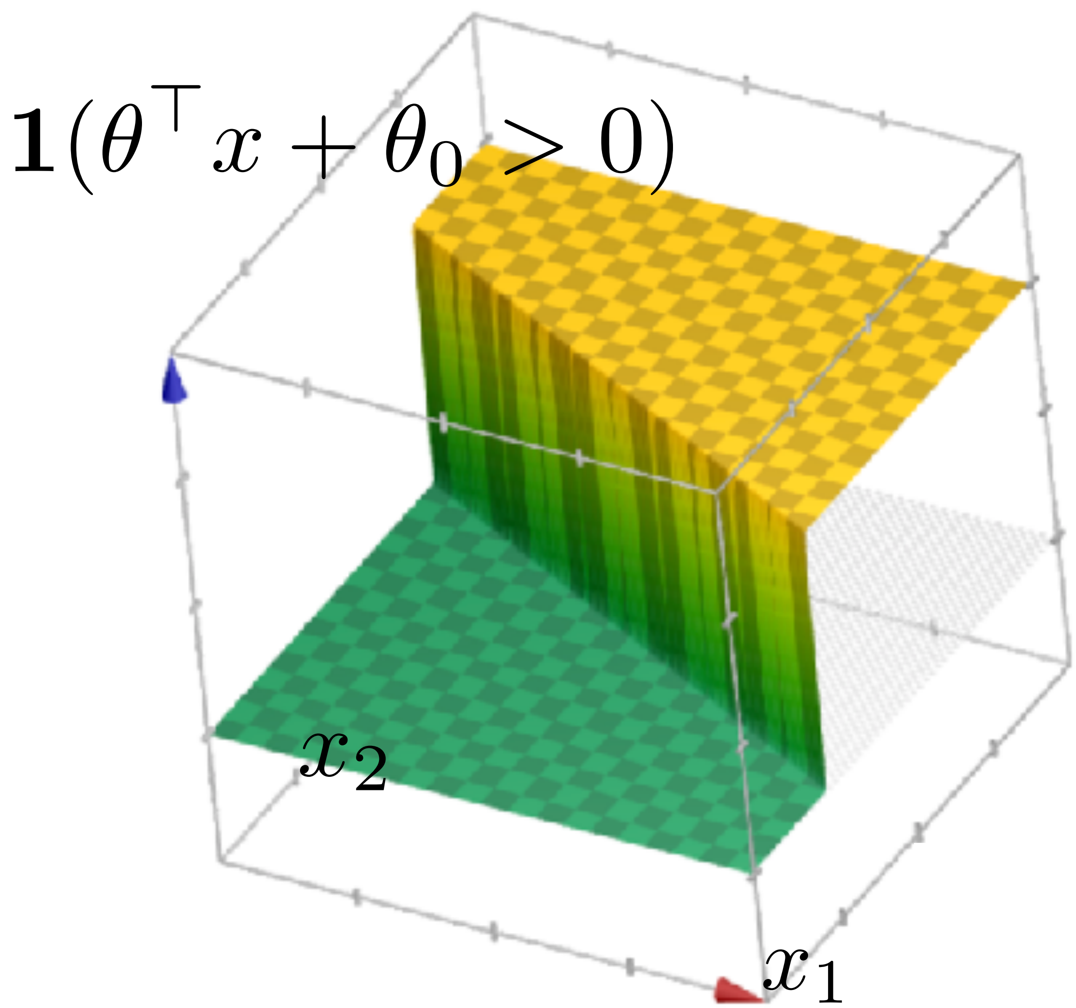


Recall

- Linear classification with default features:



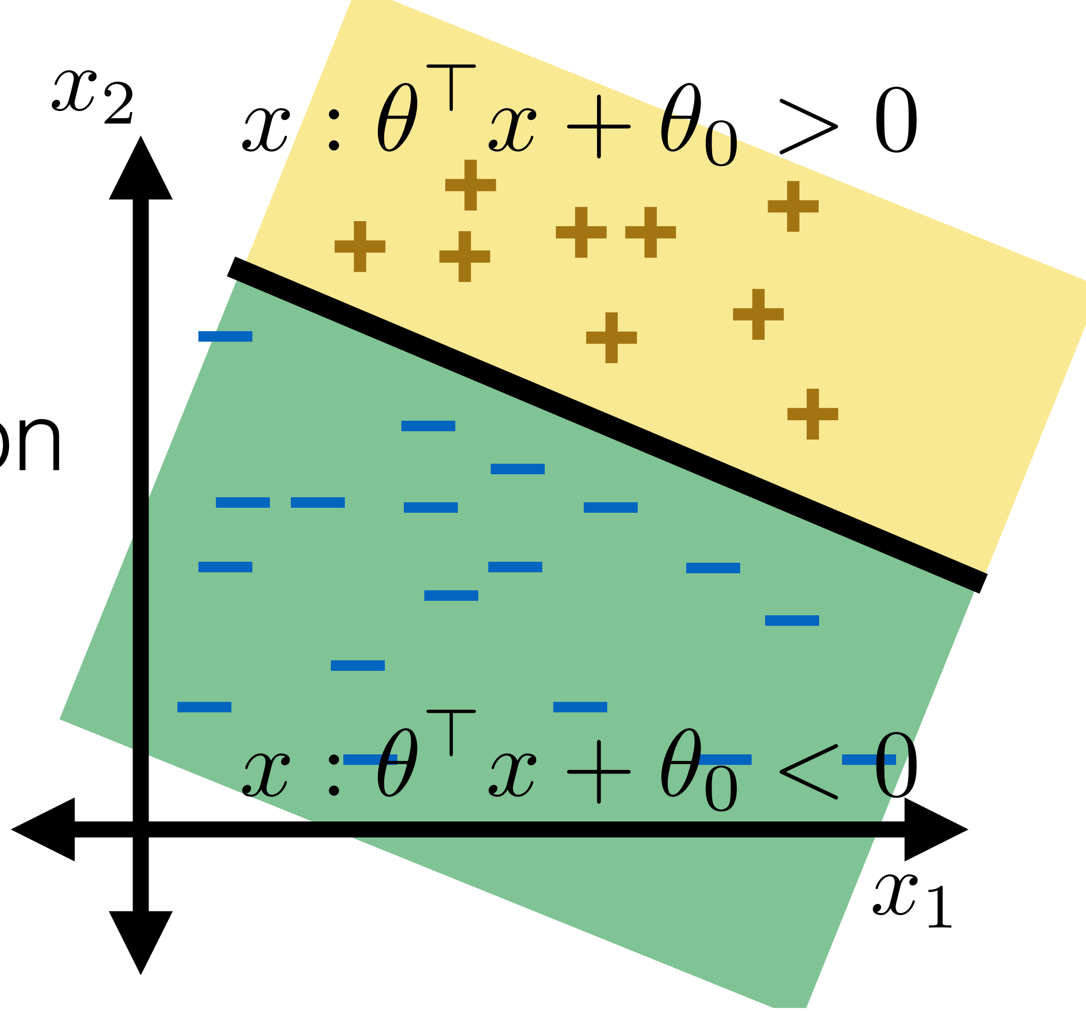


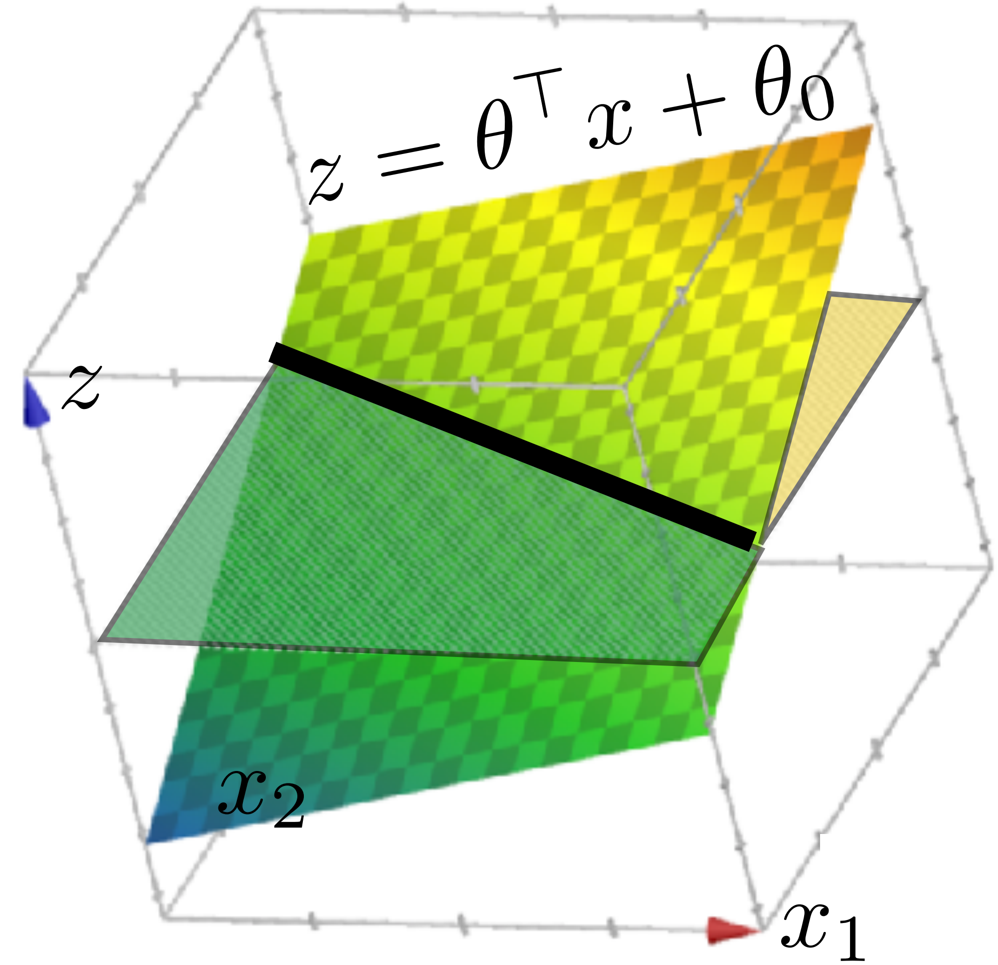


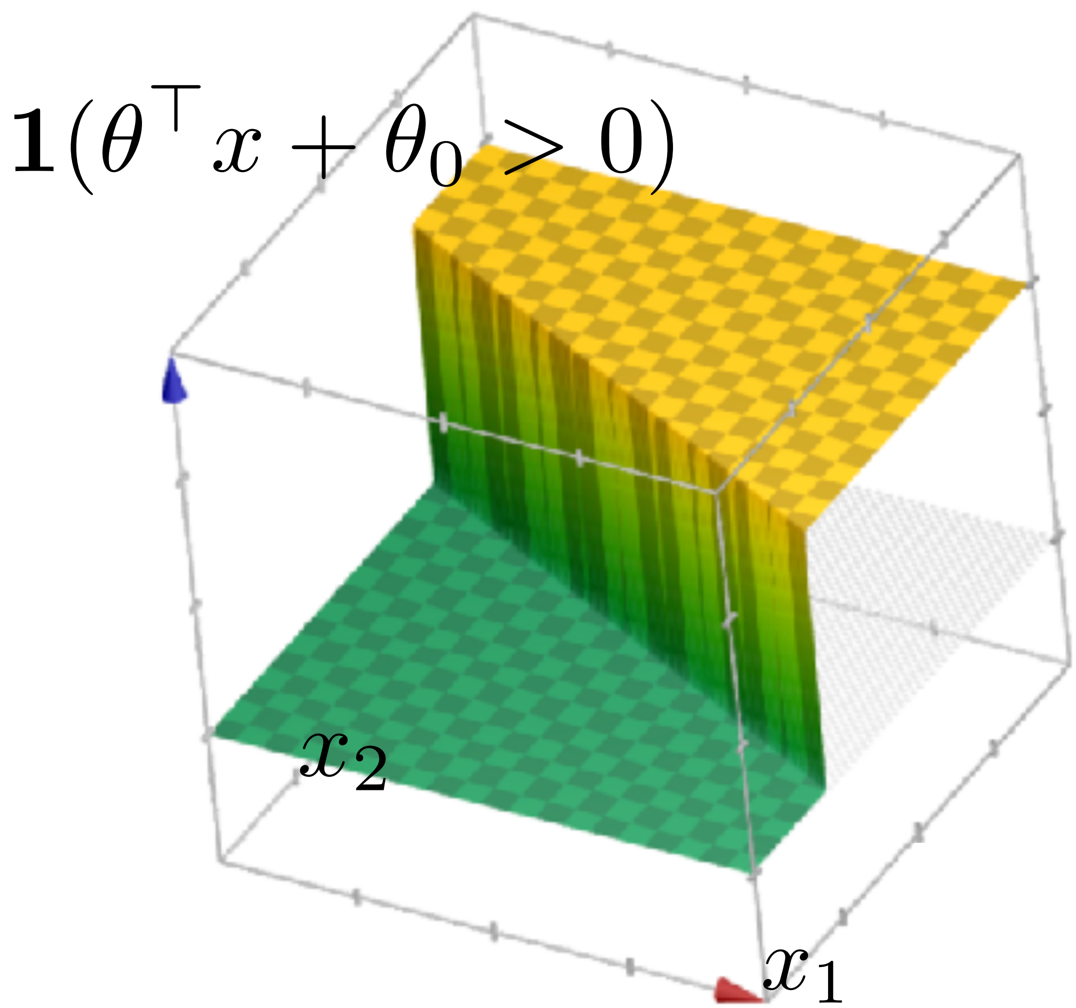
- We're used to using step functions to classify

Recall

- Linear classification with default features:







- We're used to using step functions to classify
- New idea today: we'll use step functions as *features*, with their own parameters

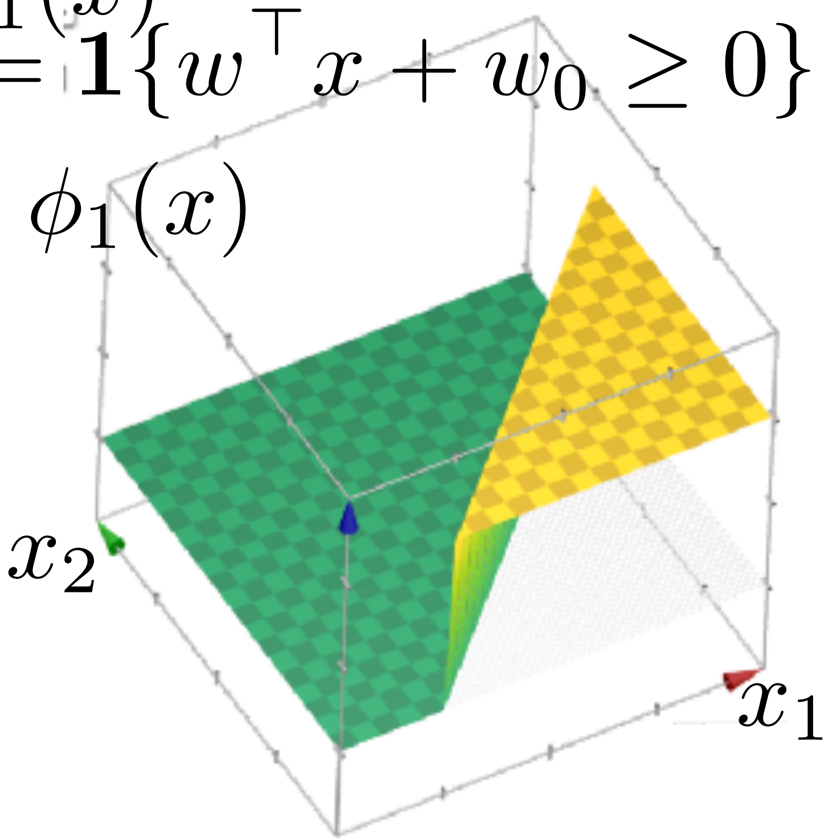
New features: step functions!

New features: step functions!

$$\phi_1(x) = \mathbf{1}\{w^\top x + w_0 \geq 0\}$$

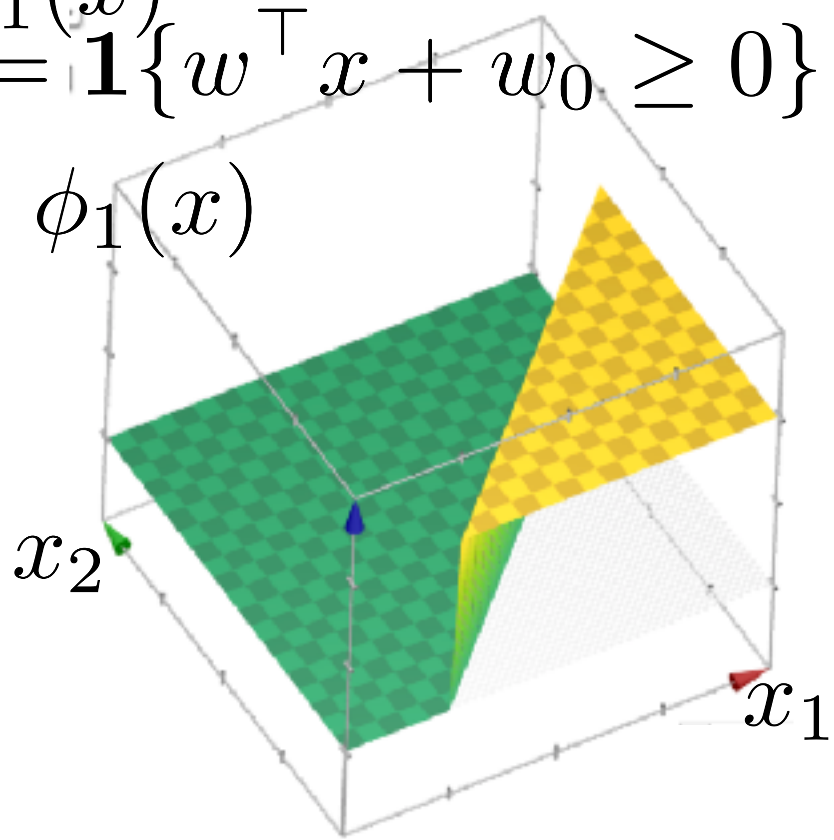
New features: step functions!

$\phi_1(x)$
$= \mathbf{1}\{w^\top x + w_0 \geq 0\}$



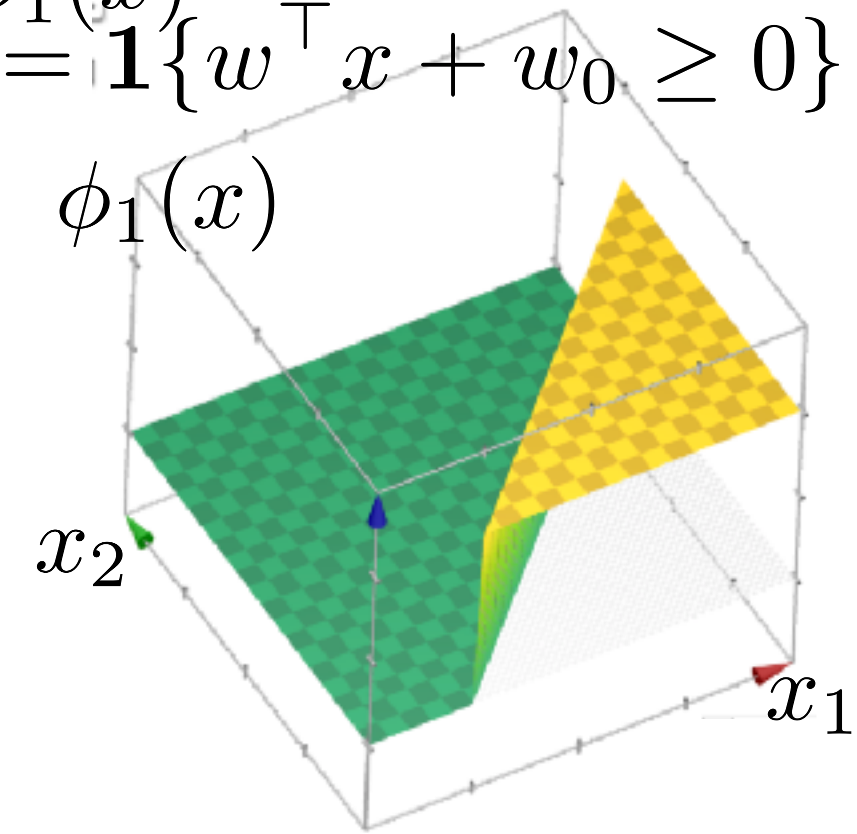
New features: step functions!

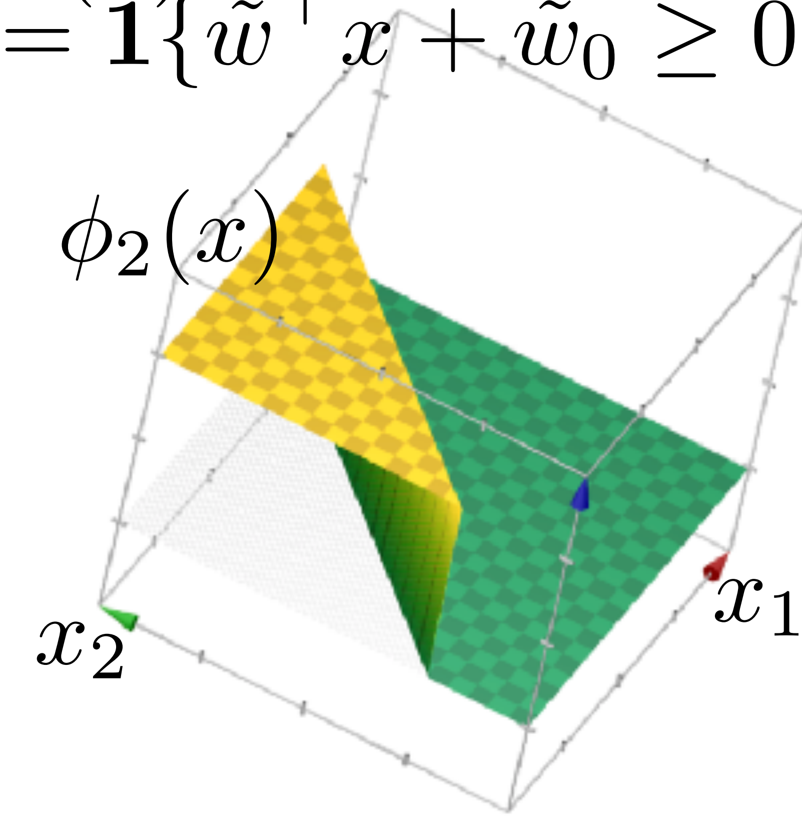
$\phi_1(x) = \mathbf{1}\{w^\top x + w_0 \geq 0\}$ $\phi_2(x) = \mathbf{1}\{\tilde{w}^\top x + \tilde{w}_0 \geq 0\}$



New features: step functions!

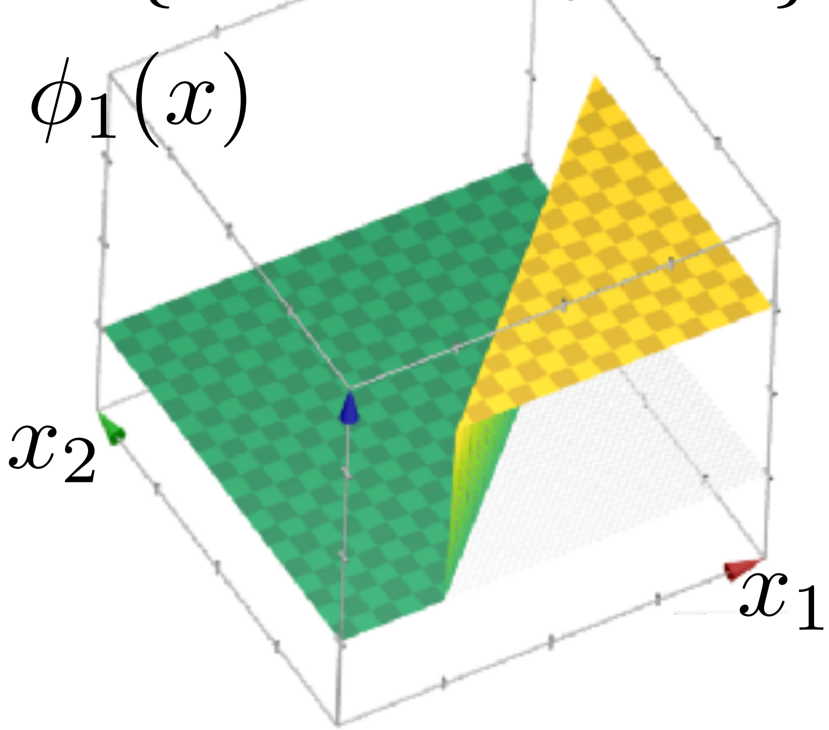
$\phi_1(x) = \mathbf{1}\{w^\top x + w_0 \geq 0\}$ $\phi_2(x) = \mathbf{1}\{\tilde{w}^\top x + \tilde{w}_0 \geq 0\}$

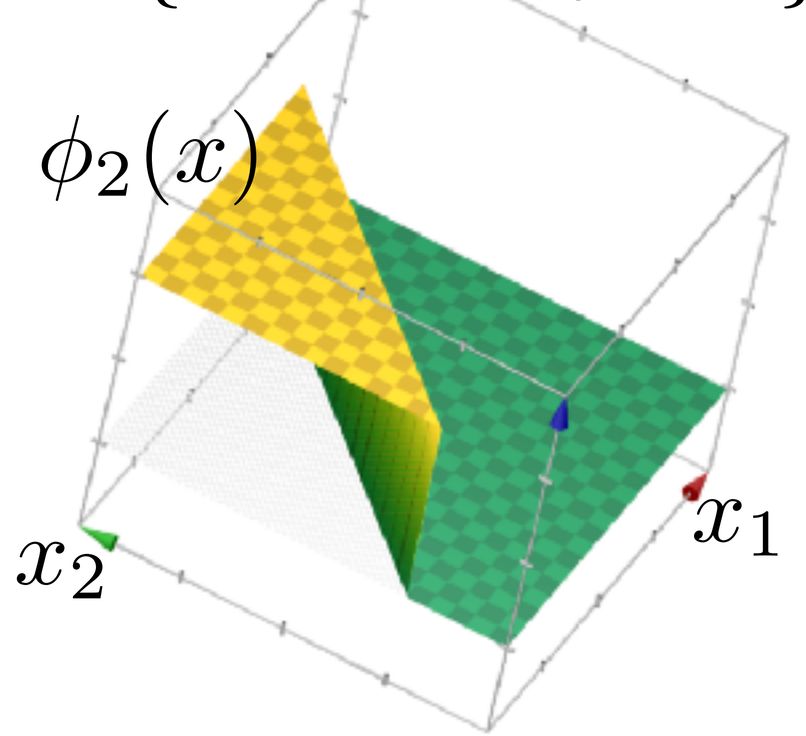




New features: step functions!

$\phi_1(x) = \mathbf{1}\{w^\top x + w_0 \geq 0\}$ $\phi_2(x) = \mathbf{1}\{\tilde{w}^\top x + \tilde{w}_0 \geq 0\}$

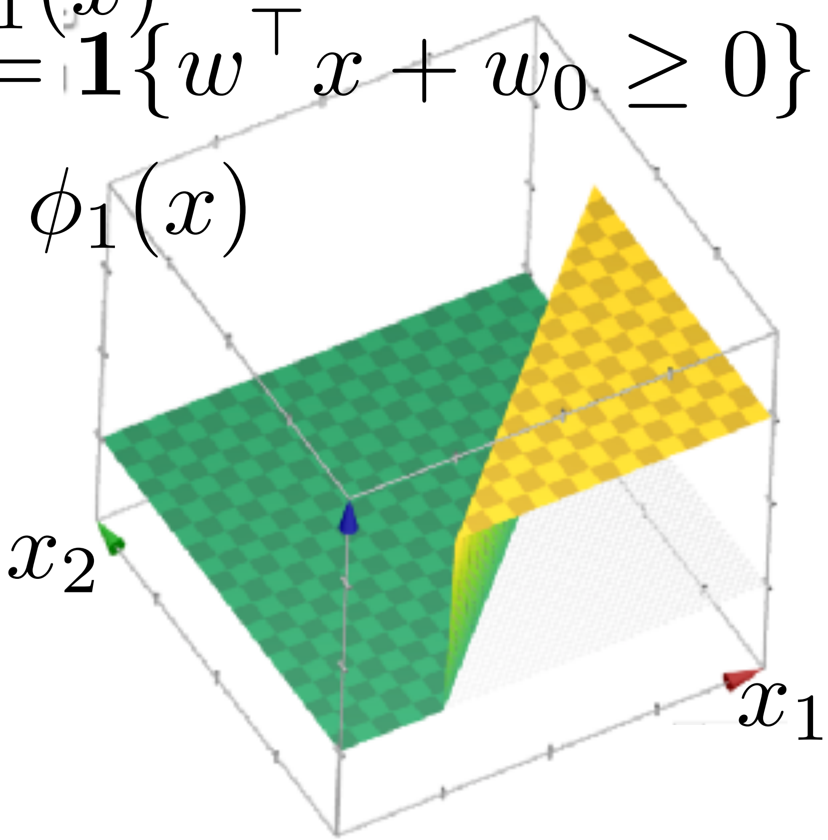


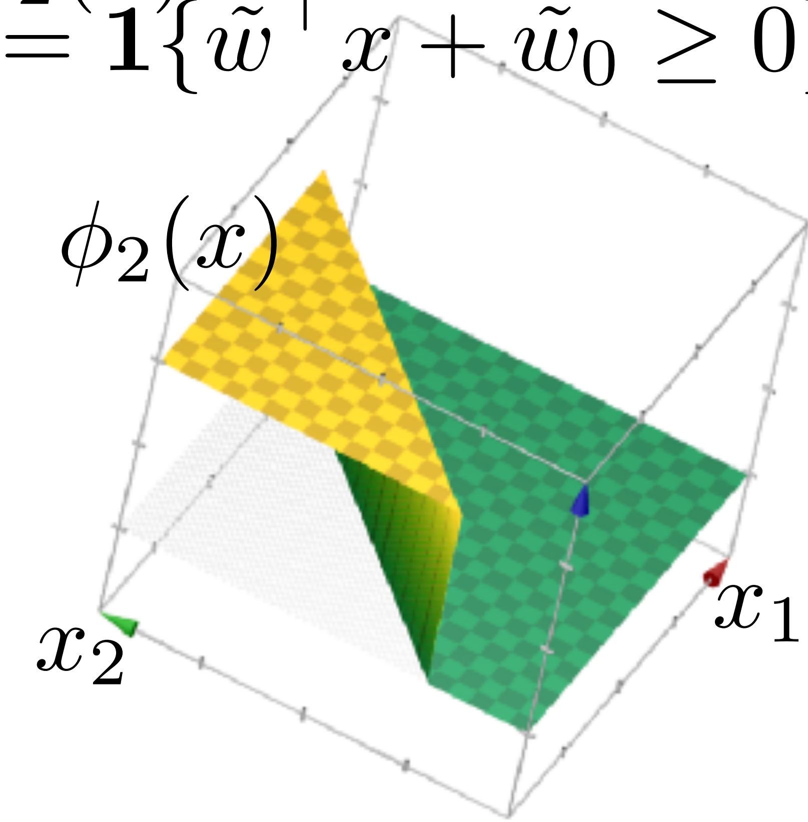


$$z = \theta^\top \phi(x) + \theta_0$$

New features: step functions!

$\phi_1(x) = \mathbf{1}\{w^\top x + w_0 \geq 0\}$ $\phi_2(x) = \mathbf{1}\{\tilde{w}^\top x + \tilde{w}_0 \geq 0\}$



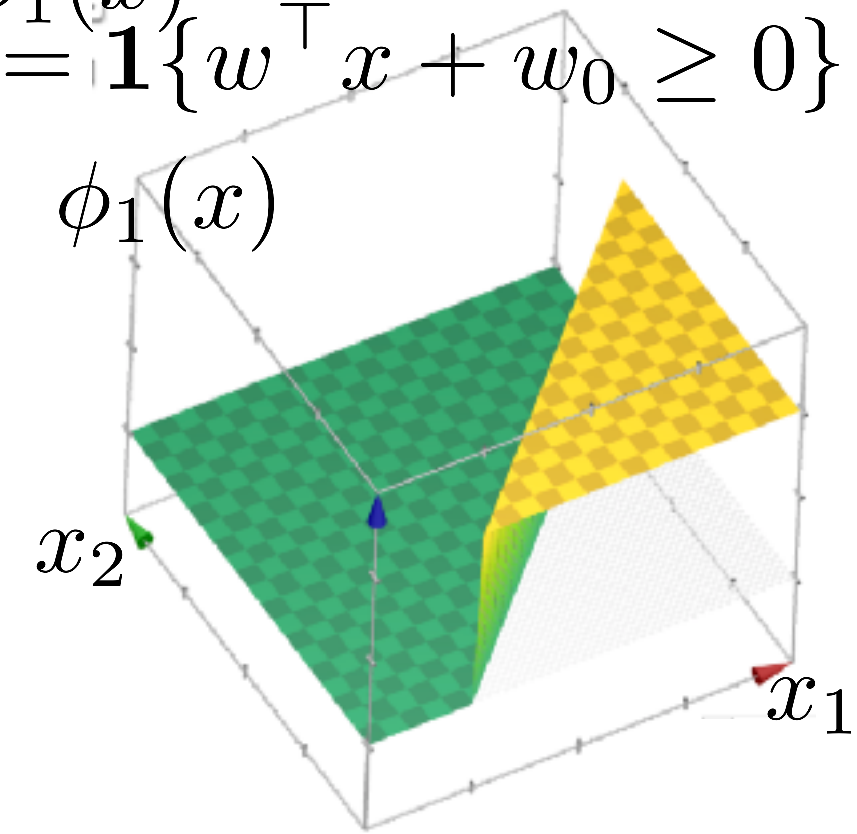


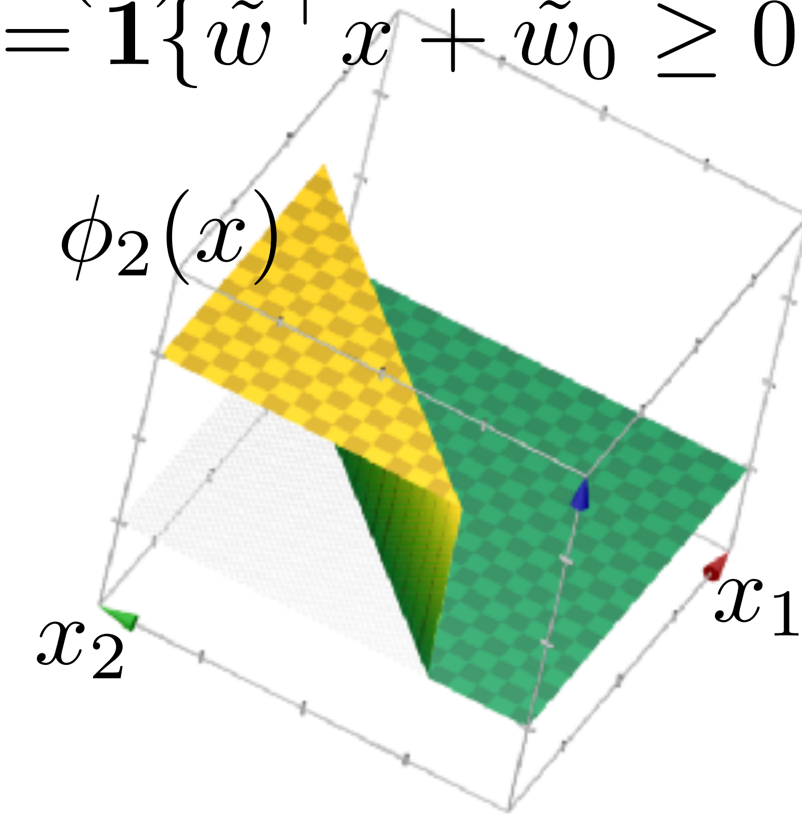
$$z = \theta^\top \phi(x) + \theta_0$$
$$= \theta_1\phi_1(x) + \theta_2\phi_2(x) + \theta_0$$

New features: step functions!

$\phi_1(x) = \mathbf{1}\{w^\top x + w_0 \geq 0\}$

$\phi_2(x) = \mathbf{1}\{\tilde{w}^\top x + \tilde{w}_0 \geq 0\}$





$$z = \theta^\top \phi(x) + \theta_0$$

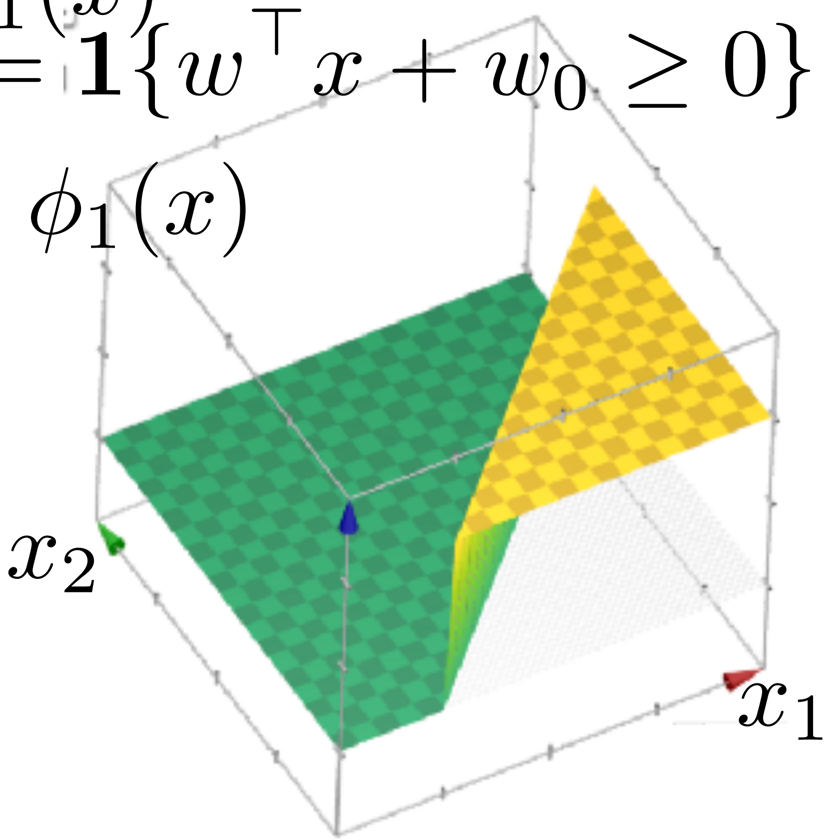
$$= \theta_1 \phi_1(x) + \theta_2 \phi_2(x) + \theta_0$$

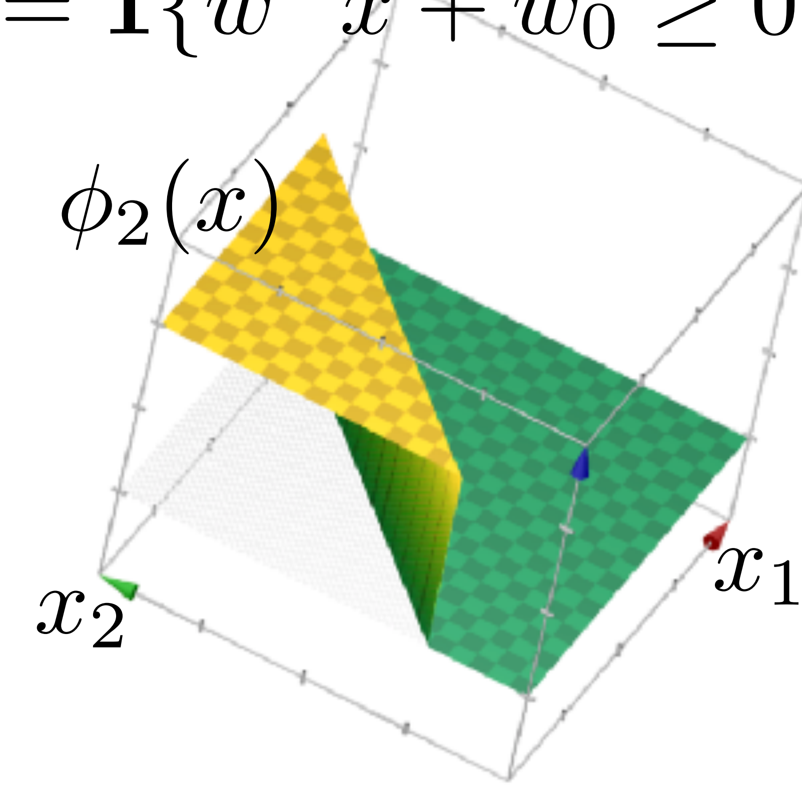
$$= 1 \cdot \phi_1(x) + 1 \cdot \phi_2(x) + (-0.5)$$

New features: step functions!

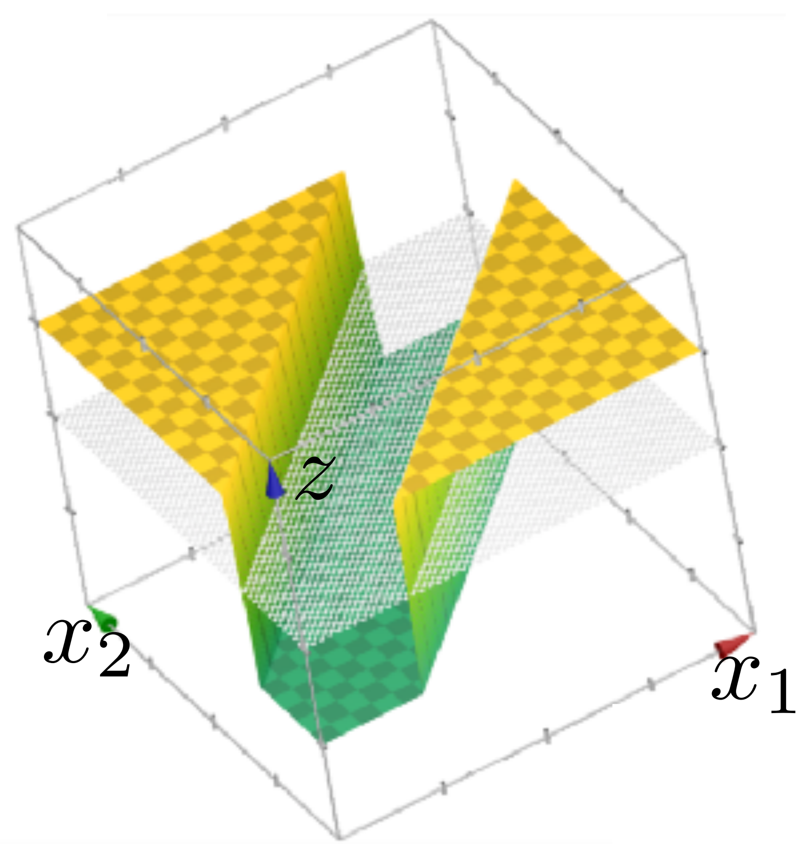
$$\phi_1(x) = \mathbf{1}\{w^\top x + w_0 \geq 0\}$$

$$\phi_2(x) = \mathbf{1}\{\tilde{w}^\top x + \tilde{w}_0 \geq 0\}$$



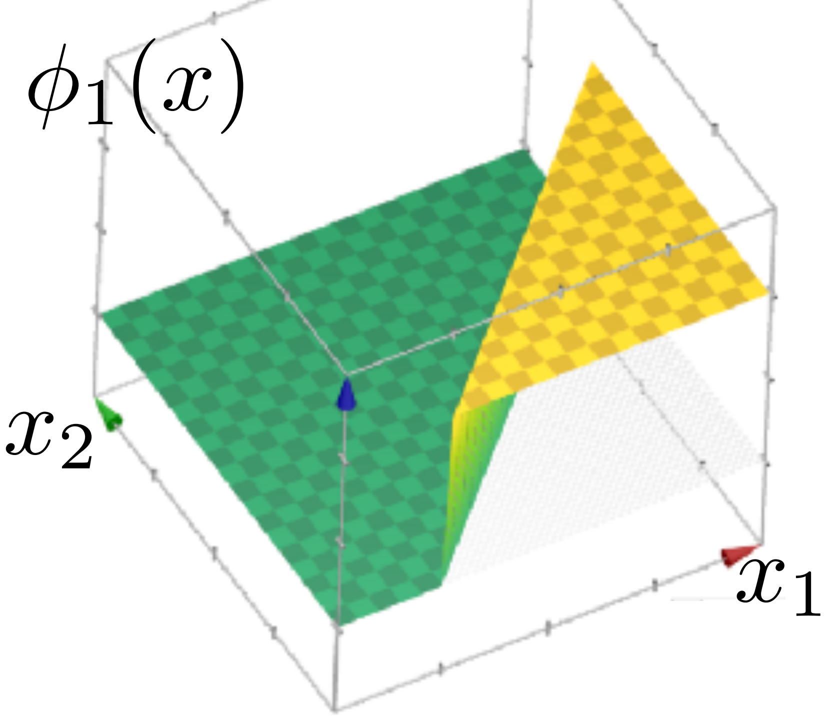


$$\begin{aligned}
z &= \theta^\top \phi(x) + \theta_0 \\
&= \theta_1 \phi_1(x) + \theta_2 \phi_2(x) \\
&\quad + \theta_0 \\
&= 1 \cdot \phi_1(x) + 1 \cdot \phi_2(x) \\
&\quad + (-0.5)
\end{aligned}$$

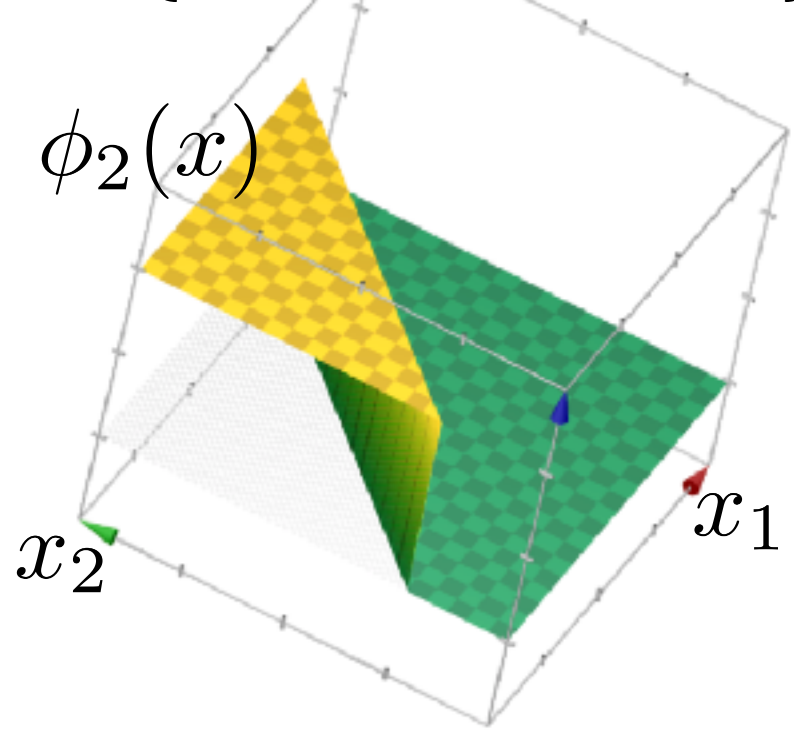


New features: step functions!

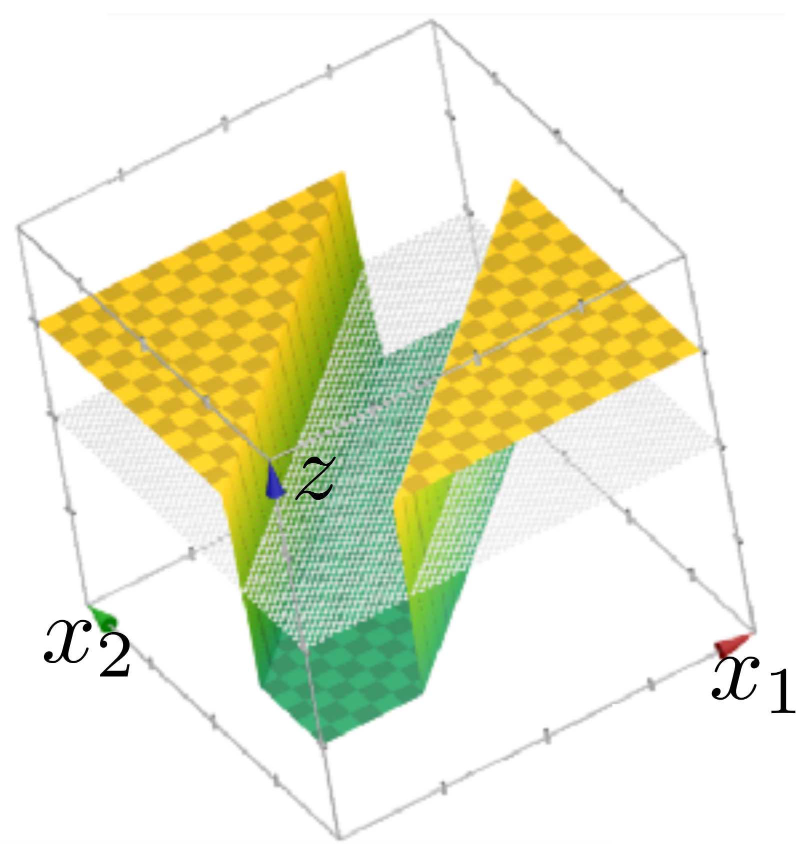
$\phi_1(x) = \mathbf{1}\{w^\top x + w_0 \geq 0\}$

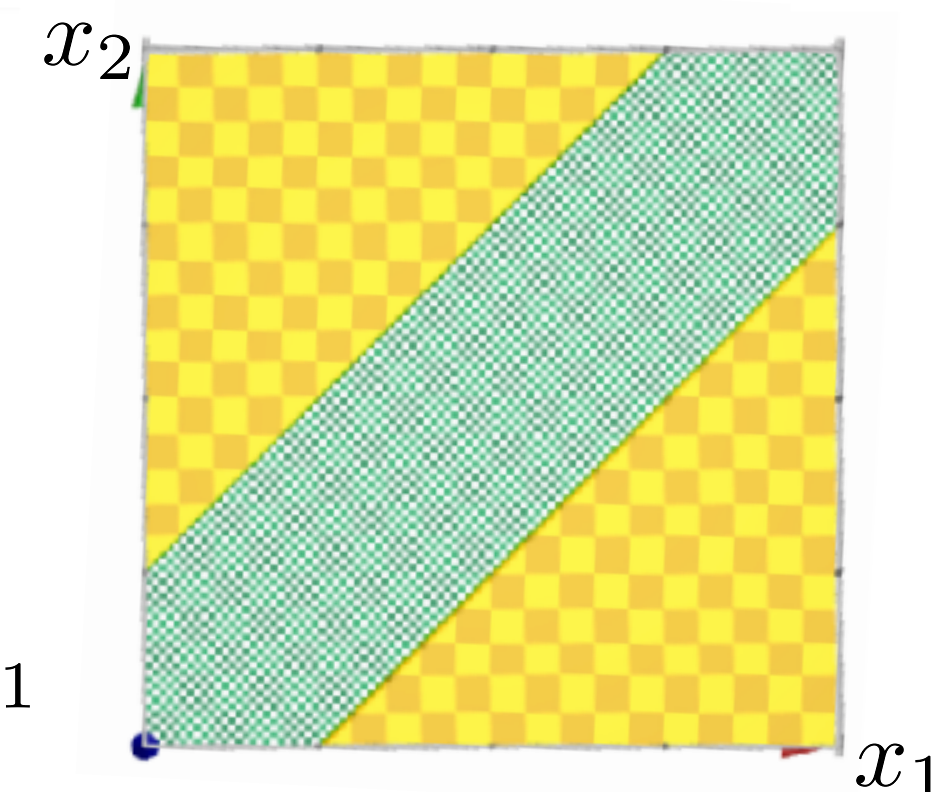


$\phi_2(x) = \mathbf{1}\{\tilde{w}^\top x + \tilde{w}_0 \geq 0\}$



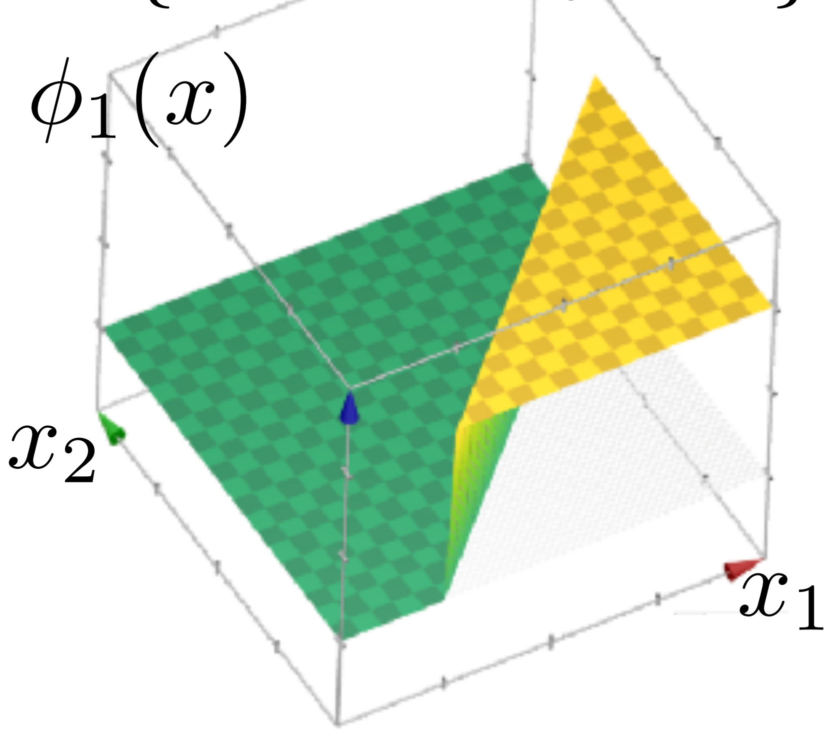
$$z = \theta^\top \phi(x) + \theta_0$$
$$= \theta_1 \phi_1(x) + \theta_2 \phi_2(x) + \theta_0$$
$$= 1 \cdot \phi_1(x) + 1 \cdot \phi_2(x) + (-0.5)$$



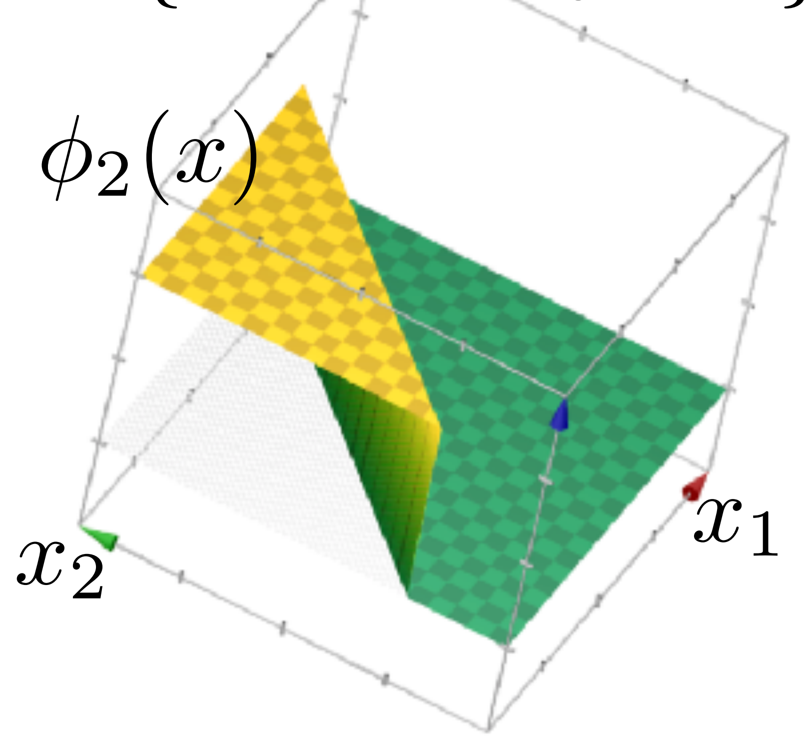


New features: step functions!

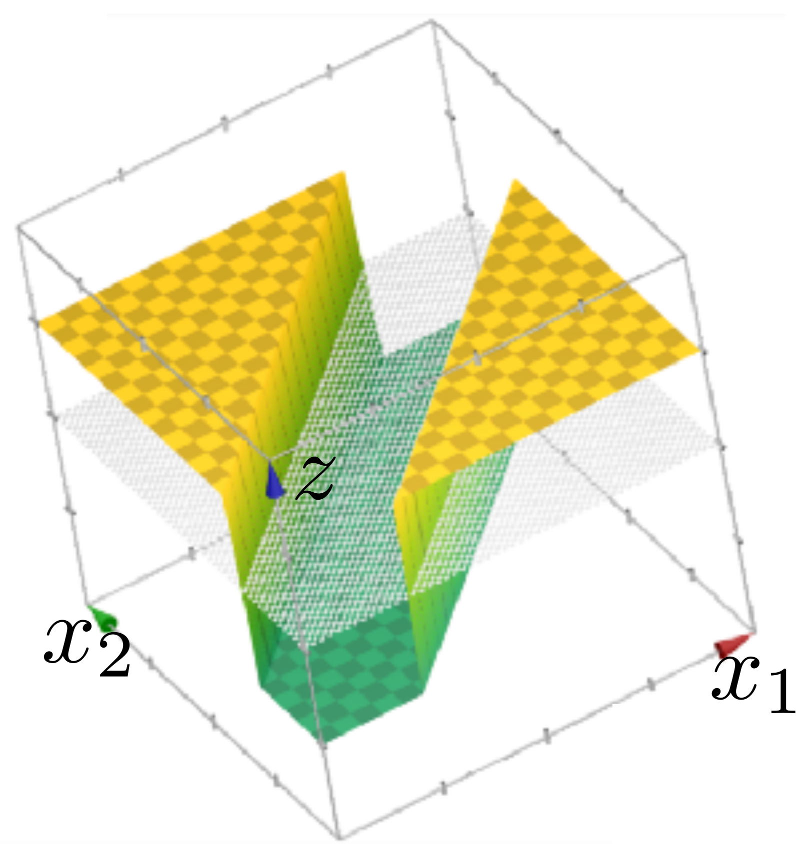
$\phi_1(x) = \mathbf{1}\{w^\top x + w_0 \geq 0\}$

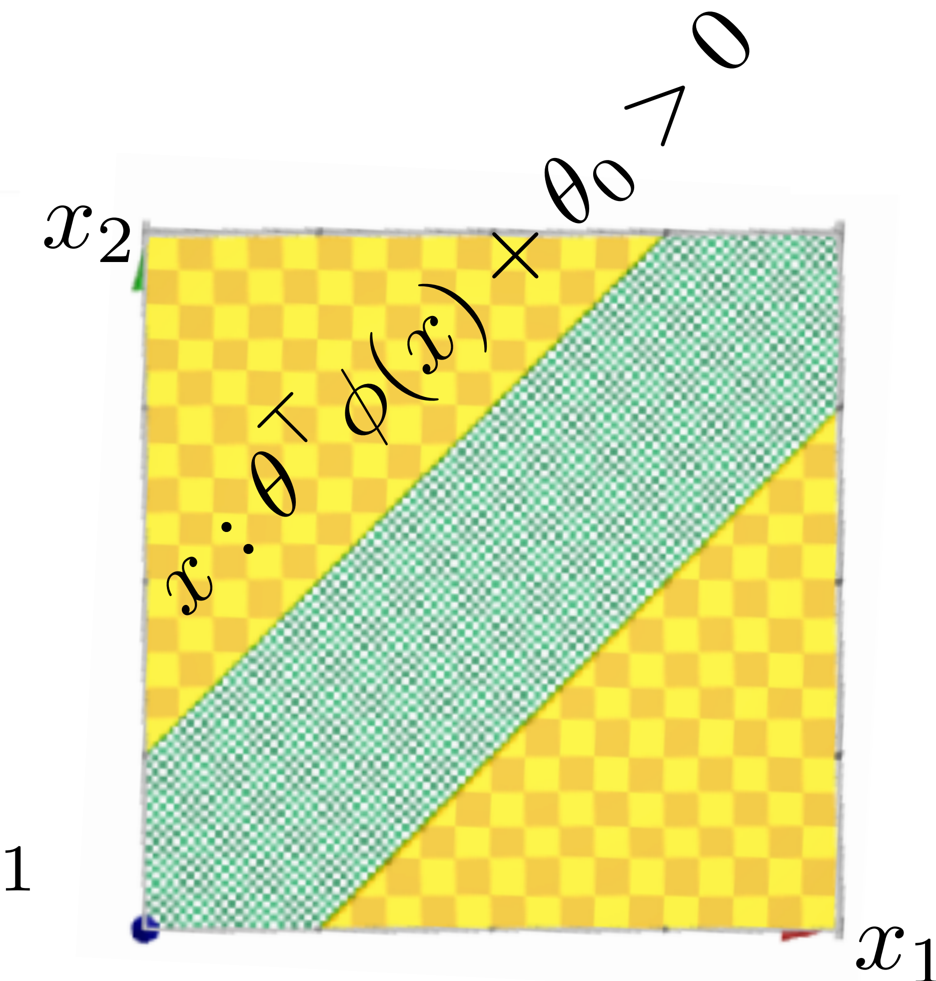


$\phi_2(x) = \mathbf{1}\{\tilde{w}^\top x + \tilde{w}_0 \geq 0\}$



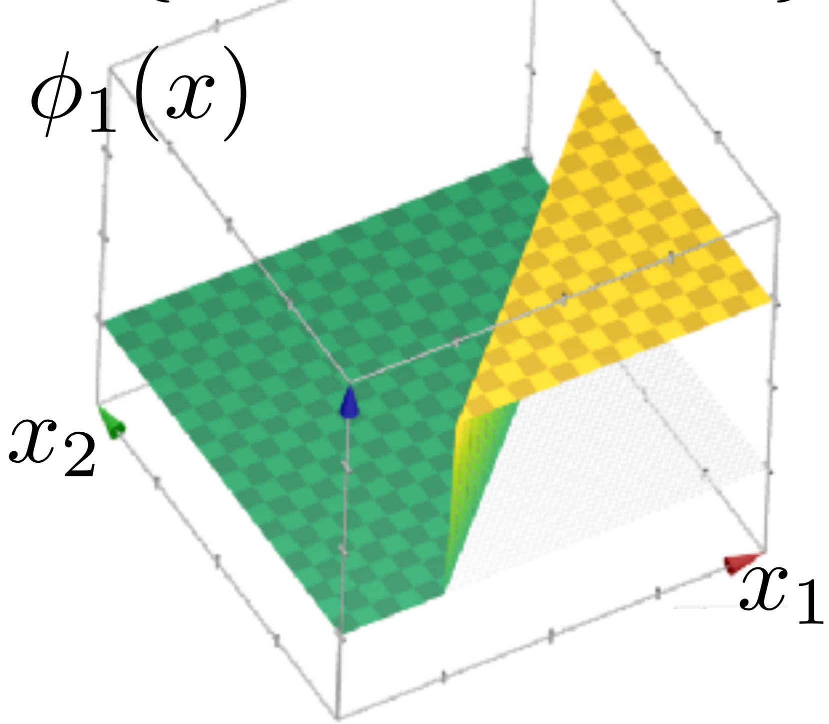
$$z = \theta^\top \phi(x) + \theta_0$$
$$= \theta_1 \phi_1(x) + \theta_2 \phi_2(x) + \theta_0$$
$$= 1 \cdot \phi_1(x) + 1 \cdot \phi_2(x) + (-0.5)$$



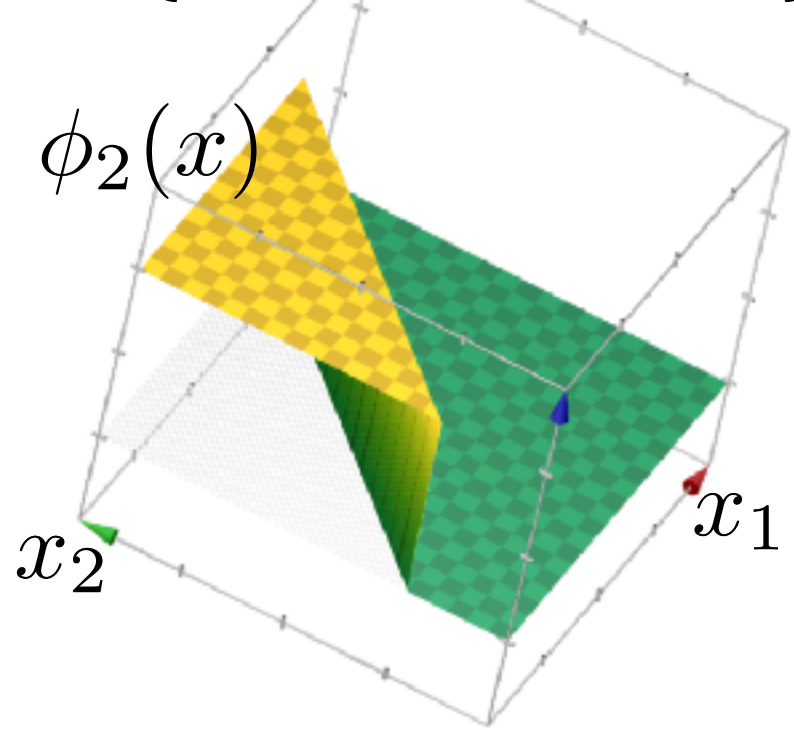


New features: step functions!

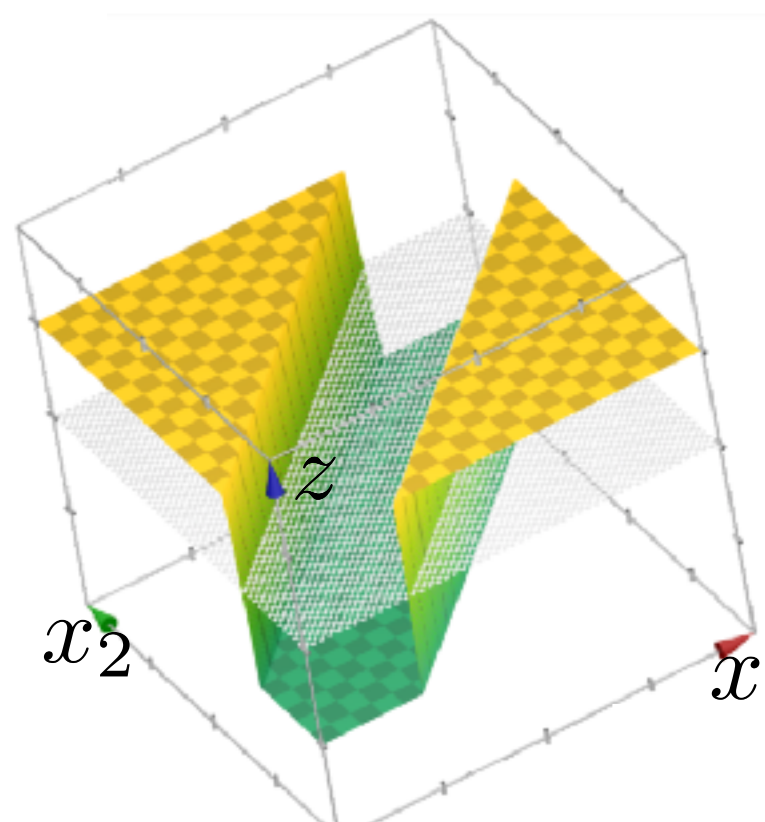
$\phi_1(x) = \mathbf{1}\{w^\top x + w_0 \geq 0\}$

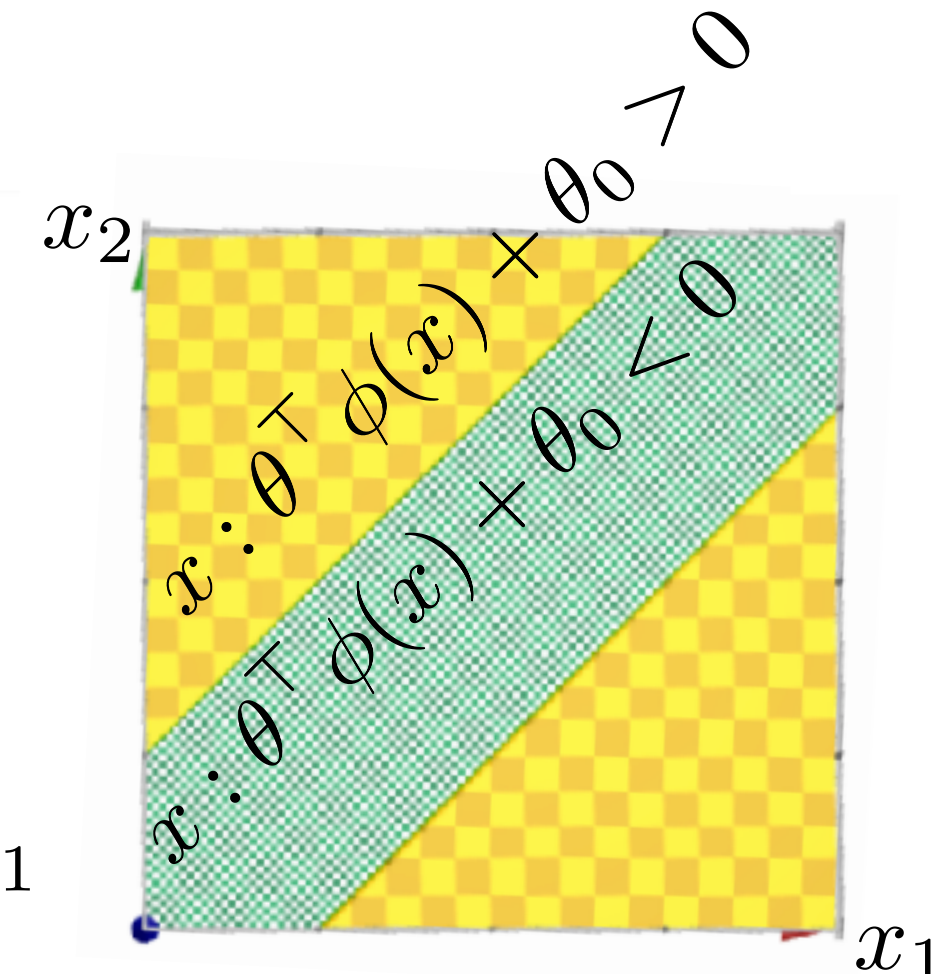


$\phi_2(x) = \mathbf{1}\{\tilde{w}^\top x + \tilde{w}_0 \geq 0\}$



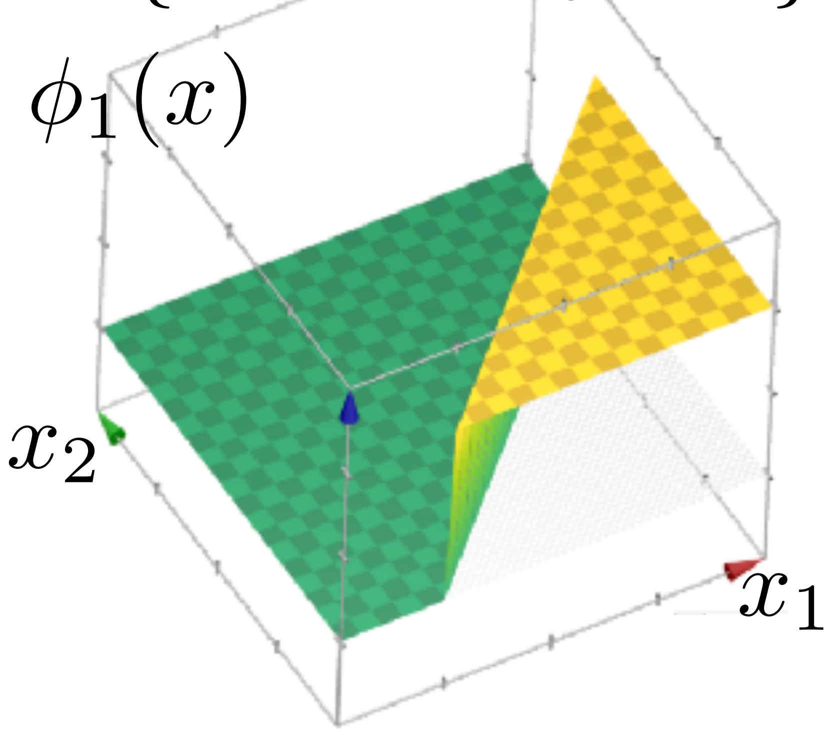
$$z = \theta^\top \phi(x) + \theta_0$$
$$= \theta_1 \phi_1(x) + \theta_2 \phi_2(x) + \theta_0$$
$$= 1 \cdot \phi_1(x) + 1 \cdot \phi_2(x) + (-0.5)$$



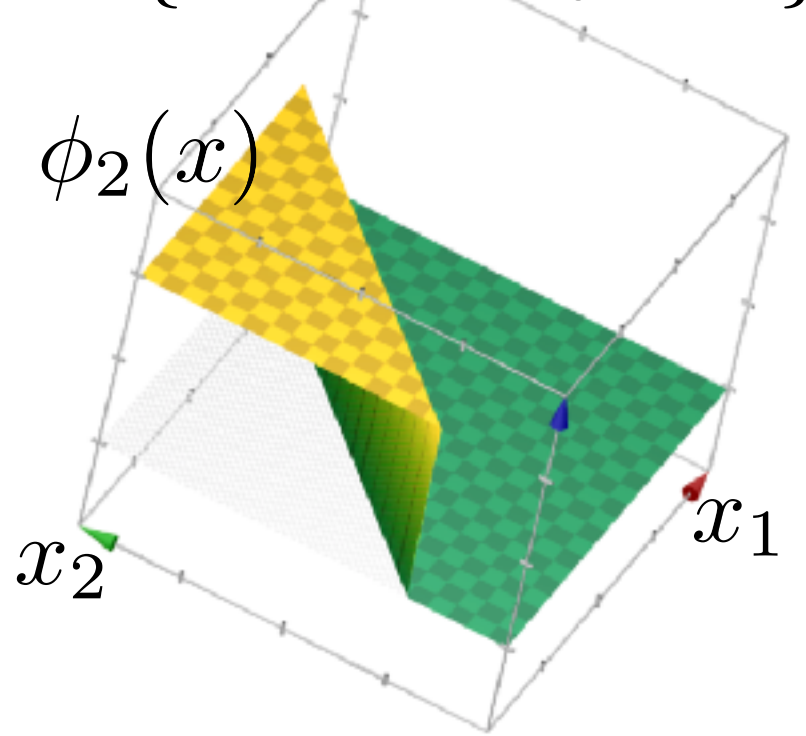


New features: step functions!

$\phi_1(x) = \mathbf{1}\{w^\top x + w_0 \geq 0\}$

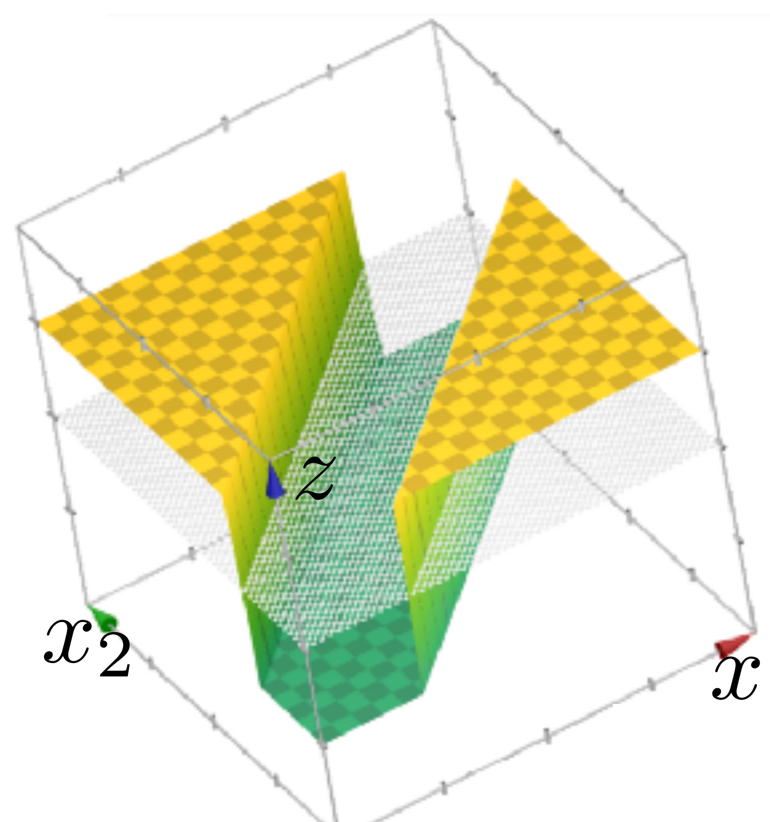


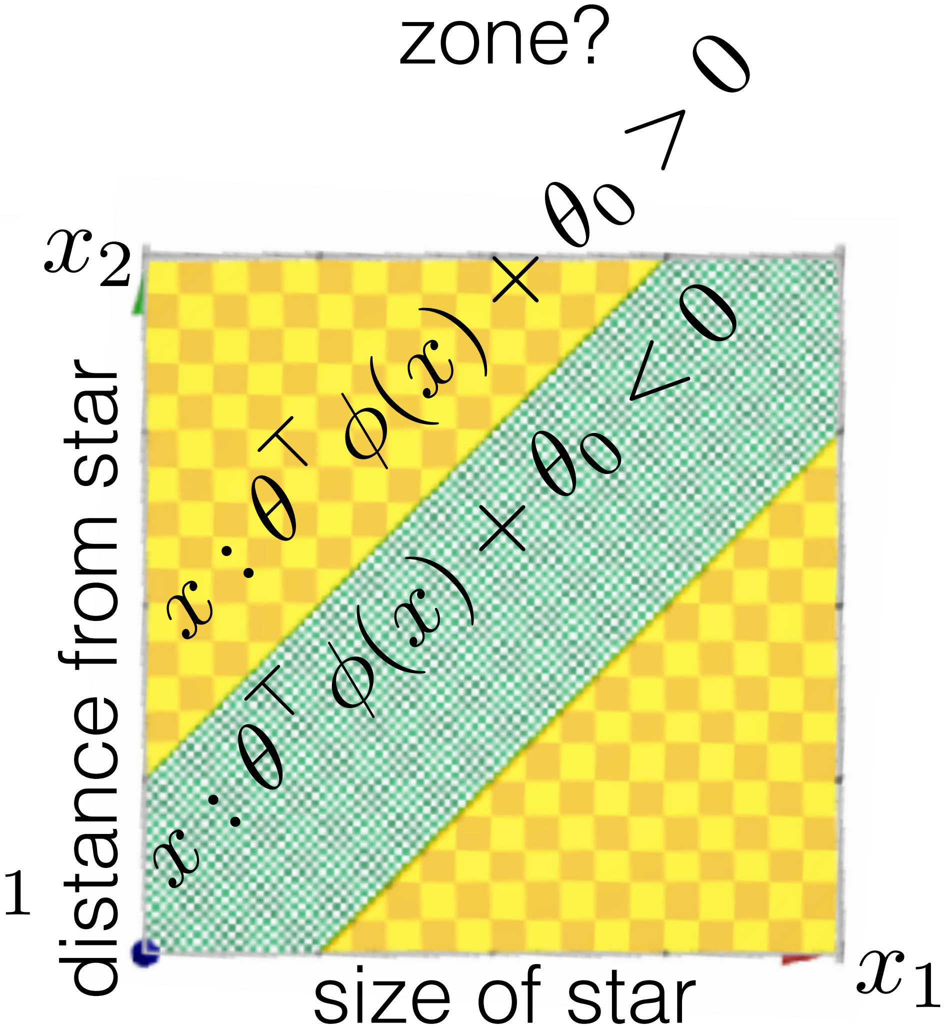
$\phi_2(x) = \mathbf{1}\{\tilde{w}^\top x + \tilde{w}_0 \geq 0\}$



Is an exoplanet in the habitable zone?

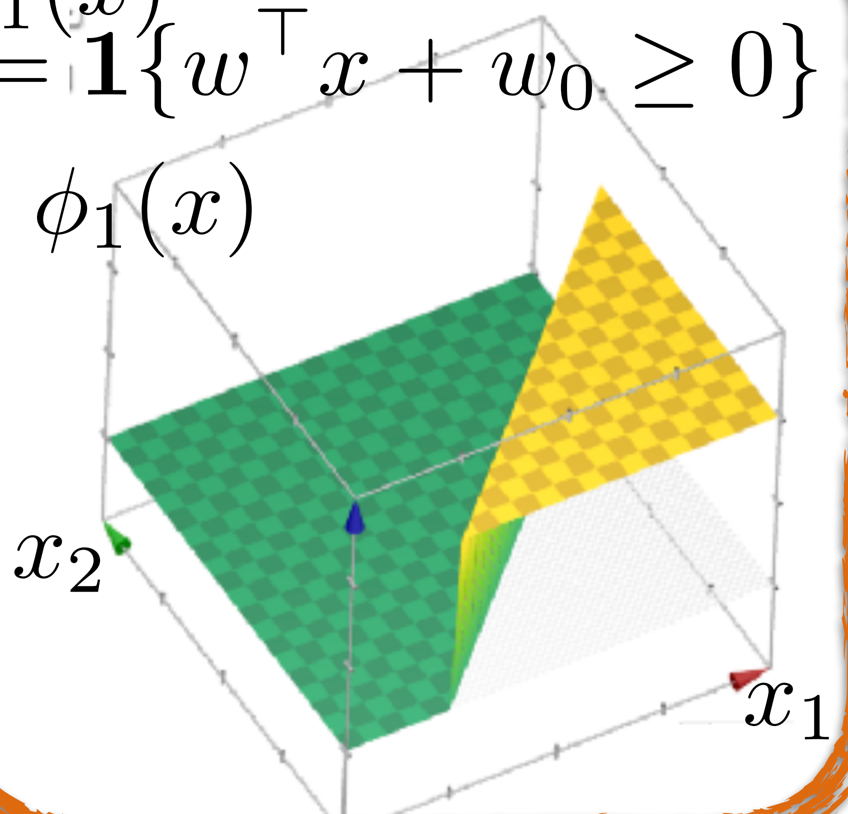
$$
\begin{aligned}
z &= \theta^\top \phi(x) + \theta_0 \\
&= \theta_1 \phi_1(x) + \theta_2 \phi_2(x) + \theta_0 \\
&= 1 \cdot \phi_1(x) + 1 \cdot \phi_2(x) + (-0.5)
\end{aligned}
$$



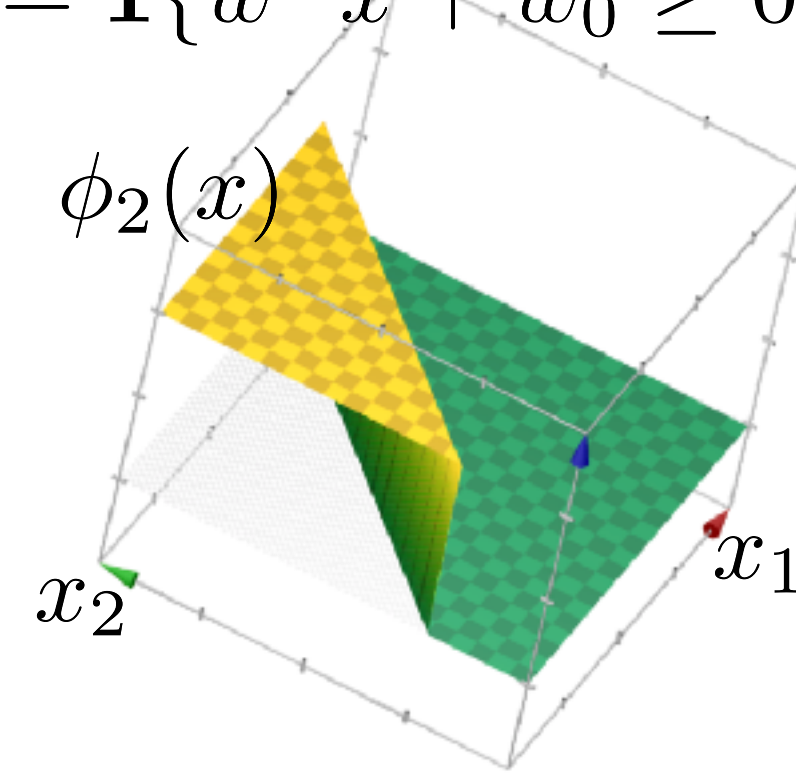


New features: step functions!

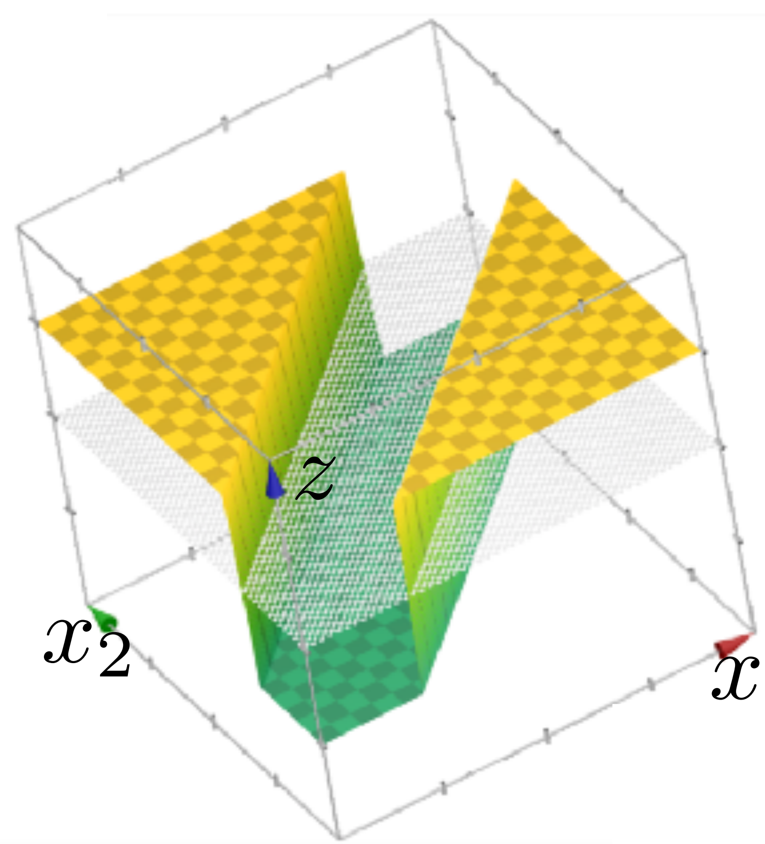
$\phi_1(x) = \mathbf{1}\{w^\top x + w_0 \geq 0\}$

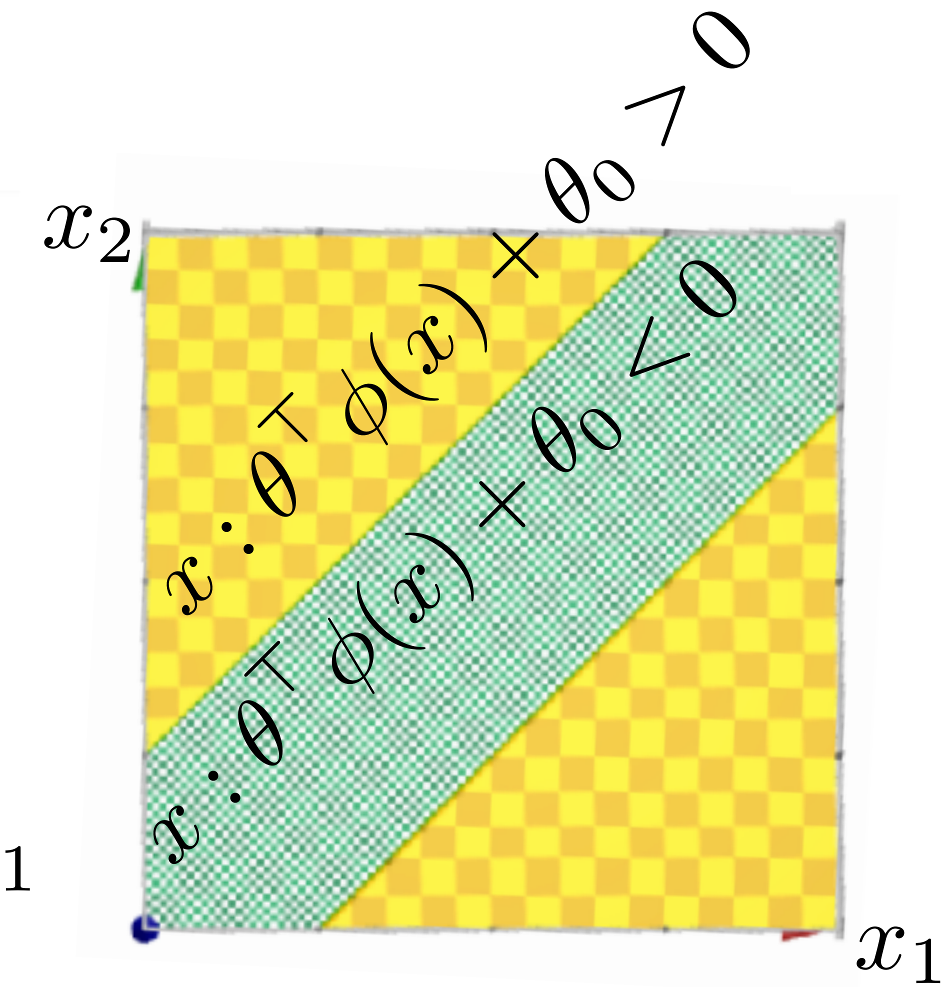


$\phi_2(x) = \mathbf{1}\{\tilde{w}^\top x + \tilde{w}_0 \geq 0\}$



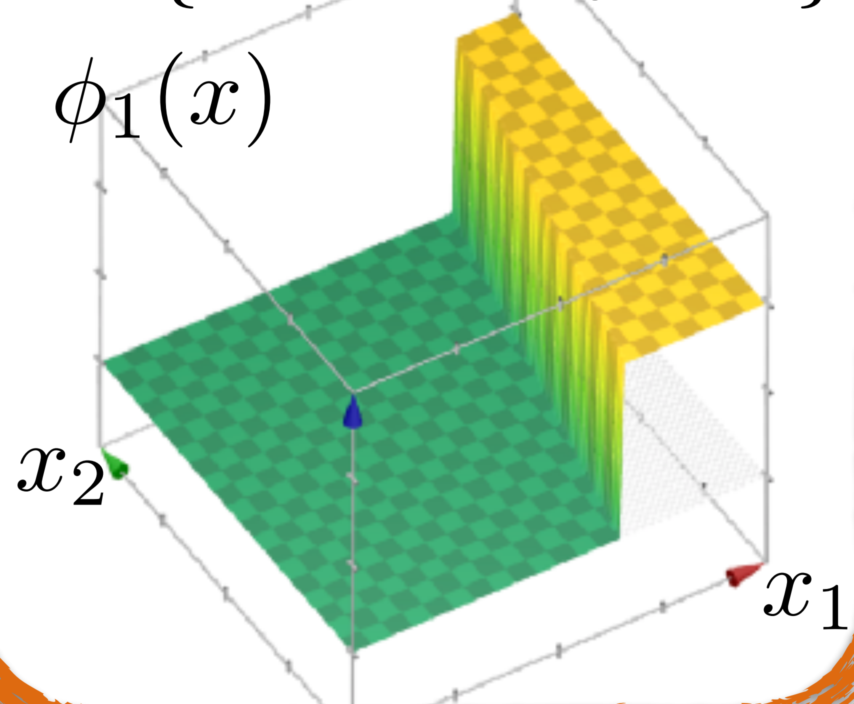
$$z = \theta^\top \phi(x) + \theta_0$$
$$= \theta_1 \phi_1(x) + \theta_2 \phi_2(x) + \theta_0$$
$$= 1 \cdot \phi_1(x) + 1 \cdot \phi_2(x) + (-0.5)$$



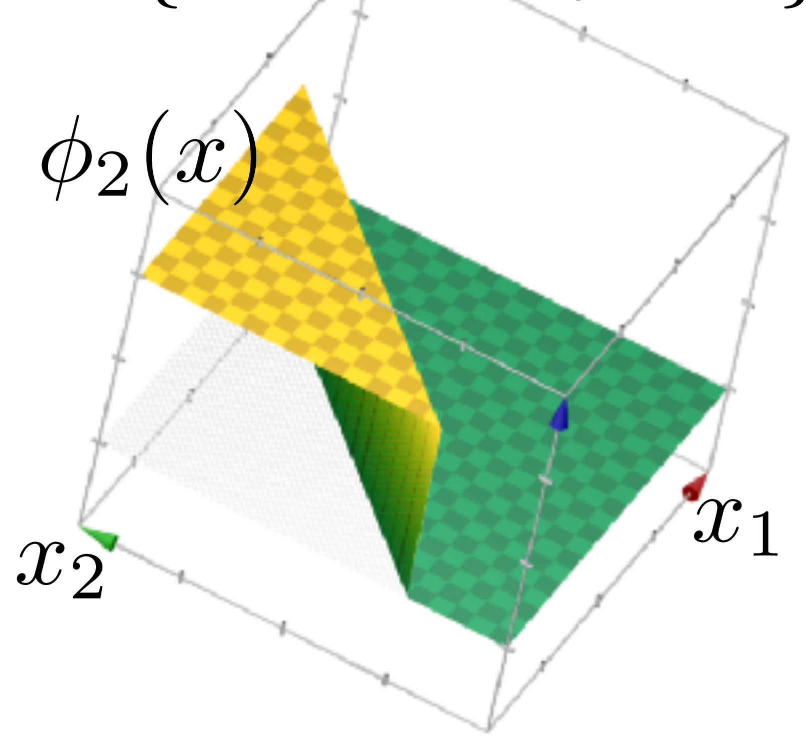


New features: step functions!

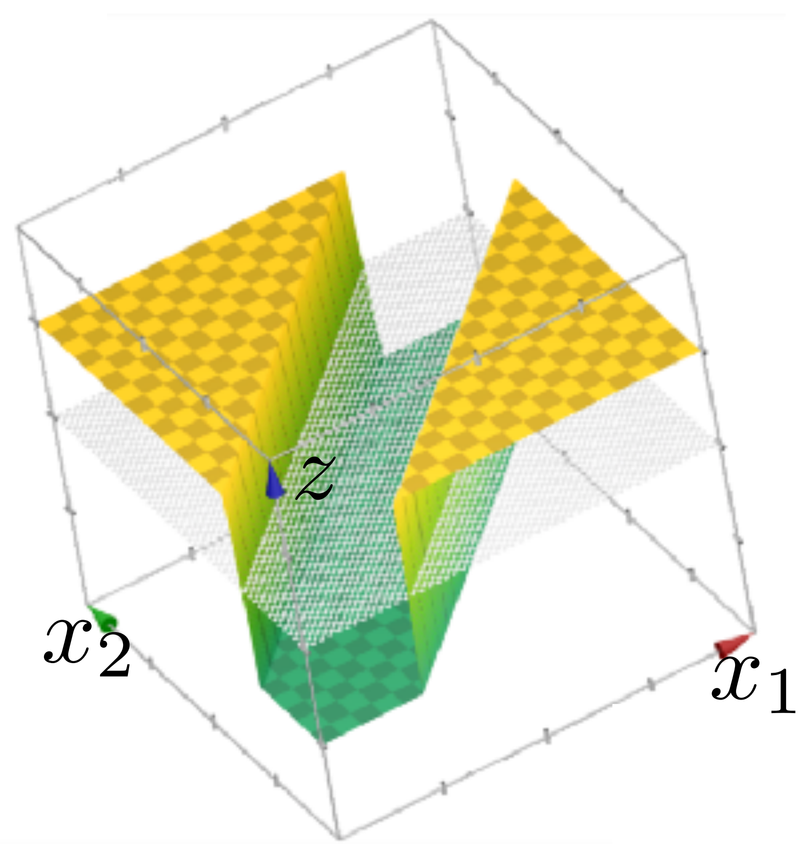
$\phi_1(x)$
$= \mathbf{1}\{w^\top x + w_0 \geq 0\}$

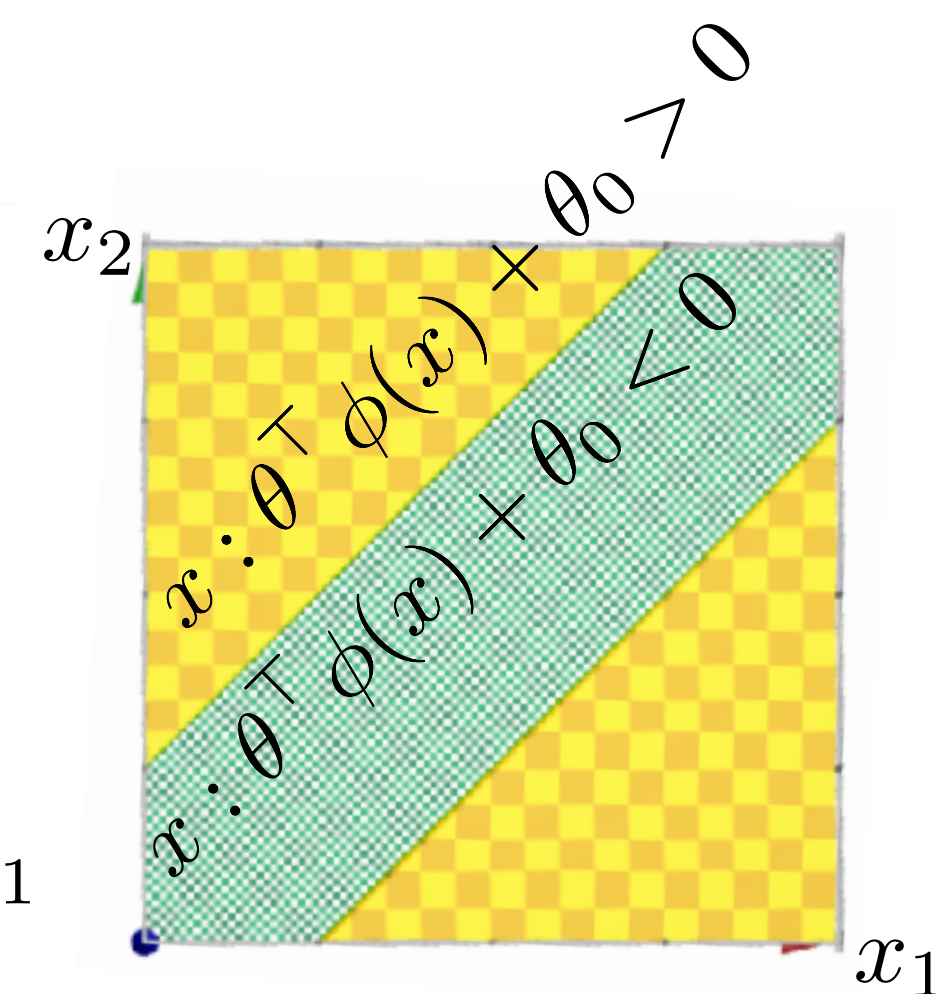


$\phi_2(x)$
$= \mathbf{1}\{\tilde{w}^\top x + \tilde{w}_0 \geq 0\}$



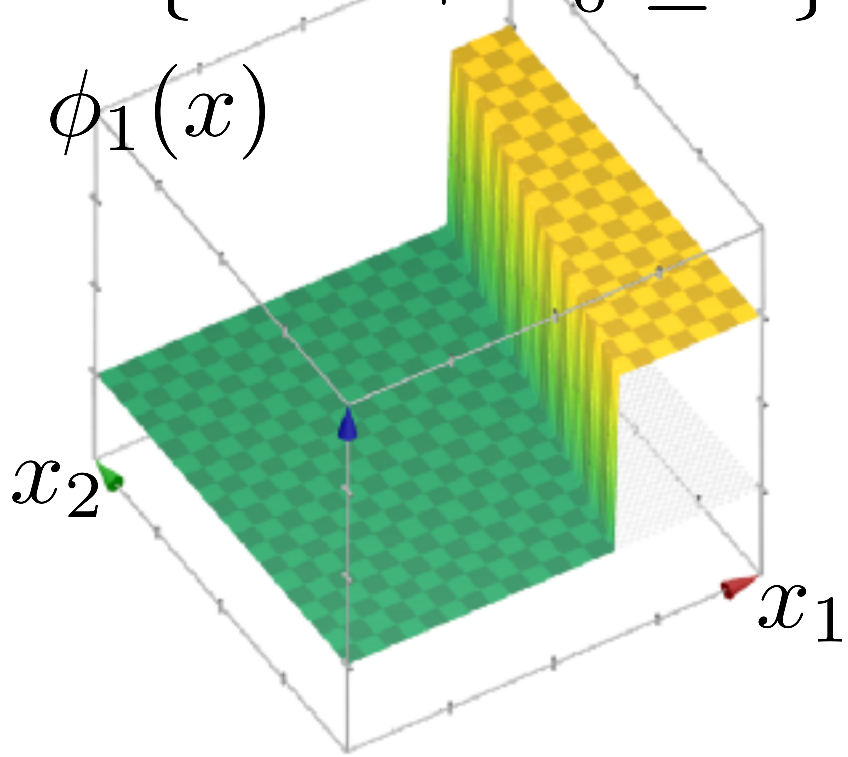
$$z = \theta^\top \phi(x) + \theta_0$$
$$= \theta_1 \phi_1(x) + \theta_2 \phi_2(x) + \theta_0$$
$$= 1 \cdot \phi_1(x) + 1 \cdot \phi_2(x) + (-0.5)$$



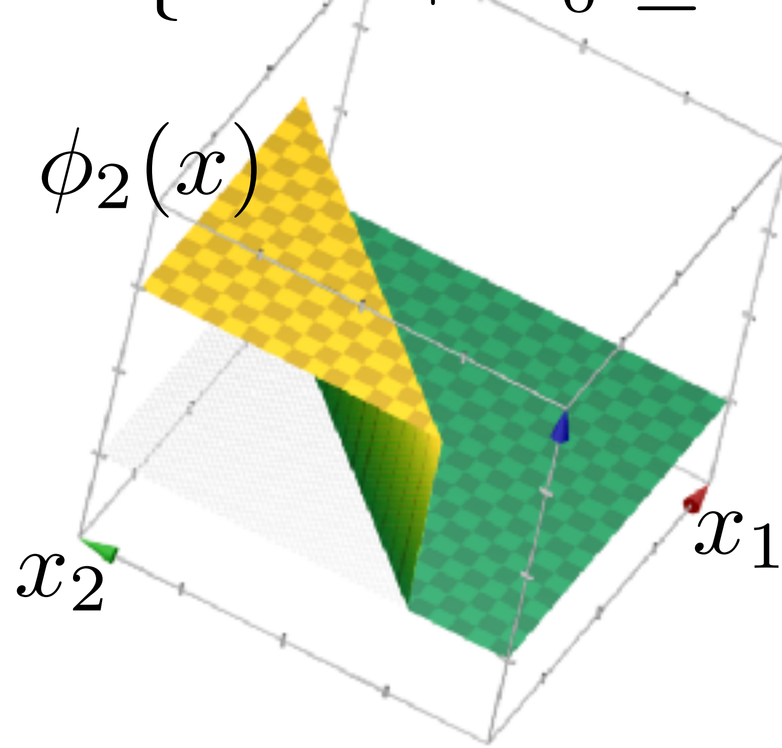


New features: step functions!

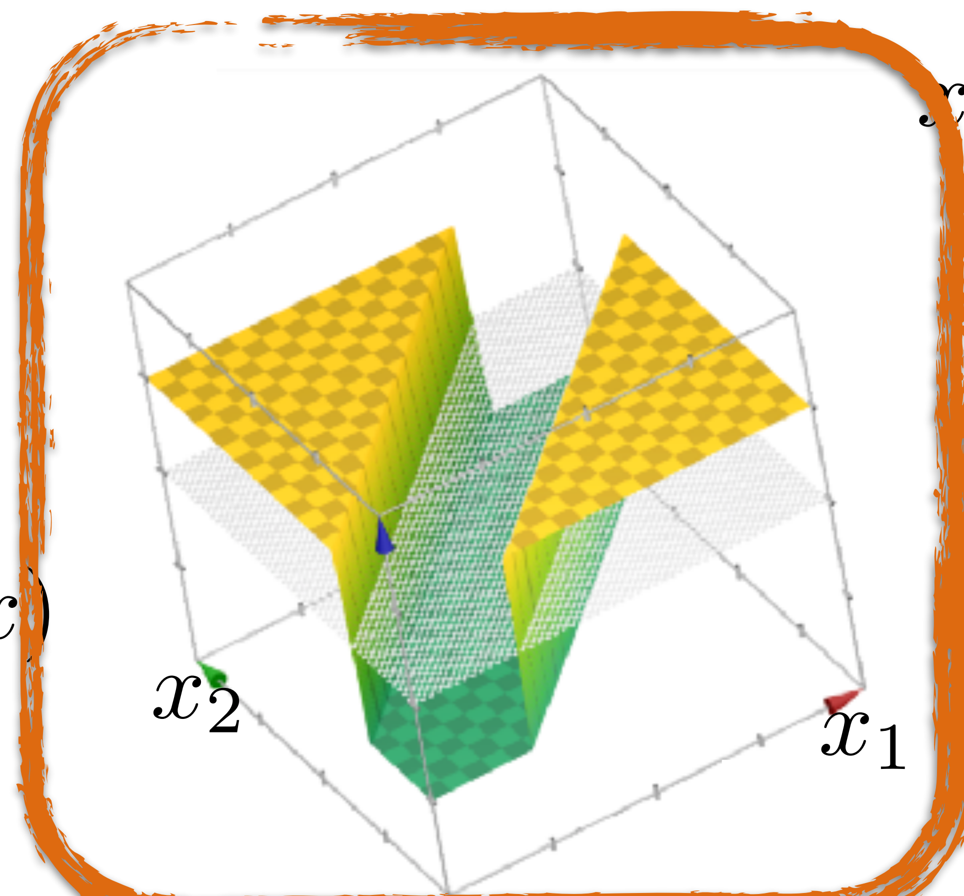
$\phi_1(x) = \mathbf{1}\{w^\top x + w_0 \geq 0\}$

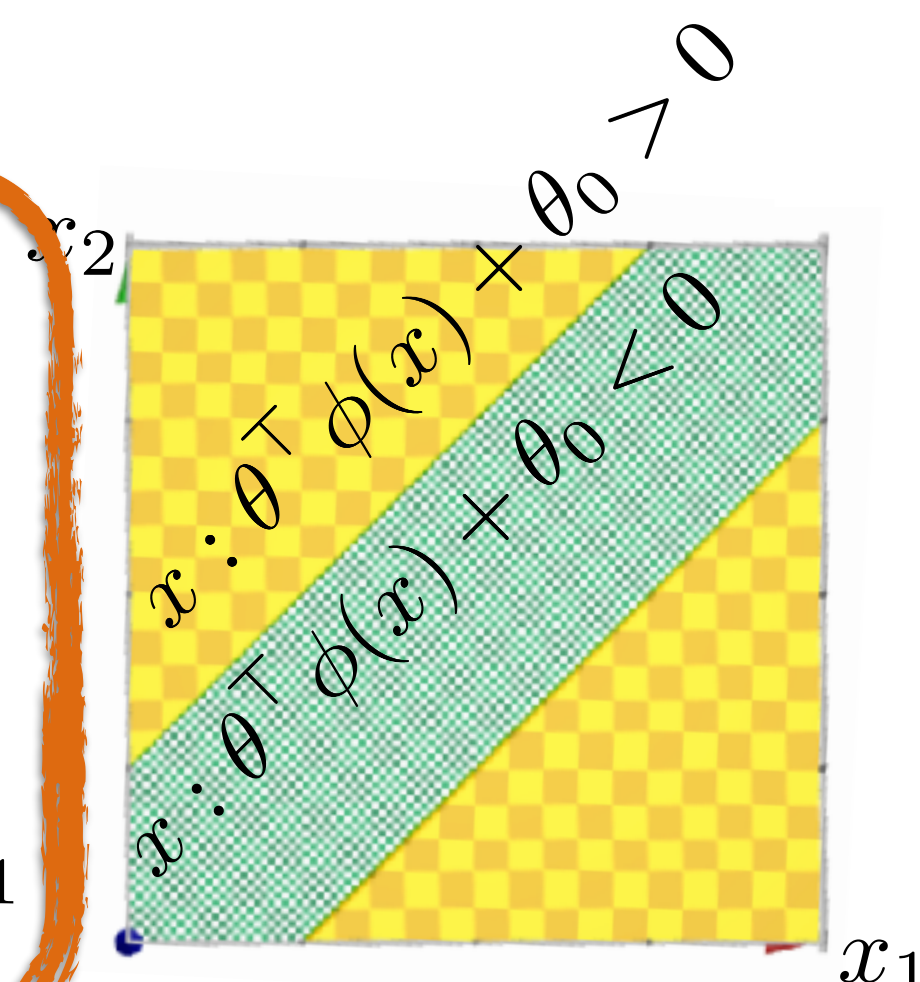


$\phi_2(x) = \mathbf{1}\{\tilde{w}^\top x + \tilde{w}_0 \geq 0\}$



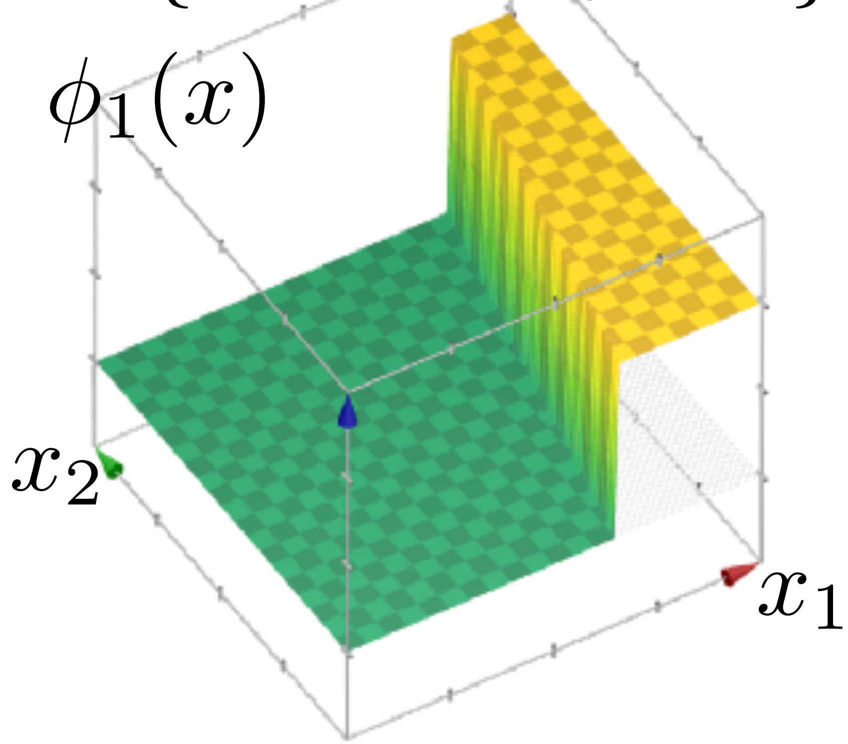
$$z = \theta^\top \phi(x) + \theta_0$$
$$= \theta_1 \phi_1(x) + \theta_2 \phi_2(x) + \theta_0$$
$$= 1 \cdot \phi_1(x) + 1 \cdot \phi_2(x) + (-0.5)$$



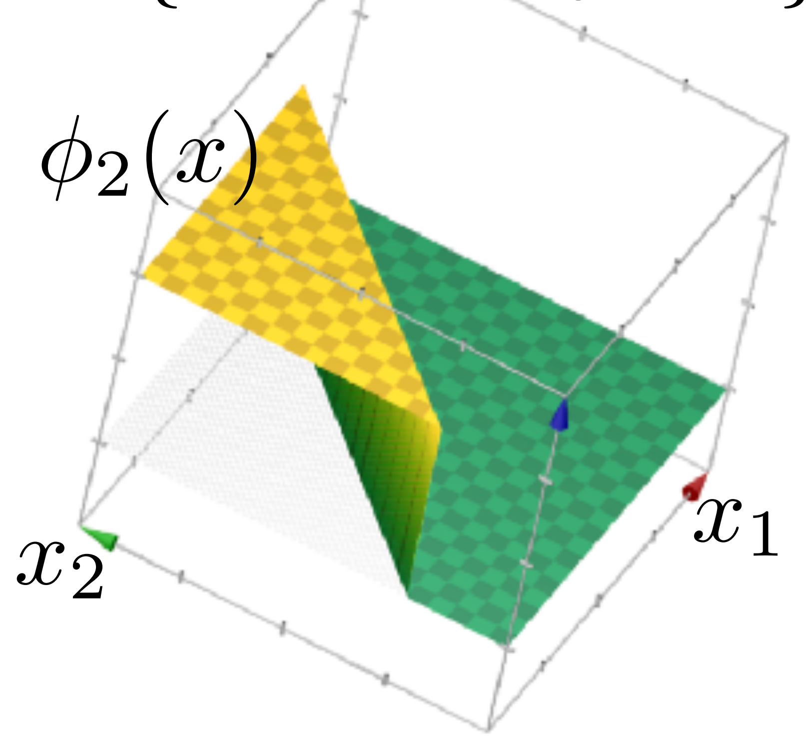


New features: step functions!

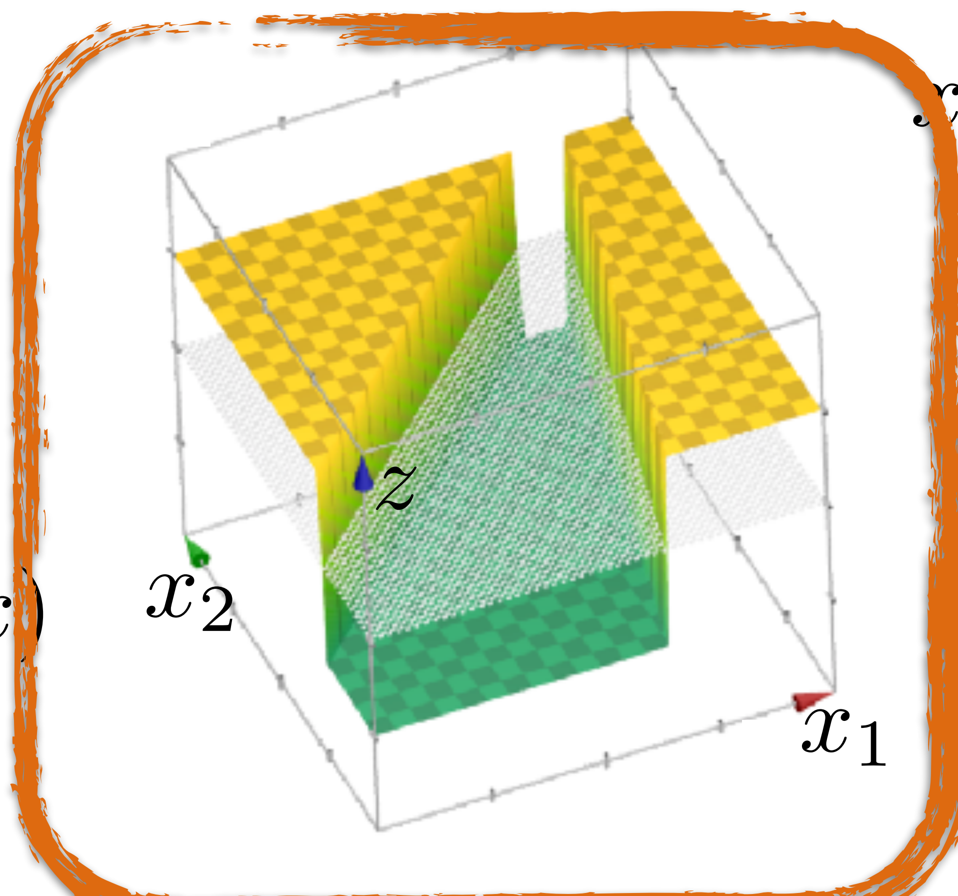
$\phi_1(x) = \mathbf{1}\{w^\top x + w_0 \geq 0\}$

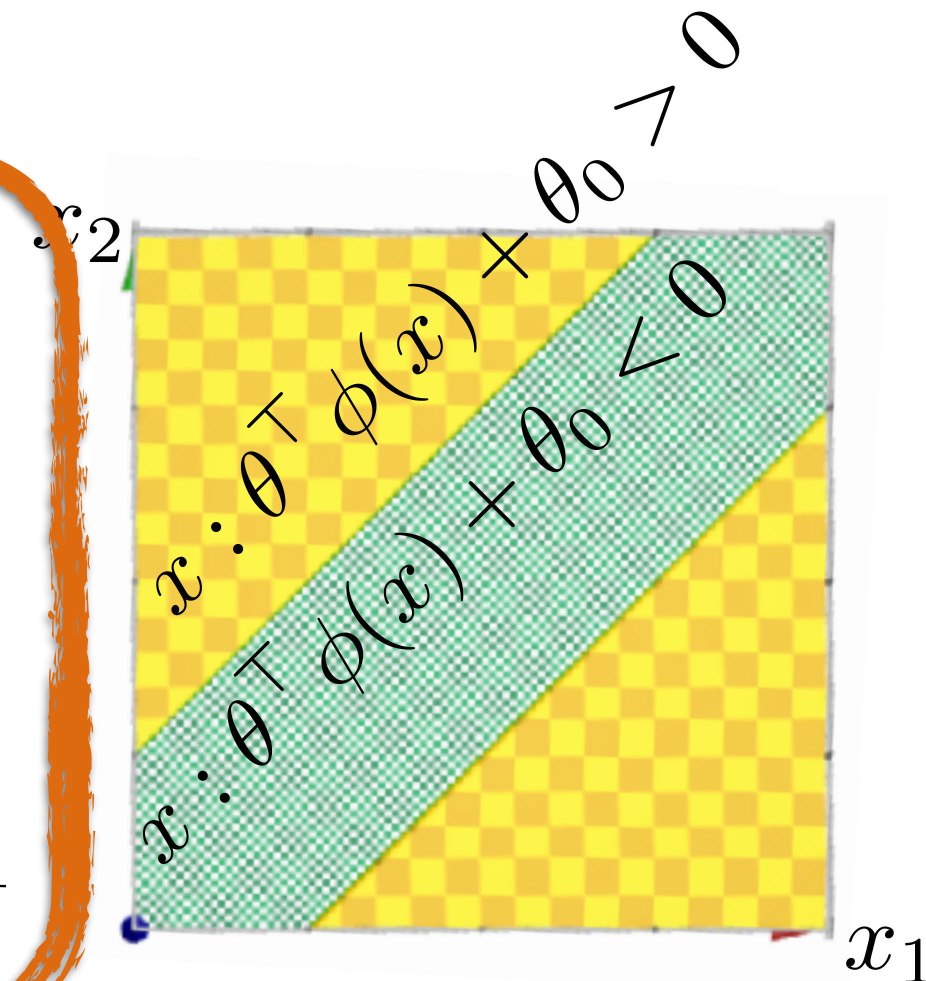


$\phi_2(x) = \mathbf{1}\{\tilde{w}^\top x + \tilde{w}_0 \geq 0\}$



$$z = \theta^\top \phi(x) + \theta_0$$
$$= \theta_1 \phi_1(x) + \theta_2 \phi_2(x) + \theta_0$$
$$= 1 \cdot \phi_1(x) + 1 \cdot \phi_2(x) + (-0.5)$$

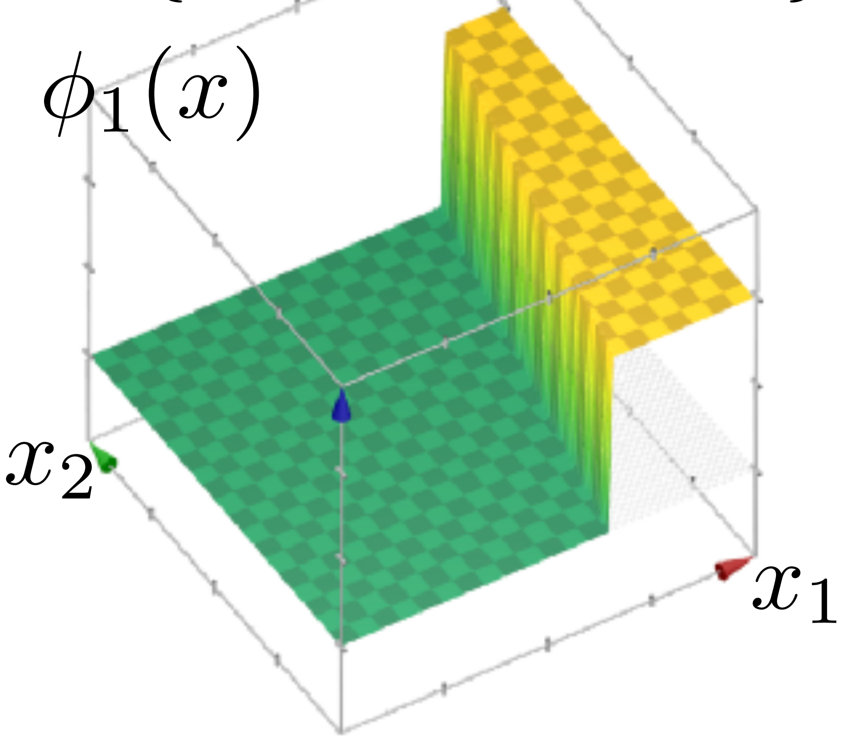


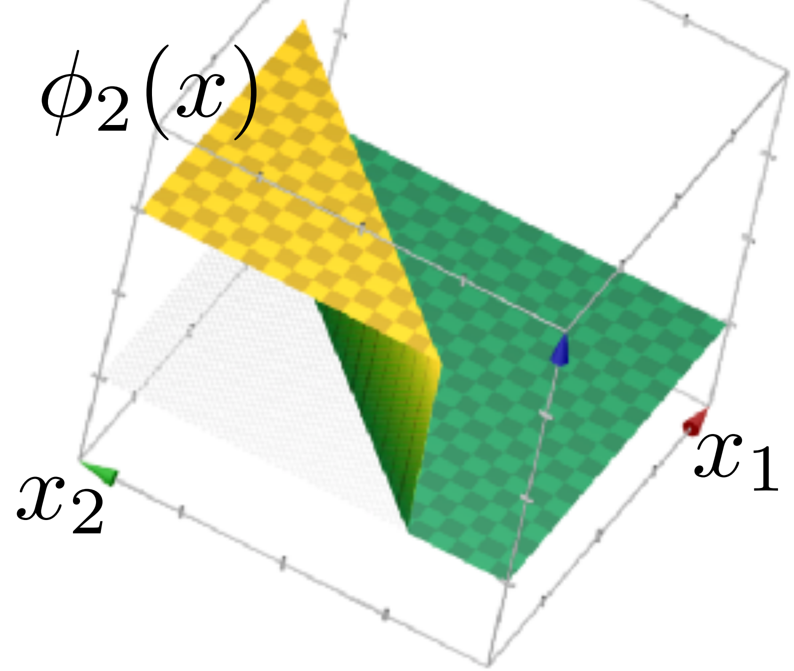


New features: step functions!

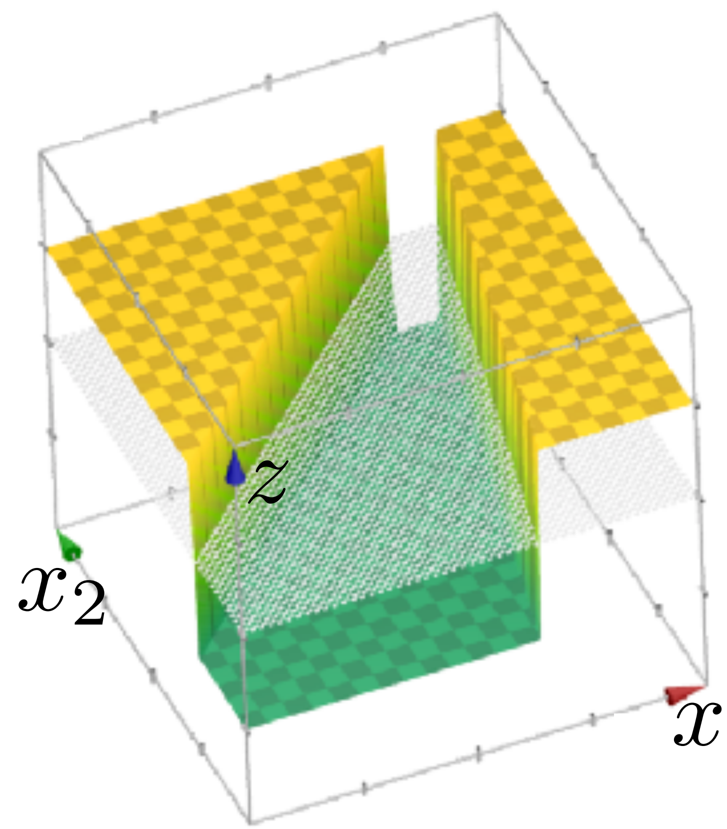
$\phi_1(x) = \mathbf{1}\{w^\top x + w_0 \geq 0\}$

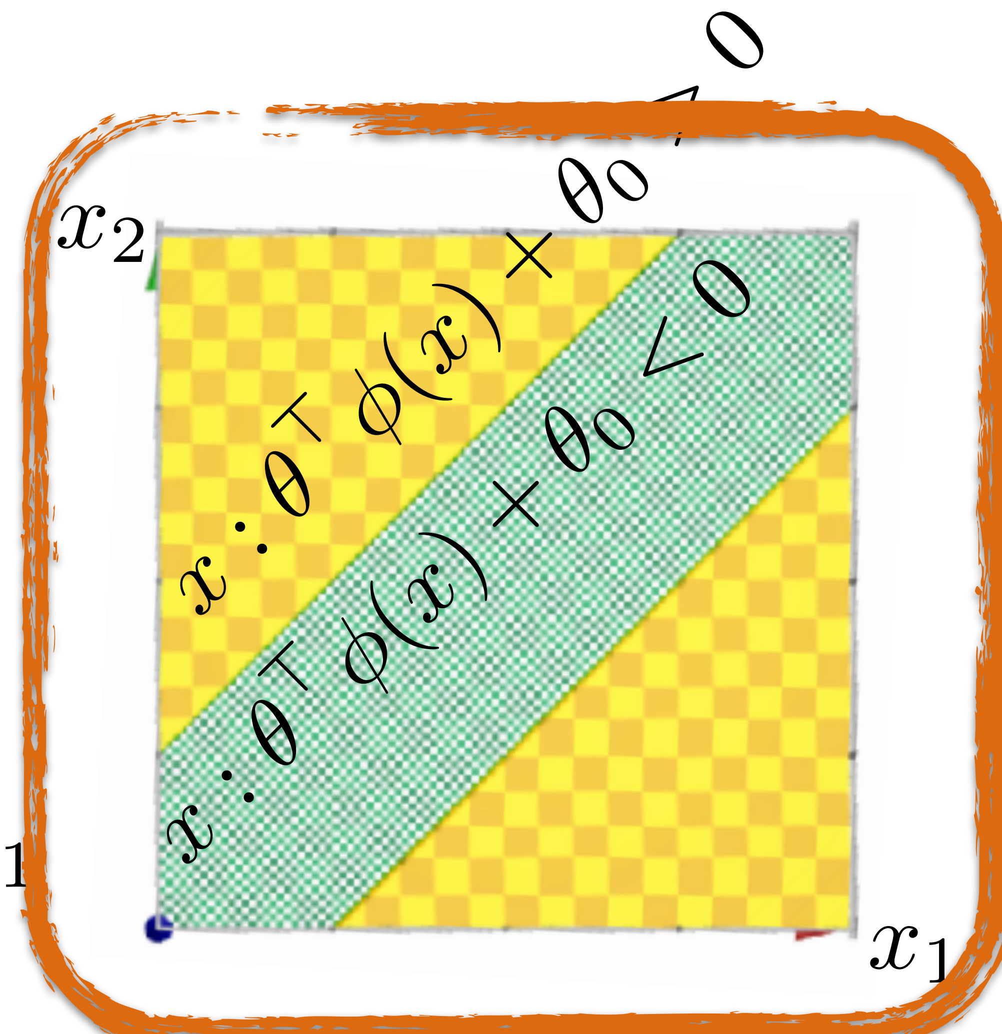
$\phi_2(x) = \mathbf{1}\{\tilde{w}^\top x + \tilde{w}_0 \geq 0\}$





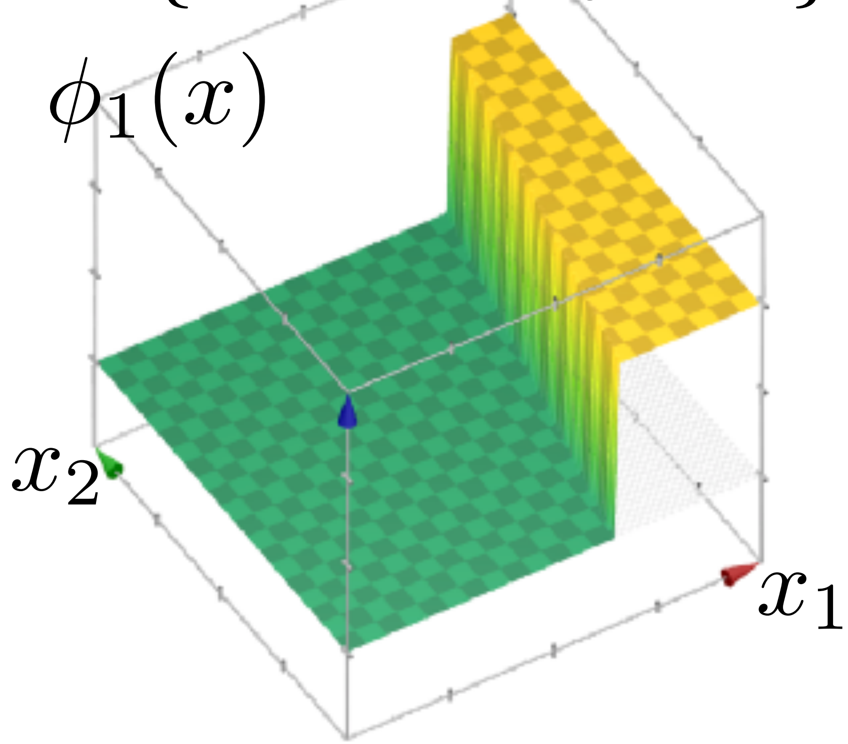
$$z = \theta^\top \phi(x) + \theta_0$$
$$= \theta_1 \phi_1(x) + \theta_2 \phi_2(x) + \theta_0$$
$$= 1 \cdot \phi_1(x) + 1 \cdot \phi_2(x) + (-0.5)$$



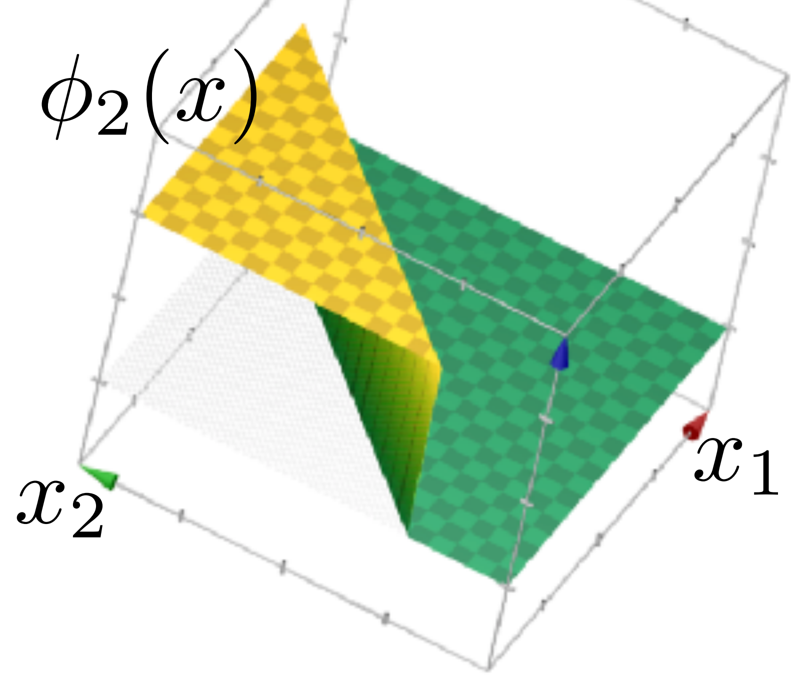


New features: step functions!

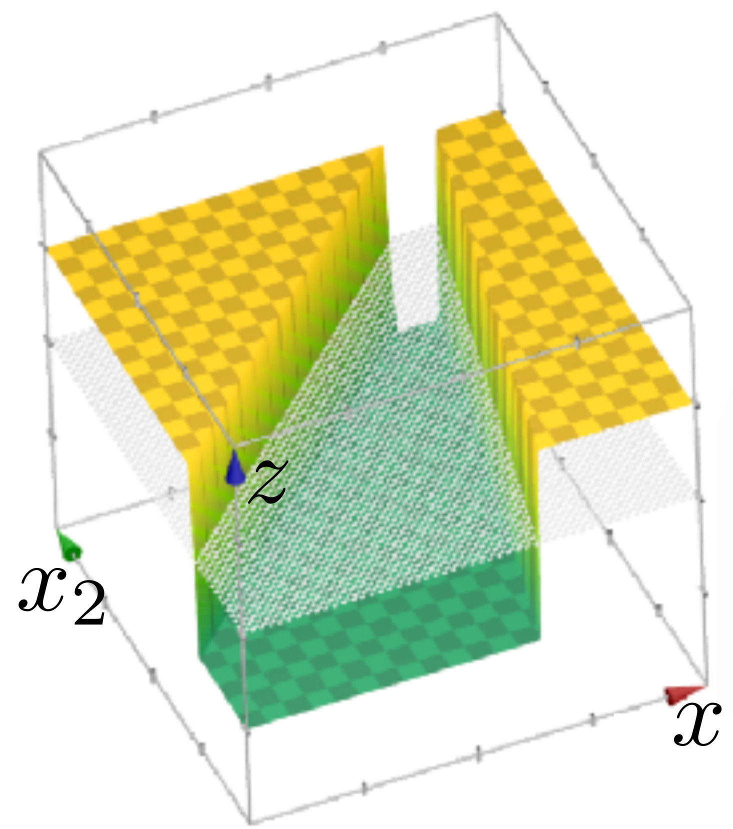
$\phi_1(x) = \mathbf{1}\{w^\top x + w_0 \geq 0\}$

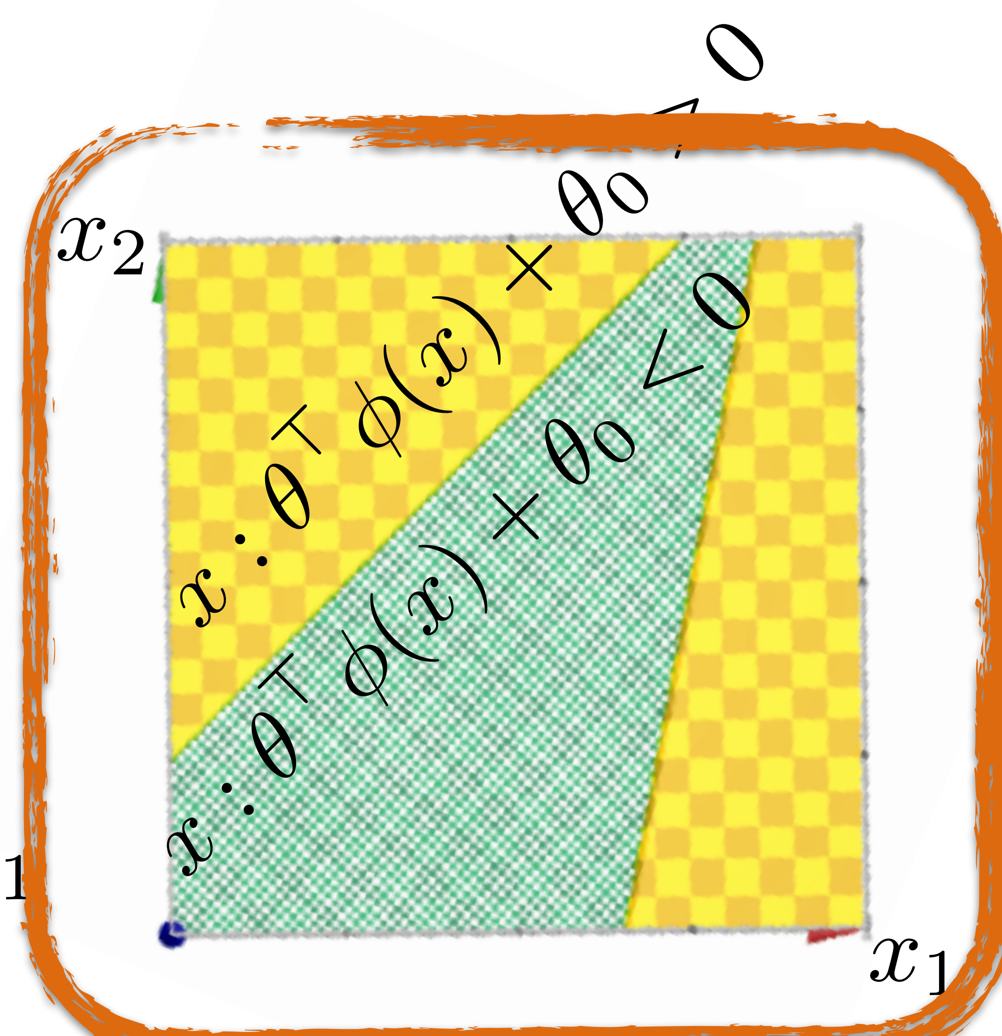


$\phi_2(x) = \mathbf{1}\{\tilde{w}^\top x + \tilde{w}_0 \geq 0\}$



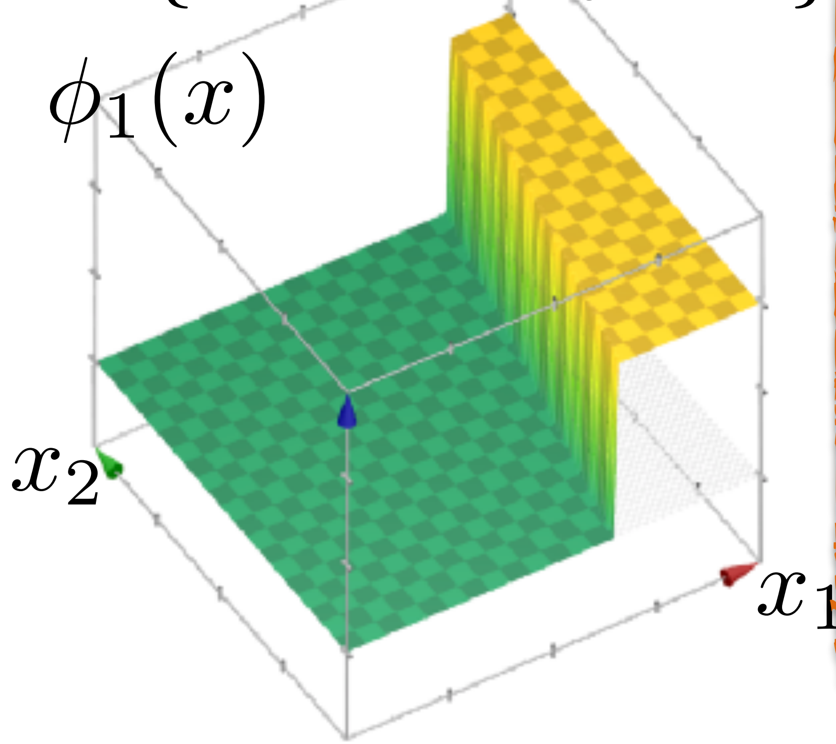
$$
\begin{aligned}
z &= \theta^\top \phi(x) + \theta_0 \\
&= \theta_1\phi_1(x) + \theta_2\phi_2(x) \\
&\quad + \theta_0 \\
&= 1 \cdot \phi_1(x) + 1 \cdot \phi_2(x) \\
&\quad + (-0.5)
\end{aligned}
$$



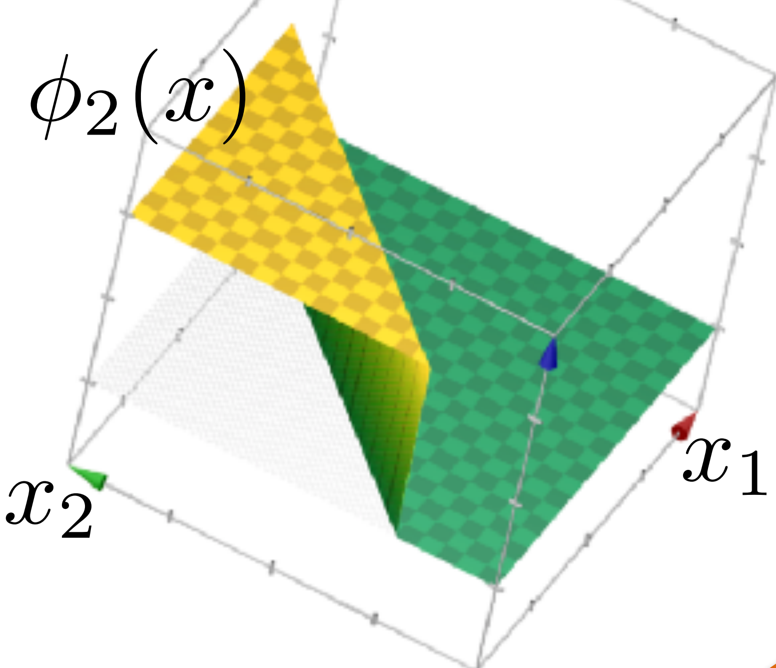


New features: step functions!

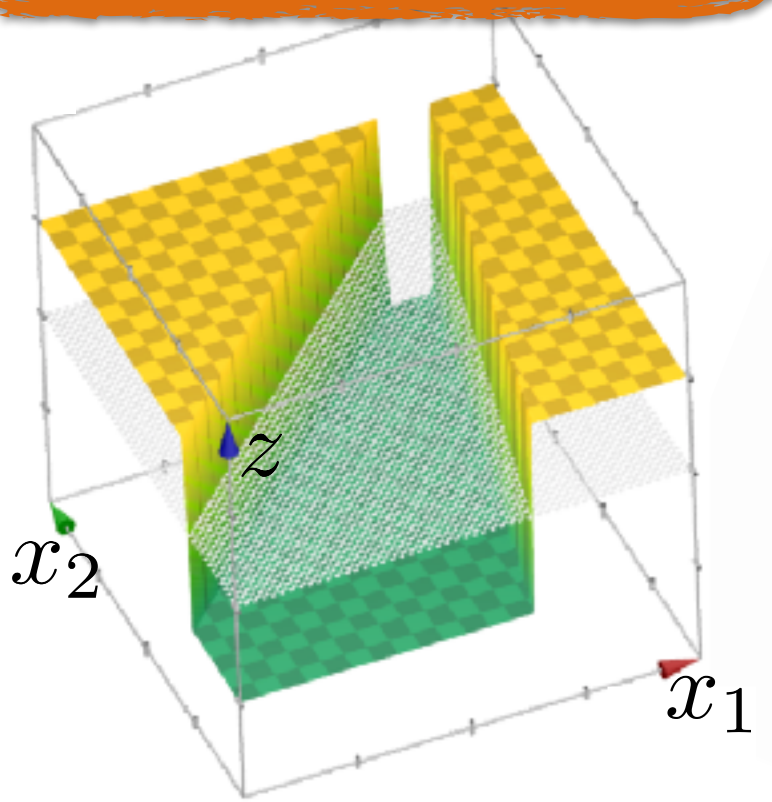
$\phi_1(x)$
$= \mathbf{1}\{w^\top x + w_0 \geq 0\}$

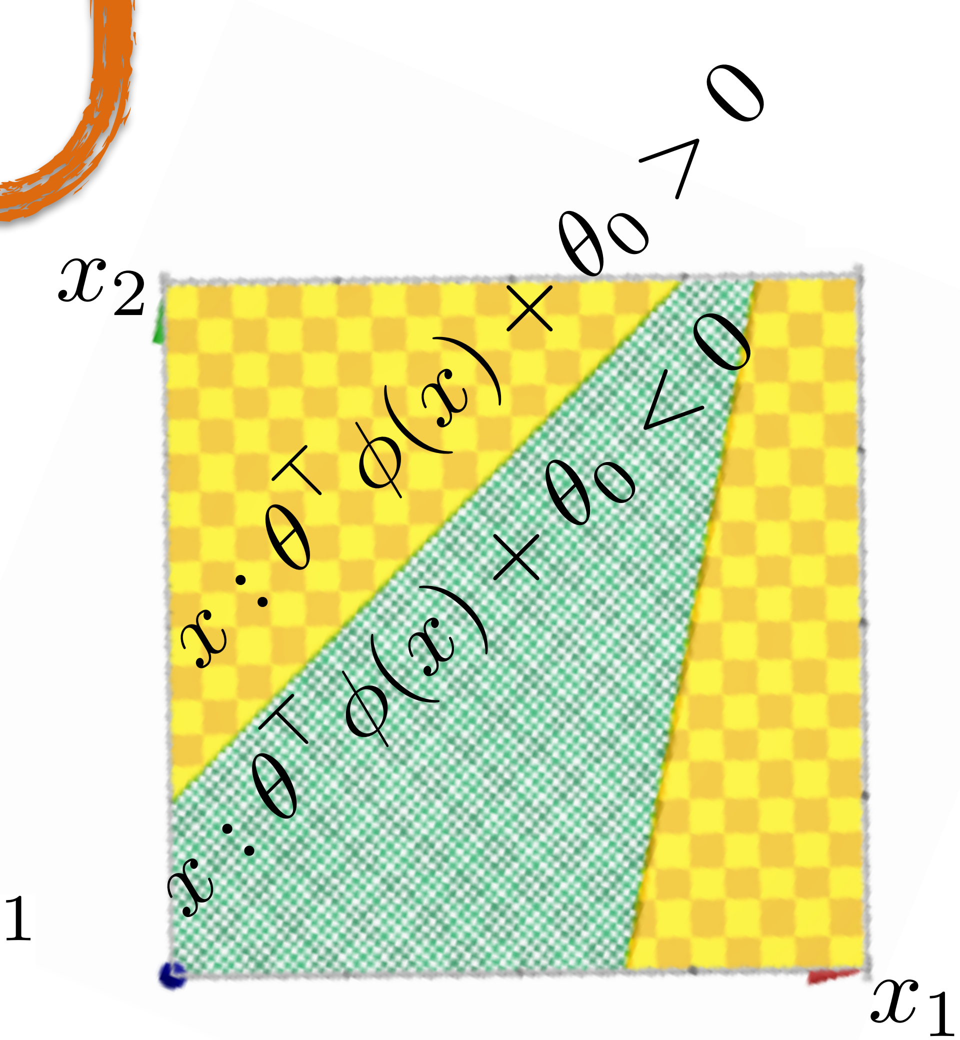


$\phi_2(x)$
$= \mathbf{1}\{\tilde{w}^\top x + \tilde{w}_0 \geq 0\}$



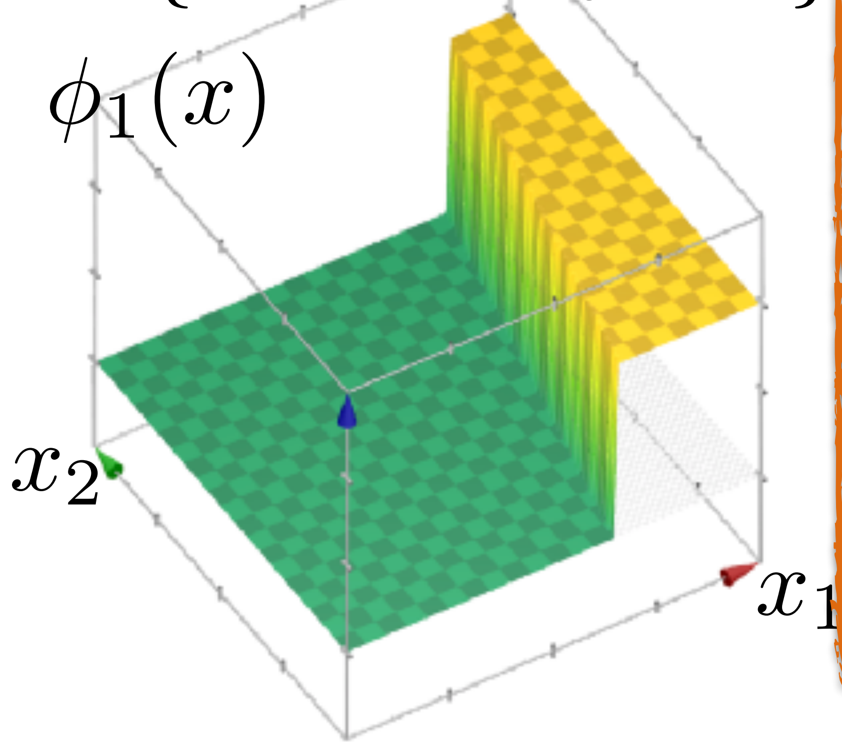
$$z = \theta^\top \phi(x) + \theta_0$$
$$= \theta_1\phi_1(x) + \theta_2\phi_2(x) + \theta_0$$
$$= 1 \cdot \phi_1(x) + 1 \cdot \phi_2(x) + (-0.5)$$



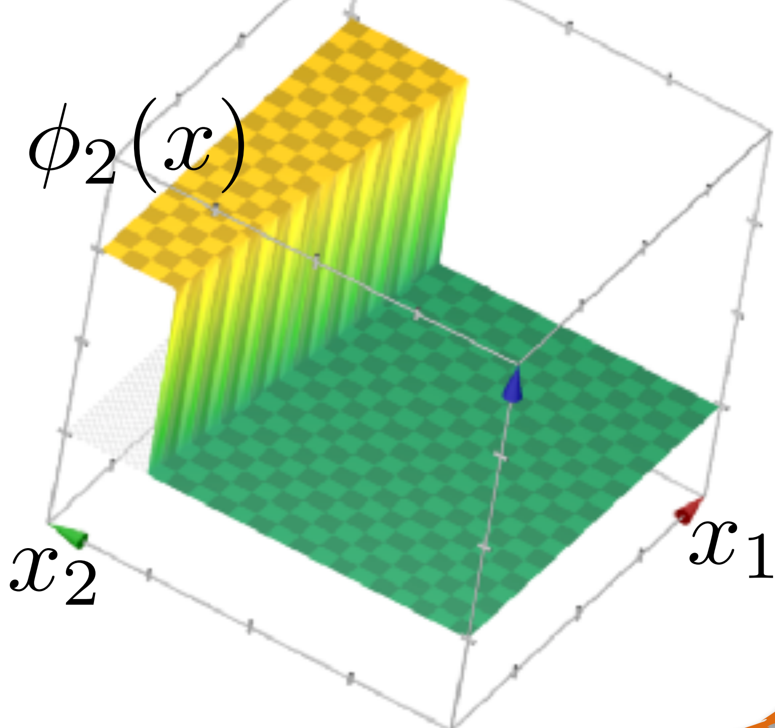


New features: step functions!

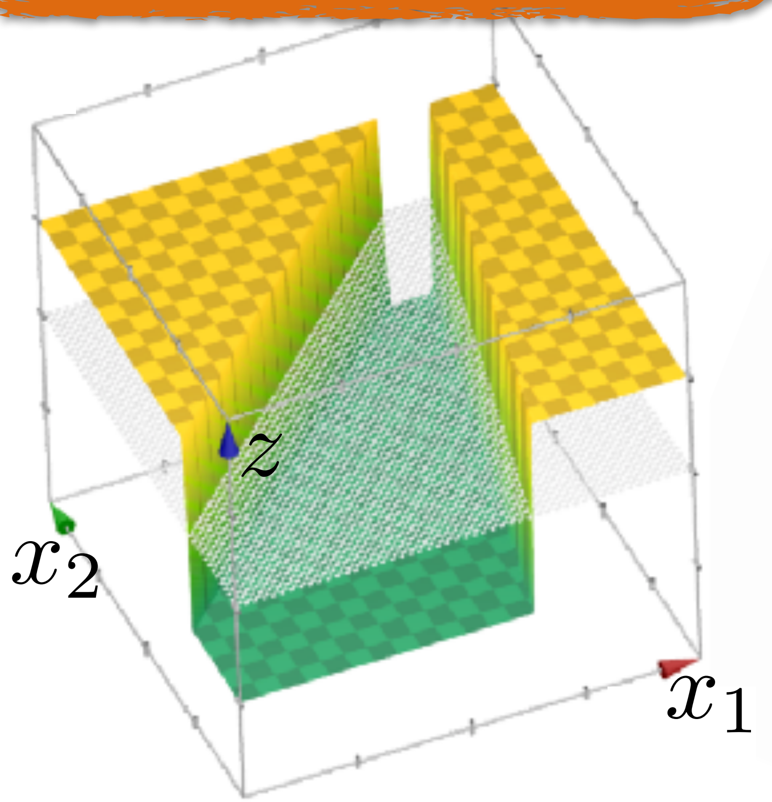
$\phi_1(x) = \mathbf{1}\{w^\top x + w_0 \geq 0\}$

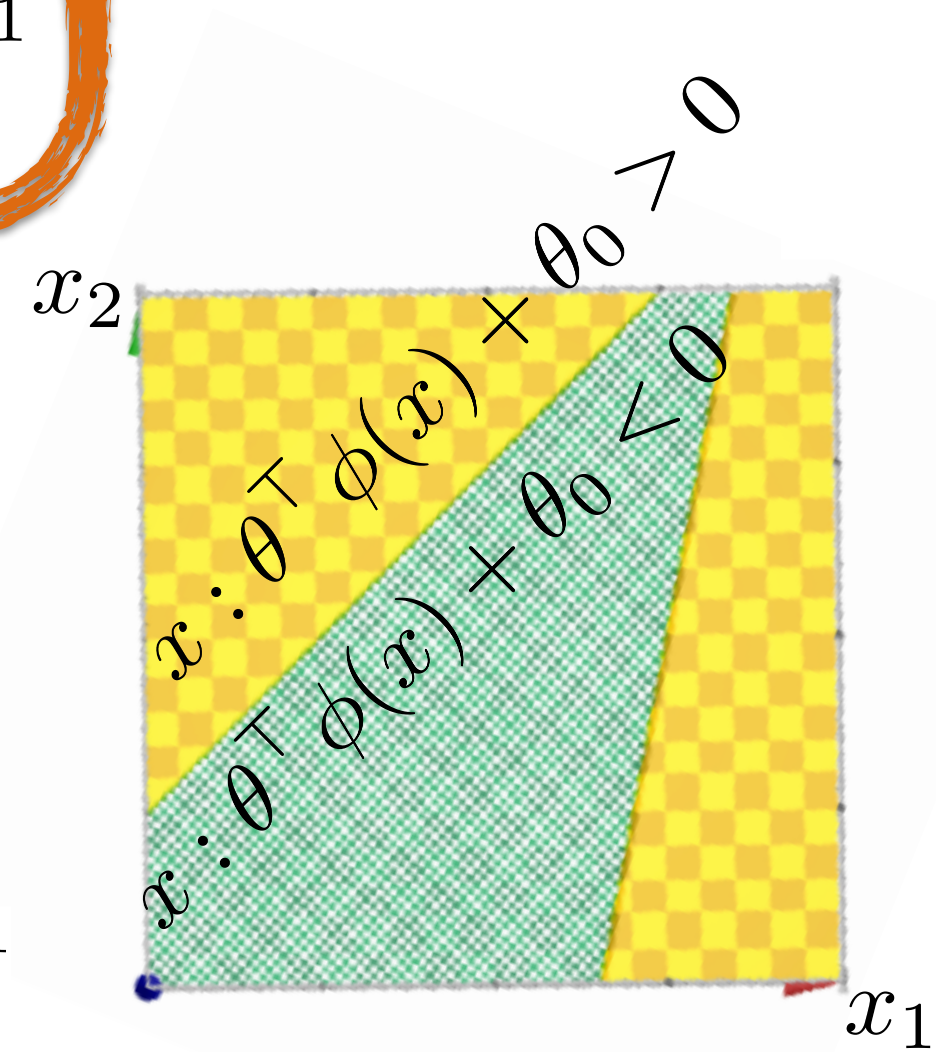


$\phi_2(x) = \mathbf{1}\{\tilde{w}^\top x + \tilde{w}_0 \geq 0\}$



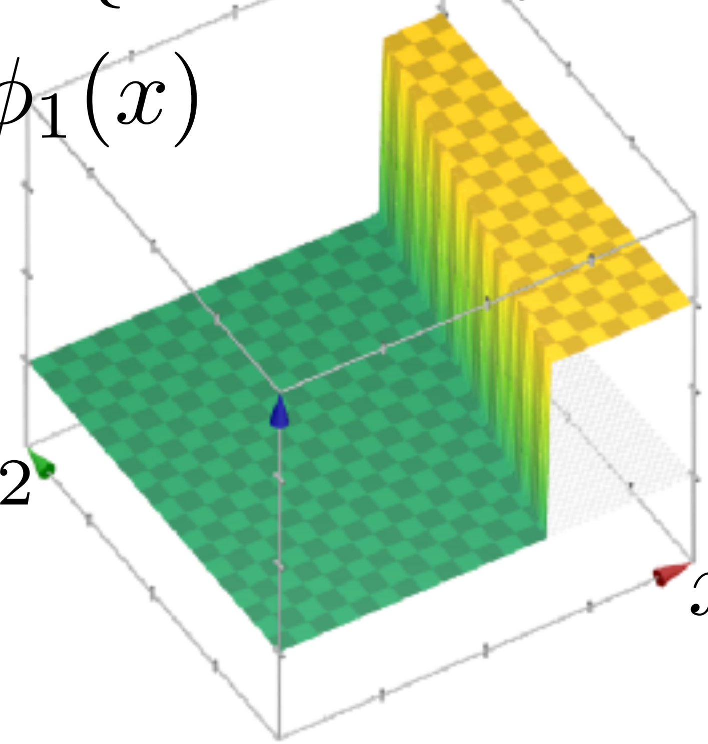
$$
\begin{aligned}
z &= \theta^\top \phi(x) + \theta_0 \\
&= \theta_1 \phi_1(x) + \theta_2 \phi_2(x) \\
&\quad + \theta_0 \\
&= 1 \cdot \phi_1(x) + 1 \cdot \phi_2(x) \\
&\quad + (-0.5)
\end{aligned}
$$



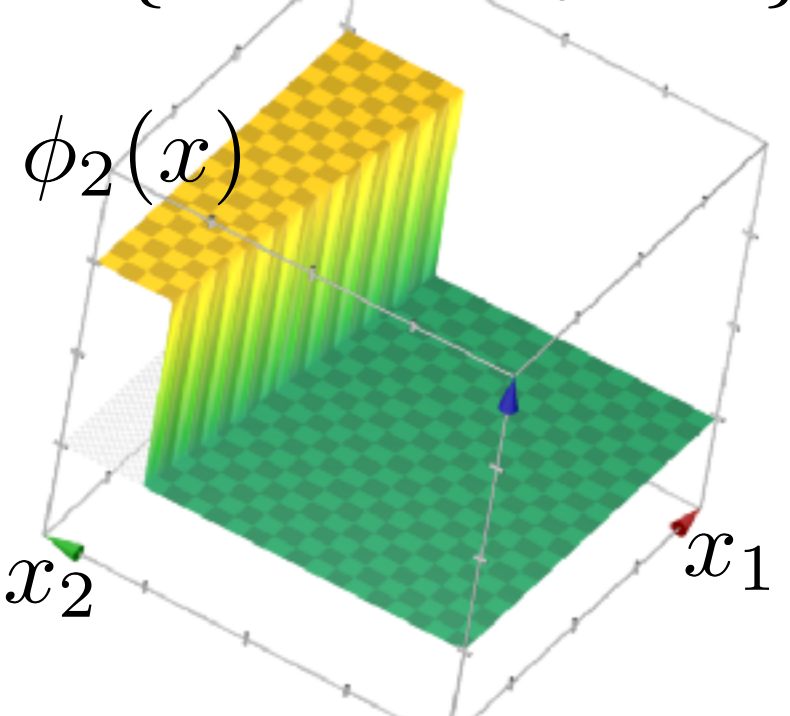


New features: step functions!

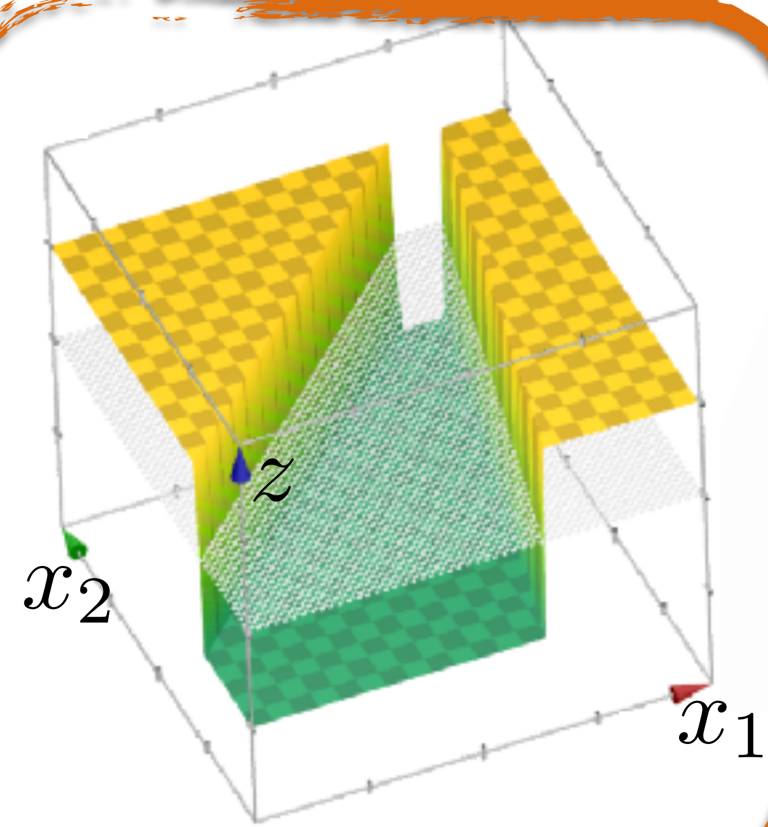
$\phi_1(x) = \mathbf{1}\{w^\top x + w_0 \geq 0\}$

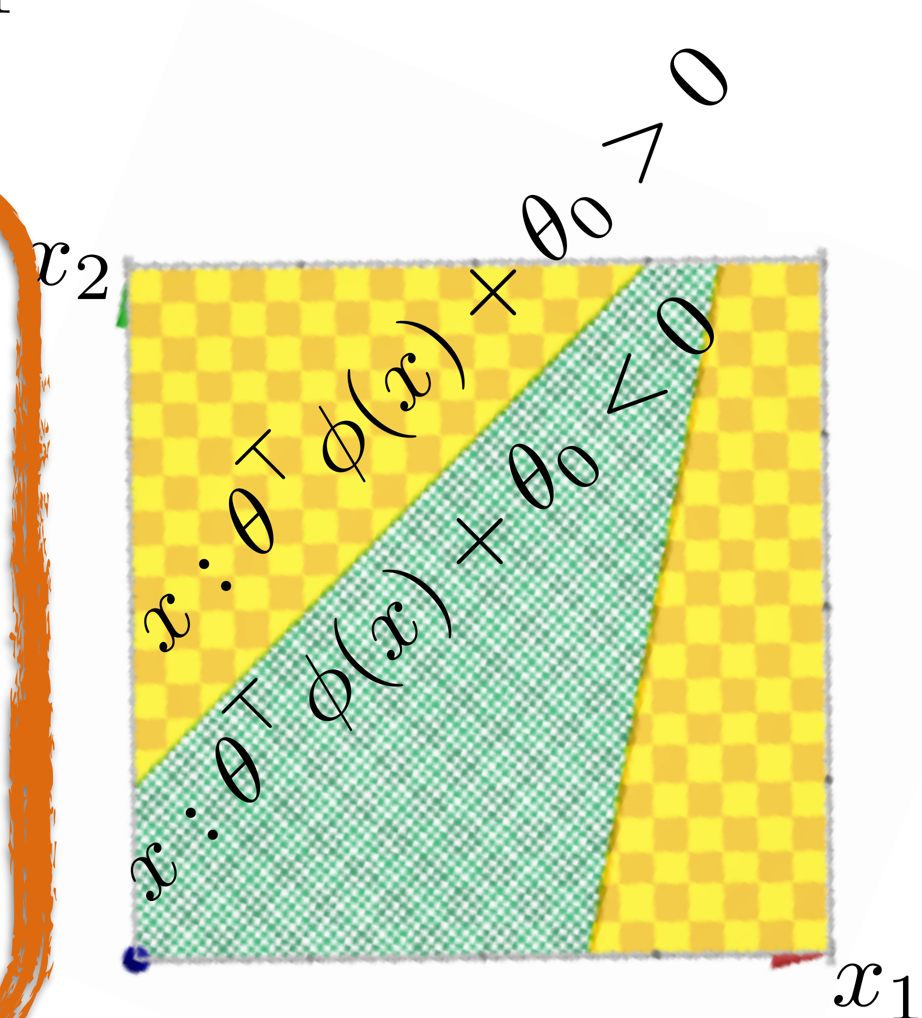


$\phi_2(x) = \mathbf{1}\{\tilde{w}^\top x + \tilde{w}_0 \geq 0\}$



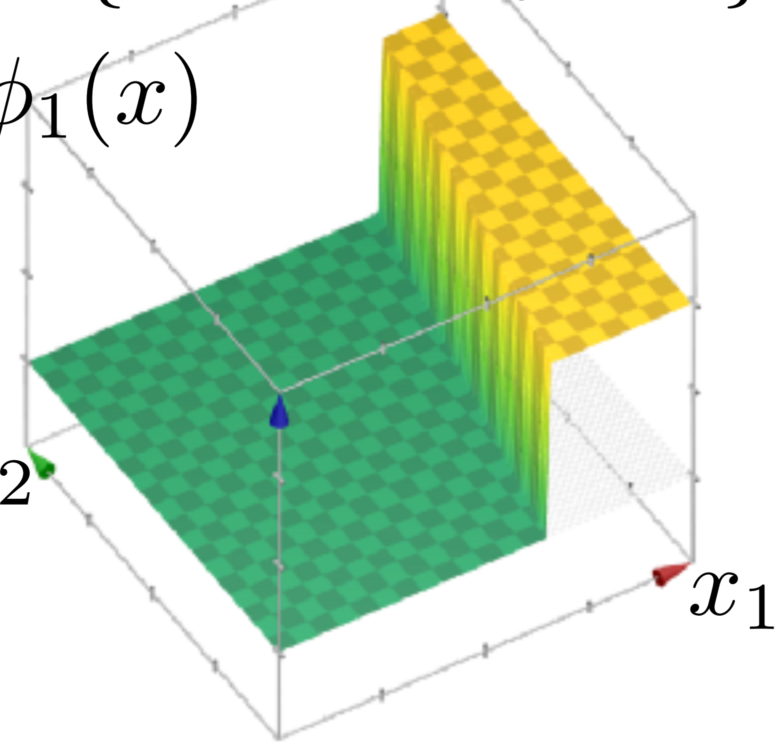
$$z = \theta^\top \phi(x) + \theta_0$$
$$= \theta_1\phi_1(x) + \theta_2\phi_2(x) + \theta_0$$
$$= 1 \cdot \phi_1(x) + 1 \cdot \phi_2(x) + (-0.5)$$



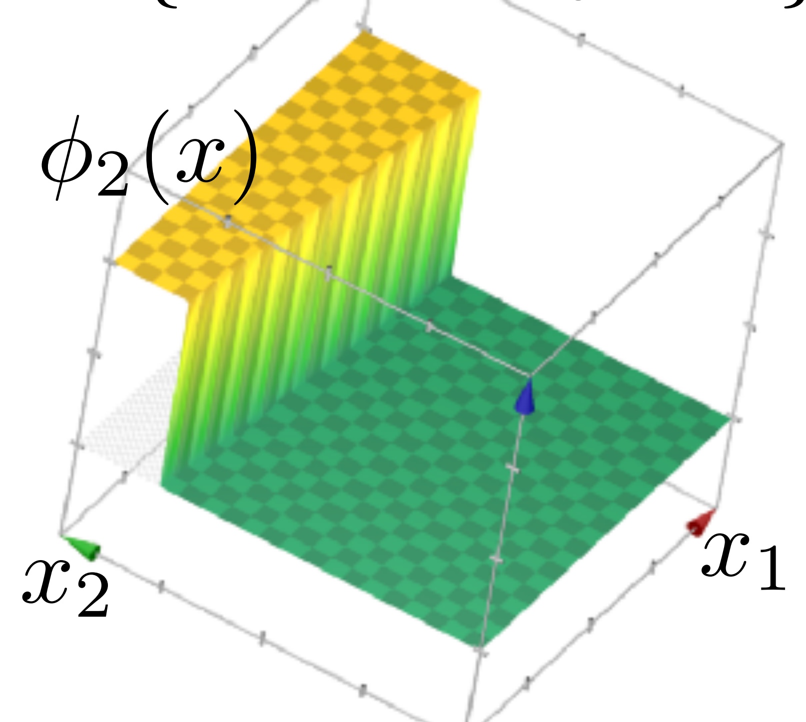


New features: step functions!

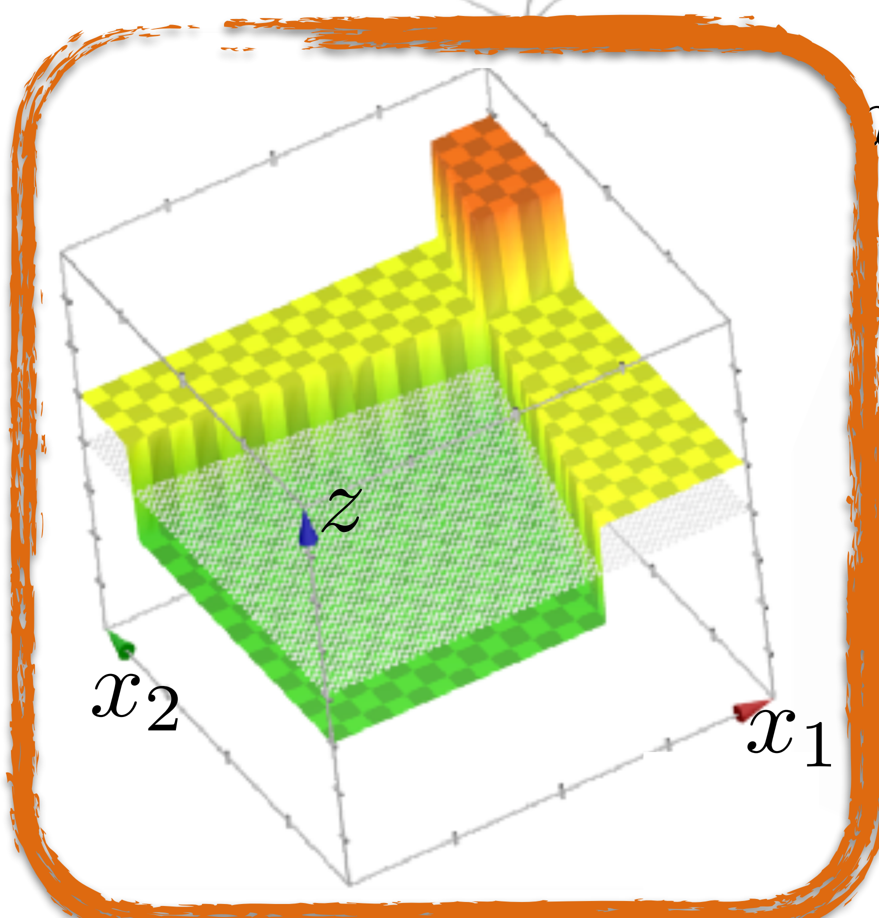
$\phi_1(x)$
$= \mathbf{1}\{w^\top x + w_0 \geq 0\}$

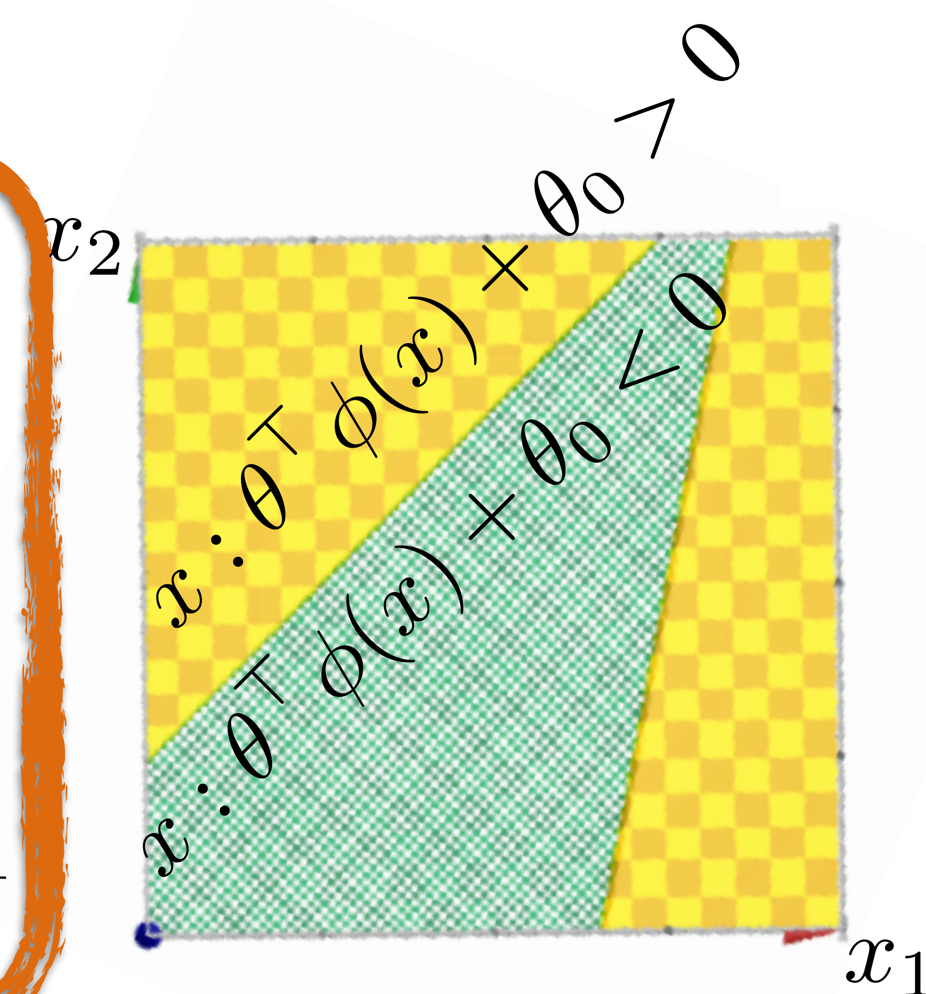


$\phi_2(x)$
$= \mathbf{1}\{\tilde{w}^\top x + \tilde{w}_0 \geq 0\}$



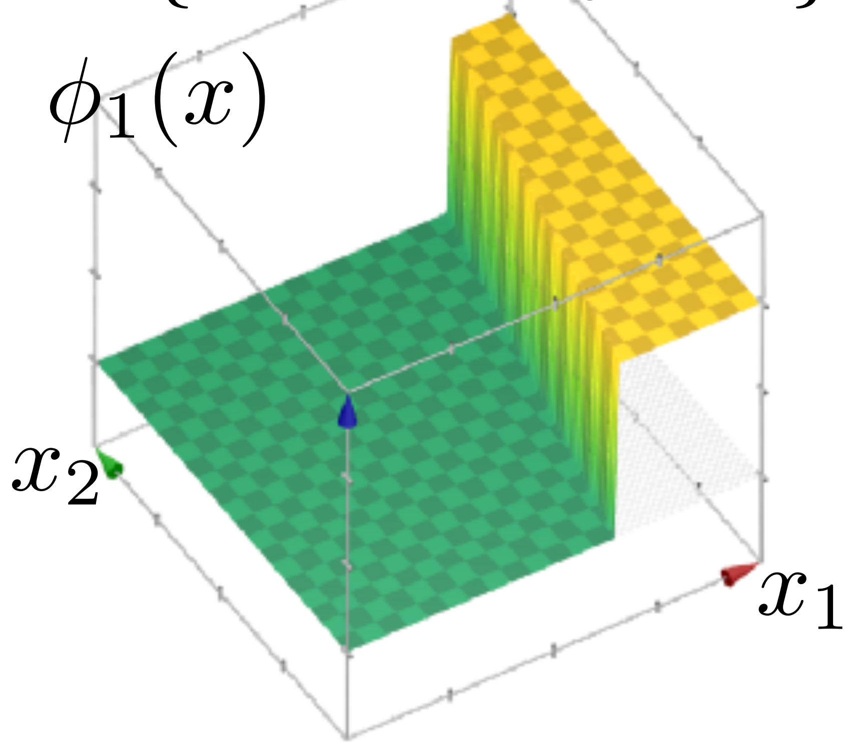
$$z = \theta^\top \phi(x) + \theta_0$$
$$= \theta_1\phi_1(x) + \theta_2\phi_2(x) + \theta_0$$
$$= 1 \cdot \phi_1(x) + 1 \cdot \phi_2(x) + (-0.5)$$



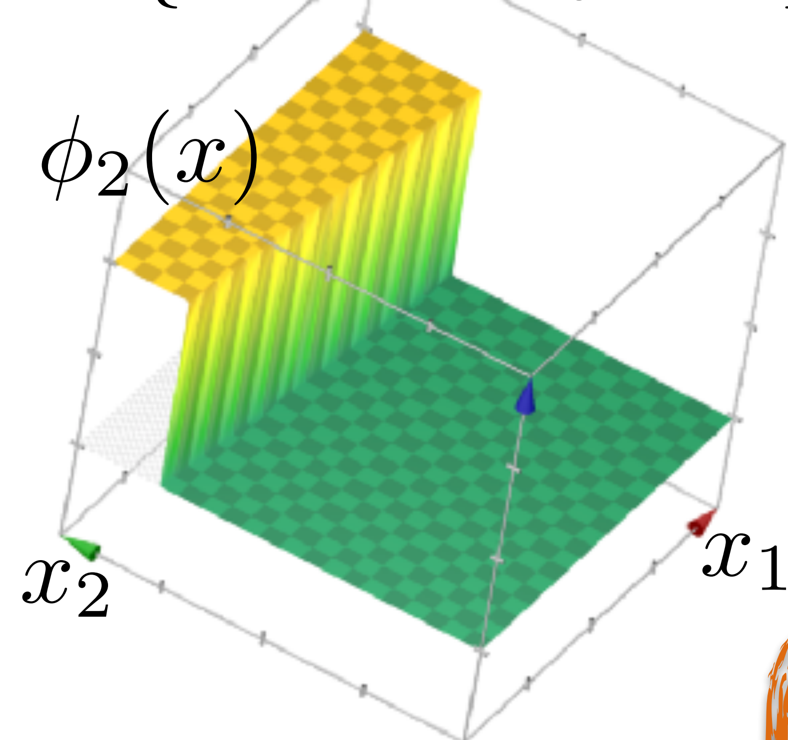


New features: step functions!

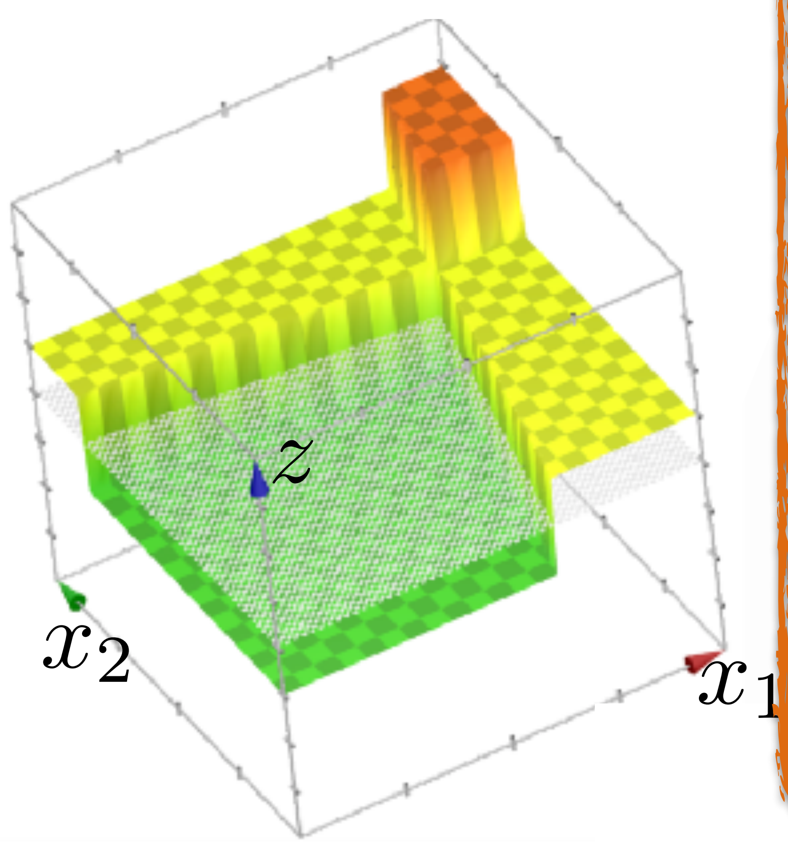
$\phi_1(x) = \mathbf{1}\{w^\top x + w_0 \geq 0\}$

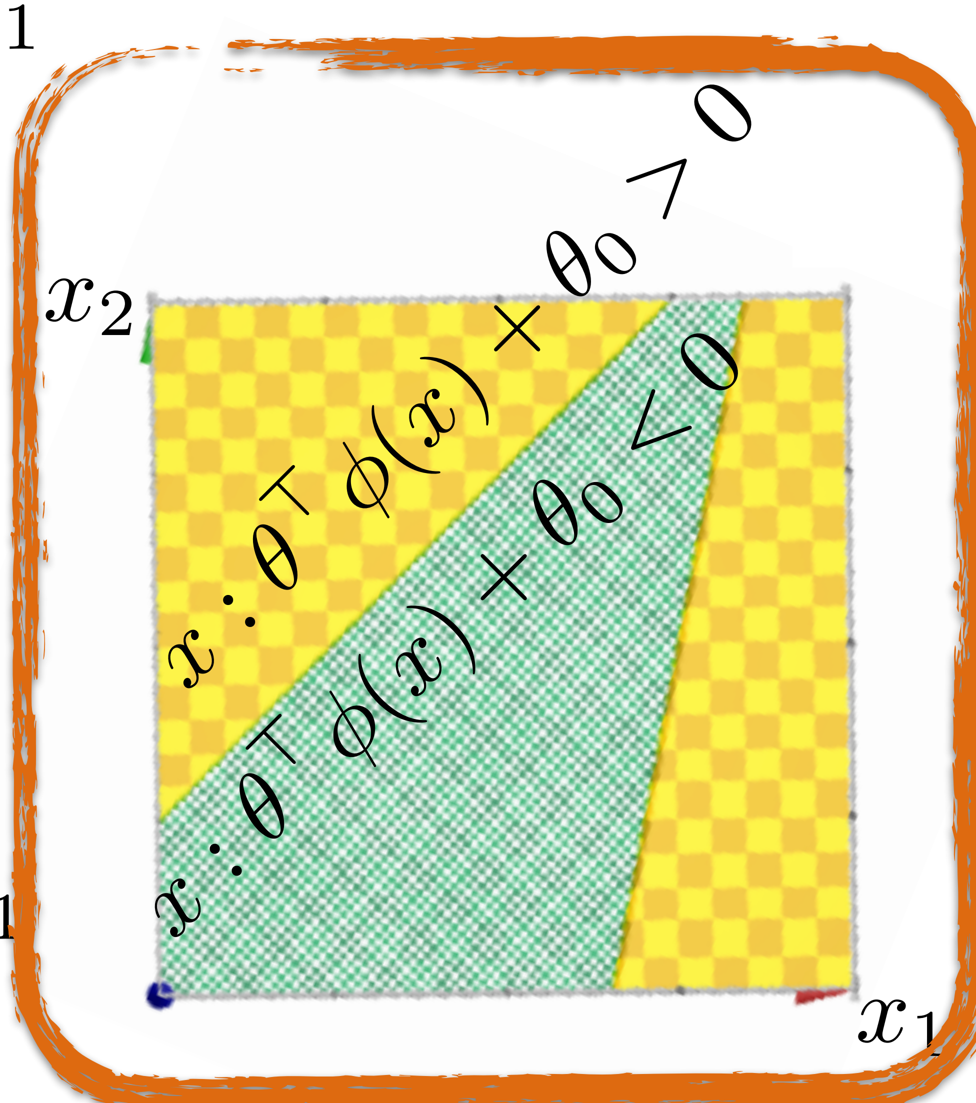


$\phi_2(x) = \mathbf{1}\{\tilde{w}^\top x + \tilde{w}_0 \geq 0\}$



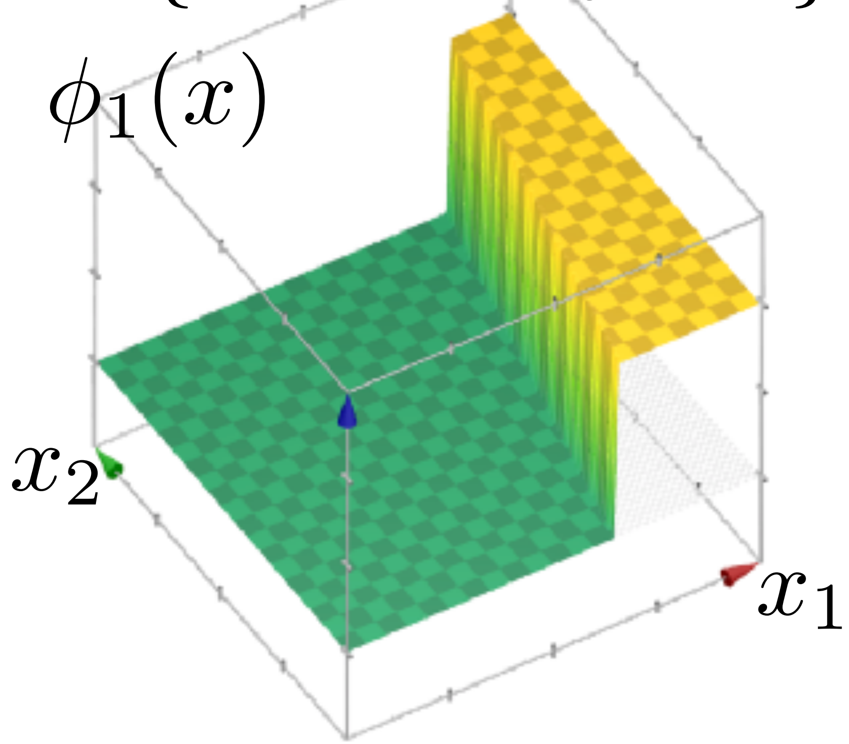
$$z = \theta^\top \phi(x) + \theta_0$$
$$= \theta_1 \phi_1(x) + \theta_2 \phi_2(x) + \theta_0$$
$$= 1 \cdot \phi_1(x) + 1 \cdot \phi_2(x) + (-0.5)$$



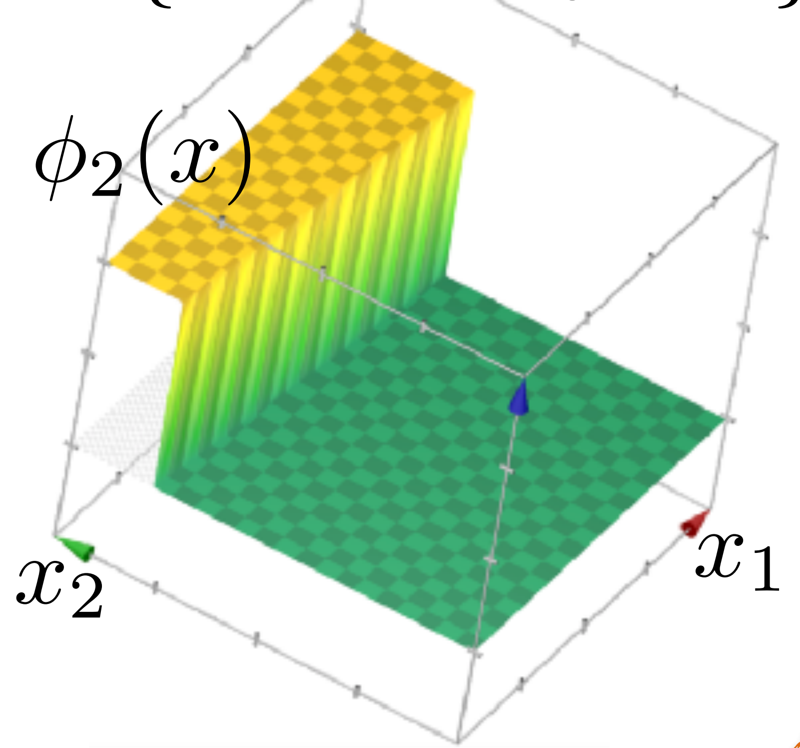


New features: step functions!

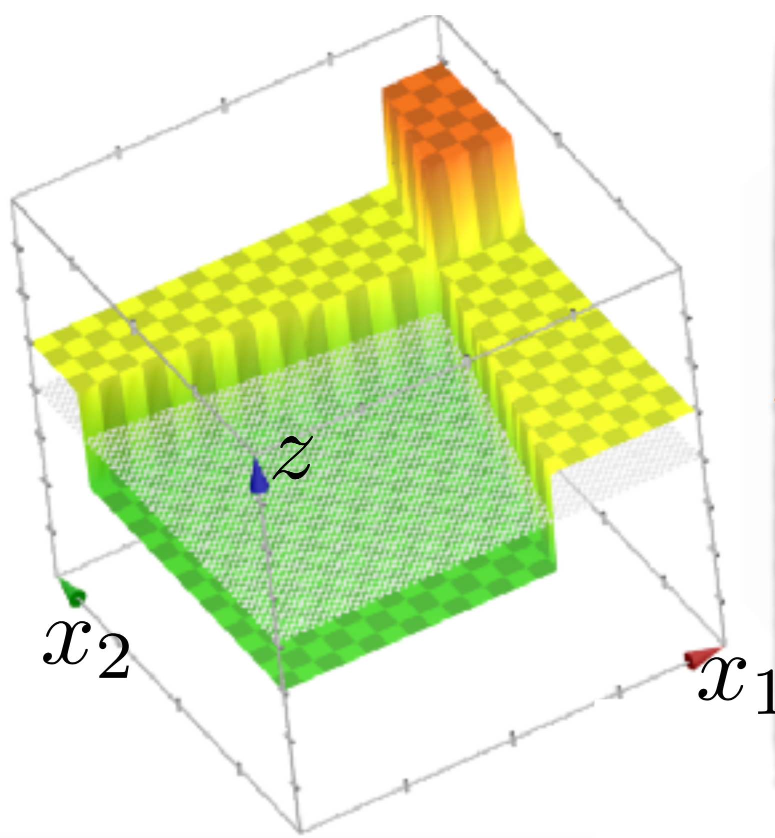
$\phi_1(x) = \mathbf{1}\{w^\top x + w_0 \geq 0\}$

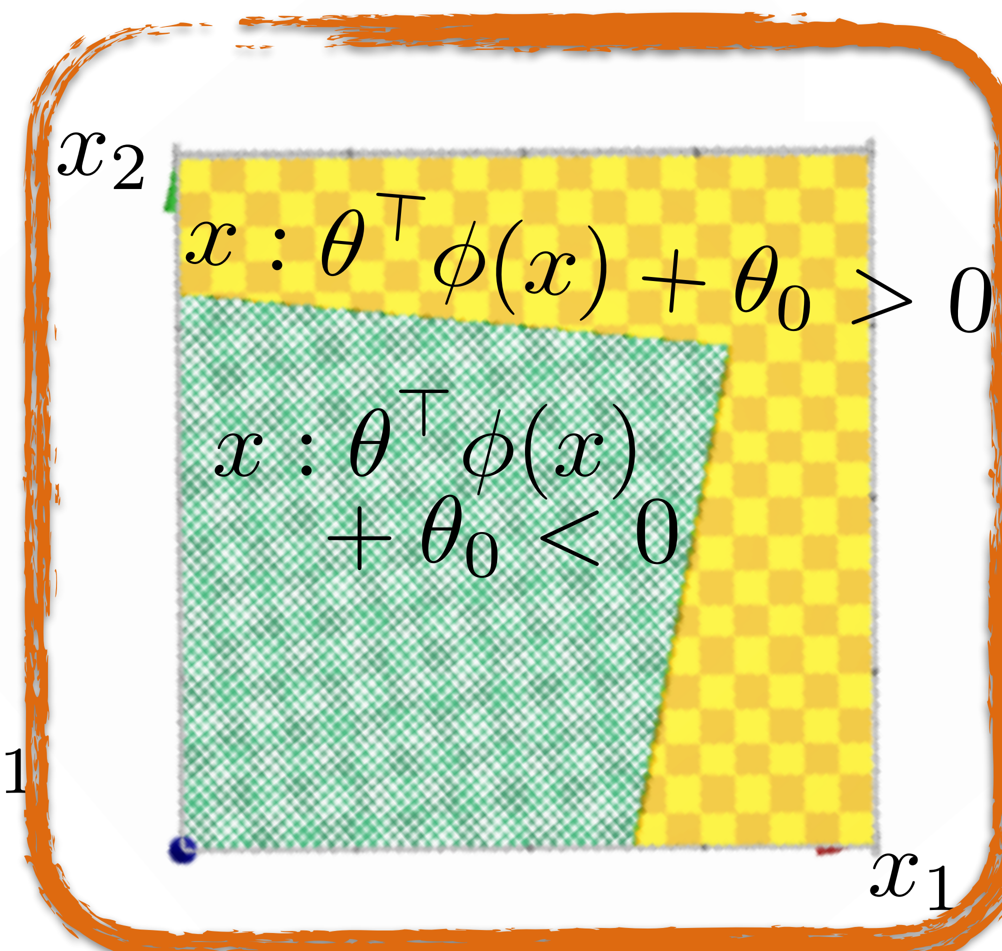


$\phi_2(x) = \mathbf{1}\{\tilde{w}^\top x + \tilde{w}_0 \geq 0\}$



$$z = \theta^\top \phi(x) + \theta_0$$
$$= \theta_1 \phi_1(x) + \theta_2 \phi_2(x) + \theta_0$$
$$= 1 \cdot \phi_1(x) + 1 \cdot \phi_2(x) + (-0.5)$$

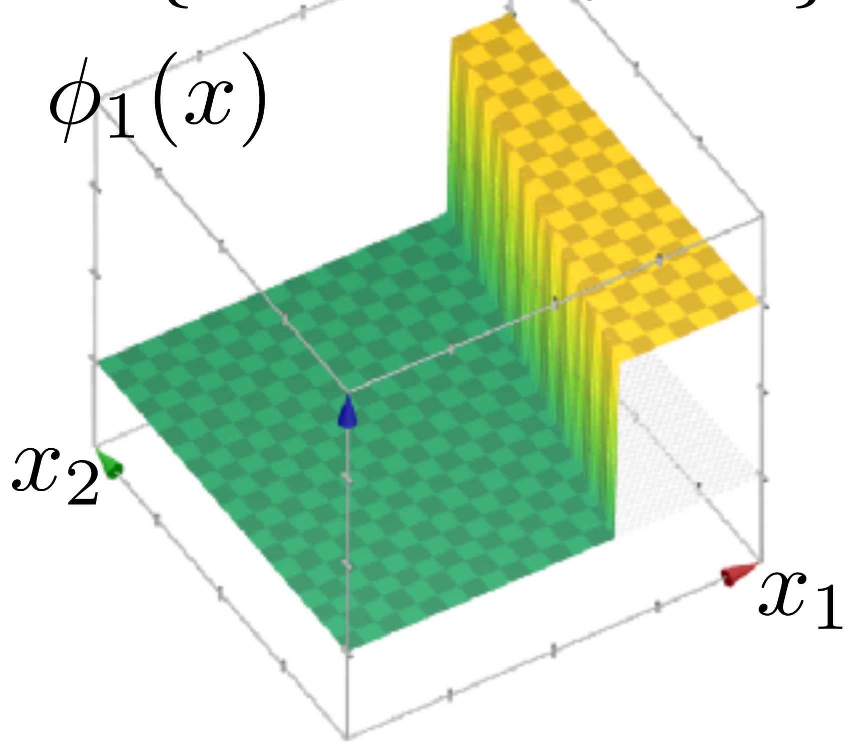


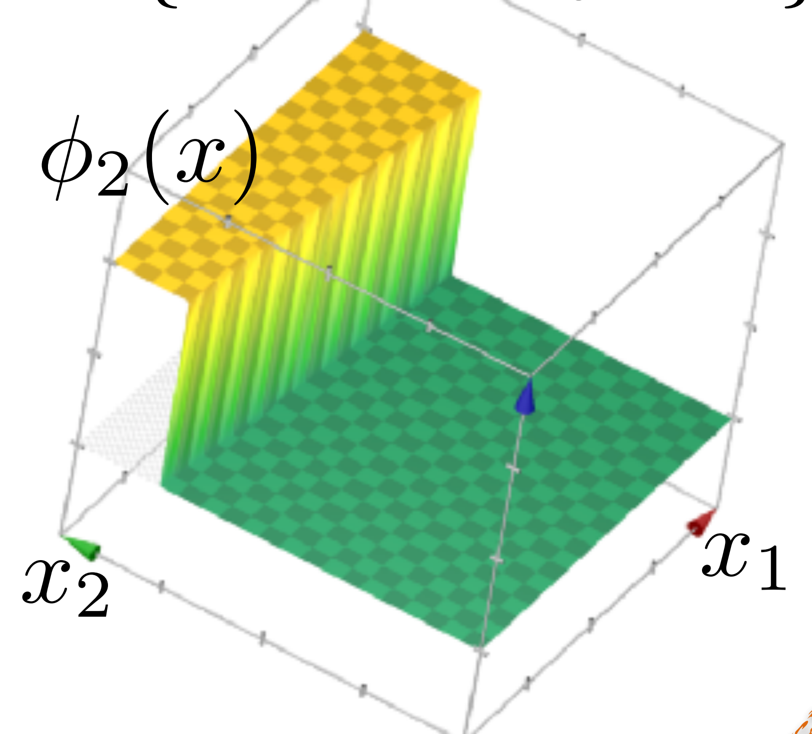


New features: step functions!

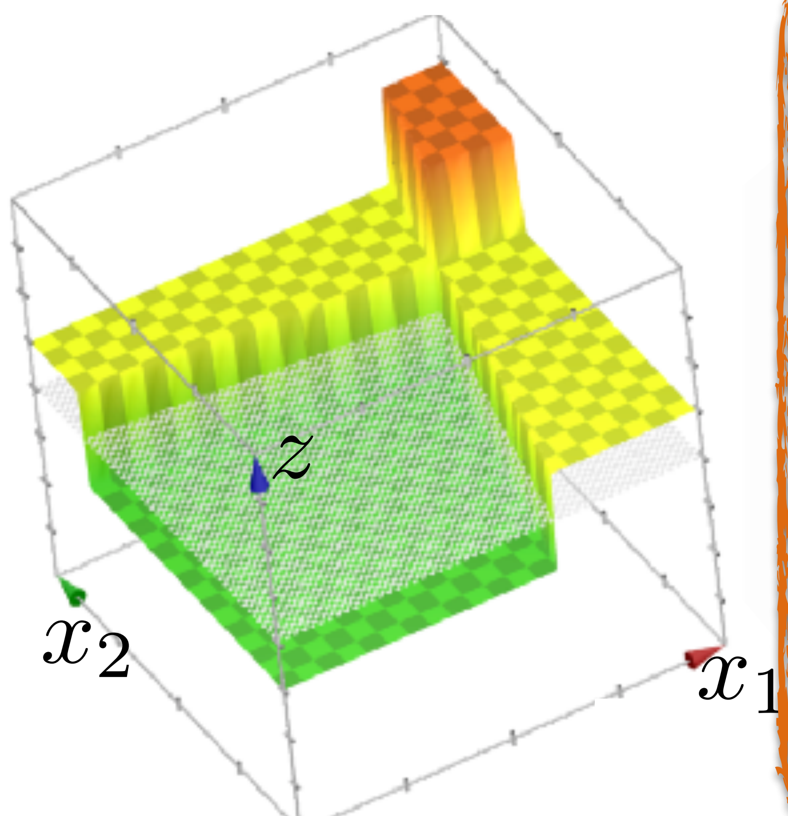
$\phi_1(x) = \mathbf{1}\{w^\top x + w_0 \geq 0\}$

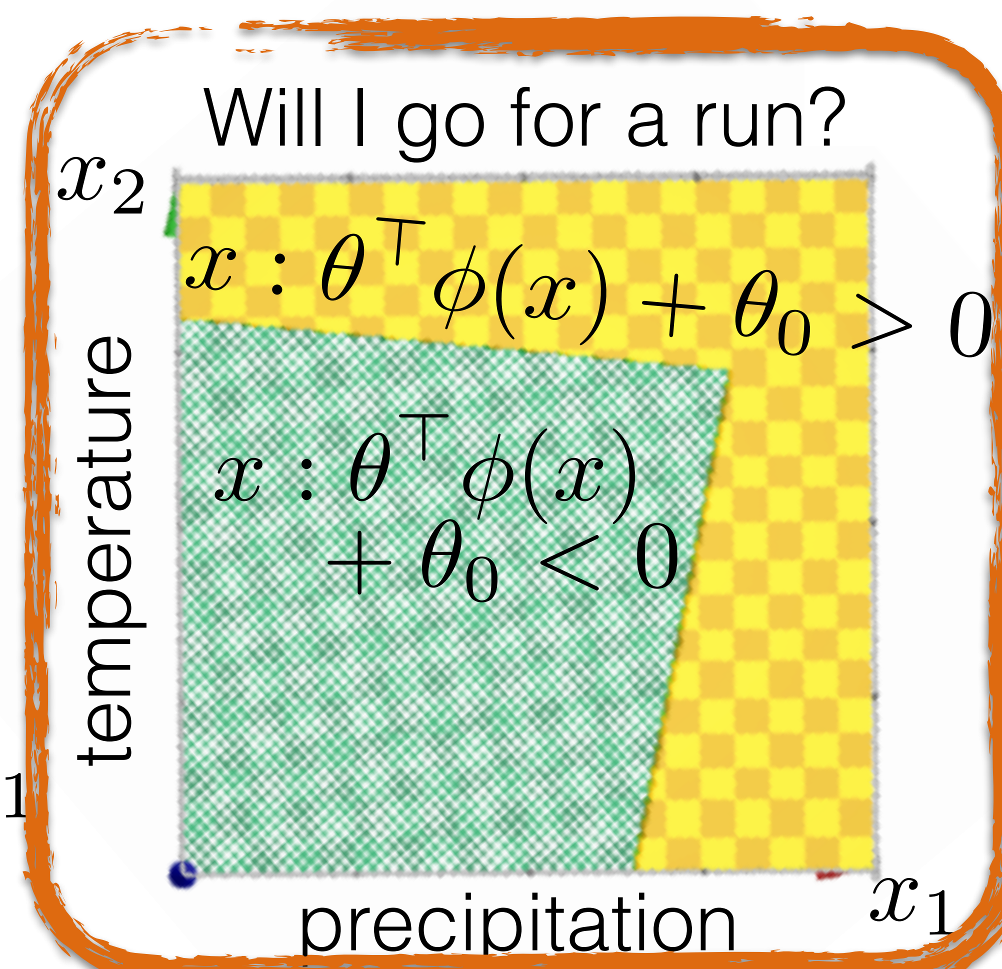
$\phi_2(x) = \mathbf{1}\{\tilde{w}^\top x + \tilde{w}_0 \geq 0\}$





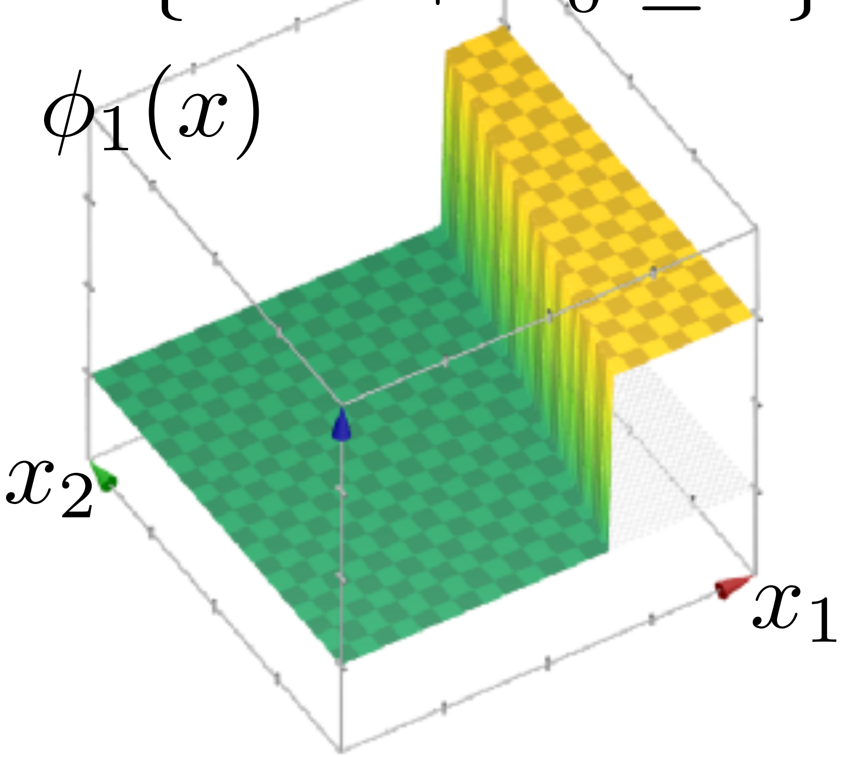
$$z = \theta^\top \phi(x) + \theta_0$$
$$= \theta_1 \phi_1(x) + \theta_2 \phi_2(x) + \theta_0$$
$$= 1 \cdot \phi_1(x) + 1 \cdot \phi_2(x) + (-0.5)$$



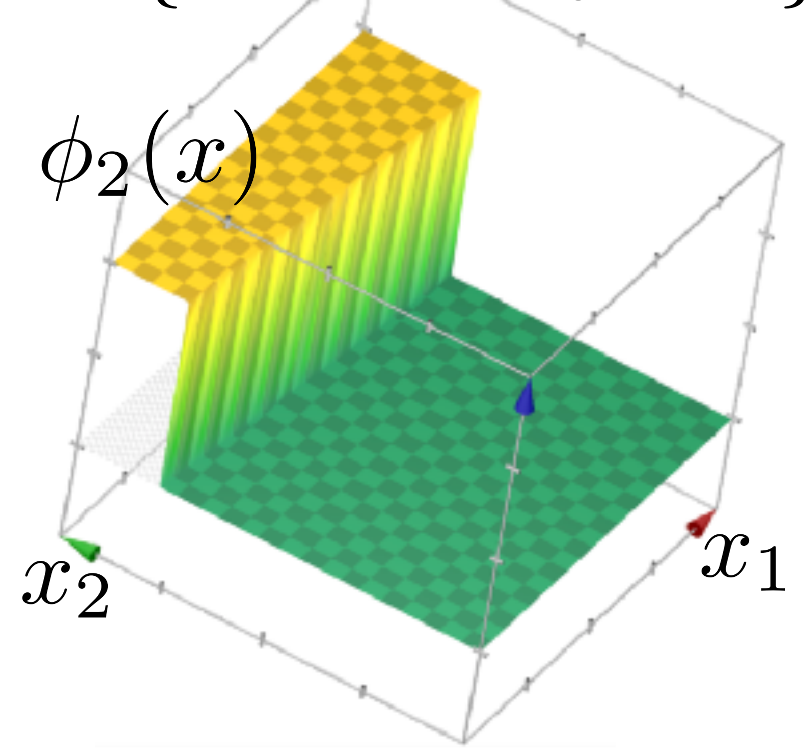


New features: step functions!

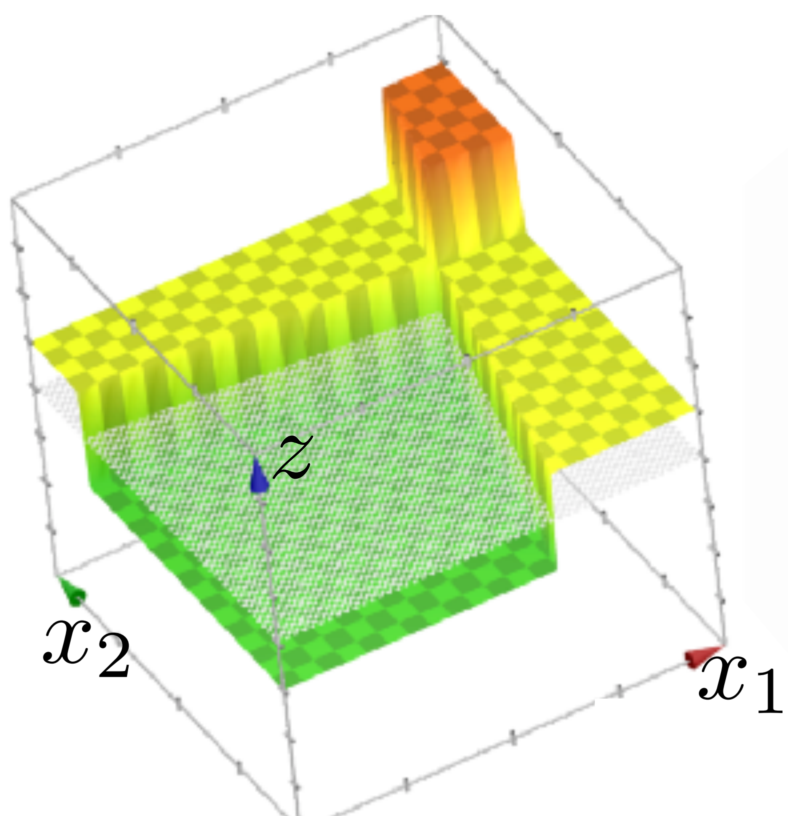
$\phi_1(x) = \mathbf{1}\{w^\top x + w_0 \geq 0\}$



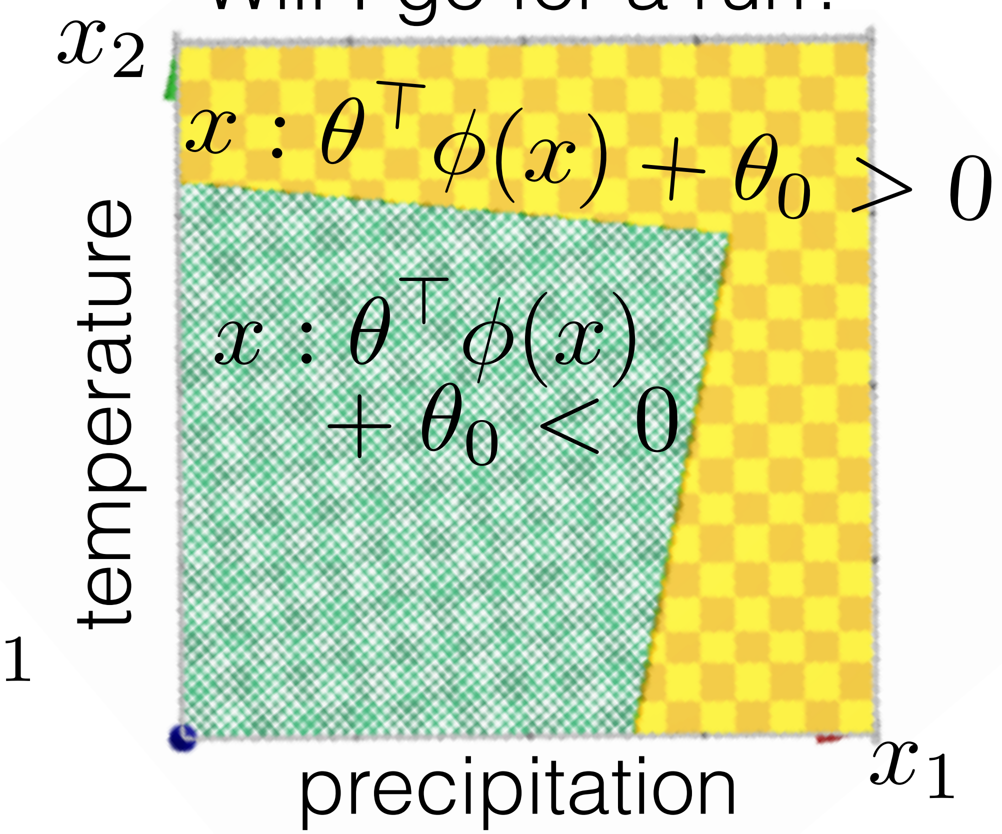
$\phi_2(x) = \mathbf{1}\{\tilde{w}^\top x + \tilde{w}_0 \geq 0\}$



$$z = \theta^\top \phi(x) + \theta_0$$
$$= \theta_1 \phi_1(x) + \theta_2 \phi_2(x) + \theta_0$$
$$= 1 \cdot \phi_1(x) + 1 \cdot \phi_2(x) + (-0.5)$$

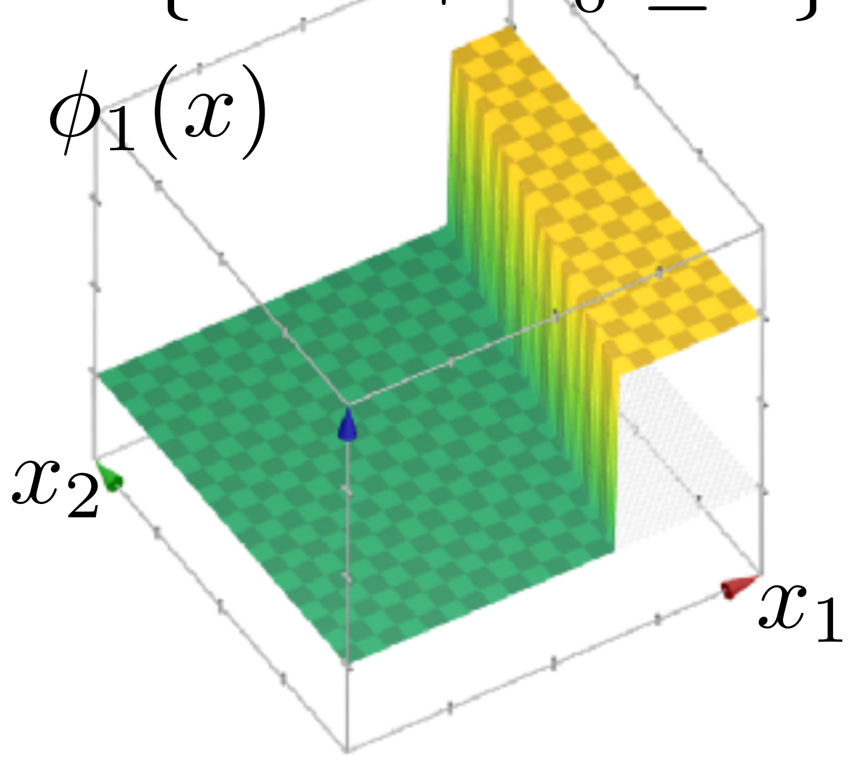


Will I go for a run?

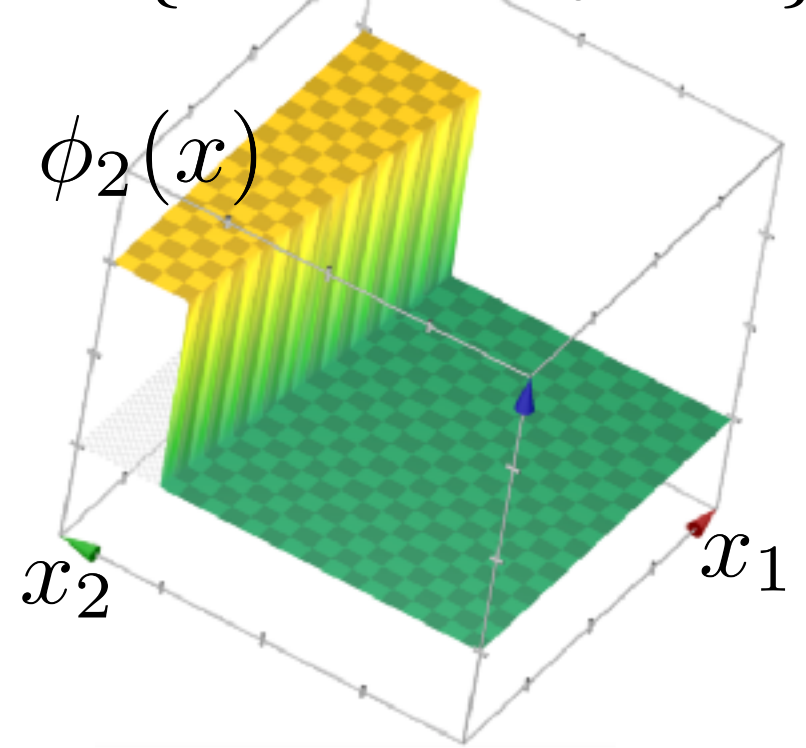


New features: step functions!

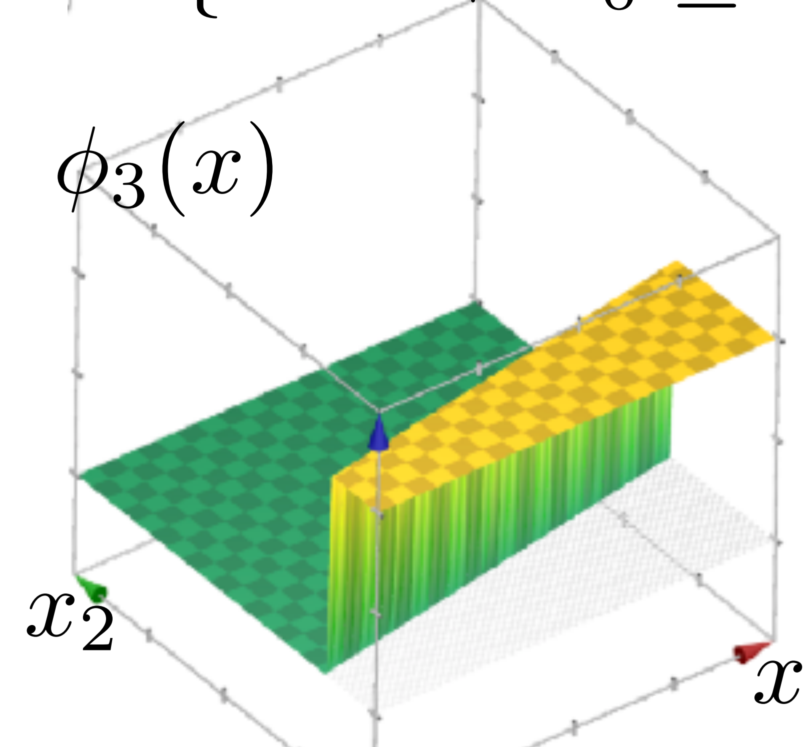
$\phi_1(x) = \mathbf{1}\{w^\top x + w_0 \geq 0\}$



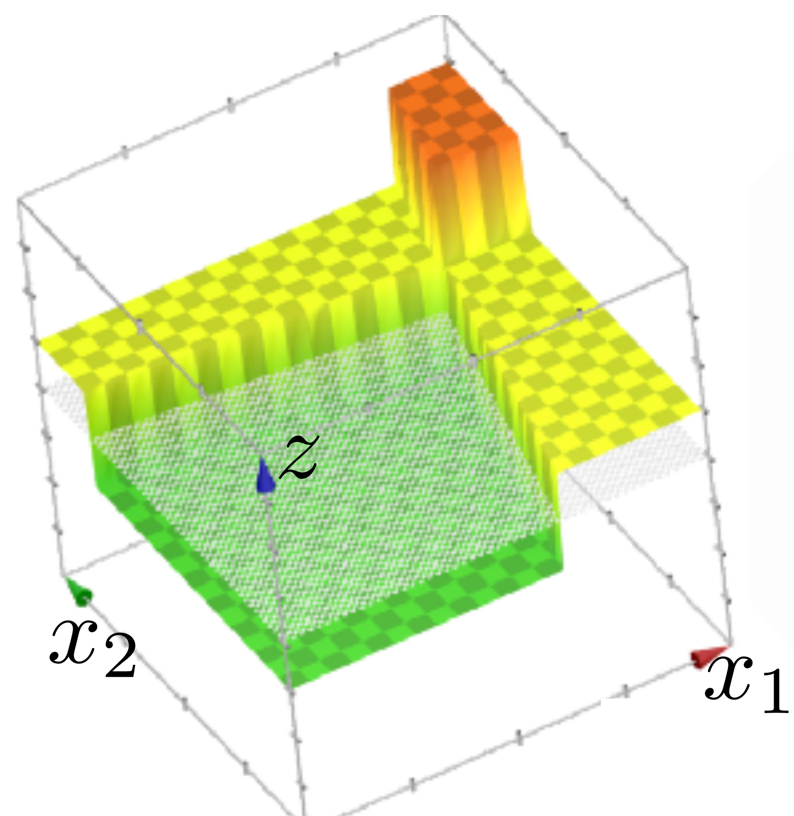
$\phi_2(x) = \mathbf{1}\{\tilde{w}^\top x + \tilde{w}_0 \geq 0\}$



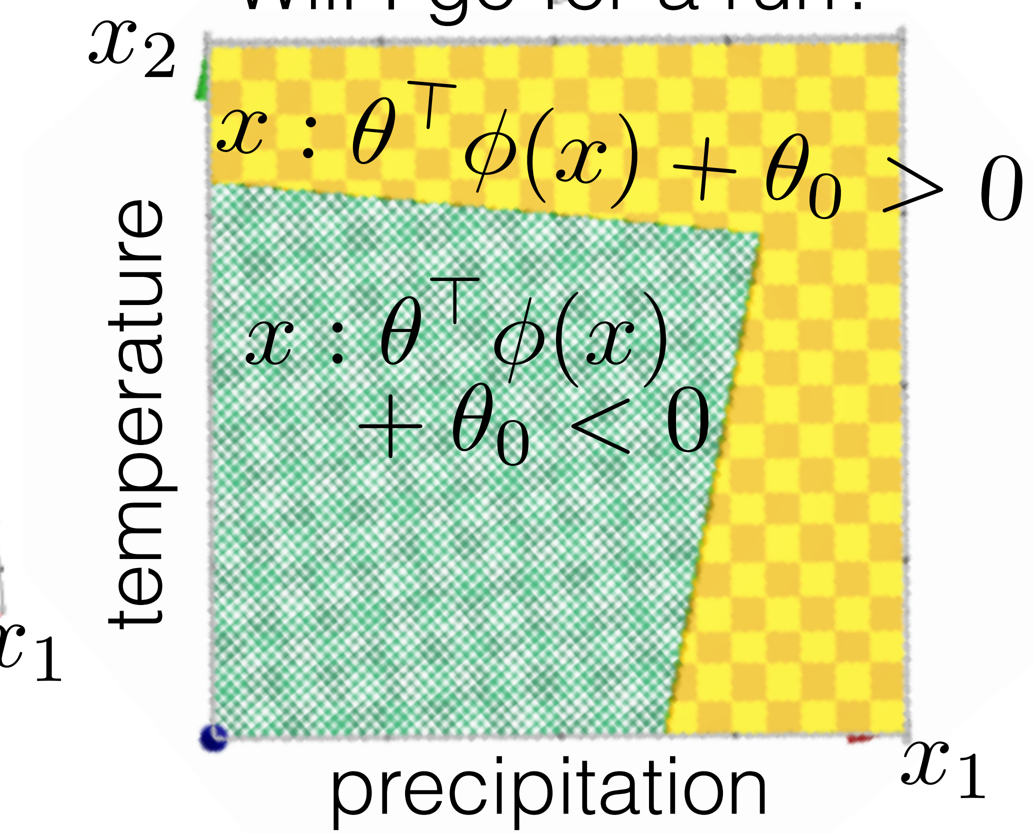
$\phi_3(x) = \mathbf{1}\{\tilde{\tilde{w}}^\top x + \tilde{\tilde{w}}_0 \geq 0\}$



$$z = \theta^\top \phi(x) + \theta_0$$
$$= \theta_1 \phi_1(x) + \theta_2 \phi_2(x) + \theta_0$$
$$= 1 \cdot \phi_1(x) + 1 \cdot \phi_2(x) + (-0.5)$$

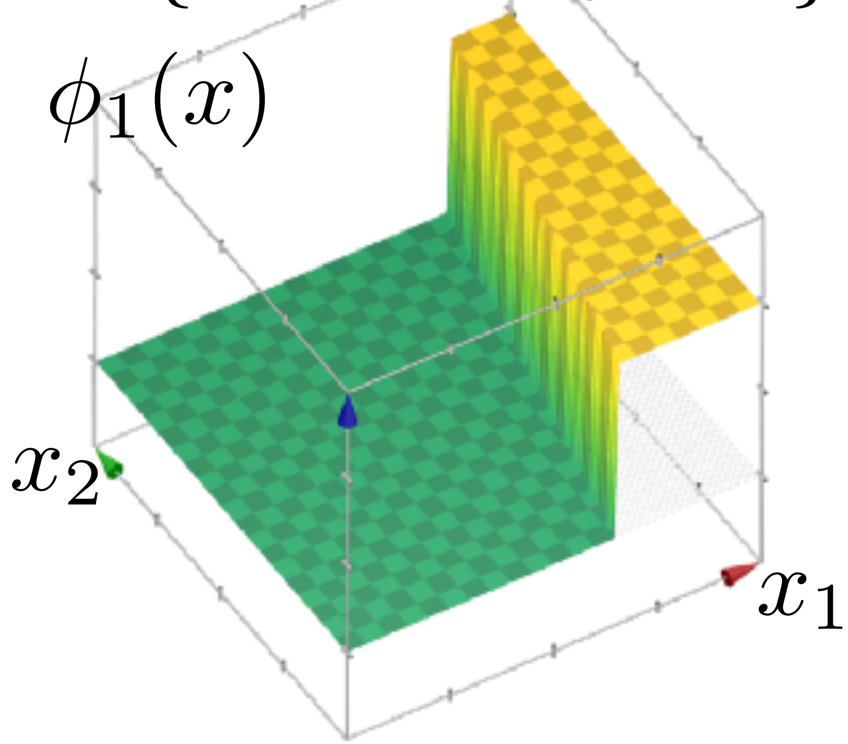


Will I go for a run?

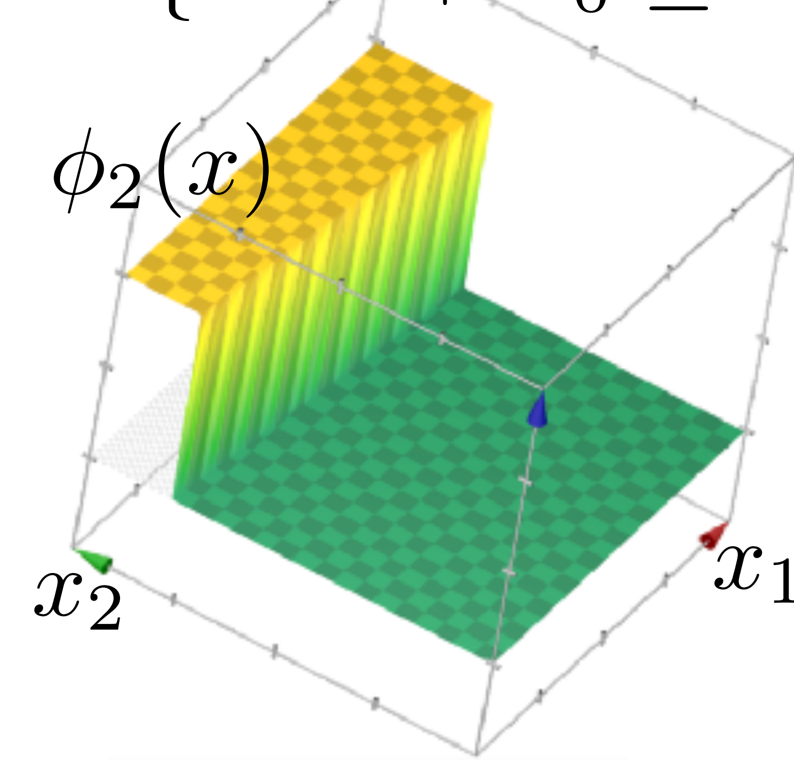


New features: step functions!

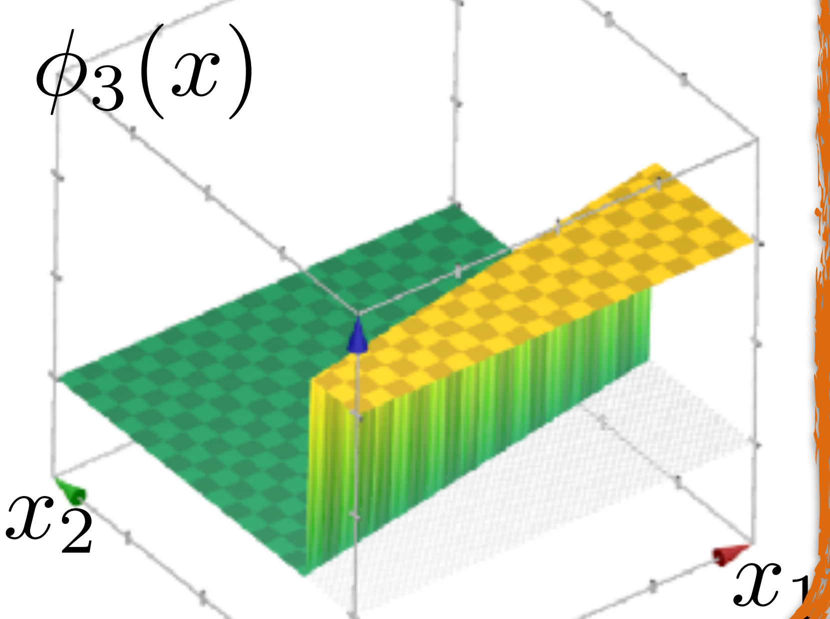
$$\phi_1(x) = \mathbf{1}\{w^\top x + w_0 \geq 0\}$$



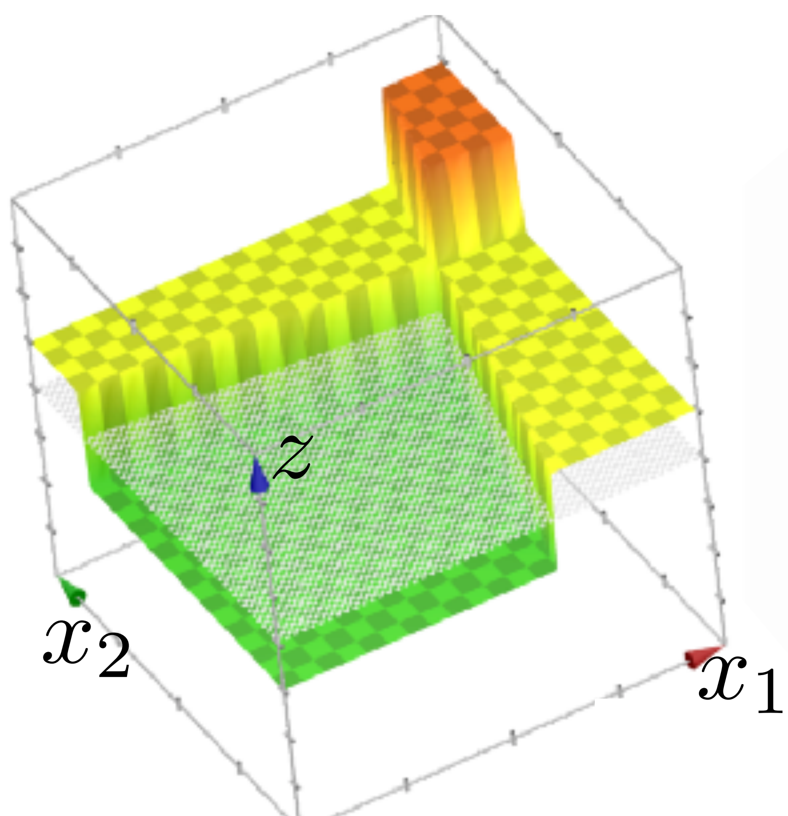
$$\phi_2(x) = \mathbf{1}\{\tilde{w}^\top x + \tilde{w}_0 \geq 0\}$$



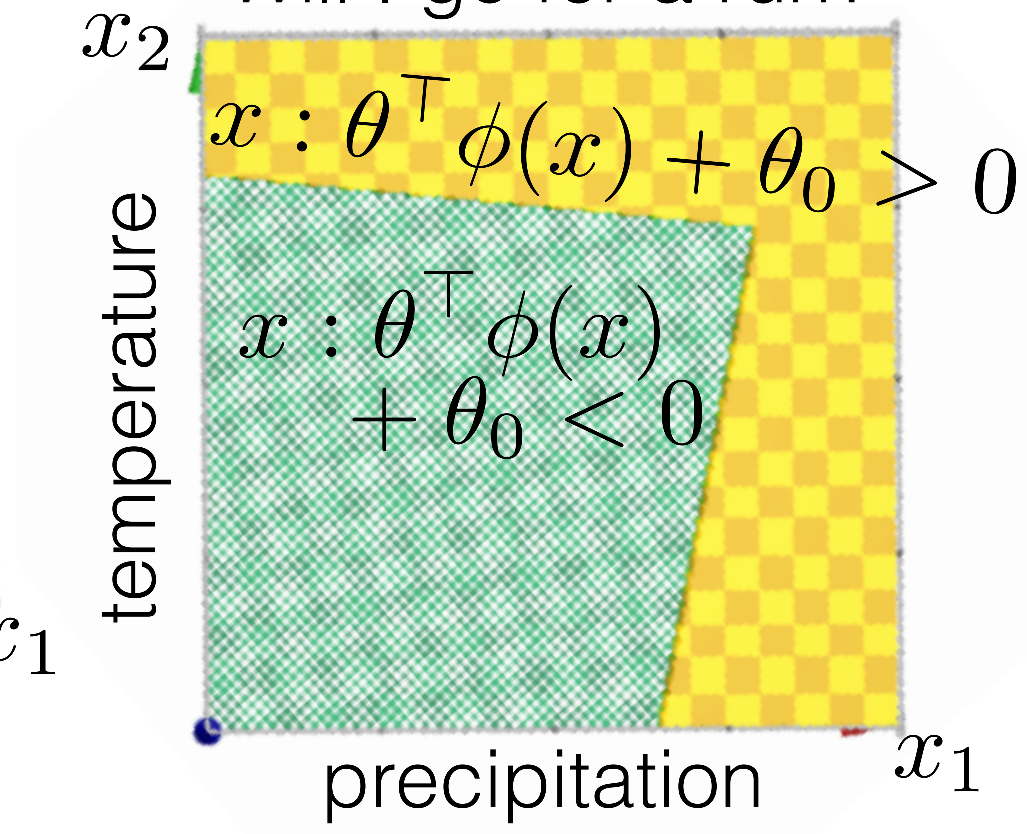
$$\phi_3(x) = \mathbf{1}\{\tilde{\tilde{w}}^\top x + \tilde{\tilde{w}}_0 \geq 0\}$$



$$z = \theta^\top \phi(x) + \theta_0$$
$$= \theta_1\phi_1(x) + \theta_2\phi_2(x) + \theta_0$$
$$= 1 \cdot \phi_1(x) + 1 \cdot \phi_2(x) + (-0.5)$$

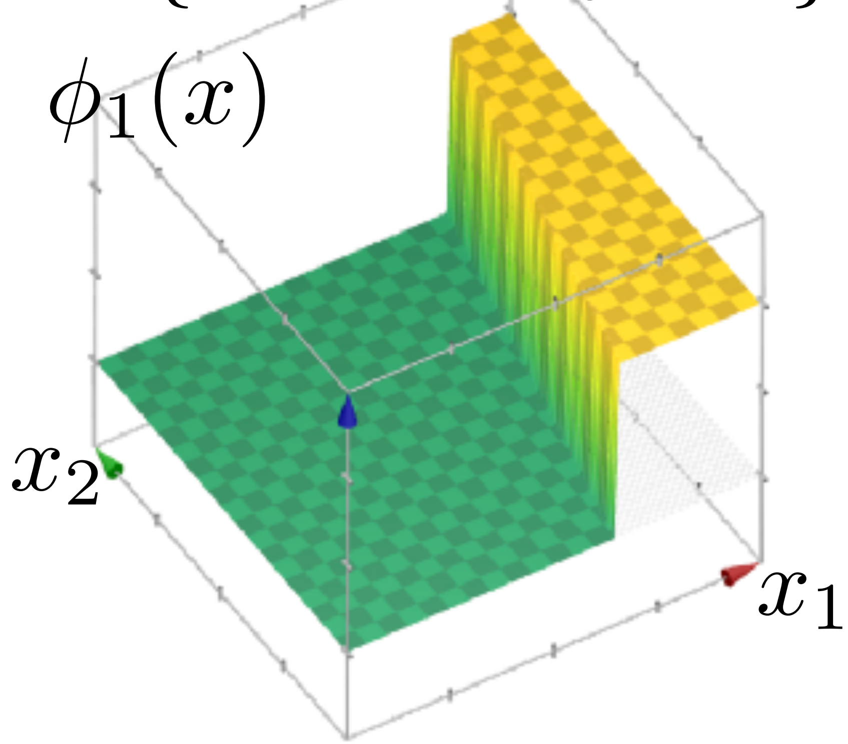


Will I go for a run?

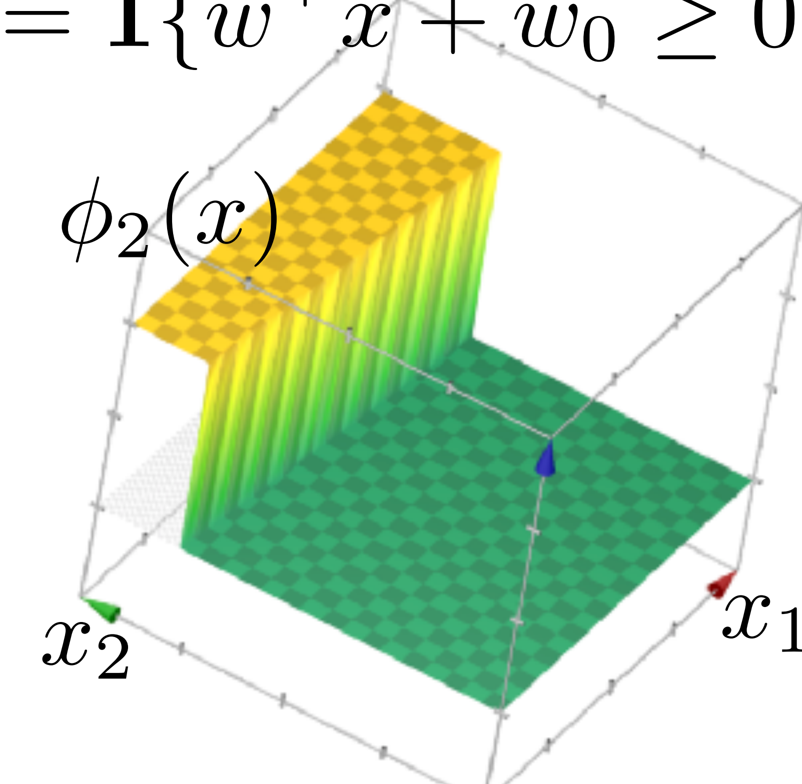


New features: step functions!

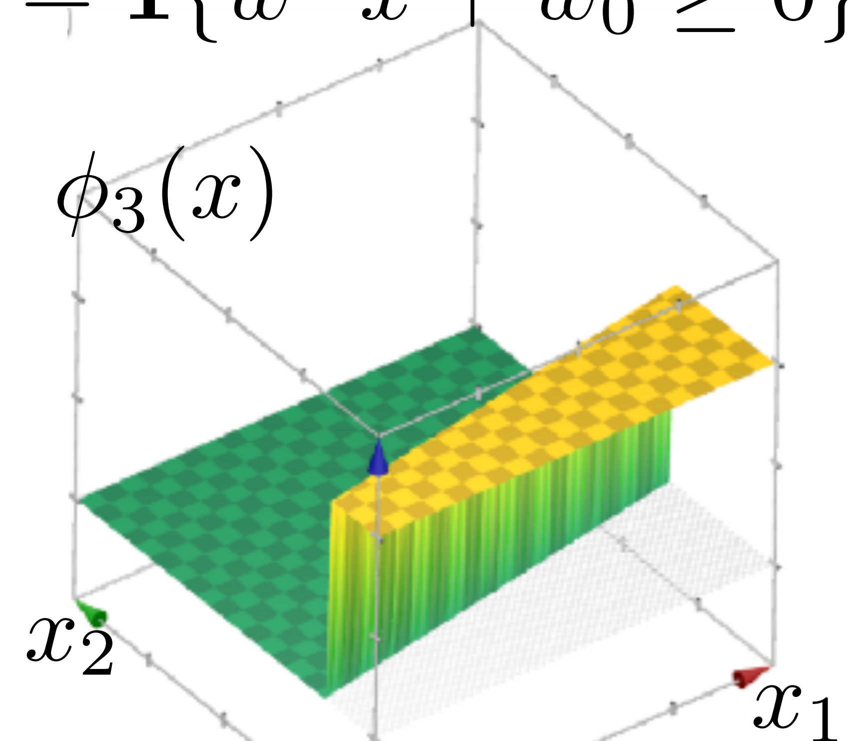
$\phi_1(x) = \mathbf{1}\{w^\top x + w_0 \geq 0\}$



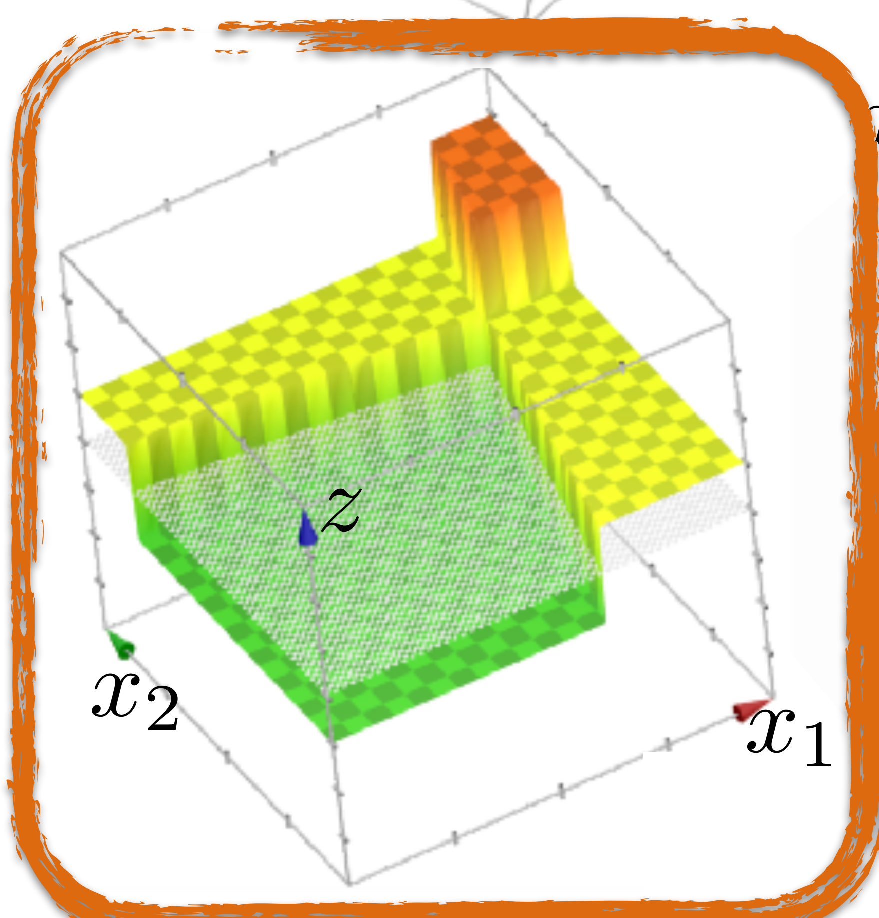
$\phi_2(x) = \mathbf{1}\{\tilde{w}^\top x + \tilde{w}_0 \geq 0\}$

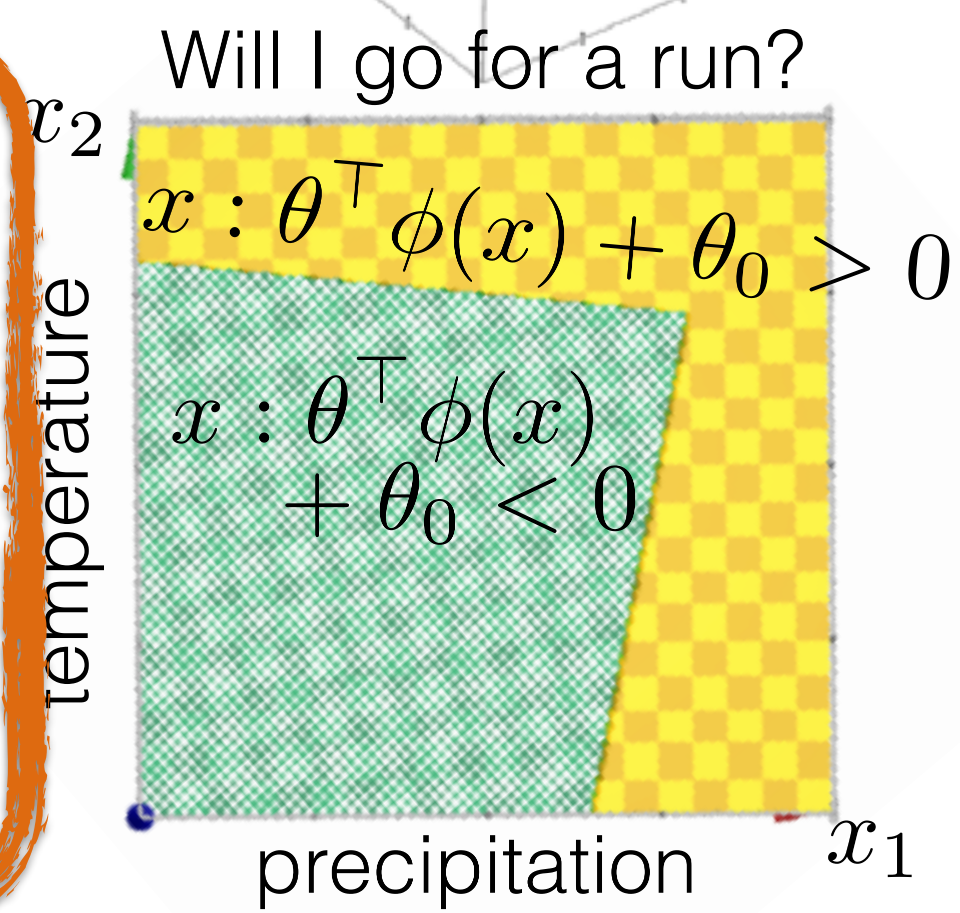


$\phi_3(x) = \mathbf{1}\{\tilde{\tilde{w}}^\top x + \tilde{\tilde{w}}_0 \geq 0\}$



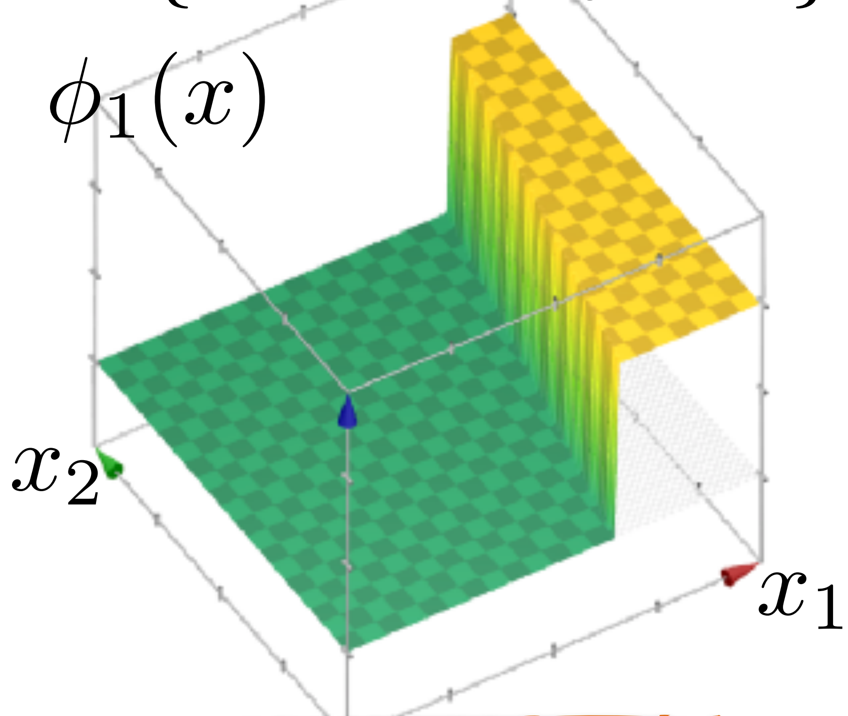
$$
\begin{aligned}
z &= \theta^\top \phi(x) + \theta_0 \\
&= \theta_1 \phi_1(x) + \theta_2 \phi_2(x) \\
&\quad + \theta_0 \\
&= 1 \cdot \phi_1(x) + 1 \cdot \phi_2(x) \\
&\quad + (-0.5)
\end{aligned}
$$



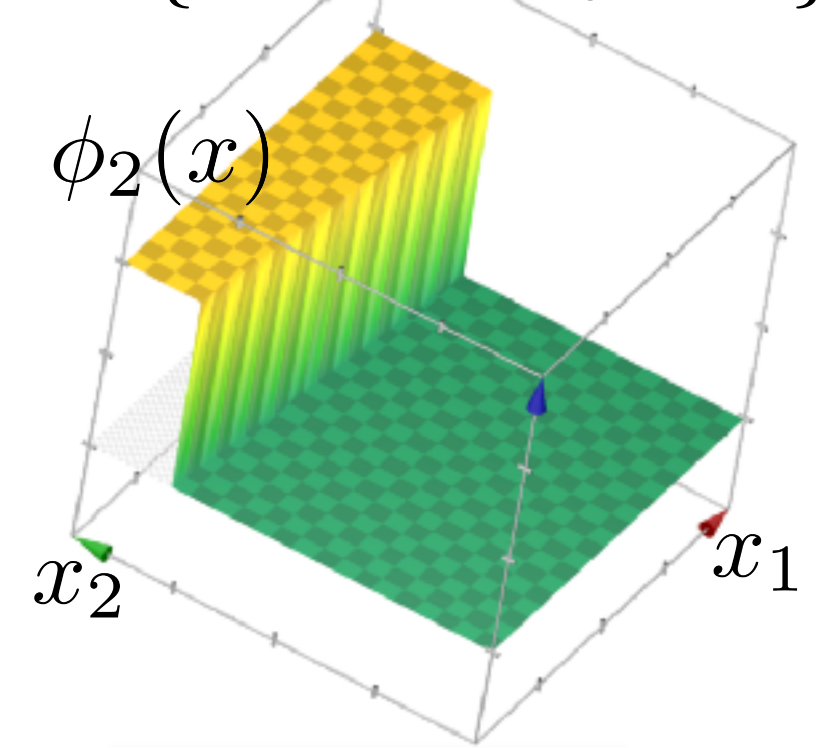


New features: step functions!

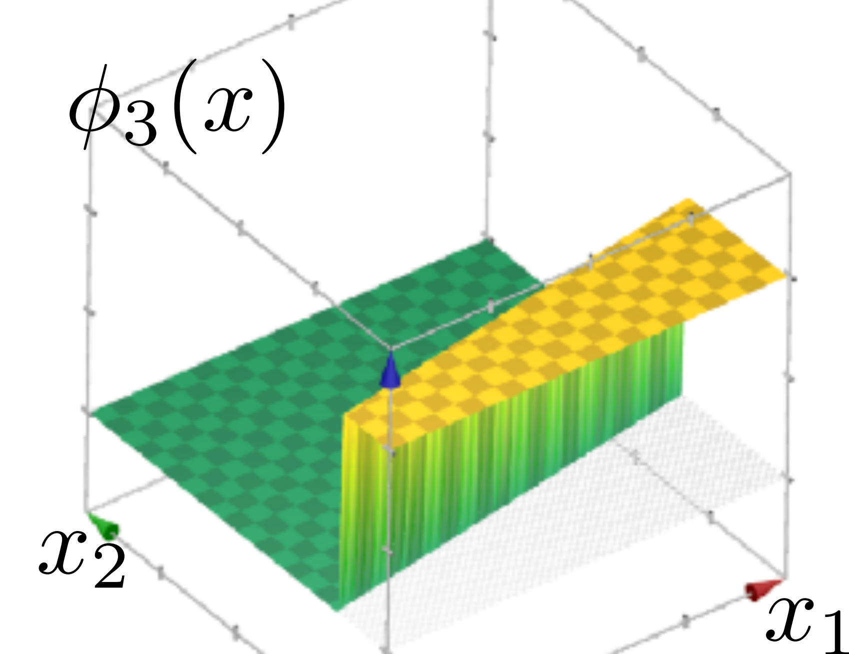
$\phi_1(x) = \mathbf{1}\{w^\top x + w_0 \geq 0\}$



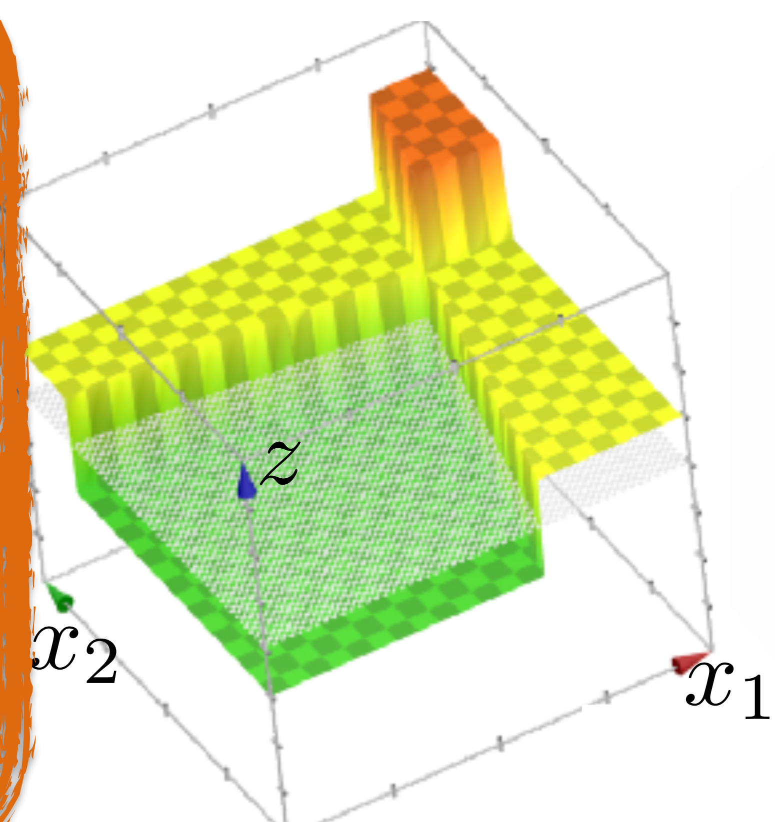
$\phi_2(x) = \mathbf{1}\{\tilde{w}^\top x + \tilde{w}_0 \geq 0\}$



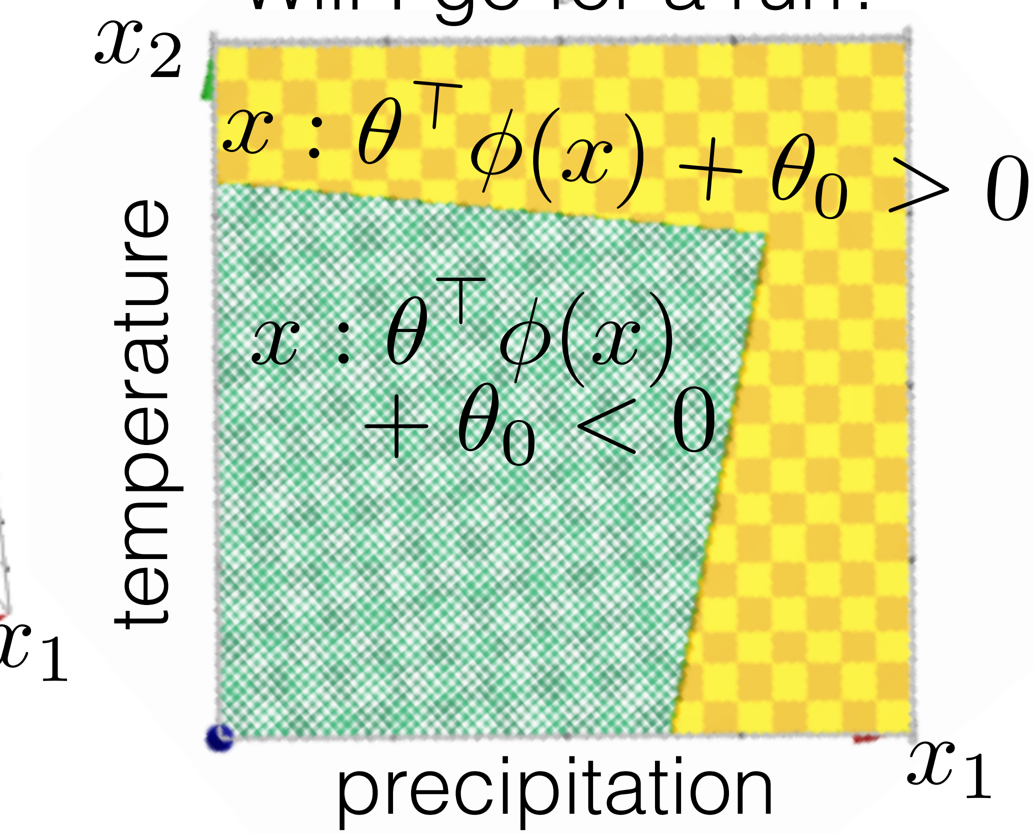
$\phi_3(x) = \mathbf{1}\{\tilde{\tilde{w}}^\top x + \tilde{\tilde{w}}_0 \geq 0\}$



$$\boxed{\begin{aligned} z &= \theta^\top \phi(x) + \theta_0 \\ &= \theta_1\phi_1(x) + \theta_2\phi_2(x) \\ &\quad + \theta_0 \\ &= 1 \cdot \phi_1(x) + 1 \cdot \phi_2(x) \\ &\quad + (-0.5) \end{aligned}}$$

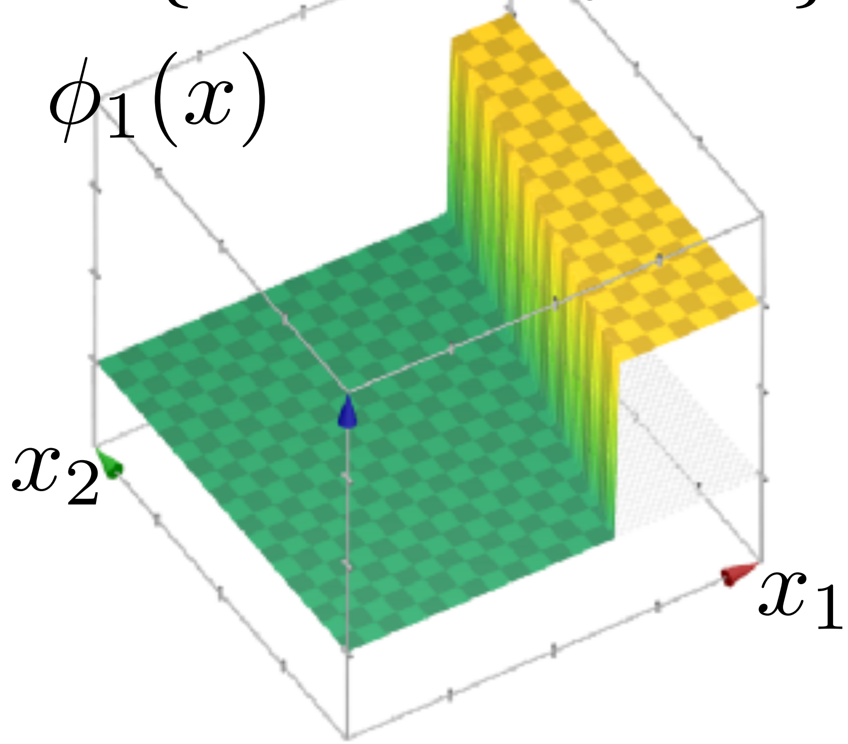


Will I go for a run?

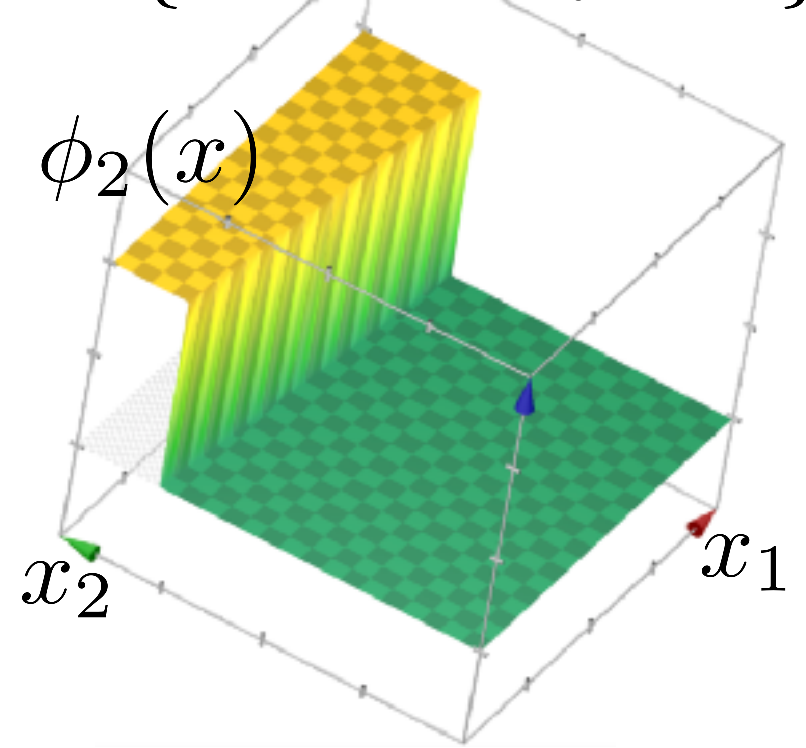


New features: step functions!

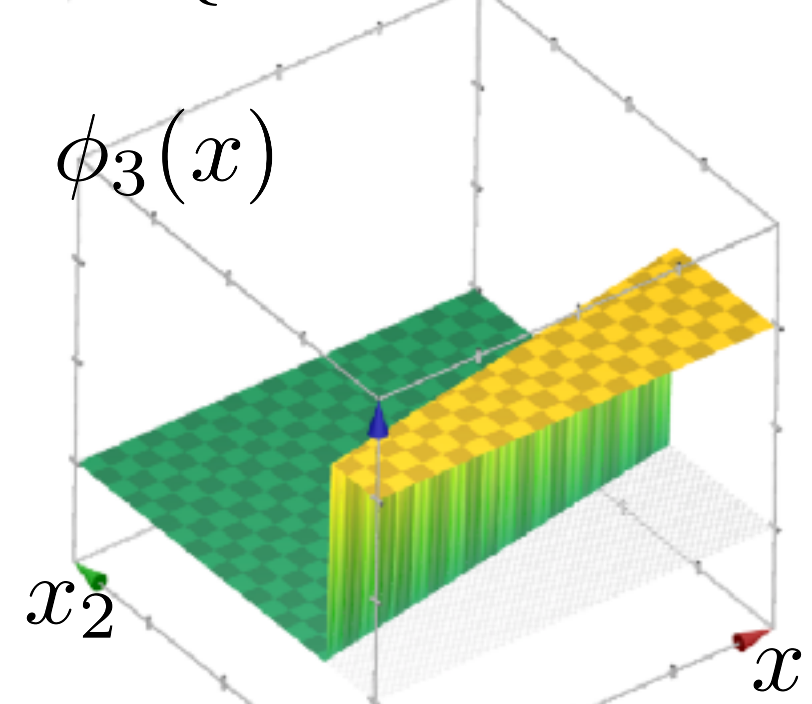
$\phi_1(x) = \mathbf{1}\{w^\top x + w_0 \geq 0\}$



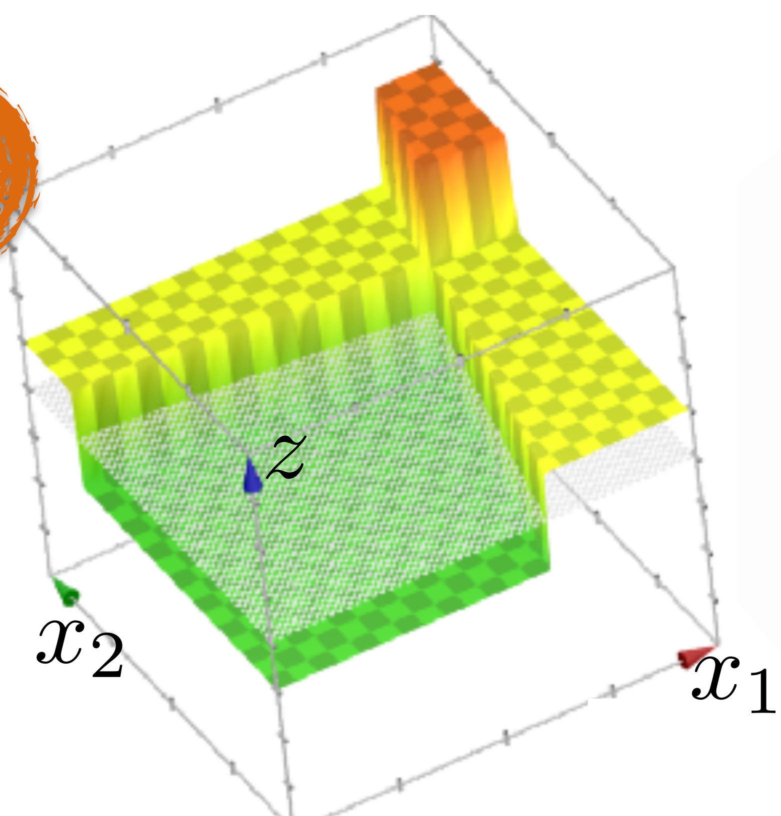
$\phi_2(x) = \mathbf{1}\{\tilde{w}^\top x + \tilde{w}_0 \geq 0\}$



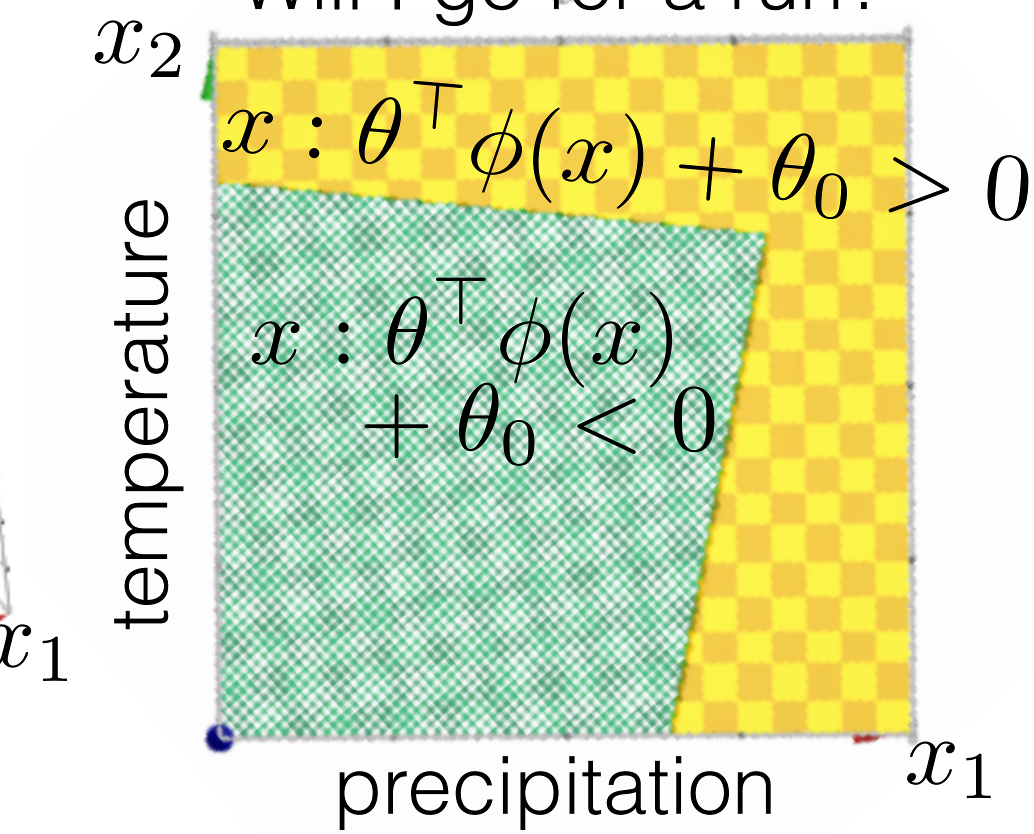
$\phi_3(x) = \mathbf{1}\{\tilde{\tilde{w}}^\top x + \tilde{\tilde{w}}_0 \geq 0\}$



$$z = \theta^\top \phi(x) + \theta_0$$
$$= \theta_1 \phi_1(x) + \theta_2 \phi_2(x) + \theta_0$$
$$= 1 \cdot \phi_1(x) + 1 \cdot \phi_2(x) + (-0.5)$$

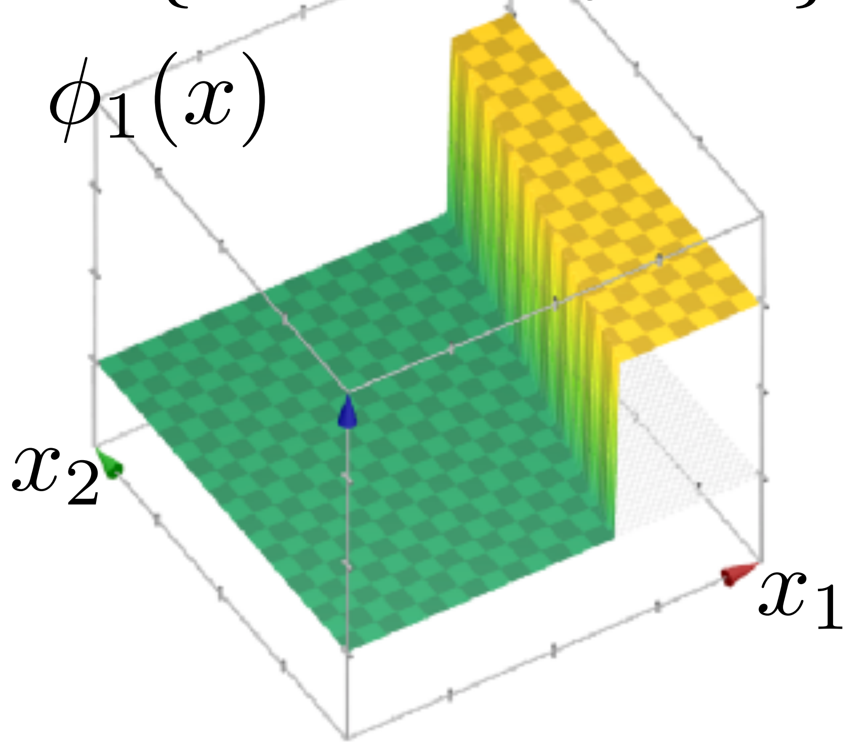


Will I go for a run?

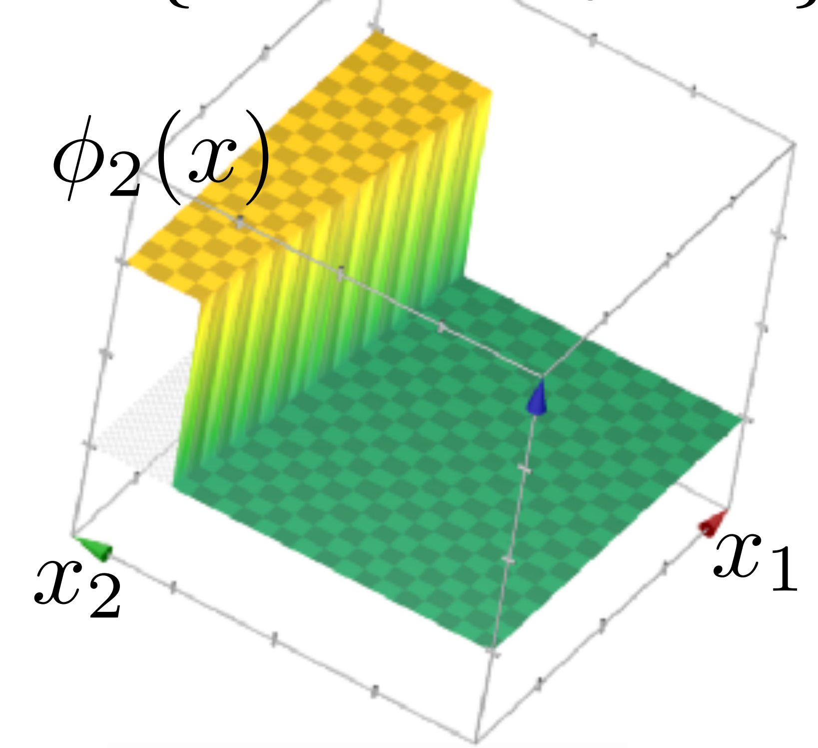


New features: step functions!

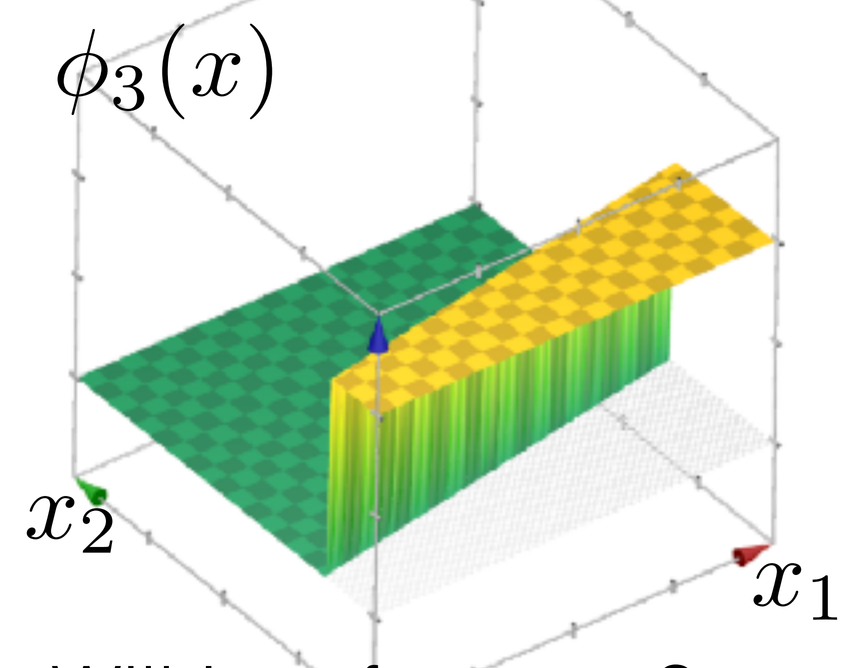
$\phi_1(x) = \mathbf{1}\{w^\top x + w_0 \geq 0\}$



$\phi_2(x) = \mathbf{1}\{\tilde{w}^\top x + \tilde{w}_0 \geq 0\}$



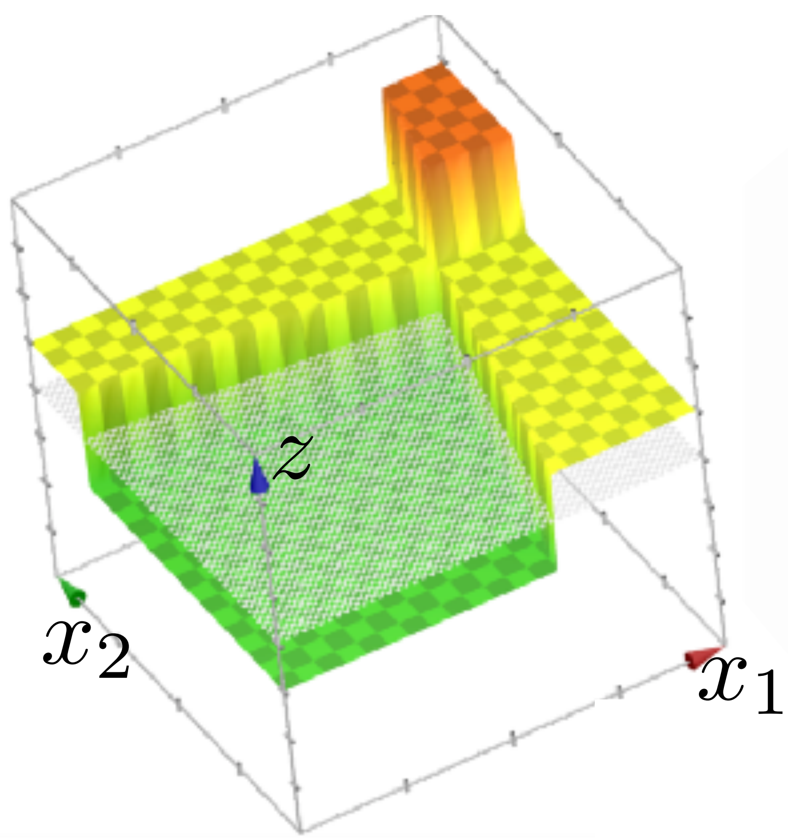
$\phi_3(x) = \mathbf{1}\{\tilde{\tilde{w}}^\top x + \tilde{\tilde{w}}_0 \geq 0\}$



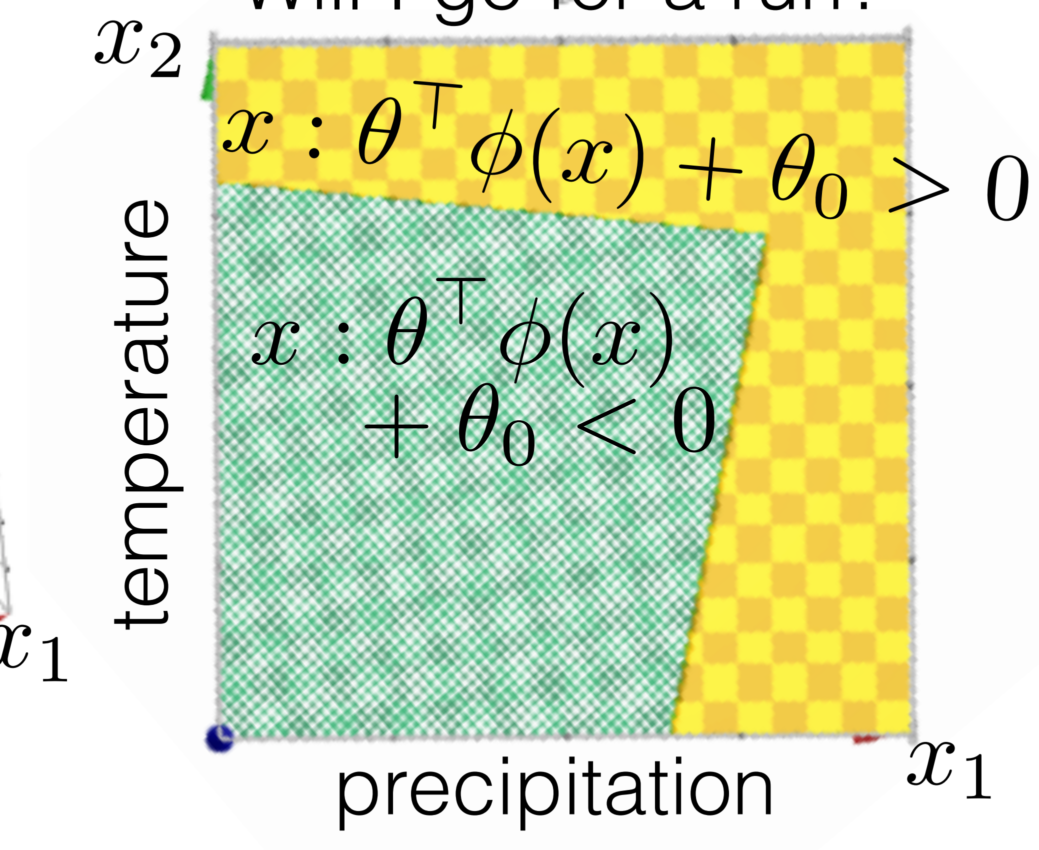
$$z = \theta^\top \phi(x) + \theta_0$$

$$= \theta_1 \phi_1(x) + \theta_2 \phi_2(x) + \theta_0$$

$$= 1 \cdot \phi_1(x) + 1 \cdot \phi_2(x) + (-0.5)$$

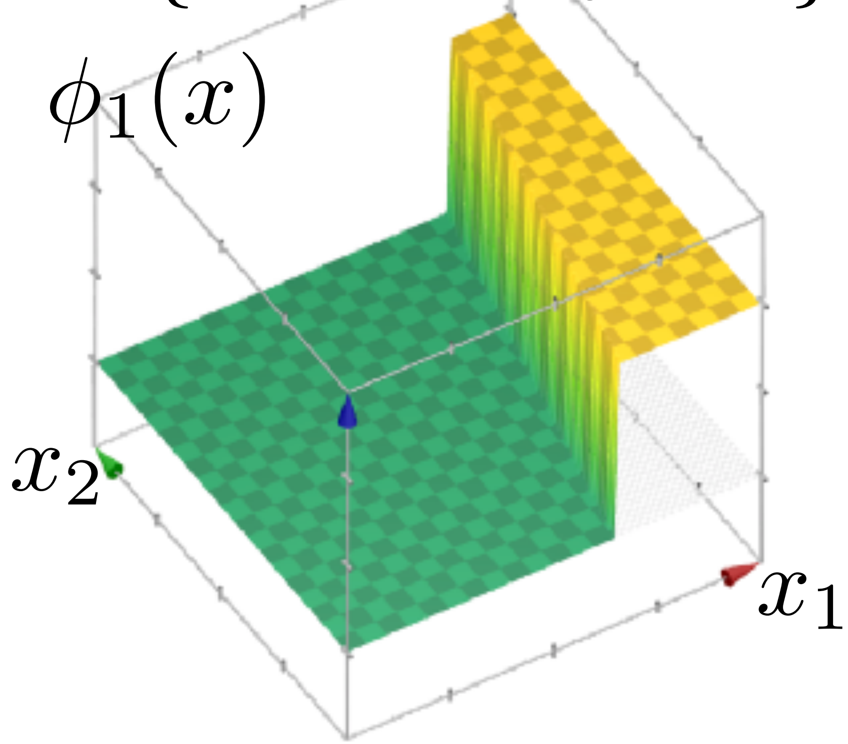


Will I go for a run?

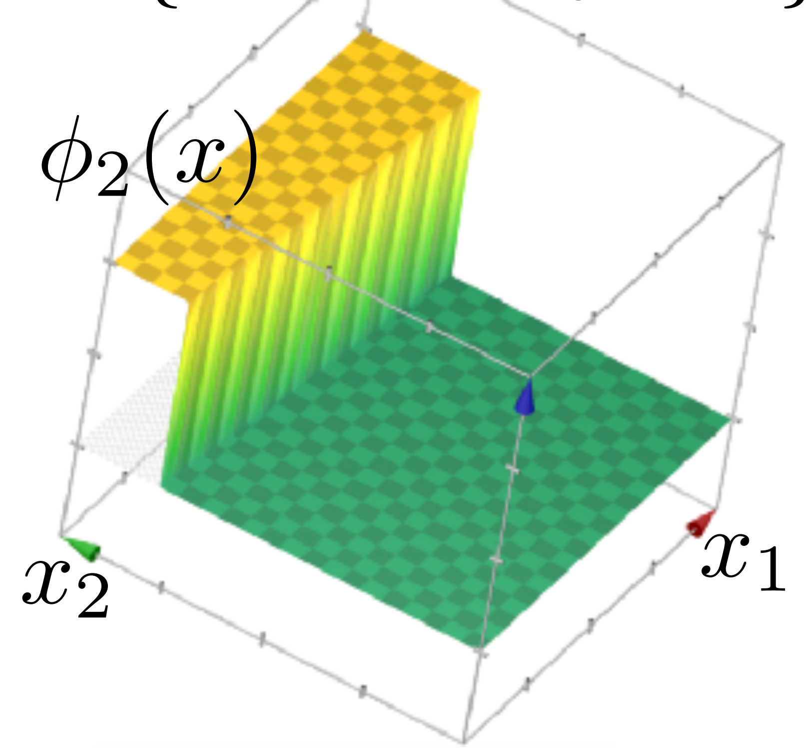


New features: step functions!

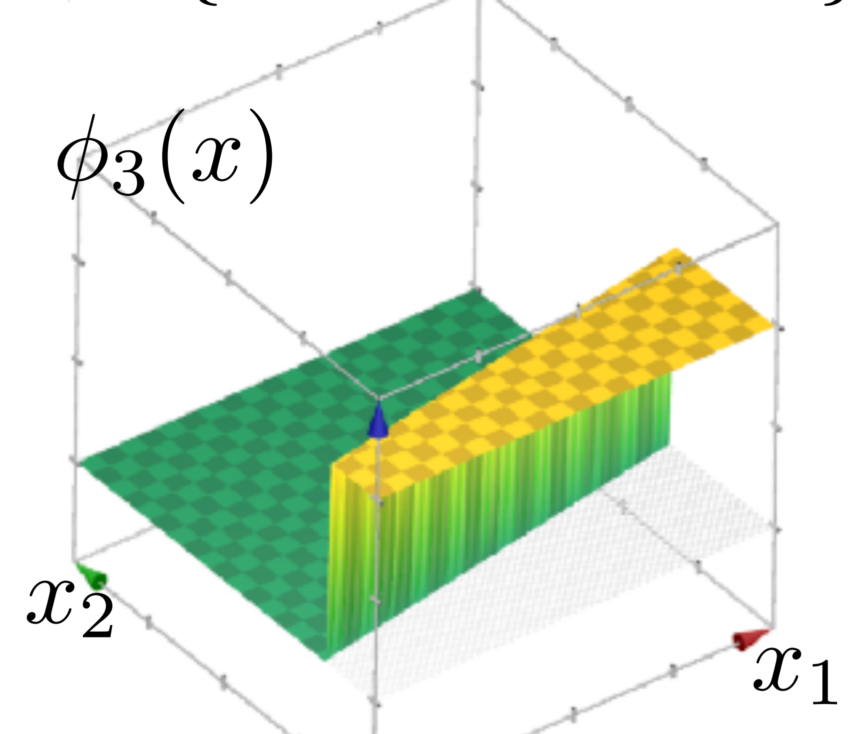
$\phi_1(x) = \mathbf{1}\{w^\top x + w_0 \geq 0\}$



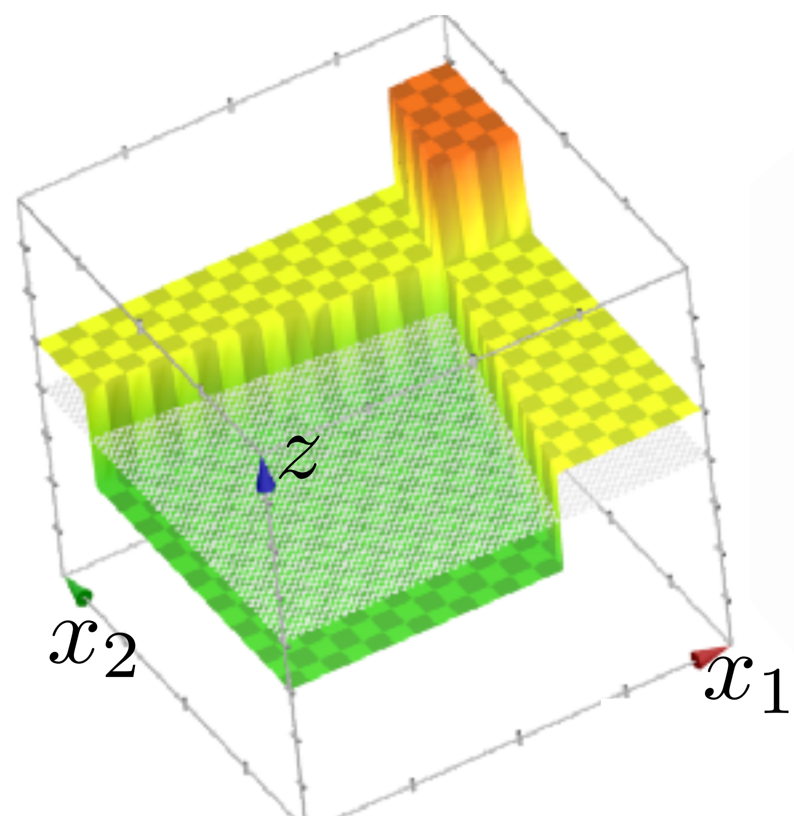
$\phi_2(x) = \mathbf{1}\{\tilde{w}^\top x + \tilde{w}_0 \geq 0\}$



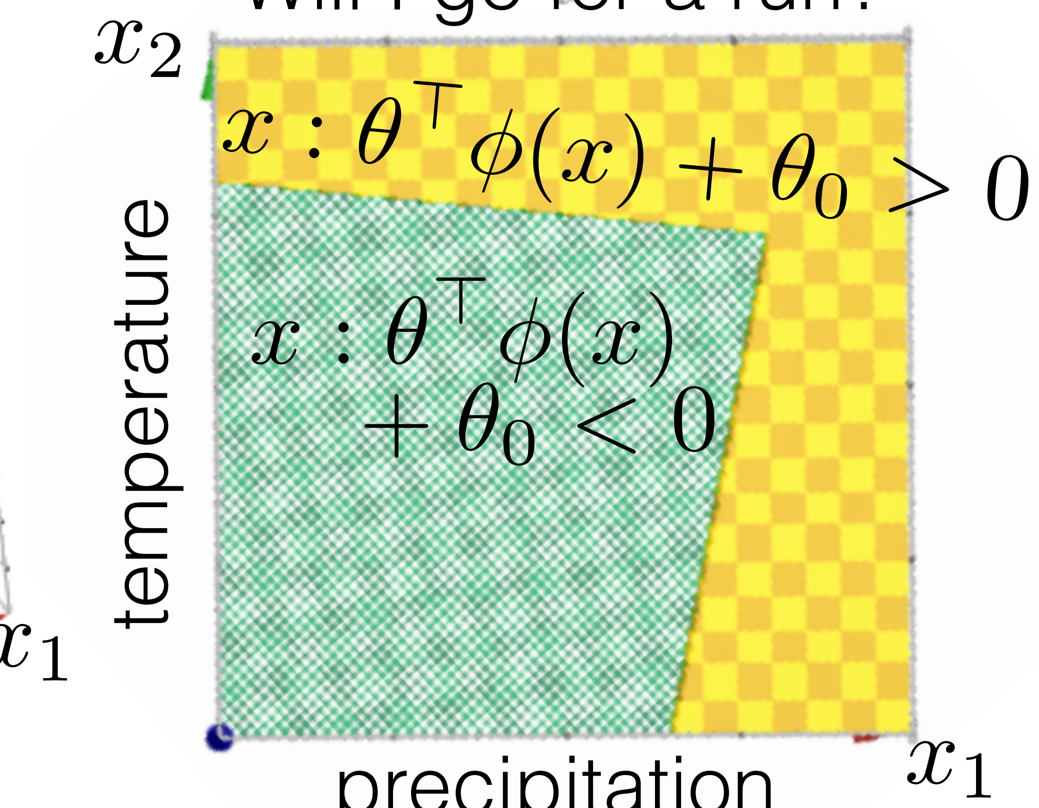
$\phi_3(x) = \mathbf{1}\{\tilde{\tilde{w}}^\top x + \tilde{\tilde{w}}_0 \geq 0\}$



$$
\begin{aligned}
z &= \theta^\top \phi(x) + \theta_0 \\
&= \theta_1 \phi_1(x) + \theta_2 \phi_2(x) + \theta_3 \phi_3(x) + \theta_0 \\
&= 1 \cdot \phi_1(x) + 1 \cdot \phi_2(x) + (-0.5)
\end{aligned}
$$

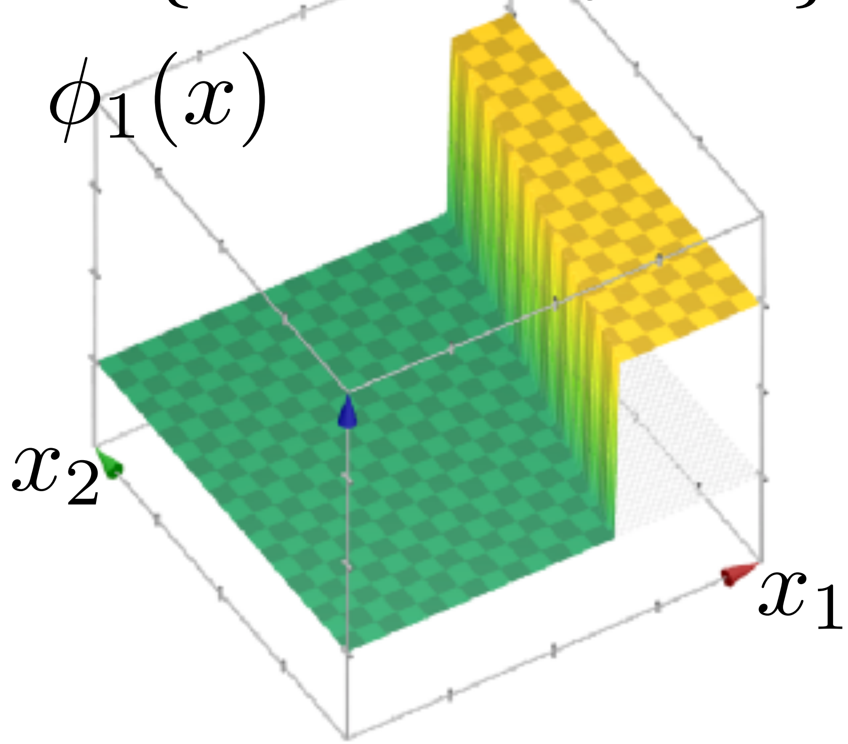


Will I go for a run?

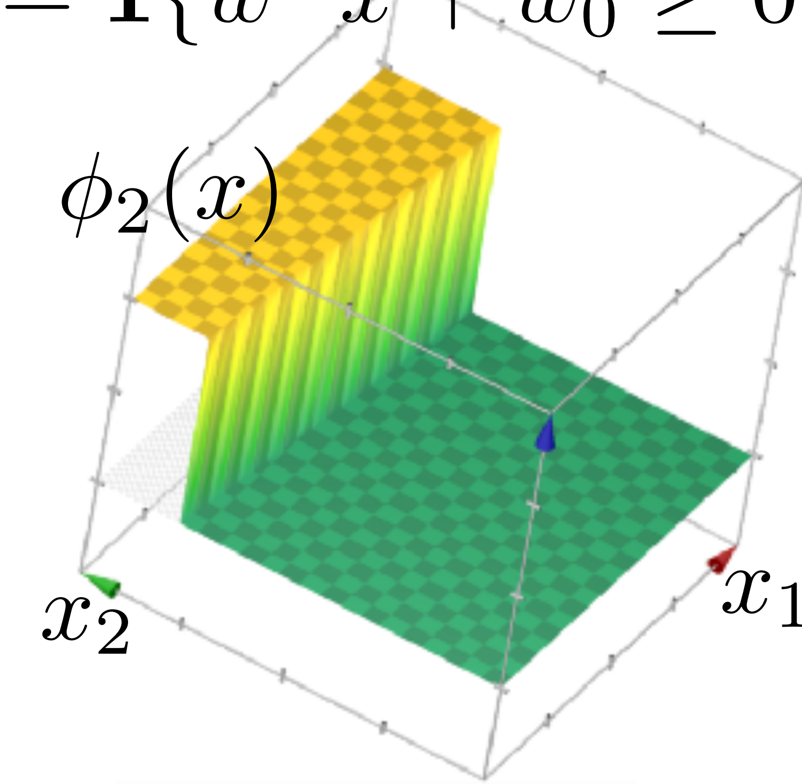


New features: step functions!

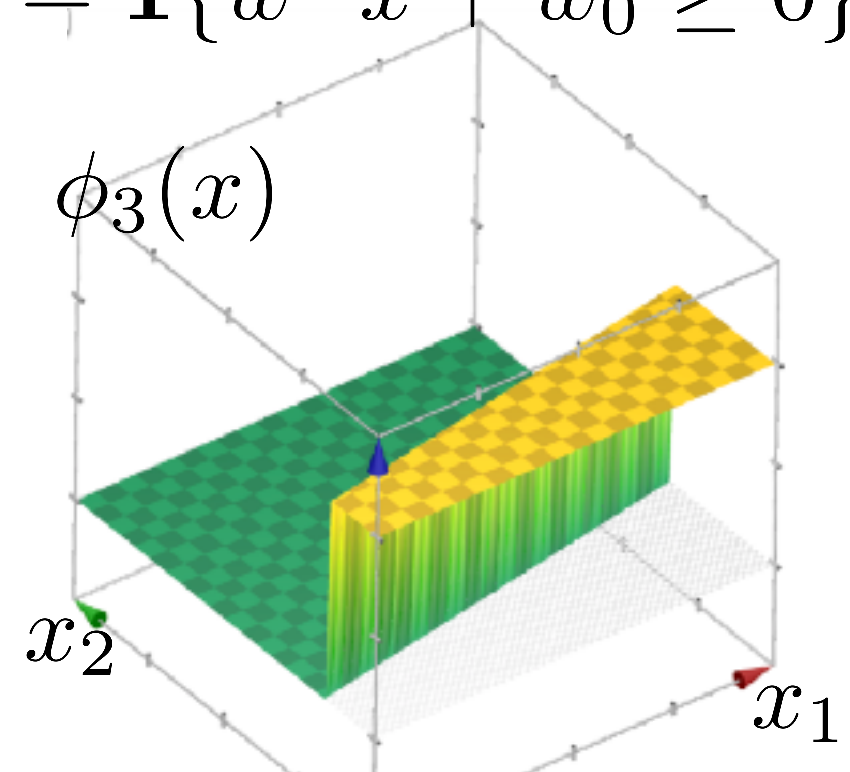
$\phi_1(x) = \mathbf{1}\{w^\top x + w_0 \geq 0\}$



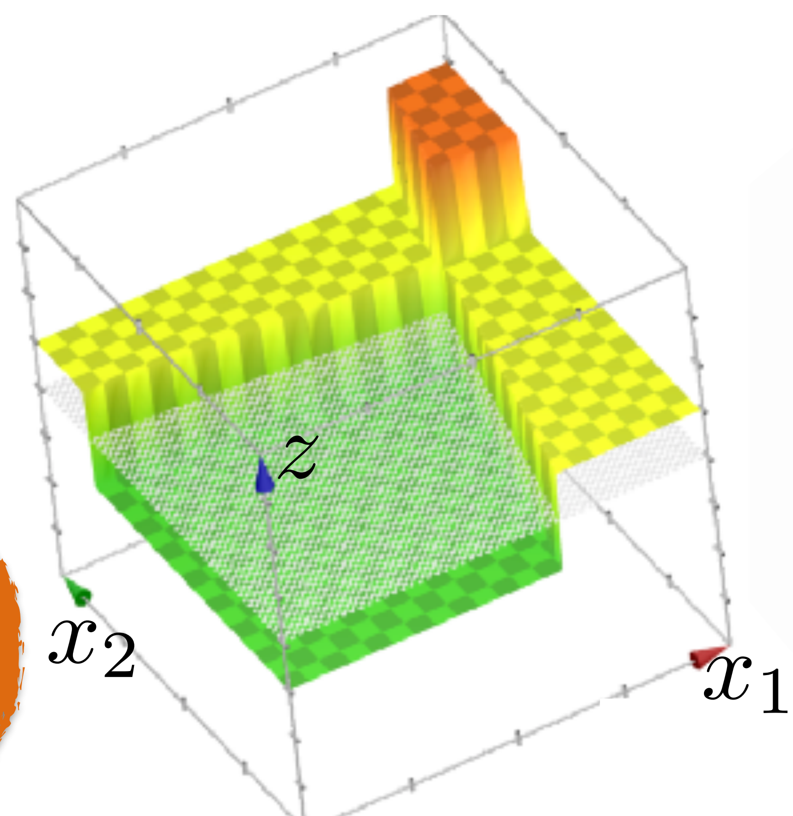
$\phi_2(x) = \mathbf{1}\{\tilde{w}^\top x + \tilde{w}_0 \geq 0\}$



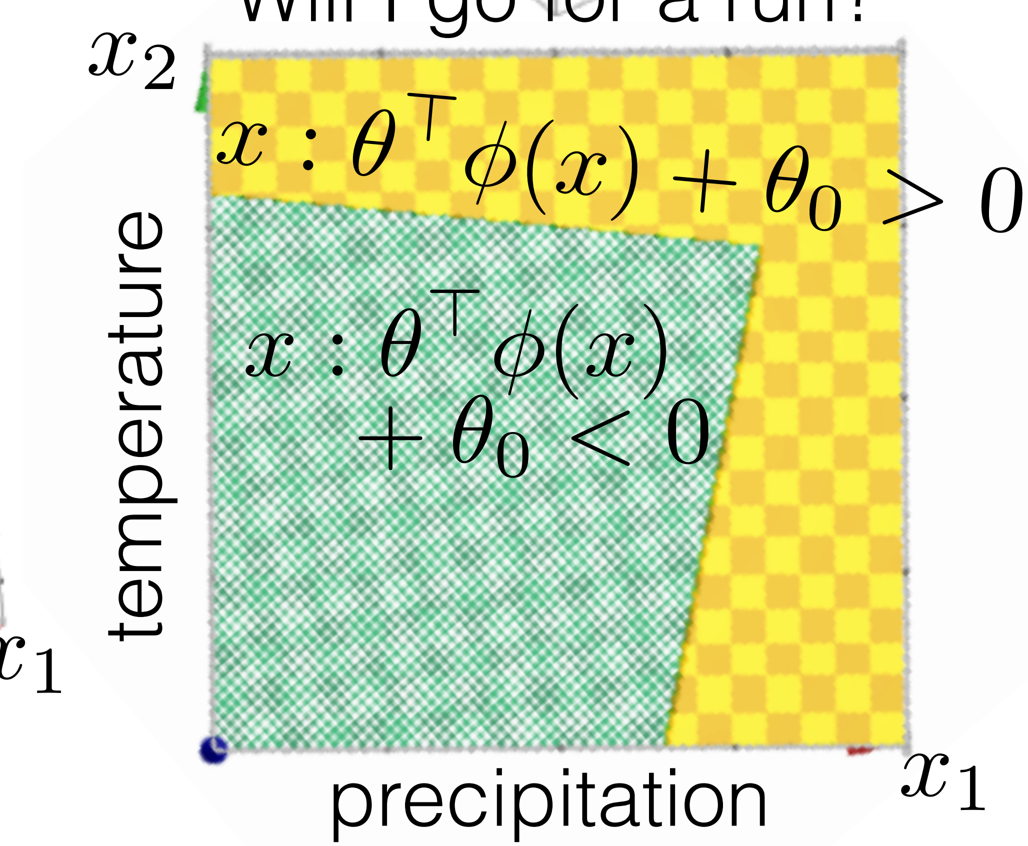
$\phi_3(x) = \mathbf{1}\{\tilde{\tilde{w}}^\top x + \tilde{\tilde{w}}_0 \geq 0\}$



$$
\begin{aligned}
z &= \theta^\top \phi(x) + \theta_0 \\
&= \theta_1 \phi_1(x) + \theta_2 \phi_2(x) \\
&\quad + \theta_3 \phi_3(x) + \theta_0 \\
&= 1 \cdot \phi_1(x) + 1 \cdot \phi_2(x) \\
&\quad + (-0.5)
\end{aligned}
$$

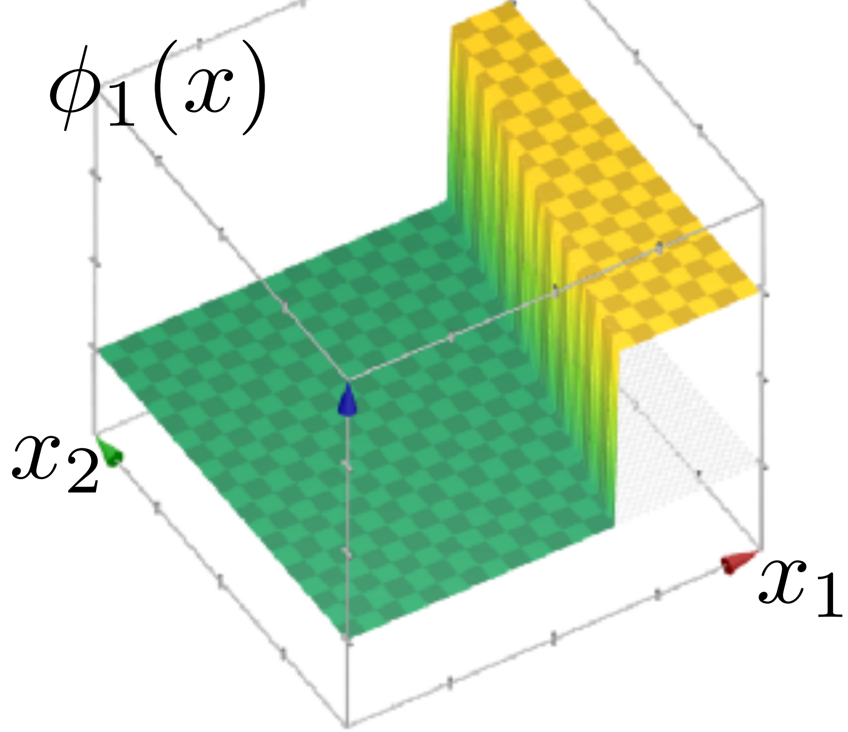


Will I go for a run?

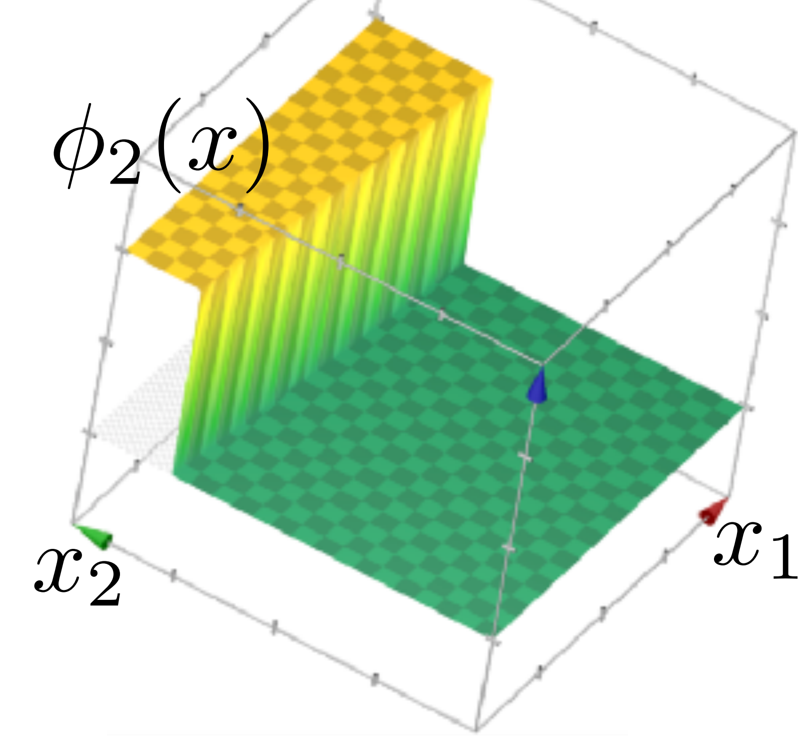


New features: step functions!

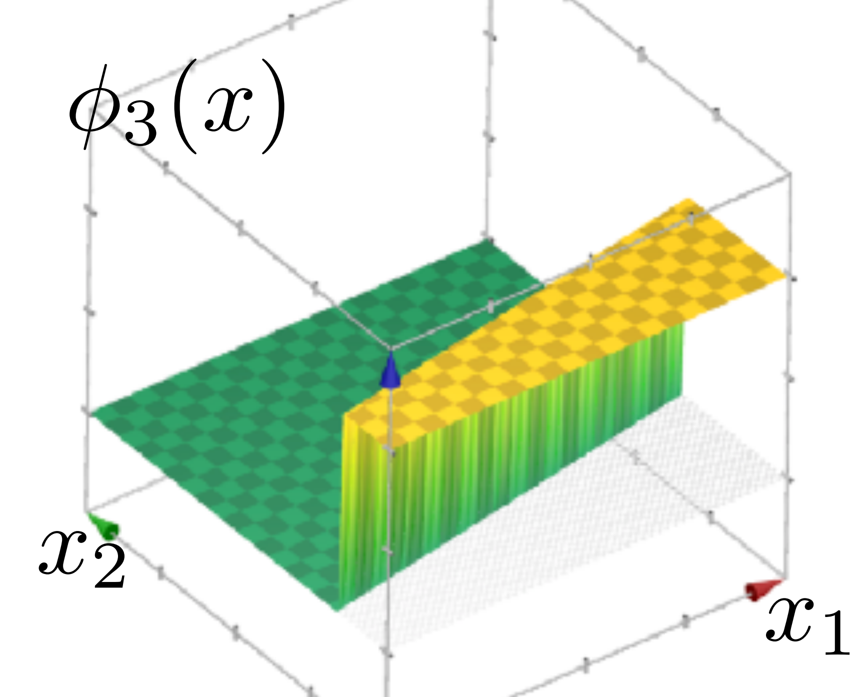
$\phi_1(x) = \mathbf{1}\{w^\top x + w_0 \geq 0\}$



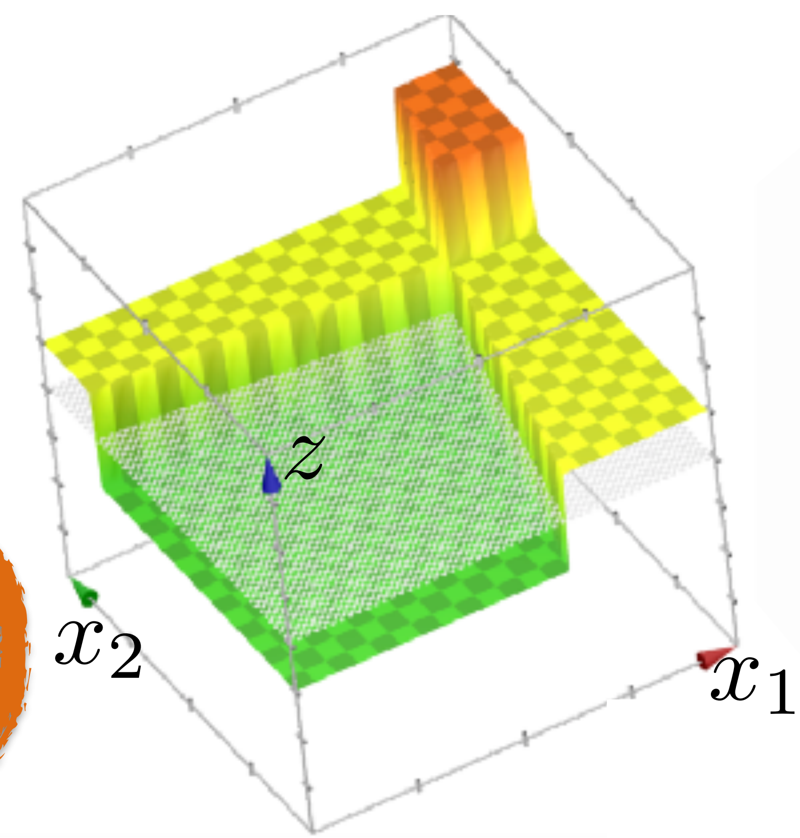
$\phi_2(x) = \mathbf{1}\{\tilde{w}^\top x + \tilde{w}_0 \geq 0\}$



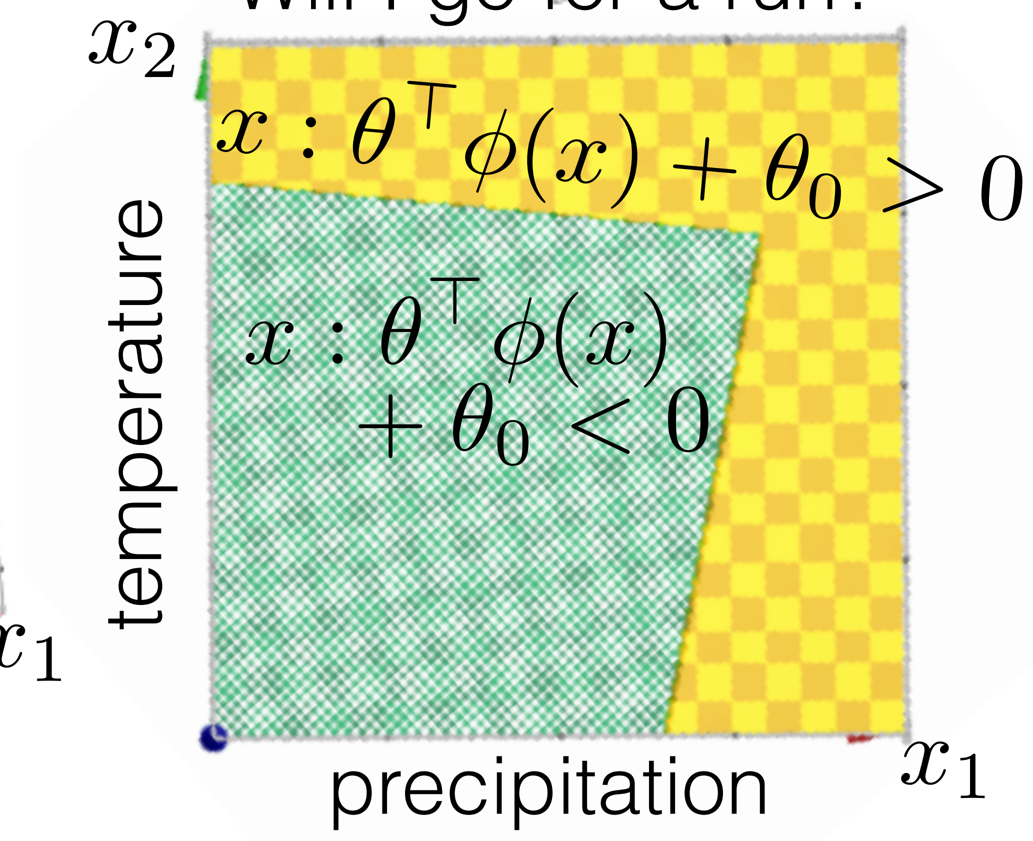
$\phi_3(x) = \mathbf{1}\{\tilde{\tilde{w}}^\top x + \tilde{\tilde{w}}_0 \geq 0\}$



$$z = \theta^\top \phi(x) + \theta_0$$
$$= \theta_1 \phi_1(x) + \theta_2 \phi_2(x) + \theta_3 \phi_3(x) + \theta_0$$
$$= 1 \cdot \phi_1(x) + 1 \cdot \phi_2(x) + 1 \cdot \phi_3(x) + (-0.5)$$

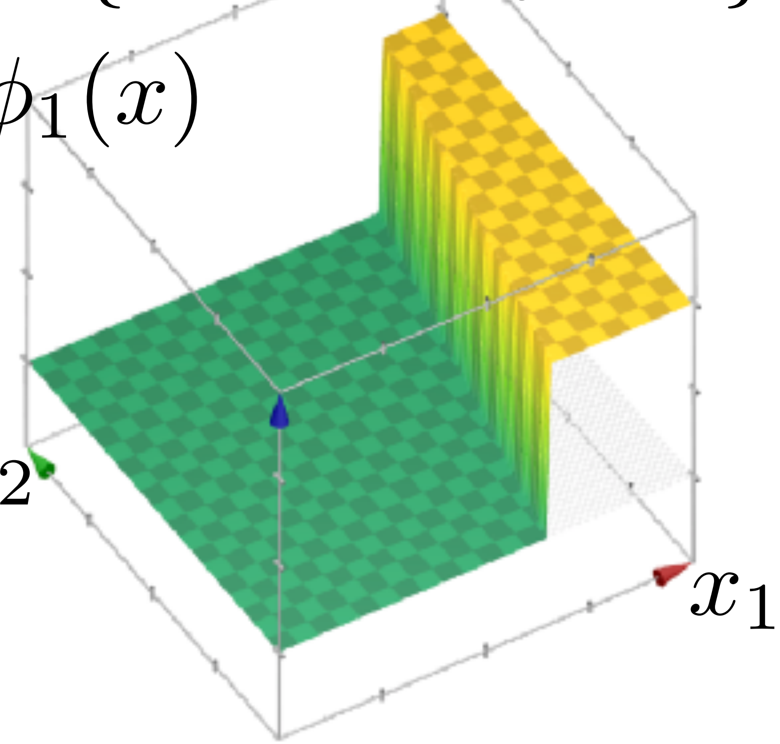


Will I go for a run?

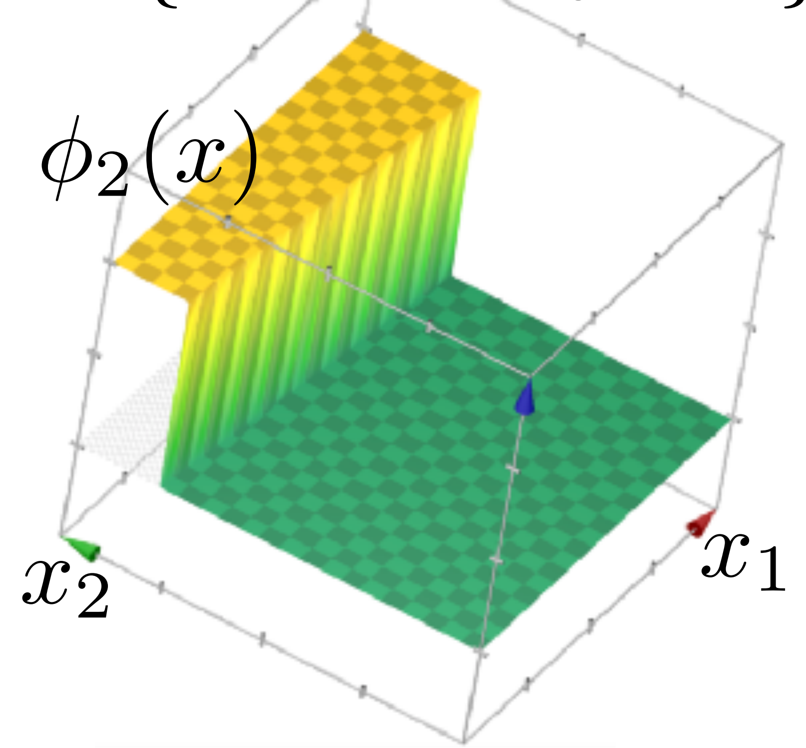


New features: step functions!

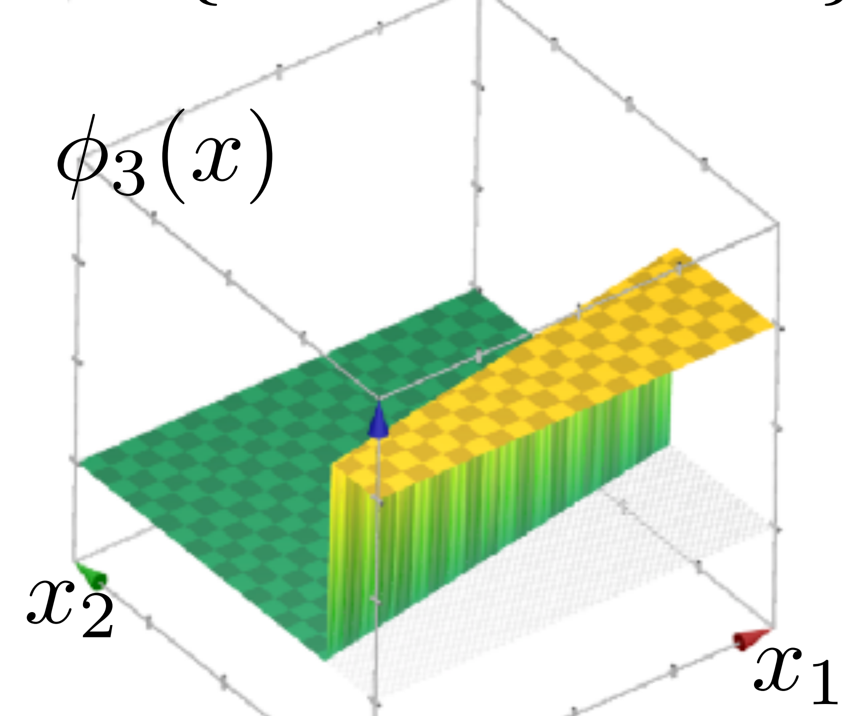
$\phi_1(x) = \mathbf{1}\{w^\top x + w_0 \geq 0\}$



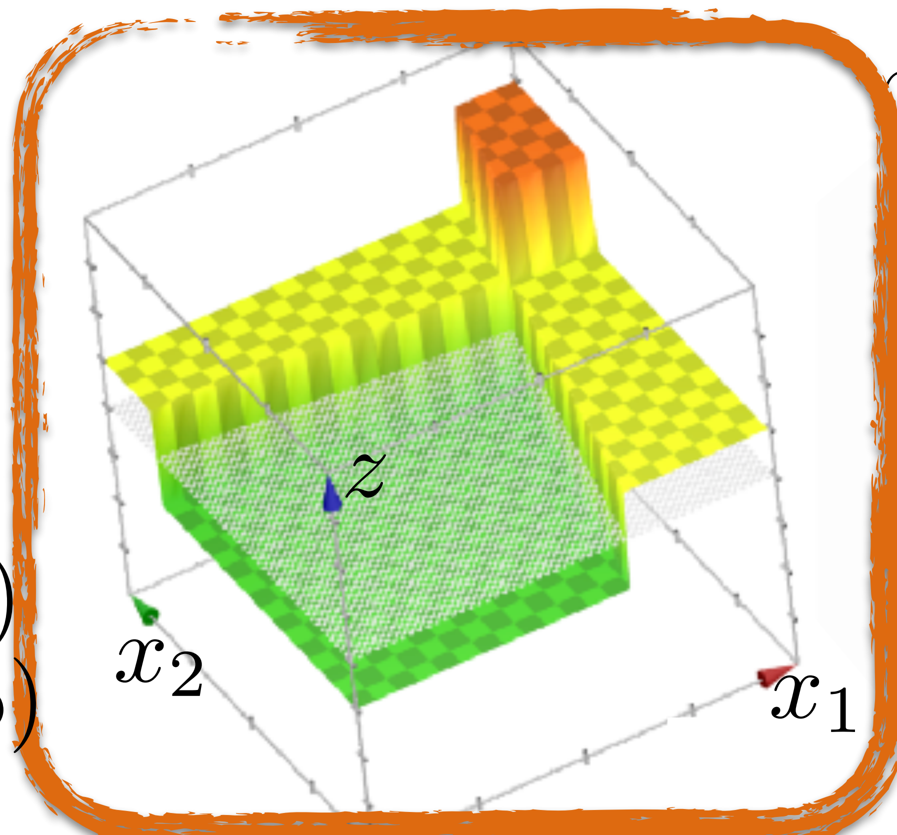
$\phi_2(x) = \mathbf{1}\{\tilde{w}^\top x + \tilde{w}_0 \geq 0\}$



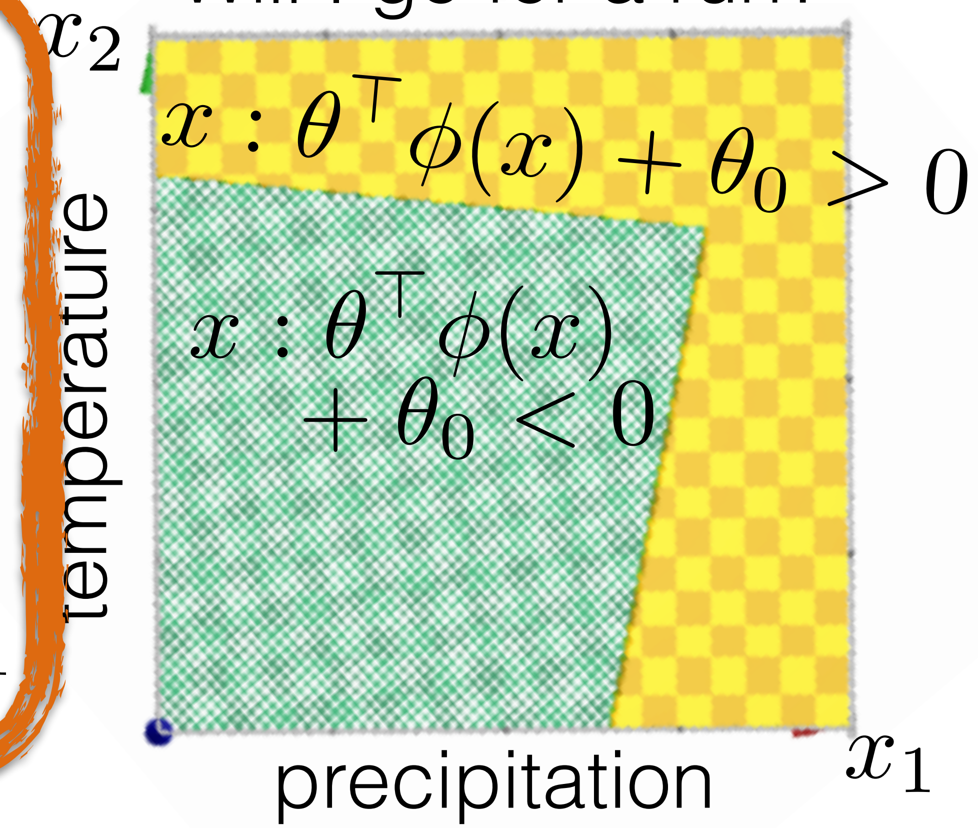
$\phi_3(x) = \mathbf{1}\{\tilde{\tilde{w}}^\top x + \tilde{\tilde{w}}_0 \geq 0\}$



$$
\begin{aligned}
z &= \theta^\top \phi(x) + \theta_0 \\
&= \theta_1 \phi_1(x) + \theta_2 \phi_2(x) \\
&\quad + \theta_3 \phi_3(x) + \theta_0 \\
&= 1 \cdot \phi_1(x) + 1 \cdot \phi_2(x) \\
&\quad + 1 \cdot \phi_3(x) + (-0.5)
\end{aligned}
$$

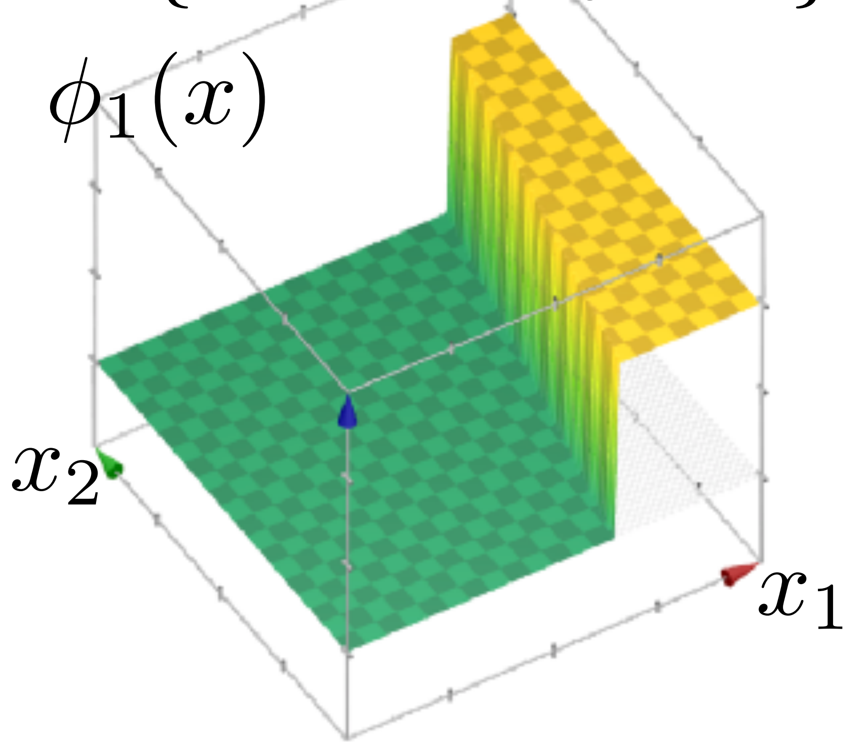


Will I go for a run?

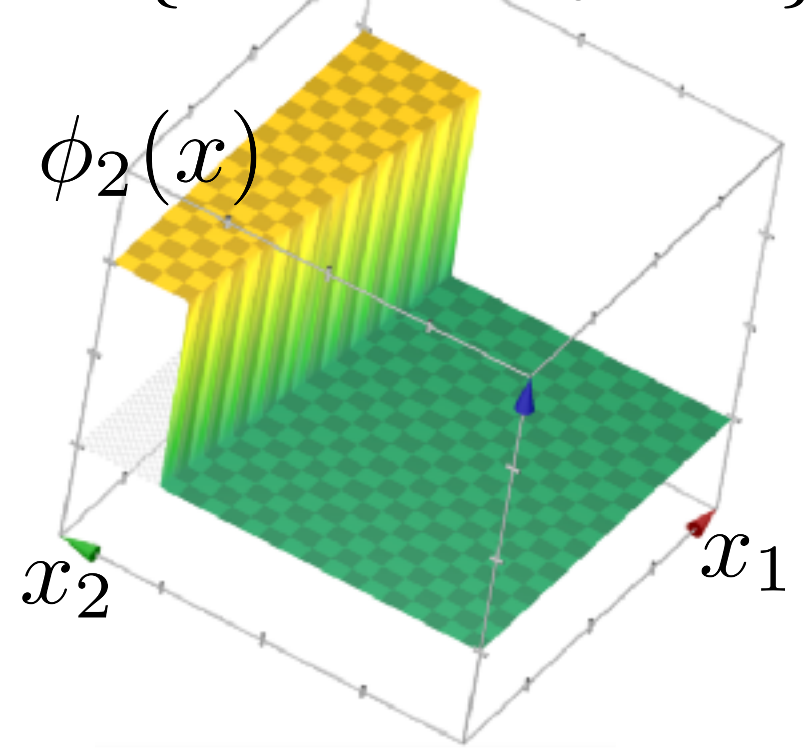


New features: step functions!

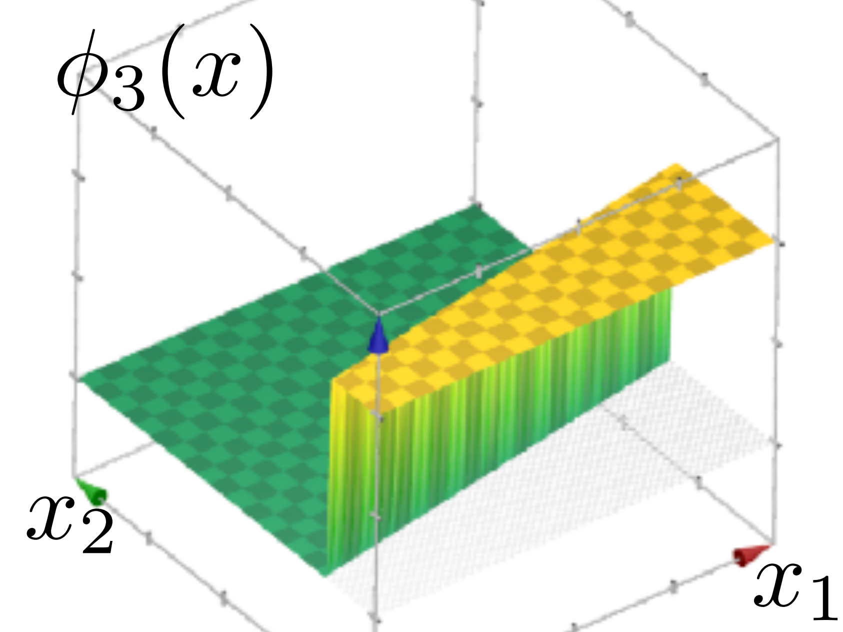
$\phi_1(x) = \mathbf{1}\{w^\top x + w_0 \geq 0\}$



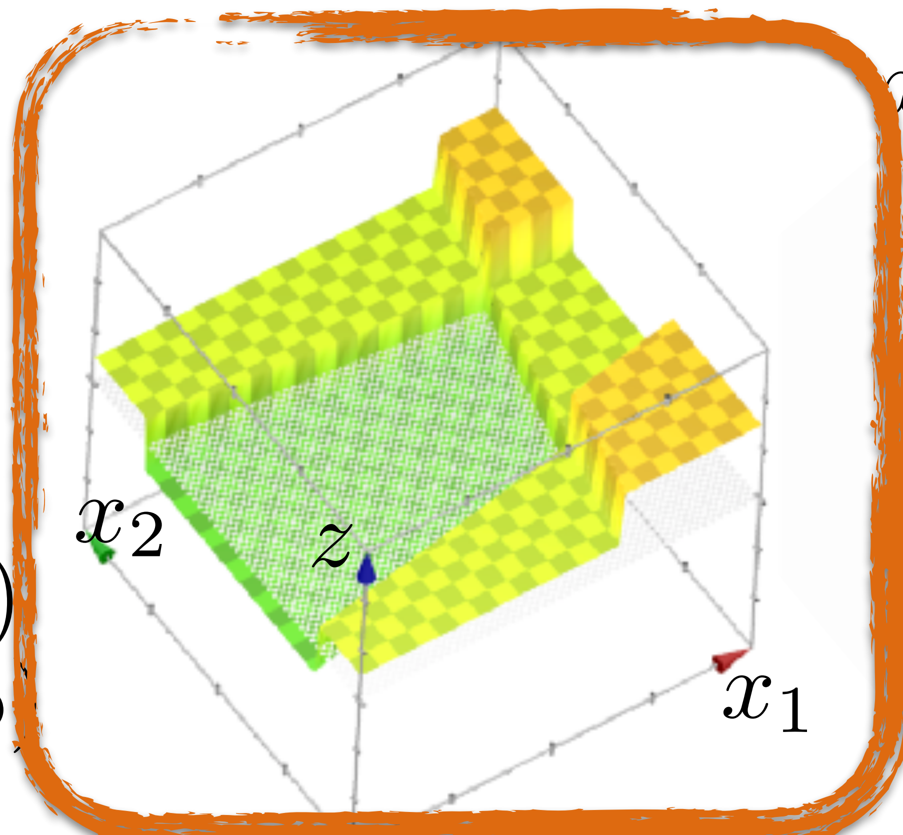
$\phi_2(x) = \mathbf{1}\{\tilde{w}^\top x + \tilde{w}_0 \geq 0\}$



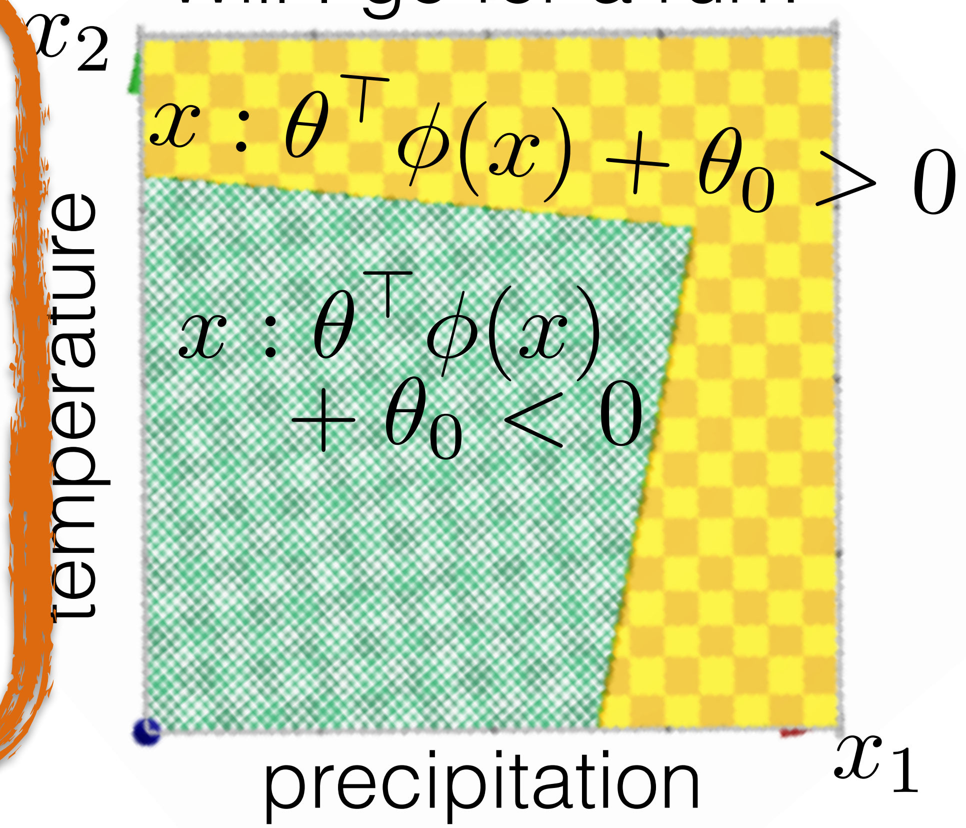
$\phi_3(x) = \mathbf{1}\{\tilde{\tilde{w}}^\top x + \tilde{\tilde{w}}_0 \geq 0\}$



$$z = \theta^\top \phi(x) + \theta_0$$
$$= \theta_1 \phi_1(x) + \theta_2 \phi_2(x) + \theta_3 \phi_3(x) + \theta_0$$
$$= 1 \cdot \phi_1(x) + 1 \cdot \phi_2(x) + 1 \cdot \phi_3(x) + (-0.5$$

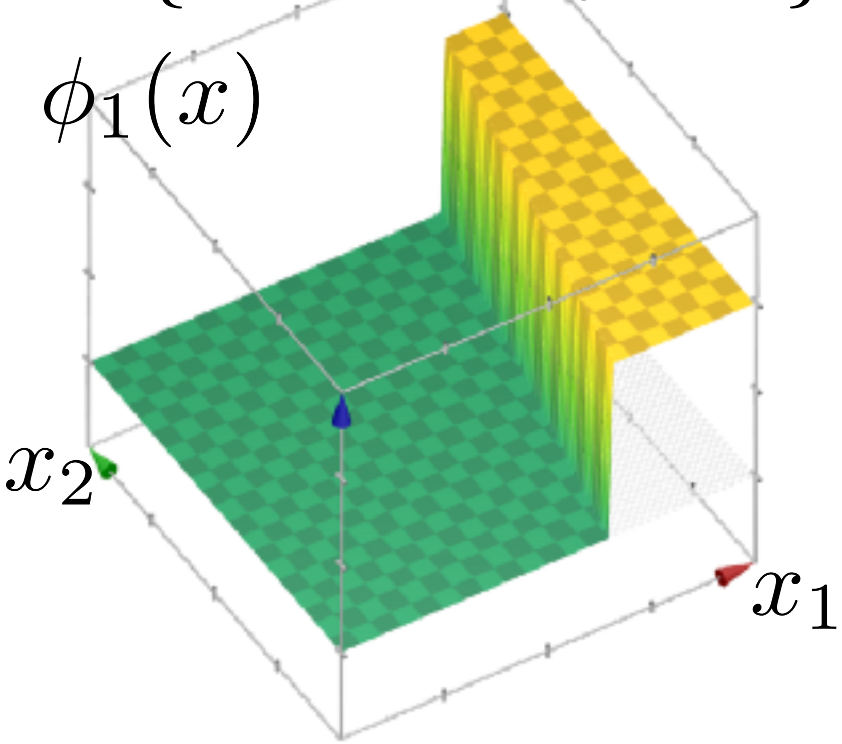


Will I go for a run?

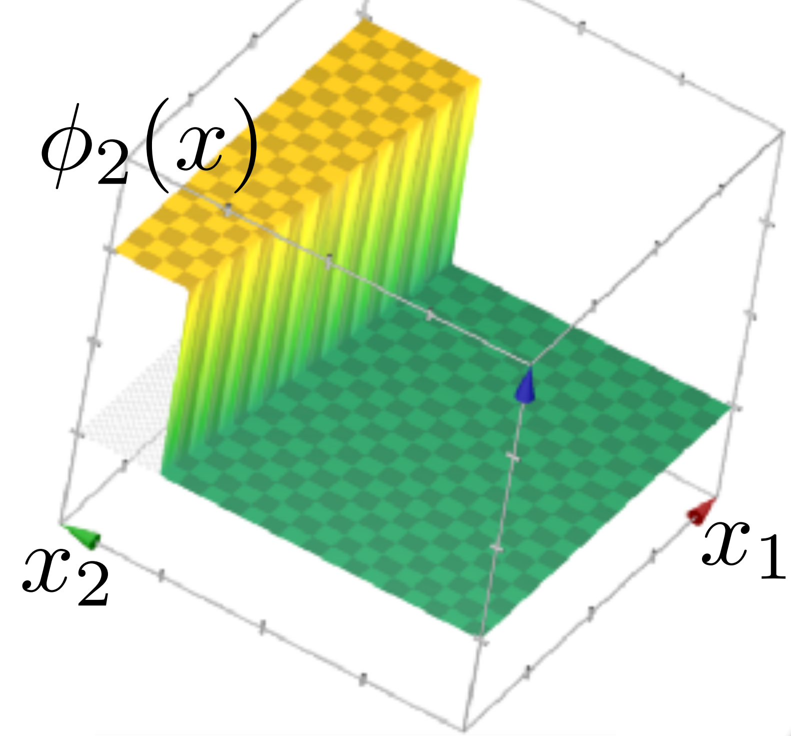


New features: step functions!

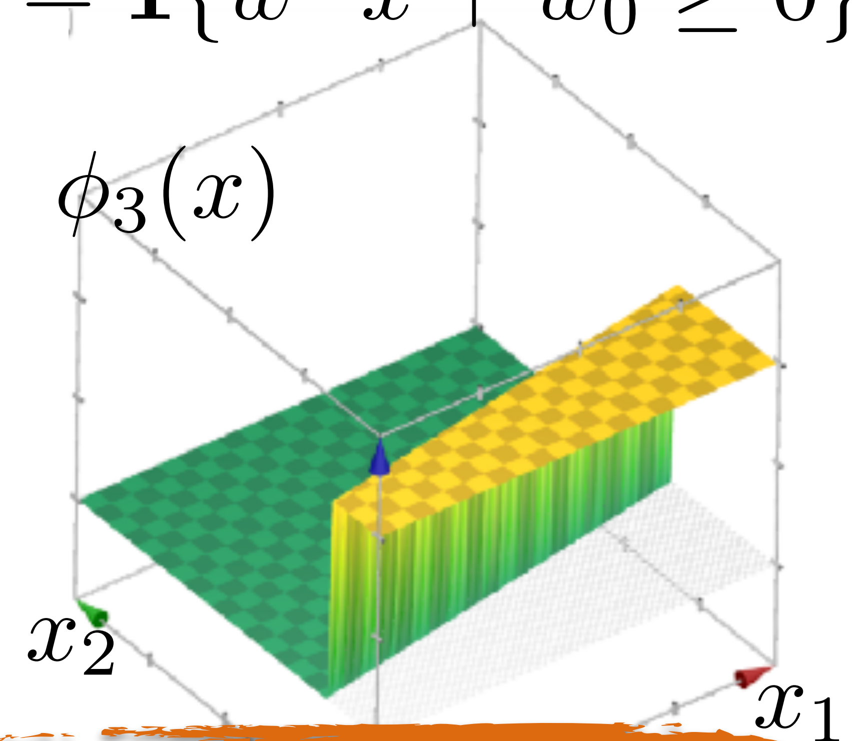
$\phi_1(x) = \mathbf{1}\{w^\top x + w_0 \geq 0\}$



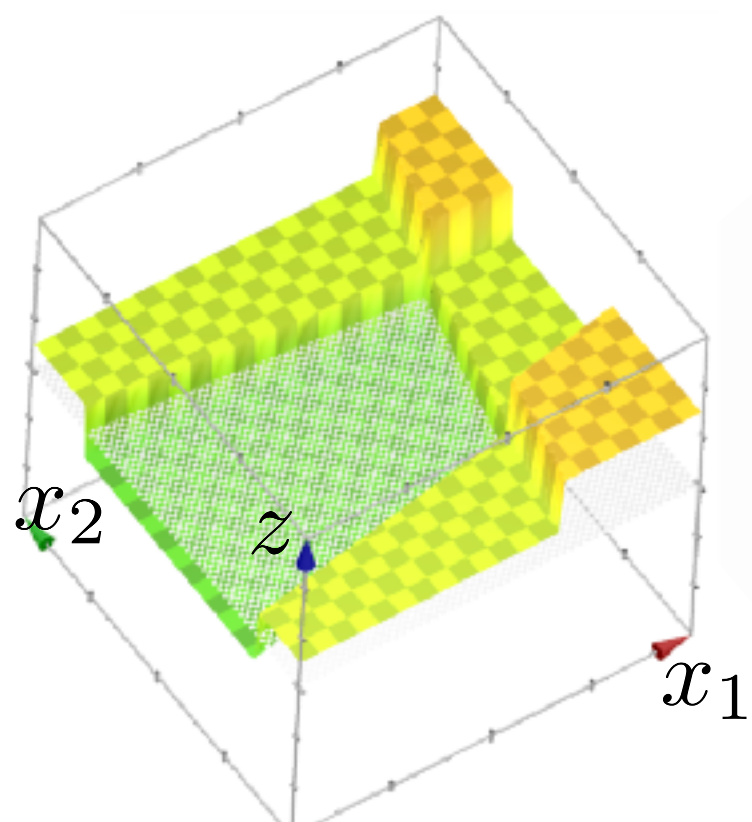
$\phi_2(x) = \mathbf{1}\{\tilde{w}^\top x + \tilde{w}_0 \geq 0\}$

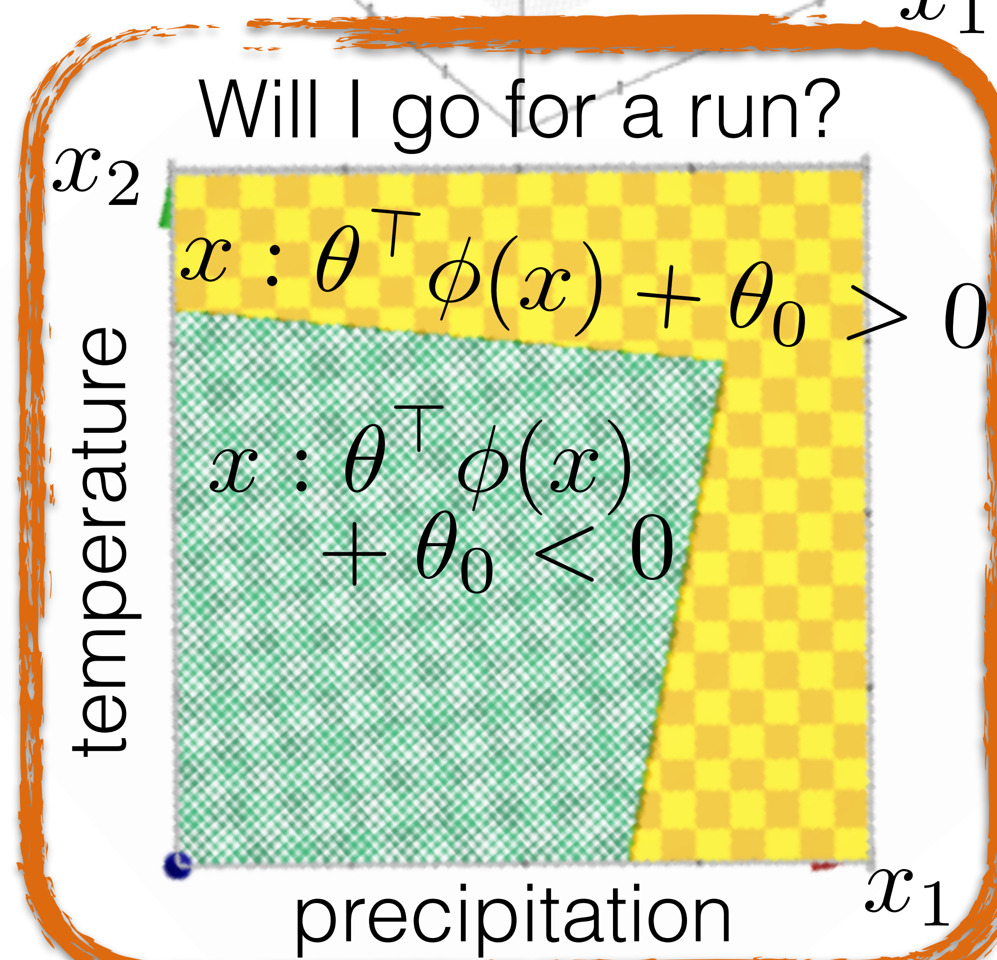


$\phi_3(x) = \mathbf{1}\{\tilde{\tilde{w}}^\top x + \tilde{\tilde{w}}_0 \geq 0\}$



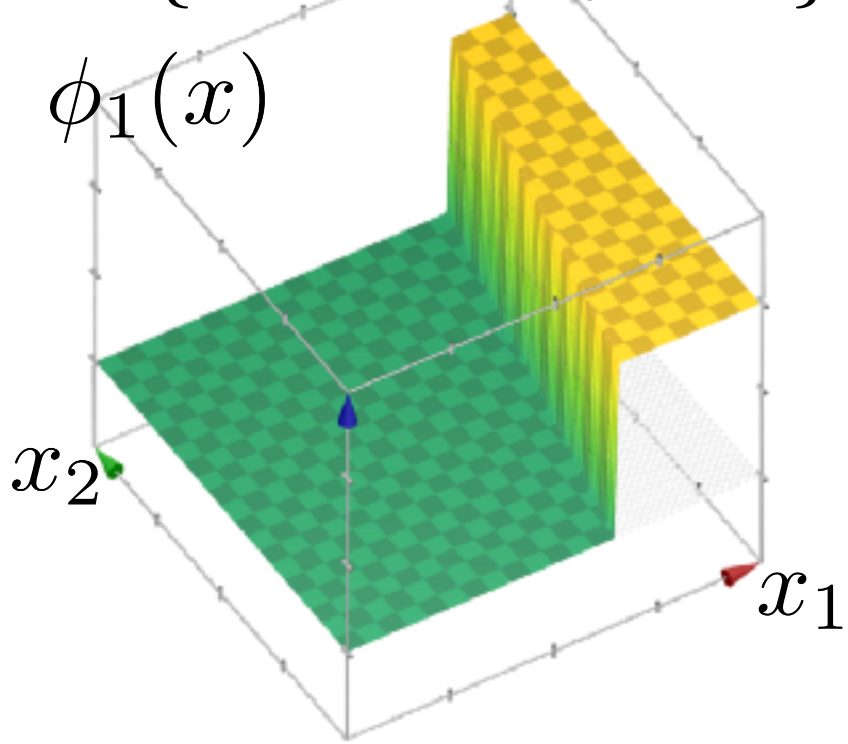
$$
\begin{aligned}
z &= \theta^\top \phi(x) + \theta_0 \\
&= \theta_1\phi_1(x) + \theta_2\phi_2(x) \\
&\quad + \theta_3\phi_3(x) + \theta_0 \\
&= 1 \cdot \phi_1(x) + 1 \cdot \phi_2(x) \\
&\quad + 1 \cdot \phi_3(x) + (-0.5)
\end{aligned}
$$



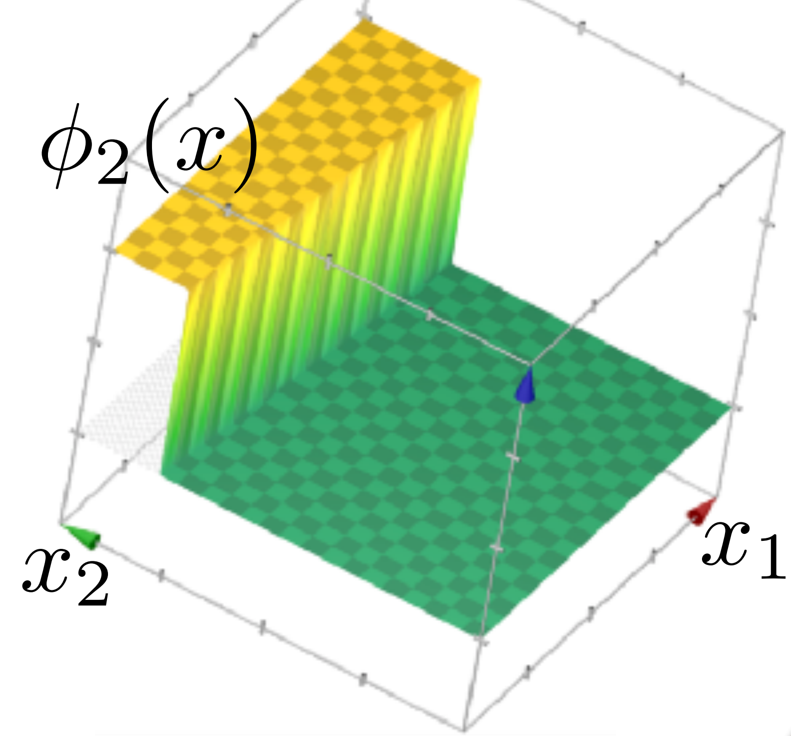


New features: step functions!

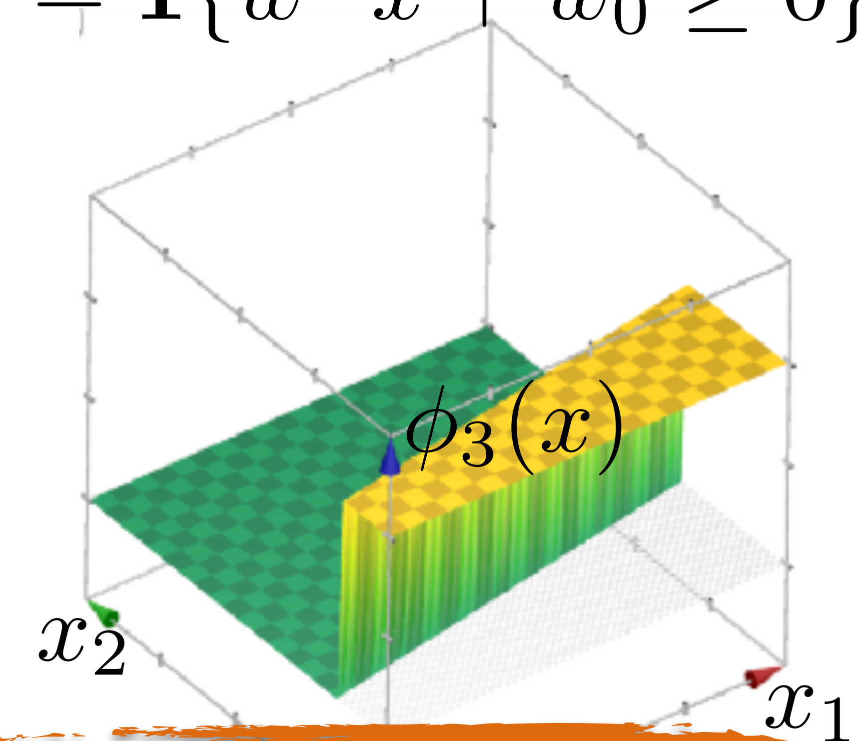
$\phi_1(x) = \mathbf{1}\{w^\top x + w_0 \geq 0\}$



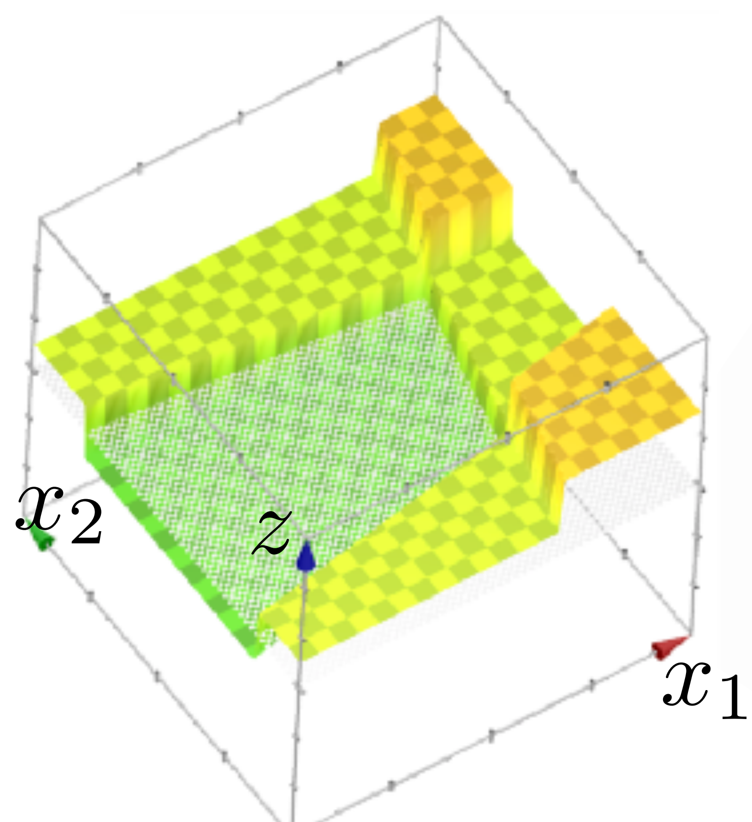
$\phi_2(x) = \mathbf{1}\{\tilde{w}^\top x + \tilde{w}_0 \geq 0\}$

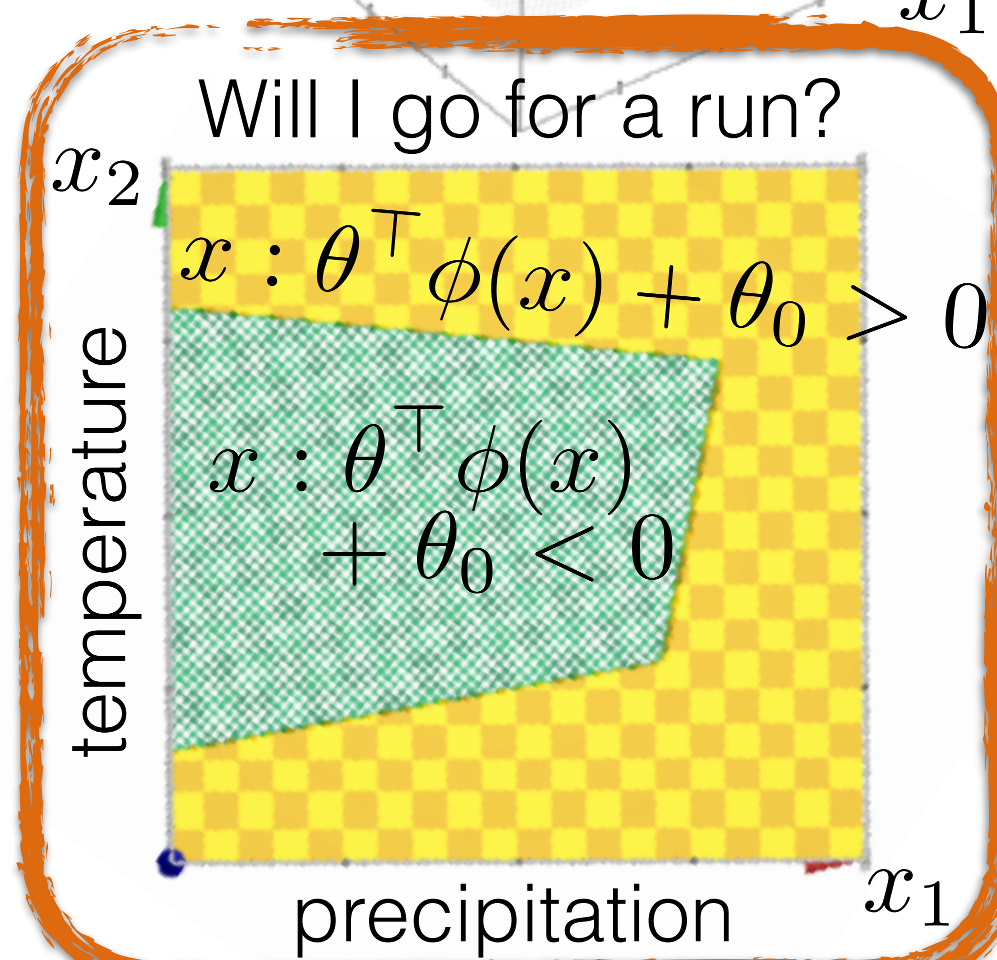


$\phi_3(x) = \mathbf{1}\{\tilde{\tilde{w}}^\top x + \tilde{\tilde{w}}_0 \geq 0\}$



$$
\begin{aligned}
z &= \theta^\top \phi(x) + \theta_0 \\
&= \theta_1 \phi_1(x) + \theta_2 \phi_2(x) \\
&\quad + \theta_3 \phi_3(x) + \theta_0 \\
&= 1 \cdot \phi_1(x) + 1 \cdot \phi_2(x) \\
&\quad + 1 \cdot \phi_3(x) + (-0.5)
\end{aligned}
$$





Let’s get some new notation

Let's get some new notation

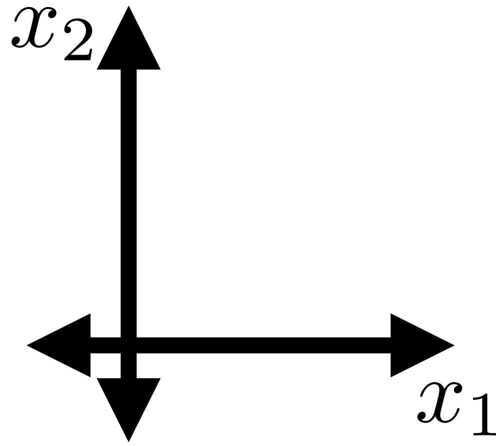
- 1st layer, constructing the features:

Let’s get some new notation

- 1st layer, constructing the features:
  - Input x (a data point)

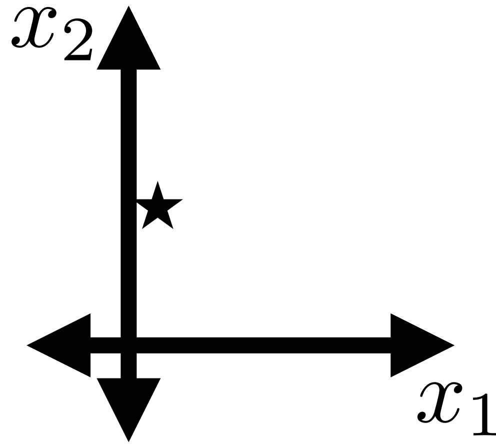
Let's get some new notation

- 1st layer, constructing the features:
  - Input x (a data point)



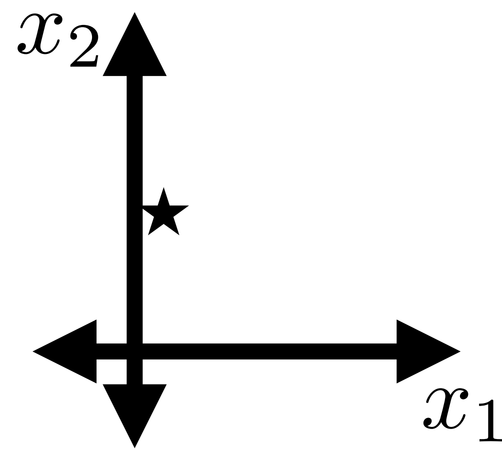
Let's get some new notation

- 1st layer, constructing the features:
  - Input x (a data point)



Let's get some new notation

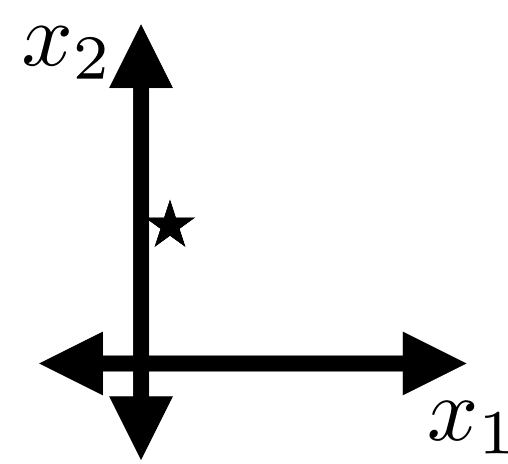
- 1st layer, constructing the features:
  - Input x (a data point): size $m^{(1)} \times 1$



Let's get some new notation

- 1st layer, constructing the features:
  - Input x (a data point): size $m^{(1)} \times 1$

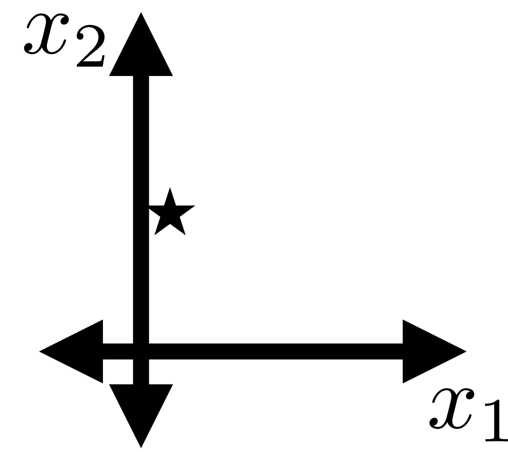
$m^{(1)} = d$



Let's get some new notation

- 1st layer, constructing the features:
  - Input x (a data point): size $m^{(1)} \times 1$
  - Output $A^{(1)}$ (vector of feature values)

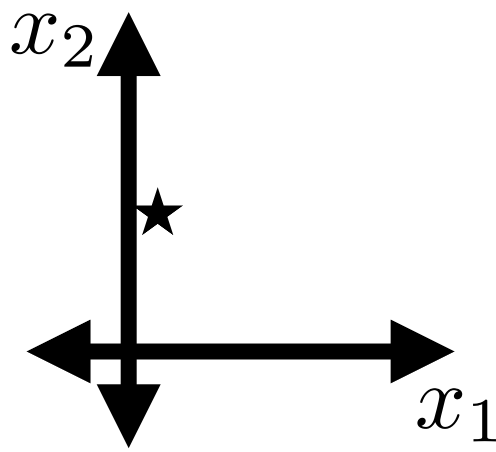
$m^{(1)} = d$

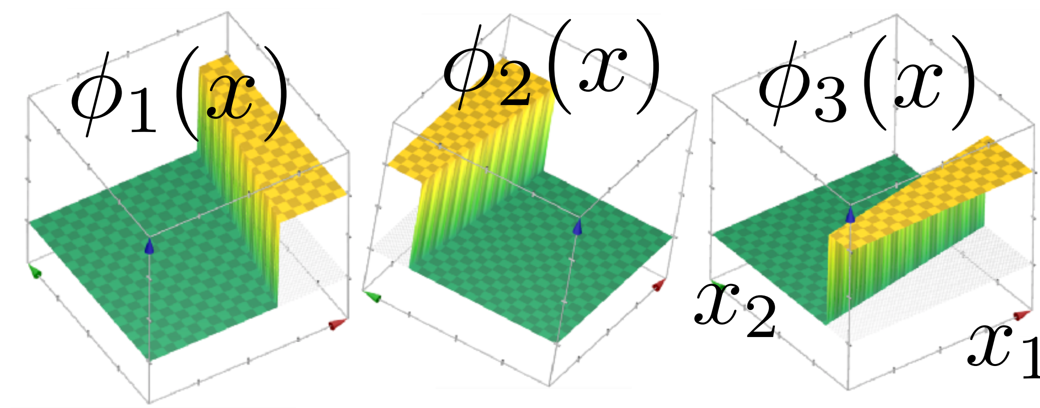


Let's get some new notation

- 1st layer, constructing the features:
  - Input x (a data point): size $m^{(1)} \times 1$
  - Output $A^{(1)}$ (vector of feature values)

$m^{(1)} = d$

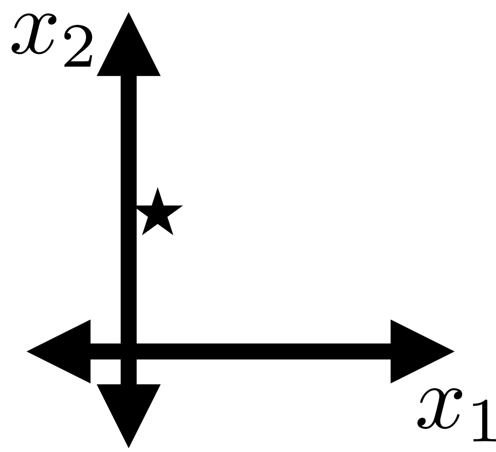


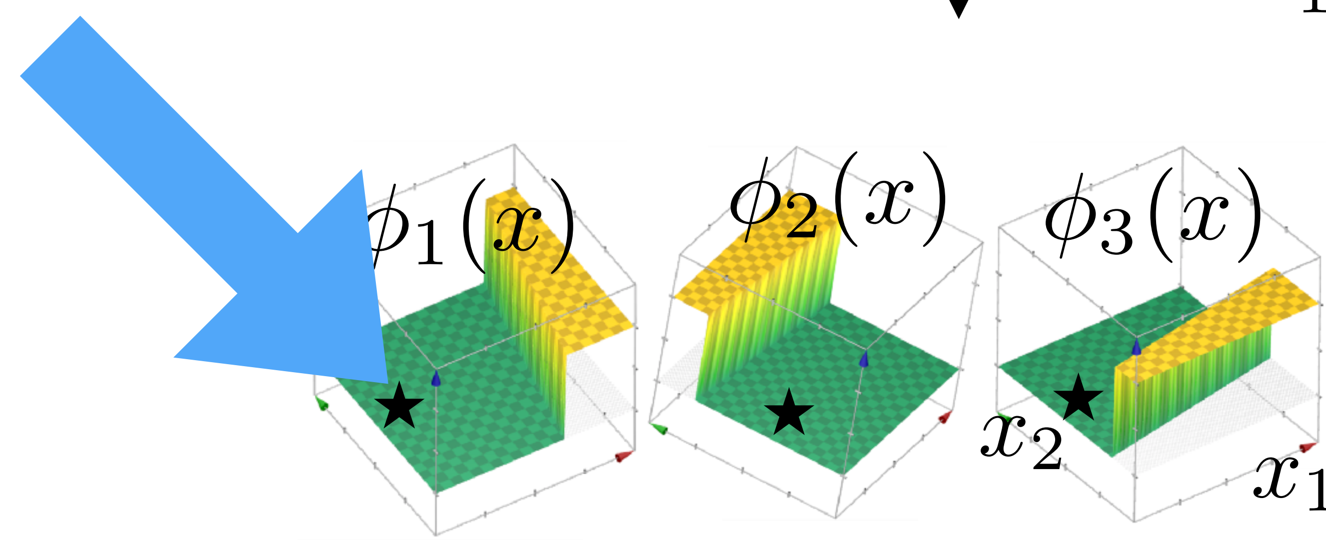


Let's get some new notation

- 1st layer, constructing the features:
  - Input x (a data point): size $m^{(1)} \times 1$
  - Output $A^{(1)}$ (vector of feature values)

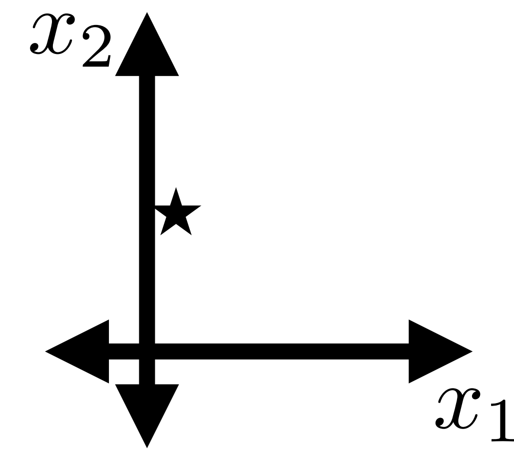
$m^{(1)} = d$

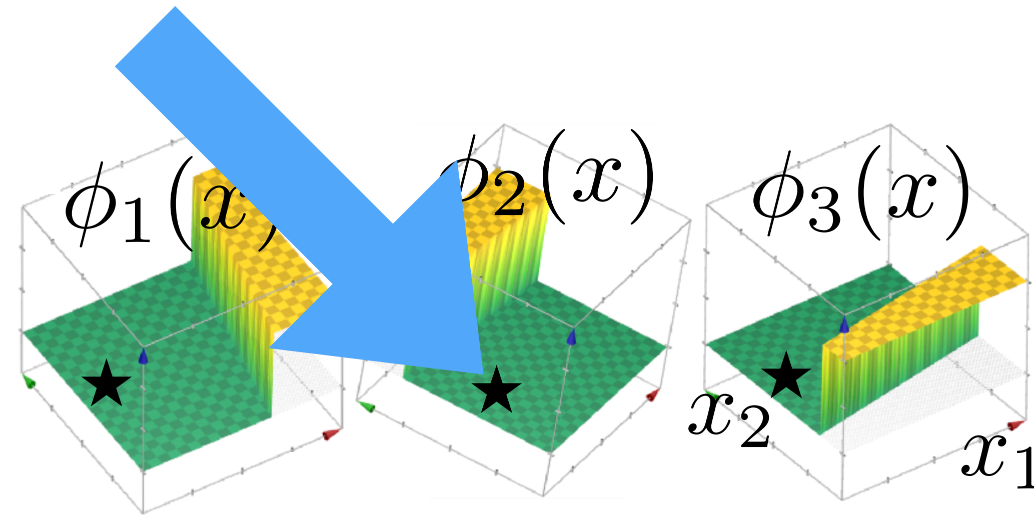




Let's get some new notation

- 1st layer, constructing the features:
  - Input x (a data point): size $m^{(1)} \times 1$ $\boxed{m^{(1)} = d}$
  - Output $A^{(1)}$ (vector of feature values)

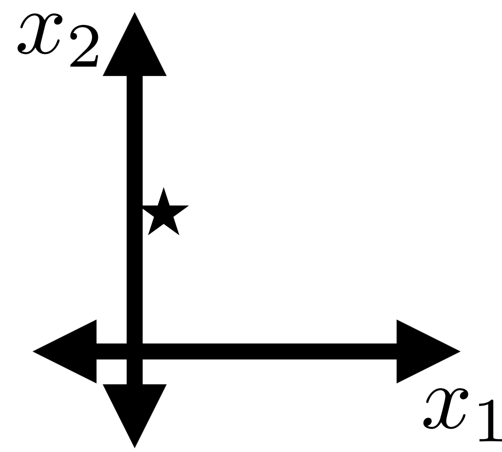


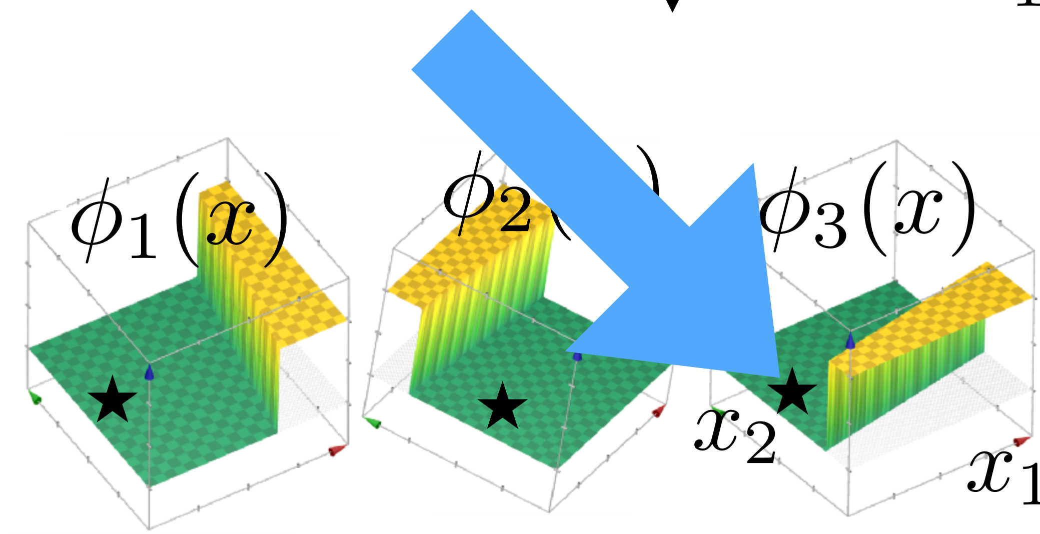


Let’s get some new notation

- 1st layer, constructing the features:
  - Input x (a data point): size $m^{(1)} \times 1$
  - Output $A^{(1)}$ (vector of feature values)

$m^{(1)} = d$

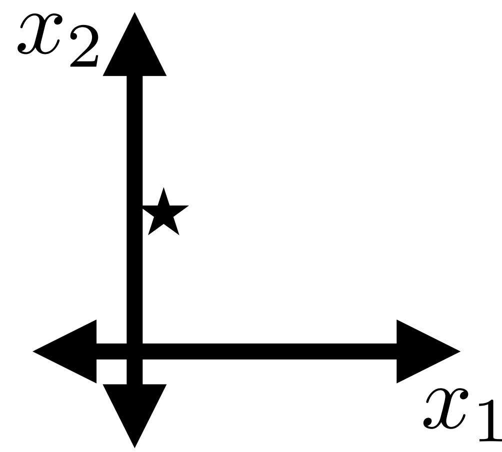


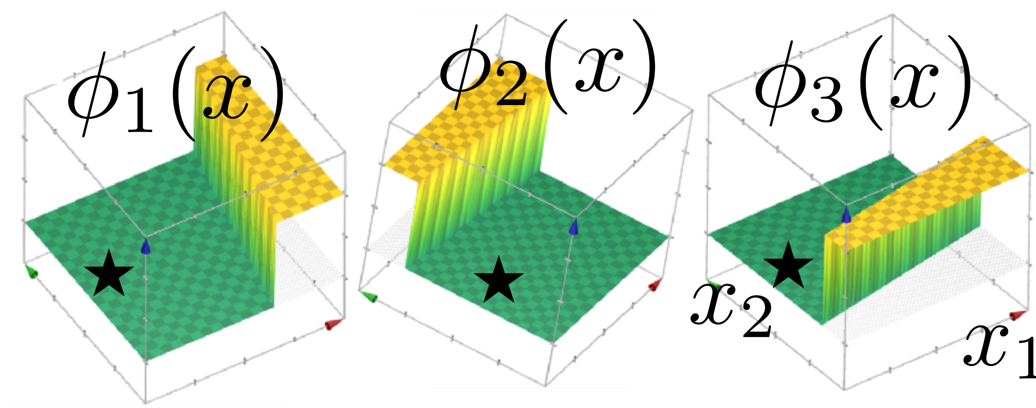


Let's get some new notation

- 1st layer, constructing the features:
  - Input x (a data point): size $m^{(1)} \times 1$
  - Output $A^{(1)}$ (vector of features)

$$m^{(1)} = d$$

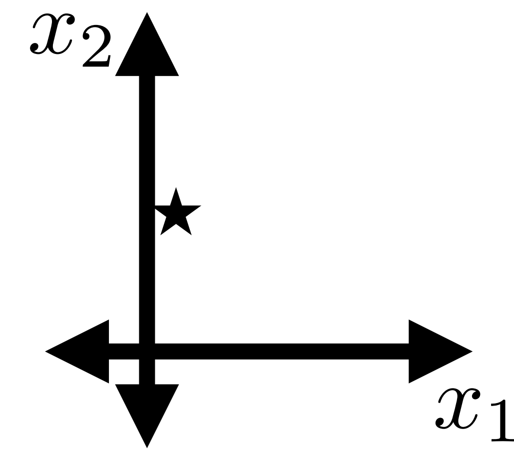


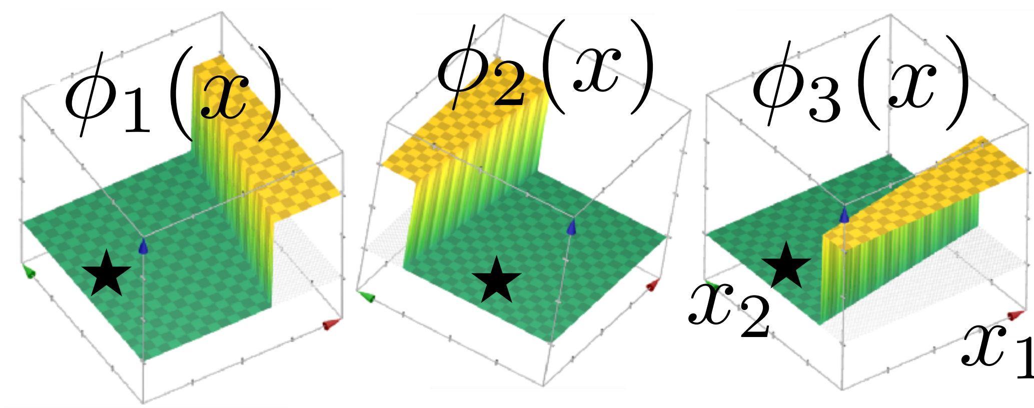


Let's get some new notation

- 1st layer, constructing the features:
  - Input x (a data point): size $m^{(1)} \times 1$
  - Output $A^{(1)}$ (vector of features): size $n^{(1)} \times 1$

$m^{(1)} = d$

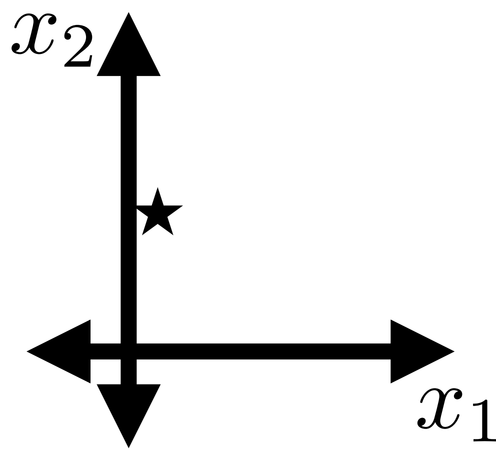


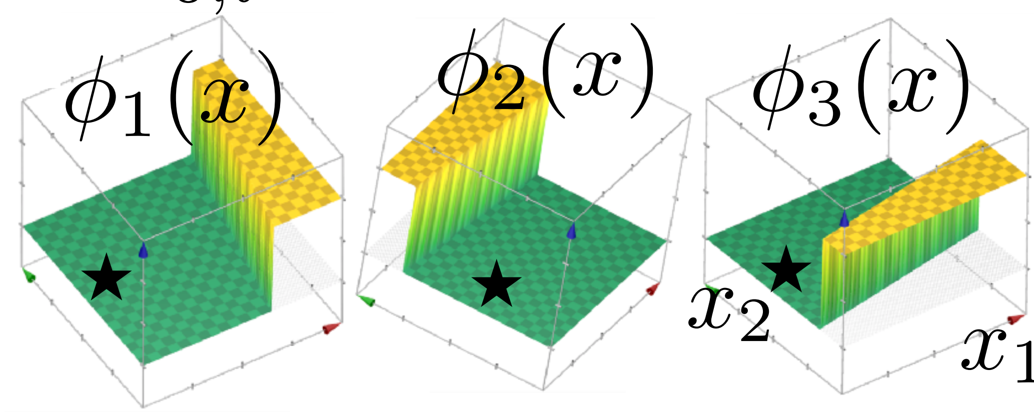


Let's get some new notation

- 1st layer, constructing the features:
  - Input x (a data point): size $m^{(1)} \times 1$
  - Output $A^{(1)}$ (vector of features): size $n^{(1)} \times 1$
  - The ith feature: $A_i^{(1)} = f^{(1)}(w_i^{(1)\top} x + w_{0,i}^{(1)})$

$\boxed{m^{(1)} = d}$

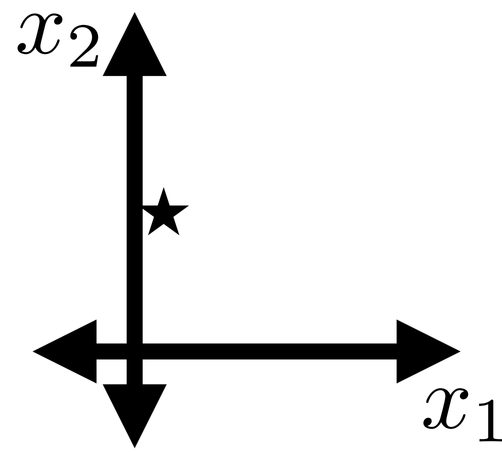


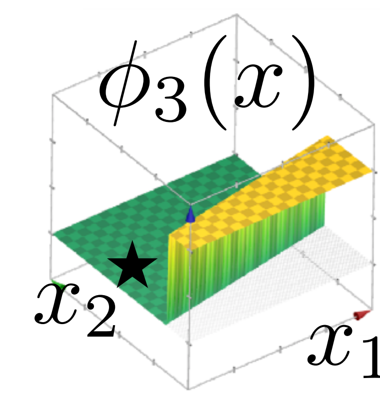


Let's get some new notation

- 1st layer, constructing the features:
  - Input x (a data point): size $m^{(1)} \times 1$

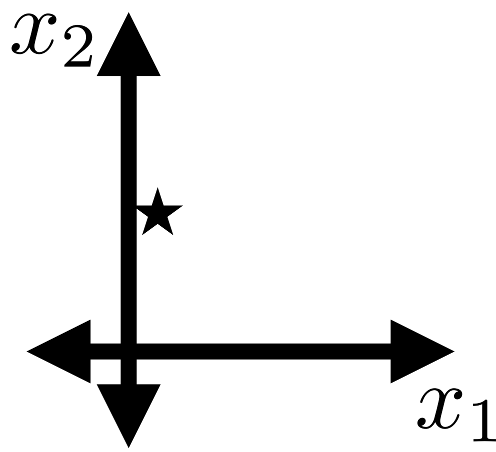
    $m^{(1)} = d$
  - Output $A^{(1)}$ (vector of features): size $n^{(1)} \times 1$
  - The *i*th feature: $A_i^{(1)} = f^{(1)}(w_i^{(1)\top} x + w_{0,i}^{(1)})$

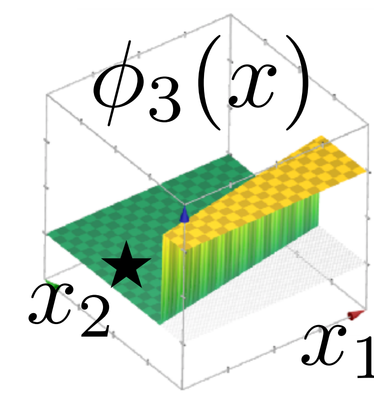




Let's get some new notation

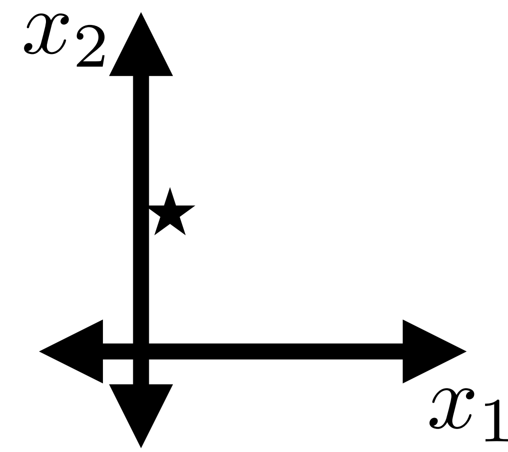
- 1st layer, constructing the features:
  - Input x (a data point): size $m^{(1)} \times 1$ $m^{(1)} = d$
  - Output $A^{(1)}$ (vector of features): size $n^{(1)} \times 1$
  - The ith feature: $A_i^{(1)} = f^{(1)}(w_i^{(1)\top} x + w_{0,i}^{(1)})$

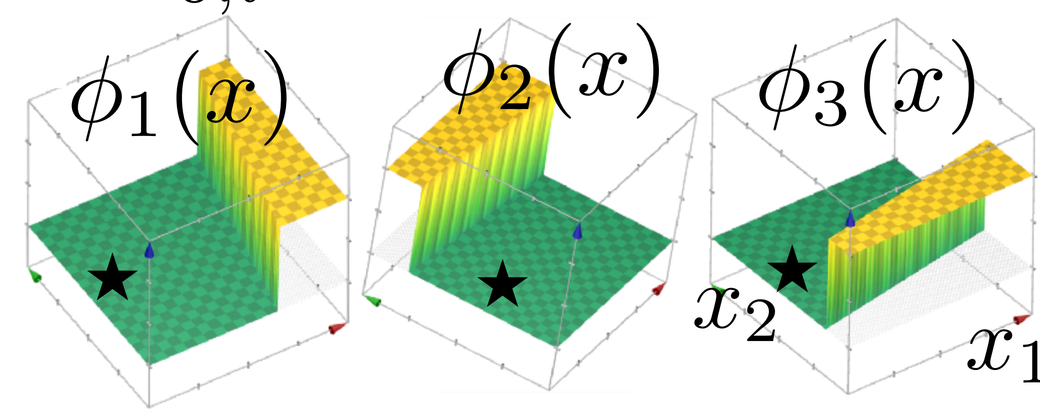




Let's get some new notation

- 1st layer, constructing the features:
  - Input x (a data point): size $m^{(1)} \times 1$ $\boxed{m^{(1)} = d}$
  - Output $A^{(1)}$ (vector of features): size $n^{(1)} \times 1$
  - The *i*th feature: $A_i^{(1)} = f^{(1)}(w_i^{(1)\top} x + w_{0,i}^{(1)})$



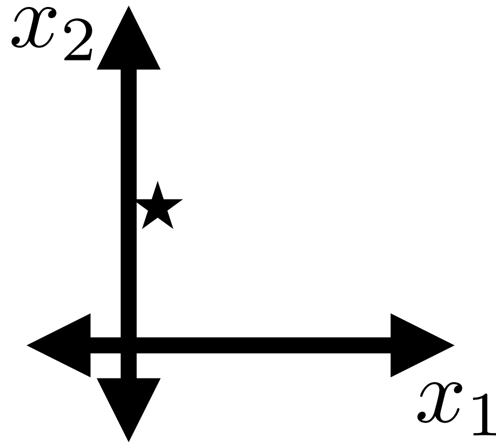


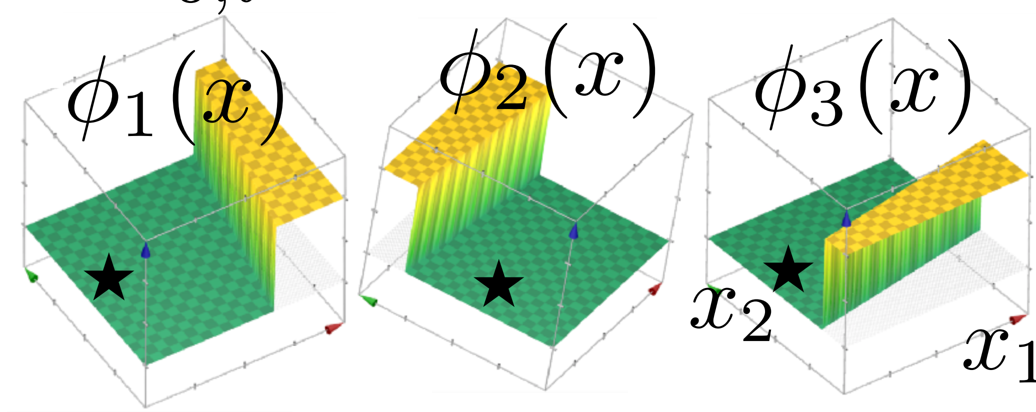
Let's get some new notation

- 1st layer, constructing the features:
  - Input x (a data point): size $m^{(1)} \times 1$

$m^{(1)} = d$

  - Output $A^{(1)}$ (vector of features): size $n^{(1)} \times 1$
  - The ith feature: $A_i^{(1)} = f^{(1)}(w_i^{(1)\top} x + w_{0,i}^{(1)})$
    - All the features at once:

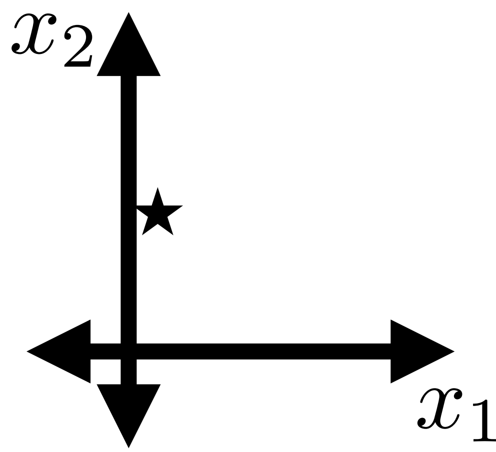


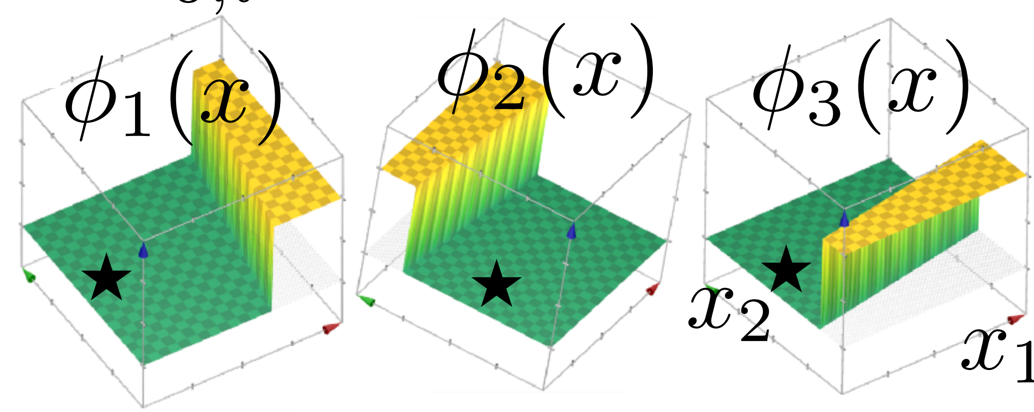


Let's get some new notation

- 1st layer, constructing the features:
  - Input x (a data point): size $m^{(1)} \times 1$
  - Output $A^{(1)}$ (vector of features): size $n^{(1)} \times 1$
  - The *i*th feature: $A_i^{(1)} = f^{(1)}(w_i^{(1)\top} x + w_{0,i}^{(1)})$
    - All the features at once:
      - $A^{(1)} = f^{(1)}(W^{(1)\top} x + W_0^{(1)})$

$m^{(1)} = d$

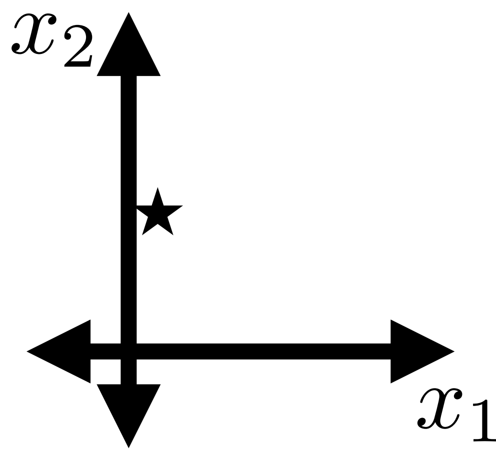


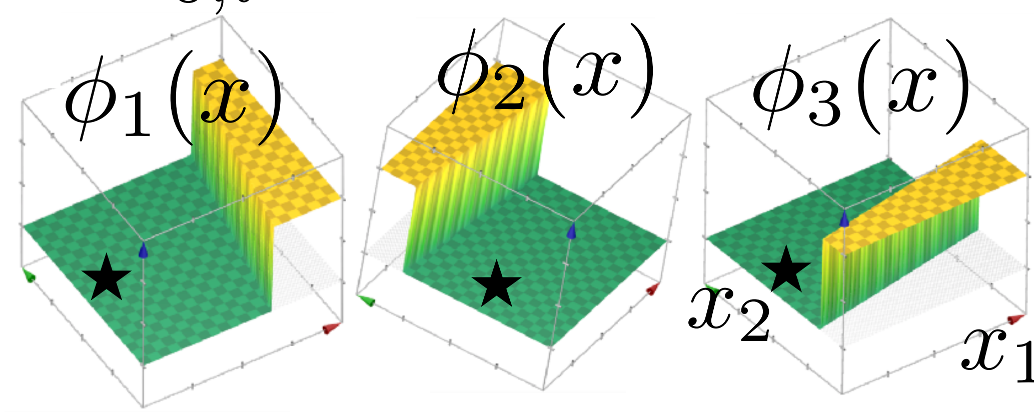


Let's get some new notation

- 1st layer, constructing the features:
  - Input x (a data point): size $m^{(1)} \times 1$
  - Output $A^{(1)}$ (vector of features): size $n^{(1)} \times 1$
  - The ith feature: $A_i^{(1)} = f^{(1)}(w_i^{(1)\top} x + w_{0,i}^{(1)})$
    - All the features at once:
      - $A^{(1)} = f^{(1)}(W^{(1)\top} x + W_0^{(1)})$
      - $W^{(1)} : m^{(1)} \times n^{(1)}$

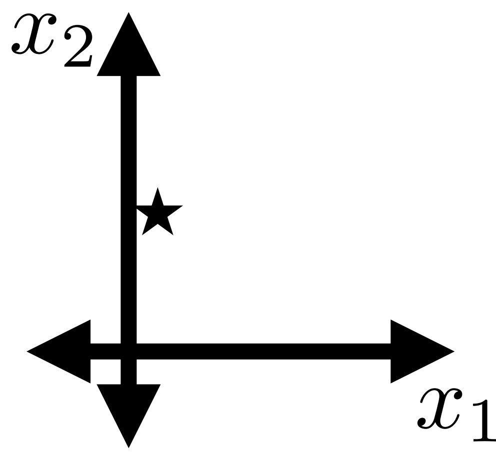
$\boxed{m^{(1)} = d}$



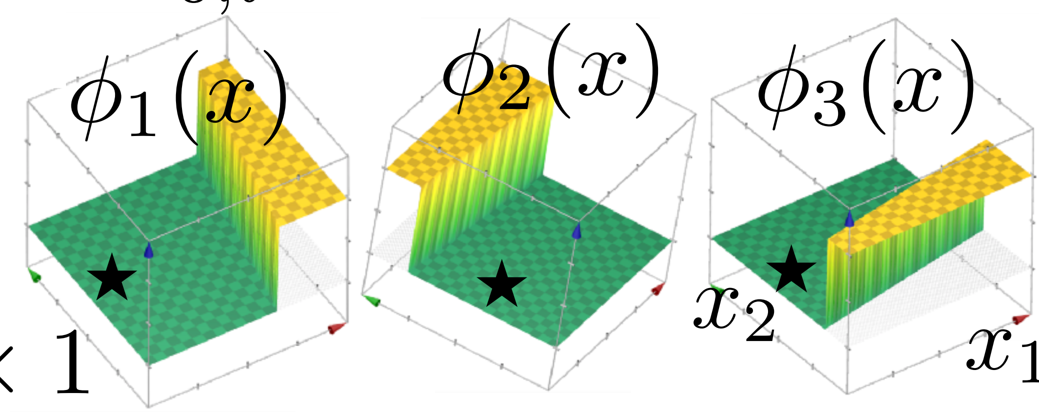


Let's get some new notation

$\boxed{m^{(1)} = d}$



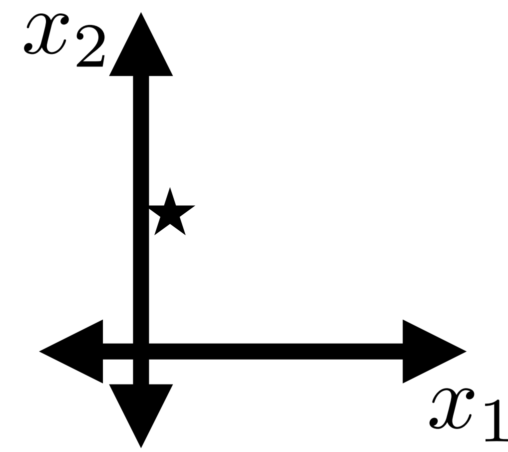
- 1st layer, constructing the features:
  - Input x (a data point): size $m^{(1)} \times 1$
  - Output $A^{(1)}$ (vector of features): size $n^{(1)} \times 1$
  - The *i*th feature: $A_i^{(1)} = f^{(1)}(w_i^{(1)\top} x + w_{0,i}^{(1)})$
    - All the features at once:
      - $A^{(1)} = f^{(1)}(W^{(1)\top} x + W_0^{(1)})$
      - $W^{(1)} : m^{(1)} \times n^{(1)}; W_0^{(1)} : n^{(1)} \times 1$

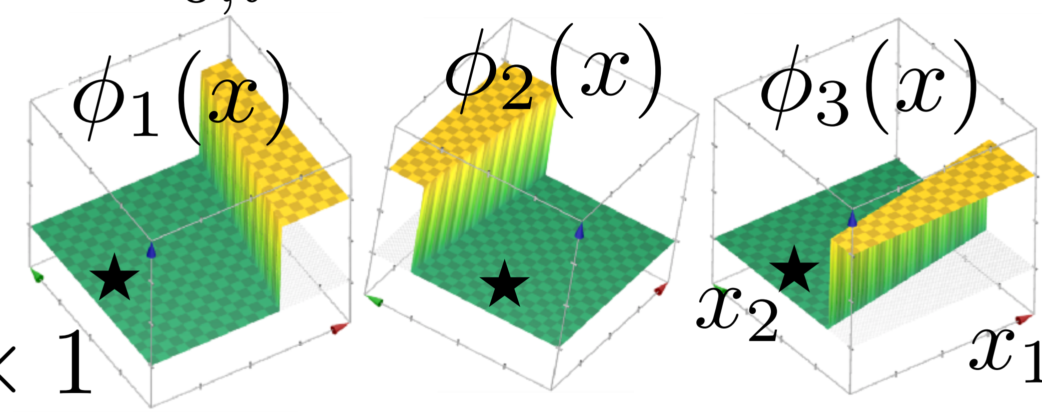


Let's get some new notation

- 1st layer, constructing the features:
  - Input x (a data point): size $m^{(1)} \times 1$
  - Output $A^{(1)}$ (vector of features): size $n^{(1)} \times 1$
  - The ith feature: $A_i^{(1)} = f^{(1)}(w_i^{(1)\top} x + w_{0,i}^{(1)})$
    - All the features at once:
      - $A^{(1)} = f^{(1)}(W^{(1)\top} x + W_0^{(1)})$
      - $W^{(1)} : m^{(1)} \times n^{(1)}; W_0^{(1)} : n^{(1)} \times 1$

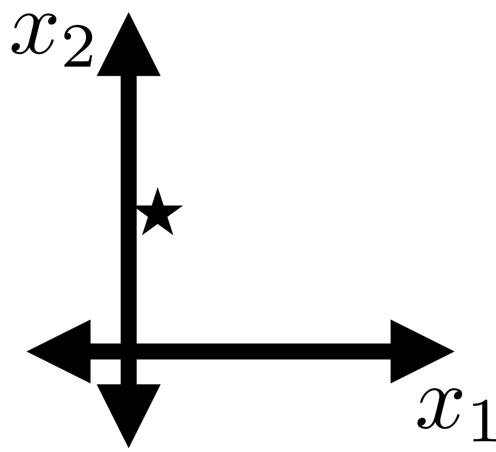
$m^{(1)} = d$

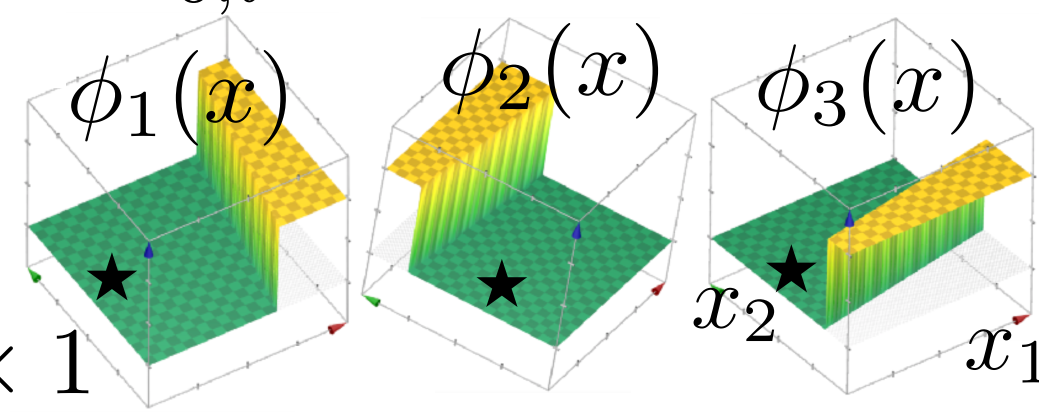




Let's get some new notation

- 1st layer, constructing the features:
  - Input x (a data point): size $m^{(1)} \times 1$ $\boxed{m^{(1)} = d}$
  - Output $A^{(1)}$ (vector of features): size $n^{(1)} \times 1$
  - The *i*th feature: $A_i^{(1)} = f^{(1)}(w_i^{(1)\top} x + w_{0,i}^{(1)})$
    - All the features at once:
      - $A^{(1)} = f^{(1)}(W^{(1)\top} x + W_0^{(1)})$
      - $W^{(1)} : m^{(1)} \times n^{(1)}; W_0^{(1)} : n^{(1)} \times 1$

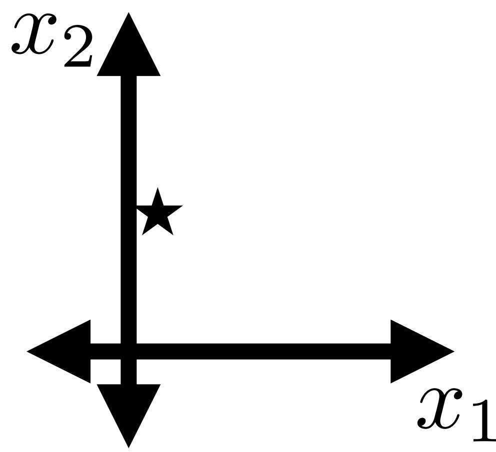


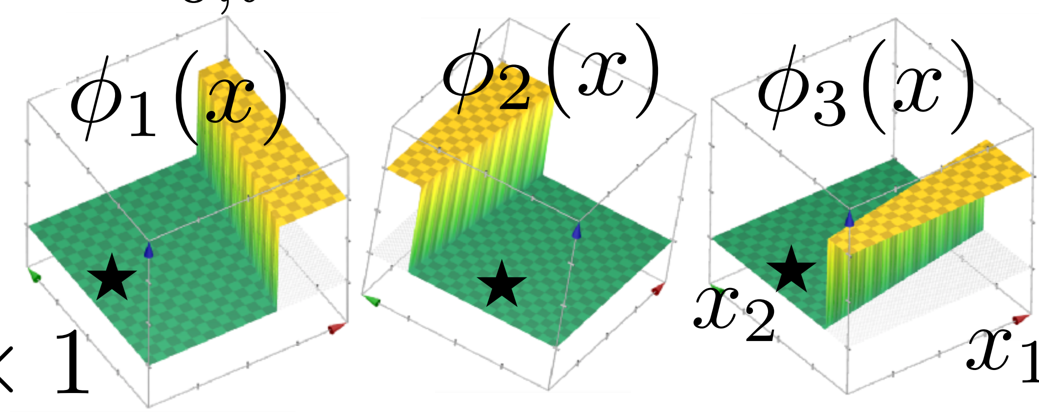


Let's get some new notation

- 1st layer, constructing the features:
  - Input x (a data point): size $m^{(1)} \times 1$
  - Output $A^{(1)}$ (vector of features): size $n^{(1)} \times 1$
  - The *i*th feature: $A_i^{(1)} = f^{(1)}(w_i^{(1)\top} x + w_{0,i}^{(1)})$
    - All the features at once:
      - $A^{(1)} = f^{(1)}(W^{(1)\top} x + W_0^{(1)})$
      - $W^{(1)} : m^{(1)} \times n^{(1)}; W_0^{(1)} : n^{(1)} \times 1$

$\boxed{m^{(1)} = d}$

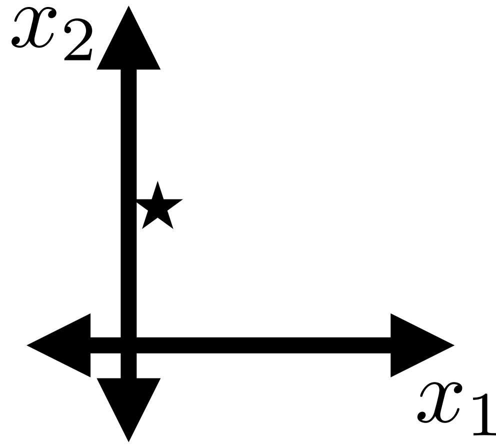


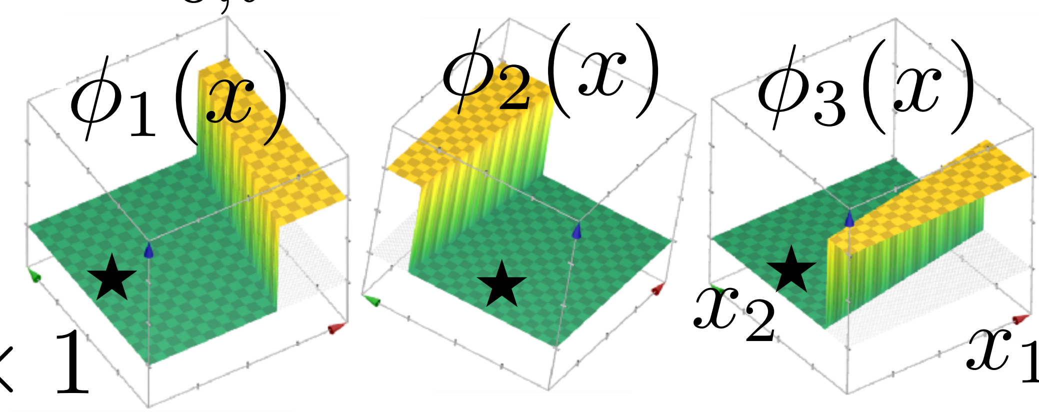


$\boxed{f^{(1)} \text{ is applied componentwise!}}$

Let's get some new notation

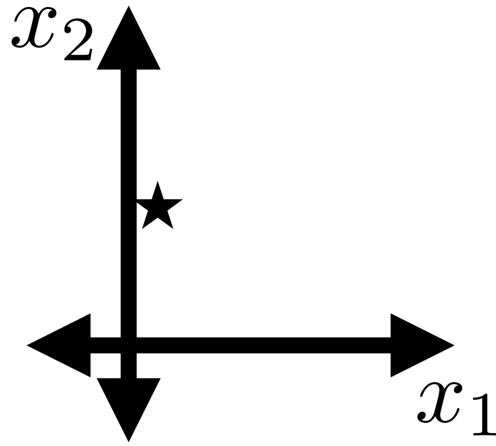
- 1st layer, constructing the features:
  - Input x (a data point): size $m^{(1)} \times 1$ $\boxed{m^{(1)} = d}$
  - Output $A^{(1)}$ (vector of features): size $n^{(1)} \times 1$
  - The *i*th feature: $A_i^{(1)} = f^{(1)}(w_i^{(1)\top} x + w_{0,i}^{(1)})$
    - All the features at once:
      - $A^{(1)} = f^{(1)}(W^{(1)\top} x + W_0^{(1)})$
      - $W^{(1)} : m^{(1)} \times n^{(1)}; W_0^{(1)} : n^{(1)} \times 1$



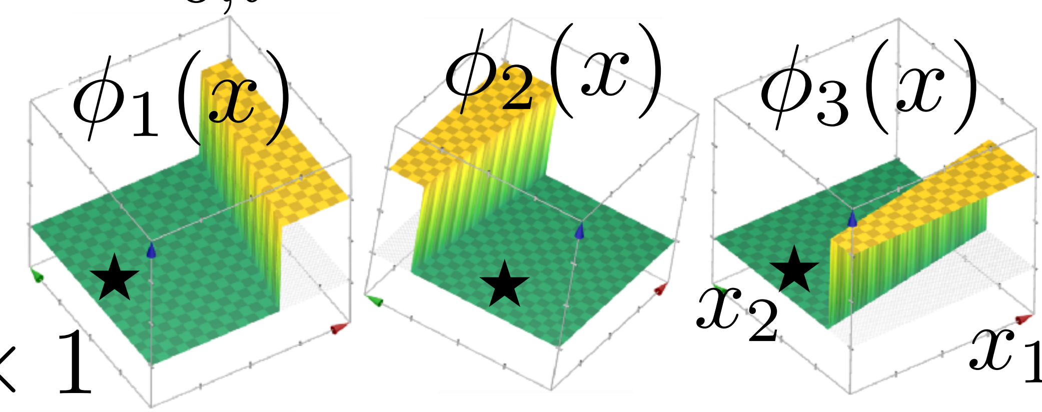


Let's get some new notation

$\boxed{m^{(1)} = d}$



- 1st layer, constructing the features:
  - Input x (a data point): size $m^{(1)} \times 1$
  - Output $A^{(1)}$ (vector of features): size $n^{(1)} \times 1$
  - The *i*th feature: $A_i^{(1)} = f^{(1)}(w_i^{(1)\top} x + w_{0,i}^{(1)})$
    - All the features at once:
      - $A^{(1)} = f^{(1)}(W^{(1)\top} x + W_0^{(1)})$
      - $W^{(1)} : m^{(1)} \times n^{(1)}; W_0^{(1)} : n^{(1)} \times 1$

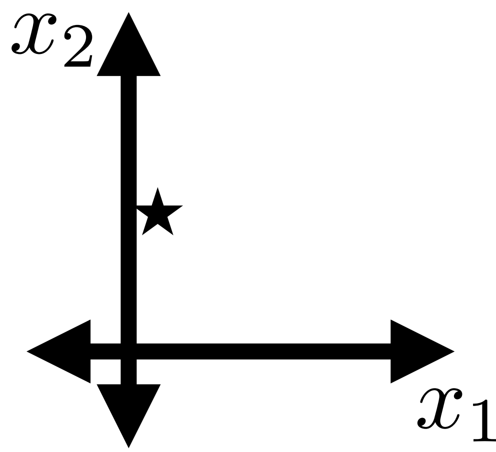


- 2nd layer, assigning a label (or labels):

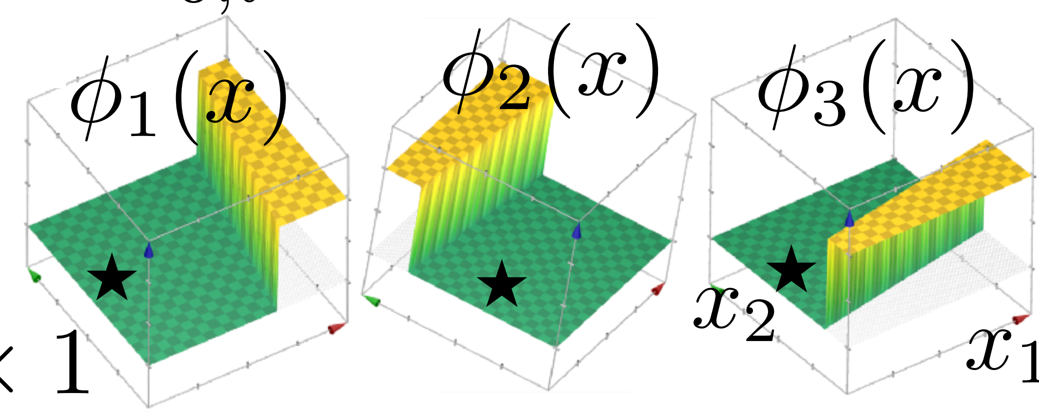
Let's get some new notation

- 1st layer, constructing the features:
  - Input x (a data point): size $m^{(1)} \times 1$

$m^{(1)} = d$



  - Output $A^{(1)}$ (vector of features): size $n^{(1)} \times 1$
  - The *i*th feature: $A_i^{(1)} = f^{(1)}(w_i^{(1)\top} x + w_{0,i}^{(1)})$
    - All the features at once:
      - $A^{(1)} = f^{(1)}(W^{(1)\top} x + W_0^{(1)})$
      - $W^{(1)} : m^{(1)} \times n^{(1)}; W_0^{(1)} : n^{(1)} \times 1$

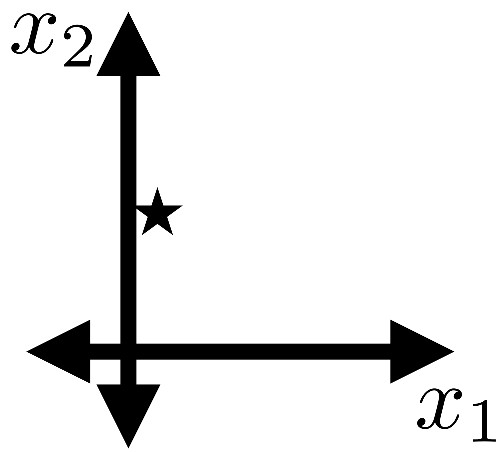


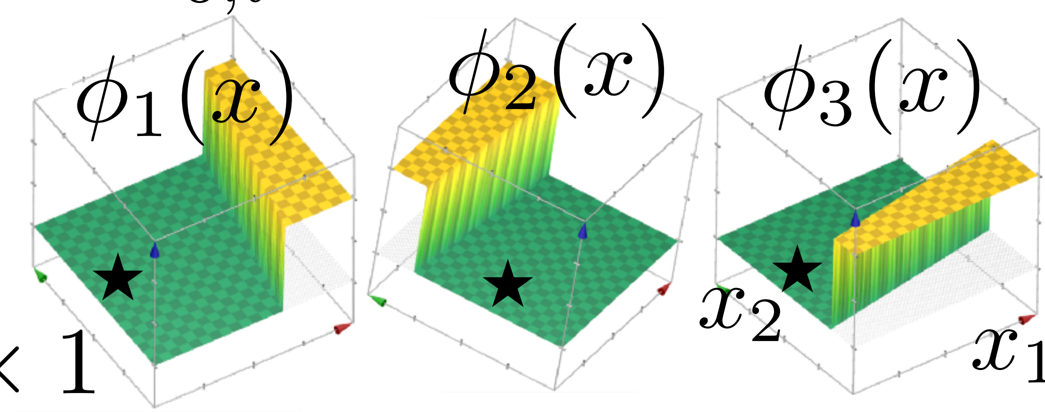
- 2nd layer, assigning a label (or labels):
  - Input (the features)

Let's get some new notation

$m^{(1)} = d$

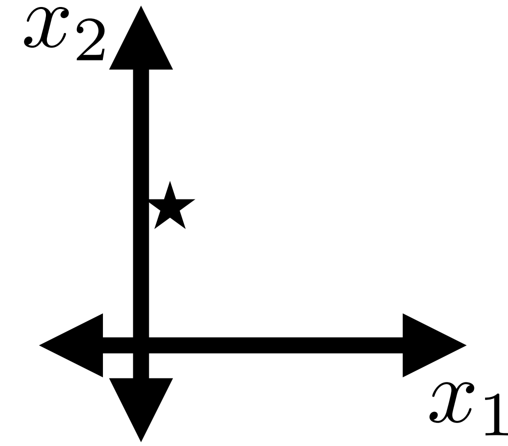
- 1st layer, constructing the features:
  - Input x (a data point): size $m^{(1)} \times 1$
  - Output $A^{(1)}$ (vector of features): size $n^{(1)} \times 1$
  - The *i*th feature: $A_i^{(1)} = f^{(1)}(w_i^{(1)\top} x + w_{0,i}^{(1)})$
    - All the features at once:
      - $A^{(1)} = f^{(1)}(W^{(1)\top} x + W_0^{(1)})$
      - $W^{(1)} : m^{(1)} \times n^{(1)}; W_0^{(1)} : n^{(1)} \times 1$
- 2nd layer, assigning a label (or labels):
  - Input (the features): size $m^{(2)} \times 1$

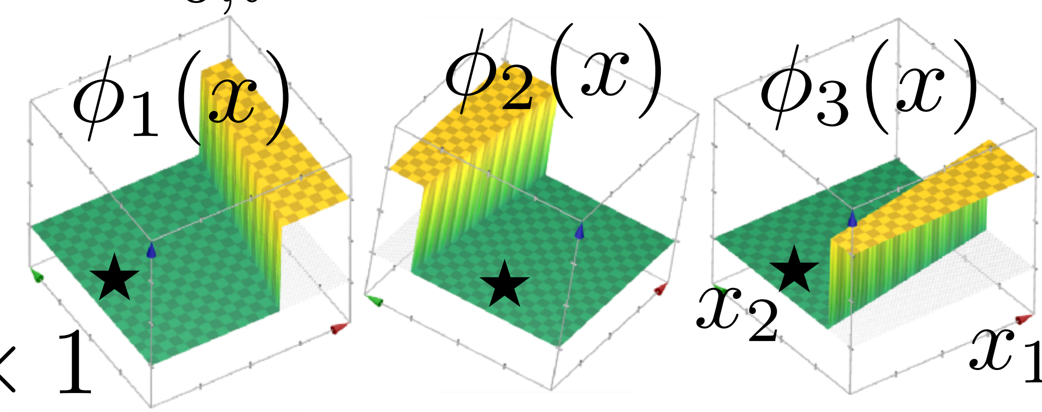




Let's get some new notation

- 1st layer, constructing the features:
  - Input x (a data point): size $m^{(1)} \times 1$ $\boxed{m^{(1)} = d}$
  - Output $A^{(1)}$ (vector of features): size $n^{(1)} \times 1$
  - The *i*th feature: $A_i^{(1)} = f^{(1)}(w_i^{(1)\top} x + w_{0,i}^{(1)})$
    - All the features at once:
      - $A^{(1)} = f^{(1)}(W^{(1)\top} x + W_0^{(1)})$
      - $W^{(1)} : m^{(1)} \times n^{(1)}; W_0^{(1)} : n^{(1)} \times 1$

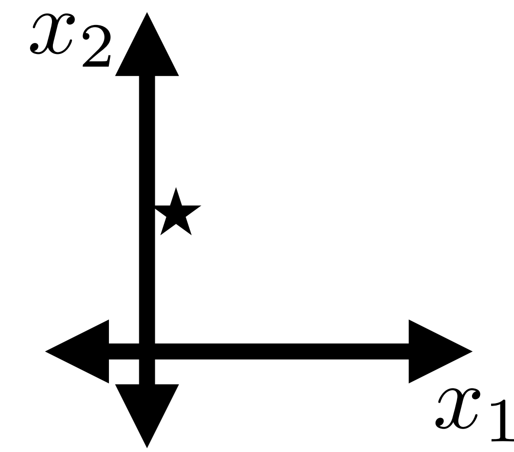


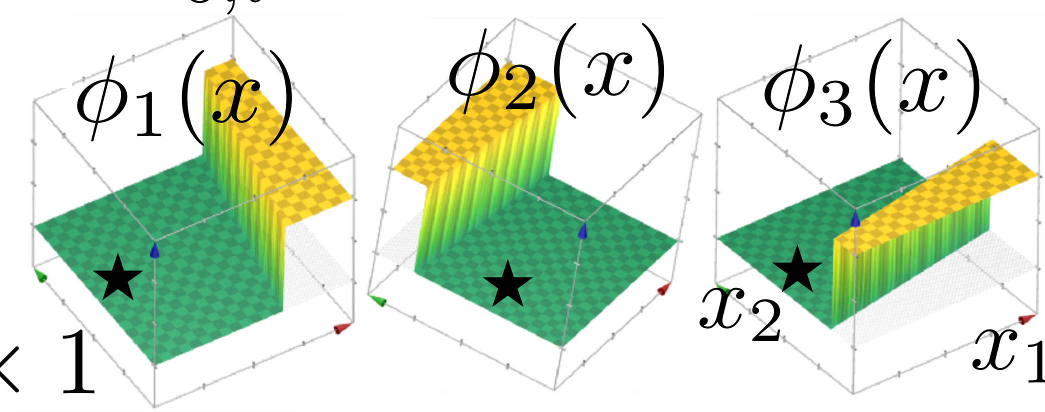


- 2nd layer, assigning a label (or labels):
  - Input (the features): size $m^{(2)} \times 1$ $\boxed{m^{(2)} = n^{(1)}}$

Let's get some new notation

- 1st layer, constructing the features:
  - Input x (a data point): size $m^{(1)} \times 1$ $m^{(1)} = d$
  - Output $A^{(1)}$ (vector of features): size $n^{(1)} \times 1$
  - The *i*th feature: $A_i^{(1)} = f^{(1)}(w_i^{(1)\top} x + w_{0,i}^{(1)})$
    - All the features at once:
      - $A^{(1)} = f^{(1)}(W^{(1)\top} x + W_0^{(1)})$
      - $W^{(1)} : m^{(1)} \times n^{(1)}; W_0^{(1)} : n^{(1)} \times 1$

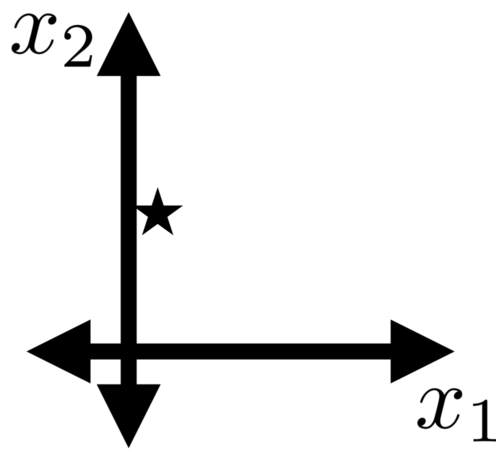


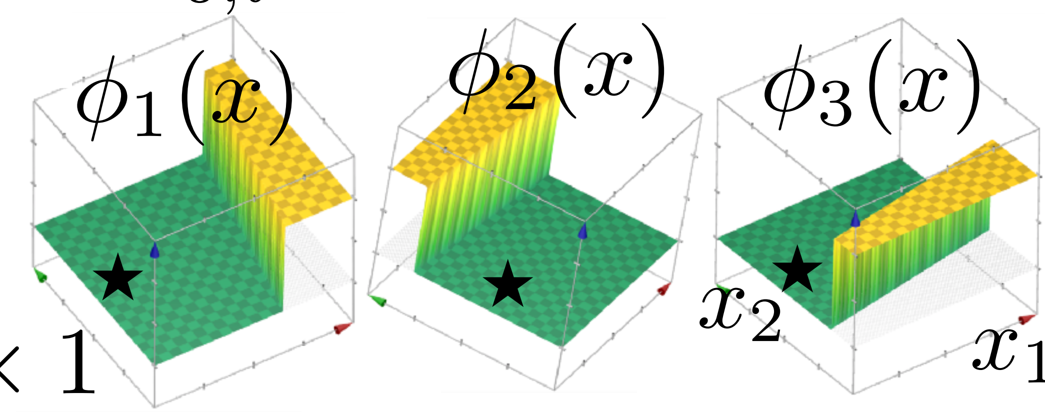


- 2nd layer, assigning a label (or labels):
  - Input (the features): size $m^{(2)} \times 1$ $m^{(2)} = n^{(1)}$
  - Output $A^{(2)}$(vector of labels)

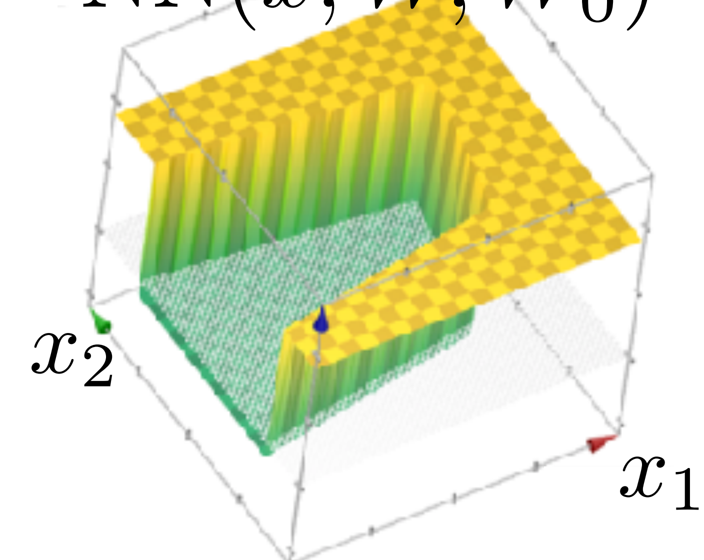
Let's get some new notation

- 1st layer, constructing the features:
  - Input x (a data point): size $m^{(1)} \times 1$ $\boxed{m^{(1)} = d}$
  - Output $A^{(1)}$ (vector of features): size $n^{(1)} \times 1$
  - The ith feature: $A_i^{(1)} = f^{(1)}(w_i^{(1)\top} x + w_{0,i}^{(1)})$
    - All the features at once:
      - $A^{(1)} = f^{(1)}(W^{(1)\top} x + W_0^{(1)})$
      - $W^{(1)} : m^{(1)} \times n^{(1)}; W_0^{(1)} : n^{(1)} \times 1$
- 2nd layer, assigning a label (or labels):
  - Input (the features): size $m^{(2)} \times 1$ $\boxed{m^{(2)} = n^{(1)}}$
  - Output $A^{(2)}$(vector of labels)



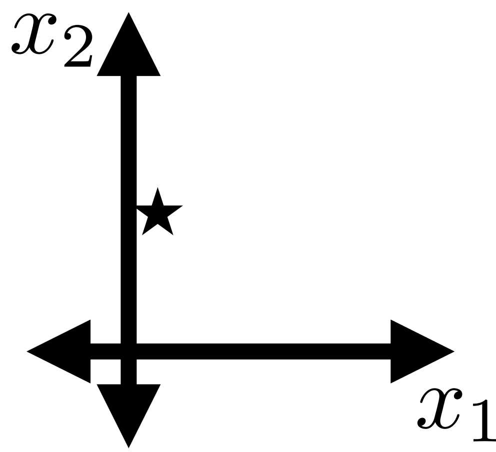


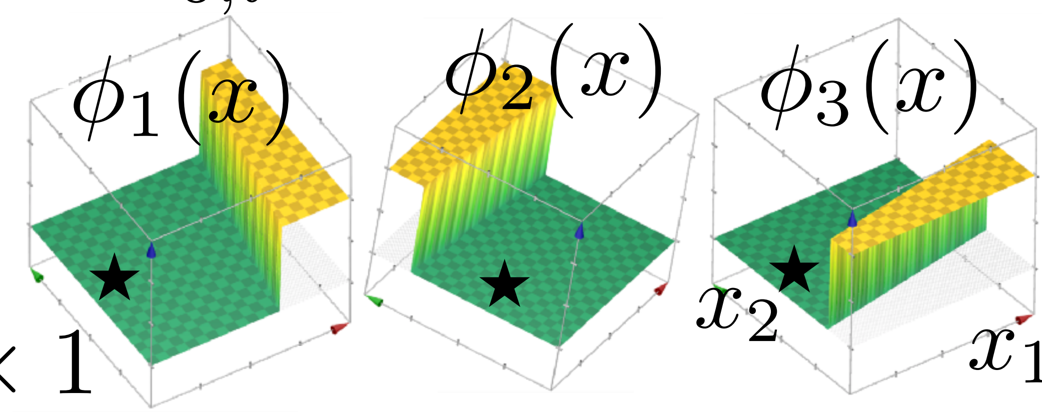
$\mathrm{NN}(x; W, W_0)$



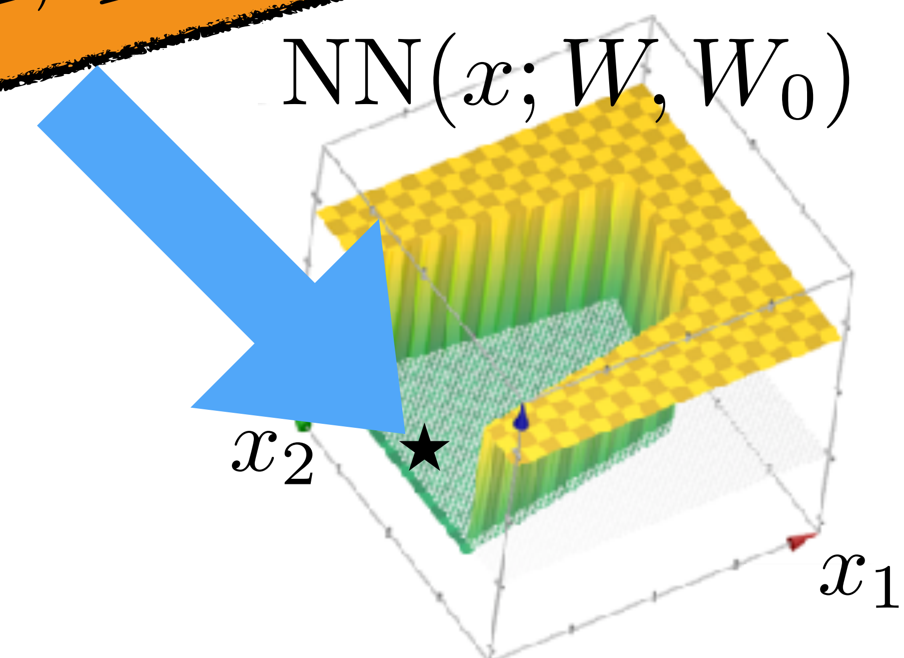
Let's get some new notation

- 1st layer, constructing the features:
  - Input x (a data point): size $m^{(1)} \times 1$ $\boxed{m^{(1)} = d}$
  - Output $A^{(1)}$ (vector of features): size $n^{(1)} \times 1$
  - The *i*th feature: $A_i^{(1)} = f^{(1)}(w_i^{(1)\top} x + w_{0,i}^{(1)})$
    - All the features at once:
      - $A^{(1)} = f^{(1)}(W^{(1)\top} x + W_0^{(1)})$
      - $W^{(1)} : m^{(1)} \times n^{(1)}; W_0^{(1)} : n^{(1)} \times 1$



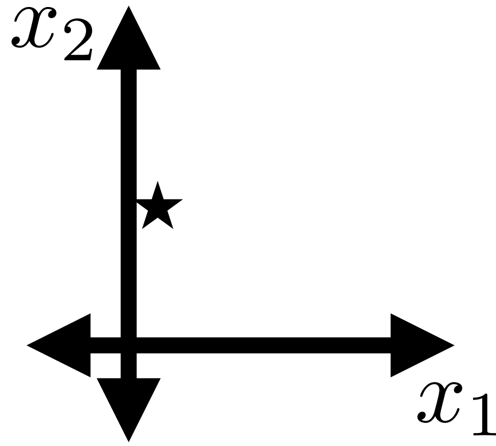


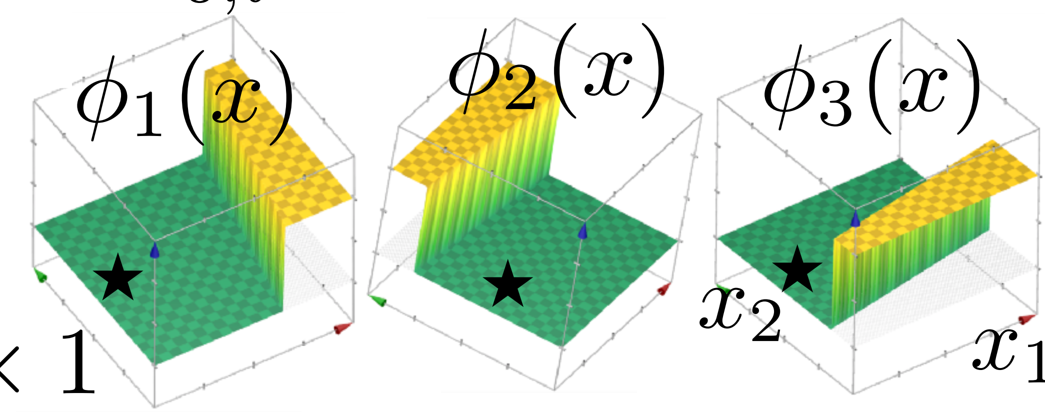
- 2nd layer, assigning a label (or labels):
  - Input (the features): size $m^{(2)} \times 1$ $\boxed{m^{(2)} = n^{(1)}}$
  - Output $A^{(2)}$(vector of labels)



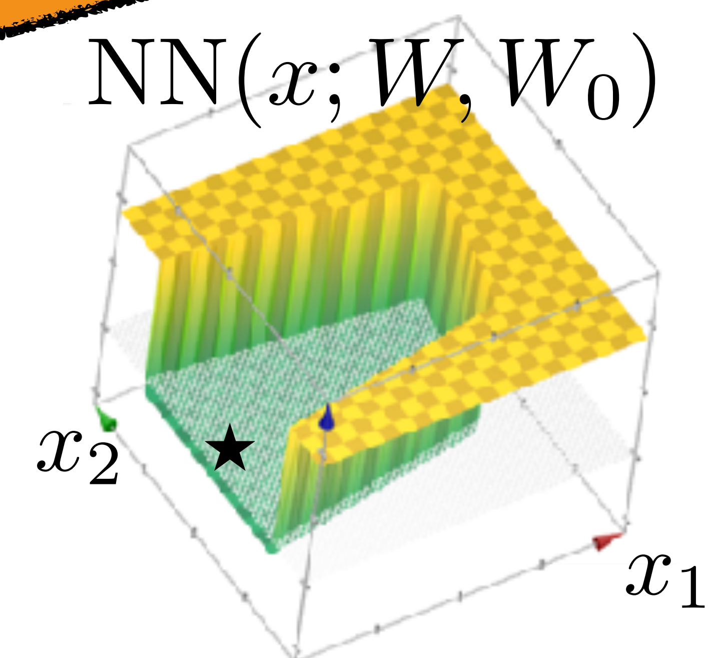
Let's get some new notation

- 1st layer, constructing the features:
  - Input x (a data point): size $m^{(1)} \times 1$ $\boxed{m^{(1)} = d}$
  - Output $A^{(1)}$ (vector of features): size $n^{(1)} \times 1$
  - The ith feature: $A_i^{(1)} = f^{(1)}(w_i^{(1)\top} x + w_{0,i}^{(1)})$
    - All the features at once:
      - $A^{(1)} = f^{(1)}(W^{(1)\top} x + W_0^{(1)})$
      - $W^{(1)} : m^{(1)} \times n^{(1)}; W_0^{(1)} : n^{(1)} \times 1$



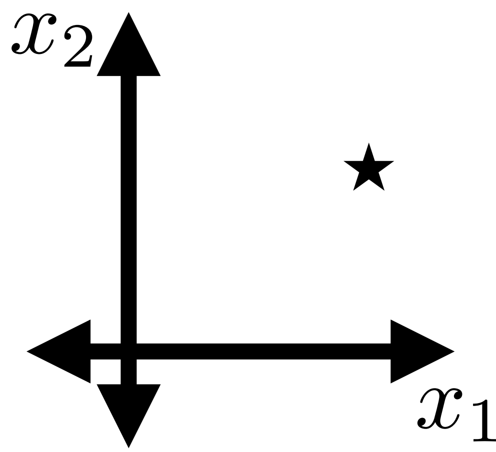


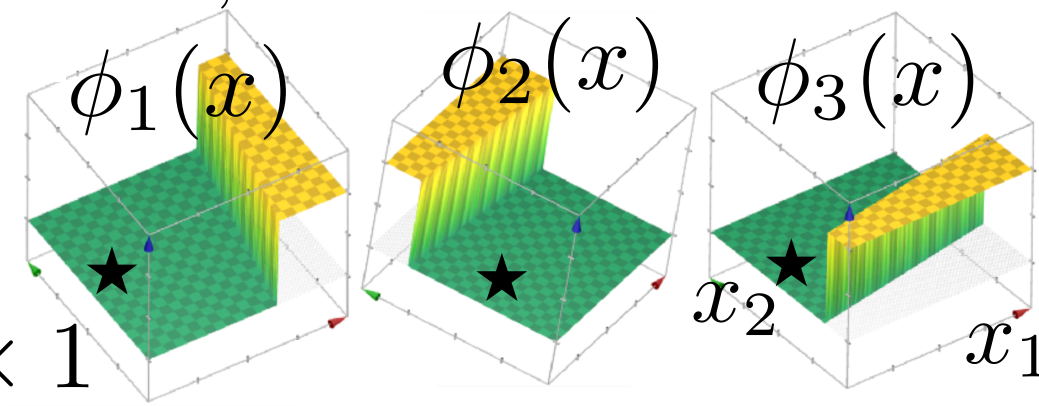
- 2nd layer, assigning a label (or labels):
  - Input (the features): size $m^{(2)} \times 1$ $\boxed{m^{(2)} = n^{(1)}}$
  - Output $A^{(2)}$(vector of labels)



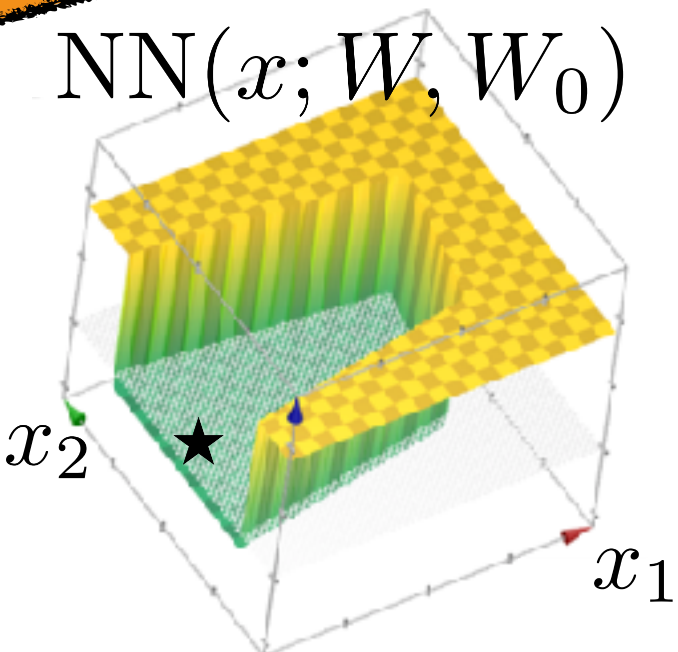
Let's get some new notation

- 1st layer, constructing the features:
  - Input x (a data point): size $m^{(1)} \times 1$ $\boxed{m^{(1)} = d}$
  - Output $A^{(1)}$ (vector of features): size $n^{(1)} \times 1$
  - The *i*th feature: $A_i^{(1)} = f^{(1)}(w_i^{(1)\top} x + w_{0,i}^{(1)})$
    - All the features at once:
      - $A^{(1)} = f^{(1)}(W^{(1)\top} x + W_0^{(1)})$
      - $W^{(1)} : m^{(1)} \times n^{(1)}; W_0^{(1)} : n^{(1)} \times 1$



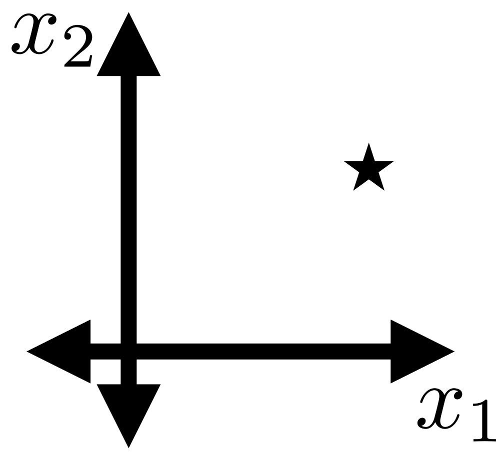


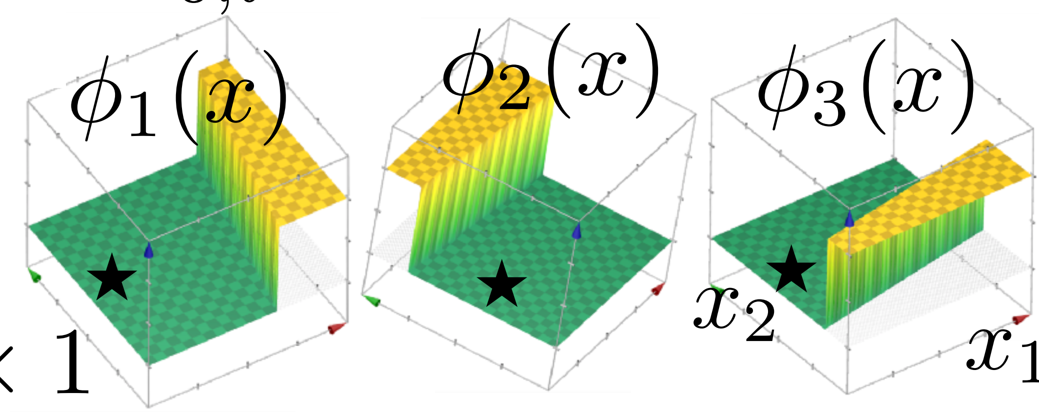
- 2nd layer, assigning a label (or labels):
  - Input (the features): size $m^{(2)} \times 1$ $\boxed{m^{(2)} = n^{(1)}}$
  - Output $A^{(2)}$(vector of labels): size $n^{(2)} \times 1$



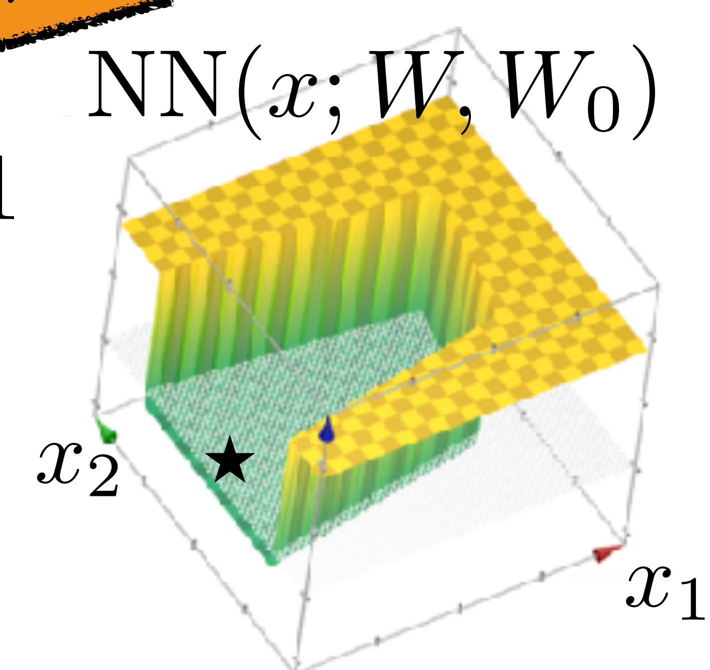
Let's get some new notation

- 1st layer, constructing the features:
  - Input x (a data point): size $m^{(1)} \times 1$ $\boxed{m^{(1)} = d}$
  - Output $A^{(1)}$ (vector of features): size $n^{(1)} \times 1$
  - The *i*th feature: $A_i^{(1)} = f^{(1)}(w_i^{(1)\top} x + w_{0,i}^{(1)})$
    - All the features at once:
      - $A^{(1)} = f^{(1)}(W^{(1)\top} x + W_0^{(1)})$
      - $W^{(1)} : m^{(1)} \times n^{(1)}; W_0^{(1)} : n^{(1)} \times 1$



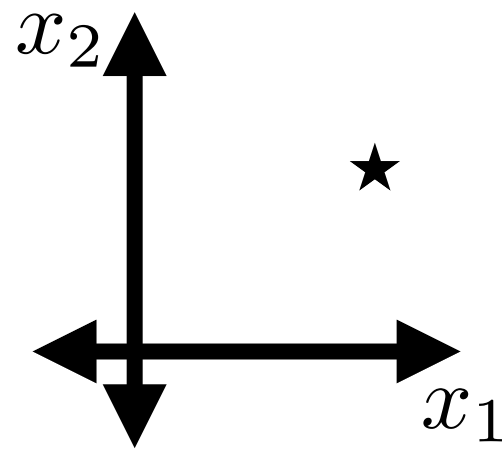


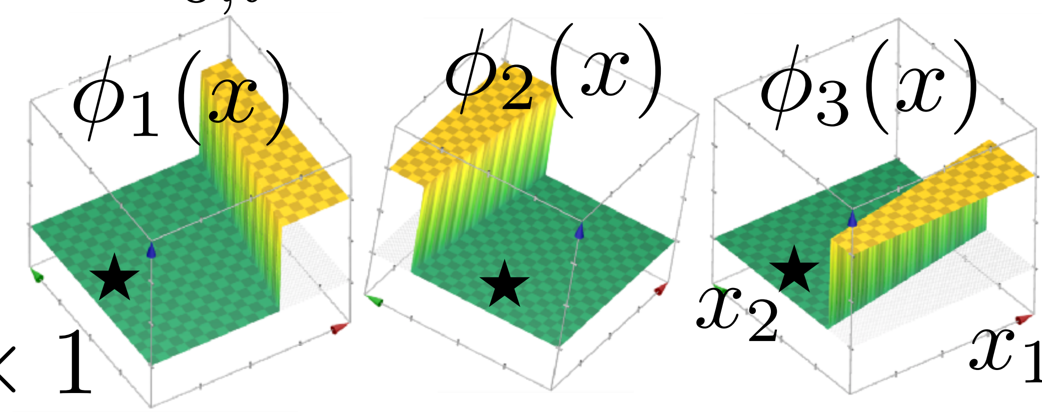
- 2nd layer, assigning a label (or labels):
  - Input (the features): size $m^{(2)} \times 1$ $\boxed{m^{(2)} = n^{(1)}}$
  - Output $A^{(2)}$(vector of labels): size $n^{(2)} \times 1$
  - The *i*th label:



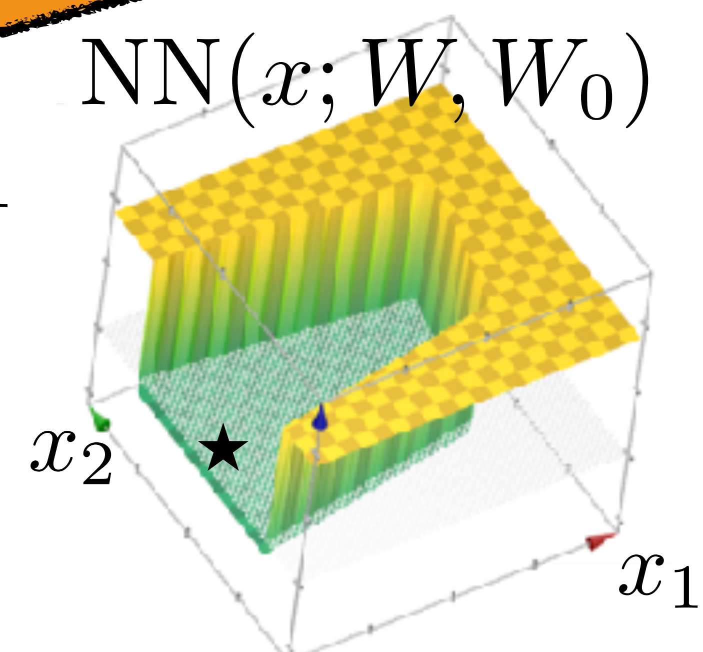
Let's get some new notation

- 1st layer, constructing the features:
  - Input x (a data point): size $m^{(1)} \times 1$ $\boxed{m^{(1)} = d}$
  - Output $A^{(1)}$ (vector of features): size $n^{(1)} \times 1$
  - The *i*th feature: $A_i^{(1)} = f^{(1)}(w_i^{(1)\top} x + w_{0,i}^{(1)})$
    - All the features at once:
      - $A^{(1)} = f^{(1)}(W^{(1)\top} x + W_0^{(1)})$
      - $W^{(1)} : m^{(1)} \times n^{(1)}; W_0^{(1)} : n^{(1)} \times 1$



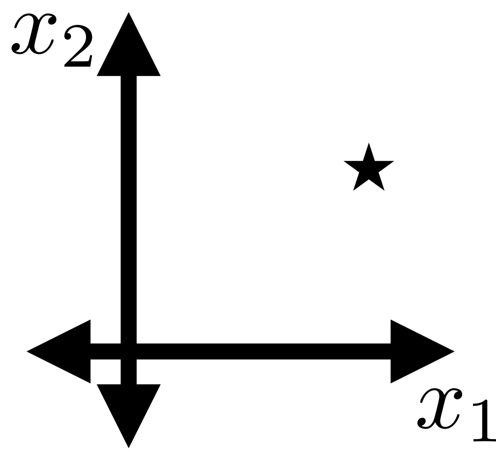


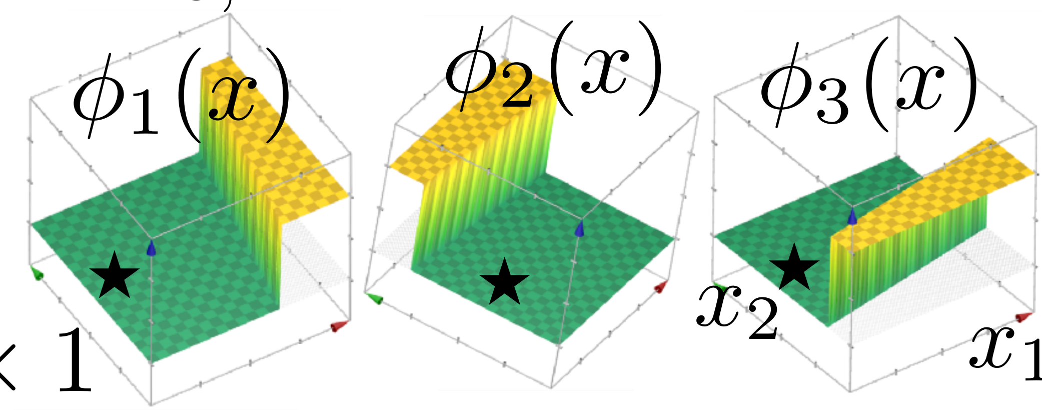
- 2nd layer, assigning a label (or labels):
  - Input (the features): size $m^{(2)} \times 1$ $\boxed{m^{(2)} = n^{(1)}}$
  - Output $A^{(2)}$(vector of labels): size $n^{(2)} \times 1$
  - The *i*th label: $A_i^{(2)} = f^{(2)}(w_i^{(2)\top} A^{(1)} + w_{0,i}^{(2)})$



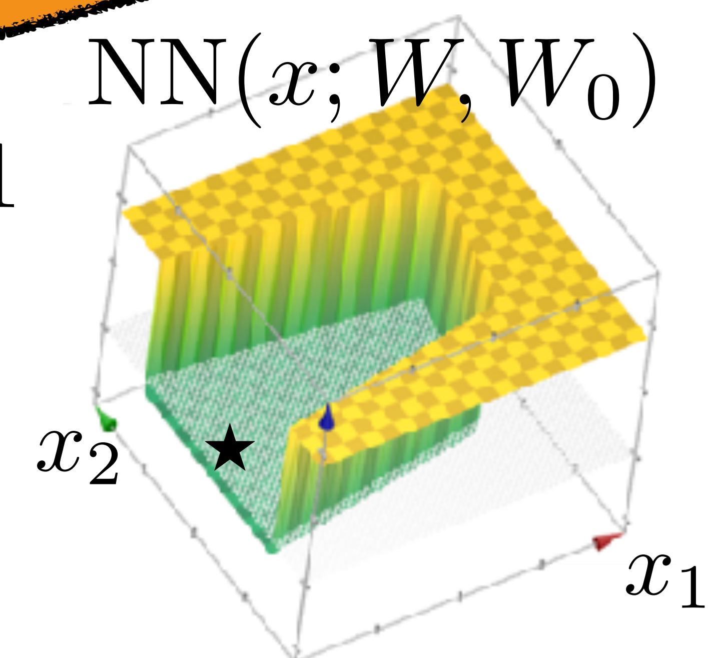
Let's get some new notation

- 1st layer, constructing the features:
  - Input x (a data point): size $m^{(1)} \times 1$ $m^{(1)} = d$
  - Output $A^{(1)}$ (vector of features): size $n^{(1)} \times 1$
  - The ith feature: $A_i^{(1)} = f^{(1)}(w_i^{(1)\top} x + w_{0,i}^{(1)})$
    - All the features at once:
      - $A^{(1)} = f^{(1)}(W^{(1)\top} x + W_0^{(1)})$
      - $W^{(1)} : m^{(1)} \times n^{(1)}; W_0^{(1)} : n^{(1)} \times 1$





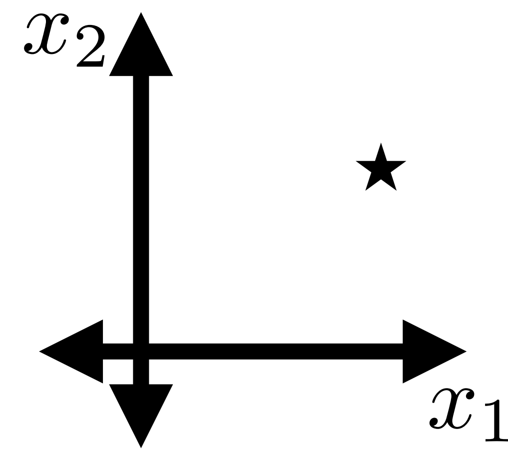
- 2nd layer, assigning a label (or labels):
  - Input (the features): size $m^{(2)} \times 1$ $m^{(2)} = n^{(1)}$
  - Output $A^{(2)}$(vector of labels): size $n^{(2)} \times 1$
  - The ith label: $A_i^{(2)} = f^{(2)}(w_i^{(2)\top} A^{(1)} + w_{0,i}^{(2)})$
    - All:

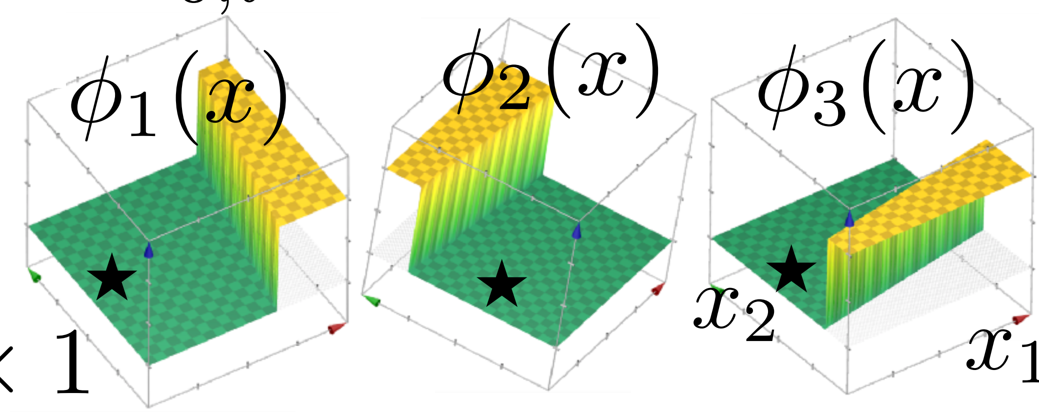


Let's get some new notation

- 1st layer, constructing the features:
  - Input x (a data point): size $m^{(1)} \times 1$
  - Output $A^{(1)}$ (vector of features): size $n^{(1)} \times 1$
  - The ith feature: $A_i^{(1)} = f^{(1)}(w_i^{(1)\top} x + w_{0,i}^{(1)})$
    - All the features at once:
      - $A^{(1)} = f^{(1)}(W^{(1)\top} x + W_0^{(1)})$
      - $W^{(1)} : m^{(1)} \times n^{(1)}; W_0^{(1)} : n^{(1)} \times 1$

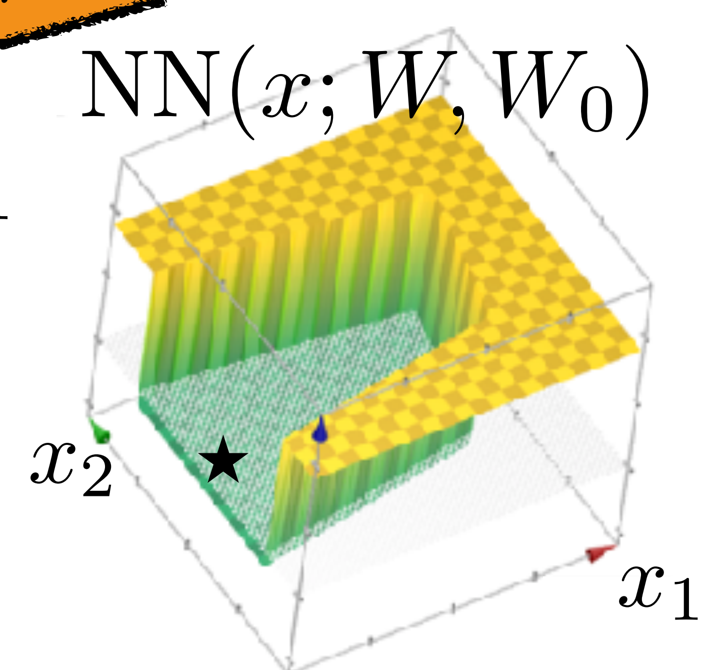
$\boxed{m^{(1)} = d}$





- 2nd layer, assigning a label (or labels):
  - Input (the features): size $m^{(2)} \times 1$
  - Output $A^{(2)}$(vector of labels): size $n^{(2)} \times 1$
  - The ith label: $A_i^{(2)} = f^{(2)}(w_i^{(2)\top} A^{(1)} + w_{0,i}^{(2)})$
    - All: $A^{(2)} = f^{(2)}(W^{(2)\top} A^{(1)} + W_0^{(2)})$

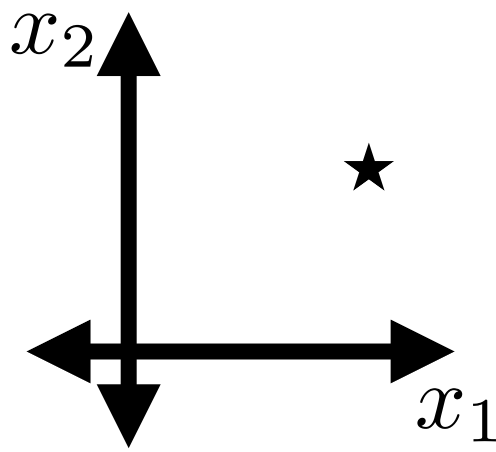
$\boxed{m^{(2)} = n^{(1)}}$

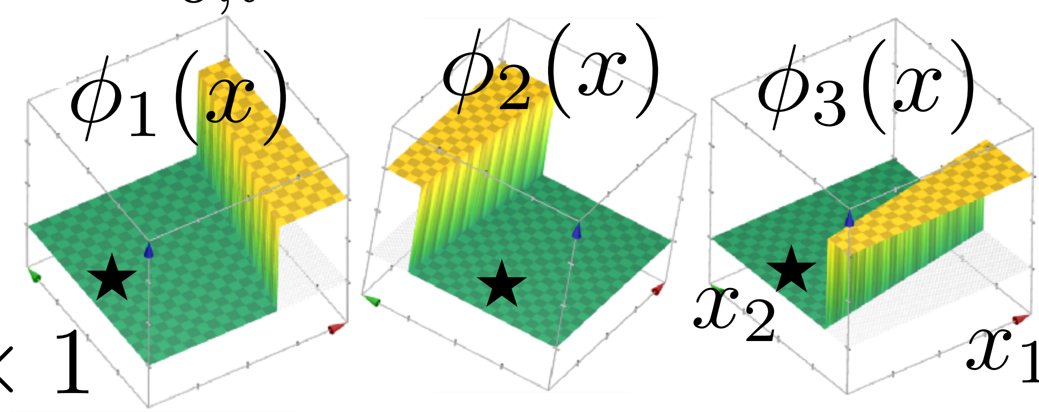


Let's get some new notation

- 1st layer, constructing the features:
  - Input x (a data point): size $m^{(1)} \times 1$
  - Output $A^{(1)}$ (vector of features): size $n^{(1)} \times 1$
  - The *i*th feature: $A_i^{(1)} = f^{(1)}(w_i^{(1)\top} x + w_{0,i}^{(1)})$
    - All the features at once:
      - $A^{(1)} = f^{(1)}(W^{(1)\top} x + W_0^{(1)})$
      - $W^{(1)} : m^{(1)} \times n^{(1)}; W_0^{(1)} : n^{(1)} \times 1$

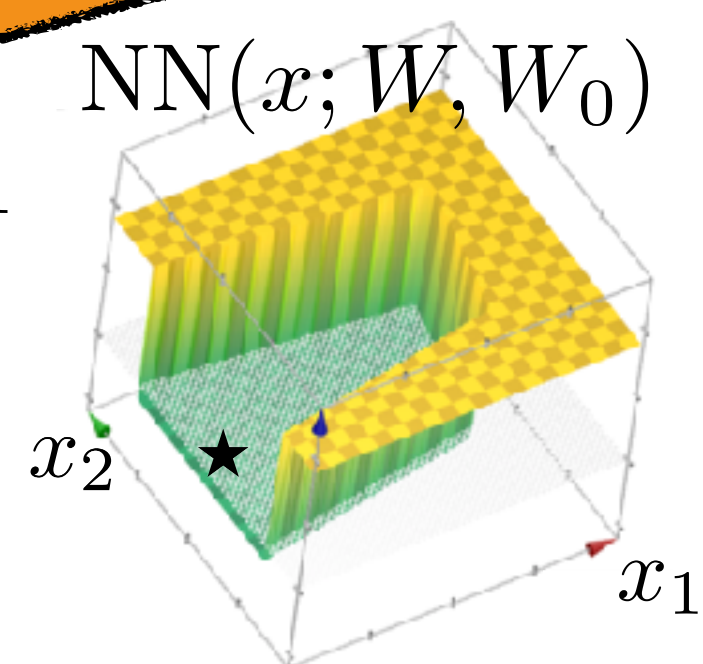
$m^{(1)} = d$





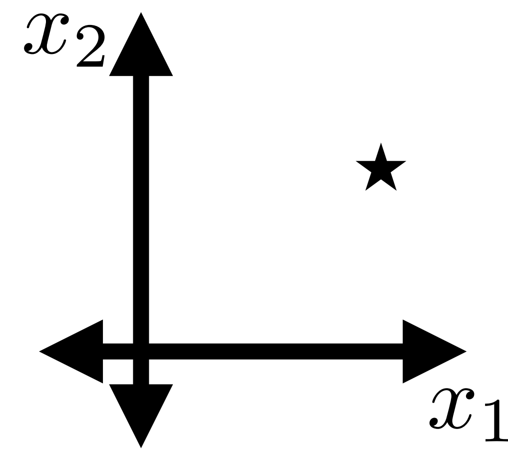
- 2nd layer, assigning a label (or labels):
  - Input (the features): size $m^{(2)} \times 1$
  - Output $A^{(2)}$(vector of labels): size $n^{(2)} \times 1$
  - The *i*th label: $A_i^{(2)} = f^{(2)}(w_i^{(2)\top} A^{(1)} + w_{0,i}^{(2)})$
    - All: $A^{(2)} = f^{(2)}(W^{(2)\top} A^{(1)} + W_0^{(2)})$
    - $W^{(2)} : m^{(2)} \times n^{(2)}$

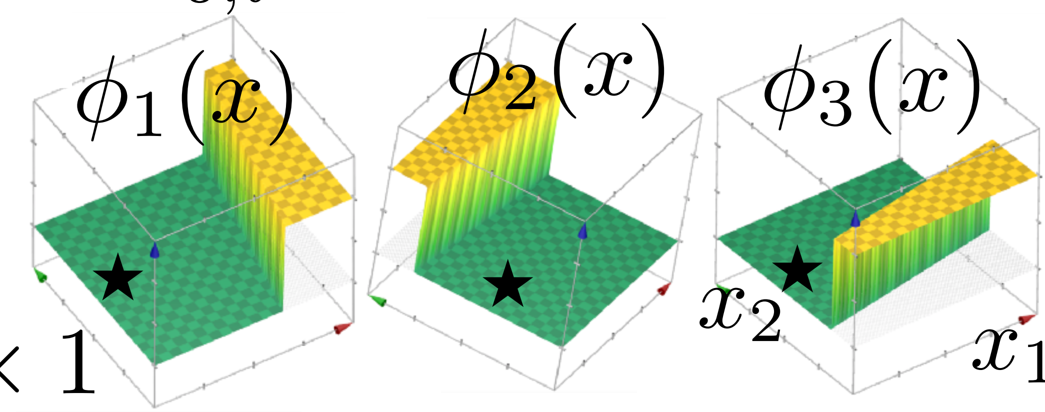
$m^{(2)} = n^{(1)}$



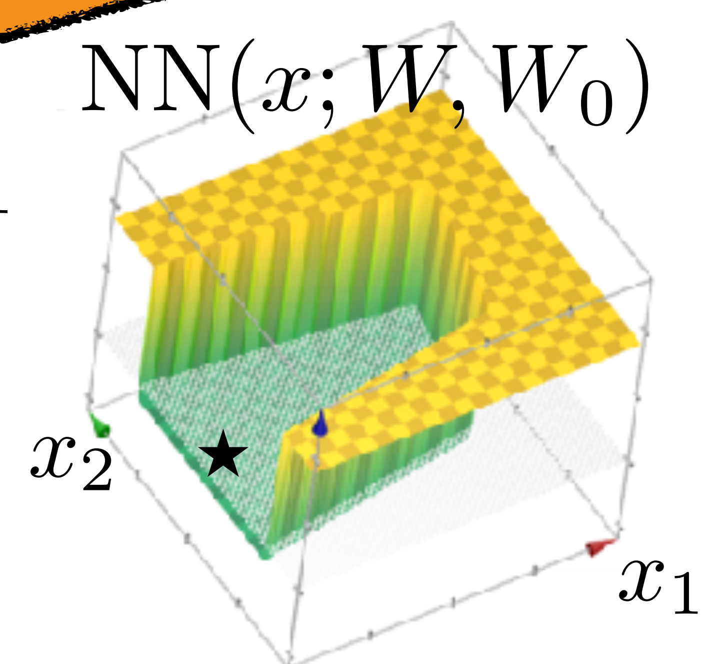
Let's get some new notation

- 1st layer, constructing the features:
  - Input x (a data point): size $m^{(1)} \times 1$ $\boxed{m^{(1)} = d}$
  - Output $A^{(1)}$ (vector of features): size $n^{(1)} \times 1$
  - The ith feature: $A_i^{(1)} = f^{(1)}(w_i^{(1)\top} x + w_{0,i}^{(1)})$
    - All the features at once:
      - $A^{(1)} = f^{(1)}(W^{(1)\top} x + W_0^{(1)})$
      - $W^{(1)} : m^{(1)} \times n^{(1)}; W_0^{(1)} : n^{(1)} \times 1$



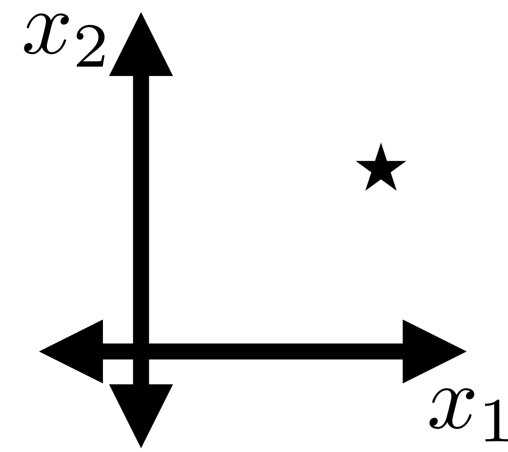


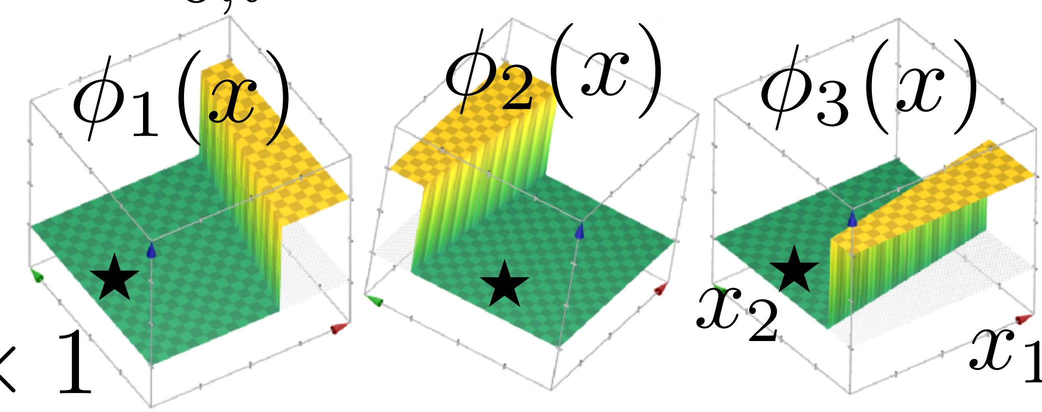
- 2nd layer, assigning a label (or labels):
  - Input (the features): size $m^{(2)} \times 1$ $\boxed{m^{(2)} = n^{(1)}}$
  - Output $A^{(2)}$(vector of labels): size $n^{(2)} \times 1$
  - The ith label: $A_i^{(2)} = f^{(2)}(w_i^{(2)\top} A^{(1)} + w_{0,i}^{(2)})$
    - All: $A^{(2)} = f^{(2)}(W^{(2)\top} A^{(1)} + W_0^{(2)})$
    - $W^{(2)} : m^{(2)} \times n^{(2)}; W_0^{(2)} : n^{(2)} \times 1$



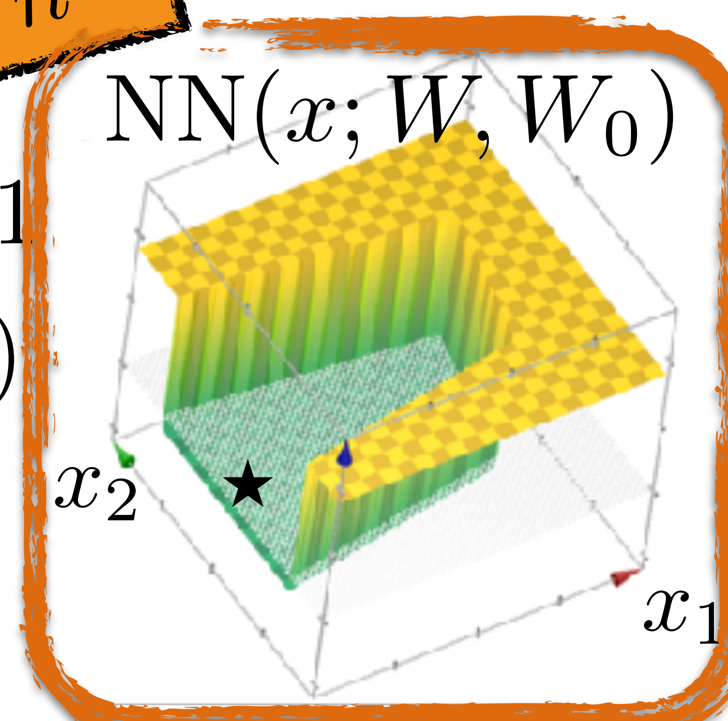
Let's get some new notation

- 1st layer, constructing the features: $m^{(1)} = d$
  - Input x (a data point): size $m^{(1)} \times 1$
  - Output $A^{(1)}$ (vector of features): size $n^{(1)} \times 1$
  - The *i*th feature: $A_i^{(1)} = f^{(1)}(w_i^{(1)\top} x + w_{0,i}^{(1)})$
    - All the features at once:
      - $A^{(1)} = f^{(1)}(W^{(1)\top} x + W_0^{(1)})$
      - $W^{(1)} : m^{(1)} \times n^{(1)}; W_0^{(1)} : n^{(1)} \times 1$





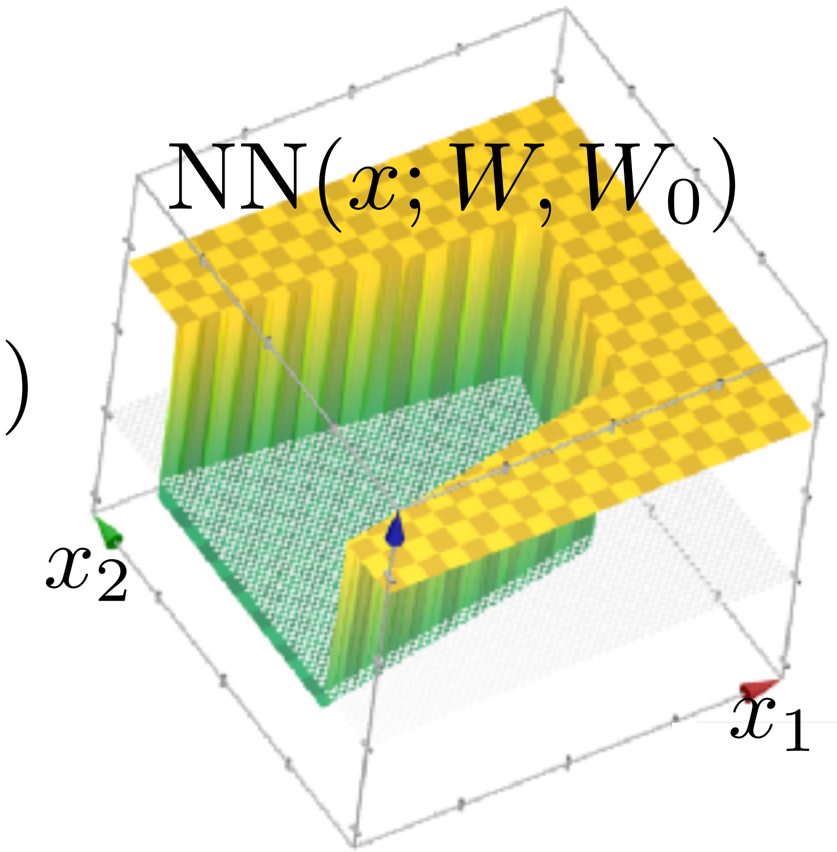
- 2nd layer, assigning a label (or labels): $m^{(2)} = n^{(1)}$
  - Input (the features): size $m^{(2)} \times 1$
  - Output $A^{(2)}$(vector of labels): size $n^{(2)} \times 1$
  - The *i*th label: $A_i^{(2)} = f^{(2)}(w_i^{(2)\top} A^{(1)} + w_{0,i}^{(2)})$
    - All: $A^{(2)} = f^{(2)}(W^{(2)\top} A^{(1)} + W_0^{(2)})$
    - $W^{(2)} : m^{(2)} \times n^{(2)}; W_0^{(2)} : n^{(2)} \times 1$



- 1st layer: $A^{(1)} = f^{(1)}(W^{(1)\top} x + W_0^{(1)})$
- 2nd layer: $A^{(2)} = f^{(2)}(W^{(2)\top} A^{(1)} + W_0^{(2)})$

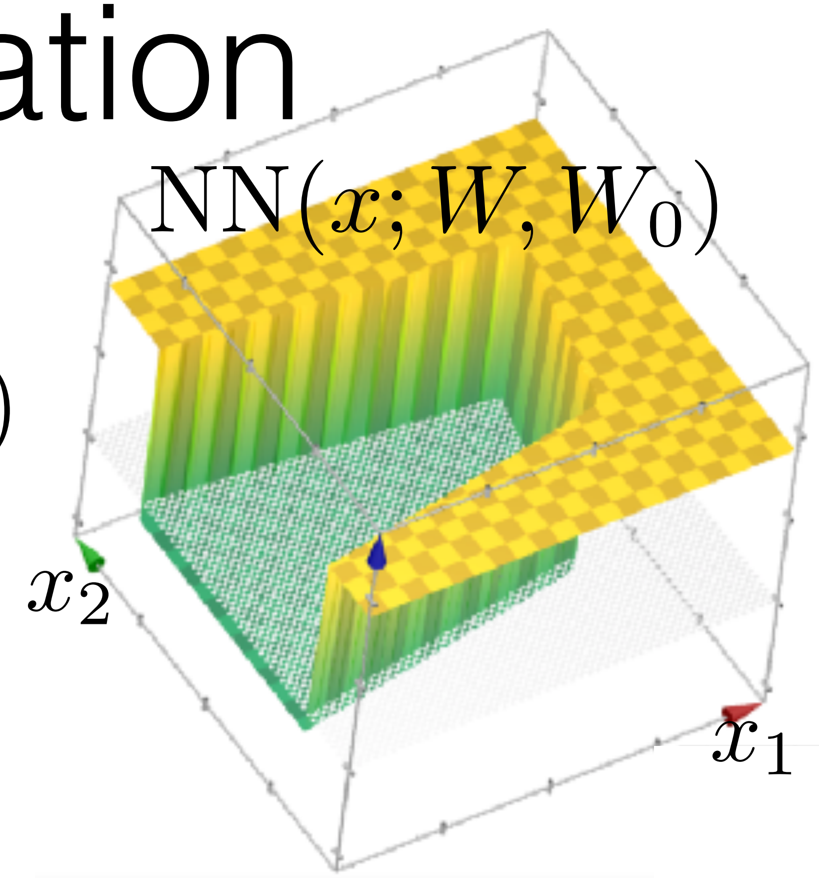
- 1st layer: $A^{(1)} = f^{(1)}(W^{(1)\top} x + W_0^{(1)})$
- 2nd layer: $A^{(2)} = f^{(2)}(W^{(2)\top} A^{(1)} + W_0^{(2)})$
- Whole thing: $A^{(2)} = \mathrm{NN}(x; W, W_0)$

- 1st layer: $A^{(1)} = f^{(1)}(W^{(1)\top} x + W_0^{(1)})$
- 2nd layer: $A^{(2)} = f^{(2)}(W^{(2)\top} A^{(1)} + W_0^{(2)})$
- Whole thing: $A^{(2)} = \mathrm{NN}(x; W, W_0)$



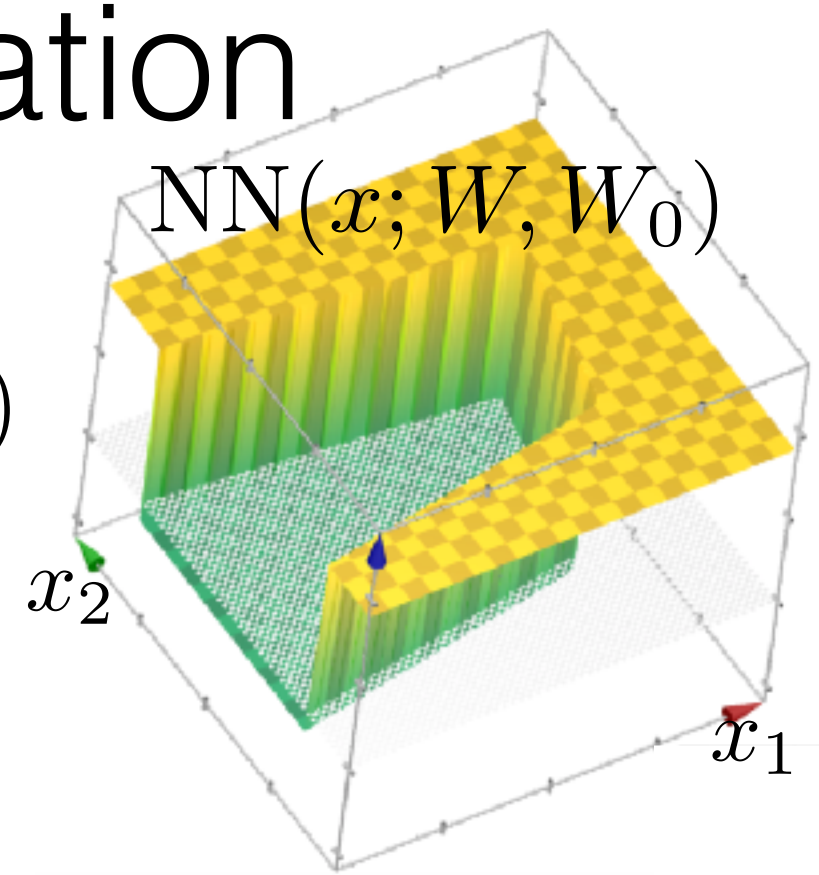
Function graph representation

- 1st layer: $A^{(1)} = f^{(1)}(W^{(1)\top}x + W_0^{(1)})$
- 2nd layer: $A^{(2)} = f^{(2)}(W^{(2)\top}A^{(1)} + W_0^{(2)})$
- Whole thing: $A^{(2)} = \mathrm{NN}(x; W, W_0)$



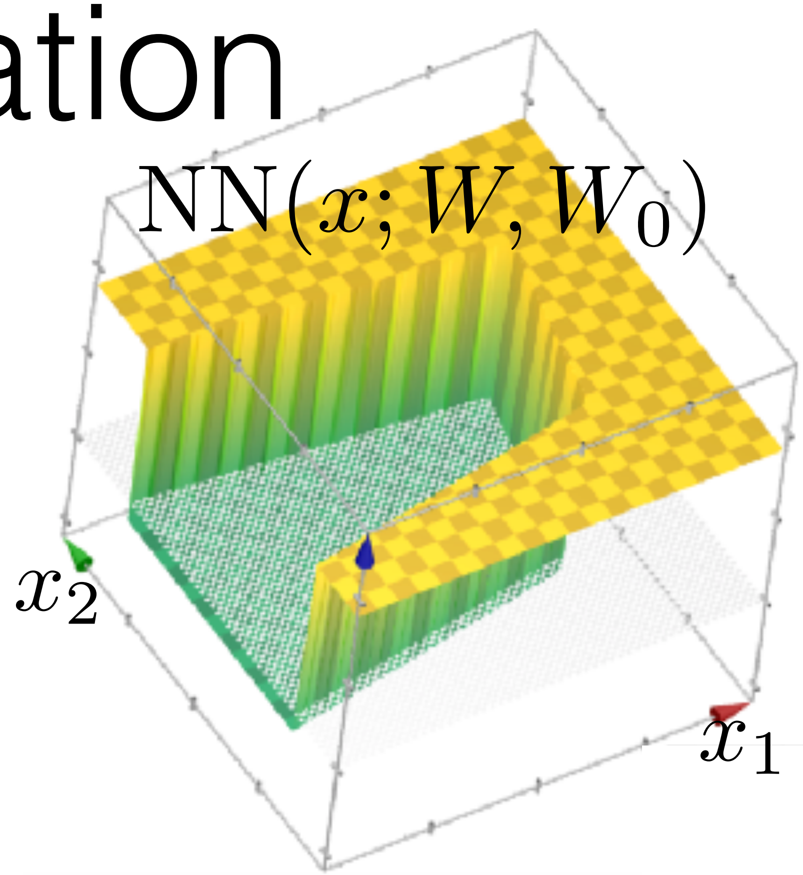
Function graph representation

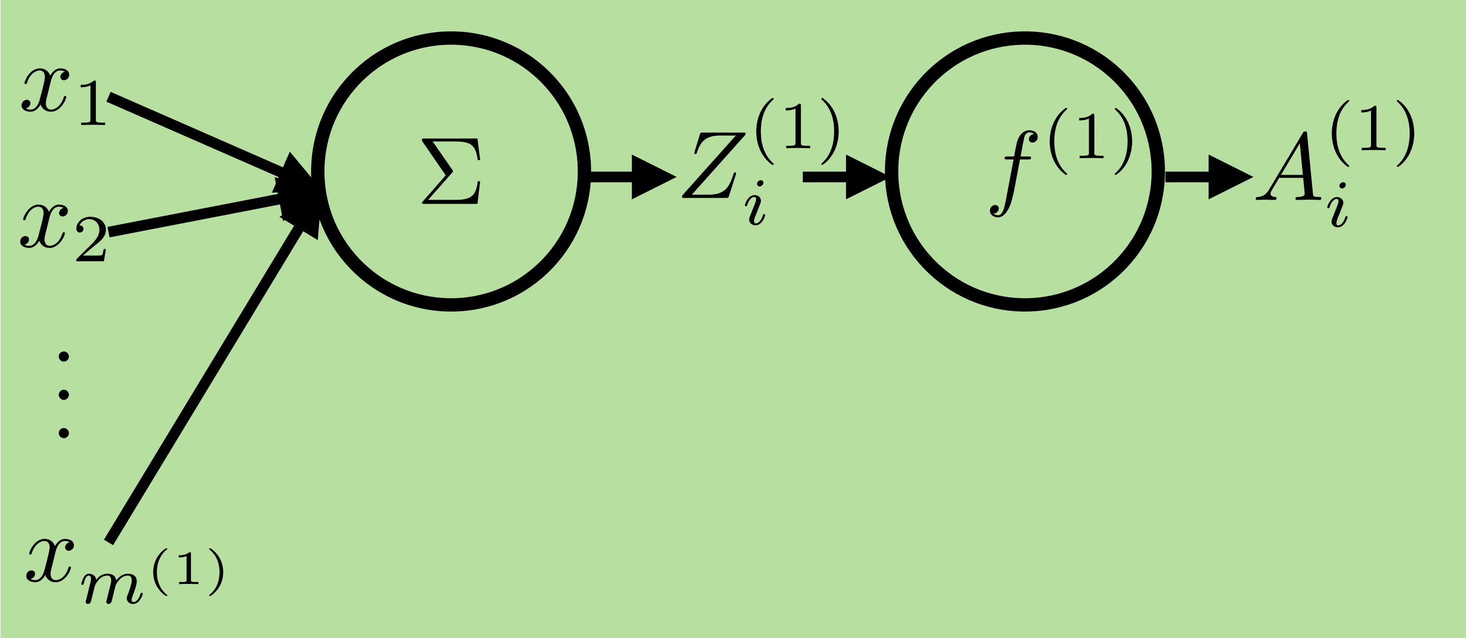
- 1st layer: $A^{(1)} = f^{(1)}(W^{(1)\top}x + W_0^{(1)})$
- 2nd layer: $A^{(2)} = f^{(2)}(W^{(2)\top}A^{(1)} + W_0^{(2)})$
- Whole thing: $A^{(2)} = \mathrm{NN}(x; W, W_0)$



Function graph representation

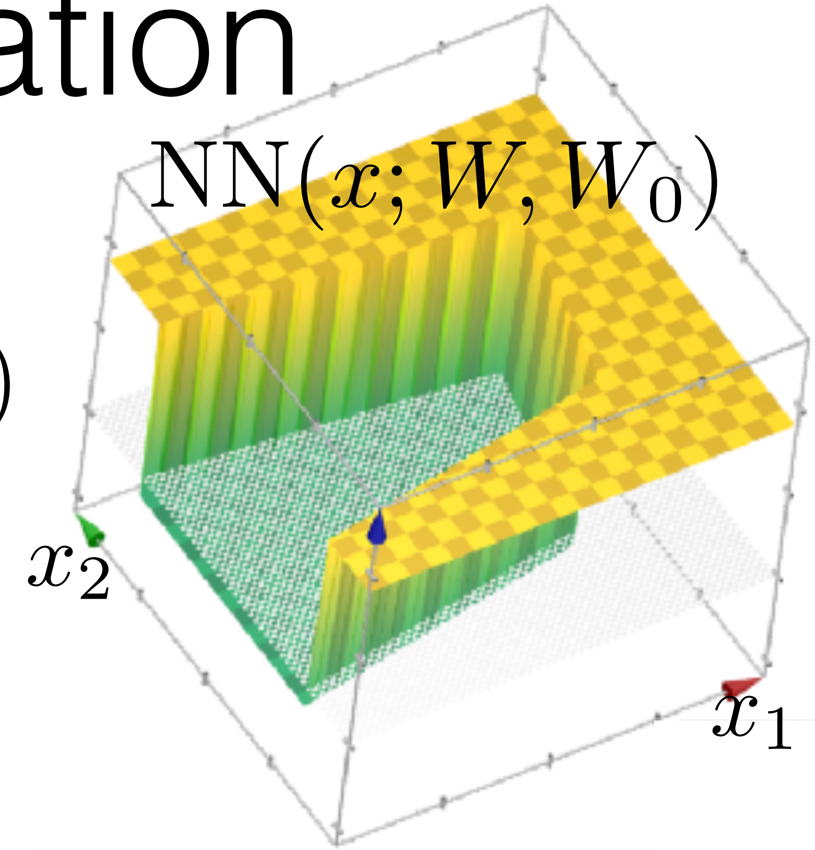
- 1st layer: $\boxed{A^{(1)} = f^{(1)}(W^{(1)\top}x + W_0^{(1)})}$
- 2nd layer: $A^{(2)} = f^{(2)}(W^{(2)\top}A^{(1)} + W_0^{(2)})$
- Whole thing: $A^{(2)} = \mathrm{NN}(x; W, W_0)$

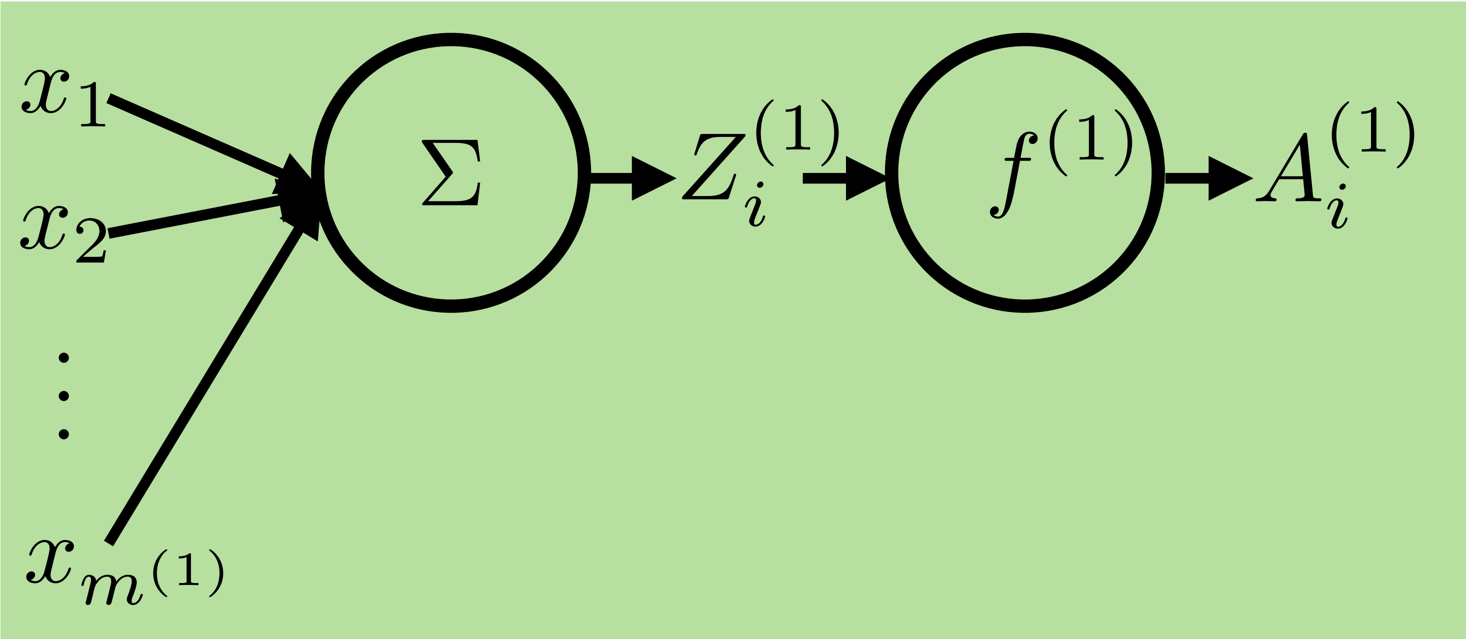




Function graph representation

- 1st layer: $A^{(1)} = f^{(1)}(W^{(1)\top} x + W_0^{(1)})$
- 2nd layer: $A^{(2)} = f^{(2)}(W^{(2)\top} A^{(1)} + W_0^{(2)})$
- Whole thing: $A^{(2)} = \mathrm{NN}(x; W, W_0)$

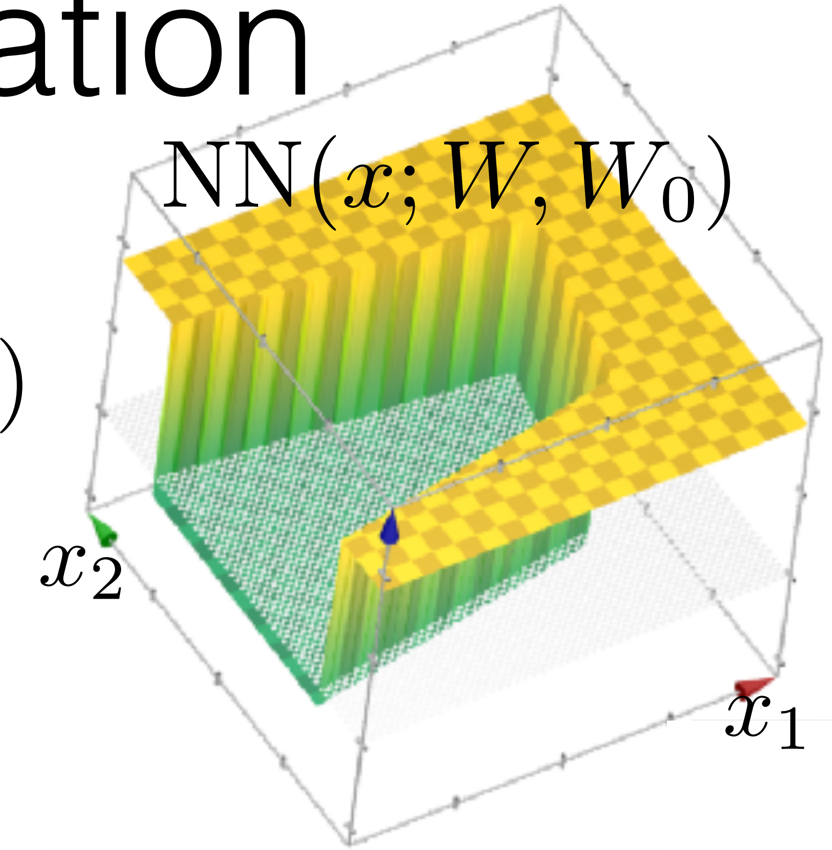


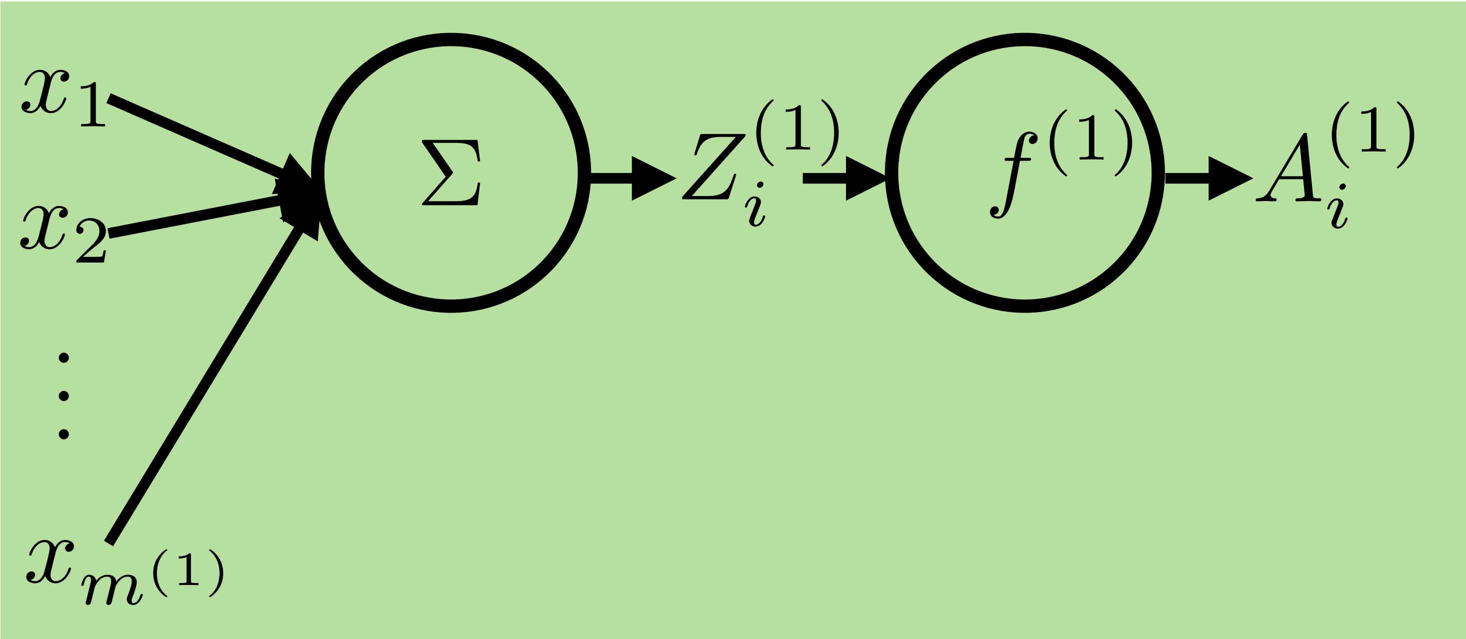


- Circle: function evaluation

Function graph representation

- 1st layer: $A^{(1)} = f^{(1)}(W^{(1)\top}x + W_0^{(1)})$
- 2nd layer: $A^{(2)} = f^{(2)}(W^{(2)\top}A^{(1)} + W_0^{(2)})$
- Whole thing: $A^{(2)} = \mathrm{NN}(x; W, W_0)$

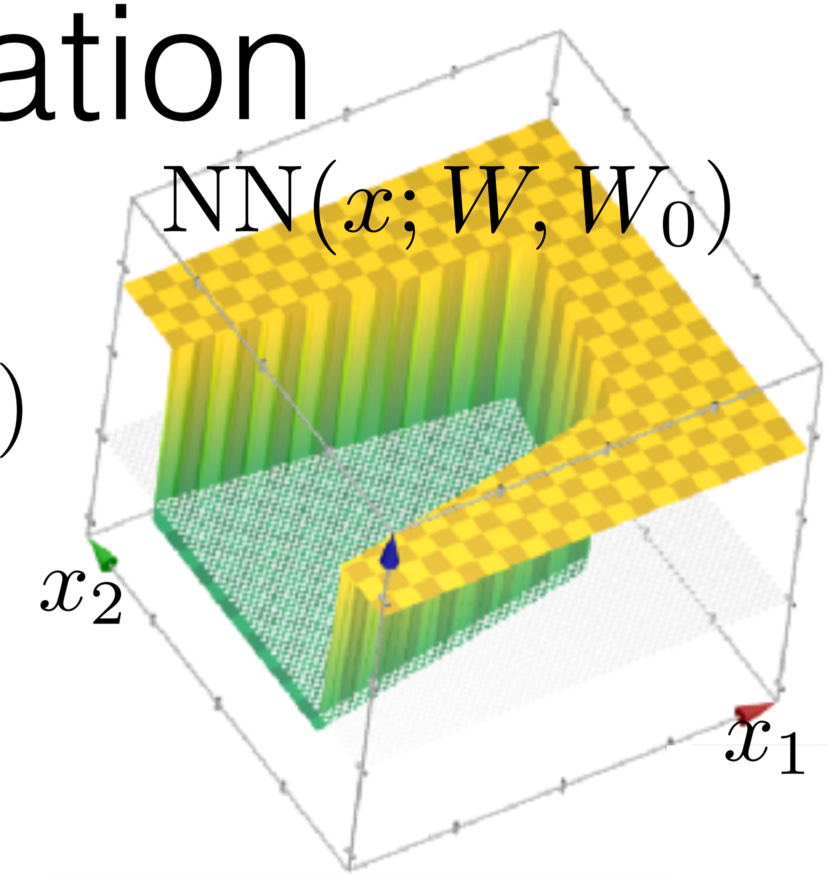


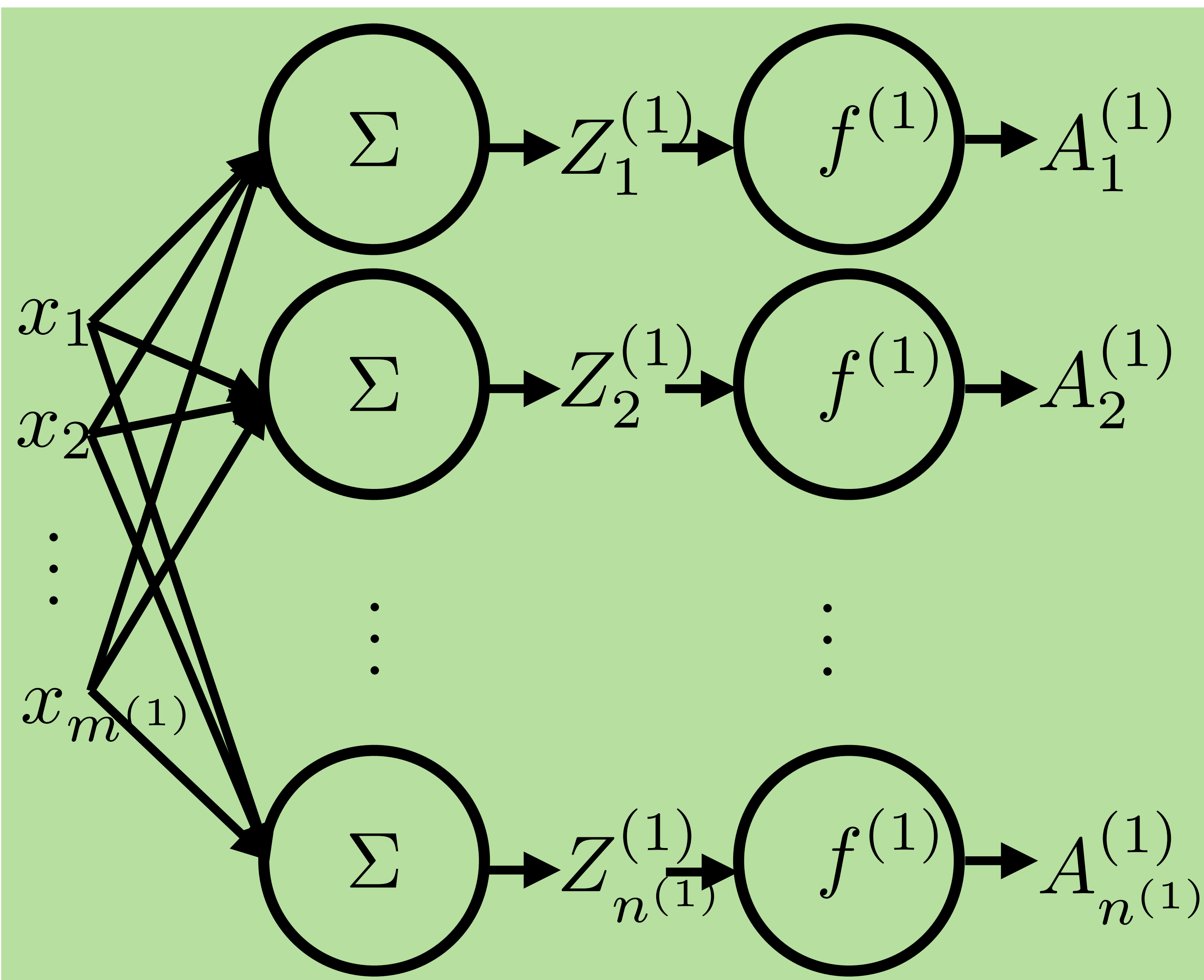


- Circle: function evaluation

Function graph representation

- 1st layer: $A^{(1)} = f^{(1)}(W^{(1)\top}x + W_0^{(1)})$
- 2nd layer: $A^{(2)} = f^{(2)}(W^{(2)\top}A^{(1)} + W_0^{(2)})$
- Whole thing: $A^{(2)} = \mathrm{NN}(x; W, W_0)$

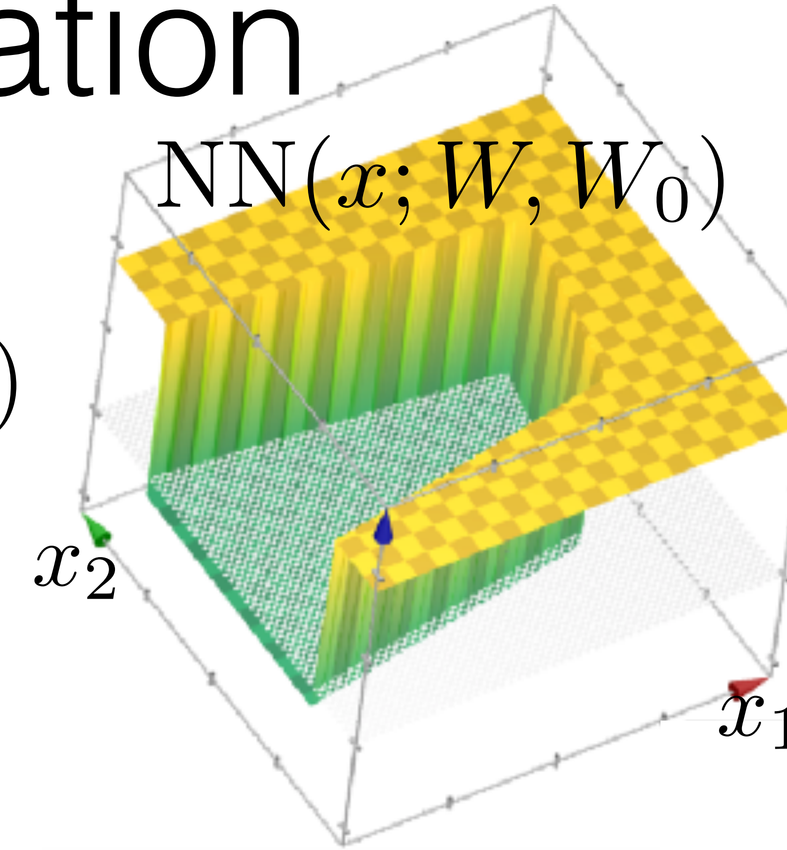


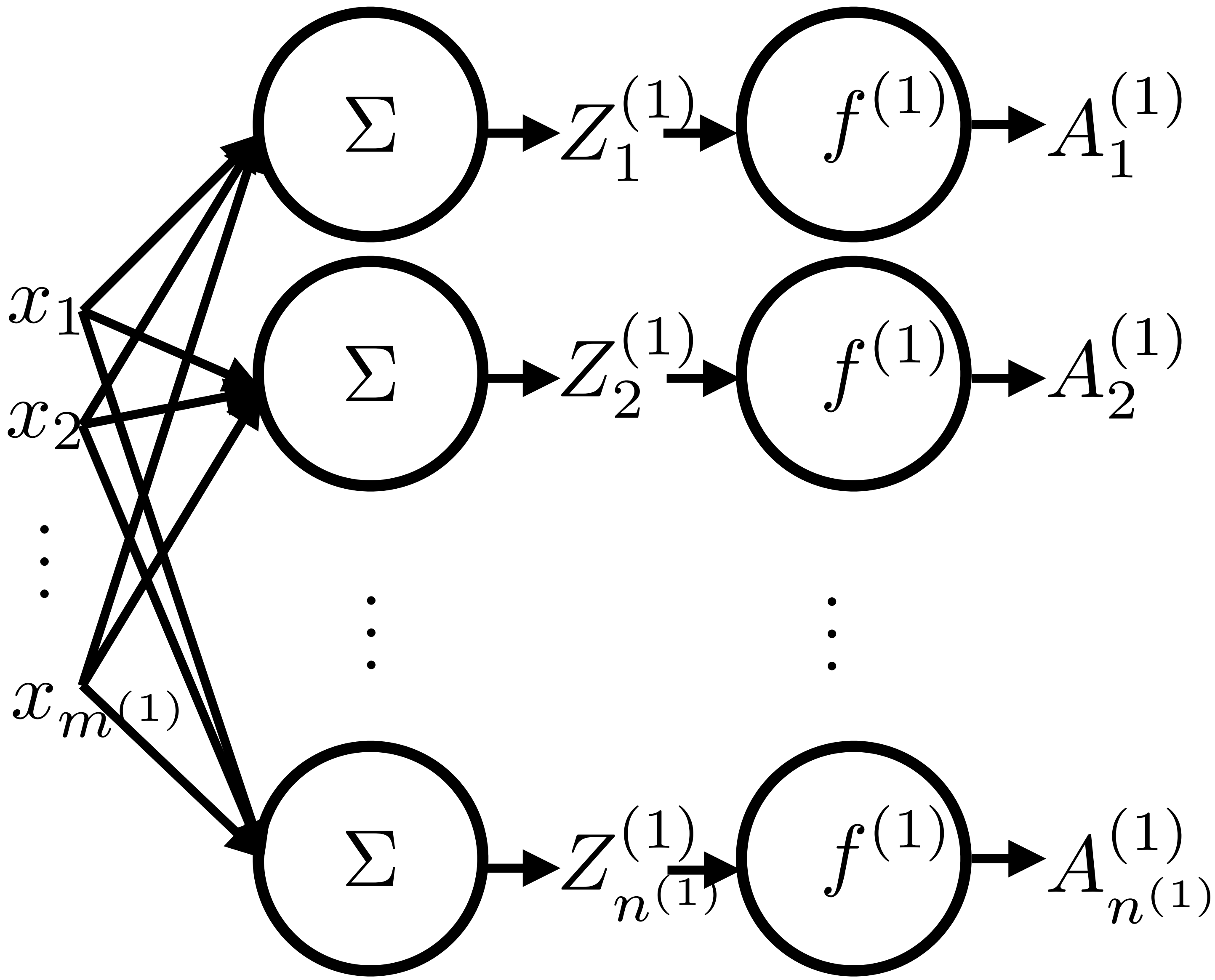


- Circle: function evaluation

Function graph representation

- 1st layer: $A^{(1)} = f^{(1)}(W^{(1)\top} x + W_0^{(1)})$
- 2nd layer: $A^{(2)} = f^{(2)}(W^{(2)\top} A^{(1)} + W_0^{(2)})$
- Whole thing: $A^{(2)} = \mathrm{NN}(x; W, W_0)$

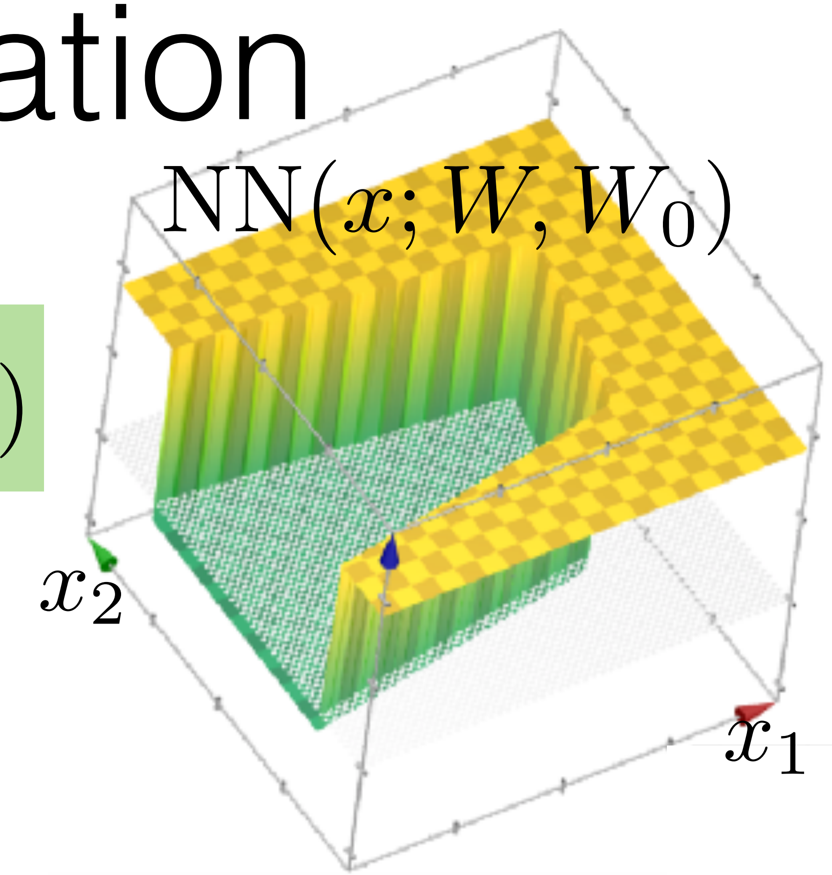


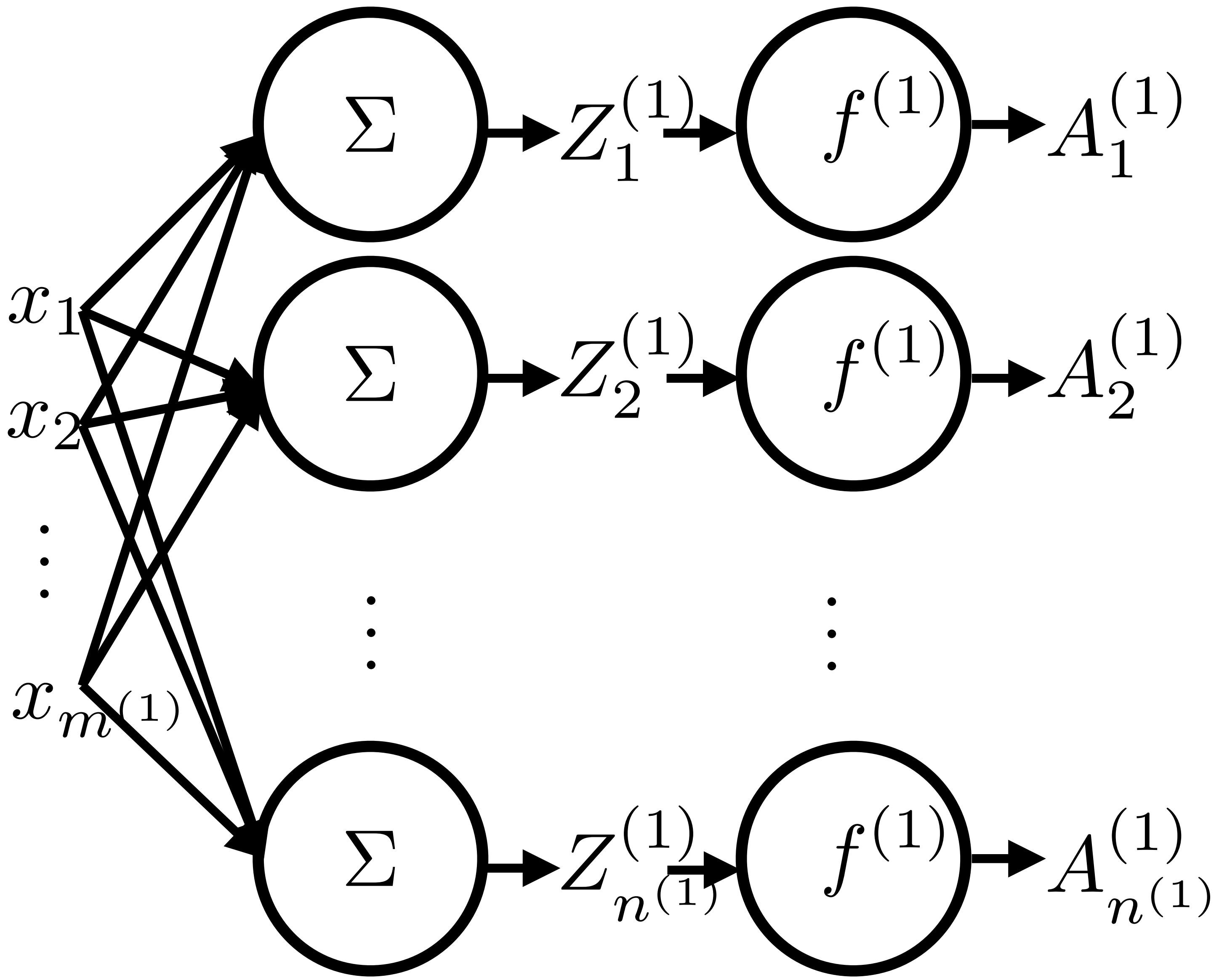


- Circle: function evaluation

Function graph representation

- 1st layer: $A^{(1)} = f^{(1)}(W^{(1)\top} x + W_0^{(1)})$
- 2nd layer: $A^{(2)} = f^{(2)}(W^{(2)\top} A^{(1)} + W_0^{(2)})$
- Whole thing: $A^{(2)} = \mathrm{NN}(x; W, W_0)$

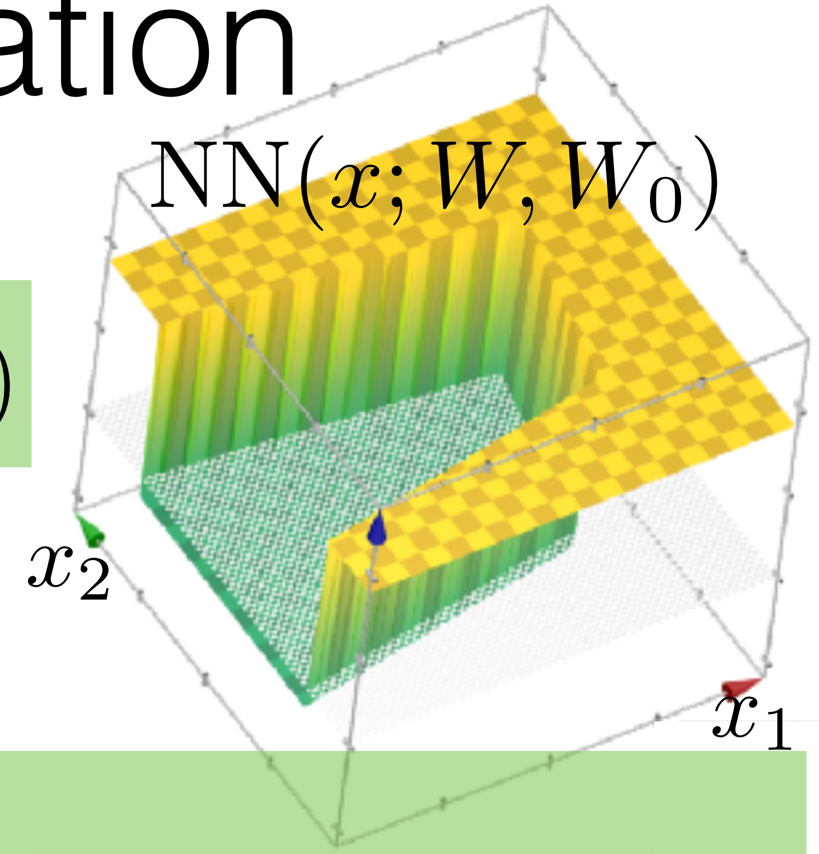


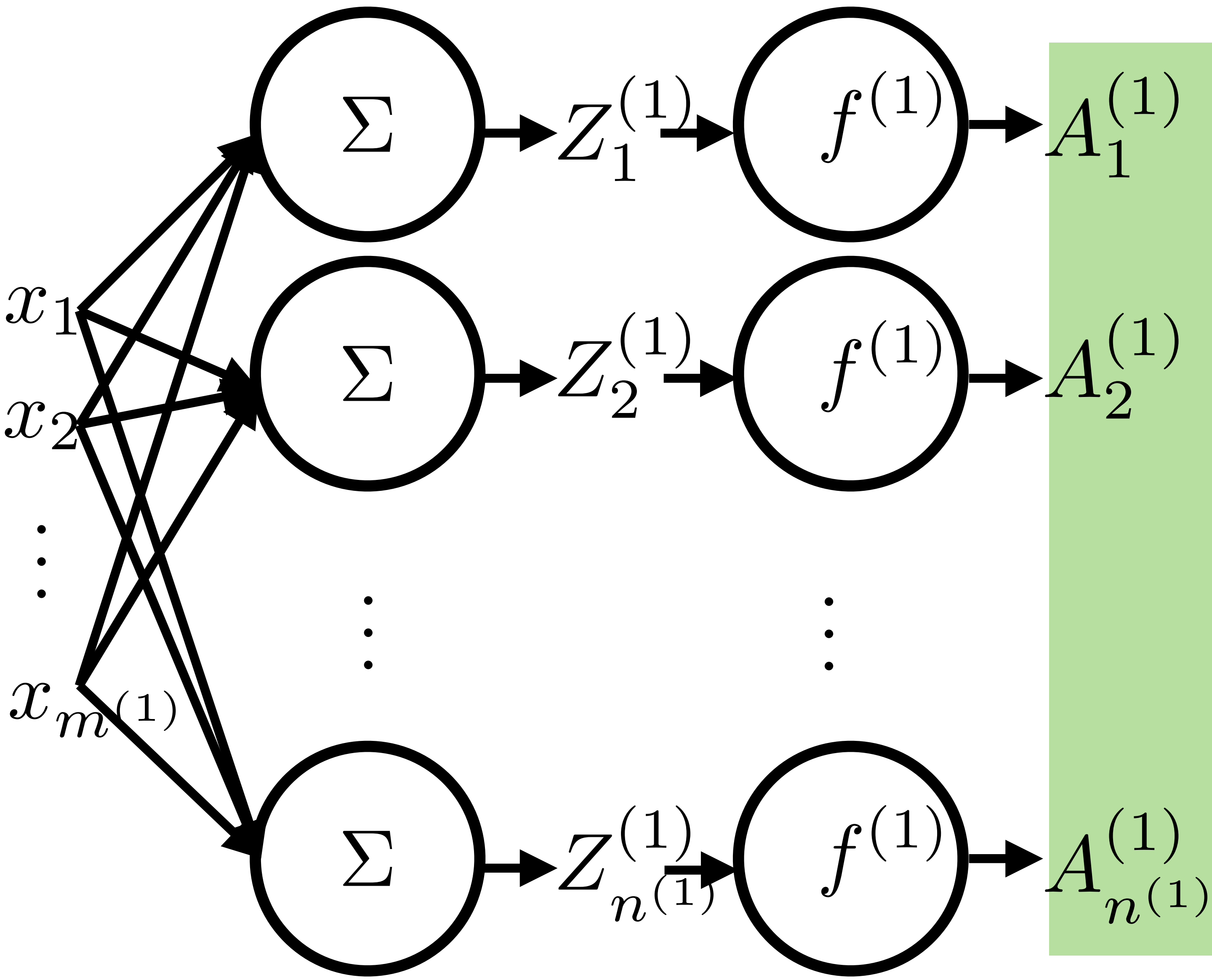


- Circle: function evaluation

Function graph representation

- 1st layer: $A^{(1)} = f^{(1)}(W^{(1)\top}x + W_0^{(1)})$
- 2nd layer: $A^{(2)} = f^{(2)}(W^{(2)\top}A^{(1)} + W_0^{(2)})$
- Whole thing: $A^{(2)} = \mathrm{NN}(x; W, W_0)$

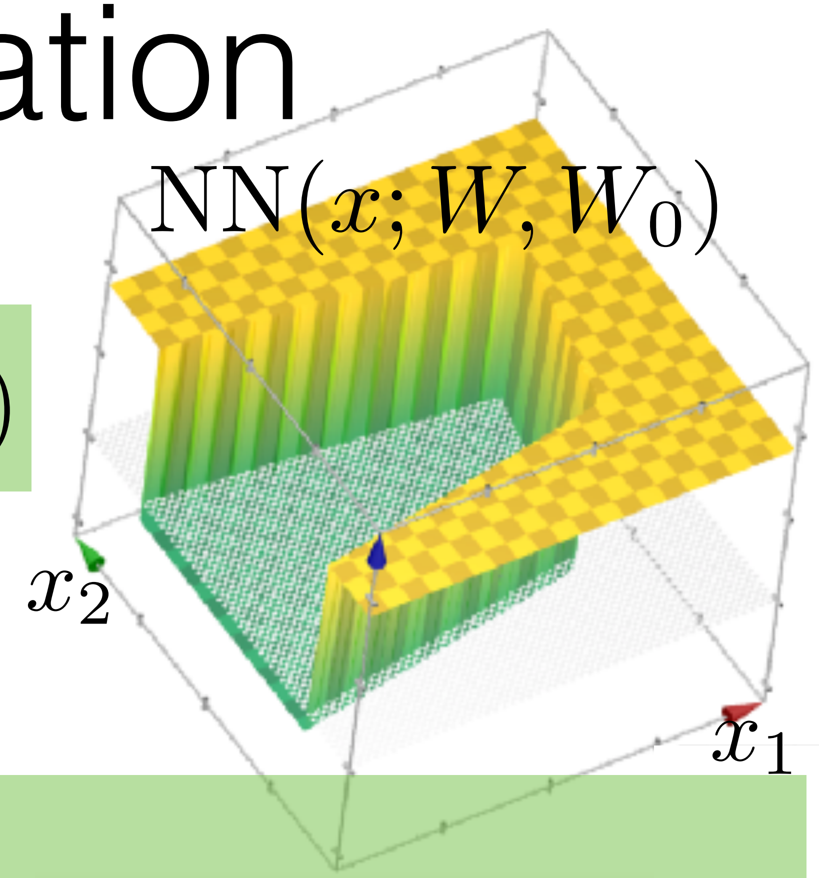


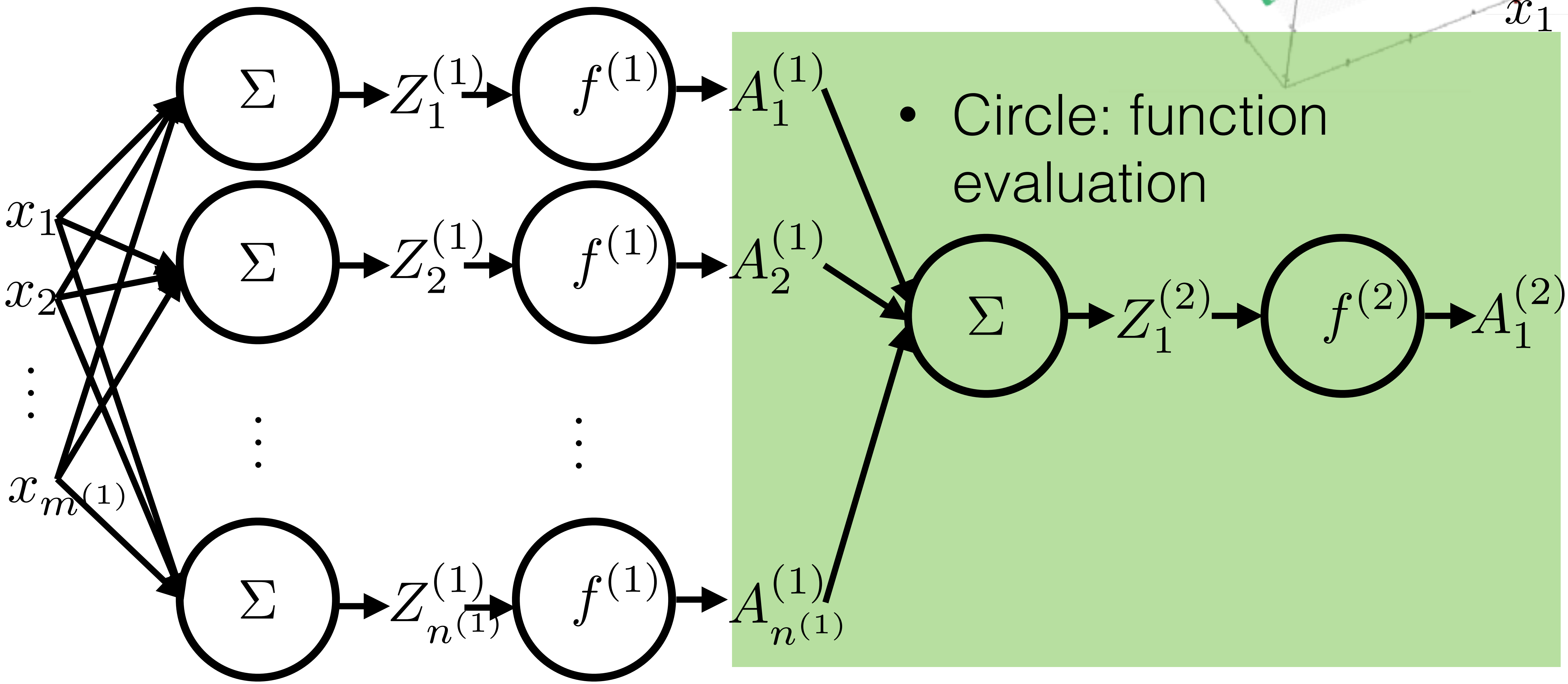


- Circle: function evaluation

Function graph representation

- 1st layer: $A^{(1)} = f^{(1)}(W^{(1)\top} x + W_0^{(1)})$
- 2nd layer: $A^{(2)} = f^{(2)}(W^{(2)\top} A^{(1)} + W_0^{(2)})$
- Whole thing: $A^{(2)} = \mathrm{NN}(x; W, W_0)$

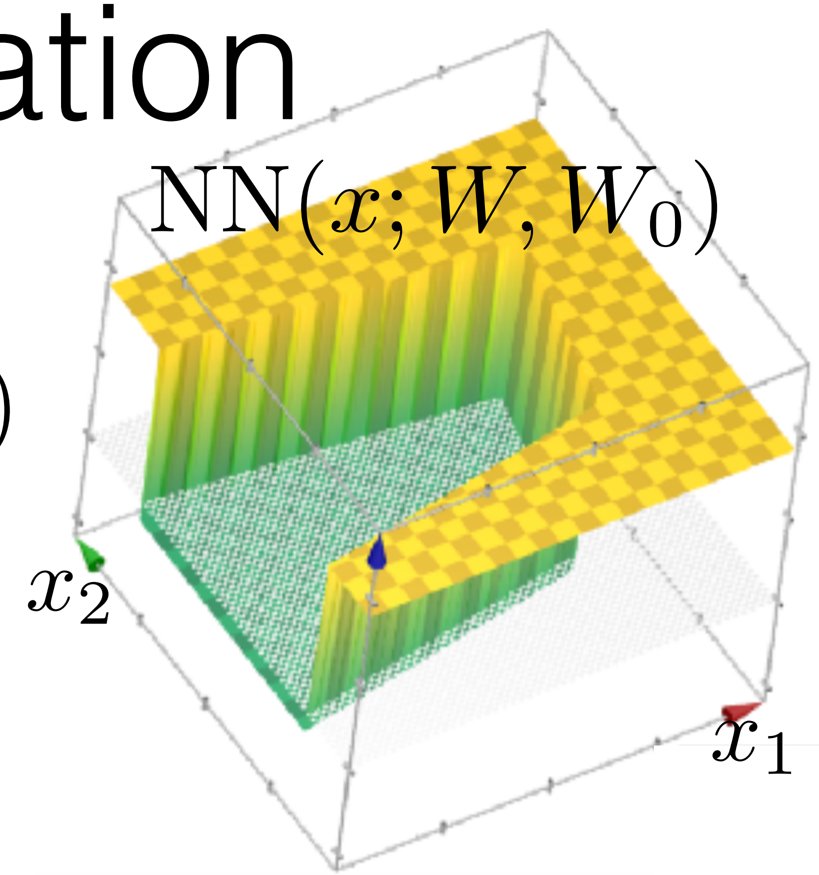


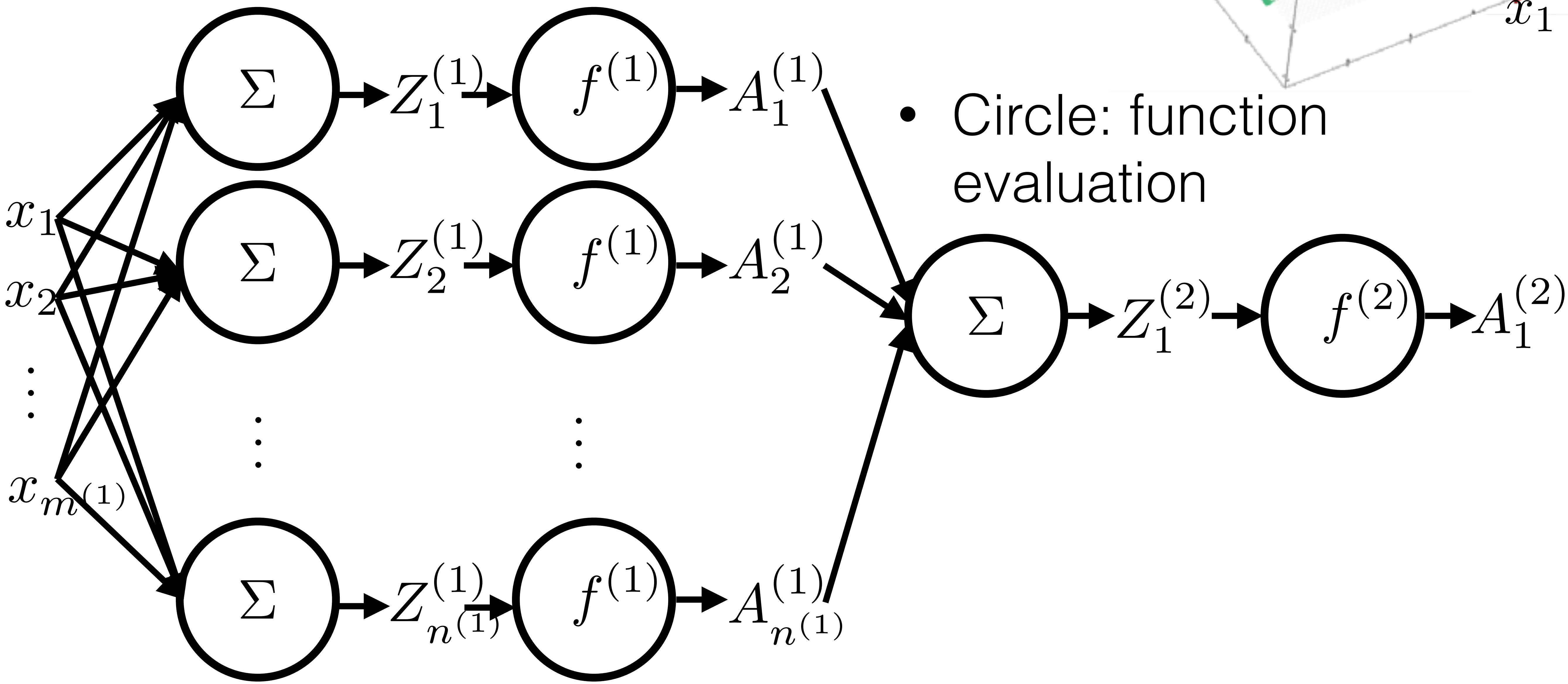


- Circle: function evaluation

Function graph representation

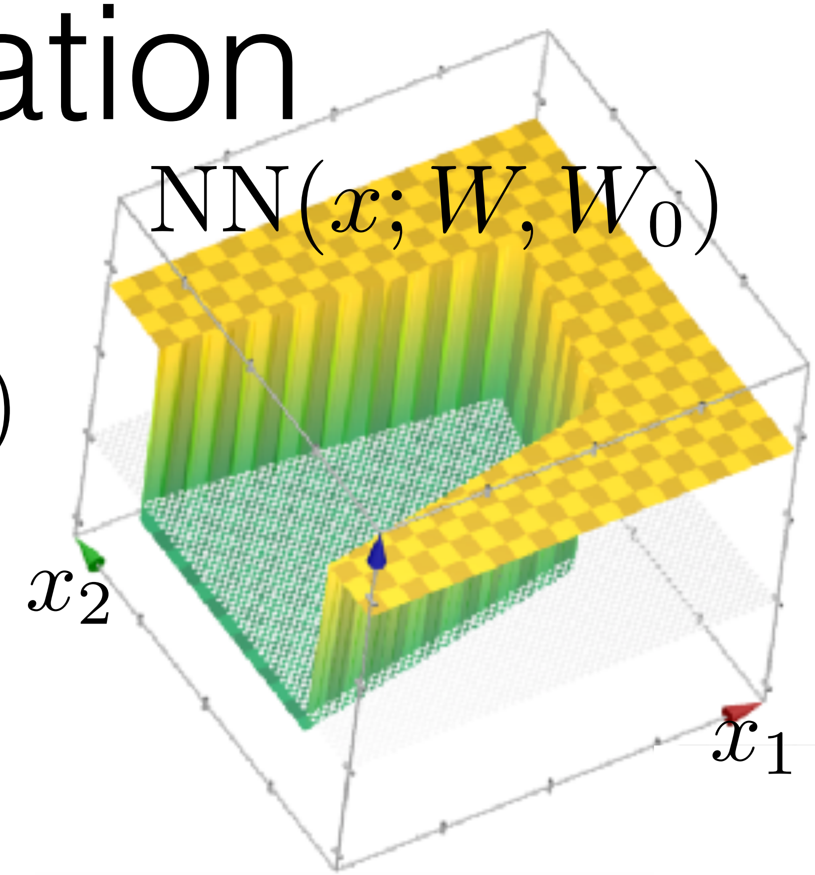
- 1st layer: $A^{(1)} = f^{(1)}(W^{(1)\top}x + W_0^{(1)})$
- 2nd layer: $A^{(2)} = f^{(2)}(W^{(2)\top}A^{(1)} + W_0^{(2)})$
- Whole thing: $A^{(2)} = \mathrm{NN}(x; W, W_0)$



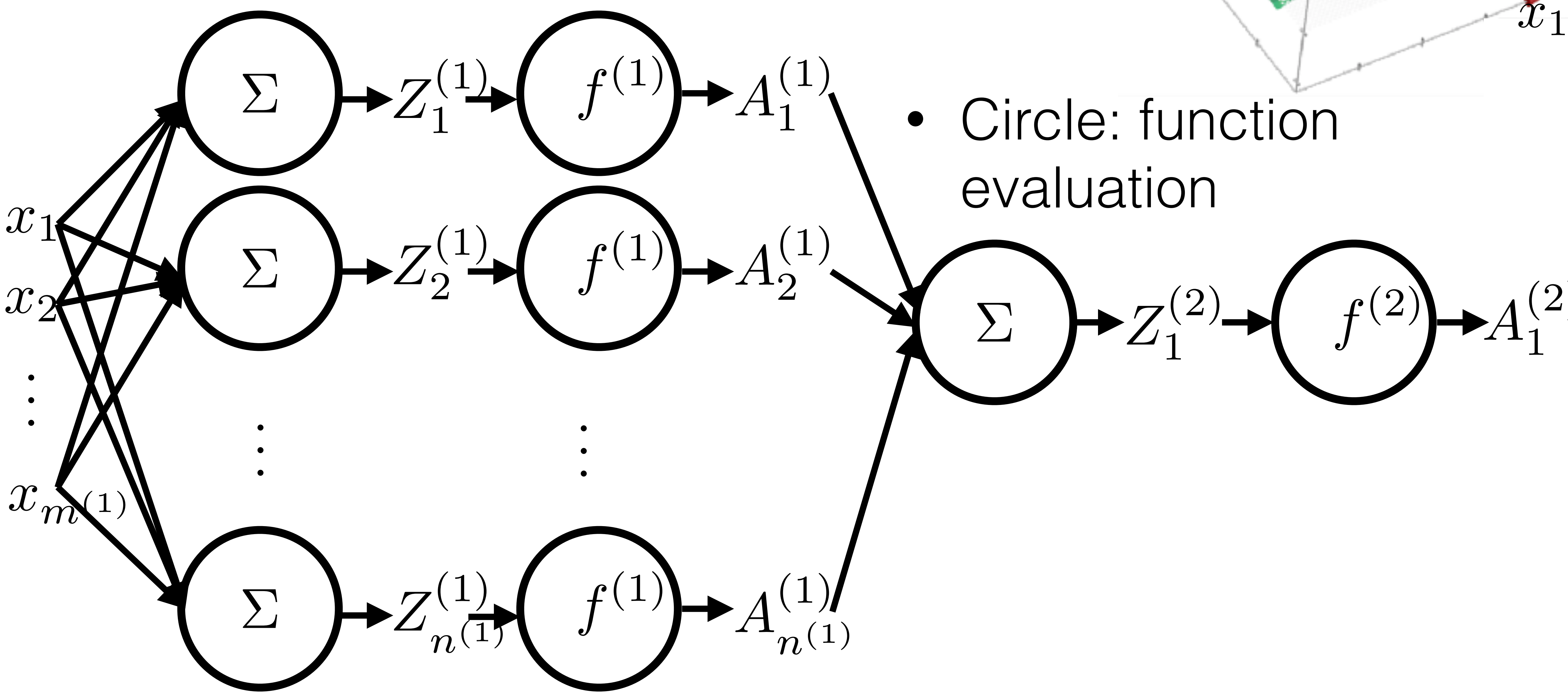


- Circle: function evaluation

Function graph representation

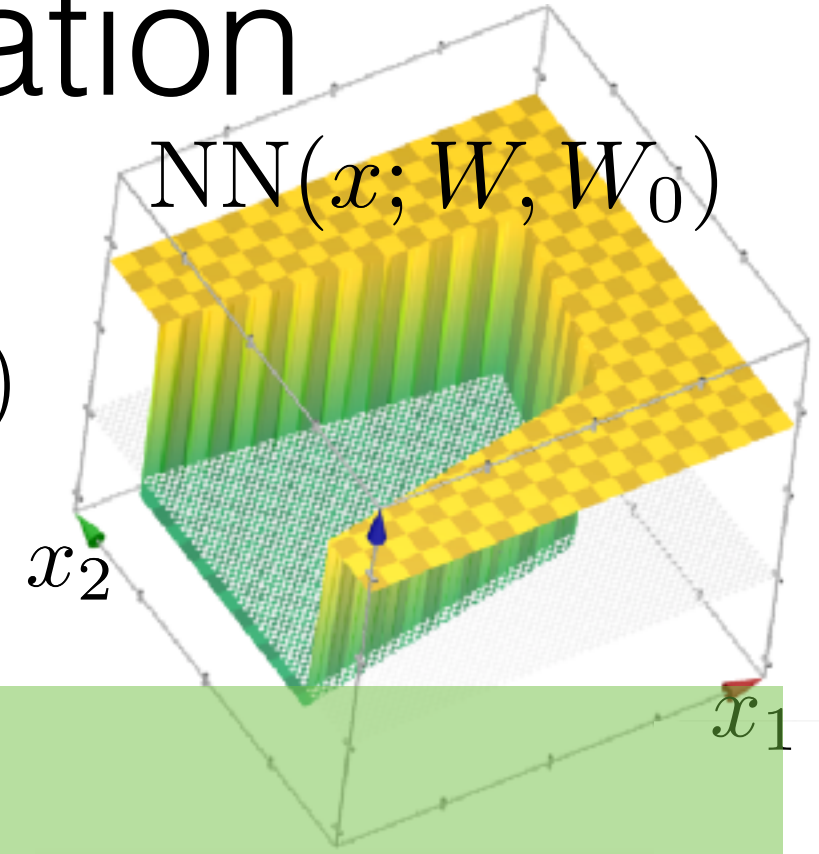


- 1st layer: $A^{(1)} = f^{(1)}(W^{(1)\top} x + W_0^{(1)})$
- 2nd layer: $A^{(2)} = f^{(2)}(W^{(2)\top} A^{(1)} + W_0^{(2)})$
- Whole thing: $A^{(2)} = \mathrm{NN}(x; W, W_0)$

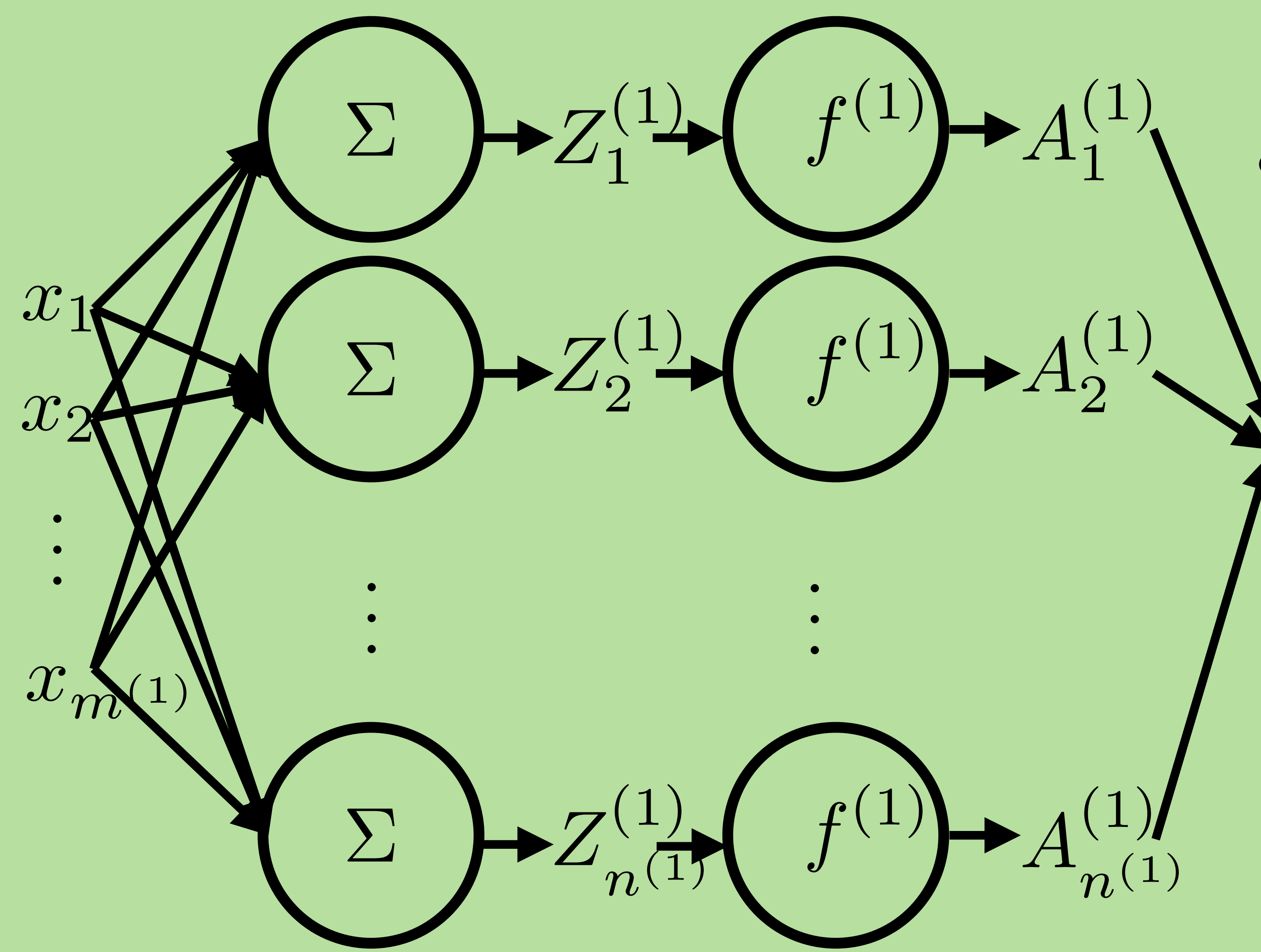


- Circle: function evaluation

Function graph representation



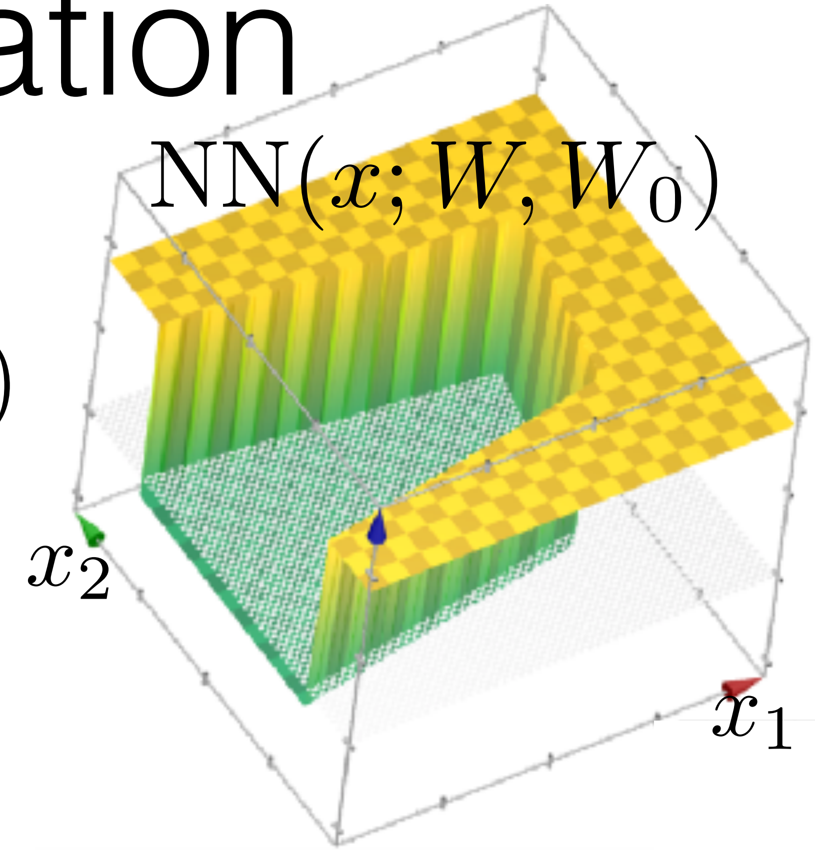
- 1st layer: $A^{(1)} = f^{(1)}(W^{(1)\top} x + W_0^{(1)})$
- 2nd layer: $A^{(2)} = f^{(2)}(W^{(2)\top} A^{(1)} + W_0^{(2)})$
- Whole thing: $A^{(2)} = \mathrm{NN}(x; W, W_0)$

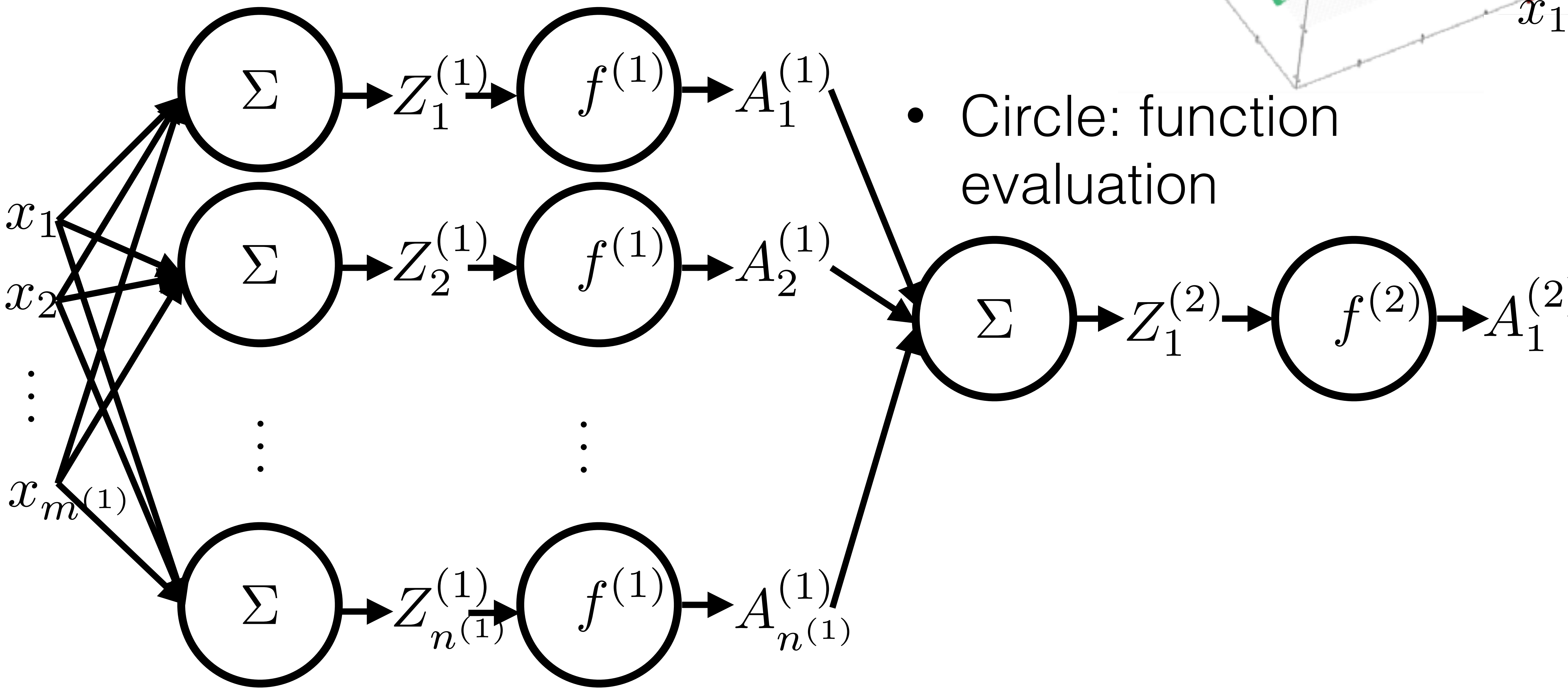


- Circle: function evaluation

Function graph representation

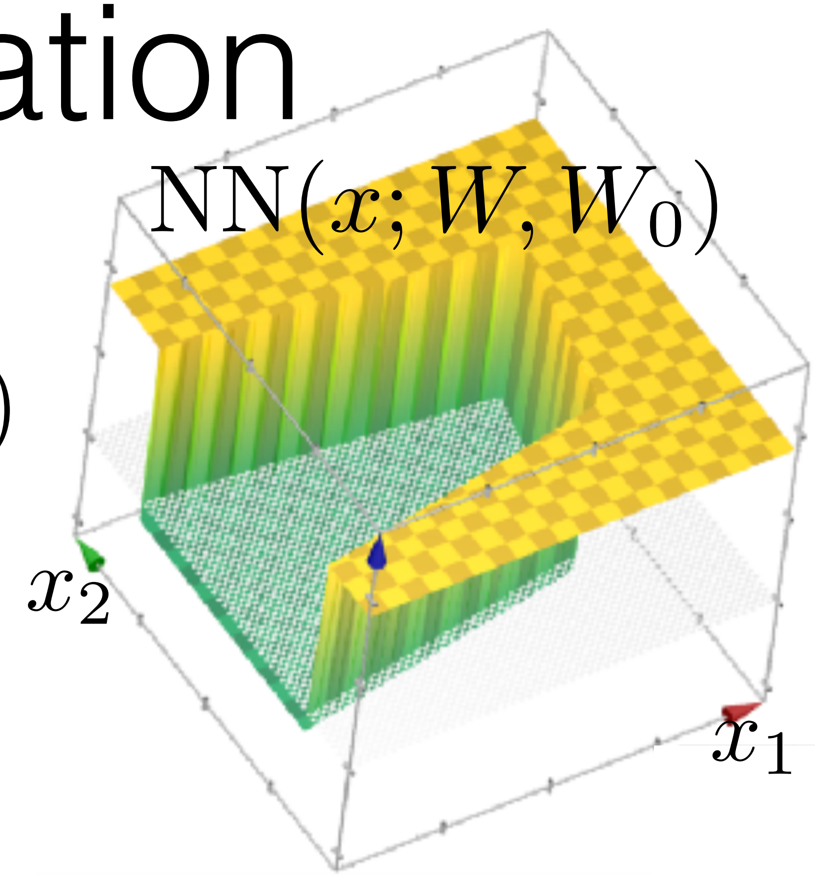
- 1st layer: $A^{(1)} = f^{(1)}(W^{(1)\top}x + W_0^{(1)})$
- 2nd layer: $A^{(2)} = f^{(2)}(W^{(2)\top}A^{(1)} + W_0^{(2)})$
- Whole thing: $A^{(2)} = \mathrm{NN}(x; W, W_0)$



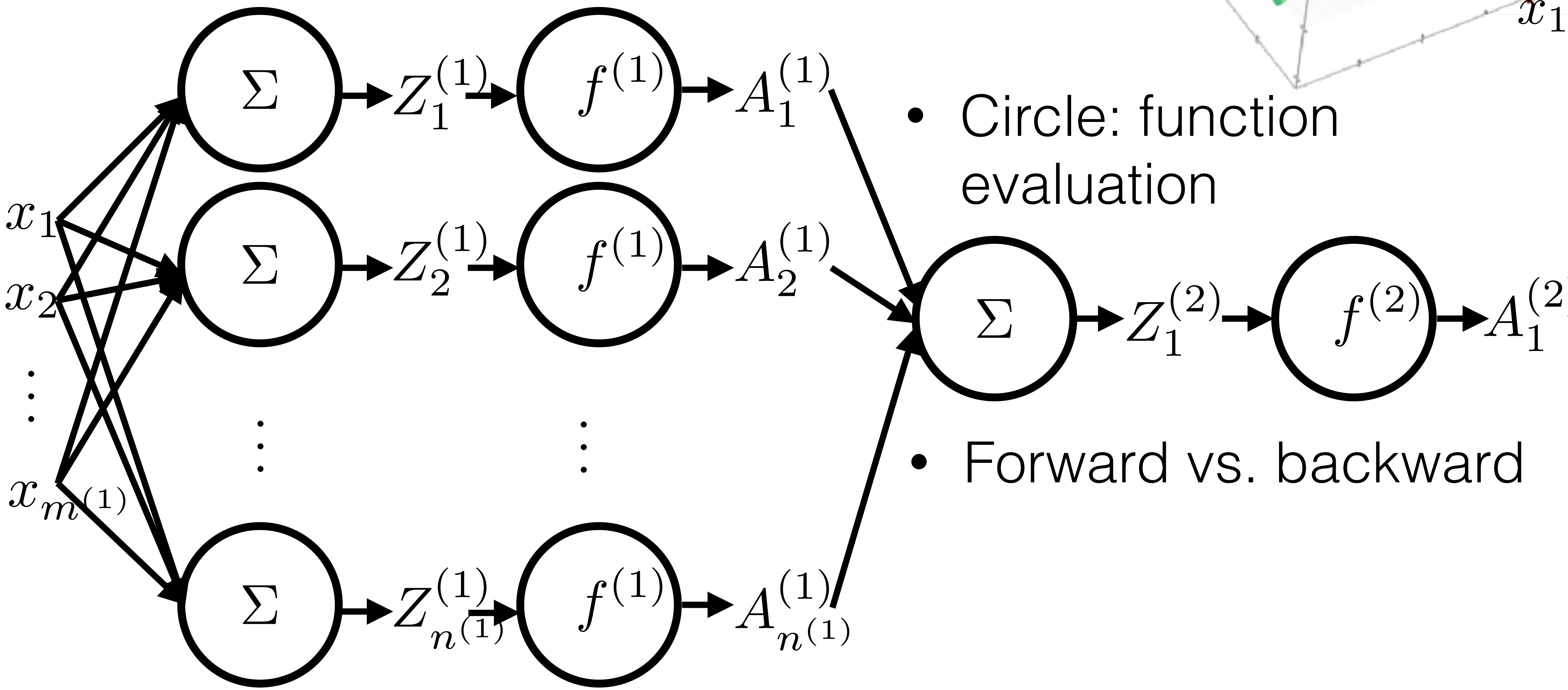


- Circle: function evaluation

Function graph representation



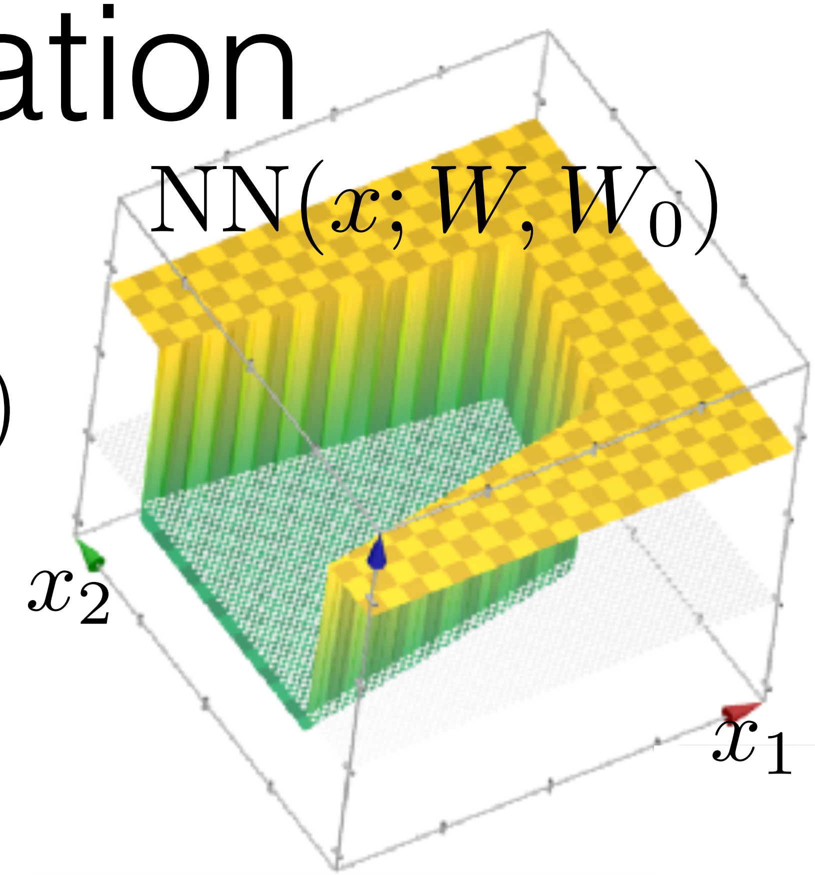
- 1st layer: $A^{(1)} = f^{(1)}(W^{(1)\top} x + W_0^{(1)})$
- 2nd layer: $A^{(2)} = f^{(2)}(W^{(2)\top} A^{(1)} + W_0^{(2)})$
- Whole thing: $A^{(2)} = \mathrm{NN}(x; W, W_0)$

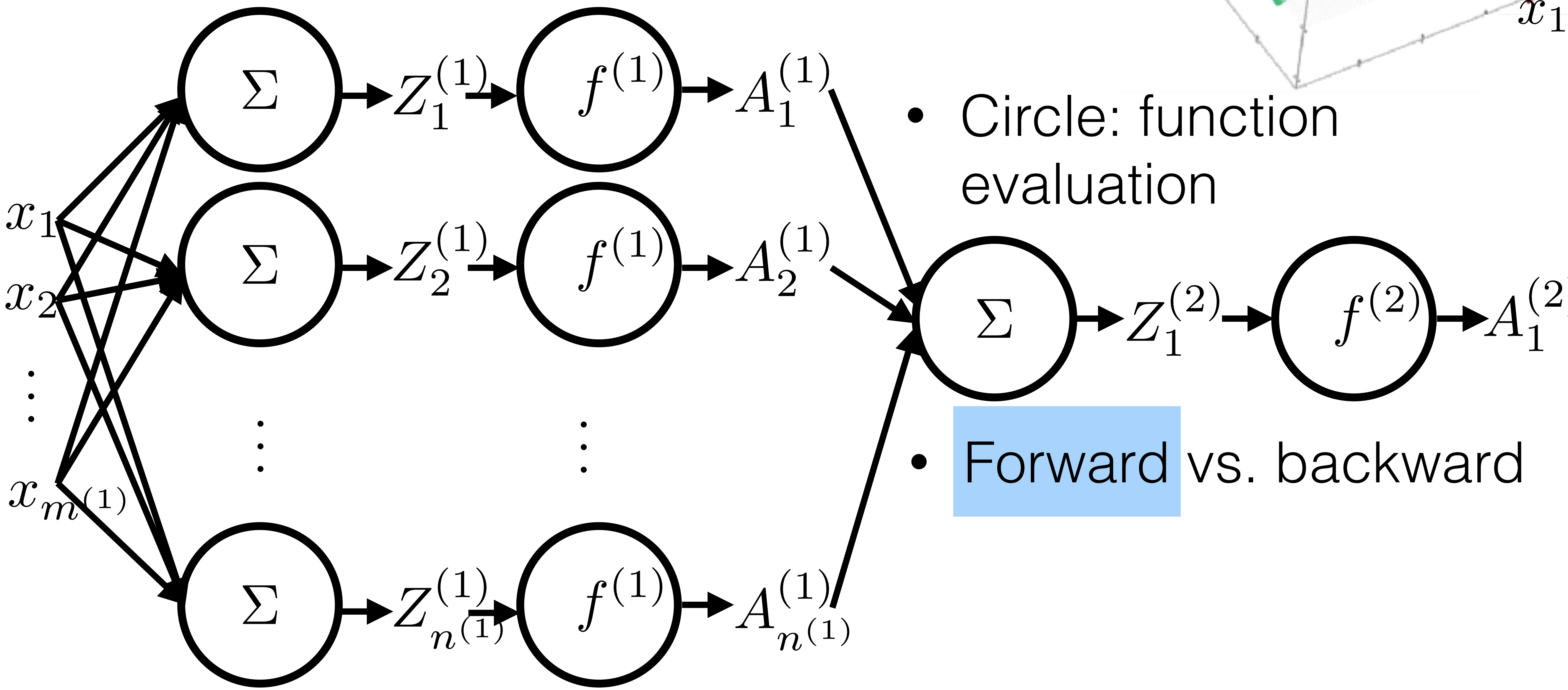


- Circle: function evaluation
- Forward vs. backward

Function graph representation

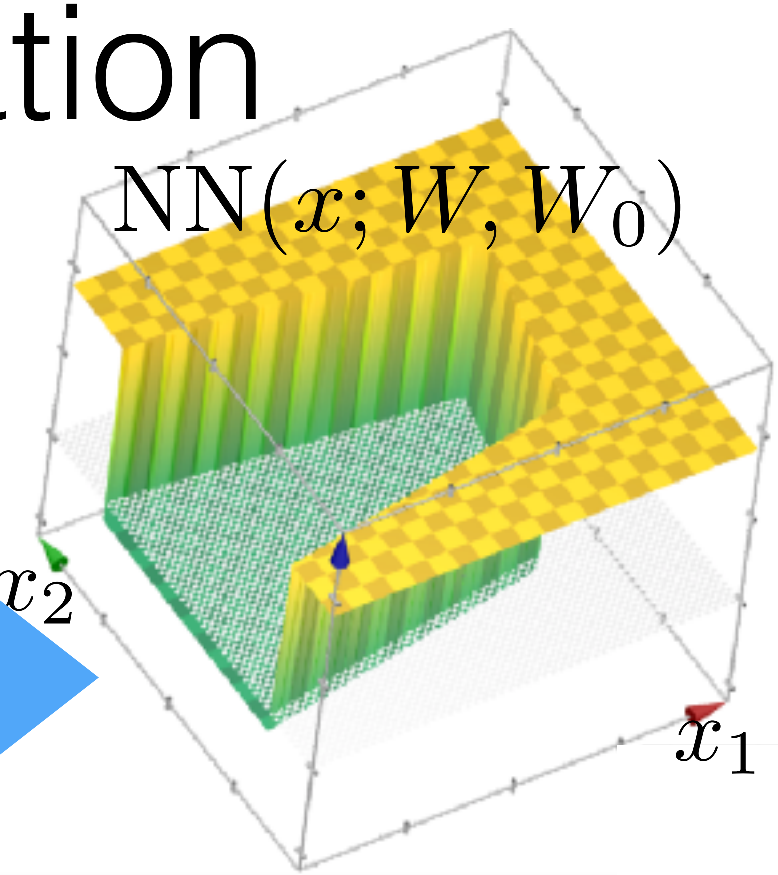
- 1st layer: $A^{(1)} = f^{(1)}(W^{(1)\top}x + W_0^{(1)})$
- 2nd layer: $A^{(2)} = f^{(2)}(W^{(2)\top}A^{(1)} + W_0^{(2)})$
- Whole thing: $A^{(2)} = \mathrm{NN}(x; W, W_0)$



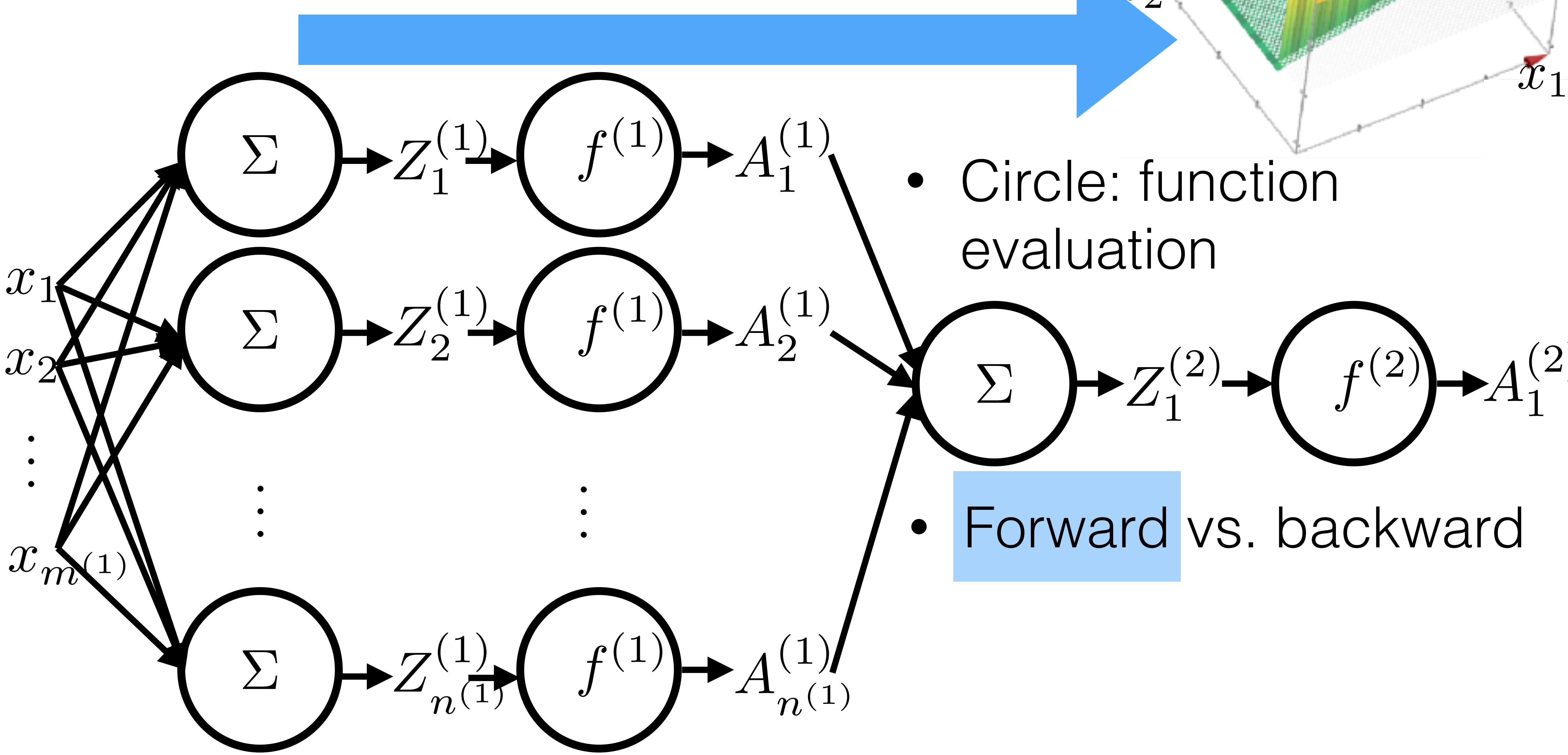


- Circle: function evaluation
- Forward vs. backward

Function graph representation



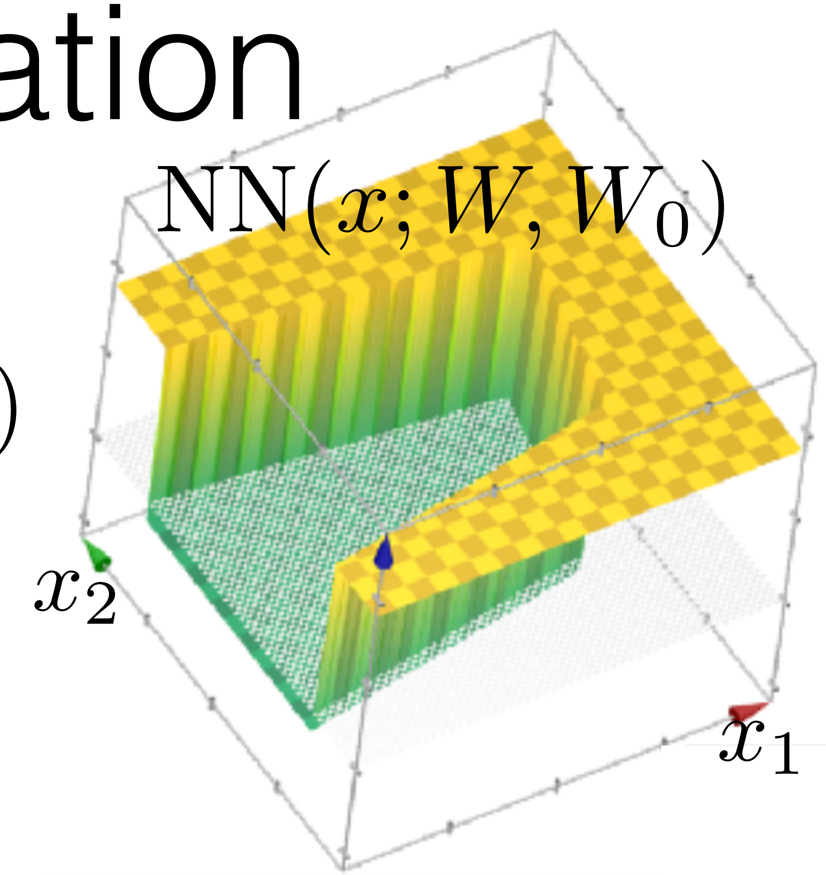
- 1st layer: $A^{(1)} = f^{(1)}(W^{(1)\top} x + W_0^{(1)})$
- 2nd layer: $A^{(2)} = f^{(2)}(W^{(2)\top} A^{(1)} + W_0^{(2)})$
- Whole thing: $A^{(2)} = \mathrm{NN}(x; W, W_0)$

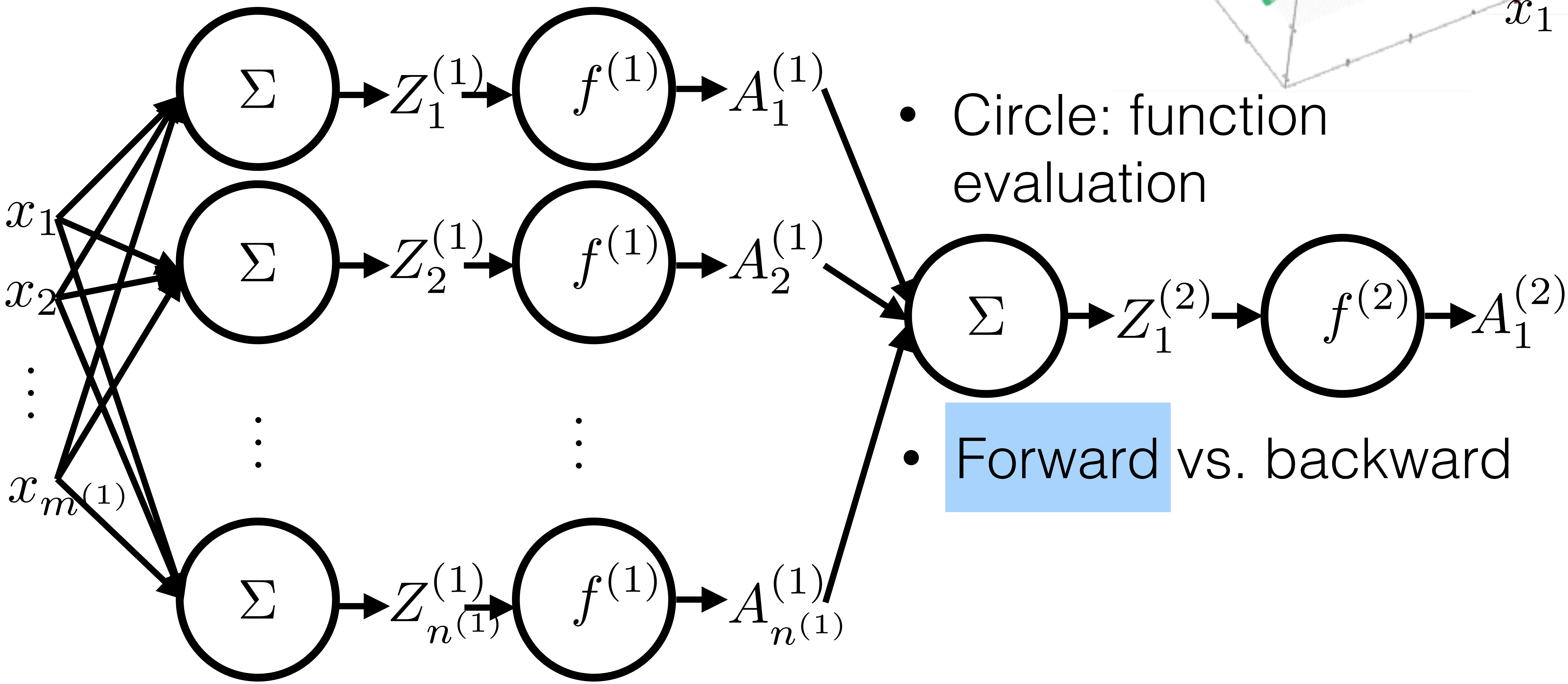


- Circle: function evaluation
- Forward vs. backward

Function graph representation

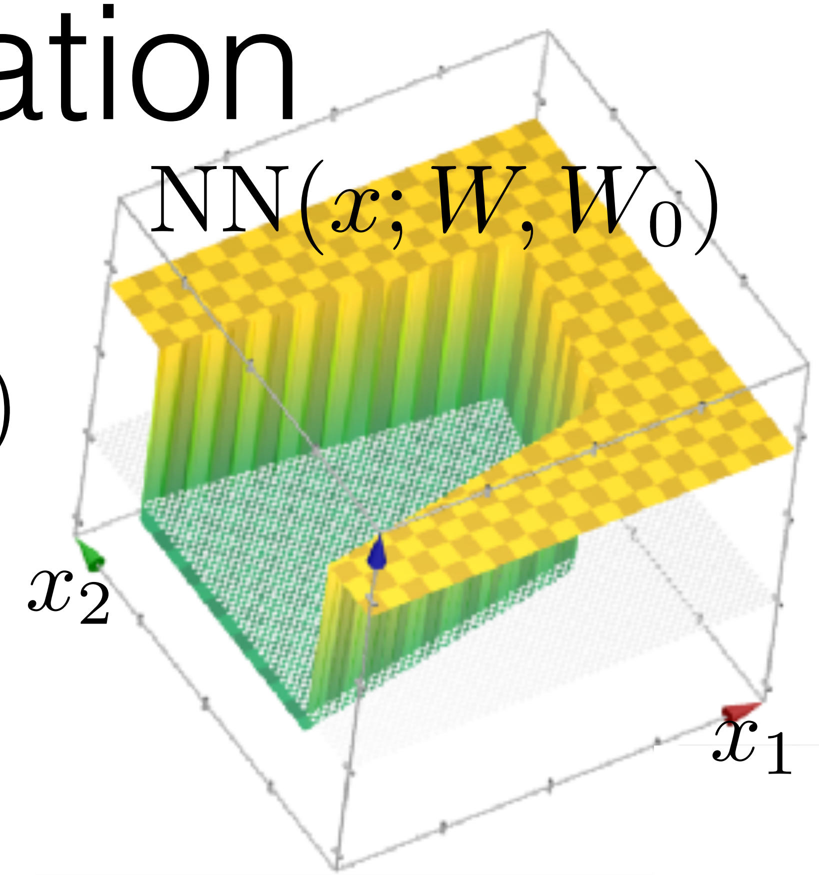
- 1st layer: $A^{(1)} = f^{(1)}(W^{(1)\top} x + W_0^{(1)})$
- 2nd layer: $A^{(2)} = f^{(2)}(W^{(2)\top} A^{(1)} + W_0^{(2)})$
- Whole thing: $A^{(2)} = \mathrm{NN}(x; W, W_0)$



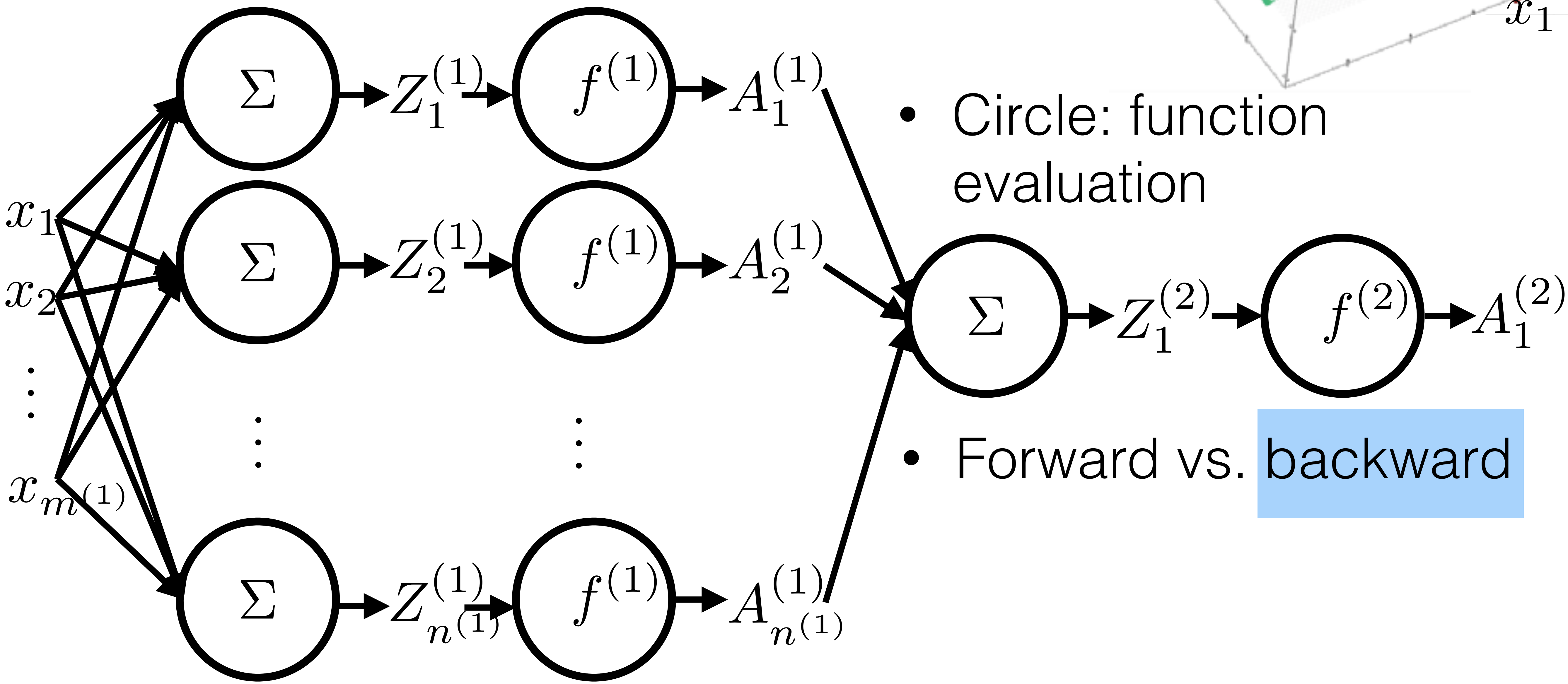


- Circle: function evaluation
- Forward vs. backward

Function graph representation

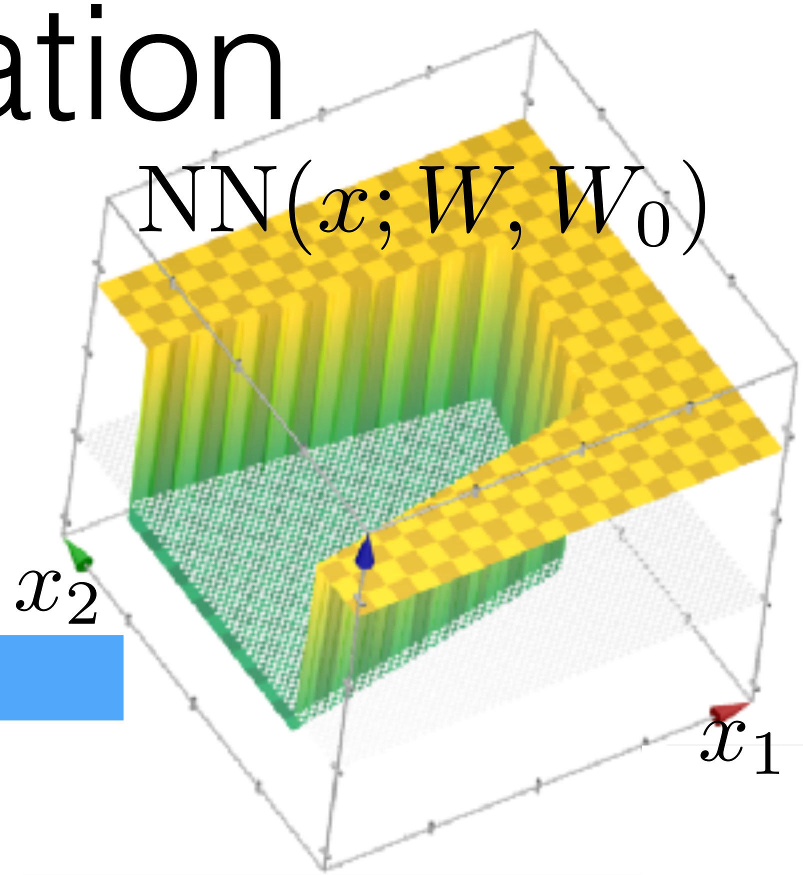


- 1st layer: $A^{(1)} = f^{(1)}(W^{(1)\top} x + W_0^{(1)})$
- 2nd layer: $A^{(2)} = f^{(2)}(W^{(2)\top} A^{(1)} + W_0^{(2)})$
- Whole thing: $A^{(2)} = \mathrm{NN}(x; W, W_0)$

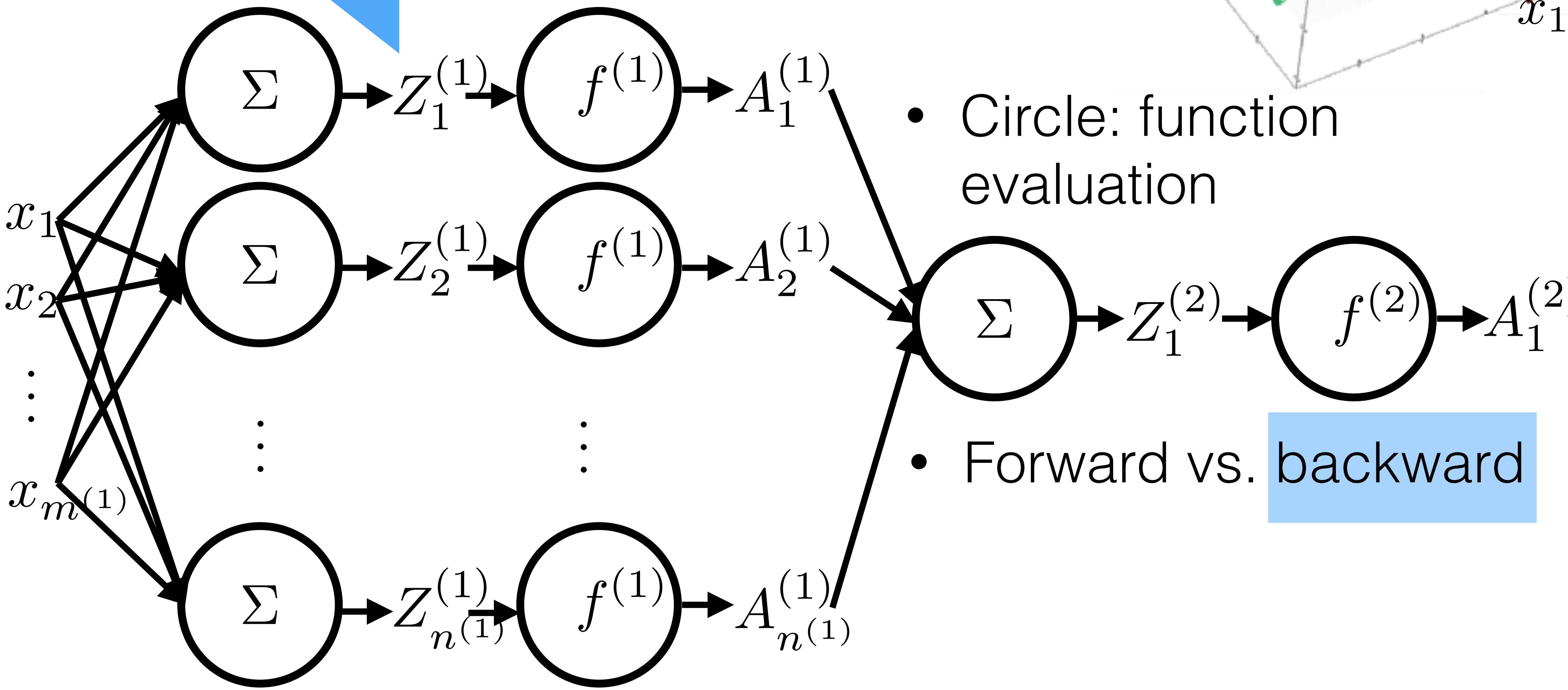


- Circle: function evaluation
- Forward vs. backward

Function graph representation

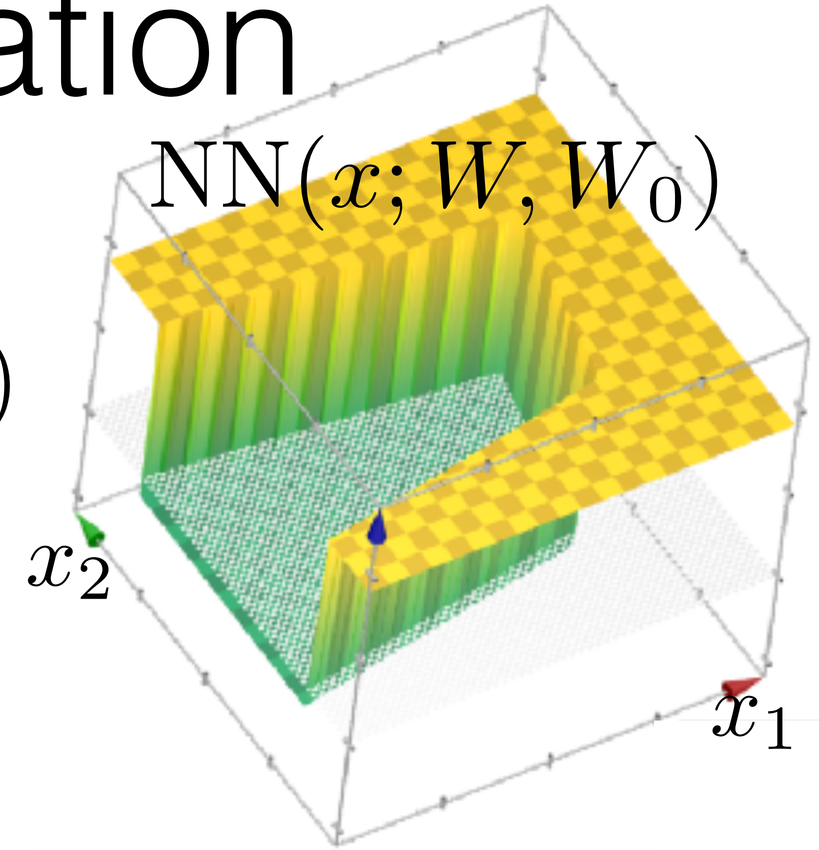


- 1st layer: $A^{(1)} = f^{(1)}(W^{(1)\top} x + W_0^{(1)})$
- 2nd layer: $A^{(2)} = f^{(2)}(W^{(2)\top} A^{(1)} + W_0^{(2)})$
- Whole thing: $A^{(2)} = \mathrm{NN}(x; W, W_0)$

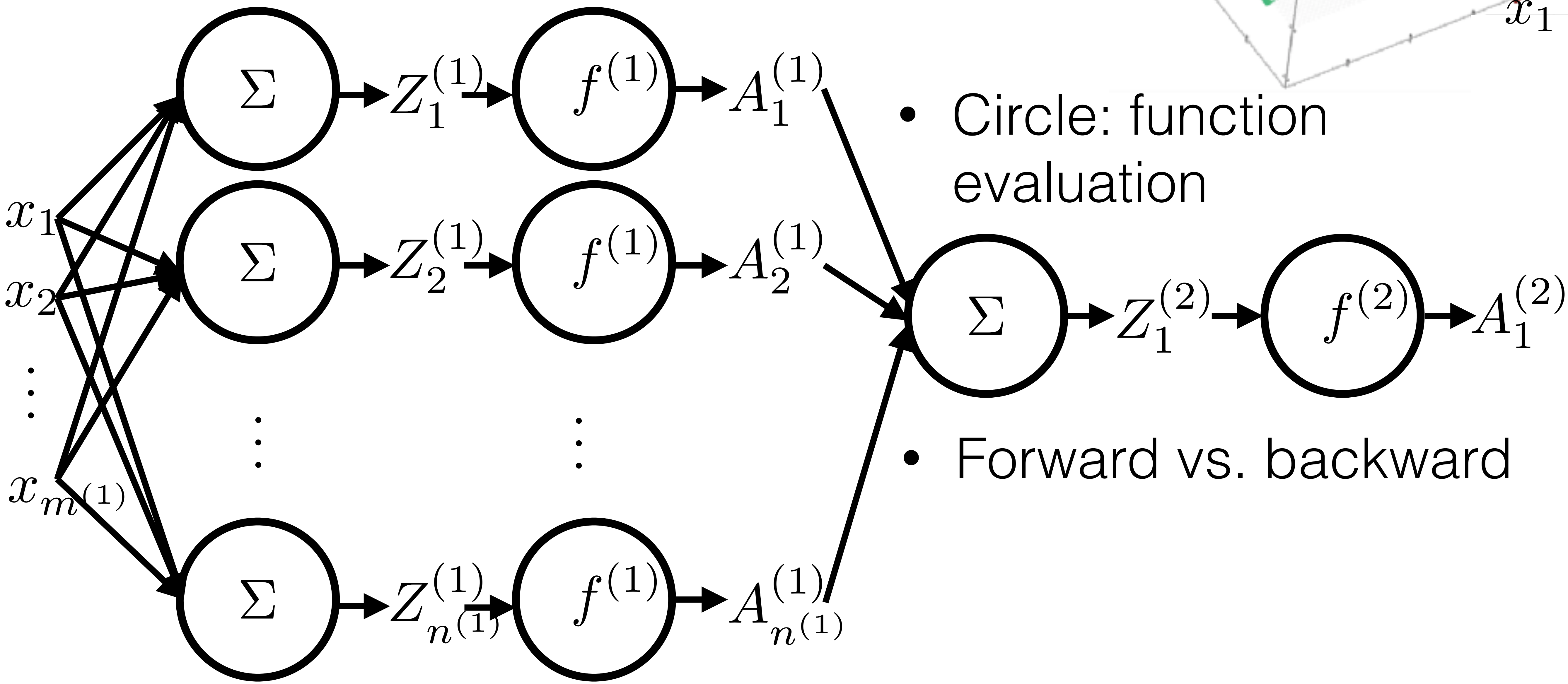


- Circle: function evaluation
- Forward vs. backward

Function graph representation



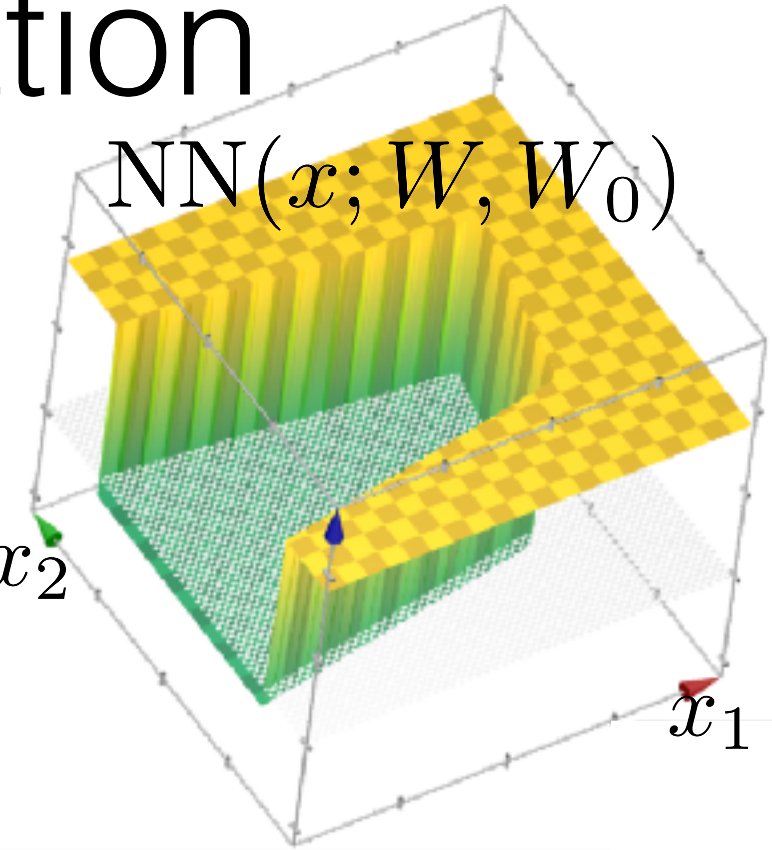
- 1st layer: $A^{(1)} = f^{(1)}(W^{(1)\top} x + W_0^{(1)})$
- 2nd layer: $A^{(2)} = f^{(2)}(W^{(2)\top} A^{(1)} + W_0^{(2)})$
- Whole thing: $A^{(2)} = \mathrm{NN}(x; W, W_0)$

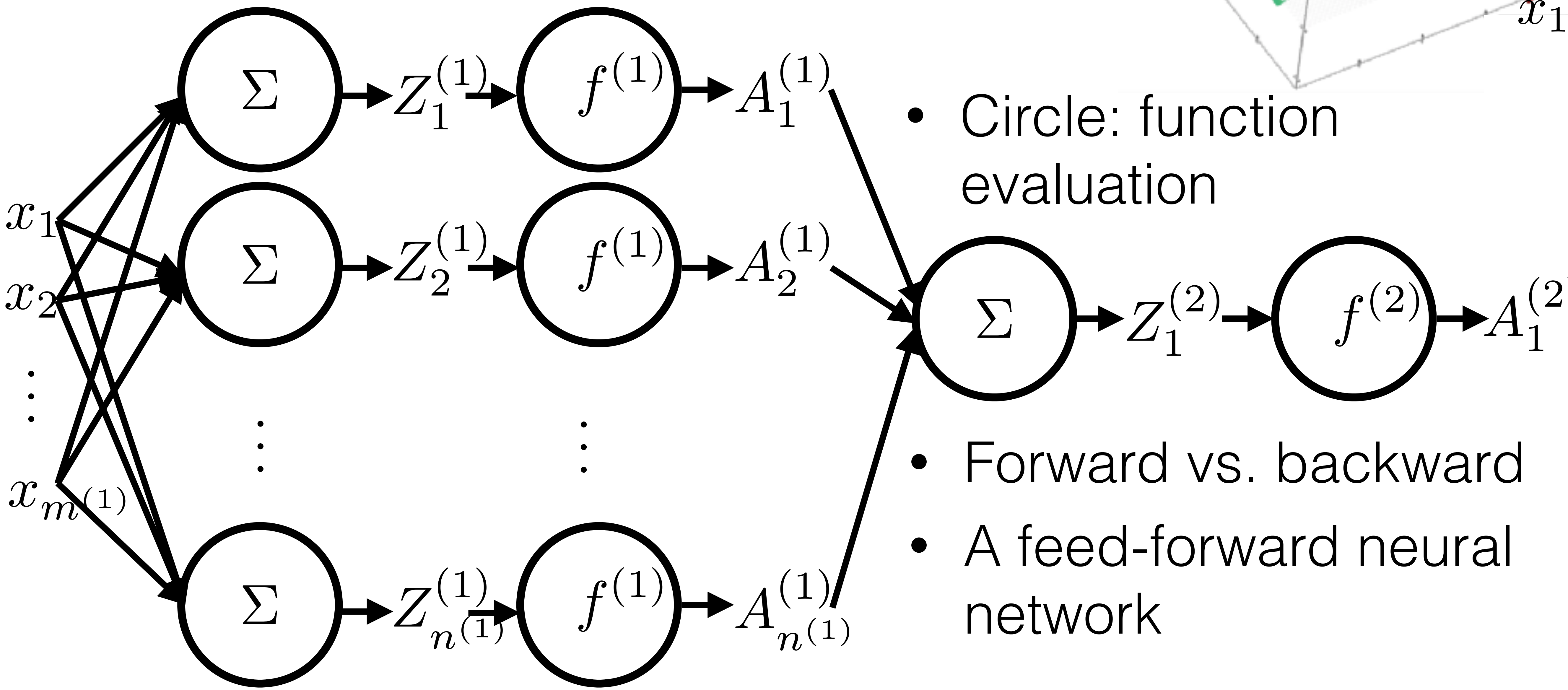


- Circle: function evaluation
- Forward vs. backward

Function graph representation

- 1st layer: $A^{(1)} = f^{(1)}(W^{(1)\top}x + W_0^{(1)})$
- 2nd layer: $A^{(2)} = f^{(2)}(W^{(2)\top}A^{(1)} + W_0^{(2)})$
- Whole thing: $A^{(2)} = \mathrm{NN}(x; W, W_0)$

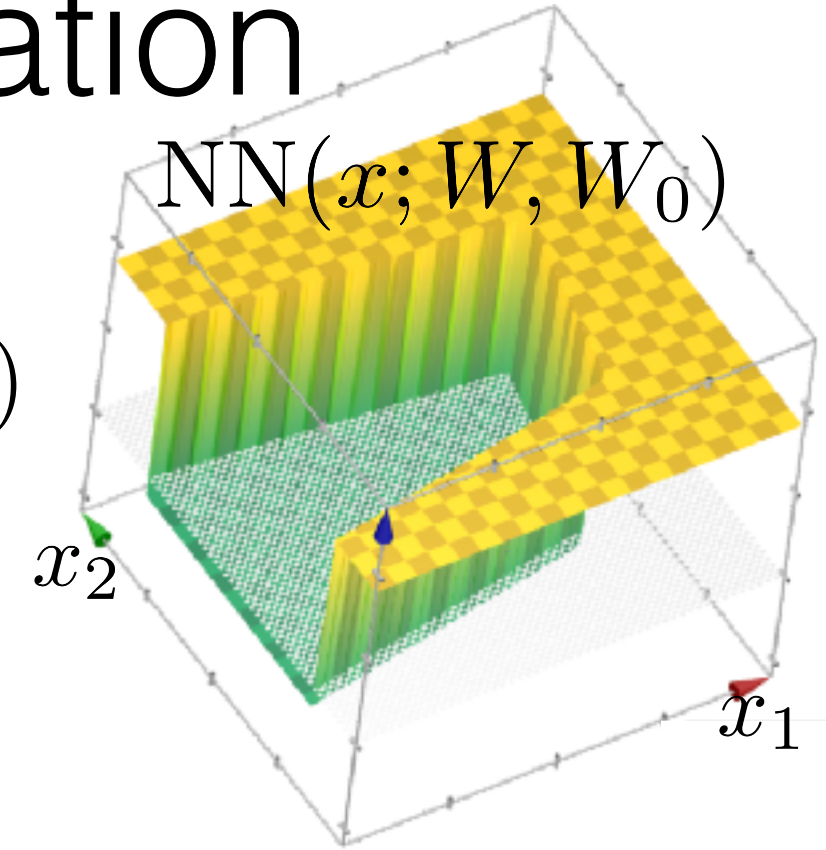


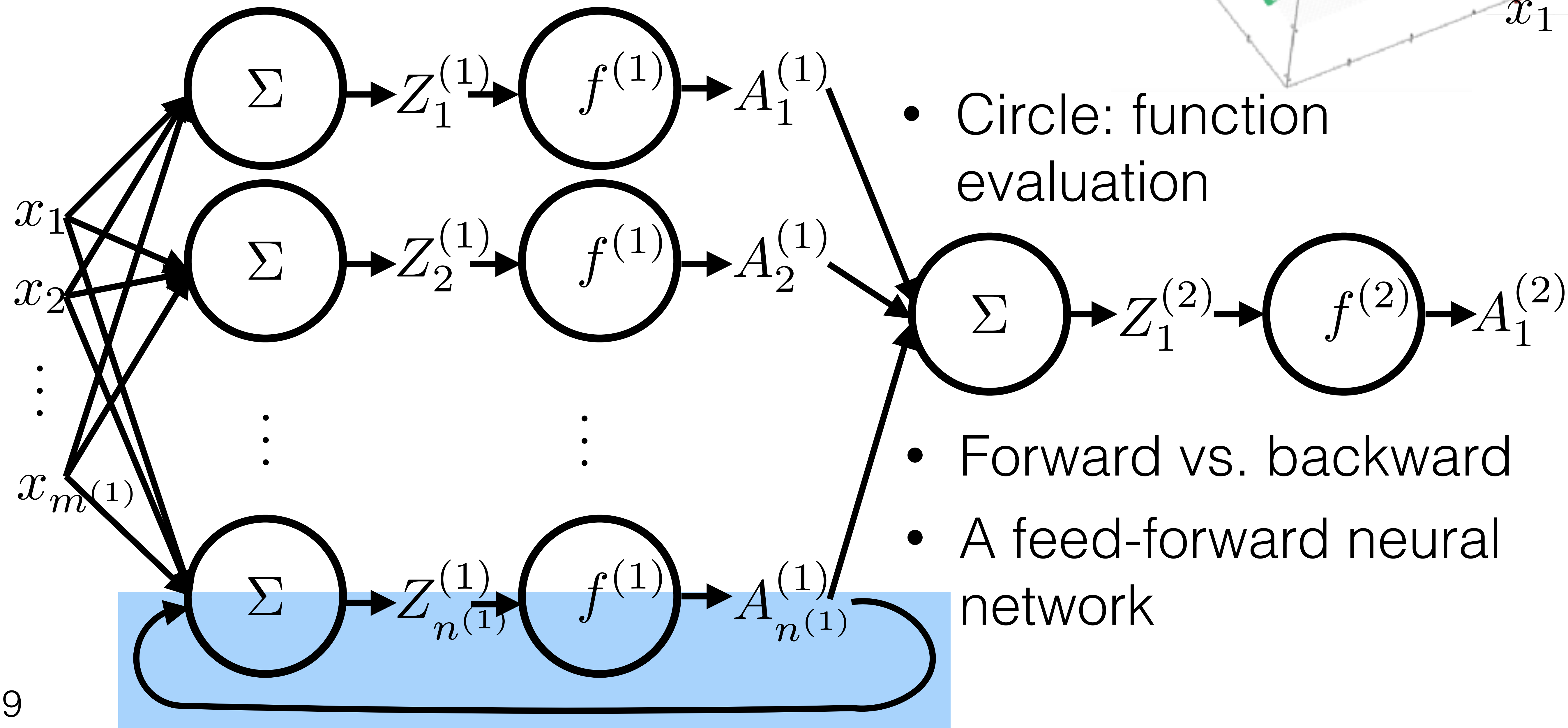


- Circle: function evaluation
- Forward vs. backward
- A feed-forward neural network

Function graph representation

- 1st layer: $A^{(1)} = f^{(1)}(W^{(1)\top} x + W_0^{(1)})$
- 2nd layer: $A^{(2)} = f^{(2)}(W^{(2)\top} A^{(1)} + W_0^{(2)})$
- Whole thing: $A^{(2)} = \mathrm{NN}(x; W, W_0)$

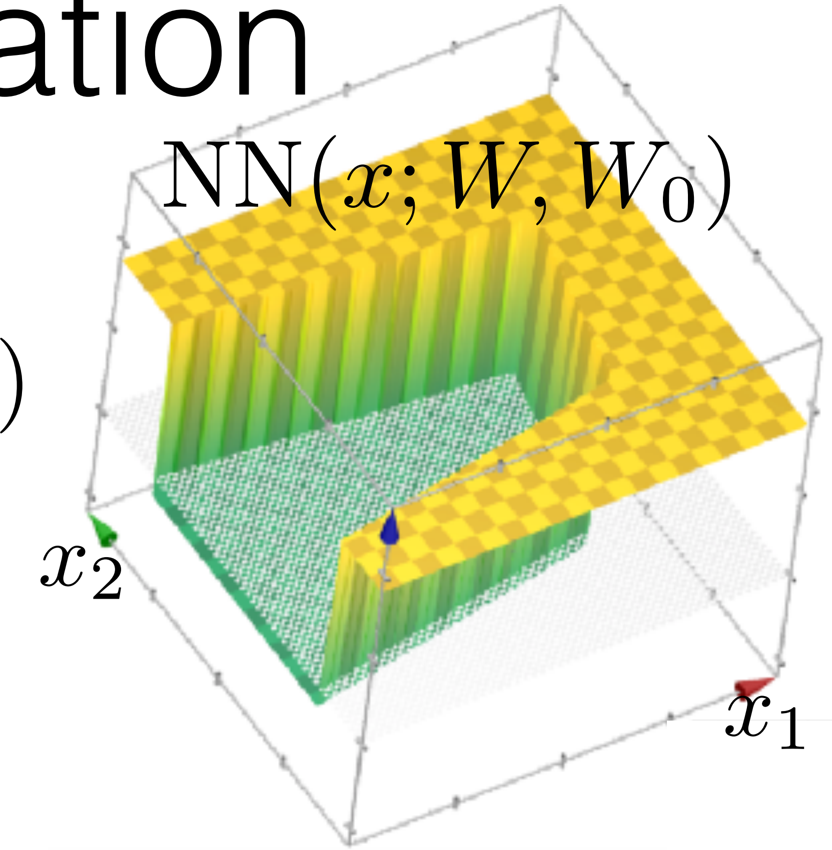


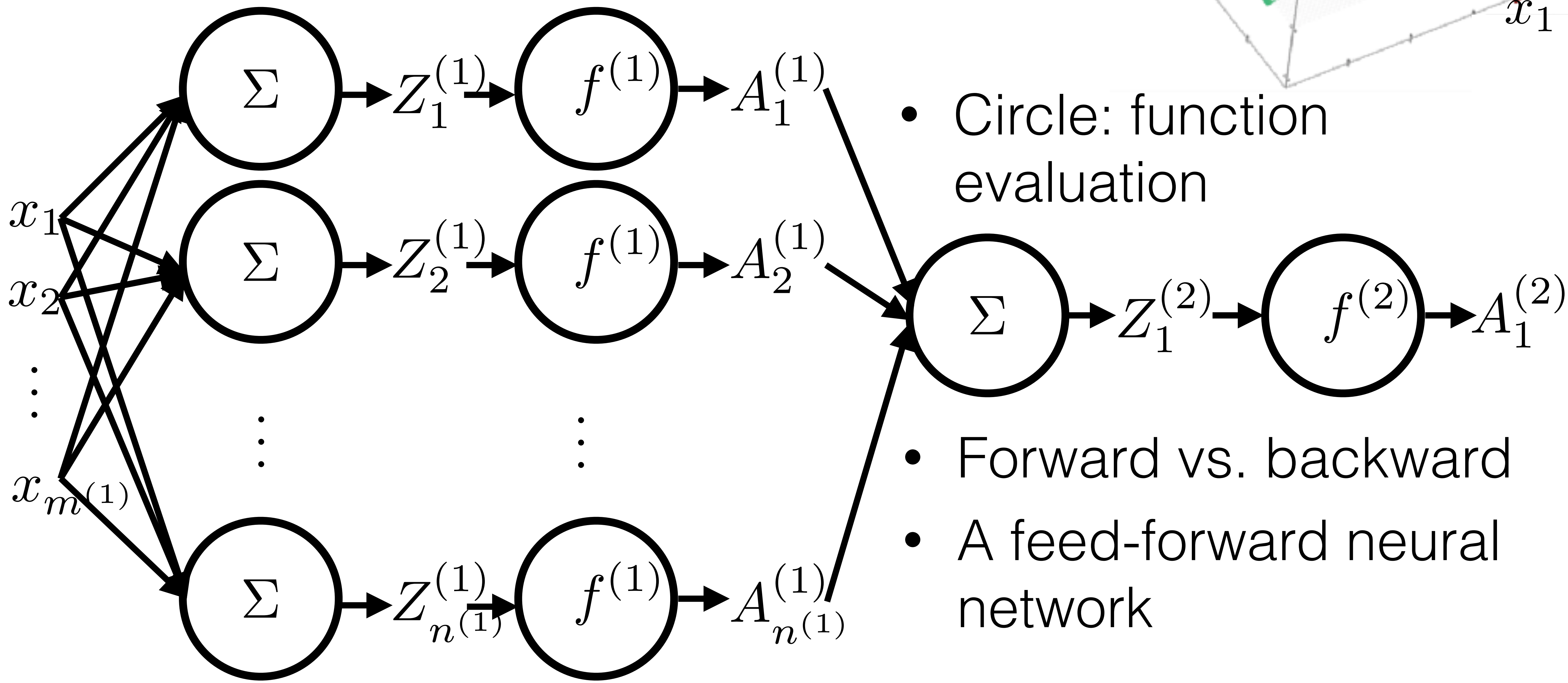


- Circle: function evaluation
- Forward vs. backward
- A feed-forward neural network

Function graph representation

- 1st layer: $A^{(1)} = f^{(1)}(W^{(1)\top} x + W_0^{(1)})$
- 2nd layer: $A^{(2)} = f^{(2)}(W^{(2)\top} A^{(1)} + W_0^{(2)})$
- Whole thing: $A^{(2)} = \mathrm{NN}(x; W, W_0)$

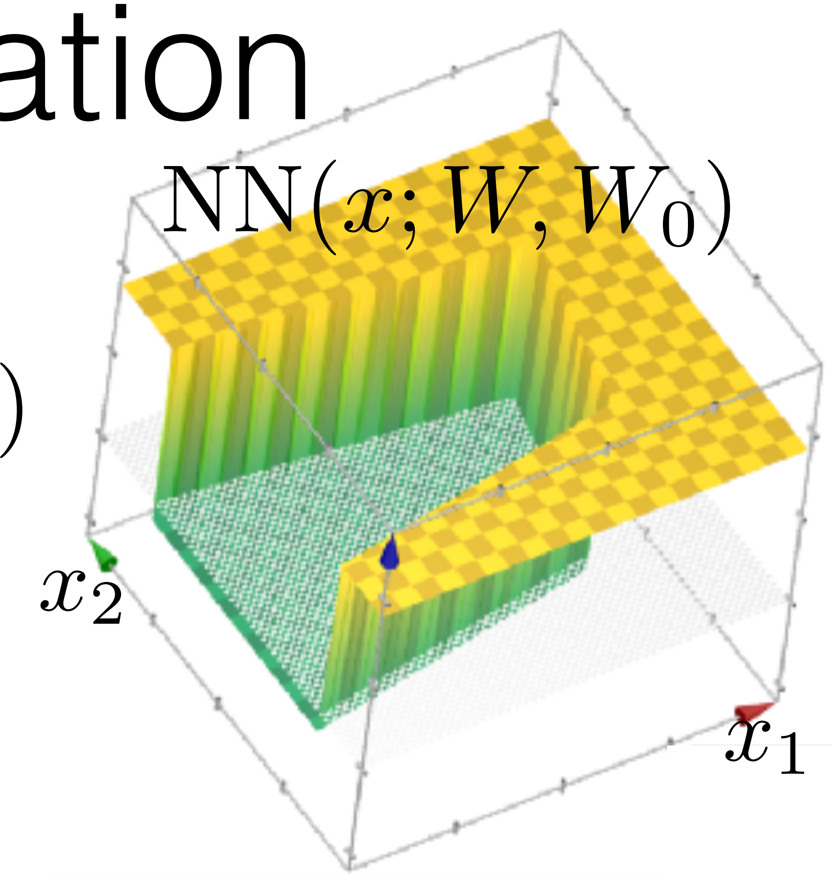


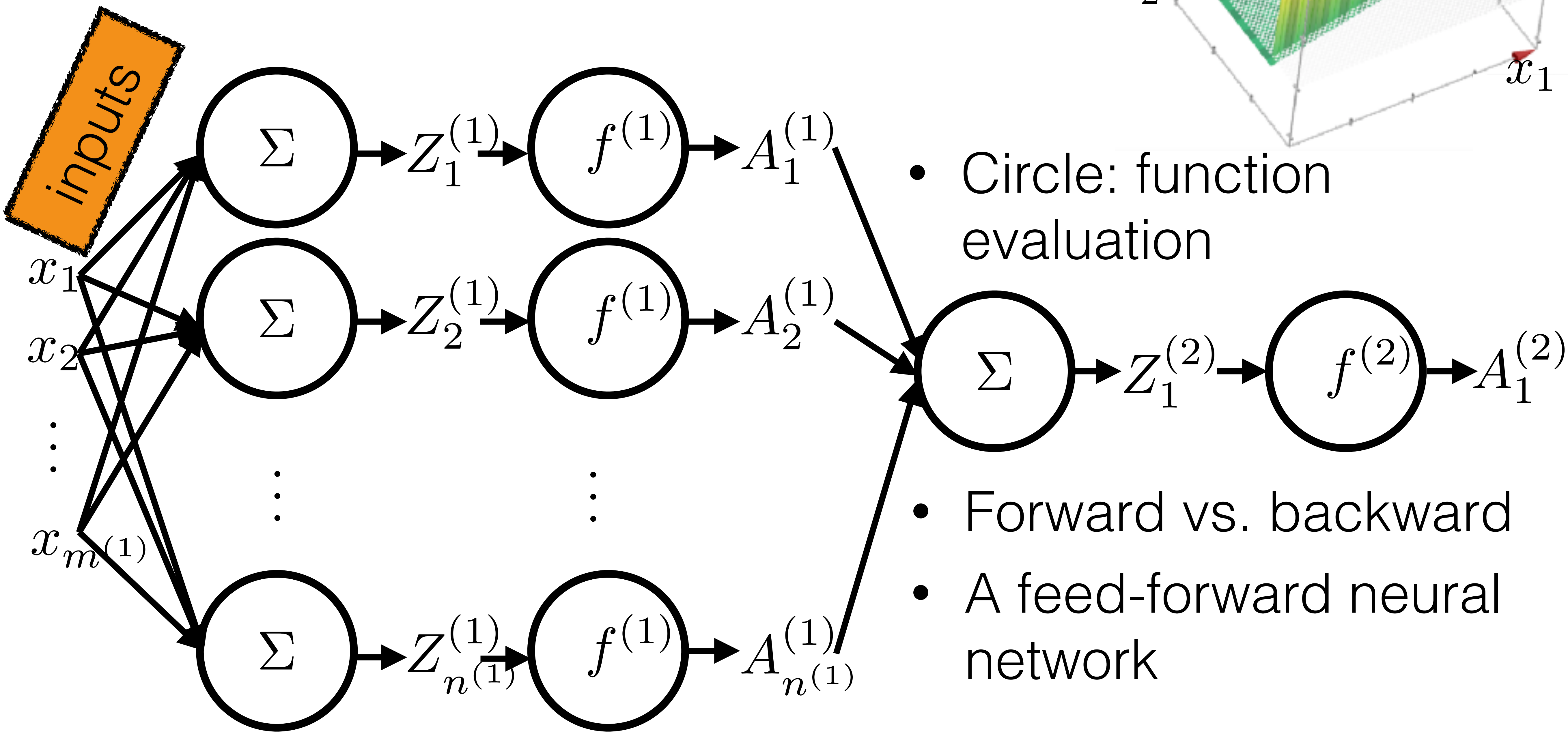


- Circle: function evaluation
- Forward vs. backward
- A feed-forward neural network

Function graph representation

- 1st layer: $A^{(1)} = f^{(1)}(W^{(1)\top} x + W_0^{(1)})$
- 2nd layer: $A^{(2)} = f^{(2)}(W^{(2)\top} A^{(1)} + W_0^{(2)})$
- Whole thing: $A^{(2)} = \mathrm{NN}(x; W, W_0)$



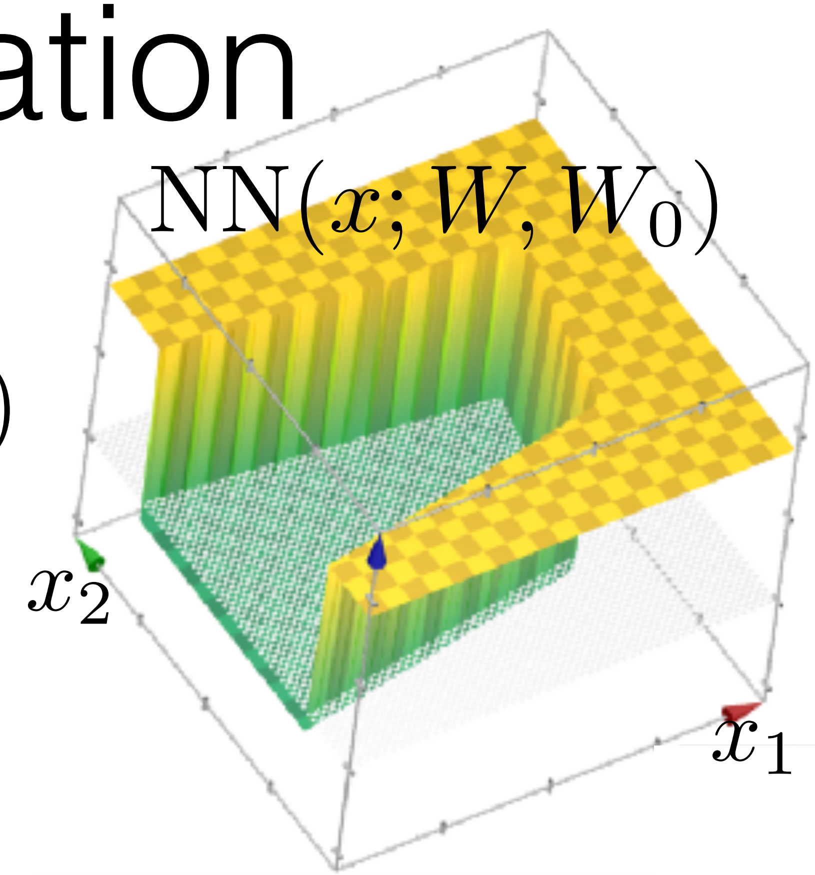


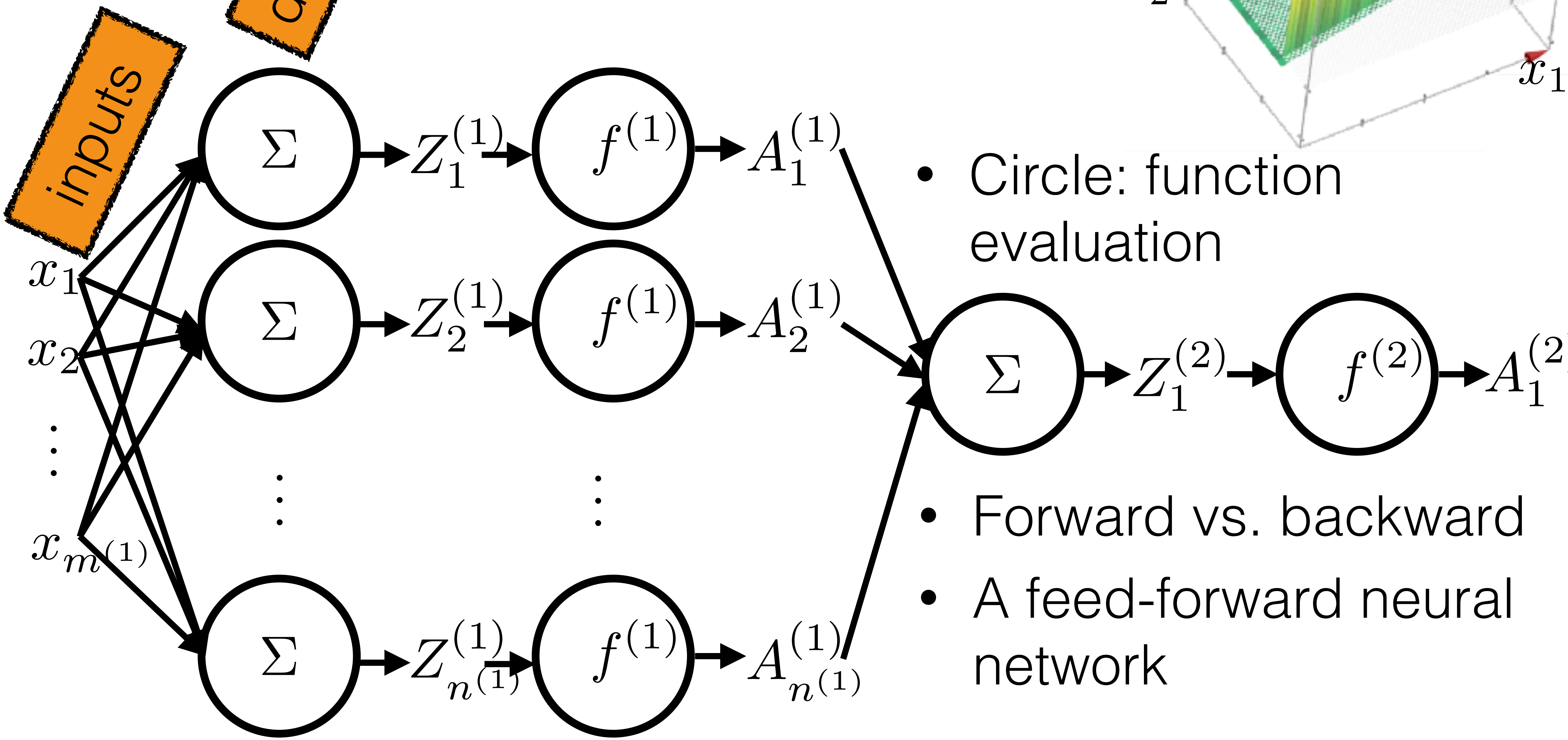
- Circle: function evaluation
- Forward vs. backward
- A feed-forward neural network

Function graph representation

- 1st layer: $= f^{(1)}(W^{(1)\top}x + W_0^{(1)})$
- 2nd layer: $A^{(2)} = f^{(2)}(W^{(2)\top}A^{(1)} + W_0^{(2)})$
- Whole thing: $A^{(2)} = \mathrm{NN}(x; W, W_0)$

dot product

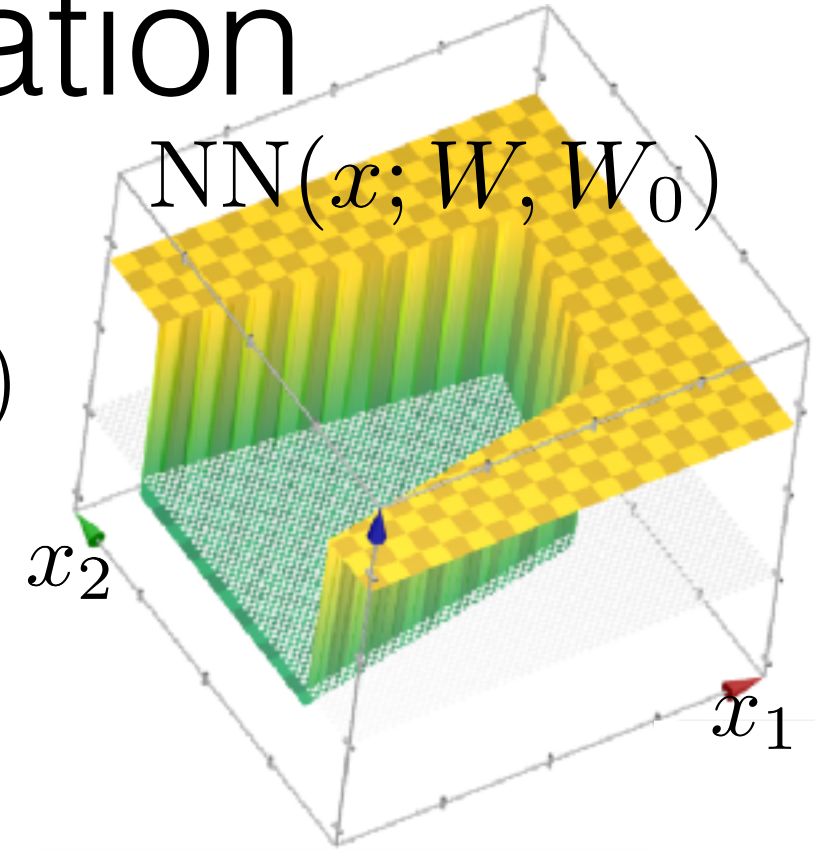


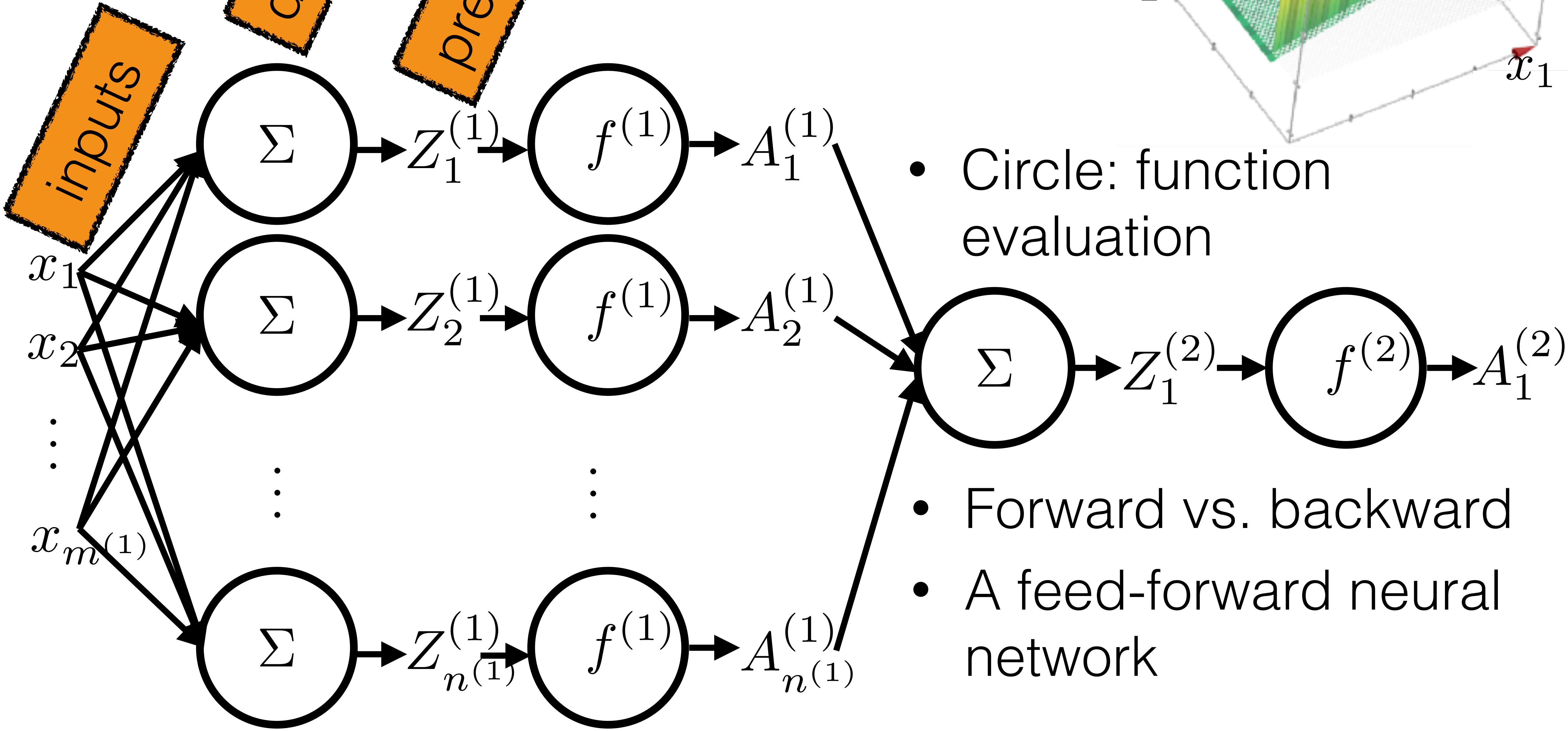


- Circle: function evaluation
- Forward vs. backward
- A feed-forward neural network

Function graph representation

- 1st layer: $= \quad (W^{(1)\top} x + W_0^{(1)})$
- 2nd layer $^{(2)} \quad ^{2)}(W^{(2)\top} A^{(1)} + W_0^{(2)})$
- Whole t: $A \quad = \mathrm{NN}(x; W, W_0)$

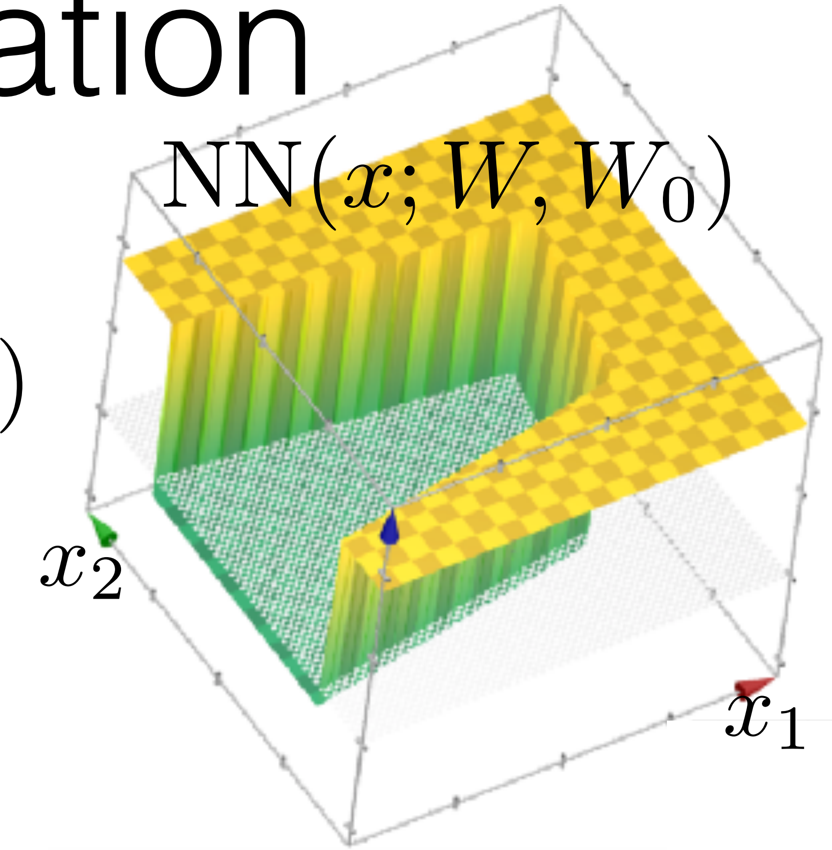


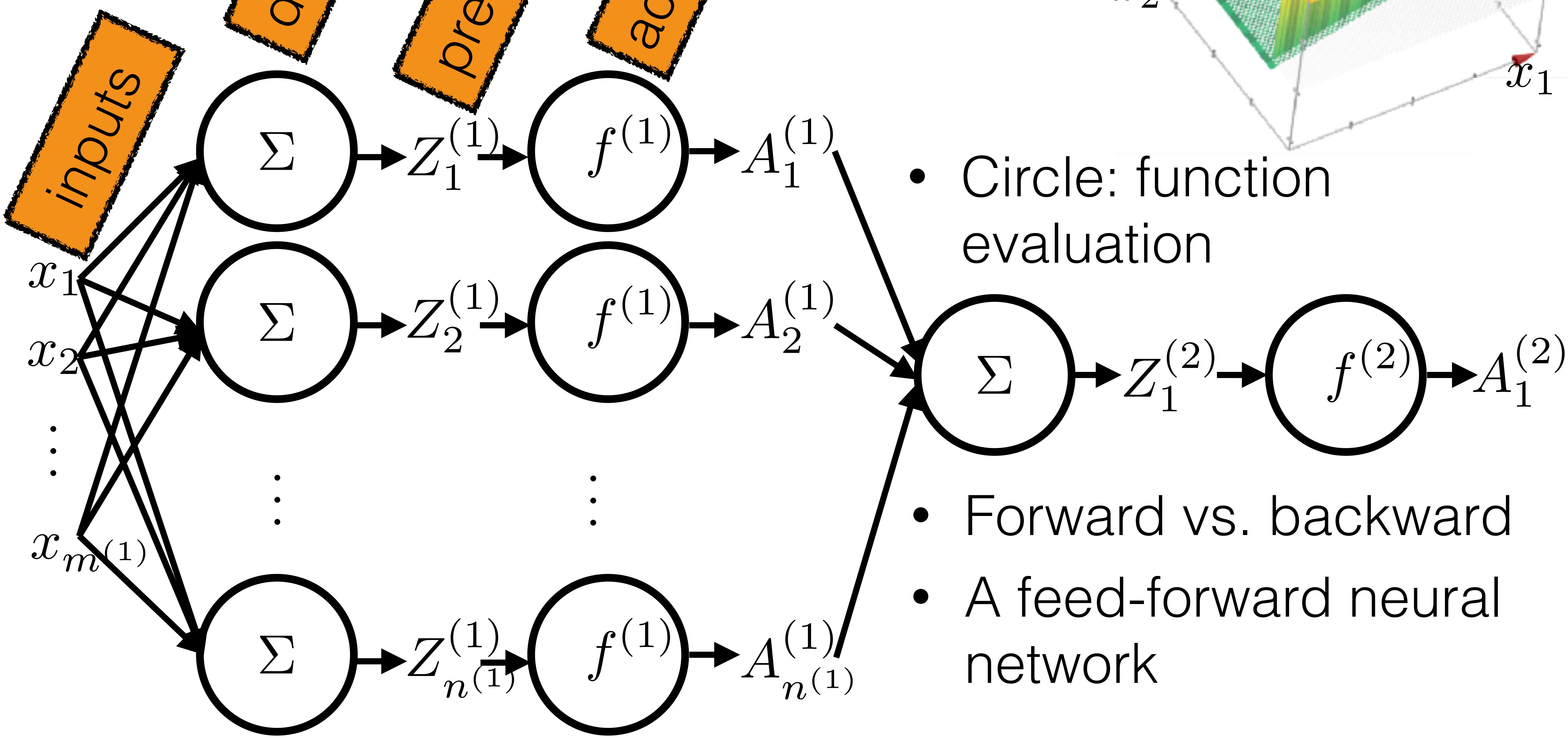


- Circle: function evaluation
- Forward vs. backward
- A feed-forward neural network

Function graph representation

- 1st layer: $A^{(1)} = f^{(1)}(W^{(1)\top} x + W_0^{(1)})$
- 2nd layer: $A^{(2)} = f^{(2)}(W^{(2)\top} A^{(1)} + W_0^{(2)})$
- Whole thing: $A^{(2)} = \mathrm{NN}(x; W, W_0)$

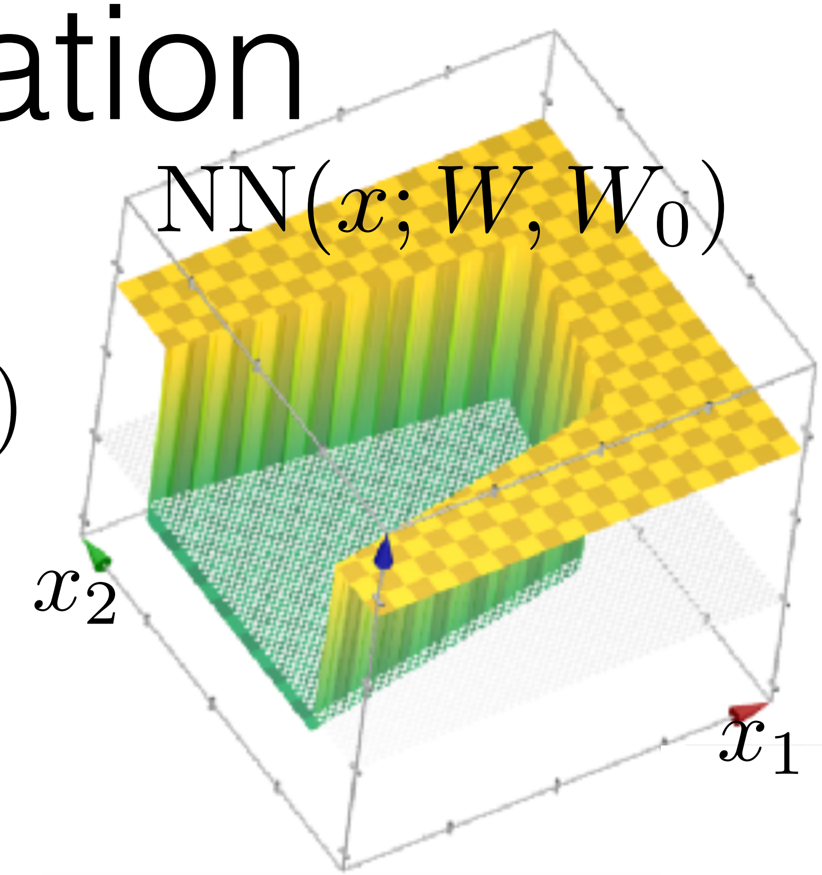


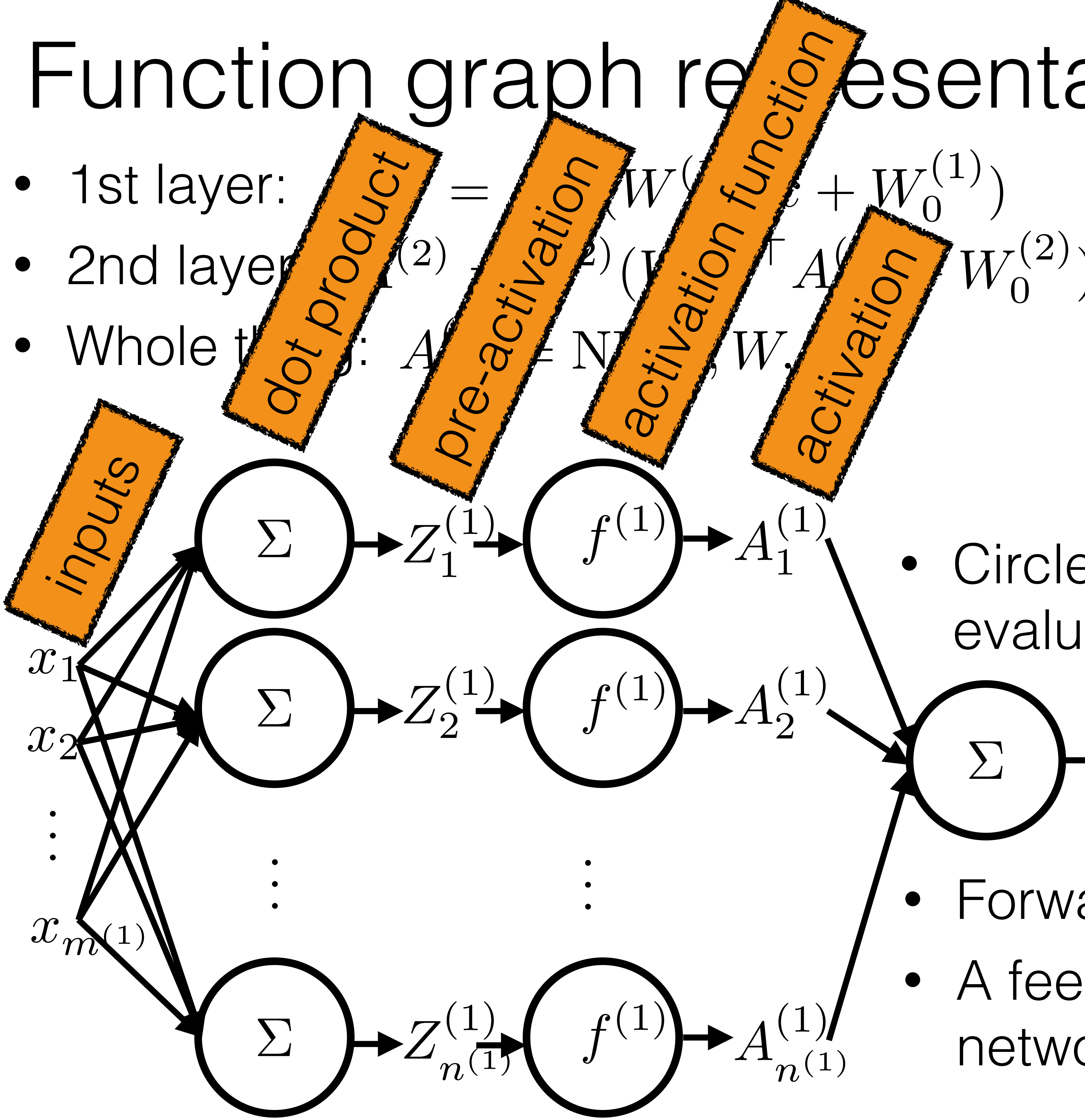


- Circle: function evaluation
- Forward vs. backward
- A feed-forward neural network

Function graph representation

- 1st layer: $A^{(1)} = f^{(1)}((W^{(1)})^\top x + W_0^{(1)})$
- 2nd layer: $A^{(2)} = f^{(2)}((W^{(2)})^\top A^{(1)} + W_0^{(2)})$
- Whole thing: $A^{(2)} = \mathrm{NN}(x; W, W_0)$

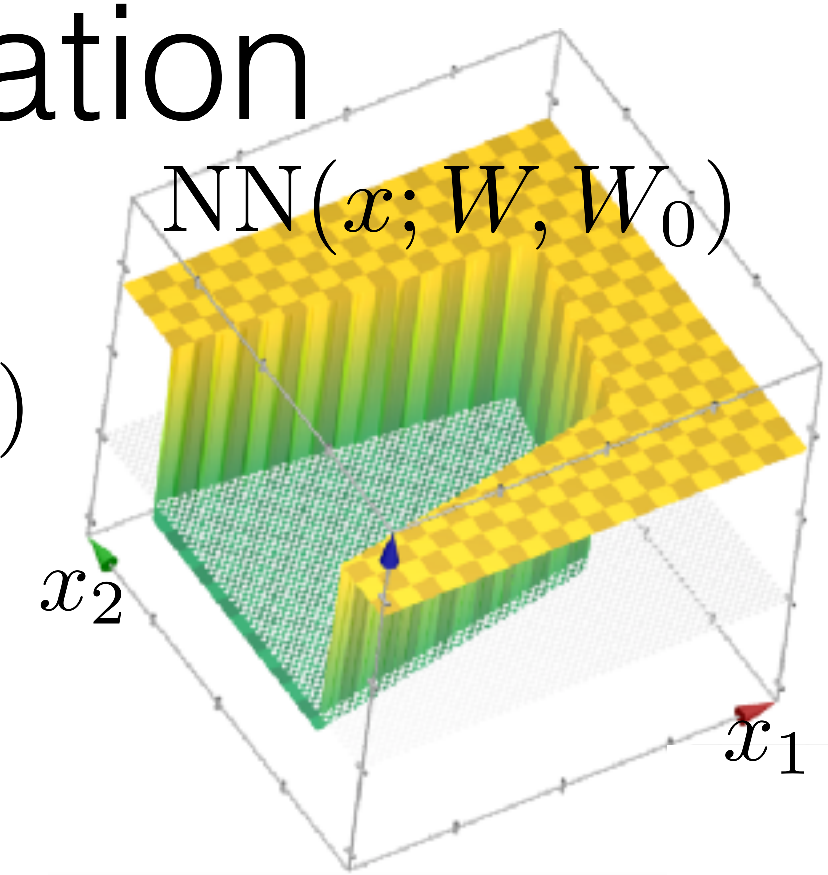


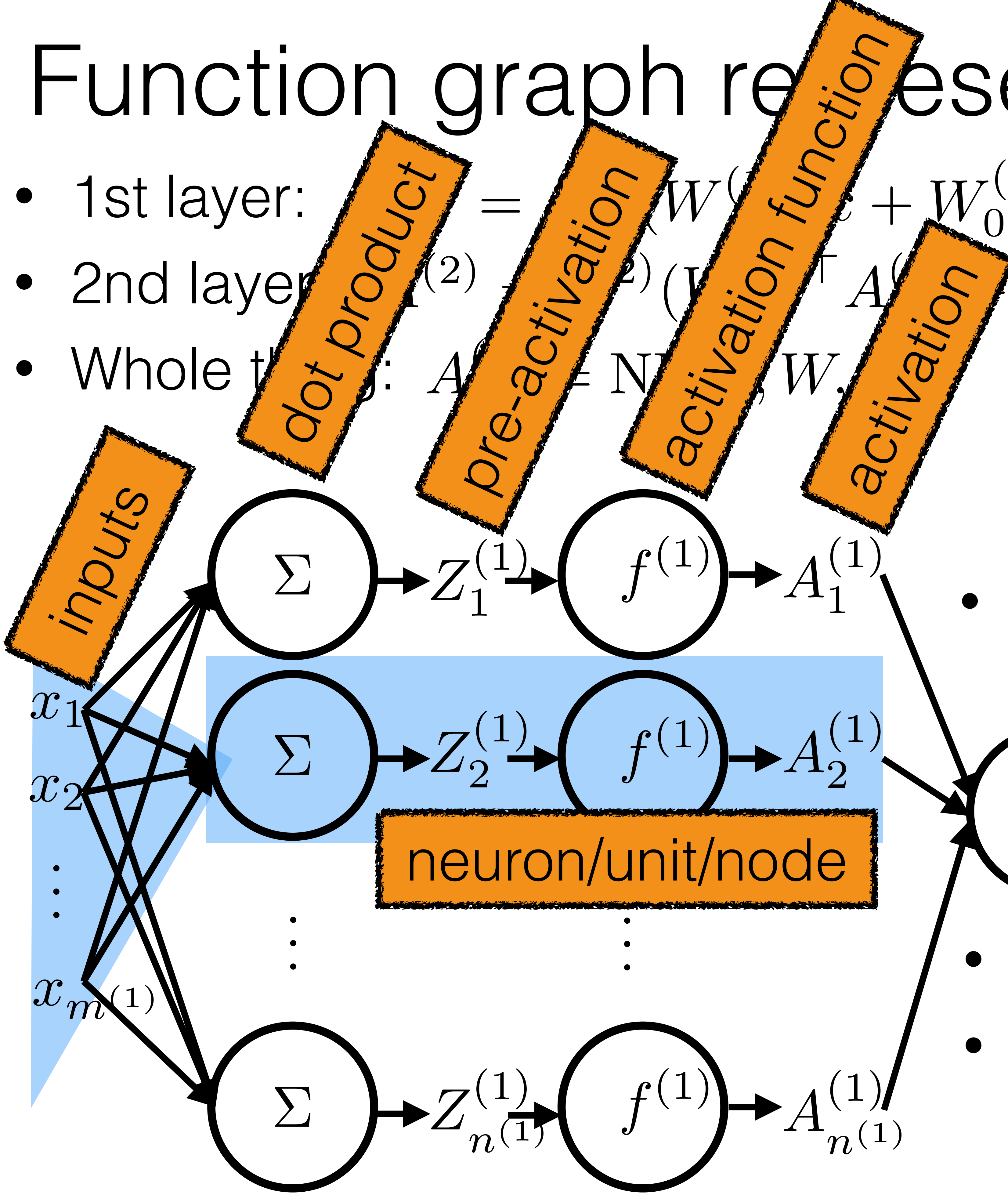


- Circle: function evaluation
- Forward vs. backward
- A feed-forward neural network

Function graph representation

- 1st layer: $A^{(1)} = f^{(1)}(W^{(1)\top} x + W_0^{(1)})$
- 2nd layer: $A^{(2)} = f^{(2)}(W^{(2)\top} A^{(1)} + W_0^{(2)})$
- Whole thing: $A^{(2)} = \mathrm{NN}(x; W, W_0)$

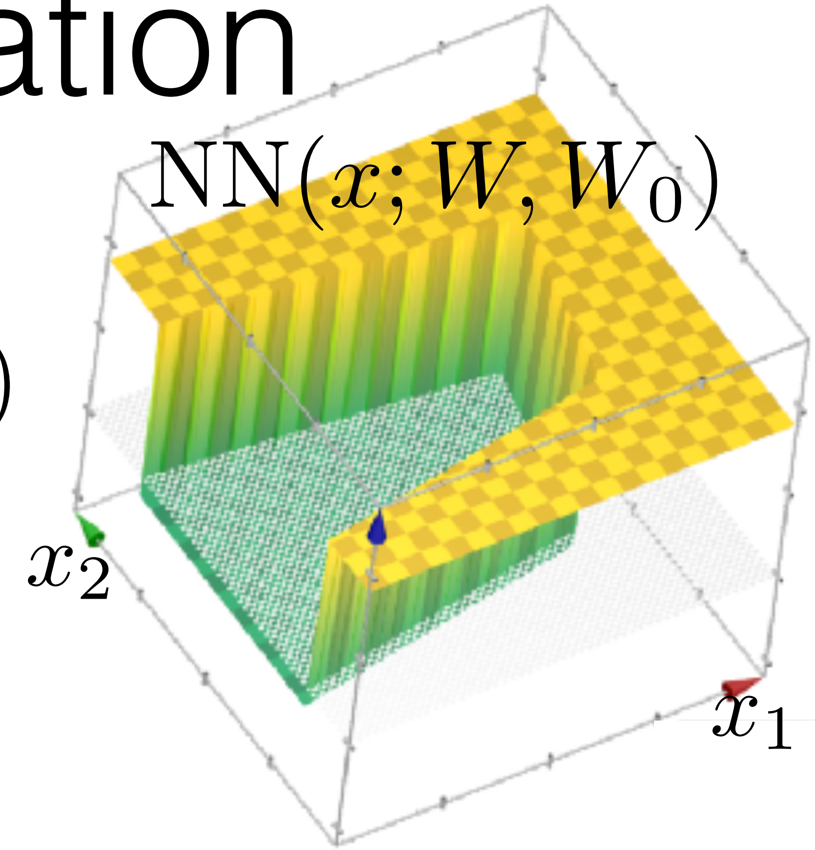


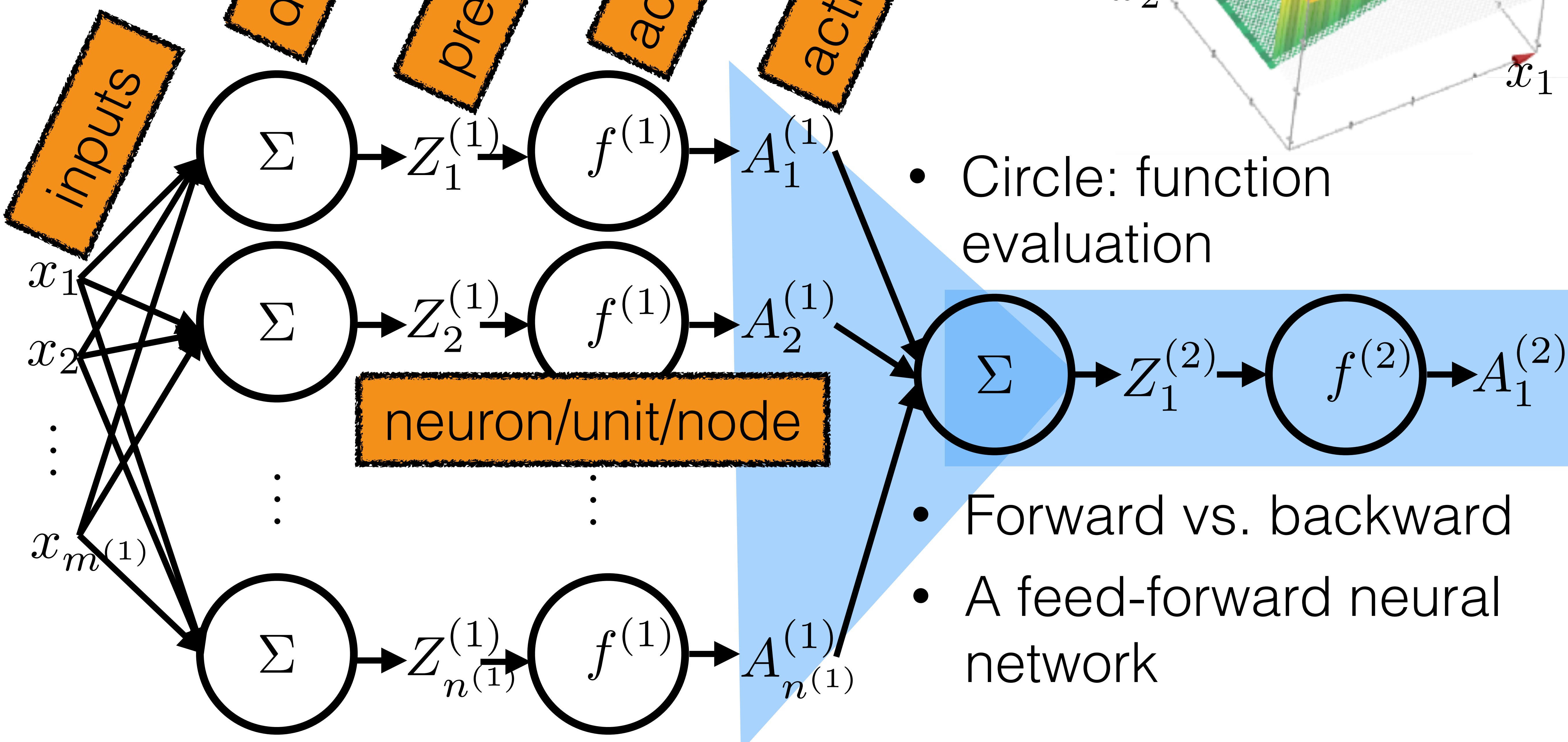


- Circle: function evaluation
- Forward vs. backward
- A feed-forward neural network

Function graph representation

- 1st layer: $= \ldots (W^{(1} \ldots + W_0^{(1)})$
- 2nd layer: $\ldots^{(2)} \ldots^{2)}(\ldots^{\top} A^{(} \ldots W_0^{(2)})$
- Whole thing: $A^{(} \ldots = \mathrm{N} \ldots, W, \ldots$

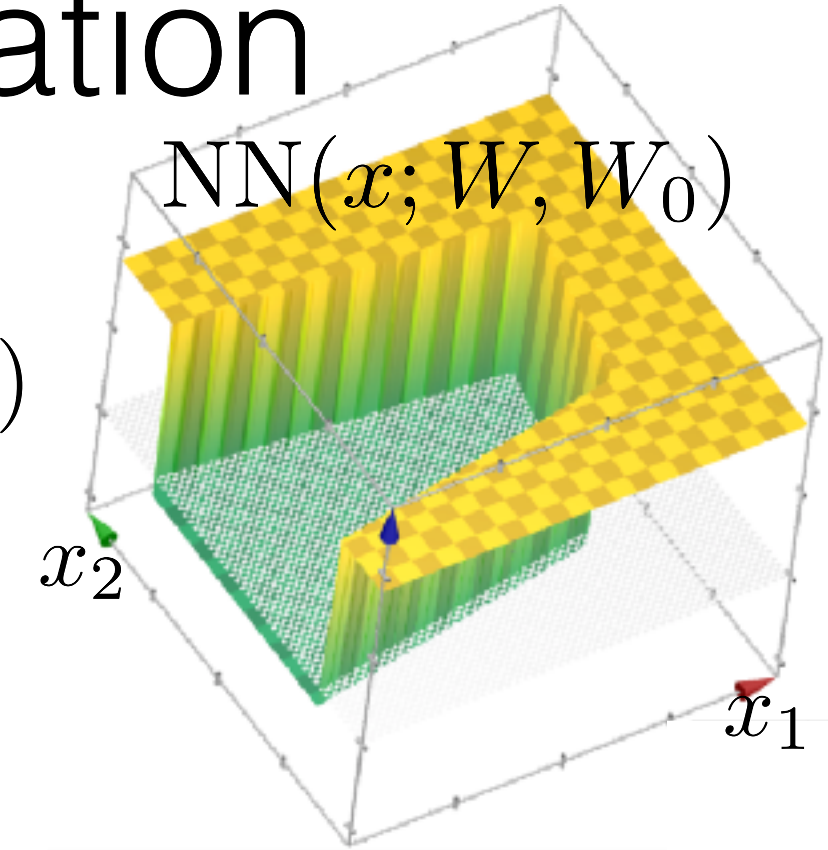


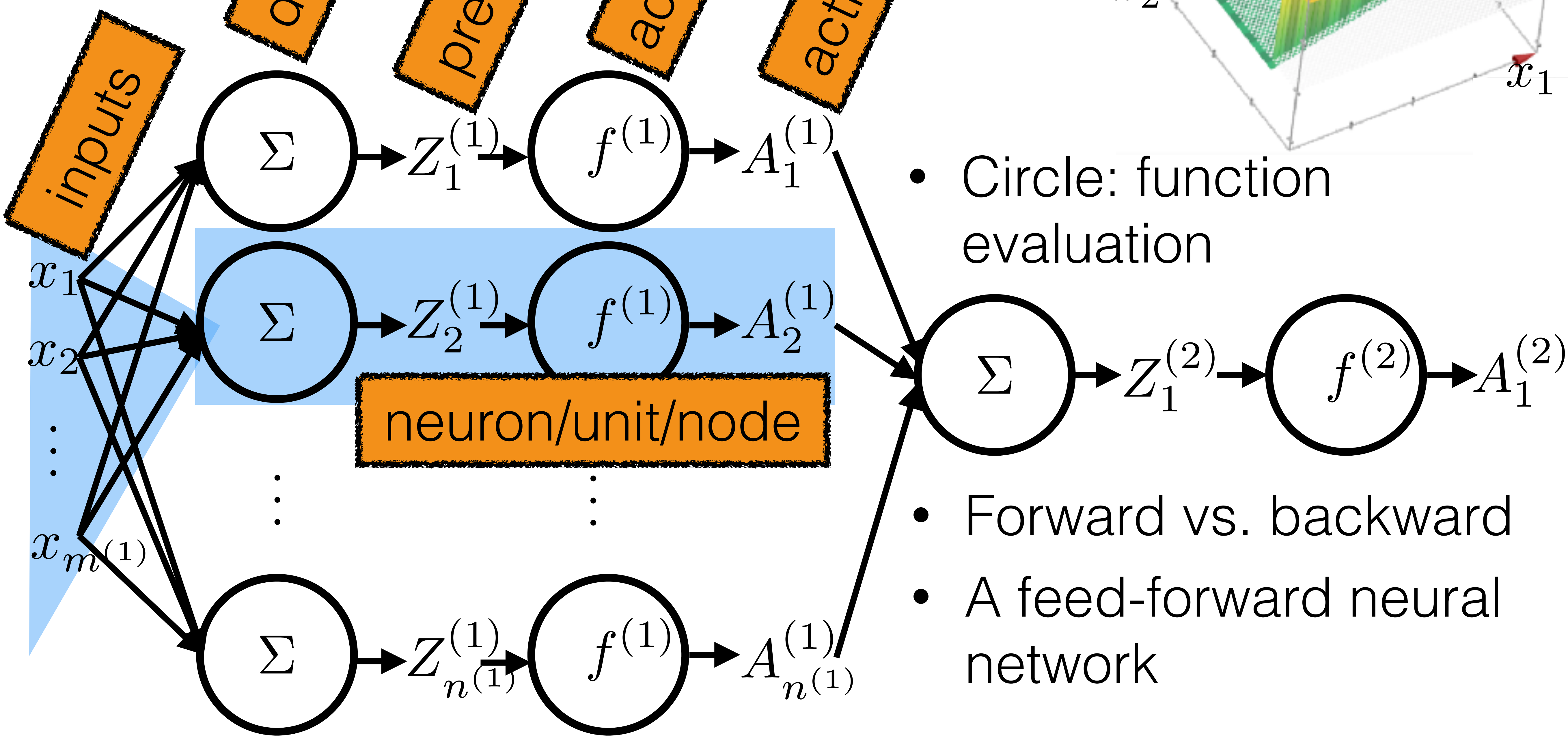


- Circle: function evaluation
- Forward vs. backward
- A feed-forward neural network

Function graph representation

- 1st layer: $A^{(1)} = f^{(1)}(W^{(1)\top} x + W_0^{(1)})$
- 2nd layer: $A^{(2)} = f^{(2)}(W^{(2)\top} A^{(1)} + W_0^{(2)})$
- Whole thing: $A^{(2)} = \mathrm{NN}(x; W, W_0)$

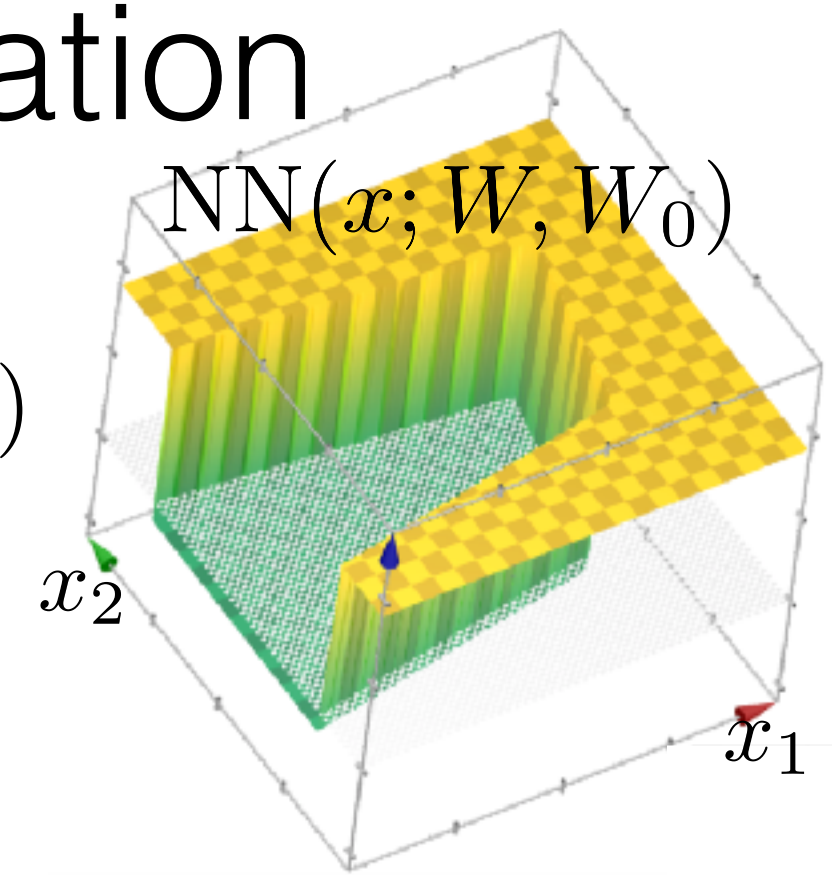


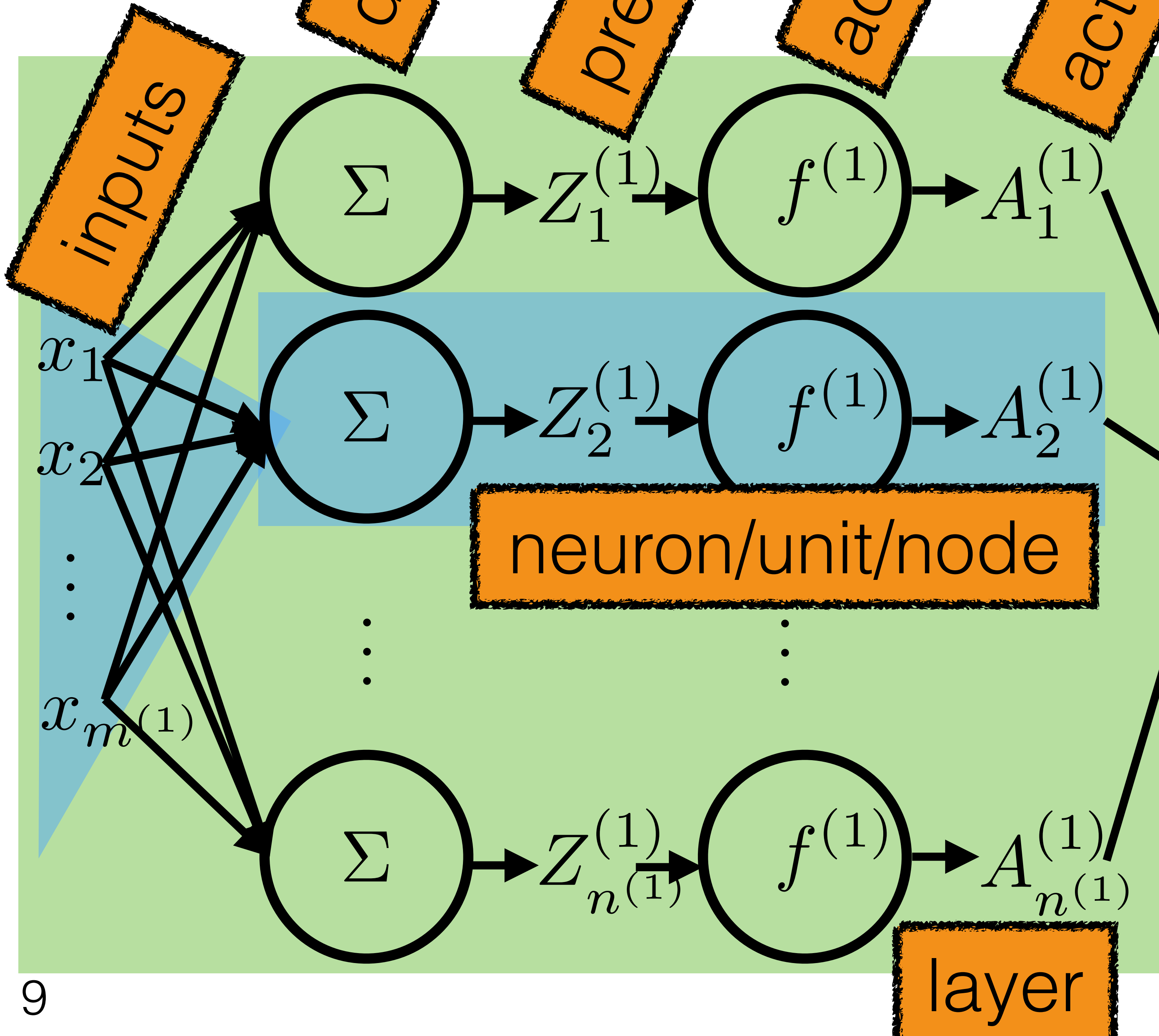


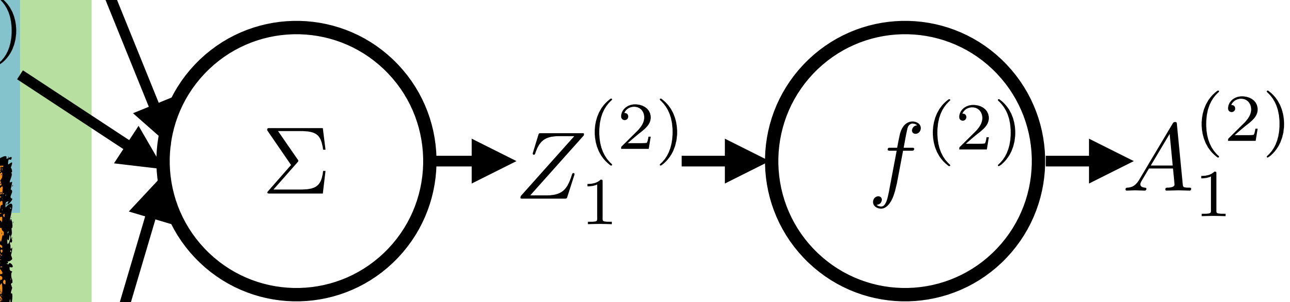
- Circle: function evaluation
- Forward vs. backward
- A feed-forward neural network

Function graph representation

- 1st layer: $= (W^{(1)\top}x + W_0^{(1)})$
- 2nd layer: $A^{(2)} = f^{(2)}(W^{(2)\top}A^{(1)} + W_0^{(2)})$
- Whole thing: $A^{(2)} = \mathrm{NN}(x; W, W_0)$



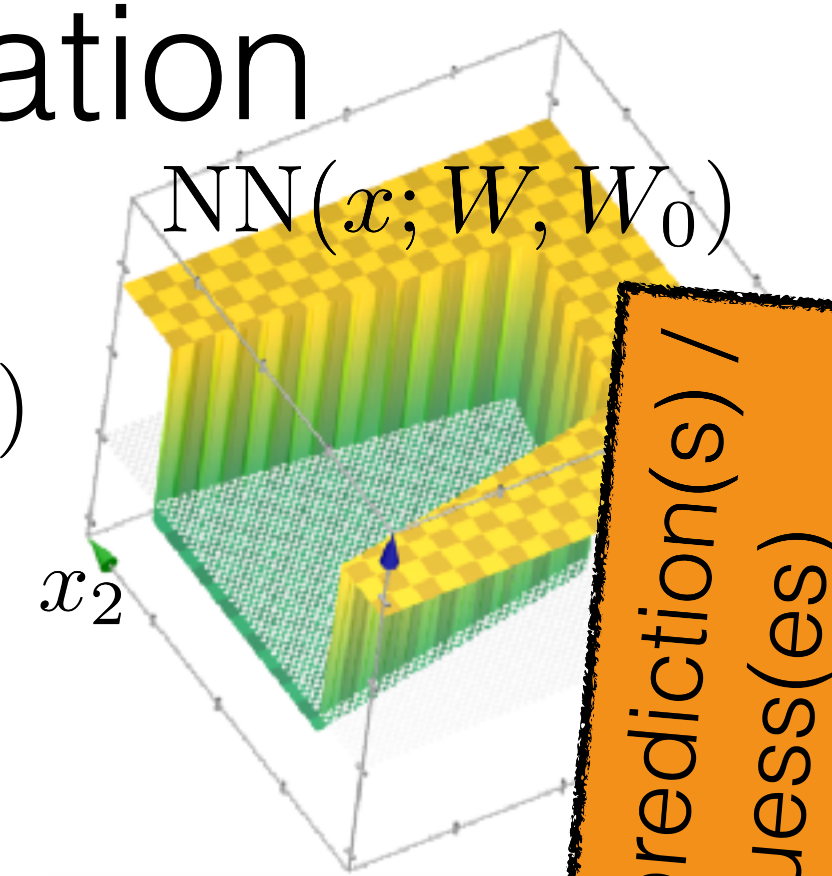


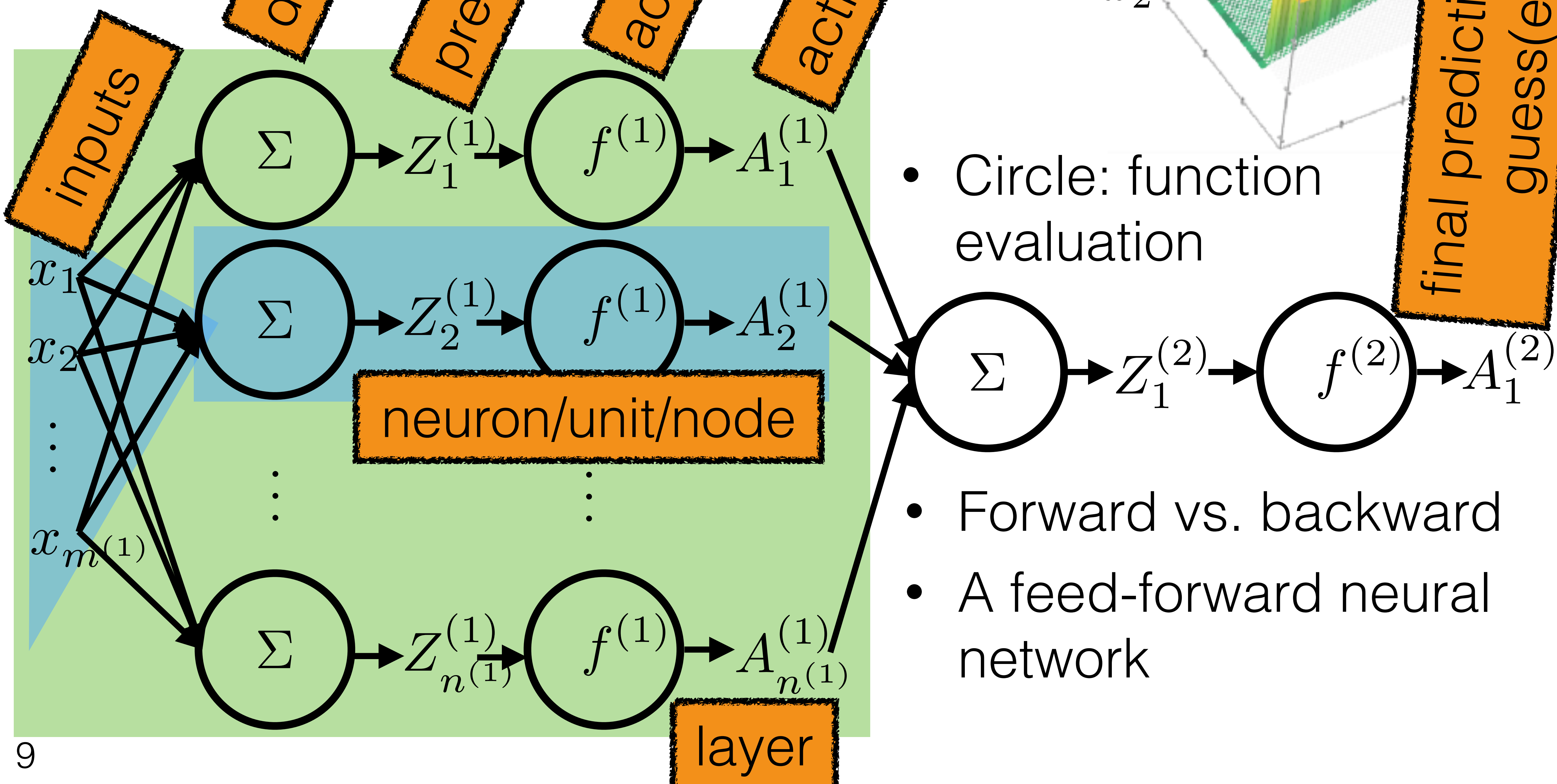


- Circle: function evaluation
- Forward vs. backward
- A feed-forward neural network

Function graph representation

- 1st layer: $A^{(1)} = f^{(1)}(W^{(1)\top} x + W_0^{(1)})$
- 2nd layer: $A^{(2)} = f^{(2)}(W^{(2)\top} A^{(1)} + W_0^{(2)})$
- Whole thing: $A^{(2)} = \mathrm{NN}(x; W, W_0)$

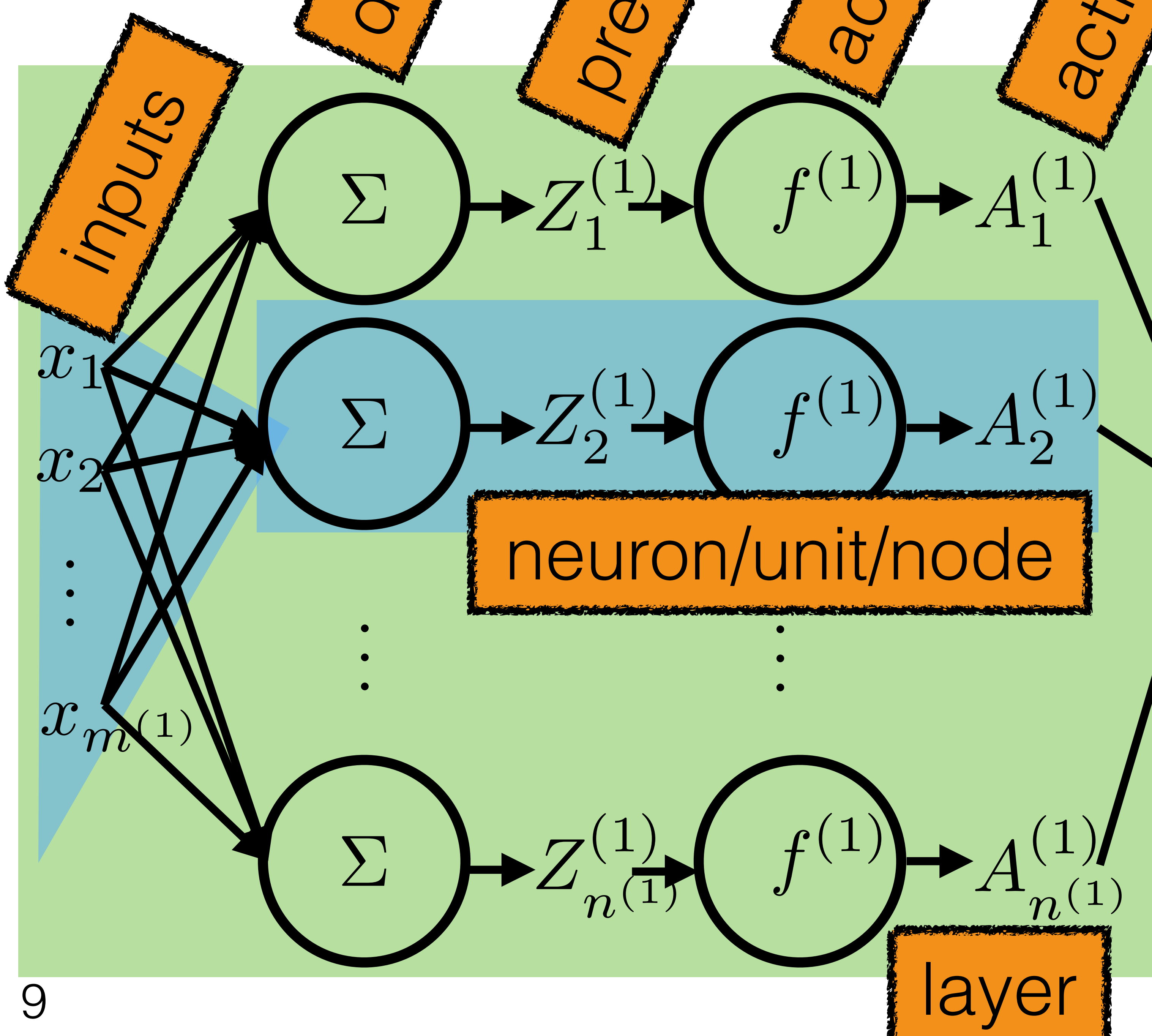


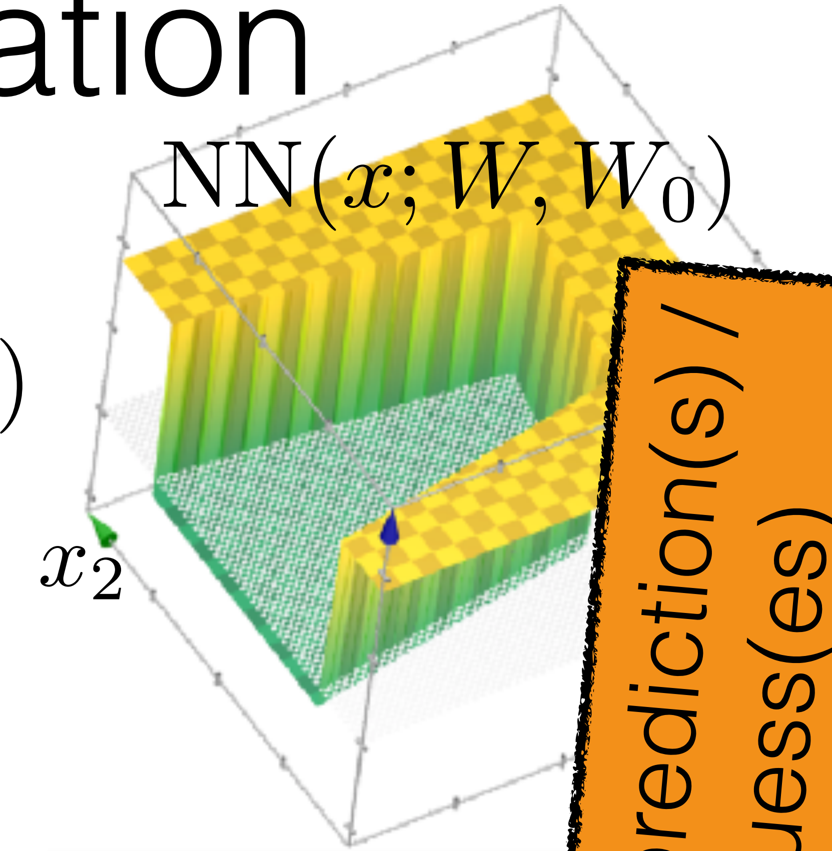


- Circle: function evaluation
- Forward vs. backward
- A feed-forward neural network

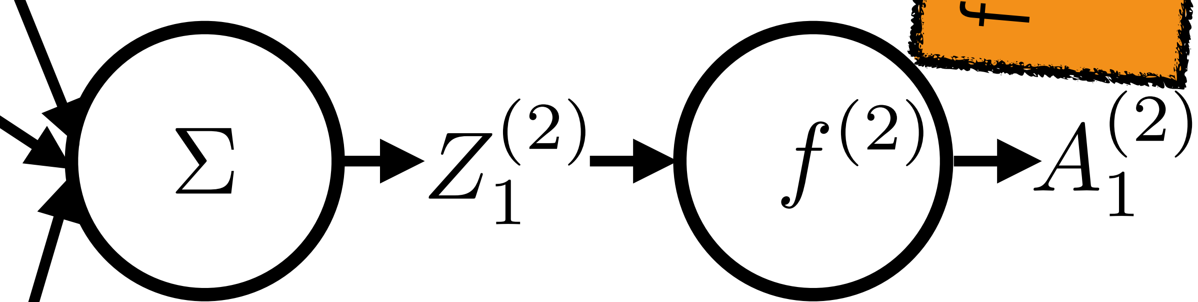
Function graph representation

- 1st layer: $\ldots = \ldots(W^{(1)\ldots} \ldots + W_0^{(1)})$
- 2nd layer: $\ldots^{(2)} = \ldots^{(2)}(W^{\ldots\top} A^{\ldots} + W_0^{(2)})$
- Whole thing: $A^{\ldots} = \mathrm{N}\ldots, W, \ldots$





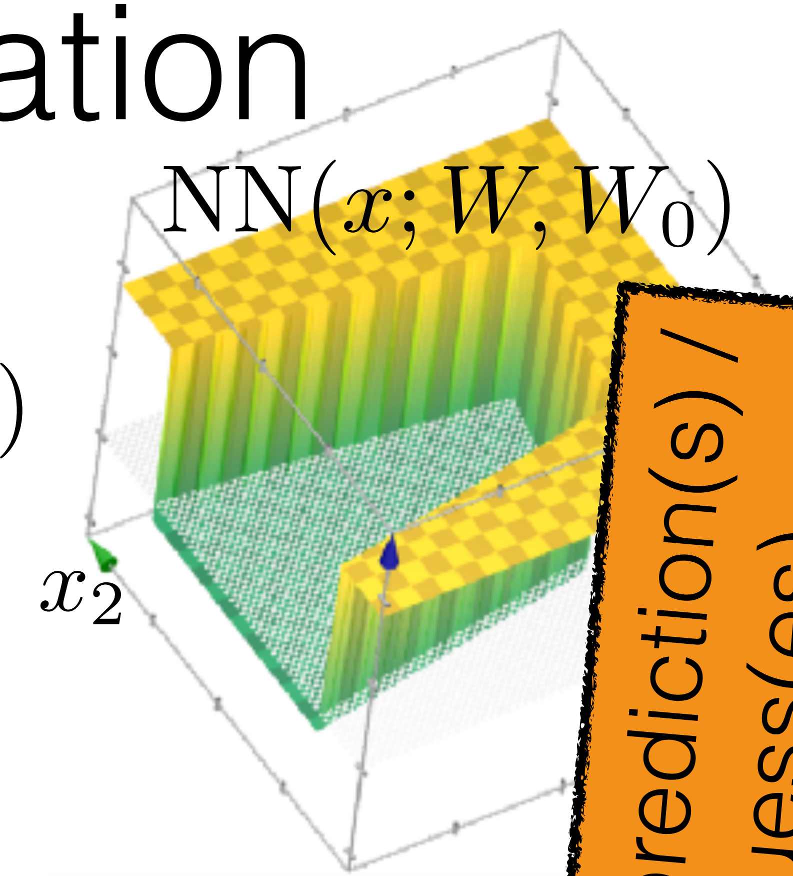
- Circle: function evaluation

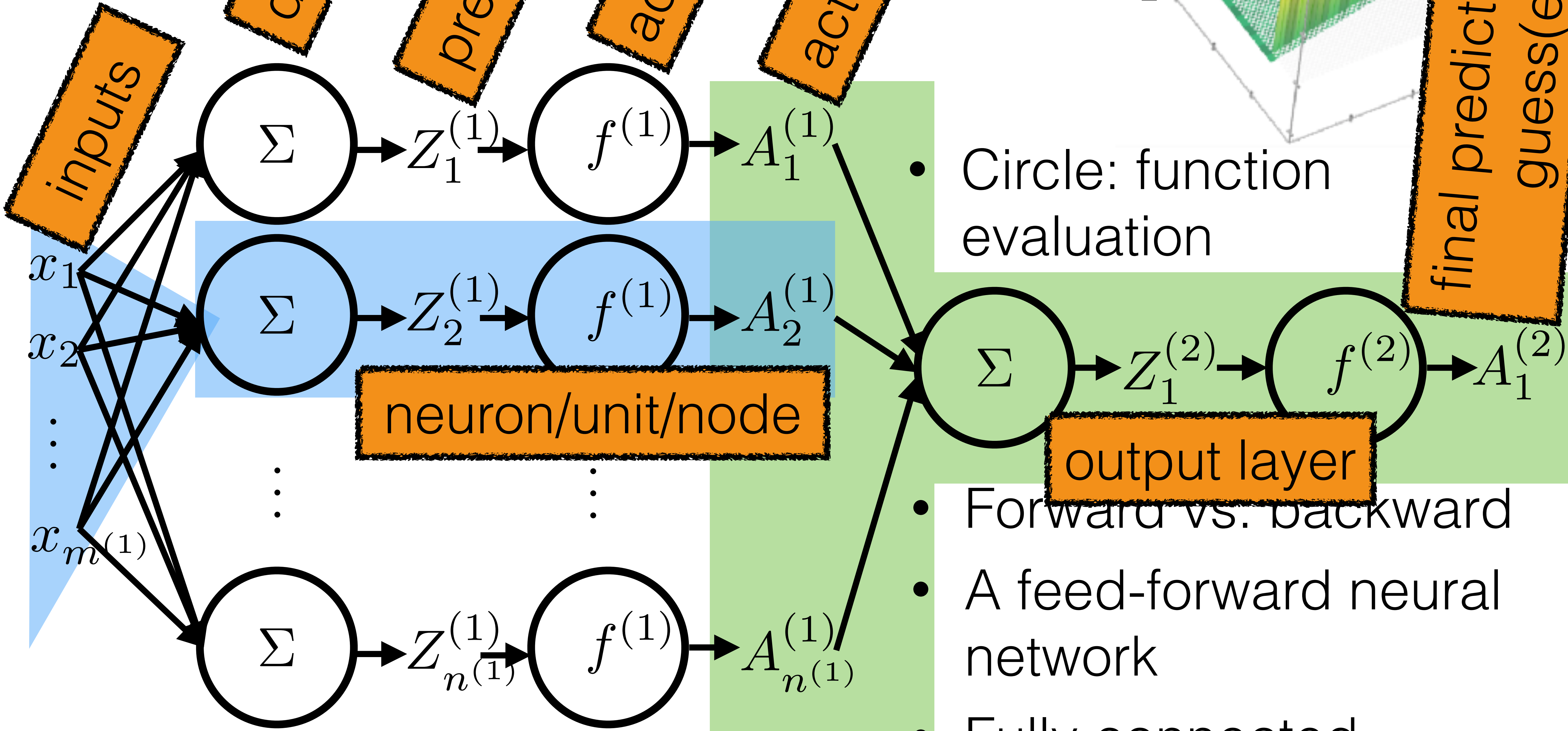


- Forward vs. backward
- A feed-forward neural network
- Fully connected

Function graph representation

- 1st layer: $= \ldots(W^{(1)\ldots} \ldots + W_0^{(1)})$
- 2nd layer: $\ldots^{(2)} = \ldots^{(2)}(\ldots^{\top} A^{(\ldots} \ldots W_0^{(2)})$
- Whole thing: $A^{(\ldots} = \mathrm{N}\ldots, W, \ldots$

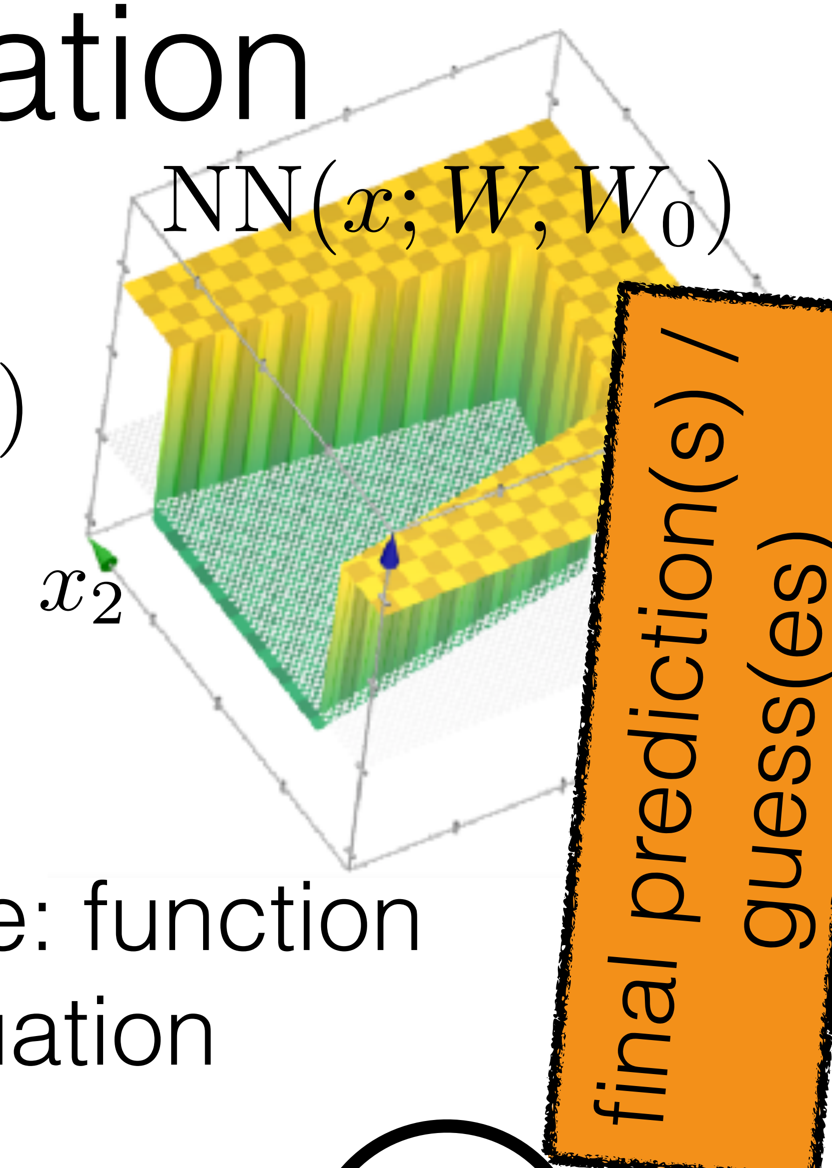


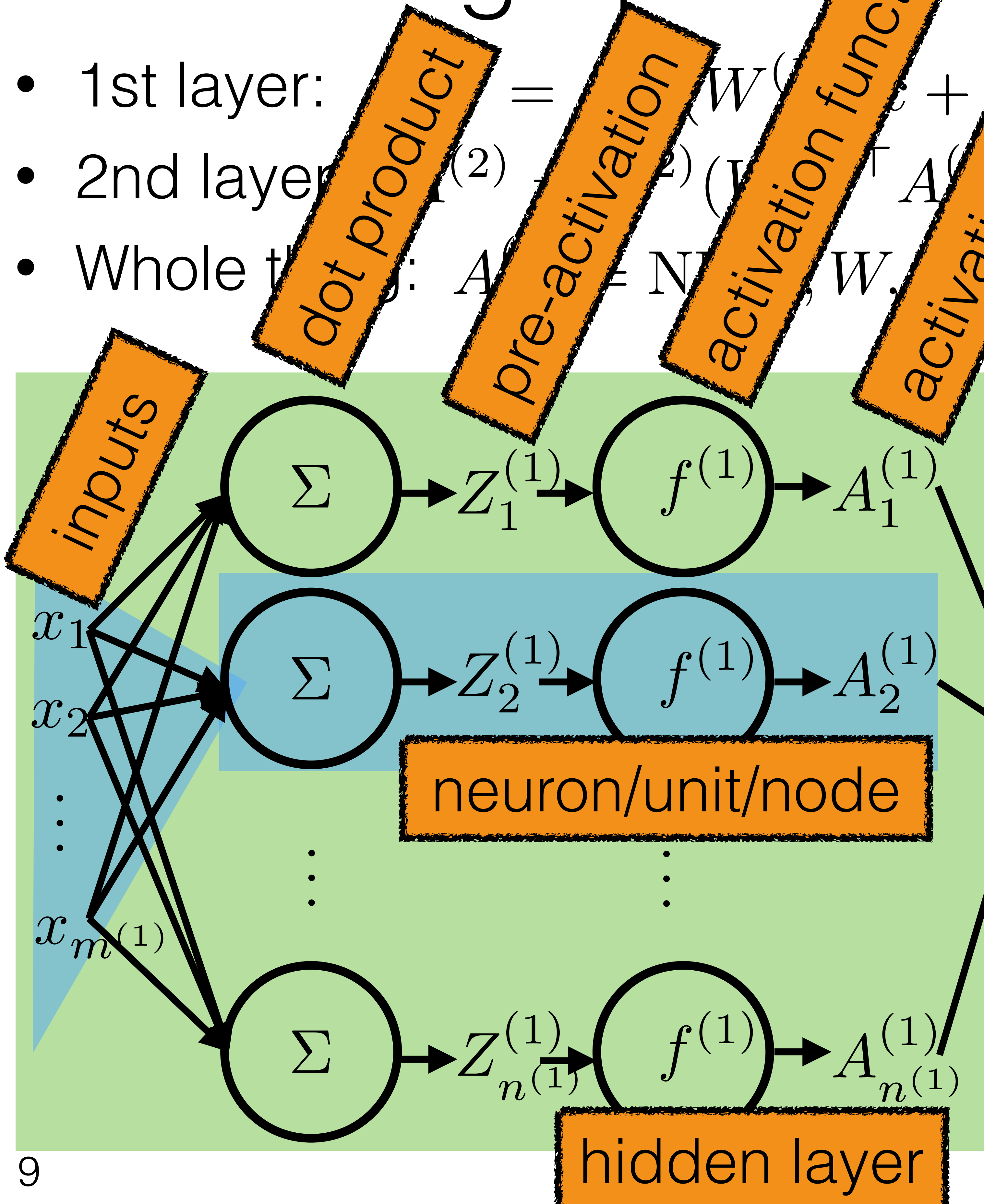


- Circle: function evaluation
- Forward vs. backward
- A feed-forward neural network
- Fully connected

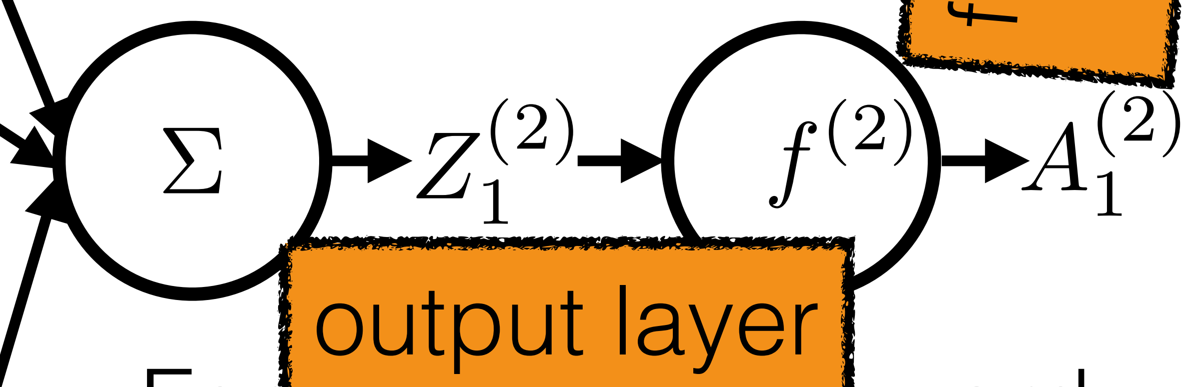
Function graph representation

- 1st layer: $A^{(1)} = f^{(1)}(W^{(1)\top}x + W_0^{(1)})$
- 2nd layer: $A^{(2)} = f^{(2)}(W^{(2)\top}A^{(1)} + W_0^{(2)})$
- Whole thing: $A^{(2)} = \mathrm{NN}(x; W, W_0)$





- Circle: function evaluation



- Forward vs. backward
- A feed-forward neural network
- Fully connected

Problem setup

Problem setup

1. Choose a hypothesis class.

Problem setup

1. Choose a hypothesis class. E.g.,

$$h(x; W, W_0) = \mathrm{NN}(x; W, W_0)$$

Problem setup

1. Choose a hypothesis class. E.g.,

$$h(x; W, W_0) = \mathrm{NN}(x; W, W_0)$$

- 1st layer: $A^{(1)} = f^{(1)}(W^{(1)\top} x + W_0^{(1)})$
- 2nd layer: $A^{(2)} = f^{(2)}(W^{(2)\top} A^{(1)} + W_0^{(2)})$

Problem setup

1. Choose a hypothesis class. E.g.,

   $$h(x; W, W_0) = \mathrm{NN}(x; W, W_0)$$

   - 1st layer: $A^{(1)} = f^{(1)}(W^{(1)\top} x + W_0^{(1)})$
   - 2nd layer: $A^{(2)} = f^{(2)}(W^{(2)\top} A^{(1)} + W_0^{(2)})$

2. Choose a loss. E.g. for classification: 0-1 loss, asymmetric, negative log likelihood

Problem setup

1. Choose a hypothesis class. E.g.,

$$h(x; W, W_0) = \mathrm{NN}(x; W, W_0)$$

   - 1st layer: $A^{(1)} = f^{(1)}(W^{(1)\top}x + W_0^{(1)})$
   - 2nd layer: $A^{(2)} = f^{(2)}(W^{(2)\top}A^{(1)} + W_0^{(2)})$

2. Choose a loss. E.g. for classification: 0-1 loss, asymmetric, negative log likelihood
3. Learn the parameters. E.g. gradient descent or SGD

Problem setup

1. Choose a hypothesis class. E.g.,

   $$h(x; W, W_0) = \mathrm{NN}(x; W, W_0)$$

   - 1st layer: $A^{(1)} = f^{(1)}(W^{(1)\top} x + W_0^{(1)})$
   - 2nd layer: $A^{(2)} = f^{(2)}(W^{(2)\top} A^{(1)} + W_0^{(2)})$

2. Choose a loss. E.g. for classification: 0-1 loss, asymmetric, negative log likelihood
3. Learn the parameters. E.g. gradient descent or SGD
4. Predict on new data using these parameters

Problem setup

1. Choose a hypothesis class. E.g.,

   $$h(x; W, W_0) = \mathrm{NN}(x; W, W_0)$$

   - 1st layer: $A^{(1)} = f^{(1)}(W^{(1)\top} x + W_0^{(1)})$
   - 2nd layer: $A^{(2)} = f^{(2)}(W^{(2)\top} A^{(1)} + W_0^{(2)})$
2. Choose a loss. E.g. for classification: 0-1 loss, asymmetric, negative log likelihood
3. Learn the parameters. E.g. gradient descent or SGD
4. Predict on new data using these parameters

Issues:

Problem setup

1. Choose a hypothesis class. E.g.,

   $$h(x; W, W_0) = \mathrm{NN}(x; W, W_0)$$

   - 1st layer: $A^{(1)} = f^{(1)}(W^{(1)\top} x + W_0^{(1)})$
   - 2nd layer: $A^{(2)} = f^{(2)}(W^{(2)\top} A^{(1)} + W_0^{(2)})$
2. Choose a loss. E.g. for classification: 0-1 loss, asymmetric, negative log likelihood
3. Learn the parameters. E.g. gradient descent or SGD
4. Predict on new data using these parameters

Issues:

- Derivatives are zero (or undefined) if we use the step function activation, so (S)GD won't do what we want

Problem setup

1. Choose a hypothesis class. E.g.,

   $$h(x; W, W_0) = \mathrm{NN}(x; W, W_0)$$

   - 1st layer: $A^{(1)} = f^{(1)}(W^{(1)\top} x + W_0^{(1)})$
   - 2nd layer: $A^{(2)} = f^{(2)}(W^{(2)\top} A^{(1)} + W_0^{(2)})$

2. Choose a loss. E.g. for classification: 0-1 loss, asymmetric, negative log likelihood
3. Learn the parameters. E.g. gradient descent or SGD
4. Predict on new data using these parameters

Issues:

- Derivatives are zero (or undefined) if we use the step function activation, so (S)GD won't do what we want
- What if I want to do regression?

Problem setup

1. Choose a hypothesis class. E.g.,

$$h(x; W, W_0) = \mathrm{NN}(x; W, W_0)$$

   - 1st layer: $A^{(1)} = f^{(1)}(W^{(1)\top} x + W_0^{(1)})$
   - 2nd layer: $A^{(2)} = f^{(2)}(W^{(2)\top} A^{(1)} + W_0^{(2)})$

2. Choose a loss. E.g. for classification: 0-1 loss, asymmetric, negative log likelihood
3. Learn the parameters. E.g. gradient descent or SGD
4. Predict on new data using these parameters

Issues:

- Derivatives are zero (or undefined) if we use the step function activation, so (S)GD won't do what we want
- What if I want to do regression?
- What if I want to use NLL loss?

Different activation functions

Different activation functions

1. Hypotheses $h(x; W, W_0) = \mathrm{NN}(x; W, W_0)$
   - 1st layer: $A^{(1)} = f^{(1)}(W^{(1)\top} x + W_0^{(1)})$
   - 2nd layer: $A^{(2)} = f^{(2)}(W^{(2)\top} A^{(1)} + W_0^{(2)})$

Different activation functions

1. Hypotheses $h(x; W, W_0) = \mathrm{NN}(x; W, W_0)$
   - 1st layer: $A^{(1)} = f^{(1)}(W^{(1)\top} x + W_0^{(1)})$
   - 2nd layer: $A^{(2)} = f^{(2)}(W^{(2)\top} A^{(1)} + W_0^{(2)})$

- What if I want to do regression?

Different activation functions

1. Hypotheses $h(x; W, W_0) = \mathrm{NN}(x; W, W_0)$
   - 1st layer: $A^{(1)} = f^{(1)}(W^{(1)\top} x + W_0^{(1)})$
   - 2nd layer: $A^{(2)} = f^{(2)}(W^{(2)\top} A^{(1)} + W_0^{(2)})$

- What if I want to do regression? $f^{(2)}(z) = z$

Different activation functions

1. Hypotheses $h(x; W, W_0) = \mathrm{NN}(x; W, W_0)$
   - 1st layer: $A^{(1)} = f^{(1)}(W^{(1)\top} x + W_0^{(1)})$
   - 2nd layer: $A^{(2)} = f^{(2)}(W^{(2)\top} A^{(1)} + W_0^{(2)})$

- What if I want to do regression? $f^{(2)}(z) = z$
- What if I want to use NLL loss?

Different activation functions

1. Hypotheses $h(x; W, W_0) = \mathrm{NN}(x; W, W_0)$
    - 1st layer: $A^{(1)} = f^{(1)}(W^{(1)\top} x + W_0^{(1)})$
    - 2nd layer: $A^{(2)} = f^{(2)}(W^{(2)\top} A^{(1)} + W_0^{(2)})$

- What if I want to do regression? $f^{(2)}(z) = z$
- What if I want to use NLL loss? $f^{(2)}(z) = \sigma(z)$

Different activation functions

1. Hypotheses $h(x; W, W_0) = \mathrm{NN}(x; W, W_0)$
    - 1st layer: $A^{(1)} = f^{(1)}(W^{(1)\top} x + W_0^{(1)})$
    - 2nd layer: $A^{(2)} = f^{(2)}(W^{(2)\top} A^{(1)} + W_0^{(2)})$

- What if I want to do regression? $f^{(2)}(z) = z$
- What if I want to use NLL loss? $f^{(2)}(z) = \sigma(z)$
- Need non-zero derivatives for (S)GD:

Different activation functions

1. Hypotheses $h(x; W, W_0) = \mathrm{NN}(x; W, W_0)$
    - 1st layer: $A^{(1)} = f^{(1)}(W^{(1)\top} x + W_0^{(1)})$
    - 2nd layer: $A^{(2)} = f^{(2)}(W^{(2)\top} A^{(1)} + W_0^{(2)})$

- What if I want to do regression? $f^{(2)}(z) = z$
- What if I want to use NLL loss? $f^{(2)}(z) = \sigma(z)$
- Need non-zero derivatives for (S)GD: Above

Different activation functions

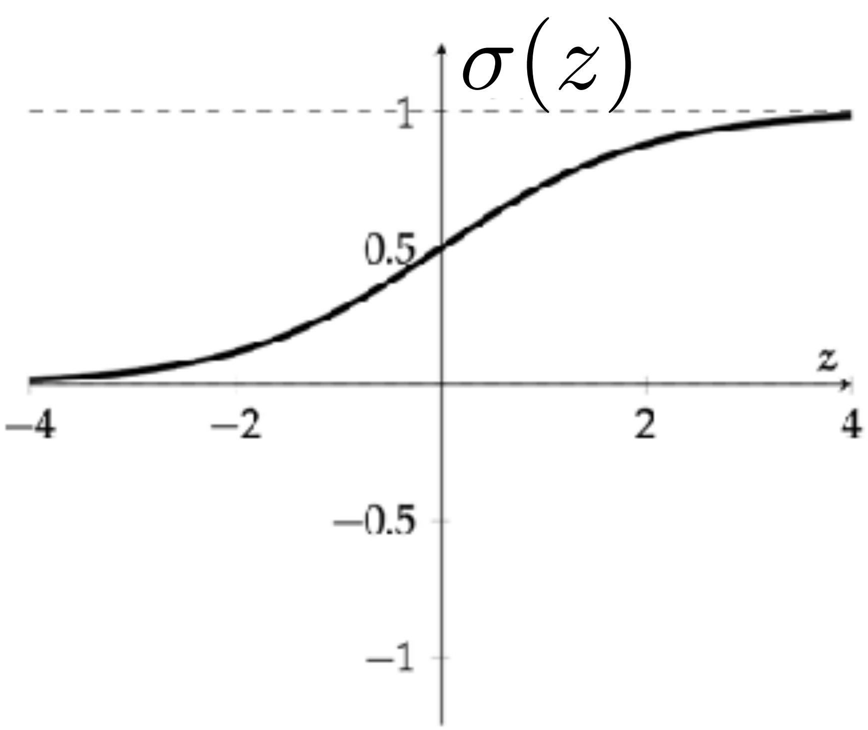
1. Hypotheses $h(x; W, W_0) = \mathrm{NN}(x; W, W_0)$
    - 1st layer: $A^{(1)} = f^{(1)}(W^{(1)\top} x + W_0^{(1)})$
    - 2nd layer: $A^{(2)} = f^{(2)}(W^{(2)\top} A^{(1)} + W_0^{(2)})$

- What if I want to do regression? $f^{(2)}(z) = z$
- What if I want to use NLL loss? $f^{(2)}(z) = \sigma(z)$
- Need non-zero derivatives for (S)GD: Above & $f^{(1)}(z) =$

Different activation functions

1. Hypotheses $h(x; W, W_0) = \mathrm{NN}(x; W, W_0)$
   - 1st layer: $A^{(1)} = f^{(1)}(W^{(1)\top} x + W_0^{(1)})$
   - 2nd layer: $A^{(2)} = f^{(2)}(W^{(2)\top} A^{(1)} + W_0^{(2)})$

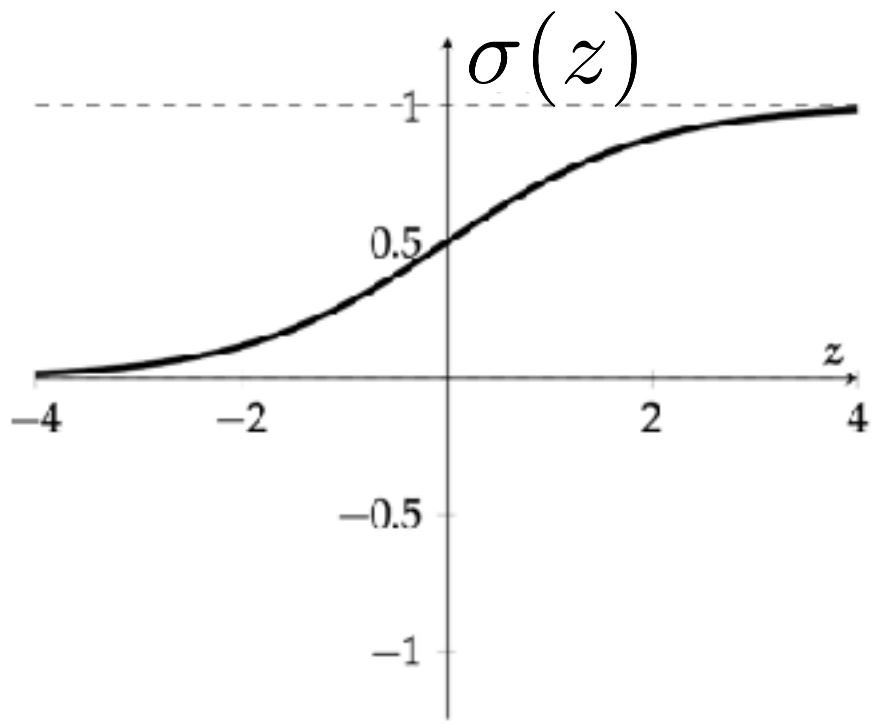
- What if I want to do regression? $f^{(2)}(z) = z$
- What if I want to use NLL loss? $f^{(2)}(z) = \sigma(z)$
- Need non-zero derivatives for (S)GD: Above & $f^{(1)}(z) =$

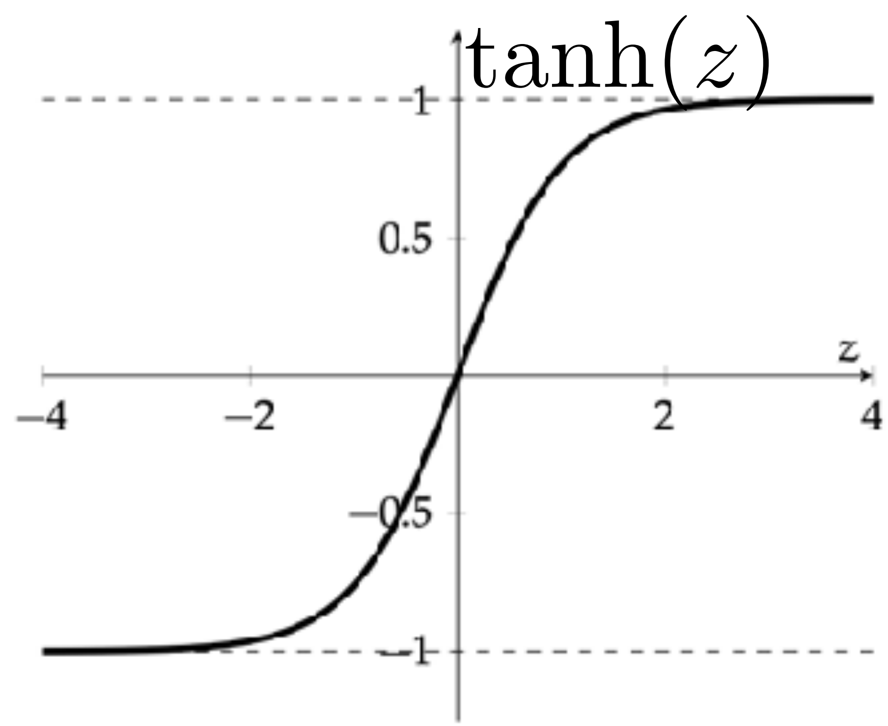


Different activation functions

1. Hypotheses $h(x; W, W_0) = \mathrm{NN}(x; W, W_0)$
    - 1st layer: $A^{(1)} = f^{(1)}(W^{(1)\top} x + W_0^{(1)})$
    - 2nd layer: $A^{(2)} = f^{(2)}(W^{(2)\top} A^{(1)} + W_0^{(2)})$

- What if I want to do regression? $f^{(2)}(z) = z$
- What if I want to use NLL loss? $f^{(2)}(z) = \sigma(z)$
- Need non-zero derivatives for (S)GD: Above & $f^{(1)}(z) =$

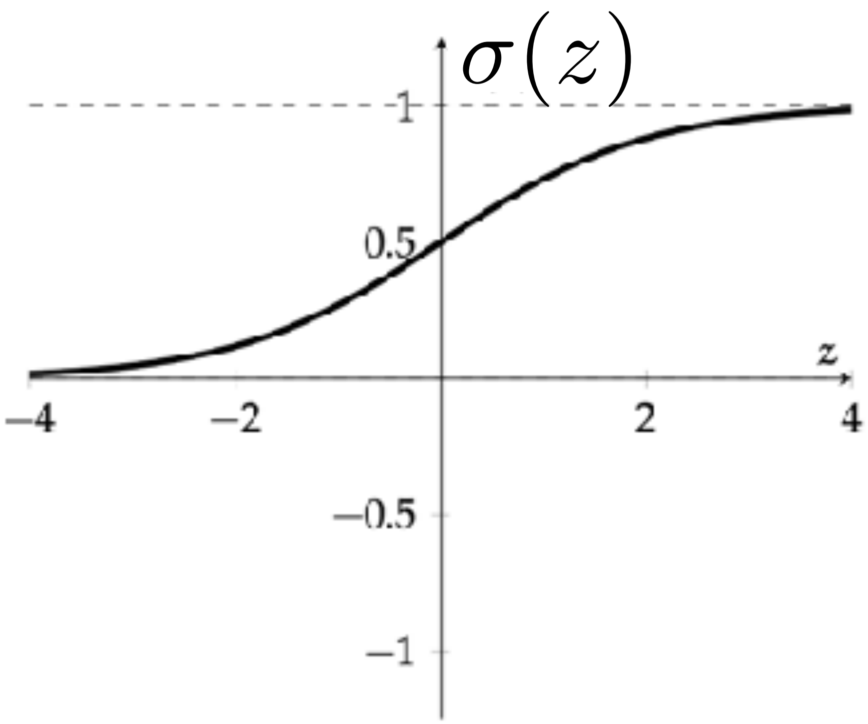


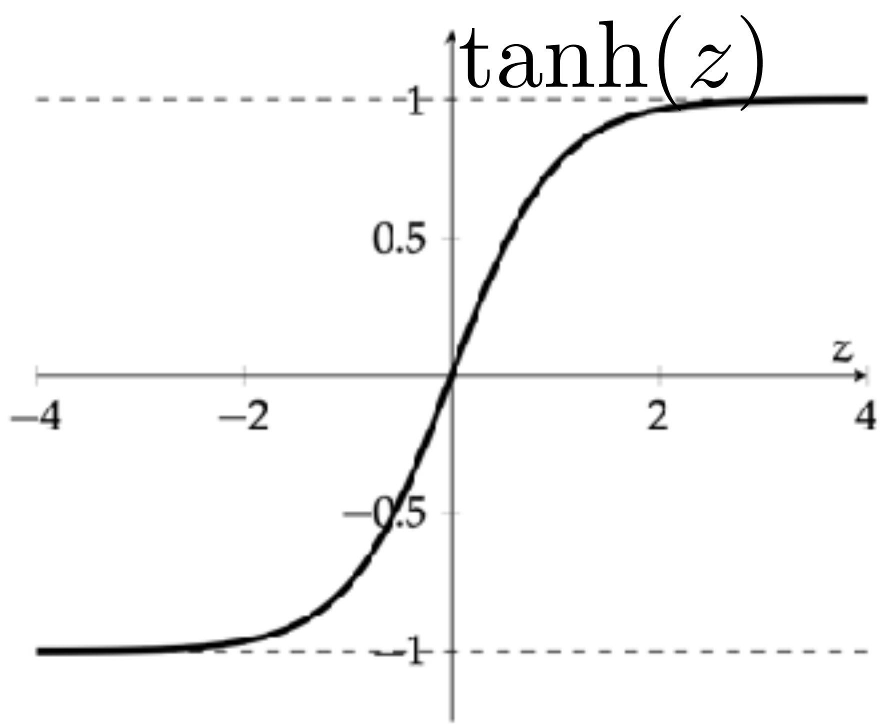


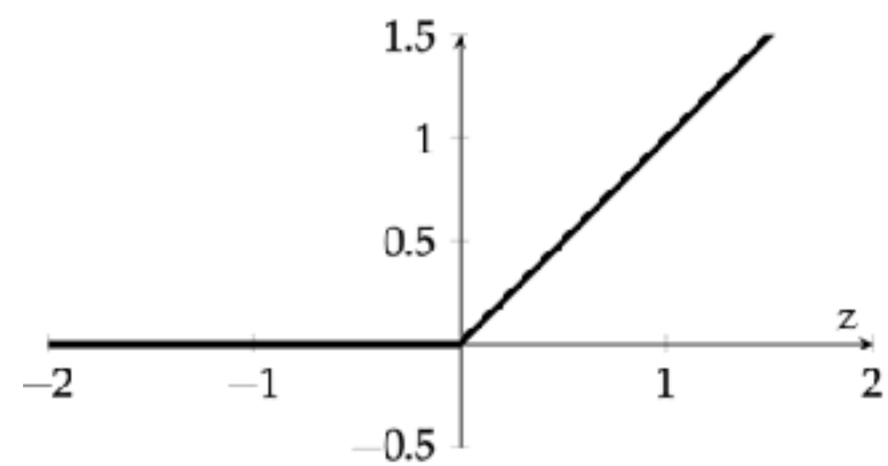
Different activation functions

1. Hypotheses $h(x; W, W_0) = \mathrm{NN}(x; W, W_0)$
    - 1st layer: $A^{(1)} = f^{(1)}(W^{(1)\top} x + W_0^{(1)})$
    - 2nd layer: $A^{(2)} = f^{(2)}(W^{(2)\top} A^{(1)} + W_0^{(2)})$

- What if I want to do regression? $f^{(2)}(z) = z$
- What if I want to use NLL loss? $f^{(2)}(z) = \sigma(z)$
- Need non-zero derivatives for (S)GD: Above & $f^{(1)}(z) =$







$\mathrm{ReLU}(z)$

Choices of activation function

Choices of activation function

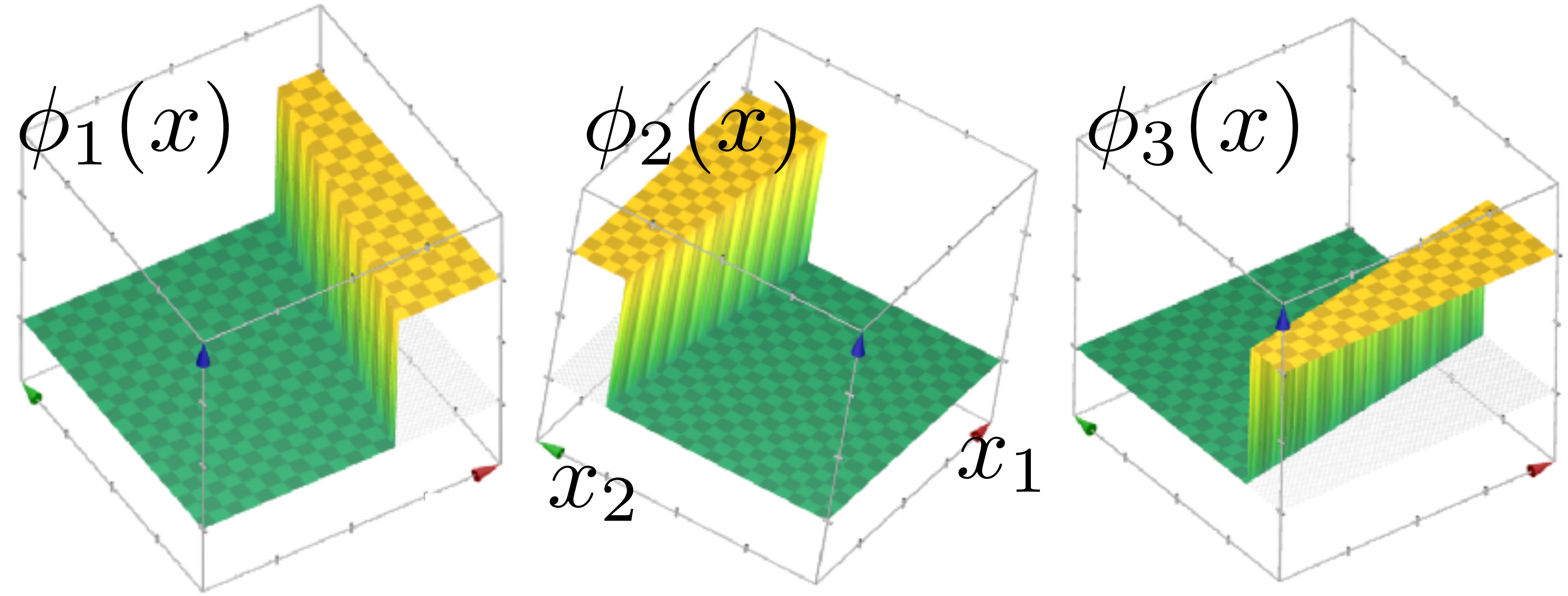
- 1st layer: $A_i^{(1)} = f^{(1)}(w_i^{(1)\top} x + w_0^{(1)})$

Choices of activation function

- 1st layer: $A_i^{(1)} = f^{(1)}(w_i^{(1)\top} x + w_0^{(1)})$
  - Choose
    $f^{(1)}(z) = \mathbf{1}\{z \geq 0\}$

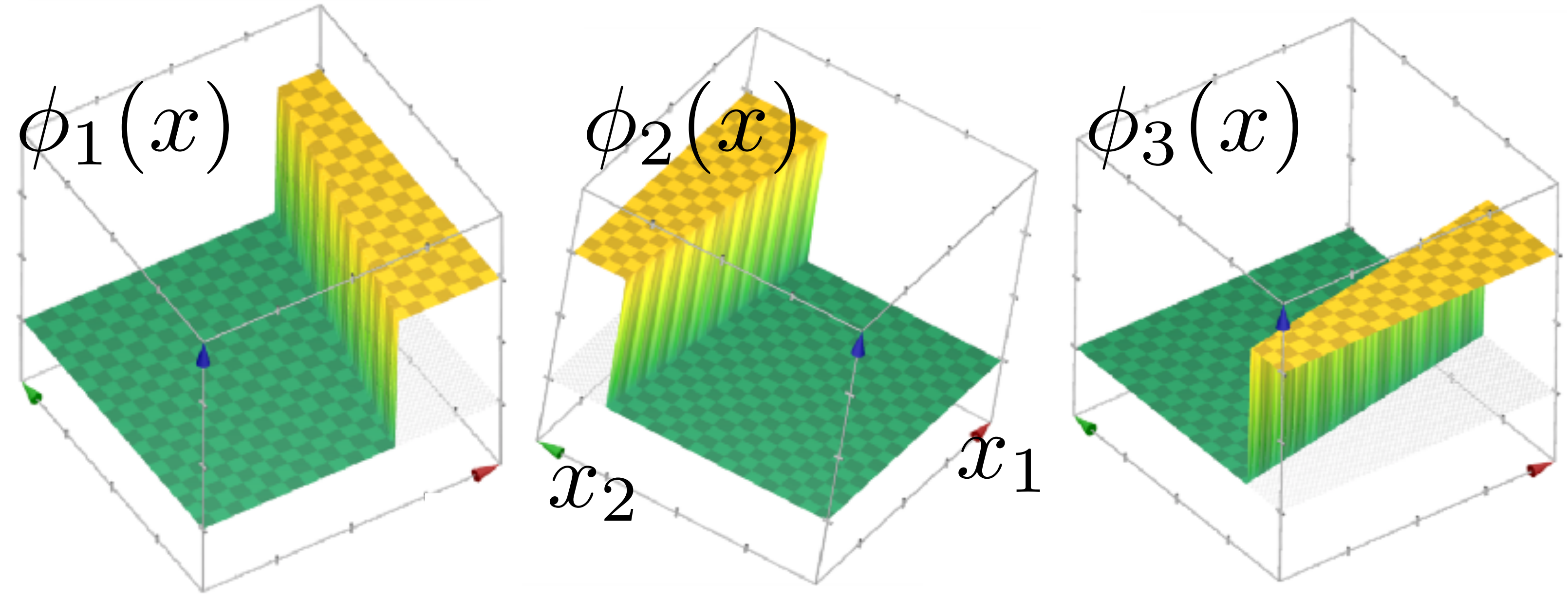
Choices of activation function

- 1st layer: $A_i^{(1)} = f^{(1)}(w_i^{(1)\top} x + w_0^{(1)})$
  - Choose $f^{(1)}(z) = \mathbf{1}\{z \geq 0\}$



Choices of activation function

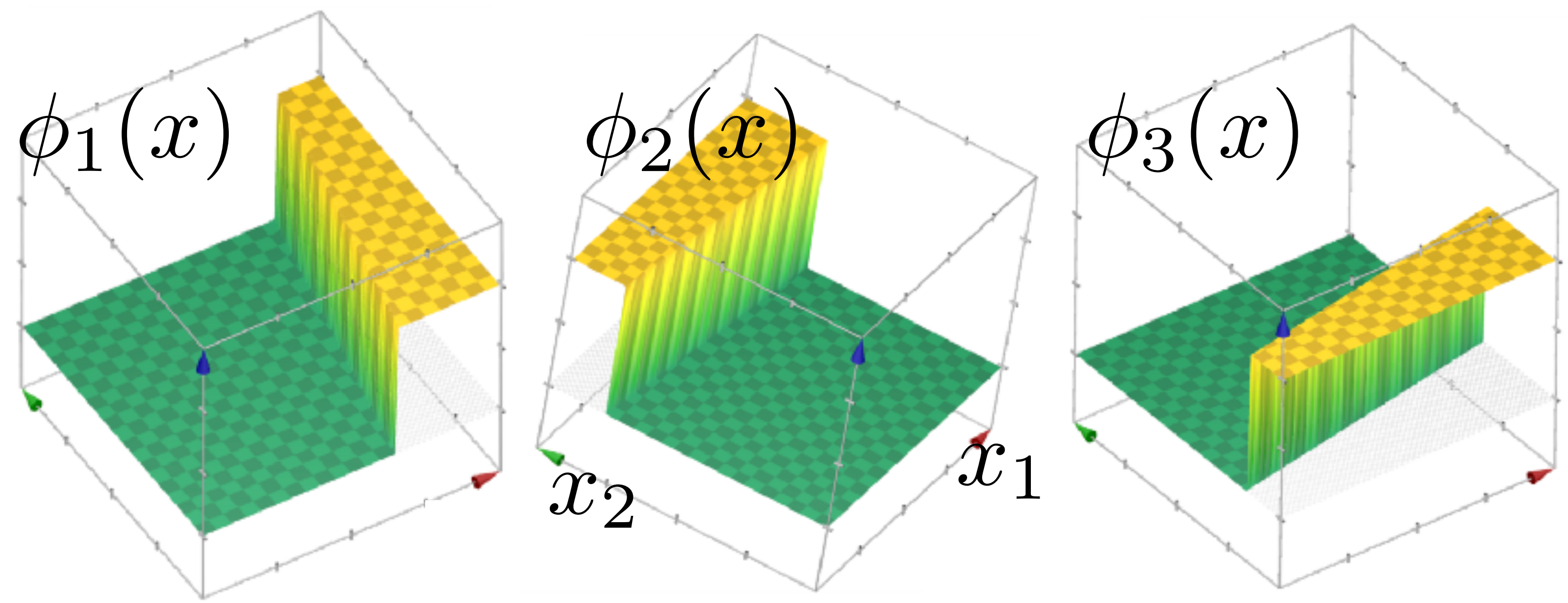
- 1st layer: $A_i^{(1)} = f^{(1)}(w_i^{(1)\top} x + w_0^{(1)})$
  - Choose
    $f^{(1)}(z) = \mathbf{1}\{z \geq 0\}$



- 2nd layer: $A_i^{(2)} = f^{(2)}(w_i^{(2)\top} A^{(1)} + w_0^{(2)})$

Choices of activation function

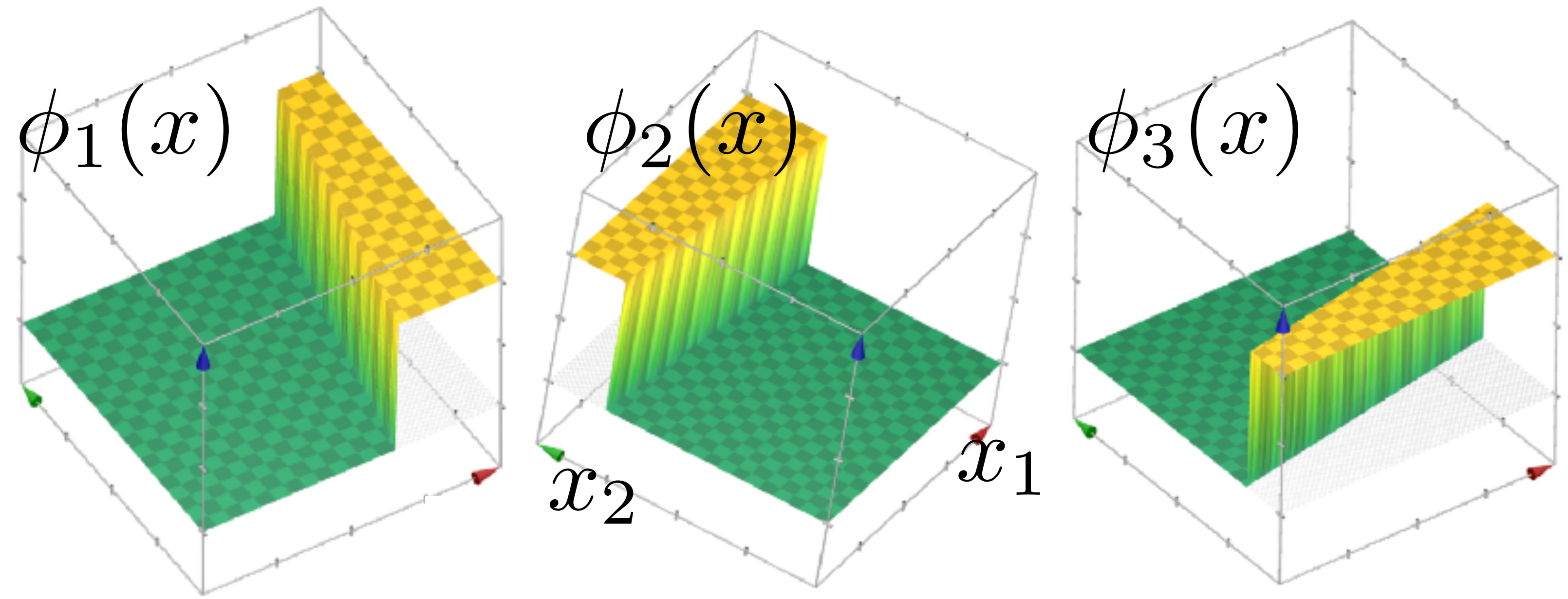
- 1st layer: $A_i^{(1)} = f^{(1)}(w_i^{(1)\top} x + w_0^{(1)})$
  - Choose
    $f^{(1)}(z) = \mathbf{1}\{z \geq 0\}$



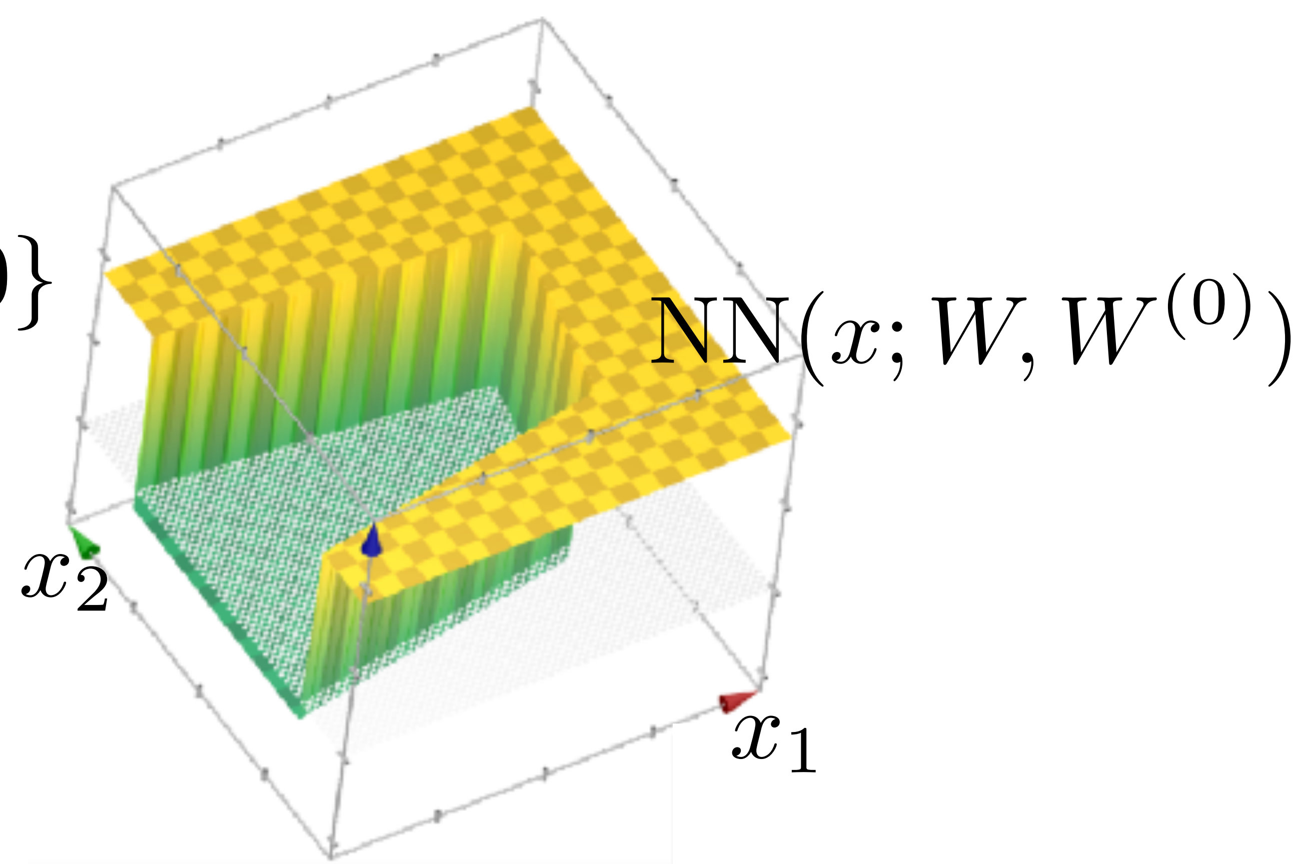
- 2nd layer: $A_i^{(2)} = f^{(2)}(w_i^{(2)\top} A^{(1)} + w_0^{(2)})$
  - Choose
    $f^{(2)}(z) = \mathbf{1}\{z \geq 0\}$

Choices of activation function

- 1st layer: $A_i^{(1)} = f^{(1)}(w_i^{(1)\top} x + w_0^{(1)})$
  - Choose $f^{(1)}(z) = \mathbf{1}\{z \geq 0\}$

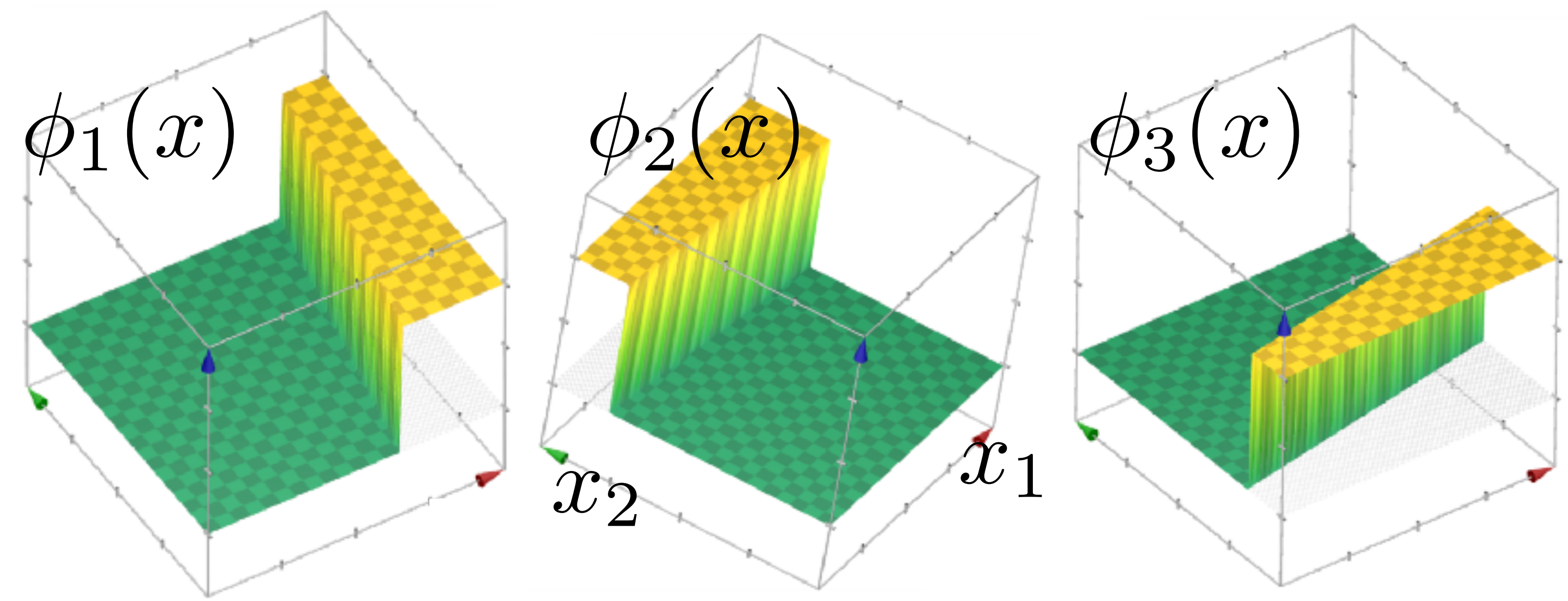


- 2nd layer: $A_i^{(2)} = f^{(2)}(w_i^{(2)\top} A^{(1)} + w_0^{(2)})$
  - Choose $f^{(2)}(z) = \mathbf{1}\{z \geq 0\}$

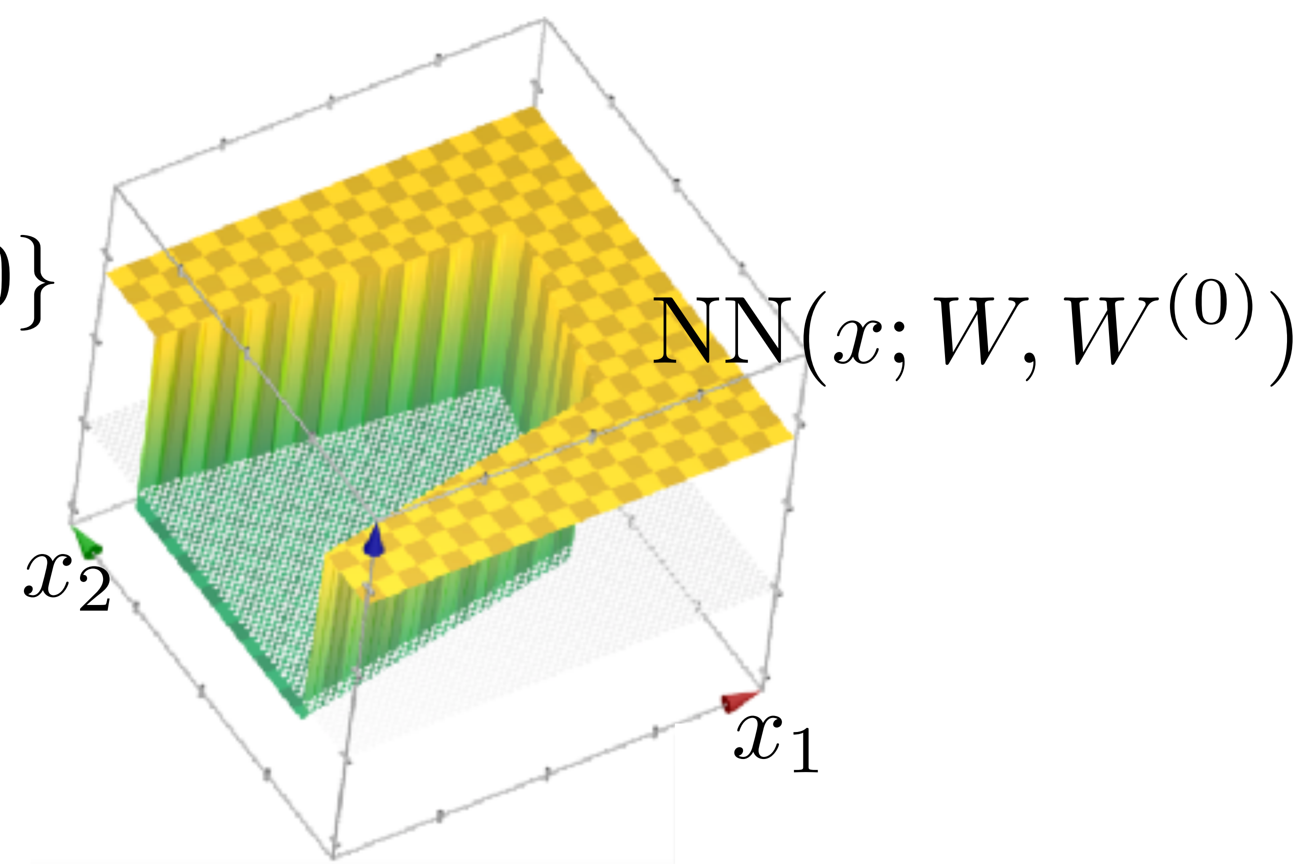


Choices of activation function

- 1st layer: $A_i^{(1)} = f^{(1)}(w_i^{(1)\top} x + w_0^{(1)})$
  - Choose $f^{(1)}(z) = \mathbf{1}\{z \geq 0\}$

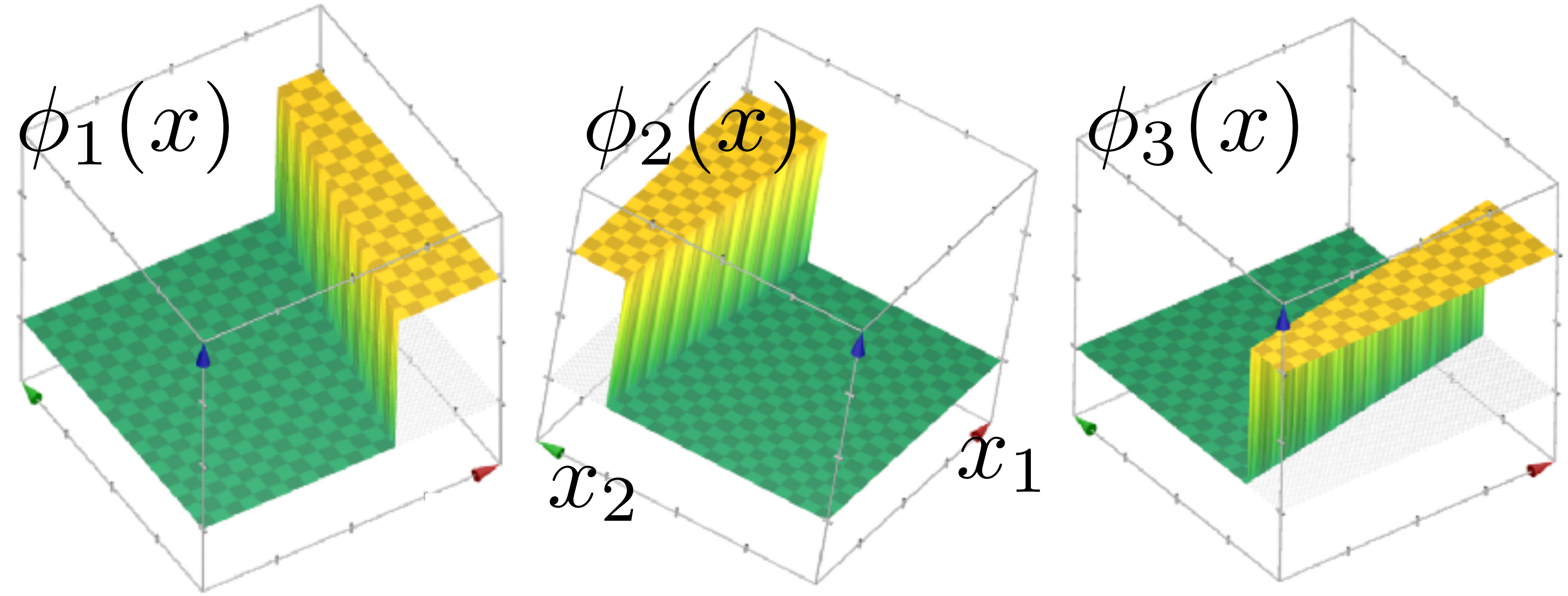


- 2nd layer: $A_i^{(2)} = f^{(2)}(w_i^{(2)\top} A^{(1)} + w_0^{(2)})$
  - Choose $f^{(2)}(z) = \mathbf{1}\{z \geq 0\}$

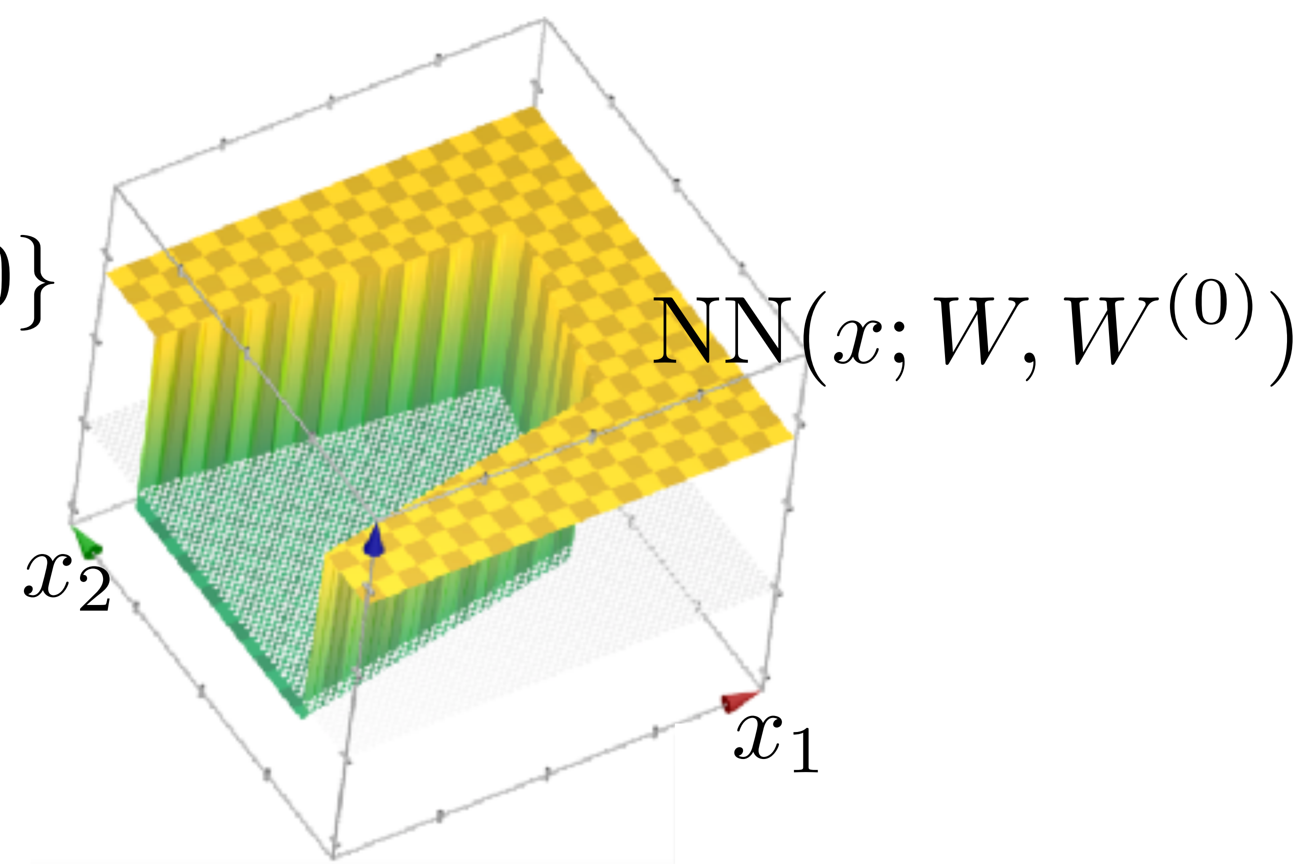


Choices of activation function

- 1st layer: $A_i^{(1)} = f^{(1)}(w_i^{(1)\top} x + w_0^{(1)})$
  - Choose $f^{(1)}(z) = \sigma(z)$

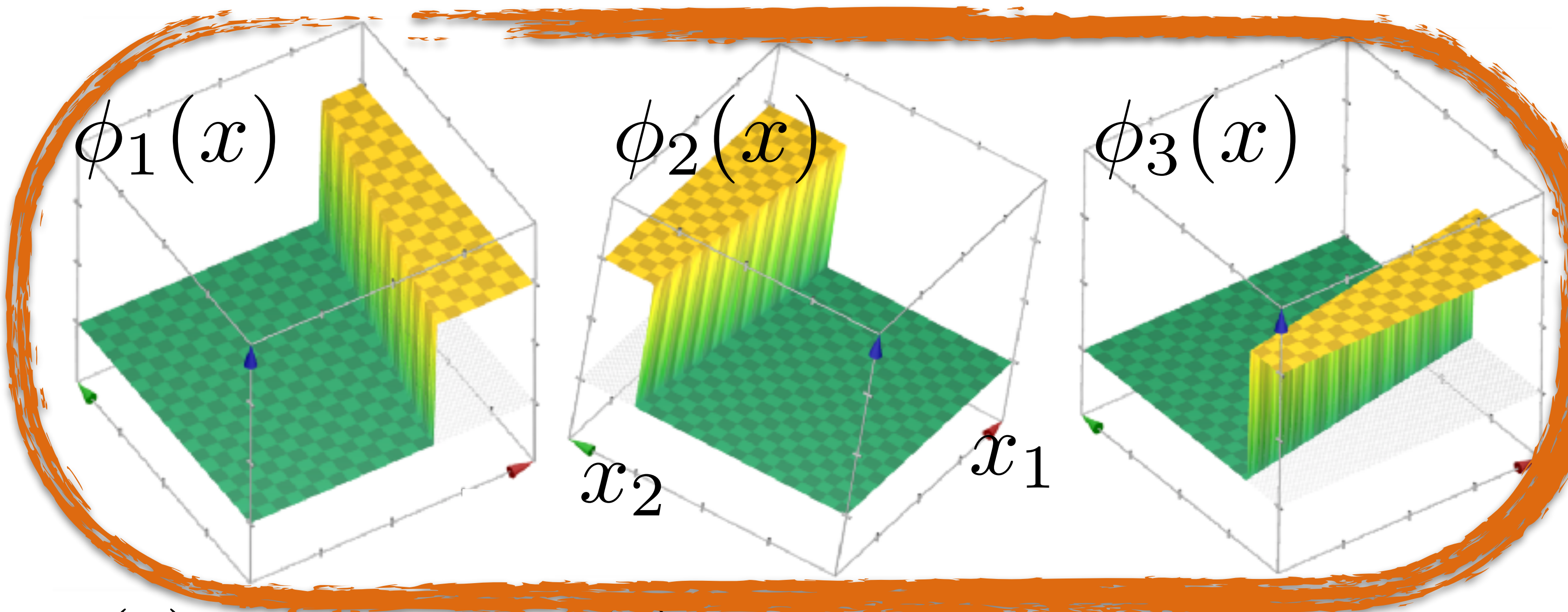


- 2nd layer: $A_i^{(2)} = f^{(2)}(w_i^{(2)\top} A^{(1)} + w_0^{(2)})$
  - Choose $f^{(2)}(z) = \mathbf{1}\{z \geq 0\}$

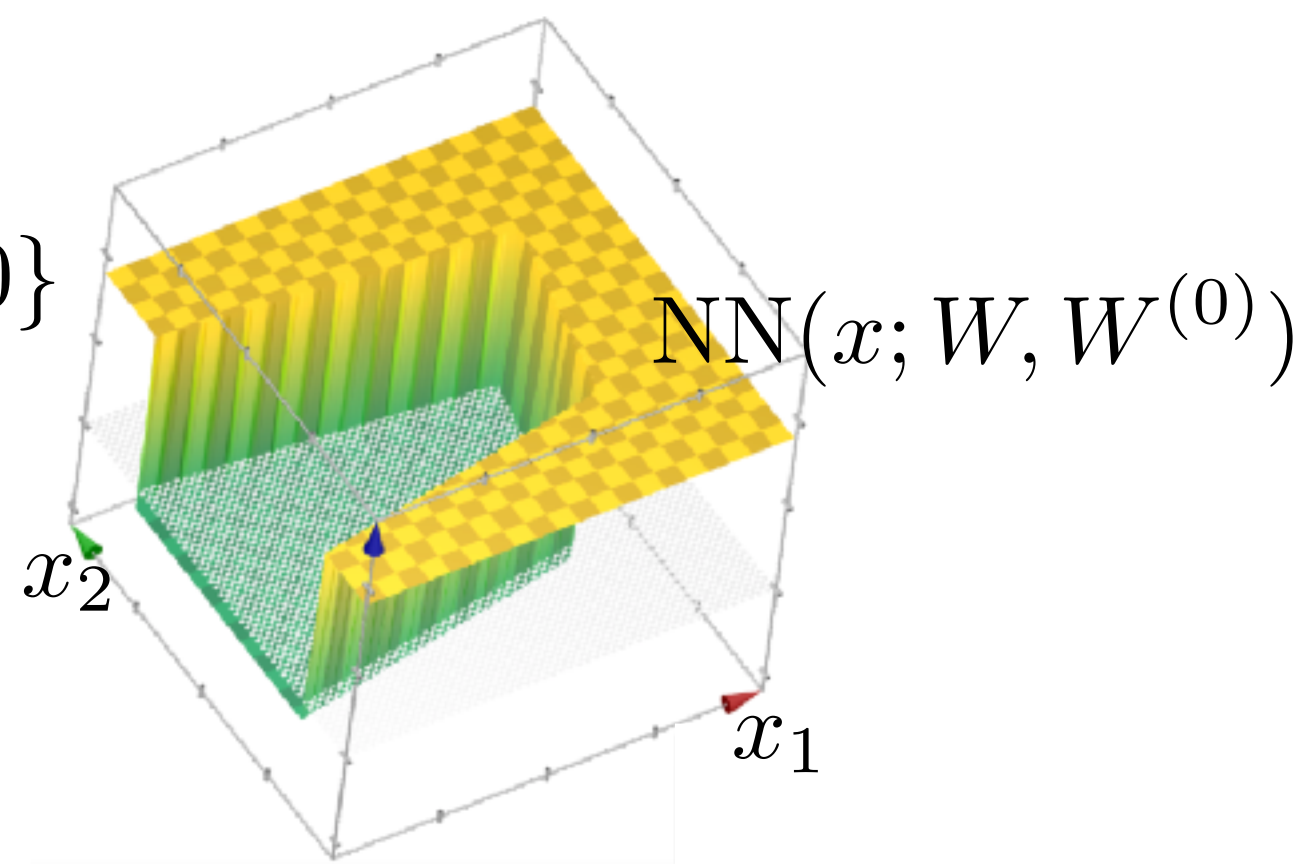


Choices of activation function

- 1st layer: $A_i^{(1)} = f^{(1)}(w_i^{(1)\top} x + w_0^{(1)})$
  - Choose $f^{(1)}(z) = \sigma(z)$

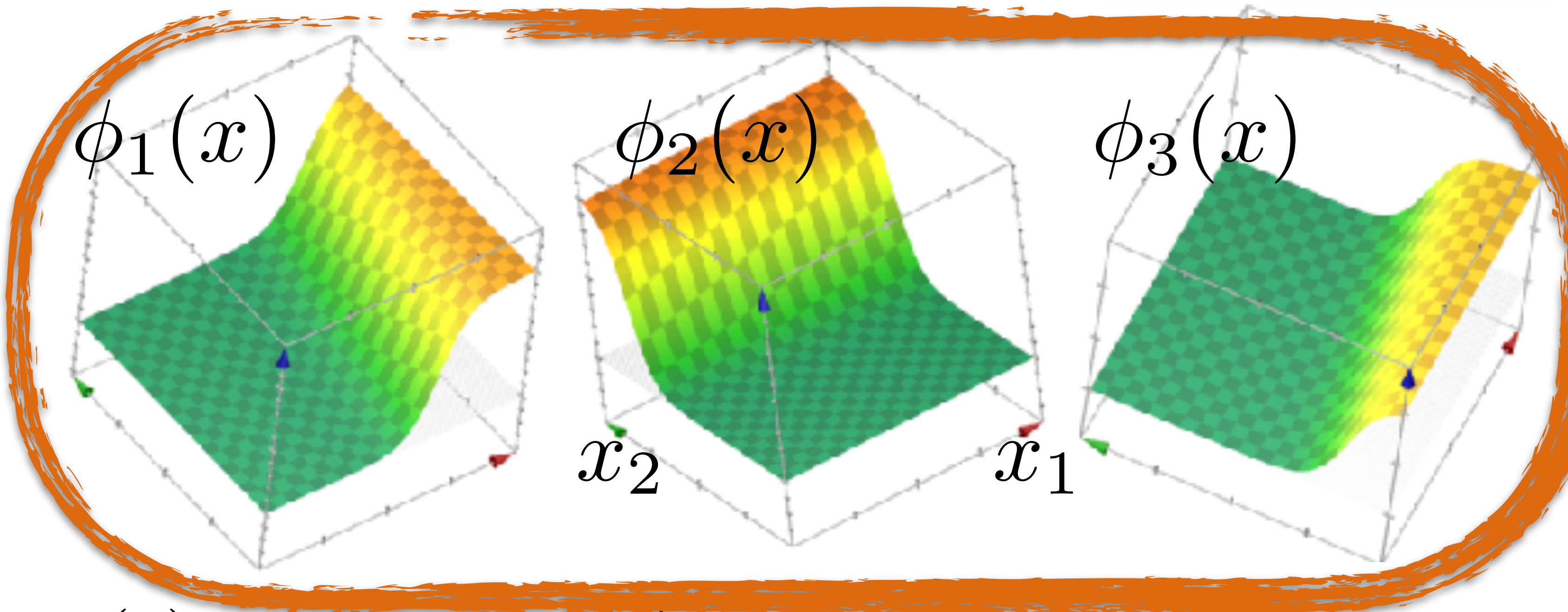


- 2nd layer: $A_i^{(2)} = f^{(2)}(w_i^{(2)\top} A^{(1)} + w_0^{(2)})$
  - Choose $f^{(2)}(z) = \mathbf{1}\{z \geq 0\}$

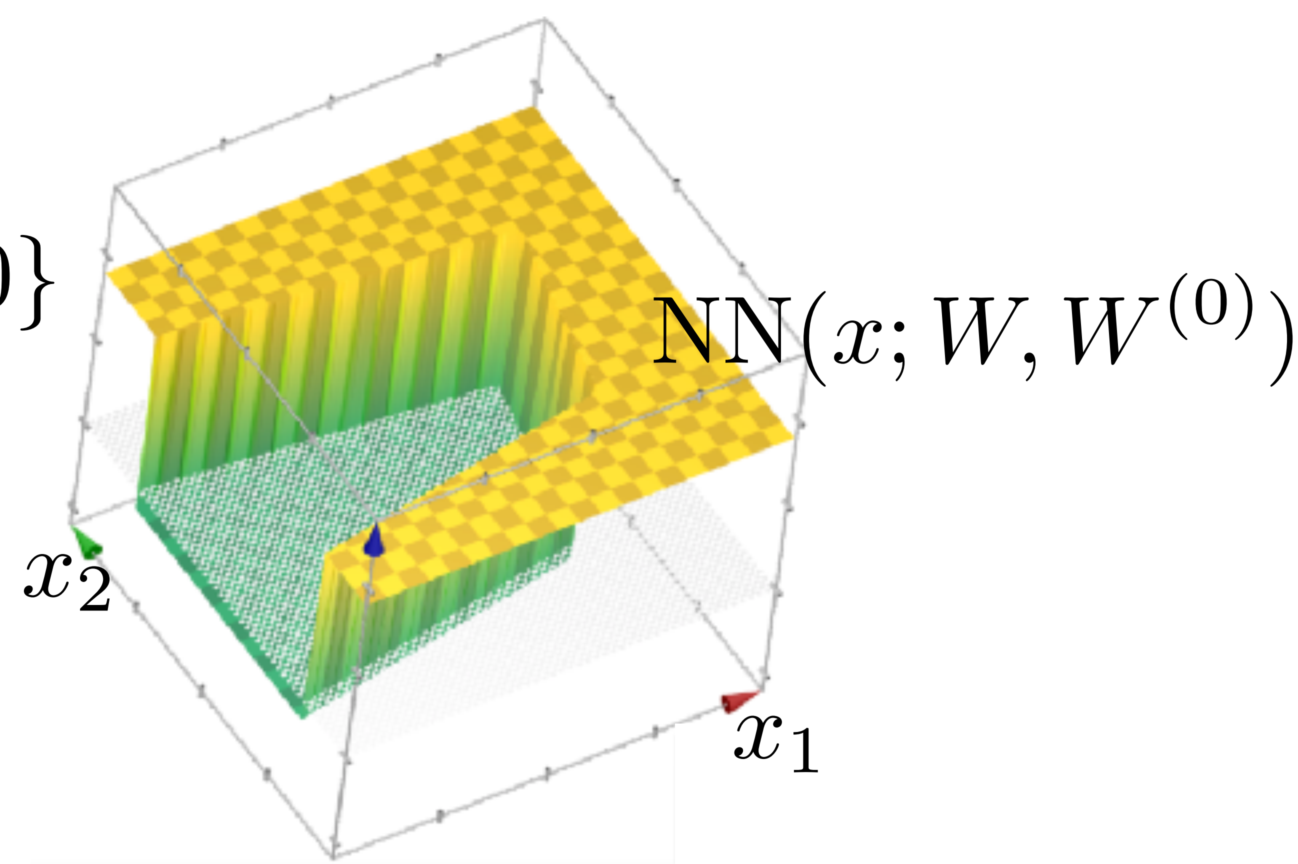


Choices of activation function

- 1st layer: $A_i^{(1)} = f^{(1)}(w_i^{(1)\top} x + w_0^{(1)})$
  - Choose $f^{(1)}(z) = \sigma(z)$

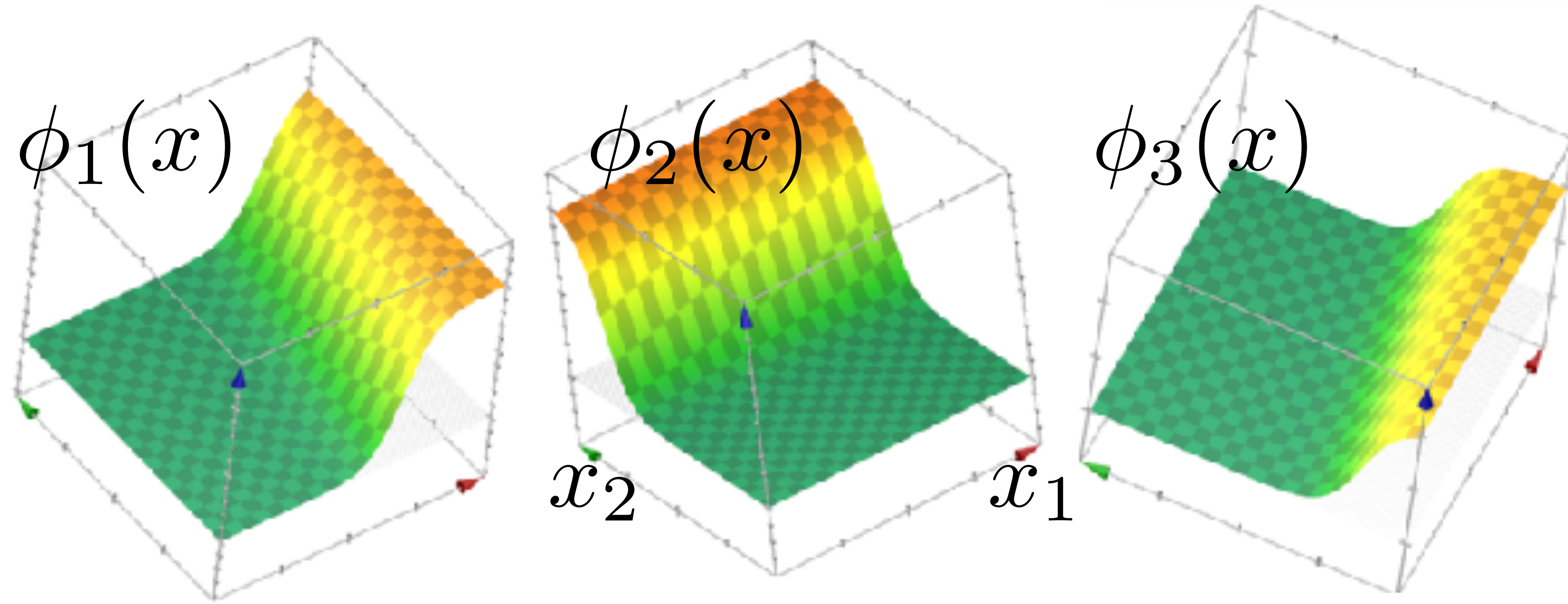


- 2nd layer: $A_i^{(2)} = f^{(2)}(w_i^{(2)\top} A^{(1)} + w_0^{(2)})$
  - Choose $f^{(2)}(z) = \mathbf{1}\{z \geq 0\}$

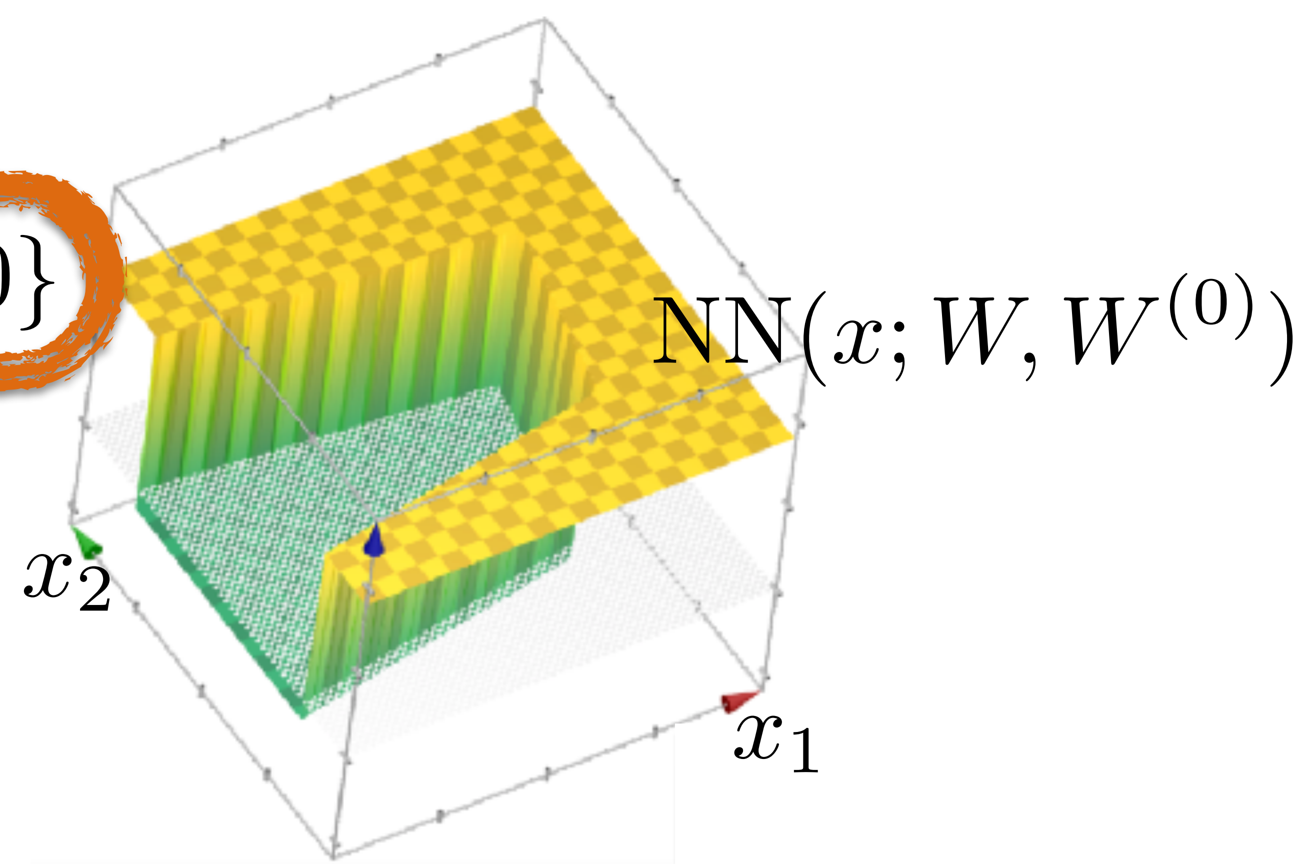


Choices of activation function

- 1st layer: $A_i^{(1)} = f^{(1)}(w_i^{(1)\top} x + w_0^{(1)})$
  - Choose $f^{(1)}(z) = \sigma(z)$

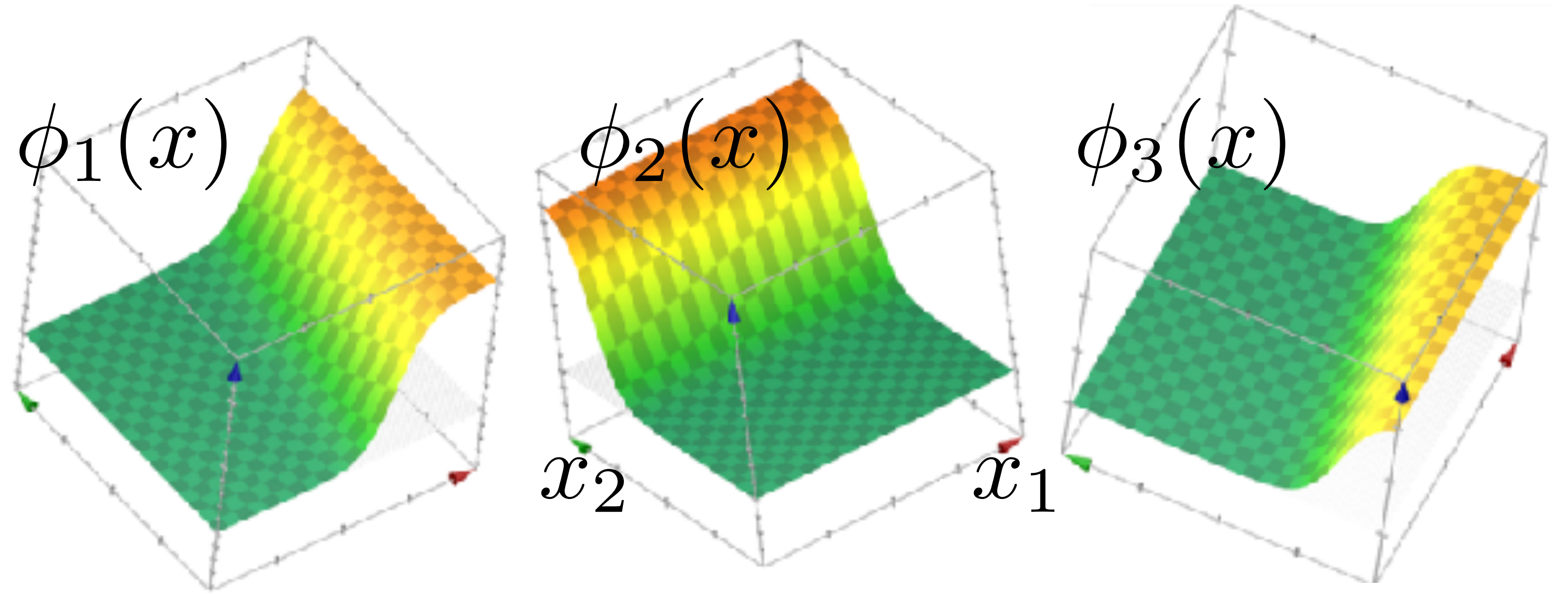


- 2nd layer: $A_i^{(2)} = f^{(2)}(w_i^{(2)\top} A^{(1)} + w_0^{(2)})$
  - Choose $f^{(2)}(z) = \mathbf{1}\{z \geq 0\}$

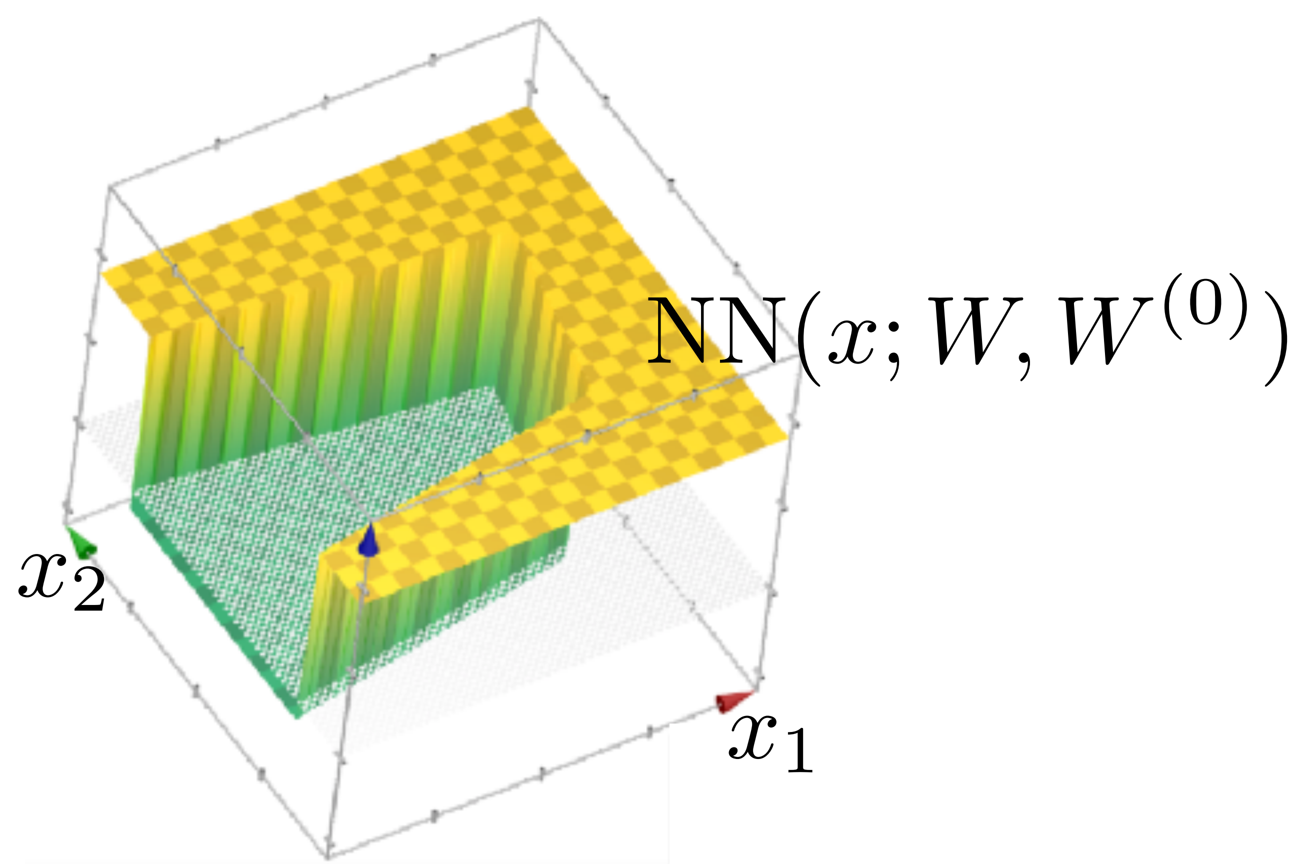


Choices of activation function

- 1st layer: $A_i^{(1)} = f^{(1)}(w_i^{(1)\top} x + w_0^{(1)})$
  - Choose $f^{(1)}(z) = \sigma(z)$

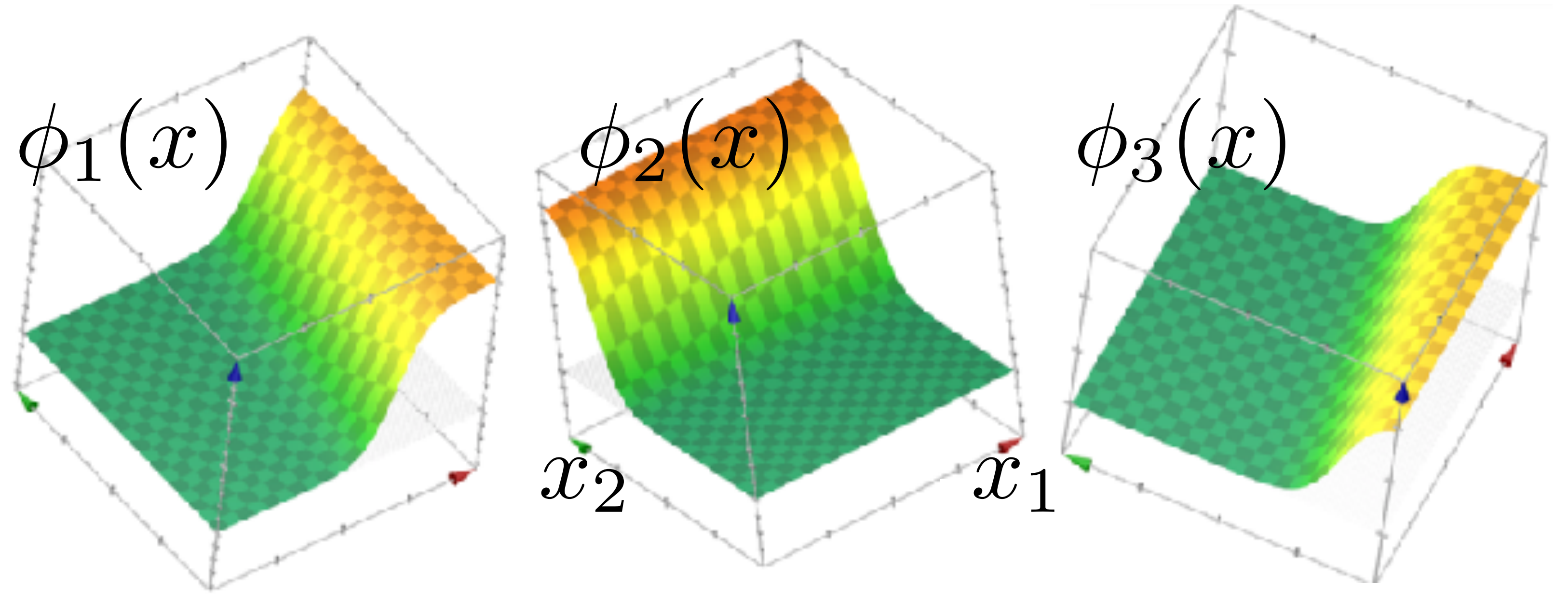


- 2nd layer: $A_i^{(2)} = f^{(2)}(w_i^{(2)\top} A^{(1)} + w_0^{(2)})$
  - Choose $f^{(2)}(z) = z$

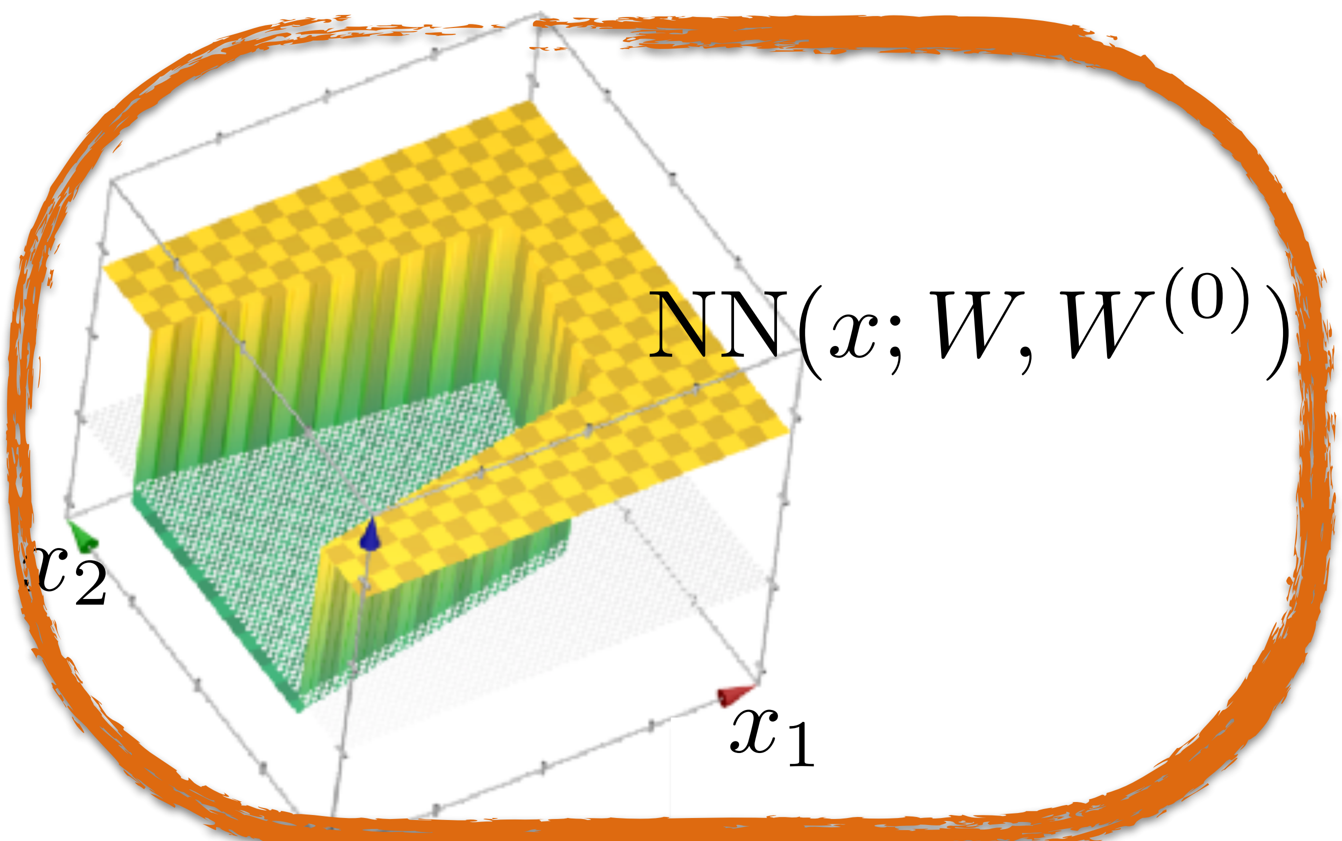


Choices of activation function

- 1st layer: $A_i^{(1)} = f^{(1)}(w_i^{(1)\top} x + w_0^{(1)})$
  - Choose $f^{(1)}(z) = \sigma(z)$

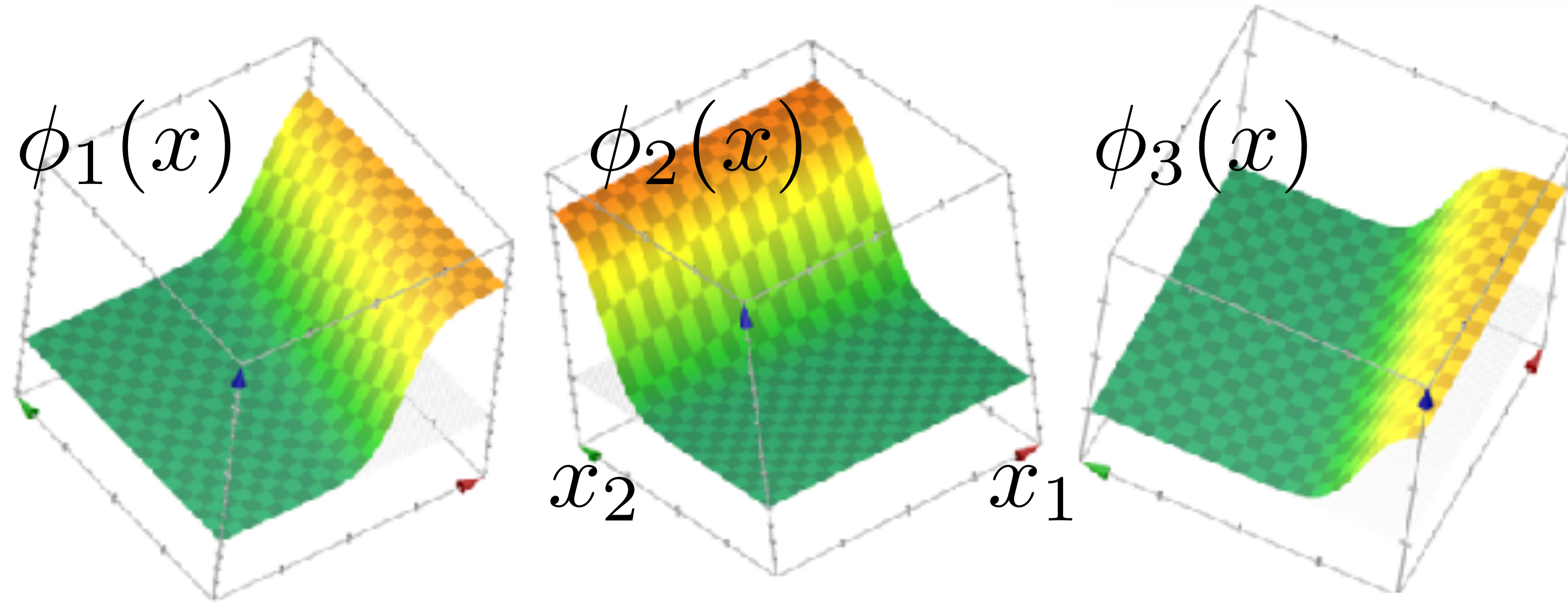


- 2nd layer: $A_i^{(2)} = f^{(2)}(w_i^{(2)\top} A^{(1)} + w_0^{(2)})$
  - Choose $f^{(2)}(z) = z$

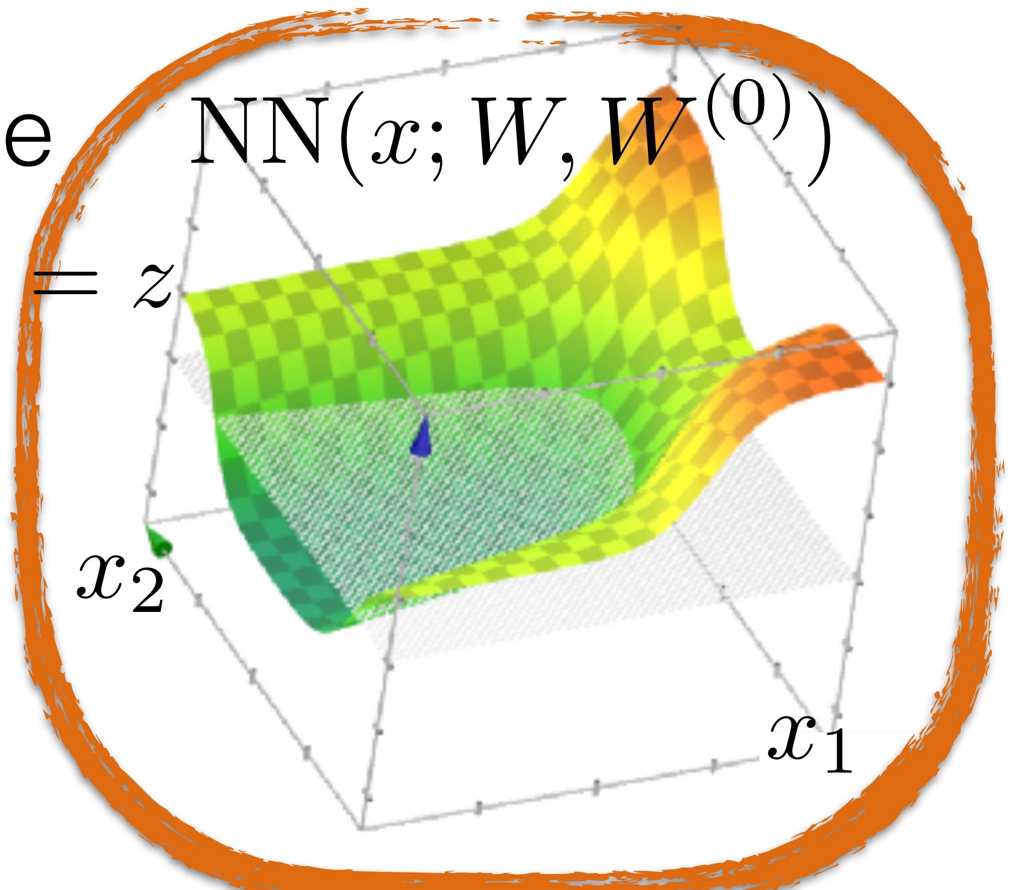


Choices of activation function

- 1st layer: $A_i^{(1)} = f^{(1)}(w_i^{(1)\top} x + w_0^{(1)})$
  - Choose
    $f^{(1)}(z) = \sigma(z)$

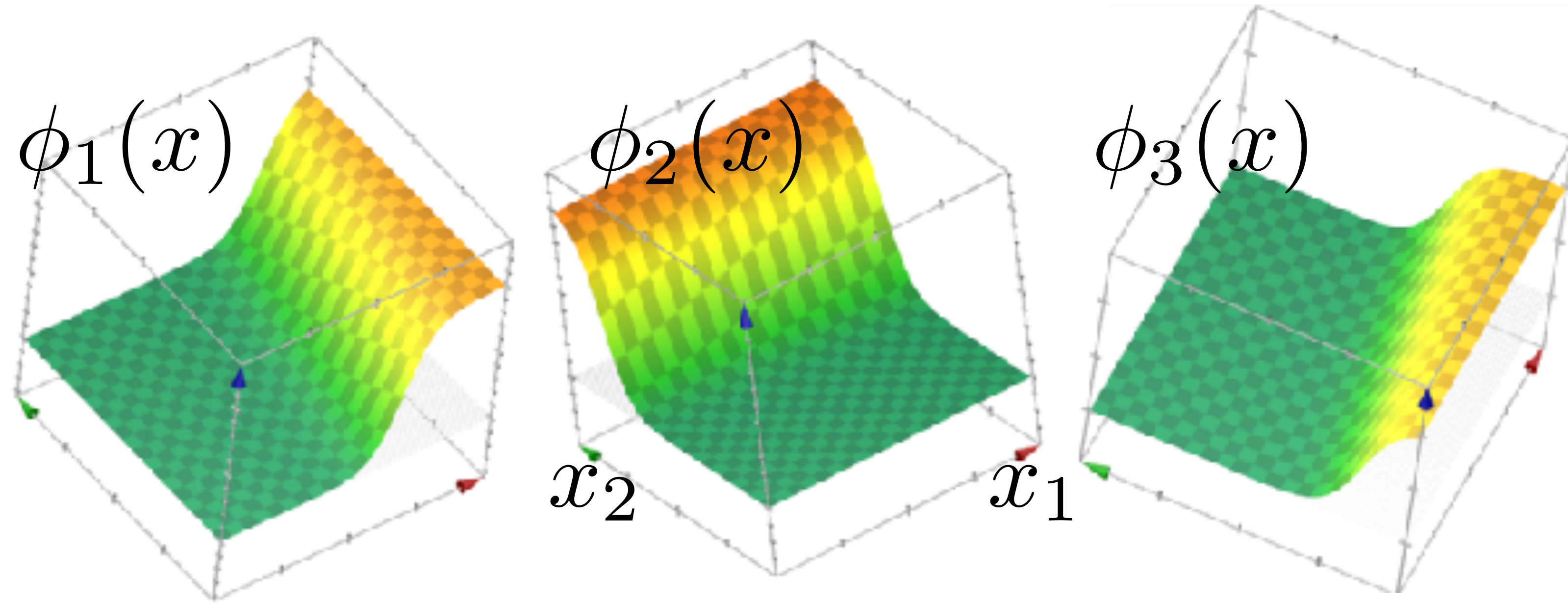


- 2nd layer: $A_i^{(2)} = f^{(2)}(w_i^{(2)\top} A^{(1)} + w_0^{(2)})$
  - Choose
    $f^{(2)}(z) = z$

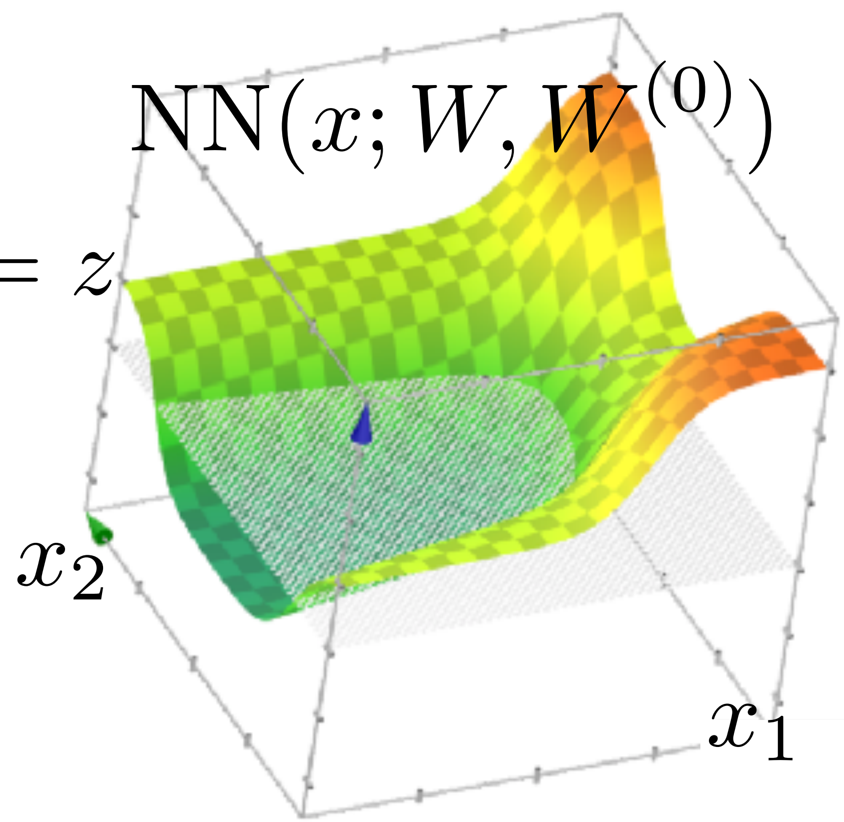


Choices of activation function

- 1st layer: $A_i^{(1)} = f^{(1)}(w_i^{(1)\top} x + w_0^{(1)})$
  - Choose $f^{(1)}(z) = \sigma(z)$

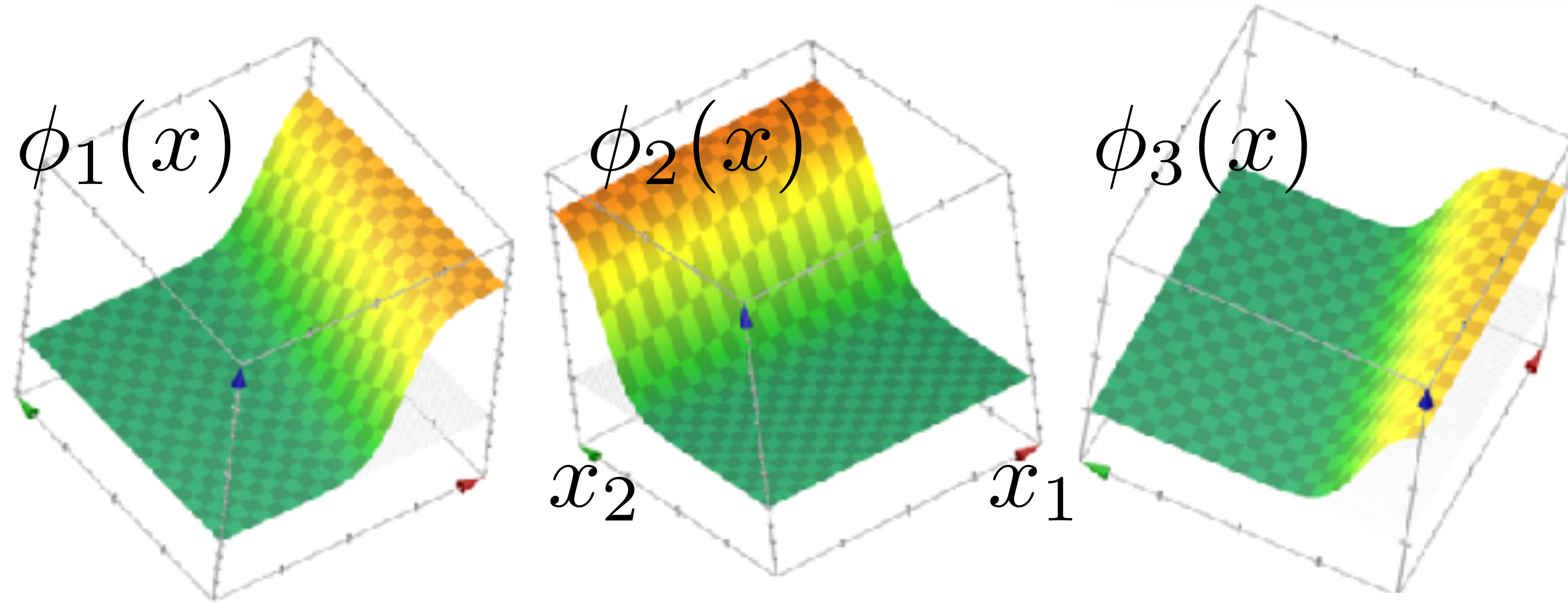


- 2nd layer: $A_i^{(2)} = f^{(2)}(w_i^{(2)\top} A^{(1)} + w_0^{(2)})$
  - Choose $f^{(2)}(z) = z$

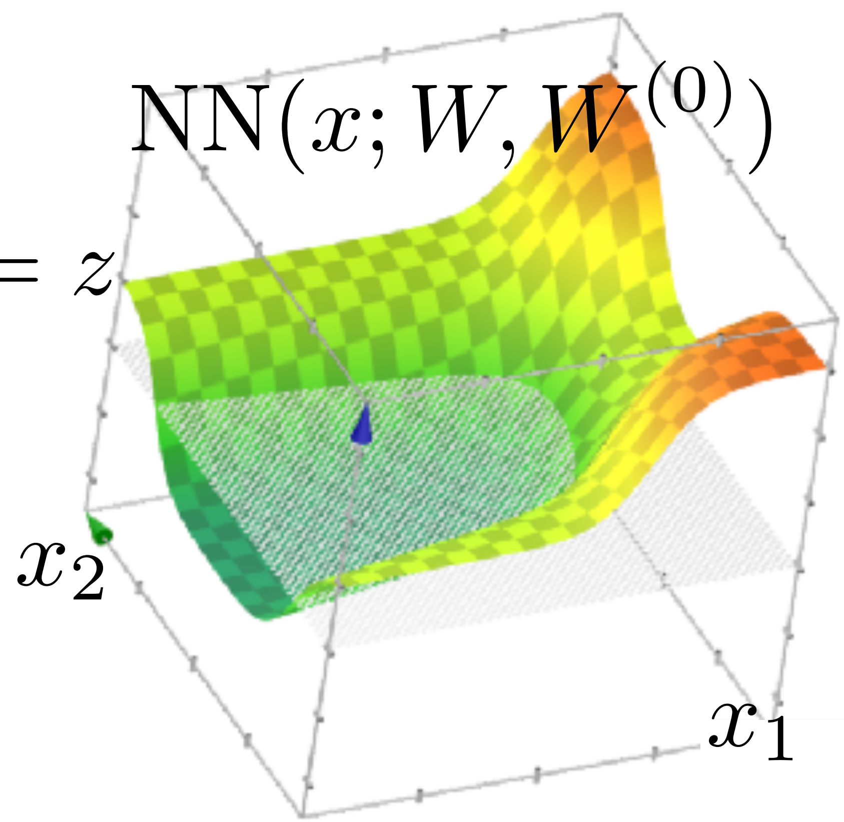


Choices of activation function

- 1st layer: $A_i^{(1)} = f^{(1)}(w_i^{(1)\top} x + w_0^{(1)})$
  - Choose $f^{(1)}(z) = \sigma(z)$



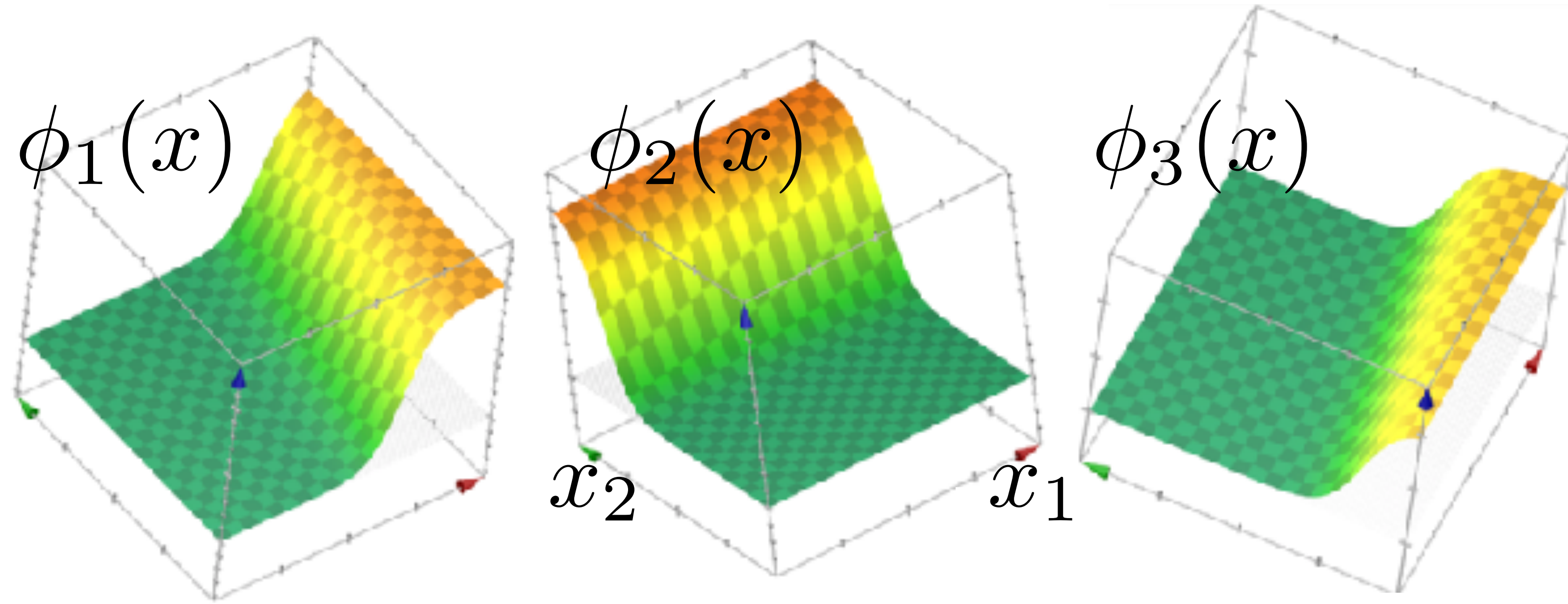
- 2nd layer: $A_i^{(2)} = f^{(2)}(w_i^{(2)\top} A^{(1)} + w_0^{(2)})$
  - Choose $f^{(2)}(z) = z$



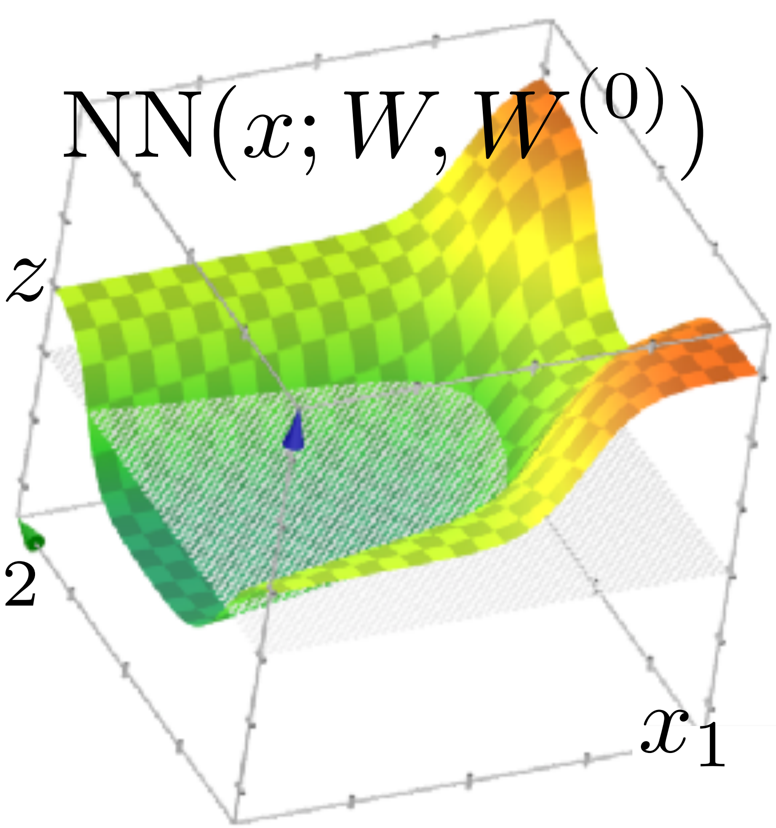
- Choose $f^{(2)}(z) = \sigma(z)$

Choices of activation function

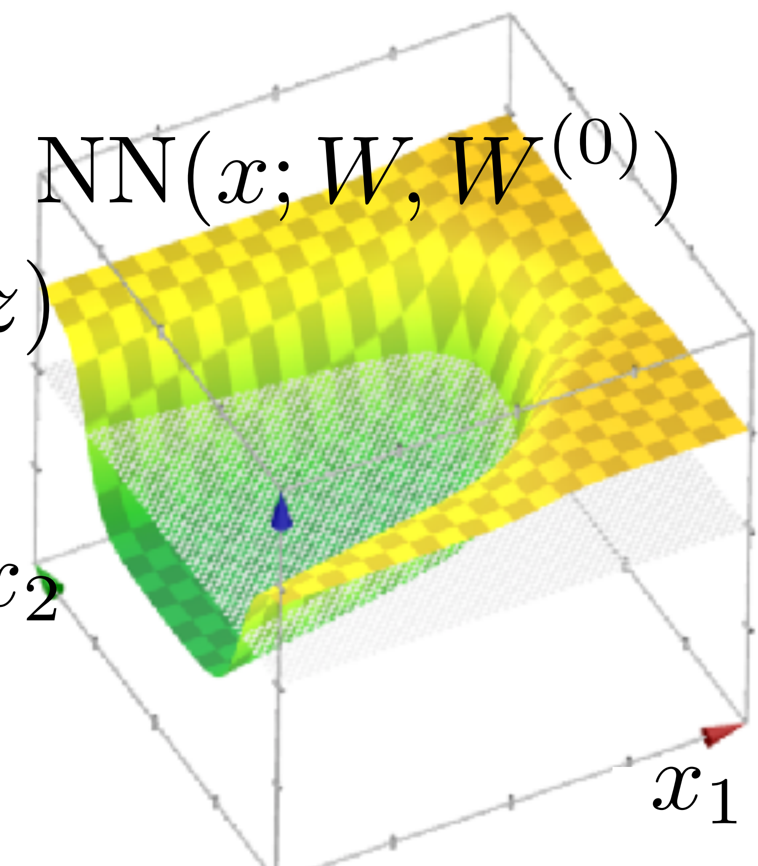
- 1st layer: $A_i^{(1)} = f^{(1)}(w_i^{(1)\top} x + w_0^{(1)})$
  - Choose $f^{(1)}(z) = \sigma(z)$



- 2nd layer: $A_i^{(2)} = f^{(2)}(w_i^{(2)\top} A^{(1)} + w_0^{(2)})$
  - Choose $f^{(2)}(z) = z$ $\quad \mathrm{NN}(x; W, W^{(0)})$



  - Choose $f^{(2)}(z) = \sigma(z)$ $\quad \mathrm{NN}(x; W, W^{(0)})$



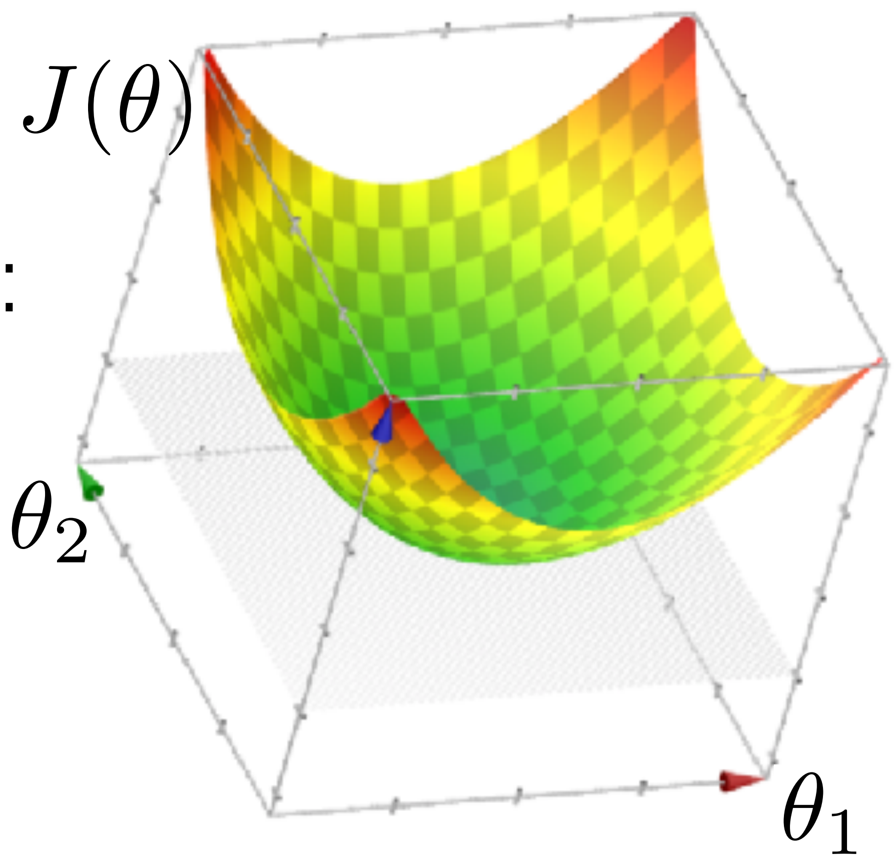
Learning the parameters

Learning the parameters

- Objective: $J(W, W_0) = \dfrac{1}{n_{\text{data}}} \displaystyle\sum_{i=1}^{n_{\text{data}}} L(h(x^{(i)}; W, W_0), y^{(i)})$

Learning the parameters

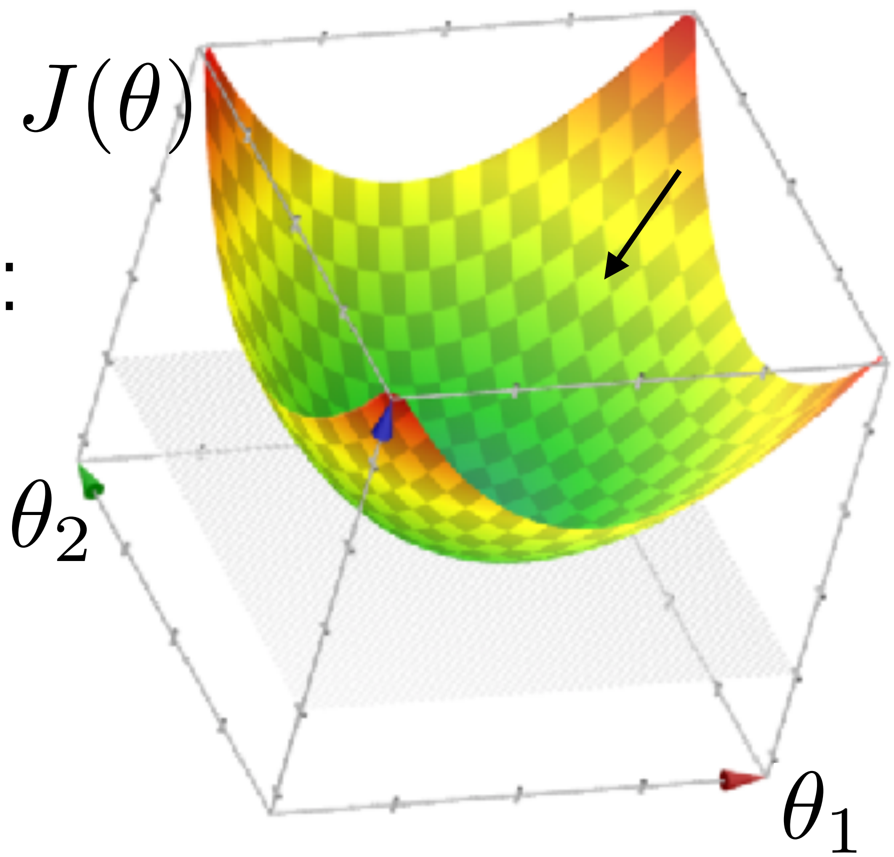
- Objective: $J(W, W_0) = \frac{1}{n_{\text{data}}} \sum_{i=1}^{n_{\text{data}}} L(h(x^{(i)}; W, W_0), y^{(i)})$
- GD:



Learning the parameters

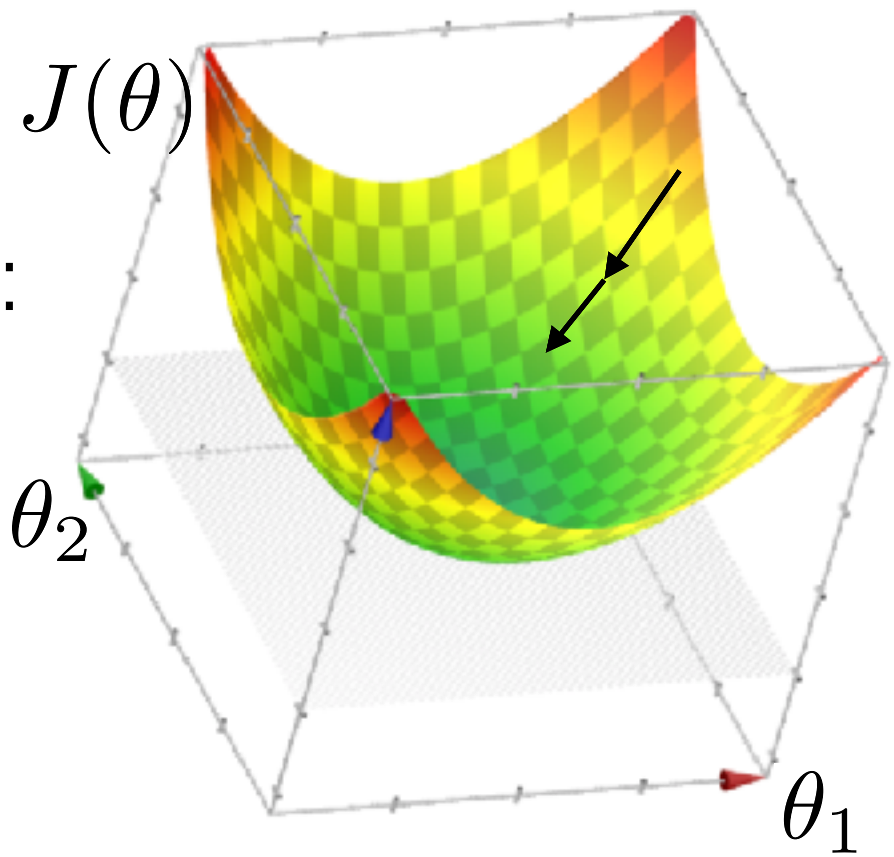
- Objective: $J(W, W_0) = \frac{1}{n_{\text{data}}} \sum_{i=1}^{n_{\text{data}}} L(h(x^{(i)}; W, W_0), y^{(i)})$

- GD:



Learning the parameters

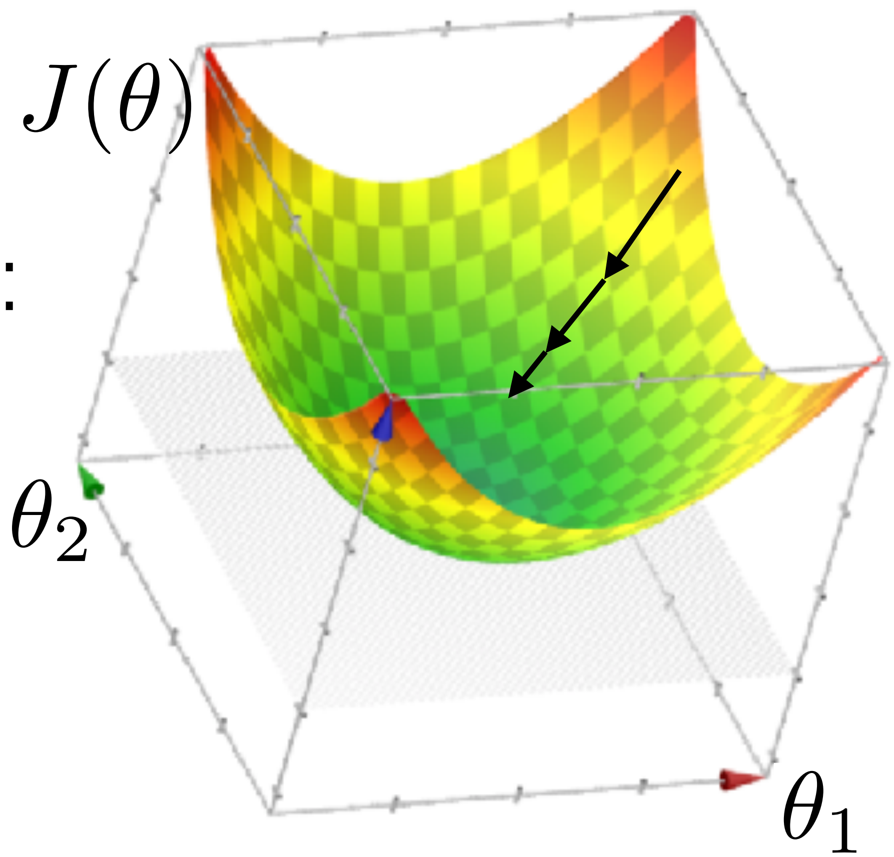
- Objective: $J(W, W_0) = \frac{1}{n_{\text{data}}} \sum_{i=1}^{n_{\text{data}}} L(h(x^{(i)}; W, W_0), y^{(i)})$
- GD:



Learning the parameters

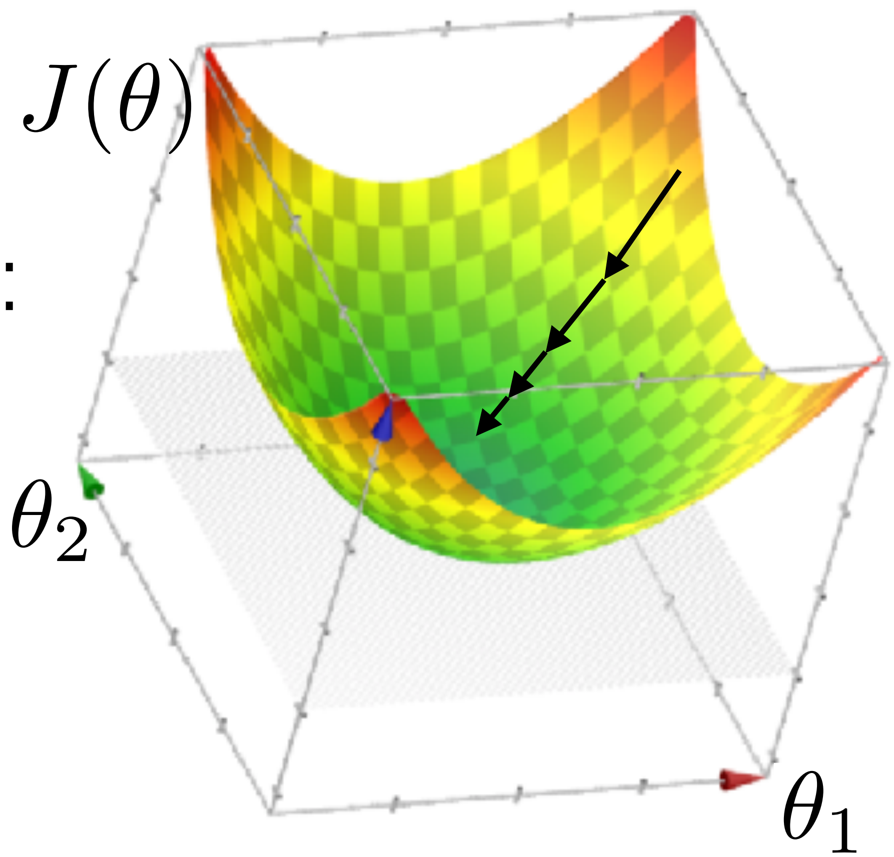
- Objective: $J(W, W_0) = \frac{1}{n_{\text{data}}} \sum_{i=1}^{n_{\text{data}}} L(h(x^{(i)}; W, W_0), y^{(i)})$

- GD:



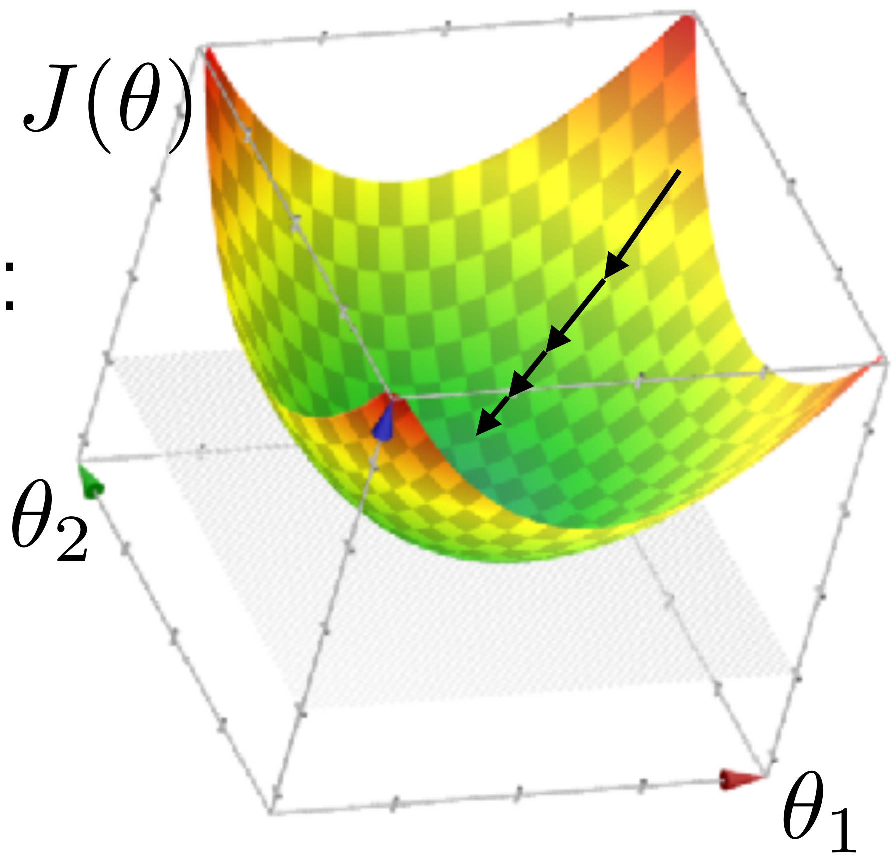
Learning the parameters

- Objective: $J(W, W_0) = \frac{1}{n_{\text{data}}} \sum_{i=1}^{n_{\text{data}}} L(h(x^{(i)}; W, W_0), y^{(i)})$
- GD:



Learning the parameters

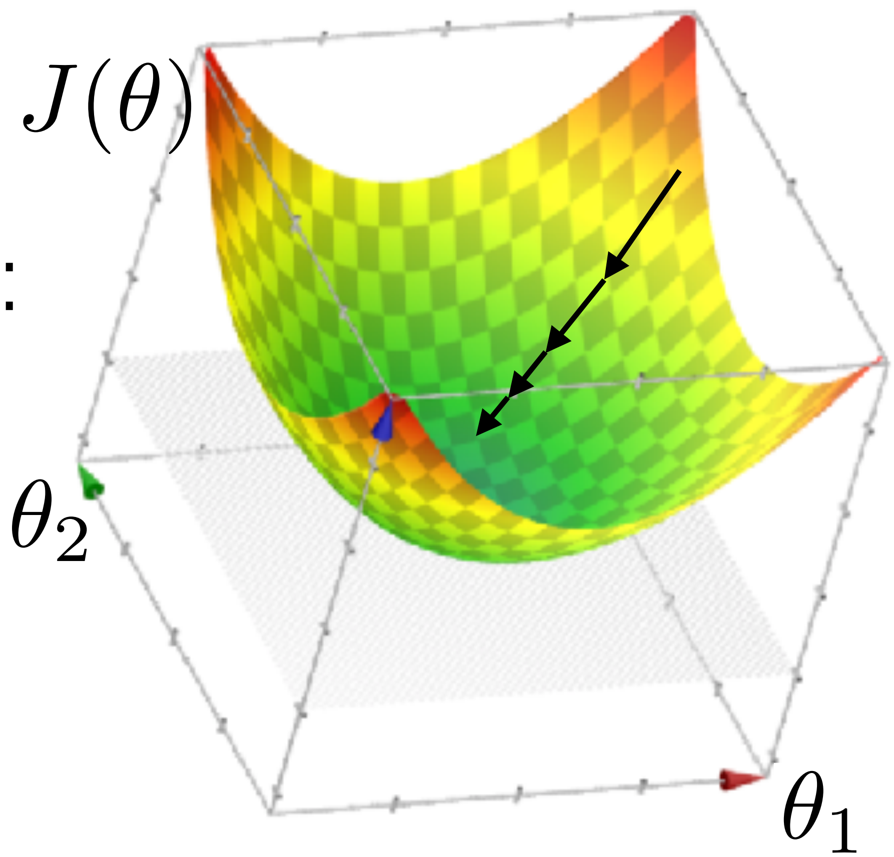
- Objective: $J(W, W_0) = \frac{1}{n_{\text{data}}} \sum_{i=1}^{n_{\text{data}}} L(h(x^{(i)}; W, W_0), y^{(i)})$



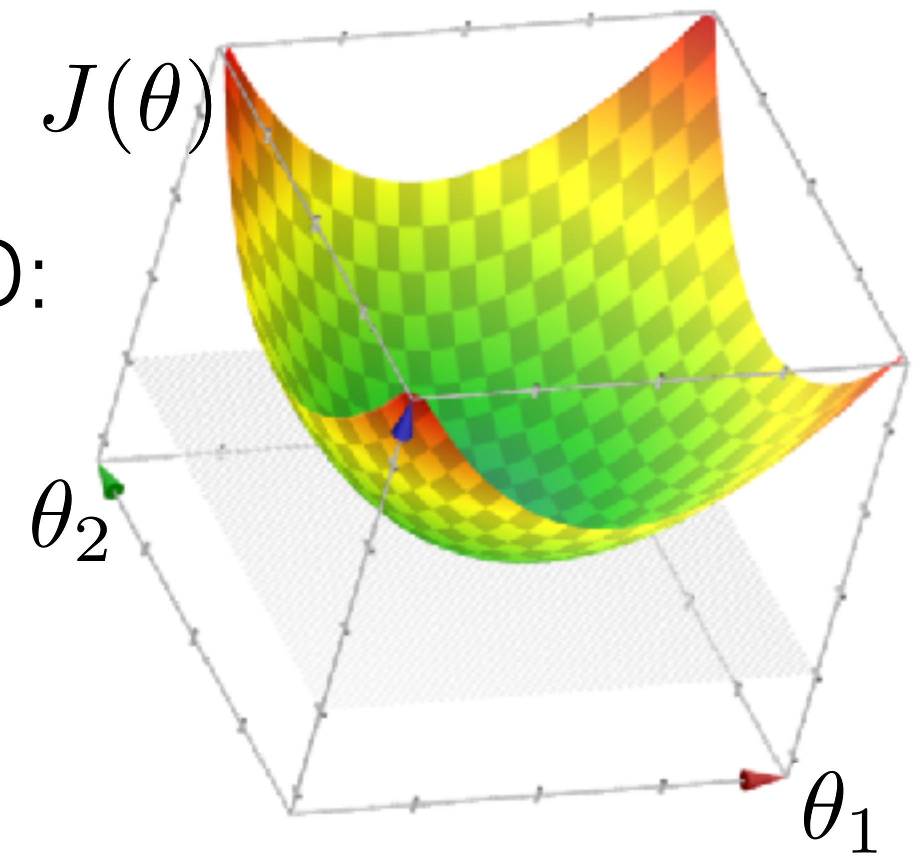
- GD:

Learning the parameters

- Objective: $J(W, W_0) = \frac{1}{n_{\text{data}}} \sum_{i=1}^{n_{\text{data}}} L(h(x^{(i)}; W, W_0), y^{(i)})$
- GD:



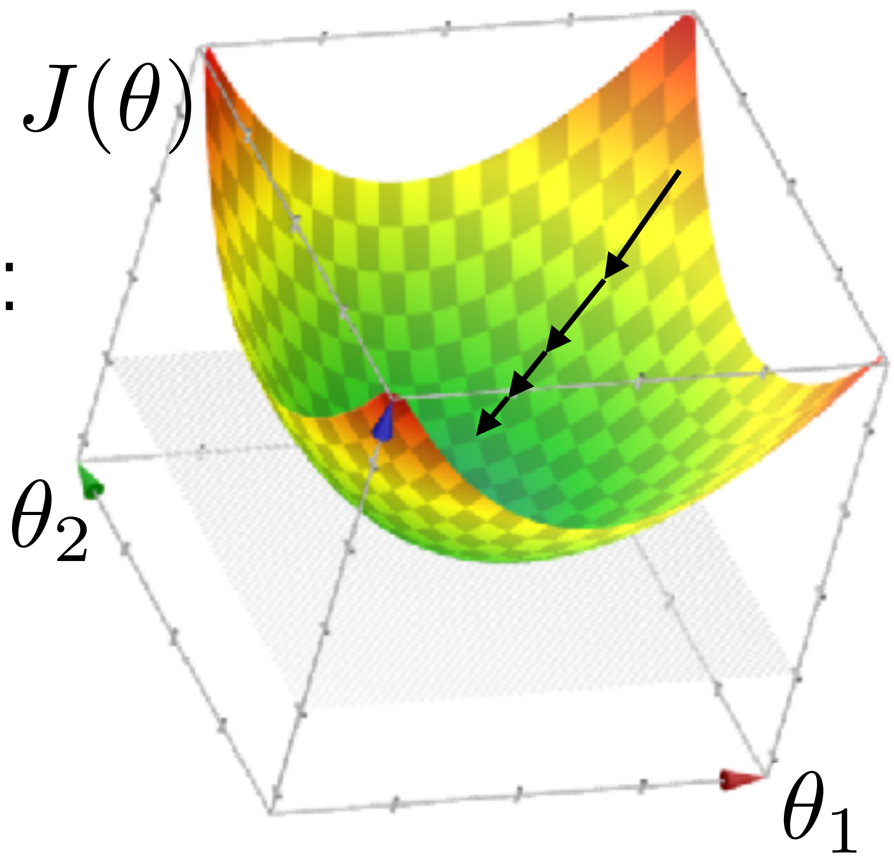
- SGD:



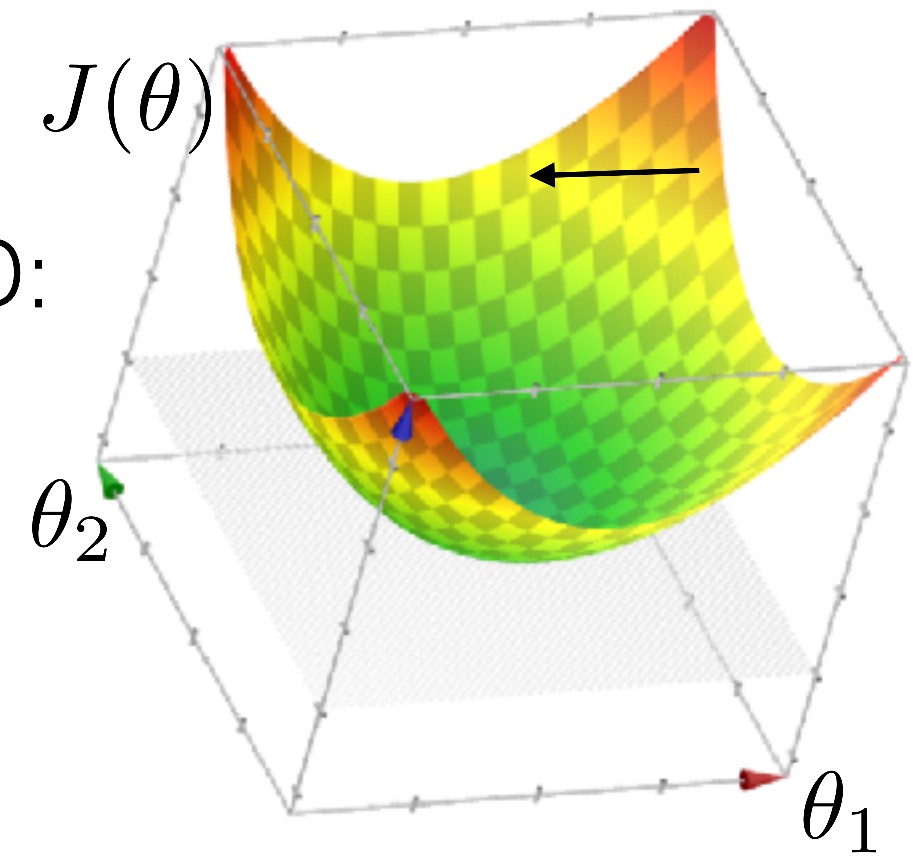
Learning the parameters

- Objective: $J(W, W_0) = \frac{1}{n_{\text{data}}} \sum_{i=1}^{n_{\text{data}}} L(h(x^{(i)}; W, W_0), y^{(i)})$

- GD:



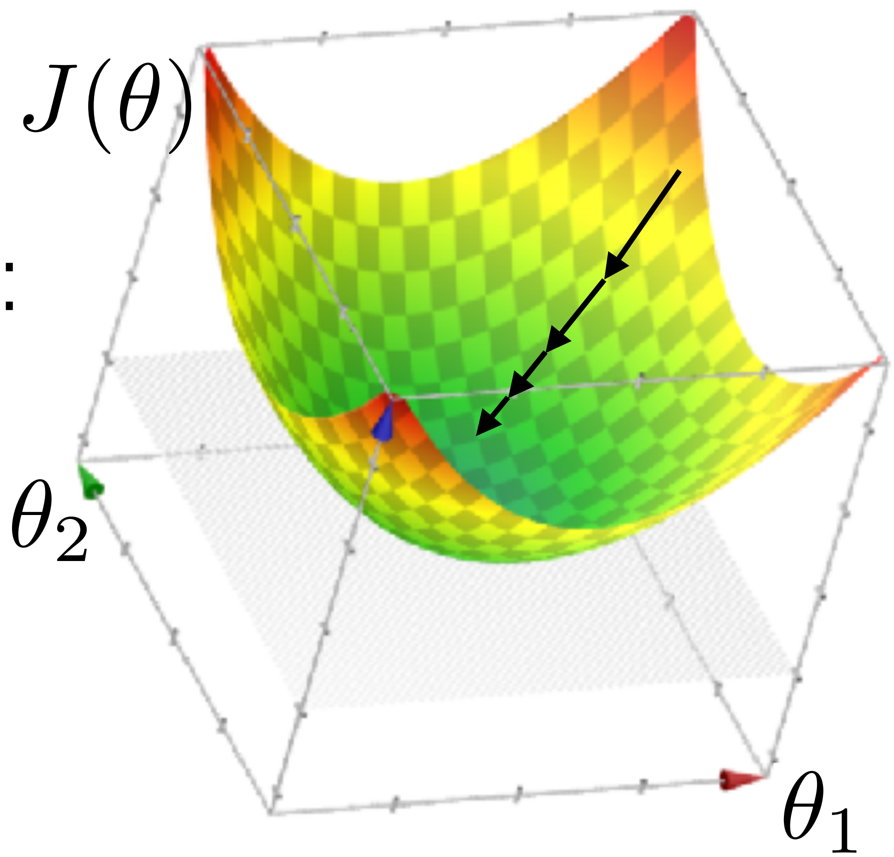
- SGD:



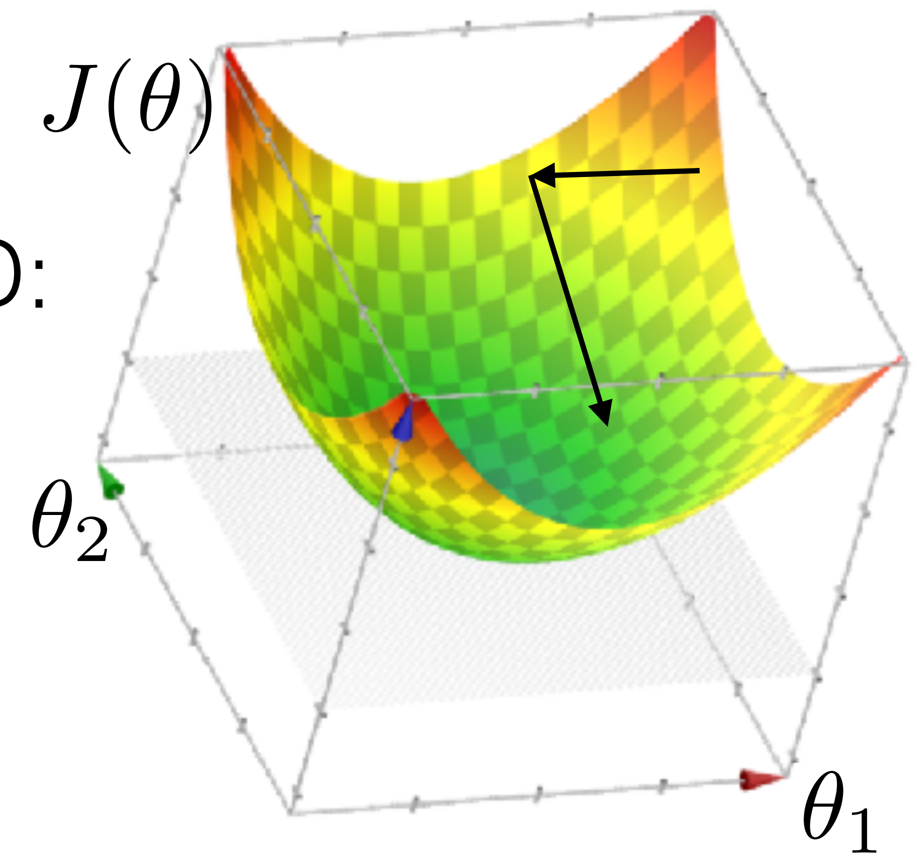
Learning the parameters

- Objective: $J(W, W_0) = \frac{1}{n_{\text{data}}} \sum_{i=1}^{n_{\text{data}}} L(h(x^{(i)}; W, W_0), y^{(i)})$

- GD:



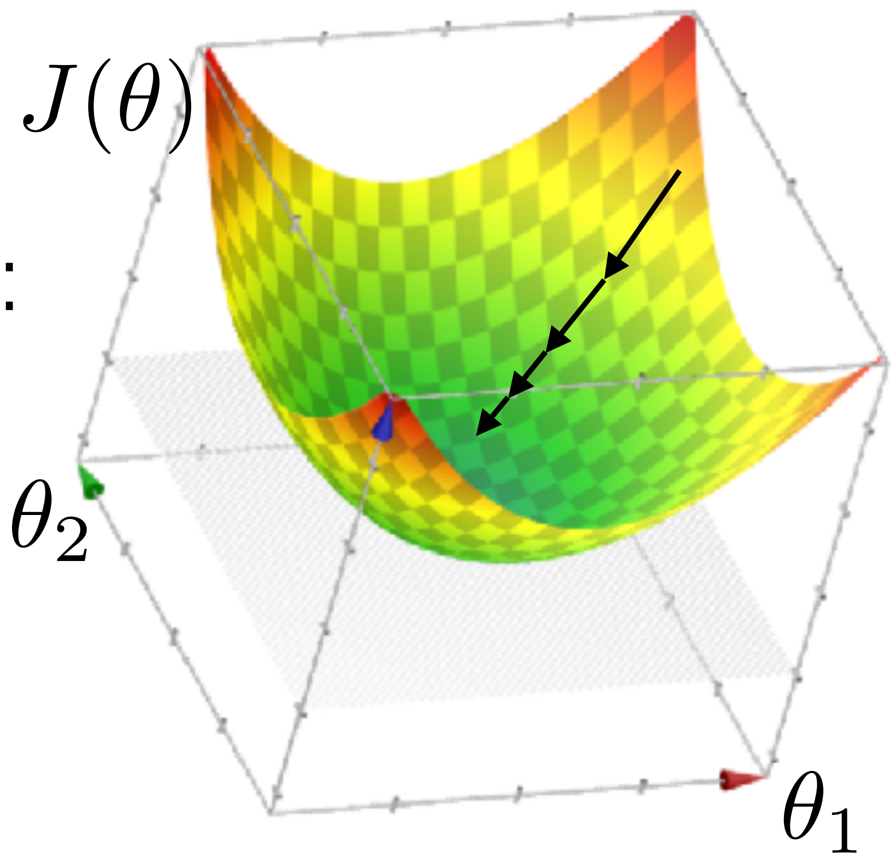
- SGD:



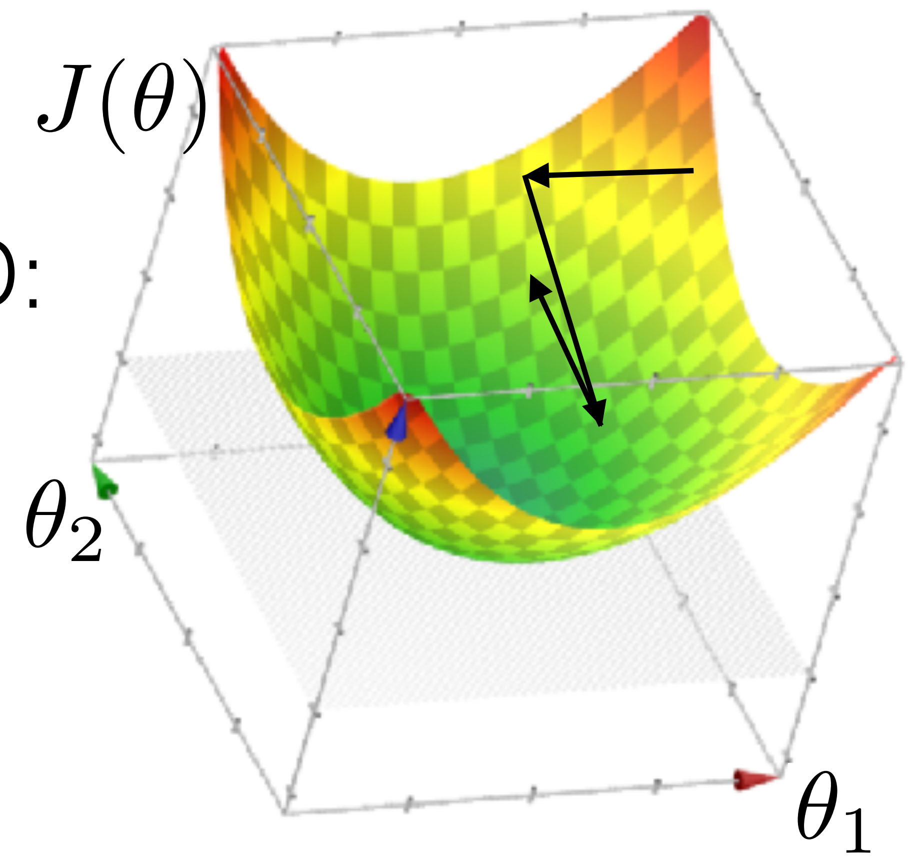
Learning the parameters

- Objective: $J(W, W_0) = \frac{1}{n_{\text{data}}} \sum_{i=1}^{n_{\text{data}}} L(h(x^{(i)}; W, W_0), y^{(i)})$

- GD:



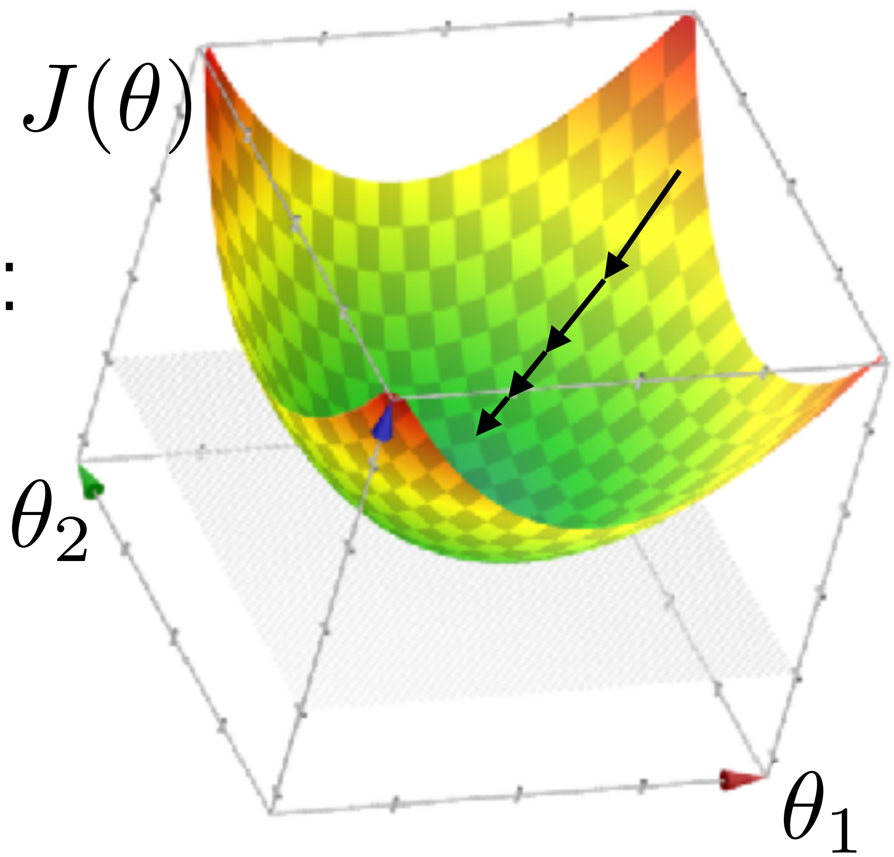
- SGD:



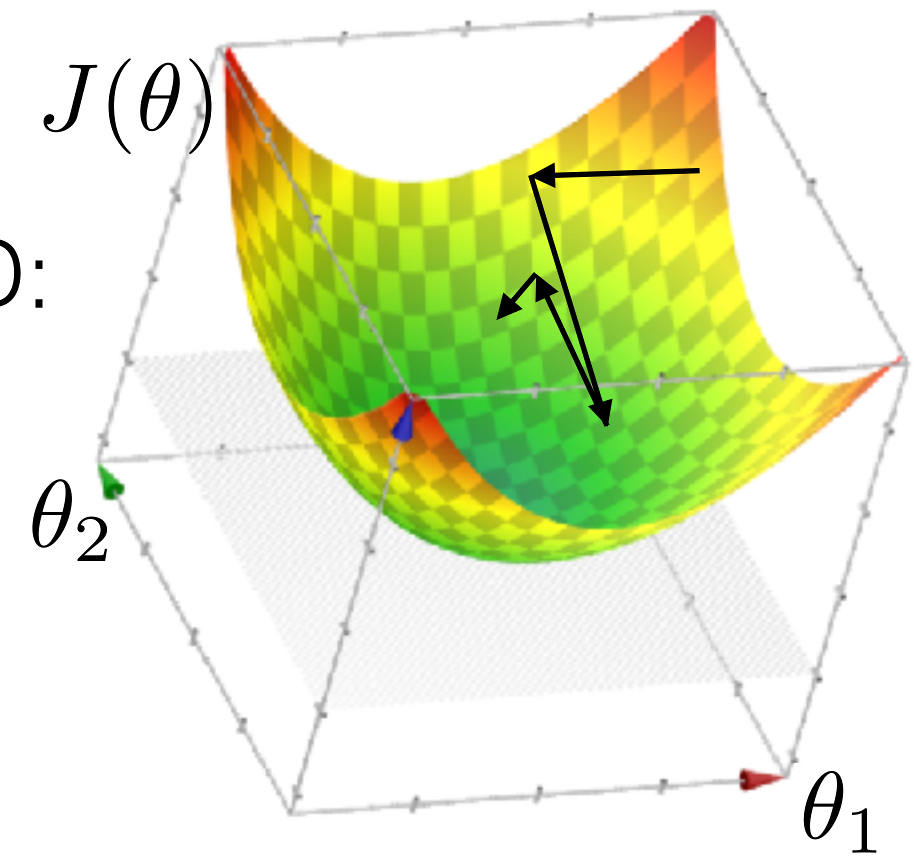
Learning the parameters

- Objective: $J(W, W_0) = \frac{1}{n_{\text{data}}} \sum_{i=1}^{n_{\text{data}}} L(h(x^{(i)}; W, W_0), y^{(i)})$

- GD:



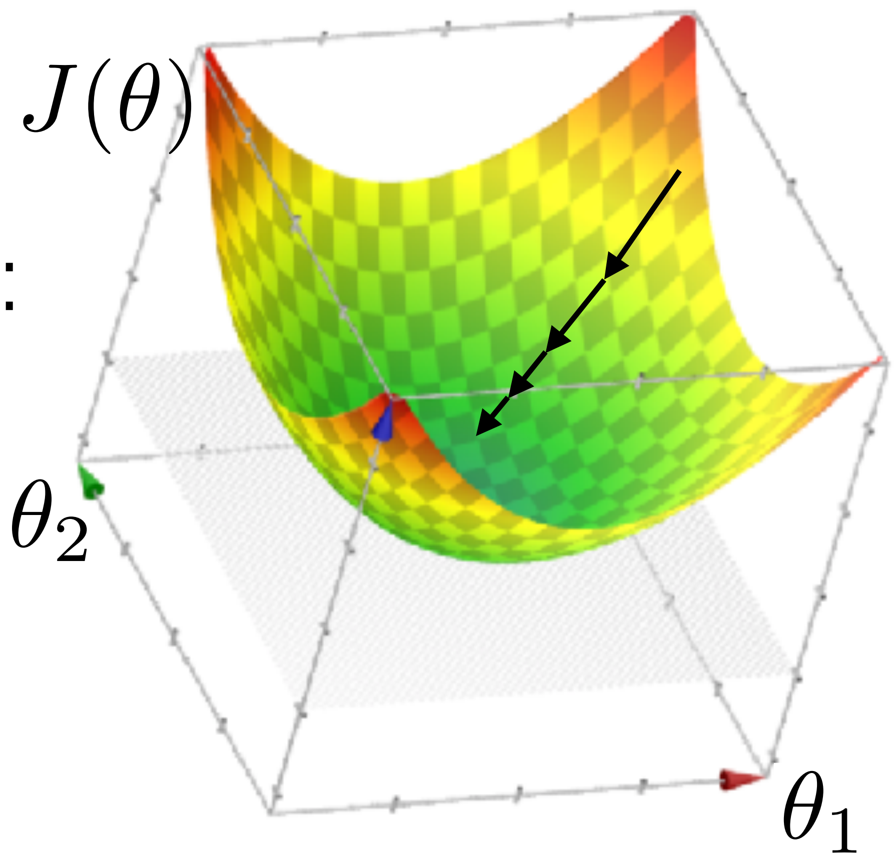
- SGD:



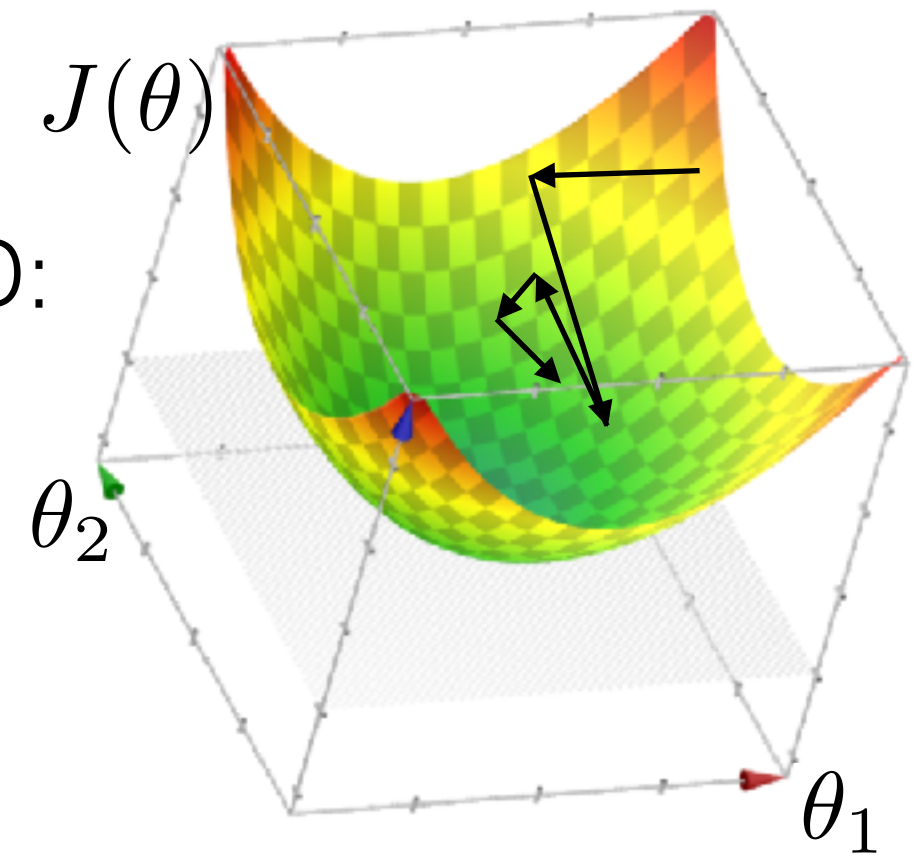
Learning the parameters

- Objective: $J(W, W_0) = \frac{1}{n_{\text{data}}} \sum_{i=1}^{n_{\text{data}}} L(h(x^{(i)}; W, W_0), y^{(i)})$

- GD:

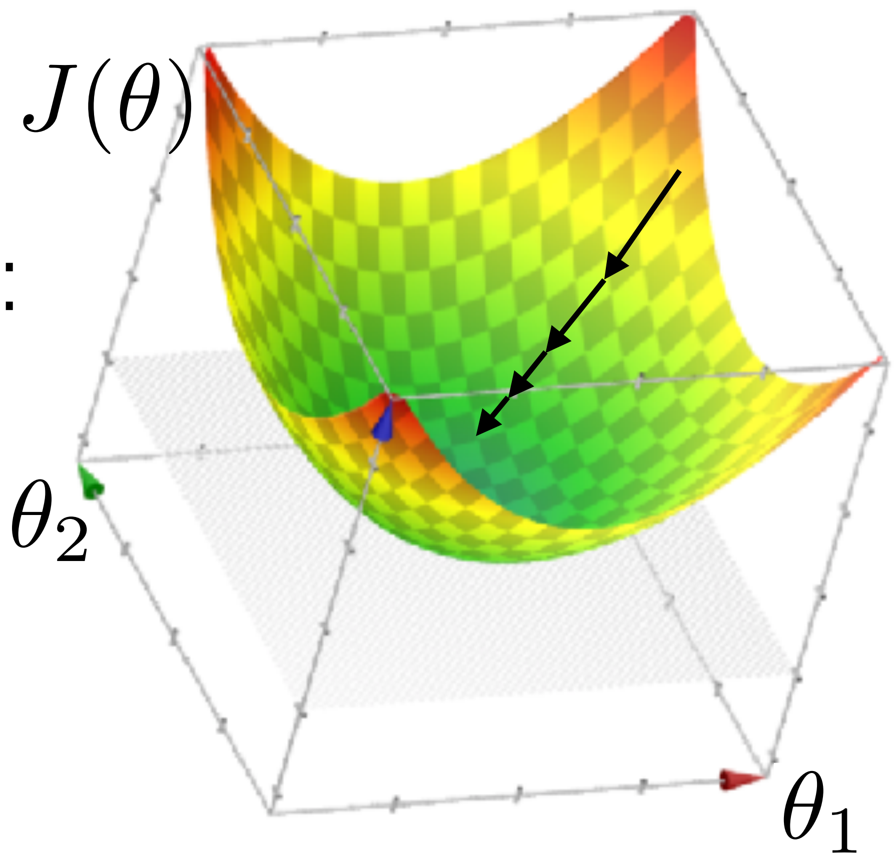


- SGD:

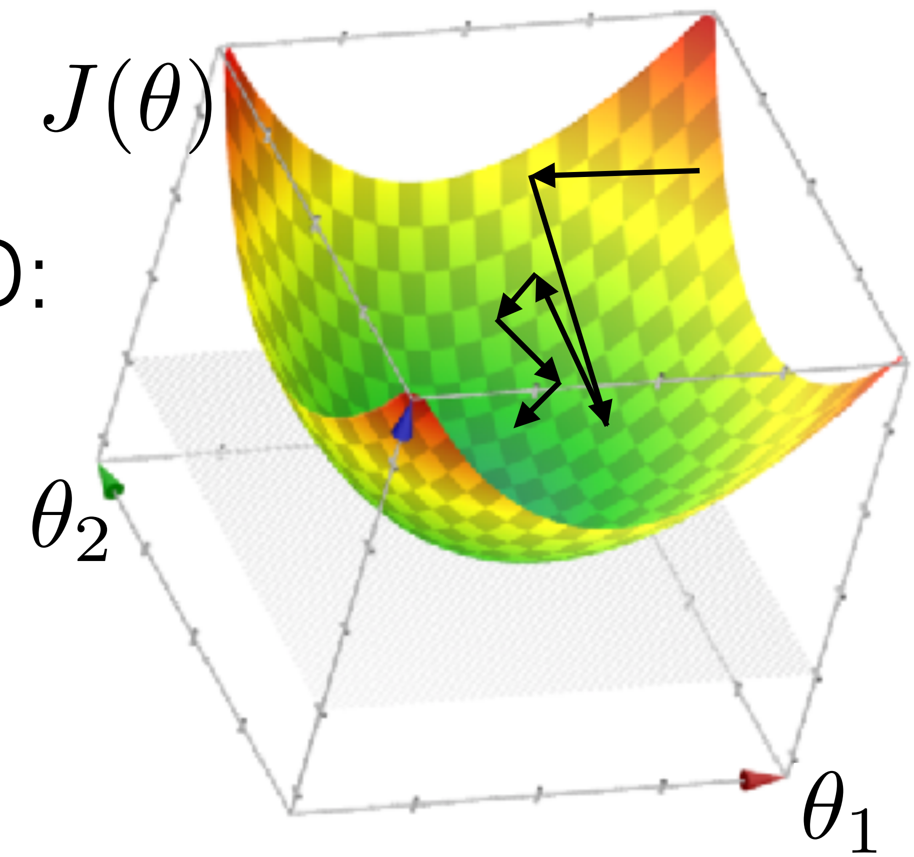


Learning the parameters

- Objective: $J(W, W_0) = \frac{1}{n_{\text{data}}} \sum_{i=1}^{n_{\text{data}}} L(h(x^{(i)}; W, W_0), y^{(i)})$
- GD:



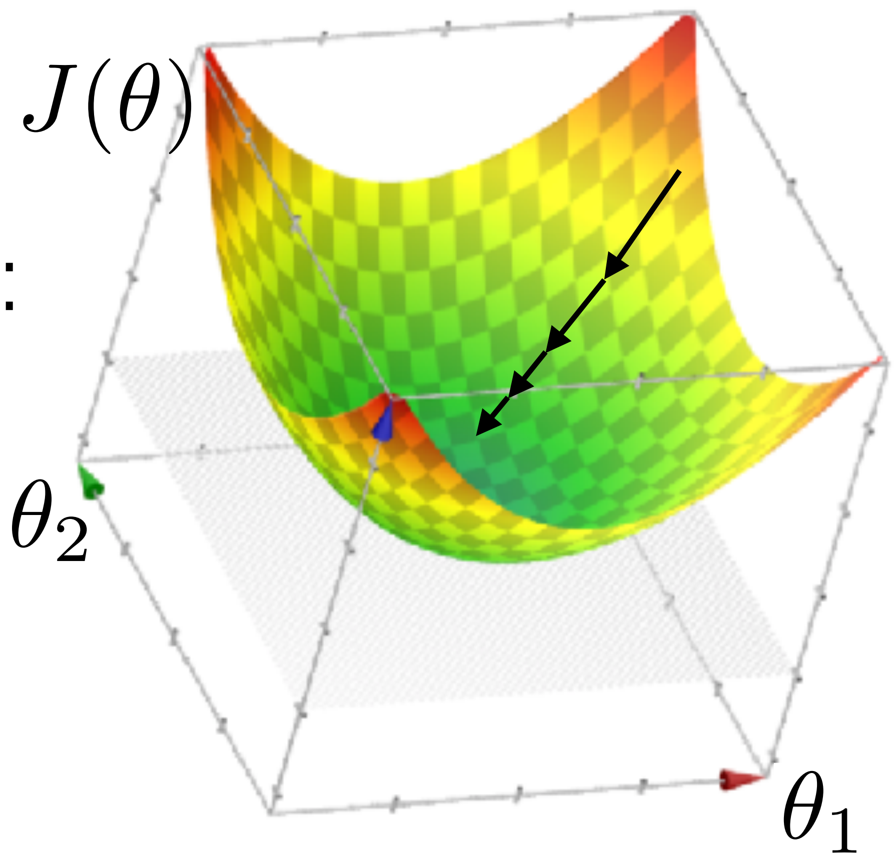
- SGD:



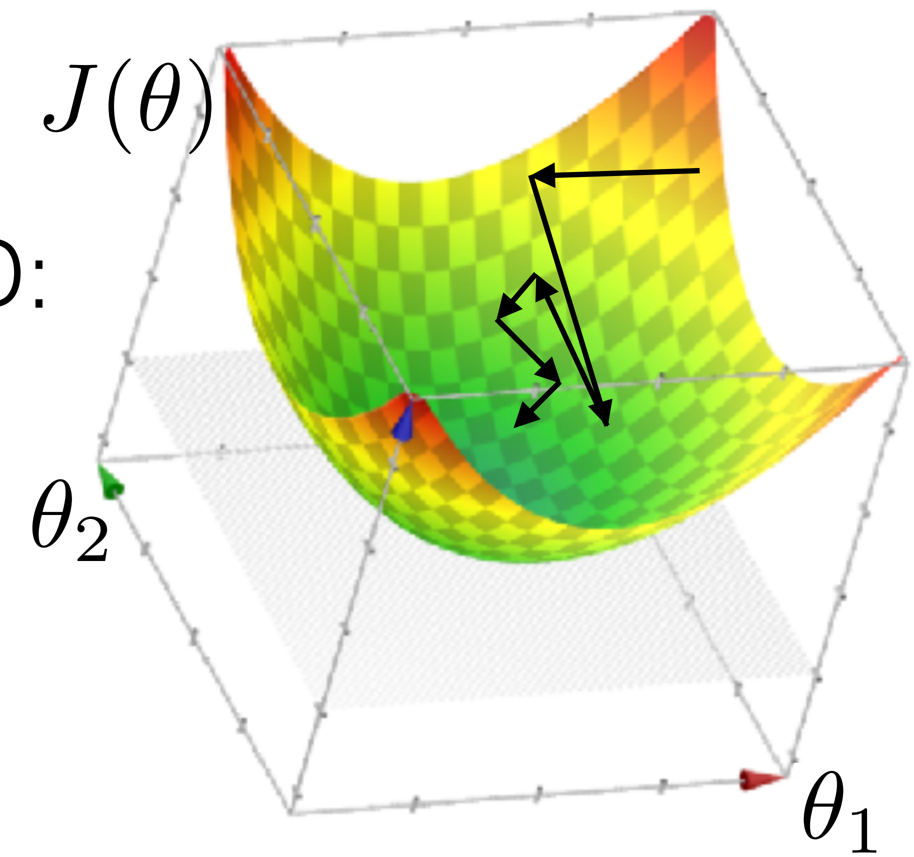
Learning the parameters

- Objective: $J(W, W_0) = \dfrac{1}{n_{\text{data}}} \displaystyle\sum_{i=1}^{n_{\text{data}}} L(h(x^{(i)}; W, W_0), y^{(i)})$

- GD:



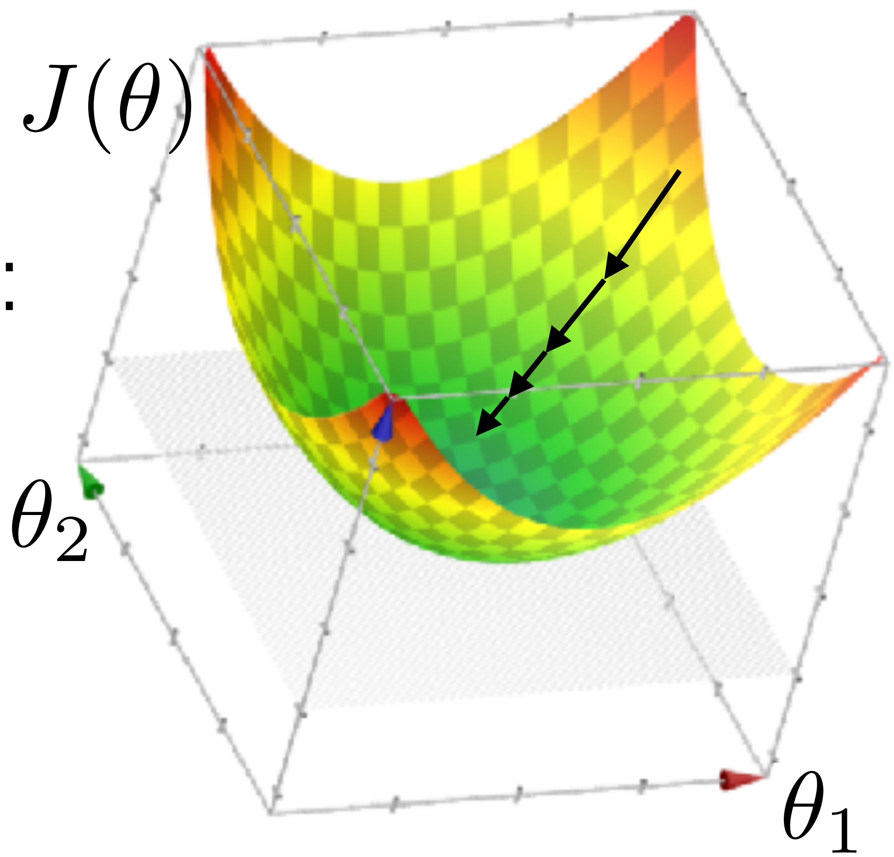
- SGD:



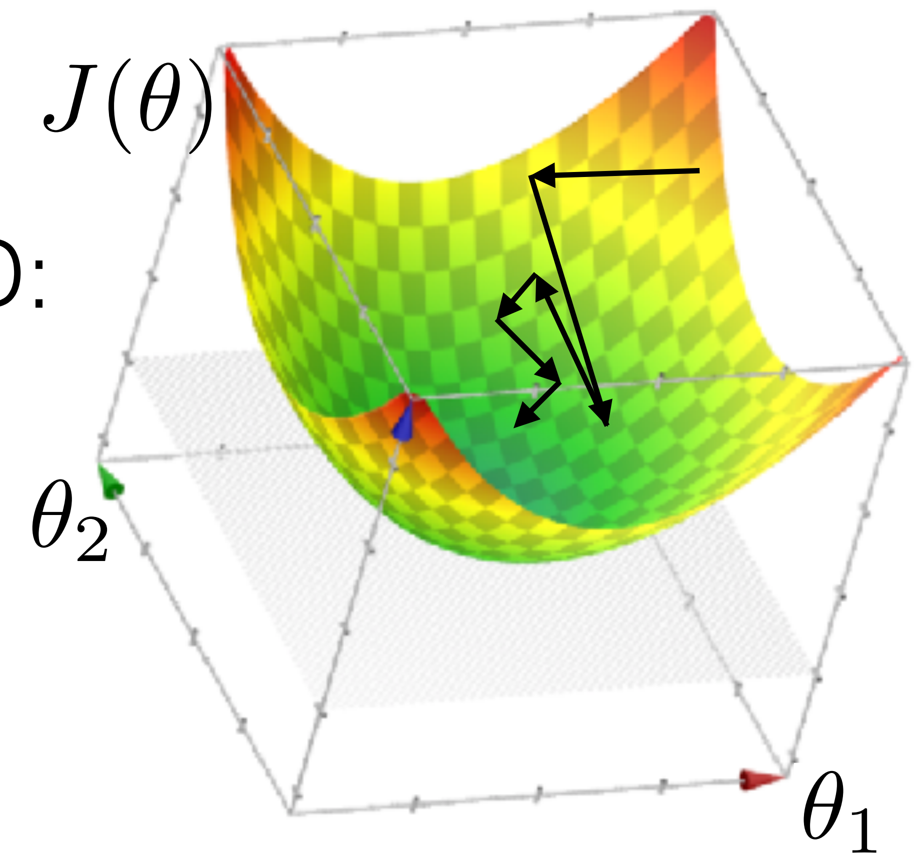
- **Theorem**: (Roughly) if the objective is nice and convex, GD and SGD perform well

Learning the parameters

- Objective: $J(W, W_0) = \frac{1}{n_{\text{data}}} \sum_{i=1}^{n_{\text{data}}} L(h(x^{(i)}; W, W_0), y^{(i)})$
- GD:



- SGD:

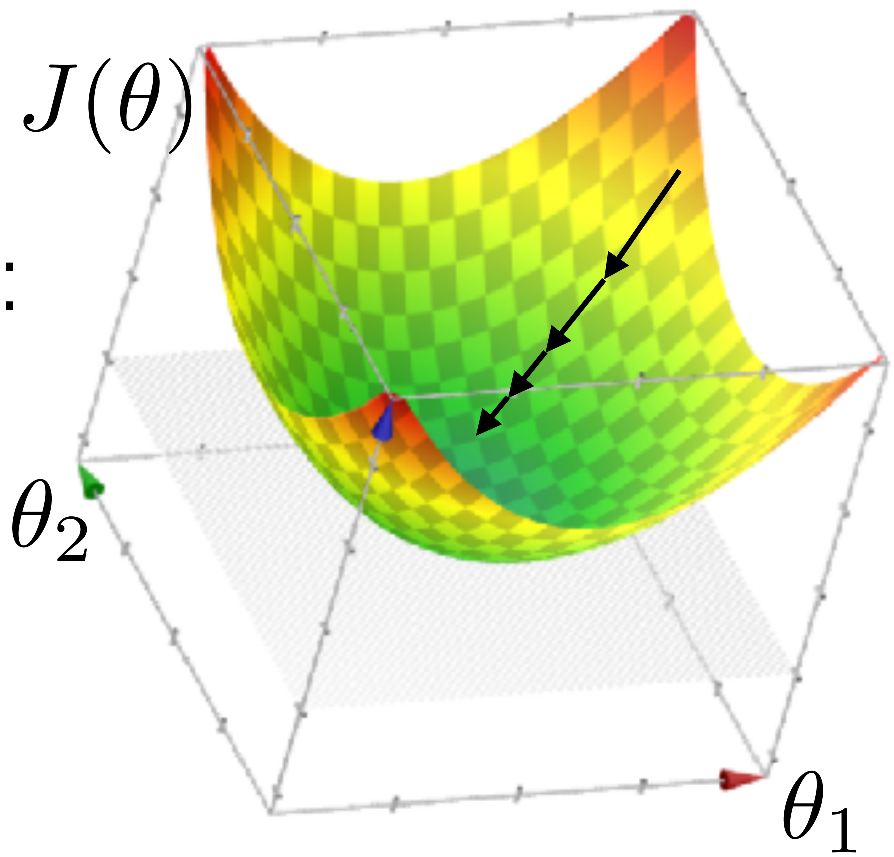


- **Theorem**: (Roughly) if the objective is nice and convex, GD and SGD perform well
- **Big challenge**: the NN objective is a (very) non-convex function of the parameters (except in e.g. 1 layer)

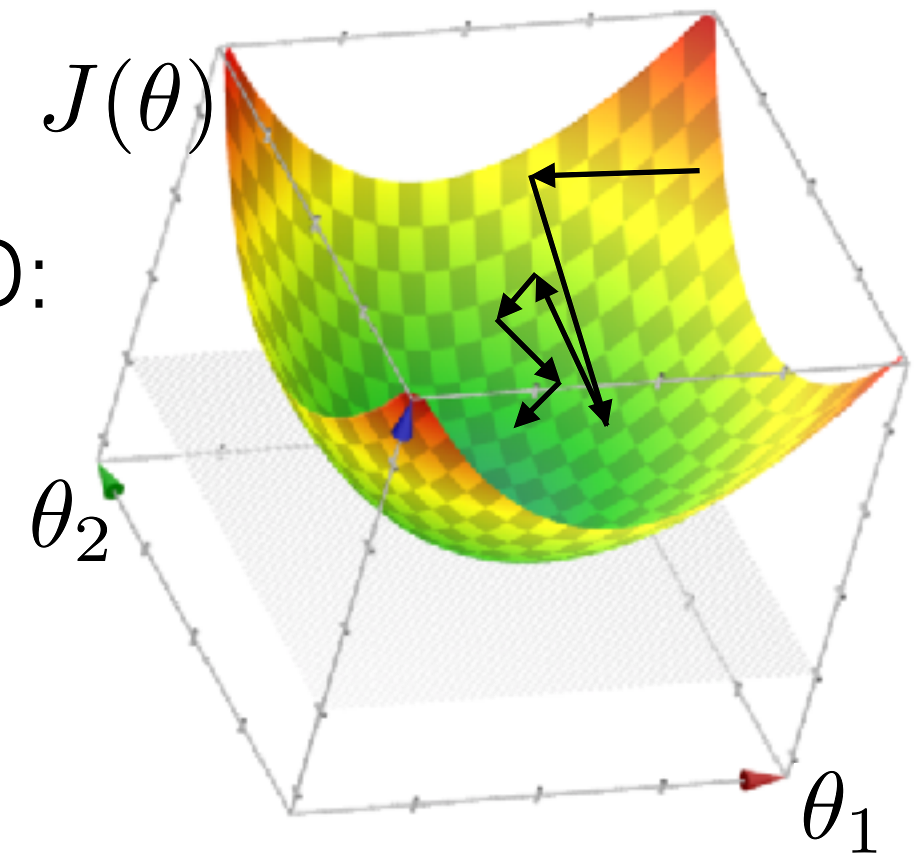
Learning the parameters

- Objective: $J(W, W_0) = \frac{1}{n_{\text{data}}} \sum_{i=1}^{n_{\text{data}}} L(h(x^{(i)}; W, W_0), y^{(i)})$

- GD:



- SGD:



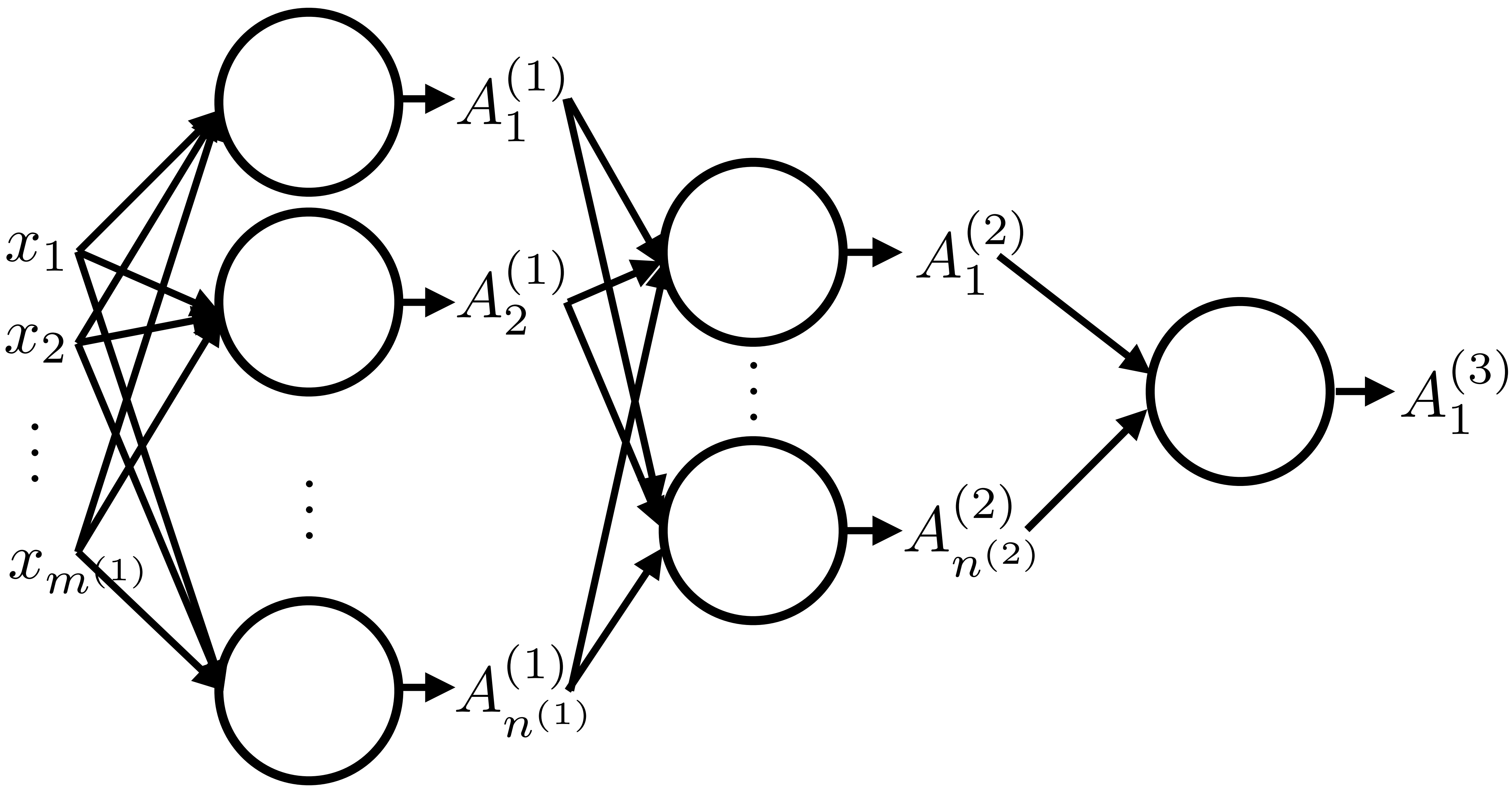
- **Theorem**: (Roughly) if the objective is nice and convex, GD and SGD perform well
- **Big challenge**: the NN objective is a (very) non-convex function of the parameters (except in e.g. 1 layer)
    - Huge bag of tricks to optimize / regularize

More layers!

- Why stop at 2 layers?

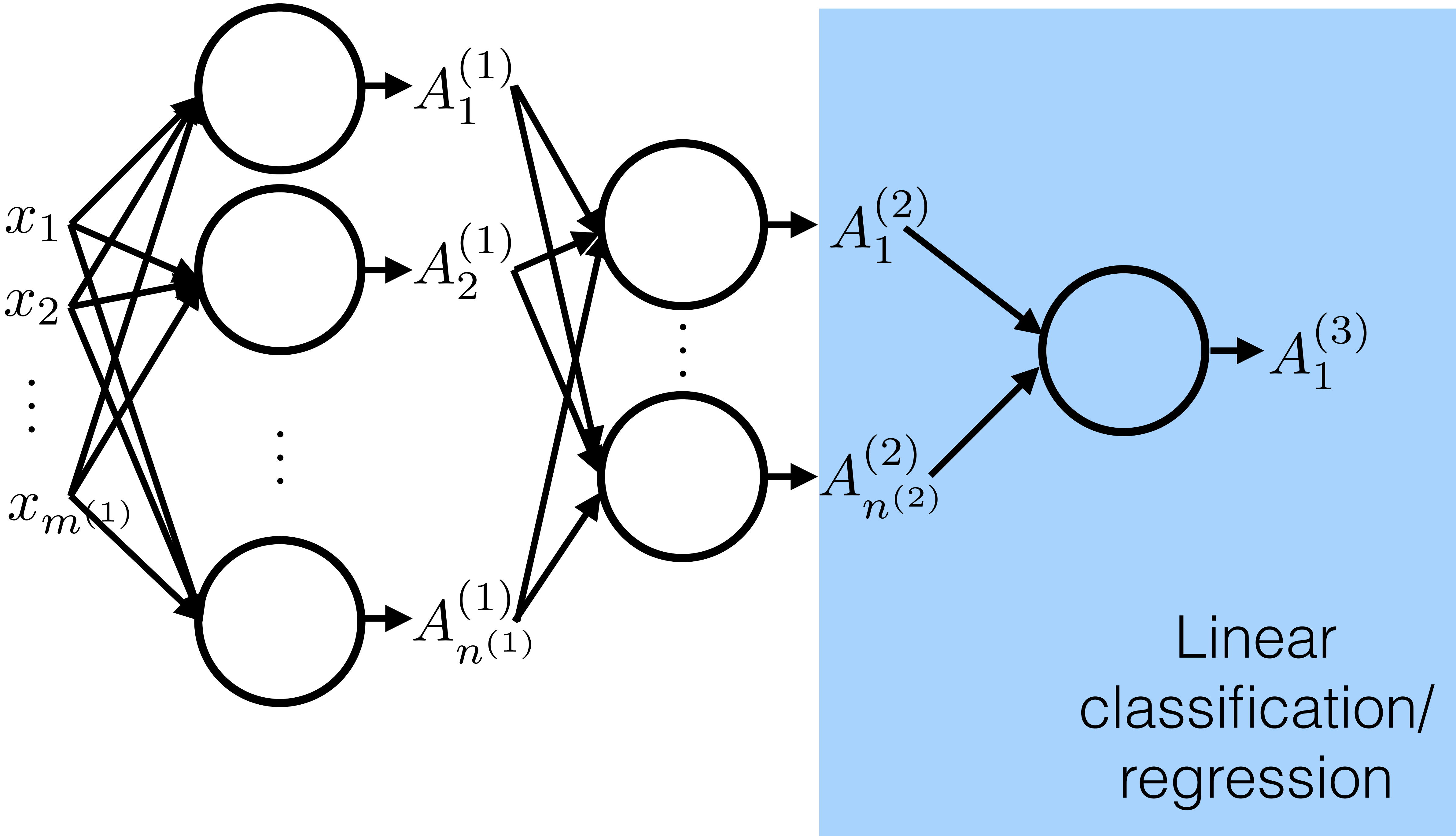
More layers!

- Why stop at 2 layers?



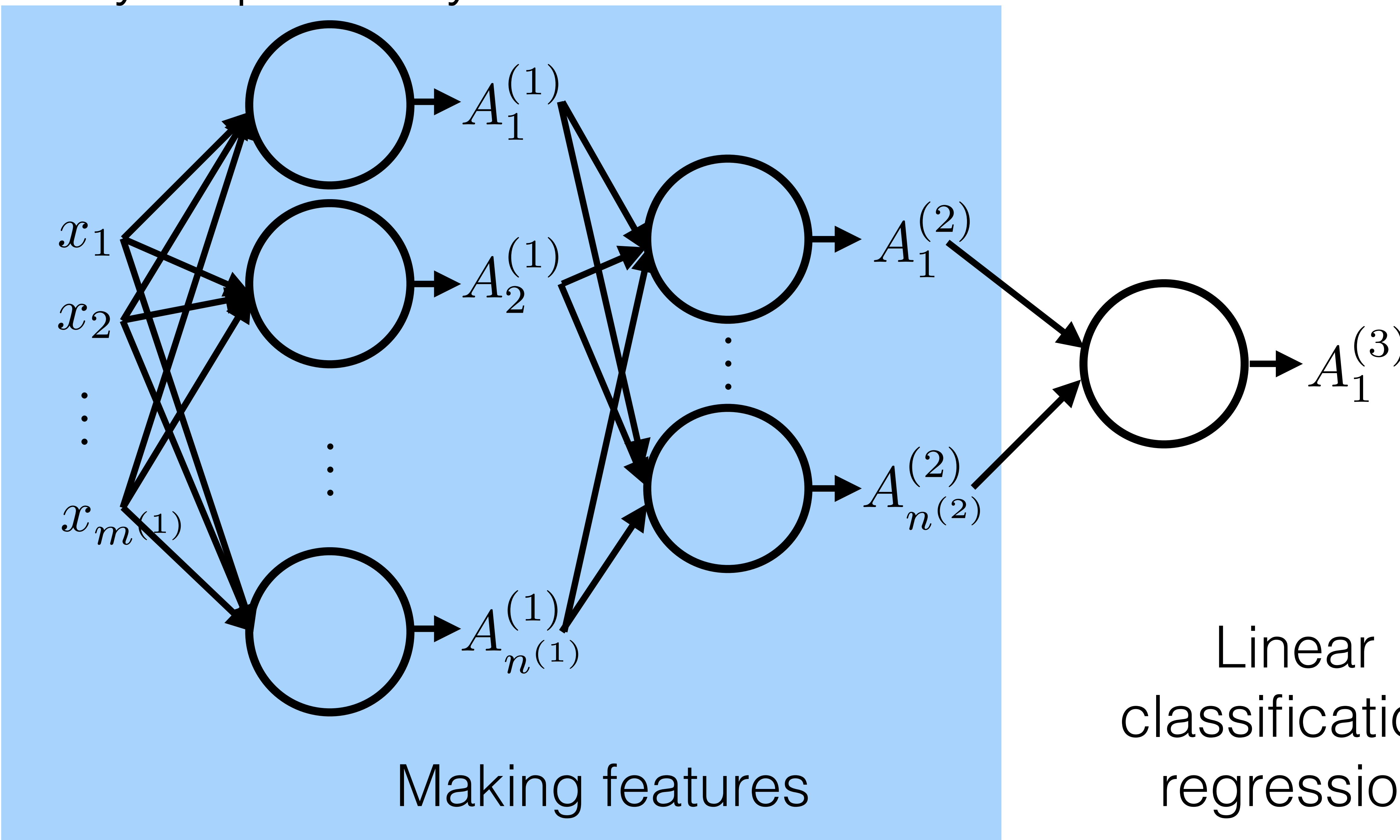
More layers!

- Why stop at 2 layers?



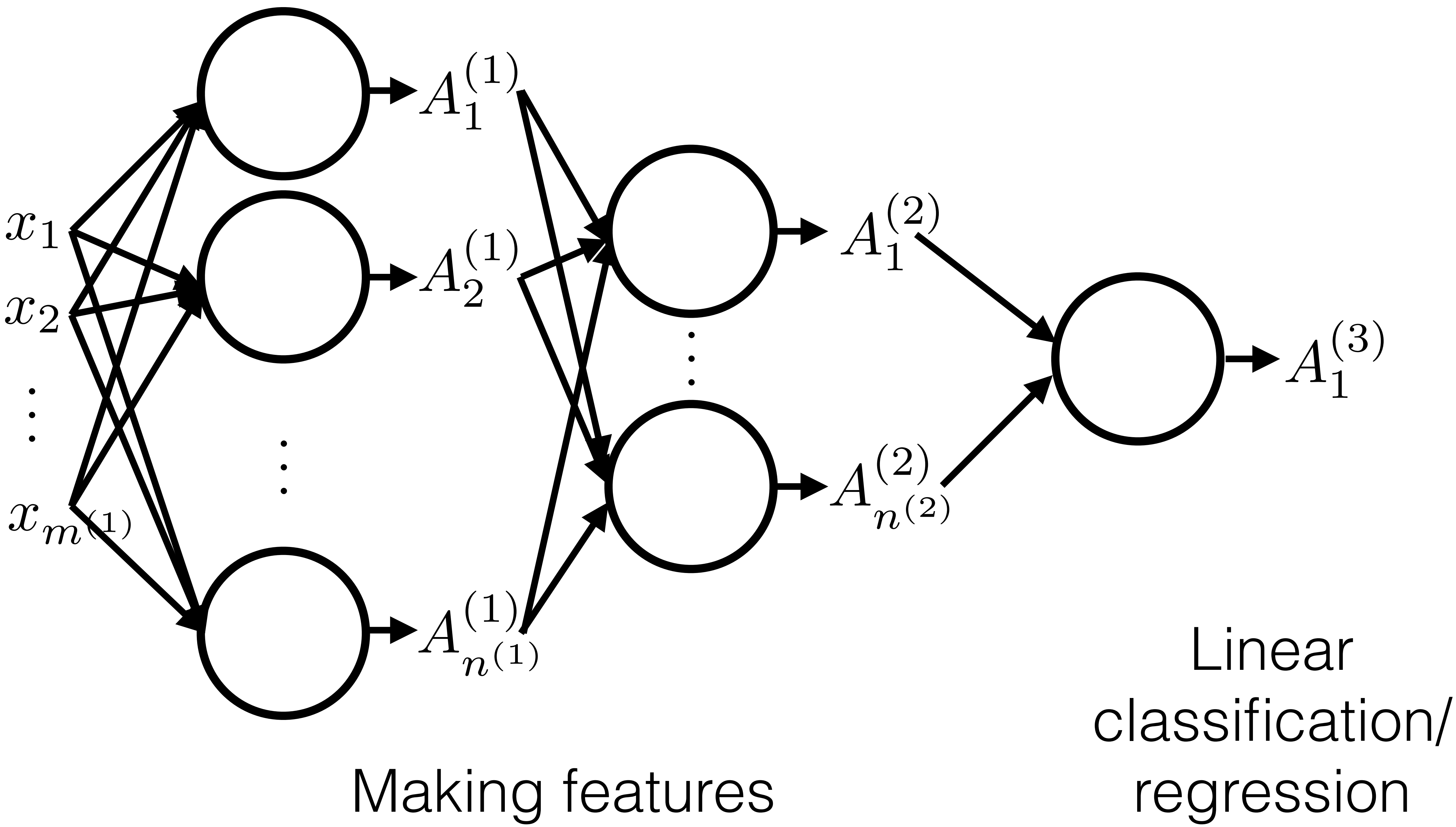
More layers!

- Why stop at 2 layers?



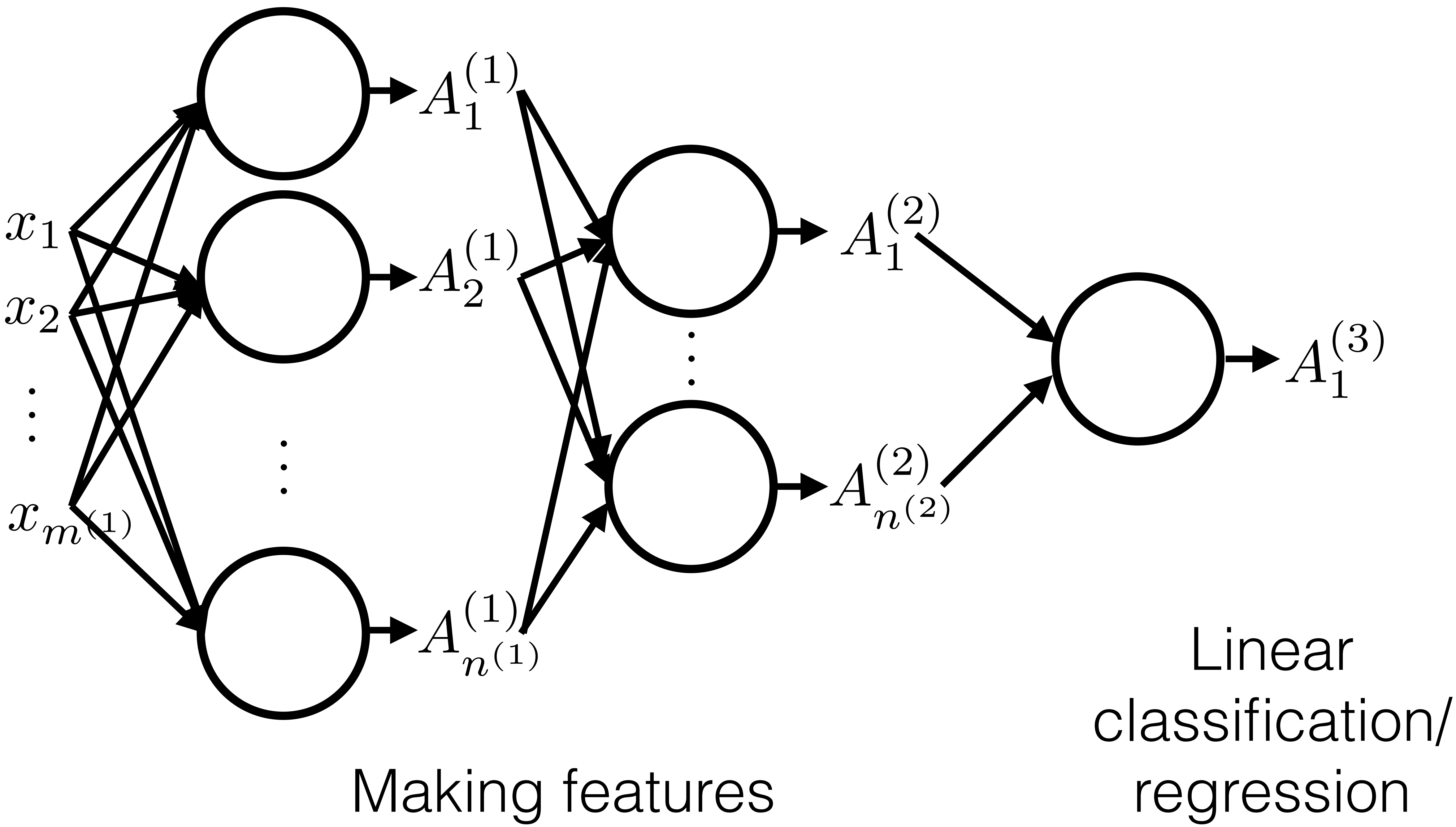
More layers!

- Why stop at 2 layers?



More layers!

- Why stop at 2 layers?



- Just one layer: linear classification/regression with default features