# 6.036: Introduction to Machine Learning

Tues 10/19 lecture time → midterm review! We'll continue neural net fun on 10/26.

**Lecture start:** Tuesdays 9:35am

**Who's talking?** Prof. Tamara Broderick

**Questions?** Ask on Piazza: "lecture (week) 6" folder

**Materials:** slides, video will all be available on Canvas

**Live Zoom feed:** https://mit.zoom.us/j/94238622313

## Last Time(s)

I. Linear regression

II. Linear classification

III. Choosing features

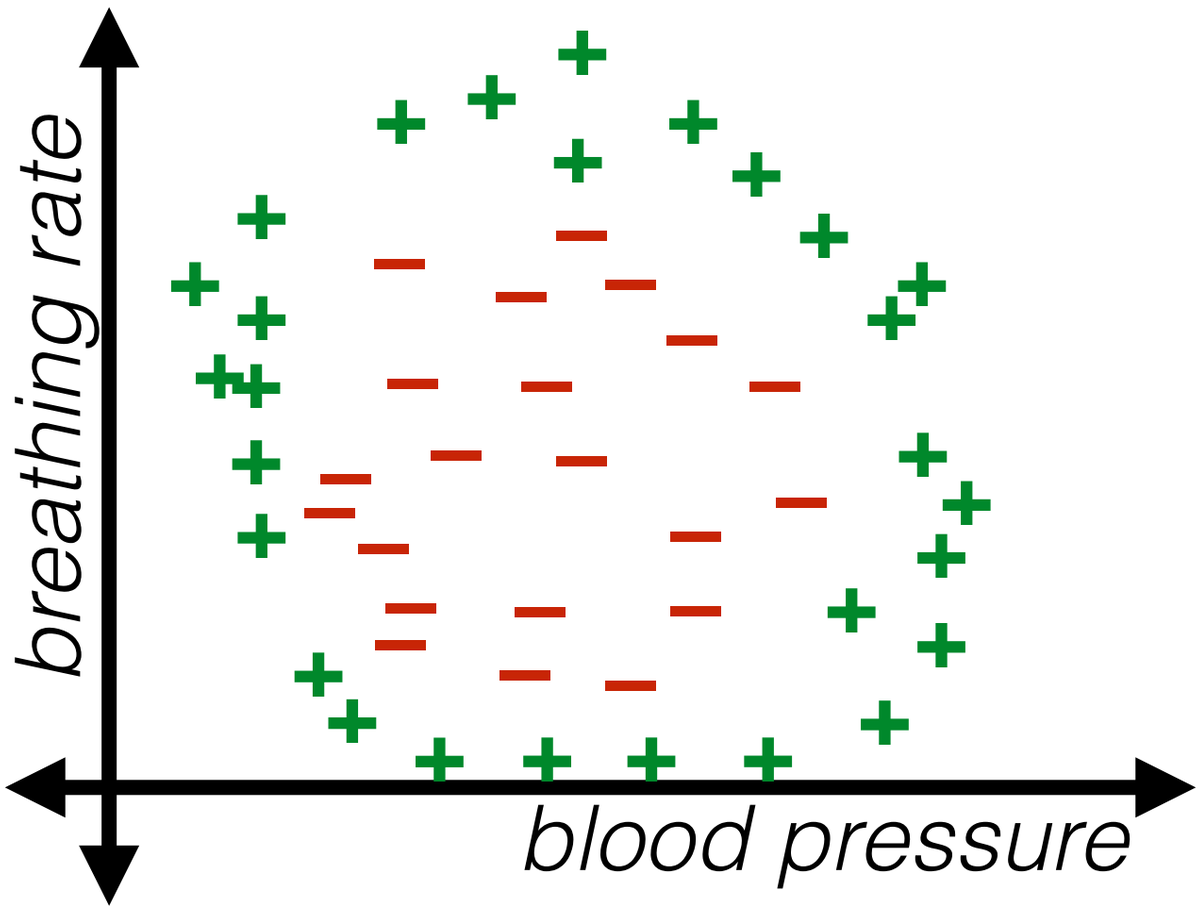
## Today's Plan

I. Polynomial features

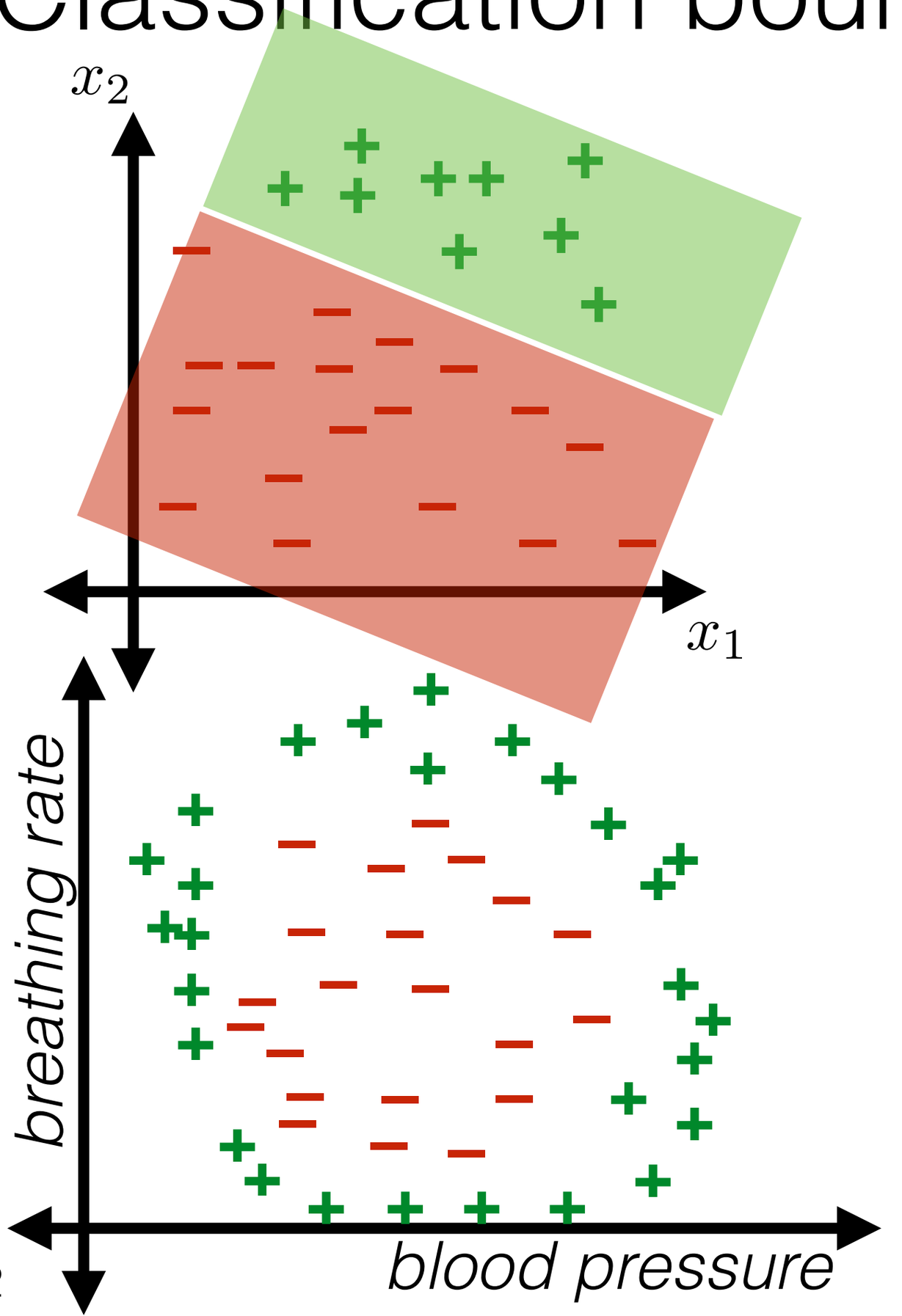
II. Step-function features

III. Neural nets: hypothesis class

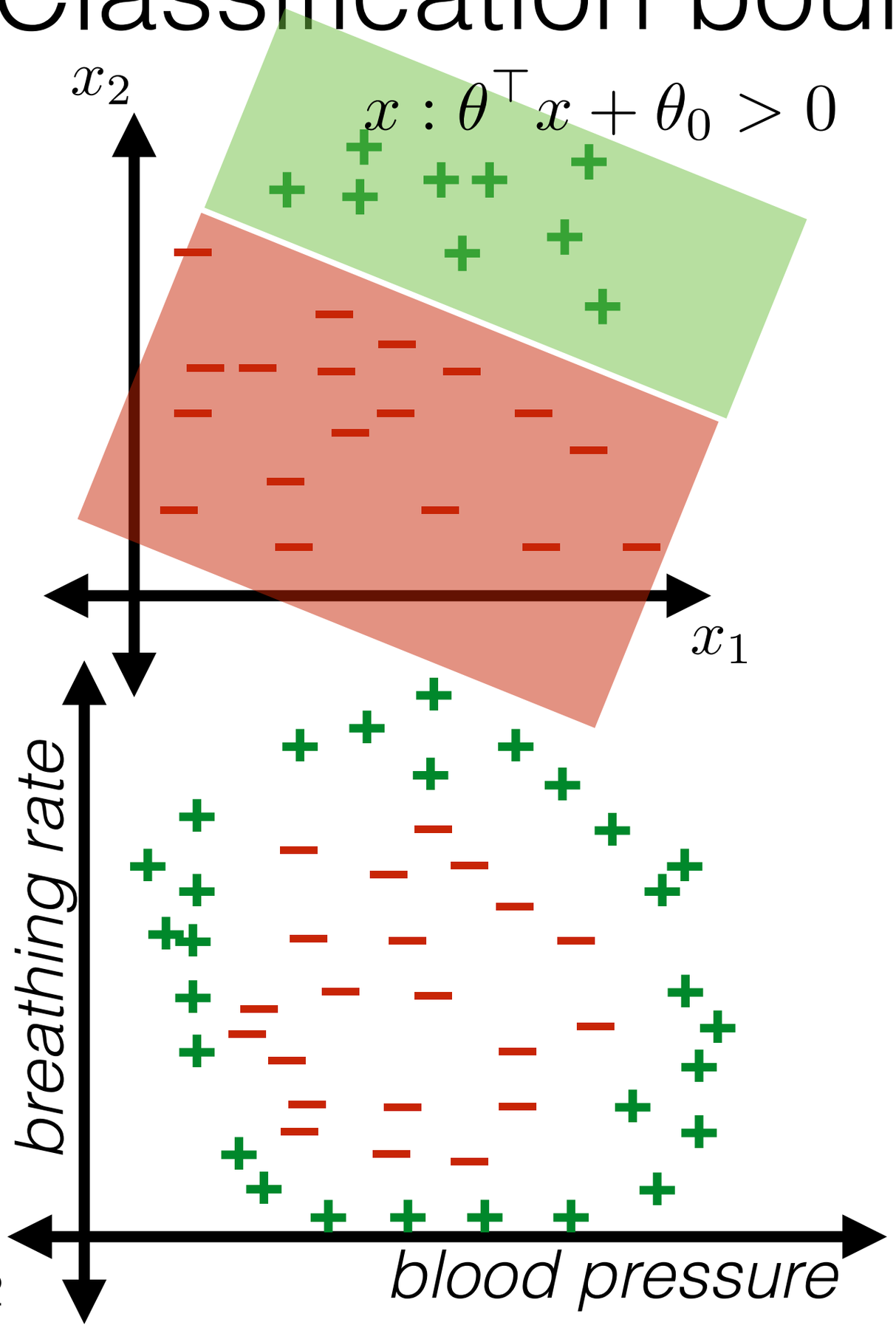
# Classification boundaries



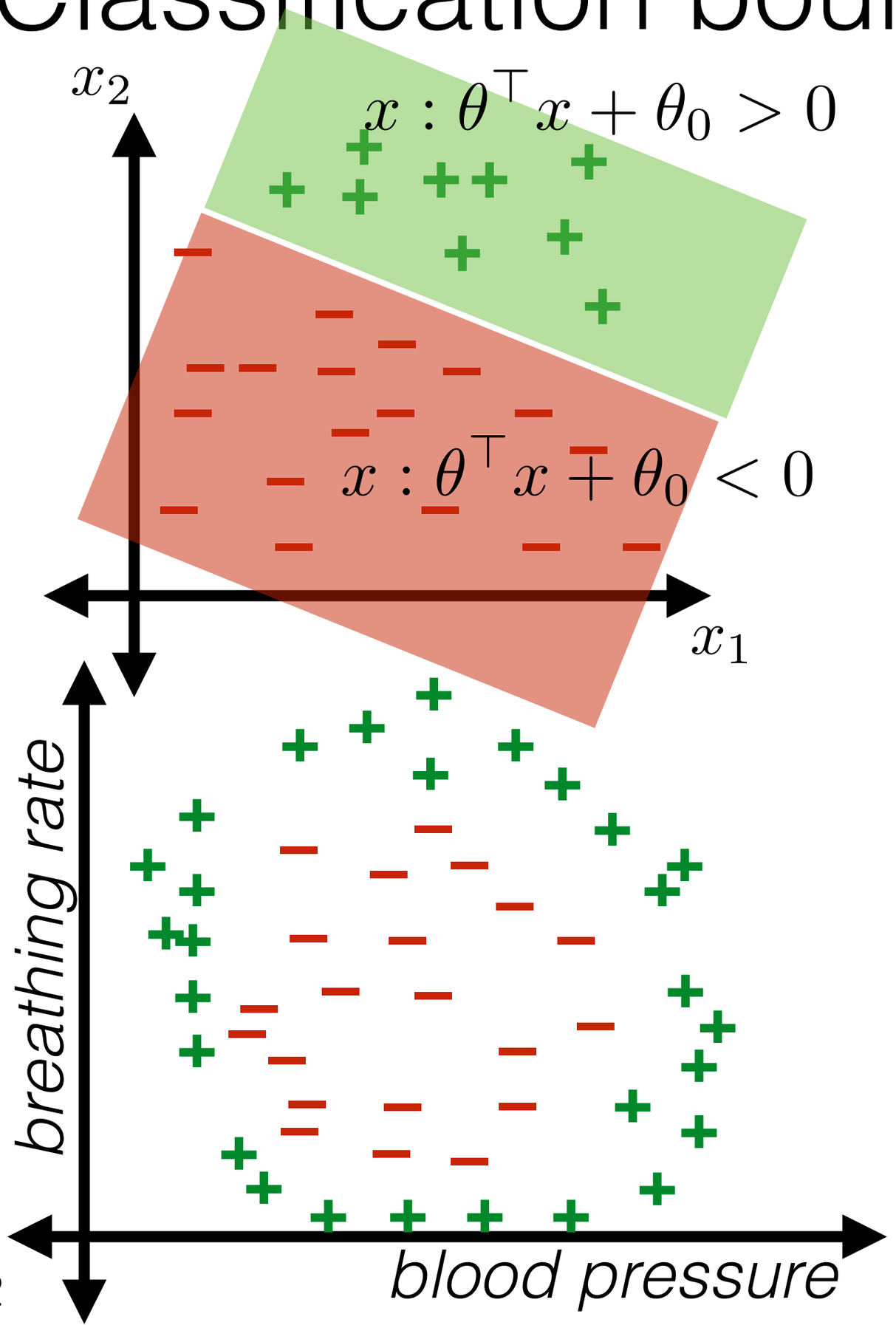
# Classification boundaries



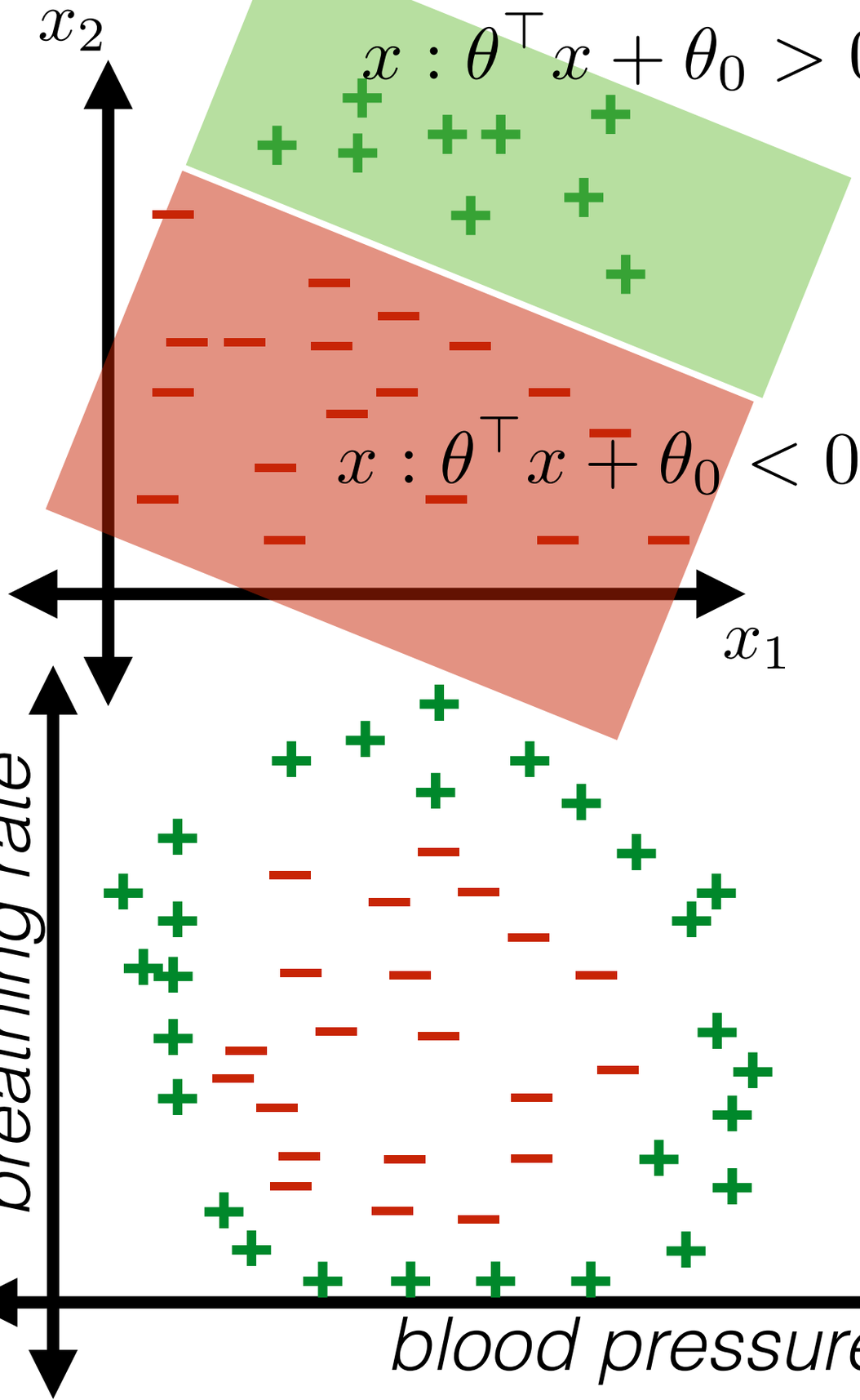
# Classification boundaries

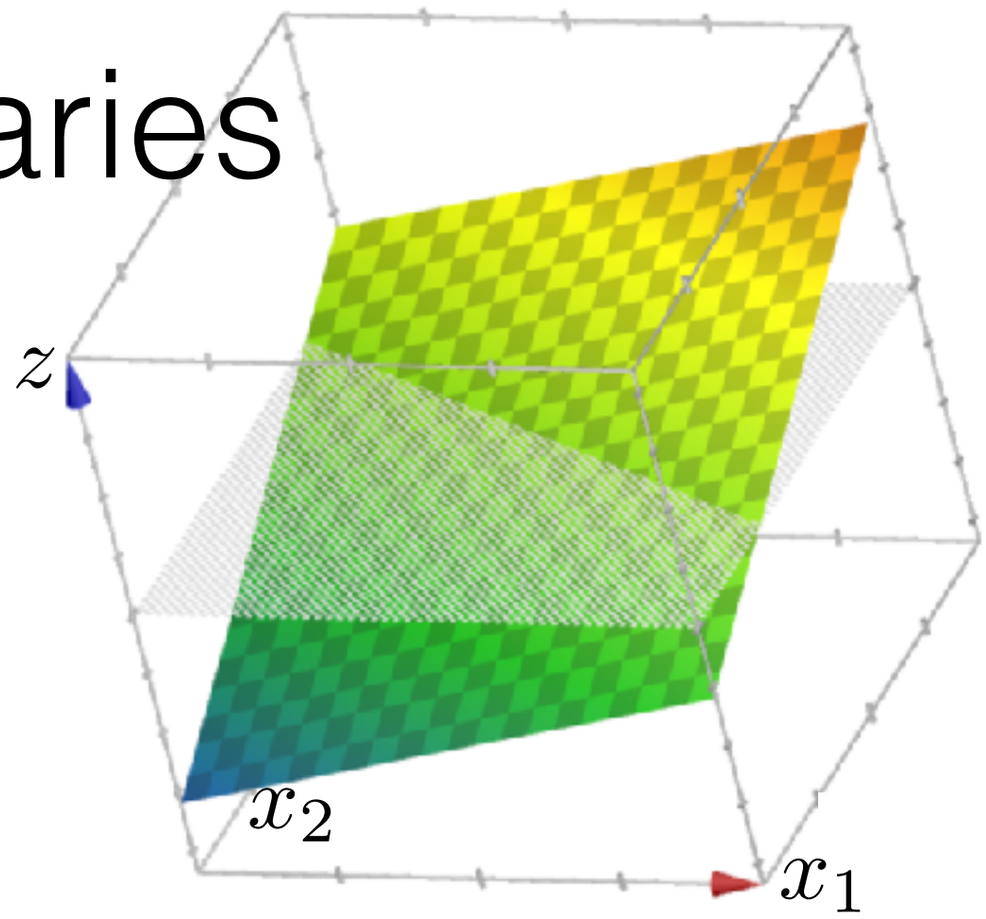


# Classification boundaries

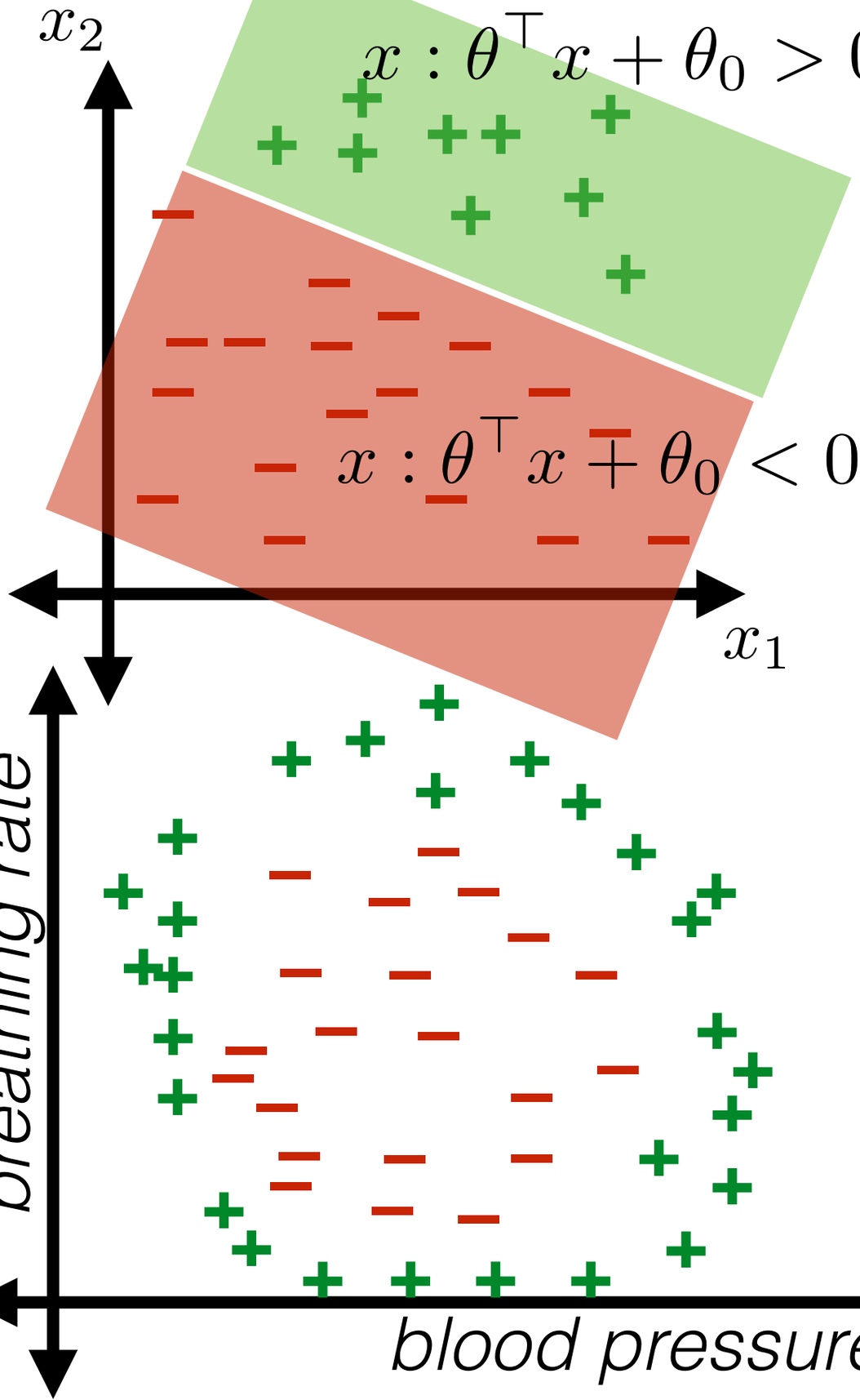


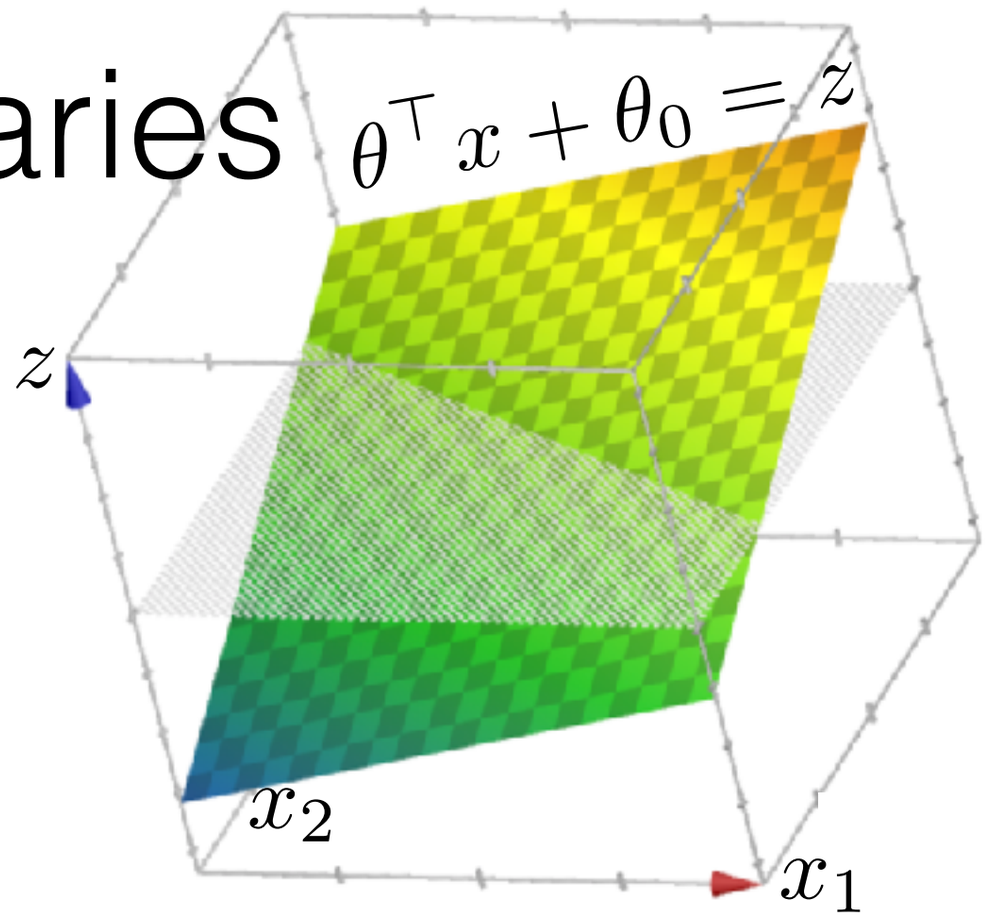
# Classification boundaries



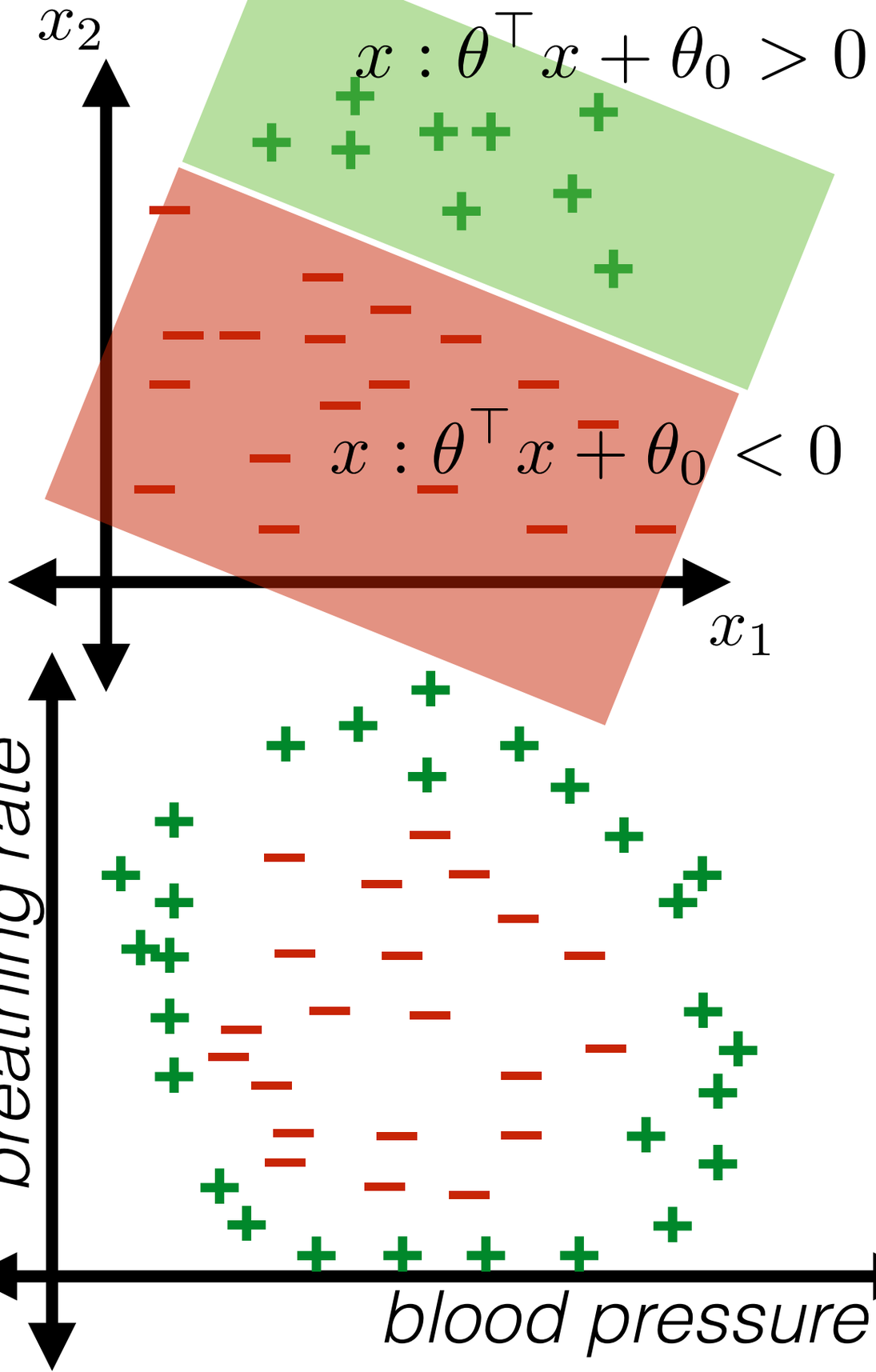



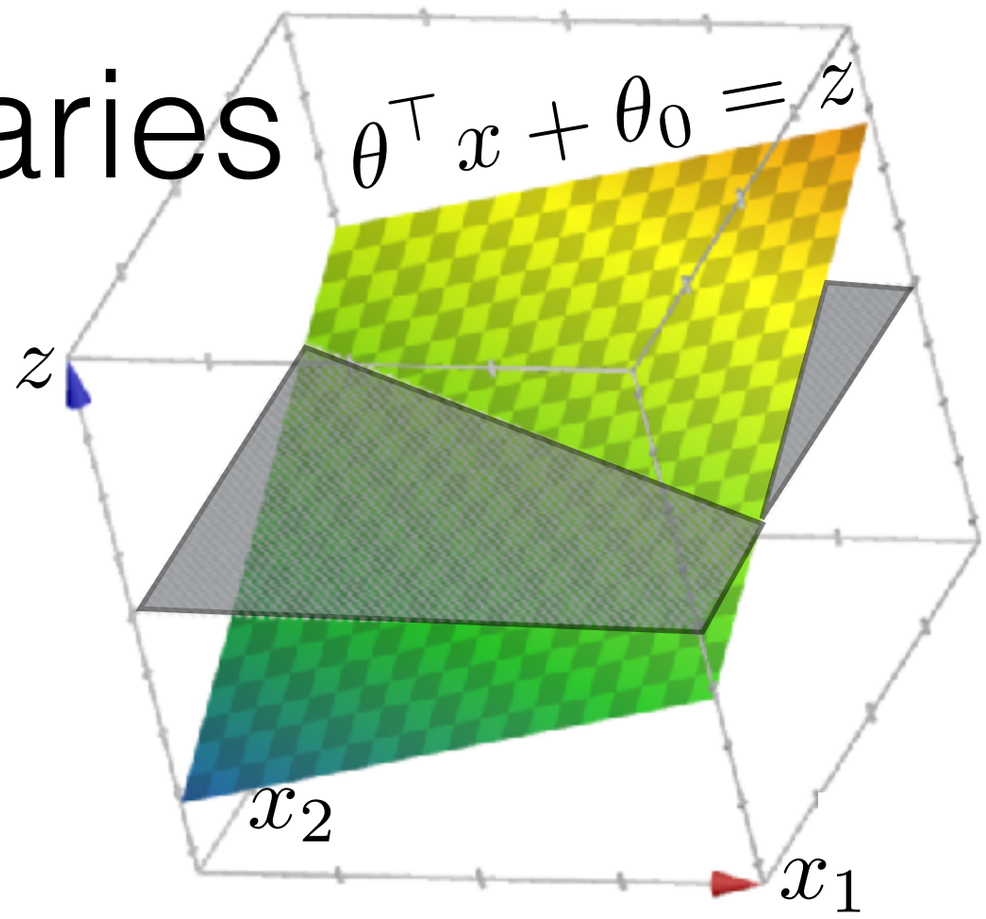
# Classification boundaries



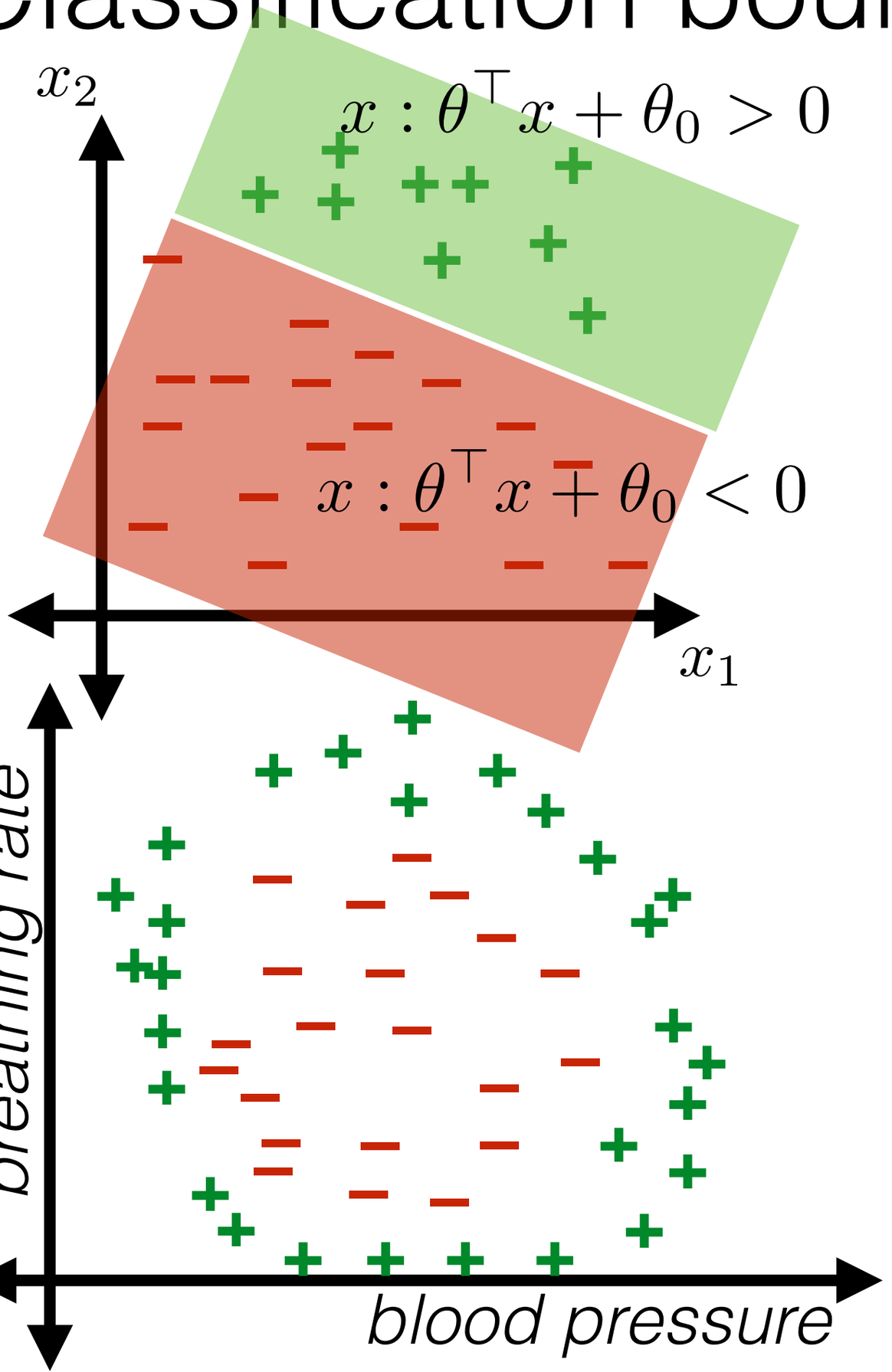



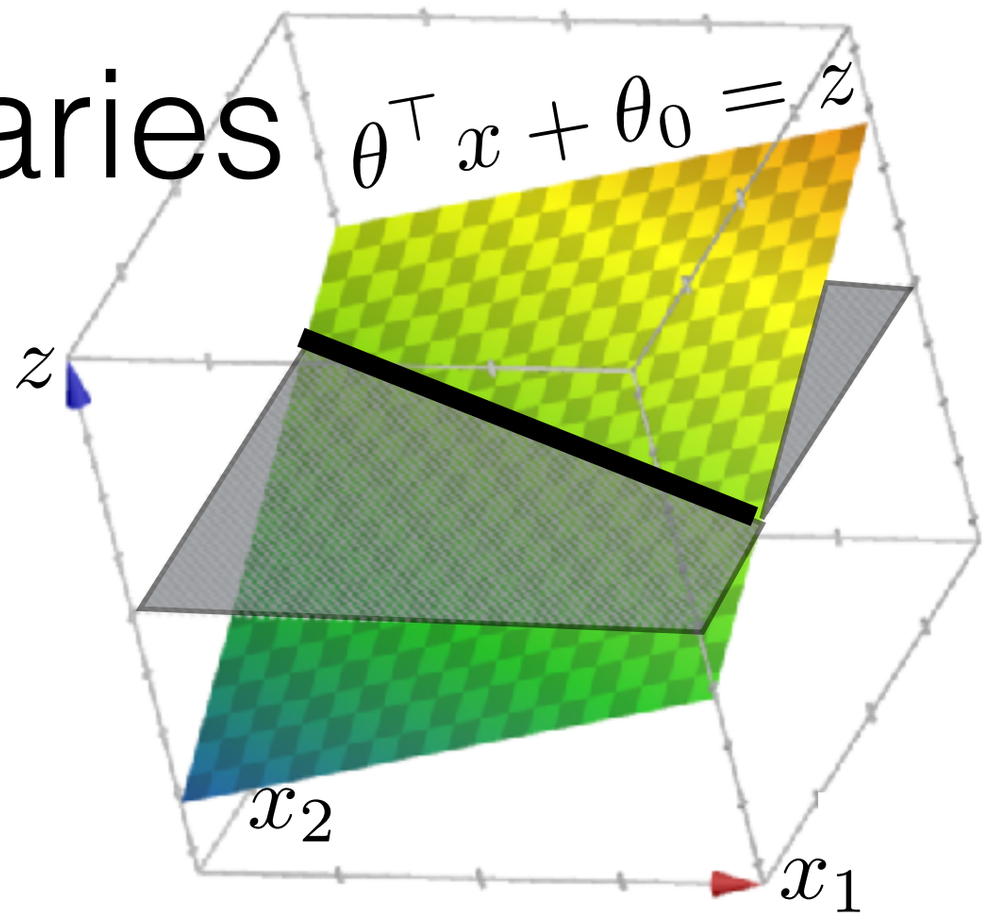
# Classification boundaries



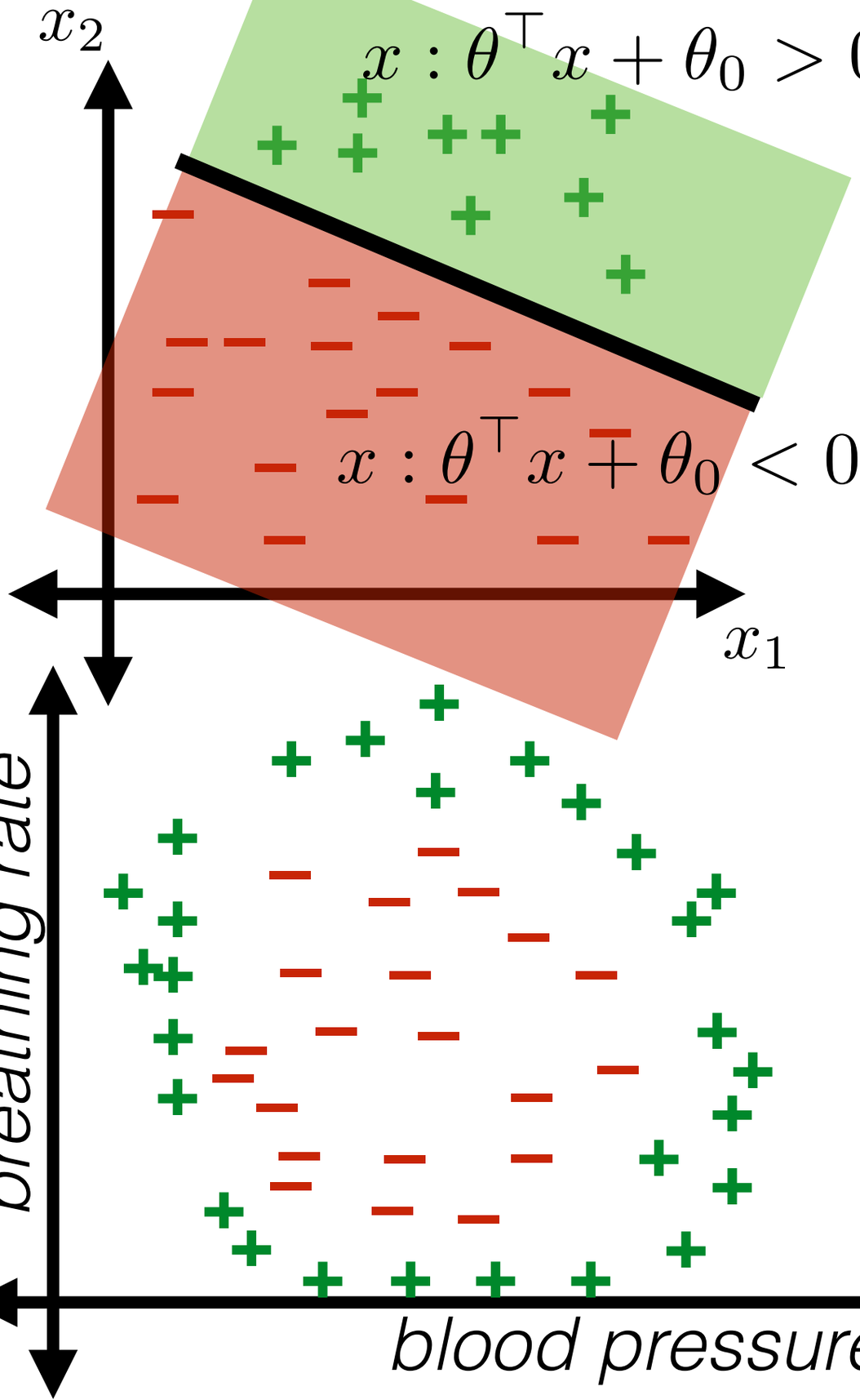



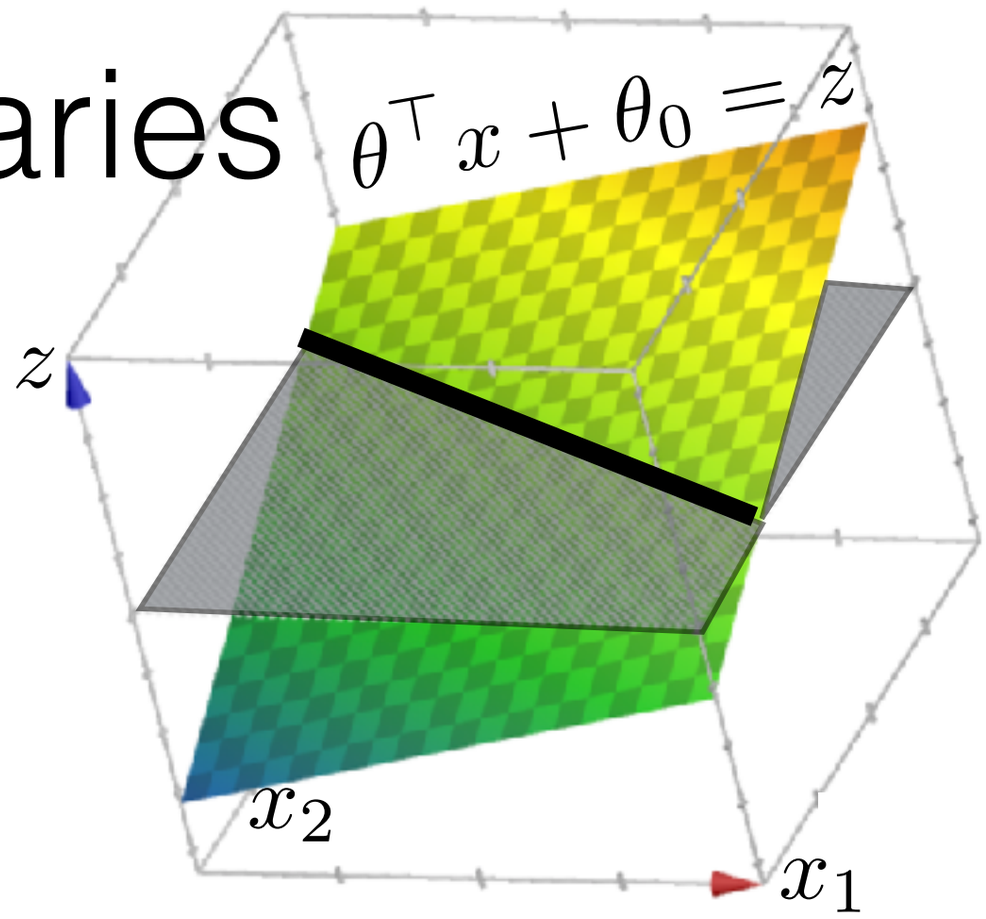
# Classification boundaries



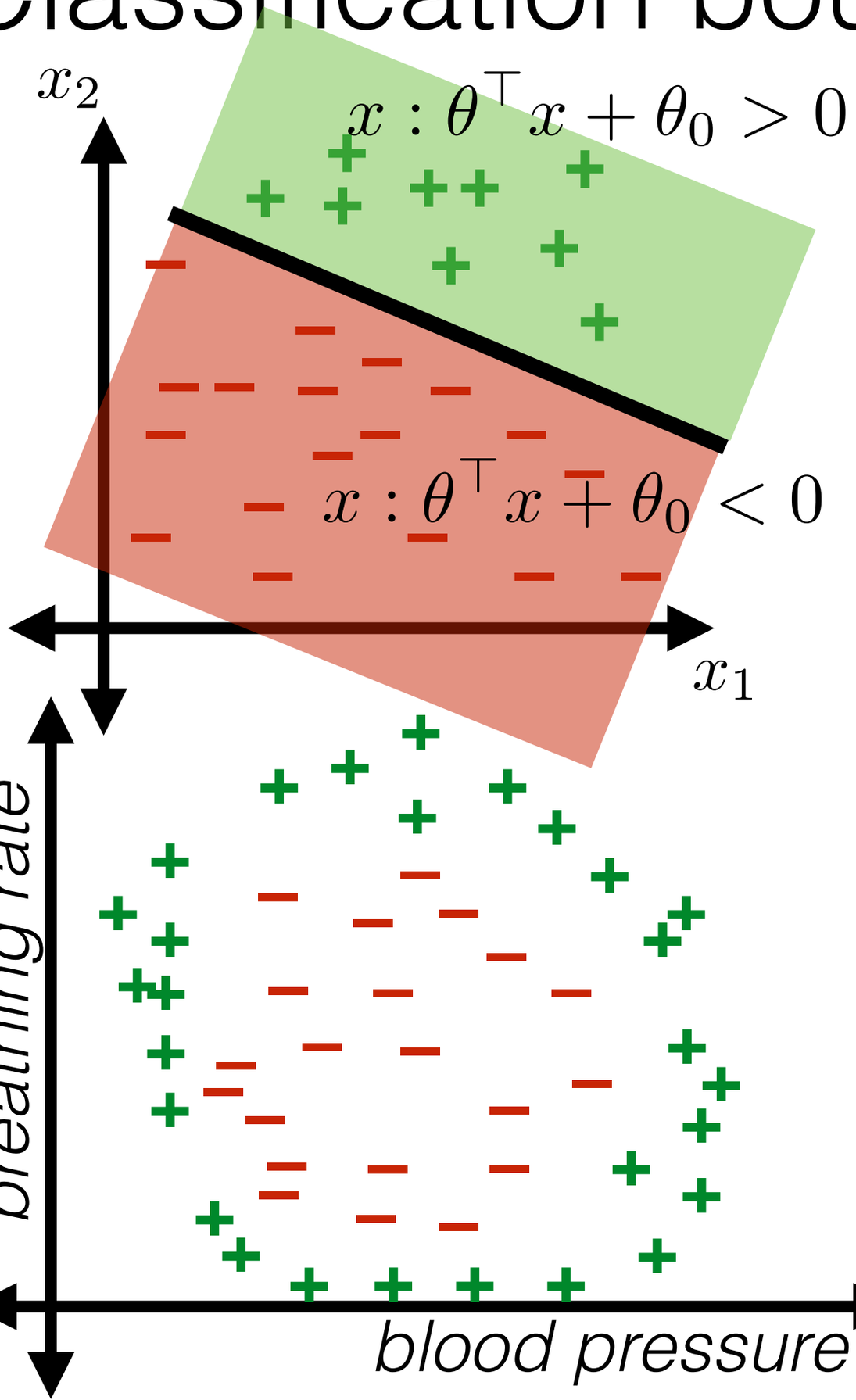



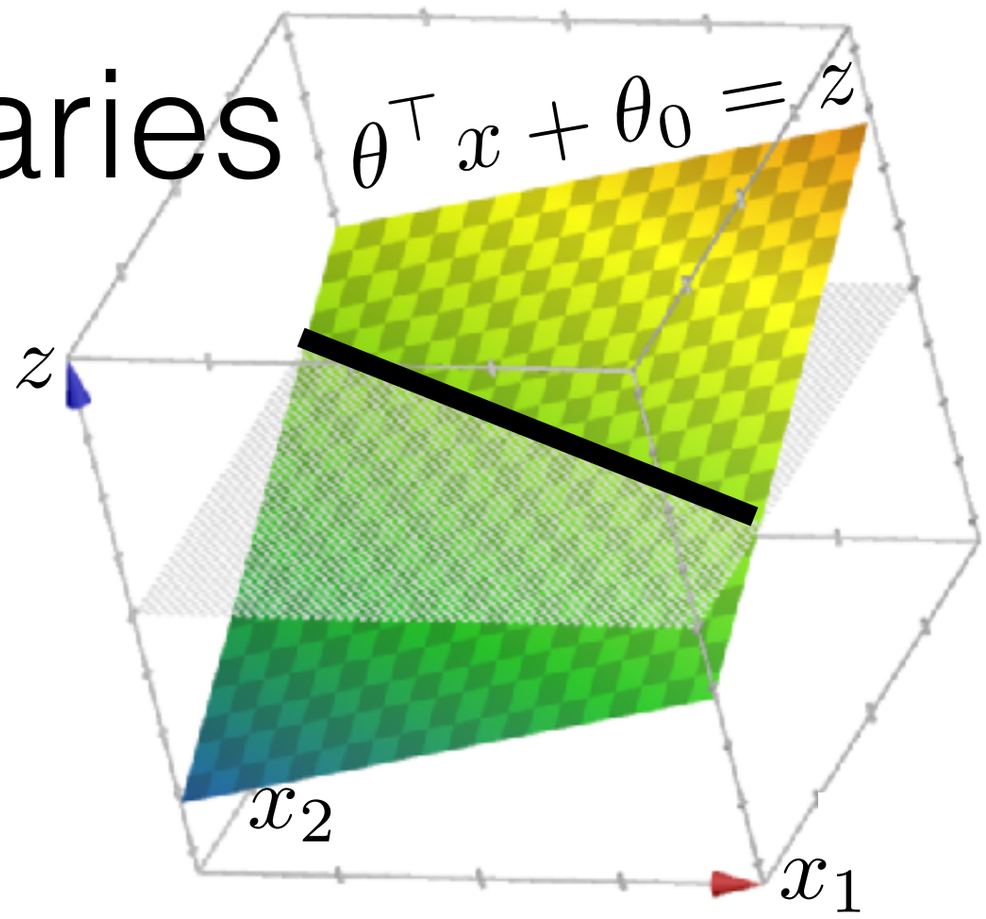
# Classification boundaries



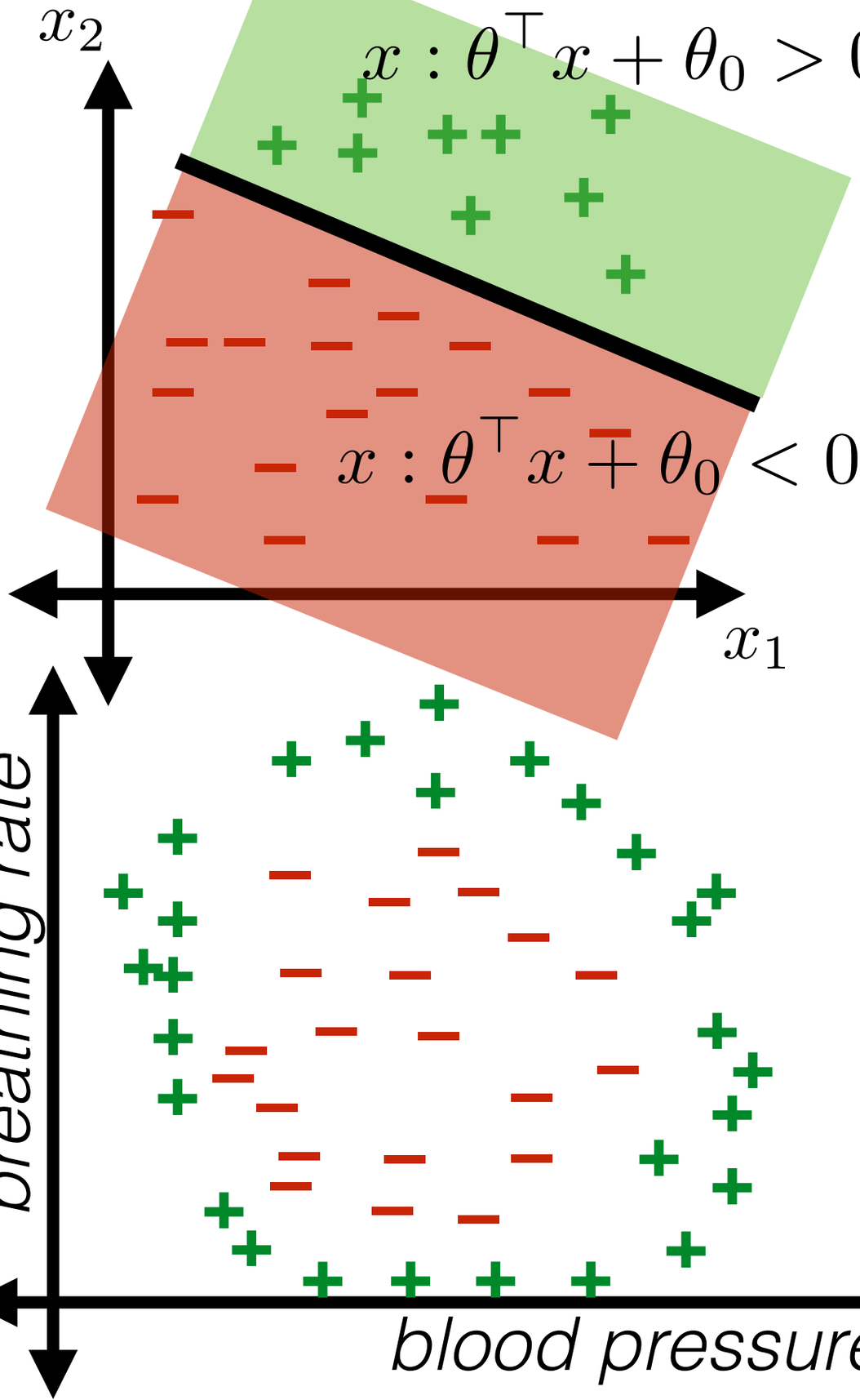



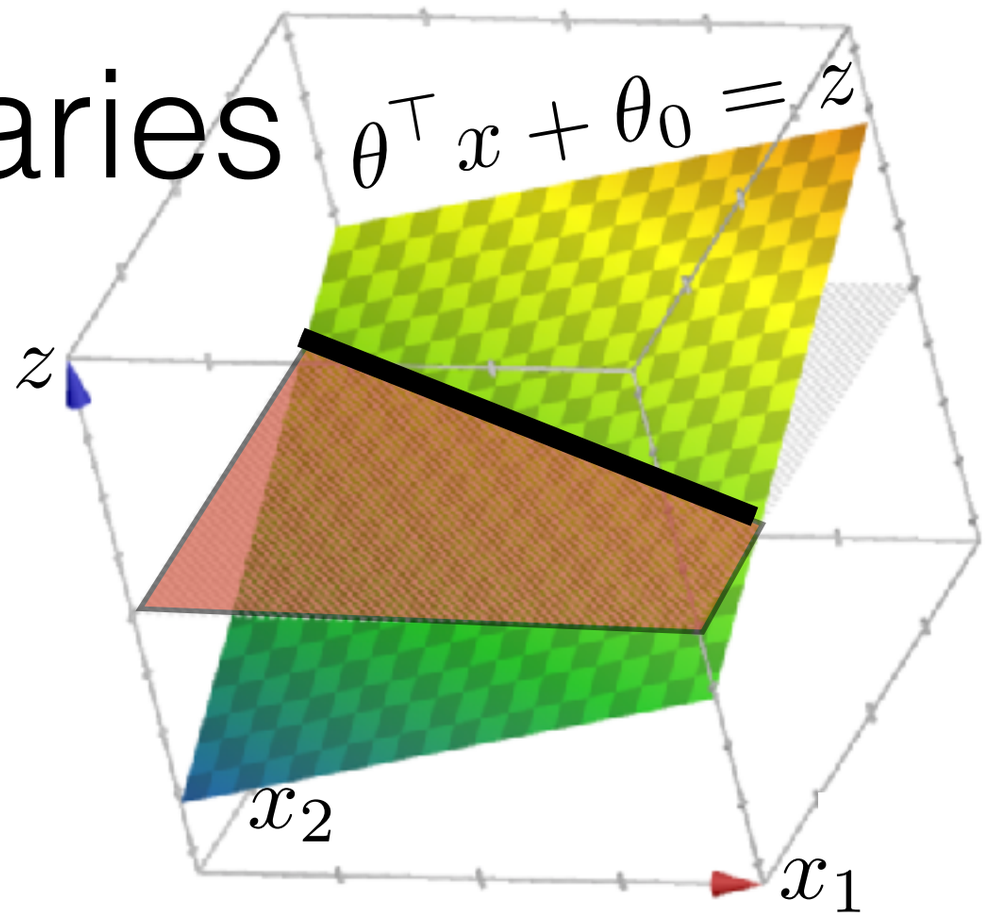
# Classification boundaries



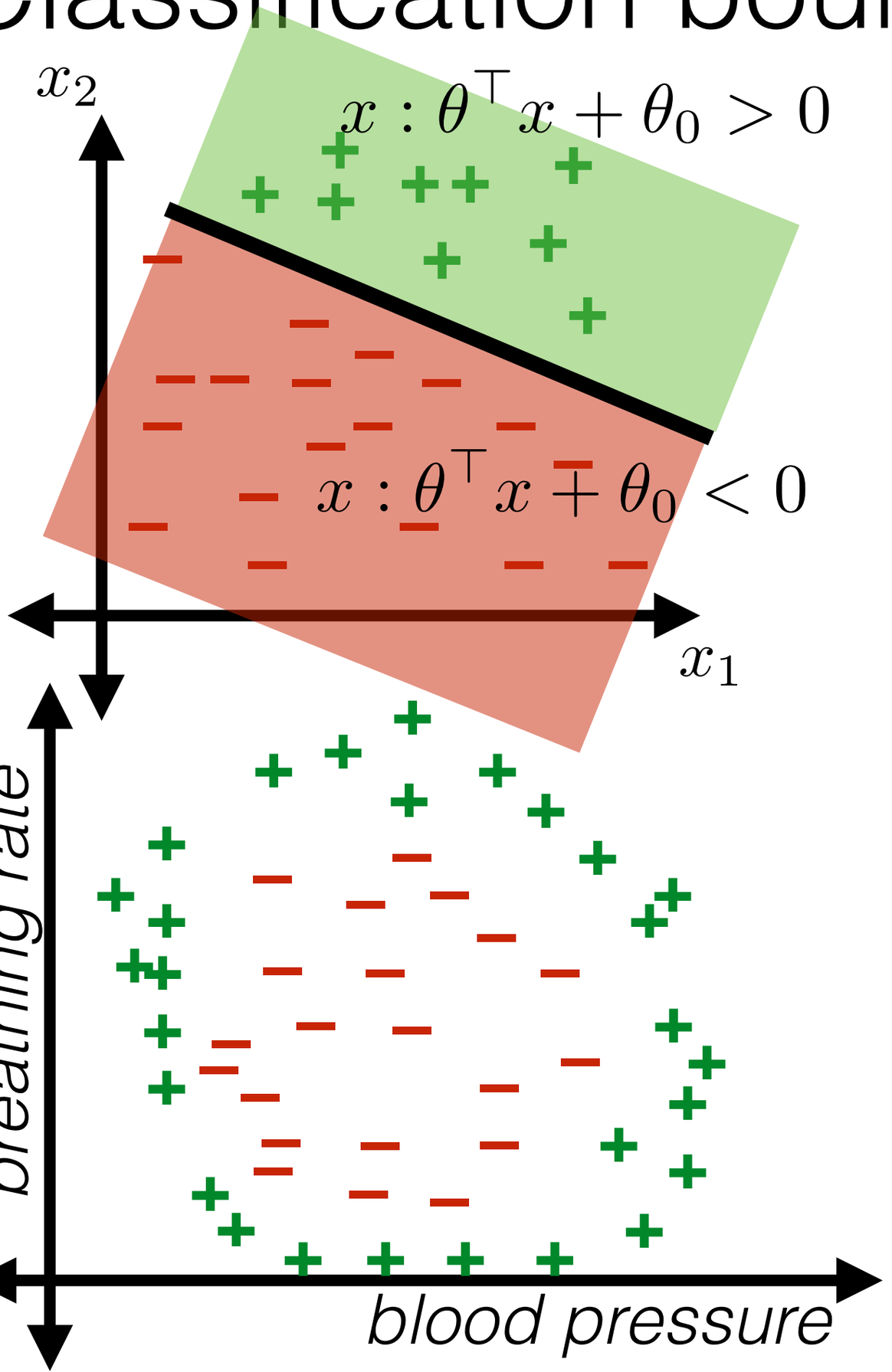



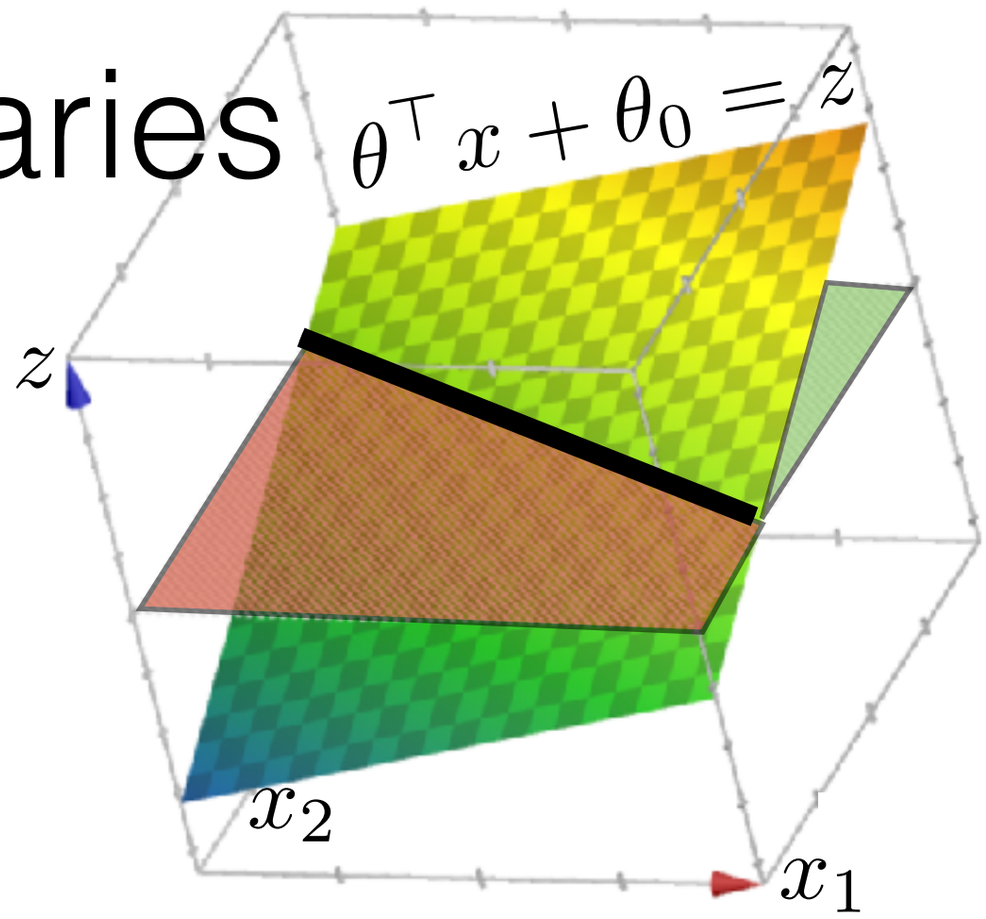
# Classification boundaries



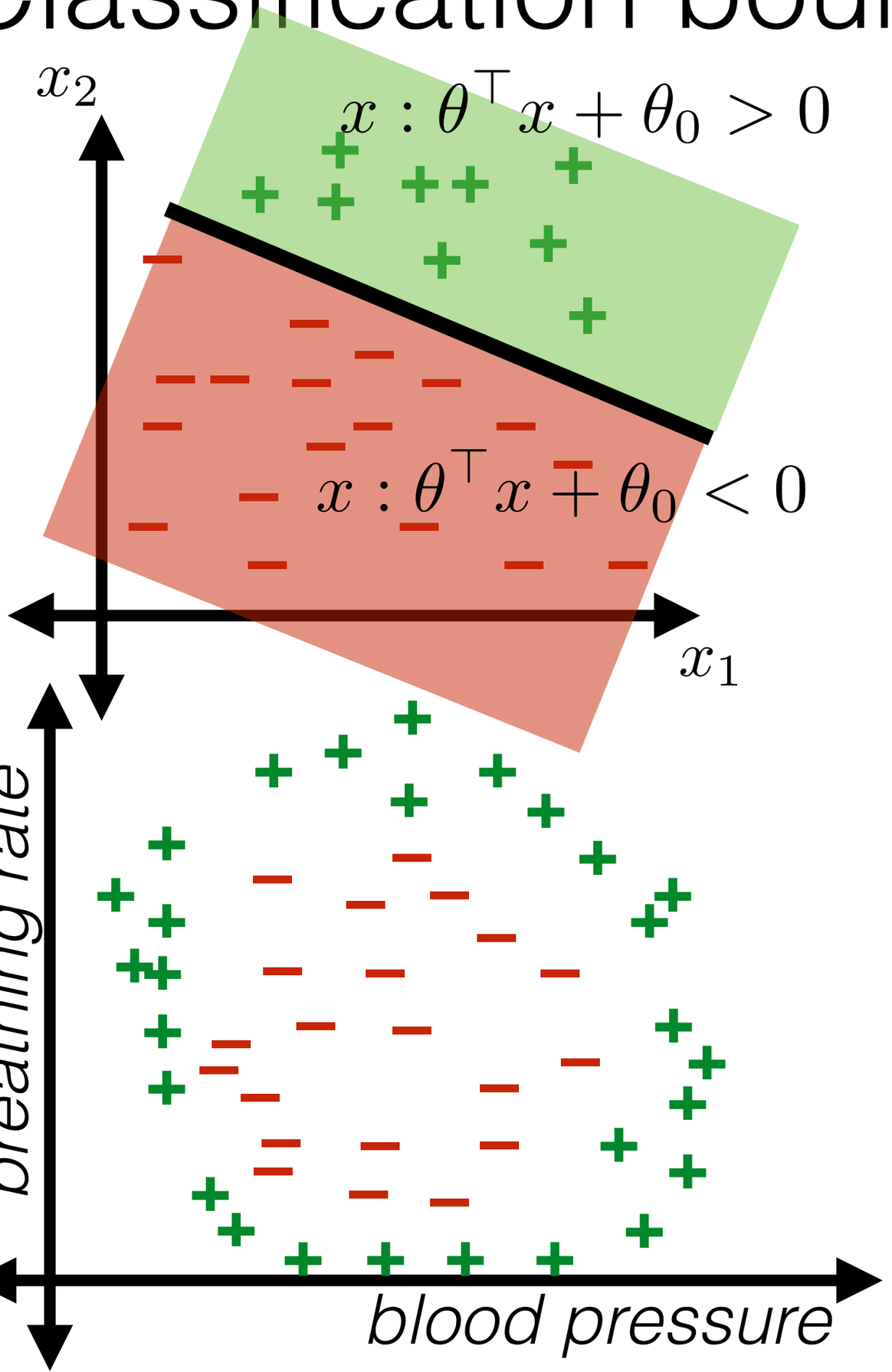



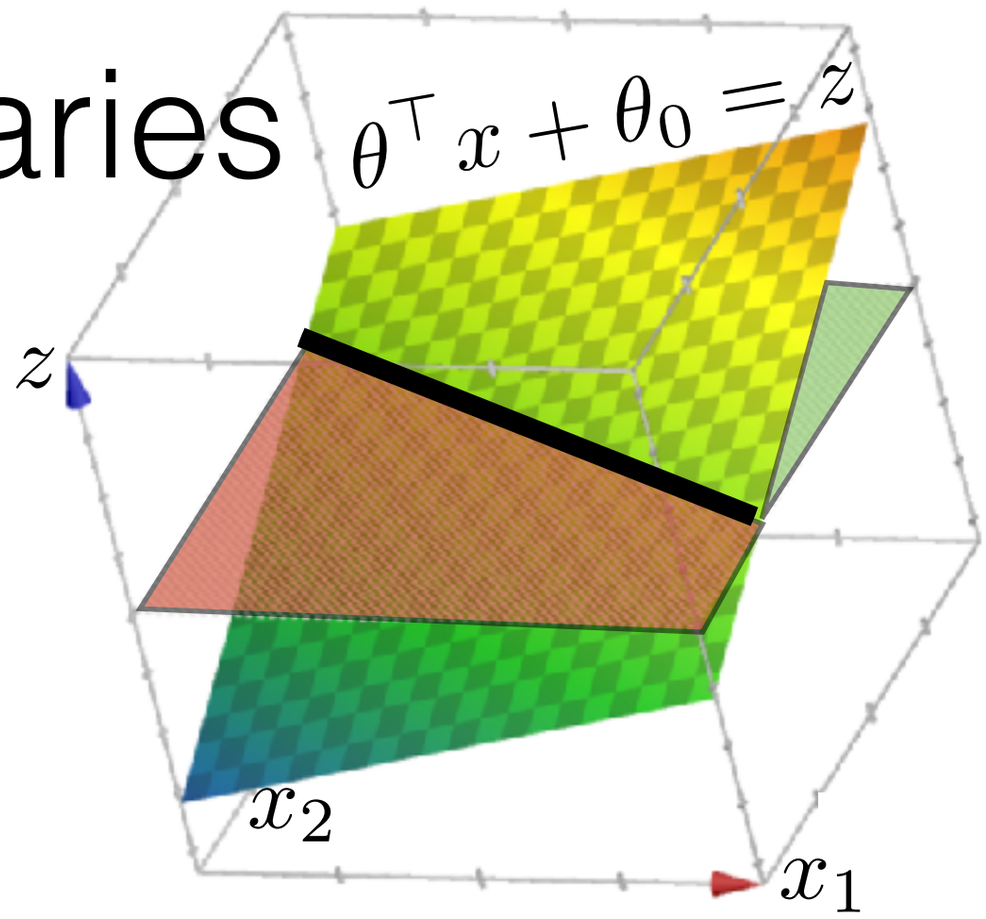
# Classification boundaries



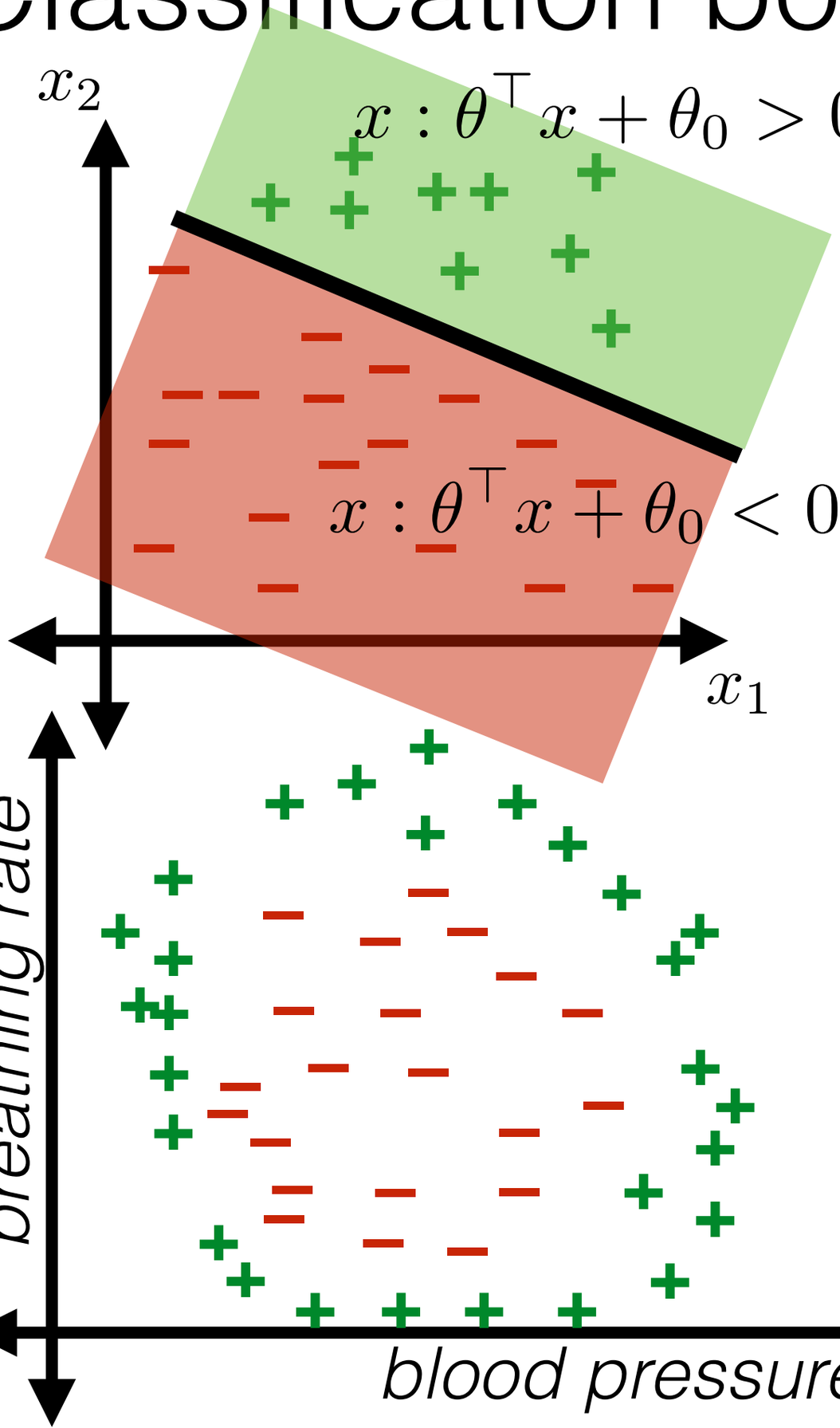



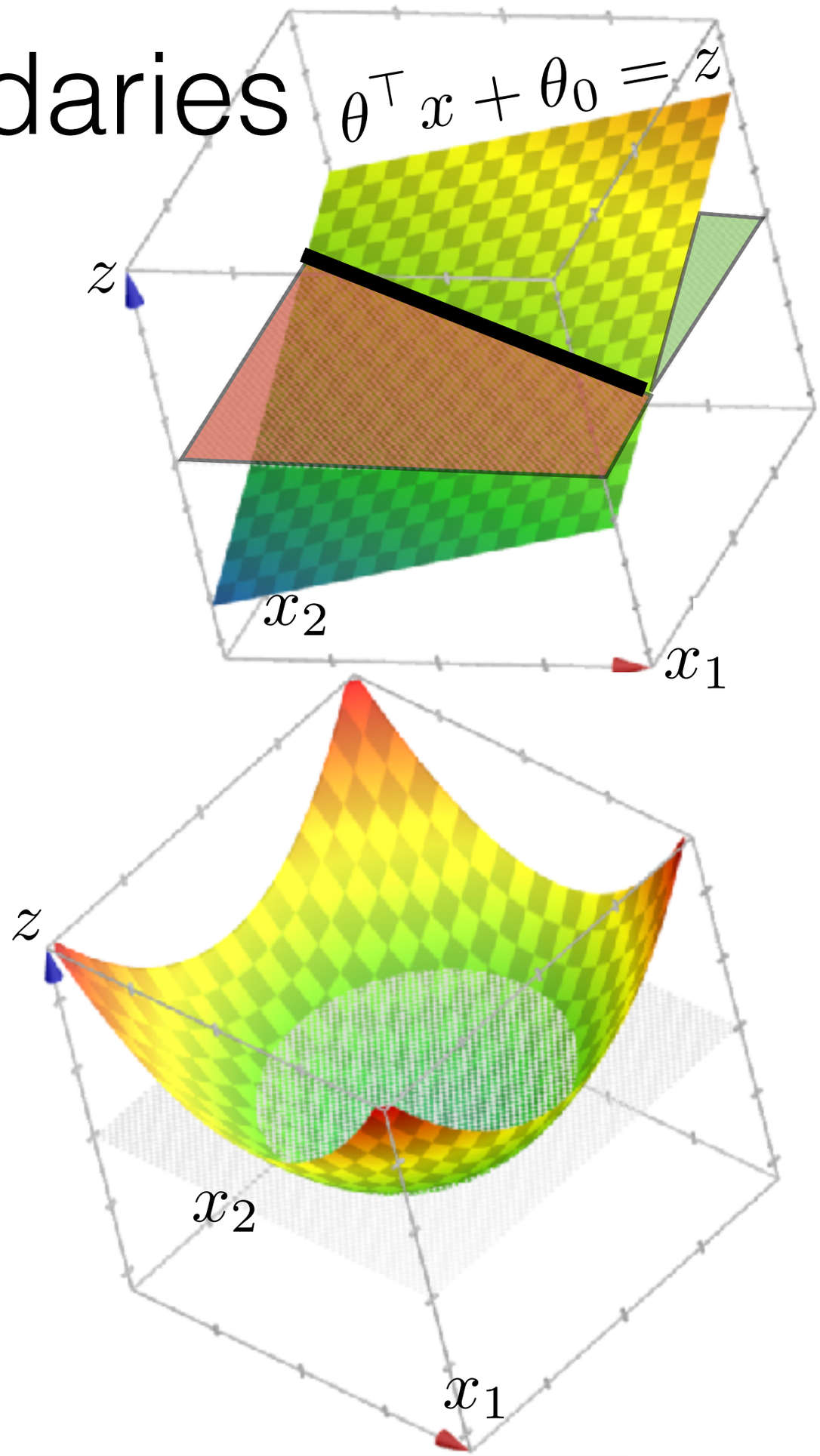
# Classification boundaries



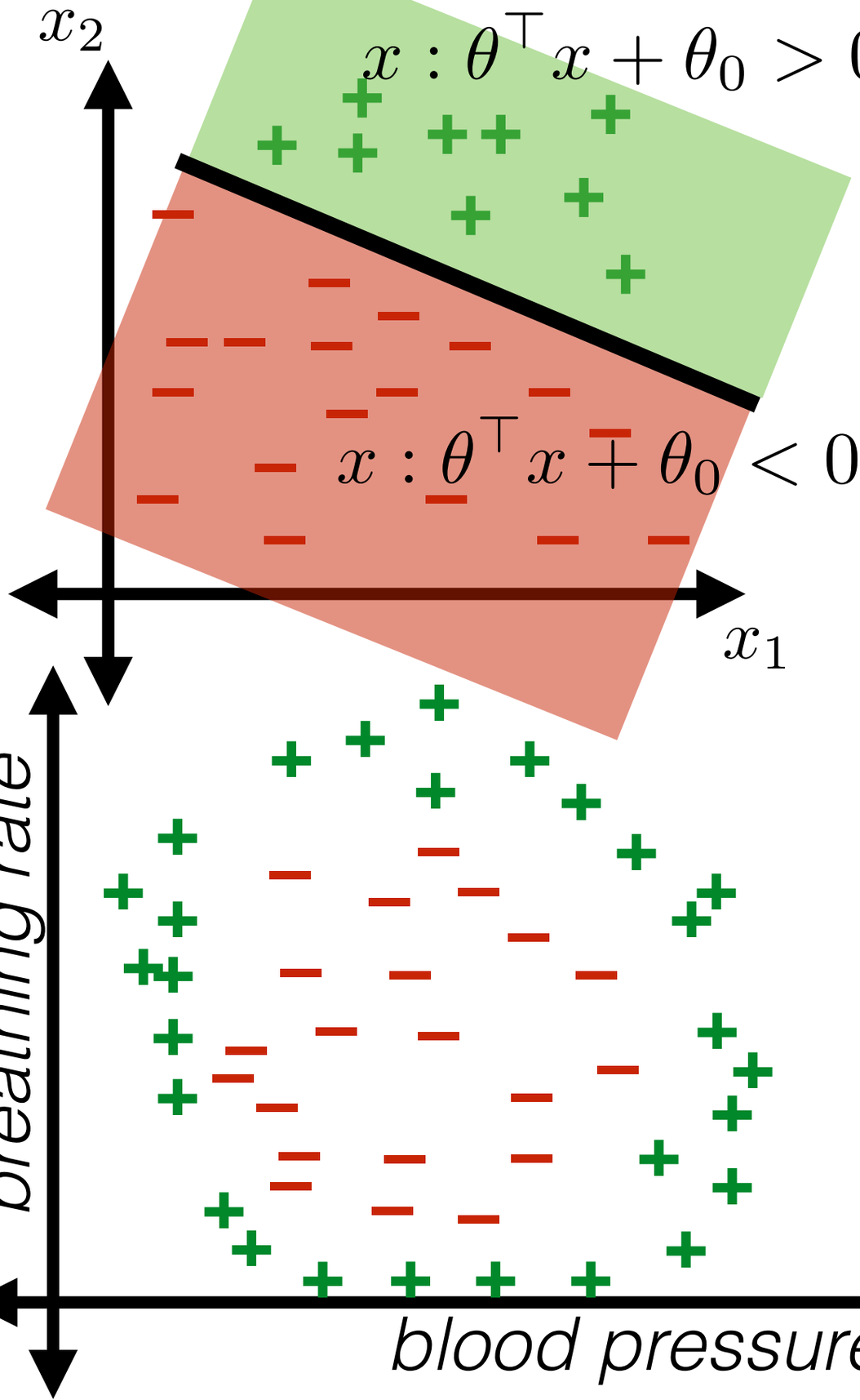



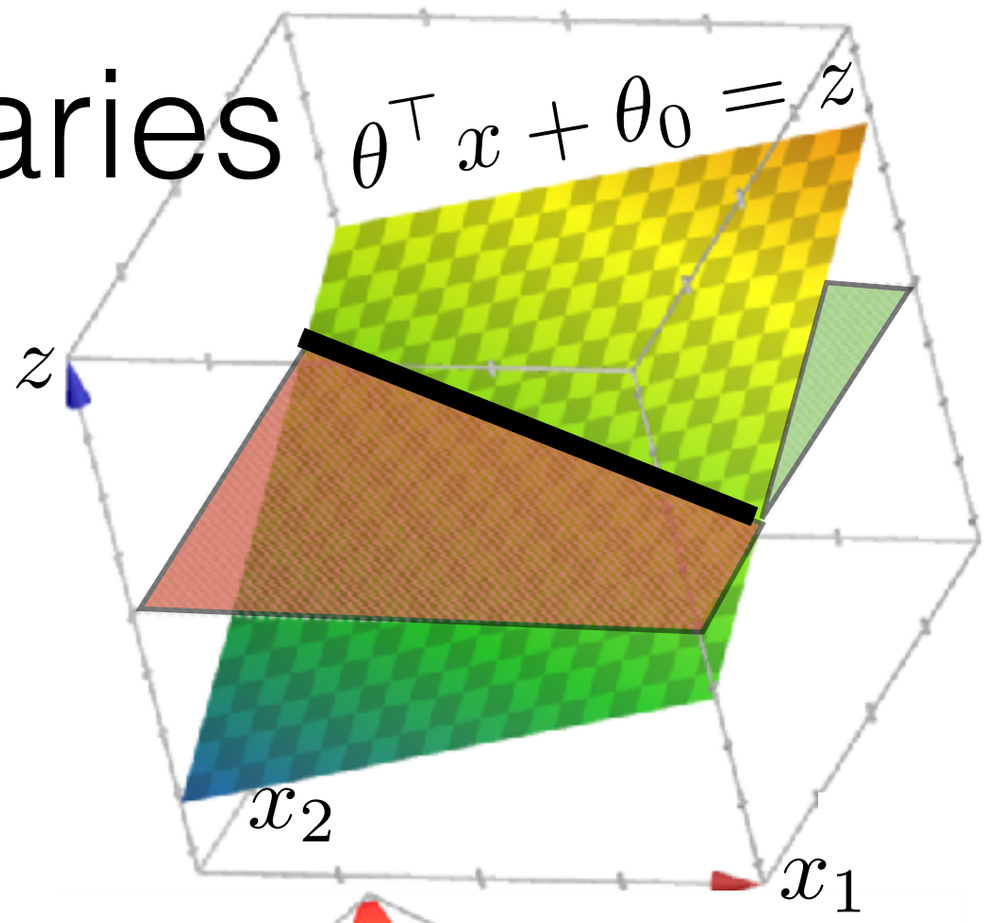
# Classification boundaries

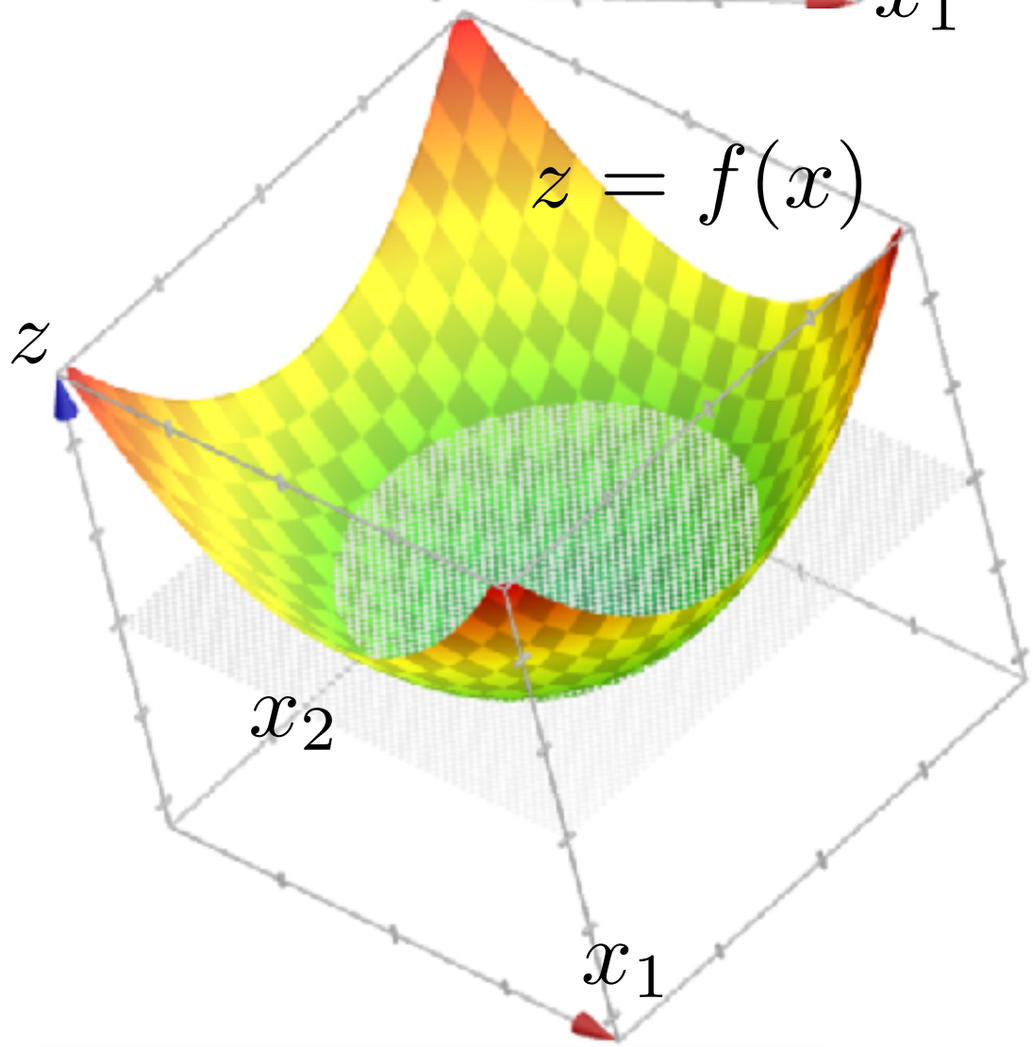


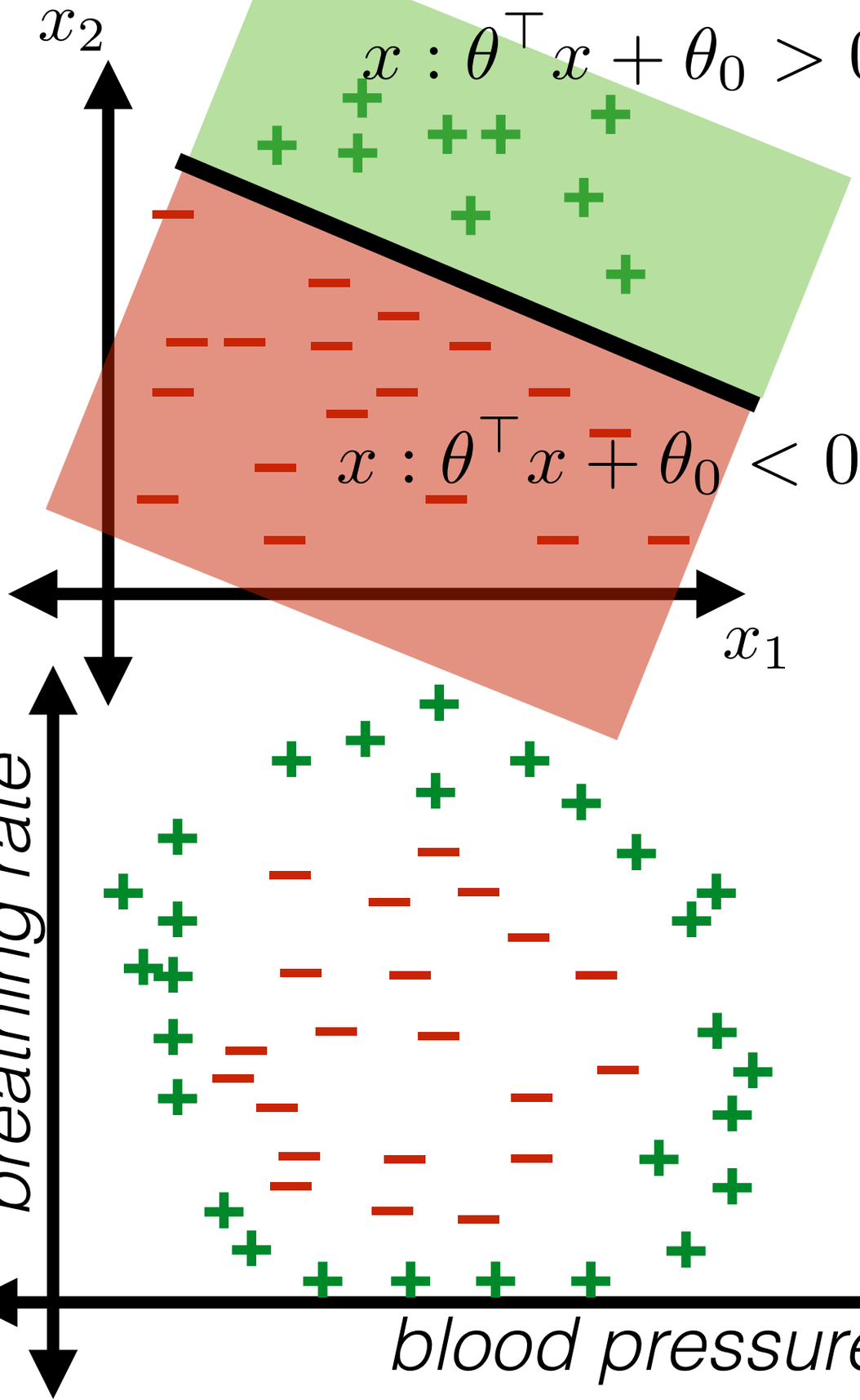
# Classification boundaries

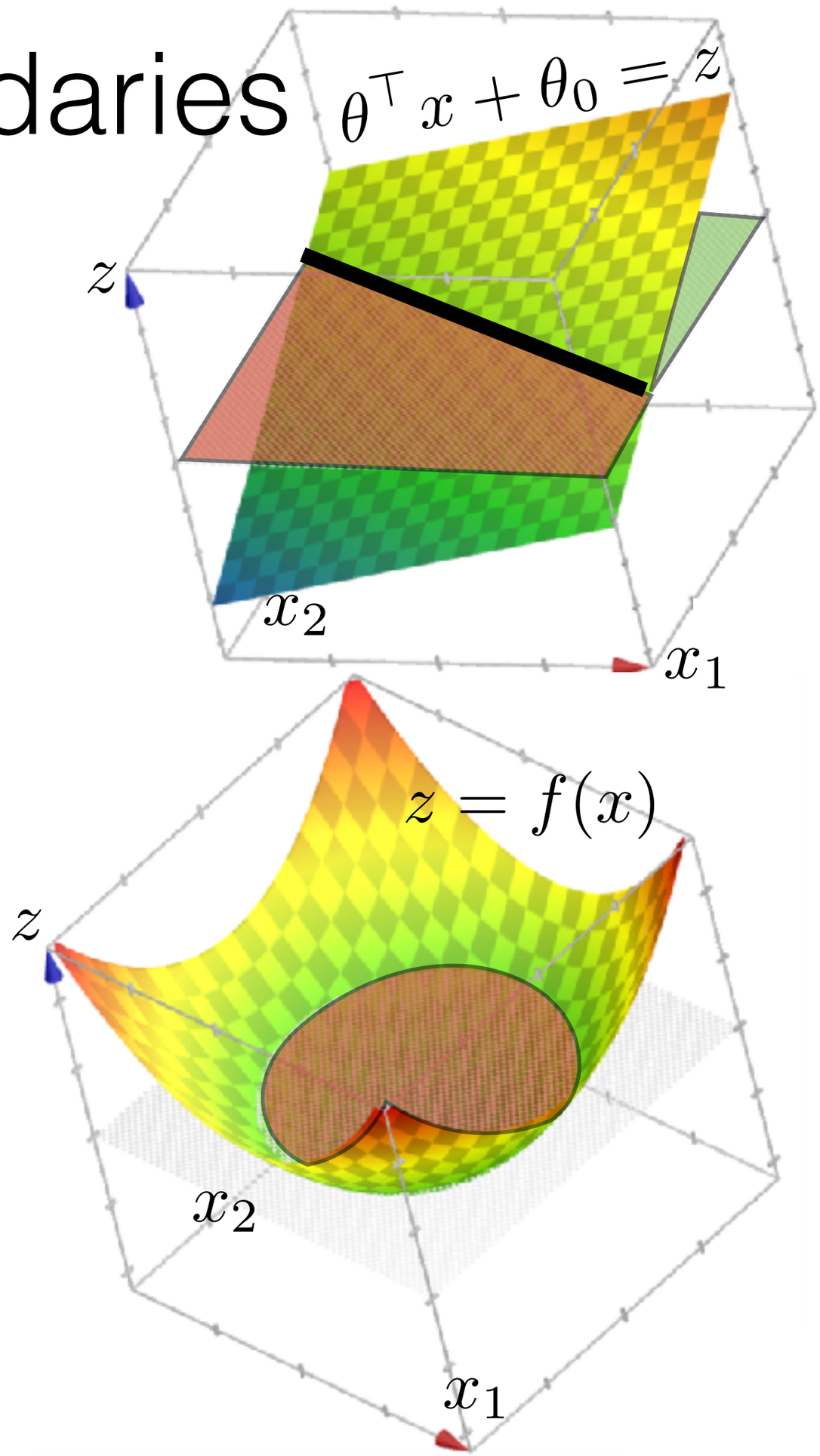


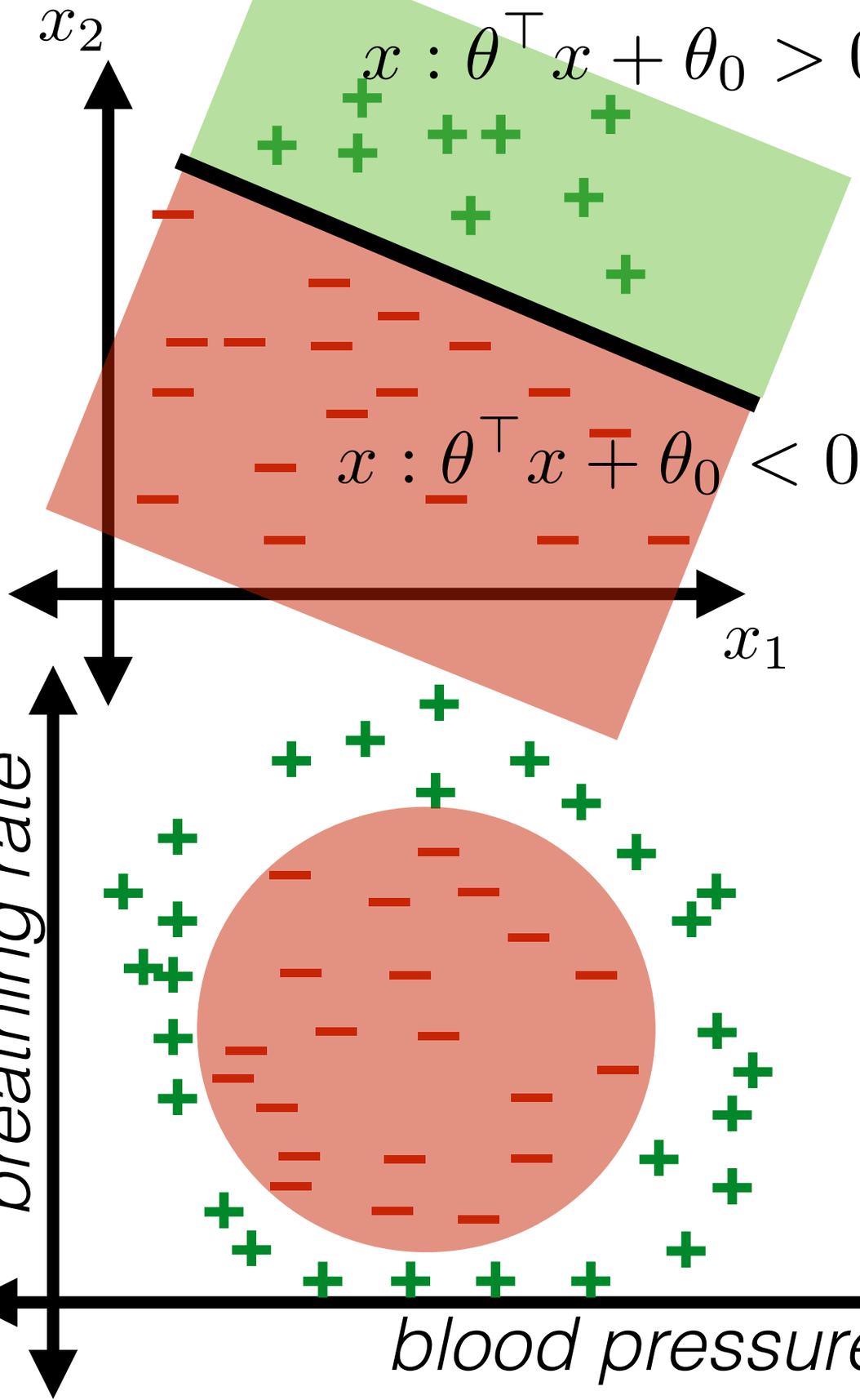



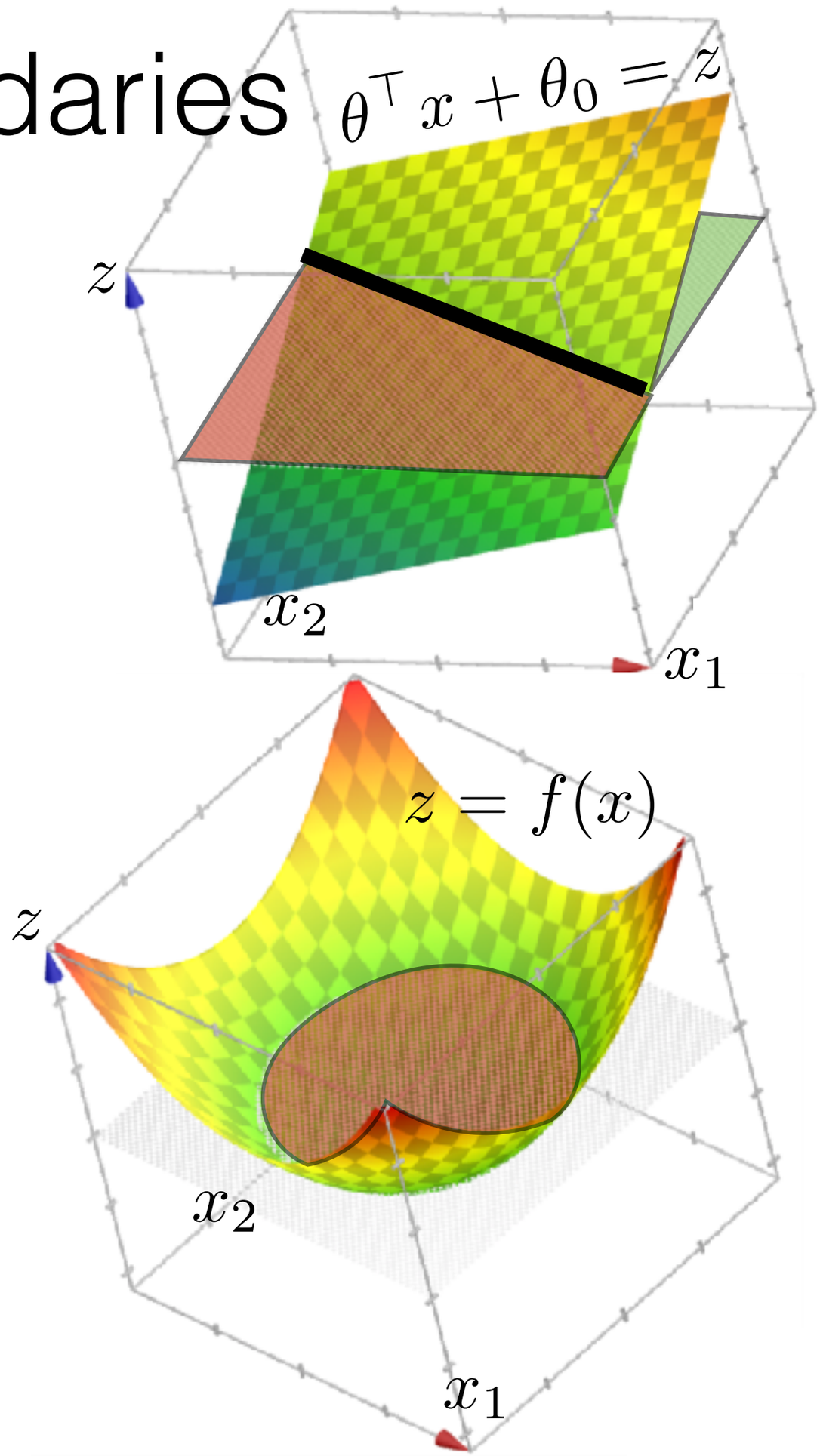
# Classification boundaries



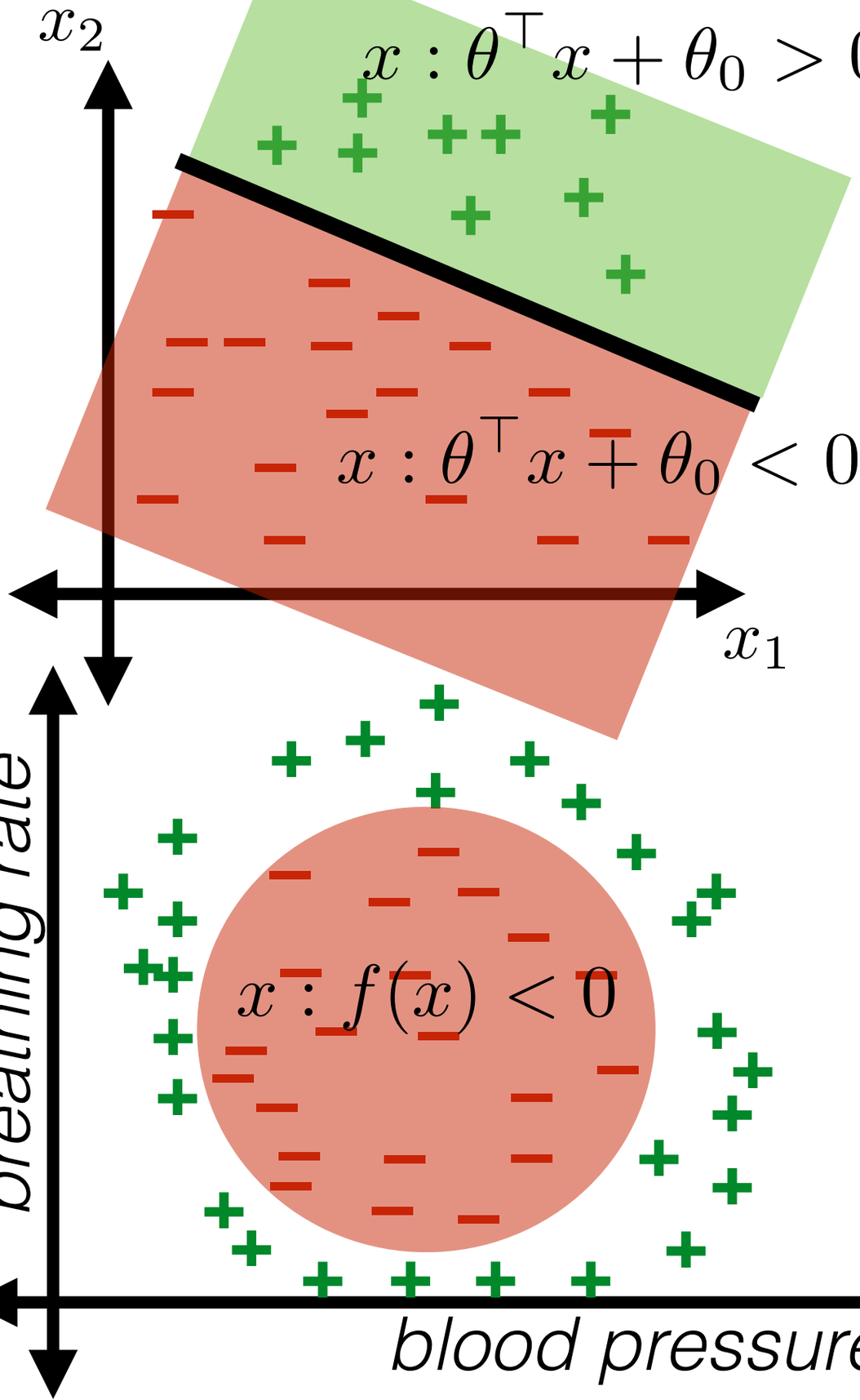



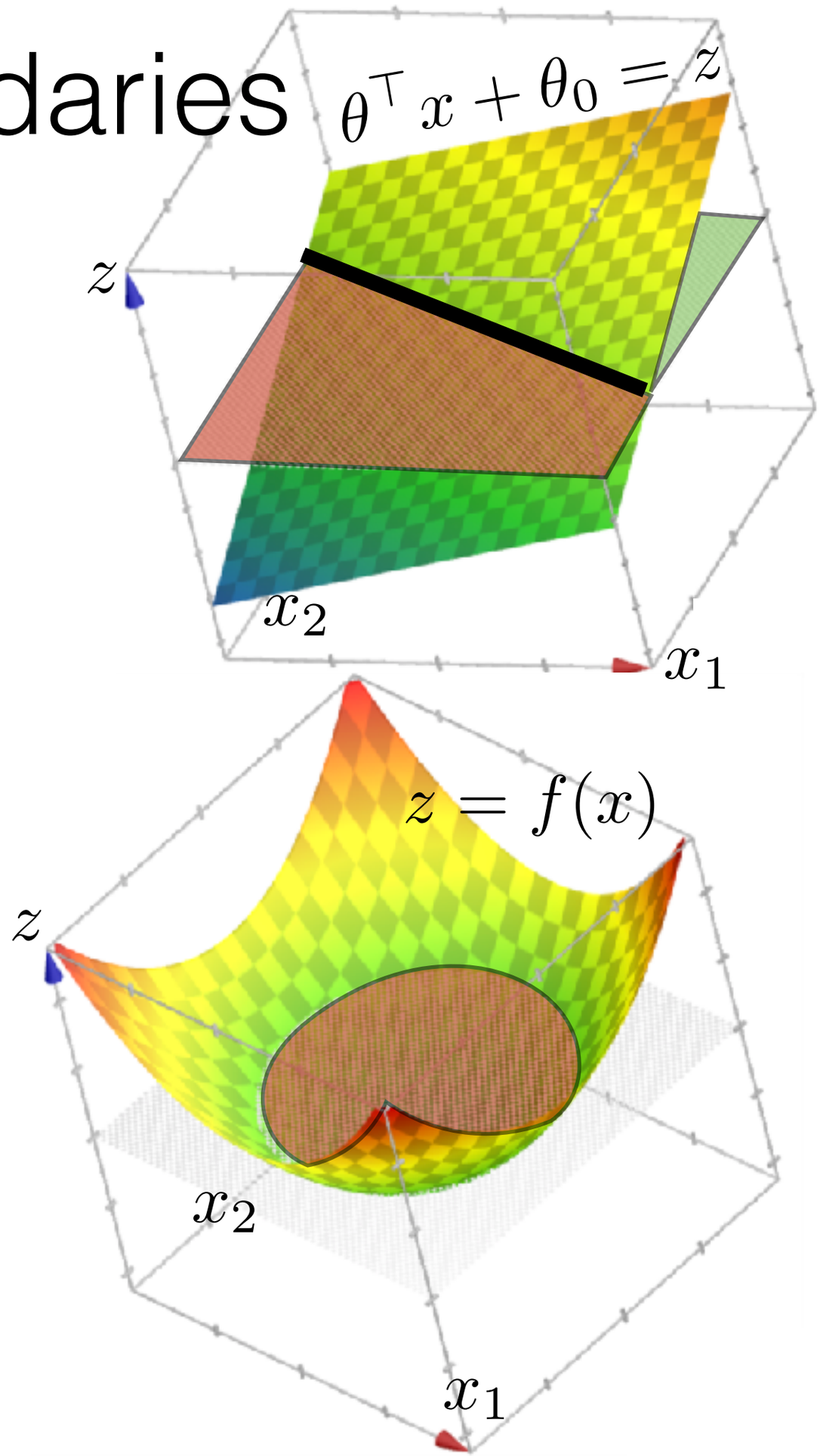
# Classification boundaries



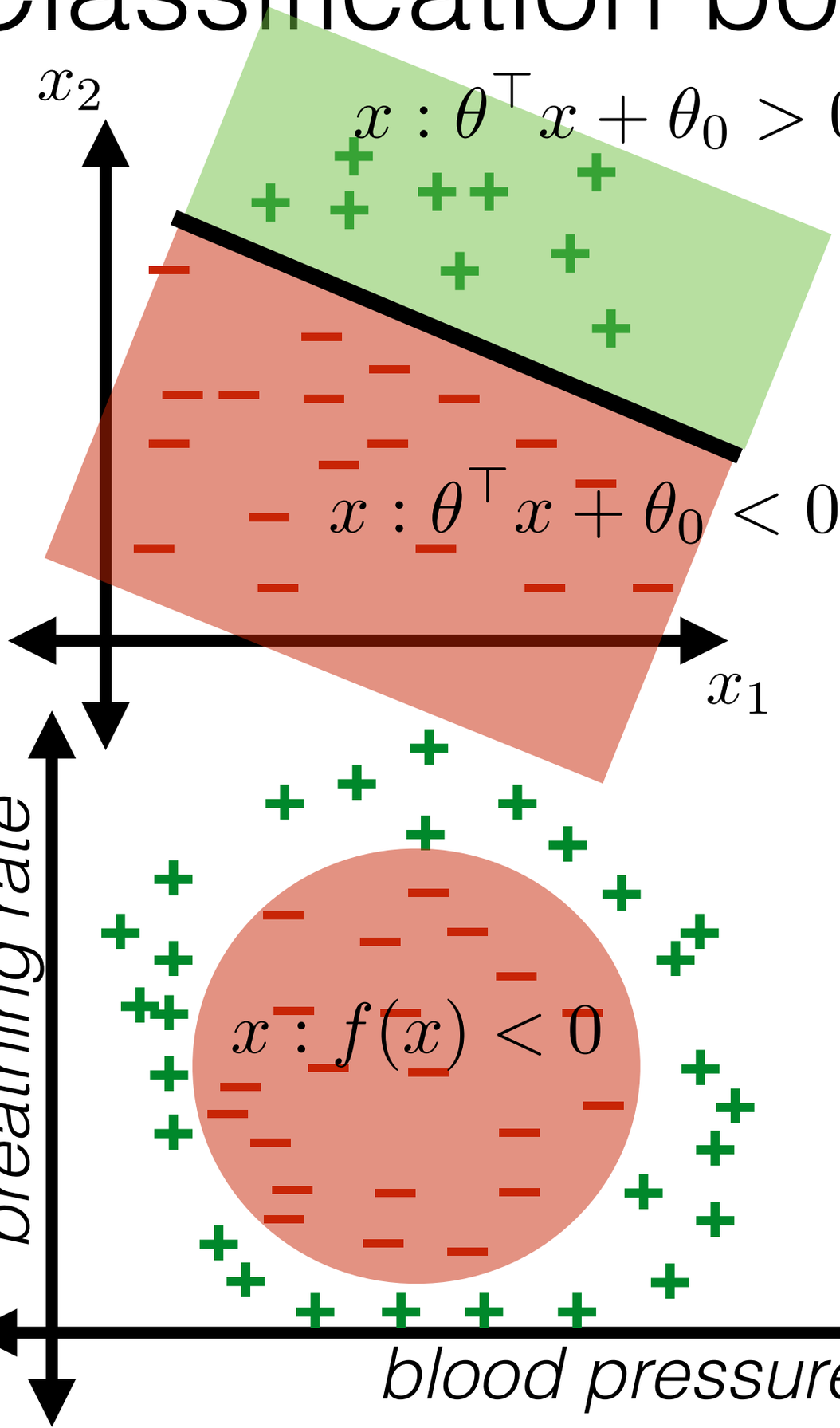



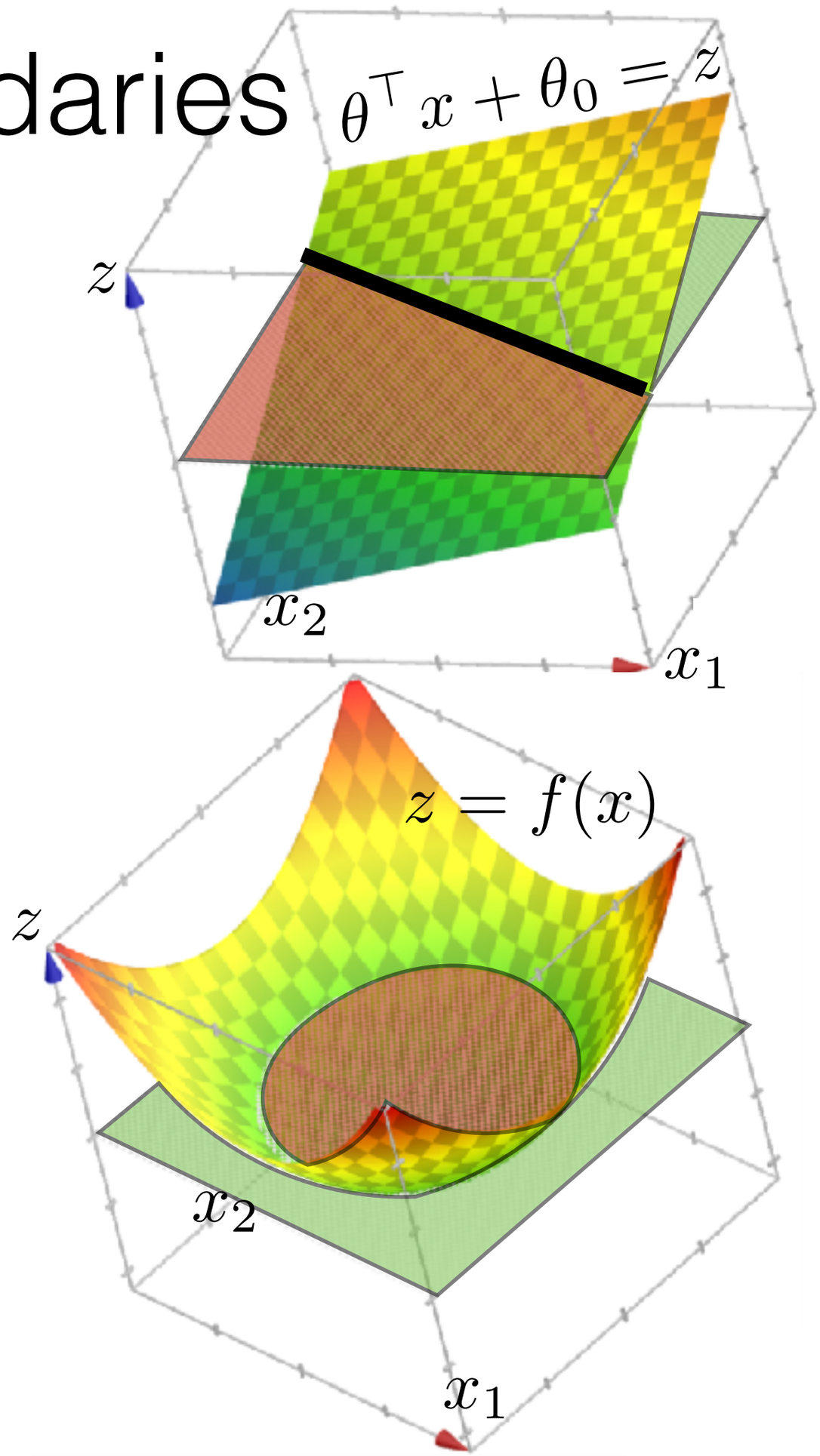
# Classification boundaries



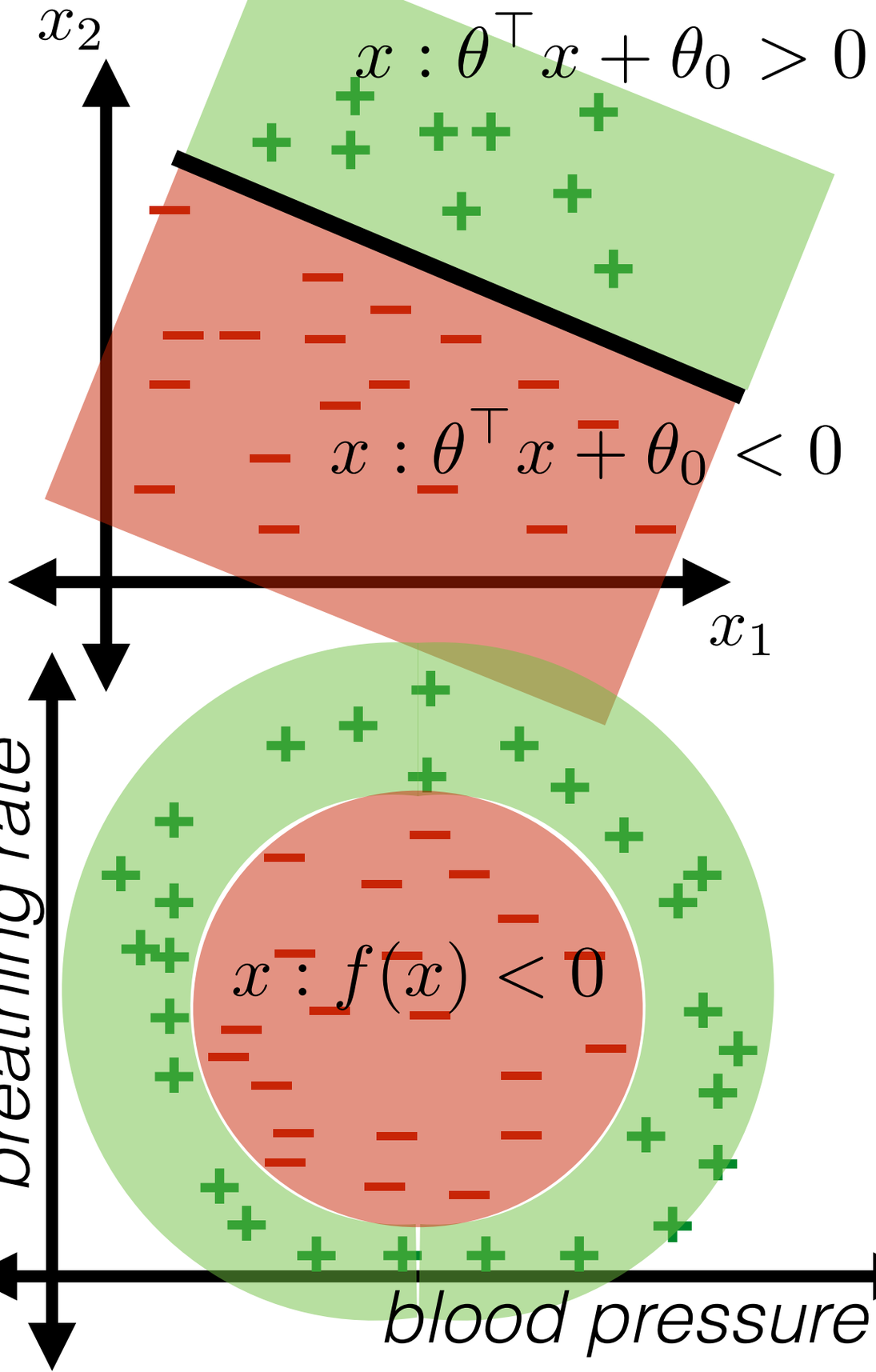



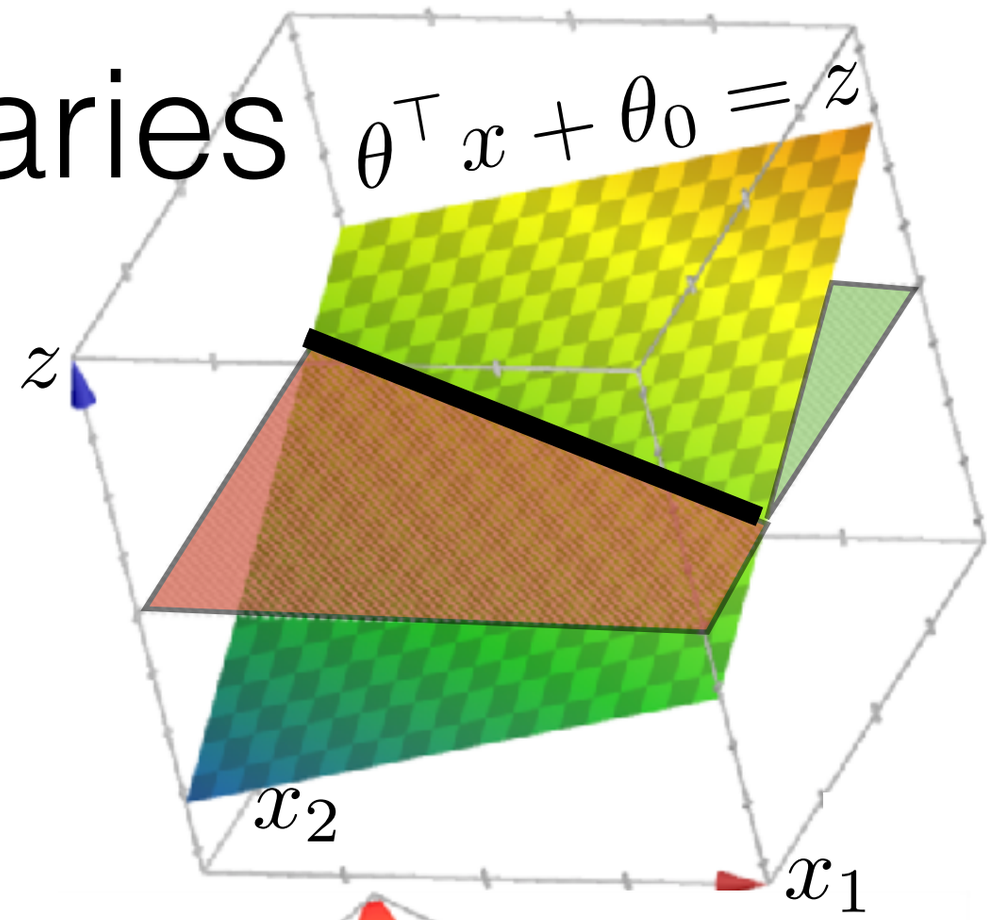
# Classification boundaries

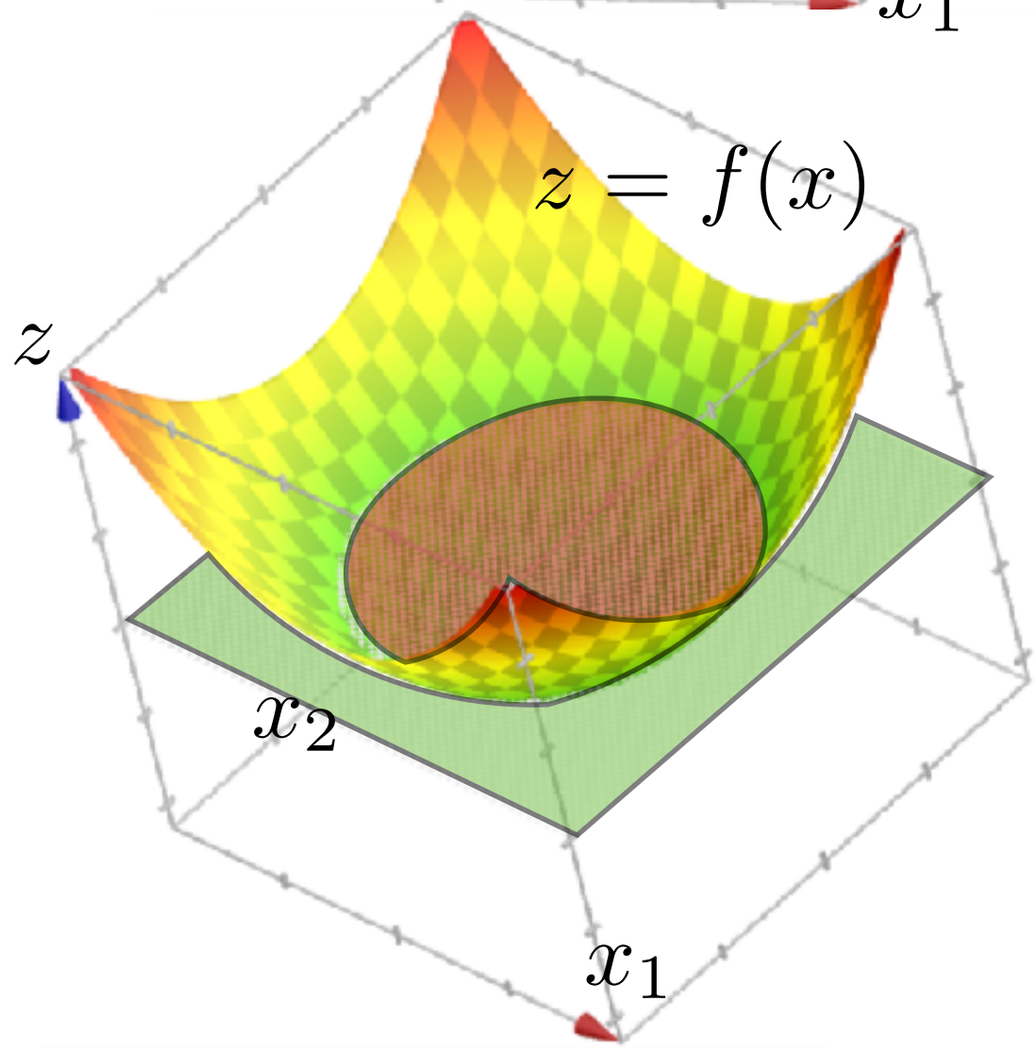


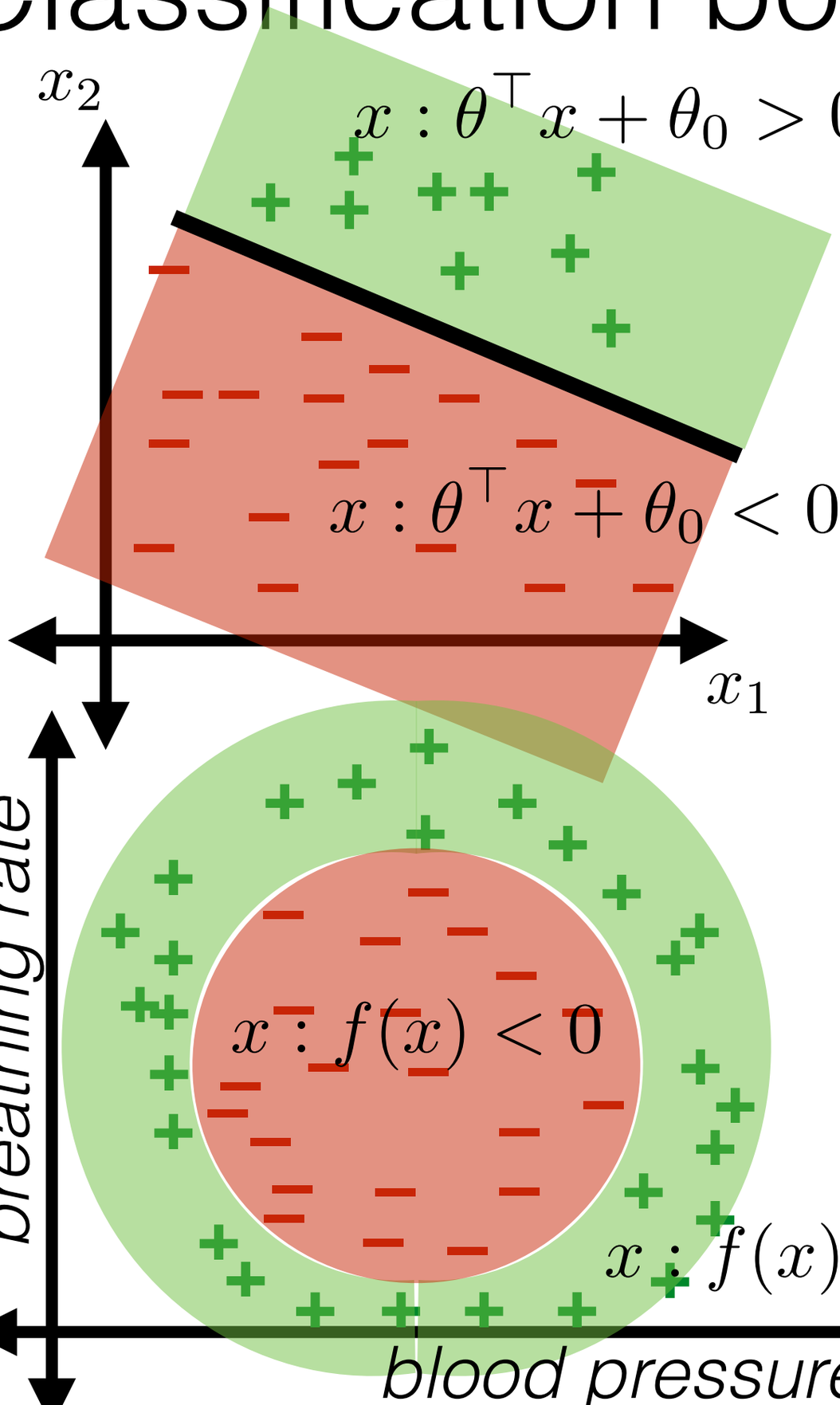
# Classification boundaries

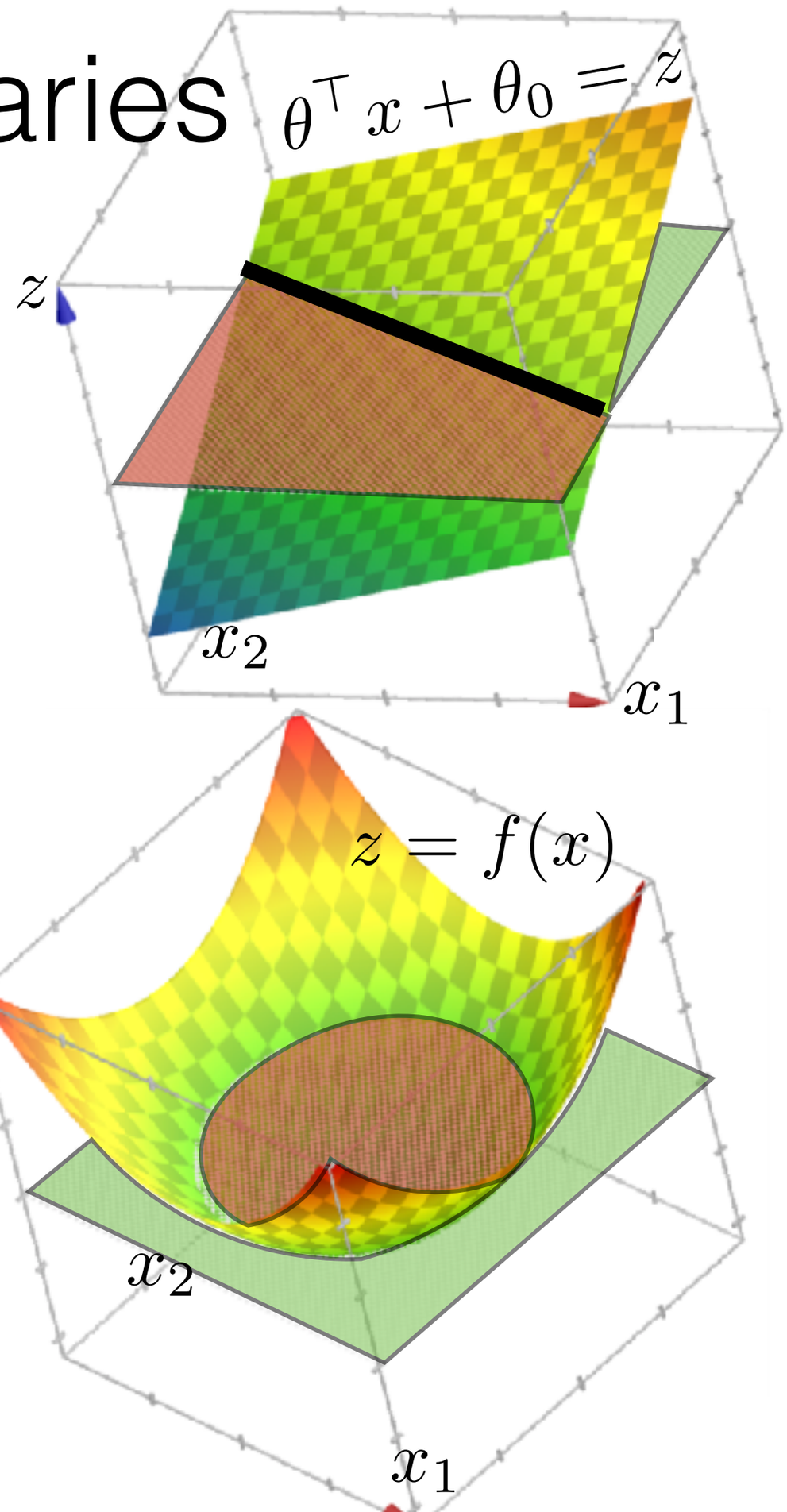


# Classification boundaries



# Nonlinear boundaries

# Nonlinear boundaries

- Idea: can approximate a smooth function with a *k*th order Taylor polynomial (e.g. around 0)

# Nonlinear boundaries

- Idea: can approximate a smooth function with a *k*th order Taylor polynomial (e.g. around 0)

| order (*k*) | terms when *d*=1 | terms for general d |
|---|---|---|
| 0 | | |
| 1 | | |
| 2 | | |
| 3 | | |

# Nonlinear boundaries

- Idea: can approximate a smooth function with a *k*th order Taylor polynomial (e.g. around 0)

| order (*k*) | terms when *d*=1 | terms for general d |
|---|---|---|
| 0 | $[1]$ | |
| 1 | | |
| 2 | | |
| 3 | | |

# Nonlinear boundaries

- Idea: can approximate a smooth function with a *k*th order Taylor polynomial (e.g. around 0)

| order (*k*) | terms when *d*=1 | terms for general d |
|---|---|---|
| 0 | $[1]$ | |
| 1 | $[1, x_1]$ | |
| 2 | | |
| 3 | | |

# Nonlinear boundaries

- Idea: can approximate a smooth function with a *k*th order Taylor polynomial (e.g. around 0)

| order (*k*) | terms when *d*=1 | terms for general d |
|---|---|---|
| 0 | $[1]$ | |
| 1 | $[1, x_1]$ | |
| 2 | $[1, x_1, x_1^2]$ | |
| 3 | | |

# Nonlinear boundaries

- Idea: can approximate a smooth function with a *k*th order Taylor polynomial (e.g. around 0)

| order (*k*) | terms when *d*=1 | terms for general d |
|---|---|---|
| 0 | $[1]$ | |
| 1 | $[1, x_1]$ | |
| 2 | $[1, x_1, x_1^2]$ | |
| 3 | $[1, x_1, x_1^2, x_1^3]$ | |

# Nonlinear boundaries

- Idea: can approximate a smooth function with a *k*th order Taylor polynomial (e.g. around 0)

| order (*k*) | terms when *d*=1 | terms for general d |
|---|---|---|
| 0 | $[1]$ | $[1]$ |
| 1 | $[1, x_1]$ | |
| 2 | $[1, x_1, x_1^2]$ | |
| 3 | $[1, x_1, x_1^2, x_1^3]$ | |

# Nonlinear boundaries

- Idea: can approximate a smooth function with a *k*th order Taylor polynomial (e.g. around 0)

| order (*k*) | terms when *d*=1 | terms for general d |
|---|---|---|
| 0 | $[1]$ | $[1]$ |
| 1 | $[1, x_1]$ | $[1, x_1, \ldots, x_d]$ |
| 2 | $[1, x_1, x_1^2]$ | |
| 3 | $[1, x_1, x_1^2, x_1^3]$ | |

# Nonlinear boundaries

- Idea: can approximate a smooth function with a *k*th order Taylor polynomial (e.g. around 0)

| order (*k*) | terms when *d*=1 | terms for general d |
|---|---|---|
| 0 | $[1]$ | $[1]$ |
| 1 | $[1, x_1]$ | $[1, x_1, \ldots, x_d]$ |
| 2 | $[1, x_1, x_1^2]$ | $[1, x_1, \ldots, x_d,$ $x_1^2, x_1x_2, \ldots, x_{d-1}x_d, x_d^2]$ |
| 3 | $[1, x_1, x_1^2, x_1^3]$ | |

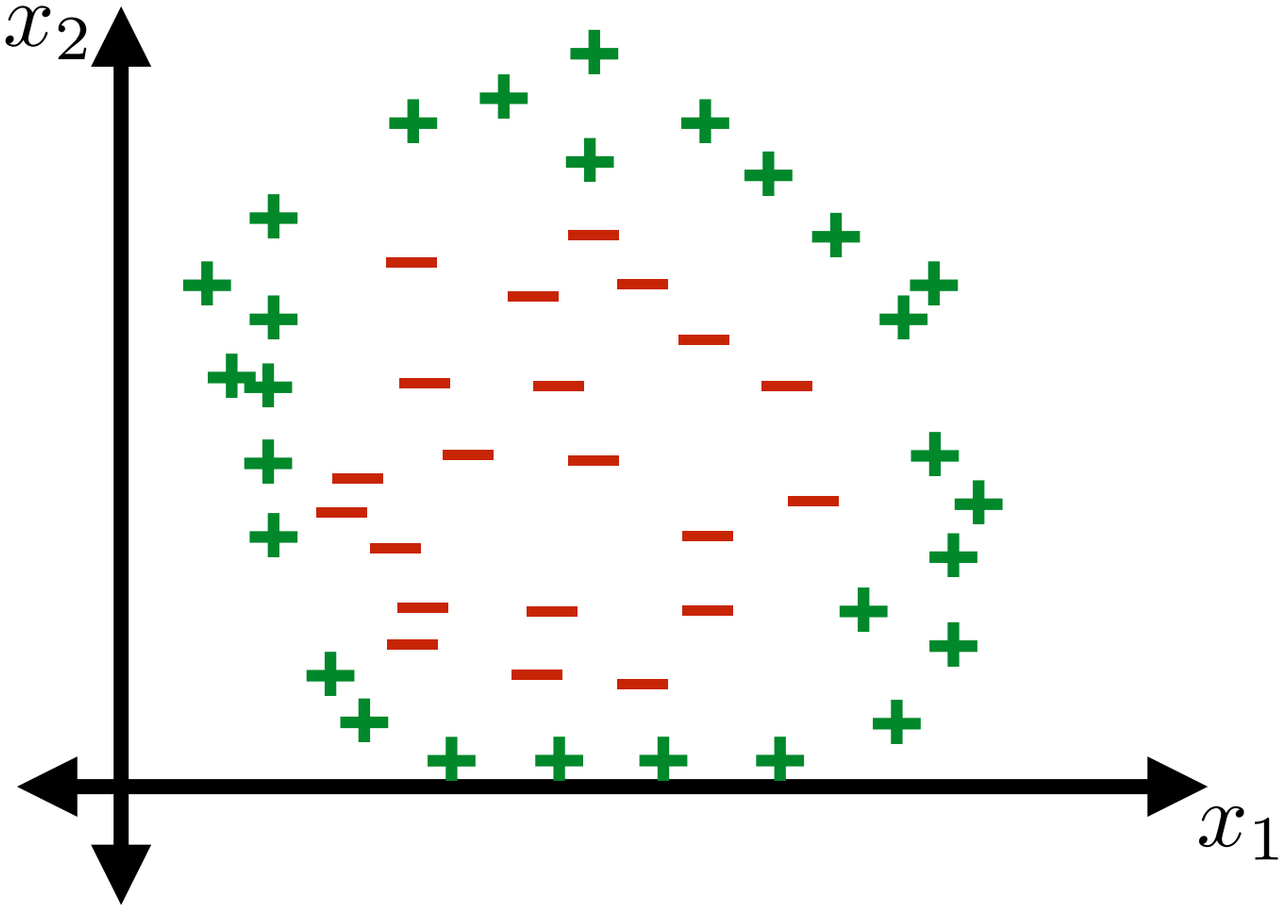
# Nonlinear boundaries

- Idea: can approximate a smooth function with a *k*th order Taylor polynomial (e.g. around 0)

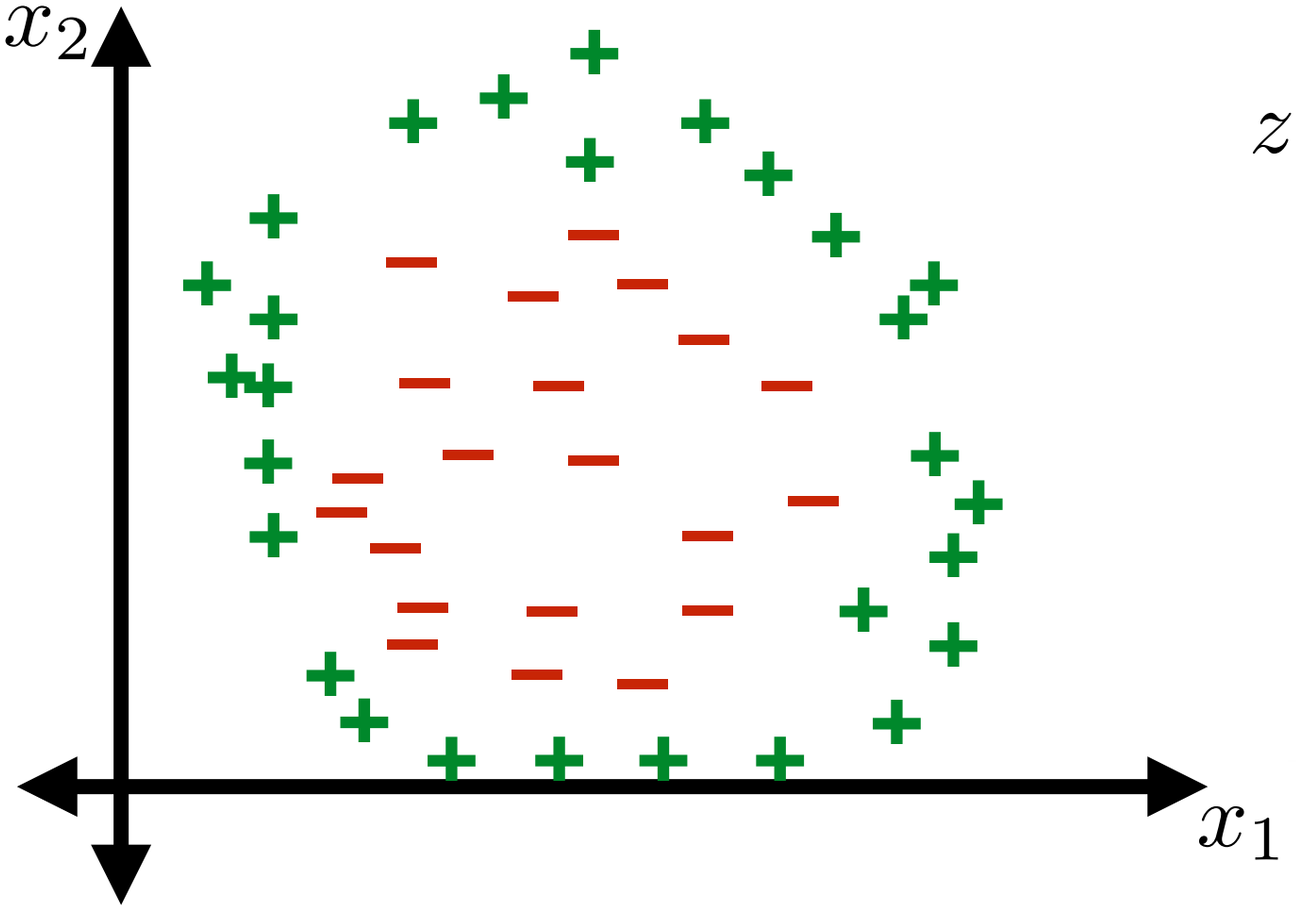
| order (*k*) | terms when *d*=1 | terms for general d |
|---|---|---|
| 0 | $[1]$ | $[1]$ |
| 1 | $[1, x_1]$ | $[1, x_1, \ldots, x_d]$ |
| 2 | $[1, x_1, x_1^2]$ | $[1, x_1, \ldots, x_d,$ $x_1^2, x_1x_2, \ldots, x_{d-1}x_d, x_d^2]$ |
| 3 | $[1, x_1, x_1^2, x_1^3]$ | $[1, x_1, \ldots, x_d,$ $x_1^2, x_1x_2, \ldots, x_{d-1}x_d, x_d^2,$ $x_1^3, x_1^2x_2, x_1x_2x_3, \ldots, x_d^3]$ |

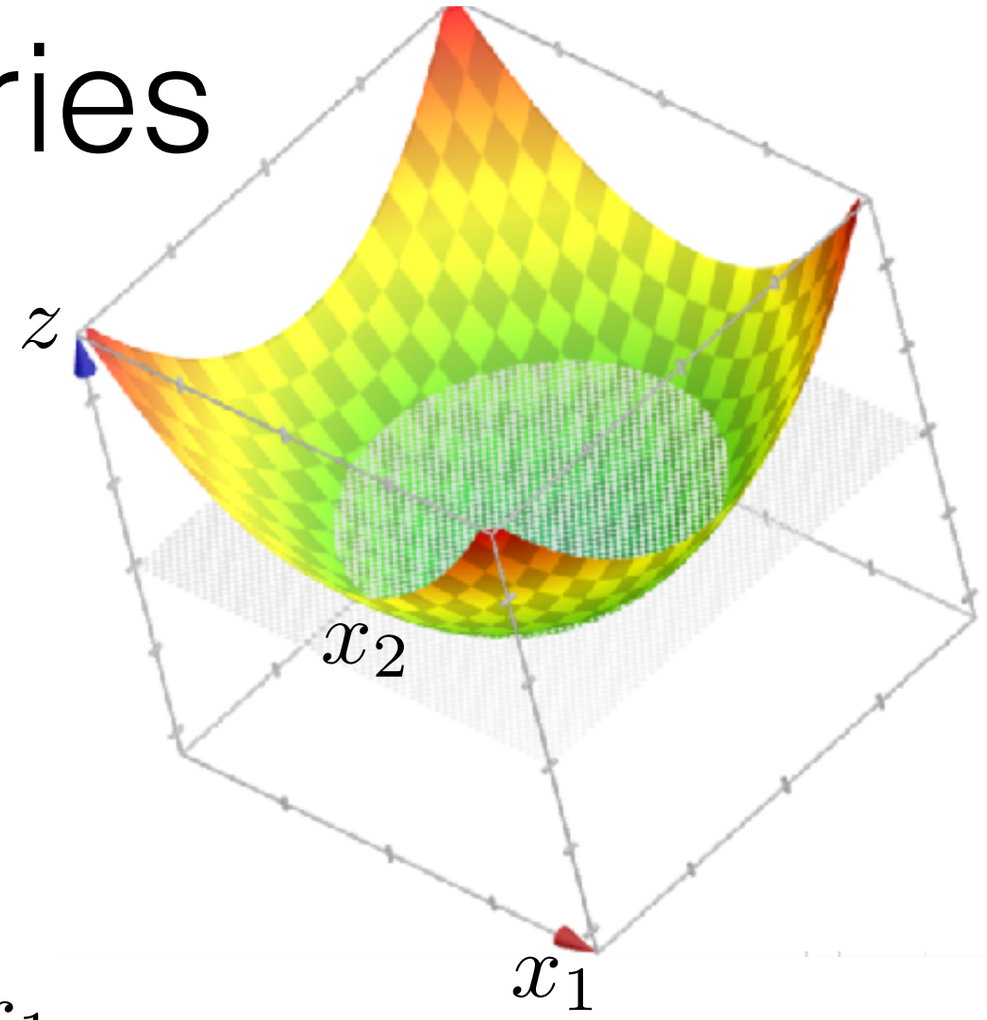
# Nonlinear boundaries

# Nonlinear boundaries

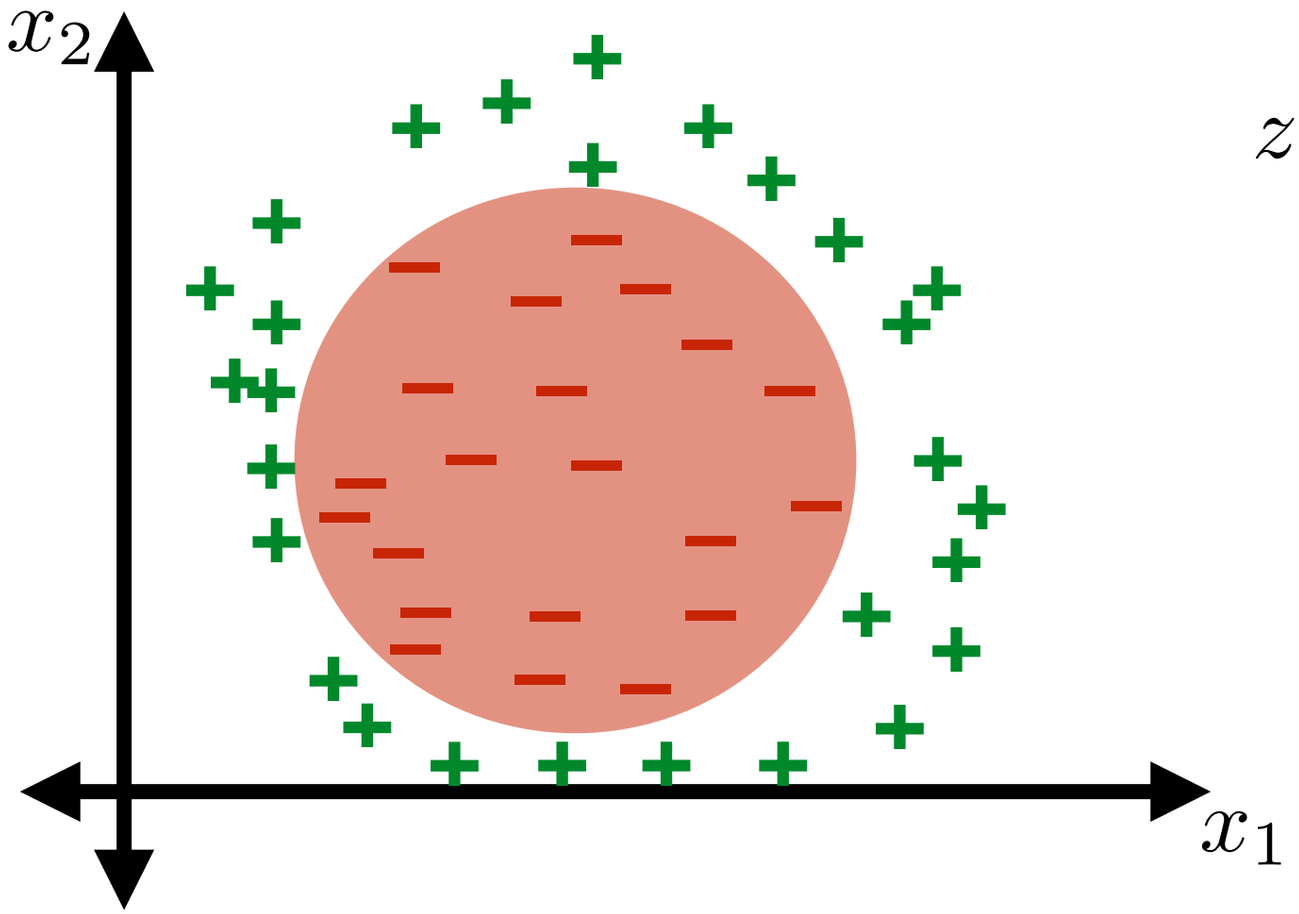


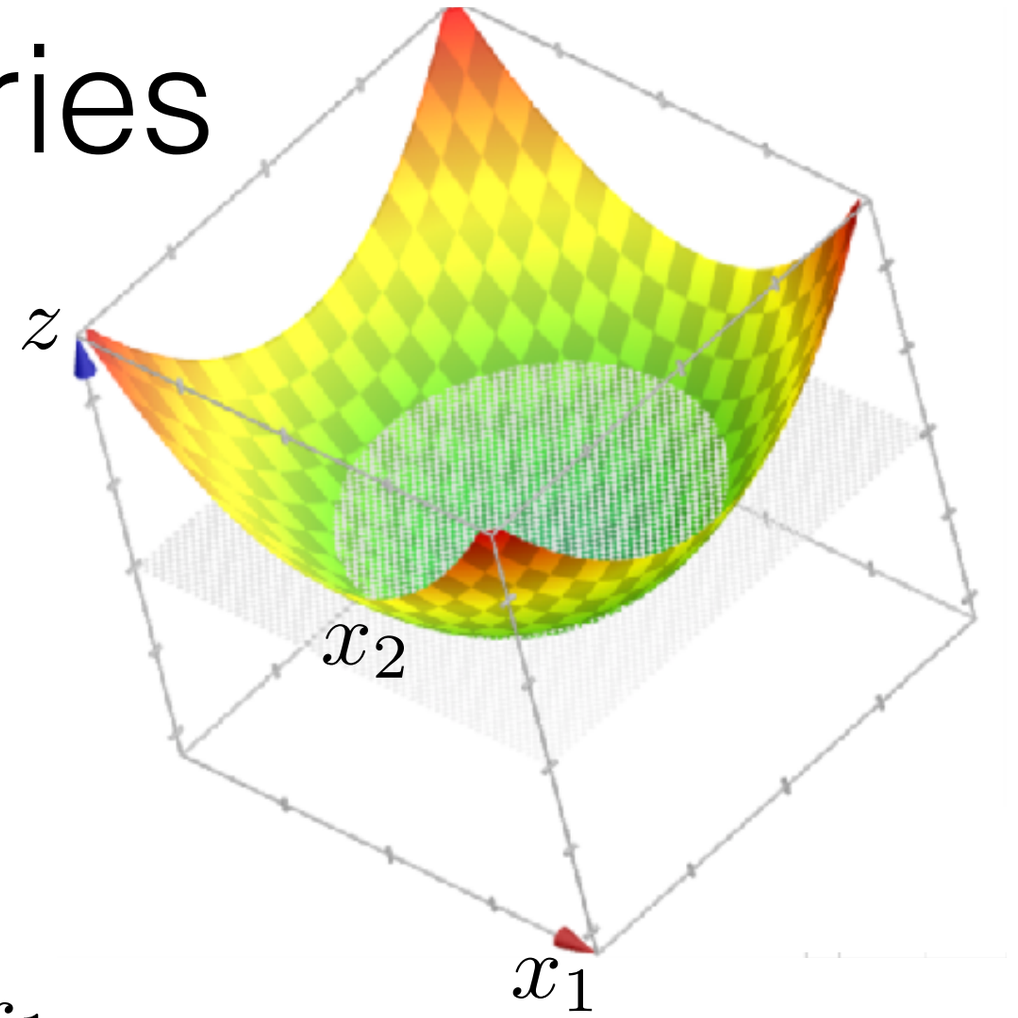
# Nonlinear boundaries



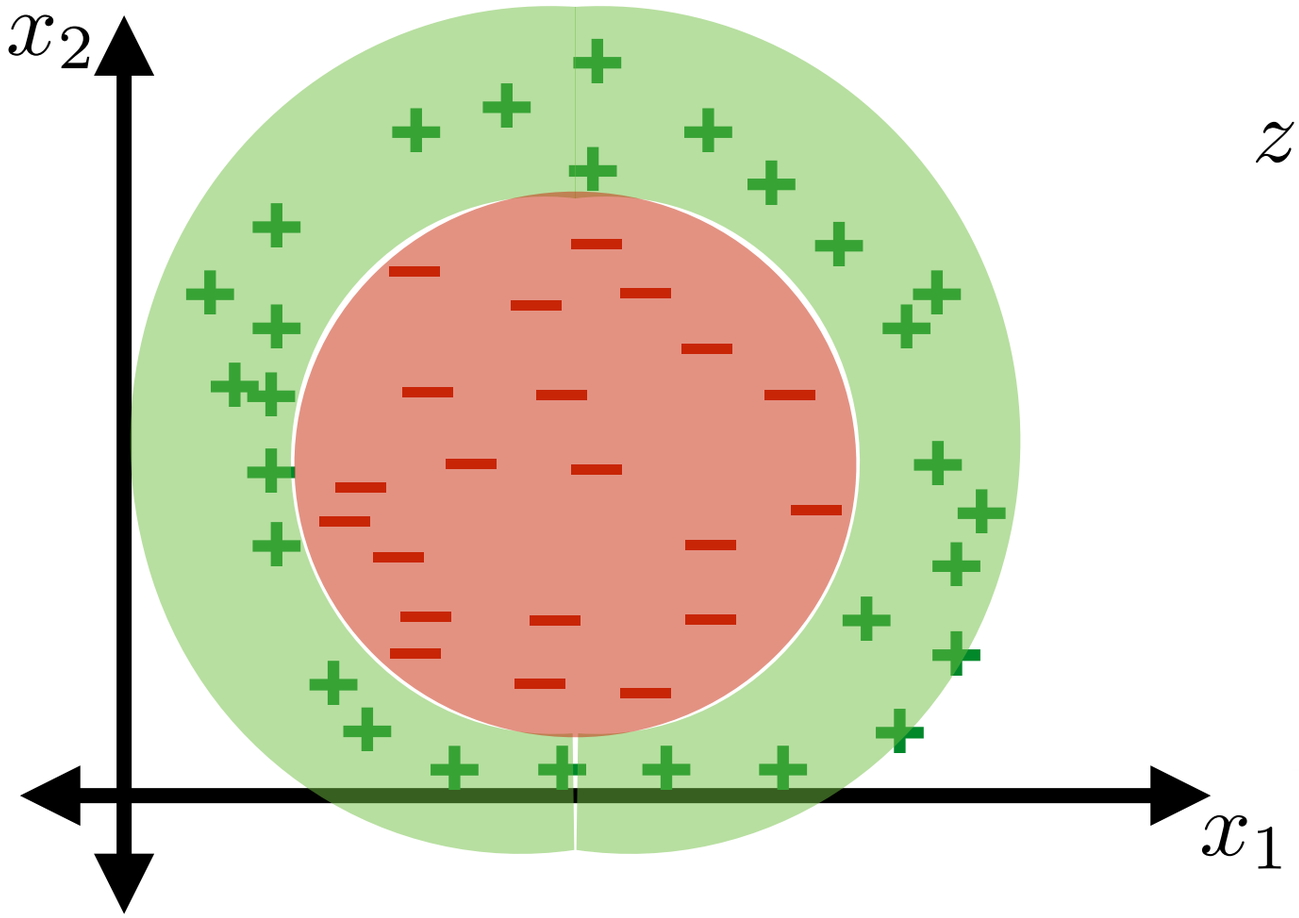



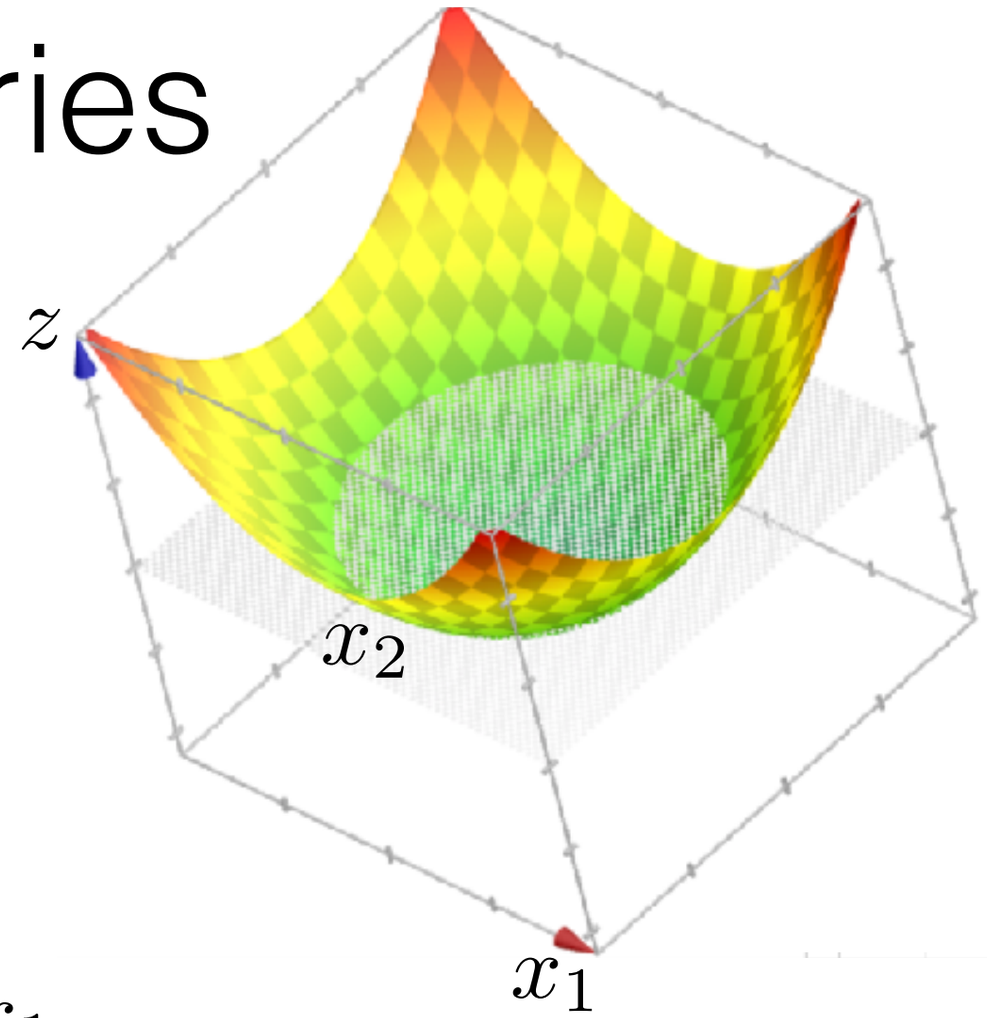
# Nonlinear boundaries



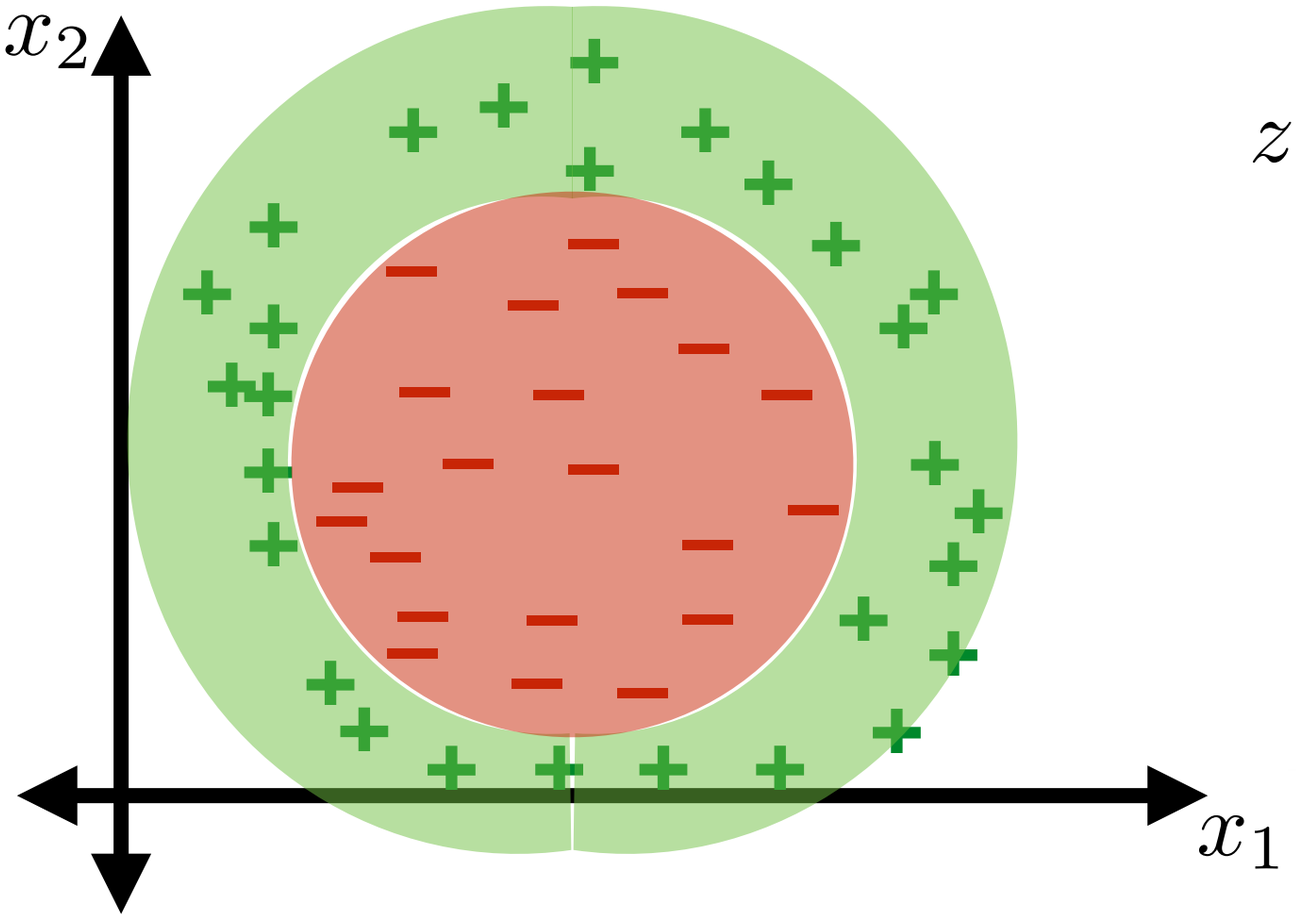



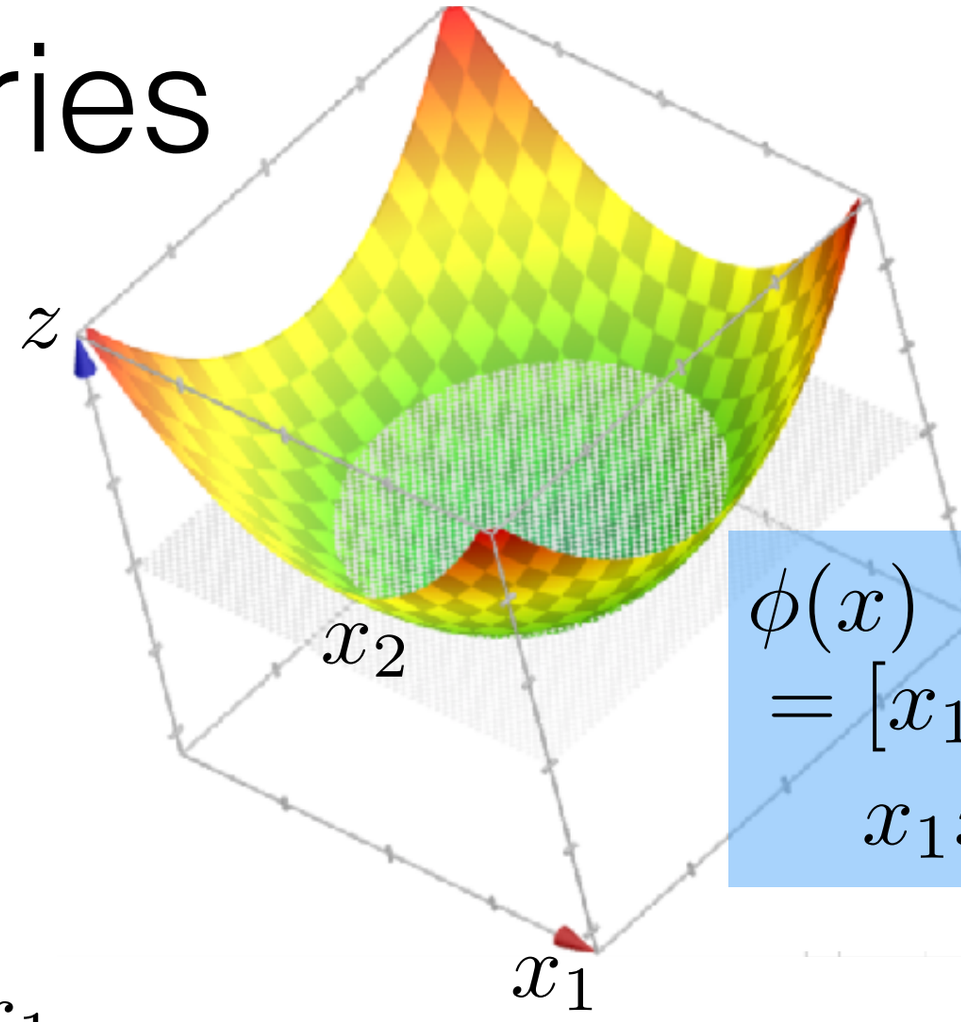
# Nonlinear boundaries



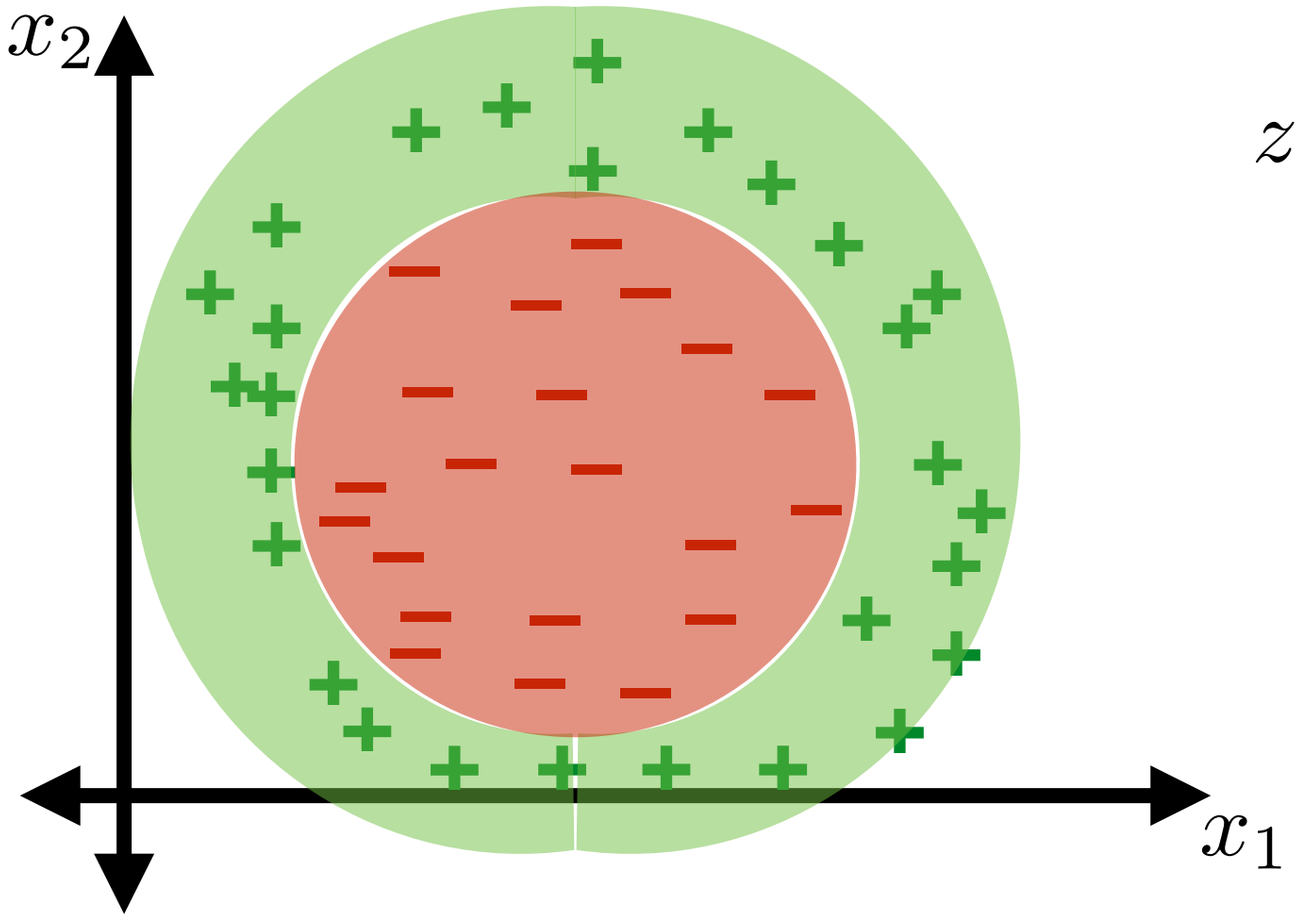
# Nonlinear boundaries



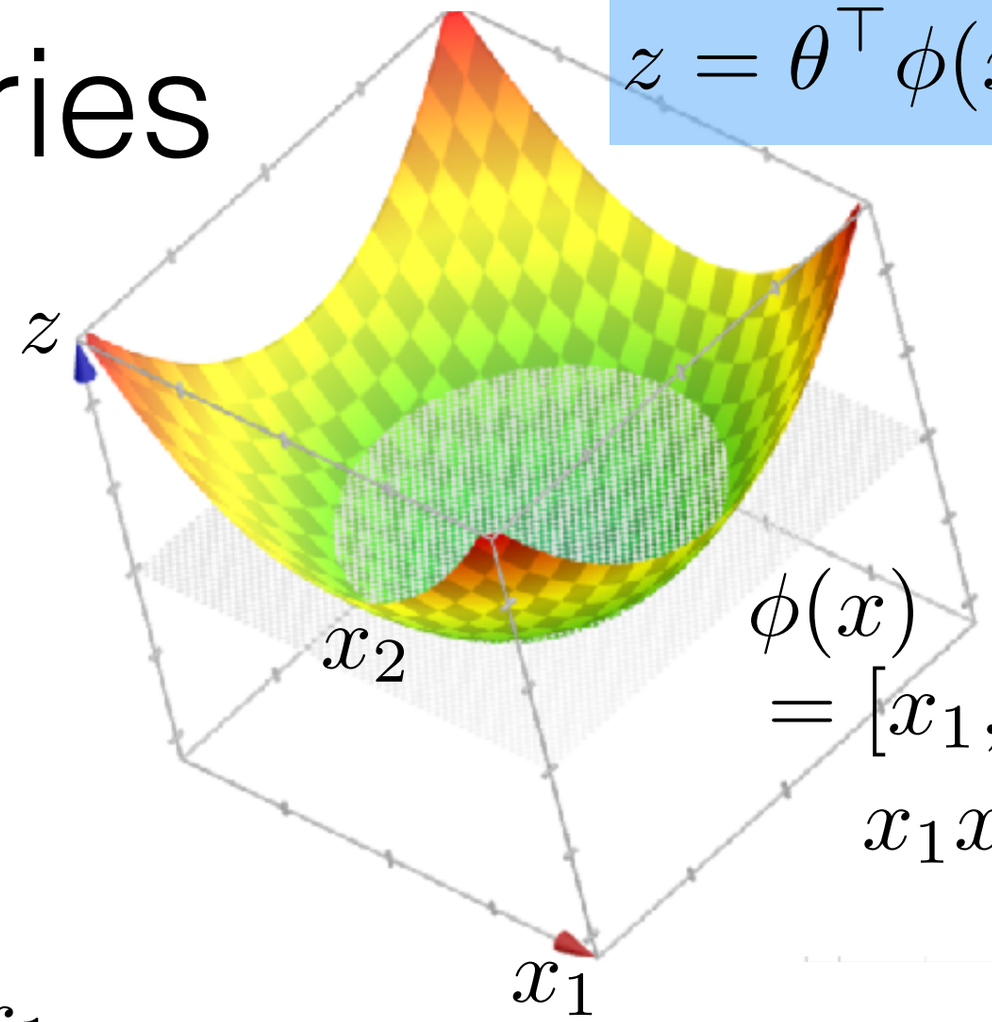




$$\phi(x) = [x_1, x_2, x_1^2, x_1 x_2, x_2^2]^\top$$

# Nonlinear boundaries

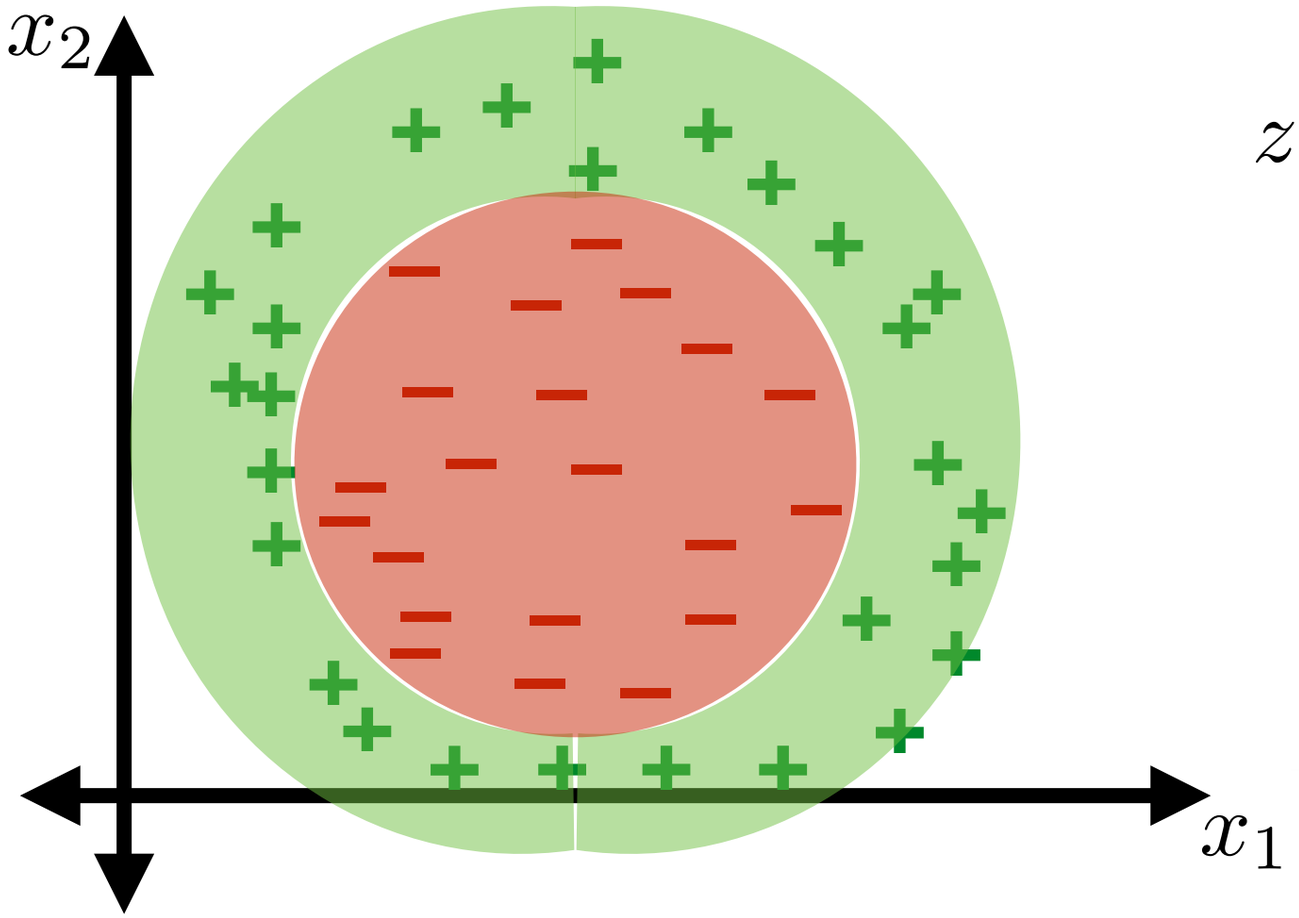



$$z = \theta^\top \phi(x) + \theta_0$$

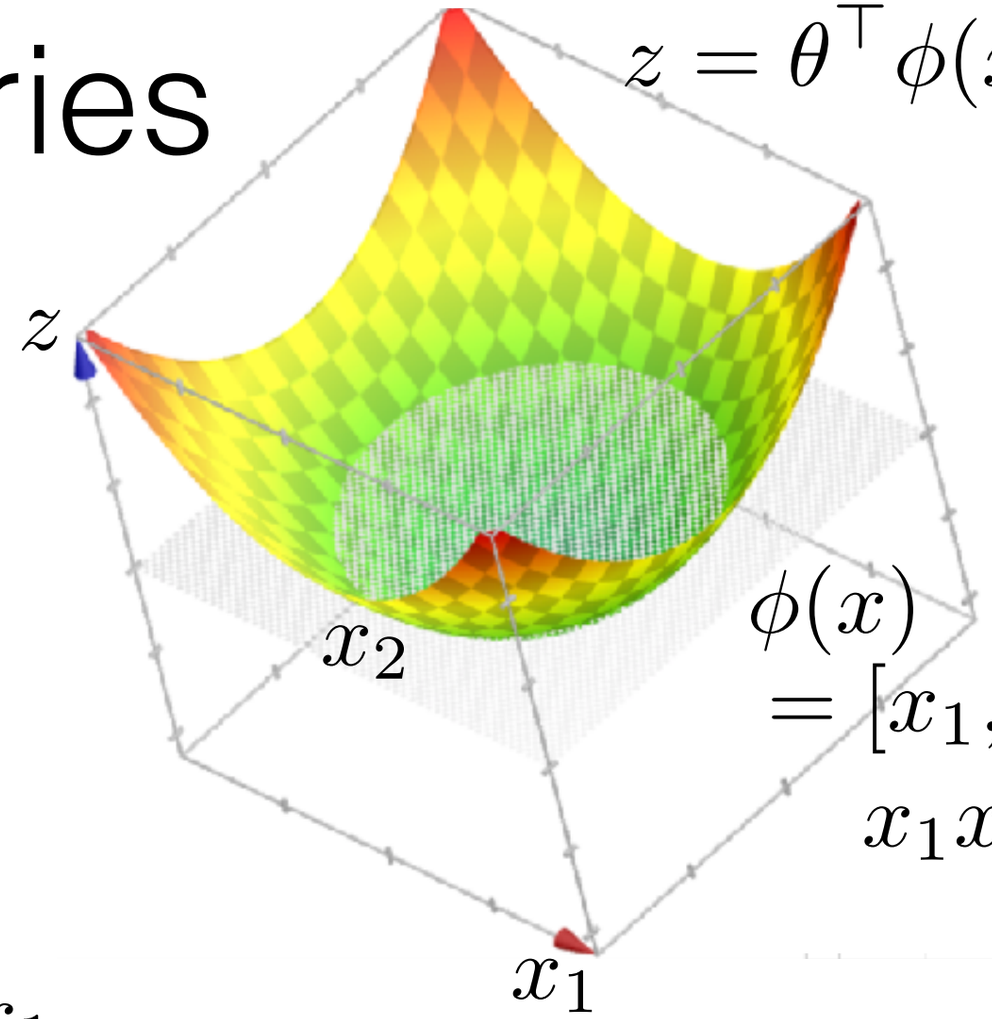



$$\phi(x) = [x_1, x_2, x_1^2, x_1 x_2, x_2^2]^\top$$

# Nonlinear boundaries

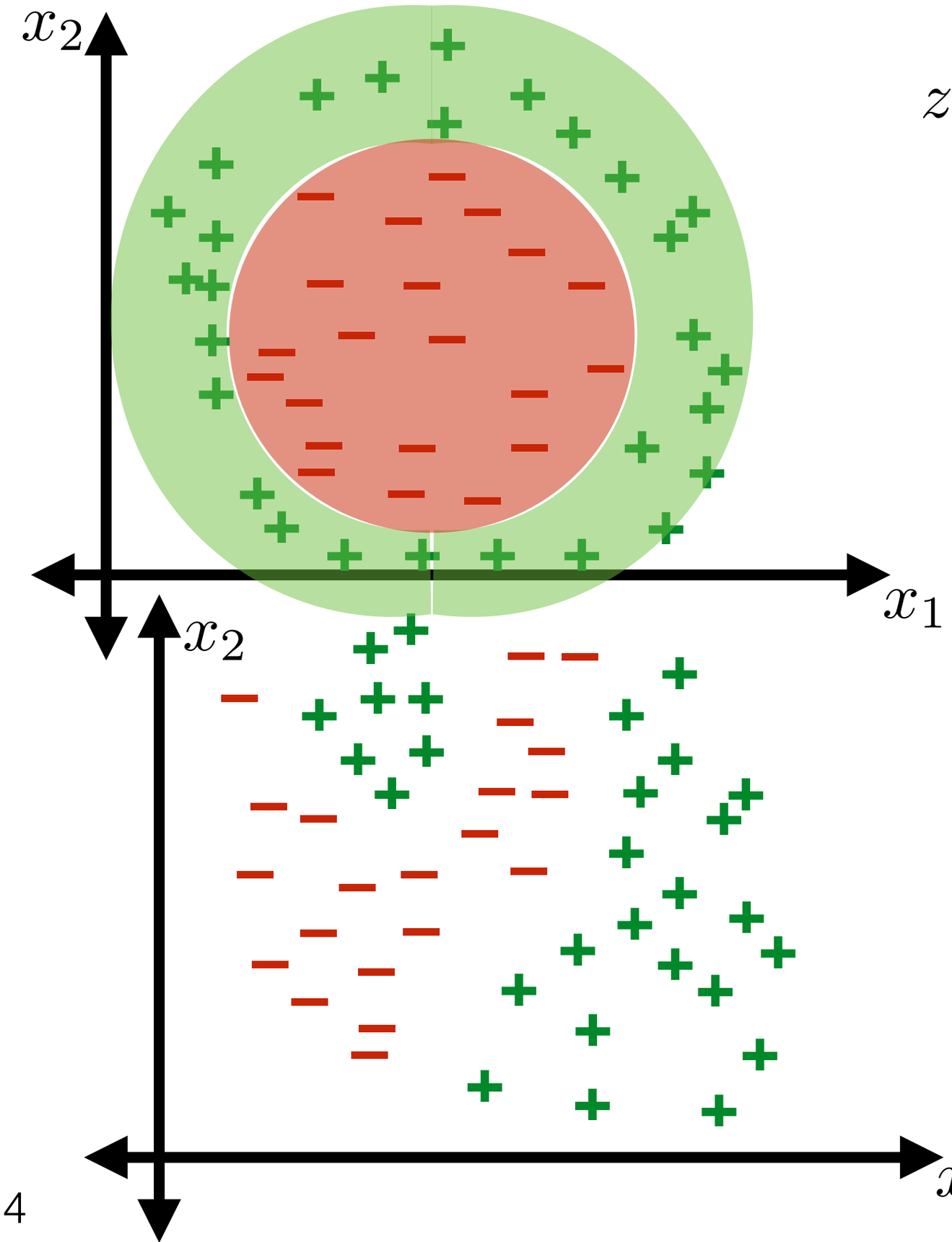


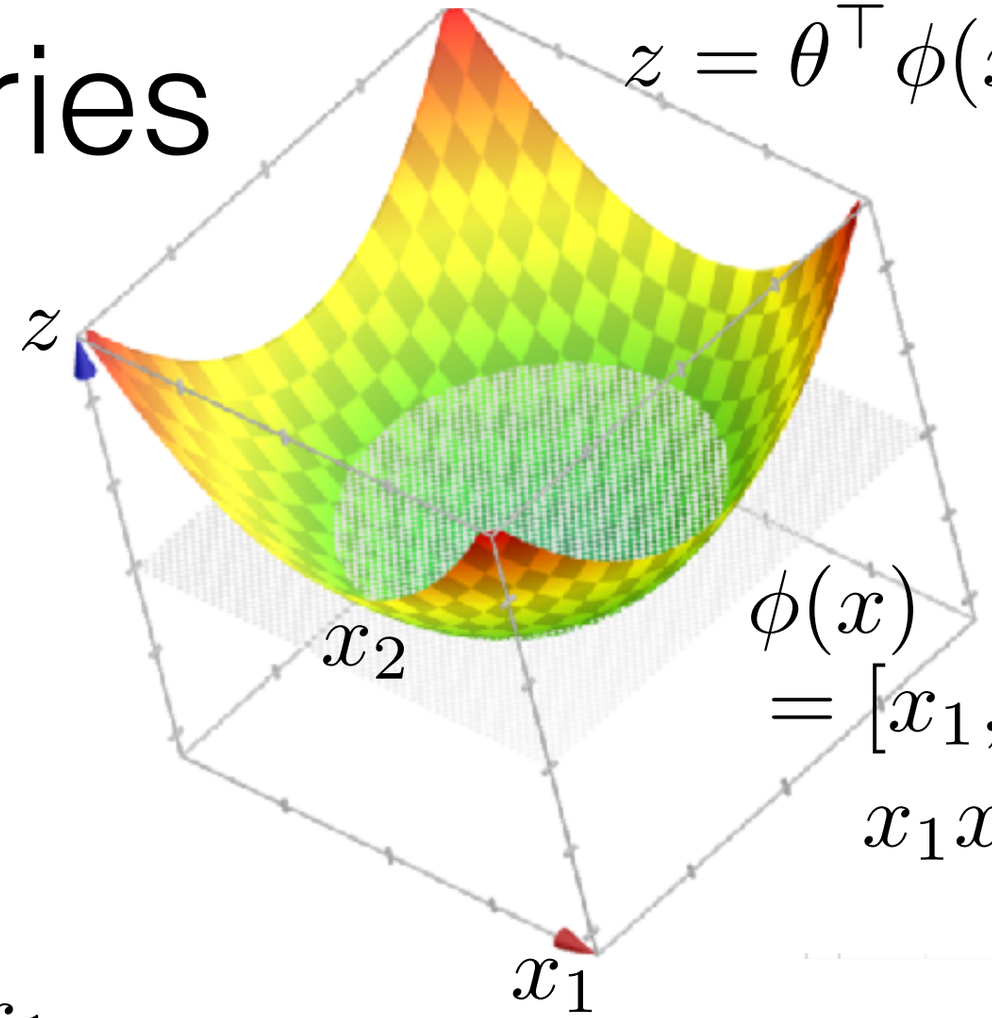

$$z = \theta^\top \phi(x) + \theta_0$$




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

# Nonlinear boundaries

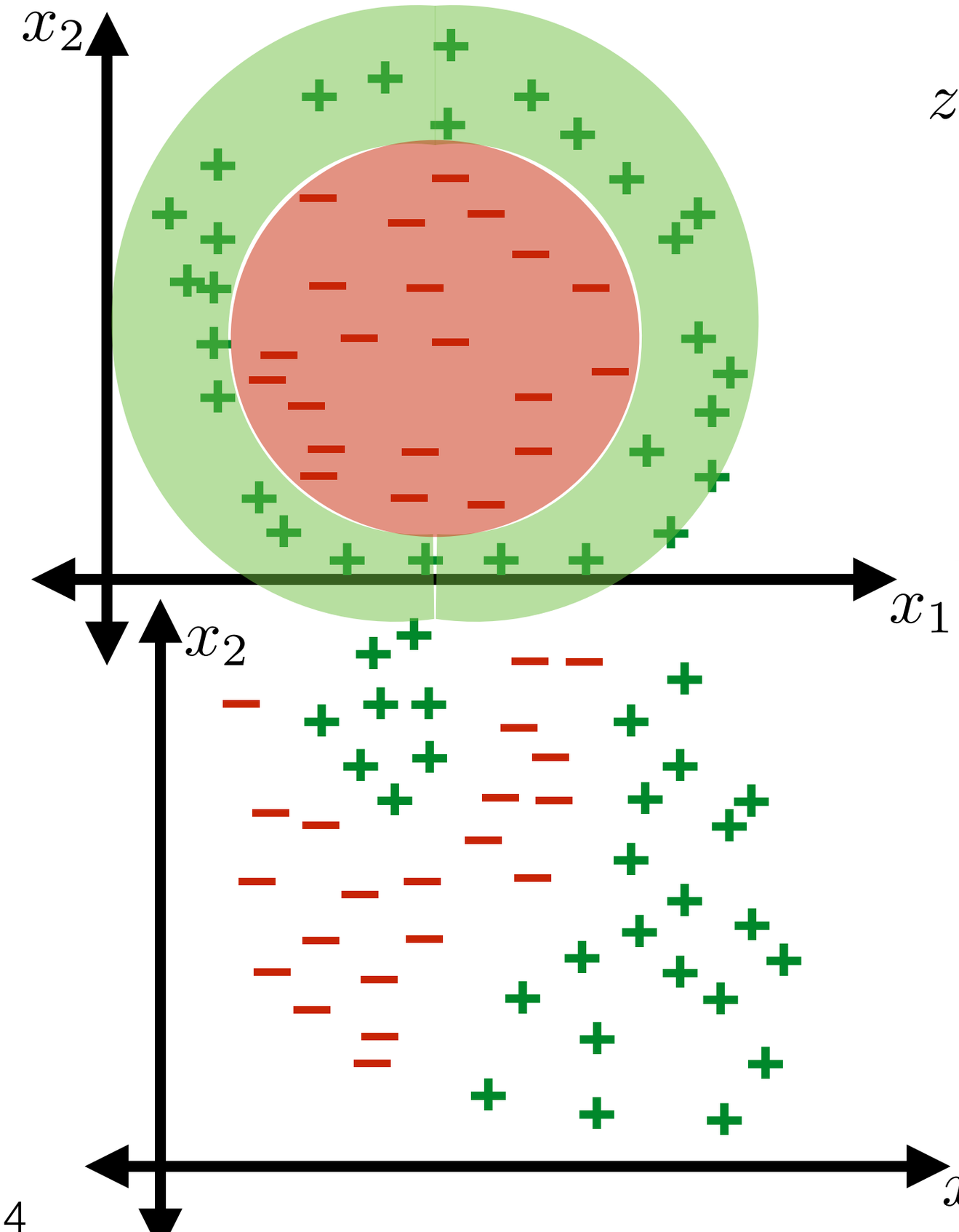


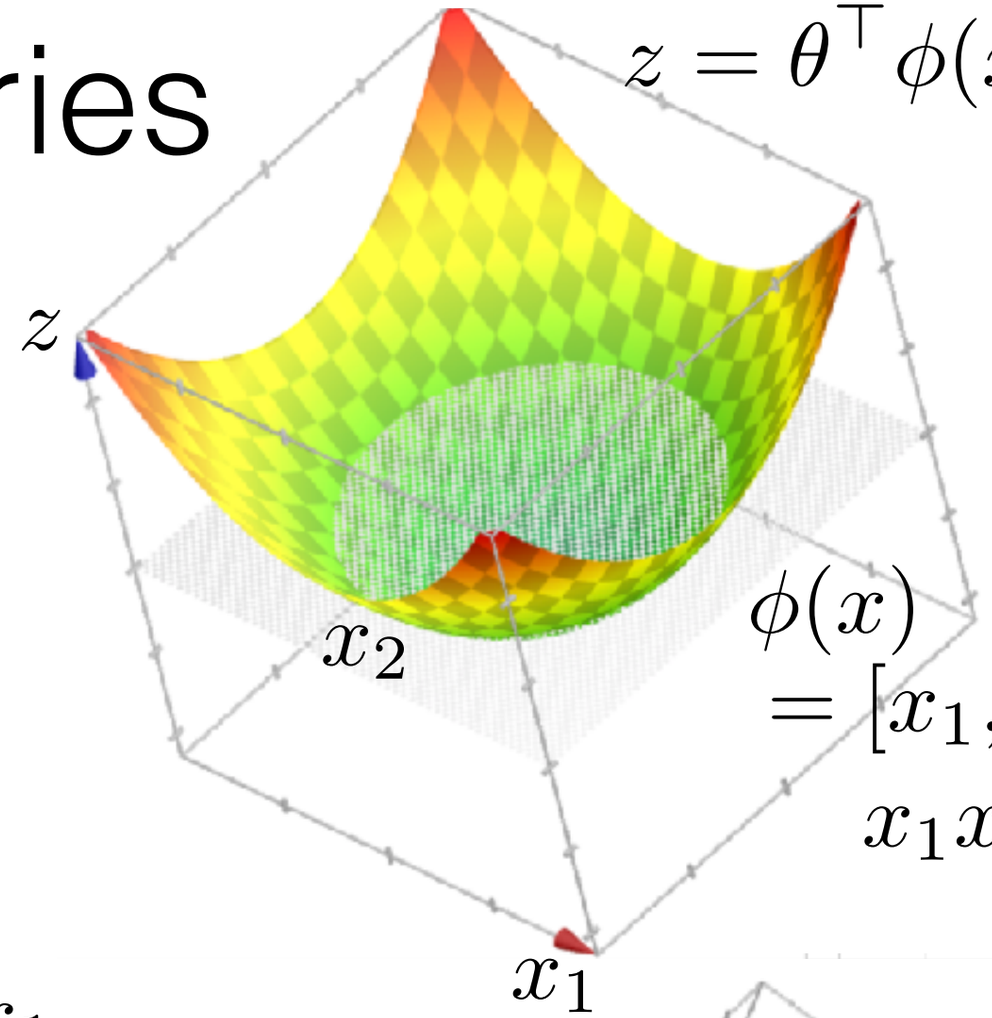

$$z = \theta^\top \phi(x) + \theta_0$$

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

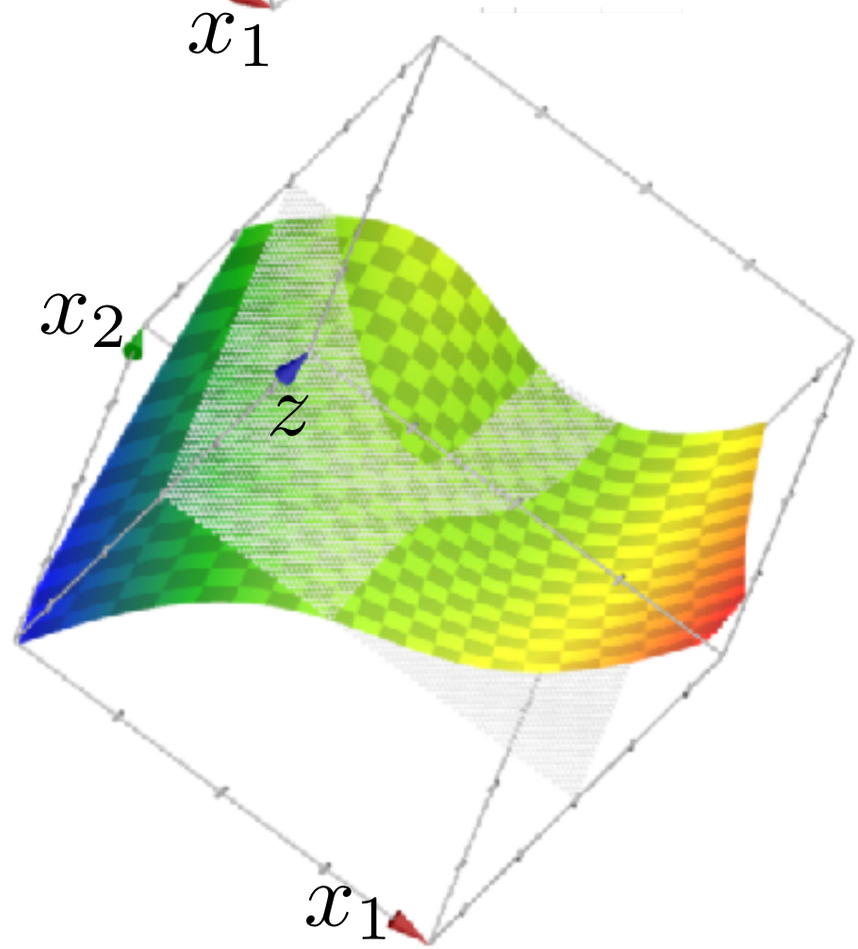
# Nonlinear boundaries



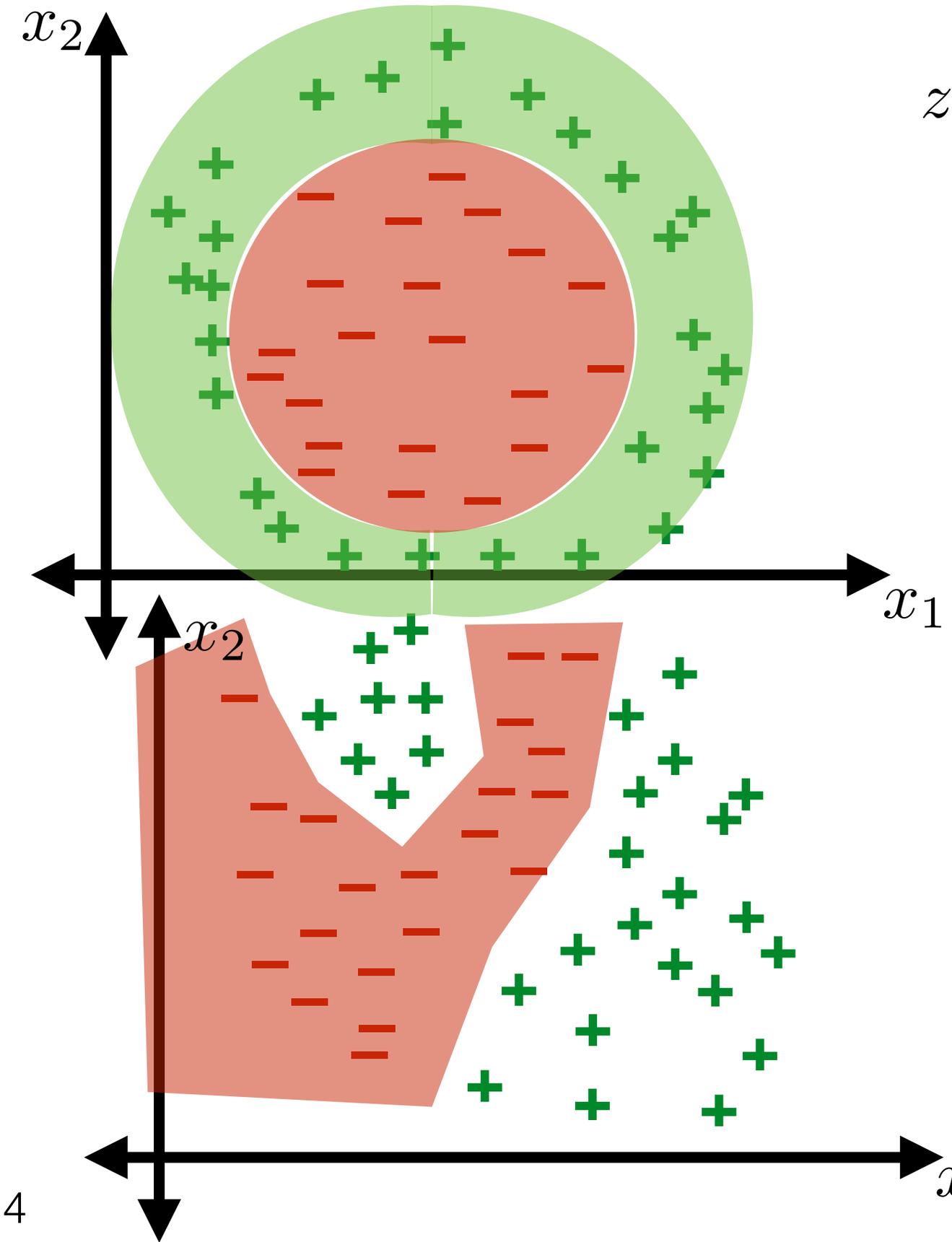

$$z = \theta^\top \phi(x) + \theta_0$$



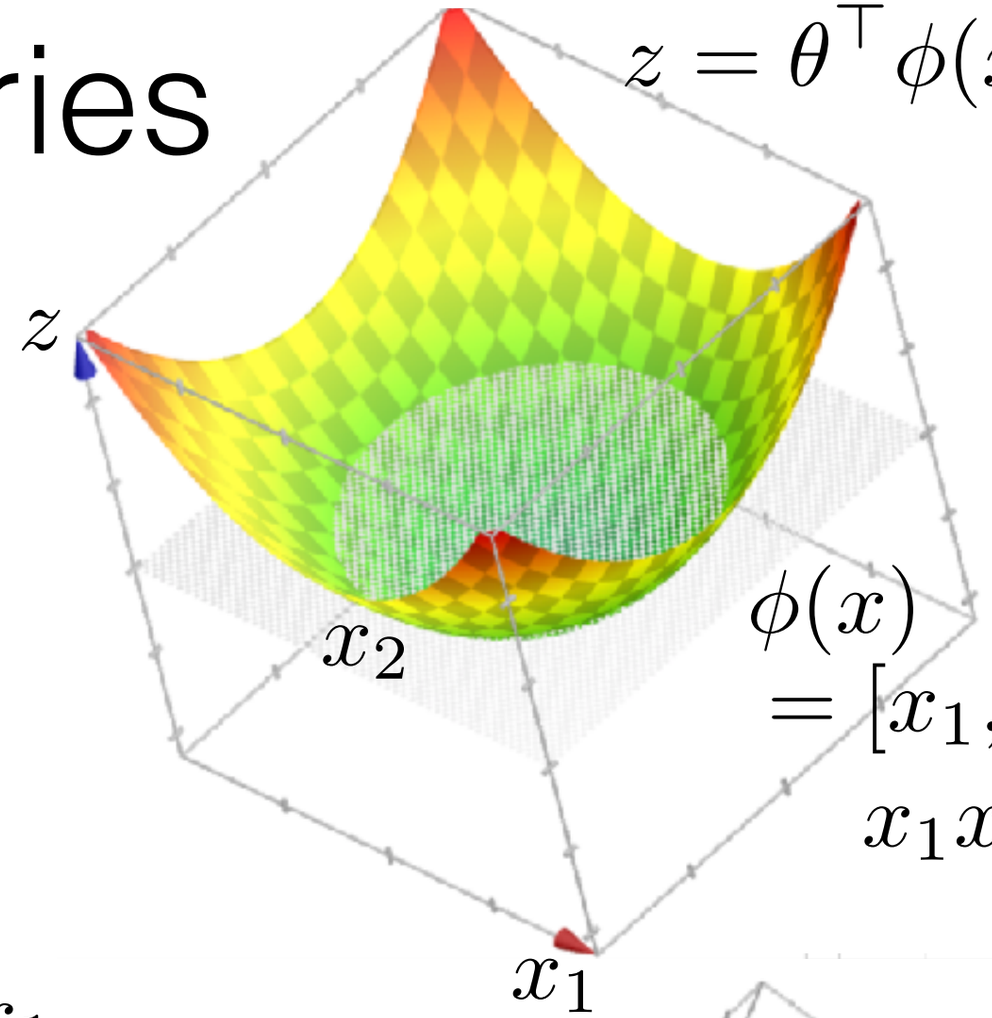

$$\phi(x) = [x_1, x_2, x_1^2, x_1 x_2, x_2^2]^\top$$



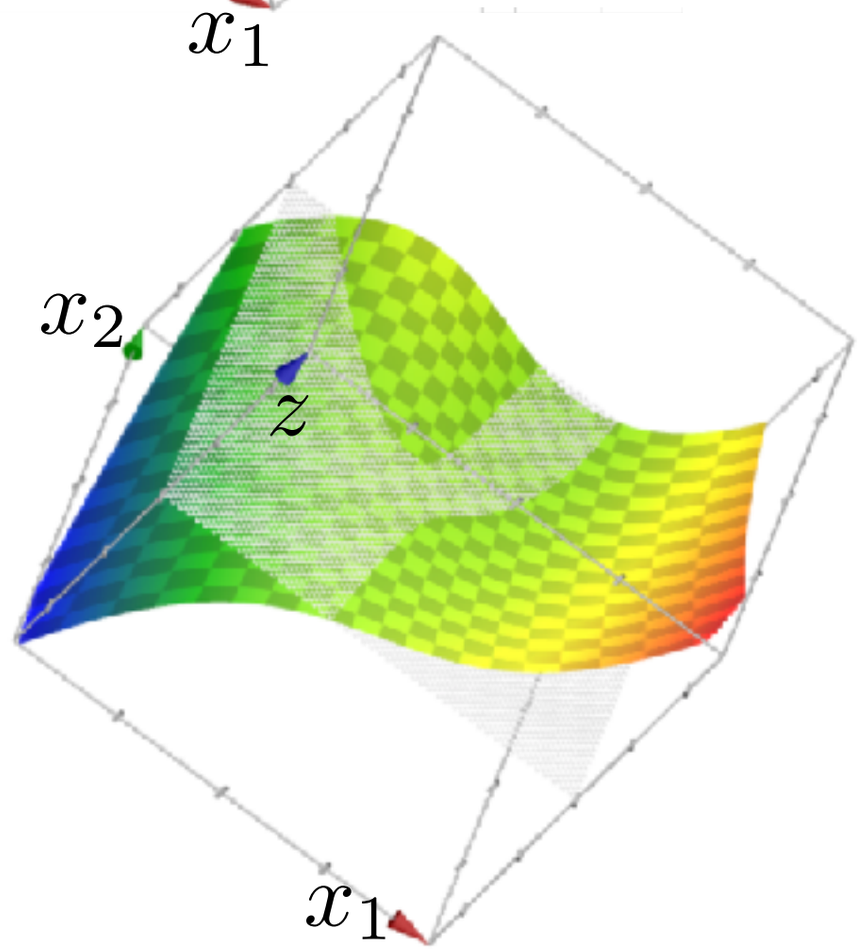
# Nonlinear boundaries



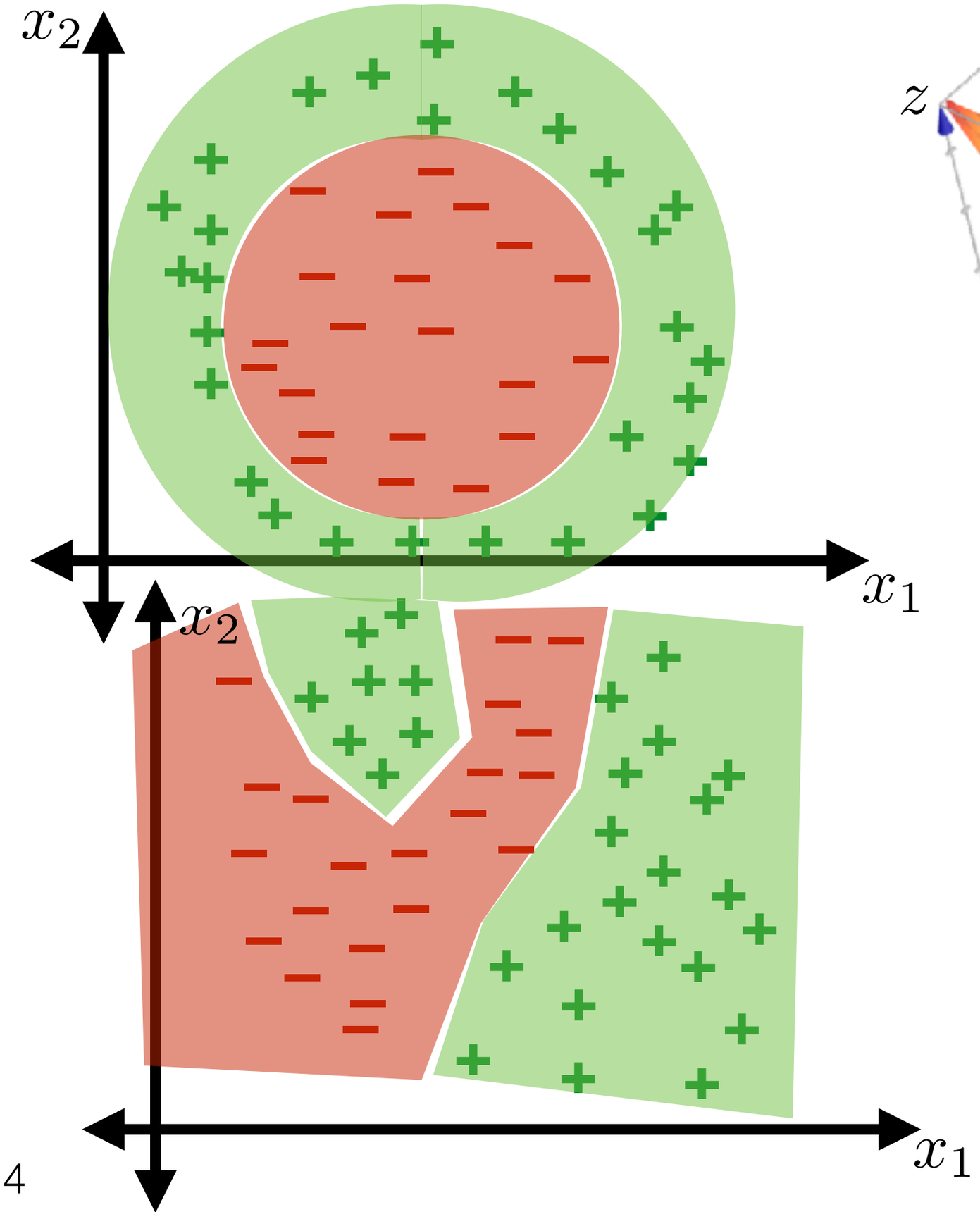

$z = \theta^\top \phi(x) + \theta_0$



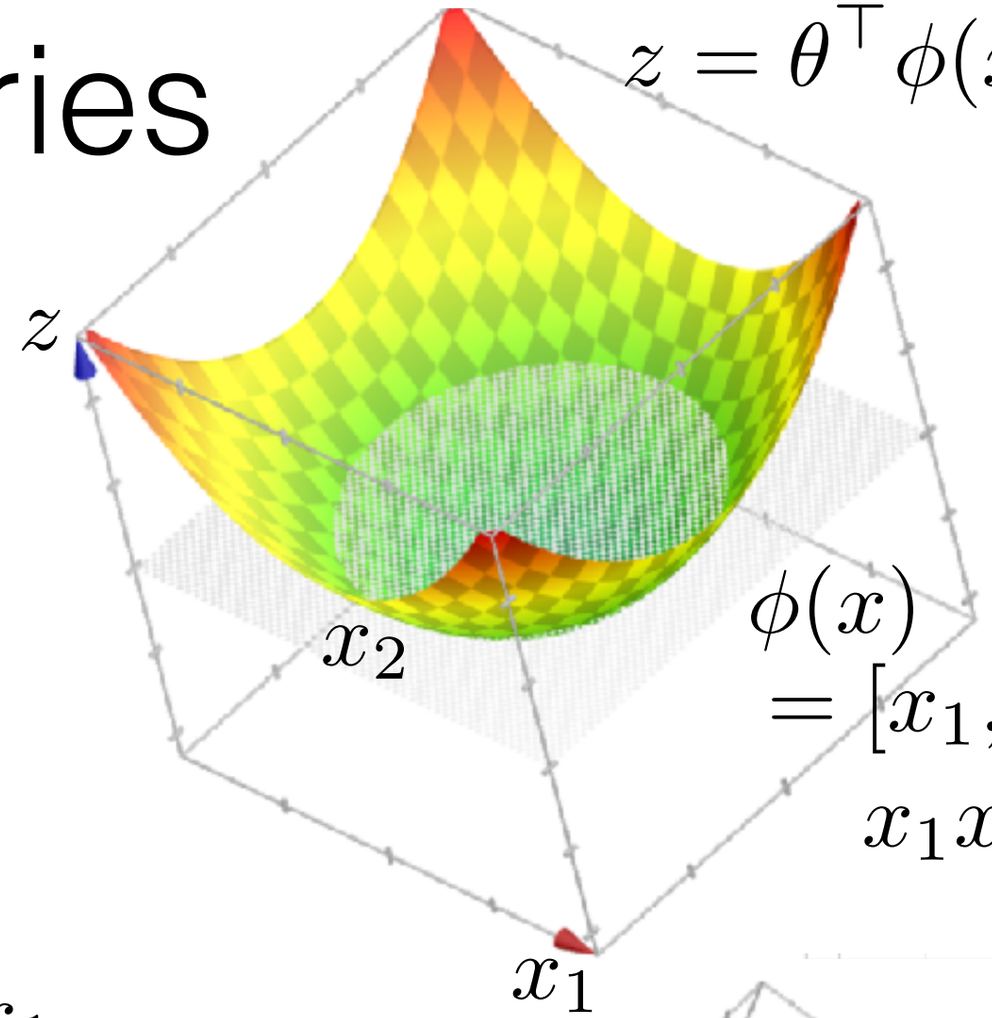

$$\phi(x) = [x_1, x_2, x_1^2, x_1 x_2, x_2^2]^\top$$



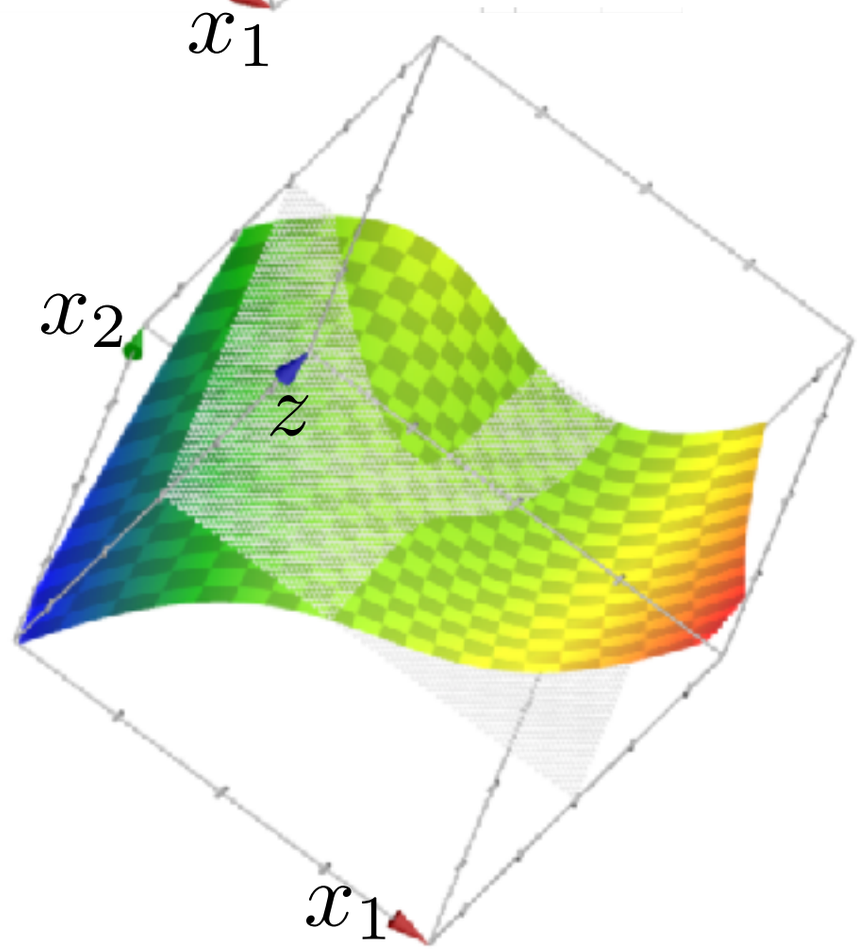
# Nonlinear boundaries



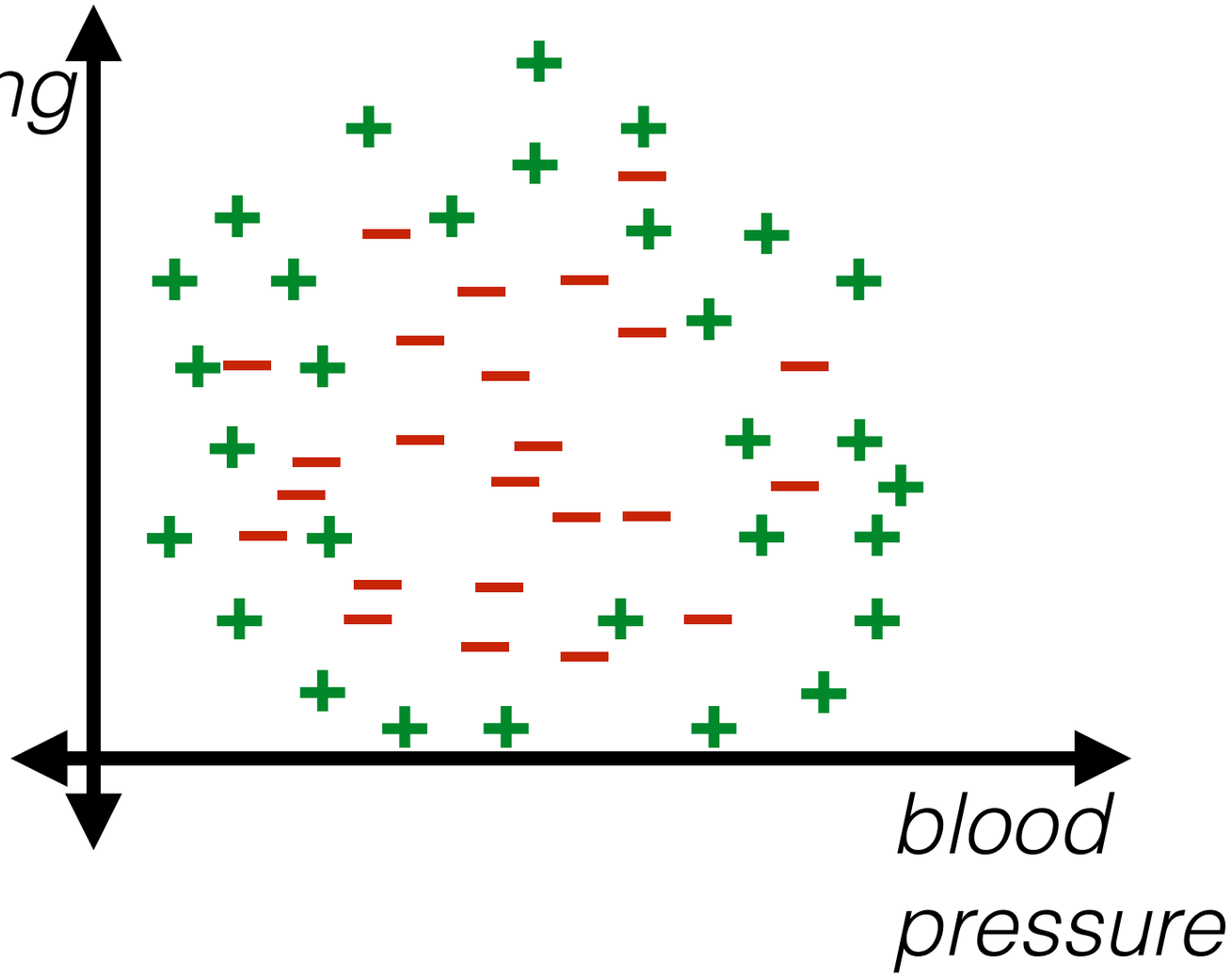

$z = \theta^\top \phi(x) + \theta_0$



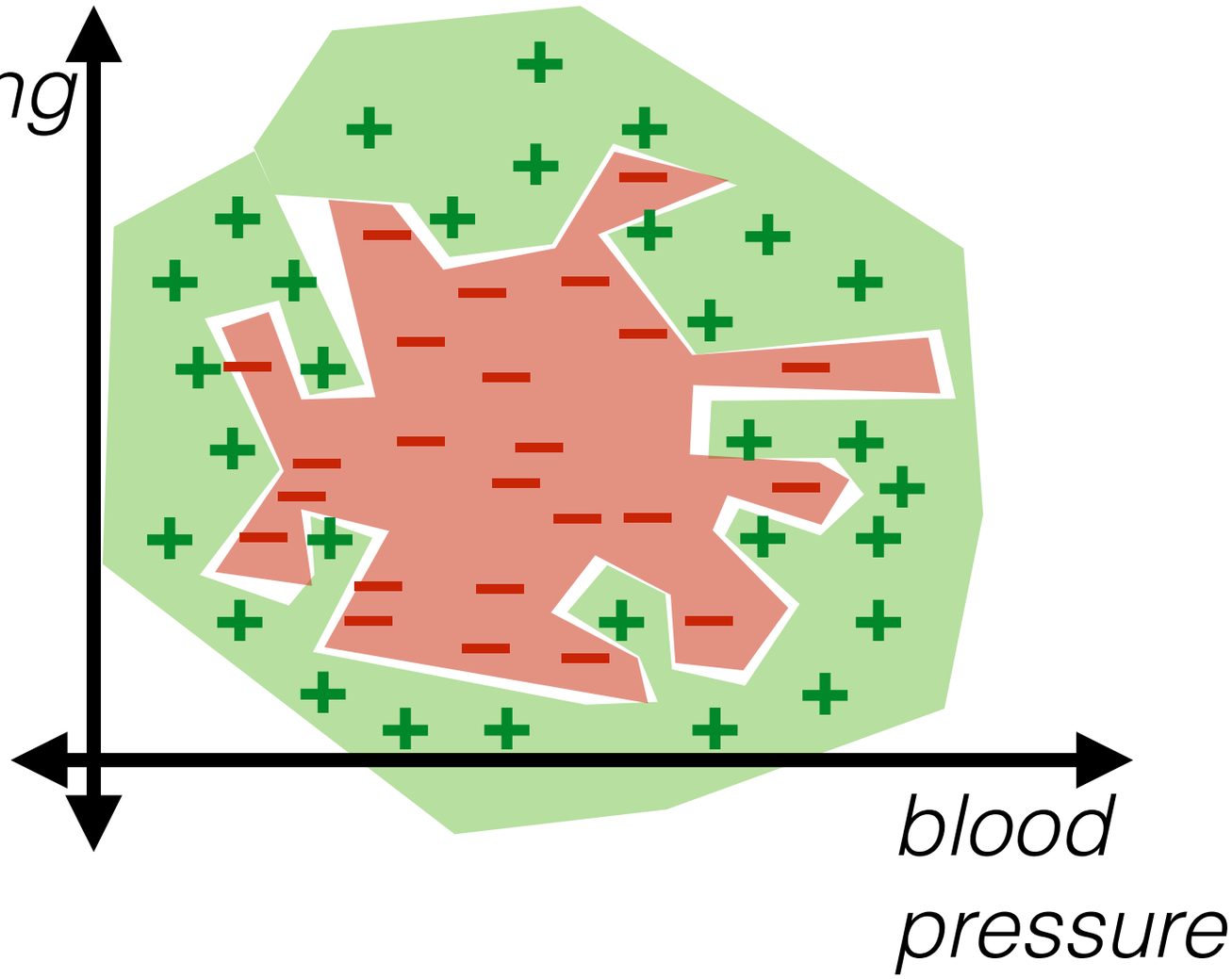

$$\phi(x) = [x_1, x_2, x_1^2, x_1 x_2, x_2^2]^\top$$



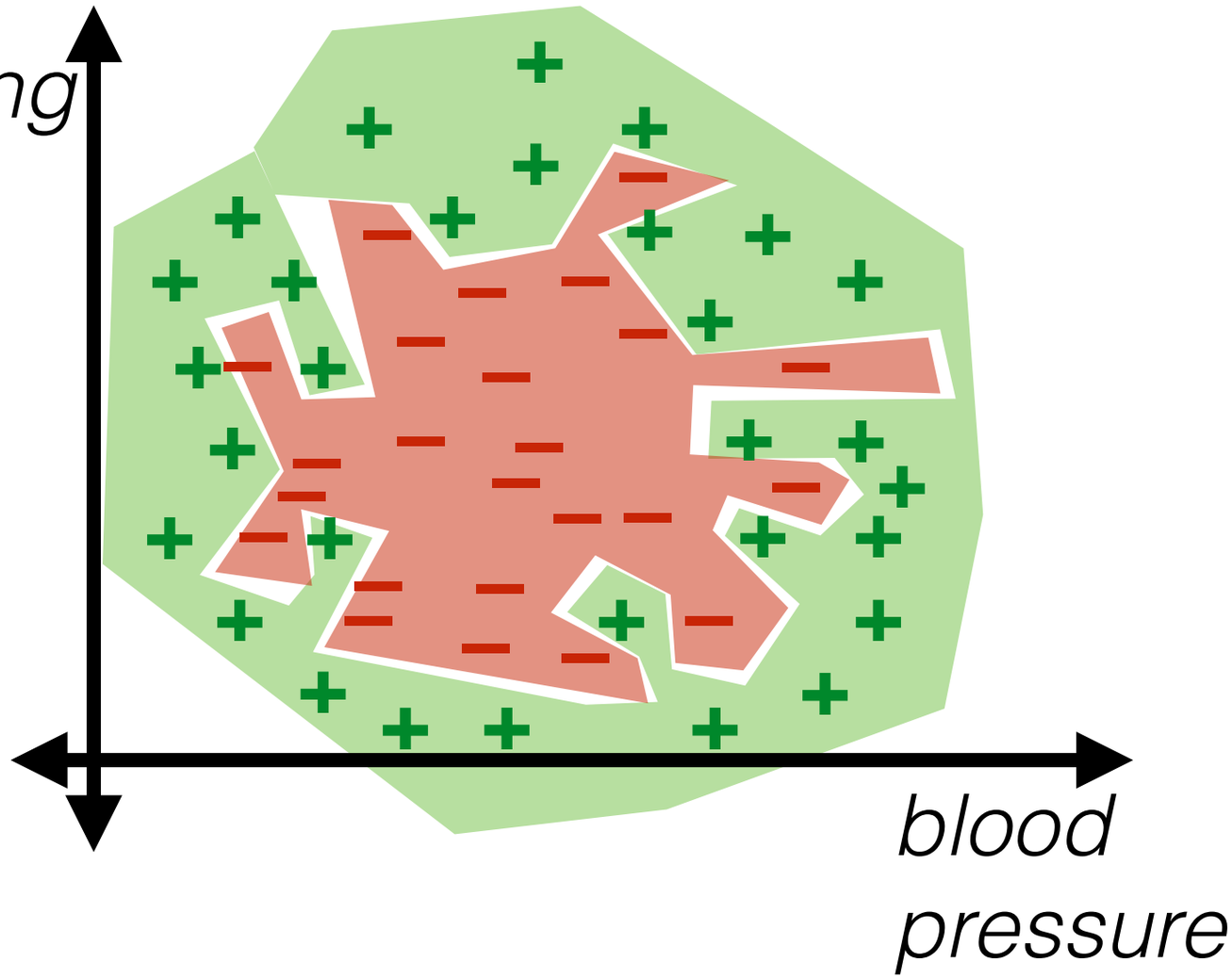
# Nonlinear boundaries



# Nonlinear boundaries

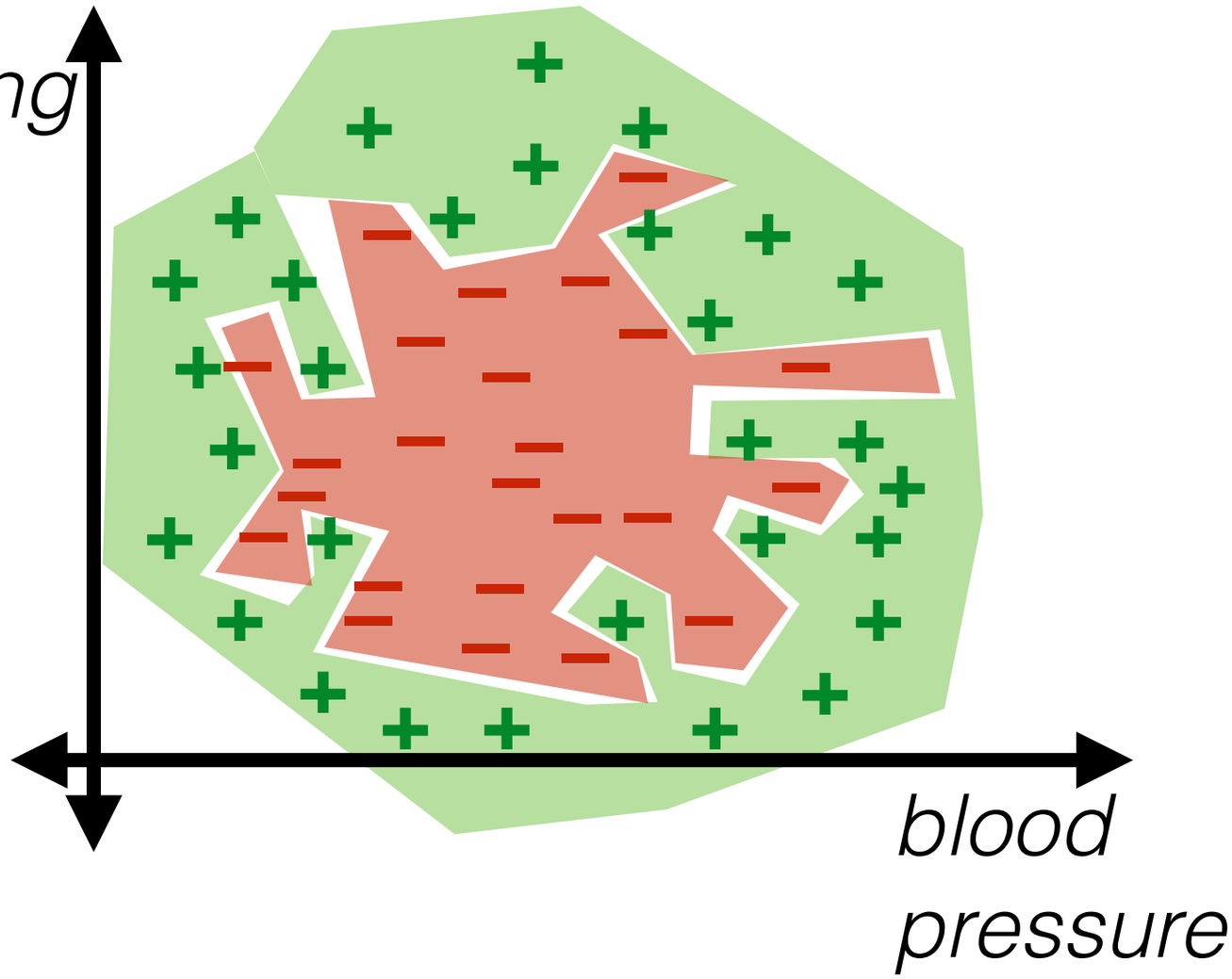


# Nonlinear boundaries



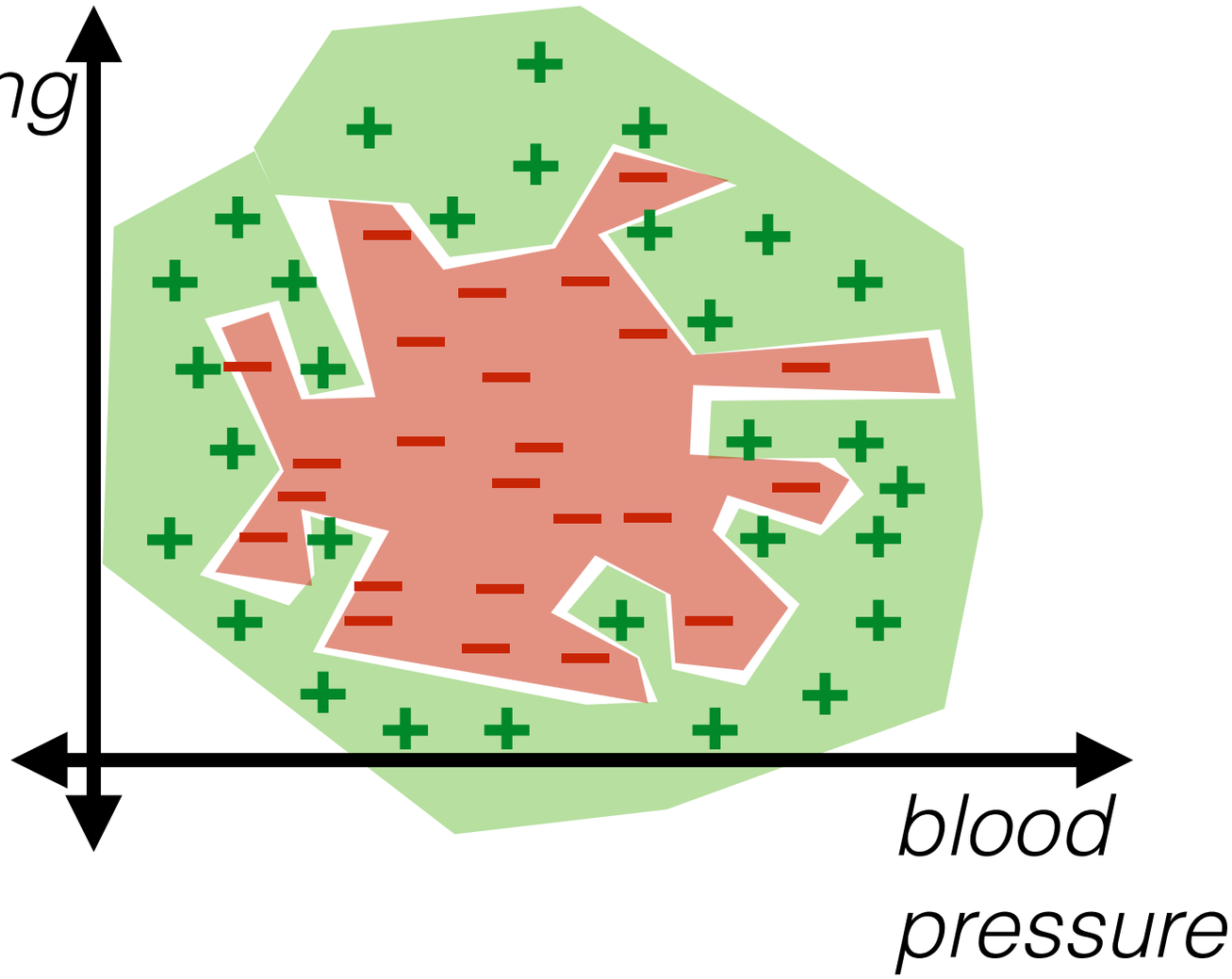

- Training error is 0!

# Nonlinear boundaries




- Training error is 0!
- But seems like our classifier is overfitting

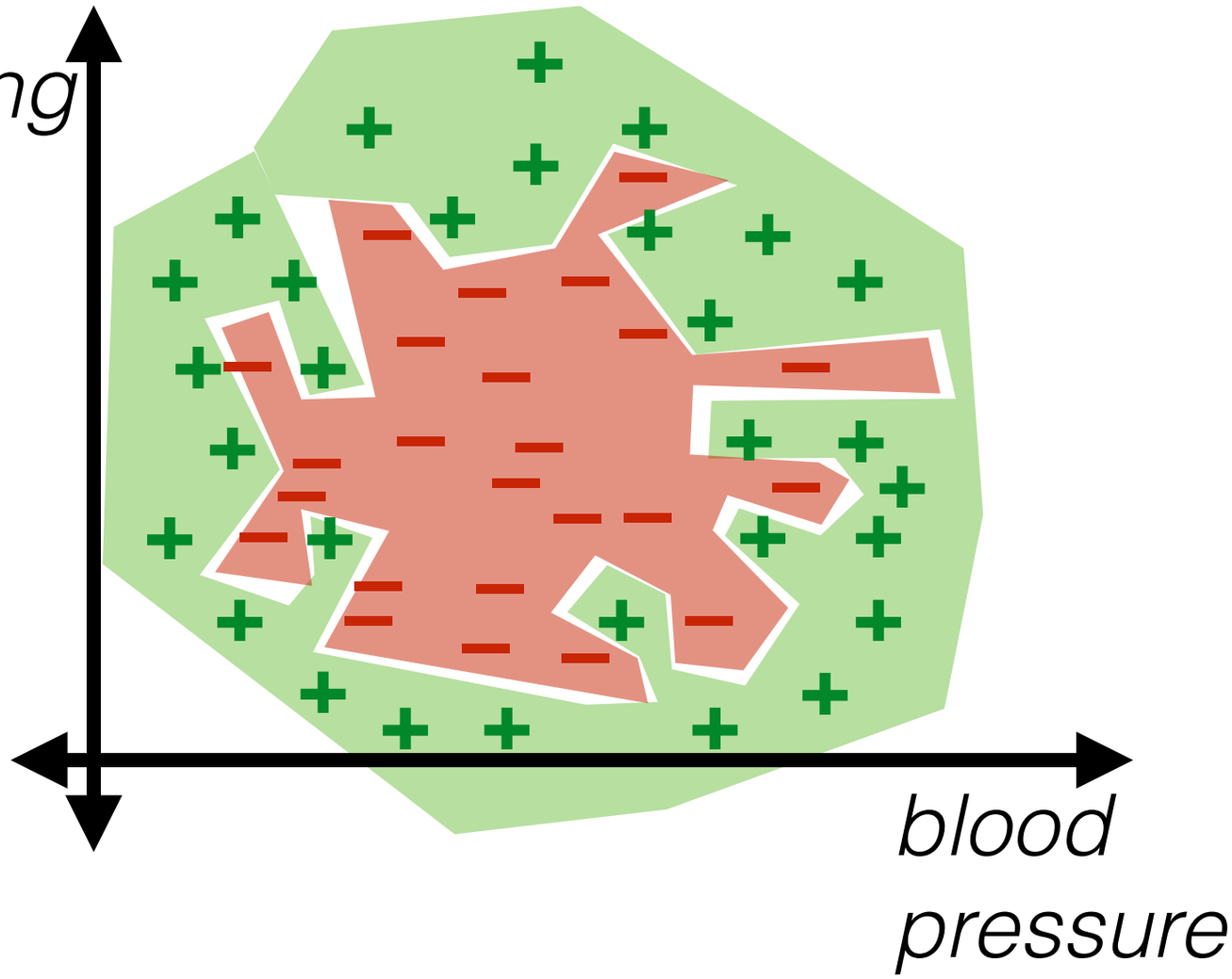
# Nonlinear boundaries




- Training error is 0!
- But seems like our classifier is overfitting

- Benefit of polynomial features: can be super flexible if we use polynomials up to a high degree

# Nonlinear boundaries

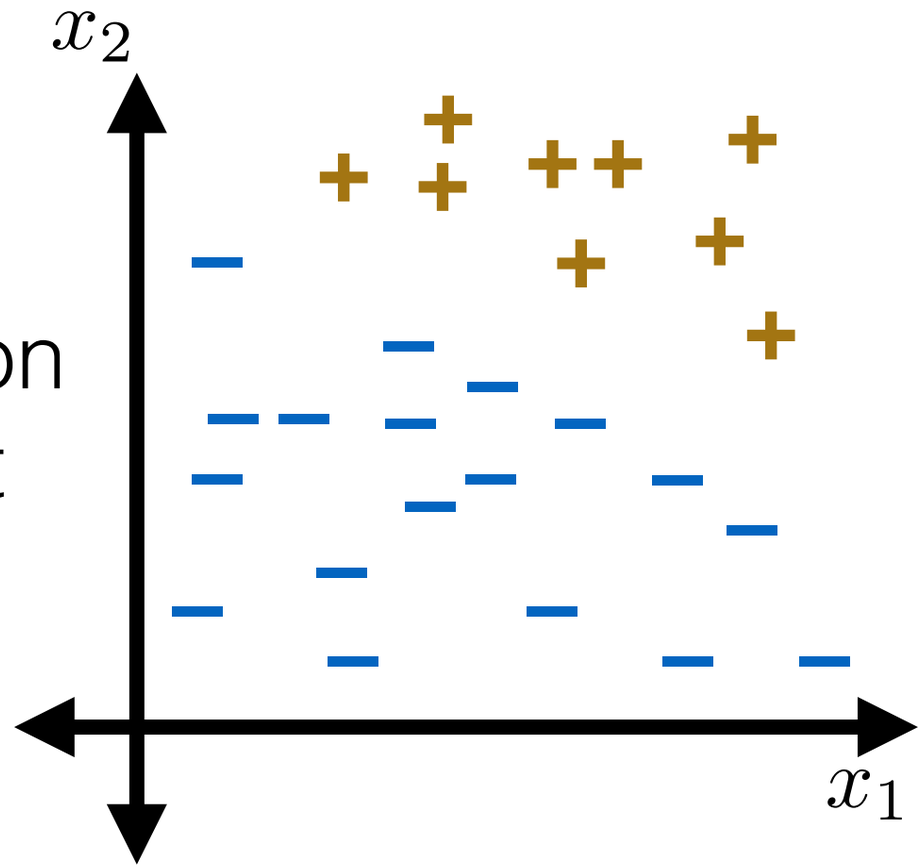



- Training error is 0!
- But seems like our classifier is overfitting

- Benefit of polynomial features: can be super flexible if we use polynomials up to a high degree
- If we use polynomials up to a high degree, we're prone to overfit

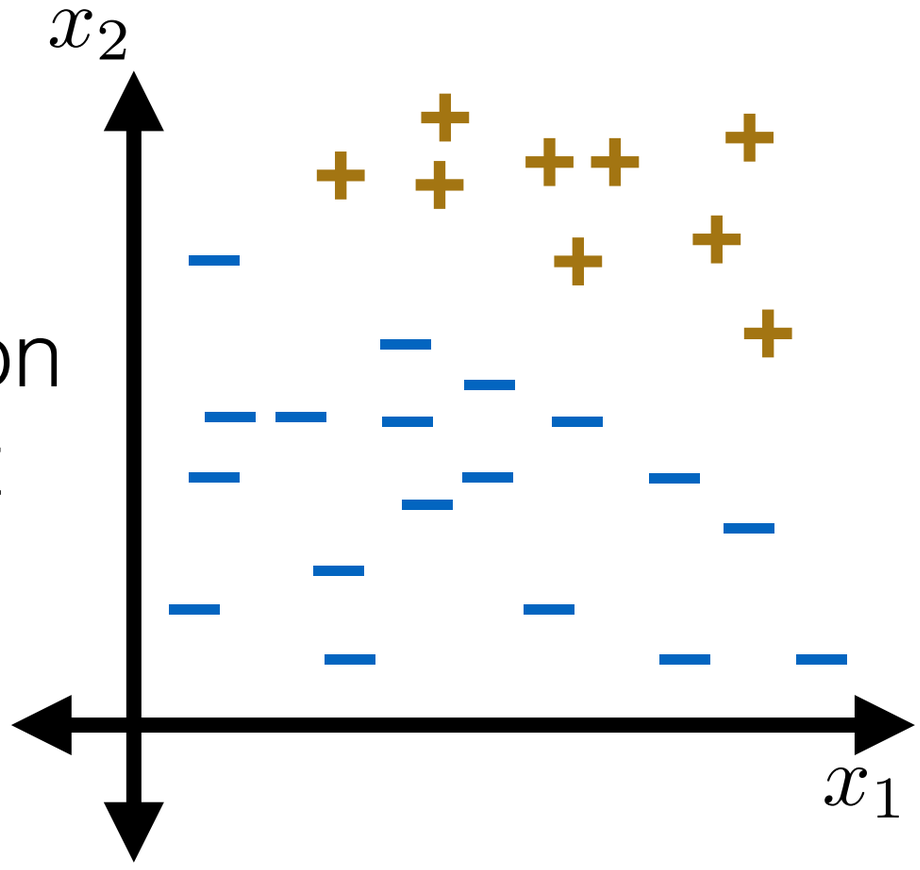
# Recall

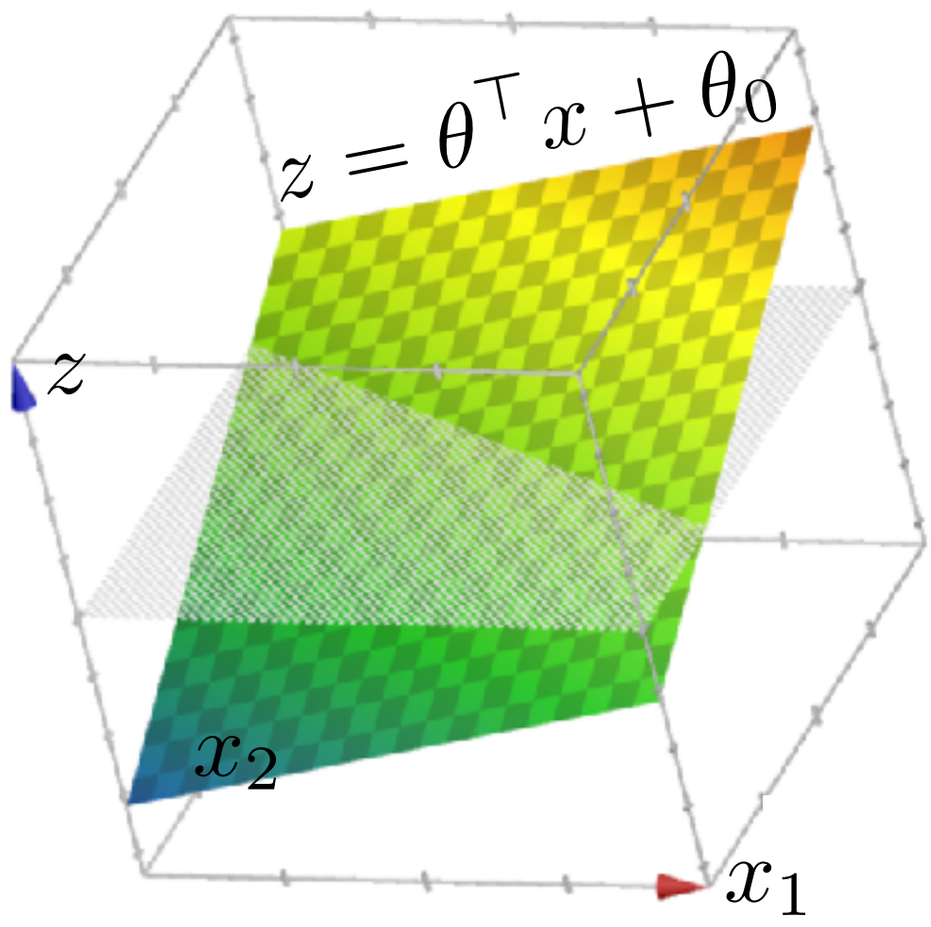
- Linear classification with default features:



# Recall

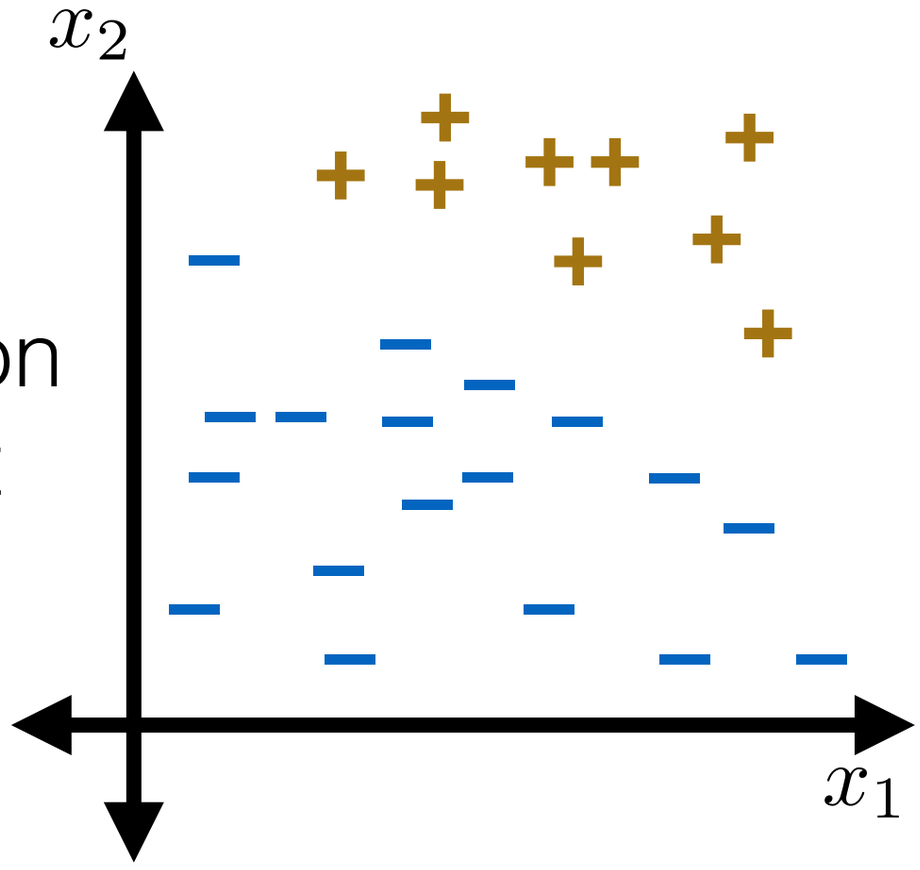
- Linear classification with default features:

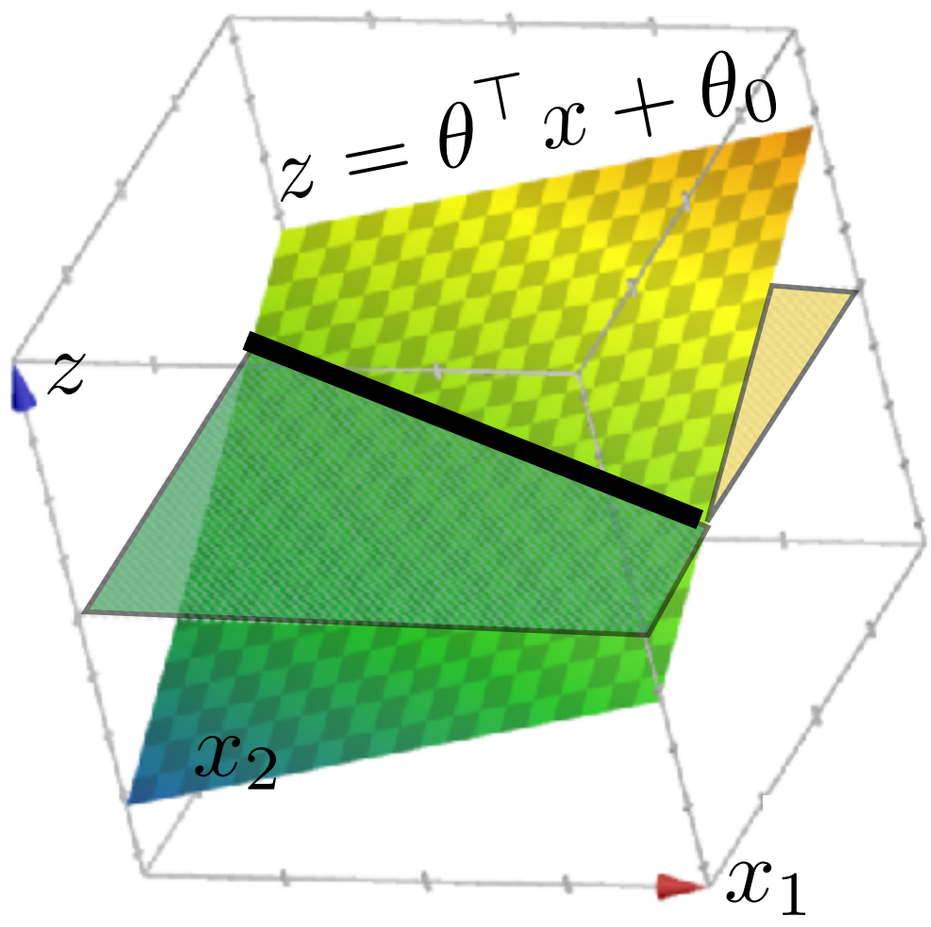


# Recall

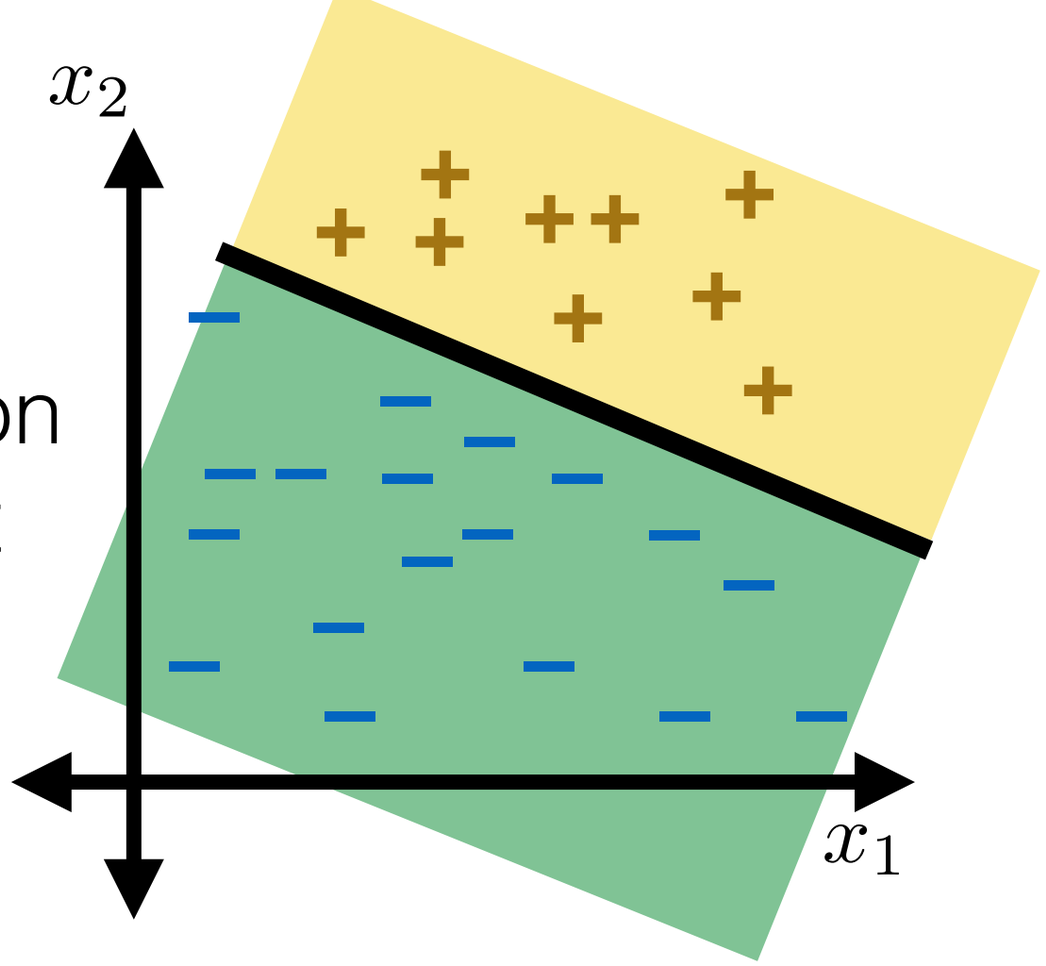
- Linear classification with default features:

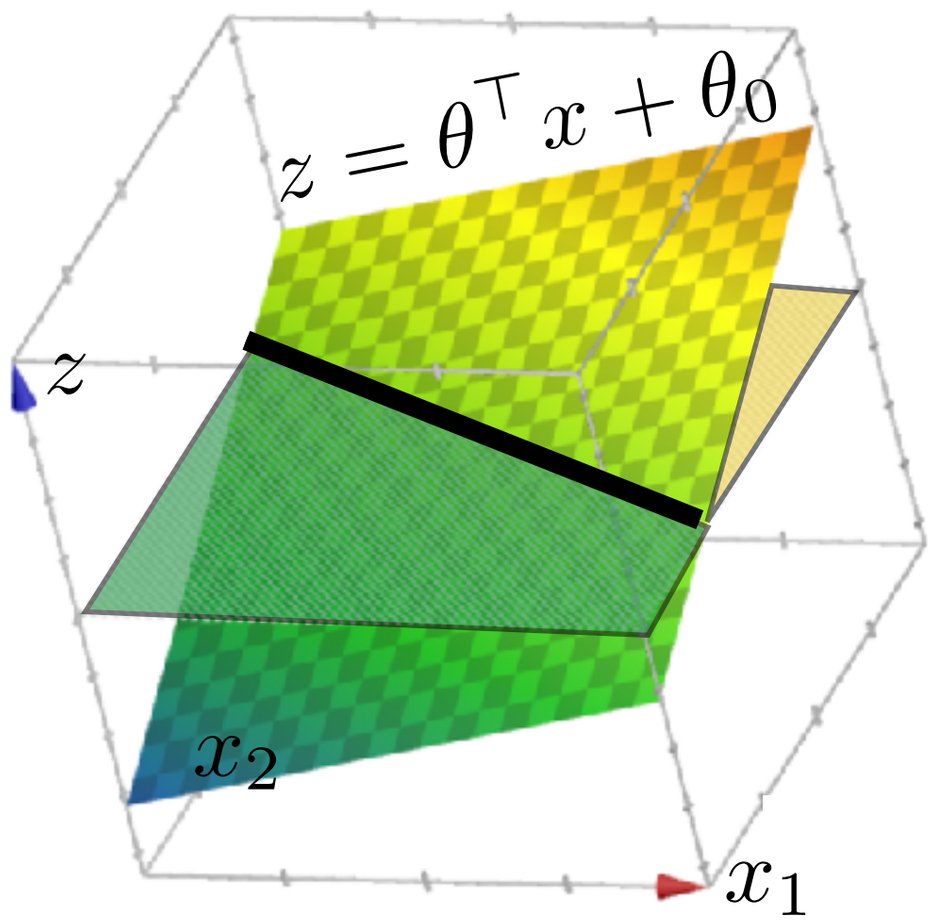


# Recall

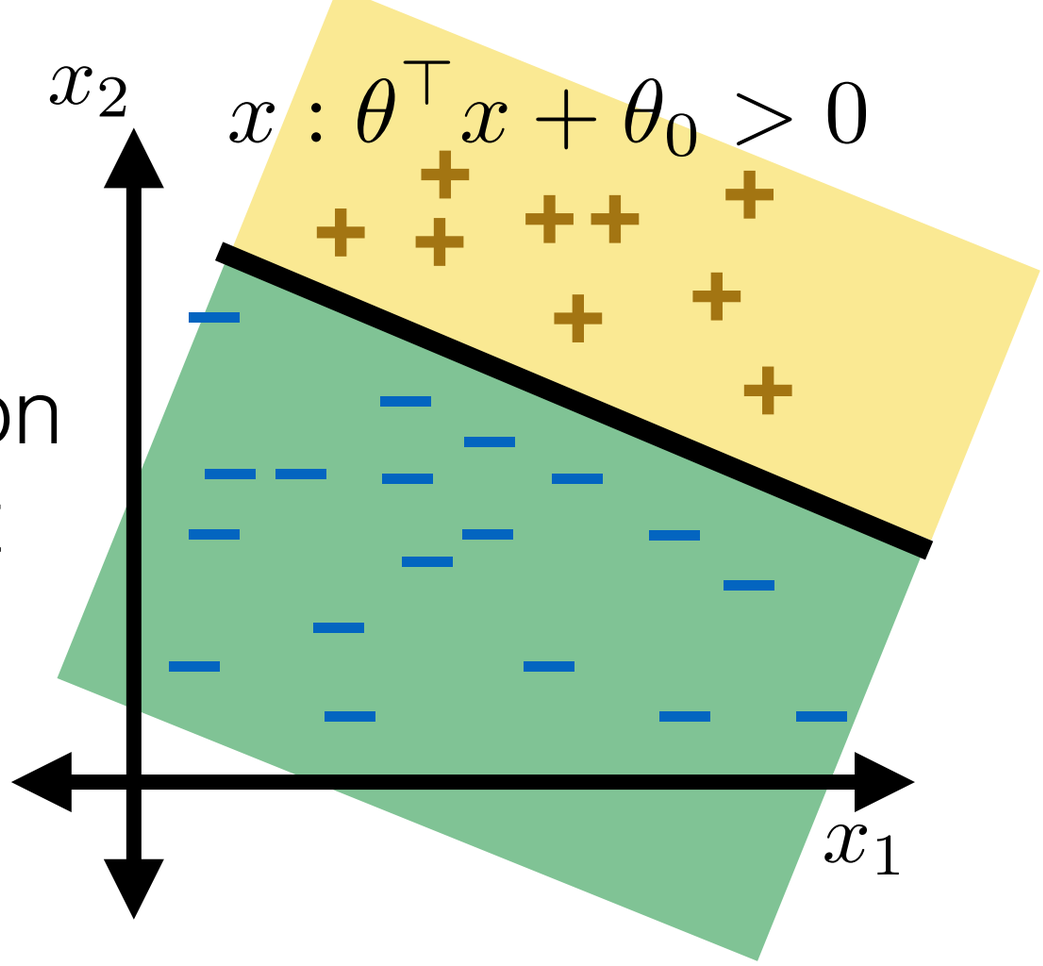
- Linear classification with default features:

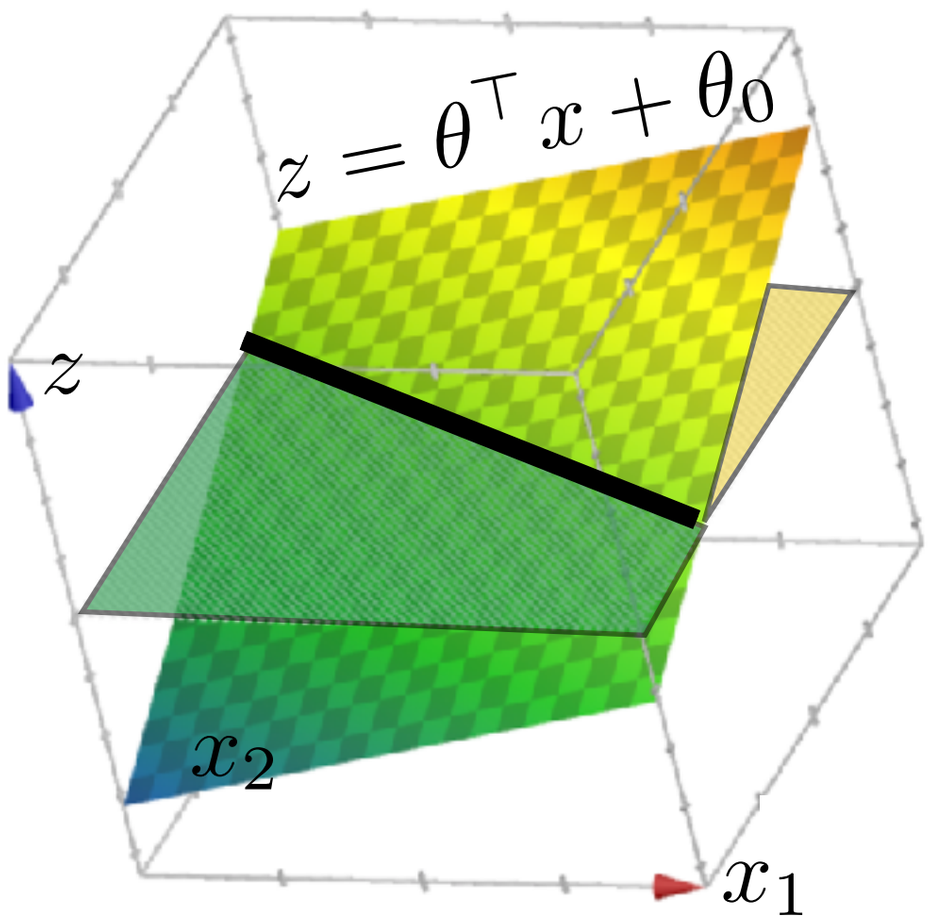


# Recall

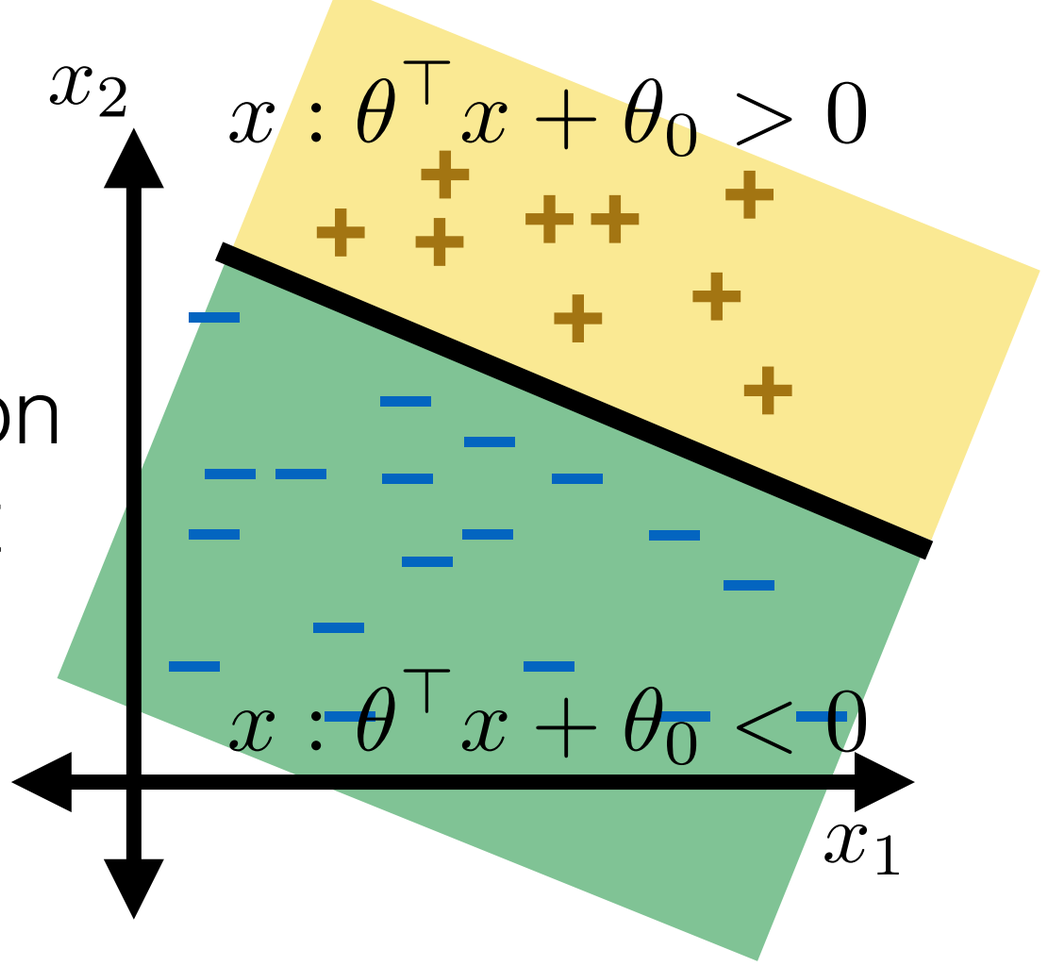
- Linear classification with default features:

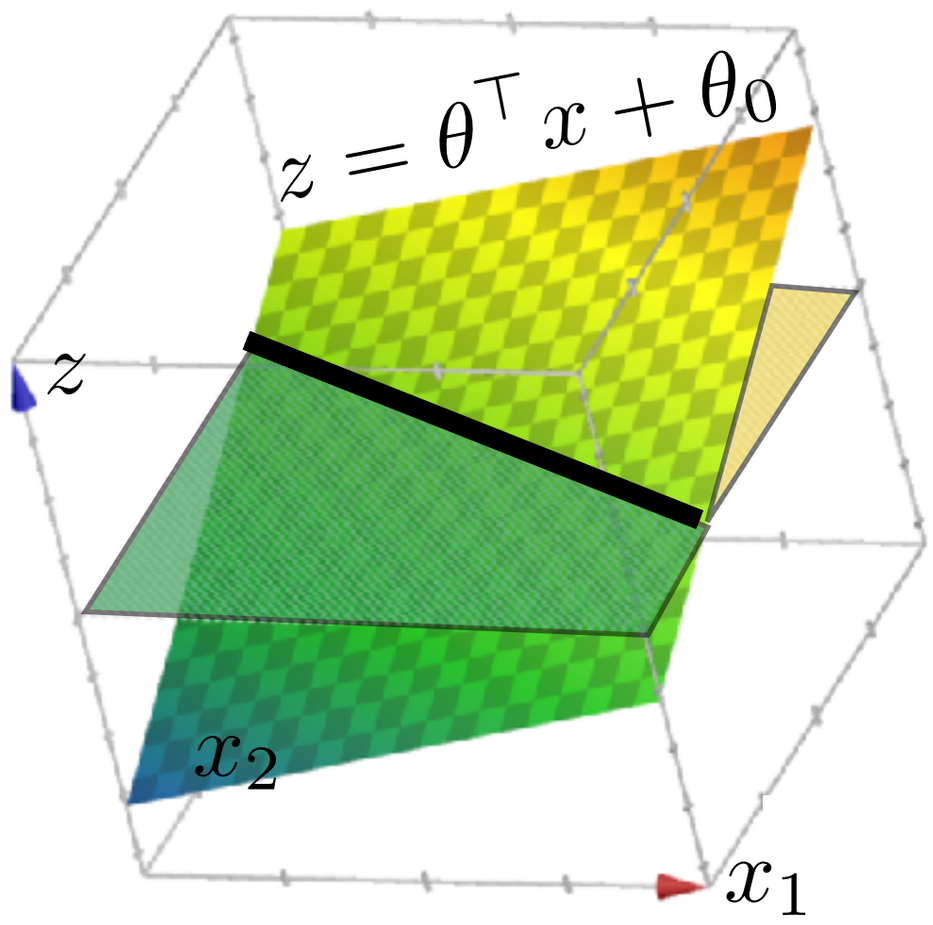


# Recall

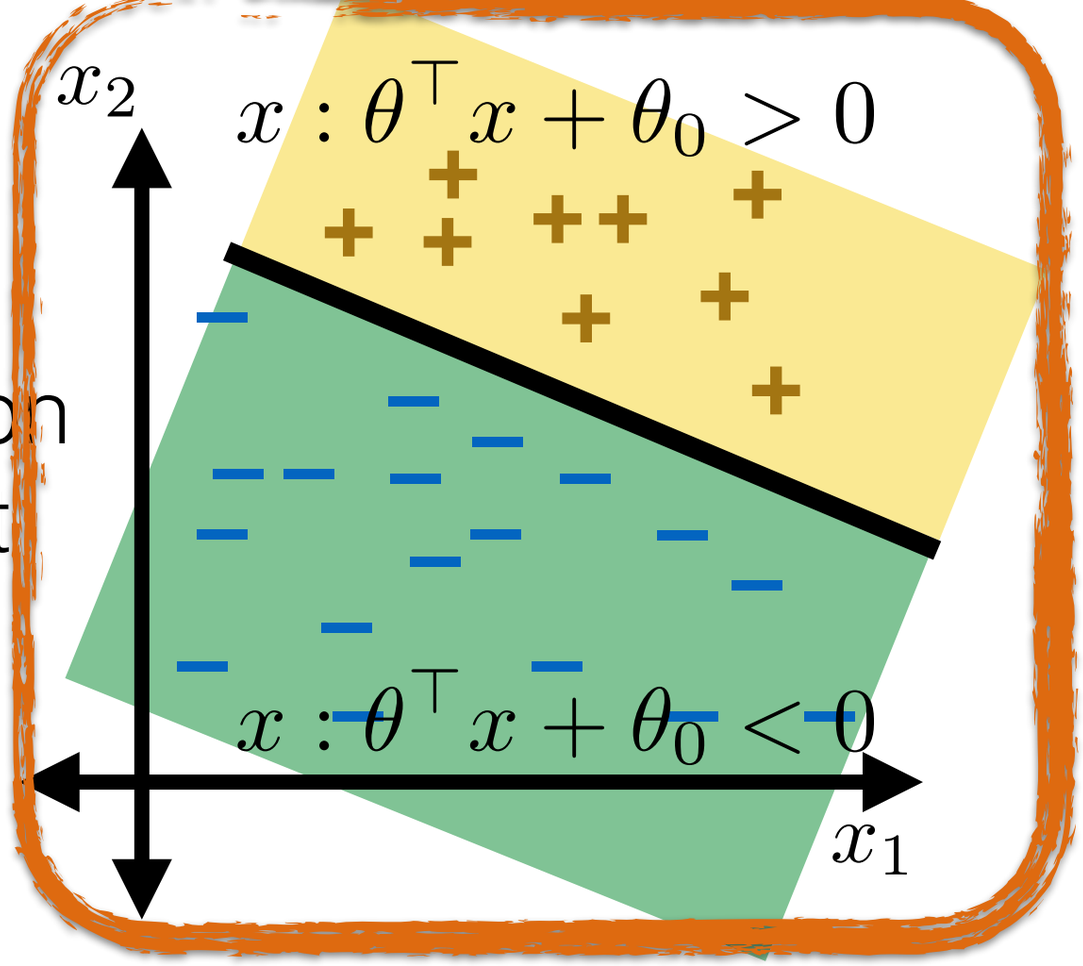
- Linear classification with default features:

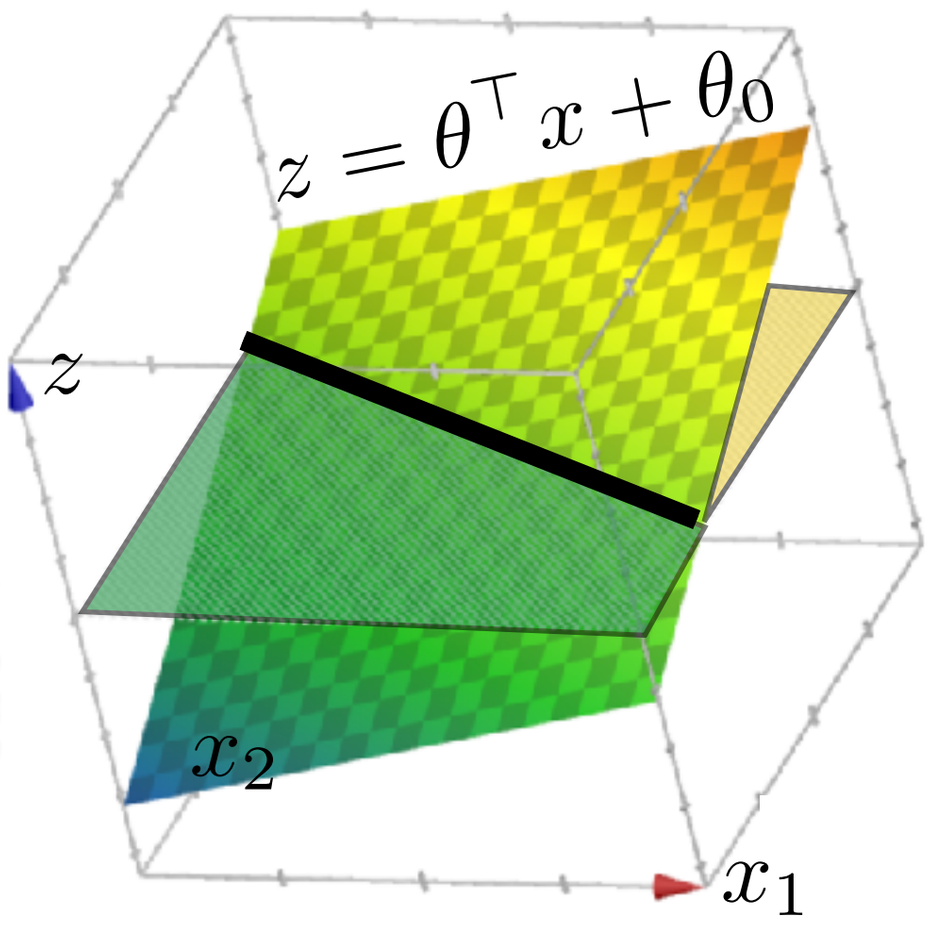


# Recall

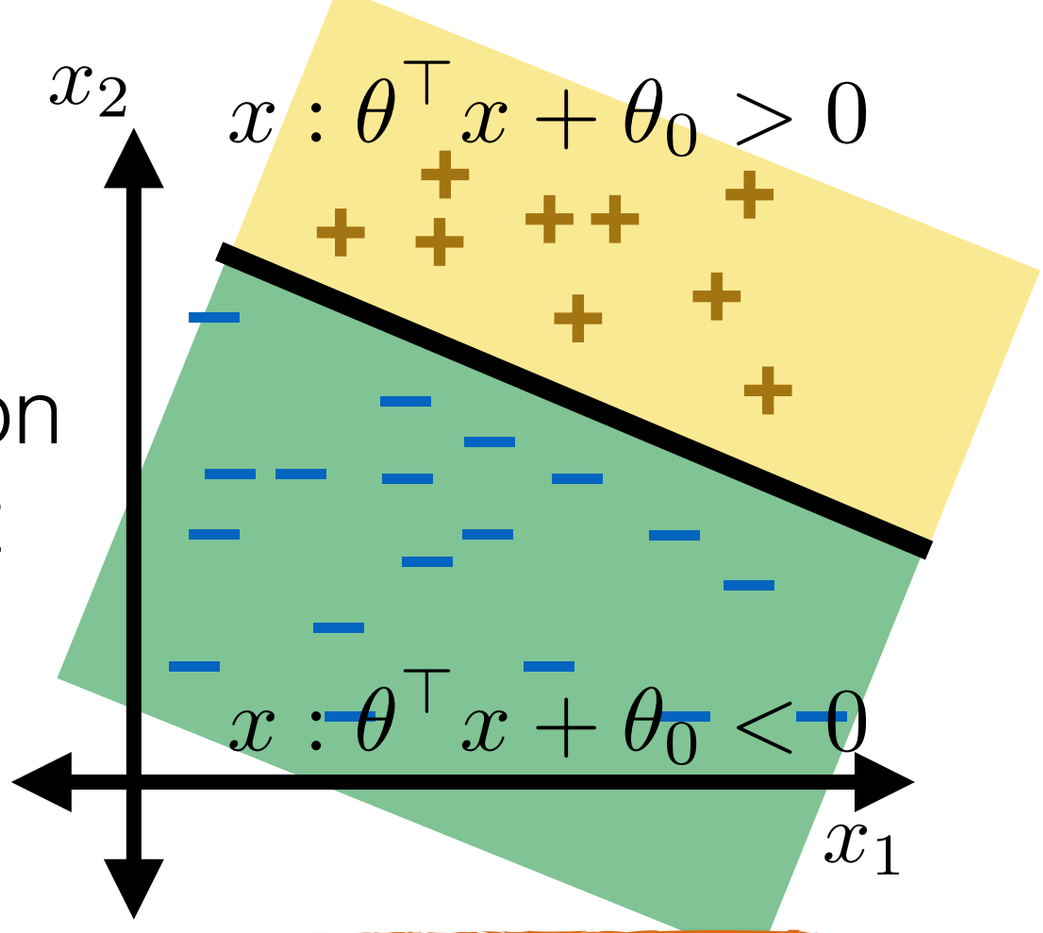
- Linear classification with default features:

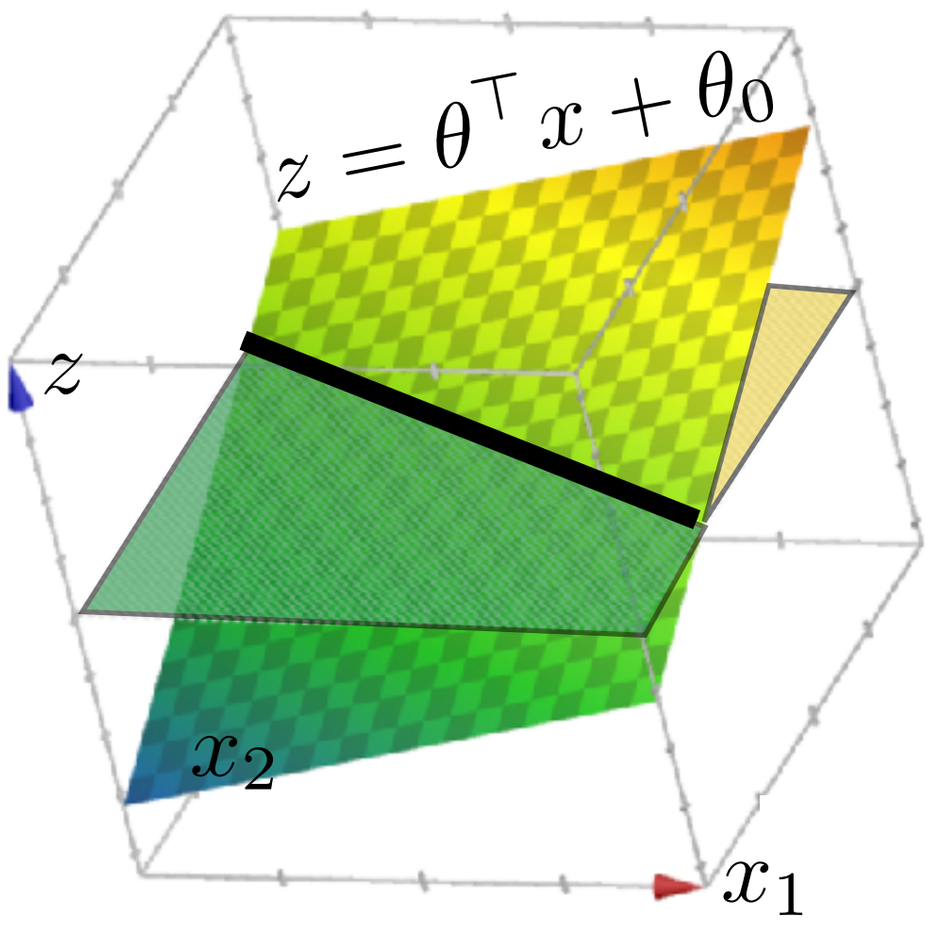


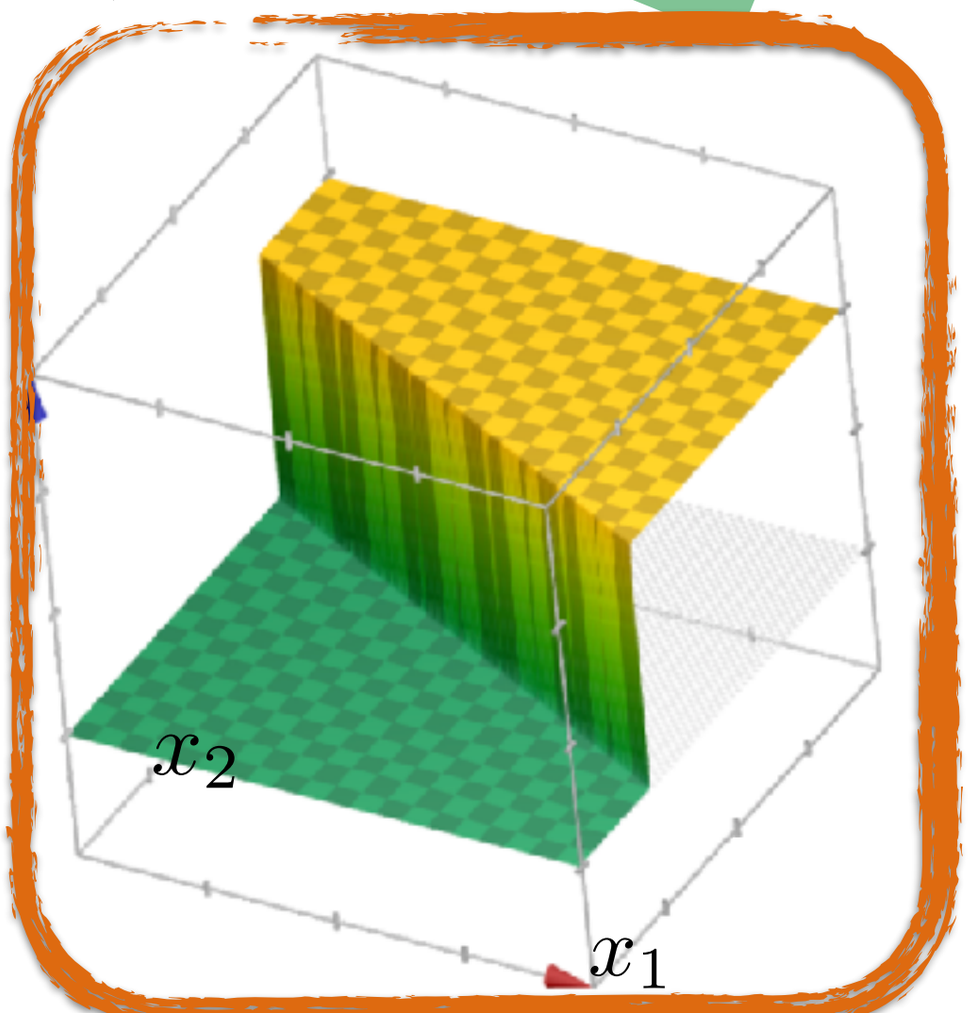



# Recall

- Linear classification with default features:

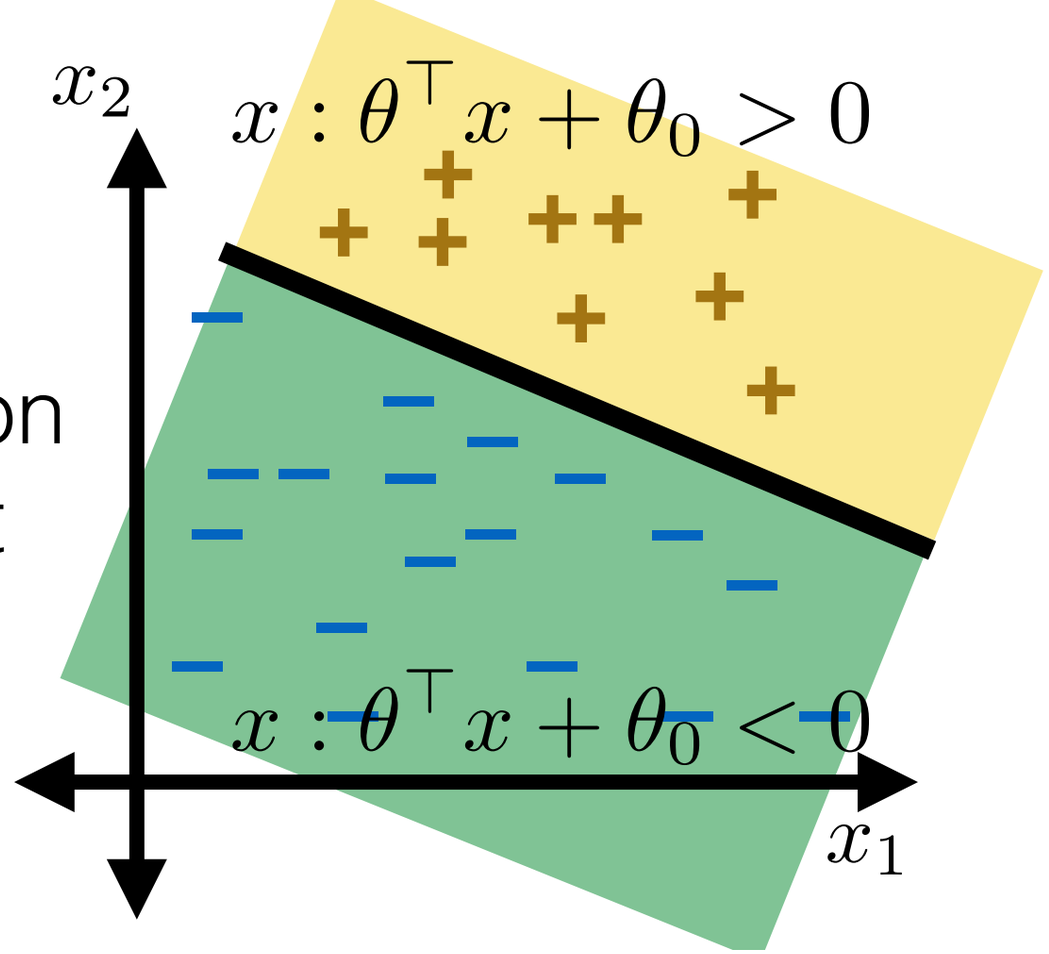


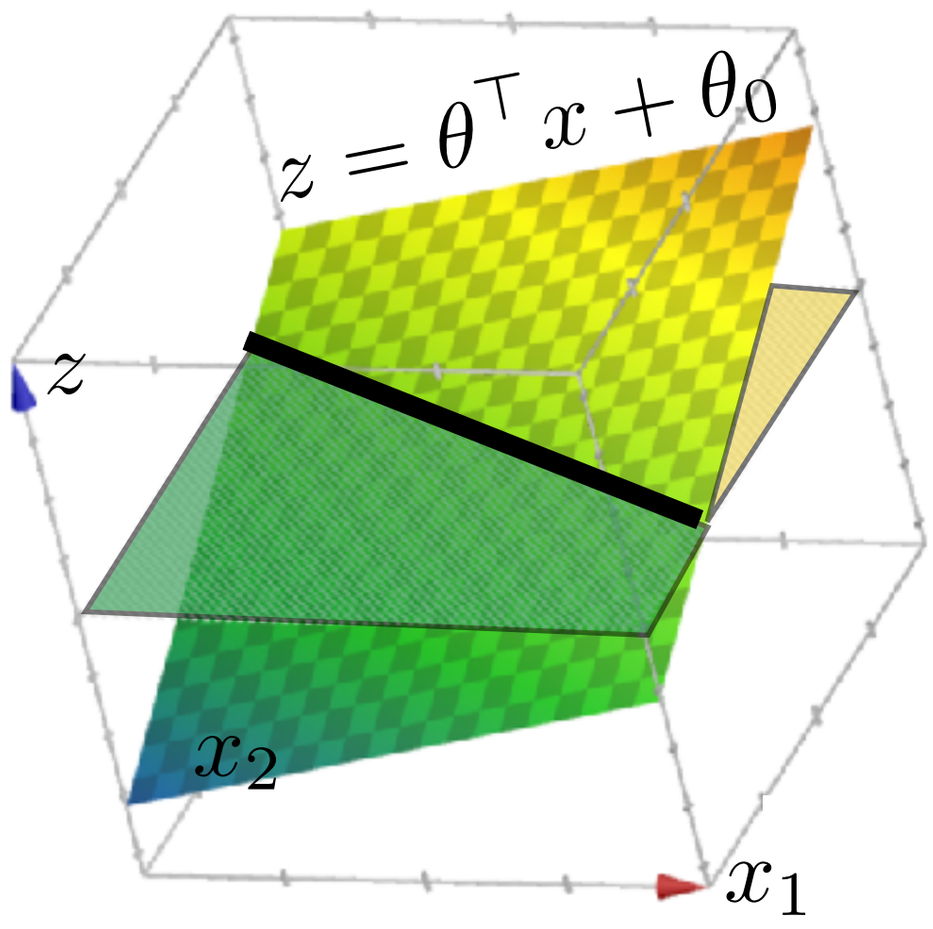



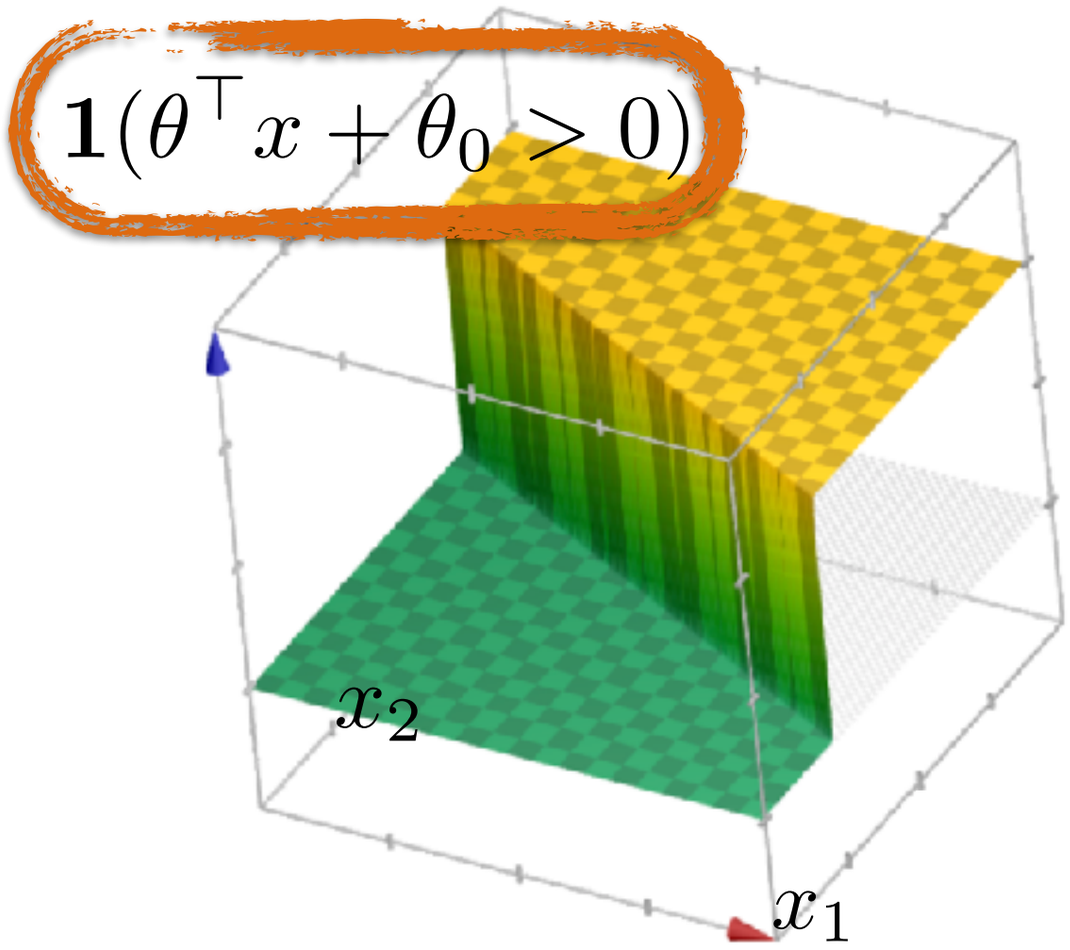



# Recall

- Linear classification with default features:

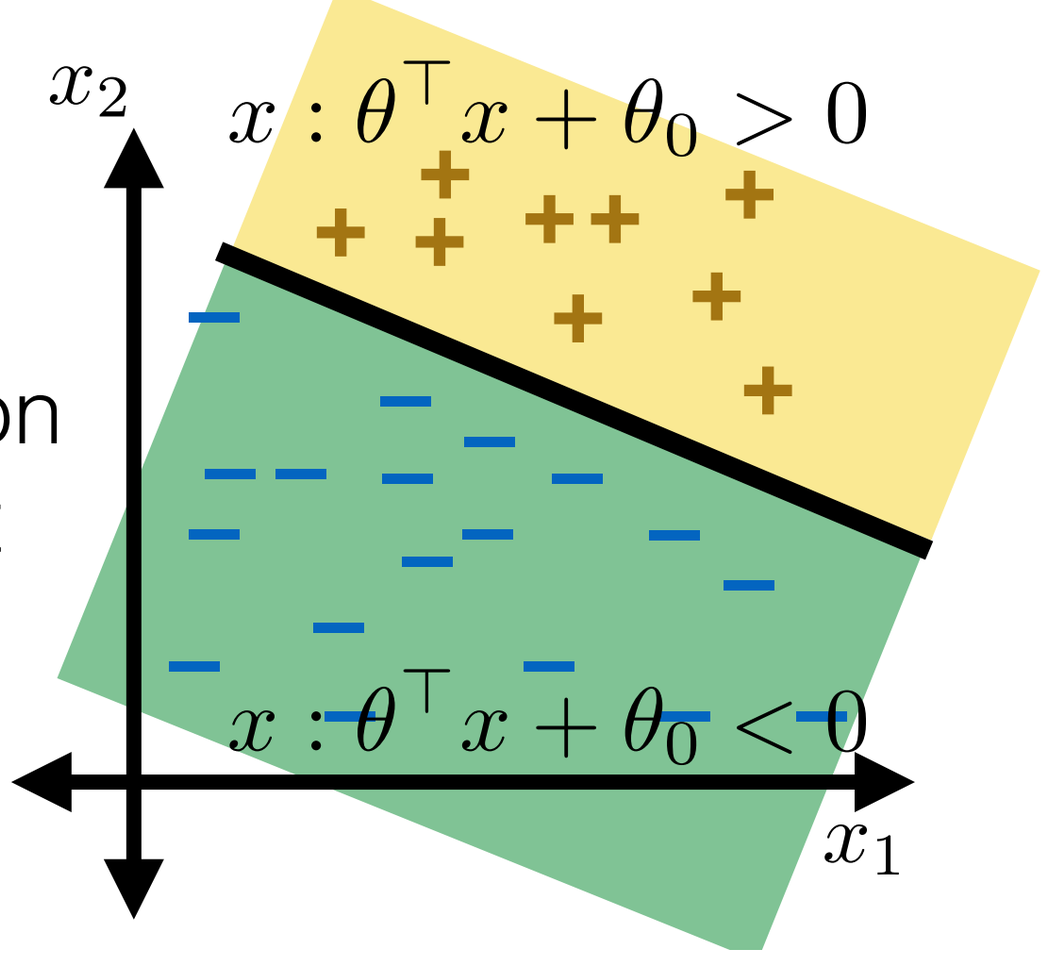


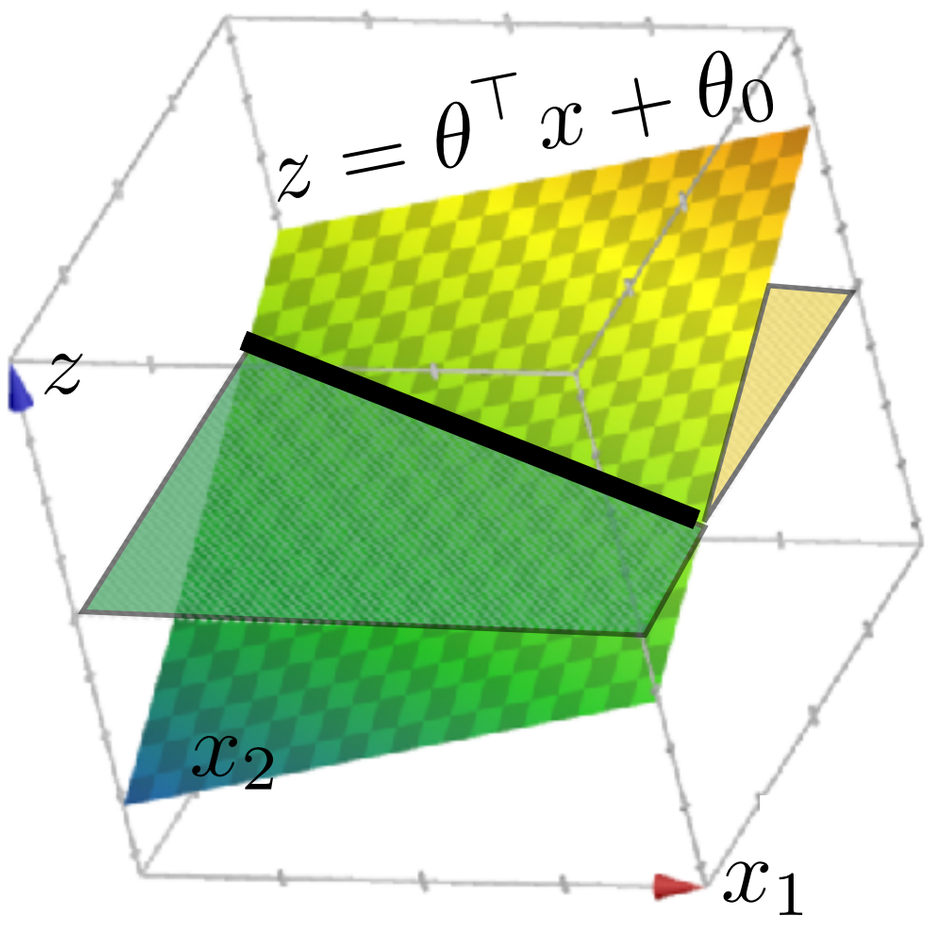



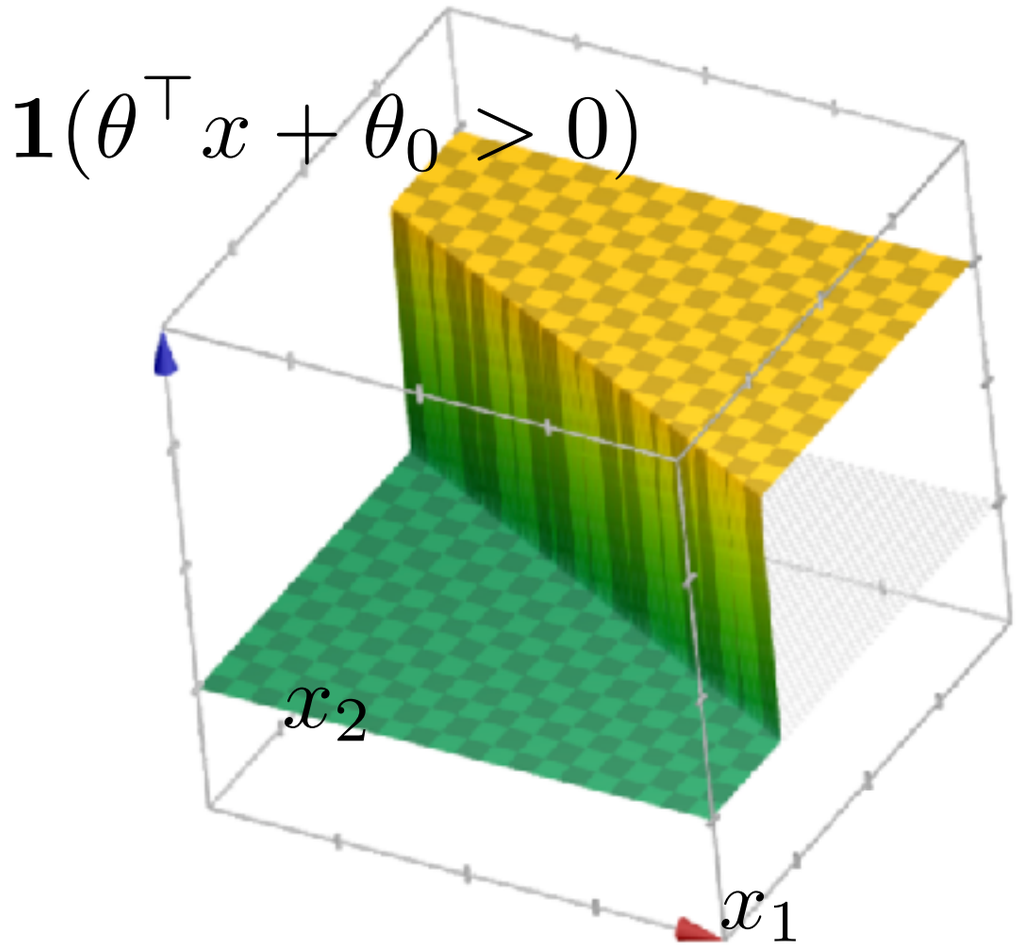



# Recall

- Linear classification with default features:



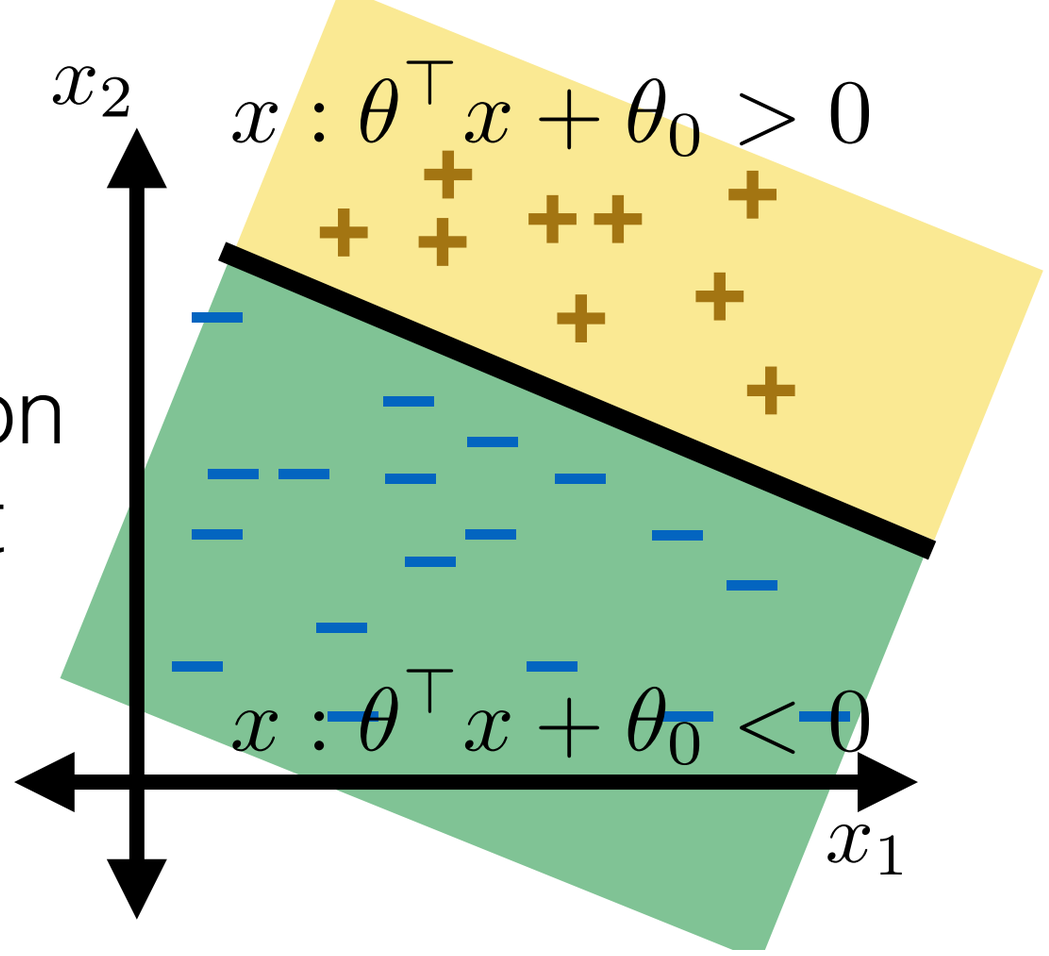



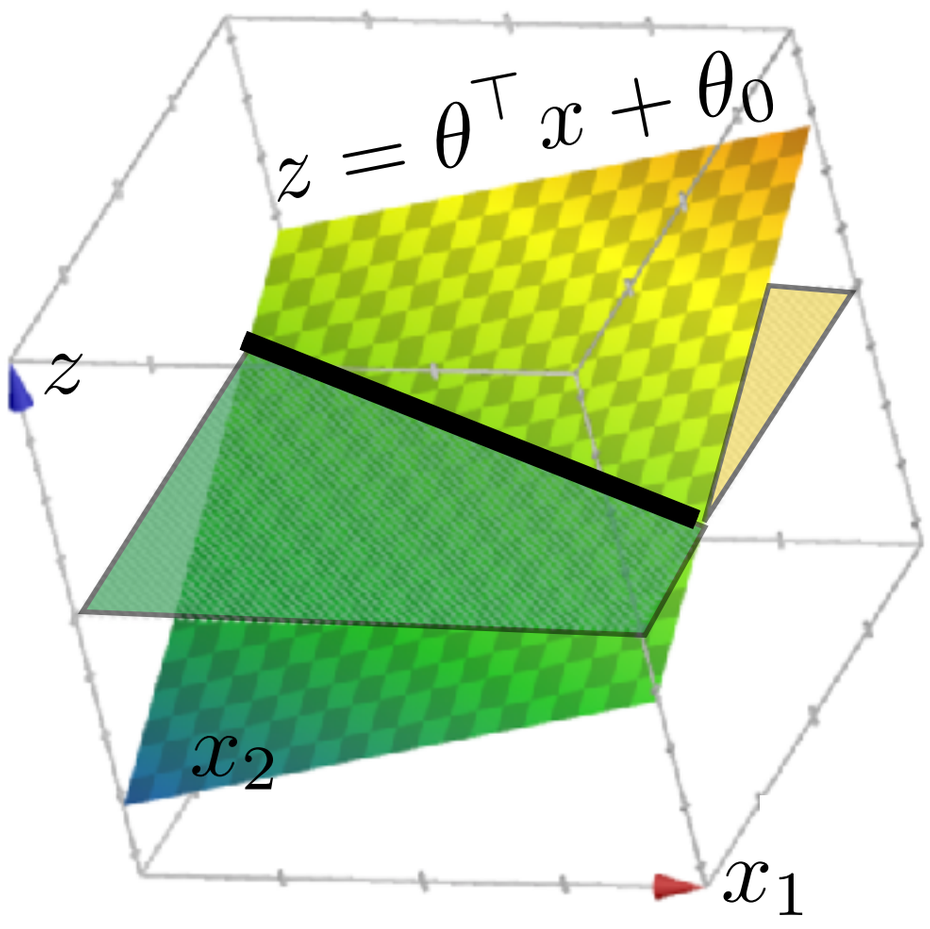



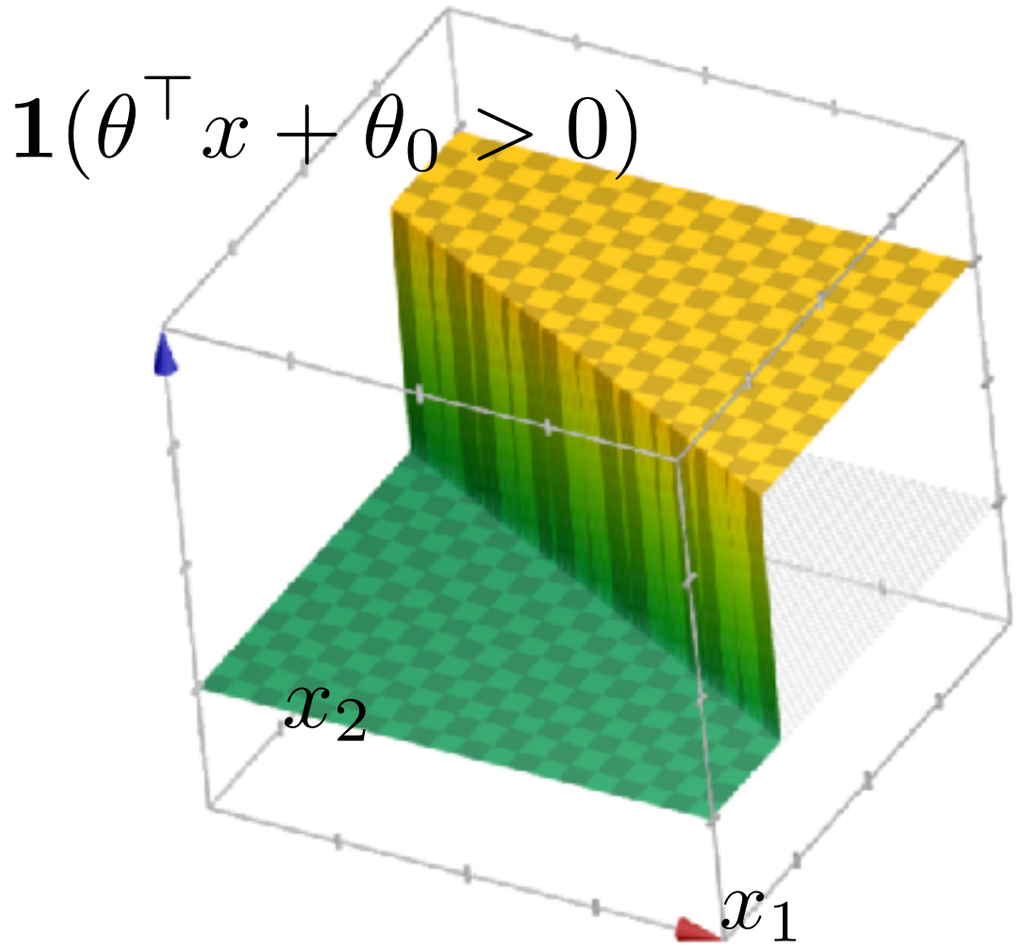

- We're used to using step functions to classify

# 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

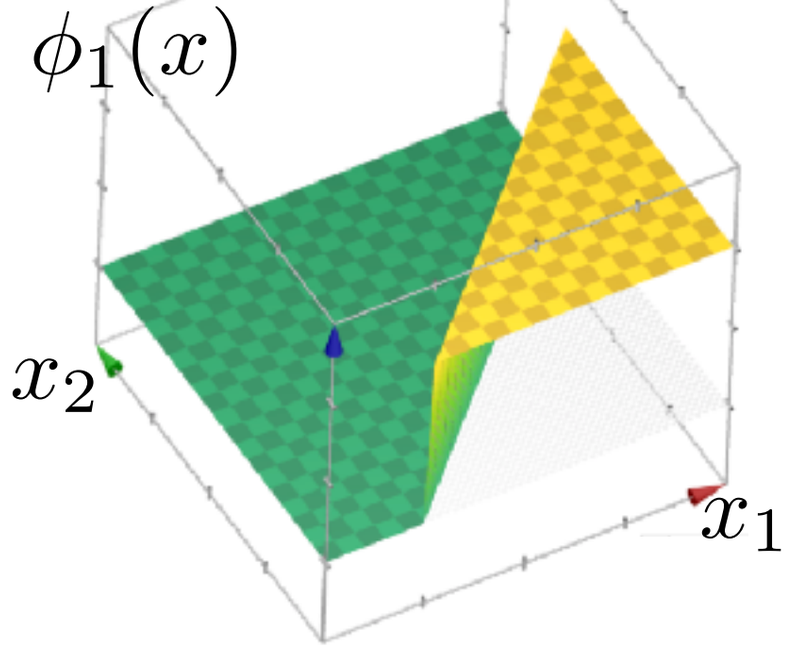
# New features: step functions!

# New features: step functions!

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

# New features: step functions!

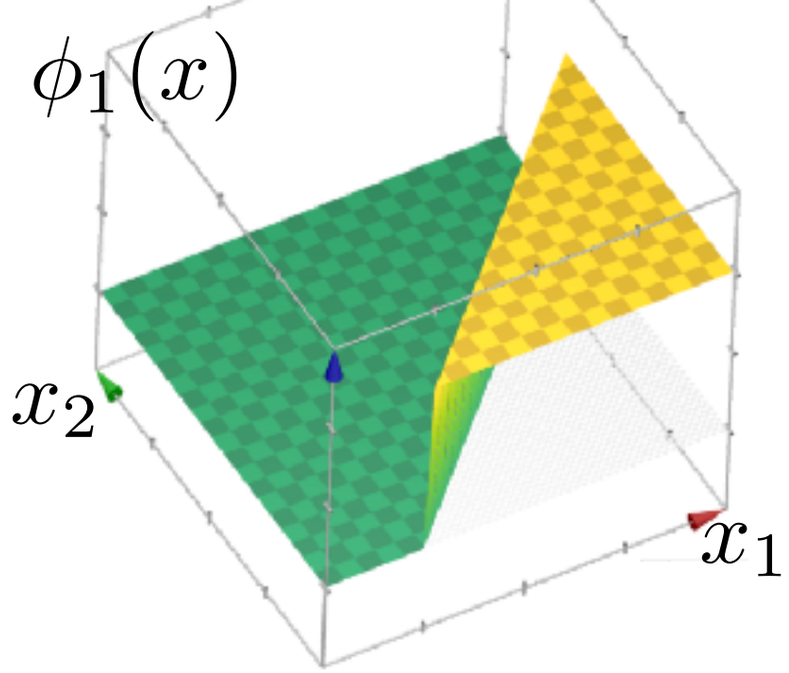
$$\phi_1(x) = \mathbf{1}\{w^\top x + w_0 \geq 0\}$$



# New features: step functions!

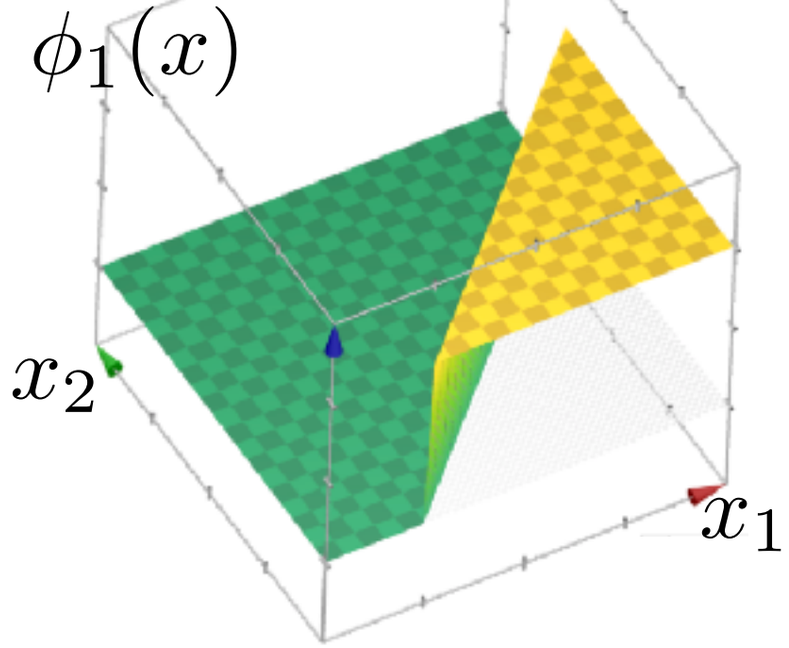
$$\phi_1(x) = \mathbf{1}\{w^\top x + w_0 \geq 0\}$$

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

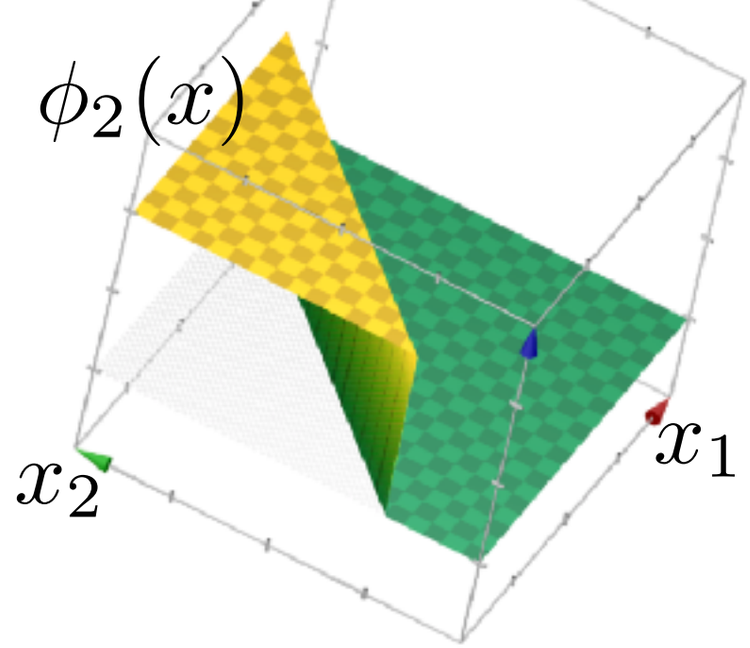


# New features: step functions!

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

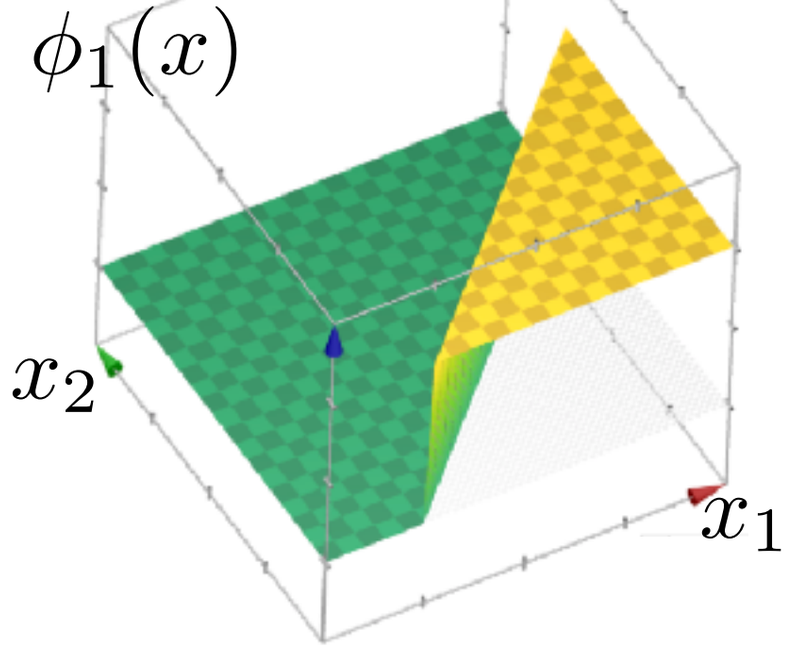



$$\phi_2(x) = \mathbf{1}\{\tilde{w}^\top x + \tilde{w}_0 \geq 0\}$$

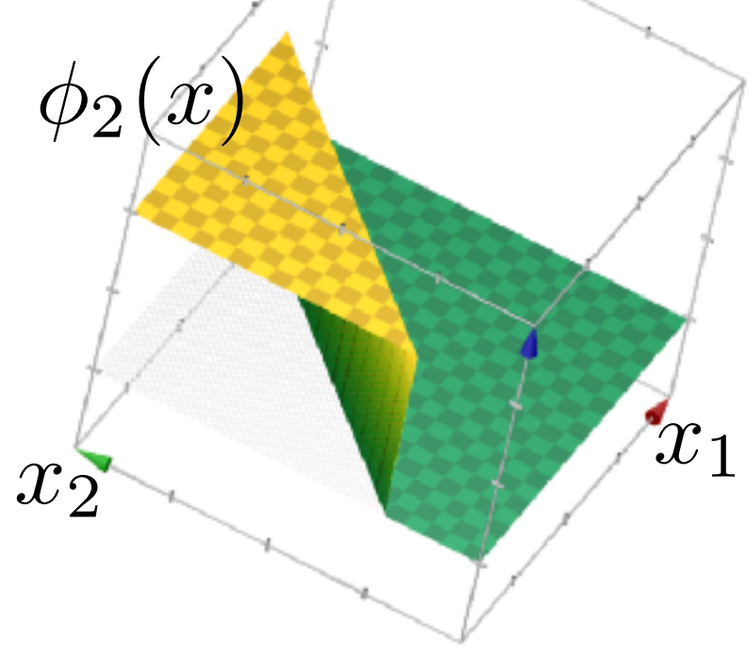


# New features: step functions!

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



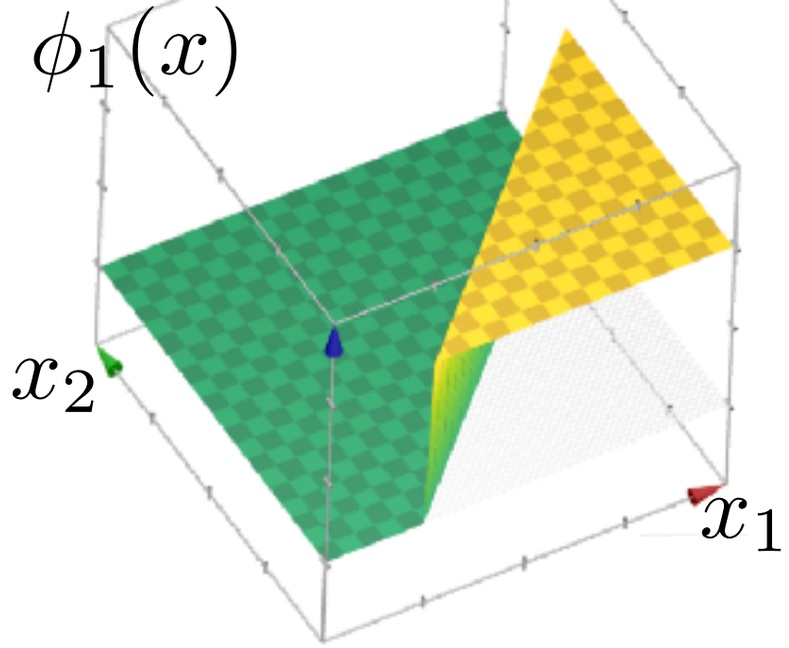

$$\phi_2(x) = \mathbf{1}\{\tilde{w}^\top x + \tilde{w}_0 \geq 0\}$$



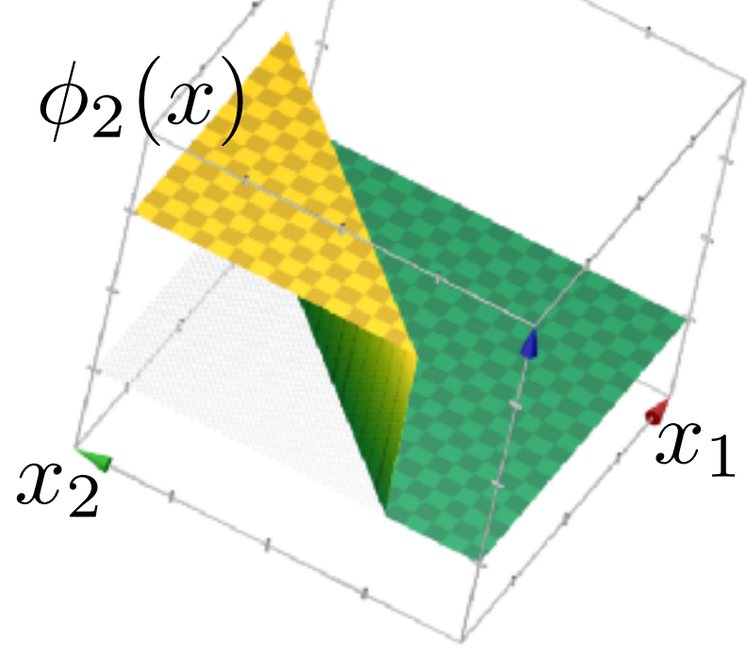

$$z = \theta^\top \phi(x) + \theta_0$$

# New features: step functions!

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



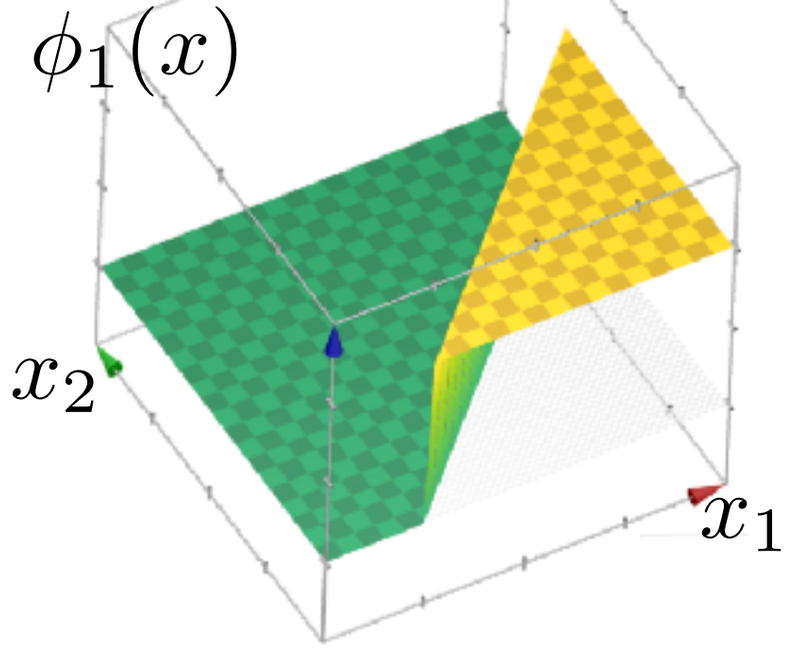

$$\phi_2(x) = \mathbf{1}\{\tilde{w}^\top x + \tilde{w}_0 \geq 0\}$$



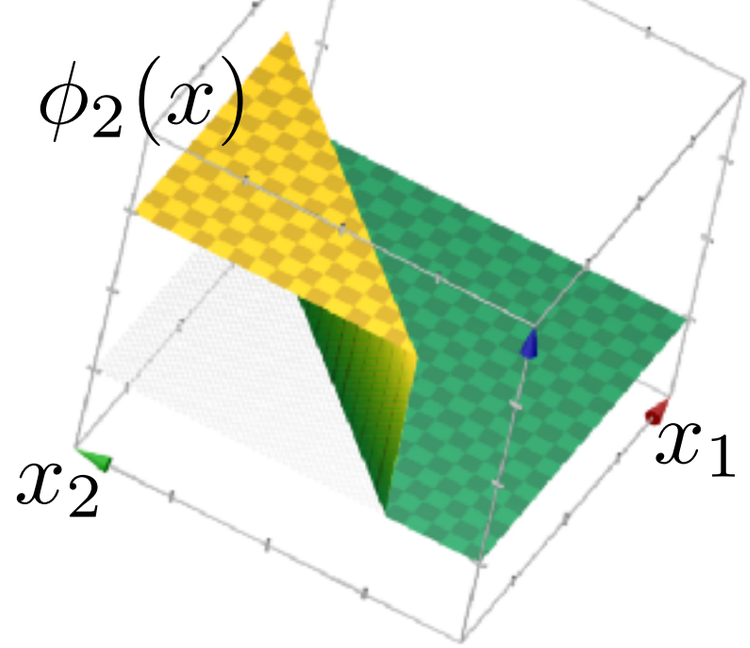

$$z = \theta^\top \phi(x) + \theta_0$$
$$= \theta_1 \phi_1(x) + \theta_2 \phi_2(x) + \theta_0$$

# New features: step functions!

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




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



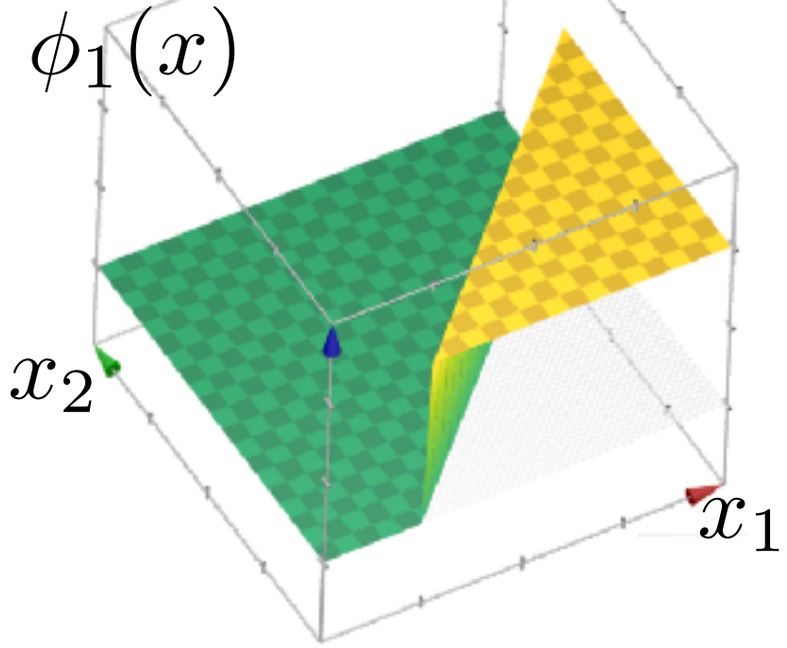

$$z = \theta^\top \phi(x) + \theta_0$$

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

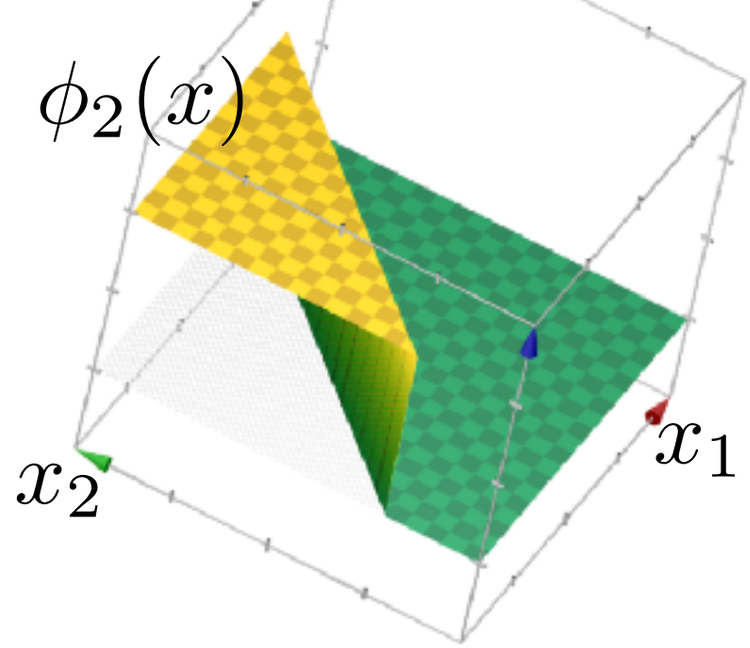
$$= 1 \cdot \phi_1(x) + 1 \cdot \phi_2(x) + (-0.5)$$

# New features: step functions!

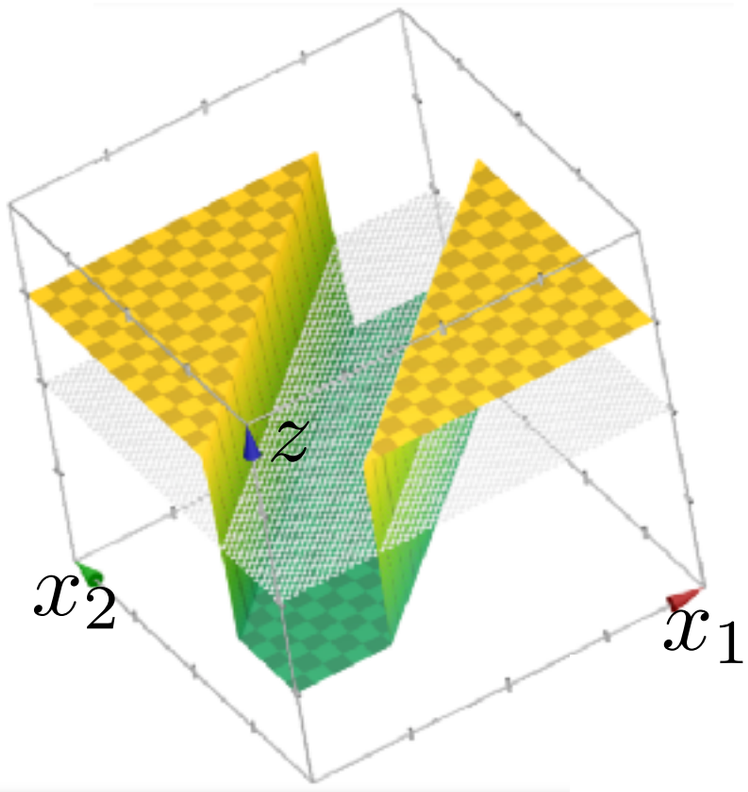
$$\phi_1(x) = \mathbf{1}\{w^\top x + w_0 \geq 0\}$$




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

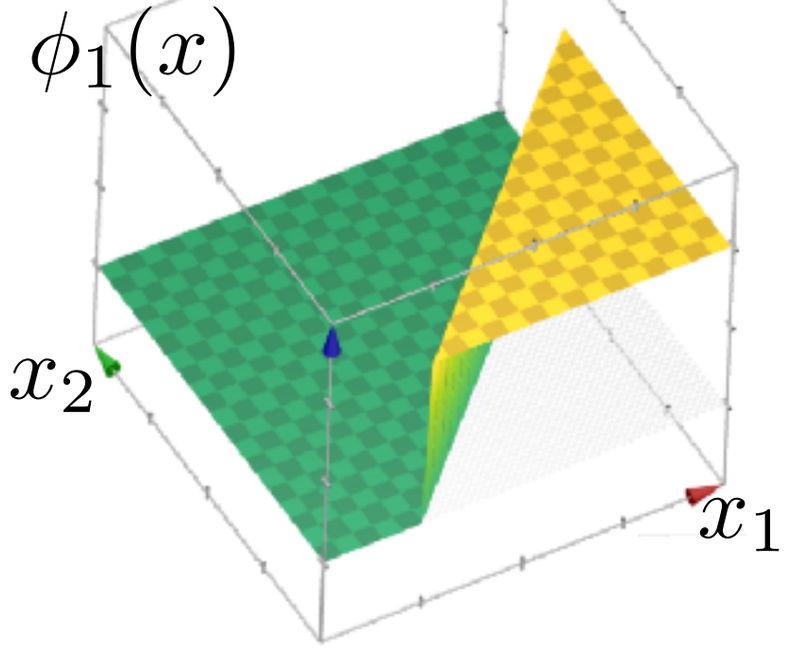



$$z = \theta^\top \phi(x) + \theta_0$$
$$= \theta_1 \phi_1(x) + \theta_2 \phi_2(x) + \theta_0$$
$$= 1 \cdot \phi_1(x) + 1 \cdot \phi_2(x) + (-0.5)$$

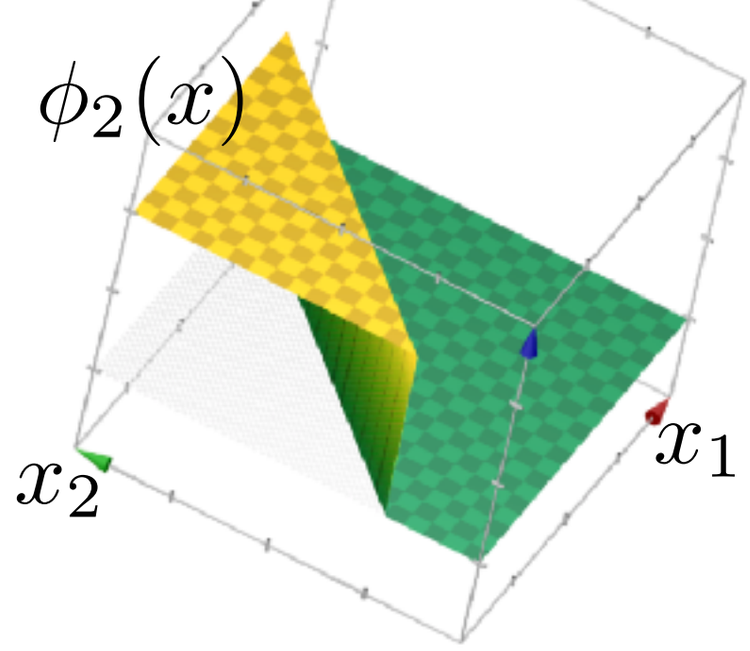


# New features: step functions!

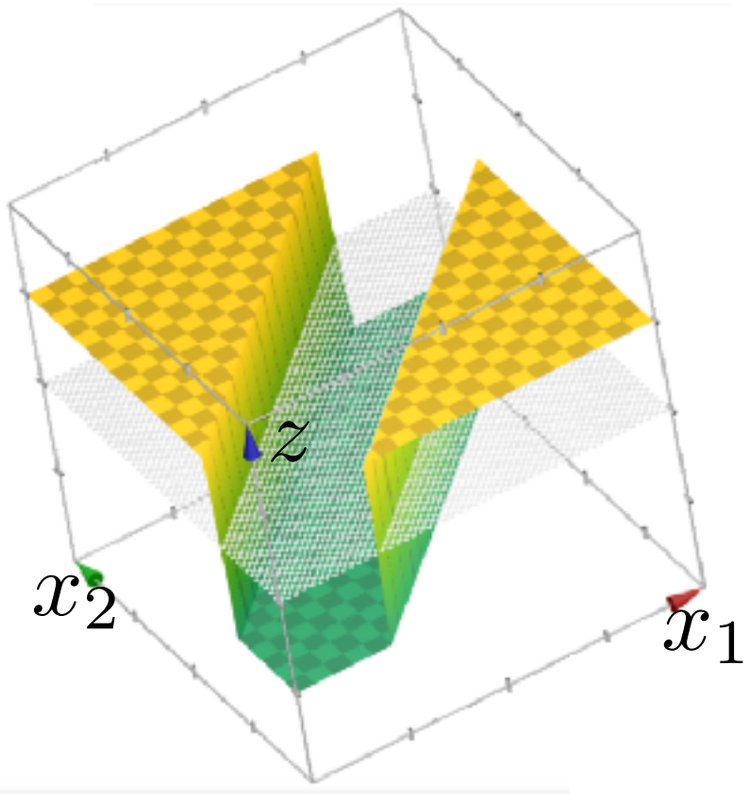
$$\phi_1(x) = \mathbf{1}\{w^\top x + w_0 \geq 0\}$$

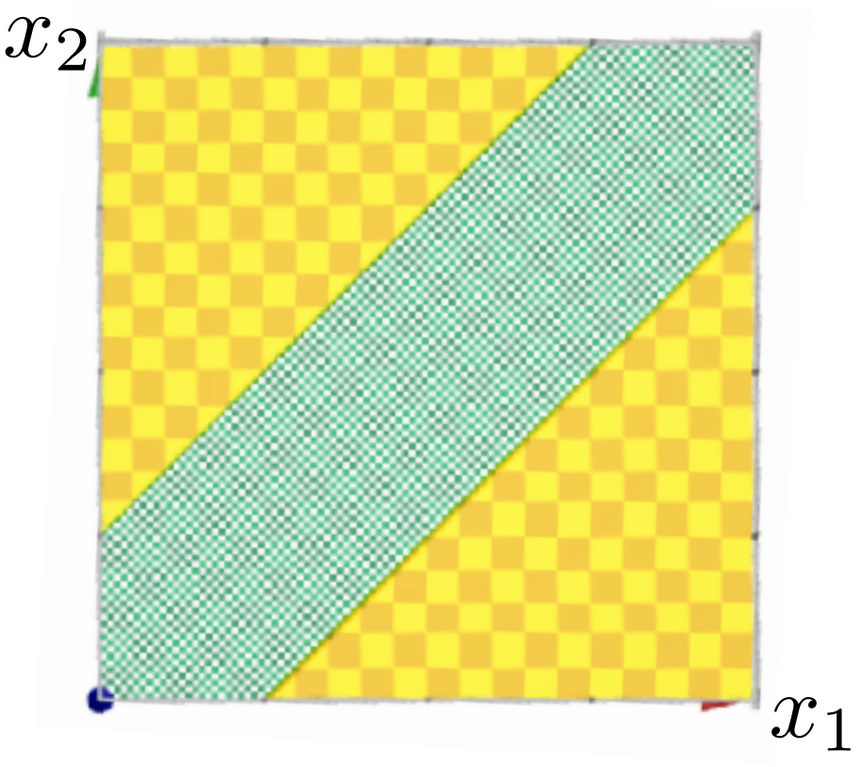



$$\phi_2(x) = \mathbf{1}\{\tilde{w}^\top x + \tilde{w}_0 \geq 0\}$$



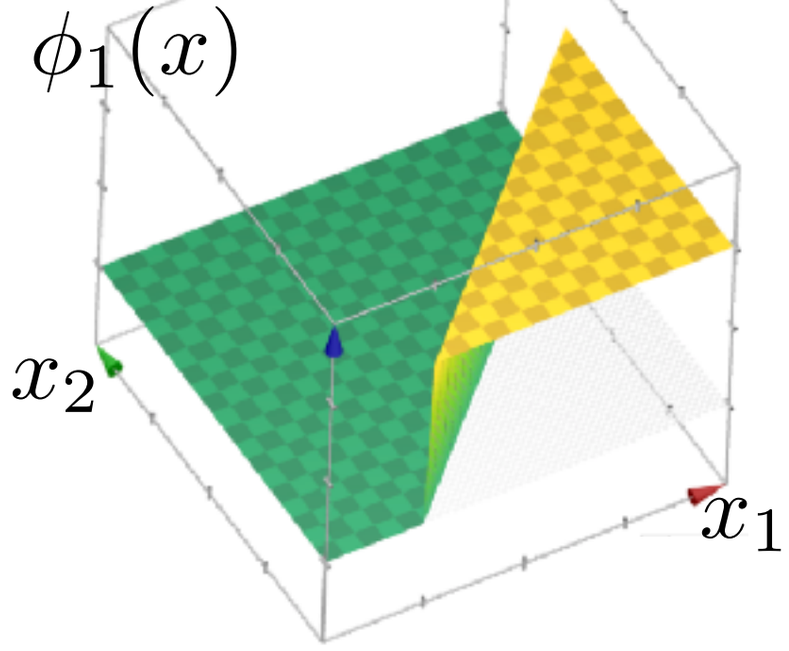

$$z = \theta^\top \phi(x) + \theta_0$$
$$= \theta_1 \phi_1(x) + \theta_2 \phi_2(x) + \theta_0$$
$$= 1 \cdot \phi_1(x) + 1 \cdot \phi_2(x) + (-0.5)$$



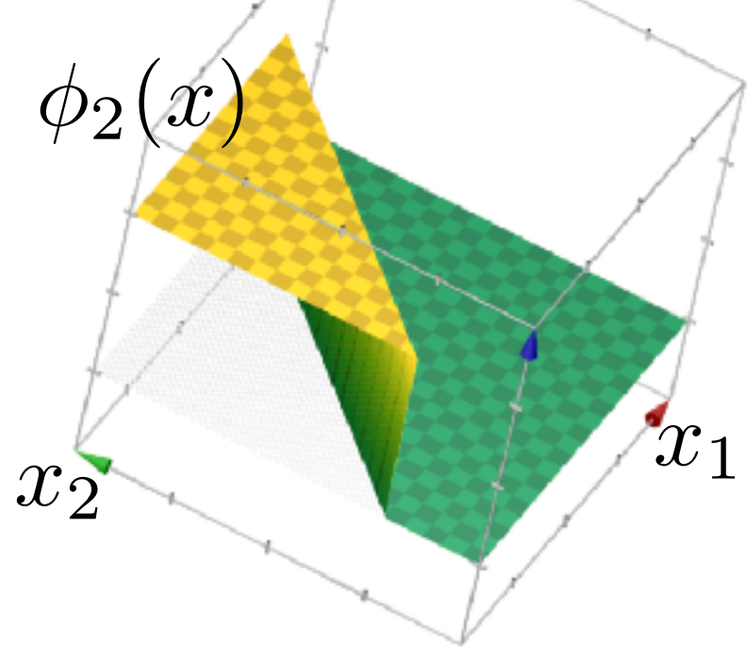



# New features: step functions!

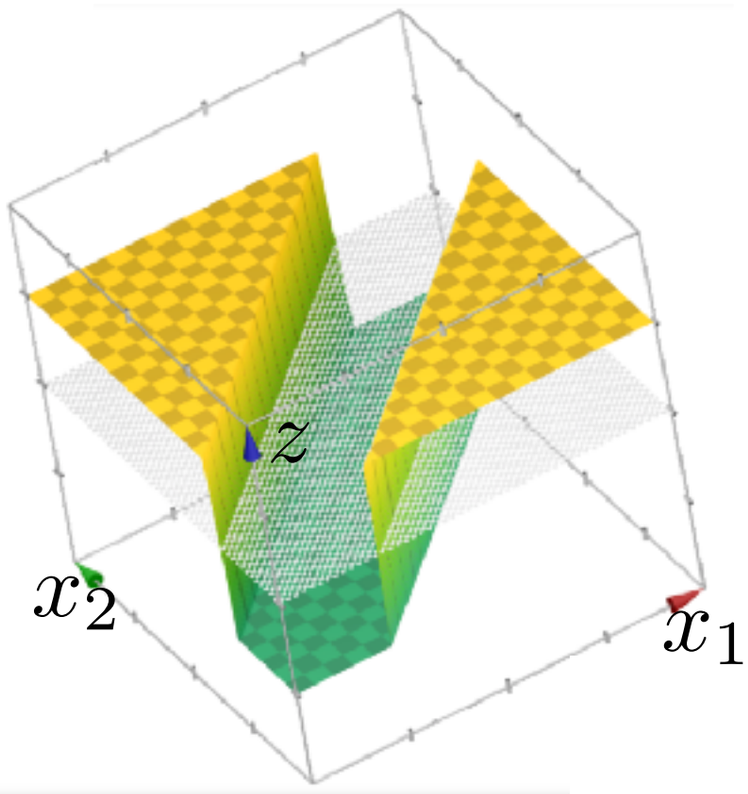
$$\phi_1(x) = \mathbf{1}\{w^\top x + w_0 \geq 0\}$$

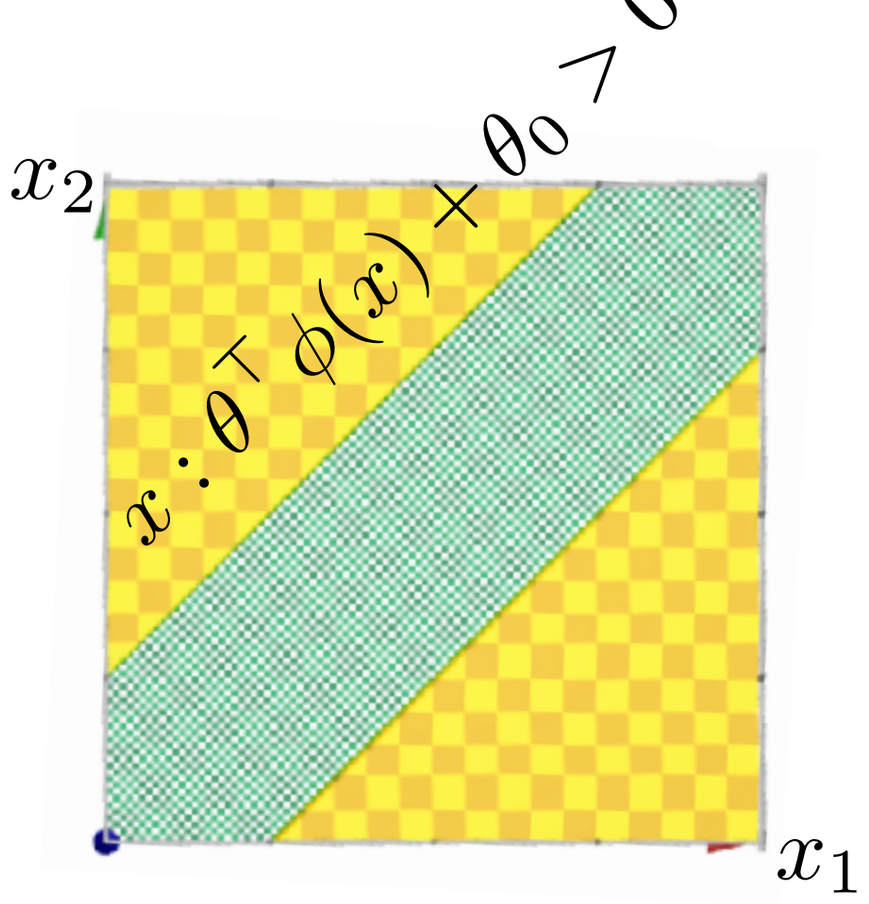



$$\phi_2(x) = \mathbf{1}\{\tilde{w}^\top x + \tilde{w}_0 \geq 0\}$$



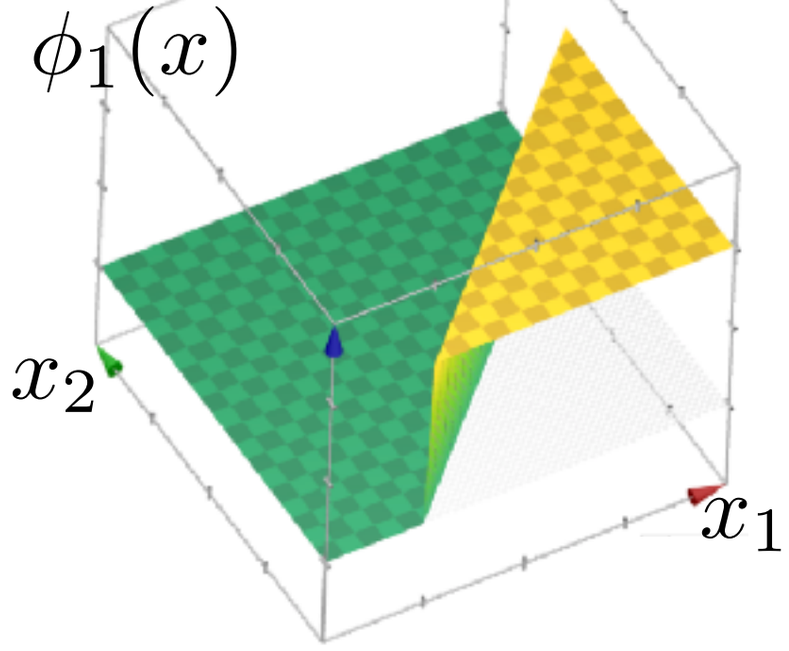

$$z = \theta^\top \phi(x) + \theta_0$$
$$= \theta_1 \phi_1(x) + \theta_2 \phi_2(x) + \theta_0$$
$$= 1 \cdot \phi_1(x) + 1 \cdot \phi_2(x) + (-0.5)$$



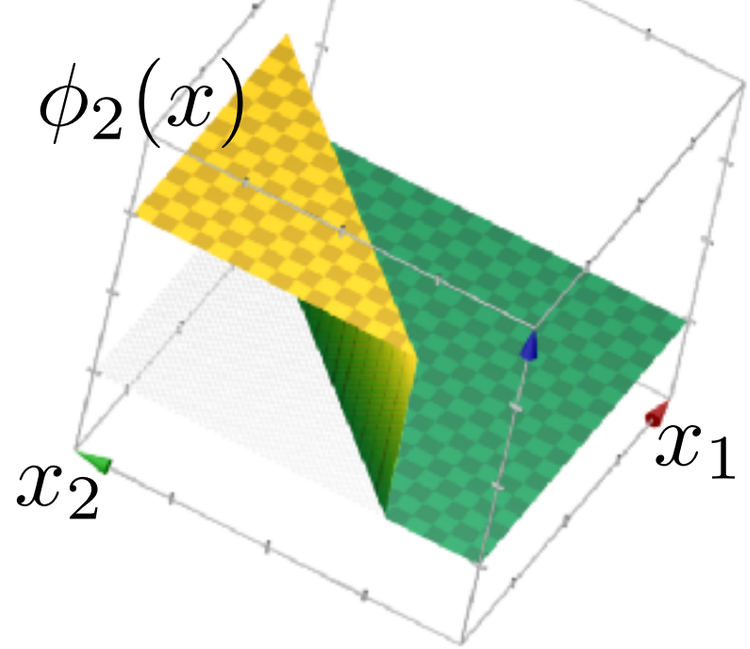



# New features: step functions!

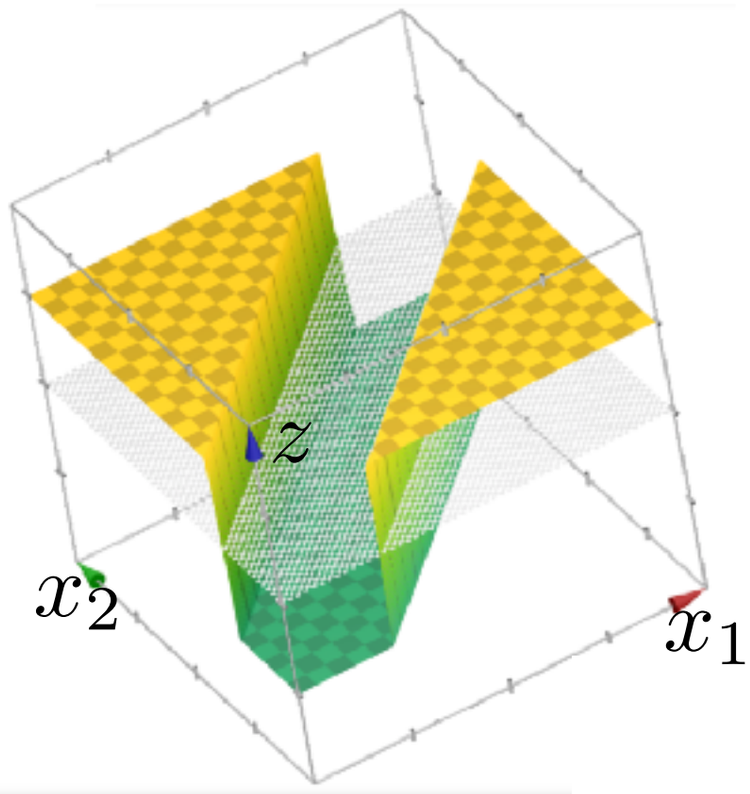
$$\phi_1(x) = \mathbf{1}\{w^\top x + w_0 \geq 0\}$$

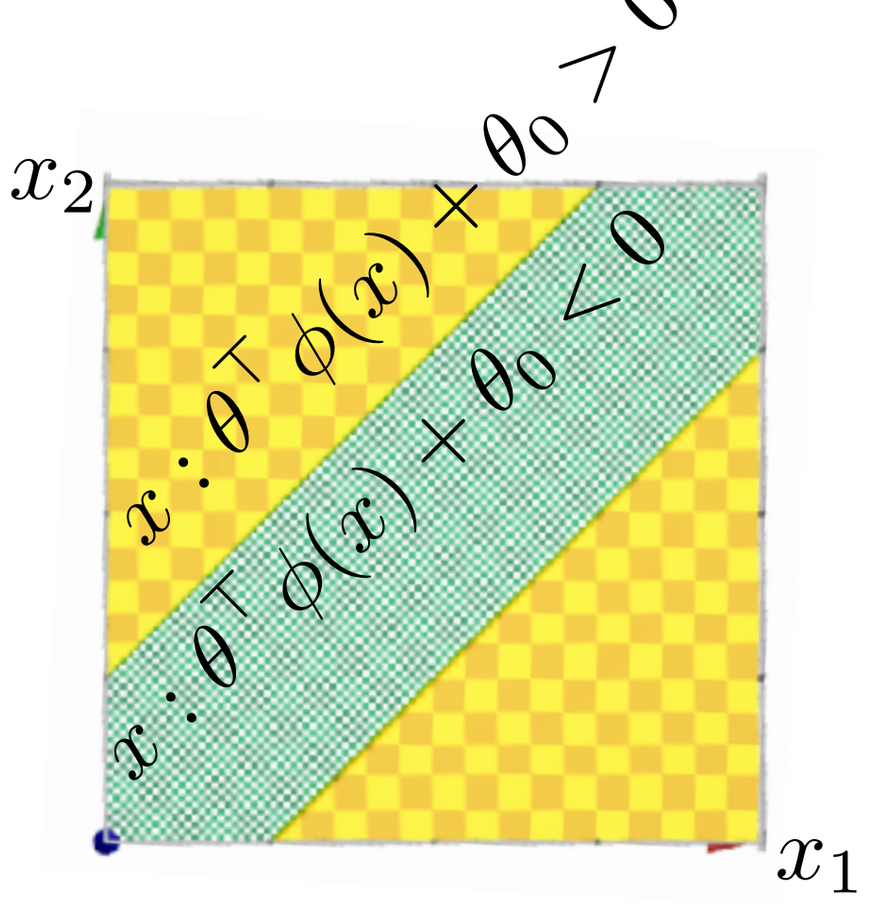



$$\phi_2(x) = \mathbf{1}\{\tilde{w}^\top x + \tilde{w}_0 \geq 0\}$$



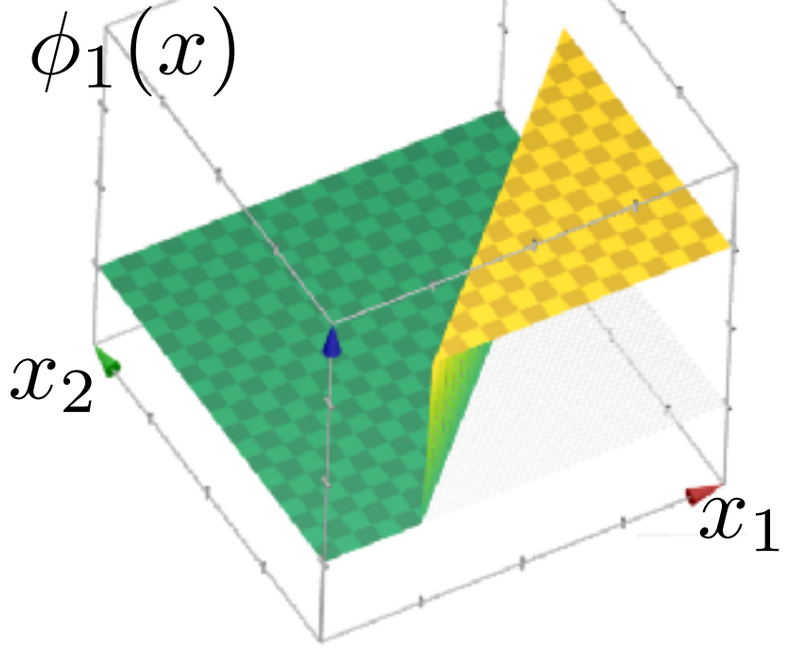

$$z = \theta^\top \phi(x) + \theta_0$$
$$= \theta_1 \phi_1(x) + \theta_2 \phi_2(x) + \theta_0$$
$$= 1 \cdot \phi_1(x) + 1 \cdot \phi_2(x) + (-0.5)$$



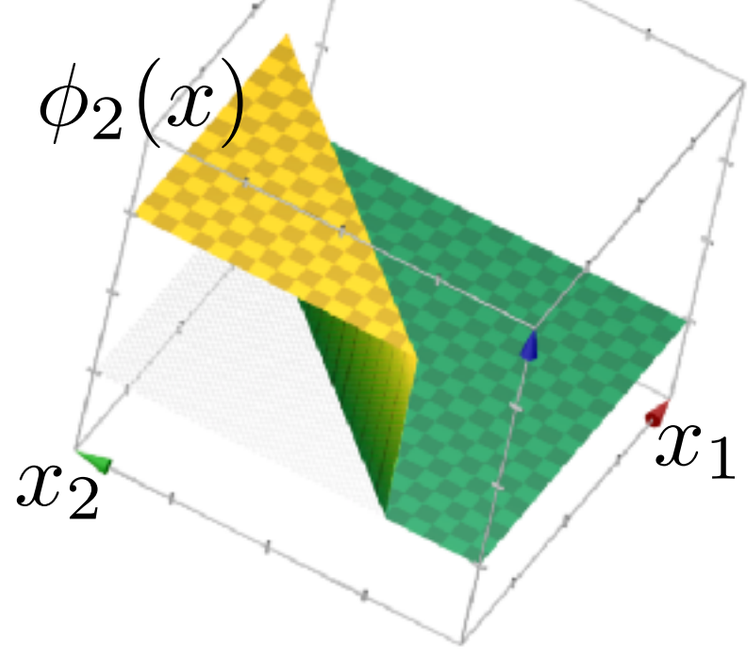



# New features: step functions!

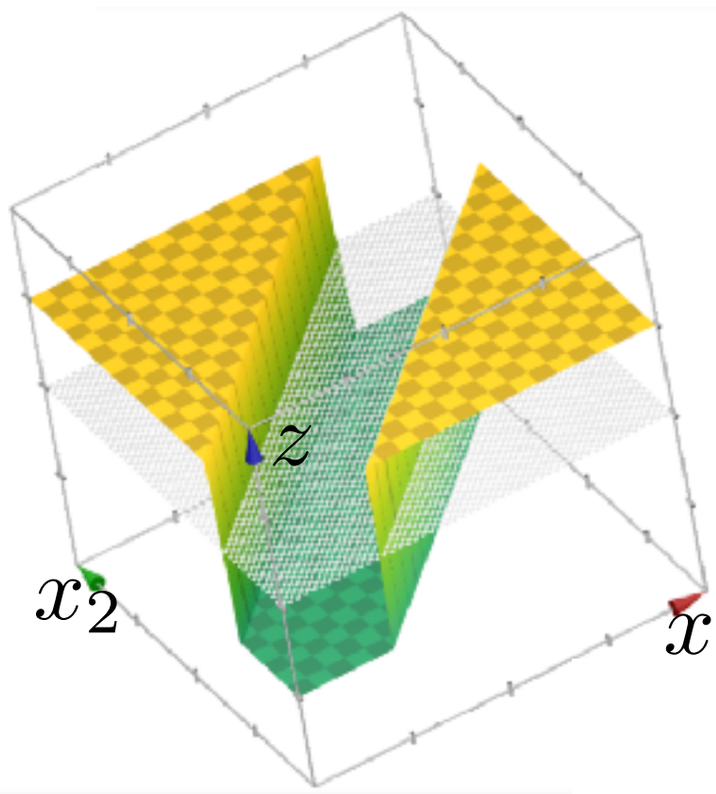
$$\phi_1(x) = \mathbf{1}\{w^\top x + w_0 \geq 0\}$$



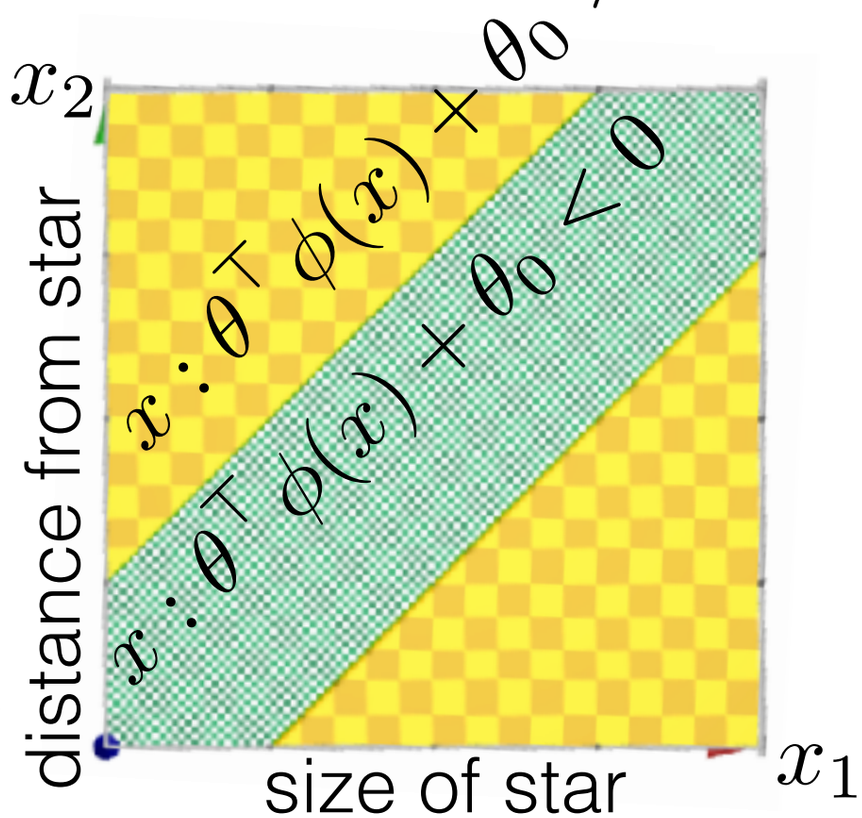

$$\phi_2(x) = \mathbf{1}\{\tilde{w}^\top x + \tilde{w}_0 \geq 0\}$$




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

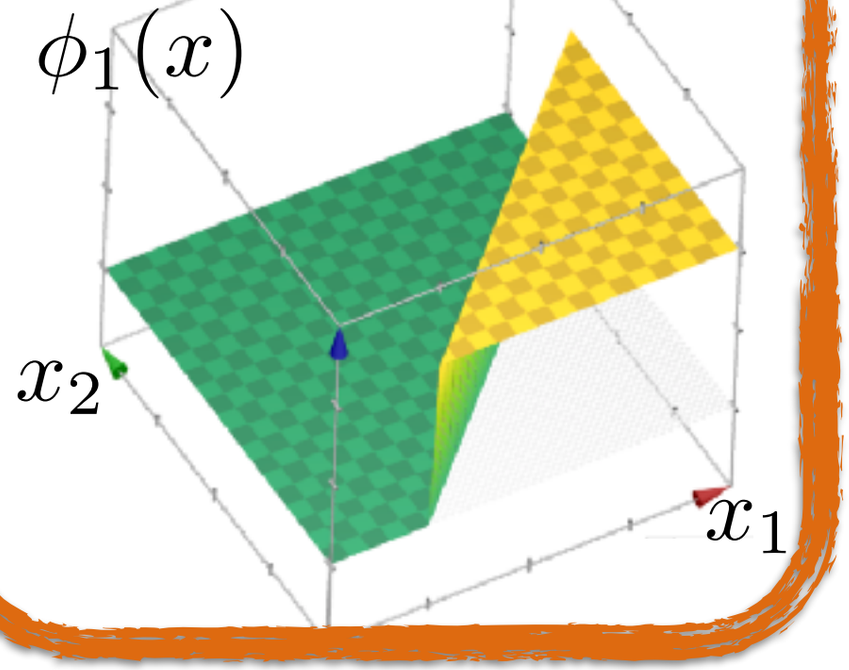



Is an exoplanet outside the habitable zone?

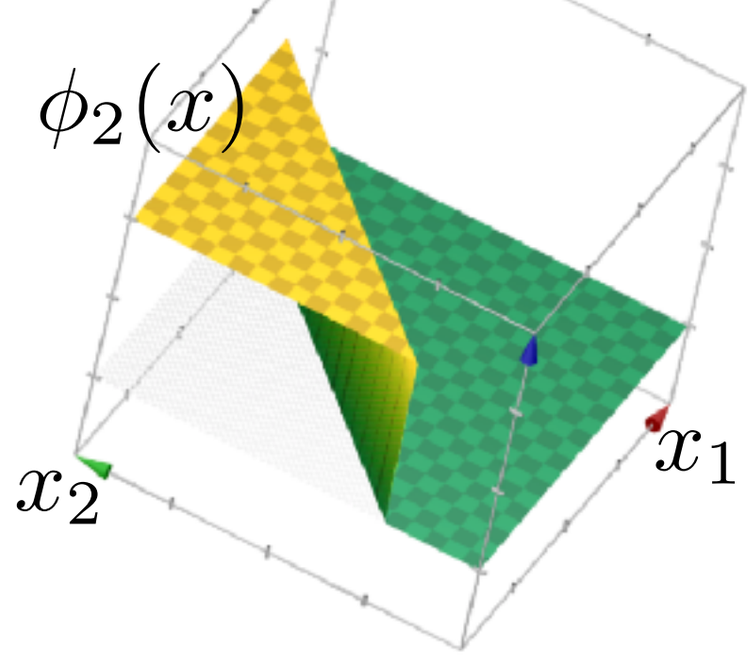


# New features: step functions!

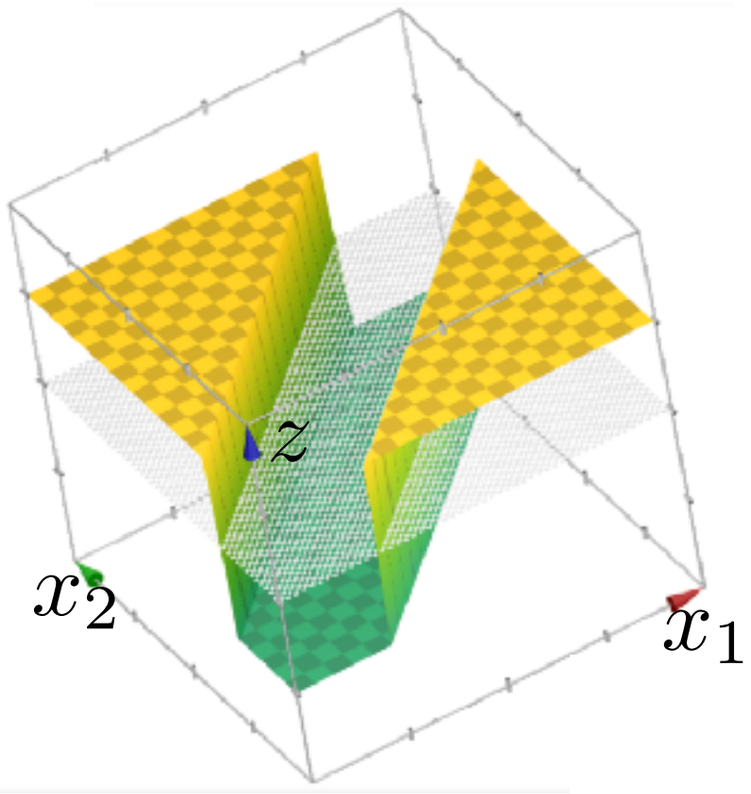
$$\phi_1(x) = \mathbf{1}\{w^\top x + w_0 \geq 0\}$$

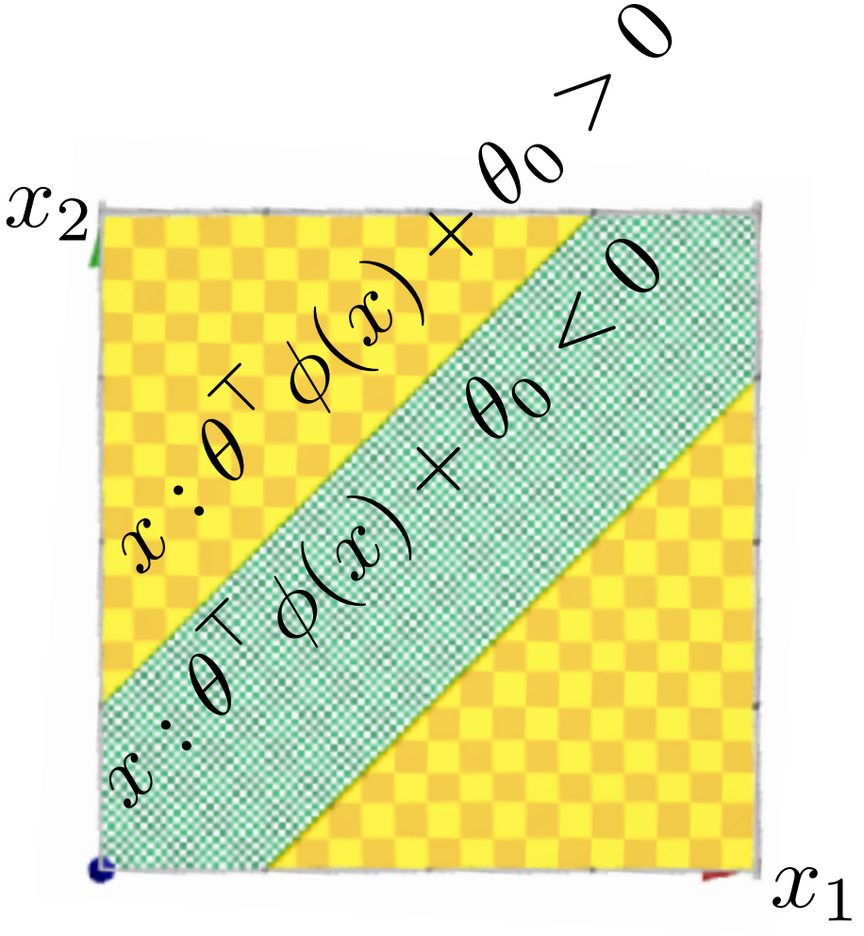



$$\phi_2(x) = \mathbf{1}\{\tilde{w}^\top x + \tilde{w}_0 \geq 0\}$$



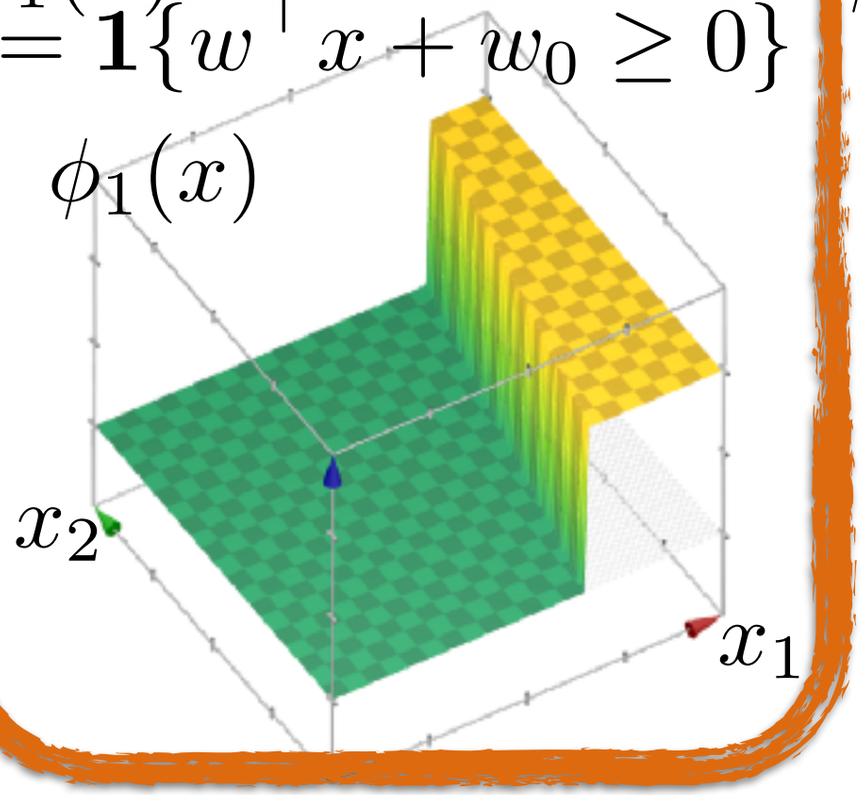

$$z = \theta^\top \phi(x) + \theta_0$$
$$= \theta_1 \phi_1(x) + \theta_2 \phi_2(x) + \theta_0$$
$$= 1 \cdot \phi_1(x) + 1 \cdot \phi_2(x) + (-0.5)$$



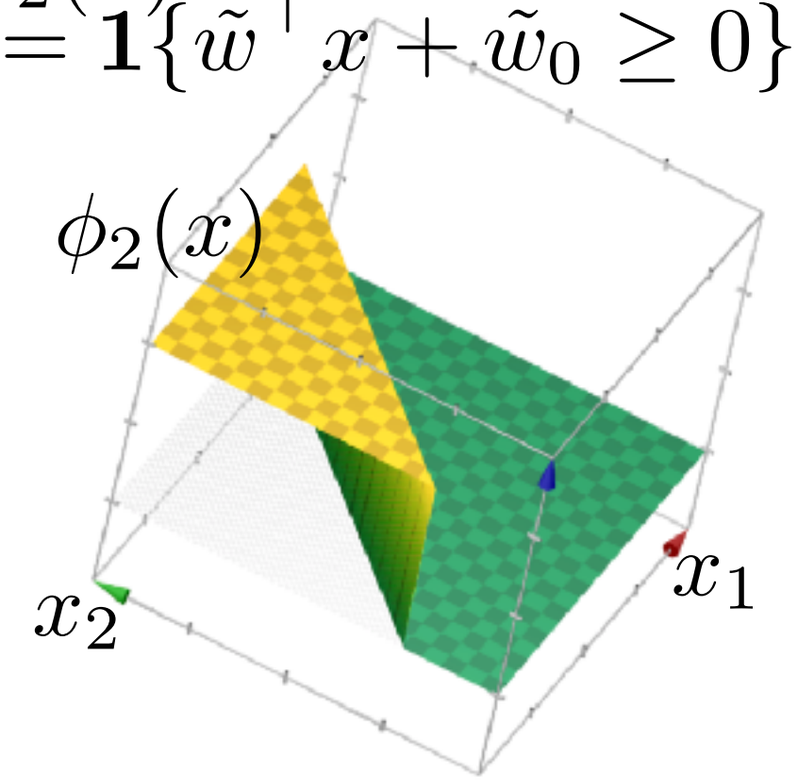



# New features: step functions!

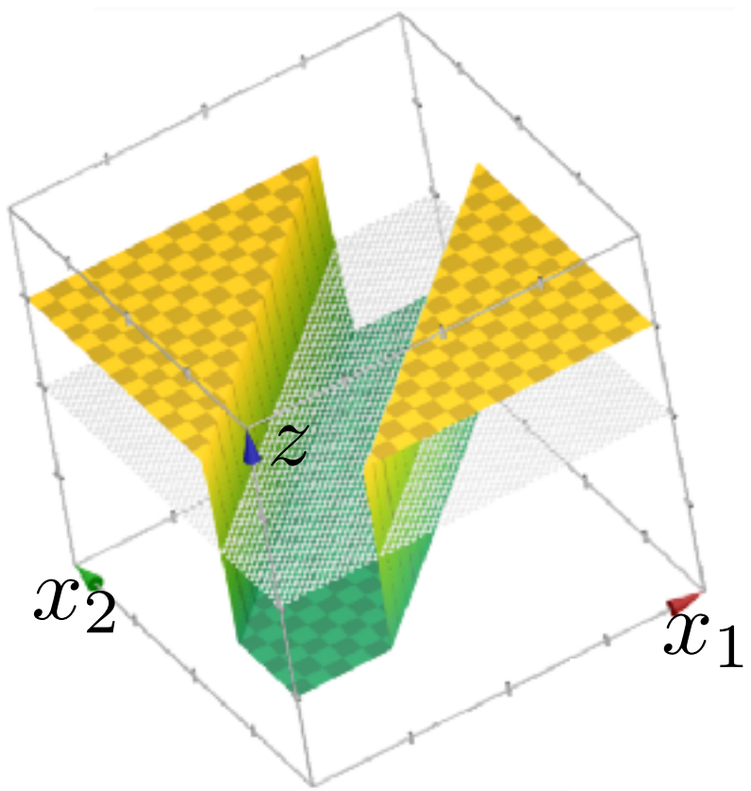
$\phi_1(x)$
$= \mathbf{1}\{w^\top x + w_0 \geq 0\}$

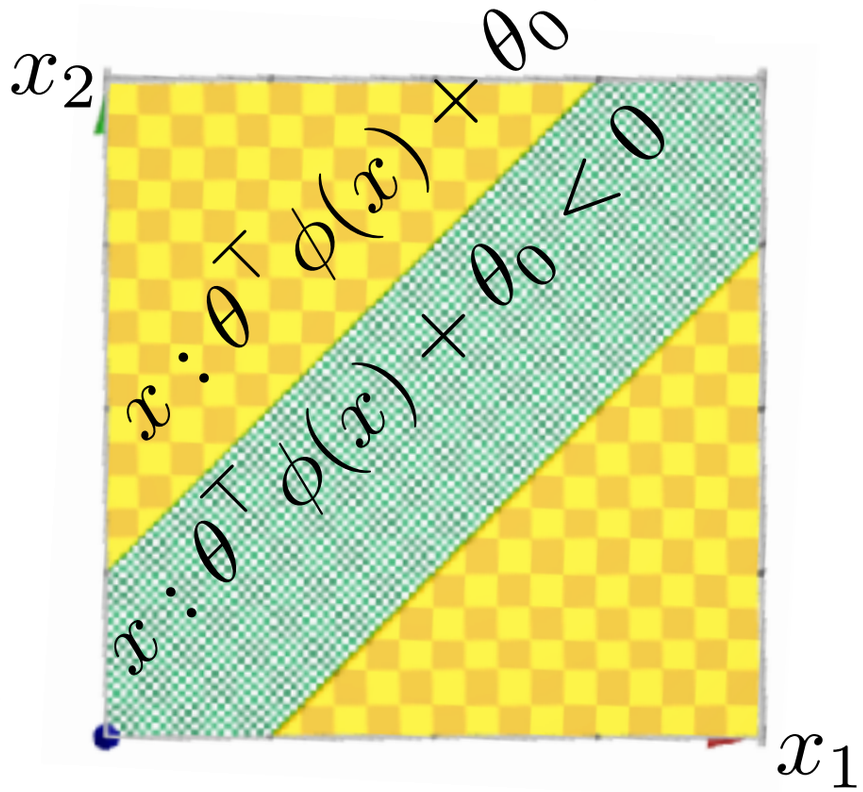



$\phi_2(x)$
$= \mathbf{1}\{\tilde{w}^\top x + \tilde{w}_0 \geq 0\}$



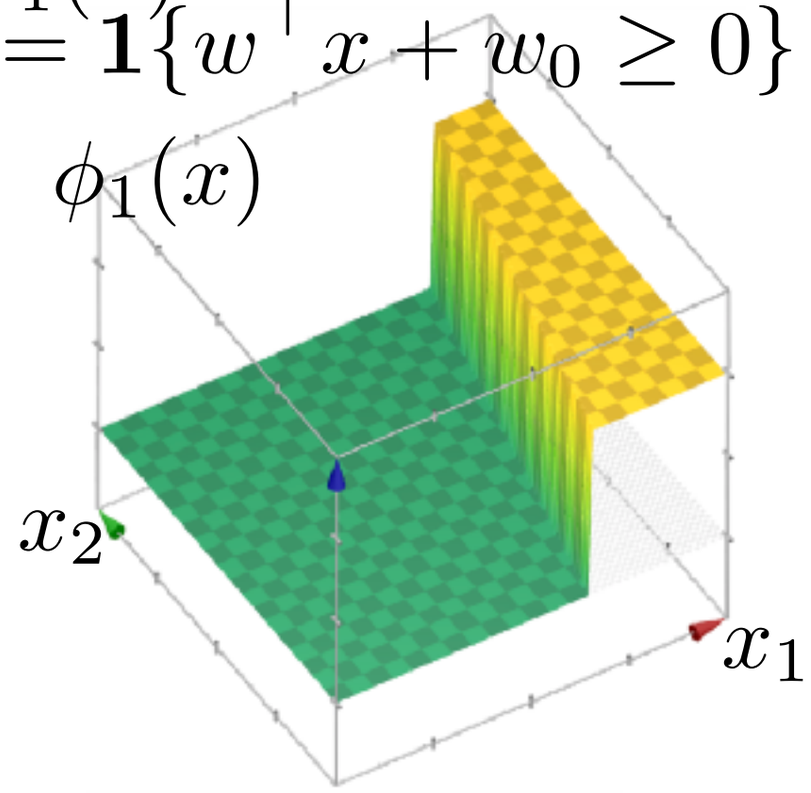

$$z = \theta^\top \phi(x) + \theta_0$$
$$= \theta_1 \phi_1(x) + \theta_2 \phi_2(x) + \theta_0$$
$$= 1 \cdot \phi_1(x) + 1 \cdot \phi_2(x) + (-0.5)$$



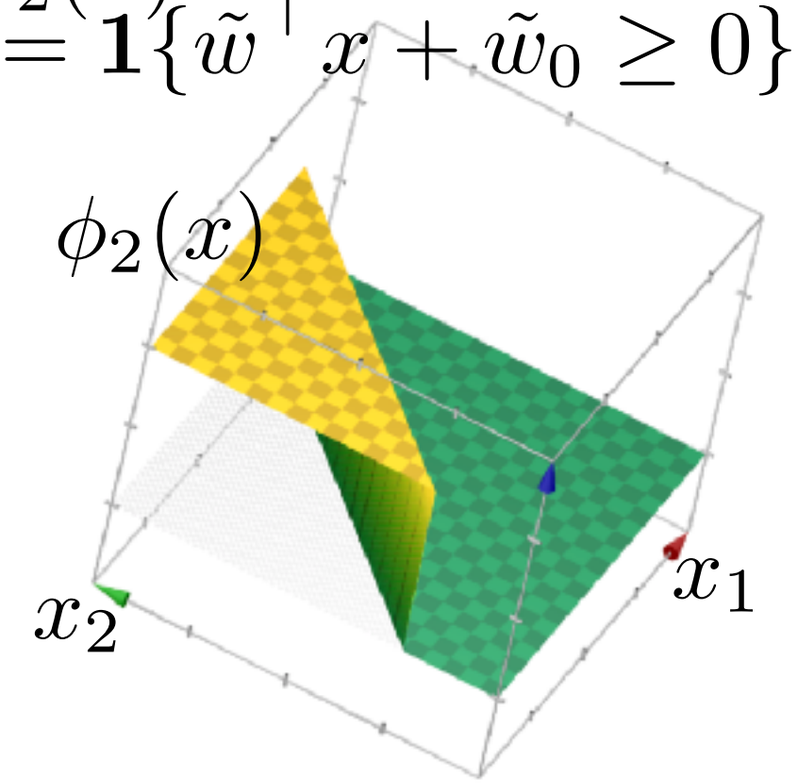



# New features: step functions!

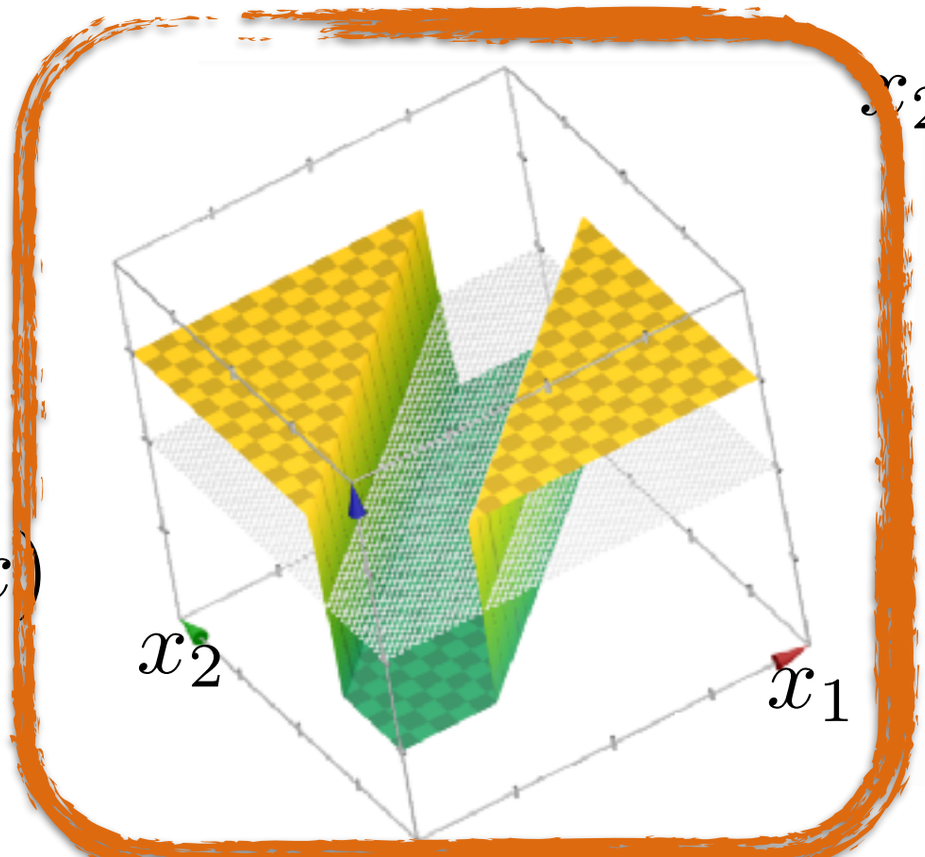
$$\phi_1(x) = \mathbf{1}\{w^\top x + w_0 \geq 0\}$$

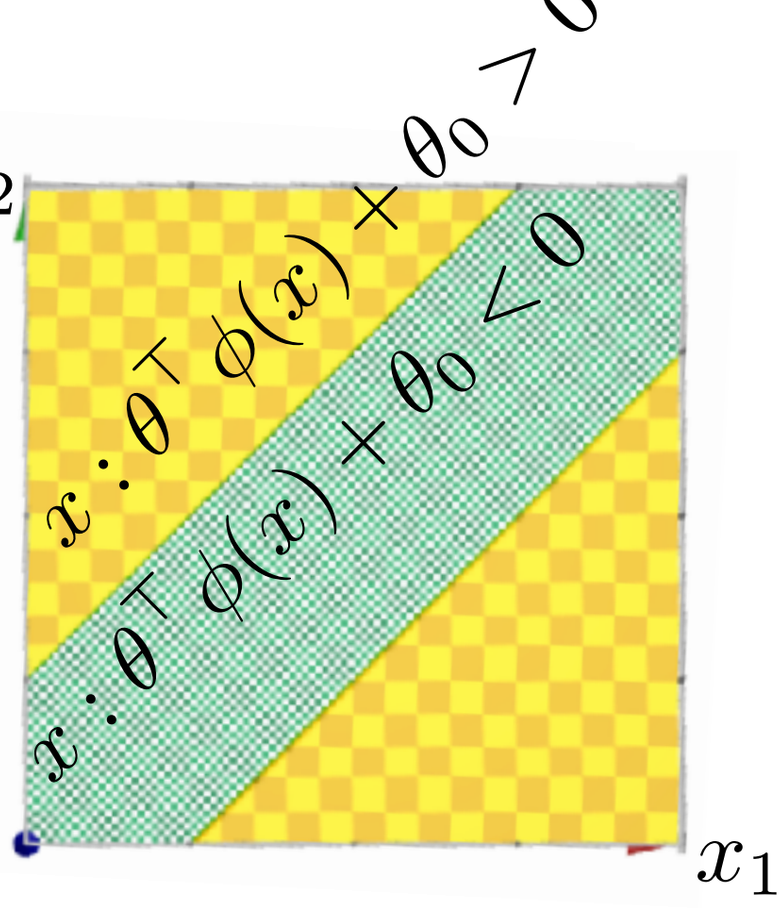



$$\phi_2(x) = \mathbf{1}\{\tilde{w}^\top x + \tilde{w}_0 \geq 0\}$$



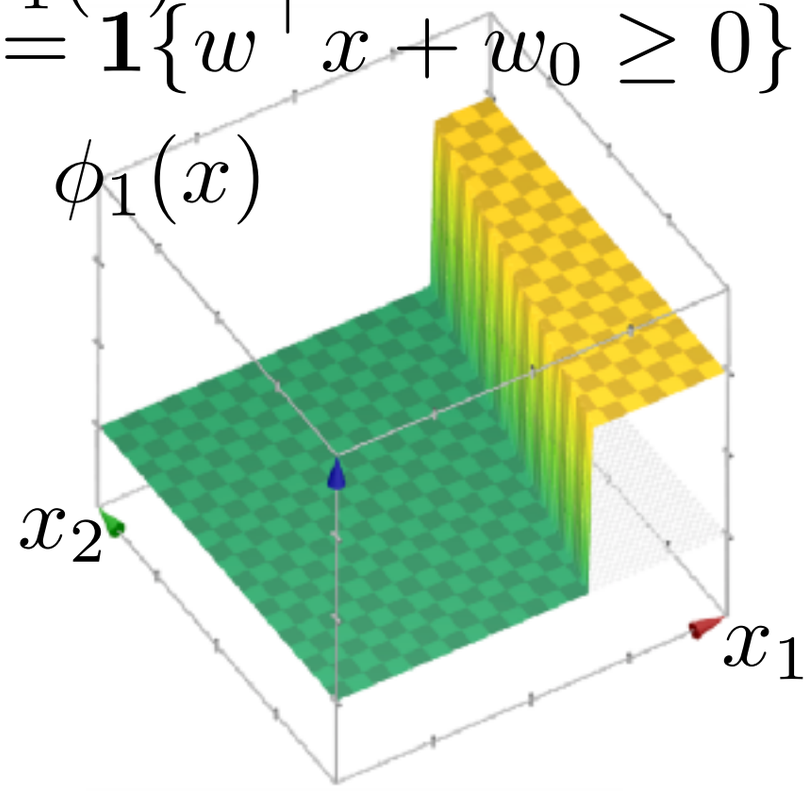

$$z = \theta^\top \phi(x) + \theta_0$$
$$= \theta_1 \phi_1(x) + \theta_2 \phi_2(x) + \theta_0$$
$$= 1 \cdot \phi_1(x) + 1 \cdot \phi_2(x) + (-0.5)$$



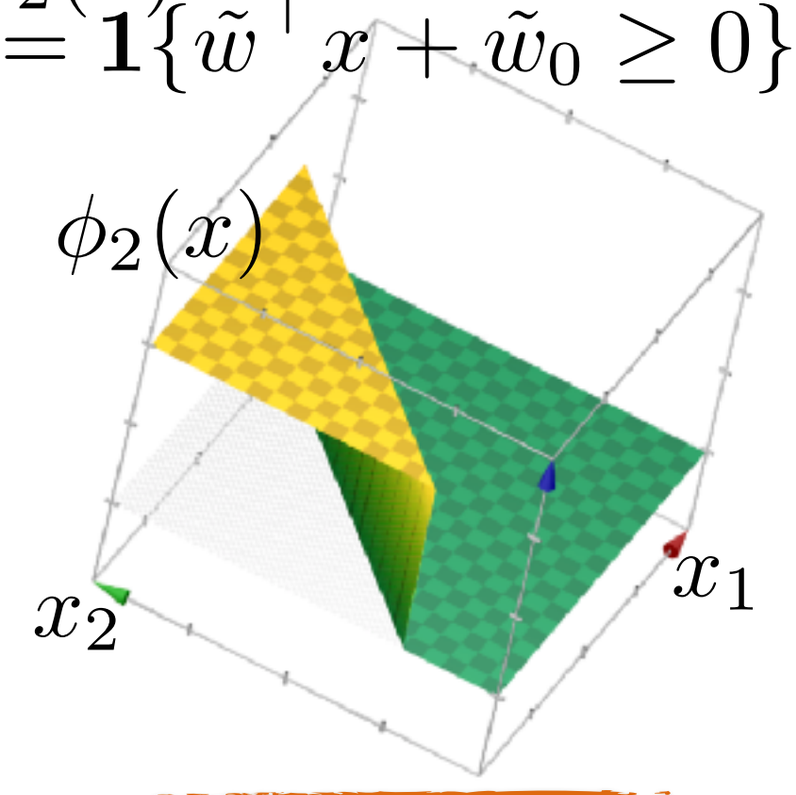



# New features: step functions!

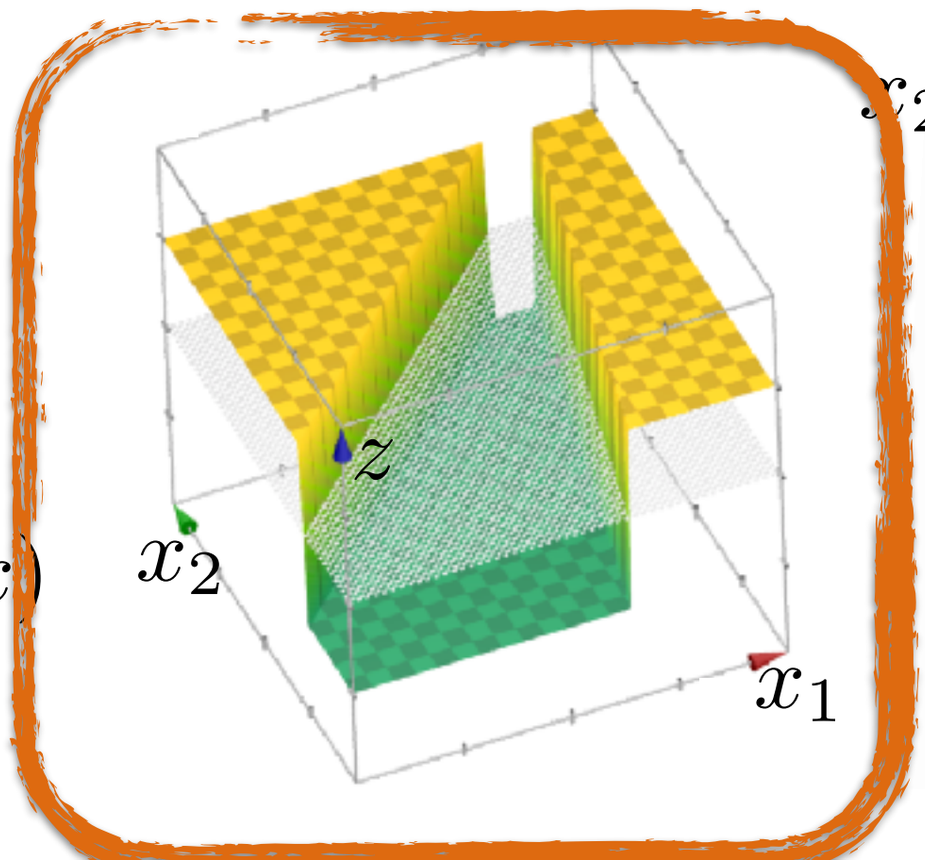
$$\phi_1(x) = \mathbf{1}\{w^\top x + w_0 \geq 0\}$$

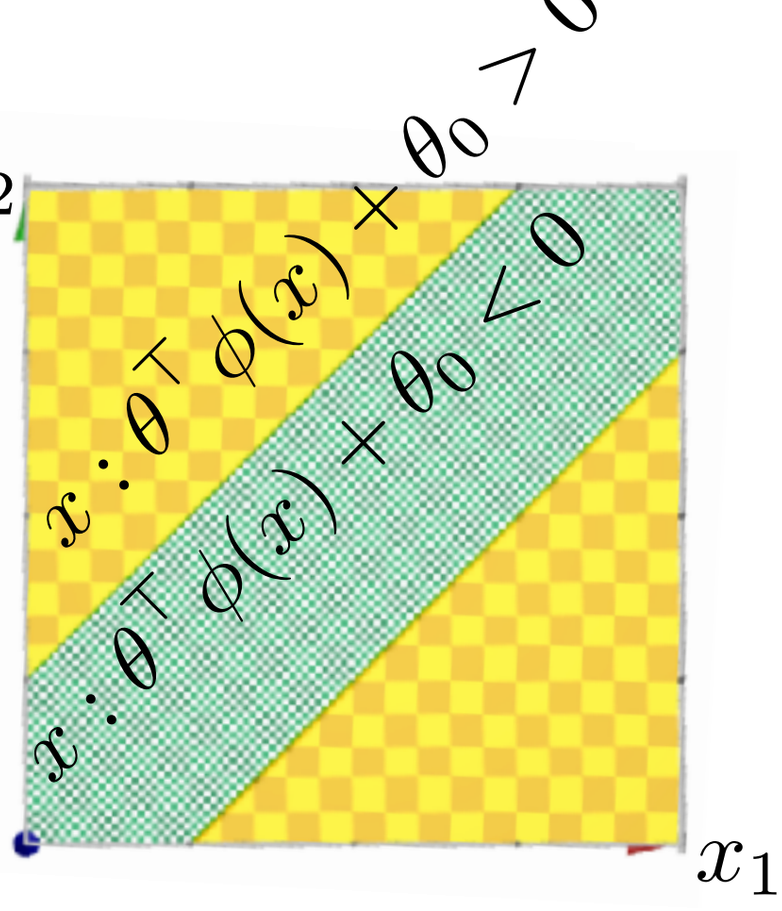



$$\phi_2(x) = \mathbf{1}\{\tilde{w}^\top x + \tilde{w}_0 \geq 0\}$$



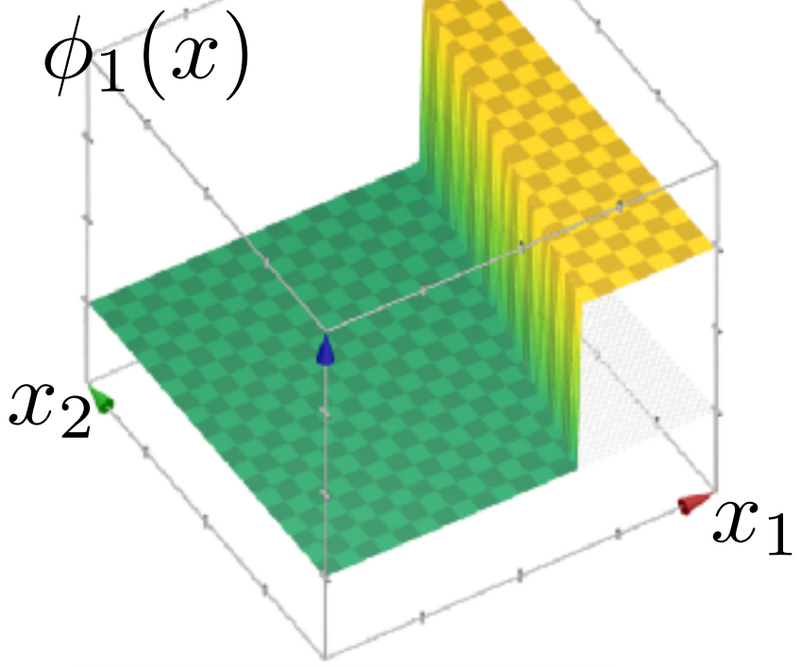

$$z = \theta^\top \phi(x) + \theta_0$$
$$= \theta_1 \phi_1(x) + \theta_2 \phi_2(x) + \theta_0$$
$$= 1 \cdot \phi_1(x) + 1 \cdot \phi_2(x) + (-0.5)$$



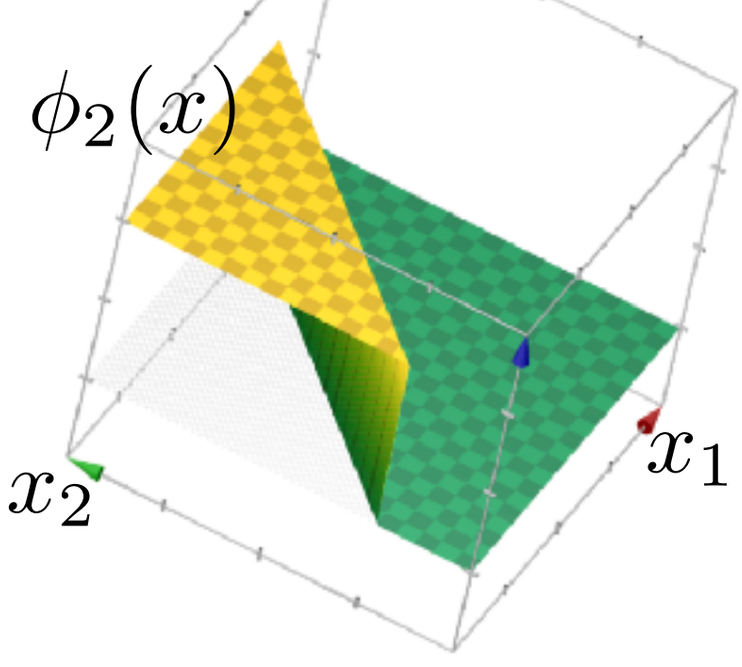



# New features: step functions!

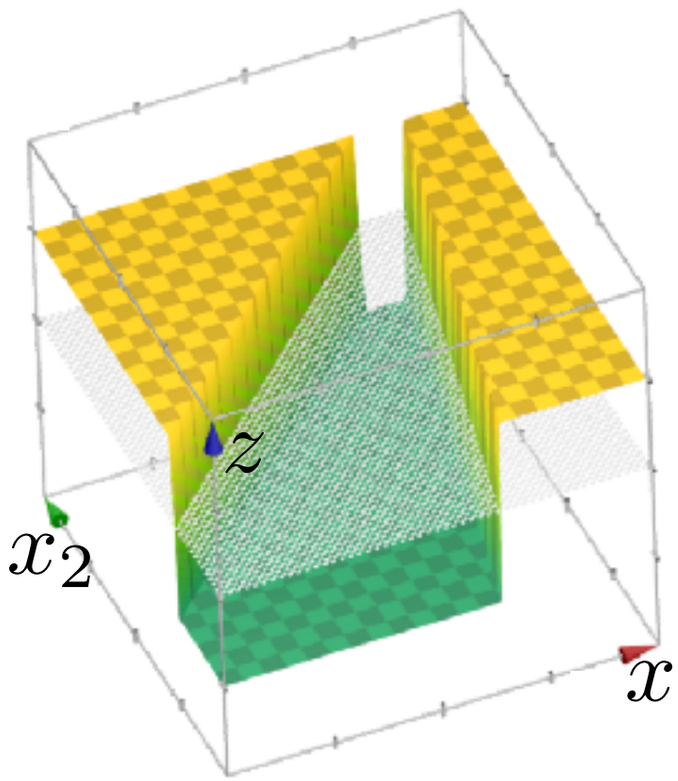
$$\phi_1(x) = \mathbf{1}\{w^\top x + w_0 \geq 0\}$$

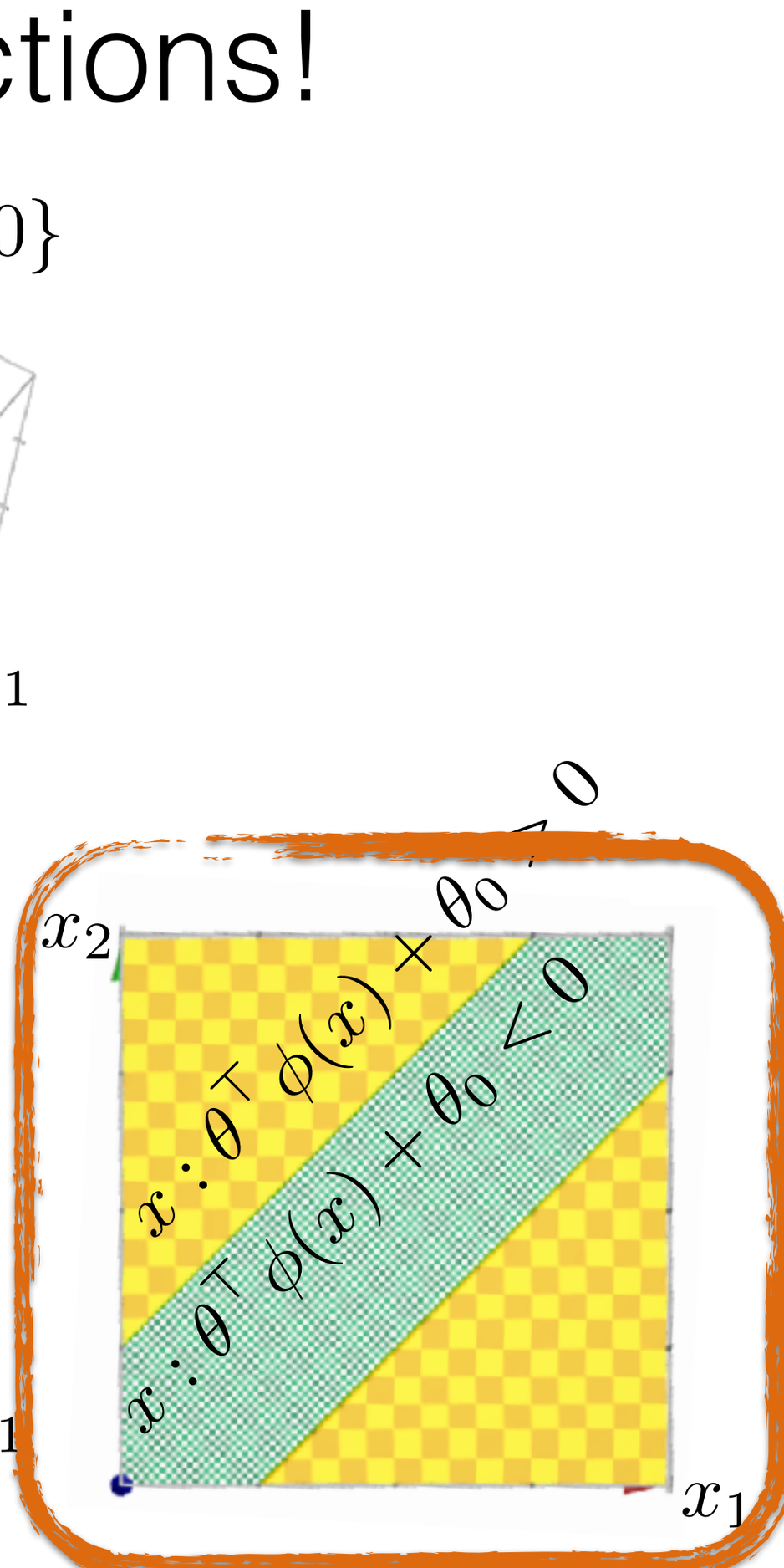



$$\phi_2(x) = \mathbf{1}\{\tilde{w}^\top x + \tilde{w}_0 \geq 0\}$$



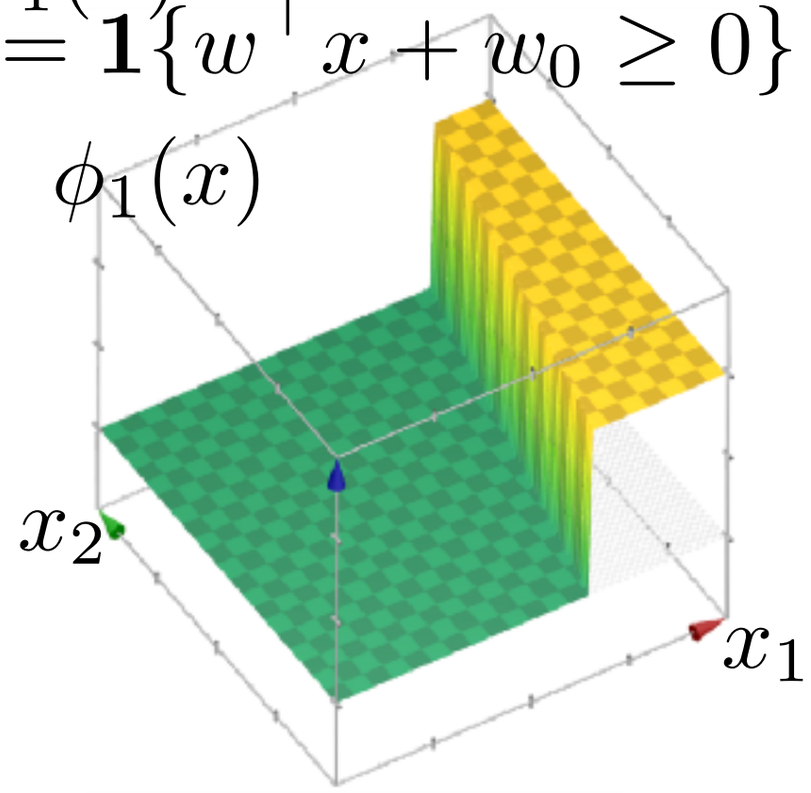

$$z = \theta^\top \phi(x) + \theta_0$$
$$= \theta_1 \phi_1(x) + \theta_2 \phi_2(x) + \theta_0$$
$$= 1 \cdot \phi_1(x) + 1 \cdot \phi_2(x) + (-0.5)$$



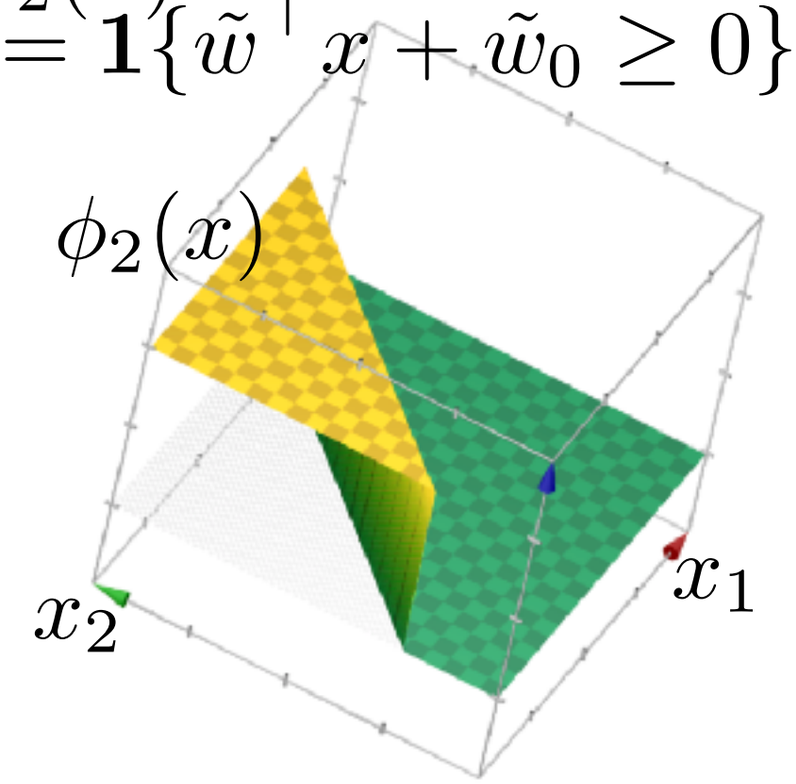



# New features: step functions!

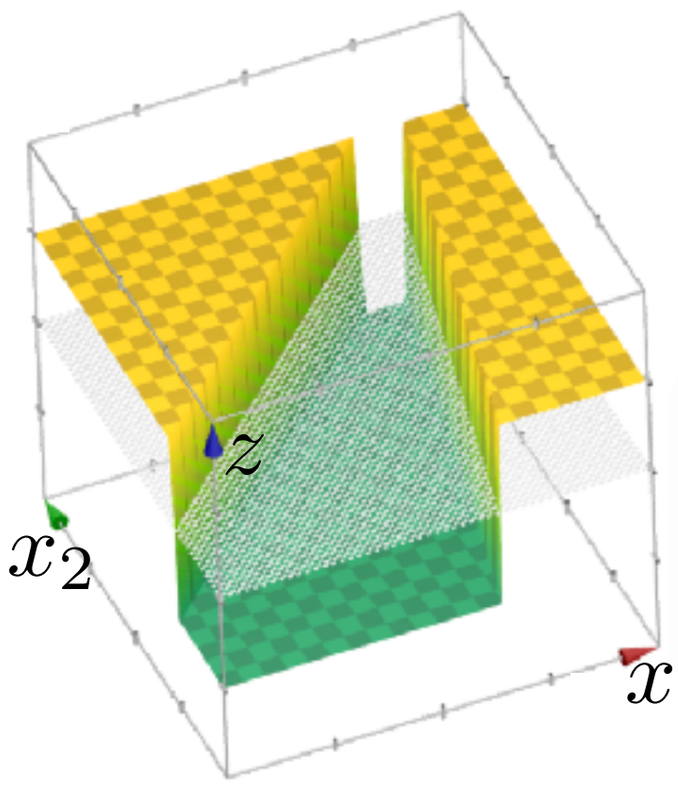
$$\phi_1(x) = \mathbf{1}\{w^\top x + w_0 \geq 0\}$$

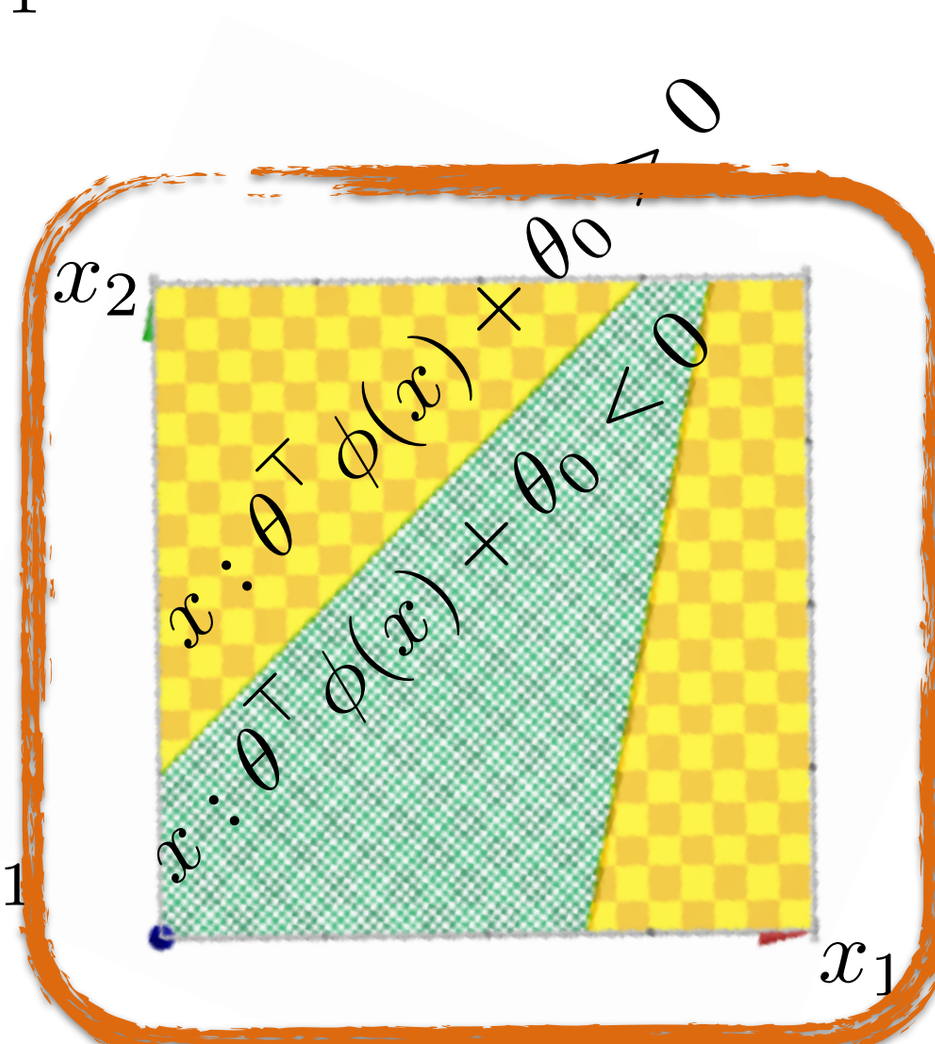



$$\phi_2(x) = \mathbf{1}\{\tilde{w}^\top x + \tilde{w}_0 \geq 0\}$$



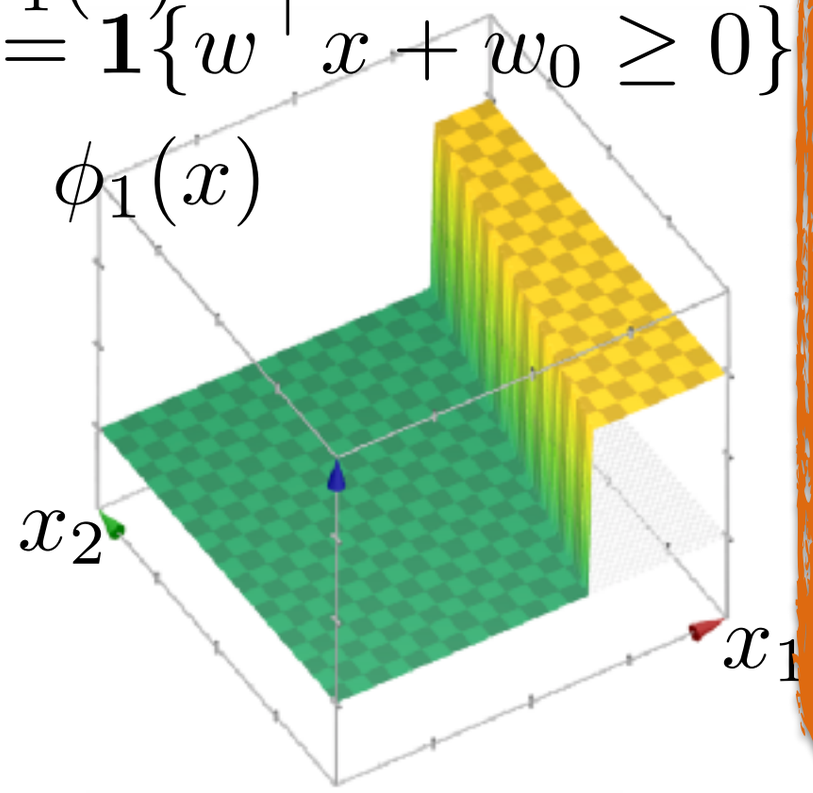

$$z = \theta^\top \phi(x) + \theta_0$$
$$= \theta_1 \phi_1(x) + \theta_2 \phi_2(x) + \theta_0$$
$$= 1 \cdot \phi_1(x) + 1 \cdot \phi_2(x) + (-0.5)$$



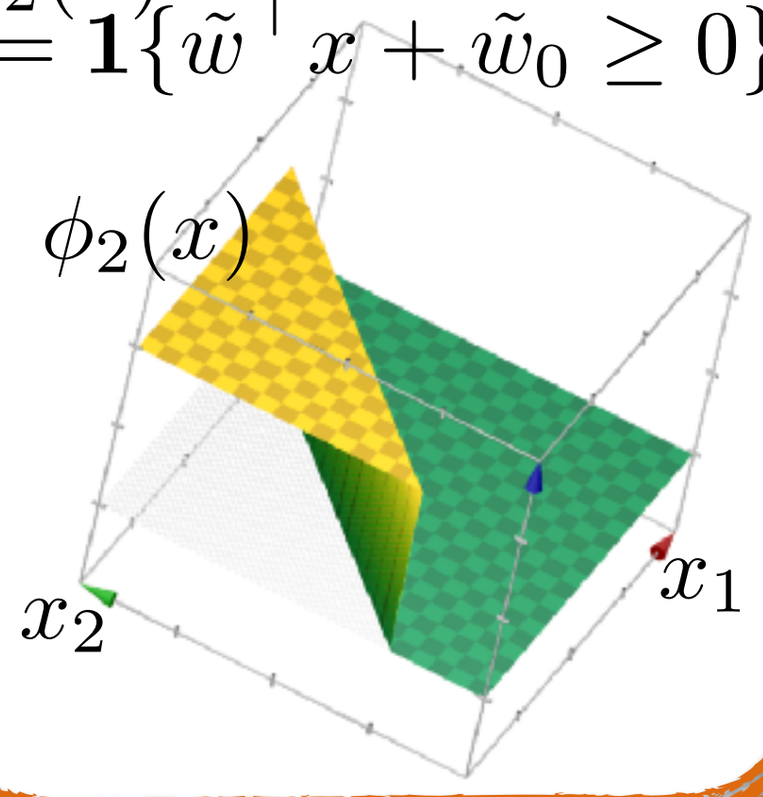



# New features: step functions!

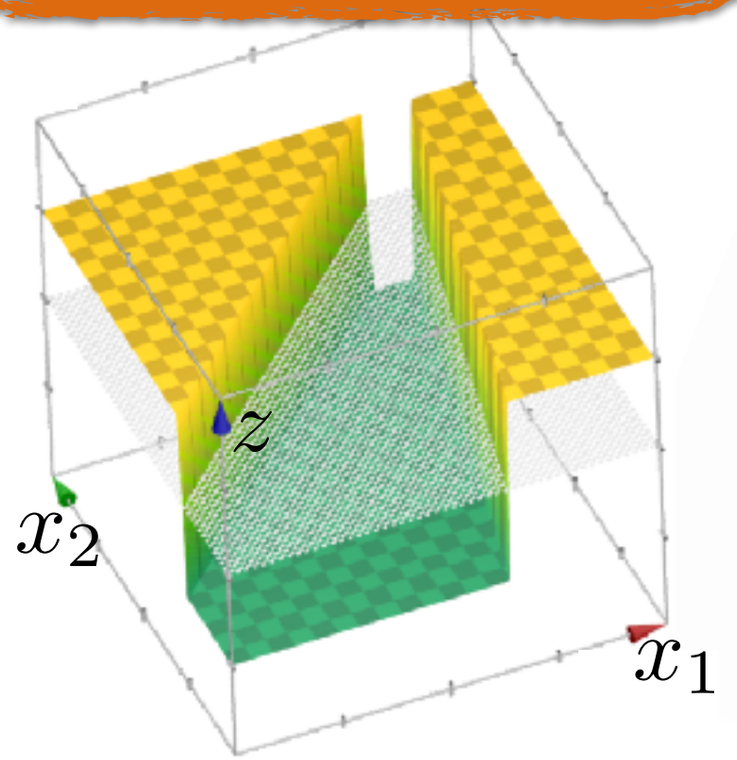
$$\phi_1(x) = \mathbf{1}\{w^\top x + w_0 \geq 0\}$$

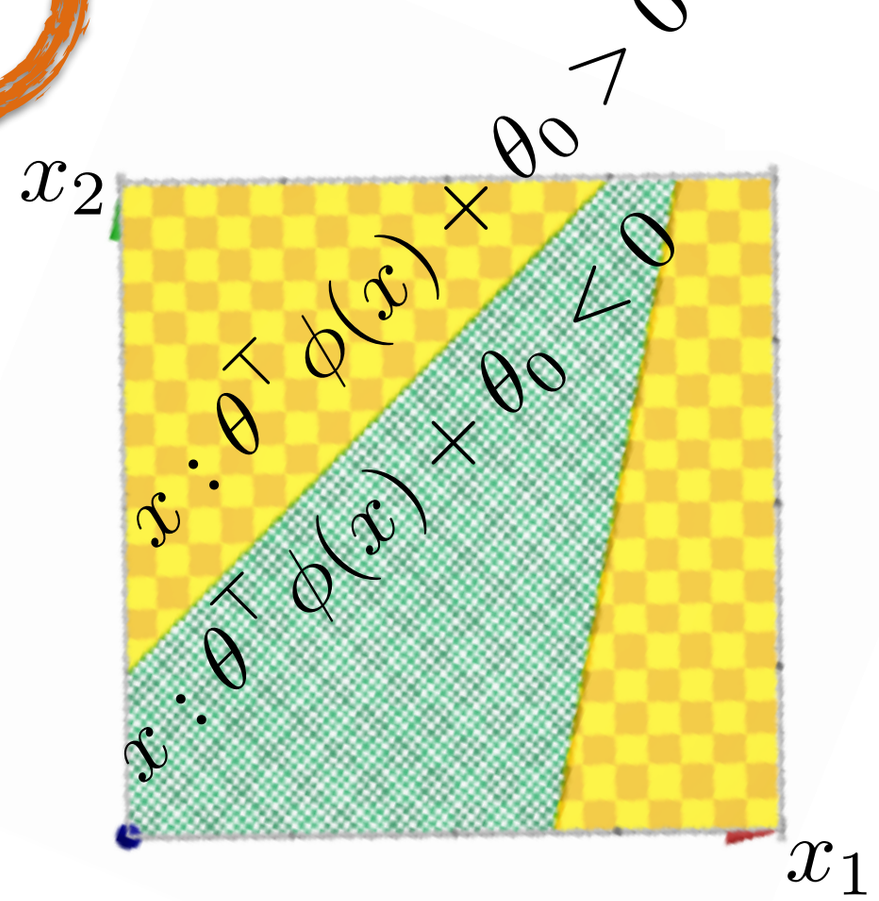



$$\phi_2(x) = \mathbf{1}\{\tilde{w}^\top x + \tilde{w}_0 \geq 0\}$$



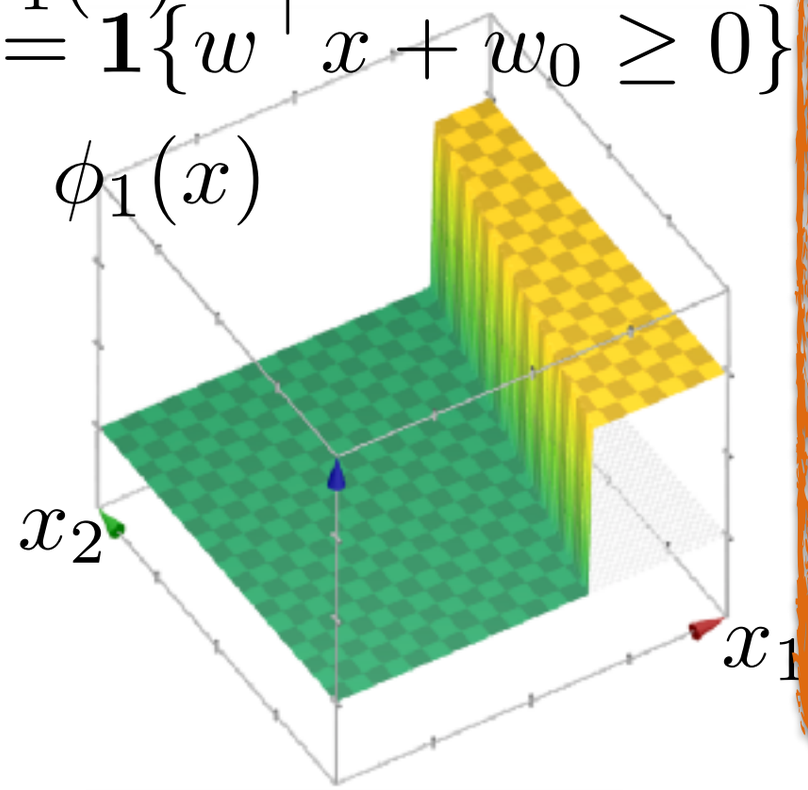

$$z = \theta^\top \phi(x) + \theta_0$$
$$= \theta_1 \phi_1(x) + \theta_2 \phi_2(x) + \theta_0$$
$$= 1 \cdot \phi_1(x) + 1 \cdot \phi_2(x) + (-0.5)$$



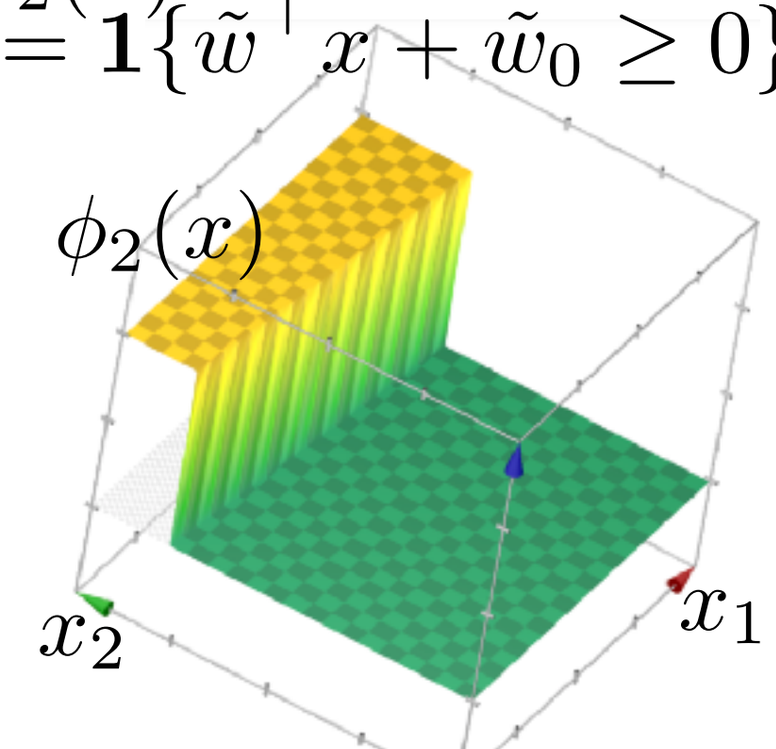



# New features: step functions!

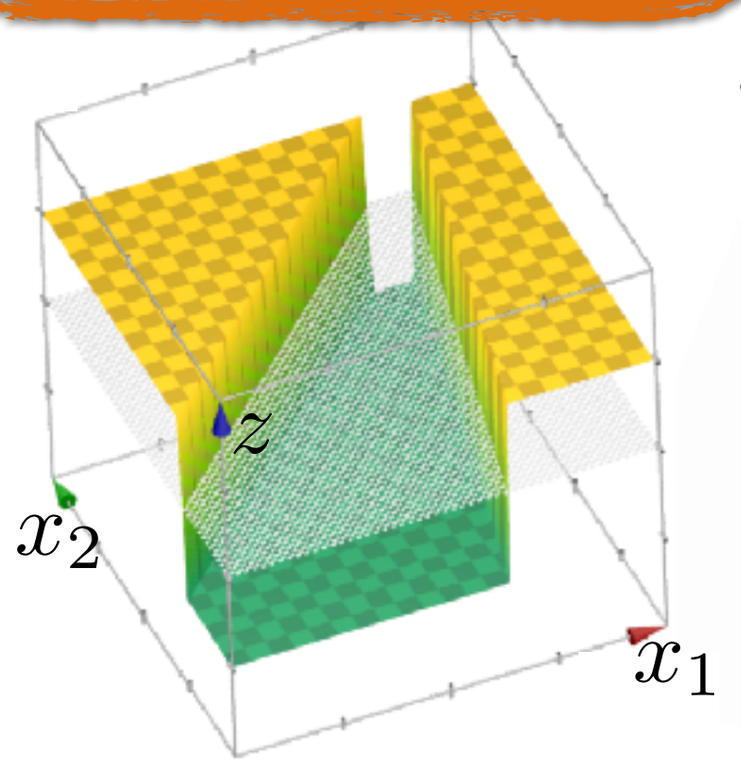
$$\phi_1(x) = \mathbf{1}\{w^\top x + w_0 \geq 0\}$$

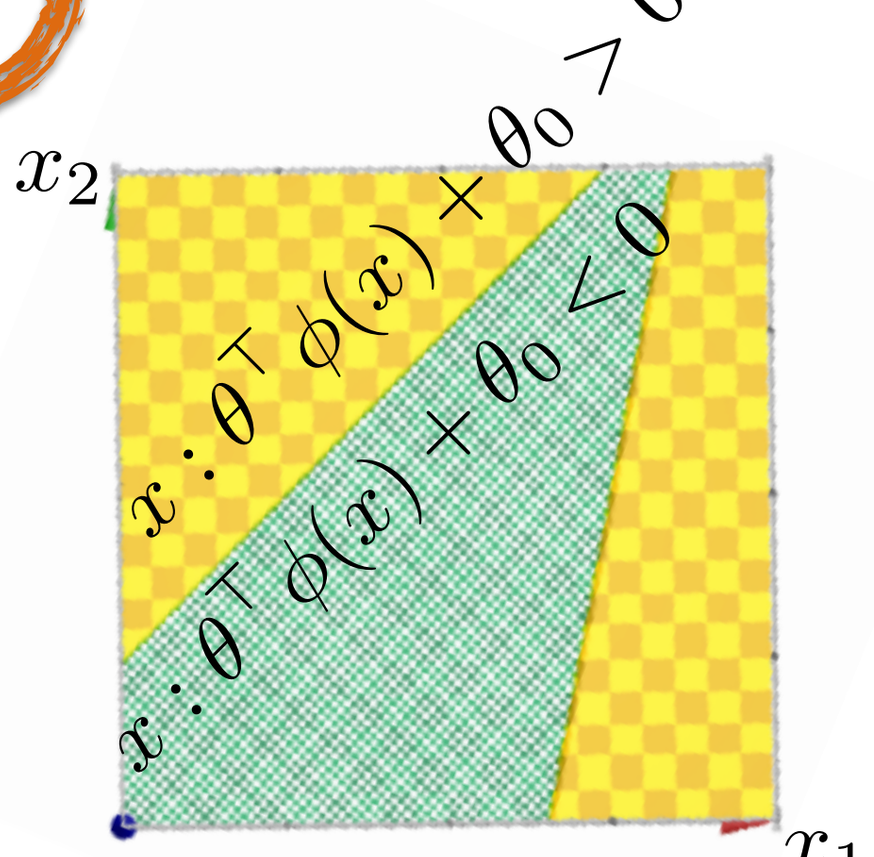



$$\phi_2(x) = \mathbf{1}\{\tilde{w}^\top x + \tilde{w}_0 \geq 0\}$$



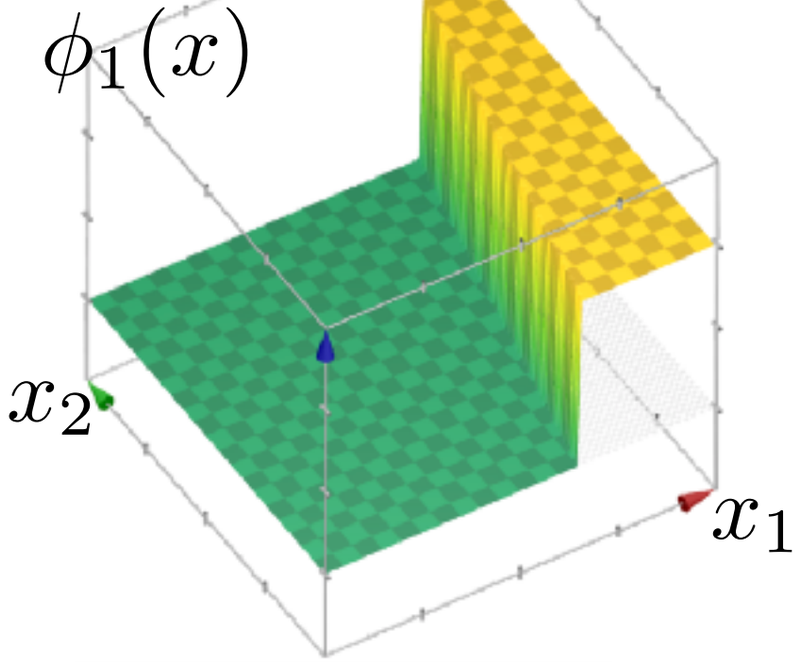

$$z = \theta^\top \phi(x) + \theta_0$$
$$= \theta_1 \phi_1(x) + \theta_2 \phi_2(x) + \theta_0$$
$$= 1 \cdot \phi_1(x) + 1 \cdot \phi_2(x) + (-0.5)$$



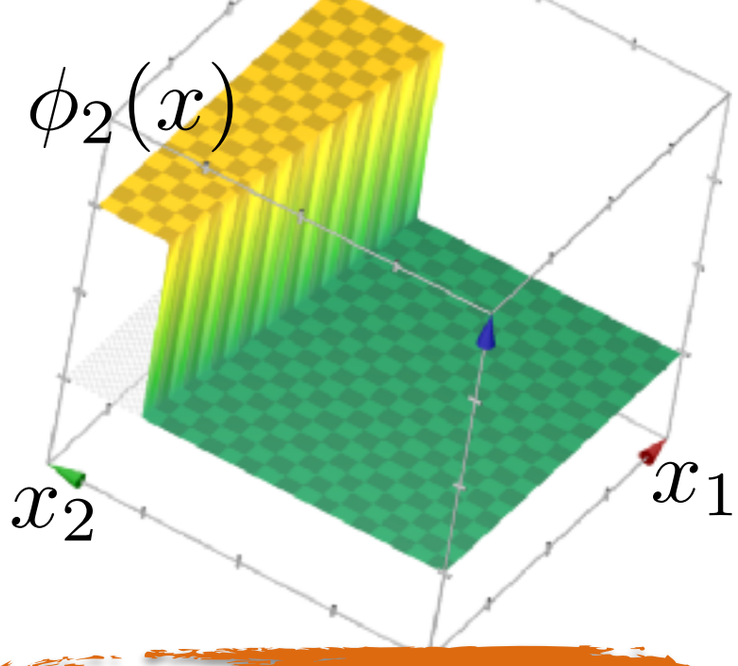



# New features: step functions!

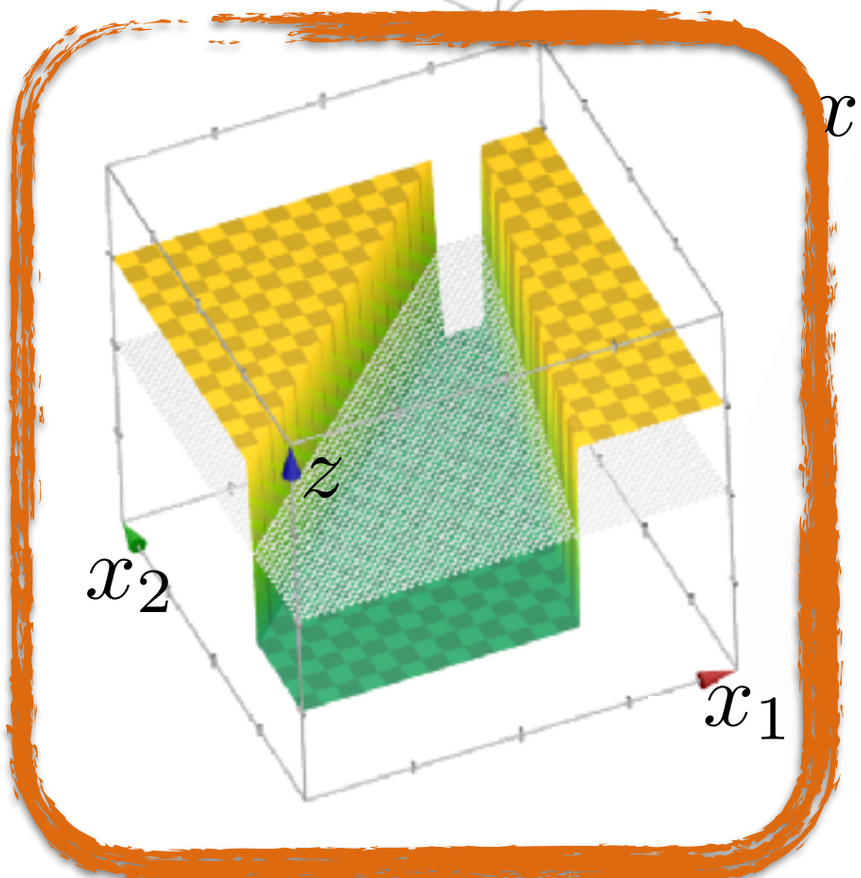
$$\phi_1(x) = \mathbf{1}\{w^\top x + w_0 \geq 0\}$$

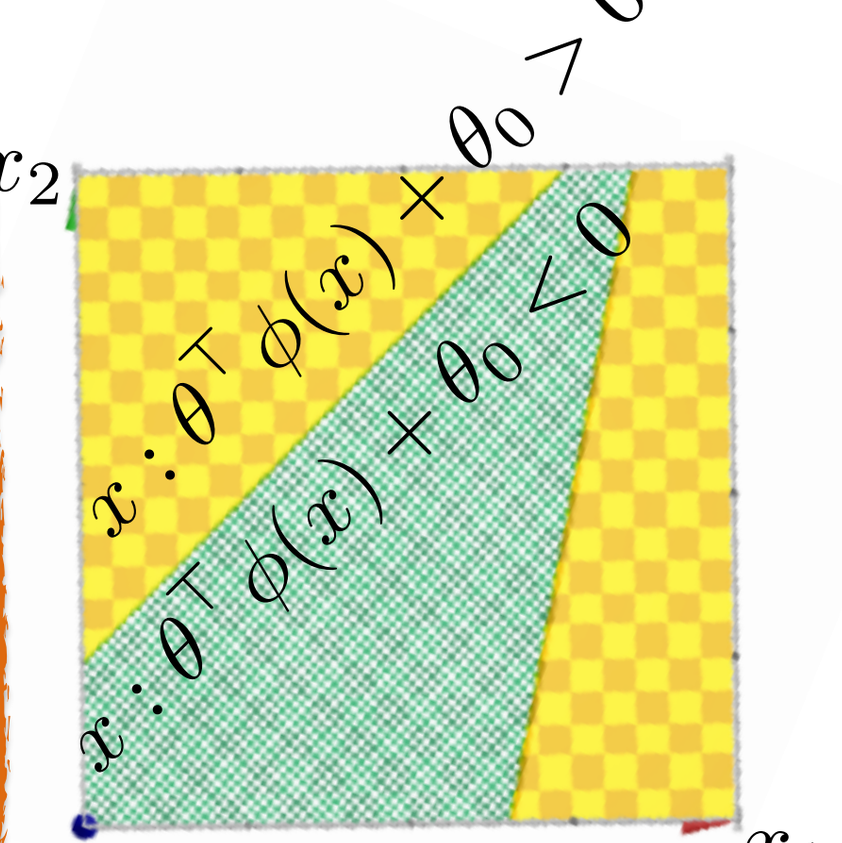



$$\phi_2(x) = \mathbf{1}\{\tilde{w}^\top x + \tilde{w}_0 \geq 0\}$$



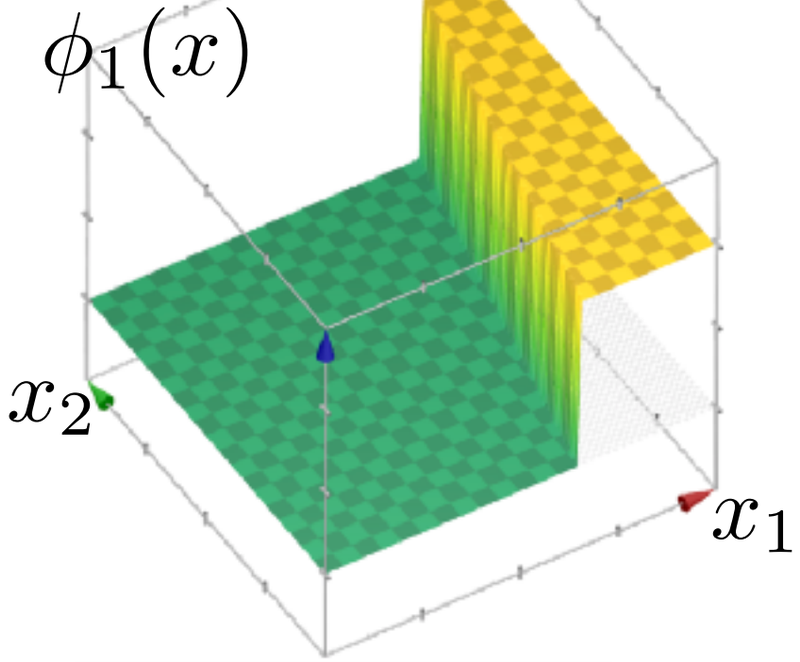

$$z = \theta^\top \phi(x) + \theta_0$$
$$= \theta_1 \phi_1(x) + \theta_2 \phi_2(x) + \theta_0$$
$$= 1 \cdot \phi_1(x) + 1 \cdot \phi_2(x) + (-0.5)$$



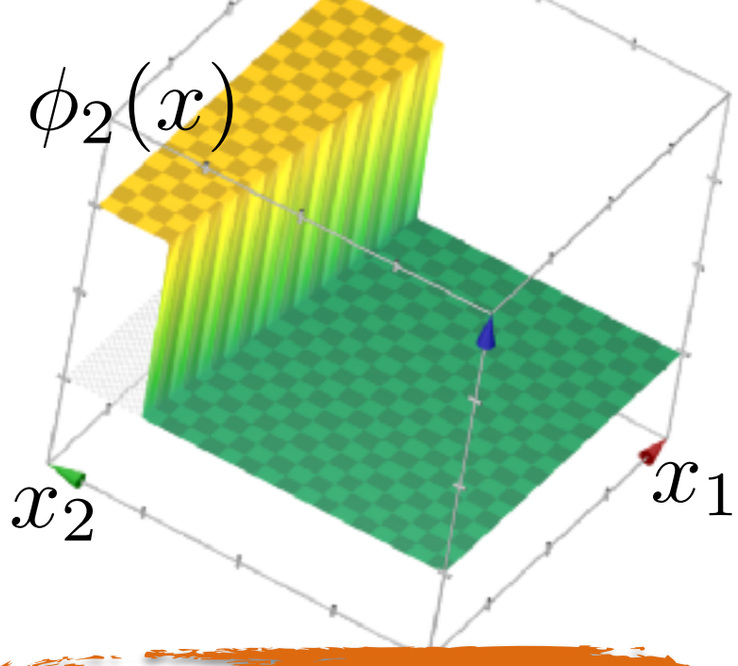



# New features: step functions!

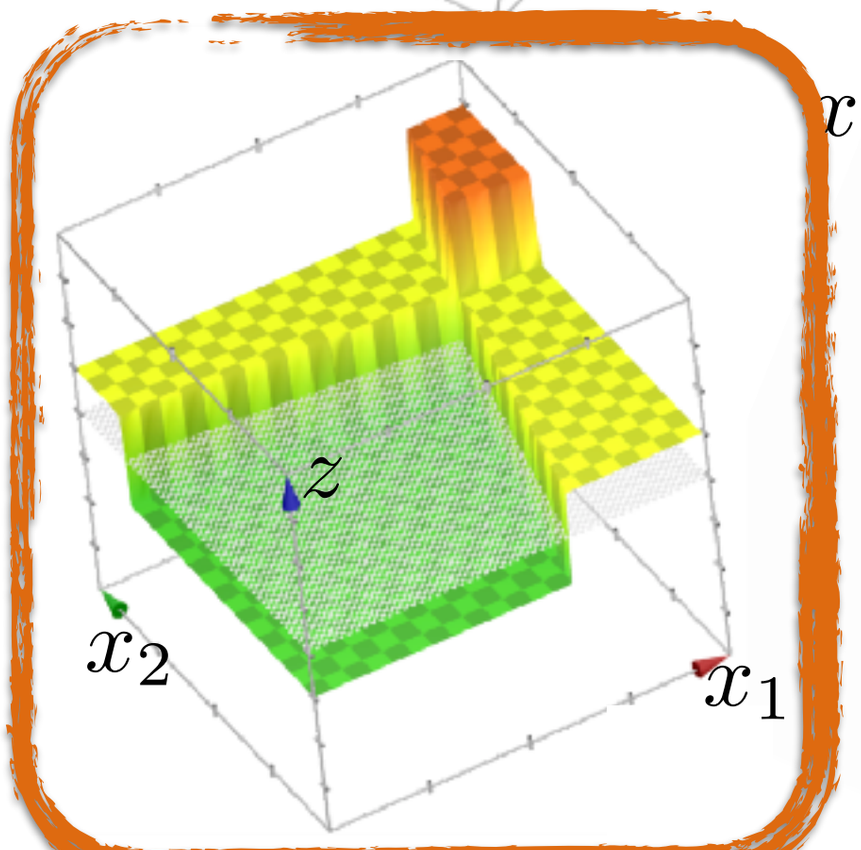
$$\phi_1(x) = \mathbf{1}\{w^\top x + w_0 \geq 0\}$$

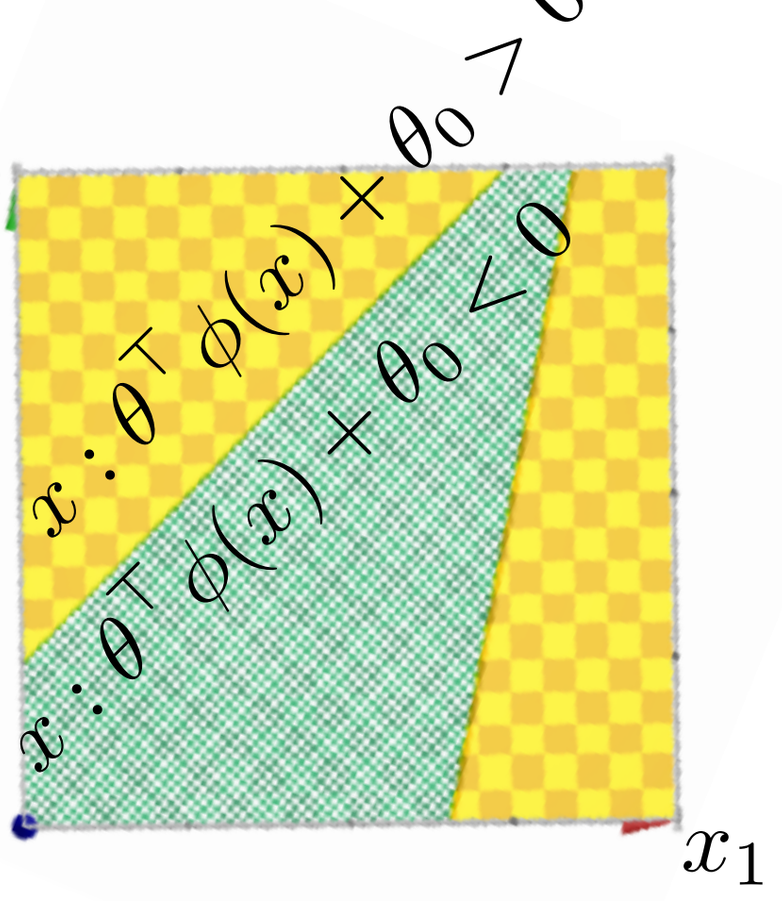



$$\phi_2(x) = \mathbf{1}\{\tilde{w}^\top x + \tilde{w}_0 \geq 0\}$$



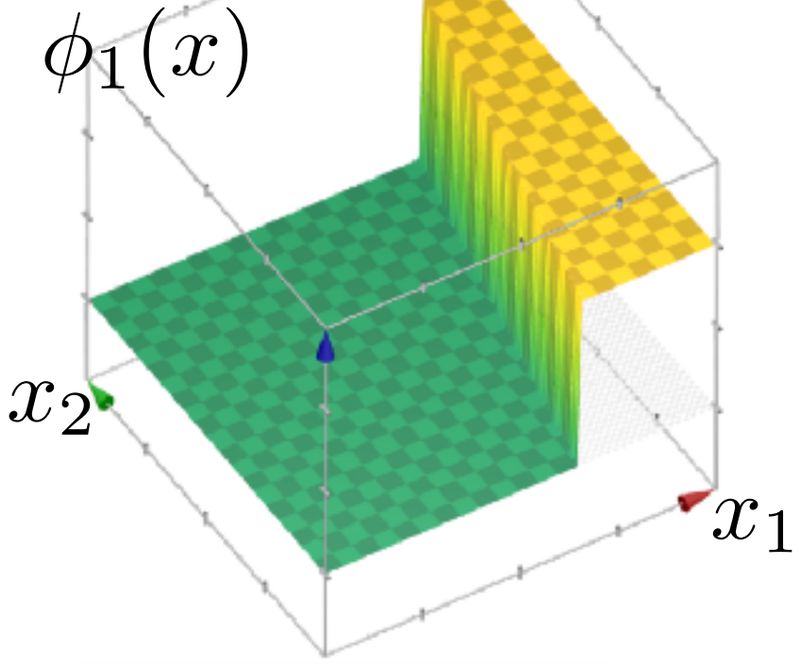

$$z = \theta^\top \phi(x) + \theta_0$$
$$= \theta_1 \phi_1(x) + \theta_2 \phi_2(x) + \theta_0$$
$$= 1 \cdot \phi_1(x) + 1 \cdot \phi_2(x) + (-0.5)$$



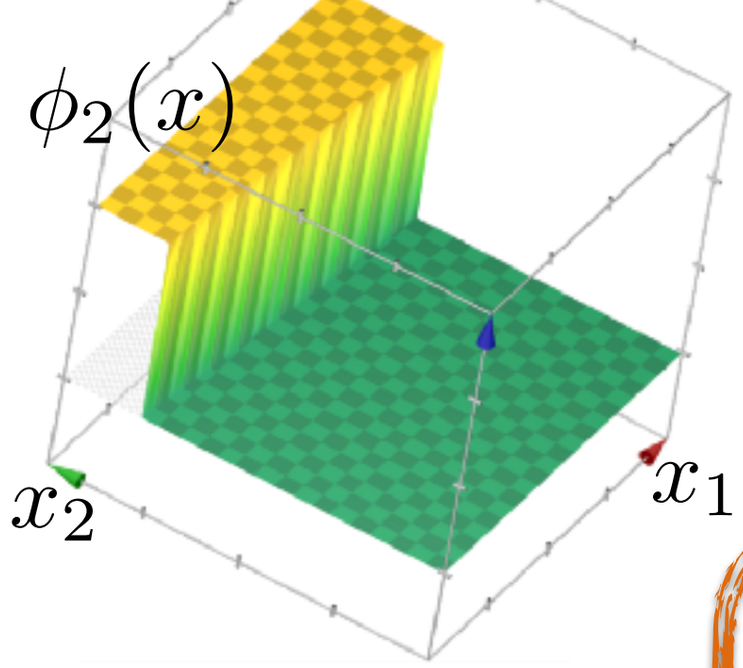



# New features: step functions!

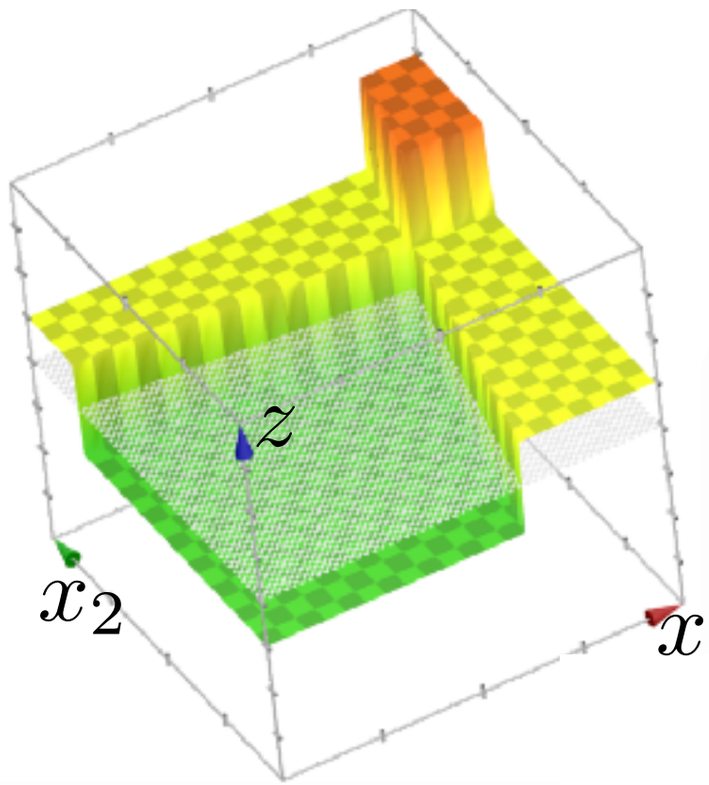
$$\phi_1(x) = \mathbf{1}\{w^\top x + w_0 \geq 0\}$$

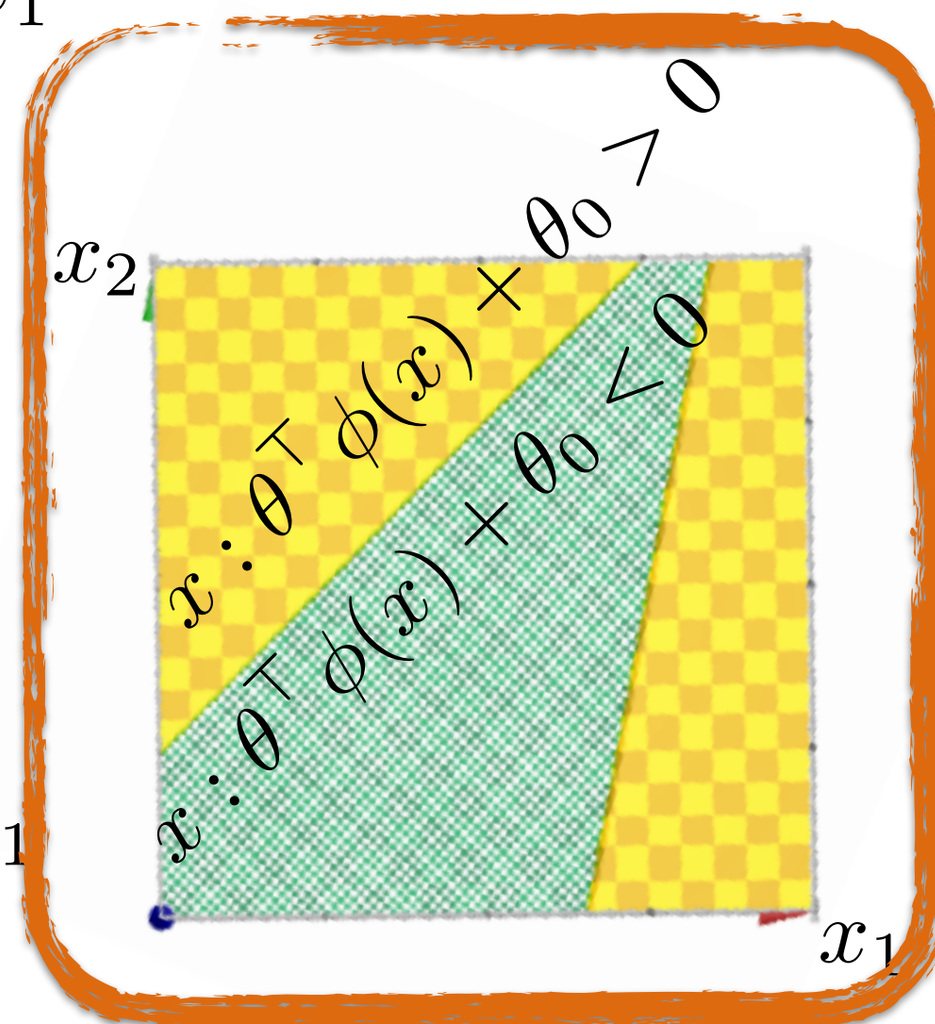



$$\phi_2(x) = \mathbf{1}\{\tilde{w}^\top x + \tilde{w}_0 \geq 0\}$$



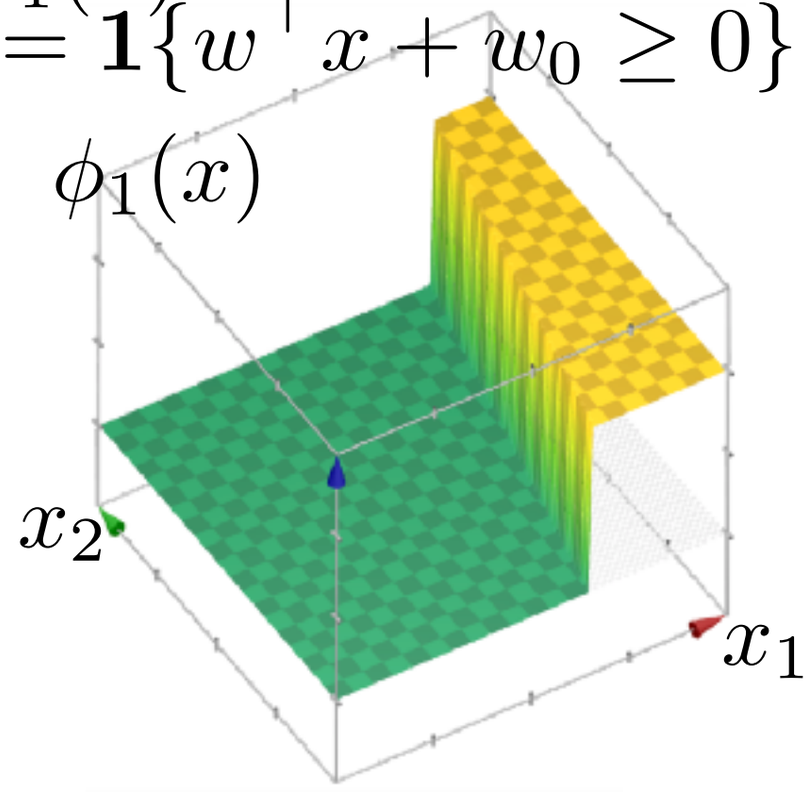

$$z = \theta^\top \phi(x) + \theta_0$$
$$= \theta_1 \phi_1(x) + \theta_2 \phi_2(x) + \theta_0$$
$$= 1 \cdot \phi_1(x) + 1 \cdot \phi_2(x) + (-0.5)$$



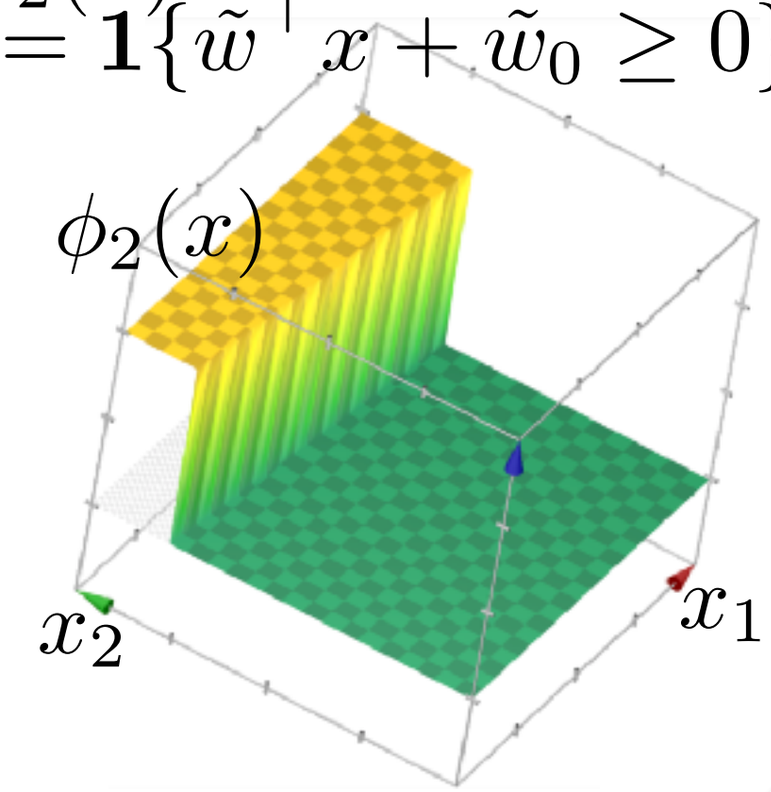



# New features: step functions!

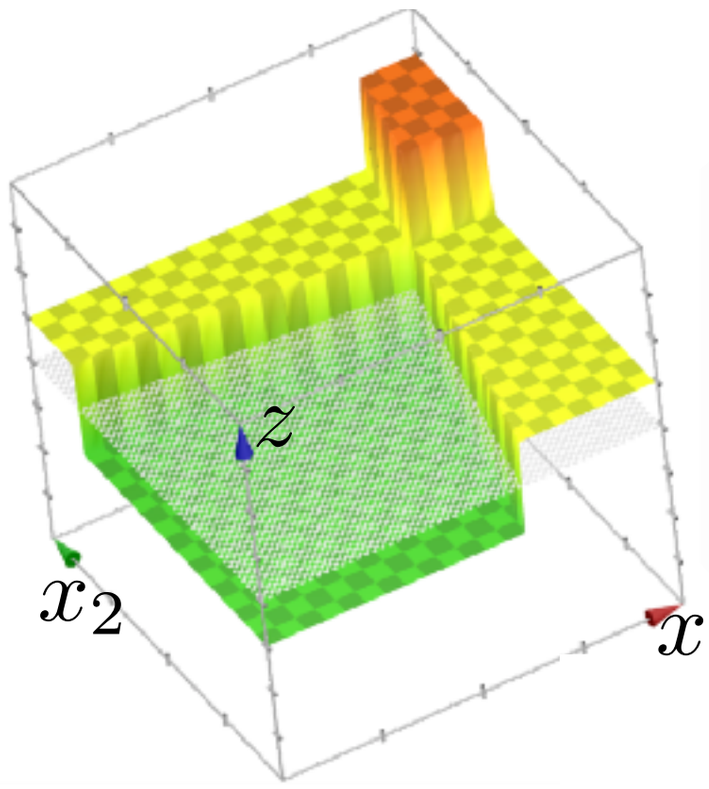
$$\phi_1(x) = \mathbf{1}\{w^\top x + w_0 \geq 0\}$$

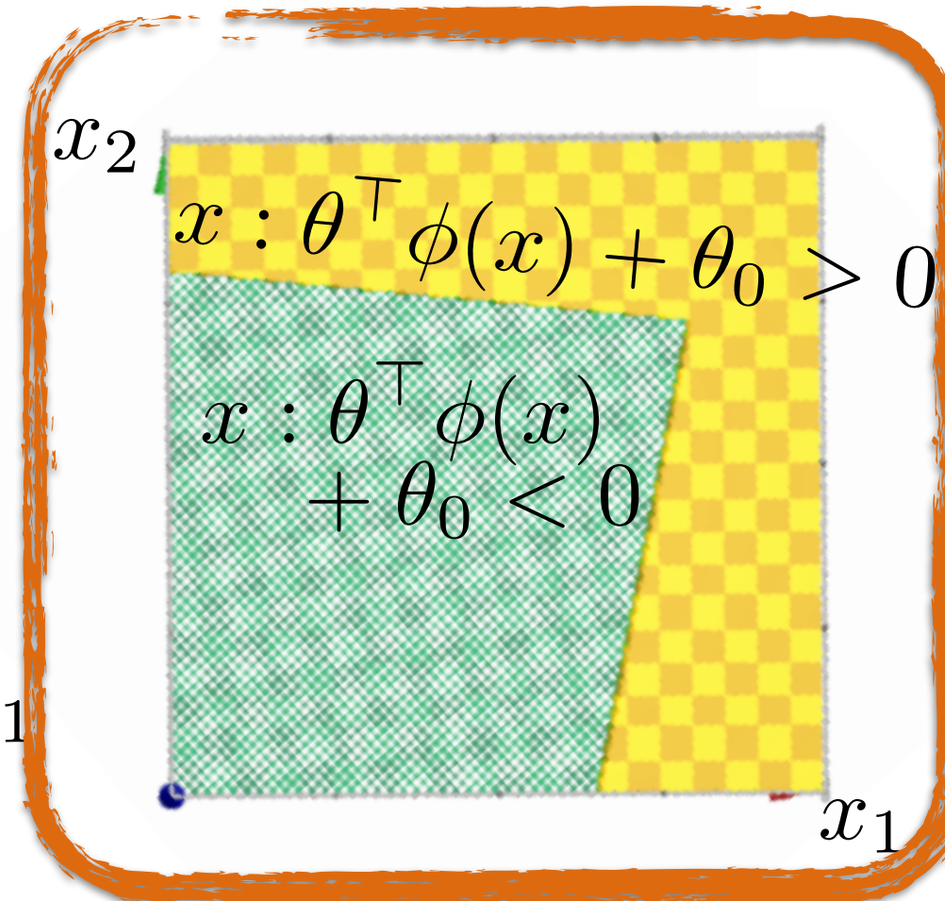



$$\phi_2(x) = \mathbf{1}\{\tilde{w}^\top x + \tilde{w}_0 \geq 0\}$$



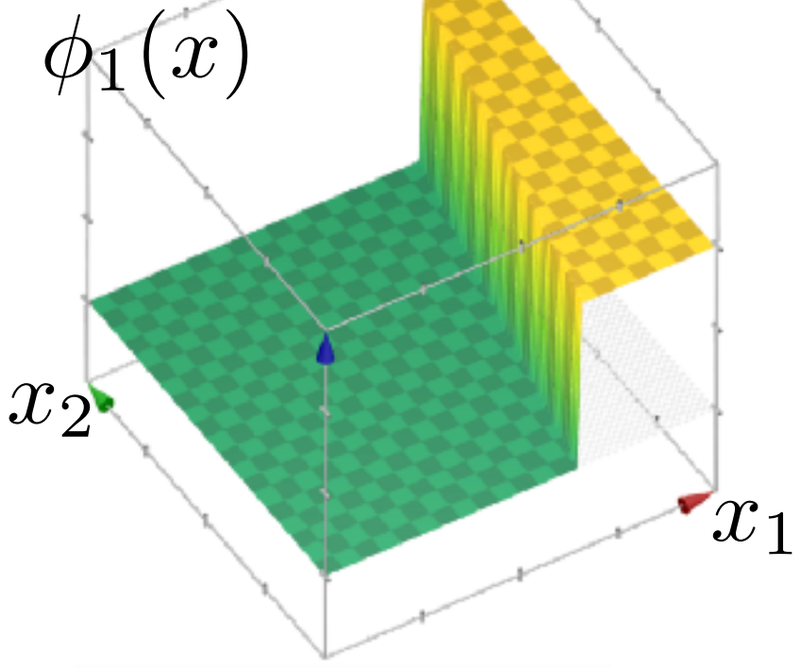

$$z = \theta^\top \phi(x) + \theta_0$$
$$= \theta_1 \phi_1(x) + \theta_2 \phi_2(x) + \theta_0$$
$$= 1 \cdot \phi_1(x) + 1 \cdot \phi_2(x) + (-0.5)$$



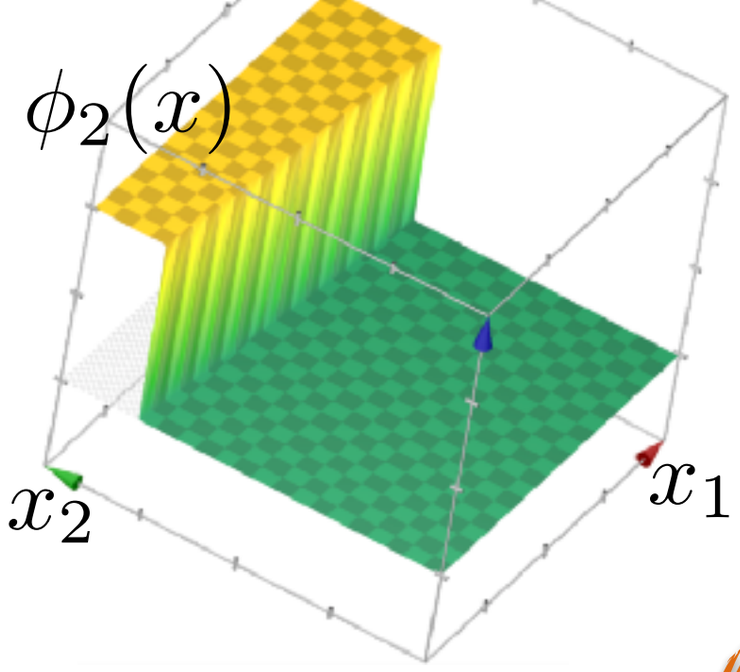



# New features: step functions!

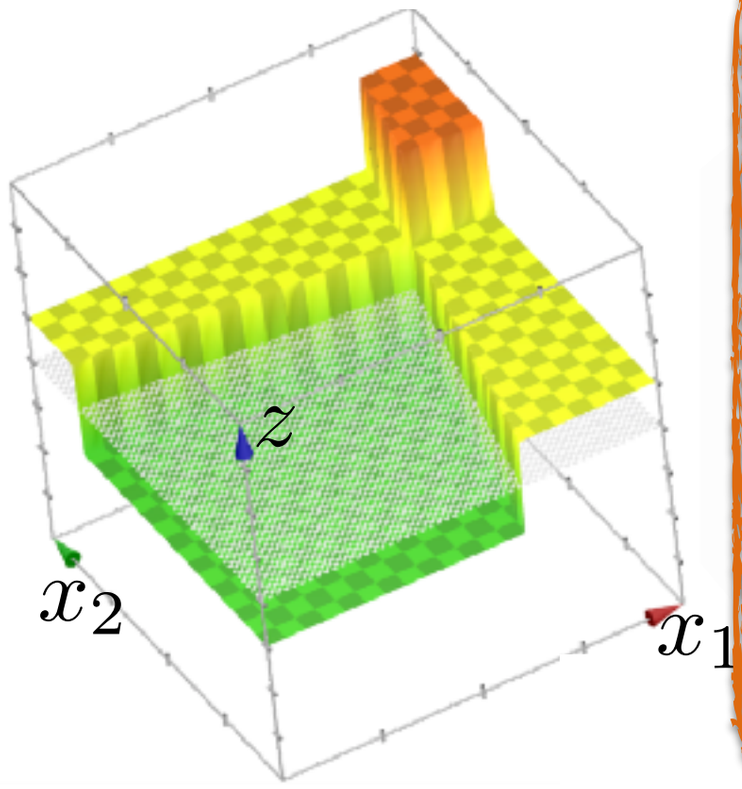
$$\phi_1(x) = \mathbf{1}\{w^\top x + w_0 \geq 0\}$$

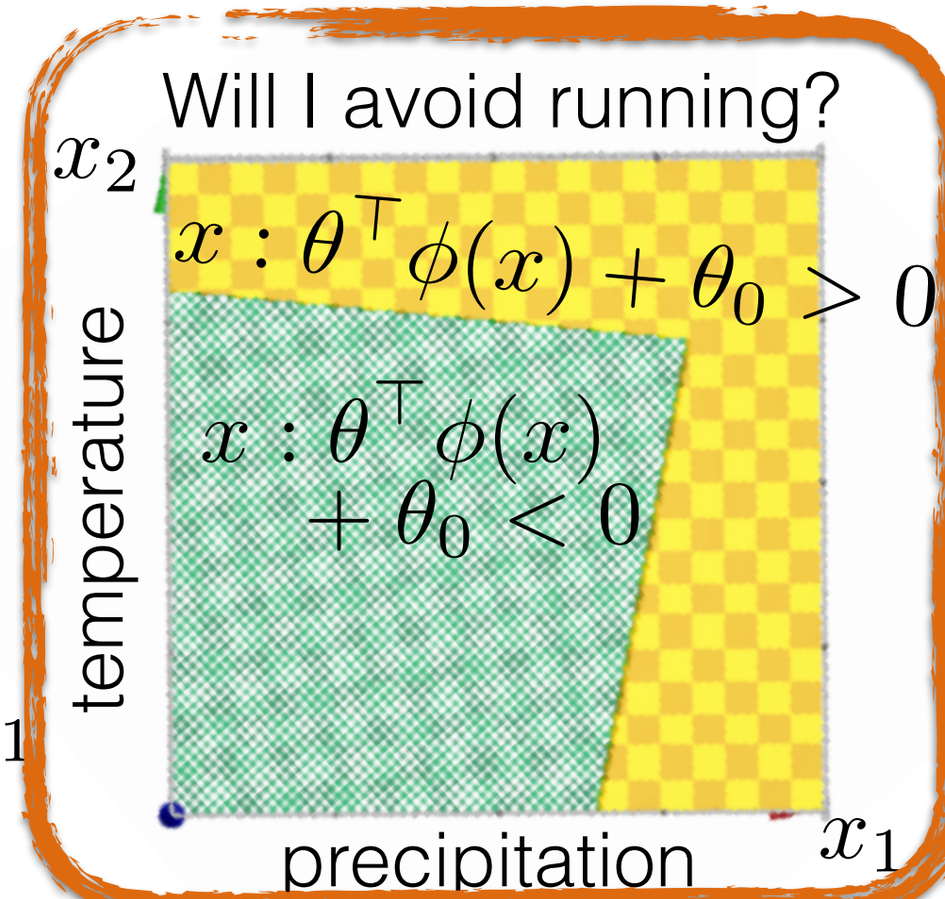



$$\phi_2(x) = \mathbf{1}\{\tilde{w}^\top x + \tilde{w}_0 \geq 0\}$$



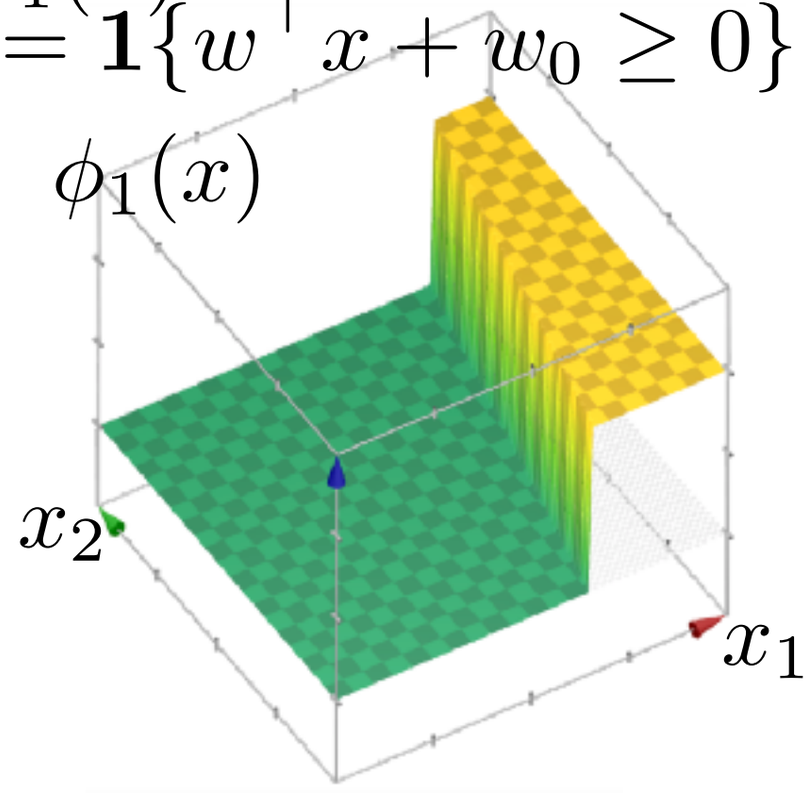

$$z = \theta^\top \phi(x) + \theta_0$$
$$= \theta_1 \phi_1(x) + \theta_2 \phi_2(x) + \theta_0$$
$$= 1 \cdot \phi_1(x) + 1 \cdot \phi_2(x) + (-0.5)$$



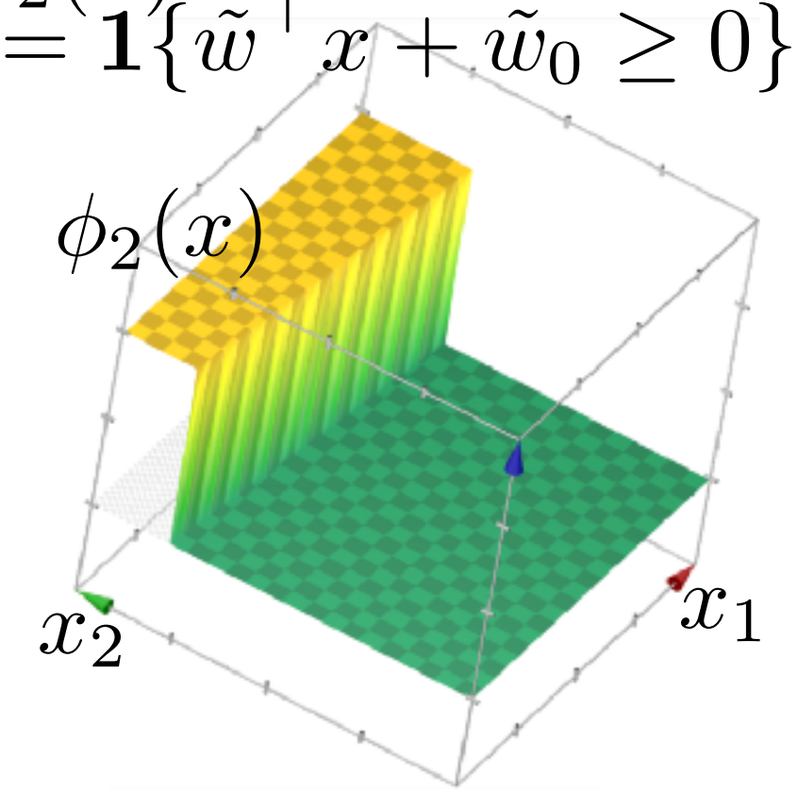



# New features: step functions!

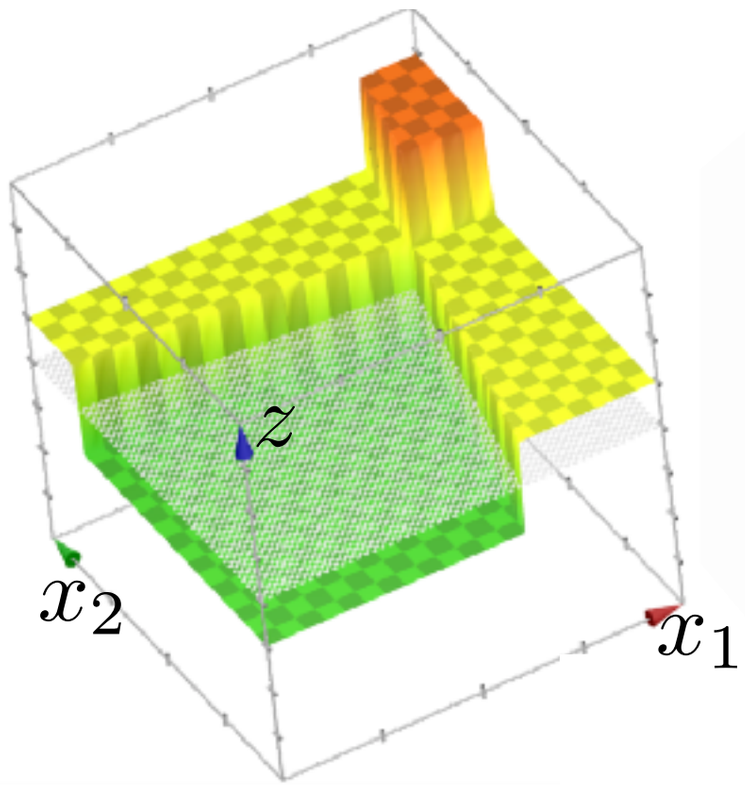
$$\phi_1(x) = \mathbf{1}\{w^\top x + w_0 \geq 0\}$$

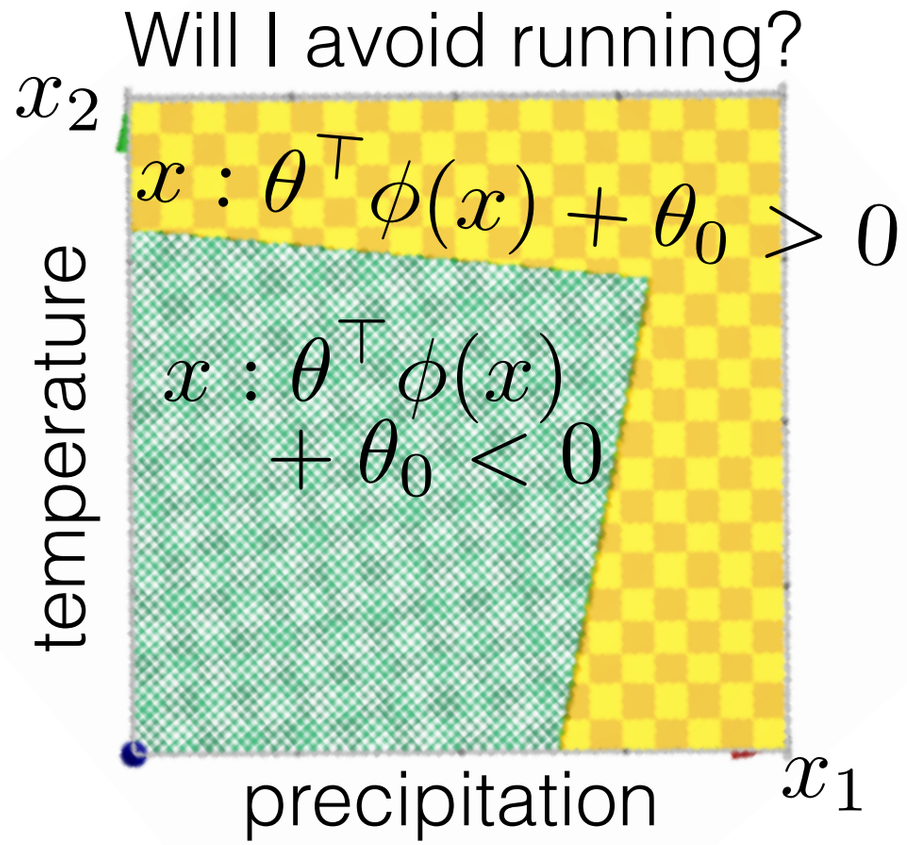



$$\phi_2(x) = \mathbf{1}\{\tilde{w}^\top x + \tilde{w}_0 \geq 0\}$$



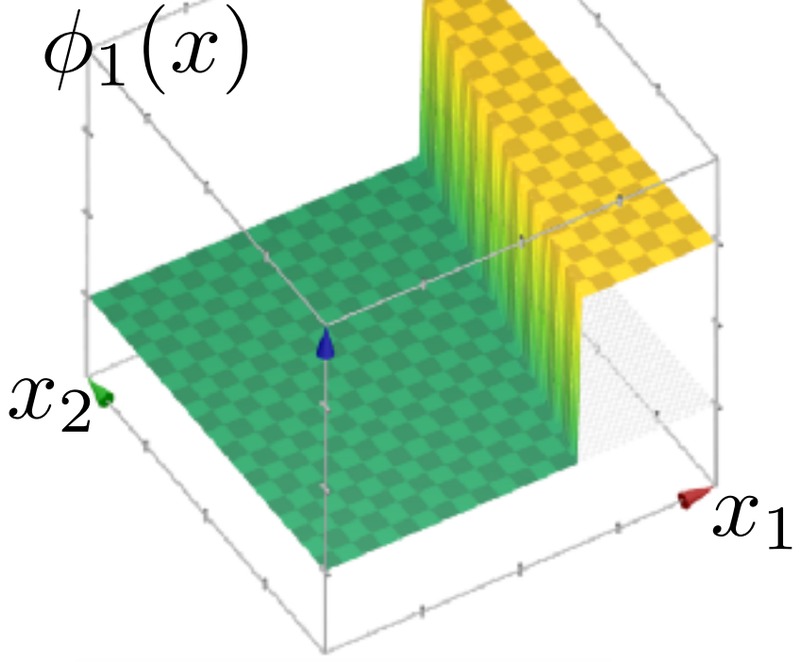

$$z = \theta^\top \phi(x) + \theta_0$$
$$= \theta_1 \phi_1(x) + \theta_2 \phi_2(x) + \theta_0$$
$$= 1 \cdot \phi_1(x) + 1 \cdot \phi_2(x) + (-0.5)$$



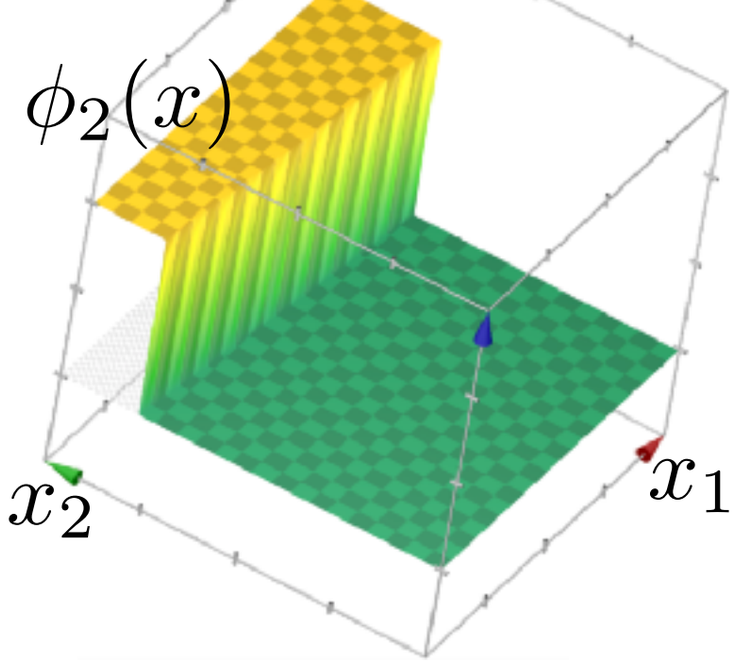



# New features: step functions!

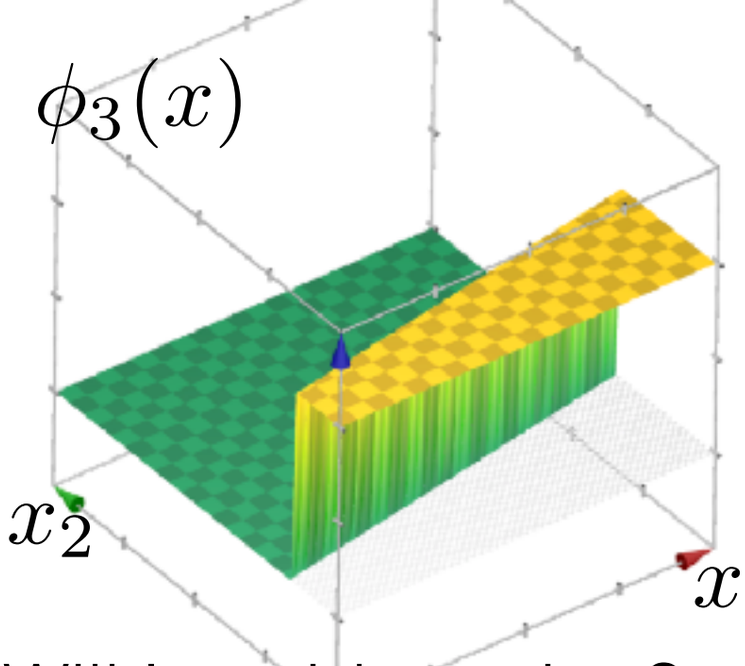
$$\phi_1(x) = \mathbf{1}\{w^\top x + w_0 \geq 0\}$$



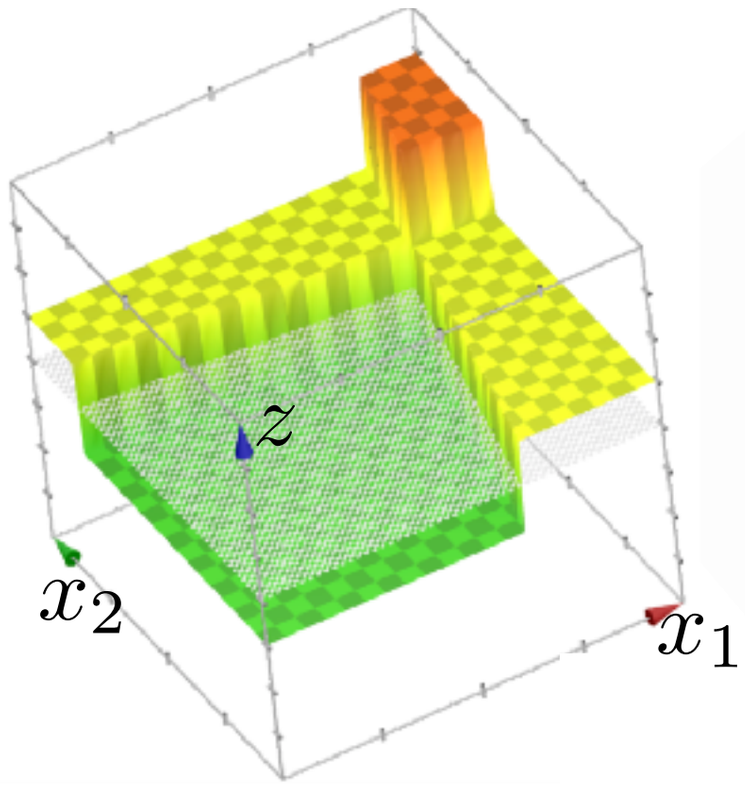

$$\phi_2(x) = \mathbf{1}\{\tilde{w}^\top x + \tilde{w}_0 \geq 0\}$$



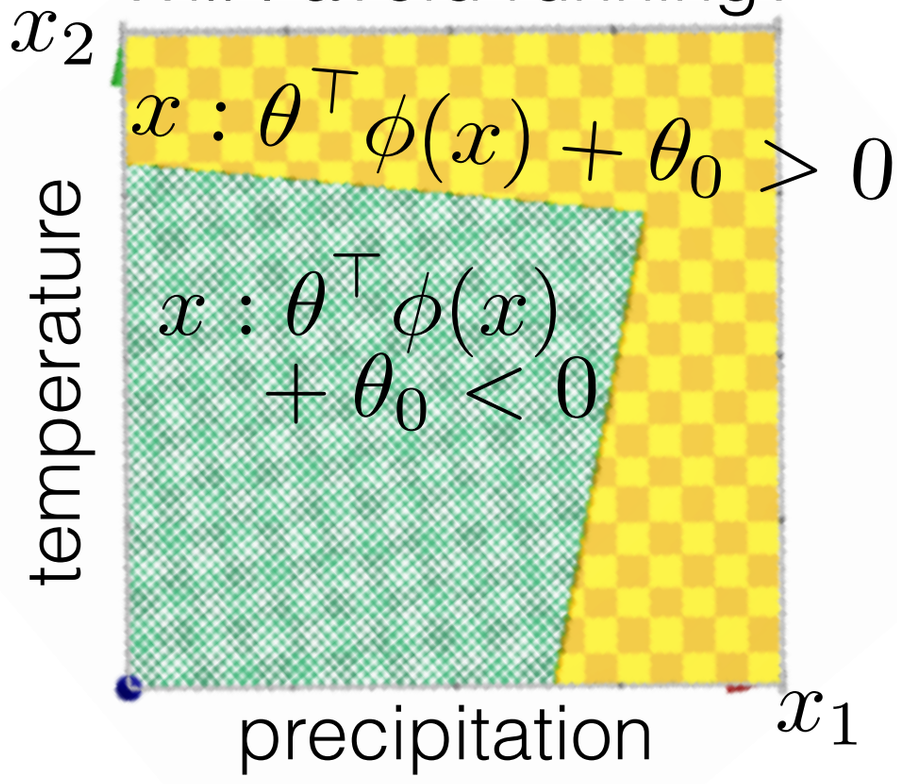

$$\phi_3(x) = \mathbf{1}\{\tilde{\tilde{w}}^\top x + \tilde{\tilde{w}}_0 \geq 0\}$$




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

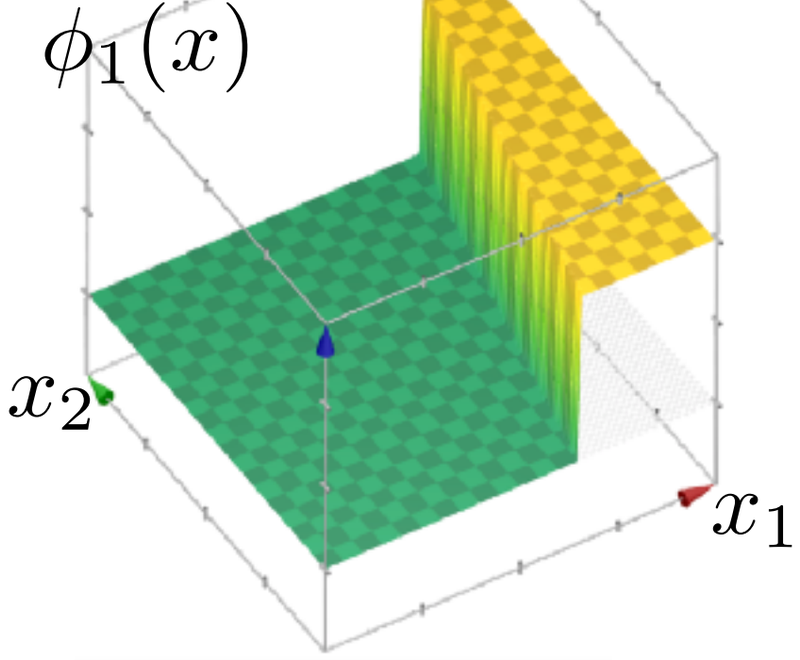



Will I avoid running?

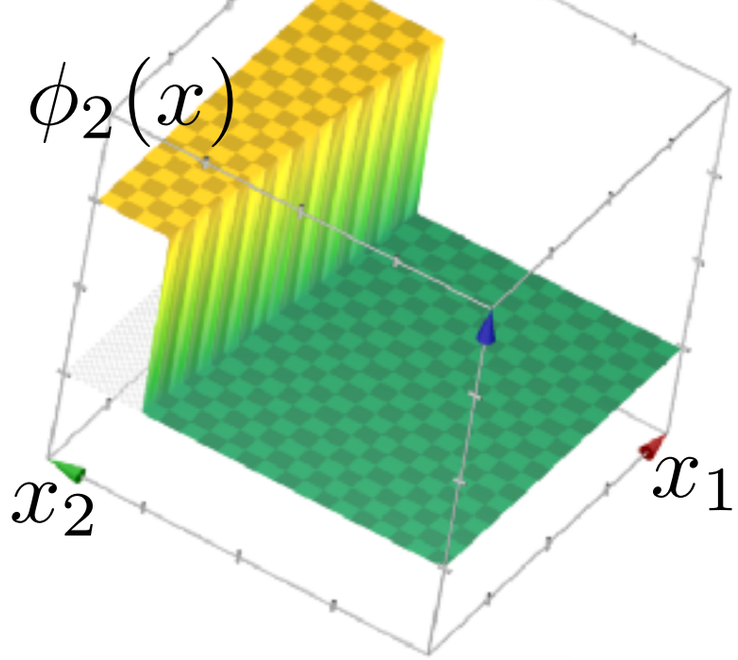


# New features: step functions!

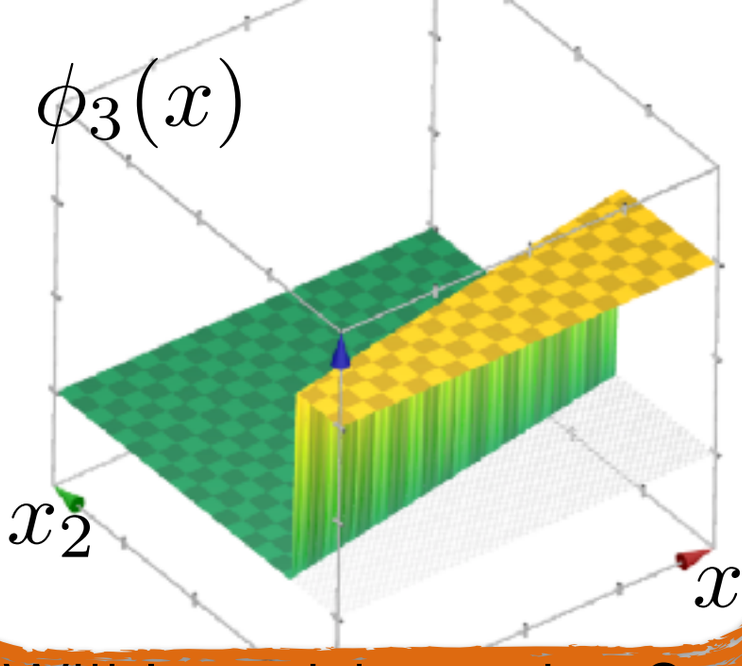
$$\phi_1(x) = \mathbf{1}\{w^\top x + w_0 \geq 0\}$$



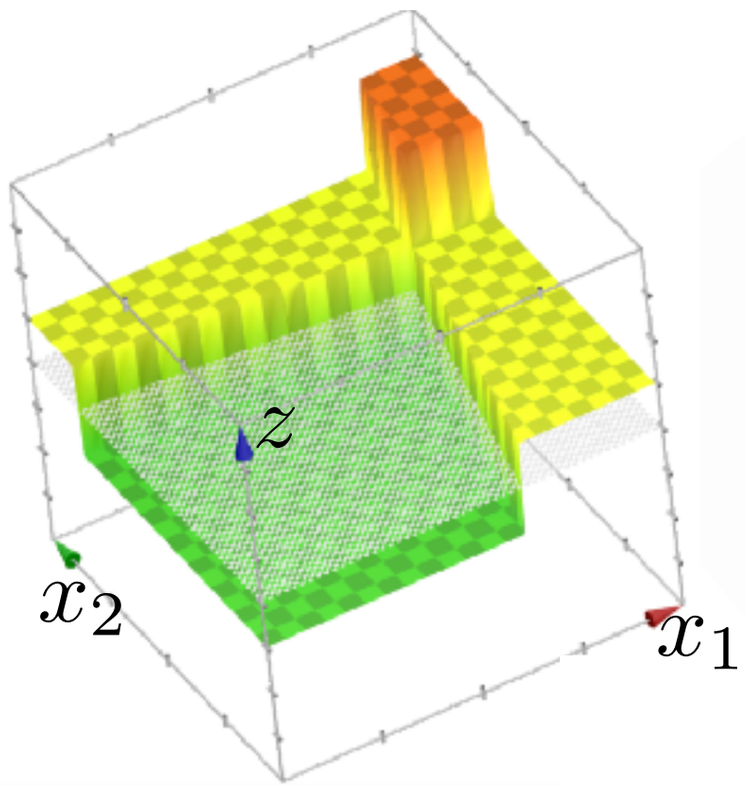

$$\phi_2(x) = \mathbf{1}\{\tilde{w}^\top x + \tilde{w}_0 \geq 0\}$$



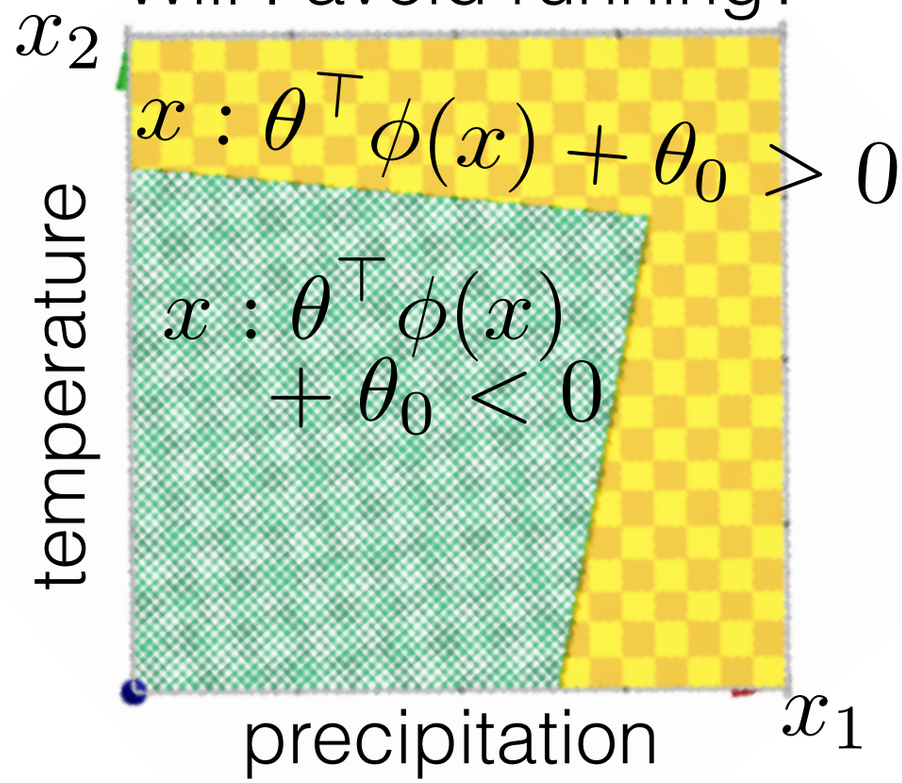

$$\phi_3(x) = \mathbf{1}\{\tilde{\tilde{w}}^\top x + \tilde{\tilde{w}}_0 \geq 0\}$$




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

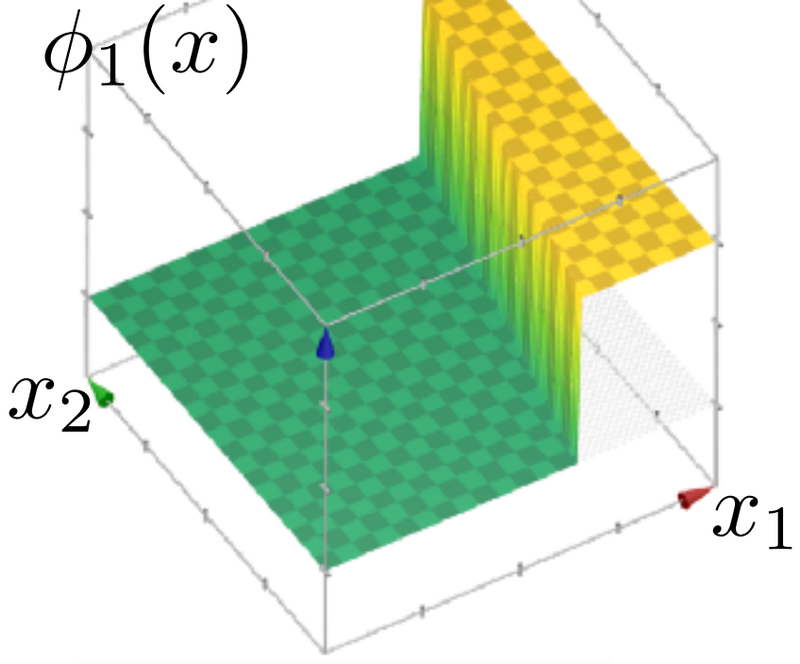



Will I avoid running?

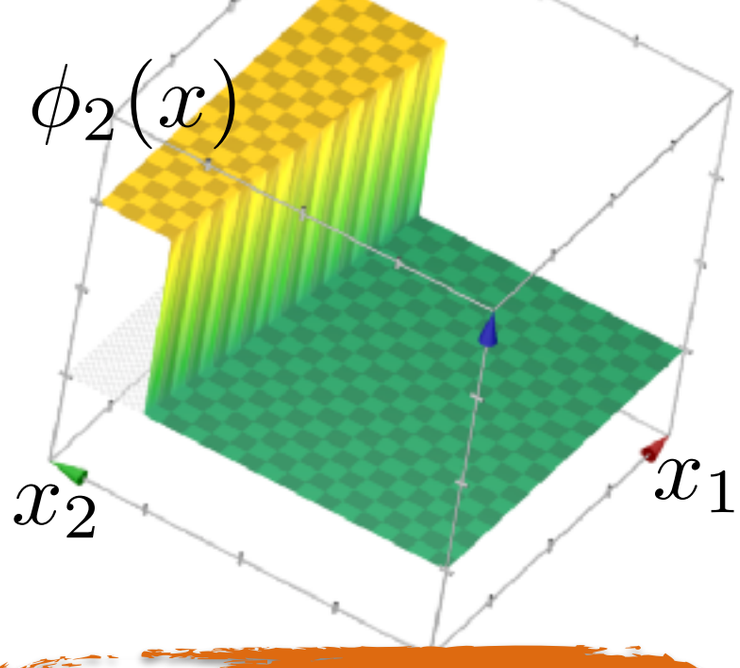


# New features: step functions!

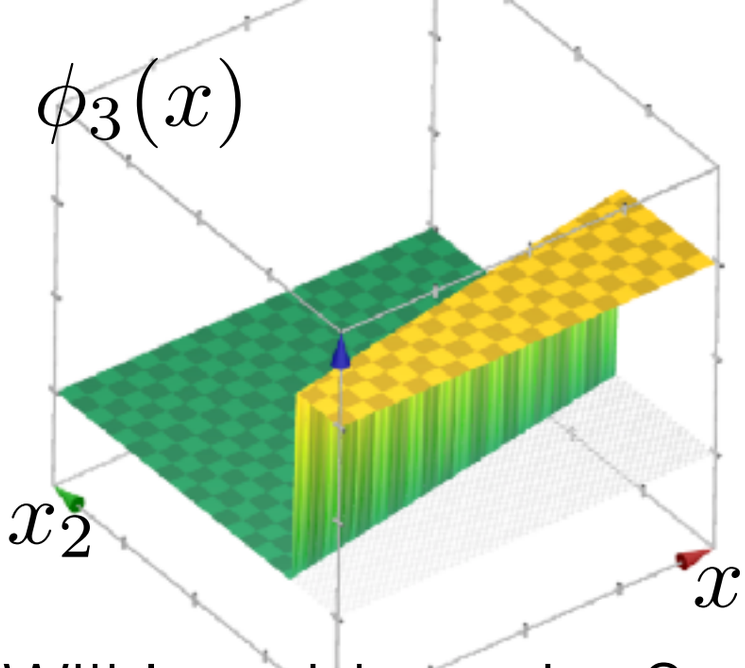
$$\phi_1(x) = \mathbf{1}\{w^\top x + w_0 \geq 0\}$$



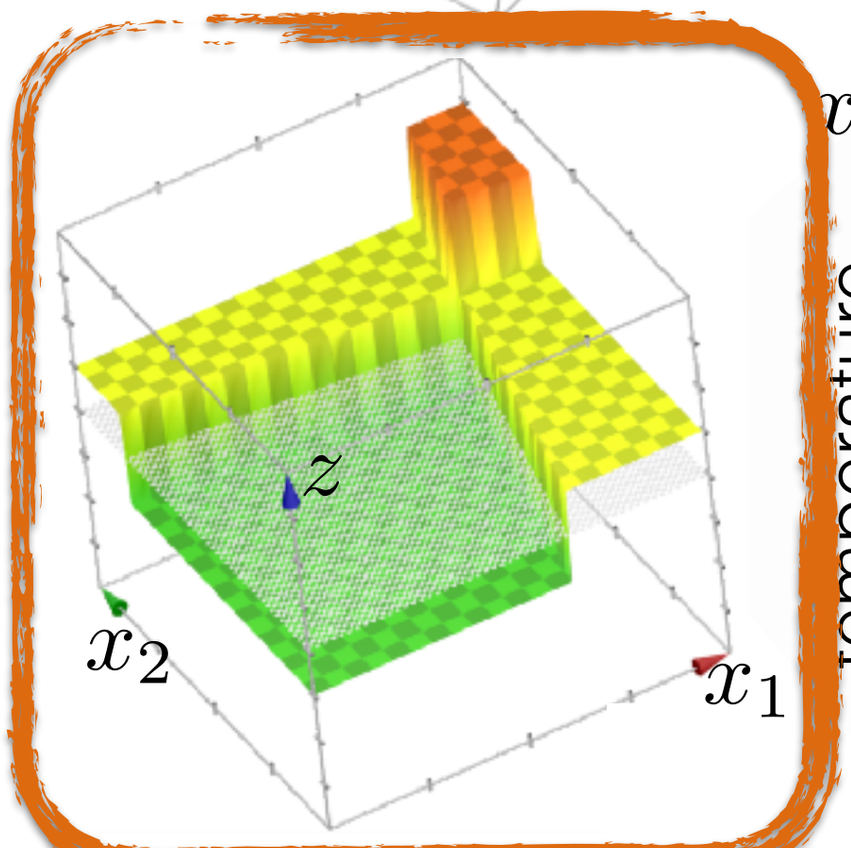

$$\phi_2(x) = \mathbf{1}\{\tilde{w}^\top x + \tilde{w}_0 \geq 0\}$$



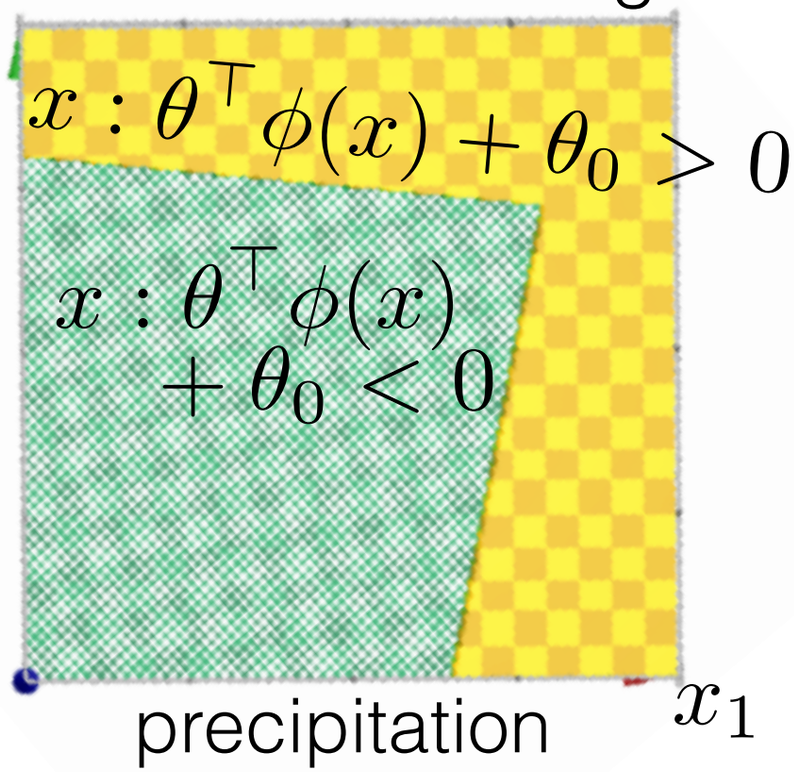

$$\phi_3(x) = \mathbf{1}\{\tilde{\tilde{w}}^\top x + \tilde{\tilde{w}}_0 \geq 0\}$$




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

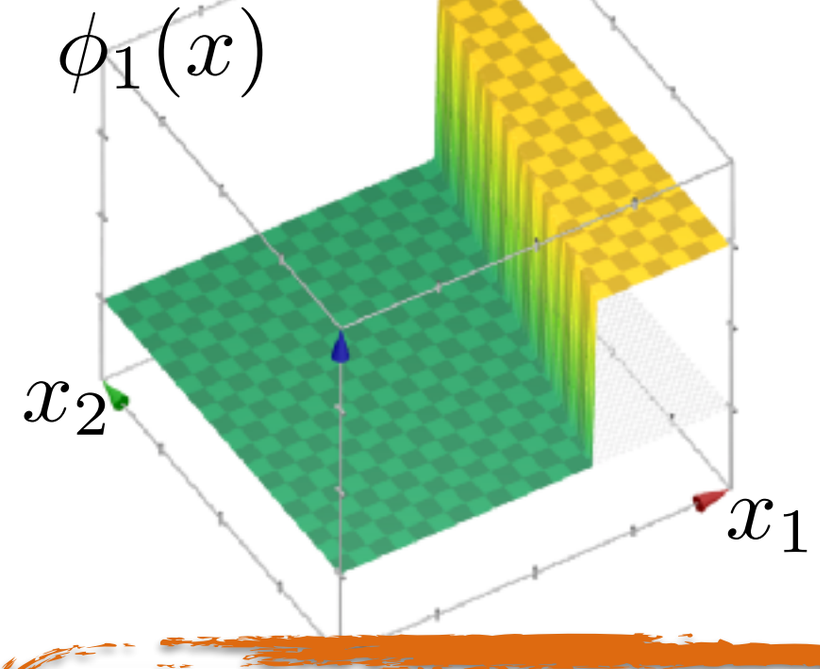



Will I avoid running?

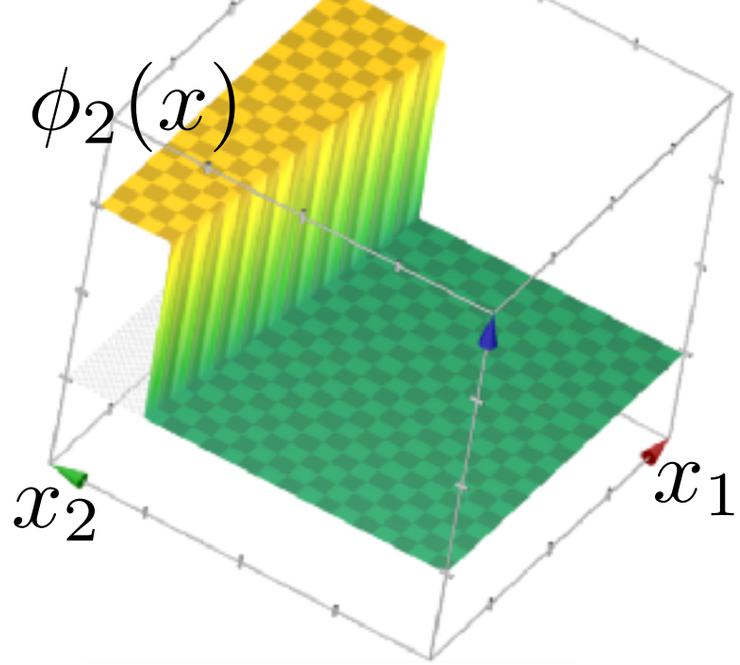


# New features: step functions!

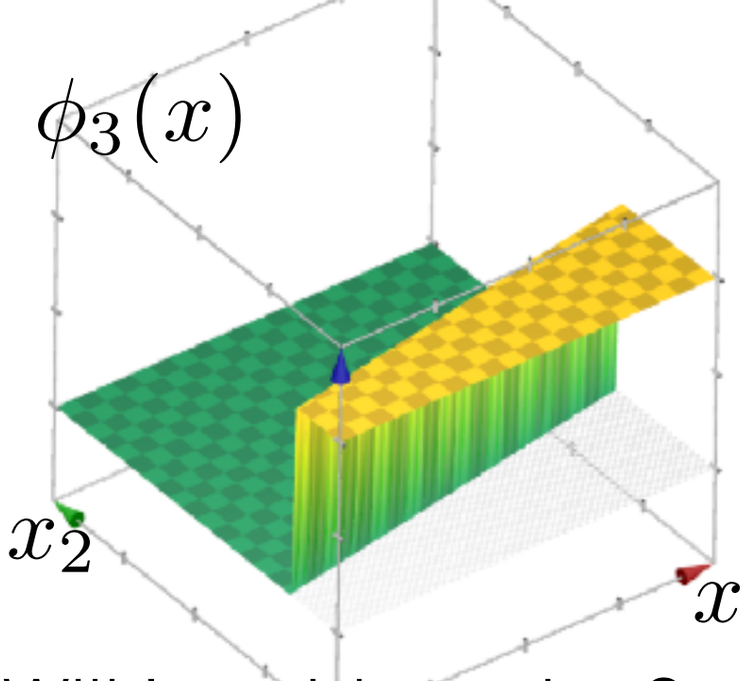
$$\phi_1(x) = \mathbf{1}\{w^\top x + w_0 \geq 0\}$$



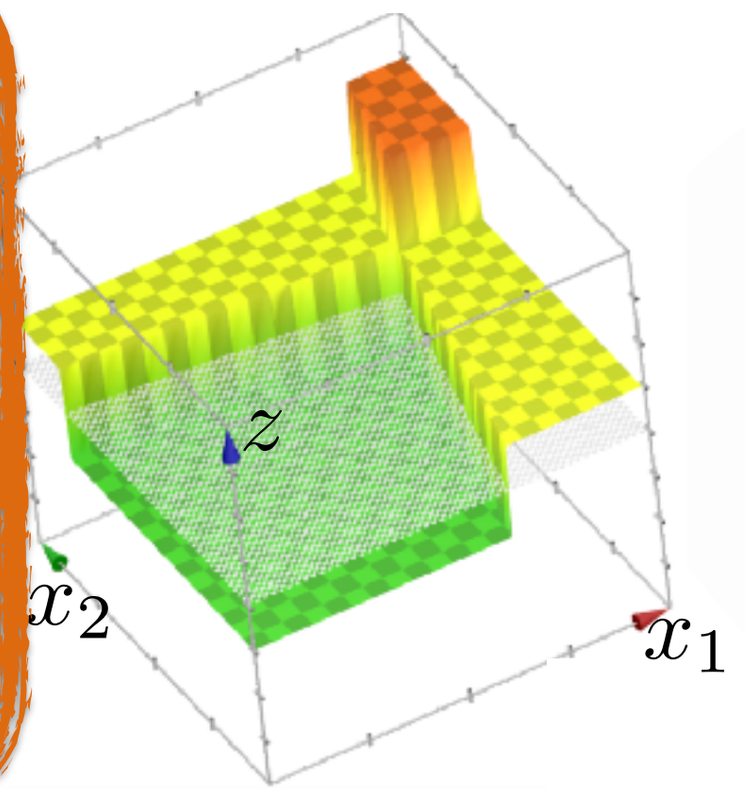

$$\phi_2(x) = \mathbf{1}\{\tilde{w}^\top x + \tilde{w}_0 \geq 0\}$$



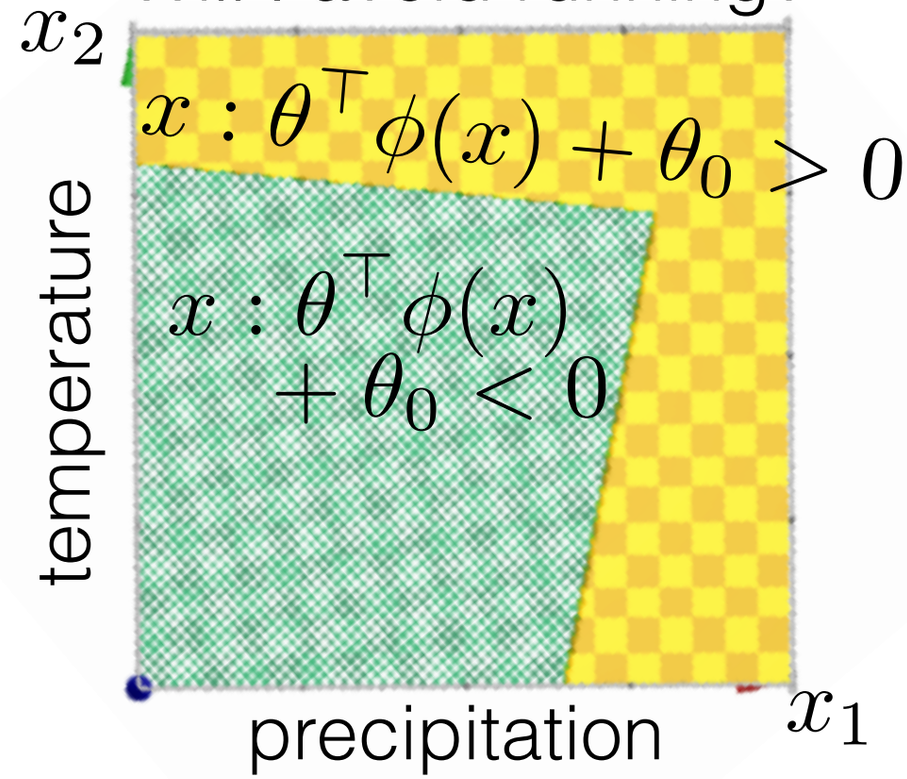

$$\phi_3(x) = \mathbf{1}\{\tilde{\tilde{w}}^\top x + \tilde{\tilde{w}}_0 \geq 0\}$$




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

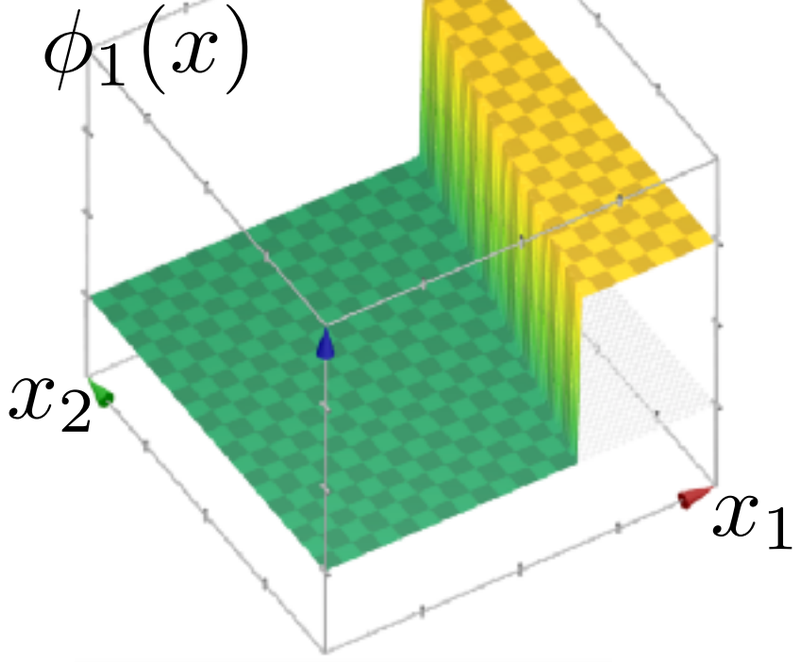



Will I avoid running?

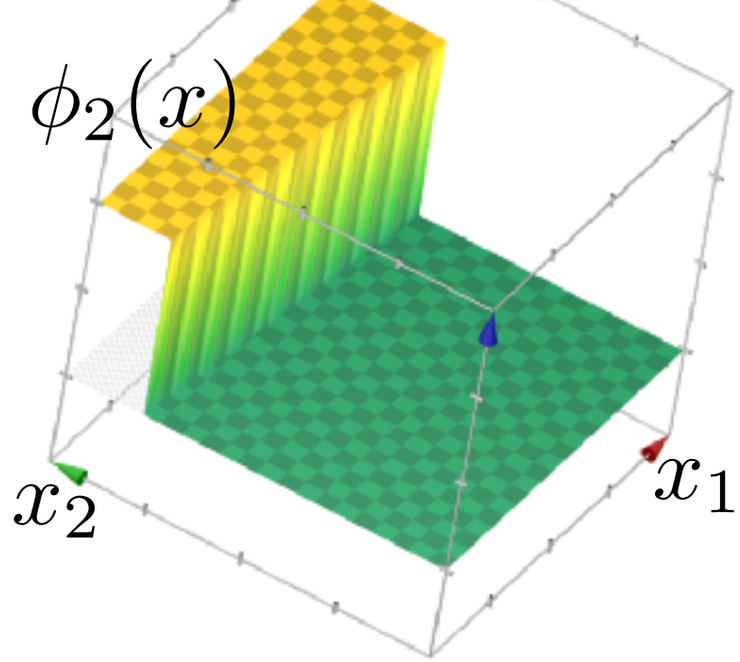


# New features: step functions!

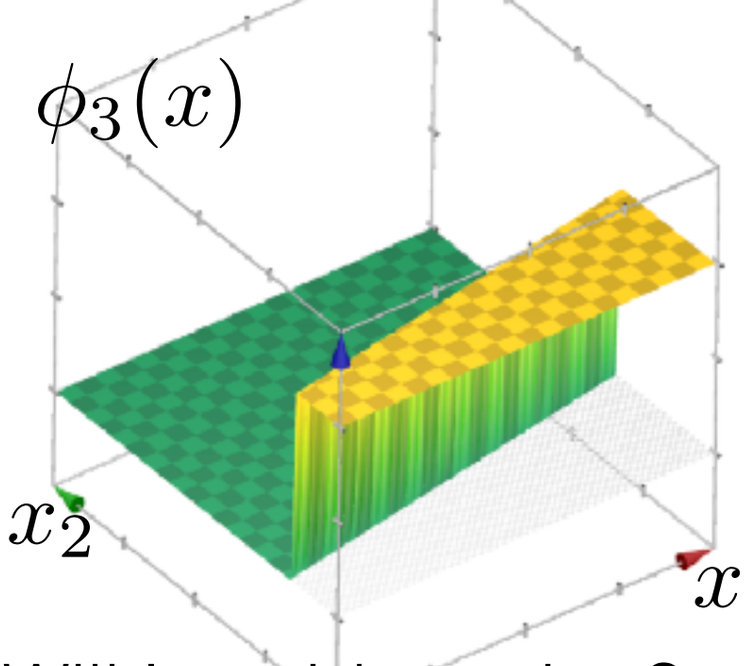
$$\phi_1(x) = \mathbf{1}\{w^\top x + w_0 \geq 0\}$$




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



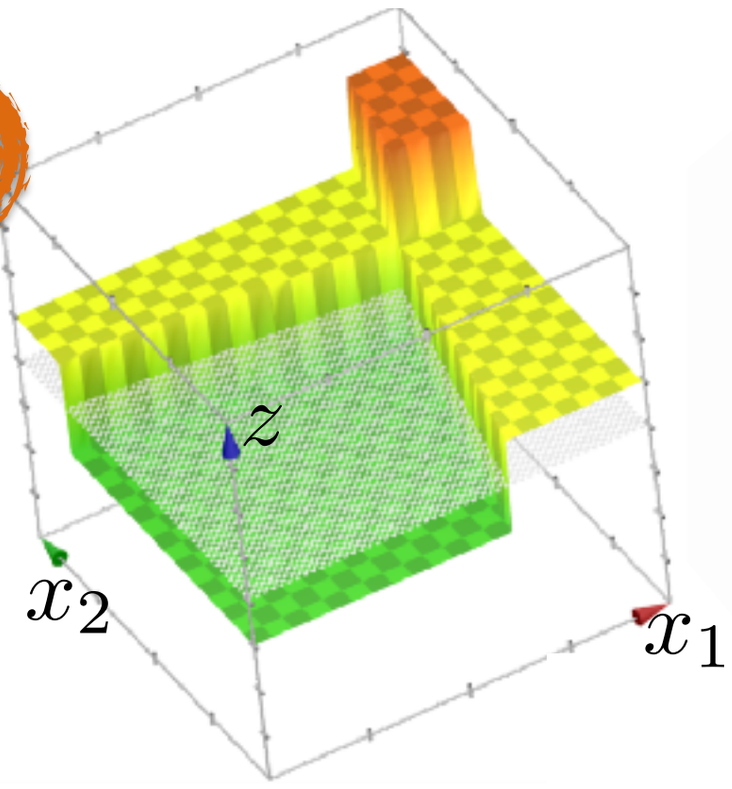

$$\phi_3(x) = \mathbf{1}\{\tilde{\tilde{w}}^\top x + \tilde{\tilde{w}}_0 \geq 0\}$$



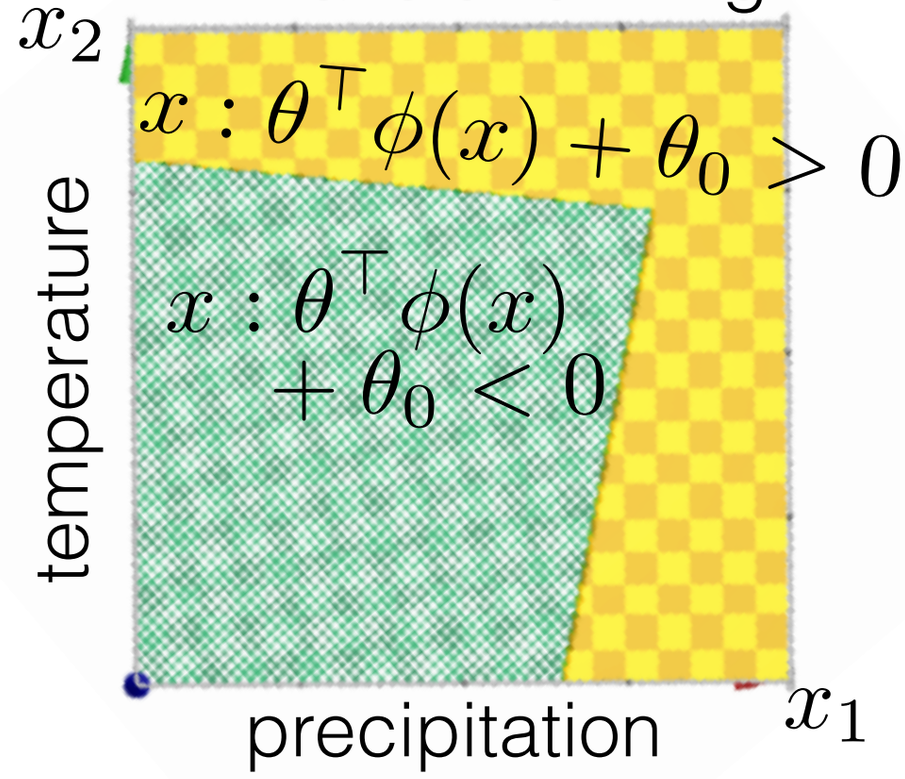

$$z = \theta^\top \phi(x) + \theta_0$$

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

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

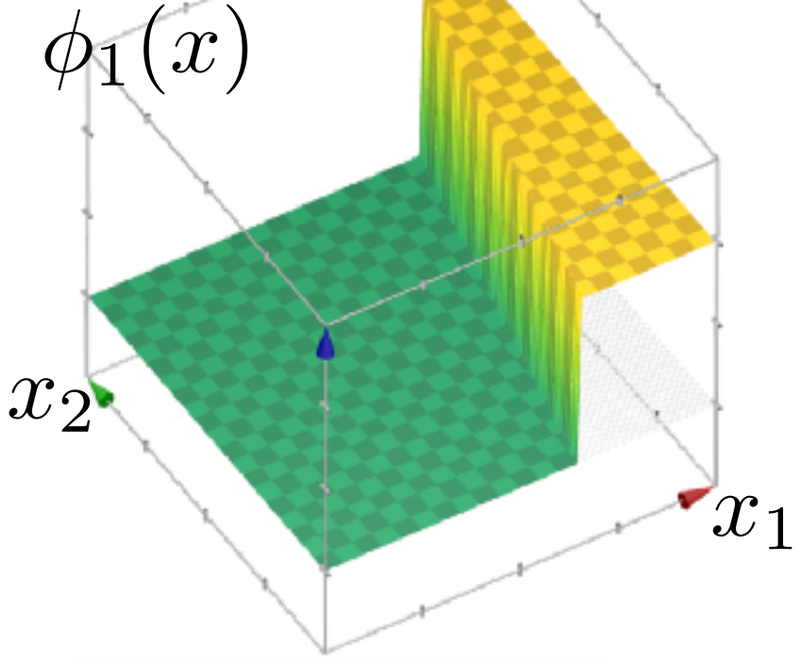



Will I avoid running?

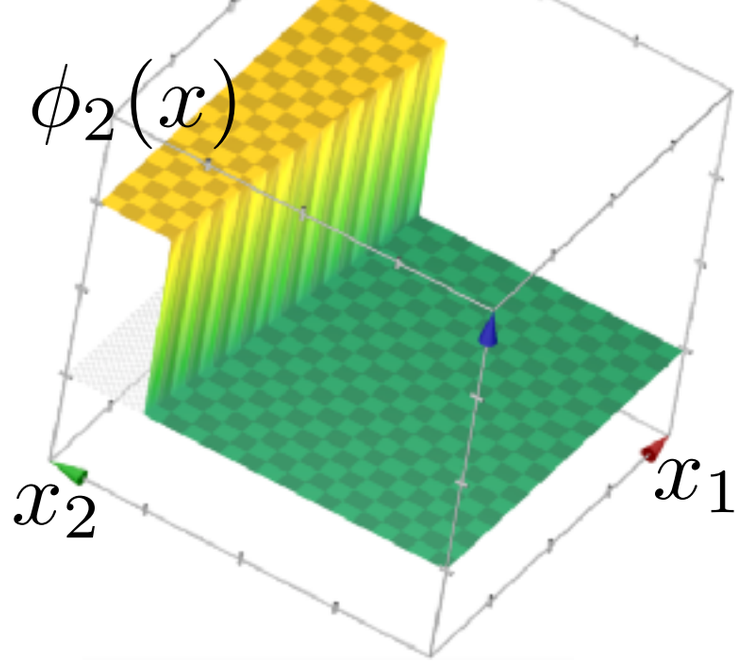


# New features: step functions!

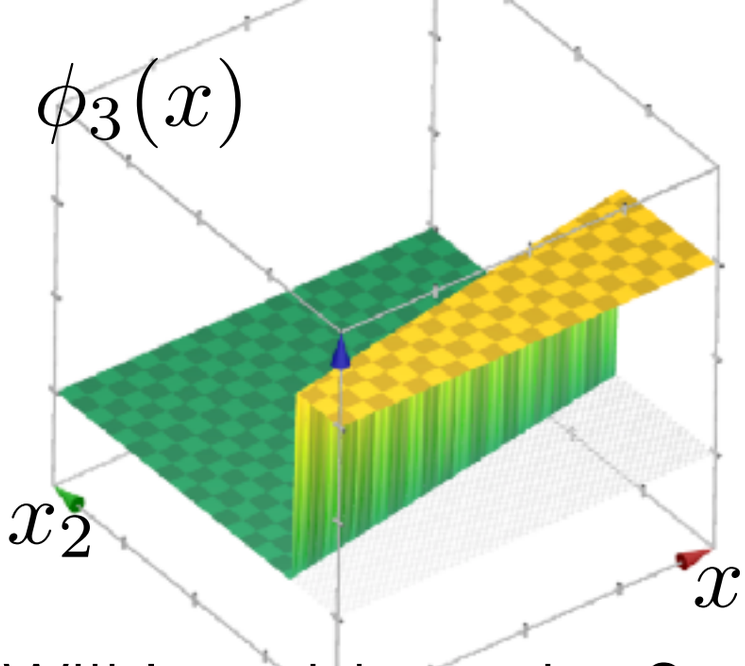
$$\phi_1(x) = \mathbf{1}\{w^\top x + w_0 \geq 0\}$$



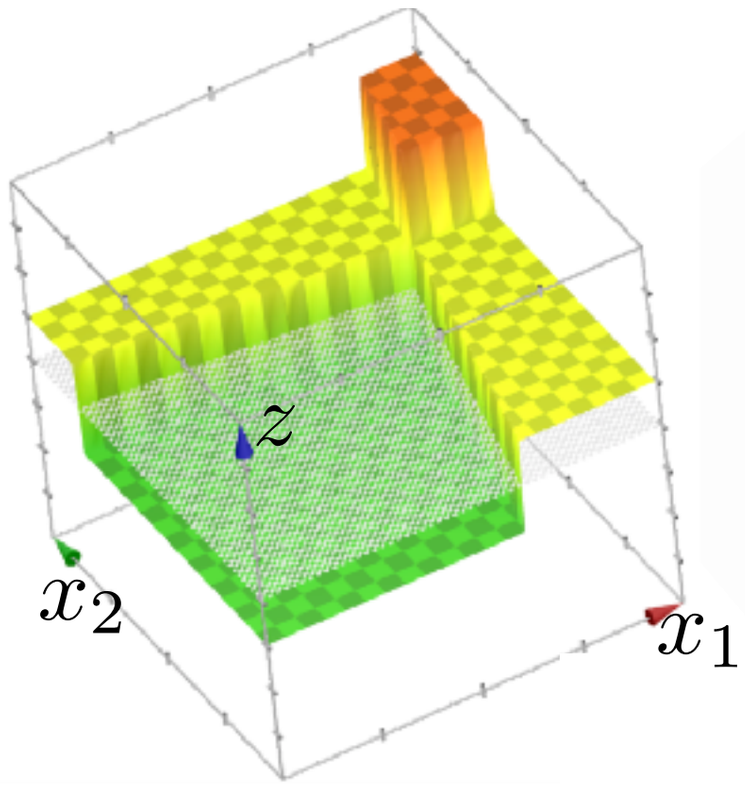

$$\phi_2(x) = \mathbf{1}\{\tilde{w}^\top x + \tilde{w}_0 \geq 0\}$$



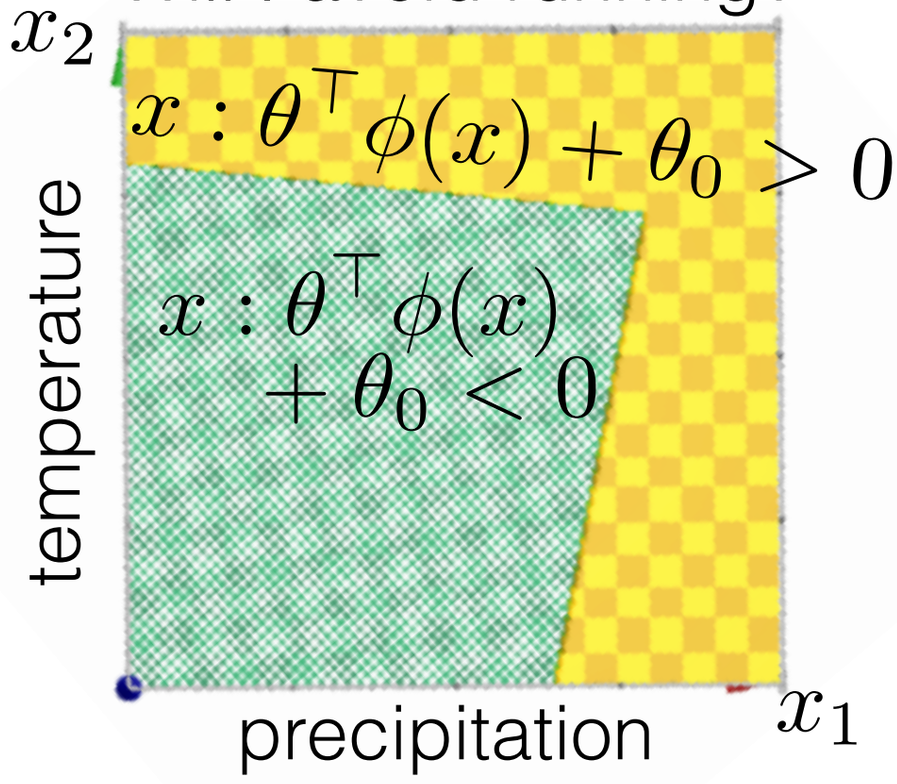

$$\phi_3(x) = \mathbf{1}\{\tilde{\tilde{w}}^\top x + \tilde{\tilde{w}}_0 \geq 0\}$$




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

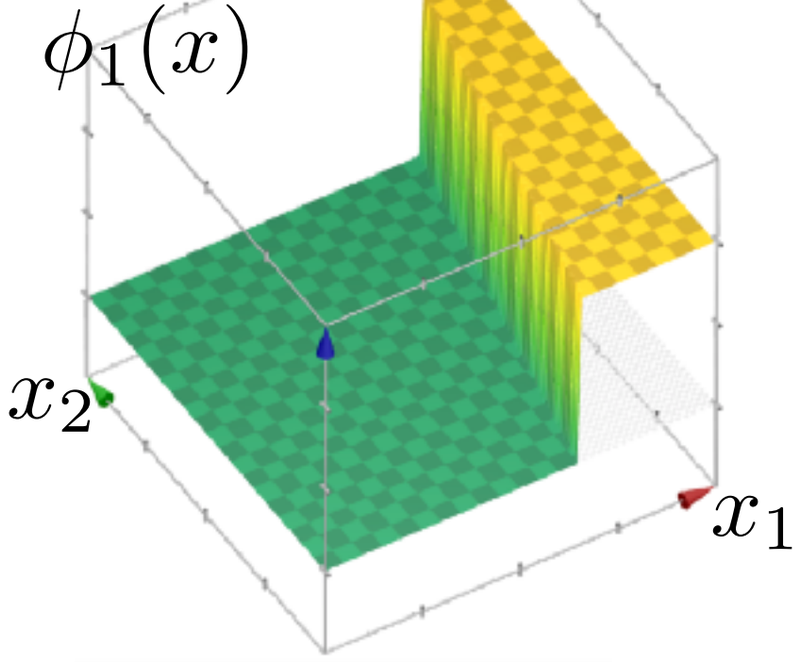



Will I avoid running?

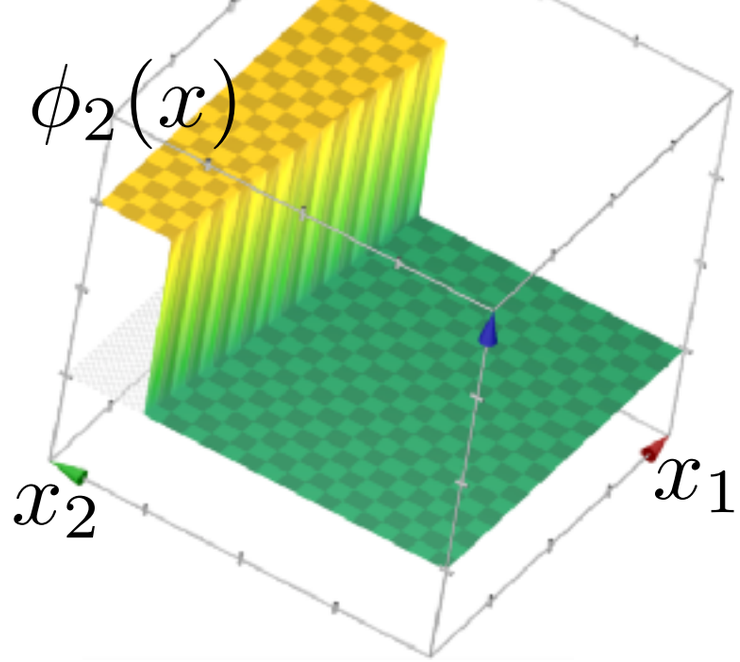


# New features: step functions!

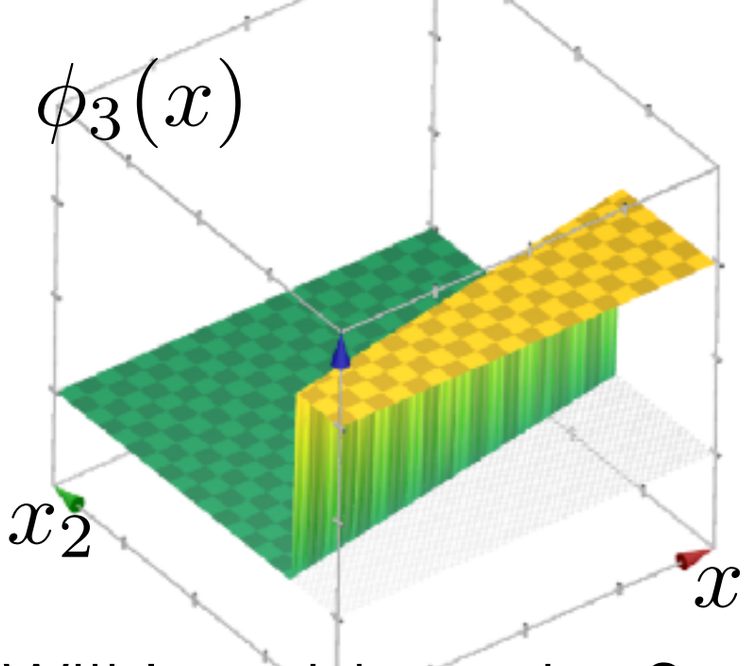
$$\phi_1(x) = \mathbf{1}\{w^\top x + w_0 \geq 0\}$$



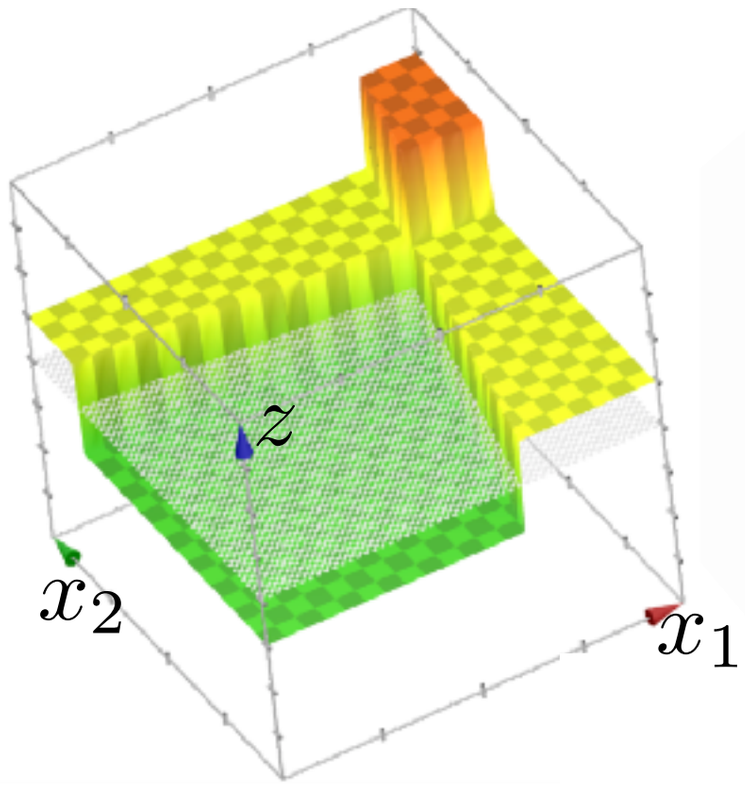

$$\phi_2(x) = \mathbf{1}\{\tilde{w}^\top x + \tilde{w}_0 \geq 0\}$$



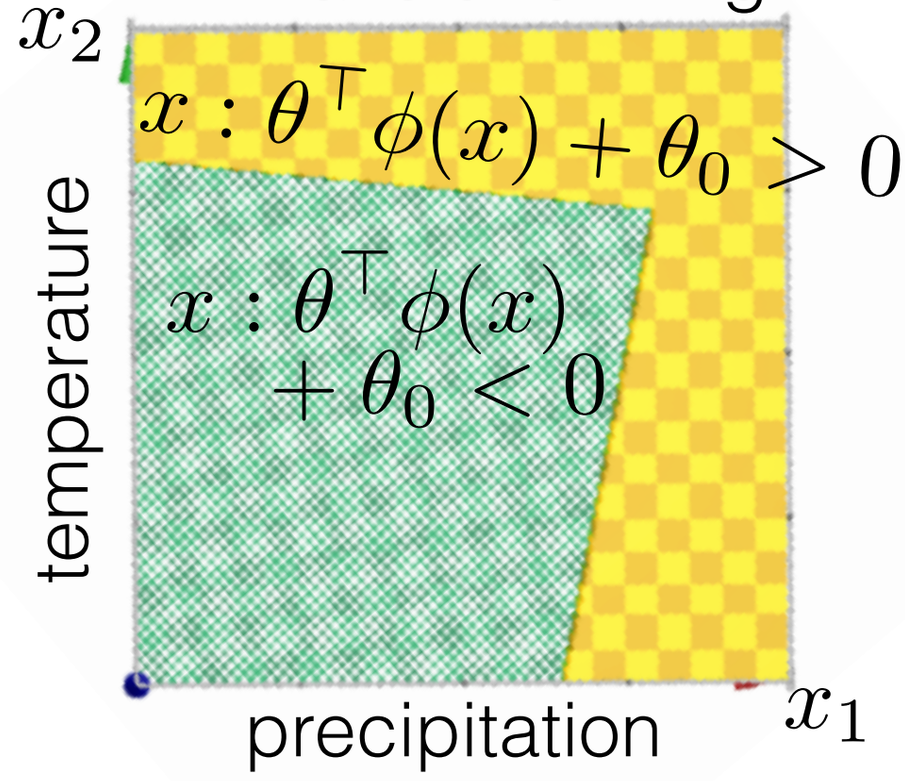

$$\phi_3(x) = \mathbf{1}\{\tilde{\tilde{w}}^\top x + \tilde{\tilde{w}}_0 \geq 0\}$$




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

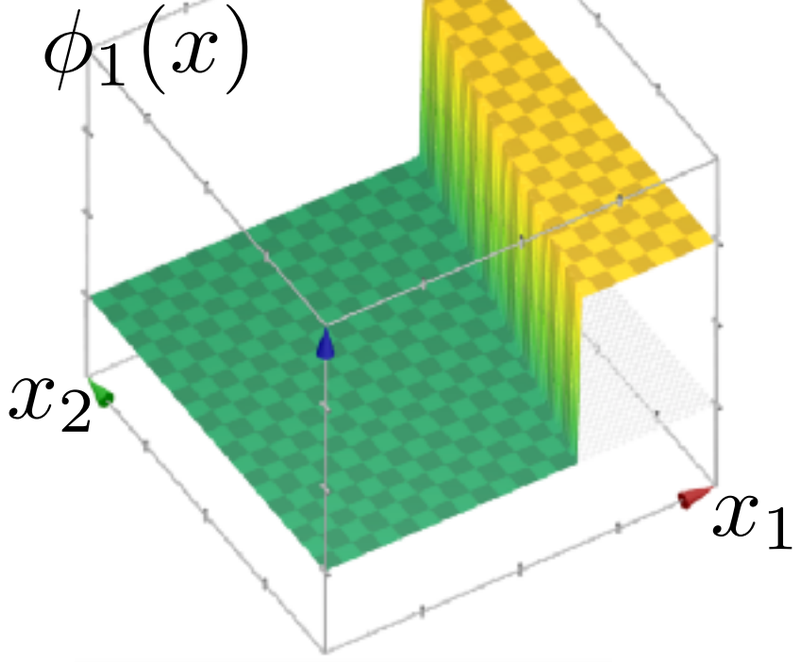



Will I avoid running?

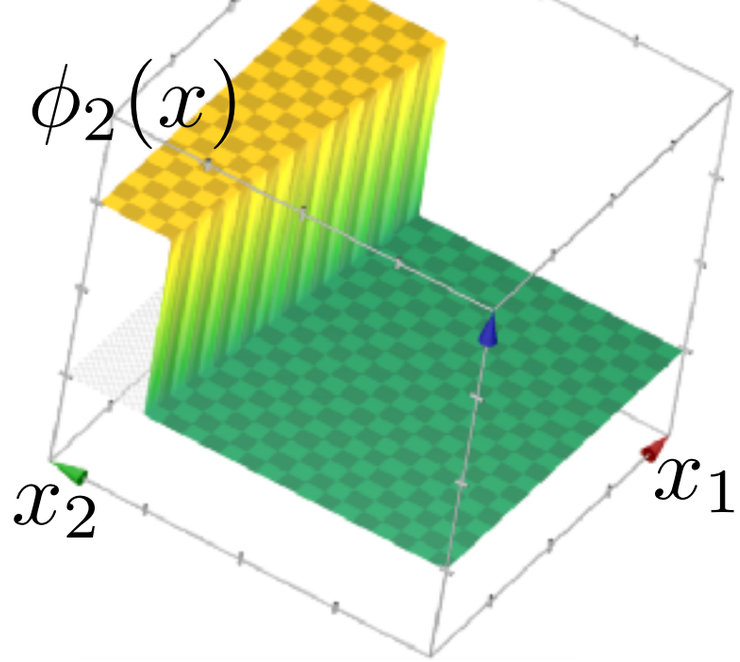


# New features: step functions!

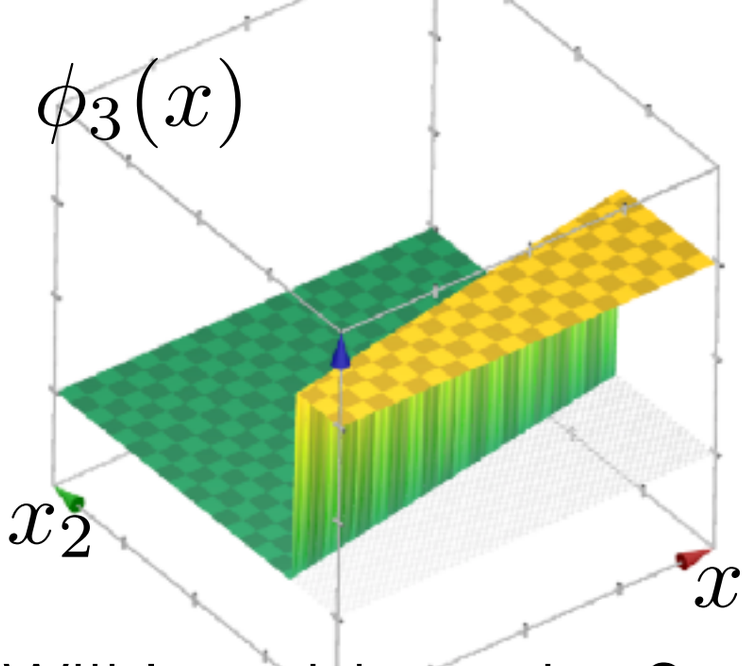
$$\phi_1(x) = \mathbf{1}\{w^\top x + w_0 \geq 0\}$$



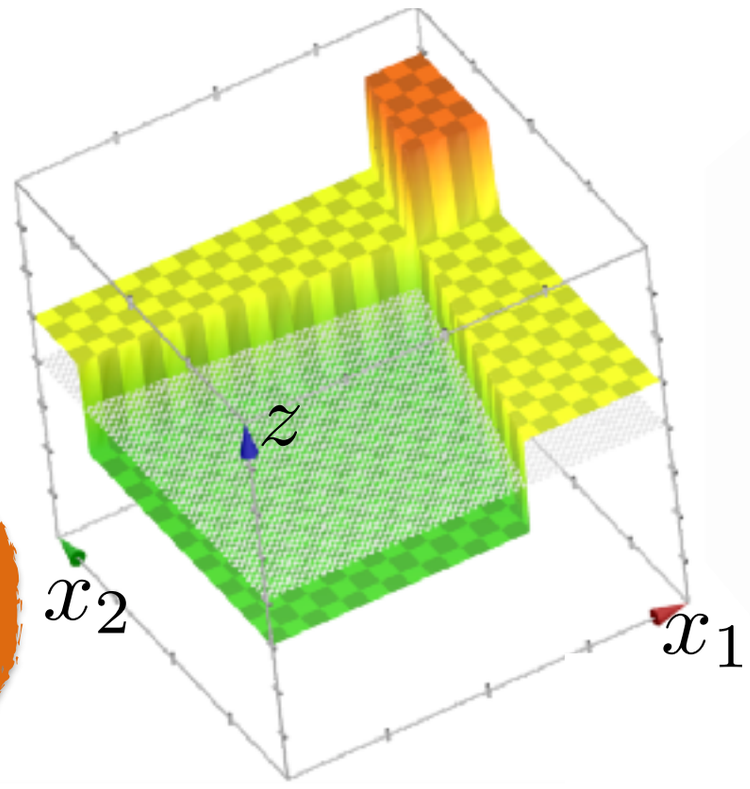

$$\phi_2(x) = \mathbf{1}\{\tilde{w}^\top x + \tilde{w}_0 \geq 0\}$$



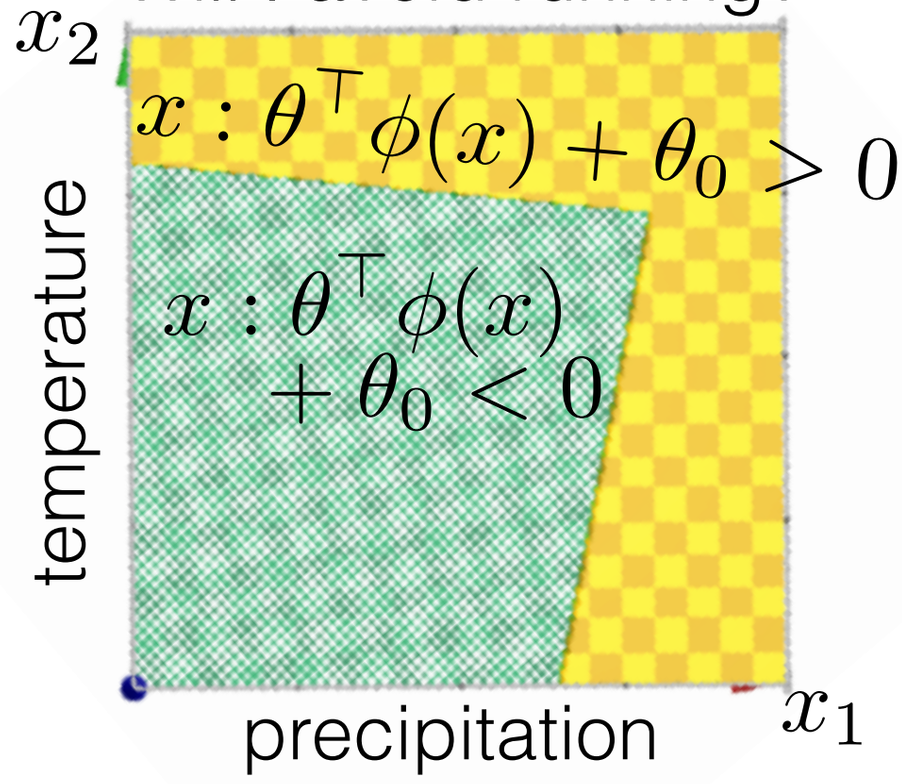

$$\phi_3(x) = \mathbf{1}\{\tilde{\tilde{w}}^\top x + \tilde{\tilde{w}}_0 \geq 0\}$$




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

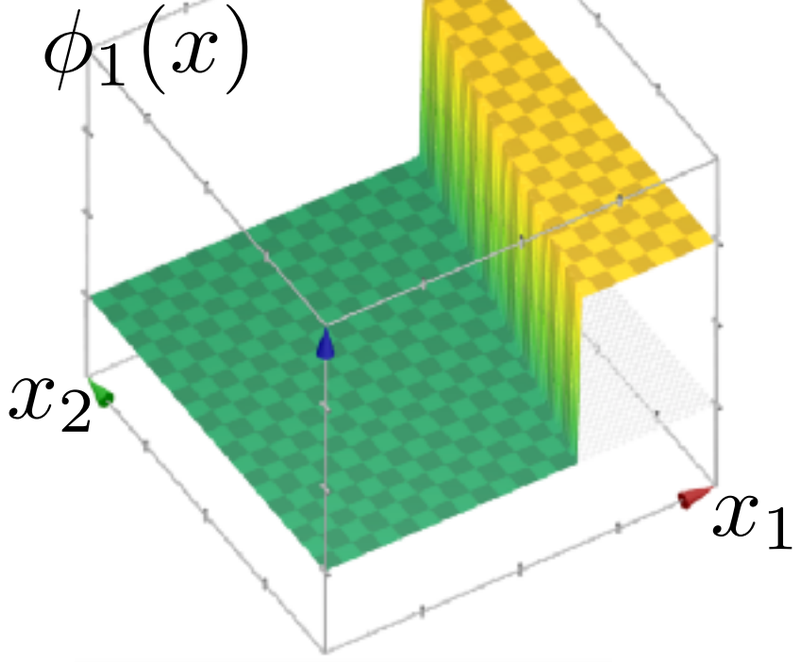



Will I avoid running?

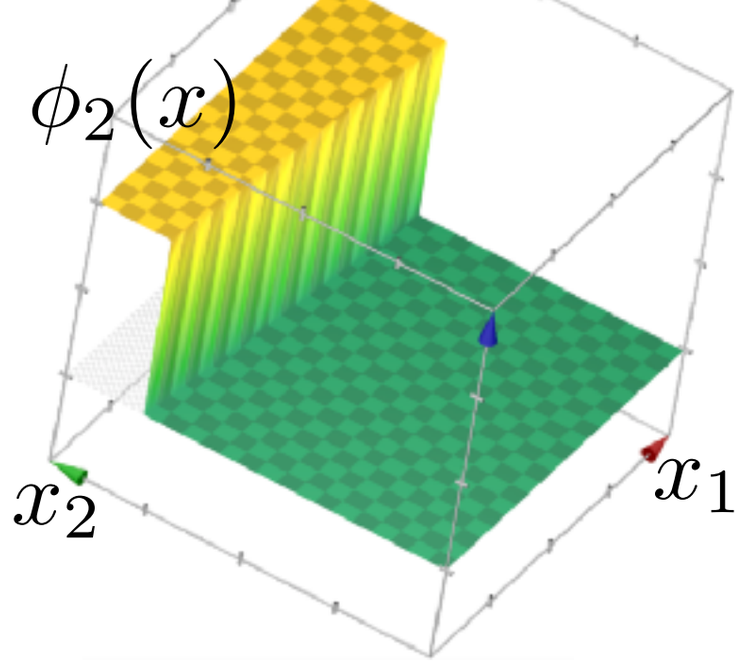


# New features: step functions!

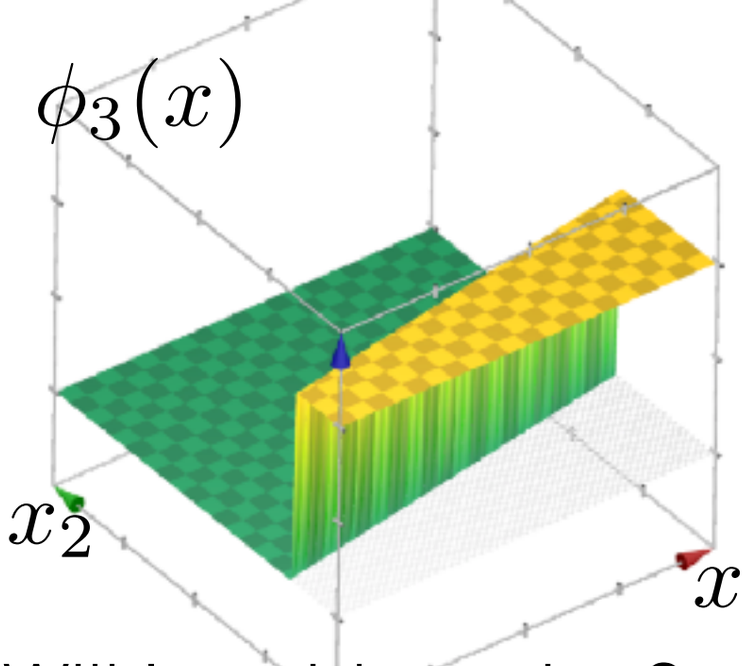
$$\phi_1(x) = \mathbf{1}\{w^\top x + w_0 \geq 0\}$$



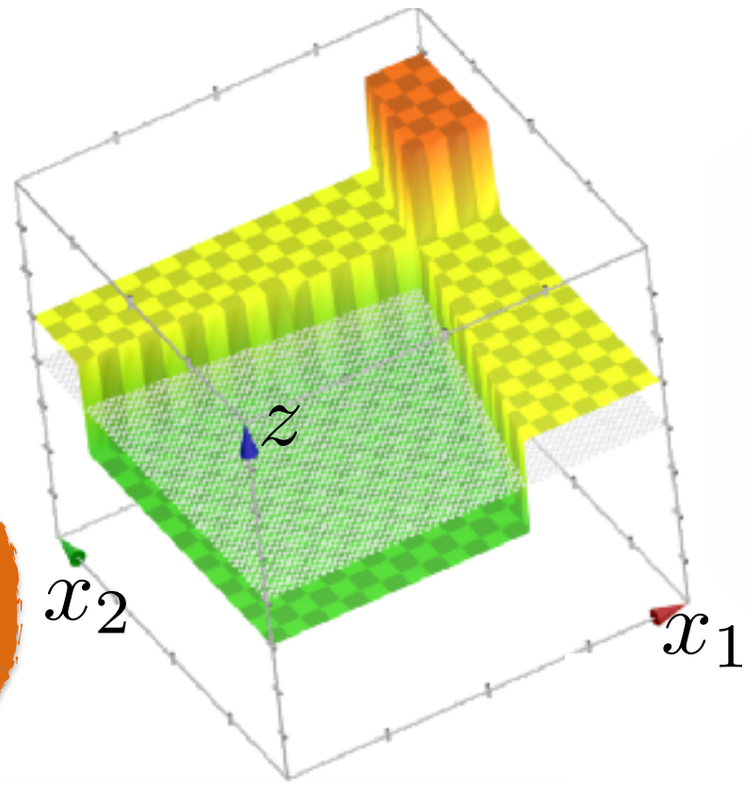

$$\phi_2(x) = \mathbf{1}\{\tilde{w}^\top x + \tilde{w}_0 \geq 0\}$$



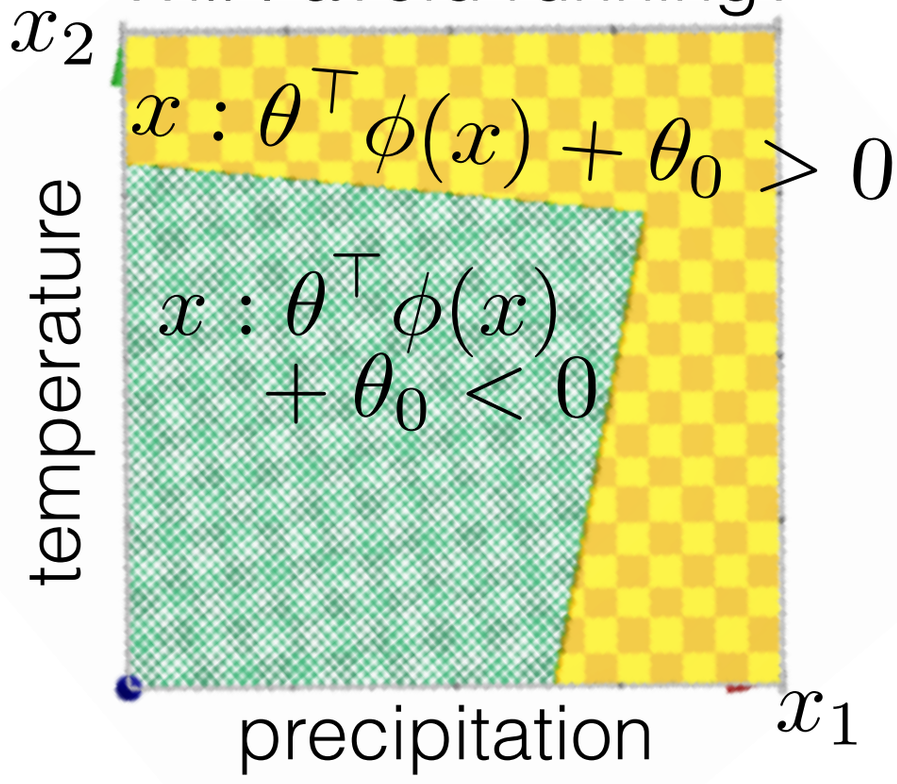

$$\phi_3(x) = \mathbf{1}\{\tilde{\tilde{w}}^\top x + \tilde{\tilde{w}}_0 \geq 0\}$$




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

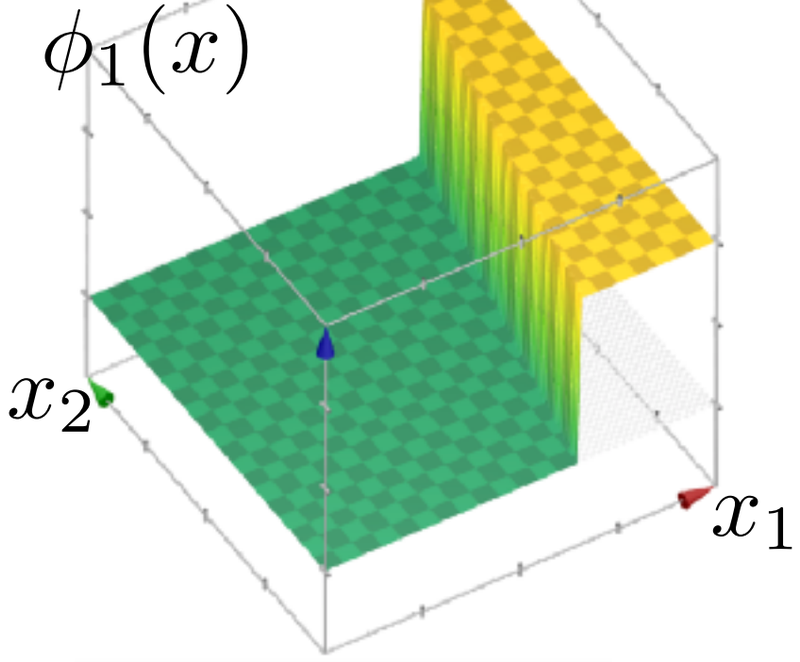



Will I avoid running?

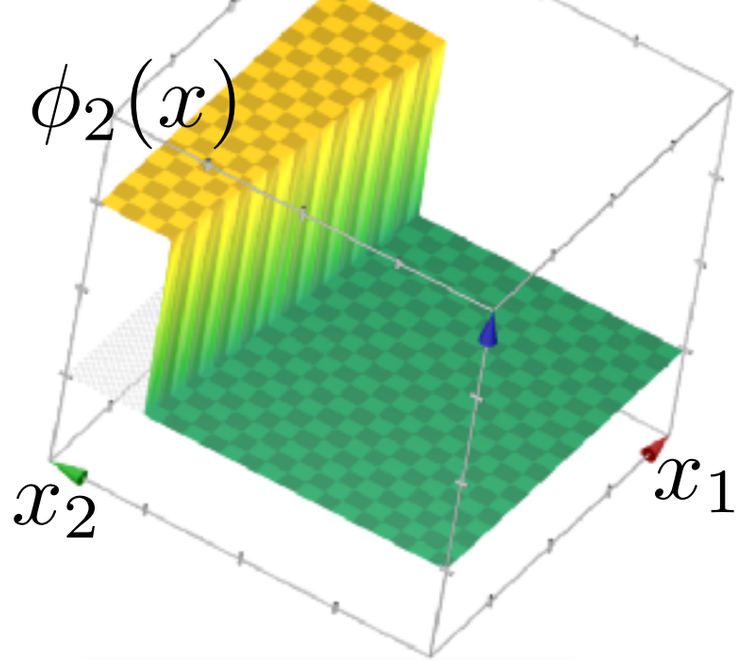


# New features: step functions!

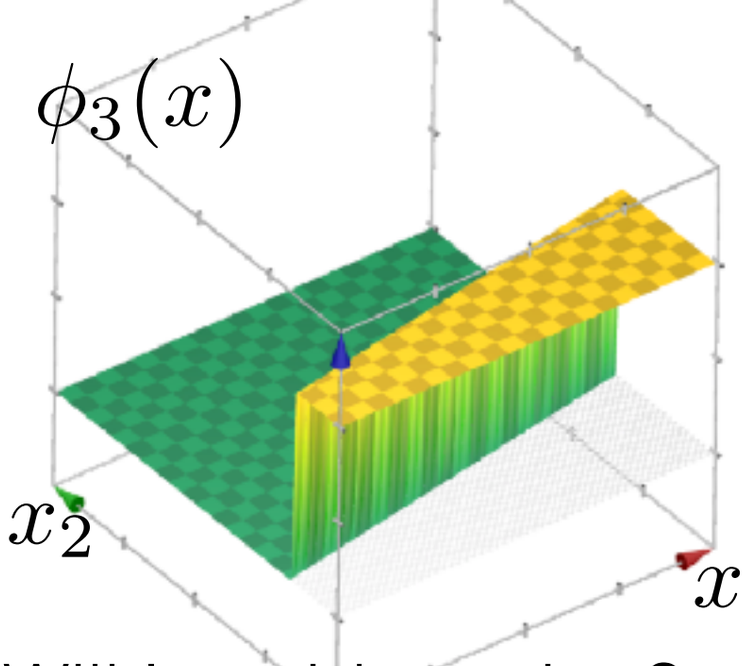
$$\phi_1(x) = \mathbf{1}\{w^\top x + w_0 \geq 0\}$$



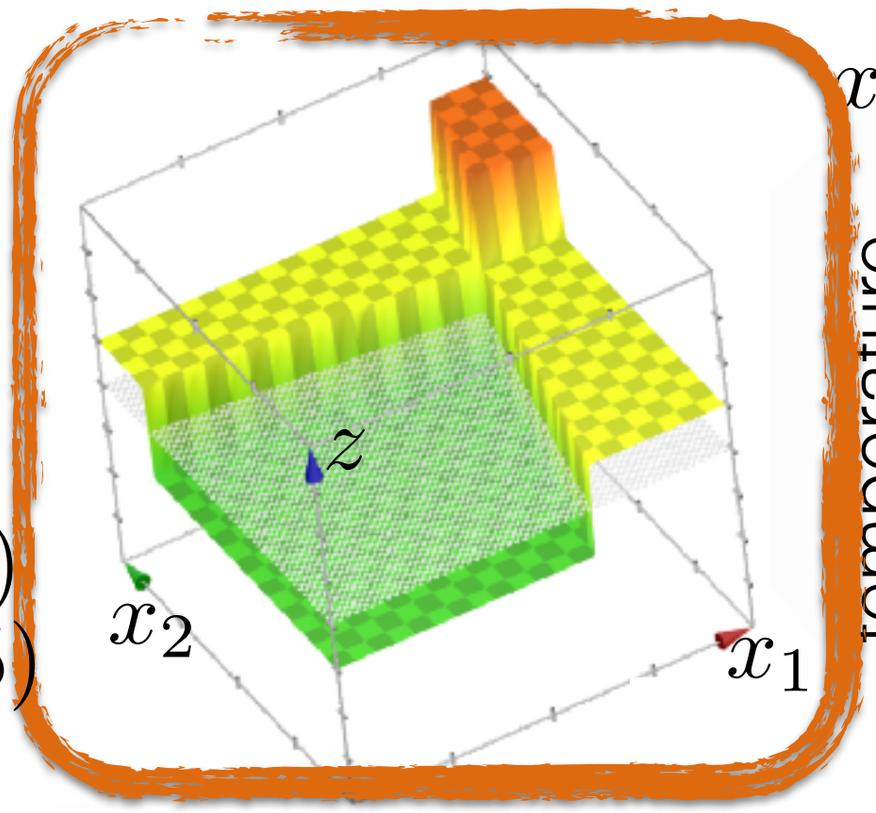

$$\phi_2(x) = \mathbf{1}\{\tilde{w}^\top x + \tilde{w}_0 \geq 0\}$$



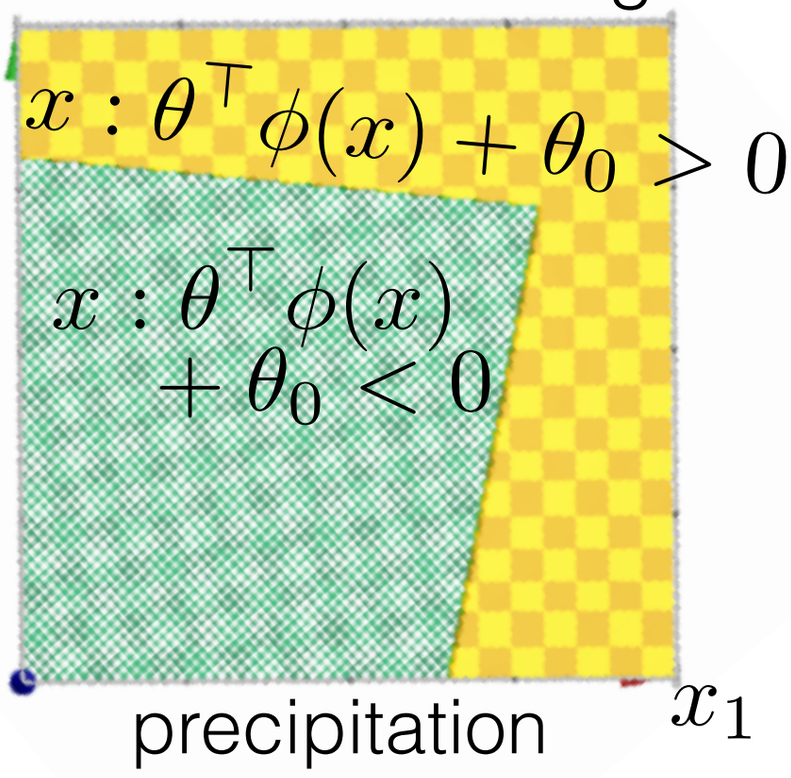

$$\phi_3(x) = \mathbf{1}\{\tilde{\tilde{w}}^\top x + \tilde{\tilde{w}}_0 \geq 0\}$$




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

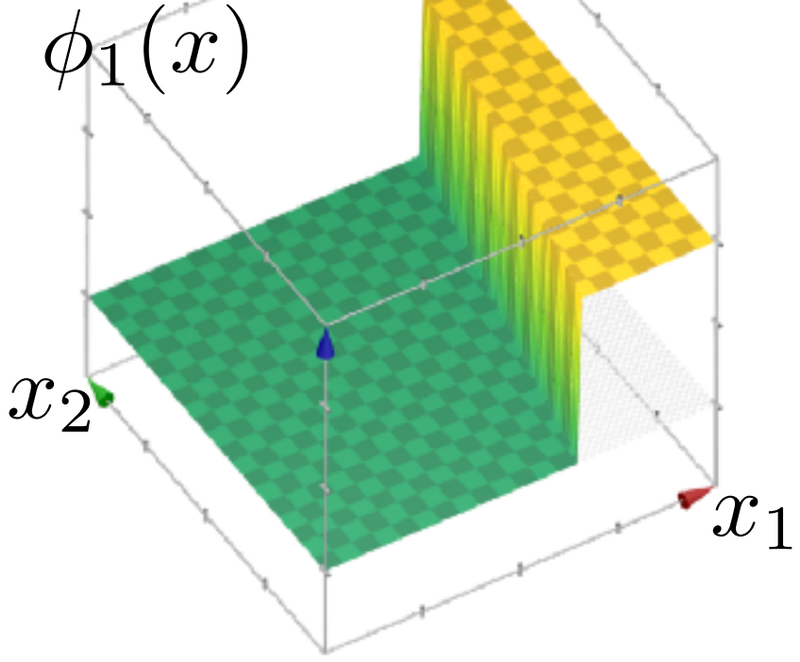



Will I avoid running?

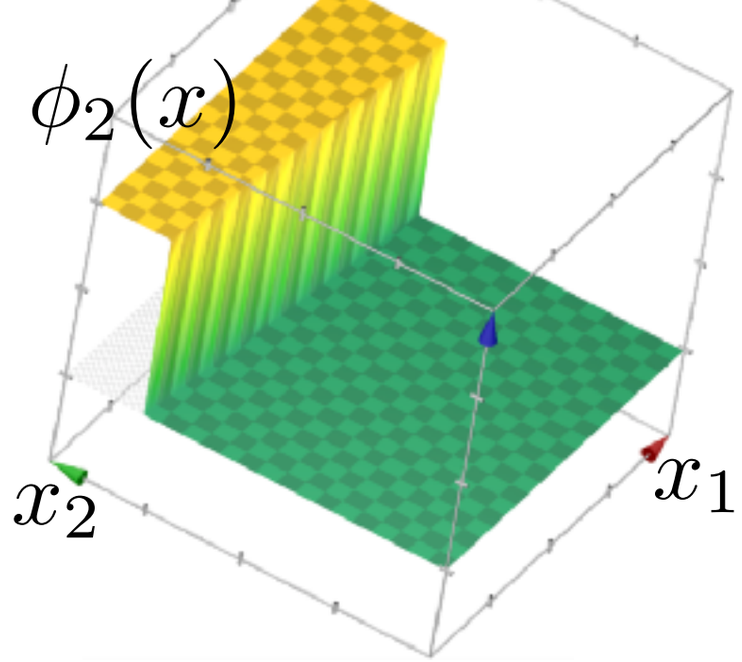


# New features: step functions!

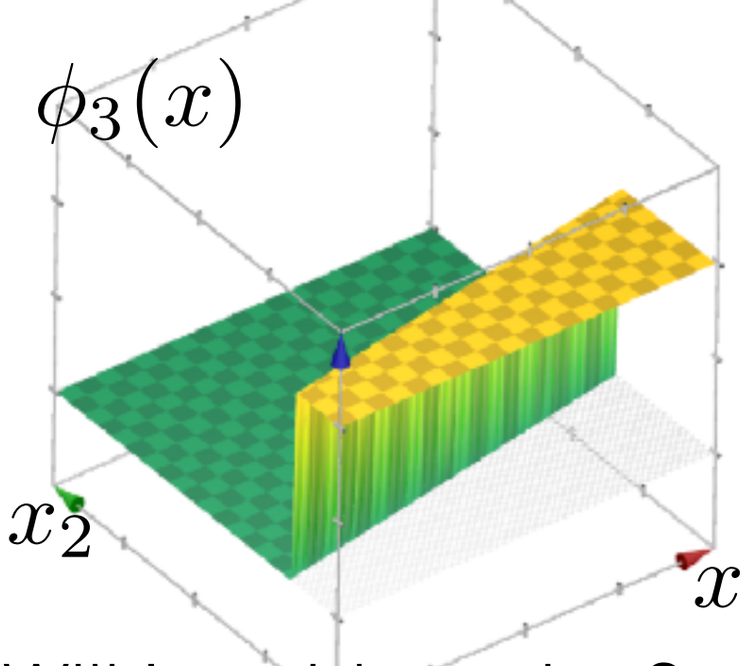
$$\phi_1(x) = \mathbf{1}\{w^\top x + w_0 \geq 0\}$$



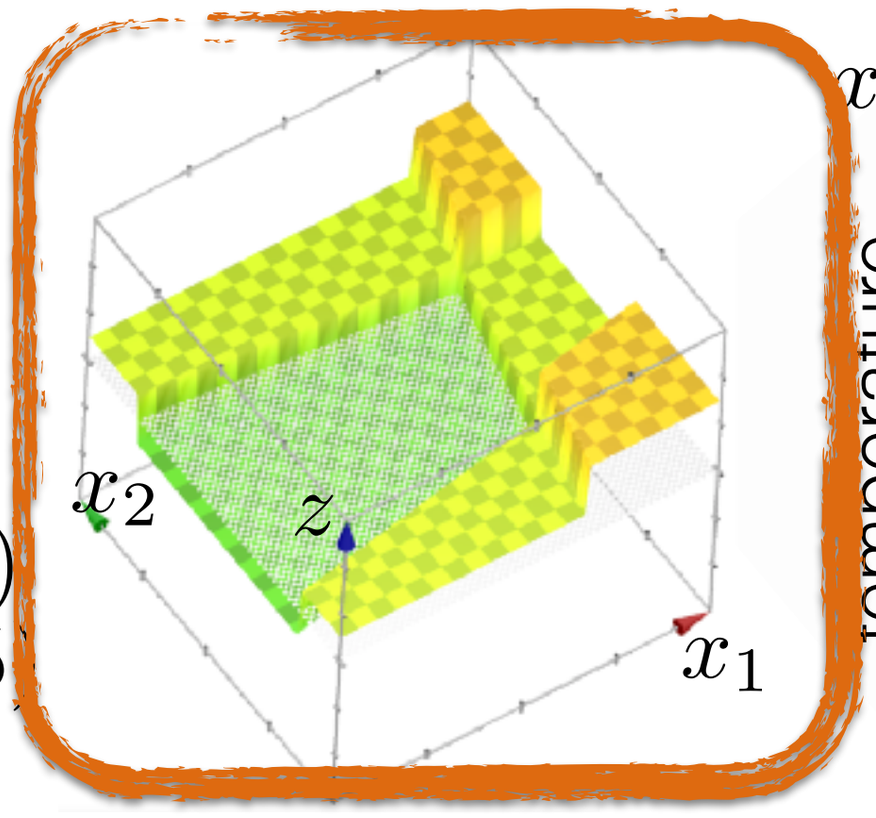

$$\phi_2(x) = \mathbf{1}\{\tilde{w}^\top x + \tilde{w}_0 \geq 0\}$$



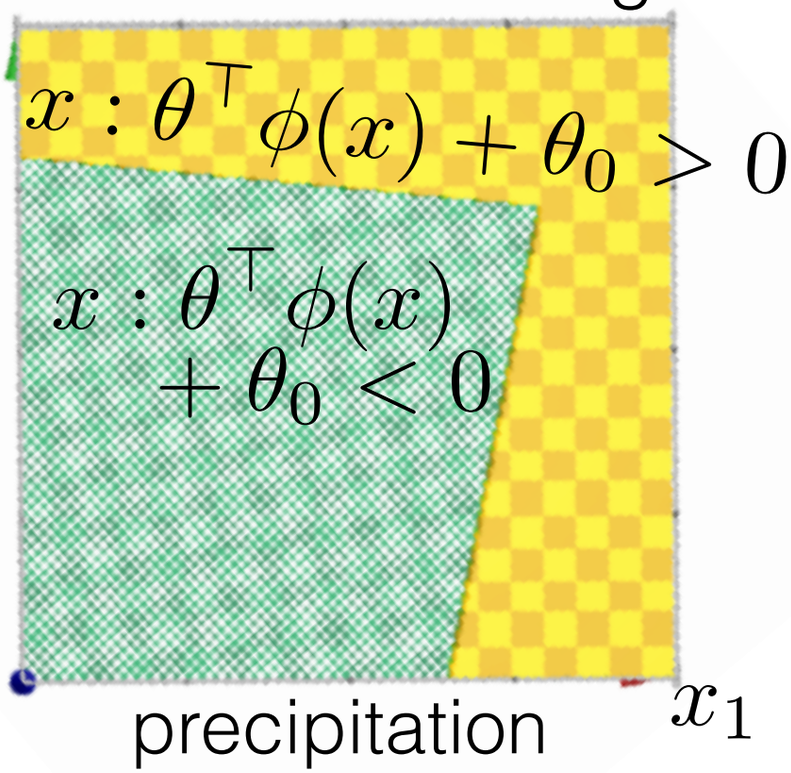

$$\phi_3(x) = \mathbf{1}\{\tilde{\tilde{w}}^\top x + \tilde{\tilde{w}}_0 \geq 0\}$$




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

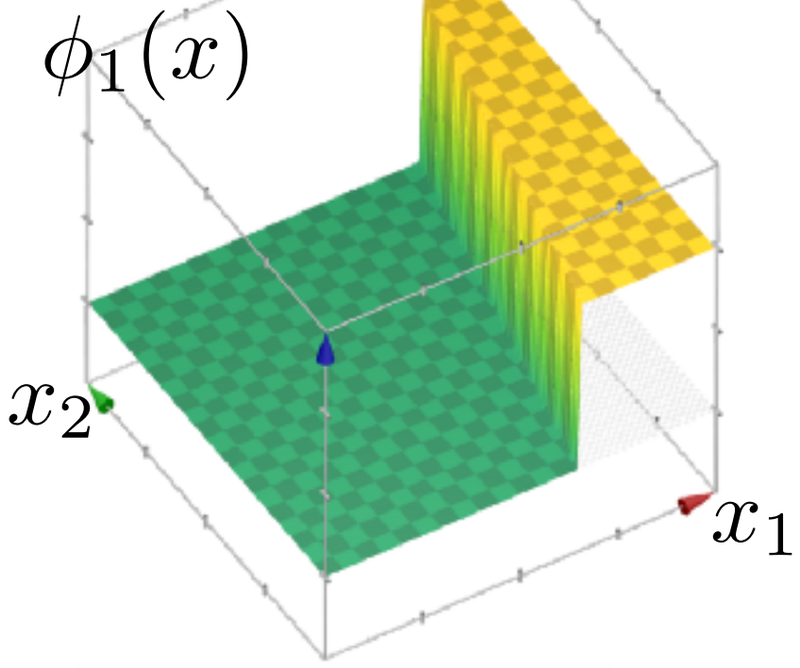



Will I avoid running?

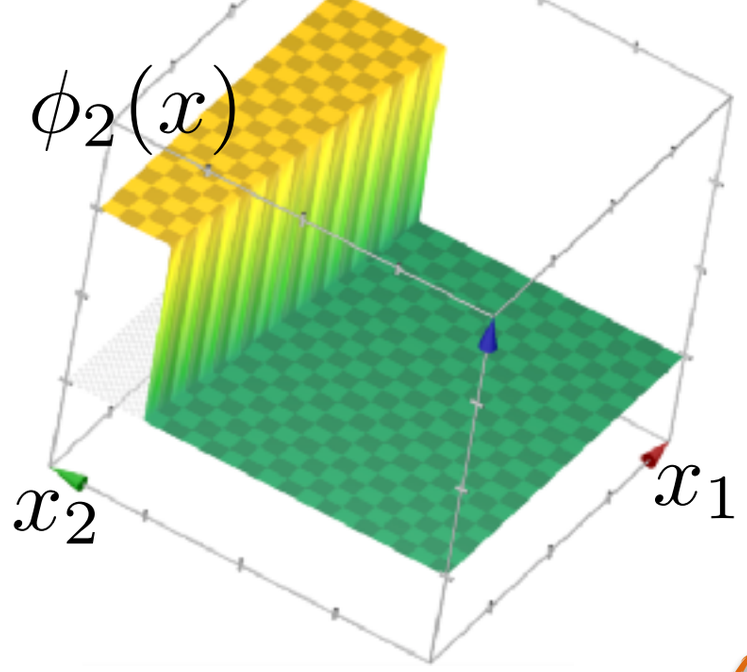


# New features: step functions!

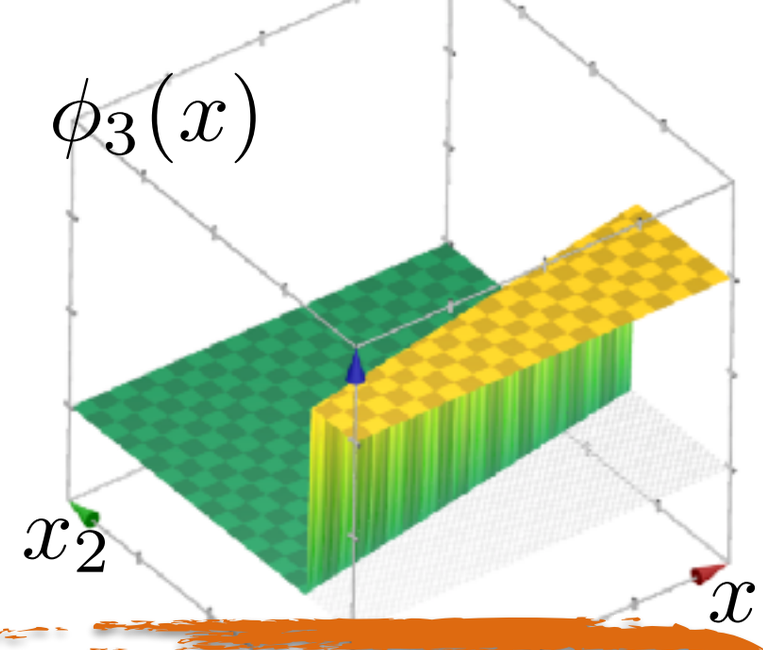
$$\phi_1(x) = \mathbf{1}\{w^\top x + w_0 \geq 0\}$$



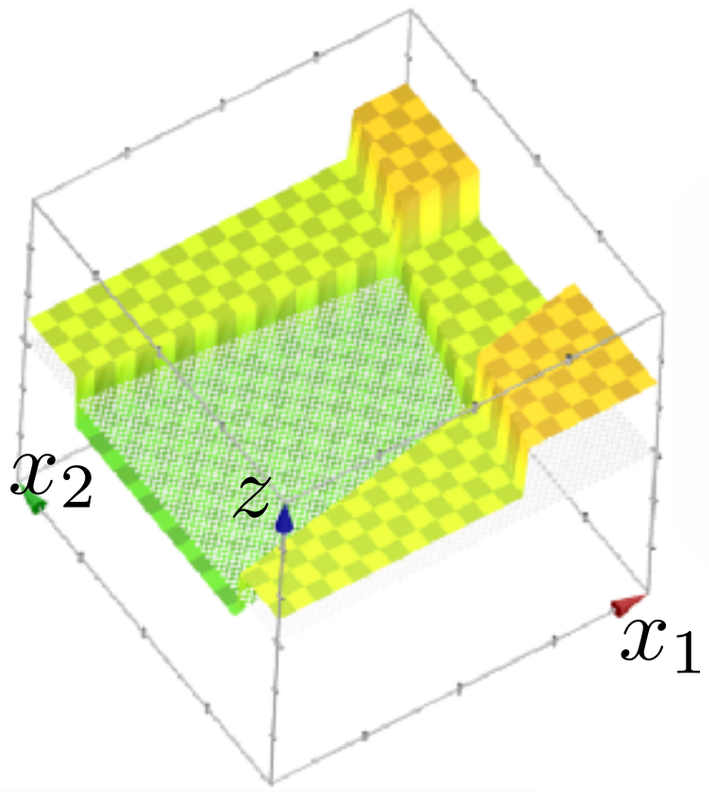

$$\phi_2(x) = \mathbf{1}\{\tilde{w}^\top x + \tilde{w}_0 \geq 0\}$$



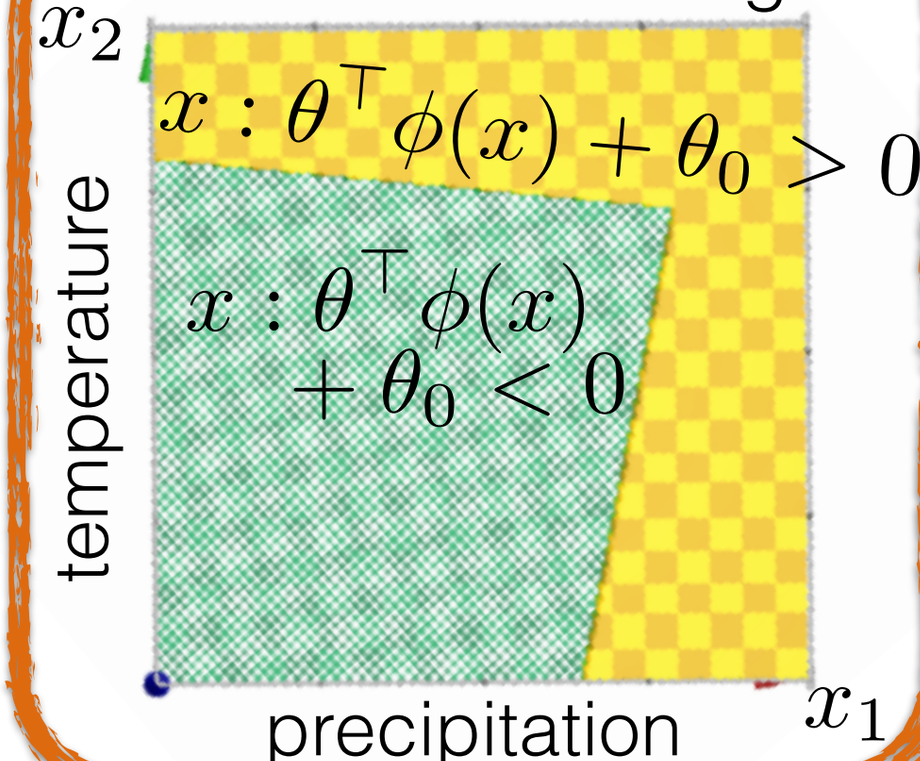

$$\phi_3(x) = \mathbf{1}\{\tilde{\tilde{w}}^\top x + \tilde{\tilde{w}}_0 \geq 0\}$$




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

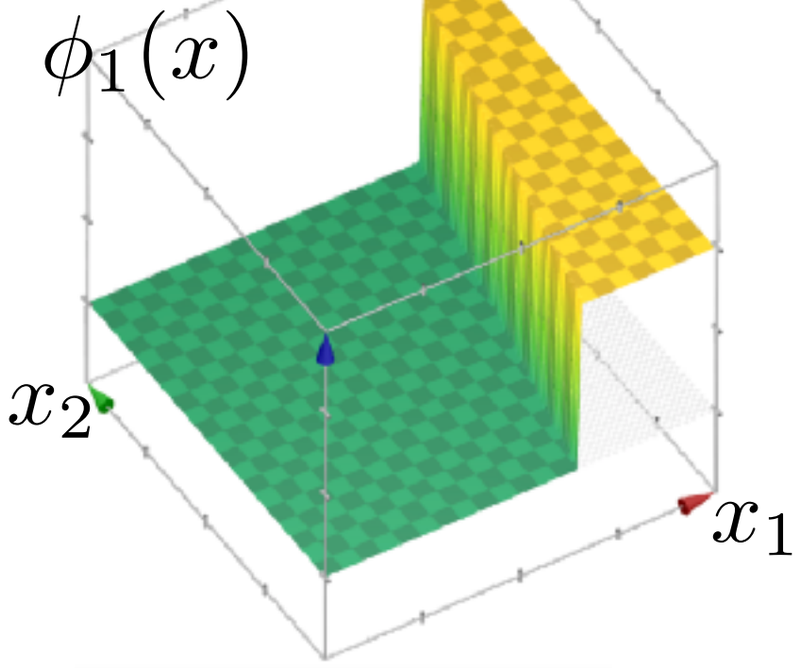



Will I avoid running?

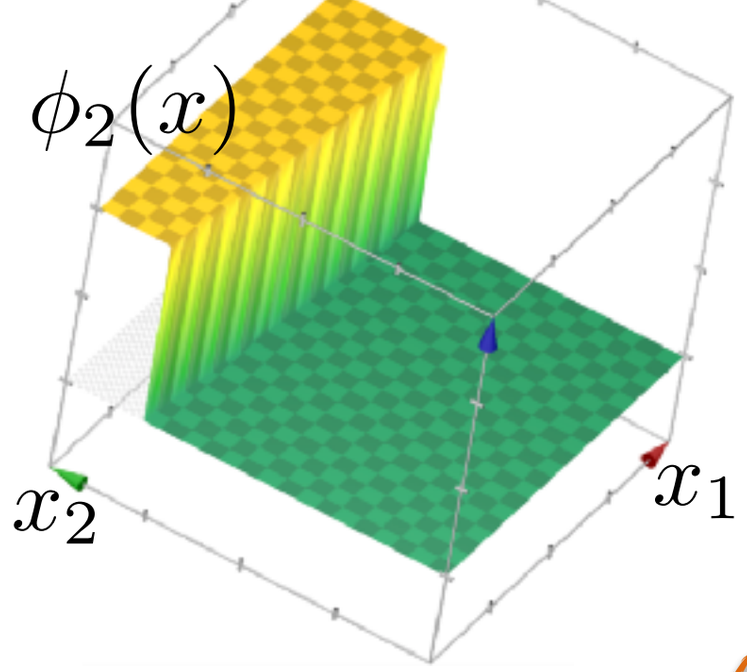


# New features: step functions!

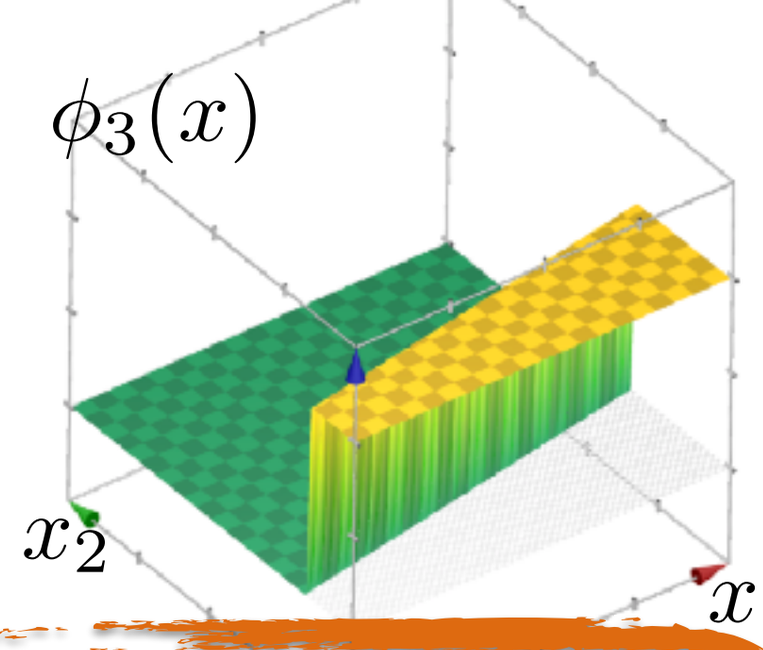
$$\phi_1(x) = \mathbf{1}\{w^\top x + w_0 \geq 0\}$$



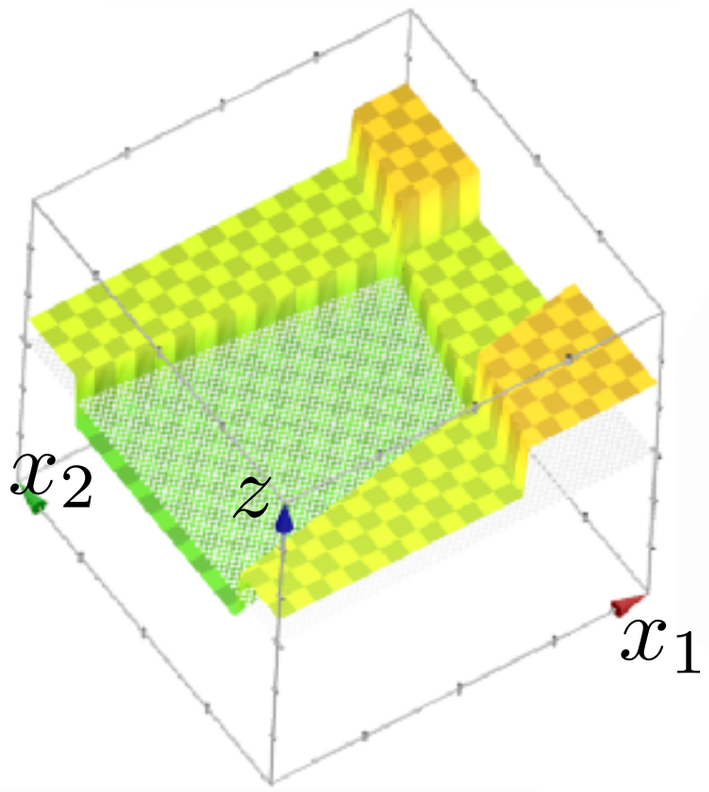

$$\phi_2(x) = \mathbf{1}\{\tilde{w}^\top x + \tilde{w}_0 \geq 0\}$$



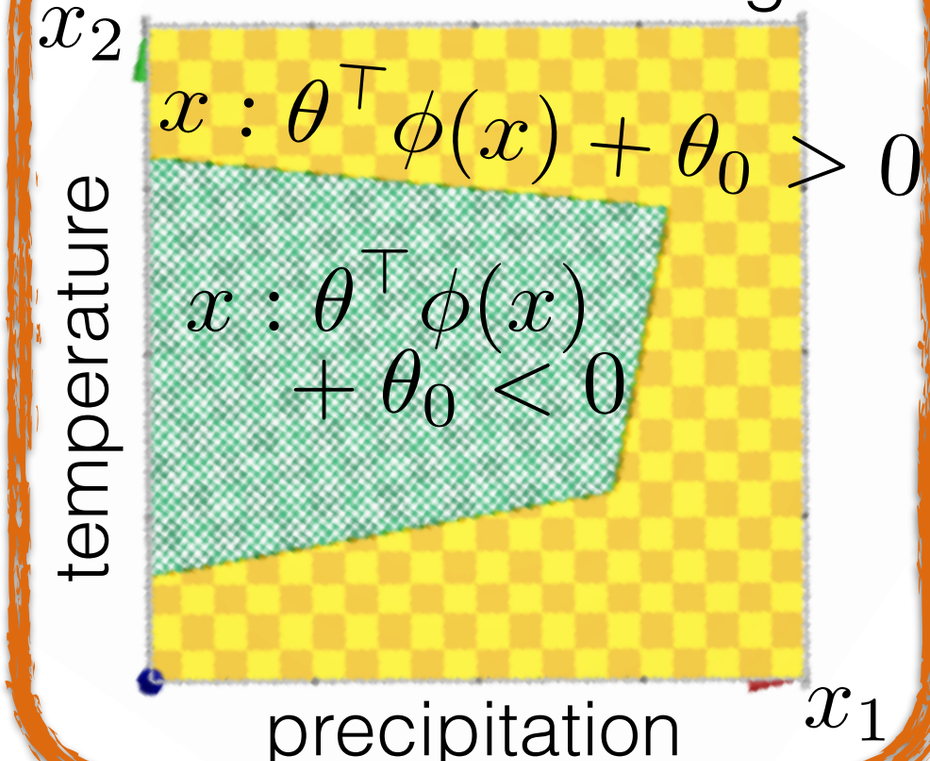

$$\phi_3(x) = \mathbf{1}\{\tilde{\tilde{w}}^\top x + \tilde{\tilde{w}}_0 \geq 0\}$$




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




Will I avoid running?



# Let’s get some new notation

# Let’s get some new notation

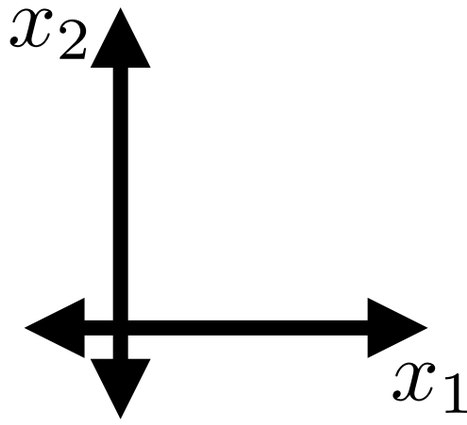
- 1st layer, constructing the features:

# Let’s get some new notation

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

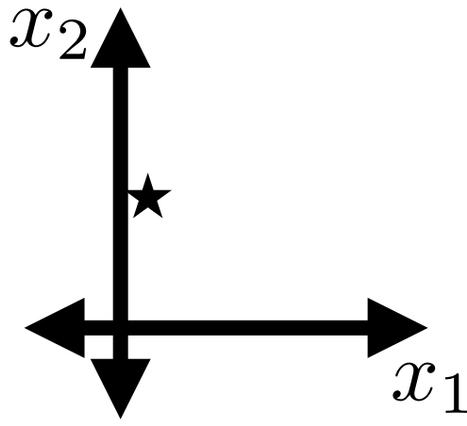
# Let's get some new notation

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



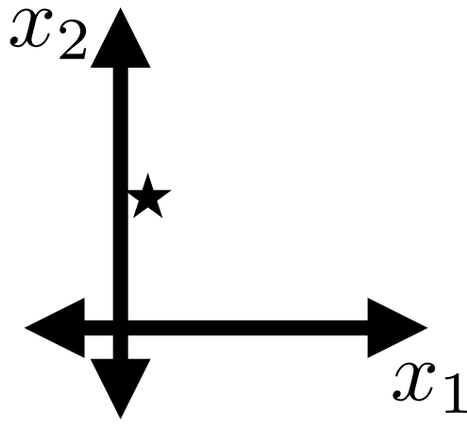
# Let's get some new notation

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



# Let's get some new notation

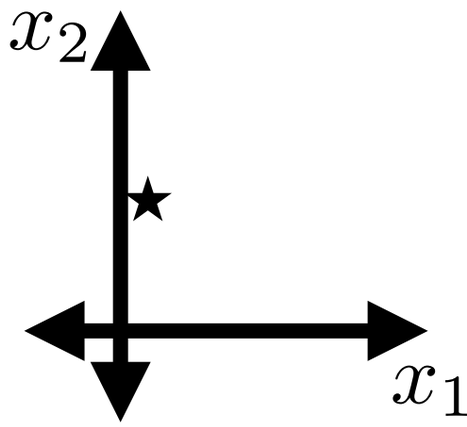
- 1st layer, constructing the features:
  - Input $x$ (a data point): size $m^{(1)} \times 1$



# Let's get some new notation

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

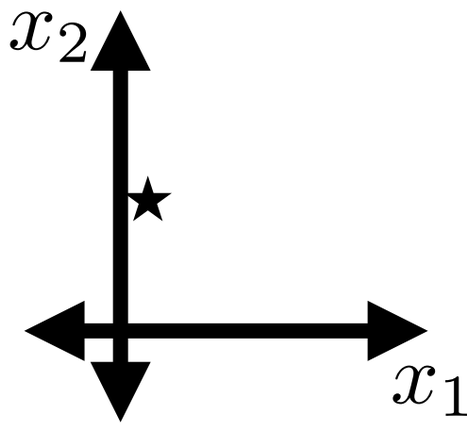
$m^{(1)} = \mathrm{d}$



# 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)

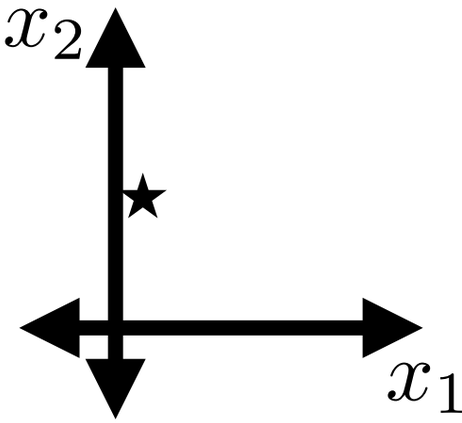
$m^{(1)} = \mathrm{d}$

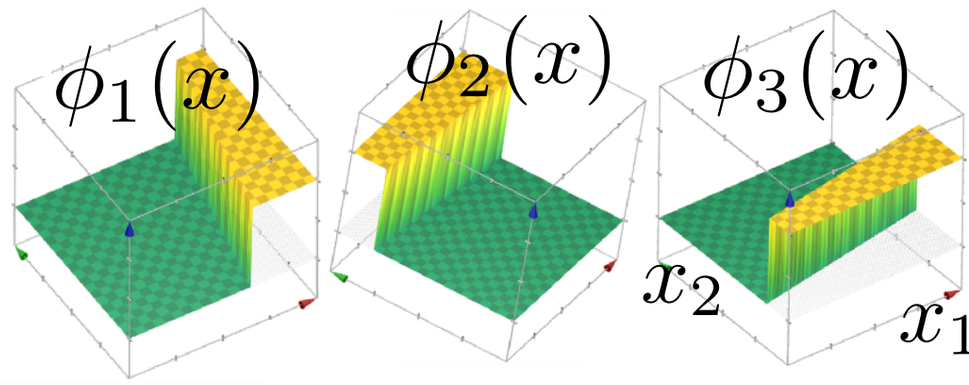


# Let's get some new notation

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

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

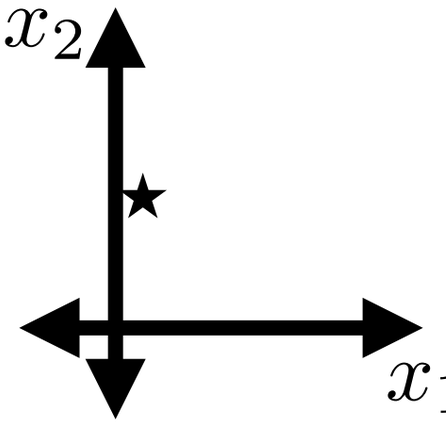


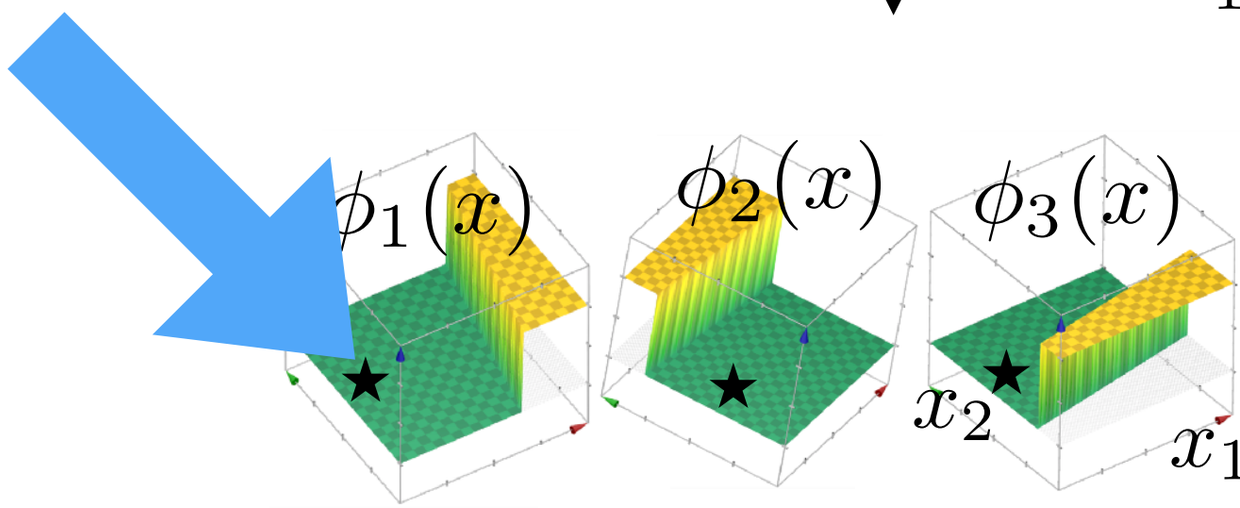



# Let's get some new notation

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

$m^{(1)} = d$

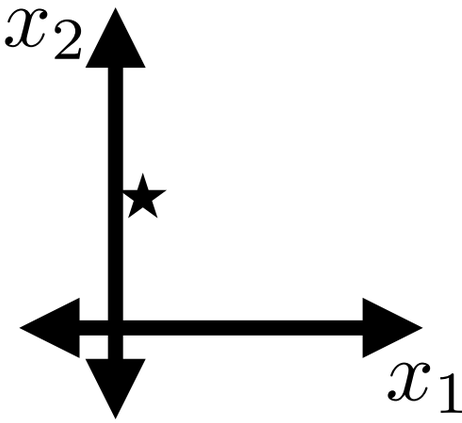


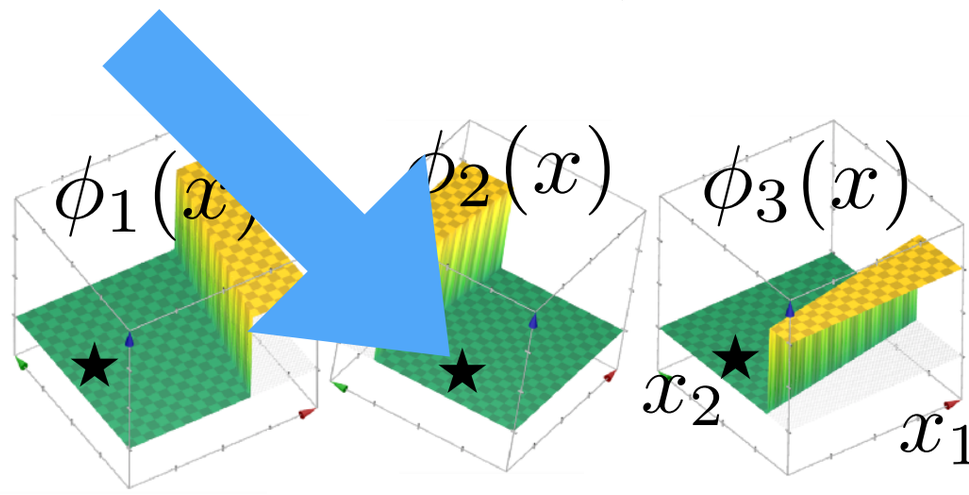



# Let's get some new notation

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

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

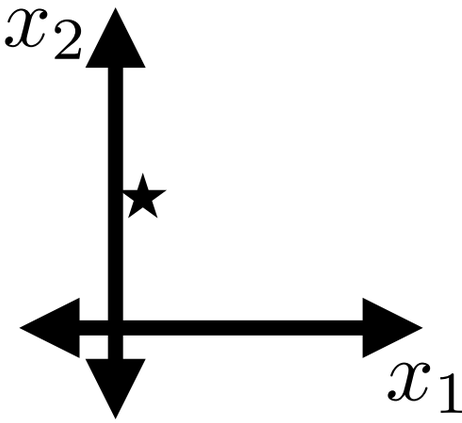


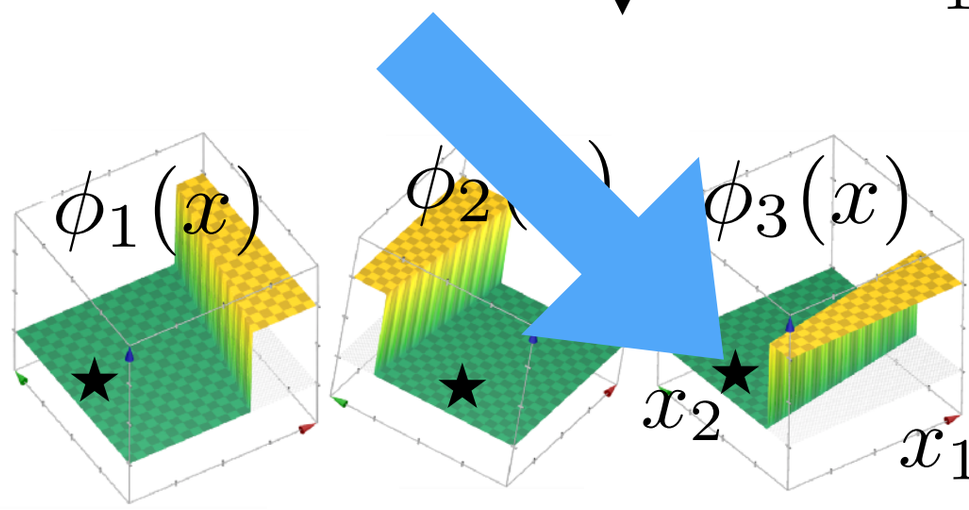



# Let's get some new notation

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

$m^{(1)} = d$

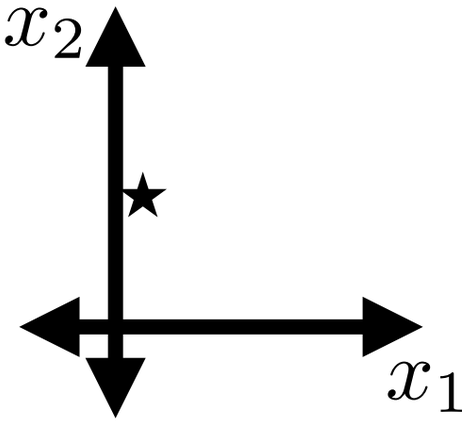


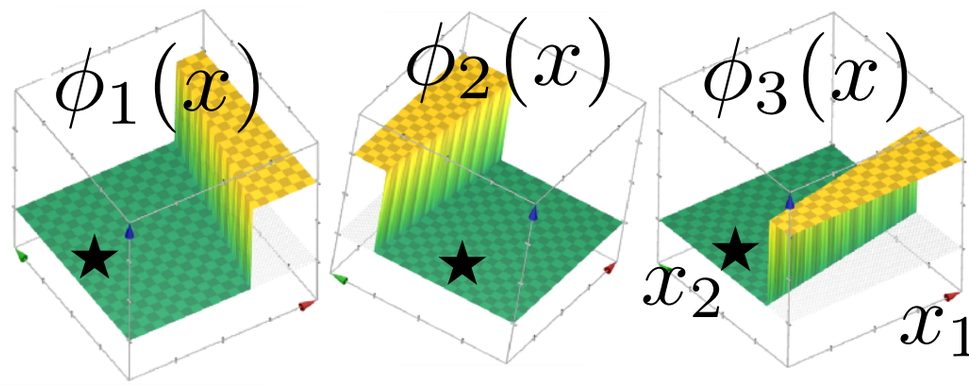



# Let's get some new notation

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

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

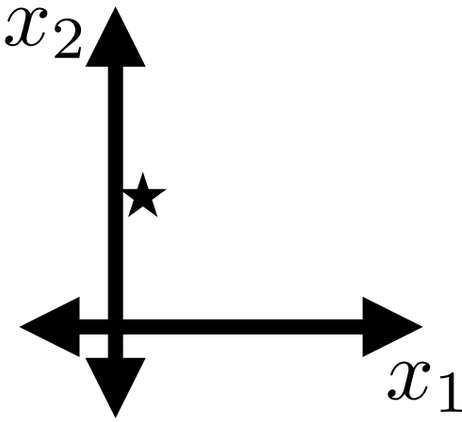


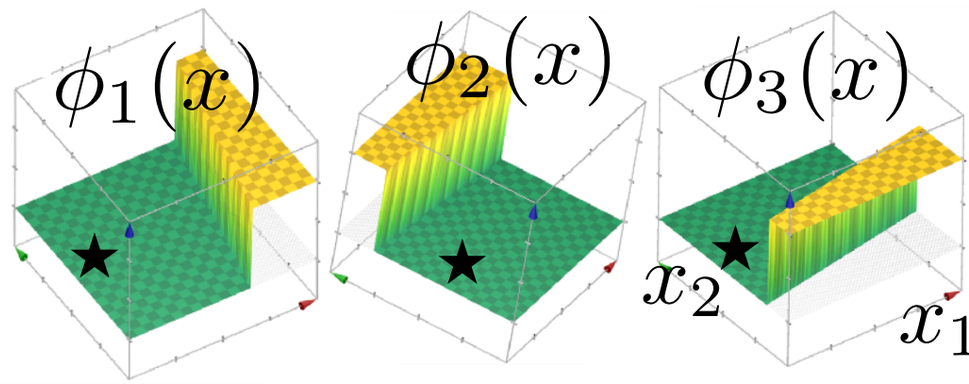



# Let's get some new notation

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

$m^{(1)} = d$

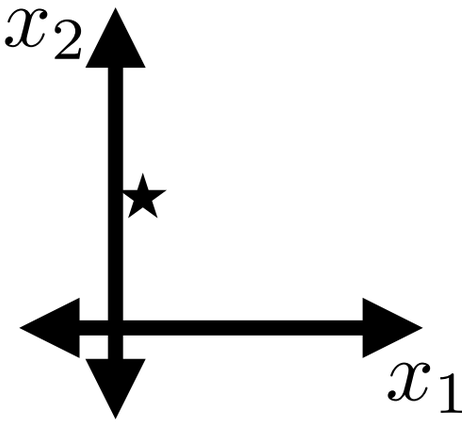


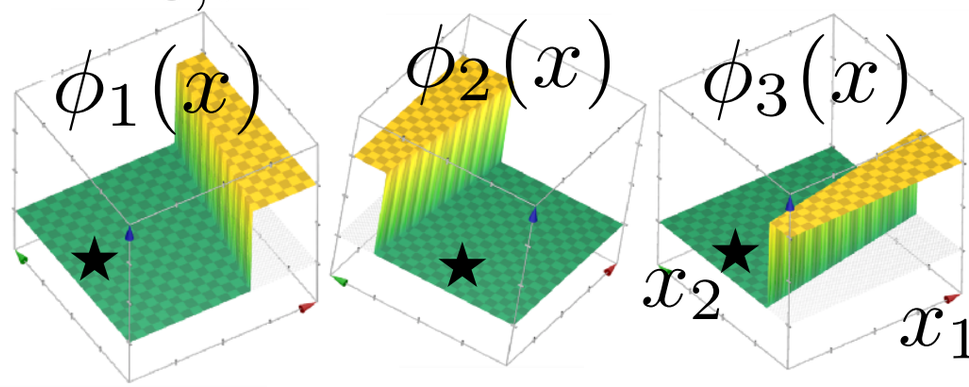



# Let's get some new notation

- 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 $i$th feature: $A_i^{(1)} = f^{(1)}(w_i^{(1)\top} x + w_{0,i}^{(1)})$

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

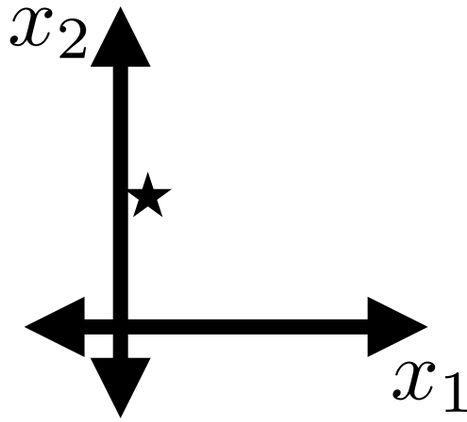


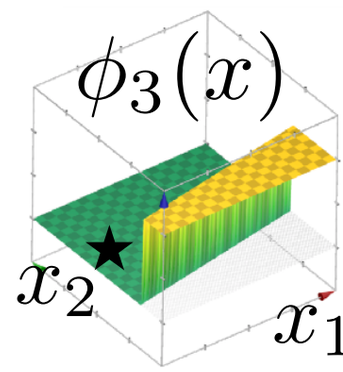



# Let's get some new notation

- 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 $i$th feature: $A_i^{(1)} = f^{(1)}(w_i^{(1)\top} x + w_{0,i}^{(1)})$

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

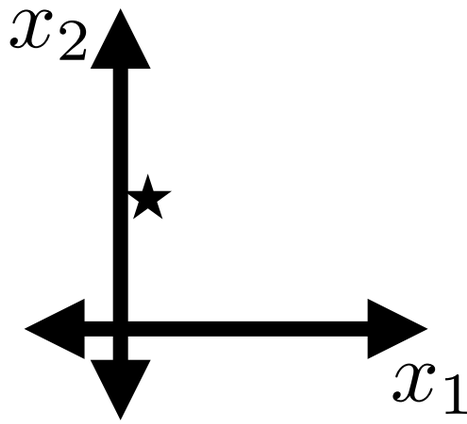


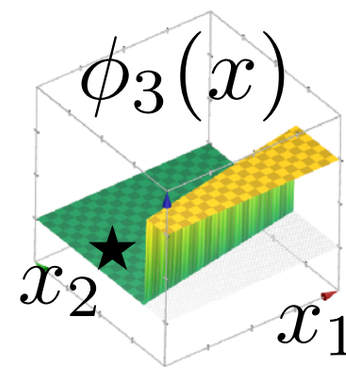



# Let's get some new notation

- 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 $i$th feature: $A_i^{(1)} = f^{(1)}(w_i^{(1)\top} x + w_{0,i}^{(1)})$

$m^{(1)} = d$

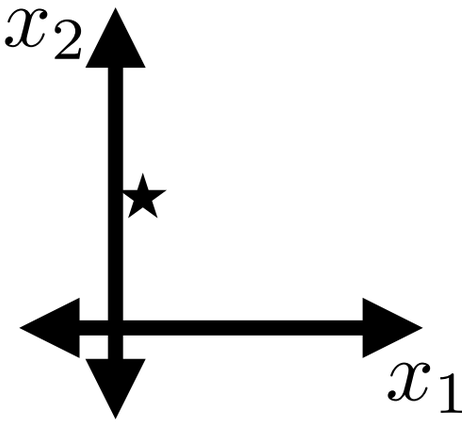


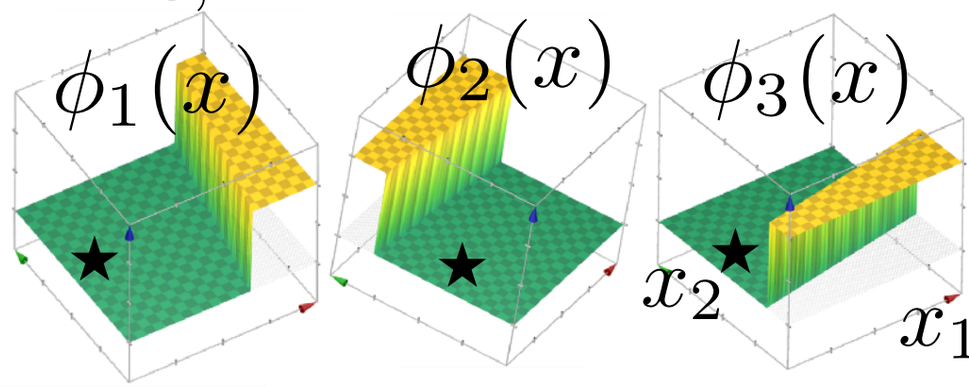



# Let's get some new notation

- 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 $i$th feature: $A_i^{(1)} = f^{(1)}(w_i^{(1)\top} x + w_{0,i}^{(1)})$

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

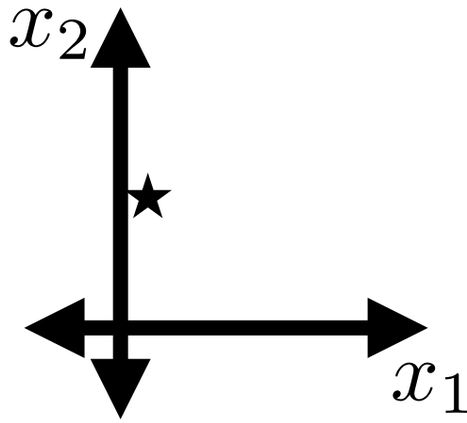


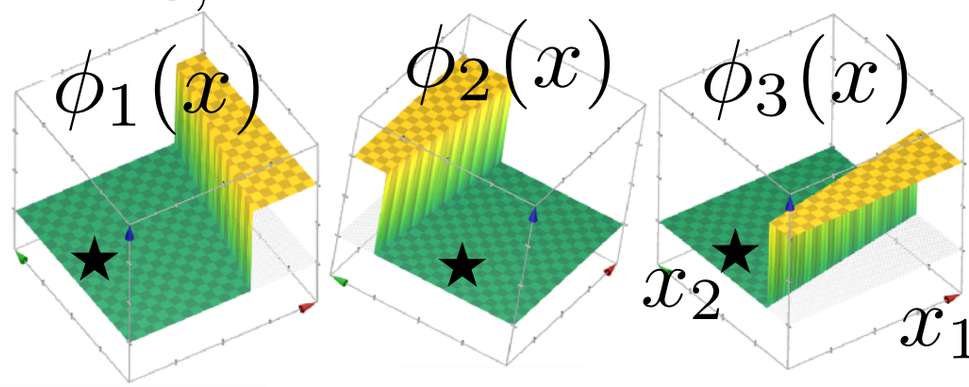



# Let's get some new notation

- 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 $i$th feature: $A_i^{(1)} = f^{(1)}(w_i^{(1)\top} x + w_{0,i}^{(1)})$
    - All the features at once:

$m^{(1)} = d$

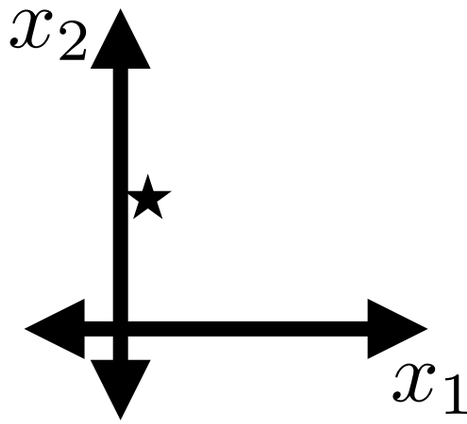


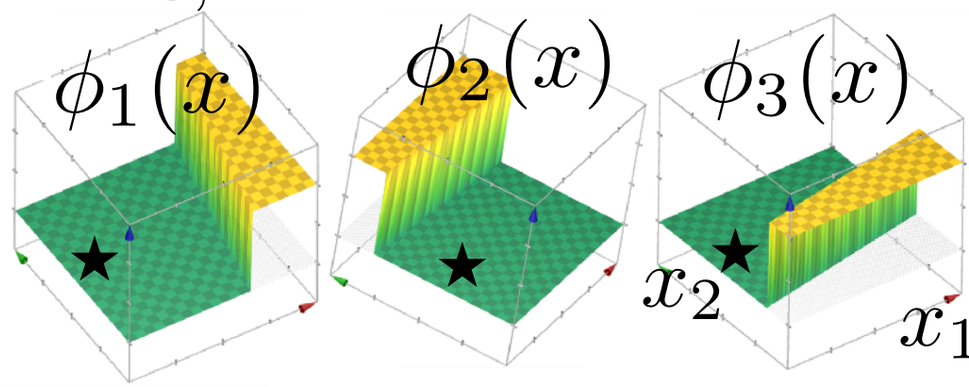



# Let's get some new notation

- 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 $i$th feature: $A_i^{(1)} = f^{(1)}(w_i^{(1)\top} x + w_{0,i}^{(1)})$
    - All the features at once:
      - $A^{(1)} = f^{(1)}(W^{(1)\top} x + W_0^{(1)})$

$m^{(1)} = d$

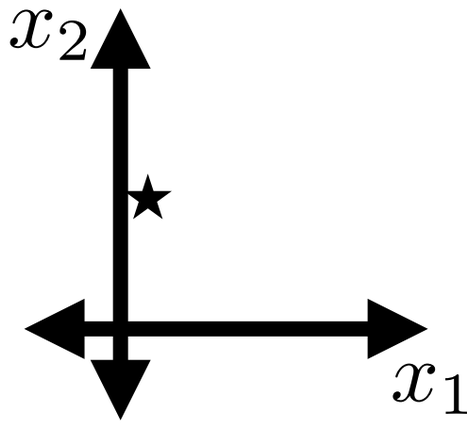


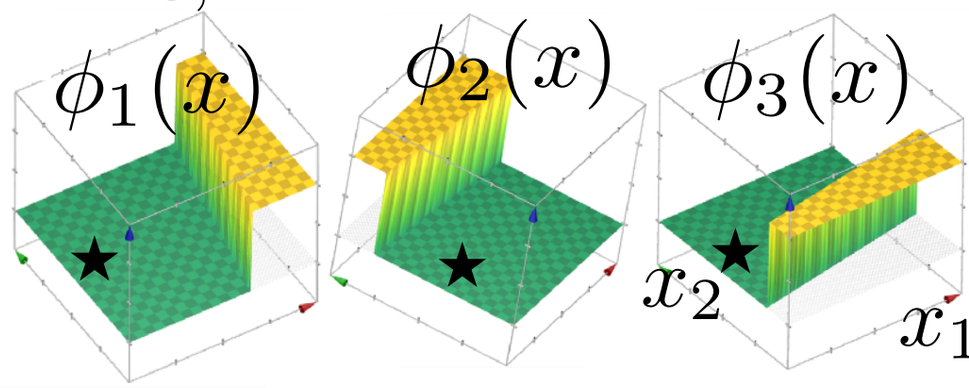



# Let's get some new notation

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

$m^{(1)} = d$

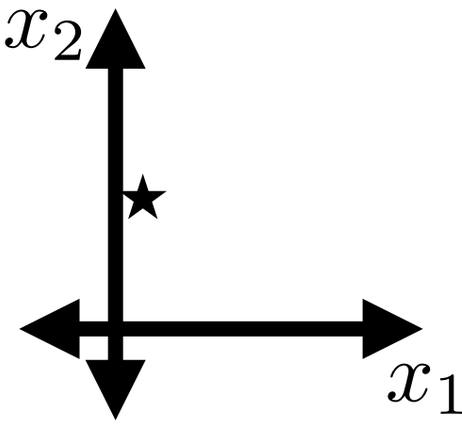


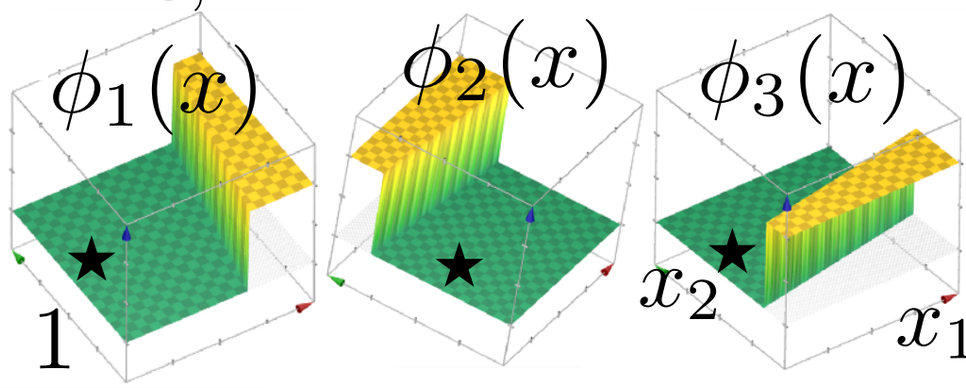



# Let's get some new notation

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

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

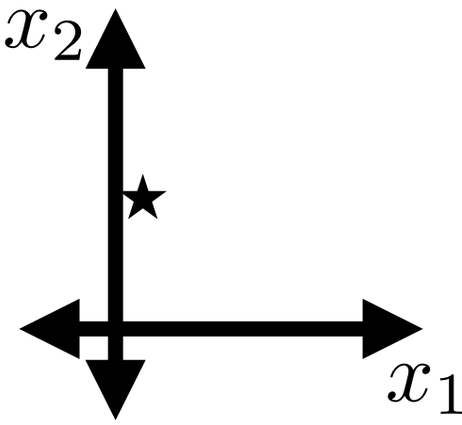


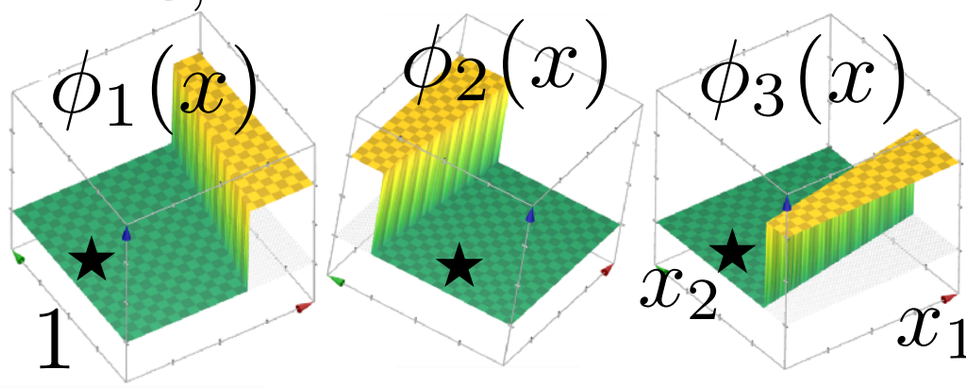



# Let's get some new notation

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

$m^{(1)} = d$

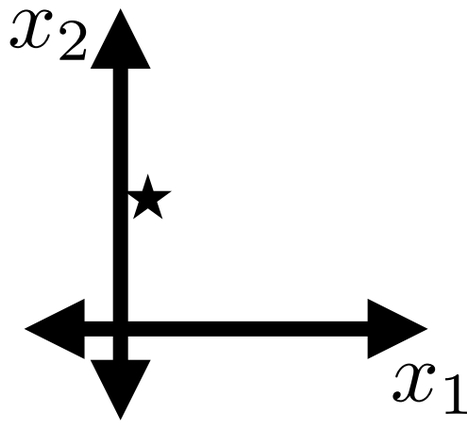


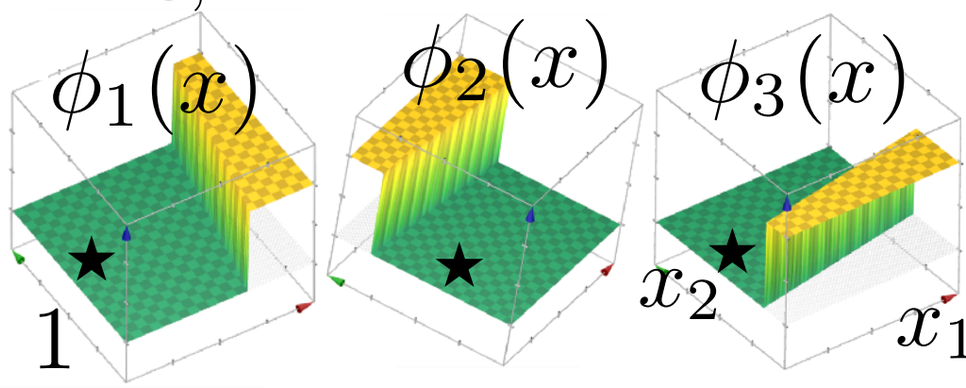



# Let's get some new notation

- 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:
      - $A^{(1)} = f^{(1)}(W^{(1)\top} x + W_0^{(1)})$
      - $W^{(1)} : m^{(1)} \times n^{(1)}; W_0^{(1)} : n^{(1)} \times 1$

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

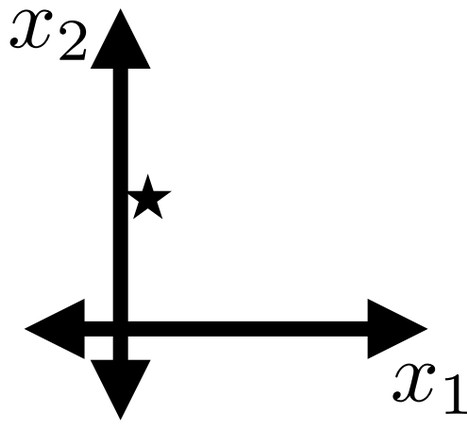


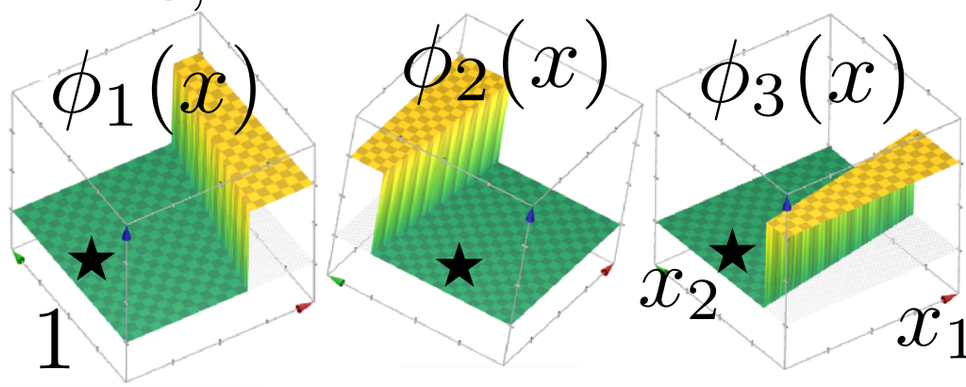



# Let's get some new notation

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

$m^{(1)} = d$



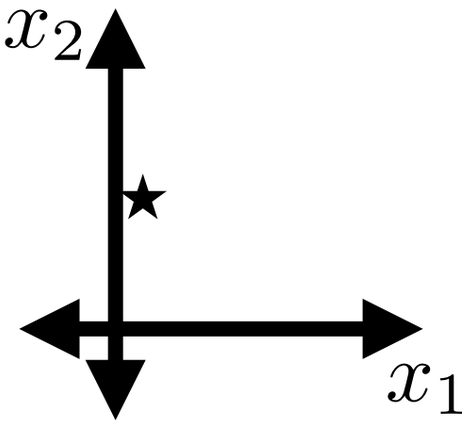



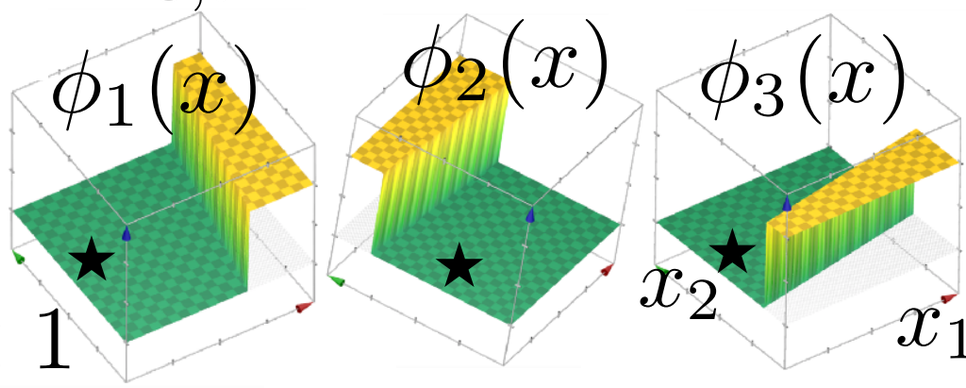

$f^{(1)}$ is applied componentwise!

# 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:
      - $A^{(1)} = f^{(1)}(W^{(1)\top} x + W_0^{(1)})$
      - $W^{(1)} : m^{(1)} \times n^{(1)}; W_0^{(1)} : n^{(1)} \times 1$

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

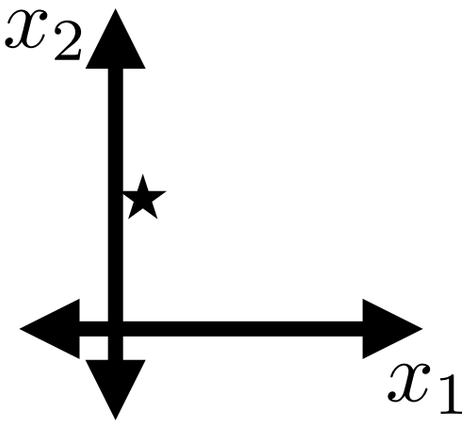


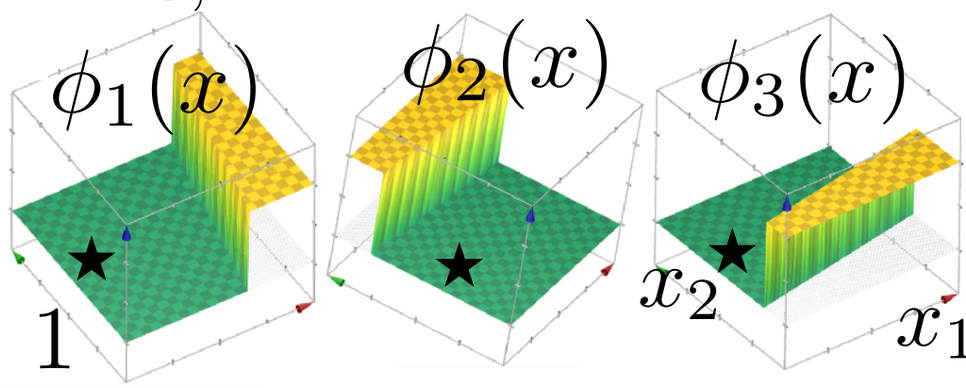



# Let's get some new notation

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

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



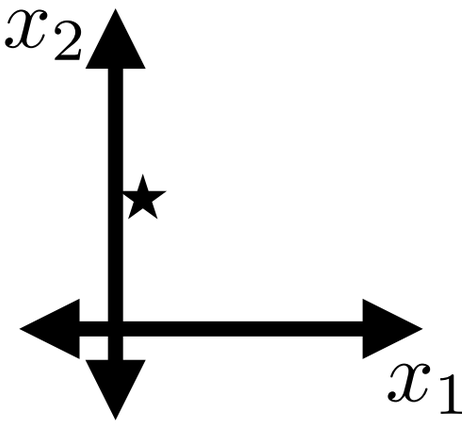



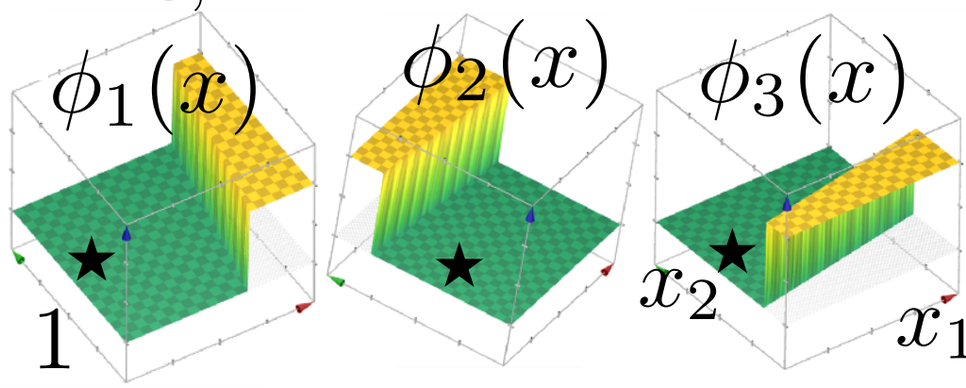

- 2nd layer, assigning a label (or labels):

# Let's get some new notation

- 1st layer, constructing the features:
  - Input $x$ (a data point): size $m^{(1)} \times 1$
  - 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:
      - $A^{(1)} = f^{(1)}(W^{(1)\top} x + W_0^{(1)})$
      - $W^{(1)} : m^{(1)} \times n^{(1)}; W_0^{(1)} : n^{(1)} \times 1$

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



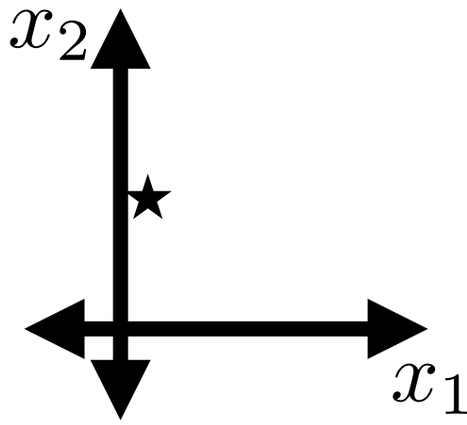



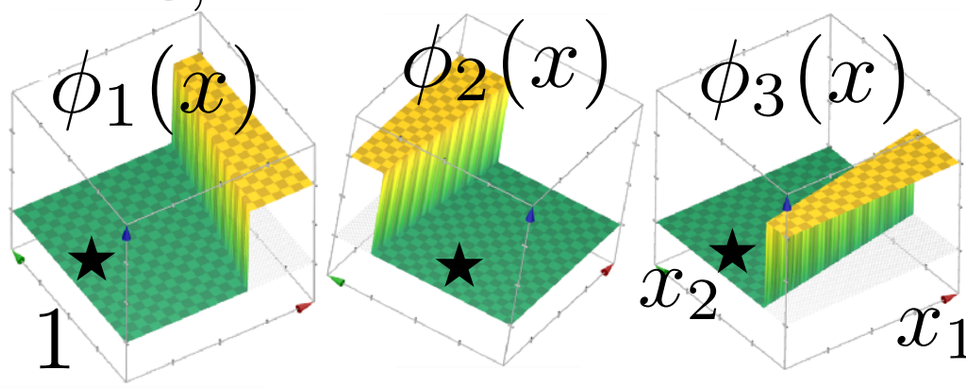

- 2nd layer, assigning a label (or labels):
  - Input (the features)

# 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:
      - $A^{(1)} = f^{(1)}(W^{(1)\top} x + W_0^{(1)})$
      - $W^{(1)} : m^{(1)} \times n^{(1)}; W_0^{(1)} : n^{(1)} \times 1$

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

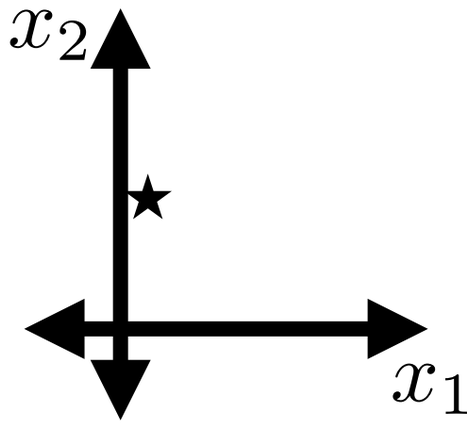


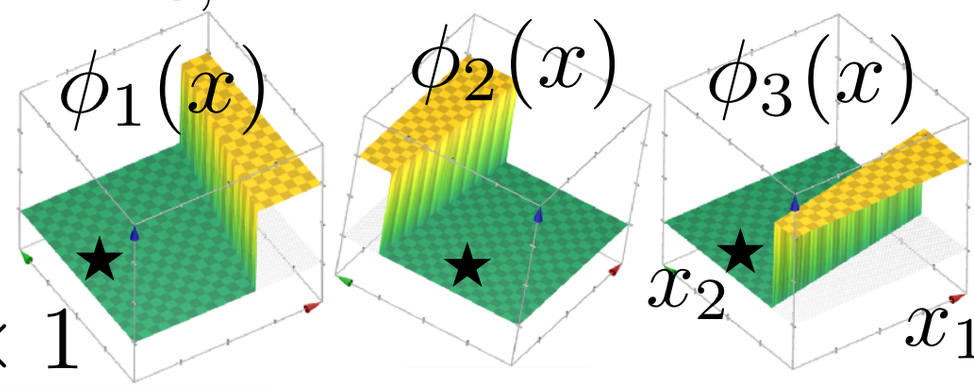




- 2nd layer, assigning a label (or labels):
  - Input (the features): size $m^{(2)} \times 1$

# Let's get some new notation

- 1st layer, constructing the features:
  - Input $x$ (a data point): size $m^{(1)} \times 1$ $m^{(1)} = d$
  - 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:
      - $A^{(1)} = f^{(1)}(W^{(1)\top} x + W_0^{(1)})$
      - $W^{(1)} : m^{(1)} \times n^{(1)}; W_0^{(1)} : n^{(1)} \times 1$



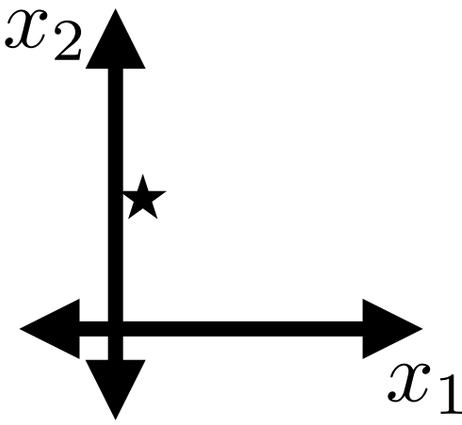



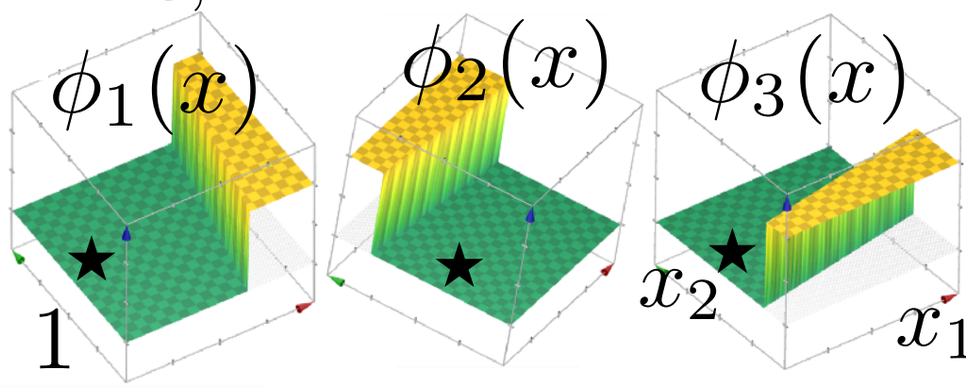

- 2nd layer, assigning a label (or labels):
  - Input (the features): size $m^{(2)} \times 1$ $m^{(2)} = n^{(1)}$

# 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:
      - $A^{(1)} = f^{(1)}(W^{(1)\top} x + W_0^{(1)})$
      - $W^{(1)} : m^{(1)} \times n^{(1)}; W_0^{(1)} : n^{(1)} \times 1$

$m^{(1)} = d$



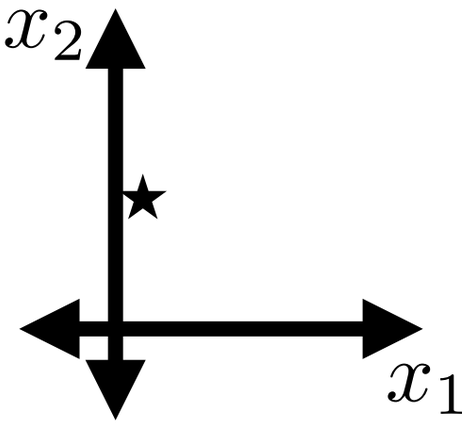

- 2nd layer, assigning a label (or labels):
  - Input (the features): size $m^{(2)} \times 1$
  - Output $A^{(2)}$

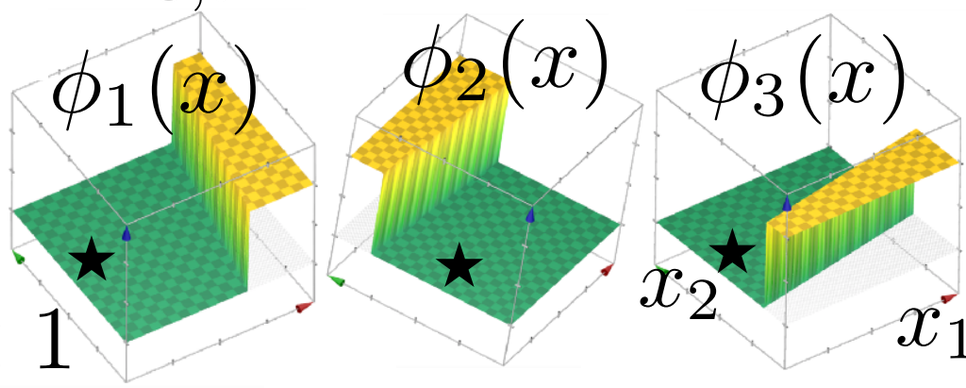
$m^{(2)} = n^{(1)}$

# 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 $A^{(2)}$

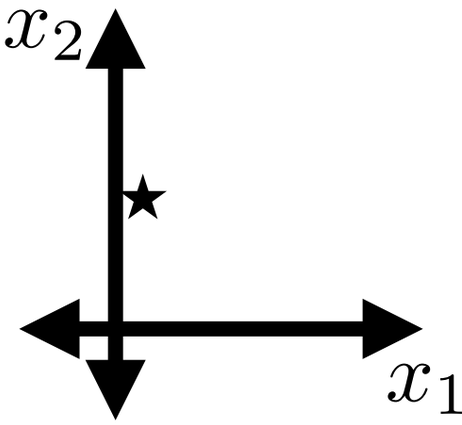
$m^{(2)} = n^{(1)}$

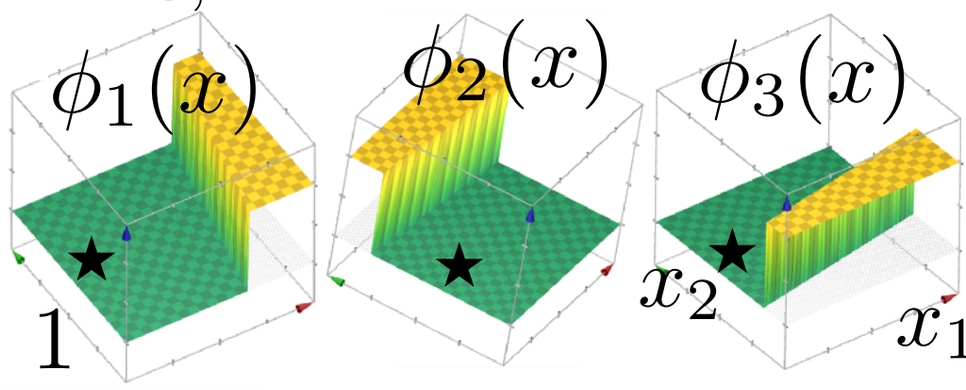
Like $y$

# 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:
      - $A^{(1)} = f^{(1)}(W^{(1)\top} x + W_0^{(1)})$
      - $W^{(1)} : m^{(1)} \times n^{(1)}; W_0^{(1)} : n^{(1)} \times 1$

$m^{(1)} = d$

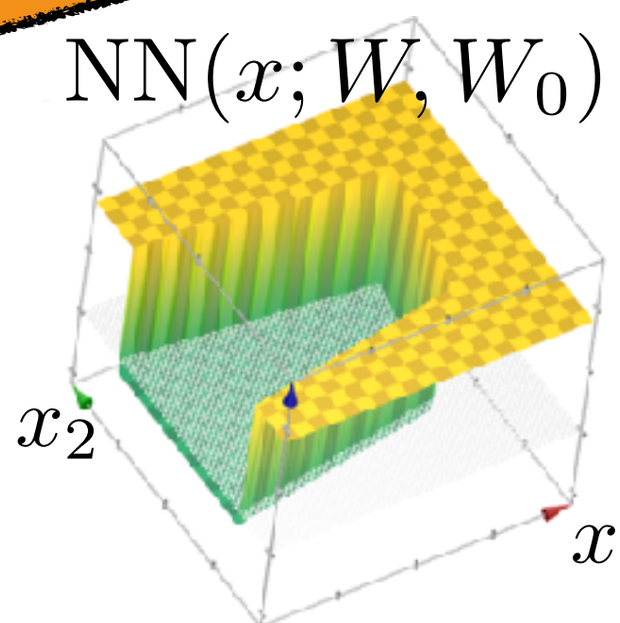



- 2nd layer, assigning a label (or labels):
  - Input (the features): size $m^{(2)} \times 1$
  - Output $A^{(2)}$

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

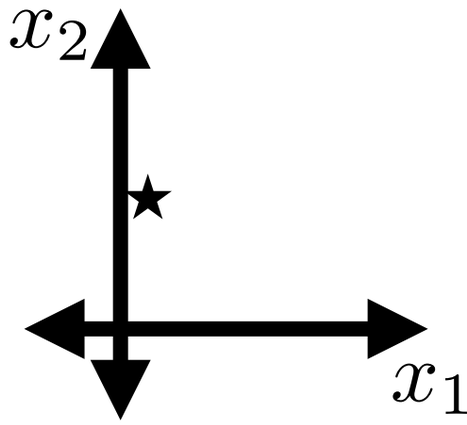
Like $y$

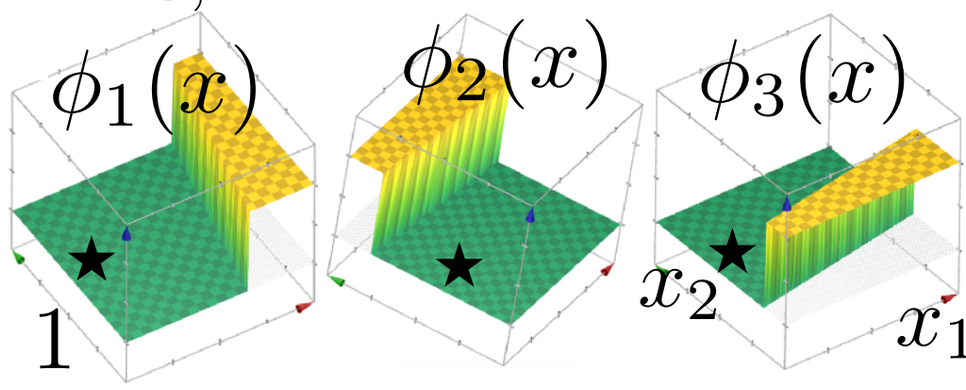


# Let's get some new notation

- 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:
      - $A^{(1)} = f^{(1)}(W^{(1)\top} x + W_0^{(1)})$
      - $W^{(1)} : m^{(1)} \times n^{(1)}; W_0^{(1)} : n^{(1)} \times 1$

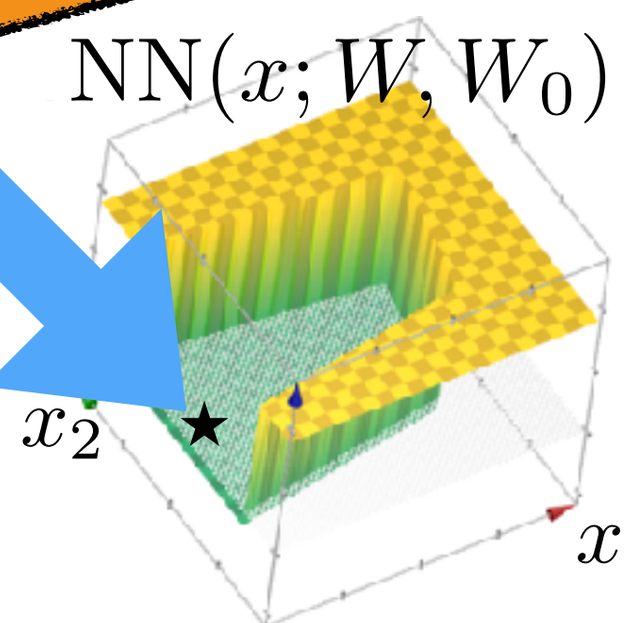
$m^{(1)} = d$




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

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

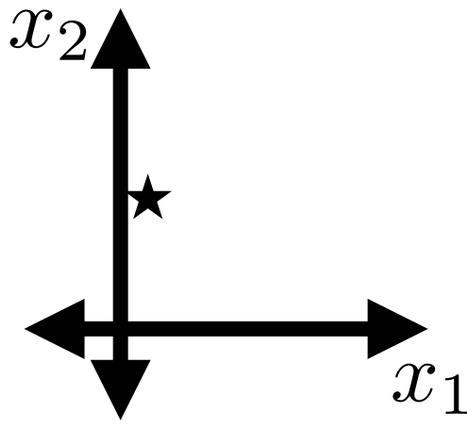


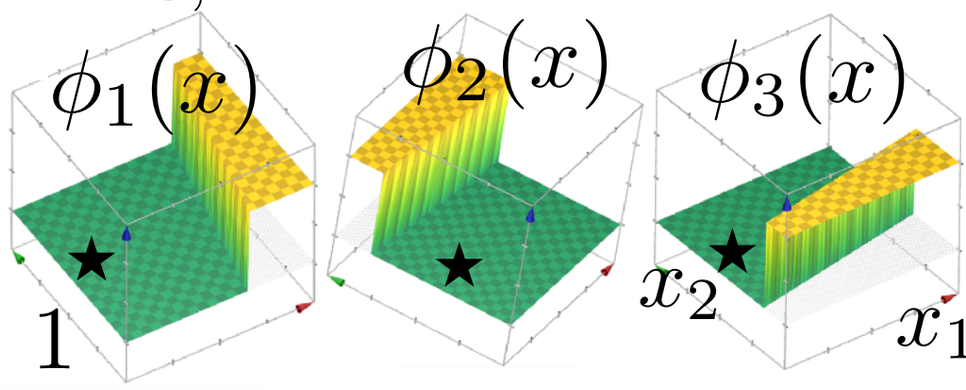

Like $y$

# 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:
      - $A^{(1)} = f^{(1)}(W^{(1)\top} x + W_0^{(1)})$
      - $W^{(1)} : m^{(1)} \times n^{(1)}; W_0^{(1)} : n^{(1)} \times 1$

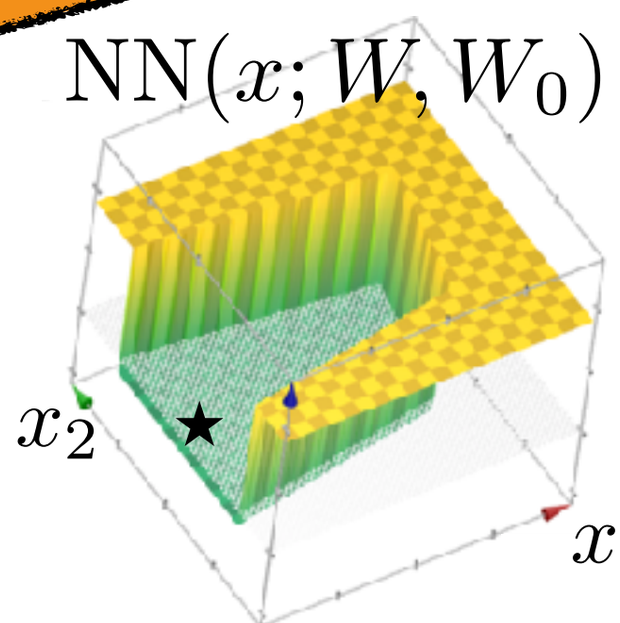
$m^{(1)} = d$




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

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

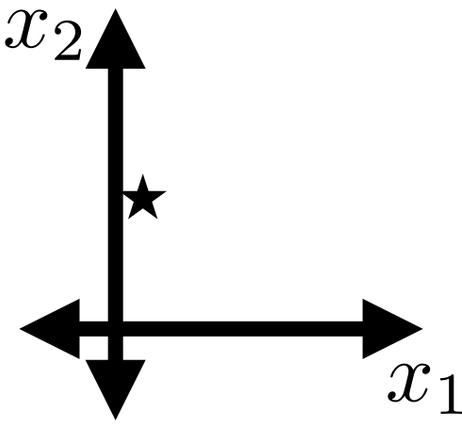


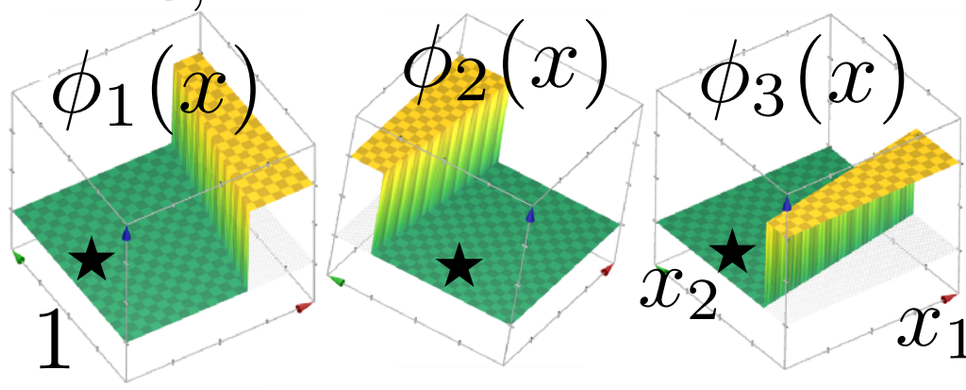

Like $y$

# 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:
      - $A^{(1)} = f^{(1)}(W^{(1)\top} x + W_0^{(1)})$
      - $W^{(1)} : m^{(1)} \times n^{(1)}; W_0^{(1)} : n^{(1)} \times 1$

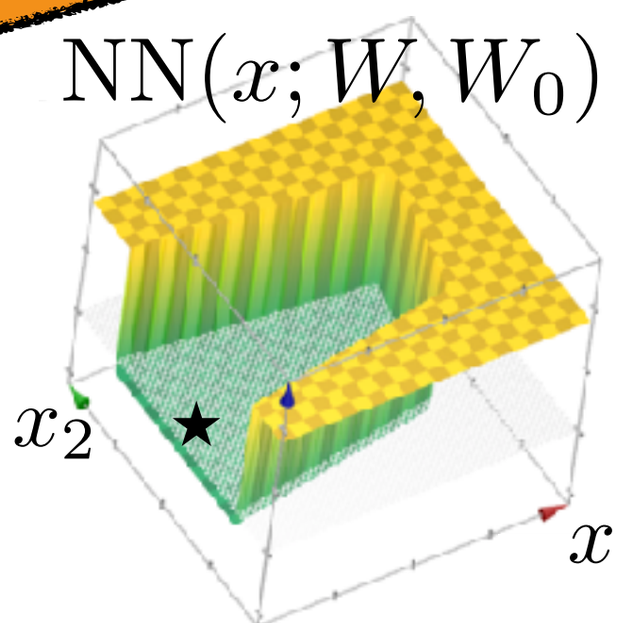
$m^{(1)} = d$




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

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

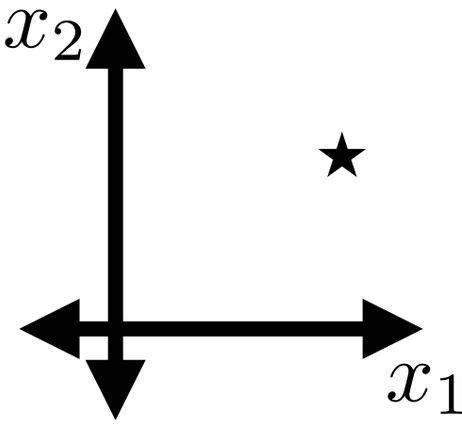


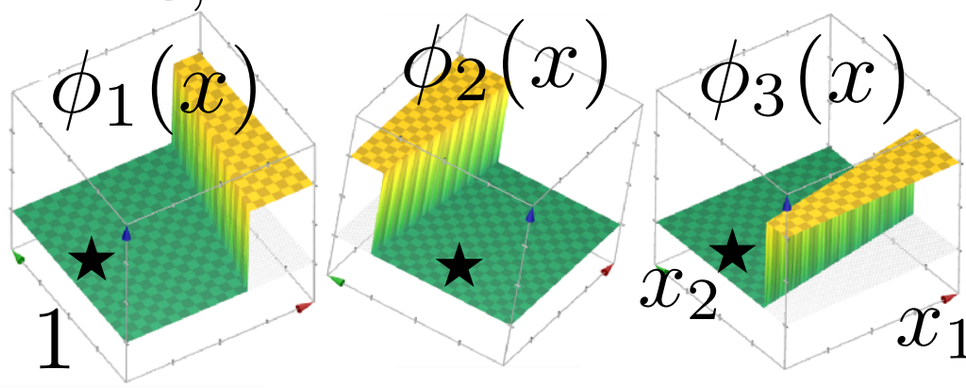

Like $y$

# 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:
      - $A^{(1)} = f^{(1)}(W^{(1)\top} x + W_0^{(1)})$
      - $W^{(1)} : m^{(1)} \times n^{(1)}; W_0^{(1)} : n^{(1)} \times 1$

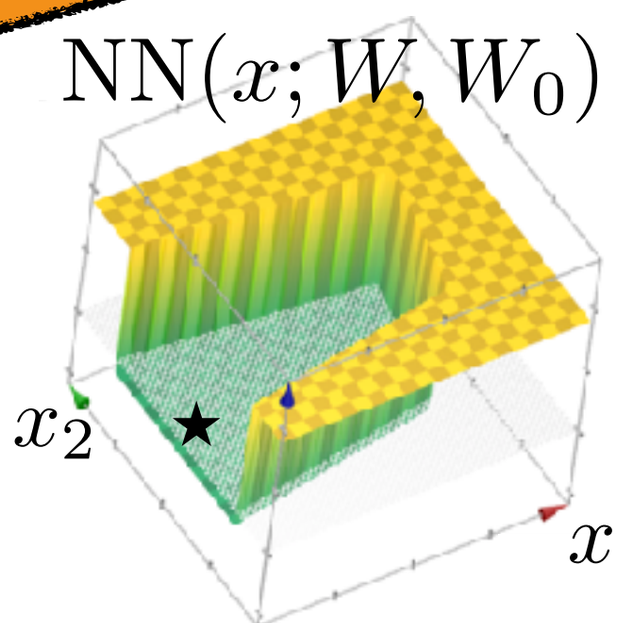
$m^{(1)} = d$




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

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

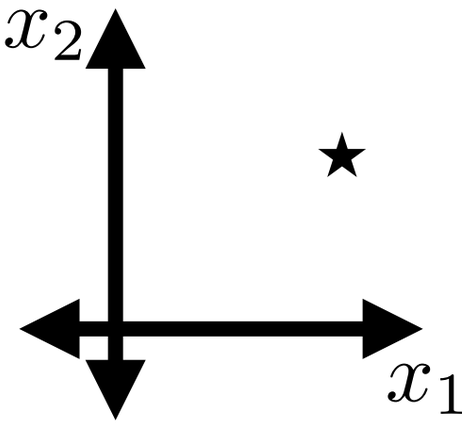


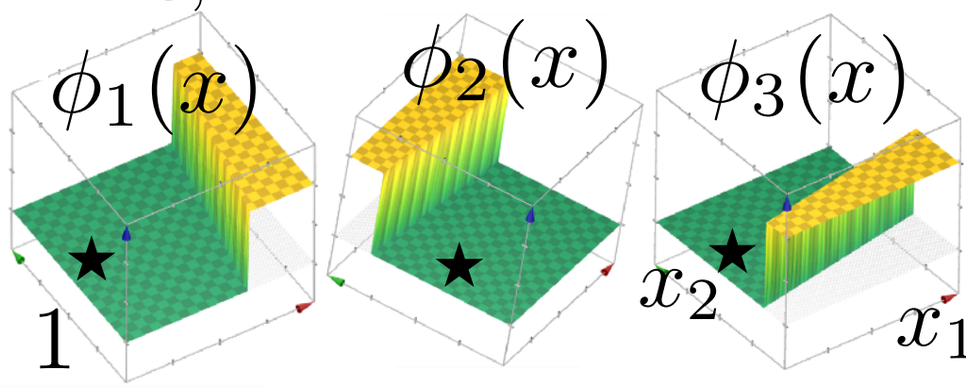

Like $y$

# 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:
      - $A^{(1)} = f^{(1)}(W^{(1)\top} x + W_0^{(1)})$
      - $W^{(1)} : m^{(1)} \times n^{(1)}; W_0^{(1)} : n^{(1)} \times 1$

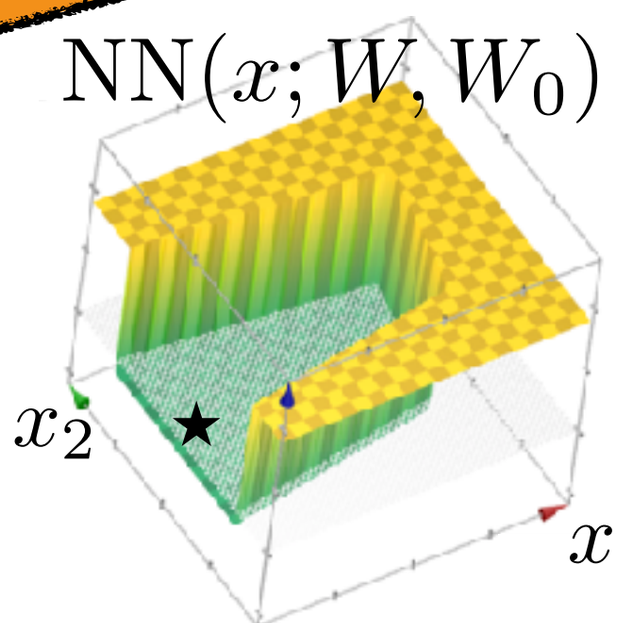
$m^{(1)} = d$




- 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:

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

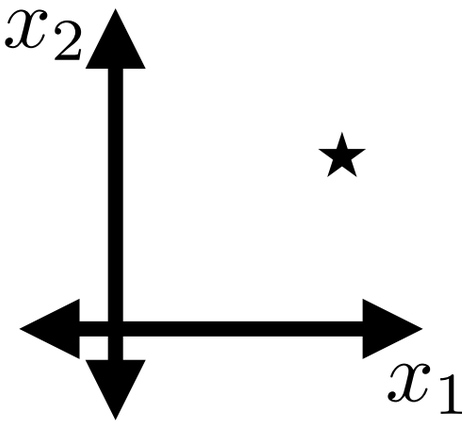


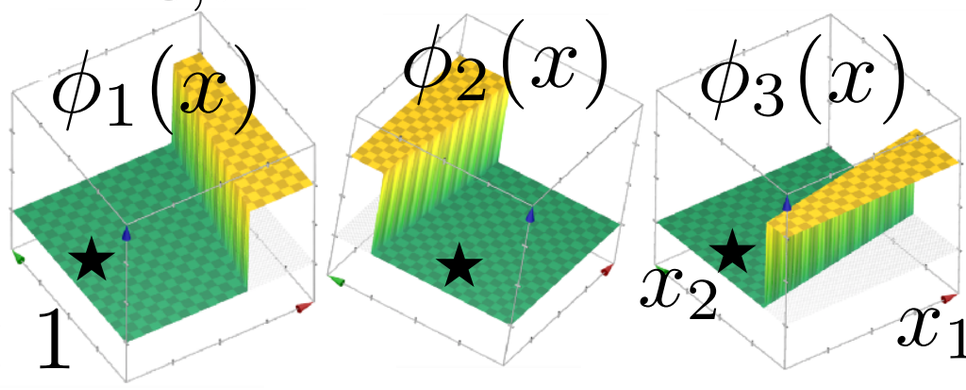

Like $y$

# 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:
      - $A^{(1)} = f^{(1)}(W^{(1)\top} x + W_0^{(1)})$
      - $W^{(1)} : m^{(1)} \times n^{(1)}; W_0^{(1)} : n^{(1)} \times 1$

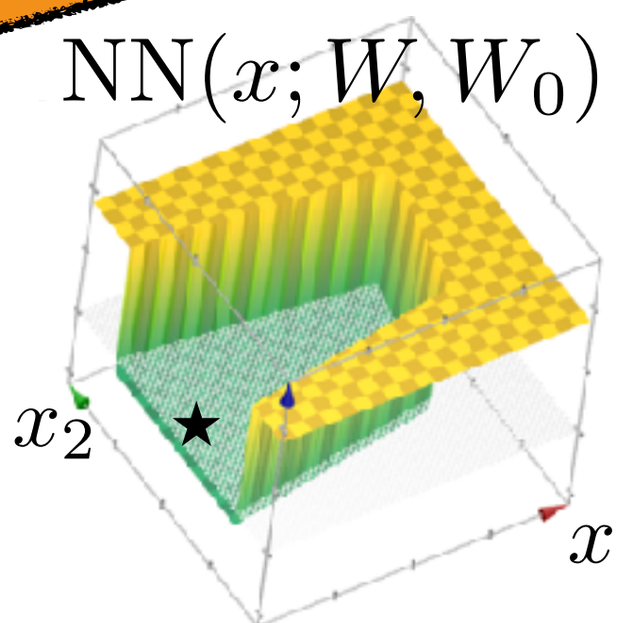
$m^{(1)} = d$




- 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: $A_i^{(2)} = f^{(2)}(w_i^{(2)\top} A^{(1)} + w_{0,i}^{(2)})$

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

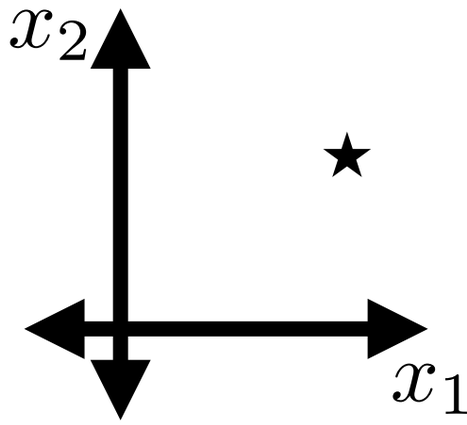


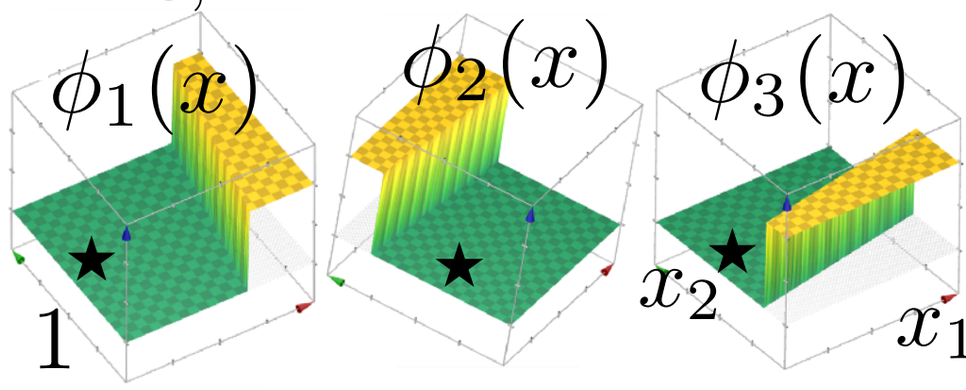

Like $y$

# 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:
      - $A^{(1)} = f^{(1)}(W^{(1)\top} x + W_0^{(1)})$
      - $W^{(1)} : m^{(1)} \times n^{(1)}; W_0^{(1)} : n^{(1)} \times 1$

$m^{(1)} = d$

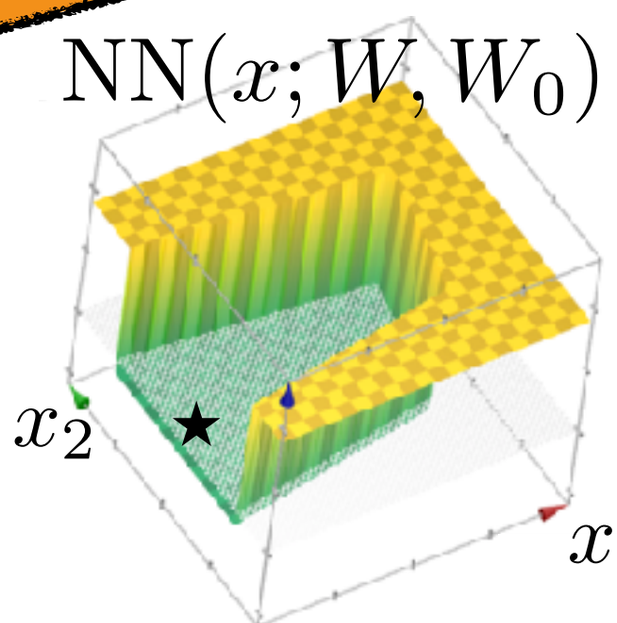



- 2nd layer, assigning a label (or labels):
  - Input (the features): size $m^{(2)} \times 1$
  - Output $A^{(2)}$(vector of labels): size $n^{(2)} \times 1$
  - The $i$th label: $A_i^{(2)} = f^{(2)}(w_i^{(2)\top} A^{(1)} + w_{0,i}^{(2)})$
    - All: $y$

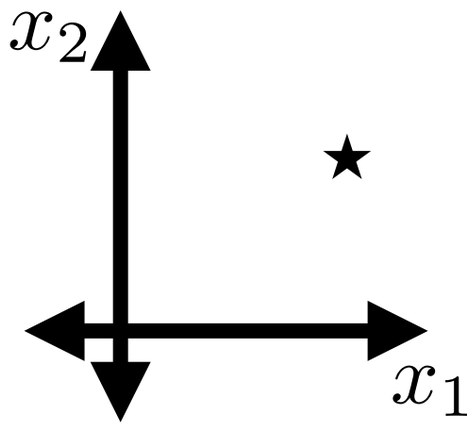
$m^{(2)} = n^{(1)}$

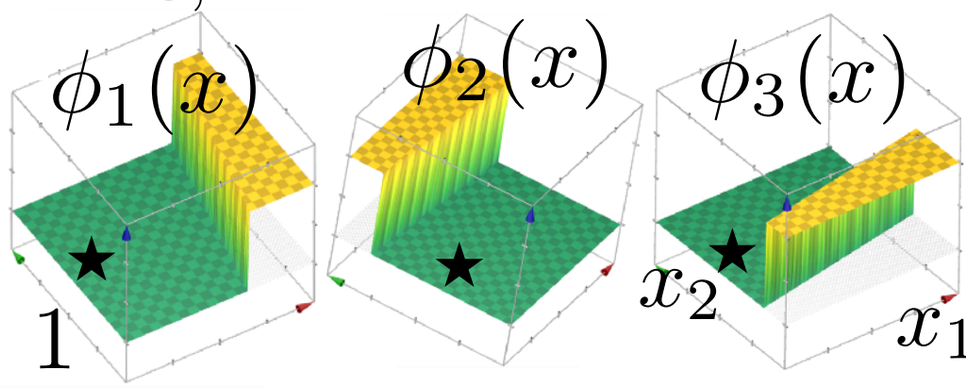
Like



# Let's get some new notation

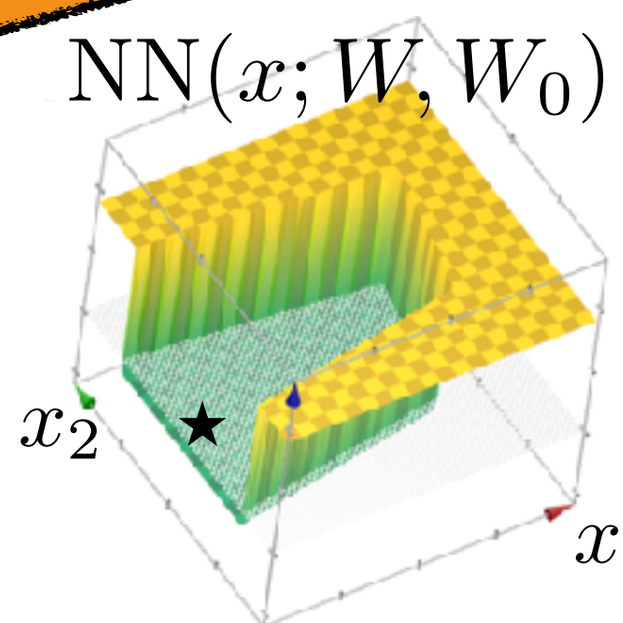
- 1st layer, constructing the features:
  - Input $x$ (a data point): size $m^{(1)} \times 1$ $m^{(1)} = d$
  - Output $A^{(1)}$ (vector of features): size $n^{(1)} \times 1$
  - The $i$th feature: $A_i^{(1)} = f^{(1)}(w_i^{(1)\top} x + w_{0,i}^{(1)})$
    - All the features at once:
      - $A^{(1)}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




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

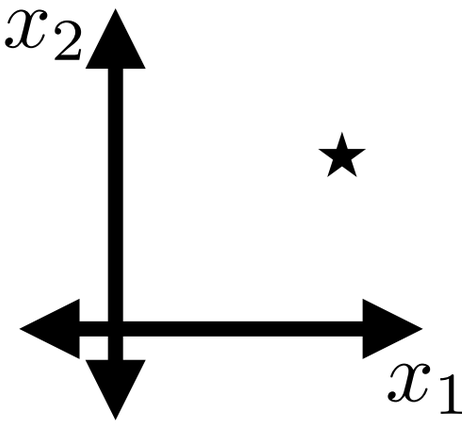
Like $y$

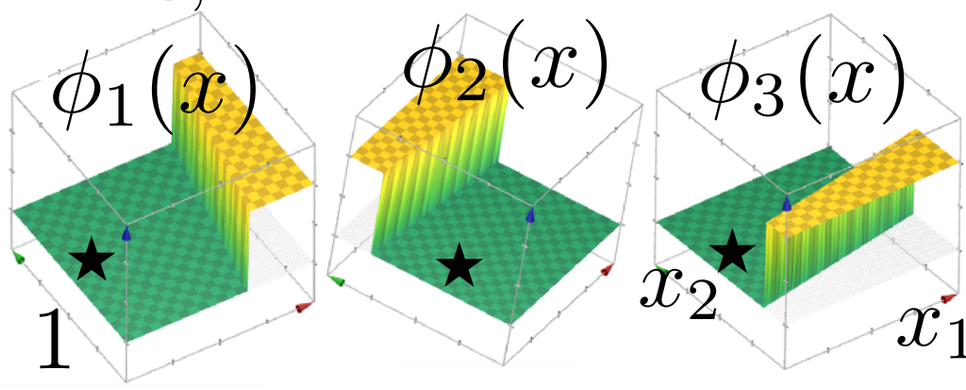


# Let's get some new notation

- 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:
      - $A^{(1)} = f^{(1)}(W^{(1)\top} x + W_0^{(1)})$
      - $W^{(1)} : m^{(1)} \times n^{(1)}; W_0^{(1)} : n^{(1)} \times 1$

$m^{(1)} = d$

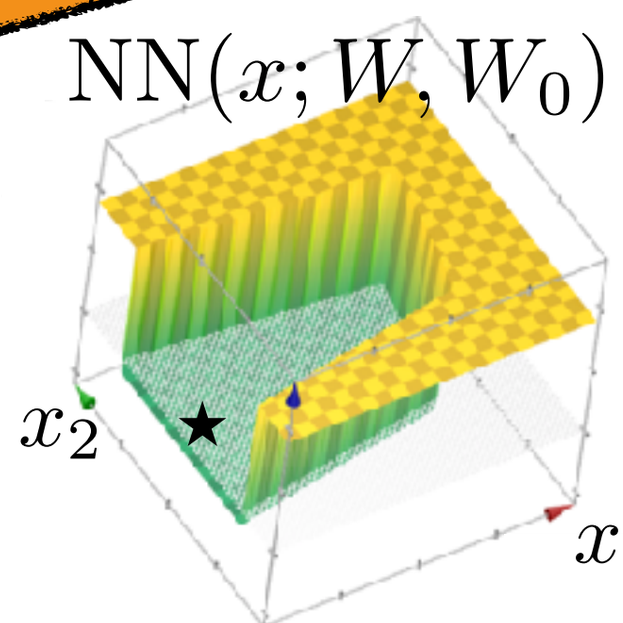



- 2nd layer, assigning a label (or labels):
  - Input (the features): size $m^{(2)} \times 1$
  - Output $A^{(2)}$(vector of labels): size $n^{(2)} \times 1$
  - The $i$th label: $A_i^{(2)} = f^{(2)}(w_i^{(2)\top} A^{(1)} + w_{0,i}^{(2)})$
    - All: $A^{(2)} = f^{(2)}(W^{(2)\top} A^{(1)} + W_0^{(2)})$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

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

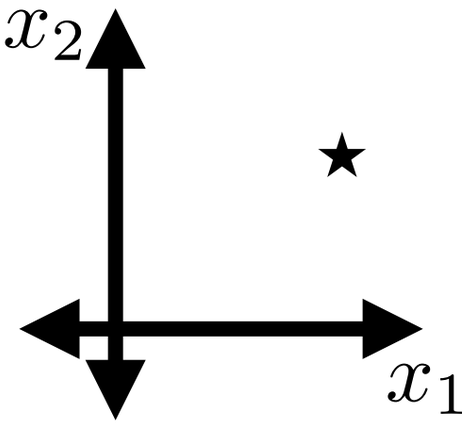
Like $y$

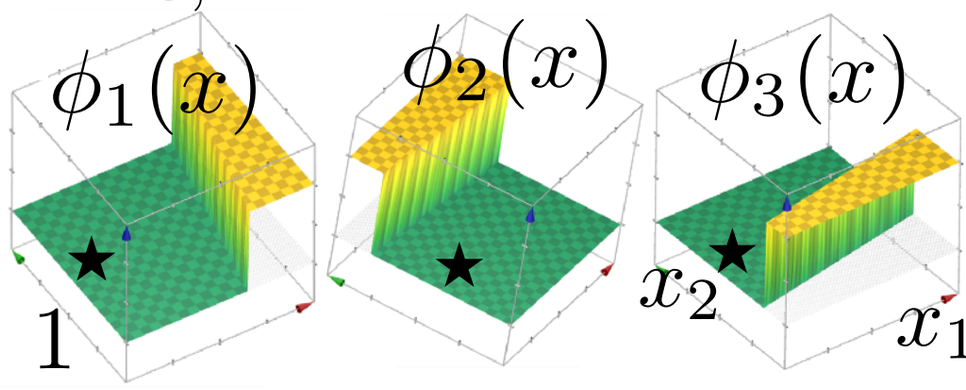


# Let's get some new notation

- 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:
      - $A^{(1)} = f^{(1)}(W^{(1)\top} x + W_0^{(1)})$
      - $W^{(1)} : m^{(1)} \times n^{(1)}; W_0^{(1)} : n^{(1)} \times 1$

$m^{(1)} = d$

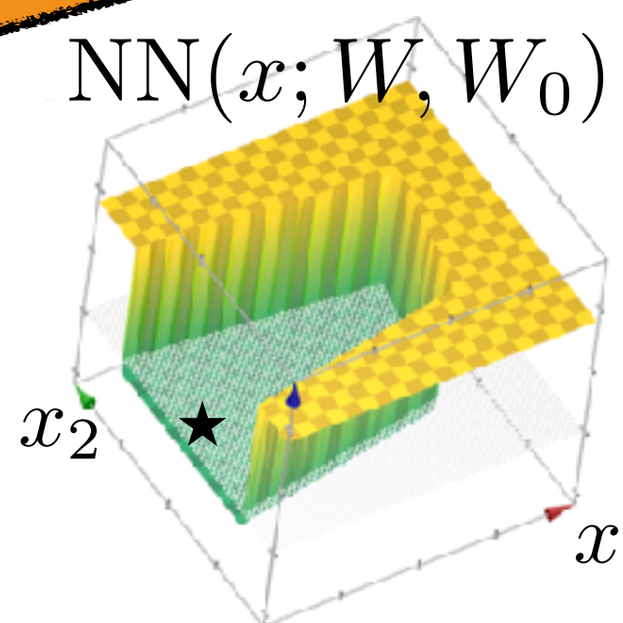



- 2nd layer, assigning a label (or labels):
  - Input (the features): size $m^{(2)} \times 1$
  - Output $A^{(2)}$(vector of labels): size $n^{(2)} \times 1$
  - The $i$th label: $A_i^{(2)} = f^{(2)}(w_i^{(2)\top} A^{(1)} + w_{0,i}^{(2)})$
    - All: $A^{(2)} = f^{(2)}(W^{(2)\top} A^{(1)} + W_0^{(2)})$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

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

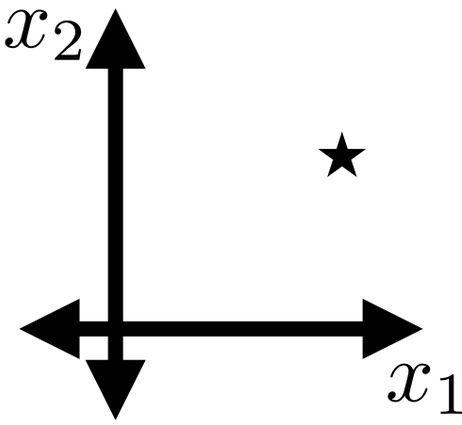
Like $y$

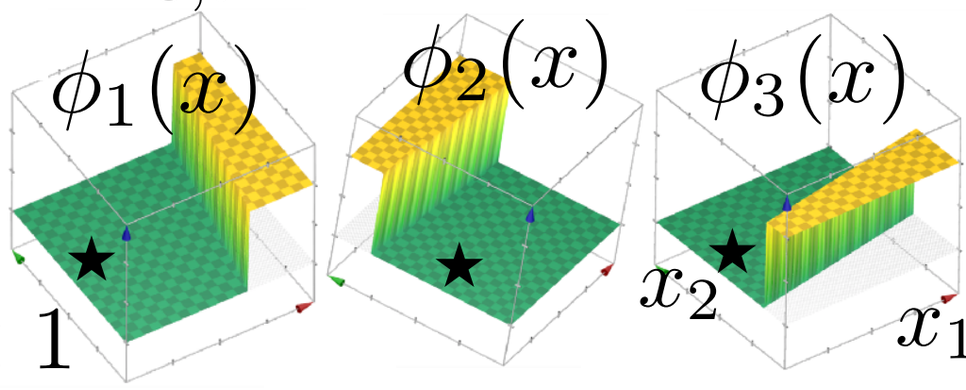


# Let's get some new notation

- 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:
      - $A^{(1)} = f^{(1)}(W^{(1)\top} x + W_0^{(1)})$
      - $W^{(1)} : m^{(1)} \times n^{(1)}; W_0^{(1)} : n^{(1)} \times 1$

$m^{(1)} = d$

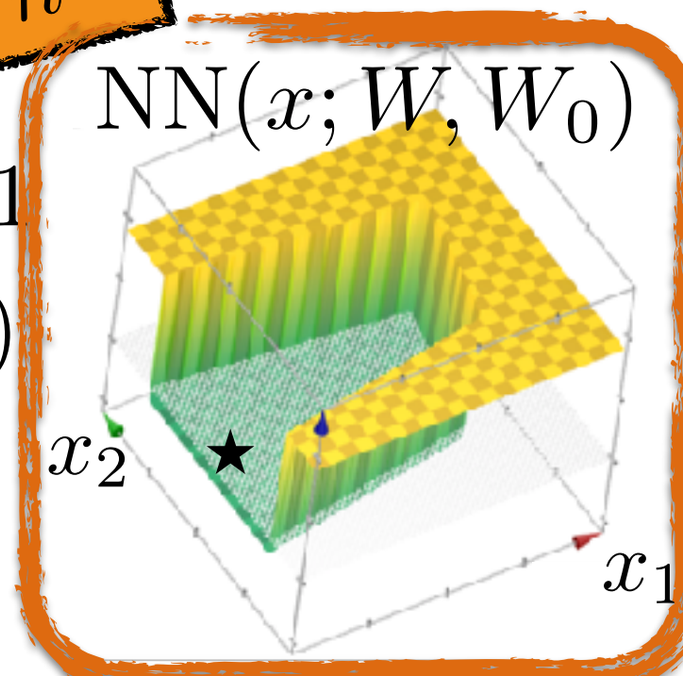



- 2nd layer, assigning a label (or labels):
  - Input (the features): size $m^{(2)} \times 1$
  - Output $A^{(2)}$(vector of labels): size $n^{(2)} \times 1$
  - The $i$th label: $A_i^{(2)} = f^{(2)}(w_i^{(2)\top} A^{(1)} + w_{0,i}^{(2)})$
    - All: $A^{(2)} = f^{(2)}(W^{(2)\top} A^{(1)} + W_0^{(2)})$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

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

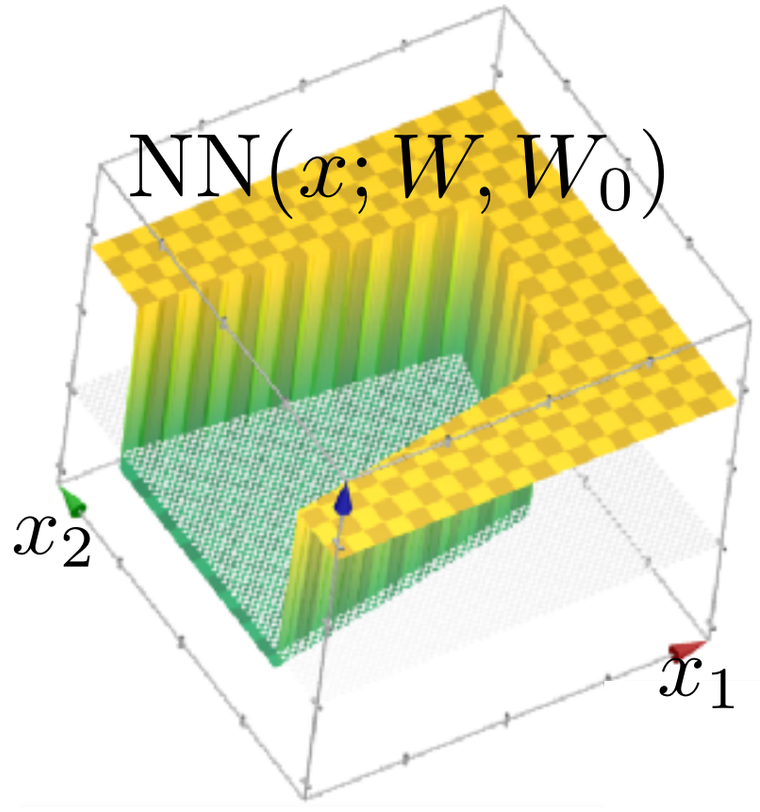
Like $y$



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

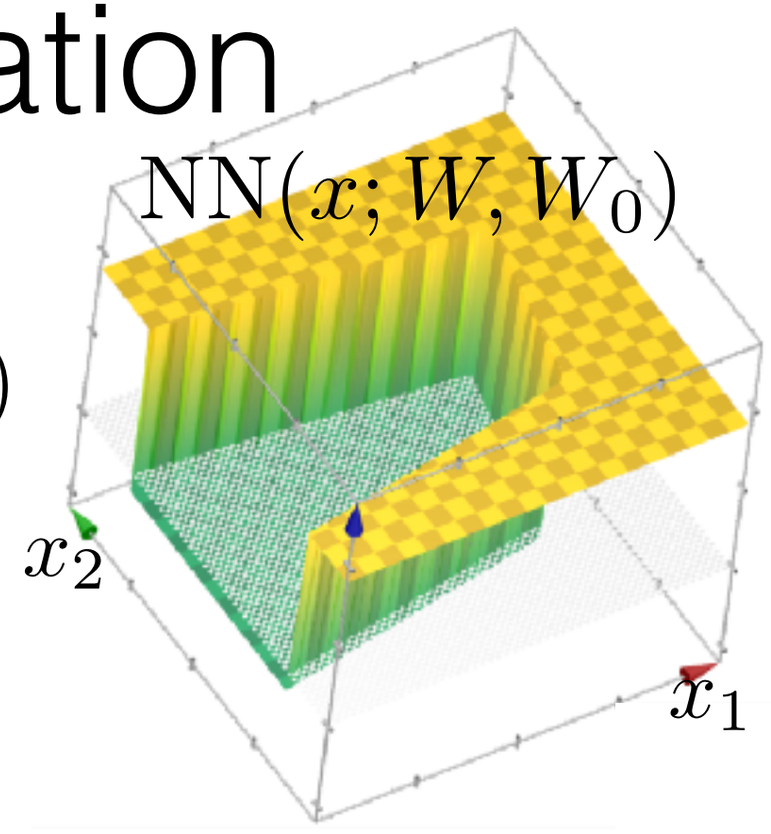
- 1st layer: $A^{(1)} = f^{(1)}(W^{(1)\top} x + W_0^{(1)})$
- 2nd layer: $A^{(2)} = f^{(2)}(W^{(2)\top} A^{(1)} + W_0^{(2)})$
- Whole thing: $A^{(2)} = \mathrm{NN}(x; W, W_0)$

- 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: $A^{(2)} = \mathrm{NN}(x; W, W_0)$



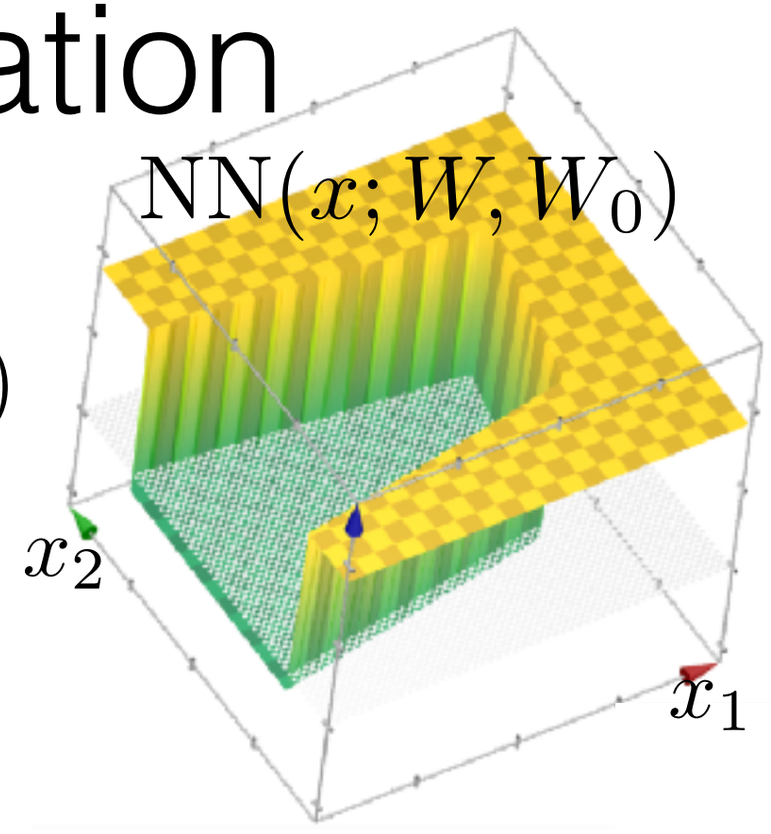
# Function graph representation

- 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: $A^{(2)} = \mathrm{NN}(x; W, W_0)$



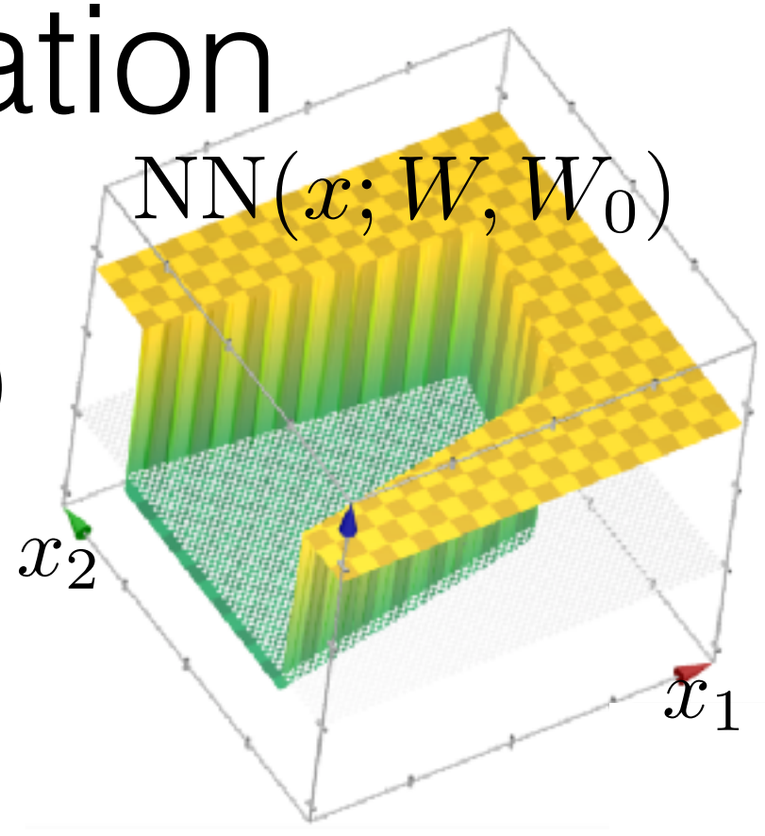
# Function graph representation

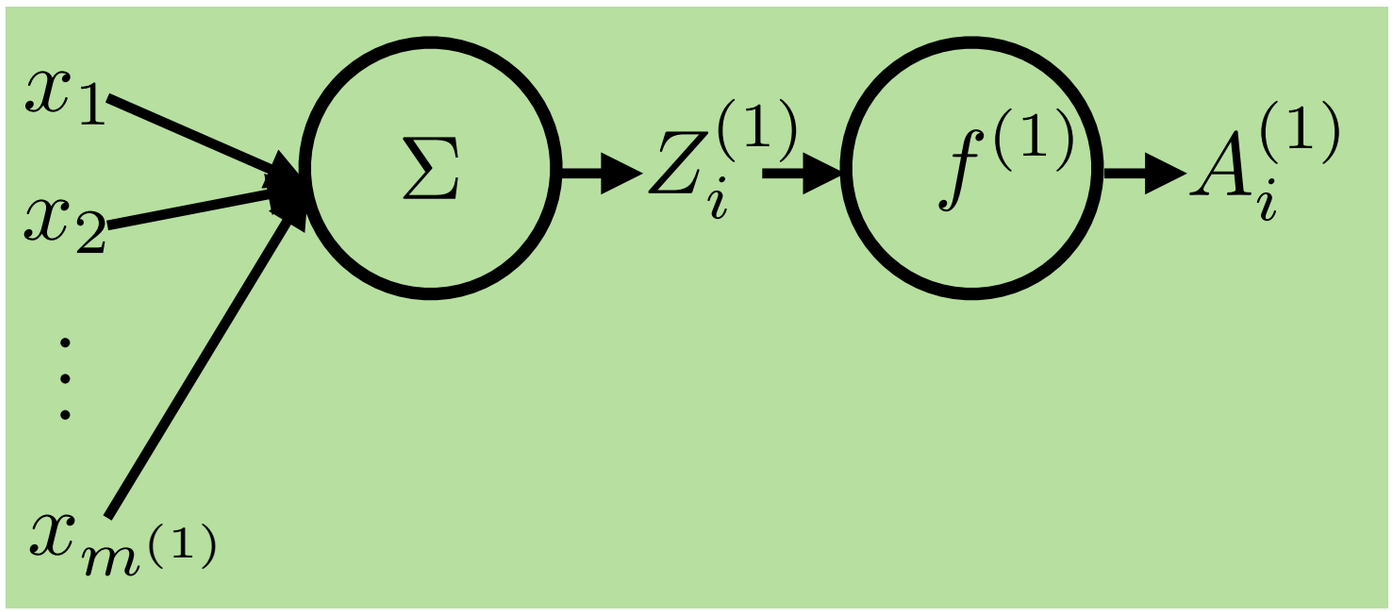
- 1st layer: $A^{(1)} = f^{(1)}(W^{(1)\top} x + W_0^{(1)})$
- 2nd layer: $A^{(2)} = f^{(2)}(W^{(2)\top} A^{(1)} + W_0^{(2)})$
- Whole thing: $A^{(2)} = \mathrm{NN}(x; W, W_0)$



# Function graph representation

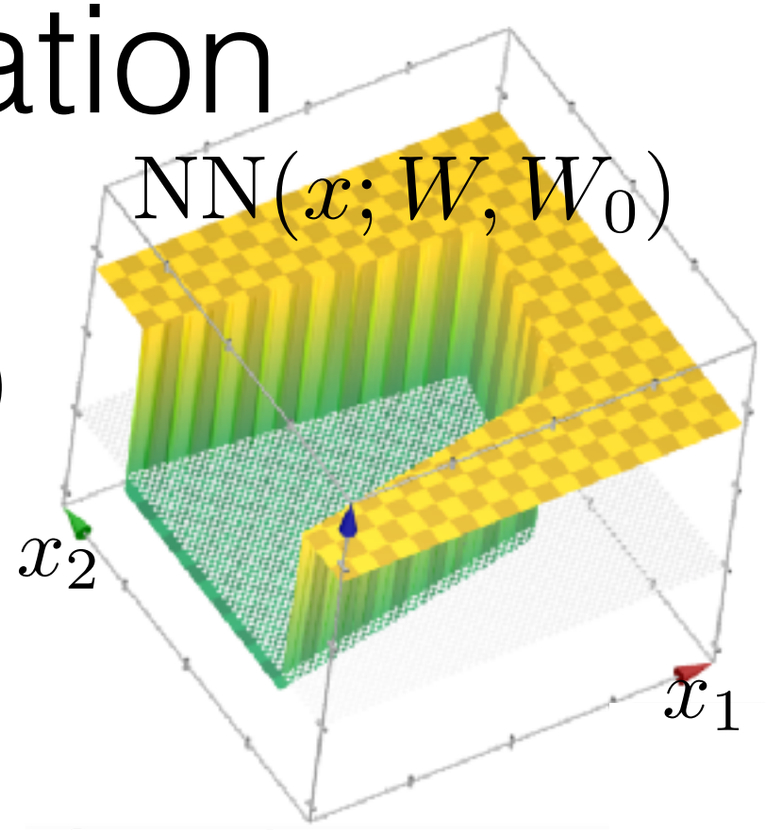
- 1st layer: $A^{(1)} = f^{(1)}(W^{(1)\top} x + W_0^{(1)})$
- 2nd layer: $A^{(2)} = f^{(2)}(W^{(2)\top} A^{(1)} + W_0^{(2)})$
- Whole thing: $A^{(2)} = \mathrm{NN}(x; W, W_0)$



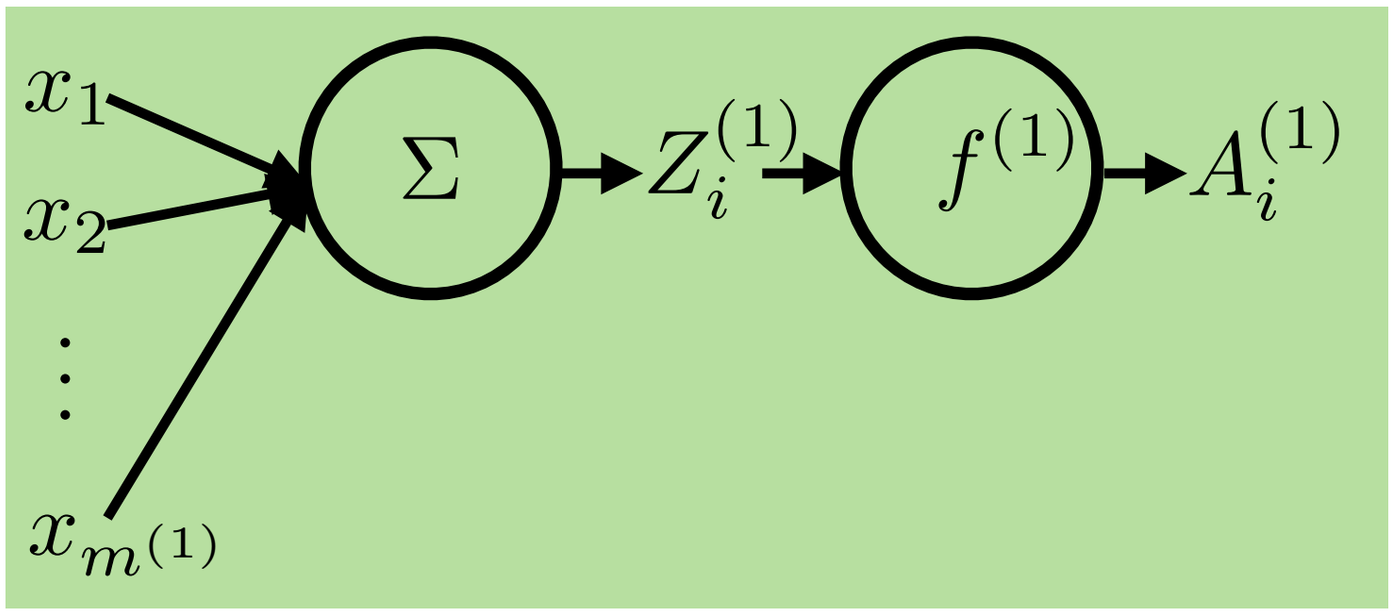



# Function graph representation

- 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: $A^{(2)} = \mathrm{NN}(x; W, W_0)$

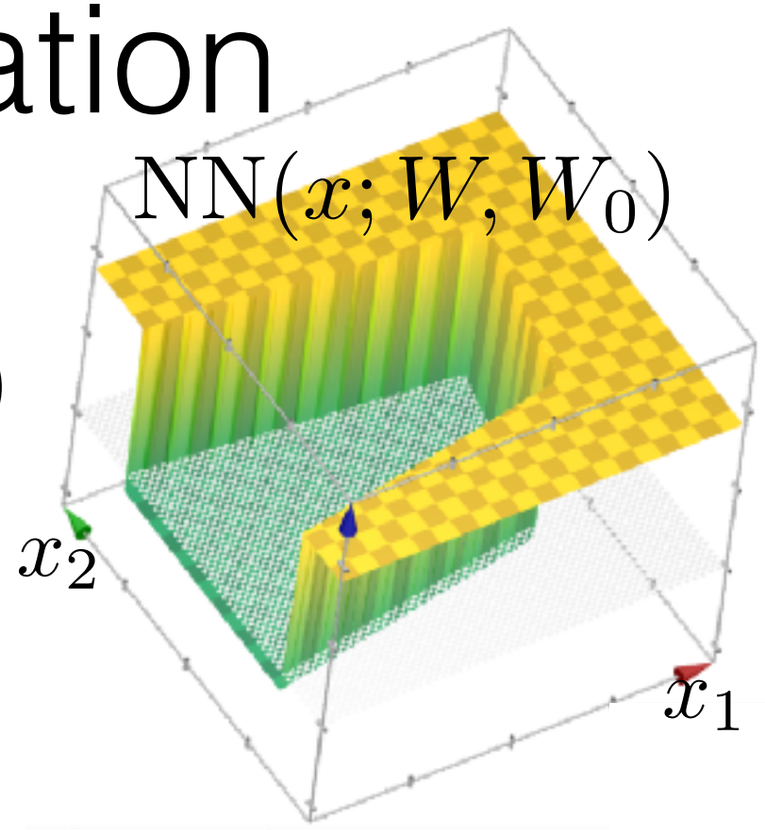



- Circle: function evaluation

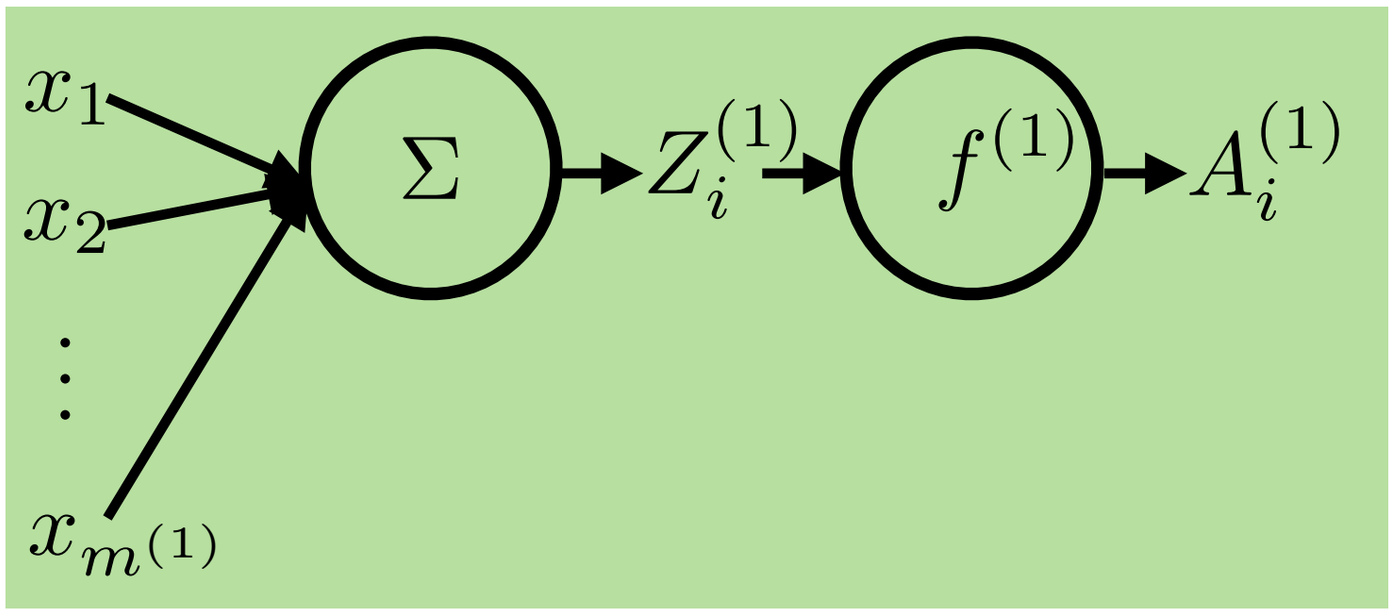


# Function graph representation

- 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: $A^{(2)} = \mathrm{NN}(x; W, W_0)$

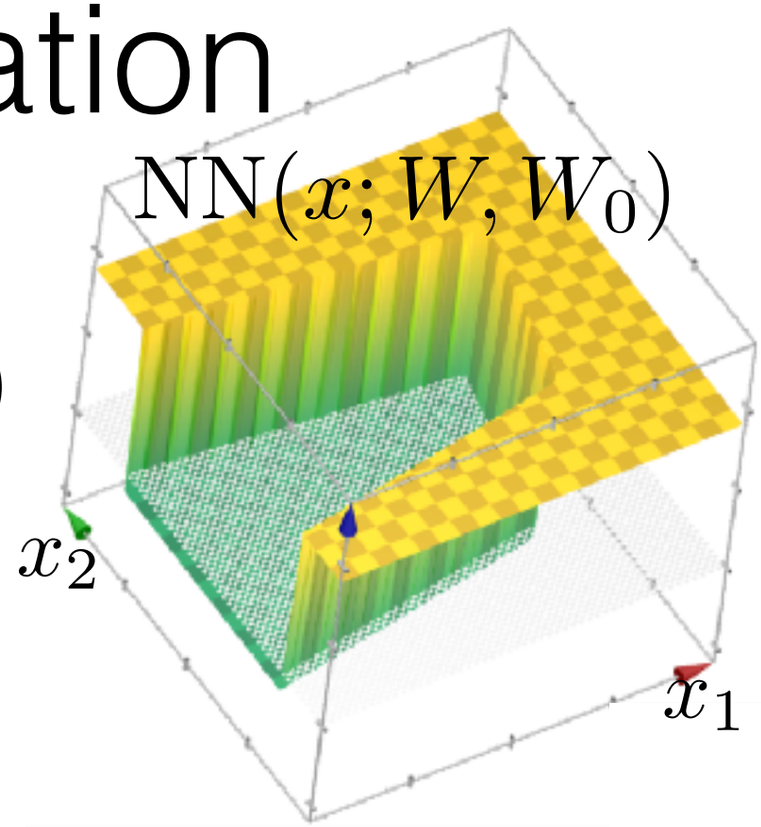


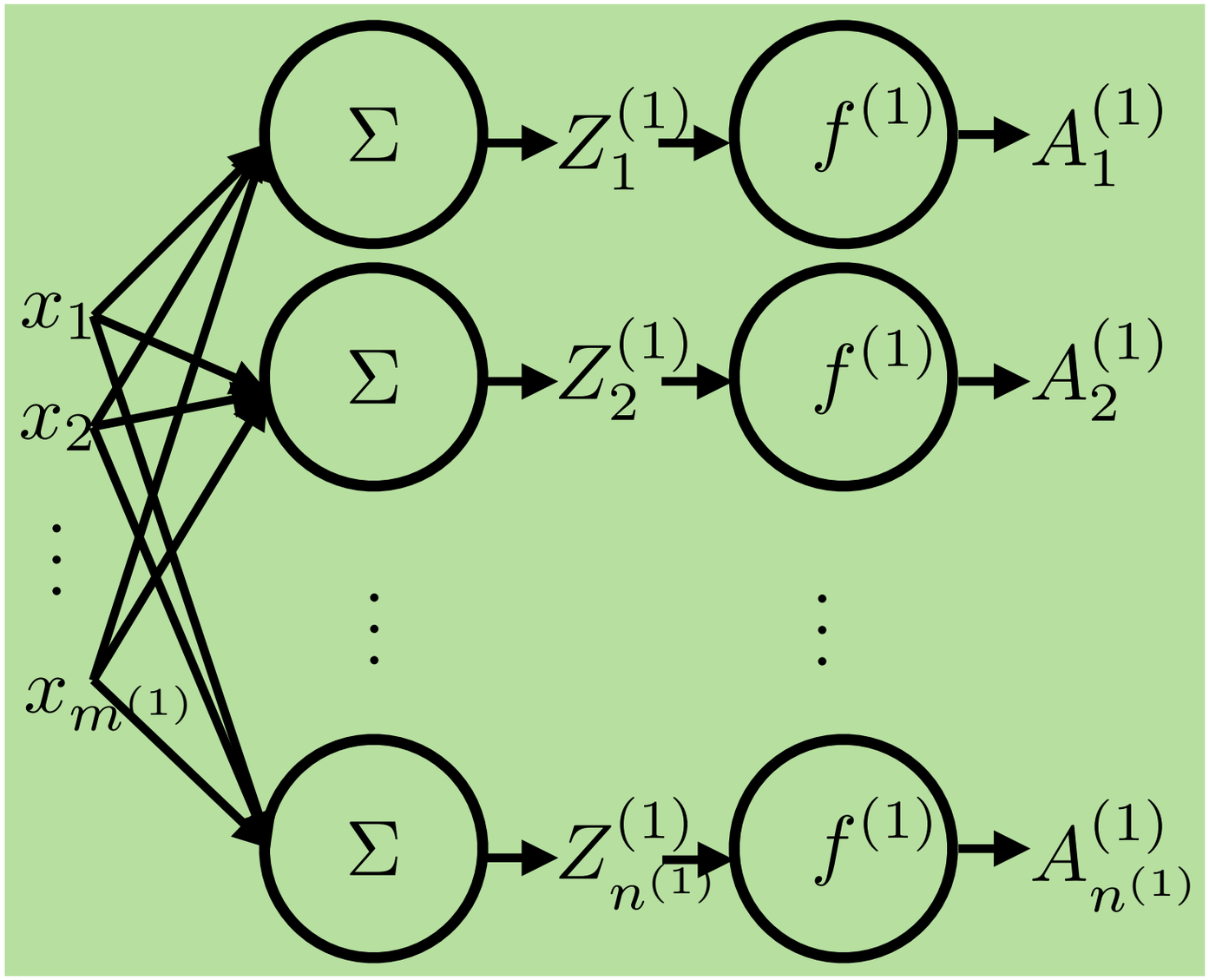

- Circle: function evaluation



# 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: $A^{(2)} = \mathrm{NN}(x; W, W_0)$

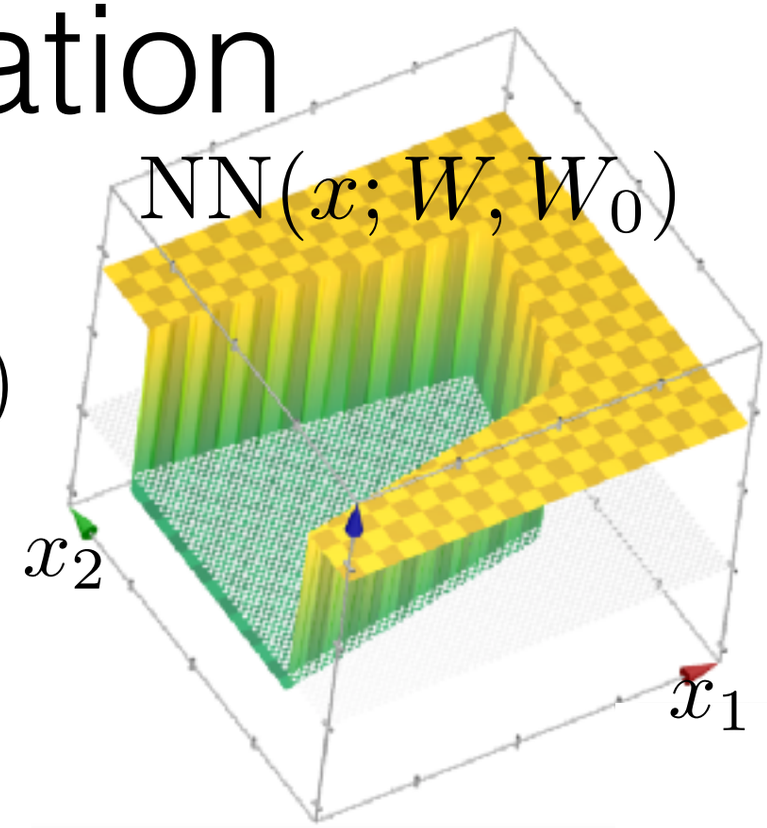


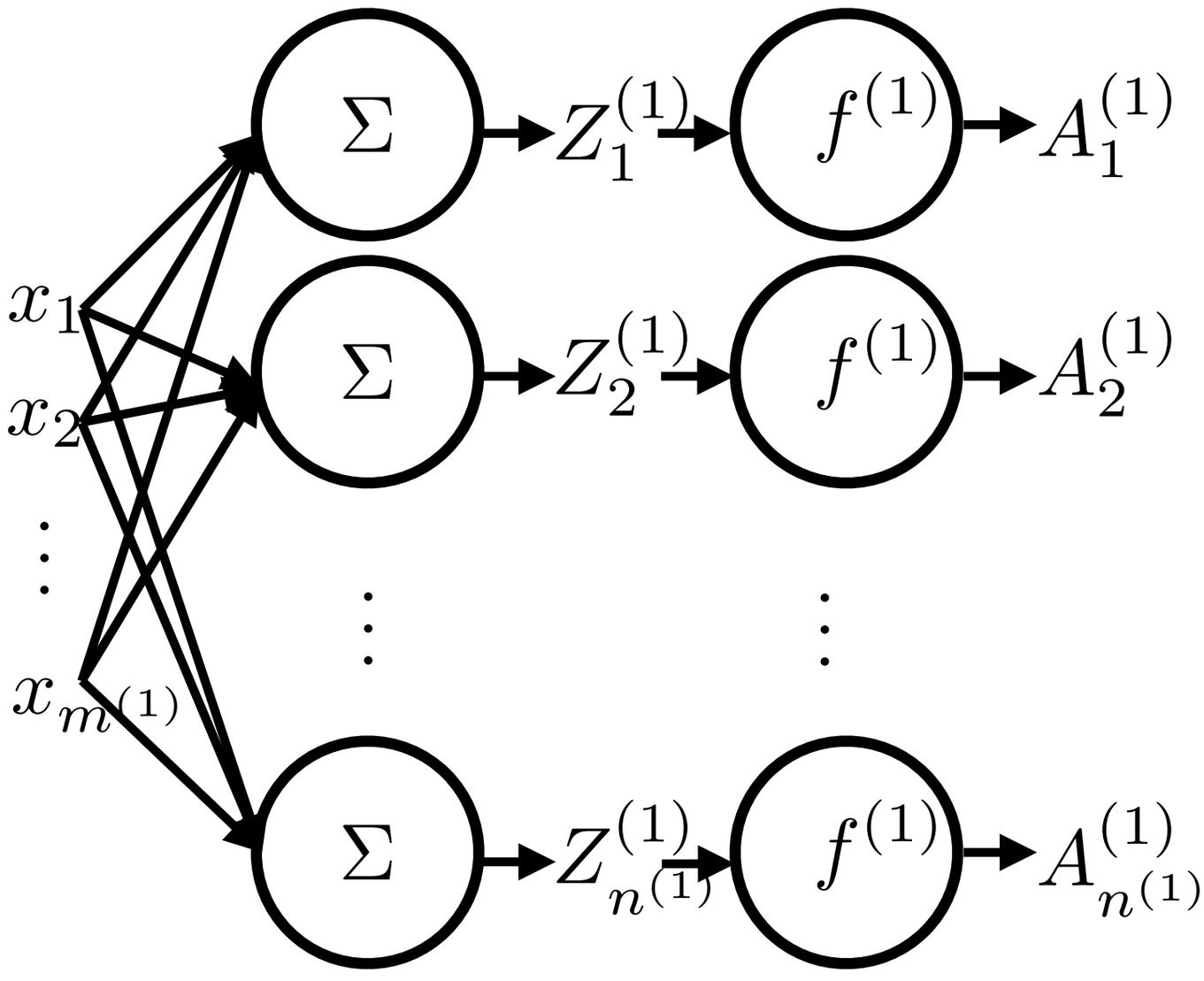




- Circle: function evaluation

# 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: $A^{(2)} = \mathrm{NN}(x; W, W_0)$

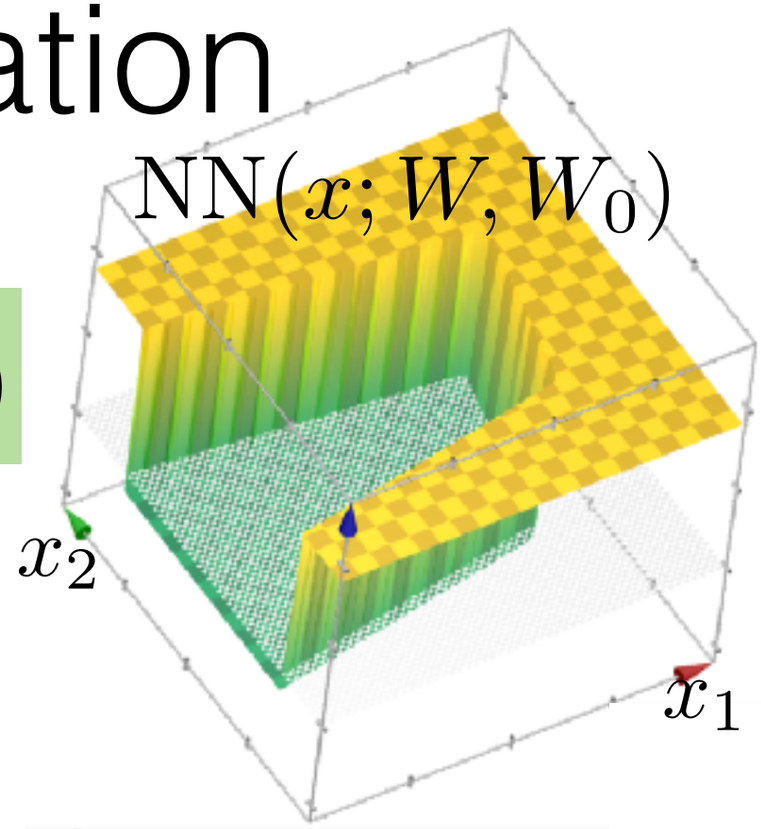


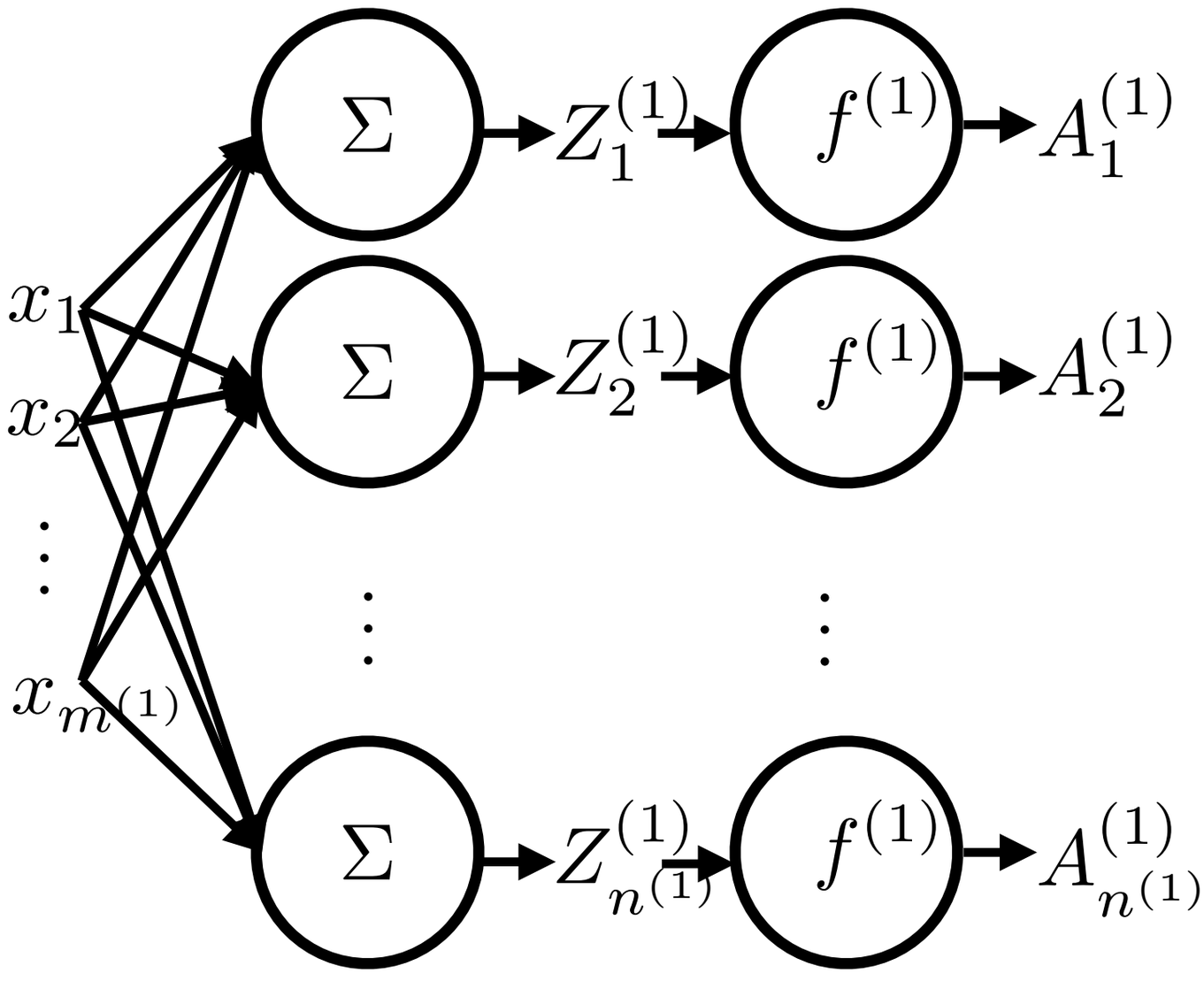




- Circle: function evaluation

# 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: $A^{(2)} = \mathrm{NN}(x; W, W_0)$

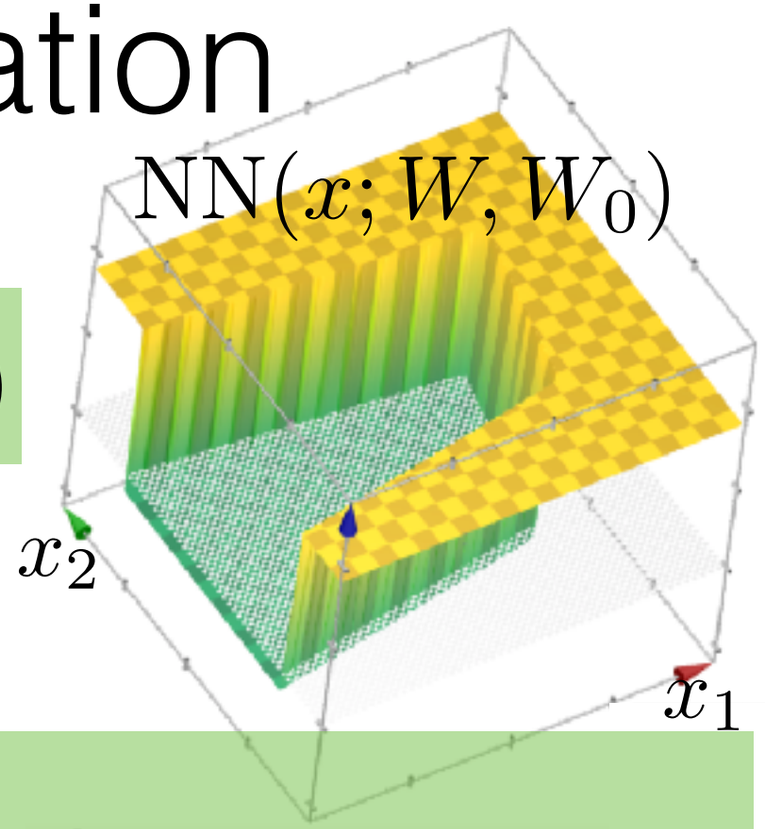


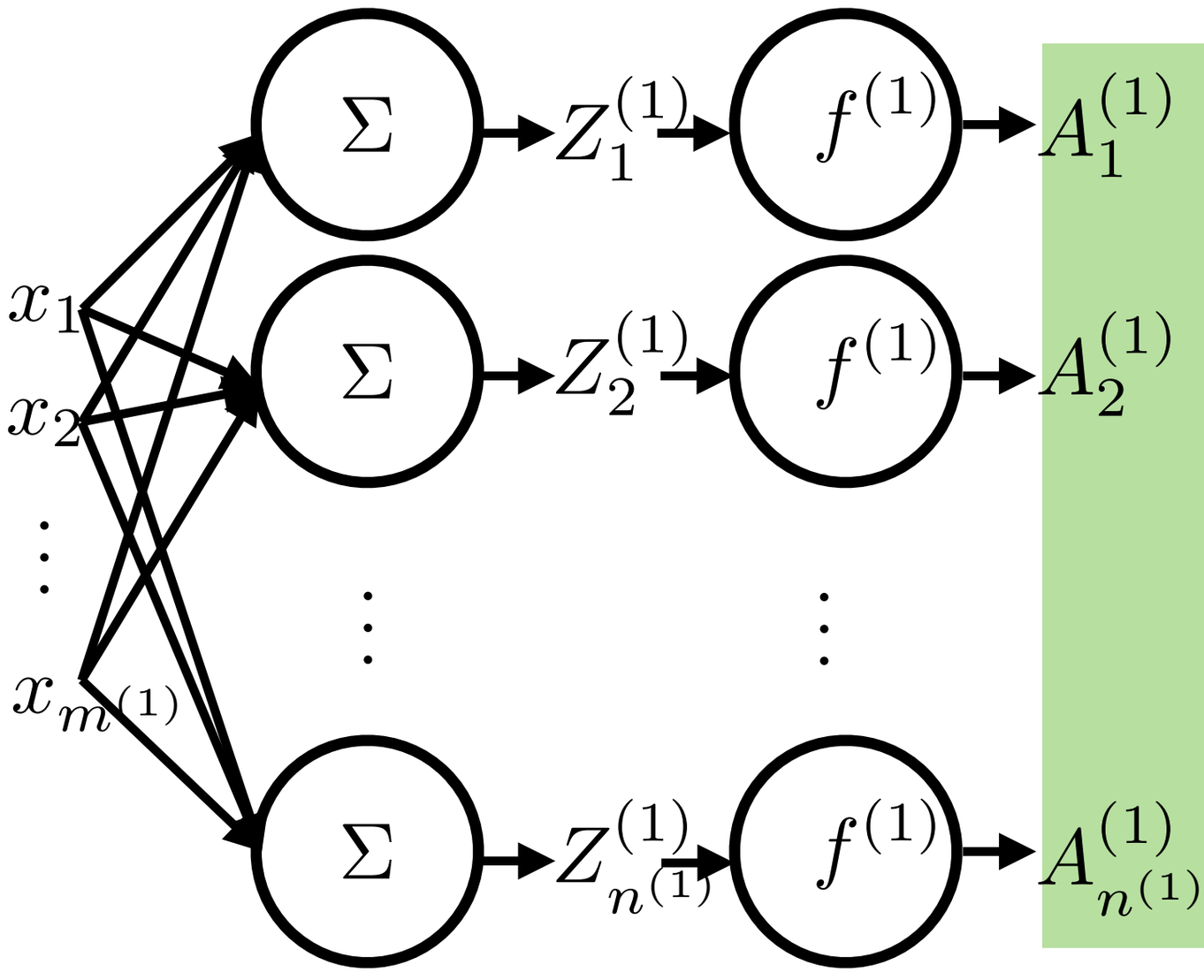




- Circle: function evaluation

# 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: $A^{(2)} = \mathrm{NN}(x; W, W_0)$

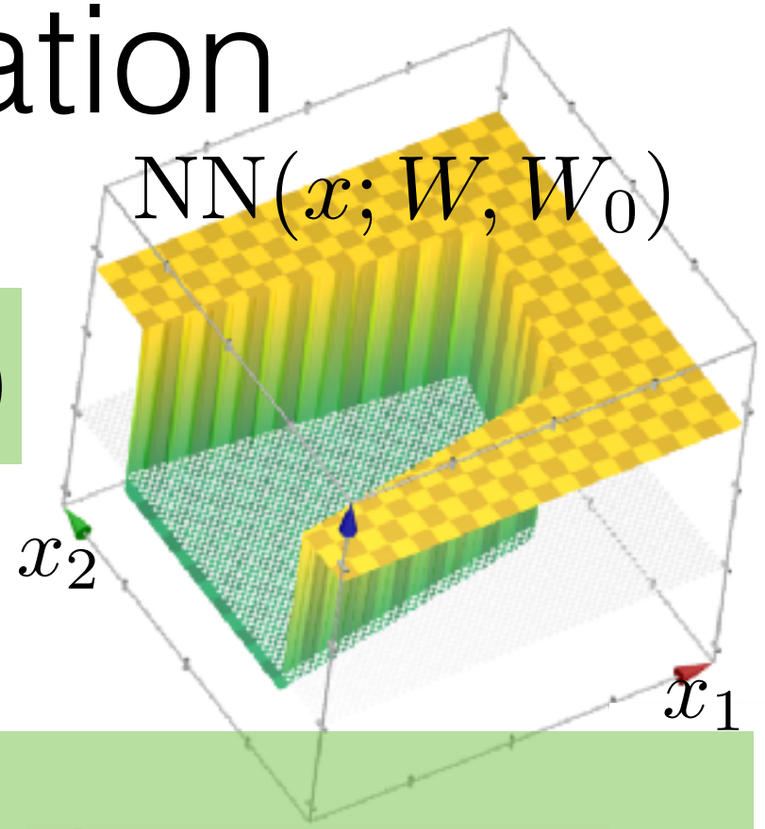


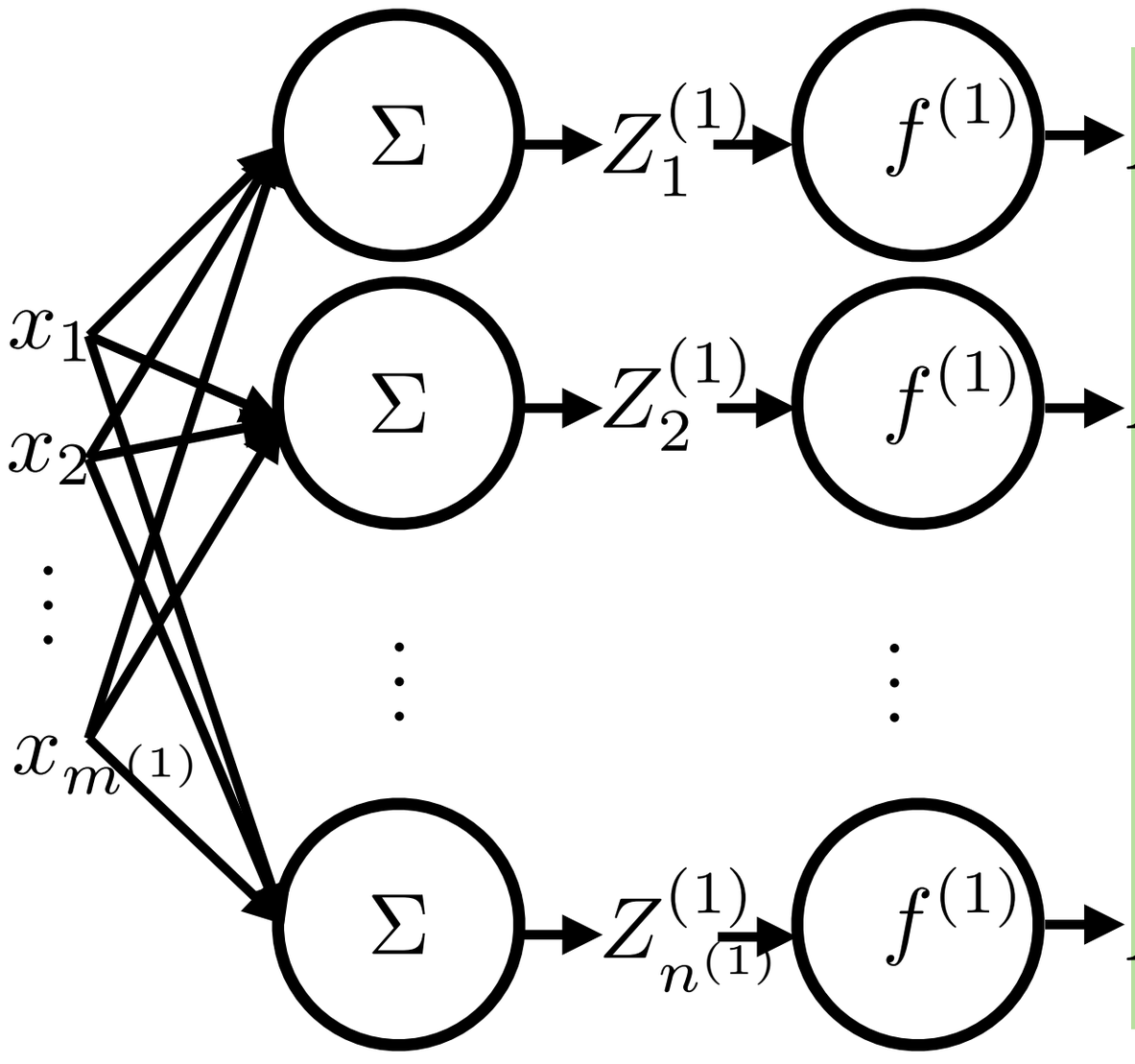




- Circle: function evaluation

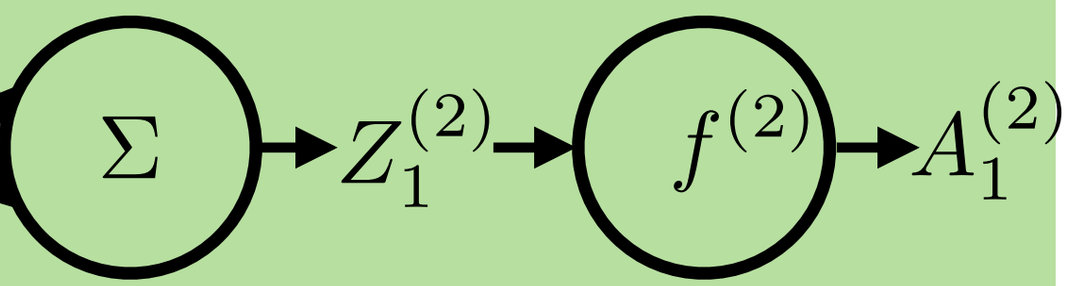
# Function graph representation

- 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: $A^{(2)} = \mathrm{NN}(x; W, W_0)$



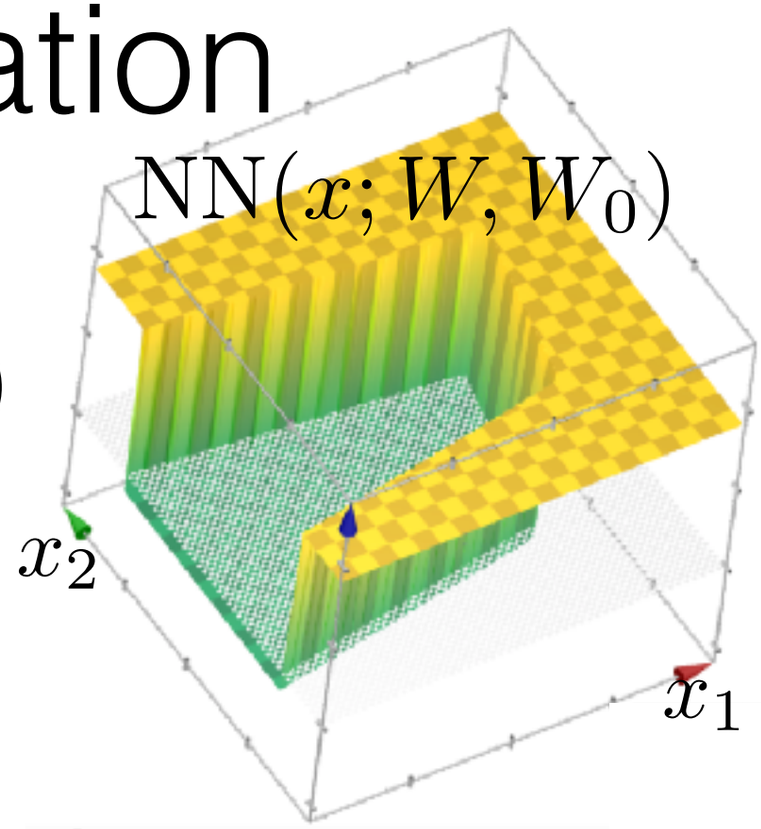



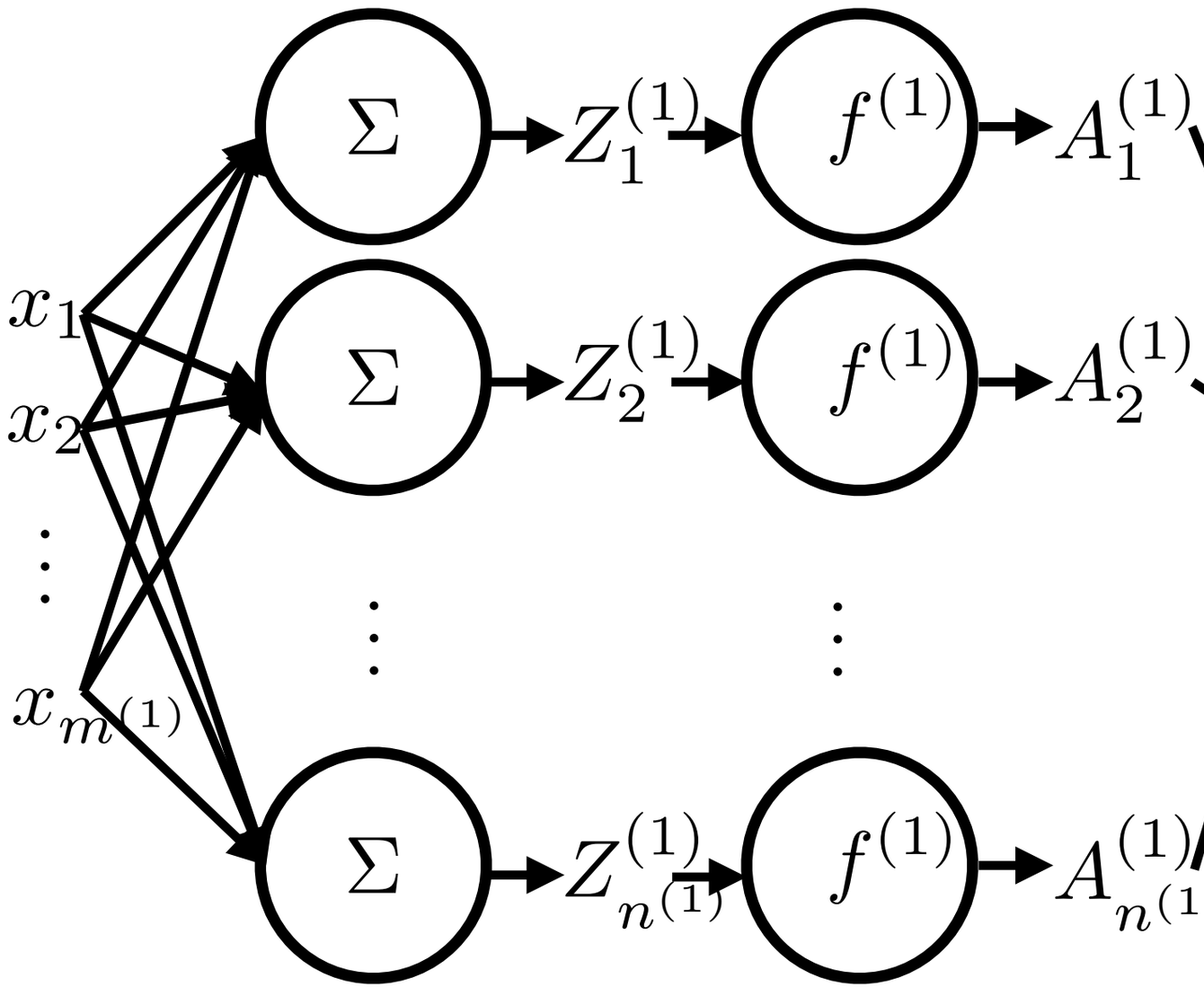

- Circle: function evaluation



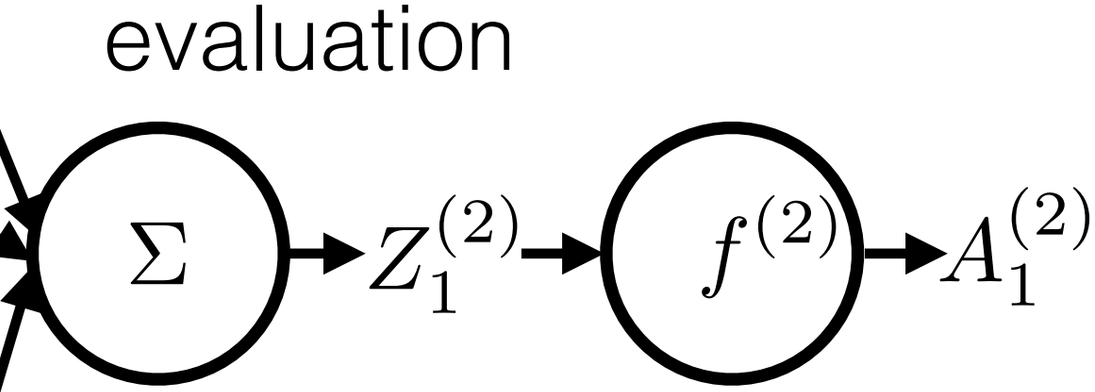
# Function graph representation

- 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: $A^{(2)} = \mathrm{NN}(x; W, W_0)$



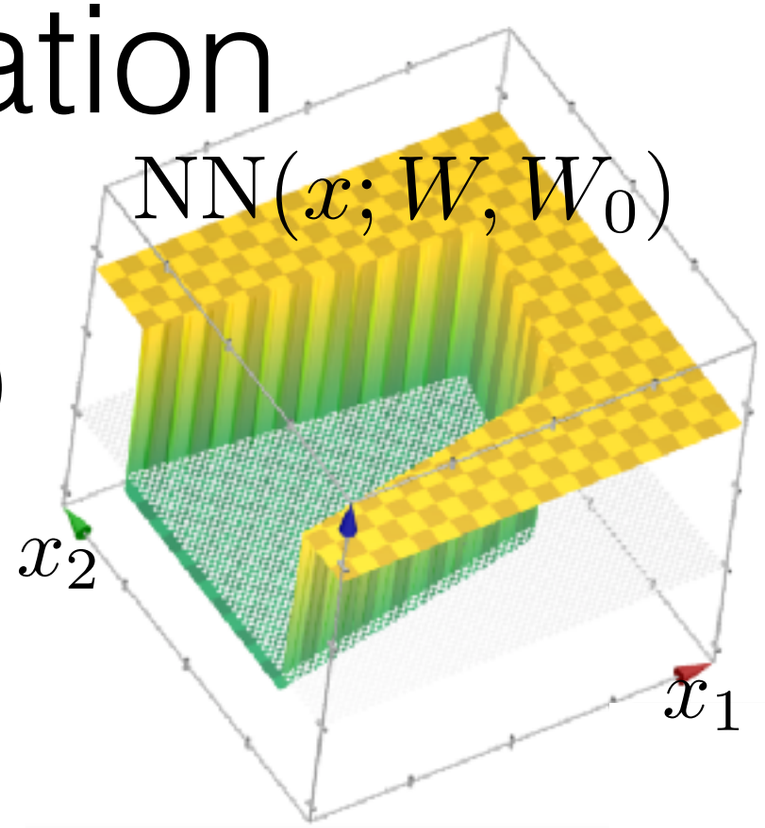



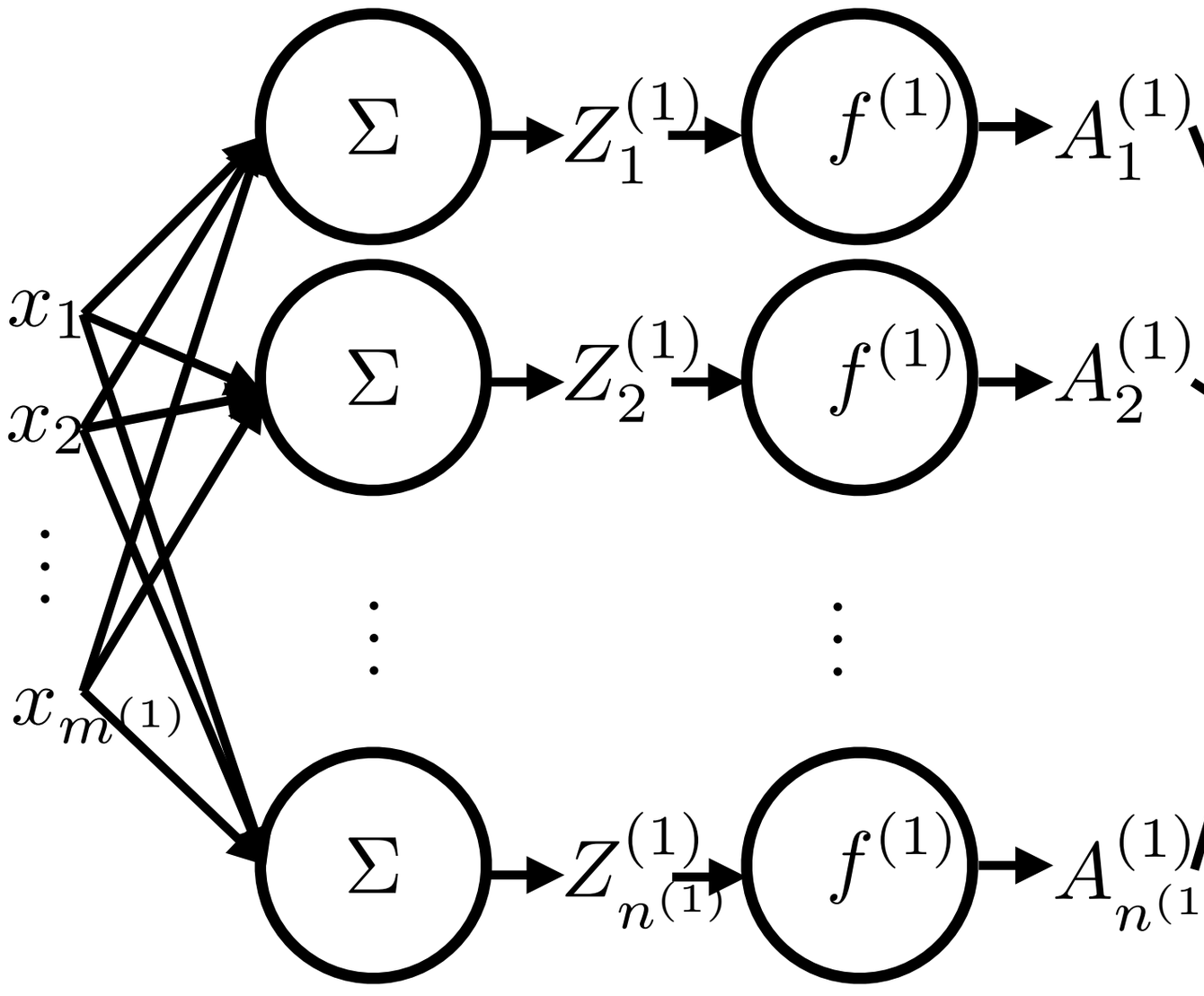

- Circle: function evaluation



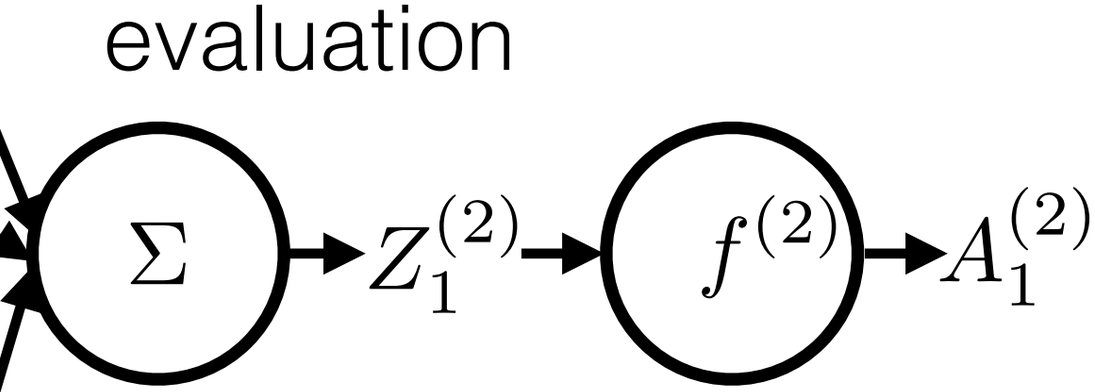
# Function graph representation

- 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: $A^{(2)} = \mathrm{NN}(x; W, W_0)$



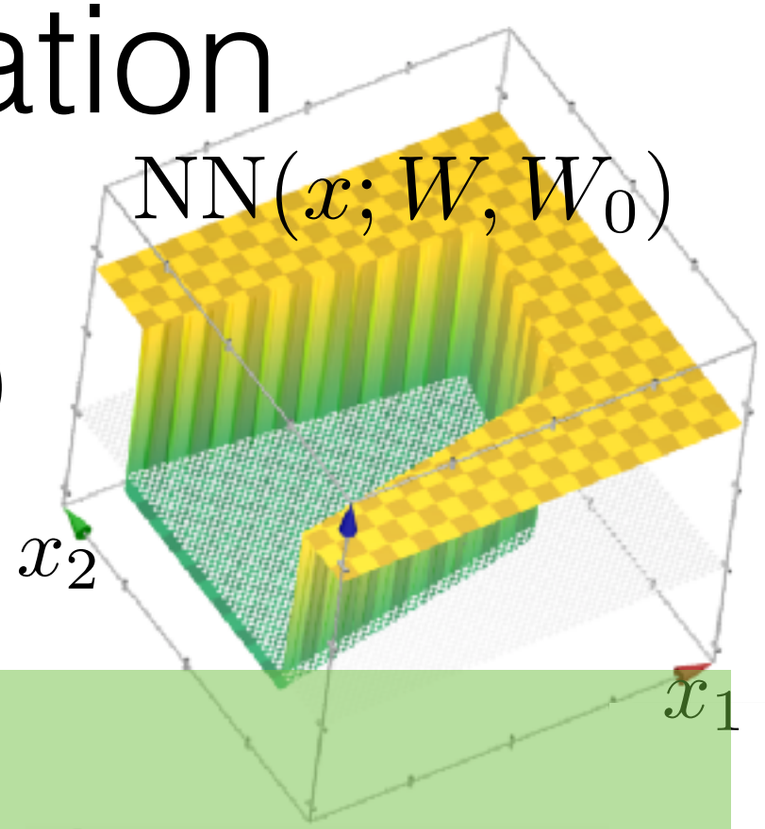



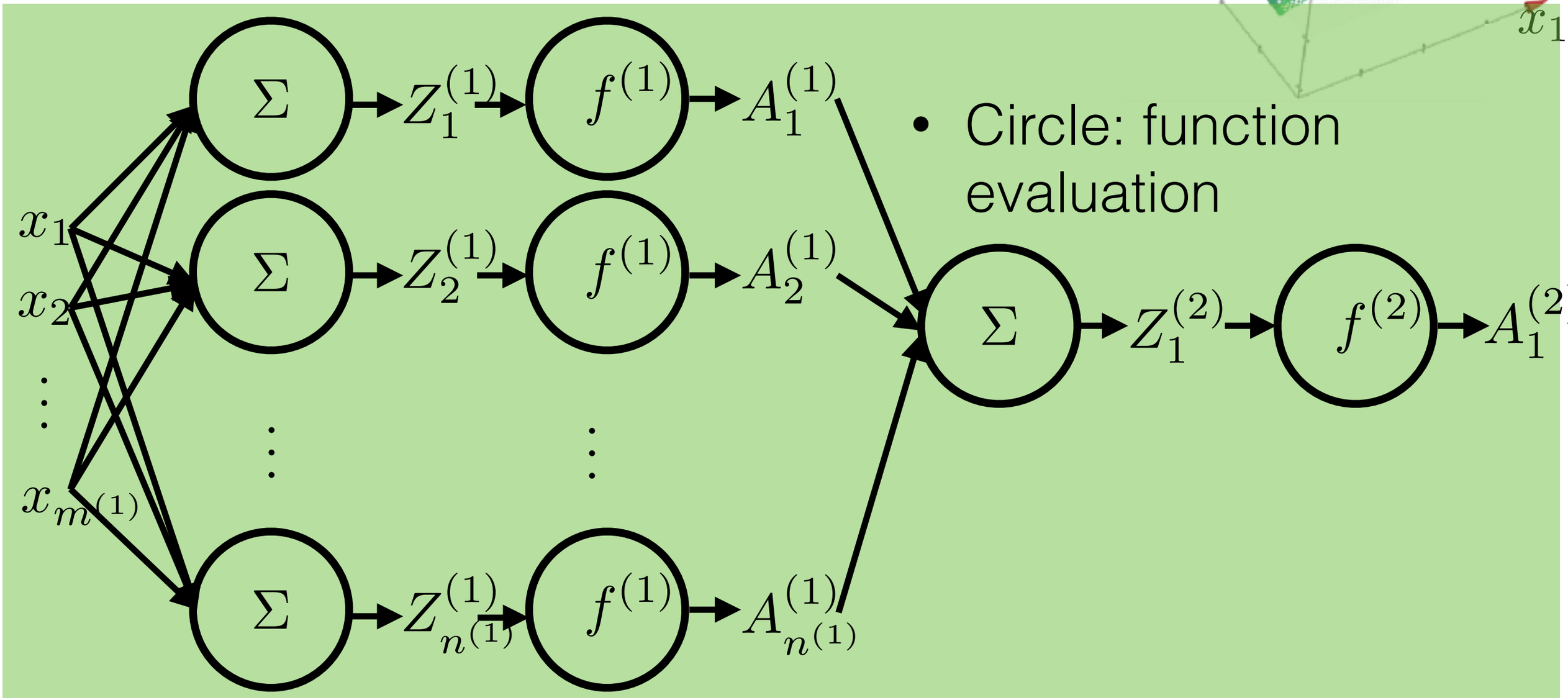

- Circle: function evaluation



# 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: $A^{(2)} = \mathrm{NN}(x; W, W_0)$

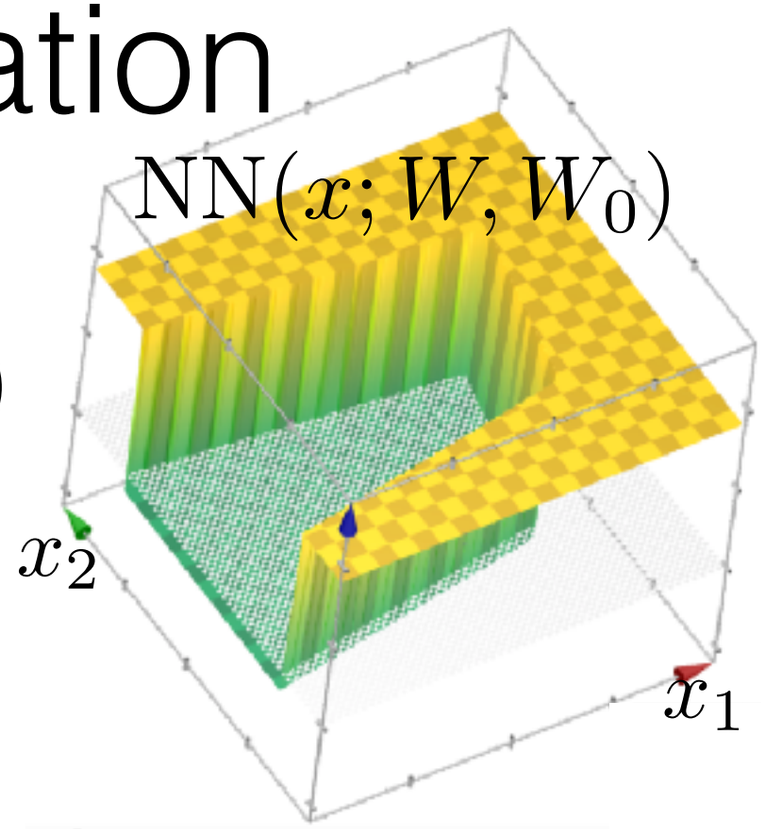


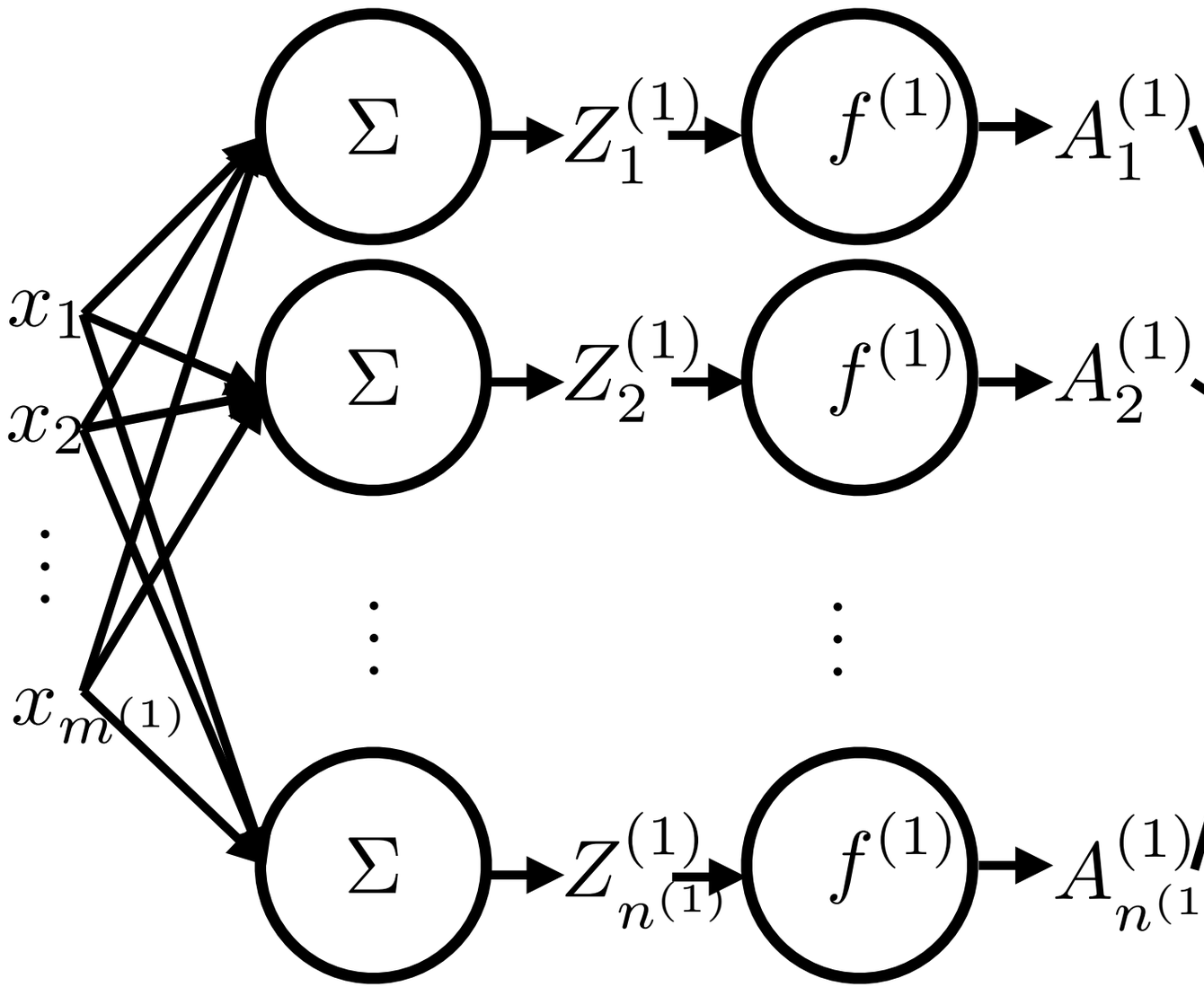




- Circle: function evaluation

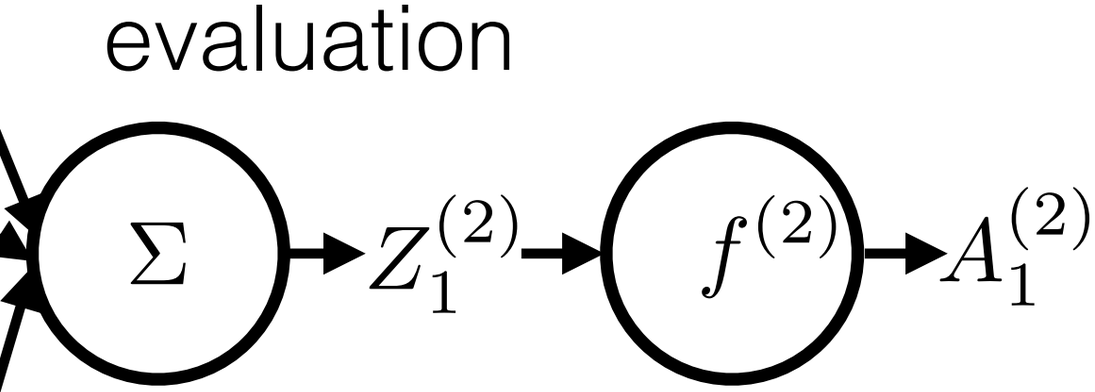
# Function graph representation

- 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: $A^{(2)} = \mathrm{NN}(x; W, W_0)$



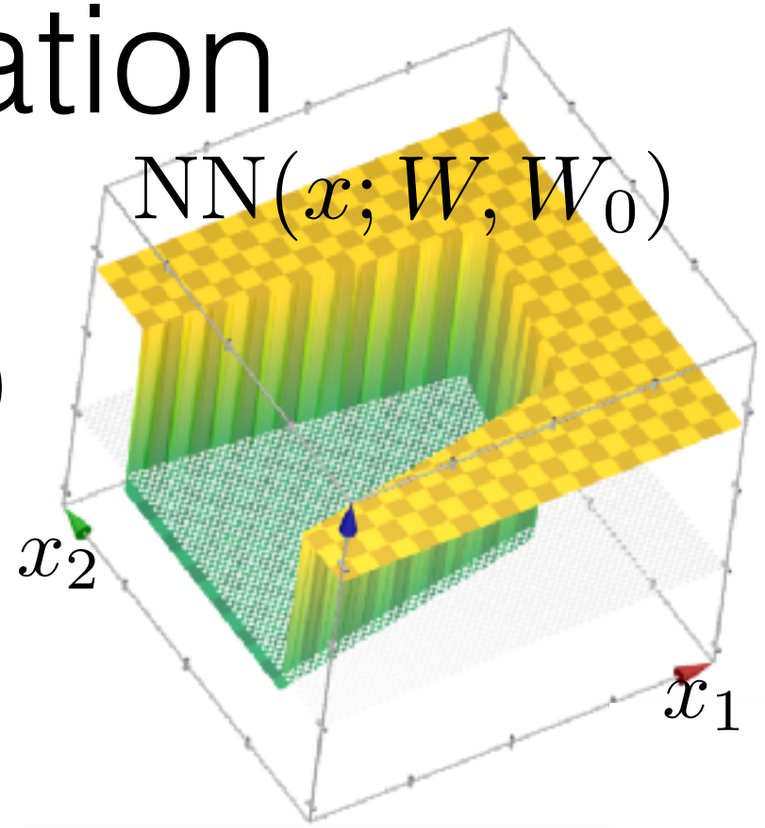



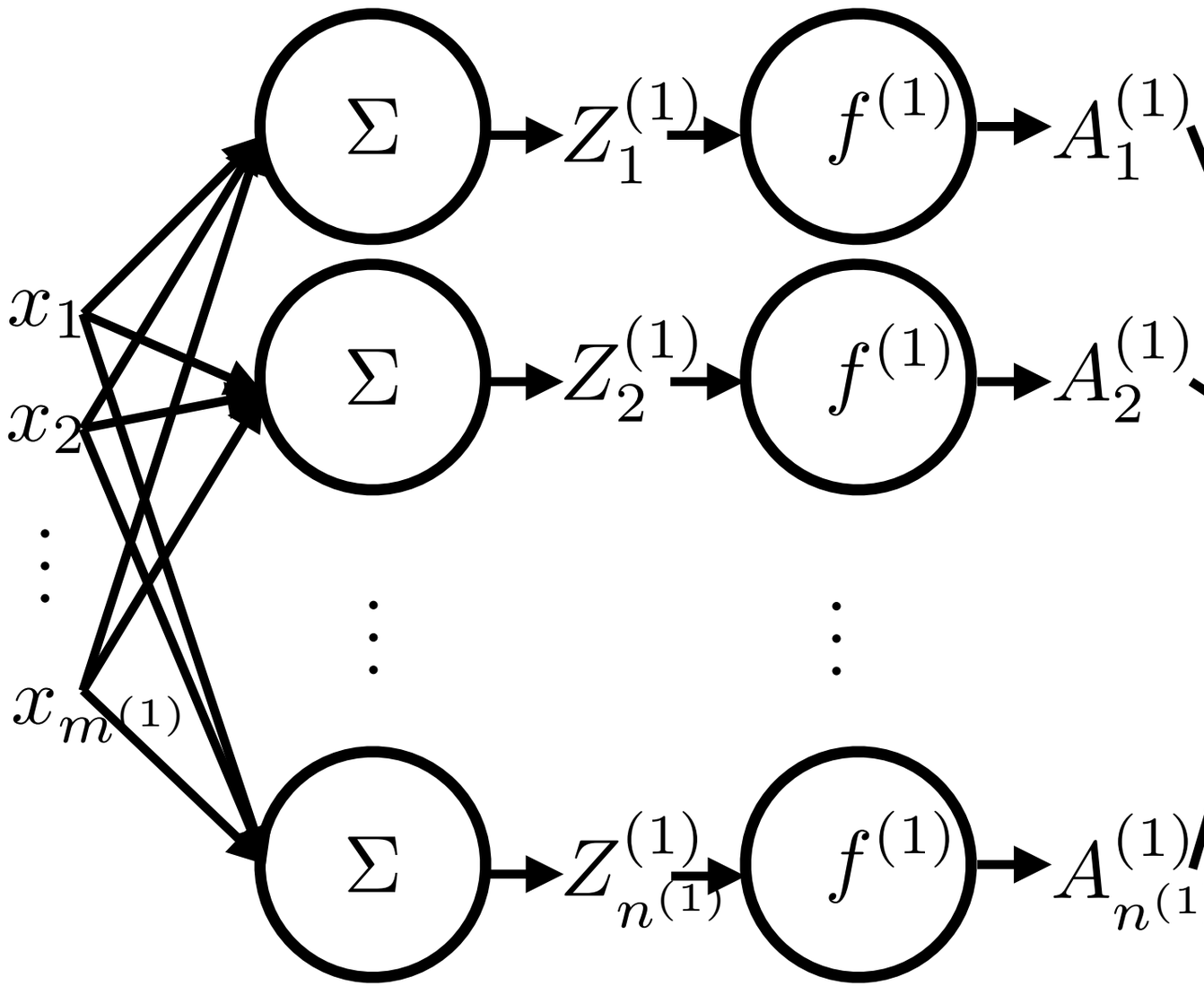

- Circle: function evaluation



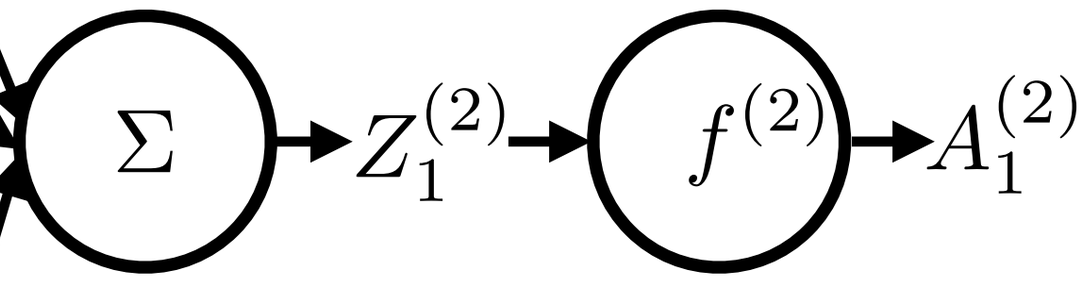
# Function graph representation

- 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: $A^{(2)} = \mathrm{NN}(x; W, W_0)$



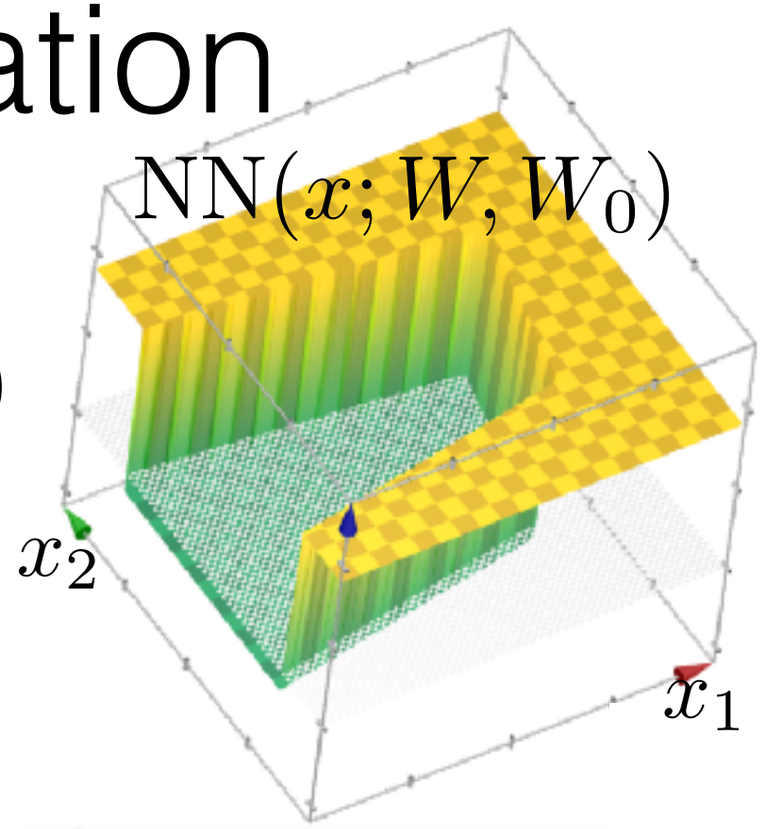

- Circle: function evaluation

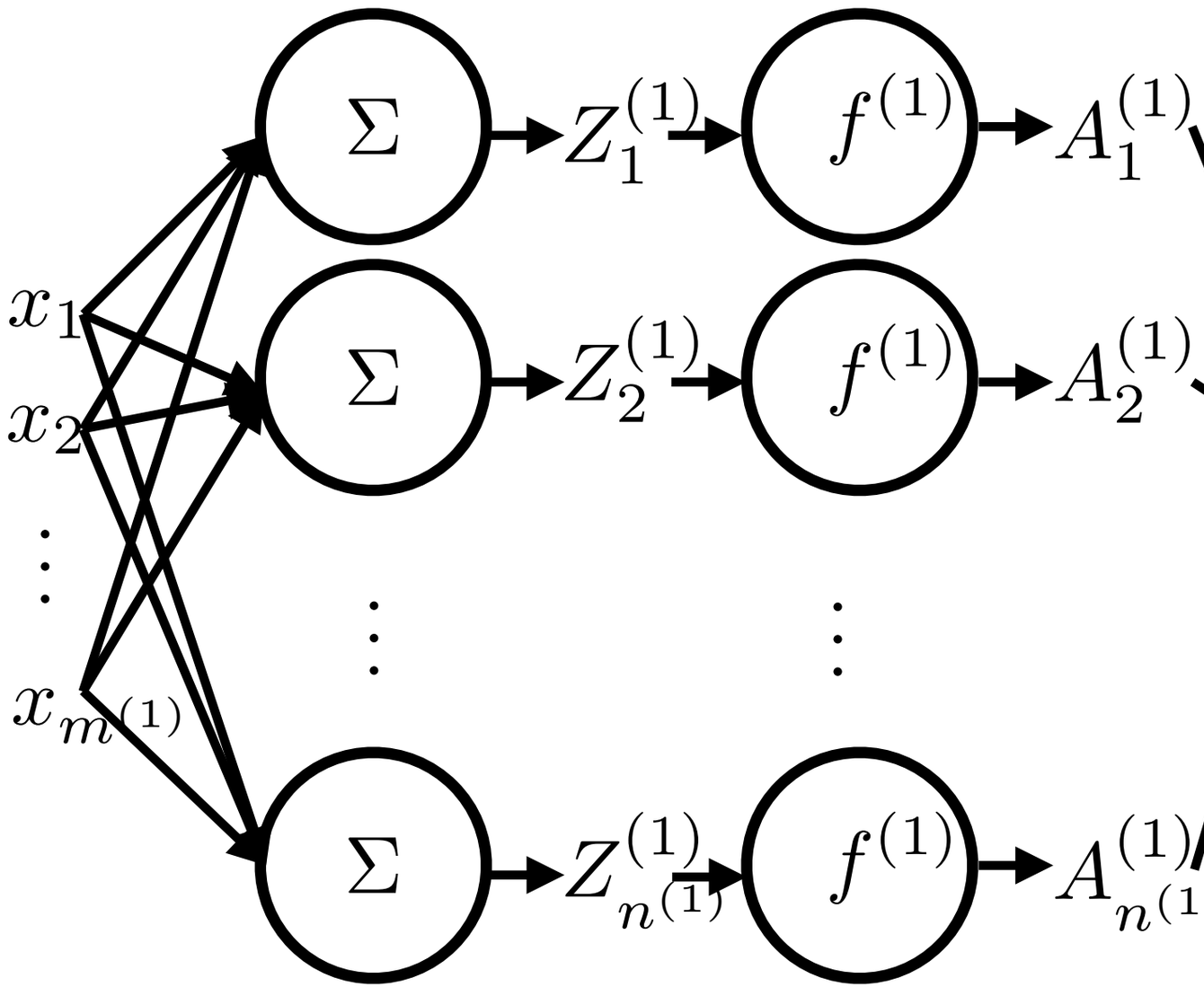



- Forward vs. backward

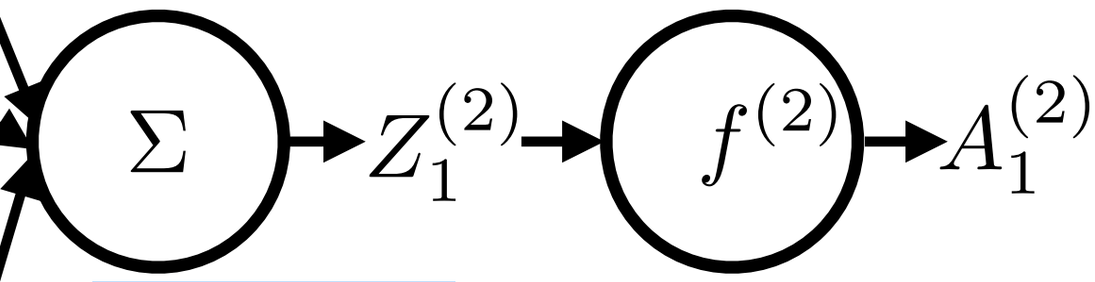
# Function graph representation

- 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: $A^{(2)} = \mathrm{NN}(x; W, W_0)$



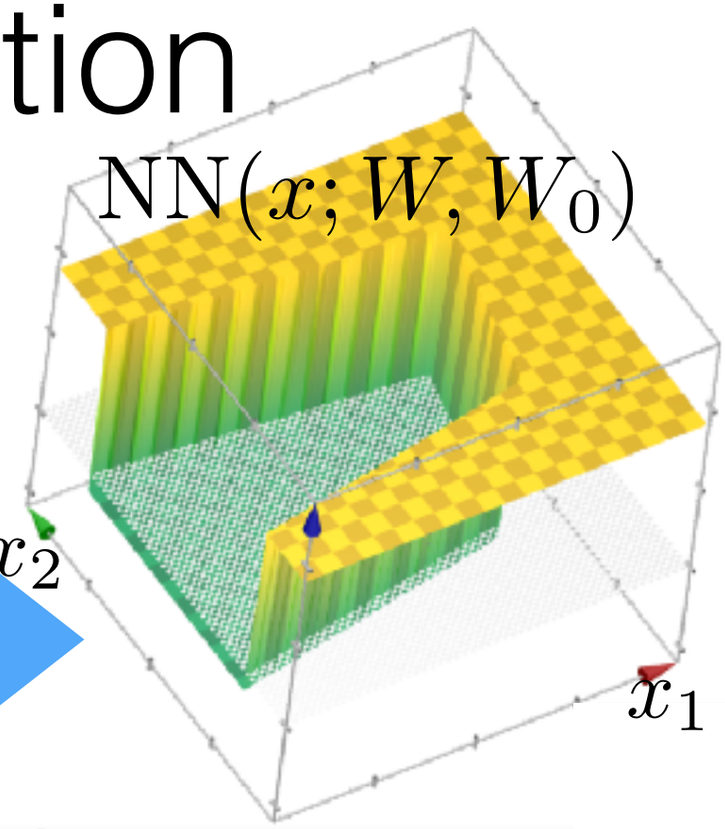

- Circle: function evaluation

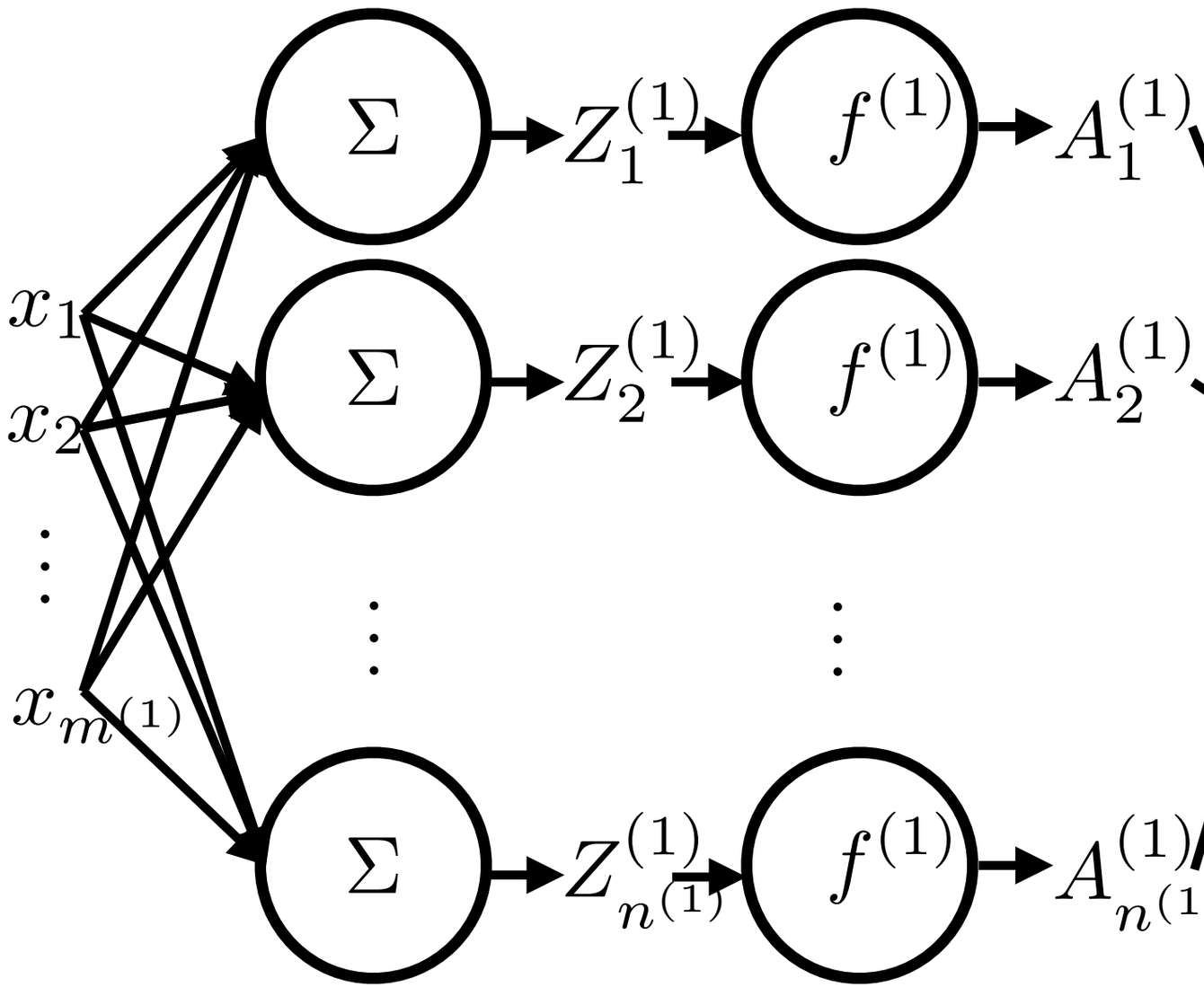



- Forward vs. backward

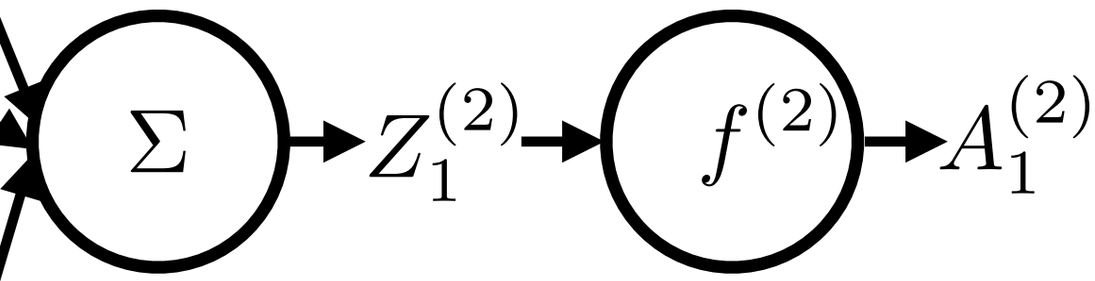
# Function graph representation

- 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: $A^{(2)} = \mathrm{NN}(x; W, W_0)$



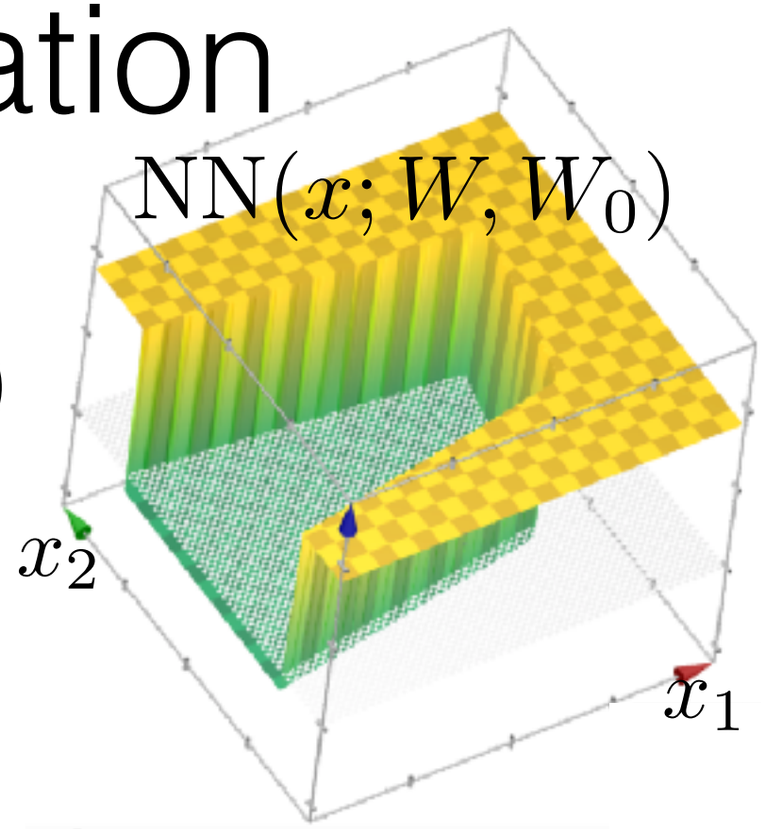

- Circle: function evaluation

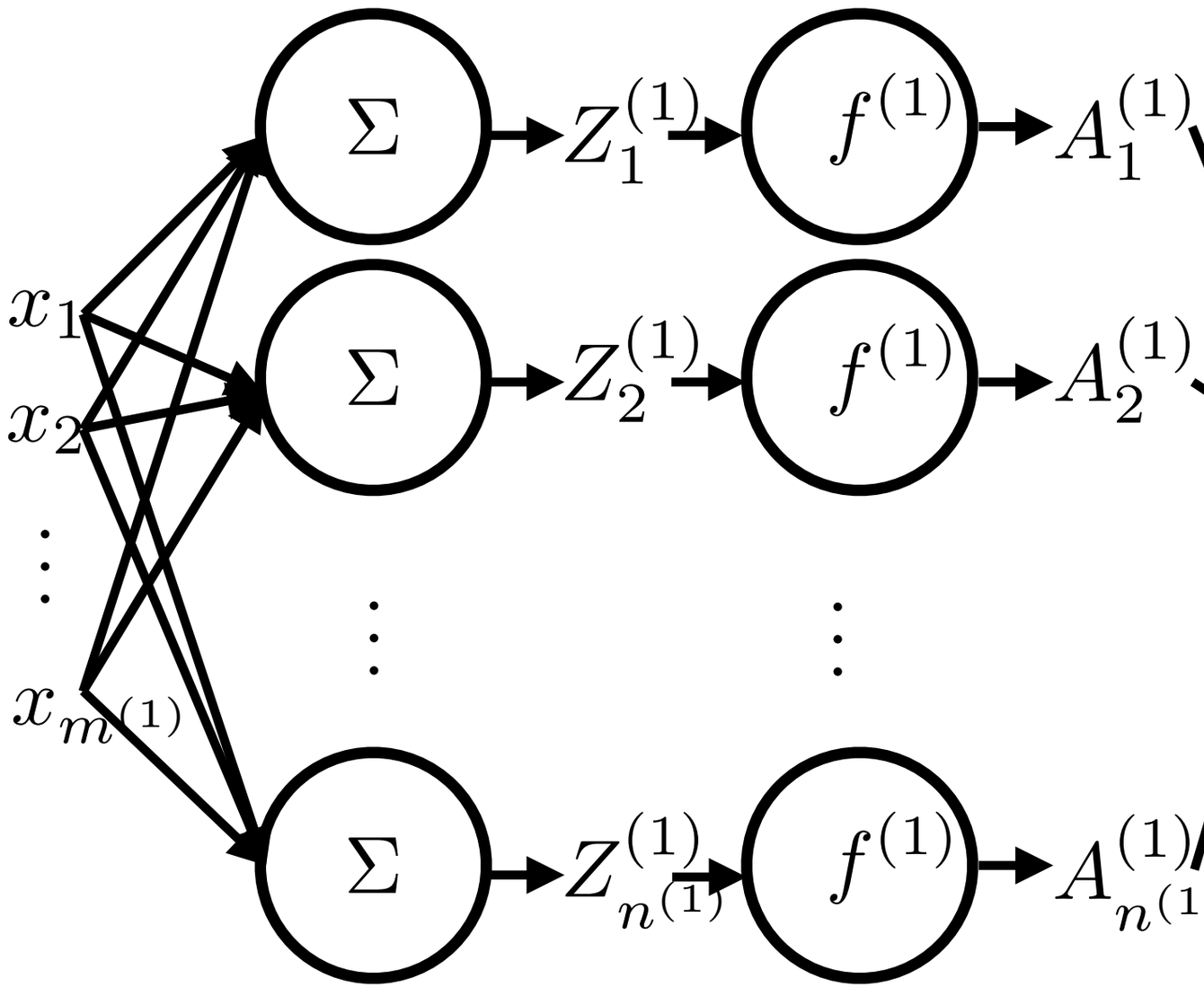



- Forward vs. backward

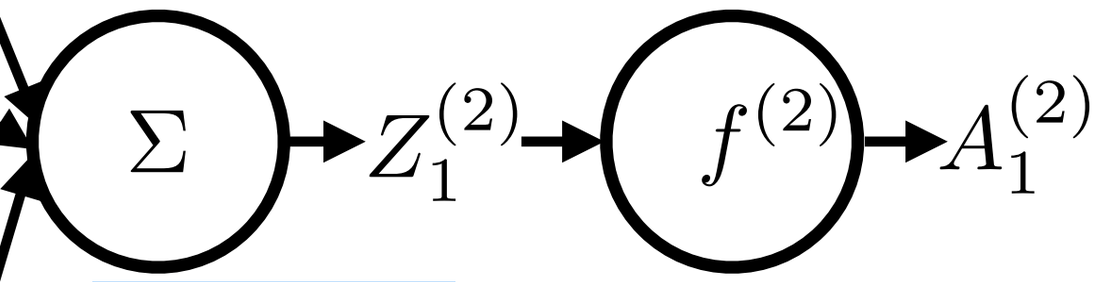
# Function graph representation

- 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: $A^{(2)} = \mathrm{NN}(x; W, W_0)$



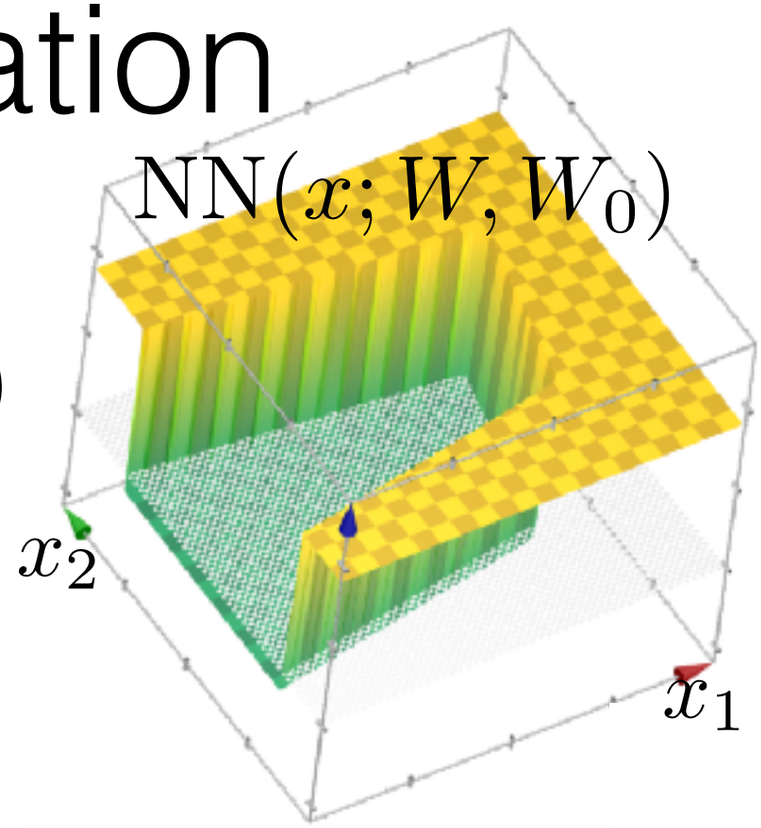

- Circle: function evaluation

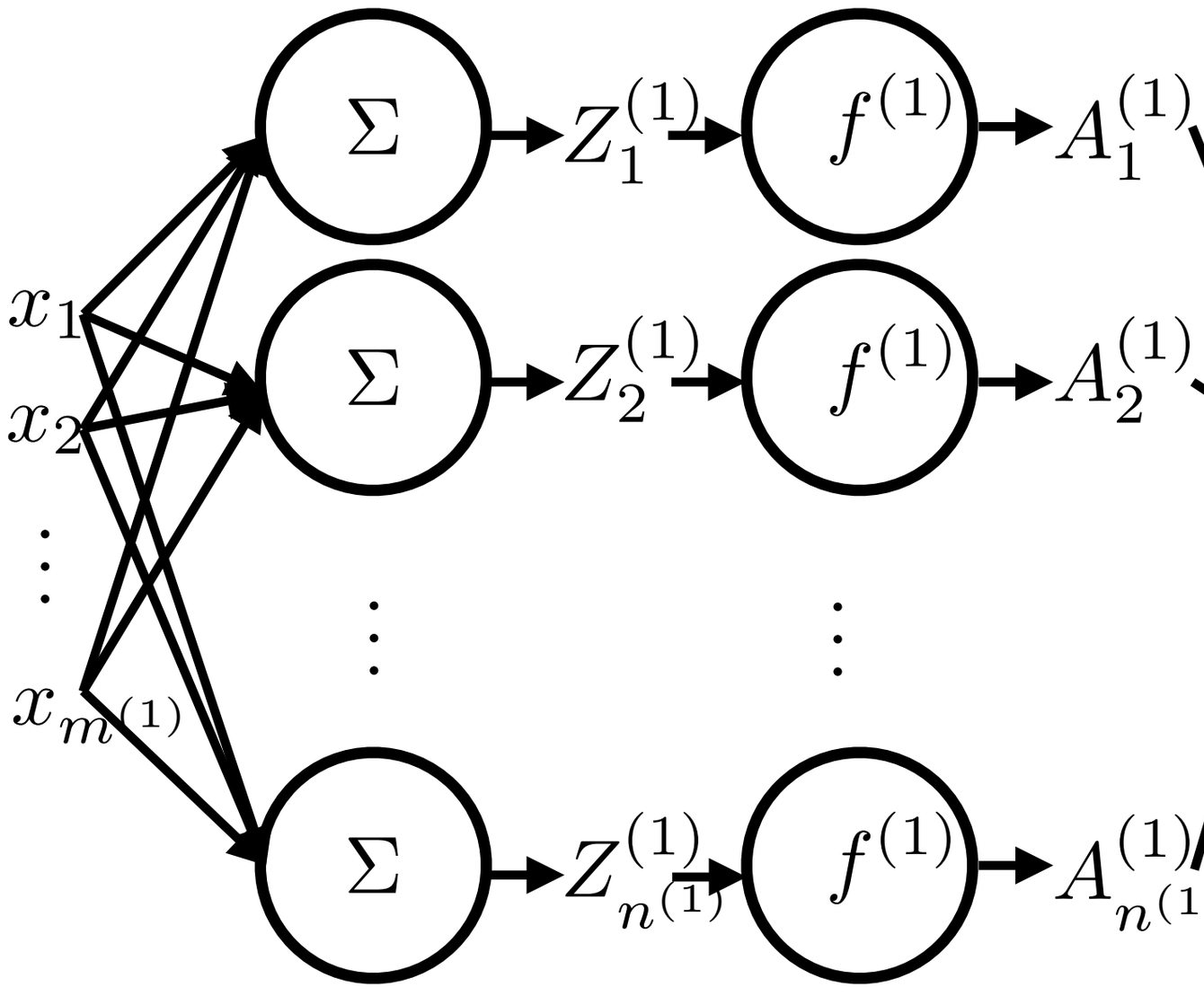



- Forward vs. backward

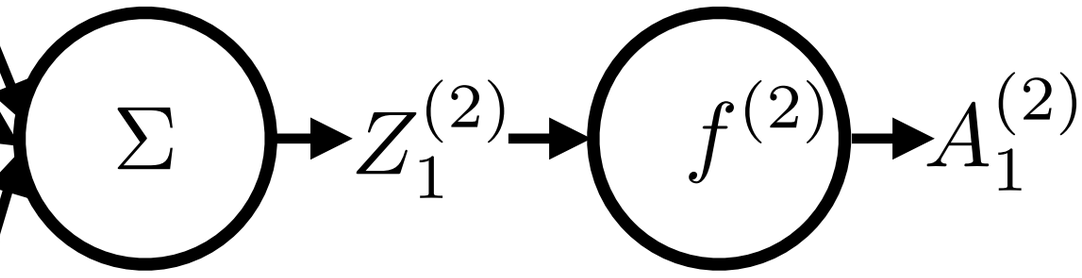
# Function graph representation

- 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: $A^{(2)} = \mathrm{NN}(x; W, W_0)$



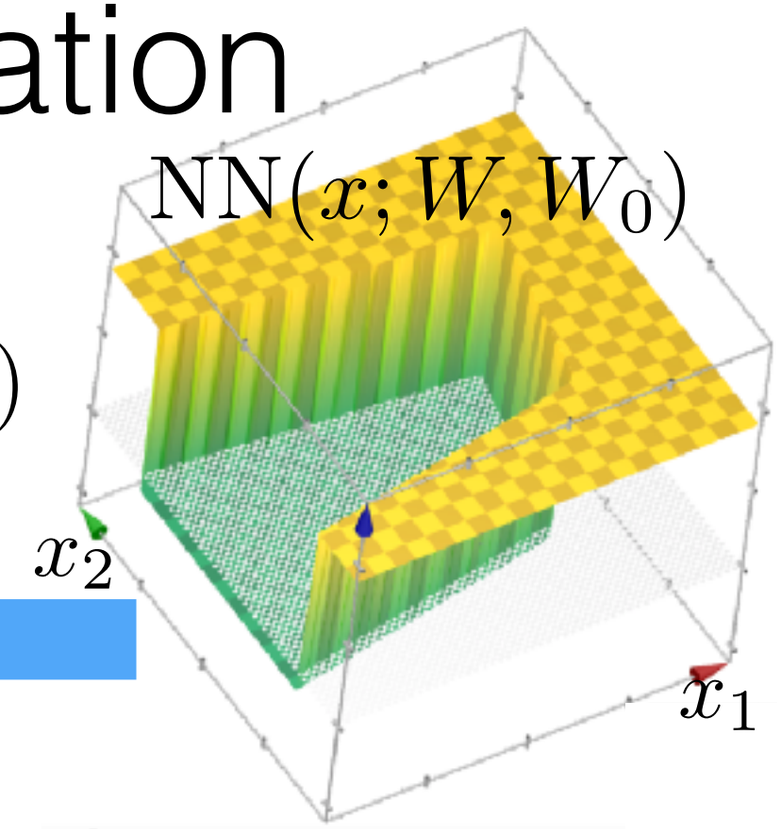

- Circle: function evaluation

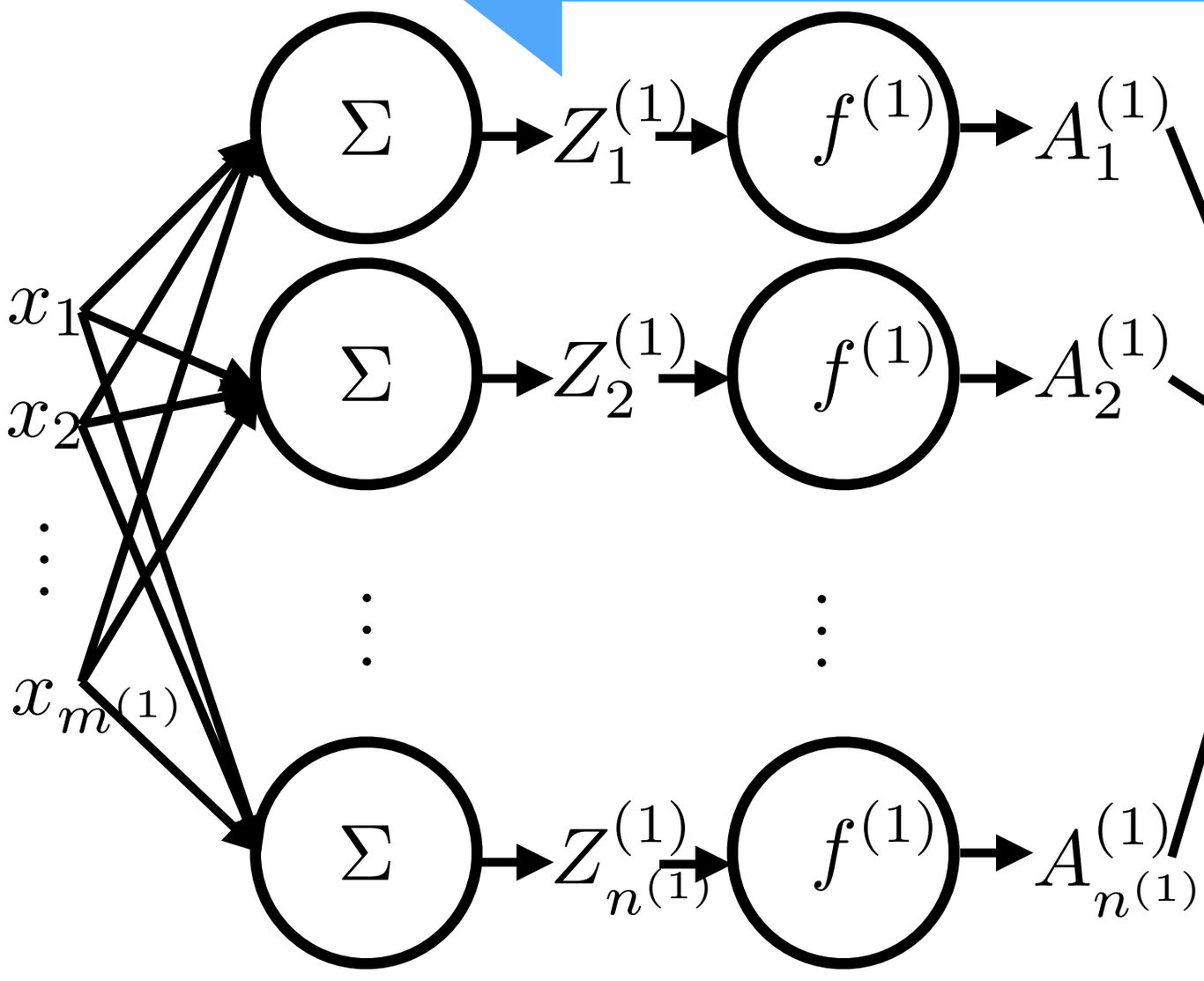



- Forward vs. backward

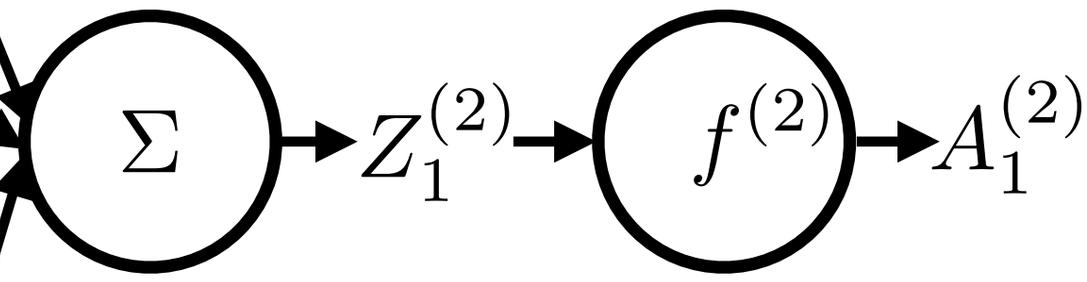
# Function graph representation

- 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: $A^{(2)} = \mathrm{NN}(x; W, W_0)$



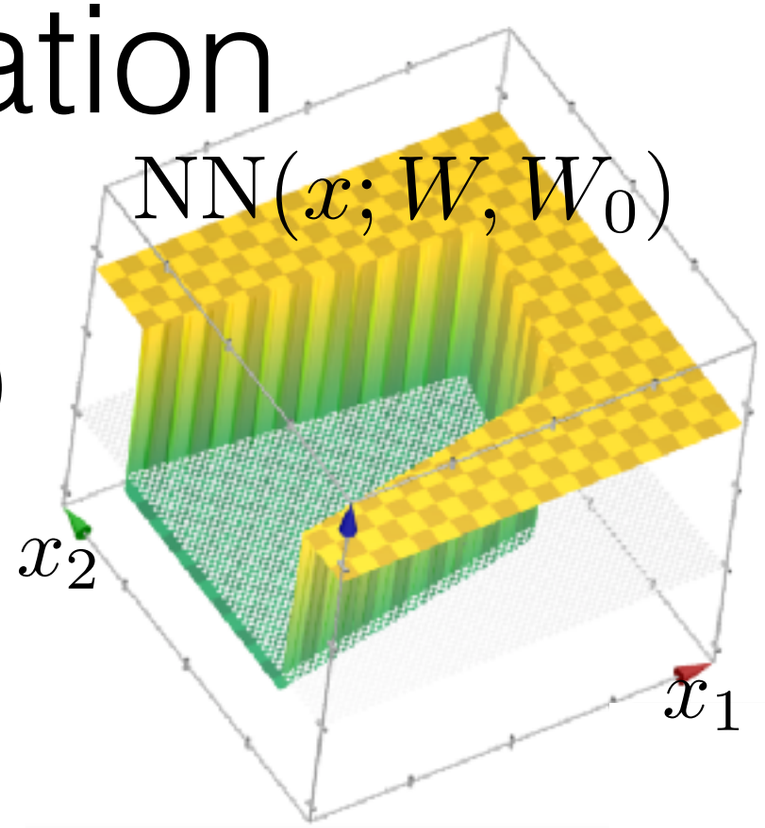

- Circle: function evaluation

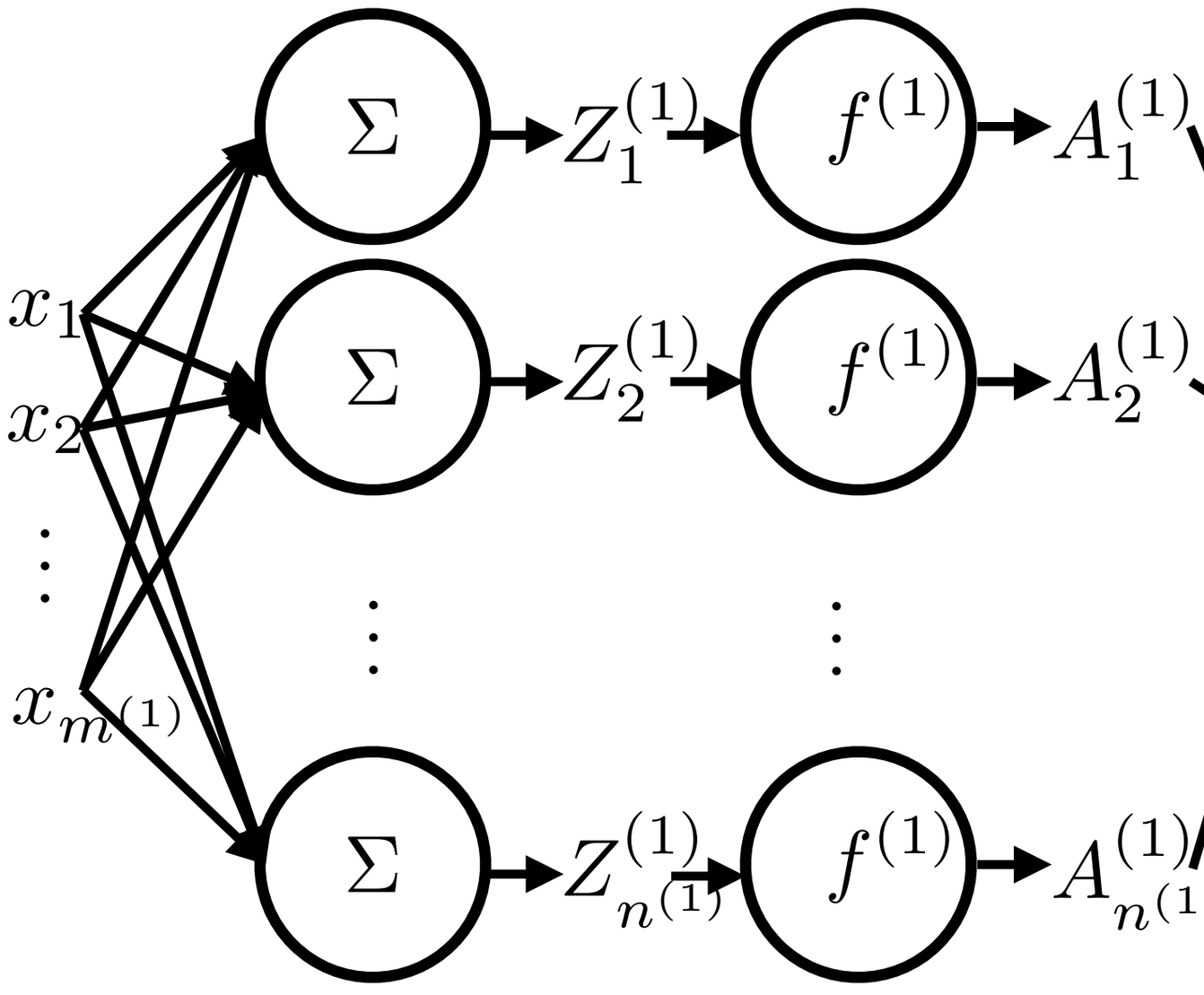



- Forward vs. backward

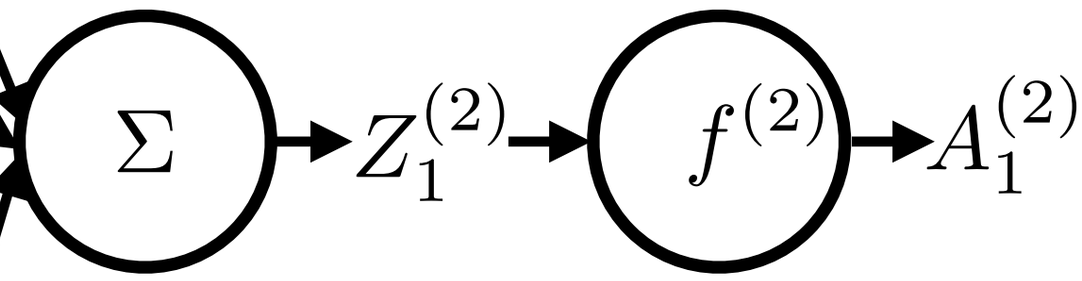
# Function graph representation

- 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: $A^{(2)} = \mathrm{NN}(x; W, W_0)$



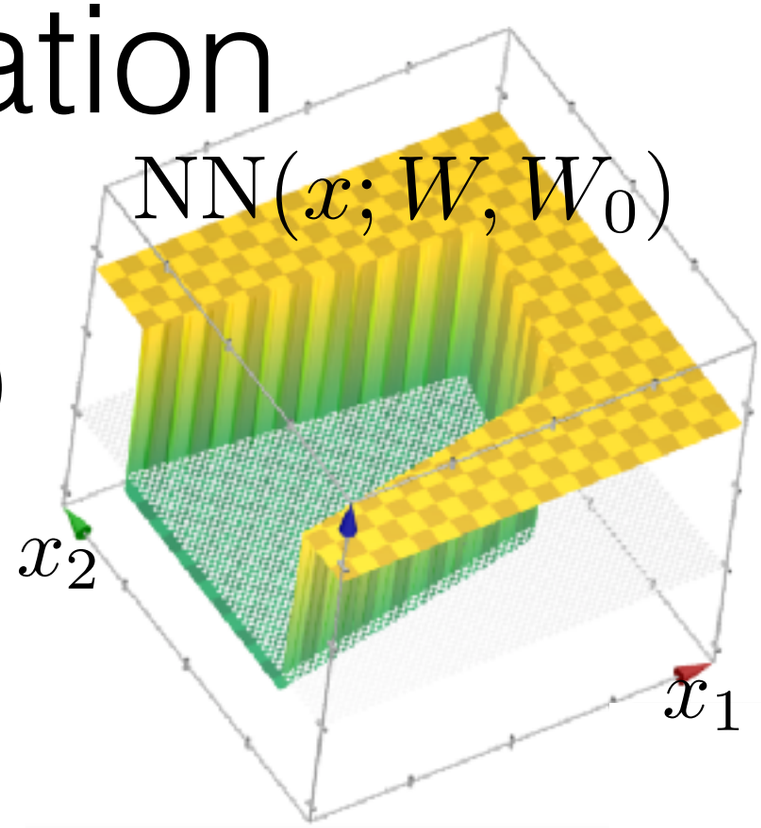

- Circle: function evaluation

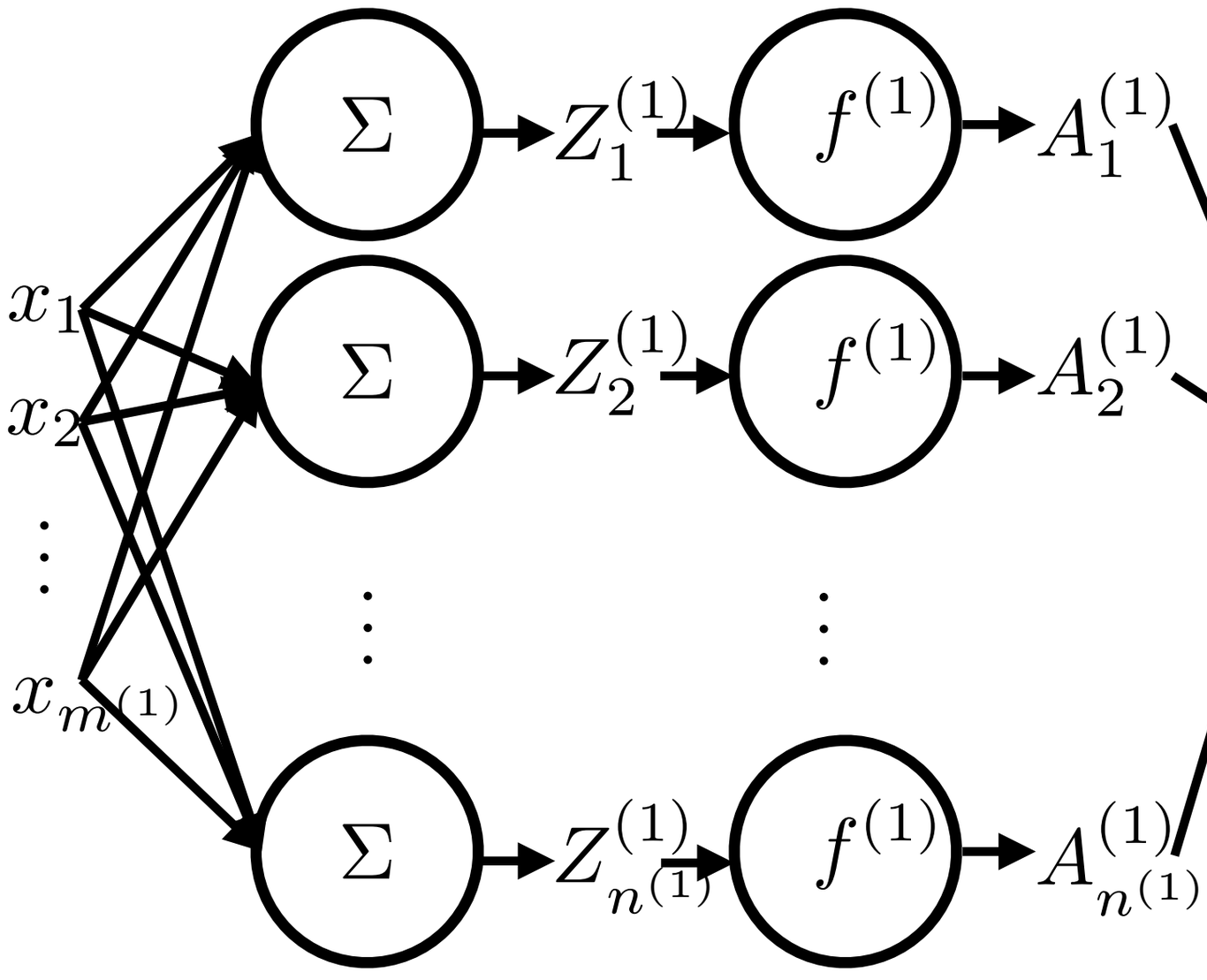



- Forward vs. backward

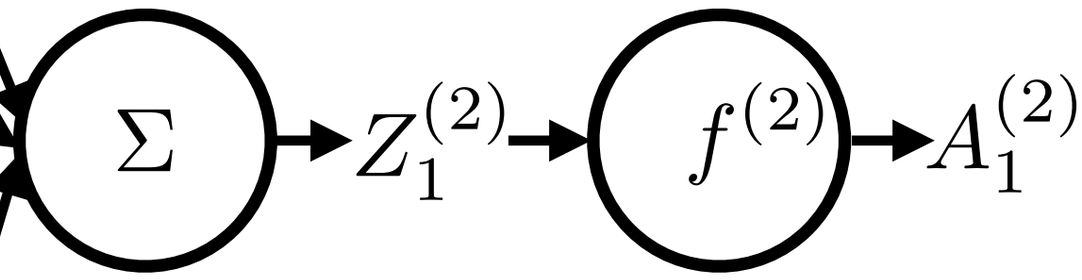
# Function graph representation

- 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: $A^{(2)} = \mathrm{NN}(x; W, W_0)$



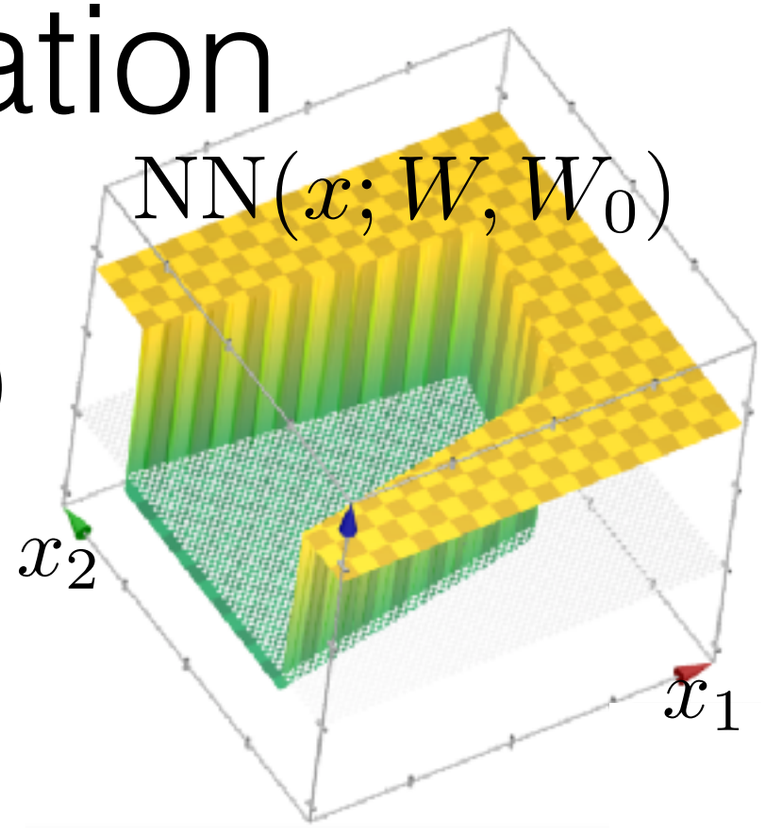

- Circle: function evaluation

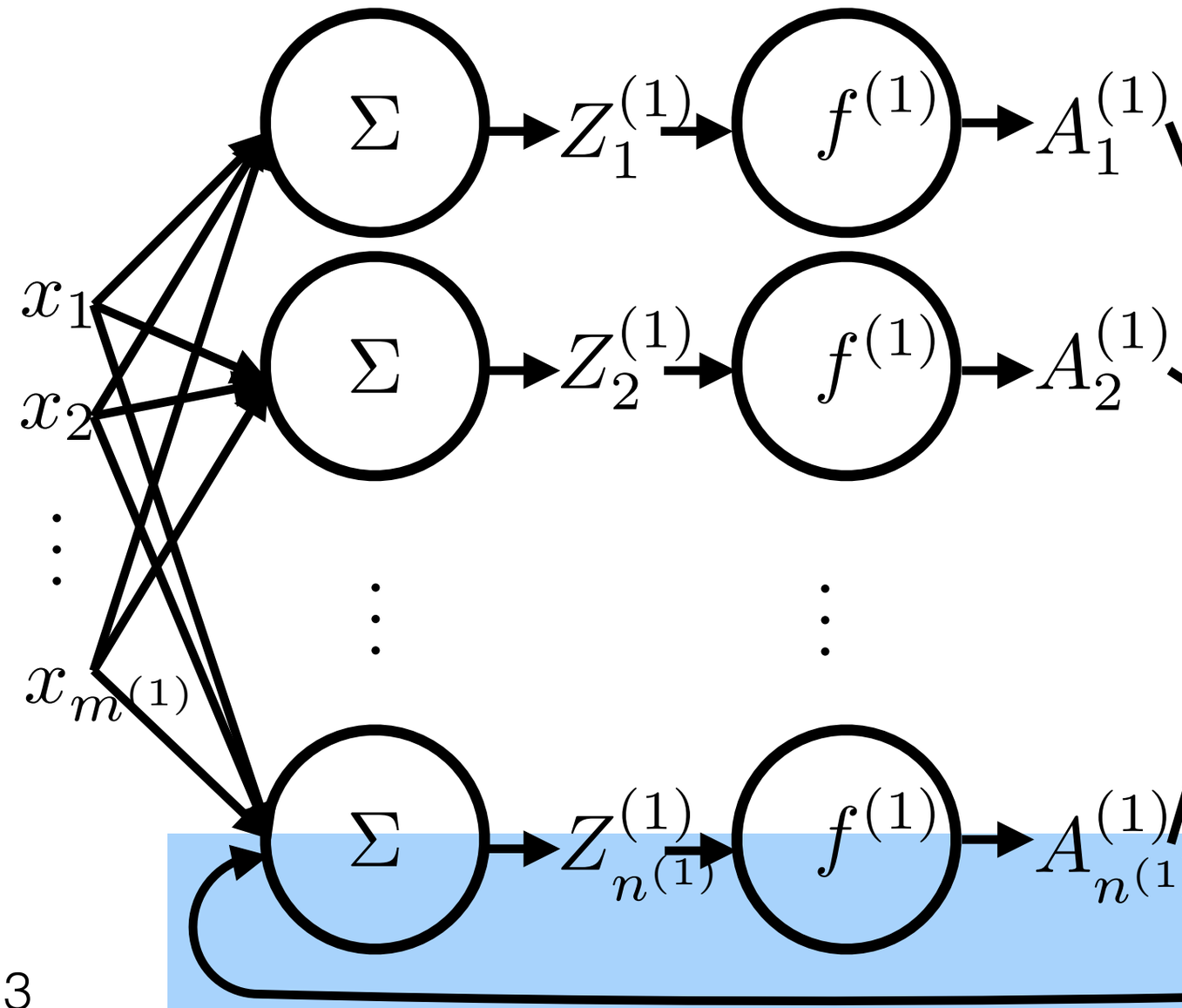



- Forward vs. backward
- A feed-forward neural network

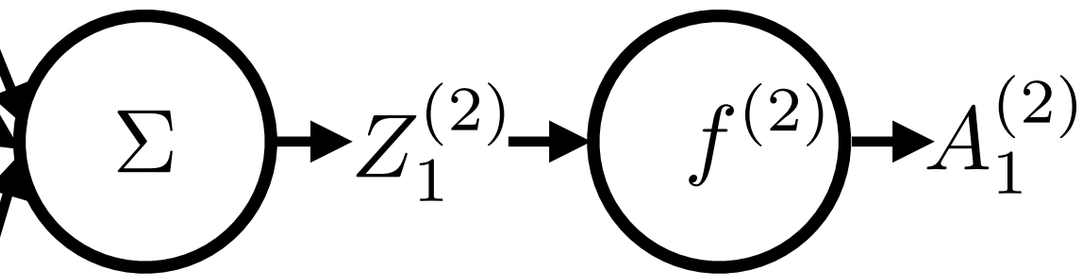
# Function graph representation

- 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

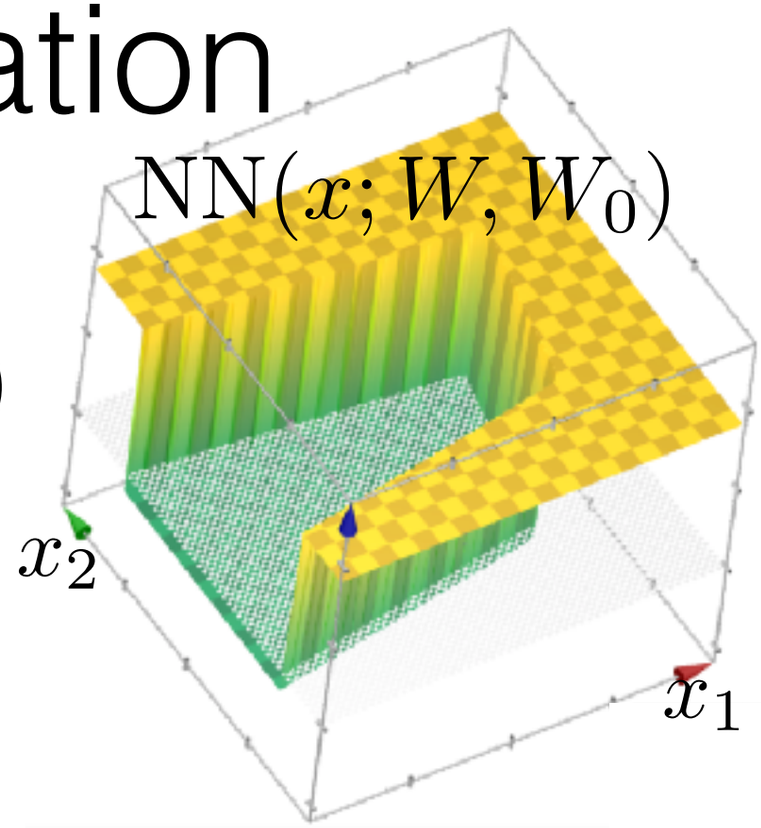


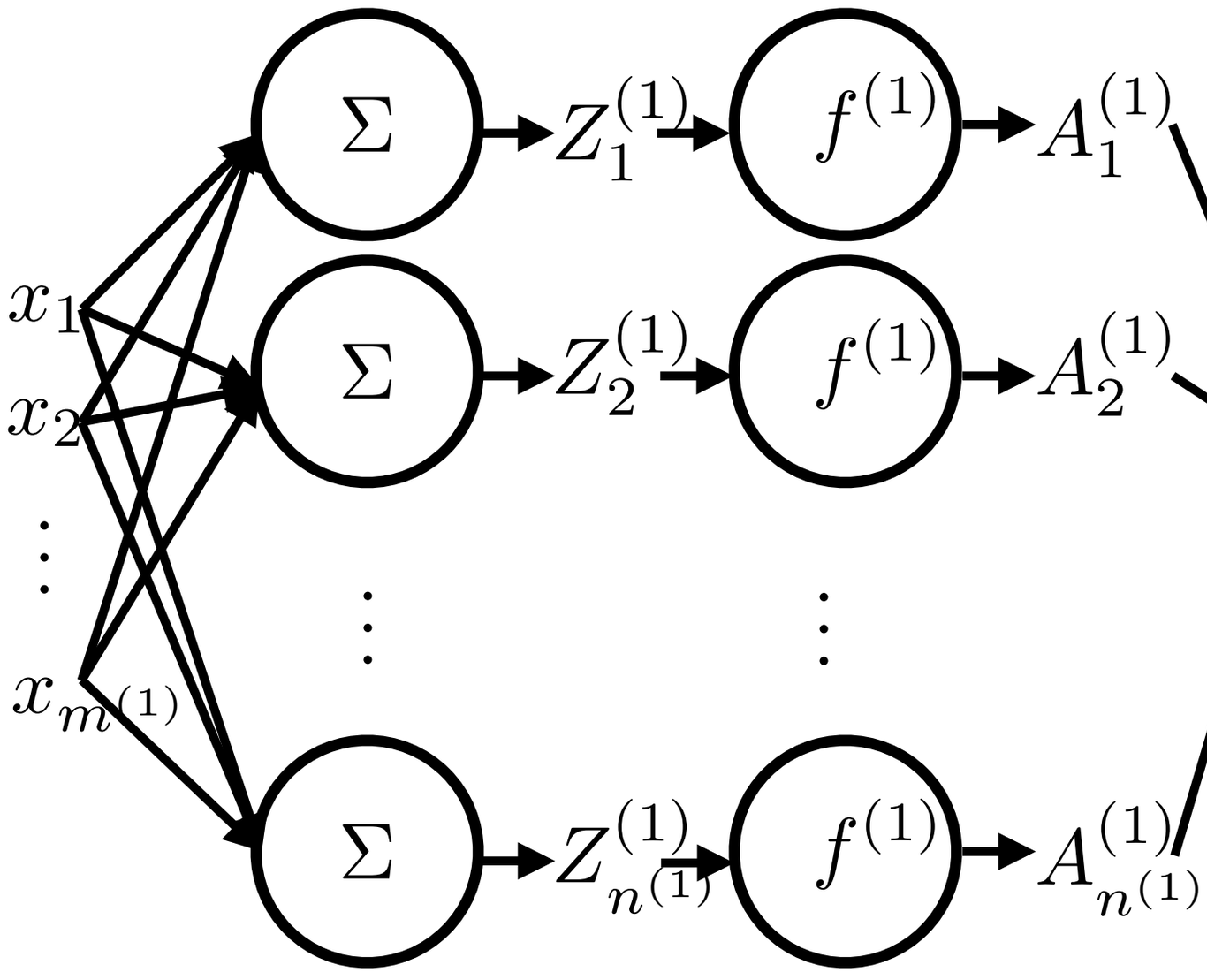

- Forward vs. backward
- A feed-forward neural network

not feed-forward

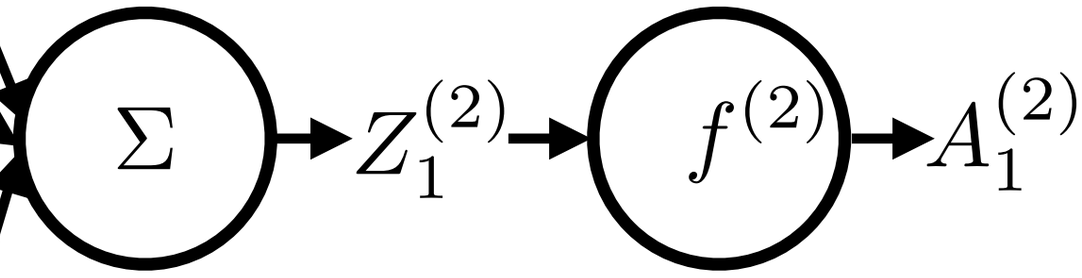
# Function graph representation

- 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: $A^{(2)} = \mathrm{NN}(x; W, W_0)$



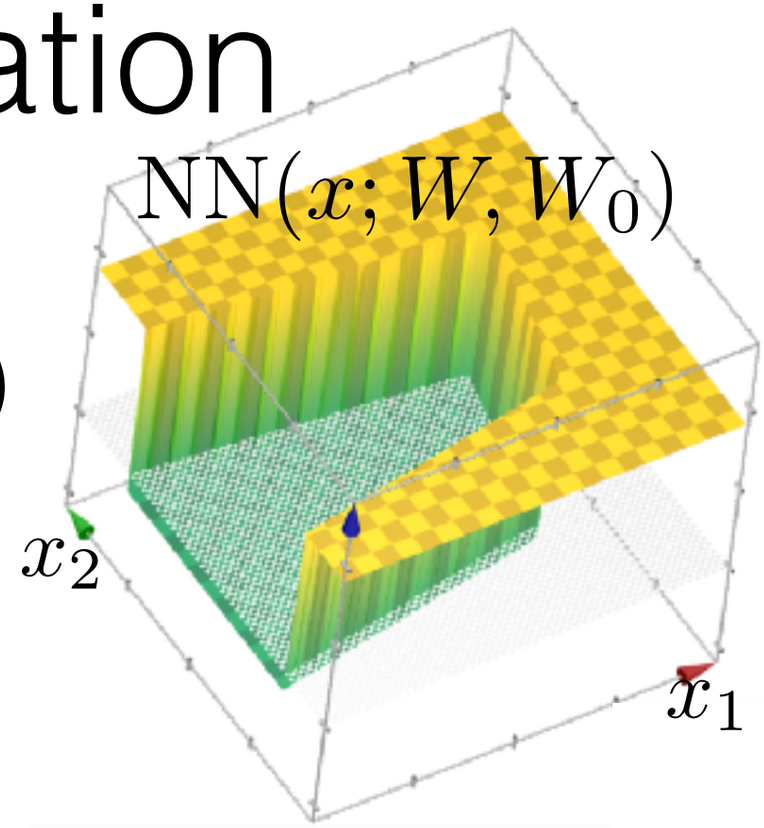

- Circle: function evaluation

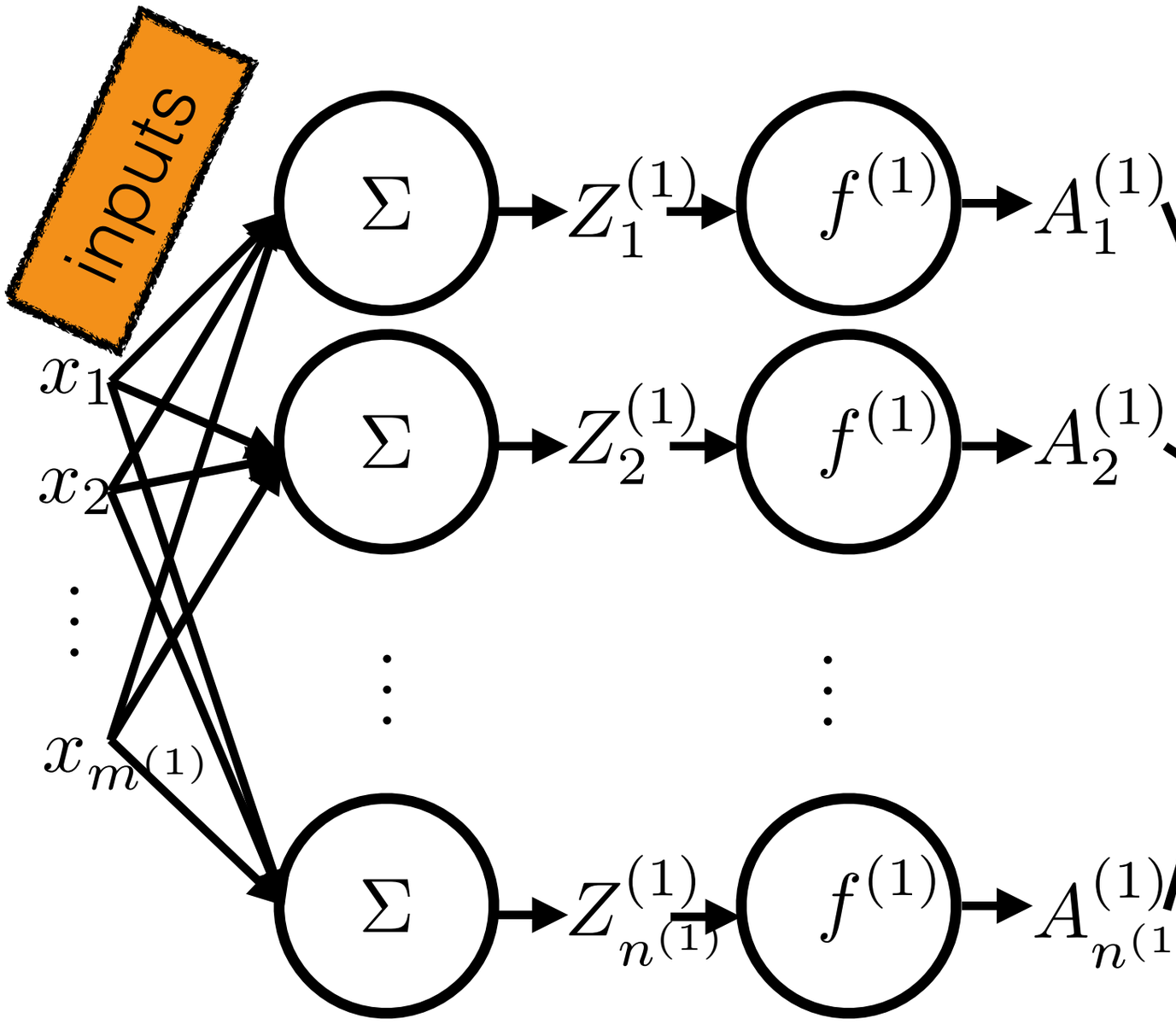



- Forward vs. backward
- A feed-forward neural network

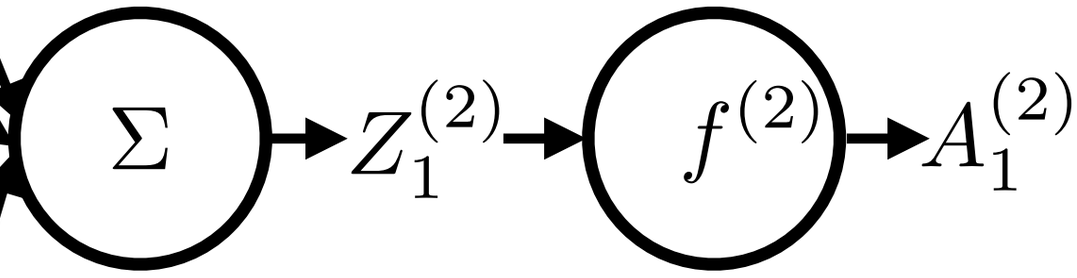
# Function graph representation

- 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: $A^{(2)} = \mathrm{NN}(x; W, W_0)$



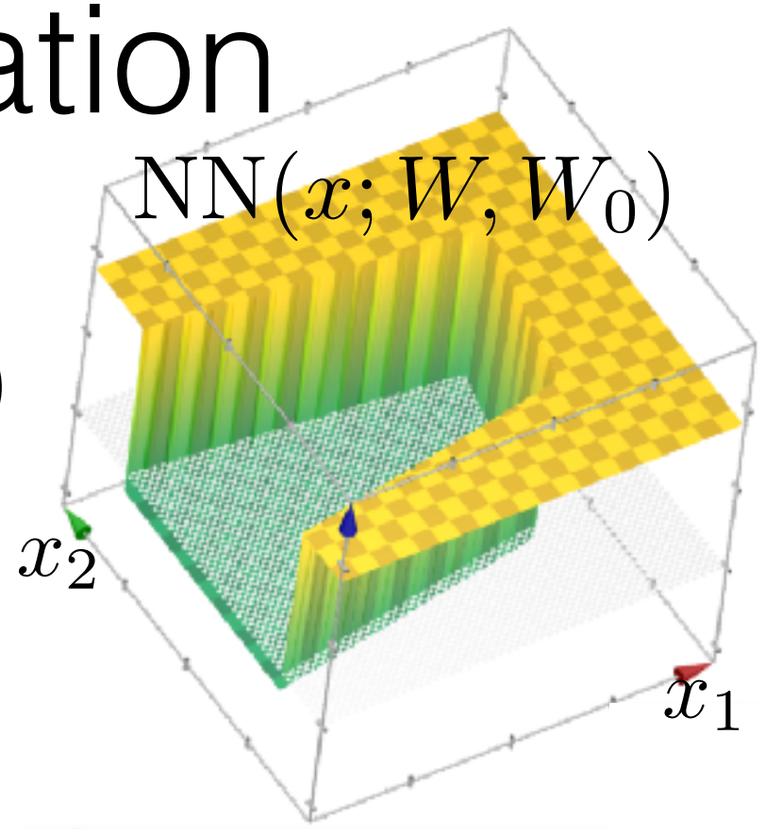

- Circle: function evaluation




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

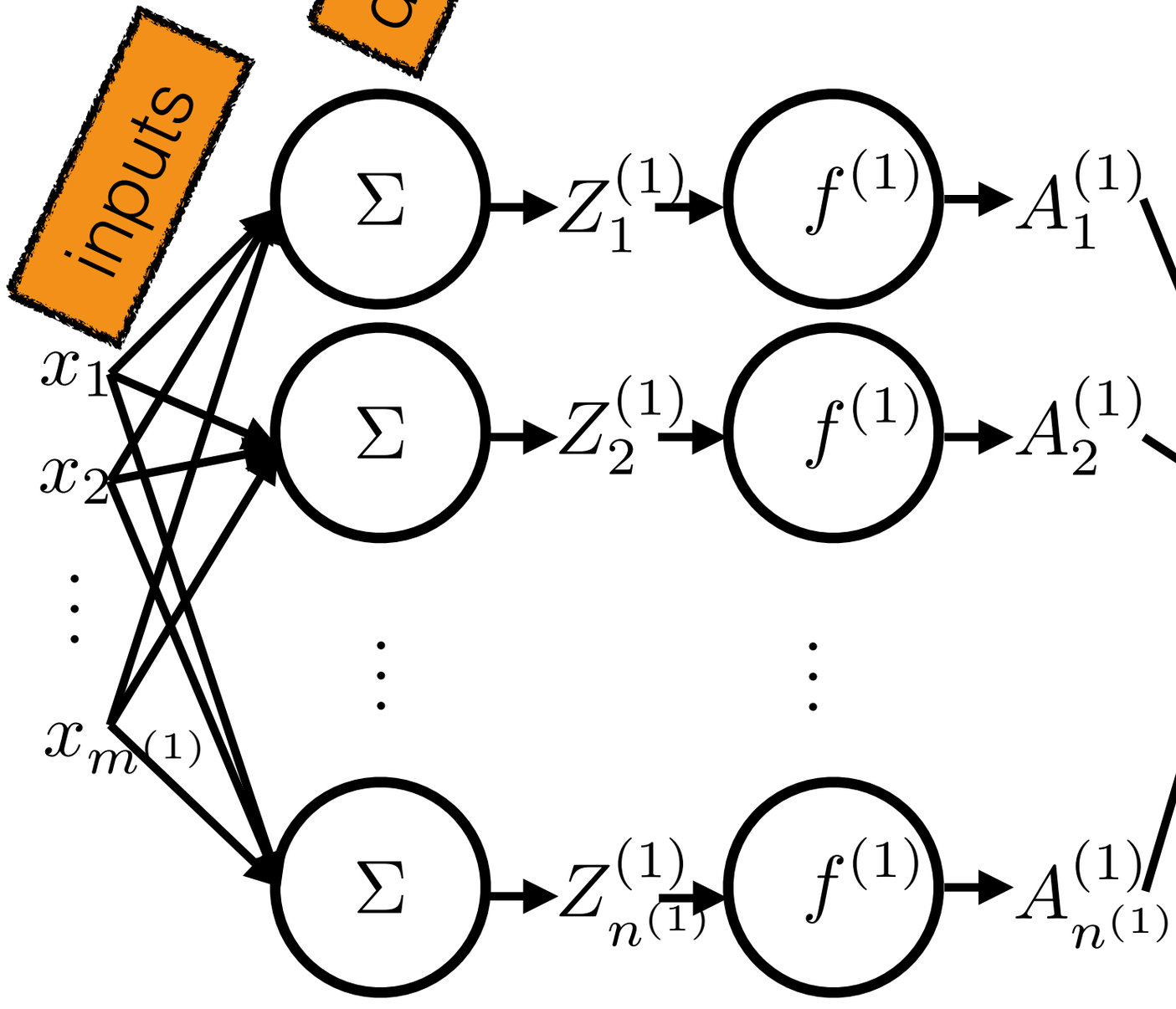
# Function graph representation

- 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: $A^{(2)} = \mathrm{NN}(x; W, W_0)$

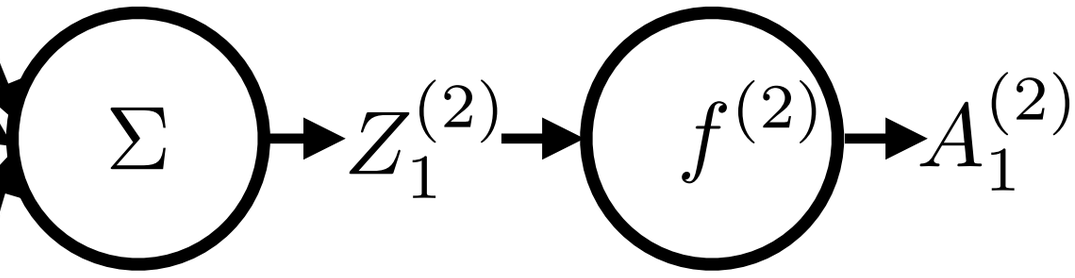



dot product

inputs



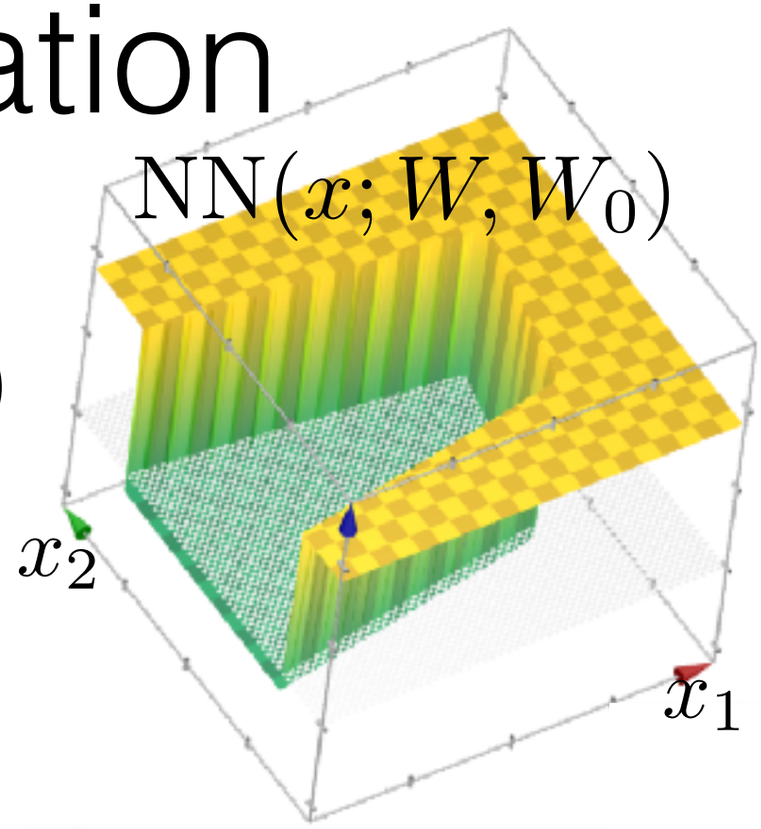

- Circle: function evaluation




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

# Function graph representation

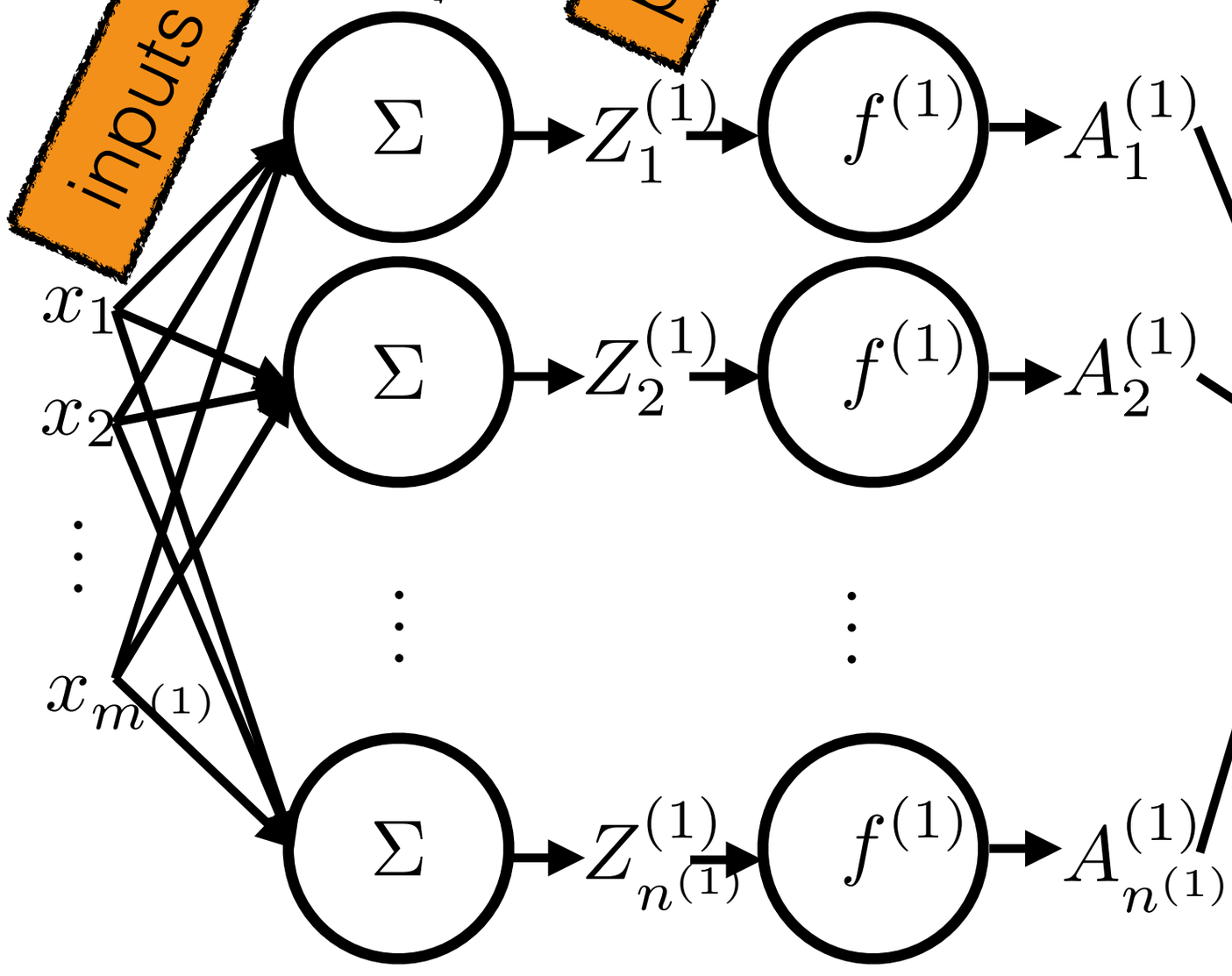
- 1st layer: $A^{(1)} = f^{(1)}(W^{(1)\top} x + W_0^{(1)})$
- 2nd layer: $A^{(2)} = f^{(2)}(W^{(2)\top} A^{(1)} + W_0^{(2)})$
- Whole thing: $A^{(2)} = \mathrm{NN}(x; W, W_0)$



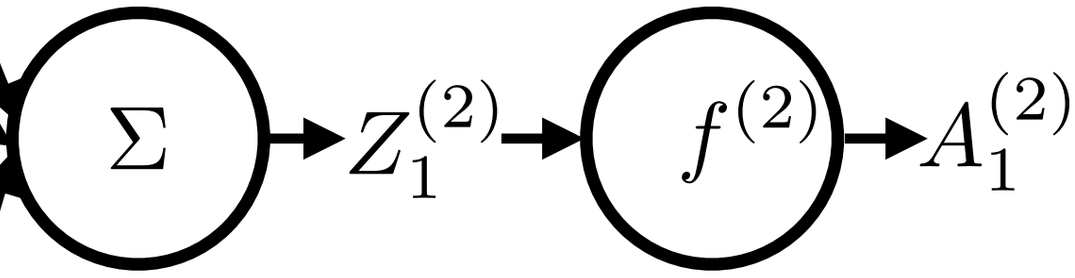

dot product

pre-activation

inputs



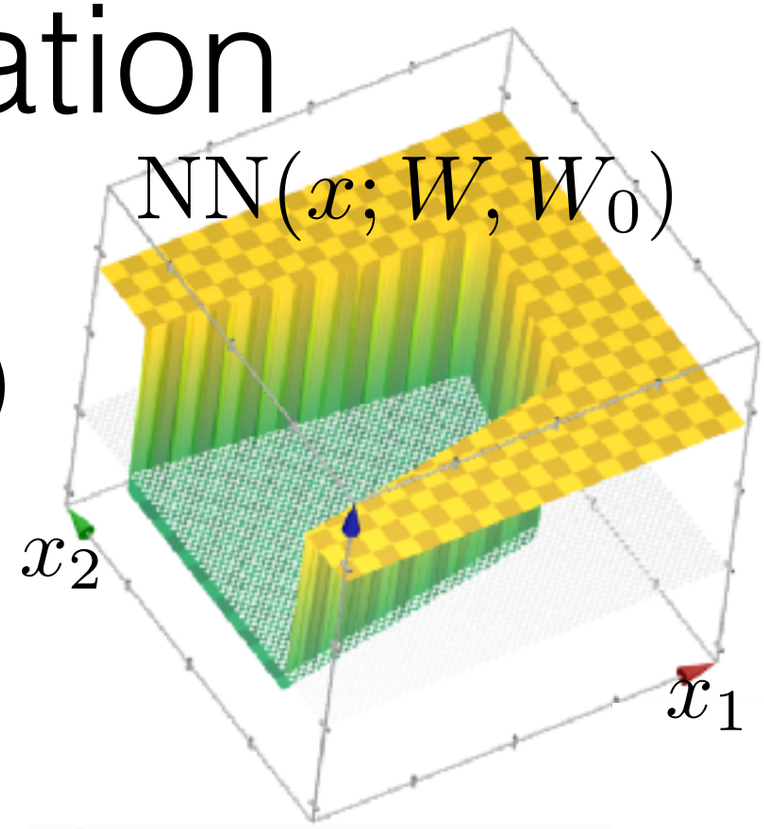

- Circle: function evaluation

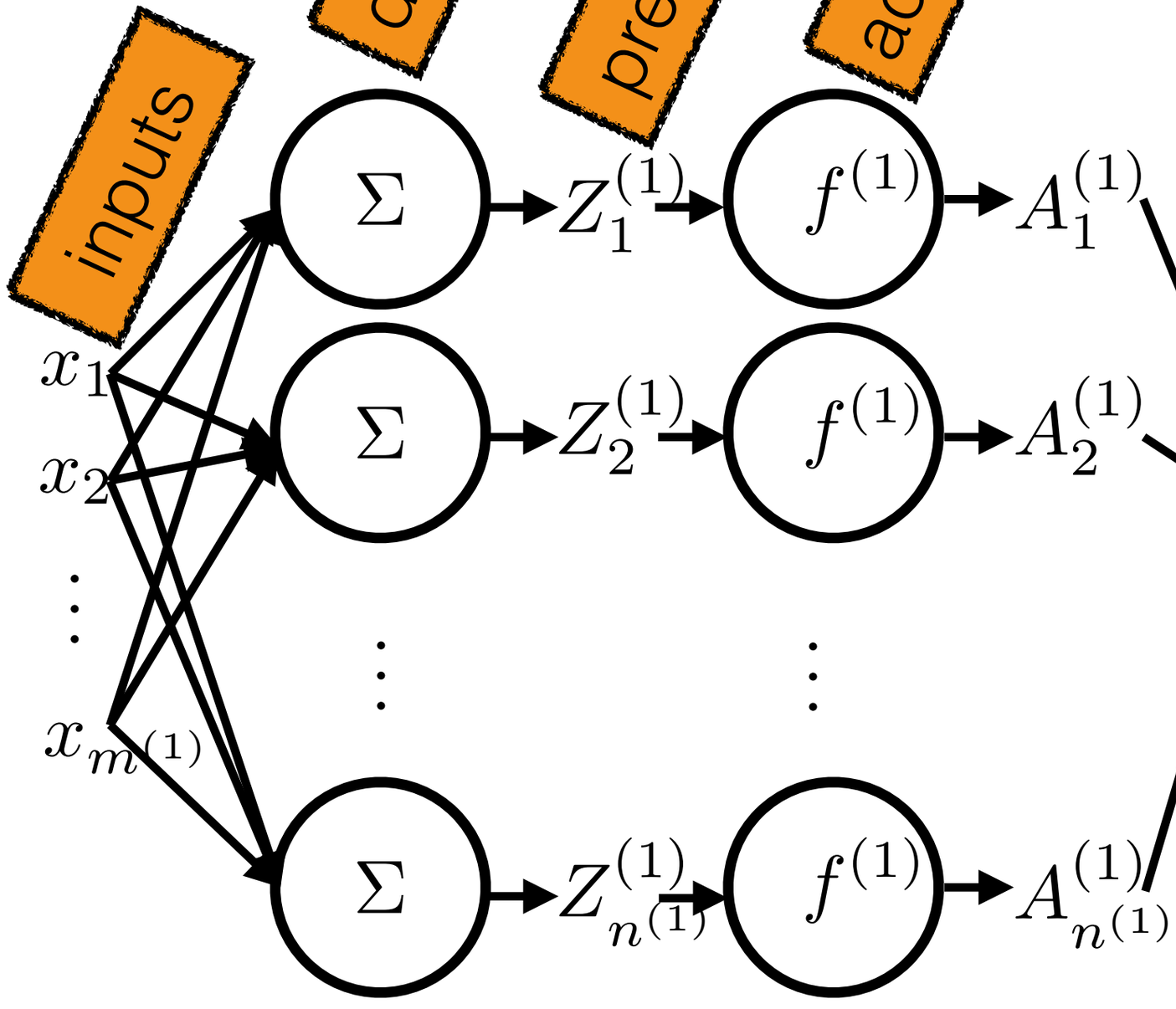



- Forward vs. backward
- A feed-forward neural network

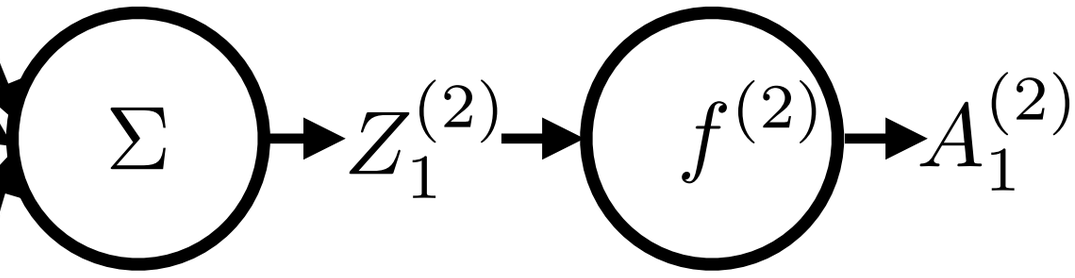
# Function graph representation

- 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: $A^{(2)} = \mathrm{NN}(x; W, W_0)$



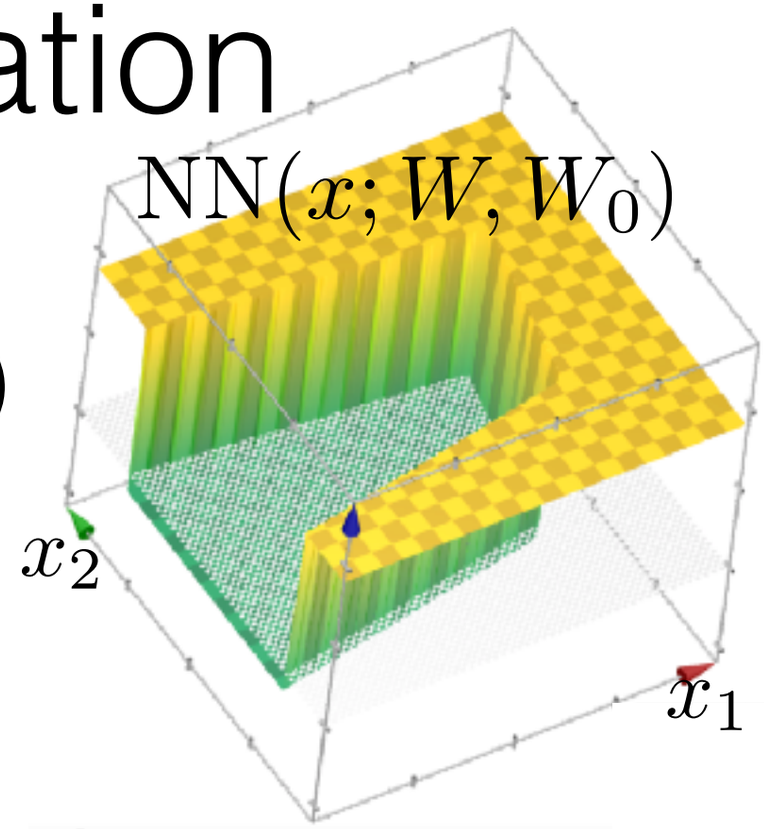

- Circle: function evaluation

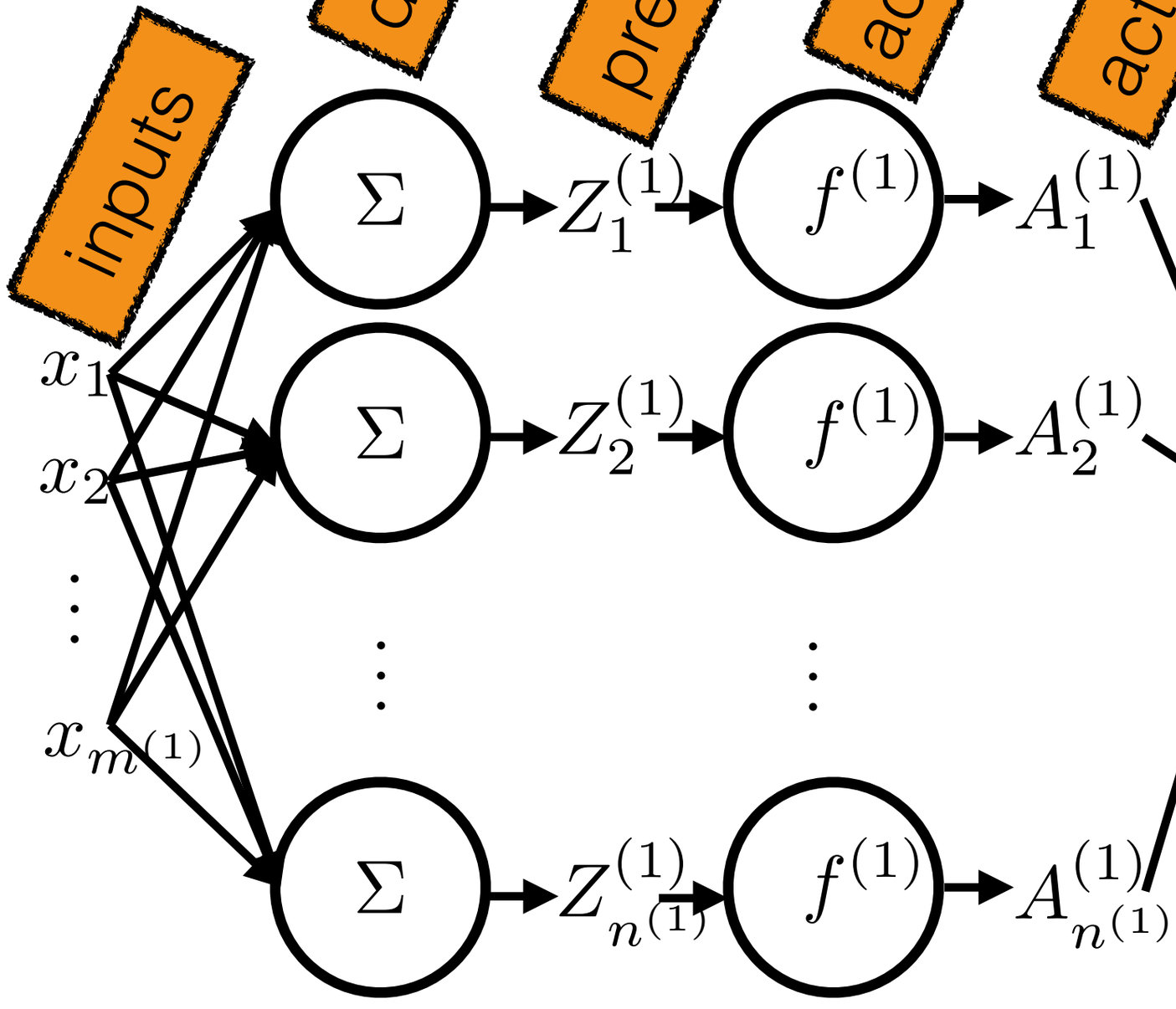



- Forward vs. backward
- A feed-forward neural network

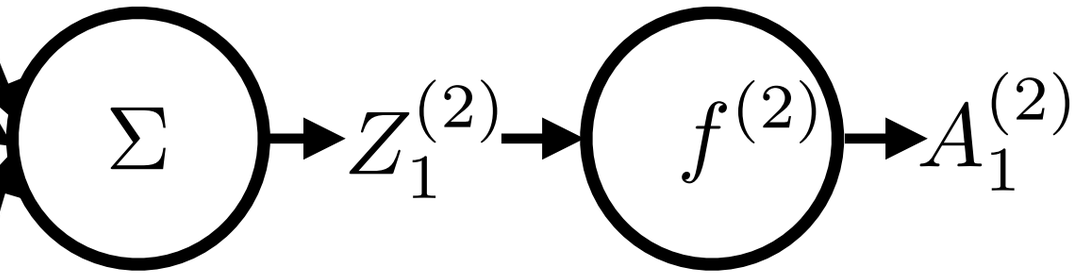
# Function graph representation

- 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: $A^{(2)} = \mathrm{NN}(x; W, W_0)$



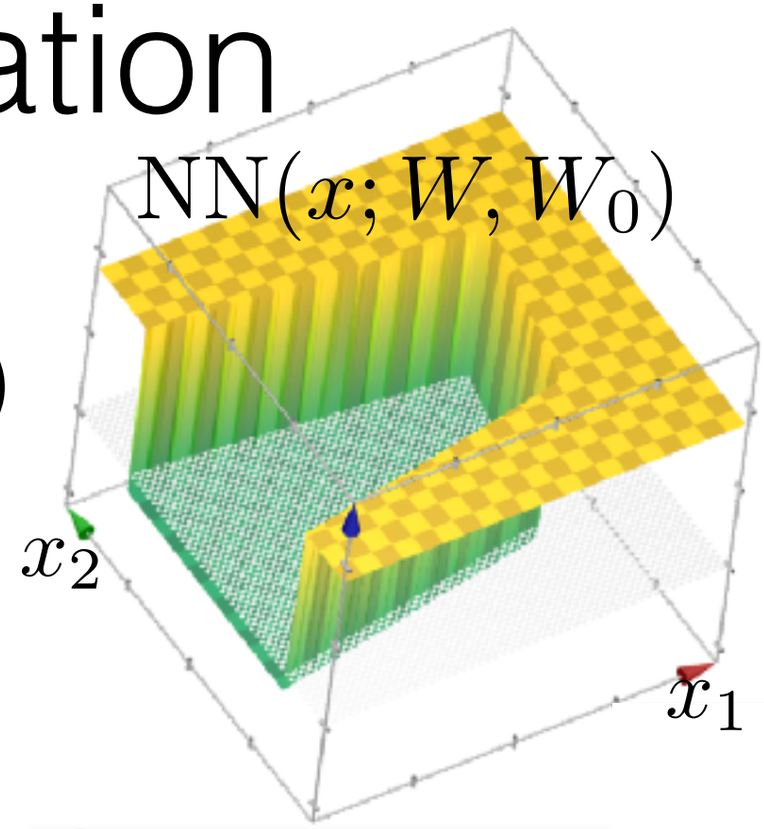

- Circle: function evaluation

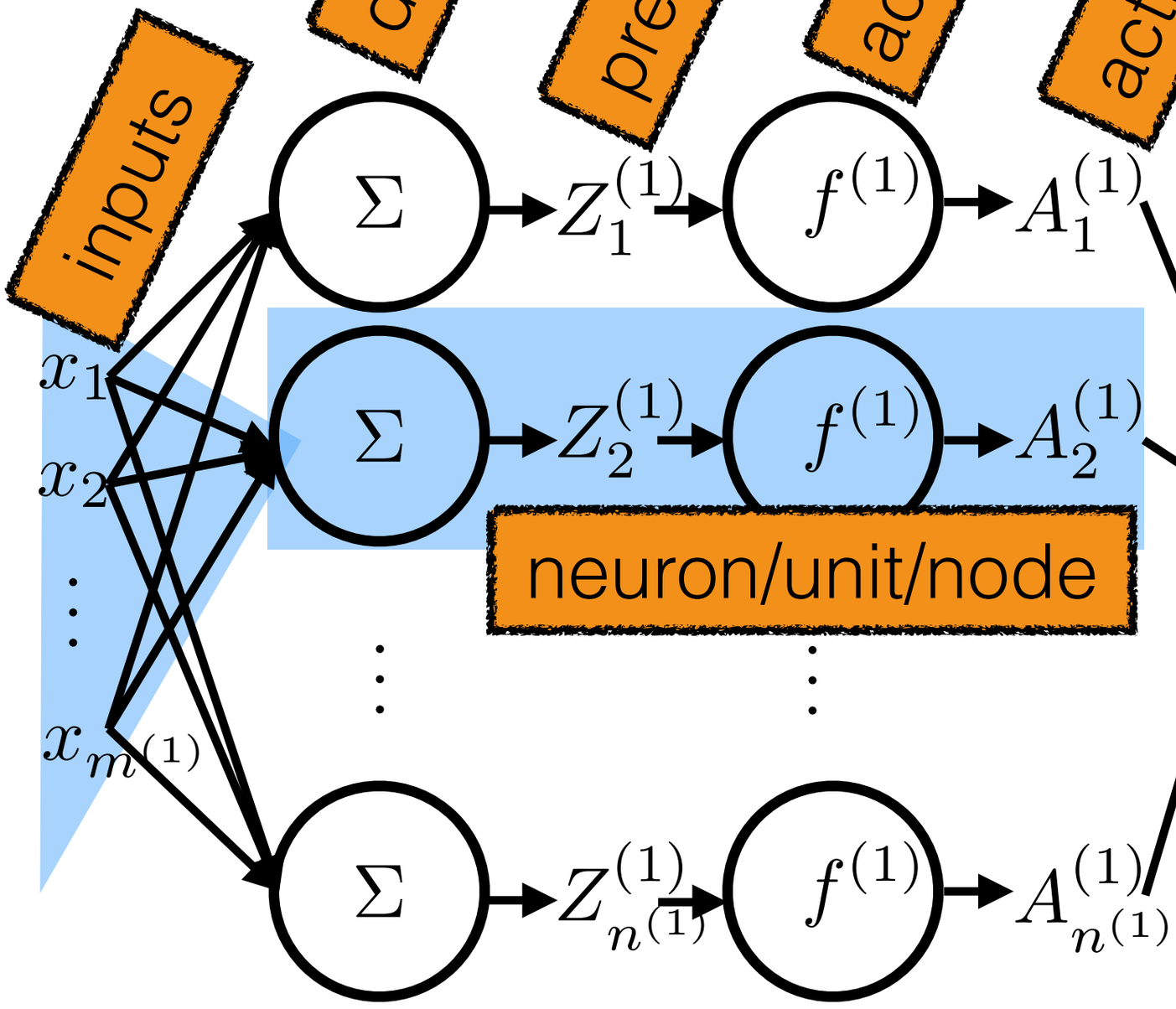



- Forward vs. backward
- A feed-forward neural network

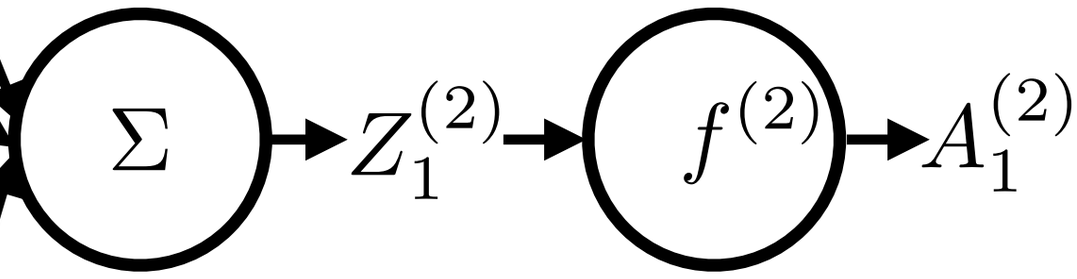
# Function graph representation

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



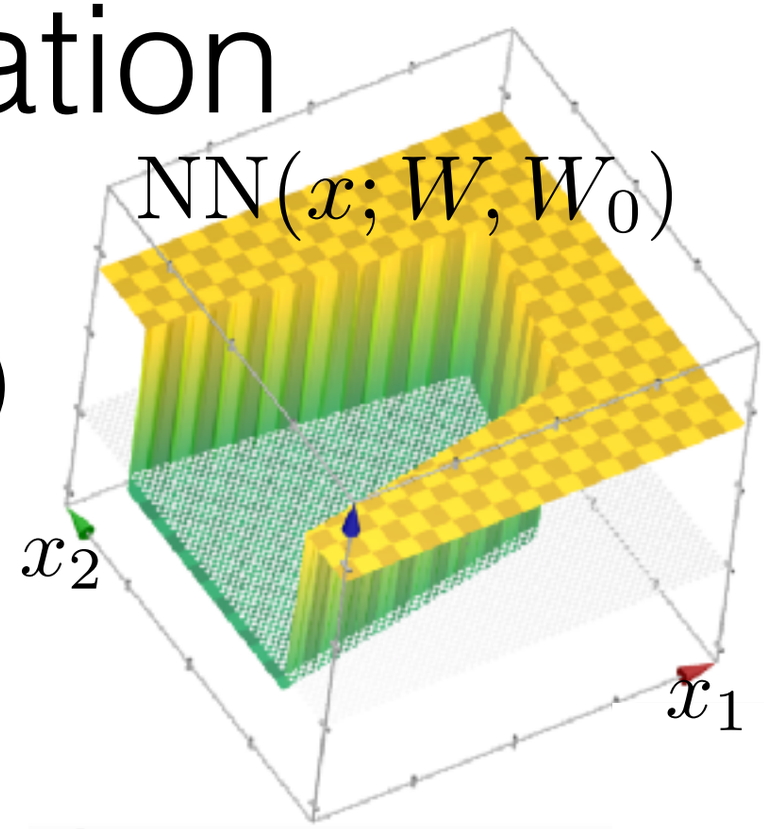

- Circle: function evaluation

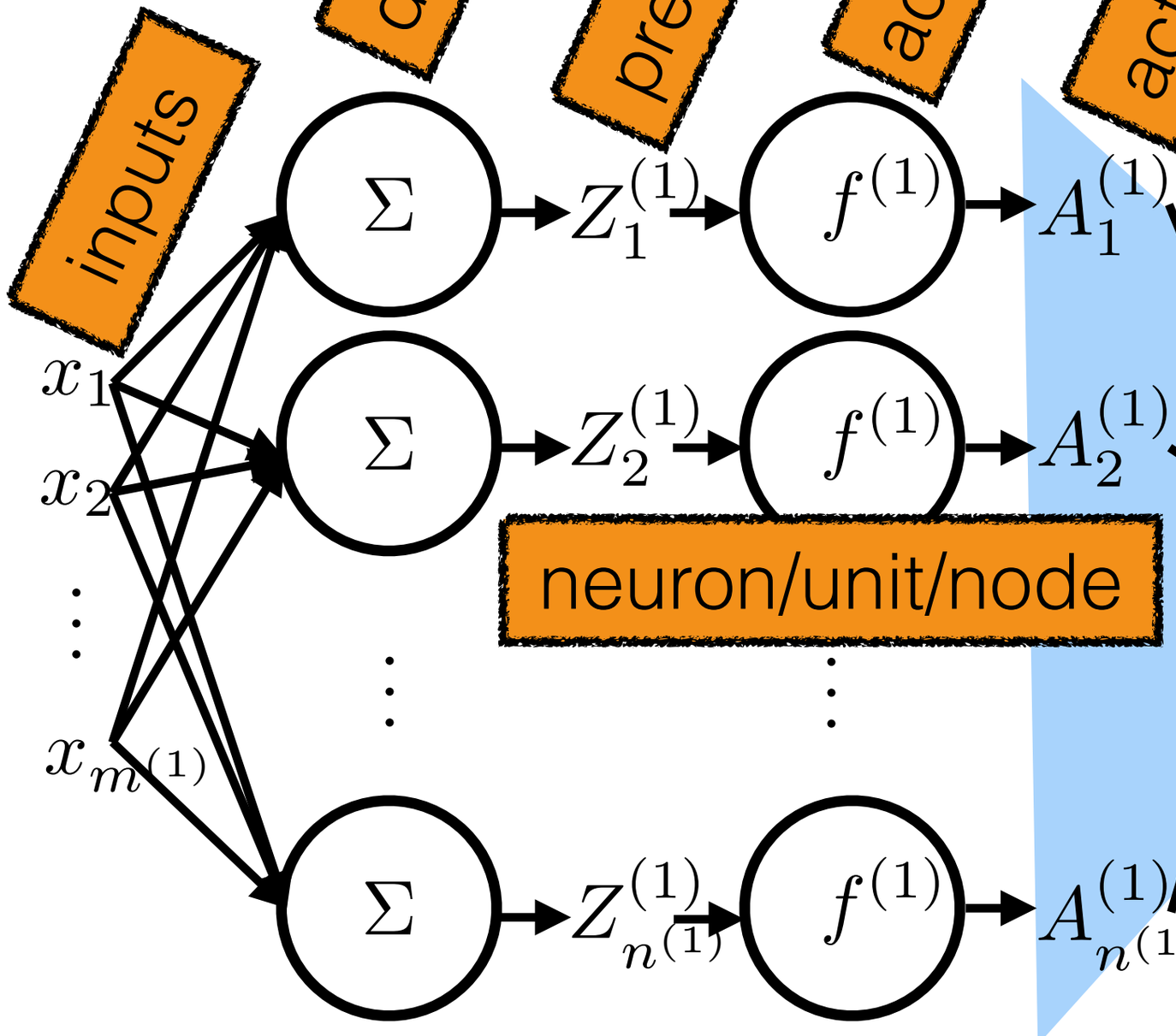



- Forward vs. backward
- A feed-forward neural network

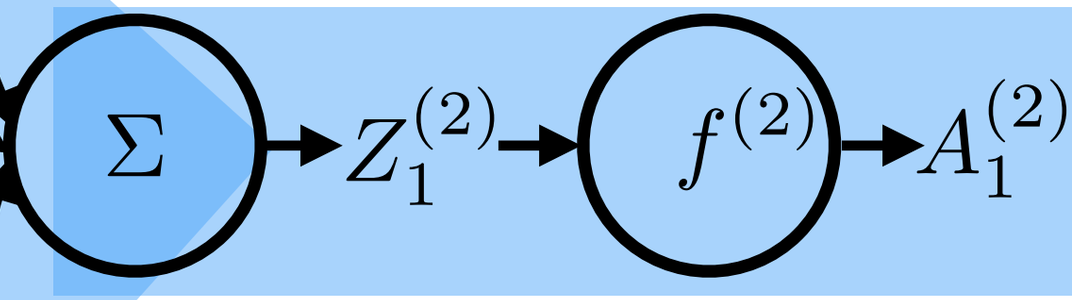
# Function graph representation

- 1st layer: $A^{(1)} = f^{(1)}(\underbrace{W^{(1)} x}_{\text{dot product}} + \underbrace{W_0^{(1)}}_{\text{pre-activation}})$
- 2nd layer: $A^{(2)} = f^{(2)}(\underbrace{W^{(2)} A^{(1)}}_{\text{activation function}} + \underbrace{W_0^{(2)}}_{\text{activation}})$
- Whole thing: $A^{(2)} = \mathrm{NN}(x; W, W_0)$



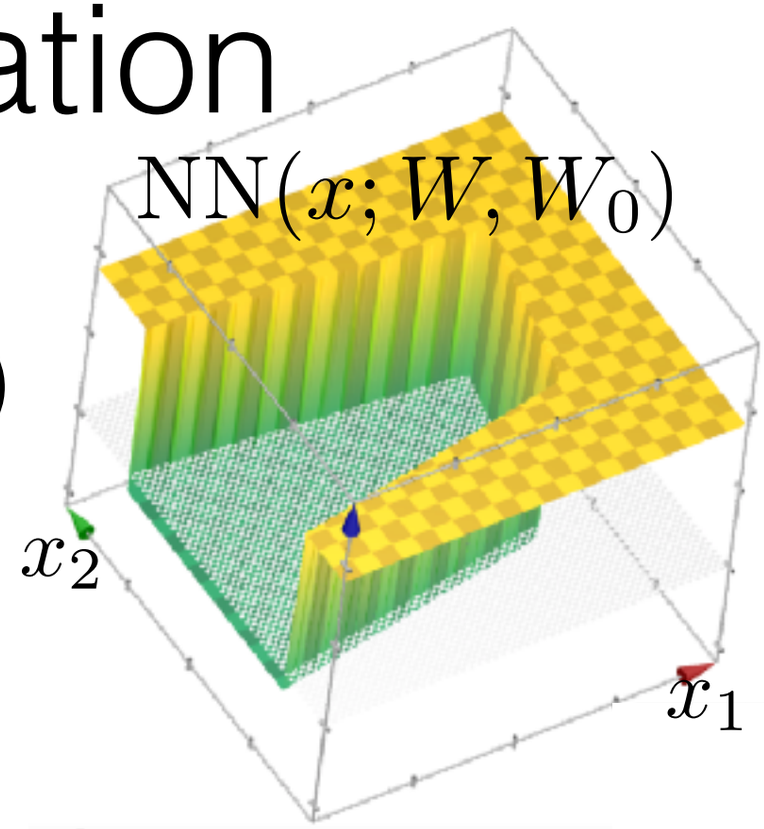

- Circle: function evaluation

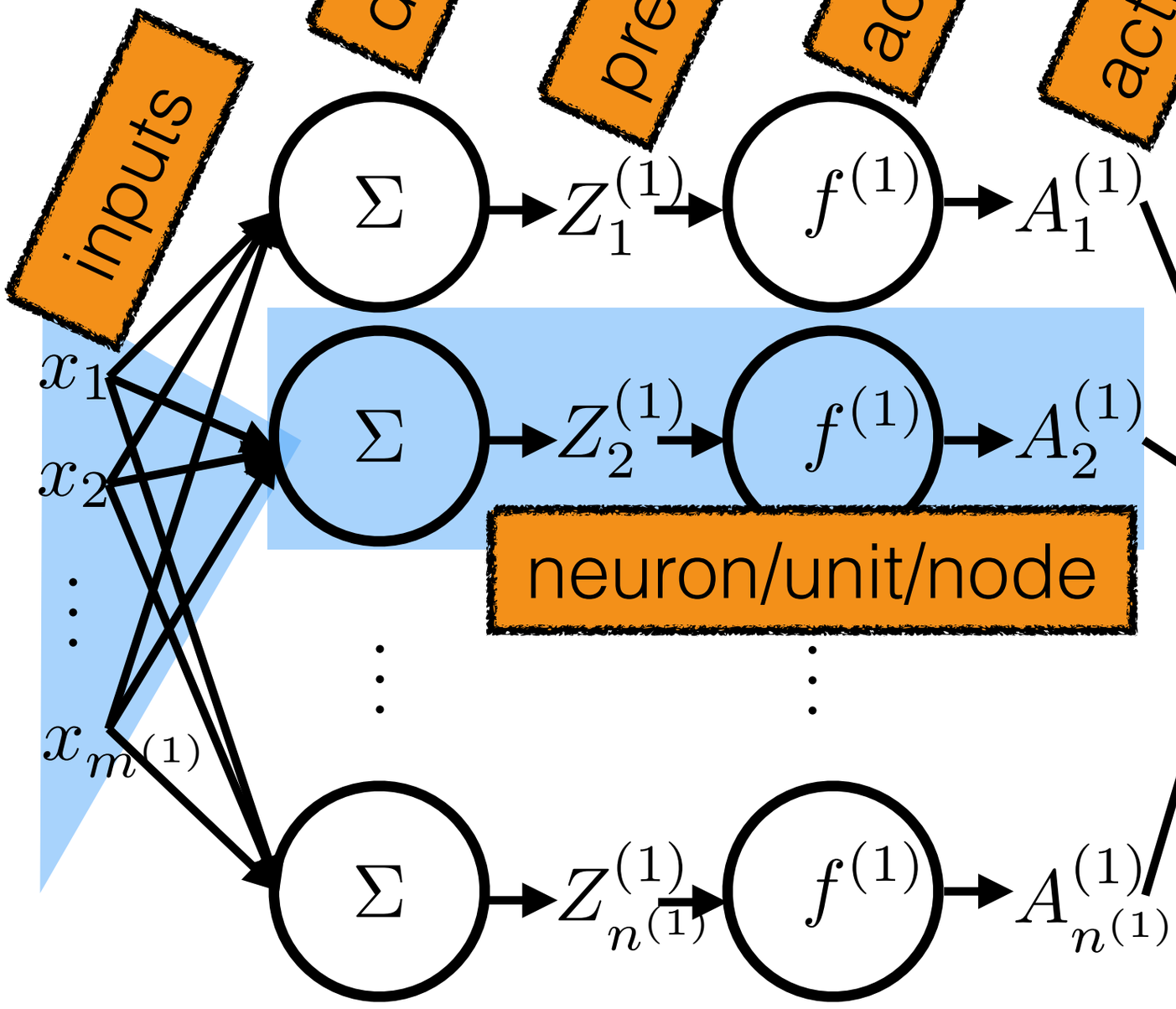



- Forward vs. backward
- A feed-forward neural network

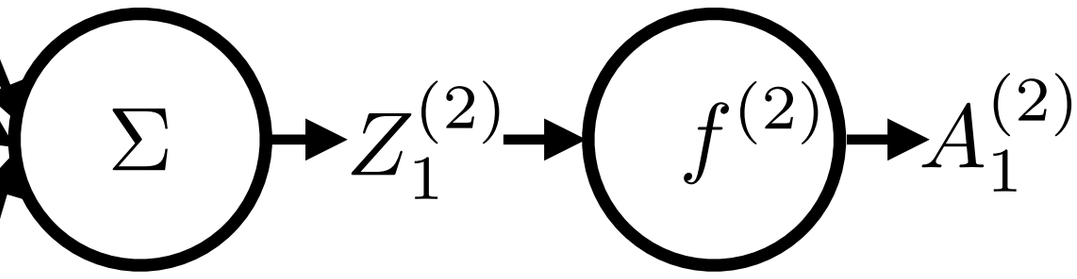
# Function graph representation

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



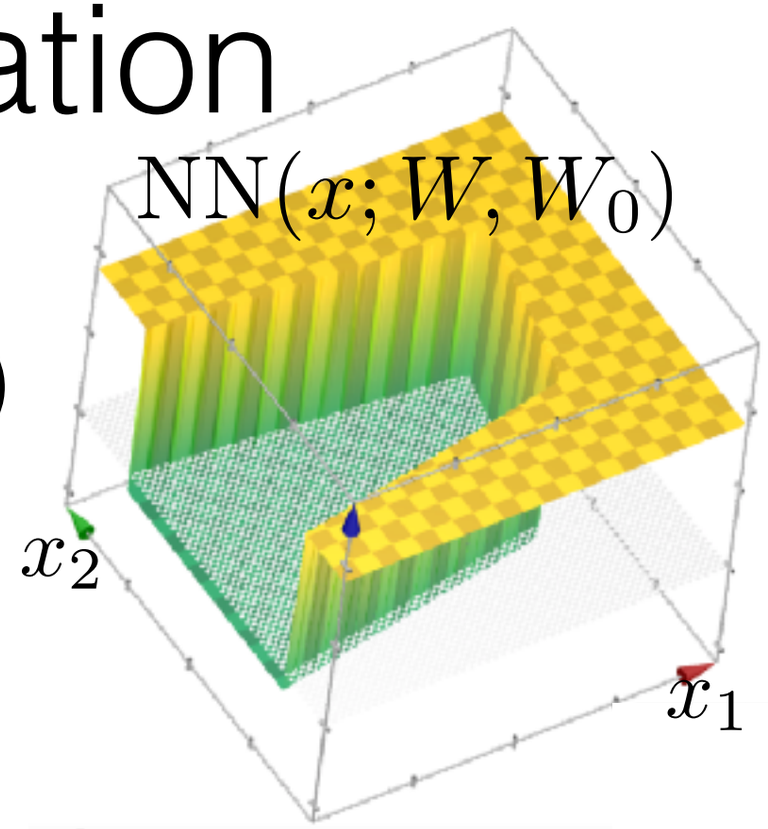

- Circle: function evaluation




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

# Function graph representation

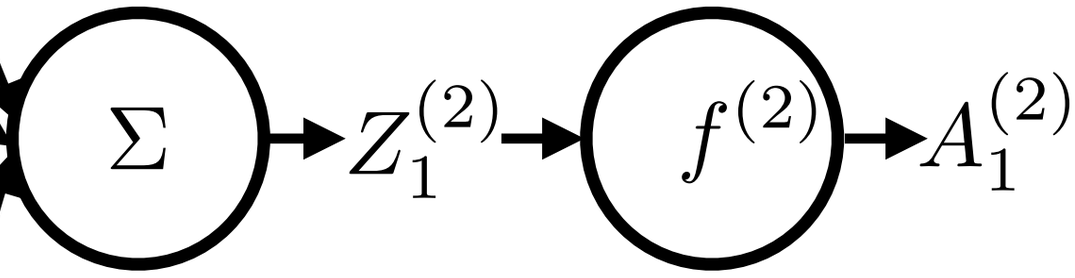
- 1st layer: $A^{(1)} = f^{(1)}(W^{(1)\top} x + W_0^{(1)})$
- 2nd layer: $A^{(2)} = f^{(2)}(W^{(2)\top} A^{(1)} + W_0^{(2)})$
- Whole thing: $A^{(2)} = \mathrm{NN}(x; W, W_0)$







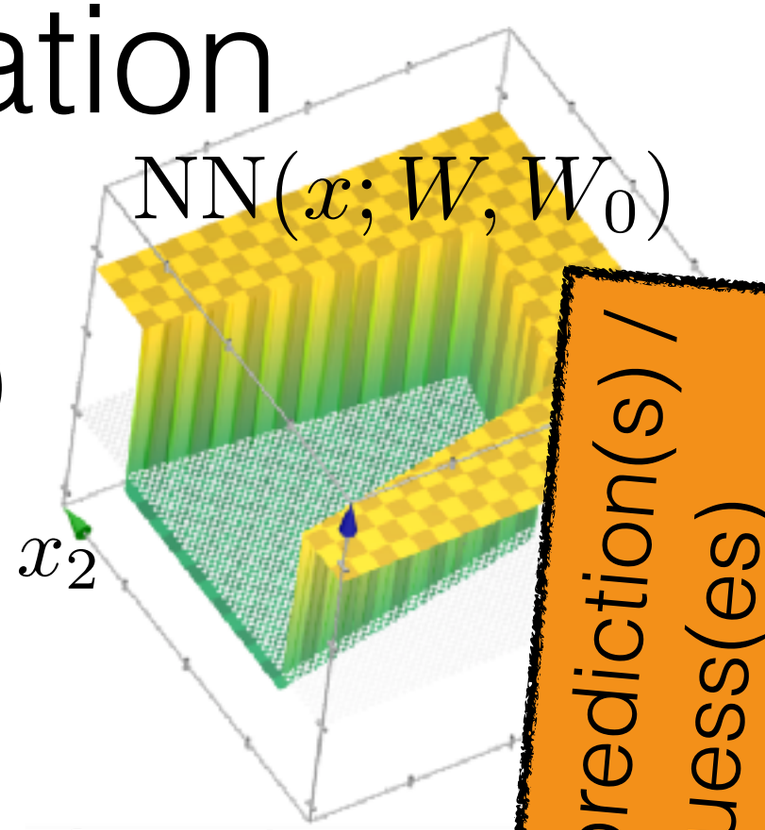
- Circle: function evaluation




- 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: $A^{(2)} = \mathrm{NN}(x; W, W_0)$



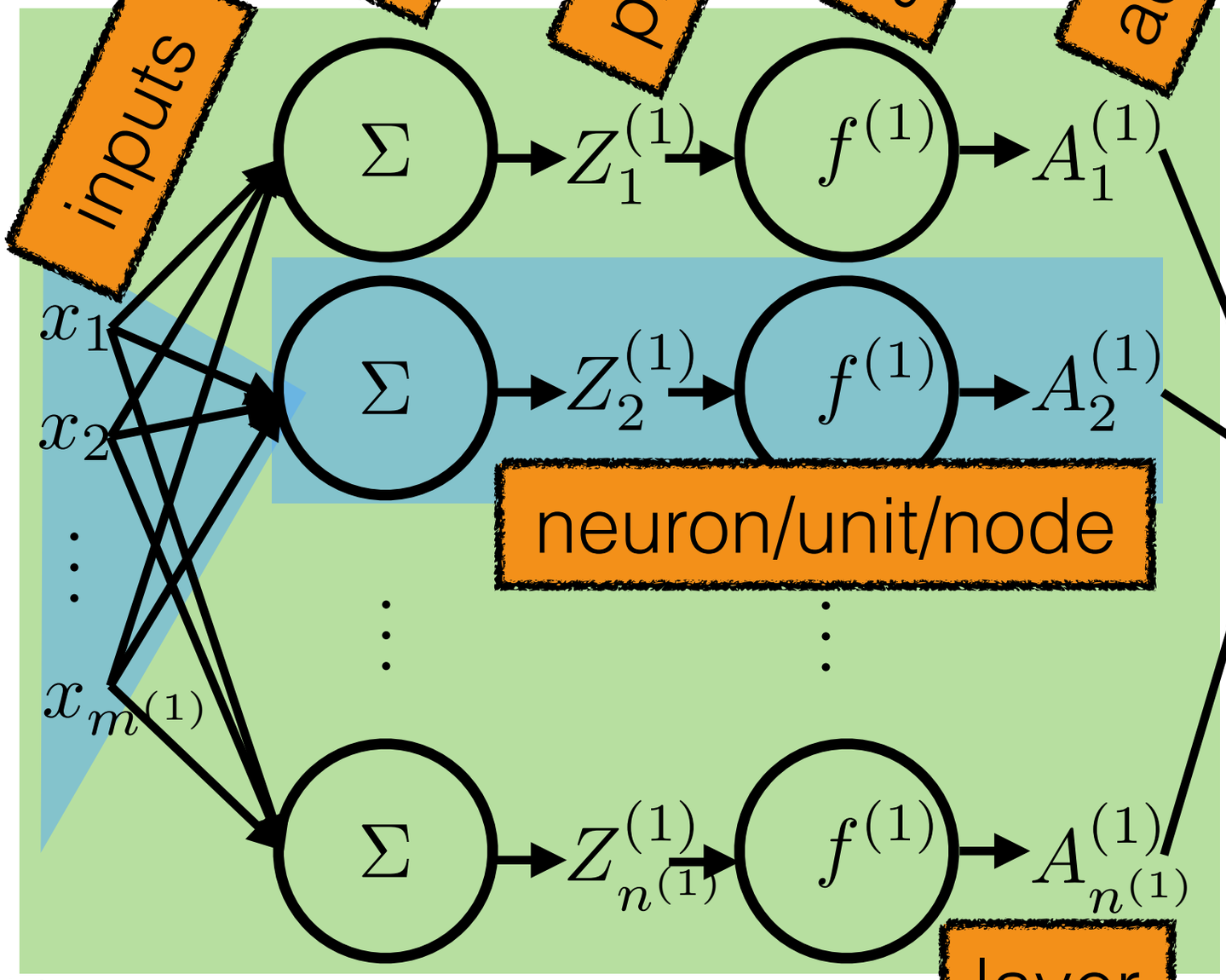

dot product

pre-activation

activation function

activation

inputs

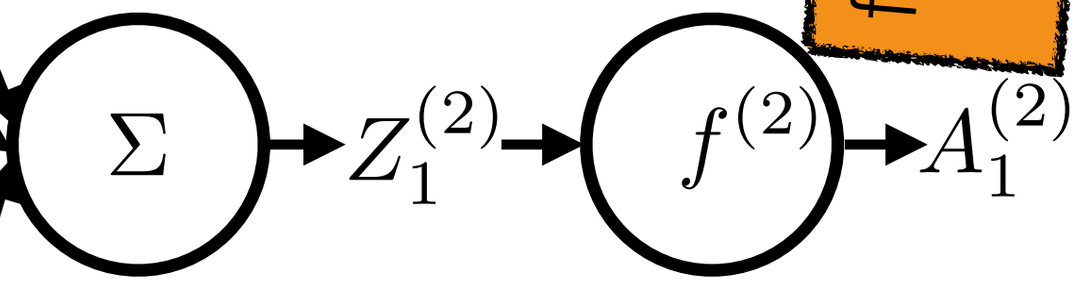



neuron/unit/node

layer

final prediction(s) / guess(es)

- Circle: function evaluation




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

# Function graph representation

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

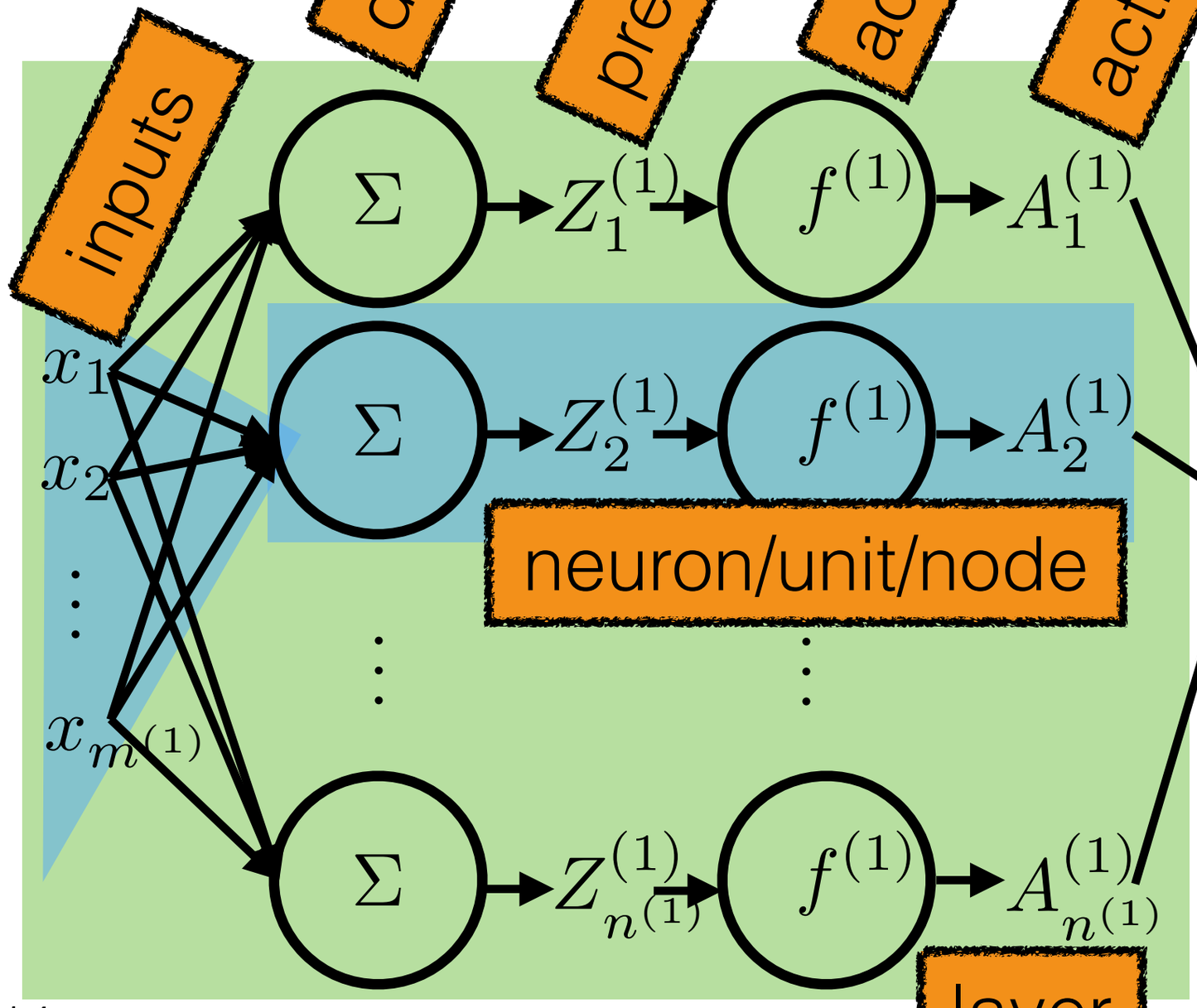
dot product

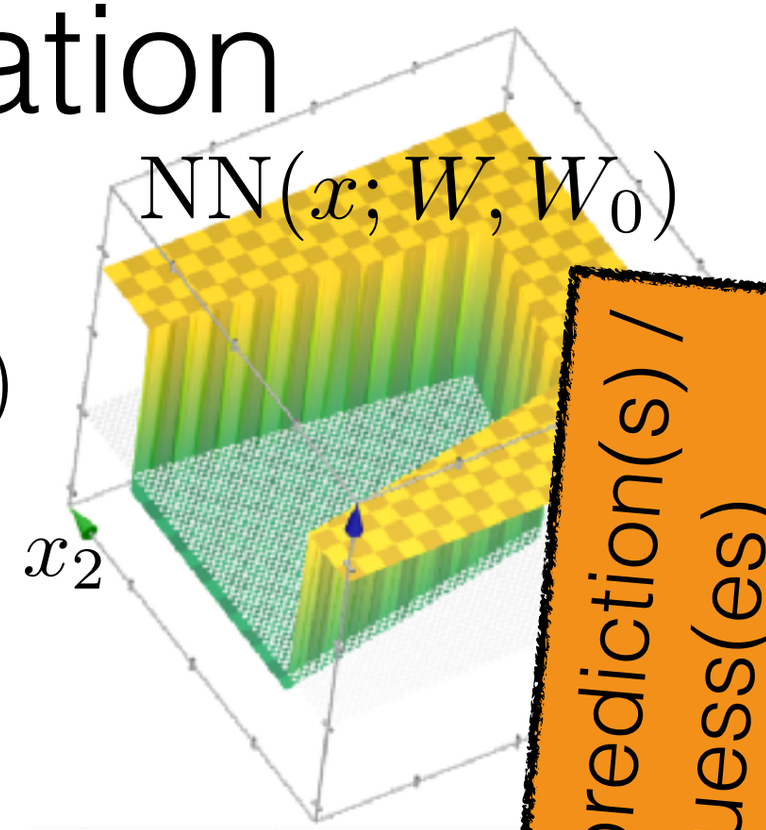
pre-activation

activation function

activation

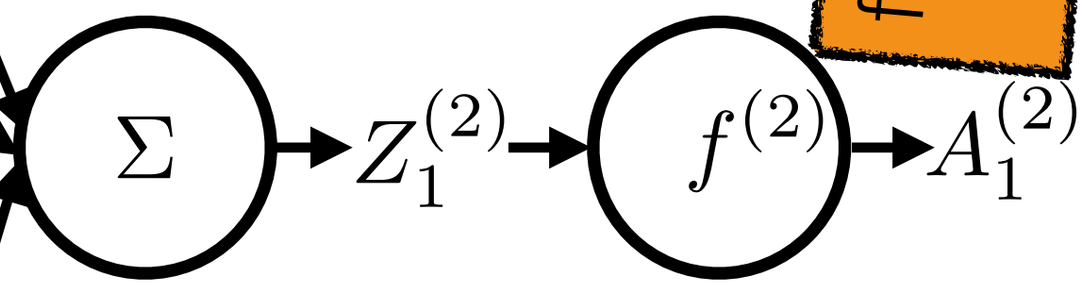
inputs




final prediction(s) / guess(es)

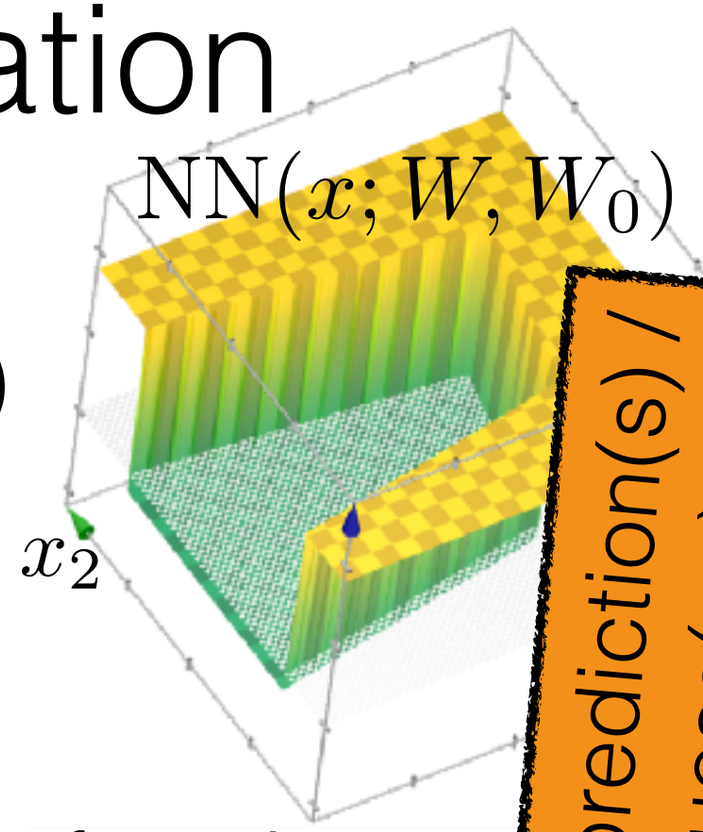
- Circle: function evaluation

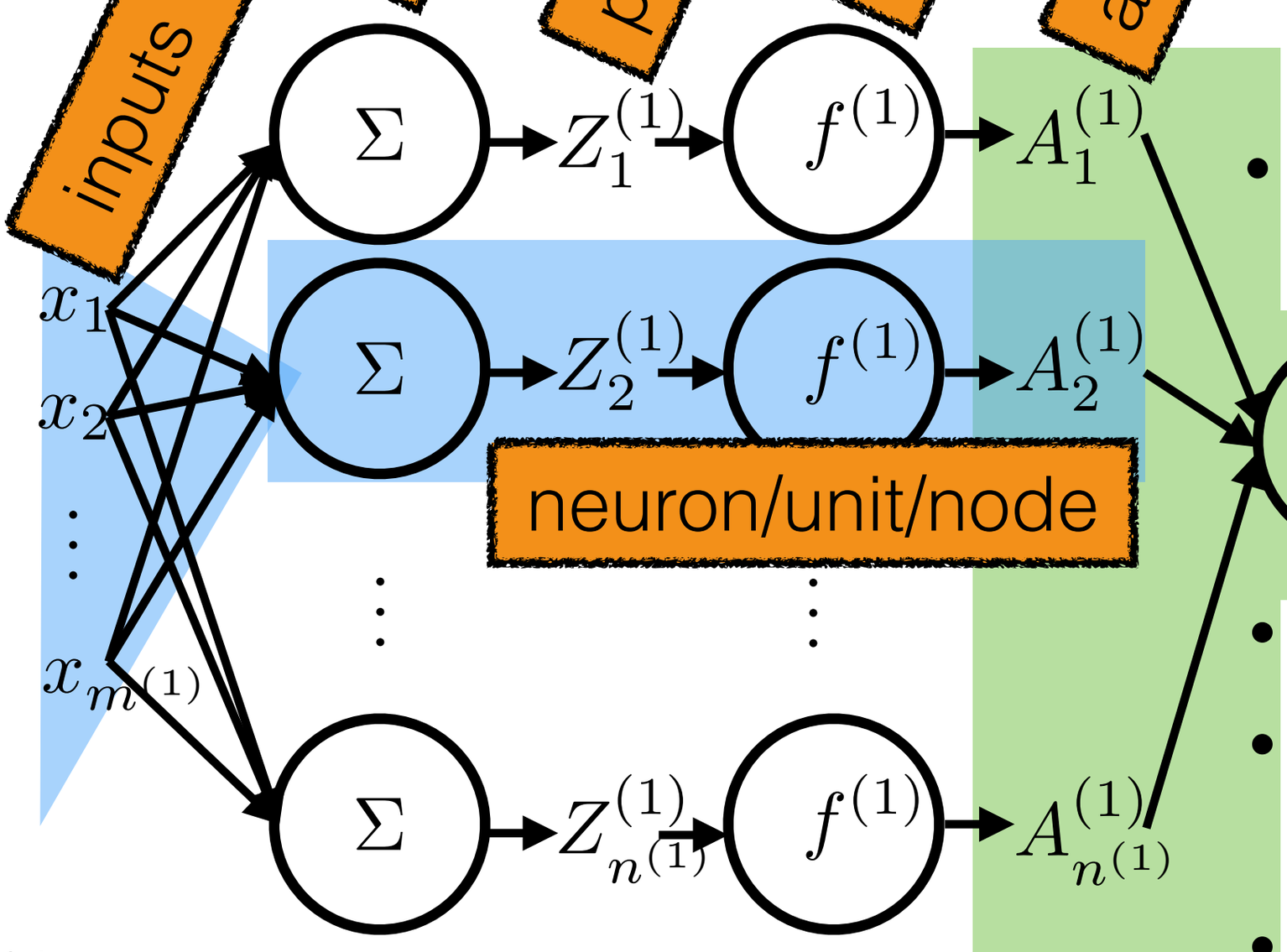


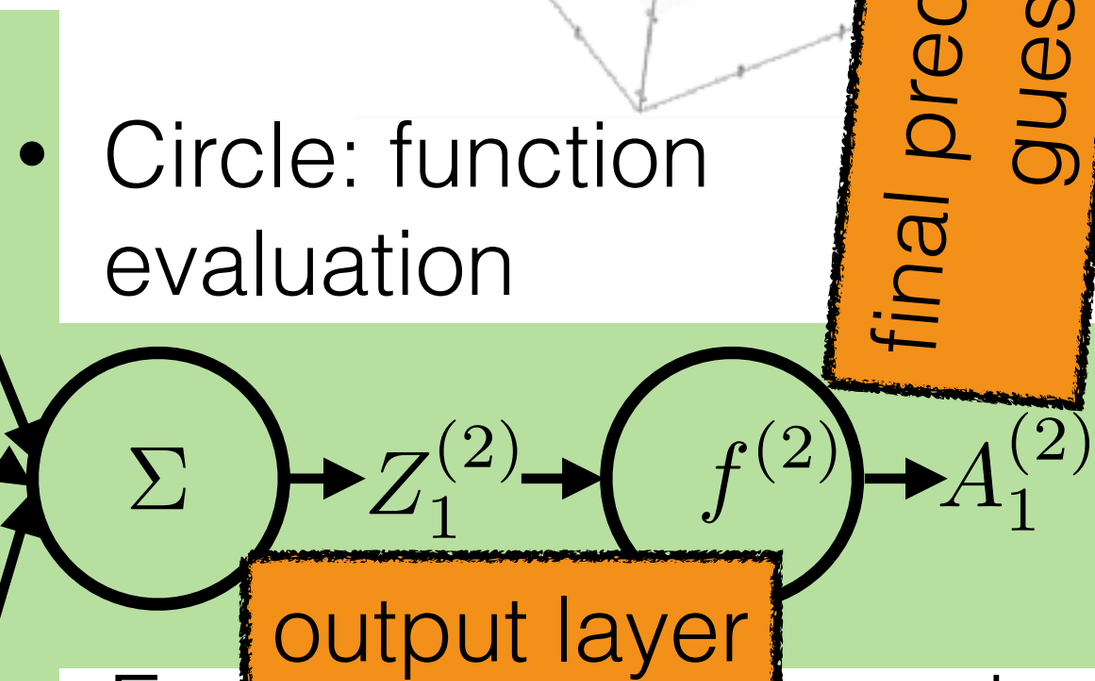

- Forward vs. backward
- A feed-forward neural network
- Fully connected

# Function graph representation

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




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

# Function graph representation

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

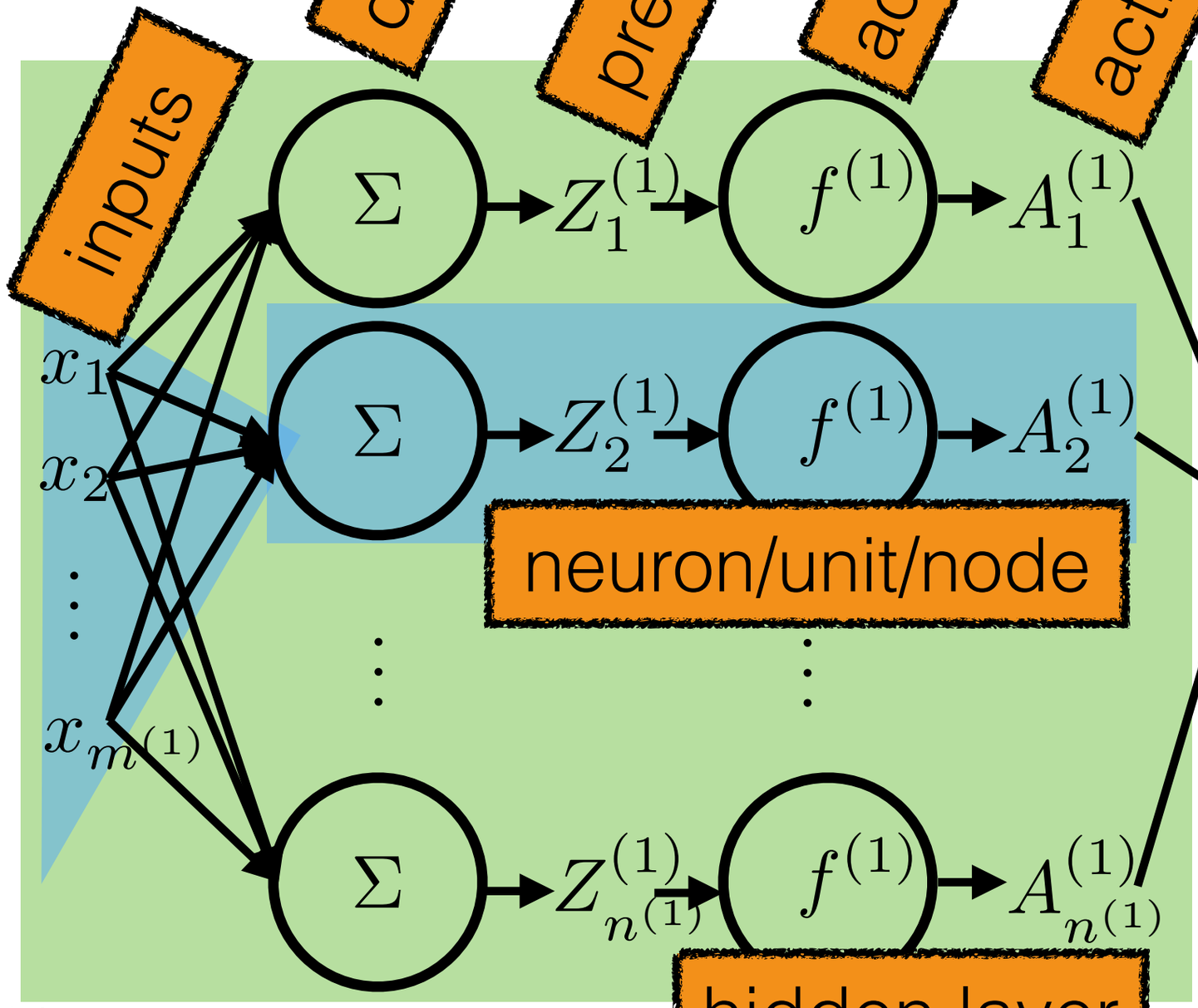
dot product

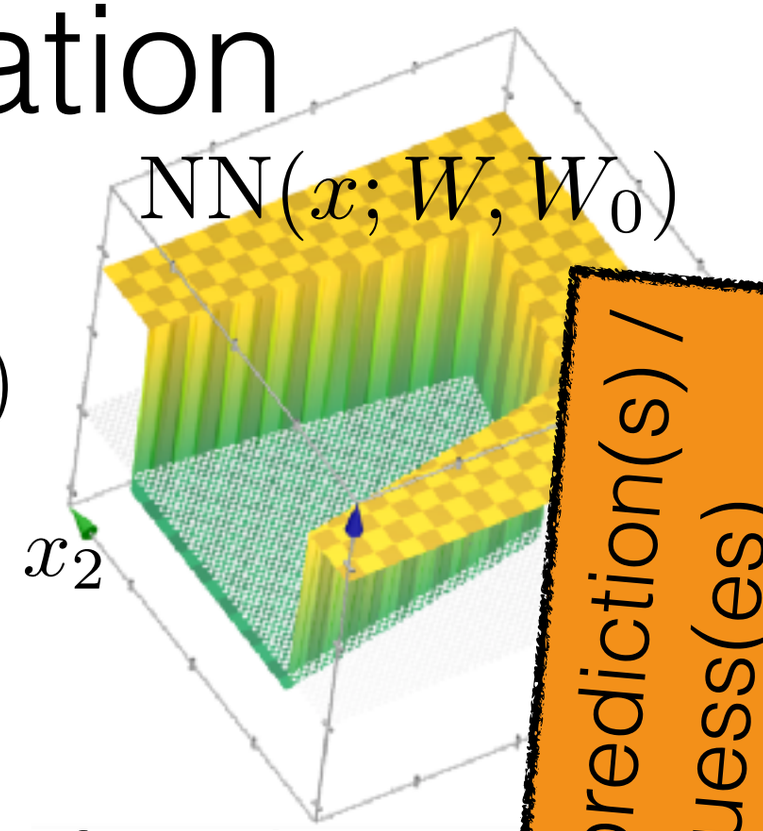
pre-activation

activation function

activation

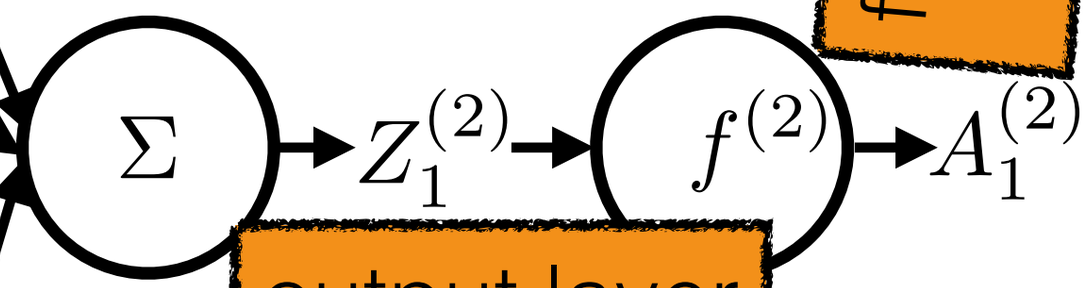
inputs




final prediction(s) / guess(es)

- Circle: function evaluation




output layer

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