# CSE 446
# Gradient Descent

Natasha Jaques

UNIVERSITY of WASHINGTON

# How do we find optimal weights?

- This is related to some questions you might have so far in this course
  - Why do we use quadratic loss, $\sum_{i=1}^{n} (y_i - w^T x_i)^2$ ?
  - Why is Gaussian noise so popular?

$$f(z) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{1}{2}\frac{(z-\mu)^2}{\sigma^2}}$$

  - Why was Ridge Regression with $L_2$ regularizer, $\|w\|_2^2$, the first to be used?
  - When we want sparsity, why do we use $L_1$ regularizer, $\|w\|_1$, and not $L_{0.5}$ regularizer, $\|w\|_{0.5}$ ?

**Convex!**

- The local minima is the global minimum

# Math is easier, but also….

# Easier to optimize! Because…?

# Why gradient descent?

- Standard ML paradigm: Define loss, then optimize

$$\hat{w}_{LS} = \arg\min_{w} ||\mathbf{y} - \mathbf{X}w||_2^2 \quad +\lambda\|w\|_1$$

$$2X^T(Xw - y) = 0$$

$$\hat{w} = (X^TX)^{-1}X^Ty$$

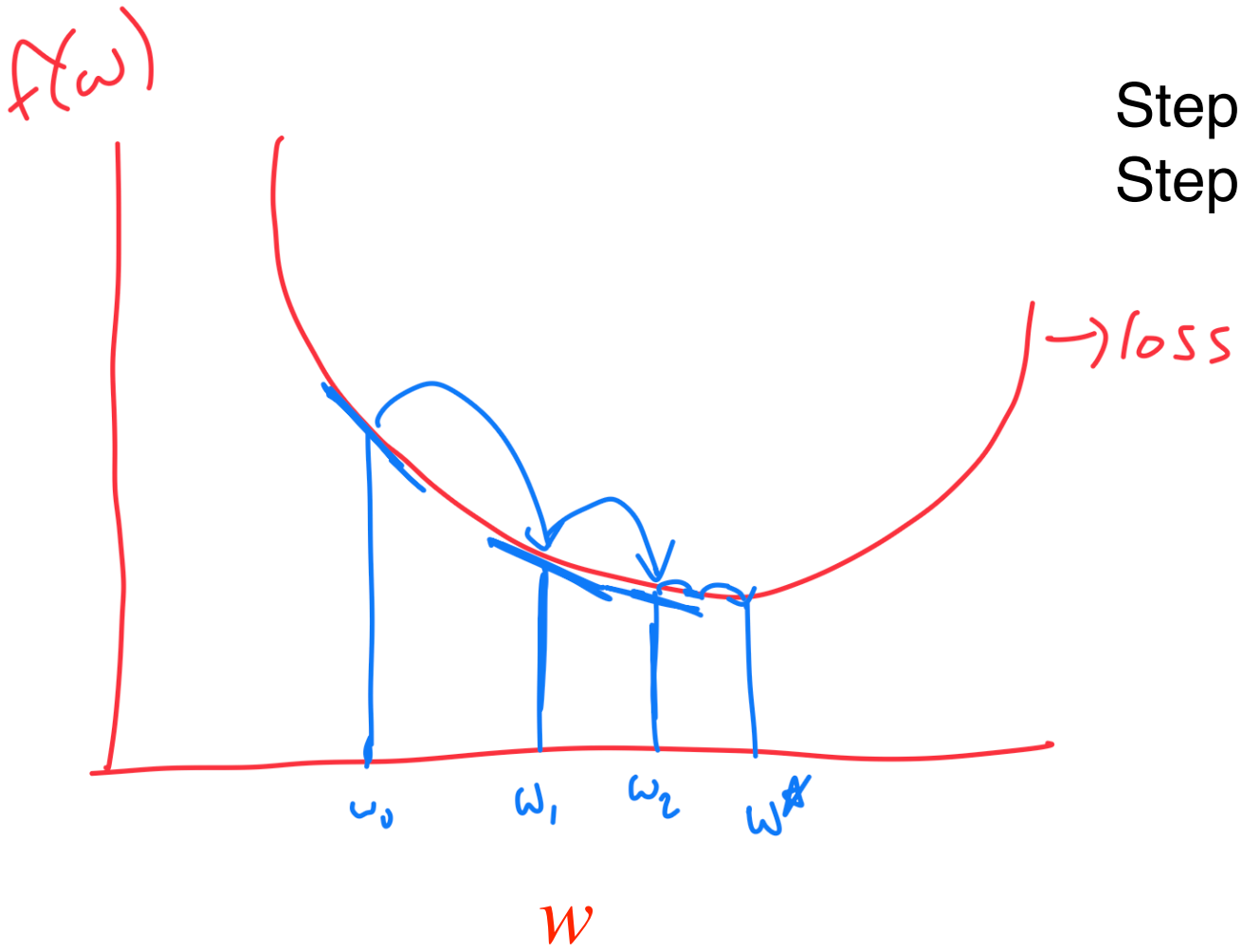
# No closed form solution!

- But, no closed-form solutions for most losses we use in practice.
- Key idea: Iterative methods
- Used everywhere!

# Start with a guess for w

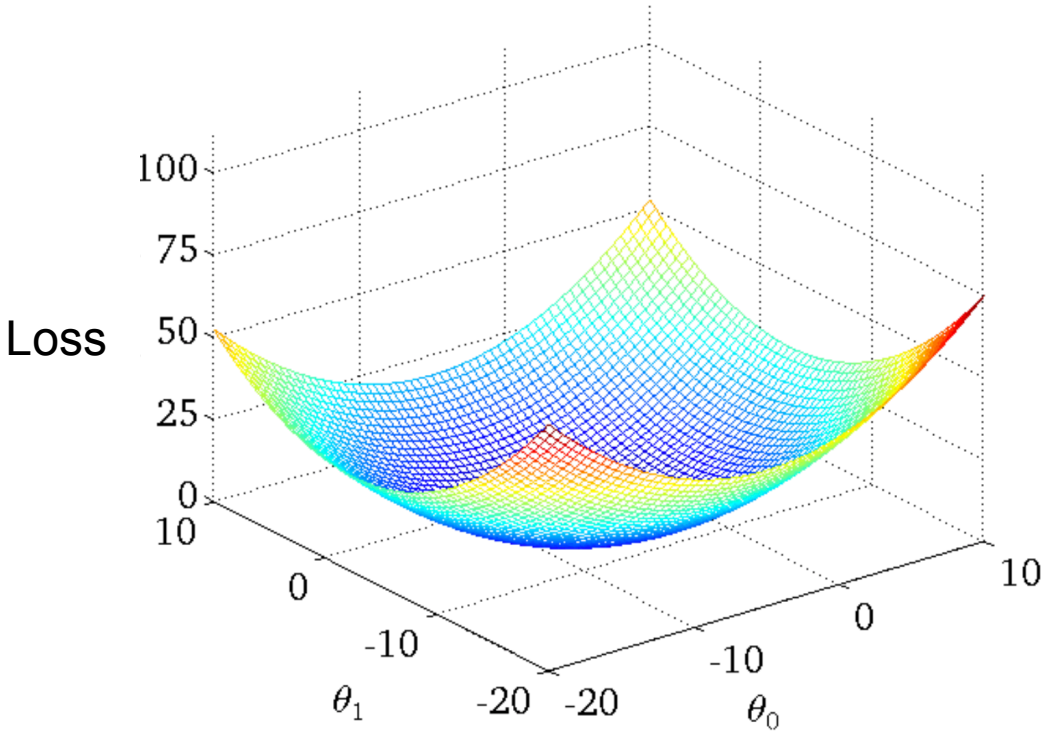
# Iteratively refine to minimize loss

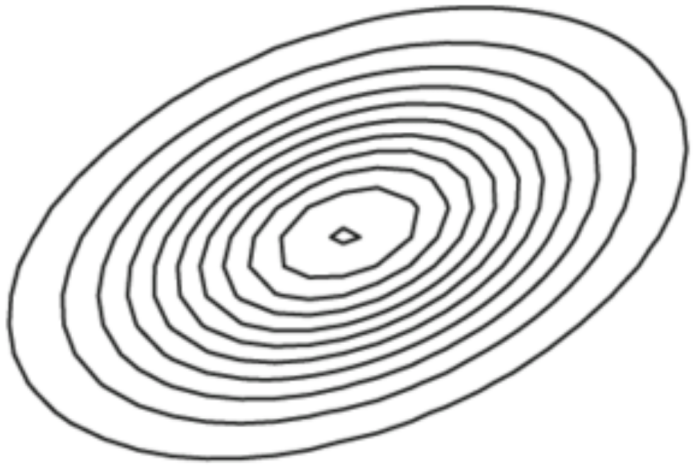
# Gradient descent in one dimension




Step direction: –gradient
Step size: η |gradient|

# Gradient descent in multiple dimensions



# Lecture plan

- Gradient descent algorithm + examples
- Theoretical analysis
  - When does it work?
  - How quickly does it converge?
  - How do we choose a step size?
  - Key idea: Convexity
- Not tested on proof details, but concepts are important & practical

# Algorithm: Gradient descent

# t is a step of the algorithm

**Algorithm**

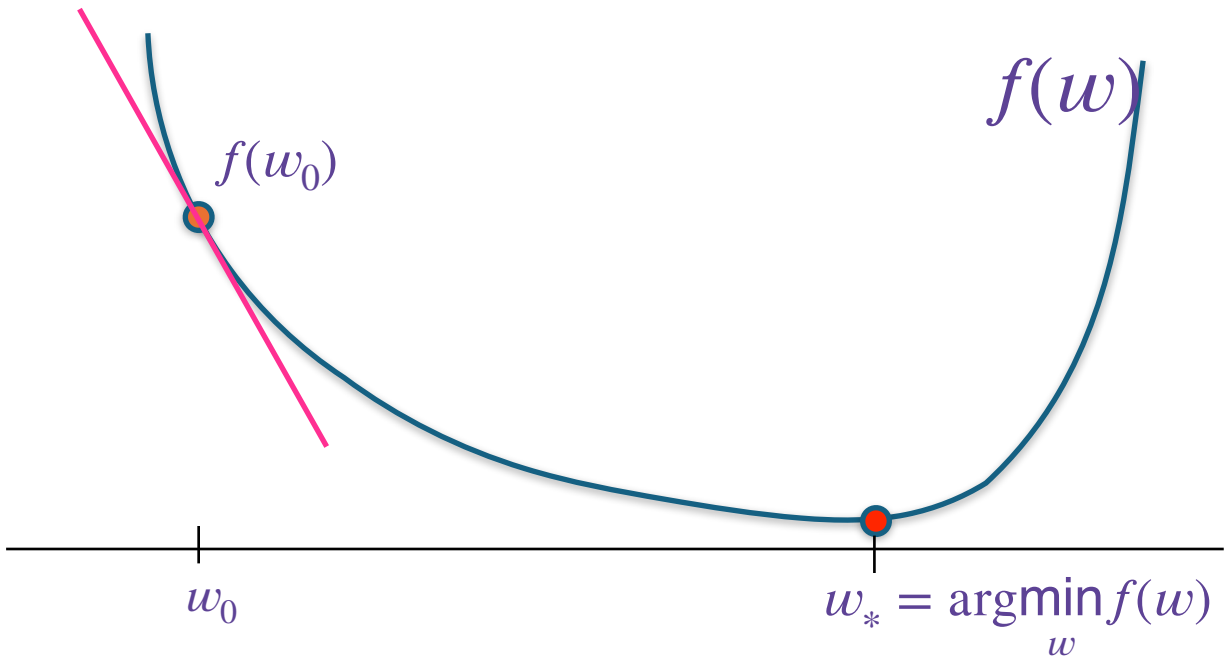
For t=0,1,2,3, …

# Gradient

$$w_{t+1} = w_t - \eta \frac{df(w)}{dw}\Big|_{w=w_t}$$

Hyperparameters:

- Initial point $w_0$
- Step size $\eta$




# How do we pick initial weights $w_0$?

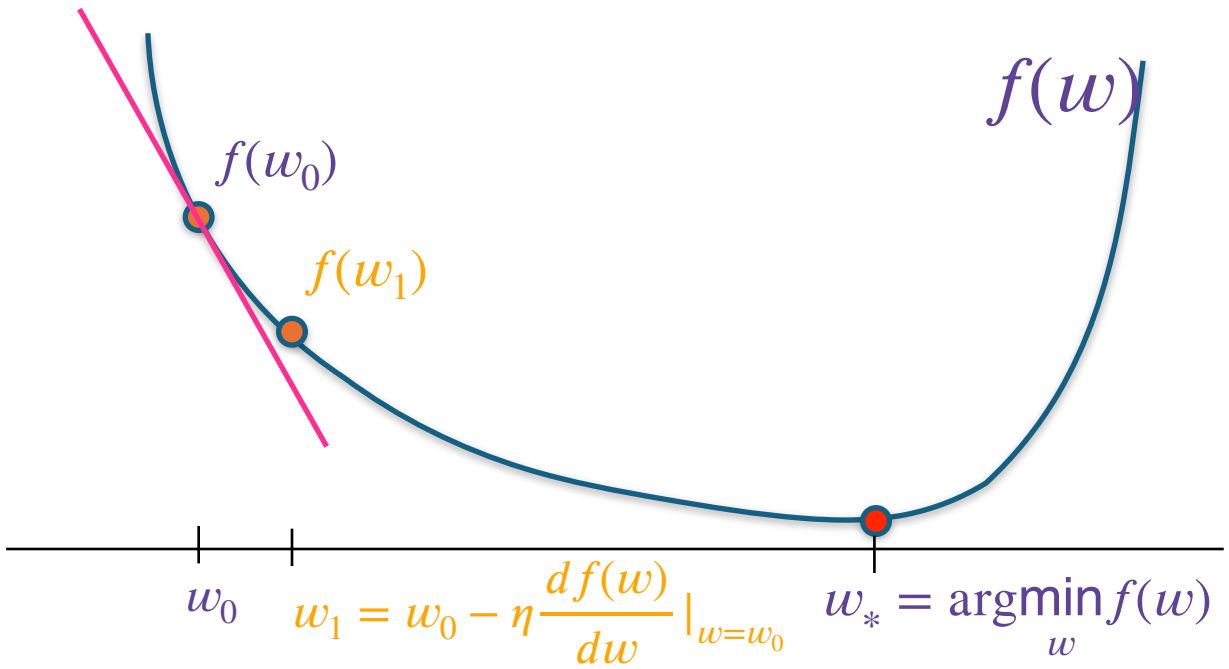
$$w_0 \sim \mathcal{N}(0, I_{d\times d}\sigma^2)$$

# Algorithm: Gradient descent

**Algorithm**
For t=0,1,2,3, …

$$w_{t+1} = w_t - \eta \frac{df(w)}{dw}|_{w=w_t}$$

Hyperparameters:
- Initial point $w_0$
- Step size $\eta$

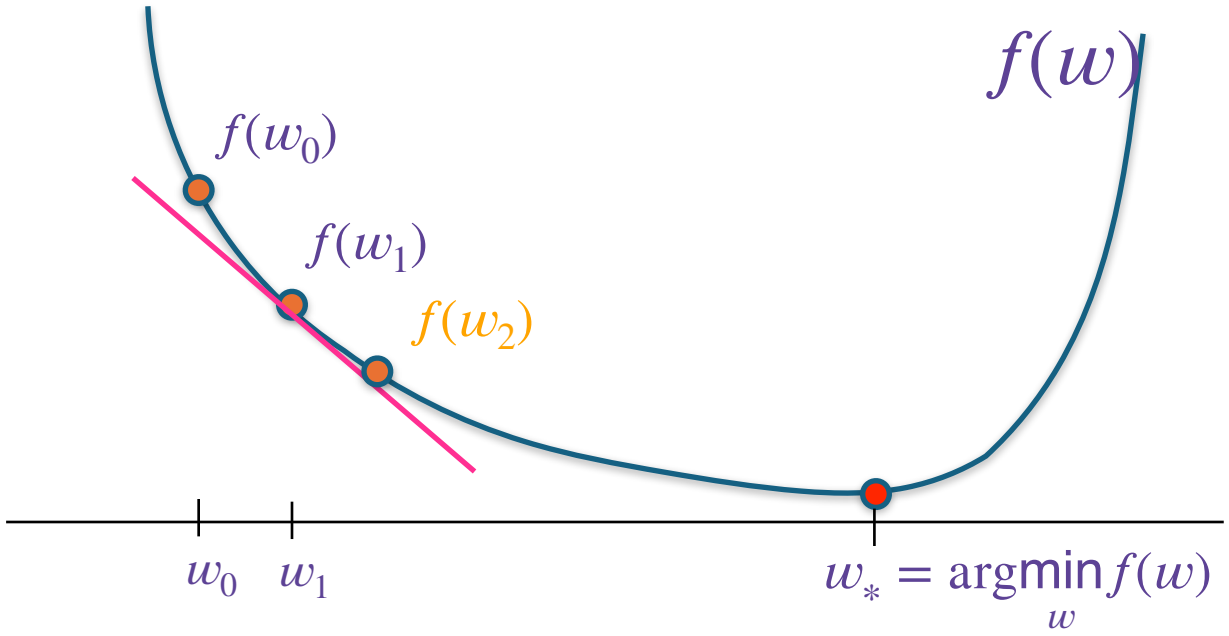


# Algorithm: Gradient descent

**Algorithm**
For t=0,1,2,3, …

$$w_{t+1} = w_t - \eta \frac{df(w)}{dw}\Big|_{w=w_t}$$

Hyperparameters:
- Initial point $w_0$
- Step size $\eta$

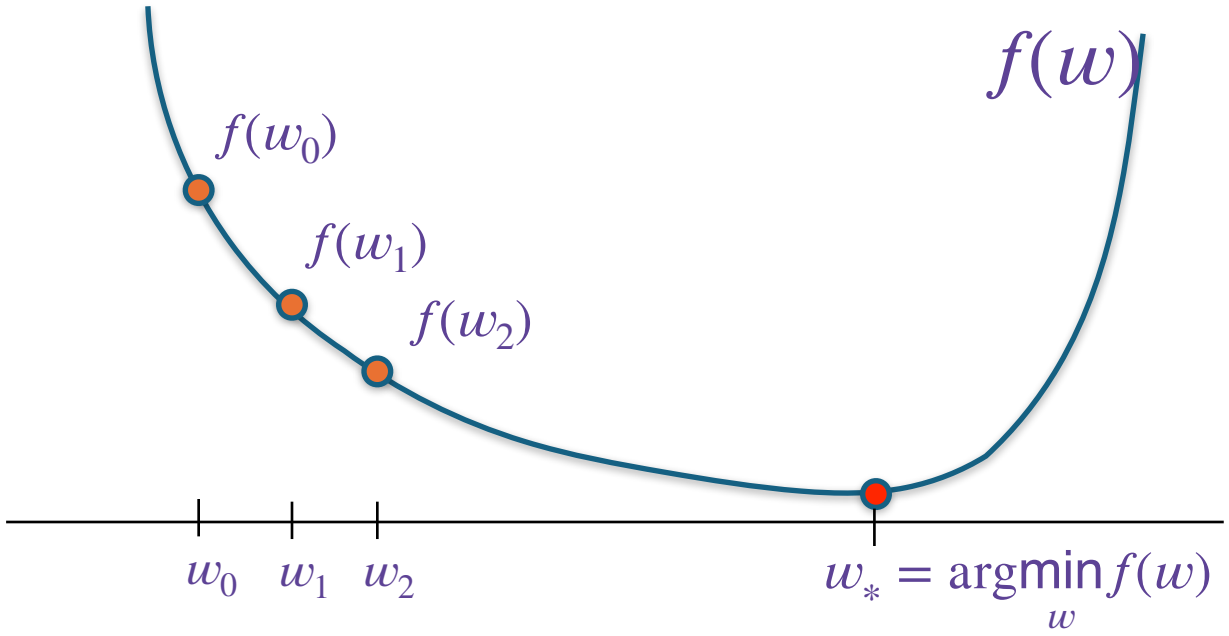


# Algorithm: Gradient descent

**Algorithm**
For t=0,1,2,3, …

$$w_{t+1} = w_t - \eta \frac{df(w)}{dw}\Big|_{w=w_t}$$

Hyperparameters:
- Initial point $w_0$
- Step size $\eta$

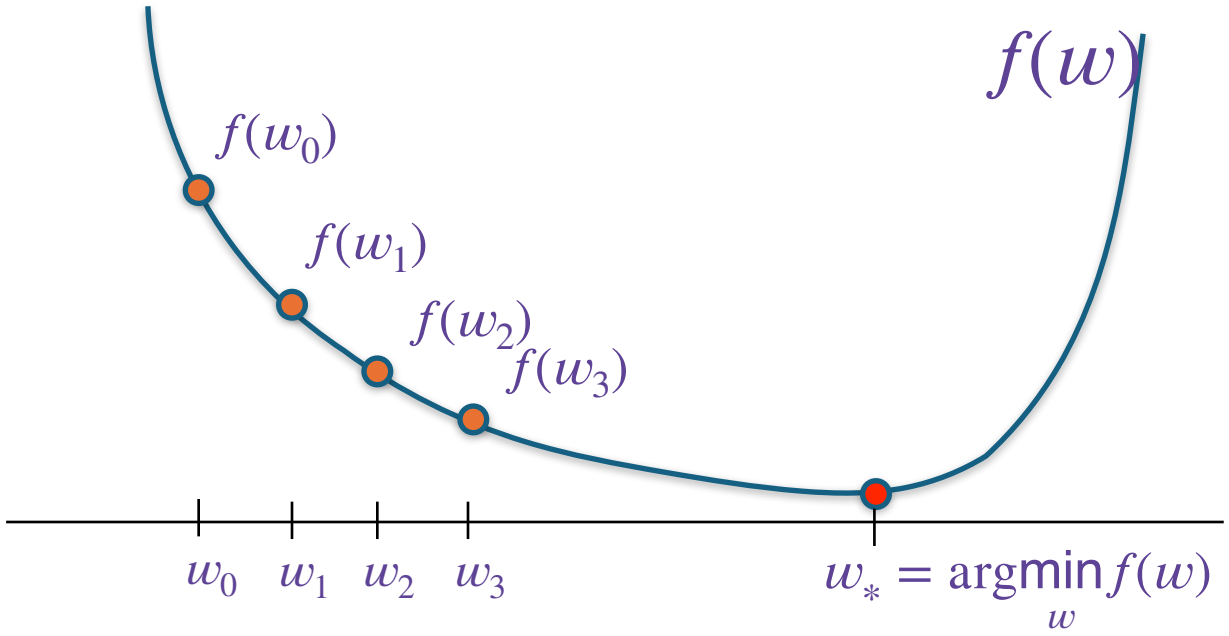


# Algorithm: Gradient descent

**Algorithm**
For t=0,1,2,3, …

$$w_{t+1} = w_t - \eta \frac{df(w)}{dw}\Big|_{w=w_t}$$

Hyperparameters:
- Initial point $w_0$
- Step size $\eta$

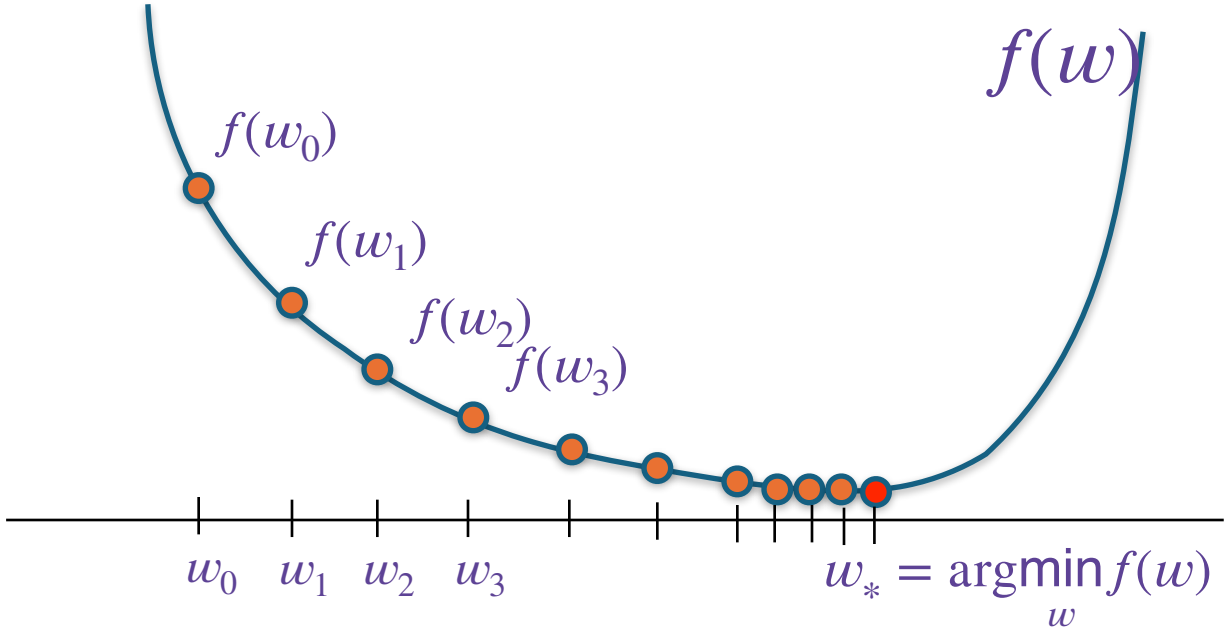


# Algorithm: Gradient descent

**Algorithm**
For t=0,1,2,3, …

$$w_{t+1} = w_t - \eta \frac{df(w)}{dw}\Big|_{w=w_t}$$

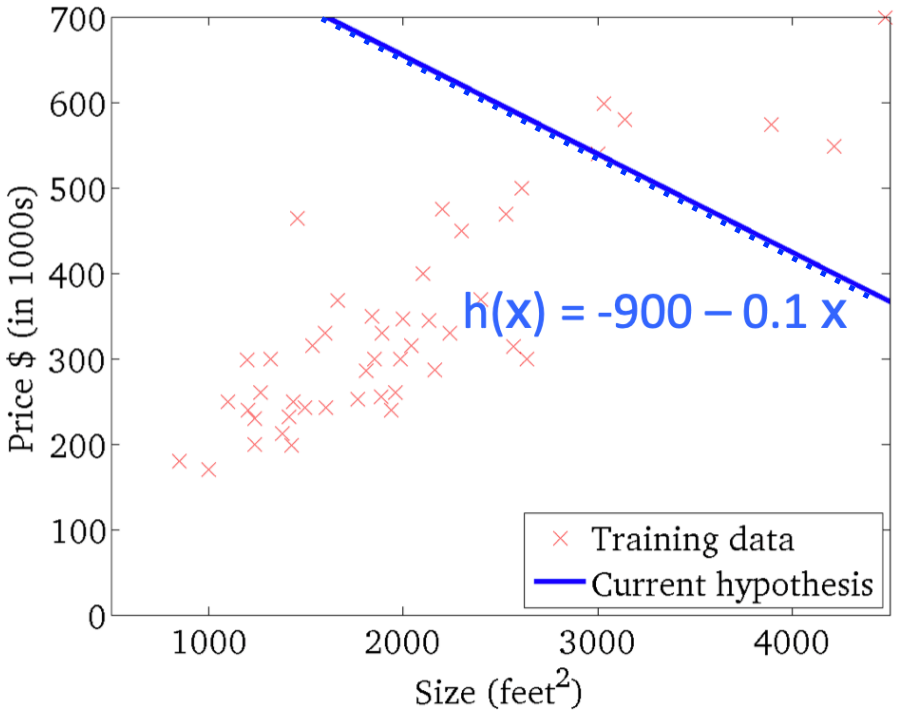
Hyperparameters:
- Initial point $w_0$
- Step size $\eta$



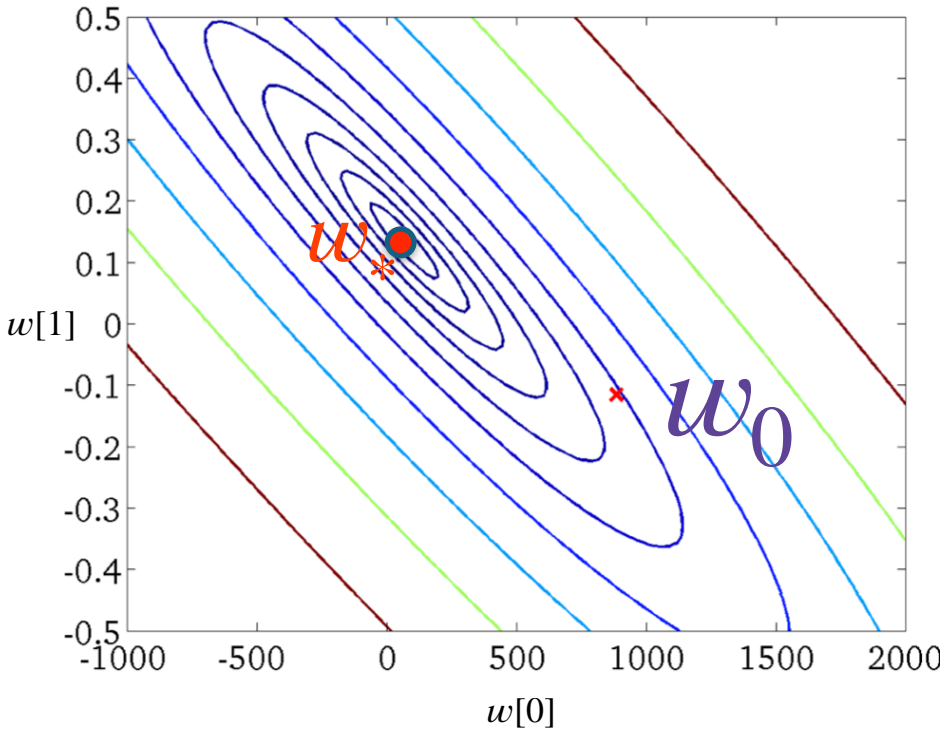

Note that as $t \to \infty$ we have $\frac{df(w)}{dw}\Big|_{w=w_t} \to 0$

# 1-dimensional linear regression with 2 parameters

$$\{(x_i, y_i)\}_{i=1}^{n}$$




$$w_{t+1} \leftarrow w_t - \eta \cdot \nabla_w f(w_t)$$

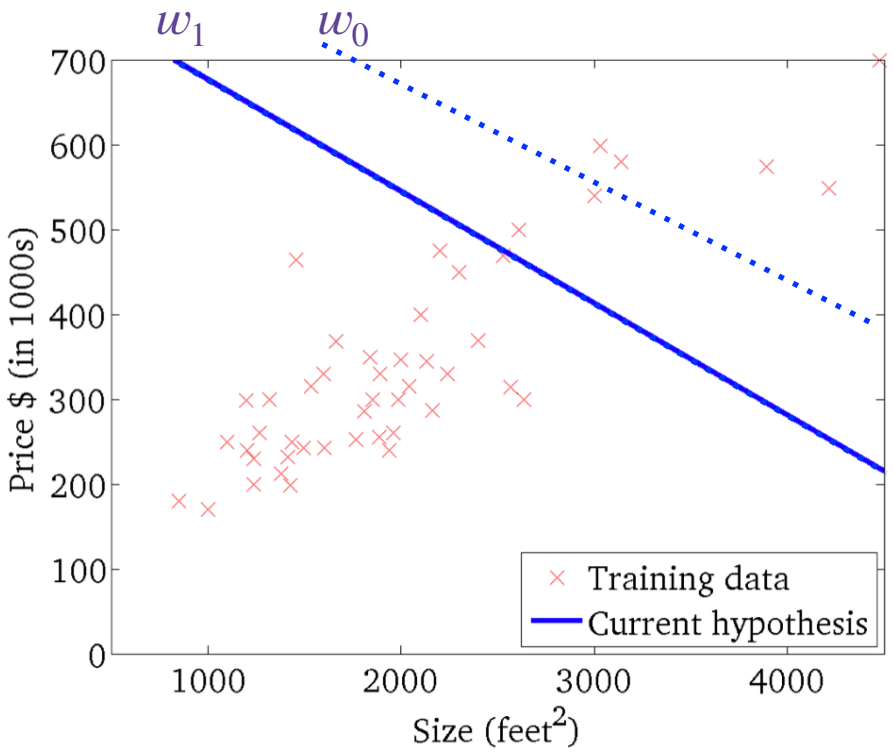


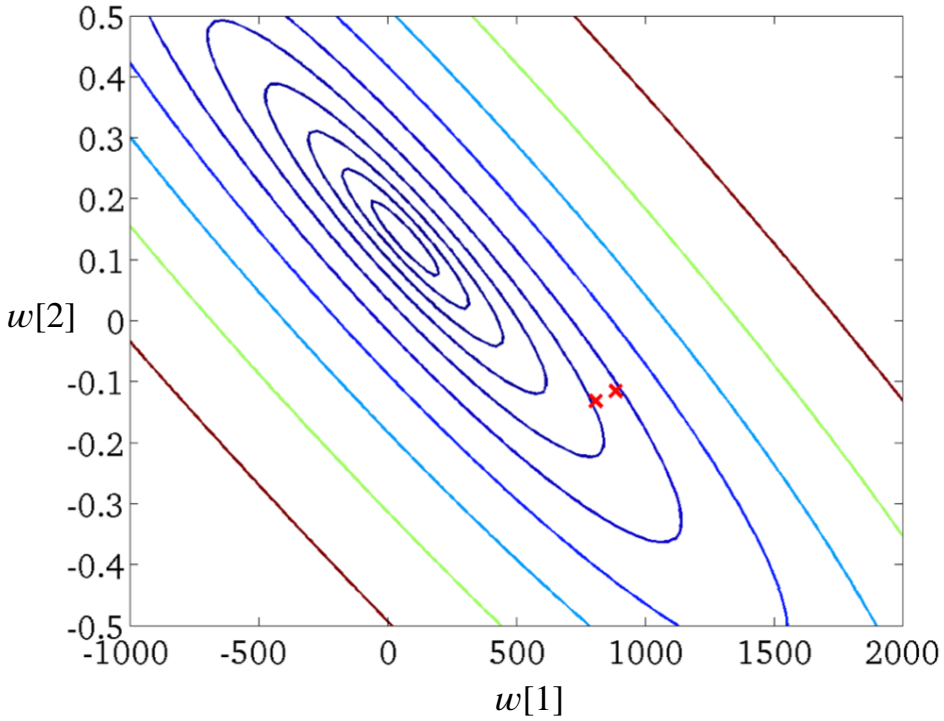

Evolution of the predictor y = w[0] + w[1]x

Gradient descent dynamics in the parameter space w
Ovals show the **level set** of the objective function

# 1-dimensional linear regression with 2 parameters

# current model

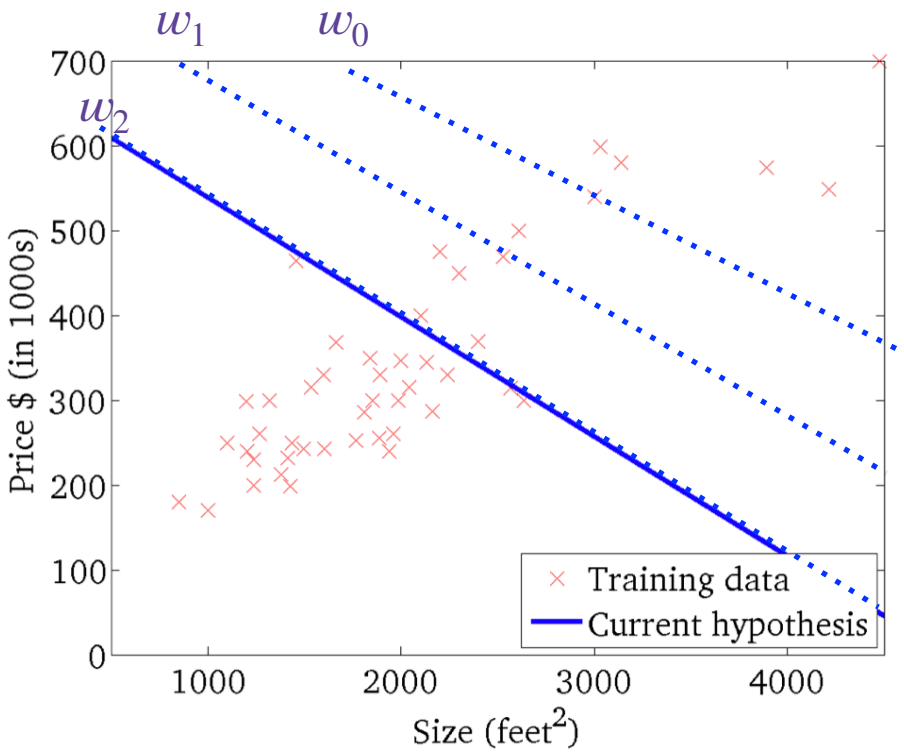


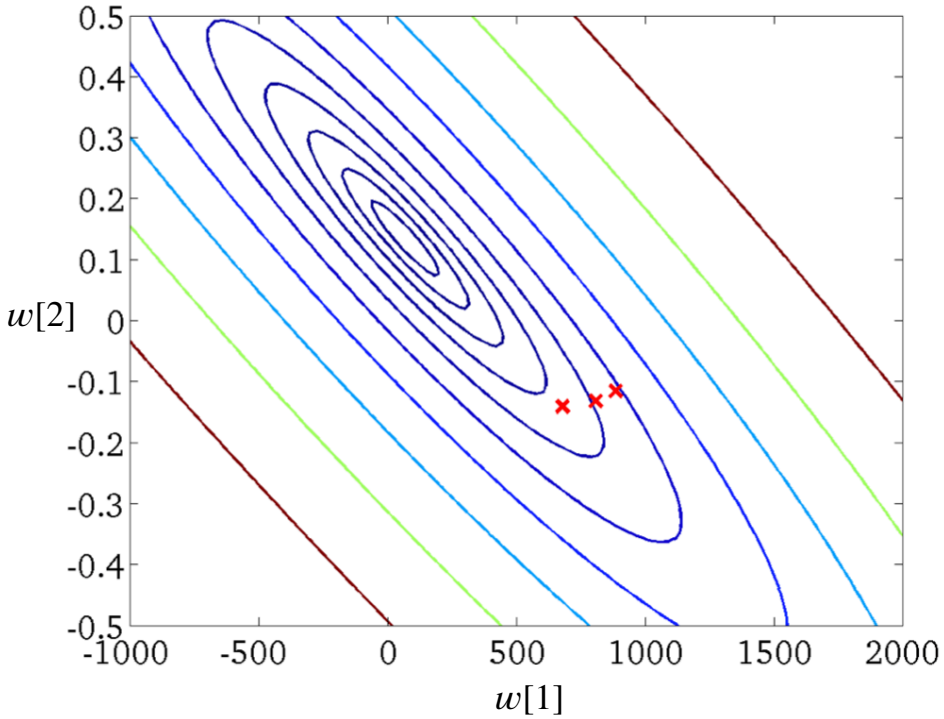

Evolution of the predictor y = w[0] + w[1]x

Gradient descent dynamics in the parameter space w
Ovals show the **level set** of the objective function

# 1-dimensional linear regression with 2 parameters

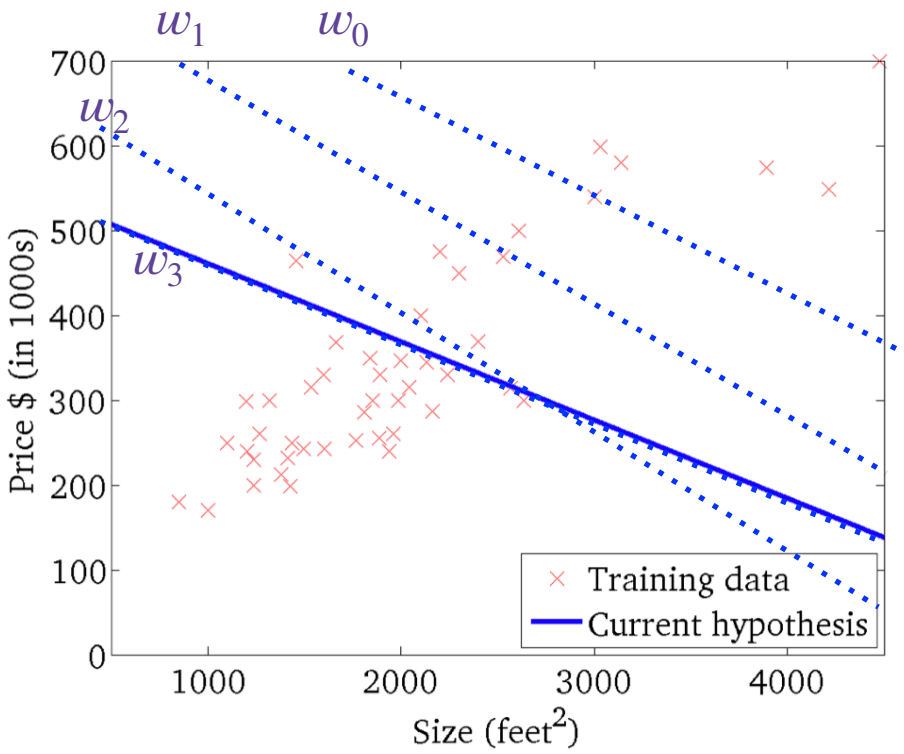



Evolution of the predictor y = w[0] + w[1]x

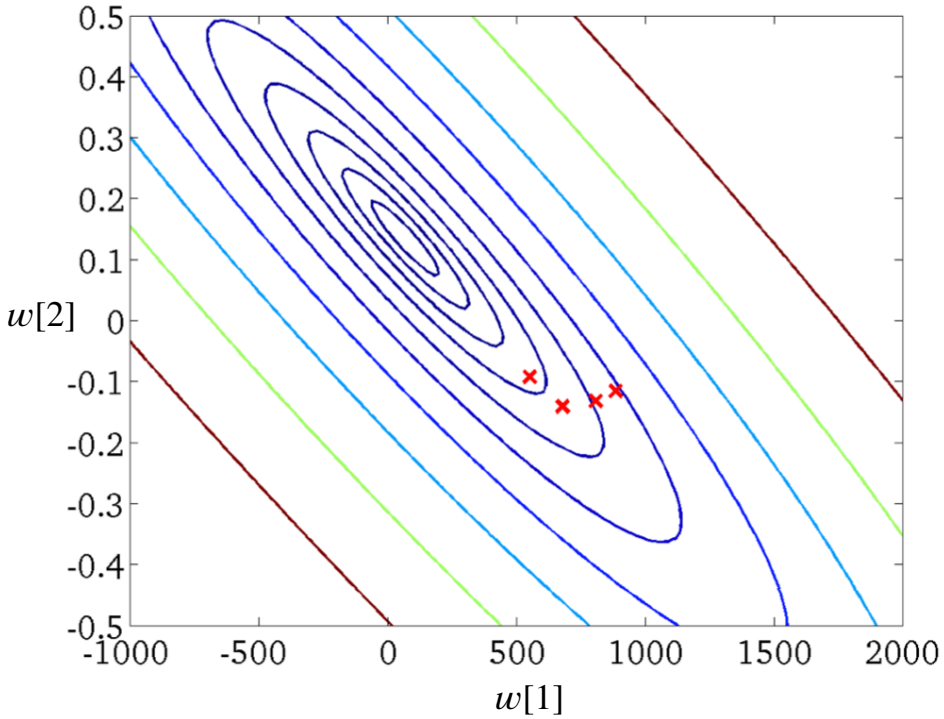



Gradient descent dynamics in the parameter space w
Ovals show the **level set** of the objective function

# 1-dimensional linear regression with 2 parameters

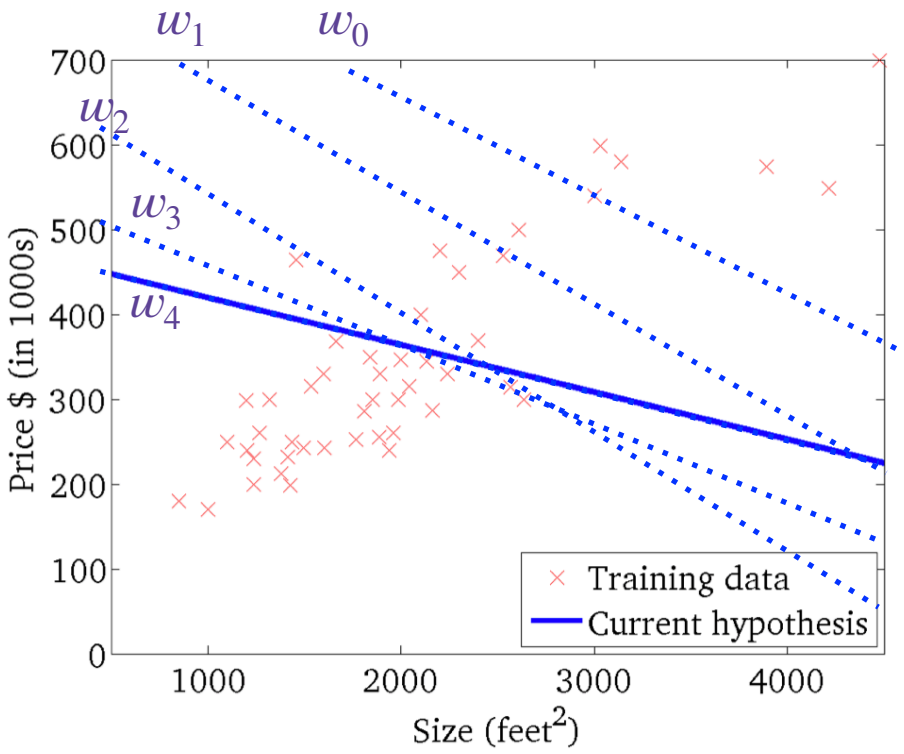



Evolution of the predictor y = w[0] + w[1]x

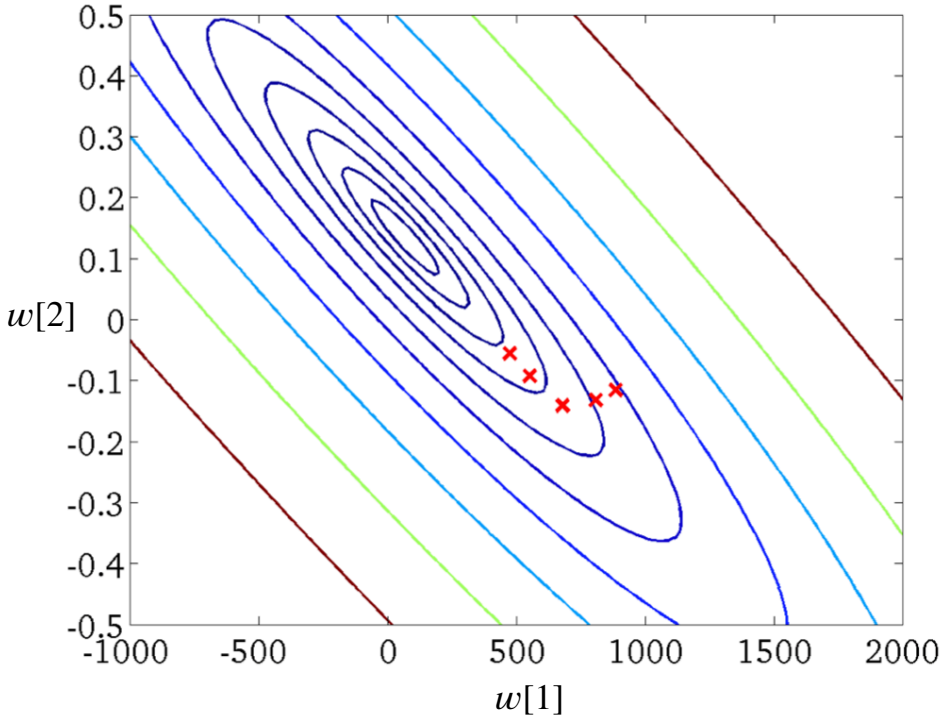



Gradient descent dynamics in the parameter space w
Ovals show the **level set** of the objective function

# 1-dimensional linear regression with 2 parameters

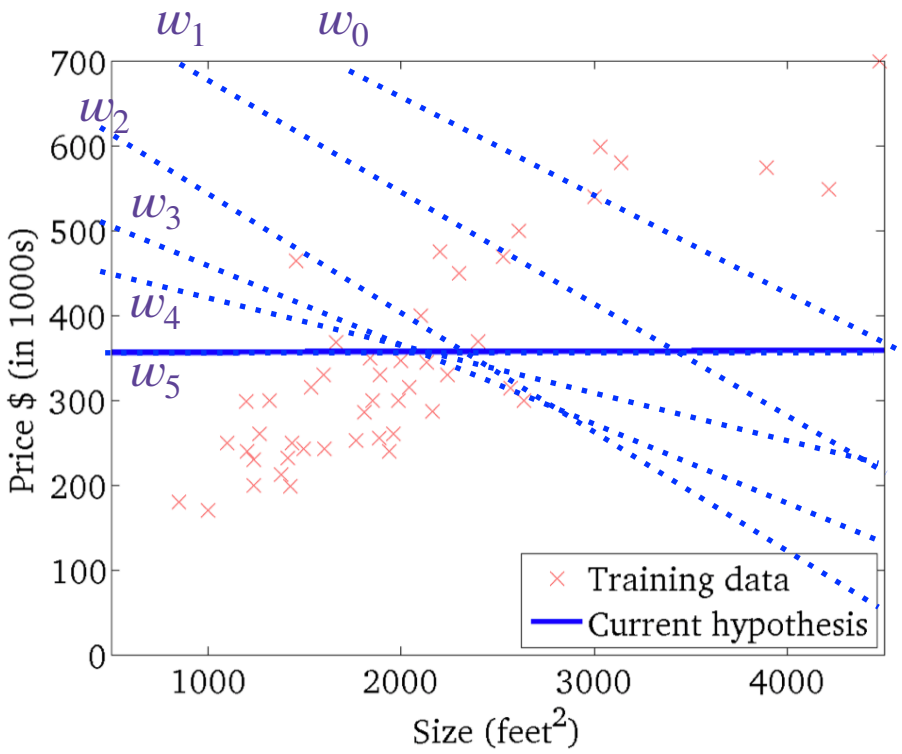



Evolution of the predictor y = w[0] + w[1]x

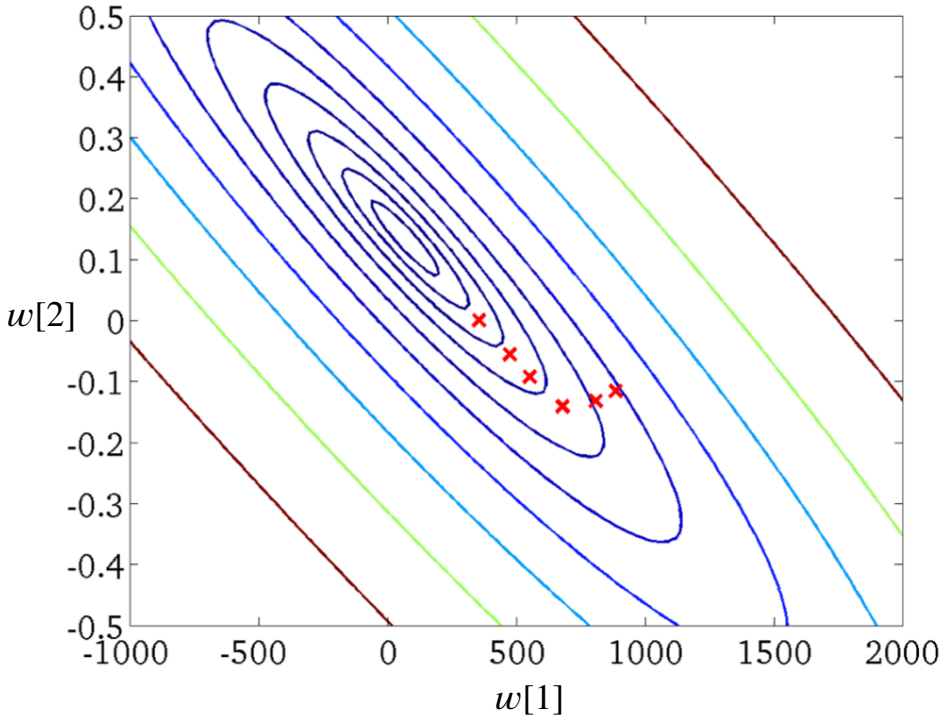



Gradient descent dynamics in the parameter space w
Ovals show the **level set** of the objective function

# 1-dimensional linear regression with 2 parameters

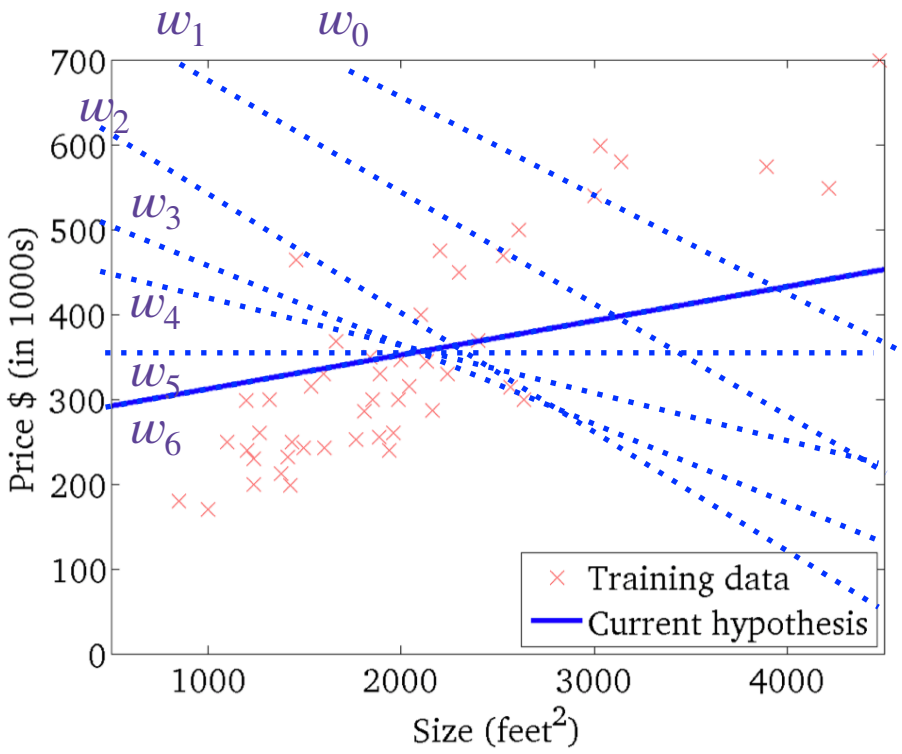



Evolution of the predictor y = w[0] + w[1]x

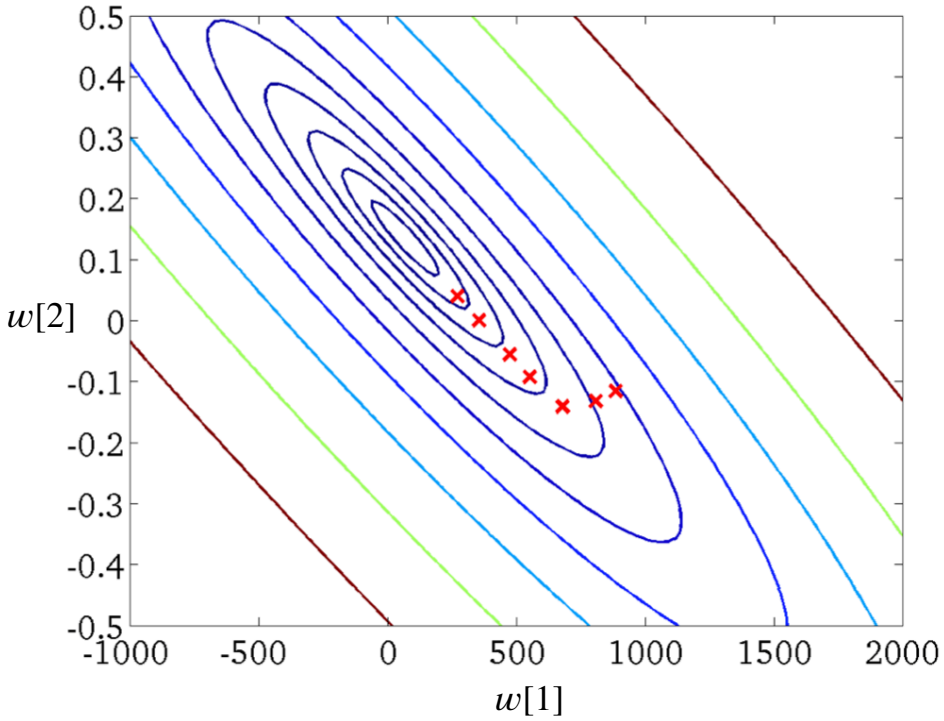



Gradient descent dynamics in the parameter space w
Ovals show the **level set** of the objective function

# 1-dimensional linear regression with 2 parameters

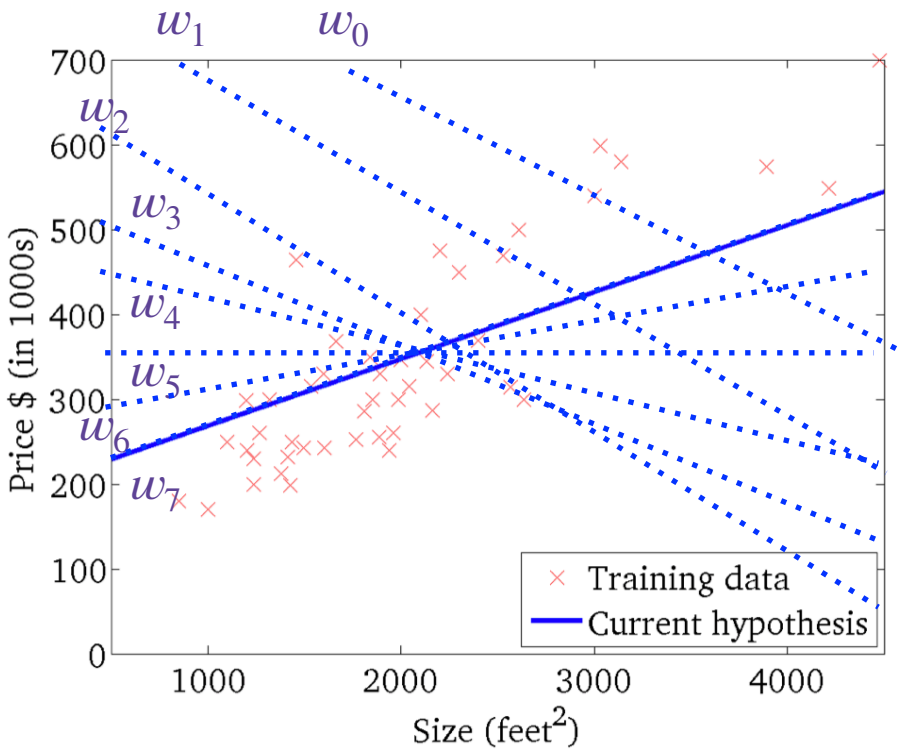



Evolution of the predictor y = w[0] + w[1]x

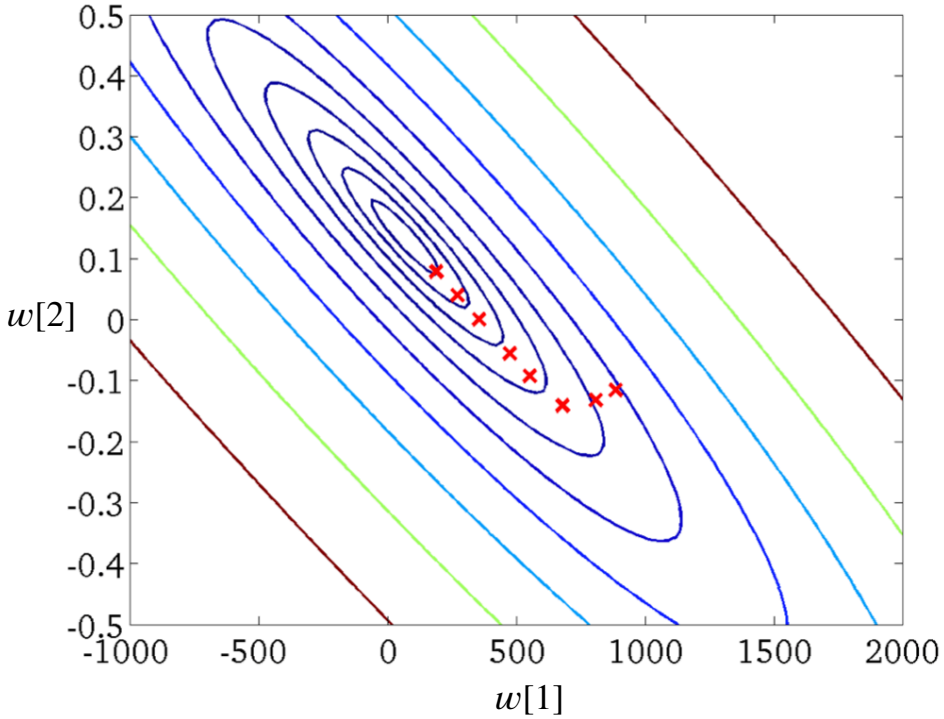



Gradient descent dynamics in the parameter space w
Ovals show the **level set** of the objective function

# 1-dimensional linear regression with 2 parameters

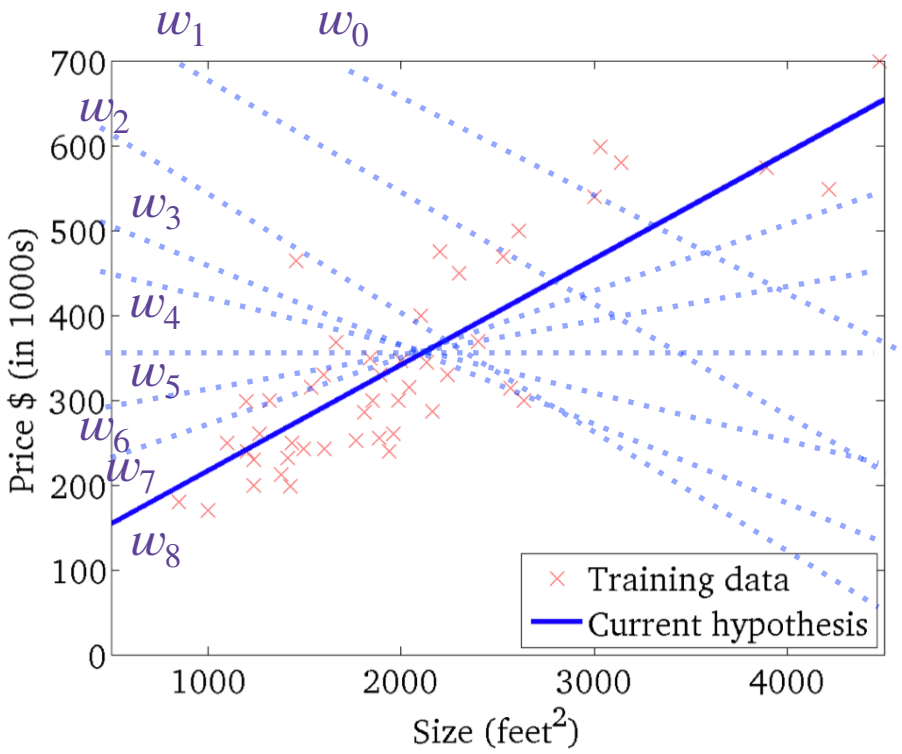



Evolution of the predictor y = w[0] + w[1]x

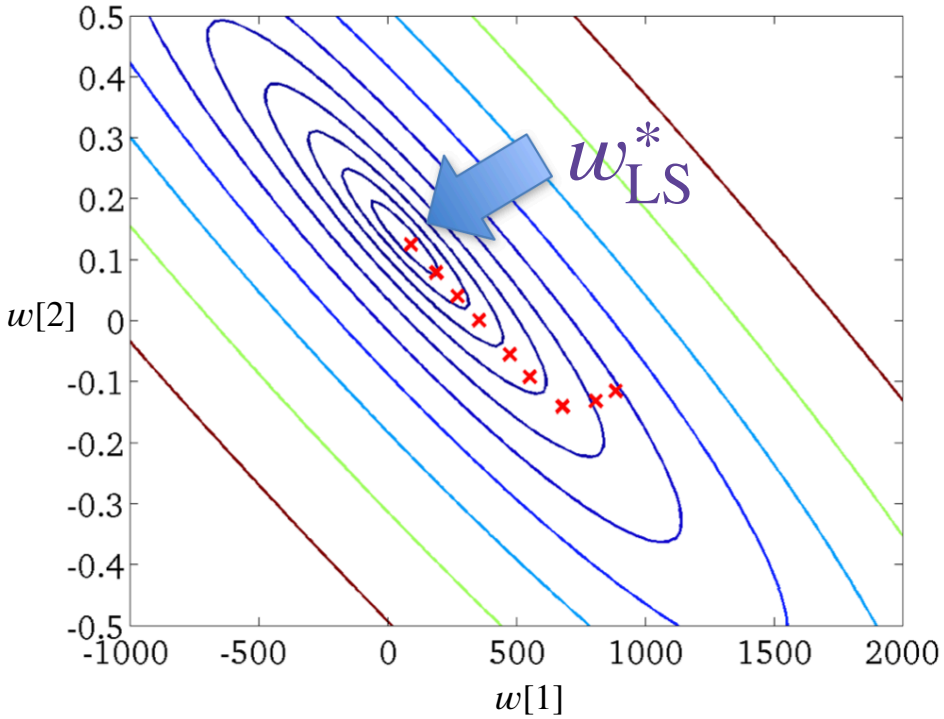



Gradient descent dynamics in the parameter space w
Ovals show the **level set** of the objective function

# 1-dimensional linear regression with 2 parameters




Evolution of the predictor y = w[0] + w[1]x




Gradient descent dynamics in the parameter space w
Ovals show the **level set** of the objective function

# Warmup: Quadratic functions

$$\hat{w} = \underset{w}{\text{argmin}} \; aw^2 + bw + c$$ # f(w) or loss function

$$w_0 \sim \mathcal{N}(0, I\sigma^2)$$

$$\left.\frac{df(w)}{dw}\right|_{w=w_0} = 2aw_0 + b$$

$$w_1 = w_0 - \eta(2aw_0 + b)$$

# Recall gradient descent eq:

$$w_{t+1} = w_t - \eta \left.\frac{df(w)}{dw}\right|_{w=w_t}$$

# Example: Linear regression

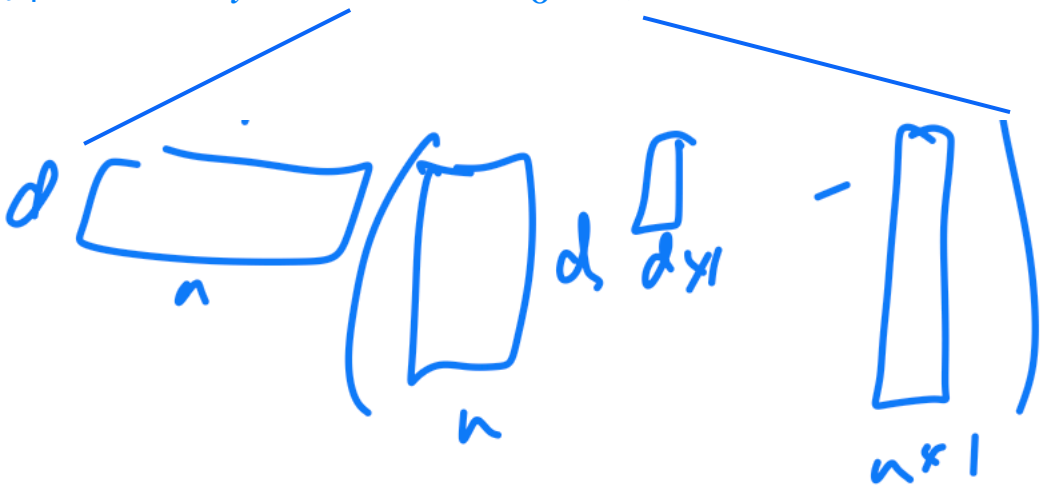
# Recall gradient descent eq:

$$w_{t+1} = w_t - \eta \frac{df(w)}{dw}\Big|_{w=w_t}$$

$$\hat{w} = \underset{w}{\operatorname{argmin}} \frac{1}{2} \sum_i \left(y_i - x_i^\top w\right)^2$$ # f(w) or loss function

$$\nabla_w f(w_0) = X^T(Xw - y)$$ # from previous lectures

$$w_{t+1} = w_t - \eta X^T(Xw_0 - y)$$ # how can we check if this is right?

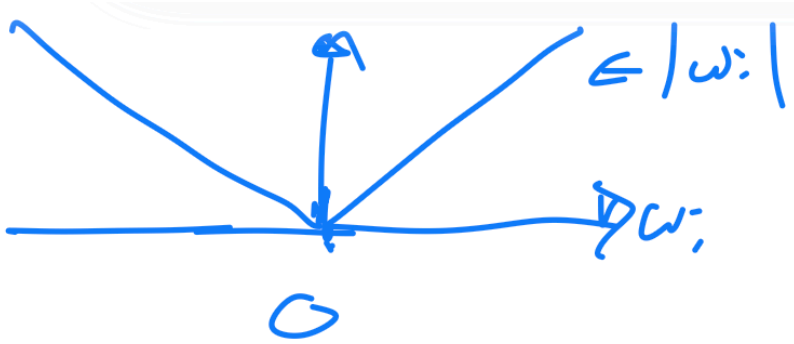


# Example: Lasso

$$\hat{w} = \underset{w}{\operatorname{argmin}} \frac{1}{2}\|y - Xw\|_2^2 + \lambda\|w\|_1$$

$$\nabla_w f = X^T(Xw - y) + \lambda \operatorname{sign}(w)$$

$$w_{t+1} \leftarrow w_t - \eta \cdot \nabla_w f_{w=w_t}$$




# LASSO regularizer is convex. So?

# Local minima = global minimum

$$\frac{d|w_i|}{dw_i} = \begin{cases} +1 & w_i > 0 \\ [-1, 1] & w_i = 0 \\ -1 & w_i < 0 \end{cases}$$

# Gradient of absolute value is undefined at w=0, so define a sub gradient

# Example: Lasso # 2

Why can't I analytically find the minimum if I can define the gradient of |w|?

$$\frac{d|w_i|}{dw_i} = \begin{cases} +1 & w_i > 0 \\ [-1, 1] & w_i = 0 \\ -1 & w_i < 0 \end{cases}$$

$$f(w) = aw^2 + bw + c + |w|$$

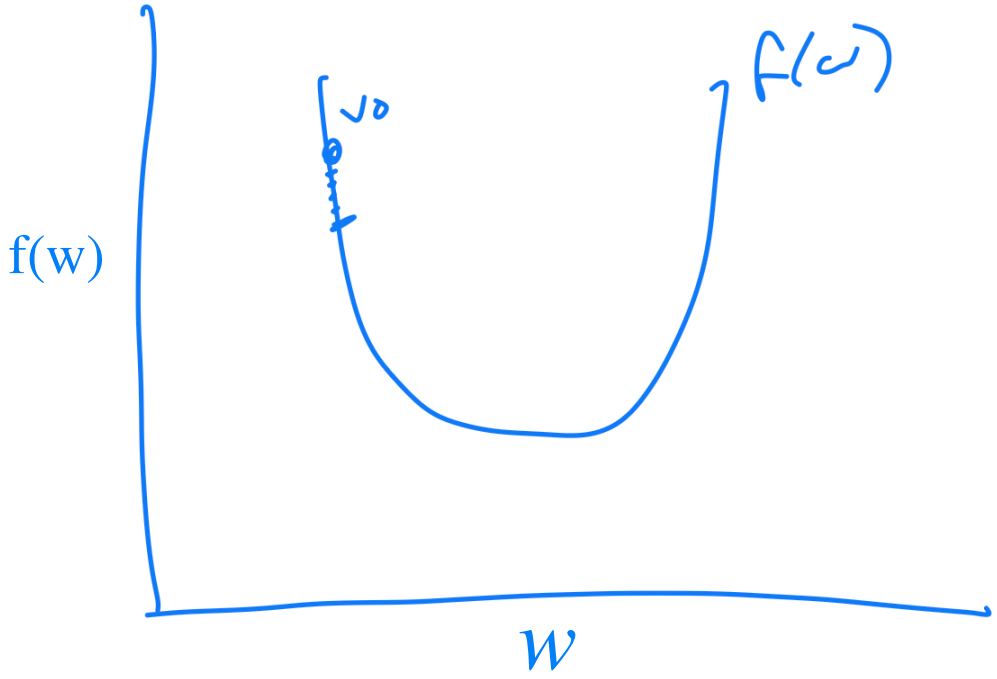
$$\frac{df}{dw} = 2aw + b + \text{sign}(w) = 0$$

# Can find if w is 1d by checking both +1 and -1 cases

But if $w \in \mathbb{R}^d$ wow many possible values for sign(w)? $2^d$

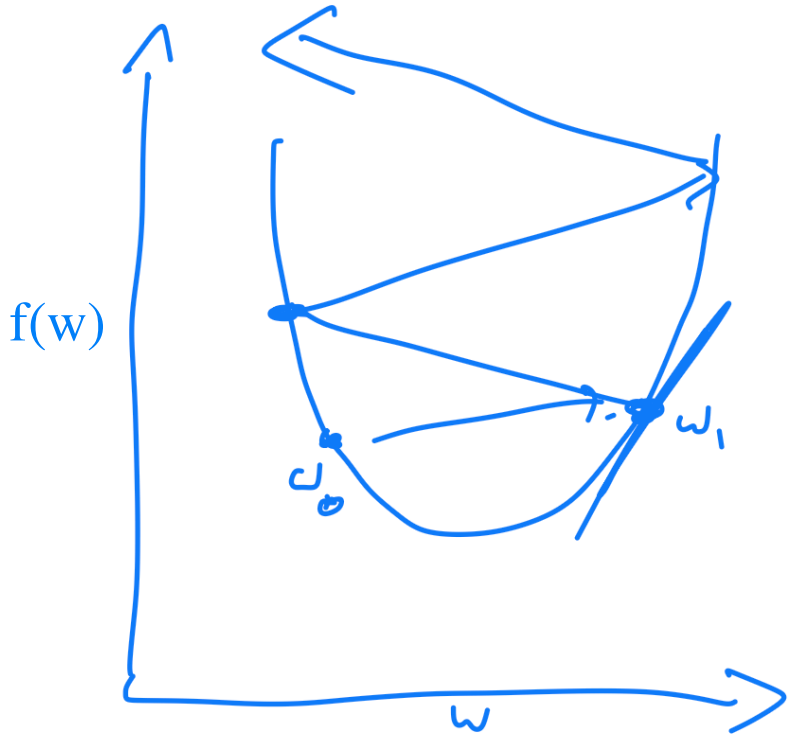
# How do you choose a step size?

What can go wrong if η is too small?




Slow convergence

What can go wrong if η is too large?




Divergence!

# How do you choose a step size?

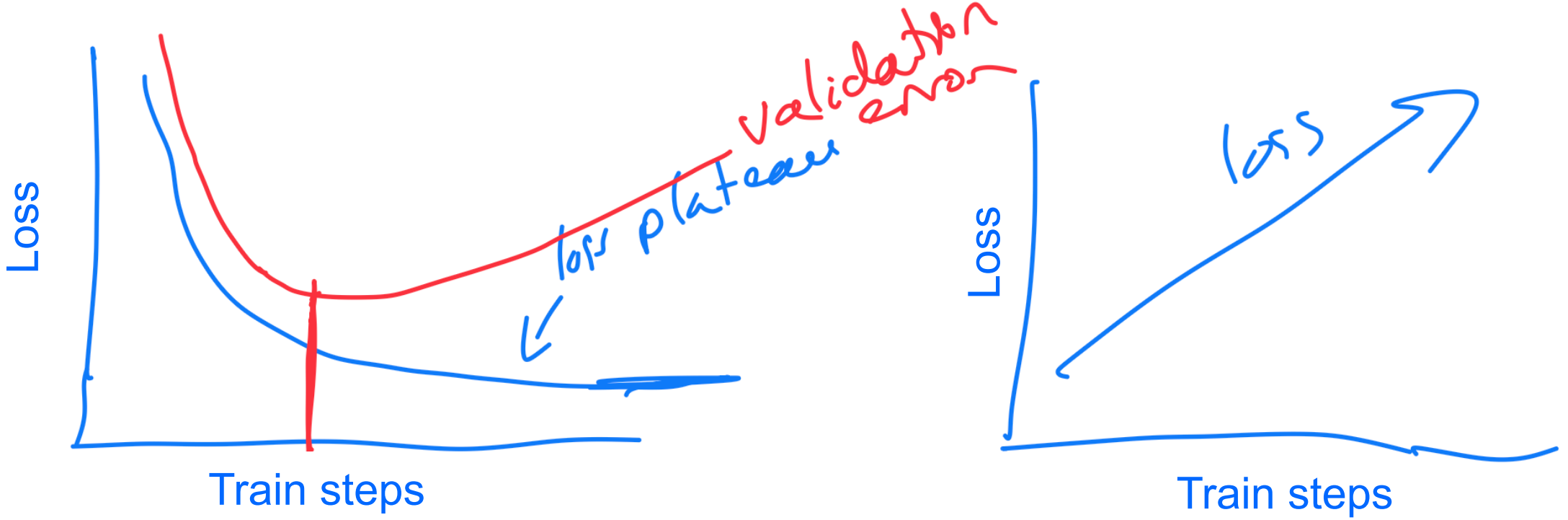
- In practice: guess and check

# MAKE PLOTS

# How do you choose a step size?

- In practice: guess and check

# MAKE PLOTS




# If I pick a learning rate η and observe the plot on the left, did I pick wrong?

# What about the plot on the right?