CSE 446
Gradient Descent

Natasha Jaques

UNIVERSITY of WASHINGTON

How do we find optimal weights?

- This is related 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, $\|w\|_{0.5}$?

Why gradient descent?

- Standard ML paradigm: Define loss, then optimize

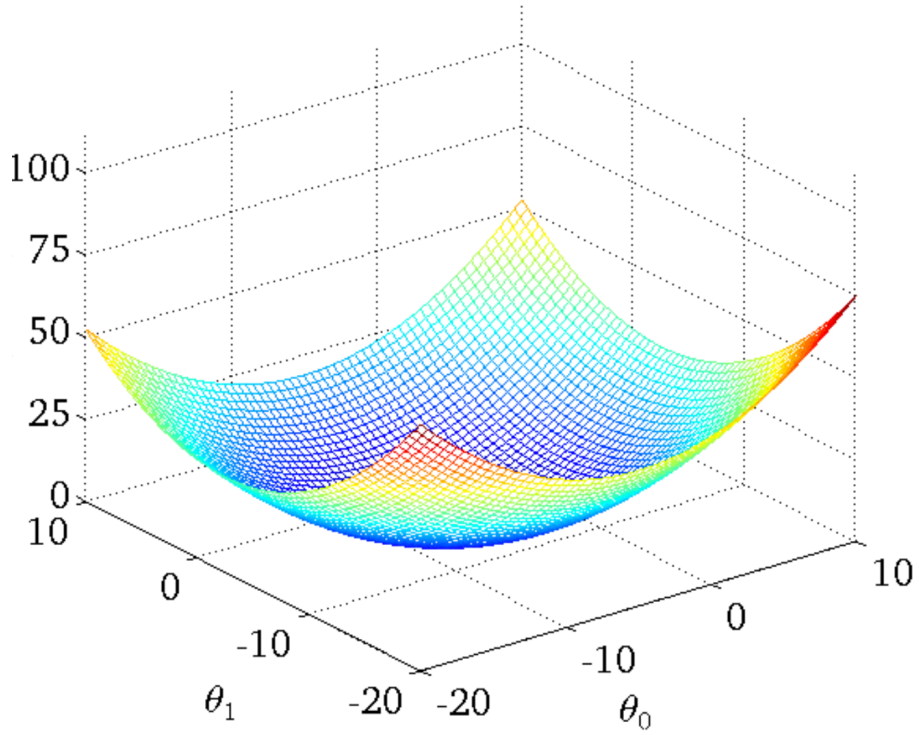
$$\widehat{w}_{LS} = \arg\min_{w} ||\mathbf{y} - \mathbf{X}w||_2^2$$

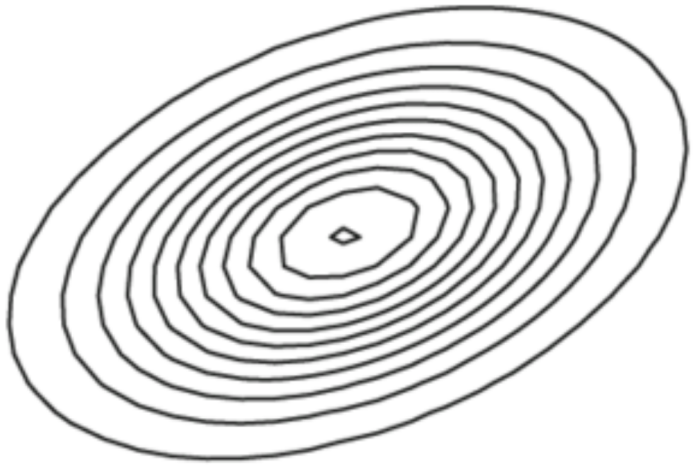
- But, no closed-form solutions for most losses we use in practice.
- Key idea: Iterative methods
- Used everywhere!

Gradient descent in one dimension

Step direction: –gradient
Step size: $\propto$ |gradient|

Gradient descent in multiple dimensions





Lecture plan

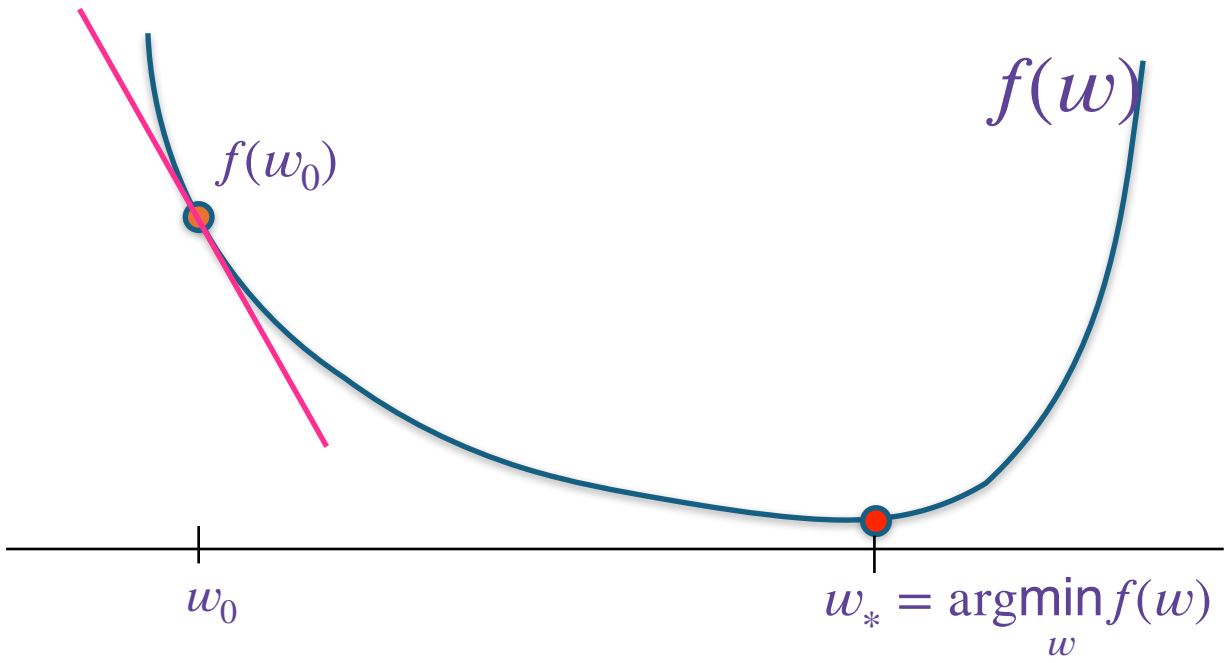
- Gradient descent algorithm + examples
- Theoretical analysis
  - When does it work?
  - How quickly does it converge?
  - How do we choose a step size?
  - Key idea: Convexity
- Not tested on proof details, but concepts are important & practical

Algorithm: Gradient descent

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 η

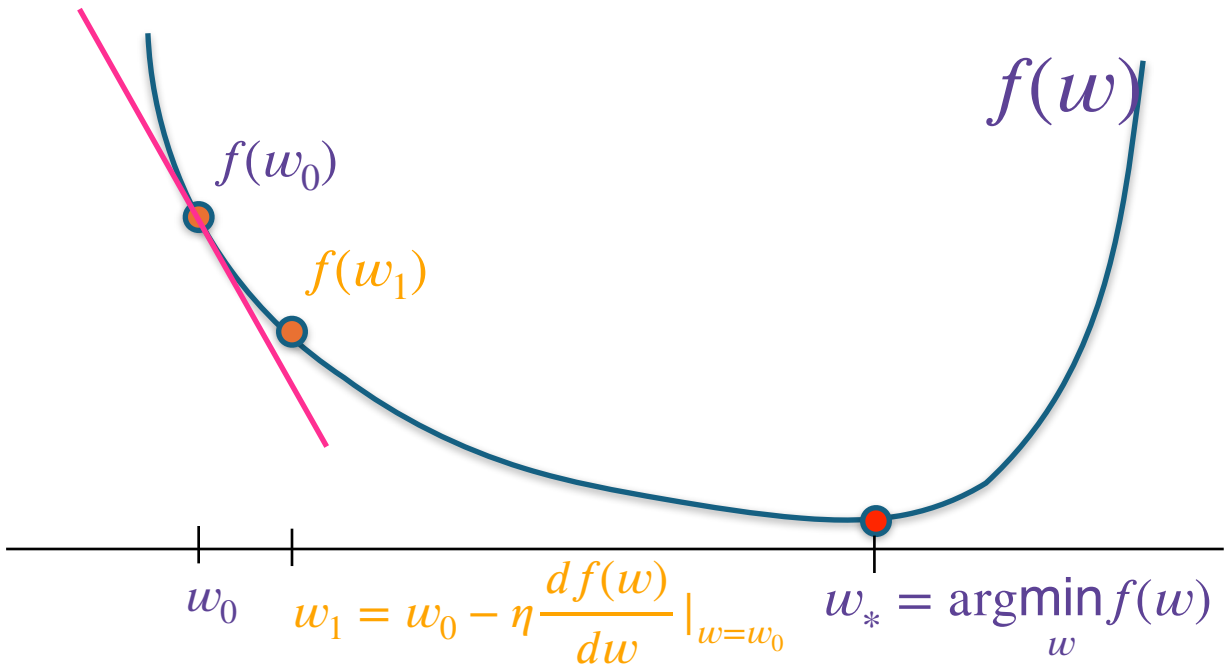


Algorithm: Gradient descent

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 η

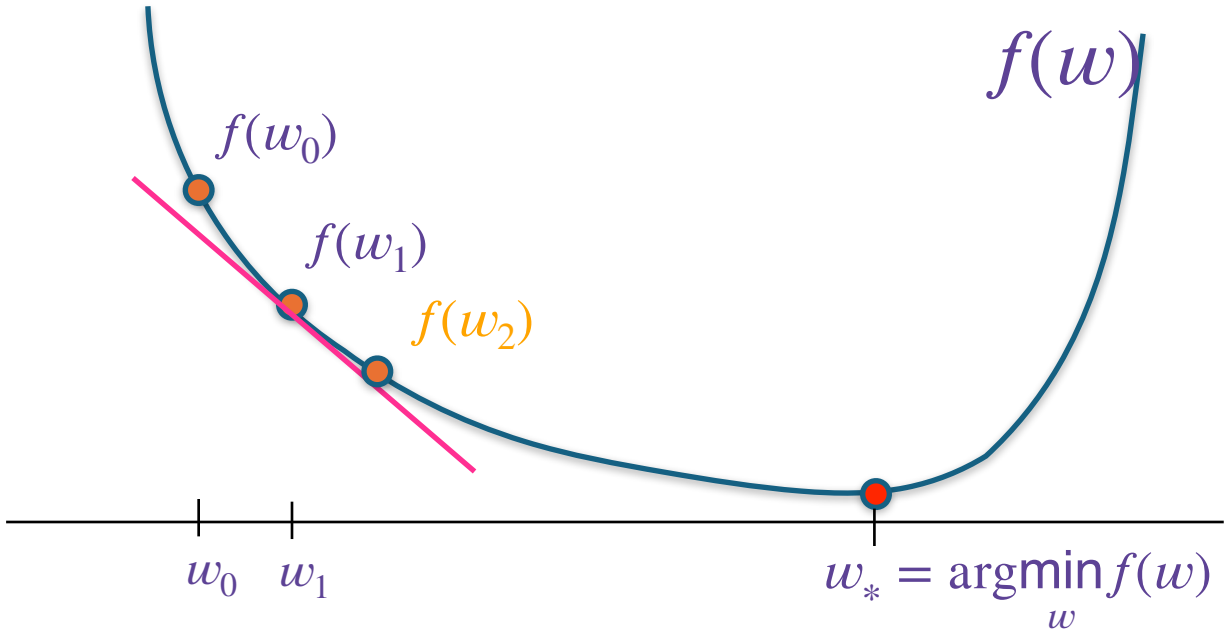


Algorithm: Gradient descent

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 η

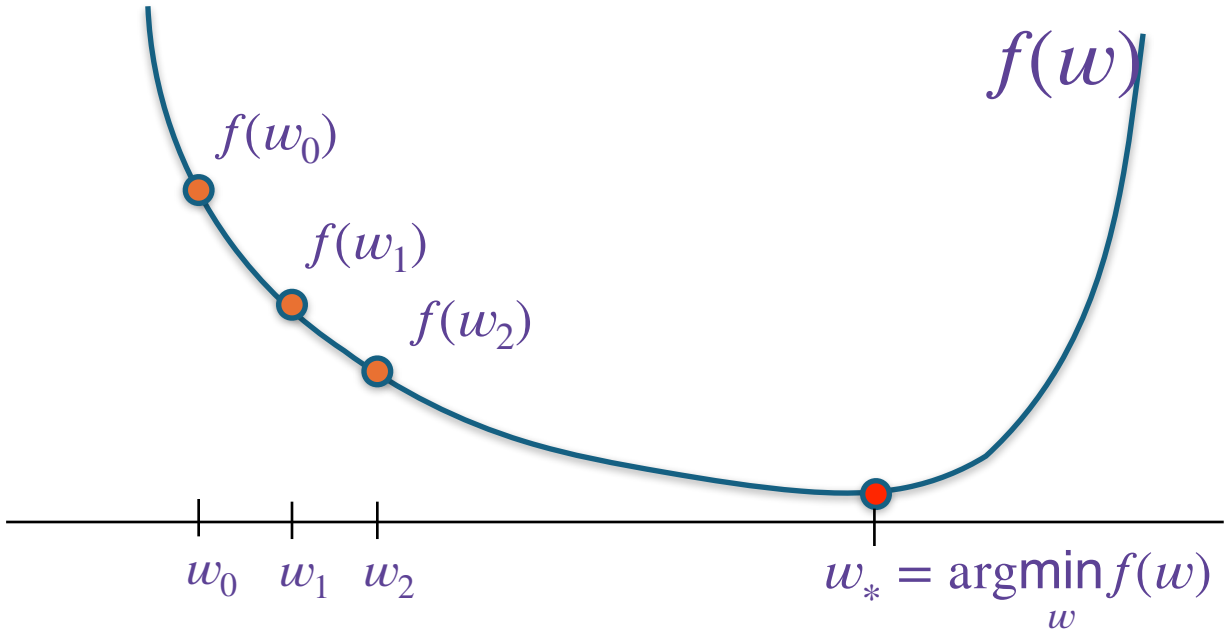


Algorithm: Gradient descent

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 η

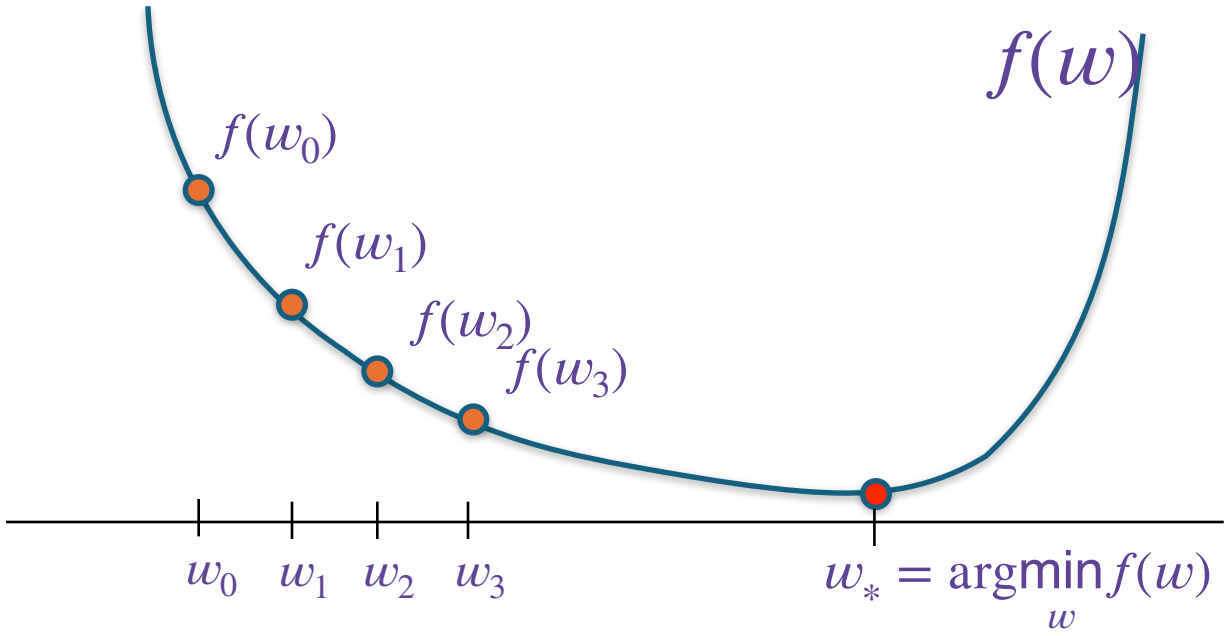


Algorithm: Gradient descent

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 η

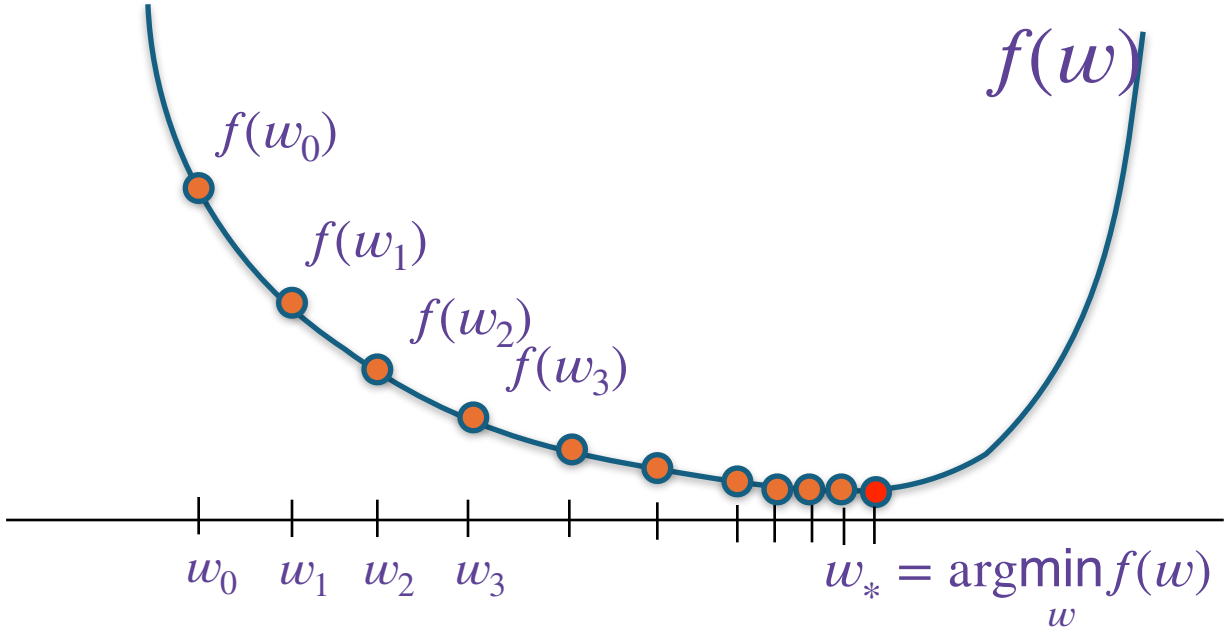


Algorithm: Gradient descent

Algorithm
For t=0,1,2,3, …

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

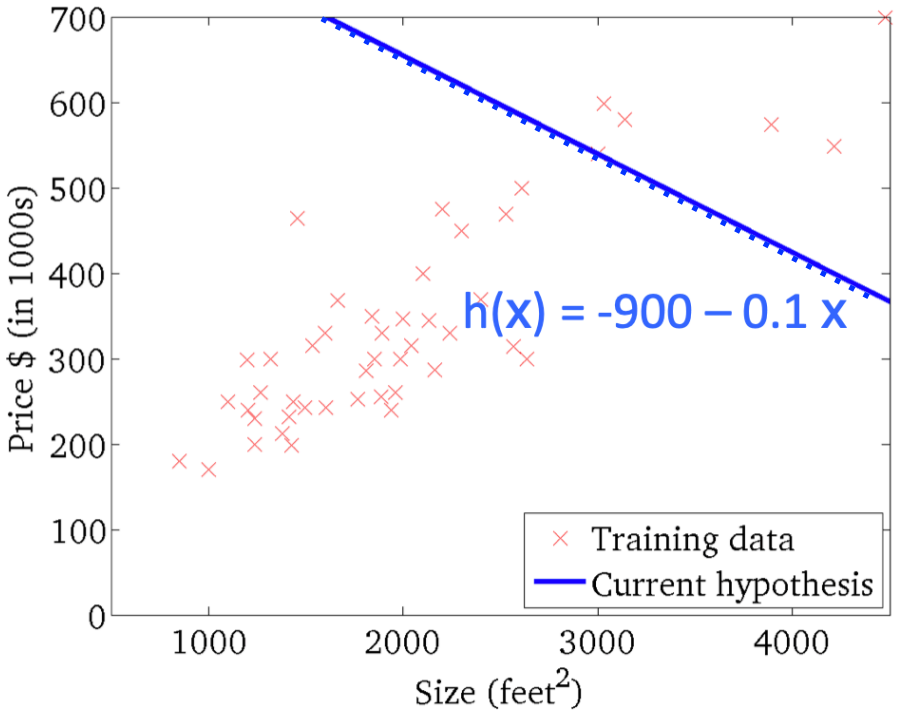
Hyperparameters:
- Initial point w_0
- Step size η



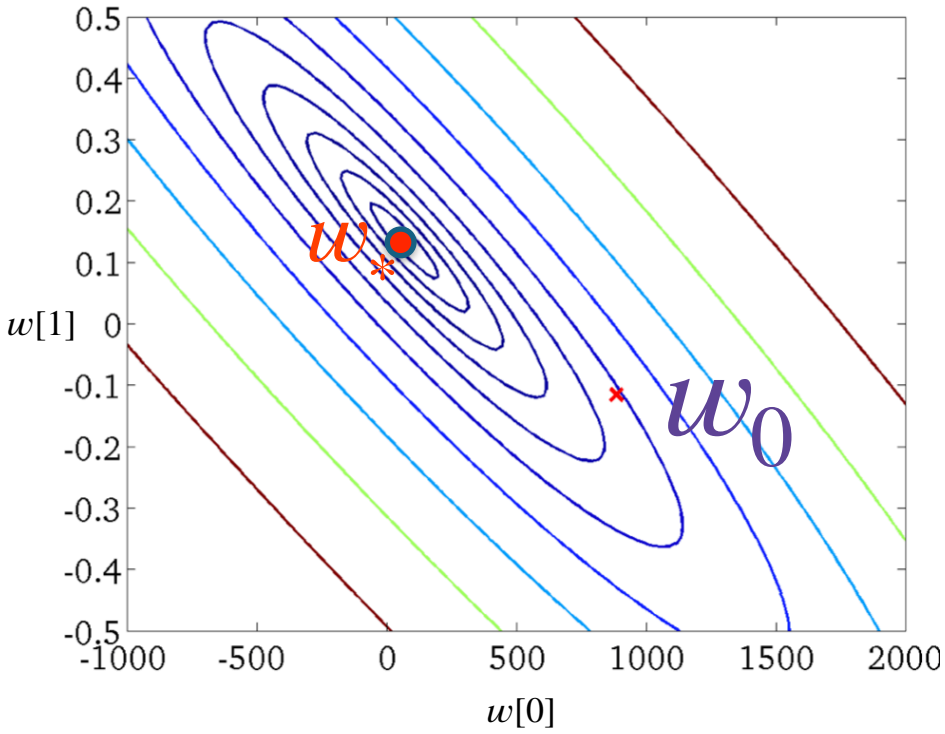
Note that as $t \to \infty$ we have $\frac{df(w)}{dw}\Big|_{w=w_t} \to 0$

1-dimensional linear regression with 2 parameters

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



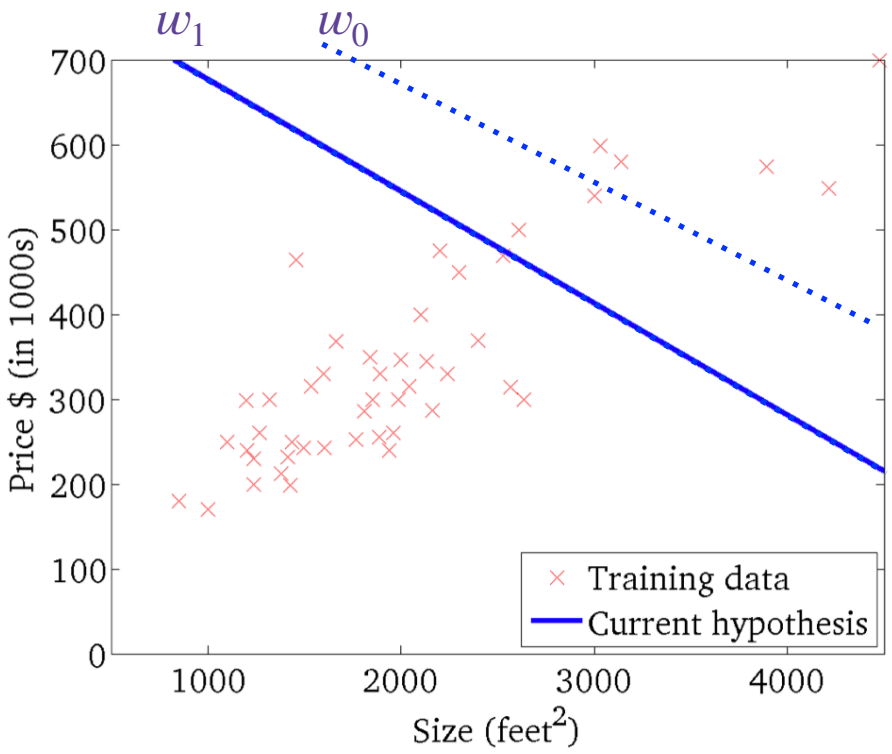
$$w_{t+1} \leftarrow w_t - \eta \cdot \nabla_w f(w_t)$$



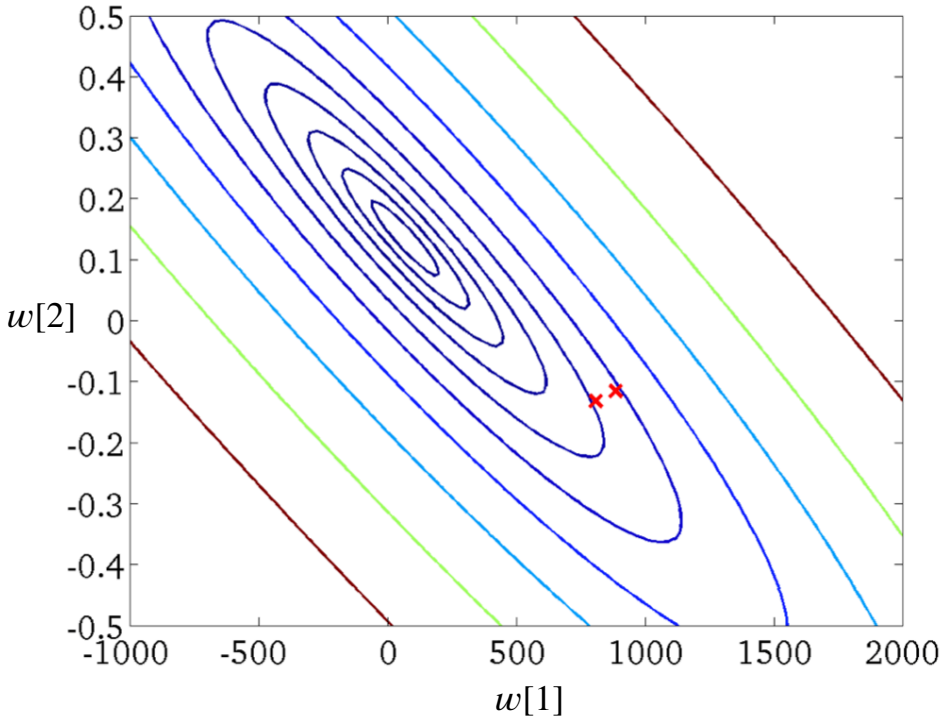
Evolution of the predictor y = w[0] + w[1]x

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

1-dimensional linear regression with 2 parameters

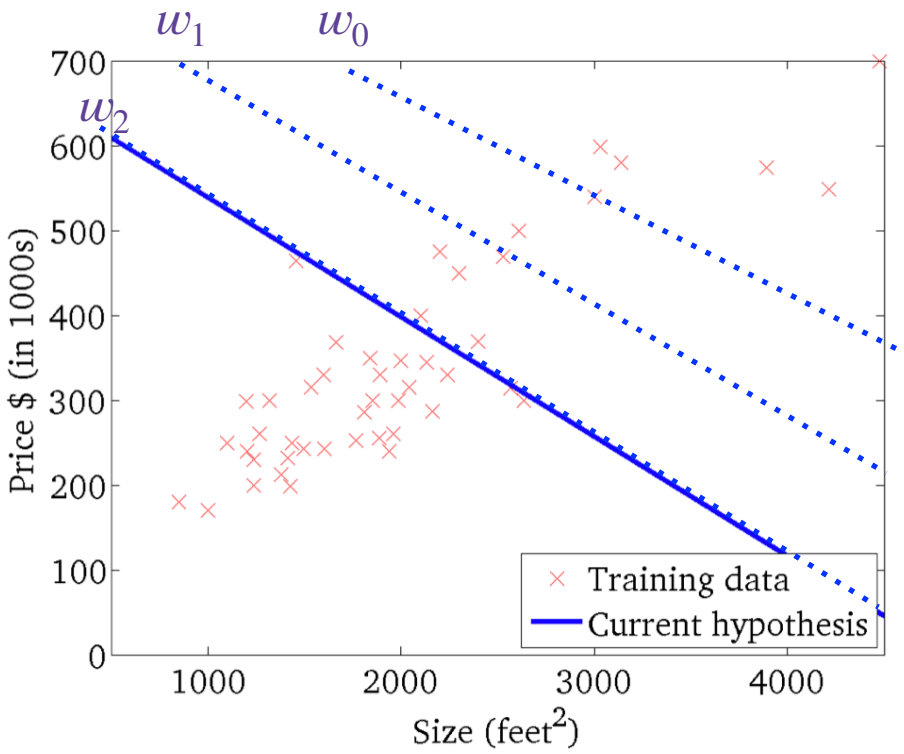


Evolution of the predictor y = w[0] + w[1]x

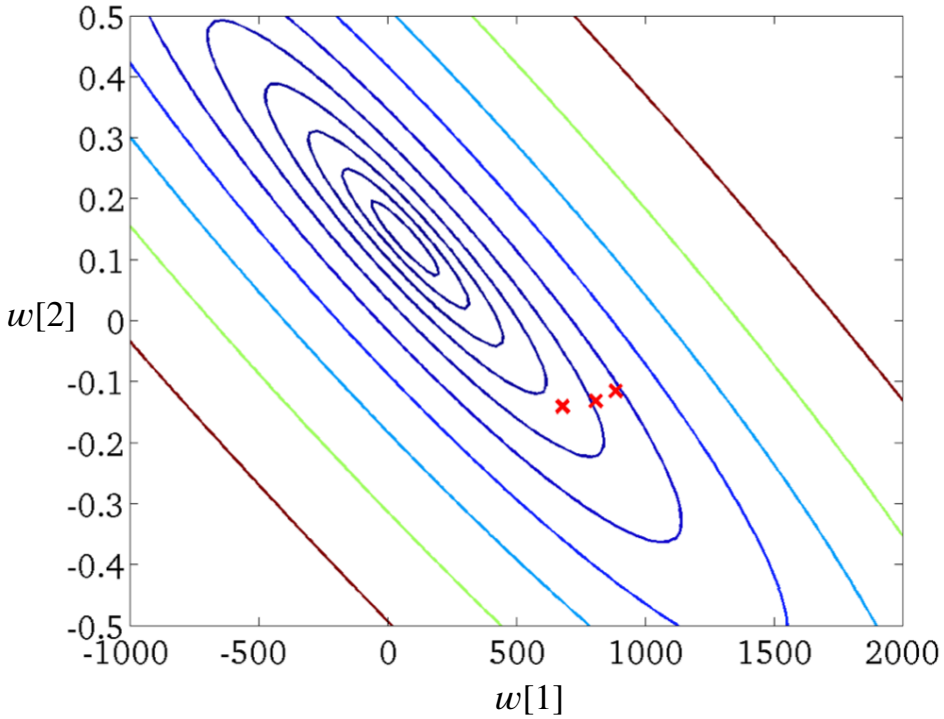


Gradient descent dynamics in the parameter space w
Ovals show the **level set** of the objective function

1-dimensional linear regression with 2 parameters

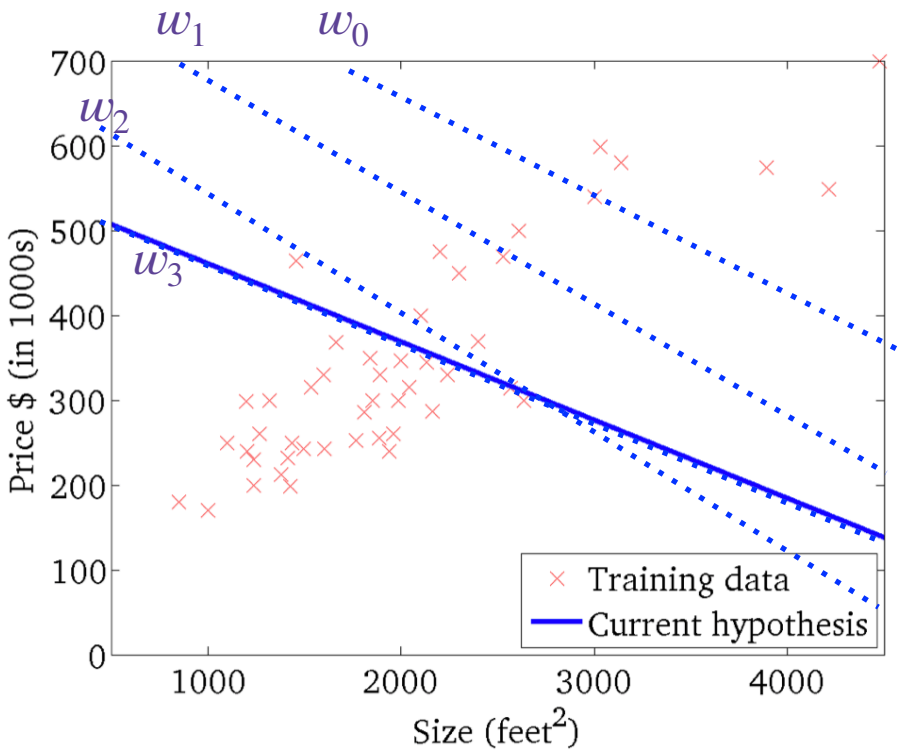


Evolution of the predictor y = w[0] + w[1]x

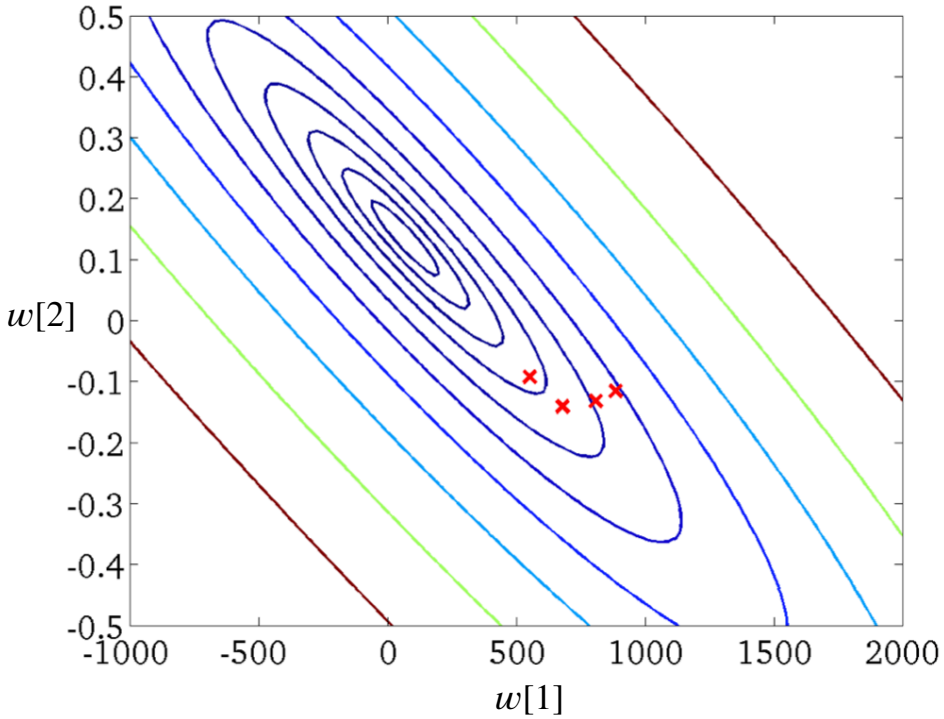


Gradient descent dynamics in the parameter space w
Ovals show the **level set** of the objective function

1-dimensional linear regression with 2 parameters

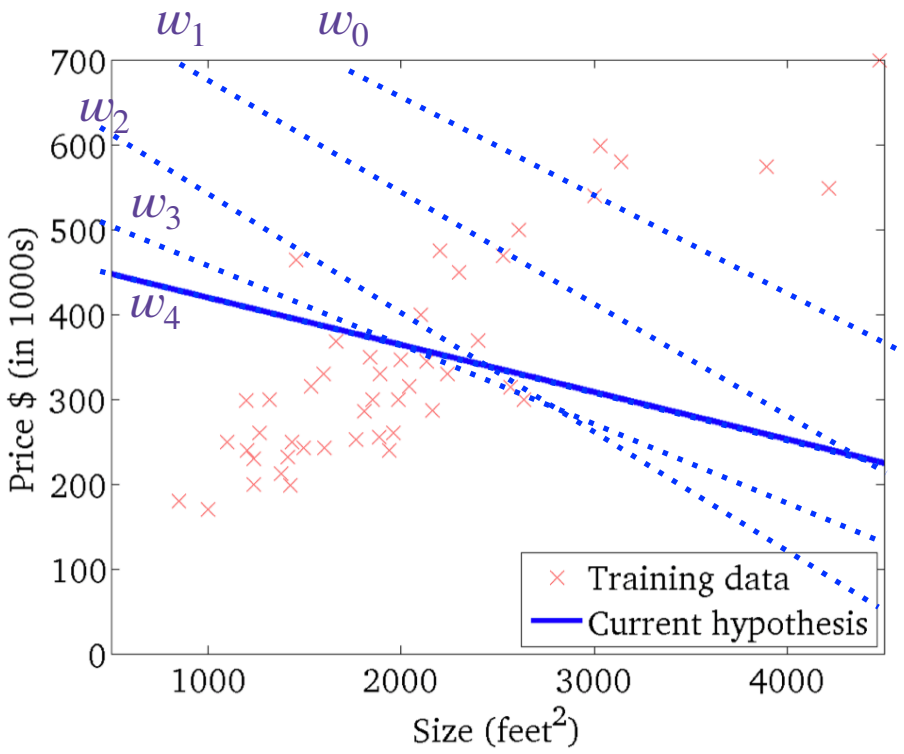


Evolution of the predictor y = w[0] + w[1]x

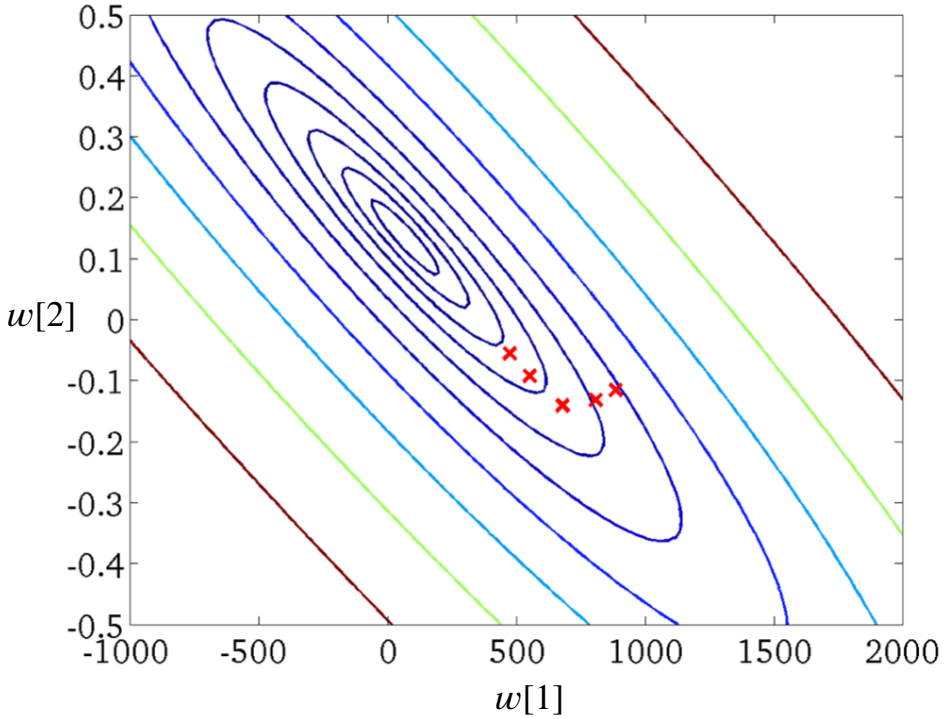


Gradient descent dynamics in the parameter space w
Ovals show the **level set** of the objective function

1-dimensional linear regression with 2 parameters

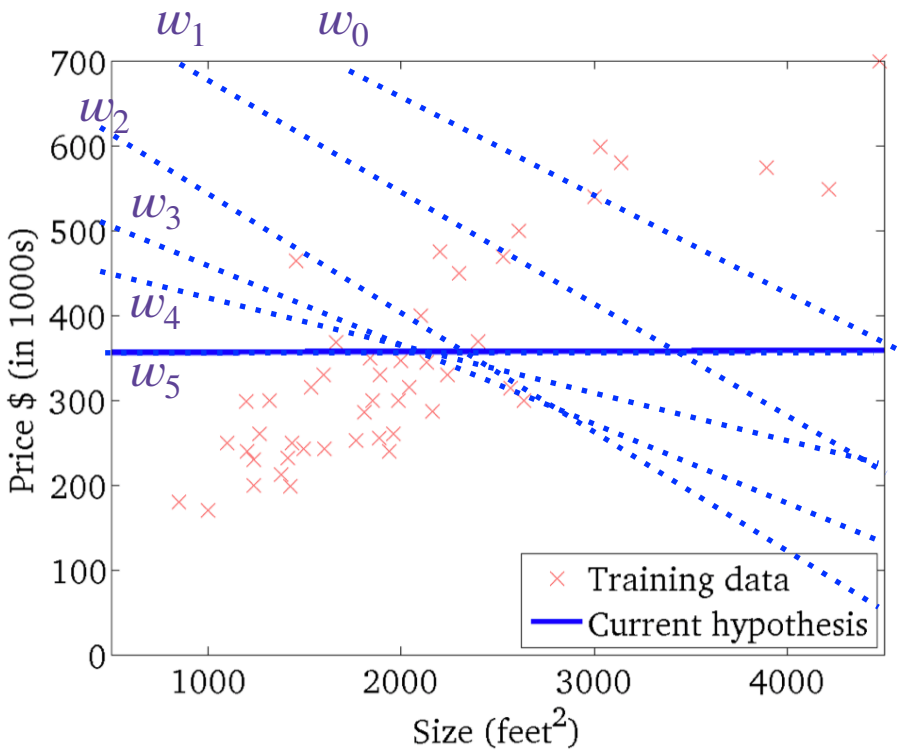


Evolution of the predictor y = w[0] + w[1]x

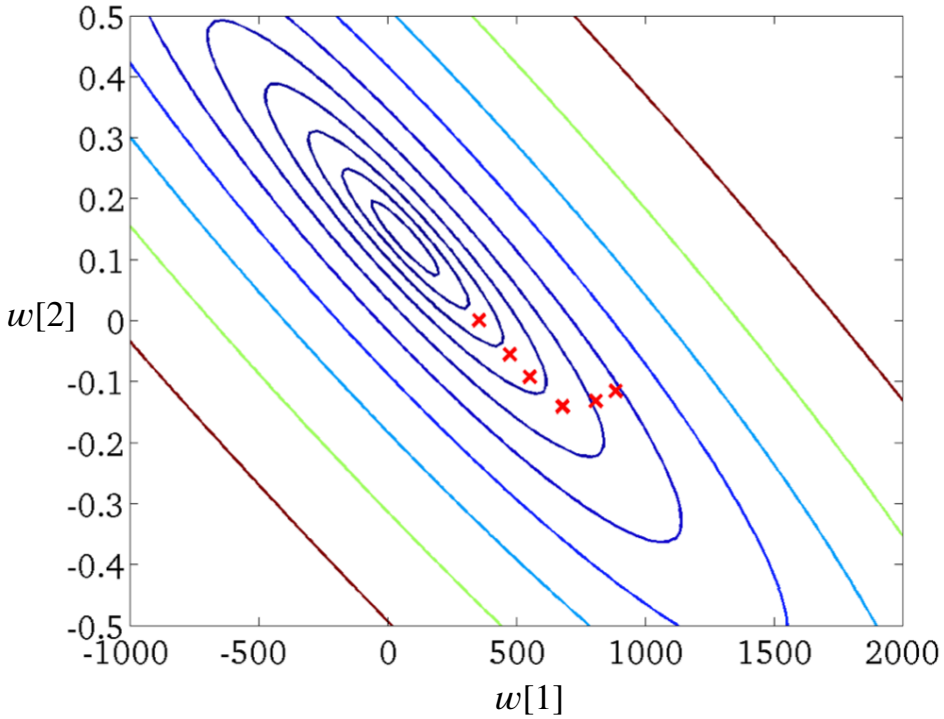


Gradient descent dynamics in the parameter space w
Ovals show the **level set** of the objective function

1-dimensional linear regression with 2 parameters

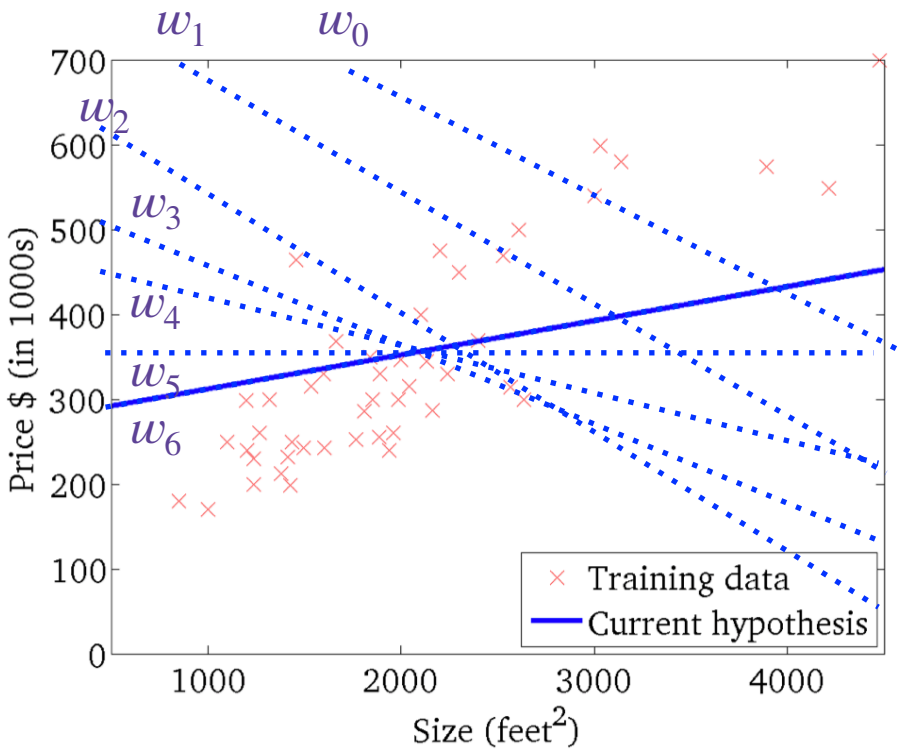


Evolution of the predictor y = w[0] + w[1]x

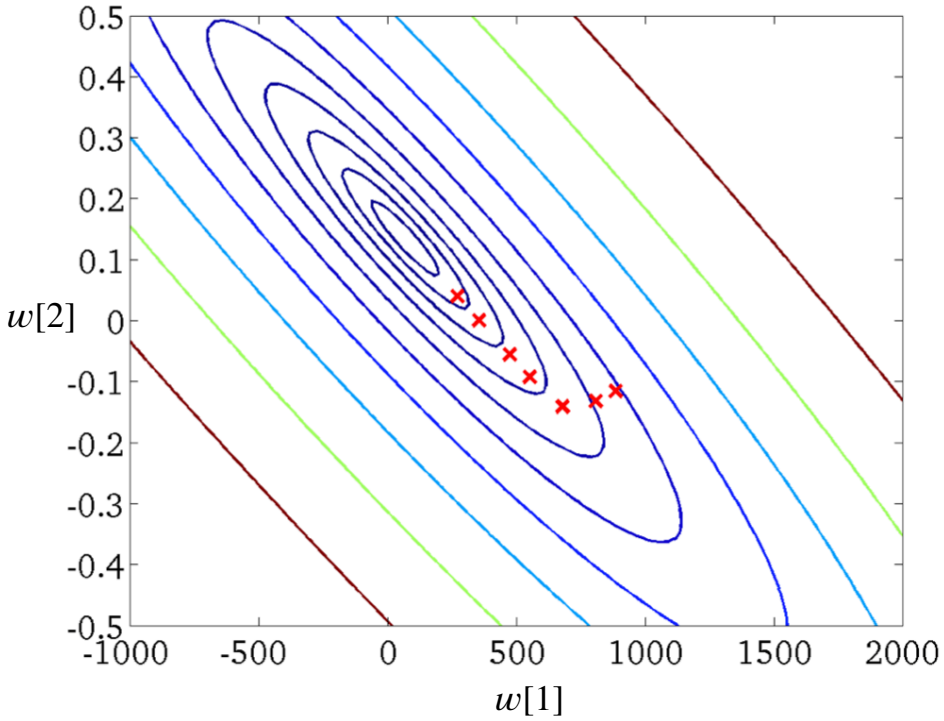


Gradient descent dynamics in the parameter space w
Ovals show the **level set** of the objective function

1-dimensional linear regression with 2 parameters

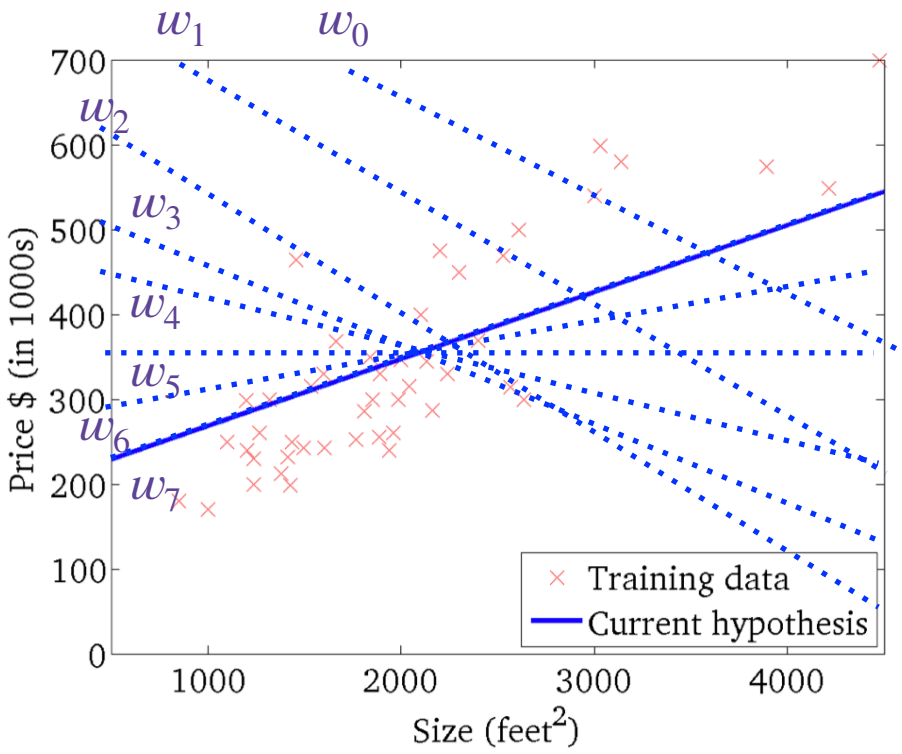


Evolution of the predictor y = w[0] + w[1]x

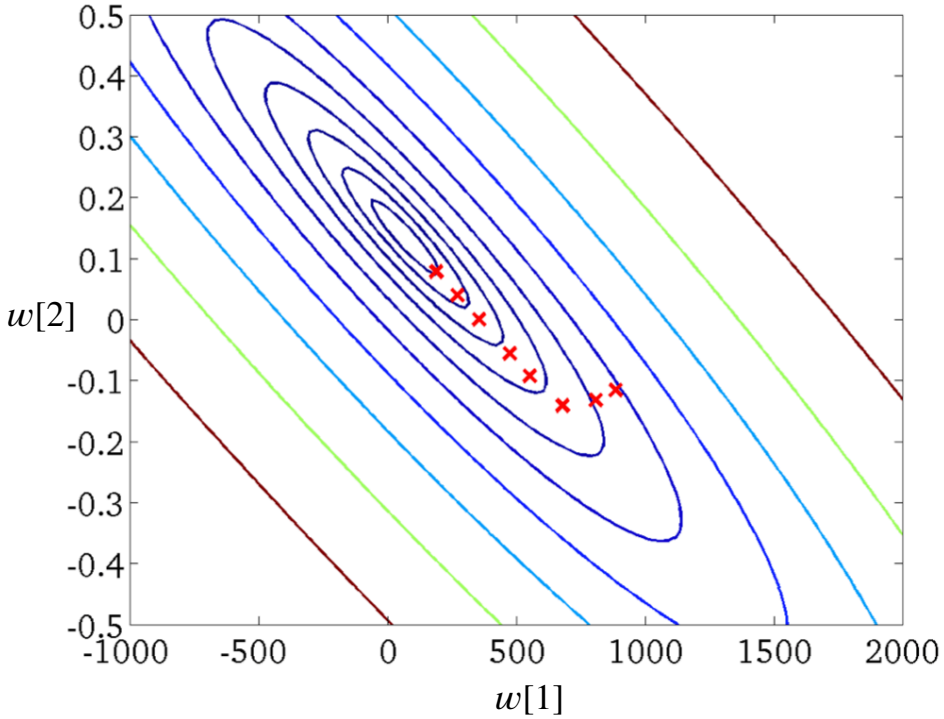


Gradient descent dynamics in the parameter space w
Ovals show the **level set** of the objective function

1-dimensional linear regression with 2 parameters

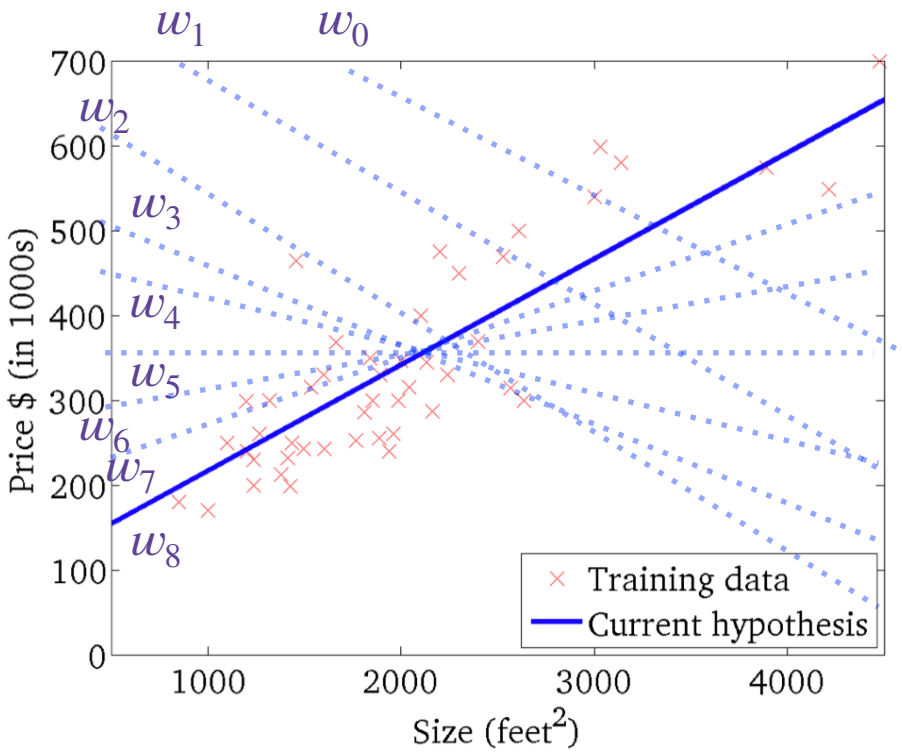


Evolution of the predictor y = w[0] + w[1]x

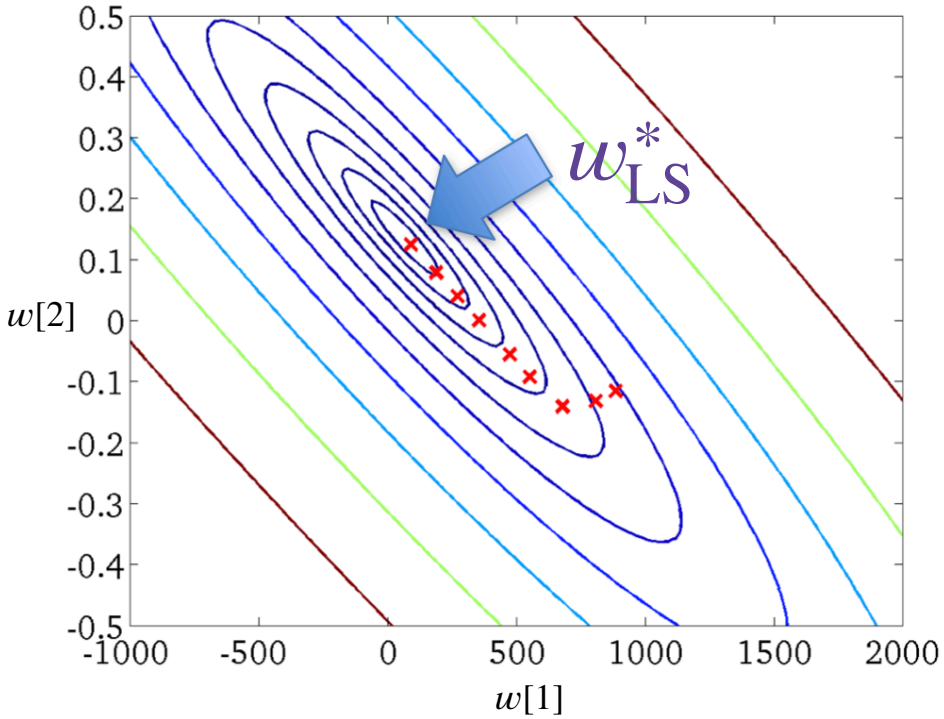


Gradient descent dynamics in the parameter space w
Ovals show the **level set** of the objective function

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

$$\hat{w} = \underset{w}{\text{argmin}} \; aw^2 + bw + c$$

Example: Linear regression

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

Example: Lasso

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

How do you choose a step size?

How do you choose a step size?

- If η too small, converges very, very slowly.
- If η too big, does not converge!
- In practice: guess and check