Gradient Descent

- how are we going to find the solution for

$$\arg\min_{b,w} \sum_{i=1}^{n} \ell(b + w^T x_i, y_i)$$

- e.g., Lasso, Logistic Regression do not have closed form solution for

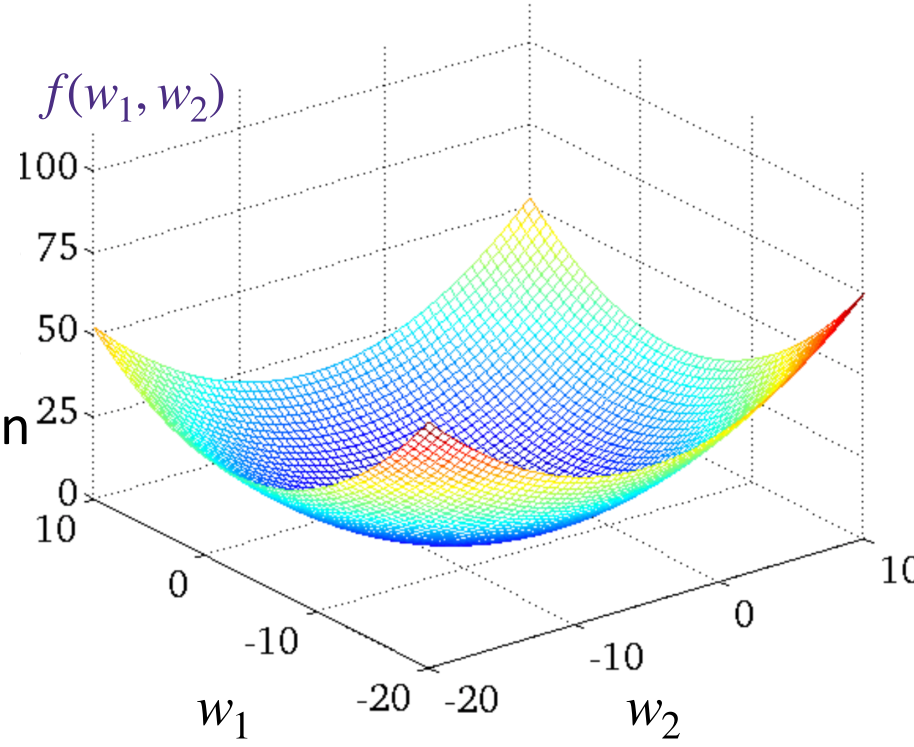
$$\nabla_{b,w} \mathscr{L}(b, w) = 0$$

Running example: linear regression

- **Given data:** $\{(x_i, y_i)\}_{i=1}^n \qquad x_i \in \mathbb{R}^d \qquad y_i \in \mathbb{R}$

- **Learning model parameters:**

$$\hat{w}_{LS} = \arg\min_{w \in \mathbb{R}^d} \underbrace{\|\mathbf{y} - \mathbf{X}w\|_2^2}_{f(w)}$$



- Although we know the optimal solution in a closed form, we will use this as a running example to understand GD

1-dimensional gradient descent

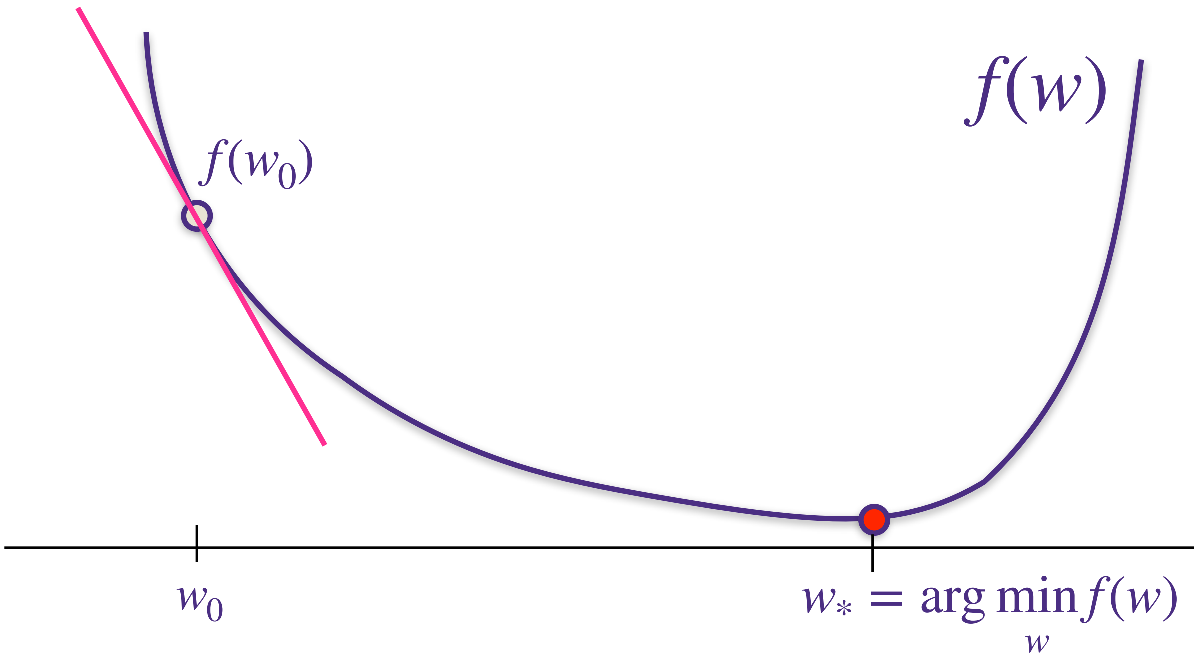
Let w_0 be an initial guess. How can we improve this solution?

Taylor series approximation:

For w very close to w_0 we have

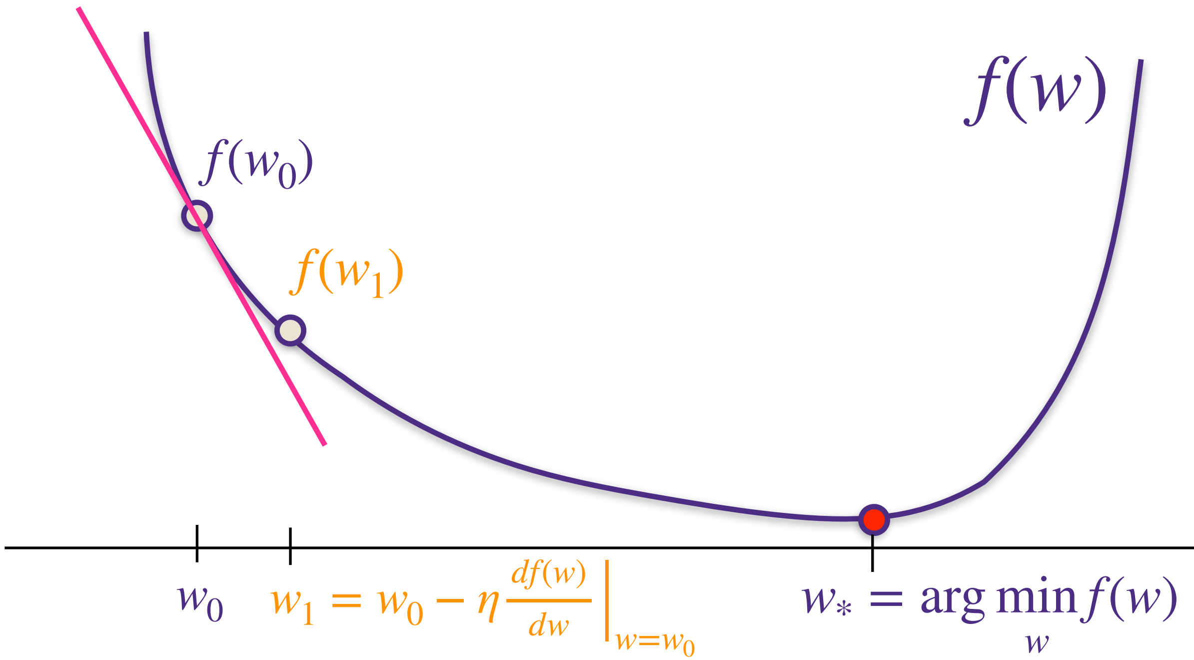
$$f(w_0) + (w - w_0)\frac{df(w)}{dw}\Big|_{w=w_0}$$

is very close to $f(w)$



1-dimensional gradient descent

Let w_0 be an initial guess. How can we improve this solution?



Taylor series approximation:
For w very close to w_0 we have

$$f(w_0) + (w - w_0)\frac{df(w)}{dw}\Big|_{w=w_0}$$

is very close to $f(w)$

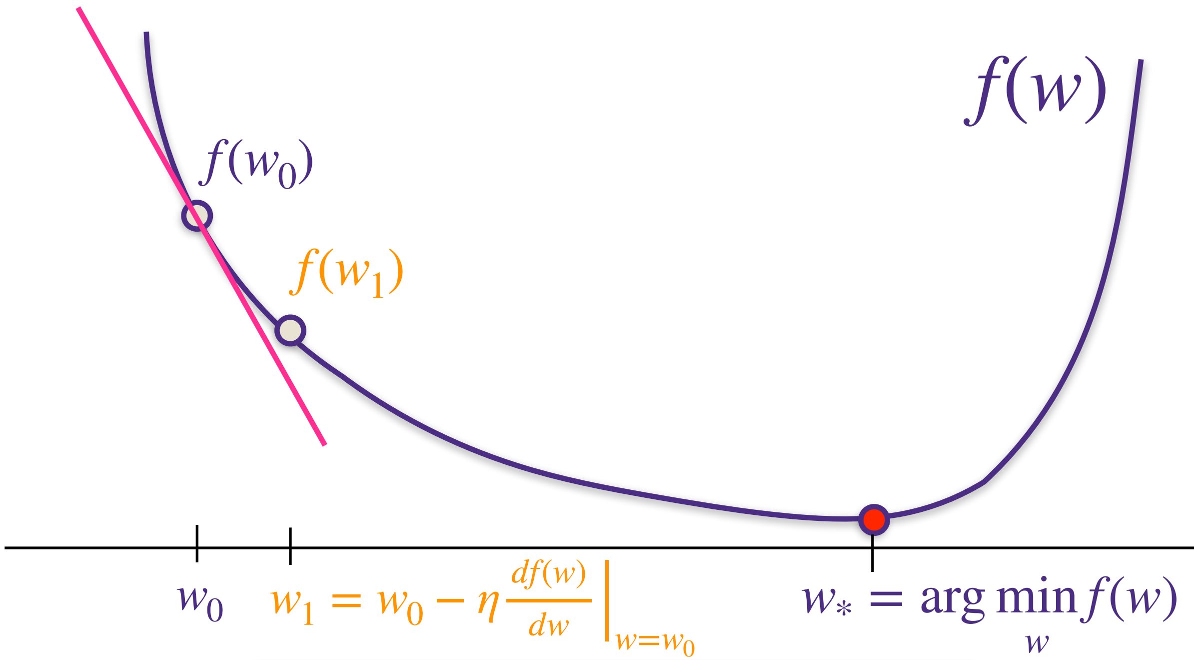
Thus, for very small $\eta > 0$,
if $w_1 = w_0 - \eta \frac{df(w)}{dw}\Big|_{w=w_0}$ then

$$f(w_0) - \eta\left(\frac{df(w)}{dw}\Big|_{w=w_0}\right)^2$$

is very close to $f(w_1) < f(w_0)$

1-dimensional gradient descent

Let w_0 be an initial guess. How can we improve this solution?



Taylor series approximation:
For w very close to w_0 we have

$$f(w_0) + (w - w_0)\frac{df(w)}{dw}\Big|_{w=w_0}$$

is very close to $f(w)$

Thus, for very small $\eta > 0$,

if $w_1 = w_0 - \eta \frac{df(w)}{dw}\Big|_{w=w_0}$ then

$$f(w_0) - \eta\left(\frac{df(w)}{dw}\Big|_{w=w_0}\right)^2$$

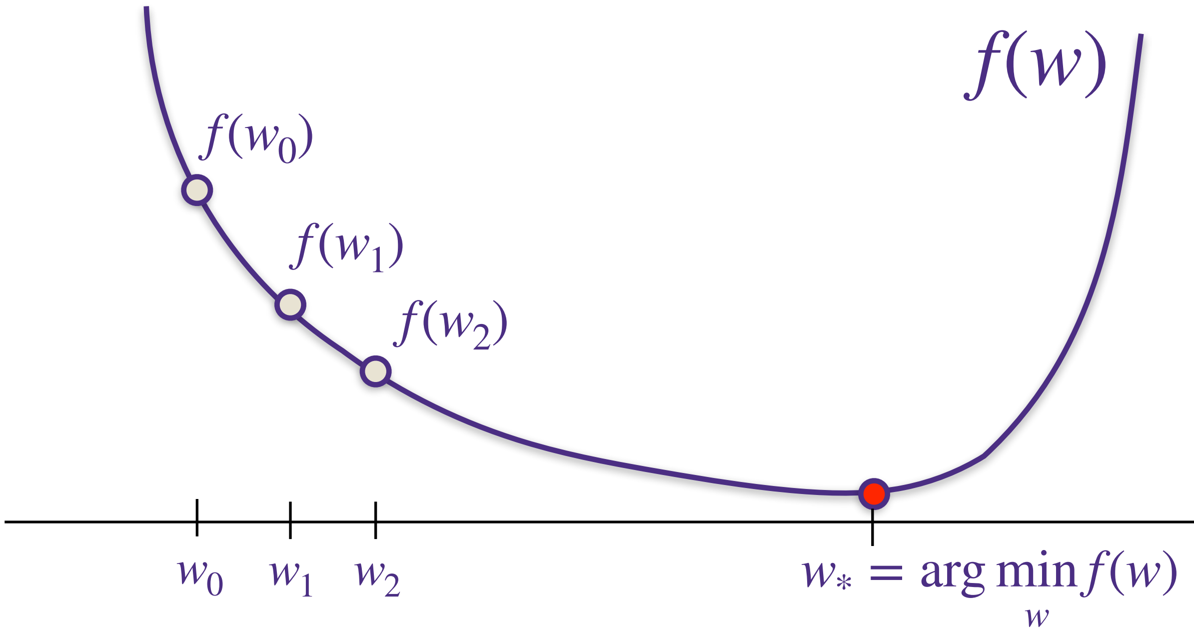
is very close to $f(w_1) < f(w_0)$

Gradient descent
For k=0,1,2,3,…

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

1-dimensional gradient descent

Let w_0 be an initial guess. How can we improve this solution?



Taylor series approximation:

For w very close to w_0 we have

$$f(w_0) + (w - w_0)\frac{df(w)}{dw}\Big|_{w=w_0}$$

is very close to $f(w)$

Thus, for very small $\eta > 0$,

if $w_1 = w_0 - \eta \frac{df(w)}{dw}\Big|_{w=w_0}$ then

$$f(w_0) - \eta\left(\frac{df(w)}{dw}\Big|_{w=w_0}\right)^2$$

is very close to $f(w_1) < f(w_0)$

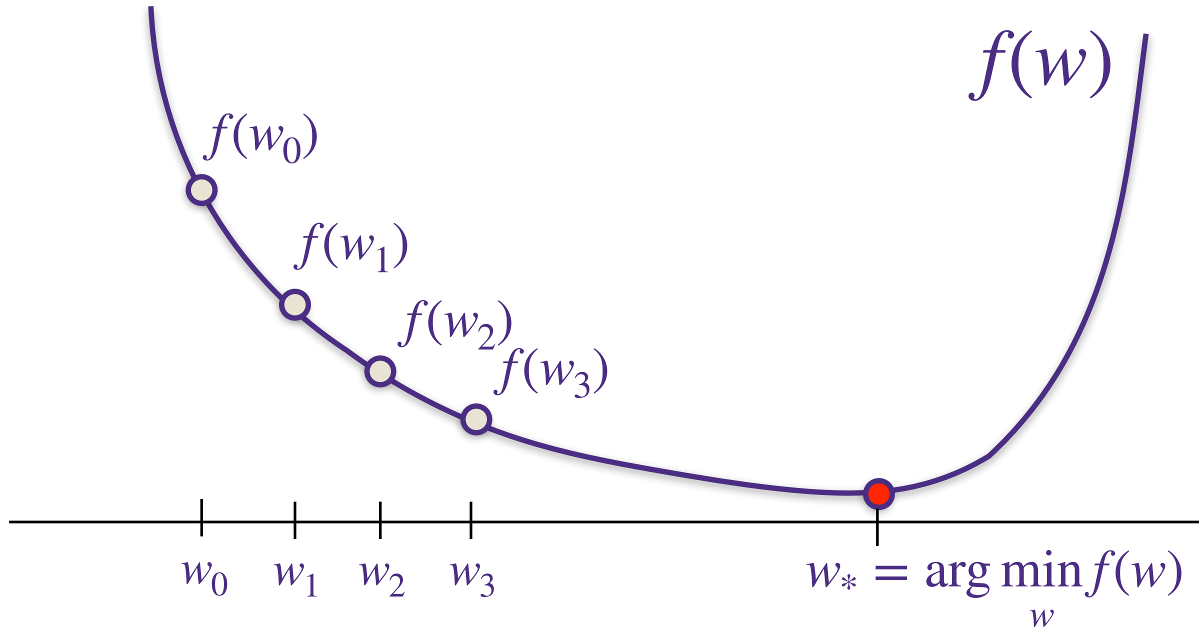
Gradient descent

For k=0,1,2,3,...

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

1-dimensional gradient descent

Let w_0 be an initial guess. How can we improve this solution?



Taylor series approximation:
For w very close to w_0 we have

$$f(w_0) + (w - w_0)\frac{df(w)}{dw}\Big|_{w=w_0}$$

is very close to $f(w)$

Thus, for very small $\eta > 0$,

if $w_1 = w_0 - \eta\frac{df(w)}{dw}\Big|_{w=w_0}$ then

$$f(w_0) - \eta\left(\frac{df(w)}{dw}\Big|_{w=w_0}\right)^2$$

is very close to $f(w_1) < f(w_0)$

Gradient descent
For k=0,1,2,3,…

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

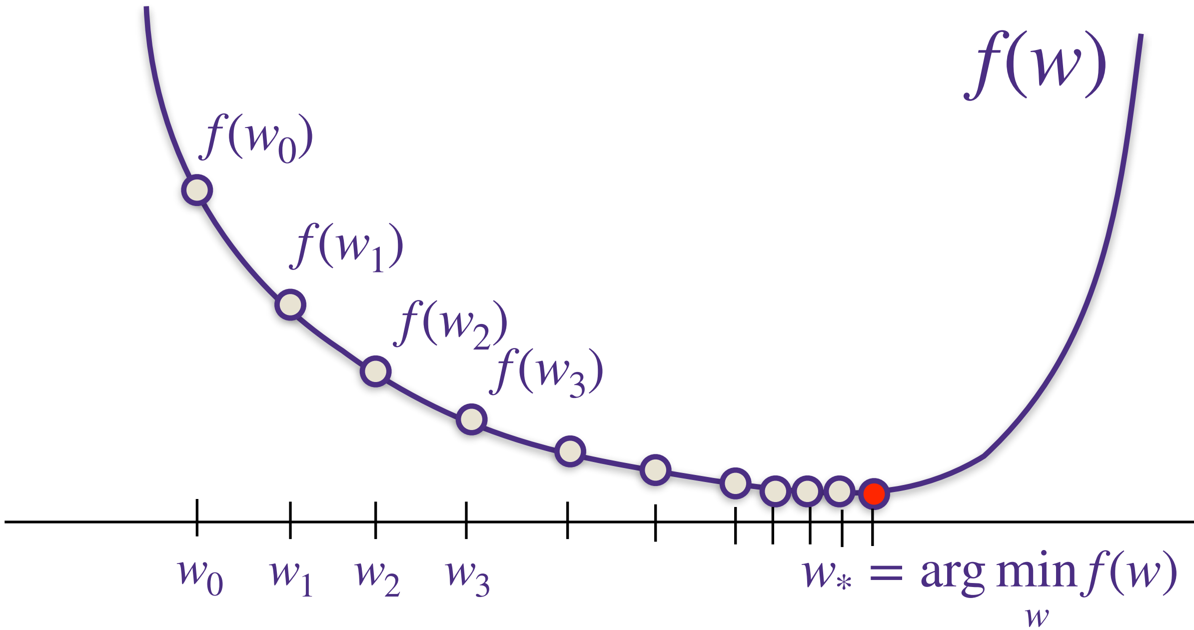
1-dimensional gradient descent

Let w_0 be an initial guess. How can we improve this solution?

Taylor series approximation:
For w very close to w_0 we have

$$f(w_0) + (w - w_0)\frac{df(w)}{dw}\Big|_{w=w_0}$$

is very close to $f(w)$



Thus, for very small $\eta > 0$,
if $w_1 = w_0 - \eta \frac{df(w)}{dw}\Big|_{w=w_0}$ then

$$f(w_0) - \eta\left(\frac{df(w)}{dw}\Big|_{w=w_0}\right)^2$$

is very close to $f(w_1) < f(w_0)$

Gradient descent
For k=0,1,2,3,...

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

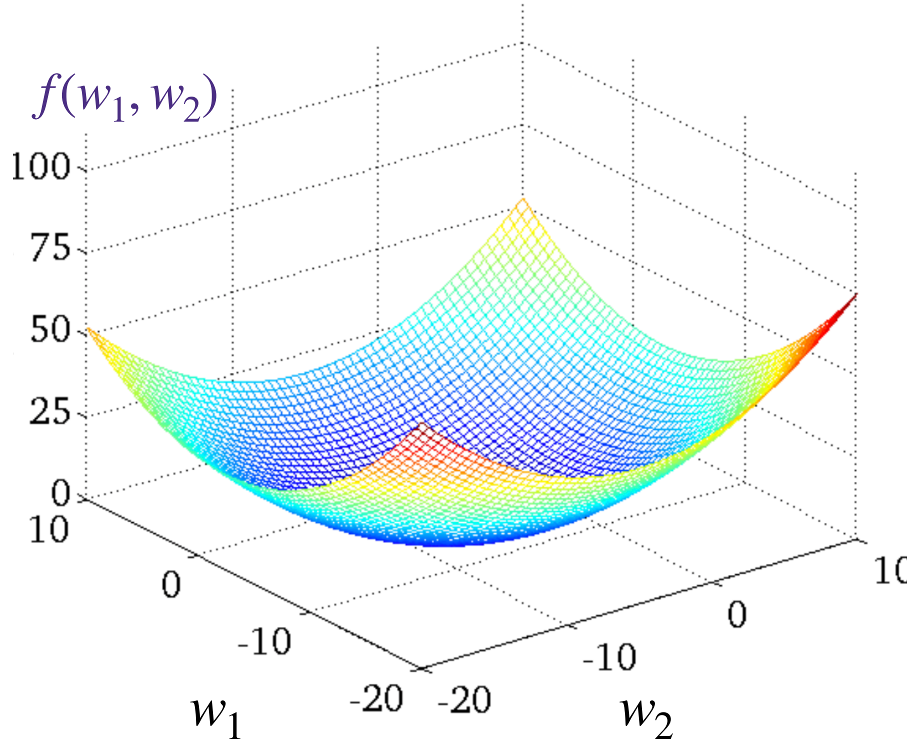
Note that as $k \to \infty$ we have $\frac{df(w)}{dw}\Big|_{w=w_k} \to 0$

Running example: linear regression

- **Given data:** $\{(x_i, y_i)\}_{i=1}^n \qquad x_i \in \mathbb{R}^d \qquad y_i \in \mathbb{R}$

- **Learning model parameters:**

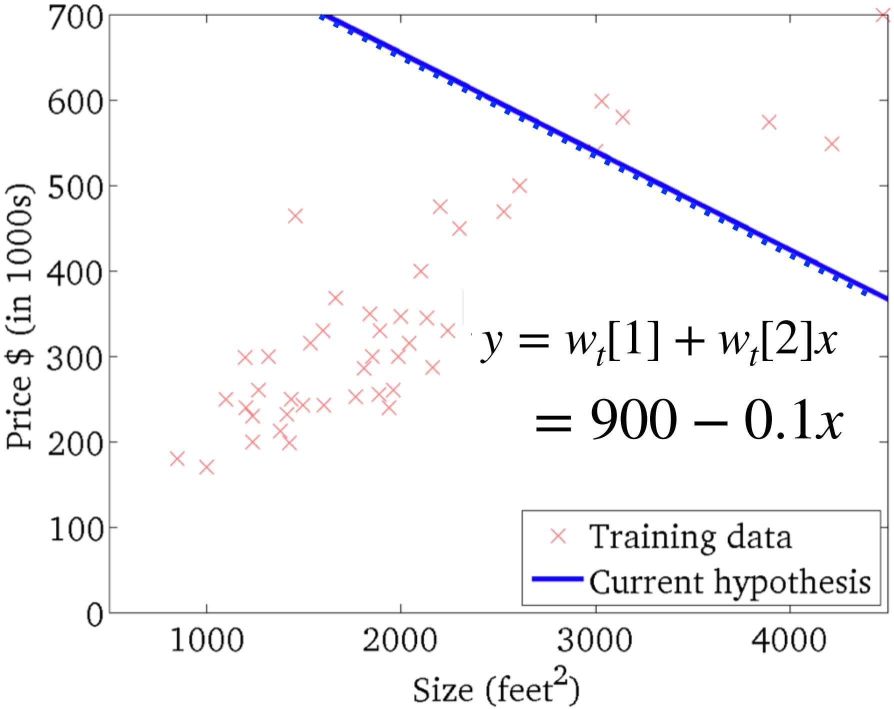
$$\hat{w}_{\text{LS}} = \arg\min_{w \in \mathbb{R}^d} \underbrace{\|\mathbf{y} - \mathbf{X}w\|_2^2}_{f(w)}$$



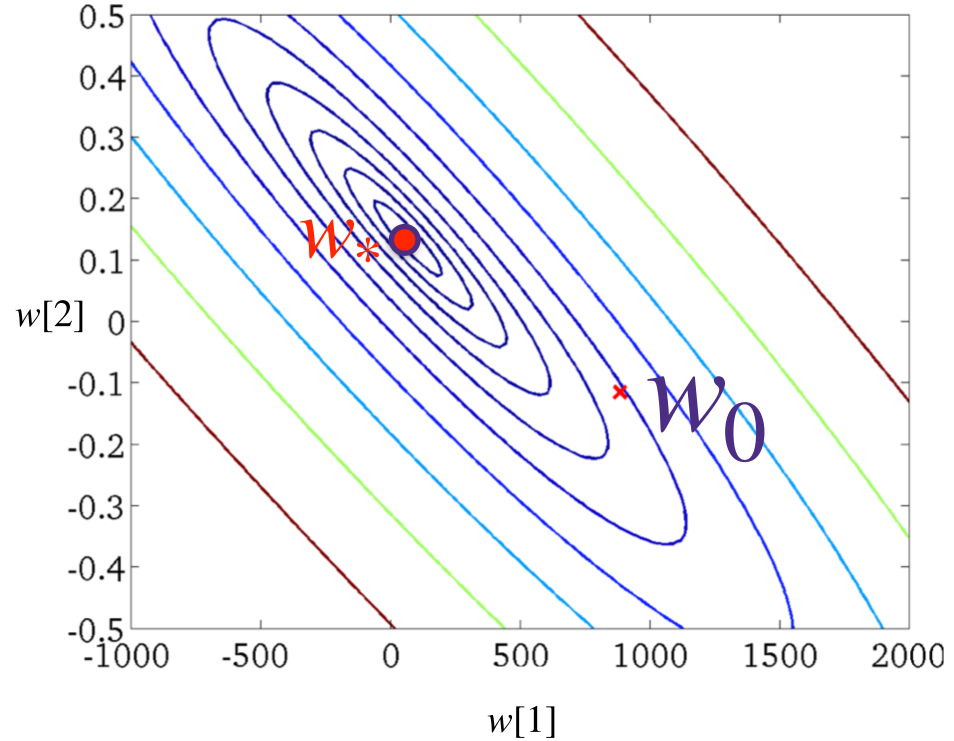
- **Gradient descent:**
  - Initialize: $w_0 = 0$
  - **For** t=0,1,2,…
    - $w_{t+1} \leftarrow w_t - \eta \cdot \nabla_w f(w_t)$

- $w_0 = (900, -0.1)$
- **For** t=0,1,2,...
  - $w_{t+1} \leftarrow w_t - \eta \cdot \nabla_w f(w_t)$

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



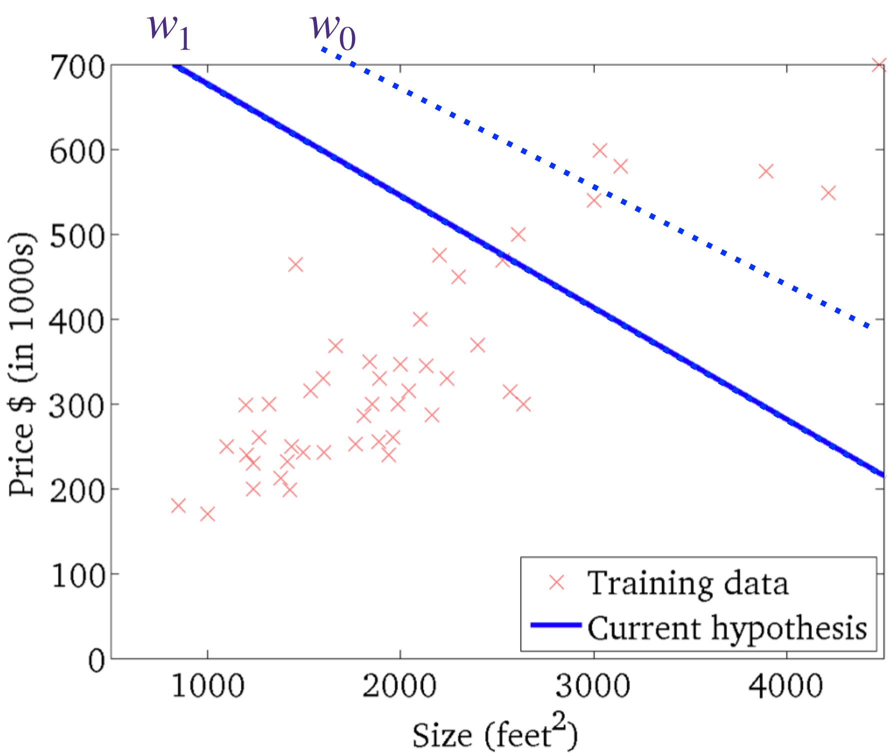
Evolution of the predictor



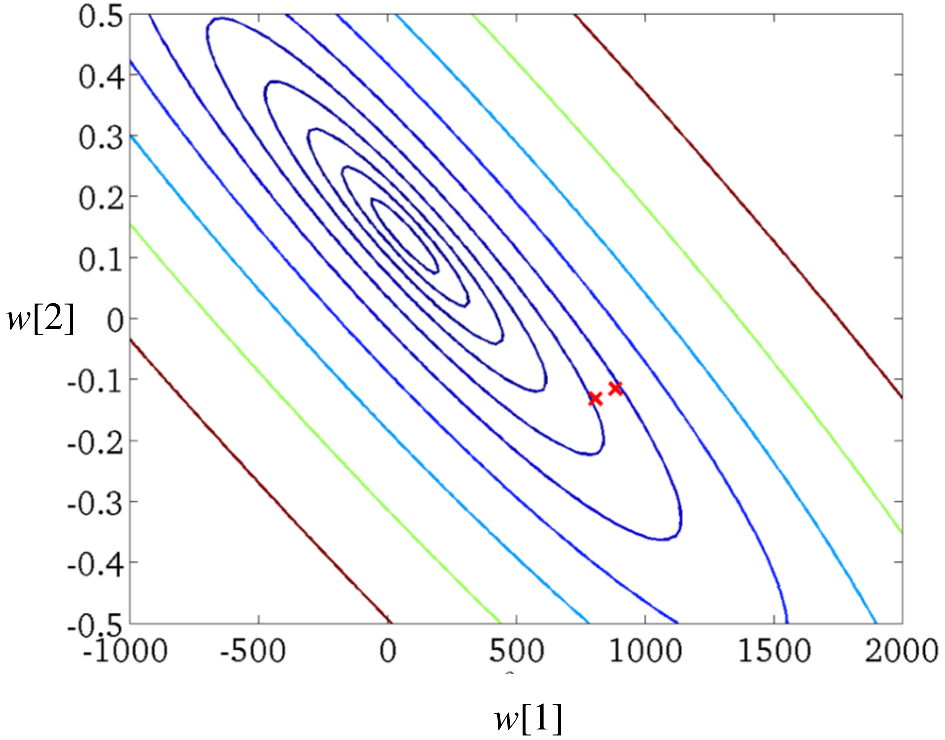
GD dynamics in the Parameter space

- Which direction will the GD move?

- $w_0 = (900, -0.1)$
- **For** t=0,1,2,…
  - $w_{t+1} \leftarrow w_t - \eta \cdot \nabla_w f(w_t)$

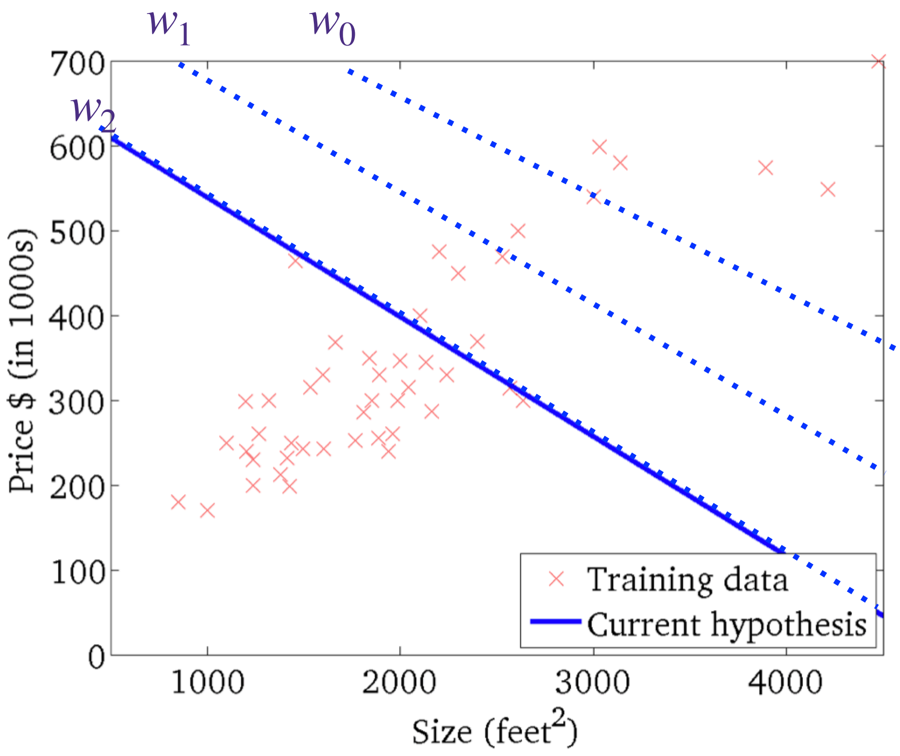


Evolution of the predictor

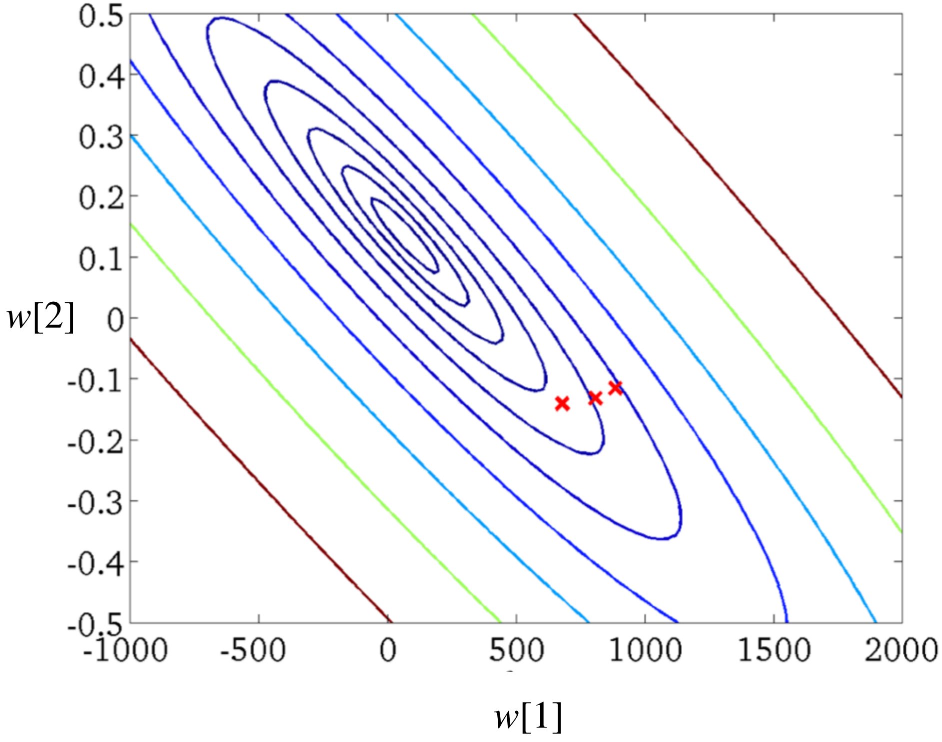


GD dynamics in the Parameter space

- $w_0 = (900, -0.1)$
- **For** t=0,1,2,…
  - $w_{t+1} \leftarrow w_t - \eta \cdot \nabla_w f(w_t)$

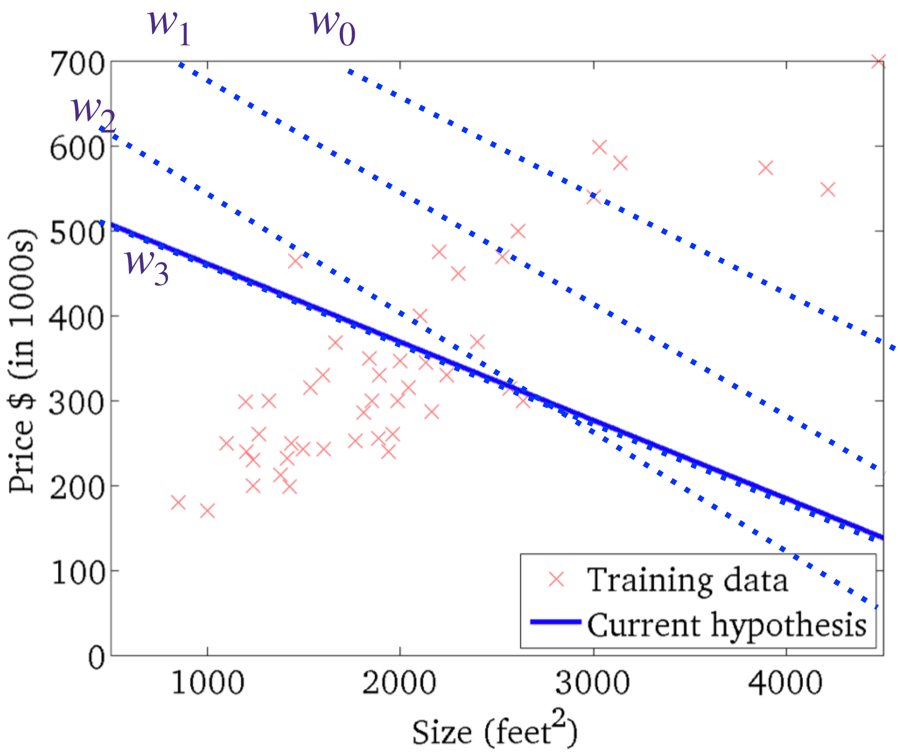


Evolution of the predictor

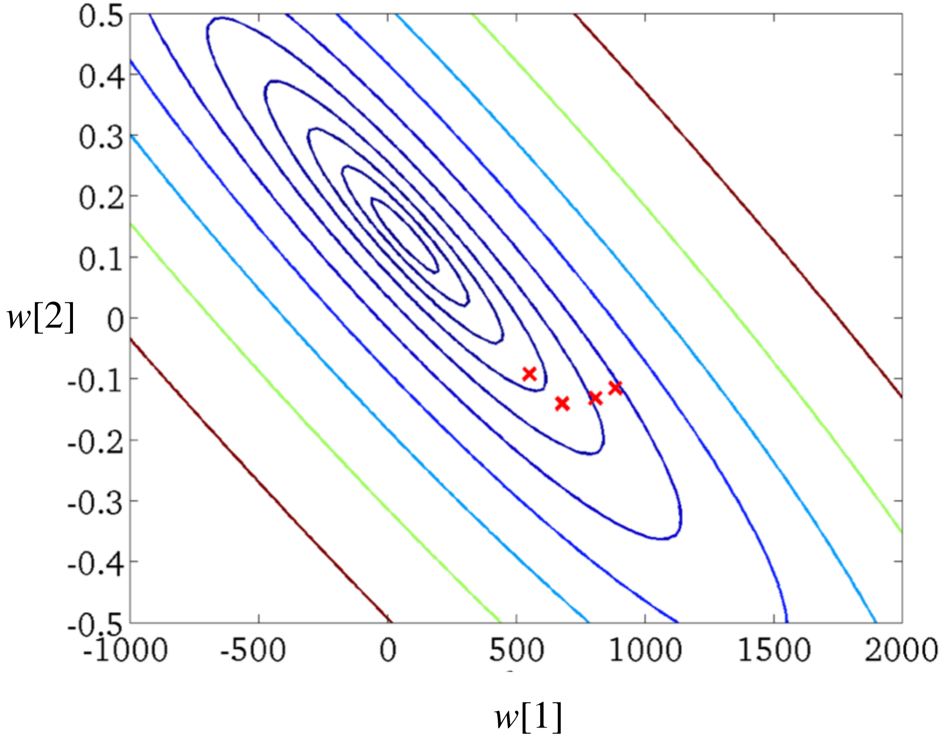


GD dynamics in the Parameter space

- $w_0 = (900, -0.1)$
- **For** t=0,1,2,…
  - $w_{t+1} \leftarrow w_t - \eta \cdot \nabla_w f(w_t)$

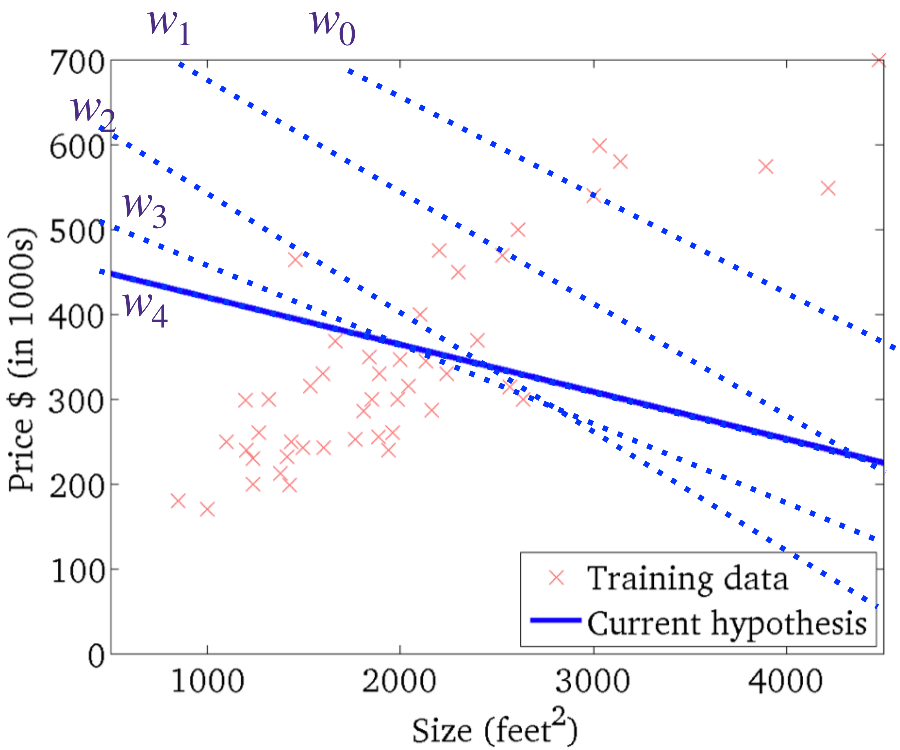


Evolution of the predictor

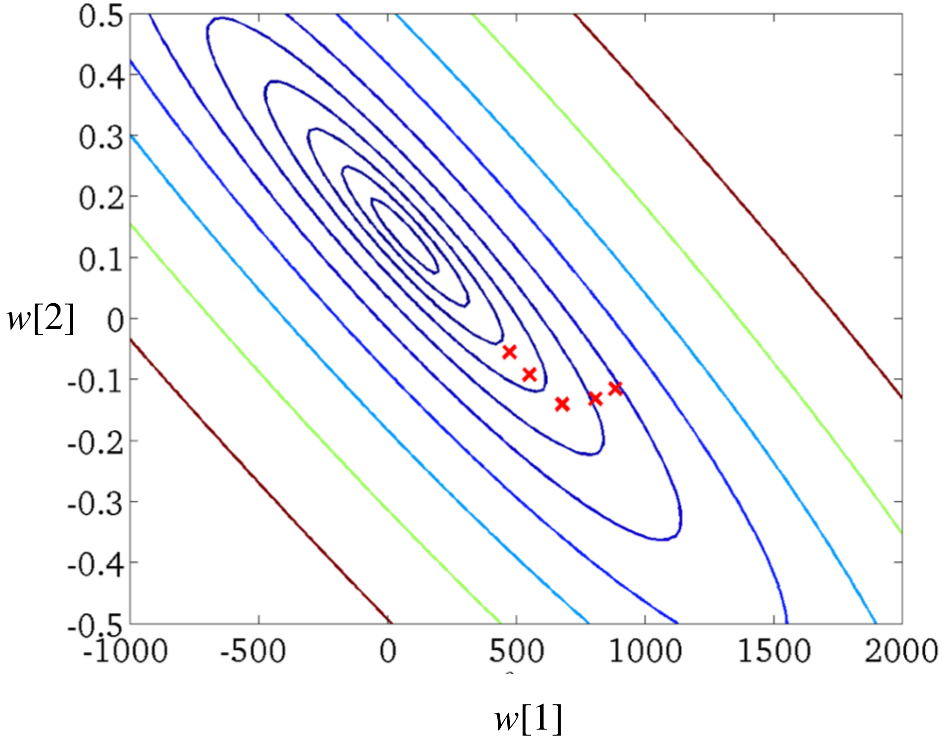


GD dynamics in the Parameter space

- $w_0 = (900, -0.1)$
- **For** t=0,1,2,…
    - $w_{t+1} \leftarrow w_t - \eta \cdot \nabla_w f(w_t)$

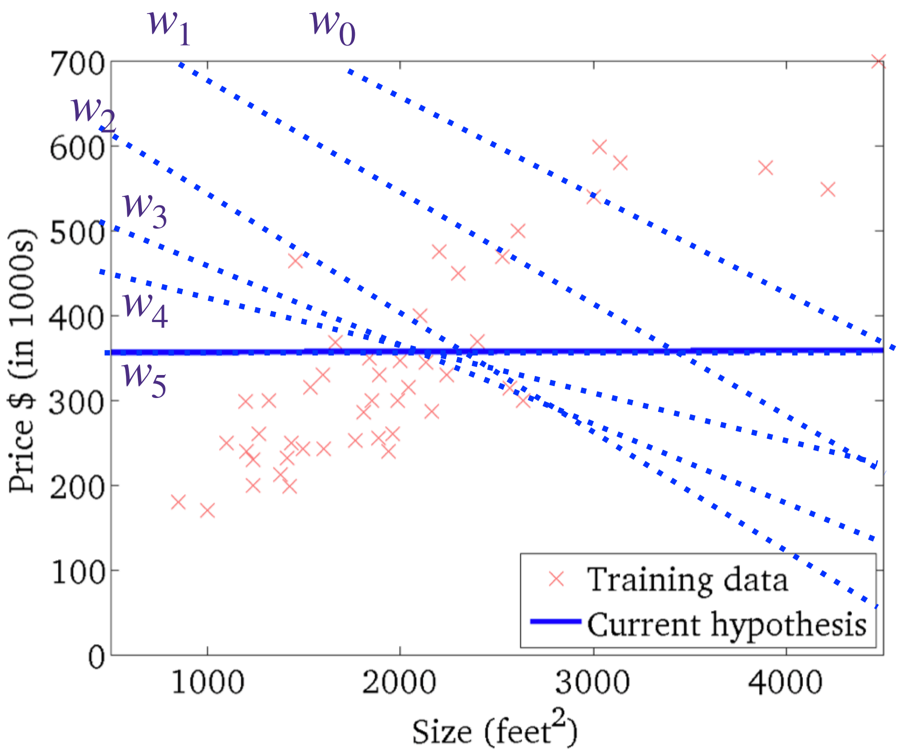


Evolution of the predictor

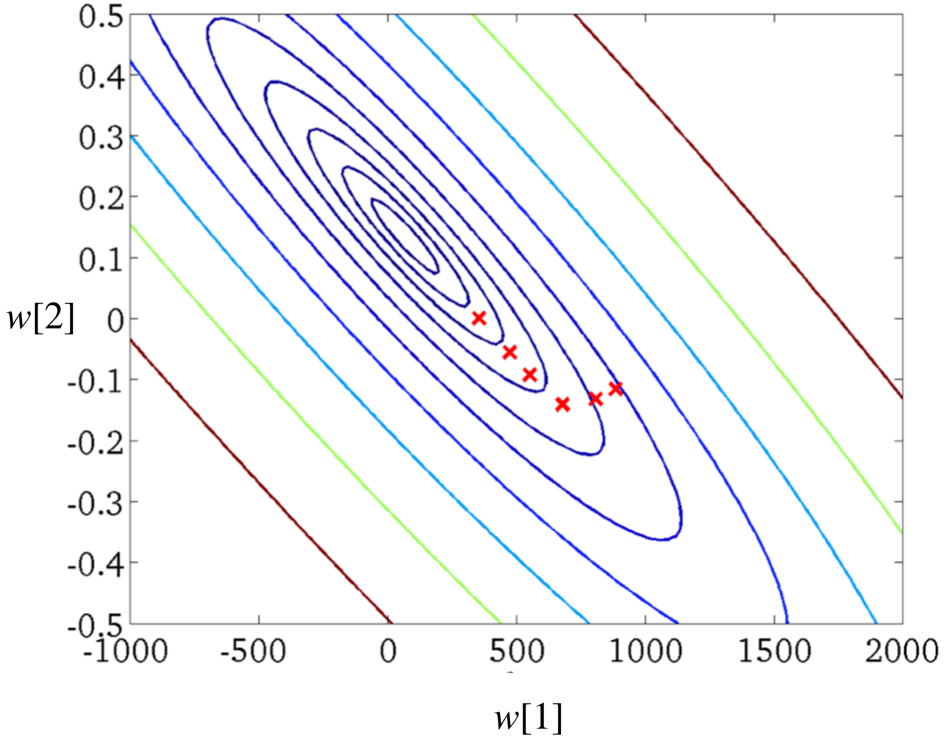


GD dynamics in the Parameter space

- $w_0 = (900, -0.1)$
- **For** t=0,1,2,...
  - $w_{t+1} \leftarrow w_t - \eta \cdot \nabla_w f(w_t)$

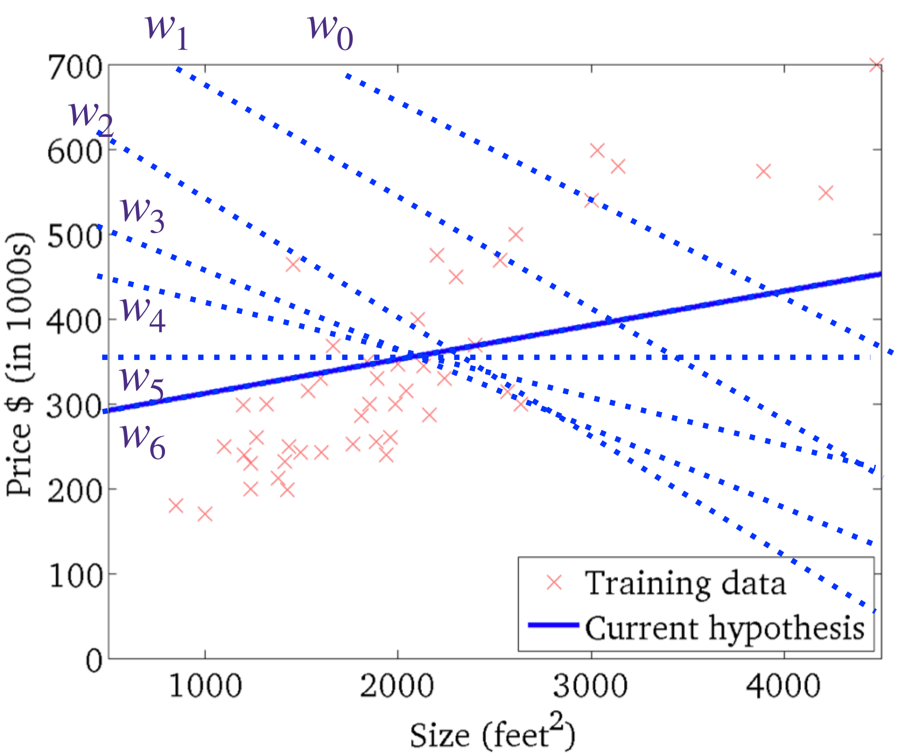


Evolution of the predictor

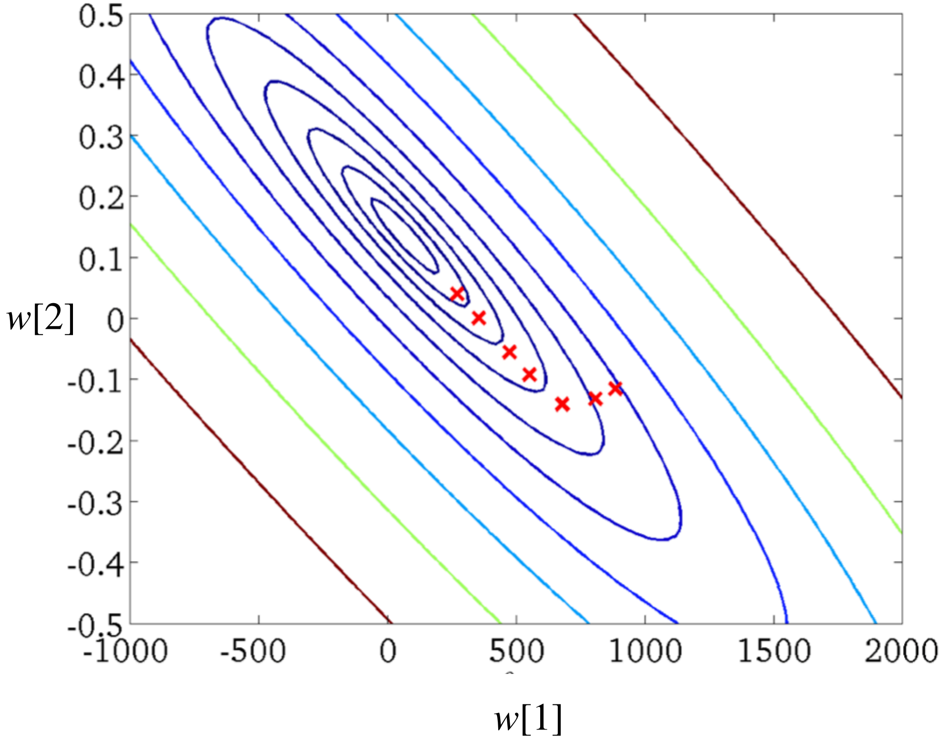


GD dynamics in the Parameter space

- $w_0 = (900, -0.1)$
- **For** t=0,1,2,…
  - $w_{t+1} \leftarrow w_t - \eta \cdot \nabla_w f(w_t)$

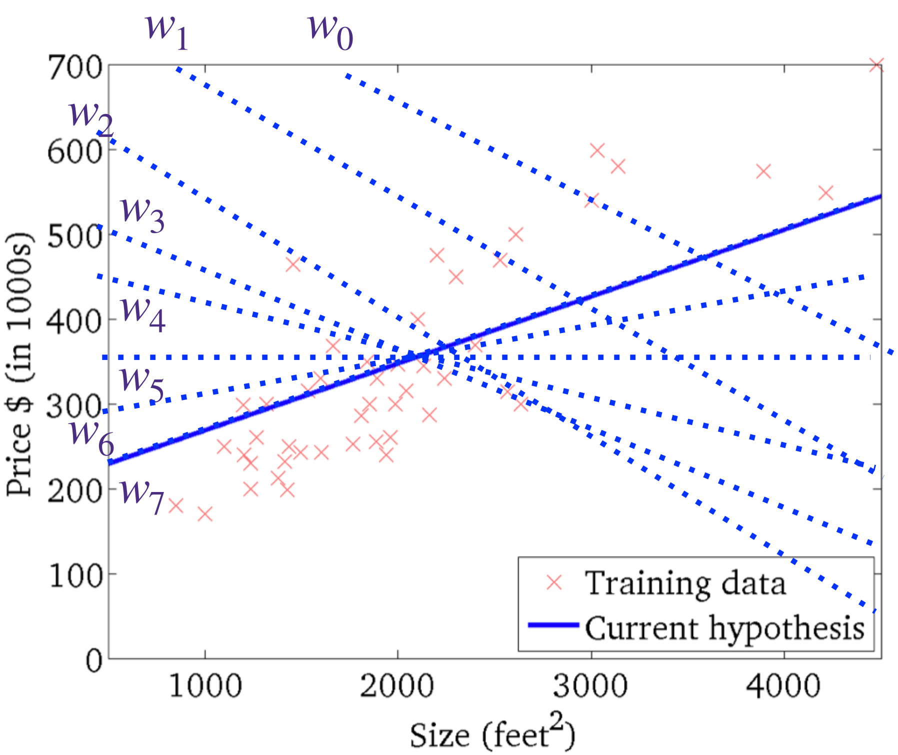


Evolution of the predictor

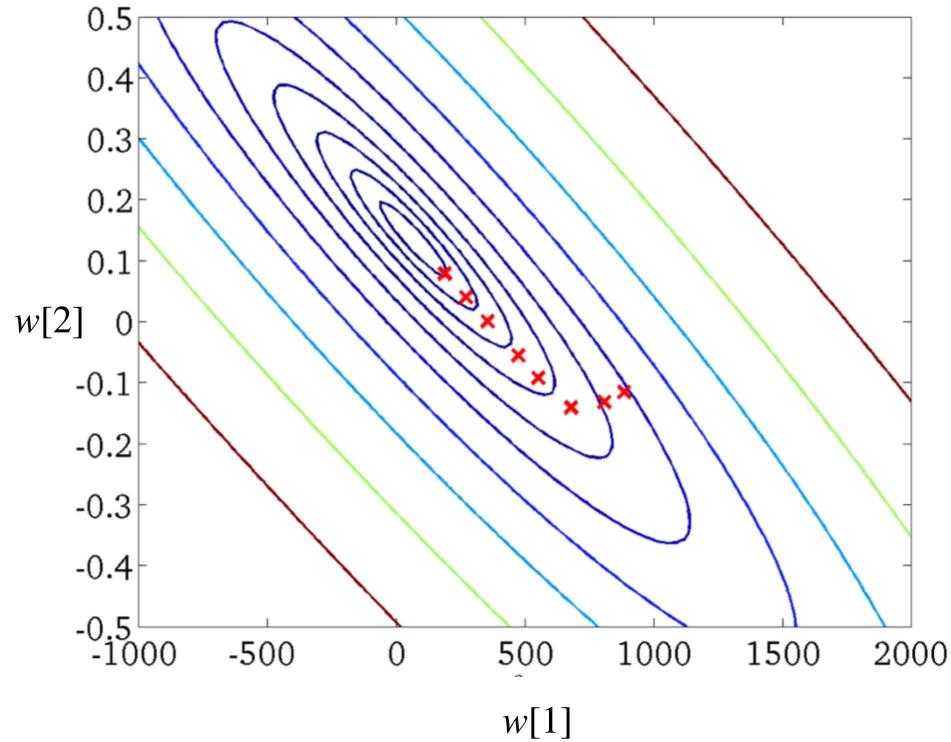


GD dynamics in the Parameter space

- $w_0 = (900, -0.1)$
- **For** t=0,1,2,…
  - $w_{t+1} \leftarrow w_t - \eta \cdot \nabla_w f(w_t)$

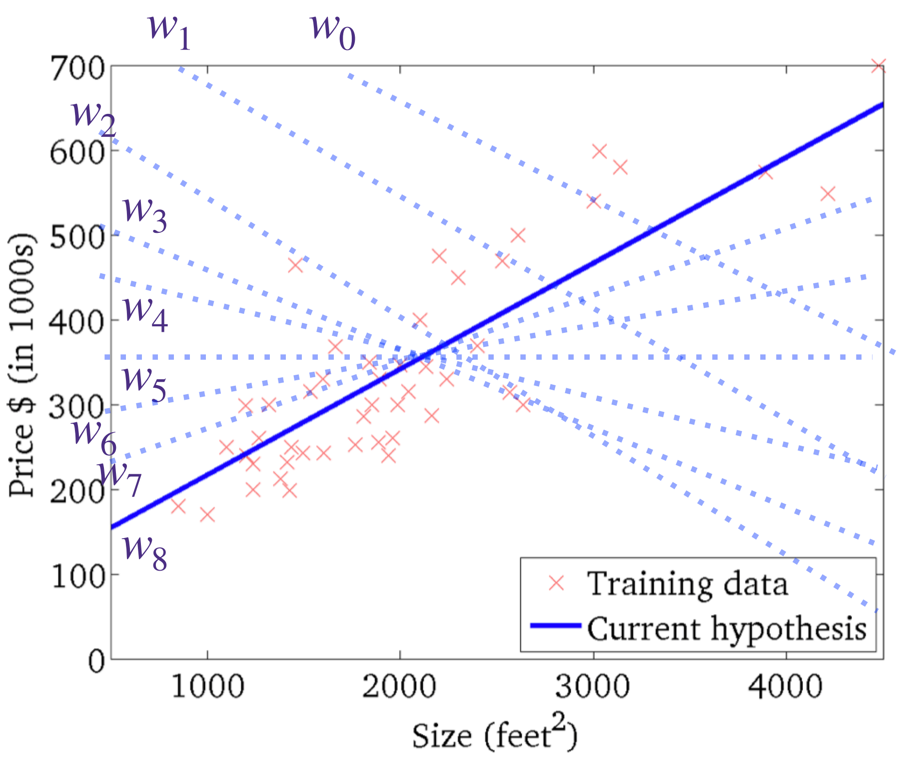


Evolution of the predictor

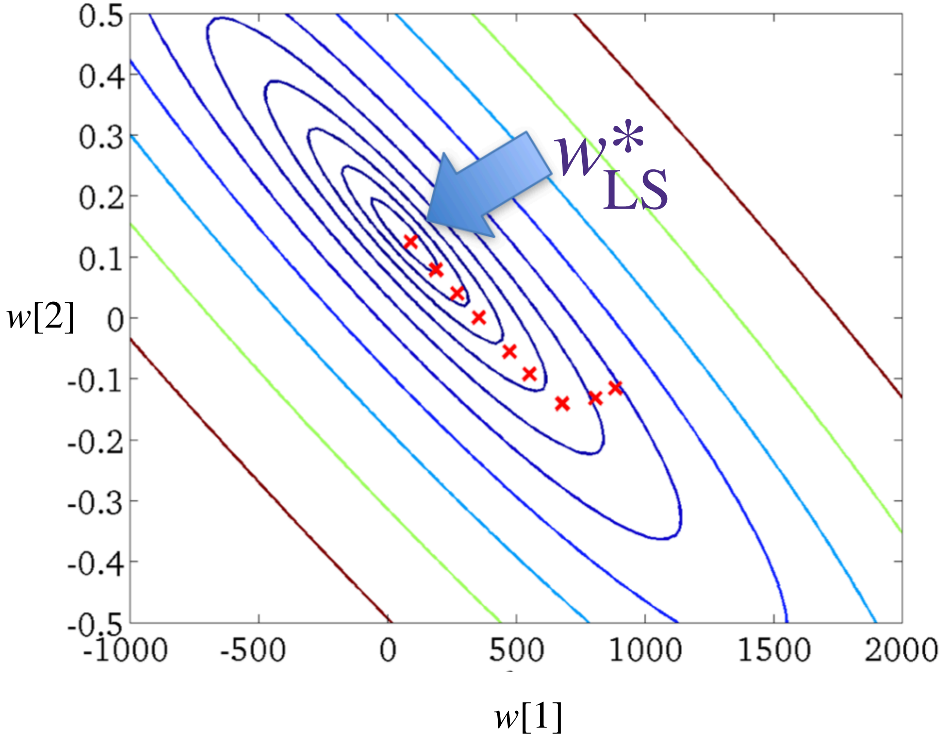


GD dynamics in the Parameter space

- $w_0 = (900, -0.1)$
- **For** t=0,1,2,…
  - $w_{t+1} \leftarrow w_t - \eta \cdot \nabla_w f(w_t)$



Evolution of the predictor



GD dynamics in the Parameter space

Gradient descent for linear regression

- In this example of linear regression, we can derive exactly the gradient descent trajectory
- Initialize: $w_0 = 0$
- **For** t=0,1,2,…
  - $w_{t+1} \leftarrow w_t - \eta \cdot \nabla_w f(w_t)$

$$\nabla f(w_t) = -2\mathbf{X}^T(\mathbf{y} - \mathbf{X}w_t)$$

For linear regression, we have

$$\hat{w}_{\mathrm{LS}} = \arg\min_{w \in \mathbb{R}^d} \underbrace{\|\mathbf{y} - \mathbf{X}w\|_2^2}_{f(w)}$$

Gradient descent for linear regression

- In this example of linear regression, we can derive exactly the gradient descent trajectory
- Initialize: $w_0 = 0$
- **For** t=0,1,2,...
  - $w_{t+1} \leftarrow w_t - \eta \cdot \nabla_w f(w_t)$

For linear regression, we have

$$\hat{w}_{\text{LS}} = \arg\min_{w \in \mathbb{R}^d} \underbrace{\|\mathbf{y} - \mathbf{X}w\|_2^2}_{f(w)}$$

$$\nabla f(w_t) = -2\mathbf{X}^T(\mathbf{y} - \mathbf{X}w_t)$$

$$w_{t+1} = w_t + \eta 2\mathbf{X}^T(\mathbf{y} - \mathbf{X}w_t) = (\mathbf{I} - 2\eta\mathbf{X}^T\mathbf{X})w_t + 2\eta\mathbf{X}^T\mathbf{y}$$

Gradient descent for linear regression

- In this example of linear regression, we can derive exactly the gradient descent trajectory
- Initialize: $w_0 = 0$
- **For** t=0,1,2,…
  - $w_{t+1} \leftarrow w_t - \eta \cdot \nabla_w f(w_t)$

For linear regression, we have

$$\hat{w}_{\text{LS}} = \arg\min_{w \in \mathbb{R}^d} \underbrace{\|\mathbf{y} - \mathbf{X}w\|_2^2}_{f(w)}$$

$$\nabla f(w_t) = -2\mathbf{X}^T(\mathbf{y} - \mathbf{X}w_t)$$

$$w_{t+1} = w_t + \eta 2\mathbf{X}^T(\mathbf{y} - \mathbf{X}w_t) = (\mathbf{I} - 2\eta\mathbf{X}^T\mathbf{X})w_t + 2\eta\mathbf{X}^T\mathbf{y}$$

Let the least-squares solution be $w^* = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{y}$

$$\begin{aligned} w_{t+1} - w^* &= (\mathbf{I} - 2\eta\mathbf{X}^T\mathbf{X})w_t + 2\eta\mathbf{X}^T\mathbf{y} - w^* \\ &= (\mathbf{I} - 2\eta\mathbf{X}^T\mathbf{X})(w_t - w^*) + 2\eta\mathbf{X}^T\mathbf{y} - 2\eta\mathbf{X}^T\mathbf{X}w^* \\ &= (\mathbf{I} - 2\eta\mathbf{X}^T\mathbf{X})(w_t - w^*) \end{aligned}$$

Gradient descent for linear regression

- Initialize: $w_0 = 0$
- **For** t=0,1,2,…
  - $w_{t+1} \leftarrow w_t - \eta \cdot \nabla_w f(w_t)$

For linear regression, we have

$$\hat{w}_{LS} = \arg\min_{w \in \mathbb{R}^d} \underbrace{\|\mathbf{y} - \mathbf{X}w\|_2^2}_{f(w)}$$

$$\nabla f(w_t) = -2\mathbf{X}^T(\mathbf{y} - \mathbf{X}w_t)$$

$$w_{t+1} = w_t + \eta 2\mathbf{X}^T(\mathbf{y} - \mathbf{X}w_t) = (\mathbf{I} - 2\eta\mathbf{X}^T\mathbf{X})w_t + 2\eta\mathbf{X}^T\mathbf{y}$$

Let the least-squares solution be $w^* = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{y}$

$$\begin{aligned} w_{t+1} - w^* &= (\mathbf{I} - 2\eta\mathbf{X}^T\mathbf{X})w_t + 2\eta\mathbf{X}^T\mathbf{y} - w^* \\ &= (\mathbf{I} - 2\eta\mathbf{X}^T\mathbf{X})(w_t - w^*) + 2\eta\mathbf{X}^T\mathbf{y} - 2\eta\mathbf{X}^T\mathbf{X}w^* \\ &= (\mathbf{I} - 2\eta\mathbf{X}^T\mathbf{X})(w_t - w^*) \end{aligned}$$

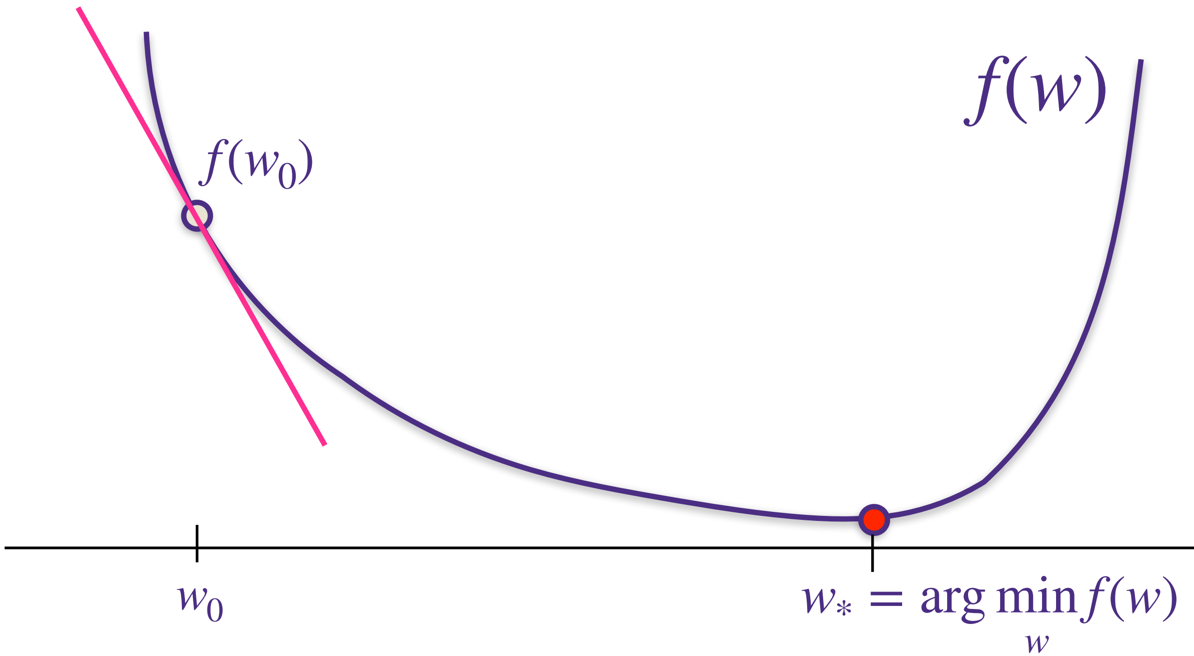
How do you choose step size?

Let w_0 be an initial guess. How can we improve this solution?

Taylor series approximation:
For w very close to w_0 we have

$$f(w_0) + (w - w_0)\frac{df(w)}{dw}\Big|_{w=w_0}$$

is very close to $f(w)$



If η too big, does not converge!
If η too small, converges very, very slowly.

In practice: choose the largest value of η that converges (guess and check)

Gradient descent for Ridge regression

- Initialize: $w_0 = 0$
- **For** t=0,1,2,…
  - $w_{t+1} \leftarrow w_t - \eta \cdot \nabla_w f(w_t)$

For Ridge we have

$$\hat{w}_{\text{Ridge}} = \arg\min_{w \in \mathbb{R}^d} \underbrace{\frac{1}{2}\|\mathbf{y} - \mathbf{X}w\|_2^2 + \frac{\lambda}{2}\|w\|_2^2}_{f(w)}$$

$$\nabla f(w_t) =$$

$$w_{t+1} =$$

Gradient descent for Ridge regression

- Initialize: $w_0 = 0$
- **For** t=0,1,2,…
    - $w_{t+1} \leftarrow w_t - \eta \cdot \nabla_w f(w_t)$

For Ridge we have

$$\hat{w}_{\text{Ridge}} = \arg\min_{w \in \mathbb{R}^d} \underbrace{\frac{1}{2}\|\mathbf{y} - \mathbf{X}w\|_2^2 + \frac{\lambda}{2}\|w\|_2^2}_{f(w)}$$

$$\nabla f(w_t) = -\mathbf{X}^T(\mathbf{y} - \mathbf{X}w_t) + \lambda w_t$$

$$w_{t+1} = (1 - \lambda)w_t + \eta \mathbf{X}^T(\mathbf{y} - \mathbf{X}w_t)$$

Gradient descent for Lasso regression

- Initialize: $w_0 = 0$
- **For** t=0,1,2,...
  - $w_{t+1} \leftarrow w_t - \eta \cdot \nabla_w f(w_t)$

For Lasso we have

$$\hat{w}_{\text{Lasso}} = \arg\min_{w \in \mathbb{R}^d} \underbrace{\frac{1}{2}\|\mathbf{y} - \mathbf{X}w\|_2^2 + \lambda\|w\|_1}_{f(w)}$$

$$\nabla f(w_t) =$$

$$w_{t+1} =$$

Gradient descent for Lasso regression

- Initialize: $w_0 = 0$
- **For** t=0,1,2,…
  - $w_{t+1} \leftarrow w_t - \eta \cdot \nabla_w f(w_t)$

For Lasso we have

$$\hat{w}_{\text{Lasso}} = \arg\min_{w \in \mathbb{R}^d} \underbrace{\frac{1}{2}\|\mathbf{y} - \mathbf{X}w\|_2^2 + \lambda\|w\|_1}_{f(w)}$$

$$\nabla f(w_t) = -\mathbf{X}^T(\mathbf{y} - \mathbf{X}w_t) + \lambda \text{sign}(w_t)$$

$$w_{t+1} = w_t + \eta \mathbf{X}^T(\mathbf{y} - \mathbf{X}w_t) - \lambda \text{sign}(w_t)$$