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

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

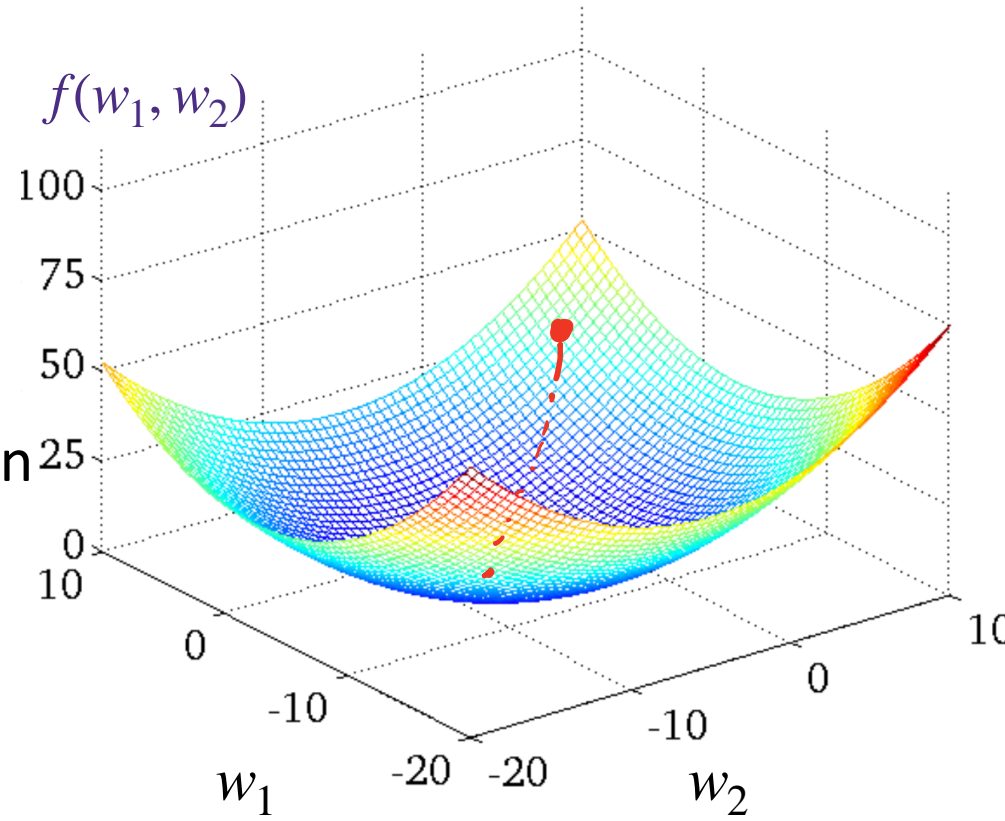
Running example: linear regression

- **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:**

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

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

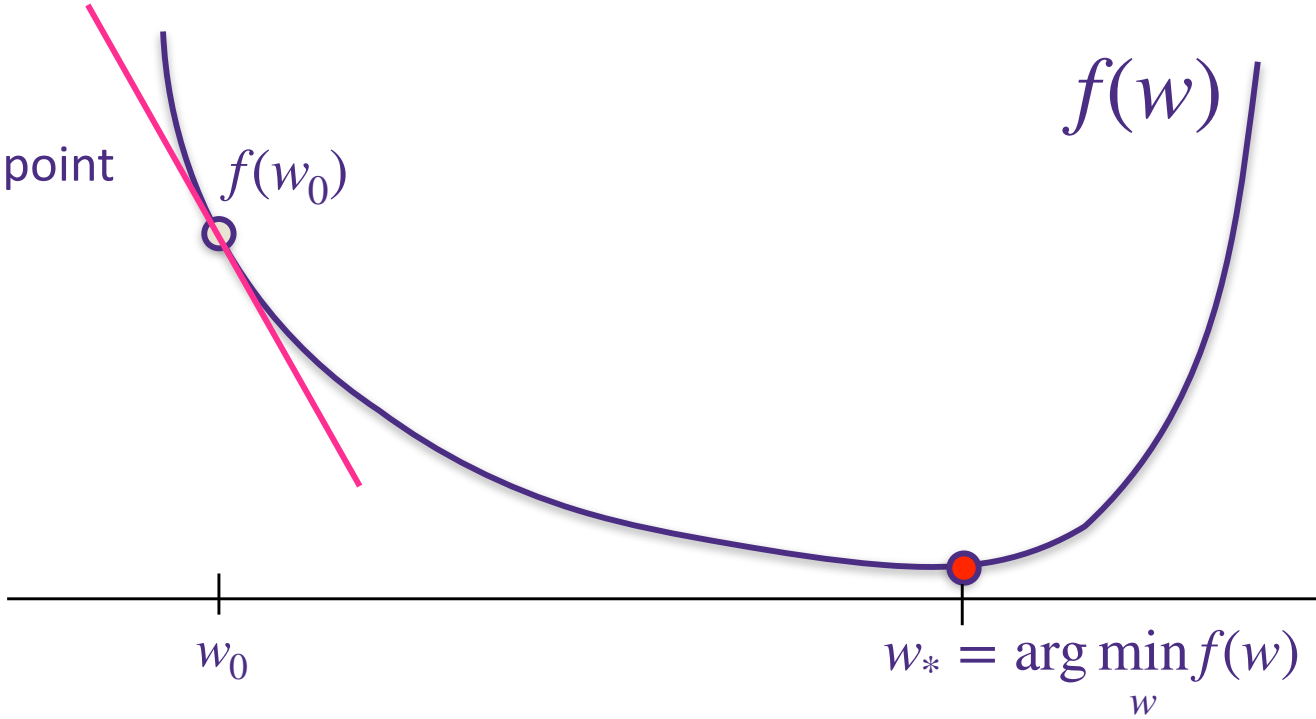


1-dimensional gradient descent

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

Derivative tells rate of change at a point

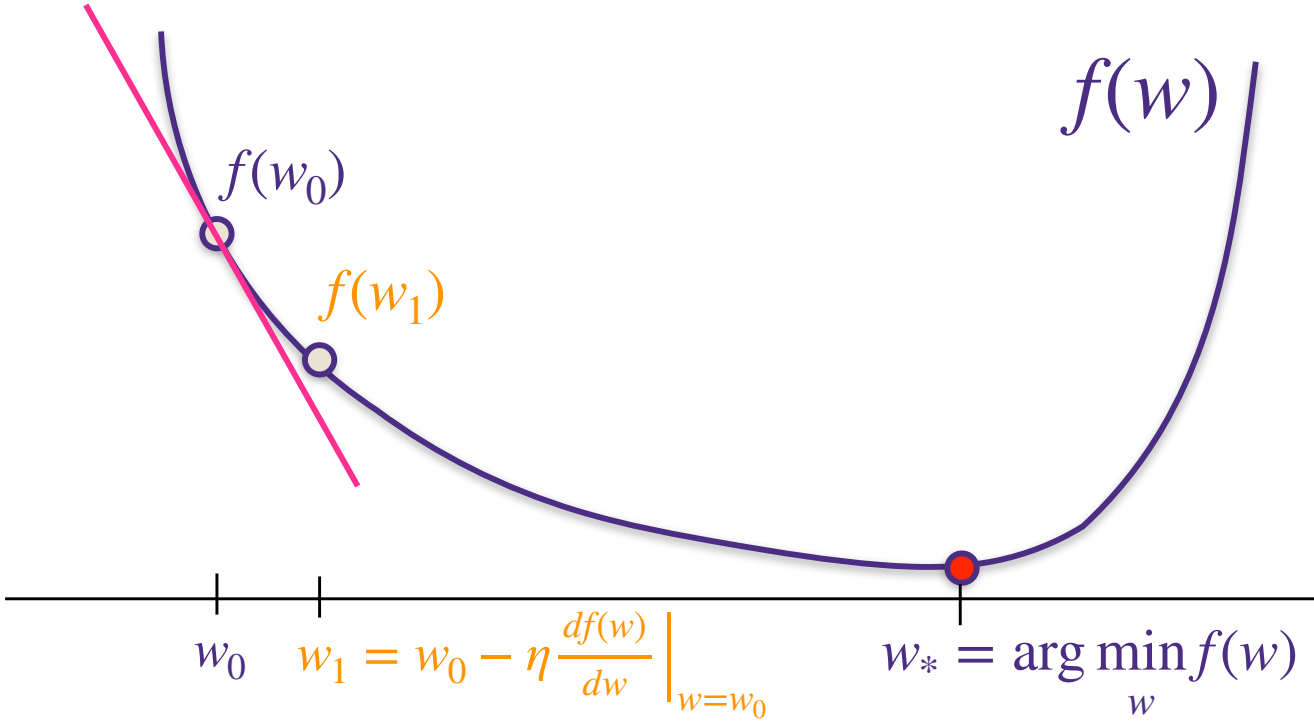
$$\frac{\partial}{\partial w} f(w)\Big|_{w=w_0}$$



Idea: If the function is convex, then stepping the *opposite* direction of the derivative gets us closer to minimizing the function

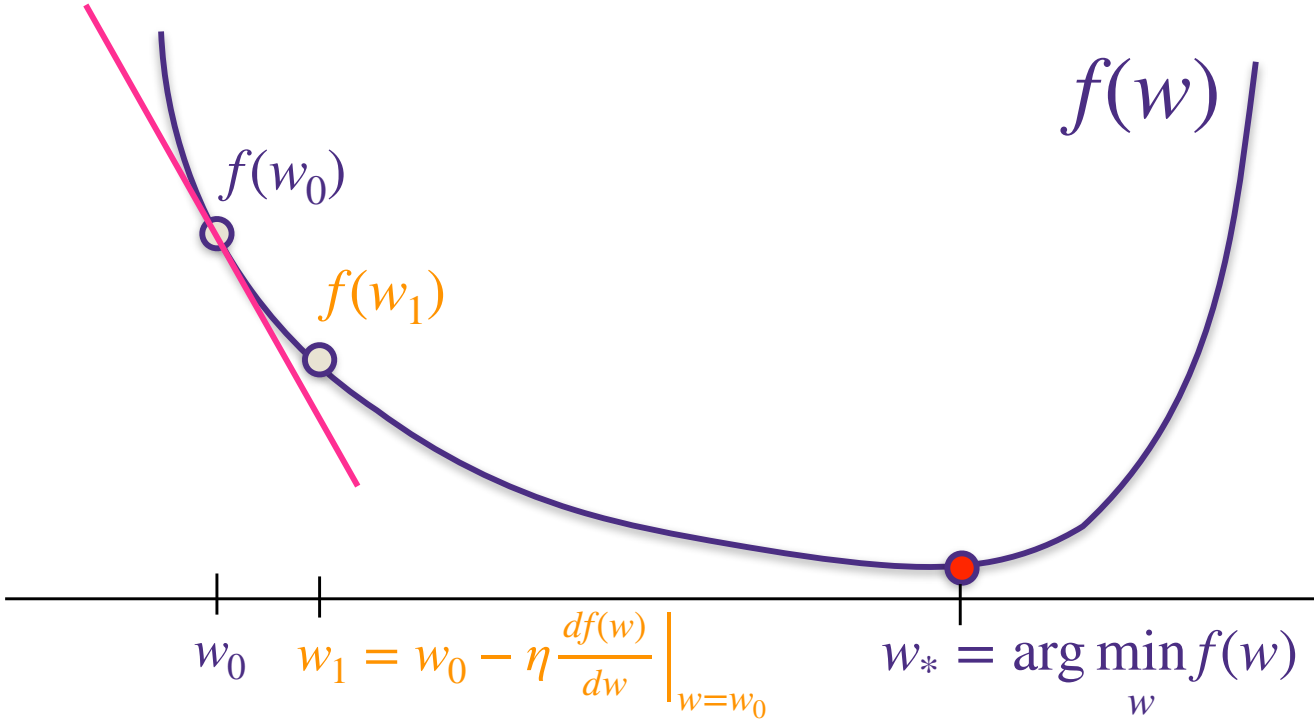
1-dimensional gradient descent

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



1-dimensional gradient descent

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



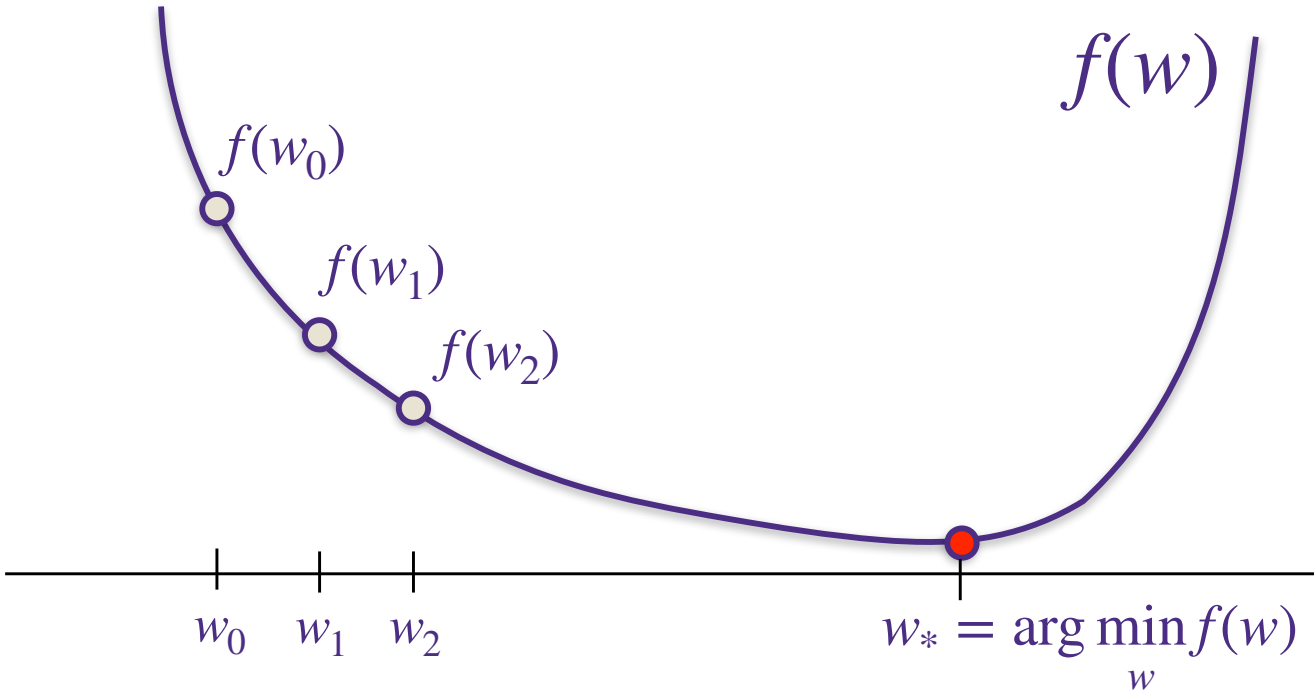
Gradient descent

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?



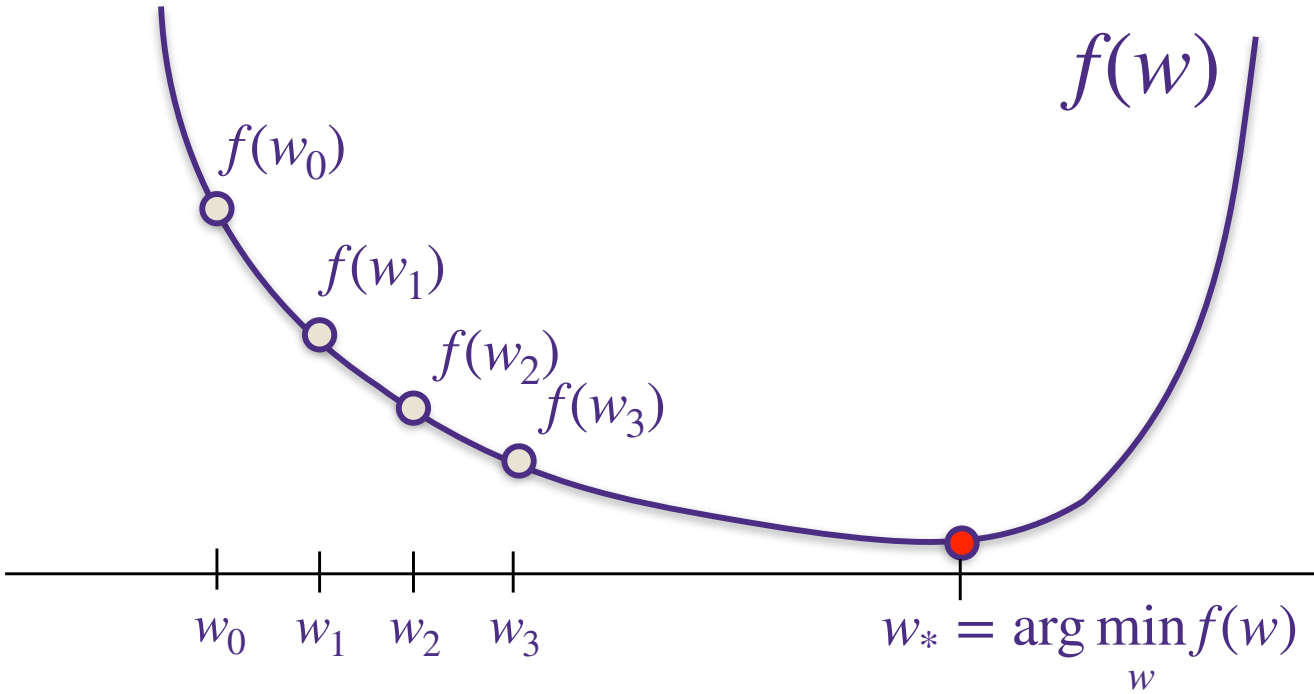
Gradient descent

For k=0,1,2,3,…

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

1-dimensional gradient descent

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



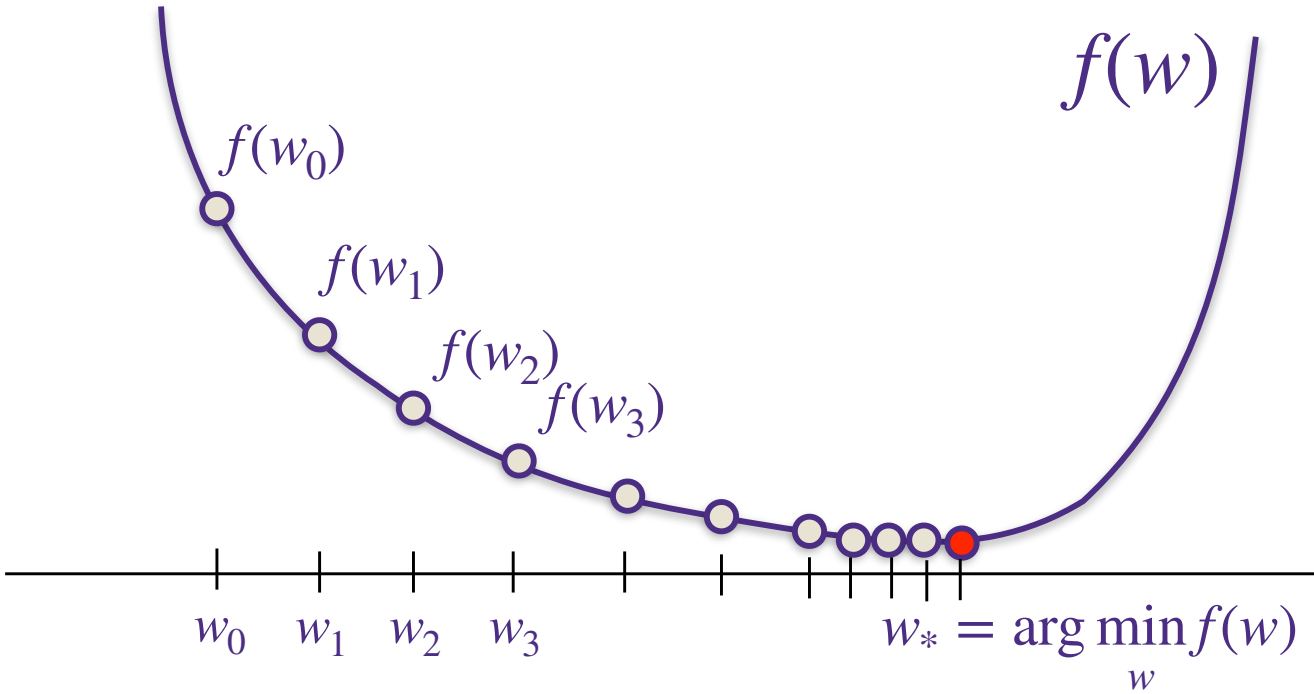
Gradient descent

For k=0,1,2,3,…

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

1-dimensional gradient descent

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



Gradient descent

For k=0,1,2,3,…

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

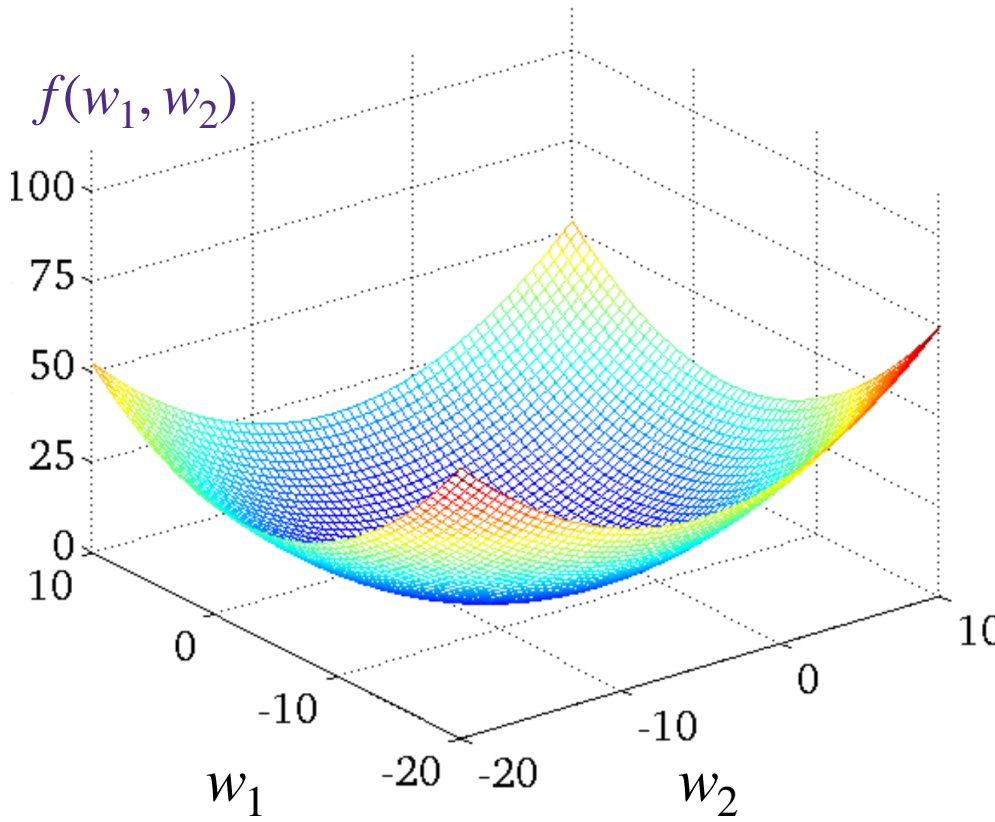
Note that as $k \to \infty$ we have $\left.\frac{df(w)}{dw}\right|_{w=w_k} \to 0$ (assuming small enough η)

Running example: Linear Regression

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

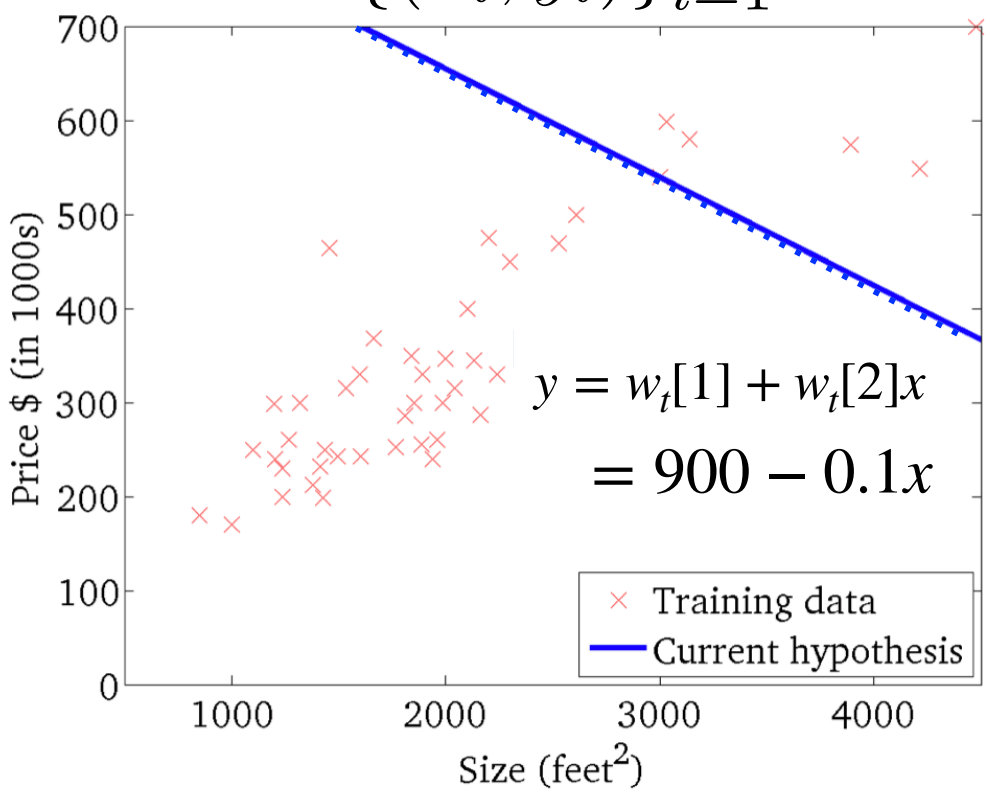
- **Learning model parameters:**

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

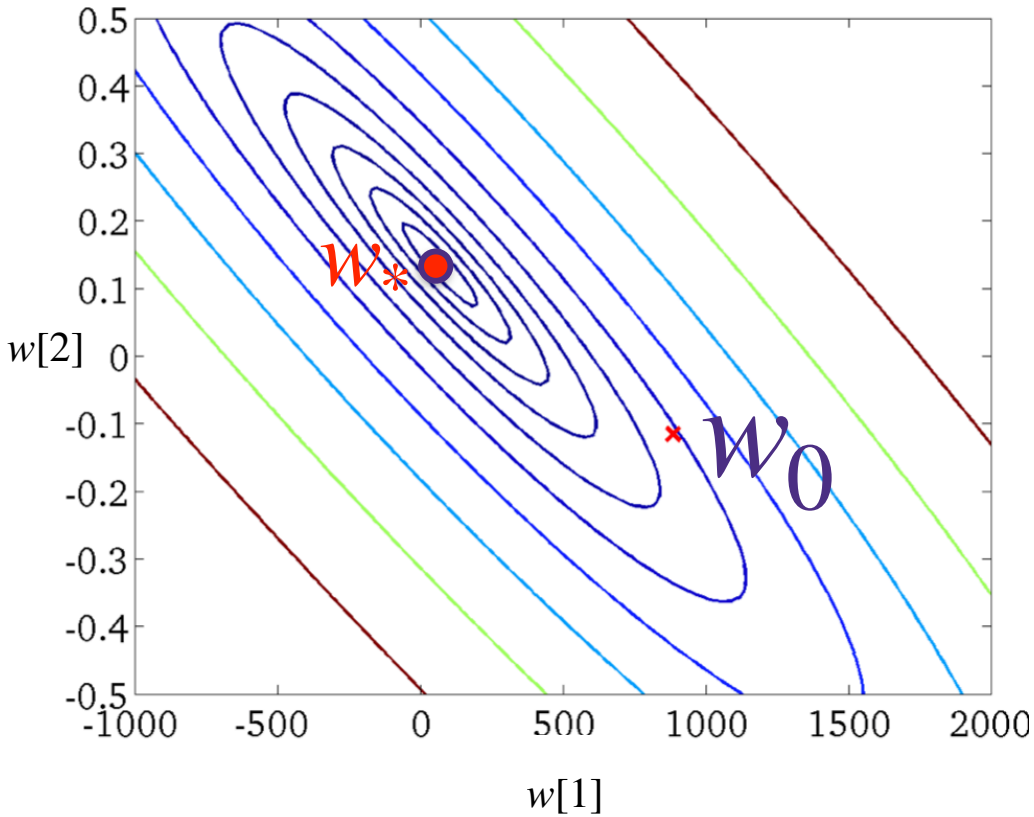


- **Gradient descent:**
  - Initialize: $w_0 = 0$
  - **For** t=0,1,2,…
    - $w_{t+1} \leftarrow w_t - \eta \cdot \nabla_w f(w_t)$

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



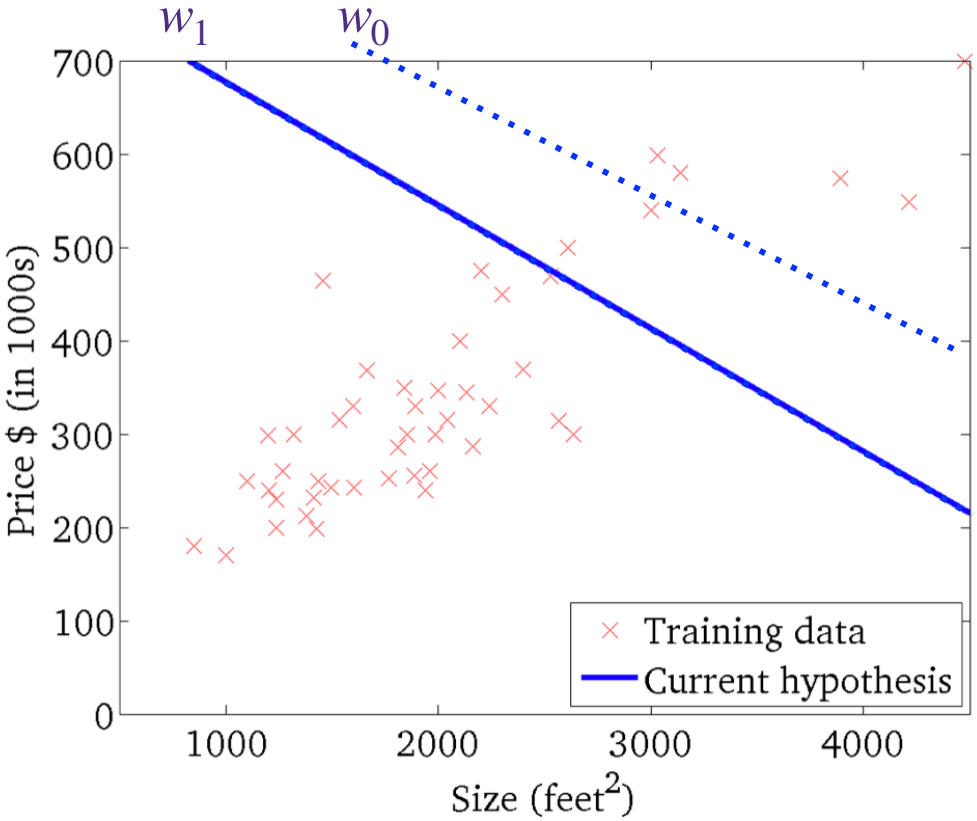
Evolution of the predictor



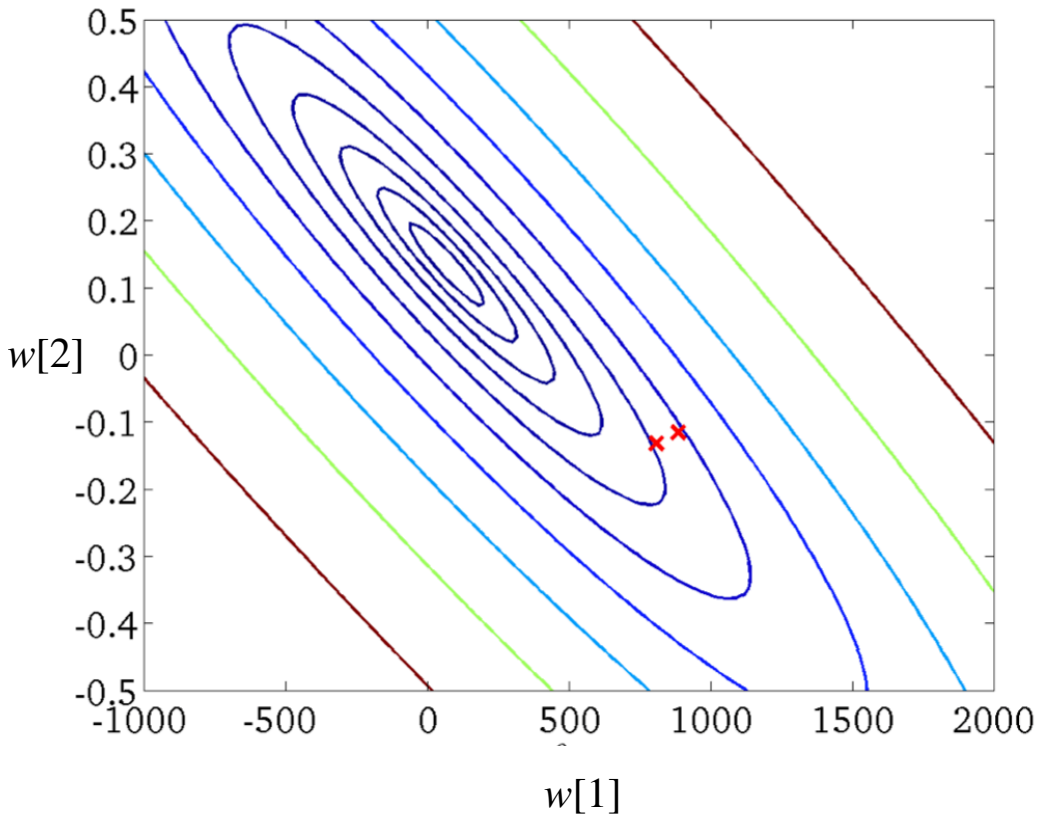
GD dynamics in the Parameter space

- Which direction will the GD move?

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

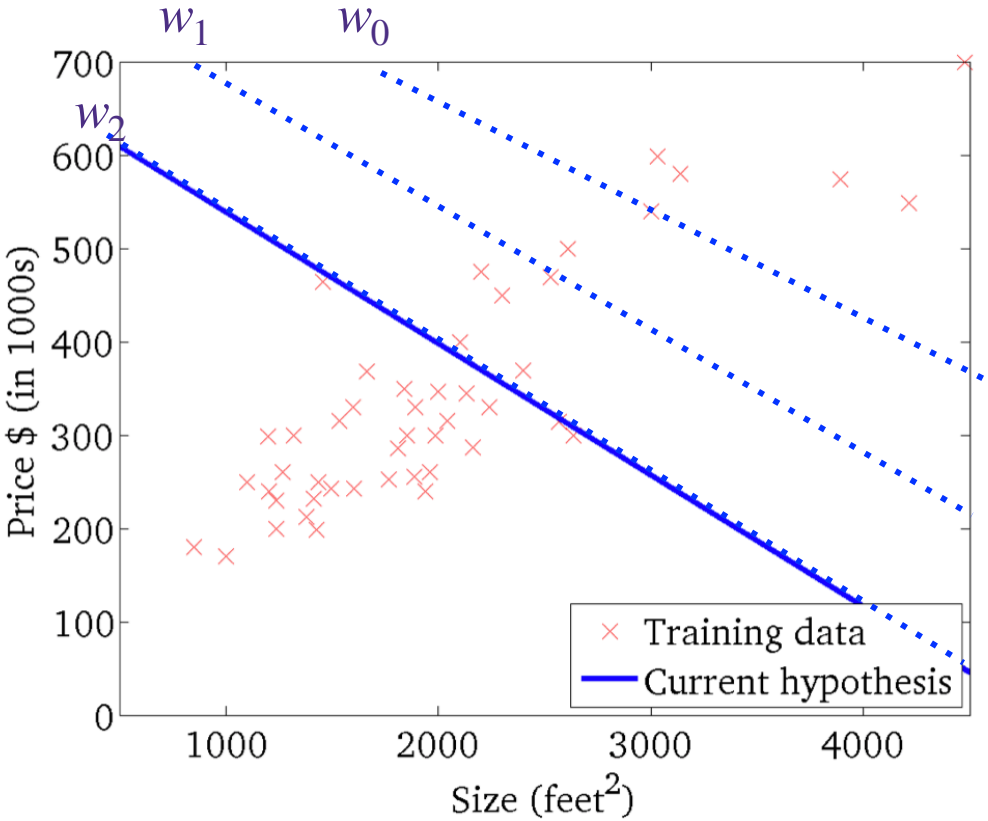


Evolution of the predictor

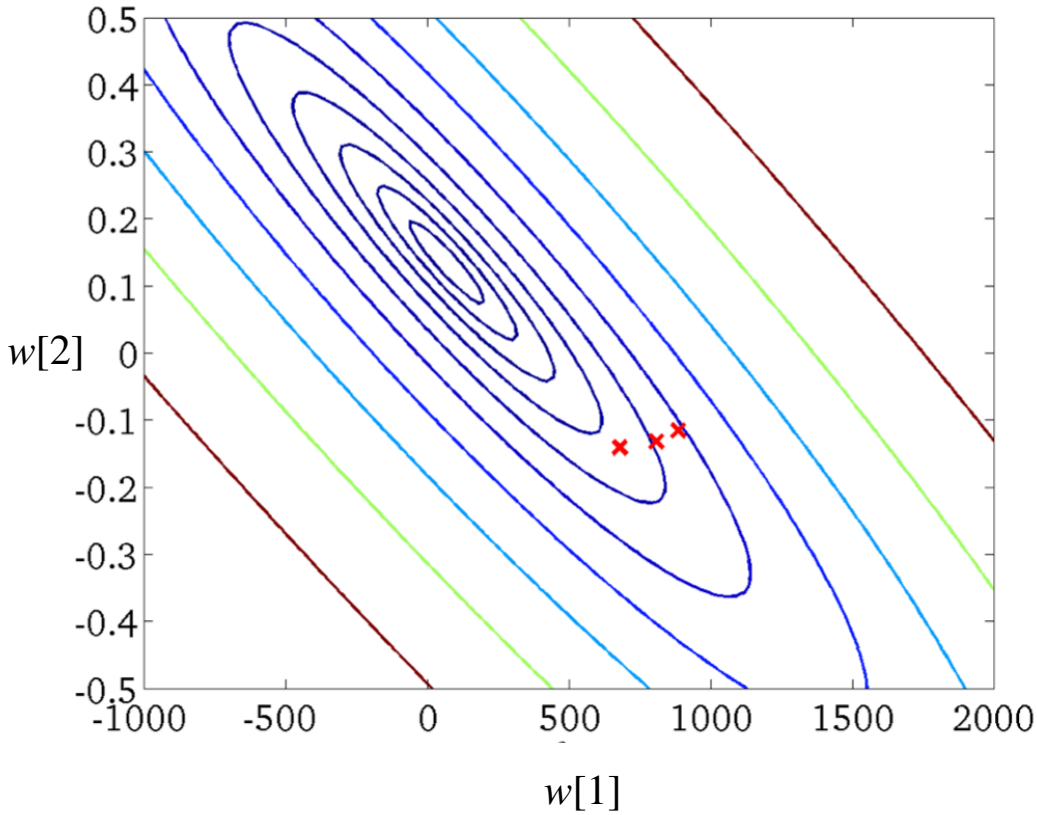


GD dynamics in the Parameter space

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

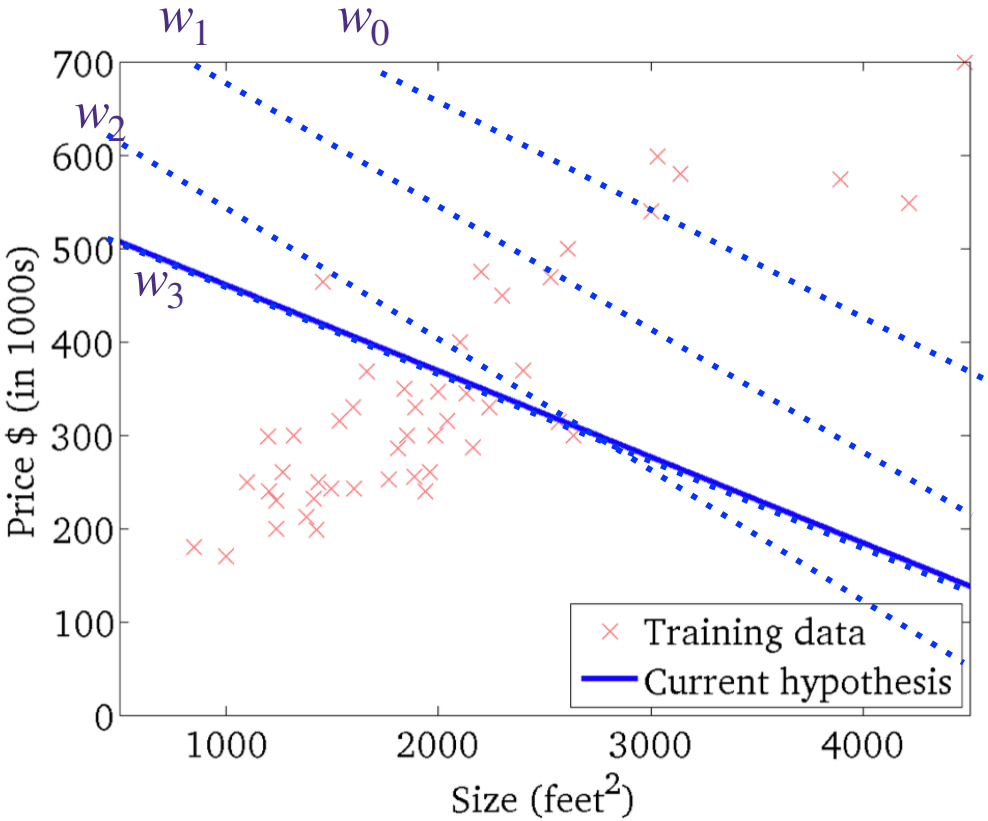


Evolution of the predictor

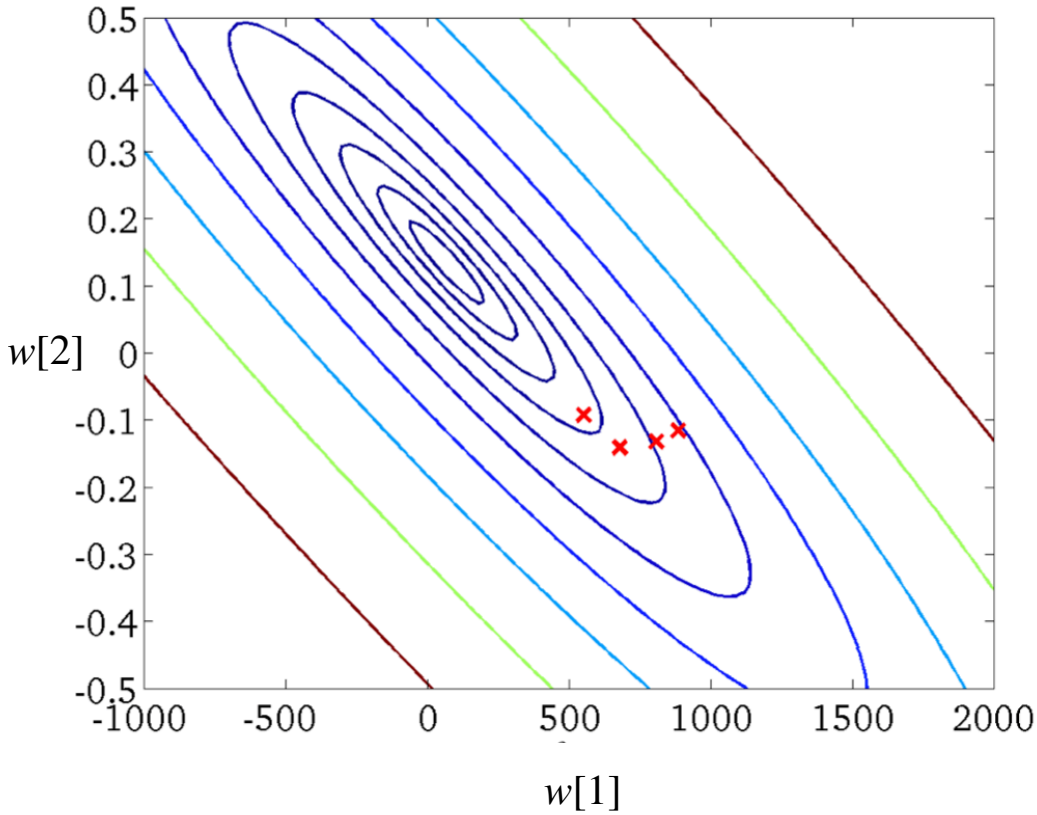


GD dynamics in the Parameter space

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

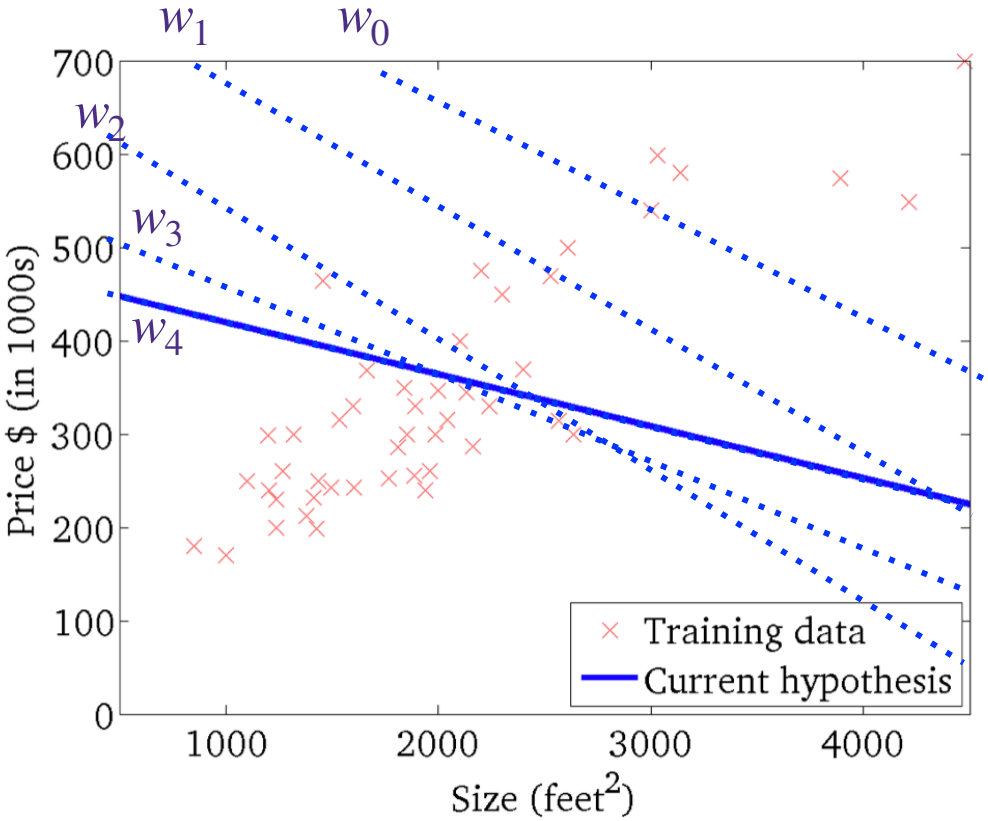


Evolution of the predictor

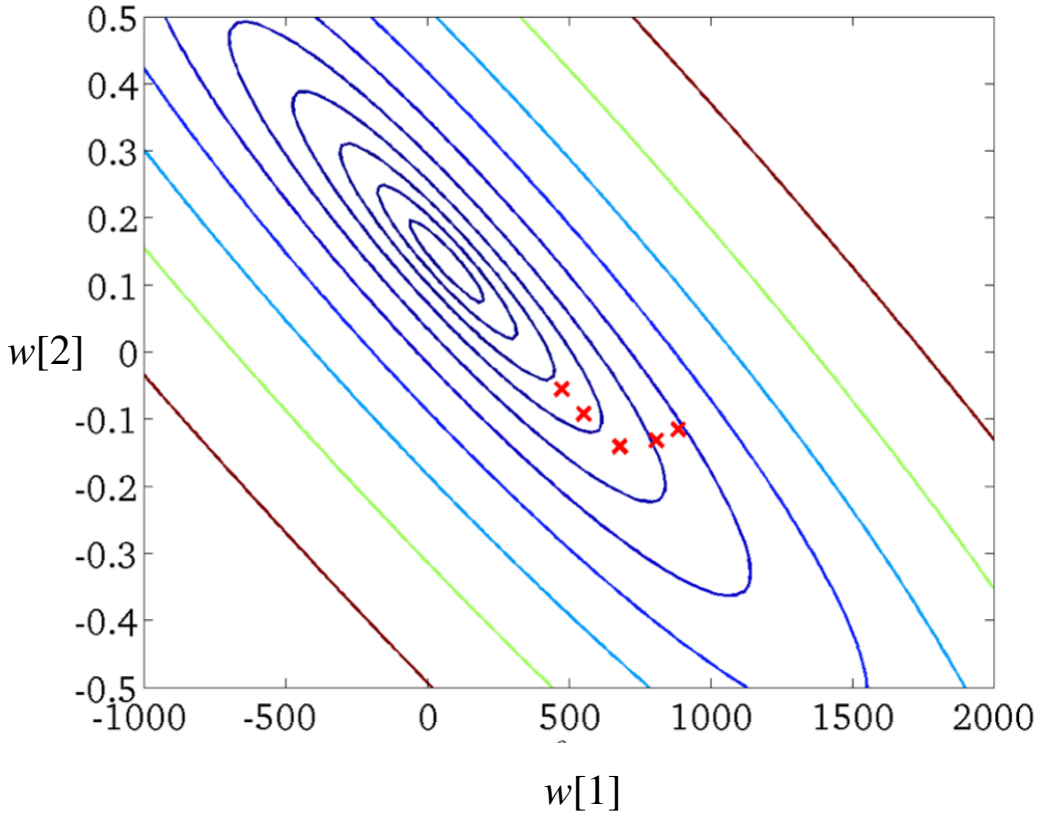


GD dynamics in the Parameter space

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

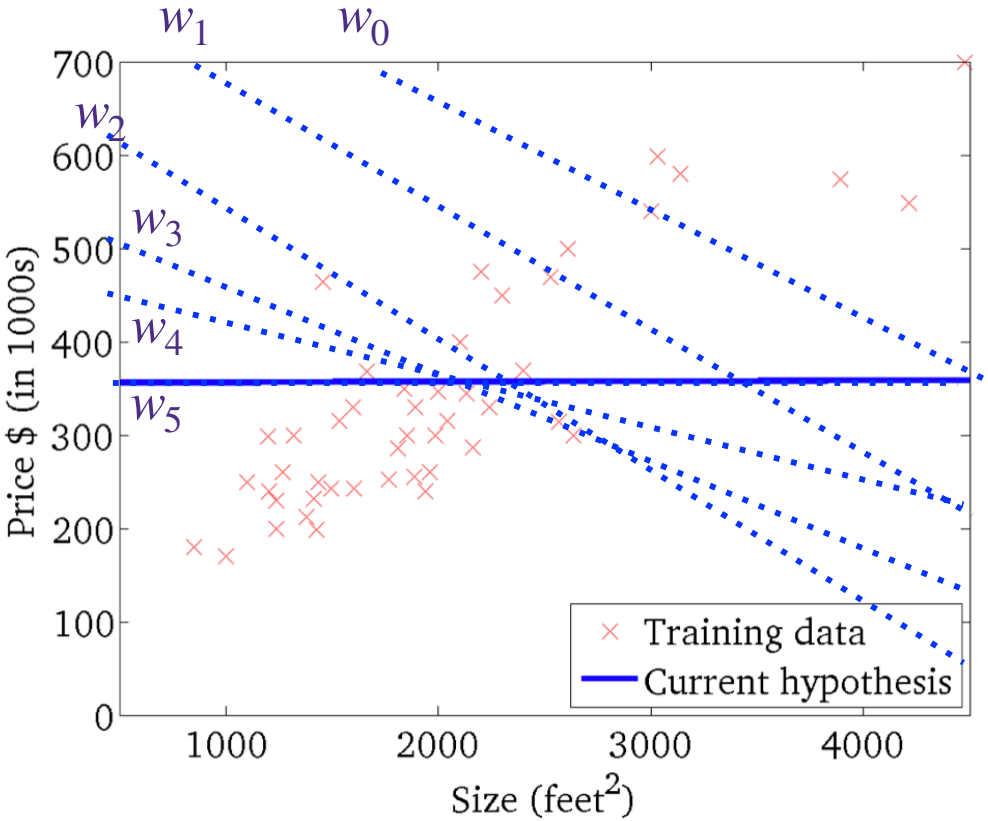


Evolution of the predictor

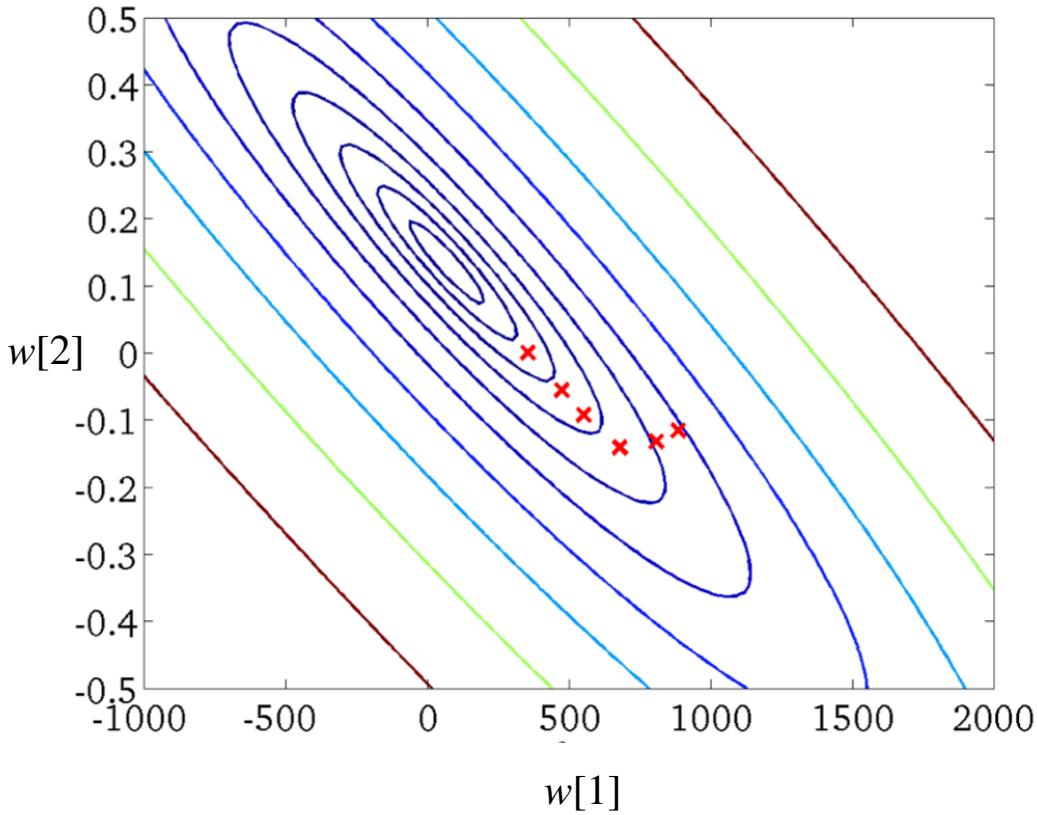


GD dynamics in the Parameter space

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

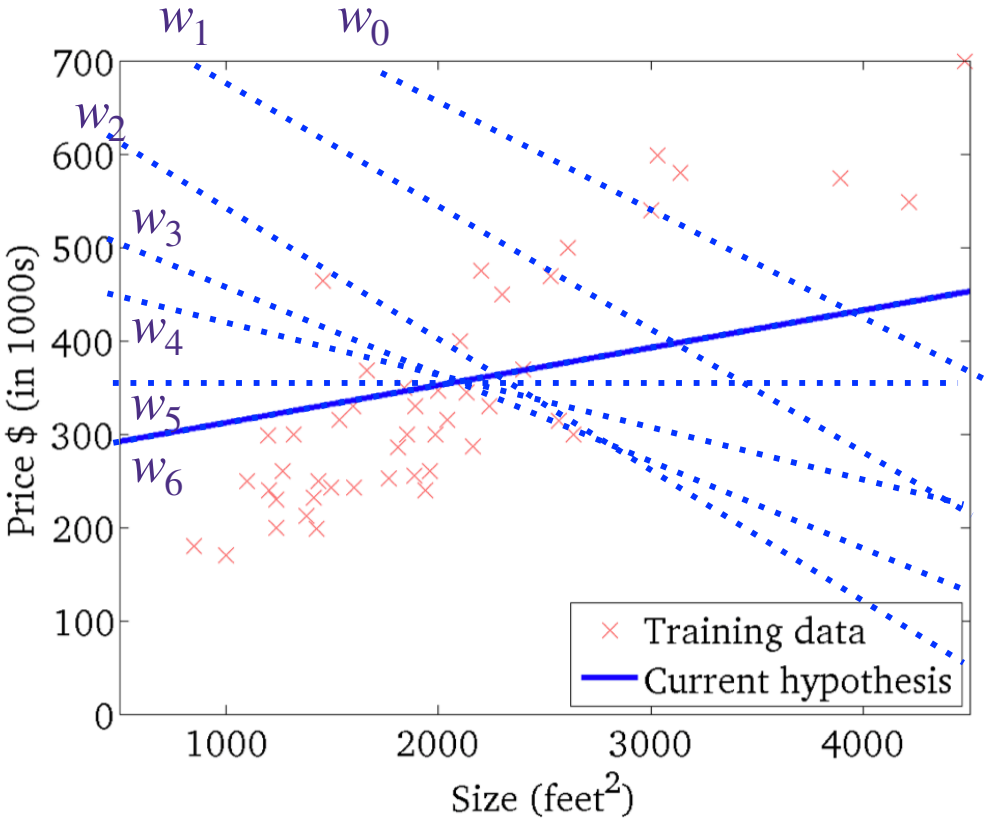


Evolution of the predictor

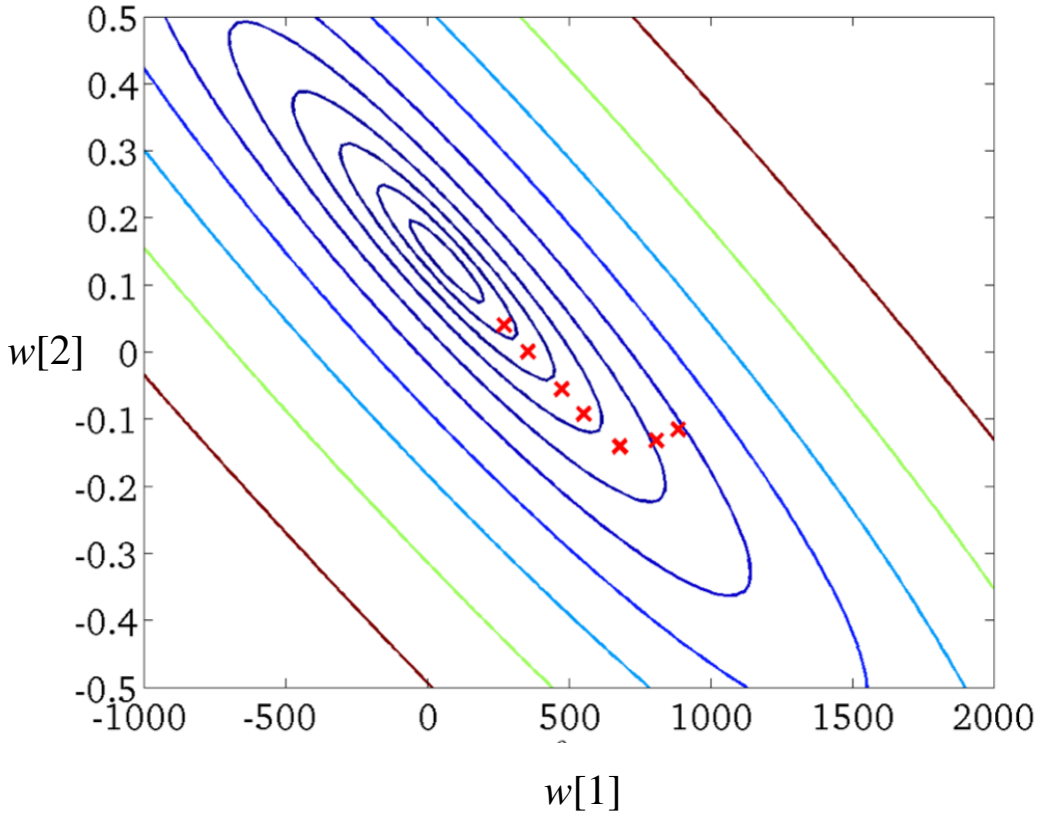


GD dynamics in the Parameter space

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

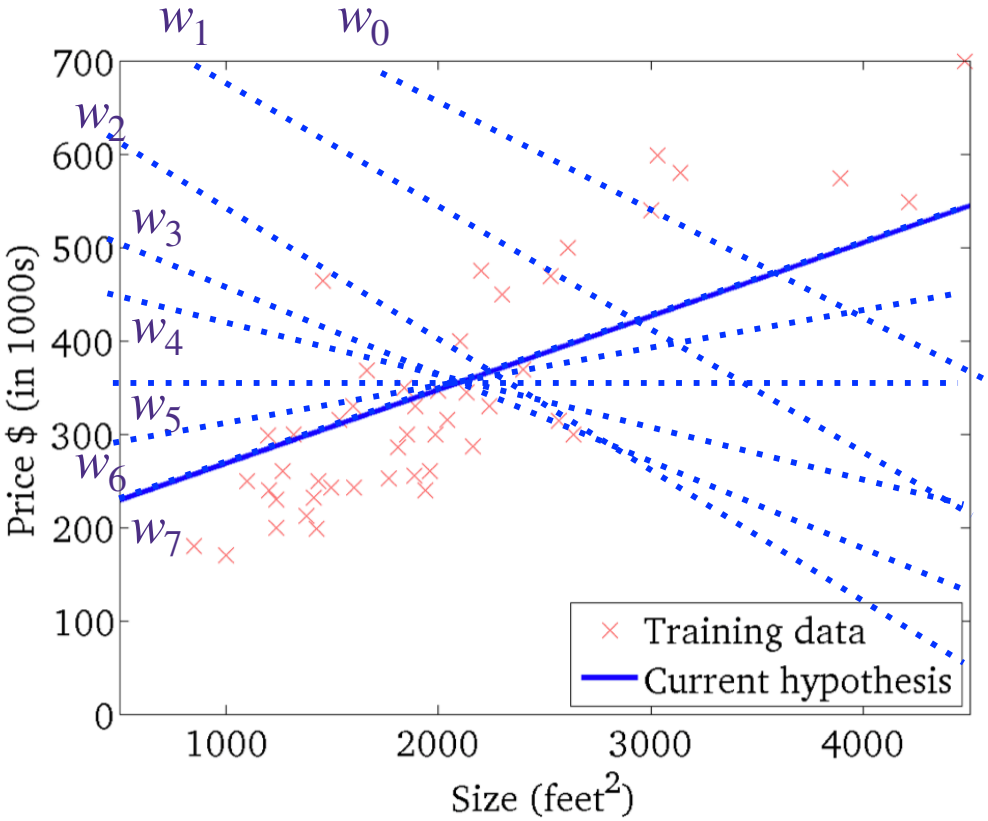


Evolution of the predictor

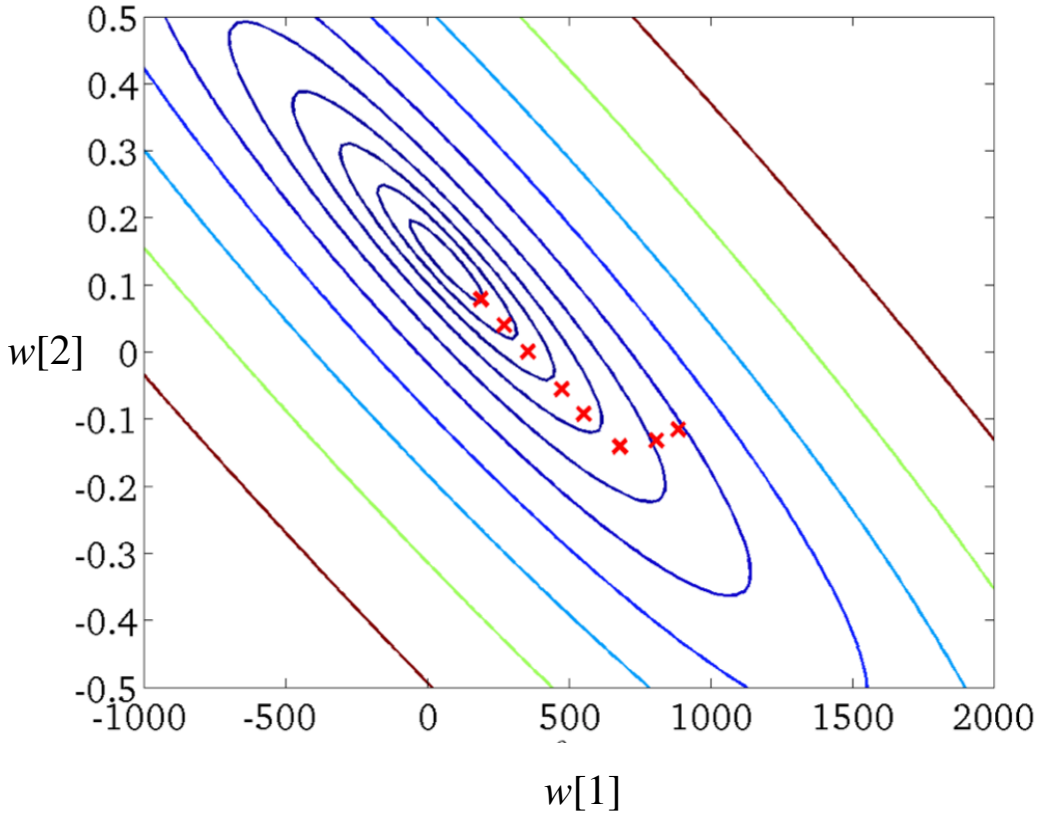


GD dynamics in the Parameter space

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



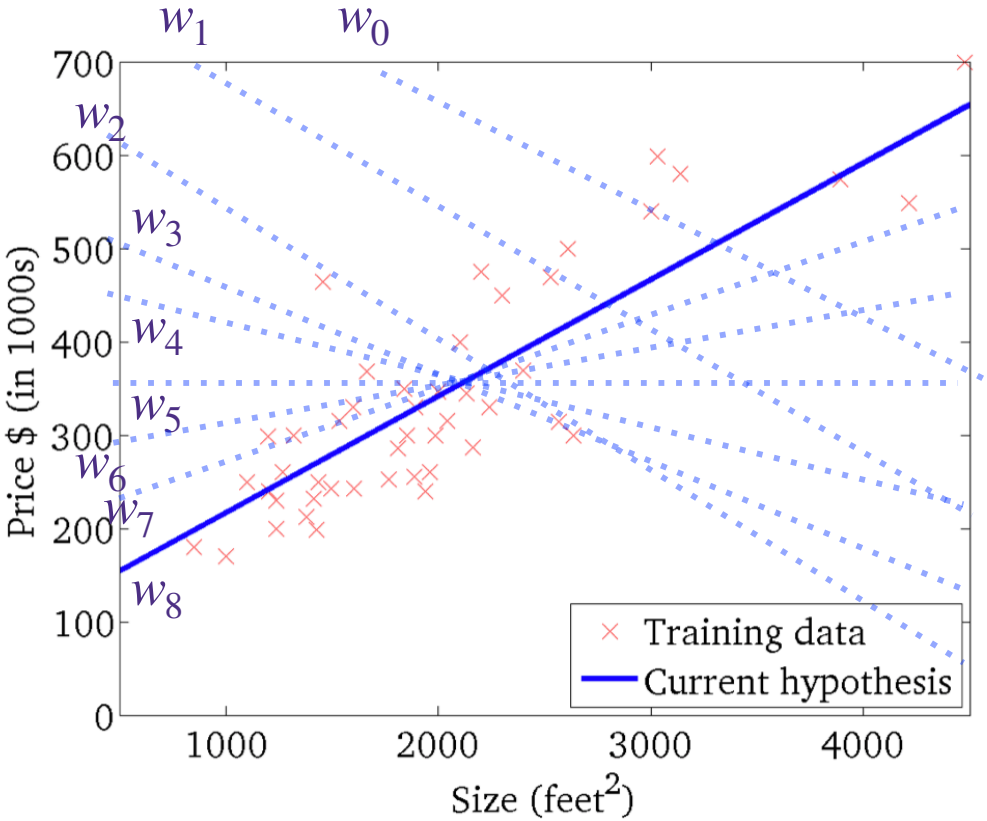
Evolution of the predictor



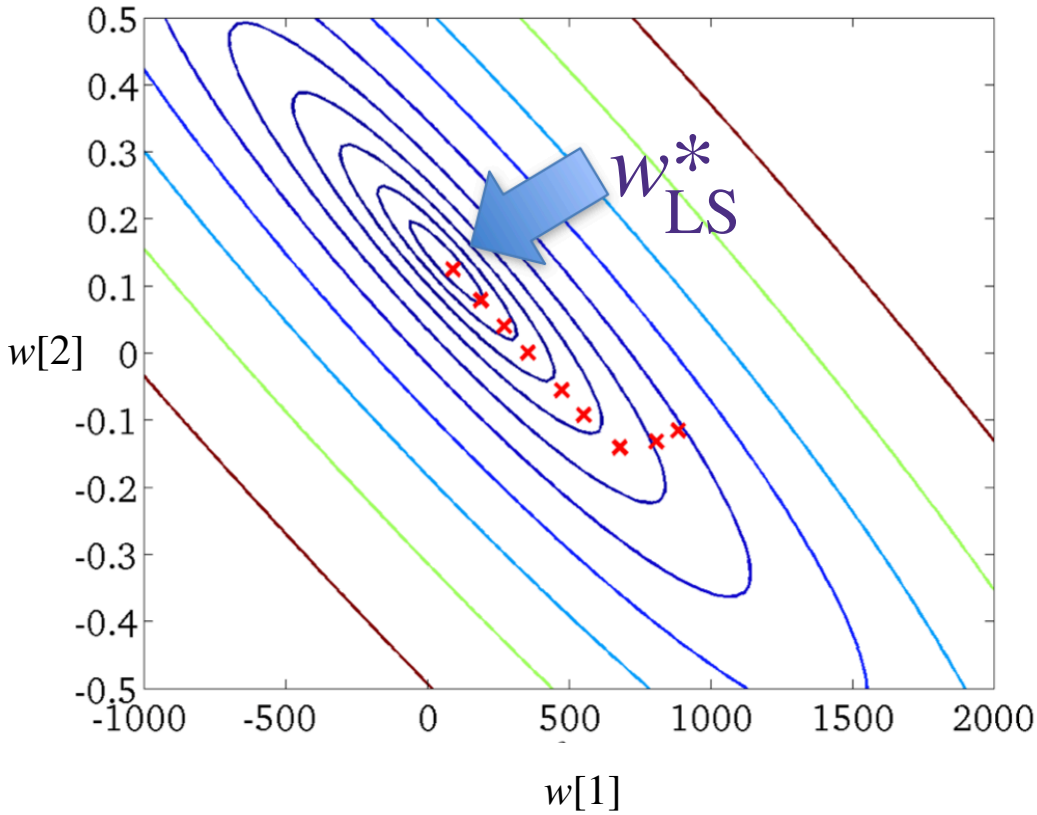
GD dynamics in the Parameter space

$X^T X = \sum_{i=1}^{n} x_i x_i^T$

- $w_0 = (900, -0.1)$
- **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 Practicalities

Practicalities

- How to initialize w_0?
  - Usually pick something at random
  - or if you have a good guess start there
- How to choose η?
  - Step size matters!
    - What happens if it is too small?
    - What happens if it is too large?
  - How to choose?
    - Special case: Solve for optimal
    - General case: Hyperparameter tuning (another one???)
- When to stop?
  - Stop when convergence is reached
  - Or stop after some fixed number of iterations (also hyperparameter 😔)

Gradient Descent Algorithm

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

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

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

$$= y^T y - 2 y^T X w + w^T X^T X w + \lambda w^T w$$

$$= y^T y - 2 y^T X w + w^T (X^T X + \lambda I) w$$

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

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

$$= \underbrace{(1 - \eta\lambda) w_t}_{\text{Shrinkage}} + \underbrace{\eta X^T (y - X w_t)}_{\text{LS gradient}}$$

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

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

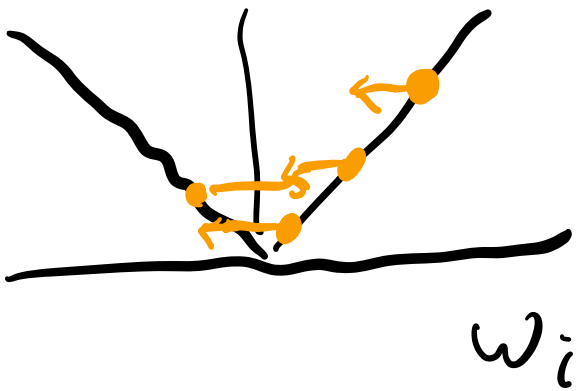
$$\nabla f(w_t) = -\mathbf{X}^T(\mathbf{y} - \mathbf{X}w_t) + \lambda w_t$$

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

Gradient descent for Lasso regression

$\|w\|_1 = \sum_{j=1}^{d} |w_j|$ $\quad [\nabla_w \|w\|_1]_i = \frac{\partial}{\partial w_i} \sum_{j=1}^{d} |w_j| = \frac{\partial |w_i|}{\partial w_i} = \text{sign}(w_i)$

- 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

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

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

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

doesn't shrink with 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

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

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

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

Gradient Descent Demo

Stochastic Gradient Descent

Machine Learning Problems

$\ell_i(w) = (y_i - x_i^T w)$

- **Given data:**

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

- **Learning a model's parameters:** $\sum_{i=1}^{n} \ell_i(w)$

$\ell(w) = \frac{1}{n} \sum_{i=1}^{n} \ell_i(w)$

Gradient Descent:

$$w_{t+1} = w_t - \eta \nabla_w \left(\frac{1}{n} \sum_{i=1}^{n} \ell_i(w) \right) \Big|_{w=w_t}$$

Machine Learning Problems

- **Given data:**

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

- **Learning a model's parameters:** $\sum_{i=1}^{n} \ell_i(w)$

Gradient Descent:

$$w_{t+1} = w_t - \eta \nabla_w \left(\frac{1}{n} \sum_{i=1}^{n} \ell_i(w) \right) \Big|_{w=w_t}$$

Stochastic Gradient Descent:

$$w_{t+1} = w_t - \eta \nabla_w \ell_{I_t}(w) \Big|_{w=w_t}$$

I_t drawn uniform at random from $\{1, \ldots, n\}$

$$\mathbb{E}[\nabla \ell_{I_t}(w)] = \sum_{i=1}^{n} P(I_t = i) \nabla \ell_i(w) = \frac{1}{n} \sum_{i=1}^{n} \nabla \ell_i(w) = \nabla \ell(w)$$

Machine Learning Problems

- **Learning a model's parameters:** $\sum_{i=1}^{n} \ell_i(w)$

Stochastic Gradient Descent:

$$w_{t+1} = w_t - \eta \nabla_w \ell_{I_t}(w)\Big|_{w=w_t}$$

I_t drawn uniform at random from $\{1, \ldots, n\}$

$$\mathbb{E}[\nabla \ell_{I_t}(w)] = \nabla \ell(w)$$

Stochastic Gradient Descent

$w_* = \arg\min_w \ell(w)$

Theorem

Let $w_{t+1} = w_t - \eta \nabla_w \ell_{I_t}(w)\Big|_{w=w_t}$ I_t drawn uniform at random from $\{1, \dots, n\}$ so that

$$\mathbb{E}\big[\nabla \ell_{I_t}(w)\big] = \frac{1}{n}\sum_{i=1}^{n} \nabla \ell_i(w) =: \nabla \ell(w)$$

If $\|w_* - w_0\|_2^2 \le R$ and $\sup_w \max_i \|\nabla \ell_i(w)\|_2 \le G$ then

$$\mathbb{E}[\ell(\bar{w}) - \ell(w_*)] \le \frac{R}{2T\eta} + \frac{\eta G}{2} \le \sqrt{\frac{RG}{T}} = \varepsilon \qquad \eta = \sqrt{\frac{R}{GT}}$$

$O(\sqrt{1/T})$ $T \sim \varepsilon^{-2}$

$$\bar{w} = \frac{1}{T}\sum_{t=1}^{T} w_t$$

(In practice use last iterate)

Stochastic Gradient Descent

Proof

$$\mathbb{E}[||w_{t+1} - w_*||_2^2] = \mathbb{E}[||w_t - \eta \nabla \ell_{I_t}(w_t) - w_*||_2^2]$$

Stochastic Gradient Descent

convexity

$f(y) \geq f(x) + \nabla f(x)^T (y - x)$

Proof

$$\mathbb{E}[||w_{t+1} - w_*||_2^2] = \mathbb{E}[||w_t - \eta \nabla \ell_{I_t}(w_t) - w_*||_2^2]$$

$$= \mathbb{E}[||w_t - w_*||_2^2] - 2\eta \mathbb{E}[\nabla \ell_{I_t}(w_t)^T (w_t - w_*)] + \eta^2 \mathbb{E}[||\nabla \ell_{I_t}(w_t)||_2^2]$$

$$\leq \mathbb{E}[||w_t - w_*||_2^2] - 2\eta \mathbb{E}[\ell(w_t) - \ell(w_*)] + \eta^2 G$$

$$\mathbb{E}[\nabla \ell_{I_t}(w_t)^T (w_t - w_*)] = \mathbb{E}\big[\mathbb{E}[\nabla \ell_{I_t}(w_t)^T (w_t - w_*) | I_1, w_1, \ldots, I_{t-1}, w_{t-1}]\big]$$

$$= \mathbb{E}\big[\nabla \ell(w_t)^T (w_t - w_*)\big]$$

$$\geq \mathbb{E}\big[\ell(w_t) - \ell(w_*)\big]$$

$$\mathbb{E}\left[\ell(w_t) - \ell(w_*)\right] \leq \frac{1}{2\eta}\Big(\mathbb{E}\left[\|w_t - w_*\|_2^2\right] - \mathbb{E}\left[\|w_{t+1} - w_*\|_2^2\right] + \eta^2 G\Big)$$

Stochastic Gradient Descent

$\sum_{t=1}^{T} x_t - x_{t+1} = x_1 - x_{T+1}$ telescoping sum

$$\sum_{t=1}^{T} \mathbb{E}\left[\ell(w_t) - \ell(w_*)\right] \leq \sum_{t=1}^{T} \frac{1}{2\eta}\left(\mathbb{E}\left[\|w_t - w_*\|_2^2\right] - \mathbb{E}\left[\|w_{t+1} - w_*\|_2^2\right] + \eta^2 G\right)$$

$$= \frac{1}{2\eta}\left(\mathbb{E}\left[\|w_1 - w_*\|_2^2\right] - \mathbb{E}\left[\|w_{T+1} - w_*\|_2^2\right] + T\eta^2 G\right)$$

$$\leq \frac{R}{2\eta} + \frac{T\eta G}{2}$$

Jensen's inequality:
For any random $Z \in \mathbb{R}^d$ and convex function $\phi : \mathbb{R}^d \to \mathbb{R}$, $\phi(\mathbb{E}[Z]) \leq \mathbb{E}[\phi(Z)]$

$$\mathbb{E}\left[\ell(\bar{w}) - \ell(w_*)\right] = \mathbb{E}\left[\ell\left(\frac{1}{T}\sum_{t=1}^{T} w_t\right) - \ell(w_*)\right] \qquad \bar{w} = \frac{1}{T}\sum_{t=1}^{T} w_t$$

$$\leq \frac{1}{T}\sum_{t=1}^{T} \mathbb{E}\left[\ell(w_t) - \ell(w_*)\right]$$

$$\leq \frac{R}{2T\eta} + \frac{\eta G}{2}$$

Stochastic Gradient Descent

Theorem

Let $w_{t+1} = w_t - \eta \nabla_w \ell_{I_t}(w)\Big|_{w=w_t}$ I_t drawn uniform at random from $\{1, \dots, n\}$ so that

$$\mathbb{E}\big[\nabla \ell_{I_t}(w)\big] = \frac{1}{n}\sum_{i=1}^{n} \nabla \ell_i(w) =: \nabla \ell(w)$$

If $\|w_1 - w_0\|_2^2 \le R$ and $\sup_w \max_i \|\nabla \ell_i(w)\|_2 \le G$ then

$$\mathbb{E}[\ell(\bar{w}) - \ell(w_*)] \le \frac{R}{2T\eta} + \frac{\eta G}{2} \le \sqrt{\frac{RG}{T}} =: \varepsilon \qquad \eta = \sqrt{\frac{R}{GT}}$$

Per iteration: d flops

Iterations T

$T \sim \varepsilon^{-2}$

Overall: $Td = d\varepsilon^{-2}$

$$\bar{w} = \frac{1}{T}\sum_{t=1}^{T} w_t$$

(In practice use last iterate)

Compare w/ closed form solution

$$X = \begin{bmatrix} -x_1^T- \\ \vdots \\ -x_n^T- \end{bmatrix} \quad y = \begin{bmatrix} y_1 \\ \vdots \\ y_n \end{bmatrix}$$

$$\hat{w} = (X^T X)^{-1} X^T y$$

To compute $X^T X$: nd^2 flops

$$\sum_{i=1}^{n} x_i x_i^T$$

To compute $(X^T X)^{-1}$: d^3 flops

Overall: $nd^2 + d^3$

Mini-batch SGD

- Instead of one iterate, average B stochastic gradient together

- Advantages:
  - Smaller variance: the variance of the stochastic gradient is smaller by a factor of $1/\sqrt{B}$ 1/B
  - Parallelization: each gradient in the mini-batch can be computed in parallel

- If you have regularizer, $\frac{1}{n}\sum_{i=1}^{n} \ell_i(w) + r(w)$, then update with the stochastic gradient of the loss and gradient of the regularizer