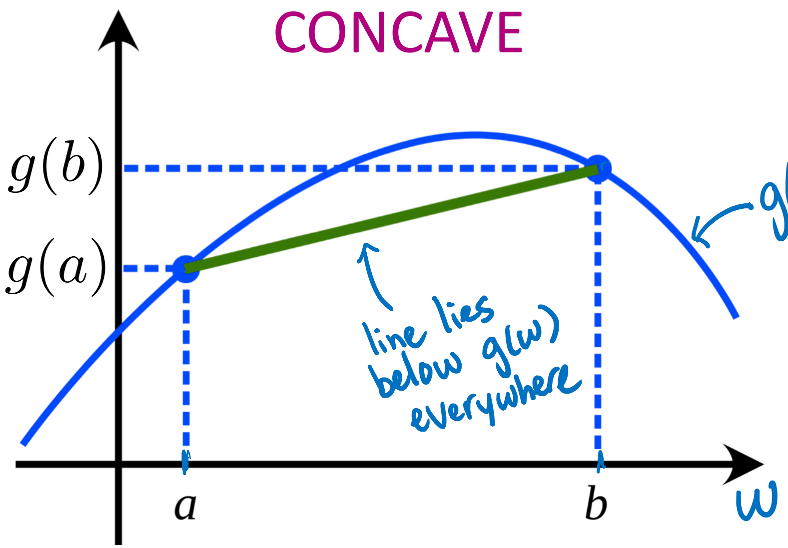
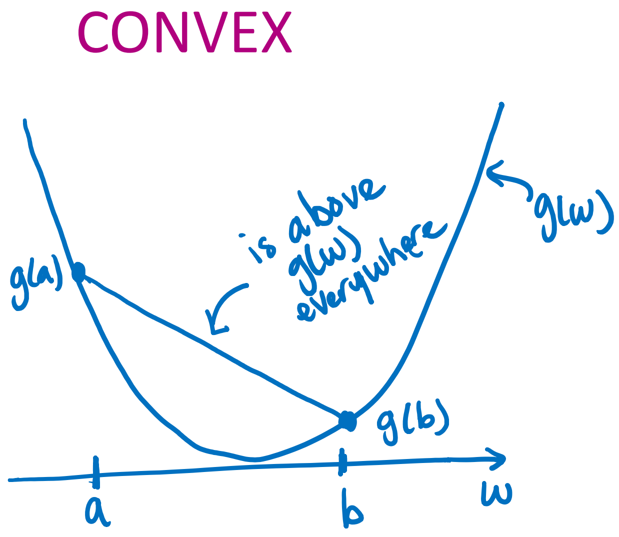
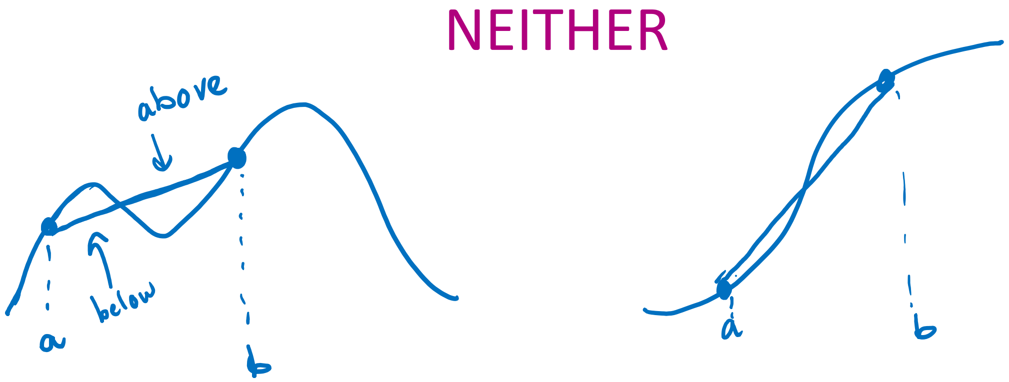
Gradient descent

Convex/concave functions

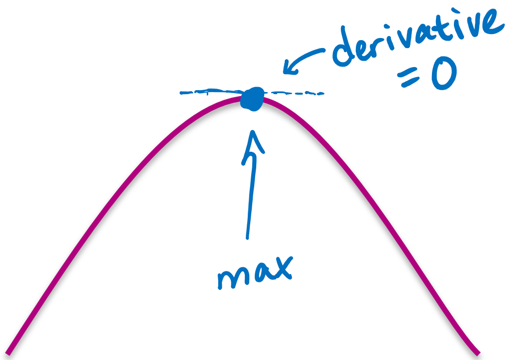






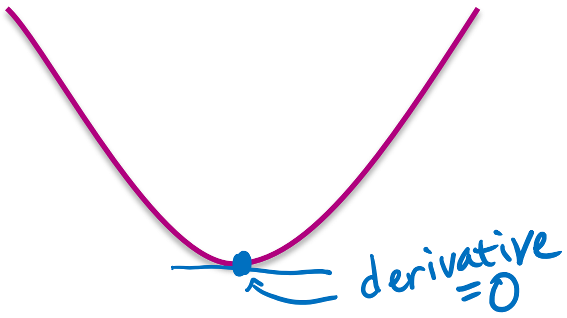
Finding the max or min analytically

CONCAVE



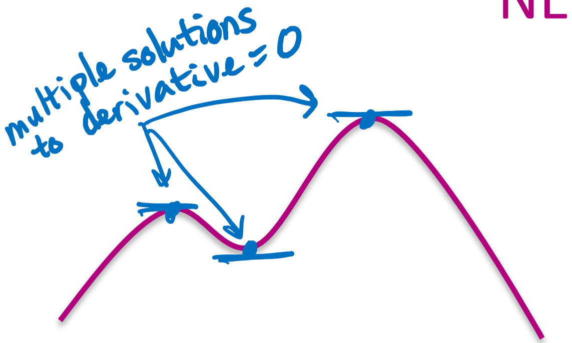
$\max_w g(w)$

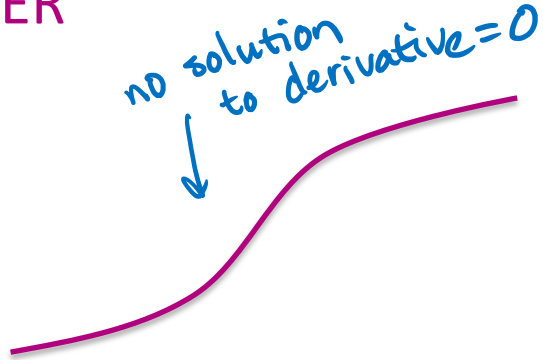
CONVEX



$\min_w g(w)$

NEITHER





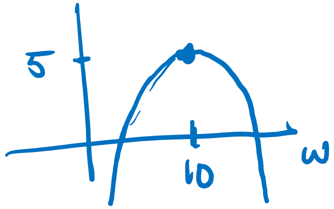
Example:

$g(w) = 5-(w-10)^2$

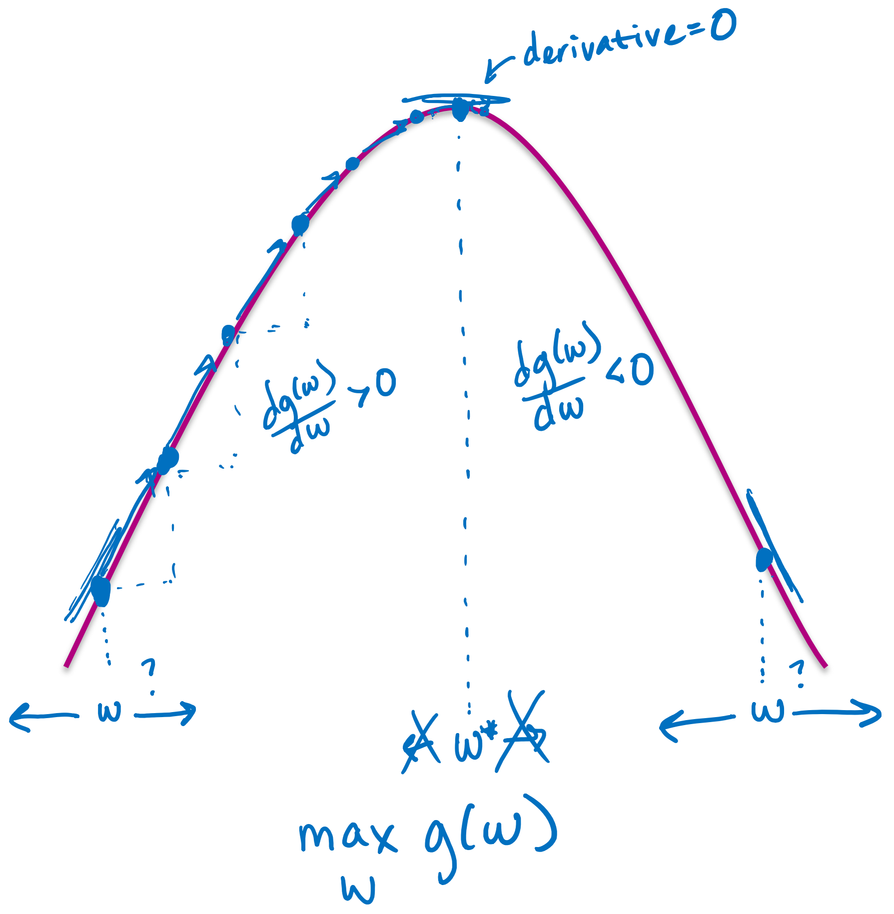
$$\frac{dg(w)}{dw} = 0 - 2(w-10)^1 \cdot 1$$
$$= -2w + 20$$

set derivate = 0:

$$-2w+20=0$$
$$w = 10$$



Finding the max via hill climbing



How do we know whether to move w to right or left?
(inc. or dec. the value of w?)

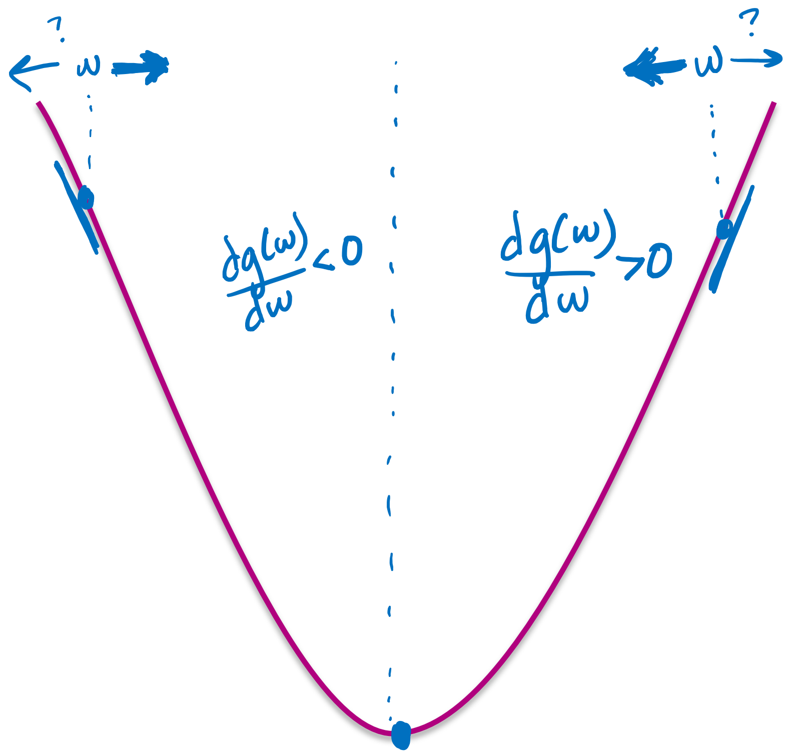
while not converged

$$w^{(t+1)} \leftarrow w^{(t)} + \eta \frac{dg(w)}{dw}$$

iteration t

stepsize

Finding the min via hill descent



$\min_w g(w)$

when derivative is positive, we want to decrease w
and when derivative is negative, we want to increase w

Algorithm:

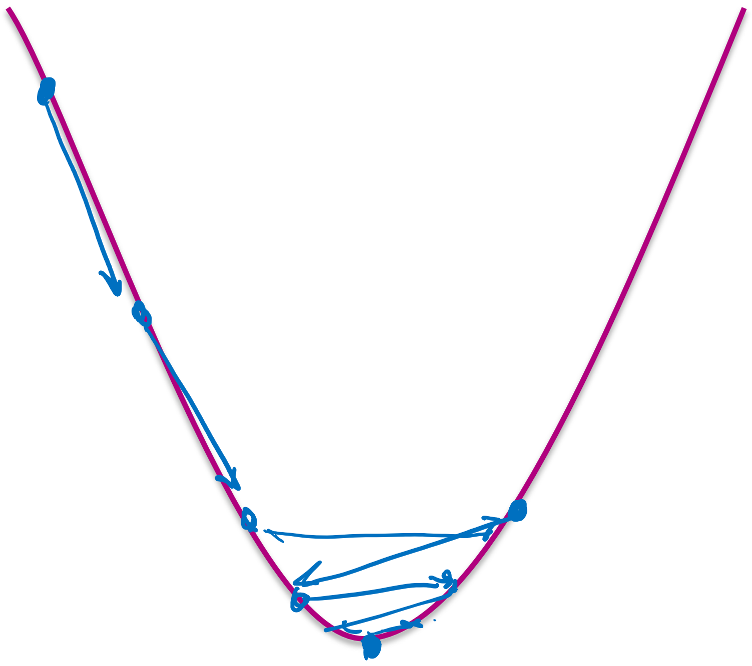
while not converged

$$w^{(t+1)} \leftarrow w^{(t)} - \eta \left.\frac{dg}{dw}\right|_{w^{(t)}}$$

Choosing the stepsize—

Fixed stepsize

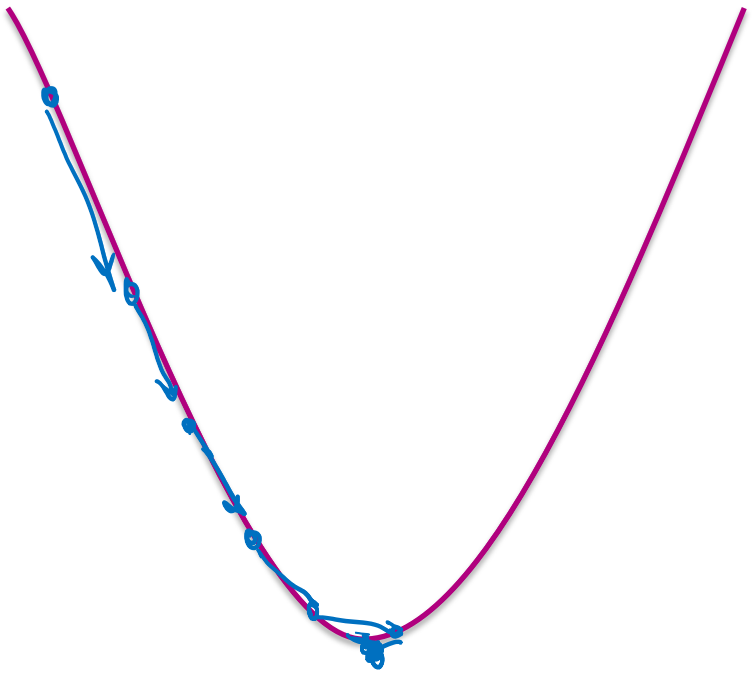
η



$\eta = 0.1$

Choosing the stepsize—
Decreasing stepsize

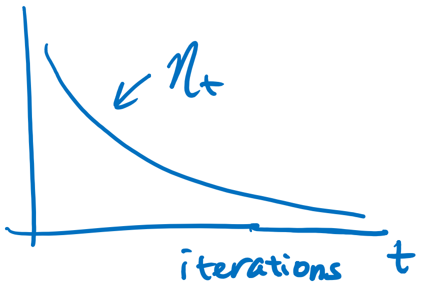
or stepsize schedule



Common choices:

$$\eta_t = \frac{\alpha}{t}$$

$$\eta_t = \frac{\alpha}{\sqrt{t}}$$



Convergence criteria

For convex functions,

optimum occurs when

$$\frac{dg(w)}{dw} = 0$$

In practice, stop when

$$\left|\frac{dg(w)}{dw}\right| < \epsilon$$

threshold to be set

Algorithm:

while not converged

$$w^{(t+1)} \leftarrow w^{(t)} - \eta \left.\frac{dg}{dw}\right|_{w^{(t)}}$$

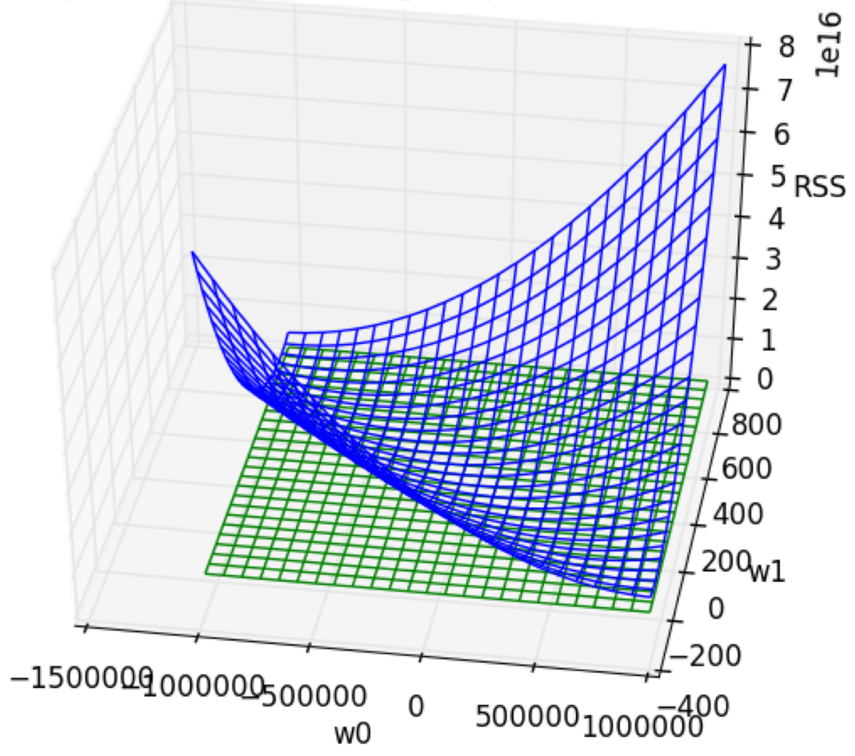
Moving to higher dimensions

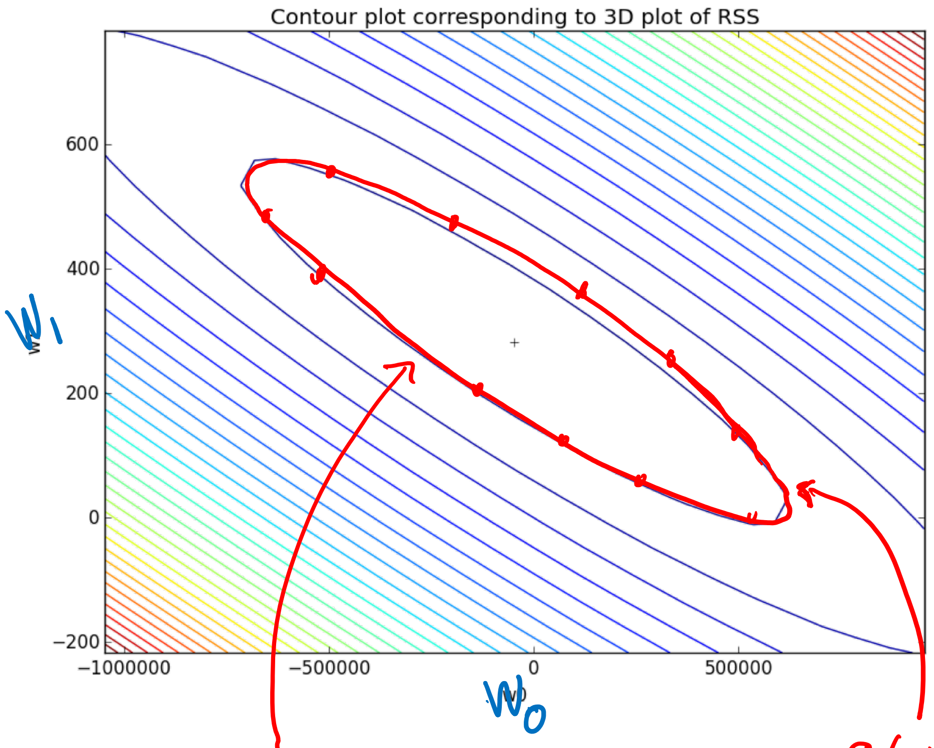
OPTIONAL

Note: We use the Optional tag to signify that you are not responsible for understanding the following material!

Contour plots

bird's eye view

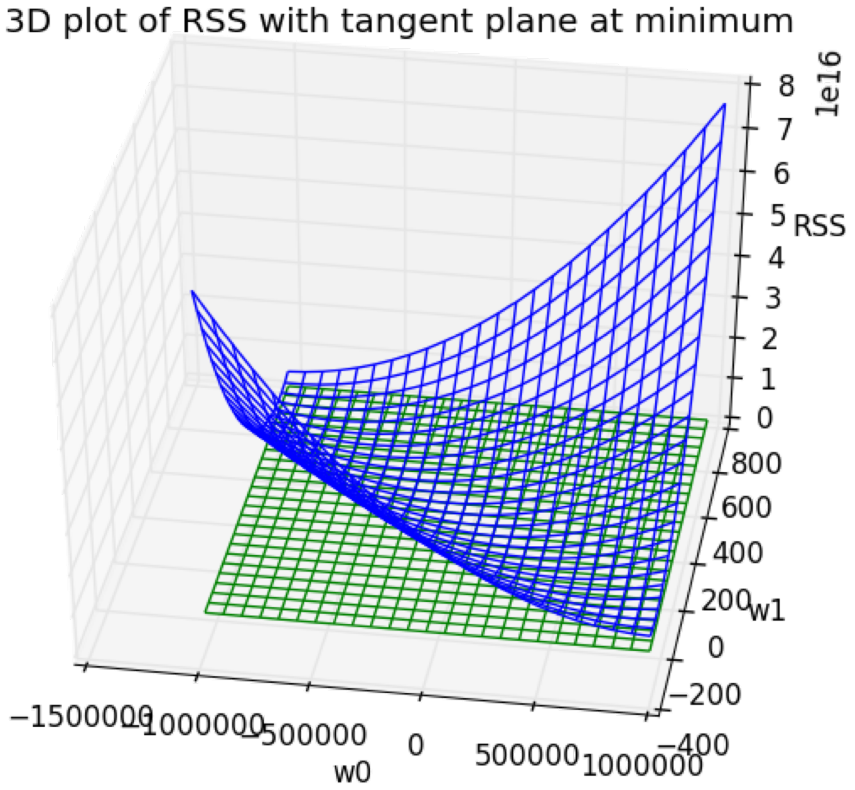




a slice of the 3D surface

$g(w_0, w_1)$

Moving to multiple dimensions: Gradients



$$\nabla g(\mathbf{w}) = \begin{bmatrix} \frac{\partial g}{\partial w_0} \\ \frac{\partial g}{\partial w_1} \\ \vdots \\ \frac{\partial g}{\partial w_p} \end{bmatrix}$$

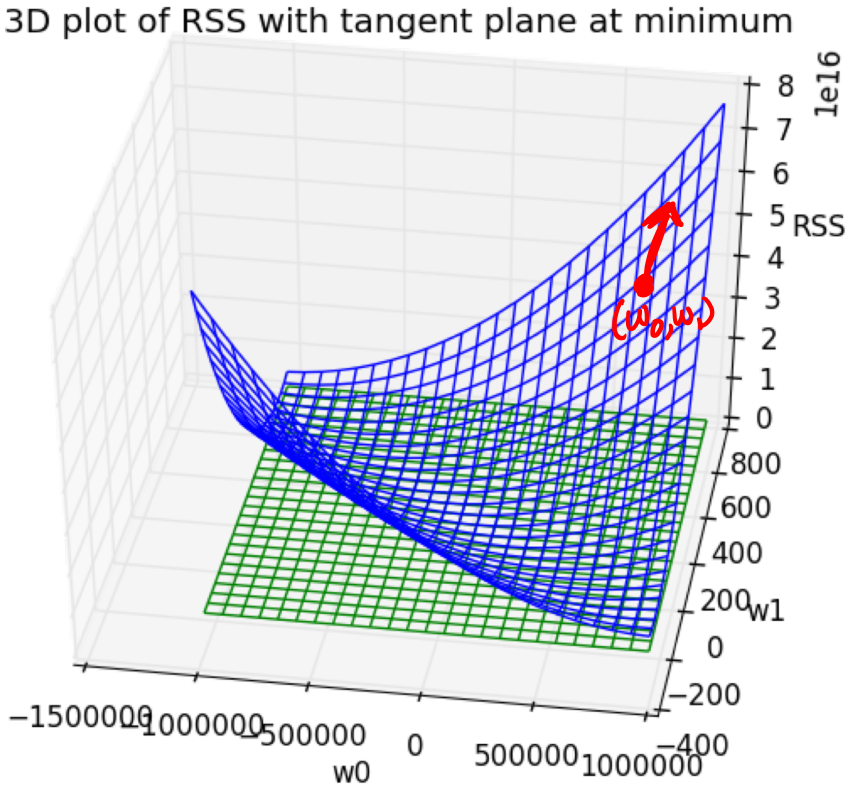
gradient

$[w_0, w_1, \ldots, w_p]$

$(p+1)$-dimensional vector

partial derivative is like a derivate with respect to w_1, treating all other variables as constants

Gradient example



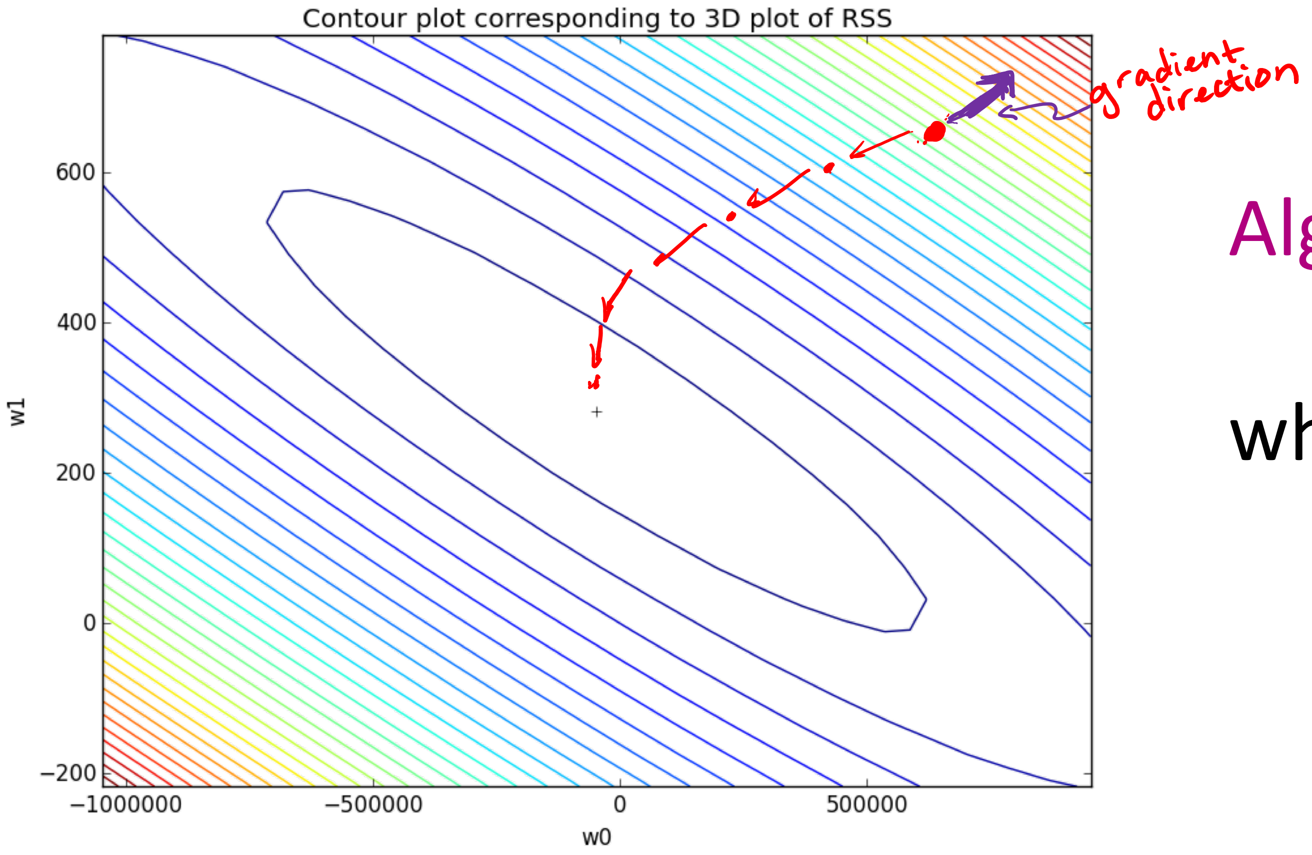
$$g(\mathbf{w}) = 5w_0 + 10\underline{w_0 w_1} + 2w_1^2$$

$$\frac{\partial g}{\partial w_0} = 5 + 10w_1$$

$$\frac{\partial g}{\partial w_1} = 10w_0 + 4w_1$$

$$\nabla g(\mathbf{w}) = \begin{bmatrix} 5 + 10w_1 \\ 10w_0 + 4w_1 \end{bmatrix}$$

Gradient descent



Algorithm:

while not converged

$w^{(t+1)} \leftarrow w^{(t)} - \eta \nabla g(w^{(t)})$

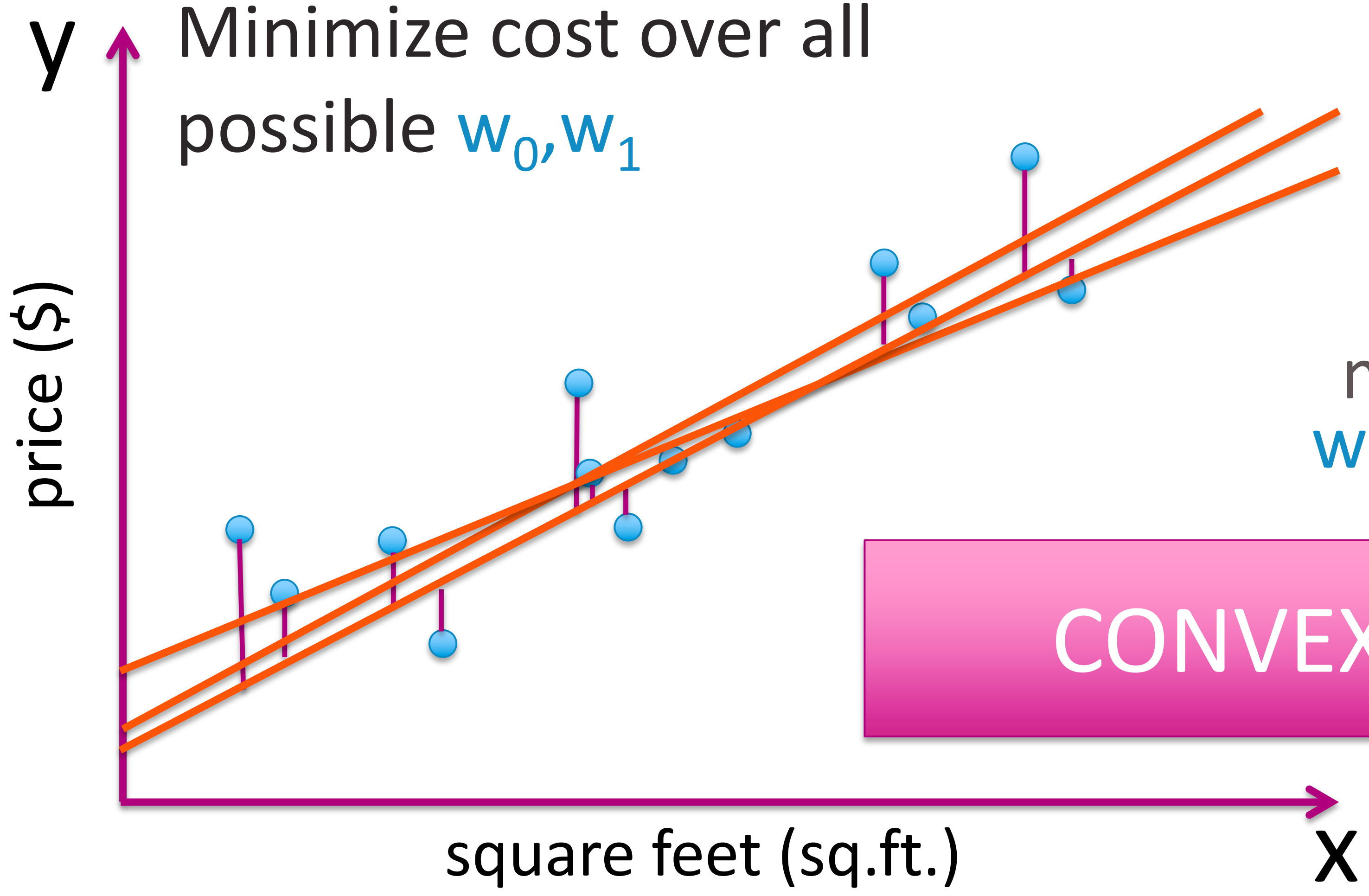
Convergence: $||\nabla g(w)|| < \epsilon$

Gradient Descent for Linear Regression

Note: We use the Cap to point out that the following section contains advanced topics, passed the level we expect from the class.

OPTIONAL

Find "best" line



Minimize cost over all possible w_0, w_1

$$\min_{w_0, w_1} \sum_{i=1}^{N} (y_i - [w_0 + w_1 x_i])^2$$

CONVEX

$\Rightarrow$ solution is unique

+ gradient descent alg. will converge to minimum

Compute the gradient

$$\text{RSS}(w_0,w_1) = \sum_{i=1}^{N}(y_i-[w_0+w_1x_i])^2$$

Aside:

$$\frac{d}{dw}\sum_{i=1}^{N} g_i(w) = \frac{d}{dw}\left(g_1(w) + g_2(w) + \ldots g_N(w)\right)$$

$$= \frac{d}{dw}g_1(w) + \frac{d}{dw}g_2(w) + \ldots \frac{d}{dw}g_N(w)$$

$$= \sum_{i=1}^{N}\frac{d}{dw}g_i(w)$$

In our case

$g_i(w) = (y_i - [w_0+w_1x_i])^2$

$$\frac{\partial\, RSS(w)}{\partial w_0} = \sum_{i=1}^{N}\frac{\partial}{\partial w_0}(y_i-[w_0+w_1x_i])^2$$

same for w_1

Compute the gradient

$$\text{RSS}(w_0,w_1) = \sum_{i=1}^{N}(y_i-[w_0+w_1x_i])^2$$

Taking the derivative w.r.t. w_0

$$\sum_{i=1}^{N} 2\,(y_i-[w_0+w_1x_i])^1 \cdot (-1)$$

$$= -2\sum_{i=1}^{N}(y_i-[w_0+w_1x_i])$$

Compute the gradient

$$\mathrm{RSS}(w_0,w_1) = \sum_{i=1}^{N}(y_i-[w_0+w_1x_i])^2$$

Taking the derivative w.r.t. w_1

$$\sum_{i=1}^{N} 2(y_i-[w_0+w_1x_i])^1 \cdot (-x_i)$$

$$= -2\sum_{i=1}^{N}(y_i-[w_0+w_1x_i])x_i$$

Compute the gradient

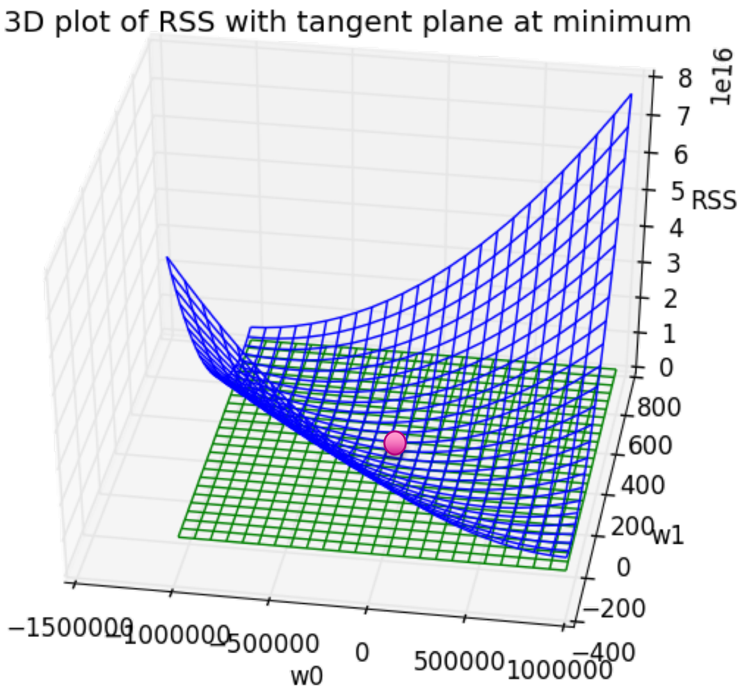
$$\text{RSS}(w_0,w_1) = \sum_{i=1}^{N}(y_i-[w_0+w_1x_i])^2$$

Putting it together:

$$\nabla\text{RSS}(w_0,w_1) = \begin{bmatrix} -2\sum_{i=1}^{N}[y_i - (w_0+w_1x_i)] \\ -2\sum_{i=1}^{N}[y_i - (w_0+w_1x_i)]x_i \end{bmatrix}$$

Approach 1: Set gradient = 0

$$\nabla RSS(w_0, w_1) = \begin{bmatrix} -2\sum_{i=1}^{N}[y_i - (w_0 + w_1 x_i)] \\ -2\sum_{i=1}^{N}[y_i - (w_0 + w_1 x_i)]x_i \end{bmatrix}$$



top term:

average house sales price

estimate of the slope

average sq. ft.

$$\hat{w}_0 = \frac{\sum_{i=1}^{N} y_i}{N} - \hat{w}_1 \frac{\sum_{i=1}^{N} x_i}{N}$$

bottom term:

$$\sum y_i x_i - \hat{w}_0 \sum x_i - \hat{w}_1 \sum x_i^2 = 0$$

plug in

$$\hat{w}_1 = \frac{\sum y_i x_i - \frac{\sum y_i \sum x_i}{N}}{\sum x_i^2 - \frac{\sum x_i \sum x_i}{N}}$$

Note:

$\sum_{i=1}^{N} y_i$

$\sum_{i=1}^{N} x_i$

$\sum_{i=1}^{N} y_i x_i$

$\sum_{i=1}^{N} x_i^2$

Approach 2: Gradient descent

Interpreting the gradient:

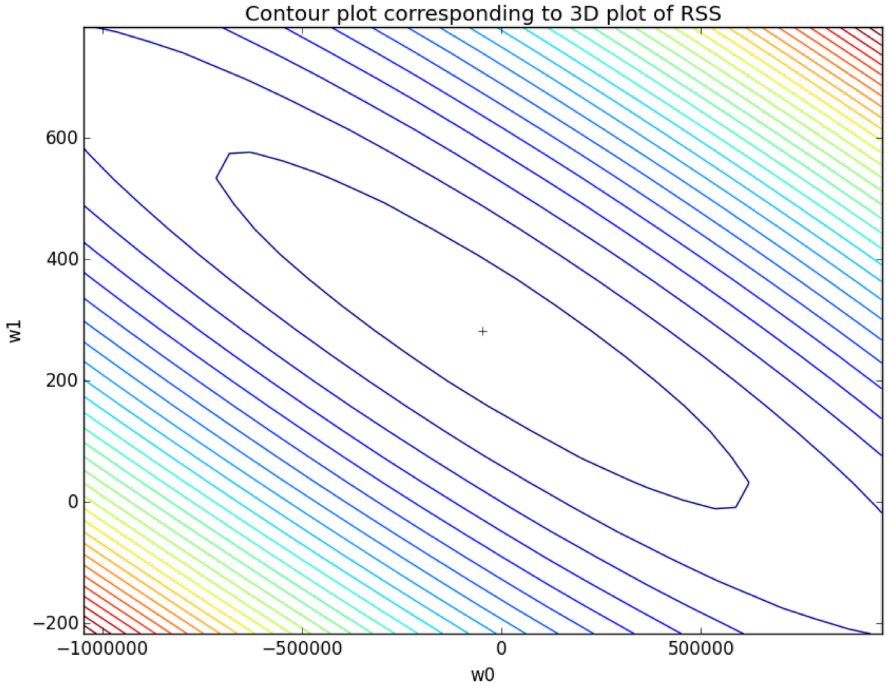
actual house sales observation

predicted value $\hat{y}_i(w_0, w_1)$

$$\nabla RSS(w_0,w_1) = \begin{bmatrix} -2\sum_{i=1}^{N}[y_i - \underline{(w_0+w_1x_i)}] \\ -2\sum_{i=1}^{N}[y_i - (w_0+w_1x_i)]x_i \end{bmatrix} = \begin{bmatrix} -2\sum_{i=1}^{N}[y_i - \hat{y}_i(w_0,w_1)] \\ -2\sum_{i=1}^{N}[y_i - \hat{y}_i(w_0,w_1)]x_i \end{bmatrix}$$

Approach 2: Gradient descent

$$\nabla RSS(w_0, w_1) = \begin{bmatrix} -2 \sum_{i=1}^{N} [y_i - \hat{y}_i(w_0, w_1)] \\ -2 \sum_{i=1}^{N} [y_i - \hat{y}_i(w_0, w_1)] x_i \end{bmatrix}$$



while not converged

$$\begin{bmatrix} w_0^{(t+1)} \\ w_1^{(t+1)} \end{bmatrix} \leftarrow \begin{bmatrix} w_0^{(t)} \\ w_1^{(t)} \end{bmatrix} + \underbrace{2\eta}_{(-2)\cdot(-\eta)} \begin{bmatrix} \sum_{i=1}^{N} [y_i - \hat{y}_i(w_0^{(t)}, w_1^{(t)})] \\ \sum_{i=1}^{N} [y_i - \hat{y}_i(w_0^{(t)}, w_1^{(t)})] x_i \end{bmatrix}$$

If overall, underpredicting $\hat{y}_i$, then $\Sigma[y_i - \hat{y}_i]$ is positive

$\rightarrow$ w_0 is going to increase

similar intuition for w_1, but multiply by x_i

Comparing the approaches

- For most ML problems, cannot solve gradient = 0
- Even if solving gradient = 0 is feasible, gradient descent can be more efficient
- Gradient descent relies on choosing stepsize and convergence criteria