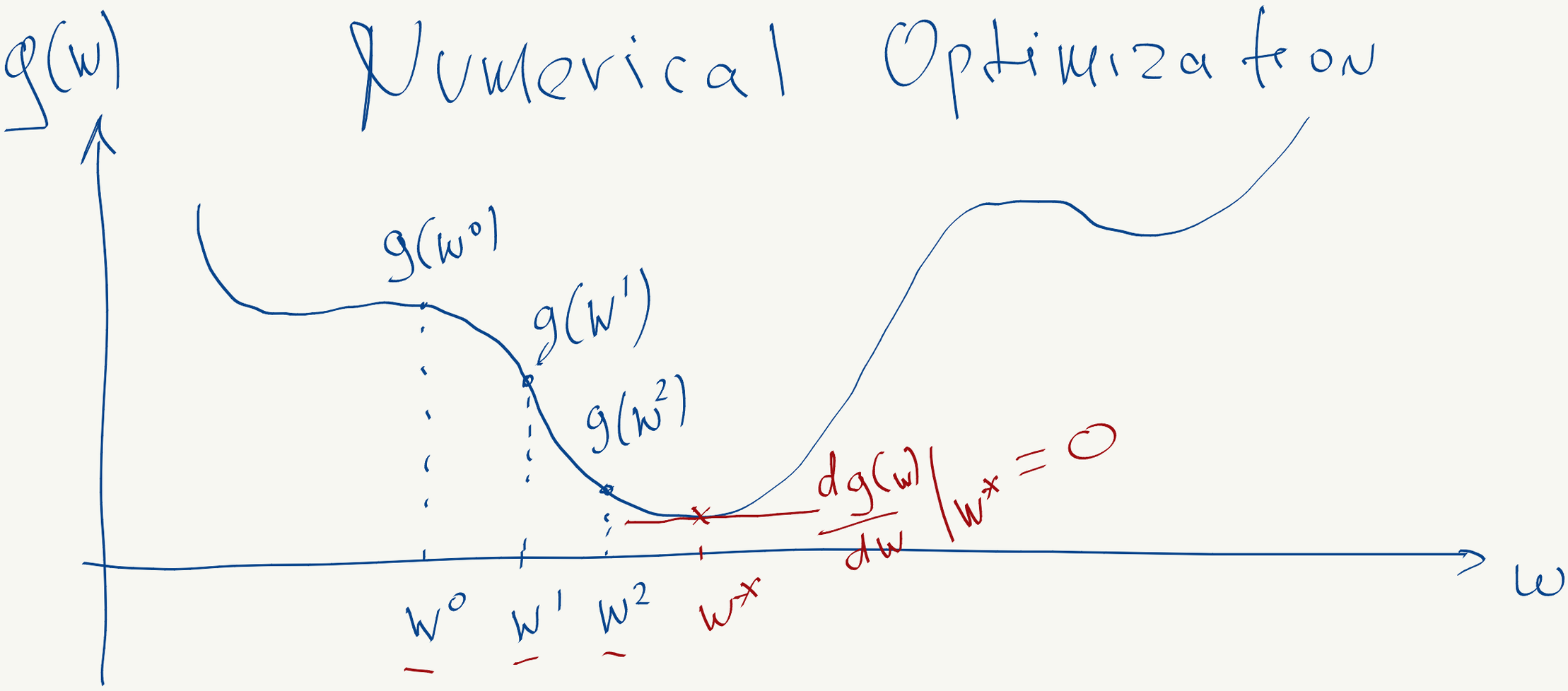
4/13/20

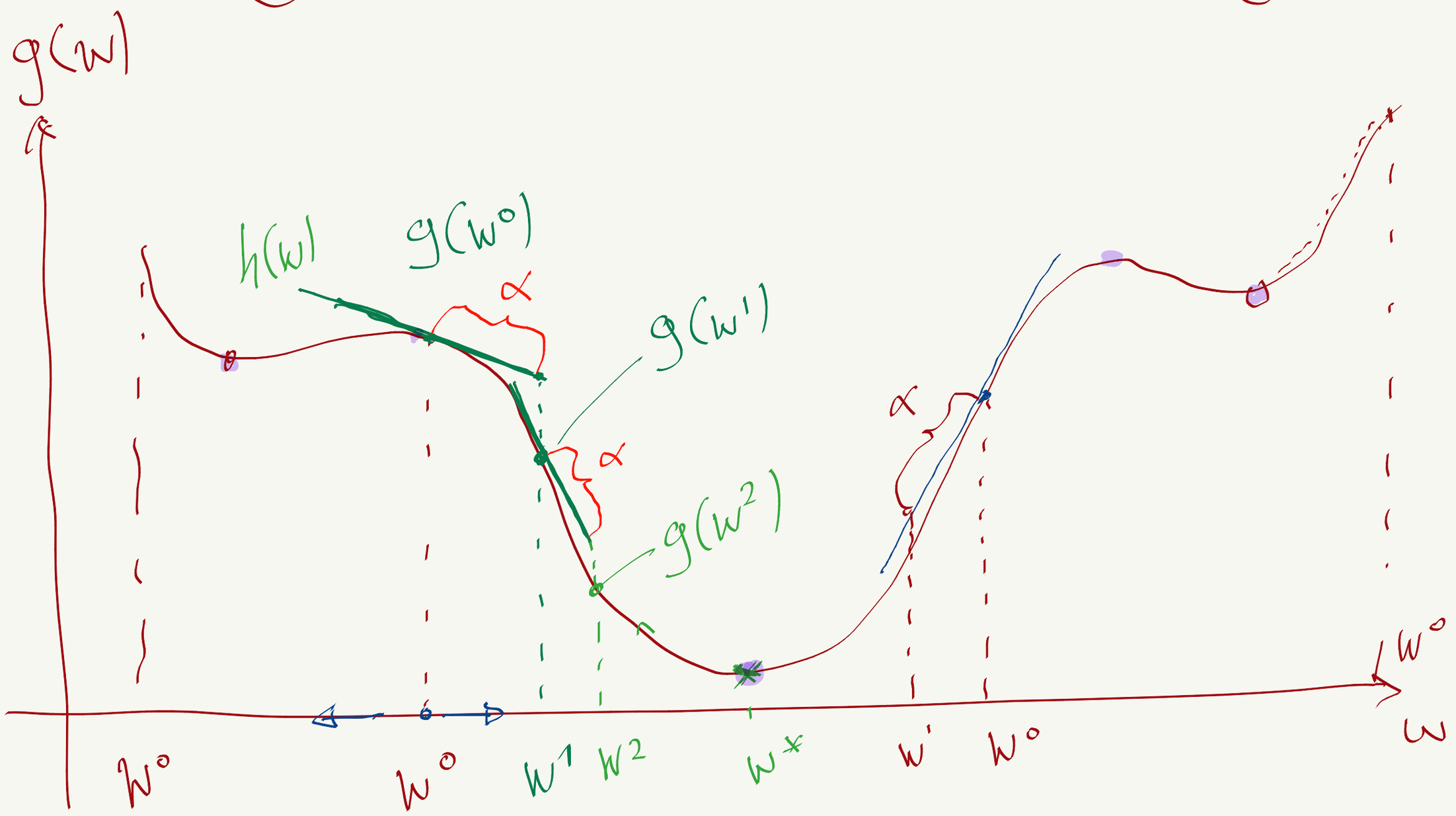
Lecture 4

Numerical Optimization



1st order optimization

Gradient Descent (GD) algorithm



- Use a linear approximation to g to determine the nex step.
- Linear approximation : 1st order Taylor approx.

$$\rightarrow \quad h(\bar{w}) \simeq g(\bar{w}^0) + \nabla g(\bar{w}^0)^T (\bar{w} - \bar{w}^0)$$

- Which direction is "downward" or "downhill"?

△ accepts "trivial" answer for $N = 1$

△ $N > 1$

$$h(\bar{w}) = g(\bar{w}^0) + \nabla g(\bar{w}^0)^T \bar{w} - \nabla g(\bar{w}^0)^T \bar{w}^0$$

do not depend on w

$$\min_{\|w\|=1} \nabla g(\bar{v})^T \bar{w} = \min_{\|w\|=1} \|\nabla g(\bar{v})\| \cdot \|\bar{w}\| \cdot \cos\theta$$

clearly

$-\nabla g(\bar{v})$ ⟷ w

when $\theta = \pi$

$$\Rightarrow \bar{w} = \frac{-\nabla g(\bar{v})}{\|\nabla g(\bar{v})\|}$$

downhill direction

$$-\nabla g(\bar{w}^0)$$

GD algorithm

Start $\bar{w}^0$

step length or learning rate

$$\bar{w}^1 = \bar{w}^0 - \alpha \nabla g(\bar{w}^0)$$

$$\bar{w}^2 = \bar{w}^1 - \alpha \nabla g(\bar{w}^1)$$

$$\Rightarrow \boxed{\bar{w}^k = \bar{w}^{k-1} - \alpha \nabla g(\bar{w}^{k-1})} \Leftarrow$$

$$\hookrightarrow \boxed{\bar{w}^k - \bar{w}^{k-1} = -\alpha \nabla g(\bar{w}^{k-1})}$$

for fixed α

More general form of GD

$$\boxed{\bar{w}^k = \bar{w}^{k-1} - \alpha^k \nabla g(\bar{w}^{k-1})}$$

$$\left.\begin{aligned} \alpha^k &= \frac{1}{k} \\ \alpha^k &= \frac{k}{k+3} \end{aligned}\right\}$$

When to stop iterating?

✓ (1.) when a stationary point is approximately reached

$$\|\nabla g(\bar{w})\| < \varepsilon \leftarrow \text{small number}$$

✓ (2.) $$\frac{\|\bar{w}^{k+1} - \bar{w}^k\|}{\|w^k\|} < \delta$$

→ 3. after a preset # of iterations

$$\begin{cases} \bar{w}^{k+1} = \bar{w}^k - \underbrace{\alpha}_{=1} \nabla g(\bar{w}^k) \\ \text{for } \nabla g(\bar{w}^k) = 0 \Rightarrow \bar{w}^{k+1} = \bar{w}^k \end{cases}$$

$$\begin{cases} \bar{w}^{k+1} - \bar{w}^k = -\alpha \nabla g(\bar{w}^k) \\ \Rightarrow \underbrace{\|\bar{w}^{k+1} - \bar{w}^k\|}_{\text{small}} = \alpha \underbrace{\|\nabla g(\bar{w}^k)\|}_{\text{small}} \end{cases}$$

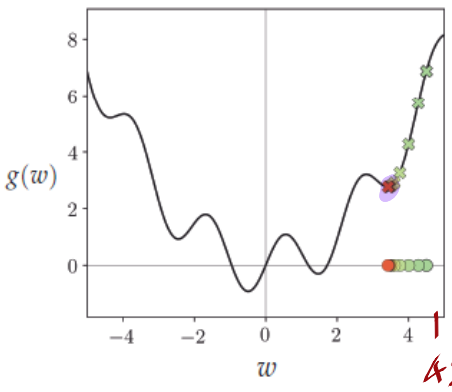
small
large

small
large

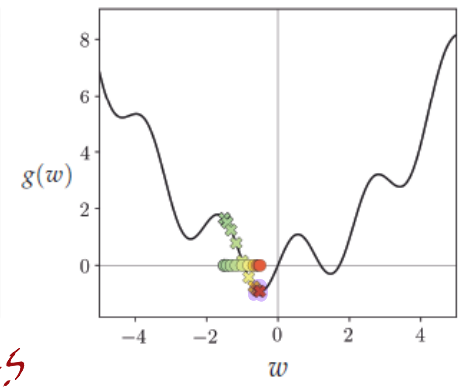
$g(w) = \sin(3w) + 0.3w^2$

$w^0 = 4.5$

$\alpha = 0.05$



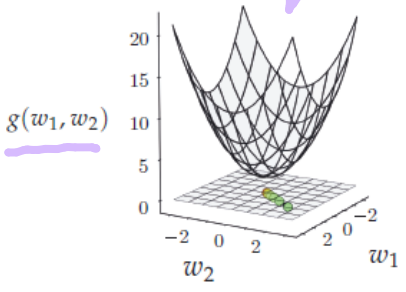
4.5

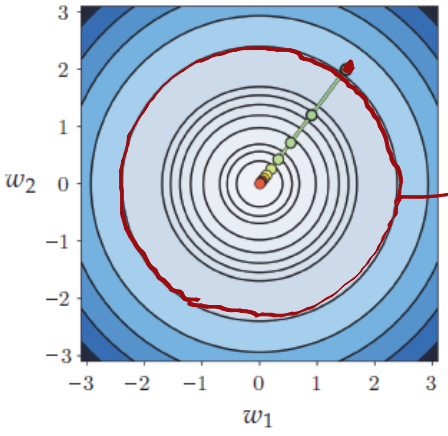


$w^0 = -1.5$

$\alpha = 0.05$

3D plot





"from above"

Iso-value curves

$\alpha = 0.1$

$g(\vec{w}) = w_1^2 + w_2^2 + 2$

Figure 3.7 (top panels) Figure associated with Example 3.7. (bottom panels) Figure associated with Example 3.8. See text for details.

$g(w) = w^2$, $w^0 = -2.5$

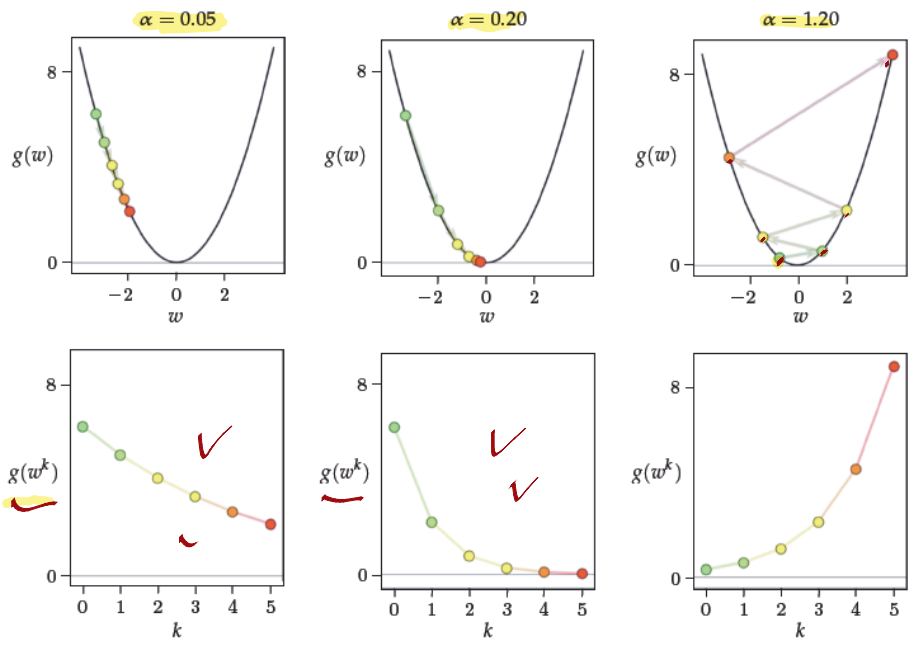


Figure 3.8 Figure associated with Example 3.9. See text for details.

slow

ascend !

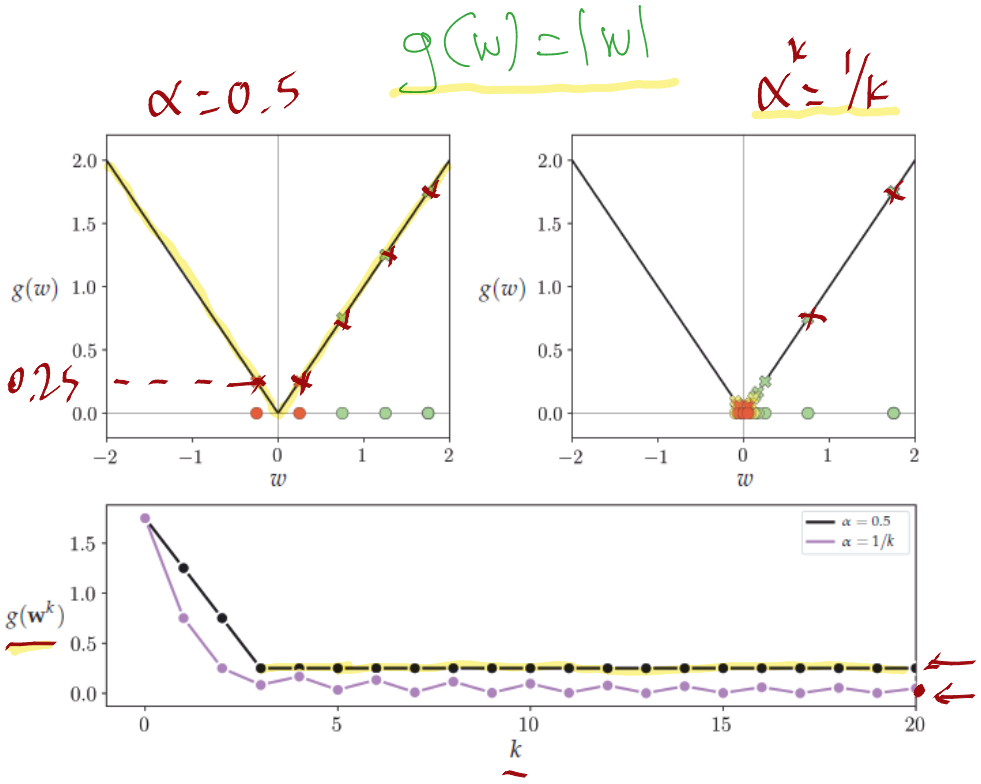


Figure 3.9 Figure associated with Example 3.10. See text for details.

20 iterations, $w^0 = 2$

$$\frac{d}{dw} g(w) = \begin{cases} +1, & w > 0 \\ -1, & w < 0 \end{cases}$$

[Hwk # 3.6]

$\Rightarrow g(\bar{w}) = w_1^2 + w_2^2 + 2\sin(1.5(w_1+w_2))^2 + 2$

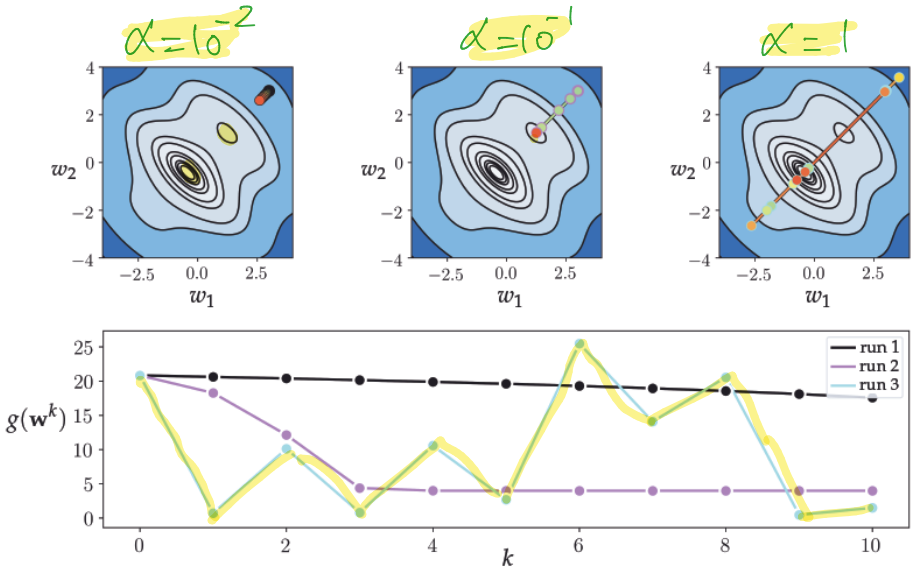


Figure 3.10 Figure associated with Example 3.11. See text for details.

→ local min ≈ $[1.5 \quad 1.5]^T$

→ global " ≈ $[-0.5 \quad -0.5]^T$

$\bar{w}^0 = [3 \quad 3]^T$, 10 steps

→ $\alpha = 10^{-2}$: too small

→ $\alpha = 10^{-1}$: local min

→ $\alpha = 1$: ✓

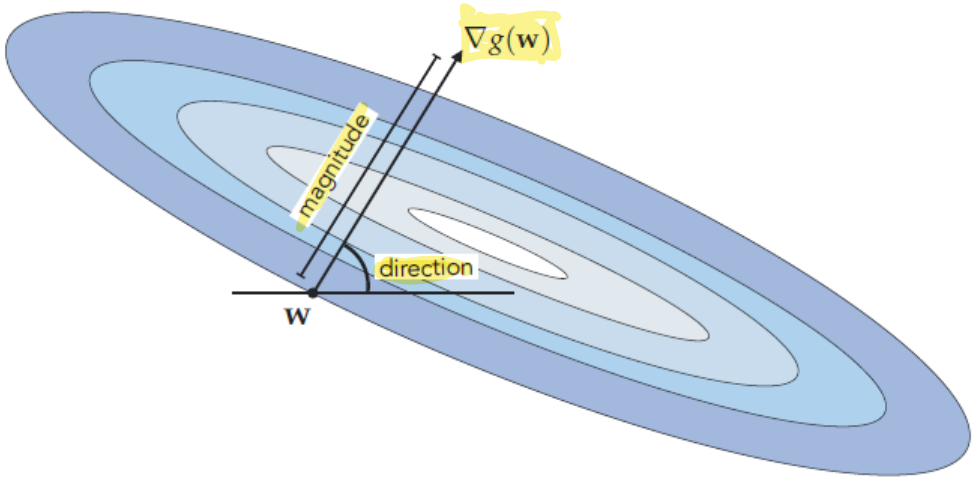


Figure 3.11 The gradient vector of any arbitrary function at any point consists of a *magnitude* and a *direction*.

direction : zig-zagging

magnitude : slow-crawl

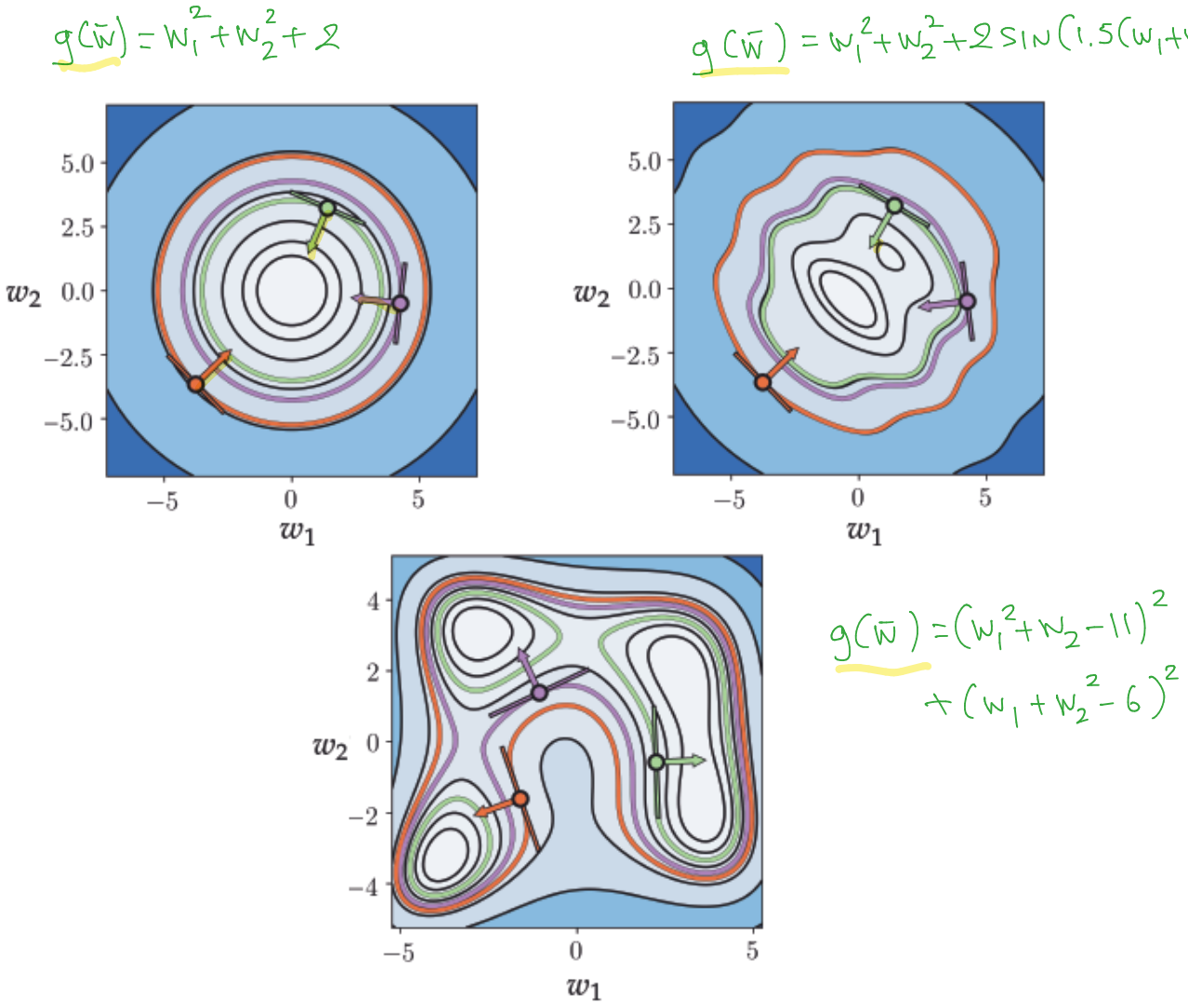
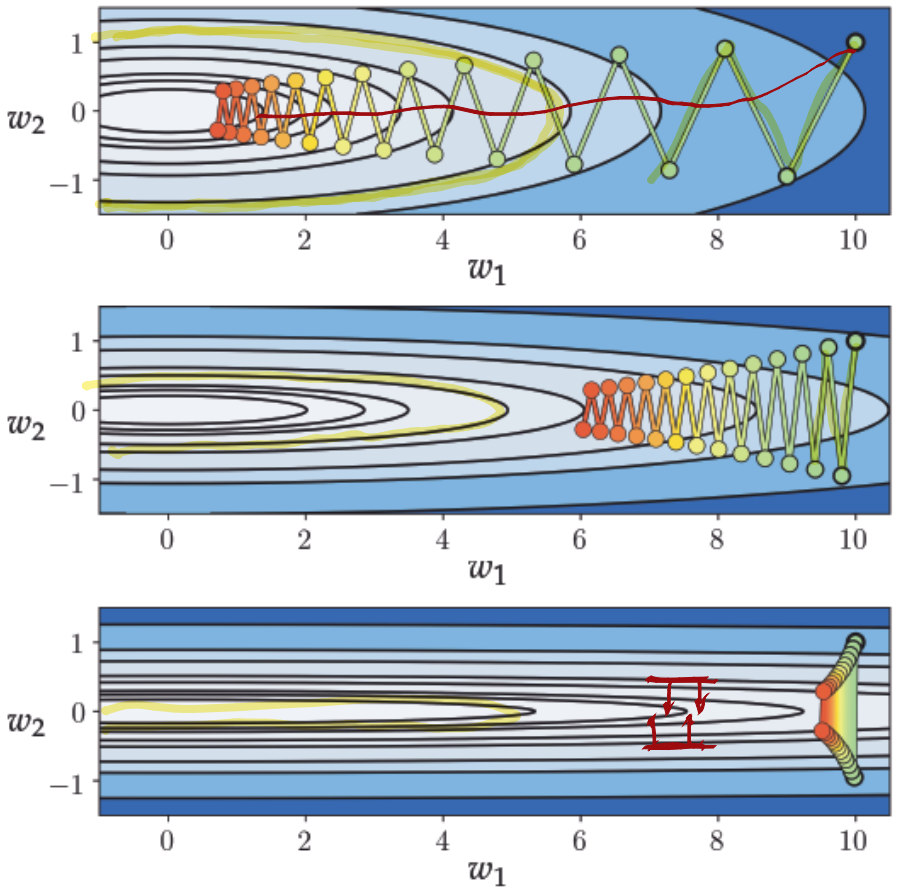


Figure 3.12 Figure associated with Example 3.12. Regardless of the function the negative gradient direction is always *perpendicular* to the function's contours. See text for further details.

$g(\bar{w}) = \bar{w}^T C \bar{w}$



$C = \begin{bmatrix} 0.5 & 0 \\ 0 & 2 \end{bmatrix}$

$C = \begin{bmatrix} 0.1 & 0 \\ 0 & 2 \end{bmatrix}$

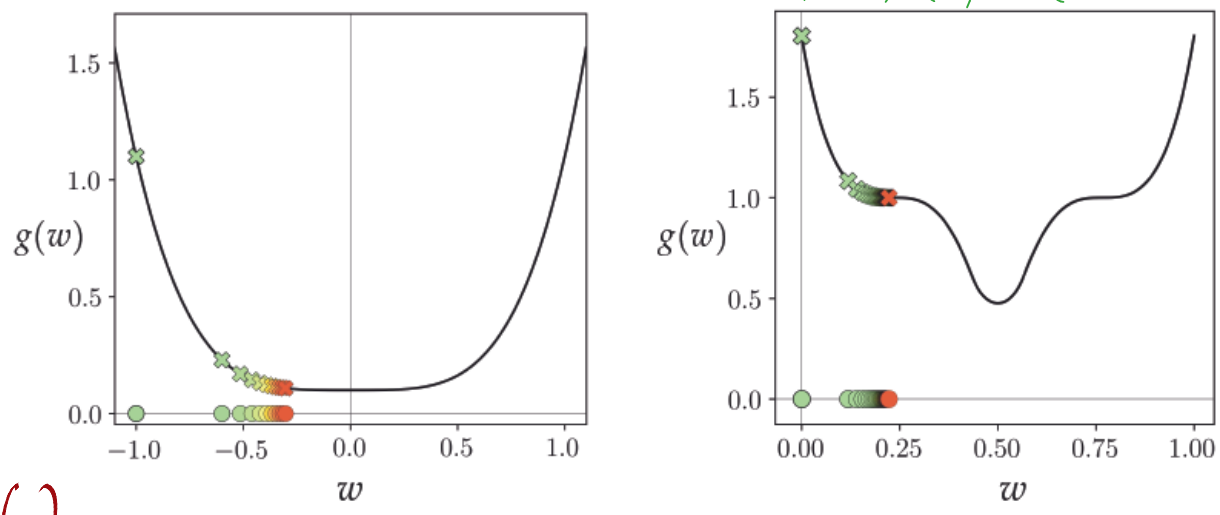
$C = \begin{bmatrix} 0.01 & 0 \\ 0 & 2 \end{bmatrix}$

Figure 3.13 Figure associated with Example 3.13, illustrating the zig-zagging behavior of gradient descent. See text for further details.

"solution": momentum-accelerated GD (A.2)

$g(w) = w^4 + 0.1$

$g(w) = \max^2(0, 1 + (3w - 2.3)^3) + \max^2(0, 1 + (-3w + 0.7)^3)$



$-\alpha \nabla g(\cdot) \approx 0$

Figure 3.14 Figure associated with Example 3.14. See text for details.

left: $\alpha = 10^{-1}$, 10 steps

right: $\alpha = 10^{-2}$, 50 steps

min at $w = 1/2$

saddle points at: $w = 7/30$, $w = 23/30$

"solution": normalized GD (A.3, A.4)

Adam optimizer (2 momentum approaches)