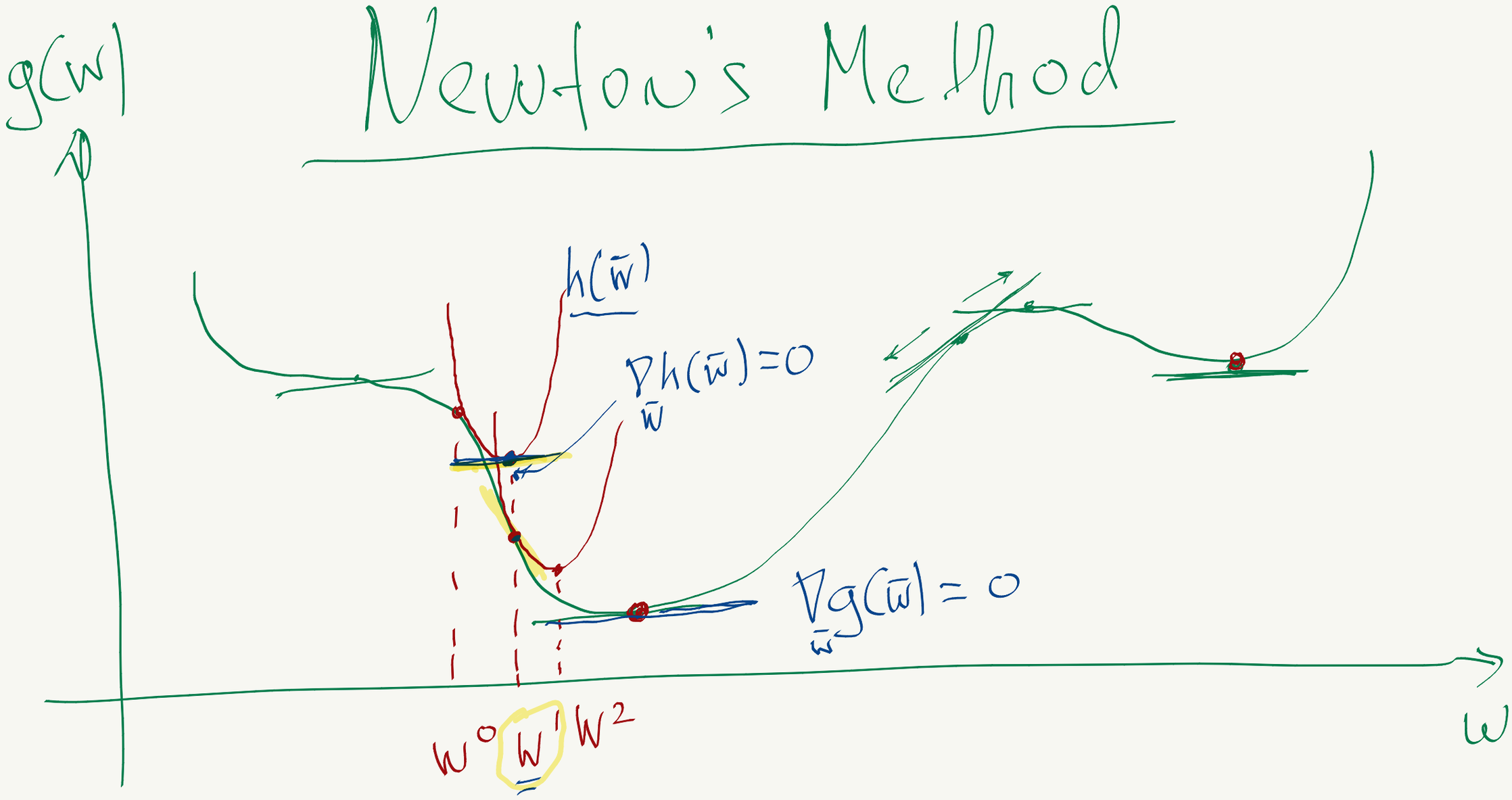
4/15/20

Newton's Method



2nd order Taylor series approximation

$$\nabla h(\bar{w}) = \underbrace{\nabla g(\bar{w}^0)}_{①} + \underbrace{\nabla\left(\nabla g(\bar{w}^0)^T(\bar{w}-\bar{w}^0)\right)}_{②} + \frac{1}{2}\underbrace{\nabla\left((\bar{w}-\bar{w}^0)^T \cdot \nabla^2 g(\bar{w}^0) \cdot (\bar{w}-\bar{w}^0)\right)}_{③}$$

Find stationary point

$$\boxed{\nabla_{\bar{w}} h(\bar{w}) = 0}$$

① $\nabla_w g(\bar{w}^0) = 0$

(T)

(B)

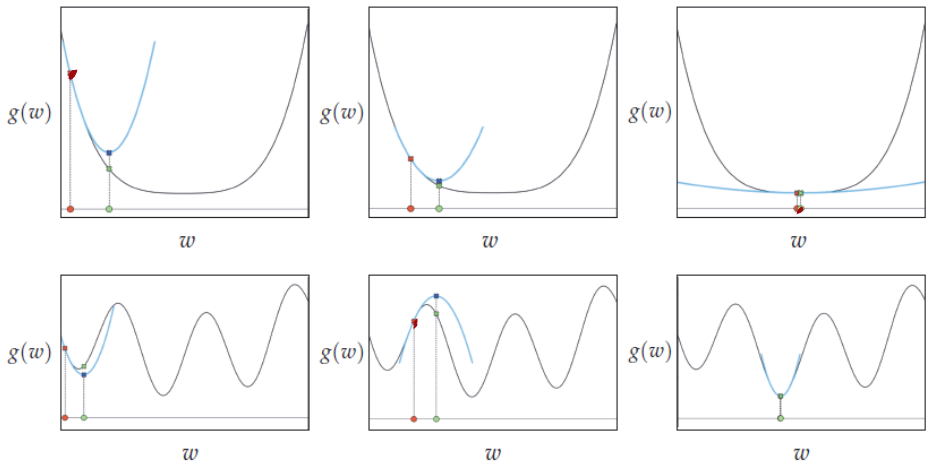


Figure 4.5 Figure associated with Example 4.4. See text for details.

⇒ (T): $g(w) = \frac{1}{50}\left(w^4 + w^2\right) + 0.5$

⇒ (B): $g(w) = \sin(3w) + 0.1w^2 + 1.5$

② $\nabla\left(\nabla g(\bar{w}^0)^T(\bar{w}-\bar{w}^0)\right)$

$= \nabla_w\left(\nabla g(\bar{w}^0)^T \bar{w}\right) - \nabla\left(\nabla g(\bar{w}^0)^T \bar{w}^0\right)$ (the second term $\to 0$)

$= \nabla(\bar{w}^T \cdot \nabla g(\bar{w}^0))$

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

$\left[\nabla(\bar{w}^T\bar{c}) = \bar{c}\right]$

$\left[\nabla(\bar{c}^T\bar{w}) = \bar{c}\right]$

$\left[\bar{w}^T\bar{c} = \bar{c}^T\bar{w}\right]$

③

$\nabla\left((\underbrace{\bar{w}-\bar{w}^0}_{\bar{x}^T})^T \underbrace{\nabla^2 g(\bar{w}^0)}_{A}(\underbrace{\bar{w}-\bar{w}^0}_{\bar{x}})\right)$

$\left[\nabla_{\bar{x}}(\bar{x}^T A \bar{x}) = 2A\bar{x}\right]$

$\left[\frac{d}{dx}(5x^2) = 2\cdot 5\cdot x\right]$

$\nabla_w\left(f^T(\bar{w}) A f(\bar{w})\right)$

$= \nabla_{f(\bar{w})}\left(f^T(\bar{w}) A f(\bar{w})\right) \nabla_w f(\bar{w})$

$\Rightarrow \underline{\underline{2}}\,\nabla^2 g(\bar{w}^0)(\bar{w}-\bar{w}^0) \cdot \nabla_w(\bar{w}-\bar{w}^0)$ (the last factor $\to 1$)

$$\rightarrow \quad \nabla h(\bar{w}) = 0$$

$$\Rightarrow \quad \boxed{\nabla g(\bar{w}^0) + \nabla^2 g(\bar{w}^0) \cdot (\bar{w} - \bar{w}^0) = 0}$$

$$\Rightarrow \quad \underbrace{\nabla^2 g(\bar{w}^0)}_{\text{constant}} \cdot \bar{w} = \underbrace{\nabla^2 g(\bar{w}^0) \cdot \bar{w}^0 - \nabla g(\bar{w}^0)}_{\text{constant}}$$

if Hessian is invertible

$$\Rightarrow \quad \bar{w} = \left(\nabla^2 g(\bar{w}^0)\right)^{-1} \left(\nabla^2 g(\bar{w}^0)\, \bar{w}^0 - \nabla g(\bar{w}^0)\right)$$

$$\Rightarrow \quad \bar{w} = \bar{w}^0 - \left(\nabla^2 g(\bar{w}^0)\right)^{-1} \cdot \nabla g(\bar{w}^0)$$

$\downarrow$

$\bar{w}'$

General Form of Newton's Method

Start $\bar{w}^0$

$$\bar{w}^k = \bar{w}^{k-1} - \left(\nabla^2 g(\bar{w}^{k-1})\right)^{-1} \cdot \nabla g(\bar{w}^{k-1})$$

Gradient Descent

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

$$\bar{w}^k = \bar{w}^{k-1} + \alpha^k \bar{d}^{k-1}$$

GD: $\bar{d}^{k-1} = -\nabla g(\bar{w}^{k-1})$; $\alpha^k = \alpha^k$

NM: $\bar{d}^{k-1} = -\left(\nabla^2 g(\bar{w}^{k-1})\right)^{-1} \cdot \nabla g(\bar{w}^{k-1})$; $\alpha^k = 1$

$$w^k = w^{k-1} - \frac{\frac{d}{dw} g(w^{k-1})}{\frac{d^2}{dw^2} g(w^{k-1}) + \epsilon}$$

Newton's method ≃ GD

⊖ need to choose α

⊕ NO α to be chosen

⊖ need inverse of Hessian
(lack of scalability
w/ input dimensions)

⊖ does not handle
well NON-convexity
(or local min
local max)

⊕ it does not suffer
from zig-zagging

⊕ typically converges
much faster
(for convex functions)

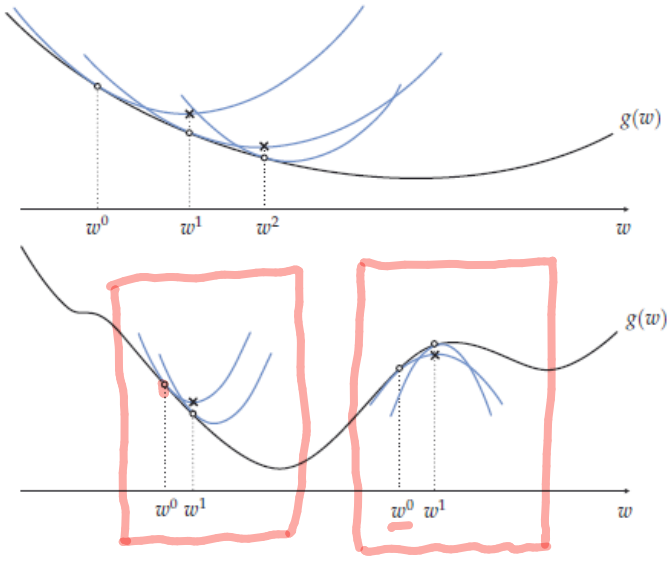


Figure 4.6 Newton's method illustrated. To find a minimum of g, Newton's method hops down the stationary points of quadratic approximations generated by its second-order Taylor series. (top panel) For convex functions these quadratic approximations are themselves always convex (whose only stationary points are minima), and the sequence leads to a minimum of the original function. (bottom panel) For nonconvex functions quadratic approximations can be concave or convex depending on where they are constructed, leading the algorithm to possibly converge to a maximum.

← CONVEX

← NON-CONVEX

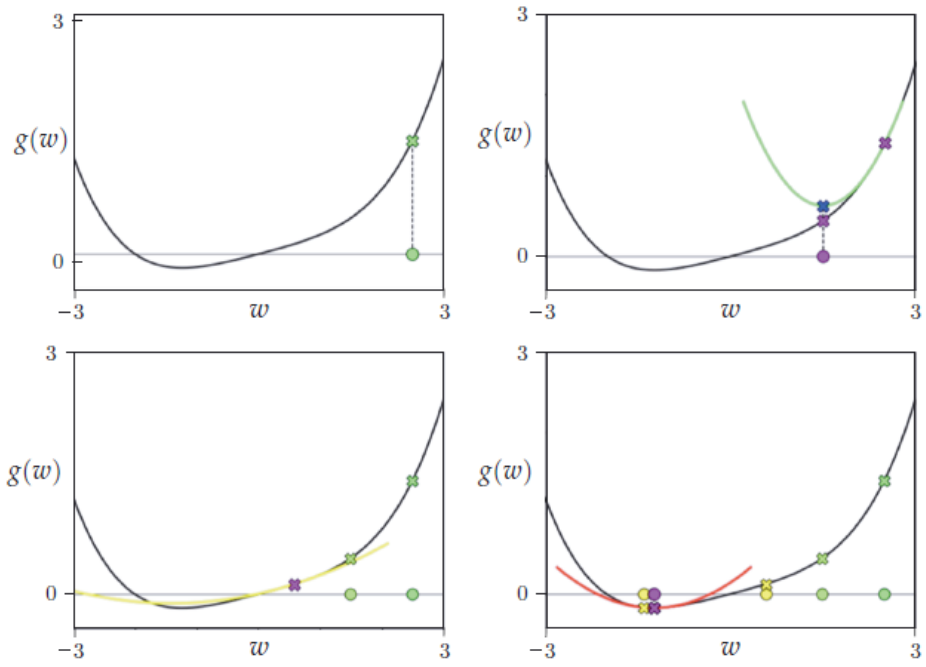


Figure 4.7 Figure associated with Example 4.5, animating a run of Newton's method applied to the function in Equation (4.18). See text for further details.

Dealing w/ NON-convexity → regularization

to avoid "division by zero"

$$\bar{w}^k = \bar{w}^{k-1} - \left(\nabla^2 g(\bar{w}^{k-1}) + \epsilon I\right)^{-1} \nabla g(\bar{w}^{k-1})$$

$$\Rightarrow \left(\nabla^2 g(\bar{w}^{k-1}) + \epsilon I\right) \bar{w}^k = \left(\nabla^2 g(\bar{w}^{k-1}) + \epsilon I\right) \bar{w}^{k-1} - \nabla g(\bar{w}^{k-1})$$

↗ this equation can be interpreted as the stationary point of a slightly adjusted 2nd order Taylor series approximation

$$\begin{cases} h(\bar{w}) = g(\bar{w}^{k-1}) + \nabla g(\bar{w}^{k-1})^T (\bar{w} - \bar{w}^{k-1}) \\ \qquad + \frac{1}{2} (\bar{w} - \bar{w}^{k-1}) \nabla^2 g(\bar{w}^{k-1}) (\bar{w} - \bar{w}^{k-1}) \\ \qquad + \frac{\epsilon}{2} \|\bar{w} - \bar{w}^{k-1}\|_2^2 \end{cases}$$

$h_1(w)$

regularizer $h_2(\bar{w})$

$$(\bar{w} - \bar{w}^{k-1})^T \left(\frac{\epsilon}{2} I\right) (\bar{w} - \bar{w}^{k-1})$$

N positive eigenvalues $= \epsilon/2$ (ϵ-values)

$$h(\bar{w}) = \underbrace{h_1(\bar{w})}_{\text{flat or non-convex}} + \underbrace{\frac{\epsilon}{2} h_2(w)}_{\text{always convex}}$$

as $\epsilon \uparrow$ we convexify $h(\bar{w})$

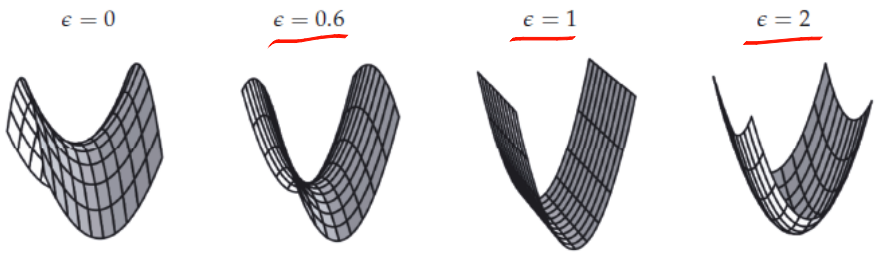


Figure A.12 Figure associated with Example A.9. From left to right, a nonconvex quadratic function is slowly turned into a convex function via the weighted addition of a convex quadratic. See text for further details.

$$h(\bar{w}) = h_1(w_1, w_2) + \epsilon \cdot h_2(w_1, w_2)$$

$$\rightarrow h_1(w_1, w_2) = w_1^2 - w_2^2$$

$$\rightarrow h_2(w_1, w_2) = w_1^2 + w_2^2 = \|\bar{w}\|_2^2$$

Dealing w/ scaling w/ input dimension

$\Rightarrow$ <u>Hessian-Free Methods</u>

(appendix A)

A. Subsampling the Hessian

- keep only diagonal value

decouple along each coordinate

$$w_n = \begin{bmatrix} w_1 \\ w_2 \\ \vdots \\ w_n \\ \vdots \\ w_N \end{bmatrix} \qquad w_n^k = w_n^{k-1} - \frac{\frac{\partial}{\partial w_n} g(\bar{w}^{k-1})}{\frac{\partial^2}{\partial w_n^2} g(\bar{w}^{k-1})}$$

B. Quasi-Newton methods

Replace Hessian by <u>Secant Matrix</u>

(use of low <u>rank</u> matrices)

$$A = u u^T$$

$A = [\ ][\ \ \ ]$

↑ rank 1