Comp Sci, Dec 13, 2022

Taylor expand $c(\beta)$

$(\hat{\beta} = \beta)$

$$c(\beta) \simeq c(\beta^{(n)}) + g_{(n)}^T (\beta - \beta^{(n)})$$

$$+ \frac{1}{2}(\beta - \beta^{(n)})^T H^{(n)} (\beta - \beta^{(n)})$$

$$\beta = \beta^{(n)} - H_{(n)}^{-1} g^{(n)}$$

$\gamma^{(n)}$ = learning rate

$$\beta = \beta^{(n)} - \gamma g^{(n)}$$

Expand around this value

$$c(\beta^{(n)} - \gamma g^{(n)}) = c(\beta^{(n)})$$

$$- \gamma g^T g + \frac{1}{2} \gamma^2 g^T H g + \ldots$$

Truncate at γ^2.

optimal γ $\quad \frac{\partial c}{\partial \gamma} = 0$

$$\gamma = \frac{g^T g}{g^T H g}$$

Taylor expansion problems when $g^T H g > g^T g$

Assume g is an eigenvector of H

$$Hg = \lambda \cdot g \qquad g^T g = 1$$

$$\gamma = \frac{1}{\lambda}$$

smallest step size (γ) given by $\gamma = \frac{1}{\lambda_{max}}$

Largest learning rate/step size given by

$$\gamma = \frac{1}{\lambda_{min}}$$

Can show $\gamma < \frac{2}{\lambda_{max}}$

λ is eigenvalue of H

<u>Steepest descent</u>

$$f(x) = \frac{1}{2} x^T A x - x^T b$$

$$\frac{\partial f}{\partial x} = 0 = Ax - b \Rightarrow$$

$$Ax = b$$

Define residual $r = b - Ax$

have solution when $r = 0$

start with a guess for

$$x = x_0$$

$$r_0 = - Ax_0 + b$$

in general

$$r_{k+1} = b - Ax_{k+1}$$

$$x_{k+1} = x_k + \alpha_k r_k$$

$$r_{k+1} = b - A(x_k + \alpha_k r_k)$$

$$= \underbrace{(b - Ax_k)}_{r_k} - \alpha_k A r_k$$

$$= r_k - \alpha_k A r_k$$

$$r_{k+1} = 0$$

$$r_k^T r_{k+1} = 0 = r_k^T r_k - \alpha_k r_k^T A r_k$$

$$\Rightarrow \quad \alpha_k = \frac{r_k^T r_k}{r_k^T A r_k}$$

A = Hessian

$$x_{k+1} = x_k + \alpha_k r_k$$

plays the role of the gradient $g(\beta^{(n)})$

$$\left(\beta_{k+1} = \beta_k - \gamma_k g_k \right)$$

Gradient descent with momentum

particle moving in a field

$$F = -\nabla V(x)$$

$$m \frac{d^2x}{dt^2} + \mu \frac{dx}{dt} = -\nabla V(x)$$

$\uparrow$
drag/friction
etc

discretize second derivative

$$\frac{d^2x}{dt^2} \simeq \frac{x_{t+\Delta t} - 2x_t + x_{t-\Delta t}}{(\Delta t)^2}$$

$$\frac{dx}{dt} \simeq \frac{x_{t+\Delta t} - x_t}{\Delta t}$$

$$m \left(\frac{x_{t+\Delta t} - 2x_t + x_{t-\Delta t}}{(\Delta t)^2}\right)$$

$$+ \mu \left(\frac{x_{t+\Delta t} - x_t}{\Delta t}\right) = -\nabla V(x)$$

$$\Delta x_{t+\Delta t} = x_{t+\Delta t} - x_t$$

$$\Delta x_t = x_t - x_{t-\Delta t}$$

$$\Delta x_{t+\Delta t} = -\frac{(\Delta t)^2}{m+\mu\Delta t}\vec{\nabla} V(x)$$

$$+ \frac{m}{m+\mu\Delta t}\Delta x_t$$

$$\delta = \frac{m}{m+\mu\Delta t} \wedge \gamma = \frac{(\Delta t)^2}{m+\mu\Delta t}$$

$$\Delta x_{t+\Delta t} = -\gamma \nabla V(x) + \underbrace{\delta \Delta x_t}_{\text{Memory term}}$$

$$x_t \to \beta_i \qquad x_{t+\Delta t} \Rightarrow \beta_{i+1}$$

$$\Delta \beta_{i+1} = \beta_{i+1} - \beta_i$$

$$\beta_{i+1} = \beta_i - \gamma g(\beta_i)$$

$$+ \delta(\beta_i - \beta_{i-1})$$