Lecture September 25

Cost function $C(\beta)$

$$\frac{\partial C}{\partial \beta} = g(\beta)$$

$$\frac{\partial^2 C}{\partial \beta \partial \beta^T} = H(\beta)$$ Hessian matrix

Can add regularization (L2-regularization)

$$C'(\beta) = C(\beta) + \lambda \beta^T \beta$$

$$g'(\beta) = g(\beta) + 2\lambda\beta$$

$$H'(\beta) = H(\beta) + 2\lambda \mathbb{I}$$

Math aside

$$* \quad \frac{\partial (a^T A a)}{\partial a} = (A + A^T) a$$

(Hessian $H = H^T$)

$$** \quad \frac{\partial (b^T a)}{\partial a} = b$$

Newton's method; iterative scheme

$$\beta^{(n+1)} = \beta^{(n)} - \left[H^{(n)}\right]^{-1} g^{(n)}$$

$$H^{(n)} = H(\beta^{(n)})$$

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

Find $\beta^{opt} = \hat{\beta}$ when

$$\|\beta^{(n+1)} - \beta^{(n)}\|_2 \leq \varepsilon \sim 10^{-8}$$

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

Learning rate.

$$\gamma^{(n)} < \frac{2}{\lambda_{max}}$$

Largest eigenvalue of H

Gradient/steepest descent.

$$C(\hat{\beta}) \cong C(\beta^{(n)}) +$$

(Taylor expand around $\hat{\beta} - \beta^{(n)}$)

$$+ \left[g^{(n)}\right]^T (\hat{\beta} - \beta^{(n)}) +$$

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

Drop

$$= C + \frac{1}{2}\beta^T A \beta + b^T \beta \Leftarrow$$

$$A = H^{(n)} = \text{known}$$

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

$$= \text{known}$$

$$C = C(\beta^{(n)}) - [g^{(n)}]^T \beta^{(n)}$$

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

$$\frac{\partial C}{\partial \beta} = A\beta + b = 0$$

(using *, **)

$$\Rightarrow$$

$$\hat{\beta} = -A^{-1}b = \beta^{(n)} - [H^{(n)}]^{-1} \cdot g^{(n)}$$

which is what we got from Newton's method.

Can generalize to a function $f(x) = \boxed{\frac{1}{2}x^T A x + b^T x}$

$+ C$

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

$$\boxed{Ax = -b = +\tilde{b}}$$

Standard steepest descent

iterative methods

guess $x_0 = x^{(0)}$ $\quad (Ax = b)$

$$\underline{r}^{(0)} = b - Ax^{(0)}$$

$$r^{(k+1)} = b - A\underline{x}^{(k+1)}$$

$$= \underline{b - A}(x^{(k)} + \alpha^{(k)} r^{(k)})$$

$$\underline{b - Ax^{(k)}} - \alpha^{(k)} A r^{(k)} = 0$$

$$(r^{(k)})^T \cdot \quad r^{(k)} = \alpha^{(k)} A r^{(k)}$$

$$[r^{(k)}]^T r^{(k)} = \alpha^{(k)} [r^{(k)}]^T A r^{(k)}$$

$$\underline{\alpha}^{(k)} = \boxed{\frac{[r^{(k)}]^T r^{(k)}}{[r^{(k)}]^T A r^{(k)}}}$$

$$\underline{x}^{(k+1)} = x^{(k)} - \alpha^{(k)} r^{(k)}$$

$$\| r^{(k+1)} - r^{(k)} \|_2 \leq \varepsilon$$

in our case A = Hessian matrix

$f(x)$ Taylor expand at a gradient with γ-step

$$f(x) \simeq f(x^{(n)}) + (x - x^{(n)})^T g^{(n)}$$
$$+ \frac{1}{2} (x - x^{(n)})^T H^{(n)} (x - x^{(n)})$$

$$\left(x = x^{(n)} - \gamma g^{(n)} \right)$$

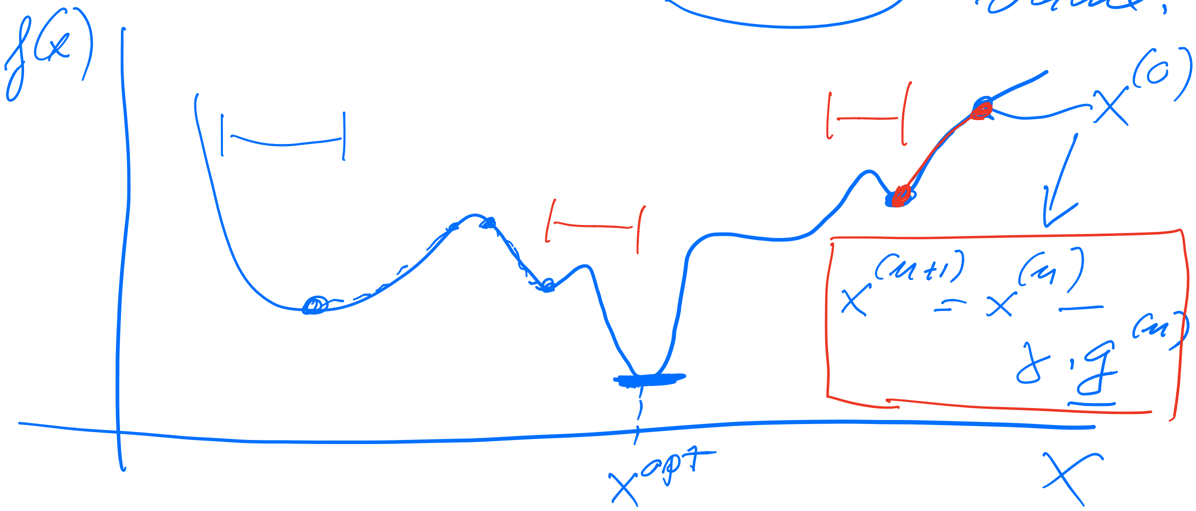
$$\simeq f(x^{(n)}) - \gamma [g^{(n)}]^T g^{(n)}$$
$$+ \frac{1}{2} \gamma^2 [g^{(n)}]^T H^{(n)} g^{(n)}$$

Minimize wrt to γ

$$\gamma^* = \frac{g^{T(n)} g^{(n)}}{g^{T(n)} H^{(n)} g^{(n)}}$$

$$\gamma^{opt} < \frac{2}{\lambda_{max}}$$

largest eigen value.



$g^{(m)}$ calculated for all points

$x_i \quad i = 0, 1, \ldots m-1$

sample a subset of x_i (stochastically)

$\mathbb{E}[g^{(m)}]$ has standard deviation $\sim \sigma / \sqrt{m}$

$m = 100 \qquad \boxed{m = 10000}$

$\sigma/10 \qquad \sigma/100$

$X^T X$