Lecture October 1

Repeat

$$C(\hat{\beta}) \simeq C(\beta^{(n)}) + (\hat{\beta} - \beta^{(n)})^T g^{(n)} + \frac{1}{2}(\hat{\beta} - \beta^{(n)})^T \times H (\hat{\beta} - \beta^{(n)})$$

$$\hat{\beta} - \beta^{(n)} = -g^{(n)} \cdot \eta$$

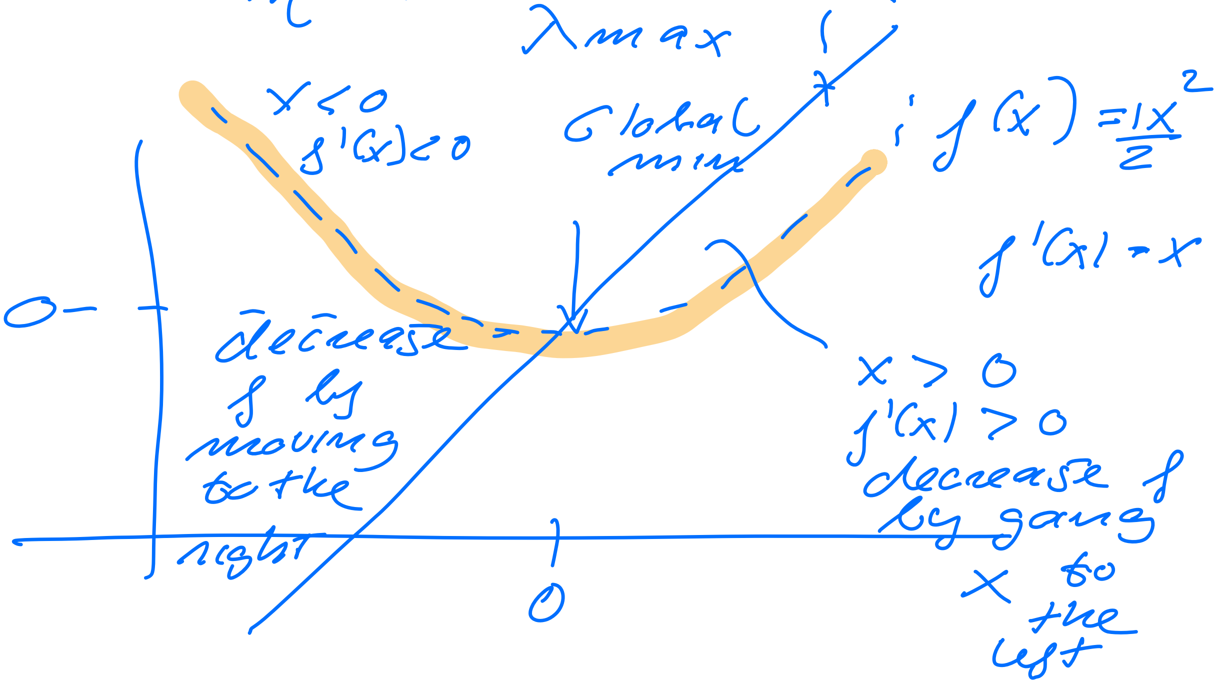
$$C(\hat{\beta}) \simeq C(\beta) - \eta \underline{(g^{(n)})^T} g^{(n)} + \frac{1}{2}(g^{(n)})^T H g^{(n)} \eta^2$$

$$\frac{\partial C}{\partial \eta} = 0 \quad \Rightarrow$$

$$\hat{\eta} = \frac{g^T g}{g^T H g}$$

$$Hg = \lambda g$$

$$\eta = \frac{1}{\lambda_{max}}$$



$$f'(x)=0 \quad \left(=\frac{\partial C(\beta)}{\partial \beta}\right)$$

$f''(x)>0$ local min

$f''(x)<0$ local max

$f''(x)=0$ inconclusive

$$b = \hat{\beta} - \beta^{(m)}$$

$$C(b) = C(\beta^{(m)}) + b^T g + \frac{1}{2} b^T H b$$

$$\left\{ f(x) = \underset{\substack{\| \\ \text{constant}}}{\tilde{C}} \pm x^T \tilde{b} + x^T A x \frac{1}{2} \right\}$$

$$\frac{\partial x^T A x}{\partial x} = (\overset{H}{A + A^T}) x$$

$$\frac{\partial x^T \tilde{b}}{\partial x} = b$$

$$\frac{\partial C}{\partial b^T} = 0 = Hb + g \Rightarrow$$

$$b = \hat{\beta} - \beta^{(m)} = -H^{-1} g$$

$$\Rightarrow \hat{\beta} = \beta^{(m)} - H^{-1} g$$

$$H^{-1} = \eta$$

$$\frac{\partial f(x)}{\partial x^T} = 0 = Ax + \tilde{b}$$

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

$$x = A^{-1} \cdot b$$

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

$$\frac{\partial f(x)}{\partial x^T} = 0 \Rightarrow \boxed{Ax = b}$$

Steepest descent

Defines a residual

$$r = b - Ax$$

Start with a guess

$$x^{(0)}$$

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

Define a recipe for the next value of x

$$x^{(1)} = x^{(0)} + \alpha^{(0)} r^{(0)}$$

to be determined

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

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

$$\left| r^{(k+1)} - r^{(k)} \right| \leq \varepsilon \sim 10^{-10}$$

$$b - A\left(x^{(k)} + \alpha^{(k)} r^{(k)}\right) = 0$$

$$\underbrace{b - A x^{(k)}}_{r^{(k)}} - A \alpha^{(k)} r^{(k)} = 0$$

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

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

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

Stochastic gradient descent (SGD)

$$\beta^{(n+1)} = \beta^{(n)} - \eta^{(n)} \nabla_\beta C(\beta^{(n)})$$

$\nabla_\beta C(\beta^{(n)})$ in linear regression involves a loop over $i = 0, 1, \ldots, n-1$

$$\nabla_\beta C \sim X^T (X\beta - y) = \sum_{i=0}^{n-1} (y_i - \tilde{y}_i) X_{ij}$$

$X \in \mathbb{R}^{n \times p}$

$X_{i*}\beta$

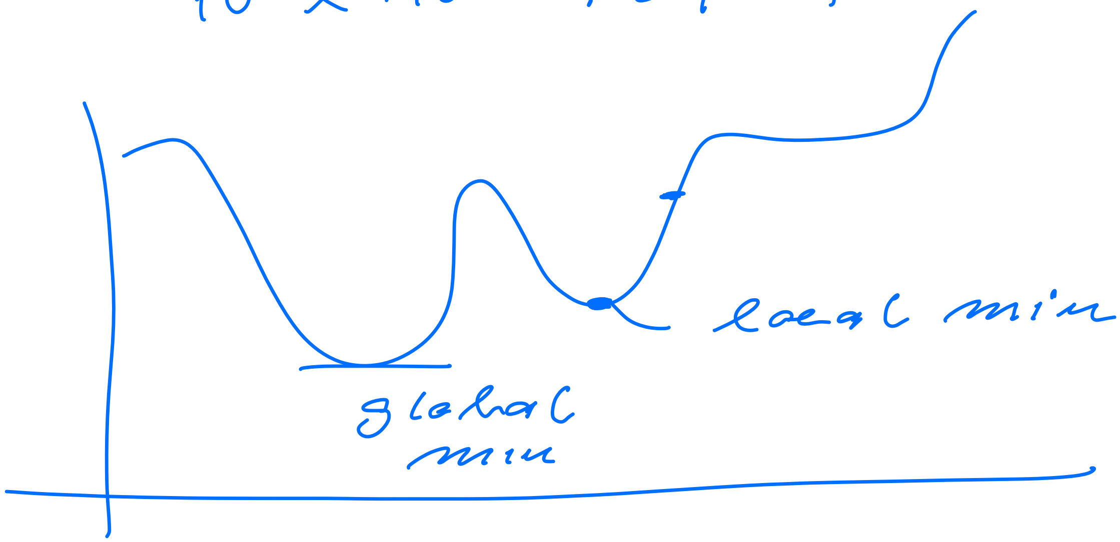
$X^T X \in \mathbb{R}^{p \times p}$

matrix-matrix multiplication is $A \in \mathbb{R}^{m \times m}$ $\sim$ FLOPS $\sim O(m^3)$

$O(p^2 n)$

$n = 10^5 \qquad p = 10^3$

$10^6 \times 10^5$ Flops.



SGD: select, at random, $n_B < n$ ($\ll$) of the training points.

place these points in a mini-batch B_1

Select randomly another n_B points and define mini-batch B_2 ($n - n_B$)

Continue till we run out of training points

Gives n/n_B mini batches

$B_1, B_2, \ldots B_{(n/n_B)}$

start with B_1 and compute gradients. And then use this as input to B_2

continue till you run out of mini-batches = 1 epoch