Lecture November 22

accuracy score $= \dfrac{\sum_{i=0}^{n-1} I(y_i = \tilde{y}_i)}{n}$

confusion matrix

TP = True Positive
equ with a correct
classification

TN = True negative, equ
with correct rejection

FP = False positive
false alarm

FN = False negative, equ
with miss

True positive rate

$$\frac{TP}{TP+FN} = TPR$$

False positive rate

$$= \frac{FP}{FP + TN} = FPR$$

True negative rate

$$= TNR = \frac{TN}{TN + FP}$$

Gains curve

$$\frac{\text{count } TP + \text{count } FP}{\text{all observations}}$$

$$FPR = 1 - TNR$$

Gradient descent

n = # data points

p = # features

$n \gg p \sim n \simeq p \quad n \leq 10^5$

suppose we have the optimal $\hat{\beta}$, in principle this is an iterative process

$$\hat{\beta} \simeq \beta_{k+1}$$

$$|\beta_{k+1} - \beta_k| \le \varepsilon \sim 10^{-10}$$

$\beta = \beta_k$, Taylor expand

$$C(\hat{\beta}) = C(\beta) + g^T(\hat{\beta} - \beta)$$

$\nabla_\beta C(\beta)$

evaluated at β

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

$$\frac{\partial^2 C(\beta)}{\partial \beta \, \partial \beta^T}$$

Logistic reg

$X^T W X$

Define $b = \hat{\beta} - \beta$

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

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

$$b = \hat{\beta} - \beta = -H^{-1} g \Rightarrow$$

$$\hat{\beta} = \beta - H^{-1} g$$

$$\beta_{k+1} = \beta_k - H^{-1}(\beta_k) \nabla_\beta C(\beta_k)$$

Newton-Raphson's method.

Newton's method is derived from a general function

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

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

$$Ax = -b = 0$$

known known

algorithm:

- start with guess β_0
- iterate till

$$|\beta_{k+1} - \beta_k| \leq \varepsilon$$

$$\beta_{k+1} = \beta_k - H^{-1}(\beta_k) \nabla_\beta C(\beta_k)$$

since we have to compute H^{-1} repeatedly, replace $H^{-1}(\beta_K)$ with a constant γ_K $\Rightarrow$

$$\beta_{K+1} = \beta_K - \gamma_K \nabla_\beta C(\beta_K)$$

$\equiv$ GRADIENT DESCENT

$$b = \hat{\beta} - \beta = -H^{-1} \cdot g = -H^{-1} \nabla_\beta C(\beta)$$

$H^{-1} \Rightarrow$ learning rate γ

$$\hat{\beta} - \beta = -\gamma \nabla_\beta C(\beta) = -\gamma g(\beta)$$

$$C(\hat{\beta}) = C(\beta) - \gamma g^T g + \frac{1}{2} \gamma^2 g^T H g$$

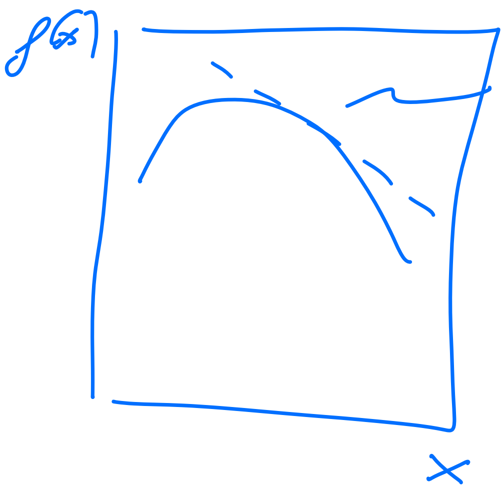
in one-dimension

$$c(\vec{\beta}') = c(\vec{\beta}) - \gamma g^2 + \frac{1}{2}\gamma^2 g^2 H$$

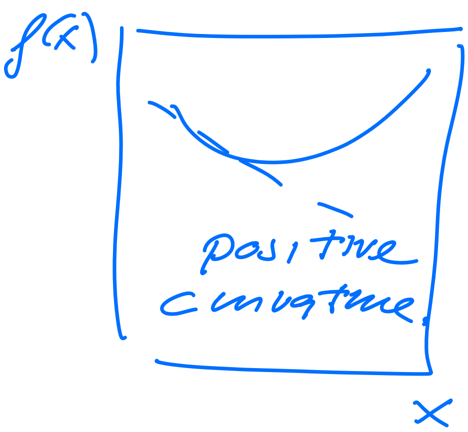
$c(\beta)$ = original/start value

γg^2 = improvement to the slope of $c(\vec{\beta}')$

$\frac{1}{2}\eta^2 g^2 H$ = correction due to curvature



negative curvature



$$\frac{\partial}{\partial \gamma} = 0 = -g^T g + \gamma g^T H g$$

or $\quad -g^T g + \gamma g^T H g \Rightarrow$

$$\gamma = \frac{g^T g}{g^T H g} \left(= \frac{g^T g}{g^T g \lambda} = \frac{1}{\lambda} \right)$$

if $\quad Hg = \lambda H$

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

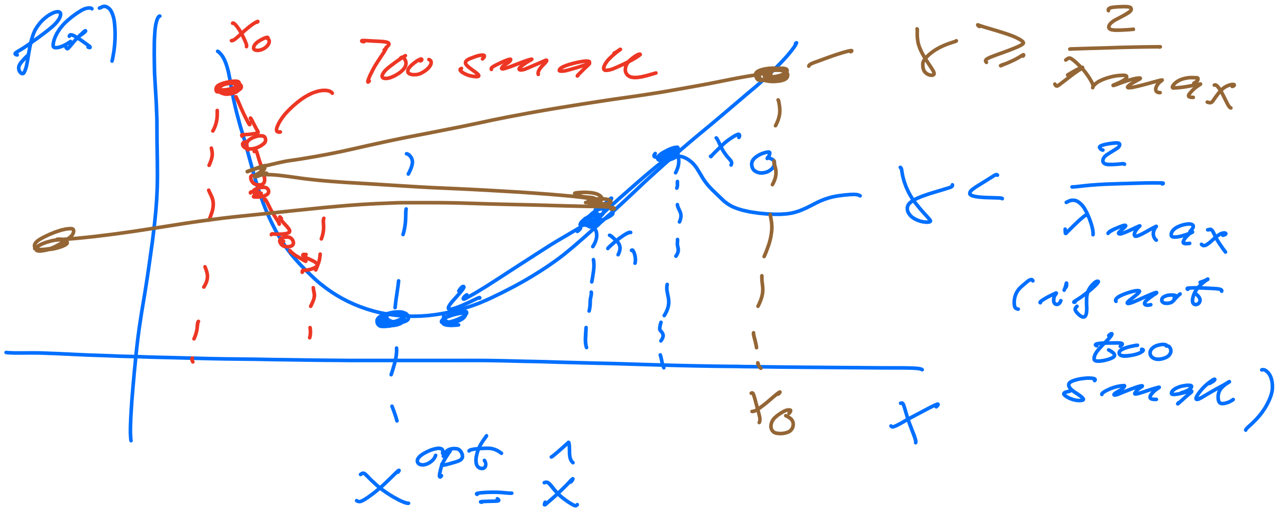
smallest $\gamma = 1/\lambda_{max}$

largest $\gamma = 1/\lambda_{min}$.

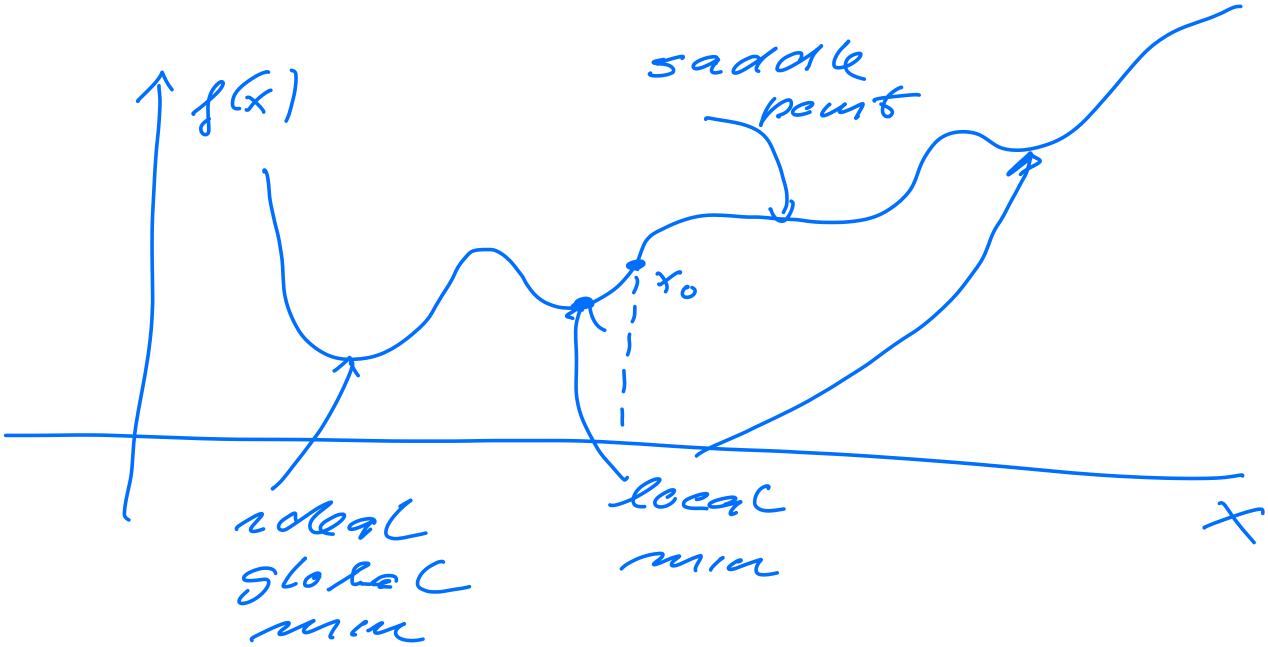
For convergence of Newton-Raphson we must have

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

where λ_{max} is the largest eigenvalue of H.



f(x)
x0
Too small
$\gamma \geq \frac{2}{\lambda_{max}}$
x0
x1
$\gamma < \frac{2}{\lambda_{max}}$
(if not too small)
x0
x
$x^{opt} = \hat{x}$



saddle point
f(x)
x0
x
ideal global min
local min