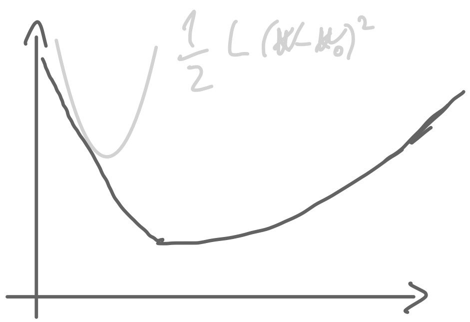
OPTIMAL LECTURE 4, 9/2/2021

$$w_{k+1} = w_k - \alpha_k g_k$$

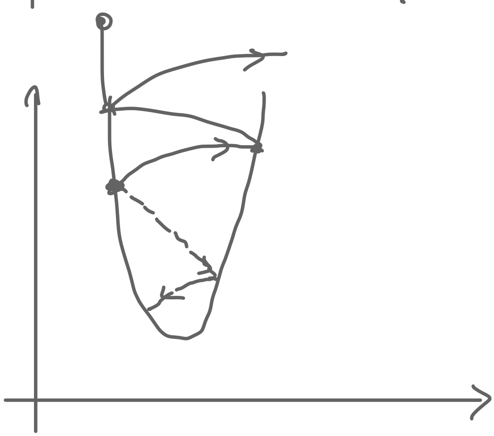
$$g_k = \frac{1}{n_k} H_k \sum_{i=1}^{n_k} \nabla f(w_k, \xi_{k,i})$$

$$H_k = I \qquad n_k = \begin{cases} 1 & \text{SGD} \\ & \text{MB} \\ |S| & \text{FB} \end{cases}$$



$$\| \nabla F(w') - \nabla F(w) \| \le L \| w' - w \|$$

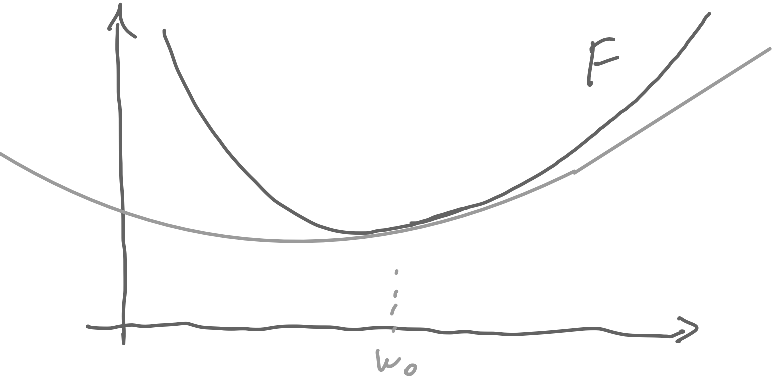
$$\frac{\| \nabla F(w') - \nabla F(w) \|}{\| w' - w \|} \le L$$



$$\boxed{\alpha < \frac{\mu}{L M_G}}$$

$$L = \infty$$

$$\alpha = 0$$



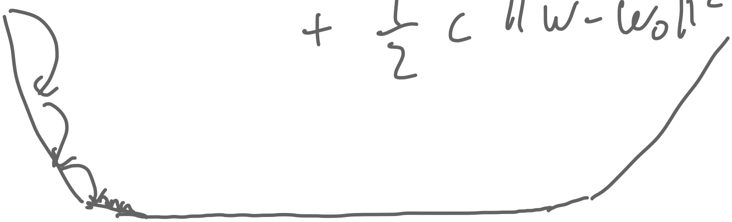
$\frac{1}{2}c(w-w_0)^2$

$F(w) = w^2$

$$F(w) \geq F(w_0) + \nabla F(w_0)^T (w-w_0) + \frac{1}{2}c\|w-w_0\|^2$$

w^2

$2w$

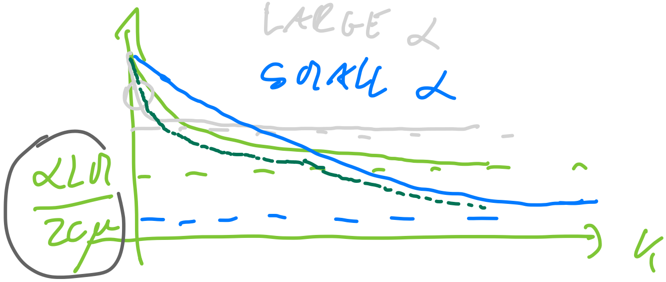


STRONGLY CONVEXITY $\Rightarrow$ GUARANTEE A MINIMUM MOVEMENT (NOT TOO SLOW CONVERGENCE)

L $\searrow$ THE BETTER

C $\nearrow$ THE BETTER

$$\mathbb{E}[F(w_k) - F^*] \leq \boxed{\frac{\alpha L M}{2c\mu}} + (1-\alpha c \mu)^{k-1} \times \left(F(w_1) - F^* - \frac{\alpha L M}{2c\mu}\right)$$



$\rho = (1 - \alpha c \mu)$ $\qquad c \nearrow \quad \rho \searrow$

M : CAPTURES THE NOISE

$\alpha \nearrow \quad (1 - \alpha c \mu) \searrow$

$\frac{\alpha L M}{2 c \mu} \nearrow$

DIMINISHING LEARNING RATE

$$\alpha_k = \frac{1}{k} \qquad \alpha_k \to 0$$

$$\underbrace{\sum_{k=1}^{\infty} \alpha_k = +\infty}_{\text{YOU CAN GO EVERYWHERE}} \qquad \underbrace{\sum_{k=1}^{\infty} \alpha_k^2 < +\infty}_{\text{FILTER OUT THE NOISE}}$$

Theorem (4.7, Bottou) Assumptions of last lecture

$$\alpha_k = \frac{\beta}{\gamma + k} \qquad \beta > \frac{1}{c\mu} \qquad \gamma > 0 \qquad \alpha_1 < \frac{\mu}{L M_G}$$

Thesis: $$\mathbb{E}\left[F(w_k) - F^*\right] \le \frac{v}{\gamma + k}$$

$$v = \max\left(\frac{\beta^2 L M}{2(\beta c \mu - 1)}, (\gamma+1)(F(w_1) - F^*)\right)$$

$$v \simeq \frac{\beta^2 L M}{2(\beta c \mu - 1)}$$

$$\left(\frac{\beta^2}{(\beta c\mu - 1)}\right)' = \frac{2\beta(\beta c\mu - 1) - \beta^2 c\mu}{(\quad)^2} = 0$$

$$2\cancel{\beta}(\beta c\mu - 1) - \beta^{\cancel{2}} c\mu = 0$$

$$2\beta c\mu - 2 - \beta c\mu = 0$$

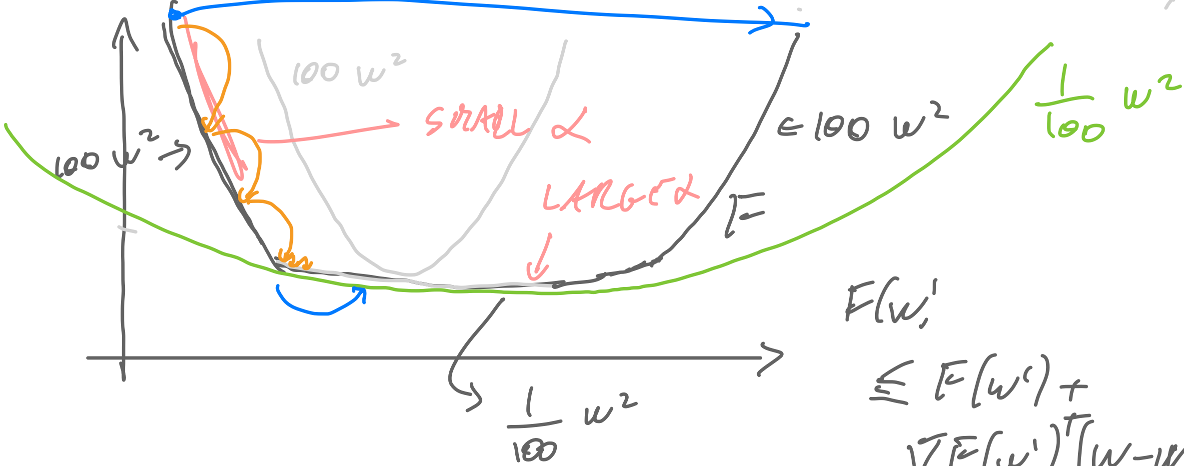
$$\beta c\mu = 2 \qquad \beta = \frac{\cancel{1}\,2}{c\mu} \Rightarrow \text{minimize } V$$

$$V = \frac{\frac{\cancel{4}\,2}{(c\mu)^2} L r}{\cancel{2}\left(\frac{2}{\cancel{c\mu}}\cancel{c\mu} - 1\right)} = \frac{2 L r}{c^2\mu^2} =$$

$$= \frac{2}{\mu^2} \frac{L}{c} \frac{r}{c}$$

$\frac{L}{c}$: CONDITION NUMBER

"THE PROBLEM IS ILL-CONDITIONED" $= \frac{L}{c}$ IS LARGE



$$F(w) \leq F(w') + \nabla F(w')^T (w - w') + \frac{1}{2} L \|w - w'\|^2$$

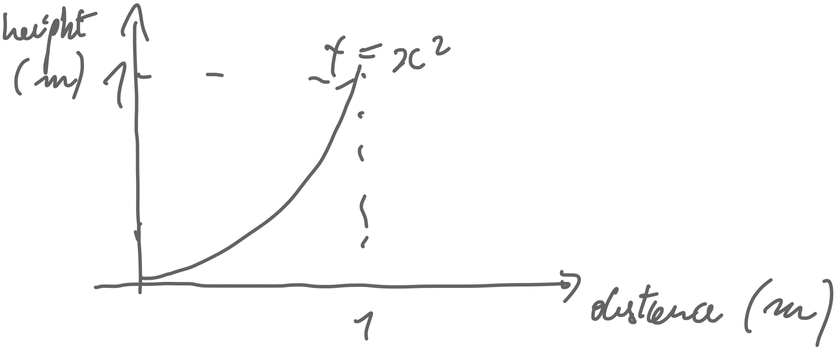
∇F is 200-Lipschitz

$$L = 200$$

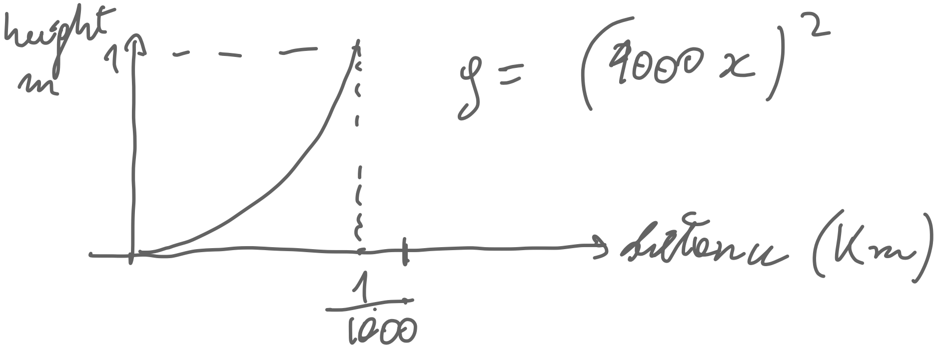
$$F(w) \geq F(w') + \nabla F(w')^T (w - w') + \frac{1}{2} c \|w - w'\|^2$$

F is 50-strongly convex $\quad c = \frac{1}{50}$

$$\frac{L}{c} = \frac{200}{\frac{1}{50}} = 10.000$$



$y = x^2$ $\quad L = 2 \quad c = 2$



$y = (1000x)^2$

$L = 2 \times 10^6 \quad c = 2 \times 10^6$

$$\frac{L}{c} = 1$$

LOOKING - FORWARDS :
THE SOLUTION WILL BE TO TAKE A LEARNING RATE THAT IS FUNCTION OF THE CURVATURE

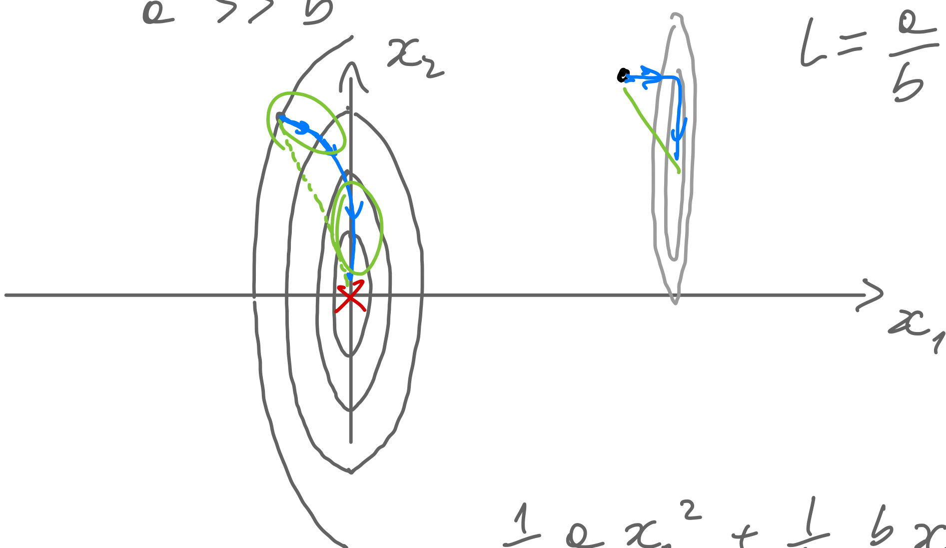
$$f = \frac{1}{2} a x_1^2 + \frac{1}{2} b x_2^2$$

$$L = \max(a,b) \qquad c = \min(a,b)$$

$$f = \frac{1}{2} a x_1^2 + \frac{1}{2} b x_2^2 \leq \frac{1}{2} \max(a,b)(x_1^2 + x_2^2)$$

$$= \frac{1}{2} L (x_1^2 + x_2^2)$$

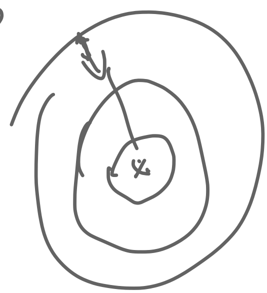
$a >> b$



$L = \frac{a}{b} >> 1$

$$\frac{1}{2} a x_1^2 + \frac{1}{2} b x_2^2 = const$$

$a = b$



$$\frac{L}{c} = 1$$

$$F(x_1, x_2) = \frac{a}{2} x_1^2 + \frac{b}{2} x_2^2$$

$$\nabla F = \begin{bmatrix} a x_1 \\ b x_2 \end{bmatrix}$$

$$\begin{bmatrix} w_{k+1,1} \\ w_{k+1,2} \end{bmatrix} = \begin{bmatrix} w_{k,1} \\ w_{k,2} \end{bmatrix} - \alpha \begin{bmatrix} a & w_{k,1} \\ b & w_{k,2} \end{bmatrix}$$

$$= \begin{cases} w_{k+1,1} = w_{k,1} - \alpha a \, w_{k,1} \\ w_{k+1,2} = w_{k,2} - \alpha b \, w_{k,2} \end{cases}$$

$$a >> b$$

$$\alpha_1 = \frac{\alpha}{a} \qquad \alpha_2 = \frac{\alpha}{b}$$

$$\begin{cases} w_{k+1,1} = w_{k,1} - \alpha \, w_{k,1} \\ w_{k+1,2} = w_{k,2} - \alpha \, w_{k,2} \end{cases}$$

$$w_{k+1} = w_k - \alpha \begin{bmatrix} \frac{1}{a} & 0 \\ 0 & \frac{1}{b} \end{bmatrix} \nabla F(w_k)$$

$$= w_k - \alpha \; H \; \nabla F(w_k)$$

$$g_k = \frac{1}{n_k} H_k \sum_{i=1}^{n_k} \nabla f(w_k, \xi_{k,i})$$

H "Hessian"

$$\text{Hessian } \cancel{f} = \begin{bmatrix} \frac{\partial^2 F}{\partial x_1^2} & \frac{\partial^2 F}{\partial x_1 \partial x_2} \\ \frac{\partial^2 F}{\partial x_1 \partial x_2} & \frac{\partial^2 F}{\partial x_2^2} \end{bmatrix} =$$

$F(x_1, x_2) =$
$= \frac{1}{2} a x_1^2 +$
$\frac{1}{2} b x_2^2$

$$= \begin{bmatrix} a & 0 \\ 0 & b \end{bmatrix}$$

$$\begin{bmatrix} \frac{1}{a} & 0 \\ 0 & \frac{1}{b} \end{bmatrix} = \begin{bmatrix} a & 0 \\ 0 & b \end{bmatrix}^{-1} = H_F^{-1}$$

$$w_{k+1} = w_k - \alpha_k H_F^{-1} \nabla F(w_k)$$

Newton Method

MUCH FASTER THAN BASIC GRADIENT

$H_F = d \times d$ $\quad$ d : # parameters in the model.

memory!

H_F^{-1} COMPLEXITY d^3

$d^{2.37}$

$\underbrace{H_F^{-1} \nabla F}$ REALLY WHAT YOU NEED

COMPUTE IT BY SOLVING

$$H_F \underbrace{s_k} = \nabla F \qquad s_k = H_F^{-1} \nabla F$$

$s_k \simeq$ A SOLUTION

$\Rightarrow$ CONJUGATE GRADIENT METHOD

$$\begin{bmatrix} \frac{1}{e} & 0 \\ 0 & \frac{1}{b} \end{bmatrix} = H_F^{-1}$$

$$\begin{bmatrix} \frac{1}{\frac{\partial^2 F}{\partial x_1^2}} & 0 & 0 & 0 & 0 \\ 0 & \frac{1}{\frac{\partial^2 F}{\partial x_2^2}} & 0 & 0 & 0 \\ - & - & - & - & \\ 0 & 0 & 0 & 0 \cdots & \frac{1}{\frac{\partial^2 F}{\partial x_d^2}} \end{bmatrix}$$

$$H_F = \begin{bmatrix} \frac{\partial^2 F}{\partial x_1^2} & 0 & 0 \\ 0 & \ddots & 0 \\ 0 & \ddots & \frac{\partial^2 F}{\partial x_d^2} \end{bmatrix}$$

TRUE ONLY IF

$$F(x_1 \dots x_d) = \sum_{i=1}^{d} f_i(x_i)$$

$$f(w,(x,y)) = \ell(h(w,x),y)$$

$$\simeq \ell(h(w',x) + J_h(w',x)(w-w'),y)$$

H_f

GAUSS-NEWTON METHOD

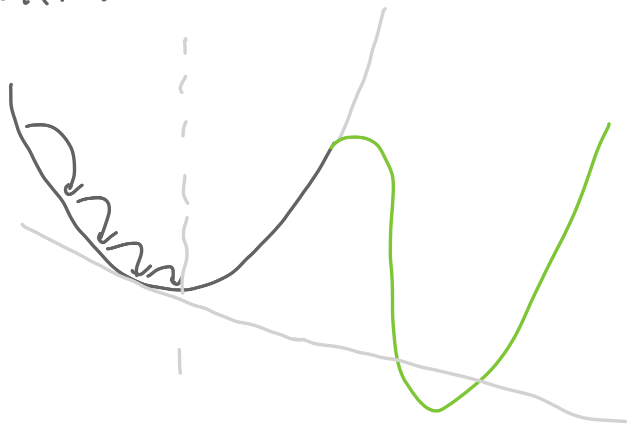
Iteratively update of H_k

$$H_k = G(H_{k-1}, w_k)$$

$\boxed{\text{BFGS}}$

UNTIL NOW WE CONSIDERED CONVEXITY

WHAT IF NO CONVEXITY?



$$\cancel{\mathbb{E}\left[F(w_k)\right] - F^* \le \text{SOMETHING}}$$

OPTIMUM

Theorem (4.8, Bottou)

ASSUMPTIONS ABOUT NOISE AND L-SMOOTHNESS

WE GIVE UP CONVEXITY.

$$\alpha \le \frac{\mu}{L M_G}$$

$$\mathbb{E}\left[\sum_{k=1}^{K} \|\nabla F(w_k)\|^2\right] \le \frac{K \alpha L M}{\mu} + \frac{2(F(w_1) - F_{inf})}{\mu \alpha}$$

$$\frac{1}{K}\mathbb{E}\left[\sum_{k=1}^{K}\|\nabla F(w_k)\|^2\right] \le \frac{\alpha L M}{\mu} + \frac{2(F(w_1)-F_{inf})}{K\mu\alpha}$$

$$F_{inf} \le F(w) \quad \forall w$$

$$F_{inf} = F^* = \min_{w} F(w)$$

SIMILAR RESULT FOR DECREASING α_k

$$\alpha_k = \frac{1}{K}$$