OPTIMAL Lecture : 02/02/2021

Recap: $w_{k+1} = w_k - \alpha_k\, g(w_k, \xi_k)$ zeta

$$\underbrace{g(w_k, \xi_k) = \frac{1}{n_k} \sum_{i=1}^{n_k} \nabla f(w_k, \xi_{k,i})}$$

$n_k = 1 \Rightarrow$ SGD

$n_k = |S| \Rightarrow$ FB

$1 < n_k < |S| \Rightarrow$ MB

Assuption :

We say F is L-smooth or the gradient of F is Lipchitz-continuous if:

$$\| \nabla F(x) - \nabla F(y) \| \leq L \| x - y \| \quad , \forall x, y$$

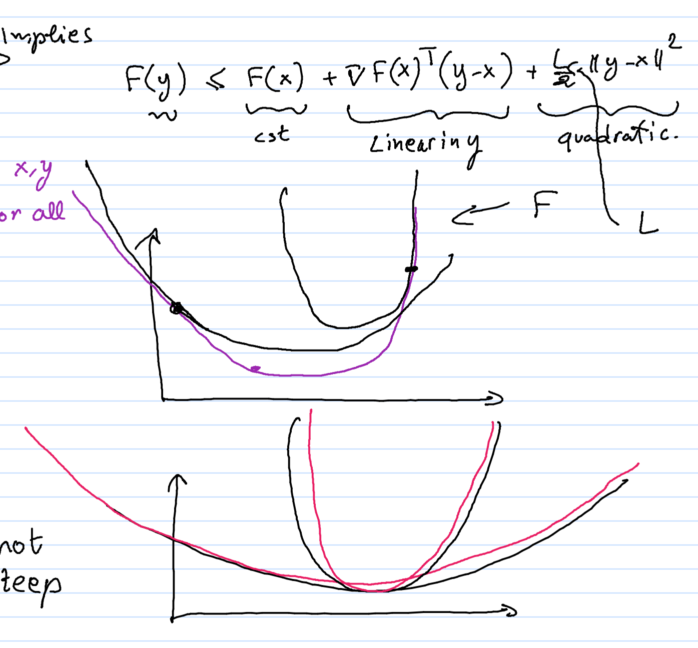
$\|x\|$: Euclidean norm $\|x\| = \sqrt{\sum_i x_i^2}$

$\xRightarrow{\text{Implies}}$

$$\tilde{F(y)} \leq \underbrace{F(x)}_{cst} + \underbrace{\nabla F(x)^T (y-x)}_{\text{Linearing}} + \underbrace{\frac{L}{2} \| y-x \|^2}_{\text{quadratic.}}$$

$\forall$ x, y

$\forall$: for all



Lemma 4.2

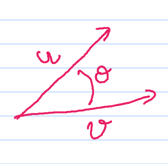
$$\mathbb{E}_{\xi_k}[F(w_{k+1})] - F(w_k) \leq -\alpha_k \nabla F(w_k)^T \mathbb{E}_{\xi_k}[g(w_k, \xi_k)] + \frac{1}{2} \alpha_k^2 L\, \mathbb{E}_{\xi_k}[\| g(w_k, \xi_k) \|^2]$$

Proof. $w_{k+1} - w_k = -\alpha_k\, g(w_k, \xi_k)$

$$F(\underbrace{w_{k+1}}_{y}) \leq F(\underbrace{w_k}_{x}) + \nabla F(\underbrace{w_k}_{x})^T (\underbrace{w_{k+1} - w_k}_{y-x}) + \frac{L}{2} \| w_{k+1} - w_k \|^2$$

$$= F(w_k) - \alpha_k \nabla F(w_k)^T g(w_k, \xi_k) + \frac{L}{2} \alpha_k^2 \| g(w_k, \xi_k) \|^2$$

$$\mathbb{E}_{\xi_k}[F(w_{k+1})] - F(w_k) \leq -\alpha_k \underbrace{\nabla F(w_k)^T \mathbb{E}[g(w_k, \xi_k)]}_{+} + \frac{L}{2} \underbrace{\alpha_k^2}_{+} \underbrace{\mathbb{E}[\| g(w_k, \xi_k) \|^2]}_{+}$$



$\underbrace{u^T v}_{\text{scalar product}} = \|u\| \|v\| \cos\theta$

if u, v are oriented in the same direction

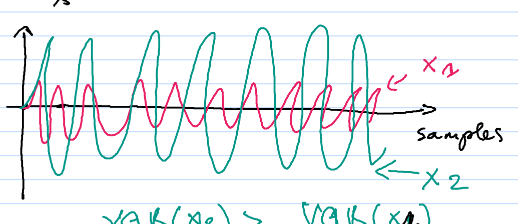
then $u^T v > 0$

$\mathbb{E}_{\xi_k}[g(w_k, \xi_k)]$: should be aligned with ∇F.

- α_k small enough.

- $\underbrace{\mathbb{V}_{\xi_k}}(g(w_k, \xi_k)) := \mathbb{E}[\| g(w_k, \xi_k) \|^2] - \| \mathbb{E}[g(w_k, \xi_k)] \|^2$
- $\mathrm{var}(X) = \mathbb{E}[X^2] - \mathbb{E}^2[X]$

$\mathbb{E}[x_2] = \mathbb{E}[x_1]$



$\mathrm{VAR}(x_2) \geq \mathrm{VAR}(x_1)$

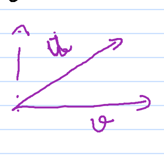
Assumption 4.3:

1) technical: $\exists$ an open set J containing all sequences $\{w_k\}$. F is Bounded below by F_{inf} ← infimum. (F has a lower bound).

2). $\exists\, \mu_G \geq \mu > 0$ such that $\forall k \in \mathbb{N}$.

$$\nabla F^T(w_k)\, \mathbb{E}[g(w_k, \xi_k)] \geq \mu \| \nabla F(w_k) \|^2 \quad (a)$$

$$\mathbb{E}[g(w_k, \xi_k)] \leq \mu_G \| \nabla F(w_k) \| \quad (b)$$



$u^T v = \|u\| \|v\| \cos\theta \geq 0$ when they are orthogonal.

If $g(w_k, \xi_k) = \nabla F(w_k)$: $n_k = |S|$

$\nabla F^T g = \nabla F^T \nabla F = \| \nabla F \|^2$

$\leq \| \nabla F \|^2 \; (\mu = 1)$

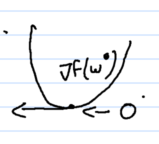
$(\mu_G = 1)$

c)

$$\mathbb{V}_{\xi_k}[g(w_k, \xi_k)] \leq M + M_V \| \nabla F(w_k) \|^2$$

if w^* is at the critical region.

$\nabla F(w^*) = 0$



$\mathbb{V}_{\xi_k}[g(w^*, \xi_k)] \leq M$ (M constant variability at critical regions

Recap : we introduced

- L-smothness on F. (Lipchitz continuity of ∇F)
- $\underbrace{\mu, \mu_G}_{\text{alignments.}}, \underbrace{M, M_V}_{\text{noise constants.}}$

$$\mathbb{E}[\| \underset{g(w_k, \xi_k)}{g} \|^2] = \underbrace{\mathbb{V}[g]} - \underbrace{\| \mathbb{E}[g] \|^2} \leftarrow (b)$$

$$\leq M + M_V \| \nabla F \|^2 + \mu_G^2 \| \nabla F \|^2$$

$$\leq M + \underbrace{M_G}_{M_V + \mu_G^2} \| \nabla F \|^2$$

Lemma 4.4:

Under assumptions 4.1 & 4.3 we have for all $k \in \mathbb{N}$.

$$\mathbb{E}_{\xi_k}[F(w_{k+1})] - F(w_k) \leq -\left(\mu - \frac{1}{2} \alpha_k L M_G\right) \alpha_k \| \nabla F(w_k) \|^2 + \frac{1}{2} \alpha_k^2 L M.$$

Proof. $k \to k+1$

$$\mathbb{E}_{\xi_k}[F(w_{k+1})] - F(w_k)$$

$$\leq -\alpha_k \nabla F^T \underbrace{\mathbb{E}_{\xi_k}[g]}_{\geq \mu \| \nabla F \|} + \frac{1}{2} \alpha_k^2 L\, \underbrace{\mathbb{E}_{\xi_k}[\| g \|^2]}_{\leq M + M_G \| \nabla F \|^2}$$

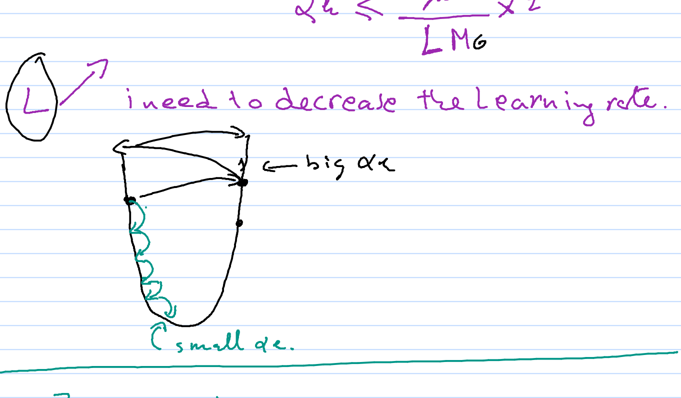
$$\leq -\alpha_k \mu \| \nabla F \|^2 + \frac{1}{2} \alpha_k^2 M L + \frac{1}{2} \alpha_k^2 L M_G \| \nabla F \|^2$$

$$\leq \underbrace{\underbrace{-\alpha_k}_{-} \underbrace{\left(\mu - \frac{1}{2} \alpha_k L M_G\right)}_{+} \underbrace{\| \nabla F \|^2}_{+}}_{-} + \underbrace{\frac{1}{2} \alpha_k^2 M L}_{+}$$

$$\mu - \frac{1}{2} \alpha_k L M_G \geq 0$$

$$\alpha_k \leq \frac{\mu}{L M_G} \times 2$$

$L \nearrow$ i need to decrease the Learning rate.



$M_G \nearrow$ I need $\searrow \alpha_k$ to be small.

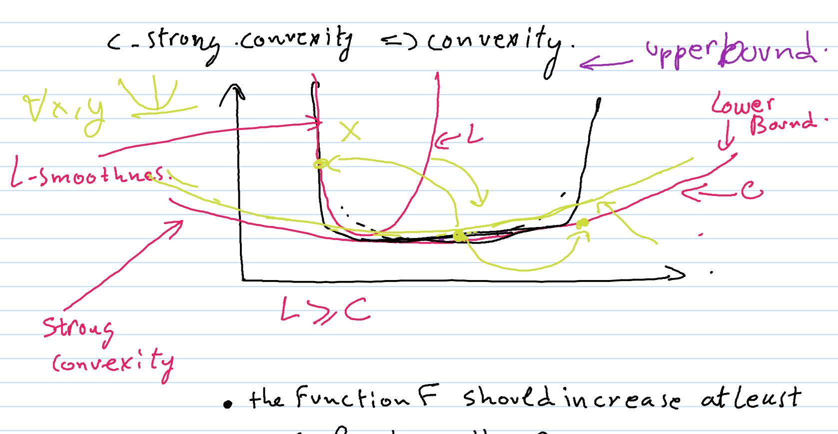
when $\mu \nearrow \Rightarrow \alpha_k \nearrow$.

Assumption 4.5 (Strong-convexity): $\underset{\sim}{c}$-strong convexity

for L-smothnes ($\leq$)

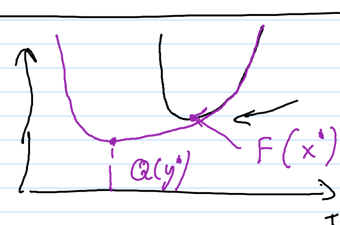
$$F(y) \geq F(x) + \nabla F(y)^T (y-x) + \frac{1}{2} c \| y - x \|^2$$

c-strong convexity $\Rightarrow$ convexity. ← upper bound.



$L \geq C$

- the function F should increase at least as fast as the lower bound. quadratic.
- // // F shouldn't increase faster than the upper Bound.



$\underbrace{Q(y^*) \leq F(x^*)}$

$$Q(y) = \underbrace{F(x)}_{\geq 0} + \nabla F(x)^T (\tilde{y} - x) + \frac{1}{2} c \| y - x \|^2$$

$$\nabla_y Q(y) = \nabla F(x) + \frac{1}{\cancel{2}} \times \cancel{2}\, (y - x) \times c$$

$$\nabla_y Q(y^*) = 0 = \nabla F(x) + (y^* - x)\, c$$

$$y^* = -\frac{\nabla F(x)}{c} + x \rightarrow \min \text{ of } Q.$$

$$\overset{\leq F(x^*)}{Q(y^*)} = F(x) + \nabla F(x)^T \left(-\frac{\nabla F(x)}{c} + \cancel{x} - \cancel{x} \right) + \frac{1}{2} \frac{\cancel{c}}{c^2} \| \nabla F(x) \|^2$$

$$\boxed{F(x) - F(x^*) \leq \frac{1}{2c} \| \nabla F(x) \|^2} \leftarrow \text{property. (1)}$$

Theorem 4.6 (Fixed step size convergence result)

$\boxed{\alpha_k = \alpha}$ it's kept constant. $F^* = \min F(x)$.

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

$\xrightarrow{k \to \infty} \frac{\alpha L M}{2 c \mu}$, with $\boxed{0 < \alpha \leq \frac{\mu}{L M_G}}$

$$0 < \alpha \leq \frac{\mu}{L M_G} \quad (2)$$

$\boxed{\text{Proof}}$ use Lemma (4.4) + (1) + (2)

$$\mathbb{E}_{\xi_k}[F(w_{k+1}) - F(w_k)]$$

$$\leq \underbrace{-\left(\mu - \underbrace{\frac{1}{2} \alpha L M_G}_{\geq}\right)}_{\leq} \alpha \| \nabla F \|^2 + \frac{1}{2} \alpha^2 L M$$

$$\overset{(2)}{\leq} -\left(\mu - \frac{1}{2} \mu\right) \alpha \| \nabla F \|^2 + \frac{1}{2} \alpha^2 L M$$

$$\leq -\frac{1}{2} \mu \alpha \| \nabla F \|^2 + \frac{1}{2} \alpha^2 L M.$$

$$\mathbb{E}_{\xi_k}[F(w_{k+1})] - F(w_k) \overset{(1)}{\leq} -\frac{1}{\cancel{2}} \mu \alpha\, \cancel{2} c\, \mathbb{E}_{\xi_k}[F(w_k) - F^*] + \frac{1}{2} \alpha^2 L M.$$

$- \alpha L M / 2 c \mu$. $= (1 - c \mu \alpha)(F(w_k) - F^*)$

$$\mathbb{E}_{\xi_k}[F(w_{k+1}) - F^*] \leq \mathbb{E}_{\xi_k}[F(w_k) - F^*] - \mu \alpha c\, \mathbb{E}(F(w_k) - F^*) + \frac{1}{2} \alpha^2 L M. - \frac{\alpha L M}{2 c \mu}$$

$$\mathbb{E}_{\xi_k}\left[F(w_{k+1}) - F^* - \frac{\alpha L M}{2 c \mu}\right]$$

$$\leq // + \frac{1}{2} \alpha^2 L M - \frac{\alpha L M}{2 c \mu}.$$

$$= \frac{1}{2} \alpha L M (c \mu \alpha - 1)$$

$$= -\frac{1}{2} \frac{\alpha L M}{c \mu} (1 - c \mu \alpha)$$

at $k+1$

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

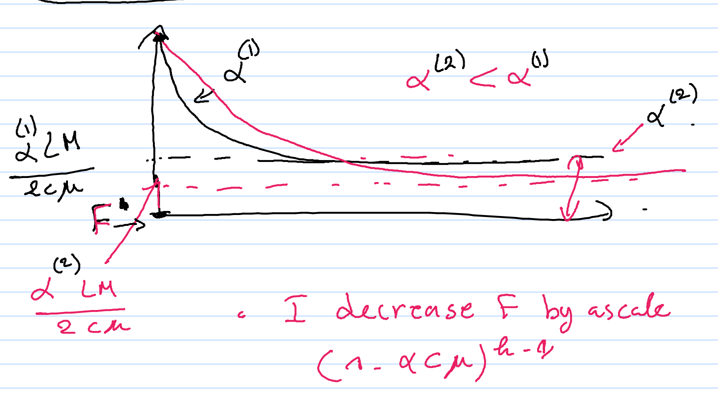
$$\leq (1 - c \mu \alpha)^2 \left(\overset{\text{at } k}{\mathbb{E}_{\xi_k}[F(w_{k-1}) - F^*]} - \frac{1}{2} \frac{\alpha L M}{c \mu} \right)$$

at $k-1$

... up to $k = 1$

$$\leq \overset{\leq 1}{(1 - c \mu \alpha)}^{\cancel{k}} \left(\mathbb{E}_{\xi_k}[F(w_{\underbrace{k-(k-1)}_{1}}) - F^*] - \frac{1}{2} \frac{\alpha L M}{c \mu} \right)$$

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



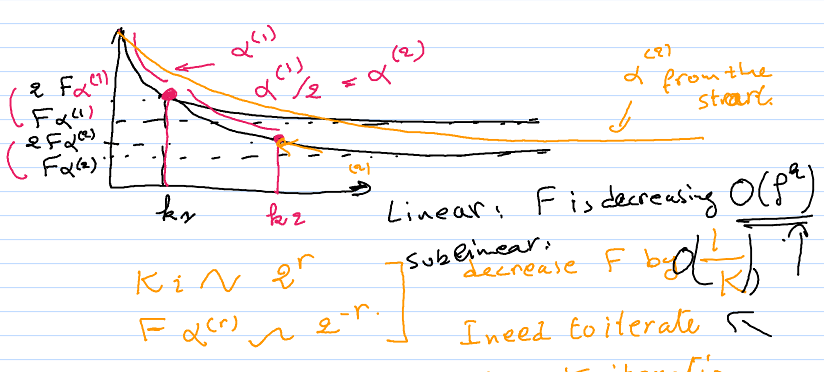
- I decrease F by a scale $(1 - \alpha c \mu)^{k-1}$
- and I converge when $k \to \infty$

to $\frac{\alpha L M}{2 c \mu}$.) M constant noise you get in critical region "minimum".

$L \to$

$\alpha \nearrow \rightarrow$ I decrease F faster because $(1 - \alpha c \mu) \Rightarrow \searrow$ smaller

I converge faster but to a bad value of F.



Linear: F is decreasing $O(\rho^k)$

$\left. \begin{array}{l} K_i \sim 2^r \\ F_{\alpha^{(r)}} \sim 2^{-r}. \end{array} \right]$ sublinear: decrease F by $O\left(\frac{1}{K}\right)$

I need to iterate for $\underbrace{K \text{ iteration}}$

= sublinear-convergence.

Eliminate completly the gap from F^* and $F(w_k)$.

$K = \left(\frac{L}{c}\right)$ the condition number

$\frac{L}{c} \geq 1$ the closer $\frac{L}{c} \to 1$ the easier the problem is.

if $\frac{L}{c} \gg 1$ it's an ill-conditioned problem.