Last week:

① binary classification with Φ-loss

$$\ell(y, f(x)) = \Phi(y f(x))$$

f: real valued function

$\Phi(u)$:
- $\log(1+e^{-u})$ logistic
- $(u-1)^2$
- $(1-u)_+$ SVM

② Estimation error f_θ models, $\theta \in \Theta$

$\hat\theta$ estimator $\in \Theta$

$$\underbrace{R(f_{\hat\theta})}_{\text{expected risk}} - \inf_{\theta\in\Theta} R(f_\theta) \leq 2 \overbrace{\sup_{\theta\in\Theta} \left| R(f_\theta) - \hat{R}(f_\theta)\right|}^{\text{uniform deviation}} \quad \frac{1}{\sqrt{n}}$$

$$+ \underbrace{\hat{R}(f_{\hat\theta}) - \inf_{\theta\in\Theta} \hat{R}(f_\theta)}$$

Focus of today : objective function

$F(\theta)$ $\quad F: \mathbb{R}^d \to \mathbb{R}$

ex: ① $F(\theta) = \frac{1}{n} \sum_{i=1}^{n} \ell(y_i, f_\theta(x_i)) + \Omega(\theta)$

↳ minimizer η_*

② $F(\theta) = \mathbb{E}[\ell(y, f_\theta(x))]$

minimiser: θ_*

⚠ dimension-independent

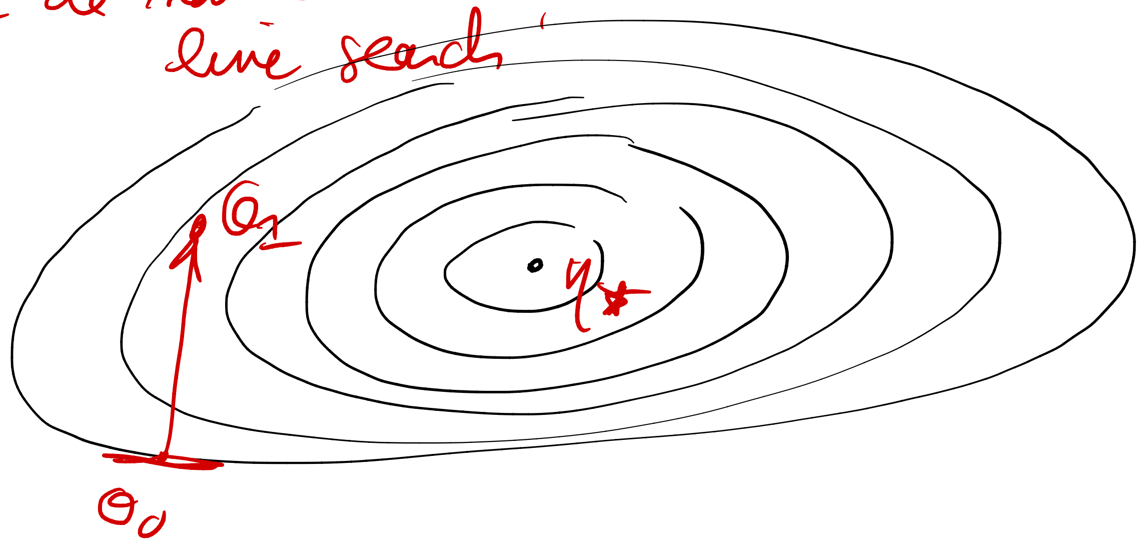
---

gradient descent : Pick $\theta_0 \in \mathbb{R}^d$

iteration : $\theta_t = \theta_{t-1} - \gamma_t \underbrace{F'(\theta_{t-1})}$

gradient = steepest descent direction

⚠ we do not consider line search !



First analysis: least-squares

$\Phi \in \mathbb{R}^{n\times d}$ design / data matrix

$y \in \mathbb{R}^n$ response vector

$$F(\theta) = \frac{1}{2n}\|y-\Phi\theta\|_2^2, \quad F'(\theta) = \frac{1}{n}\Phi^T(\Phi\theta - y)$$

$$= \frac{1}{2n}\|y\|_2^2 + \frac{1}{2n}\theta^T\Phi^T\Phi\theta - \frac{1}{n}y^T\Phi\theta$$

$$F'(\theta) = 0 \quad \frac{1}{n}\Phi^T\Phi\theta \quad -\frac{1}{n}\Phi^T y$$

gradient $\in \mathbb{R}^d$

$$F''(\theta) = \frac{1}{n}\Phi^T\Phi = H$$

Hessian matrix $\in \mathbb{R}^{d\times d}$

optimality conditions: $F'(\eta_*) = 0 \Leftrightarrow H\eta_* = \frac{1}{n}\Phi^T y$ "normal" equations

Excess objective function: $F(\theta) - F(\eta_*) = \cancel{F'(\eta_*)^T(\theta-\eta_*)}$

$$+ \boxed{\frac{1}{2}(\theta-\eta_*)^T H(\theta-\eta_*)}$$

Taylor expansion

Eigenvalues of $H \subset [\mu, L]$. μ smallest ≥ 0, L largest, $\kappa = \frac{L}{\mu} \geq 1$ condition number condition number $= \dfrac{\lambda_{max}(H)}{\lambda_{min}(H)}$ and $\lambda_{min}(H) = \dfrac{1}{\lambda_{max}(H^{-1})}$

$$H = \sum_{i=1}^{d} \lambda_i \underbrace{u_i}_{\text{eigenvector}} u_i^T, \quad H^{-1} = \sum_{i=1}^{n} \frac{1}{\lambda_i} u_i u_i^T$$

$$\lambda_{max}(H) \cdot \lambda_{max}(H^{-1})$$

GD: $$\theta_t = \theta_{t-1} - \underbrace{\gamma}_{\text{constant}} F'(\theta_{t-1}) = \theta_{t-1} - \gamma\Big[H\theta_{t-1} - \underbrace{\tfrac{1}{n}\Phi^T y}_{H\eta_*}\Big]$$

$$\theta_t = \theta_{t-1} - \gamma H\left[\theta_{t-1} - \eta_*\right]$$

$$\theta_t - \eta_* = (I - \gamma H)\left[\theta_{t-1} - \eta_*\right] = (I-\gamma H)^t(\theta_0 - \eta_*)$$

$$\Rightarrow F(\theta_t) - F(\eta_*) = \frac{1}{2}(\theta_t - \eta_*)^T H(\theta_t - \eta_*)$$

$$= \frac{1}{2}(\theta_0 - \eta_*)^T \underbrace{(I-\gamma H)^t H (I-\gamma H)^t}_{H(I-\gamma H)^{2t}} (\theta_0 - \eta_*)$$

$$= \frac{1}{2}(\theta_0 - \eta_*)^T \sum_{i=1}^{d} \lambda_i (1-\gamma\lambda_i)^{2t} u_i u_i^T (\theta_0 - \eta_*) = \frac{1}{2}\sum_{i=1}^{d} \lambda_i (1-\gamma\lambda_i)^{2t} \cdot \left[u_i^T(\theta_0 - \eta_*)\right]^2$$

$$F(\theta_t) - F(\eta_*) = \frac{1}{2}\sum_{i=1}^{d} \lambda_i (1-\gamma\lambda_i)^{2t} \left[u_i^T(\theta_0 - \eta_*)\right]^2$$

convergence if $|1-\gamma\lambda_i| \le 1$ for all i.

remember that $\lambda_i \in [\mu, L]$

we consider $\gamma = \frac{1}{L} \Rightarrow 1-\gamma\lambda_i \in [0,1]$

<u>Analysis-1</u>: $\lambda_i \ge \mu \Rightarrow 1-\gamma\lambda_i \le 1-\gamma\mu = 1-\frac{\mu}{L}$

$$F(\theta_t) - F(\theta_*) \le \underbrace{\frac{1}{2}\sum_{i=1}^{d} \lambda_i \left[u_i^T(\theta_0 - \eta_*)\right]^2}_{F(\theta_0) - F(\theta_*)} \underbrace{\left(1-\frac{\mu}{L}\right)^{2t}}$$

$$\Rightarrow F(\theta_t) - F(\eta_*) \le \left(1-\frac{\mu}{L}\right)^{2t}\left[F(\theta_0) - F(\eta_*)\right]$$

$$F(\theta_t) - F(\eta_*) \leq \underbrace{\left(1-\frac{\mu}{L}\right)^{2t}}_{\varepsilon} \left[F(\theta_0) - F(\eta_*)\right]$$

vanishes if $\mu = 0$

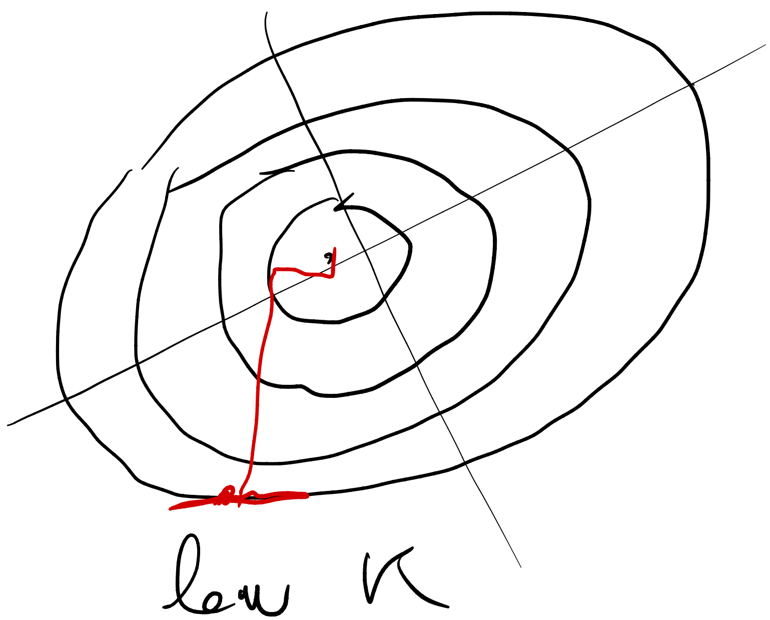
exponential convergence
linear
geometric

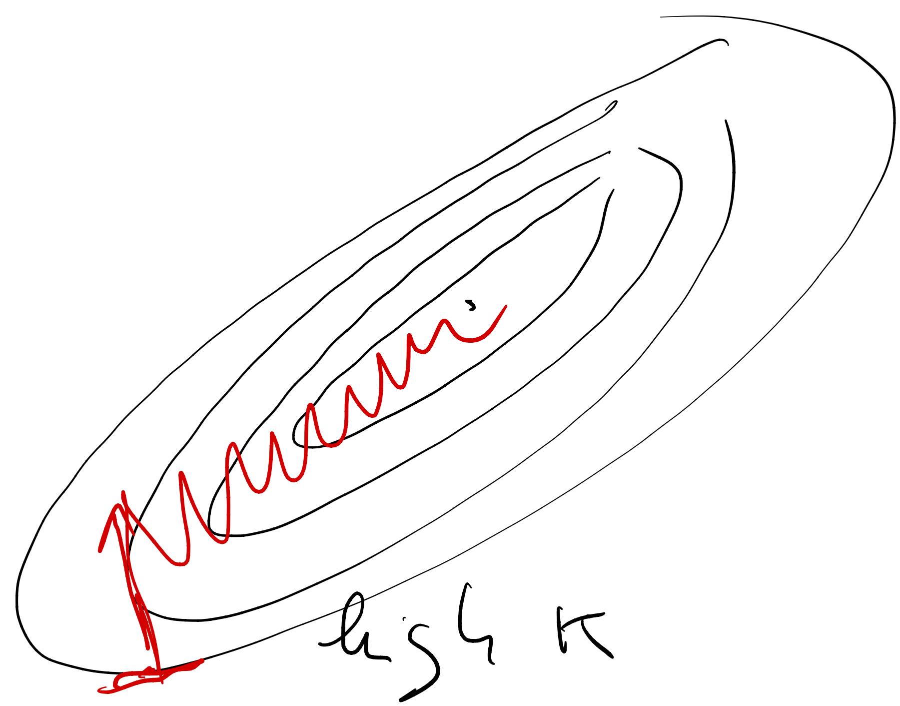
$$\varepsilon = \left(1-\frac{\mu}{L}\right)^{2t} \leq \left(e^{-\mu/L}\right)^{2t} \text{ because } 1-\alpha \leq e^{-\alpha} \text{ for } \alpha \geq 0$$

$$= e^{-\frac{2\mu t}{L}} = e^{-2t/\kappa}$$

$\kappa = \frac{L}{\mu}$ condition number

$$\Leftrightarrow t \geq \frac{\kappa}{2} \log\left(\frac{1}{\varepsilon}\right)$$



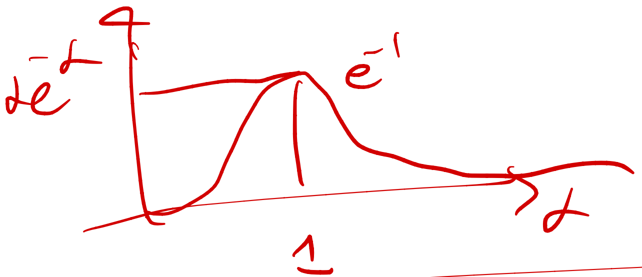


$$F(\theta_t) - F(\eta_*) = \frac{1}{2}\sum_{i=1}^{d} \lambda_i (1-\gamma\lambda_i)^{2t} \left[u_i^T(\theta_0 - \eta_*)\right]^2$$

Analysis-2: $\lambda_i (1-\gamma\lambda_i)^{2t} \le \lambda_i \left(e^{-\gamma\lambda_i}\right)^{2t}$

$$= \frac{\lambda_i 2\gamma t \, e^{-\gamma\lambda_i 2t}}{2\gamma t} \le \frac{1}{2}\,\frac{1}{2\gamma t} = \frac{1}{4\gamma t}$$

Lemma: $\max_{x \ge 0} x e^{-x} = e^{-1} \le \frac{1}{2}$



Consequence: $F(\theta_t) - F(\eta_*) \le \frac{1}{8\gamma t} \sum_{i=1}^{n} \left(u_i^T(\theta_0 - \eta_*)\right)^2$

$(\theta_0 - \theta_*)^T u_1 u_1^T (\theta_0 - \theta_*)$

$$(\theta_0 - \eta_*)^T \underbrace{\sum_{i=1}^{d} u_i u_i^T}_{I} (\theta_0 - \eta_*) = \|\theta_0 - \eta_*\|_2^2$$

$$\le \frac{L \|\theta_0 - \eta_*\|_2^2}{8t}$$

sublinear rate

μ / L for Machine learning : $F(\theta) = \frac{1}{2n} \|y - \Phi\theta\|_2^2$

$H = \frac{1}{n} \Phi^T \Phi = \hat{\Sigma}$ empirical covariance matrix.

if $n < d$, $\mu = 0$

if $n \geq d$, μ is very small

For ridge regression, $F(\theta) = \frac{1}{2n} \|y - \Phi\theta\|_2^2 + \frac{\lambda}{2} \|\theta\|_2^2$

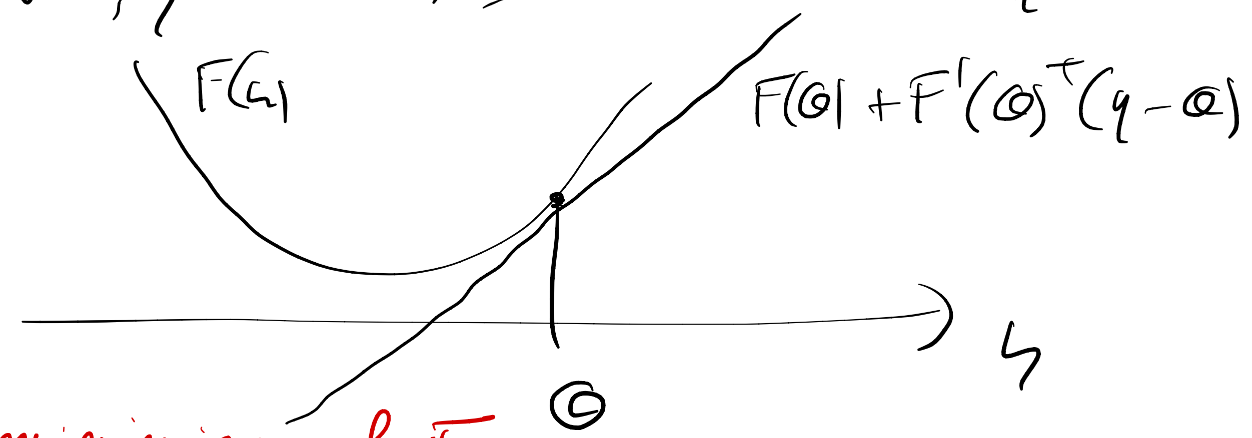
$H = \hat{\Sigma} + \lambda I \Rightarrow \mu \geq \lambda$

$\boxed{10^{-4} 17 - 10^{-4} 37}$

ex: $\lambda \simeq \frac{1}{\sqrt{n}}, \frac{1}{n}$

Convex functions : $F : \mathbb{R}^d \to \mathbb{R}$ differentiable

- F is convex $\iff$ $\forall \theta, \eta$ $\quad F(\eta) \geq F(\theta) + F'(\theta)^T(\eta - \theta)$



lemma : if η_* is a minimizer of F

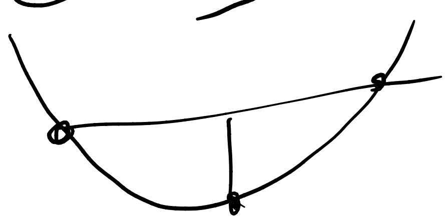
$$F(\theta) - F(\eta_*) \leq F'(\theta)^T(\theta - \eta_*)$$

$\Rightarrow$ ($F'(\theta) = 0$ $\iff$ θ minimizer)

- Jensen's inequality

if F is convex, θ random vector

$$\mathbb{E}[F(\theta)] \geq F(\mathbb{E}\theta)$$



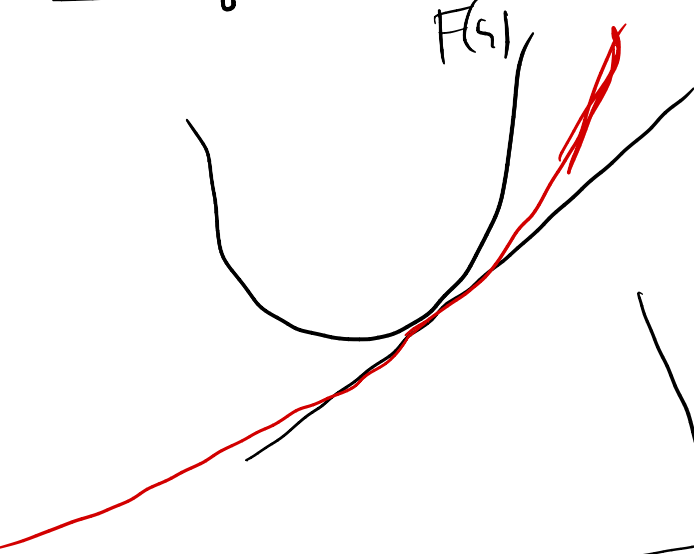
$F(\theta) = \mathbb{E}\left[\ell(y, f_\theta(x))\right]$

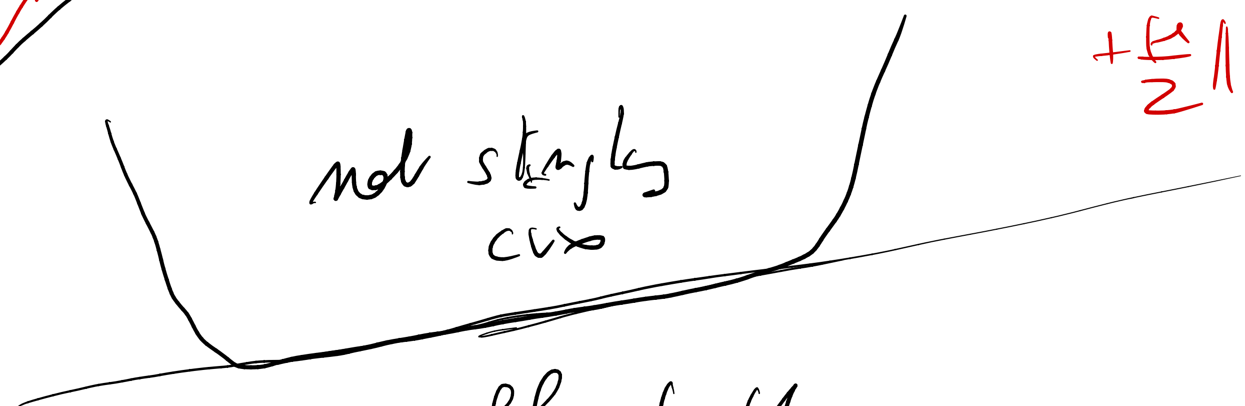
- for ML : F is cvx

if loss is cvx and f_θ = linear in θ

Strong convexity : F is μ-strongly-cvx iff $\forall \theta, \eta$

$$F(\eta) \geq F(\theta) + F'(\theta)^T(\eta - \theta) + \frac{\mu}{2}\|\theta - \eta\|_2^2$$





Property : ① if F is twice differentiable

- F is cvx $\iff$ $\forall \theta$ $F''(\theta)$ is positive semi definite $\iff$ $\forall \theta$ $F''(\theta)$ has non negative eigenvalues
- F is μ-strongly cvx $\iff$ $\forall \theta$, $F''(\theta)$ has eigenvalues $\geq \mu$

② F cvx $\Rightarrow$ $\theta \mapsto F(\theta) + \frac{\mu}{2}\|\theta\|^2$ is μ-strongly cvx

③ is $\mu > 0$, $\mu \geq 0$?

ŁOJASIEWICZ's inequality:

if F is μ-strongly cvx then, $\forall \theta$, $\dfrac{\|F'(\theta)\|_2^2}{2\mu} \geq F(\theta) - F(\eta_*)$

proof: definition: $\forall \eta, \theta$, $F(\eta) \geq \underbrace{F(\theta) + F'(\theta)^T(\eta-\theta) + \frac{\mu}{2}\|\theta-\eta\|_2^2}$

minimize wrt η.

$$F(\eta_*) \geq F(\theta) - \frac{1}{2\mu}\|F'(\theta)\|_2^2$$

lemma: $\inf_x a^T x + \frac{\mu}{2}\|x\|_2^2$

$= -\frac{1}{2\mu}\|a\|_2^2$

proof: gradient $= a + \mu x = 0$

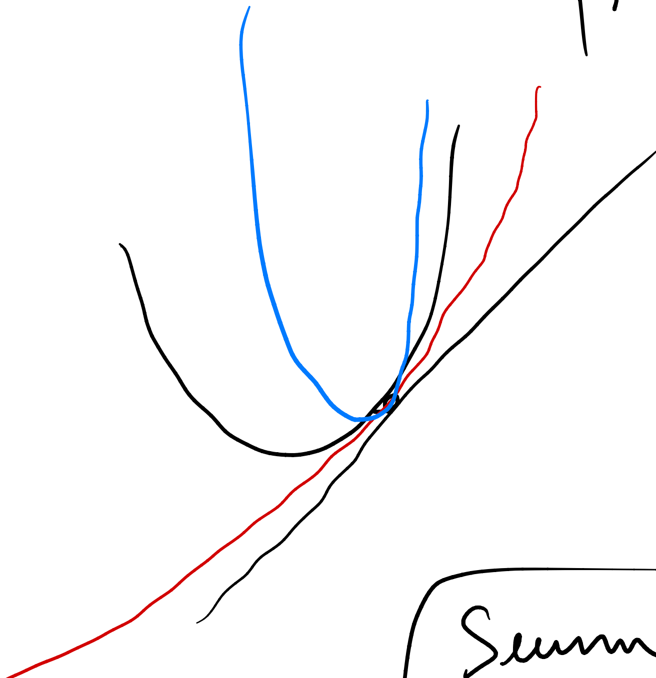
$x = -\frac{a}{\mu}$

then $a^T x + \frac{\mu}{2}\|x\|_2^2 = -\frac{a^T a}{\mu} + \frac{\mu}{2}\left\|\frac{a}{\mu}\right\|_2^2$

$= -\frac{1}{2\mu}\|a\|_2^2$

Smoothness: F is L-smooth $\Leftrightarrow$ $\forall \theta, \eta$

$$\left| F(\eta) - F(\theta) - F'(\theta)^T(\eta - \theta) \right| \leq \frac{L}{2} \|\theta - \eta\|_2^2$$

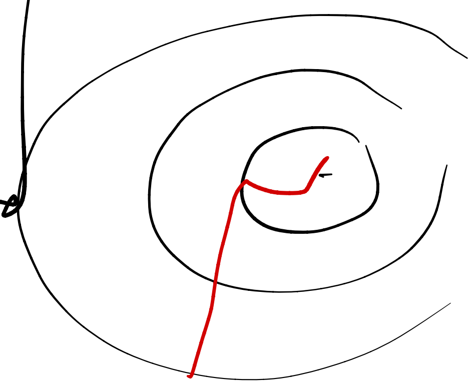


strong convexity

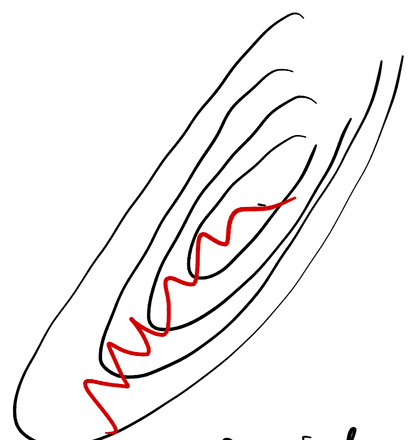
- if F is twice differentiable

F is L-smooth $\Leftrightarrow$ eigenvalues of $F''(\theta) \in [-L, L]$ $\forall \theta$.

Summary: L-smooth, μ-strongly convex $\Big\}$ eigenvalues of $f''(\theta) \in [\mu, L]$



low $\frac{L}{\mu}$



high $\frac{L}{\mu}$

Proof of exponential convergence of GD: $\theta_t = \theta_{t-1} - \underbrace{\frac{1}{L}}_{\text{step size}} F'(\theta_{t-1})$

(1) smoothness

$$F(\theta_t) \le F(\theta_{t-1}) + F'(\theta_{t-1})^T(\theta_t - \theta_{t-1}) + \frac{L}{2}\|\theta_t - \theta_{t-1}\|_2^2$$

$$\le F(\theta_{t-1}) - \frac{1}{L}F'(\theta_{t-1})^T F'(\theta_{t-1}) + \frac{L}{2}\left\|\frac{1}{L} \nabla F'(\theta_{t-1})\right\|_2^2$$

$$F(\theta_t) - F(\eta_*) \le F(\theta_{t-1}) - F(\eta_*) - \frac{1}{2L}\underbrace{\|F'(\theta_{t-1})\|_2^2}_{\ge 2\mu[F(\theta_{t-1}) - F(\eta_*)]}$$

Loja...

$$\Rightarrow F(\theta_t) - F(\theta_*) \le \left(1 - \frac{\mu}{L}\right)\left(F(\theta_{t-1}) - F(\theta_*)\right)$$

$$\boxed{F(\theta_t) - F(\theta_*) \le \left(1 - \frac{\mu}{L}\right)^t \left(F(\theta_0) - F(\theta_*)\right)}$$

ADAPTIVE

NB: we also have $\le \frac{L}{2t}\|\theta_0 - \theta_*\|_2^2$

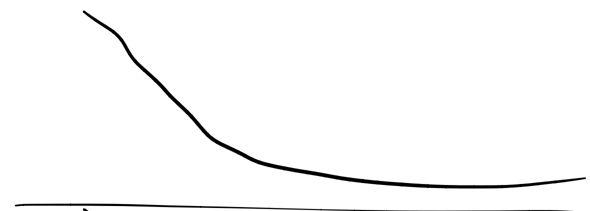
__Newton__ : $\theta_t = \theta_{t-1} - F''(\theta_{t-1})^{-1} F'(\theta_{t-1})$

NONSMOOTH OPTIMIZATION

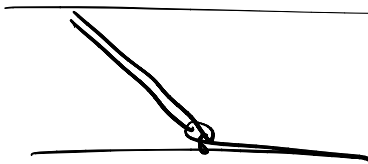
F convex and differentiable with gradients bounded by B.

$$\|F'(\theta)\|_2 \leq B$$

which is true for logistic regression



logistic. $\varphi(u) = \log(1+e^{-u})$



hinge. $\varphi(u) = (1-u)_+$

GD: $\theta_t = \theta_{t-1} - \gamma F'(\theta_{t-1})$

Convergence proof

Need a Lyapunov $V(\theta_t) \downarrow$

$$\|\theta_t - \eta_*\|_2^2 = \|\theta_{t-1} - \eta_* - \gamma F'(\theta_{t-1})\|_2^2$$

$$= \|\theta_{t-1} - \eta_*\|_2^2 + \gamma^2 \underbrace{\|F'(\theta_{t-1})\|_2^2}_{\leq B^2 \text{ Bounded gradients}} - 2\gamma \underbrace{(\theta_{t-1} - \eta_*)^\top F'(\theta_{t-1})}_{\geq F(\theta_{t-1}) - F(\eta_*) \text{ Convexity}}$$

$$2\gamma\left[F(\theta_{t-1}) - F(\eta_*)\right] \leq \|\theta_{t-1} - \eta_*\|_2^2 - \|\theta_t - \eta_*\|_2^2 + \gamma^2 B^2$$

$$\sum_{t=1}^{T} 2\gamma\left(F(\theta_{t-1}) - F(\eta_*)\right) \leq \|\theta_0 - \eta_*\|^2 - \|\theta_T - \eta_*\|_2^2 + T\gamma^2 B^2$$

$$\frac{1}{T}\sum_{t=1}^{T} F(\theta_{t-1}) - F(\eta_*) \leq \frac{\|\theta_0 - \eta_*\|^2}{2\gamma T} + \frac{\gamma B^2}{2}$$

$$F\left(\frac{1}{T}\sum_{t=1}^{T}\theta_{t-1}\right) \leq$$ ↳ Jensen (convexity)

$$\frac{1}{T}\sum_{t=1}^{T} F(\theta_{t-1}) - F(\eta_*) \leq \frac{\|\theta_0-\eta_*\|^2}{2\gamma T} + \frac{\gamma B^2}{2}$$

$$F\left(\frac{1}{T}\sum_{t=1}^{T}\theta_{t-1}\right) \leq$$ ↳ Jensen (convexity)

$$\min_{t\in\{1,T\}} F(\theta_{t-1}) \leq$$

$$\gamma = \frac{C}{\sqrt{T}}$$

$$\leq \left[\frac{\|\theta_0-\eta_*\|^2}{2\gamma C} + \frac{CB^2}{2}\right]\frac{1}{\sqrt{T}}$$

γ depends on the "horizon" T

$\nexists$ "anytime" algorithm $\theta_t = \theta_{t-1} - \gamma_t F'(\theta_{t-1})$

$\gamma_t = \frac{C}{\sqrt{t}}$

Application to ML: G-Lipschitz continuous loss

linear model $\|\theta_*\| \leq D$, $\|\varphi(x)\| \leq R$

estimation error $\leq \frac{GDR}{\sqrt{n}}$ optimization error $\frac{GDR}{\sqrt{t}}$

$$F(\theta) = \frac{1}{n} \sum_{i=1}^{n} \ell(y_i, \varphi(x_i)^T \theta)$$

$$F'(\theta) = \frac{1}{n} \sum_{i=1}^{n} \ell'(y_i, \varphi(x_i)^T \theta)\, \varphi(x_i)$$

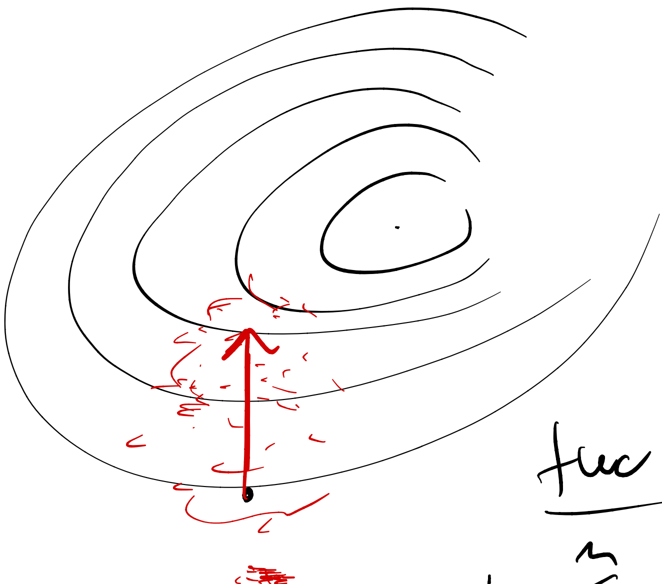
SGD: $F(\theta)$ objective function

$$\theta_t = \theta_{t-1} - \gamma g_t$$

with $\mathbb{E}[g_t \mid \theta_{t-1}] = F'(\theta_{t-1})$

unbiased gradient

$\|g_t\| \leq B$



two set-ups:

(1) $F(\theta) = \frac{1}{n} \sum_{i=1}^{n} \ell(y_i, f_\theta(x_i))$

empirical risk

g_t = gradient of $\ell(y_i, f_\theta(x_i))$

for $i = i(t)$ selected at random

multiple pass SGD

(2) $F(\theta) = \mathbb{E}[\ell(y, f_\theta(x))]$

Expected risk

g_t = gradient of $\ell(y_t, f_\theta(x_t))$ at θ_{t-1}

$t \leq n$

single pass

Proof of SGD : ingredients : $\theta_t = \theta_{t-1} - \gamma g_t$ $\quad \|g_t\|^2 \le B^2$

$\mathbb{E}(g_t \mid \theta_{t-1})$

$= F'(\theta_{t-1})$

$$\mathbb{E}\left[\|\theta_t - \theta_*\|_2^2 \mid \theta_{t-1}\right] = \|\theta_{t-1} - \theta_* - \gamma g_t\|_2^2$$

$$= \|\theta_{t-1} - \theta_*\|^2 + \gamma^2 \underbrace{\|g_t\|_2^2}_{\le B^2} - 2\gamma \underbrace{g_t}^{T} (\theta_{t-1} - \theta_*)$$

$$-2\gamma \underbrace{F'(\theta_{t-1})^T (\theta_{t-1} - \theta_*)}_{\ge F(\theta_{t-1}) - F(\theta_*)}$$

$$F(\theta_{t-1}) - F(\theta_*) \le \frac{\|\theta_{t-1} - \theta_*\|^2}{2\gamma} - \mathbb{E}\left(\frac{\|\theta_t - \theta_*\|^2}{2\gamma} \Big| \theta_{t-1}\right) + \frac{\gamma B^2}{2}$$

$$\mathbb{E}\left(F(\theta_{t-1})\right) - F(\theta_*) \le \mathbb{E}(\quad) - \mathbb{E}(\quad) + \frac{\gamma B^2}{2}$$

...

$$\mathbb{E}\left[F\left(\frac{1}{T}\sum_{t=1}^{T} \theta_{t-1}\right)\right] - F(\theta_*) = \frac{C}{\sqrt{T}}$$