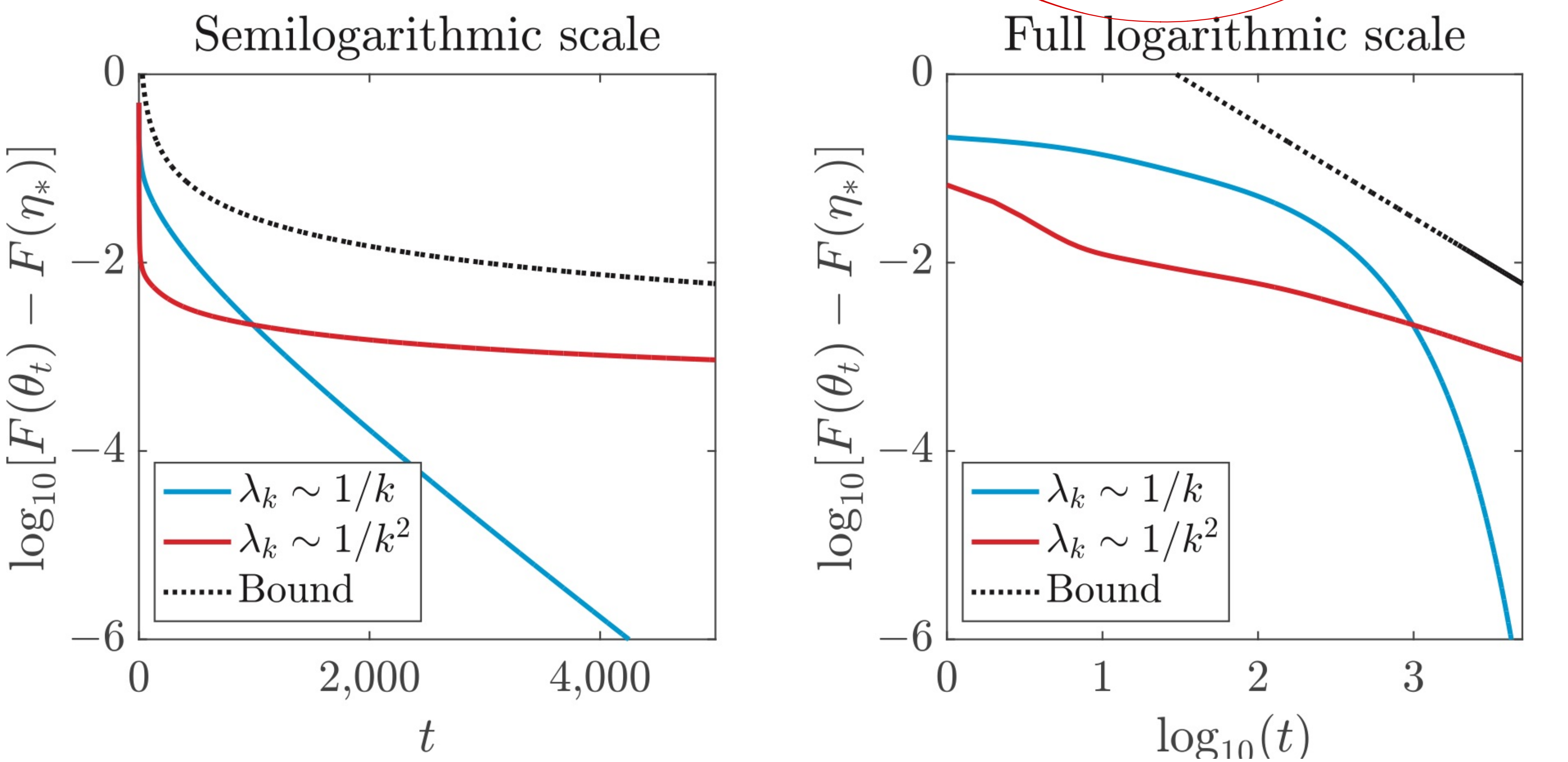
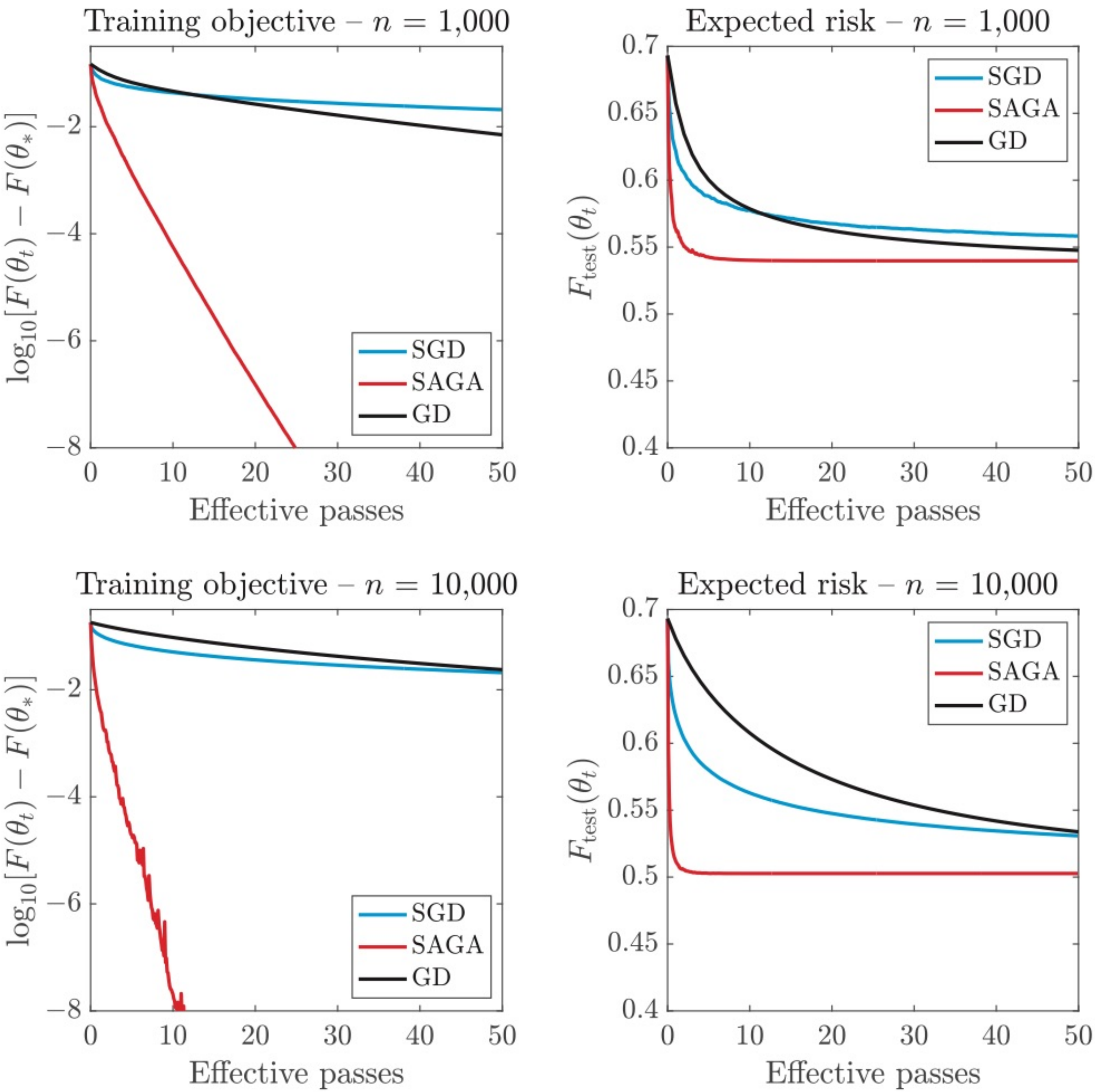
Blog post



eigenvalues λ_k, $k \in \{1, \dots, d\}$ $\hookrightarrow 1000$ $r = \frac{1}{1000}$ & $\lambda_k = \frac{1}{k}$


Training objective – n = 1,000
$\log_{10}[F(\theta_t) - F(\theta_*)]$
SGD
SAGA
GD
Effective passes
Expected risk – n = 1,000
$F_{\text{test}}(\theta_t)$
Training objective – n = 10,000
Expected risk – n = 10,000

Optimization

data $(x_i, y_i)_{i=1,\dots,n} \in \mathcal{X} \times \mathbb{R}$

Expected risk $\mathcal{R}(f) = \mathbb{E}[\ell(y, f(x))]$

$\hookrightarrow$ convex wrt $f(x)$

"weak"

$$\hat{\mathcal{R}}(f) = \frac{1}{n} \sum_{i=1}^{n} \ell(y_i, f(x_i))$$ Empirical Risk min.

Model $f_\theta : \mathcal{X} \to \mathbb{R}$ $\quad \theta \in \Theta \subset \mathbb{R}^d$

Algorithm: minimize $F(\theta) = \hat{\mathcal{R}}(f_\theta) + \underbrace{\text{regularizer}}_{\Omega(\theta)}$

"objective function"

$\hat{\theta}$ approximate minimizer.

Estimation error

$$R(f_{\hat\theta}) - \inf_{\theta\in\mathbb{R}^d} R(f_\theta) = R(f_{\hat\theta}) - \hat R(f_{\hat\theta})$$

$$\boxed{+\hat R(f_{\hat\theta}) - \hat R(f_{\theta_*})}$$

$$+\hat R(f_{\theta_*}) - R(f_{\theta_*})$$

uniform deviations $O\left(\frac{1}{\sqrt{n}}\right)$

θ_* minimizes

$$\leq \hat R(f_{\hat\theta}) - \inf_{\theta\in\mathbb{R}^d} \hat R(f_\theta)$$

optimization error

η_* minimizes

⚠ $\eta_* \neq \theta_*$

⚠ precision: high precision $\Leftrightarrow$ low error

Goal: bounds on optimization error: dimension-indep

Goal: Objective function: $F: \mathbb{R}^d \to \mathbb{R}$

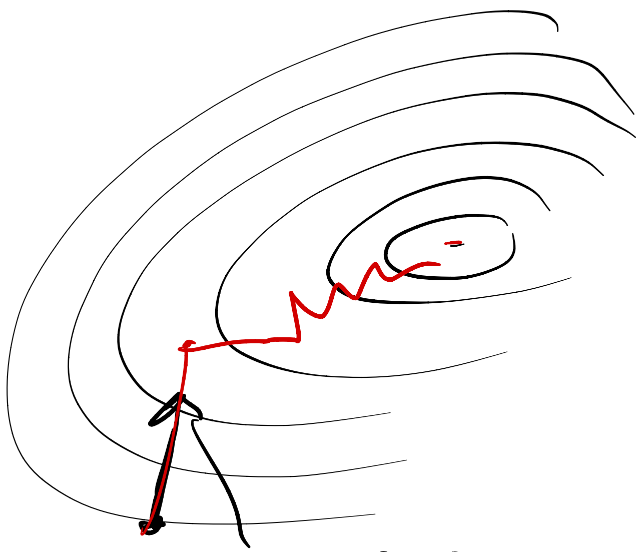
Find an approximate minimizer: $\min_{\theta \in \mathbb{R}^d} F(\theta)$

Algo: gradient descent

θ_0 initial point.

$$\theta_t = \theta_{t-1} - \gamma_t F'(\theta_{t-1})$$ → gradient

↳ step-size



two strategies

① constant or time dependent

② line search

← this class

least-squares: $F(\theta) = \frac{1}{2n} \|\Phi\theta - y\|_2^2$ response vector $y \in \mathbb{R}^n$

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

to avoid large values

gradient: $F'(\theta) = \frac{1}{n} \Phi^T (\Phi\theta - y) \in \mathbb{R}^d$

minimizers: $F'(\eta_*) = 0 \iff \frac{1}{n}\Phi^T\Phi \, \eta_* = \frac{1}{n}\Phi^T y$ normal equations

$$\boxed{H\eta_* = \frac{1}{n}\Phi^T y} \iff \hat{\Sigma} = H$$ Hessian matrix

symmetric positive semi-definite

Characterizing convergence: ① "iterates" $\|\theta_t - \eta_*\|_2^2$

t-th iterate — minimizer.

② "Function values"

$$F(\theta_t) - F(\eta_*) = \cancel{F'(\eta_*)^T(\theta_t - \eta_*)} + \frac{1}{2}(\theta_t - \eta_*)^T H (\theta_t - \eta_*)$$

$= 0$

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

$$\theta_t = \theta_{t-1} - \gamma F'(\theta_{t-1}) = \theta_{t-1} - \gamma H\theta_{t-1} + \gamma \underbrace{\frac{1}{n}\Phi^T y}_{H\eta_*}$$

$$= \theta_{t-1} - \gamma H(\theta_{t-1} - \eta_*)$$

$$\theta_t - \eta_* = \theta_{t-1} - \eta_* - \gamma H(\theta_{t-1} - \eta_*) = (I - \gamma H)(\theta_{t-1} - \eta_*)$$

$$\theta_t - \eta_* = (I - \gamma H)^t(\theta_0 - \eta_*)$$

Function value: $\frac{1}{2}(\theta_t - \eta_*)^T H (\theta_t - \eta_*)$

$$H = \sum_{i=1}^{d} \lambda_i u_i u_i^T$$ eigenvalue decomposition

λ_i: eigenvalues

u_i: eigenvectors (orthogonal)

$$(I - \gamma H)^t = \sum_{i=1}^{d} (1 - \gamma\lambda_i)^t u_i u_i^T$$

converging to 0 $\iff |1 - \gamma\lambda_i| < 1 \quad \forall i \Rightarrow$ depends on smallest μ and largest L eigenvalues > 0

sufficient: $1 - \gamma\lambda_i \in [0, 1)$

$$\iff \begin{cases} 1 - \gamma L \in (0,1) \\ 1 - \gamma \mu \in (0,1) \end{cases} \iff \boxed{\begin{matrix} \gamma L \leq 1 \\ \gamma > 0 \end{matrix}}$$

assumptions

$\boxed{\gamma \leq 1/L} \Rightarrow$ eigenvalues of $(I - \gamma H)^t \in [0, (1 - \gamma\mu)^t]$

Summary: $\theta_t - \eta_* = (1-\gamma H)^t(\theta_0 - \eta_*)$.

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

eigenvalues of $(1-\gamma H)^t \le (1-\gamma \mu)^t$

$$F(\theta_t) - F(\eta_*) = \frac{1}{2}(\theta_0 - \eta_*)^\top H (1-\gamma H)^{2t}(\theta_0 - \eta_*)$$

$$\le \frac{1}{2}(\theta_0 - \eta_*)^\top H (\theta_0 - \eta_*) \cdot (1-\gamma\mu)^{2t}$$

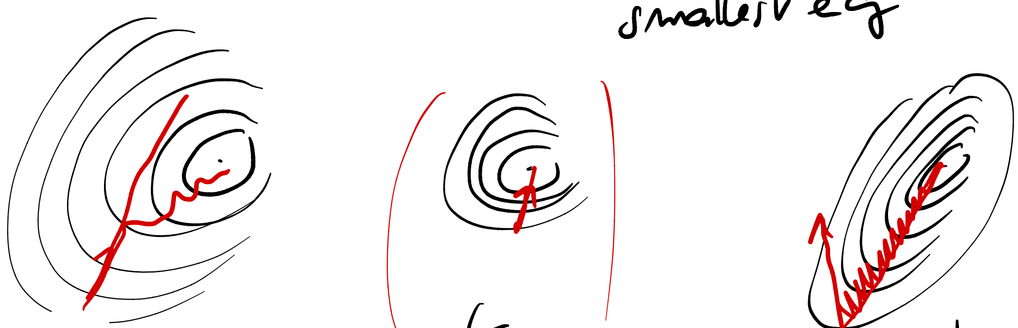
↳ eigenvalues of $H(1-\gamma H)^{2t} \le \lambda_i (1-\gamma\mu)^{2t}$

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

↳ if $\gamma = \frac{1}{L} \Rightarrow \left(1-\frac{\mu}{L}\right)^{2t} \le \exp\left(-\frac{\mu}{L}2t\right) = \varepsilon$ because $1-\alpha \le e^{-\alpha}$

$t = \frac{L}{\mu}\frac{1}{2}\log 1/\varepsilon$

⚠ $\frac{\mu}{L}$ can be tiny. $\frac{L}{\mu}$ = condition number $= \frac{\text{largest eig}}{\text{smallest eig}}$



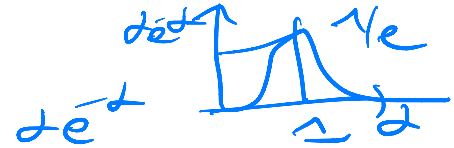
$\frac{L}{\mu}$ small

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

$\frac{L}{\mu}$ large

$\frac{L}{\mu} = +\infty$ $(\mu = 0)$

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



$$\le \lambda_i \exp(-2t\lambda_i\gamma) = \frac{2t\lambda_i\gamma e^{-2t\lambda_i\gamma}}{2t\gamma}$$

$$\le \frac{1}{2t\gamma}\cdot\frac{1}{e}$$

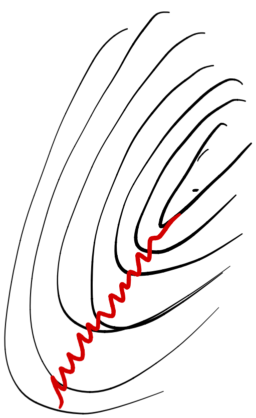
$$\le \frac{1}{4t\gamma e}\|\theta_0-\eta_*\|_2^2 \le \frac{1}{8t\gamma}\|\theta_0-\eta_*\|^2$$

$e \ge 2$

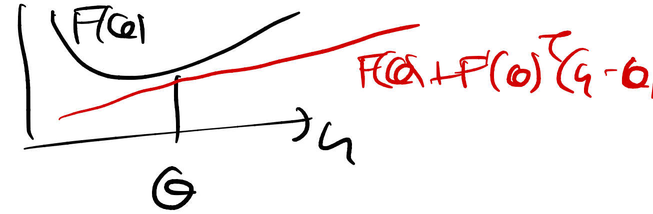
$\to 0$ when $t \to +\infty$

slow but almost unavoidable

H has disappeared



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

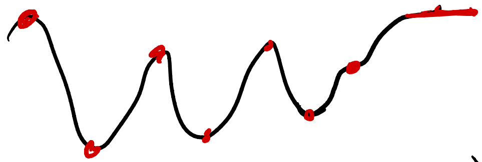


F cvx $\Leftrightarrow \forall \theta, \eta.\quad F(\eta) \geq F(\theta) + F'(\theta)^T(\eta-\theta)$

if F twice differentiable $\Leftrightarrow \forall\theta,\ F''(\theta)$ positive semi-definite

$$\forall \theta,\ F''(\theta) \succcurlyeq 0$$

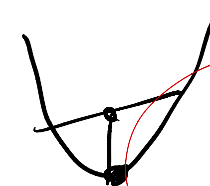
1st consequence: if $F'(\theta)=0$ (stationary point) then $F(\eta) \geq F(\theta)\ \forall \eta$.
$\Leftrightarrow \theta$ is a global minimizer.



2nd consequence: $F(\eta_*) \geq F(\theta) + F'(\theta)^T(\eta_* - \theta) \Rightarrow \boxed{F(\theta) - F(\eta_*) \leq F'(\theta)^T(\theta - \eta_*)}$

Jensen's inequality: F cvx, μ proba. measure on $\mathbb{R}^d$

$$F(\mathbb{E}_\mu \theta) \leq \mathbb{E}_\mu[F(\theta)]$$



Blog post

Properties of cvx fct: Boyd & Van den Berghe

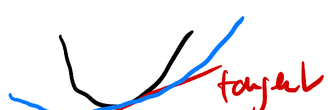
When are ML objective functions cvx? $F(\theta) = \frac{1}{n}\sum_{i=1}^{n} \ell(y_i, h_\theta(x_i))$

Sufficient condition: ℓ is cvx wrt to 2nd var
$h_\theta(x)$ is linear in θ

Strong convexity: $\exists \mu > 0$

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

$\Leftrightarrow F''(\theta)$ has eigenvalues $\geq \mu,\ \forall\theta$



3 examples: linear fct; quadratic; $F(\theta) + \frac{\mu}{2}\|\theta\|_2^2$

lc 31 → lc 54

.F is strongly cvx $\iff F(\eta) \geq F(\theta) + F'(\theta)^T(\eta-\theta) + \frac{\mu}{2}\|\theta-\eta\|_2^2 \quad \forall \theta,\eta$

$$\inf_\eta \Big(\quad F(\eta_*) \geq F(\theta) - \frac{1}{2\mu}\|F'(\theta)\|_2^2 \qquad \begin{pmatrix} a = F'(\theta) \\ z = \eta - \theta \end{pmatrix}$$

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

proof: gradient $a - \mu z = 0$, $z^* = a/\mu$

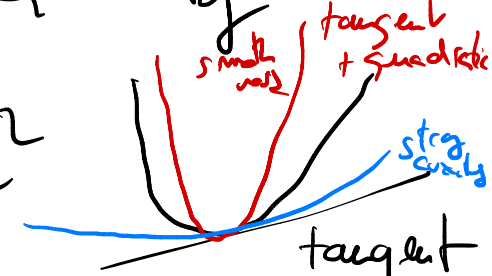
$z^{*T} a - \frac{\mu}{2}\|z^*\|_2^2 = \frac{1}{2\mu}\|a\|_2^2$

$\Rightarrow \forall \theta : \boxed{F(\theta) - F(\eta_*) \leq \frac{1}{2\mu}\|F'(\theta)\|_2^2}$ Łojasiewicz' inequality

smoothness: $F : \mathbb{R}^d \to \mathbb{R}$ is L-smooth iff $\forall \eta, \theta \quad |F(\eta) - F(\theta) - F'(\theta)^T(\eta-\theta)| \leq \frac{L}{2}\|\eta-\theta\|_2^2$

$\iff F''(\theta)$ has eigenvalues in $[-L, L]$

if twice differentiable



Bound for GD (μ-strg cvxity, L-smoothness): $\theta_t = \theta_{t-1} - \frac{1}{L} F'(\theta_{t-1})$ GD

$$F(\theta_t) \leq F(\theta_{t-1}) + F'(\theta_{t-1})^T(\theta_t - \theta_{t-1}) + \frac{L}{2}\|\theta_t - \theta_{t-1}\|^2 \quad \text{(smoothness)}$$

$$= F(\theta_{t-1}) - \frac{1}{L}\|F'(\theta_{t-1})\|_2^2 + \frac{L}{2}\Big\|\frac{1}{L}F'(\theta_{t-1})\Big\|_2^2 = F(\theta_{t-1}) - \frac{1}{2L}\|F'(\theta_{t-1})\|_2^2$$

$$\leq F(\theta_{t-1}) - \frac{\mu}{L}\big(F(\theta_{t-1}) - F(\eta_*)\big) \quad \text{(Łojasiewicz)}$$

$$\Rightarrow F(\theta_t) - F(\eta_*) \leq \Big(1 - \frac{\mu}{L}\Big)\big(F(\theta_{t-1}) - F(\eta_*)\big) \leq \boxed{\Big(1-\frac{\mu}{L}\Big)^t \big(F(\theta_0) - F(\eta_*)\big)}$$

↳ exponential convergence | linear

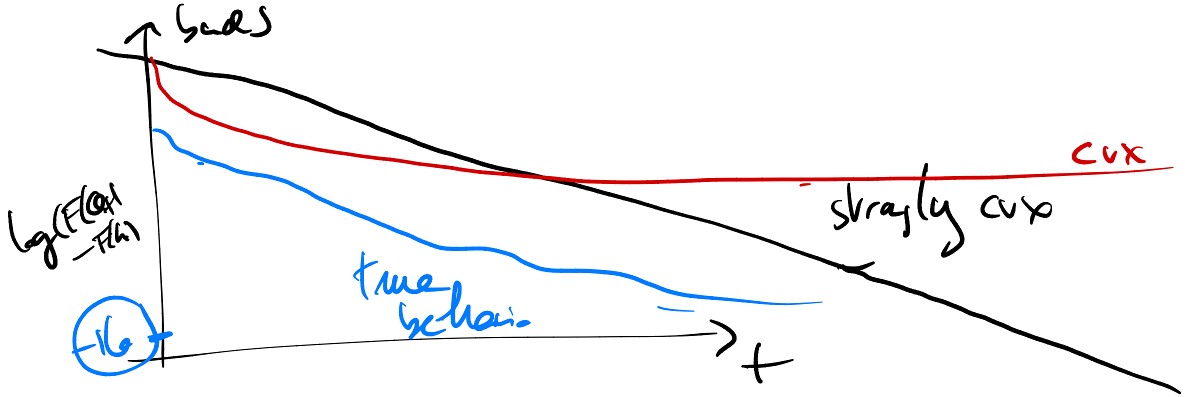
L/μ = condition number

Bound for GD: $F(\theta_t) - F(\eta_*) \leq \boxed{\frac{L}{2t}\|\theta_0 - \eta_*\|_2^2}$ (see back)

$\left(1 - \sqrt{\frac{\mu}{L}}\right)^t$ accelerated

$\frac{1}{t^2}$ Nesterov

① Adaptivity ② "optimal" ③ Newton ④ proximal


bounds
cvx
strongly cvx
true behavior
$\log(F(x_t) - F(x_*))$
t

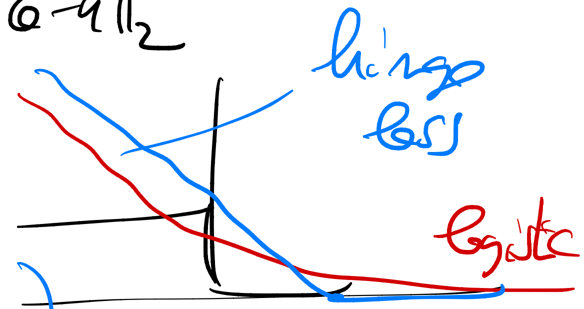
Non-smooth problems : $F:\mathbb{R}^d\to\mathbb{R}$ B-Lipschitz continuous

iff $\forall\theta,\eta,\ |F(\theta)-F(\eta)|\le B\|\theta-\eta\|_2$

$\iff \forall\theta,\ \|F'(\theta)\|_2\le B$

if differentiable

(⚠ excluding quadratic fct)



See notion of subgradient in book

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

assuming F has a minimizer η_*

Looking for a "Lyapunov fct"

$$\|\theta_t-\eta_*\|_2^2=\|\theta_{t-1}-\eta_*-\gamma_t F'(\theta_{t-1})\|_2^2 \quad \text{(definition of GD)}$$

$$=\|\theta_{t-1}-\eta_*\|_2^2-2\gamma_t\underbrace{F'(\theta_{t-1})^T(\theta_{t-1}-\eta_*)}_{\ge F(\theta_{t-1})-F(\eta_*)\ge 0 \text{ (by convexity)}}+\gamma_t^2\underbrace{\|F'(\theta_{t-1})\|_2^2}_{\le B^2 \text{ (Lipschitz continuity)}}$$

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

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

if $\gamma_t=\gamma$

(telescoping sum)

(Jensen)

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

$$F\Big(\frac{1}{T}\sum_{t}^{T}\theta_{t-1}\Big) - F(\eta_*) \le \frac{1}{2\gamma T}\|\theta_0-\eta_*\|_2^2 + \frac{B^2}{2}\gamma$$

Does not go to zero if $T \to +\infty$.

if $\gamma = \frac{\delta}{\sqrt{T}}$

$$\le \frac{1}{\sqrt{T}}\left[\frac{1}{2\delta}\|\theta_0-\eta_*\|^2 + \frac{B^2\delta}{2}\right]$$

slow but "optimal"

→ not an anytime algorithm.

- step-size depends on horizon T
- see back for $\gamma_t = \frac{\delta}{\sqrt{t}}$ $\Rightarrow$ bound $\propto \frac{\log t}{\sqrt{t}}$
- Bound on $F(\theta_t) - F(\eta_*)$? Blog post by F. Orabona.
- Application to ML: $F(\theta) = \frac{1}{n}\sum_{i=1}^{n} \ell(y_i, \theta^\top \varphi(x_i))$

G-Lipschitz

$\|\varphi(x)\|_2$ bounded by R

$$\|F'(\theta)\|_2 \le GR$$

optim. error $\le \frac{1}{2\sqrt{T}}\left[\frac{1}{\delta}\|\theta_0-\eta_*\|^2 + \delta G^2R^2\right] = \frac{GR}{\sqrt{T}}\|\underline{\theta_0}-\eta_*\|_2$ with proper choice of δ

$\theta_0 = 0$

$= \frac{GR\|\eta_*\|_2}{\sqrt{T}}$

Last week: uniform deviation $\simeq \frac{GR\|\eta_*\|_2}{\sqrt{n}}$

$\boxed{T \simeq n}$

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

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

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

- SGD: $\theta_t = \theta_{t-1} - \gamma \overbrace{\frac{\partial}{\partial\theta}\ell(y_{i(t)}, f_{\theta_{t-1}}(x_{i(t)}))}^{g_t}$

multiple pass

$i(t)$ uniform in $\{1,\dots,n\}$.

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

(wrt $i(t)$)

$$\mathbb{E}\big(F(\bar{\theta}_T) - F(\theta_*)\big) \le \frac{D}{\sqrt{T}}$$

wrt $i(1),\dots,i(t)$

as follows

- Single pass SGD: $F(\theta) = \mathbb{E}\,\ell(y, f_\theta(x))$

$$\theta_t = \theta_{t-1} - \gamma \overbrace{\frac{\partial}{\partial\theta}\ell(y_t, f_{\theta_{t-1}}(x_t))}^{g_t}$$

$\boxed{\text{unbiased}}$ $\mathbb{E}[g_t \mid \theta_{t-1}] = F'(\theta_{t-1})$ $\qquad t \in \{1,\dots,n\}$