Optimisation Stochastique

Charles Vin

S1-2023

RATRAPER COURS 1

Nouveau cours du 16/11

CCL du cours de la dernière fois

$$R^{\phi}(\hat{h}^{\phi-\mathbb{E}R?}) - R^{\phi}(h^{\star}, \phi).$$

0.1 Relation between R^{ϕ} and $R^{0/1}$

In this section, no empirical proof, no n

- $R^{\phi}(h) = \mathbb{E}[\phi(-Yh(X))]$
- $R^{0/1}(h) = \mathbb{E}[\mathbb{1}_{Y \neq sign(h(X))}]$
- ϕ = hinge / logistic / least square

Lemme 1

If ϕ is diff, convex, increasing, then $sign(h^{\star,\phi}) = f^{\star,Bayes}$ with $h^{\star,\phi} \in \arg\min_h R^{\phi}(h)$

Preuve : 1.

$$\begin{aligned} R^{\phi}(h) &= \mathbb{E}[\phi(-Yh(X))(\mathbb{1}_{Y=1} + \mathbb{1}_{Y=-1})|X] \\ &= \mathbb{E}[\phi(-h(X))\eta(X) + \phi(h(X))(1-\eta(X))] \end{aligned}$$

with $\eta(X) = P(Y = 1|X)$

2. Define $H_{\phi}(p, \eta) := \eta\phi(-p) + (1-\eta)\phi(p)$ and $p^{\star,\phi}(\eta) = \arg\min H_{\phi}(p, \eta)$ (assuming existence for now)
   $h^{\star,\phi}$ minimizes R^{ϕ} and is such that for any fixed x

$$h^{\star,\phi}(x) = p^{\star,\phi}(\eta(x)).$$

   $\forall h, R^{\phi}(h) - R^{\phi}(h^{\star,\phi}) = \mathbb{E}[H_{\phi}(h(X), \eta(X)) - H_{\phi}(h^{\star,\phi}(X), \eta(X))]$

3. Example for Least Square :

$$\begin{aligned} H_{\phi}(p, \eta) &= \eta(1-p)^2 + (1-p)(1+p)^2 \\ \frac{\partial H_{\phi}}{\partial p}(p, \eta) &= 2(p-1)\eta + 2(1-\eta)(1+p) \\ &= 0 \Leftrightarrow p = 2\eta - 1 \end{aligned}$$

   See Table 0.1

   In all cases, $sign(p^{\star,\phi}(\eta(X)) = sign(\eta(X) - 1/2)) = sign(h^{\star,\phi}(X)) = f^{\star,Bayes}$

4. In general with ϕ strictly increasing, diff, convex, when $\phi(t) \to_{t\to+\infty} +\infty \; \forall \eta \in]0,1[, H_{\phi}(\eta, p) \to_{p\to\pm\infty} +\infty$. Thus $p^{\star,\phi}(\eta)$ exists. And $p \mapsto H_{\phi}(p, \eta)$ is diff

$$\frac{\partial H_{\phi}}{\partial p}(p, \eta) = 0 \Leftrightarrow \eta\phi'(-p^{\star,\phi}(\eta)) = (1-\eta)\phi(p^{\star,\phi}(\eta)).$$

   (a) If $\eta < 1/2$, then $\eta < 1 - \eta \Rightarrow \phi'(p^{\star,\phi}(\eta)) > \phi'(p^{\star,\phi}(\eta)) \Rightarrow p^{\star,\phi}(\eta) \leq 0$

   (b) If $\eta > 1/2$ ... $\Rightarrow p^{\star,\phi} \geq 0$

Finally, $sign(p^{\star,\phi}(\eta) = sign(\eta - 1/2))$ and thus $sign(h^{\star,\phi}(X)) = f^{\star,Bayes}(X)$

□

Loss	$p^{\star,\phi}(\eta)$	$\min H_\phi(p,\eta)$
LS : $(1+v)^2$	$2\eta-1$	$4\eta(1-\eta)$
Hinge	sign	a
Logistic	a	a

Lemme 2 (Zhang)

Assume *phi* increasing, convex such that $\phi(0) = 1$. For $\gamma \geq 1$ we have $|\eta - 1/2|^\gamma \geq c\left|1 - H_\phi(p^{\star,\phi}(\eta),\eta)\right|$.
$\forall h$ classifier $h : \mathcal{X} \to \mathbb{R}$

$$R^{0/1}(sign(h)) - R^{0/1}(f^{\star,Bayes}) \leq 2c^{1/\gamma}(R^\phi(h) - R^\phi(h^{\star,\phi})).$$

When h approximately minimizes the relaxed excess risk its $sign(h)$ behaves well in terms of the initial excess risk !!.

Note. Note that $\gamma = 2$ for the square loss and the logistic loss. And that $\gamma = 1$ for the hinge loss.
(we do not care about c)

Preuve :

$$\begin{aligned} R^{0/1}(sign(h)) - R^{0/1}(f^{\star,Bayes}) &= \mathbb{E}[\mathbb{1}_{sign(h(X))\neq f^{\star,Bayes}(X)2|\eta(X)-1/2|}] \\ \text{(jensen, (1))} &\leq \mathbb{E}[\mathbb{1}_{sign(h(X))\neq f^{\star,Bayes}(X)2^\gamma|\eta(X)-1/2|^\gamma}]^{1/\gamma} \\ &\leq 2c^{1/\gamma}\mathbb{E}[\mathbb{1}_{sign(h(X))\neq f^{\star,Bayes}(X)}(1 - H_\phi(p^\star_\phi(\eta(X)),\eta(X))]^{1/\gamma} (\eta(X) = P(Y=1|X)) \end{aligned}$$

Note. Note that when $sign(h(X)) \neq sign(\eta(X) - 1/2)$, then $H'_\phi(h(X),\eta(X)) > 1$. Indeed, $\eta\phi(-p) + (1-\eta)\phi(p) \geq \phi(-\eta p + (1-\eta)p) = \phi((1-2\eta)p)$ because ϕ convex. And now $\phi((1-2\eta)p) \geq \phi(0) = 1$ because ϕ increasing ≥ 0 when $sign(p) \neq sign(\eta - 1/2)$

$$\begin{aligned} (1) &\leq 2c^{1/\gamma}(\mathbb{E}[H(h(X),\eta(X)) - H(p^{\star,\phi}(\eta(X)),\eta(X))])^{1/\gamma} \\ &= 2c^{1/\gamma}(R^\phi(h) - R^\phi(h^{\star,\phi}))^{1/\gamma} \end{aligned}$$

□

CCL : $\forall \hat{h}$

$$R^{0/1}(sign(\hat{h})) - R^{0/1}(f^{\star,Bayes}) \leq c^{1/\gamma}(R^\phi(\hat{h}) - R^\phi(h^{\star,\phi}))^{1/\gamma}$$
$$R^\phi(\hat{h}) - R^\phi(h^{\star,\phi}) = R^\phi(\hat{h}) - R^\phi(h^{\star,\phi}_{\mathcal{F}}) + R^\phi(h^{\star,\phi}_{\mathcal{F}}) - R^\phi(h^{\star,\phi})$$

where

- $h^{\star,\phi}_{\mathcal{F}} \in \arg\min R^\phi(h)$
- $R^\phi(h^{\star,\phi}_{\mathcal{F}}) - R^\phi(h^{\star,\phi})$ approx error

$$\begin{aligned} R^phi(\hat{h}) - R^\phi(h^{\star,\phi}_{\mathcal{F}}) = R^\phi(\hat{h}) - \hat{R}^\phi_n(\hat{h}) &(\leq \sup_{\mathcal{F}} \hat{R}_n - R^\phi) \\ &+ \hat{R}^\phi_n(\hat{h}) - \hat{R}^\phi_n(\hat{h}^{\phi ERM}) (\text{"optim error"}) \\ &+ \hat{R}^\phi_n(\hat{h}^{\phi-ERM}) - \hat{R}^\phi_n(\hat{h}^{\star,\phi}_{\mathcal{F}}) (\leq 0) \\ &+ \hat{R}^\phi_n(h^{\star,\phi}_{\mathcal{F}}) - R^\phi(h^{\star,\phi}_{\mathcal{F}}) (\leq \sup_{\mathcal{F}} \hat{R}^\phi_n - R^\phi) \end{aligned}$$

Since the estimation error typically scales in $O(\frac{1}{\sqrt{n}})$, no need to reach the ERM using our optimization algo !!.

Note. When using Lipschitz fonctions, we obtain slow rates $O(\frac{1}{\sqrt{n}})$. Is there a path towards fast rates ? Let's take the example of the mean estimation.

1. Method 1 :

$$\hat{\theta} = \frac{1}{n}\sum_{i=1}^{n} Z_i = \arg\min_{\theta} \frac{1}{2n}\sum_{i=1}^{n}(Z_i - \theta)^2$$
$$\theta^\star = \arg\min \frac{1}{2}\mathbb{E}[(\theta - Z)^2] = \mathbb{E}[Z]$$

   From the developpement before on the estimation error

$$R(\hat{\theta}) - R(\theta^\star) = O(\frac{1}{\sqrt{n}}).$$

2. Method 2 : Direct computation

$$R(\theta) = \frac{1}{2}\mathbb{E}[(\theta - Z)^2] = \frac{1}{2}(\theta - \mathbb{E}[Z])^2 + \frac{1}{2}Var(Z)$$
$$\Rightarrow R(\hat{\theta}) - R(\theta^\star) = R(\hat{\theta})(R(\mathbb{E}[Z])) = \frac{1}{2}(\hat{\theta} - \mathbb{E}(Z))^2 (\text{conditionallty to } \mathcal{D}_n)$$
$$\mathbb{E}_{D_n}[] = \frac{1}{2}\mathbb{E}[(\frac{1}{n}\sum Z_i - \mathbb{E}[Z])^2] = \frac{1}{2\mathbf{n}}Var(Z) (\mathbf{n} \text{ is FAST RATE } O(\frac{1}{n}))$$

   Bound only for this specific $\hat{\theta}$ and because I also have stong convexity.

   In supervised learning, fast rates can be established for strongly convex function (in our relaxed risks)

Chapter 1

Basics of deterministic optimisation

In ML, construct a predictor often boils down to minimize an empirical risk using iterative algorithms.

1.1 First order method

1.1.1 Basics of convex analysis

$F : \mathbb{R}^d \to \mathbb{R}$ convex, diff, L-smooth (its gradient is L-Lipschitz).

- convexity (under chords) : $F(\eta\theta + (1-\eta)\theta') \leq \eta F(\theta) + (1-\eta)F(\theta'), \forall \theta, \theta', \forall \eta \in [0,1]$
- If we add diff (tangent lie below) we have $F(\theta') \geq F(\theta) + \langle \nabla F(\theta), \theta' - \theta \rangle, \forall \theta, \theta'$
- (increasing slopes) $\langle \nabla F(\theta) - \nabla F(\theta'), \theta - \theta' \rangle \geq 0$ (∇F is said to be a monotone operator)
- if we add $\mathcal{C}^2$ (curves upwards) $\forall \theta, Hess_F(\theta) \succeq 0$ (SDP)

μ-strongly convex, $\mu > 0$.

- convexity (**"well"** under chords) : $F(\eta\theta+(1-\eta)\theta') \leq \eta F(\theta)+(1-\eta)F(\theta'), \forall \theta, \theta' \frac{\mu(1-\mu)}{2} \|\theta - \theta'\|_2^2, \forall \eta \in [0,1]$
- If we add diff (tangent lie **"well"** below) we have $F(\theta') \geq F(\theta) + < \nabla F(\theta), \theta' - \theta > \forall \theta, \theta' + \frac{\mu}{2}\|\theta - \theta'\|_2^2$
- (**"well"**increasing slopes) $\langle \nabla F(\theta) - \nabla F(\theta'), \theta - \theta' \rangle \geq 0 + \mu \|\theta - \theta'\|$
- if we add $\mathcal{C}^2$ (curves upwards) $\forall \theta, Hess_F(\theta) \succeq \mu Id$ (SDP)

F is μ-strongly convex $\forall \theta_0, \theta \mapsto F(\theta) - \frac{\mu}{2}\|\theta - \theta_0\|_2^2$ is convex.
L-Smooth : $\forall \theta, \theta', \|\nabla F(\theta) - \nabla F(\theta')\| \leq L \|\theta - \theta'\|$

Lemme 3 (Descent lemma)

Assume that F is L-Smooth. Therefore $\forall \theta, \theta' \in$ dommain of f

$$F(\theta') \leq F(\theta) + \langle \nabla F(\theta), \theta' - \theta \rangle + \frac{L}{2} \|\theta' - \theta\| .$$

Preuve :

$$\begin{aligned}
F(\theta') &= F(\theta) + \int_0^1 < \nabla F(\theta + t(\theta' - \theta)), \theta' - \theta > dt \\
&= F(\theta) + < \nabla F(\theta), \theta' - \theta > + \int_0^1 < \nabla F(\theta + t(\theta' - \theta)) - \nabla F(\theta), \theta' - \theta > dt \\
&\leq F(\theta) + \langle \nabla F(\theta), \theta' - \theta \rangle + \int_0^1 \|\nabla F(\theta + t(\theta' - \theta)) - \nabla F(\theta)\| \, \|\theta' - \theta\| \, dt \\
&\leq F(\theta) + \langle \nabla F(\theta), \theta' - \theta \rangle + \int_0^1 tL \|\theta' - \theta\|^2 \, dt \\
&\leq F(\theta) + \langle \nabla F(\theta), \theta' - \theta \rangle + \frac{1}{2} L \|\theta' - \theta\|_2^2
\end{aligned}$$

□

Consequence of this quadratics upper bound

1.

$$F(\theta) \leq F(\theta^\star) + \langle \nabla F(\theta^\star), \theta - \theta^\star \rangle + \frac{L}{2} \|\theta - \theta^\star\|^2$$

$$F(\theta) - F(\theta^\star) \leq \frac{L}{2} \|\theta - \theta^\star\|^2$$

2.

$$\min_\theta F(\theta) \leq \min_\theta F(\theta) + \langle \nabla F(\theta), \theta' - \theta \rangle + \frac{L}{2} \|\theta' - \theta\|^2 .$$

$\min_\theta F(\theta) + \langle \nabla F(\theta), \theta' - \theta \rangle + \frac{L}{2} \|\theta' - \theta\|^2$ is reach for $\theta' = \theta - \frac{1}{L} \nabla F(\theta)$

$$\begin{aligned}
&\leq F(\theta) + \langle \nabla F(\theta), \theta - \frac{1}{L} \nabla F(\theta) - \theta \rangle + \frac{L}{2} \left\| \theta - \frac{1}{L} \nabla F(\theta) - \theta \right\|^2 \\
&= F(\theta) - \frac{1}{2L} \|\nabla F(\theta)\|^2
\end{aligned}$$

All in all, $\forall \theta$

$$\frac{1}{2L} \|\nabla F(\theta)\|^2 \leq F(\theta) - F(\theta^\star) \leq \frac{L}{2} \|\theta - \theta^\star\|^2 .$$

Note. In what follows, we could easily extend the study to non-diff function by involving **subgradients**.
$F : \mathbb{R}^D \mapsto \mathbb{R}$ A vector $\eta \in \mathbb{R}^d$ is a subgradient of F at θ if

$$\forall \theta', F(\theta') \geq F(\theta) + \langle \eta, \theta' - \theta \rangle.$$

$\partial F(\theta)$ is the subdifferential of F at θ a,d gathers all the subgradients of F at 0 i.e. the direction of hyperplanes passing through $(\theta, F(\theta))$ but remaining below the graph of F

1.1.2 Gradient algorithms

$\theta^\star = \arg\min F$ assuming existence and uniqueness.

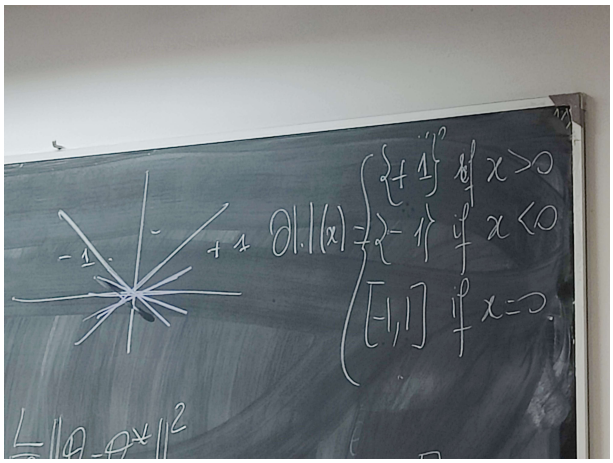


Figure 1.1: subgradients

Gradient algo

1. Init $\theta_0 \in \mathbb{R}^d$
2. $\forall t \geq 0, \theta_{t+1} = \theta_t - \gamma_{t+1} \nabla F(\theta_t)$ with γ_{t+1} gradient steps / learning rates

Choice of steps :

- Constant step sizes $\gamma_t = \gamma, \forall t$ it may depend on the time horizons $T : \forall t \in [0,1], \gamma_t = \gamma(T)$
- Line search : optimal step size at each iteration. $\gamma_t = \arg\min_{\gamma > 0} F(\theta_{t-1} - \gamma \nabla F(\theta_{t-1})))$. You can forget about that case in online algo!

Link with the gradient flow

The iterates of Gradient Descent (GD, Euler, XVIIIe)

$$\theta_{t+1} = \theta_t - \gamma_t \nabla F(\theta_t).$$

can be rewrittent as

$$\frac{\theta_{t+1} - \theta_t}{\gamma_t} = -\nabla F(\theta_t).$$

Make the step size γ_t shrink to 0, we obtain the ODE

$$\frac{\partial \theta}{\partial t}(t) = -\nabla F(\theta(t)).$$

This continuous version is called the Gradient Flow (GF). Thus GD is a discretization of GF (using finite differences).

$\nabla F(\theta)$ is orthogonal to $\{\theta' : F(\theta') = F(\theta)\}$ (level set) so that $\frac{\partial \theta}{\partial t}(t) = \theta(t)$ point inwards $\{\theta' : F(\theta') \leq F(\theta)\}$ which guarantees that $F(\theta(t))$ is decreasing.
Indeed $\frac{\partial (F \circ \theta)}{\partial t}(t) = \langle \nabla F(\theta(t)), \dot{\theta}(t) \rangle = - \left\| \nabla F(\theta(t)) \right\|^2$

Théorème 4

For F an L-Smooth. for $\gamma_t = \gamma, \forall t$ with $\gamma < 2/L$

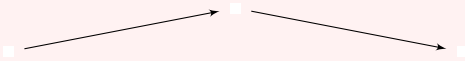
$$F(\theta_t) - F(\theta^\star) \leq \frac{\|\theta_0 \theta^\star\|}{2\gamma(1 - \frac{\gamma L}{2})T}.$$

For $\gamma = \frac{1}{L}$ we have

$$F(\theta_t) - F(\theta^\star) \leq \frac{\|\theta_0 - \theta^\star\|}{2\gamma(1 - \frac{\gamma L}{2})T} = \frac{L\|\theta_0 - \theta^\star\|^2}{T}.$$

Note. 1. This is a sublinear rate $O(1/T)$

2. Using a constant step size.

γ	0	$1/L$	$2/L$
the rate			

3. Optimal "constant" step size $= \frac{1}{L}$

Note (Interpolation of GD with $\gamma = \frac{1}{L}$). Note that

$$\begin{aligned}\tilde{\theta}_t &= \arg\min F(\tilde{\theta}_{t-1}) + \langle \nabla F(\tilde{\theta}_{t-1}), \theta - \tilde{\theta}_{t-1}\rangle + \frac{L}{2}\left\|\theta - \tilde{\theta}_{t-1}\right\|^2 \\ &= \tilde{\theta}_{t-1} - \frac{1}{L}\nabla F(\tilde{\theta}_{t-1})\end{aligned}$$

Using GD with $\gamma = \frac{1}{L}$ amounts to minimizer a quadratic upper bound (provided by smoothness). This idea is a the heart of the Majorize-Minimize algo.

Preuve :

$$\begin{aligned}\|\theta_{t+1} - \theta^\star\|_2^2 &=^{(\mathrm{GD})} \|\theta_t - \gamma\nabla F(\theta_t) - \theta^\star\|_2^2 \\ &= \|\theta_t - \theta^\star\|_2^2 - 2\gamma\langle \nabla F(\theta_t), \theta_t - \theta^\star\rangle + \gamma^2\|\nabla F(\theta_t)\|_2^2\end{aligned}$$

Function convexe + L-Smooth : $\|\nabla F(\theta)\|^2 \leq L\langle \nabla F(\theta), \theta - \theta^\star\rangle$. This is a consequence of the co-coercivity of ∇F (with param $1/L$)

Note (Co-coercivity). F convex, L-Smooth, then θ, θ'

$$\langle \nabla F(\theta) - \nabla F(\theta'), \theta - \theta'\rangle \geq_{\text{co-coercivity}} \frac{1}{L}\|\nabla F(\theta) - \nabla F(\theta')\|_2^2.$$

Proof of the note on co-coercivity: Define two function

$$\begin{aligned}G(\theta') &= F(\theta') - \langle \nabla F(\theta), \theta'\rangle \\ H(\theta') &= F(\theta) - \langle \nabla F(\theta'), \theta\rangle\end{aligned}$$

G and H are smooth. $\theta' = \theta$ minimize $\theta' \mapsto G(\theta')$ and

$$\begin{aligned}F(\theta') - F(\theta) - \langle \nabla F(\theta), \theta' - \theta\rangle &= G(\theta') - G(\theta) \\ &\geq \frac{1}{2L}\|\nabla G(\theta')\|^2 \text{ (by LHS, 1) and where "all in all")} \\ &= \frac{1}{2L}\|\nabla F(\theta') - \nabla F(\theta)\|^2\end{aligned}$$

Idem, $\theta = \theta'$ minimizes $\theta \mapsto H(\theta)$

$$\begin{aligned} F(\theta) - F(\theta') - \langle \nabla F(\theta'), \theta - \theta' \rangle &= H(\theta) - H(\theta') \\ &\geq \frac{1}{2L} \left\| \nabla H(\theta) \right\|^2 \\ &= \frac{1}{2L} \left\| \nabla F(\theta') - \nabla F(\theta) \right\|^2 \end{aligned}$$

Sum the 2 inequalities to conclude □

End of the co-coercivity note

$$\begin{aligned} \left\| \theta_{t+1} - \theta^\star \right\|^2 &= \left\| \theta_t - \theta^\star \right\|^2 - 2\gamma \langle \nabla F(\theta_t), \theta_t - \theta^\star \rangle + \gamma^2 \left\| \nabla F(\theta_t) \right\|^2 \\ &\geq \left\| \theta_t - \theta^\star \right\|^2 - 2\gamma (1 - \frac{\gamma L}{2}) \langle \nabla F(\theta - t), \theta_t - \theta^\star \rangle \\ &\Rightarrow 2\gamma (1 - \frac{\gamma L}{2}) \langle \nabla F(\theta_t), \theta_t - \theta^\star \rangle \leq \left\| \theta_{t+1} - \theta^\star \right\|^2 - \left\| \theta_t - \theta^\star \right\|^2 \\ &\Rightarrow 2\gamma (1 - \frac{\gamma L}{2}) (F(\theta_t) - F(\theta^\star)) \leq \left\| \theta_{t+1} - \theta^\star \right\|^2 - \left\| \theta_t - \theta^\star \right\|^2 \\ F(\theta_t) - F^\star &\leq \frac{1}{T} \sum_{t=1}^{T} F(\theta_t) - F(\theta^\star) \\ &\leq \frac{\left\| \theta_0 - \theta^\star \right\|^2}{2\gamma (1 - \frac{\gamma L}{2}) T} \end{aligned}$$

□

Nouveau cours du 23/11

RAPPEL : On regarde

- $\theta_{t+1} = \theta_t - \gamma_t \nabla F(\theta_t)$
- $\theta_0 \in \mathbb{R}^d$

Théorème 5

F L-smooth, diff
For $\gamma_t = \gamma$ for all $t \leq 0$

$$\begin{aligned} F(\theta_T) - F(\theta^\infty) &\leq \frac{\|\theta_0 - \theta^\infty\|^2}{2\gamma(1 - \frac{\gamma L}{2})T} \\ &= L\frac{\|\theta_0 - \theta^\infty\|^2}{T} (\gamma = 1/L) \end{aligned}$$

- $\gamma = \dfrac{1}{L}$ It is the largest constant step size ensuring the most decrease of the objective fct at each iteration.
- L-smooth, diff $\mathcal{C}^2 \Leftrightarrow \lambda_{MAX}(H_F(\theta)) \leq L \forall \theta$

$$\begin{aligned} \Leftarrow \|\nabla F(\theta) - \nabla F(\theta')\| &= \left\| \int_0^1 H_F(\theta' + t(\theta - \theta'))(\theta - \theta') dt \right\| \\ &\leq \int_0^1 \|H_F(\theta + t(\theta - \theta'))(\theta - \theta')\| \, dt \\ &\leq L \|\theta - \theta'\|_2 \end{aligned}$$

Théorème 6

If F is L-Smooth, diff and μ - strongly convexe, then for all step size $\gamma \leq 1/L$

$$\begin{aligned} \|\theta_T - \theta^\star\|^2 &\leq (1 - \gamma\mu)^T \|\theta_0 - \theta^\star\|^2 \\ (\text{for } \gamma = 1/L) &= (1 - \frac{\mu}{L})^T \|\theta_0 - \theta^\star\|^2 \, (\text{for } \gamma = 1/L) \end{aligned}$$

Note.

1. The algorithm is the same so the CV rate is improved only by properties of F. In such a casen the rate is said to be linear.
2. CV rate on the iterates (!!) and not only on the objective rate

$$\begin{aligned} F(\theta_T) - F(\theta^\star) &\leq \langle \nabla F(\theta^\infty), \theta_T - \theta^\star \rangle + \frac{L}{2} \|\theta_T - \theta^\star\| \\ &= 0 + \frac{L}{2} \|\theta_T - \theta^\star\| \end{aligned}$$

$$\frac{\mu}{2} \|\theta_T - \theta^\star\|^2 \leq_{\text{strong cvxty}} F(\theta_T) - F(\theta^\star) \leq + \frac{L}{2} \|\theta_T - \theta^\star\| .$$

3. Choice of γ : the largest possible.
4. $\mu \leq L$. ($\mu = L$ iff $F(\theta) = \dfrac{L}{2} \|\theta - \theta^\star\|$)

$\kappa = \dfrac{\mu}{L}$ is called the condition number of F.
$\kappa << 1$ "Bad conditioning"
$\kappa \simeq 1$ "good conditioning"

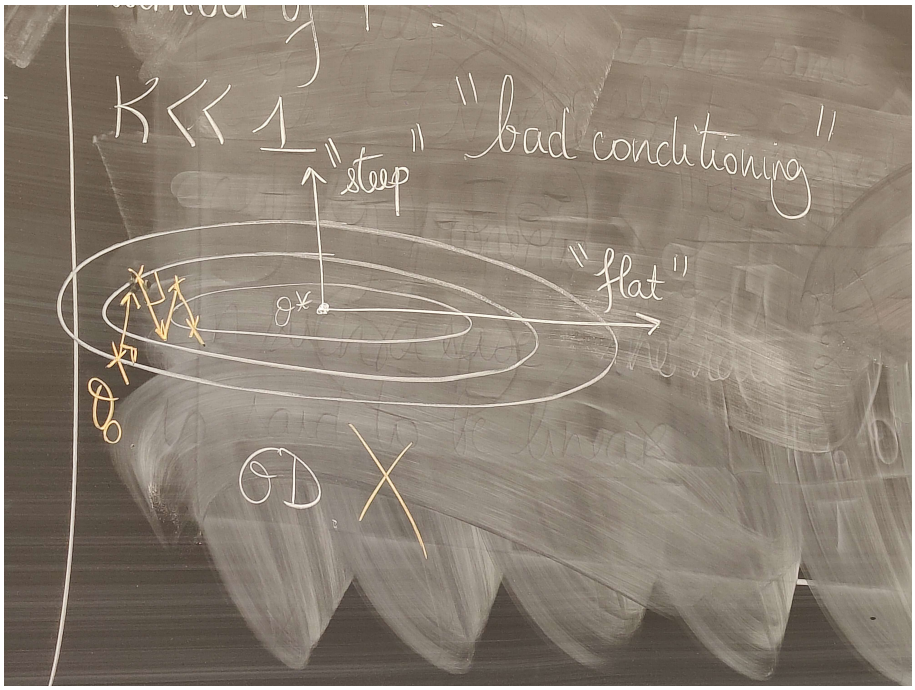


Figure 1.2: With $\kappa \ll 1$ "bad conditioning"

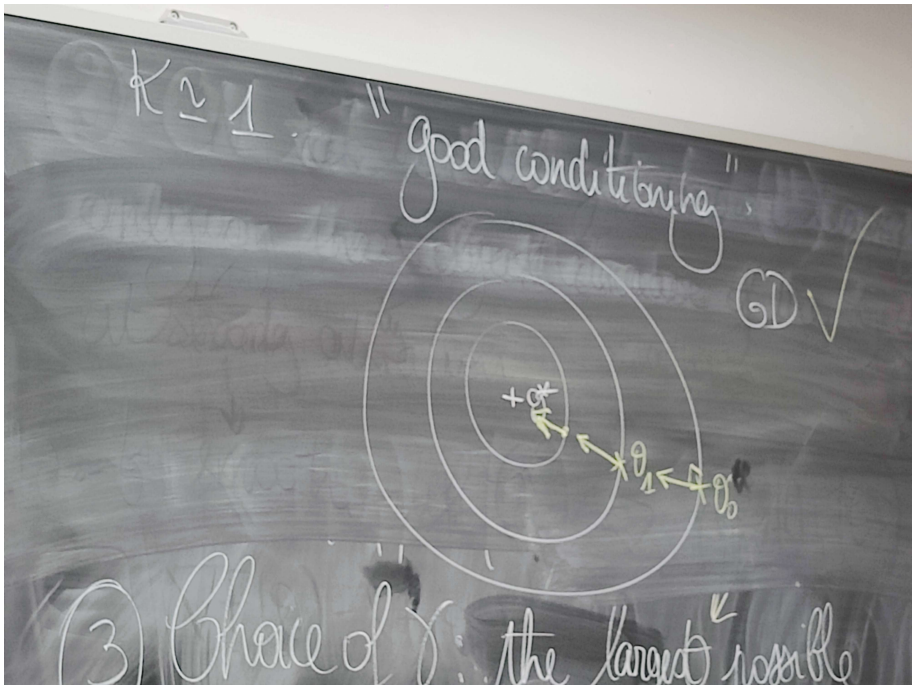


Figure 1.3: With $\kappa \approx 1$ "good conditioning"

Preuve :

$$\begin{aligned}
\|\theta_{t+1} - \theta^\star\|^2 &= \|\theta_t - \gamma \nabla F(\theta_t) - \theta^\star\|^2 \\
&= \|\theta_t - \theta^\star\|^2 - 2\gamma \langle \nabla F(\theta_t), \theta_t - \theta^\star \rangle + \gamma^2 \|\nabla F(\theta_t)\|^2 \\
&= \|\theta_t - \theta^\star\|^2 - 2\gamma \langle \nabla F(\theta_t), \theta^\star - \theta_t \rangle + \gamma^2 \|\nabla F(\theta_t)\|^2
\end{aligned}$$

By μ -strong convexity, we got

$$F(\theta^\star) \geq F(\theta_t) + \langle \nabla F(\theta_t), \theta^\star - \theta_t \rangle + \frac{\mu}{2} \|\theta^\star - \theta_t\|^2$$
$$\Rightarrow \langle \nabla F(\theta_t), \theta^\star - \theta_t \rangle \leq F(\theta^\star) - F(\theta_t) - \frac{\mu}{2} \|\theta^\star - \theta_t\|^2$$

Therefore $\|\theta_{t+1} - \theta^\star\|^2 \leq \|\theta_t - \theta^\star\|^2 - 2\gamma(F(\theta_t) - F(\theta^\star) + \frac{\mu}{2} \|\theta^\star - theta_t\|^2) + \gamma^2 \|\nabla F(\theta_t)\|^2$
Beside, by L-smoothness, we get

$$\begin{aligned}
F(\theta_{t+1}) - F(\theta_t) &= F(\theta_t - \gamma \nabla F(\theta_t)) - F(\theta_t) \\
&= [F(\theta_t \tau \nabla F(\theta_t))]_{\tau=0}^{\gamma} \\
&= -\int_0^\gamma \langle \nabla F(\theta_t), \nabla F(\theta_t - \tau \nabla F(\theta_t)) \rangle \, d\tau \qquad = -\int_0^\gamma \langle \nabla F(\theta_t), \nabla F(\theta_t - \tau \nabla F(\theta_t)) \\
&= -\gamma \|\nabla F(\theta_t)\|^2 + \int_0^\gamma \langle \nabla F(\theta_t), \nabla F(\theta_t) - \nabla F(\theta_t - \tau \nabla F(\theta_t)) \rangle \, d\tau \\
&\leq -\gamma \|\nabla F(\theta_t)\|^2 + \int_0^\gamma \tau L \|\nabla F(\theta_t)\|^2 \, d\tau \\
&\leq -(\gamma - \frac{\gamma^2 L}{2}) \|\nabla (F(\theta_t))\|^2 \text{ using CS + L-smooth}
\end{aligned}$$

Combining the 2 previous inequalities,

$$\begin{aligned}
\left\|\theta_{t+1} - \theta^{star}\right\|^2 &\leq \left\|\theta_t - \theta^{star}\right\|^2 (1 - \gamma\mu) - 2\gamma(F(\theta_t) - F^\star) + \frac{\gamma^2}{\gamma - \gamma^2 \frac{L}{2}} \\
&\leq (1 - \gamma\mu) \|\theta_t - \theta^\star\|^2 - \gamma(\frac{2\gamma - \gamma^2 \frac{L}{2} - \gamma}{\gamma - \gamma^2 \frac{L}{2}})(F(\theta_t) - F^\star)
\end{aligned}$$

using that $F(\theta) \geq F(\theta^\star) \Rightarrow F(\theta_t) - F(\theta_{t+1}) \leq F(\theta_t) - F(\theta^\star)$

- Numerator > 0 when $0 < \gamma \leq 1/L$
- Denominator > 0 when $0 < \gamma < 2/L$

Then by assuming $\gamma \leq \frac{1}{L}$ just ignore the last term and conclude □

Subgradient method

Théorème 7 (GD for non-smooth fonctions)

Hypothese : F convexe, has subgradients, β -Lipschitz

$$\begin{cases} \|\nabla F(\theta)\|^2 \leq \beta^2 \\ \forall \eta \in \partial F(\theta), \|\eta\|^2 \leq \beta^2 \end{cases}.$$

Then GD iterates with Polyak-Ruppert averaging enjoy the following error bound

$$\bar{\theta}_T = \frac{1}{T} \sum_{t=1}^{T} \theta_t.$$

$$\begin{aligned}
F(\bar{\theta_T}) - F(\theta^{star}) &\leq \frac{\|\theta_0 - \theta^\star\|^2}{2\gamma T} + \frac{\gamma \beta^2}{2} \\
&= \left\| \frac{\theta_0 - \theta^\star}{\sqrt{T}} \right\| \text{ for } \gamma = \gamma^\star \text{ (when looking at below figures)}
\end{aligned}$$

$$F(\bar{\theta_T}) - F(\theta^{star}) \leq \frac{\|\theta_0 - \theta^\star\|^2}{2\gamma T} + \frac{\gamma \beta^2}{2}.$$

NB: now there is a trade-off on the choice of γ. Now we have two terms :

- $\frac{\|\theta_0-\theta^\star\|^2}{2\gamma T}$ in purple in Figure 1.4
- $\frac{\gamma\beta^2}{2}$ in green in Figure 1.4

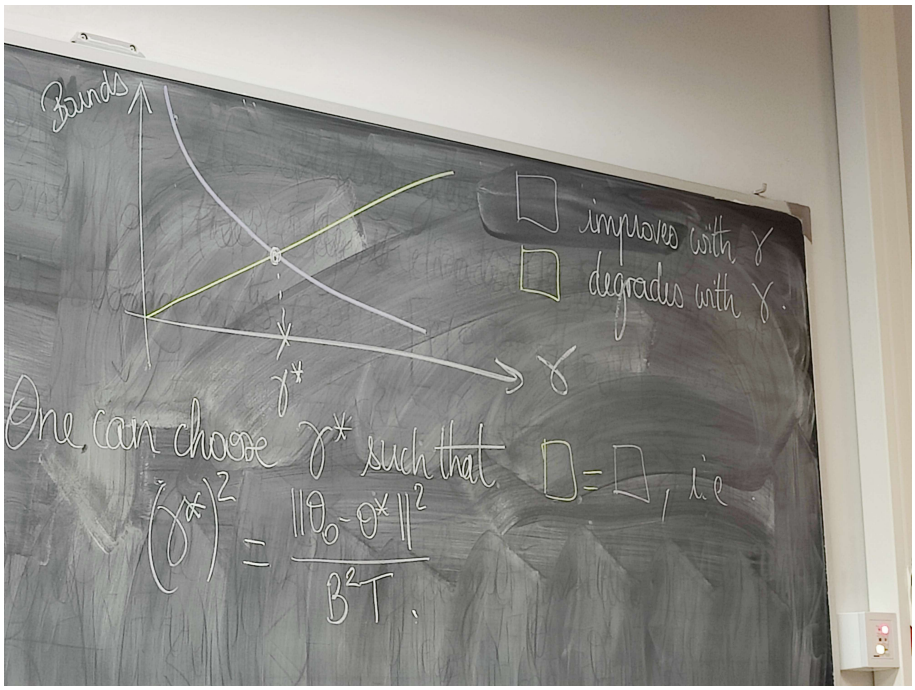


Figure 1.4:

One can choose $\gamma^\star$ such that "purple" = "green" (Figre 1.4), i.e.

$$(\gamma^\star)^2 = \frac{\|\theta_0-\theta^\star\|^2}{\beta^2 T}.$$

- Non-smoothness is paid through a $O(\frac{1}{\sqrt{T}})$ rate.
- Guarantee for $\bar{\theta}_T$
- CCL Big picture : BD-Based strategies
  - convex non-smooth $O(1/\sqrt{T})$
  - convex L-smooth $O(1/T)$
  - mu-strongly convex non-smooth $O((1-\frac{\mu}{L})^T)$

$F(\frac{1}{T}\sum_{t=1}^T \theta_t) - F^\star \leq \frac{1}{T}\sum_{t=1}^T (F(\theta_t)-F^\star)$ by convexity.
And $(F(\theta_t)-F^\star)$ is on $\frac{1}{t}$
So, $F(\frac{1}{T}\sum_{t=1}^T \theta_t) - F^\star \leq \frac{1}{T}\sum_{t=1}^T (F(\theta_t)-F^\star) \lesssim \mathcal{O}(\frac{\log_T}{T})$.

Preuve :

$$\begin{aligned}
\|\theta_{t+1}-\theta^\star\|^2 &= \|\theta_t - \gamma_t g_t - \theta^\star\|^2 \text{ with } g_t \in \partial F(\theta_t) \\
&= \|\theta_t-\theta^\star\|^2 - 2\gamma_t \langle g_t, \theta_t-\theta^\star\rangle + \gamma_t^2 \|g_t\|_2^2 \\
\text{by def of subgradient } &\leq \|\theta_t-\theta^\star\|^2 - 2\gamma_t (F(\theta_t)-F^\star) + \gamma_t^2 \|g_t\|_2^2
\end{aligned}$$

Recursively we obtain

$$\|\theta_{t+1}-\theta^\star\|^2 \leq \|\theta_1-\theta^\star\|^2 - 2\sum_{s=1}^t \gamma_s (F(\theta_s)-F^\star) + \sum_{s=1}^t \gamma_s^2 \|g_s\|_2^2 .$$

Combining this with $\sum_{s=1}^{t} \gamma_s (F(\theta_s) - F^\star) \geq \sum_{s=1}^{t} \gamma_s . \min_{1 \leq s \leq t} (F(\theta_s) - F^\star)$
γ cte + polyak- Ruppert
$t \sum_{s=1}^{t} \frac{\gamma_s}{t} (F(\theta_s) - F^\star) \geq t\gamma (F(\bar{\theta}_t) - F^\star)$
Finally,

$$\begin{aligned} \min_{1 \leq s \leq t} F(\theta_s) - F^\star &\leq \frac{\|\theta_1 - \theta^\star\|_2^2 + \sum_{s=1}^{t} \gamma_s \|g_s\|_2^2}{2 \sum_{s=1}^{t} \gamma_s} \\ &\leq \frac{\|\theta_1 - \theta^\star\|_2^2 + \beta^2 \sum_{s=1}^{t} \gamma_s}{2 \sum_{s=1}^{t} \gamma_s} \\ F(\bar{\theta}_t) - F^\star &\leq \frac{\|\theta_1 - \theta^\star\|^2 + t\gamma^2\beta^2}{2t\gamma} \end{aligned}$$

□

Note (Implicit gradient method). Subgradient method = generalization of GD in the non-smooth case but O is typically slow ($\frac{1}{\sqrt{T}}$).

The essential reason is that there are plenty of subgradients that are large near and event at the solution.

$g \in \partial F(\theta)$ if $\forall \theta', F(\theta') \geq F(\theta) + \langle g, \theta' - \theta \rangle$

$$\partial ?(\theta) = \begin{cases} \{+1\} & \text{if } \theta > 0 \\ \{-1\} & \text{if } \theta < 0 \\ [-1, 1] & \text{if } \theta = 0 \end{cases}.$$

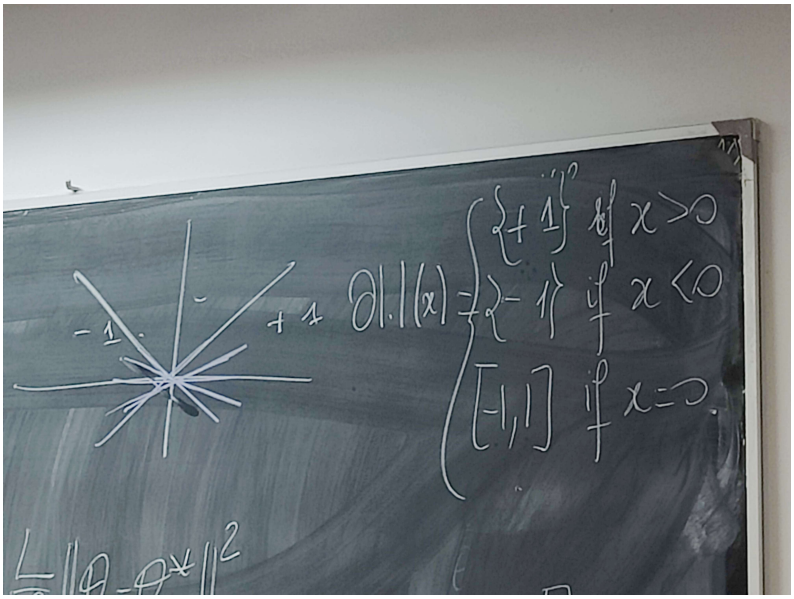


Figure 1.5: sub gradiens

Another way to deal with this is to add a smooth regularized term. In particular, if $\theta^\star$ is minimizer of F then it minimizes as well

$$\theta \mapsto F(\theta) + \gamma \|\theta - \theta^\star\| \text{ for } gamma > 0$$

Now the regulatized fonction is strongly convexe and the only subgradient at the solution is the zero vector :

- **Good** : It addresses tje main drawback of subgrad methods
- **Bad** : We have to know $\theta^\star$

One can implement an iterative version of it, this is the proximal algo :

$$\theta_{t+1} = \arg\min_{\theta} F(\theta) + \frac{1}{2\gamma_t} \|\theta - \theta_t\|^2$$

When F is convex, $F + \frac{1}{2\gamma_t} \|\circ - \theta_t\|^2$ is strictly convexe so the mapping is well defined. This gives the proximal operator / Moreau envelope.

$$prox_{\gamma_t F}(\theta) = \arg\min_{\tilde{\theta}} F(\tilde{\theta}) + \frac{1}{2\gamma_t} \left\|\theta - \tilde{\theta}\right\|^2.$$

The proximal operator can be interpreted as a variation of gradient methods

$$\begin{cases} \frac{d\theta}{dt}(t) = -\nabla F(\theta) \\ \theta(0) = \theta_0 \in \mathbb{R}^d \end{cases}.$$

The equilibrium points of this system are the θ's such that $\nabla F(\theta) = 0$, i.e the minimizers of F when F is convex

GD = 1st order numerical method for tracing the path from θ_0 to $\theta^\star$

$$\frac{\theta(t+h) - \theta(t)}{h} \approx -\nabla F(\theta(t)).$$

GD $\equiv$ Forward Euler discretization.
But we could use Backward instead

$$\frac{\theta(t) - \theta(t-h)}{h} \approx -\nabla F(\theta(t)).$$

And now the iterates obey :

$$\theta_{t+1} = \theta_t - h\nabla F(\theta_{t+1}) \qquad \text{"Implicit".}$$

Their construction is not straight forward anymore. But this is what the prox operator actually computes

$$\begin{aligned} \theta_{t+1} &= \arg\min F(\theta_t) + \frac{1}{\gamma_t} \|\theta - \theta_t\|^2 \\ \Leftrightarrow 0 &= \nabla F(\theta_{t+1}) + \frac{1}{\gamma_t}(\theta_{t+1} - \theta_t) \end{aligned}$$

Note (Newton's method). Given θ_{t-1}, the Newtons's method minimizes the 2nd ordre Taylor expansion arount θ_{t-1}

$$\theta \mapsto F(\theta_{t-1}) + \langle \nabla F(\theta_{t-1}, \theta - \theta_{t-1}) \rangle + \frac{1}{2}(\theta - \theta_{t-1})^T Hess_F(\theta_{t-1})(\theta - \theta_{t-1}).$$

the gradient of this quadratic form is

$$\nabla F(\theta_{t-1}) + H_F(\theta_{t-1})^{-1} \nabla F(\theta_{t-1}).$$

Exercise : Check that $-H_F(\theta_{t-1})^{-1}\nabla F(\theta_{t-1})$ is indeed a descent direction of F at θ_{t-1}.

Newton's method are methods of order 2 : using the gradient (order 1) and the Hessian (order 2). Running-time complexity is $O(d^3)$ in general to solve the linear system.

It leads to local quadratic CV :

$$(C \|\theta_t - \theta^\star\|) \leq (C \|\theta_t - \theta^\star\|)^2.$$

For global convergence guarantees, see *Boyd & Vandenberghe (2004)* in particular using the self-concordance relating 3rd and 2nd order derivatives.

1.2 Inertial methods

1.2.1 Préliminaries

So far we have

- convex, L-smooth : $O(1/k)$
- strongly convex, L-smooth : $O((1-\frac{\mu}{L})^k)$

Can we do better with a **gradient-like** algo ?

Définition 8

A gradient-like algo is an algo such taht

$$\theta_{t+1} \in span\{\theta_0, \dots, \theta_t, \nabla F(\theta_0), \dots, \nabla F(\theta_t)\}.$$

Théorème 9 (Nemirovski-Rudin 1983)

$\forall \theta_0 \in \mathbb{R}^d, \forall 0 \leq t \leq \frac{d-1}{2}$
$\exists F$ convex, L-smooth such that for every gradient-like algon we have

$$F(\theta_t) - \inf F \geq \frac{3L \left\|\theta^0 - \theta^\infty\right\|}{32(\mathbf{t+1})^\mathbf{2}}.$$

Théorème 10 (Nesterov 2003)

$\forall \theta_0 \in \mathbb{R}^d, \mu > 0, L > 0, \exists F$ mu-strongly convexe and L-smooth such that for every gradient-like algo

1. $F(\theta_t) - \inf F \geq \frac{\mu}{2}(\frac{1-\sqrt{\kappa}}{1+\sqrt{\kappa}})^{2t} \left\|\theta_0 - \theta^\star\right\|$
2. $\|\theta_t - \theta^\star\| \geq (\frac{1-\sqrt{\kappa}}{1+\sqrt{\kappa}})^t \left\|\right\| \theta_0 - \theta^\star \quad$ with $\kappa = \frac{\mu}{L}$

Can we design first-order strategies that achieve convergence rates matching these lower bounds ?

1.2.2 Heavy ball dynamics

$$\ddot{\theta}(t) = -\alpha(t)\dot{\theta} - \nabla F(\theta(t)), (\alpha(t) > 0).$$

We add a function term to the gradient flow.

We can have a look at the quantity

$$\epsilon(t) = F(\theta(t)) - \inf F + \frac{1}{2}\left\|\dot{\theta}(t)\right\|^2 = E_{pot} + E_{cin}.$$

We can show that $\epsilon(t)$ is decreasing (this is a Lyapunov energy)

$$\begin{aligned}
&\dot{\epsilon}(t) \\
&\quad = \left\langle \nabla F(\theta(t)), \dot{\theta}(t) \right\rangle + \left\langle \ddot{\theta}(t), \dot{\theta}(t) \right\rangle \\
&\quad = \left\langle \ddot{\theta}(t) + \nabla F(\theta(t)), \dot{\theta}(t) \right\rangle \\
&\quad = -\alpha(t)\left\|\dot{\theta}(t)\right\|^2 \; () \leq \mathbf{0})
\end{aligned}$$

Note. $\alpha(t) \equiv 0$ gives a conservative dynamics with aliittle hope of CV.

$$F(\theta) = \frac{1}{2}\theta^2, \alpha = 0$$
$$\ddot{\theta}(t) = -\theta(t) \Leftrightarrow \theta(t) = c_1 \sin(t) + c_2 \cos(t)$$

Why it can help? Gabriel Goh "Why momentum really works".

$$\ddot{\theta}(t) = -\alpha(t)\dot{\theta}(t) - \nabla F(\theta(t)).$$

Discretization

$$\theta(t_k) \approx \theta_k$$
$$\dot{\theta}(t_k) \approx \frac{\theta_k - \theta_{k-1}}{h}$$
$$\ddot{\theta}(t_k) \approx \frac{\dot{\theta}(t_{k+1}) - \dot{\theta}(t_k)}{h}$$
$$\frac{\theta_{k+1} - 2\theta_k + \theta_{k+1}}{h^2} + \alpha(t_k)\frac{\theta_k - \theta_{k-1}}{h} + \nabla F(\theta_k) = 0$$

Define $\gamma = h^2$ $\alpha_k = \frac{\alpha(t_k)}{\sqrt{\gamma}}$ we get :

$$\theta_{k+1} = \theta_k - \gamma \nabla F(\theta_k) + (1 - \alpha_k)(\theta_k - \theta_{k-1}).$$

where $\gamma \nabla F(\theta_k)$ is the gradient step and
$(1 - \alpha_k)(\theta_k - \theta_{k-1})$ is the inertia : memory of the last iterates. [polyak 64]

HEAVYBALL [Polyak, 64]

$$\beta_k = \theta_k + (1 - \alpha_k)(\theta_k - \theta_{k-1})$$
$$\theta_{k+1} = \beta_k - \gamma \nabla F(\theta_k)$$

NESTEROV ALGO [83]

$$\beta_k = \theta_k + (1 - \alpha_k)(\theta_k - \theta_{k-1})$$
$$\theta_{k+1} = \beta_k - \gamma \nabla F(\beta_k)$$

They look the same, the only difference is where the gradient is evaluated. Both algo come with 2 choices for the friction α_k

- constant friction $\alpha_k \equiv \alpha\sqrt{\gamma}$ (for good functions)
- vanishing friction $\alpha_k \equiv \frac{\alpha}{k}$ (for bad functions)

HEAVY BALL

Théorème 11 (polyak 64, écrit vite fait parce que c'est la fin du cours)

F quadratic -smooth, mμ- strongly cvx, $\kappa = \frac{\mu}{L}$ with,

$$\begin{cases} \gamma = \frac{4}{L(1+K)^2} \\ \alpha_k = \frac{2\sqrt{\mu\gamma}}{1+\sqrt{\kappa}} \end{cases}.$$

CV rate $\mathcal{O}((\frac{1-\kappa}{1+\kappa})^t)$
Cool : We have Optimal rate and constant friction is enough
But : HB can fail on general strongly convexe fonction and need to know μ (and L)

NESTEROV

Théorème 12

F L-smooth, μ-strongly cvx Choose $\gamma = 1/L, \alpha = \frac{\sqrt{L}-\sqrt{\mu}}{\sqrt{L}+\sqrt{\mu}}$ to get $(1-\sqrt{\frac{\mu}{L}})$ -linear CV (convergence)
Cool : Better GD
Questionnable : Not optimal

Théorème 13 (Nesterov 83, Chambolle-Dossal 2015)

F convex, L-smooth $\gamma \leq 1/L, \alpha_k = \alpha/k$ with $\alpha \geq 3$

$$F(\theta_k) - F^\star \leq O(\frac{1}{k^2}).$$

Cool : Optimal

We can take other choices for decreasing $(\alpha_k)_k$, the historical choice is

$$(Friction :)(1-\alpha_k) = \frac{t_k - 1}{t_{k+1}} \text{ with } \begin{cases} t_1 = 1 \\ t_{k+1} = \frac{1+\sqrt{1+4t_k^2}}{2} \end{cases} .$$

CCL : Essayer les deux méthodes : speed upped or not

	GD	Nesterov $\alpha_k \searrow$	Nesterov $\alpha_k = \alpha$	Heavy ball $\alpha_k \searrow$	Heavy ball $\alpha_k = \alpha$
convex + smooth	$O(1/k)$	$O(1/k^2)$ ✓	$O(1/k)$ ✗		$O(1/k)$ ✗
smooth + strongly convex ↓	$O((\frac{1-\kappa}{1+\kappa})^k)$		good but not opti ✓ $O((1-\sqrt{\kappa})^k)$		✗ may not converge [Guelluim] [Diakonikolas]
+ quadratic	$O((\frac{1-\kappa}{1+\kappa})^k)$	[Boyd, Su, Candès] $O(1/k^3)$ ✗	good but not opti ✓ $O((1-\sqrt{\kappa})^k)$		✓ optimal $O((\frac{1-\sqrt{\kappa}}{1-\sqrt{\kappa}})^k)$
quadratic but not strongly convex	Linear rate on $\ker^\perp$ Otherwise $O(1/k)$	$O(1/k^2)$		$O(1/k^2)$	

Figure 1.6: Tableau de la vitesse de convergence des algos

CCL : vanishing friction helps on the worst fcts

Chapter 2

Stochastic Gradient Algorithms (SGD)

$$\min_{\theta \in \mathbb{R}^d} F(\theta).$$

At any step, assume that we have access to a "random" direction / gradient : $g_t : \mathbb{R}^d \to \mathbb{R}^d$
$\forall t \geq 0$, $\theta_{t+1} = \theta_t - \gamma_t g_{t+1}(\theta_t)$ and $\gamma_t =$ (learning rate) / (step size)
Think of g_t as a noisy estimate of the "true" gradient, we would like to use instead

Figure 2.1: Noisy gradient descent

Hypothesis: [Unbiased estimates of the gradient]

$$\mathbb{E}[g_t(\theta_{t-1})|\theta_{t-1}] = \nabla F(\theta_{t-1}).$$

θ_{t-1} encapsulates all the randomness due to> the past iterations, so we only require "fresh" randomness at time t.

2.1 SDG in machine learning

There are 2 ways to use SGB in supervised learning

2.1.1 Empirical risk minimization

If $F(\theta) = \frac{1}{n}\sum_{i=1}^n l(Y_i, f_\theta(X_i))$ then at iteration t, we can choose uniformly at random $i(t) \approx \mathcal{U}([1,n])$ and define g_t as the gradient of $l_{i(t)} : \theta \mapsto l(Y_{i(t)}, f_\theta(X_{i(t)}))$ A full GD would use $\nabla F(\theta) = \frac{1}{n}\sum_{i=1}^n \nabla l_i(\theta), g_t := \nabla l_{i(t)}(\theta)$ for $l_i(\theta) = l(Y_i, f_\theta(X_i))$, i.e. the n gradients of the terms composing the sum. SGD relies on a "moisy" estimate of $\nabla F(\theta)$ by selecting at random only one term $\nabla l_{i(t)}$

Conditionnally to the training data, we aim at minimizing a deterministic fonctions using a stochastic algo to help on complexity issues. Indeed, the randomness comes from the random indeces $(i_(t))_t$.
There exist minibatch versions where at each iteration the gradient is estimated over a random subset of indices

- reducing the variance of the estimated gradient
- increasing the running time

The theoritical analysis focuses on the CV to the ERM $\theta^\star$:

$$I \sim \mathcal{U}([1;n])$$

$$\mathbb{E}[g_t(\theta)|\theta] = \mathbb{E}[\nabla l_I(\theta)|\theta] = \sum_{i=1}^n \mathbb{P}(I=i)\nabla l_i(\theta)$$

$$= \frac{1}{n}\sum_{i=1}^n \nabla l_i(\theta) = \nabla F(\theta)$$

We can select several times the same ∇l_i even within n iterations sampling without replacement can be possible but its analysis is more involved (need to handle the bias) see Nagaraj et al 2019

2.1.2 Expected risk minimization

$$F(\theta) = \mathbb{E}[l(Y, f_\theta(X))].$$

expected (non-observable) risk, then at each iteration t, we can take (X_t, Y_t) and define g_t as the gradient of $\theta \mapsto l(Y_t, f_\theta(X_t))$ By swapping the order of expectation and differentiation, we can get unbiased estimators

$$\mathbb{E}_{(X_t,Y_t)}[\nabla_\theta l(Y_t, f_\theta(X_t))] = \nabla_\theta \mathbb{E}[l(Y_t, f_\theta(X_t))]_t.$$

Sanity-check for linear regression :

$$F(\theta) = \mathbb{E}[(Y - \langle X, \theta\rangle)^2] = \mathbb{E}[f(\theta)].$$

$$\nabla f(\theta) = 2(\langle X, \theta\rangle)X.$$

$$\|\nabla f(\theta)\| \leq 2\left|\langle X, \theta\rangle - Y\right| \|X\|.$$

If $\forall\theta$, $\mathbb{E}[|\langle X, \theta\rangle| \|X\|] < +\infty$, the $\nabla_\theta F(\theta) = \nabla_\theta \mathbb{E}[(Y - \langle X, \theta\rangle)^2] = \mathbb{E}[\nabla_\theta f(\theta)]$

Note that to preserve the unbiasedness, only a **signle pass** is allowed.

Here, we directly minimize the generalization risk. As we perfom only one pass, with n data, we can run only n SDG iteration. As one can hope that $(\theta_t)_t$ converge to $\omega\theta^\star$ a minimizer of the expected risk.

In practice, multiple passes are used (and theorelical guarantees fall)

Note (**warning**). SGD is not a descent method : the function values often go up but in **expectaiton** they go down

In what follows we will handle both situations with a unified view.

2.1.3 First impressions on SGD

Set for $i \geq 1, F_i(\theta) = \frac{1}{2}(\theta - a_i)^2, a_i \sim \mathcal{U}([-1, 1])$.

This means that when the data come in a streaming fashion, our goal is to minimize $\theta \mapsto^F \mathbb{E}[\frac{1}{2}(\theta - a)^2]$ that we know to be optimal at $\theta^\star = \mathbb{E}[a]$.

Without knowing the distribution of $(a_i)_i$ one can use SGD strategy to estimate $\theta^\star = \mathbb{E}[a]$

$$\forall t \geq 0, \begin{cases} \theta_t = \theta_{t-1} - \gamma_t g_t(\theta_[t-1]) \\ \theta_0 = cst \end{cases}$$

$$g_t(\theta_{t-1} = \theta_{t-1} - a_t)$$

$$\theta_t = (1 - \gamma_t)\theta_{t-1} + \gamma_t a_t$$

If we choose $\gamma_t = \gamma$ (cst),

$$\theta_t = ... = (1 - \gamma)^t\theta_0 + \gamma\sum_{k=0}^{t}(1 - \gamma)^k a_{t-k}.$$

The first term shrinks to 0 (we forget the initial condition) if $\gamma \leq 1(= 1/L), L = 1$

$$\begin{aligned} \nabla F(\theta) &= \mathbb{E}[\theta - a] \\ &= \theta(\text{which is 1-Lip}) \end{aligned}$$

Note that $\forall\theta$

$$\begin{aligned} \mathbb{E}[(g_t(\theta) - \nabla F(\theta))^2] &= \mathbb{E}[(\theta - a - \theta)^2] \\ &= \mathbb{E}[a^2], a \sim \mathcal{U}([-1, 1]) \\ &= 1/3(2^2/12) \end{aligned}$$

Our gradients enjoy a uniform bound on their variance.

If we continue the calculation

$$F(\theta^\star) = F(0) = \mathbb{E}[\frac{1}{2}a^2] = \frac{1}{6}$$

$$\begin{aligned} \mathbb{E}[F(\theta_t) - F(\theta^\star)] &= \mathbb{E}[\frac{1}{2}(\theta_t - a)^2] - \frac{1}{6} \\ &= \frac{1}{2}\mathbb{E}[\theta_t^2] \end{aligned}$$

$$\mathbb{E}[\theta_t^2] = Var((1-\gamma)^t\theta_0 + \gamma\sum_{k=1}^{t}(1-\gamma)^k a_{t-k}) + (\mathbb{E}[\theta_t])^2$$

$$= \frac{1}{3}\gamma\frac{1-(1-\gamma)^{2(t+1)}}{1-(1-\gamma)^2} + (1-\gamma)^{2t}\theta_0^L$$

$$\rightarrow_{t\rightarrow+\infty} \begin{cases} \frac{1}{3}\gamma & \text{if } \gamma = 1 \\ \frac{1}{3}\frac{\gamma}{2\gamma-\gamma^2} & \text{if } 0 < \gamma < 1 \end{cases}$$

WHICH DOES NOT TEND TO 0 WHEN $t \rightarrow +\infty$
Obviously the variance $Var[\nabla F_1(\theta^*)] = 1/3$ at the solution is a big problem. Having a vanishing step size could help ! What about Polyak-Reppert averaging ?

Nouveau cours du 29/11

Rappel du cours précédent je crois

$$\theta_{t+1} = \theta_t - \gamma_{t+1} g_{t+1}(\theta_t)$$
$$\theta_t \in \mathbb{R}^d$$

$(g_t)_t$ noisy estimations of ∇F of the true objective fct

Hypothesis : $\mathbb{E}[g_{t+1}(\theta_t)|\theta_t] = \nabla F(\theta_t)$ Unbiased estimates

1. ERM : $F(\theta)\frac{1}{n}\sum_{i=1}^n F_i(\theta)$
2. True risk minimization $F(\theta) = \mathbb{E}[l(y, f_\theta(X))] = \mathbb{E}[l(Y_i, f_\theta(X_i))] = \mathbb{E}[F_i(\theta)]$

First impression : $F_i(\theta) = \frac{1}{2}(\theta - a_i)^2$, $a_i \sim \mathcal{U}([-1,1])$

$$\begin{cases} \theta_t = \theta_{t-1} - \gamma_t(\theta_{t-1} - a_t) \\ \theta_0 = cste \end{cases}.$$

$$\mathbb{E}[F(\theta_t) - F^\star] \not\to_{t\to+\infty} 0.$$

Because of $Var[\nabla F_i(theta^\star)] = \dfrac{1}{3}$

- Vanishing step size
- Polyak-Ruppert av $\bar{\theta_T} = \frac{1}{T+1}\sum_{t=0}^T \theta_t$

$$\mathbb{E}[F(\bar{\theta}_t) - F^\star] = \frac{1}{2}\mathbb{E}[\bar{\theta}_2^2]$$

$$\bar{\theta}_t = \frac{1}{T+1}\sum_{t=0}^T \theta_t$$
$$= \frac{1}{T+1}\sum_{t=0}^T (1-\gamma)^t\theta_0 + \frac{\gamma}{T+1}\sum_{t=0}^T\sum_{k=0}^t (1-\gamma)^k a_{t-k}$$

$$\sum_{t=0}^T\sum_{k=0}^t (1-\gamma)^{t-k}a_k = \sum_{k=0}^T\sum_{t=k}^T (1-\gamma)^{t-k}a_k$$
$$= \sum_{k=0}^T \frac{a_k}{(1-\gamma)^k}\sum_{t=k}^T (1-\gamma)^t$$
$$= \frac{1}{\gamma}\sum_{k=0}^T (1-(1-\gamma)^{T-k-1})a_k$$

In consequence,

$$\mathbb{E}[(\bar{\theta_T} - 0)^2] = (\frac{1}{T+1}\frac{1-(1-\gamma)^{T+1}}{1-(1-\gamma)}\theta_0)^2$$
$$+ \mathbb{E}[(\frac{\gamma}{T+1}\sum_{k=0}^T (1-(1-\gamma)^{T-k-1})a_k)^2]$$
$$....$$
$$\leq (\frac{1}{T+1}\frac{1-(1-\gamma)^{T+1}}{1-(1-\gamma)}\theta_0)^2 + \frac{\gamma^2}{(T+1)^2}(T+1)\frac{1}{3}$$
$$= (\frac{1}{T+1}\frac{1-(1-\gamma)^{T+1}}{1-(1-\gamma)}\theta_0)^2 + \frac{\gamma^2}{(T+1)}\frac{1}{3}$$
$$\to_{T\to+\infty} 0$$

In this *specific* quadratic setting, the polyak-Ruppert averaging is enough to average the noise out around the solution despite a constant step-size.
Warning: Valid only for quadratic fonction.

More generally,

Théorème 14

Hypothesis :

1. F is L-Smooth and convex
2. Unbiased gradients : $\mathbb{E}[g_t(\theta_{t-1})|\theta_{t-1}] = \nabla F(\theta_{t-1})$
3. Bounded variance uniformly : $\forall\theta\mathbb{E}[\|g_t(\theta) - \nabla F(\theta)\|^2 |\mathcal{F}_{t-1}] \leq \sigma^2$ with $\mathcal{F}_{t-1}$ the filtration such that θ_t is $\mathcal{F}_t$ -measurable.

More explanation on $\mathcal{F}_t$ in Figure 2.2 Then $\forall\gamma \leq 1/L$, the SGD iterates with Pdeak-Ruppert averagin satisfy

$$\mathbb{E}[F(\bar{\theta}_T) - F(\theta^\star)] \leq \frac{\|\theta_0 - \theta^\star\|^2}{2\gamma(1 - \frac{\gamma^2}{2})T} + \frac{\gamma\sigma^2}{2}.$$

For $\forall t \geq 1$

$$\begin{cases} \theta_t = \theta_{t-1} - \gamma g_t(\theta_[t-1]) \\ \theta_0 \in \mathbb{R}^d \end{cases}.$$

$$\bar{\theta}_T = \frac{1}{T}\sum_{t=1}^{T} \theta_t.$$

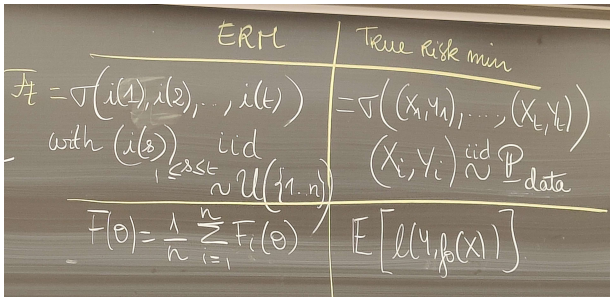


Figure 2.2: More explaination of $\mathcal{F}_t$

Note.

- 2 terms
  - optimization term $\frac{\|\theta_0-\theta^\star\|^2}{2\gamma(1-\frac{\gamma^2}{2})T}$ similar to that of GD int the smooth case
  - The variance term $\frac{\gamma\sigma^2}{2}$ impact of the noise which increase with γ and σ^2
- Behaviour w.r.t γ
  - Because of optimisation : $\frac{\|\theta_0-\theta^\star\|^2}{2\gamma(1-\frac{\gamma^2}{2})T}$ it goes up
  - Bacause of Variance term : $\frac{\gamma\sigma^2}{2}$ it goes down
- Best trade-off is for $\gamma \approx \frac{\|\theta_0-\theta^\star\|}{4L\sqrt{T}\sigma}$ (constant step size but depending on the finit horizon)
- Comment the assumptions in the case ERM

$$F(\theta) = \frac{1}{n}\sum_{i=1}^{n} F_i(\theta).$$

(ii) is satisfied whenever $\forall t \geq 1, g_t = \nabla F_{i(t)}$ with $i(t) \sim \mathcal{U}(\{1, \dots, n\})$.
Assume that (iii) holds for $g_t = \nabla F_{i(t)}$ (usual SGD).
One my use as gradient estimates $g_t^{|B|} = \frac{1}{|B|} \sum_{i \in B_t} \nabla F_i$ where B_t is of cardinality $|B_t| = |B|$ uniformly ddrawn at random in $\{1, \dots, n\}$
This is called a **mini batch strategy**

$$\begin{aligned}
\mathbb{E}[\left\| g_t^{|B|}(\theta) - \nabla F(\theta) \right\|_2^2 |\mathcal{F}_{t-1}] &= \mathbb{E}[\left\| \frac{1}{|B|} \sum_{i \in B_t} \nabla F_i(\theta) \right\|_2^2 |\mathcal{F}_{t-1}] \\
&= \mathbb{E}[\left\| \frac{1}{|B|} \sum_{i \in B_t} (\nabla F_i(\theta) - \nabla F(\theta)) \right\|_2^2 |\mathcal{F}_{t-1}] \\
&= \frac{1}{|B|^2} |B| \sigma^2 = \frac{\sigma^2}{|B|}
\end{aligned}$$

Proof.

$$\|\theta_{t+1} - \theta^\star\|_2^2 = \|\theta_t - \theta^\star\|^2 - 2\gamma_t \langle g_{t+1}(\theta_t), \theta_t - \theta^\star \rangle + \gamma_t^2 \|g_{t+1}(\theta_t)\|_2^2 .$$

Appying $\mathbb{E}[-1\mathcal{F}_t]$ gives

$$\mathbb{E}[\|\theta_{t+1} - \theta^\star\|_2^2 |\mathcal{F}_t] = \|\theta_t - \theta^\star\|_2^2 - \mathbb{E}[2\gamma_t \langle \nabla F(\theta_t), \theta_t - \theta^\star \rangle |\mathcal{F}_t] + \gamma_t^2 \mathbb{E}[\|g_{t+1}(\theta_t)\|_2^2 |\mathcal{F}_t].$$

$\|\theta_t - \theta^\star\|_2^2$ is $\mathcal{F}_t$-mesurable. and $\mathbb{E}[-2\gamma_t \langle \nabla F(\theta_t), \theta_t - \theta^\star \rangle |\mathcal{F}_t] = -2\gamma_t \langle \mathbb{E}[\nabla F(\theta_t)|\mathcal{F}_t], \theta_t - \theta^\star \rangle = -2\gamma_t \langle g_{t+1}(\theta_t), \theta_t - \theta^\star \rangle$ as $\theta_t - \theta^\star$ is $\mathcal{F}_t$-mesurable and with (ii).

$$\begin{aligned}
\gamma_t^2 \mathbb{E}[\|g_{t+1}(\theta_t)\|_2^2 |\mathcal{F}_t] &= \gamma_t^2 \mathbb{E}[\|g_{t+1}(\theta_t) - \nabla F(\theta_t) + \nabla F(\theta_t)\|_2^2 |\mathcal{F}_t] \\
&= \gamma_t^2 \mathbb{E}[\|g_{t+1}(\theta_t) - \nabla F(\theta_t)\|_2^2 |\mathcal{F}_t] + \gamma_t^2 \|\nabla F(\theta_t)\|_2^2 + 2\gamma_t^2 \langle \mathbb{E}[g_{t+1}(\theta_t)|\mathcal{F}_t] - \nabla F(\theta_t), \nabla F(\theta_t) \rangle \\
&= \gamma_t^2 \mathbb{E}[\|g_{t+1}(\theta_t) - \nabla F(\theta_t)\|_2^2 |\mathcal{F}_t] + \gamma_t^2 \|\nabla F(\theta_t)\|_2^2 + 2\gamma_t^2 * 0 \text{ not sure it's what she mean} \\
&\leq \gamma_t^2 \sigma^2 + \gamma_t \|\nabla F(\theta_t)\|_2^2 \\
&\leq \gamma_t^2 \sigma^2 + \gamma_t^2 L \langle \nabla F(\theta_t), \theta_t - \theta^\star \rangle \\
&\text{by cocoercivity of the gradient } \nabla F
\end{aligned}$$

We get

$$\mathbb{E}[\|\theta_{t+1} - \theta^\star\|_2^2 |\mathcal{F}_t] \leq \|\theta_t - \theta^\star\|_2^2 + (-2\gamma_t + \gamma_t^2 L) \langle \nabla F(\theta_t), \theta_t - \theta^\star \rangle + \gamma_t^2 \sigma^2.$$

By convexity of F,

$$\begin{aligned}
F(\theta^\star) &\geq F(\theta e_t) + \langle \nabla F(\theta_t), \theta^\star - \theta_t \rangle \\
\text{i.e. } F(\theta_t) F^\star &\leq \langle \nabla F(\theta_t), \theta_t - \theta^\star \rangle
\end{aligned}$$

If $\gamma_t = \gamma \leq 1/L$, then

$$\begin{aligned}
0 \leq \gamma_t L &\leq 1 \\
-2 \leq -2 + \gamma_t L &\leq -1 \\
-2\gamma_t + \gamma_t^2 L &\leq -\gamma_t
\end{aligned}$$

Therefore

$$\gamma \mathbb{E}[F(\theta_t) - F^\star] \leq \mathbb{E}b[\|\theta_t - \theta^\star\|_2^2] - \mathbb{E}[\|\theta_{t+1} - \theta^\star\|_2^2] + \gamma^2 \sigma^2.$$

Using Jensen's inequality, $F(\bar{\theta_T}) = \frac{1}{\gamma T} \sum_{t=1}^T F(\theta_t)$.
Finaly

$$\mathbb{E}[F(\bar{\theta}_T - F^\star)] \leq \frac{1}{T} \sum_{t=1}^T \mathbb{E}[F(\theta_t) - F^\star] \leq \frac{1}{\gamma T} \|\theta_0 - \theta^\star\|_2^2 + \gamma^2 \sigma^2.$$

□

Théorème 15

F μ-strongly cvx L-smooth. Ball "$\kappa = L/\mu$"

Choose $\gamma_t = \begin{cases} \frac{1}{2L} & \text{for } t \leq 4\lceil\kappa\rceil \\ \frac{2t+1}{(t+1)^2\mu} & \text{for } t > 4\lceil\kappa\rceil \end{cases}$

If $t \geq 4\lceil\kappa\rceil$, then

$$\mathbb{E}[\|\theta_t - \theta^\star\|_2^2] \leq \frac{\sigma^2 4}{\mu t} + \frac{16\lceil\kappa\rceil^2}{ct^2}\|\theta_0 - \theta^\star\|_2^2.$$

Preuve : See [Gower.] 2014 / 2016.
TD : In the case μ -strongly convex

- $\gamma = \frac{2}{\mu(t+1)}$
- $\|g_t(\theta)\| \leq b$ a.s. $\forall\theta$
- $\theta_t = proj_B(\theta_{t-1} - \gamma_t g_t(\theta_{t-1}))$

□

Note.

- **Good:** The result hold for the objective fonction

$$F(\theta) - F(\theta^\star) \leq \langle\nabla F(\theta^\star), \theta - \theta^\star\rangle + \frac{L}{2}\|\theta - \theta^\star\|_2^2.$$

$$\mathbb{E}[\|\theta_t - \theta^\star\|_2^2] = O(1/t) \Rightarrow \mathbb{E}[F(\theta_t) - F^\star] = O(1/t).$$

- **Bad:** With this proof strategy, we do not see the benefit of P-R averaging

$$\mathbb{E}[F(\bar{\theta}_T) - F^\star] \leq \frac{1}{T}\mathbb{E}[\sum_{t=1}^{T} F(\theta_t) - F^\star] \leq 0(\frac{\log T}{T}).$$

We have proven this

Théorème 16

F smooth, unbiased gradients, uniformly bounded variance

$$\mathbb{E}[F(\bar{\theta}_T) - F^\star] \leq \frac{\|\theta_0 - \theta^\star\|_2^2}{\gamma T} + \gamma\sigma^2.$$

For $\gamma \propto 1/\sqrt{T}$

$$= O(1/\sqrt{T}).$$

$$\mathbb{E}[F(\theta_t) - F^\star] \leq \frac{1}{\gamma T}\|\theta_0 - \theta^\star\|^2 + \gamma\sigma^2.$$

Exercice TD : In the case μ -strongly convex

- $\gamma = \frac{2}{\mu(t+1)}$
- $\|g_t(\theta)\| \leq b$ a.s. $\forall\theta$
- $\theta_t = proj_B(\theta_{t-1} - \gamma_t g_t(\theta_{t-1}))$

Exercice 1 - TD3

1.

$$\begin{aligned}\|\theta_t - \theta^\star\|_2^2 &= \|proj_B(\theta_{t-1} - \gamma_t g_t(\theta_{t-1})) - \theta^\star\|_2^2 \\ &= \|proj_B(\theta_{t-1} - \gamma_t g_t(\theta_{t-1})) - proj_B(\theta^\star)\|_2^2 \\ &\leq \|\theta_{t-1} - \gamma_t g_t(\theta_{t-1}) - \theta^\star\|_2^2 \\ &= \|\theta_t - \theta^\star\|_2^2 + \gamma_t^2 \|g_t(\theta_{t-1})\|_2^2 - 2\gamma_t \langle g_t(\theta_{t-1}), \theta_{t-1} - \theta^\star\rangle\end{aligned}$$

$$\mathbb{E}[\|\theta_t - \theta^\star\|_2^2 |\mathcal{F}_{t-1}] \leq \mathbb{E}[\|\theta_{t-1} - \theta^\star\|_2^2 + \gamma_t^2 \|g_t(\theta_{t-1})\|_2^2 - 2\gamma_t \langle g_t(\theta_{t-1}), \theta_{t-1} - \theta^\star\rangle |\mathcal{F}_{t-1}] \quad \leq \|\theta_{t-1} - \theta^\star\|_2^2 + \gamma_t^2 b^2 - 2\gamma$$

2. By μ - strong convex, $F(y) - F(x) \geq \langle \nabla F(x), y - x\rangle + \frac{\mu}{2}\|y - x\|_2^2, \forall x, y$
for $x = \theta_{t-1}$ and $y = \theta^\star$,

$$F(\theta_{t-1}) - F(\theta^\star) \leq \langle \nabla F(\theta_{t-1}), \theta_{t-1} - \theta^\star\rangle + \frac{\mu}{2}\|\theta_{t-1} - \theta^\star\|_2^2.$$

(a) $\leq \frac{1}{2\gamma_t}\|\theta_{t-1} - \theta^\star\|_2^2 + \frac{\gamma + b^2}{2} - \frac{1}{2\gamma_t}\mathbb{E}[\|\theta_t - \theta^\star\|_2^2 |\mathcal{F}_{t-1}] - \frac{\mu}{2}\|\theta_{t-1} - \theta^\star\|_2^2$

$$\mathbb{E}[F(\theta_{t-1} - F(\theta^\star))] \leq \frac{\mu(t+1)}{4}\mathbb{E}[\|\theta_{t-1} - \theta^\star\|^2] + \frac{b^2}{\mu(t+1)} - \frac{\mu(t+1)}{4}\|\theta_t - \theta^s tar\|^2 - \frac{\mu}{2}\mathbb{E}[\|\theta_{t-1-\theta^\star}\|^2]$$

$$\mathbb{E}[F(\theta_{t-1}) - F(\theta^\star)] \leq \frac{\mu(t-1)}{4}\mathbb{E}[\|\theta_{t-1} - \theta^\star\|^2] - \frac{\mu(t+1)}{4}\mathbb{E}[\|\theta_t - \theta^\star\|^2] + \frac{b^2}{\mu(t+1)}$$

$$\begin{aligned}\sum_{s=1}^{t} s\mathbb{E}[F(\theta_{s-1}) - F(\theta^\star)] &\leq \frac{\mu}{4}[\sum_{s=1}^{t} s(s-1)\mathbb{E}[\|\theta_{s-1} - \theta^\star\|_2^2] - s(s-1)\mathbb{E}[\|\theta_s - \theta^\star\|^2]] + \frac{b^2}{\mu}t \\ &\leq \frac{b^2}{\mu}\end{aligned}$$

By convexity of F :

$$\begin{aligned}\mathbb{E}[F(\frac{2}{t(t+1)}\sum_{s=1}^{t} s\theta_{s-1}) - F(\theta^\star)] &\leq \frac{2}{t(t+1)}\sum_{s=1}^{t} s\mathbb{E}[F(\theta_{s-1}) - F(\theta^\star)] \\ &\leq \frac{2}{t(t+1)}\sum_{s=1}^{t} s\mathbb{E}[F(\theta_{s-1}) - F(\theta^\star)] \\ &\leq \frac{2b^2}{\mu(t+1)}\end{aligned}$$

Note.
- L-smooth constant $\gamma = mathcalO(1/\sqrt{T})$, rate = $mathcalO(1/\sqrt{T})$
- L-smooth & *mu*-strongly convex, $\gamma_t \propto \frac{1}{t}$: rate = $mathcalO(1/t)$

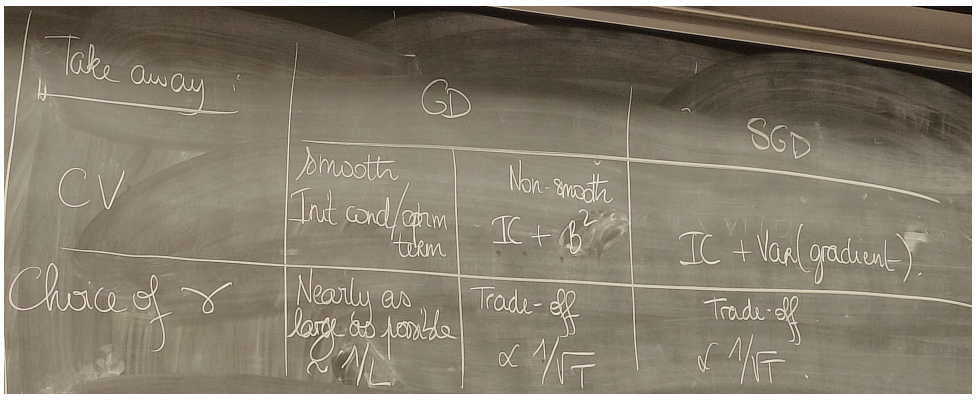


Figure 2.3: Take Away table

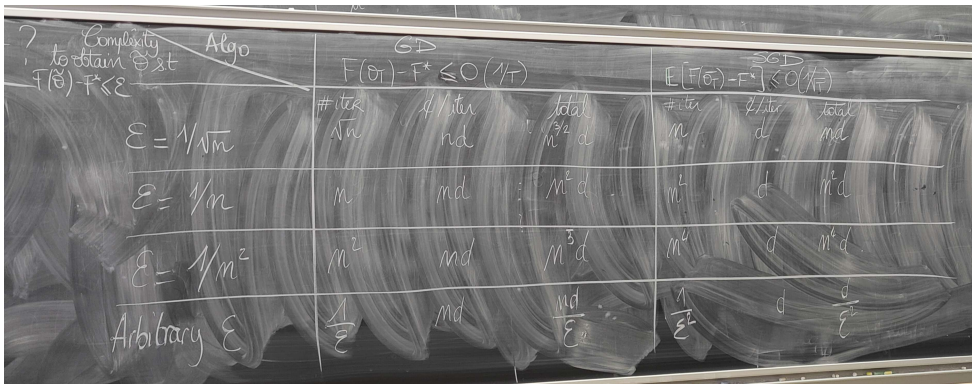


Figure 2.4: In terms of optimization, SGD vs GD : who is best?

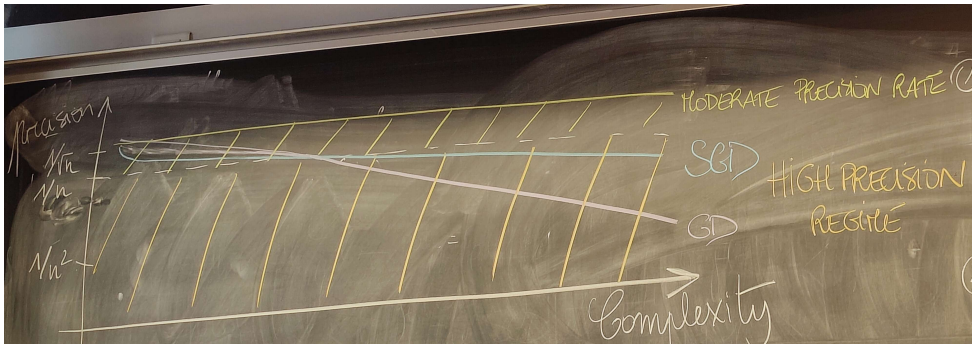


Figure 2.5: In terms of optimization, SGD vs GD : who is best? Graph

In terms of optimization, SGD vs GD : who is best? Constext : F smooth, ERM $F(\theta) = \frac{1}{n}\sum_{i=}^{n} F_i(\theta)$

1. GD tends to outperform SGD (in termes of complexity) for hig presion regimes
2. SGD outperforms GD for low to moderate precision regimes

So none is best. It depends to the precision.
The frontier between "moderate" & "high" precision has been fixed to $\frac{1}{n}$. In ML, one aims to optimize up to the **statistical** precision, ranging from $1/\sqrt{n}$ to $1/n$, thus moderate precision is enough!

CCL

- For optimization, no best method between GD and SGD
- For ML , moderate precision, better choose SGD

In terms of generalization?

- Running SGD for the expected risk minimization with n samples leads to a generatization error $\propto 1/\sqrt{n}$. This has to be compared with classical bounds of stat/ML SGD on the expected risk avoids estimation pb. One pass of SGD may be competitive without exact ERM.
  In term of running complexities, we get :

$$\mathcal{O}(tnd) = \mathcal{O}(n^2 d) \text{ vs } \mathcal{O}(nd)$$
$$(GD \text{ for } ERM) \text{ vs } (SGD)$$

Warning We are only comparing upperbounds!

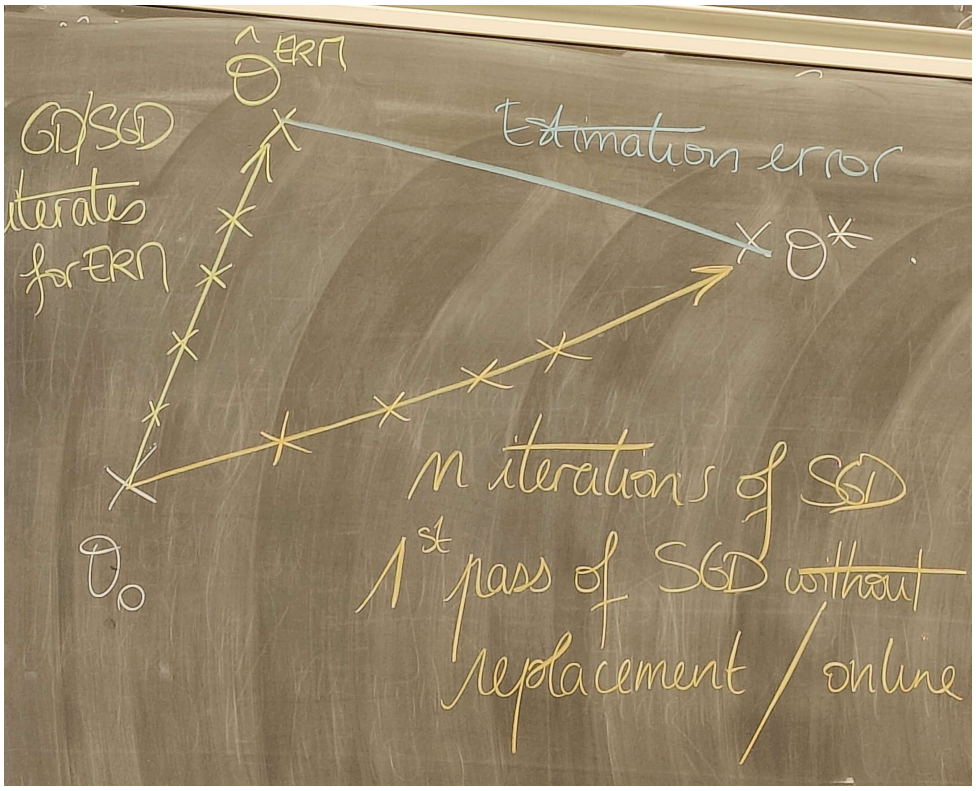


Figure 2.6: In terms of optimization, SGD vs GD : who is best? Graph