Optimisation Stochastique

Charles Vin

S1-2023

Chapter 1

Some basics of statistical learning

Nouveau cours du 15/11

$\mathcal{D}_n = (X_1, Y_1), \ldots, (X_n, Y_n)$ training samples iid copies of (X, Y). $X \in \mathcal{X}$ and $Y \in \mathcal{Y}$

- **goal** : fing a predictor surch that $f(X_i) \simeq Y_i$ on the training set.
  but above all $f(X_{\text{new}}) \simeq Y_{\text{new}}$ (test set)

- **Risk** : $\mathcal{R}^{(l)}(f) = \mathbb{E}[\ell(Y, f(X))]$
  loss ↓ ; ↑ on (X,Y)

- **Bayes predictor** : $f^* \in \arg\min \mathcal{R}(f)$
  $\mathcal{R}^* = \mathcal{R}(f^*) = \inf_f \mathcal{R}(f)$

- **Empirical risk** : $\hat{\mathcal{R}}_n(f) := \frac{1}{n} \sum_{i=1}^{n} \ell(Y_i, f(X_i))$
  $\hat{f}^{\text{ERM}} \in \arg\min_f \hat{\mathcal{R}}_n(f)$ (On some class of predictor)

Statistical learning theory focuses on controlling

$$\mathcal{R}_n(\hat{f}_n) - \mathcal{R}^* \text{ or } \hat{\mathcal{R}}_n(\hat{f}_n) - \mathcal{R}^*$$

for $\hat{f}_n$ a constructed predictor on the training set of size n.

Classical error decomposition :

$$\mathcal{R}(\hat{f}_n) - \mathcal{R}^* = \underbrace{\inf_{f \in \mathcal{F}} \mathcal{R}(f) - \mathcal{R}^*}_{\text{approximation error}} + \underbrace{\mathcal{R}(\hat{f}^{\text{ERM}}) - \inf_{f \in \mathcal{F}} \mathcal{R}(f)}_{\text{estimation error}} + \underbrace{\mathcal{R}(\hat{f}_n) - \mathcal{R}(\hat{f}^{\text{ERM}})}_{\substack{\text{error due to the use of an} \\ \text{optimisation algo for ERM}}} .$$

Control of the estimation error
"Overfitting" : we have

$$\underbrace{\hat{\mathcal{R}}_n(\hat{f}_n)}_{\text{train risk}} <<< \underbrace{\mathcal{R}_n(\hat{f}_n)}_{\text{test risk}}.$$

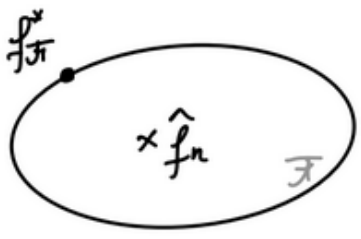


Figure 1.1: Overfitting for ERM.

Lemma 1

For the ERM predictor $\hat{f}_n = \hat{f}^{\mathrm{ERM}}$, one has

$$\mathcal{R}(\hat{f}^{\mathrm{ERM}}) - \mathcal{R}(f^*_{\mathcal{F}}) \leq 2 \sup_{f \in \mathcal{F}} \left| \hat{\mathcal{R}}_n - \mathcal{R} \right| (f).$$

with $f^*_{\mathcal{F}} \in \arg\min_{f \in \mathcal{F}} \mathcal{R}(f)$

Meaning the there is a link between estimation error and control of empirical process.

Preuve :

$$\mathcal{R}(\hat{f}^{\mathrm{ERM}}) - \mathcal{R}(f^*_{\mathcal{F}}) = \underbrace{\mathcal{R}(\hat{f}^{\mathrm{ERM}}) - \hat{\mathcal{R}}_n(\hat{f}^{\mathrm{ERM}})}_{\leq \sup_{f \in \mathcal{F}} (\hat{\mathcal{R}}_n - \mathcal{R})(f)} + \underbrace{\hat{\mathcal{R}}_n(\hat{f}^{\mathrm{ERM}}) - \hat{\mathcal{R}}_n(f^*_{\mathcal{F}})}_{\leq 0 \text{ by optimality of } \hat{f}^{\mathrm{ERM}}} + \underbrace{\hat{\mathcal{R}}_n(f^*_{\mathcal{F}}) - \mathcal{R}(f^*_{\mathcal{F}})}_{\leq \sup_{f \in \mathcal{F}} (\hat{\mathcal{R}}_n - \mathcal{R})(f)}$$

$$\leq 2 \sup_{f \in \mathcal{F}} \left| \hat{\mathcal{R}}_n - \mathcal{R} \right| (f)$$

□

Technics to control the supremum of empirical process.

- In expectation : $\mathbb{E}[\mathcal{R}(\hat{f}^{\mathrm{ERM}}) - \mathcal{R}(f^*_{\mathcal{F}})]$ small?
  tool : Pisier's lemma
- in proba : $\mathbb{P}(\mathcal{R}(\hat{f}^{\mathrm{ERM}}) - \mathcal{R}(f^*_{\mathcal{F}}) > \varepsilon) \leq \delta$
  tool : union bound + Hoeffding inequality

Theorem 2 (Hoeffding's inequality)

$X_1, \dots X_n \perp\!\!\!\perp$ r.v. and $X_i \in [a_i, b_i]$ a.s.
then

$$\mathbb{P}\left(\left| \frac{1}{n} \sum_{i=1}^{n} (X_i - \mathbb{E}[X_i]) \right| > \varepsilon \right) \leq 2e^{-\frac{2n^2\varepsilon^2}{\sum_i (b_i - a_i)^2}}$$

$$\text{when} \begin{cases} a_i = a \\ b_i = b \end{cases} \to \leq 2e^{-\frac{2n\varepsilon^2}{(b-a)^2}}$$

1. When $nto + \infty$ (ε fixed) the bound goes to 0.
2. When $b - a \to +\infty$ the bound goes to 1.
3. The R.H.S (right hand side) should depend on $\frac{\varepsilon}{b-a}$ (One could rescale X_i by $\frac{1}{b-a}$)
4. By the CLT, the R.H.S should go to a constant if $n \to \infty$ and $\varepsilon = \frac{\tau}{\sqrt{n}}$

$$\mathbb{P}\left(\frac{1}{n} \sum_{i=1}^{n} (X_i - \mathbb{E}[X_i]) > \frac{\tau}{\sqrt{n}} \right) = \mathbb{P}\left(\frac{1}{\sqrt{n}} \sum_{i=1}^{n} (X_i - \mathbb{E}[X_i]) > \tau \right)$$

$$\xrightarrow[n \to \infty]{} \mathbb{P}(\mathcal{N}(0, .) > \tau) = \text{ constant } (\tau)$$

Proposition 3

In the case of binary classification ($y = \{+1, -1\}$), concider a finite class $\mathcal{F}$ of predictors, i.e. of cardinality $|\mathcal{F}| < +\infty$.

$$\mathbb{P}\left(\sup_{f \in \mathcal{F}} \left| \hat{\mathcal{R}}_n - \mathcal{R} \right| (f) > \varepsilon \right) \leq 2 \left| \mathcal{F} \right| e^{-2n\varepsilon^2}.$$

Preuve :

$$\mathbb{P}(\sup_{f in \mathcal{F}} \left|\hat{\mathcal{R}}_n - \mathcal{R}\right|(f) > \varepsilon) = \mathbb{P}(\bigcup_{f \in \mathcal{F}} \{\left|\hat{\mathcal{R}}_n - \mathcal{R}\right|(f) > \varepsilon\})$$
$$\leq \sum_{f in \mathcal{F}} \mathbb{P}(\left|\hat{\mathcal{R}}_n - \mathcal{R}\right|(f) > \varepsilon) \qquad \text{[Union bound]}$$

But $\left|\hat{\mathcal{R}}_n - \mathcal{R}\right|(f) = \frac{1}{n}\left|\sum_{i=1}^n \mathbb{1}_{f(X_i)\neq Y_i} - \mathbb{E}[\mathbb{1}_{f(X_i)\neq Y_i}]\right| = \frac{1}{n}\left|\sum_{i=1}^n Z_i - \mathbb{E}[Z_i]\right|$
with $Z_i \in \{0,1\} \in [0,1]$
Therefore $\forall f \in \mathcal{F}, \mathbb{P}(\left|\hat{\mathcal{R}}_n - \mathcal{R}\right|(f) > \varepsilon) \leq 2e^{-2n\varepsilon^2}$.
Finally,

$$\mathbb{P}(\sup_{f \in \mathcal{F}} \left|\hat{\mathcal{R}}_n - \mathcal{R}\right|(f) > \varepsilon) \leq 2\left|\mathcal{F}\right| e^{-2n\varepsilon^2}$$

□

TO BE COMPLETED

Nouveau cours du 16/11

CCL du cours de la dernière fois

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

1.1 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

Lemma 4

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

Proof: 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 **??**

   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

Lemma 5 (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)

Proof:

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

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 : $\forall \hat{h}$

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

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 2

Basics of deterministic optimisation

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

2.1 First order method

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

Other definition:

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

Lemma 6 (Descent lemma)

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

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

Proof:

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

□

Consequence of this quadratics upper bound

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

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

2.1.2 Gradient algorithms

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

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!

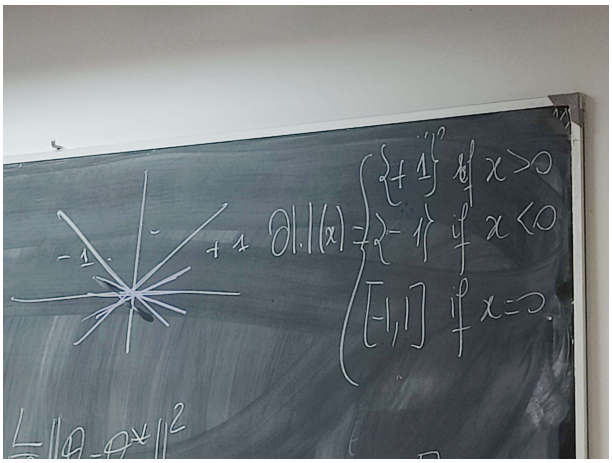


Figure 2.1: subgradients

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 $\frac{\partial (F \circ \theta)}{\partial t}(t) = \langle \nabla F(\theta(t)), \dot{\theta}(t) \rangle = - \left\| \nabla F(\theta(t)) \right\|^2$

Theorem 7

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

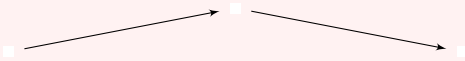
$$F(\theta_t) - F(\theta^\star) \leq \frac{\|\theta_0 - \theta^\star\|}{2\gamma(1 - \frac{\gamma L}{2})T}.$$

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.

Proof:

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

Function convex + 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

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

□

Nouveau cours du 23/11

RAPPEL : On regarde

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

Theorem 8

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 $\mathcal{C}^2 \Leftrightarrow \lambda_{MAX}(H_F(\theta)) \leq L \forall \theta$

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

Theorem 9

If F is L-Smooth, diff and μ - strongly convex, 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 case the rate is said to be linear.
2. CV rate on the iterates (!!) and not only on the objective rate

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

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

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

So :

- $\kappa = \dfrac{\mu}{L}$ is called the condition number of F.

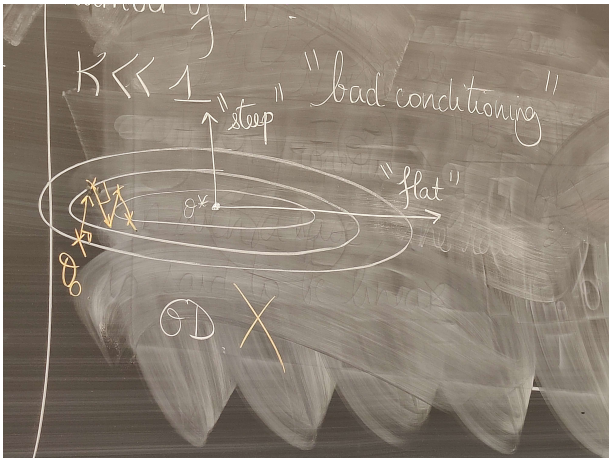


Figure 2.2: With $\kappa \ll 1$ "bad conditioning"

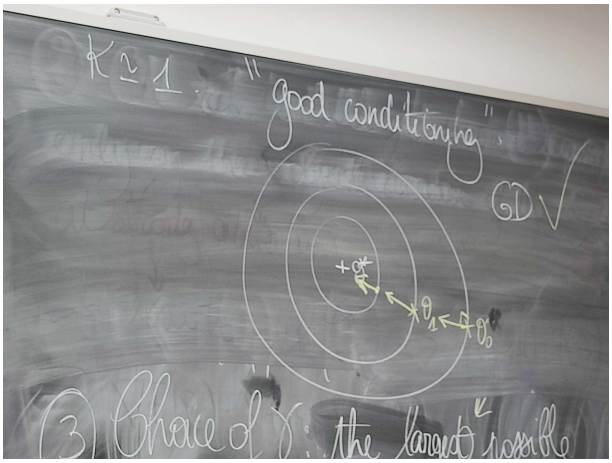


Figure 2.3: With $\kappa \approx 1$ "good conditioning"

- $\kappa << 1$ "Bad conditioning"
- $\kappa \simeq 1$ "good conditioning"

Proof:

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

By μ-strong convexity, we got

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

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

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

Combining the 2 previous inequalities,

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

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

Theorem 10 (GD for non-smooth fonctions)

Hypothese : F convex, has subgradients, β -Lipschitz

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

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 **??**
- $\frac{\gamma\beta^2}{2}$ in green in Figure **??**

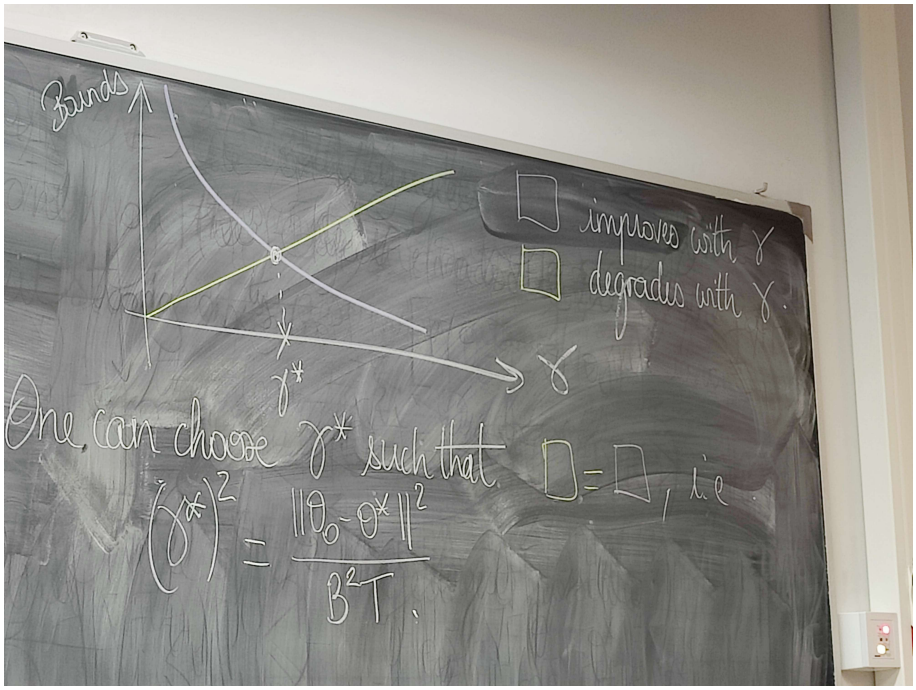


Figure 2.4:

One can choose $\gamma^\star$ such that "purple" = "green" (Figre **??**), 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, $F(\frac{1}{T}\sum_{t=1}^T \theta_t) - F^\star \leq \frac{1}{T}\sum_{t=1}^T (F(\theta_t) - F^\star) \lesssim \mathcal{O}(\frac{\log_T}{T})$.

Proof:

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

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

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

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 convex and the only subgradient at the solution is the zero vector :

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

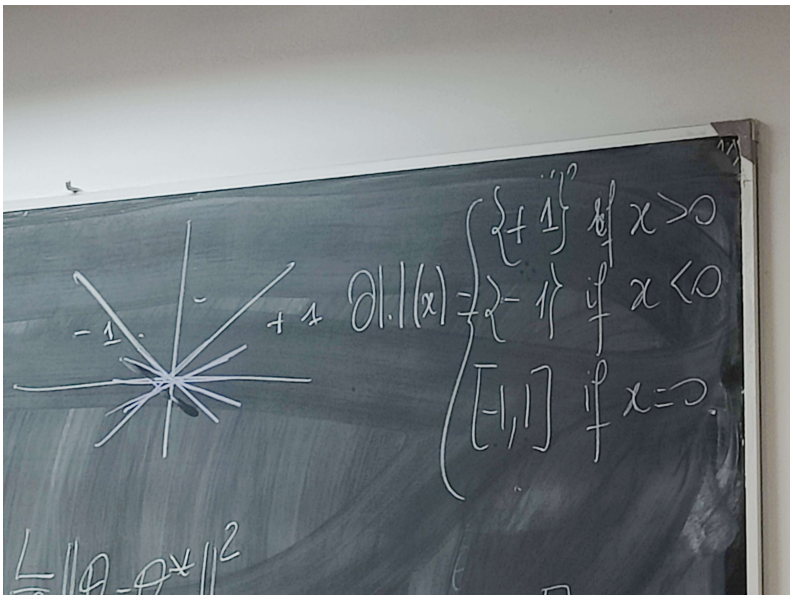


Figure 2.5: sub gradiens

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 convex 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 :

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

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

2.2 Inertial methods

2.2.1 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 11

A gradient-like algo is an algo such taht

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

Theorem 12 (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

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

Theorem 13 (Nesterov 2003)

$\forall \theta_0 \in \mathbb{R}^d, \mu > 0, L > 0, \exists F$ *mu*-strongly convex 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. $\left\|\theta_t - \theta^\star\right\| \geq (\frac{1-\sqrt{\kappa}}{1+\sqrt{\kappa}})^t \left\|\right\| \theta_0 - \theta^\star \quad$ with $\kappa = \frac{\mu}{L}$

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

2.2.2 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.

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

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

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

Discretization

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

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]

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

NESTEROV ALGO [83]

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

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

Theorem 14 (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 convex fonction and need to know μ (and L)

NESTEROV

Theorem 15

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

Theorem 16 (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
CCL : vanishing friction helps on the worst fcts

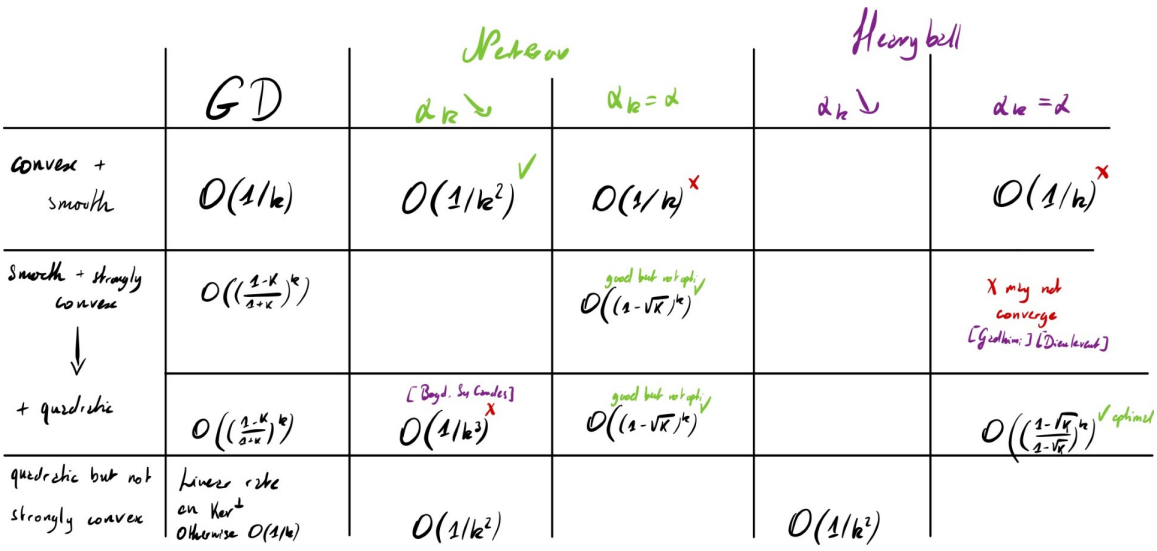


Figure 2.6: Tableau de la vitesse de convergence des algos

Chapter 3

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 3.1: 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.

3.1 SDG in machine learning

There are 2 ways to use SGB in supervised learning

3.1.1 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

3.1.2 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.

3.1.3 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,

Theorem 17

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 **??** 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 $\forall t \geq 1$

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

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

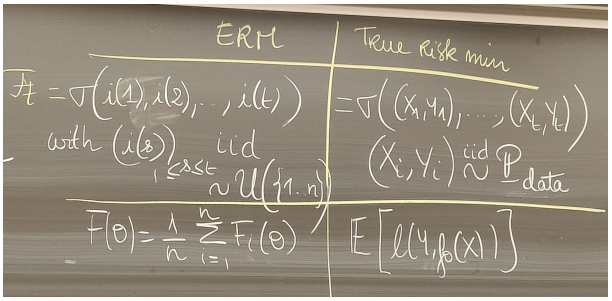


Figure 3.2: More explaination of $\mathcal{F}_t$

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

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

□

Theorem 18

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

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

Proof: 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

Theorem 19

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 = $mathcalO(1/t)$

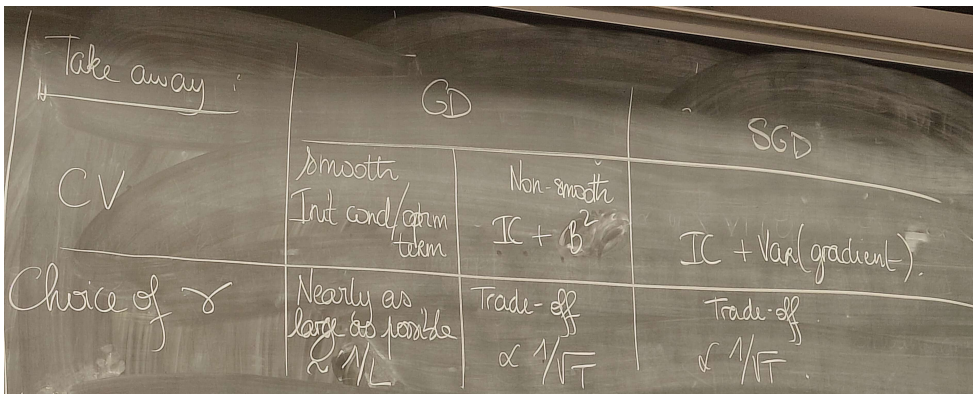


Figure 3.3: Take Away table

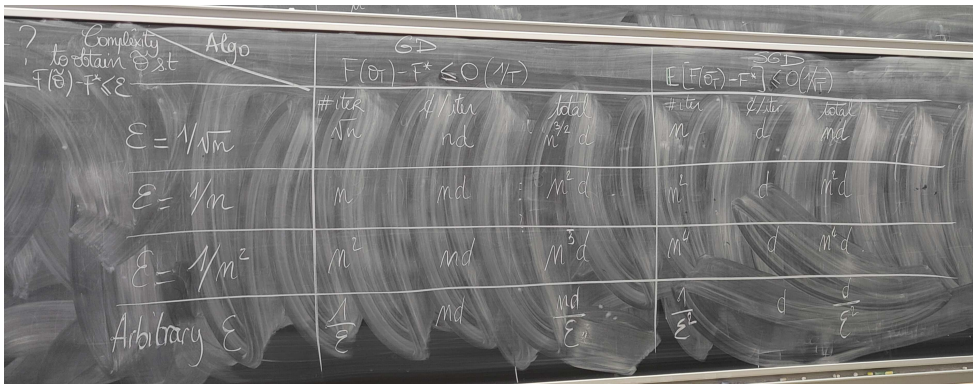


Figure 3.4: In terms of optimization, SGD vs GD : who is best?

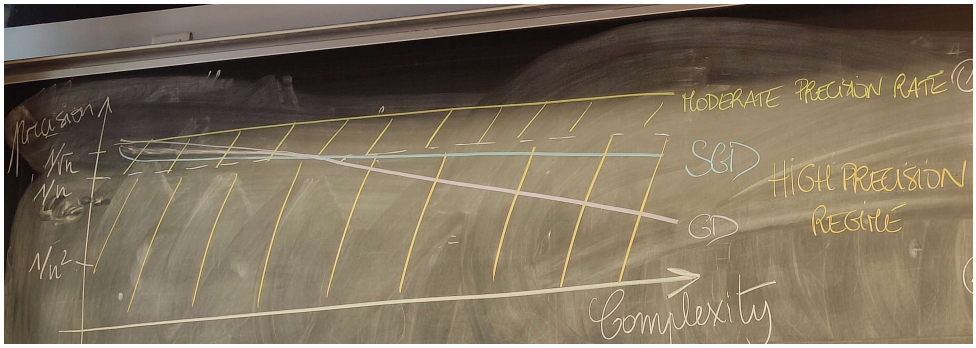


Figure 3.5: In terms of optimization, SGD vs GD : who is best? Graph

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!

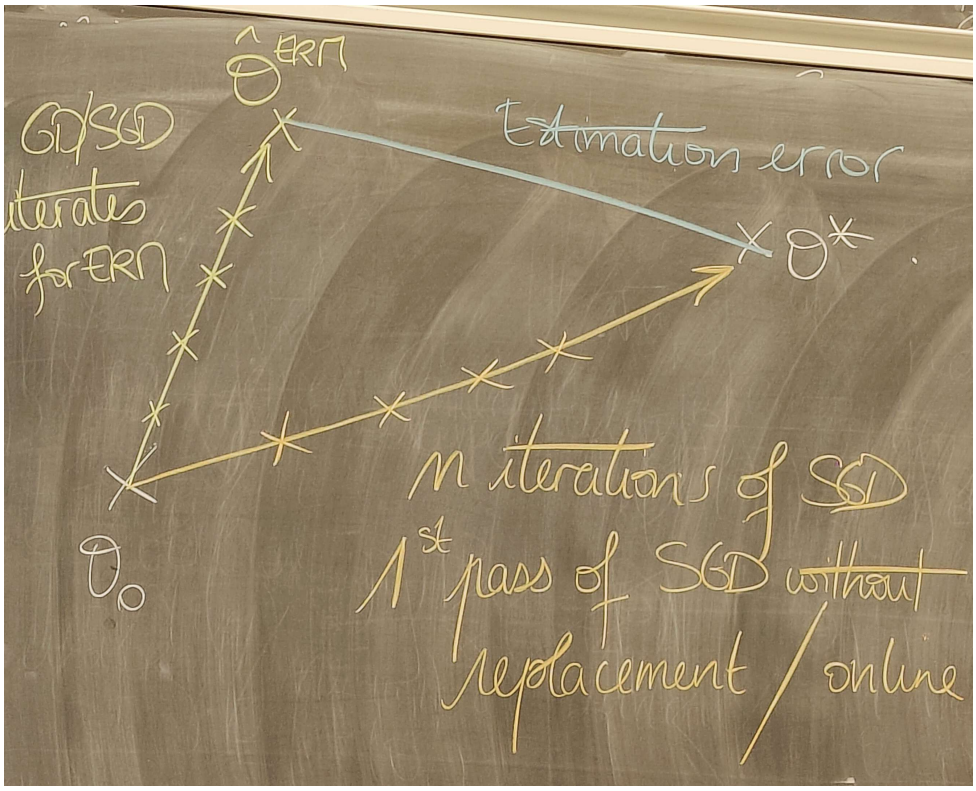


Figure 3.6: In terms of optimization, SGD vs GD : who is best? Graph

Nouveau cours du 30/11

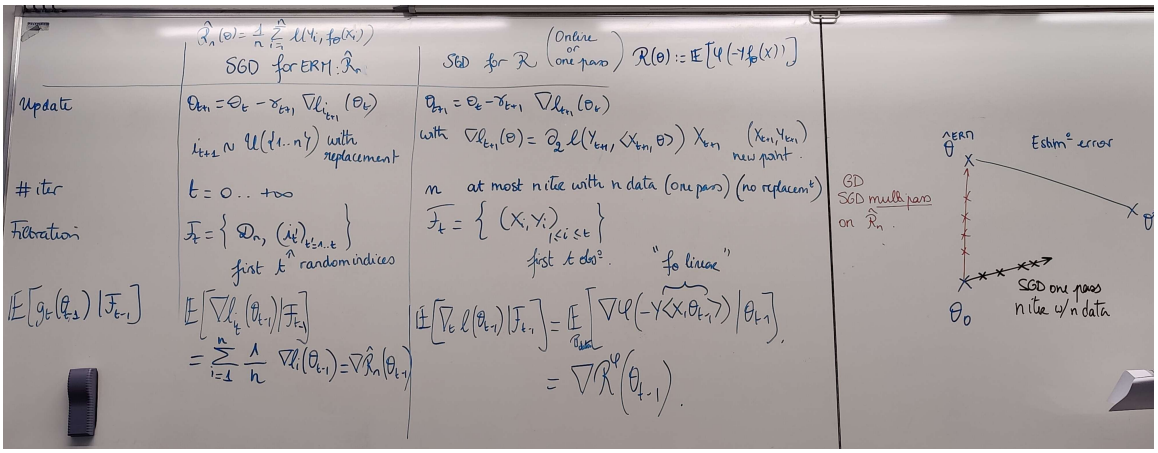


Figure 3.7: Comparaison between SGD for ERM $\hat{\mathcal{R}}_n$ and SGD for $\mathcal{R}$

Averaging : In the case of strongly cvx fonction & smooth

- SGD with no averaging $\gamma_t = Ct^{-\alpha}, \alpha = 1$
- SGD + averaging $\gamma_t = Ct^{\alpha}$, $\frac{1}{2} \leq \alpha \leq 1$ rate : $\mathcal{O}(t^{-1}\mu^{-1})$

=> Averaging help in case of "bad choice" of step size.

Take-home message : use $\alpha = \frac{1}{2}$ + averaging.

Chapter 4

Stability of learning Algorithm & generalizations

4.1 Motivation

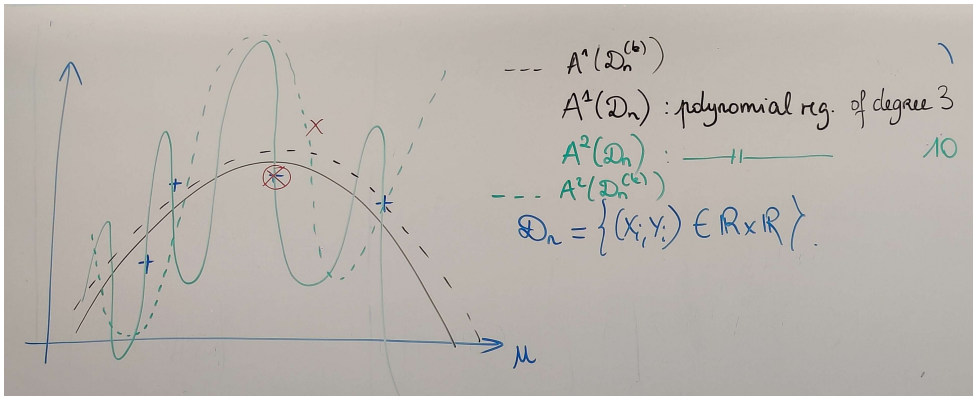


Figure 4.1: Chapter 4: Motivation. The black algorythm seem better (better generalization, less overfitting)

Generalization : for a test point (X, Y) of same distrib and $\perp\!\!\!\perp$ of $\mathcal{D}_n$, evaluation of $\mathbb{E}[l(Y, A(\mathcal{D}_n)(X))]$

Another point of view of generalization : compare $A(\mathcal{D}_n)$ vs $A(\mathcal{D}_n^{(k)})$ where $\mathcal{D}_n^{(k)}$ is the training set in which we change the k-th point.

$$\mathcal{D}_n^{(k)\leftarrow(X_k', y_k')} = (X_1, Y_1)...(X_{k-1}, Y_{k-1}), (x_k', y_k'), (X_{k+1}, Y_{k+1}), ...(X_n, Y_n).$$

In the picture

- $A^1(\mathcal{D}_n)$ is "close" to $A^1(\mathcal{D}_n^{(k)})$
- $A^2(\mathcal{D}_n)$ is very different from $A^2(\mathcal{D}_n^{(k)}) \rightarrow A^2$ is unstable.

Claim : Stability implies generalization

Définition 20

A learning algo $(A_n)_n$ is a sequence of mapping

$$A_n : (\mathcal{X}, \mathcal{Y})^{\times n} \rightarrow \mathcal{C}$$
$$\mathcal{D}_n \mapsto A_n(\mathcal{D}_n)(\text{ possibly random})$$

from any number of points into the set $\mathcal{C}$ of classifier.

Exemple 21 (1)

Empirical risk minimizer over Θ is a learning algo

$$A^{ERM}(\mathcal{D}_n) \equiv \arg\min_{\theta\in\Theta} \hat{\mathcal{R}}_{n,\mathcal{D}_n}(\theta).$$

Exemple 22 (This is a **randomized algo**)

SGD with T steps and $(\gamma_t)_{t=1\dots T}$ is a learning algo

$$A^{SGD} \equiv \begin{cases} \theta_0 & = 0 \\ \theta_{t+1} & = \theta_t - \gamma_{t+1} g_{t+1}(\theta_t) \text{ for } t = 0, \dots, T-1 \end{cases}.$$

where $g_t(\theta_{t-1})$ is the gradient of

$$\theta \mapsto l(Y_{i_t}, \langle \theta, X_{i_t} \rangle), i_t \sim \mathcal{U}(\{1, \dots, n\}).$$

Définition 23 (Uniform stability (deterministic algo))

Concider a deterministic algo A
A is uniformly stable for $\beta : \mathbb{N} \to \mathbb{R}_+$ if for **any** $\mathcal{D}_n$, for **any** k, for **any** $(X'_k, Y'_k) \in \mathcal{X} \times \mathcal{Y}$, for any test point $(X_{test}, T_{test}) \in \mathcal{X}, \mathcal{Y}$

$$\left| l(Y_{test}, A(\mathcal{D}_n)(X_{test})) - l(Y_{test}, A(\mathcal{D}_n^{(k)\leftarrow(X'_k,Y'_k)})(X_{test})) \right| \leq \beta(n).$$

Définition 24 (Uniform stability (randomized algo))

Concider a randomized algo A
A is uniformly stable for $\beta_{av} : \mathbb{N} \to \mathbb{R}_+$ if for **any** $\mathcal{D}_n$, for **any** k, for **any** $(X'_k, Y'_k) \in \mathcal{X} \times \mathcal{Y}$, for any test point $(X_{test}, T_{test}) \in \mathcal{X}, \mathcal{Y}$

$$\mathbb{E}[\left| l(Y_{test}, A(\mathcal{D}_n)(X_{test})) - l(Y_{test}, A(\mathcal{D}_n^{(k)\leftarrow(X'_k,Y'_k)})(X_{test})) \right|] \leq \beta_{av}(n).$$

Thi only random thing is the algo

Note. We could extend thses definitions to EXPECTED stability (kind of a less stronger version) : for $(X_i, Y_i)_{i=1}^n, (X'_k, Y'_k), (X_{test}, Y_{test}) \sim \mathbb{P}_{data}^{\otimes(n+2)}$

Ref :

- Stability introduced in Bousquet & Elisseeff (2002)
- Re-introduced by HArdt, Recht, Singer "Train faster, generalize bettter" (2016)
- Widely used since then

Note. Related to algorithm sensibility, privacy-preserving algo with DP (differential privacy)

$$(\epsilon, \delta)\text{-DP} : \forall S, \mathbb{P}(A(\mathcal{D}_n) \in S) \leq (1+\epsilon)\mathbb{P}(A(\mathcal{D}_n^{(k)}) \in S) + \delta.$$

$\rightsquigarrow$ Aurélien Bellet, french expert in Monpelier.

4.2 Stability implies generalization

Warning: A stupid algo A such that $\forall \mathcal{D}_n, A(\mathcal{D}_\setminus) = \theta_0 \perp\!\!\!\perp \mathcal{D}_n$ is β-stable for $\beta = 0$.

Intuition: the generalization ability should be controlled by : 1. Stability term ; 2. Data-fitting term.

Lemma 25

Call $\theta^\star = \arg\min_{\theta\in\Theta} \mathcal{R}(\theta)$ and $\theta^{ERM} = \arg\min_{\theta\in\Theta} \hat{\mathcal{R}}_{n,\mathcal{D}_n}(\theta)$

$$\mathbb{E}[\mathcal{R}(A(\mathcal{D}_n)) - \mathcal{R}(\theta^\star)] \leq \mathbb{E}[\underbrace{\mathcal{R}(A(\mathcal{D}_n)) - \hat{\mathcal{R}}_{n,\mathcal{D}_n}(A(\mathcal{D}_n))}_{\text{Difference between generalisation / empirical errors} \approx \text{Stability}} + \underbrace{\mathbb{E}[\hat{\mathcal{R}}_{n,\mathcal{D}_n}(A(\mathcal{D}_n)) - \hat{\mathcal{R}}_{n,\mathcal{D}_n}(\hat{\theta}^{ERN})]}_{\text{optim error} \rightsquigarrow \text{ data fitting term}}]$$

Note. For any predictor $\theta_0 \perp\!\!\!\perp \mathcal{D}_n$

$$\mathcal{R}(\theta_0) = \mathbb{E}[\hat{\mathcal{R}}_{n,\mathcal{D}_n}(\theta_0)].$$

Proof:

$$\begin{aligned}\mathcal{R}(A(\mathcal{D}_n)) - \mathcal{R}(\theta^\star) &= \mathcal{R}(A(\mathcal{D}_n)) - \hat{\mathcal{R}}_{n,\mathcal{D}_n}(A(\mathcal{D}_n)) \\ &+ \hat{\mathcal{R}}_{n,\mathcal{D}_n}(A(\mathcal{D}_n)) - \hat{\mathcal{R}}_{n,\mathcal{D}_n}(\hat{\theta}^{ERM}) \\ &+ \hat{\mathcal{R}}_{n,\mathcal{D}_n}(\hat{\theta}^{ERM}) - \hat{\mathcal{R}}_{n,\mathcal{D}_n}(\hat{\theta}^\star) \quad \leq 0 \text{ since } \hat{\theta}^{ERM} \in \arg\min_\Theta \hat{\mathcal{R}}_{n,\mathcal{D}_n} \\ &+ \hat{\mathcal{R}}_{n,\mathcal{D}_n}(\hat{\theta}^\star) - \mathcal{R}(\theta^\star)\end{aligned}$$

And

$$\begin{aligned}\mathbb{E}[\mathcal{R}(A(\mathcal{D}_n)) - \mathcal{R}(\theta^\star)] &= \mathbb{E}[\mathcal{R}(A(\mathcal{D}_n)) - \hat{\mathcal{R}}_{n,\mathcal{D}_n}(A(\mathcal{D}_n))] \\ &+ \mathbb{E}[\hat{\mathcal{R}}_{n,\mathcal{D}_n}(A(\mathcal{D}_n)) - \hat{\mathcal{R}}_{n,\mathcal{D}_n}(\hat{\theta}^{ERM})] \\ &+ \mathbb{E}[\hat{\mathcal{R}}_{n,\mathcal{D}_n}(\hat{\theta}^{ERM}) - \hat{\mathcal{R}}_{n,\mathcal{D}_n}(\hat{\theta}^\star)] \\ &+ \mathbb{E}[\hat{\mathcal{R}}_{n,\mathcal{D}_n}(\hat{\theta}^\star) - \mathcal{R}(\theta^\star)] \quad = 0 \text{ since } \theta^\star \perp\!\!\!\perp \mathcal{D}_n\end{aligned}$$

□

Note. For example 1, $A(\mathcal{D}_n) \equiv \hat{\theta}^{ERN}$, Lemma 22 gives

$$\mathbb{E}[\mathcal{R}(A(\mathcal{D}_n))] - \mathcal{R}(\theta^\star) \leq \mathbb{E}[\mathcal{R}(A(\mathcal{D}_n)) - \hat{\mathcal{R}}_{n,\mathcal{D}_n}(A(\mathcal{D}_n))].$$

the optim error vanishes

Theorem 26 (Stability → Generalization)

If A is a learning algo which is β-uniformly stable then

$$\mathbb{E}[\mathcal{R}(A(\mathcal{D}_n)) - \hat{\mathcal{R}}_{n,\mathcal{D}_n}] \leq \beta(n).$$

Proof: Reminder

- $\mathcal{R}(A(\mathcal{D}_n)) = \mathbb{E}_{(X,Y)\sim p}[l(Y, A(\mathcal{D}_n))|\mathcal{D}_n]$
- $\mathbb{E}_{\mathcal{D}_n}[\mathcal{R}(A(\mathcal{D}_n))]$
- $\hat{\mathcal{R}}_{n,\mathcal{D}_n}(A(\mathcal{D}_n)) = \frac{1}{n}\sum_{i=1}^n l(Y_i, A(\mathcal{D}_n)(X_i))$

- $\mathbb{E}_{\mathcal{D}_n}[\hat{\mathcal{R}}_{n,\mathcal{D}_n}(A(\mathcal{D}_n))]$

Observe that for any $k \in \{1,\dots,n\}$

$$\begin{aligned}\mathbb{E}[\mathcal{R}(A(\mathcal{D}_n))] &= \mathbb{E}_{\mathcal{D}_n\sim p^{\otimes n},,\perp\!\!\!\perp(X,Y)\sim p}[l(Y, A(\mathcal{D}_n)(X))] \\ &= \mathbb{E}_{\mathcal{D}_n\sim p^{\otimes n},,\perp\!\!\!\perp(X,Y)\sim p}[l(Y_k, A(\mathcal{D}_n^{(k)})(X_k))]\end{aligned}$$

Since $((X_1,Y_1),\dots,(X_k,Y_k),\dots,(X_n,Y_n),(X,Y)) =^{\text{dist}} ((X_1,Y_1),\dots,(X,Y)\dots,(X_n,Y_n),(X_k,Y_k))$
Therefore

$$\begin{aligned}\mathbb{E}[\mathcal{R}(A(\mathcal{D}_n)) - \hat{\mathcal{R}}_{n,\mathcal{D}_n}(A(\mathcal{D}_n))] &= \mathbb{E}[\frac{1}{n}\sum_{k=1}^{n}(\mathcal{R}(A(\mathcal{D}_n)) - l(Y_k, A(\mathcal{D}_n)(X_k)))] \\ &= \frac{1}{n}\sum_{k=1}^{n}\mathbb{E}[\underbrace{l(Y_k, A(\mathcal{D}_n^{(k)(X,Y)})(X_k)) - l(Y_k, A(\mathcal{D}_n)(X_k))}_{\leq\beta(n) \text{ by stability assumption}}] \\ &\leq \beta(n)\end{aligned}$$

□

CCL :

- (BEFORE) $gen \leq \underbrace{\text{state rate}}_{\text{Worst } \hat{\mathcal{R}}_n - \mathcal{R} \text{ on } \theta\in\Theta(\text{ Uniform bounds})} + \text{optim rate}$
- (NOW) $gen \leq \underbrace{stab}_{\text{This term is algo-dependent}} + \text{optimization.}$

4.3 Computing the stability β for ERM with a strongly cvx risk

Hypothesis : $\theta \mapsto \hat{\mathcal{R}}_{n,\mathcal{D}_n}(\theta)$ is μ-strongly cvx
Remark: Is the relaxed risk strongly cvx in ML?
When $F \in \mathcal{C}^2$, $\forall\theta, \nabla^2 F(\theta) \succcurlyeq \mu$ Id. i.e $\lambda_{\text{MIN}}(\nabla^2 F(\theta)) \geq \mu$
When $\hat{\mathcal{R}}_n(\theta) = \frac{1}{n}\sum_{i=1}^{n}\varphi(-X_i^T\theta Y_i), Y_i \in \{\pm 1\}, Y_i^2 = 1$ then

$$\nabla\hat{\mathcal{R}}_n(\theta) = \frac{1}{n}\sum_{i=1}^{n}(-\varphi'(-X_i^T\theta Y_i)Y_i X_i)$$

$$\nabla^2\hat{\mathcal{R}}_n(\theta) = \frac{1}{n}\sum_{i=1}^{n}\varphi''(-X_i^T\theta Y_i)\underbrace{X_i X_i^T}_{\in\mathbb{R}^{d\times d}}$$

If φ is μ-strongly cvx from $\mathbb{R}$ to $\mathbb{R}$, the $\forall u, \varphi''(u) \geq \mu$ and $\nabla^2\hat{\mathcal{R}}_n(\theta) \succcurlyeq \mu\frac{1}{n}\sum_{i=1}^{n}X_i X_i^T$ i.e. $\lambda_{\text{MIN}}(\nabla^2\hat{\mathcal{R}}_n(\theta)) \geq \mu\lambda_{\text{MIN}}(\frac{1}{n}\sum_{i=1}^{n}X_i X_i^T)$
And $\frac{1}{n}\sum_{i=1}^{n}X_i X_i^T \to_{n\to+\infty} \text{Cov}(X)$ by the LLN if $((X_i)_i$ centered.

The strong convexity of $\hat{\mathcal{R}}_n$ is linked to the lowest eigenvalue of $Cov(X)$. It can be small (!!) and if $d > n$ then $\lambda_{\text{MIN}}(\frac{1}{n}\sum_{i=1}^{n}X_i X_i^T) = 0$ almost surely.

Finally strong convexity is not always true or μ can be small. Note that regularization can bring strong convexity

$$\min_\theta \hat{\mathcal{R}}(\theta) + \frac{\mu}{2}\|\theta\|_2^2 .$$

The ridge regularized risk is μ strongly cvx.

Theorem 27

Let $\hat{\theta}^{ORN} = \arg\min_{\theta\in\mathbb{R}^d} \hat{\mathcal{R}}_{n,\mathcal{D}_n}(\theta)$ with $\hat{\mathcal{R}}_{n,\mathcal{D}_n}$ μ-strongly convex.
Assume that the loss function l is β-Lipschitz w.r.t. it's second argument i.e.

$$\forall y, z, z', |l(y,z) - l(y,z')| \leq \beta |z - z'| .$$

Assume that $\|X\|_2 \leq \mathcal{R}$ a.s.
Then, the ERM algo $A(\mathcal{D}_n) \equiv \hat{\theta}^{ERM}$ is $\frac{4\beta^2R^2}{\mu n}$ uniformly stable.
$\rightarrow$ **Good:** this is a fast rate !! (for ERM) [see Srebro, Sridahran, Shalev-shwartz]

Proof: 1. By strong convexity, $\hat{\mathcal{R}}_{n,\mathcal{D}_n}(\theta) - \hat{\mathcal{R}}_{n,\mathcal{D}_n}(\overbrace{A(\mathcal{D}_n)}^{\hat{\theta}^{ERM}}) \geq \frac{\mu}{2}\|\theta - A(\mathcal{D}_n)\|_2^2$
Let $k \in \{1,\dots,n\}$, apply this inequation to $\theta = A(\mathcal{D}_n^{(k)})$ then $\hat{\mathcal{R}}_{n,\mathcal{D}_n}(A(\mathcal{D}_n)) - \hat{\mathcal{R}}_{n,\mathcal{D}_n}(A(\mathcal{D}_n)) \geq \frac{\mu}{2}\left\|A(\mathcal{D}_n^{(k)}) - A(\mathcal{D}_n)\right\|_2^2$

2.

$$\begin{aligned}
\hat{\mathcal{R}}_{n,\mathcal{D}_n}(A(\mathcal{D}_n^{(k)})) - \hat{\mathcal{R}}_{n,\mathcal{D}_n}(A(\mathcal{D}_n)) = \frac{1}{n}\sum_{i=1,i\neq k}^{n} l(Y_i, A(\mathcal{D}_n)(X_i)) - l(Y_i, A(\mathcal{D}_n)(X_i)) \quad \text{(L1)}\\
+ \frac{1}{n}[l(Y_k', A(\mathcal{D}_n^{(k)})(X_k')) - l(Y_k', A(\mathcal{D}_n)(X_k'))] \quad \text{(L2)}\\
- \frac{1}{n}[l(Y_k', A(\mathcal{D}_n^{(k)})(X_k')) - l(Y_k', A(\mathcal{D}_n)(X_k'))] \quad \text{(L3)}\\
+ \frac{1}{n}[l(Y_k, A(\mathcal{D}_n^{(k)})(X_k)) - l(Y_k, A(\mathcal{D}_n)(X_k))] \quad \text{(L4)}
\end{aligned}$$

- L1 + L2 $\rightarrow \hat{\mathcal{R}}_{n,\mathcal{D}_n^{(k)}}(A(\mathcal{D}_n^{(k)})) - \hat{\mathcal{R}}_{n,\mathcal{D}_n^{(k)}}(A(\mathcal{D}_n^{)}) \leq 0$ a.s
- L3 & L4 $\leq \frac{1}{n}.B.\left|A(\mathcal{D}_n^{(k)})(X_k) - A(\mathcal{D}_n)(X_k)\right| \leq \frac{B}{n}\left|A(\mathcal{D}_n^{(k)})(X_k) - A(\mathcal{D}_n)\right|_2 \underbrace{|Xk|}_{\leq R} 2$ (with $A(\mathcal{D}_n)(X_k) = A(\mathcal{D}_n)^T X_k$
- 1 & 2 $\rightarrow$

$$\frac{\mu}{2}\left\|A(\mathcal{D}_n^{(k)}) - A(\mathcal{D}_n)\right\|_2^2 \leq_{a.s.} \underbrace{2}_{\text{from L3+L4}} \frac{B.R}{n}\left\|A(\mathcal{D}_n^{(k)}) - A(\mathcal{D}_n)\right\|_2 \underbrace{\|X_k\|_2}_{\leq R}$$

$$\left\|A(\mathcal{D}_n^{(k)}) - A(\mathcal{D}_n)\right\|_2 \leq \frac{4BR}{\mu n}$$

$$\Rightarrow \forall (x,y), \left|l(y, A(\mathcal{D}_n^{(k)})) - l(y, A(\mathcal{D}_n)(x))\right| \leq \frac{4B^2RR^2}{\mu n}$$

Finaly, A is β-univormly stable. □

Note. We have provent a stronger condition than stability, i.e. $A(\mathcal{D}_n)$ is close to $A(\mathcal{D}_n^{(k)})$ in Euclidien norm

4.4 Stability of SGD ()

Algo

$$\begin{aligned}
&\theta_0, (\gamma_t)_{t=1}^T\\
&\forall 0 \leq t \leq T-1, \quad \theta_{t+1} = \theta_t - \gamma_t g_{t+1}(\theta_t)\\
&\text{with } g_{t+1}(\theta) = \nabla_\theta l(Y_{i_{t+1}}, \theta^T X_{i_{t+1}}) (\text{ linear predictors })\\
&i_{t+1} \sim \mathcal{U}(\{1,\dots,n\})
\end{aligned}$$

Question studying the stability of SGD requires to understand $A(\mathcal{D}_n)$ vs $A(\mathcal{D}_n)^{(k)}$

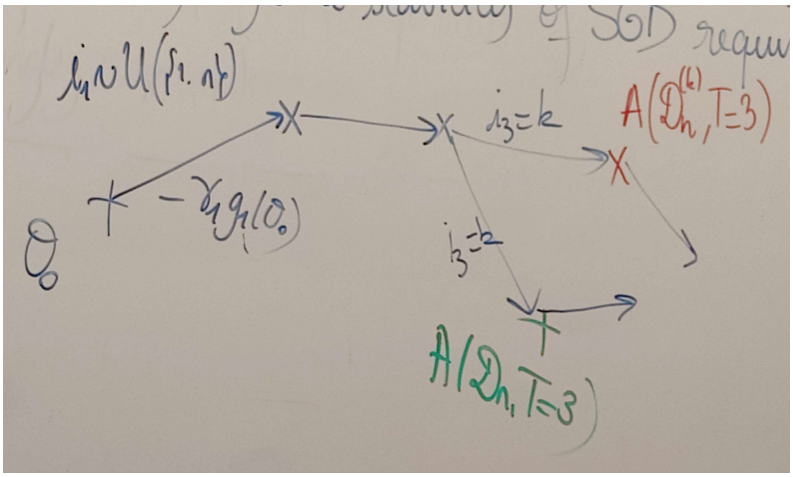


Figure 4.2: Difference between $A(\mathcal{D}_n)$ vs $A(\mathcal{D}_n)^{(k)}$

goal Control $A(\mathcal{D}_n, T_{\text{steps}}) - A(\mathcal{D}_n^{(k)}, T_{\text{steps}})$ where

- for most of the steps with $\mathbb{P} = \frac{n-1}{n}$,I use the same gradients (g_{i_t}), $i_t \neq k$ (at different points)
- for some steps with $\mathbb{P} = \frac{1}{n}$, I use g_k for $A(\mathcal{D}_n)$ and g'_k for $A(\mathcal{D}_n^{(k)})$

Key argument: for any convex and L-smooth F, then $\forall \theta, \eta$

$$\|\theta - \gamma\nabla F(\theta) - \eta + \gamma\nabla F(\eta)\|_2^2 = \|\theta - \eta\|_2^2 - 2\gamma\langle\theta - \eta, \nabla F(\theta) - \nabla F(\eta)\rangle + \gamma^2 \underbrace{\|\nabla F(\theta) - \nabla F(\eta)\|_2^2}_{\leq L\langle\nabla F(\theta)-\nabla F(\eta),\theta-\eta\rangle \text{ cocoercivite}}$$

$$\leq \|\theta - \eta\|^2 - \underbrace{2\gamma(1 - \frac{\gamma L}{2})}_{\geq 0 \text{ if } \gamma L \leq 2} \underbrace{\langle\theta - \eta, \nabla F(\theta) - \nabla F(\eta)\rangle}_{\geq 0 \text{ when } F \text{ cvx}}$$

$$(\text{ if } \gamma L \leq 2) \leq \|\theta - \eta\|^2$$

With stong convexity, we would get $\leq (1 - 2\gamma\mu(1 - \frac{\gamma L}{2}))\,|\theta - \eta|^2$

Theorem 28

Consider $A \equiv$ Algo 2, i.e. SGD with T steps & step size $(\gamma_t)_{t=1,\dots,T}$.
Assume that $\forall k \in \{1, \dots, n\}, \psi_k : \theta \mapsto l(Y_k, X_k^T\theta)$ is convex, L-Smooth, BR Lipschitz (which is ok for l B-Lip & $\|X\| \leq R$ a.s.)
Then A is uniformly stable with $\beta(n) = \frac{2B^2R^2\sum_{t=1}^T \gamma_t}{n}$

Nouveau cours du 07/12

Proof. We have to use **averaged** uniform stab (this is a randomized algo!).
Notation shortcuts :

- A($\mathcal{D}_n$, t steps) $\equiv \theta_t$
- A($\mathcal{D}_n^{(k)}$, t steps) $\equiv \tilde{\theta}_t$

$\forall t \geq 1$,

$$\begin{aligned}
&\mathbb{E}[\left\|\theta_{t+1} - \tilde{\theta_{t+1}}\right\| |\mathcal{F}_t] \\
&= \mathbb{E}[\left\|\theta_{t+1} - \tilde{\theta_{t+1}}\right\| \mathbb{1}_{i(t+1)=k} + \left\|\theta_{t+1} - \tilde{\theta_{t+1}}\right\| \mathbb{1}_{i(t+1)\neq k} |\mathcal{F}_t] \\
&= \frac{1}{n}\mathbb{E}[\left\|\theta_{t+1} - \tilde{\theta_{t+1}}\right\| |\mathcal{F}_t, i(t+1) = k] + \frac{n-1}{n}\mathbb{E}[\left\|\theta_{t+1} - \tilde{\theta_{t+1}}\right\| |\mathcal{F}_t, i(t+1) \neq k]
\end{aligned}$$

- Conditionally to $i(t+1) = k$, $(\mathbb{E}[\left\|\theta_{t+1} - \tilde{\theta_{t+1}}\right\| |\mathcal{F}_t, i(t+1) = k])$,

$$\left\|\theta_{t+1} - \tilde{\theta_{t+1}}\right\| \leq \left\|\theta_t - \tilde{\theta}_t\right\| + \gamma_t(\underbrace{\|\nabla\psi_k(\theta_k)\|}_{\leq BR} + \underbrace{\left\|\nabla\tilde{\psi}_k(\tilde{\theta}_k)\right\|}_{\leq BR}).$$

$\|\nabla\psi_k(\theta_k)\| \leq BR$
$\left\|\nabla\tilde{\psi}_k(\tilde{\theta}_k)\right\| \leq BR$

- Conditionally to $i(t+1) \neq k$,
the non-expensivness property gives that $\left\|\theta_{t+1} - \tilde{\theta_{t+1}}\right\| \leq \left\|\theta_t - \tilde{\theta}_t\right\|$
Therefore,

$$\begin{aligned}
\mathbb{E}[\left\|\tilde{\theta_{t+1}} - \theta_{t+1}\right\| |\mathcal{F}_t] &\leq \frac{1}{n}\left\|\theta_t - \tilde{\theta}_t\right\| + \frac{n-1}{n}\left\|\theta_t - \tilde{\theta}_t\right\| \\
&= \left\|\theta_t - \tilde{\theta}_t\right\| + \frac{\gamma_{t+1}2BR}{n} \\
&\cdots \\
&\leq \left\|\theta_0 - \tilde{\theta}_0\right\| + \frac{\sum_{u=1}^{t+1}}{n}
\end{aligned}$$

As previously, we have obtained a bound on $\left\|A(\mathcal{D}_n, Tsteps) - A(\mathcal{D}_n^{(k)}, Tsteps)\right\|$ witch leads to uniform stability with $\beta(n) = \frac{2B^2R^2\sum_{t=1}^{T}\gamma_t}{n}$

□

Conclusion What can be said about the generalization of SGD after T steps?

- Recall that for $T \leq n$, without replacement (one pass over the data), we actually minimize the generalization risk directly : we already obtained a rate in the previous lecture.
- For $T \geq n$ (multipasses over the data) with rempalcement, Lemma 1 of this lecture give

$$\begin{aligned}
\mathcal{R}(A(\mathcal{D}_n, T)) - \mathcal{R}^\star &\leq \underbrace{\beta(n,T)}_{\text{stability}} + \underbrace{Optim(T \text{ iter})}_{Optimisation} \\
&\leq \underbrace{\frac{2\beta^2R^2}{n}\sum\gamma_t}_{\text{Today}} + \underbrace{\frac{\|\theta_0 - \theta^\star\|}{\gamma T} + \gamma\sigma^2}_{\text{with } \gamma \text{ cst (cf previus lecture)}}
\end{aligned}$$

✓ bla

× bla

Good: We have obtained a generalization bound for any nb T of iteration
Trade-off between stability and optimisation. This suggest the existence of an optimal "EARLY STOPPING" time $T^\star$, such that the risk decreases up to $T^\star$

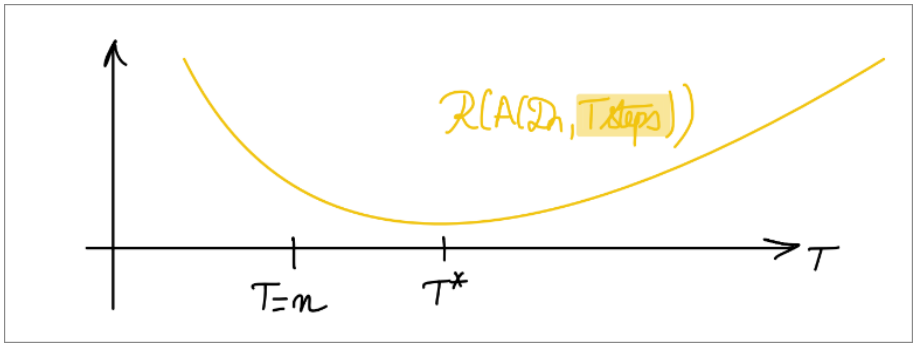


Figure 4.3: Early Stopping : $\min_\gamma \frac{C_1 T\gamma}{n} + \frac{C_2}{\gamma T} + \gamma\sigma^2 \Rightarrow \gamma = \sqrt{\frac{C_2}{T(\frac{C_1 T}{n}+\sigma^2)}}$

Nouveau cours du 13/12

4.5 Big sum up of what have been done

Lecture 1 : $\mathcal{R}(\hat{f} - \mathcal{R}^\star)$ with the help of **uniform bound** over the class of predictors

$$\mathcal{R}(\hat{f} - \mathcal{R}^\star) \leq \text{approx error} + \text{Stochastic error } + \text{ Optim error.}$$

- Stochastic error : $\mathcal{O}(1/\sqrt{n})$ due to uniform bound and due to working with **finite** samples
- optimisation error : **NO NEED TO CONVERGE PRECISELY TOWARDS THE ERM**

Lecture 2 : Basics of deterministic optim

- Definition of convexity, convexe function, μ-strongly convexe, L-smooth ($\rightarrow$ cocoercivity of the gradient).
- Convergence analysis of GD: L-Smooth ($\mathcal{O}(1/T)$) + μ-strongly convex $\mathcal{O}((1 - \underbrace{\kappa}_{=\mu/L})^t)$
- Subgradient algo: 2 terms in the bound $\rightarrow$ need of trade-off $\rightarrow$ Choice of the step size
- Inertial methods : Nesterov, Heavy ball

Lecture 3 : SGD $\theta_{t+1} = \theta_t - \gamma_{t+1}, g_{t+1}(\theta_t)$

- $\rightarrow$ ERM
- $\rightarrow$ True risk minimization
- $\Rightarrow \mathbb{E}[g_t|\mathcal{F}_{t-1}] = \nabla F(\theta_{t-1})$
- Convergence analysis (with gradient estimates of boundede variance) $\rightarrow$ Again trade off between 2 terms that is reflected bu the choice in the step size.
  - L-Smooth : $\gamma = \mathcal{O}(1/\sqrt{t})$
  - + μ-strongly convex : $\gamma = \mathcal{O}(1/t)$

Lecture 4 : Stability of learning algo

$$\text{Gen bound} \leq \frac{\text{Stability } + \text{ Optimisation}}{\text{Algorithm- dependent}}.$$

- ERM
- SGD with multi-pass !! + early stopping as algorithmic regularization.

Lecture 5 : Better stochastic methods
In case of ERM

1. Variance reduction methods
2. Inertial stochastic methods (bof)
3. Adaptive learning rate ($\approx$ Newton quand on peut pas se l'offrir)

2+3 ADAM ♡

Chapter 5

Learning in interpolation regimes

It is usual in ML to resort to an explicit regularization (ridge / losses, ect) to control the "size" of the hypothesis space, and above all the stochastic error.

This approach was the paradigm in ML until a few years ago. However, large scale models are often trained without such a regularization & still achieve state-of-the-art *test* performance in certain task.
→ Counter-intuitive, all the more so as they are over-parameterized and achieve *zero* in training error.

Statistical wisdom suggests that a method that takes advantage of too many degrees of freedom by perfectly interpolating noisy data will be poor at predicting new outcomes.
In deep learning, training algo seem to induce a bias that break the equivalence among all models that interpolate the observed data → **Implicit bias**

5.1 Implicit bias of (S)GD in interpolation regimes

5.1.1 Regression with least squares

Settings : Linear model with quadratic cost $F(\theta) = \frac{1}{2n}\sum_{i=1}^{n}(Y_i - X_i^T\theta)^2$ with $(X_i, Y_i)_{1\leq i\leq n}$ training set, $\theta \in \mathbb{R}^d, X_i \in \mathbb{R}^d, Y_i \in \mathbb{R}$

$$Y = \begin{pmatrix} Y_1 \\ \vdots \\ Y_n \end{pmatrix} \in \mathbb{R}^n, \mathbb{X} = \begin{pmatrix} X_1^T \\ \vdots \\ X_n^T \end{pmatrix}, \rightarrow F(\theta) = \frac{1}{2n}\left\|Y - \mathbb{X}\theta\right\|_2^2 .$$

with $n \ll d$ (overparametrisation). The kernel matrix $\mathbb{X}\mathbb{X}^T \in \mathbb{R}^{n\times n}$ is assumed to be invertible.

Therefore, there exist an infinity of minimizer of F, corresponding to the solutions of the system $Y = \mathbb{X}\theta$. The set of minimizers is actually an affine space (given a solution θ_0 to "$Y = \mathbb{X}\theta$"), $\theta_0 + \ker(\mathbb{X})$ is the entire set of solution

Note. All the minizer achieve zero training error!

Running GD Imagine that you run (S)DG to minimize F without explicite regularization (L2 norm or L1 norm, ect)

$$\begin{cases} \theta_0 \in \mathbb{R}^d \\ \theta_{t+1} = \theta_t - \gamma\nabla F(\theta_t) \end{cases} .$$

Note. Computation of Lip const

$$\nabla F(\theta) = \frac{1}{n}\mathbb{X}^T(\mathbb{X}\theta - Y)$$

$$H_F(\theta) = \frac{1}{n}\mathbb{X}\mathbb{X}^T$$

$$L = \lambda_{max}(\mathbb{X}\mathbb{X}^T)/n$$

When F is convex & L-smooth (with $L == \lambda_{max}(\mathbb{X}\mathbb{X}^T)/n$) the GD CV provided $\gamma < 2/L$ and typically choosing $\gamma = 1/L$ ("optimal" constant step size).

In particular when $\theta_0 = 0$ and $\gamma \leq 1/L$, the GD iterates are $\theta_{t+1} = \theta_t - \frac{\gamma}{n}\mathbb{X}^T(\mathbb{X}\theta_T - Y)$. Thus

$$\begin{aligned}
\mathbb{X}\theta_t - Y &= \mathbb{X}\theta_{t-1} - \frac{\gamma}{n}\mathbb{X}\mathbb{X}^T(\mathbb{X}\theta_{t-1} - Y) - Y \\
&= (I - \frac{\gamma}{n}\mathbb{X}\mathbb{X}^T)(\mathbb{X}\theta_{t-1} - Y) \\
&= (I - \frac{\gamma}{n}\mathbb{X}\mathbb{X}^T)^t(\mathbb{X}\theta_0 - Y) \\
&= -(I - \frac{\gamma}{n}\mathbb{X}\mathbb{X}^T)^t Y(\text{ with } \theta_0 = 0)
\end{aligned}$$

This leads to

$$\|x\theta_t - Y\|_2^2 - 0 \leq (1 - \frac{\gamma}{n}\underbrace{\lambda_{min}}_{\neq 0; \geq 0}(\mathbb{X}\mathbb{X}^T))^{2t}\|Y\|_2^2 .$$

$\times$: Linear convergence of $\mathbb{X}\theta_t$ towards Y Note that when $\theta_0 = 0, \theta_t \in In\mathbb{X}^T = span(\{X_1, \ldots, X_n\})$ for all t. Indeed GD iterates are always linear combination of "$\mathbb{X}^T$ something". We can write $\forall t > 0$ for some $\alpha_t, \theta_t = \mathbb{X}^T\alpha_t$ (*"representer thm in an algorithmic version"*)

Since

$$\begin{aligned}
\|\mathbb{X}\theta_t - Y\|_2^2 &\to_{t\to+\infty} 0 \\
\mathbb{X}\theta_t &\to_{t\to+\infty} Y \\
\mathbb{X}\mathbb{X}^T &\to_{t\to+\infty} Y \\
\left\|\alpha_t - (\mathbb{X}\mathbb{X}^T)^{-1}Y\right\|_2^2 &= \left\|(\mathbb{X}\mathbb{X}^T)^{-1}\mathbb{X}\mathbb{X}^T\alpha_t - (\mathbb{X}\mathbb{X}^T)^{-1}Y\right\|_2^2 \\
&\leq (\lambda_{max}((\mathbb{X}\mathbb{X}^T)^{-1}))^2\|\mathbb{X}\theta_t - Y\|_2^2 \\
\text{then } \alpha_t &\to_{t\to+\infty} (\mathbb{X}\mathbb{X}^T)^{-1}Y
\end{aligned}$$

And finally

$$\left\|\theta_t - \mathbb{X}^T(\mathbb{X}\mathbb{X}^T)^{-1}Y\right\|_2^2 \to_{t\to+\infty} 0.$$

(sanity check: $\mathbb{X}^T(\mathbb{X}\mathbb{X}^T)^{-1}Y$ is solution of the system).

What is $\theta^\star := \mathbb{X}^T(\mathbb{X}\mathbb{X}^T)^{-1}Y$? $\mathbb{X}^T(\mathbb{X}\mathbb{X}^T)^{-1}$ is the pseudo-inverse of $\mathbb{X}$. Indeed when $\mathbb{X} \in \mathbb{R}^{n\times d}$, the SVD decompostion of $\mathbb{X}$ reads as

$$\begin{aligned}
\mathbb{X} &= UDV^T \\
U &\in \mathbb{R}^{n\times n} \text{ orthogonal } U^T = T^{-1} \\
V &\in \mathbb{R}^{d\times d} \text{ orthogonal} \\
D &\in \mathbb{R}^{n\times d}, D = [\begin{pmatrix} \sigma_1 & \ldots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \ldots & \sigma_n \end{pmatrix} |0] \\
\sigma_1 \geq \sigma_2 \geq \ldots \sigma_n &\underbrace{>}_{rk(\mathbb{X})=n} \text{ singular values of } \mathbb{X} \\
\mathbb{X}^T(\mathbb{X}\mathbb{X}^T)^{-1} &= VD^TU^T(UDV^TVD^TU^T)^{-1} \\
&= VD^TU^T(U\begin{pmatrix} \sigma_1^2 & \ldots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \ldots & \sigma_n^2 \end{pmatrix}U^T) \\
&= VD^TU^T(U\begin{pmatrix} 1/\sigma_1^2 & \ldots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \ldots & 1/\sigma_n^2 \end{pmatrix}U^T)
\end{aligned}$$

$\in \underbrace{\mathbb{X}^T(\mathbb{X}\mathbb{X})^{-1}}_{X^\dagger}$ is the pseudo-inverse of the *fat* matrix $\mathbb{X}$

Note. When $d \ll n$, $\mathbb{X}$ is long and $\mathbb{X}^\dagger = (\mathbb{X}^T\mathbb{X})^{-1}\mathbb{X}^T$

We can show that $\theta^\star = \mathbb{X}^T(\mathbb{X}\mathbb{X})^{-1}$ is the solution of "$Y = \mathbb{X}\theta$" of the least l^2-norm

$$\min \frac{1}{2}\|\theta\|_2^2 = \min_{\theta\in\mathbb{R}^d}\max_{\Lambda\in\mathbb{R}^n} \frac{1}{2}\|\theta\|_2^2 + \underbrace{\langle \Lambda, \mathbb{X}\theta - Y\rangle}_{\Lambda^T\mathbb{X}\theta - \Lambda^T Y}$$

KKT conditionns

$$\begin{cases} \theta + \mathbb{X}\Lambda = 1 & (1) \\ Y = \mathbb{X}\theta & (2) \end{cases}.$$

$$\begin{aligned} (1) &\Rightarrow \mathbb{X}\theta + \mathbb{X}\mathbb{X}^T\Lambda = 0 \\ &\Rightarrow \Lambda = -(\mathbb{X}\mathbb{X}^T)^{-1}\underbrace{\mathbb{X}\theta}_{Y} = -(\mathbb{X}\mathbb{X}^T)^{-1}Y \\ \theta^\star &= -\mathbb{X}^T\Lambda = +\mathbb{X}^T(\mathbb{X}\mathbb{X}^T)^{-1}Y \end{aligned}$$

Take-home message : In the case of overparameterized linear regression, the GD initialized at 0 converge towards the solution of "$\mathbb{X}\theta = Y$" of *minimal* l^2-norm, a.k.a. minimal-l^2-norm interpolator.

This can be interpreted as an implicit bias / regularization of GD.

Note. This result holds for gradient-based methods in general using linear combination of current & past gradient :

- ✓ (S)GD
- ✓ (S)GD with momentum
- ✓ Nesterov's acceleration
- × quasi Newton methods
- × diagonally preconditioned methods (Adam/Adagrad)

Note. In the overparameterized regime, SGD will also converge to the min-l^2-norm interpolation, *even* with a *fixed* learning rate. Indeed the minimizers of F also minimize the f_i's! Therefore the stochastic noise in the gradient estimates at the optimum is 0 in the overparameterized regime.

5.1.2 Classification in the interpolation regime/separable case

In this section, we are interested in the behaviour of GD for unregularized logistic regression in the separable setting → interpolation (zero classification error on the training set).

Goal:

$$\min_{\theta\in\mathbb{R}^d} F(\theta) := \frac{1}{n}\sum_{i)1}^{n} \log(1 + \exp(-Y_i X_i^T\theta)).$$

(MLE approach associated to the model $\mathbb{P}(Y = 1|X) = \sigma(X^T\theta^\star)$ with σ sigmoid).

In the case of separable data, $\exists\theta_{sep} \in \mathbb{R}^d, \forall i = 1, \dots, n$

$$Y_i X_i^T\theta_{sep} > 0.$$

here $X_i^t\theta a_{sep}$ is

- > 0 for $Y_i = 1$
- < 0 for $Y_i = -1$

$\frac{1}{n}\sum_{i=1}^n \log(1+\exp(-Y_i X_i^T(\lambda\theta_{sep}))) \to_{\lambda\to+\infty} 0$ (pas sur d'où point la flèche de la limite) → *No* minimizer, Only infimum.

To focus on the key aspect of the problem, we make some simplifications

(i) consider the gradient flow (GF) from some initialisation point $\theta(0) = \theta_0 \in \mathbb{R}^d$ and $\theta'(t) = -\nabla F(\theta(t))$

(ii) we replacethe logistic loss by exponential loss

$$F(\theta) = \frac{1}{n}\sum_{i=1}^{n} \exp(-Y_i X_i^T\theta).$$

Path and time parametrization : Here we only care about the optimization path $\{\theta(t) : t \geq 0\} \subset \mathbb{R}^d$, and more particulary its limit. We remark that the optimization path is unchanged if the objective fct is composed with a differentiable function $h : \mathbb{R} \to \mathbb{R}$ such that $h'(u) > 0$ for $u \in ImF$.

Indeed let $G = h \circ F$ and call θ_h the corresponding GF

$$\begin{aligned} &\theta_h'(t) = -\nabla G(\theta_h(t)) \\ \Leftrightarrow &\theta_h' = -\nabla F(\theta_h(t))h'(F(\theta_h(t))) \end{aligned}$$

Let $s(t) := \int_0^t [h'(F(\theta_h(s)))]^{-1} ds$, then

$$\frac{d}{dt}\theta_h(s(t)) = \theta_h'(s(t))s'(t) \underbrace{=}_{(GF)} -\nabla F(\theta_h(s(t)))\frac{h'(F(\theta_h(t)))}{h'(F(\theta_h(t)))} = \frac{d}{dt}\theta_h(t).$$

which show that the paths $\{\theta_h(t)\}_{t\geq 0}$ and $\{\theta(s)\}_{s\geq 0}$ are the same (up to a reparameterization of time).

Keeping this remark in mind, one can consider (finally we study)

$$F(\theta) := -\log(\frac{1}{n}\sum_{i=1}^n \exp(-Y_i X_i^T \theta)).$$

and the dynamics

$$\theta'(t) = \nabla F(\theta(t)).$$

We remove the minus sign in the definition of GD as $-\log$ has a negative derivative.

Separable data & margin : When (X_i, Y_i) is linearly separable

$$\gamma := \max_{\|\theta\|_2 \leq 1} \underbrace{\min_i Y_i X_i^T \theta}_{\text{margin}}.$$

satisfies $\gamma > 0$

Equivalently, there existe a linear classifier $\theta_{sep} \in \mathbb{R}^d$ which makes no mistake on the training set, i.e. for all i

$$Y_i = sign(X_i^T \theta_{sep}).$$

For such a dataset, a natural predictor is the (unique) max margin predictof that achieves the max.

Recall that by Lagrage duality

$$\begin{aligned} \sup_{\|\theta\|_2 \leq 1} \inf_{1\leq i\leq n} Y_i X_i^T \theta &= \sup_{\|\theta\|_2 \leq 1} \text{ s.t. } \forall i, Y_i X_i^T \theta \geq t \\ &= \inf_{\alpha \in \mathbb{R}_+^n} \sup_{\|\theta\|_2 \leq 1} t + \sum \alpha_i (Y_i X_i^T \theta - t) \\ &= \inf_{\alpha \in \mathbb{R}_+^n} \left\| \sum_{i=1}^n \alpha_i Y_i X_i \right\|_2 \text{ such that } \sum_{i=1}^n \alpha_i = 1 \end{aligned}$$

KKT :

$$\mathcal{L}(t, \theta, \alpha) = t + \sum_i \alpha_i (Y_i X_i^T \theta - t) + \mu(\|\theta\|_2^2 - 1).$$

KKT

1. $\nabla_t \mathcal{L} = 0 \Rightarrow \sum \alpha_i = 1$
2. $\nabla_\theta \mathcal{L} = 0 \Rightarrow \sum \alpha_i Y_i X_i + 2\mu\theta = 1$
3. $\mu = 0$ or $\|\theta\|_2^2 = 1$

So that $\theta \propto \sum_i \alpha_i Y_i X_i$ at the optimum.

By complementary slackness, non-negative α_i is non zeor only for i such that at the optimum $t = Y_i X_i^T \theta$, i.e. for i attaining the minimum $\min_{1\leq i\leq} Y_i X_i^T \theta$ corresponding to the so called *support vectors*.

Nouveau cours du 14/12

Lemma 29

With F defined as $(F(\theta) := -\log(\frac{1}{n}\sum_{i=1}^n \exp(-Y_i X_i^T\theta)))$, it holds

1. $\min_i Y_i X_i^T\theta \leq F(\theta) \leq \min_i Y_i X_i^T\theta + \log_n, \forall\theta \in \mathbb{R}^d$
2. $\|\nabla F(\theta)\|_2 \geq \gamma \forall\theta \in \mathbb{R}^d$

Preuve : (TD, exo tasse de café). $\forall\theta \operatorname{arsinh} \mathbb{R}^d, m_\theta := \min_i Y_i X_i^t\theta$

$$\begin{aligned} e^{-m_\theta} &= \exp(-\min_i Y_i X_i^T\theta) \\ &= \frac{1}{n}\sum_{j=1}^n \underbrace{\exp(-\min_i Y_i X_i^T\theta)}_{\geq Y_j X_j^T\theta, \forall j\in[1,n]} \\ &\geq \frac{1}{n} \underbrace{\sum_{i=1}^n \exp(-Y_i X_i^T\theta)}_{e^{-m_\theta} + \sum_{i=1, i\neq \arg\min X_i}^n e^{\cdots}} \end{aligned}$$

Finally, $e^{-m_\theta} \geq \frac{1}{n}\sum_{i=1}^n e^{-Y_i X_i^T\theta} \geq \frac{1}{n} e^{-m_\theta}$ Apply $-\log$ (a decrease fct) to conclude.

2 - $\mathbb{Z} = \begin{pmatrix} Y_1 X_1^T \\ \dots \\ Y_n X_n^T \end{pmatrix} \in \mathbb{R}^{n\times d}$

$\Delta_n = \{p \in \mathbb{R}_+^n : \sum_{i=1}^n p_i = 1\}$

a -

$$\begin{aligned} \gamma &:= \max_{\|\theta\|_2\leq 1} \min_i Y_i X_i\theta = e_i^T Z\theta \\ &= \max_{\|\theta\|_2\leq 1} \min_{p\in\Delta_n} p^T Z\theta \quad \text{(since } \Delta_n \text{is the convex hull of } \{e_i\}_{i=1\dots n}) \\ &= \min_{p\in\Delta_n} \max_{\|\theta\|_2\leq 1} p^T Z\theta quad \text{ (by the minimax thm for bilinear function in game theory)} \end{aligned}$$

Therefore,

$$\begin{aligned} \gamma &= \min_{p\in\delta_n} \max_{\|\theta_2\|\leq 1} p^T Z\theta = \langle Z^T p, \theta\rangle \\ &= \min_{p\in\delta_n} \left\langle Z^T p, \frac{Z^T p}{\|Z^T p\|_2}\right\rangle \\ &= \min_{p\in\delta_n} \|Z^T p\|_2 \end{aligned}$$

$$F(\theta) = -\log(\sum_{i=1}^n \exp(-Y_i X_i^T\theta))$$

Therefore $\forall\theta$

$$\nabla F(\theta) = +\frac{1}{\sum_{i=1}^n \exp(Y_i X_i^T\theta)} \sum_{i=1}^n \exp(-Y_i X_i^T\theta) Y_i X_i$$

2. $\mathbb{Z} = \begin{pmatrix} Y_1 X_1^T \\ \vdots \\ Y_n X_n^T \end{pmatrix} \in \mathbb{R}^{n\times d}$

Thus $\nabla F(\theta) = Z^T p^{(\theta)}$ with $p^{(\theta)} = (p_i^{(\theta)})_{i=1,\dots,n}, p_i^{(\theta)} = \frac{\exp(-Y_i X_i^T\theta)}{\sum_j \exp(-Y_j X_j^T\theta)}$ □

Theorem 30

Assume that the training sample $(X_i, Y_i)_i$ is seprable (i.e. $\gamma > 0$). For any initial point $\theta_0 \in \mathbb{R}^d, \|\theta(t)\|_2 \xrightarrow[t\to+\infty]{} +\infty$.

Moreover, the renormalized predictor $\hat{\theta}(t) = \frac{\theta'(t)}{\|\theta(t)\|_2}$ converge to the optimal margin at a rate of $\mathcal{O}(1/t)$

Assuming $\theta_0 = \theta$, it holds that for $t \geq t^* = \log(n)/\gamma^2$

$$\min_i Y_i X_i^T \theta(t) \geq \gamma - \frac{\log(n)}{t}.$$

Preuve : The equivalent of the gradient descent leamma in continuous time

$$\frac{d}{dt} F(\theta(t)) = \nabla F(\theta(t))^T \frac{d\theta(t)}{dt} \underbrace{=}_{(GF)} \|\nabla F(\theta(t))\|_2^2 \underbrace{\geq}_{(i)\text{Lemma}} \gamma^2.$$

F grows unbounded and this $\|\theta(t)\|_2 \to +\infty$

It holds that

$$\begin{aligned} F(\theta(t)) - F(\theta_0) = [F(\theta(s))]_0^t &= \int_0^t \nabla F(\theta(s))^T \frac{d\theta}{ds}(s) ds \\ \text{(GF)} \quad &= \int_0^t \|\nabla F(\theta(s))\|_2^2 \, ds \\ &\geq \gamma \int_0^t \|\nabla F(\theta(s))\|_2 \, ds \text{ (Lemma ??)} \end{aligned}$$

With $\theta_0 = 0, F(\theta_0) = 0$

1.

$$\begin{aligned} \min_i Y_i X_i^T \theta(t) &\underbrace{\geq}_{\text{Lemma}} F(\theta(t)) - \log n \\ &\geq \underbrace{\gamma \int_0^t \|\nabla F(\theta(s))\|_2 \, ds - \log n}_{\text{which is non-negative for } t \geq t^*} \end{aligned}$$

and is $\geq \gamma^2 - \log n \geq 0$ when $t \geq t^*$

Note that

$$\begin{aligned} \|\theta(t)\|_2 &= \int_0^t \underbrace{\frac{d}{ds} \|\theta(s)\|_2}_{\frac{\theta(s)^T}{\|\theta(s)\|_2} \frac{d\theta}{ds}(s)} ds \\ &\leq \int_0^t \left\| \frac{d}{ds} \theta(s) \right\|_2 ds \\ \text{(GF)} \quad &= \int_0^t \|\nabla F(\theta(s))\|_2 \, ds \end{aligned}$$

2. Then, $\frac{1}{\|\theta(t)\|_2} \geq \frac{1}{\int_0^t \|\nabla F(\theta(s))\|_2} ds$

(1)x(2) , for $t \geq t^*$

$$\begin{aligned} \min_i Y_i X_i^T \theta(t) &\geq \gamma - \frac{\log n}{\int_0^t \|\nabla F(\theta(s))\|_2 \, ds} \\ &\underbrace{\geq}_{\text{Lemma}} \gamma - \frac{\log n}{\gamma t} \end{aligned}$$

□

✓ For classification task, we only care about the sign of the prediction at test time, so the fact that $\|\theta(t)\|_2 \to +\infty$ is not really an issue.

✓ GD diverge but towards the maxmargin classifier.

✓ This **algorithmic implicit** bias is preserved for GD and the logistic loss. With hands, with separable data, the logistic loss F has an infinimum equal to 0. For any sequence θ_t such for all $Y_i X_i^T \theta \xrightarrow[t\to+\infty]{} +\infty$

$$F(\theta_t) \xrightarrow[t\to+\infty]{} \inf_\theta F(\theta) = 0.$$

It turns out that GD diverges along a direction that is

$$\begin{cases} \|\theta_t\|_2 & \xrightarrow[t\to+\infty]{} +\infty \\ \frac{\theta_t}{\|\theta_2\|_2} & \xrightarrow[t\to+\infty]{} \eta \in \mathbb{R}^d \end{cases}.$$

for some ηt of unit-norm.

$$\nabla F(\theta) = -\frac{1}{n}\sum_{i=1}^n \frac{\exp(-Y_i X_i^T\theta)}{1+\exp(-Y_i X_i^T\theta)} Y_i X_i.$$

By the structure of the sum of exponentials, the dominant term in $\nabla F(\theta_t)$ corresponds to the indeces i for which $-Y_i X_i^T \eta$ is the largest $\rightsquigarrow$ the support vectors !
Biblio [Lyu & Li](2019) extention to homogenous NN.

5.2 Statistical analysis of overparametrization: the double descent phenomenon

Recall that classical error bounds scale as follows

Proposition 31

B-Lipschitz loss ℓ
$\mathcal{F} = \{f_\theta : f_\theta(x) = \theta^T\phi(x) \, \|\theta\|_2 \le D\}$ where $\mathbb{E}\,\|\phi(x)\|_2^2 \le R^2$

$$\mathbb{E}[\mathcal{R}(f_{\hat\theta}^{\mathrm{ERM}})] \le \inf_{\|\theta\|_2 \le D} \mathcal{R}(f_\theta) + \frac{2BRD}{\sqrt{n}}.$$

The "capacity" of the class $\mathcal{F}$ of learners is controlled here by the norm of its parameter ; it oculd be the number of parameters as well. This type of bounds is in accordance with the following classical scheme

Figure 5.1: Traditional trade-off between approximation and stochastic errors

However modern architectures involve always more param (achieving zero training error & achieve state-of-the-art test perf.

When the capacity of $\mathcal{F} \nearrow$, the model becomes over-parameterized, a phenomenon occurs : after the test error explodes, it goes down again.
$\to$ this is so-called **double descent**. We are going to analyze it for a linear model! (with gaussian features)

Model $Y_i = X_i^T\theta^\star + \epsilon_i, \theta^\star \in \mathbb{R}^d$, $X_i \sim \mathcal{N}(0, Id_d), \epsilon_i \sim \mathcal{N}(0,\sigma^2) \perp\!\!\!\perp X_i, \theta^\star \in \arg\min_\theta \mathbb{E}[(Y - X^T\theta)^2]$

Estimator : ERM $\hat{\theta} \in \arg\min_\theta \frac{1}{n}\sum_{i=1}^n (Y_i - X_i^T\theta)^2 = \frac{1}{n}\|Y - \mathbb{X}\theta\|_2^2$.
Optimality condition : $\mathbb{X}^T\mathbb{X}\theta = \mathbb{X}^T Y$.

When $d < n$ (underparametrized regime) $\mathbb{X}^T\mathbb{X} \in \mathbb{R}^{d\times d}$ is almost surely invertible. The traditional least-square estimator is

$$\hat{\theta} = (\mathbb{X}^T\mathbb{X})^{-1}\mathbb{X}^T Y.$$

Then,

$$\begin{aligned}
\mathcal{R}(\hat{\theta}) &= \mathbb{E}[(X^T\hat{\theta} - Y)^2] \\
&= \mathbb{E}[(X^T\hat{\theta} - Y + X^T\theta^\star - X^T\theta^\star)^2] \\
&= \mathbb{E}[(X^T\hat{\theta} - X^T\theta^\star)^2] + \mathbb{E}[(X^T\theta^\star - Y)^2] \\
\mathcal{R}(\hat{\theta}) - \mathcal{R}(\theta^\star) &= \mathbb{E}[(X^T\hat{\theta} - X^t\theta^\star)^2] \\
&= \mathbb{E}[\left\langle X, \hat{\theta} - \theta^\star \right\rangle^2] \\
&= \mathbb{E}[(\hat{\theta} - \theta^\star)^T XX^T(\hat{\theta} - \theta^\star)] \\
&= (\hat{\theta} - \theta^\star)^T \underbrace{\mathbb{E}[XX^T]}_{Id}(\hat{\theta} - \theta^\star) \text{ conditionally to } \mathcal{D}_n \\
&= \left\|\hat{\theta} - \theta^\star\right\|_2^2 \\
\mathbb{E}_{\mathcal{D}_n}[\mathcal{R}(\hat{\theta}) - \mathcal{R}(\theta^\star)] &= \mathbb{E}_{\mathcal{D}_n}[\left\|(\mathbb{X}^T\mathbb{X})^{-1}\mathbb{X}^T Y - \theta^\star\right\|_2^2] \\
&= \mathbb{E}_{\mathcal{D}_n}[\left\|(\mathbb{X}^T\mathbb{X})^{-1}\mathbb{X}^T\mathbb{X}\theta^\star - \theta^\star + (\mathbb{X}^T\mathbb{X})^{-1}\mathbb{X}^T\epsilon\right\|_2^2] \\
&= \mathbb{E}_{\mathcal{D}_n}[\left\|Id.\theta^\star - \theta^\star + (\mathbb{X}^T\mathbb{X})^{-1}\mathbb{X}^T\epsilon\right\|_2^2] \\
&= \mathbb{E}_{\mathcal{D}_n}[\left\|(\mathbb{X}^T\mathbb{X})^{-1}\mathbb{X}^T\epsilon\right\|_2^2] \\
&= \mathbb{E}_{\mathcal{D}_n}[Tr(\underbrace{\epsilon^T\mathbb{X}(\mathbb{X}^T\mathbb{X})^{-1}(\mathbb{X}^T\mathbb{X})^{-1}\mathbb{X}^T\epsilon}_{\mathbb{R}})] \\
&= \mathbb{E}_{\mathcal{D}_n}[Tr(\mathbb{X}(\mathbb{X}^T\mathbb{X})^{-2}\mathbb{X}^T\epsilon\epsilon^T)] \\
&= \sigma^2\mathbb{E}_{\mathbb{X}}[Tr((\mathbb{X}^T\mathbb{X})^{-2}\mathbb{X}^T\mathbb{X})] \\
&= \sigma^2\mathbb{E}_{\mathbb{X}}[Tr((\mathbb{X}^T\mathbb{X})^{-\not{2}-1}\ \not{\mathbb{X}^T}\ \not{\mathbb{X}})] \\
&= \sigma^2\mathbb{E}_{\mathbb{X}}[Tr[(\mathbb{X}^T\mathbb{X})^{-1}]]
\end{aligned}$$

$\mathbb{X}^T\mathbb{X} \in \mathbb{R}^D$ is a Wishart matrix with n degrees of freedom almost surely invertible

$$\begin{aligned}
&\text{(admitted)} \\
&= \begin{cases} \sigma^2\frac{d}{n-d-1} & \text{if } n \geq d+2 \\ +\infty & \text{if } n = d \text{ or } n = d+1 \end{cases}
\end{aligned}$$

When $d \geq n$ (over-parametrized regime) The kernel matrix $\mathbb{X}\mathbb{X}^T \in \mathbb{R}^{n\times n}$ is a.s. invertible. There exist plenty of ERM, we are going to focus on a solution in particular : the minimum ℓ^2 norm interpolator (limit iterate of (S)GD strategies!)

In this case

$$\begin{aligned}
\hat{\theta} &= \mathbb{X}^\dagger Y = \mathbb{X}^T(\mathbb{X}\mathbb{X}^T)^{-1}Y \\
&= \mathbb{X}^T(\mathbb{X}\mathbb{X}^T)^{-1}\mathbb{X}\theta^* + \mathbb{X}^T(\mathbb{X}\mathbb{X}^T)^{-1}\varepsilon
\end{aligned}$$

The expected risk can be computed

$$\begin{aligned}
\mathbb{E}_{\mathcal{D}_n}[\mathcal{R}(\hat{\theta}) - \mathcal{R}(\theta^\star)] &= \mathbb{E}_{\mathcal{D}_n}[\left\|\hat{\theta} - \theta^\star\right\|_2^2] \\
&= \mathbb{E}_{\mathcal{D}_n}[\left\|\mathbb{X}^T(\mathbb{X}\mathbb{X}^T)^{-1}\mathbb{X}\theta^* + \mathbb{X}^T(\mathbb{X}\mathbb{X}^T)^{-1}\varepsilon - \theta^\star\right\|_2^2] \\
&= \mathbb{E}_{\mathcal{D}_n}[\left\|\mathbb{X}^T(\mathbb{X}\mathbb{X}^T)^{-1}\mathbb{X}\theta^\star - \theta^\star\right\|_2^2] + \mathbb{E}_{\mathcal{D}_n}[\left\|\mathbb{X}^T(\mathbb{X}\mathbb{X}^T)^{-1}\epsilon\right\|_2^2]
\end{aligned}$$

We find back two term : bias & variance

- Variance

$$\mathbb{E}_{\mathcal{D}_n}[\left\|\mathbb{X}^T(\mathbb{X}\mathbb{X}^T)^{-1}\epsilon\right\|_2^2] = \mathbb{E}[Tr(\epsilon^T(\mathbb{X}\mathbb{X}^T)^{-1}\mathbb{X}\mathbb{X}^T(\mathbb{X}\mathbb{X}^T)^{-1}\epsilon)] \qquad = \sigma^2\mathbb{E}[Tr((\mathbb{X}\mathbb{X}^T)^{-1}$$

$$= \begin{cases} \sigma^2 \frac{n}{d-n-1} & \text{if } d \geq n+2 \\ +\infty & \text{if } d = n \text{ or } d = n+1 \end{cases} \text{ similarly to previous calculation}$$

- Bias

$$\mathbb{E}_{\mathcal{D}_n}[\left\|\mathbb{X}^T(\mathbb{X}\mathbb{X}^T)^{-1}\mathbb{X}\theta^\star - \theta^\star\right\|_2^2] = \mathbb{E}_{\mathcal{D}_n}[\left\| \underbrace{(\mathbb{X}^T(\mathbb{X}\mathbb{X}^T)^{-1}\mathbb{X} - I)}_{\text{projection matrix on } \ker\mathbb{X} \text{ (up to a sign)}} \theta^\star \right\|_2^2]$$
$$= \mathbb{E}[\|Proj_{\ker\mathbb{X}}(\theta^\star)\|_2^2]$$
$$= \mathbb{E}[(\theta^\star)^T(I - \mathbb{X}^T(\mathbb{X}\mathbb{X}^T)^{-1}\mathbb{X})\theta^\star]$$

(projection matrices are independent $P^2 = P$).
Introduce $\mathcal{R}^{(\ell)}$ to be the rotation such that $\theta^* = \|\theta^*\|_2 \mathcal{R}^{(\ell)} e_\ell$; i.e. which rotates the ℓ-th vector of the canonical basis over θ^*.

$$\mathbb{E}[(\theta^\star)^T\mathbb{X}^T(\mathbb{X}\mathbb{X}^T)^{-1}\mathbb{X}\theta^\star] = \|\theta^\star\|_2^2 \mathbb{E}[e_\ell^T\mathcal{R}^{(\ell)T}\mathbb{X}^T(\mathbb{X}\mathbb{X}^T)^{-1}\mathbb{X}\mathcal{R}^{(\ell)}e_\ell]$$
$$= \|\theta^\star\|_2^2 e_\ell \mathbb{E}[(\mathbb{X}\mathcal{R}^{(\ell)})^T(\mathbb{X}\underbrace{\mathcal{R}^{(\ell)}\mathcal{R}^{(\ell)T}}_{Id}\mathbb{X}^T)^{-1}\mathbb{X}\mathcal{R}^{(\ell)}]e_\ell$$

$\mathbb{X}\mathcal{R}^{(\ell)}$ has the same distribution as $\mathbb{X}$

$$= \|\theta^\star\|_2^2 e_\ell^T\mathbb{E}[\mathbb{X}^T(\mathbb{X}\mathbb{X}^T)^{-1}\mathbb{X}]e_\ell \text{ for all } \ell$$
$$= \frac{\|\theta^\star\|_2^2}{d}\sum_{\ell=1}^{d} e_\ell^T\mathbb{E}[\mathbb{X}^T(\mathbb{X}\mathbb{X}^T)^{-1}\mathbb{X}]e_\ell$$
$$= \frac{\|\theta^\star\|_2^2}{d}Tr(\mathbb{X}^T(\mathbb{X}\mathbb{X}^T)^{-1}\mathbb{X}) \text{(by linearity of the trace)}$$
$$= \frac{\|\theta^*\|_2^2}{d}\mathbb{E}[Tr(Id_n)]$$
$$= \frac{\|\theta^*\|_2^2 n}{d}$$

(bias) $\mathbb{E}_{\mathcal{D}_n}[\left\|\mathbb{X}^T(\mathbb{X}\mathbb{X}^T)^{-1}\mathbb{X}\theta^\star - \theta^\star\right\|_2^2] = \frac{\|\theta^\star\|_2^2}{d}(d-n)$

Overall,

$$\mathbb{E}_{\mathcal{D}_n}[\mathcal{R}(\hat{\theta}) - \mathcal{R}(\theta^*)] = \begin{cases} \sigma^2\frac{n}{d-n-1} + \|\theta^*\|_2^2\frac{d-n}{d} & \text{if } d \geq n+2 \\ +\infty & \text{o.w} \end{cases}.$$

In the overparameterized regime, the risk does not necessarily explode when d. But this holds for a particular ERM, and not any one !!