Optimisation Stochastique

Charles Vin

S1-2023

Nouveau cours du 16/11

CCL du cours de la dernière fois

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

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

In this section, no empirical proof, no n

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

Lemme 1

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

Preuve : 1.

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

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

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

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

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

3. Example for Least Square :

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

   See Table 0.1

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

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

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

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

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

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

□

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

Lemme 2 (Zhang)

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

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

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

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

Preuve :

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

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

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

□

CCL : $\forall \hat{h}$

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

where

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

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

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

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

1. Method 1 :

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

From the developpement before on the estimation error

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

2. Method 2 : Direct computation

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

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

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

Chapter 1

Basics of deterministic optimisation

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

1.1 First order method

1.1.1 Basics of convex analysis

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

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

μ-strongly convex, $\mu > 0$.

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

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

Lemme 3 (Descent lemma)

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

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

Preuve :

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

□

Consequence of this quadratics upper bound

1.

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

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

1.1.2 Gradient algorithms

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

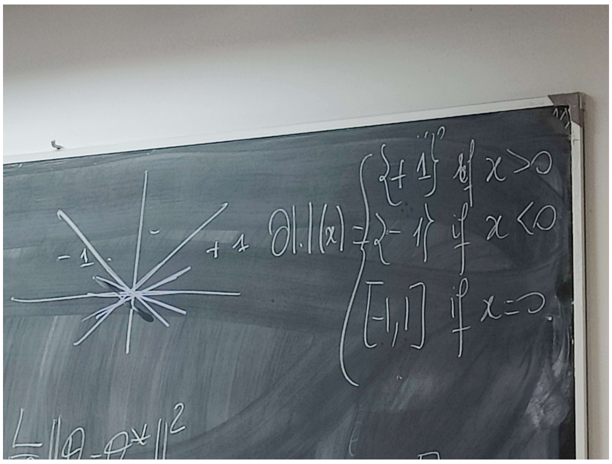

Figure 1.1: subgradients

Gradient algo

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

Choice of steps :

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

Link with the gradient flow

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

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

can be rewrittent as

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

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

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

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

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

Théorème 4

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

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

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

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

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

2. Using a constant step size.

γ	0	$1/L$	$2/L$
the rate			

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

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

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

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

Preuve :

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

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

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

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

Preuve : Define two function

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

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

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

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

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

Sum the 2 inequalities to conclude □

End of the co-coercivity note

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

□