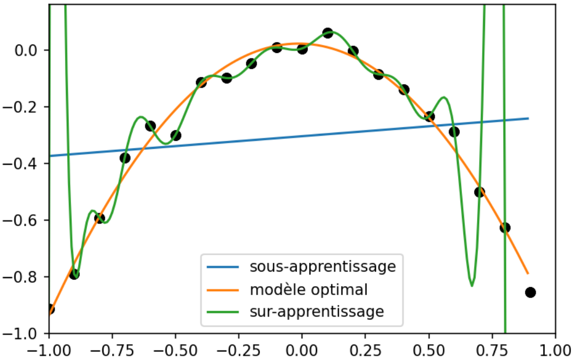
Stabilité en apprentissage statistique

Axel Forveille
Arnaud Gardille
Sofiane Dakhmouche

Université Paris-Saclay
Département de mathémathiques

16 juin 2021

Sur-apprentissage, Compromis biais-variance



Cadre formel

Apprentissage supervisé

- $Y = h^*(X)$
- Classe de fonctions d'approximation $\mathcal{H} = \{h_\theta, \ \theta \in \Theta\}$
- Échantillon $\quad (X_i, Y_i)_{1 \leq i \leq n} \quad$ avec $X_i \in \mathcal{X}, \quad Y_i \in \mathbb{R}$

Cadre formel

- Déterminer θ tel que h_θ soit proche de h^*
- Fonction de perte

$$\ell : \mathcal{H} \times \mathcal{Z} \longrightarrow \mathbb{R}_+$$

.

- *Algorithme d'apprentissage*

$$\mathcal{A} : \begin{cases} (\mathcal{X} \times \mathcal{Y})^n & \longrightarrow (\mathcal{X} \longrightarrow \mathcal{Y}) \\ S & \longmapsto h_{\mathcal{A},S} \end{cases}$$

- Risque

$$R(h) := \mathbb{E}_{Z\sim\mathbb{P}}\left[\ell(h, Z)\right].$$

Cadre formel

- $\mathcal{X}$ est un espace de Banach séparable.
- $\mathcal{Y} = \mathbb{R}$.
- $\mathcal{H} \subseteq \{h : \mathcal{X} \longrightarrow \mathcal{Y} \mid h \text{ est linéaire continue } \} =: \mathfrak{B}$ (dual topologique de $\mathcal{X}$)
- $\mathcal{H}$ est donc muni de la norme opérateur,

$$\forall h \in \mathcal{H}, \ \|h\| = \sup_{x \in \mathcal{X} - 0} \frac{\|h(x)\|_{\mathcal{Y}}}{\|x\|_{\mathcal{X}}}$$

Cadre formel

Notations :
$S := (Z_1, \ldots, Z_n)$ et $S^i := (Z_1, \ldots, Z_{i-1}, Z_i', Z_{i+1}, \ldots, Z_n)$

Définitions

- Stabilité uniforme $|\ell(h_S, Z) - \ell(h_{S^i}, Z)| \leq \beta(n)$ p.s.
- Stabilité uniforme des arguments
  $\forall i \in \{1, \ldots, n\},\ \|h_S - h_{S^i}\| \leq \alpha(n)$
- Stabilité des arguments
  $\forall i \in \{1, \ldots, n\},\ \mathbb{E}[\|h_S - h_{S^i}\| \, |S] \leq \alpha(n)$ p.s

Cadre

- $(\mathfrak{B}, \|.\|)$ une Banach séparable et $(2, D)$-smooth :

$$\forall h, h' \in \mathfrak{B}, \|h + h'\|^2 + \|h - h'\|^2 \leq 2\|h\|^2 + 2D^2\|h'\|^2.$$

- Soit $\mathbb{M} = (M_n)_{n\geq 0}$ une martingale par rapport à sa filtration naturelle $\mathbb{F} = (F_n)_{n\geq 0}$
- $\mathbb{D} = (D_n)_{n\geq 0}$ le processus définit par

$$\begin{cases} D_0 = M_0 \\ D_n = M_n - M_{n-1} \text{ pour tout } n \geq 1 \end{cases}$$

Observations

- $\sum_{i=0}^{n} D_i = M_n$
- le processus $\mathbb{D}$ est adapté à la filtration $\mathbb{F}$,
- pour tout $n \geq 0$, $\mathbb{E}[D_n] < \infty$,
- pour tout $n \geq 1$, $\mathbb{E}[D_n | F_{n-1}] = 0$.

On dit que le processus D est une suite de différences de martingales.

Proposotion 1, Pinelis 1994

- Notons $D_*^2 := \max(1, D^2)$
- Supposons qu'il existe $C^2 > 0$ tel que : $\sum_{i \geq 0} \|D_i\|_\infty^2 \leq C^2$

Alors, pour tout $r \geq 0$,

$$\mathbb{P}\left(\sup_{n \geq 1} \left\| \sum_{i=0}^{n} D_i \right\| \geq r \right) \leq 2 \exp\left(-\frac{r^2}{2 D_*^2 C^2}\right).$$

Preuve

Notons pour tout $t \in \mathbb{R}$ et $j \geq 1$,

$$\phi(t) := \mathbb{E}\left[\mathrm{ch}(\lambda \| M_{j-1} + tD_j \|) | F_{j-1}\right] = \mathbb{E}\left[\mathrm{ch}(u(t)) | F_{j-1}\right],$$

où $u(t) := \lambda \| M_{j-1} + tD_j \|$.
Alors ϕ est deux fois dérivable sur $[-1, 1]$ et :

$$\phi'(t) = \lambda \mathbb{E}\left[\mathrm{sh}(u(t)) D_j d_{M_{j-1}+tD_j} \|.\|(1) | F_{j-1}\right],$$

$$\phi''(t) = \mathbb{E}\left[\mathrm{ch}(u(t))'' | F_{j-1}\right].$$

Preuve

Donc,

$$\begin{aligned}\phi'(0) &= \lambda\mathbb{E}\left[\mathrm{sh}(M_{j-1})D_j d_{M_{j-1}}\|.\|(1)|F_{j-1}\right]\\ &= \lambda\mathrm{sh}(M_{j-1})d_{M_{j-1}}\|.\|(1)\mathbb{E}\left[D_j|F_{j-1}\right] = 0.\end{aligned}$$

Et par $\mathrm{ch}u(t)'' \leq D_*^2\|v\|^2\mathrm{ch}u(t)$,

$$\begin{aligned}\phi''(t) &\leq D_*^2\mathbb{E}[\|\lambda D_j\|^2\mathrm{ch}u(t)|F_{j-1}]\\ &\leq \lambda^2 D_*^2\|D_j\|_\infty^2\mathbb{E}[\mathrm{ch}u(t)|F_{j-1}]\\ &= \lambda^2 D_*^2\|D_j\|_\infty^2\phi(t).\end{aligned}$$

Preuve

Lemme

Soit $f : \mathbb{R} \to [0, \infty[$ C^2 et $R \in \mathbb{R}$ telle que :

- $f'(0) = 0$,
- $f'' \leq R^2 f$.

Alors pour tout $t \in \mathbb{R}$: $f(t) \leq f(0)\mathrm{ch}(Rt) \leq f(0) \exp \frac{(Rt)^2}{2}$.

$$\left\{ \begin{array}{l} \phi'(0) = 0 \\ \phi''(t) \leq \lambda^2 D_*^2 \|D_j\|_\infty^2 \phi(t) \\ \text{Lemme en } t = 1 \end{array} \right. \implies \phi(1) \leq \exp\left(\frac{\lambda^2 D_*^2 \|D_j\|_\infty^2}{2}\right) \phi(0).$$

Preuve

Donc, en remplaçant par les valeurs de ϕ :

$$\mathbb{E}[\mathrm{ch}(\lambda \| M_j \|) | F_{j-1}] \leq \exp\left(\frac{\lambda^2 D_*^2 \| D_j \|_\infty^2}{2}\right) \mathrm{ch}(\lambda \| M_{j-1} \|). \qquad (1)$$

Construisons, compte tenu de (1), la surmartingale $\mathbb{G}$ définie par :

$$G_j = \exp\left(-\frac{\lambda^2 D_*^2 s_j^2}{2}\right) \mathrm{ch}(\lambda \| M_j \|),$$

où $s_j = \sqrt{\sum_{i=0}^{j} \| D_i \|_\infty^2}$.

Preuve

Soit $\tau_r := \inf\{i \in \mathbb{N} | \|M_i\| \geq r\}$ le temps d'arrêt pour la filtration $\mathbb{F}$. Alors, $(G_{n\wedge\tau_r})_n$ est une surmatingale positive donc converge presque sûrement et donc :

$$\mathbb{E}[G_{\tau_r}] \leq \mathbb{E}[G_0] = 1. \tag{2}$$

Preuve

Finalement pour tout $\lambda > 0$, par croissance de ch et ch$u > e^u/2$:

$$
\begin{aligned}
\mathbb{P}\left(\sup_{n\geq 1}\left\|\sum_{i=0}^{n} D_i\right\| \geq r\right) &\leq \mathbb{P}\left(G_{\tau_r} \geq \exp\left(-\frac{\lambda^2 D_*^2 s_{\tau_r}^2}{2}\right)\mathrm{ch}(\lambda r)\right) \\
&\leq \frac{\exp\left(\frac{\lambda^2 D_*^2 s_{\tau_r}^2}{2}\right)}{\mathrm{ch}(\lambda r)}\mathbb{E}[G_{\tau_r}] \text{ par Markov} \\
&\leq 2\exp\left(-\lambda r + \frac{\lambda^2 D_*^2 C^2}{2}\right) \quad \text{par (2)} \\
&\leq 2\exp\left(-\frac{r^2}{2D_*^2 C^2}\right) \quad \text{par minimisation.}
\end{aligned}
$$

Proposotion 1, Pinelis 1994

- Notons $D_*^2 := \max(1, D^2)$
- Supposons qu'il existe $C^2 > 0$ tel que : $\sum_{i\geq 0} \|D_i\|_\infty^2 \leq C^2$

Alors, pour tout $r \geq 0$,

$$\mathbb{P}\left(\sup_{n\geq 1}\left\|\sum_{i=0}^{n} D_i\right\| \geq r\right) \leq 2\exp\left(-\frac{r^2}{2D_*^2 C^2}\right).$$

Lemme 1, Concentration des hypothèses

Supposons que,

- $(\mathfrak{B}, \|.\|)$ est $(2, D)$-smooth,
- $\mathcal{A}$ est un algorithme $\alpha(n)$-argument stable.

Alors pour tout échantillon S et tout $\delta > 0$, avec probabilité au moins $1 - \delta$:

$$\|h_S - \mathbb{E}[h_S]\| \leq D_* \alpha(n) \sqrt{2Bn \log(2\delta^{-1})}.$$

Preuve

- Soient $(Z_i)_{i\geq 1}$ des V.A. de loi $\mathbb{P}$ et $S := (Z_1, ..., Z_n)$,
- Notons $\begin{cases} D_t = \mathbb{E}[h_S | Z_1, ..., Z_t] - \mathbb{E}[h_S | Z_1, ..., Z_{t-1}] \\ D_1 = \mathbb{E}[h_S] \end{cases}$

Alors,

- $(D_t)_{t\geq 1}$ est une suite de différences de martingales,
- $h_S - \mathbb{E}[h_S] = \sum_{t=1}^{n} D_t$.

Preuve

$$
\begin{aligned}
\sum_{t\geq 1} \|D_t\|_\infty^2 &= \sum_{t=1}^{n} \|\mathbb{E}[h_S|Z_1, ..., Z_t] - \mathbb{E}[h_S|Z_1, ..., Z_{t-1}]\|_\infty^2 \\
&\text{car } D_t = 0 \text{ pour tout } t \geq n+1 \\
&= \sum_{t=1}^{n} \|\mathbb{E}[h_S - h_{S^t}|Z_1, ..., Z_t]\|_\infty^2 \\
&\text{car } \mathbb{E}[h_S|Z_1, ..., Z_{t-1}] = \mathbb{E}[h_{S^t}|Z_1, ..., Z_t] \\
&\leq \sum_{t=1}^{n} \mathbb{E}[\|h_S - h_{S^t}\|_\infty|Z_1, ..., Z_t]^2
\end{aligned}
$$

Preuve

$$
\begin{aligned}
\sum_{t\geq 1} \|D_t\|_\infty^2 &\leq \sum_{t=1}^{n} \mathbb{E}[\|h_S - h_{S^t}\|_\infty | Z_1, \ldots, Z_t]^2 \\
&= \sum_{t=1}^{n} \mathbb{E}[\mathbb{E}[\|h_S - h_{S^t}\|_\infty | S] | Z_1, \ldots, Z_t]^2 \\
&\text{car } \sigma(Z_1, \ldots, Z_t) \subset \sigma(S) \\
&\leq Bn\alpha(n)^2 \\
&\text{car } \|h_S - h_{S^t}\|_\infty \leq B\|h_S - h_{S^t}\|.
\end{aligned}
$$

Preuve

Proposotion 1, Pinelis 1994

- Notons $D_*^2 := \max(1, D^2)$
- Supposons qu'il existe $C^2 > 0$ tel que : $\sum_{i \geq 0} \|D_i\|_\infty^2 \leq C^2$

Alors, pour tout $r \geq 0$,

$$\mathbb{P}\left(\sup_{n \geq 1} \left\| \sum_{i=0}^{n} D_i \right\| \geq r\right) \leq 2 \exp\left(-\frac{r^2}{2 D_*^2 C^2}\right).$$

Preuve

Donc, pour tout $r > 0$:

$$\begin{aligned}\mathbb{P}\left(\|h_S - \mathbb{E}[h_S]\| \leq r\right) &\geq \mathbb{P}\left(\sup_{n\geq 1}\left\|\sum_{i=0}^{n} D_i\right\| \leq r\right) \\ &\geq 1 - 2\exp\left(-\frac{r^2}{2D_*^2 Bn\alpha(n)^2}\right).\end{aligned}$$

D'où, pour tout $\delta > 0$:

$$\mathbb{P}\left(\|h_S - \mathbb{E}[h_S]\| \leq D_*\alpha(n)\sqrt{2Bn\log(2\delta^{-1})}\right) \geq 1 - \delta.$$

Notation

- Classe algorithmique d'hypothèses

$$B_r := \left\{ h \in \mathcal{H} \mid \quad \|h - \mathbb{E} h_S\| \leq \underbrace{D\alpha(n)\sqrt{2n\log(2/\delta)}}_{:=r(n,\delta)} \right\}.$$

Outils

- Complexité de Rademacher

$$\mathcal{R}(\mathcal{H}) = \mathbb{E} \sup_{h \in \mathcal{H}} \frac{1}{n} \sum_{i=1}^{n} \sigma_i \langle h, X_i \rangle$$

où $\mathbb{P}(\sigma_i = \pm 1) = \frac{1}{2}$

- Le Banach $(\mathcal{X}, \|.\|)$ est de type $p \geq 1$ si,

$$\exists C_p > 0, \forall x_1, \cdots, x_n, \quad \mathbb{E} \left\| \sum_{i=1}^{n} \sigma_i x_i \right\| \leq C_p \left(\sum_{i=1}^{n} \|x_i\|^p \right)^{1/p}$$

Borne sur la complexité de Rademacher

Théorème 1, [Liu et al., 2017]

Supposons,

- $\mathcal{X}$ Banach séparable de type $p \geq 1$
- le dual topologique de $\mathcal{X}$ est $(2, D)$-smooth
- $\exists B > 0, \ \|X_i\| \leq B \ p.s.$
- h_S est construite par un algorithme $\alpha(n)$-argument stable.

Alors,

$$\mathcal{R}(B_r) \leq DC_p B \sqrt{2 \log(2/\delta)} \alpha(n) n^{-1/2+1/p}$$

Borne sur la complexité de Rademacher

Preuve :

$$\begin{aligned}
\mathcal{R}(B_r) &= \mathbb{E} \sup_{h \in B_r} \frac{1}{n} \sum_{i=1}^{n} \sigma_i \langle h, X_i \rangle \\
&= \mathbb{E} \sup_{h \in B_r} \frac{1}{n} \sum_{i=1}^{n} \sigma_i \langle h, X_i \rangle - \sigma_i \mathbb{E}[\langle h_{S^i}, X_i \rangle | X_i] + \sigma_i \mathbb{E}[\langle h_{S^i}, X_i \rangle | X_i] \\
&= \mathbb{E} \sup_{h \in B_r} \frac{1}{n} \sum_{i=1}^{n} \sigma_i (\langle h, X_i \rangle - \mathbb{E}[\langle h_{S^i}, X_i \rangle | X_i]) \\
&= \mathbb{E} \sup_{h \in B_r} \frac{1}{n} \sum_{i=1}^{n} \sigma_i (\langle h, X_i \rangle - \langle \mathbb{E} h_{S^i}, X_i \rangle)
\end{aligned}$$

Borne sur la complexité de Rademacher

Preuve :

$$\mathcal{R}(B_r) = \mathbb{E} \sup_{h \in B_r} \frac{1}{n} \sum_{i=1}^{n} \sigma_i \langle h, X_i \rangle$$

$$= \mathbb{E} \sup_{h \in B_r} \frac{1}{n} \sum_{i=1}^{n} \sigma_i \langle h, X_i \rangle - \sigma_i \mathbb{E}[\langle h_{S^i}, X_i \rangle | X_i] + \sigma_i \mathbb{E}[\langle h_{S^i}, X_i \rangle | X_i]$$

$$= \mathbb{E} \sup_{h \in B_r} \frac{1}{n} \sum_{i=1}^{n} \sigma_i (\langle h, X_i \rangle - \langle \mathbb{E} h_{S^i}, X_i \rangle)$$

Borne sur la complexité de Rademacher

Preuve :

$$\begin{aligned}
\mathcal{R}(B_r) &= \mathbb{E} \sup_{h \in B_r} \frac{1}{n} \sum_{i=1}^{n} \sigma_i \langle h, X_i \rangle \\
&= \mathbb{E} \sup_{h \in B_r} \frac{1}{n} \sum_{i=1}^{n} \sigma_i \langle h, X_i \rangle - \sigma_i \mathbb{E}[\langle h_{S^i}, X_i \rangle | X_i] + \sigma_i \mathbb{E}[\langle h_{S^i}, X_i \rangle | X_i] \\
&= \mathbb{E} \sup_{h \in B_r} \frac{1}{n} \sum_{i=1}^{n} \sigma_i (\langle h, X_i \rangle - \mathbb{E}[\langle h_{S^i}, X_i \rangle | X_i]) \\
&= \mathbb{E} \sup_{h \in B_r} \frac{1}{n} \sum_{i=1}^{n} \sigma_i (\langle h, X_i \rangle - \langle \mathbb{E} h_{S^i}, X_i \rangle)
\end{aligned}$$

Borne sur la complexité de Rademacher

$$
\text{i.e.} \quad \mathcal{R}(B_r) = \mathbb{E} \sup_{h \in B_r} \frac{1}{n} \sum_{i=1}^{n} \sigma_i \langle h - \mathbb{E} h_{S^i}, X_i \rangle
$$

$$
\leq \mathbb{E} \sup_{h \in B_r} \frac{1}{n} \| h - \mathbb{E} h_S \| . \left\| \sum_{i=1}^{n} \sigma_i X_i \right\|
$$

$$
\leq \frac{r}{n} \mathbb{E} \left\| \sum_{i=1}^{n} \sigma_i X_i \right\|
$$

$$
\leq \frac{1}{n} \alpha(n) D \sqrt{2n \log(2/\delta)} C_p \left(\sum_{i=1}^{n} \| X_i \|^p \right)^{1/p}
$$

$$
\leq D C_p B \sqrt{2 \log(2/\delta)} \alpha(n) n^{-1/2+1/p},
$$

Borne sur la complexité de Rademacher

i.e. $$\begin{aligned}\mathcal{R}(B_r) &= \mathbb{E}\sup_{h\in B_r}\frac{1}{n}\sum_{i=1}^{n}\sigma_i\langle h-\mathbb{E}h_{S^i}, X_i\rangle \\ &\leq \mathbb{E}\sup_{h\in B_r}\frac{1}{n}\|h-\mathbb{E}h_S\|.\left\|\sum_{i=1}^{n}\sigma_i X_i\right\| \\ &\leq \frac{r}{n}\mathbb{E}\left\|\sum_{i=1}^{n}\sigma_i X_i\right\| \\ &\leq \frac{1}{n}\alpha(n)D\sqrt{2n\log(2/\delta)}C_p\left(\sum_{i=1}^{n}\|X_i\|^p\right)^{1/p} \\ &\leq DC_pB\sqrt{2\log(2/\delta)}\alpha(n)n^{-1/2+1/p},\end{aligned}$$

Borne sur l'erreur de généralisation classique

Corollaire 1, [Liu et al., 2017]

Supposons,

- les hypothèses du théorème 1 vérifiées
- ℓ est majorée par M et $L-$admissible
- h_S est construite par un algorithme $\alpha(n)$-uniformément argument stable.

Alors, avec probabilité au moins $1-2\delta$,

$$R(h_S) - R_S(h_S) \leq 2L\sqrt{2\log(2/\delta)}\alpha(n) + M\sqrt{\frac{\log(1/\delta)}{2n}}$$

.

Borne sur l'erreur de généralisation classique

Lemme 1, Concentration de la prédiction

Supposons que,

- $(\mathfrak{B}, \|.\|)$ est $(2, D)$-smooth,
- $\mathcal{A}$ est un algorithme $\alpha(n)$-argument stable.

Alors pour tout échantillon S et tout $\delta > 0$, avec probabilité au moins $1 - \delta$:

$$\|h_S - \mathbb{E}[h_S]\| \leq D_* \alpha(n) \sqrt{2Bn \log(2\delta^{-1})}.$$

Borne sur l'erreur de généralisation classique

Preuve :

- D'après le lemme 1, avec probabilité au moins $1-\delta$,

$$R(h_S) - R_S(h_S) \leq \sup_{h \in B_r} (R(h) - R_S(h)).$$

- D'autre part,

$$\begin{aligned} \sup_{h \in B_r} (R(h) - R_S(h)) &\leq \mathbb{E} \sup_{h \in B_r} (R(h) - R_S(h)) + M\sqrt{\frac{\log(1/\delta)}{2n}} \\ &\leq 2\mathcal{R}(\ell \circ B_r) + M\sqrt{\frac{\log(1/\delta)}{2n}}. \end{aligned}$$

Borne sur l'erreur de généralisation classique

où $\ell \circ B_r := \{(x, y) \longmapsto \ell(h(x), x, y),\ h \in B_r\}$.

- Or,

$$\begin{aligned}\mathcal{R}(\ell \circ B_r) &\leq L.\mathcal{R}(B_r)\\ &\leq LB\sqrt{2\log(2/\delta)}\alpha(n),\end{aligned}$$

- donc, on a bien

$$R(h_S) - R_S(h_S) \leq 2L\sqrt{2\log(2/\delta)}\alpha(n) + M\sqrt{\frac{\log(1/\delta)}{2n}}$$

Borne sur l'erreur de généralisation déformée

Théorème 3, [Liu et al., 2017]

Supposons,

- $\mathcal{X}$ Hilbert séparable
- $\exists B > 0, \ \|X_i| \ \leq B \ p.s.$
- ℓ est majorée par M et $L-$admissible
- h_S est construite par un algorithme $\alpha(n)$- argument stable.

Alors pour tout $a > 1$, avec probabilité au moins $1 - 2\delta$,

$$R(h_S) - \frac{a}{a-1} R_S(h_S) \leq 8LB\sqrt{2\log(2/\delta)}\alpha(n) + \frac{(6a+8)M\log(1/\delta)}{3n}$$

Cadre du second théorème de concentration

Soient,

- $(X_1, ..., X_n)$ des V.A. de loi $\mathbb{P}$,
- $Z := F(X_1, ..., X_n)$,
- $(Z_1, ...Z_n)$ des V.A. $(X_1, ..., X_n)$-mesurables,
- $(Z_1', ...Z_n')$ des V.A. $(X_1, ..., X_{i-1}, X_{i+1}, .., X_n)$-mesurables respectivement.

Second théorème de concentration, Bousquet 2001

Supposons que :

$$\begin{cases} Z_k' \leq Z - Z_k \leq 1 \text{ p.s.} \\ \sum_{k=1}^n Z - Z_k \leq Z \text{ p.s.} \\ \mathbb{E}_k[Z_k'] \geq 0 \text{ p.s.} \\ Z_k' \leq u \text{ p.s.} \\ \frac{1}{n}\sum_{k=1}^n \mathbb{E}_k[(Z_k')^2] \leq \sigma^2 \text{ p.s.} \end{cases}$$

Alors, pour tout $\lambda \geq 0$:

$$\log \mathbb{E}\left[\exp(\lambda(Z - \mathbb{E}[Z]))\right] \leq v\psi(-\lambda),$$

où $v := n\sigma^2 + (1+u)\mathbb{E}[Z]$.

Corollaire

Dans le cadre du théorème de concentration, on a pour tout $t \geq 0$:

$$\mathbb{P}(Z \geq \mathbb{E}[Z] + t) \leq \exp\left(-vh\left(\frac{t}{v}\right)\right),$$

$$\mathbb{P}\left(Z \geq \mathbb{E}[Z] + \sqrt{2vt} + \frac{t}{3}\right) \leq \exp(-t).$$

Majoration uniforme pour une classe de fonctions

Objectif : majorer

$$\sup_{f \in \mathcal{F}} \left[\mathbb{E}[f(X_i)] - \frac{1}{n} \sum_{k=1}^{n} f(X_i) \right]$$

Majoration uniforme pour une classe de fonctions

Objectif : majorer

$$\sup_{f \in \mathcal{F}} \left[\mathbb{E}[f(X_i)] - \frac{1}{n} \sum_{k=1}^{n} f(X_i) \right]$$

Pour commencer : concentration de

$$Z = \sup_{f \in \mathcal{F}} \sum_{k=1}^{n} f(X_k)$$

Concentration pour une classe de fonctions

Soit $\mathcal{F} \subset \{\mathfrak{X} \to \mathbb{R}\}$ telle que $\forall f \in \mathcal{F}$:

- $(\mathcal{F}, \|.\|_\infty)$ est séparable
- $\forall i \in \{1, \dots, n\}, \mathbb{E}[f(X_i)] = 0$ et $\mathbb{V}(f(X_i)) \leq \sigma^2$
- $\|f\|_\infty \leq c$

Alors $\forall t \geq 0$, avec proba $\leq \exp(-t)$:

$$Z > \mathbb{E}[Z] + \sqrt{2t\left(n\sigma^2 + 2c\mathbb{E}[Z]\right)} + \frac{ct}{3}$$

Suite extraites + cv monotone $\implies$ cas dénombrable
Passage au cas général par densité

Majoration uniforme de l'excès de risque

Soit $\mathcal{F} \subset \{\mathfrak{X} \to [0, M]\}$ telle que $\forall f \in \mathcal{F}$:

- $(\mathcal{F}, \|.\|_\infty)$ est séparable
- $\forall i \in \{1, \dots, n\}$, $\mathbb{V}(f(X_i)) \leq \sigma^2$

Alors, $\forall \delta \in]0, 1]$, avec proba $\geq 1 - \delta$:

$$\sup_{f \in \mathcal{F}} \left[\mathbb{E}[f(X_i)] - \frac{1}{n} \sum_{k=1}^n f(X_i)\right] \leq 4\mathcal{R}_n(\mathcal{F}) + \sqrt{\frac{2\rho \log(\delta^{-1})}{n}} + \frac{4M \log(\delta^{-1})}{3n}$$

Minimisation du risque empirique régularisé

Cadre :

$$\mathcal{H} \subset \mathcal{B} \quad \text{Hilbert séparable}$$

$$\underset{h \in \mathcal{H}}{argmin} \frac{1}{n} \sum_{i=1}^{n} \ell(h, Z_i) + \lambda N(h)$$

Notations :

$$R_r(h) := \frac{1}{n} \sum_{i=1}^{n} \ell(h, z_i) + \lambda N(h)$$

$$R_r^{\setminus j}(h) := \frac{1}{n} \sum_{i \neq j} \ell(h, z_i) + \frac{1}{n} \ell(h, z^{'}) + \lambda N(h)$$

Proposition 4, [Wibisono et al., 2009]

Supposons que,

- ℓ convexe en sa première variable, majorée par $M > 0$ et $L-$admissible
- $\exists\, B > 0, \|X_i\| \leq B$ $p.s.$
- il existe $C > 0$ et $\xi > 1$ tels que,

$$N(h_S) + N(h_{S^i}) - 2N\left(\frac{h_S + h_{S^i}}{2}\right) \geq C\|h_S - h_{S^i}\|_{\ell_2}^{\xi}.$$

Lemme, [Wibisono et al., 2009]

Pour $1 < p \leq 2$, la régularisation avec la norme ℓ_p, vérifie le critère suivant,

$$\|h_S\|_{\ell_p} + \|h_{S^i}\|_{\ell_p} - 2\left\|\frac{h_S + h_{S^i}}{2}\right\|_{\ell_p} \geq C\|h_S - h_{S^i}\|_{\ell_2}^{\xi}.$$

avec $\xi = 2$ et $C = \frac{1}{4}p(p-1)\left(\frac{B}{\lambda}\right)^{\frac{p-2}{p}}$.

Proposition 4, [Wibisono et al., 2009]

Alors,

- l'algorithme ERM régularisé est $\beta(n)$-uniformément stable avec : $\beta(n) = \left(\frac{B^\xi L^\xi}{C\lambda n}\right)^{\frac{1}{\xi-1}}$
- et même $\alpha(n)-$ uniformément argument stable avec $\alpha(n) = \left(\frac{BL}{C\lambda n}\right)^{\frac{1}{\xi-1}}$.

Preuve :

Lemme, [Bousquet and Elisseeff, 2002]

Supposons que,

- ℓ est majorée par $M > 0$
- N est convexe
- R_r et $R_r^{\setminus j}$ admettent des minima.

Alors,

- $\exists\ \tau(\lambda) = \frac{M}{\lambda} + N(0) \geq 0$ tel que $N(h_S) \leq \tau(\lambda)$
- $N(h) - N(h + t\Delta h) + N(h^{\setminus j} - t\Delta h) - N(h^{\setminus j}) \leq \frac{2tL}{\lambda n}|\Delta h(x_j)|$

Preuve :

En se restreignant à $\mathrm{B}(0,B) \subset \mathcal{X}$, puisque p.s. $\|X_i\| \leq B$:

$$\|h_S - h_{S^j}\|_\infty = \sup_{x \in \mathrm{B}(0,B)} |h_S(x) - h_{S^j}(x)| \leq B\|h_S - h_{S^j}\|_{\ell_2}.$$

d'où, par le lemme précédent avec $t = 1/2$:

$$N(h_S) + N(h_{S^j}) - 2N\left(\frac{h_S + h_{S^j}}{2}\right) \leq \frac{L}{n\lambda}\|h_S - h_{S^j}\|_\infty$$

En combinant ceci avec l'hypothèse **??** sur N, on obtient :

$$\begin{aligned} \|h_S - h_{S^j}\|_{\ell_2}^{\xi} &\leq \frac{L}{n\lambda C}\|h_S - h_{S^j}\|_\infty \\ &\leq \frac{LB}{n\lambda C}\|h_S - h_{S^j}\|_{\ell_2}. \end{aligned}$$

Preuve :

D'où

$$\|h_S - h_{S^j}\|_{\ell_2} \leq \left(\frac{LB}{n\lambda C}\right)^{\frac{1}{\xi-1}}$$

de plus, puisque ℓ est supposée $L-$admissible, on déduit aussi que :

$$\forall i \in \{1, \cdots, n\}, \quad |\ell(h - S, Z_i) - \ell(h_{S^j}, Z)| \begin{aligned}[t] &\leq L\|h_S - h_{S^j}\|_\infty \\ &\leq LB\|h_S - h_{S^j}\|_{\ell_2}\end{aligned}$$

d'où,

$$\forall i \in \{1, \cdots, n\}, \quad |\ell(h - S, Z_i) - \ell(h_{S^j}, Z)| \leq \left(\frac{L^\xi B^\xi}{n\lambda C}\right)^{\frac{1}{\xi-1}}$$

Résultat

Théorème 4, [Liu et al., 2017]

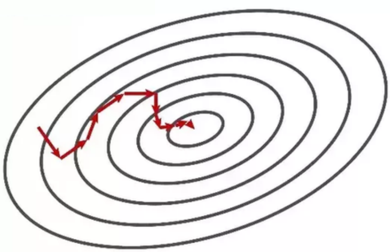
Avec les hypothèses de la proposition précédente, pour tout $\delta > 0$, et $a > 1$, on a

$$R(h_S) - \frac{a}{a-1} R_S(h_S) \leq 8LB \left(\frac{LB}{C\lambda n}\right)^{\frac{1}{\xi - 1}} \sqrt{2\log(2/\delta)} + \frac{(6a+8)M\log(1/\delta)}{3n}$$

En particulier, lorsque $N(h) = \|h\|^2$, la condition suffisante précédemment donnée est vérifiée pour $\xi = 2$ et $C = \frac{1}{2}\left(\frac{M}{\lambda}\right)^{\frac{1}{2}}$.

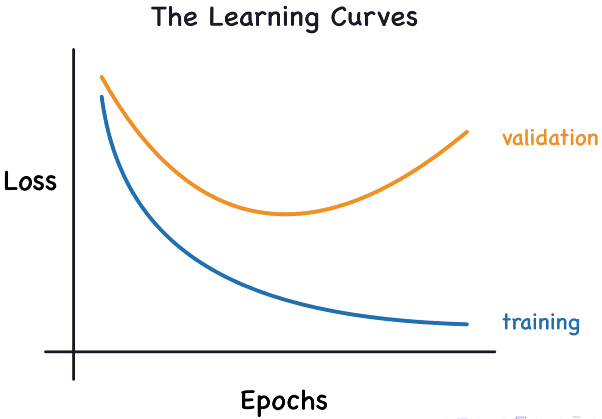
Definition

descente de gradient stochastique
$h_{t+1} = h_t - \alpha_t \nabla_h \ell(h_t, Z_{i_t})$



$$\mathbb{E}_{i_t}[\nabla_h \ell(h_t, Z_{i_t})] = \nabla_h R_S(h)$$

Sur-apprentissage, Compromis biais-variance



Definition

(Smooth)
Une fonction de perte différentiable l est

- *s-smooth* si son gradient est s-Lipschitz :

$$\forall h, h' \in H, \|\nabla_h \ell(h, \cdot) - \nabla_{h'} \ell(h', \cdot)\| \leq s\|h - h'\|$$

- *γ-fortement convexe*, avec $\gamma > 0$, si :

$$(\nabla_h \ell(h, \cdot) - \nabla_{h'} \ell(h', \cdot))^T (h - h') \geq \gamma \|h - h'\|^2$$

Definition

(Smooth)
la *règle de mise à jour du gradient* à la t^{ieme} itération, $G_t(h) = h - \alpha \nabla l(h, Z_{i_t})$, est

- *η-expansive* si :

$$\sup_{h,h' \in \mathcal{H}} \frac{\|G(h) - G(h')\|}{\|h - h'\|} \leq \eta$$

- σ-bornée si :

$$\sup_{h \in \mathcal{H}} \|h - G(h)\| \leq \sigma$$

Évolution de l'écart maximal entre hypothèses

Deux SGD $\perp$: $h_{t+1} = G_t(h_t)$ et $h'_{t+1} = G_t(h_t)'$
$\delta_t = \|h'_t - h_t\|$

- Si $G_t = G'_t$, et est η-expansive, alors :

$$\delta_{t+1} \leq \eta \delta_t$$

- Si G_t et G'_t sont σ-bornées, et que G_t est η-expansive, alors :

$$\delta_{t+1} \leq \min(\eta, 1)\delta_t + 2\sigma_t$$

Expansivité pour une perte régulière

Si f est s-smooth, alors :

1. **Cas général**
   $G_{f,\alpha}$ est $(1+\alpha s)$-expansive.
2. **Cas convexe**
   Si de plus, f est convexe,
   alors pour tout $\alpha \leq 2/s$, alors $G_{f,\alpha}$ est 1-expansive.
3. **Cas strictement convexe**
   Si de plus, f est γ-fortement convexe,
   alors pour tout $\alpha \leq \frac{2}{s+\gamma}$, $G_{f,\alpha}$ est $\left(1 - \frac{\alpha s \gamma}{s+\gamma}\right)$- expansive.

Expansivité pour une perte régulière

Lemme de Baillon-Haddad

Si f est convexe et s-smooth,
alors $\langle \nabla f(h') - \nabla f(h), h - h' \rangle \geq \frac{1}{s} \|\nabla f(h') - \nabla f(h)\|^2$

Expansivité pour une perte régulière

Lemme de Baillon-Haddad

Si f est convexe et s-smooth,
alors $\langle \nabla f(h') - \nabla f(h), h - h' \rangle \geq \frac{1}{s} \|\nabla f(h') - \nabla f(h)\|^2$

$$\|G_{f,\alpha}(h) - G_{f,\alpha}(h')\|^2$$

Expansivité pour une perte régulière

Lemme de Baillon-Haddad

Si f est convexe et s-smooth,
alors $\langle \nabla f(h') - \nabla f(h), h - h' \rangle \geq \frac{1}{s} \|\nabla f(h') - \nabla f(h)\|^2$

$$\|G_{f,\alpha}(h) - G_{f,\alpha}(h')\|^2$$
$$= \|h - h'\|^2 - 2\alpha\langle \nabla f(h) - \nabla f(h'), h - h' \rangle + \alpha^2 \|\nabla f(h) - \nabla f(h')\|^2$$

Expansivité pour une perte régulière

Lemme de Baillon-Haddad

Si f est convexe et s-smooth,
alors $\langle \nabla f(h') - \nabla f(h), h - h' \rangle \geq \frac{1}{s} \|\nabla f(h') - \nabla f(h)\|^2$

$$\begin{aligned}
&\|G_{f,\alpha}(h) - G_{f,\alpha}(h')\|^2 \\
&= \|h - h'\|^2 - 2\alpha \langle \nabla f(h) - \nabla f(h'), h - h' \rangle + \alpha^2 \|\nabla f(h) - \nabla f(h')\|^2 \\
&\leq \|h - h'\|^2 - \left(\frac{2\alpha}{s} - \alpha^2\right) \|\nabla f(h) - \nabla f(h')\|^2
\end{aligned}$$

Expansivité pour une perte régulière

Lemme de Baillon-Haddad

Si f est convexe et s-smooth,
alors $\langle \nabla f(h') - \nabla f(h), h - h' \rangle \geq \frac{1}{s} \|\nabla f(h') - \nabla f(h)\|^2$

$$
\begin{aligned}
&\|G_{f,\alpha}(h) - G_{f,\alpha}(h')\|^2 \\
&= \|h - h'\|^2 - 2\alpha \langle \nabla f(h) - \nabla f(h'), h - h' \rangle + \alpha^2 \|\nabla f(h) - \nabla f(h')\|^2 \\
&\leq \|h - h'\|^2 - \left(\frac{2\alpha}{s} - \alpha^2\right) \|\nabla f(h) - \nabla f(h')\|^2 \\
&\leq \|h - h'\|^2
\end{aligned}
$$

Expansivité pour une perte régulière

Argument stability des algorithmes entraînés par SGD

Supposons que l est s-Smooth et L-Admissible

1. **Cas général**
   Si $\alpha_t \leq c/t$ et $\mathcal{H}$ est borné par B :

$$\mathbb{E}[\|h_T - h'_T\||S] \leq (c + \frac{1}{s})\frac{2LT^{\frac{sc}{sc+1}}}{n+1}$$

Argument stability des algorithmes entraînés par SGD

Supposons que l est s-Smooth et L-Admissible

1. **Cas général**
   Si $\alpha_t \leq c/t$ et $\mathcal{H}$ est borné par B :

$$\mathbb{E}[\|h_T - h'_T\||S] \leq (c + \frac{1}{s})\frac{2LT^{\frac{sc}{sc+1}}}{n+1}$$

- Réseaux de neurones

Argument stability des algorithmes entraînés par SGD

Supposons que l est s-Smooth et L-Admissible

1. **Cas général**
   Si $\alpha_t \leq c/t$ et $\mathcal{H}$ est borné par B :

$$\mathbb{E}[\|h_T - h'_T\||S] \leq (c + \frac{1}{s})\frac{2LT^{\frac{sc}{sc+1}}}{n+1}$$

- Réseaux de neurones
- $\neq$ article d'origine

Argument stability des algorithmes entraînés par SGD

2. **Cas convexe**
   Si l est convexe, et $\alpha_t \leq 2/s$:

$$\mathbb{E}[\|h_T - h'_T\||S] \leq \frac{2BL}{n}\sum_{t=1}^{T}\alpha_t$$

Argument stability des algorithmes entraînés par SGD

2 **Cas convexe**
Si l est convexe, et $\alpha_t \leq 2/s$:

$$\mathbb{E}[\|h_T - h'_T\||S] \leq \frac{2BL}{n} \sum_{t=1}^{T} \alpha_t$$

- Avec proba $1 - 1/n$, même exemple $G_t = G'_t$ + lemme $\implies$ Maj 1-expansive.

Argument stability des algorithmes entraînés par SGD

2 **Cas convexe**

Si l est convexe, et $\alpha_t \leq 2/s$:

$$\mathbb{E}[\|h_T - h'_T\||S] \leq \frac{2BL}{n} \sum_{t=1}^{T} \alpha_t$$

- Avec proba $1 - 1/n$, même exemple $G_t = G'_t$ + lemme $\implies$ Maj 1-expansive.
- Avec proba $1/n$, exemple différent.

Argument stability des algorithmes entraînés par SGD

2 **Cas convexe**
Si l est convexe, et $\alpha_t \leq 2/s$:

$$\mathbb{E}[\|h_T - h'_T\| | S] \leq \frac{2BL}{n} \sum_{t=1}^{T} \alpha_t$$

- Avec proba $1 - 1/n$, même exemple
  $G_t = G'_t$ + lemme $\implies$ Maj 1-expansive.
- Avec proba $1/n$, exemple différent.
  G_t et G'_t sont $(\alpha_t L)$-bornés : $\|h - G_{f,\alpha}(h)\| = \|\alpha \nabla f(h)\| \leq \alpha L$
  lemme $\implies \delta_t \leq \delta_t + 2\alpha_t L$

Ainsi, par linéarité de l'espérance, on a $\forall t \geq t_0$,

$$\begin{aligned}\mathbb{E}[\delta_{t+1}|S] &\leq \left(1 - \frac{1}{n}\right)\mathbb{E}[\delta_t|S] + \frac{1}{n}\mathbb{E}[\delta_t|S] + \frac{2\alpha_t L}{n} \\ &= \mathbb{E}[\delta_t|S] + \frac{2\alpha_t L}{n}\end{aligned}$$

Ainsi, par linéarité de l'espérance, on a $\forall t \geq t_0$,

$$\begin{aligned}\mathbb{E}[\delta_{t+1}|S] &\leq \left(1 - \frac{1}{n}\right)\mathbb{E}[\delta_t|S] + \frac{1}{n}\mathbb{E}[\delta_t|S] + \frac{2\alpha_t L}{n} \\ &= \mathbb{E}[\delta_t|S] + \frac{2\alpha_t L}{n}\end{aligned}$$

$\delta_0 = 0$, par téléscopage :

$$\mathbb{E}[\delta_T|S] = \sum_{t=1}^{T} \mathbb{E}[\delta_t|S] \leq \frac{2L}{n}\sum_{t=1}^{T} \alpha_t$$

Argument stability des algorithmes entraînés par SGD

3 **Cas strictement convexe** avec projection sur un convexe compacte
Pour $h_{t+1} = \Pi_\Omega(h_t - \alpha_t \nabla_h \ell(h_t, Z_{i_t}))$
Si l est γ-fortement convexe, et $\alpha_t \leq 1/s$:

$$\mathbb{E}[\|h_T - h'_T\||S] \leq \frac{2BL}{\gamma n}$$

Argument stability des algorithmes entraînés par SGD

3 **Cas strictement convexe** avec projection sur un convexe compacte
Pour $h_{t+1} = \Pi_\Omega(h_t - \alpha_t \nabla_h \ell(h_t, Z_{i_t}))$
Si l est γ-fortement convexe, et $\alpha_t \leq 1/s$:

$$\mathbb{E}[\|h_T - h'_T\||S] \leq \frac{2BL}{\gamma n}$$

- Ne dépend plus de T

Borne probabiliste de *l'erreur de généralisation déformée* des algorithmes entraînés par SGD

On suppose que

- nos données sont bornées : $||X|| \leq B$ p.s.
- La perte l est s-Smooth, L-Admissible et bornée par M.

Borne probabiliste de *l'erreur de généralisation déformée* des algorithmes entraînés par SGD

On suppose que

- nos données sont bornées : $||X|| \leq B$ p.s.
- La perte l est s-Smooth, L-Admissible et bornée par M.

1. **Cas général :**
   Si $\alpha_t \leq c/t$ Alors $\forall \delta > 0, \forall a > 1$, avec proba $\geq 1 - 2\delta$:
   $R(h_T) - \frac{a}{a-1} R_S(h_T) \leq$

$$16(c + \frac{1}{s}) \frac{BLT^{1+\frac{sc}{sc+1}}}{n+1} \sqrt{2\ln(2/\delta)} + \frac{(6a+8)M\ln(1/\delta)}{3n}$$

Borne probabiliste de *l'erreur de généralisation déformée* des algorithmes entraînés par SGD

2 **Cas convexe**
Si l est convexe et $\alpha_t \leq 2/s$
Alors $\forall \delta > 0, \forall a > 1$, avec proba $\geq 1 - 2\delta$:
$R(h_T) - \frac{a}{a-1} R_S(h_T) \leq$

$$\frac{16B^2L^2}{n} \sum_{t=1}^{T} \alpha_t \sqrt{2\ln(2/\delta)} + \frac{(6a+8)M\ln(1/\delta)}{3n}$$

Borne probabiliste de *l'erreur de généralisation déformée* des algorithmes entraînés par SGD

3 **Cas strictement convexe avec projection sur un convexe compact**

$h_{t+1} = \Pi_\Omega(h_t - \alpha_t \nabla_h \ell(h_t, Z_{i_t}))$

Si l est γ-fortement convexe et $\alpha_t \leq 1/s$ Alors

$\forall \delta > 0, \forall a > 1$, avec proba $\geq 1 - 2\delta$:

$R(h_T) - \frac{a}{a-1} R_S(h_T) \leq$

$$\frac{16B^2L^2}{\gamma n} \sum_{t=1}^{T} \alpha_t \sqrt{2\ln(2/\delta)} + \frac{(6a+8)M\ln(1/\delta)}{3n}$$

Algorithme du gradient proximal : problème

Problème : Minimiser $F = f + g$ lorsque :

- $f, g \in \Gamma_0(\mathcal{H}) :=$ $\{$fcts convexes, semi-continues inférieurements et propres$\}$.
- f est différentiable,
- F est coercive (i.e $F(h) \longrightarrow_{\|h\| \longrightarrow \infty} = \infty$).

Algorithme du gradient proximal : définition

Opérateur proximal de ϕ : L'unique point de $\mathcal{H}$ tel que, pour $\mu > 0$ fixé :

$$\inf_{h \in \mathcal{H}} \left(\phi(h) + \frac{\|y-h\|^2}{2\mu} \right) = f(\mathrm{prox}_{\mu\phi}(y)) + \frac{\|y - \mathrm{prox}_{\mu\phi}(y)\|^2}{2\mu}.$$

Minimum de F :

$$h \text{ est un minimum de } F \Leftrightarrow h = \mathrm{prox}_{\mu g}(h - \mu \nabla f(h)).$$

Algorithme de minimisation de F :

$$h_{T+1} = \mathrm{prox}_{\mu_T g}(h_T - \mu_T \nabla f(h_T)).$$

Algorithme du gradient proximal : convergence

Convergence de l'algorithme

Si

- $f, g \in \Gamma_0(\mathcal{H})$,
- f différentiable et s-smooth,
- $F = f + g$ coercive,
- $h_0 \in \mathcal{H}$, $\beta \in]0,1[$ et μ_0 tel que $\mu_0 s \geq 1$.

Alors :

$$\forall T \geq 0,\ F(h_T) - F(h_*) \leq \frac{s}{2\beta T}\|h_0 - h_*\|^2.$$

Application à la RERM : définition du problème

On voudrait minimiser :

$$R_{S,\lambda}(h) = \frac{1}{n}\sum_{i=1}^{n} \ell(h, X_i) + \lambda N(h).$$

Sous les hypothèses suivantes :

- Les données sont presque sûrement bornées par $B > 0$,
- $N \in \Gamma_0(\mathcal{H})$,
- $\ell \in \Gamma_0(\mathcal{H})$, ℓ est L-admissible et bornée par $M > 0$,
- $R_{S,\lambda}$ est coercive au sens de la norme $\|.\|$.

RERM : pénalité différentiable

Hypothèses initiales :

- Les données sont presque sûrement bornées par $B > 0$,
- $N \in \Gamma_0(\mathcal{H})$,
- $\ell \in \Gamma_0(\mathcal{H})$, ℓ est L-admissible et bornée par $M > 0$,
- $R_{S,\lambda}$ est coercive au sens de la norme $\|.\|$.

Hypothèses supplémentaires :

- N est différentiable,
- N est s-smooth,
- N est K-Lipschitz.

Définition de l'algorithme

Pour tout échantillon S et tout indice T :

$$h_{S,T+1} = \mathrm{prox}_{\mu_T \ell_{i_T}}(h_{S,T} - \mu_T \nabla N(h_{S,T})),$$

où

- $\ell_i : h \longmapsto l(h, X_i)$,
- i_T est un indice choisit uniformément aléatoirement dans $\{1, ..., n\}$.

Convergence en moyenne de l'algorithme

Proposition

Sous les hypothèses précédentes, soient :

- $h_{S,0} \in \mathcal{H}$,
- $\beta \in]0, 1[$,
- μ_0 tel que $\mu_0 \geq 1/s$.

Alors :

$$\forall T \geq 0,\ \mathbb{E}[R_{S,\lambda}(h_{S,T}) - R_{S,\lambda}(h_*)] \leq \frac{s}{2\beta T}\|h_{S,0} - h_*\|^2.$$

Stabilité de l'algorithme

Proposition

Sous les hypothèses précédente et si $\mu_0 \leq 2/s$, alors pour tout $T \geq 0$:

$$\mathbb{E}\left[\|h_{S,T} - h_{S^i,T}\| | S\right] \leq \frac{2\sqrt{KL}}{n} \sum_{t=0}^{T} \mu_T \text{ p.s.}$$

Stabilité de l'algorithme : Preuve

Notons $G_{S,T}$ la fonction de mise à jour du gradient, alors :

$$h_{S,T+1} = \text{prox}_{\mu_T \ell_{i_T}}(G_{S,T}(h_{S,T})).$$

Sur l'évènement $\{G_{S,T} = G_{S^i,T}\}$:

$$\begin{aligned}
\delta_{T+1} &= \|h_{S,T+1} - h_{S^i,T+1}\| \\
&\leq \|G_{S,T}(h_{S,T}) - G_{S,T}(h_{S^i,T})\| \text{ par concentration fermée} \\
&\leq \delta_T \text{ car } G_{S,T} \text{ est 1-expansive.}
\end{aligned}$$

Stabilité de l'algorithme : Preuve

Sur l'évènement $\{G_{S,T} \neq G_{S^i,T}\}$:

$$\begin{aligned}
\delta_{T+1} &\leq \|h_{S,T+1} - G_{S,T}(h_{S,T})\| + \|G_{S,T}(h_{S,T}) - G_{S,T}(h_{S^i,T})\| \\
&+ \|G_{S,T}(h_{S^i,T}) - h_{S^i,T+1}\| \\
&\leq \delta_T + \|\mathrm{prox}_{\mu_T \ell_{i_T}}(G_{S,T}(h_{S,T})) - G_{S,T}(h_{S,T})\| \\
&+ \|\mathrm{prox}_{\mu_T \ell_{i_T}}(G_{S^i,T}(h_{S^i,T})) - G_{S^i,T}(h_{S^i,T})\| \\
&\leq \delta_T + \sqrt{\mu_T \left|\ell_{i_T}(h_{S,T+1}) - \ell_{i_T}(G_{S,T}(h_{S,T}))\right|} \\
&+ \sqrt{\mu_T \left|\ell^i_{i_T}(h_{S^i,T+1}) - \ell^i_{i_T}(G_{S^i,T}(h_{S^i,T}))\right|}
\end{aligned}$$

Stabilité de l'algorithme : Preuve

Sur l'évènement $\{G_{S,T} \neq G_{S^i,T}\}$:

$$\begin{aligned}
\delta_{T+1} \leq \delta_T &+ \sqrt{\mu_T \left|\ell_{i_T}(h_{S,T+1}) - \ell_{i_T}(G_{S,T}(h_{S,T}))\right|} \\
&+ \sqrt{\mu_T \left|\ell^i_{i_T}(h_{S^i,T+1}) - \ell^i_{i_T}(G_{S,T}(h_{S^i,T}))\right|} \\
&\leq \delta_T + \sqrt{\mu_T} L \left|h_{S,T+1} - G_{S,T}(h_{S,T})\right| \\
&+ \sqrt{\mu_T} L \left|h_{S^i,T+1} - G_{S^i,T}(h_{S^i,T})\right| \\
&\leq \delta_T + L\sqrt{K}\mu_T.
\end{aligned}$$

Stabilité de l'algorithme : Preuve

Donc

$$\delta_{T+1} \leq \delta_T + 2\mu_T L \mathbb{1}_{\{G_{S,T} \neq G_{S^i,T}\}} \text{ p.s..}$$

Or,

$$\mathbb{P}(G_{S,T} \neq G_{S^i,T} | S) = \frac{1}{n} \text{ p.s.}$$

Donc,

$$\mathbb{E}[\delta_{T+1}|S] \leq \mathbb{E}[\delta_T|S] + \frac{2\mu_T L}{n} \text{ p.s.}$$

Stabilité de l'algorithme

Proposition

Sous les hypothèses précédente et si $\mu_0 \leq 2/s$, alors pour tout $T \geq 0$:

$$\mathbb{E}\left[\|h_{S,T} - h_{S^i,T}\| | S\right] \leq \frac{2\sqrt{KL}}{n} \sum_{t=0}^{T} \mu_T \text{ p.s.}$$

Borne sur l'erreur de généralisation déformée

Théorème

Si $\mu_0 \leq 2/s$, alors pour tout $\delta > 0$, tout $a > 0$ et tout $T \geq 0$ avec probabilité au moins $1 - 2\delta$:

$$R(h_{S,T}) - \frac{a}{a-1} R_S(h_{S,T}) \leq 16LB\sqrt{2KL\log(2/\delta)}\frac{\sum_{t=0}^{T}\mu_T}{n} + \frac{(6a+8)M\log(1/\delta)}{3n}.$$

Remarques

Terme majorant

$$16LB\sqrt{2KL\log(2/\delta)}\frac{\sum_{t=0}^{T}\mu_T}{n}+\frac{(6a+8)M\log(1/\delta)}{3n}.$$

- Par construction $\mu_T \leq \beta\mu_T \leq \beta^T\mu_0$ p.s., donc

$$\sum_{t=0}^{T}\mu_T \leq \frac{\mu_0}{1-\beta} \text{ p.s.}$$

- pour avoir convergence de l'algorithme il faut avoir $1/s \leq \mu_0$,
- pour avoir stabilité de l'algorithme il faut avoir $\mu_0 \leq 2/s$.

Intérêts de l'algorithme

- Plus besoin d'avoir la condition :

$$N(h_S) + N(h_{S^i}) - 2N\left(\frac{h_S + h_{S^i}}{2}\right) \geq C\|h_S - h_{S^i}\|_{\ell_2}^{\xi}.$$

- Si la perte n'est pas différentiable : on ne peut pas appliquer la descente de gradient stochastique.
- Mais il faut N smooth et Lipschitz.

Notions de stabilité

Définition, [Shalev-Shwartz et al., 2010]

Soit $\epsilon : \mathbb{N} \longrightarrow \mathbb{R}$ une fonction décroissante. On dit qu'un algorithme $\mathcal{A}$ est On-Average-Replace-One ϵ-stable si, pour tout $n \in \mathbb{N}$

$$\mathop{\mathbb{E}}_{i \sim Unif(1,n)} [\ell(h_{S^i}, Z_i) - \ell(h_S, Z_i)] \leq \epsilon(n).$$

Remarque :

Cette notion de stabilité est clairement plus faible que celles vues précédemment.

Notions de stabilité, stabilité et généralisation

Théorème 9, [Shalev-Shwartz et al., 2010]

Pour tout algorithme $\mathcal{A}$, on a

$$\mathbb{E}_S\left[\mathbb{E}_Z[\ell(h_S, Z)] - \frac{1}{n}\sum_{i=1}^{n}\ell(h_S, Z_i)\right] = \underset{i\sim Unif(1,n)}{\mathbb{E}}[\ell(h_{S^i}, Z_i) - \ell(h_S, Z_i)]$$

Notions de stabilité

Théorème 10, [Shalev-Shwartz et al., 2010]

Supposons que,

- la fonction de perte ℓ est convexe et L-admissible.

Alors, l'ERM régularisé est On-Average-Replace-One stable de rapport $\frac{2L^2}{\lambda n}$. Il s'ensuit que,

$$\mathbb{E}_S\left[\mathbb{E}_Z[\ell(h_S, Z)] - \frac{1}{n}\sum_{i=1}^{n}\ell(h_S, Z_i)\right] \leq \frac{2L^2}{\lambda n}$$

Bornes étudiées

Corollaire, [Shalev-Shwartz et al., 2010]

Supposons que,

- ℓ est convexe en sa première variable et $L-$admissible
- $\mathcal{H} \subset \mathbb{R}^d$ est convexe et bornée par $B > 0$

Alors,

$$\mathbb{E}[\ell(h_S, Z)] \leq \inf_{h \in \mathcal{H}} \mathbb{E}[\ell(h, Z)] + LB\sqrt{\frac{8}{n}}$$

Bornes étudiées

Théorème, [Bousquet and Elisseeff, 2002]

Supposons que,

- ℓ est majorée par $M > 0$
- $\beta(n)$ est le coefficient d'uniforme stabilité de l'algorithme construisant h_S.

Alors,

$$R(h_S) \leq R_S(h_S) + 2\beta(n) + (4n\beta(n) + M)\sqrt{\frac{\log(1/\delta)}{2n}}$$

Techniques utilisées

- 1^{ere} Inégalité de concentration, [Pinelis, 1994]
- Majoration de la complexité de Rademacher, [Liu et al., 2017]
- Borne de généralisation classique, [Liu et al., 2017]

Techniques utilisées

- Borne sur l'erreur de généralisation déformée, [Bartlett et al., 2005]
- Point fixe de ψ vérifiant une inégalité de la forme :

$$\psi(r) \geq \frac{c_1}{n} + c_2 \mathbb{E}_\sigma \sup_{h \in \mathcal{F}_n} \frac{1}{n} \sum_{i=1}^{n} \sigma_i h(X_i)$$

Techniques utilisées

- 2de Inégalité de concentration, [Bousquet, 2001]
- Basée sur des méthodes d'entropie
- Nouvelle technique pour une telle borne, [Liu et al., 2017]

Perspectives

- Etude de la stabilité de l'algorithme SGD avec réduction de la variance
- Exploration d'autre propriétés algorithmique que la stabilité résultant en une classe algorithmique d'hypothèses petite, [Liu et al., 2017]

Bartlett, P. L., Bousquet, O., Mendelson, S., et al. (2005).
Local rademacher complexities.
The Annals of Statistics, 33(4) :1497–1537.

Bousquet, O. (2001).
A bennett concentration inequality and its application to suprema of empirical processes.
C.R. Acad. Sci. Paris, 332.

Bousquet, O. and Elisseeff, A. (2002).
Stability and generalization.
The Journal of Machine Learning Research, 2 :499–526.

Liu, T., Lugosi, G., Neu, G., and Tao, D. (2017).
Algorithmic stability and hypothesis complexity.
In *International Conference on Machine Learning*, pages 2159–2167. PMLR.

Pinelis, I. (1994).
Optimum bounds for the distributions of martingales in banach spaces.
The Annals of Probability, 22(4) :1679–1706.

Shalev-Shwartz, S., Shamir, O., Srebro, N., and Sridharan, K. (2010).
Learnability, stability and uniform convergence.
The Journal of Machine Learning Research, 11 :2635–2670.

Wibisono, A., Rosasco, L., and Poggio, T. (2009).
Sufficient conditions for uniform stability of regularization algorithms.
Computer Science and Artificial Intelligence Laboratory Technical Report, MIT-CSAIL-TR-2009-060.