Regression à noyau, Neural Tangent Kernel et Overfitting bénin

Eddie Aamari
28/09/2023

I Régression et espaces à noyau reproduisant

1) Régression linéaire

Données $(X_i, Y_i) \in \mathcal{X} \times \mathcal{Y} = \mathbb{R}^d \times \mathbb{R}$

On cherche à trouver un minimiseur pour

$$\min_{\beta \in \mathbb{R}^d} \sum_{i=1}^{m} (Y_i - \langle X_i, \beta \rangle)^2 = \min_{\beta \in \mathbb{R}^d} \|Y - \mathbb{X}\beta\|_2^2$$

où $\left\{ \begin{array}{l} \mathbb{X} = \begin{pmatrix} X_1^T \\ \vdots \\ X_m^T \end{pmatrix} \in \mathbb{R}^{m \times d} \\ Y = \begin{pmatrix} Y_1 \\ \vdots \\ Y_m \end{pmatrix} \end{array} \right.$

Points critiques $\Leftrightarrow \mathbb{X}^T(\mathbb{X}\hat{\beta} - Y) = 0$

$\Leftrightarrow \mathbb{X}^T\mathbb{X}\hat{\beta} = \mathbb{X}^T Y$

$(m \gg d)$ • Si $\mathbb{X}^T\mathbb{X}$ inversible, $\hat{\beta} = (\mathbb{X}^T\mathbb{X})^{-1}\mathbb{X}^T Y$

$(m \ll d)$ • Sinon, infinité de solutions à $Y = \mathbb{X}\beta$

↳ Solution de norme minimale $\hat{\beta} = \underset{\mathbb{X}\beta = Y}{\operatorname{argmin}} \dfrac{\|\beta\|_2^2}{2}$

$$= \mathbb{X}^T \underbrace{(\mathbb{X}\mathbb{X}^T)^{-1} Y}_{\hat{\alpha} \in \mathbb{R}^n} = \sum_{i=1}^{m} \hat{\alpha}_i X_i$$

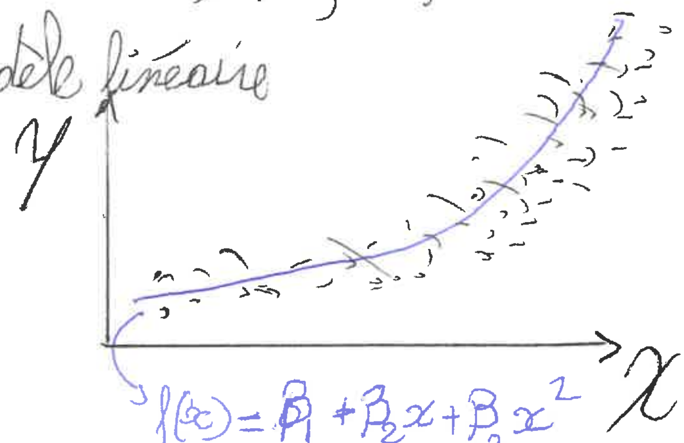
$\Rightarrow \hat{\beta}$ combinaison linéaire des $(X_i)_i$

Rq: $\mathbb{X}\mathbb{X}^T = (\langle X_i, X_j \rangle)_{i,j \leq m}$

matrice de Gram

2) Machines à noyau

Au lieu de travailler directement avec les $(X_i)_{i\leq n}$, on peut d'abord les transformer, puis considérer le modèle linéaire



$$f(x) = \beta_1 + \beta_2 x + \beta_3 x^2 = \left\langle \begin{pmatrix} \beta_1 \\ \beta_2 \\ \beta_3 \end{pmatrix}, \begin{pmatrix} 1 \\ x \\ x^2 \end{pmatrix} \right\rangle$$

linéaire en $\Phi(x) := \begin{pmatrix} 1 \\ x \\ x^2 \end{pmatrix} \in \mathcal{X}' = \mathbb{R}^3$

Idée : Envoyer les $(X_i)_i$ dans un espace de Hilbert $\mathcal{H}$ via

$$\begin{aligned} \Phi : \mathcal{X} &\longrightarrow \mathcal{H} \\ X_i &\longmapsto \Phi(X_i) \end{aligned}$$

Feature space

Feature map

Feature

Travailler dans le nouvel espace $\mathcal{H}$ et résoudre le problème linéaire associé semble nécessiter de calculer $\left(\langle \Phi(X_i), \Phi(X_j) \rangle_{\mathcal{H}}\right)_{i,j \leq n}$ via le calcul des $\Phi(X_i)$ et des $\langle \cdot, \cdot \rangle_{\mathcal{H}}$.

Astuce du noyau (Kernel trick)

Il est parfois plus facile de calculer le noyau

$$\begin{aligned} k : \mathcal{X} \times \mathcal{X} &\longrightarrow \mathbb{R} \\ (x_i, x_j) &\longmapsto \langle \Phi(X_i), \Phi(X_j) \rangle_{\mathcal{H}} \end{aligned}$$

que de calculer Φ, qui vit possiblement en dimension infinie.

<u>Inversement</u> : On va voir que se donner un "noyau" $k : \mathcal{X}\times\mathcal{X} \to \mathbb{R}$ est équivalent à se donner Φ.

<u>Def</u> : (Noyau) $k : \mathcal{X}\times\mathcal{X} \to \mathbb{R}$ est appelée <u>noyau</u> lorsque

- k est symétrique, continue
- k est positive :

$$\forall n \geq 1, \begin{cases} \forall X_1, \dots, X_n \in \mathcal{X} \\ \forall a_1, \dots, a_n \in \mathbb{R} \end{cases}, \quad \sum_{i=1}^{n} \sum_{\ell=1}^{n} a_i a_k k(X_i, X_\ell) \geq 0$$

On appelle <u>matrice de Gram</u> $\boxed{K := \left(k(X_i, X_j) \right)_{i,j \leq n} \in \mathbb{R}^{n \times n}}$

<u>Rq</u> : • la définition de noyau est exactement la contrainte $K \succcurlyeq 0$

• $K(X_i, X_j) := g(\|X_i - X_j\|)$ est un noyau $\overset{\text{Bochner}}{\underset{\cancel{\text{Moore}}}{\iff}} \hat{g} \geq 0$

<u>Thm</u> : (Moore-Aronszajn)

Pour tout noyau $k : \mathcal{X}\times\mathcal{X} \to \mathbb{R}$, il existe
- un hilbert $\mathcal{H} \subset \mathcal{C}(\mathcal{X}, \mathbb{R})$
- $\Phi : \mathcal{X} \to \mathcal{H}$

tels que $\forall x, x' \in \mathcal{X}$, $k(x, x') = \langle \Phi(x), \Phi(x') \rangle_{\mathcal{H}}$

De plus, on a la propriété de noyau reproduisant

$$\forall f \in \mathcal{H}, \quad f(x) = \langle f, k(x, \cdot) \rangle_{\mathcal{H}}$$

Dem : On pose $\mathcal{H} := \overline{\left\{ \sum_{j=1}^{N} a_j K(x_i, \cdot) \;\middle|\; \begin{array}{l} N \geq 1 \\ a_j \in \mathbb{R} \\ x_i \in \mathcal{X} \end{array} \right\}}$

[où la clôture est prise pour $\left\langle \sum a_j k(x_j, \cdot), \sum_j b_j k(y_j, \cdot) \right\rangle_{\mathcal{H}}$

$$:= \sum_{j,\ell} a_j b_\ell k(x_j, y_\ell) \;]$$

Ex : • Noyau polynômial $k(x,x') = (\langle x, x' \rangle + c)^r$

↳ Feature map $\Phi(x) = \begin{pmatrix} x_1^2 \\ x_2^2 \\ \sqrt{2} x_1 x_2 \\ \sqrt{2} x_1 \\ \sqrt{2} x_2 \\ 1 \end{pmatrix}$ $(r = 2,\ d = 2)$

Complexité : $k(x,x') : O(d)$

$\Phi(x) : O\left(\binom{d+r}{r}\right)$

↑ $\dim \mathbb{R}_{\leq r}[X_1, \dots, X_d]$

• Radial Basis Function : $k(x,x') = \exp\left(-\frac{\|x - x'\|^2}{2\sigma^2}\right)$

$$\Phi(x) = e^{-\frac{x^2}{2\sigma^2}} \left(1, \sqrt{\frac{1}{1!\,\sigma^2}}\, x, \sqrt{\frac{1}{2!\,\sigma^2}}\, x^2, \dots \right)$$

$\dim \mathcal{H} = +\infty$

Il se trouve que bien que travaillant désormais (potentiellement) en dimension infinie, on peut se limiter à la dimension n pour des problèmes d'optimisation

Thm: (Théorème du représentant)

Soient $X_1, \dots, X_n \in \mathcal{X}$, $\Phi : \mathcal{X} \to \mathcal{H}$ et $g : \mathbb{R} \to \mathbb{R}$ strictement croissante

Alors

$$\underset{h \in \mathcal{H}}{\operatorname{Min}}\; L\big(\underbrace{h(X_1)}_{\substack{= \langle h, k(X_1, \cdot)\rangle_{\mathcal{H}} \\ = \langle h, \phi(X_1)\rangle_{\mathcal{H}}}}, \dots, h(X_n)\big) + g\left(\|h\|_{\mathcal{H}}\right)$$

n'admet que des solutions de la forme

$$h^* = \sum_{i=1}^{n} \alpha_i k(X_i, \cdot) = \sum_{i=1}^{n} \alpha_i \Phi(X_i)$$

↳ Analogie avec la régression linéaire :
les solutions de norme minimale sont une combinaison linéaire des descripteurs $\phi(X_i)$ (et donc vivent en dim $\leq n$)

Application : Kernel ridge regression

$$\underset{h \in \mathcal{H}}{\operatorname{Min}} \sum_{i=1}^{n} \left(Y_i - \langle h, \Phi(X_i)\rangle\right)^2 + \frac{\lambda}{2} \|h\|_{\mathcal{H}}^2$$

En notant $\mathbb{X} : \mathcal{H} \to \mathbb{R}^n$, $h \mapsto \begin{pmatrix} \langle h, \Phi(X_1)\rangle_{\mathcal{H}} \\ \vdots \\ \langle h, \Phi(X_n)\rangle_{\mathcal{H}} \end{pmatrix}$ et $\mathbb{X}^T : \mathbb{R}^n \to \mathcal{H}$, $\alpha \mapsto \sum_{i=1}^{n} \alpha_i \Phi(X_i) = \sum_{i=1}^{n} \alpha_i \Phi(X_i)$

on pose $\hat{\Sigma} := \mathbb{X}^T \mathbb{X} = \sum_{i=1}^{n} \Phi(X_i) \otimes \Phi(X_i)$

On obtient $h^* = (\hat{\Sigma} + \lambda I_{d_H})^{-1} \mathbb{X}^T Y$

$$= \mathbb{X}^T (\lambda I d_{\mathbb{R}^n} + \mathbb{X}\mathbb{X}^T)^{-1} Y$$

le ~~regresseur~~ final est alors

$$\boxed{\begin{aligned} \hat{f}_{KRR}(x) &:= \langle h^*, \phi(x) \rangle_H \\ &= \langle \mathbb{X}^T (\lambda I d_n + \mathbb{X}\mathbb{X}^T)^{-1} Y, \Phi(x) \rangle_H \\ &= \langle (\lambda I d_n + \mathbb{X}\mathbb{X}^T)^{-1} Y, \mathbb{X}(\Phi(x)) \rangle_{\mathbb{R}^n} \\ &= \left\langle (\lambda I d_n + \mathbb{X}\mathbb{X}^T)^{-1} Y, \begin{pmatrix} k(X_1, x) \\ \vdots \\ k(X_n, x) \end{pmatrix} \right\rangle_{\mathbb{R}^n} \end{aligned}}$$

qui ne dépend que de $k(\cdot,\cdot)$.

Rq: Pour $\lambda = 0$, on obtient $\hat{f}_{KR}(x) = \langle K^{-1} Y, K(X, x) \rangle_{\mathbb{R}^n}$

où $\begin{cases} K = (k(X_i, X_j))_{i,j \le n} \\ K(X, x) = \begin{pmatrix} k(X_1, x) \\ \vdots \\ k(X_n, x) \end{pmatrix} \end{cases}$

II Réseaux de neurones et Neural Tangent Kernel

1) Régression avec réseaux de neurones feedforward

Def : Un réseau de neurones (feedforward) à L couches cachées de fonction d'activation $\Psi : \mathbb{R} \to \mathbb{R}$

et une fonction $f^{(L)}(\cdot, W) : \mathbb{R}^d \to \mathbb{R}$ définie par

$$f^{(L)}(x, W) = W^{(L+1)} \frac{1}{\sqrt{k_L}} \Psi\left(W^{(L)} \frac{1}{\sqrt{k_{L-1}}} \Psi\left(\cdots W^{(2)} \frac{1}{\sqrt{k_1}} \Psi\left(W^{(1)} x\right)\right)\right),$$

où $\begin{cases} W^{(i)} \in \mathbb{R}^{k_i \times k_{i-1}} \\ k_0 = d \ , k_L = 1 \end{cases}$

Exemple d'utilisation : la classe $\left(f^{(L)}(\cdot, W)\right)_{W \in \mathbb{R}^p}$ peut être insérée dans une méthode par moindres carrés, cette fois-ci non-linéaire si Ψ est non-linéaire :

$$W^* \in \underset{W \in \mathbb{R}^p}{\operatorname{argmin}} \sum_{i=1}^{n} \left(Y_i - f^{(L)}(X_i, W)\right)^2,$$

qui peut être obtenu par descente de gradient (stochastique) autour d'une valeur initiale $W_0 \in \mathbb{R}^p$.

2) Linéarisation : le Neural Tangent Kernel

Si la SGD a des trajectoires qui restent dans le domaine de validité du developpement limité

$$f^{(L)}(x,W) \simeq f^{(L)}(x,W_0) + \langle \nabla_W f^{(L)}(x,W_0), x - x_0 \rangle$$

alors

$$W^*_{SGD} \in \underset{W \in \mathbb{R}^p}{\operatorname{argmin}} \sum_{i=1}^{n} \left(Y_i - \left(f^{(L)}(X_i,W_0) + \langle \nabla_W f^{(L)}(X_i,W_0), W - W_0 \rangle \right) \right)^2$$

ce qui revient essentiellement à faire une régression linéaire sur les features

$$\Phi(x) = \nabla_W f^{(L)}(x,W_0)$$

le noyau associé à ce descripteur est appelé Neural Tangent Kernel ou NTK

Def: (NTK) $\forall x, \tilde{x} \in \mathbb{R}^d$, $\forall W \in \mathbb{R}^p$, $K^{(L)}_W(x,\tilde{x}) := \langle \nabla_W f^{(L)}(x,W), \nabla_W f^{(L)}(\tilde{x},W) \rangle$

↳ [Jacot et all, Neurips 2018]

3) Validité de la linéarisation en grande largeur

Dans [Liu, Zhu, Belkin - Neurips 2020] les auteurs démontrent une validité uniforme de la constance du NTK pour des initialisations W_0 gaussiennes

Thm : Si $\begin{cases} \bullet\ L \text{ est fixé} \\ \bullet\ k_1, \dots, k_L \longrightarrow \infty \end{cases}$ on a pour tout $x, \tilde{x} \in \mathbb{R}^d$

$$W_{a,b}^{(i)} \overset{iid}{\sim} \mathcal{N}(0,1) \Rightarrow \begin{cases} \bullet\ \sup_{W' \in B(W,1)} \left| K_{W'}^{(L)}(x,\tilde{x}) - K_{W}^{(L)}(x,\tilde{x}) \right| \longrightarrow 0 \\ \bullet\ K_{W'}^{(L)}(x,x) = \Omega(1) \quad \forall W' \in B(W,1) \end{cases}$$

Autrement dit, le NTK est constant et non-trivial sur une boule de rayon constant :

$$\begin{aligned} B(W,1) &\longrightarrow \mathbb{R}^p \\ W' &\longmapsto \nabla_W f^{(L)}(x, W') \end{aligned}$$

"constant"

Abusif car $p \to \infty$

En particulier, si on utilise une méthode d'optimisation à l'ordre 1,

$$\text{Regression } L^2 \text{ sur } \{f^{(L)}(x,W)\}_{W \in \mathbb{R}^p} \Longleftrightarrow \text{Régression à noyau } K_W^{(L)}(\cdot,\cdot)$$

Informel^{++}

Une fois ceci acté, on peut se demander

- Quelle tête a $K_W^{(L)}$ en grande largeur ?
- $\quad$ —— $K_W^{(L)}$ —— profondeur ?
- À la limite, obtient-on une méthode statistiquement intéressante ?

III Convergence du NTK et application en classification

[Radhakrishnan, Belkin, Uhler, PNAS 2023 - Wide & Deep neural networks achieve consistency for classification]

1) NTK en largeur infinie et fonction d'activation duale

a) Une couche cachée

Pour commencer, on considère $L=1$ et

$$f(x,w)=\frac{1}{\sqrt{k}}A\,\psi(Bx)$$

où $\begin{cases} A\in\mathbb{R}^{1\times k} \\ B\in\mathbb{R}^{k\times d} \end{cases}$, $B=\begin{pmatrix} B_{1:} \\ \vdots \\ B_{k:} \end{pmatrix}$ $\boxed{w=(A_{11},\dots,A_{1k},B_{11},\dots,B_{kd}) \in\mathbb{R}^{k+kd} \quad (p=k+kd)}$

On suppose que $\begin{cases} A_{1i},B_{ij}\overset{iid}{\sim}N(0,1) \\ x,\tilde{x}\in\mathbb{R}^d \text{ sont normalisés } \|x\|=\|\tilde{x}\|=1 \end{cases}$ (Initialisation standard des NN)

Par définition $f(x,w)=\frac{1}{\sqrt{k}}\sum_{i=1}^{k}A_{1,i}\,\psi(\underbrace{B_{i:}x}_{=\sum_{j=1}^{d}B_{ij}x_j})$

d'où $\frac{\partial f}{\partial A_{1i}}(x,w)=\frac{1}{\sqrt{k}}\psi(B_{i:}x)$ et $\frac{\partial f}{\partial B_{ij}}(x,w)=\frac{1}{\sqrt{k}}A_{1i}\,\psi'(B_{i:}x)x_j$

En particulier,

$$K^{(1)}(x,\tilde{x}) = \left\langle \nabla_W f^{(1)}(x,W), \nabla_W f^{(1)}(\tilde{x},W) \right\rangle$$

$$= \underbrace{\frac{1}{k}\sum_{i=1}^{k} \psi(B_{i:}x)\psi(B_{i:}\tilde{x})}_{(a)}$$

$$+ \underbrace{\frac{1}{k}\sum_{i=1}^{k}\sum_{j=1}^{d} A_{1i}^2\, \psi'(B_{i:}x)\psi'(B_{i:}\tilde{x})\, x_j \tilde{x}_j}_{(b)}$$

(a) et (b) font intervenir des sommes iid de variables aléatoires dépendant de $(B_{i:}x, B_{i:}\tilde{x})$.

Comme $B_{i:} \sim N(0, Id_{d\times d})$, $\underbrace{(B_{i:}x, B_{i:}\tilde{x})}_{\in \mathbb{R}^2}$ est un vecteur gaussien

De plus, $E[B_{i:}x] = E[B_{i:}]x = 0$ et

$$Cov(B_{i:}x, B_{i:}\tilde{x}) = E\left[x^T B_{i:}^T B_{i:}\tilde{x}\right]$$

$$= x^T\tilde{x} \quad \text{car } E[B_{i:}^T B_{i:}] = Id_{d\times d}$$

$$= \langle x, \tilde{x}\rangle$$

Comme $\langle x,x\rangle = \langle \tilde{x},\tilde{x}\rangle = 1$, on a donc

$$\begin{pmatrix} B_{i:}x \\ B_{i:}\tilde{x} \end{pmatrix} \sim N(0,\Lambda) \quad \text{où} \quad \boxed{\Lambda := \begin{pmatrix} 1 & \langle x,\tilde{x}\rangle \\ \langle x,\tilde{x}\rangle & 1 \end{pmatrix}}$$

Par la loi des grands nombres :

$$(a) \xrightarrow[k\to\infty]{p.s} \underset{(u,v)\sim N(0,\Lambda)}{\mathbb{E}}\left[\psi(u)\psi(v)\right]$$

$$\mathbb{E}[A^2]=1$$

$$(b) \xrightarrow[k\to\infty]{p.s} \sum_{j=1}^{d} \underset{\substack{A\sim N(0,1) \\ \perp\!\!\!\perp \\ (u,v)\sim N(0,\Lambda)}}{\mathbb{E}}\left[A^2\psi'(u)\psi'(v)\right] x_j\tilde{x}_j = \mathbb{E}\left[\psi'(u)\psi'(v)\right]\langle x,\tilde{x}\rangle$$

La transformation $\psi \longmapsto \underset{(u,v)\sim N(0,\Lambda)}{\mathbb{E}}\left[\psi(u)\psi(v)\right]$ apparaît comme centrale !

$$\Lambda = \begin{pmatrix} 1 & \langle x,\tilde{x}\rangle \\ \langle x,\tilde{x}\rangle & 1 \end{pmatrix}$$

b) Fonction d'activation duale

Déf : (Activation duale)

Pour $\psi : \mathbb{R} \to \mathbb{R}$ fonction d'activation, sa fonction duale

$$\check{\psi} : [-1,1] \longrightarrow \mathbb{R}$$

est définie par $\check{\psi}(\xi) := \underset{(u,v)\sim N(0,\Lambda)}{\mathbb{E}}\left[\psi(u)\psi(v)\right]$, où $\Lambda = \begin{pmatrix} 1 & \xi \\ \xi & 1 \end{pmatrix}$

Rq : • Pour $\psi(z) = e^{iz}$, $K(x,\tilde{x}) = e^{-\frac{\|x-\tilde{x}\|^2}{2}}$

• le Neural Network Gaussian Process est $K(x,\tilde{x}) = \underset{(u,v)\sim N\left(0,\begin{pmatrix} \|x\|^2 & \langle x,\tilde{x}\rangle \\ \langle x,\tilde{x}\rangle & \|\tilde{x}\|^2 \end{pmatrix}\right)}{\mathbb{E}}\left[\psi(u)\psi(v)\right]$

donne $K(x,\tilde{x}) = \check{\psi}(\langle x,\tilde{x}\rangle)$ $\forall x,\tilde{x} \in \mathcal{S}^{d-1}$

les fonctions $\check{\psi}$ ont beaucoup de bonnes propriétés, et s'étudient via les polynômes de Hermite $h'_i(x) = \sqrt{i}\, h_i(x)$ qui forment une base orthogonale de $L^2(\mathcal{N}(0,1))$.

<u>Prop</u>: Si $\psi \in L^2(\mathcal{N}(0,1))$, alors

- $\psi(x) = \sum_{i=0}^{\infty} a_i h_i(x)$ a pour duale $\check{\psi}(\xi) = \sum_{i=0}^{\infty} a_i^2 \xi^i$
- $\check{\psi}$ est croissante convexe sur $[0,1]$, ainsi que toutes ses dérivées
- $(\check{\psi})' = (\check{\psi'})$ ($\vee$ commute avec la dérivation)
- $\check{\psi} \in \mathcal{C}^0([-1,1])$ et $\check{\psi} \in \mathcal{C}^\infty((-1,1))$
- $\check{\psi}$ est définie positive, au sens où $K : \mathbb{S}^{d-1} \times \mathbb{S}^{d-1} \to \mathbb{R}$, $(x, \tilde{x}) \mapsto \check{\psi}(\langle x, \tilde{x} \rangle)$ est un noyau (positif)
- Inversement, toute fonction $\sigma : [-1,1] \to \mathbb{R}$ définie positive est la duale $\check{\psi} = \sigma$ d'une fonction d'activation

c) Une formule de récurrence pour le NTK en largeur infinie

Pour simplifier, on normalise $\begin{cases} \psi \text{ de sorte que } \check{\psi}(1) = 1 \\ x, \tilde{x} \in \mathbb{R}^d \longrightarrow \|x\| = \|\tilde{x}\| = 1 \end{cases}$

Thm : Pour des entrées $W^{(i)} \in \mathbb{R}^{k_i \times k_{i-1}}$ ont des entrées $\mathcal{N}(0,1)$ iid, alors quand $k_1 \to \infty$, puis $k_2 \to \infty$, ..., puis $k_L \to \infty$, les NTK convergent presque sûrement vers

$$K^{(L)}(x, \tilde{x}) = \Sigma^{(L)}(x, \tilde{x}) + K^{(L-1)}(x, \tilde{x})\, \check{\psi}'\left(\Sigma^{(L-1)}(x, \tilde{x})\right)$$

avec $K^{(0)}(x, \tilde{x}) = \langle x, x \rangle$

où

$$\Sigma^{(L)}(x, \tilde{x}) = \check{\psi}\left(\Sigma^{(L-1)}(x, \tilde{x})\right)$$

avec $\Sigma^{(0)}(x, \tilde{x}) = \langle x, \tilde{x} \rangle$

Rq : • $\Sigma^{(L)}(x, \tilde{x})$ correspond à la dérivation par rapport à la dernière couche (linéaire)

• $K^{(L-1)}(x, \tilde{x})\, \check{\psi}'\left(\Sigma^{(L-1)}(x, \tilde{x})\right)$ à la dérivation par rapport aux autres couches.

2) NTK en profondeur infinie

a) Enoncé du résultat

On étudie maintenant le NTK lorsque la profondeur $L \to \infty$, en laissant à chaque fois toutes les largeurs tendre vers ∞

Dans tout ce qui suit, $\begin{cases} x, \tilde{x} \in \mathbb{S}^{d-1} \text{ sont fixés} \\ \boxed{\psi(1) = 1 = \mathbb{E}_{z \sim N(0,1)}\left[\psi(z)^2\right]} \end{cases}$

L'asymptotique de $\bar{K}^{(L)}$ est gouvernée par trois quantités :

$$\sqrt{\psi(0)} = \boxed{\begin{gathered} A := \mathbb{E}_{z \sim N(0,1)}\left[\psi(z)\right], \quad A' := \mathbb{E}_{z \sim N(0,1)}\left[\psi'(z)\right] \\ B' := \mathbb{E}_{z \sim N(0,1)}\left[\psi'(z)\right] = \psi'(1) \end{gathered}} = \sqrt{\psi'(0)}$$

__Rq:__ L'article étudie le classifieur limite $m_n(x) := \mathrm{sign}\left(\langle K^{-1} Y, K(X, x)\rangle\right)$ pour $(X_i, Y_i) \in \mathbb{R}^d \times \{-1, 1\}$ des données de classification fixées.

Les résultats sur l'asymptotique de $K^{(L)}$ se répercutent sur $m_n(x)$

Thm :

$(A=0, A'\neq 0)$ $\displaystyle\lim_{L\to\infty} \frac{K^{(L)}(x,\tilde{x})}{(A')^{2L}(L+1)} = \frac{R(\|x-\tilde{x}\|)}{\|x-\tilde{x}\|^{\alpha}}$, où $\begin{cases} \alpha = -\dfrac{2\log(A'^2)}{\log(B')} \\ R(\cdot)\geq 0 \text{ bornée loin de zero en } 0. \end{cases}$

↳ Classifieur à noyau singulier pour $\alpha > 0$

$(A=0, A'=0)$ Si $\|x - X_i\| = \min_{j\leq m} \|x - X_j\|$ et que ce minimum est unique,

$$\lim_{L\to\infty} \frac{K^{(L)}(X_j, x)}{K^{(L)}(X_i, x)} = 0 \quad \forall j \neq i$$

↳ Classifieur au plus proche voisin $m(x) = \text{sign}(X_i(x))$

$(A\neq 0)$ $\forall x \neq \tilde{x}$, $\displaystyle\lim_{L\to\infty} \frac{K^{(L)}(x,\tilde{x})}{C(L)} = C_1 \neq \lim_{L\to\infty} \frac{K^{(L)}(x,x)}{C(L)}$

↳ Classifieur par vote majoritaire $m(x) = \text{sign}\left(\sum_{i=1}^{M} Y_i\right)$

↳ ReLU $\varphi(x) = x_+$ tombe dans ce cas

Rq : Dans le cas $A=0, A'\neq 0$, le classifieur peut être consistant quand $n\to\infty$ cas il est equivalent au classifieur de Hilbert [Devroye, Györfi, Krzyniak - 1998] pour $\alpha = \frac{d}{2}$

↳ Et cet estimateur overfitte ! $(m(X_i) = Y_i)$

b) Idée de démonstration pour $A=0$, $A'\neq 0$

On note $\eta = \langle x, \tilde{x} \rangle$ et $\check{\psi}^{(L)}(\eta) = \underbrace{\check{\psi} \circ \cdots \circ \check{\psi}}_{(L+1) \text{ fois}}(\eta)$.

La formule de récurrence sur $K^{(L)}$ donne l'expression explicite

$$\boxed{K^{(L)}(\eta) = \sum_{i=0}^{L} \check{\psi}^{(i)}(\eta) \prod_{j=i}^{L-1} \check{\psi}'\left(\check{\psi}^{(j)}(\eta)\right)}$$

Pour étudier cette somme, on commence d'abord par examiner l'asymptotique de $\check{\psi}^{(l)}(\eta)$

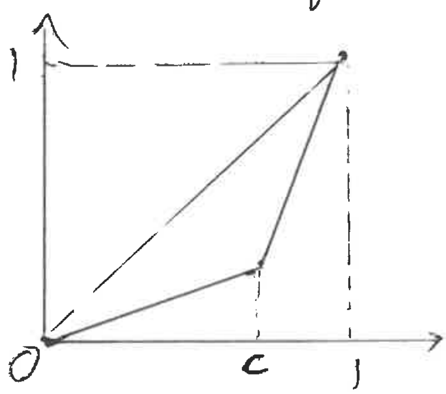
Lemme: $$\lim_{L\to\infty} \frac{\check{\psi}^{(L)}(\eta)}{\check{\psi}'(0)^L} = \frac{\tilde{R}(\eta)}{(1-\eta)^{\frac{\alpha}{2}}}$$, où $\alpha = -\frac{2\log \check{\psi}'(0)}{\log \check{\psi}'(1)} = -\frac{2\log A'^2}{\log(B')}$

$0 \neq A'^2 =$

Dem: On traite le cas simple où $f = \check{\psi}$ est telle que $f : [0,1] \to [0,1]$

$$f(x) = \begin{cases} ax & x \in [0,c] \\ 1-b(1-x) & x \in [c,1] \end{cases}$$

avec $c := \frac{b-1}{b-a}$ $\left(f'(0) = a,\ f'(1) = b \right)$



• Pour $x \in [0, c]$, $f^{(L)}(x) = a^L x$, d'où $\lim\limits_{L\to\infty} \dfrac{f^{(L)}(x)}{a^L} = x$

• Pour $x \in [c, 1]$, il existe $L_0 \in \mathbb{N}$ (défini comme le plus petit L tq que $f^{(L)}(x) \leq$

$\hookrightarrow$ Existe car $\begin{cases} f(x) < x \\ f(0) = 0 \leq c \end{cases}$

On écrit donc
$$\lim_{L\to\infty} \frac{f^{(L)}(x)}{a^L} = \lim_{L\to\infty} \frac{f^{(L-L_0)}\left(f^{(L_0)}(x)\right)}{a^{L-L_0}}\, a^{L_0}$$
$$= \frac{f^{(L_0)}(x)}{a^{L_0}}, \quad \text{car } f^{(L_0)}(x) \leq c \text{ et ce qui précède}$$

Un calcul simple donne
$$\frac{1}{a^{L_0}} \in \left[\left(\frac{1-c}{1-x}\right)^{-\frac{\log a}{\log b}}, \frac{1}{a}\left(\frac{1-c}{1-x}\right)^{-\frac{\log a}{\log b}} \right]$$

On obtient donc
$$\lim_{L\to\infty} \frac{f^{(L)}(x)}{a^L} = \frac{R(x)}{(1-x)^{-\frac{\log a}{\log b}}},$$

avec $\boxed{R(x) := \dfrac{f^{(L_0)}(x)\,(1-x)^{-\frac{\log a}{\log b}}}{a^{L_0}}}$, où $\boxed{\begin{cases} f^{(L_0)}(x) \in [ac, c] \\ \dfrac{(1-x)^{-\frac{\log a}{\log b}}}{a^{L_0}} \in \left[(1-c)^{-\frac{\log a}{\log b}}, \dfrac{(1-c)^{-\frac{\log a}{\log b}}}{a} \right] \end{cases}}$

Pour conclure sur $K^{(L)}(\eta) = \sum_{i=0}^{L} \check{\varphi}^{(i)}(\eta) \prod_{k=i}^{L-1} \check{\varphi}'(\check{\varphi}^{(k)}(\eta))$, les auteurs procèdent par $\underbrace{\text{majoration / minoration}}$

↳ Existence des limites laissées sous le tapis

* Majoration : long & technique
* Minoration : On utilise les propriétés de positivité de $\check{\varphi}$.

$\boxed{1}$ $\forall \eta \in [0,1]$, $\check{\varphi}'(\eta)\eta \geq \check{\varphi}(\eta)$

En effet, $\check{\varphi}(\eta) = \sum_{j=1}^{\infty} a_j \eta^j$ car $\check{\varphi}(0)=0$

$$\Rightarrow \check{\varphi}'(\eta)\eta = \eta \sum_{j=1}^{\infty} a_j \times j \eta^{j-1} = \sum_{j=1}^{\infty} \underbrace{j}_{\geq 1} \underbrace{a_j \eta^j}_{\geq 0} \geq \check{\varphi}(\eta) \square$$

$\boxed{2}$ On écrit

$$K^{(L)}(\eta) = \sum_{i=0}^{L} \check{\varphi}^{(i)}(\eta) \prod_{k=i}^{L-1} \check{\varphi}'(\underbrace{\check{\varphi}^{(k)}(\eta)}_{=\tilde{\eta}_k})$$

$$\geq \sum_{i=0}^{L} \check{\varphi}^{(i)}(\eta) \prod_{k=i}^{L-1} \frac{\check{\varphi}(\check{\varphi}^{(k)}(\eta))}{\check{\varphi}^{(k)}(\eta)}$$

$$= \sum_{i=0}^{L} \check{\varphi}^{(L)}(\eta)$$

$$= (L+1)\check{\varphi}^{(L)}(\eta).$$

Ainsi, $\lim_{L\to\infty} \frac{K^{(L)}(\eta)}{\check{\varphi}'(0)^L (L+1)} \geq \lim_{L\to\infty} \frac{\check{\varphi}^{(L)}(\eta)}{\check{\varphi}'(0)} = \frac{\tilde{R}(\eta)}{(1-\eta)^{\frac{\alpha}{2}}}$

IV. Quelques questions ouvertes

- Peut-on obtenir des résultats similaires avec des structures de réseaux de neurones différentes ?
  ↳ Convolutional NN ?

- La partie traitant de limite $L \to \infty$ est très artisanale et ne fait pas apparaître de structure à la limite.
  ↳ Formaliser la limite de noyaux généraux
  ([Aronszajn 1950] le fait pour des noyaux ↗ ou ↘)

  ↳ Donner une structure type RKHS pour des "noyaux" tels que
  $K(x,x) = +\infty$.

  ⚠ $\frac{\mathbb{1}_{[-1,1]}}{|x|^{\alpha}}$ pas RKHS

  (Partir de Fourier pour $K(x,\tilde{x}) = g(\|x - x'\|)$
  avec $\hat{g} \geq 0$ au sens des distributions ?)

  ↳ Comprendre plus finement le reste $R(\|x - \tilde{x}\|)$ au numérateur