TER M1 :
NN and RKHS

Matthieu Denis

7 juin 2021

Table des matières

Introduction : Réseau de neurone simple

Commencons par étudier un NN très simple : une fonction $\Phi : (\mathbb{R}^m \times \mathbb{R}^{m\times m} \times \mathbb{R}^m) \times \mathbb{R} \to \mathbb{R}$ combinaison d'applications linéaires, sans non linéarités intermédiaires :

$$\Phi((\beta, A, u), x) \coloneqq \frac{1}{m^\alpha}\beta^T \left(\frac{1}{m^\gamma} A\right) ux$$

On initialise $\theta^0 \coloneqq (\beta^0, A^0, u^0)$ de manière standarde : $\forall i, j \in \{1, \cdots, m\},\ u_i^0, A_{ij}^0, \beta_i^0 \sim_{iid} N(0,1)$

Nous montrerons quelques propriétés asymptotiques en la largeur des couches, et sur l'évolution des paramètres lors du premier pas de la descente de gradient.

Lois suivant la largeur des couches m

— Loi de $||u^0||_2^2$ pour m grand

Comme $||u^0||_2^2 = \sum_{i=1}^m (u_i^0)^2 \sim \chi^2(m)$ et $u_i^0 \sim \chi^2(1)$, en appliquant le TCL aux u_i^0, on a :

$$\frac{||u^0||_2^2 - m}{\sqrt{2m}} \sim_{m\to\infty} N(0,1)$$

C'est-à-dire que $||u^0||_2^2 \sim N(m, 2m)$ pour m grand.

— Loi de $\left(\frac{1}{m^\gamma}A^0\right) u^0 x$ sachant u^0

$(A^0u^0)_i = \sum_{j=1}^m A_{ij}^0 u_j^0$. En sachant u^0, comme $A_{i.}^0$ est un vecteur gaussien, $(A^0u^0)_i \sim N(0, ||u^0||_2^2)$.
De même, part indépendance des A_{ij}^0, les $(A^0u^0)_i$ sont indépendants et $A^0u^0 \sim N(0_m, ||u^0||_2^2\ Id_m)$.
Ainsi, $\left(\frac{1}{m^\gamma}A^0\right) u^0 x$ — $u^0 \sim N(0_m, \left(\frac{x}{m^\gamma}\right)^2 ||u^0||_2^2\ Id_m)$.

— Loi de $\left(\frac{1}{m^\gamma}A^0\right) u^0 x$

HISTOIRE DE MEME TRIBU ENGENDREE PAR U0 ET LE RESTE IMPLIQUE QUE CEST LA MEME CHOSE DECONDITIONNEE
Donc $\left(\frac{1}{m^\gamma}A^0\right) u^0 x \sim N(0_m, \left(\frac{x}{m^\gamma}\right)^2 ||u^0||_2^2\ Id_m)$.

— Loi de $||\left(\frac{1}{m^\gamma}A^0\right) u^0 x\,||_2^2$ pour m grand

On a $\left(\left(\frac{1}{m^\gamma}A^0\right) u^0 x\right)_i^2 \sim \left(\frac{x}{m^\gamma}\right)^2 ||u^0||_2^2 \cdot \chi^2(1)$, d'espérance $\mu \coloneqq \left(\frac{x}{m^\gamma}\right)^2 ||u^0||_2^2$ et de variance $\sigma^2 \coloneqq 2\left(\left(\frac{x}{m^\gamma}\right)^2 ||u^0||_2^2\right)^2$.
Donc en appliquant le TCL à ceux ci, on a :

$$\frac{||\left(\frac{1}{m^\gamma}A^0\right) u^0 x\,||_2^2 - m\mu}{\sigma\sqrt{m}} \sim_{m\to\infty} N(0,1)$$

C'est-à-dire que $||\left(\frac{1}{m^\gamma}A^0\right) u^0 x\,||_2^2 \sim N(m\mu, m\sigma^2)$ pour m grand.

— Loi de $\frac{1}{m^\alpha}(\beta^0)^T x_2$ sachant x_2, avec $x_2 = \left(\frac{1}{m^\gamma} A^0\right) u^0 x$

On a $\frac{1}{m^\alpha}(\beta^0)^T x_2 | x_2 \sim N(0, \frac{1}{m^{2\alpha}}||x_2||_2^2)$

— Loi de $\frac{1}{m^\alpha}(\beta^0)^T x_2$

On a $\frac{1}{m^\alpha}(\beta^0)^T x_2 \sim N(0, \frac{1}{m^{2\alpha}}||x_2||_2^2)$

Choix de α et γ

On regarde la variance de chaque composante de $\left(\frac{1}{m^\gamma} A^0\right) u^0 x$ pour m grand : $Var = x^2 m^{-2\gamma} m$ comme $||u^0||_2^2$ se comporte en m pour m grand. Or on ne veut pas qu'elle tende vers 0 ou l' infini lorsque m tend vers l'infini car Φ prendrait des valeurs de 0 ou l'infini, ce qui impose le choix $\gamma = 1/2$

Quant au choix de α, on a accès à la variance de $\frac{1}{m^\alpha}(\beta^0)^T x_2$, qui est une v.a gaussienne pour m grand. En prenant cette approximation, l'espérance de la variance de $\frac{1}{m^\alpha}(\beta^0)^T x_2$ est $m^{-2\alpha} m\mu$ pour m grand, i.e $x^2 m^{-2\alpha} m \cdot m^{-2\gamma} m = x^2 m^{-2\alpha+1}$ en prenant $\gamma = 1/2$. De la même manière que pour γ, on se retrouve avec le choix $\alpha = 1/2$ pour que la variance de Φ n'explose pas ni ne tende vers 0.

Gradients

Trivialement,

$$\nabla_u \Phi = \frac{x}{m^{\alpha+\gamma}} \beta^T A \in \mathbb{R}^m$$

$$\nabla_\beta \Phi = \frac{x}{m^{\alpha+\gamma}} A u \in \mathbb{R}^m$$

$$\nabla_A \Phi = \frac{x}{m^{\alpha+\gamma}} \beta u^T \in \mathbb{R}^{m \times m}$$

Descente de gradient

On va étudier ici le premier pas de descente de gradient.

Posons une fonction de perte $F : \mathbb{R} \to \mathbb{R}$ t.q $F'(0) \neq 0$ et $\Delta F \coloneqq F(\Phi(\theta^1, x)) - F(\Phi(\theta^0, x))$, avec $\theta^1 \coloneqq \theta^0 - \eta \nabla_\theta F(\Phi(\theta^0, x))$

Il semble honnête de prendre η dépendant de m, le produit scalaire final ayant plus de chance d'exploser en grande dimension. Prenons $\eta \coloneqq m^a$, $a \in \mathbb{R}$

— Choix de η

On veut que ΔF ne diverge pas ni ne tende vers 0 lorsque m tend vers l'infini.
Pour cela, on utilise l'approximation $\Delta F \simeq < \Delta\theta, \nabla_\theta F(\Phi(\theta^0, x)) >$.
On a

$$\Delta F \simeq < -\eta \nabla_\theta F(\Phi(\theta^0, x)), \nabla_\theta F(\Phi(\theta^0, x)) >= -\eta ||\nabla_\theta F(\Phi(\theta^0, x))||^2$$

$$\nabla_\theta F(\Phi(\theta^0, x)) = \underbrace{F'(\Phi(\theta^0, x))}_{\text{constante en } m} \cdot \nabla_\theta \Phi(\theta^0, x)$$

Or $||\nabla_\theta \Phi(\theta^0,x)||^2 = ||\nabla_u \Phi(\theta^0,x)||^2 + ||\nabla_A \Phi(\theta^0,x)||^2 + ||\nabla_\beta \Phi(\theta^0,x)||^2$

En faisant les mêmes types de calculs que dans la partie précédente, et en utilisant la loi des grands nombres, on trouve que pour m grand :

$$||\nabla_u \Phi(\theta^0,x)||^2 = C^2 \cdot ||(\beta^0)^T A||^2 \simeq cste \cdot C^2 \cdot m||\beta^0||^2 \simeq cste \cdot C \cdot m^2 \;\text{ avec } C := \frac{x}{m^{\alpha+\gamma}}$$

$$||\nabla_\beta \Phi(\theta^0,x)||^2 = C^2 \cdot ||Au||^2 \simeq cste \cdot C^2 \cdot m||u^0||^2 \simeq cste \cdot C^2 \cdot m^2$$

Pour $||\nabla_u \Phi(\theta^0,x)||^2$, en considérant le gradient en A comme un vecteur de la matrice des dérivées partielles applatie, on a $||\beta^0(u^0)^T||^2 = \sum_{i,j=1}^m (\beta_i^0 u_j^0)^2$. On fait face à une gaussiène puissance 4 : elle admet une espérance finie indépendante de m, donc en appliquant la LGN, on a :

$$||\nabla_u \Phi(\theta^0,x)||^2 = C^2 \cdot ||(\beta^0)^T A|| \simeq cste \cdot C^2 \cdot m^2$$

Ainsi, pour m grand :

$$\begin{aligned}
\Delta F &\simeq -\eta ||\nabla_\theta F(\Phi(\theta^0,x))||^2 \\
&= -cste \cdot \eta ||\nabla_\theta \Phi(\theta^0,x)||^2 \\
&\simeq -cste \cdot C^2 \eta (m^2 + m^2 + m^2) \\
&= -cste \cdot m^{-2} m^a m^2 \;\text{ en prenant } \alpha = \gamma = 1/2 \\
&= -cste \cdot m^a
\end{aligned}$$

Ce qui nous force le choix $a = 0$ pour que tout se passe bien.

A présent, regardons les comment les paramètres ont évolué après ce premier pas de descente de gradient, i.e les ordres de grandeur des Δ. On prend $\eta = \mathcal{O}(1)$, les résultats ci-dessus ne changent pas.

$$\begin{aligned}
\Delta u &= -\eta \nabla_u F(\Phi(\theta^0,x)) \\
&= -\eta F'(\Phi(\theta^0,x)) \cdot \nabla_u \Phi(\theta^0,x) \\
&= -\eta\, C\, d\, (\beta^0)^T A \text{ avec } d := F'(\Phi(\theta^0,x)) \\
&= -\eta\, x\, d\, m^{-1}\, (\beta^0)^T A
\end{aligned}$$

$$\begin{aligned}
|(\Delta u)_i| &= \eta\, |x\, d|\, m^{-1} \left| \sum_{j=1}^m A_{ji}\beta_j \right| \\
&\leq \eta\, |x\, d|\, m^{-1} \sum_{j=1}^m |A_{ji}\beta_j| \\
&\leq \eta\, |x\, d|\, m^{-1} \cdot m \sup_j |A_{ji}\beta_j| \\
&\leq \eta\, |x\, d| \sup_{ij} |A_{ji}\beta_j|
\end{aligned}$$

Donc

$$\begin{aligned}
||\Delta u|| &= \left(\sum_{i=1}^{m} |(\Delta u)_i|^2 \right)^{1/2} \\
&\leq (m\,(\eta\,|x\,d|\,\sup_{ij}|A_{ji}\beta_j|)^2)^{1/2} \\
&= \eta\,|x\,d|\,\sup_{ij}|A_{ji}\beta_j| \cdot m^{1/2} \\
&= \mathcal{O}(m^{1/2})
\end{aligned}$$

De la même manière :

$$\begin{aligned}
||\Delta\beta|| &\leq \eta\,|x\,d|\,\sup_{ij}|A_{ij}u_j| \cdot m^{1/2} \\
&= \mathcal{O}(m^{1/2})
\end{aligned}$$

$$\begin{aligned}
|(\Delta A)_{ij}| &= |-\eta\,x\,d\,m^{-1}\,(\beta^0(u^0)^T)_{ij}| \\
&= \eta\,|x\,d|\,m^{-1}\,|\beta_i^0 u_j^0| \\
&\leq \eta\,|x\,d|\,m^{-1}\,\sup_{ij}|\beta_i^0 u_j^0| \\
&= \mathcal{O}(m^{-1})
\end{aligned}$$

Donc

$$\begin{aligned}
||\Delta A||_F &= \left(\sum_{i,j=1}^{m} |(\Delta A)_{ij}|^2 \right)^{1/2} \\
&\leq (m^2(\eta\,|x\,d|\,m^{-1}\,\sup_{ij}|\beta_i^0 u_j^0|)^2)^{1/2} \\
&= \eta\,|x\,d|\,\sup_{ij}|\beta_i^0 u_j^0| \\
&= \mathcal{O}(1)
\end{aligned}$$