

Fondements théoriques du deep-learning

François Malgouyres¹, Edouard Pauwels², Nicolas Thome³ et Sébastien Gerchinovitz^{4,1}

¹Institut de Mathématiques de Toulouse, Université Paul Sabatier

²Mathématiques de la décision et statistique, TSE, Université Toulouse Capitole

³Institut des systèmes intelligents et de robotique, Sorbonne Université

⁴Institut de Recherche Technologique, Saint-Exupéry

2023

Plan

- 1 Introduction
 - Classification et régression
 - Le contrôle du risque
 - Les réseaux de neurones

2 Questions liées à l'optimisation et la régularisation implicite

Outline

1 Introduction

- **Classification et régression**
- Le contrôle du risque
- Les réseaux de neurones
 - Multi-Layer Perceptron/fully-connected neural network
 - Réseaux convolutifs
 - Beaucoup de variantes importantes

2 Questions liées à l'optimisation et la régularisation implicite

- Les réseaux comme une fonction des paramètres
- La Rétropropagation
- Géométrie locale des réseaux ReLU fully-connected
- Le paysage de la fonction objectif
 - Introduction
 - Paysage pour les réseaux larges
 - Paysage pour les réseaux linéaires
 - Paysage pour les réseaux linéaires (cas à 1 couche cachée)

Classification et régression

Il existe un couple de variables aléatoires (X, Y)

- On connaît \mathcal{X} et \mathcal{Y} tels que $\mathbb{P}(X \in \mathcal{X}) = \mathbb{P}(Y \in \mathcal{Y}) = 1$
- On sait que l'on va observer x d'une réalisation (x, y) de (X, Y)
- On veut construire une fonction $g: \mathcal{X} \rightarrow \mathcal{Y}$ prédisant y

Plus précisément, pour une fonction coût $L: \mathcal{Y} \times \mathcal{Y} \rightarrow \mathbb{R}$, on veut trouver

$$g^* \in \underset{L(g(x), y)}{\operatorname{argmin}} R(g)$$

On appelle $R(g) = \mathbb{E}(L(g(X), Y))$ le risque de g .
" " en population

Car : La personnes utilisant g observe le coût

$$R_{test}(g) \simeq \frac{1}{n'} \sum_{i=1}^{n'} L(g(x'_i), y'_i)$$

pour un échantillon i.i.d $(x'_i, y'_i)_{i=1..n'}$ suivant la loi (X, Y) , **indépendant de l'échantillon d'apprentissage**. La loi des grands nombres conduit à considérer R .

Classification

En classification, \mathcal{Y} est fini.

- Classification binaire : $\mathcal{Y} = \{-1, +1\}$
- Classification multi-classes : $\mathcal{Y} = \{1, \dots, C\}$, $C \in \mathbb{N}$, $C > 2$

En deep-learning, pour $C \geq 2$, on construit C fonctions

$$g_c : \mathcal{X} \rightarrow \mathbb{R}, c = 1..C$$

une note pour la classe c

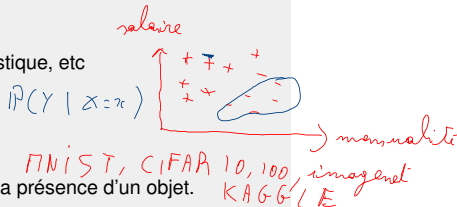
on retourne

$$y \in \operatorname{argmax}_{c=1..C} g_c(x)$$

$$L(y, y') = \mathbb{1}_{y \neq y'} \quad \text{ou la perte logistique, etc}$$

Exemples:

- $\mathcal{X} = \mathbb{R}^2$, $\mathcal{Y} = \{-1, +1\}$
- $\mathcal{X} = \mathbb{R}^n$: n est le nombre de pixels d'une image, $\mathcal{Y} = \{1, \dots, C\}$: chaque sortie est une "note" pour la présence d'un objet.
- $\mathcal{X} = \mathbb{R}^n$: n est la taille d'un "état" (ex: un damier), $\mathcal{Y} = \{1, \dots, C\}$: chaque sortie est l'espérance de gain de l'action $c = 1..C$.



Régression

En régression,

- $\mathcal{Y} = \mathbb{R}$
- $\mathcal{Y} = \mathbb{R}^C$

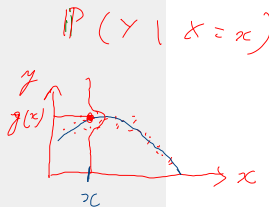
On construit C fonctions

$$g_c : \mathcal{X} \rightarrow \mathbb{R}, c = 1..C$$

$$L(y, y') = \|y - y'\|^2 \quad (\text{Il existe beaucoup de variantes})$$

Exemples:

- $\mathcal{X} = \mathbb{R}, \mathcal{Y} = \mathbb{R}$
- $\mathcal{X} = \mathbb{R}^d$: d est le nombre de mesures concernant un individu,
 $\mathcal{Y} = \mathbb{R}^C$: sa propension à avoir des maladies $c = 1..C$.
- $\mathcal{X} = \mathbb{R}^d$: d la taille d'un vocabulaire, $(0, \dots, 0, 1, 0, \dots, 0)$ est un mot
 $\mathcal{Y} = \mathbb{R}^C$: espace d'une représentation des mots du vocabulaire.



Décision Bayésienne

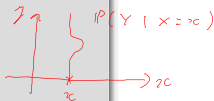
Sous des hypothèses faibles sur la loi de (X, Y) et si elle existe, la fonction définie pour $x \in \mathcal{X}$ par

$$g_b(x) \in \operatorname{argmin}_{y \in \mathcal{Y}} \mathbb{E}(L(y, Y) | X = x)$$

minimise le risque :

$$\forall g: \mathcal{X} \rightarrow \mathcal{Y}, \quad R(g) \geq R(g_b)$$

On appelle g_b la **décision Bayésienne**.



Remarques :

- On n'a pas forcément

$$\forall x \in \mathcal{X}, \exists a \in \mathcal{Y}, \quad \mathbb{P}(Y = a | X = x) = 1.$$

En général,

$$R(g_b) \neq 0.$$

- On ne peut pas calculer g_b car:
 - On ne connaît pas la loi de (X, Y)
 - On devrait représenter g_b avec un ordinateur

preuve: Soit $g: \mathcal{X} \rightarrow \mathcal{Y}$

$$R(g) = \mathbb{E}(L(Y, g(X)))$$

$$= \int_{\mathcal{X} \times \mathcal{Y}} L(g, g(x)) dP_{(X, Y)}(x, y)$$

$$= \int_{\mathcal{X}} \left[\int_{\mathcal{Y}} L(g, g(x)) dP_{Y|X}(y|x) \right] dP_X(x)$$

$$= \int_{\mathcal{X}} \mathbb{E}(L(Y, g(x)) | X = x) dP_X(x)$$

$$\geq \int_{\mathcal{X}} \mathbb{E}(L(Y, g_b(x)) | X = x) dP_X(x)$$

$$= R(g_b)$$

Classification et régression : les principes

apprentissage supervisé

- On observe $(x_i, y_i)_{i=1..n}$ un échantillon i.i.d. de même loi que (X, Y)
- On considère une famille \mathcal{F} de fonctions g_w paramétrés par des paramètres w d'un espace Euclidien.
 - ▶ Pour nous : elle est basée sur les réseaux de neurones
- On voudrait résoudre

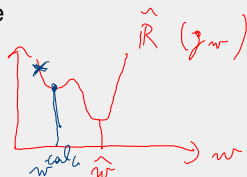
$$w^* \in \operatorname{argmin}_w R(g_w) = \mathbb{E} (L(g(x), y))$$

- La minimisation du risque empirique consiste à résoudre

$$\hat{w} \in \operatorname{argmin}_w \hat{R}(g_w)$$

où

$$\hat{R}(g_w) = \frac{1}{n} \sum_{i=1}^n L(g_w(x_i), y_i)$$



Ci-dessous, pour simplifier, on note $R(w)$ au lieu de $R(g_w)$ (idem $\hat{R}(w)$).

Classification et régression : les principes

Cela pose plusieurs problèmes:

- **Optimisation:** Est-ce que $\mathbf{w}^{calc} \sim \hat{\mathbf{w}}$?

Au sens où

$$\hat{R}(\mathbf{w}^{calc}) \approx \hat{R}(\hat{\mathbf{w}})$$

- ▶ les **structures de la fonction objectif** ne se prêtent pas forcément à l'optimisation
- ▶ le calcul de la fonction objectif dépend de **beaucoup d'entrées** (n grand)
- ▶ \mathbf{w} vit dans un espace de **grande dimension** (d grand)

- **Généralisation** de $(x_i, y_i)_{i=1..n}$: Est-ce que $\hat{\mathbf{w}} \sim \mathbf{w}^*$?

Au sens où :

$$R(\hat{\mathbf{w}}) \approx R(\mathbf{w}^*)$$

- ▶ Contrôle t'on

$$\sup_{\mathbf{w}} |R(\mathbf{w}) - \hat{R}(\mathbf{w})|?$$

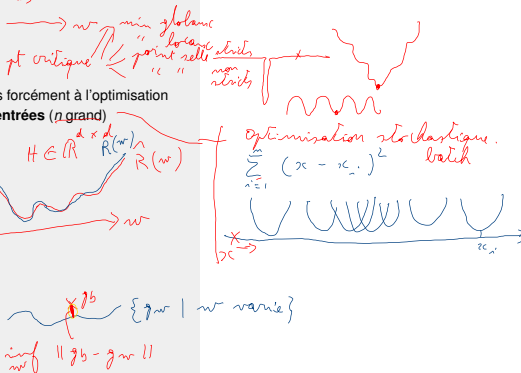
- **Approximation:** Est-ce que $g_{w^*} \sim g_b$?

Au sens où :

$$R(\mathbf{w}^*) \approx R(g_b)$$

- ▶ Que sait-on de g_b ?

- ▶ Contrôle t'on $\inf_{\mathbf{w}} |R(\mathbf{w}) - R(g_b)|?$



Outline

1 Introduction

- Classification et régression
- **Le contrôle du risque**
- Les réseaux de neurones
 - Multi-Layer Perceptron/fully-connected neural network
 - Réseaux convolutifs
 - Beaucoup de variantes importantes

2 Questions liées à l'optimisation et la régularisation implicite

- Les réseaux comme une fonction des paramètres
- La Rétropropagation
- Géométrie locale des réseaux ReLU fully-connected
- Le paysage de la fonction objectif
 - Introduction
 - Paysage pour les réseaux larges
 - Paysage pour les réseaux linéaires
 - Paysage pour les réseaux linéaires (cas à 1 couche cachée)

Le contrôle du risque

- Pour toute fonction g , on appelle **risque en population** et **risque empirique** de g :

$$R(g) = \mathbb{E}(L(g(X), Y)) \quad \text{et} \quad \hat{R}(g) = \frac{1}{n} \sum_{i=1}^n L(g(x_i), y_i)$$

- $R^* = \inf_g R(g)$ le risque optimal
- Pour $\varepsilon > 0$, on fixe \mathbf{w}^* et $\hat{\mathbf{w}}$ tels que :

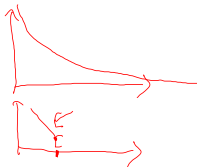
$$R(g_{\mathbf{w}^*}) \leq \inf_{\mathbf{w}} R(g_{\mathbf{w}}) + \varepsilon \quad \text{et} \quad \hat{R}(g_{\hat{\mathbf{w}}}) \leq \inf_{\mathbf{w}} \hat{R}(g_{\mathbf{w}}) + \varepsilon$$

- Pour \mathbf{w} retourné par un algorithme *on les ont il n'existe pas de minimiseur*

On **décompose**

$$\begin{aligned} 0 &\leq \underbrace{|R(g_{\hat{\mathbf{w}}}) - R^*|}_{\text{(l'excès de risque)}} \\ &= |R(g_{\hat{\mathbf{w}}}) - \hat{R}(g_{\hat{\mathbf{w}}})| \quad \text{(erreur de généralisation)} \\ &+ |\hat{R}(g_{\hat{\mathbf{w}}}) - \hat{R}(g_{\mathbf{w}^*})| \quad \text{(erreur d'optimisation)} \\ &+ |\hat{R}(g_{\mathbf{w}^*}) - R(g_{\mathbf{w}^*})| \quad \leq \varepsilon \\ &+ |R(g_{\mathbf{w}^*}) - R^*| \quad \text{(erreur de généralisation)} \\ &\quad \text{(erreur d'approximation)} \end{aligned}$$

Puis on **majoré** indépendamment chaque terme.



Le contrôle du risque

$E(w+h) = E(w) + \nabla E(w) \cdot h$
 $+ \frac{1}{2} (\nabla^2 E(w) \cdot h, h) + o(\|h\|^2)$
en m p. c
sur paramètres
 $\lambda_i < 0$

● **Erreur d'optimisation** : $|\hat{R}(g_w) - \hat{R}(g_w^*)|$

- Est-ce que l'optimisation a réussi : Optimisation non-convexe, paysage de la fonction objectif...
- En deep-learning** : Elle est petite pour les "gros" réseaux.

● **Erreur de généralisation** : $|R(g_w) - \hat{R}(g_w)|$

- Quelles conditions sur le réseau et l'échantillon pour avoir la **convergence uniforme**

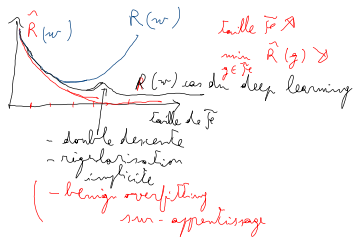
$\sup_{w \in \mathcal{C}} |\hat{R}(g_w) - R(g_w)|$ soit petit?

- Quelle dimension de Vapnik-Chervonenkis, Complexité de Rademacher ...
- En deep-learning** : Reste faible même pour les "gros" réseaux

~ régularisation implicite.

● **Erreur d'approximation** : $|R(g_{w^*}) - R(g_b)|$

- Il suffit que $\|g_{w^*} - g_b\|$ soit petit.
- Quelles fonctions peut-on approximer avec notre classe de fonctions? ("Expressive power", "Expressivité", ...)
- En deep-learning** : Elle est petite pour les "gros" réseaux.



Le contrôle du risque: En pratique

Pour \mathbf{w} retourné par un algorithme

$$R(g_{\mathbf{w}}) = \underbrace{(R(g_{\mathbf{w}}) - \hat{R}(g_{\mathbf{w}}))}_{\text{atteignable par la minimisation}} + \underbrace{\hat{R}(g_{\mathbf{w}})}_{\text{(erreur de généralisation) observable}}$$

- N'a d'intérêt que si on est capable de trouver un réseau et un \mathbf{w} tels que

$\hat{R}(g_{\mathbf{w}})$ soit petit.

- ▶ i.e. : On a résolu les problèmes de l'expressivité et de l'optimisation.
- **Erreur de généralisation** : même problématique que précédemment.

Classification et régression : les principes

ATTENTION : Cette modélisation **ne** tient **pas** compte :

- du **biais** dans les données: $(x_i, y_i)_{i=1..n}$ est rarement i.i.d. selon (X, Y)
 - ▶ Les x_i peuvent ne couvrir que partiellement le domaine, être corrompus, etc
 - ▶ Les y_i peuvent être une décision biaisée, bruitée, etc
- du besoin de **robustesse**
 - ▶ On impose que y_i soit prédit pour tout $x_i + e$, où $\|e\| \leq \varepsilon$
- d'une **régularisation** du réseau
 - ▶ w est parfois **quantifié** \implies **interprétabilité**, faible **coût énergétique/calculatoire/mémoire**
 - ▶ w est tel que g_w est **1-Lipschitz** \implies garantie de **robustesse**
 - ▶ **Régularisations explicites** : poids parcimonieux, pénalisation quadratique, dropout
- Problèmes plus complexes : apprentissage non-supervisé, semi-supervisé, par renforcement, apprentissage de représentations ...

Outline

1 Introduction

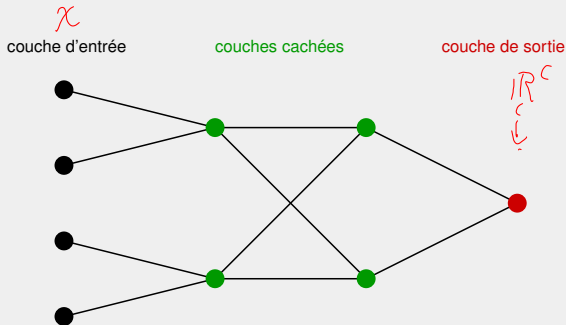
- Classification et régression
- Le contrôle du risque
- **Les réseaux de neurones**
 - Multi-Layer Perceptron/fully-connected neural network
 - Réseaux convolutifs
 - Beaucoup de variantes importantes

2 Questions liées à l'optimisation et la régularisation implicite

- Les réseaux comme une fonction des paramètres
- La Rétropropagation
- Géométrie locale des réseaux ReLU fully-connected
- Le paysage de la fonction objectif
 - Introduction
 - Paysage pour les réseaux larges
 - Paysage pour les réseaux linéaires
 - Paysage pour les réseaux linéaires (cas à 1 couche cachée)

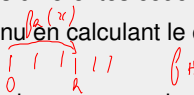
Multi-Layer Perceptron/fully-connected neural network

Feedforward



- Couche 0: couche d'entrée;
Couches 1 et 2: couches cachées;
Couche 3: couche de sortie.
- Chaque sommet (= neurone) contient un réel
- Chaque arête contient un poids

Multi-Layer Perceptron/fully-connected neural network

- H couches (on dit aussi $H - 1$ couches cachées)
- les tailles $\hat{n}_0, n_1, \dots, n_H \in \mathbb{N}$ des différentes couches
- On note $f_h(x)$: le résultat obtenu en calculant le contenu de la couche h pour l'entrée $x \in \mathbb{R}^{n_0}$

- On note $W_h \in \mathbb{R}^{n_h \times n_{h-1}}$: la matrice contenant les poids sur les arcs entre la couche $h-1$ et la couche h
- On note $b_h \in \mathbb{R}^{n_h}$: le biais ajouté à la couche h
- On note σ_h : la fonction d'activation appliquée à chaque couche
 - ▶ Typiquement, elle applique la même fonction à chaque entrée d'un vecteur

La prédiction/l'inférence : la fonction $f_H: \mathbb{R}^{n_0} \rightarrow \mathbb{R}^{n_H}$

$$\begin{cases} f_0(x) = x \\ f_h(x) = \sigma_h(W_h f_{h-1}(x) + b_h) \end{cases}, \forall h = 1, \dots, H$$



Multi-Layer Perceptron/fully-connected neural network : Exemple

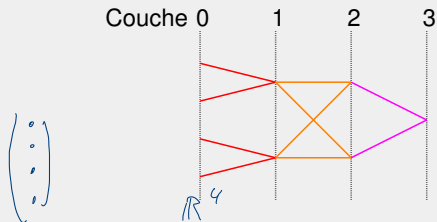


Figure: Réseau Fully-connected

Ci-dessous, chaque * est un nombre

$$W_1 = \begin{pmatrix} * & * & 0 & 0 \\ 0 & 0 & * & * \end{pmatrix}, b_1 = \begin{pmatrix} * \\ * \end{pmatrix}, W_2 = \begin{pmatrix} * & * \\ * & * \end{pmatrix}, b_2 = \begin{pmatrix} * \\ * \end{pmatrix}, W_3 = \begin{pmatrix} * & * \end{pmatrix},$$

$$b_3 = \begin{pmatrix} * \end{pmatrix},$$

$$f_3(x) = \sigma_3 \left(W_3 \sigma_2 \left(W_2 \sigma_1 \left(W_1 x + b_1 \right) + b_2 \right) + b_3 \right)$$

Multi-Layer Perceptron/fully-connected neural network: Vocabulaire et variantes

- **Fully-connected layer** : La matrice est “pleine”
- Fully connected **linear** network: On a $\sigma(t) = t, \forall t \in \mathbb{R}$, sans biais $b_h = 0$, pour tout $h = 1..H$.
- **Fonctions d'activation** :
 - ▶ Une zoologie importante : tanh, heavyside, identité, sigmoïde, etc
 - ▶ Des non-locales : group-sort, max_pooling
 - ▶ La plus utilisée est "Rectified Linear Unit" (ReLU) : $\sigma(t) = \max(0, t)$
 - ▶ Souvent, celle associée à la dernière couche est l'identité ou softmax (pour la classification).
- **“Batch-normalisation”** : Au lieu d'optimiser les biais \mathbf{b} , on règle un couple $(\text{diag}(\gamma), \mathbf{b})$ permettant de centrer les données et fixer leur variance.
- **“Dropout”** : On minimise le risque moyen pour des sous-réseaux aléatoires.
- **L'architecture du réseau**: La donnée des hyperparamètres. Ex : Un réseau ReLU de largeur 100 et de profondeur 10.

Regression

Multi-Layer Perceptron/fully-connected neural network

Proposition : Propriétés de f_H , cas linéaire avec biais

Pour tout réseaux linéaire fully-connected, f_H **est affine**:

$$f_H(x) = W_H \cdots W_1 x + b'_H$$

où

$$b'_H = W_H \cdots W_2 b_1 + \cdots + W_H b_{H-1} + b_H$$

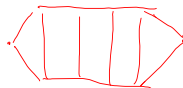
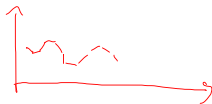
Preuve:

- Une composition de fonctions affines est affine.
- La formule est facile établir avec une récurrence sur H . (Exercice)

Proposition : Propriétés de f_H , cas ReLU (On verra la preuve plus tard.)

Pour tout réseaux fully-connected, avec la fonction d'activation ReLU et l'identité sur la dernière couche

- 1 La fonction f_H est continue
- 2 La fonction f_H est affine par morceaux
- 3 Il existe une partition compatible ayant moins de $2^{n_1 + \cdots + n_{H-1}}$ morceaux dont les morceaux sont des polyèdres ayant au plus $n_1 + \cdots + n_{H-1}$ faces.



Réseaux convolutifs

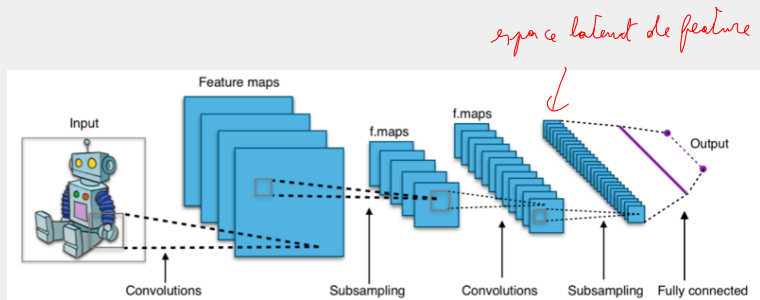
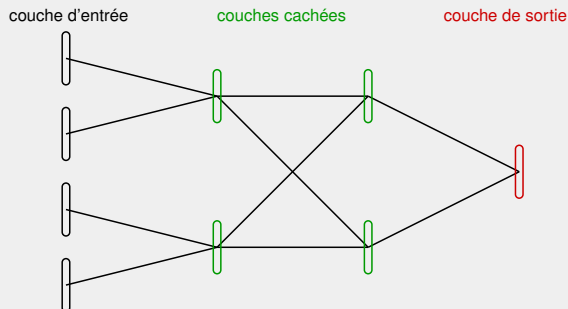


Figure: Source : wikipedia page "apprentissage profond"

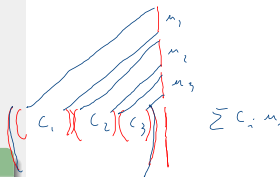
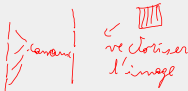
Réseaux convolutifs



- Chaque sommet contient un signal/une image. On parle de **canal**.
- Les canaux de la couche d'entrée correspondent aux canaux (ex (r, g, b)) de x .

Réseaux convolutifs

- **des signaux de taille N**
- H couches (on dit aussi $H - 1$ couches cachées)
- les tailles $n_0, n_1, \dots, n_H \in \mathbb{N}$ des différentes couches (= $N \times$ nombre de canaux)
- On note $f_h(x)$: le résultat obtenu en calculant le contenu de la couche h pour l'entrée x
- On note $W_h \in \mathbb{R}^{n_h \times n_{h-1}}$: la matrice **concaténant des matrices de convolutions** pour passer de la couche $h - 1$ et la couche h
- On note $b_h \in \mathbb{R}^{n_h}$: le biais ajouté à la couche h
- On note σ : la fonction d'activation appliquée à chaque couche
 - elle applique la même fonction à chaque entrée d'un vecteur



L'action du réseau : la fonction f_H

$$\begin{cases} f_0(x) = \text{vect}(x) \\ f_h(x) = \sigma(W_h f_{h-1}(x) + b_h) \end{cases}, \forall h = 1, \dots, H$$

Réseaux convolutifs: Matrice circulante, Toeplitz, bloc circulante... calculant des convolutions

On suppose $v, x \in \mathbb{R}^N$. Le signal v est supposé (N) -périodique de période:

$$v_{-n'} = v_{N-n'} \quad \text{pour tout } n' = 0..N-1$$



On a:

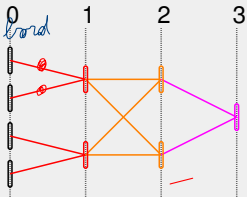
$$\begin{pmatrix}
 v_0 & v_{N-1} & \cdots & \cdots & \cdots & v_1 \\
 v_1 & v_0 & v_{N-1} & \cdots & \cdots & v_2 \\
 \vdots & \ddots & \ddots & \ddots & \ddots & \vdots \\
 \vdots & & \ddots & \ddots & \ddots & \vdots \\
 \vdots & & & \ddots & \ddots & v_{N-1} \\
 v_{N-1} & \cdots & \cdots & \cdots & v_1 & v_0
 \end{pmatrix}
 \begin{pmatrix}
 x_0 \\
 x_1 \\
 \vdots \\
 \vdots \\
 \vdots \\
 x_{N-1}
 \end{pmatrix}
 =
 \begin{pmatrix}
 \sum_{n'=0}^{N-1} v_{0-n'} x_{n'} \\
 \sum_{n'=0}^{N-1} v_{1-n'} x_{n'} \\
 \vdots \\
 \vdots \\
 \vdots \\
 \sum_{n'=0}^{N-1} v_{N-1-n'} x_{n'}
 \end{pmatrix}
 = \underline{v * x}$$

Réseaux convolutifs: Exemple

kernel
 $K_{c,m,i,j} \in \mathbb{R}$
 n_c canal en entrée
 n_c canal sortie



Couche 0
 Couche 1
 Couche 2
 Couche 3



$$(v * u)_{k,h} = \sum_{i,j} v_{i,j} u_{i-k, j-h}$$



Figure: Réseau convolutif

Ci-dessous,

- chaque * est une matrice composant **une convolution et un échantillonnage**
- chaque \circ est un vecteur colonne de taille N (cas sans sous-échantillonnage)

$$W_1 = \begin{pmatrix} * & * & 0 & 0 \\ 0 & 0 & * & * \end{pmatrix}, b_1 = \begin{pmatrix} \circ \\ \circ \end{pmatrix}, W_2 = \begin{pmatrix} * & * \\ * & * \end{pmatrix}, b_2 = \begin{pmatrix} \circ \\ \circ \end{pmatrix}, W_3 = \begin{pmatrix} * & * \end{pmatrix}, b_3 = \begin{pmatrix} \circ \end{pmatrix}$$

$W(k) (n_c \cdot)$

$$f_3(x) = \sigma \left(W_3 \sigma \left(W_2 \sigma \left(W_1 x + b_1 \right) + b_2 \right) + b_3 \right)$$

Réseaux convolutifs: Propriétés

Remarques

- Les formules sont les mêmes que dans le cas fully-connected. Seules les façons de construire les W_h et $f_0(x)$ diffèrent. $W_a (w) \quad W_a : \mathbb{R}^s \longrightarrow \mathbb{R}^{a \times a-1}$
- Un réseau convolutif contient des couches convolutives et des couches fully-connected.
- **Un réseau convolutif est un réseau fully-connected dans lequel on impose une structure.**

Proposition : Propriétés de f_H , cas linéaire

Pour tout réseaux convolutif linéaire , f_H **est affine**:

$$f_H(x) = W_H \cdots W_1 x + b'_H$$

où

$$b'_H = W_H \cdots W_2 b_1 + \cdots + W_H b_{H-1} + b_H$$

De plus, $x \longmapsto W_H \cdots W_1 x$ est une concaténation de convolutions du signal de départ.
(Preuve en exercice)

Réseaux convolutifs: Propriétés

Proposition : Propriétés de f_H , cas ReLU

Pour tout réseau formé de couches convolutives et fully-connected, avec la fonction d'activation ReLU et l'identité sur la dernière couche

- 1 La fonction f_H est continue
 - 2 La fonction f_H est affine par morceaux
 - 3 Il existe une partition compatible ayant moins de $2^{n_1 + \dots + n_{H-1}}$ morceaux dont les morceaux sont des polyèdres ayant au plus $n_1 + \dots + n_{H-1}$ faces.
-
- Réciproque dans : Arora, Raman, et al. "Understanding Deep Neural Networks with Rectified Linear Units." ICLR, 2018:
Toute fonction continue, affine sur un nombre fini de morceaux est implémentable par un réseau de neurone ReLU.

Beaucoup de variantes importantes

- Des architectures pour éviter le “vanishing/exploding gradient”: ex : Couches "Res-net"

$$f_h(x) = \sigma\left(W_h \sigma\left(W_{h-1} f_{h-2}(x) + b_{h-1}\right) + b_h + f_{h-2}(x)\right)$$

- Des architectures plus évoluées, dédiés à des applications.
 - ▶ En vision : U-Net (segmentation), YOLO (detection), transformer...
 - ▶ Séries temporelles et traitement du langage naturel : URNN, LSTM, transformer, GPT-3...
 - ▶ etc
- etc

Les combinaisons de réseaux

- **Autoencoder et VAE** : Un réseau compresse, un réseau décompresse.
- **Création d'embeddings** : Deux réseaux envoient une image/un texte dans un même espace de caractéristiques. (Ex : Visual Query Answering (VQA))
- **Generative Adversarial Networks (GAN)** : un réseau génère des données; un réseau discrimine les données générées de vraies données.
- **Réduction de biais dans les données (FairGAN)** : Un réseaux envoie les données dans un espace de caractéristiques et ses poids sont optimisés pour qu'un autre réseau classifiant sur un critère pertinent fonctionne, un troisième réseau classifiant sur un critère non-pertinent échoue.
- etc

Plan

1 Introduction

2 Questions liées à l'optimisation et la régularisation implicite

- Les réseaux comme une fonction des paramètres
- La Rétropropagation
- Géométrie locale des réseaux ReLU fully-connected
- Le paysage de la fonction objectif

Outline

1 Introduction

- Classification et régression
- Le contrôle du risque
- Les réseaux de neurones
 - Multi-Layer Perceptron/fully-connected neural network
 - Réseaux convolutifs
 - Beaucoup de variantes importantes

2 Questions liées à l'optimisation et la régularisation implicite

- **Les réseaux comme une fonction des paramètres**
- La Rétropropagation
- Géométrie locale des réseaux ReLU fully-connected
- Le paysage de la fonction objectif
 - Introduction
 - Paysage pour les réseaux larges
 - Paysage pour les réseaux linéaires
 - Paysage pour les réseaux linéaires (cas à 1 couche cachée)

Structure de la fonction objectif

- On dispose d'un échantillon $(x_i, y_i)_{i=1..n}$.
- On dispose d'un réseau de profondeur H , d'architecture fixée, dont
 - ▶ les arcs sont paramétrés par $\mathbf{w} \in \mathbb{R}^{H \times S}$. La matrice permettant de passer d'une couche à l'autre est $W_h(\mathbf{w}_h)$ où $W_h: \mathbb{R}^S \rightarrow \mathbb{R}^{n_h \times n_{h-1}}$ est (souvent) linéaire.
 - ▶ les nœuds sont paramétrés par $\mathbf{b} \in \mathbb{R}^{n_1 + \dots + n_H}$
 - ▶ les fonctions d'activation sont notées σ_h

La prédiction est la fonction

$$f_{\mathbf{w}, \mathbf{b}}(x) = \sigma_H(W_H(\mathbf{w}_H) \cdots \sigma_2(W_2(\mathbf{w}_2) \sigma_1(W_1(\mathbf{w}_1)x + b_1) + b_2) \cdots + b_H)$$

- Les réseaux ReLU:
 - ▶ $\sigma_h = \text{ReLU}$, pour $h = 1, \dots, H-1$
 - ▶ Cas de la **régression** : $\sigma_H = \text{Id}$
 - ▶ Cas de la **classification** : $\sigma_H(y) \in \text{argmax}_{i=1, \dots, n_H} y_i$ ou un softmax.
- On considère une fonction coût définie par

$$\begin{aligned} E: \mathbb{R}^{H \times S} \times \mathbb{R}^{n_1 + \dots + n_H} &\longrightarrow \mathbb{R} \\ (\mathbf{w}, \mathbf{b}) &\longmapsto E(\mathbf{w}, \mathbf{b}) = \sum_{i=1}^n L(f_{\mathbf{w}, \mathbf{b}}(x_i), y_i) \end{aligned}$$

pour une fonction coût $L: \mathcal{Y} \times \mathcal{Y} \rightarrow \mathbb{R}$.

Structure de la fonction objectif

Proposition: Non-convexité

Pour la loss quadratique $L(y', y) = (y' - y)^2$, le réseaux ReLU fully connected d'architecture $(1, 1, 1)$ et l'échantillon constitué de l'exemple $(x, y) = (1, 0)$, le risque empirique est non convexe.

Preuve: On a pour tout $(w_1, w_2) \in \mathbb{R}^2$, avec $w_1 \geq 0$,

$$E(w_1, w_2) = (w_2 w_1)^2.$$

Cette fonction n'est pas convexe car:

- En $(0, 1)$, $E(0, 1) = 0$
- En $(1, 0)$, $E(1, 0) = 0$
- En $(0.5, 0.5) = 0.5 \times (1, 0) + 0.5 \times (0, 1)$,

$$E(0.5, 0.5) = 0.25^2 > 0 = 0.5 \times E(1, 0) + 0.5 \times E(0, 1).$$



Outline

1 Introduction

- Classification et régression
- Le contrôle du risque
- Les réseaux de neurones
 - Multi-Layer Perceptron/fully-connected neural network
 - Réseaux convolutifs
 - Beaucoup de variantes importantes

2 Questions liées à l'optimisation et la régularisation implicite

- Les réseaux comme une fonction des paramètres
- **La Rétropropagation**
- Géométrie locale des réseaux ReLU fully-connected
- Le paysage de la fonction objectif
 - Introduction
 - Paysage pour les réseaux larges
 - Paysage pour les réseaux linéaires
 - Paysage pour les réseaux linéaires (cas à 1 couche cachée)

La Rétropropagation

On suppose:

- La fonction coût est de la forme :

$$(\mathbf{w}, \mathbf{b}) \longmapsto E(\mathbf{w}, \mathbf{b}) = \sum_{i=1}^n L(f_{\mathbf{w}, \mathbf{b}}(x_i), y_i)$$

où L est C^1 .

- On suppose que σ est C^1
 - ▶ On peut régulariser σ
 - ▶ Il existe un cadre formel pour faire du calcul différentiel et de l'optimisation avec des activations comme ReLU
- Sous ces hypothèses: E **est différentiable**.
- Les frameworks de deep learning utilisent la **différentiation automatique**.
- Ci-dessous, on ne considère qu'un unique exemple (x, y) :
 - ▶ On somme si besoin plusieurs gradients $\nabla L(f_{\mathbf{w}, \mathbf{b}}(x_i), y_i)$

Rétropropagation: cas fully-connected

- \mathbf{W}_h la matrice pour passer de la couche $h-1$ à h ,
- $f_{\mathbf{w},\mathbf{b}}^h(x)$, le contenu de la couche h ,
- $d_{\mathbf{w},\mathbf{b}}^h(x) = \sigma'(\mathbf{W}_h f_{\mathbf{w},\mathbf{b}}^{h-1}(x) + \mathbf{b}_h)$.

Proposition: Gradient pour un réseau fully-connected

On a pour $h = 1..H$, $i = 1..n_h$, $j = 1..n_{h-1}$

$$\frac{\partial E}{\partial \mathbf{W}_{h,i,j}}(\mathbf{w}, \mathbf{b}) = \left[f_{\mathbf{w},\mathbf{b}}^{h-1}(x) \right]_j \left[d_{\mathbf{w},\mathbf{b}}^h(x) \right]_i \Delta_i^h(x)$$

$$\frac{\partial E}{\partial \mathbf{b}_{h,i}}(\mathbf{w}, \mathbf{b}) = \left[d_{\mathbf{w},\mathbf{b}}^h(x) \right]_i \Delta_i^h(x)$$

où $\Delta^h(x) \in \mathbb{R}^{n_h}$ est défini par

$$\Delta^h(x) = \begin{cases} \nabla_{y_1} L(f_{\mathbf{w},\mathbf{b}}(x), y) & , \text{ si } h = H \\ \mathbf{W}_{h+1}^T \cdot \text{diag} \left(d_{\mathbf{w},\mathbf{b}}^{h+1}(x) \right) \cdot \Delta^{h+1}(x) & , \text{ sinon} \end{cases}$$

Rq: \mathbf{W}_{h+1}^T remonte d'un niveau dans le réseau: **Rétropropagation**.

Rétropropagation: Problèmes connus

- “Vanishing/Exploding gradient” :

- ▶ Si $\mathbf{W}_{h+1}^T \cdot \text{diag}(d_{\mathbf{w},\mathbf{b}}^{h+1}(x))$ est systématiquement une contraction \implies “Vanishing gradient”
- ▶ Si $\mathbf{W}_{h+1}^T \cdot \text{diag}(d_{\mathbf{w},\mathbf{b}}^{h+1}(x))$ augmente systématiquement la norme \implies “Exploding gradient”
- ▶ Solutions:
 - ★ ResNet : voir précédemment
 - ★ Batch normalisation
 - ★ Contraindre les matrices \mathbf{W}_h à être orthogonales ou presque.

Rétropropagation: Problèmes connus

Dans certaines régions de l'espace, la fonction objectif est très irrégulière :

Pour une fonction d'activation homogène¹ (notamment ReLU):

Pour $\lambda > 0$, $\lambda \sim 0$

$$\mathbf{W}_1^\lambda = \lambda^{H-1} \mathbf{W}_1 \quad \text{et} \quad \mathbf{W}_h^\lambda = \lambda^{-1} \mathbf{W}_h, \forall h = 2..H$$

$$\mathbf{b}_1^\lambda = \lambda^{H-1} \mathbf{b}_1 \quad \text{et} \quad \mathbf{b}_h^\lambda = \lambda^{H-h} \mathbf{b}_h, \forall h = 2..H$$

On a, pour tout x , $f_{\mathbf{w}^\lambda, \mathbf{b}^\lambda}(x) = f_{\mathbf{w}, \mathbf{b}}(x)$, pour $h > 1$

$$\begin{aligned} \frac{\partial E}{\partial \mathbf{W}_{h,i,j}}(\mathbf{w}^\lambda, \mathbf{b}^\lambda) &= \left[f_{\mathbf{w}^\lambda, \mathbf{b}^\lambda}^{h-1}(x) \right]_j \left[d_{\mathbf{w}^\lambda, \mathbf{b}^\lambda}^h(x) \right]_i (\Delta_\lambda^h)_i(x) \\ &= \lambda^{H-(h-1)} \left[f_{\mathbf{w}, \mathbf{b}}^{h-1}(x) \right]_j \left[d_{\mathbf{w}, \mathbf{b}}^h(x) \right]_i \lambda^{-(H-h)} \Delta_i^h(x) \\ &= \lambda \frac{\partial E}{\partial \mathbf{W}_{h,i,j}}(\mathbf{w}, \mathbf{b}) \end{aligned}$$

mais $\frac{\partial E}{\partial \mathbf{b}_{h,i}}(\mathbf{w}^\lambda, \mathbf{b}^\lambda) = \lambda^{-(H-h)} \frac{\partial E}{\partial \mathbf{b}_{h,i}}(\mathbf{w}, \mathbf{b})$ et $\frac{\partial E}{\partial \mathbf{W}_{1,i,j}}(\mathbf{w}^\lambda, \mathbf{b}^\lambda) = \lambda^{-(H-1)} \frac{\partial E}{\partial \mathbf{W}_{1,i,j}}(\mathbf{w}, \mathbf{b})$.

Les petits b_h ont de forts gradients et seront multipliés par des forts W_h .

¹ Attention à la non-différentiabilité en 0

Outline

1 Introduction

- Classification et régression
- Le contrôle du risque
- Les réseaux de neurones
 - Multi-Layer Perceptron/fully-connected neural network
 - Réseaux convolutifs
 - Beaucoup de variantes importantes

2 Questions liées à l'optimisation et la régularisation implicite

- Les réseaux comme une fonction des paramètres
- La Rétropropagation
- **Géométrie locale des réseaux ReLU fully-connected**
- Le paysage de la fonction objectif
 - Introduction
 - Paysage pour les réseaux larges
 - Paysage pour les réseaux linéaires
 - Paysage pour les réseaux linéaires (cas à 1 couche cachée)

Géométrie locale des réseaux ReLU fully-connected.

On note E les arcs du réseaux et B les neurones des couches 1 jusqu'à H .

On note $\theta \in \mathbb{R}^E \times \mathbb{R}^B$ que l'on identifie, si besoin, à

$$(W_1, \dots, W_H, b^1, \dots, b^H) \in (\mathbb{R}^{n_1 \times n_0} \times \dots \times \mathbb{R}^{n_H \times n_{H-1}}) \times (\mathbb{R}^{n_1} \times \dots \times \mathbb{R}^{n_H}).$$

Pour $n \in \mathbb{N}^*$, on définit la fonction **pattern d'activation**

$$\begin{aligned} a: \mathbb{R}^{n_0 \times n} \times (\mathbb{R}^E \times \mathbb{R}^B) &\longrightarrow \{0, 1\}^{(n_1 + \dots + n_{H-1}) \times n} \\ (X, \theta) &\longmapsto a(X, \theta) \end{aligned}$$

où

$$a(X, \theta)_{i,j} = \begin{cases} 1 & \text{si } [W_h f_{h-1}(x^j) + b_h]_v \geq 0 \\ 0 & \text{sinon,} \end{cases}$$

pour v le neurone correspondant à i , $h \in \{1, \dots, H-1\}$ la couche de v , et la j^{ieme} colonne x^j de X .

Pour $X \in \mathbb{R}^{n_0 \times n}$ fixé, elle atteint un nombre fini de valeur que l'on note $\Delta_1^X, \dots, \Delta_{q_X}^X$.

Géométrie locale des réseaux ReLU fully-connected.

On note, pour $X \in \mathbb{R}^{n_0 \times n}$ fixé et pour $j \in \{1, \dots, q_X\}$,

$$\widetilde{\mathcal{U}}_j^X = \text{Int} \left\{ \theta \in \mathbb{R}^E \times \mathbb{R}^B \mid a(X, \theta) = \Delta_j^X \right\},$$

et $m_X = |\{j \in \{1, \dots, q\} \mid \widetilde{\mathcal{U}}_j^X \neq \emptyset\}|$.

Quitte à changer l'ordre, on suppose que les $\widetilde{\mathcal{U}}_1^X, \dots, \widetilde{\mathcal{U}}_{m_X}^X$ sont non-vides.

Lemma (Prédiction polynomiale par morceaux en θ)

Pour tout $n \in \mathbb{N}^*$, $X \in \mathbb{R}^{n_0 \times n}$, les ensembles $\widetilde{\mathcal{U}}_1^X, \dots, \widetilde{\mathcal{U}}_{m_X}^X$ sont non-vides, ouverts et disjoints

- Pour tout $j = 1, \dots, m_X$, $\theta \mapsto f_\theta(X)$ est une fonction polynomiale de degré inférieur à H sur $\widetilde{\mathcal{U}}_j^X$.
- Le complémentaire $\left(\bigcup_{j=1}^{m_X} \widetilde{\mathcal{U}}_j^X \right)^c$ est fermé et de mesure de Lebesgue nulle.

Rq : La fonction $\theta \mapsto f_\theta(X)$ est une composition de fonctions continues, elle est continue.

Géométrie locale des réseaux ReLU fully-connected.

Pour $n \in \mathbb{N}^*$, $X \in \mathbb{R}^{n_0 \times n}$, et $j \in \{1, \dots, m_X\}$, on note

$$r_j^X = \max_{\theta \in \widetilde{\mathcal{U}}_j^X} \text{rk}(\mathbf{D}f_\theta(X)) \quad \text{et} \quad \mathcal{U}_j^X = \{\theta \in \widetilde{\mathcal{U}}_j^X \mid \text{rk}(\mathbf{D}f_\theta(X)) = r_j^X\}.$$

Theorem (Bona-Pellissier, F. Bachoc, F. Malgouyres 2023)

Pour tout réseau ReLU, $n \in \mathbb{N}^*$, $X \in \mathbb{R}^{n_0 \times n}$, par définition

- $\mathcal{U}_1^X, \dots, \mathcal{U}_{m_X}^X$ sont non-vides et disjoints,
- Pour tout $j \in \{1, \dots, m_X\}$, la fonction $\theta \mapsto a(X, \theta)$ est constante sur \mathcal{U}_j^X et prend des valeurs distinctes pour $j' \neq j$;
- Pour tout $j \in \{1, \dots, m_X\}$, $\theta \mapsto \text{rk}(\mathbf{D}f_\theta(X))$ est constante sur \mathcal{U}_j^n et vaut r_j^X .

De plus,

- Les ensembles $\mathcal{U}_1^X, \dots, \mathcal{U}_{m_X}^X$ sont ouverts,
- $\left(\bigcup_{j=1}^{m_X} \mathcal{U}_j^X\right)^c$ est fermé et de mesure de Lebesgue nulle;
- Pour tout $j \in \{1, \dots, m_X\}$, $\theta \mapsto f_\theta(X)$ est une fonction polynomiale de degré inférieur à H sur \mathcal{U}_j^X .

Structure de la fonction objectif

Proposition: Structure de la fonction objectif pour les réseaux linéaires structurés

Pour tout réseau linéaire structuré (i.e. $\sigma = \text{Id}$), pour tout $x \in \mathbb{R}^{n_0}$, pour tout $j = 1 \cdots n_H$, la fonction

$$\begin{aligned} \mathbb{R}^{H \times S} \times \mathbb{R}^{n_1 + \cdots + n_H} &\longrightarrow \mathbb{R} \\ (\mathbf{w}, \mathbf{b}) &\longmapsto [f_{\mathbf{w}, \mathbf{b}}(x)]_j \end{aligned}$$

est un polynôme de degré H .

Pour la régression, si l est quadratique, la fonction E est un polynôme de degré $2H$.

Structure de la fonction objectif

Proposition: Structure de la fonction objectif: le cas ReLU

Pour tout réseau, avec le fonction d'activation ReLU, pour tout $x \in \mathbb{R}^{n_0}$, pour tout $j = 1 \cdots n_H$, la fonction

$$\begin{aligned} \mathbb{R}^{H \times S} \times \mathbb{R}^{n_1 + \cdots + n_H} &\longrightarrow \mathbb{R} \\ (\mathbf{w}, \mathbf{b}) &\longmapsto [f_{\mathbf{w}, \mathbf{b}}(x)]_j \end{aligned}$$

est **continu**. C'est un **polynôme de degré H par morceaux**.

Il existe une partition compatible avec moins de $2^{n_1 + \cdots + n_H}$ morceaux (pas forcément connexes).

Pour la régression, si l est quadratique, la fonction E est **continu et c'est un polynôme de degré $2H$ par morceaux**.

Ex.: Le bord $\partial C = \overline{C} - \overset{\circ}{C}$ d'un morceau C est toujours de mesure nulle.

Outline

1 Introduction

- Classification et régression
- Le contrôle du risque
- Les réseaux de neurones
 - Multi-Layer Perceptron/fully-connected neural network
 - Réseaux convolutifs
 - Beaucoup de variantes importantes

2 Questions liées à l'optimisation et la régularisation implicite

- Les réseaux comme une fonction des paramètres
- La Rétropropagation
- Géométrie locale des réseaux ReLU fully-connected
- **Le paysage de la fonction objectif**
 - Introduction
 - Paysage pour les réseaux larges
 - Paysage pour les réseaux linéaires
 - Paysage pour les réseaux linéaires (cas à 1 couche cachée)

Paysage de la fonction objectif : Introduction

Rappel, pour tout² \mathbf{w} :

$$\begin{aligned} 0 &\leq R(f_{\mathbf{w}}) - R^* && \text{(l'excès de risque)} \\ &= R(f_{\mathbf{w}}) - \widehat{R}(f_{\mathbf{w}}) && \text{(erreur de généralisation)} \\ &+ \widehat{R}(f_{\widehat{\mathbf{w}}}) - \widehat{R}(f_{\widehat{\mathbf{w}}}) && \text{(erreur d'optimisation)} \\ &+ \widehat{R}(f_{\widehat{\mathbf{w}}}) - \widehat{R}(f_{\mathbf{w}^*}) && \leq \varepsilon \\ &+ \widehat{R}(f_{\mathbf{w}^*}) - R(f_{\mathbf{w}^*}) && \text{(erreur de généralisation)} \\ &+ R(f_{\mathbf{w}^*}) - R^* && \text{(erreur d'approximation)} \end{aligned}$$

On utilise un algorithme d'optimisation pour trouver un \mathbf{w} , on veut que

$$\widehat{R}(f_{\mathbf{w}}) - \inf_{\mathbf{w}} \widehat{R}(f_{\mathbf{w}})$$

soit le plus faible possible.

Idéalement, on voudrait que cette quantité soit nulle ou au moins on voudrait la borner supérieurement.

²Pour alléger les notations, on oublie \mathbf{b} .

Paysage de la fonction objectif : Introduction

On suppose que $\mathbf{w} \mapsto \widehat{R}(f_{\mathbf{w}})$ est C^2 partout. On a

$$\widehat{R}(f_{\mathbf{w}}) = \widehat{R}(f_{\mathbf{w}^*}) + \langle \nabla_{\mathbf{w}} \widehat{R}(f_{\mathbf{w}^*}), \mathbf{w} - \mathbf{w}^* \rangle + \frac{1}{2} \langle \nabla_{\mathbf{w}}^2 \widehat{R}(f_{\mathbf{w}^*})(\mathbf{w} - \mathbf{w}^*), \mathbf{w} - \mathbf{w}^* \rangle + o(\|\mathbf{w} - \mathbf{w}^*\|^2)$$

On distingue:

- **\mathbf{w}^* est un minimiseur global:**

- ▶ $\forall \mathbf{w}, \quad \widehat{R}(f_{\mathbf{w}^*}) \leq \widehat{R}(f_{\mathbf{w}})$
- ▶ $\widehat{R}(f_{\mathbf{w}^*}) = \min_{\mathbf{w}} \widehat{R}(f_{\mathbf{w}})$

- **\mathbf{w}^* est un minimiseur local:**

- ▶ Il existe un voisinage ouvert \mathcal{O} de \mathbf{w}^* tel que

$$\forall \mathbf{w} \in \mathcal{O}, \quad \widehat{R}(f_{\mathbf{w}^*}) \leq \widehat{R}(f_{\mathbf{w}})$$

- **\mathbf{w}^* est un point critique du second ordre:**

- ▶ On a

$$\nabla \widehat{R}(f_{\mathbf{w}^*}) = 0 \quad \text{et} \quad \nabla^2 \widehat{R}(f_{\mathbf{w}^*}) \geq 0$$

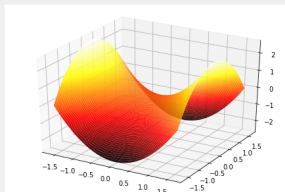
- **\mathbf{w}^* est un point critique du premier ordre:**

- ▶ On a

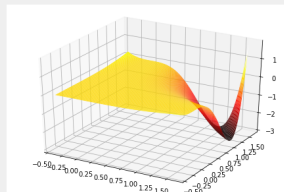
$$\nabla \widehat{R}(f_{\mathbf{w}^*}) = 0$$

Paysage de la fonction objectif : Introduction

- **w^* est un point selle** si c'est un point critique qui n'est ni un minimiseur local, ni un maximiseur local
 - ▶ **Un point selle w^* est strict** : si ce n'est pas un point critique du second ordre (i.e., le Hessian a une v.p. négative).
 - ▶ **Un point selle w^* est non-strict**: si c'est un point critique du second ordre (i.e. le Hessian est semi-défini positif et a une v.p. égale à 0. Typiquement, un terme d'ordre supérieur en fait un point selle.).



(a) Point selle strict



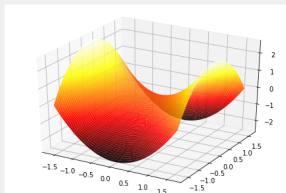
(b) Point selle non-strict

Paysage de la fonction objectif : Introduction

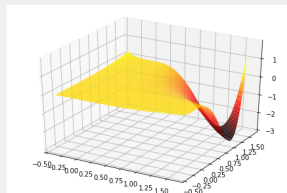
Optimisation non convexe avec les mains

Pour des fonctions non-convexes, on sait montrer que

- **dans un cadre assez vaste**, l'algorithme du gradient (ou gradient stochastique) **converge vers un point critique du premier ordre**
- **dans un cadre plus restreint**, l'algorithme du gradient **converge vers un point critique du second ordre**



(a) Point selle strict



(b) Point selle non-strict

- S. Gadat, F. Panloup, S. Saadane. "Stochastic heavy ball." *Electronic Journal of Statistics* 12.1 (2018): 461-529.
- J. Lee, M. Simchowitz, M. Jordan, B. Recht. "Gradient Descent Converges to Minimizers." *COLT 2016*.

Paysage pour les réseaux larges

Différents énoncés décrits dans

- Gori, Tesi, "On the problem of local minima in backpropagation", IEEE PAMI, 1992
- Yu, Chen, "On the local minima free condition of backpropagation learning", IEEE Trans. Neural Networks, 1995
- Nguyen, Hein, "The loss surface of deep and wide neural networks", ICML, 2017
-

On considère:

- un problème de régression
- un réseau fully-connected
- des observations $(x_i, y_i)_{i=1..n}$:

$$E(\mathbf{w}, \mathbf{b}) = \widehat{R}(\mathbf{f}_{\mathbf{w}, \mathbf{b}}) = \sum_{i=1}^n L(\mathbf{f}_{\mathbf{w}, \mathbf{b}}(x_i) - y_i)$$

Paysage pour les réseaux larges

Le Paysage pour les réseaux larges

On suppose que σ est C^1 et que, pour tout $t \in \mathbb{R}$, $\sigma'(t) \neq 0$. On suppose que le coût L est C^1 , à valeur dans \mathbb{R}^+ et tel que $L(0) = 0$. On suppose aussi que $\nabla L(y) = 0$ si et seulement si $y = 0$. On note $X = [x_1 \cdots x_n] \in \mathbb{R}^{n_0 \times n}$ et $A = \begin{pmatrix} X \\ \mathbb{1}_n^T \end{pmatrix}$.

On considère un point critique du premier ordre (\mathbf{w}, \mathbf{b}) de E .

On suppose que $\text{rang}(A) = n$ et que, pour tout $h = 1..H$, $\text{rang}(\mathbf{W}_h) = n_h$.

Alors, on a

$$\widehat{R}(f_{\mathbf{w}, \mathbf{b}}) = 0$$

et (\mathbf{w}, \mathbf{b}) est un minimum global.

Les hypothèses fortes sont celles sur le rang. Elles impliquent notamment

$$n \leq n_0 + 1 \quad \text{et} \quad n_H \leq n_{H-1} \leq \cdots \leq n_0$$

Nb: Ci-dessus l'indice 0 est arbitraire car on pourrait supposer que les x_i sont le résultat des premières couches d'un réseau.

Pointeurs bibliographiques

- Baldi, Hornik, "Neural networks and principal component analysis: Learning from examples without local minima", Neural networks, 1989.
 - Baldi, Hornik, "Learning in linear neural networks: A survey", IEEE Transactions on neural networks, 1995.
 - Kawaguchi, "Deep learning without poor local minima", NIPS 2016
 - (Notamment dans les groupes de S. Arora à Princeton, de F. Bach à l'ENS)
 - En collaboration avec **E. Achour** (Polytechnique, M2 MVA, actuellement en 3^{ème} année de thèse), **S. Gerchinovitz** (IRT Saint Exupéry)
-
- un problème de régression
 - un réseau fully-connected, linéaire (notamment sans biais)
 - des observations $(x_i, y_i)_{i=1..n}$

Paysage pour les réseaux linéaires

On note

- $X \in \mathbb{R}^{d_x \times n}$ et $Y \in \mathbb{R}^{d_y \times n}$ les matrices contenant les données.

- $\hat{R}(\mathbf{W}) = \sum_{i=1}^m \|W_H W_{H-1} \cdots W_2 W_1 x_i - y_i\|_2^2 = \|W_H \cdots W_1 X - Y\|^2$

-

$$\Sigma_{XX} = \sum_{i=1}^n x_i x_i^T = XX^T \in \mathbb{R}^{d_x \times d_x}, \quad \Sigma_{YY} = \sum_{i=1}^n y_i y_i^T = YY^T \in \mathbb{R}^{d_y \times d_y},$$

$$\Sigma_{XY} = \sum_{i=1}^n x_i y_i^T = XY^T \in \mathbb{R}^{d_x \times d_y}, \quad \Sigma_{YX} = \sum_{i=1}^n y_i x_i^T = YX^T \in \mathbb{R}^{d_y \times d_x},$$

-

$$\Sigma^{1/2} = \Sigma_{YX} \Sigma_{XX}^{-1} X \in \mathbb{R}^{d_y \times n},$$

sa SVD

$$\Sigma^{1/2} = U \Delta V^T,$$

avec $U \in \mathbb{R}^{d_y \times d_y}$, $\Delta = \text{diag}((\delta_i)_{i=1..d_y}) \in \mathbb{R}^{d_y \times n}$, $V \in \mathbb{R}^{n \times n}$.

- $r_{max} = \min(d_x, n_1, \dots, n_{H-1}, d_y)$

Paysage pour les réseaux linéaires

On suppose

- $d_y \leq d_x \leq n$
- Σ_{XX} est inversible et Σ_{XY} est de rang plein
- les valeurs singulières de $\Sigma^{1/2}$ sont distinctes

Lemma (Inférence et valeurs critiques)

Soit \mathbf{W} un point critique du premier ordre de \hat{R} . On pose $r = rk(W_H \cdots W_1)$.
Il existe un unique sous-ensemble $\mathcal{S} \subset \llbracket 1, d_y \rrbracket$ de taille r tel que:

$$W_H \cdots W_1 = U_{\mathcal{S}} U_{\mathcal{S}}^T \Sigma_{YX} \Sigma_{XX}^{-1}.$$

La valeur critique correspondante vaut $\hat{R}(\mathbf{W}) = \text{tr}(\Sigma_{YY}) - \sum_{i \in \mathcal{S}} \delta_i^2$.
On dit que le point critique \mathbf{W} est associé à \mathcal{S} .

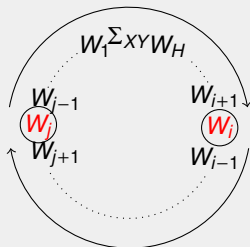
Proposition

Pour tout $\mathcal{S} \subset \llbracket 1, d_y \rrbracket$ de taille $r \in \llbracket 0, r_{max} \rrbracket$, il existe un point critique \mathbf{W} associé à \mathcal{S} .

Paysage pour les réseaux linéaires

- **Pivot** $(i, j) \in \llbracket 1, H \rrbracket^2$, avec $i > j$
- **Blocs complémentaires du pivot** (i, j) :

Premier bloc complémentaire : $W_{j-1} \cdots W_1 \Sigma_{XY} W_H \cdots W_{i+1}$



Second bloc complémentaire : $W_{i-1} \cdots W_{j+1}$

- **Pivot tenu pour W** : L'un des blocs complémentaires est de rang $\text{rk}(W_H \dots W_1)$
- **W est point critique tenu**: W est un point critique et tous les pivots sont tenus pour W

Paysage pour les réseaux linéaires

W p. c. du premier ordre de \hat{R}

$$r := \text{rk}(W_H \cdots W_1)$$

$\exists! \mathcal{S} \subset \llbracket 1, d_Y \rrbracket$ de taille r t.q. $W_H \cdots W_1 = U_{\mathcal{S}} U_{\mathcal{S}}^T \Sigma_{YX} \Sigma_{XX}^{-1}$ et $\hat{R}(W) = \text{tr}(\Sigma_{YY}) - \sum_{i \in \mathcal{S}} \delta_i^2$

$$r = r_{\max}$$

On regarde \mathcal{S}

$$\mathcal{S} = \llbracket 1, r_{\max} \rrbracket$$

W est un minimiseur global

$$\mathcal{S} \neq \llbracket 1, r_{\max} \rrbracket$$

W est un p.s. strict

$$r < r_{\max}$$

W est un point selle

$$\mathcal{S} \neq \llbracket 1, r \rrbracket$$

W est un p.s. strict

$$\mathcal{S} = \llbracket 1, r \rrbracket$$

$$W_H \cdots W_1 = \underset{\text{rk}(R) \leq r}{\text{argmin}} \|RX - Y\|^2$$

W non tenu

W est un p.s. strict

W tenu

W est un p.s. NON strict

Figure: Theorem [E. Achour et al.]

Paysage pour les réseaux linéaires

Proposition (E. Achour et al.)

- Pour $H = 2$, il n'existe pas de point selle non-strict.
- Pour $H \geq 3$, pour tout $r < r_{\max}$, il existe des points critiques associés à $\llbracket 1, r \rrbracket$ tenus et non-tenus.

Proposition (E. Achour et al.)

On note $\mathcal{S}_{\max} = \llbracket 1, r_{\max} \rrbracket$ et $Q_{\max} = \llbracket 1, d_y \rrbracket \setminus \mathcal{S}_{\max} = \llbracket r_{\max} + 1, d_y \rrbracket$.

W est un **minimiseur global** de \widehat{R} si et seulement si il existe des matrices inversibles $D_{H-1} \in \mathbb{R}^{d_{H-1} \times d_{H-1}}, \dots, D_1 \in \mathbb{R}^{d_1 \times d_1}$, et des matrices $A_R \in \mathbb{R}^{(d_y - r_{\max}) \times (d_{H-1} - r_{\max})}$, $(W_h)_{DR} \in \mathbb{R}^{(d_h - r_{\max}) \times (d_{h-1} - r_{\max})}$ pour $h \in \llbracket 2, H-1 \rrbracket$, et $M_D \in \mathbb{R}^{(d_1 - r_{\max}) \times d_x}$ tels que:

$$W_H = [U_{\mathcal{S}_{\max}}, U_{Q_{\max}} A_R] D_{H-1}^{-1}$$

$$W_h = D_h \begin{bmatrix} I_{r_{\max}} & 0 \\ 0 & (W_h)_{DR} \end{bmatrix} D_{h-1}^{-1} \quad \forall h \in \llbracket 2, H-1 \rrbracket$$

$$W_1 = D_1 \begin{bmatrix} U_{\mathcal{S}_{\max}}^T & \Sigma_{YX} \Sigma_{XX}^{-1} \\ & M_D \end{bmatrix}.$$

Paysage pour les réseaux linéaires

Conclusion [E. Achour et al.]

- **Classification** de l'ensemble des points critiques en: minimiseurs globaux; points selles stricts; point selles non-stricts.
- Tout **point critique du second ordre** qui n'est pas un minimiseur global conduit à une solution de la régression linéaire sous **contrainte de rang**.
- Les points selles non-stricts sont associés avec r_{\max} **valeurs plateau** pour le risque empirique
- **Paramétrisation des minimiseurs globaux**

Paysage pour les réseaux linéaires (cas à 1 couche cachée)

On simplifie

- Cas $H = 2$
- On note pour $A \in \mathbb{R}^{n_2 \times n_1}$ et $B \in \mathbb{R}^{n_1 \times n_0}$

$$E(A, B) = \sum_{i=1}^L \|y_i - ABx_i\|^2$$

- On note pour $D \in \mathbb{R}^{n_2 \times n_0}$

$$F(D) = \sum_{i=1}^L \|y_i - Dx_i\|^2$$

- On note

$$\Sigma_{XX} = \sum_{i=1}^L x_i x_i^T \in \mathbb{R}^{n_0 \times n_0} \quad , \quad \Sigma_{XY} = \sum_{i=1}^L x_i y_i^T \in \mathbb{R}^{n_0 \times n_2}$$

$$\Sigma_{YX} = \sum_{i=1}^L y_i x_i^T \in \mathbb{R}^{n_2 \times n_0} \quad , \quad \Sigma_{YY} = \sum_{i=1}^L y_i y_i^T \in \mathbb{R}^{n_2 \times n_2}$$

Paysage pour les réseaux linéaires (cas à 1 couche cachée)

Remarques

On a

- 1 Pour tout $C \in \mathbb{R}^{n_1 \times n_1}$ inversible, $AB = (AC)(C^{-1}B) = A'B'$
- 2 Si Σ_{XX} est inversible alors

$$\Sigma_{YX}\Sigma_{XX}^{-1} \in \operatorname{argmin}_D F(D)$$

- 3 Soit $M \in \mathbb{R}^{n \times p}$ avec $p \leq n$ de rang p . Pour tout $x \in \mathbb{R}^n$, la projection $P_M(x)$ de x sur l'espace vectoriel généré par les colonnes de M vaut

$$P_M(x) = M(M^T M)^{-1} M^T x$$

Paysage pour les réseaux linéaires (cas à 1 couche cachée)

Lemme 1

Si Σ_{XX} est inversible. Soient $A \in \mathbb{R}^{n_2 \times n_1}$ telle que $\text{rang}(A) = n_1$ et $B \in \mathbb{R}^{n_1 \times n_0}$. Alors, (A, B) est un point critique du premier ordre de E si et seulement si

$$AB\Sigma_{XX}B^T = \Sigma_{YX}B^T \quad \text{et} \quad B = (A^T A)^{-1} A^T \Sigma_{YX} \Sigma_{XX}^{-1}$$

Lemme 2

Si Σ_{XX} est inversible. Soient $A \in \mathbb{R}^{n_2 \times n_1}$ telle que $\text{rang}(A) = n_1$ et $B \in \mathbb{R}^{n_1 \times n_0}$. Les deux assertions suivantes sont équivalentes:

① (A, B) est un point critique du premier ordre de E

②

$$AB = P_A \Sigma_{YX} \Sigma_{XX}^{-1}$$

et

$$P_A \Sigma = P_A \Sigma P_A = \Sigma P_A$$

pour

$$\Sigma = \Sigma_{YX} \Sigma_{XX}^{-1} \Sigma_{XY} \in \mathbb{R}^{n_2 \times n_2}$$

Paysage pour les réseaux linéaires (cas à 1 couche cachée)

On diagonalise (Σ est symétrique)

$$\Sigma = U\Lambda U^T$$

avec $\Lambda \in \mathbb{R}^{n_2 \times n_2}$ diagonale et $U \in \mathbb{R}^{n_2 \times n_2}$ unitaire.

Pour $\mathcal{S} \subset \{1, \dots, n_2\}$, on note $U_{\mathcal{S}}$ la matrice extraite de U en prenant les colonnes d'indice dans \mathcal{S} .

Proposition 1: Paramétrisation des points critiques

Si Σ_{XX} est inversible. Soient $A \in \mathbb{R}^{n_2 \times n_1}$ et $B \in \mathbb{R}^{n_1 \times n_0}$ avec A telle que $\text{rang}(A) = n_1$.

On suppose que les valeurs propres de Σ sont distinctes.

Alors, (A, B) est un point critique du premier ordre de E si et seulement si il existe $C \in \mathbb{R}^{n_1 \times n_1}$ inversible et $\mathcal{S} \subset \{1, \dots, n_2\}$ de taille n_1 tels que

$$A = U_{\mathcal{S}} C \quad \text{et} \quad B = C^{-1} U_{\mathcal{S}}^T \Sigma_{YX} \Sigma_{XX}^{-1}$$

Paysage pour les réseaux linéaires (cas à 1 couche cachée)

Pour simplifier les notations, on suppose les valeurs propres de Σ ordonnées:

$$\lambda_1 > \lambda_2 > \dots > \lambda_{n_2}$$

Proposition 2: minimiseur global \Leftrightarrow minimiseur local

Si Σ_{XX} est inversible. Soient $A \in \mathbb{R}^{n_2 \times n_1}$ et $B \in \mathbb{R}^{n_1 \times n_0}$ avec A telle que $\text{rang}(A) = n_1$.

On suppose que les valeurs propres de Σ sont distinctes.

- 1 pour tout point critique du premier ordre (A, B) de E et pour \mathcal{S} définissant A et B (voir la proposition précédente), on a

1

$$AB = P_{U_{\mathcal{S}}} \Sigma_{YX} \Sigma_{XX}^{-1}$$

2

$$E(A, B) = \text{trace}(\Sigma_{YY}) - \sum_{i \in \mathcal{S}} \lambda_i$$

- 2 Si (A, B) est un minimiseur global alors c'est un point critique du premier ordre associé à $\mathcal{S} = \{1, \dots, n_1\}$.

- 3 **(A, B) est minimiseur local si et seulement si c'est un minimiseur global.**

Merci pour votre attention !