Une Analyse des Méthodes de Projections Aléatoires par la Théorie des Matrices Aléatoires

Colloque GRETSI 2017

Zhenyu Liao, Romain Couillet

CentraleSupélec
Université Paris-Saclay

Juan-Les-Pins, 05 Septembre, 2017

CentraleSupélec

Plan

Plan

Motivation: Extraire des features en utilisant des projections aléatoires

Apprentissage = Représentation + Évaluation + Optimisation.[1]

[1] Domingos, Pedro. "A few useful things to know about machine learning." Communications of the ACM 55.10 (2012): 78-87.

Motivation: Extraire des features en utilisant des projections aléatoires

$$\text{Apprentissage} = \text{Représentation} + \text{Évaluation} + \text{Optimisation}.^{1}$$

Comment trouver les "bonnes" features?

- apprendre les features à partir des données:

[1] Domingos, Pedro. "A few useful things to know about machine learning." Communications of the ACM 55.10 (2012): 78-87.

Motivation: Extraire des features en utilisant des projections aléatoires

$$\text{Apprentissage} = \text{Représentation} + \text{Évaluation} + \text{Optimisation}.^{1}$$

Comment trouver les "bonnes" features?

- apprendre les features à partir des données:
  - réseaux de neurones: backprop coûteux

[1] Domingos, Pedro. "A few useful things to know about machine learning." Communications of the ACM 55.10 (2012): 78-87.

Motivation: Extraire des features en utilisant des projections aléatoires

$$\text{Apprentissage} = \text{Représentation} + \text{Évaluation} + \text{Optimisation}.^{1}$$

Comment trouver les "bonnes" features?

- apprendre les features à partir des données:
  - réseaux de neurones: backprop coûteux
  - problème de sur-apprentissage: jeux de données très grands

[1] Domingos, Pedro. "A few useful things to know about machine learning." Communications of the ACM 55.10 (2012): 78-87.

Motivation: Extraire des features en utilisant des projections aléatoires

$$\text{Apprentissage} = \text{Représentation} + \text{Évaluation} + \text{Optimisation}.^{1}$$

Comment trouver les "bonnes" features?

- apprendre les features à partir des données:
  - réseaux de neurones: backprop coûteux
  - problème de sur-apprentissage: jeux de données très grands
  - absence de compréhension théorique

[1] Domingos, Pedro. "A few useful things to know about machine learning." Communications of the ACM 55.10 (2012): 78-87.

Motivation: Extraire des features en utilisant des projections aléatoires

$$\text{Apprentissage} = \text{Représentation} + \text{Évaluation} + \text{Optimisation}.^{1}$$

Comment trouver les "bonnes" features?

- apprendre les features à partir des données:
  - réseaux de neurones: backprop coûteux
  - problème de sur-apprentissage: jeux de données très grands
  - absence de compréhension théorique
- méthodes aléatoires: projections aléatoires, compressive sensing

[1] Domingos, Pedro. "A few useful things to know about machine learning." Communications of the ACM 55.10 (2012): 78-87.

Motivation: Extraire des features en utilisant des projections aléatoires

$$\text{Apprentissage} = \text{Représentation} + \text{Évaluation} + \text{Optimisation}.^{1}$$

Comment trouver les "bonnes" features?

- apprendre les features à partir des données:
  - réseaux de neurones: backprop coûteux
  - problème de sur-apprentissage: jeux de données très grands
  - absence de compréhension théorique
- méthodes aléatoires: projections aléatoires, compressive sensing
  - simples et rapides

[1] Domingos, Pedro. "A few useful things to know about machine learning." Communications of the ACM 55.10 (2012): 78-87.

Motivation: Extraire des features en utilisant des projections aléatoires

$$\text{Apprentissage} = \text{Représentation} + \text{Évaluation} + \text{Optimisation}.^{1}$$

Comment trouver les "bonnes" features?

- apprendre les features à partir des données:
  - réseaux de neurones: backprop coûteux
  - problème de sur-apprentissage: jeux de données très grands
  - absence de compréhension théorique
- méthodes aléatoires: projections aléatoires, compressive sensing
  - simples et rapides
  - adaptées aux applications en ligne (e.g., moindres carrées récursifs)

[1] Domingos, Pedro. "A few useful things to know about machine learning." Communications of the ACM 55.10 (2012): 78-87.

Motivation: Extraire des features en utilisant des projections aléatoires

$$\text{Apprentissage} = \text{Représentation} + \text{Évaluation} + \text{Optimisation}.^{1}$$

Comment trouver les "bonnes" features?

- apprendre les features à partir des données:
  - réseaux de neurones: backprop coûteux
  - problème de sur-apprentissage: jeux de données très grands
  - absence de compréhension théorique
- méthodes aléatoires: projections aléatoires, compressive sensing
  - simples et rapides
  - adaptées aux applications en ligne (e.g., moindres carrées récursifs)
  - analyse théorique possible

[1] Domingos, Pedro. "A few useful things to know about machine learning." Communications of the ACM 55.10 (2012): 78-87.

Motivation: Extraire des features en utilisant des projections aléatoires

$$\text{Apprentissage} = \text{Représentation} + \text{Évaluation} + \text{Optimisation}.^{1}$$

Comment trouver les "bonnes" features?

- apprendre les features à partir des données:
  - réseaux de neurones: backprop coûteux
  - problème de sur-apprentissage: jeux de données très grands
  - absence de compréhension théorique
- méthodes aléatoires: projections aléatoires, compressive sensing
  - simples et rapides
  - adaptées aux applications en ligne (e.g., moindres carrées récursifs)
  - analyse théorique possible
  - mais difficultés théoriques dans le cas non linéaire

[1] Domingos, Pedro. "A few useful things to know about machine learning." Communications of the ACM 55.10 (2012): 78-87.

Motivation: Extraire des features en utilisant des projections aléatoires

$$\text{Apprentissage} = \text{Représentation} + \text{Évaluation} + \text{Optimisation}.^{1}$$

Comment trouver les "bonnes" features?

- apprendre les features à partir des données:
  - réseaux de neurones: backprop coûteux
  - problème de sur-apprentissage: jeux de données très grands
  - absence de compréhension théorique
- méthodes aléatoires: projections aléatoires, compressive sensing
  - simples et rapides
  - adaptées aux applications en ligne (e.g., moindres carrées récursifs)
  - analyse théorique possible
  - mais difficultés théoriques dans le cas non linéaire

Exemples dans la littérature:

- *random features maps*: projections non linéaires aléatoires

[1] Domingos, Pedro. "A few useful things to know about machine learning." Communications of the ACM 55.10 (2012): 78-87.

Motivation: Extraire des features en utilisant des projections aléatoires

$$\text{Apprentissage} = \text{Représentation} + \text{Évaluation} + \text{Optimisation}.^{1}$$

Comment trouver les "bonnes" features?

- apprendre les features à partir des données:
  - réseaux de neurones: backprop coûteux
  - problème de sur-apprentissage: jeux de données très grands
  - absence de compréhension théorique
- méthodes aléatoires: projections aléatoires, compressive sensing
  - simples et rapides
  - adaptées aux applications en ligne (e.g., moindres carrées récursifs)
  - analyse théorique possible
  - mais difficultés théoriques dans le cas non linéaire

Exemples dans la littérature:

- *random features maps*: projections non linéaires aléatoires
- *extreme learning machine (ELM)*: réseaux de neurones aléatoires simples

[1] Domingos, Pedro. "A few useful things to know about machine learning." Communications of the ACM 55.10 (2012): 78-87.

Motivation: Extraire des features en utilisant des projections aléatoires

$$\text{Apprentissage} = \text{Représentation} + \text{Évaluation} + \text{Optimisation}.^{1}$$

Comment trouver les "bonnes" features?

- apprendre les features à partir des données:
  - réseaux de neurones: backprop coûteux
  - problème de sur-apprentissage: jeux de données très grands
  - absence de compréhension théorique
- méthodes aléatoires: projections aléatoires, compressive sensing
  - simples et rapides
  - adaptées aux applications en ligne (e.g., moindres carrées récursifs)
  - analyse théorique possible
  - mais difficultés théoriques dans le cas non linéaire

Exemples dans la littérature:

- *random features maps*: projections non linéaires aléatoires
- *extreme learning machine (ELM)*: réseaux de neurones aléatoires simples
- *echo state nets*: réseaux de neurones récurrents simples

[1] Domingos, Pedro. "A few useful things to know about machine learning." Communications of the ACM 55.10 (2012): 78-87.

Motivation: Extraire des features en utilisant des projections aléatoires

$$\text{Apprentissage} = \text{Représentation} + \text{Évaluation} + \text{Optimisation}.^{1}$$

Comment trouver les "bonnes" features?

- apprendre les features à partir des données:
  - réseaux de neurones: backprop coûteux
  - problème de sur-apprentissage: jeux de données très grands
  - absence de compréhension théorique
- méthodes aléatoires: projections aléatoires, compressive sensing
  - simples et rapides
  - adaptées aux applications en ligne (e.g., moindres carrées récursifs)
  - analyse théorique possible
  - mais difficultés théoriques dans le cas non linéaire

Exemples dans la littérature:

- *random features maps*: projections non linéaires aléatoires
- *extreme learning machine (ELM)*: réseaux de neurones aléatoires simples
- *echo state nets*: réseaux de neurones récurrents simples

Dans ce travail:

- analyse spectrale des matrices de projections aléatoires non linéaire en grande dimension

[1] Domingos, Pedro. "A few useful things to know about machine learning." Communications of the ACM 55.10 (2012): 78-87.

Motivation: Extraire des features en utilisant des projections aléatoires

$$\text{Apprentissage} = \text{Représentation} + \text{Évaluation} + \text{Optimisation}.^{1}$$

Comment trouver les "bonnes" features?

- apprendre les features à partir des données:
  - réseaux de neurones: backprop coûteux
  - problème de sur-apprentissage: jeux de données très grands
  - absence de compréhension théorique
- méthodes aléatoires: projections aléatoires, compressive sensing
  - simples et rapides
  - adaptées aux applications en ligne (e.g., moindres carrées récursifs)
  - analyse théorique possible
  - mais difficultés théoriques dans le cas non linéaire

Exemples dans la littérature:

- *random features maps*: projections non linéaires aléatoires
- *extreme learning machine (ELM)*: réseaux de neurones aléatoires simples
- *echo state nets*: réseaux de neurones récurrents simples

Dans ce travail:

- analyse spectrale des matrices de projections aléatoires non linéaire en grande dimension
- application à l'analyse des performances asymptotiques de l'ELM

[1] Domingos, Pedro. "A few useful things to know about machine learning." Communications of the ACM 55.10 (2012): 78-87.

Projections aléatoires: une exemple

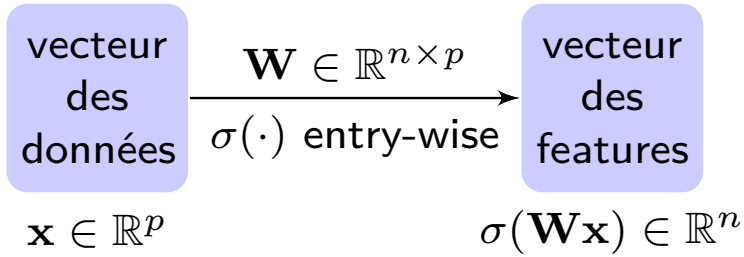

Figure: Illustration de la méthode de projections aléatoires

avec la matrice de poids $\mathbf{W} \in \mathbb{R}^{n \times p}$ telle que les $\mathbf{W}_{ij}$'s sont des variables aléatoires i.i.d. et la *fonction d'activation* $\sigma(\cdot)$ appliquée élément par élément.

Projections aléatoires: une exemple

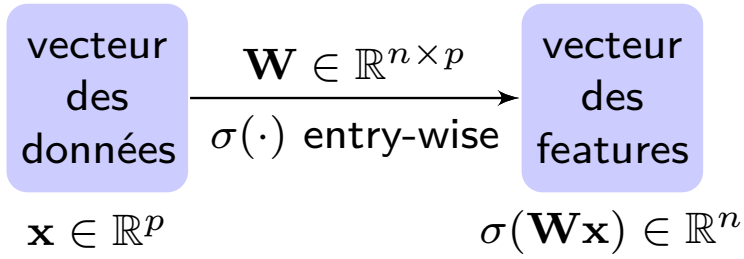

Figure: Illustration de la méthode de projections aléatoires

avec la matrice de poids $\mathbf{W} \in \mathbb{R}^{n \times p}$ telle que les $\mathbf{W}_{ij}$'s sont des variables aléatoires i.i.d. et la *fonction d'activation* $\sigma(\cdot)$ appliquée élément par élément.

Exemple: Classification des spams

$$\underbrace{\begin{bmatrix} 0 & 0 & 0 & 1 & 1 \\ 0 & 1 & 0 & 0 & 0 \\ 1 & 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 1 & 0 \end{bmatrix}}_{\mathbf{W}} \underbrace{\begin{bmatrix} \textit{Cadeaux} \\ \textit{gratuits} \\ \textit{pour} \\ \textit{votre} \\ \textit{famille} \end{bmatrix}}_{\mathbf{x}} \xrightarrow{\sigma(t)=t} \underbrace{\begin{bmatrix} \textit{votre+famille} \\ \textit{Cadeaux} \\ \textit{Cadeaux+gratuits} \\ \textit{pour+famille} \end{bmatrix}}_{\sigma(\mathbf{W}\mathbf{x})}$$

Quand $n, p \to \infty$, on obtient de "bonnes" features, ensuite un prédicteur peut être appris de manière supervisée, e.g., en appliquant une régression normalisée.

Comprendre la méthode de projections aléatoires

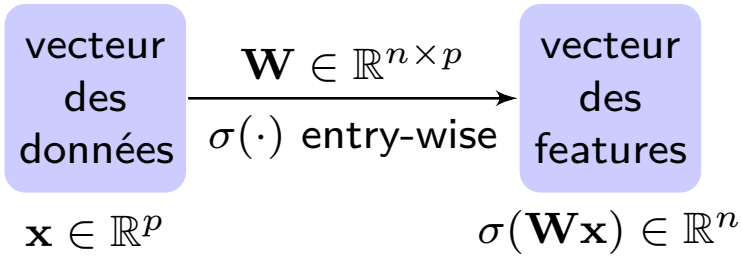

Figure: Illustration de la méthode de projections aléatoires

Comprendre la méthode de projections aléatoires

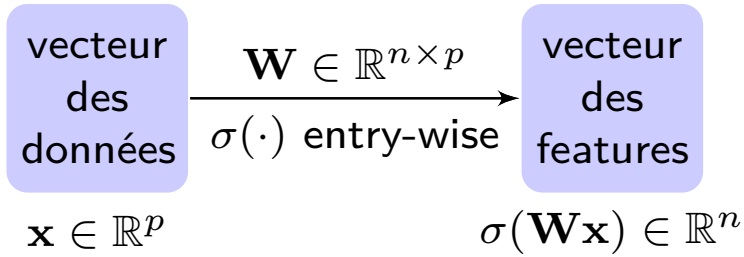

Figure: Illustration de la méthode de projections aléatoires

- Pour matrice de données $\mathbf{X} = [\mathbf{x}_1, \ldots, \mathbf{x}_T] \in \mathbb{R}^{p \times T}$, on note la matrice de feature $\mathbf{\Sigma} \equiv \sigma(\mathbf{W}\mathbf{X}) = [\sigma(\mathbf{W}\mathbf{x}_1), \ldots, \sigma(\mathbf{W}\mathbf{x}_T)] \in \mathbb{R}^{n \times T}$

Comprendre la méthode de projections aléatoires

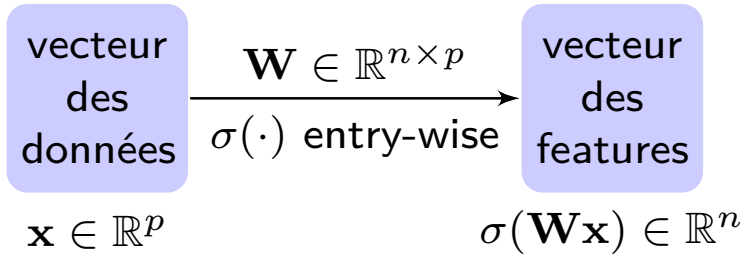

Figure: Illustration de la méthode de projections aléatoires

- Pour matrice de données $\mathbf{X} = [\mathbf{x}_1, \ldots, \mathbf{x}_T] \in \mathbb{R}^{p \times T}$, on note la matrice de feature $\mathbf{\Sigma} \equiv \sigma(\mathbf{W}\mathbf{X}) = [\sigma(\mathbf{W}\mathbf{x}_1), \ldots, \sigma(\mathbf{W}\mathbf{x}_T)] \in \mathbb{R}^{n \times T}$
- Objet clé: la matrice de Gram $\mathbf{G} = \frac{1}{T}\mathbf{\Sigma}^\mathsf{T}\mathbf{\Sigma}$: matrice de corrélation dans l'espace "features"

Comprendre la méthode de projections aléatoires

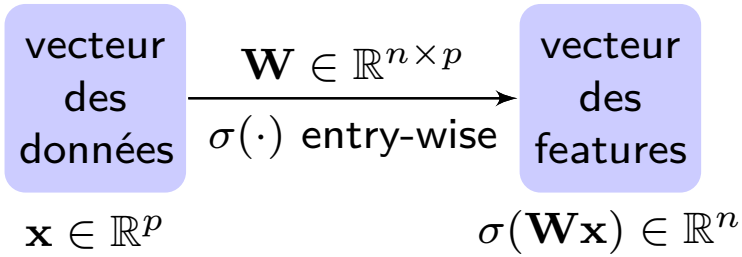

Figure: Illustration de la méthode de projections aléatoires

- Pour matrice de données $\mathbf{X} = [\mathbf{x}_1, \ldots, \mathbf{x}_T] \in \mathbb{R}^{p \times T}$, on note la matrice de feature $\mathbf{\Sigma} \equiv \sigma(\mathbf{W}\mathbf{X}) = [\sigma(\mathbf{W}\mathbf{x}_1), \ldots, \sigma(\mathbf{W}\mathbf{x}_T)] \in \mathbb{R}^{n \times T}$
- Objet clé: la matrice de Gram $\mathbf{G} = \frac{1}{T}\mathbf{\Sigma}^\mathsf{T}\mathbf{\Sigma}$: matrice de corrélation dans l'espace "features"
- RMT $\Rightarrow$ analyse de la résolvante associée

$$\mathbf{Q}(z) = (\mathbf{G} - z\mathbf{I}_T)^{-1} = \left(\frac{1}{T}\mathbf{\Sigma}^\mathsf{T}\mathbf{\Sigma} - z\mathbf{I}_T\right)^{-1}, \; z \in \mathbb{C} \setminus \mathbb{R}^+$$

donne accès à:

Comprendre la méthode de projections aléatoires

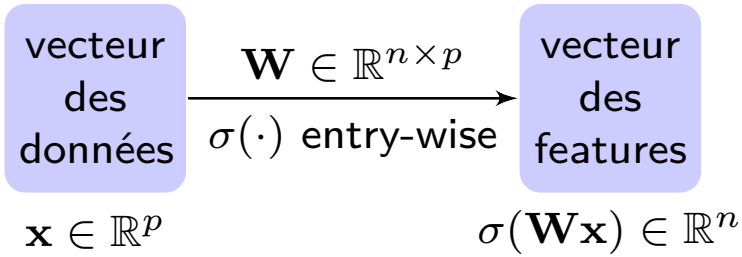

Figure: Illustration de la méthode de projections aléatoires

- Pour matrice de données $\mathbf{X} = [\mathbf{x}_1, \ldots, \mathbf{x}_T] \in \mathbb{R}^{p \times T}$, on note la matrice de feature $\mathbf{\Sigma} \equiv \sigma(\mathbf{W}\mathbf{X}) = [\sigma(\mathbf{W}\mathbf{x}_1), \ldots, \sigma(\mathbf{W}\mathbf{x}_T)] \in \mathbb{R}^{n \times T}$
- Objet clé: la matrice de Gram $\mathbf{G} = \frac{1}{T}\mathbf{\Sigma}^\mathsf{T}\mathbf{\Sigma}$: matrice de corrélation dans l'espace "features"
- RMT $\Rightarrow$ analyse de la résolvante associée

$$\mathbf{Q}(z) = (\mathbf{G} - z\mathbf{I}_T)^{-1} = \left(\frac{1}{T}\mathbf{\Sigma}^\mathsf{T}\mathbf{\Sigma} - z\mathbf{I}_T\right)^{-1}, \; z \in \mathbb{C} \setminus \mathbb{R}^+$$

donne accès à:

- valeurs propres de $\mathbf{G}$ (via transformée de Stieltjes $\frac{1}{T}\operatorname{tr}\mathbf{Q}$)

Comprendre la méthode de projections aléatoires

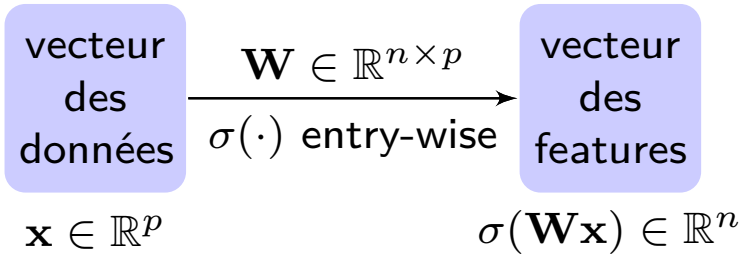

Figure: Illustration de la méthode de projections aléatoires

- Pour matrice de données $\mathbf{X} = [\mathbf{x}_1, \ldots, \mathbf{x}_T] \in \mathbb{R}^{p \times T}$, on note la matrice de feature $\mathbf{\Sigma} \equiv \sigma(\mathbf{W}\mathbf{X}) = [\sigma(\mathbf{W}\mathbf{x}_1), \ldots, \sigma(\mathbf{W}\mathbf{x}_T)] \in \mathbb{R}^{n \times T}$
- Objet clé: la matrice de Gram $\mathbf{G} = \frac{1}{T}\mathbf{\Sigma}^\mathsf{T}\mathbf{\Sigma}$: matrice de corrélation dans l'espace "features"
- RMT $\Rightarrow$ analyse de la résolvante associée

$$\mathbf{Q}(z) = (\mathbf{G} - z\mathbf{I}_T)^{-1} = \left(\frac{1}{T}\mathbf{\Sigma}^\mathsf{T}\mathbf{\Sigma} - z\mathbf{I}_T\right)^{-1}, \; z \in \mathbb{C} \setminus \mathbb{R}^+$$

donne accès à:

- valeurs propres de $\mathbf{G}$ (via transformée de Stieltjes $\frac{1}{T}\operatorname{tr}\mathbf{Q}$)
- vecteurs propres isolés de $\mathbf{G}$ (via intégrale complexe sur $\mathbf{Q}$)

Comprendre la méthode de projections aléatoires

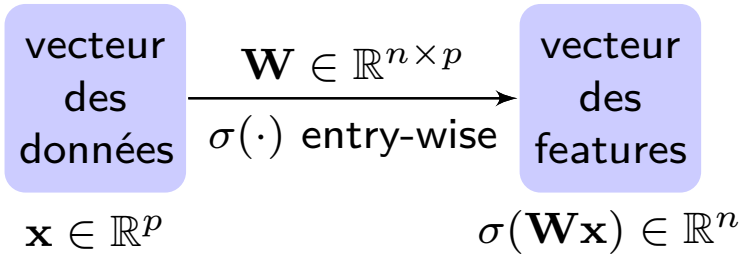

Figure: Illustration de la méthode de projections aléatoires

- Pour matrice de données $\mathbf{X} = [\mathbf{x}_1, \ldots, \mathbf{x}_T] \in \mathbb{R}^{p \times T}$, on note la matrice de feature $\mathbf{\Sigma} \equiv \sigma(\mathbf{W}\mathbf{X}) = [\sigma(\mathbf{W}\mathbf{x}_1), \ldots, \sigma(\mathbf{W}\mathbf{x}_T)] \in \mathbb{R}^{n \times T}$
- Objet clé: la matrice de Gram $\mathbf{G} = \frac{1}{T}\mathbf{\Sigma}^\mathsf{T}\mathbf{\Sigma}$: matrice de corrélation dans l'espace "features"
- RMT $\Rightarrow$ analyse de la résolvante associée

$$\mathbf{Q}(z) = (\mathbf{G} - z\mathbf{I}_T)^{-1} = \left(\frac{1}{T}\mathbf{\Sigma}^\mathsf{T}\mathbf{\Sigma} - z\mathbf{I}_T\right)^{-1}, \; z \in \mathbb{C} \setminus \mathbb{R}^+$$

donne accès à:

- valeurs propres de $\mathbf{G}$ (via transformée de Stieltjes $\frac{1}{T}\operatorname{tr}\mathbf{Q}$)
- vecteurs propres isolés de $\mathbf{G}$ (via intégrale complexe sur $\mathbf{Q}$)

- dans le régime où $n, p, T \to \infty$, on cherche un équivalent déterministe asymptotique de $\mathbf{Q}$

Plan

Hypothèses

Hypothèse 1: Taux de Croissance

Lorsque $n \to \infty$, on a

- $\frac{p}{n} \to c_0 \in (0, \infty)$

Hypothèses

Hypothèse 1: Taux de Croissance

Lorsque $n \to \infty$, on a

- $\frac{p}{n} \to c_0 \in (0, \infty)$
- $\frac{T}{n} \to c_T \in (0, \infty)$

Hypothèses

Hypothèse 1: Taux de Croissance

Lorsque $n \to \infty$, on a

- $\frac{p}{n} \to c_0 \in (0, \infty)$
- $\frac{T}{n} \to c_T \in (0, \infty)$
- $\|\mathbf{X}\| = O(1)$ [norme opérateur]

Hypothèses

Hypothèse 1: Taux de Croissance

Lorsque $n \to \infty$, on a

- $\frac{p}{n} \to c_0 \in (0, \infty)$
- $\frac{T}{n} \to c_T \in (0, \infty)$
- $\|\mathbf{X}\| = O(1)$ [norme opérateur]

Hypothèse 2: Fonction d'activation

La fonction σ est Lipschitzienne avec constante de Lipschitz λ_σ indépendante de n.

Hypothèses

Hypothèse 1: Taux de Croissance

Lorsque $n \to \infty$, on a

- $\frac{p}{n} \to c_0 \in (0, \infty)$
- $\frac{T}{n} \to c_T \in (0, \infty)$
- $\|\mathbf{X}\| = O(1)$ [norme opérateur]

Hypothèse 2: Fonction d'activation

La fonction σ est Lipschitzienne avec constante de Lipschitz λ_σ indépendante de n.

Hypothèse 3: $\mathbf{W}$ Gaussienne

La matrice $\mathbf{W}$ a ses entrées i.i.d. $\mathcal{N}(0, 1)$.

Équivalent déterministe de $\mathbf{Q}$ (1)

Theorem (Équivalent déterministe de $\mathbf{Q}$)

Sous hypothèses 1 - 3, lorsque $n \to \infty$, on a

$$\|\mathbb{E}[\mathbf{Q}] - \bar{\mathbf{Q}}\| \to 0$$

où $\bar{\mathbf{Q}}$ est donnée par

$$\bar{\mathbf{Q}} \equiv \left(\frac{n}{T} \frac{\mathbf{\Phi}}{1+\delta} - z\mathbf{I}_T \right)^{-1}$$

avec δ l'unique solution positive $\delta = \frac{1}{T}\operatorname{tr}(\mathbf{\Phi}\bar{\mathbf{Q}})$ et $\mathbf{\Phi} \equiv \mathbb{E}[\sigma(\mathbf{X}^\mathsf{T}\mathbf{w})\sigma(\mathbf{w}^\mathsf{T}\mathbf{X})]$ pour $\mathbf{w} \sim \mathcal{N}(\mathbf{0}, \mathbf{I}_p)$.

Équivalent déterministe de $\mathbf{Q}$ (1)

Theorem (Équivalent déterministe de $\mathbf{Q}$)

Sous hypothèses 1 - 3, lorsque $n \to \infty$, on a

$$\|\mathbb{E}[\mathbf{Q}] - \bar{\mathbf{Q}}\| \to 0$$

où $\bar{\mathbf{Q}}$ est donnée par

$$\bar{\mathbf{Q}} \equiv \left(\frac{n}{T} \frac{\mathbf{\Phi}}{1+\delta} - z\mathbf{I}_T \right)^{-1}$$

avec δ l'unique solution positive $\delta = \frac{1}{T}\operatorname{tr}(\mathbf{\Phi}\bar{\mathbf{Q}})$ et $\mathbf{\Phi} \equiv \mathbb{E}[\sigma(\mathbf{X}^\mathsf{T}\mathbf{w})\sigma(\mathbf{w}^\mathsf{T}\mathbf{X})]$ pour $\mathbf{w} \sim \mathcal{N}(\mathbf{0}, \mathbf{I}_p)$.

Intuition:

- pour $\sigma(t) = t$, [Silverstein et Bai'95]: $\|\mathbb{E}[\mathbf{Q}] - \bar{\mathbf{Q}}\| \to 0$, $\bar{\mathbf{Q}} \equiv \left(\frac{n}{T} \frac{\mathbf{X}^\mathsf{T}\mathbf{X}}{1+\delta} - z\mathbf{I}_T \right)^{-1}$

Équivalent déterministe de $\mathbf{Q}$ (1)

Theorem (Équivalent déterministe de $\mathbf{Q}$)

Sous hypothèses 1 - 3, lorsque $n \to \infty$, on a

$$\|\mathbb{E}[\mathbf{Q}] - \bar{\mathbf{Q}}\| \to 0$$

où $\bar{\mathbf{Q}}$ est donnée par

$$\bar{\mathbf{Q}} \equiv \left(\frac{n}{T} \frac{\mathbf{\Phi}}{1+\delta} - z\mathbf{I}_T \right)^{-1}$$

avec δ l'unique solution positive $\delta = \frac{1}{T}\operatorname{tr}(\mathbf{\Phi}\bar{\mathbf{Q}})$ et $\mathbf{\Phi} \equiv \mathbb{E}[\sigma(\mathbf{X}^\mathsf{T}\mathbf{w})\sigma(\mathbf{w}^\mathsf{T}\mathbf{X})]$ pour $\mathbf{w} \sim \mathcal{N}(\mathbf{0}, \mathbf{I}_p)$.

Intuition:

- pour $\sigma(t) = t$, [Silverstein et Bai'95]: $\|\mathbb{E}[\mathbf{Q}] - \bar{\mathbf{Q}}\| \to 0$, $\bar{\mathbf{Q}} \equiv \left(\frac{n}{T} \frac{\mathbf{X}^\mathsf{T}\mathbf{X}}{1+\delta} - z\mathbf{I}_T \right)^{-1}$
- repose essentiellement sur lemme de la trace: pour $\mathbf{A}$ de norme borné et indépendante de $\mathbf{W}$,

$$\left| \frac{1}{n}\mathbf{w}_i^\mathsf{T}\mathbf{A}\mathbf{w}_i - \frac{1}{n}\operatorname{tr}\mathbf{A} \right| \to 0$$

Équivalent déterministe de $\mathbf{Q}$ (1)

Theorem (Équivalent déterministe de $\mathbf{Q}$)

Sous hypothèses 1 - 3, lorsque $n \to \infty$, on a

$$\|\mathbb{E}[\mathbf{Q}] - \bar{\mathbf{Q}}\| \to 0$$

où $\bar{\mathbf{Q}}$ est donnée par

$$\bar{\mathbf{Q}} \equiv \left(\frac{n}{T} \frac{\mathbf{\Phi}}{1+\delta} - z\mathbf{I}_T \right)^{-1}$$

avec δ l'unique solution positive $\delta = \frac{1}{T}\operatorname{tr}(\mathbf{\Phi}\bar{\mathbf{Q}})$ et $\mathbf{\Phi} \equiv \mathbb{E}[\sigma(\mathbf{X}^\mathsf{T}\mathbf{w})\sigma(\mathbf{w}^\mathsf{T}\mathbf{X})]$ pour $\mathbf{w} \sim \mathcal{N}(\mathbf{0}, \mathbf{I}_p)$.

Intuition:

- pour $\sigma(t) = t$, [Silverstein et Bai'95]: $\|\mathbb{E}[\mathbf{Q}] - \bar{\mathbf{Q}}\| \to 0$, $\bar{\mathbf{Q}} \equiv \left(\frac{n}{T} \frac{\mathbf{X}^\mathsf{T}\mathbf{X}}{1+\delta} - z\mathbf{I}_T \right)^{-1}$
- repose essentiellement sur lemme de la trace: pour $\mathbf{A}$ de norme borné et indépendante de $\mathbf{W}$,

$$\left| \frac{1}{n}\mathbf{w}_i^\mathsf{T}\mathbf{A}\mathbf{w}_i - \frac{1}{n}\operatorname{tr}\mathbf{A} \right| \to 0$$

  - ✓ $\mathbf{w}_i^\mathsf{T}$ (colonne de $\mathbf{W}$) indépendants

Équivalent déterministe de $\mathbf{Q}$ (1)

Theorem (Équivalent déterministe de $\mathbf{Q}$)

Sous hypothèses 1 - 3, lorsque $n \to \infty$, on a

$$\|\mathbb{E}[\mathbf{Q}] - \bar{\mathbf{Q}}\| \to 0$$

où $\bar{\mathbf{Q}}$ est donnée par

$$\bar{\mathbf{Q}} \equiv \left(\frac{n}{T} \frac{\mathbf{\Phi}}{1+\delta} - z\mathbf{I}_T \right)^{-1}$$

avec δ l'unique solution positive $\delta = \frac{1}{T}\operatorname{tr}(\mathbf{\Phi}\bar{\mathbf{Q}})$ et $\mathbf{\Phi} \equiv \mathbb{E}[\sigma(\mathbf{X}^\mathsf{T}\mathbf{w})\sigma(\mathbf{w}^\mathsf{T}\mathbf{X})]$ pour $\mathbf{w} \sim \mathcal{N}(\mathbf{0}, \mathbf{I}_p)$.

Intuition:

- pour $\sigma(t) = t$, [Silverstein et Bai'95]: $\|\mathbb{E}[\mathbf{Q}] - \bar{\mathbf{Q}}\| \to 0$, $\bar{\mathbf{Q}} \equiv \left(\frac{n}{T} \frac{\mathbf{X}^\mathsf{T}\mathbf{X}}{1+\delta} - z\mathbf{I}_T \right)^{-1}$
- repose essentiellement sur lemme de la trace: pour $\mathbf{A}$ de norme borné et indépendante de $\mathbf{W}$,

$$\left| \frac{1}{n}\mathbf{w}_i^\mathsf{T}\mathbf{A}\mathbf{w}_i - \frac{1}{n}\operatorname{tr}\mathbf{A} \right| \to 0$$

  - ✓ $\mathbf{w}_i^\mathsf{T}$ (colonne de $\mathbf{W}$) indépendants
  - ✓ $\mathbf{w}_i$ à entrées i.i.d. $\mathcal{N}(0,1)$,

Équivalent déterministe de $\mathbf{Q}$ (1)

Theorem (Équivalent déterministe de $\mathbf{Q}$)

Sous hypothèses 1 - 3, lorsque $n \to \infty$, on a

$$\|\mathbb{E}[\mathbf{Q}] - \bar{\mathbf{Q}}\| \to 0$$

où $\bar{\mathbf{Q}}$ est donnée par

$$\bar{\mathbf{Q}} \equiv \left(\frac{n}{T} \frac{\mathbf{\Phi}}{1+\delta} - z\mathbf{I}_T \right)^{-1}$$

avec δ l'unique solution positive $\delta = \frac{1}{T}\operatorname{tr}(\mathbf{\Phi}\bar{\mathbf{Q}})$ et $\mathbf{\Phi} \equiv \mathbb{E}[\sigma(\mathbf{X}^\mathsf{T}\mathbf{w})\sigma(\mathbf{w}^\mathsf{T}\mathbf{X})]$ pour $\mathbf{w} \sim \mathcal{N}(\mathbf{0}, \mathbf{I}_p)$.

Intuition:

- pour $\sigma(t) = t$, [Silverstein et Bai'95]: $\|\mathbb{E}[\mathbf{Q}] - \bar{\mathbf{Q}}\| \to 0$, $\bar{\mathbf{Q}} \equiv \left(\frac{n}{T} \frac{\mathbf{X}^\mathsf{T}\mathbf{X}}{1+\delta} - z\mathbf{I}_T \right)^{-1}$
- repose essentiellement sur lemme de la trace: pour $\mathbf{A}$ de norme borné et indépendante de $\mathbf{W}$,

$$\left| \frac{1}{n}\mathbf{w}_i^\mathsf{T}\mathbf{A}\mathbf{w}_i - \frac{1}{n}\operatorname{tr}\mathbf{A} \right| \to 0$$

  - ✓ $\mathbf{w}_i^\mathsf{T}$ (colonne de $\mathbf{W}$) indépendants
  - ✓ $\mathbf{w}_i$ à entrées i.i.d. $\mathcal{N}(0,1)$,
- Ici, pour $\frac{1}{n}\sigma(\mathbf{w}_i^\mathsf{T}\mathbf{X})\mathbf{A}\sigma(\mathbf{X}^\mathsf{T}\mathbf{w}_i)$,

Équivalent déterministe de $\mathbf{Q}$ (1)

Theorem (Équivalent déterministe de $\mathbf{Q}$)

Sous hypothèses 1 - 3, lorsque $n \to \infty$, on a

$$\|\mathbb{E}[\mathbf{Q}] - \bar{\mathbf{Q}}\| \to 0$$

où $\bar{\mathbf{Q}}$ est donnée par

$$\bar{\mathbf{Q}} \equiv \left(\frac{n}{T} \frac{\mathbf{\Phi}}{1+\delta} - z\mathbf{I}_T \right)^{-1}$$

avec δ l'unique solution positive $\delta = \frac{1}{T}\operatorname{tr}(\mathbf{\Phi}\bar{\mathbf{Q}})$ et $\mathbf{\Phi} \equiv \mathbb{E}[\sigma(\mathbf{X}^\mathsf{T}\mathbf{w})\sigma(\mathbf{w}^\mathsf{T}\mathbf{X})]$ pour $\mathbf{w} \sim \mathcal{N}(\mathbf{0}, \mathbf{I}_p)$.

Intuition:

- pour $\sigma(t) = t$, [Silverstein et Bai'95]: $\|\mathbb{E}[\mathbf{Q}] - \bar{\mathbf{Q}}\| \to 0$, $\bar{\mathbf{Q}} \equiv \left(\frac{n}{T} \frac{\mathbf{X}^\mathsf{T}\mathbf{X}}{1+\delta} - z\mathbf{I}_T \right)^{-1}$
- repose essentiellement sur lemme de la trace: pour $\mathbf{A}$ de norme borné et indépendante de $\mathbf{W}$,

$$\left| \frac{1}{n}\mathbf{w}_i^\mathsf{T}\mathbf{A}\mathbf{w}_i - \frac{1}{n}\operatorname{tr}\mathbf{A} \right| \to 0$$

  - ✓ $\mathbf{w}_i^\mathsf{T}$ (colonne de $\mathbf{W}$) indépendants
  - ✓ $\mathbf{w}_i$ à entrées i.i.d. $\mathcal{N}(0,1)$,
- Ici, pour $\frac{1}{n}\sigma(\mathbf{w}_i^\mathsf{T}\mathbf{X})\mathbf{A}\sigma(\mathbf{X}^\mathsf{T}\mathbf{w}_i)$,
  - ✓ les colonnes de $\mathbf{\Sigma} \equiv \sigma(\mathbf{W}\mathbf{X})$: $\sigma(\mathbf{X}^\mathsf{T}\mathbf{w}_i)$ i.i.d.

Équivalent déterministe de $\mathbf{Q}$ (1)

Theorem (Équivalent déterministe de $\mathbf{Q}$)

Sous hypothèses 1 - 3, lorsque $n \to \infty$, on a

$$\|\mathbb{E}[\mathbf{Q}] - \bar{\mathbf{Q}}\| \to 0$$

où $\bar{\mathbf{Q}}$ est donnée par

$$\bar{\mathbf{Q}} \equiv \left(\frac{n}{T} \frac{\mathbf{\Phi}}{1+\delta} - z\mathbf{I}_T \right)^{-1}$$

avec δ l'unique solution positive $\delta = \frac{1}{T}\operatorname{tr}(\mathbf{\Phi}\bar{\mathbf{Q}})$ et $\mathbf{\Phi} \equiv \mathbb{E}[\sigma(\mathbf{X}^\mathsf{T}\mathbf{w})\sigma(\mathbf{w}^\mathsf{T}\mathbf{X})]$ pour $\mathbf{w} \sim \mathcal{N}(\mathbf{0}, \mathbf{I}_p)$.

Intuition:

- pour $\sigma(t) = t$, [Silverstein et Bai'95]: $\|\mathbb{E}[\mathbf{Q}] - \bar{\mathbf{Q}}\| \to 0$, $\bar{\mathbf{Q}} \equiv \left(\frac{n}{T} \frac{\mathbf{X}^\mathsf{T}\mathbf{X}}{1+\delta} - z\mathbf{I}_T \right)^{-1}$
- repose essentiellement sur lemme de la trace: pour $\mathbf{A}$ de norme borné et indépendante de $\mathbf{W}$,

$$\left| \frac{1}{n}\mathbf{w}_i^\mathsf{T}\mathbf{A}\mathbf{w}_i - \frac{1}{n}\operatorname{tr}\mathbf{A} \right| \to 0$$

  - ✓ $\mathbf{w}_i^\mathsf{T}$ (colonne de $\mathbf{W}$) indépendants
  - ✓ $\mathbf{w}_i$ à entrées i.i.d. $\mathcal{N}(0,1)$,
- Ici, pour $\frac{1}{n}\sigma(\mathbf{w}_i^\mathsf{T}\mathbf{X})\mathbf{A}\sigma(\mathbf{X}^\mathsf{T}\mathbf{w}_i)$,
  - ✓ les colonnes de $\mathbf{\Sigma} \equiv \sigma(\mathbf{W}\mathbf{X})$: $\sigma(\mathbf{X}^\mathsf{T}\mathbf{w}_i)$ i.i.d.
  - × mais $\sigma(\mathbf{X}^\mathsf{T}\mathbf{w}_i)$ à entrées non-indépendants

Équivalent déterministe de $\mathbf{Q}$ (2)

Theorem (Équivalent déterministe de $\mathbf{Q}$)

Sous hypothèses 1 - 3, lorsque $n \to \infty$, on a

$$\|\mathbb{E}[\mathbf{Q}] - \bar{\mathbf{Q}}\| \to 0$$

où $\bar{\mathbf{Q}}$ est donnée par

$$\bar{\mathbf{Q}} \equiv \left(\frac{n}{T} \frac{\mathbf{\Phi}}{1+\delta} - z\mathbf{I}_T \right)^{-1}$$

avec δ l'unique solution positive $\delta = \frac{1}{T}\operatorname{tr}(\mathbf{\Phi}\bar{\mathbf{Q}})$ et $\mathbf{\Phi} \equiv \mathbb{E}[\sigma(\mathbf{X}^\mathsf{T}\mathbf{w})\sigma(\mathbf{w}^\mathsf{T}\mathbf{X})]$ pour $\mathbf{w} \sim \mathcal{N}(\mathbf{0}, \mathbf{I}_p)$.

Équivalent déterministe de $\mathbf{Q}$ (2)

Theorem (Équivalent déterministe de $\mathbf{Q}$)

Sous hypothèses 1 - 3, lorsque $n \to \infty$, on a

$$\|\mathbb{E}[\mathbf{Q}] - \bar{\mathbf{Q}}\| \to 0$$

où $\bar{\mathbf{Q}}$ est donnée par

$$\bar{\mathbf{Q}} \equiv \left(\frac{n}{T} \frac{\mathbf{\Phi}}{1+\delta} - z\mathbf{I}_T \right)^{-1}$$

avec δ l'unique solution positive $\delta = \frac{1}{T}\operatorname{tr}(\mathbf{\Phi}\bar{\mathbf{Q}})$ et $\mathbf{\Phi} \equiv \mathbb{E}[\sigma(\mathbf{X}^\mathsf{T}\mathbf{w})\sigma(\mathbf{w}^\mathsf{T}\mathbf{X})]$ pour $\mathbf{w} \sim \mathcal{N}(\mathbf{0}, \mathbf{I}_p)$.

Intuition:

- On remplace le lemme de la trace par concentration de la mesure:

$$P\left(\left| \frac{1}{n}\sigma(\mathbf{w}_i^\mathsf{T}\mathbf{X})\mathbf{A}\sigma(\mathbf{X}^\mathsf{T}\mathbf{w}_i) - \frac{1}{n}\operatorname{tr}(\mathbf{\Phi}\mathbf{A}) \right| > t \right) \leq Ce^{-cn\max(t,t^2)}$$

avec $\mathbf{\Phi} \equiv \mathbb{E}_{\mathbf{w}}[\sigma(\mathbf{X}^\mathsf{T}\mathbf{w})\sigma(\mathbf{w}^\mathsf{T}\mathbf{X})]$ pour $\mathbf{w} \sim \mathcal{N}(\mathbf{0}, \mathbf{I}_p)$.

Application: Analyse de la performance de l'ELM

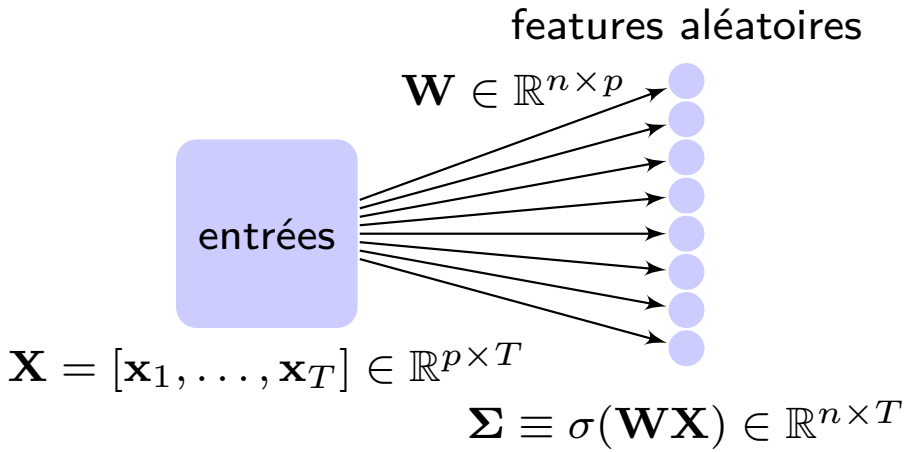

Figure: Illustration de l'ELM

Phase d'apprentissage:

Application: Analyse de la performance de l'ELM

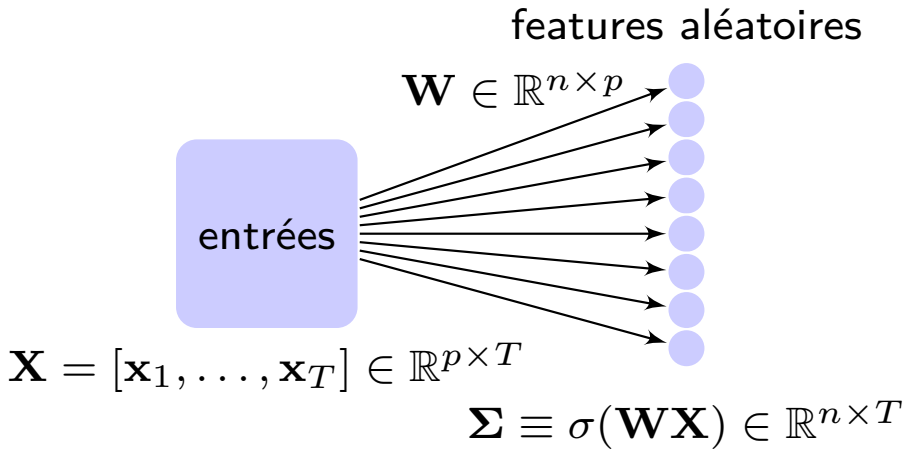

Figure: Illustration de l'ELM

Phase d'apprentissage:

- génération de features aléatoires pour les données d'entraînement $\mathbf{X}$: $\mathbf{\Sigma} \equiv \sigma(\mathbf{WX})$

Application: Analyse de la performance de l'ELM

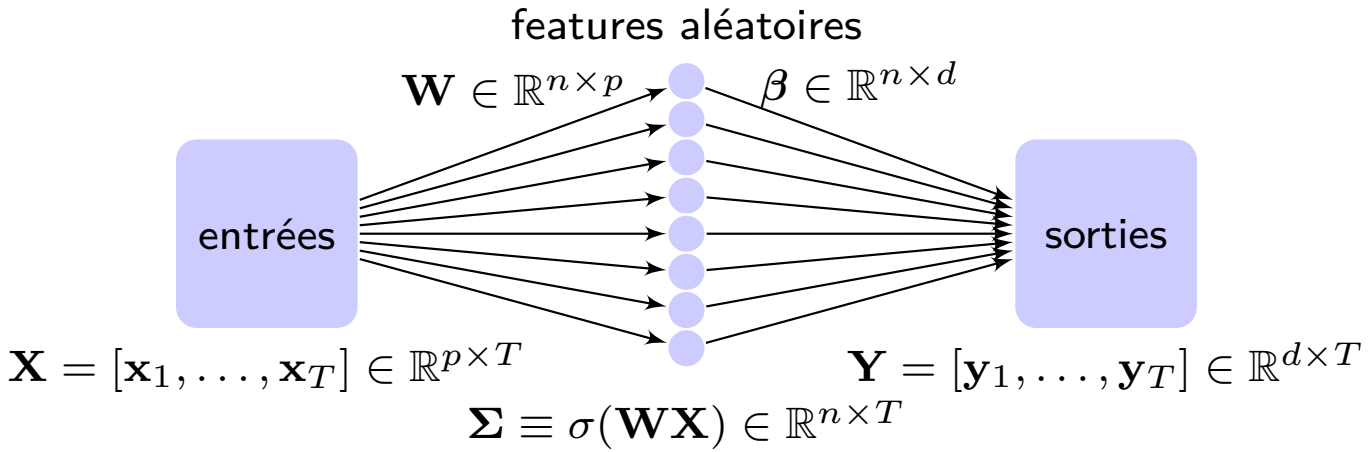

Figure: Illustration de l'ELM

Phase d'apprentissage:

- génération de features aléatoires pour les données d'entraînement $\mathbf{X}$: $\boldsymbol{\Sigma} \equiv \sigma(\mathbf{WX})$
- régression normalisée de $\boldsymbol{\Sigma}$ aux sorties associées $\mathbf{Y}$:
  $\boldsymbol{\beta} = \frac{1}{T}\boldsymbol{\Sigma}\left(\frac{1}{T}\boldsymbol{\Sigma}^{\mathsf{T}}\boldsymbol{\Sigma} + \gamma \mathbf{I}_T\right)^{-1}\mathbf{Y}^{\mathsf{T}} = \frac{1}{T}\boldsymbol{\Sigma}\mathbf{Q}(-\gamma)\mathbf{Y}^{\mathsf{T}}$

Application: Analyse de la performance de l'ELM

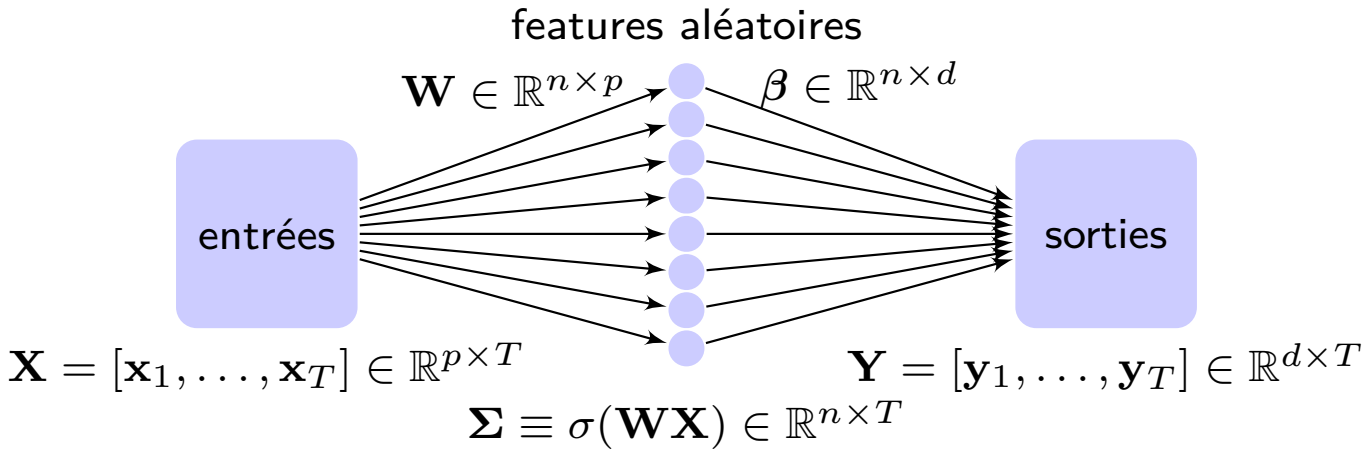

Figure: Illustration de l'ELM

Phase d'apprentissage:

- génération de features aléatoires pour les données d'entraînement $\mathbf{X}$: $\boldsymbol{\Sigma} \equiv \sigma(\mathbf{WX})$
- régression normalisée de $\boldsymbol{\Sigma}$ aux sorties associées $\mathbf{Y}$:
  $\boldsymbol{\beta} = \frac{1}{T}\boldsymbol{\Sigma}\left(\frac{1}{T}\boldsymbol{\Sigma}^{\mathsf{T}}\boldsymbol{\Sigma} + \gamma \mathbf{I}_T\right)^{-1}\mathbf{Y}^{\mathsf{T}} = \frac{1}{T}\boldsymbol{\Sigma}\mathbf{Q}(-\gamma)\mathbf{Y}^{\mathsf{T}}$

Phase de test:

Application: Analyse de la performance de l'ELM

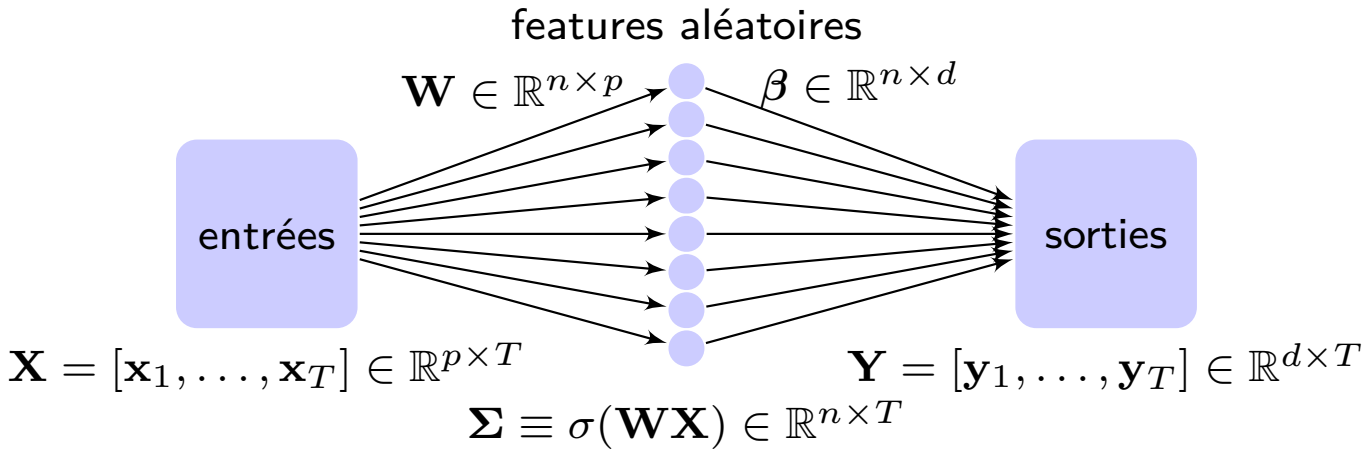

Figure: Illustration de l'ELM

Phase d'apprentissage:

- génération de features aléatoires pour les données d'entraînement $\mathbf{X}$: $\boldsymbol{\Sigma} \equiv \sigma(\mathbf{WX})$
- régression normalisée de $\boldsymbol{\Sigma}$ aux sorties associées $\mathbf{Y}$:
  $\beta = \frac{1}{T}\boldsymbol{\Sigma}\left(\frac{1}{T}\boldsymbol{\Sigma}^{\mathsf{T}}\boldsymbol{\Sigma} + \gamma\mathbf{I}_T\right)^{-1}\mathbf{Y}^{\mathsf{T}} = \frac{1}{T}\boldsymbol{\Sigma}\mathbf{Q}(-\gamma)\mathbf{Y}^{\mathsf{T}}$

Phase de test:

- appliquer le régresseur aux nouvelles données $\hat{\mathbf{X}}$, sorties $\hat{\mathbf{Y}}$ inconnues

Application: Analyse de la performance de l'ELM

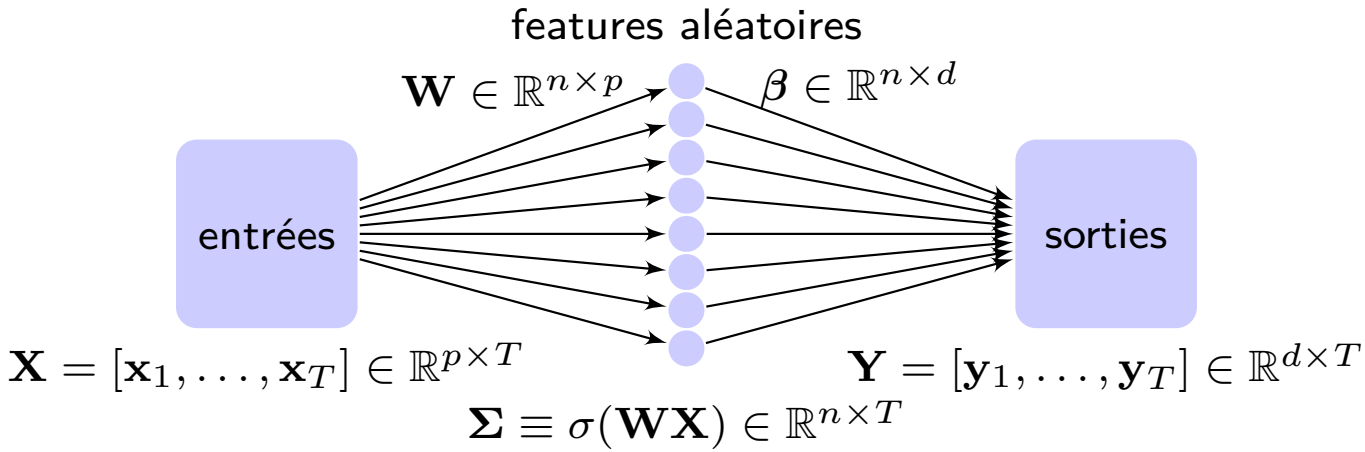

Figure: Illustration de l'ELM

Phase d'apprentissage:

- génération de features aléatoires pour les données d'entraînement $\mathbf{X}$: $\boldsymbol{\Sigma} \equiv \sigma(\mathbf{WX})$
- régression normalisée de $\boldsymbol{\Sigma}$ aux sorties associées $\mathbf{Y}$:
  $\beta = \frac{1}{T}\boldsymbol{\Sigma}\left(\frac{1}{T}\boldsymbol{\Sigma}^{\mathsf{T}}\boldsymbol{\Sigma} + \gamma\mathbf{I}_T\right)^{-1}\mathbf{Y}^{\mathsf{T}} = \frac{1}{T}\boldsymbol{\Sigma}\mathbf{Q}(-\gamma)\mathbf{Y}^{\mathsf{T}}$

Phase de test:

- appliquer le régresseur aux nouvelles données $\hat{\mathbf{X}}$, sorties $\hat{\mathbf{Y}}$ inconnues
- sortie de l'ELM donnée par $\mathbf{Z} = \beta^{\mathsf{T}}\hat{\boldsymbol{\Sigma}}$, avec $\hat{\boldsymbol{\Sigma}} \equiv \sigma(\mathbf{W}\hat{\mathbf{X}})$

Compréhension la performance de l'ELM

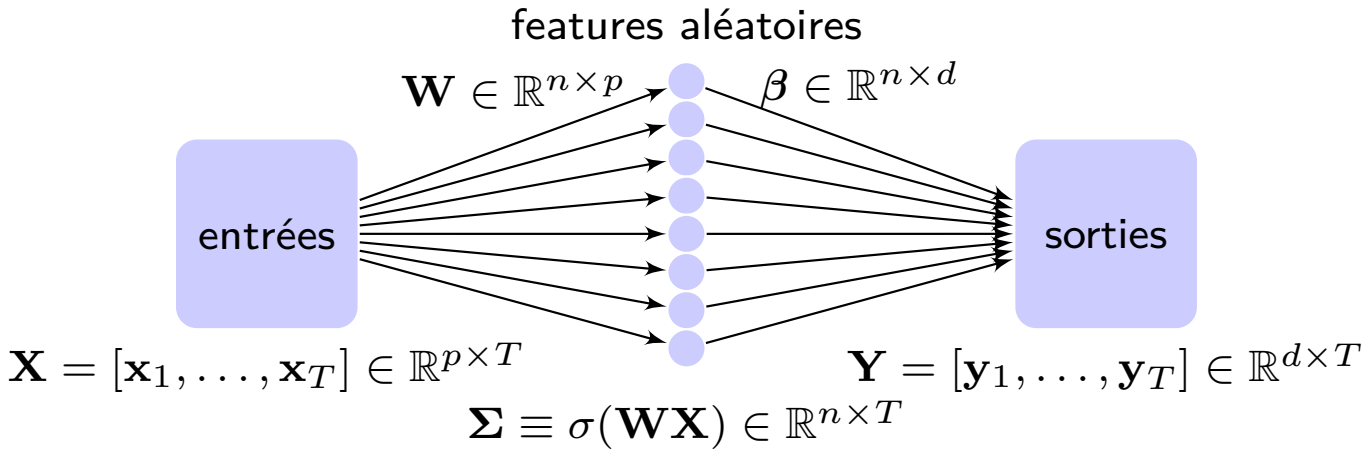

Figure: Illustration of ELM

avec un *ridge-regressor* $\beta = \frac{1}{T}\boldsymbol{\Sigma}\left(\frac{1}{T}\boldsymbol{\Sigma}^{\mathsf{T}}\boldsymbol{\Sigma} + \gamma \mathbf{I}_T\right)^{-1}\mathbf{Y}^{\mathsf{T}} = \frac{1}{T}\boldsymbol{\Sigma}\mathbf{Q}\mathbf{Y}^{\mathsf{T}}$.

Compréhension la performance de l'ELM

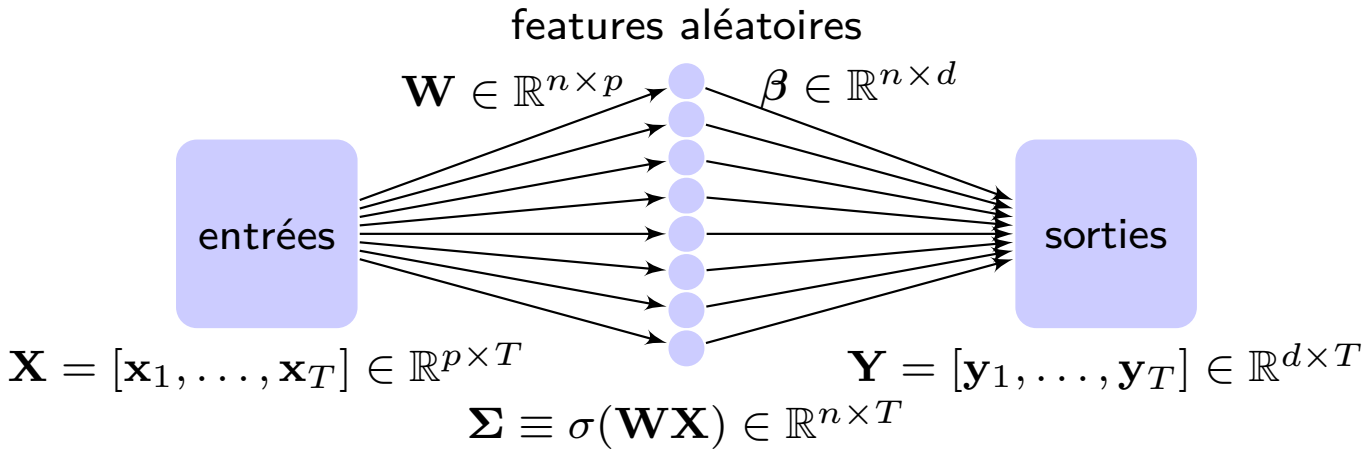

Figure: Illustration of ELM

avec un *ridge-regressor* $\beta = \frac{1}{T}\boldsymbol{\Sigma}\left(\frac{1}{T}\boldsymbol{\Sigma}^{\mathsf{T}}\boldsymbol{\Sigma} + \gamma \mathbf{I}_T\right)^{-1}\mathbf{Y}^{\mathsf{T}} = \frac{1}{T}\boldsymbol{\Sigma}\mathbf{Q}\mathbf{Y}^{\mathsf{T}}$.
Pour comprendre le MSE (mean-square error) d'apprentissage:

$$E_{\text{train}} = \frac{1}{T}\left\|\mathbf{Y}^{\mathsf{T}} - \boldsymbol{\Sigma}^{\mathsf{T}}\beta\right\|_F^2 = \frac{\gamma^2}{T}\operatorname{tr}(\mathbf{Y}\mathbf{Q}^2\mathbf{Y}^{\mathsf{T}})$$

Compréhension la performance de l'ELM

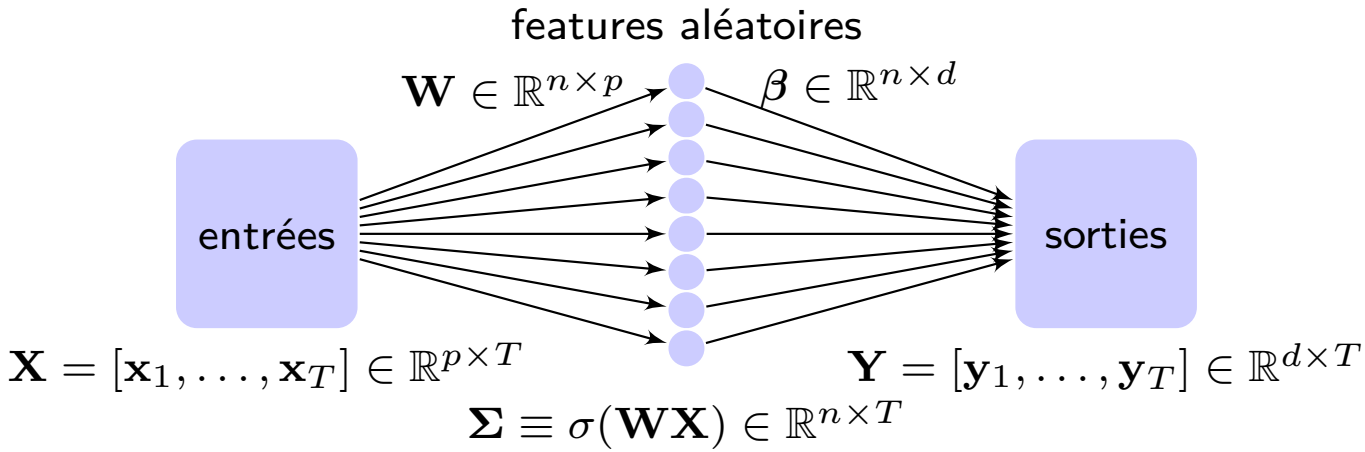

Figure: Illustration of ELM

avec un *ridge-regressor* $\beta = \frac{1}{T}\boldsymbol{\Sigma}\left(\frac{1}{T}\boldsymbol{\Sigma}^{\mathsf{T}}\boldsymbol{\Sigma} + \gamma \mathbf{I}_T\right)^{-1}\mathbf{Y}^{\mathsf{T}} = \frac{1}{T}\boldsymbol{\Sigma}\mathbf{Q}\mathbf{Y}^{\mathsf{T}}$.
Pour comprendre le MSE (mean-square error) d'apprentissage:

$$E_{\text{train}} = \frac{1}{T}\left\|\mathbf{Y}^{\mathsf{T}} - \boldsymbol{\Sigma}^{\mathsf{T}}\beta\right\|_F^2 = \frac{\gamma^2}{T}\operatorname{tr}(\mathbf{Y}\mathbf{Q}^2\mathbf{Y}^{\mathsf{T}})$$

aussi la performance de test pour les nouvelles données $\hat{\mathbf{X}}$ et les sorties associées $\hat{\mathbf{Y}}$:

$$E_{\text{test}} = \frac{1}{T}\left\|\hat{\mathbf{Y}}^{\mathsf{T}} - \hat{\boldsymbol{\Sigma}}^{\mathsf{T}}\beta\right\|_F^2, \ \hat{\boldsymbol{\Sigma}} = \sigma(\mathbf{W}\hat{\mathbf{X}})$$

Équivalent déterministe de MSE

Theorem (Équivalent déterministe de E_{train})

Sous hypothèses 1-3, lorsque $n \to \infty$, on a

$$\left| E_{\text{train}} - \bar{E}_{\text{train}} \right| \to 0$$

avec

$$\bar{E}_{\text{train}} = \frac{\gamma^2}{T} \operatorname{tr} \left(\mathbf{Y} \bar{\mathbf{Q}} \left[\frac{\frac{1}{n} \operatorname{tr}(\bar{\mathbf{Q}} \boldsymbol{\Psi} \bar{\mathbf{Q}})}{1 - \frac{1}{n} \operatorname{tr}(\boldsymbol{\Psi}^2 \bar{\mathbf{Q}}^2)} \boldsymbol{\Psi} + \mathbf{I}_T \right] \bar{\mathbf{Q}} \mathbf{Y}^\mathsf{T} \right), \quad \boldsymbol{\Psi} \equiv \frac{n}{T} \frac{\boldsymbol{\Phi}}{1+\delta}$$

Équivalent déterministe de MSE

Theorem (Équivalent déterministe de E_{train})

Sous hypothèses 1-3, lorsque $n \to \infty$, on a

$$\left| E_{\text{train}} - \bar{E}_{\text{train}} \right| \to 0$$

avec

$$\bar{E}_{\text{train}} = \frac{\gamma^2}{T} \operatorname{tr} \left(\mathbf{Y}\bar{\mathbf{Q}} \left[\frac{\frac{1}{n}\operatorname{tr}(\bar{\mathbf{Q}}\boldsymbol{\Psi}\bar{\mathbf{Q}})}{1 - \frac{1}{n}\operatorname{tr}(\boldsymbol{\Psi}^2\bar{\mathbf{Q}}^2)} \boldsymbol{\Psi} + \mathbf{I}_T \right] \bar{\mathbf{Q}}\mathbf{Y}^\mathsf{T} \right), \quad \boldsymbol{\Psi} \equiv \frac{n}{T} \frac{\boldsymbol{\Phi}}{1+\delta}$$

Dernière mission: évaluer $\boldsymbol{\Phi}_{\mathbf{ab}} \equiv \mathbb{E}_{\mathbf{w}}[\sigma(\mathbf{w}^\mathsf{T}\mathbf{a})\sigma(\mathbf{w}^\mathsf{T}\mathbf{b})]$ pour des $\sigma(\cdot)$ différents:

Équivalent déterministe de MSE

Theorem (Équivalent déterministe de E_{train})

Sous hypothèses 1-3, lorsque $n \to \infty$, on a

$$\left| E_{\text{train}} - \bar{E}_{\text{train}} \right| \to 0$$

avec

$$\bar{E}_{\text{train}} = \frac{\gamma^2}{T} \operatorname{tr}\left(\mathbf{Y}\bar{\mathbf{Q}} \left[\frac{\frac{1}{n}\operatorname{tr}(\bar{\mathbf{Q}}\boldsymbol{\Psi}\bar{\mathbf{Q}})}{1 - \frac{1}{n}\operatorname{tr}(\boldsymbol{\Psi}^2\bar{\mathbf{Q}}^2)} \boldsymbol{\Psi} + \mathbf{I}_T \right] \bar{\mathbf{Q}}\mathbf{Y}^\mathsf{T} \right), \quad \boldsymbol{\Psi} \equiv \frac{n}{T}\frac{\boldsymbol{\Phi}}{1+\delta}$$

Dernière mission: évaluer $\boldsymbol{\Phi}_{\mathbf{ab}} \equiv \mathbb{E}_{\mathbf{w}}[\sigma(\mathbf{w}^\mathsf{T}\mathbf{a})\sigma(\mathbf{w}^\mathsf{T}\mathbf{b})]$ pour des $\sigma(\cdot)$ différents:

$\sigma(t)$	$\Phi_{\mathbf{ab}}$
t	$\mathbf{a}^\mathsf{T}\mathbf{b}$
$\max(t,0)$	$\frac{1}{2\pi}\|\mathbf{a}\|\|\mathbf{b}\| \left(\angle(\mathbf{a},\mathbf{b}) \arccos(-\angle(\mathbf{a},\mathbf{b})) + \sqrt{1-\angle(\mathbf{a},\mathbf{b})^2} \right)$
$\operatorname{erf}(t)$	$\frac{2}{\pi}\arcsin\left(\frac{2\mathbf{a}^\mathsf{T}\mathbf{b}}{\sqrt{(1+2\|\mathbf{a}\|^2)(1+2\|\mathbf{b}\|^2)}} \right)$
$1_{t>0}$	$\frac{1}{2} - \frac{1}{2\pi}\arccos(\angle(\mathbf{a},\mathbf{b}))$
$\operatorname{sign}(t)$	$\frac{2}{\pi}\arcsin(\angle(\mathbf{a},\mathbf{b}))$
$\cos(t)$	$\exp\left(-\frac{1}{2}\left(\|\mathbf{a}\|^2 + \|\mathbf{b}\|^2\right)\right)\cosh(\mathbf{a}^\mathsf{T}\mathbf{b})$
$\sin(t)$	$\exp\left(-\frac{1}{2}\left(\|\mathbf{a}\|^2 + \|\mathbf{b}\|^2\right)\right)\sinh(\mathbf{a}^\mathsf{T}\mathbf{b})$

Table: Valeurs de $\Phi_{\mathbf{ab}}$ pour $w \sim \mathcal{N}(0, \mathbf{I}_p)$, avec $\angle(\mathbf{a},\mathbf{b}) \equiv \frac{\mathbf{a}^\mathsf{T}\mathbf{b}}{\|\mathbf{a}\|\|\mathbf{b}\|}$.

Validation numérique sur la base de données MNIST[2]

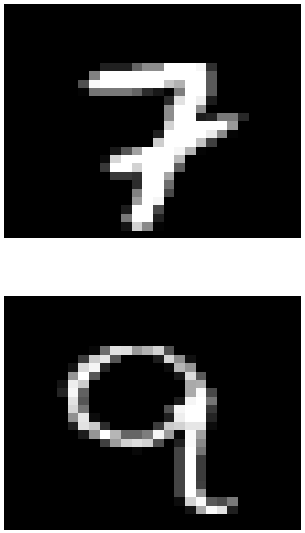

Figure: Exemples de la base de données MNIST

[2] Les images des chiffres manuscrits de taille 28 28×28.

Validation numérique sur la base de données MNIST[2]

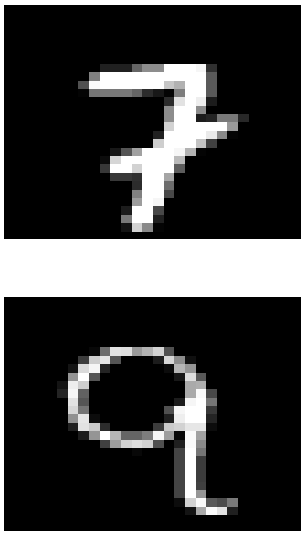

Figure: Exemples de la base de données MNIST

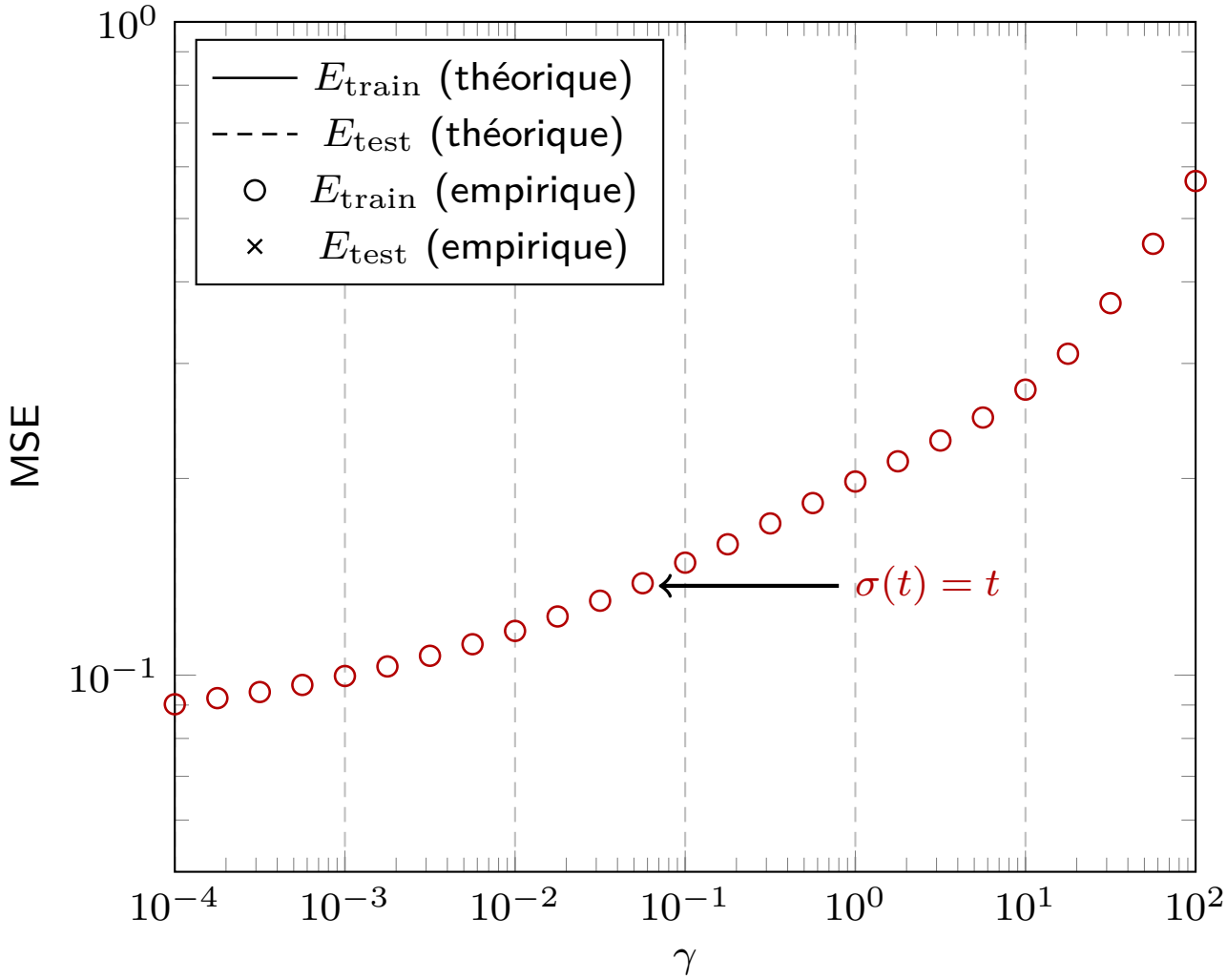

Figure: Performance de l'ELM pour σ Lipschitzienne, en fonction de γ, pour 2 classes de données de MNIST (sept et neuf), $n = 512$, $T = \hat{T} = 1024$, $p = 784$.

[2]Les images des chiffres manuscrits de taille 28 28×28.

Validation numérique sur la base de données MNIST[2]

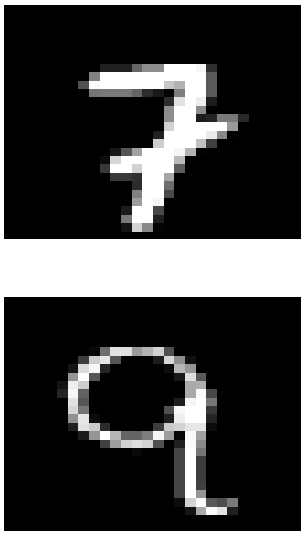

Figure: Exemples de la base de données MNIST

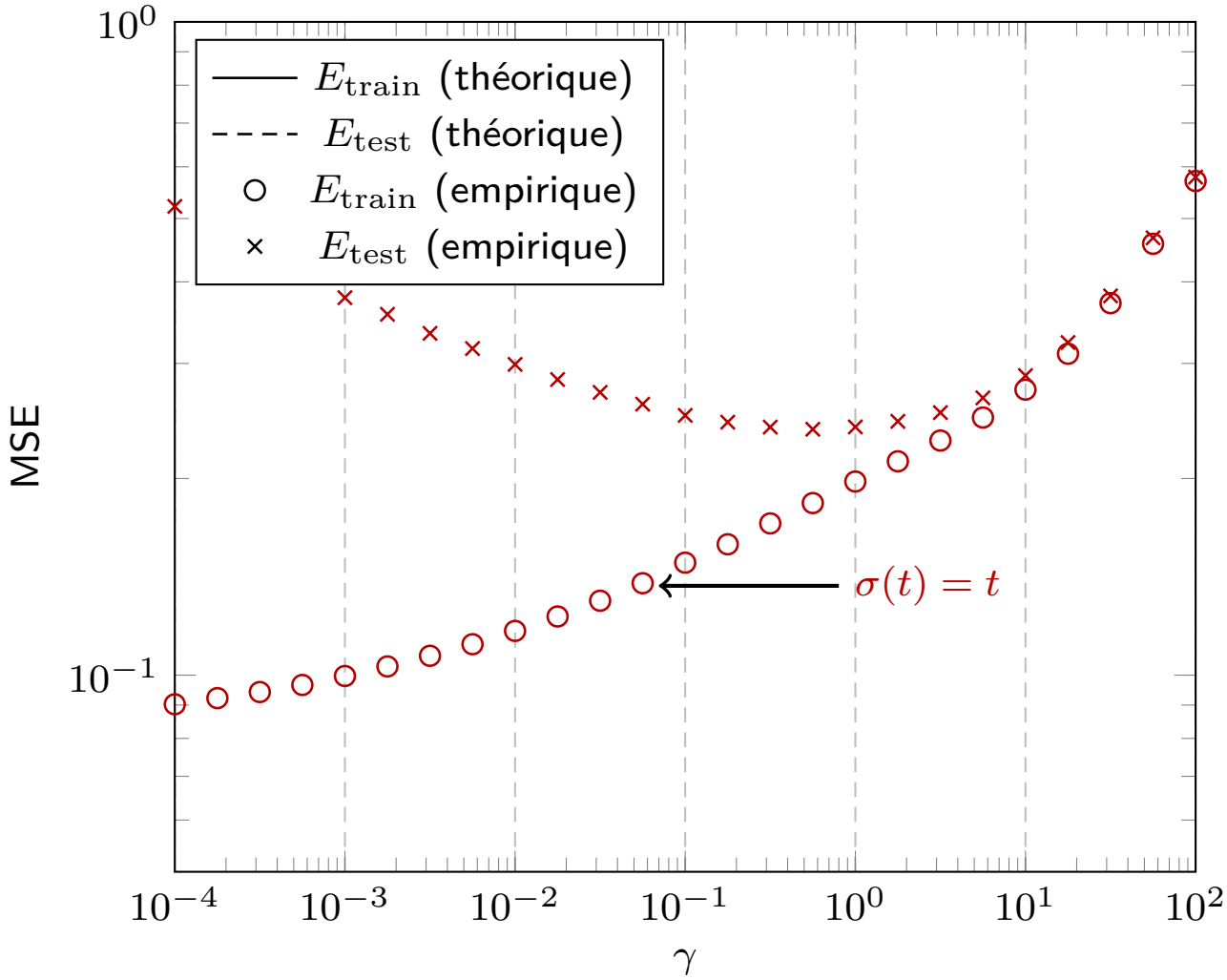

Figure: Performance de l'ELM pour σ Lipschitzienne, en fonction de γ, pour 2 classes de données de MNIST (sept et neuf), $n = 512$, $T = \hat{T} = 1024$, $p = 784$.

[2]Les images des chiffres manuscrits de taille 28 28×28.

Validation numérique sur la base de données MNIST[2]

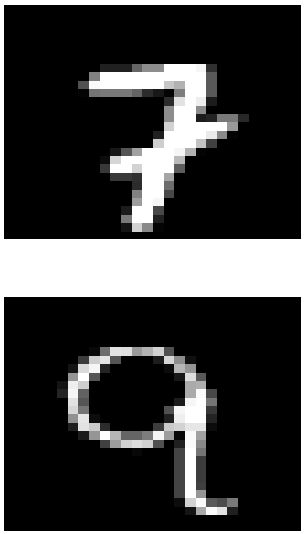

Figure: Exemples de la base de données MNIST

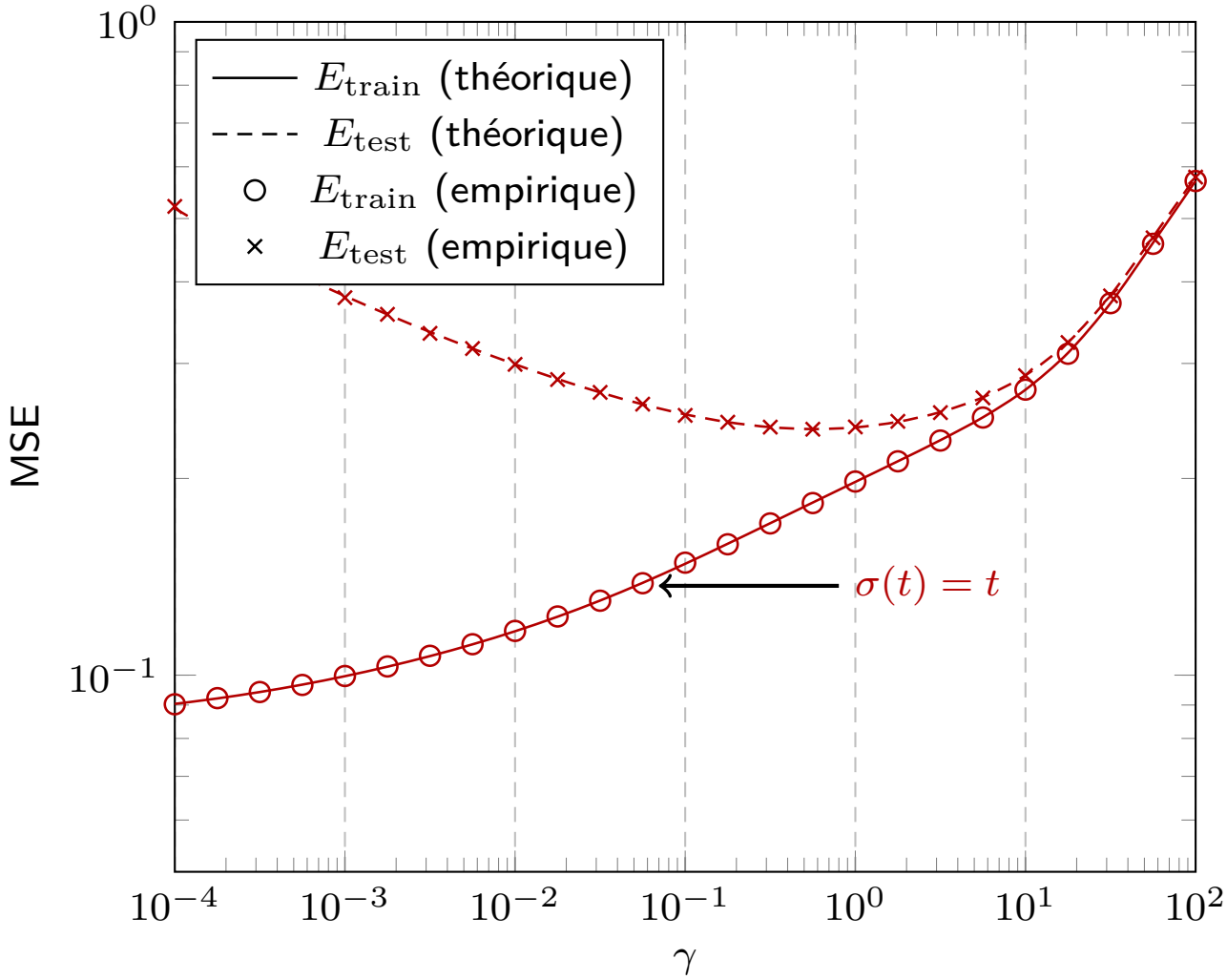

Figure: Performance de l'ELM pour σ Lipschitzienne, en fonction de γ, pour 2 classes de données de MNIST (sept et neuf), $n = 512$, $T = \hat{T} = 1024$, $p = 784$.

[2]Les images des chiffres manuscrits de taille 28 28 $\times$ 28.

Validation numérique sur la base de données MNIST[2]

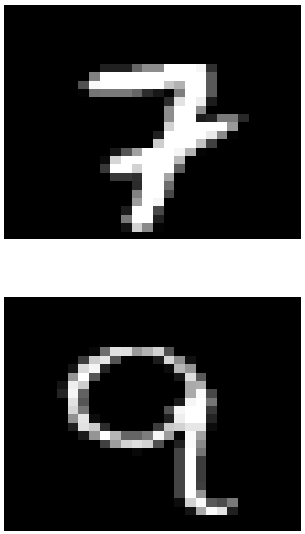

Figure: Exemples de la base de données MNIST

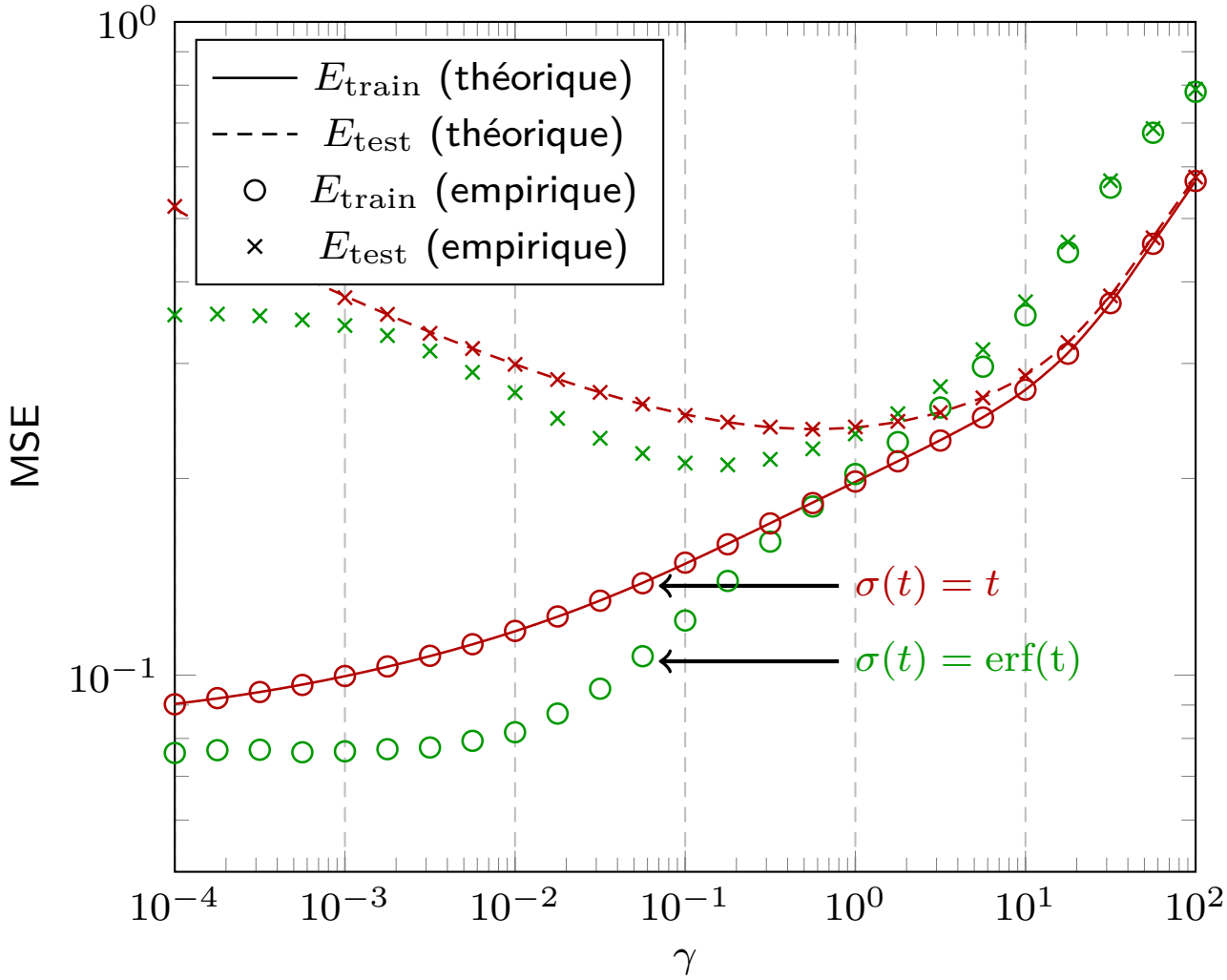

Figure: Performance de l'ELM pour σ Lipschitzienne, en fonction de γ, pour 2 classes de données de MNIST (sept et neuf), $n = 512$, $T = \hat{T} = 1024$, $p = 784$.

[2]Les images des chiffres manuscrits de taille 28 28 $\times$ 28.

Validation numérique sur la base de données MNIST[2]

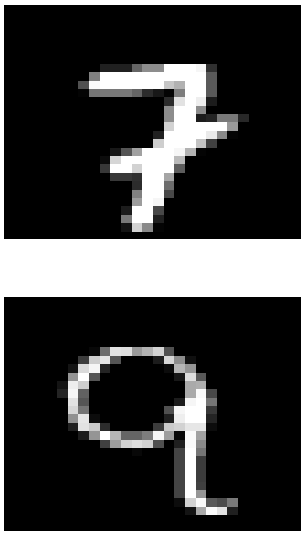

Figure: Exemples de la base de données MNIST

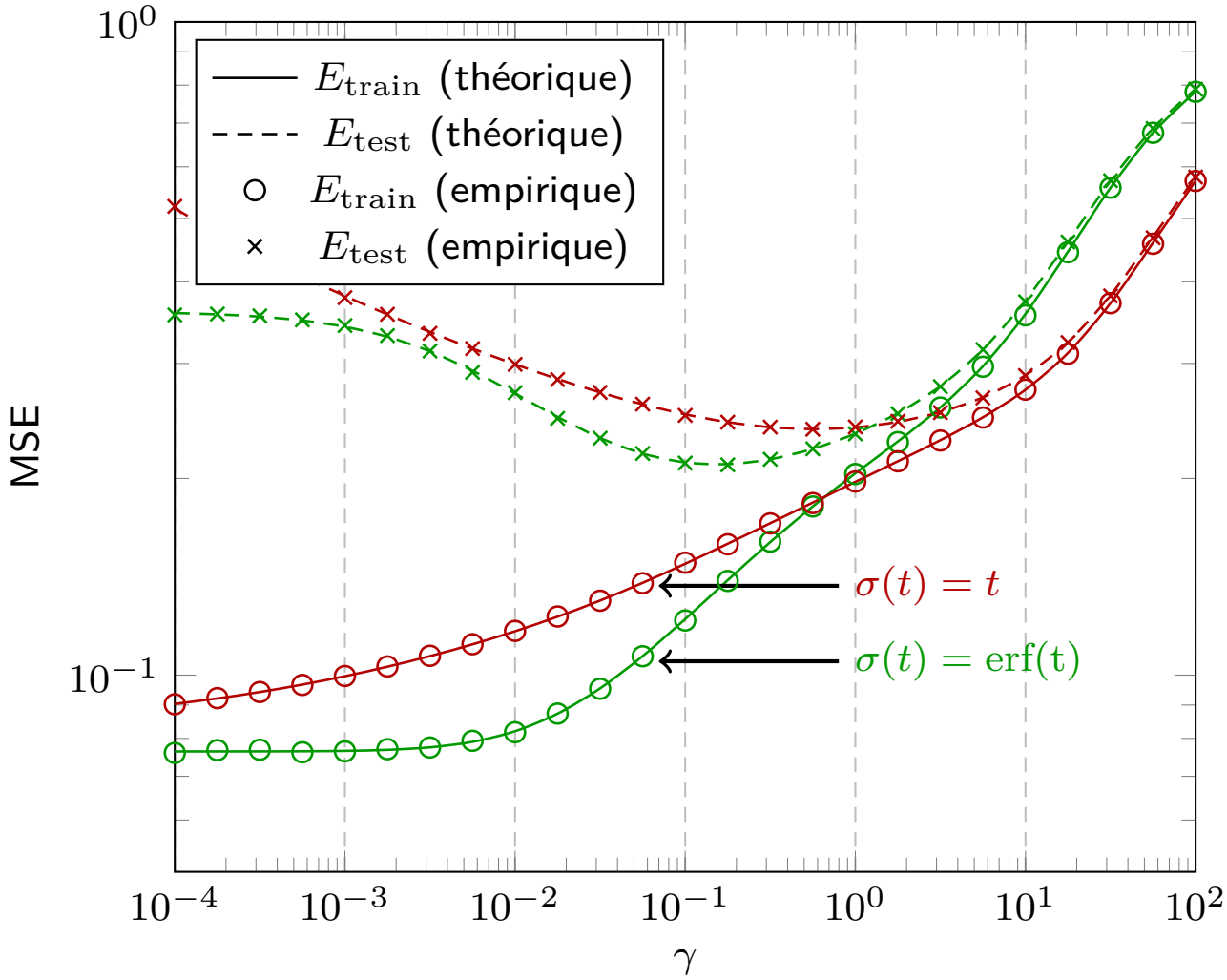

Figure: Performance de l'ELM pour σ Lipschitzienne, en fonction de γ, pour 2 classes de données de MNIST (sept et neuf), $n = 512$, $T = \hat{T} = 1024$, $p = 784$.

[2]Les images des chiffres manuscrits de taille 28 28 $\times$ 28.

Validation numérique sur la base de données MNIST[2]

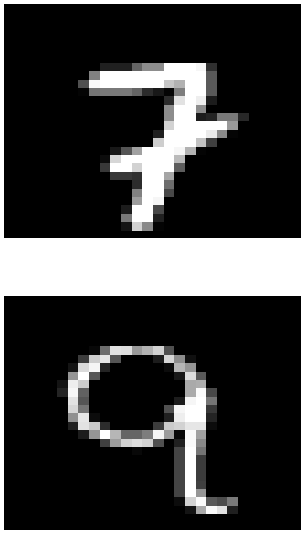

Figure: Exemples de la base de données MNIST

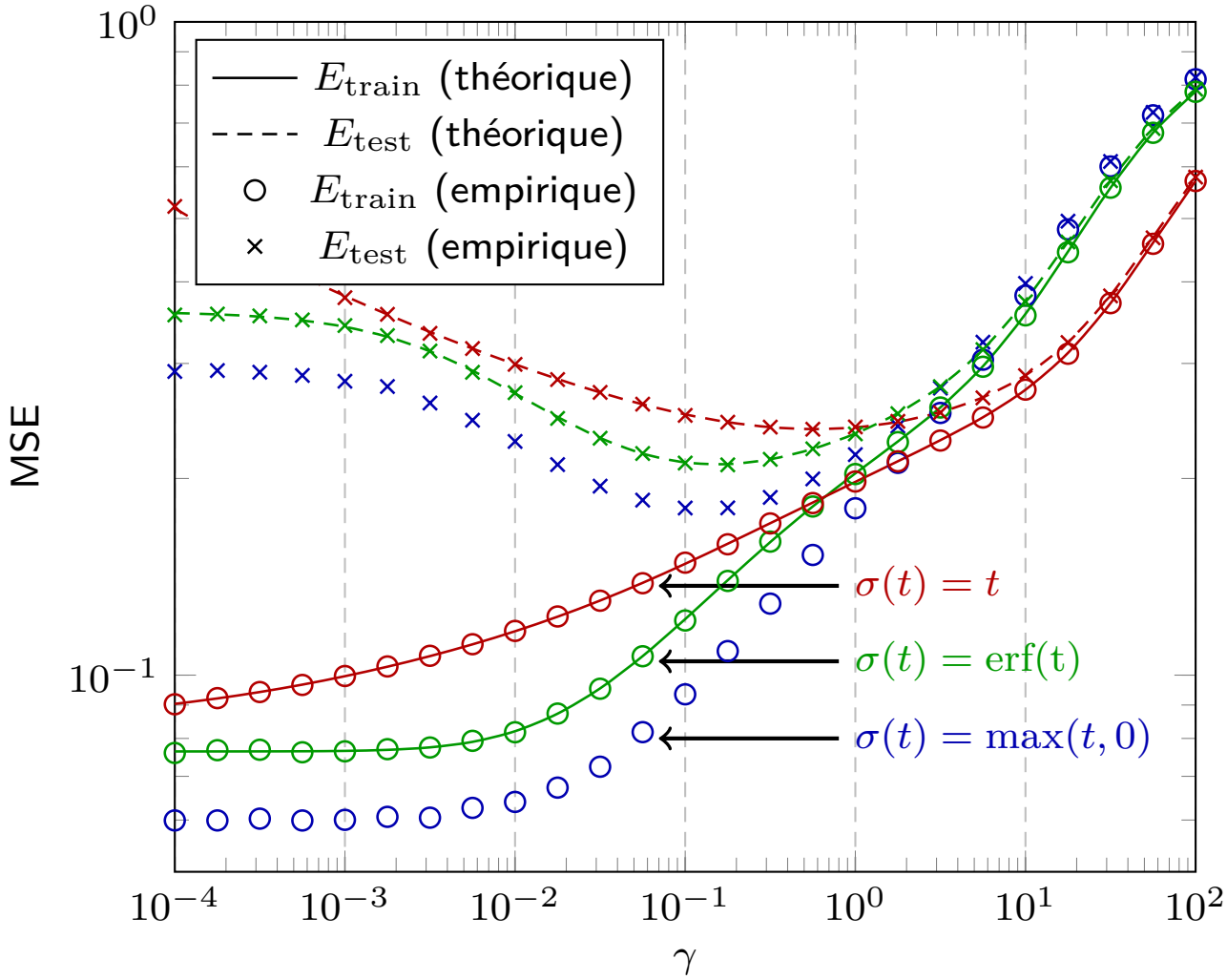

Figure: Performance de l'ELM pour σ Lipschitzienne, en fonction de γ, pour 2 classes de données de MNIST (sept et neuf), $n = 512$, $T = \hat{T} = 1024$, $p = 784$.

[2]Les images des chiffres manuscrits de taille 28 28 $\times$ 28.

Validation numérique sur la base de données MNIST[2]

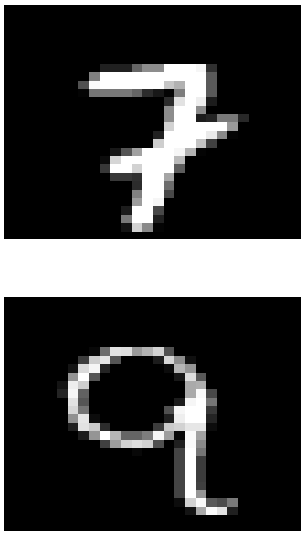

Figure: Exemples de la base de données MNIST

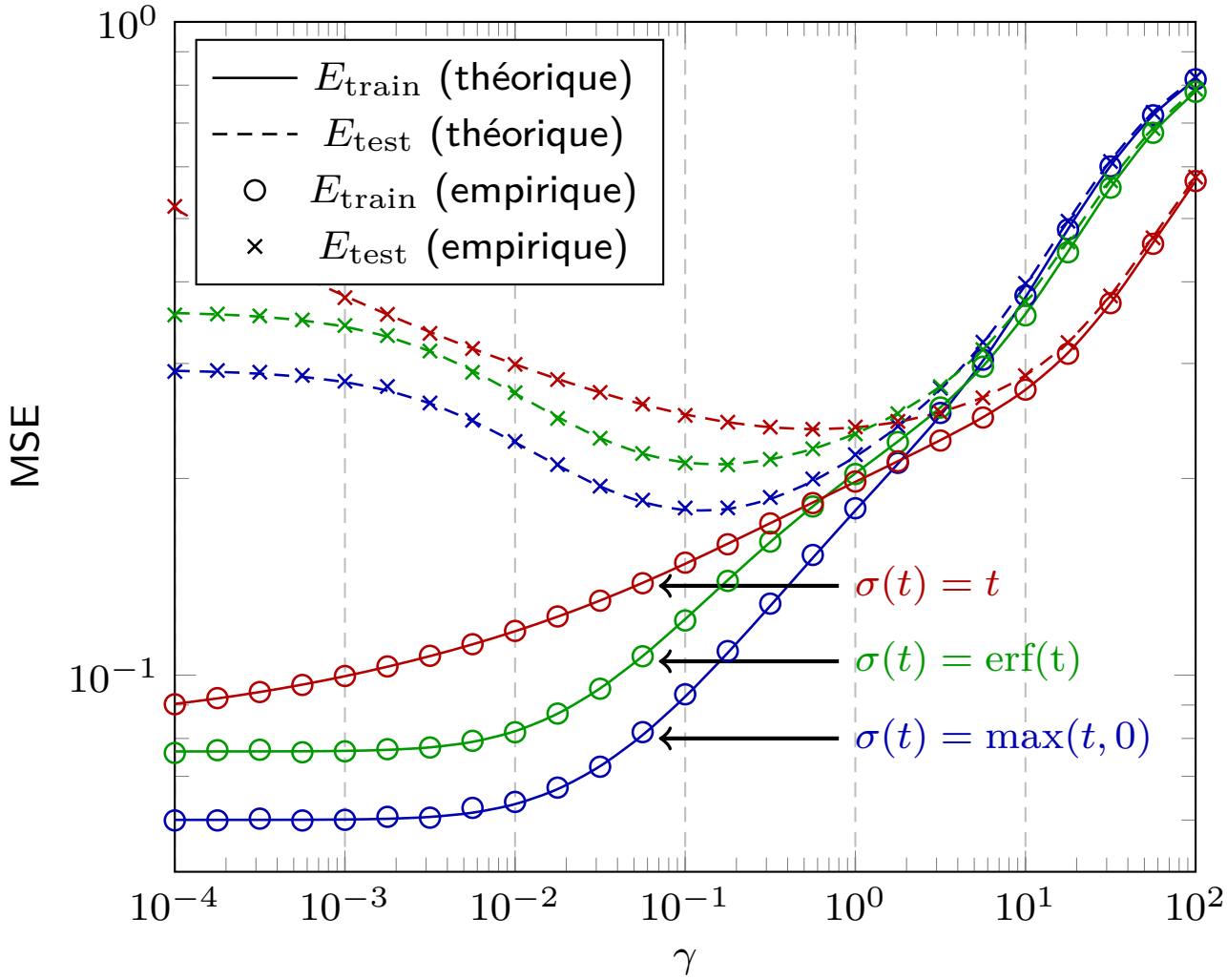

Figure: Performance de l'ELM pour σ Lipschitzienne, en fonction de γ, pour 2 classes de données de MNIST (sept et neuf), $n = 512$, $T = \hat{T} = 1024$, $p = 784$.

[2]Les images des chiffres manuscrits de taille 28 28 $\times$ 28.

Plan

Conclusion

Messages:

- étude de la matrice de Gram non-linéaire: $\mathbf{G} \equiv \frac{1}{T}\boldsymbol{\Sigma}^{\mathsf{T}}\boldsymbol{\Sigma}$

Conclusion

Messages:

- étude de la matrice de Gram non-linéaire: $\mathbf{G} \equiv \frac{1}{T}\boldsymbol{\Sigma}^{\mathsf{T}}\boldsymbol{\Sigma}$
- objet clé: la résolvante $\mathbf{Q}$ et son équivalent déterministe asymptotique $\bar{\mathbf{Q}}$

Conclusion

Messages:

- étude de la matrice de Gram non-linéaire: $\mathbf{G} \equiv \frac{1}{T}\mathbf{\Sigma}^{\mathsf{T}}\mathbf{\Sigma}$
- objet clé: la résolvante $\mathbf{Q}$ et son équivalent déterministe asymptotique $\bar{\mathbf{Q}}$
- permet la compréhension de la performance de l'ELM et projections aléatoires

Conclusion

Messages:

- étude de la matrice de Gram non-linéaire: $\mathbf{G} \equiv \frac{1}{T}\mathbf{\Sigma}^{\mathsf{T}}\mathbf{\Sigma}$
- objet clé: la résolvante $\mathbf{Q}$ et son équivalent déterministe asymptotique $\bar{\mathbf{Q}}$
- permet la compréhension de la performance de l'ELM et projections aléatoires
- optimisation théorique des hyper-paramètres (ici γ)

Conclusion

Messages:

- étude de la matrice de Gram non-linéaire: $\mathbf{G} \equiv \frac{1}{T}\mathbf{\Sigma}^{\mathsf{T}}\mathbf{\Sigma}$
- objet clé: la résolvante $\mathbf{Q}$ et son équivalent déterministe asymptotique $\bar{\mathbf{Q}}$
- permet la compréhension de la performance de l'ELM et projections aléatoires
- optimisation théorique des hyper-paramètres (ici γ)

Travaux à venir:

- analyse des méthodes plus sophistiquées basées sur les projections aléatoires, e.g., projections aléatoires + régression logistique, projections aléatoires + SVM, etc.

Conclusion

Messages:

- étude de la matrice de Gram non-linéaire: $\mathbf{G} \equiv \frac{1}{T}\mathbf{\Sigma}^{\mathsf{T}}\mathbf{\Sigma}$
- objet clé: la résolvante $\mathbf{Q}$ et son équivalent déterministe asymptotique $\bar{\mathbf{Q}}$
- permet la compréhension de la performance de l'ELM et projections aléatoires
- optimisation théorique des hyper-paramètres (ici γ)

Travaux à venir:

- analyse des méthodes plus sophistiquées basées sur les projections aléatoires, e.g., projections aléatoires + régression logistique, projections aléatoires + SVM, etc.
- analyse de $\mathbf{Q}$ pour un modèle statistique des données: choisir $\sigma(\cdot)$ en fonction des données

Conclusion

Messages:

- étude de la matrice de Gram non-linéaire: $\mathbf{G} \equiv \frac{1}{T}\boldsymbol{\Sigma}^{\mathsf{T}}\boldsymbol{\Sigma}$
- objet clé: la résolvante $\mathbf{Q}$ et son équivalent déterministe asymptotique $\bar{\mathbf{Q}}$
- permet la compréhension de la performance de l'ELM et projections aléatoires
- optimisation théorique des hyper-paramètres (ici γ)

Travaux à venir:

- analyse des méthodes plus sophistiquées basées sur les projections aléatoires, e.g., projections aléatoires + régression logistique, projections aléatoires + SVM, etc.
- analyse de $\mathbf{Q}$ pour un modèle statistique des données: choisir $\sigma(\cdot)$ en fonction des données
- combinaison de plusieurs types de non-linéarités, e.g., random Fourier features ($\sin + \cos \approx$ noyau Gaussien)

Conclusion

Messages:

- étude de la matrice de Gram non-linéaire: $\mathbf{G} \equiv \frac{1}{T}\boldsymbol{\Sigma}^{\mathsf{T}}\boldsymbol{\Sigma}$
- objet clé: la résolvante $\mathbf{Q}$ et son équivalent déterministe asymptotique $\bar{\mathbf{Q}}$
- permet la compréhension de la performance de l'ELM et projections aléatoires
- optimisation théorique des hyper-paramètres (ici γ)

Travaux à venir:

- analyse des méthodes plus sophistiquées basées sur les projections aléatoires, e.g., projections aléatoires + régression logistique, projections aléatoires + SVM, etc.
- analyse de $\mathbf{Q}$ pour un modèle statistique des données: choisir $\sigma(\cdot)$ en fonction des données
- combinaison de plusieurs types de non-linéarités, e.g., random Fourier features ($\sin + \cos \approx$ noyau Gaussien)

References:

C. Louart, Z. Liao, R. Couillet, "**A Random Matrix Approach to Neural Networks**", (à paraître) Annals of Applied Probability, 2017.

Merci!