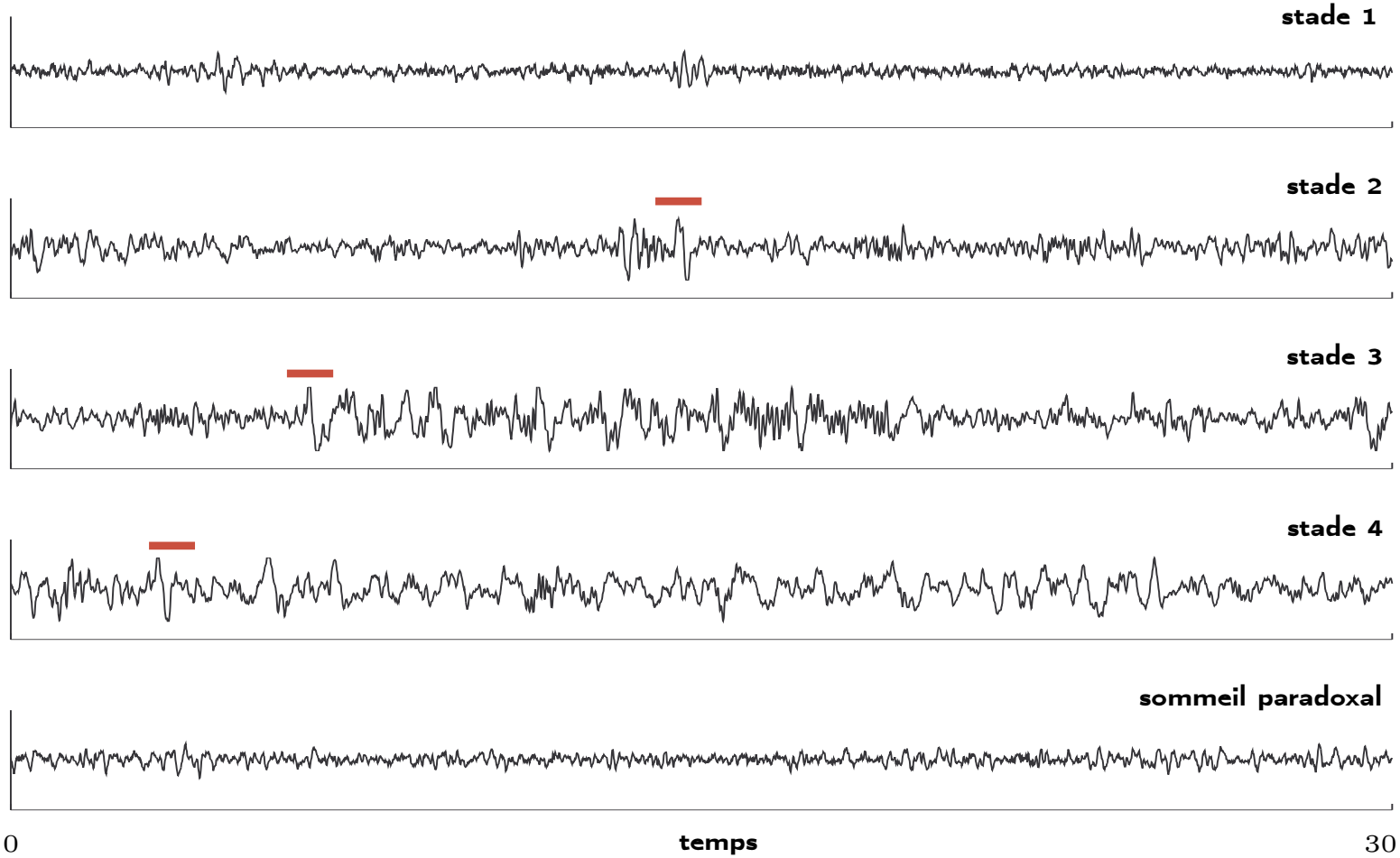
Reconnaissance des formes : applications en surveillance

Paul HONEINE

– Université de technologie de Troyes –

— 2013 —



Exemple : détection du complexe K dans l'EEG de sommeil.

Le problème considéré peut s'écrire :

$$\begin{cases} \omega_0 : \boldsymbol{x} = \boldsymbol{b} & \text{hypothèse "bruit seul"} \\ \omega_1 : \boldsymbol{x} = \boldsymbol{b} + \boldsymbol{s} & \text{hypothèse "signal et bruit"} \end{cases}$$

Il s'agit d'élaborer un détecteur d, par exemple de probabilité d'erreur minimale

$$P_e(d) = p(d(\boldsymbol{X}) \neq Y),$$

où $\boldsymbol{X}$ désigne une observation et Y l'hypothèse associée.

La stratégie à adopter pour apporter une solution à ce problème dépend de la nature de l'information disponible sur $(\boldsymbol{X}, Y)$.

Détection à structure libre.

En se limitant à des hypothèses simples, l'application d'une règle de décision telle que celle de Bayes conduit à

$$d^*(\boldsymbol{x}) = \begin{cases} 1 & \text{si} \quad p(\boldsymbol{x}|\omega_1)/p(\boldsymbol{x}|\omega_0) \geq \lambda_0 \\ 0 & \text{sinon,} \end{cases}$$

à condition de connaître au moins $p(\boldsymbol{x}|\omega_0)$ et $p(\boldsymbol{x}|\omega_1)$. Le seuil λ_0 est le seul paramètre dépendant de la règle choisie.

Ainsi, le détecteur n'est assujetti à aucune contrainte structurelle mais résulte du choix d'un critère.

Détection à structure imposée.

L'ignorance des propriétés statistiques de l'échantillon impose la mise en œuvre d'une stratégie alternative, qui peut être

1. définir une classe de détecteurs $\mathcal{D} = \{d(\boldsymbol{x}, \theta) \ : \ \theta \in \Theta\}$
2. sélectionner l'élément de $\mathcal{D}$ le plus performant

Simple en apparence, cette approche suppose toutefois que l'on réponde de façon satisfaisante aux questions qui suivent :

1. Comment choisir la classe de détecteurs $\mathcal{D}$?
2. Quelles sont les fonctionnelles de risque pertinentes pour le problème traité ?
3. Quelle procédure d'optimisation adopter ?

Part 1

Chapitre 1 *: Eléments de théorie statistique de l'apprentissage*
→ *apprentissage fonctionnel, consistance, capacité en généralisation, etc.*

Chapitre 2 *: Régularisation*
→ *problèmes bien et mal posés, régularisations de Tiknonov, etc.*

Chapitre 3 *: Méthodes à noyau*
→ *RKHS, condition de Mercer, exemple de kernelisation, etc.*

Chapitre 4 *: Support Vector Machines*
→ *optimisation sous contrainte, hyperplan optimum, extension au cas non-séparable, etc.*

Part 2

...

Eléments de théorie statistique de l'apprentissage

La connaissance d'un modèle probabiliste est remplacée par celle d'un ensemble d'apprentissage $\mathcal{A}_n$:

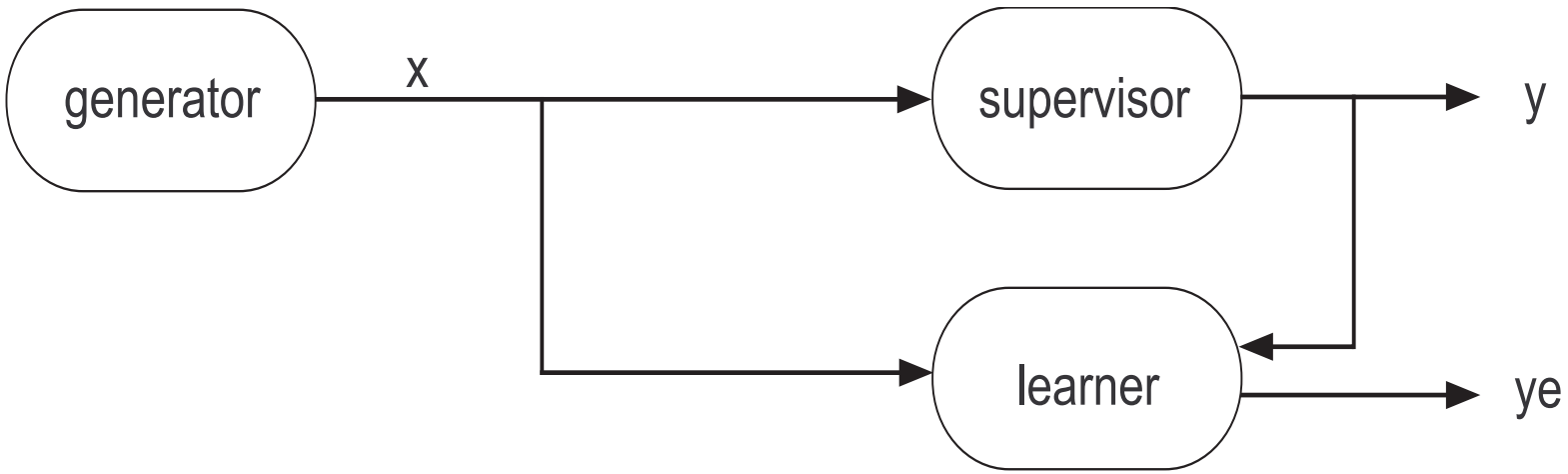
$$\mathcal{A}_n = \{(\boldsymbol{x}_1, y_1), (\boldsymbol{x}_2, y_2), \ldots, (\boldsymbol{x}_n, y_n)\}.$$

L'élaboration d'une règle de décision consiste à rechercher une partition de l'espace des observations $\mathcal{X}$ qui soit optimale au sens du critère de performance choisi.

On distingue principalement deux approches possibles :

1. Choix préalable de la structure de la règle de décision, puis optimisation des paramètres caractéristiques selon le critère retenu.
2. Utilisation directe de l'ensemble d'apprentissage pour la prise de décision.

Le modèle d'apprentissage comporte 3 éléments :



1. Générateur : $\boldsymbol{X} \in \mathcal{X} \subset \mathbb{R}^l$, des vecteurs aléatoires *i.i.d.*
2. Superviseur : $Y \in \mathcal{Y} \subset \mathbb{R}$, des variables aléatoires
3. Apprenti : représenté par $d(\boldsymbol{x}; \theta) \in \mathcal{D}$

– **Polynômes de degré p**

$$d(\boldsymbol{x};\boldsymbol{a}) = \sum_{\substack{i_1,\ldots,i_l \in \mathbb{N} \\ i_1+\ldots+i_l \leq p}} a_{i_1,\ldots,i_l}\ x[1]^{i_1} \ldots x[l]^{i_l}$$

..., et autres décompositions sur une base de Fourier, de Haar, ...

– **Splines**

$$d(\boldsymbol{x};c) \in \mathcal{L}^2(\mathbb{R}^l) \text{ tel que } d' \in \mathcal{L}^2(\mathbb{R}^l), \|d'\|^2 \leq c$$

– **Nadaraya-Watson**

$$d(\boldsymbol{x};\sigma) = \frac{\sum_{i=1}^n y_i\ K_\sigma(\boldsymbol{x},\boldsymbol{x}_i)}{\sum_{i=1}^n\ K_\sigma(\boldsymbol{x},\boldsymbol{x}_i)}$$

– **MLP, RBF, ...**

$$d(\boldsymbol{x};\boldsymbol{a},\boldsymbol{\theta}) = \sum_k a_k\ g_k(\boldsymbol{x};\boldsymbol{\theta}_k)$$

Objectif

Rechercher au sein de $\mathcal{D} = \{d(\boldsymbol{x}, \theta) \,:\, \theta \in \Theta\}$ une fonction réalisant la meilleure approximation de y au sens d'une fonctionnelle de risque de la forme

$$J(d) = \int Q(d(\boldsymbol{x}, \theta), y)\, p(\boldsymbol{x}, y)\, d\boldsymbol{x}\, dy,$$

où Q représente le coût associé à chaque couple $(\boldsymbol{x}, y)$.

Exemple de fonction coût : probabilité d'erreur
Lorsqu'il s'agit d'élaborer une structure de décision de probabilité d'erreur minimale, le risque s'exprime ainsi

$$P_e(d) = \int \mathbb{1}_{d(\boldsymbol{x},\theta)\neq y}\, p(\boldsymbol{x}, y)\, d\boldsymbol{x}\, dy,$$

où $\mathbb{1}$ désigne la fonction indicatrice.

– **Coût quadratique**

$$Q(\boldsymbol{x}, y) = (y - d(\boldsymbol{x}; \theta))^2 \quad \rightarrow \quad d^*(\boldsymbol{x}; \theta) = \mathrm{E}(y \,|\, \boldsymbol{x})$$

– **Coût absolu**

$$Q(\boldsymbol{x}, y) = |y - d(\boldsymbol{x}; \theta)|$$

– **Entropie croisée**

$$Q(\boldsymbol{x}, y) = -y \log(d(\boldsymbol{x}; \theta)) - (1 - y) \log(1 - d(\boldsymbol{x}; \theta)) \quad \rightarrow \quad d^*(\boldsymbol{x}; \theta) = \mathrm{P}(y = 1 \,|\, \boldsymbol{x})$$

Il s'agit de minimiser la fonctionnelle de risque

$$J(d) = \int Q(d(\boldsymbol{x};\theta), y)\, p(\boldsymbol{x}, y)\, d\boldsymbol{x}\, dy,$$

la densité $p(\boldsymbol{x}, y)$ étant inconnue.

Minimisation du risque empirique (MRE)

La minimisation de $J(d)$ se traduit par celle du risque empirique

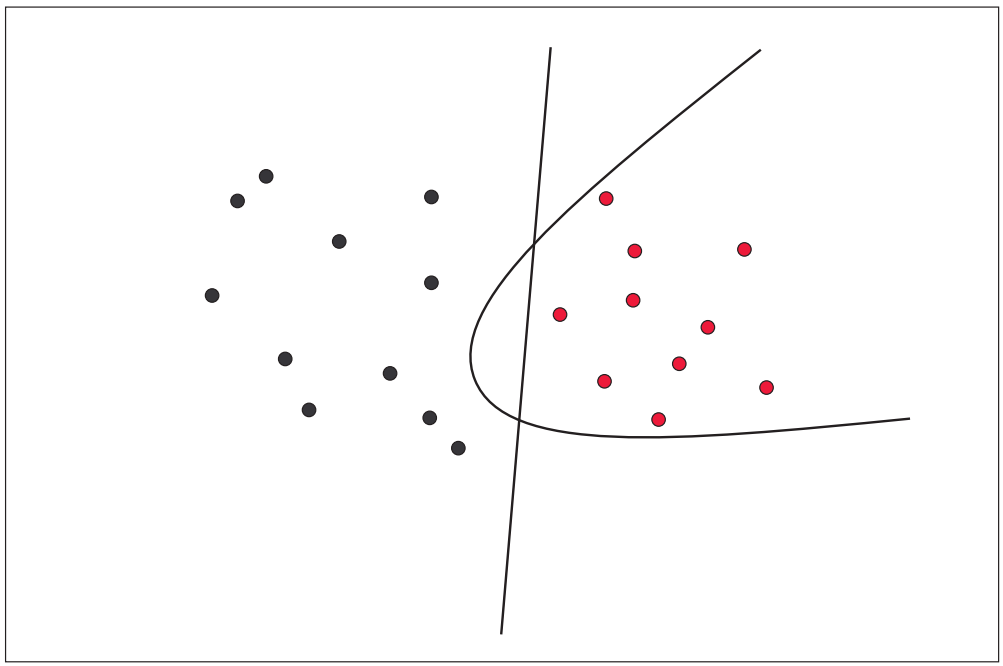
$$J_{emp}(d) = \frac{1}{n} \sum_{k=1}^{n} Q(d(\boldsymbol{x}_k;\theta), y_k),$$

calculable sur les données constituant l'ensemble d'apprentissage $\mathcal{A}_n$.

Probabilité d'erreur empirique.
Le risque empirique associé à la probabilité d'erreur correspond au nombre d'erreurs d'affectation commises par $d(\boldsymbol{x};\theta)$ sur $\mathcal{A}_n$

$$P_{emp}(d) = \frac{1}{n} \sum_{k=1}^{n} \mathbb{1}_{d(\boldsymbol{x}_k;\theta) \neq y_k}.$$

Problème. Deux familles gaussiennes ω_0 et ω_1 dans $\mathbb{R}^2$, de moyennes et matrices de covariance distinctes, constituées de 10 échantillons chacune.



Quelle frontière choisir ?
Que dire de $\hat{P}_e(\text{linéaire}) = 5\%$ tandis que $\hat{P}_e(\text{quadratique}) = 9\%$?

On note $d^* = \arg\min J(d)$ la règle de risque minimum, et $d_n^* = \arg\min_{d\in\mathcal{D}} J_{emp}(d)$ celle obtenue par minimisation du risque empirique sur $\mathcal{D}$ à partir de $\mathcal{A}_n$.

Définition (Erreur d'estimation)

C'est la différence de performance entre la meilleure règle de $\mathcal{D}$ et celle obtenue au terme de l'apprentissage :

$$J_{estim} = J_e(d_n^*) - \inf_{d\in\mathcal{D}} J_e(d)$$

▷ *pertinence du critère empirique et performance de l'algorithme*

Définition (Erreur d'approximation)

Elle est donnée par la différence de performance entre la règle optimum d^* et la meilleure au sein de $\mathcal{D}$:

$$J_{approx} = \inf_{d\in\mathcal{D}} J_e(d) - J_e(d^*)$$

▷ *choix de la classe $\mathcal{D}$*

Apprentissage

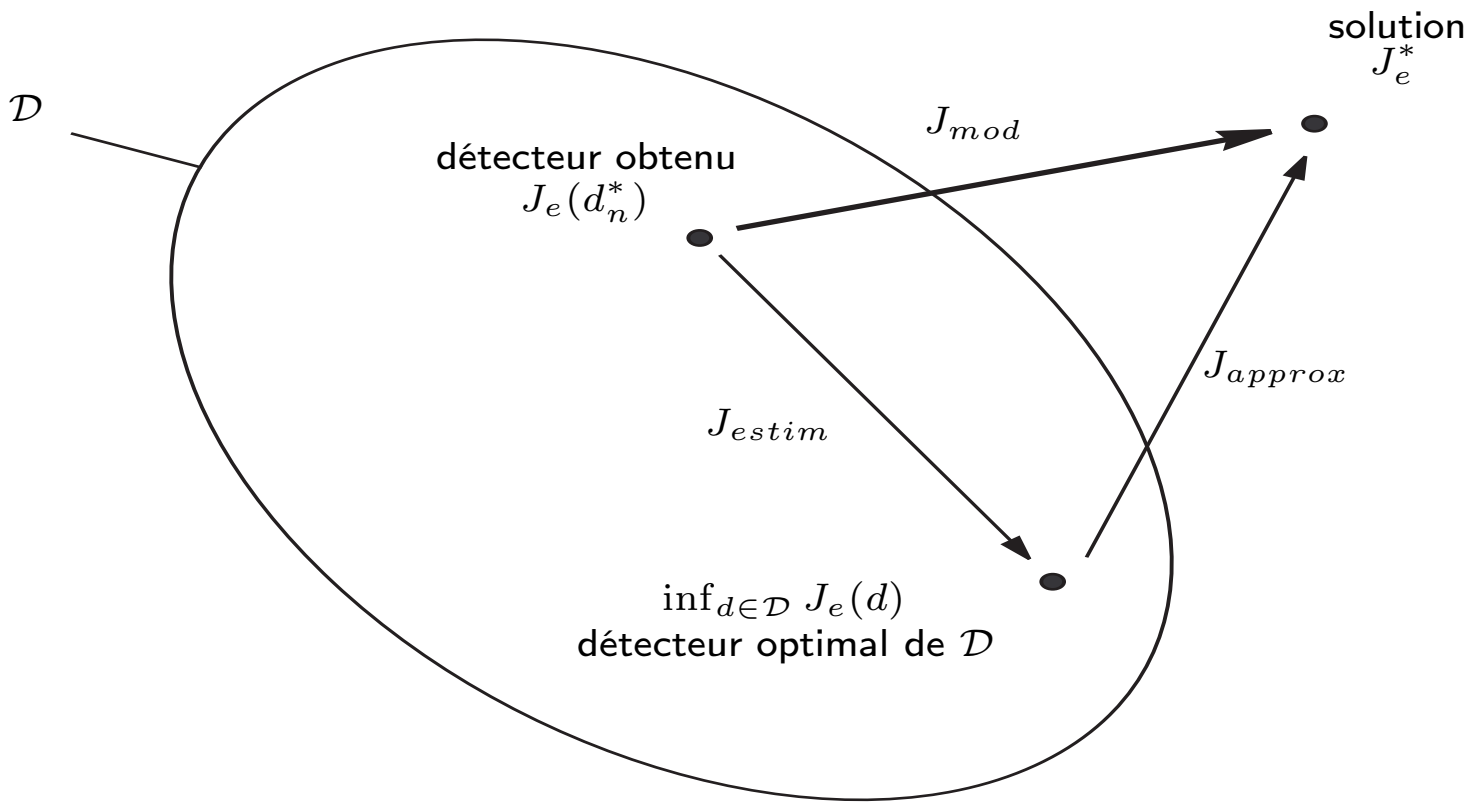
L'objectif de l'apprentissage est de minimiser l'erreur de modélisation, définie par :

$$J_{mod}(d_n^*) = J_e(d_n^*) - J_e(d^*).$$

On distingue deux contributions de natures différentes dans cette erreur :

$$J_{mod}(d_n^*) = \underbrace{\left(J_e(d_n^*) - \inf_{d \in \mathcal{D}} J_e(d) \right)}_{J_{estim}} + \underbrace{\left(\inf_{d \in \mathcal{D}} J_e(d) - J_e(d^*) \right)}_{J_{approx}}.$$

La minimisation de J_{mod} repose sur la recherche d'un compromis entre ces deux termes antagonistes : l'augmentation du nombre de tests de $\mathcal{D}$ conduit à un accroissement de J_{estim} tandis que J_{approx} décroît, et inversement.


solution
J_e^*
$\mathcal{D}$
J_{mod}
détecteur obtenu
$J_e(d_n^*)$
J_{approx}
J_{estim}
$\inf_{d\in\mathcal{D}} J_e(d)$
détecteur optimal de $\mathcal{D}$

1. *L'objectif est-il réalisable ?*

$\rightarrow$ *Consistance de la règle de décision*

$\rightarrow$ *Consistance du principe d'induction*

$\rightarrow$ *Vitesse de convergence*

2. *: Si oui, comment en pratique ?*

On peut espérer qu'il existe dans la classe $\mathcal{D}$ considérée, une suite $\{d_n^*(\boldsymbol{X};\theta)\}_{n>0}$ de détecteurs optimaux au sens du critère retenu telle que $P_e(d_n^*)$ puisse être rendue arbitrairement proche de P_e^* lorsque n tend vers l'infini.

Définition (Consistance et consistance forte)

Etant donnée une base $\mathcal{A}_n$, une suite $\{d_n^*(\boldsymbol{X};\theta)\}_{n>0}$ de détecteurs optimaux au sens d'un critère donné est dite consistante pour une loi $p(\boldsymbol{x},y)$ si

$$\lim_{n\to\infty} \mathrm{E}\{P_e(d_n^*;\mathcal{A}_n)\} = P_e^*.$$

On dit qu'elle est fortement consistante si, avec une probabilité égale à 1,

$$\lim_{n\to\infty} P_e(d_n^*;\mathcal{A}_n) = P_e^*.$$

On peut distinguer le cas où la propriété de consistance n'est vérifiée que pour une loi $p(\boldsymbol{x}, y)$ donnée, du cas où elle reste vraie indépendamment de celle-ci.

Définition (Consistance universelle)

La suite $\{d_n^*(\boldsymbol{X}; \theta)\}_{n>0}$ est dite universellement (fortement) consistante si elle est (fortement) consistante pour toute loi de probabilité $p(\boldsymbol{x}, y)$.

Cette propriété a été observée pour la première fois en 1977 par Stone dans le cadre de la méthode des *k plus proches voisins*, à la condition que le paramètre k croisse moins vite que la taille n de la base d'apprentissage. Depuis, il a été démontré que d'autres règles de décision y satisfont :

- *fonctions à noyaux réguliers*
- *certains détecteurs linéaires généralisés*
- *(...)*

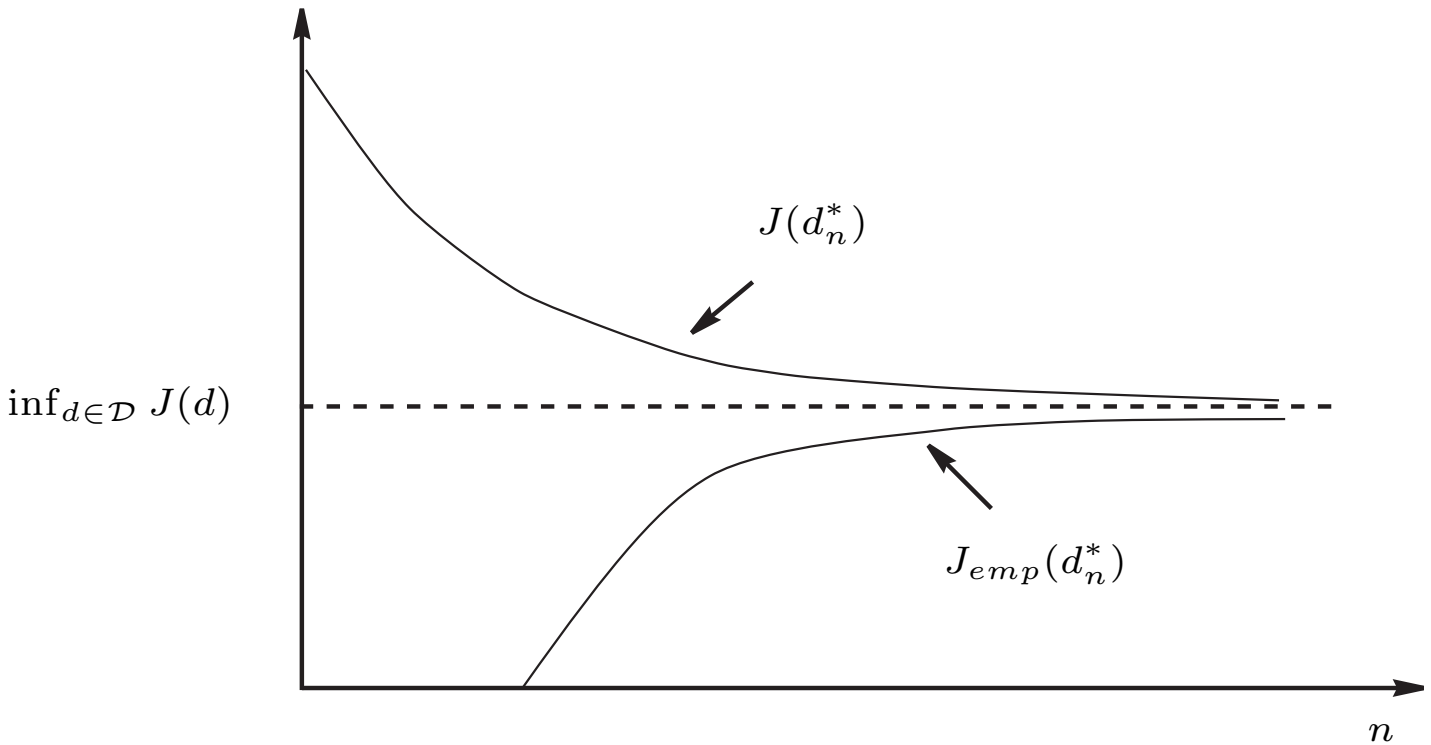
Le principe de MRE est consistant pour la fonction objectif choisie et le problème, si l'apprenti fait du mieux possible quand la taille de l'échantillon tend vers l'infini.

Consistance du principe de MRE

Le principe de MRE est consistant pour un coût Q, une famille de fonctions $\mathcal{D} = \{d(\boldsymbol{x}; \theta) : \theta \in \Theta\}$ et une distribution $p(\boldsymbol{x}, y)$ si, appliqué à chaque taille n d'échantillon, il engendre une suite $\{d_n^*(\boldsymbol{x}; \theta) : \theta \in \Theta\}_{n>0}$ telle que

$$J(d_n^*) \xrightarrow[n\to\infty]{p} \inf_{d\in\mathcal{D}} J(d)$$

$$J_{emp}(d_n^*) \xrightarrow[n\to\infty]{p} \inf_{d\in\mathcal{D}} J(d).$$



$$J(d_n^*) \xrightarrow[n\to\infty]{p} \inf_{d\in\mathcal{D}} J(d)$$

$$J_{emp}(d_n^*) \xrightarrow[n\to\infty]{p} \inf_{d\in\mathcal{D}} J(d)$$

Par soucis de clarté, on considère dans la suite de cette section que le coût Q prend la forme d'une fonction indicatrice, soit

$$Q(d(\boldsymbol{x};\theta);y) = \mathbb{1}_{d(\boldsymbol{x};\theta)\neq y} \triangleq \begin{cases} 0 & \text{si} \quad y = d(\boldsymbol{x};\theta) \\ 1 & \text{si} \quad y \neq d(\boldsymbol{x};\theta), \end{cases}$$

Définition (VC-dimension)

La dimension de Vapnik-Chervonenkis d'une classe $\mathcal{D}$ donnée est définie par le plus grand nombre d'éléments $\boldsymbol{x}_k$ de l'espace des réalisations $\mathcal{X}$ dont les détecteurs de $\mathcal{D}$ peuvent réaliser toutes les dichotomies.

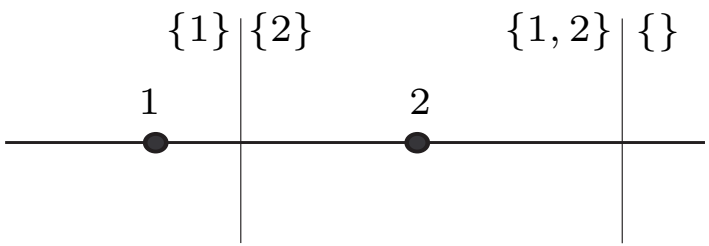
Exemple 1. On considère la classe $\mathcal{D}$ des détecteurs linéaires opérant dans $\mathbb{R}^l$ définis par $d(\boldsymbol{x};\theta) = \text{sign}(\sum_{k=1}^{l} \theta_k\, x(k) + \theta_0)$, les paramètres θ_k étant réels et $\text{sign}(\cdot)$ désignant la fonction « signe ». On montre que

$$h_{\mathcal{D}} = l + 1$$

Exemple 2. On considère la classe des détecteurs $\{d(x;\theta) = \text{sign}(\sin(\theta x)) \,:\, \theta \in \mathbb{R}\}$ opérant dans $\mathbb{R}$. Il est aisé de démontrer que

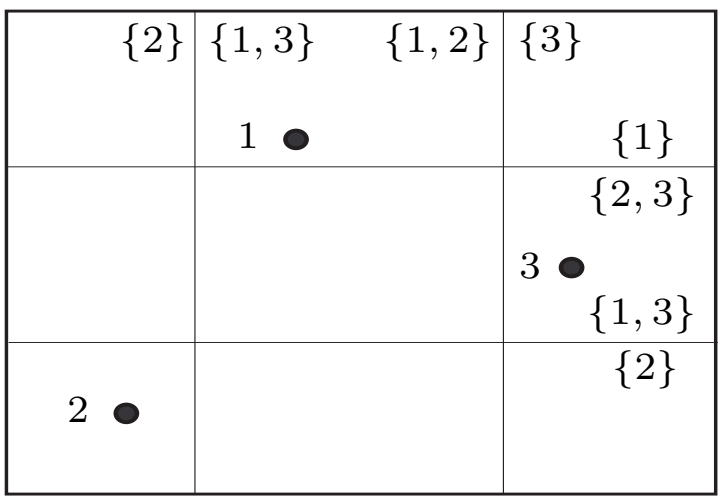
$$h_{\mathcal{D}} = +\infty$$

dans $\mathbb{R}$

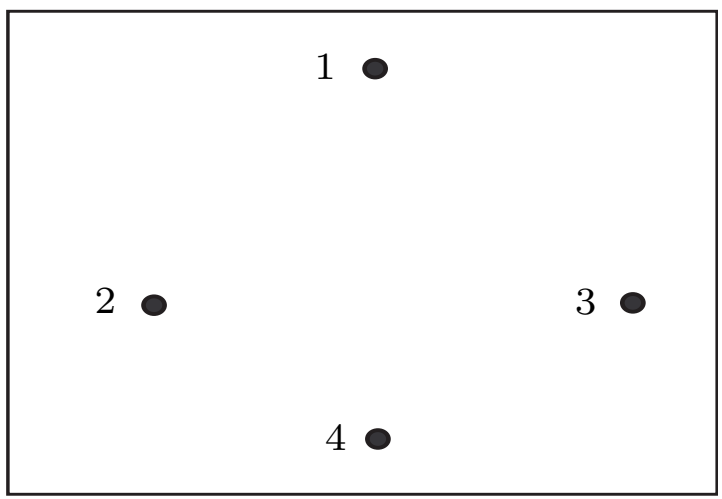


dans $\mathbb{R}$

dans $\mathbb{R}^2$



dans $\mathbb{R}^2$



Théorème

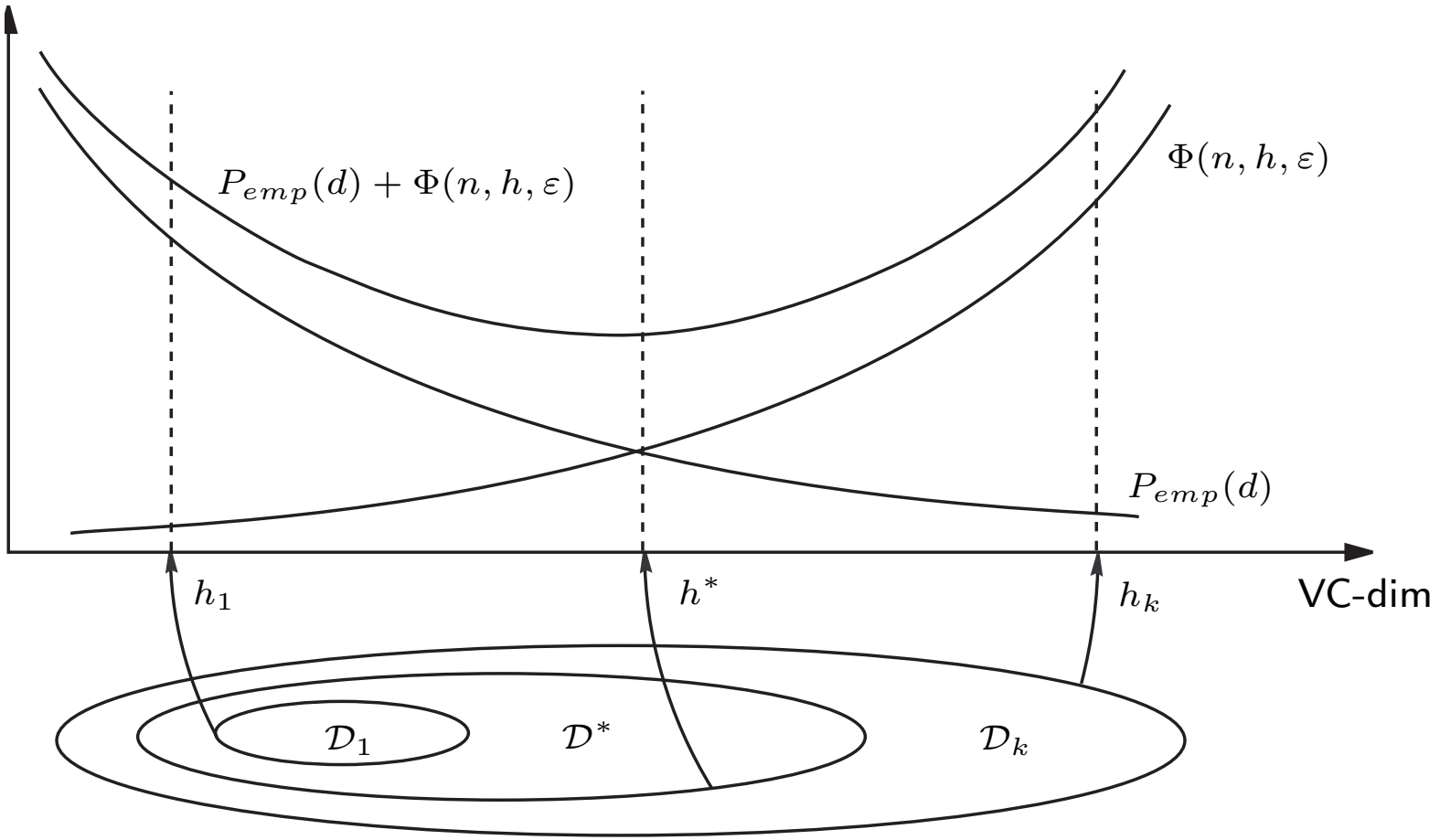
Pour que le principe MRE soit consistant indépendamment de la distribution de probabilité gouvernant les observations, il suffit que la classe de détecteurs considérée soit de VC-dimension h finie.

Les travaux précurseurs de Vapnik et Chervonenkis (1971) ont également apporté des enseignements quantitatifs relatifs à la vitesse de convergence de P_{emp} vers P_e.

Inégalité de Vapnik-Chervonenkis.
Avec une probabilité égale à $1 - \varepsilon$ au moins, on a :

$$P_e(d) \leq P_{emp}(d) + \sqrt{\frac{h \ln\left(\frac{2en}{h}\right) - \ln\frac{\varepsilon}{4}}{n}}.$$

Attention ! Majoration souvent grossière... mais indépendante de toute loi $p(\boldsymbol{x}, y)$.


$P_{emp}(d) + \Phi(n, h, \varepsilon)$
$\Phi(n, h, \varepsilon)$
$P_{emp}(d)$
h_1
h^*
h_k
VC-dim
$\mathcal{D}_1$
$\mathcal{D}^*$
$\mathcal{D}_k$

Le principe de *minimisation du risque structurel* préconisé par Vapnik suppose la construction, au sein de la classe $\mathcal{D}$, d'une séquence de sous-ensembles imbriqués $\mathcal{D}_k$

$$\mathcal{D}_1 \subset \ldots \subset \mathcal{D}_k \subset \ldots \subset \mathcal{D}.$$

Cette structure étant établie, la phase d'apprentissage est menée en deux étapes :

1. Recherche du détecteur d'erreur empirique minimale dans chaque sous-ensemble $\mathcal{D}_k$:
$$d^*_{n,k} = \arg \min_{d \in \mathcal{D}_k} P_{emp}(d).$$
2. Sélection du détecteur présentant l'erreur garantie $P_{emp}(d^*_{n,k}) + \Phi(n, h_k, \varepsilon)$ la plus favorable :
$$d^*_n = \arg \min_{k \geq 1} \{P_{emp}(d^*_{n,k}) + \Phi(n, h_k, \varepsilon)\}.$$

Régularisation et noyau reproduisant

Problème d'apprentissage :

On est à la recherche d'une fonction ψ d'un espace $\mathcal{H}$ de fonctions candidates de $\mathcal{X}$ dans $\mathcal{Y}$, qui, pour un $\boldsymbol{x}$, prédit l'étiquette correspondante y, soit

$$y = \psi(\boldsymbol{x})$$

On dispose d'un ensemble d'apprentissage $\mathcal{A}_n = \{(\boldsymbol{x}_1, y_1), \ldots, (\boldsymbol{x}_n, y_n)\}$

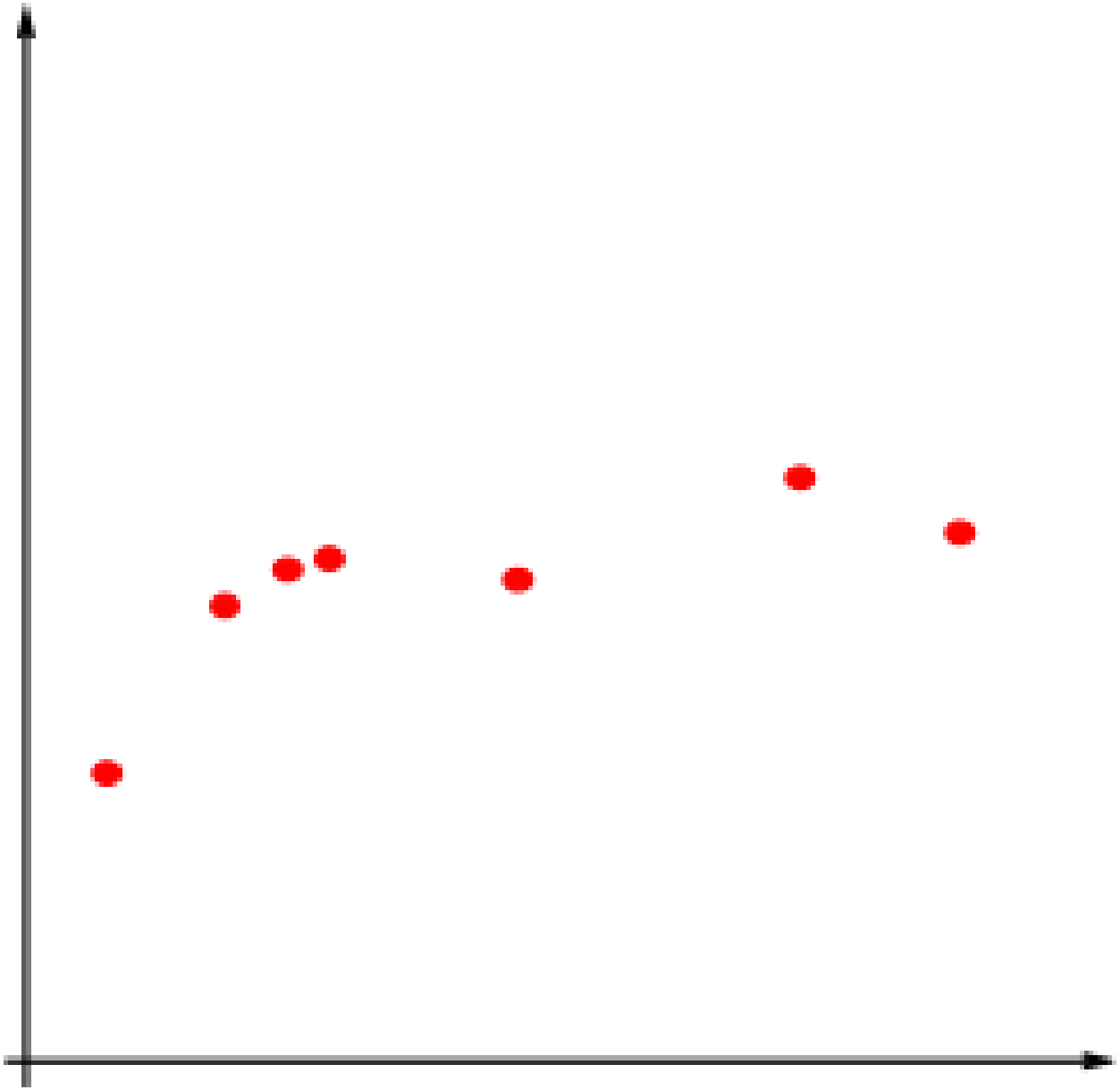
▷ *Minimisation du risque empirique et Généralisation !*

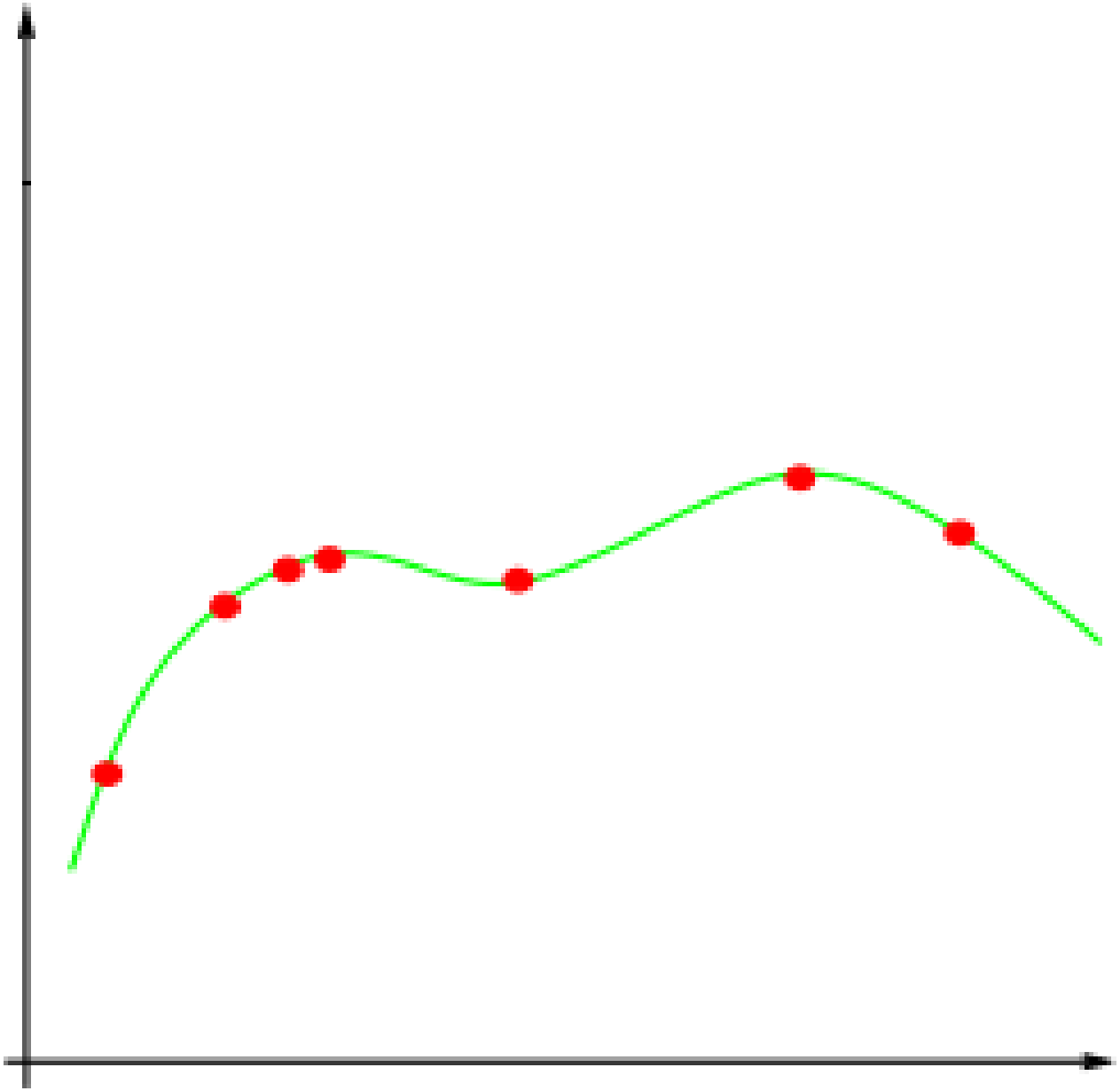
Definition (Problème bien-posé / problème mal-posé (Hadamard))

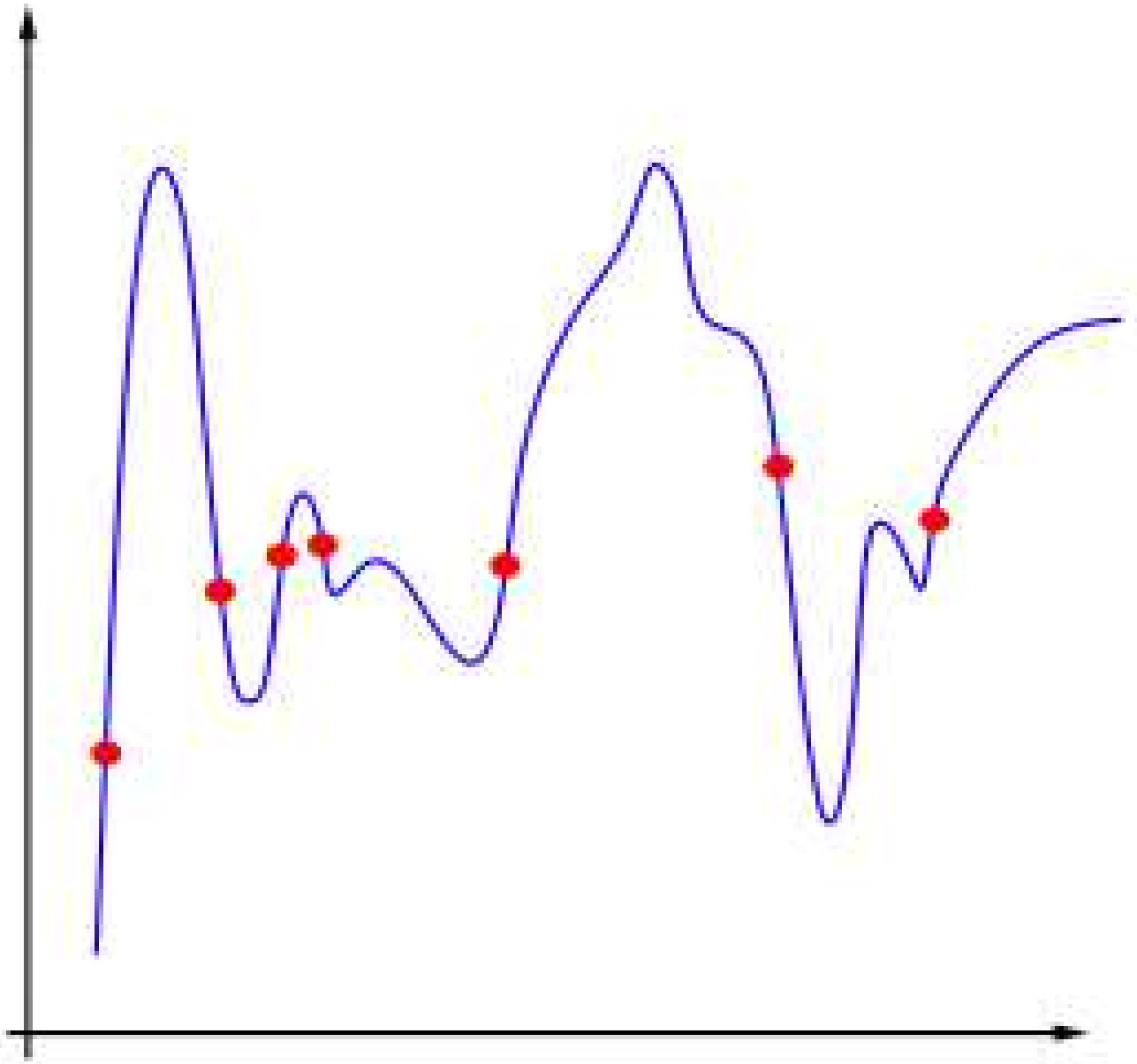
Un problème est dit *bien-posé* si

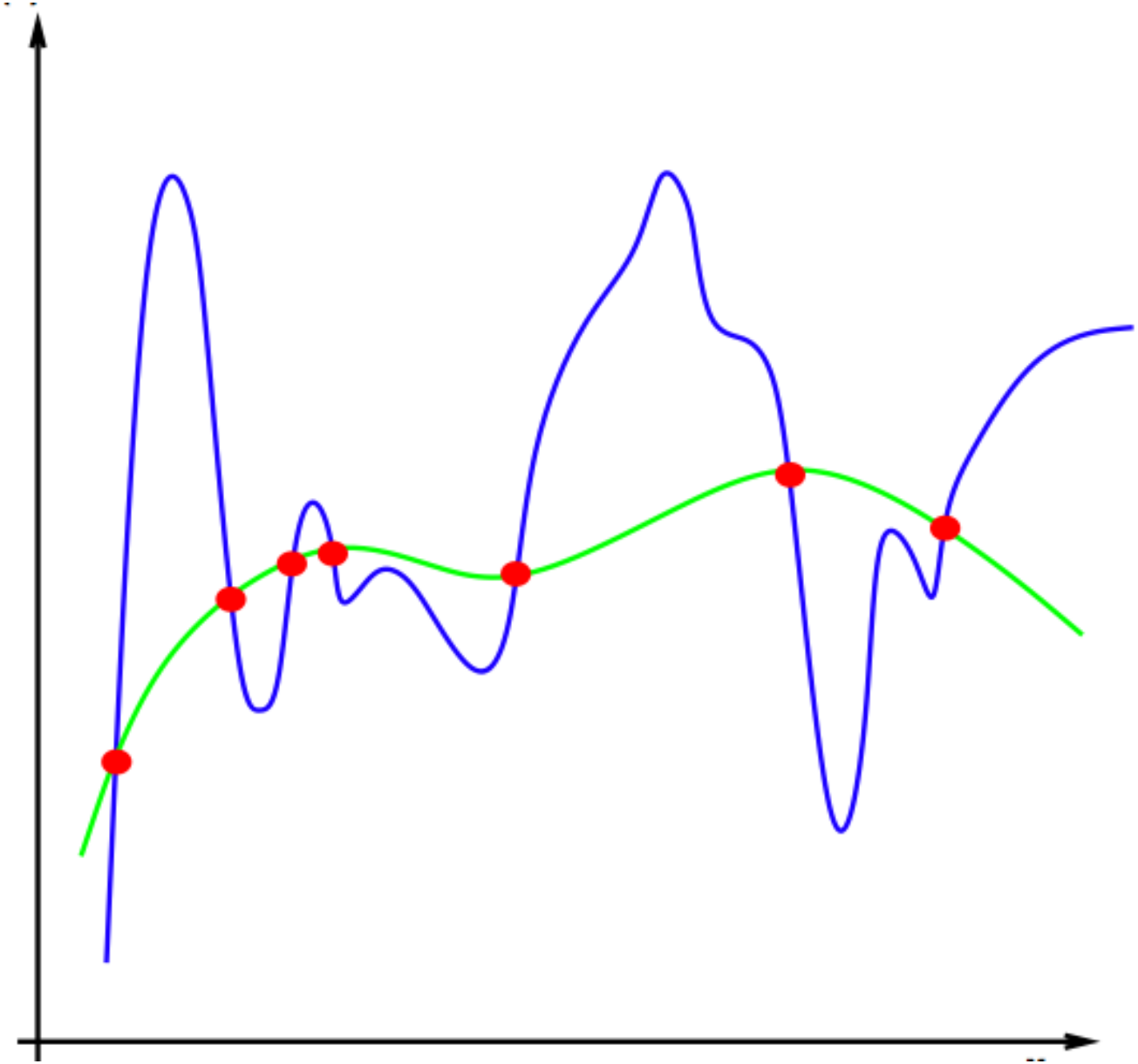
- la solution existe
- la solution est unique
- la solution est une fonction continue des données (une faible perturbation des données conduit à une faible perturbation de la solution)

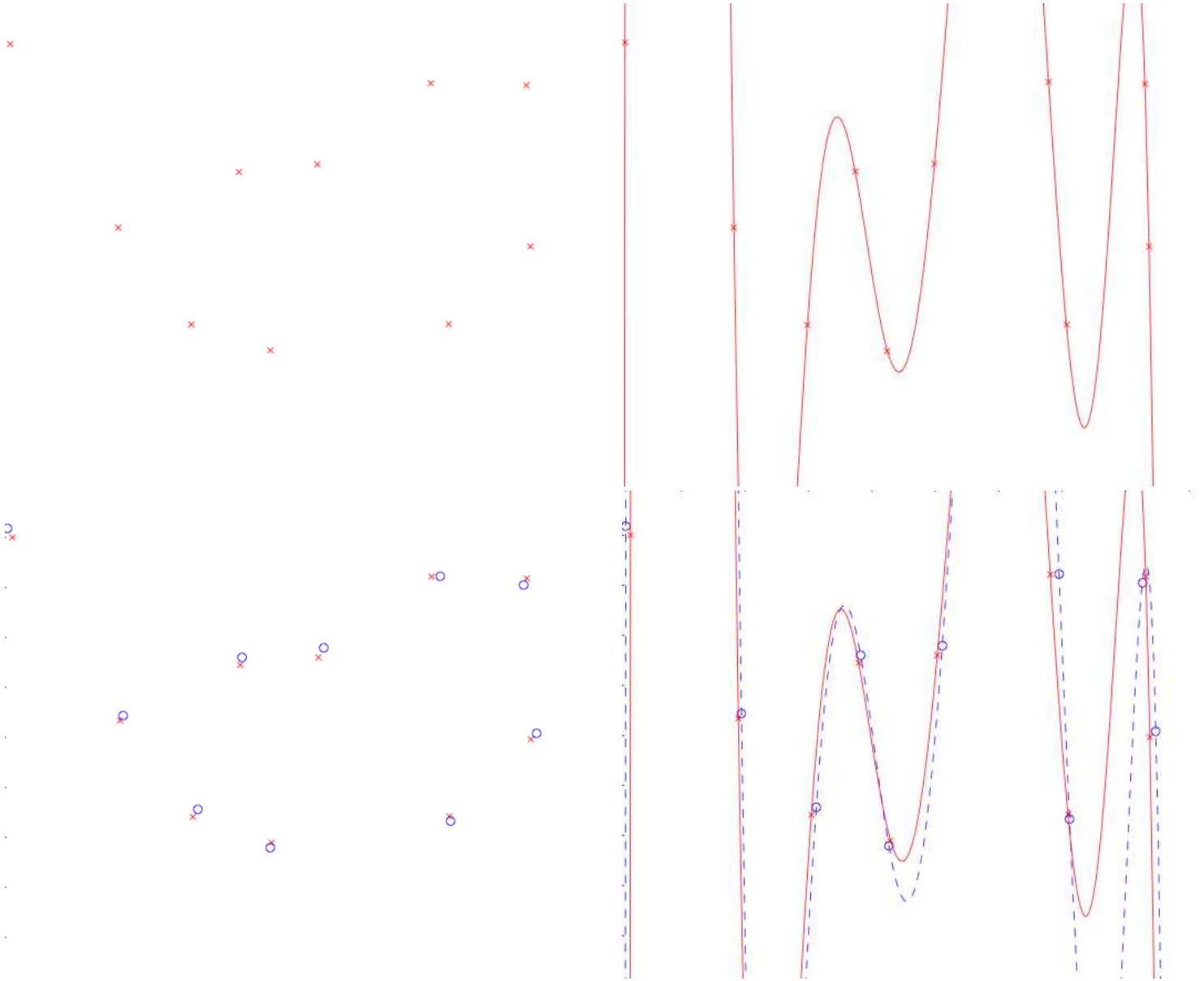
Un problème est dit *mal-posé* s'il n'est pas bien-posé

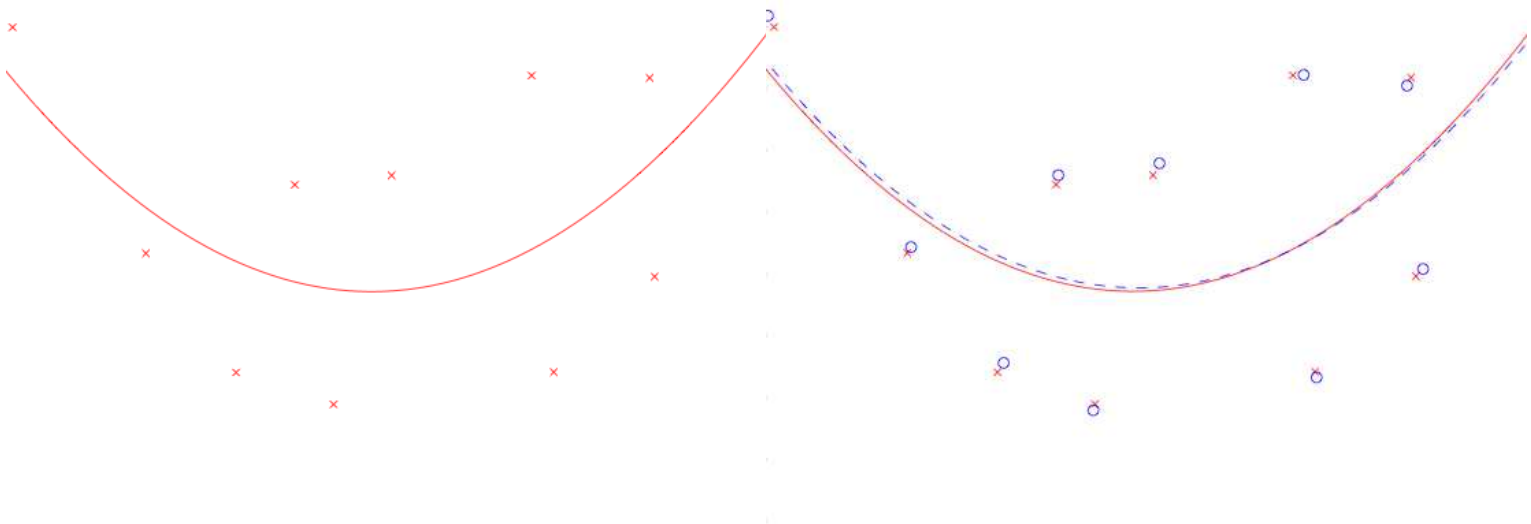












La minimisation du risque empirique

$$J_{emp}(\psi) = \frac{1}{n} \sum_{k=1}^{n} Q(\psi(\boldsymbol{x}_k), y_k),$$

est un problème mal-posé.

Solution : **Régularisation**

Régularisation d'Ivanov

Déterminer la fonction ψ qui minimise

$$\frac{1}{n}\sum_{k=1}^{n} Q(\psi(\boldsymbol{x}_k), y_k),$$

sous la contrainte

$$\|\psi\|^2 \leq A$$

Pénalisation du risque empirique :

$$\text{RisqEmp}(\psi) + \eta\ \text{Pén}(\psi),$$

avec η est un paramètre positif controlant le compromis entre ces deux termes.

▷ *Le terme de pénalisation permet d'incorporer un effet de lissage*

Régularisation de Tikhonov

Déterminer la fonction ψ d'un espace $\mathcal{H}$ de fonctions candidates, minimisant

$$\frac{1}{n}\sum_{k=1}^{n} Q(\psi(\boldsymbol{x}_k), y_k) + \eta\|\psi\|_{\mathcal{H}}^2,$$

pour un paramètre $\eta > 0$, et où $\|\psi\|_{\mathcal{H}}$ est la norme fonctionnelle dans l'espace $\mathcal{H}$.

Ce problème est bien-posé.

Espace de Hilbert à noyau reproduisant

Espace fonctionnel $\mathcal{H}$

Une *norme* $\|\cdot\|$sur un espace $\mathcal{H}$ est une application de $\mathcal{H}$ dans $\mathbb{R}$, nonnégative, vérifiant les propriétés suivantes, pour tout $\psi, \phi \in \mathcal{H}$,

- Positivité : $\|\psi\| \geq 0$, avec égalité si et seulement si $\psi = 0$
- Homogénéité : $\|\alpha\psi\| = |\alpha| \ \|\psi\|$ pour tout $\alpha \in \mathbb{R}$
- Inégalité triangulaire : $\|\psi + \phi\| \leq \|\psi\| + \|\phi\|$

Un *produit scalaire* $\langle \cdot, \cdot \rangle$ est une application de $\mathcal{H} \times \mathcal{H}$ dans $\mathbb{R}$ qui vérifie les propriétés suivantes

- Positivité : $\langle \psi, \psi \rangle \geq 0$, avec égalié si et seulement si $\psi = 0$
- Bilinéarité : $\langle \alpha\psi, \phi \rangle = \alpha \langle \phi, \psi \rangle$ et $\langle \psi_1 + \psi_2, \phi \rangle = \langle \psi_1, \phi \rangle + \langle \psi_2, \phi \rangle$
- Symétrie : $\langle \psi, \phi \rangle = \langle \phi, \psi \rangle$

On peut définir une norme à partir d'un produit scalaire, avec $\|\psi\|^2 = \langle \psi, \psi \rangle$.

L'espace fonctionnel muni d'un produit scalaire (et complet pour la norme associée) est dit *espace de Hilbert*.

Exemple : $\mathcal{L}_2[a, b] = \{\psi \mid \int_a^b |\psi^2(x)| dx < \infty\}$ est un espace de Hilbert où le produit scalaire est donné par

$$\langle \psi, \phi \rangle = \int_a^b \psi(x)\phi(x)dx$$

Fonctionnelle d'évaluation :
Une fonctionnelle (linéaire) $\delta_{\boldsymbol{x}} : \mathcal{H} \to \mathbb{R}$ est dite d'évaluation si elle évalue toute fonction ψ de cet espace $\mathcal{H}$ en tout $\boldsymbol{x} \in \mathcal{X}$, soit

$$\delta_{\boldsymbol{x}}(\psi) = \psi(\boldsymbol{x})$$

Definition (espace de Hilbert à noyau reproduisant (RKHS))

Un espace de Hilbert est dit *à noyau reprodusant* si et seulement si, pour tout $\boldsymbol{x} \in \mathcal{X}$, la fonctionnelle d'évaluation $\delta_{\boldsymbol{x}}$ est bornée.

En d'autres termes, il existe un M tel que, pour tout $\psi \in \mathcal{H}$,

$$|\delta_{\boldsymbol{x}}(\psi)| = |\psi(\boldsymbol{x})| \leq M \|\psi\|_{\mathcal{H}}.$$

Théorème de (représentation de) Riesz :
Si $\mathcal{H}$ est un RKHS, et d"après le théorème (de représentation) de Riesz, il existe pour tout $\boldsymbol{x} \in \mathcal{X}$ une fonction unique $\kappa(\cdot, \boldsymbol{x})$ de $\mathcal{H}$ telle que

$$\psi(\boldsymbol{x}) = \langle \psi, \kappa(\cdot, \boldsymbol{x}) \rangle_{\mathcal{H}} \qquad \forall \psi \in \mathcal{H}$$

Propriété reproduisante :

$$\kappa(\boldsymbol{x}_i, \boldsymbol{x}_j) = \langle \kappa(\cdot, \boldsymbol{x}_i), \kappa(\cdot, \boldsymbol{x}_j) \rangle_{\mathcal{H}} \qquad \kappa(\boldsymbol{x}_i, \boldsymbol{x}_i) = \|\kappa(\cdot, \boldsymbol{x}_i)\|^2_{\mathcal{H}}$$

Unicité : Pour un RKHS, son noyau reproduisant est unique
Eléments de preuve :
$0 \leq \|\kappa_1(\cdot, \boldsymbol{x}) - \kappa_2(\cdot, \boldsymbol{x})\|^2_{\mathcal{H}} = \kappa_1(\boldsymbol{x}, \boldsymbol{x}) - \kappa_2(\boldsymbol{x}, \boldsymbol{x}) - \kappa_1(\boldsymbol{x}, \boldsymbol{x}) + \kappa_2(\boldsymbol{x}, \boldsymbol{x}) = 0$

Définition : Un noyau est dit *défini positif* si

$$\sum_{i,j} \alpha_i \alpha_j \kappa(\boldsymbol{x}_i, \boldsymbol{x}_j) \geq 0$$

pour tout $\alpha_i, \alpha_j \in \mathbb{R}$ et $\boldsymbol{x}_i, \boldsymbol{x}_j \in \mathcal{X}$.

Theorem

Un noyau reproduisant est un noyau défini positif

Eléments de preuve :
$\sum_{i,j} \alpha_i \alpha_j \kappa(\boldsymbol{x}_i, \boldsymbol{x}_j) = \sum_{i,j} \alpha_i \alpha_j \langle \kappa(\cdot, \boldsymbol{x}_i), \kappa(\cdot, \boldsymbol{x}_j) \rangle_{\mathcal{H}} = \| \sum_i \alpha_i \kappa(\cdot, \boldsymbol{x}_i) \|^2_{\mathcal{H}}$

Theorem (Aronszajn)

A tout noyau κ défini positif correspond un unique espace de Hilbert à noyau reproduisant, dont le noyau reproduisant est κ.

Eléments de preuve :
Il suffit de compléter $\mathcal{H}_0$, l'espace engendré par les fonctions noyau, de sorte que toute suite de Cauchy y converge, avec $\mathcal{H}_0 = \{\psi \mid \psi = \sum_i \alpha_i \kappa(\cdot, \boldsymbol{x}_i), \alpha_i \in \mathbb{R}, \boldsymbol{x}_i \in \mathcal{X}\}$.
L'espace fonctionnel résultant $\mathcal{H}$ est muni du produit scalaire défini pour tout $\psi = \sum_i \alpha_i \kappa(\cdot, \boldsymbol{x}_i)$ et $\phi = \sum_j \beta_j \kappa(\cdot, \boldsymbol{x}_j)$ de $\mathcal{H}$ par $\langle \psi, \phi \rangle_{\mathcal{H}} = \sum_{i,j} \alpha_i \beta_j \kappa(\boldsymbol{x}_j, \boldsymbol{x}_i)$

On s'intéresse aux fonctions $\kappa(\boldsymbol{x}, \boldsymbol{x}')$ pouvant faire fonction de produit scalaire dans un espace $\mathcal{H}$. On appelle *noyau* une fonction symétrique κ de $\mathcal{X} \times \mathcal{X}$ dans $\mathbb{R}$.

Theorem (Mercer)

Si κ est un noyau continu d'un opérateur intégral défini positif, ce qui signifie que

$$\iint \varphi(\boldsymbol{x})\, \kappa(\boldsymbol{x}, \boldsymbol{x}')\, \varphi^*(\boldsymbol{x}')\, d\boldsymbol{x}\, d\boldsymbol{x}' \geq 0$$

pour tout $\varphi \in \mathcal{L}^2(\mathcal{X})$, il peut être décomposé sous la forme

$$\kappa(\boldsymbol{x}, \boldsymbol{x}') = \sum_{i=1}^{\infty} \lambda_i\, \psi_i(\boldsymbol{x})\, \psi_i(\boldsymbol{x}'),$$

où ψ_i et λ_i sont les fonctions propres (orthogonales) et valeurs propres (positives) du noyau κ, respectivement, telles que

$$\int \kappa(\boldsymbol{x}, \boldsymbol{x}')\, \psi_i(\boldsymbol{x})\, d\boldsymbol{x} = \lambda_i\, \psi_i(\boldsymbol{x}').$$

Il est aisé de voir qu'un noyau κ satisfaisant au théorème de Mercer peut faire fonction de produit scalaire dans un espace transformé $\mathcal{H}$. Il suffit d'écrire :

$$\boldsymbol{\phi}(\boldsymbol{x}) = \begin{pmatrix} \sqrt{\lambda_1}\,\psi_1(\boldsymbol{x}) \\ \sqrt{\lambda_2}\,\psi_2(\boldsymbol{x}) \\ \cdots \end{pmatrix}$$

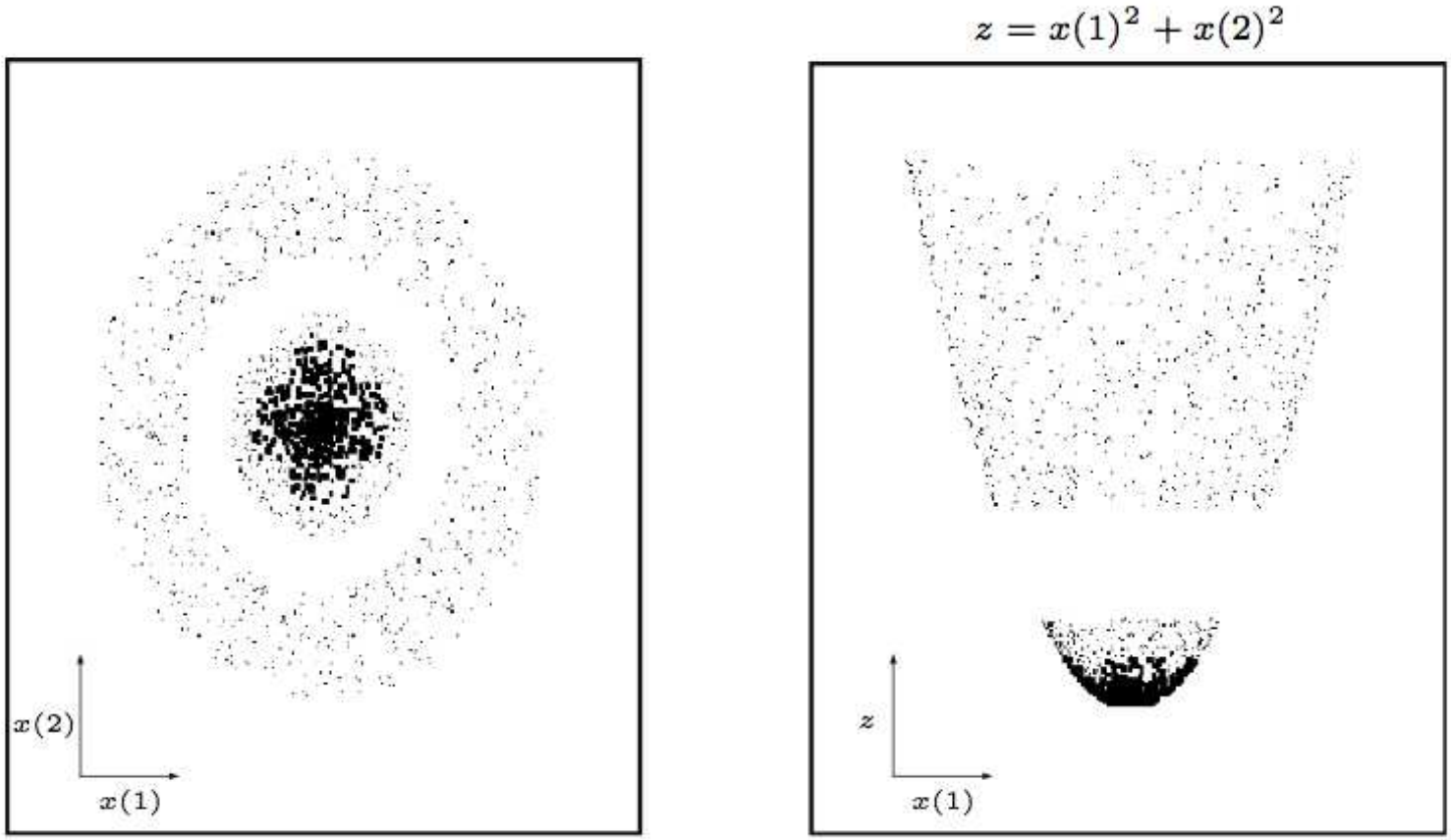
Dans ces conditions, on vérifie bien que l'on retrouve : $\langle \boldsymbol{\phi}(\boldsymbol{x}), \boldsymbol{\phi}(\boldsymbol{x}') \rangle = \kappa(\boldsymbol{x}, \boldsymbol{x}')$.

On définit l'espace $\mathcal{H}$ comme étant celui engendré par les fonctions propres ψ_i du noyau κ, c'est-à-dire

$$\mathcal{H} = \{ f(\cdot) \mid f(x) = \sum_{i=1}^{\infty} \alpha_i \; \psi_i(x), \; \alpha_i \in \mathbb{R} \}.$$

Propriété

$\phi(x)$ est souvent de grande dimension, parfois infinie.



La transformation polynomiale rend les données linéairement séparables.

Un classifieur linéaire en $\phi(x)$ est non-linéaire par rapport à x

Propriété

On n'a jamais besoin de calculer explicitement $\phi(\boldsymbol{x})$

Dans le cas de la transformation polynomiale de degré 2, on montre aisément que :

$$\langle \phi(\boldsymbol{x}), \phi(\boldsymbol{x}') \rangle = (1 + \langle \boldsymbol{x}, \boldsymbol{x}' \rangle)^2 \triangleq \kappa(\boldsymbol{x}, \boldsymbol{x}')$$

▷ **Le calcul de produit scalaire peut s'effectuer dans $\mathbb{R}^2$!**

Plus généralement, on s'intéresse à $\kappa(\boldsymbol{x}, \boldsymbol{x}') = (1 + \langle \phi(\boldsymbol{x}), \phi(\boldsymbol{x}') \rangle)^q$, avec $\boldsymbol{x} \in \mathbb{R}^l$.

$$\kappa(\boldsymbol{x}, \boldsymbol{x}') = (1 + \langle \boldsymbol{x}, \boldsymbol{x}' \rangle)^q = \sum_{j=0}^{q} \binom{q}{j} \langle \boldsymbol{x}, \boldsymbol{x}' \rangle^j .$$

Chaque composante $\langle \boldsymbol{x}, \boldsymbol{x}' \rangle^j = [x(1)\, x'(1) + \ldots + x(l)\, x'(l)]^j$ de cette expression peut être développée en une somme pondérée de monômes de degré j de la forme

$$[x(1)\, x'(1)]^{j_1} [x(2)\, x'(2)]^{j_2} \ \ldots \ [x(l)\, x'(l)]^{j_l}$$

avec $\sum_{i=1}^{l} j_i = j$. Ceci mène directement à l'expression de $\phi(\boldsymbol{x})$...

On peut montrer que les noyaux suivants vérifie la condition de Mercer, et correspondent donc à un produit scalaire dans un espace $\mathcal{H}$.

Noyaux projectifs	
monomial de degré q	$\langle \boldsymbol{x}, \boldsymbol{x}' \rangle^q$
polynomial de degré q	$(1 + \langle \boldsymbol{x}, \boldsymbol{x}' \rangle)^q$
sigmoidal	$\frac{1}{\eta_0} \tanh(\beta_0 \langle \boldsymbol{x}, \boldsymbol{x}' \rangle - \alpha_0)$

Noyaux radiaux	
Gaussien	$\exp(-\frac{1}{2\sigma_0^2} \|\boldsymbol{x} - \boldsymbol{x}'\|^2)$
exponentiel	$\exp(-\frac{1}{2\sigma_0^2} \|\boldsymbol{x} - \boldsymbol{x}'\|)$
uniforme	$\frac{1}{\eta_0} \mathbb{1}_{\|\boldsymbol{x}-\boldsymbol{x}'\| \leq \beta_0}$
Epanechnikov	$\frac{1}{\eta_0} (\beta_0^2 - \|\boldsymbol{x} - \boldsymbol{x}'\|^2) \mathbb{1}_{\|\boldsymbol{x}-\boldsymbol{x}'\| \leq \beta_0}$
Cauchy	$\frac{1}{\eta_0} \frac{1}{1+\|\boldsymbol{x}-\boldsymbol{x}'\|^2/\beta_0^2}$

... et encore $\kappa_1(\boldsymbol{x}, \boldsymbol{x}') + \kappa_2(\boldsymbol{x}, \boldsymbol{x}')$, $\kappa_1(\boldsymbol{x}, \boldsymbol{x}') \cdot \kappa_2(\boldsymbol{x}, \boldsymbol{x}')$, ...

Theorem (Théorème de Représentation)

Toute fonction ψ d'un espace de Hilbert à noyau reproduisant $\mathcal{H}$, de noyau κ, qui minimise le risque empirique régularisé

$$\frac{1}{n}\sum_{k=1}^{n} Q(\psi(\boldsymbol{x}_k), y_k) + \eta\ g(\|\psi\|^2_{\mathcal{H}}),$$

impliquant n sorties $\psi(\boldsymbol{x}_k)$ obtenues pour des entrées $\boldsymbol{x}_k$, et (éventuellement) n sorties désirées y_k, avec g une fonction monotone croissante sur $\mathbb{R}_+$, peut s'écrire sous la forme

$$\psi(\cdot) = \sum_{i=1}^{n} \alpha_i \kappa(\cdot, \boldsymbol{x}_i).$$

Eléments de preuve :
Toute fonction ψ de $\mathcal{H}$ se décompose selon $\psi = \sum_{i=1}^{n} \alpha_i\, \kappa(\cdot, \boldsymbol{x}_i) + \psi^{\perp}$, avec $\langle \psi^{\perp}, \kappa(\cdot, \boldsymbol{x}_i)\rangle_{\mathcal{H}} = 0$ pour tout $i = 1, \ldots, n$. Puisque $\psi(x_j) = \langle \psi, \kappa(\cdot, \boldsymbol{x}_j)\rangle$, la valeur de $\psi(x_j)$ n'est donc pas affectée par $\psi^{\perp}$, pour $j = 1, \ldots, n$.

Consequences :
La minimisation sur un espace fonctionnel Hilbertien (parfois de dimension infinie) aboutit à une minimisation sur $\mathbb{R}^n$