# Apprentissage: cours 10
# Estimateur de Stein, régularisation et pénalisation

Simon Lacoste-Julien

28 novembre 2014

**Résumé**

Sous la perspective du risque fréquentiste, nous donnons de nouvelles perspectives sur la régularisation pour la perte quadratique. Le paradoxe de Stein motive le "shrinkage" (une sorte de régularisation). Nous analysons ensuite le risque fréquentiste pour des estimateurs linéaires, qui donnent lieu à la pénalité de Mallows qui peut nous aider à faire de la sélection de modèle en régularisant le risque empirique.

Dans le cours 6, nous avons vu une motivation pour la régularisation avec la perspective PAC. Aujourd'hui, nous analysons les estimateurs avec le risque fréquentiste $\mathbb{E}[\mathcal{R}\left(\hat{f}\right)]$. Nous allons seulement considérer la perte quadratique. Il est donc utile de répéter la décomposition biais-variance pour le risque fréquentiste sous la perte quadratique telle que mentionné dans le premier cours.

**Décomposition biais-variance** du risque fréquentiste pour la perte quadratique :

$$\mathbb{E}\big[\mathcal{R}(\widehat{f})\big] = \mathbb{E}[(\widehat{f}(X) - Y)^2] = \underbrace{\mathbb{E}[(\widehat{f}(X) - \mathbb{E}[\widehat{f}(X)|X])^2]}_{\text{variance de } \widehat{f}} + \underbrace{\mathbb{E}[(\mathbb{E}[\widehat{f}(X)|X] - \overbrace{\mathbb{E}[Y|X]}^{f^\star(X)})^2]}_{\text{biais de } \widehat{f}} + \underbrace{\mathbb{E}[(Y - \overbrace{\mathbb{E}[Y|X]}^{f^\star(X)})^2]}_{\text{variance du "bruit"}}$$

Le dernier terme est le risque de l'estimateur optimal, et donc les deux premiers termes (la variance et le biais) forment l'espérance de l'excès de risque pour l'estimateur $\widehat{f}$.

## 1 Estimateur de James et Stein

Soit $X = (X^{(1)}, \dots, X^{(p)})^\top$ un vecteur Gaussien d'espérance $\boldsymbol{\mu} \in \mathcal{R}^p$ et de matrice de covariance $\sigma^2\mathbf{I}$. L'estimateur du maximum de vraisemblance est $\bar{X} \sim \mathcal{N}(\boldsymbol{\mu}, \frac{1}{n}\sigma^2\mathbf{I})$.

On s'intéresse à la possibilité de produire un estimateur dont le risque quadratique est uniformément meilleur que celui de $\bar{X}$.

Comme $\bar{X} \sim \mathcal{N}(\boldsymbol{\mu}, \bar{\sigma}^2\mathbf{I})$ avec $\bar{\sigma}^2 = \sigma^2/n$ on peut se restreindre au cas de l'échantillon de taille 1.

**Définition 1** (Estimateur de James et Stein)**.** L'estimateur de James et Stein pour l'espérance $\boldsymbol{\mu}$ d'une variable aléatoire $X$ est défini par :

$$\hat{\boldsymbol{\mu}}^{(\mathrm{JS})} = \Big(1 - \frac{\sigma^2(p-2)}{\|X\|^2}\Big)\, X$$

**Théorème 1.** *Pour $p \geqslant 3$, l'excès de risque de l'estimateur de James-Stein est*

$$\mathbb{E}\big[\mathcal{R}\left(\hat{\boldsymbol{\mu}}^{(JS)}\right) - \mathcal{R}\left(\boldsymbol{\mu}\right)\big] = \mathbb{E}[(\hat{\boldsymbol{\mu}}^{(JS)} - \boldsymbol{\mu})^2] = p\sigma^2 - (p-2)^2\sigma^4\mathbb{E}\big[\|X\|^{-2}\big] < p\sigma^2 = \mathbb{E}[(X - \boldsymbol{\mu})^2]$$

*Exercice* 1. Montrer à partir du résultat du théorème que :

$$\mathbb{E}[(\hat{\boldsymbol{\mu}}^{(\mathrm{JS})} - \boldsymbol{\mu})^2] \leqslant 4\sigma^2 + \frac{p\sigma^2\|\boldsymbol{\mu}\|^2}{p\sigma^2 + \|\boldsymbol{\mu}\|^2} \leqslant 4\sigma^2 + \sigma^2 \min\Big(p, \frac{\|\boldsymbol{\mu}\|^2}{\sigma^2}\Big)$$

On peut en fait avec une analyse plus fine montrer que

$$\mathbb{E}[(\hat{\boldsymbol{\mu}}^{(\mathrm{JS})} - \boldsymbol{\mu})^2] \leqslant 2\sigma^2 + \frac{(p-2)\sigma^2\|\boldsymbol{\mu}\|^2}{(p-2)\sigma^2 + \|\boldsymbol{\mu}\|^2} \leqslant \sigma^2 \min(p, 2 + \frac{\|\boldsymbol{\mu}\|^2}{\sigma^2})$$

**Paradoxe de Stein**

Supposons qu'on cherche à estimer
- la vitesse de la lumière $\mu_1$,
- la consommation de thé annuelle en Chine $\mu_2$,
- la quantité de précipitation annuelle moyenne à Paris $\mu_3$.

L'estimateur de Stein permet d'obtenir des estimateurs $\hat{\mu}_1^{(\mathrm{JS})}$, $\hat{\mu}_2^{(\mathrm{JS})}$ et $\hat{\mu}_3^{(\mathrm{JS})}$ de ces trois quantités telles que l'erreur quadratique globale $\sum_{j=1}^3 \mathbb{E}[(\hat{\mu}_j^{(\mathrm{JS})} - \mu_j)^2]$ soit *strictement plus faible* que celle de la moyenne empirique !

L'estimateur de James-Stein obtient un meilleur risque que l'estimateur de maximum de vraisemblance en augmentant le biais (le biais est 0 pour l'estimateur de MV), pour plus diminuer la variance grâce au "shrinkage". L'estimateur de James-Stein peut s'interpréter comme un estimateur MAP avec un à-priori bayésien $\boldsymbol{\mu} \sim \mathcal{N}(\mathbf{0}, \hat{a}^2\mathbf{I})$ où l'hyperparamètre $\hat{a}^2 := \frac{\|X\|^2}{p-2} - \sigma^2$ dépend des données (méthodologie bayésienne empirique). Cette interprétation n'est valide que pour $\frac{\|X\|^2}{p-2} > \sigma^2$. À noter que le shrinkage peut se faire vers n'importe quelle valeur de référence fixe $\boldsymbol{\mu}_0$, pas seulement le zéro, et il y aura quand-même meilleure risque que l'estimateur de maximum de vraisemblance.

# 2 Régression avec design fixe

En apprentissage supervisé, nous avons défini jusqu'ici le risque d'un prédicteur comme l'espérance sur les valeurs possible d'un nouveau couple entrée-sortie :

$$\mathcal{R}(f) = \mathbb{E}[\ell(f(X), Y)].$$

Nous allons maintenant modifier légèrement ce cadre et considérer le cadre du design fixe pour pouvoir facilemen calculer les quantités dans le compromis biais-variance du risque fréquentiste.

Dans ce cadre, on considère un ensemble d'entraînement $D_n = \{(x_1, y_1, \ldots, x_n, y_n)\}$, et on souhaite faire en sorte que la prédiction soit la plus exacte possible pour les valeurs d'entrées $x_i$ considérées, mais pour de nouvelles valeurs $y_i'$ tirée de la même loi conditionnelle que les $y_i$, c'est-à-dire selon $\mathbb{P}(Y = \cdot | X = x_i)$.

**Définition 2.** Le *risque à design fixe* est défini comme

$$\mathcal{R}_n^{DF}(f) := \mathbb{E}_n\big[\,\mathbb{E}\big[\ell(f(x_i), Y) \mid X = x_i\big]\big] = \frac{1}{n}\sum_{i=1}^n \mathbb{E}\big[\ell(f(x_i), Y) \mid X = x_i\big]$$

De façon équivalente, étant donné notre définition du risque empirique $\widehat{\mathcal{R}}_n(f) = \mathbb{E}_n[\ell(f(X), Y)] = \frac{1}{n}\sum_{i=1}^n \ell(f(X_i), Y_i)$, le risque à design fixe est l'espérance du risque empirique lorsqu'on conditionne sur les données d'entrée :

$$\mathcal{R}_n^{DF}(f) = \mathbb{E}\big[\widehat{\mathcal{R}}_n \mid X_1, \ldots, X_n\big].$$

Une troisième manière d'envisager $\mathcal{R}_n^{DF}$ est que c'est le risque si l'on suppose que la distribution des données d'entrées est uniforme sur les valeurs déjà observées.

Comme on s'en doute, l'étude à design fixe est plus adaptée au débruitage qu'à la prédiction.

Dans le cadre à design fixe, comme on ne cherche pas à généraliser à de nouvelles valeurs en entrée (mais seulement à bien généraliser sur de nouvelles valeurs en sortie), le prédicteur est entièrement spécifié par l'ensemble des valeurs qu'il prend sur l'ensemble $x_1, \ldots, x_n$. Par conséquent, plutôt que de raisonner sur des prédicteurs qui sont des fonctions, étant donné un ensemble de données d'entrées $(x_1, \ldots, x_n)$, il suffit de

raisonner sur les vecteurs de valeurs $\mathbf{f} = (f(x_1), \dots, f(x_n))^\top$. On notera $\mathbf{y} = (y_1, \dots, y_n)^\top$ le vecteur dont les entrées sont les valeurs de sortie de l'ensemble d'entraînement.

De même que dans le cadre initial, on peut définir dans ce cadre une fonction cible $f^* = \operatorname{argmin}_f \mathcal{R}_n^{DF}(f)$ (lorsque ce minimum existe) et l'excès de risque d'un prédicteur $f$ :

$$\mathcal{E}_n^{DF}(f) = \mathcal{R}_n^{DF}(f) - \mathcal{R}_n^{DF}(f^*).$$

## 2.1 Estimation linéaire

**Définition 3** (Estimateur par transformation linéaire)**.** Un estimateur par transformation linéaire (ETL) est un estimateur dont les valeurs sont obtenues en appliquant une transformation linéaire (dépendant en général des données d'entrée) aux données de sortie.

$$\hat{\mathbf{f}} = \mathbf{A}\mathbf{y} \quad \text{pour} \quad \mathbf{A} \in \mathcal{R}^{n \times n}.$$

**Définition 4** (Estimateur par projection)**.** Un estimateur par projection est un estimateur par transformation linéaire pour lequel $\mathbf{A}$ est une matrice de projection (on rappelle qu'une matrice $\mathbf{A}$ est une projection si et seulement si $\mathbf{A}^2 = \mathbf{A}$).

Exemples d'ETL :
- les estimateurs par moyennage local (histogrammes, Nadaraya-Watson, $k$-p.p.v.),
- l'estimateur de la régression linéaire ordinaire, de la régression ridge et de la régression dans un RKHS.

Exemples d'estimateurs par projection :
- estimateurs par histogrammes,
- estimateur de la régression linéaire ordinaire.

## 2.2 Pénalité $C_L$ de Mallows

On s'intéresse aux estimateurs par transformation linéaire pour la perte des moindres carrés. Pour un prédicteur $\mathbf{f}$ fixe, le risque empirique et le risque (au sense de Vapnik) à design fixe sont respectivement [1] :

$$\widehat{\mathcal{R}}_n(\mathbf{f}) = \frac{1}{n}\|\mathbf{y} - \mathbf{f}\|_2^2 \quad \text{et} \quad \mathcal{R}_n^{DF} = \frac{1}{n}\mathbb{E}\big[\|\mathbf{y} - \mathbf{f}\|_2^2\big] = \frac{1}{n}\|\mathbf{f} - \mathbf{f}^*\|_2^2 + \frac{1}{n}\mathbb{E}\big[\|\mathbf{y} - \mathbf{f}^*\|_2^2\big]$$

(Rappelez-vous que $\mathbf{f}^* = \mathbb{E}[\mathbf{y}]$.) Le dernier terme est la variance du bruit, c'est le risque du prédicteur optimal pour la perte quadratique sous design fixe.

Supposons que $\mathbf{y} = \mathbf{f}^* + \boldsymbol{\varepsilon}$ avec $\varepsilon_1, \dots, \varepsilon_n$ i.i.d. $\mathbb{E}[\varepsilon_i] = 0$ et $\mathbb{E}[\varepsilon_i^2] = \sigma^2$.

Soit $\hat{\mathbf{f}} = \mathbf{A}\mathbf{y}$ un estimateur par transformation linéaire. Si l'on suppose que l'on connaît la variance du bruit $\sigma^2$, nous allons voir si nous pouvons utiliser cette information pour "corriger" le risque empirique en le comparant au risque fréquentiste.

$$n\,\mathbb{E}[\widehat{\mathcal{R}}_n(\hat{\mathbf{f}})] = \mathbb{E}\big[\|\mathbf{y} - \hat{\mathbf{f}}\|_2^2\big] = \mathbb{E}\big[\|(\mathbf{I} - \mathbf{A})\mathbf{y}\|_2^2\big] = \|(\mathbf{I} - \mathbf{A})\mathbf{f}^*\|_2^2 + \sigma^2\|\mathbf{I} - \mathbf{A}\|_F^2$$

et comme $\mathcal{E}_n^{DF}(\hat{\mathbf{f}}) = \frac{1}{n}\|\hat{\mathbf{f}} - \mathbf{f}^*\|_2^2$, on a :

$$n\mathbb{E}[\mathcal{E}_n^{DF}] = \mathbb{E}[\|\hat{\mathbf{f}} - \mathbf{f}^*\|_2^2] = \mathbb{E}[\|\mathbf{A}\,\mathbf{y} - \mathbf{f}^*\|_2^2] = \mathbb{E}[\|(\mathbf{A} - \mathbf{I})\mathbf{f}^* + \mathbf{A}\boldsymbol{\varepsilon}\|_2^2] = \underbrace{\|(\mathbf{A} - \mathbf{I})\mathbf{f}^*\|_2^2}_{\text{biais}} + \underbrace{\sigma^2\,\|\mathbf{A}\|_F^2}_{\text{variance}}$$

Nous avons encore pour le risque fréquentiste :

$$\mathbb{E}[\mathcal{R}_n^{DF}(\hat{\mathbf{f}})] = \mathbb{E}[\mathcal{E}_n^{DF}] + \frac{\sigma^2}{n}\|\mathbf{I}\|_F^2$$

1. À noter que la perte quadratique est parfois $\ell(a, y) = (a - y)^2$ et parfois $\ell(a, y) = \frac{1}{2}(a - y)^2$ dépendant de la convention ; ici on le prend sans le facteur 1/2. Il faut faire attention à la convention utilisée quand on fait la régression ridge (les différents facteurs font un re-scaling de $\lambda$).

En comparant les expressions

$$\mathbb{E}[\widehat{\mathcal{R}}_n(\hat{\mathbf{f}})] = \frac{1}{n}\|(\mathbf{I}-\mathbf{A})\mathbf{f}^*\|_2^2 + \frac{\sigma^2}{n}\|\mathbf{I}-\mathbf{A}\|_F^2$$

et

$$\mathbb{E}[\mathcal{R}_n^{DF}(\hat{\mathbf{f}})] = \frac{1}{n}\|(\mathbf{I}-\mathbf{A})\mathbf{f}^*\|_2^2 + \frac{\sigma^2}{n}(\|\mathbf{I}\|_F^2 + \|\mathbf{A}\|_F^2)$$

on obtient

$$\boxed{\mathbb{E}[\mathcal{R}_n^{DF}(\hat{\mathbf{f}})] = \mathbb{E}[\widehat{\mathcal{R}}_n(\hat{\mathbf{f}})] + 2\frac{\sigma^2}{n}\operatorname{tr}(\mathbf{A})}$$

Si $\operatorname{tr}(\mathbf{A}) > 0$, ce qui typiquement le cas en pratique, cela montre que le risque empirique sous-estime le risque fréquentiset de l'estimateur. On voit également que le biais est estimé correctement et que c'est la variance de l'estimateur qui n'est pas juste en espérance. L'estimation de la variance est biaisée.

**Définition 5** (Degrés de liberté, pénalité $C_L$ de Mallows)**.**

On appelle *nombre de degrés de liberté* de l'estimateur linéaire (ETL) basé sur $\mathbf{A}$ la quantité $\operatorname{tr}(\mathbf{A})$.

On appelle *pénalité $C_L$ de Mallows* le terme $2\frac{\sigma^2}{n}\operatorname{tr}(\mathbf{A})$.

Le résulat ci-dessous montre que la quantité obtenue en pénalisant le risque empirique avec le $C_L$ de Mallows, soit

$$\widehat{\mathcal{R}}_n(\hat{\mathbf{f}}) + 2\frac{\sigma^2}{n}\operatorname{tr}(\mathbf{A}),$$

est un estimateur sans biais du risque à design fixe. À noter que $\mathbf{A}$ ne dépend que des données $X_i$, pas les sorties $y_i$.

## 2.3 Pénalité $C_p$ de Mallows

La pénalité $C_p$ de Mallows est un cas particulier de la pénalité $C_L$ lorsque, $\hat{\mathbf{f}}$ est un estimateur par projection.

Dans ce cas $\mathbf{A}$ est un projecteur et $\operatorname{tr}(\mathbf{A})$ est donc la dimension du sous-espace sur lequel on projette (car les valeurs propres d'une matrice de projection sont seulement 0 ou 1). Soit $p = \operatorname{tr}(\mathbf{A})$ cette dimension, le $C_p$ de Mallows est la quantité $2p\frac{\sigma^2}{n}$. La pénalisation de Mallows est alors

$$\widehat{\mathcal{R}}_n(\hat{\mathbf{f}}) + 2p\frac{\sigma^2}{n}$$

**Sélection de modèle.** Cette pénalisation pourrait être utilisé pour faire de la sélection de modèle en considérant différents nombre de degrés de liberté $p$. Cette pénalité peut être interpréter comme de la régularisation pour empêcher le surapprentissage quand on augmente $p$. Dans le TD, vous aller relier la pénalité avec la régularisation ridge.

**TD** Ces exercices seront couverts durant le prochain TD.

*Exercice* 2. Montrer que le prédicteur de la régression linéaire ordinaire est un estimateur par projection. Quelle est la dimension $p$ correspondante ?

*Exercice* 3. Supposer que les données sont générées selon $y_i = x_i^\top \beta + \varepsilon_i$. Montrer que le prédicteur de la régression ridge est un ETL. Montrer qu'à design fixe et sous les hypothèses de bruit i.i.d. , on peut exprimer la variance du prédicteur seulement en fonction de $\sigma^2$, des valeurs singulières de la matrice de design $\lambda_i$ et du paramètre de régularisation $\lambda$. Montrer que s'il on note $\tilde{\beta}$ la représentation de $\beta$ dans une base orthonormale bien choisie, le biais de la régression ridge s'exprime en fonction de $\tilde{\beta}$, $\sigma^2$, $\lambda$ et les valeurs singulières de $\mathbf{A}$.

*Exercice* 4. Comment utiliser le $C_L$ de Mallows pour choisir $\lambda$ pour la régression ridge si l'on connaît $\sigma^2$ ?