Debriefing du TP n°1.

1/ Soit η la variable aléatoire de Bernouilli qui vaut 1 quand la discrimination est réussie et 0 sinon. Soit τ_{gen} la valeur moyenne de η obtenue en calculant la moyenne statistique sur différentes réalisations des vecteurs de la base de généralisation. La moyenne empirique τ_g (si besoin, revoir sa définition dans l'énoncé du TP) est un estimateur de τ_{gen}.

2/ Il est possible d'estimer la barre d'erreur de l'estimation τ_g avec $\widehat{\sigma}_{\tau_g} = \sqrt{\frac{\tau_g(1-\tau_g)}{P_{gen}}}$, où P_{gen} est la taille de la base de généralisation.

3/ Sur la courbe $P_{app} \mapsto \tau_g^{(PI)}$, on observe avec un minimum local en $P_{app} = N$. On explique ci-dessous l'origine de ce phénomène, mais avant, on a besoin de quelque rappels mathématiques.
Comme expliqué en cours, pour PI le frontière $\boldsymbol{w}$ est donnée par

$$\boldsymbol{w} = \left(\boldsymbol{X}\,\boldsymbol{X}^T\right)^{-1}\boldsymbol{X}\,\boldsymbol{d}$$

où $\left(\boldsymbol{X}\,\boldsymbol{X}^T\right)^{-1}$ est remplacée par $\boldsymbol{U}\,\boldsymbol{\Delta}_{1/\lambda_{1:r}}\boldsymbol{U}^T$ quand $\boldsymbol{X}\,\boldsymbol{X}^T$ n'est pas inversible. On rappelle que

$$\boldsymbol{X}\,\boldsymbol{X}^T = \boldsymbol{U}\,\boldsymbol{\Delta}_{\lambda_{1:r}}\boldsymbol{U}^T$$

est une décomposition en valeurs propres. Ainsi $\boldsymbol{U}\,\boldsymbol{U}^T = \boldsymbol{Id}$ et $\boldsymbol{\Delta}_{\lambda_{1:r}}$ est une matrice diagonale avec sur les r premiers termes de la diagonale $\lambda_1 \geq \lambda_2 \geq ... \geq \lambda_r > 0$.

Si la matrice est inversible, que peut-on dire de r ?

Remarque. Si, quand la matrice $\boldsymbol{X}\,\boldsymbol{X}^T$ est singulière, on complète la suite de $(\lambda_n)_{n=1,...,r}$ avec des zéros, alors on peut aussi s'écrire

$$\boldsymbol{X}\,\boldsymbol{X}^T = \sum_{n=1}^{N} \lambda_n \boldsymbol{u}_n \boldsymbol{u}_n^T \tag{42}$$

où $(\boldsymbol{u}_n)_{n=1,..,N}$ sont des vecteurs propres qui forment une base orthonormée.

Ainsi, quand $\left(\boldsymbol{X}\,\boldsymbol{X}^T\right)^{-1}$ est remplacée par $\boldsymbol{U}\,\boldsymbol{\Delta}_{1/\lambda_{1:r}}\boldsymbol{U}^T$, tout se passe comme si on n'inverse *uniquement* les valeurs propres strictement positives.

A présent, expliquons le phénomène observé. Par définition $\boldsymbol{X} = \left(\boldsymbol{x}^{(1)}, \boldsymbol{x}^{(2)}, ..., \boldsymbol{x}^{(P_{app})}\right)$ donc

$$\boldsymbol{X}\,\boldsymbol{X}^T = \sum_{p=1}^{P_{app}} \boldsymbol{x}^{(p)} \left(\boldsymbol{x}^{(p)}\right)^T$$

que l'on peut comparer à l'équation (42).

A priori, les P_{app} vecteurs $\boldsymbol{x}^{(p)}$ forment une famille libre, donc

- si $P_{app} >> N$, alors $\lambda_N > 0$. Dans ce cas, la matrice $\boldsymbol{X}\,\boldsymbol{X}^T$ est inversible. RAS.
- si $P_{app} < N$, alors $r = P_{app}$. La matrice $\boldsymbol{X}\boldsymbol{X}^T$ n'est pas inversible et $\left(\boldsymbol{X}\,\boldsymbol{X}^T\right)^{-1}$ est remplacée par $\boldsymbol{U}\,\boldsymbol{\Delta}_{1/\lambda_{1:r}}\boldsymbol{U}^T$. RAS.
- si $P_{app} = N$, alors $r = N = P_{app}$ car, a priori, la famille des $\boldsymbol{x}$ de la base d'apprentissage est libre. Néanmoins, quand N est grand et qu'on tire de manière aléatoire une famille de taille N, on a une probabilité forte que l'un des vecteurs soit proche d'une combinaison linéaire des autres. Dans ce cas, $\lambda_N \mapsto 0$, sans l'atteindre. Par conséquent, $\lambda_N^{-1} \mapsto +\infty$. Ceci est un *gros* problème, car on obtient alors une estimation *très bruitée* de l'inverse de $\boldsymbol{X}\,\boldsymbol{X}^T$. C'est pour cette raison que l'on a une chute de performance en $P_{app} = N$.

Une solution possible à ce problème est de *régulariser* l'inversion de la matrice $\boldsymbol{X}\,\boldsymbol{X}^T$.

Mathématiquement, cela peut revenir à inverser les valeurs propres de $\boldsymbol{X}\,\boldsymbol{X}^T$ uniquement si celle-ci sont plus grandes qu'un certain seuil, par exemple $\lambda_1/100$.

4/ Avec la ridge approximation (RA), on s'intéresse à

$$\boldsymbol{w} = \left(\boldsymbol{X}\,\boldsymbol{X}^T + P_{app}\sigma^2\boldsymbol{Id}\right)^{-1}\boldsymbol{X}\,\boldsymbol{d}$$

et donc à l'inverse de $\boldsymbol{X}\,\boldsymbol{X}^T + P\sigma^2\boldsymbol{I}_d$. On a

$$\begin{array}{l} w_{RA} \mapsto w_{PI} \quad \text{quand } \sigma \mapsto 0 \\ w_{RA} \mapsto w_{Hebb} \quad \text{quand } \sigma \mapsto \infty \end{array}$$

Plus intéressant, on a $\boldsymbol{X}\,\boldsymbol{X}^T = \boldsymbol{U}\Delta_{\lambda_{1:N}}\boldsymbol{U}^T$ et aussi

$$\boldsymbol{X}\,\boldsymbol{X}^T + P_{app}\sigma^2\boldsymbol{I}_d = \boldsymbol{U}\Delta_{\lambda_{1:N}+P_{app}\sigma^2}\boldsymbol{U}^T$$

Que peut-on dire de la plus petite valeur propre de $\left(\boldsymbol{X}\,\boldsymbol{X}^T + P_{app}\sigma^2\boldsymbol{I}_d\right)$?

Ainsi fixer la valeur de σ permet de régulariser l'inversion de la matrice.

Il reste néanmoins à trouver la "bonne" valeur du paramètre σ. Pour cela, on ne doit pas utiliser la base de généralisation, car la base de généralisation est réservée à l'analyse des performances. En revanche, empiriquement, il semble que fixer σ quand le taux d'apprentissage commence à baisser semble une solution intéressante.

Synthèse du TP n°1 : on a comparé une approche empirique (Hebb) avec une approche mathématique (PI). Finalement, la méthode mathématique s'avère intéressante, mais cela a nécessité de faire attention à l'inversion de matrice.