# M2-ISL PVP
# Application d'un mécanisme de LDP pour de l'apprentissage

Jean-François COUCHOT
couchot[arobase]femto-st[point]fr

5 décembre 2022

## 1 Estimation de fréquences à partir de données netoyées selon $\mathcal{M}_{GRR}$

**Exercice 1.1 (Estimateur : implantation et vérification).** *On considère le jeu de données adult de l'UCI déjà vu en TD et particulièrement l'attribut* `marital-status`.

1. *Afficher les fréquences des différentes valeurs de cet attribut.*
2. *Implanter le mécanisme $\mathcal{M}_{GRR}$ et l'estimateur $\hat{f}$.*
3. *Appliquer le mécanisme $\mathcal{M}_{GRR}$ à chaque réponse avec une valeur de $\epsilon$ fixée à 1.*
4. *Vérifier que l'estimateur $\hat{f}$ permet de retrouver des fréquences similaires à celles de la question 1.*

**Exercice 1.2 (Estimateur : statistiques).** *Toujours sur le même jeu de données adult de l'UCI et le même attribut* `marital-status`, *contruire des statistiques de dispersion de cet estimateur.*

## 2 Application du mécanisme $\mathcal{M}_{GRR}$ à du Machine Learning

L'objectif ici est d'évaluer l'application du mécanisme $\mathcal{M}_{GRR}$ comme un prétraitement pour du Machine learning, "Categorical Naive Bayes" par exemple.

On rappelle que $\mathcal{M}_{GRR}$ ajoute du bruit à des valeurs dans un ensemble. Dans ce qui suit, on considérera donc les attributs discrets ['workclass', 'education', 'occupation', 'relationship', 'race', 'sex','native-country'] et l'on essaiera de deviner la valeur de 'marital-status'.

**Exercice 2.1 (Sélection des attributs et valeur de baseline).**

1. *A partir du jeu de données complet, construire un dataset où* `X` *et* `y` *ne contiennent que les attributs discrets à utiliser pour l'apprentissage et la valeur à prédire.*
2. *Découper "honnêtement"* `X` *et* `y` *en* `X_train`, `X_test`, `y_train` *et* `y_test`.
3. *Evaluer l'algorithme d'apprentissage bayésien naïf catégoriel sur ce jeu de données. En déduire une valeur de baseline.*

**Exercice 2.2 (Evaluation d'un apprentissage de données bruitées par le mécanisme $\mathcal{M}_{GRR}$).**

1. *Montrer théoriquement que si l'on veut évaluer le mécanisme $\mathcal{M}_{GRR}$, il faut*
   (a) *appliquer le mécanisme $\mathcal{M}_{GRR}$ sur* `X_train`, `X_test` *et* `y_train` *et obtenir respectivement ;* `Xp_train`, `Xp_test` *et* `yp_train`
   (b) *faire l'apprentissage sur le couple (*`Xp_train`, `yp_train`*);*
   (c) *prédire les réponses correspondantes à* `Xp_test` *;*
   (d) *comparer celles-ci à* `y_test` *qui n'a pas été modifié.*
2. *Mettre en place cette démarche et afficher une valeur de précision.*

**Exercice 2.3 (Analyse statistiques d'un apprentissage de données bruitées par le mécanisme $\mathcal{M}_{GRR}$).**

1. *Répéter un grand nombre de fois la mesure de précision dont la démarche est détaillée à l'exercice précédent.*
2. *Construire des statistiques de moyenne de précision et de dispersion.*
3. *Etendre cette étude en faisant varier $\epsilon$ dans l'ensemble $\{10^{-2}, 5 * 10^{-2}, 10^{-1}, 0.5, 1, 5, 10\}$.*
4. *Illustrer ceci au moyen de courbes.*