Approximate full conformal prediction via influence functions in regression

Davidson Lova Razafindrakoto [1,2] & Alain Celisse [2] & Jérôme Lacaille [1]

[1] *Safran Aircraft Engines, France*
[2] *Laboratoire SAMM, Université Paris 1 Panthéon-Sorbonne, France*
Davidson-Lova.Razafindrakoto@edu.univ-paris1.fr
alain.celisse@univ-paris1.fr
jerome.lacaille@safrangroup.com

Résumé. La prédiction conforme est un cadre qui fournit implicitement des régions prédictives de confiance sur la prédiction fournie par tout estimateur. Dans ce contexte, un problème majeur est le calcul de cette région prédictive de confiance, habituellement définie de façon implicite. Le point de vue "full conformal" où toute les observations d'entraînement sont utilisées pour cette région prédictive de confiance est d'ailleurs souvent abandonné pour cette raison au profit du point de vue "split conformal". La principale contribution de ce travail repose sur l'utilisation de fonctions d'influence afin de bâtir des régions prédictives de confiance approchées dans le cadre "full conformal" qui soient calculables efficacement. Parmi d'autres avantages, notre approche permet d'obtenir des régions prédictives de confiance plus informatives que celles obtenues par l'approche concurrente "split conformal". L'un des objectifs du présent travail consiste à bâtir des régions prédictives de confiance (approchées) pour la prédiction obtenue au terme de T itérations d'un algorithme de type descente de gradient (GD). Les performances de notre approche seront illustrées au travers de modèles tels que : la régression linéaire, les réseaux de neurones de type MLP,...

Mots-clés. Apprentissage statistique, Réseaux de neurones, Régression, Quantification d'incertitudes, Prédiction conforme, Fonction d'influence

Abstract. Conformal prediction is the framework implicitly provides confidence predictive regions for the prediction of any estimator. In this context, a major issue is that the explicit computation of that predictive confidence region, is usually defined implicitly. The "full conformal" procedure where all training observations are used in building the confidence predictive region is often dismissed in favor of the "split conformal" procedure. The main contribution of this work lays on the use of influence function in order to build approximate confidence predictive regions in the "full conformal" framework which are efficiently calculable. Among other advantages, our approach enables to have, more informative confidence predictive regions than the competing approach "split conformal". One the of the aims of the present work consists at building (approximate) confidence predictive regions for the prediction given at the end of T iterations of a gradient descent algorithm (GD). The performances of our approach will be illustrated through models such a: linear regression, neural networks (MLP),...

Keywords. Statistical learning, Neural networks, Regression, Conformal prediction, Influence function

1 État de l'art

La prédiction conforme est un cadre qui fournit implicitement des régions prédictives de confiance sur la prédiction fournie par tout prédicteur. Dans le cadre dit "full conformal", la région prédictive de confiance selon Vovk et al. (2005) est un sous-ensemble de l'ensemble des prédictions possibles en lesquelles la "p-valeur conforme" dépasse un seuil de confiance spécifié. Ces régions prédictives conformes dépendent de nombreux calculs de "p-valeurs conformes". Le nombre de calculs à effectuer est le cardinal de l'espace des prédictions possibles (Section 2.2.4 dans Vovk et al. (2022)).

Le calcul de cette "p-valeur conforme" requiert potentiellement plusieurs entraînements de prédicteurs, ce qui est coûteux sur le plan calculatoire. Cette difficulté est exacerbée par le nombre de calculs à effectuer. En particulier, c'est la raison pour laquelle l'approche "full conformal" est abandonnée au profit d'autres approches alternatives, notamment "split conformal" (Papadopoulos, 2008), "cross conformal", "Jackknife+" (Vovk, 2015),... Dans l'approche "split conformal", un seul entraînement est requis pour le calcul les "p-valeurs conformes" tandis que pour l'approche "cross conformal", le nombre d'entraînements peut aller d'un seul jusqu'au nombre total d'observations du jeu de données d'entraînement. Cependant, ces approximations motivées par les aspects calculatoires peuvent induire une perte d'informativité des régions fournies (Chapitre 4 de Vovk et al. (2022)).

Pour éviter cette perte, Nouretdinov et al. (2001) ont développé une méthode qui permet pour calculer efficacement les régions "full conformal" dans le cadre de la régression ridge. Récemment Ndiaye and Takeuchi (2019); Lei (2019) ont étendu ces méthodes pour d'autres variantes de la régression linéaire telles que Elastic Net et LASSO. Ndiaye (2022) utilise les propriétés de stabilité de l'entraînement pour approcher les régions "full conformal". Ndiaye and Takeuchi (2023) exploite la forme des régions prédictives de confiance et des algorithmes de recherche de racines pour les estimer. En classification, Martinez et al. (2023) utilisent les fonctions d'influence pour quantifier l'apport d'une nouvelle variable y à prédire lorsqu'elle est incluse dans l'ensemble d'entraînement. Cette approximation est ensuite utilisée pour bâtir des régions prédictives de confiance approchées. Dans ce cadre, le prédicteur entraîné est celui fourni par la minimisation du risque empirique. Cependant plus généralement, le prédicteur entraîné est celui fourni au terme de T itérations d'un algorithme d'optimisation de type descente de gradient (GD).

Contributions. Le présent travail traite de l'approximation par les fonctions d'influence de régions prédictives de confiance en régression dans le cadre "full conformal". Deux situations sont considérées ici : (i) le prédicteur entraîné est celui qui minimise le risque empirique, (ii) le prédicteur entraîné est celui fourni à l'issue de T itérations de l'algorithme de descente de gradient. Ensuite pour chaque situation, les régions prédictives de confiance approchées sont explicitées dans le cas de la fonction de perte quadratique. Enfin, les approches développées sont évaluées au moyen de garanties théoriques de performances, de même qu'empiriquement par comparaison avec des approches concurrentes telles que l'oracle de Ndiaye and Takeuchi (2019), "split conformal" de Papadopoulos (2008) et "cross conformal" de Vovk (2015).

2 Prédiction conforme

2.1 Cadre "full conformal"

Soient $(X_1, Y_1), \ldots, (X_N, Y_N), (X_{N+1}, Y_{N+1})$ un ensemble d'observations *échangeables* (Section 2.1.5 de Vovk et al. (2022)). Pour un niveau de confiance α, la région prédictive de confiance "full conformal" $\hat{C}_\alpha(X_{N+1})$ pour Y_{N+1} a la garantie suivante (Vovk et al., 2005) :

$$\mathbb{P}(Y_{N+1} \in \hat{C}_\alpha(X_{N+1})) \geq 1 - \alpha.$$

Cette région $\hat{C}(X_{N+1})$ est définie à partir de l'entrée X_{N+1} et des observations d'entraînement $(X_1, Y_1), \ldots, (X_N, Y_N)$ de la manière suivante :

$$\hat{C}(X_{N+1}) = \{y \in \mathcal{Y} : \pi_{N+1}(X_{N+1}, y) > \alpha\},$$

où $y \in Y$ est la prédiction à tester, et la p-valeur conforme $\pi_{N+1}(X_{N+1}, y)$ évaluée en le couple (X_{N+1}, y) est donnée par

$$\pi_{N+1}(X_{N+1}, y) = \frac{1}{N+1}\text{Card}(\{i = 1, \ldots, N+1; S_i^y \geq S_{N+1}^y\}),$$

où S_i^y et S_{N+1}^y sont des scores de "non-conformité" définis ci-après (**Score**).

Posons $Z_i := (X_i, Y_i)$, $Z_{N+1}^y := (X_{N+1}, y)$ et

$$\mathcal{B}_{N+1}^{y;-i} = \wr Z_1, \ldots, Z_{i-1}, Z_{i+1}, \ldots, Z_N, Z_{N+1}^y \int \mathcal{B}_N = \wr Z_1, \ldots, Z_N \int .$$

Remarquons que $\pi_{N+1}(X_{N+1}, y)$ est bâtie à partir d'une collection de $N+1$ scores dits de *non-conformité* S_i^y. Pour $i = 1, \ldots, N$, chaque score S_i^y dépend de (X_{N+1}, y) et de toute les autres observations sauf (X_i, Y_i). Intuitivement, c'est une mesure de l'inadéquation entre la réponse cible Y_i et la prédiction fournie par le prédicteur M entraîné sur l'ensemble d'entraînement $\mathcal{B}_{N+1}^{y,-i}$

$$S_i^y = S(Y_i, M(X_i; \mathcal{B}_{N+1}^{y,-i})). \qquad \textbf{(Score)}$$

Le score S_{N+1}^y quant à lui dépend de (X_{N+1}, y) et de toutes les autres observations. C'est une mesure de l'inadéquation entre la prédiction à tester y et la prédiction fournie par le prédicteur M entraîné sur l'ensemble d'entraînement $\mathcal{B}_N$

$$S_{N+1}^y = S(y, M(X_{N+1}; \mathcal{B}_N))$$

avec S une fonction appelée "mesure de non-conformité", par exemple le carré du résidu en régression (Kato et al. (2023); Balasubramanian et al. (2014), Section 2.9.5 de Vovk et al. (2022)).

En général, il n'existe pas de formule fermée qui permette de décrire le prédicteur $M(.; \mathcal{B}_{N+1}^{y,-i})$ entraîné sur $\mathcal{B}_{N+1}^{y,-i}$, en fonction de y et de i. Il faut alors effectuer N entraînements (autant que d'indices i) pour chaque prédiction y, ce qui est d'autant plus coûteux avec des prédicteurs "complexes". Le coût induit est ensuite multiplié par la taille de l'ensemble des prédictions y possibles. Par exemple, pour $y \in \mathbb{R}$, le coût computationel est infini. C'est pour cela que l'approche "full conformal" est abandonnée au profit d'autres approches alternatives moins coûteuses telles que "split conformal" et "cross conformal" présentées dans les sections suivantes.

2.2 Cadre "split conformal"

La "p-valeur conforme" dans le cadre "split conformal" de Papadopoulos (2008) est calculée en deux temps. D'abord, une partie $\mathcal{B}_{\text{Train}}$ de l'ensemble d'entraînement est utilisée pour ajuster le prédicteur M. Ensuite, les scores de "non-conformité" sont évalués sur le sous-ensemble $\mathcal{B}_{\text{Calib}} = \mathcal{B}_N \setminus \mathcal{B}_{\text{Train}}$ et en (X_{N+1}, y). Pour $(X_i, Y_i) \in \mathcal{B}_{\text{Calib}}$

$$S_i = S(Y_i, M(X_i; \mathcal{B}_{\text{Train}})) \text{ et } S_{N+1}^y = S(y, M(X_{N+1}; \mathcal{B}_{\text{Train}})).$$

La "p-valeur conforme" est donnée par

$$\pi_{N+1}^{\text{SCP}}(X_{N+1}, y) = \frac{\text{Card}(\{(X_i, Y_i) \in \mathcal{B}_{\text{Calib}} : S_i \geq S_{N+1}^y\}) + 1}{\text{Card}(\mathcal{B}_{\text{Calib}}) + 1}.$$

Un seul entraînement est fait sur $\text{Card}(\mathcal{B}_{\text{Train}})$ observations, et le nombre de scores de "non-conformité" est $\text{Card}(\mathcal{B}_{\text{Calib}})$. La région $\hat{C}_\alpha^{\text{SCP}}(X_{N+1})$ a alors la garantie

$$\mathbb{P}(Y_{N+1} \in \hat{C}_\alpha^{\text{SCP}}(X_{N+1})) \geq 1 - \alpha.$$

2.3 Cadre "cross conformal"

La "p-valeur conforme" dans la cadre "cross conformal" de Vovk (2015) est calculée en trois temps. D'abord, $\mathcal{B}_N$ est séparé en K sous-ensembles $\mathcal{B}^k$ disjoints. Ensuite, pour chaque sous-ensemble $\mathcal{B}^k$, une "p-valeur" $\pi^k(X_{N+1}, y)$ est calculée comme dans "split conformal"

$$\pi_k(X_{N+1}, y) = \frac{\text{Card}(\{(X_i, Y_i)) \in \mathcal{B}^k : S_i^k \geq S_{N+1}^{y,k}\}) + 1}{\text{Card}(\mathcal{B}^k) + 1},$$

avec

$$S_i^k = S(Y_i, M(X_i; \mathcal{B}_N \setminus \mathcal{B}_k) \text{ et } S_{N+1}^{y,k} = S(y, M(X_{N+1}; \mathcal{B}_N \setminus \mathcal{B}_k).$$

Enfin, la "p-valeur conforme" $\pi_{N+1}^{K,\text{CV}}(X_{N+1}, y)$ est la moyenne des $\pi_k(X_{N+1}, y)$

$$\pi_{N+1}^{K;\text{CV}}(X_{N+1}, y) = \frac{1}{K} \sum_{k=1}^{K} \pi_k(X_{N+1}, y).$$

K entraînements sont faits sur $N + 1 - \text{Card}(\mathcal{B}^k)$ observations pour un nombre de scores de "non-conformité" de $\text{Card}(\mathcal{B}^k)$. Pour $K = N$, la région $\hat{C}_\alpha^{\text{N;CV}}(X_{N+1})$ a la garantie suivante (Théorème 1 dans Barber et al. (2021))

$$\mathbb{P}(Y_{N+1} \in \hat{C}_\alpha^{\text{N;CV}}(X_{N+1})) \geq 1 - 2\alpha.$$

3 Approximation des scores par fonctions d'influence

Pour des prédicteurs "complexes", il n'existe en général pas de formule fermée pour $(y, i) \mapsto M(.; \mathcal{B}_{N+1}^{y,-i})$. En classification, Martinez et al. (2023) ont utilisé les fonctions d'influence pour construire une approximation du prédicteur $M(.; \mathcal{B}_{N+1}^{y,-i})$ à partir du prédicteur entraîné sur l'ensemble d'entraînement $M(.; \mathcal{B}_N)$ en quantifiant "l'influence" du rajout de (X_{N+1}, y) et du retrait de (X_i, Y_i) de $\mathcal{B}_N$ sur les scores de "non-conformité". Cette approximation a le mérite de *rendre accessible le calcul de la région prédictive de confiance approchée.*

3.1 Minimisation du risque empirique

Dans la suite, pour un modèle prédictif M paramétré par θ, le risque empirique $\hat{R}$ évalué sur $\mathcal{B}$ est défini par

$$\hat{R}_{\mathcal{B}}(\theta) = \frac{1}{|\mathcal{B}|} \sum_{(x,y)\in\mathcal{B}} l(y, M(x; \theta)) = \frac{1}{|\mathcal{B}|} \sum_{z\in\mathcal{B}} l(z; \theta),$$

avec l une fonction de perte et $z = (x, y)$. Dans le cadre de la minimisation du risque empirique, le prédicteur $M(.; \mathcal{B})$ entraîné sur $\mathcal{B}$ est celui dont le paramètre $\hat{\theta}(\mathcal{B})$ minimise le risque empirique

$$\hat{\theta}(\mathcal{B}) \in \arg\min_{\theta\in\Theta} \hat{R}_{\mathcal{B}}(\theta).$$

Les scores de "non-conformité" pour le minimiseur du risque empirique sont donnés par

$$S_i^y = S(Y_i, M(X_i; \hat{\theta}(\mathcal{B}_{N+1}^{y,-i}))) = S(Z_i; \hat{\theta}(\mathcal{B}_{N+1}^{y,-i}))$$
$$S_{N+1}^y = S(y, M(X_{N+1}; \hat{\theta}(\mathcal{B}_N))) = S(Z_{N+1}^y; \hat{\theta}(\mathcal{B}_N)).$$

Martinez et al. (2023) définissent l'approximation $\tilde{S}_i^y$ du score S_i^y, à l'aide de la fonction d'influence d'une observation sur le score pour $i = 1, \ldots, N$ comme

$$S_i^y \approx \tilde{S}_i^y = S(Z_i; \hat{\theta}(\mathcal{B}_N)) + \mathcal{I}_{S;Z_i}(Z_{N+1}^y; \hat{\theta}(\mathcal{B}_N)) - \mathcal{I}_{S;Z_i}(Z_i; \hat{\theta}(\mathcal{B}_N)),$$

où l'influence $\mathcal{I}_{S;u}(z, \hat{\theta}(\mathcal{B}_N))$ du couple $z = (x, y)$ sur le score S évalué en $u = (u_{\text{in}}, u_{\text{out}})$ est donnée par

$$\mathcal{I}_{S;u}(z; \hat{\theta}(\mathcal{B}_N)) = [\nabla_\theta s(u; \theta)]^T_{\theta=\hat{\theta}(\mathcal{B}_N)} I_{\hat{\theta};B_N}(z).$$

Ici, $I_{\hat{\theta};\mathcal{B}_N}(z)$ représente l'influence du couple $z = (x, y)$ sur le minimiseur du risque empirique $\hat{\theta}$ entraîné sur $\mathcal{B}_N$

$$I_{\hat{\theta};\mathcal{B}_N}(z) = -\frac{1}{N} \left[(\nabla_\theta^2 \hat{R}_{\mathcal{B}_N}(\theta))^{-1} \nabla_\theta l(z; \theta) \right]_{\theta=\hat{\theta}(\mathcal{B}_N)}. \tag{1}$$

Remarquons que l'inverse de la hessienne du risque empirique intervient dans l'expression de $I_{\hat{\theta},\mathcal{B}_N}(z)$. Le calcul et l'inversion de cette matrice sont coûteux sur le plan computationel et d'autant plus à mesure que la dimension du paramètre θ augmente. Cependant, des approximations telles que K-FAC (Ba et al., 2016) et GGN (Gargiani et al., 2020) pourraient diminuer ce coût de sorte permettre d'obtenir des régions approchées pour des modèles hautement paramétrés du type réseaux de neurones profonds.

3.2 Descente de gradient

Pour des prédicteurs "complexes", en général nous ne disposons du minimiseur du risque empirique $\hat{\theta}(\mathcal{B})$. Le paramètre $\theta_T(\mathcal{B})$ n'est plus le minimiseur du risque empirique, mais est plutôt fourni au terme de T itérations d'un algorithme d'optimisation de type descente de gradient appliquée à la minimisation de $\theta \mapsto \hat{R}_{\mathcal{B}}(\theta)$.

Les itérés successifs sont donnés par

$$\begin{aligned} \theta_0 &= \theta^{(0)} \\ \text{Pour } t = 0, \dots, T-1, \qquad \theta_{t+1} &= \theta_t - \eta_t [\nabla_\theta \hat{R}_{\mathcal{B}}(\theta)]_{\theta=\theta_t}. \end{aligned} \tag{2}$$

Définition 1 *Dans ce cadre, les scores de "non-conformité" S_i^y pour $i = 1, \dots, N+1$ sont donnés par*

$$\begin{aligned} S_i^y &= S(Y_i, M(X_i; \theta_T(\mathcal{B}_{N+1}^{y,-i}))) = S(Z_i; \theta_T(\mathcal{B}_{N+1}^{y,-i})) \\ S_{N+1}^y &= S(y, M(X_{N+1}; \theta_T(\mathcal{B}_N))) = S(Z_{N+1}^y; \theta_T(\mathcal{B}_N)). \end{aligned}$$

En appliquant la stratégie précédemment décrite basée sur les fonctions d'influence, l'approximation $\tilde{S}_i^y$ par fonctions d'influence du scores S_i^y est donnée pour $i = 1, \dots, N$ par

$$S_i^y \approx \tilde{S}_i^y = S(Z_i, \theta_T(\mathcal{B}_N)) + \mathcal{I}_{S;Z_i}(Z_{N+1}^y; \theta_T(\mathcal{B}_N)) - \mathcal{I}_{S;Z_i}(Z_i; \theta_T(\mathcal{B}_N)).$$

où $\mathcal{I}_{S;u}(z; \theta_T(\mathcal{B}_N))$ de $z = (x, y)$ sur le score S évalué en $u = (u_{\text{in}}, u_{\text{out}})$ est donnée par

$$\mathcal{I}_{S;u}(z; \theta_T(\mathcal{B}_N)) = [\nabla_\theta s(u, \theta)]_{\theta=\theta_T(\mathcal{B}_N)}^T I_{\theta_T;\mathcal{B}_N}(z).$$

Ici $I_{\theta_T;\mathcal{B}_N}(z)$ représente l'influence de l'observation $z = (x, y)$ sur l'itération θ_T de l'algorithme (2) qui lui, vise à minimiser le risque empirique $\hat{R}$ évalué sur $\mathcal{B}_N$. Il est calculer itérativement en parallèle avec la descente de gradient à l'aide de ses T itérés $(\theta_t)_{t=0}^{T-1}$.

Lemme 1 *L'influence $I_{\theta_T;\mathcal{B}_N}(z)$ de $z = (x, y)$ sur le paramètre θ_T entraîné sur $\mathcal{B}_N$ est donnée par*

$$I_{\theta_T;\mathcal{B}_N}(z) = \frac{1}{N} C_T(z),$$

où $C_T(z)$ est l'itéré T du schéma itératif suivant

$$\begin{aligned} C_1(z) &= -\eta_0 [\nabla_\theta l(z; \theta)]_{\theta=\theta_0} \\ \textit{Pour } t = 2, \dots, T, \qquad C_t(z) &= (I - \eta_{t-1} [\nabla_\theta^2 \hat{R}_{\mathcal{B}_N}(\theta)]_{\theta=\theta_{t-1}}) C_{t-1}(z) \\ &\quad - \eta_{t-1} [\nabla_\theta l(z; \theta)]_{\theta=\theta_{t-1}}. \end{aligned} \tag{3}$$

Remarquons que la matrice hessienne du risque empirique intervient à chaque itération. Cependant, contrairement à la méthode précédente (1), elle n'est pas inversée. La méthode peut alors être appliquée même quand la matrice hessienne est non-inversible.

Éléments de preuve pour le lemme 1. Posons $(\theta_{t+1}^{\epsilon,z})_{t=1}^{T}$ comme étant les itérés d'un algorithme d'optimisation de type descente de gradient appliquée à la minimisation de $\hat{R}_{\mathcal{B}_N}(\theta) + \epsilon l(z;\theta)$, le risque empirique $\hat{R}_{\mathcal{B}_N}(\theta)$ auquel est ajouté $\epsilon l(z,\theta)$. Les itérés successifs sont donnés par

$$\begin{aligned} \theta_0^{\epsilon,z} &= \theta^{(0)} \\ \text{Pour } t = 0,\dots,T-1, \qquad \theta_{t+1}^{\epsilon,z} &= \theta_t^{\epsilon,z} - \eta_t[\nabla_\theta(\hat{R}_{\mathcal{B}_N}(\theta) + \epsilon l(z;\theta))]_{\theta=\theta_t^{\epsilon,z}}. \end{aligned} \tag{4}$$

Pour $\epsilon = \frac{1}{N}$, $\theta_T^{\frac{1}{N},z}$ est le paramètre fourni à l'issue de T itérations du schéma (2) le risque empirique à minimiser est évalué sur $\mathcal{B}_N \cup \{z\} = \{Z_1,\dots,Z_N,z\}$. $\theta_T^{\frac{1}{N},z}$ est approché par une approximation de Taylor d'ordre 1

$$\theta_T^{\frac{1}{N},z} = \theta_T^{0,z} + \frac{1}{N}\left[\frac{d}{d\epsilon}\theta_T^{\epsilon,z}\right]_{\epsilon=0} = \theta_T + \frac{1}{N}C_T(z).$$

Le schéma (3) est ensuite déduit de (4). Notons que le même raisonnement est appliqué pour $\mathcal{B}_N \setminus \{Z_i\} = \{Z_1,\dots,Z_{i-1},Z_{i+1},\dots,Z_N\}$ en prenant $z = Z_i$ et $\epsilon = -\frac{1}{N}$.

4 Régions prédictives de confiance approchées

4.1 Calcul des régions approchées

Nous supposons dans la suite que la fonction de perte l et la fonction de score S sont quadratiques. Dans ce cas, l'approximation des scores $\tilde{S}_i^y$ par fonctions d'influence est donnée pour $i = 1,\dots,N$ par

$$\tilde{S}_i^y = a_i y + b_i \text{ et } \tilde{S}_{N+1}^y = S_{N+1}^y = (y - M(X_{N+1};\theta))^2$$

où les coefficients a_i et b_i ont des expressions complètement connues qui dépendent des estimateurs envisagés (minimiseur du risque empirique ou descente de gradient). Nous obtenons la forme suivante pour l'approximation de la région conforme

$$\tilde{C}_\alpha(X_{N+1}) = \left(\bigcup_{k=0,\dots,K:p_k>\alpha}]y_k, y_{k+1}[\right) \cup \left(\bigcup_{k=1,\dots,K:p_k>\alpha} \{y_k\}\right),$$

où la "p-valeur conforme" p_k associée à l'intervalle $]y_k, y_{k+1}[$ est donnée pour $k = 0,\dots,K$ par

$$p_k = \frac{\text{Card}(\{i = 1,\dots,N :]y_k, y_{k+1}[\subseteq C_i\}) + 1}{N+1},$$

et celle associée au singleton $\{y_k\}$ est donnée pour $k = 1,\dots,K$ par

$$p_k = \frac{\text{Card}(\{i = 1,\dots,N : y_k \in C_i\}) + 1}{N+1}.$$

et enfin les y_k sont définis à partir des C_i.

L'intervalle C_i est celui sur lequel $\tilde{S}_i^y$ est supérieur à S_{N+1}^y. De manière équivalente celui sur lequel $P_i(y) := S_{N+1}^y - \tilde{S}_i^y$, défini dans le lemme 4.1, est négatif.

Lemme 2 *Quand l'estimateur envisagé est le minimiseur du risque empirique, P_i est un polynôme en y qui est donné pour $i = 1, \ldots, N$ par*

$$P_i(y) = S_{N+1}^y - \tilde{S}_i^y = y^2 - (2M(X_{N+1}; \hat{\theta}(\mathcal{B}_N)) + a_i)y + M(X_{N+1}; \hat{\theta}(\mathcal{B}_N))^2 - b_i$$

Quand l'estimateur envisagé est fourni à l'issue de T itérations d'un algorithme d'optimisation de type descente de gradient, P_i est un polynôme en y qui est donné pour $i = 1, \ldots, N$ par

$$P_i(y) = S_{N+1}^y - \tilde{S}_i^y = y^2 - (2M(X_{N+1}; \theta_T) + a_i)y + M(X_{N+1}; \theta_T)^2 - b_i$$

Ce qui implique que l'intervalle C_i pour $i = 1, \ldots, N$ est donnée par

$$C_i = \begin{cases} [u_i, v_i] & \text{si } P_i \text{ admet deux racines réelles distinctes,} \\ \{w_i\} & \text{si } P_i \text{ n'admet qu'une seule racine réelle,} \\ \emptyset & \text{si } P_i \text{ n'admet pas de racines réelles.} \end{cases}$$

Enfin, $y_0 = -\infty$, $y_{K+1} = +\infty$ et $y_1, \ldots, y_K$ sont les u_i, v_i, w_i ordonnés dans l'ordre croissant.

4.2 Illustration

Les (X_i, Y_i) sont tirées indépendamment et sont données par

$$X_i \sim \mathcal{U}([-1, 1]) \text{ et } Y_i = f(1.8(X_i + 0.8)) + \epsilon_i,$$

où $\epsilon_i \sim \mathcal{N}(0, 0.1)$ et

$$f : x \mapsto f(x) = 3\exp\left(-\frac{(x-1.8)^2}{0.28}\right) + 1.6\exp\left(-\frac{(x-0.7)^2}{0.13}\right) + 2\exp\left(-2.8x\right).$$

Le prédicteur est un réseau de neurones à une couche cachée à 10 neurones, ajusté par 200 itérations de la descente de gradient avec un pas de 0.05. "Split conformal" (a) est appliqué avec 100 observations d'entraînement, 100 de calibration. Les approches 3.1 (b) et 3.2 (c) sont appliqués avec 200 observations d'entraînement. Pour chaque méthode, les régions à niveau de confiance $1 - \alpha = 90\%$ sont évaluées sur 200 points de validation.

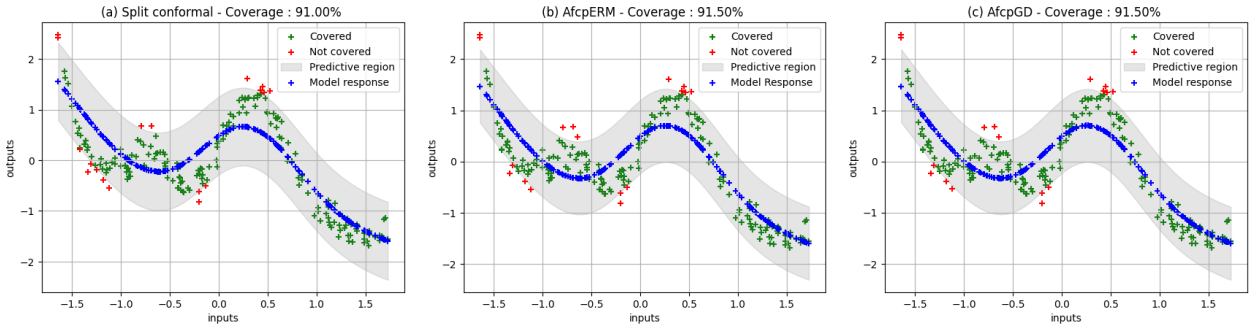

Figure 1: Régions prédictives et taux de couverture

Remarquons que pour chaque méthode, la figure 1 montre que le taux de couverture, c'est-à-dire la proportion de points de validation inclus dans les régions prédictives (ici en vert) est proche de $1 - \alpha$, le niveau de confiance spécifié.

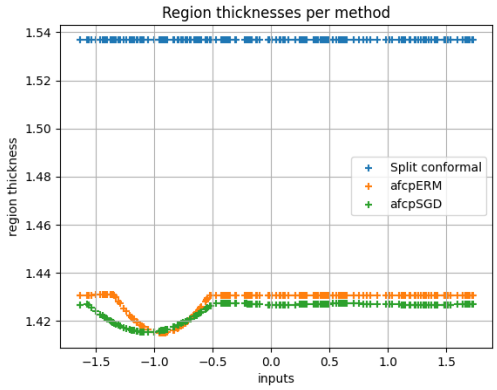

Figure 2: Épaisseur des régions par méthode

De plus, la figure 2 montre que la méthode 3.2 donne des régions moins épaisses que celles de la méthode 3.1, qui elles sont moins épaisses que celles données par "split conformal". Cependant, il faudra davantage d'expérimentations pour trancher.

5 Discussion

Le présent travail sera suivi d'une étude théorique et empirique de la précision l'approximation des scores, la précision de l'approximation des p-valeurs, le taux de couverture (validité) et la taille des régions prescrites (précision) en fonction de la taille de l'échantillon d'entraînement. Ces performances seront illustrées suivant des critères d'intérêt tels que la "complexité" du modèle prédictif, la dimension des covariables, la valeur du paramètre de régularisation dans le risque empirique et la valeur du seuil de confiance. L'approche proposé sera enfin comparée au prédicteur conforme oracle de Ndiaye (2022), "split conformal" de Papadopoulos (2008) et "cross conformal" de Vovk (2015). Nous envisageons aussi d'effectuer ces évaluations quand la matrice hessienne est remplacée par une approximation (Gargiani et al., 2020; Ba et al., 2016).

Bibliographie

Ba, J., Grosse, R., and Martens, J. (2016). Distributed second-order optimization using kronecker-factored approximations. In *International Conference on Learning Representations*.

Balasubramanian, V., Ho, S.-S., and Vovk, V. (2014). *Conformal prediction for reliable machine learning: theory, adaptations and applications*. Newnes.

Barber, R. F., Candes, E. J., Ramdas, A., and Tibshirani, R. J. (2021). Predictive inference with the jackknife+.

Gargiani, M., Zanelli, A., Diehl, M., and Hutter, F. (2020). On the promise of the stochastic generalized gauss-newton method for training dnns. *arXiv preprint arXiv:2006.02409*.

Kato, Y., Tax, D. M., and Loog, M. (2023). A review of nonconformity measures for conformal prediction in regression. *Conformal and Probabilistic Prediction with Applications*, pages 369–383.

Lei, J. (2019). Fast exact conformalization of the lasso using piecewise linear homotopy. *Biometrika*, 106(4):749–764.

Martinez, J. A., Bhatt, U., Weller, A., and Cherubin, G. (2023). Approximating full conformal prediction at scale via influence functions. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 37, pages 6631–6639.

Ndiaye, E. (2022). Stable conformal prediction sets. In *International Conference on Machine Learning*, pages 16462–16479. PMLR.

Ndiaye, E. and Takeuchi, I. (2019). Computing full conformal prediction set with approximate homotopy. *Advances in Neural Information Processing Systems*, 32.

Ndiaye, E. and Takeuchi, I. (2023). Root-finding approaches for computing conformal prediction set. *Machine Learning*, 112(1):151–176.

Nouretdinov, I., Melluish, T., and Vovk, V. (2001). Ridge regression confidence machine. In *ICML*, pages 385–392. Citeseer.

Papadopoulos, H. (2008). Inductive conformal prediction: Theory and application to neural networks. In *Tools in artificial intelligence*. Citeseer.

Vovk, V. (2015). Cross-conformal predictors. *Annals of Mathematics and Artificial Intelligence*, 74:9–28.

Vovk, V., Gammerman, A., and Shafer, G. (2005). *Algorithmic learning in a random world*, volume 29. Springer.

Vovk, V., Gammerman, A., and Shafer, G. (2022). *Algorithmic Learning in a Random World*. Second edition.