Prévision Conforme adaptative avec un pas de gradient explicite

Guillaume Principato[1,2,*]

[1] *Laboratoire de Mathématiques d'Orsay (LMO), Université Paris-Saclay, France*
[2] *EDF R&D, France*
[*] *guillaume.principato@universite-paris-saclay.fr*

Résumé. La prévision conforme est une méthode qui permet de construire des intervalles de prévision théoriquement valides à partir d'une prévision boîte noire. Ces dernières années, des méthodes ont été proposées pour l'appliquer à des séries temporelles qui, par définition, ne vérifient pas l'hypothèse d'échangeabilité nécessaire à la théorie initiale. Nous présentons les méthodes adaptatives de la littérature en insistant sur l'interprétation des garanties que l'on peut obtenir. Ensuite, nous montrons de nouveaux résultats sur un algorithme de descente de gradient à pas de gradient adaptatif qui n'avait, jusqu'alors, qu'été utilisé comme intermédiaire. Nous verrons, en le comparant aux autres algorithmes adaptatifs, que son utilisation en tant que tel est justifiée.

Mots-clés. Prévision conforme, inférence conforme adaptative, série temporelle, optimisation convexe en ligne.

Abstract. Conformal prediction is a method for constructing theoretically valid prediction sets from any black-box forecaster. In recent years, methods have been proposed to apply it to time series which, by definition, do not verify the exchangeability assumption required by the initial theory. We present adaptive methods from the literature, focusing on the interpretation of the guarantees that can be obtained. Next, we show new results on an gradient descent algorithm with adaptive step size that had hitherto only been used as an intermediary. By comparing it with other adaptive algorithms, we show that its use as such is justified.

Keywords. Conformal prediction, adaptive conformal inference, time series, online convex optimization.

1 Introduction

Considérons une série temporelle $(X_t, Y_t)_{t\in[\![1,T]\!]}$ avec $Y_t \in \mathbb{R}$ la variable que l'on cherche à prédire et $X_t \in \mathbb{R}^m$ un ensemble de m covariables qui expliquent Y_t. L'objectif est de construire un intervalle de prévision qui a une probabilité $1-\alpha$ de contenir l'observation Y_t, où α est un paramètre que l'utilisateur choisit en fonction du type d'intervalle qu'il souhaite construire. Le cadre général des prévisions conformes introduit par Vovk et al. (2005) permet cela dans le cadre échangeable.

Pour ce faire, il faut construire une distribution empirique de scores, que l'on note $\mathcal{D}_t$, pour ensuite obtenir un intervalle en sélectionnant y tel que son score soit plus petit qu'une proportion $1-\alpha$ des scores : $\hat{C}_t(\alpha) := \{y : S_t^y \leq \text{ Quantile}\,(1-\alpha, \mathcal{D}_t)\}$. Par exemple, si on note $\hat{\mu}(X_t)$ la prévision de la moyenne associée à Y_t à partir des covariables X_t on peut choisir comme score $S_t^y = |\hat{\mu}(X_t) - y|$.

L'inférence conforme adaptative (Adaptive Conformal Inference, ACI) est une méthode proposée par Gibbs and Candes (2021) qui se fonde sur le schéma classique de la prévision conforme mais qui intègre un aspect adaptatif. Cela permet alors d'obtenir des résultats même lorsque l'hypothèse d'échangeabilité n'est pas vérifiée.

Dans ce résumé, nous présentons différentes variantes d'algorithmes de type ACI introduites récemment par la littérature, en détaillant des critères qui attestent de la performance de ces algorithmes. La contribution principale est l'étude plus approfondie d'un algorithme à pas de gradient adaptatif (Algorithme 2). Nous montrons qu'il vérifie simultanément un ensemble de critères.

1.1 Présentation de l'algorithme ACI

Sans aucune hypothèse sur la distribution des observations Y_t, il est impossible d'avoir des garanties sur la probabilité de couverture de l'intervalle de prévision $\hat{C}_t^{\alpha}$ pour chacune des observations. Cependant, la littérature considère alors l'hypothèse suivante :

$$\text{Il existe un } \alpha_t^* \text{ tel que : } \mathbb{P}\left(Y_t \in \hat{C}_t(\alpha_t^*)\right) = 1-\alpha \tag{1}$$

où α_t^* et α ne sont pas nécessairement égaux ce qui permet de modéliser les changements de lois des Y_t. L'idée est alors de ne plus considérer l'intervalle $\hat{C}_t(\alpha)$ mais $\hat{C}_t(\alpha_t)$ avec α_t qui est appris sur les données afin de pouvoir s'adapter à un potentiel changement dans la distribution. Pour ce faire, Gibbs and Candes (2021) proposent une formule récurrente très interprétable :

Algorithme 1. ***(ACI)***

$$\alpha_{t+1} = \alpha_t + \gamma\,(\alpha - \text{err}_t) \qquad \textit{avec } \text{err}_t := \begin{cases} 0, & \textit{si } Y_t \in \hat{C}_t\,(\alpha_t) \\ 1, & \textit{si } Y_t \notin \hat{C}_t\,(\alpha_t) \end{cases} \tag{2}$$

Ainsi, si l'observation à l'instant précédent n'est pas couverte par l'intervalle, c'est probablement que ce dernier est trop étroit. Par conséquent, on l'agrandit en regardant un $\alpha_{t+1} < \alpha_t$. A l'inverse, si l'intervalle couvre l'observation, c'est qu'il est probablement trop large et donc $\alpha_{t+1} > \alpha_t$, ce qui réduit la largeur de l'intervalle. La formule récurrente (2) peut être vue comme un pas de descente de gradient par rapport à la fonction de perte quantile (3).

1.2 Garanties et limitations

L'hypothèse d'échangeabilité n'étant pas vérifiée dans le cas des séries temporelles, on passe d'un critère de validité en probabilité (Vovk et al., 2005) à un critère empirique de validité asymptotique.

Critère 1. ***(Validité)***
Un intervalle de prévision $\left(\hat{C}_t^{\alpha}\right)_{t\in[\![1,T]\!]}$ *est dit valide asymptotiquement si sa couverture asymptotique vaut* $1-\alpha$ *:*

$$\lim_{T\to\infty}\frac{1}{T}\sum_{t=1}^{T}\mathbb{1}_{\{Y_t\in\hat{C}_t^{\alpha}\}}\overset{p.s.}{=}1-\alpha$$

Gibbs and Candes (2021) montrent que l'intervalle produit par ACI est valide asymptotiquement avec une vitesse de convergence en $O\left(1/T\right)$ pour tout $\gamma\in\mathbb{R}^{\star}$ fixe, ce qui est la meilleure vitesse parmi les algorithmes du même type. Ce résultat étant valable pour tout $\gamma\in\mathbb{R}^{*}$ fixe, il ne donne pas de moyen de déterminer la valeur du pas de gradient γ. Ce paramètre est pourtant très important en pratique comme le soulignent Gibbs and Candes (2021). Une manière de déterminer γ pourrait être de le contraindre par un second critère. En l'occurrence, notons $\beta_t := \sup\{\beta : Y_t\in\hat{C}_t(\beta)\}$ le plus grand niveau de quantile tel qu' Y_t soit dans l'intervalle de prévision et la perte quantile de niveau $1-\alpha$:

$$\ell(\beta_t,\theta) := (\alpha-\mathrm{err}_t)(\beta_t-\theta) \tag{3}$$

On remarque que $\nabla_{\theta}\ell\left(\beta_t,\alpha_t\right)=\mathrm{err}_t-\alpha$. Ainsi, (2) s'écrit également $\alpha_{t+1}=\alpha_t-\gamma\nabla_{\theta}\ell\left(\beta_t,\alpha_t\right)$, ce qui correspond à une étape de descente de gradient. Par conséquent, il est possible d'utiliser les méthodes classiques d'optimisation convexe en ligne pour obtenir des garanties théoriques comme, par exemple, des bornes de regret sur la perte quantile.

Critère 2. ***(Efficacité)***
On définit le regret d'un algorithme de type ACI par rapport à la perte quantile par :

$$\mathrm{Reg}_T(ACI)=\sum_{t=1}^{T}\ell(\beta_t,\alpha_t)-\inf_{\theta}\ell(\beta_t,\theta)$$

On considère alors qu'un algorithme est performant lorsque ce regret peut être borné par un terme sous-linéaire.

Les travaux de Zinkevich (2003) suggèrent alors de choisir γ fixe tel que $\gamma=O\big(1/\sqrt{T}\big)$ pour (2) ce qui permet d'obtenir $\mathrm{Reg}_T(ACI)\leq O\big(\sqrt{T}\big)$. Cependant, cette approche n'est pas satisfaisante car cela ne garantit pas de bons résultats à tout instant, mais uniquement pour un horizon T fixé.

2 Améliorations d'ACI

Nous avons conclu la partie précédente en insistant sur le fait que le choix du pas de gradient γ était un enjeu majeur dans la construction d'un algorithme de type ACI. La littérature a proposé dans un premier temps de choisir une valeur constante, et a étudié dans un second temps des valeurs adaptatives (changeant au cours du temps, et déterminées selon les observations), ce dont nous rendons compte dans cette partie.

2.1 Méthodes d'agrégation d'experts

Une façon d'obtenir un pas de gradient γ qui varie en fonction de l'instant est de procéder par agrégation d'experts. Il s'agit de définir une grille de γ candidats et de considérer une instance d'ACI pour chaque γ. On obtient alors un ensemble d'experts $\{\alpha_t^\gamma\}$ que l'on agrège en considérant comme fonction de coût la perte quantile (3). Cette méthode est celle choisie par Zaffran et al. (2022) et Gibbs and Candès (2022) avec respectivement les algorithmes Online Expert Aggregation on ACI (AgACI) et Dynamically-tuned ACI (DtACI).

Ces algorithmes donnent d'excellents résultats pratiques (même si cette façon de construire l'intervalle rend difficile l'obtention d'une garantie par rapport au critère 1). Pour expliquer cette réussite, Gibbs and Candès (2022) mettent en avant le fait que leur algorithme produit des niveaux de quantile $1-\alpha_t$ qui sont proches des niveaux $1-\alpha_t^*$[1]. Plus formellement, on peut introduire le critère suivant :

Critère 3. ***(Efficacité)*** *Soit α_t^* tel que $\mathbb{P}\left(Y_t \in \hat{C}_t\left(\alpha_t^*\right) \mid \{\beta_s\}_{s<t}\right) = 1-\alpha$ et α_t produit par un algorithme de type ACI. On souhaite pouvoir contrôler le terme :*

$$\frac{1}{T}\sum_{t=1}^{T}\frac{\mathbb{E}\left[\left(\alpha_t-\alpha_t^*\right)^2\right]}{2}$$

avec l'espérance qui porte sur Y_t et les $(Y_s)_{s\in[\![1,t-1]\!]}$.

Ce critère valorise les algorithmes qui s'adaptent rapidement aux changements de distribution et permet, sous réserve d'hypothèses supplémentaires, d'obtenir des garanties plus interprétables.

[1] Le niveau α_t^* défini dans le critère 3 n'est pas en adéquation avec sa précédente définition (1) pour laquelle il n'y avait pas de conditionnement par rapport aux $(\beta_s)_{s<t}$. Nous conservons cette double notation pour nous conformer au reste de la littérature.

2.2 Algorithmes à pas de gradient γ_t explicite

Des approches différentes sont proposées par Bhatnagar et al. (2023) et Angelopoulos et al. (2024) pour définir séquentiellement le paramètre γ_t. Dans les deux cas, l'idée est de proposer un γ_t qui, dans un premier temps, est suffisamment grand pour s'adapter à un potentiel changement de loi des observations Y_t, puis, dans un second temps, décroît lentement pour permettre la convergence. Bhatnagar et al. (2023) utilisent un algorithme qu'ils nomment Scale-Free Online Gradient Descent (SF-OGD) qui se caractérise par un pas $\gamma_t = O\big(1/\sqrt{t}\big)$ inspiré du pas γ_t qui minimise le regret à tout instant, tandis que Angelopoulos et al. (2024) proposent une version plus générale avec γ_t décroissant.

Ces derniers montrent qu'en prenant un tel γ_t, il est possible d'obtenir simultanément de bonnes garanties dans le cas adversarial ainsi que dans le cas *i.i.d.* tout en montrant que cela aurait été impossible à γ fixe. Ces résultats récents sont dans le même esprit que la contribution de cet article qui met en valeur le fait que l'algorithme de descente de gradient à pas adaptatif vérifie simultanément différents critères.

3 Descente de gradient à pas adaptatif

Dans cette partie, nous présentons de nouveaux résultats obtenus sur un algorithme déjà proposé dans la littérature de la prévision conforme adaptative par Bhatnagar et al. (2023). Nous montrons qu'il satisfait à la fois un critère de validité (critère 1) et des critères d'efficacité (critères 2 et 3). Ceci constitue une nouveauté dans le sens où aucun des algorithmes proposés précédemment dans la littérature ne satisfait à la fois le critère 1 et 3.

Remarque : Bhatnagar et al. (2023) proposent cet algorithme dans l'optique de servir d'intermédiaire à un algorithme fortement adaptatif (Orabona, 2019) (Strongly Adaptive Online Conformal Prediction, SAOCP). Ces derniers n'ont donc pas étudié en détail l'algorithme de descente de gradient à pas adaptatif, ce que nous faisons ici.

3.1 Présentation de l'algorithme

D'abord appliqué aux prévisions conformes sous le nom SF-OGD, l'algorithme de descente de gradient à pas adaptatif introduit par Zinkevich (2003) pour l'optimisation convexe séquentielle, est une façon d'obtenir un pas de gradient γ_t qui ait une expression explicite et donc pour laquelle il est plus simple d'obtenir de bonnes garanties théoriques.

Algorithme 2.

$$\alpha_{t+1} = \alpha_t + \frac{\eta}{\sqrt{\sum_{s=1}^{t} \nabla_\theta \ell\left(\beta_s, \alpha_s\right)^2}} \left(\alpha - \mathrm{err}_t\right) \tag{4}$$

Notons D le maximum des α_t. On peut choisir $\eta = D/\sqrt{2}$ pour minimiser la borne du regret à tout instant (Orabona, 2019).

Remarque : La formule (4) revient à écrire $\gamma_t = \dfrac{\eta}{\sqrt{\sum_{s=1}^{t} \nabla_\theta \ell\left(\beta_s, \alpha_s\right)^2}}$, ce qui est un $O\left(\frac{1}{\sqrt{t}}\right)$.

3.2 Garantie de validité

Comme souligné par Angelopoulos et al. (2024), sous réserve d'avoir des γ_t décroissants, un algorithme construit d'après la formule (2) répond au critère 1. Nous vérifions cela en nous inspirant de la preuve de Gibbs and Candes (2021) pour l'algorithme 2 :

Proposition 1. *Prenons la convention $\hat{Q}_t(x) = -\infty$ pour $x < 0$ et $\hat{Q}_t(x) = +\infty$ pour $x > 1$. Alors, avec probabilité 1, on a pour tout $T \in \mathbb{N}$,*

$$\left| \frac{1}{T} \sum_{t=1}^{T} \mathrm{err}_t - \alpha \right| \leq O\left(\frac{1}{\sqrt{T}}\right)$$

et donc

$$\lim_{T \to \infty} \frac{1}{T} \sum_{t=1}^{T} \mathbb{1}_{\{Y_t \in \hat{C}_t^\alpha\}} \overset{p.s.}{=} 1 - \alpha$$

La contrepartie par rapport au cas à γ fixe est que l'on perd en vitesse de convergence ($1/T$) (il est à noter que la vitesse obtenue par Bhatnagar et al. (2023) était encore pire, de l'ordre de $T^{-1/4}$ à facteurs logarithmiques près). Une question que nous voulons continuer d'explorer est l'amélioration de cette vitesse de convergence dans le cas adaptatif.

3.3 Garanties d'efficacité

L'algorithme 2 est en réalité défini de telle sorte que par construction (cf. les résultats de pas de gradient adaptatifs (Zinkevich, 2003) (Orabona, 2019)), le critère 2 soit vérifié. Le coût de l'adaptativité est uniquement une constante multiplicative $\sqrt{2}$ par rapport à la borne de regret que l'on obtient à γ fixe. De plus, ce résultat est beaucoup plus fort que celui que l'on obtient à γ fixe car il est valable à tout instant.

La principale nouveauté est inspirée de ce qui est fait par Gibbs and Candès (2022) et réside dans l'obtention d'une borne qui montre que l'algorithme 2 vérifie le critère 3 :

Proposition 2. *Soit* α_t^* *tel que* $\mathbb{P}\left(Y_t \in \hat{C}_t\left(\alpha_t^*\right) \mid \{\beta_s\}_{s<t}\right) = 1-\alpha$ *et* p *un minorant positif de la densité de chacun des* β_t*. Les* α_t *construits par l'algorithme 2 vérifient :*

$$\frac{1}{T}\sum_{t=1}^{T}\frac{\mathbb{E}\left[\left(\alpha_t-\alpha_t^*\right)^2\right]}{2} \leq O\left(\frac{\sum_{t=2}^{T}\mathbb{E}\left[\left|\alpha_t^*-\alpha_{t-1}^*\right|\right]}{p\sqrt{T}}\right)$$

Cette proposition permet de mesurer à quel point les α_t sont proches des α_t^* en bornant le terme de gauche par un terme qui dépend uniquement des variations relatives des α_t^*. On peut alors interpréter la borne comme étant un $O\left(\sqrt{T}\right)$ dans le pire des cas qui correspond à un environnement totalement adversarial avec la distribution des Y_t qui change significativement à chaque instant. A l'inverse, si la distribution des Y_t ne subit qu'un nombre fini de changements sur un horizon infini, on obtient donc une borne en $O\left(1/\sqrt{T}\right)$. On peut alors raisonnablement espérer qu'en pratique, nous sommes plutôt dans le second cas, ce qui rend cette proposition très intéressante pour les cas d'application.

Remarque : Cette seconde interprétation est, en fait, plus générale que ce qui est présenté par Angelopoulos et al. (2024). En effet, ils présentent pour leur méthode des garanties dans le cadre *i.i.d.* tandis que ce résultat permet également d'obtenir des résultats de convergence de l'estimateur dès lors que le nombre de changements de distribution est faible.

4 Comparaison

Dans cette section, il s'agit de comparer les garanties théoriques des différentes méthodes de type ACI. Parmi elles, on considère ACI, DtACI et surtout l'algorithme 2 pour lequel, nous avons montré de nouvelles garanties.

Méthode	ACI	SF-OGD	DtACI	Algorithme 2
γ adaptatif	Non	Oui	Oui	Oui
Couverture 1	$O\left(\frac{1}{T}\right)$	$O\left(\frac{\log T}{T^{1/4}}\right)$	$O(1)$	$O\left(\frac{1}{T^{1/2}}\right)$
Regret 2	$DG\sqrt{T}$	$(\sqrt{3}+1)DG\sqrt{t}$	?	$\sqrt{2}DG\sqrt{t}$
Efficacité 3	?	?	$\sqrt{\frac{\log T+\sum_{t=2}^{T}\mathbb{E}\left[\left\|\alpha_t^*-\alpha_{t-1}^*\right\|\right]}{T}}$	$\frac{\sum_{t=2}^{T}\mathbb{E}\left[\left\|\alpha_t^*-\alpha_{t-1}^*\right\|\right]}{\sqrt{T}}$

Ce tableau permet de conclure que l'algorithme 2 offre parmi les meilleures garanties selon les critères que nous avons définis. Susmann et al. (2023) comparent également ces méthodes

en les appliquant sur des données simulées à l'aide de leur package R.

Remarque : Nous avons décidé de ne pas inclure SAOCP (Bhatnagar et al., 2023) dans le tableau comparatif car la nature des résultats et des hypothèses rendait difficile toute comparaison vis à vis des critères que nous avons énoncés.

5 Remerciements

Ce travail s'inscrit dans le cadre d'une thèse qui a débuté le 1er Décembre 2023 sous la supervision d'Yvenn Amara-Ouali, Bachir Hamrouche, Yannig Goude, Gilles Stoltz et Jean-Michel Poggi. Je les remercie pour les discussions ainsi que pour leur aide pour la rédaction.

References

Amara-Ouali, Y., Hamrouche, B., Principato, G., and Goude, Y. (2024). Quantifying the uncertainty of electric vehicle charging with probabilistic load forecasting. In preparation.

Angelopoulos, A. N., Barber, R. F., and Bates, S. (2024). Online conformal prediction with decaying step sizes. *arXiv preprint arXiv:2402.01139*.

Bhatnagar, A., Wang, H., Xiong, C., and Bai, Y. (2023). Improved online conformal prediction via strongly adaptive online learning. *arXiv preprint arXiv:2302.07869*.

Gibbs, I. and Candes, E. (2021). Adaptive conformal inference under distribution shift. *Advances in Neural Information Processing Systems*, 34:1660–1672.

Gibbs, I. and Candès, E. (2022). Conformal inference for online prediction with arbitrary distribution shifts. *arXiv preprint arXiv:2208.08401*.

Orabona, F. (2019). A modern introduction to online learning. *arXiv preprint arXiv:1912.13213*.

Susmann, H., Chambaz, A., and Josse, J. (2023). Adaptiveconformal: An R package for adaptive conformal inference. *arXiv preprint arXiv:2312.00448*.

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

Zaffran, M., Féron, O., Goude, Y., Josse, J., and Dieuleveut, A. (2022). Adaptive conformal predictions for time series. In *International Conference on Machine Learning*, pages 25834–25866.

Zinkevich, M. (2003). Online convex programming and generalized infinitesimal gradient ascent. In *International Conference on Machine Learning*, pages 928–936.