# Convergence d'un score d'ensemble en ligne : étude empirique

Benoît Lalloué, Jean-Marie Monnez, Eliane Albuisson

# Convergence d'un score d'ensemble en ligne : étude empirique

Benoît Lalloué [1,3,*], Jean-Marie Monnez [1,3,†], Éliane Albuisson [2,4,5,‡]

[1] *Université de Lorraine, CNRS, Inria*, IECL**, F-54000 Nancy, France*
**Inria, Project-Team BIGS*
***Institut Elie Cartan de Lorraine, Vandoeuvre-lès-Nancy, France*
[2] *Université de Lorraine, CNRS, IECL**, F-54000 Nancy, France*
[3] *Inserm U1116, Centre d'Investigation Clinique Plurithématique 1433, Université de Lorraine, Nancy, France*
[4] *BIOBASE, Pôle S2R, CHRU de Nancy, Vandoeuvre-lès-Nancy, France*
[5] *Faculté de Médecine, InSciDenS, Vandoeuvre-lès-Nancy, France*
∗ *benoit.lalloue@univ-lorraine.fr;* † *jean-marie.monnez@univ-lorraine.fr;*
‡ *eliane.albuisson@univ-lorraine.fr*
*Financement : Programme Investissement d'Avenir ANR-15-RHU-0004*

**Résumé.** Dans un contexte en ligne où des données arrivent de façon continue, on souhaite actualiser les paramètres d'un score "batch" construit à l'aide d'une méthode d'ensemble. On utilise pour cela des processus d'approximation stochastique, dont la convergence a été établie théoriquement par les auteurs, permettant d'actualiser les estimations des paramètres lors de la prise en compte de nouvelles observations sans avoir à conserver toutes les données obtenues précédemment. Nous étudions ici empiriquement la convergence du score en ligne vers le score "batch", en utilisant différents jeux de données à partir desquels on simule des flux de données et différents types de processus.

**Mots-clés.** Apprentissage pour les données massives, approximation stochastique, médecine, méthode d'ensemble, score en ligne.

**Abstract.** In an online setting, where data arrives continuously, we want to update the parameters of a "batch" score constructed with an ensemble method. To do so, we use stochastic approximation processes, the convergence of which has been theoretically established by the authors, so that parameter estimates can be updated when new observations are taken into account without the need to store all the data obtained previously. Here we study empirically the convergence of the online score to the "batch" score, using different datasets from which data streams are simulated and using different types of processes.

**Keywords.** Learning for big data, stochastic approximation, medicine, ensemble method, online score.

# 1 Introduction

Afin d'établir un score d'événement en ligne actualisé lors de l'arrivée de nouvelles données d'apprentissage sans avoir à stocker l'ensemble des données obtenues, nous avons entrepris une étude en plusieurs étapes.

Pour cela, deux classifieurs ont été utilisés, l'analyse discriminante linéaire et la régression logistique, ainsi qu'une méthode de construction d'un score d'ensemble qui a été définie dans Duarte et al (2018a). Les méthodes d'ensemble, en construisant une collection de prédicteurs "de base" (en faisant varier l'échantillon initial, les variables sélectionnées, la méthode de régression ou de classification utilisée...) puis en agrégeant leurs prédictions, permettent souvent d'obtenir de meilleurs résultats que les prédicteurs individuels (si ceux-ci sont relativement bons et suffisamment différents, Genuer et Poggi 2017).

Nous avons tout d'abord défini et démontré la convergence de plusieurs types de processus d'approximation stochastique pour actualiser en ligne les paramètres d'une fonction de régression linéaire (Duarte *et al.* 2018b) ou logistique (Lalloué *et al.* 2019b) et montré l'intérêt d'utiliser des données standardisées en ligne plutôt que des données brutes, en particulier pour éviter une explosion numérique.

Le principe général de construction d'un score en ligne a ensuite été présenté dans Lalloué *et al.* (2019a).

Pour terminer cette étude, nous avons implémenté en R la construction de ce score en ligne et en avons testé empiriquement la convergence sur plusieurs jeux de données, en utilisant pour chaque classifieur plusieurs processus d'approximation stochastique et en comparant la précision des estimations obtenues. Nous présentons ici certains de ces résultats empiriques.

# 2 Expérimentations

## 2.1 Données et score "batch" de référence

Cinq jeux de données ont été utilisés : quatre disponibles sur Internet et un dérivé de l'étude EPHESUS (Pitt 2003), tous déjà utilisés pour tester les performances de processus de gradient stochastique (Duarte *et al.* 2018b, Lalloué *et al.* 2019b). `Twonorm` et `Ringnorm` sont des jeux de données simulées avec des variables homogènes (Breiman 1996). `Quantum` contient des données observées réelles, sans valeurs aberrantes et dont la plupart des variables sont sur la même échelle. `Adult2` et `HOSPHF30D` contiennent des données observées réelles, avec des valeurs aberrantes et des variables de différents types et sur différentes échelles. La table 1 résume ces données. Les détails des pré-traitements sont donnés dans Lalloué *et al.* (2019b)

Pour chaque jeu de données, on construit un score d'ensemble de référence en utilisant la méthode définie dans Duarte *et al.* (2018a) avec les paramètres suivants :

Table 1: Description of the datasets.

| **Dataset name** | $N_a$ | $N$ | $p_a$ | $p$ | **Source** |
|---|---|---|---|---|---|
| **Twonorm** | 7400 | 7400 | 20 | 20 | www.cs.toronto.edu/ delve/data/datasets.html |
| **Ringnorm** | 7400 | 7400 | 20 | 20 | www.cs.toronto.edu/ delve/data/datasets.html |
| **Quantum** | 50000 | 15798 | 78 | 12 | dérivé de www.osmot.cs.cornell.edu/kddcup |
| **Adult2** | 45222 | 45222 | 14 | 38 | dérivé de www.cs.toronto.edu/ delve/data/datasets.html |
| **HOSPHF30D** | 21382 | 21382 | 29 | 13 | dérivé de EPHESUS study |

$N_a$: nombre d'observations disponibles; $N$: nombre d'observations sélectionnées; $p_a$: nombre de paramètres disponibles; $p$: nombre de paramètres sélectionnés.

- Deux règles de classification sont utilisées : analyse discriminante linéaire (LDA) et régression logistique (LR).
- 100 échantillons bootstrap sont générés pour les deux règles.
- Toutes les variables disponibles sont incluses (des expérimentations avec des modalités de sélection et des nombres différents de variables tirées au sort sont en cours).
- Pour chaque règle de classification, les 100 prédicteurs associés sont agrégés par moyenne arithmétique puis les coefficients sont normalisés de manière à ce que le score varie entre 0 et 100.
- L'agrégation entre les deux scores synthétiques $S_1$ (LDA) et $S_2$ (LR) est faite par combinaison convexe : $S = \lambda S_1 + (1-\lambda)S_2$ avec ici $\lambda = 0.5$ (ultérieurement, une valeur optimale de $\lambda$ pourra être déterminée).

Le score ainsi obtenu pour chaque jeu de données est utilisé comme "gold standard" pour évaluer la convergence des processus en ligne testés.

A partir de chaque jeu de données, un flux de données est simulé en effectuant à chaque étape un tirage au hasard avec remise de 100 nouvelles observations. Les scores en ligne sont alors construits et mis à jour à partir de ces flux.

## 2.2 Processus

**Types de processus**

Les processus stochastiques $(X_n)$ utilisés sont de trois types différents :

- gradient stochastique "classique" (notation C$___$). À l'étape $n$, card $I_n = m_n$ observations $(Z_j, S_j)$ sont prises en compte et on calcule récursivement :

$$X_{n+1} = X_n - a_n \frac{1}{m_n} \sum_{j \in I_n} Z_j \left( h(Z_j' X_n) - S_j \right) ;$$

avec $Z_j$ vecteur des variables explicatives ; $S_j \in \{0,1\}$ ; $h(u) = u$ pour la LDA, $h(u) = \frac{e^u}{1+e^u}$ pour la LR.

- gradient stochastique "moyennisé" (notation A␣␣␣) : $\overline{X}_{n+1} = \frac{1}{n+1}\sum_{i=1}^{n+1} X_i$.
- uniquement dans le cas de la LDA : processus prenant en compte à chaque étape toutes les observations $(Z_j, S_j)$ jusqu'à cette étape, $j \in I_1 \cup ... \cup I_n$ (mention finale "all")(Duarte *et al.* 2018b).

Dans tous les cas, les variables explicatives sont standardisées en ligne (notation ␣S␣) : le principe et l'intérêt pratique de cette méthode pour éviter des explosions numériques ont été montrés dans Duarte *et al.* (2018b) et dans Lalloué *et al.* (2019b). En effet, pour certains jeux de données (`Adult2`, `HOSPHF30D`) les processus avec données brutes conduisent à une explosion numérique, contrairement à ceux avec données standardisées en ligne.

**Choix du pas**

Le pas $a_n$ peut être :

- continûment décroissant : $a_n = \frac{c}{(b+n^\alpha)}$ (notation ␣␣V).
- constant : $a_n = 1/p$ (avec $p$ le nombre de variables explicatives) (notation ␣␣C).
- constant par paliers (Bach 2014) : $a_n = \frac{c}{(b+\lfloor \frac{n}{\tau} \rfloor)^\alpha}$ ($\lfloor . \rfloor$ est la partie entière, $\tau$ la taille des paliers) (notation ␣␣P).

On prend dans tous les cas $\alpha = 2/3$, $b = 1$ et $c = 1$.

**Notation des processus**

Dans un couple de processus, le premier est celui utilisé pour la LDA, le second celui pour la LR.

Par exemple, AS100Call_AS100P200 désigne le couple formé pour la LDA d'un processus moyennisé (A) avec données standardisées en ligne (S), 100 nouvelles observations par étape (100), à pas constant(C) et prenant en compte toutes les observations jusqu'à l'étape en cours (all) ; et pour la LR d'un processus moyennisé (A) avec données standardisées en ligne (S), 100 nouvelles observations par étape (100) et à pas constant par paliers de taille 200 (P200).

**Processus testés**

Les six couples de processus testés font partie de ceux ayant eu les meilleures performances lors des études dédiées à la LDA en ligne (Duarte *et al.* 2018b) et à la régression logistique en ligne (Lalloué *et al.* 2019b), ou sont des processus "classiques" habituellement utilisés (mis à part la standardisation en ligne des données).

On utilise 100 nouvelles observations par étape. Chaque processus a été appliqué sur chacun des flux issus des jeux de données, avec un nombre d'observations utilisées de $100N$, correspondant à $N$ itérations.

**Critère de convergence**

On a utilisé comme critère de convergence la différence relative des normes $\frac{\|\theta^b - \hat{\theta}_N\|}{\|\theta^b\|}$ entre le vecteur $\theta^b$ des coefficients obtenus pour le score "batch" et le vecteur $\hat{\theta}_N$ des coefficients estimés par un processus après $N$ itérations. On considère qu'il y a eu convergence lorsque la valeur de ce critère est inférieure au seuil arbitraire de 0.05.

## 2.3 Résultats

Trois résultats sont comparés pour chaque couple de processus : la valeur du critère sur les coefficients normalisés (modifiés pour que le score varie entre 0 et 100) et standardisés (divisés par l'écart-type de la variable associée) pour le score synthétique $S_1$ obtenu par l'agrégation des LDA, celle pour le score synthétique $S_2$ obtenu par agrégation des régressions logistiques, et celle pour le score final $S$ (table 2).

Table 2: Différences relatives des normes après $100N$ observations utilisées

| **Processus** | | **Twonorm** | **Ringnorm** | **Quantum** | **Adult2** | **HOSPHF30D** |
|---|---|---|---|---|---|---|
| **CS100V _CS100V** | $S_1$ | 0.0010* | 0.0020* | 0.0073* | 0.0076* | 0.0165* |
| | $S_2$ | 0.0033* | 0.0009* | 0.0168* | 0.1002 | 0.0566 |
| | $S$ | **0.0015*** | **0.0014*** | **0.0083*** | **0.0414*** | **0.0289*** |
| **AS100P50 _AS100P50** | $S_1$ | 0.0006* | 0.0007* | 0.0027* | 2.7560 | 0.0176* |
| | $S_2$ | 0.0006* | 0.0007* | 0.0032* | 0.0346* | 0.0203* |
| | $S$ | **0.0005*** | **0.0007*** | **0.0029*** | **1.6968** | **0.0192*** |
| **AS100C _AS100P200** | $S_1$ | 0.0006* | 0.0007* | 0.0028* | 0.0066* | 0.0165* |
| | $S_2$ | 0.0007* | 0.0007* | 0.0033* | 0.0069* | 0.0206* |
| | $S$ | **0.0006*** | **0.0007*** | **0.0030*** | **0.0067*** | **0.0190*** |
| **CS100Vall _CS100V** | $S_1$ | 0.0005* | 0.0006* | 0.0033* | 0.0287* | 0.0153* |
| | $S_2$ | 0.0033* | 0.0009* | 0.0168* | 0.1002 | 0.0566 |
| | $S$ | **0.0017*** | **0.0007*** | **0.0090*** | **0.0281*** | **0.0290*** |
| **AS100P50all _AS100P50** | $S_1$ | 0.0006* | 0.0007* | 0.0046* | 0.0100* | 0.0060* |
| | $S_2$ | 0.0006* | 0.0007* | 0.0032* | 0.0346* | 0.0203* |
| | $S$ | **0.0005*** | **0.0007*** | **0.0039*** | **0.0193*** | **0.0147*** |
| **AS100Call _AS100P200** | $S_1$ | 0.0006* | 0.0007* | 0.0046* | 0.0153* | 0.0060* |
| | $S_2$ | 0.0007* | 0.0007* | 0.0033* | 0.0069* | 0.0206* |
| | $S$ | **0.0005*** | **0.0007*** | **0.0039*** | **0.0120*** | **0.0149*** |

* marque les valeurs du critère $< 0.05$.

Première abréviation : processus pour la LDA; Deuxième : processus pour la régression logistique.

*Type de processus :* C pour SGD classique, A pour ASGD.

*Données :* R pour brutes, S pour standardisées en ligne (*1er nombre* : nombre de nouvelles observations par étape).

*Pas :* V pour variable, C pour constant, P pour constant par palier (*2e nombre* : taille des paliers).

On constate que, pour tous les couples de processus testés, le score en ligne final $S$ est très proche du score de référence "batch" sur quatre des cinq jeux de données testés. Seul le couple de processus AS100P50_AS100P50 appliqué au jeu de données `Adult` ne converge pas en $N$ itérations. Dans la plupart des cas, la valeur du critère pour le score final $S$ est comprise entre celles des deux scores intermédiaires. Ceci conduit, pour deux couples de processus (CS100V_CS100V et CS100Vall_CS100V) appliqués à `Adult2` et `HOSPHF30D`, à

une convergence du score final alors que le score intermédiaire $S_2$ n'a pas encore convergé. Si l'on range les couples de processus du plus performant au moins performant pour chaque jeu de données puis qu'on calcule le rang moyen sur l'ensemble des jeux de données, le meilleur couple de processus est AS100P50all_AS100P50.

# 3 Conclusion

Nous avons mis au point un programme de construction d'un score en ligne en associant des processus avec données standardisées en ligne dont la convergence a déjà été établie théoriquement. On observe empiriquement la convergence de ce score d'ensemble en ligne vers le score "batch" avec plusieurs processus, ainsi que la supériorité de certains choix : processus moyennisés et pas constant par paliers notamment. On a inclus ici toutes les variables disponibles, le score est ainsi obtenu par bagging et agrégation de deux méthodes. D'autres expérimentations ont été effectuées avec des modalités de sélection aléatoire des variables.

# Bibliographie


Bach, F. (2014). Adaptivity of averaged stochastic gradient descent to local strong convexity for logistic regression. *Journal of Machine Learning Research* 15, pp. 595-627.

Breiman, L. (1996). Bias, variance, and arcing classifiers. *Technical Report 460*, Departement of Statistics, University of California, Berkeley.

Duarte, K., Monnez J.-M. and Albuisson E. (2018a). Methodology for Constructing a Short-Term Event Risk Score in Heart Failure Patients. *Appl Math*, 09(08):954-74.

Duarte, K., Monnez, J.-M. and Albuisson, E. (2018b). Sequential linear regression with online standardized data, *PloS One*, 13 (1) e0191186.

Genuer, R. and Poggi, J.-M. (2017). Arbres CART et Forêts aléatoires, Importance et sélection de variables, *hal-01387654*.

Lalloué, B., Monnez, J.-M and Albuisson, E. (2019a). Actualisation en ligne d'un score d'ensemble. *51e Journées de Statistique. hal-02152352.*

Lalloué, B., Monnez, J.-M and Albuisson, E. (2019b). Streaming constrained binary logistic regression with online standardized data, *hal-02156324*.

Oza, N.C. and Russell, S.J. (2001). Online Bagging and Boosting. In: *Proceedings of the Eighth International Workshop on Artificial Intelligence and Statistics, AISTATS 2001.*

Pitt, B., Remme, W., Zannad, F., Neaton, J., Martinez, F., Roniker, B., Bittman, R., Hurley, S., Kleiman, S., and Gatlin, M. (2003). Eplerenone, a selective aldosterone blocker, in patients with left ventricular dysfunction after myocardial infarction. *New England Journal of Medicine*, 348, pp. 1309-1321.