Streaming constrained binary logistic regression with online standardized data. Application to scoring heart failure

Benoît Lalloué, Jean-Marie Monnez, Eliane Albuisson

Streaming constrained binary logistic regression with online standardized data. Application to scoring heart failure.

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

1 Université de Lorraine, CNRS, Inria[1], IECL[2], F-54000 Nancy, France

2 CHRU Nancy, INSERM, Université de Lorraine, CIC[3], Plurithématique, F-54000 Nancy, France

3 Université de Lorraine, CNRS, IECL, F-54000 Nancy, France

4 BIOBASE, Pôle S2R, CHRU de Nancy, Vandœuvre-lès-Nancy, France

5 Faculté de Médecine, InSciDenS, Vandœuvre-lès-Nancy, France

* *benoit.lalloue@univ-lorraine.fr* ** *jean-marie.monnez@univ-lorraine.fr*

*** *eliane.albuisson@univ-lorraine.fr*

Abstract

We study a stochastic gradient algorithm for performing online a constrained binary logistic regression in the case of streaming or massive data. Assuming that observed data are realizations of a random vector, these data are standardized online in particular to avoid a numerical explosion or when a shrinkage method such as LASSO is used. We prove the almost sure convergence of a variable step-size constrained stochastic gradient process with averaging when a varying

[1]Inria, Project-team BIGS
[2]Institut Elie Cartan de Lorraine, BP 239, F-54506 Vandœuvre-lès-Nancy Cedex, France
[3]Centre d'Investigation Clinique

number of new data is introduced at each step. 24 stochastic approximation processes are compared on real or simulated datasets, classical processes with raw data, processes with online standardized data, with or without averaging and with variable or piecewise constant step-sizes. The best results are obtained by processes with online standardized data, with averaging and piecewise constant step-sizes. This can be used to update online an event rate score in heart failure patients.

1 Introduction

Three types of methods can be used to analyze very large datasets. First, subsampling i.e. analyzing only a subset of data to approximate results that would be obtained on the whole set. Second, partitioning the dataset in subsets, performing an analysis on each subset, then aggregating the results of all analyses when it is possible. Third, online learning which proceeds in successive steps, the results of the analysis being updated at each step taking into account a batch of new data.

The third type of method is particularly adapted to data streams. Data stream online analysis concerns data that arrive continuously such as process control data, web data, telecommunication data, medical data, financial data.... Recursive stochastic algorithms can be used for observations arriving sequentially to estimate for example parameters of a linear regression model [1] or principal components of a factorial analysis [2] or centres of classes in non-hierarchical clustering [3], whose estimations are updated by each new arriving data batch. When using such recursive processes, it is not necessary to store the data and, due to the relative simplicity of the computation involved, much more data than with classical methods can be taken into account during the same period of time. For very large datasets, the capacities of statistical learning methods can be

limited by the computing time. Recursive algorithms can also be used profitably in this context by randomly drawing at each step a data batch from the dataset.

Why use online standardized data (each continuous variable is standardized with respect to the estimations at the current step of its expectation and its standard deviation computed online) and a constrained process?

We have studied in [1] the sequential least square multidimensional linear regression using a stochastic approximation process, particularly in the case of a data stream. We have proved the convergence of three processes with online standardized data instead of raw data used in particular *to avoid the phenomenon of numerical explosion* that can be encountered [4]. The experiments conducted have shown better performance of processes with online standardized data compared to those with raw data. We use in the present study the same approach in the case of logistic regression and prove the convergence of a process of the Robbins-Monro type [5] with a varying number of new observations introduced at each step and a variable step-size. Moreover, *when using a shrinkage method such as LASSO or ridge*, we have first to standardize the explanatory variables. In the case of a data stream, when the mathematical expectation and the variance of each variable are a priori unknown, these variables can be standardized online and a process of the same type can be used but with a projection at each step on the convex set defined by the constraint on the parameters of the regression function. More generally *this type of process can be used for any convex set,* for example if it is imposed that the parameters associated with the explanatory variables are positive. Finally we can consider a case where at step n of the process, when introducing new data, *explanatory variables have their expectation and their variance that may depend on n or on the values of controlled variables* according to a specific model and a logistic regression model with standardized explanatory variables is defined. Assuming that we can estimate online the

expectation and the variance of these variables, we can use the same type of process to estimate the parameters of the logistic regression function. Note that, when the only objective is to avoid a numerical explosion without use of a shrinkage method or online estimation of distribution parameters evolving in time, we can use a pseudo-standardization of the data after a certain step n_0, standardizing the explanatory vatiables with respect to the estimations obtained at step n_0 of their expectation and their standard deviation. This reduces the computing time since the estimations of expectations and variances are no more computed online after the sep n_0.

A suitable choice of step-size is often crucial for obtaining good performance of a stochastic gradient process. If the step-size is too small, the convergence will be slower. Conversely, if the step-size is too large, a numerical explosion may occur during the first iterations. We do not use here a constant step-size stochastic gradient process with averaging studied in particular by Bach and Moulines [6] who have shown that such a process does not converge to the true value of the parameter because the gradient of the loss function is not linear in the case of logistic regression and who have defined other processes with a Newton-approximation scheme. But we use in our experiments *a piecewise constant step-size* as suggested in [7] in order that the step-size does not decrease too quickly and reduces the speed of convergence.

We also consider *an averaged stochastic gradient process* which may be trivially computed online.

Section 2 is devoted to the formulation of the problem, Section 3 to the definition of the stochastic gradient process, Section 4 to a theorem of almost sure (a.s.) convergence of this process and its proof, Section 5 to results of experiments where we have compared processes with raw data and with online standardized data, Section 6 to the presentation of an application to online

updating of an event score in heart failure patients defined by an ensemble method [8] that we actually study.

2 Formulation of the problem

Suppose that observed data are realizations of a random vector $(R^1, ..., R^p, S)$ in $\mathbb{R}^p \times \{0, 1\}$.

Let A' be the transpose of a matrix A and:

R the random column vector $\left(R^1 .....R^p\ 1\right)'$,

m the random column vector $\left(E\left[R^1\right] .....E\left[R^p\right]\ 0\right)'$,

R^c the random column vector $R - m$, r^c a realization of R^c,

σ^k the standard deviation of R^k, $k = 1, ..., p$,

Γ the diagonal $(p+1, p+1)$ matrix whose diagonal elements are $\frac{1}{\sigma^1}, ..., \frac{1}{\sigma^p}, 1$, taking by convention $\sigma^k = 1$ for a discrete variable,

$Z = \Gamma R^c$, whose continuous components are standardized, $z = \Gamma r^c$ a realization of Z,

$\theta = \left(\theta^1 .....\theta^P \theta^{P+1}\right)'$ a $(p+1, 1)$ column vector of real parameters,

Consider the logistic model with standardized covariates:

$$P\left(S = s \ \mid R = r\right) = f(s; z, \theta) = \left(\frac{e^{z'\theta}}{1 + e^{z'\theta}}\right)^s \left(\frac{1}{1 + e^{z'\theta}}\right)^{1-s} = \frac{e^{z'\theta s}}{1 + e^{z'\theta}}.$$

$E\left[S \mid R\right] = h\left(Z'\theta\right)$ with $h(u) = \frac{e^u}{1+e^u} = \frac{1}{1+e^{-u}}$.

Define the loss function -ln $f\left(s; z, x\right) = -z'xs + \ln\left(1 + e^{z'x}\right)$. The cost function

$$F(x) = -E\left[\ln f\left(S; Z, x\right)\right] = E\left[-Z'xS + \ln\left(1 + e^{Z'x}\right)\right]$$

has θ for unique minimizer since it is a convex function with positive hessian

$$F''(x) = E\left[ZZ'\frac{e^{Z'x}}{(1+e^{Z'x})^2}\right].$$

Note that there is no uniform strictly positive lower-bound on $F''(x)$, then F is not strongly convex [6,9], unless restricted to a convex set $\{\|x\| \leq c\}$ with $\|Z\|$ uniformly bounded and no linear relation between the components of Z.

θ is the unique solution of:

$$F'(x) = E\left[-ZS + \frac{Ze^{Z'x}}{1+e^{Z'x}}\right] = E\left[Z\left(h\left(Z'x\right) - S\right)\right] = 0.$$

The purpose of this study is to recursively estimate θ using a stochastic gradient algorithm with online standardized data.

Note that if θ_0 is the column vector of the parameters of the logistic regression function of S with respect to R, $f(s;z,\theta) = f_0(s;r,\theta_0) = \frac{e^{r'\theta_0 s}}{1+e^{r'\theta_0}}$,

$$\theta_0^j = \frac{\theta^j}{\sigma^j}, j = 1,...,p, \theta_0^{p+1} = \theta^{p+1} - m^1\frac{\theta^1}{\sigma^1} - ... - m^p\frac{\theta^p}{\sigma^p}$$

$$\text{with } m^j = E[R^j], j = 1,...,p \Leftrightarrow \theta_0 = \begin{pmatrix} \frac{1}{\sigma^1} & & & \\ & \ddots & & \\ & & & \frac{1}{\sigma^p} \\ -\frac{m^1}{\sigma^1} & \cdots & -\frac{m^p}{\sigma^p} & 1 \end{pmatrix}\theta.$$

3 Definition of a stochastic gradient process

Let $\left((R_n^1,.....,R_n^p,S), n \geqslant 1\right)$ be an i.i.d. sample of $(R^1,.....,R^p,S)$. Let, for $n \geqslant 1$:

R_n be the random column vector $\left(R_n^1.....R_n^p\ 1\right)'$,

R_n^c the random column vector $R_n - m$, $Z_n = \Gamma R_n^c$,

for $k = 1,...,p$, $\overline{R}_n^k$ the mean of the sample $\left(R_1^k,...,R_n^k\right)$ of R^k and

$\left(V_n^k\right)^2 = \frac{1}{n}\sum_{i=1}^{n}\left(R_i^k - \overline{R_n^k}\right)^2$ its variance, both recursively computed,

$\overline{R}_n$ the random column vector $\left(\overline{R_n^1}.....\overline{R_n^p}\ 0\right)'$ and Γ_n the $(p+1, p+1)$ diagonal matrix with diagonal elements

$$\frac{1}{\sqrt{\frac{n}{n-1}}V_n^1}, ..., \frac{1}{\sqrt{\frac{n}{n-1}}V_n^p}, 1.$$

Suppose that m_n observations (R_i, S_i) are taken into account at step n of the following defined process.

Let $\mu_n = \sum_{i=1}^{n} m_i, I_n = \{\mu_{n-1}+1, ..., \mu_n\}$, $\widehat{R}_n = \overline{R}_{\mu_n}$, $\widehat{\Gamma}_n = \Gamma_{\mu_n}$ and

$$\text{for } j \in I_n, \widetilde{Z}_j = \widehat{\Gamma}_{n-1}\left(R_j - \widehat{R}_{n-1}\right).$$

For $k = 1, ..., p$, each component R_j^k of R_j is pseudo-standardized with respect to the empirical mean $\widehat{R}_{n-1}^k$ and to the empirical estimation of σ^k, $\sqrt{\frac{n}{n-1}}V_{\mu_{n-1}}^k$. Note that:

$$\widetilde{Z}_j = \widehat{\Gamma}_{n-1}\left(R_j^c - \widehat{R}_{n-1}^c\right), \text{ with } \widehat{R}_{n-1}^c = \widehat{R}_{n-1} - m.$$

Suppose that θ is constrained to belong to a convex subset K of $\mathbb{R}^{p+1}$. Let Π be the projection operator on K.

Recursively define the stochastic approximation processes (X_n) and $\left(\overline{X}_n\right)$ in $\mathbb{R}^{p+1}$:

$$\begin{aligned} X_{n+1} &= \Pi\left(X_n - a_n \frac{1}{m_n}\sum_{j\in I_n}\widetilde{Z}_j\left(h\left(\widetilde{Z}_j' X_n\right) - S_j\right)\right), \\ \overline{X}_{n+1} &= \frac{1}{n+1}\sum_{1}^{n+1} X_i = \overline{X}_n - \frac{1}{n+1}\left(\overline{X}_n - X_{n+1}\right). \end{aligned}$$

Almost sure convergence of (X_n) and $\left(\overline{X}_n\right)$ to θ is proved in the next section.

4 Almost sure convergence of the process

Make the following assumptions:

(H1a) There is no affine relation between the components of R.

(H1b) The moments of order 4 of R exist.

(H2) $a_n > 0$, $\sum_{n=1}^{\infty} a_n = \infty$, $\sum_{n=1}^{\infty} \frac{a_n}{\sqrt{n}} < \infty$, $\sum_{n=1}^{\infty} a_n^2 < \infty$.

Theorem 1 *Suppose H1a,b and H2 hold. Then* (X_n) *and* $\left(\overline{X}_n\right)$ *converge to* θ *a.s.*

Let us state the Robbins-Siegmund lemma [10] and another lemma [1] used in the proof.

Lemma 2 *Let* (Ω, A, P) *be a probability space and* (T_n) *a non-decreasing sequence of sub-*σ*-fields of* A*. Suppose for all* n*,* z_n*,* α_n*,* β_n *and* γ_n *are four integrable non-negative* T_n*-measurable random variables defined on* (Ω, A, P) *such that:*

$$E\left[z_{n+1}|T_n\right] \leq z_n\left(1+\alpha_n\right) + \beta_n - \gamma_n \ a.s.$$

Then, in the set $\left\{\sum_{n=1}^{\infty} \alpha_n < \infty, \sum_{n=1}^{\infty} \beta_n < \infty\right\}$*,* (z_n) *converges to a finite random variable and* $\sum_{n=1}^{\infty} \gamma_n < \infty$ *a.s.*

Lemma 3 *Suppose H1b holds and* $a_n > 0$, $\sum_{n=1}^{\infty} \frac{a_n}{\sqrt{n}} < \infty$*. Then:*

$\sum_{n=1}^{\infty} a_n \left\|\widehat{R}_{n-1}^c\right\| < \infty$ *and* $\sum_{n=1}^{\infty} a_n \left\|\widehat{\Gamma}_{n-1} - \Gamma\right\| < \infty$ *a.s.*

Proof

The usual Euclidean norm in $\mathbb{R}^{p+1}$ and the spectral norm for matrices are used in this proof.

Part 1

Let T_n be the σ-field generated by the events before time n: $X_1, ..., X_n$ are T_n-measurable, as $\widehat{R}_{n-1}$ and $\widehat{\Gamma}_{n-1}$.

$$\begin{aligned}\|X_{n+1}-\theta\| &= \left\|\Pi\left(X_n - a_n\frac{1}{m_n}\sum_{j\in I_n}\widetilde{Z}_j\left(h\left(\widetilde{Z}_j'X_n\right)-S_j\right)\right)-\Pi\theta\right\| \\ &\leq \left\|X_n - a_n\frac{1}{m_n}\sum_{j\in I_n}\widetilde{Z}_j\left(h\left(\widetilde{Z}_j'X_n\right)-S_j\right)-\theta\right\|.\end{aligned}$$

Taking the conditional expectation with respect to T_n gives a.s.:

$$\begin{aligned}E\left[\|X_{n+1}-\theta\|^2 \mid T_n\right] \leq\ & \|X_n-\theta\|^2 \\ & -2a_n\left\langle X_n-\theta, \frac{1}{m_n}\sum_{j\in I_n}E\left[\widetilde{Z}_j\left(h\left(\widetilde{Z}_j'X_n\right)-S_j\right)\mid T_n\right]\right\rangle \\ & +a_n^2\ E\left[\left\|\frac{1}{m_n}\sum_{j\in I_n}\widetilde{Z}_j\left(h\left(\widetilde{Z}_j'X_n\right)-S_j\right)\right\|^2 \mid T_n\right] \text{ a.s.}\end{aligned}$$

Part 2 Decomposition of $E\left[\widetilde{Z}_j\left(h\left(\widetilde{Z}_j'X_n\right)-S_j\right)\mid T_n\right], j\in I_n$.

$E\left[\widetilde{Z}_j\left(h\left(\widetilde{Z}_j'X_n\right)-S_j\right)\mid T_n\right] = E\left[\widetilde{Z}_j\left(h\left(\widetilde{Z}_j'X_n\right)-E\left[S_j\mid R_j\right]\right)\mid T_n\right] - E\left[\widetilde{Z}_j\left(S_j-E\left[S_j\mid R_j\right]\right)\mid T_n\right].$

$$\begin{aligned}E\left[\widetilde{Z}_j\left(S_j-E\left[S_j\mid R_j\right]\right)\mid T_n\right] &= E\left[\widehat{\Gamma}_{n-1}\left(R_j-\widehat{R}_{n-1}\right)\left(S_j-E\left[S_j\mid R_j\right]\right)\ \mid T_n\right] \\ &= \widehat{\Gamma}_{n-1}E\left[R\left(S-E\left[S\mid R\right]\right)\right]-\widehat{\Gamma}_{n-1}\widehat{R}_{n-1}E\left[S-E\left[S\mid R\right]\right] \\ &= 0 \text{ a.s.}\end{aligned}$$

Then:

$$E\left[\widetilde{Z}_j\left(h\left(\widetilde{Z}_j'X_n\right)-S_j\right)\mid T_n\right] = E\left[\widetilde{Z}_j\left(h\left(\widetilde{Z}_j'X_n\right)-h\left(Z_j'\theta\right)\right)\mid T_n\right] \text{ a.s.}$$

Consider the decomposition $E\left[\widetilde{Z}_j\left(h\left(\widetilde{Z}_j'X_n\right)-h\left(Z_j'\theta\right)\right)\mid T_n\right] = E\left[Z_j\left(h\left(Z_j'X_n\right)-h\left(Z_j'\theta\right)\right)\mid T_n\right] + E\left[V_j\mid T_n\right]$ with

$$V_j = \left(\widetilde{Z}_j - Z_j\right)\left(h\left(Z_j' X_n\right) - h\left(Z_j'\theta\right)\right) + \widetilde{Z}_j\left(h\left(\widetilde{Z}_j' X_n\right) - h\left(Z_j' X_n\right)\right).$$

For $j \in I_n$, there exist ξ_j^1 and ξ_j^2 such that:

$$\begin{aligned}
h\left(\widetilde{Z}_j' X_n\right) &= h\left(\left(R_j^c - \widehat{R}_{n-1}^c\right)' \widehat{\Gamma}_{n-1} X_n\right) \\
&= h\left(\left(R_j^c - \widehat{R}_{n-1}^c\right)' \Gamma X_n\right) + \left(R_j^c - \widehat{R}_{n-1}^c\right)' \left(\widehat{\Gamma}_{n-1} - \Gamma\right) X_n h'\left(\xi_j^1\right) \\
&= h\left(Z_j' X_n - \widehat{R}_{n-1}^{c\prime} \Gamma X_n\right) + \left(R_j^c - \widehat{R}_{n-1}^c\right)' \left(\widehat{\Gamma}_{n-1} - \Gamma\right) X_n h'\left(\xi_j^1\right) \\
&= h\left(Z_j' X_n\right) - \widehat{R}_{n-1}^{c\prime} \Gamma X_n h'\left(\xi_j^2\right) + \left(R_j^c - \widehat{R}_{n-1}^c\right)' \left(\widehat{\Gamma}_{n-1} - \Gamma\right) X_n h'\left(\xi_j^1\right).
\end{aligned}$$

As $\widetilde{Z}_j - Z_j = \left(\widehat{\Gamma}_{n-1} - \Gamma\right) R_j^c - \widehat{\Gamma}_{n-1} \widehat{R}_{n-1}^c$, it follows that:

$$\begin{aligned}
V_j &= \left(\left(\widehat{\Gamma}_{n-1} - \Gamma\right) R_j^c - \widehat{\Gamma}_{n-1} \widehat{R}_{n-1}^c\right)\left(h\left(Z_j' X_n\right) - h\left(Z_j'\theta\right)\right) \\
&\quad + \widehat{\Gamma}_{n-1}\left(R_j^c - \widehat{R}_{n-1}^c\right)\left(-\widehat{R}_{n-1}^{c\prime} \Gamma X_n h'\left(\xi_j^2\right) + \left(R_j^c - \widehat{R}_{n-1}^c\right)' \left(\widehat{\Gamma}_{n-1} - \Gamma\right) X_n h'\left(\xi_j^1\right)\right).
\end{aligned}$$

Part 3 Application of Robbins-Siegmund lemma.

It follows from Part 1 and Part 2 that:

$$\begin{aligned}
E\left[\left\|X_{n+1} - \theta\right\|^2 \mid T_n\right] &\leq \left\|X_n - \theta\right\|^2 - 2a_n \frac{1}{m_n} \sum_{j \in I_n} \left\langle X_n - \theta, E\left[V_j \mid T_n\right]\right\rangle \\
&\quad + a_n^2 \; E\left[\left\|\frac{1}{m_n} \sum_{j \in I_n} \widetilde{Z}_j \left(h\left(\widetilde{Z}_j' X_n\right) - S_j\right)\right\|^2 \mid T_n\right] \\
&\quad - 2a_n \frac{1}{m_n} \sum_{j \in I_n} \left\langle X_n - \theta, E\left[Z_j \left(h\left(Z_j' X_n\right) - h\left(Z_j'\theta\right)\right) \mid T_n\right]\right\rangle \text{ a.s.}
\end{aligned}$$

(a) For $j \in I_n$, there exists $0 \leq \lambda_j \leq 1$ such that:

$$h\left(Z_j' X_n\right) - h\left(Z_j'\theta\right) = Z_j'\left(X_n - \theta\right) h'\left(\xi_j\right), \text{with } \xi_j = \lambda_j Z_j' X_n + \left(1 - \lambda_j\right) Z_j'\theta.$$

Then as h is an increasing function:

$$\left\langle X_n-\theta, E\left[Z_j\left(h\left(Z_j'X_n\right)-h\left(Z_j'\theta\right)\right)\mid T_n\right]\right\rangle = E\left[\left\|Z_j'\left(X_n-\theta\right)\right\|^2 h'\left(\xi_j\right)\mid T_n\right] \geqslant 0 \text{ a.s.}$$

(b) For $j\in I_n$, by definition of V_j, as $0<h(x)<1$ and $0<h'(x)\leq\frac{1}{4}$:

$$\begin{aligned} E\left[\|V_j\|\mid T_n\right] &\leq \left\|\widehat{\Gamma}_{n-1}-\Gamma\right\| E\left[\|R^c\|\right]+\left\|\widehat{\Gamma}_{n-1}\right\|\left\|\widehat{R}^c_{n-1}\right\| \\ &+\frac{1}{4}\left\|\widehat{\Gamma}_{n-1}\right\|\left(E\left[\|R^c\|\right]+\left\|\widehat{R}^c_{n-1}\right\|\right)\left\|\widehat{R}^c_{n-1}\right\|\|\Gamma\|\left(\|X_n-\theta\|+\|\theta\|\right) \\ &+\frac{1}{2}\left\|\widehat{\Gamma}_{n-1}\right\|\left(E\left[\|R^c\|^2\right]+\left\|\widehat{R}^c_{n-1}\right\|^2\right)\left\|\widehat{\Gamma}_{n-1}-\Gamma\right\|\left(\|X_n-\theta\|+\|\theta\|\right). \end{aligned}$$

As $\widehat{\Gamma}_{n-1}$ and $\widehat{R}^c_{n-1}$ are T_n-measurable and converge respectively to Γ and 0, as $\sum\limits_{n=1}^{\infty}a_n\left\|\widehat{R}^c_{n-1}\right\|<\infty$ and $\sum\limits_{n=1}^{\infty}a_n\left\|\widehat{\Gamma}_{n-1}-\Gamma\right\|<\infty$ a.s. by Lemma 3, it follows that there exist two non-negative T_n- measurable random variables D_n and E_n such that for $j\in I_n$:

$$\|E\left[V_j\mid T_n\right]\|\leq D_n\|X_n-\theta\|+E_n,\ \sum_{n=1}^{\infty}a_nD_n<\infty, \sum_{n=1}^{\infty}a_nE_n<\infty \text{ a.s.}$$

Then:

$$\begin{aligned} \left|\frac{1}{m_n}\sum_{j\in I_n}\left\langle X_n-\theta, E\left[V_j\mid Tn\right]\right\rangle\right| &\leq \|X_n-\theta\|\left(D_n\|X_n-\theta\|+E_n\right) \\ &\leq \left(D_n+E_n\right)\|X_n-\theta\|^2+E_n \text{ a.s.} \end{aligned}$$

(c) $$E\left[\left\|\widetilde{Z}_j\left(h\left(\widetilde{Z}_j'X_n\right)-S_j\right)\right\|^2\mid T_n\right]\leq E\left[\left\|\widetilde{Z}_j\right\|^2\mid T_n\right]$$

$$\leq E\left[\left\|\widehat{\Gamma}_{n-1}\left(R^c_j-\widehat{R}^c_{n-1}\right)\right\|^2\mid T_n\right]\leq 2\left\|\widehat{\Gamma}_{n-1}\right\|^2\left(E\left[\|R^c\|^2\right]+\left\|\widehat{R}^c_{n-1}\right\|^2\right) \text{ a.s.}$$

By H2, as $\widehat{\Gamma}_{n-1}$ and $\widehat{R}^c_{n-1}$ converge a.s. respectively to Γ and 0, we have:

$$\sum_{n=1}^{\infty} a_n^2 E\left[\left\|\frac{1}{m_n}\sum_{j\in I_n}\widetilde{Z}_j\left(h\left(\widetilde{Z}_j' X_n\right)-S_j\right)\right\|^2 \mid T_n\right] < \infty \text{ a.s.}$$

(d) Conclusion

$$\begin{aligned} E\left[\|X_{n+1}-\theta\|^2 \mid T_n\right] \;\leq\; & \|X_n-\theta\|^2(1+D_n+E_n)+2a_nE_n \\ & +a_n^2\; E\left[\left\|\frac{1}{m_n}\sum_{j\in I_n}\widetilde{Z}_j\left(h\left(\widetilde{Z}_j' X_n\right)-S_j\right)\right\|^2 \mid T_n\right] \\ & -2a_n\frac{1}{m_n}\sum_{j\in I_n}E\left[\left\|Z_j'(X_n-\theta)\right\|^2 h'(\xi_j)\mid T_n\right] \text{ a.s.} \end{aligned}$$

Applying Robbins-Siegmund lemma yields that there exists a non-negative random variable T such that a.s.:

$$\|X_n-\theta\|^2 \longrightarrow T \text{ and } \sum_{n=1}^{\infty} a_n\frac{1}{m_n}\sum_{j\in I_n}E\left[\left\|Z_j'(X_n-\theta)\right\|^2 h'(\xi_j)\mid T_n\right] < \infty.$$

Part 4 Prove that $T=0$ a.s.

Let ω be fixed belonging to the intersection of the convergence sets. The writing of ω will be omitted in the following.

Suppose $T\neq 0$. There exists $0<\epsilon<1$ such that $\epsilon < \|X_n-\theta\| < \frac{1}{\epsilon}$.

As for $j\in I_n$, $\xi_j = \lambda_j Z_j' X_n + (1-\lambda_j) Z_j'\theta = \lambda_j Z_j'(X_n-\theta) + Z_j'\theta$,

$|\xi_j| \leq \left\|R_j^c\right\| b$, with $b = \|\Gamma\|\left(\frac{1}{\epsilon}+\|\theta\|\right)$.

Remember that h' is an even function, decreasing for $x>0$, and $h'(x) \geqslant \frac{1}{4}e^{-x}$ for $x>0$. Then, $h'(\xi_j) \geqslant h'\left(\left\|R_j^c\right\| b\right) \geqslant \frac{1}{4}e^{-\left\|R_j^c\right\| b}$.

Therefore, denoting by $\lambda_{\min}(A)$ the lowest eigenvalue of a matrix A, we have for $j\in I_n$ as $Z_j = \Gamma R_j^c$:

$$E\left[\left\|Z_j'\left(X_n-\theta\right)\right\|^2 h'\left(\xi_j\right) \mid T_n\right] \geqslant \tfrac{1}{4}\left(X_n-\theta\right)' \Gamma E\left[R_j^c R_j^{c\prime} e^{-\left\|R_j^c\right\| b} \mid T_n\right] \Gamma\left(X_n-\theta\right)$$
$$\geqslant \tfrac{1}{4}\lambda_{\min}\left(E\left[R^c R^{c\prime} e^{-\|R^c\| b}\right]\right)\left\|\Gamma\left(X_n-\theta\right)\right\|^2$$
$$\geqslant \tfrac{1}{4}\lambda_{\min}\left(E\left[R^c e^{-\frac{1}{2}\|R^c\| b}\left(R^c e^{-\frac{1}{2}\|R^c\| b}\right)'\right]\right)\left(\lambda_{\min}\left(\Gamma\right)\right)^2 \epsilon^2.$$

The symmetric matrix $E\left[R^c R^{c\prime} e^{-\|R^c\| b}\right]$ is positive definite since by H1a there is no linear relation between the components of R^c, consequently between the components of $R^c e^{-\frac{1}{2}\|R^c\| b}$; its lowest eigenvalue is strictly positive. By H2, it follows that:

$$\begin{aligned} & \sum_{n=1}^{\infty} a_n \frac{1}{m_n} \sum_{j \in I_n} E\left[\left\|Z_j'\left(X_n-\theta\right)\right\|^2 h'\left(\xi_j\right) \mid T_n\right] \\ \geqslant \quad & \frac{1}{4}\lambda_{\min}\left(E\left[R^c R^{c\prime} e^{-\|R^c\| b}\right]\right)\left(\lambda_{\min}\left(\Gamma\right)\right)^2 \epsilon^2 \sum_{n=1}^{\infty} a_n = \infty. \end{aligned}$$

This is a contradiction as ω belongs to the convergence set of this series. Thus $T = 0$. We deduce immediately the convergence of $\left(\overline{X}_n\right)$ to θ.■

5 Experiments

24 stochastic approximation processes were compared, including classic stochastic gradient descent (SGD), averaged stochastic gradient descent (ASGD) with a piecewise constant step-size with different level sizes as suggested in [7], and the same processes but with online standardization of the data (Section 3). The processes studied and their respective parameters are described in Table 1.

5.1 Step-size

For processes with a variable step-size (processes C1 to C3 and SC1 to SC3), we have defined

$$a_n = \frac{c}{(b+n)^{\alpha}}$$

Table 1. Description of the processes.

Method type	Abbreviation	Type of data	Number of observations used at each step of the process	Step-size	Levels size	Use of the averaged process
Classic	C1	Raw data	1	Variable	-	No
	C2		10			
	C3		100			
ASGD with piecewise constant step-size	L11		1	Piecewise constant	50	Yes
	L12		10			
	L13		100			
	L21		1		100	
	L22		10			
	L23		100			
	L31		1		200	
	L32		10			
	L33		100			
Standardization 1	SC1	Online standardized data	1	Variable	-	No
	SC2		10			
	SC3		100			
Standardization 2	SL11		1	Piecewise constant	50	Yes
	SL12		10			
	SL13		100			
	SL21		1		100	
	SL22		10			
	SL23		100			
	SL31		1		200	
	SL32		10			
	SL33		100			

For processes with a piecewise constant step-size (processes L11 to L33 and SL11 to SL33), we have chosen

$$a_n = \frac{c}{(b + \lfloor \frac{n}{\tau} \rfloor)^\alpha}$$

where $\lfloor . \rfloor$ denotes the integer part and τ is the size of the levels. For both cases, we set $\alpha = 2/3$ (this value was suggested by Xu [11] in the case of linear regression), $b = 1$ and $c = 1$.

Bach and Moulines [6] have shown that averaged processes with constant step-size do not converge to the true value of the parameter in the case of logistic regression, therefore we have not tested this type of process.

5.2 Initialization

All processes were initialized with $X_1 = \underline{0}$. For processes with online standardization, a random sample of 1000 observations (drawn with replacement from the dataset) was used to compute a first estimation of the means and standard deviations of the explanatory variables before the beginning of the iterations. For averaged processes, the first 1000 iterations were used as a burn-in period and were not included in the computation of the average.

5.3 Convergence criteria

We used as "gold standard" the coefficients obtained by classical logistic regression (using R's `glm` function) on a dataset $((r_i^1, ..., r_i^p, s_i), i = 1, ..., N)$ to assess the convergence of the processes. Let θ^c be the vector of coefficients obtained with this method and $\hat{\theta}_{n+1}$ the estimated vector obtained by a tested process after n iterations.

$$\text{As } \theta_0 = \begin{pmatrix} \frac{1}{\sigma^1} & & & \\ & \ddots & & \\ & & \frac{1}{\sigma^p} & \\ -\frac{m^1}{\sigma^1} & \cdots & -\frac{m^p}{\sigma^p} & 1 \end{pmatrix} \theta, \hat{\theta}_{n+1} = \begin{pmatrix} \hat{\Gamma}_n(1,1) & & & \\ & \ddots & & \\ & & \hat{\Gamma}_n(p,p) & \\ -\hat{\Gamma}_n(1,1)\hat{r}_n^1 & \cdots & -\hat{\Gamma}_n(p,p)\hat{r}_n^p & 1 \end{pmatrix} \bar{x}_{n+1},$$

($\bar{x}_{n+1}$, realization of $\bar{X}_{n+1}$, is the estimation of θ at step n).

The cosine of the angle between θ^c and $\hat{\theta}_{n+1}$ was used as a convergence criterion:

$$cos(\theta^c, \hat{\theta}_{n+1}) = \frac{\theta^{c\prime}\hat{\theta}_{n+1}}{\|\theta^c\|\|\hat{\theta}_{n+1}\|}$$

The coefficient of correlation between the predictions obtained with the classical method and the process as well as the ratio $\frac{\hat{F}(\hat{\theta}_{n+1})-\hat{F}(\theta^c)}{\hat{F}(\theta^c)}$, $\hat{F}(\hat{\theta}_{n+1}) = \frac{1}{N}\sum_{i=1}^{N}\left(-r_i'\hat{\theta}_{n+1}s_i + ln(1+e^{r_i'\hat{\theta}_{n+1}})\right)$ being an estimation of the cost function F at $\hat{\theta}_{n+1}$, were also used as criteria (results not shown).

5.4 Datasets

The processes were tested on five datasets $((r_i^1, ..., r_i^p, s_i), i = 1, ..., N)$ available on the Internet and one dataset derived from the EPHESUS study [12], all already used to test the performance of stochastic approximation processes with online standardized data in the case of online linear regression [1]. `Twonorm`, `Ringnorm`, `Quantum` and `Adult` are commonly used to test classification methods (the first two were introduced by Breiman [13]). Table 2 summarizes these datasets. For each dataset, a data stream was simulated by randomly drawing a data batch at each step.

Table 2. 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	derived from `www.osmot.cs.cornell.edu/kddcup`
Adult2	45222	45222	14	38	derived from `www.cs.toronto.edu/~delve/data/datasets.html`
EEG	14980	14977	14	14	`https://archive.ics.uci.edu/ml/datasets/EEG+Eye+State`
HOSPHF30D	21382	21382	29	13	derived from EPHESUS study

N_a: number of available observations; N: number of selected observations; p_a: number of available parameters; p: number of selected parameters.

The following preprocessings were done on the data:

- `Twonorm` and `Ringnorm`: no preprocessing.
- `Quantum`: a stepwise variable selection (using AIC) was performed on the 6197 observations without any missing value. The dataset with complete observations for the 12 selected variables was used.

- `Adult2`: from the Adult dataset, modalities of several categorical variables were merged (in order to have a larger number of observations for each modality) and all categorical variables were then replaced by sets of binary variables, leading to a dataset with 38 variables.
- `EEG`: Three outliers were excluded.
- `HOSPHF30D`: 13 variables were selected using stepwise selection.

All processes were applied on all datasets for a fixed number of observations used and for a fixed processing time (the cumulative time to compute the process updates, excluding operations such as data sampling, data management, formatting and recording of results...). For each dataset and at each recording point (see below), processes were ranked from the best (highest cosine) to the worst (lowest cosine). The mean rank over all datasets was used to compare the processes at a given recorded point and globally.

Processing time to treat $10N$ observations and average number of observations used per second were also studied. Note that it is preferable to consider only the order of magnitude of these indicators, as CPU and memory usage by other applications were not controlled while the processes were running and could explain small differences.

Processes were implemented with the R 3.5.2 software (64bits version) and tested on a Windows 10 computer with an Intel Core i7-8650U CPU and 32Go of memory.

5.5 Comparison for a fixed number of observations

As in [1], the values of criteria for each process were recorded every N observations used, from $1N$ to $100N$. For the cosine criterion, results for $10N$ observations are shown in Table 3. Note that since the number of observations used at each

step differs from one process to another, the number of iterations is not the same for each process (e.g. to use $100N$ observations, C1 will run for $100N$ iterations whereas C3 will run for N iterations).

Table 3. Cosines for 10N observations used

Process	Twonorm	Ringnorm	Quantum	Adult	EEG	HOSPHF30D	Mean rank
C1	0.9991	0.9990	0.9020	Expl	Expl	Expl	20.0
C2	0.9979	0.9994	0.8155	Expl	Expl	Expl	22.8
C3	0.9964	0.9989	0.7023	Expl	Expl	Expl	26.8
L11	0.9993	0.9997	0.9952	Expl	Expl	Expl	15.8
L12	0.9997	0.9998	0.9849	Expl	Expl	Expl	16.2
L13	0.9995	0.9972	0.9566	Expl	Expl	Expl	25.5
L21	0.9991	0.9995	0.9971	Expl	Expl	Expl	15.7
L22	0.9997	0.9998	0.9906	Expl	Expl	Expl	15.3
L23	0.9994	0.9962	0.9745	Expl	Expl	Expl	24.2
L31	0.9991	0.9993	0.9988	Expl	Expl	Expl	16.7
L32	0.9997	0.9998	0.9928	Expl	Expl	Expl	16.2
L33	0.9992	0.9943	0.9836	Expl	Expl	Expl	25.2
SC1	0.9986	0.9993	0.9500	0.9974	-0.9968	0.9826	16.5
SC2	0.9972	0.9997	0.9575	0.9939	-0.9959	0.9548	17.0
SC3	0.9964	0.9999	0.9484	0.9892	0.9987	0.5511	15.7
SL11	0.9992	0.9998	0.9971	0.9964	0.9994	0.9707	9.0
SL12	0.9996	0.9998	0.9980	0.9993	0.9997	0.9833	4.0
SL13	0.9996	0.9984	0.9727	0.9988	0.9994	0.9843	11.2
SL21	0.9992	0.9998	0.9966	0.9882	0.9992	0.9695	11.2
SL22	0.9997	0.9998	0.9965	0.9987	0.9998	0.9827	5.2
SL23	0.9995	0.9973	0.9703	0.9993	0.9994	0.9893	11.7
SL31	0.9991	0.9997	0.9932	0.9815	0.9985	0.9643	14.3
SL32	0.9997	0.9998	0.9933	0.9968	1.0000	0.9813	6.7
SL33	0.9994	0.9950	0.9664	0.9993	0.9994	0.9727	12.5

Expl: numerical explosion

All tested processes using raw data had a numerical explosion for half of the datasets (especially datasets with real data and different types of variables). Over all datasets, the eight processes with the lowest mean rankings are averaged processes with online standardization and piecewise constant step-sizes, the best one with levels of size 50 and 10 new observations per step (SL12). The five processes with the worst mean rankings are processes on raw data and 100 new

observations at each step. These conclusions remain valid if we use $\frac{\hat{F}(\hat{\theta}_{n+1})-\hat{F}(\theta^c)}{\hat{F}(\theta^c)}$ as criterion instead of the cosine.

Processing times for $10N$ observations are shown in Table 4. For all processes, the processing time decreases as the number of observations used at each step increases (and therefore as the number of iterations decreases for a given total number of observation used). A process with online standardization has a 4 to 21 times longer processing time than its equivalent on raw data, the ratio increasing with the number of observations used at each step of the two processes. Nevertheless, we will see below that for a fixed processing time the best processes remain those with online standardization. Thus, the main factors affecting the processing time are the number of observations used at each step, the online standardization and the dataset used.

Note that if the estimation of the expectations and standard deviations is stopped after a certain step and a pseudo-standardization with respect to the last obtained estimations used afterwards, the processing times improve for all processes.

Table 4. Processing time to treat 10N observations (in seconds)

Process	Twonorm	Ringnorm	Quantum	Adult	EEG	HOSPHF30D
C1	2.04	1.94	4.22	Expl	Expl	Expl
C2	0.23	0.23	0.45	Expl	Expl	Expl
C3	0.05	0.04	0.05	Expl	Expl	Expl
L11	2.54	2.04	4.47	Expl	Expl	Expl
L12	0.24	0.19	0.60	Expl	Expl	Expl
L13	0.03	0.04	0.06	Expl	Expl	Expl
L21	2.30	2.03	4.10	Expl	Expl	Expl
L22	0.25	0.22	0.50	Expl	Expl	Expl
L23	0.04	0.03	0.07	Expl	Expl	Expl
L31	2.04	2.12	4.47	Expl	Expl	Expl
L32	0.27	0.27	0.53	Expl	Expl	Expl
L33	0.03	0.02	0.06	Expl	Expl	Expl
SC1	10.84	11.41	20.23	59.75	19.57	26.73
SC2	1.96	2.00	3.65	10.74	3.90	4.63
SC3	0.48	0.49	0.92	2.74	0.83	1.11
SL11	10.52	10.17	20.59	60.36	19.31	26.37
SL12	1.89	1.72	3.84	11.03	3.50	5.00
SL13	0.64	0.46	0.85	2.89	0.85	1.17
SL21	10.83	9.71	19.94	60.37	19.41	27.18
SL22	1.85	1.86	3.56	10.94	3.35	4.91
SL23	0.46	0.48	0.82	2.84	0.87	1.13
SL31	9.64	9.78	20.11	61.79	20.18	27.54
SL32	1.88	1.61	3.51	11.09	3.96	4.74
SL33	0.50	0.41	0.87	2.75	0.91	1.11

Expl: numerical explosion

5.6 Comparison for a varying number of observations

When studying the variation of the rankings with the number of observations used from N to $100N$(Fig. 1), there is instability in the rankings until about $25N$ observations, after which most processes are in a stable position. Over all the numbers of observations used, the best process appears to be the averaged process with online standardization with 100 new observations at each step and piecewise constant step-size with levels of size 50 (SL13), followed by the same process with levels of sizes 100 and 200 (SL23 and SL33). Processes with online standardization are constantly better than processes using raw data.

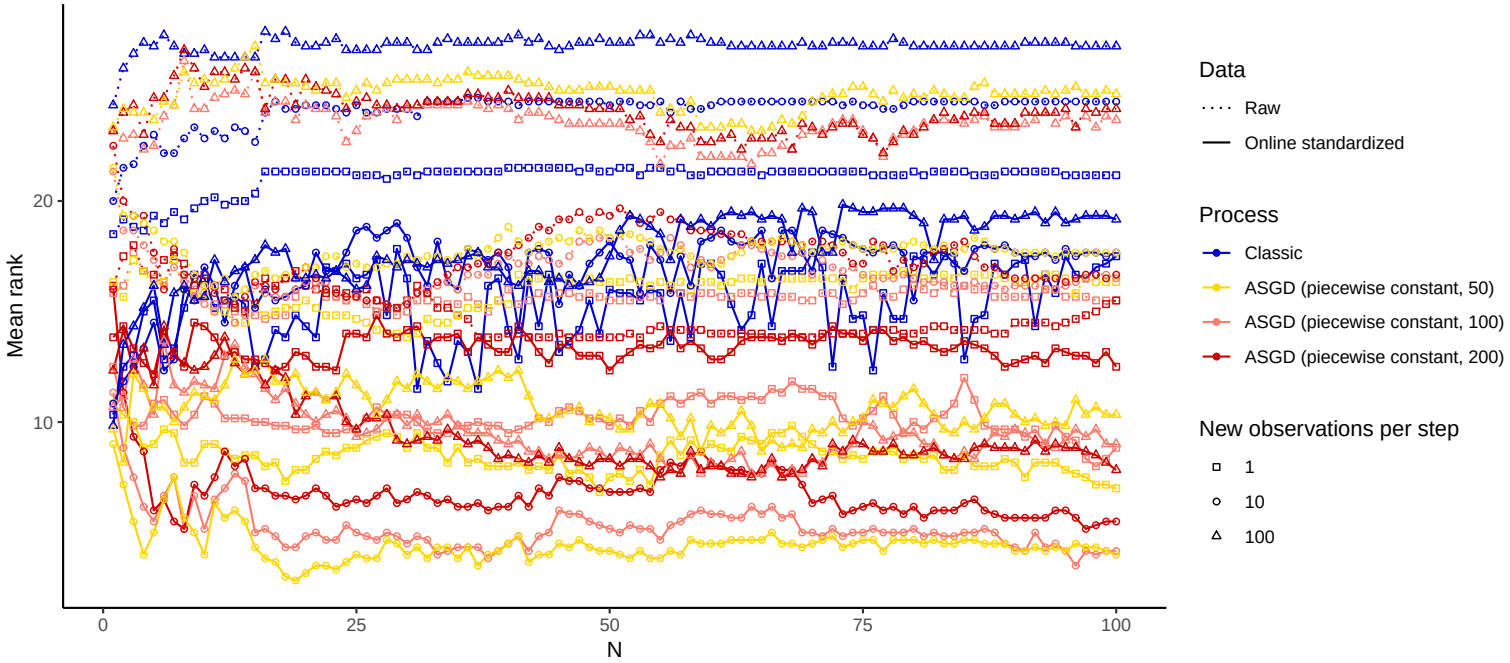

Fig 1. Evolution with the number of observations

5.7 Comparison for a fixed processing time

As in [1], the values of the criteria for each process were then recorded every second of processing time from 1 to 120s. For the cosine criterion, results for 60s observations are shown in Table 5.

Again, all tested processes using raw data had a numerical explosion for half of the datasets. Over all datasets, the six processes with the lowest mean rankings are averaged processes with online standardization and piecewise constant step-sizes, the best one with levels of size 200 and 100 new observations per step (SL33). These conclusions remain valid if we use $\frac{\hat{F}(\hat{\theta}_{n+1})-\hat{F}(\theta^c)}{\hat{F}(\theta^c)}$ as criterion with the exception that the best process uses levels of size 100 (SL23).

Average number of observations used per second for 60s of processing time are shown in Table 6. For all processes, the number of observations used by second increases with the number of observations used at each step. A process with

Table 5. Cosines for 1 minute of processing time

Process	Twonorm	Ringnorm	Quantum	Adult	EEG	HOSPHF30D	Mean rank
C1	0.9999	0.9999	0.9709	Expl	Expl	Expl	20.8
C2	1.0000	1.0000	0.9683	Expl	Expl	Expl	21.8
C3	1.0000	1.0000	0.9659	Expl	Expl	Expl	22.5
L11	1.0000	1.0000	0.9978	Expl	Expl	Expl	19.3
L12	1.0000	1.0000	0.9960	Expl	Expl	Expl	18.3
L13	1.0000	1.0000	0.9948	Expl	Expl	Expl	20.0
L21	1.0000	1.0000	0.9991	Expl	Expl	Expl	18.0
L22	1.0000	1.0000	0.9972	Expl	Expl	Expl	17.5
L23	1.0000	1.0000	0.9959	Expl	Expl	Expl	19.0
L31	1.0000	1.0000	0.9999	Expl	Expl	Expl	16.8
L32	1.0000	1.0000	0.9981	Expl	Expl	Expl	16.5
L33	1.0000	1.0000	0.9970	Expl	Expl	Expl	18.2
SC1	0.9997	0.9998	0.9987	0.9979	0.9988	0.9898	17.5
SC2	0.9996	1.0000	0.9989	0.9968	0.9994	0.9932	15.8
SC3	0.9994	1.0000	0.9992	0.9953	0.9993	0.9840	15.3
SL11	0.9999	1.0000	0.9959	0.9964	0.9993	0.9854	17.2
SL12	1.0000	1.0000	0.9999	0.9998	0.9997	0.9986	8.2
SL13	1.0000	1.0000	0.9999	0.9999	1.0000	0.9999	6.5
SL21	0.9999	1.0000	0.9948	0.9888	0.9992	0.9841	19.2
SL22	1.0000	1.0000	0.9999	0.9998	0.9996	0.9987	8.8
SL23	1.0000	1.0000	0.9999	0.9999	1.0000	0.9999	6.7
SL31	0.9999	0.9999	0.9934	0.9823	0.9987	0.9812	19.8
SL32	1.0000	1.0000	0.9999	0.9996	0.9996	0.9986	8.2
SL33	1.0000	1.0000	0.9999	1.0000	1.0000	0.9999	4.8

Expl: numerical explosion

online standardization treats 4 to 18 times less observations per second than its equivalent on raw data, the ratio increasing with the number of observations used at each step of the two processes. Thus, the main factors affecting the average number of observations used per second are the number of new observations used at each step and the online standardization.

Table 6. Average number of observations used per second for 60s of processing time

Process	Twonorm	Ringnorm	Quantum	Adult	EEG	HOSPHF30D
C1	35 657	26 185	32 210	Expl	Expl	Expl
C2	313 352	279 990	296 649	Expl	Expl	Expl
C3	2 162 220	1 773 088	1 995 762	Expl	Expl	Expl
L11	32 264	25 079	33 284	Expl	Expl	Expl
L12	298 995	273 666	314 852	Expl	Expl	Expl
L13	2 271 557	1 905 687	2 424 623	Expl	Expl	Expl
L21	26 591	27 205	34 679	Expl	Expl	Expl
L22	238 334	249 312	310 214	Expl	Expl	Expl
L23	1 787 677	1 850 917	2 399 433	Expl	Expl	Expl
L31	27 101	24 647	33 817	Expl	Expl	Expl
L32	230 269	264 637	315 603	Expl	Expl	Expl
L33	1 746 310	2 098 472	2 330 235	Expl	Expl	Expl
SC1	6 800	5 555	6 591	7 377	5 723	7 478
SC2	39 522	33 279	38 751	38 554	33 236	44 129
SC3	181 780	137 465	172 560	141 615	145 198	184 768
SL11	7 315	5 571	6 925	7 157	5 437	7 329
SL12	41 879	31 104	39 294	39 991	28 847	42 858
SL13	161 653	136 027	187 612	153 328	142 280	186 263
SL21	7 159	4 955	7 144	6 997	5 596	7 207
SL22	41 351	30 742	42 447	37 901	31 831	42 858
SL23	148 013	130 098	195 987	129 703	134 973	186 263
SL31	5 575	5 570	7 398	6 547	5 628	7 293
SL32	30 273	32 825	43 051	38 486	31 186	43 260
SL33	131 130	119 582	194 177	142 552	143 285	190 840

Expl: numerical explosion

5.8 Comparison for a varying processing time

When studying the evolution of the rankings with the processing time from 1 to 120s (Figure 2), two groups of processes appear clearly from the beginning and remain during all the studied period. The group with the worst rankings contains all processes using raw data, all processes using only one new observation at each step, and all "classic" processes. The group with the best rankings contains all averaged processes with online standardization and using 10 or 100 new observations at each step. Within this group, a clear difference appears after about 10s of processing time between processes using 10 new observations and

processes using 100 new observations. Over all the processing times recorded, the best process appears to be the averaged process with online standardization and piecewise constant step-size with levels of size 200 using 100 new observations at each step (SL33).

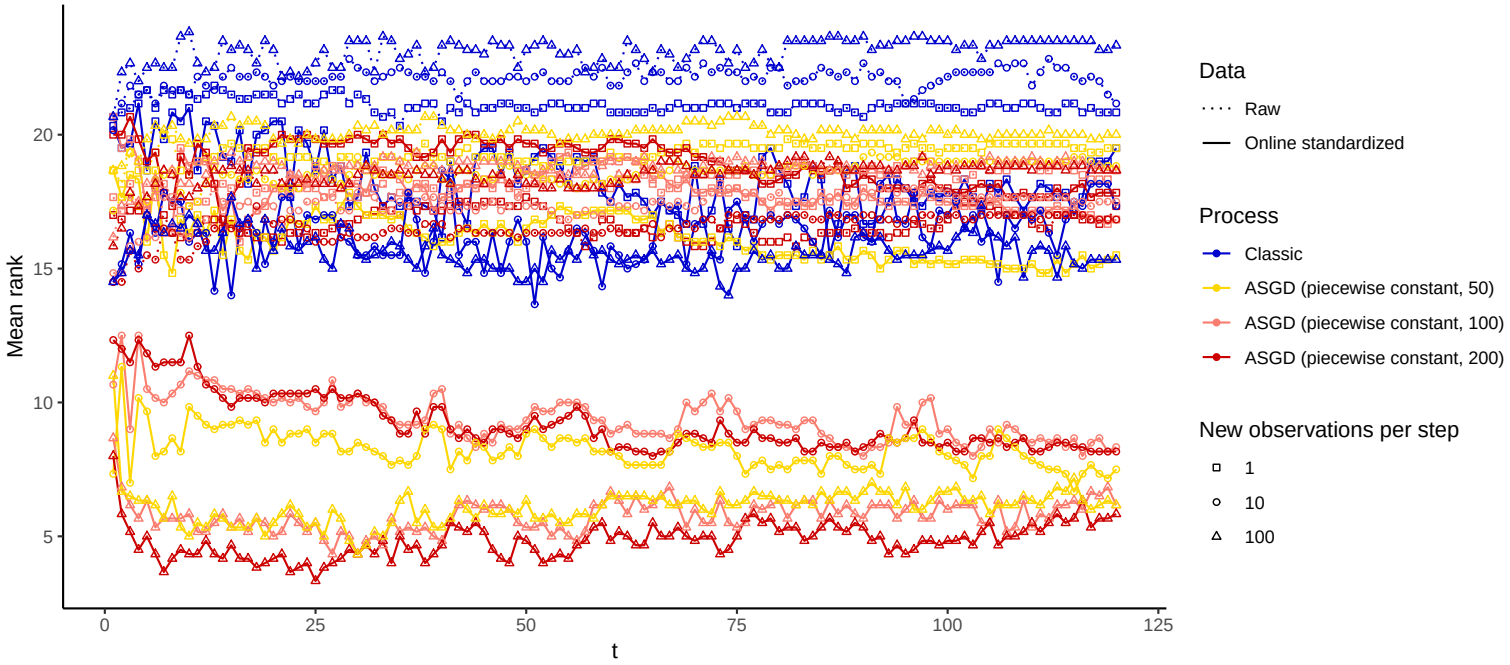

Fig 2. Evolution with the processing time

6 Application to online updating of a score in heart failure patients

In [8], we have presented a methodology for constructing a short-term event (death or hospitalization for heart failure) risk score in heart failure patients, based on an ensemble predictor built in several steps:

- $n_1 = 2$ classification rules (logistic regression and linear discriminant

analysis for mixed data) are used.

- n_2 bootstrap samples are drawn.
- For each sample, a fixed number of explanatory variables are selected according to n_3 modalities of random selection.

This yields a set of $n_1 n_2 n_3$ predictors. As logistic model is a generalized linear model, a score function that is an affine combination of the explanatory variables, as in linear discriminant analysis, can be built.

Then the $n_1 n_2 n_3$ score functions obtained are aggregated in two steps:

- The $n_2 n_3$ score functions for each fixed classification rule are aggregated by averaging and finally reduced on a scale from 0 to 100.
- A single score function is obtained by an optimal weighted averaging of the two previous reduced score functions.

This methodology has been used for EPHESUS trial [12] patients data on whom biological, clinical and medical history variables have been measured.

Let us show that this methodology can be adapted to the case of a data stream.

Suppose that new data for heart failure patients arrive continuously. At step n of the process, a batch of new data is taken into account and allocated to bootstrap samples using Poisson bootstrap [14]. The set of randomly selected variables is fixed for each bootstrap sample; then:

- A predictor based on logistic regression can be updated using the stochastic gradient algorithm with online standardized data studied here. Thus each of the $n_2 n_3$ score functions obtained by logistic regression can be updated online.

- In [1], we have studied stochastic algorithms for updating online a predictor based on linear regression, in particular binary linear discriminant analysis, when new data arrive which are standardized online. Thus each of the $n_2 n_3$ score functions obtained by linear discriminant analysis can be updated online.

- Thus each of the $n_1 n_2 n_3$ score functions can be updated online. By aggregating them according to the method described above, the ensemble score can be updated online.

7 Conclusion

We have studied an averaged constrained stochastic gradient algorithm for performing online a constrained binary logistic regression in the case of streaming or massive data. We have proposed to use an online standardization of the data to avoid a numerical explosion, or when a shrinkage method (such as LASSO) is used, or even when expectations or variances of explanatory variables change (varying with time or depending on the values of controlled variables) and can be estimated online. We have proposed to use a decreasing piecewise constant step-size in order that it does not decrease too quickly and consequently reduces the speed of convergence of the process. We have made experiments on real and simulated datasets. The results confirm the validity of the choices made: online standardization of the data, averaged process and piecewise constant step-size.

Acknowledgments

Results incorporated in this article received funding from the investments for the Future program under grant agreement No ANR-15-RHU-0004. The authors

thank Mr. Edward Sismey for editing this manuscript.

References

1. Duarte K, Monnez JM, Albuisson E. Sequential linear regression with online standardized data. PLOS ONE. 2018;13(1):e0191186. doi:10.1371/journal.pone.0191186.

2. Monnez JM, Skiredj A. Convergence of a normed eigenvector stochastic approximation process and application to online principal component analysis of a data stream; 2018. Available from: `https://hal.archives-ouvertes.fr/hal-01844419`.

3. Cardot H, Cénac P, Monnez JM. A fast and recursive algorithm for clustering large datasets with k-medians. Computational Statistics & Data Analysis. 2012;56(6):1434–1449. doi:10.1016/j.csda.2011.11.019.

4. Pascanu R, Mikolov T, Bengio Y. On the difficulty of training recurrent neural networks. arXiv:12115063 [cs]. 2012.

5. Robbins H, Monro S. A stochastic approximation method. The Annals of Mathematical Statistics. 1951;22(3):400–407.

6. Bach F, Moulines E. Non-strongly-convex smooth stochastic approximation with convergence rate O(1/n). In: Burges CJC, Bottou L, Welling M, Ghahramani Z, Weinberger KQ, editors. Advances in Neural Information Processing Systems 26. Curran Associates, Inc.; 2013. p. 773–781.

7. Bach F. Adaptivity of averaged stochastic gradient descent to local strong convexity for logistic regression. Journal of Machine Learning Research. 2014;15:595–627.

8. Duarte K, Monnez JM, Albuisson E. Methodology for constructing a short-term event risk score in heart failure patients. Applied Mathematics. 2018;09(08):954–974. doi:10.4236/am.2018.98065.

9. Bottou L, Curtis F, Nocedal J. Optimization methods for large-scale machine learning. SIAM Review. 2018;60(2):223–311. doi:10.1137/16M1080173.

10. Robbins H, Siegmund D. A convergence theorem for nonnegative almost supermartingales and some applications. In: Rustagi JS, editor. Optimizing Methods in Statistics. Academic Press; 1971. p. 233–257.

11. Xu W. Towards optimal one pass large scale learning with averaged stochastic gradient descent. arXiv:11072490 [cs]. 2011.

12. Pitt B, Remme W, Zannad F, Neaton J, Martinez F, Roniker B, et al. Eplerenone, a selective Aldosterone blocker, in patients with left ventricular dysfunction after myocardial infarction. New England Journal of Medicine. 2003;348(14):1309–1321. doi:10.1056/NEJMoa030207.

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

14. Oza NC, Russell SJ. Online bagging and boosting. In: Proceedings of the Eighth International Workshop on Artificial Intelligence and Statistics, AISTATS 2001, Key West, Florida, USA, January 4-7, 2001.