Streaming constrained binary logistic regression with online standardized data.

Benoît Lalloué, Jean-Marie Monnez, Eliane Albuisson

ORIGINAL RESEARCH ARTICLE

Streaming constrained binary logistic regression with online standardized data.

Benoît Lalloué[a,b,*], Jean-Marie Monnez[a,b,†] and Eliane Albuisson[c,d,e,‡]

[a]Université de Lorraine, CNRS, Inria (Project-team BIGS), IECL (Institut Elie Cartan de Lorraine), BP239, F-54506 Vandœuvre-lès-Nancy, France
[b]CHRU Nancy, INSERM, Université de Lorraine, CIC Plurithématique, F-54000 Nancy, France
[c]Université de Lorraine, CNRS, IECL, F-54000 Nancy, France
[d]BIOBASE, Pôle S2R, CHRU de Nancy, Vandœuvre-lès-Nancy, France
[e]Faculté de Médecine, InSciDenS, Vandœuvre-lès-Nancy, France

ARTICLE HISTORY
Compiled July 1, 2020

ABSTRACT
Online learning is a method for analyzing very large datasets ("big data") as well as data streams. In this article, we consider the case of constrained binary logistic regression and show the interest of using processes with an online standardization of the data, in particular to avoid numerical explosions or to allow the use of shrinkage methods. We prove the almost sure convergence of such a process and propose using a piecewise constant step-size such that the latter does not decrease too quickly and does not reduce the speed of convergence. We compare twenty-four stochastic approximation processes with raw or online standardized data on five real or simulated data sets. Results show that, unlike processes with raw data, processes with online standardized data can prevent numerical explosions and yield the best results.

KEYWORDS
Big data; Data stream; Logistic regression; Online learning; Stochastic approximation; Stochastic gradient

1. Introduction

1.1. Background

Data stream online analysis concerns data that arrive continuously such as process control data, web data, telecommunication data, medical data or financial data. Online learning, which proceeds in successive steps, the results of which are being updated at each step taking into account a batch of new data, is particularly adapted to data streams. For observations arriving sequentially, recursive stochastic algorithms can be used to estimate, for instance, parameters of a linear regression model [8, 10], principal components of a factorial analysis [18] or centers of classes in non-hierarchical clustering [6], whose estimations are updated by each new arriving data batch. When

* benoit.lalloue@univ-lorraine.fr ; https://orcid.org/0000-0002-7433-9678
† jean-marie.monnez@univ-lorraine.fr
‡ eliane.albuisson@univ-lorraine.fr

using such recursive processes, it is not necessary to store the data and, due to the relative simplicity of the computation involved, a much greater number of data can be taken into account than with classical methods during the same amount of time. Therefore, recursive algorithms can also profitably be used for very large datasets (by randomly drawing a data batch from the dataset at each step) when the capacities of statistical learning methods are potentially limited by computing time. Batch gradient and stochastic gradient methods are presented and compared in [4] and reviewed in [3].

When using a stochastic gradient process, a *numerical explosion* can be encountered [19]. To avoid such phenomenon, methods of gradient variance reduction [3, 13], such as gradient clipping [19], can be used. The idea underpinning gradient clipping is to limit the norm of the gradient to a maximum number called threshold. This number must be chosen and a poor choice of threshold can affect computing speed. In our approach, the limitation of the gradient is implicitly obtained by online standardization of the data: each continuous variable is standardized with respect to the estimations at the current step of its expectation and its standard deviation computed online. Indeed, in the case of a data stream, the mathematical expectation and the variance of each variable are a priori unknown and the usual offline standardization cannot be performed. This may also be an issue when using a *shrinkage method such as LASSO or ridge*, which first necessitates standardizing the explanatory variables. Again, it is not possible to perform the offline standardization in the case of a data stream and an online process can be used, 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 *the expectations and the variances of the explanatory variables depend on the step n or on the values of controlled variables* and a regression model with standardized explanatory variables is defined. Assuming that we can estimate online the expectation and the variance of these variables, we can also use the same type of process to estimate the parameters of the regression function.

In a previous study [10] addressing sequential least square multidimensional linear regression using a stochastic approximation process, we proved the convergence of three processes with online standardized data instead of raw data, discussed the advantages of this approach compared to other methods, and experimentally showed that processes with online standardized data were superior to processes with raw data. In the present study, we use a similar approach in the case of *constrained binary logistic regression, using a stochastic gradient process with online standardization of the data*. Herein, the second Lyapounov method ([14], p.9) is used in the proof of convergence, the additional problem being that the expectations and variances of the explanatory variables are unknown but replaced by convergent online estimations, as in the case of sequential linear regression [10]. We consider *an averaged stochastic gradient process*: intuitively, when the algorithm is not too distant from the solution, averaging allows decreasing the variability of the initial algorithm, which can oscillate around the true solution, and thereby improve its performance [21][6].

Since a suitable choice of step-size is often crucial for obtaining good performances for stochastic gradient processes, we examined various choices of step-size. Bach and Moulines [2] already showed that a constant step-size averaged stochastic gradient 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 alternatively defined other processes with a Newton-approximation scheme. However, Bach [1] suggested

using *a decreasing piecewise constant step-size*, in order that the step-size does not decrease too quickly and does not reduce the speed of convergence, which we test in the present experiments and compare the latter to a more classical decreasing step-size.

Subsection 1.2 is devoted to the formulation of the problem, which is a problem of Stochastic Optimization of the expectation $F(x,a)$ of a random variable $Y(x,a)$ depending on an unknown parameter a which is estimated online along with the stochastic gradient process, Subsection 1.3 to a comparison with other recent formulations such as Stochastic Compositional Optimization [23] and Conditional Stochastic Optimization [12], Subsection 2.1 to a definition and to a theorem of almost sure convergence of the stochastic gradient process, Subsection 2.2 to a comparison with the Stochastic Compositional Gradient Descent algorithm [23] as well as some possible extensions of our work, while Section 3 is devoted to the results of experiments where processes with raw data are compared to processes with online standardized data. The article ends with a conclusion (Section 4) and two appendices: Appendix A contains the proof of the theorem and Appendix B features additional experimental results.

1.2. Formulation of the problem

Let A' be the transpose of a matrix A. The abbreviation a.s. stands for almost surely.

Consider a data stream and assume that the observed data are realizations of a random vector $(R^1,...,R^p,S)$ in $\mathbb{R}^p \times \{0,1\}$. Let R be the random column vector $\left(R^1 \ ... \ R^p \ 1\right)'$, $m = \left(E\left[R^1\right] \ ... \ E\left[R^p\right] \ 0\right)'$, $R^c = R - m$ (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 with diagonal elements $\frac{1}{\sigma^1},...,\frac{1}{\sigma^p},1$ (taking by convention $\sigma^k = 1$ for a categorical variable), $Z = \Gamma R^c$ the vector R whose continuous components are standardized ($z = \Gamma r^c$ a realization of Z) and $\theta = \left(\theta^1 \ ... \ \theta^p \ \theta^{p+1}\right)'$ a column vector of real parameters.

Consider the logistic model with standardized covariates:

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

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

Remark 1. Let $m^j = E[R^j], j = 1,...,p$. Note that if θ_0 is the column vector of the parameters of the logistic regression function of S with respect to R, then $f(s;z,\theta) = f_0(s;r,\theta_0) = \frac{e^{r'\theta_0 s}}{1+e^{r'\theta_0}}$, with for $j = 1,...,p$:

$$\theta_0^j = \frac{\theta^j}{\sigma^j}, \ \theta_0^{p+1} = \theta^{p+1} - \left(\sum_{j=1}^p m^j \frac{\theta^j}{\sigma^j}\right) \Leftrightarrow \theta_0 = \begin{pmatrix} \frac{1}{\sigma^1} & & & \\ & \ddots & & \\ & & \frac{1}{\sigma^p} & \\ -\frac{m^1}{\sigma^1} & \cdots & -\frac{m^p}{\sigma^p} & 1 \end{pmatrix} \theta. \qquad (2)$$

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

$$F(x) = -E[\ln f(S;Z,x)] = E\left[\ln \frac{1+e^{Z'x}}{e^{Z'xS}}\right] = E\left[-Z'xS + \ln\left(1+e^{Z'x}\right)\right] \qquad (3)$$

has θ for unique minimizer since F is a convex function with positive Hessian

$$F''(x) = E\left[ZZ'\frac{e^{Z'x}}{\left(1+e^{Z'x}\right)^2}\right]. \tag{4}$$

θ is the unique solution of:

$$F'(x) = E\left[-ZS + \frac{Ze^{Z'x}}{1+e^{Z'x}}\right] = E\left[Z\left(h\left(Z'x\right) - S\right)\right] = 0. \tag{5}$$

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

1.3. Comparison with other formulations

Let R^2 denote the random column p-vector $\left(\left(R^1\right)^2 \dots \left(R^p\right)^2\right)'$ and $g(R)$ denote the random column $2p+1$-vector $\left(R'R^{2\prime}\right)'$.

The diagonal matrix Γ is a function of $E[R]$ and $E\left[R^2\right]$, thus of $E[g(R)]$: $\Gamma = C\left(E\left[g\left(R\right)\right]\right)$. Moreover $E[R] = AE[g(R)]$, with $A = \left(I_{p+1}\left(0\right)\right)$, I_{p+1} the identity matrix of order $p+1$ and (0) the null $(p+1, p)$ matrix. We can then write $F(x) = F\left(x, E\left[g\left(R\right)\right]\right)$ as the expectation of a function f of x parametrized by $E\left[g(R)\right], R, S$:

$$\begin{aligned} F(x) = F\left(x, E\left[g\left(R\right)\right]\right) &= E\left[\ln\frac{1+\exp\left(\left(R - AE\left[g\left(R\right)\right]\right)'C\left(E\left[g\left(R\right)\right]\right)x\right)}{\exp\left(S\left(R - AE\left[g\left(R\right)\right]\right)'C\left(E\left[g\left(R\right)\right]\right)x\right)}\right] \\ &= E\left[f\left(x, E\left[g\left(R\right)\right]; R, S\right)\right]. \end{aligned} \tag{6}$$

A. The minimization of $F(x)$ is a problem of Stochastic Optimization of the expectation of a random variable $Y(x, a)$ that depends on an unknown parameter $a = E\left[g\left(R\right)\right]$ which is estimated online by iterative averaging along with the solution θ. We have dealt with this type of problem in other settings, for example in a streaming multiple factor analysis of a random vector Z [15] or in a streaming generalized canonical correlation analysis [17] where unknown elements such as an expectation or a covariance matrix or a metric are estimated online along with the principal components. We extended this approach in [16] to the case of principal component analysis of a random vector with a time-varying expectation. We present in Subsection 2.2 the same type of extension.

B. In [7], the authors studied the minimization on a compact set of composite risk functions, in particular

$$F(x) = E_V\left[f\left(x, E_V g\left(x; V\right); V\right)\right]. \tag{7}$$

Formulation (6) can be considered as a particular case of (7) with g not depending on x. However, the authors only established a central limit theorem for the empirical estimator minimizing the empirical risk obtained by replacing the expectations in (7) by empirical means over a finite sample. Here, we solve the minimization of $F(x)$ (6) directly by using a stochastic approximation process including the sequential estimation of $E\left[g\left(R\right)\right]$ by iterative averaging.

C. The aim of Stochastic Compositional Optimization (SCO [23], [11], [25] for multi-level compositional optimization, and references therein) is to minimize the composition of two expected-value functions:

$$F(x) = E_V\left[f\left(E_W g(x;W);V\right)\right]. \tag{8}$$

Taking $V=(R,S)$, $W=R$ and $g(x;R)=\left(x', g(R)'\right)'$, this formulation is formally the same as in equation (6). Thus formulation (6) could be considered as a particular case of (8) and the convergence results of SCO applied to the problem (6). However, we show in Subsection 2.2 that a larger choice of step-sizes than in [23] is allowed by our convergence analysis using a classical method of stochastic approximation.

D. The aim of Conditional Stochastic Optimization (CSO, [12]) is to minimize

$$F(x) = E_V\left[f\left(E_{W/V} g(x;V,W);V\right)\right]. \tag{9}$$

This formulation does not apply to the present case since there is no conditional expectation in formulation (6).

In conclusion, the sequential minimization of $F(x)$ in (6) is a stochastic approximation problem involving a stochastic gradient process and requiring the simultaneous online estimation of the unknown expectation of $g(R)$.

2. Approach: Definition of a stochastic gradient process

2.1. Definition and convergence

Let $\left((R_n^1,...,R_n^p,S_n), n\geqslant 1\right)$ be an i.i.d. sample of $(R^1,...,R^p,S)$ and, for $n\geqslant 1$, $R_n = \left(R_n^1\ ...\ R_n^p\ 1\right)'$, $R_n^c = R_n - m$ and $Z_n = \Gamma R_n^c$. For $k=1,...,p$, let $\overline{R}_n^k$ be 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), $\overline{R}_n = \left(\overline{R_n^1}\ ...\ \overline{R_n^p}\ 0\right)'$ and Γ_n the $(p+1,p+1)$ diagonal matrix with diagonal elements $\frac{1}{\sqrt{\frac{n}{n-1}V_n^1}},...,\frac{1}{\sqrt{\frac{n}{n-1}V_n^p}},1$.

Assume that m_n observations (R_i,S_i) are taken into account at step n of the following defined process. Let $\mu_n = \sum_{i=1}^{n} m_i$, $I_n = \{\mu_{n-1}+1,...,\mu_n\}$ be the set of indices of the observations taken into account at step n, $\widehat{R}_n = \overline{R}_{\mu_n}$, $\widehat{\Gamma}_n = \Gamma_{\mu_n}$ and for $j\in I_n$:

$$\widetilde{Z}_j = \widehat{\Gamma}_{n-1}\left(R_j - \widehat{R}_{n-1}\right) = \widehat{\Gamma}_{n-1}\left(R_j^c - \widehat{R}_{n-1}^c\right) \text{ with } \widehat{R}_{n-1}^c = \widehat{R}_{n-1} - m. \tag{10}$$

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, $\sqrt{\frac{\mu_{n-1}}{\mu_{n-1}-1}}V_{\mu_{n-1}}^k$. Note that all data up to step $n-1$ are used to estimate m and Γ at step n by $\hat{R}_{n-1}$ and $\hat{\Gamma}_{n-1}$ respectively, which are recursively computed.

Assume that θ is constrained to belong to a convex subset K of $\mathbb{R}^{p+1}$ (if there is no constraint, $K=\mathbb{R}^{p+1}$). Let Π be the projection operator on K. Recursively define the stochastic approximation process $(X_n, n\geq 1)$ and the averaged process $\left(\overline{X}_n, n\geq 1\right)$

in $\mathbb{R}^{p+1}$ such that:

$$X_{n+1} = \Pi\left(X_n - a_n \frac{1}{m_n} \sum_{j \in I_n} \widetilde{Z}_j \left(h\left(\widetilde{Z}_j' X_n\right) - S_j\right)\right), \tag{11}$$

$$\overline{X}_{n+1} = \frac{1}{n+1} \sum_{i=1}^{n+1} X_i = \overline{X}_n - \frac{1}{n+1}\left(\overline{X}_n - X_{n+1}\right). \tag{12}$$

Remark 2. The use of the projection operator Π is only necessary if θ is constrained to belong to a convex set K, for example when using a shrinkage method. It is not the case if we wish only to avoid a numerical explosion, as in the experiments conducted in Section 3.

Assume:
(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 2.1. *Under H1a,b and H2, (X_n) and $\left(\overline{X}_n\right)$ converge almost surely to θ.*

The proof using the second Lyapounov method ([14], p.9), also valid in the case of linear regression, is shown in Appendix A.

2.2. Discussion and possible extensions

A. The SCGD (Stochastic Compositional Gradient Descent) algorithm used to solve the minimization of $F(x)$ in (8) is the composition of a stochastic gradient descent algorithm (X_n) with step-size (α_n) and of an iterative weighted algorithm (Y_n) with step-size (β_n) such that:

$$X_{n+1} = \Pi\left(X_n - \alpha_n \nabla g\left(X_n; W_n\right) \nabla f\left(Y_{n+1}; V_n\right)\right), \tag{13}$$

$$Y_{n+1} = \left(1 - \beta_n\right) Y_n + \beta_n g\left(X_n; W_n\right), \tag{14}$$

Π being the projection operator on a closed convex set, (V_n, W_n) an i.i.d. observation of (V, W), (X_n) depending on (Y_n) and (Y_n) on (X_n), α_n and β_n verifying the assumptions

$$\sum_{n=1}^{\infty} \alpha_n = \infty, \sum_{n=1}^{\infty} \beta_n = \infty, \sum_{n=1}^{\infty} \left(\alpha_n^2 + \beta_n^2 + \frac{\alpha_n^2}{\beta_n}\right) < \infty \tag{15}$$

to ensure the a.s. convergence of the algorithm (X_n) ([23], Theorem 1). It requires that the algorithm (X_n) with step-size (α_n) must be slower than the algorithm (Y_n) with step-size (β_n), which decreases the convergence rate and creates practical difficulties according to [11].

B. Consider the algorithm (X_n) defined in equation (11). To estimate $E\left[R^j\right]$ and $Var\left[R^j\right]$ for $j = 1, \ldots, p$, we estimate $E\left[g\left(R\right)\right]$ whose unknown components are $E\left[R^j\right]$ and $E\left[(R^j)^2\right]$, $j = 1, \ldots, p$, by iterative averaging, introducing at each step n a mini-batch of m_n observations R_k, $k \in I_n$, and estimating $E\left[g\left(R\right)\right]$ at step n by M_n such

that

$$M_{n+1} = (1-\beta_n) M_n + \beta_n \frac{1}{m_n} \sum_{i \in I_n} g(R_i), M_1 = 0, \text{ with } \beta_n = \frac{m_n}{\mu_n}, \text{ thus} \tag{16}$$

$$M_{n+1} = \frac{1}{\mu_n} \sum_{i=1}^{\mu_n} g(R_i). \tag{17}$$

Then $\widehat{R}_{n-1} = \left(M_n^1 \dots M_n^p \ 1\right)$ and $\widehat{V}_{n-1}^i = M_n^{p+1+i} - \left(M_n^i\right)^2$, $i = 1, ..., p$. Thus the algorithm (X_n) defined by equation (11) depends on (M_n).

C. Consider the case where $m_n = 1$ for all n and compare (11) and (12) with SCGD defined by (13) and (14) taking $\alpha_n = a_n$ and $\beta_n = \frac{m_n}{\mu_n} = \frac{1}{n}$.

According to definition (14) of (Y_n), as $g(x; R) = \binom{x}{g(R)}$ and $E[g(x; R)] = \binom{x}{E[g(R)]}$, we have $Y_n = \binom{Y_n^1}{Y_n^2}$ and:

a) $Y_{n+1}^1 = (1-\beta_n) Y_n^1 + \beta_n X_n = \frac{1}{n} \sum_{i=1}^n X_i = \overline{X}_n$, with $\mathrm{Y}_1^1 = 0$;
b) $Y_{n+1}^2 = (1-\beta_n) Y_n^2 + \beta_n g(R_n)$, thus $Y_{n+1}^2 = M_{n+1}$, with $\mathrm{Y}_1^2 = 0$;
c) $X_{n+1} = \Pi\left(X_n - a_n \widetilde{Z}_n \left(h\left(\widetilde{Z}_n' \overline{X}_n\right) - S_n\right)\right)$;

thus, SCGD is different from the algorithm defined by (11) and (12) as $h\left(\widetilde{Z}_n' X_n\right)$ in (11) is replaced by $h\left(\widetilde{Z}_n' \overline{X}_n\right)$;

d) if we use in definition (11) of (X_n) with $m_n = 1$, the step-size $a_n = \frac{1}{n^\alpha}$, $\frac{1}{2} < \alpha \leq 1$, and take $\beta_n = \frac{1}{n}$, then $\frac{a_n^2}{\beta_n} = \frac{1}{n^{2\alpha-1}} \geqslant \frac{1}{n}$, thus (15) is not verified; however, (a_n) verifies the assumption $\sum_{n=1}^\infty a_n = \infty, \sum_{n=1}^\infty a_n^2 < \infty, \sum_{n=1}^\infty \frac{a_n}{\sqrt{n}} < \infty$.

D. A more general step-size β_n could be taken and, in the definition of $\widetilde{Z}_j$, $j \in I_n$, $\widehat{R}_{n-1}$ could be replaced by $(M_n^1 \dots M_n^p \ 1)'$ and $\widehat{\Gamma}_{n-1}$ by a $(p+1, p+1)$ diagonal matrix with diagonal elements $\frac{1}{\sqrt{M_n^{p+1+i} - (M_n^i)^2}}$, $i = 1, \dots, p$, and 1. Then it can be proved that (M_n) converges a.s. to $E[g(R)]$ and (X_n) defined by (11) to θ under the following assumptions on (β_n):

$$\begin{aligned} &\beta_n > 0, \frac{\beta_n}{\beta_{n+1}} \leq 1 + \gamma\beta_n + \gamma_n + o(\beta_n), \gamma < 2, \gamma_n \geq 0, \sum_{n=1}^\infty \gamma_n < \infty, \\ &\sum_{n=1}^\infty \beta_n = \infty, \sum_{n=1}^\infty \beta_n^2 < \infty \end{aligned} \tag{18}$$

and assumptions H1a, H1b, H2 with $\sum_{n=1}^\infty \frac{a_n}{\sqrt{n}} < \infty$ replaced by $\sum_{n=1}^\infty a_n \sqrt{\beta_n} < \infty$. Condition (18) is verified for example when $\beta_n = \frac{1}{n}$ or from a certain rank when $\beta_n = \frac{1}{n^\alpha}$, $\frac{1}{2} < \alpha < 1$. Note that the same time-scale can be taken for the two processes (X_n) and (M_n), $\beta_n = a_n$ verifying $\sum_{n=1}^\infty (a_n)^{\frac{3}{2}} < \infty$.

In conclusion, solving the minimization of $F(x)$ in equations (6) and (8) involves the composition of a stochastic gradient descent algorithm and an iterative weighted algorithm, although the performed convergence analyses are different, allowing a larger choice of step-sizes in the present study without assumption on the comparison of the convergence rates of (a_n) and (β_n).

E. Note that Theorem 2.1 can be extended to the case where the explanatory

variables have an expectation and a variance depending on n or on the values of controlled variables, thus m and Γ depend on n and are denoted by $m(n)$ and $\Gamma(n)$. Then $\hat{R}_{n-1}$ and $\hat{\Gamma}_{n-1}$ must be replaced by estimators of $m(n)$ and $\Gamma(n)$, respectively Θ_n and Φ_n depending on data up to step $n-1$ (as in [16] in the case of a time-varying expectation). To ensure the a.s. convergence of the process (X_n), we assume that:

$$\sup_n \|m(n)\| < \infty, \ \sup_n \|\Gamma(n)\| < \infty, \tag{19}$$

$$\Theta_n - m(n) \longrightarrow 0, \ \sum_{n=1}^{\infty} a_n \|\Theta_n - m(n)\| < \infty, \tag{20}$$

$$\Gamma_n - \Gamma(n) \longrightarrow 0, \ \sum_{n=1}^{\infty} a_n \|\Gamma_n - \Gamma(n)\| < \infty \text{ a.s.} \tag{21}$$

3. Application

Twenty-four stochastic approximation processes were compared, including classical stochastic gradient descent (SGD), averaged stochastic gradient descent (ASGD) with a piecewise constant step-size with different level sizes as suggested in [1], as well as the same processes but with the online standardization of the data defined in Subsection 2.1.

The processes and their respective parameters are described in Table 1. Abbreviations used to name the processes are as follows: C for classical SGD or A for ASGD; R for raw data or S for online standardized data; V for variable step-size or P for piecewise constant step-size. For instance, AR1P50 is the averaged process with raw data, 1 observation per step, piecewise constant step-size with level size 50; CS1V is the classical process with online standardized data, 1 observation per step and variable step-size. Processes on raw data (in particular "CR.") are those currently used, while those using online standardization of the data (particularly averaged stochastic gradient descent with piecewise constant step-size, "AS.P."), are introduced in this article.

3.1. Step-size

For processes with a variable step-size (V), we have defined $a_n = \frac{c}{(b+n)^\alpha}$. For processes with a piecewise constant step-size (P), we have chosen $a_n = \frac{c}{(b+\lfloor \frac{n}{\tau} \rfloor)^\alpha}$ where $\lfloor . \rfloor$ denotes the integer part while τ is the size of the levels. For both cases, we set $\alpha = 2/3$ (as suggested by Xu [24] in the case of linear regression), $b = 1$ and $c = 1$.

3.2. Initialization and simulation of a data stream

All processes were initialized with $X_1 = 0$. 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 prior to 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.

Table 1. Description of the processes.

Abbreviation	Method type	Type of data	Number of observations used at each step of the process	Step-size	Levels size	Use of the averaged process
CR1V			1			
CR10V	***C**lassical (C)*		10	**V**ariable (V)	-	No
CR100V			100			
AR1P50			1			
AR10P50			10		50	
AR100P50		**R**aw data (R)	100			
AR1P100			1	**P**iecewise constant (P)		
AR10P100	***A**veraged (A) SGD*		10		100	Yes
AR100P100			100			
AR1P200			1			
AR10P200			10		200	
AR100P200			100			
CS1V			1			
CS10V	***C**lassical (C)*		10	**V**ariable (V)	-	No
CS100V			100			
AS1P50			1			
AS10P50			10		50	
AS100P50		Online **S**tandardized data (S)	100			
AS1P100			1	**P**iecewise constant (P)		
AS10P100	***A**veraged (A) SGD*		10		100	Yes
AS100P100			100			
AS1P200			1			
AS10P200			10		200	
AS100P200			100			

Then, for each dataset, a data stream was simulated by randomly sampling with replacement a data batch of 1, 10 or 100 observations (depending on the process studied) at each step of the process.

Processes were implemented with the R 3.6.1 software (64-bit version).

3.3. Convergence criteria

The coefficients obtained by the usual "offline" logistic regression (using R's `glm` function) on a dataset $((r_i^1, ..., r_i^p, s_i), i = 1, ..., N)$ were used as "gold standard" 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.

$$\text{As } \theta_0 = \begin{pmatrix} \frac{1}{\sigma^1} & & & \\ & \ddots & & \\ & & \frac{1}{\sigma^p} & \\ -\frac{m^1}{\sigma^1} & \cdots & -\frac{m^p}{\sigma^p} & 1 \end{pmatrix} \theta,\ \hat{\theta}_{n+1} = \begin{pmatrix} \hat{\Gamma}_n(1,1) & & & \\ & \ddots & & \\ & & \hat{\Gamma}_n(p,p) & \\ -\hat{\Gamma}_n(1,1)\hat{r}_n^1 & \cdots & -\hat{\Gamma}_n(p,p)\hat{r}_n^p & 1 \end{pmatrix} \bar{x}_{n+1} \tag{22}$$

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

The relative norm of the difference between θ^c and $\hat{\theta}_{n+1}$, $\frac{\|\theta^c - \hat{\theta}_{n+1}\|}{\|\theta^c\|}$, was used as a convergence criterion.

The cosine of the angle between θ^c and $\hat{\theta}_{n+1}$, $\frac{\theta^{c\prime}\hat{\theta}_{n+1}}{\|\theta^c\|\|\hat{\theta}_{n+1}\|}$, the coefficient of correlation between the predictions obtained with the usual 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).

3.4. Datasets

The processes were tested on four datasets available on the Internet and one dataset derived from the EPHESUS study [20][1], all of which have already been used to test the performance of stochastic approximation processes with online standardized data in the case of online linear regression [10]. `Twonorm`, `Ringnorm`, `Quantum` and `Adult` datasets are commonly used to test classification methods. `Twonorm`[2] and `Ringnorm`[3], introduced by Breiman [5], contain simulated data with homogeneous variables. `Quantum` contains observed "clean" data, without outliers and with most of its variables on a similar scale. `Adult` and `HOSPHF30D` contain observed data with outliers, heterogeneous variables of different types and scales. Table 2 summarizes these datasets.

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
HOSPHF30D	21382	21382	29	13	derived from EPHESUS study [20]

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 performed 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 obtain 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.
- `HOSPHF30D`: 13 variables were selected using stepwise selection.

All processes were applied on all datasets for a fixed number of $100N$ observations used and for a fixed processing time of 60s (the cumulative time to compute the process updates, excluding operations such as data sampling, data management, formatting and recording of results, etc.).

For each dataset and at each recording point (see below), processes that did not explode were ranked from the best (lowest relative norm) to the worst (highest relative

[1]Due to legal restrictions, data from the EPHESUS study are only available upon request. Interested researchers may request access to data upon approval from the EPHESUS Executive Steering Committee of the study.

[2]*`Twonorm` "is 20 dimension, 2 class data. Each class is drawn from a multivariate normal distribution with unit covariance matrix. Class 1 has mean (a,a, ... ,a) and class 2 has mean (-a,-a, ... ,-a).*"(extract from [5])

[3]*`Ringnorm` "is 20 dimension, 2 class data. Class 1 is multivariate normal with mean zero and covariance matrix 4 times the identity. Class 2 has unit covariance matrix and mean (a,a, ...,a).*"(extract from [5])

norm). The mean rank over all datasets was used to compare the global performance of the processes without any numerical explosion.

Processing time to treat $100N$ 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, since CPU and memory usage by other software were not controlled during the running of the processes which may explain small differences.

3.5. Comparison for a fixed number of observations

As in [10], the criteria values for each process were recorded every N observations used, from $1N$ to $100N$, N being the number of selected observations after preprocessing in the studied dataset. For the relative norm criterion, results for $100N$ 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, CR1V will run for $100N$ iterations whereas CR100V will run for N iterations).

Table 3. Relative norms for $100N$ observations used

Process	Twonorm	Ringnorm	Quantum	Adult	HOSPHF30D	Mean rank
CR1V	0.085	0.026*	0.304	**EXPL**	**EXPL**	-
CR10V	0.206	0.017*	0.500	**EXPL**	**EXPL**	-
CR100V	0.335	0.019*	0.661	**EXPL**	**EXPL**	-
AR1P50	0.028*	0.037*	0.087	**EXPL**	**EXPL**	-
AR10P50	0.010*	0.013*	0.118	**EXPL**	**EXPL**	-
AR100P50	0.034*	0.006*	0.191	**EXPL**	**EXPL**	-
AR1P100	0.035*	0.061	0.065	**EXPL**	**EXPL**	-
AR10P100	0.010*	0.021*	0.102	**EXPL**	**EXPL**	-
AR100P100	0.014*	0.007*	0.144	**EXPL**	**EXPL**	-
AR1P200	0.040*	0.102	0.041*	**EXPL**	**EXPL**	-
AR10P200	0.012*	0.034*	0.090	**EXPL**	**EXPL**	-
AR100P200	0.009*	0.011*	0.123	**EXPL**	**EXPL**	-
CS1V	0.108	0.016*	0.074	0.057	0.088	10.2
CS10V	0.234	0.011*	0.041*	0.085	0.260	9.8
CS100V	0.364	0.009*	0.138	0.120	0.629	10.6
AS1P50	0.016*	0.015*	0.027*	0.035*	0.064	6.6
AS10P50	0.011*	0.008*	0.024*	0.011*	0.060	3.2
AS100P50	0.048*	0.006*	0.024*	0.021*	0.067	5.8
AS1P100	0.018*	0.023*	0.035*	0.040*	0.062	7.6
AS10P100	0.010*	0.010*	0.024*	0.014*	0.060	3.8
AS100P100	0.021*	0.006*	0.023*	0.013*	0.065	4.2
AS1P200	0.026*	0.036*	0.049*	0.055	0.057	8.0
AS10P200	0.011*	0.015*	0.028*	0.018*	0.059	5.2
AS100P200	0.010*	0.007*	0.022*	0.011*	0.066	3.0

* denotes a criterion value < 0.05.
EXPL: numerical explosion.
Process type: C for classical SGD, A for ASGD. *Data type:* R for raw data, S for online standardized data.
First number: number of new observations at each step.
Step-size: V for variable, P for piecewise constant (*second number* denotes the level size).

Globally, processes R and S converged similarly on simulated datasets with homogeneous data (especially `Twonorm` which already contains standardized variables). However, it was not verified for observed datasets for which processes S yielded better results, especially with heterogeneous data (`Adult2`, `HOSPHF30D`): all tested processes R had a numerical explosion for these two datasets.

Over all datasets, the processes S with the lowest mean rankings were averaged processes with piecewise constant step-sizes, the best being AS100P200 (Table 4). Among these AS.P. processes, those with the highest mean rankings were processes

with one observation per step (AS1P.). Note that for `HOSPHF30D`, all AS.P. processes had a criterion value lower than 0.05 after $300N$ observations used (Appendix Table B1).

Table 4. Processes S ordered by mean ranks for $100N$ observations used

Process	AS100P200	AS10P50	AS10P100	AS100P100	AS10P200	AS100P50
Mean rank	3.0	3.2	3.8	4.2	5.2	5.8
Process	AS1P50	AS1P100	AS1P200	CS10V	CS1V	CS100V
Mean rank	6.6	7.6	8.0	9.8	10.2	10.6

Additional results for both fixed and varying numbers of observations are available as Supplemental online material.

3.6. Comparison for a fixed processing time

As in [10], the values of the criteria for each process were then recorded every second of processing time from 1 to 120s. Results for the relative norm criterion after 60s of processing time are shown in Table 5.

Table 5. Relative norms for 60s of processing time

Process	**Twonorm**	**Ringnorm**	**Quantum**	**Adult**	**HOSPHF30D**	**Mean rank**
CR1V	0.055	0.019*	0.288	**EXPL**	**EXPL**	-
CR10V	0.061	0.005*	0.310	**EXPL**	**EXPL**	-
CR100V	0.073	0.002*	0.333	**EXPL**	**EXPL**	-
AR1P50	0.011*	0.019*	0.086	**EXPL**	**EXPL**	-
AR10P50	0.002*	0.002*	0.095	**EXPL**	**EXPL**	-
AR100P50	0.001*	0.001*	0.102	**EXPL**	**EXPL**	-
AR1P100	0.015*	0.029*	0.064	**EXPL**	**EXPL**	-
AR10P100	0.002*	0.003*	0.079	**EXPL**	**EXPL**	-
AR100P100	0.001*	0.001*	0.090	**EXPL**	**EXPL**	-
AR1P200	0.018*	0.052	0.040*	**EXPL**	**EXPL**	-
AR10P200	0.002*	0.005*	0.064	**EXPL**	**EXPL**	-
AR100P200	0.001*	0.001*	0.076	**EXPL**	**EXPL**	-
CS1V	0.139	0.023*	0.173	0.134	0.153	10.0
CS10V	0.182	0.011*	0.057	0.101	0.228	9.0
CS100V	0.227	0.004*	0.071	0.108	0.326	9.0
AS1P50	0.027*	0.025*	0.042*	0.389	0.095	8.6
AS10P50	0.006*	0.005*	0.014*	0.020*	0.053	4.8
AS100P50	0.009*	0.002*	0.007*	0.017*	0.014*	3.2
AS1P100	0.032*	0.037*	0.071	0.386	0.087	9.2
AS10P100	0.005*	0.006*	0.014*	0.025*	0.050*	4.8
AS100P100	0.004*	0.002*	0.007*	0.011*	0.011*	1.8
AS1P200	0.046*	0.060	0.121	0.498	0.112	10.6
AS10P200	0.005*	0.008*	0.017*	0.035*	0.049*	5.4
AS100P200	0.003*	0.002*	0.007*	0.009*	0.012*	1.6

* denotes a criterion value < 0.05.
EXPL: numerical explosion.
Process type: C for classical SGD, A for ASGD. *Data type:* R for raw data, S for online standardized data.
First number: number of new observations at each step.
Step-size: V for variable, P for piecewise constant (*second number* denotes the level size).

Again, all tested processes R had a numerical explosion for the same two datasets (`Adult2` and `HOSHF30D`).

Over all datasets, processes S with the lowest mean rankings were averaged processes with piecewise constant step-sizes (Table 6), except AS1P100 and AS1P200 with one observation per step (which use few observations per second, see Table 7). The best

process was AS100P200. Of note, for `HOSPHF30D`, processes AS10P. and AS100P. had a criterion value lower than 0.05 at 120s (Appendix Table B2).

Table 6. Processes S ordered by mean ranks for 60s of processing time

Process	AS100P200	AS100P100	AS100P50	AS10P50	AS10P100	AS10P200
Mean rank	1.6	1.8	3.2	4.8	4.8	5.4
Process	AS1P50	CS10V	CS100V	AS1P100	CS1V	AS1P200
Mean rank	8.6	9.0	9.0	9.2	10.0	10.6

The average numbers of observations used per second for 60s of processing time are shown in Table 7. For all processes, the number of observations used per second increased with the number of observations used at each step. Due to the online updating of expectations and variances, processes S treated less observations per second than their equivalent on raw data, the ratio n_R/n_S increasing with the number of observations used at each step (n_R, resp. n_S, the number of observations treated per second by a process R, resp. S).

These results combined with those of Table 5 indicate that processes S with piecewise constant step-sizes, particularly AS100P. and AS10P., achieved a better performance using less observations.

Table 7. Average number of observations used per second for 60s of processing time

Process	**Twonorm**	**Ringnorm**	**Quantum**	**Adult**	**HOSPHF30D**
CR1V	32 464	31 661	32 632	**EXPL**	**EXPL**
CR10V	307 385	291 475	292 206	**EXPL**	**EXPL**
CR100V	2 189 952	2 016 338	2 248 173	**EXPL**	**EXPL**
AR1P50	32 019	29 106	30 247	**EXPL**	**EXPL**
AR10P50	273 377	303 805	283 962	**EXPL**	**EXPL**
AR100P50	2 169 318	2 226 950	2 189 662	**EXPL**	**EXPL**
AR1P100	31 433	33 362	30 591	**EXPL**	**EXPL**
AR10P100	282 423	292 126	283 747	**EXPL**	**EXPL**
AR100P100	2 127 127	2 248 095	2 210 533	**EXPL**	**EXPL**
AR1P200	29 100	30 919	29 420	**EXPL**	**EXPL**
AR10P200	271 102	275 657	273 606	**EXPL**	**EXPL**
AR100P200	2 077 237	2 019 090	2 129 145	**EXPL**	**EXPL**
CS1V	5 970	6 061	6 162	5 799	6 715
CS10V	33 469	34 216	35 776	29 430	41 679
CS100V	142 573	141 815	154 233	119 468	169 937
AS1P50	5 862	5 856	5 635	5 613	6 560
AS10P50	33 780	34 336	35 548	31 498	43 315
AS100P50	139 907	142 993	154 767	119 863	190 613
AS1P100	6 030	6 031	5 994	5 819	7 059
AS10P100	33 884	34 075	34 904	32 019	42 846
AS100P100	141 037	144 380	151 340	121 745	180 893
AS1P200	5 819	5 871	5 978	5 606	7 016
AS10P200	34 032	34 474	34 638	31 910	44 620
AS100P200	133 773	140 265	149 313	108 710	193 833

EXPL: numerical explosion.
Process type: C for classical SGD, A for ASGD. *Data type:* R for raw data, S for online standardized data.
First number: number of new observations at each step.
Step-size: V for variable, P for piecewise constant (*second number* denotes the level size).

3.7. Comparison of processes S for a varying processing time

When studying the evolution of the mean rankings of processes S with processing times from 1 to 120s (Figure 1), two groups of processes clearly appeared from the outset and

remained apparent throughout the entire studied period. The group with the worst rankings contains all processes using only one new observation at each step as well as all "classical" processes. The group with the best rankings contained all averaged processes S using 10 or 100 new observations at each step. Within this group, a clear difference appeared after approximately 10s of processing time between processes using 10 new observations and those using 100 new observations. Of all the processing times recorded, the two best processes S appeared to be AS100P100 and AS100P200.

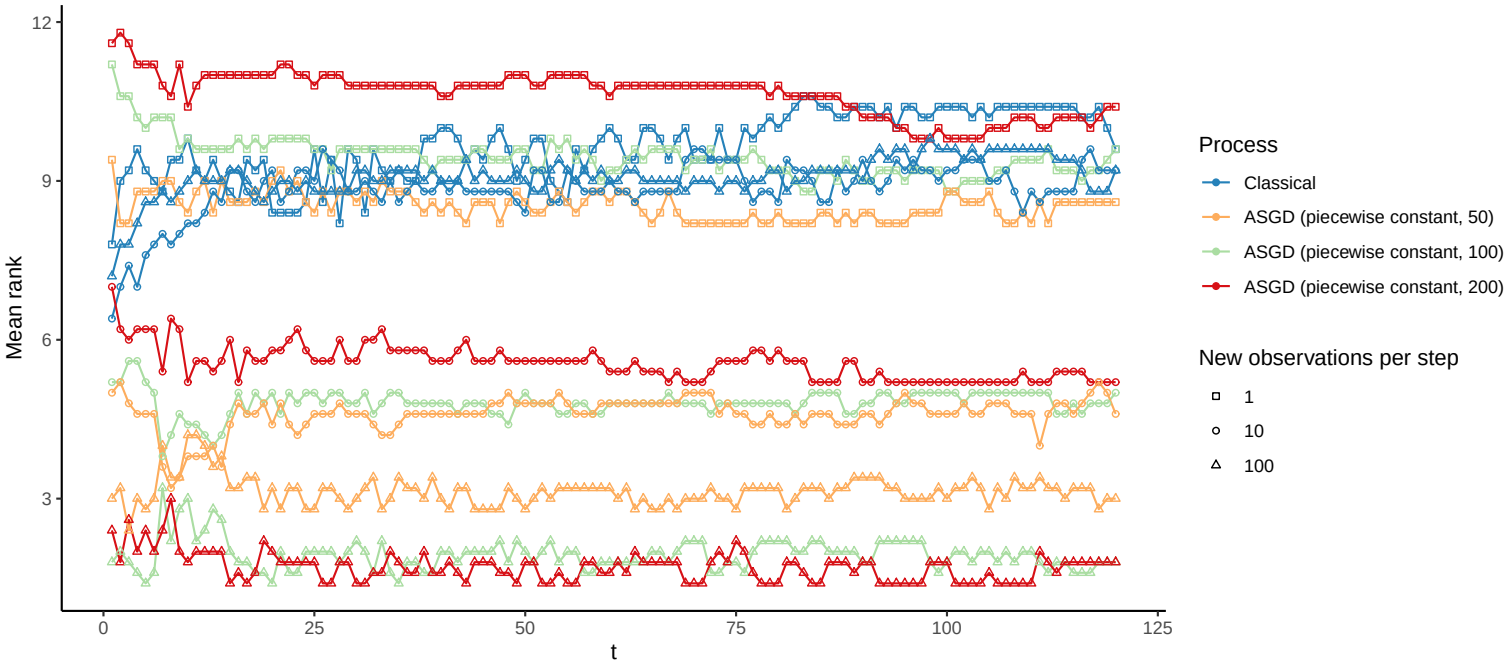

Figure 1. Evolution according to processing time

4. Conclusion

In the present analysis, we studied a constrained averaged stochastic gradient algorithm with online standardized data for performing an online constrained binary logistic regression in the case of streaming or massive data. The proposed approach included using an online standardization of the data to avoid a numerical explosion, or when using a shrinkage method (such as LASSO), 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 also proposed using a decreasing piecewise constant step-size in order for the latter to not decrease too quickly and therefore not reduce the speed of convergence of the process. The results of the experiments conducted on real and simulated datasets confirm the validity of the choices made: online standardization of the data, averaged process and piecewise constant step-size. This work will be applied in an ongoing study to update an ensemble score online to detect adverse events in the case of heart failure [8, 9].

Acknowledgments

The authors thank Mr. Pierre Pothier and Mr. Edward Sismey for editing this manuscript.

Disclosure statement

No potential conflict of interest was reported by the authors.

Funding

This work was supported by the investments for the Future Program under grant ANR-15-RHU-0004.

References

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

[2] F. Bach and E. Moulines, *Non-strongly-convex smooth stochastic approximation with convergence rate O(1/n)*, in *Advances in Neural Information Processing Systems 26*, C.J.C. Burges, L. Bottou, M. Welling, Z. Ghahramani, and K.Q. Weinberger, eds., Curran Associates, Inc., 2013, pp. 773–781.

[3] L. Bottou, F. Curtis, and J. Nocedal, *Optimization methods for large-scale machine learning*, SIAM Review 60 (2018), pp. 223–311.

[4] L. Bottou and Y. Le Cun, *On-line learning for very large data sets*, Applied stochastic models in business and industry 21 (2005), pp. 137–151.

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

[6] H. Cardot, P. Cénac, and J.M. Monnez, *A fast and recursive algorithm for clustering large datasets with k-medians*, Computational Statistics & Data Analysis 56 (2012), pp. 1434–1449.

[7] D. Dentcheva, S. Penev, and A. Ruszczyński, *Statistical estimation of composite risk functionals and risk optimization problems*, Annals of the Institute of Statistical Mathematics 69 (2017), pp. 737–760.

[8] K. Duarte, *Medical decision support and telemedecine in the monitoring of heart failure. Ph.D. thesis.*, Université de Lorraine (France) (2018). Available at https://hal.univ-lorraine.fr/tel-02096008.

[9] K. Duarte, J.M. Monnez, and E. Albuisson, *Methodology for constructing a short-term event risk score in heart failure patients*, Applied Mathematics 09 (2018), pp. 954–974.

[10] K. Duarte, J.M. Monnez, and E. Albuisson, *Sequential linear regression with online standardized data*, PLOS ONE 13 (2018), p. e0191186.

[11] S. Ghadimi, A. Ruszczyński, and M. Wang, *A Single Time-Scale Stochastic Approximation Method for Nested Stochastic Optimization*, arXiv:1812.01094 [math] (2019).

[12] Y. Hu, X. Chen, and N. He, *Sample Complexity of Sample Average Approximation for Conditional Stochastic Optimization*, arXiv:1905.11957 [math, stat] (2020).

[13] R. Johnson and T. Zhang, *Accelerating stochastic gradient descent using predictive variance reduction*, in *Advances in Neural Information Processing Systems 26*, C.J.C. Burges, L. Bottou, M. Welling, Z. Ghahramani, and K.Q. Weinberger, eds., Curran Associates, Inc., 2013, pp. 315–323.

[14] L. Ljung, G.C. Pflug, and H. Walk, *Stochastic approximation and optimization of random systems*, DMV Seminar, Band 17, Birkhäuser, Basel, 1992.
[15] J.M. Monnez, *Approximation stochastique en analyse factorielle multiple (Stochastic approximation in multiple factor analysis)*, Annales de l'ISUP 50 (2006), pp. 27–45.
[16] J.M. Monnez, *Analyse en composantes principales d'un flux de données d'espérance variable dans le temps (Principal component analysis of a data stream with time-varying expectation)*, Revue des Nouvelles Technologies de l'Information (2008), pp. 43–56.
[17] J.M. Monnez, *Stochastic approximation of the factors of a generalized canonical correlation analysis*, Statistics and Probability Letters 78 (2008), pp. 2210–2216.
[18] J.M. Monnez and A. Skiredj, *Convergence of a normed eigenvector stochastic approximation process and application to online principal component analysis of a data stream*, 2018. Available at https://hal.archives-ouvertes.fr/hal-01844419.
[19] R. Pascanu, T. Mikolov, and Y. Bengio, *Understanding the exploding gradient problem*, arXiv:1211.5063 (2012). Available at http://arxiv.org/abs/1211.5063.
[20] B. Pitt, W. Remme, F. Zannad, J. Neaton, F. Martinez, B. Roniker, R. Bittman, S. Hurley, J. Kleiman, and M. Gatlin, *Eplerenone, a selective aldosterone blocker, in patients with left ventricular dysfunction after myocardial infarction*, New England Journal of Medicine 348 (2003), pp. 1309–1321.
[21] B.T. Polyak and A.B. Juditsky, *Acceleration of stochastic approximation by averaging*, SIAM Journal on Control and Optimization 30 (1992), pp. 838–855.
[22] H. Robbins and D. Siegmund, *A convergence theorem for nonnegative almost supermartingales and some applications*, in *Optimizing Methods in Statistics*, J.S. Rustagi, ed., Academic Press, 1971, pp. 233–257.
[23] M. Wang, E.X. Fang, and H. Liu, *Stochastic compositional gradient descent: algorithms for minimizing compositions of expected-value functions*, Mathematical Programming 161 (2017), pp. 419–449.
[24] W. Xu, *Towards optimal one pass large scale learning with averaged stochastic gradient descent*, arXiv:1107.2490 (2011). Available at http://arxiv.org/abs/1107.2490.
[25] S. Yang, M. Wang, and X.E. Fang, *Multilevel stochastic gradient methods for nested composition optimization*, SIAM Journal on Optimization 29 (2019), pp. 616–659.

Appendix A. Proof of the convergence theorem

The usual Euclidean norm in $\mathbb{R}^{p+1}$ and the spectral norm for matrices are used in this proof. Let us state the Robbins-Siegmund lemma [22] and another lemma ([10], Lemma 5) used in this proof. This proof is also valid in the case of linear regression.

Lemma A.1. *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 a.s. to a finite random variable and* $\sum_{n=1}^{\infty}\gamma_n<\infty$ *a.s.*

Lemma A.2. *Suppose H1b holds and $a_n>0$,* $\sum_{n=1}^{\infty}\frac{a_n}{\sqrt{n}}<\infty$. *Then:*

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

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 yields a.s.:

$$\begin{aligned} E\left[\|X_{n+1} - \theta\|^2 \mid T_n \right] \leq \|X_n - \theta\|^2 & \\ - 2a_n \left\langle X_n - \theta, \frac{1}{m_n} \sum_{j \in I_n} E\left[\widetilde{Z}_j \left(h\left(\widetilde{Z}_j' X_n \right) - S_j \right) \mid T_n \right] \right\rangle & \\ + a_n^2 E\left[\left\| \frac{1}{m_n} \sum_{j \in I_n} \widetilde{Z}_j \left(h\left(\widetilde{Z}_j' X_n \right) - S_j \right) \right\|^2 \mid T_n \right] \text{ a.s.} \qquad \text{(A1)} \end{aligned}$$

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

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

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

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

Consider the decomposition

$$E\left[\widetilde{Z}_j \left(h\left(\widetilde{Z}_j' X_n \right) - h\left(Z_j' \theta \right) \right) \mid T_n \right] = E\left[Z_j \left(h\left(Z_j' X_n \right) - h\left(Z_j' \theta \right) \right) \mid T_n \right] + E\left[V_j \mid T_n \right] \qquad \text{(A2)}$$

$$\text{with } V_j = \left(\widetilde{Z}_j - Z_j \right) \left(h\left(Z_j' X_n \right) - h\left(Z_j' \theta \right) \right) + \widetilde{Z}_j \left(h\left(\widetilde{Z}_j' X_n \right) - h\left(Z_j' X_n \right) \right) \qquad \text{(A3)}$$

For $j \in I_n$, there exist $0 \leq \lambda_j \leq 1$, ξ_j^1 and ξ_j^2 such that:

a) $h\left(Z_j'X_n\right) - h\left(Z_j'\theta\right) = Z_j'\left(X_n-\theta\right)h'\left(\xi_j\right)$, with $\xi_j = \lambda_j Z_j'X_n + \left(1-\lambda_j\right)Z_j'\theta$, $Z_j = \Gamma R_j^c$;

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

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

$$\begin{aligned}
V_j = &\left(\left(\widehat{\Gamma}_{n-1}-\Gamma\right)R_j^c - \widehat{\Gamma}_{n-1}\widehat{R}_{n-1}^c\right)R_j^{c\prime}\Gamma\left(X_n-\theta\right)h'\left(\xi_j\right)\\
&+\widehat{\Gamma}_{n-1}\left(R_j^c - \widehat{R}_{n-1}^c\right)\left(-\widehat{R}_{n-1}^{c\prime}\Gamma X_n h'\left(\xi_j^2\right) + \left(R_j^c - \widehat{R}_{n-1}^c\right)'\left(\widehat{\Gamma}_{n-1}-\Gamma\right)X_n h'\left(\xi_j^1\right)\right).
\end{aligned} \tag{A4}$$

Part 3. For $0 < h'(x) \le c$ ($c = \frac{1}{4}$ for logistic regression, $c = 1$ for linear regression), for $j \in I_n$:

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

Since $\widehat{\Gamma}_{n-1}$ and $\widehat{R}_{n-1}^c$ are T_n-measurable and converge respectively to Γ and 0, since $\sum_{n=1}^{\infty} a_n\left\|\widehat{R}_{n-1}^c\right\| < \infty$ and $\sum_{n=1}^{\infty} a_n\left\|\widehat{\Gamma}_{n-1}-\Gamma\right\| < \infty$ a.s. by Lemma A.2, it follows that there exist two non-negative T_n-measurable random variables D_n and E_n such that for $j \in I_n$:

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

$$\begin{aligned}
\text{Then } \left|\frac{1}{m_n}\sum_{j\in I_n}\left\langle X_n-\theta, E\left[V_j \mid Tn\right]\right\rangle\right| &\le \|X_n-\theta\|\left(D_n\|X_n-\theta\| + E_n\right)\\
&\le \left(D_n+E_n\right)\|X_n-\theta\|^2 + E_n \text{ a.s.}
\end{aligned} \tag{A5}$$

Part 4. For $|h(x)| \le d|x| + e$ ($d = 0$, $e = 1$ for logistic regression, $d = 1$, $e = 0$ for

linear regression):

$$E\left[\left\|\widetilde{Z}_j\left(h\left(\widetilde{Z}_j'X_n\right)-S_j\right)\right\|^2 \mid T_n\right] \leq E\left[\left\|\widetilde{Z}_j\right\|^2\left(d\left\|\widetilde{Z}_j\right\|\left(\|X_n-\theta\|+\|\theta\|\right)+e+1\right)^2 \mid T_n\right]$$
$$\leq 3d^2E\left[\left\|\widetilde{Z}_j\right\|^4 \mid T_n\right]\left(\|X_n-\theta\|^2+\|\theta\|^2\right)+3\left(e+1\right)^2E\left[\left\|\widetilde{Z}_j\right\|^2 \mid T_n\right] \text{ a.s.}$$

For $\gamma \geqslant 1$,

$$E\left[\left\|\widetilde{Z}_j\right\|^\gamma \mid T_n\right]=E\left[\left\|\widehat{\Gamma}_{n-1}\left(R_j-\widehat{R}_{n-1}\right)\right\|^\gamma \mid T_n\right] \leq 2^{\gamma-1}\left\|\widehat{\Gamma}_{n-1}\right\|^\gamma\left(E\left[\|R\|^\gamma\right]+\left\|\widehat{R}_{n-1}\right\|^\gamma\right) \text{ a.s.}$$

Since $\widehat{\Gamma}_{n-1}$ and $\widehat{R}_{n-1}$ are T_n-measurable and converge respectively to Γ and m and since $\sum_{n=1}^{\infty} a_n^2 < \infty$, there exist two non-negative T_n-measurable random variables F_n and G_n such that for $j \in I_n$:

$$E\left[\left\|\widetilde{Z}_j\left(h\left(\widetilde{Z}_j'X_n\right)-S_j\right)\right\|^2 \mid T_n\right] \leq F_n\left\|X_n-\theta\right\|^2+G_n, \sum_{n=1}^{\infty}a_n^2F_n<\infty, \sum_{n=1}^{\infty}a_n^2G_n<\infty,$$

$$\text{then } E\left[\left\|\frac{1}{m_n}\sum_{j\in I_n}\widetilde{Z}_j\left(h\left(\widetilde{Z}_j'X_n\right)-S_j\right)\right\|^2 \mid T_n\right] \leq F_n\left\|X_n-\theta\right\|^2+G_n \text{ a.s.} \tag{A6}$$

Part 5. Application of Robbins-Siegmund lemma (Lemma A.1).
By (A1) and (A2):

$$\begin{aligned}E\left[\left\|X_{n+1}-\theta\right\|^2 \mid T_n\right] &= \left\|X_n-\theta\right\|^2-2a_n\frac{1}{m_n}\sum_{j\in I_n}\left\langle X_n-\theta, E\left[V_j \mid T_n\right]\right\rangle \\ &+a_n^2E\left[\left\|\frac{1}{m_n}\sum_{j\in I_n}\widetilde{Z}_j\left(h\left(\widetilde{Z}_j'X_n\right)-S_j\right)\right\|^2 \mid T_n\right] \\ &-2a_n\frac{1}{m_n}\sum_{j\in I_n}\left\langle X_n-\theta, E\left[Z_j\left(h\left(Z_j'X_n\right)-h\left(Z_j'\theta\right)\right) \mid T_n\right]\right\rangle \text{ a.s.}\end{aligned} \tag{A7}$$

Since h is an increasing function:

$$\left\langle X_n-\theta, E\left[Z_j\left(h\left(Z_j'X_n\right)-h\left(Z_j'\theta\right)\right) \mid Tn\right]\right\rangle = E\left[\left\|Z_j'\left(X_n-\theta\right)\right\|^2h'\left(\xi_j\right) \mid T_n\right] \geqslant 0 \text{ a.s.} \tag{A8}$$

By (A5,A6,A7,A8):

$$E\left[\|X_{n+1}-\theta\|^2 \mid T_n\right] \leq \|X_n-\theta\|^2\left(1+2a_nD_n+2a_nE_n+a_n^2F_n\right)+2a_nE_n+a_n^2G_n$$
$$-2a_n\frac{1}{m_n}\sum_{j\in I_n}E\left[\left\|Z_j'\left(X_n-\theta\right)\right\|^2 h'\left(\xi_j\right) \mid T_n\right],$$
$$\sum_{n=1}^{\infty}a_nD_n<\infty,\ \sum_{n=1}^{\infty}a_nE_n<\infty,\ \sum_{n=1}^{\infty}a_n^2F_n<\infty,\ \sum_{n=1}^{\infty}a_n^2G_n<\infty \text{ a.s.} \quad \text{(A9)}$$

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

$$\|X_n-\theta\|^2\longrightarrow T,\ \sum_{n=1}^{\infty}a_n\frac{1}{m_n}\sum_{j\in I_n}E\left[\left\|Z_j'\left(X_n-\theta\right)\right\|^2 h'\left(\xi_j\right) \mid T_n\right]<\infty \text{ a.s.} \quad \text{(A10)}$$

Part 6. 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 $\epsilon<\|X_n-\theta\|<\frac{1}{\epsilon}$.

Since for $j\in I_n$, $\xi_j=\lambda_jZ_j'X_n+\left(1-\lambda_j\right)Z_j'\theta=\lambda_jZ_j'\left(X_n-\theta\right)+Z_j'\theta$ with $Z_j=\Gamma R_j^c$, $|\xi_j|\leq\left\|R_j^c\right\|b$, with $b=\|\Gamma\|\left(\frac{1}{\epsilon}+\|\theta\|\right)$.

For logistic regression, $h'\left(u\right)=\frac{e^u}{(1+e^u)^2}$ is an even positive function, decreasing for $u>0$, then $h'\left(\xi_j\right)\geqslant h'\left(\left\|R_j^c\right\|b\right)>0$. For linear regression, $h'\left(u\right)=1$.

Let $\lambda_{\min}\left(A\right)$ denote the lowest eigenvalue of a matrix A; we have for $j\in I_n$:

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

The symmetric matrix $E\left[R^cR^{c\prime}h'\left(\left\|R_j^c\right\|b\right)\right]$ is positive definite since by H1a there is no linear relationship between the components of R^c, thus between the components of $R^c\left(h'\left(\left\|R_j^c\right\|b\right)\right)^{\frac{1}{2}}$; its lowest eigenvalue is strictly positive. By H2, it follows that:

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

This is a contradiction since 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. □

Appendix B. Additional results regarding HOSPHF30D

For HOSPHF30D, all processes AS.P. have a criterion value lower than 0.05 after $300N$ observations used (Table B1).

Table B1. Evolution of the relative norms for HOSPHF30D after $100N$ observations used

Process	$100N$	$200N$	$300N$	$400N$	$500N$
CS1V	0.088	0.045*	0.034*	0.021*	0.023*
CS10V	0.260	0.178	0.115	0.098	0.088
CS100V	0.629	0.499	0.406	0.357	0.322
AS1P50	0.064	0.049*	0.013*	0.016*	0.009*
AS10P50	0.060	0.051	0.013*	0.016*	0.010*
AS100P50	0.067	0.056	0.019*	0.018*	0.013*
AS1P100	0.062	0.045*	0.013*	0.016*	0.009*
AS10P100	0.060	0.048*	0.012*	0.015*	0.009*
AS100P100	0.065	0.055	0.015*	0.016*	0.011*
AS1P200	0.057	0.040*	0.015*	0.015*	0.011*
AS10P200	0.059	0.045*	0.012*	0.015*	0.008*
AS100P200	0.066	0.053	0.013*	0.016*	0.010*

* denotes a criterion value < 0.05.

For HOSPHF30D, processes AS10P. and AS100P. have a criterion value lower than 0.05 at 120s (Table B2).

Table B2. Evolution of the relative norms for HOSPHF30D after 60s of processing time

Processus	**60s**	**120s**	**180s**	**240s**
CS1V	0.153	0.161	0.125	0.064
CS10V	0.228	0.158	0.103	0.087
CS100V	0.326	0.231	0.167	0.141
AS1P50	0.095	0.088	0.064	0.073
AS10P50	0.053	0.034*	0.018*	0.010*
AS100P50	0.014*	0.010*	0.012*	0.014*
AS1P100	0.087	0.086	0.062	0.063
AS10P100	0.050*	0.032*	0.017*	0.009*
AS100P100	0.011*	0.009*	0.014*	0.015*
AS1P200	0.112	0.081	0.062	0.050*
AS10P200	0.049*	0.027*	0.010*	0.009*
AS100P200	0.012*	0.005*	0.015*	0.012*

* denotes a criterion value < 0.05.