One-step corrected projected stochastic gradient descent for statistical estimation

Alexandre Brouste, Youssef Esstafa

One-step corrected projected stochastic gradient descent for statistical estimation

Alexandre Brouste[1*] and Youssef Esstafa[1]

[1]Laboratoire Manceau de Mathématiques, Le Mans Université, France.

*Corresponding author(s). E-mail(s): alexandre.brouste@univ-lemans.fr;
Contributing authors: youssef.esstafa@univ-lemans.fr;

Abstract

A generic, fast and asymptotically efficient method for parametric estimation is described. It is based on the projected stochastic gradient descent on the log-likelihood function corrected by a single step of the Fisher scoring algorithm. We show theoretically and by simulations that it is an interesting alternative to the usual stochastic gradient descent with averaging or the adaptative stochastic gradient descent.

Keywords: Statistical inference, Numerical optimization, Stochastic gradient descent.

1 Introduction

The stochastic gradient descent [1] for finding the root of a given functional is a widely used method in statistical learning. In the parametric estimation setting, this method leads to a (strongly) consistent estimator but which is not asymptotically efficient in term of converging rate or in term of asymptotic variance depending on the conditions retained. For sublinear functionals, consistency has been shown with probabilistic arguments in [2–4] and asymptotic normality in [5–8]. For more general functionals, the ordinary differential equation method has been developed [9, 10] with a boundedness assumption of the random sequence. In order to avoid this quite restrictive assumptions, truncated (or projected) stochastic gradient descent has been proposed [11, 12].

The stochastic gradient descent has been improved in two direction to obtain a statistical procedure with optimal asymptotic rate and variance. On the one hand, the

stochastic gradient with averaging has been studied [13–15]. On the other hand, the adaptative stochastic gradient [16, 17] has been suggested.

In this paper, we propose a fast and asymptotically efficient alternative to averaging or adaptivity. It is based on the one-step procedure.

The one-step procedure was initially considered in [18] for the estimation of parameters in independent and identically distributed (i.i.d.) samples. In this procedure, an initial guess estimator is proposed which is fast to be computed but not asymptotically efficient. Then, a single step of the gradient descent method is done on the log-likelihood function in order to correct the initial estimation and reach asymptotic efficiency. With some recent developments, the one-step procedure has been successfully generalized to more sophisticated statistical experiments as diffusion processes [19, 20], ergodic Markov chains [21], inhomogeneous Poisson and Hawkes counting processes [22, 23], fractional Gaussian and stable noises observed at high frequency [24, 25].

In the following, Section 2 is dedicated to notations and known results of convergence rates for stochastic gradient descent (SGD), stochastic gradient descent with averaging (AVSGD), adaptative gradient descent (ADSGD) and maximum likelihood estimation (MLE). The main result on (strong) consistency and asymptotic normality of the one-step procedure in the multidimensional parameter setting is given in Section 3. Monte Carlo simulations are done in Section 4 to assess the performance of the proposed statistical procedure (OSSGD) in comparison with SGD, AVSGD, ADSGD and MLE in terms of computation time and asymptotic variance for samples of finite size.

2 Notations

In our parametric estimation problem, the observation sample is denoted $X^{(n)} = (X_1, \dots, X_n)$ and is composed of independent and identically distributed random variables. The probability density $f(\cdot, u)$ (with respect to some σ-finite measure) of X_1 is parametrized by $u \in \Theta \subset \mathbb{R}^p$ where Θ is an open set. The true parameter $\vartheta \in \Theta$ is to be estimated.

The estimation problem of the unknown parameter ϑ can be seen as finding the minimum of an unknown function $G(u) = \mathbf{E}_\vartheta\left(-\log f(X_1, u)\right)$ or the root of its gradient

$$M(u) = \mathbf{E}_\vartheta\left(-\nabla_u \log f(X_1, u)\right). \tag{1}$$

The standard statistical procedure to estimate the parameter ϑ is the maximum likelihood estimator (MLE) defined by

$$\widehat{\vartheta}_n = \arg\max_{u\in\Theta} \frac{1}{n}\sum_{i=1}^{n} \log f(X_i, u). \tag{2}$$

Under regularity assumptions, the MLE is consistent, asymptotically normal

$$\sqrt{n}\left(\widehat{\vartheta}_n - \vartheta\right) \Longrightarrow \mathcal{N}(0, \mathcal{I}(\vartheta)^{-1})$$

where $\mathcal{I}(\vartheta)$ stands for the Fisher information matrix

$$\mathcal{I}(\vartheta) = -\mathbf{E}_\vartheta \left[\nabla^2_{u,u} \log f\left(X_1, \vartheta\right)\right] \tag{3}$$

and asymptotically efficient in the minimax sense [26]. Here $\Longrightarrow$ is the convergence in law as $n \to \infty$. But the MLE is generally not in a closed form and its approximation by a classical gradient descent method can be time consuming for large samples. The moment estimator, which is an other generic methodology, when it has closed-form, is generally not asymptotically efficient [26].

Consequently, the Robins-Monro algorithm [1] could be considered to find this root. It is defined recursively by

$$\vartheta_{i+1} = \vartheta_i - \gamma_i(-\nabla_u \log f(X_{i+1}, \vartheta_i)), \quad 0 \leq i \leq n-1,$$

where $(\gamma_i)_i$ is the step sequence and ϑ_0 is the initial value (it may be random but square integrable) of the procedure.

In our estimation problem, the functional M is generally not sublinear and the sequence $(\vartheta_i)_i$ cannot be considered as bounded in probability. For instance, direct computations when the distribution of X_1 is exponential of rate parameter ϑ give

$$M(u) = \frac{1}{\vartheta} - \frac{1}{u}, \quad u > 0.$$

Consequently, projected stochastic gradient descent will be considered in the following. Namely,

$$\vartheta_{i+1} = \Pi_K \left[\vartheta_i - \gamma_i(-\nabla_u \log f(X_{i+1}, \vartheta_i))\right], \quad 0 \leq i \leq n-1, \tag{4}$$

where Π_K is the projection onto the constraint set $K = \{u : a_j \leq u_j \leq b_j\}$ for $-\infty < a_j < b_j < \infty$, $j = 1, \ldots, p$. It can be reformulated as

$$\vartheta_{i+1} = \vartheta_i - \gamma_i(-\nabla_u \log f(X_{i+1}, \vartheta_i)) + \gamma_i Z_i, \quad 0 \leq i \leq n-1, \tag{5}$$

where $\gamma_i Z_i$ is the shortest Euclidian length to take back $\vartheta_i - \gamma_i(-\nabla_u \log f(X_{i+1}, \vartheta_i))$ to the constraint set K if it is not in K.

Under general assumptions, this procedure leads to a strongly consistent estimator [12] for

$$\gamma_i \geq 0, \quad \sum_i \gamma_i^2 < \infty \quad \text{and} \quad \sum_i \gamma_i = \infty, \tag{6}$$

that is $\vartheta_n \longrightarrow \vartheta$ as $n \to \infty$ with probability one. This algorithm is fast but is not asymptotically efficient, neither in terms of converging rate nor in terms of asymptotic variance. For the sequence $\gamma_i = i^{-r}$ and $r \in (1/2, 1)$, it leads to an asymptotically normal estimator for which

$$n^{\frac{r}{2}} \left(\vartheta_n - \vartheta\right) \Longrightarrow \mathcal{N}\left(0, \frac{1}{2} I_p\right) \tag{7}$$

where I_p stands for the $p \times p$ identity matrix. It is worth mentioning that the asymptotic variance does not depend on ϑ in the i.i.d. setting.

It had also been shown that the stochastic gradient descent with $\gamma_i = ci^{-1}$ and

$$c > \frac{1}{2\lambda_{min}(\mathcal{I}(\vartheta))}, \tag{8}$$

where $\lambda_{min}(A)$ is the lowest eigenvalue of the matrix A, is asymptotically rate efficient but is still not asymptotically variance efficient with

$$\sqrt{n}\,(\vartheta_n - \vartheta) \Longrightarrow \mathcal{N}\left(0, c^2\mathcal{I}(\vartheta)(2c\mathcal{I}(\vartheta) - I_p)^{-1}\right).$$

The constraint (8) depends on the unknown parameter and cannot be used in practice.

Consequently, in order to speed up the estimation convergence rate in (7), two methods are classically used: averaging and adaptivity.

2.1 Averaging

The averaging method was proposed (see [12, 13, 15]) to reach variance efficiency with

$$\overline{\vartheta}_n = \frac{1}{n}\sum_{i=1}^{n} \vartheta_i.$$

This estimator is consistent, asymptotically normal with efficient rate and variance, namely

$$\sqrt{n}\left(\overline{\vartheta}_n - \vartheta\right) \Longrightarrow \mathcal{N}(0, \mathcal{I}(\vartheta)^{-1}).$$

2.2 Adaptivity

In the simple setting of i.i.d. samples, the adaptative stochastic gradient descent writes

$$\widetilde{\vartheta}_{i+1} = \widetilde{\vartheta}_i - i^{-1}\mathcal{I}(\widetilde{\vartheta}_i)^{-1}\left(-\nabla_u \log f(X_{i+1}, \widetilde{\vartheta}_i)\right), \quad 0 \leq i \leq n-1.$$

When the classical assumptions are fulfilled, it leads also to a consistent and asymptotical normal estimators with optimal limit variance (see [27] and the references therein), namely

$$\sqrt{n}\left(\widetilde{\vartheta}_n - \vartheta\right) \Longrightarrow \mathcal{N}(0, \mathcal{I}(\vartheta)^{-1}).$$

3 One-step correction

In order to improve the convergence rate of the gradient descent algorithm, we propose in the following the one-step procedure starting from an initial guess estimator taken from the projected stochastic gradient algorithm. This procedure is shown to be faster than the classical computation of the MLE but still asymptotically efficient. It is an interesting alternative to the stochastic gradient algorithm with averaging or adaptative gradient descent and shows nice properties also on samples of finite size.

In the one-step estimation procedure, the estimation ϑ_n given at step n by the projected stochastic gradient descent (2) is corrected by

$$\vartheta_n^* = \vartheta_n + \mathcal{I}(\vartheta_n)^{-1} \cdot \frac{1}{n} \sum_{i=1}^{n} \nabla_u \log f(X_i, \vartheta_n). \tag{9}$$

It leads to a consistent, asymptotically normal and asymptotically efficient estimator of ϑ (see Theorem 1 below).

In the following, we recall the slow convergence of the projected stochastic gradient descent in the multidimensional setting which is the initial guess estimator in the one-step procedure.

Let $Y_i = \nabla_u \log f(X_{i+1}, \vartheta_i)$ and $\mathbf{E}_j$ the conditional expectation with respect to the σ-algebra generated by $\{\vartheta_0, (Y_i, i < j)\}$. Let $\gamma_i = i^{-r}$ and $r \in (1/2, 1)$ in the algorithm (4). The classical assumptions are formulated in [12, Section 10.4 p. 341], namely

A.1 The true value ϑ is in the interior of the constraint set K and $\vartheta_n \to \vartheta$ as $n \to \infty$ with probability one;

A.2 For small $\rho > 0$, $\left\{Y_n \mathbb{I}_{\{|\vartheta_n - \vartheta| \le \rho\}}\right\}$ is uniformly integrable and there is a function g such that for $|\vartheta_n - \vartheta| \le \rho$,

$$\mathbf{E}_n Y_n = g(\vartheta_n);$$

A.3 There exists a constant $0 < C < \infty$ such that for small $\rho > 0$,

$$\sup_n \mathbf{E}_n |Y_n|^2 \mathbb{I}_{\{|\vartheta_n - \vartheta| \le \rho\}} < C \quad \text{w.p.1;}$$

A.4 There is a Hurwitz matrix A such that

$$g(u) = A(u - \vartheta) + o(|u - \vartheta|).$$

The following proposition gives the $n^{\frac{r}{2}}$-consistency of the initial guess estimator which is the first key ingredient in order to prove our next Theorem 1.

Proposition 1 ([12]). *Under aforementioned assumptions, the sequence $n^{\frac{r}{2}} (\vartheta_n - \vartheta)$ is tight.*

Remark 1. *Additive assumptions in order to fullfill A.1 are given in [12, Section 5.2 p 125].*

Remark 2. *Additive assumptions to obtain the asymptotic normality,*

$$n^{\frac{r}{2}} (\vartheta_n - \vartheta) \Longrightarrow \mathcal{N}\left(0, \frac{1}{2} I_p\right), \tag{10}$$

are given in [12, Section 10.2 p. 329].

This algorithm is fast but is not asymptotically efficient, neither in terms of converging rate nor in terms of asymptotic variance. In order to obtain asymptotic normality with optimal rate and variance for the one-step corrected projected stochastic gradient descent, we also suppose that

A.5. The matrix valued function $\mathcal{I}(\vartheta)$ is Lipschitz continuous, *i.e.* there exists a constant $L > 0$ such that

$$\|\mathcal{I}(x) - \mathcal{I}(y)\|_m \leq L\|x - y\|, \quad x, y \in \Theta,$$

where $\|\cdot\|_m$ and $\|\cdot\|$ stand for Euclidean norms in the space of matrices and vectors respectively.

With this condition, we can state the main result:

Theorem 1. *The sequence $(\vartheta_n^*, n \geq 1)$ of one-step estimators of ϑ defined by (9) is consistent and asympotically normal, i.e.*

$$\sqrt{n}(\vartheta_n^* - \vartheta) \Longrightarrow \mathcal{N}(0, \mathcal{I}(\vartheta)^{-1}). \tag{11}$$

It is worth emphasizing that both speed and asymptotic variance are improved in this one-step procedure due to the regularity of the Fisher information matrix.

Proof. The proof is postponed in Appendix A. □

4 Simulations

The joint estimation of the shape parameter α and scale parameter β is considered in the statistical experiment generated by a sample $X^{(n)} = (X_1, X_2, \ldots, X_n)$ of i.i.d. Gamma random variables whose probability density function is given by

$$f(x) = \frac{\beta^\alpha}{\Gamma(\alpha)} x^{\alpha-1} \exp(-\beta x), \quad x > 0.$$

Let us denote $\vartheta = (\alpha, \beta)$. In this statistical experiment, the sequence of maximum likelihood estimators $(\widehat{\vartheta}_n)_{n\geq 1}$ of ϑ is not in a closed-form. The sequence of MLE satisfies

$$\sqrt{n}\left(\widehat{\vartheta}_n - \vartheta\right) \to \mathcal{N}\left(0, \mathcal{I}(\vartheta)^{-1}\right),$$

where

$$\mathcal{I}(\vartheta) = \begin{pmatrix} \psi^{(2)}(\alpha) & -\frac{1}{\beta} \\ -\frac{1}{\beta} & \frac{\alpha}{\beta^2} \end{pmatrix}.$$

Here, $\psi^{(n)}$ is the polygamma functions (see [28, Section 6.4.1, page 260]) defined by $\psi^{(n)}(\alpha) = \frac{\partial^n}{\partial \alpha^n} \log \Gamma(\alpha)$.

The different estimators (MLE, SGD, OSSGD, AVSGD, ADSGD) have been compared in terms of variance and computation time on $B = 2 \times 10^3$ Monte Carlo simulations for samples of size $n = 2 \times 10^4$. The SGD is done with $\gamma_i = i^{-r}$ where r is chosen to be equal to 0.6. It is worth mentioning that the results are similar for all values of $\frac{1}{2} < r < 1$.

We can see on Figure 1 that the optimal variance is reached for the OSSGD (as for the MLE, AVSGD and ADSGD) that naturally overperforms the non-optimal variance of the slowly converging SGD. It is worth noting the relative bias for samples of finite size of the AVSGD when the initial value ϑ_0 is fixed.

In terms of computation time, the OSSGD (as the AVSGD) is more than 3 times faster than the MLE. In comparison, the ADSGD is more than two times faster.

For these reasons, the fast and asymptotically efficient OSSGD is a proper alternative to the averaged and the adapted stochastic gradient descent methods.

	MLE	SGD	OSSGD	AVSGD	ADSGD
time (s)	198.22	63.87	64.25	64.22	88.20

It can also be noticed that, for the specific case of the estimation of the parameters in the Gamma distribution, moment estimators [29] or other original explicit estimators [30] could have been considered as initial guess estimation in the one-step procedure instead of the SGD.

5 Conclusion

In this paper, we propose to apply the one-step procedure to the slowly converging stochastic gradient descent in order to improve the convergence rate and reach asymptotical efficiency. It is a fast and asymptotically efficient alternative to averaging or adaptivity.

The one-step procedure for the stochastic gradient descent is considered here in the i.i.d. setting but it will be extended in a further work to the regression setting (linear regression, logistic regression (see also [31] for an adaptative procedure), generalized linear models) for larger applications.

A One-step procedure

For an observation sample $(X_1, \ldots, X_n)$, let us denote $\ell_n(u) = \sum_{i=1}^n \log f(X_i, u)$. Recall that ϑ is the true parameter and

$$\vartheta_n^* = \vartheta_n + \mathcal{I}(\vartheta_n)^{-1} \cdot \frac{1}{n} \nabla_u \ell_n(\vartheta_n), \quad n \geq 1. \tag{12}$$

Consistency:

The consistency of the sequence of initial guess estimators gives, as $n \to \infty$, $\vartheta_n \longrightarrow \vartheta$ in probability. Since $\mathbf{E}_\vartheta \nabla_u \ell_n(\vartheta) = 0$, the uniform law of large numbers gives, as $n \to \infty$,

$$\frac{1}{n} \nabla_u \ell_n(\vartheta_n) \longrightarrow 0_{\mathbb{R}^p}$$

in probability. The uniform continuity of the Fisher information matrix gives the result. Since the initial stochastic gradient descent is also strongly consistent [12], we can also obtain the strong consistency with the strong law of large numbers.

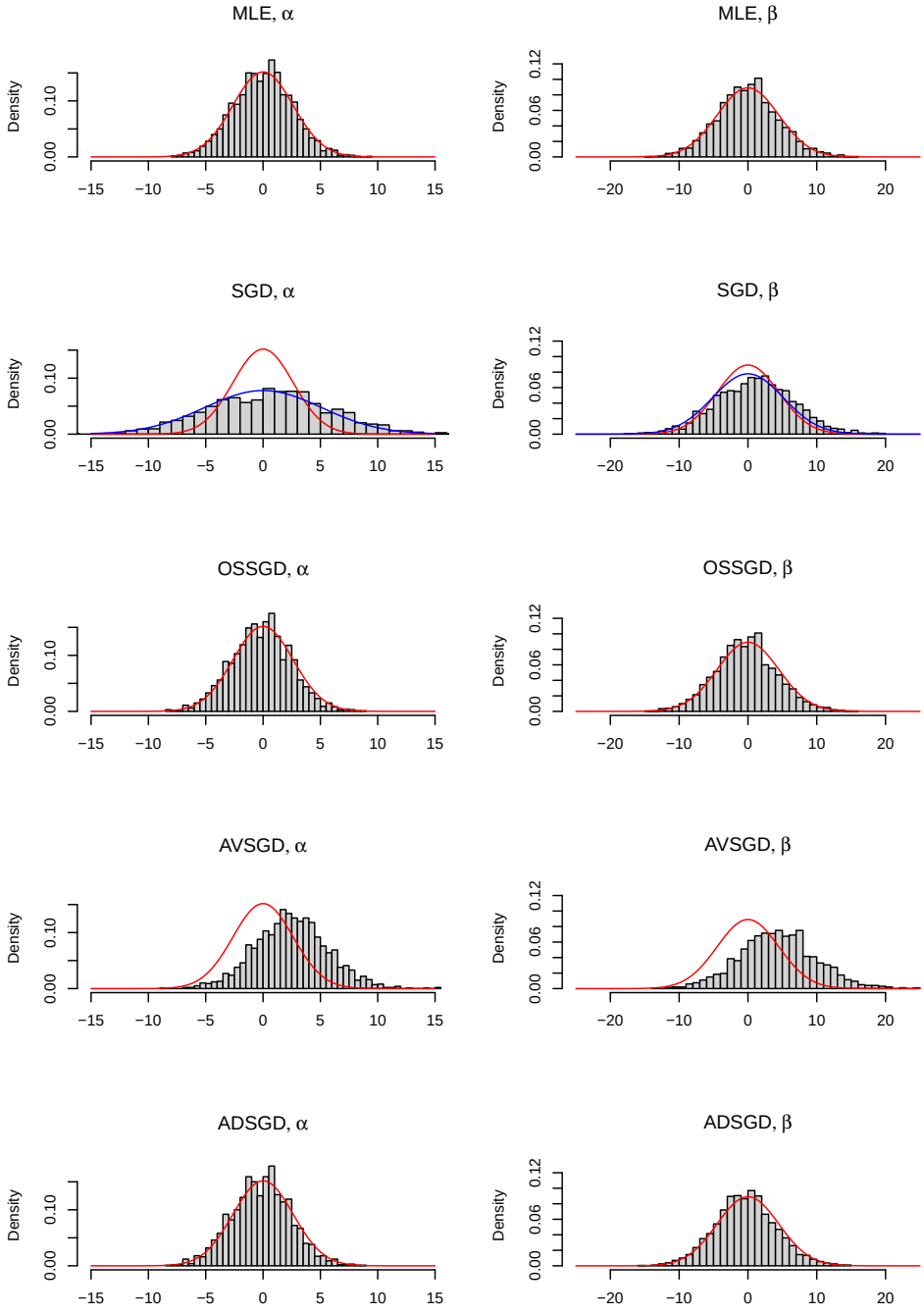

Fig. 1 Statistical errors renormalized by $\sqrt{n}$ for MLE, SGD, OSSGD, AVSGD and ADSGD for $n = 2 \times 10^4$ and $B = 2 \times 10^3$ Monte-Carlo simulations. Theoretical optimal variance (in red) and variance of SGD (in blue) are superimposed.

Asymptotic normality:
From (12), we have

$$\sqrt{n}\,(\vartheta_n^* - \vartheta) = \sqrt{n}\,(\vartheta_n - \vartheta) + \mathcal{I}(\vartheta_n)^{-1} \cdot \frac{1}{\sqrt{n}} \nabla_u \ell_n(\vartheta_n).$$

The mean-value theorem gives

$$\nabla_u \ell_n(\vartheta_n) = \nabla_u \ell_n(\vartheta) + \int_0^1 \nabla^2_{u,u} \ell_n\left(\vartheta + \tau(\vartheta_n - \vartheta)\right) d\tau \cdot (\vartheta_n - \vartheta)$$

and

$$\begin{aligned}\sqrt{n}\,(\vartheta_n^* - \vartheta) = n^{\frac{r}{2}} \left\{ I_p + \mathcal{I}(\vartheta_n)^{-1} \frac{1}{n} \int_0^1 \nabla^2_{u,u} \ell_n\left(\vartheta + \tau(\vartheta_n - \vartheta)\right) d\tau \right\} n^{\frac{r}{2}}\,(\vartheta_n - \vartheta)\, n^{\frac{1}{2}-r} \\ + \mathcal{I}(\vartheta_n)^{-1} \cdot \frac{1}{\sqrt{n}} \nabla_u \ell_n(\vartheta), \qquad (13)\end{aligned}$$

where I_p is the $p \times p$ identity matrix.
The central limit theorem gives, as $n \to \infty$,

$$\frac{1}{\sqrt{n}} \nabla_u \ell_n(\vartheta) \Longrightarrow \mathcal{N}\left(0, \mathcal{I}(\vartheta)\right)$$

in law and the proper convergence of the second term in the r.h.s. of Equation (13).

Considering the first right-hand term, we have that $(\vartheta_n)_{n\geq 1}$ is $n^{\frac{r}{2}}$-consistent by assumption and $n^{\frac{1}{2}-r} \to 0$, as $n \to \infty$, for $\frac{1}{2} < r \leq 1$. Then, we need to show that

$$n^{\frac{r}{2}} \left(I_p + \mathcal{I}(\vartheta_n)^{-1} \frac{1}{n} \int_0^1 \nabla^2_{u,u} \ell_n\left(\vartheta + \tau(\vartheta_n - \vartheta)\right) d\tau \right) = n^{\frac{r}{2}} A$$

is bounded in probability as $n \to \infty$ with

$$\begin{aligned} A &= \mathcal{I}(\vartheta_n)^{-1} \left(\mathcal{I}(\vartheta_n) + \frac{1}{n} \int_0^1 \nabla^2_{u,u} \ell_n\left(\vartheta + \tau(\vartheta_n - \vartheta)\right) d\tau \right) \\ &= \mathcal{I}(\vartheta_n)^{-1} \cdot \left(\left[\mathcal{I}(\vartheta_n) - \mathcal{I}(\vartheta)\right] + \left[\frac{1}{n} \nabla^2_{u,u} \ell_n(\vartheta) + \mathcal{I}(\vartheta) \right] \right. \\ &\quad \left. + \frac{1}{n} \int_0^1 \left[\frac{\partial^2}{\partial \vartheta^2} \ell_n\left(\vartheta + \tau(\vartheta_n - \vartheta)\right) - \nabla^2_{u,u} \ell_n(\vartheta) \right] d\tau \right). \end{aligned}$$

The second terms in the r.h.s. converges to zero at rate $\sqrt{n}$. The Lipschitz continuity of the Fisher information allows to control the first and third terms by $C\|\vartheta_n - \vartheta\|$ where C is a generic constant. Since $(\vartheta_n)_{n\geq 1}$ is $n^{\frac{r}{2}}$-consistent, the quantity $n^{\frac{r}{2}} A$ is bounded in probability as $n \to \infty$. The Slutsky theorem gives the final result.

Acknowledgments

We would like to thank Alain Bensoussan for all the fruitful discussions on this work. This research partially benefited from the support of the ANR project 'Efficient inference for large and high-frequency data' (ANR-21-CE40-0021) and the 'Efficience et Sobriété Numériques' research program under the aegis of Fondation du Risque, a joint initiative by Le Mans University and EREN Groupe.

References

[1] Robbins, H., Monro, S.: A stochastic approximation method. Ann. Math. Statist. **22**, 400–407 (1951) https://doi.org/10.1214/aoms/1177729586

[2] Wolfowitz, J.: On the stochastic approximation method of Robbins and Monro. Ann. Math. Statist. **23**, 457–461 (1952) https://doi.org/10.1214/aoms/1177729391

[3] Blum, J.R.: Approximation methods which converge with probability one. Ann. Math. Statist. **25**, 382–386 (1954) https://doi.org/10.1214/aoms/1177728794

[4] Goodsell, C., Hanson, D.: Almost sure convergence for the Robbins-Monro process. The Annals of Probability **4**(6), 890–901 (1976)

[5] Chung, K.L.: On a stochastic approximation method. Ann. Math. Statist. **25**, 463–483 (1954) https://doi.org/10.1214/aoms/1177728716

[6] Hodges, J.L. Jr., Lehmann, E.L.: Two approximations to the Robbins-Monro process. In: Proc. Third Berkeley Symp. Math. Statist. Prob., vol. 1, pp. 95–104 (1956)

[7] Sacks, J.: Asymptotic distribution of stochastic approximation procedures. Ann. Math. Statist. **29**, 373–405 (1958) https://doi.org/10.1214/aoms/1177706619

[8] Fabian, V.: On asymptotic normality in stochastic approximation. Ann. Math. Statist. **39**, 1327–1332 (1968) https://doi.org/10.1214/aoms/1177698258

[9] Ljung, L.: Analysis of recursive stochastic algorithms. IEEE Transactions on Automatic Control **22**(4), 551–575 (1977)

[10] Kushner, H., Clark, D.: Stochastic Approximation for Constrained and Unconstrained Systems. Springer, New-York (1978)

[11] Chen, H.-F.: Recent developments in stochastic approximation. 13th Triennal World Congress, 1815–1820 (1996)

[12] Kushner, H., Yin, G.: Stochastic Approximation and Recursive Algorithms and Applications. Springer, New-York (2003)

[13] Ruppert, D.: Efficient estimations from a slowly convergent Robbins-Monro process. Technical report, Cornell University (1988)

[14] Polyak, B.T.: New stochastic approximation type procedure. Automat. Remote Control **51**(7) (1990)

[15] Polyak, B.T., Juditsky, A.B.: Acceleration of stochastic approximation by averaging. SIAM J. Control Optim. **30**(4), 838–855 (1992) https://doi.org/10.1137/0330046

[16] Lai, T.L., Robbins, H.: Adaptive design and stochastic approximation. The Annals of Statistics **7**(6), 1196–1221 (1979) https://doi.org/10.1214/aos/1176344840

[17] Venter, J.H.: An extension of the Robbins-Monro procedure. Ann. Math. Statist. **38**, 181–190 (1967) https://doi.org/10.1214/aoms/1177699069

[18] Le Cam, L.: On the asymptotic theory of estimation and testing hypotheses. In: Proc. Third Berkeley Symp. Math. Statist. Prob., vol. 1, pp. 355–368 (1956)

[19] Gloter, A., Yoshida, N.: Adaptive estimation for degenerate diffusion processes. Electron. J. Stat. **15**(1), 1424–1472 (2021) https://doi.org/10.1214/20-ejs1777

[20] Kamatani, K., Uchida, M.: Hybrid multi-step estimators for stochastic differential equations based on sampled data. Stat. Inference Stoch. Process. **18**(2), 177–204 (2015) https://doi.org/10.1007/s11203-014-9107-4

[21] Kutoyants, Y.A., Motrunich, A.: On multi-step MLE-process for Markov sequences. Metrika **79**(6), 705–724 (2016) https://doi.org/10.1007/s00184-015-0574-4

[22] Brouste, A., Farinetto, C.: Fast and asymptotically efficient estimation in the Hawkes processes. Jpn. J. Stat. Data Sci. **6**(1), 361–379 (2023) https://doi.org/10.1007/s42081-023-00186-2

[23] Dabye, A.S., Gounoung, A.A., Kutoyants, Y.A.: Method of moments estimators and multi-step MLE for Poisson processes. Izv. Nats. Akad. Nauk Armenii Mat. **53**(4), 31–45 (2018)

[24] Brouste, A., Masuda, H.: Efficient estimation of stable Lévy process with symmetric jumps. Stat. Inference Stoch. Process. **21**(2), 289–307 (2018) https://doi.org/10.1007/s11203-018-9181-0

[25] Brouste, A., Soltane, M., Votsi, I.: One-step estimation for the fractional Gaussian noise at high-frequency. ESAIM Probab. Stat. **24**, 827–841 (2020) https://doi.org/10.1051/ps/2020022

[26] Ibragimov, I., Has'minskii, R.: Statistical Estimation: Asymptotic Theory.

Springer, New-York (1981)

[27] Amari, S.: Natural gradient works efficiently in learning. Neural Computation **10**(2), 251–276 (1998) https://direct.mit.edu/neco/article-pdf/10/2/251/813415/089976698300017746.pdf

[28] Abramowitz, M., Stegun, I.A. (eds.): Handbook of Mathematical Functions with Formulas, Graphs, and Mathematical Tables, p. 1046. Dover Publications Inc., New-York (1992)

[29] Brouste, A., Dutang, C., Mieniedou, D.N.: The R journal: OneStep : Le Cam's one-step estimation procedure. The R Journal **13**, 383–394 (2021) https://doi.org/10.32614/RJ-2021-044

[30] Ye, Z.-S., Chen, N.: Closed-form estimators for the gamma distribution derived from likelihood equations. Amer. Statist. **71**(2), 177–181 (2017) https://doi.org/10.1080/00031305.2016.1209129

[31] Bercu, B., Godichon, A., Portier, B.: An efficient stochastic Newton algorithm for parameter estimation in logistic regressions. SIAM J. Control Optim. **58**(1), 348–367 (2020) https://doi.org/10.1137/19M1261717