Hindawi
Journal of Mathematics
Volume 2021, Article ID 7660113, 7 pages
https://doi.org/10.1155/2021/7660113

Research Article

Estimation of the Conditional Probability Using a Stochastic Gradient Process

Ali Labriji, Abdelkrim Bennar, and Mostafa Rachik

Department of Mathematics and Computer Science, Faculty of Sciences, Ben M'sik Hassan II University, Casablanca, Morocco

Correspondence should be addressed to Ali Labriji; alilabriji@gmail.com

Received 22 August 2021; Revised 3 November 2021; Accepted 13 November 2021; Published 6 December 2021

Academic Editor: Barbara Martinucci

The use of conditional probabilities has gained in popularity in various fields such as medicine, finance, and imaging processing. This has occurred especially with the availability of large datasets that allow us to extract the full potential of the available estimation algorithms. Nevertheless, such a large volume of data is often accompanied by a significant need for computational capacity as well as a consequent compilation time. In this article, we propose a low-cost estimation method: we first demonstrate analytically the convergence of our method to the desired probability and then we perform a simulation to support our point.

1. Introduction

The likelihood that an event B will occur knowing that event A has already occurred is called the conditional probability, denoted by $P(B|A)$ or $P_A(B)$. For example, if a card is randomly drawn from a deck, there is a one in four chance of getting a heart suit, but if a red reflection is seen on the table, there is now a one in two chance of getting it. If events A and B have nonzero probabilities, then Bayes theorem states that $P(B|A) = P(A \cap B)/P(A)$. That was for the scientific part, but in daily life also conditional probability is useful in various fields and is even gaining more and more interest. For example, banks estimate the probability of default of a borrower or bond issuer using conditional probability estimation methods based on Basel II regulations (see [1] for more information). The estimation of this probability is crucial since it allows the banks to compute the expected losses and therefore to cover the consequences. Another area where the estimation of conditional probabilities is important is marketing, where it is used to estimate the interest of a customer in a given product or service. Therefore, they are able to focus on the most attractive population in order to optimize the marketing costs [2]. The estimation of this probability is also frequently used in the field of medicine, as doctors need to estimate the likelihood of a patient being affected by a given disease based on the symptoms the patient presents [3] and many more areas, such as drug discovery, computer vision, speech recognition, handwriting recognition, biometric identification, document classification, Internet search engines, pattern recognition, and recommender system [4–11].

In practice, we can divide conditional probability estimation methods into two categories, linear and nonlinear classifiers. The linear classifiers can be split into two subcategories, the generative and discriminative models [12, 13], and the most commonly used are

(i) Fisher's linear discriminant
(ii) Logistic regression
(iii) Naive Bayes classifier

Nonlinear classifiers can be grouped into the following groups of methods:

(i) Linear classifier with transformed data such as a discretization of continuous variables
(ii) Support vector machines
(iii) Quadratic classifiers
(iv) K-nearest neighbor
(v) Decision trees

(vi) Neural networks

(vii) Learning vector quantization

To learn more about these different algorithms, see [14–20].

Let us consider an observable random binary variable U and a random variable V. We define U such that

$$U = \begin{cases} 1, & \text{if the studied event occured,} \\ 0, & \text{if not.} \end{cases} \quad (1)$$

We are willing to estimate the vector θ such that the conditional probability $P(U = 1|V)$ is written in the form:

$$P(U = 1|V) = \frac{1}{1 + \exp(-\theta' V)}. \quad (2)$$

We are looking for a simple method of estimating the parameter θ that will be less demanding in terms of computational capacity. This is useful especially in the Big Data era, where the datasets can be massive and any common iterative estimation can take a lot of time. To do this, we use the stochastic approximation, which has been introduced by Herbert Robbins and Sutton Monro in 1951 [21]. The goal is to find the unique root θ^* of a function $M(\theta) = \alpha$, while $M(\theta)$ cannot be directly observed. Yet, we assume that we can observe a variable $Y(\theta)$ such that $E[Y(\theta)] = M(\theta)$. According to [21], there exists a sequence a_n that satisfies

$$\sum_{n=1}^{\infty} a_n = \infty, \sum_{n=1}^{\infty} a_n^2 < \infty, \quad (3)$$

such as the process θ_n defined by

$$\theta_{n+1} = \theta_n - a_n (M(\theta_n) - \alpha), \quad (4)$$

converges to the unique root of $M(\theta) = \alpha$. In our case, we start from the work of Bennar et al. [22] who established the conditions for almost sure convergence, as well as the quadratic mean convergence of a stochastic gradient process θ_n to the parameter θ that allows us to estimate $E[U|V]$. Here, we are interested in the case of binary random variables, where $E[U|V]$ is equivalent to $P(U = 1|V)$, as we can see in the following:

$$E[U|V] = \sum_{i=1}^{2} u_i P(U = u_i|V) = P(U = 1|V). \quad (5)$$

We also chose these results as the basis of our work since the stochastic gradient process performs a sampling at each iteration in order to achieve the estimates without relying on all the available data.

In this article, we first present the convergence results elaborated by Bennar et al., then we show that these results are also valid in the framework of estimating the conditional probability. We also present a simulation to highlight the obtained results, and finally, we conclude our work by addressing development perspectives.

2. Preliminaries

Let us consider an observable random variable U and a random variable V, both have values in $\mathbb{R}^k$ of law μ. We try to estimate the parameter θ in $\mathbb{R}^p$ such that $\phi(V, \theta)$ approaches $E[U|V]$ in the least squares sense. It should also be noted that the estimation of the parameters of a logistic regression in the sense of least squares is already achieved through the iterative weighted least squares method [23] which, unlike our purpose, is heavy and employs huge computing capacities in the case of large dataset.

Let f be the real positive function defined in $\mathbb{R}^p$ by

$$f(\theta) = E\left[(E[U|V] - \phi(V, \theta))^2\right], \quad (6)$$

we are looking for the value of θ that minimizes the function f.

Let us define the real positive function g in $\mathbb{R}^p$ by

$$g(\theta) = E\left[(U - \phi(V, \theta))^2\right]. \quad (7)$$

We have

$$g(\theta) = f(\theta) + E\left[(U - E[U|V])^2\right], \quad (8)$$

thus the problem reduces to looking for θ that minimizes the function g. We have

$$\nabla_\theta g(\theta) = 2E[\nabla_\theta \phi(V, \theta)(\phi(V, \theta) - U)]. \quad (9)$$

To estimate θ in a sequential way, we use a stochastic gradient algorithm. We consider a random θ_n in $\mathbb{R}^p$ defined by

$$\theta_{n+1} = \theta_n - a_n \nabla_\theta \phi(V_n, \theta_n)(\phi(V_n, \theta_n) - U_n), \quad (10)$$

with

(i) (a_n) is a sequence of positive real numbers

(ii) $(U_1, V_1), (U_2, V_2), \ldots, (U_n, V_n)$ is a sample of independent random variable couples with the same probability law that (U, V)

(iii) $\phi(.,.)$ is a real known measurable function in $\mathbb{R}^k \times \mathbb{R}^p$

In the following, the abbreviation *a.s* means almost sure convergence and *q.m* means quadratic mean convergence.

2.1. Almost Sure Convergence.

Bennar et al. have considered the following assumptions:

(H_1) $a_n > 0, \sum_{n=1}^{\infty} a_n^2 < \infty$,

(H_1') $a_n > 0, \sum_{n=1}^{\infty} a_n = \infty, \sum_{n=1}^{\infty} a_n^2 < \infty$,

(H_2): there exist a and b such that for all $\theta = (\theta_1, \theta_2, \ldots, \theta_p)' \in \mathbb{R}^p$,

$$\operatorname{Var}\left[\frac{\partial \phi(V,\theta)}{\partial \theta_i}(\phi(V,\theta) - U)\right] < ag(\theta) + b, \quad \text{for all } i = 1, 2, \ldots, p, \tag{11}$$

(H_3): there exists $K > 0$ such that for all $\theta = (\theta_1, \theta_2, \ldots, \theta_p)' \in \mathbb{R}^p$,

$$\frac{\partial^2 g(\theta)}{\partial \theta_i \partial \theta_j} < K, \quad \text{for all } i = 1, 2, \ldots, p, \tag{12}$$

$(H_4)\, \theta^*$ is a local minimum of g:

$$\exists\, \alpha > 0\colon \left(\theta \neq \theta^*, \left\|\theta - \theta^*\right\| < \alpha \Rightarrow g(\theta^*) < g(\theta)\right). \tag{13}$$

$(H_5)\, \theta^*$ is the unique stationary point of g:

$$\forall \theta \in \mathbb{R}^p, (\theta \neq \theta^*) \Leftrightarrow \nabla_\theta g(\theta) \neq 0). \tag{14}$$

Lemma 1. *Under the assumptions H_1', H_2, H_3, H_4, H_5, we have*

$$\theta_n \longrightarrow \theta^* \text{ a.s or } \left\|\theta_n\right\| \longrightarrow +\infty \text{ a.s.} \tag{15}$$

Proof. See [22]. □

2.2. Quadratic Mean Convergence. Bennar et al. have considered the following assumptions:

$(H_6)\phi(V,\theta)$ and $\nabla_\theta \phi(V,\theta)$ are uniformly bounded in V and θ.

(H_7): there exist two real positive functions h and h' defined in $\mathbb{R}^p$ such that

$\forall \theta,\ \theta' \in \mathbb{R}^p,\ \forall V \in \mathbb{R}^p$,

$$\begin{aligned} |\phi(V,\theta) - \phi(V,\theta')| &\leq h(V)\|\theta - \theta'\|, \\ \left\|\nabla_\theta \phi(V,\theta) - \nabla_{\theta'} \phi(V,\theta')\right\| &\leq h'(V)\|\theta - \theta'\|, \\ E[h(V)] < \infty, \quad & E[h'(V)] < \infty. \end{aligned} \tag{16}$$

$(H_8)U$ is a real random bounded variable.

Lemma 2. *Under the assumptions H_1', H_3, H_6, H_7, H_8, we have*

$$\nabla_\theta g(\theta_n) \longrightarrow 0 \text{ a.s and } \nabla_\theta g(\theta_n) \longrightarrow 0\ q.m. \tag{17}$$

Proof. See [22]. □

3. Application

3.1. Proof of Process Convergence. Let us assume $\rho_1, \rho_2, \ldots, \rho_p$ be p functions of q measurable real variables. We note

$$\rho = \left(\rho_1, \rho_2, \ldots, \rho_p\right)'. \tag{18}$$

In order to estimate the value of θ that minimizes $E[(E[U|V] - (1/1 + \exp(-\rho(V)'\theta)))^2]$, we consider the following stochastic approximation process (θ_n) in $\mathbb{R}^p$ defined by

$$\theta_{n+1} = \theta_n - a_n \frac{\rho(V_n)\exp\left(-\rho(V_n)'\theta_n\right)}{\left(1 + \exp\left(-\rho(V_n)'\theta_n\right)\right)^2}\left(\frac{1}{1 + \exp\left(-\rho(V_n)'\theta_n\right)} - U_n\right), \tag{19}$$

with

$$\frac{\rho(V_n)\exp\left(-\rho(V_n)'\theta_n\right)}{\left(1+\exp\left(-\rho(V_n)'\theta_n\right)\right)^2}=\begin{pmatrix}\dfrac{\rho_1\left(V_{1,n}\right)\exp\left(-\rho(V_n)'\theta_n\right)}{\left(1+\exp\left(-\rho(V_n)'\theta_n\right)\right)^2}\\ \vdots \\ \dfrac{\rho_p\left(V_{p,n}\right)\exp\left(-\rho(V_n)'\theta_n\right)}{\left(1+\exp\left(-\rho(V_n)'\theta_n\right)\right)^2}\end{pmatrix}, \tag{20}$$

where $(U_1, V_1), (U_2, V_2), \ldots, (U_n, V_n)$ is a sample of (U, V) formed of independent random variable and distributed identically.

We assume the following assertions:

$(H_9)\, \rho_1(V_1), \rho_2(V_2), \ldots, \rho_p(V_p)$ are observed in a finite way

$(H_{10})\, \theta_1$ is a random variable such that $E[\|\theta_1\|^2] < \infty$

Theorem 3. *Under the assumptions H_1', H_9, H_{10}, we have*

$$\nabla_\theta g(\theta_n) \longrightarrow 0\, a.s \text{ and } \nabla_\theta g(\theta_n) \longrightarrow 0\, q.m. \tag{21}$$

Proof. Let ϕ be the real function of $\mathbb{R}^p \times \mathbb{R}^p$ defined by

$$\phi(V,\theta)=\frac{1}{1+\exp(-\rho(V)'\theta)}=\frac{1}{1+\exp\left(-\sum_{j=1}^{p}\theta_j\rho_j(V)\right)}. \tag{22}$$

Let us prove that the assumption 3 is true.

We have

$$g(\theta)=E\left[\left(U-\frac{1}{1+\exp(-\rho(V)'\theta)}\right)^2\right]. \tag{23}$$

For $i = 1, 2, \ldots, p$, we have

$$\begin{aligned}\frac{\partial g(\theta)}{\partial\theta_i}&=-2E\left[\frac{\rho_i(V_i)\exp(-\rho(V)'\theta)}{(1+\exp(-\rho(V)'\theta))^2}\left(U-\frac{1}{1+\exp(-\rho(V)'\theta)}\right)\right],\\&=-2E\left[\frac{\rho_i(V_i)}{(2+\exp(\rho(V)'\theta)+\exp(-\rho(V)'\theta))}\left(U-\frac{1}{1+\exp(-\rho(V)'\theta)}\right)\right].\end{aligned} \tag{24}$$

Thus, for $i, j = 1, 2, \ldots, p$, we have

$$\begin{aligned}\frac{\partial^2 g(\theta)}{\partial\theta_i\partial\theta_j}&=2E\left[\frac{\rho_i(V_i)\rho_j\left(V_j\right)(\exp(\rho(V)'\theta)-\exp(-\rho(V)'\theta))}{(2+\exp(\rho(V)'\theta)+\exp(-\rho(V)'\theta))^2}\left(U-\frac{1}{1+\exp(-\rho(V)'\theta)}\right)\right]\\&\quad+2\mathrm{E}\left[\frac{\rho_i(V_i)}{(2+\exp(\rho(V)'\theta)+\exp(-\rho(V)'\theta))}\left(\frac{\rho_j(Vj)}{\left(2+\exp(\rho(V)'\theta)+\exp(-\rho(V)'\theta)\right)}\right)\right]\\&\cdot\left|\frac{\partial^2 g(\theta)}{\partial\theta_i\partial\theta_j}\right|\\&\le 2E\left[\left|\rho_i(V_i)\rho_j\left(V_j\right)\left(U-\frac{1}{1+\exp(-\rho(V)'\theta)}\right)\right|\right]\\&\quad+2\mathrm{E}\left[\left|\rho_i(V_i)\rho_j(Vj)\left(\frac{1}{(2+\exp(\rho(V)'\theta)+\exp(-\rho(V)'\theta))}\right)\right|\right],\end{aligned} \tag{25}$$

as the $\rho_1(V_1), \rho_2(V_2), \ldots, \rho_p(V_p)$ are observed in a finite way, and

$$0<\frac{1}{1+\exp(-\rho(V)'\theta)}<1,\ 0<\frac{1}{(2+\exp(\rho(V)'\theta)+\exp(-\rho(V)'\theta))}<1. \tag{26}$$

Then, there exists $K > 0$ such that for all $\theta = (\theta_1, \theta_2, \ldots, \theta_p)' \in \mathbb{R}^p$,

$$\frac{\partial^2 g(\theta)}{\partial\theta_i \partial\theta_j} < K, \quad \text{for all}\, i = 1, 2, \ldots, p. \tag{27}$$

Let us prove that the assumption 6 is true.

We have

$$\phi(V, \theta) = \frac{1}{1 + \exp(-\rho(V)'\theta)},$$
$$\nabla_\theta \phi(V, \theta) = \frac{\rho(V)}{(2 + \exp(\rho(V)'\theta) + \exp(-\rho(V)'\theta))}, \tag{28}$$

with $\rho(V)/(2 + \exp(\rho(V)'\theta) + \exp(-\rho(V)'\theta)) = \begin{pmatrix} \rho_1(V_1)/(2 + \exp(\rho(V)'\theta) + \exp(-\rho(V)'\theta)) \\ \vdots \\ \rho_p(V_p)/(2 + \exp(\rho(V)'\theta) + \exp(-\rho(V)'\theta)) \end{pmatrix}$;

then, $\phi(V, \theta) < 1$ and $\|\nabla_\theta \phi(V, \theta)\| < \|\rho(V)\|$, and since $\rho_1(V_1), \rho_2(V_2), \ldots, \rho_p(V_p)$ are observed in a finite way, then there exists $K_2 > 0$ such that for all $\theta = (\theta_1, \theta_2, \ldots, \theta_p)' \in \mathbb{R}^p$ and $V = (V_1, V_2, \ldots, V_p)' \in \mathbb{R}^p$, $\|\nabla_\theta \phi(V, \theta)\| < K_2$. Then, $\phi(V, \theta)$ and $\nabla_\theta \phi(V, \theta)$ are uniformly bounded in V and θ.

Let us prove that the assumption 7 is true. To do this, we use the following result. □

Lemma 4 (mean value inequalities). *Let E and F be two real normed vector spaces, U an open of E, and $f: U \longrightarrow F$ a differentiable application. For any segment $[a, b]$ included in U, we have*

$$\|f(b) - f(a)\|_F \leq \sup_{x \in [a,b]} \left(\left\|f'(x)\right\|\right) \|b - a\|_E, \tag{29}$$

where, for any point x of U, $\|f'(x)\|$ is the operator norm of the differential of f at point x.

Proof. See [24], p. 31.

Then, there exist two real positive functions h and h' defined in $\mathbb{R}^p$ such that

$\forall\theta,\ \theta\prime \in \mathbb{R}^p,\ \forall V \in \mathbb{R}^p,$

$$|\phi(V, \theta) - \phi(V, \theta\prime)| \leq h(V)\|\theta - \theta\prime\|,$$
$$\left\|\nabla_\theta \phi(V, \theta) - \nabla_{\theta'} \phi(V, \theta\prime)\right\| \leq h'(V)\|\theta - \theta\prime\|. \tag{30}$$

Let us prove that $E[h(V)] < \infty$, and $E[h'(V)] < \infty$.

We have already seen that $\|\nabla_\theta \phi(V, \theta)\| \leq \|\rho(V)\|$ and since

$\rho_1(V_1), \rho_2(V_2), \ldots, \rho_p(V_p)$ are observed in a finite way, then $E[h(V)] < \infty$.

Furthermore, we have that $\partial^2 \phi(V, \theta)/\partial\theta_i \partial\theta_j = -\rho_i(V_i)\rho_j(V_j)\ (\exp(\rho(V)'\theta) - \exp(-\rho(V)'\theta))/(2 + \exp(\rho(V)'\theta) + \exp(-\rho(V)'\theta))^2$, then $|\partial^2 \phi(V, \theta)/\partial\theta_i \partial\theta_j| \leq |\rho_i(V_i)\rho_j(V_j)|$, and since $\rho_1(V_1), \rho_2(V_2), \ldots, \rho_p(V_p)$ are observed in a finite way, then $E[h'(V)] < \infty$.

Moreover, since U is a binary random variable, then assumption 8 is true.

Table 1: Fisher scoring algorithm outputs.

	Dependent variable
	U
θ	0.567*** (0.013)
Constant	0.012 (0.040)
	Observations
Observations	10,000
Log likelihood	−2,091.211
Akaike Inf. Crit.	4,186.422

Note: $^{*}p < 0.1$; $^{**}p < 0.05$; $^{***}p < 0.01$.

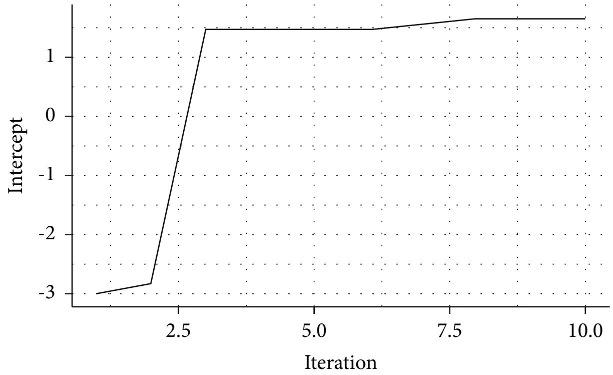

Figure 1: Intercept estimation process.

Then, under assumptions H_1', H_9, H_{10}, we have

$$\nabla_\theta g(\theta_n) \longrightarrow 0\, a.s\, \text{and}\, \nabla_\theta g(\theta_n) \longrightarrow 0\, q.m. \tag{31}$$

□

3.2. Simulation. In order to illustrate our work, we perform a simulation in which we estimate the different parameters of a logistic regression. Our simulations are performed using the programming language "R." We simulate 10 000 observations of the random variable $V \backslash \text{leads to}\, N(3, 10)$, and we define U such that

$$U = \begin{cases} 1, & \text{if}\, V + \epsilon > 0, \\ 0, & \text{if not,} \end{cases} \tag{32}$$

with $\varepsilon \backslash \text{leads to}\, N(0, 3)$, to avoid having a perfectly fitted model. Then, we fitted a classical logistic regression with the Fisher scoring algorithm, which converged in 12 iterations. We define the accuracy rate as the number of correctly classified observations over the total number of our observations, and the classic model has an accuracy of 90.34%. Table 1 shows all the remaining outputs of the model.

Regarding the proposed process, we initiate it with the following randomly chosen values, Intercept = −3, $\theta = -3$, and we choose $a_n = 1 + \exp(-\rho(V)'\theta/n)$; as $\rho(V)$ and θ are finite, we can see that assumption H_1' is verified, and we also randomly draw a sample of one observation to perform our calculations at each iteration. Finally, we have set an accuracy of 10^{-12}. Following simulations, we obtain the results as follows.

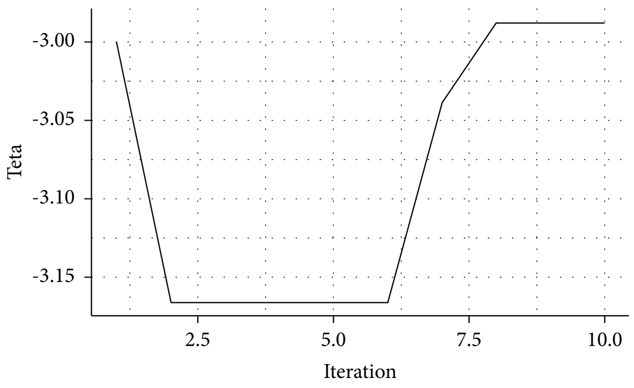

Figure 2: θ estimation process.

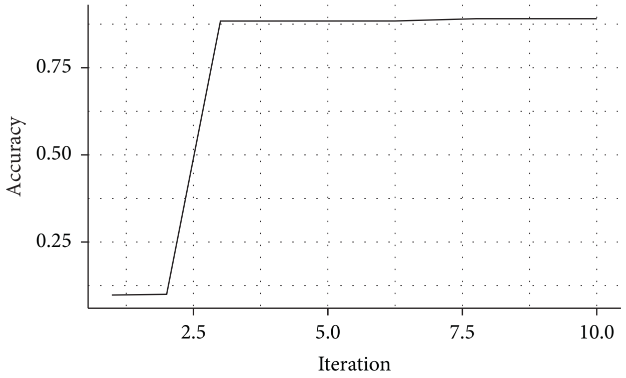

Figure 3: Accuracy of forecasts during iterations.

Table 2: Summary of the iterations.

Iteration	Accuracy	θ	Intercept
1	0.097 9	−3.000 000	−3.000 000
2	0.099 1	−3.166 397	−2.827 981
3	0.883 3	−3.166 397	1.467 527
4	0.883 3	−3.166 396	1.467 534
5	0.883 3	−3.166 396	1.467 534
6	0.883 3	−3.166 395	1.467 538
7	0.887 4	−3.039 216	1.547 544
8	0.890 7	−2.987 531	1.648 300
9	0.890 7	−2.987 531	1.648 300
10	0.890 7	−2.987 531	1.648 300

We can see through Figures 1 and 2, as well as Figure 3, that the process converged in 10 iterations. Therefore, we only needed 10 samples of one observation to obtain a robust estimation of the coefficients. Moreover, we can see in Figure 3 as well as in the summary of the process, in Table 2, that the latter records a prediction accuracy on the set of simulated observations of 89%, hence a loss of 1% in accuracy, but, in return, we gained greatly in terms of computing capacity.

4. Conclusion

In this work, we have demonstrated the convergence of the process studied towards the values that minimize the function $g(\theta)$, and following our simulations, we can see that this theoretical result is also valid on the empirical level. Nevertheless, this simulation required that we arbitrarily set a starting point, which leads to a possible slow convergence of the process in case the initial point is far from the targeted value. Moreover, the speed of convergence is also greatly affected by the choice of the a_n. Thus, a possible improvement would be to find the optimal sequence a_n that provides the fastest convergence.

Data Availability

No data were used to support this study.

Conflicts of Interest

The authors declare no conflicts of interest.

References

[1] Basel Committee on Banking Supervision, *Consultative Document "The Internal Ratings Based Approach" Supporting Document to the New Basel Capital Accord Posted: 2001-01*, Basel Committee on Banking Supervision, Basel, Switzerland, 2001.

[2] R. Michel, I. Schnakenburg, and T. von Martens, "Effective customer selection for marketing campaigns based on net scores," *Journal of Research in Interactive Marketing*, vol. 11, no. 1, 2017.

[3] D. S. W. Ho, W. Schierding, M. Wake, R. Saffery, and J. O'Sullivan, "Machine learning SNP based prediction for precision medicine," *Frontiers in Genetics*, vol. 10, p. 267, 2019.

[4] R. Abel, S. Mondal, C. Masse et al., "Accelerating drug discovery through tight integration of expert molecular design and predictive scoring," *Current Opinion in Structural Biology*, vol. 43, pp. 38–44, 2017.

[5] P. Hu, D. Cai, S. Wang, A. Yao, and Y. Chen, "Learning supervised scoring ensemble for emotion recognition in the wild," in *Proceedings of the 19th ACM International Conference on Multimodal Interaction*, pp. 553–560, Glasgow, UK, November 2017.

[6] J. Shin, Y. Lee, and K. Jung, "Effective sentence scoring method using bert for speech recognition," in *Proceedings of the Asian Conference on Machine Learning*, pp. 1081–1093, PMLR, Nagoya, Japan, October 2019.

[7] J. Calvo-Zaragoza, A. H. Toselli, and E. Vidal, "Probabilistic music-symbol spotting in handwritten scores," in *Proceedings of the 2018 16th International Conference on Frontiers in Handwriting Recognition (ICFHR)*, pp. 558–563, IEEE, Niagara Falls, NY, USA, August 2018.

[8] A. Abozaid, A. Haggag, H. Kasban, and M. Eltokhy, "Multimodal biometric scheme for human authentication technique based on voice and face recognition fusion," *Multimedia Tools and Applications*, vol. 78, no. 12, pp. 16345–16361, 2019.

[9] R. S. Perdana and A. Pinandito, "Combining likes-retweet analysis and naive bayes classifier within twitter for sentiment analysis," *Journal of Telecommunication, Electronic and Computer Engineering (JTEC)*, vol. 10, no. 1-8, pp. 41–46, 2018.

[10] B. C. Searle, M. Turner, and A. I. Nesvizhskii, "Improving sensitivity by probabilistically combining results from

multiple MS/MS search methodologies," *Journal of Proteome Research*, vol. 7, no. 1, pp. 245–253, 2008.

[11] K. Yoshii, M. Goto, K. Komatani, T. Ogata, and H. G. Okuno, "An efficient hybrid music recommender system using an incrementally trainable probabilistic generative model," *IEEE Transactions on Audio Speech and Language Processing*, vol. 16, no. 2, pp. 435–447, 2008.

[12] T. Mitchell, *Generative and Discriminative Classifiers: Naive Bayes and Logistic Regression*, Carnegie Mellon University, Pittsburgh, PA, USA, 2005.

[13] A. Y. Ng and M. I. Jordan, "On discriminative vs. generative classifiers: a comparison of logistic regression and naive bayes," *Advances in Neural Information Processing Systems*, vol. 2, pp. 841–848, 2002.

[14] J. Le, A tour of the top 10 algorithms for machine learning newbies, 2018.

[15] C. Cortes and V. Vapnik, "Support-vector networks," *Machine Learning*, vol. 20, no. 3, pp. 273–297, 1995.

[16] A. Tharwat, "Linear vs. quadratic discriminant analysis classifier: a tutorial," *International Journal of Applied Pattern Recognition*, vol. 3, no. 2, pp. 145–180, 2016.

[17] E. Fix and J. L. Hodges, "Discriminatory analysis. Nonparametric discrimination: consistency properties," *International Statistical Review/Revue Internationale de Statistique*, vol. 57, no. 3, pp. 238–247, 1989.

[18] S. Shalev-Shwartz and S. Ben-David, *Understanding Machine Learning: From Theory to Algorithms*, Cambridge University Press, Cambridge, UK, 2014.

[19] F. Rosenblatt, "The perceptron: a probabilistic model for information storage and organization in the brain," *Psychological Review*, vol. 65, no. 6, pp. 386–408, 1958.

[20] T. Kohonen, "Learning vector quantization," in *Self-organizing Maps*, pp. 175–189, Springer, Berlin, Heidelberg, 1995.

[21] H. Robbins and S. Monro, "A stochastic approximation method," *The Annals of Mathematical Statistics*, vol. 22, no. 3, pp. 400–407, 1951.

[22] A. Bennar, A. Bouamaine, and A. Namir, "Almost sure Convergence and in quadratic mean of the gradient stochastic process for the sequential estimation of a conditional expectation," *Applied Mathematical Sciences*, vol. 2, no. 8, pp. 387–395, 2008.

[23] J. A. Nelder and R. W. M. Wedderburn, "Generalized linear models," *Journal of the Royal Statistical Society: Series A*, vol. 135, no. 3, pp. 370–384, 1972.

[24] S. Benzoni-Gavage, *Calcul Différentiel et équations Différentielles-2e éd.: Cours et Exercices Corrigés*, Dunod, Paris, France, 2014.