Advances and Applications in Statistics

Published Online: September 2015
http://dx.doi.org/10.17654/ADASJul2015_037_055
Volume 46, Number 1, 2015, Pages 37-55 **ISSN: 0972-3617**

GEOMETRIC ERGODICITY OF THE NORMALIZED PERCEPTRON ALGORITHM FOR THE CLASSIFICATION OF NON-SEPARABLE GAUSSIAN INPUT VECTORS

Rieken S. Venema

Department of Mathematics and Statistics
University of Alaska Anchorage
3211 Providence Drive
Anchorage, Alaska 99508, U. S. A.
e-mail: rvenema@alaska.edu

Abstract

In this paper, we study the asymptotic behavior of the normalized weight sequence of a single-layer perceptron. If the perceptron is used for the classification of two infinite populations that cannot be linearly separated, then the weights do not converge but under certain conditions approach a steady state. Assuming that the input vectors are two-dimensional and normally distributed, we will show that the normalized perceptron weight process is geometrically ergodic.

1. Introduction

Neural networks have been applied successfully in classification problems. The simplest type of neural network is the single-layer perceptron,

Received: May 14, 2015; Accepted: July 18, 2015
2010 Mathematics Subject Classification: 60J05, 62G20, 62M45.
Keywords and phrases: single-layer perceptron, non-separable classifications, normalized perceptron algorithm, Markov chains, geometrically ergodic.
Communicated by K. K. Azad

invented in 1962 by Rosenblatt [4] to classify image patterns. Perceptrons can learn by example and training. After each classification, the perceptron is told the correct answer. There exists a simple learning algorithm for finite classification problems that guarantee a correct solution given that a solution exists. Specifically, when the input vectors of the single-layer perceptron can be linearly separated into two categories, this network can be trained to correctly classify these input vectors after a finite number of misclassifications. The separable case is well understood but rare; in most practical classification problems, the linearly separability assumption is not satisfied. Therefore, in this paper, we focus on the case where the input vectors are not linearly separable.

The single-layer perceptron is an artificial neural network composed of k input units, where information from the environment is sent in, and one processing unit, which is the output of the system. Each input neuron i synapses onto the output neuron and a weight $w(i)$ is assigned to this connection. These weights are considered adjustable and will be changed by the system in response to experience (the "learning" of the network). Suppose that an input vector $x = (x(1), x(2), ..., x(k)) \in R^k$ is presented to the network. The input neuron i sends the activity $x(i)$ to the output neuron which is scaled by the weight $w(i)$. If we define the weight vector $w = (w(1), w(2), ..., w(k)) \in R^k$, then the output neuron of the single-layer perceptron produces the output $Y_w(x) = 1_{\{w \cdot x \geq \theta\}}$, i.e., $Y_w(x) = 1$ if the inner product $w \cdot x = \sum_{i=1}^{k} [w(i)x(i)]$ is greater than or equal to a threshold θ, and $Y_w(x) = 0$ otherwise. Formally, we may add an additional input element always set equal to -1, so that the input $(x(1), x(2), ..., x(k))$ is replaced by $(-1, x(1), x(2), ..., x(k))$. The corresponding weight setting will be (θ, w). This means we can assume with no loss of generality that there is no threshold, and that the output of the single-layer perceptron is $Y_w(x) = 1_{\{w \cdot x \geq 0\}}$.

Let the set of pairs $\{(x_n, t_n), n \geq 1\}$ be taken from a space $\Gamma \subseteq R^k \times \{0, 1\}$. The goal of the single-layer perceptron is, given the value of the input vector x_n, to learn the value of the corresponding t_n, which is the desired output of the network. We say that the input vector x_n is classified correctly by the network (with weight vector w) if its observed output $Y_w(x_n)$ equals the desired output t_n. The *perceptron algorithm*, introduced by Rosenblatt [4], in which the network adjusts its weight vector to try to learn the classification is the following:

$$w_n = w_{n-1} + \eta(t_n - Y_{w_{n-1}}(x_n))x_n, \tag{1.1}$$

where $n \geq 1$, η is a fixed positive parameter (called the *learning rate*), and w_0 is the initial arbitrary weight vector in R^k. In the nth iteration, the input vector $x_n \in R^k$ is presented to the network which has weight setting w_{n-1}. If the vector x_n is classified correctly (i.e., if $Y_{w_{n-1}}(x_n) = t_n$), then there is no need to change the weight setting and algorithm (1.1) gives $w_n = w_{n-1}$. In case $Y_{w_{n-1}}(x_n) = 0$ but $t_n = 1$, the algorithm gives $w_n = w_{n-1} + \eta x_n$, so w_n is closer to the direction of x_n, yielding a larger inner product. A similar situation holds if $Y_{w_{n-1}}(x_n) = 1$ but $t_n = 0$. This process of adjustments of the weights results in the "training" of the neural network. The single-layer perceptron is "learning" itself so that it can correctly classify the input vectors.

Now suppose that the space Γ is finite. Also, suppose that Γ is linearly separable, i.e., there exists a vector $w^* \in R^k \setminus \{0\}$ such that

$$w^* \cdot x \geq 0 \text{ for all } (x, 1) \in \Gamma,$$

$$w^* \cdot x < 0 \text{ for all } (x, 0) \in \Gamma.$$

Then the Perceptron Convergence Theorem (Rosenblatt [4]) states that, *after a finite number of misclassifications*, *every input vector will be correctly*

classified. In other words, if there exists a weight vector w^* that can linearly separate the input vectors, then the perceptron algorithm will converge to such a vector after a finite number of weight adjustments.

Most natural classification problems, however, are not linearly separable (i.e., in two-group classification problems, the "positive examples" cannot be separated from the "negative examples" by a hyperplane (a linear subspace in R^k of dimension $k-1$)). Minsky and Papert [8] showed that in these cases the perceptron learning rule (1.1) will perform infinitely many changes of the weight vector, and the perceptron algorithm will not converge. Learning will never reach a point where all vectors are classified properly. The most famous example of the perceptron's inability to solve problems with linearly non-separable vectors is the Boolean exclusive-or problem. There have been several studies of the performance of the perceptron learning algorithm, or variants of it, in non-separable classification problems, see, for instance, Bershad and Shynk [9], Cortes and Vapnik [2], Freund and Schapire [16], Gallant [14], Roychowdhury et al. [15], Shynk and Bershad [6] and Yang et al. [5].

The Perceptron Cycling Theorem (Minsky and Papert [8]) states that *for finite non-separable classification problems the perceptron weight sequence* $\{w_n, n \geq 0\}$ *stays in a uniformly bounded region and does not diverge to infinity*. The first reported proof of this theorem is by Efron [1]. In the present paper, we want to study the behavior of the single-layer perceptron weight algorithm (1.1) for infinite non-separable classification problems. If the input vectors $\{x_n, n \geq 1\}$ form an independent identically distributed (i.i.d.) sequence, then the weight process is a Markov chain with stationary transition probabilities. In the next section, we will give sufficient conditions under which the perceptron weight process is ergodic.

2. Sufficient Conditions for Ergodicity of the Perceptron Weight Process

In this section, we want to study the weight sequence generated by the perceptron algorithm (1.1) in case the input vectors are not linearly separable.

We assume that the input-teacher pairs are realizations of an i.i.d. sequence of random variables (X_n, T_n) taken from an infinite space $\Phi \subseteq R^k \times \{0, 1\}$. If $W_0 \in R^k$ is an arbitrary vector, then the perceptron algorithm

$$W_n = W_{n-1} + \eta X_n(T_n - 1_{\{W_{n-1} \cdot X_n \geq 0\}}) \tag{2.1}$$

defines a Markov process with stationary transition probabilities. The conditional probability distribution of W_n given the values of $\{W_0, ..., W_{n-2}, W_{n-1}\}$ just depends on the value of W_{n-1} : $P(W_n \in A | W_0 = w_0, ..., W_{n-2} = w_{n-2}, W_{n-1} = w) = P(W_n \in A | W_{n-1} = w)$, and these one-step transition probabilities are independent of the time variable n : $P(W_n \in A | W_{n-1} = w) = P(W_1 \in A | W_0 = w) = P(w, A)$ for all positive integers n.

To simplify (2.1), we transform an input-teacher pair (X_n, T_n) with $T_n = 0$ to $(-X_n, 1)$ and we leave (X_n, T_n) unchanged in case $T_n = 1$. If we denote the new pair by $(Z_n, 1)$, then we get $Z_n = -X_n$ in case $T_n = 0$ and hence equation (2.1) becomes:

$$W_n = W_{n-1} + \eta Z_n 1_{\{W_{n-1} \cdot Z_n \leq 0\}}. \tag{2.2}$$

In the other case, i.e., when $T_n = 1$, we get $Z_n = X_n$ and hence equation (2.1) becomes $W_n = W_{n-1} + \eta Z_n 1_{\{W_{n-1} \cdot Z_n < 0\}}$. As we will assume later on that hyperplanes have Z-measure zero, this recursion might also be written as (2.2). Note that equation (2.2) defines the Markov process $\{W_n, n \geq 0\}$ naturally as a random dynamical system with iteration map $g_Z(w) = w + \eta Z 1_{\{w \cdot Z \leq 0\}}$.

Observe that the transformed input vectors $\{Z_n, n \geq 1\}$ form an i.i.d. sequence, defined on an infinite space $\Psi \subseteq R^k$. Let ν be the induced probability measure of Z_1. In what follows, we will denote by Z a generic random variable with distribution ν. Define for the vector $w \in R^k \setminus \{0\}$ the

$(k-1)$-dimensional hyperplane $H_w = \{Z \in R^k : w \cdot Z = 0\}$. In this paper, we assume that the measure ν satisfies the following:

$$\nu(H_w) = P(Z \in H_w) = P(Z \in R^k : w \cdot Z = 0) = 0 \text{ for all } w \in R^k \backslash \{0\}. \quad (2.3)$$

This assumption assures that $w \to g_Z(w)$ is continuous in distribution - at least for all w not equal to the zero vector in R^k, which is why we have to restrict the state space to $R^k \backslash \{0\}$. Note that, since H_w has Lebesgue measure zero in R^k for all $w \in R^k \backslash \{0\}$, assumption (2.3) is satisfied if the measure ν is absolutely continuous with respect to Lebesgue measure λ on R^k (i.e., $\forall A \subset R^k : \lambda(A) = 0$ implies $\nu(A) = 0$).

For $w \in R^k \backslash \{0\}$ and $A \subset R^k$, we denote by

$$P(w, A) = P(W_n \in A | W_{n-1} = w)$$

the transition kernel associated with perceptron algorithm (2.2). Observe that $P(w, A) = P(Z \in R^k : w + \eta Z 1_{\{w \cdot Z \le 0\}} \in A)$. For $w \in R^k \backslash \{0\}$, $A \subset R^k$, and n a positive integer, we let $P^n(w, A)$ denote the n-step transition probabilities: $P^n(w, A) = P(W_n \in A | W_0 = w)$, where $P^1(w, A) = P(w, A)$.

Definition 2.1. A Markov chain $\{W_n, n \ge 0\}$ on a state space X is *ergodic* if there exists a unique stationary distribution μ such that $P^n(w, A) \to \mu(A)$ as $n \to \infty$, for any measurable set $A \subset X$ and any initial value $w \in X$.

Ergodicity says that $P^n(w, \cdot)$ *converges* to μ in total variation for each $w \in X$. The main result of this section gives mild conditions for the perceptron weight sequence $\{W_n, n \ge 0\}$, defined by (2.2), to be ergodic. Burton et al. [12] showed that assumption (2.3) implies that the perceptron weight sequence is a weak Feller chain, i.e., the map $w \mapsto P(w, A)$ is lower

semicontinuous for any open set A, or equivalently, the map $w \mapsto \int h(y)P(w, dy)$ is continuous for any bounded continuous function $h(\cdot)$ on the state space. Venema [13] further showed that assumption (2.3) also implies that the transition law $\{P(w, \cdot)\}$ is strongly continuous, i.e., for all Borel measurable $A \subset R^k$: $P(w, A)$ is a continuous function in w. This form of continuity is sometimes expressed by calling $\{W_n, n \geq 0\}$ strongly Feller.

If we define for $w \in R^k \setminus \{0\}$ the half-space $C_w = \{Z \in R^k : w \cdot Z \leq 0\}$, then, in addition to assumption (2.3), we further assume that

$$\nu(C_w) = P(Z \in C_w) = P(Z \in R^k : w \cdot Z \leq 0) > 0 \text{ for all } w \in R^k \setminus \{0\}. \quad (2.4)$$

Assumption (2.4) is equivalent to non-separability of the classification problem, since it says that *no vector* $w \in R^k \setminus \{0\}$ correctly classifies, and is thus crucial for the analysis. If we let $\| \cdot \|$ denote the Euclidian norm in R^k, then we finally assume that for input vector $Z_n \in R^k$:

$$E(\| Z_n \|) = \int_{R^k} \| z \| d\nu(z) < \infty. \quad (2.5)$$

Under assumptions (2.3), (2.4) and (2.5), Venema [13] proved that, for the perceptron weight sequence $\{W_n, n \geq 0\}$ generated by equation (2.2), there exists a compact set $K \subset R^k$, with Lebesgue measure $\lambda(K) > 0$, which satisfies that

- $\exists 0 < c < \infty$ such that $E(\| W_n \| - \| W_{n-1} \| | W_{n-1} = w) \leq -c$ for all $w \notin K$.
- $\exists 0 < B < \infty$ such that $E(\| W_n \| - \| W_{n-1} \| | W_{n-1} = w) \leq B$ for all $w \in K$.

The existence of this compact set K, together with the strong continuity of $\{P(w, \cdot)\}$, implies that the conditions of Theorem 5.1 in Tweedie [11]

are satisfied. These conditions are sufficient conditions for our Markov process $\{W_n, n \geq 0\}$ to be ergodic. Therefore, as stated in Venema [13], the following conclusion can be drawn:

Theorem 2.2. *Define the perceptron weight sequence in* $R^k : W_n = W_{n-1} + \eta Z_n 1_{\{W_{n-1} \cdot Z_n \leq 0\}}$ *for* $n = 1, 2, ...,$ *with* $W_0 \in R^k$ *arbitrary and* $\eta > 0$ *fixed. Suppose that the input vectors* $\{Z_n, n \geq 1\}$ *form an i.i.d. sequence, defined on an infinite space* $\Psi \subseteq R^k$, *with* $E(\| Z_n \|) < \infty$. *Assume that* $\forall w \in R^k \setminus \{0\} : [P(Z \in R^k : w \cdot Z = 0) = 0$ *and* $P(Z \in R^k : w \cdot Z \leq 0) > 0]$. *Then the Markov process* $\{W_n, n \geq 0\}$ *is ergodic.*

A key conclusion from Theorem 2.2 is the existence of a unique stationary distribution μ such that $\forall A \in B(R^k) : \left[\mu(A) = \int P(w, A) d\mu(w) \right.$ and $\left. P(W_n \in A | W_0 = w) \to \mu(A) \text{ as } n \to \infty \text{ for any } w \in R^k \right]$. By applying a result in Tweedie [10], we get the following corollary, as a consequence of Theorem 2.2.

Corollary 2.3. *Assume that the conditions in Theorem* 2.2 *hold. Then, as* $N \to \infty$, *we have*: $\frac{1}{N} \sum_{n=1}^{N} P^n(w, A) \overset{a.s.}{\to} \mu(A)$ *for any* $A \in B(R^k)$, $w \in R^k$.

Actually, a result in Pollard and Tweedie [3] allows us to replace this corollary by the following stronger statement:

Corollary 2.4. *Assume that the conditions in Theorem* 2.2 *hold. Then for any initial distribution* ν *on* R^k *we have*:

$$\left\| \frac{1}{N} \int \sum_{n=1}^{N} P^n(w, \cdot) d\nu(w) - \mu(\cdot) \right\|_{TV} \to 0 \text{ as } N \to \infty,$$

where $\| \cdot \|_{TV}$ *denotes the total variation norm.*

A special application, where the conditions in Theorem 2.2 are satisfied, is the case where the input vectors $\{Z_n, n \geq 1\}$ are normally distributed. We will study this case in the next section. We will show that, under the assumption that the input vectors are in R^2 and normally distributed, the normalized perceptron algorithm is geometrically ergodic.

3. Geometric Ergodicity of the Normalized Perceptron Weight Process

Consider again the standard perceptron algorithm $W_n = W_{n-1} + \eta Z_n 1_{\{W_{n-1} \cdot Z_n \leq 0\}}$. In Section 2, we saw sufficient conditions for this weight process to be ergodic in the non-separable case. We are now interested in non-separable cases where the weight process is geometrically ergodic, that is the convergence of the *n*-step transition probability $P^n(w, A)$ to its limit $\mu(A)$ is geometrically fast.

Definition 3.1. A Markov chain $\{W_n, n \geq 0\}$ on a state space X is *geometrically ergodic* if there exist a unique stationary distribution μ and a number $\beta \in (0, 1)$ such that $| P^n(w, A) - \mu(A) | \leq M_{w, A} \beta^n$ for all measurable sets $A \subset X$, $w \in X$, and $n \geq 1$ ($M_{w, A}$ denotes a finite positive constant depending on *w* and A).

One example where the conditions in Theorem 2.2 are satisfied is the case where the input vectors are normally distributed. So, applying this theorem, we conclude that in the Gaussian case our weight process is ergodic. We will see in this section that a slightly modified weight process is geometrically ergodic in the two-dimensional Gaussian case.

From now on, assume $k = 2$, i.e., the input vectors $\{Z_n, n \geq 1\}$ and the weights $\{W_n, n \geq 0\}$ are vectors in R^2. The aim of perceptron learning is to find a weight vector *w* so that the classification procedure $Y_w(x) = 1_{\{w \cdot x \geq 0\}}$ minimizes the misclassification probability. In other words, the perceptron algorithm attempts to construct a vector *w* such that the separation line

$\{x : x \cdot w = 0\}$ results in a minimal number of classification errors. In the separable case, we concluded from the Perceptron Convergence Theorem that $W_n \to w^*$, where the line $\{x : x \cdot w^* = 0\}$ is such that we do not make any classification errors. In the construction of the optimal separation line, the length of the vector w is of no importance. We are only interested in the angle that w makes with the x-axis, because this angle determines the direction of the separation line. It does not matter for the separation line if we make our vector w longer or shorter to a vector with the same direction and with length 1. Therefore, without loss of generality, we can restrict to the case where $\| w \| = 1$ for our weight vectors. So, from now on, we will work on the unit circle $S^1 = \{u \in R^2 : \| u \| = 1\}$, a compact set, instead of the space R^2.

We observed that the perceptron output only depends on the direction of the weight vector, and not on its length. Thus if we replace w by $w/\| w \|$, we get the same classifications. This motivates the following *normalized perceptron algorithm*:

$$W_n = \frac{W_{n-1} + \eta Z_n 1_{\{W_{n-1} \cdot Z_n \le 0\}}}{\| W_{n-1} + \eta Z_n 1_{\{W_{n-1} \cdot Z_n \le 0\}} \|} \tag{3.1}$$

for $n = 1, 2, ...,$ where W_0 is an arbitrary vector in S^1. This modified perceptron algorithm was originally proposed by Burton et al. [12]. Compared to the standard perceptron algorithm (2.2), the angle between W_n and the x-axis stays the same but the length of W_n becomes 1. The separation line $\{x : x \cdot W_n = 0\}$ does not change. We will study the resulting weight sequence again under the assumption that $\{Z_n, n \ge 1\}$ forms an i.i.d. sequence (now in R^2). The Markov chain $\{W_n, n \ge 0\}$ generated by (3.1) has the compact state space S^1. We will show that this weight process is geometrically ergodic when the two-dimensional input vectors are normally distributed.

Definition 3.2. Let F_{S^1} be the collection of Borel sets on S^1. A function $q : S^1 \times F_{S^1} \to [0, 1]$ is called a *stochastic transition function* if it satisfies the following properties:

(1) $q(w, \cdot)$ determines, for a fixed $w \in S^1$, a probability measure on F_{S^1}.

(2) $q(\cdot, A)$ determines, for a fixed $A \in F_{S^1}$, a F_{S^1}-measurable function.

The n-step transition probabilities are examples of stochastic transition functions, and they are calculated in the following inductive way:

- $P^1(w, A) = P(w, A) = P(W_{k+1} \in A | W_k = w)$.
- $P^{n+1}(w, A) = P(W_{k+n+1} \in A | W_k = w) = \int_{S^1} P^n(\alpha, A) P(w, d\alpha)$, for $n = 1, 2, \ldots$.

The probability that W_n belongs to the set A is then calculated as follows: $P(W_n \in A) = \int_{S^1} P^n(\alpha, A) dp(\alpha)$, for $n \geq 1$, where $p(A) = P(W_0 \in A)$. Note that when this probability does not depend on n, the $\{W_n\}$-process is called *strictly stationary* and $p(\cdot)$ a stationary or invariant probability distribution.

For all $w \in S^1$ and $n \geq 1$, $P^n(w, \cdot)$ is a measure on F_{S^1} and therefore by the Lebesgue decomposition it has an absolutely continuous and a singular component with respect to Lebesgue measure φ on S^1, i.e., we can write $P^n(w, A) = \int_A p_0^{(n)}(w, \alpha) d\varphi(\alpha) + \Delta^{(n)}(w, A)$ with for all $w \in S^1$: $p_0^{(n)}(w, \cdot)$ a F_{S^1}-measurable function and $\Delta^{(n)}(w, \cdot)$ a measure on F_{S^1}. The following lemma can be found in Doob [7]:

Lemma 3.3. *Let* $\{W_n, n \geq 0\}$ *be a Markov process on a space* Ω *with n-step transition probabilities* $P^n(w, A) = P(W_{k+n} \in A | W_k = w)$. *Define* $p_0^{(1)}(\cdot, \cdot)$ *to be the absolute continuous component of* $P^1(\cdot, \cdot)$ *with respect to Lebesgue measure on* Ω. *Suppose* $\exists \delta \in (0, 1)$ *such that* $p_0^{(1)}(w, \alpha) \geq \delta$, $\forall w, \alpha \in \Omega$. *Then there exists a unique stationary probability distribution* $\mu(\cdot)$ *such that* $| P^n(w, A) - \mu(A) | \leq (1 - \delta)^{n-1}$ *for all* $n \geq 1$, $w \in \Omega$, *and* $A \in F_\Omega$.

Can we apply this lemma to the weight process generated by the normalized perceptron algorithm (3.1)? For this, we have to determine whether the condition in Lemma 3.3 holds for this process. In the proof of the next theorem, we will show that this is indeed the case when the two-dimensional input vectors are normally distributed.

Theorem 3.4. *Suppose that the Markov process* $\{W_n, n \geq 0\}$ *on* S^1 *is generated by the algorithm* $W_n = \dfrac{W_{n-1} + \eta Z_n 1_{\{W_{n-1} \cdot Z_n \leq 0\}}}{\| W_{n-1} + \eta Z_n 1_{\{W_{n-1} \cdot Z_n \leq 0\}} \|}$ *for* $n = 1, 2, \ldots,$ *where* W_0 *is an arbitrary vector in* S^1 *and* $\eta > 0$ *is fixed. Let the input vectors* $\{Z_n, n \geq 1\}$ *in* R^2 *be normally* $N_2(\mu, I)$ *distributed. Then* $\{W_n, n \geq 0\}$ *is geometrically ergodic.*

Proof. Let $w = (w_1, w_2) \in S^1$, $A \in F_{S^1}$, and let $P^n(w, A) = P(W_{k+n} \in A | W_k = w)$ denote the *n*-step transition probability for the Markov process, with $P(w, A) = P^1(w, A)$. Let $Z_n \in R^2$ denote the *n*th input vector presented to the network. Define the function:

$$f(w, Z_n) = \frac{w + \eta Z_n 1_{\{w \cdot Z_n \leq 0\}}}{\| w + \eta Z_n 1_{\{w \cdot Z_n \leq 0\}} \|}. \tag{3.2}$$

Then we have:

$$P(w, A) = P(W_n \in A \mid W_{n-1} = w) = P(f(W_{n-1}, Z_n) \in A \mid W_{n-1} = w),$$

and because Z_n is independent of W_{n-1} we get that

$$P(w, A) = P(f(w, Z_n) \in A). \tag{3.3}$$

From Definition 3.2, it follows that $f(w, Z_n) = w$ with probability $p^* = P(w \cdot Z_n > 0)$, and $f(w, Z_n) = \frac{w + \eta Z_n}{\| w + \eta Z_n \|}$ with probability $1 - p^*$.

We will try to find the distribution of $\frac{w + \eta Z_n}{\| w + \eta Z_n \|}$. Define the random variable $Y_n = w + \eta Z_n$. Since we assume that $Z_n \sim N_2(\mu, I)$, we have $Y_n \sim N_2(w + \eta\mu, \eta^2 I)$ and therefore the density of Y_n is $p(y) = \frac{1}{2\pi\eta} e^{\frac{-1}{2\eta^2}(y-w-\eta\mu)^T(y-w-\eta\mu)}$. The random variable $\frac{Y_n}{\| Y_n \|}$ is a point on S^1, with coordinates $(\cos\theta, \sin\theta)$. Define the density of $(\cos\theta, \sin\theta)$ on S^1 by the density of θ on $[0, 2\pi]$.

Let $(y_1, y_2) \in R^2$ be the coordinates of Y_n and consider the function: $h : (y_1, y_2) \to \left(\sqrt{y_1^2 + y_2^2}, \arctan\left(\frac{y_2}{y_1}\right) \right)$, or: $h : (r\cos\theta, r\sin\theta) \to (r, \theta)$. Define $q(\theta)$ to be the density function of $\theta = \arctan\left(\frac{y_2}{y_1}\right)$. We are trying to calculate $q(\theta)$. First calculate the density function of $(r, \theta) = \left(\sqrt{y_1^2 + y_2^2}, \arctan\left(\frac{y_2}{y_1}\right) \right)$. Denote this function by $s(r, \theta)$. Then:

$$\begin{aligned} s(r, \theta) &= p_{h(y_1, y_2)}(r, \theta) \\ &= p_{(y_1, y_2)}(h^{-1}(r, \theta)) \cdot \left| \det Jac_{h^{-1}}(r, \theta) \right| \end{aligned}$$

$$= p_{(y_1, y_2)}(r\cos\theta,\ r\sin\theta)\cdot r.$$

So we get that

$$s(r,\theta) = \frac{r}{2\pi\eta} e^{\frac{-1}{2\eta^2}(r\cos\theta - w_1 - \eta\mu_1, r\sin\theta - w_2 - \eta\mu_2)(r\cos\theta - w_1 - \eta\mu_1, r\sin\theta - w_2 - \eta\mu_2)^T}$$

and the density function of $\frac{Y_n}{\| Y_n \|}$ is $q(\theta) = \int_0^\infty s(r,\ \theta)dr$.

Because $p(\cdot, \cdot)$ is a normal density function, we have: $s(r,\ \theta) = r \cdot p(r\cos\theta,\ r\sin\theta) > 0,\ \forall r > 0,\ \forall\theta \in [0,\ 2\pi]$. Therefore:

$$q(\theta) = \int_0^\infty s(r,\ \theta)dr > 0, \quad \forall\theta \in [0,\ 2\pi]. \tag{3.4}$$

Also, $q(\theta)$ is a continuous function because:

- $r \to s(r,\ \theta)$ is integrable over $[0,\ \infty)$ for all $\theta \in [0,\ 2\pi]$, since $s(r,\ \theta) = r \cdot p(r,\ \cos\theta,\ r\sin\theta)$, with $p(\cdot, \cdot)$ a normal density function.

- $\theta \to s(r,\ \theta)$ is continuous for almost all $r \in [0,\ \infty)$, since $\cos(\theta)$ and $\sin(\theta)$ are continuous functions, so $\lim_{h\to 0} p(r\cos(\theta + h),\ r\sin(\theta + h)) = p(r\cos\theta,\ r\sin\theta)$ because $p(\cdot, \cdot)$ is a normal density function, so it is continuous in both variables.

- There exists a function $g(r)$ integrable, non-negative, and independent of θ such that $|\, s(r,\ \theta)| \le g(r),\ \forall r > 0,\ \forall\theta \in [0,\ 2\pi]$. Intuitively, it is clear that such a function $g(r)$ exists, because if $r \to \infty$, then $p(r\cos\theta,\ r\sin\theta) \to 0,\ \forall\theta \in [0,\ 2\pi]$ and so $s(r,\ \theta) \to 0,\ \forall\theta \in [0,\ 2\pi]$. However, it is hard to determine such a function explicitly.

Now we can apply the lemma, an application of the dominated convergence theorem, which states that the three upper conditions imply that $q(\theta) = \int_0^\infty s(r,\ \theta)dr$ is continuous on $[0,\ 2\pi]$. The continuity of $q(\theta)$ and

(3.4) now implies that

$$\exists 0 < \delta_0 < 1 \text{ such that } q(\theta) > \delta_0, \forall \theta \in [0, 2\pi]. \tag{3.5}$$

We can conclude from (3.2) and (3.3) that

$$\begin{aligned} P^1(w, A) &= P(w, A) = P(W_n \in A \mid W_{n-1} = w) \\ &= P\left(\left(\frac{w + \eta Z_n}{\| w + \eta Z_n \|} \in A\right) \cap (wZ_n \le 0)\right) \\ &\quad + P((w \in A) \cap (wZ_n > 0)). \end{aligned}$$

Since the vectors Z_n and $W_{n-1} = w$ are independent, we get:

$$\begin{aligned} P^1(w, A) &= P\left(\frac{w + \eta Z_n}{\| w + \eta Z_n \|} \in A \mid wZ_n \le 0\right) P(wZ_n \le 0) \\ &\quad + P(w \in A) P(wZ_n > 0) \\ &= (1 - p^*) P\left(\frac{w + \eta Z_n}{\| w + \eta Z_n \|} \in A \mid wZ_n \le 0\right) + p^* P(w \in A). \end{aligned}$$

Because $w = (w_1, w_2)$ is a fixed vector in S^1, we can write the following:

$$P^1(w, A) = (1 - p^*) \int_A q(\theta) 1_{(w_1 r \cos\theta + w_2 r \sin\theta \le 0)} d\theta + p^* \delta_w(A),$$

with $\delta_w(A)$ denoting the Dirac measure of w on $A \subset S^1$.

Now define $p_0^{(1)}(w, \alpha) = (1 - p^*) q(\theta) 1_{(w_1 r \cos\theta + w_2 r \sin\theta \le 0)}$, where $\alpha = (r, \theta)$. Further, define $\Delta^{(1)}(w, A) = p^* \delta_w(A)$, then we observe that $P^1(w, A) = \int_A p_0^{(1)}(w, \alpha) d\varphi(\alpha) + \Delta^{(1)}(w, A)$, with φ the Lebesgue measure on S^1. From (3.5), we can conclude that

$$\begin{aligned} p_0^{(1)}(w, \alpha) &= (1 - p^*) q(\theta) 1_{(w_1 r \cos\theta + w_2 r \sin\theta \le 0)} \\ &> (1 - p^*) \delta_0, \forall w, \alpha \in S^1. \end{aligned}$$

Let $\delta = (1 - p^*)\delta_0$. Then $0 < \delta < 1$ and we have shown that, in the case of two-dimensional Gaussian input vectors, the normalized perceptron weight sequence results in an absolutely continuous component $p_0^{(1)}(w, \alpha)$ of its one-step transition probabilities, for which there exists a $\delta \in (0, 1)$ such that $p_0^{(1)}(w, \alpha) \geq \delta$, $\forall w, \alpha \in S^1$. Hence, we can apply Lemma 3.3 and conclude that there exists a unique stationary probability distribution $\mu(\cdot)$ such that $| P^n(w, A) - \mu(A) | \leq (1 - \delta)^{n-1}$ for all $n \geq 1$, $w \in S^1$, and $A \in F_{S^1}$. If we take $\beta = 1 - \delta \in (0, 1)$ and $M = \frac{1}{1 - \delta}$, then we have shown that $| P^n(w, A) - \mu(A) | \leq M\beta^n$ for all measurable sets $A \subset S^1$, $w \in S^1$, and $n \geq 1$. Therefore, $\{W_n, n \geq 0\}$ is geometrically ergodic. □

4. Conclusions and Future Research

In this paper, we gave certain conditions under which the weight process generated by the standard perceptron algorithm is ergodic in infinite non-separable classification problems. Further, we proved that a slightly modified weight sequence is geometrically ergodic in the two-dimensional Gaussian case. In a future paper, we will present simulations that illustrate our main result, Theorem 3.4. We first considered the standard perceptron algorithm

$$W_n = W_{n-1} + \eta X_n (t(X_n) - 1_{\{W_{n-1} \cdot X_n \geq 0\}})$$

and we let the input vectors X with $t(X) = 1$ be $N_2((a, 0), I)$-distributed and the vectors X with $t(X) = 0$ be $N_2((-a, 0), I)$-distributed, with $a > 0$. The optimal separation line, which results in a minimum number of expected classification errors, is the y-axis. Our main interest is the size of $\varphi_n = \arctan(W_n(2)/W_n(1))$, the angle between the weight W_n and the x-axis. This is because the separation line $\{x : x \cdot W_n = 0\}$ is only depending on the angle φ_n and not on the length of the vector W_n. The optimal angle is in our case

0. Our simulations showed that, using different values of the learning rate η and different initial vectors W_0, the angle initially approaches the optimal angle 0, but after a certain number of iterations the weights get stuck in a small neighborhood of the origin and the process is oscillating under the influence of the discontinuity point $(0, 0)$. The angle is fluctuating and does not stay in a neighborhood of 0.

Next, we considered the normalized perceptron algorithm

$$W_n = \frac{W_{n-1} + \eta X_n (t(X_n) - 1_{\{W_{n-1} \cdot X_n \geq 0\}})}{\| W_{n-1} + \eta X_n (t(X_n) - 1_{\{W_{n-1} \cdot X_n \geq 0\}}) \|}$$

and again we let the input vectors X with $t(X) = 1$ be $N_2((a, 0), I)$-distributed and the vectors X with $t(X) = 0$ be $N_2((-a, 0), I)$-distributed, with $a > 0$. The optimal weight vector is $(1, 0)$, because for this vector the separation line is the y-axis and the input vectors with positive x-coordinate are classified to be 1, and the vectors with negative x-coordinate to be 0. Using the results from Theorem 3.4, we know that the conditional distribution of W_n, given the value of W_0, converges to the invariant distribution $\mu = \mu^\eta$. Assume that $W_0 = (1, 0)$. This is no restriction because in using the algorithm we could, after one step, end up in any other point on the circle with positive probability. The simulations showed that, for different values of η, the weights live in a small neighborhood of the asymptotically stable equilibrium point $(1, 0)$ on the circle, and the corresponding separation lines in a neighborhood of the y-axis. This is an illustration of the advantage of not working in R^2, but on the circle S^1.

Now, if the function $\eta \to \mu^\eta$ is continuous, then this would imply that $\mu^\eta \to \mu^0$, if $\eta \to 0$. We have $\mu^0 = \delta_{(1,0)}$ because if $\eta = 0$, then after applying the algorithm the weights stay to be $(1, 0)$. So, if $\eta \to \mu^\eta$ is continuous, then we have convergence of the weight vectors to the optimal

vector $(1, 0)$, if we let $\eta \to 0$ (as, for instance, in simulated annealing schedules). The question remains whether this assumption of continuity is true. We conjecture that the distribution of μ^{η} converges to the distribution concentrated at the optimal weight vector $(1, 0)$. This is actually supported by computer simulations, but we have not been able (yet) to find a rigorous proof.

References

[1] B. Efron, The perceptron correction procedure in non-separable situations, Technical Report RADC-TDR-63-533, Rome Air Development Center, Rome, NY, 1964.

[2] C. Cortes and V. Vapnik, Support-vector networks, Mach. Learn. 20(3) (1995), 273-297.

[3] D. B. Pollard and R. L. Tweedie, *R*-theory for Markov chains on a topological state space I, J. London Math. Soc. 10 (1975), 389-400.

[4] F. Rosenblatt, Principles of Neurodynamics: Perceptrons and the Theory of Brain Mechanisms, Spartan Books, New York, NY, 1962.

[5] J. Yang, R. Parekh and V. Honavar, Comparison of performance of variants of single-layer perceptron algorithms on non-separable data, Neural Parallel Sci. Comput. 8 (2000), 415-438.

[6] J. J. Shynk and N. J. Bershad, Stationary points of a single-layer perceptron for nonseparable data models, Neural Networks 6(2) (1993), 189-202.

[7] J. L. Doob, Stochastic Processes, Wiley, New York, NY, 1953.

[8] M. Minsky and S. Papert, Perceptrons: An Introduction to Computational Geometry, The MIT Press, Cambridge MA, 1969.

[9] N. J. Bershad and J. J. Shynk, Performance analysis of a converged single-layer perceptron for nonseparable data models with bias terms, IEEE Trans. Signal Process. 42(1) (1994), 175-188.

[10] R. L. Tweedie, *R*-theory for Markov chains on a general state space I: solidarity properties and *R*-recurrent chains, Ann. Probab. 2 (1974), 840-864.

[11] R. L. Tweedie, Sufficient conditions for ergodicity and recurrence of Markov chains on a general state space, Stoch. Proc. Appl. 3(4) (1975), 385-403.

[12] R. M. Burton, H. G. Dehling and R. S. Venema, Perceptron algorithms for the classification of non-separable populations, Commun. Stat. - Stoch. Models 13(2) (1997), 205-222.

[13] R. S. Venema, Sufficient conditions for ergodicity of the single-layer perceptron weight sequence in infinite non-separable classification problems (manuscript submitted for publication).

[14] S. I. Gallant, Perceptron-based learning algorithms, IEEE T. Neural Networ. 1(2) (1990), 179-191.

[15] V. P. Roychowdhury, K. Y. Siu and T. Kailath, Classification of linearly nonseparable patterns by linear threshold elements, IEEE T. Neural Networ. 6(2) (1995), 318-331.

[16] Y. Freund and R. Schapire, Large margin classification using the perceptron algorithm, Mach. Learn. 37(3) (1999), 277-296.