On the Classification Capability of Sign-Constrained Perceptrons

Robert Legenstein
legi@igi.tugraz.at
Wolfgang Maass
maass@igi.tugraz.at
Institute for Theoretical Computer Science, Technische Universitaet Graz,
A-8010 Graz, Austria

The perceptron (also referred to as McCulloch-Pitts neuron, or linear threshold gate) is commonly used as a simplified model for the discrimination and learning capability of a biological neuron. Criteria that tell us when a perceptron can implement (or learn to implement) all possible dichotomies over a given set of input patterns are well known, but only for the idealized case, where one assumes that the sign of a synaptic weight can be switched during learning. We present in this letter an analysis of the classification capability of the biologically more realistic model of a sign-constrained perceptron, where the signs of synaptic weights remain fixed during learning (which is the case for most types of biological synapses). In particular, the VC-dimension of sign-constrained perceptrons is determined, and a necessary and sufficient criterion is provided that tells us when all 2^m dichotomies over a given set of m patterns can be learned by a sign-constrained perceptron. We also show that uniformity of L_1 norms of input patterns is a sufficient condition for full representation power in the case where all weights are required to be nonnegative. Finally, we exhibit cases where the sign constraint of a perceptron drastically reduces its classification capability. Our theoretical analysis is complemented by computer simulations, which demonstrate in particular that sparse input patterns improve the classification capability of sign-constrained perceptrons.

1 Introduction

A simple mathematical model for a neuron is a perceptron, which computes a function $f_{th} : \mathbb{R}^n \to \{-1, 1\}$ of the form $f_{th}(\mathbf{x}) = \varphi(\mathbf{w}^T \mathbf{x} - \theta)$ (McCulloch & Pitts, 1943; Minsky & Papert, 1988; Rosenblatt, 1962; Haykin, 1999). The activation function φ is defined by $\varphi(z) = 1$ if $z \geq 0$, else $\varphi(z) = -1$. $\mathbf{w} = (w_1, \ldots, w_n)^T$ is the weight vector, and θ is the threshold. θ and $\mathbf{w}$ are the adjustable parameters of a perceptron that can be changed during learning.

If one uses a perceptron as the model for a readout neuron from a neural circuit, then the vector $\mathbf{x} = (x_1, \ldots, x_n)^T$ represents the synaptic inputs that this readout receives at a certain time t from a set of n neurons in the circuit. If the neurons in the circuit are modeled as spiking neurons, then x_i could, for example, be defined as the number of spikes that the ith presynaptic neuron emitted during a preceding time interval $[t - \Delta, t]$ of length Δ. In this case x_i represents an estimate of the current firing rate of the ith presynaptic neuron.

For an unconstrained perceptron, it is quite clear how the neural circuit that contains the n presynaptic neurons of this readout could optimally support the discrimination capability of the perceptron for m activation patterns $\mathbf{x}(1), \ldots, \mathbf{x}(m)$ of the n presynaptic neurons: it should make sure that the n-dimensional inputs $\mathbf{x}(1), \ldots, \mathbf{x}(m)$ are linearly independent. If that is the case and $m \leq n$, then a perceptron can learn (with the help of the perceptron learning rule) to compute any of the 2^m possible classification functions (or dichotomies) $h : \{\mathbf{x}(1), \ldots, \mathbf{x}(m)\} \rightarrow \{-1, 1\}$.

However there exists one significant discrepancy between physiological reality and the perceptron learning rule. In the perceptron learning rule, and also in most other commonly considered learning algorithm for linear neurons or perceptrons, weights can assume values of any sign and can even change their sign in the learning process. However, biological synapses are either excitatory or inhibitory and usually do not switch between excitation and inhibition. This fact is commonly referred to as Dale's law. In fact, many neurophysiologists prefer the assumption that only excitatory synapses are directly used for learning, whereas inhibitory synapses are tuned for other tasks (such as gain regulation or regulation of the firing threshold of a neuron, or timing of firing). In the latter case, one arrives at a perceptron with nonnegative weights as a more realistic model for a readout neuron.

We consider in this letter the case where the readout is modeled by a perceptron with sign-constrained weights (i.e., by a linear threshold gate that obeys Dale's law). For any vector $\mathbf{s} = (s_1, \ldots, s_n) \in \{-1, 1\}^n$, we define $\mathbb{R}^n_{\mathbf{s}} \subseteq \mathbb{R}^n$ as

$$\mathbb{R}^n_{\mathbf{s}} = \{\mathbf{w} = (w_1, \ldots, w_n) \in \mathbb{R}^n | w_i s_i \geq 0 \text{ for all } i \in \{1, \ldots, n\}\}. \qquad (1.1)$$

A set $\mathbb{R}^n_{\mathbf{s}}$ of this type represents the space of possible weight assignments for such a perceptron, with $s_i = 1$ if the ith presynaptic neuron is excitatory and $s_i = -1$ if it is inhibitory. We define $\mathbb{R}^n_+$ as the nonnegative subset of $\mathbb{R}^n$: $\mathbb{R}^n_+ = \mathbb{R}^n_{(1,\ldots,1)}$. An n-dimensional point (vector) $\mathbf{w}$ is called nonnegative if $\mathbf{w} \in \mathbb{R}^n_+$.

For a list $\langle \langle \mathbf{x}(1), t(1) \rangle, \ldots, \langle \mathbf{x}(m), t(m) \rangle \rangle$ of training examples $\langle \mathbf{x}(i), t(i) \rangle$ with target outputs $t(i) \in \{-1, 1\}$, the goal of perceptron learning is to find a weight vector $\mathbf{w} \in \mathbb{R}^n$ and a bias θ such that $\varphi(\mathbf{w}^T \mathbf{x}(i) - \theta) = t(i)$ for $i = 1, \ldots, m$. For unconstrained weights, this is achieved, according

to the perceptron convergence theorem, by the perceptron learning algorithm, provided that such weight vector exists (Rosenblatt, 1962; Minsky & Papert, 1988). A variation of the perceptron convergence theorem for sign-constrained weights was proven in Amit, Wong, and Campbell (1989). However this result provides no information about the discrimination capability of a sign-constrained perceptron. It tells us only that if a sign-constrained perceptron can implement a given dichotomy, then it can learn it. Hence the result of Amit, Wong, et al. (1989) reduces all questions regarding learnability to realizability of given dichotomies by a sign-constrained perceptron. The latter is the topic of this letter.

The storage capacity of recurrent attractor neural networks with sign-constrained weights was investigated in Amit, Campbell, and Wong (1989), based on the classical result by Gardner (1987) for unconstrained weights. For a single perceptron with input dimension n, Gardner calculated the relative volume of the weight space, which implements a randomly drawn dichotomy on p randomly drawn patterns from $\{-1, 1\}^n$. The volume was calculated as a function of the loading level $\alpha = p/n$. Gardner (1987) showed that there is a critical value $\alpha_c = 2$ of the loading level in the following sense: in the limit of large n, the relative volume of weights that implement a particular dichotomy on p patterns (averaged over randomly drawn dichotomies and randomly drawn patterns) vanishes for a loading level above $\alpha_c = 2$. Amit, Campbell, et al. (1989) showed that for sign-constrained weights, α_c is exactly one-half of the value in the unconstrained case. However, this result does not provide information about the number of dichotomies that can be implemented by sign-constrained perceptrons for a given set of input patterns for finite n and does not determine the structure of point sets that maximize this number. Hence, it does not provide any information regarding the questions investigated in this letter. However, we show that for the interesting case of a critical loading level $\alpha = 1$, there exist point sets for which any dichotomy can be implemented with sign-constrained weights and that one can identify those point sets.

Building on the work of Gardner (1987) and Brunel, Hakim, Isope, Nadal, and Barbour (2004) calculated the weight distribution of a perceptron with nonnegative weights that stores a random set of patterns near its critical loading level. The authors showed that the perceptron with nonnegative weights is a good model for cerebellar Purkinje cells (with respect to linear summation and a small time window of temporal integration of inputs). Fitting their analytically derived weight distributions to that measured for granule cell–Purkinje cell synapses, a good agreement could be achieved. The main conclusion of this work was that optimal weight distributions for perceptrons with nonnegative weights contain a significant amount (more than 50%) of silent synapses (those with zero weight). The fact that such silent synapses are not pruned (they obviously do not contribute to the network behavior), points to the importance of sustained flexibility of neurons even in the adult (Brunel et al., 2004). The question how the

(positive) weight constraint of these synapses reduces the flexibility of computation and learning for a Purkinje cell (modeled by a perceptron) was not addressed in that article. This question is the main topic of this letter.

So far, little has been known about the expressive power and generalization capability of sign-constrained perceptrons. The VC-dimension (which has value $n + 1$ for an unconstrained perceptron with input dimension n) is a standard measure for the expressive power and generalization capability of a learning device (see Vapnik, 1998, and Bartlett & Maass, 2003). The VC-dimension of a class $\mathcal{H}$ of hypotheses (the class of hypotheses which can be implemented by the learner) is defined as the size of the largest set S of points on which hypotheses from $\mathcal{H}$ have unrestricted expressive power (in the sense that any $h' : S \rightarrow \{-1, 1\}$ can be realized by some $h \in \mathcal{H}$). Hence, it informs us about the size of the largest possible best-case input point set for the hypothesis class $\mathcal{H}$. We show in section 2 that the VC-dimension of sign-constrained perceptrons is only by 1 smaller than that of unconstrained perceptrons. This implies that there exist also for sign-constrained perceptrons sets $S = \{\mathbf{x}(1), \ldots, \mathbf{x}(n)\} \subseteq \mathbb{R}^n$ of n inputs on which a perceptron can produce (and also learn according to Amit, Wong, et al., 1989) each of the 2^n theoretically possible dichotomies—in spite of its sign constraint. However, we show in section 3 that such sets S for which sign-constrained perceptrons achieve their maximal possible flexibility (we call such sets S shatterable) are much sparser than for unconstrained perceptrons. Whereas unconstrained perceptrons achieve this for any set S of $\leq n$ linearly independent vectors $\mathbf{x}(i)$, we show in theorem 2 that linear independence does not entail any significant flexibility for sign-constrained perceptrons. This fact gives rise to the question whether there exists any property of sets $S = \{\mathbf{x}(1), \ldots, \mathbf{x}(m)\} \subseteq \mathbb{R}^n$ that guarantees for $m \leq n$ that a sign-constrained perceptron can learn to compute all of the 2^m possible dichotomies over S. We present such property in section 3 (see theorem 4). In fact we present a sufficient and necessary condition that requires checking whether a system of m linear equations is solvable (the straightforward approach would require checking 2^m systems, which is infeasible for reasonable m). In the case of $m = n$, this condition can be stated as: a sign-constrained perceptron can learn to compute all of the 2^m possible dichotomies over $\{\mathbf{x}(1), \ldots, \mathbf{x}(m)\}$ if and only if a scaling vector $\mathbf{v}$ exists that respects the sign constraints with nonzero entries such that $\mathbf{v}^T \mathbf{x}(i) = 1$ for all i.

In the case of perceptrons with nonnegative weights and inputs, this condition can be interpreted as a requirement that all vectors $\mathbf{x}(i)$ in the set have the same norm for a scaled version of the L_1 norm (where the components of $\mathbf{v}$ are the scaling factors). Furthermore, in this case, uniformity of the L_1 norms of the input vectors $\mathbf{x}(i)$ is a sufficient condition for shatterability. We analyze in section 4 through computer experiments to what extent it suffices when this property is just approximately satisfied. Furthermore we show that the classification capability of sign-constrained perceptrons is

substantially larger for sparse input patterns, thereby exhibiting an unexpected new benefit of sparse coding for neural computation.

2 An Optimal Bound for the VC-Dimension of Sign-Constrained Perceptrons

For a vector $\mathbf{s} \in \{-1, 1\}^n$, we denote by $\mathcal{H}_\mathbf{s}^n$ the class of functions $f : \mathbb{R}^n \to \{-1, 1\}$ computable by perceptrons with weights constrained to have signs given by $\mathbf{s}$, formally:

Definition 1. *For a vector* $\mathbf{s} \in \{-1, 1\}^n$, $\mathcal{H}_\mathbf{s}^n$ *is the set of functions* $f : \mathbb{R}^n \to \{-1, 1\}$ *of the form* $f(\mathbf{x}) = \varphi(\mathbf{w}^T \mathbf{x} - \theta)$ *with* $\mathbf{w} \in \mathbb{R}_\mathbf{s}^n$ *and* $\theta \in \mathbb{R}$.

For any class $\mathcal{H}$ of functions from some domain D into $\{-1, 1\}$, the VC-dimension of $\mathcal{H}$ ($VC - Dim(\mathcal{H})$) is defined as the size of the largest[1] subset D' of its domain D, which is shattered by $\mathcal{H}$, that is, for which every dichotomy h' over D' (each of the $2^{|D'|}$ many functions $h' : D' \to \{-1, 1\}$) can be represented as a restriction of some function $h \in \mathcal{H}$ to the subdomain D' (there exists some h in $\mathcal{H}$ so that $h'(x) = h(x)$ for all $x \in D'$). The VC-dimension of a class $\mathcal{H}$ of functions had been introduced by Vapnik and Chervonenkis (1971) as the main tool for estimating the number of training examples needed for training a learning algorithm (that outputs hypotheses from $\mathcal{H}$) in order to achieve a given error probability on test examples (see Vapnik, 1998; a short review is given in Bartlett & Maass, 2003). It is well known that the VC-dimension of unconstrained perceptrons with n inputs is $n + 1$. Figure 1 illustrates how the class of unconstrained perceptrons shatters a point set consisting of three points in $\mathbb{R}^2$. We show in this letter that constraining the weights of a perceptron reduces its VC-dimension only by 1. This small reduction of the VC-dimension is insofar surprising as the hypothesis space $\mathcal{H}$ itself is drastically reduced through the introduction of a sign constraint vector $\mathbf{s}$, for example, to the class of all *monotone* linear threshold functions for $\mathbf{s} = (1, 1, \ldots, 1)$. It implies that a sign-constrained perceptron needs (for a worst-case probability distribution of examples) almost the same number of training examples as an unconstrained perceptron in order to achieve a given error rate on test examples, in spite of its significantly reduced expressive power.

Theorem 1. $VC - Dim(\mathcal{H}_\mathbf{s}^n) = n$, *for every* $\mathbf{s} \in \{-1, 1\}^n$.

We give a formal proof of $VC - Dim(\mathcal{H}_\mathbf{s}^n) \geq n$ in appendix A. Here we will give the main idea of this proof in order to provide some intuition,

---

[1] In case that subsets D' of arbitrarily large finite size are shattered by $\mathcal{H}$, one says that the VC-dimension of $\mathcal{H}$ is infinite.
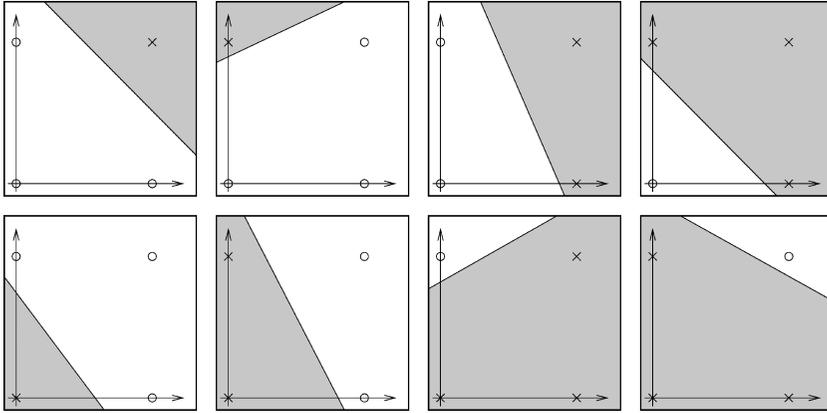
Figure 1: Eight dichotomies for a set $D := \{(0, 0), (0, 1), (1, 0), (1, 1)\}$ of four points in $\mathbb{R}^2$ can be realized by the class $\mathcal{H}^2$ of unconstrained perceptrons. For each of the eight functions $h \in \mathcal{H}^2$ illustrated, the shaded region represents the half-space where $h(\mathbf{x}) = 1$. When a point $\mathbf{x} \in D$ satisfies $h(\mathbf{x}) = 1$, it is marked as an $\mathcal{X}$; when it satisfies $h(\mathbf{x}) = -1$, it is marked as a circle. The functions illustrated show that the subset $D' := \{(0, 0), (0, 1), (1, 0)\}$ of D is shattered by $\mathcal{H}^2$. Since each $h \in \mathcal{H}^2$ defines a half-plane in $\mathbb{R}^2$, one can see easily that the set D itself cannot be shattered by $\mathcal{H}^2$. This is true for any set D of four points in $\mathbb{R}^2$.

which will prove useful also for the following sections. To simplify matters, we restrict our intuitive arguments to nonnegative weights (i.e., $\mathbf{s} = (1, 1, \ldots, 1)$). We can assume without loss of generality that the weight vector is not the null vector.[2] A weight vector $\mathbf{w} \in \mathbb{R}^n$ and threshold θ define a hyperplane by the linear equation $\mathbf{w}^T \mathbf{x} = \theta$. We say that a point $\mathbf{x}$ is above the hyperplane if $\mathbf{w}^T \mathbf{x} > \theta$ and below the hyperplane if $\mathbf{w}^T \mathbf{x} < \theta$. A nonnegative weight vector is constrained to the nonnegative quadrant of $\mathbb{R}^n$, and the hyperplane is perpendicular to this weight vector. We call a hyperplane constrained if its weight vector is constrained and nonnegative if its weight vector is nonnegative.

In order to shatter a point set S with nonnegative perceptrons, all dichotomies on S have to be implemented with such restricted hyperplanes. This can easily be done for n points, which are placed directly on the axes of the coordinate system (e.g., the point set $S = \{\mathbf{e}_i | i = 1, \ldots, n\}$). Here, $\mathbf{e}_i$ is the ith unit vector with the ith entry being 1 and other entries being zero (see Figure 2A). We can find a nonnegative hyperplane that goes through these points. Then we can slightly tilt the hyperplane such that an arbitrary

---

[2]A perceptron with the null weight vector can implement only the two constant functions, and for a given set of input patterns, these functions can also be implemented by perceptrons with a nonnegative weight vector of nonzero length.
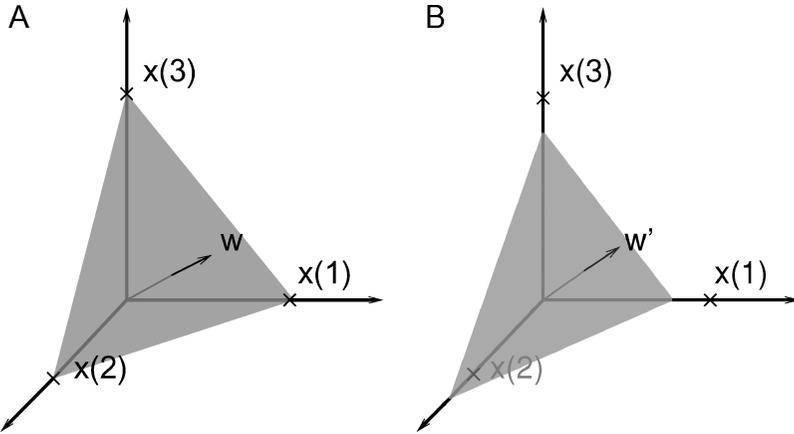
Figure 2: Shattering of point sets with nonnegative weights in three dimensions. (A) A nonnegative weight vector $\mathbf{w}$ is restricted to the positive quadrant of weight space. The corresponding hyperplane is indicated by the gray triangle. Lines "behind the hyperplane" are plotted in dark gray. $\mathbf{x}(1)$, $\mathbf{x}(2)$, and $\mathbf{x}(3)$ are points on the axes of the coordinate system. The hyperplane goes through these points. (B) A given dichotomy on the points $\mathbf{x}(1)$, $\mathbf{x}(2)$, and $\mathbf{x}(3)$ can be realized by slightly tilting the hyperplane from A). The hyperplane shown here corresponds to a perceptron that classifies $\mathbf{x}(2)$ negative and $\mathbf{x}(1)$ $\mathbf{x}(3)$ positive.

subset of points in S is above the hyperplane, and the other points of S are below it (see Figure 2B). This finishes our description of the idea for proving $VC - Dim(\mathcal{H}_{\mathbf{s}}^{n}) \geq n$.

To prove that $VC - Dim(\mathcal{H}_{\mathbf{s}}^{n}) \leq n$, we show that for each point set S that can be shattered by perceptrons with sign-constrained weights, one can construct a point set S' with one additional point, which can be shattered by some unconstrained perceptron. Hence, if one could shatter a point set of $n+1$ points with constrained perceptrons, the VC-dimension of unconstrained perceptrons would be at least $n+2$, which is a contradiction. We sketch the construction of S' for nonnegative weights (i.e., $\mathbf{s} = (1, 1, \ldots, 1)$) in the following. Since we can shatter S, we can fix for each dichotomy $h_i : S \to \{-1, 1\}$ one nonnegative hyperplane H_i defined by $\mathbf{w}(i), \theta(i)$, which realizes this dichotomy ($i = 1, \ldots, 2^{|S|}$). We can assume that no point in S is on one of these hyperplanes (a point $\mathbf{x}$ is on the hyperplane H_i if $\mathbf{w}(i)^T \mathbf{x} = \theta(i)$). Note that for any point $\mathbf{x} \in S$, if $\mathbf{w}^T \mathbf{x} > \theta$, then $-\mathbf{w}^T \mathbf{x} < -\theta$. Hence, the "flipped" hyperplane H_i' defined by $-\mathbf{w}(i)$ and $-\theta(i)$ realizes the dichotomy $-h_i$ for $i \in \{1, \ldots, 2^{|S|}\}$. Hence, the point set S can also be shattered with the set of flipped hyperplanes. Furthermore, there exists a point $\mathbf{p}$ that is above all hyperplanes $H_1, \ldots, H_{2^{|S|}}$ (therefore,

$\mathbf{p} \notin S$).[3] This point $\mathbf{p}$ is below all flipped hyperplanes $H'_1, \ldots H'_{2^{|S|}}$. Hence, there exists a separating hyperplane for each dichotomy on $S' = S \cup \{\mathbf{p}\}$ in $\{H_1, \ldots, H_{2^{|S|}}, H'_1, \ldots, H'_{2^{|S|}}\}$. It follows that S' can be shattered by unconstrained perceptrons. A rigorous proof can be found in appendix A.

Theorem 1 implies that the generalization capabilities of sign-constrained perceptrons (for worst-case probability distribution of examples) are only marginally better than for unconstrained perceptrons. Hence, if a perceptron learns a classification on some input distribution, there cannot exist a general guarantee that the number of errors after learning on data coming from the same distribution is much better for sign-constrained perceptrons than for unconstrained ones.

3 A Characterization of Those Sets of Patterns on Which Sign-Constrained Perceptrons Have Unlimited Classification Capability

The VC-dimension of a class $\mathcal{H}$ informs us only about the size of the largest set S of points that can be shattered by $\mathcal{H}$, that is, on which hypotheses from $\mathcal{H}$ have unrestricted expressive power (in the sense that any $h' : S \to \{-1, 1\}$ can be realized by some $h \in \mathcal{H}$). But the VC-dimension does not provide any information about the frequency (or mathematical structure) of these point sets S that are "ideal" with regard to the expressive power of $\mathcal{H}$. It could, for example, be the case that these "ideal" sets S are very rare and that practically, the expressive power of a learning algorithm with hypothesis space $\mathcal{H}$ is much less than suggested by the value of $VC - Dim(\mathcal{H})$. This question is, for example, relevant if one considers a neuron v (modeled by a linear threshold gate) in a neural circuit and the sign-vector $\mathbf{s}$ is defined by the type (excitatory or inhibitory) of the n synapses onto this neuron v. If one wants to know how much this sign constraint $\mathbf{s}$ reduces the potential of this neuron v to learn an arbitrary dichotomy on a given set S of m activation patterns of the n presynaptic neurons, one needs to know under what conditions on S this neuron v can shatter this set S. For the classical theory without a sign constraint, there is a clear answer: a set $S \subseteq \mathbb{R}^n$ can be shattered by a perceptron if and only if the points in S are linearly independent. But for the case of sign-constrained perceptrons, the situation is completely different.

Figure 3A shows that for linearly independent points drawn uniformly from $[0, 1]^n$, the fraction F_{shat} of point sets S that can be shattered with nonnegative weights drops exponentially with the dimension n. The same holds true if the vectors in S are normalized to have the same L_2 norm. The

---

[3]Consider points of the form $\gamma \mathbf{1}$ with $\mathbf{1} = (1, 1, \ldots, 1)^T$ and $\gamma > 0$. For large enough γ, we have $\mathbf{w}(i)^T (\gamma \mathbf{1}) > \theta(i)$ for all $i \in \{1, \ldots, 2^{|S|}\}$, since all $\mathbf{w}(i)$ are nonnegative and not the null vector.
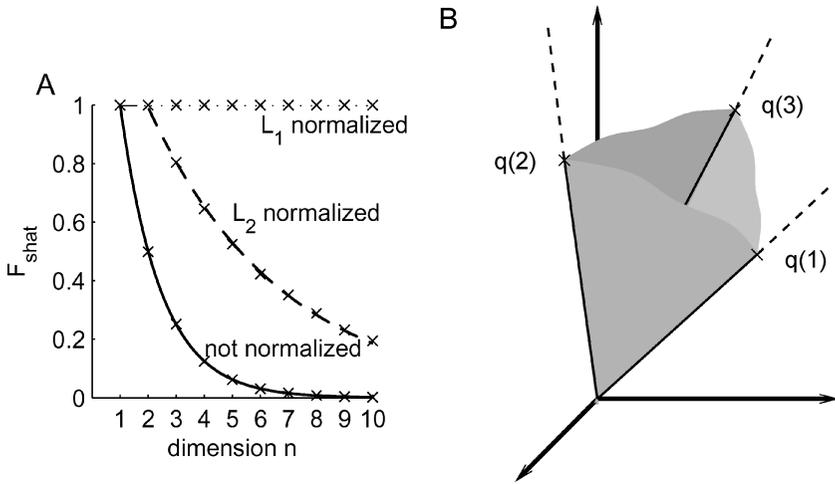
Figure 3: (A) Only an exponentially decreasing fraction F_{shat} of sets S of points drawn uniformly from $[0, 1]^n$ can be shattered by nonnegative perceptrons (full line). The same holds true if the points in S are normalized to have the same L_2 norm (dashed line). Each set S consisted of n n-dimensional nonnegative vectors drawn from a uniform distribution on $[0, 1]^n$. For each data point, we have drawn 10^5 sets S and determined shatterability with the use of corollary 1 given below. Lines show fitted exponentials (least mean squares fit). (B) Positive combinations of some randomly drawn points $\mathbf{q}(1), \mathbf{q}(2), \mathbf{q}(3)$ in $\mathbb{R}^3$ (which could, for example, be the columns of the matrix X in equation 3.1) constitute a cone C (surfaces of C indicated by gray shading). The cone expands unboundedly in the direction indicated by the dashed lines.

following theorem shows that the number of computable dichotomies for linearly independent points can in fact be extremely small (logarithmic in the number of all possible dichotomies) if the sign of weights is constrained.

Theorem 2. *For every $\mathbf{s} \in \{-1, 1\}^n$ and every $m \le n$, there exists a set S of m linearly independent vectors $\mathbf{x} \in \mathbb{R}^n$, such that at most $m + 1$ dichotomies over S can be implemented by perceptrons from $\mathcal{H}_s^n$.*

For the case $\mathbf{s} = (1, \ldots, 1)$ one can use here the fact that perceptrons with nonnegative weights can compute only monotone functions. We construct a set S of linearly independent points $\mathbf{x}(1), \ldots, \mathbf{x}(m)$ that are monotonously linearly ordered in every dimension: $\mathbf{x}(i + 1)$ has in any dimension a value not less than $\mathbf{x}(i)$ in that dimension, for all $i \in \{1, \ldots, m - 1\}$. Obviously, if $\mathbf{x}(i)$ is classified positive by a perceptron with nonnegative weights, then $\mathbf{x}(j)$ is also classified positive for all $j > i$. Therefore, nonnegative perceptrons can only compute $m + 1$ dichotomies over S. The full proof can be found in appendix B.

Perceptrons with unconstrained weights can produce all 2^m dichotomies over these sets S of linearly independent points. Therefore, sign-constrained perceptrons are much less powerful than unconstrained ones in the worst case. Since the VC-dimension of sign-constrained perceptrons is n, there is a huge gap between their classification capability for the best and the worst set S of points. Indeed, our simulation results indicate that most point sets in high-dimensional input spaces are not shatterable by sign-constrained perceptrons (see below). Furthermore, theorem 2 shows that the actual structure of the point set has a strong influence on its shatterability. In the following, we analyze which types of point sets are optimal from the perspective of shatterability. In other words, we analyze how input patterns should be represented in order to give a sign-constrained perceptron maximal flexibility for classification of these patterns.

Consider a set S of $m \leq n$ linearly independent points $\mathbf{x}(1), \ldots, \mathbf{x}(m)$ in $\mathbb{R}^n$. We examine for an arbitrarily fixed $\mathbf{s} \in \{-1, 1\}^n$ the question of whether it is possible to compute all 2^m possible classifications on these m points (i.e., one can shatter them) with sign-constrained perceptrons from $\mathcal{H}^n_\mathbf{s}$. In the following discussion, we consider for simplicity the case $\mathbf{s} = (1, \ldots, 1)$. To determine the set of all dichotomies on S that could possibly be implemented by a nonnegative perceptron, we analyze the set of scalar products of points in S with nonnegative vectors $\mathbf{v}$. For a nonnegative weight vector $\mathbf{v}$, consider the vector

$$(\mathbf{x}(1)^T\mathbf{v}, \mathbf{x}(2)^T\mathbf{v}, \ldots, \mathbf{x}(m)^T\mathbf{v})^T = X\mathbf{v},$$

where the ith row of the matrix X is given by $\mathbf{x}(i)^T$. For nonnegative $\mathbf{v}$, $X\mathbf{v}$ is called a positive combination of the columns of X (for short a positive combination of X).[4] Note that one linearly combines here the columns of X (n points in m-dimensional space) whereas our point set S is given by the rows of X. The set of all possible positive combinations of X has the form of a cone. We therefore refer to this set as the *cone* C_X of X:

$$C_X = \{\mathbf{y} \in \mathbb{R}^m \mid \mathbf{y} = X\mathbf{v} \text{ for some } \mathbf{v} \in \mathbb{R}^n_+\}. \tag{3.1}$$

The cone is a convex set and closed under positive combinations (see Figure 3B). It is the set of points that lie on a ray that passes through the origin and the convex region spanned by the n points defined by the columns of X. This cone C_X provides complete information about which dichotomies h can be computed over $S = \{\mathbf{x}(1), \ldots, \mathbf{x}(m)\}$ by perceptrons with nonnegative weights: these are exactly those dichotomies for which there exists a point $\mathbf{q} \in C_X$ and a threshold θ so that $q_i \geq \theta \Leftrightarrow h(\mathbf{x}(i)) = 1$. We can write

---

[4]We use the term *positive combination* for consistency with the literature. However, *nonnegative combination* would be more appropriate because factors can be zero.
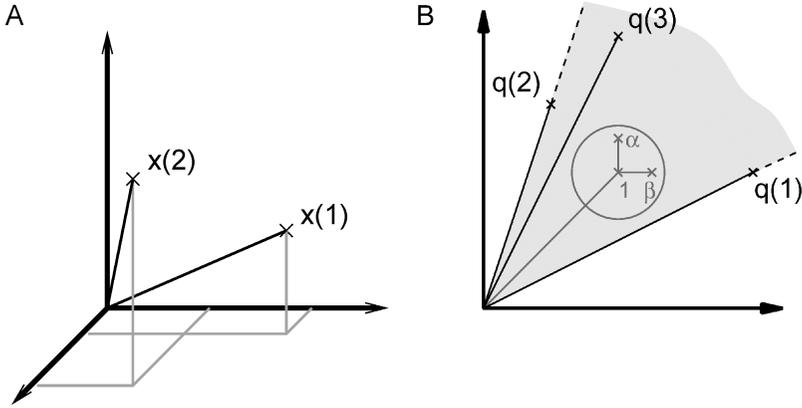
A

B

Figure 4: For two points $\mathbf{x}(1), \mathbf{x}(2)$ in three-dimensional space, the column space of X consists of three points $\mathbf{q}(1), \mathbf{q}(2), \mathbf{q}(3)$ in $\mathbb{R}^2$. (A) Can one shatter the two points $\mathbf{x}(1) = (4, 1, 2)$ and $\mathbf{x}(2) = (2, 3, 4)$ by perceptrons with nonnegative weights? (B) In the column space of X, these two points transform to $\mathbf{q}(1) = (4, 2), \mathbf{q}(2) = (2, 3)$, and $\mathbf{q}(3) = (2, 4)$. The cone C_X is indicated by the gray shaded region. An ϵ-ball around $\mathbf{1}$ is inside the cone (gray circle). Therefore, one can shatter $\mathbf{x}(1), \mathbf{x}(2)$ by perceptrons with nonnegative weights. Since the ϵ-ball around $\mathbf{1}$ is in the cone, the two points $\alpha = (1, 1 + \gamma)$ and $\beta = (1 + \gamma, 1)$ are also in the cone for some small $\gamma > 0$. $\alpha \in C_X$ implies that a nonnegative weight vector $\mathbf{w}$ exists such that $\mathbf{w}^T \mathbf{x}(2) > \mathbf{w}^T \mathbf{x}(1)$, and $\mathbf{x}(2)$ can be classified positive, while $\mathbf{x}(1)$ can be classified negative with this weight vector. Furthermore, $\beta \in C_x$ implies that a nonnegative weight vector exists such that $\mathbf{x}(1)$ is classified positive and $\mathbf{x}(2)$ is classified negative.

this condition without explicitly referring to the threshold θ: exactly those dichotomies h can be computed over $S = \{\mathbf{x}(1), \ldots, \mathbf{x}(m)\}$ by perceptrons with nonnegative weights for which there exists a point $\mathbf{q} \in C_X$ so that $h(\mathbf{x}(i)) = 1 \wedge h(\mathbf{x}(j)) = -1 \Rightarrow q_i > q_j$.

We illustrate the statement of the subsequent theorem 3 in an example. Two points in three-dimensional space are shown in Figure 4A. The situation in the two-dimensional column space of X together with the corresponding cone is shown in Figure 4B.

The observation above indicates a special role of points having the same value in all dimensions, for example, $\mathbf{1}$ (the vector having entry 1 in all dimensions). Consider, for example, the set of points $Q = \{1, 1 + \gamma\}^m$ for some small $\gamma > 0$. This set consists of points that differ from 1 by γ in an arbitrary set of dimensions, that is, points $\mathbf{q}$ of the form $\mathbf{q} = (1 + \gamma_1, 1 + \gamma_2, \ldots, 1 + \gamma_m)^T$, where γ_i can be either 0 or γ. If $Q \subset C_X$, then it follows from the preceding observations that all dichotomies can be computed over S by perceptrons with nonnegative weights. If a small ball around $\mathbf{1} = (1, 1, \ldots, 1)^T$ is in C_X, then for small enough γ, Q is a subset of the

cone. Hence, any dichotomy can be computed over S by perceptrons with nonnegative weights if a small ball around $\mathbf{1}$ is in the cone C_X.

We formalize and generalize these ideas. An ϵ-ball around a vector $\mathbf{x}$ is the set of points $\{\mathbf{y} \mid \|\mathbf{x} - \mathbf{y}\| < \epsilon\}$, where $\|\cdot\|$ denotes the L_2 norm. For a matrix X, let C_X be the cone spanned by the columns of X and $\mathbf{1}$ be the vector having entry 1 in all dimensions. Let $D_\mathbf{s}$ denote the diagonal matrix with the sign constraints $\mathbf{s}$ in its diagonal, that is, the elements of $D_\mathbf{s} = [d_{ij}]_{i,j=1,...,n}$ are $d_{ii} = s_i$ for $i = 1, \ldots, n$, and $d_{ij} = 0$ for $i \neq j$. Note that for a matrix X, the multiplication $X \cdot D_\mathbf{s}$ multiplies the ith column of X with the sign constraint s_i, $i = 1, \ldots, n$.

Theorem 3. *Fix any $\mathbf{s} \in \{-1, 1\}^n$. Let $S = \{\mathbf{x}(1), \ldots, \mathbf{x}(m)\}$ be a set of $m \leq n$ linearly independent n-dimensional points, and let the ith row of the $m \times n$ matrix X be given by $\mathbf{x}(i)^T$ for $i = 1, \ldots, m$. S can be shattered by $\mathcal{H}_\mathbf{s}^n$ if and only if an ϵ-ball around $\mathbf{1}$ lies inside the cone $C_{(X \cdot D_s)}$.*

The proof of the "if"-part of this theorem was sketched in the preceding discussion (a full proof is given in appendix C). For the "only if" part, one intuitive idea is to construct an ϵ-ball from the points defined by the 2^m dichotomies, thereby taking the reverse route in the preceding argument. However, the proof that such a construction is possible is not straightforward. Instead we employ a well-known theorem from economics: the theorem of separating hyperplanes. The full proof of theorem 3 is given in appendix C.

The condition given by theorem 3 is not easy to verify in practice for a given sets of points. The criterion can be stated in a simpler form with the help of the following lemma. For an $m \times n$ matrix X and a vector $\mathbf{v} \in \mathbb{R}^n$, let $R_\mathbf{v}(X)$ be the set of columns of X for which $\mathbf{v}$ has nonzero entries, that is, $R_\mathbf{v}(X) = \{\text{column } i \text{ of } X | v_i \neq 0\}$.

Lemma 1. *Let $S = \{\mathbf{x}(1), \ldots, \mathbf{x}(m)\}$ be a set of $m \leq n$ linearly independent n-dimensional points, and let the ith row of the $m \times n$ matrix X be given by $\mathbf{x}(i)^T$ for $i = 1, \ldots, m$. An ϵ-ball around $\mathbf{1}$ is in the cone C_X if and only if there exists a nonnegative $\mathbf{v}$ such that $X \cdot \mathbf{v} = \mathbf{1}$ and $R_\mathbf{v}(X)$ contains m linearly independent vectors.*

A rigorous proof of this lemma is given in appendix D. Theorem 3 states that S can be shattered by $\mathcal{H}_\mathbf{s}^n$ if and only if an ϵ-ball around $\mathbf{1}$ lies inside the cone $C_{X \cdot D_s}$. With the help of lemma 1, we see that S can be shattered by $\mathcal{H}_\mathbf{s}^n$ if and only if there exists a nonnegative $\mathbf{v}$ such that $(XD_\mathbf{s})\mathbf{v} = \mathbf{1}$ and $R_\mathbf{v}(X)$ contains m linearly independent vectors. We note that $(XD_\mathbf{s})\mathbf{v} = X(D_\mathbf{s}\mathbf{v})$ and that $\mathbf{v}$ is nonnegative if and only if $D_\mathbf{s}\mathbf{v} \in \mathbb{R}_\mathbf{s}^n$. Hence, S can be shattered by $\mathcal{H}_\mathbf{s}^n$ if and only if there exists a weight vector $\mathbf{v} \in \mathbb{R}_\mathbf{s}^n$ such that $X\mathbf{v} = \mathbf{1}$ and $\mathbf{v}$ combines (i.e., has nonzero entries for) m linearly independent columns of X.

Theorem 4. *Let $\mathbf{s} \in \{-1, 1\}^n$. Let $S = \{\mathbf{x}(1), \ldots, \mathbf{x}(m)\}$ be a set of $m \leq n$ linearly independent n-dimensional points, and let the ith row of the $m \times n$ matrix X be given by $\mathbf{x}(i)^T$ for $i = 1, \ldots, m$. S can be shattered by $\mathcal{H}_{\mathbf{s}}^n$ if and only if there exists a $\mathbf{v} \in \mathbb{R}_{\mathbf{s}}^n$ such that $X\mathbf{v} = \mathbf{1}$ and $R_{\mathbf{v}}(X)$ contains m linearly independent vectors.*

As a consequence of theorem 4, we get the following simple criterion for $m = n$.

Corollary 1. *Let $\mathbf{s} \in \{-1, 1\}^n$, and $S = \{\mathbf{x}(1), \ldots, \mathbf{x}(n)\}$ be a set of n linearly independent n-dimensional points. S can be shattered by $\mathcal{H}_{\mathbf{s}}^n$ if and only if there exists a $\mathbf{v} \in \mathbb{R}_{\mathbf{s}}^n$ with all entries being nonzero such that $\mathbf{x}(i)^T\mathbf{v} = 1$ for $i = 1, \ldots, n$.*

Corollary 1 gives a simple criterion for predicting whether a set S of n n-dimensional points can be shattered by sign-constrained perceptrons. If the matrix X that consists of the rows $\mathbf{x}(1)^T, \ldots, \mathbf{x}(n)^T$ does not have full rank, this is not possible. If X has full rank, compute $\mathbf{v} = X^{-1}\mathbf{1}$. Since the solution to this equation is unique, S can be shattered by sign-constrained perceptrons from $\mathcal{H}_{\mathbf{s}}^n$ if and only if $\mathbf{v}$ satisfies the sign constraints given by $\mathbf{s}$. Obviously this criterion is much easier to test than checking for each of the 2^n possible dichotomies h over S whether h can be realized with the given sign constraint.

We add another corollary for the case of perceptrons with nonnegative weights and $n = m$:

Corollary 2. *Let $S = \{\mathbf{x}(1), \ldots, \mathbf{x}(n)\}$ be a set of n linearly independent n-dimensional points. S can be shattered by $\mathcal{H}_{(1,1,\ldots,1)}^n$ if and only if there exists a positive $\mathbf{v}$ such that $\mathbf{x}(i)^T\mathbf{v} = 1$ for $i = 1, \ldots, n$.*

For the biologically important case of nonnegative inputs, the product $\mathbf{x}(i)^T\mathbf{v}$ in corollary 2 can be interpreted as a scaled L_1 norm of $\mathbf{x}(i)$. In this interpretation, the condition of corollary 2 is that there exists a positive scaling factor v_i for each dimension i of the nonnegative input vectors such that the L_1 norm of the scaled vectors is uniform, that is, that there exists some positive scaling vector $\mathbf{v}$ such that $\sum_{j=1}^n v_j x_j(i) = 1$ for $i = 1, \ldots, n$. Note that corollary 2 implies a simple sufficient condition for shatterability of $S \subseteq \mathbb{R}_+^n$ with nonnegative weights: the condition that all points in S have the same (nonzero) L_1 norm.

4 Approximate Uniformization of the L_1 Norm

We concentrate in the following on perceptrons with nonnegative weights and inputs for simplicity. Analogous arguments apply to general sign constraints. Corollary 2 implies that a set S of nonnegative vectors can be shattered by perceptrons with nonnegative weights if all of these vectors have
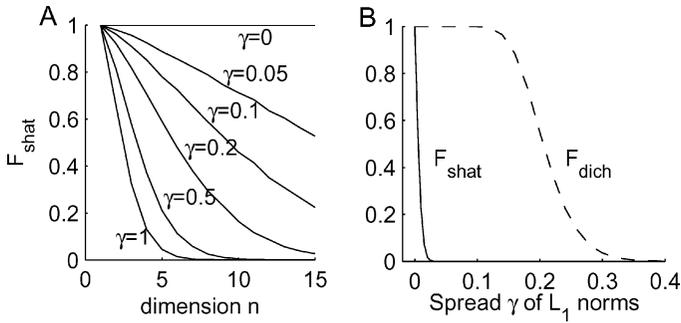
Figure 5: Fraction F_{shat} of randomly drawn point sets S that can be shattered by perceptrons with nonnegative weights (solid lines) and fraction F_{dich} of dichotomies that can be implemented by perceptrons with nonnegative weights over randomly drawn sets S (dashed line). Each set S consisted of n n-dimensional nonnegative vectors drawn from a uniform distribution on $[0, 1]^n$. The parameter γ determines the spread of different L_1 norms in S (see text). Dependency of F_{shat} on dimension (A) and dependency on γ for dimension $n = 100$ (B, solid line) are shown. F_{dich} was estimated by testing 200 dichotomies that were randomly drawn from a uniform distribution on over all possible dichotomies.

the same L_1 norm, because then we get $X \cdot (c \cdot \mathbf{1}) = \mathbf{1}$ for some scalar $c > 0$. This is not true if the vectors in S have the same L_2 norm. This confirms our simulation results that sets S of nonnegative vectors with uniform L_1 norm can be shattered by perceptrons with nonnegative weights, whereas sets S of nonnegative vectors with uniform L_2 norm do not have this property in general (see Figure 3A). What happens if the L_1 norms of the vectors in S are just slightly jittered around a constant value? We have performed computer simulations to elucidate this question. Vectors were drawn randomly from a uniform distribution over $[0, 1]^n$. Each vector $\mathbf{x}$ was then normalized to $(1 + r) \cdot \mathbf{x}/||\mathbf{x}||_{L_1}$, where r is a random number (different for each vector) drawn from the uniform distribution on the interval $[-\gamma/2, +\gamma/2]$. For each dimension and each γ, 30,000 such sets S, were drawn. Then for each such set S we determined with the help of corollary 1 whether the set can be shattered by a perceptron with nonnegative weights. The fraction F_{shat} of sets S that could be shattered with nonnegative weights is shown for different γ values as a function of n in Figure 5A. Figure 5B shows that it drops for fixed $n = 100$ very fast with increasing γ.

Even if a set S cannot be shattered, the number of dichotomies over S that can be computed by sign-constrained perceptrons could still be quite large. This was tested in further computer simulations. Let F_{dich} denote the fraction of dichotomies that can be computed by a perceptron with nonnegative weights. We show in Figure 5B that F_{dich} decays more
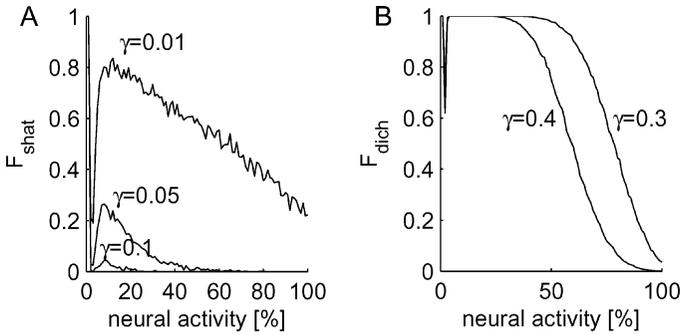
Figure 6: The sparsity of input vectors influences the classification capabilities of perceptrons with nonnegative weights. The "neural activity" on the horizontal axis is the percentage of nonzero entries in the vectors of the point sets S. Point sets S consisted of 100 linearly independent 100-dimensional nonnegative vectors. (A) Fraction F_{shat} of point sets S that can be shattered by perceptrons with nonnegative weights for different neural activities. (Sets of vectors with only one nonzero dimension per vector can always be shattered in a trivial way. However, very sparse vectors with more than one nonzero dimension have a very low probability of being shatterable, an effect for which we could not find a full explanation.) For each point in the plot, we tested 500 randomly chosen point sets. (B) Fraction F_{dich} of dichotomies that can be realized by perceptrons with nonnegative weights on point sets S for different neural activities. For each point in the plot, we tested 200 randomly chosen dichotomies on each of the 250 randomly chosen point sets. The parameter γ determines the spread of L_1 norms in the set (see text).

gracefully with γ than the fraction F_{shat} of shatterable sets, but still approaches 0 if the L_1 norm of the patterns $\mathbf{x}$ is not sufficiently uniform.

Neural activity in biological neural circuits is known to be sparse. We therefore also considered the impact of sparse input vectors on the classification capability of a sign-constrained perceptron. Sparse vectors were produced by setting all except a given percentage of randomly chosen components[5] of each input vector $\mathbf{x}$ to zero (we term this percentage "neural activity").[6] Figure 6A shows that sparse input sets have a higher probability of being shatterable than nonsparse ones. This effect is even more pronounced if one probes the number of dichotomies that can be realized with nonnegative weights on sparse sets (see Figure 6B).

---

[5]Implemented by drawing random permutations of the components $(1, \ldots, n)$.

[6]Very sparse sets tend to produce singular matrices (in our case, this tends to happen up to a dimension of 10). We therefore added only vectors to the set that were not linearly dependent on the previously chosen vectors.

5 Discussion

We have presented theoretical results on the generalization capability (the VC-dimension) and expressive power of sign-constrained perceptrons. We have also exhibited a simple necessary and sufficient condition for points sets S that tells us when exactly sign-constrained perceptrons have the same full classification capability over S as unconstrained perceptrons. A resulting sufficient criterion is the uniformity of L_1 norms in S. Computer tests with randomly drawn point sets S show also that an approximate satisfaction of the uniformity of L_1 norms increases the chance that S can be shattered with sign-constrained perceptrons. We have also demonstrated that the experimentally observed sparsity of neural activity contributes to the classification capability of sign-constrained perceptrons. A more detailed analysis of implications of the results of this letter for learning in circuits of spiking neurons is given in Legenstein and Maass (2006).

Appendix A: Proof of Theorem 1

We first show $VC - Dim(\mathcal{H}_s^n) \geq n$. We do this by constructing a set of n points that can be shattered by $\mathcal{H}_s^n$.

Consider the set of points $S = \{s_i \mathbf{e}_i | i = 1, \ldots, n\}$. Here, $\mathbf{e}_i$ is the ith unit vector, with the ith entry being 1 and other entries being zero. Any partition of S into P, N (P and N are a partition of S if $P \cup N = S$ and $P \cap N = \{\}$) can be implemented by a perceptron (vectors in P are assumed to be classified positive) with $\theta = 1/2$ and the weight vector

$$\mathbf{w} = \sum_{\mathbf{x} \in P} \mathbf{x}. \tag{A.1}$$

We get $\mathbf{w}^T \mathbf{x} = 1 > \theta$ for $\mathbf{x} \in P$ and $\mathbf{w}^T \mathbf{x} = 0 < \theta$ for $\mathbf{x} \in N$.

Now we show $VC - Dim(\mathcal{H}_s^n) \leq n$. We use the following strategy. We show that if one can shatter a set of $n + 1$ points in n-dimensional space by $\mathcal{H}_s^n$, then one can shatter a set of $n + 2$ points with perceptrons without weight restrictions. This is a contradiction to well-known results about the VC-dimension of perceptrons.

Suppose that $VC - Dim(\mathcal{H}_s^n) = n + 1$. We can then shatter a set S of $n + 1$ points in n-dimensional space. For each partition P, N of S, let $\mathbf{w}^{P,N}, \theta^{P,N}$ be a weight vector and threshold that separate P from N (points in P are classified positive; note that we fix one weight vector and threshold for each pair P, N). We can assume without loss of generality (w.l.o.g.) that $\mathbf{w}^{P,N} \neq (0, \ldots, 0)$ for all partitions P, N of S and $(\mathbf{w}^{P,N})^T \mathbf{x} - \theta^{P,N} \neq 0$ for all partitions P, N of S and $\mathbf{x} \in S$.

There exists a point $p \in \mathbb{R}^n$ that is classified positive by all $\mathbf{w}^{P,N}, \theta^{P,N}$. To show this, we consider the point $\mathbf{p} = \alpha \mathbf{s}$ for some $\alpha > 0$. This point is

classified positive for all partitions P, N of S, if $\sum_{i=1}^{n} w_i^{P,N} p_i = \alpha \sum_{i=1}^{n} |w_i^{P,N}| > \theta^{P,N}$ for all P, N. Since $\mathbf{w}^{P,N} \neq (0, \ldots, 0)$ for all partitions P, N of S, this is true for

$$\alpha > \max_{P,N} \left\{ \frac{\theta^{P,N}}{\sum_{i=1}^{n} |w_i^{P,N}|} \right\}. \tag{A.2}$$

Hence, $\mathbf{p}$ exists, and since it is classified positive by all weight vectors and thresholds, $\mathbf{p}$ is not in S.

One can shatter $S \cup \{\mathbf{p}\}$ with unconstrained weights in the following way. In order to implement any partition $P \cup \{\mathbf{p}\}, N$, one can simply use $\mathbf{w}^{P,N}$, and $\theta^{P,N}$ since $\mathbf{p}$ is classified positive by these weights and thresholds. In order to implement any partition $P, N \cup \{\mathbf{p}\}$, one can use $-\mathbf{w}^{N,P}$ and $-\theta^{N,P}$ (note the changed superscripts in $\mathbf{w}^{N,P}$ and $\theta^{N,P}$), since we have:

$$\forall \mathbf{x} \in P : \mathbf{x}^T \mathbf{w}^{N,P} < \theta^{N,P} \Rightarrow \forall \mathbf{x} \in P : -\mathbf{x}^T \mathbf{w}^{N,P} > -\theta^{N,P} \Rightarrow$$

$$\Rightarrow \forall \mathbf{x} \in P : \mathbf{x} \text{ is classified positive,}$$

$$\forall \mathbf{x} \in N : \mathbf{x}^T \mathbf{w}^{N,P} > \theta^{N,P} \Rightarrow \forall \mathbf{x} \in N : -\mathbf{x}^T \mathbf{w}^{N,P} < -\theta^{N,P} \Rightarrow$$

$$\Rightarrow \forall \mathbf{x} \in N : \mathbf{x} \text{ is classified negative,}$$

$$\mathbf{p}^T \mathbf{w}^{N,P} > \theta^{N,P} \Rightarrow \mathbf{p}^T (-\mathbf{w}^{N,P}) < -\theta^{N,P} \Rightarrow \mathbf{p} \text{ is classified negative.}$$

It follows that the class of perceptrons in n dimensions has a VC-dimension at least $n + 2$. This in a contradiction to the fact that the VC-dimension of perceptrons in n dimensions is $n + 1$. Therefore, the assumption that $VC - Dim(\mathcal{H}_{\mathbf{s}}^n) \geq n + 1$ was wrong, and the claim of theorem 1 follows.

Appendix B: Proof of Theorem 2

We can assume w.l.o.g. that all weights are restricted to be nonnegative, that is, $\mathbf{w} \in \mathbb{R}_+^n$. The result for $\mathbf{w} \in \mathbb{R}_{\mathbf{s}}^n$ simply follows by symmetry (see also the last paragraph in the proof). The proof idea is the following. We consider a set of m points in $\mathbb{R}^n$ that are ordered according to their index in every dimension, that is, $\mathbf{x}(i + 1)$ has in any dimension a value not less than $\mathbf{x}(i)$ in that dimension for all i. m such points exist that are linearly independent. Since all components of $\mathbf{w}$ are nonnegative, it follows that $\mathbf{x}(i + 1)^T \mathbf{w}$ is not smaller than $\mathbf{x}(i)^T \mathbf{w}$ for all i. Hence, if $\mathbf{x}(i)$ is classified positive, then for all $j > i$, $\mathbf{x}(j)$ is classified positive. The theorem follows. We make this mathematically explicit:

Consider a set $S = \{\mathbf{x}(1), \ldots, \mathbf{x}(m)\}$ of m linearly independent vectors such that for all $i \in \{1, \ldots, m - 1\}$ and all $j \in \{1, \ldots, n\}$, we have $x_j(i) \leq x_j(i + 1)$. Such a set exists. The vectors $\mathbf{x}(1), \ldots, \mathbf{x}(m)$ can, for example, be defined as $x_j(i) = \min\{1, \max\{0, i - j + 1\}\}$ $(i = 1, \ldots, n, \ j = 1, \ldots m)$.

Hence, in $\mathbf{x}(i)$, the first i components have value 1, and the other components have value 0. These vectors are obviously linearly independent, and the criterion $x_j(i) \leq x_j(i+1)$ is fulfilled. For any nonnegative weight vector $\mathbf{w}$ and any i, k with $i < k$, we have $\sum_j x_j(i) w_j \leq \sum_j x_j(k) w_j$. Hence, for any nonnegative weight vector $\mathbf{w}$, any threshold θ, and any i, k with $i < k$, if $\mathbf{x}(i)$ is classified positive, then $\mathbf{x}(k)$ is classified positive. Therefore, no more than $m + 1$ dichotomies can be produced with nonnegative weight vectors on S.

For arbitrary sign-constrained weights, one can simply multiply $\mathbf{x}(i)$ with the ith weight sign. Hence the result is true also for arbitrary sign-constrained threshold functions.

Appendix C: Proof of Theorem 3

The theorem of separating hyperplanes is the basis of the following proof (see Strang, 1988, p. 420):

Lemma 2 (theorem of separating hyperplanes). *Either $X\mathbf{v} = \mathbf{b}$ has a nonnegative solution, or there is a $\mathbf{y}$ such that $\mathbf{y}X \geq 0$, $\mathbf{y}\mathbf{b} < 0$.*

Proof: "If"-part: Consider a dichotomy $h : S \to \{-1, 1\}$. Because an ϵ-ball around $\mathbf{1}$ lies inside $C_{(X \cdot D_s)}$, there exists an $\epsilon' > 0$ such that the vector $\mathbf{b}$ of the form $\mathbf{b} = \mathbf{1} + \epsilon' \sum_{i=1}^{m} \mathbf{e}_i h(\mathbf{x}_i)$ lies inside $C_{(X \cdot D_s)}$, where $\mathbf{e}_i$ is the ith unit vector. Since $C_{(X \cdot D_s)}$ is the set of positive combinations of $(X \cdot D_s)$, there exists a nonnegative weight vector $\mathbf{v}$ such that $(X \cdot D_s)\mathbf{v} = \mathbf{b}$. Hence, the vector $\tilde{\mathbf{v}} = D_s \mathbf{v}$ satisfies the weight constraints given by $\mathbf{s}$ and $X\tilde{\mathbf{v}} = \mathbf{b}$. The dichotomy h is accomplished by setting the threshold θ to 1.

Now we show the "only if" part. We show that if all dichotomies are possible, then an ϵ-ball around $\mathbf{1}$ lies inside $C_{(X \cdot D_s)}$. If all dichotomies are possible, then for an arbitrary dichotomy $h : S \to \{-1, 1\}$, there exists a weight vector $\mathbf{v} \in \mathbb{R}_+^n$, a threshold $\theta^h > 0$, and a $\tilde{\mathbf{p}}^h \in \mathbb{R}^m$ such that $XD_s\mathbf{v} = \tilde{\mathbf{p}}^h$ and $\tilde{p}_i^h > \theta^h$ if $h(x(i)) = 1$ and $\tilde{p}_i^h < \theta^h$ if $h(x(i)) = -1$. Because a nonnegative weight vector can only project onto points in $C_{(X \cdot D_s)}$, the point $\mathbf{p}^h$ has to lie inside $C_{(X \cdot D_s)}$. Since for any point, any scaling of the point is in $C_{(X \cdot D_s)}$, the point $\mathbf{p}^h = \tilde{\mathbf{p}}^h / \theta^h$ is in $C_{(X \cdot D_s)}$ and $p_i^h > 1$ if $h(x(i)) = 1$ and $p_i^h < 1$ if $h(x(i)) = -1$.

Consider some ordering of the 2^m possible dichotomies $h_1, \ldots, h_{2^m}$. Since we can find such a point $\mathbf{p}^{h_i}$ for any of these dichotomies h_i, we can construct an $m \times 2^m$ matrix P with these points. In P, the jth column corresponds to the jth dichotomy h_j. Hence, there exists an $\epsilon' > 0$ such that for any $j \in \{1, \ldots, 2^m\}$, the jth column $\mathbf{p}(j)$ of P is such that $p_i(j) \geq 1 + \epsilon'$ for $h_j(\mathbf{x}(i)) = 1$ and $p_i(j) \leq 1 - \epsilon'$ for $h_j(\mathbf{x}(i)) = -1$.

Note that any column of P lies in $C_{(X \cdot D_s)}$. Hence, any positive combination of columns of P also lies in $C_{(X \cdot D_s)}$, which is closed under such combinations. We show that any vector $\mathbf{b}$ of the form $\mathbf{b} = \mathbf{1} + \epsilon \sum_{i=1}^{m} a_i \mathbf{e}_i$ (for $a_i \in [-1, 1]$)

lies inside $C_{(X \cdot D_s)}$ for $\epsilon \leq \epsilon'$ (this does actually not define an ϵ-ball but a hypercube of side length 2ϵ, but still a ball is inside the cube). To do this, we use lemma 1 and show that a $\mathbf{y}$ as given there does not exist, which implies that a nonnegative $\mathbf{v}$ exists and $\mathbf{b}$ lies inside $C_{(X \cdot D_s)}$. In order to achieve $\mathbf{y}^T \mathbf{b} < 0$,

$$\sum_{i=1}^{m} y_i(1 + \epsilon a_i) < 0$$

has to hold. We get

$$\sum_{i=1}^{m} y_i < -\epsilon \sum_{i=1}^{m} a_i \cdot y_i$$

$$\leq \epsilon \cdot \|\mathbf{y}\|_{L1}. \tag{C.1}$$

Let S_{pos} be the set of indices of nonnegative entries in $\mathbf{y}$ and S_{neg} be the set of indices of negative entries in $\mathbf{y}$ (i.e., $S_{pos} \cup S_{neg} = \{1, \ldots, m\}$, $y_i \geq 0$ for $i \in S_{pos}$ and $y_i < 0$ for $i \in S_{neg}$). Choose j such that $p_i(j) \geq 1 + \epsilon'$ for $i \in S_{neg}$ and $p_i(j) \leq 1 - \epsilon'$ for $i \in S_{pos}$. By lemma 2, either there exists a $\mathbf{y}$ of the form given by equation C.1 such that the product $\mathbf{y}^T \mathbf{p}(j)$ is nonnegative or $P\mathbf{v} = \mathbf{b}$ has a nonnegative solution. The product $\mathbf{y}^T \mathbf{p}(j)$ is

$$\mathbf{y}^T \mathbf{p}(j) = \sum_{i \in S_{pos}} y_i p_i(j) - \sum_{i \in S_{neg}} |y_i| p_i(j)$$

$$\leq \sum_{i \in S_{pos}} y_i(1 - \epsilon') - \sum_{i \in S_{neg}} |y_i|(1 + \epsilon')$$

$$= \sum_{i \in S_{pos}} y_i - \sum_{i \in S_{neg}} |y_i| - \sum_{i \in S_{pos}} y_i \epsilon' - \sum_{i \in S_{neg}} |y_i| \epsilon'$$

$$< \epsilon \cdot \|y\|_{L1} - \epsilon' \sum_{i} |y_i|$$

$$= \epsilon \cdot \|y\|_{L1} - \epsilon' \|y\|_{L1}.$$

Hence, for $\epsilon \leq \epsilon'$, we have $\mathbf{y}^T \mathbf{p}(j) < 0$. It follows from lemma 2 that $P\mathbf{v} = \mathbf{b}$ has a nonnegative solution. This implies that $\mathbf{b}$ lies inside $C_{(X \cdot D_s)}$.

Appendix D: Proof of Lemma 1

We first show the "if" part. We can choose m linearly independent columns in X out of $R_v(X)$ and write them in an $m \times m$ matrix X^*. We write the corresponding entries of $\mathbf{v}$ in a vector $\mathbf{v}^*$. The remaining columns of X are

collected in an $m \times (n - m)$ matrix $\bar{X}$ and the corresponding entries in $\mathbf{v}$ in $\bar{\mathbf{v}}$. We can now write the linear equation as

$$X\mathbf{v} = X^*\mathbf{v}^* + \bar{X}\bar{\mathbf{v}} = \mathbf{1}.$$

Since X^* is invertible, $\mathbf{v}^*$ is given by

$$\mathbf{v}^* = X^{*-1}(\mathbf{1} - \bar{X}\bar{\mathbf{v}}).$$

We decompose the equation $X \cdot \mathbf{v}' = \mathbf{b}$ for a vector $\mathbf{b} \in \mathbb{R}^n$ in a similar manner:

$$X\mathbf{v}' = X^*\mathbf{v}'^* + \bar{X}\bar{\mathbf{v}} = \mathbf{b}.$$

We consider vectors $\mathbf{b}$ of the form $\mathbf{b} = \mathbf{1} + \epsilon' \sum_{i=1}^{m} \mathbf{e}_i \, p_i$ with $p_i \in [-1, 1]$ defining a hypercube around $\mathbf{1}$. If for small enough ϵ', there exists a non-negative solution for $\mathbf{v}'$ for any such vector, then an ϵ-ball around $\mathbf{1}$ lies inside the cone C_X. For such $\mathbf{b}$, we have

$$\mathbf{v}'^* = X^{*-1} \cdot \left(\mathbf{1} + \epsilon' \sum_{i=1}^{m} \mathbf{e}_i \, p_i - \bar{X}\bar{\mathbf{v}} \right)$$

$$= \mathbf{v}^* + X^{*-1} \cdot \epsilon' \sum_{i=1}^{m} \mathbf{e}_i \, p_i.$$

Let x_{ij}^{-1} denote the entry in the ith row and jth column of X^{*-1}. For $0 < \epsilon' < \frac{\min_i \{v_i^*\}}{m \cdot \max_{i,j} \{|x_{ij}^{-1}|\}}$, we get for arbitrary $k \in \{1, \dots, m\}$

$$v_k'^* \geq v_k^* - \epsilon' \cdot m \cdot \max_{i,j} \left\{ |x_{ij}^{-1}| \right\}$$

$$> v_k^* - \min_i \{v_i^*\} \geq 0.$$

It follows that $\mathbf{v}'$ is nonnegative for any $\mathbf{b}$ of the form $\mathbf{b} = \mathbf{1} + \epsilon' \sum_{i=1}^{m} \mathbf{e}_i \, p_i$ with $p_i \in [-1, 1]$. Therefore, an ϵ-ball around $\mathbf{1}$ lies inside C_X.

Now we show the "only-if" part of the lemma. Since an ϵ-ball lies inside C_X, there exist m linearly independent points $\mathbf{y}(1), \dots, \mathbf{y}(m)$ such that an ϵ'-ball around $\mathbf{1}$ lies inside the cone of these points. These points could, for example, be $\mathbf{y}(i) = \mathbf{1} + \epsilon^* \mathbf{e}_i$ for $i = 1, \dots, m$ and some $\epsilon^* > 0$. Let the ith column of the $m \times m$ matrix Y be given by $\mathbf{y}(i)$ for $i = 1, \dots, m$. Y has full rank m and is given by

$$Y = X \cdot W$$

for some nonnegative matrix W. W might have rows with zero entries only. However, rows of W with nonzero entries correspond to a set of columns of X consisting of m linearly independent vectors. More precisely, the set

$$R_W(X) = \{\text{column } i \text{ of } X | \exists j : w_{ij} > 0\}$$

contains m linearly independent vectors. Otherwise, Y could not have full rank.

There exists a unique and nonnegative $\mathbf{v}^*$ such that $Y \cdot \mathbf{v}^* = \mathbf{1}$. Since $Y = X \cdot W$, we have

$$X \cdot W \cdot \mathbf{v}^* = \mathbf{1}.$$

We consider the weight vector $\mathbf{v} = W \cdot \mathbf{v}^*$ (hence, we have $X \cdot \mathbf{v} = \mathbf{1}$). We assume that $\mathbf{v}^*$ is positive in the following (this will be proven at the end of this proof). Since W is nonnegative, it follows from the assumption that $\mathbf{v}^*$ is positive that $\mathbf{v}$ has nonzero entries for rows of W that are not entirely consisting of zeros. Hence,

$$R_{\mathbf{v}}(X) = \{\text{column } i \text{ of } X | v_i \neq 0\} = \{\text{column } i \text{ of } X | \exists j : w_{ij} > 0\} = R_W(X).$$

Hence, under the assumption that $\mathbf{v}^*$ is positive, $R_{\mathbf{v}}(X)$ contains m linearly independent vectors (see arguments about $R_W(X)$ given above).

It remains to be shown that $\mathbf{v}^*$ is positive. The solution to $Y\mathbf{v}^* = \mathbf{1}$ is unique, and $v_i^* \geq 0$ holds for $i = 1, \ldots, m$. In the following we consider the case of zero entries in $\mathbf{v}^*$. Recall that an ϵ-ball around $\mathbf{1}$ lies inside C_Y and that $\mathbf{v}^*$ is the solution to $Y\mathbf{v}^* = \mathbf{1}$. Assume that $v_i^* = 0$ for some $i \in \{1, \ldots, m\}$. Since an ϵ-ball around $\mathbf{1}$ lies inside C_Y, there exists an $\epsilon' > 0$ such that for arbitrary $j \in \{1, \ldots, m\}$ and $p \in \{-1, 1\}$, there exists a nonnegative $\mathbf{v}'(j, p)$ such that

$$\mathbf{v}'(j, p) = Y^{-1}\left(\mathbf{1} + \epsilon' \mathbf{e}_j p\right)$$
$$= Y^{-1}\mathbf{1} + Y^{-1}\epsilon' \mathbf{e}_j p$$
$$= \mathbf{v}^* + Y^{-1}\epsilon' \mathbf{e}_j p.$$

The ith component of $\mathbf{v}'(j, p)$ is therefore (note that i was chosen such that $v_i^* = 0$)

$$v_i'(j, p) = \epsilon' \mathbf{e}_i^T Y^{-1} \mathbf{e}_j p$$
$$= \epsilon' \cdot p \cdot c(j),$$

where $c(j) = \mathbf{e}_i^T Y^{-1} \mathbf{e}_j$. We can find some j such that $c(j) \neq 0$ since otherwise Y^{-1} would have a row consisting of zeros only, which implies that it is not invertible. Hence, we can choose a $j \in \{1, \ldots, m\}$ and a $p \in \{-1, 1\}$ such that $v_i'(j, p) < 0$ which leads to a contradiction. Hence, the assumption that $\mathbf{v}^*$ is positive holds, which finally proves the lemma.

Acknowledgments

References

Amit, D. J., Campbell, C., & Wong, K. Y. M. (1989). The interaction space of neural networks with sign-constrained synapses. *J. Phys. A: Math. Gen., 22*, 4687–4693.

Amit, D. J., Wong, K. Y. M., & Campbell, C. (1989). Perceptron learning with sign-constrained weights. *J. Phys. A: Math. Gen., 22*, 2039–2045.

Bartlett, P. L., & Maass, W. (2003). Vapnik-Chervonenkis dimension of neural nets. In M. A. Arbib (Ed.), *The handbook of brain theory and neural networks* (2nd ed., pp. 1188–1192). Cambridge: MIT Press.

Brunel, N., Hakim, V., Isope, P., Nadal, J. P., & Barbour, B. (2004). Optimal information storage and the distribution of synaptic weights: Perceptron versus Purkinje cell. *Neuron, 48*, 745–757.

Gardner, E. (1987). The space of interactions in neural network models. *Journal of Physics A, 21*, 257–270.

Haykin, S. (1999). *Neural networks: A comprehensive foundation* (2nd ed.). Upper Saddle River, NJ: Prentice Hall.

Legenstein, R. A., & Maass, W. (2006). *Neural codes that enhance the discrimination capability of readout neurons*. in preparation.

McCulloch, W. S., & Pitts, W. (1943). A logical calculus of ideas immanent in nervous activity. *Bulletin of Mathematical Biophysics, 5*, 115–133.

Minsky, M., & Papert, S. (1988). *Perceptrons: An introduction to computational geometry*. Cambridge, MA: MIT Press.

Rosenblatt, J. F. (1962). *Principles of neurodynamics*. New York: Spartan Books.

Strang, G. (1988). *Linear algebra and its applications* (3rd ed.). San Diego, CA: Harcourt Brace Jovanovich.

Vapnik, V. N. (1998). *Statistical learning theory*. New York: Wiley.

Vapnik, V. N., & Chervonenkis, A. Y. (1971). On the uniform convergence of relative frequencies of events to their probabilities. *Theory of Probability and Its Applications, 16*(2), 264–280.