Iterative Learning and Denoising in Convolutional Neural Associative Memories

Abstract

The task of a neural associative memory is to retrieve a set of previously memorized patterns from their noisy versions by using a network of neurons. Hence, an ideal network should be able to 1) gradually learn a set of patterns, 2) retrieve the correct pattern from noisy queries and 3) maximize the number of memorized patterns while maintaining the reliability in responding to queries. We show that by considering the inherent redundancy in the memorized patterns, one can obtain all the mentioned properties at once. This is in sharp contrast with the previous work that could only improve one or two aspects at the expense of the third. More specifically, we devise an iterative algorithm that learns the redundancy among the patterns. The resulting network has a retrieval capacity that is exponential in the size of the network. Lastly, by considering the local structures of the network, the asymptotic error correction performance can be made linear in the size of the network.

1. Introduction

The ability to memorize a large set of patterns and reliably retrieve them in the presence of noise, are among the main reasons that attracted a large body of research to neural networks for the past three decades. Ideally, a perfect neural associative memory should be able to *learn* patterns, have a large pattern retrieval *capacity* and be *noise-tolerant*. This problem, called "associative memory", is in spirit very similar to reliable information transmission faced in communication systems where the goal is to efficiently decode a set of transmitted patterns over a noisy channel.

Despite this similarity and common methods deployed in both fields (e.g., graphical models, iterative algorithms, to name a few), we have witnessed a huge gap between the efficiency achieved by them. More specifically, by deploying modern coding techniques, it was shown that the number of reliably transmitted patterns over a noisy channel can be made *exponential* in n, the length of the patterns. This was achieved by intelligently imposing redundancy among transmitted patterns. In contrast, the maximum number of patterns that can be reliably memorized by most current neural networks scales *linearly* in the size of the network. This is due to the common assumption that a neural network should be able to memorize *any* subset of patterns drawn randomly from the set of all possible vectors of length n (see, for example, (Hopfield, 1982), (Venkatesh & Psaltis, 1989), (Jankowski et al., 1996), (Muezzinoglu et al., 2003)).

Recently, (Kumar et al., 2011) suggested a new formulation of the problem where only a suitable set of patterns was considered for storing. To enforce the set of constraints, they formed a bipartite graph (as opposed to a complete graph considered in the earlier work) where one layer feeds the patterns to the network and the other takes into account the inherent structure. The role of bipartite graph is indeed similar to the Tanner graphs used in modern coding techniques (R. Tanner , 1982). Using this model, (Kumar et al., 2011) provided evidence that the resulting network can potentially memorize a substantial number of patterns at the expense of correcting only a *single* error during the recall phase. Later, (Salavati & Karbasi, 2012) further improved this result by introducing a multi-layer structure that can correct a *constant* number of errors.

In this work, we introduce a novel neural architecture equipped with an online learning algorithm. This architecture can correct a *linear* fraction of errors while keeping the storage capacity *exponential* in the size of the network. This is achieved by memorizing a set of patterns with some degrees of redundancy, i.e., those with very weak minor components. In other words, we find dual vectors that are (almost) orthogonal to particular parts of input patterns. By making use of this inherent redundancy, we can increase the number of memorized patterns from linear to exponen-

Preliminary work. Under review by the International Conference on Machine Learning (ICML). Do not distribute.

tial. Our learning algorithm is an extension of the subspace learning method proposed by (Oja & Kohonen, 1986), with an additional property of imposing the learned vectors *sparse*. The sparsity property will become helpful during the noise-elimination phase.

We will provide theoretical analysis to support our claims and also assess the accuracy of our results through simulations. In particular, we evaluate the performance of our proposed algorithms over a dataset of spoken English words, as well as synthetic datasets, and compare them with prior art.

2. Related Work

Arguably, the Hopfield network is the first auto-associative neural mechanism capable of learning a set of patterns and recalling them later (Hopfield, 1982). By utilizing the Hebbian learning rule, Hopfield considered a neural network of size n with binary state neurons. It was shown by McEliece et al. (1987) that the capacity of a Hopfield network is bounded by $\mathcal{C} = (n/2\log(n))$.

With the hope of increasing the capacity of the Hopfield network, an extension of associative memories to non-binary states has also been explored in the past. In particular, Jankowski et al. (1996) investigated a complex-valued neural associative memory where each neuron can be assigned a multivalued state from the set of complex numbers. It was shown by Muezzinoglu et al. (2003) that the capacity of such networks can be increased to $\mathcal{C} = n$ at the cost of a prohibitive weight computation mechanism.

Recently, in order to increase the capacity and robustness, a line of work considered exploiting the inherent structure of the patterns. This is done by either making use of the correlations among the pattern or memorizing only those patterns that have some sort of redundancy. Note that they differ from the previous methods in one important aspect: not any possible set of patterns is considered for learning. By utilizing neural cliques, (Gripon & Berrou, 2011) were among the first who demonstrated that considerable improvements in the pattern retrieval capacity of Hopfield networks is possible, albeit still not passing the polynomial boundary on the capacity, i.e. $C = O(n^2)$.

By deploying higher order neural models, in contrast to the pairwise correlation considered in Hopfield networks, (Peretto & Niez, 1986) showed that the storage capacity can be improved to $\mathcal{C} = O(n^{p-2})$, where p is the degree of correlation. In such models, the state of the neurons not only depends on the state of their neighbors, but also on the correlations among them. However, the main drawback of this work lies in the prohibitive computational complexity of the learning phase. To address this difficulty, while being able to capture higher-order correlations, a new model based on bipartite graphs was introduced by (Kumar et al., 2011), and was further explored by (Salavati & Karbasi, 2012). Under the restrictive assumptions that the bipartite graph is fully known, sparse, and expander, the proposed algorithm by (Kumar et al., 2011) increased the pattern retrieval capacity to $\mathcal{C} = O(a^n)$, for some $a > 1$. The lack of a learning algorithm that satisfies all the mentioned requirements are among the reasons that reduces the capabilities of this model.

In this paper, we propose a novel architecture to capture the internal redundancy by deviding the input patterns into *overlapping* clusters/patches. We first devise an online learning algorithm that effectively learns the structure of the bipartite graph and show that the capacity of the resulting network is exponential in its size. We then propose a novel error correction algorithm that corrects a linear fraction of errors in the recall phase. Our analytical results, supported by simulations, demonstrate that introducing overlaps among clusters improves the error correction performance significantly.

It is worth mentioning that learning a set of input patterns with robustness against noise is not just the focus of neural associative memory. For instance, (Vincent et al., 2008) proposed an interesting approach to extract robust features in autoencoders. Their approach is based on *artificially introducing* noise during the learning phase and let the network learn the mapping between the corrupted input and the correct version. This way, they shift the burden from the recall phase to the learning phase. We, on the other hand, consider a more particular form of redundancy and enforce a suitable structure which helps us design algorithms that are faster and *guaranteed* to correct a linear fraction of noise without previously being exposed to.

We should also mention that learning linear constraints by a neural network is hardly a new topic. It is shown by (Xu et al., 1991) and (Oja & Kohonen, 1986) that one can learn a matrix orthogonal to a set of patterns in the training set by using simple neural learning rules. However, to our knowledge, finding such a matrix subject to sparsity has not been addressed before. This problem can also be regarded as an instance of compressed sensing (Candes & Tao, 2006), (Donoho, 2006). With few exceptions, including (Donoho et al., 2009) and (Tropp & Wright, 2010), the decoders proposed in the compressed sensing area are far too complicated to be implementable by a neural network only

capable of performing simple operations, i.e., weighted sum of a nonlinear function (see Section 3).

Deep Belief Networks: Our neural architecture is in spirit similar to those of Deep Belief Networks (DBN). DBNs are typically used to extract/classify features by the means of several consecutive stages (e.g., pooling, rectification, etc). Having multiple stages help the network to learn more interesting and complex features. An important class of DBNs are convolutional DBNs where the input layer (or the receptive field) is divided into multiple possibly-overlapping patches, and the network extract features from each patch (Jarrett et al., 2009).

Since we divide the input patterns into a few overlapping smaller clusters, our model is similar to those of convolutional DBNs. Furthermore, we also learn multiple *features* (i.e., dual vectors) from each patch where the feature extractions differ over different patches. This is indeed very similar to (Le et al., 2010).

In contrast to convolutional DBNs, the focus of this work is not classification but rather recognition of the *exact* patterns from their noisy versions. Moreover, in most DBNs, we not only have to find the proper dictionary for classification, but we also need to calculate the features for each input pattern. This alone increases the complexity of the whole system, especially if denoising is part of the objective. In our model, however, the dictionary is defined in terms of dual vectors. Consequently, previously memorized patterns are computationally easy to recognize as they yield the all-zero vector in the output of the feature extraction stage. In other words, a non-zero output can only happen if the input pattern is noisy. Another advantage of our model over DBNs is a much faster learning phase. More precisely, by using a single overlapping layer in our model the information diffuses gradually in the network. The same criteria is achieved in DBNs (Socher et al., 2011) by constructing several stages.

3. Notation and Definitions

In neural associative memories - the subject of this work - the goal is to design a neural network capable of memorizing a large set of patterns from a data set $\mathcal{X}$ (learning phase), and recalling them later in presence of noise (recall phase).

Learning phase: Each pattern $x = (x_1, x_2, \ldots, x_n)$ is a vector of length n, where $x_i \in \mathcal{S}$ for $i \in [n]$. In this work, our focus is on memorizing the patterns with strong *local correlation* among the entries. More specifically, we divide the entries of each pattern x into L *overlapping* sub-patterns of lengths $n_1, \ldots, n_L$, so that $\sum n_i \geq n$. Note that due to overlaps, a pattern node can be a member of multiple sub-patterns, as shown in Figure 1. We denote the i-th sub-pattern by $x^{(i)} = (x_1^{(i)}, x_2^{(i)}, \ldots, x_{n_i}^{(i)})$. To enforce local correlations, we assume that the sub-patterns $x^{(i)}$ form a subspace of dimension $k_i < n_i$. This is done by impose linear constraints on each cluster. These linear constraints are captured in the form of dual vectors as follows. In the learning phase, we would like to memorize these patterns by finding a set of non-zero vectors $w_1^{(i)}, w_2^{(i)}, \ldots, w_{m_i}^{(i)}$ that are orthogonal to the set of sub-patterns $x^{(i)}$, i.e.,

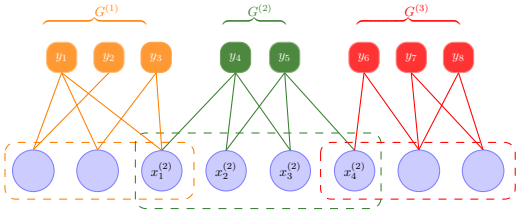

Figure 1. Bipartite graph G.

$$y_j^{(i)} = (w_j^{(i)})^\top \cdot x^{(i)} = 0, \quad \forall j \in [m_i] \,\&\, \forall i \in [L], \quad (1)$$

where $[q]$ denote the set of first q integers, namely, $[q] = \{1, 2, \cdots, q\}$. The weight matrix $W^{(i)} = [w_1^{(i)} | w_2^{(i)} | \ldots | w_{m_i}^{(i)}]^\top$ of cluster i is created by putting all the dual vectors next to each other. Equation (1) can be written equivalently as $W^{(i)} \cdot x^{(i)} = 0$. Therefore, the goal is to devise a low complexity algorithm that updates the weights of $W^{(i)}$'s (for $i \in [L]$) once it encounters a new pattern $x \in \mathcal{X}$. We should stress here that contrary to our iterative algorithm, many proposed learning algorithms in the literature can learn a set of patterns only after the whole set is presented to the neural network. More importantly, they are usually unable to learn a new set of patterns. Removing this restriction is one of the major contributions of our work.

One can easily map the local constraints imposed by $W^{(i)}$'s into a global constrain by introducing a global weight matrix W of size $m \times n$. The first m_1 rows of the matrix W correspond to the constraints in the first cluster, the rows $m_1 + 1$ to $m_1 + m_2$ correspond to the constraints in the second cluster, and so forth. Hence, by inserting the zero entries at proper positions, we can construct the global constraint matrix W. We will use both the local and global connectivity matrices to eliminate noise in the recall phase.

Recall phase: Once the set of patterns $\mathcal{X}$ have been memorized, a desirable neural network should be able to retrieve an already memorized pattern from its corrupted version (of course, if the intensity of noise is

not moderate, there is no hope for recovery). Needless to say that the recall process should be implementable by simple operations mentioned earlier.

In the recall, phase a noisy version, say y, of an already learned pattern $x \in \mathcal{X}$ is given. Here, we assume that the noise is an additive vector of size n, denoted by e, whose entries assume values independently from $\{-1, 0, +1\}$ with corresponding probabilities $p_{-1} = p_{+1} = p_e/2$ and $p_0 = 1 - p_e$. To ensure that the amount of noise is moderate, we also need to assume that the error probability, p_e, is less than p_0. As before, we denote by $e^{(i)}$, the realization of noise on the sub-pattern $x^{(i)}$. In formula, $y = x + e$ (modulo S). Note that $W \cdot y = W \cdot e$ and $W^{(i)} \cdot y^{(i)} = W^{(i)} \cdot e^{(i)}$. Therefore, the goal will be to remove the noise e and obtain the desired pattern x as the true states of the pattern neurons. This task will be accomplished by exploiting the fact that we have chosen the set of patterns $\mathcal{X}$ to satisfy the set of constraints $W^{(i)} \cdot x^{(i)} = 0$.

Capacity: The last issue we look at in this work is the retrieval capacity $\mathcal{C}$ of our proposed method. Formally, the retrieval capacity is defined in terms of the maximum number of patterns that a neural network can learn and distinguish later. By construction, we show that the retrieval capacity of our network is exponential in the size of the network.

4. The Learning Algorithm

In this section, we discuss the learning algorithm for a given cluster ℓ. The case of other clusters would be a straightforward extension of the suggested algorithm. Since the patterns lie in a subspace of dimension $k_\ell \leq n_\ell$, we adapt the algorithm proposed in (Oja & Karhunen, 1985) and (Xu et al., 1991) to learn the null-space of the subspace defined by the patterns. Due to requirements of the denoising algorithm used in the recall phase, we also need the dual vectors be sparse. To this end, we add an additional term to penalize non-sparse solutions during the learning phase. Furthermore, we are not looking for an orthogonal basis as in (Xu et al., 1991). Instead, we would like to find m_ℓ vectors that are orthogonal to the patterns. The problem to find a constraint *vector* $w^{(\ell)}$ is given by

$$\min_{w^{(\ell)}} \sum_{x^{(\ell)} \subset x \in \mathcal{X}} |x^{(\ell)} \cdot w^{(\ell)}|^2 + \beta g(w^{(\ell)}), \quad \text{s.t. } \|w^{(\ell)}\|_2 = 1. \tag{2}$$

In the above problem, $x \in \mathcal{X}$ is a pattern drawn from the training set, β is a positive constant and $g(.)$ is the penalty term to favor sparse results. For instance one can pick $g(w^{(\ell)}) = \|w^{(\ell)}\|_1$, which leads to the ℓ_1-norm penalty, widely used in compressed sensing (Donoho, 2006), (Candes & Tao, 2006). In this paper, however, we find it more appropriate to consider

$$g(w^{(\ell)}) = \sum_{i=1}^{n} \tanh(\sigma(w_i^{(\ell)})^2).$$

Intuitively, for large σ, $\tanh(\sigma(w_i^{(\ell)})^2)$ approximates $|\text{sign}(w_i^{(\ell)})|$. Therefore, the larger σ gets, the closer $g(w^{(\ell)})$ will be to $\|\cdot\|_0$. By calculating the derivative of the objective function, and by considering the update required for each randomly picked pattern x, we will obtain the following iterative algorithm:

$$y^{(\ell)}(t) = x^{(\ell)}(t) \cdot w^{(\ell)}(t), \tag{3}$$

$$\tilde{w}^{(\ell)}(t+1) = w^{(\ell)}(t) - \alpha_t \Big(2y^{(\ell)}(t)x^{(\ell)}(t) + \beta\Gamma(w^{(\ell)}(t))\Big), \tag{4}$$

$$w^{(\ell)}(t+1) = \frac{\tilde{w^{(\ell)}}(t+1)}{\|\tilde{w}^{(\ell)}(t+1)\|_2}. \tag{5}$$

Where t is the iteration number, $x^{(\ell)}(t)$ is the sub-pattern of a pattern $x(t)$ drawn at iteration t, α_t is a small positive constant and $\Gamma(w^{(\ell)}) = \nabla g(w^{(\ell)})$, the gradient of the penalty term. This function has the interesting property that suppresses very small values of $w_i^{(\ell)}(t)$. To see why, consider the i-th entry of $\Gamma(w^{(\ell)}(t)))$, namely,

$$\begin{aligned}\Gamma_i(w^{(\ell)}(t)) &= \partial g(w^{(\ell)}(t))/\partial w_i^{(\ell)}(t) \\ &= 2\sigma_t w_i^{(\ell)}(t)(1 - \tanh^2(\sigma w_i^{(\ell)}(t)^2)).\end{aligned}$$

It is easy to see that $\Gamma_i(w(t)) \simeq 2\sigma w_i(t)$ for relatively small values of $w_i(t)$, and $\Gamma_i(w^{(\ell)}(t)) \simeq 0$ for larger values of $w_i^{(\ell)}(t)$. Thus, for proper choices of β and σ, Eq (4) suppresses small entries of $w^{(\ell)}(t)$ towards zero and favors sparser results. To simplify the analysis, with some abuse of notations, we approximate the function $\Gamma(w^{(\ell)}(t))$ with the following function:

$$\Gamma_i(w^{(\ell)}(t)) = \begin{cases} w_i^{(\ell)}(t) & \text{if } |w_i^{(\ell)}(t)| \leq \theta_t; \\ 0 & \text{otherwise.} \end{cases} \tag{6}$$

Where θ_t is a small positive threshold.

Following the same approach as in (Oja & Karhunen, 1985) we assume α_t to be small enough such that Eq (5) can be expanded as powers of α_t. This leads us to a simpler version of equations (3-5) as follows.

$$y^{(\ell)}(t) = x^{(\ell)}(t) \cdot w^{(\ell)}(t) \tag{7}$$

$$\begin{aligned} w^{(\ell)}(t+1) &= w^{(\ell)}(t) - \alpha_t(y^{(\ell)}(t)(x^{(\ell)}(t) \\ &- \frac{y^{(\ell)}(t)w^{(\ell)}(t)}{\|w^{(\ell)}(t)\|_2^2}) + \beta\Gamma(w^{(\ell)}(t))) \end{aligned} \tag{8}$$

Algorithm 1 Iterative Learning

Input: Dataset $\mathcal{X}$ with $|\mathcal{X}| = \mathcal{C}$, stopping point ε.
Output: $w^{(\ell)}$

while $\sum_{x\in\mathcal{X}} |x^{(\ell)} \cdot w^{(\ell)}(t)|^2 > \varepsilon$ **do**
  Choose pattern $x(t)$ uniformly at from $\mathcal{X}$.
  Compute $y^{(\ell)}(t) = x^{(\ell)}(t) \cdot w(t)$.
  Update $w^{(\ell)}(t)$ according to Eq (8).
  $t \leftarrow t + 1$.
end while

In the above approximation, we also omitted the term $\alpha_t\beta\left(w^{(\ell)}(t)\cdot\Gamma(w^{(\ell)}(t))\right)w^{(\ell)}(t)$ since $w^{(\ell)}(t)\cdot\Gamma(w^{(\ell)}(t))$ would be negligible.

The resulting learning algorithm for one constraint node is shown in Algorithm 1. In words, $y^{(\ell)}(t)$ is the projection of $x^{(\ell)}(t)$ onto $w^{(\ell)}(t)$. If for a given data vector $x^{(\ell)}(t)$, the projection $y^{(\ell)}(t)$ is non-zero, then the weight vector will be updated in order to reduce this projection.

To prove the convergence of Algorithm 1, we benefit from the convergence of Stochastic Gradient Descent (SGD) algorithms (Bottou, 1998). Let $E(w^{(\ell)}) = \sum_{x\in\mathcal{X}} |x^{(\ell)} \cdot w^{(\ell)}|^2$ denote the cost function. Furthermore, let $A_x = x^{(\ell)}(x^{(\ell)})^\top$, and $A = \mathbb{E}\{A_x | x \in \mathcal{X}\}$ represent the correlation among patterns in the training set $\mathcal{X}$. Since patterns are uniformly sampled from $\mathcal{X}$, one can rewrite $E(w^{(\ell)}) = (w^{(\ell)})^\top A w^{(\ell)}/\mathcal{C}$.

Theorem 1. *Under the following regularity conditions*

- *[A1.]* $\|A\|_2 \le \Upsilon < \infty$,
- *[A2.]* $\sup_{x\in\mathcal{X}} \|A_x\|_2 \le \zeta < \infty$,
- *[A3.]* $\alpha_t \ge 0$, $\sum \alpha_t = \infty$,
- *[A4.]* $\sum \alpha_t^2 < \infty$,
- *[A5.] at each iteration* t, $2\alpha_t\beta < 1$,

Algorithm 1 converges to a local minimum $\hat{w}^{(\ell)}$ *for which* $\nabla E(\hat{w}^{(\ell)}) = 0$. *Furthermore, at this local minimum, the vector* $\hat{w}^{(\ell)}$ *is orthogonal to all the patterns of the training set, i.e.* $A\hat{w}^{(\ell)} = 0$.

The proof uses the results of (Bottou, 1998) to show the convergence to a local minimum proving that the weight vector at this local minimum is orthogonal to the patterns of the training set. Furthermore, condition A5 ensures that Algorithm 1 does not converge to the trivial solution $\hat{w}^{(\ell)} = 0$. The full proof could be found in the supplementary materials.

In order to find m_ℓ constraints required by the learning phase, we need to run Algorithm 1 several times.

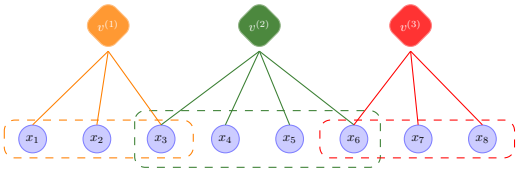

Figure 2. Contraction graph $\widetilde{G}$ corresponding to graph G in Figure 1.

In practice, we can perform this process in parallel, to speed up the learning phase. It is also more meaningful from a biological point of view, as each constraint neuron can act independently from the others. Although running Algorithm 1 in parallel may result in redundant constraints, our experimental results show that by starting from different random initial points, the algorithm converges to linearly independent constraints.

5. Recall Phase

The recall phase of our proposed method consists of two parts, intra-module and inter-module. In the intra-module part, each cluster tries to remove noise from its sub-pattern. As we will show in Section 5.1, this is indeed possible if the sub-patterns experience a single error. In case of successful decoding, the pattern neurons keep their states and revert back to their original states otherwise. During the inter-module decoding, once the correction in cluster ℓ finishes, cluster $\ell + 1$ starts and this process continues in several rounds. Here, the overlapping structure of the clusters helps them correct a linear number of errors together.

5.1. Intra-Module Recall Algorithm

In the intra-module part, we exploit the fact that the connectivity matrix of the neural network in each cluster is sparse and orthogonal to the memorized patterns. For a cluster ℓ, let $x^{(\ell)} + e^{(\ell)}$ be the given noisy input pattern. Recall that $W^{(\ell)}(x^{(\ell)}+e^{(\ell)}) = W^{(\ell)}e^{(\ell)}$.

Algorithm 2 performs a series of forward and backward iterations to remove $e^{(\ell)}$. At each iteration, the pattern neurons decide locally to whether update their current state or not: if the amount of feedback received by a pattern neuron exceeds a threshold, the neuron updates its state, and remains intact, otherwise.[1]

Theorem 2. *Algorithm 2 corrects at least a single error of cluster* $G^{(\ell)}$ *with probability at least* $1 -$

[1] In order to maintain the current value of a neuron, we can add self-loops to pattern neurons in Figure 1 (the self-loops are not shown in the figure for the sake of clarity).

Algorithm 2 Intra-Module Error Correction

Input: Training set $\mathcal{X}$, threshold φ, iteration $t_{\max}$
Output: $x_1^{(\ell)}, x_2^{(\ell)}, \ldots, x_{n_\ell}^{(\ell)}$

1: **for** $t = 1 \to t_{\max}$ **do**
2: *Forward iteration:* Calculate the weighted input sum $h_i^{(\ell)} = \sum_{j=1}^{n_\ell} W_{ij}^{(\ell)} x_j^{(\ell)}$, for each neuron $y_i^{(\ell)}$ and set $y_i^{(\ell)} = \text{sign}(h_i^{(\ell)})$.
3: *Backward iteration:* Each neuron $x_j^{(\ell)}$ computes

$$g_j^{(\ell)} = \frac{\sum_{i=1}^{m_\ell} W_{ij}^{(\ell)} y_i^{(\ell)}}{\sum_{i=1}^{m_\ell} |W_{ij}^{(\ell)}|}.$$

4: Update the state of each pattern neuron j according to

$$x_j^{(\ell)} = x_j^{(\ell)} + \text{sign}(g_j^{(\ell)})$$

only if $|g_j^{(\ell)}| > \varphi$.
5: $t \leftarrow t + 1$
6: **end for**

$\left(\frac{d_{avg}^{(\ell)}}{m}\right)^{d_{\min}^{(\ell)}}$ *as* $\varphi \to 1$, *where* $d_{avg}^{(\ell)}$ *and* $d_{\min}^{(\ell)}$ *are the average and minimum pattern-node degrees.*

The proof can be found in the supplementary material. In short, we bound the probability of correcting a single error, P_{c_1}, in terms of the probability that two nodes share the same neighborhood in a cluster.

Theorem 2 implies that to have one error corrected with high probability, we must make sure to have small average and large minimum degrees among the pattern nodes. To simplify the analysis of Section 5.2, we make a conservative assumption that a cluster is capable of correcting only a single error with overwhelming probability, and declares failure in case of multiple errors. In practice, as it is confirmed by our simulations, clusters are able to correct more than one error.

As stated earlier, the efficiency of Algorithm 2 relies on the assumption that the neural network is sparse. To gain some insight, consider an extreme case where the bipartite graph is complete. Then, a single error results in the violation of all constraint neurons in the forward iteration. Therefore, in the backward iteration, all the pattern neurons receive feedback from their neighbors. This makes it impossible to tell which pattern neuron is the noisy one. On the other hand, once the graph is sparse, a single error makes only a small set of constraint neurons unsatisfied. Consequently, in the backward iteration, only the pattern nodes that share the same neighborhood with the noisy one receive feedback. Note that in this case, the fraction of the received feedback will be much larger for the true noisy neuron. Therefore, by merely looking at the fraction of received feedback from the constraint neurons, one can identify the noisy pattern neuron.

Algorithm 3 Sequential Peeling Algorithm

Input: $\widetilde{G}, G^{(1)}, G^{(2)}, \ldots, G^{(L)}$.
Output: $x_1, x_2, \ldots, x_n$

1: **while** there is an unsatisfied $v^{(\ell)}$ **do**
2: **for** $\ell = 1 \to L$ **do**
3: If $v^{(\ell)}$ is unsatisfied, apply Algorithm 2 to cluster $G^{(l)}$.
4: If $v^{(\ell)}$ remained unsatisfied, revert the state of pattern neurons connected to $v^{(\ell)}$ to their initial state. Otherwise, keep their current states.
5: **end for**
6: **end while**
7: Declare $x_1, x_2, \ldots, x_n$ if all $v^{(\ell)}$'s are satisfied. Otherwise, declare failure.

5.2. Inter-Module Recall Algorithm

In the previous section, we showed that a single error inside a cluster can be corrected with high probability. In fact, a finer analysis reveals that there exists a threshold $T_e > 1$ such that any number of errors less than T_e can be corrected by the whole network. Since clusters have overlaps with one another, removing an error from a cluster, can potentially help the others.

Our proposed algorithm is based on a famous error correction decoder called *peeling algorithm*. To begin, we consider the contracted graph $\widetilde{G}$ in which for each cluster $G^{(\ell)}$, we contract its set of constraint nodes $y_1^{(\ell)}, \ldots, y_{m_\ell}^{(\ell)}$ into a single node $v^{(\ell)}$ (see Figure 2). The contraction graph $\widetilde{G}$ is closely related to recursive autoencoders introduced in (Socher et al., 2011). Let us denote the degree distributions of $\widetilde{G}$ (from the edge perspective) by $\widetilde{\lambda}$ and $\widetilde{\rho}$. We say that the node $v^{(\ell)}$ is unsatisfied if it is connected to a noisy pattern node.

The asymptotic performance of our error recovery method, shown in Algorithm 3, is given by the following theorem.

Theorem 3. *Under the assumptions that graph* $\widetilde{G}$ *grows large and it is chosen randomly with degree distributions given by* $\widetilde{\lambda}$ *and* $\widetilde{\rho}$, *Algorithm 3 is successful if* $p_e \cdot \widetilde{\lambda}(1 - \widetilde{\rho}(1 - z)) < z$ *for* $z \in (0, p_e)$.

The proof is based on the density evolution technique (Richardson & Urbanke, 2008; Luby et al., 2001), detailed in the Supplementary Materials.

The condition given in Theorem 3 can be used to calculate the maximal fraction of errors Algorithm 3 can correct for the given degree distributions. For instance, for the the degree distribution pair $(\widetilde{\lambda}(z) = z^2, \widetilde{\rho}(z) = z^5)$, the threshold is $p_e \approx 0.429$, below which Algorithm 3 corrects all the errors with high probability. Note that the predicted threshold by Theorem 3 is based on the assumption that a cluster can only correct a single error. In practice, as we noted earlier, a cluster can correct more. Hence, the threshold predicted by Theorem 3 is a lower bound on the overall recall performance of our neural network (as the size of the network grows).

6. Pattern Retrieval Capacity

Note that the number of patterns $\mathcal{C}$ does not have any effect on the learning or recall algorithm except for its obvious influence on the learning time. As long as the patterns come from a subspace, learning Algorithm 1 will yield a matrix W which is orthogonal to all the patterns of the training set. In the recall phase, the proposed denoising algorithms 2 and 3 deal only with $W \cdot e$, where e is the noise vector. In order to show that the pattern retrieval capacity is exponential in n, all we need to demonstrate is that there exists a training set $\mathcal{X}$ with $\mathcal{C}$ patterns of length n for which $\mathcal{C} \propto a^{rn}$, for some $a > 1$ and $0 < r$.

Theorem 4. *Let $\mathcal{X}$ be a $\mathcal{C} \times n$ matrix, formed by C vectors of length n with entries from the set $\mathcal{S}$. Furthermore, let $k = rn$ for some $0 < r < 1$. Then, there exists a set of vectors for which $\mathcal{C} = a^{rn}$, with $a > 1$, and rank$(\mathcal{X}) = k < n$.*

The proof of this theorem is by construction. The details are outlined in the Supplementary Materials.

7. Simulation Results

We have performed simulations over synthetic and natural databases to investigate the performance of the proposed algorithm and confirm the accuracy of our theoretical analysis.

7.1. Synthetic Dataset

There is a systematic way of generating patterns satisfying a set of linear constraints. More specifically, we generate a matrix $G \in \mathbb{R}^{k \times n}$ of rank $k = r \cdot n$ $(0 < r < 1)$ such that all the entries be non-negative and lie between 0 and $\gamma - 1 > 0$. We construct the patterns of the dataset by setting $x = u \cdot G$, where $u \in \mathbb{R}^k$ is an integer-valued random vector whose entries lie between 0 and $\upsilon - 1 > 0$. We select γ, υ, and G such that all entries of x be less than S. This realization is based on the constructive proof of Theorem 4 and more details can be found in the supplementary materials.

In our simulations, we assume that each pattern neuron is connected to approximately 5 clusters. The number of connections should be neither too small (to ensure information propagation) nor too big (to adhere to the sparsity requirement).

In the learning phase, Algorithm 1 is run in parallel for each cluster which results in learning the constraints. In the recall phase, at each round, a pattern x is sampled uniformly at random from the training set. Then, each of its entries gets corrupted independently with probability p_e. Afterwards, Algorithm 3 is used to denoise the corrupted pattern. We repeat this process many times to calculate the error rate, and compare it to the bound derived in section 5.2.

Learning Results: The left panels in Figures 3 illustrate the degree distributions of pattern and constraint neurons, respectively, over an ensemble of 5 randomly generated simulation setups. The horizontal axis shows the normalized degree of pattern (resp., constraint) neurons and the vertical axis represents the fraction of neurons with the given normalized degree. The parameters for the learning algorithm are $\alpha_t \propto 0.95/t$, $\beta = 0.75/\alpha_t$ and $\theta_t = 0.05$.

We have executed the learning algorithm for different network sizes and learning parameters. Qualitatively, they all look similar to Figure 3. Moreover, in almost all cases, the learning phase converged within two learning iterations, i.e. by going over the data set only twice.

Recall Results: The top panel of Figure 3 illustrates the performance of the recall algorithm. The horizontal and vertical axes represent the number of initial erroneous neurons and the final pattern error rate, respectively. The performance is compared with the theoretical bound derived in section 5.2, as well as the results of the algorithms proposed in (Salavati & Karbasi, 2012) and (Kumar et al., 2011). Note the slight difference between the theoretical and simulation results in the low noise regime (emphasized in the right panel of Figure 3), which is due to the following facts: 1) we stop Algorithm 3 after a limited number of iterations, $t_{\max}$, and 2) the network size is small. The threshold predicted by Theorem 3 becomes accurate as $n \to \infty$.

7.2. Spoken Words Dataset

Encouraged by the results of (Kohonen et al., 1980) which suggests that phonemes of natural languages

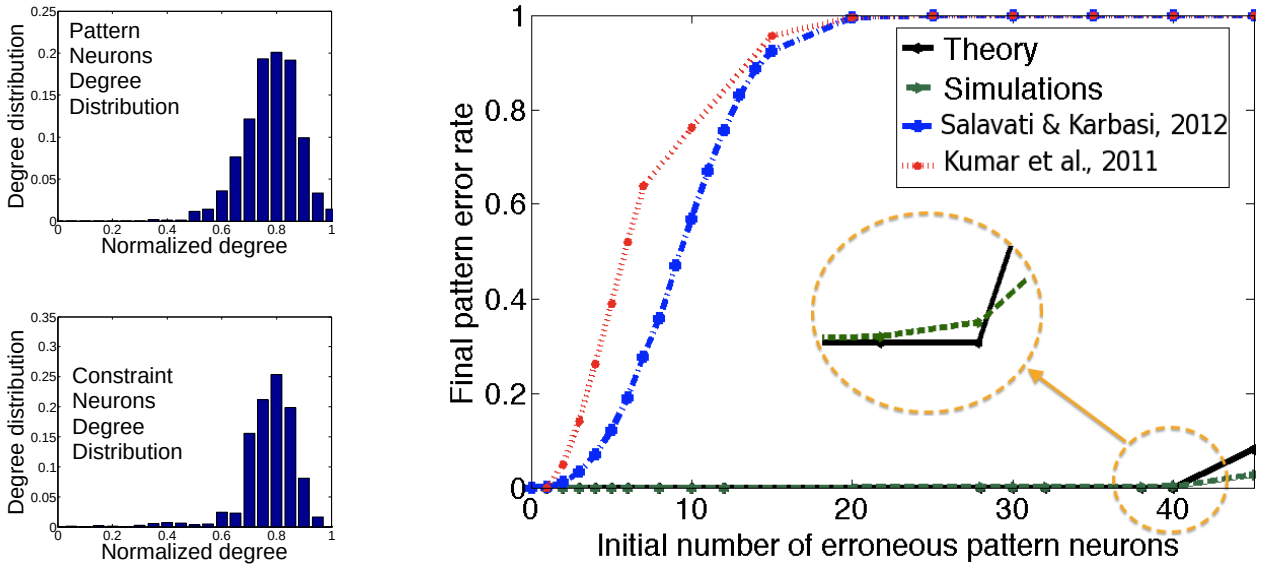

Figure 3. Pattern and constraint neuron degree distributions for $n = 400$, $L = 50$, and on average of 20 constraints per cluster. The learning parameters are $\alpha_t \propto 0.95/t$, $\beta = 0.75/\alpha_t$ and $\theta_t \propto 0.05/t$.

form a subspace in the spectral domain, we have applied our proposed method to the *Noun Pool* dataset of 482 spoken English words[2]. We first transform the words into the spectral domain (by applying Fourier transform), and then digitize the spectral signals. This is done by dividing the frequency domain into n bands and quantize the value of the spectral signal using an S-level non-uniform quantizer. As a result, we will have a signal with length n and elements varying from 0 to $S-1$. We divide the digitized signals into random clusters and apply Algorithm 1 to learn the dual weight matrix of each cluster.

During the recall phase, we add a random noise to the *digitized spectral signal* and investigate the performance of our algorithm with respect to the theoretical analysis. The noise is added in the spectral domain rather than the time domain for the ease of theoretical verifications and does not serve any other purpose.

Figure 4 illustrates the results. Firstly, Theorem 3 provides a nice upper bound on the error rate of the algorithm. Secondly, as the gap suggests our assumption that a cluster can correct a single error is too conservative; they can correct multiple errors.

8. Future Works

We believe that the same method could easily be extend to any other natural datasets with weak minor components in their correlation matrix, as was the case here, since these datasets could *approximately* be mapped to a subspace which could be then learned by our algorithm. A good case in point is the dataset of natural images belonging to a particular class, which is in fact part of our ongoing research.

[2] http://memory.psych.upenn.edu/WordPools

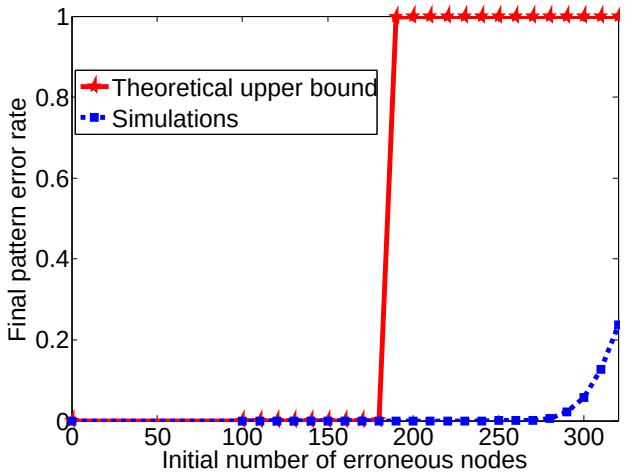

Figure 4. Pattern error rates for the speech dataset with $n = 1200$, $L = 150$ and $S = 40$. Simulations parameters are $\alpha_t \propto 0.95/t$, $\beta = 1/\alpha_t$ and $\theta_t \propto 0.01/t$.

References

Bottou, L. *Online algorithms and stochastic approximations.* in David Saad, editor, *Online Learning and Neural Networks.* Cambridge University Press, 1998.

Candes, E., Tao, T. *Near optimal signal recovery from random projections: Universal encoding strategies?.* IEEE Trans. Inf. Theory, Vol. 52, No. 12, 2006, pp. 5406 -

5425.

Donoho, D. L. *Compressed Sensing.* IEEE Transactions on Information Theory.

Donoho, D. L., Maleki, A., Montanari, A. *Message passing algorithms for compressed sensing.* PNAS, 106, 18914-18919.

Gripon, V., Berrou, C. *Sparse neural networks with large learning diversity.* IEEE Trans. Neur. Net., 22 (7), 10871096.

Hopfield, J. J. *Neural networks and physical systems with emergent collective computational abilities.* PNAS, 79, 2554-2558.

Jankowski S., Lozowski, A., Zurada, J. M. *Complex-valued multistate neural associative memory.* IEEE Tran. Neur. Net., 1 (6), 1491-1496.

Jarrett, K., Kavukcuoglu, K., Ranzato, M. A., LeCun, Y. *What is the best multi-stage architecture for object recognition?* Proc. ICCV 2009, 2146-2153.

Kohonen, T., Riittinen, M. J., Reuhkala, E., Haltsonen, S., *A 1000 word recognition system based on the learning subspace method and redundant hash addressing.* Proc. Pattern Recognition 1980, 168-166.

Kumar, K. R., Salavati, A. H., Shokrollahi, A. *Exponential pattern retrieval capacity with non-binary associative memory.* Proc. ITW 2011, 80-84.

Le, Q. V., Ngiam, J., Chen, Z., Chia, D., Koh, P. W., Ng, A. Y. *Tiled convolutional neural.* Proc. NIPS 2010, 1279-1287.

Luby, M. G., Mitzenmacher, M., Shokrollahi, M. A., Spielman, D. A. *Efficient erasure correcting codes.* IEEE Trans. Inf. Theory, 47 (2), 569-584.

McEliece, R., Posner, E., Rodemich, E., Venkatesh, S. *The capacity of the Hopfield associative memory.* IEEE Trans. Inf. Theory, 33 (4), 461-482.

Muezzinoglu, M. K., Guzelis, C., Zurada, J. M. *A new design method for the complex-valued multistate Hopfield associative memory.* IEEE Trans. Neur. Net., 14 (4), 891-899.

Oja, E., Karhunen, J. *On stochastic approximation of eigenvectors and eigenvalues of the expectation of a random matrix.* J. Math. Anal. & App., 106 (1), 69-84.

Oja, E., Kohonen, T. *The subspace learning algorithm as a formalism for pattern recognition and neural networks.* Proc. Neur. Net., 1, 277-284.

Peretto, P., Niez, J. J. *Long term memory storage capacity of multiconnected neural networks.* Bio. Cyb., 54 (1), 53-63.

Richardson, T., Urbanke, R. *Modern coding theory.* Cambridge University Press, 2008.

Salavati, A. H., Karbasi, A. *Multi-Level error-resilient neural networks.* Proc. ISIT 2012, 1064-1068.

Socher, R., Pennington, J., Huang, E. H., Ng, A. Y., Manning, C. D. *Semi-supervised recursive autoencoders for predicting sentiment distributions.* Proc. EMNLP 2011, 151-161.

R. Tanner. *A recursive approach to low complexity codes* . IEEE Transactions on Information Theory.

Tropp, J., Wright, S. J. *Computational methods for sparse solution of linear inverse problems.* Proc. IEEE, Vol. 98, No. 6, 2010, pp. 948-958.

Venkatesh, S. S., Psaltis, D. *Linear and logarithmic capacities in associative neural networks.* IEEE Trans. Inf. Theory, 35 (3), 558-568.

Vincent, P., Larochelle, H., Bengio, Y., Manzagol, P. A., Koh, P. W., Ng, A. Y. *Extracting and composing robust features with denoising autoencoders.* Proc. ICML 2008, 1096-1103.

Xu, L., Krzyzak, A., Oja, E. *Neural nets for dual subspace pattern recognition method.* J. Neur. Syst., 2 (3), 169-184.

Proof of Theorem 1

To prove the theorem, we use the convergence results in (Bottou, 1998) and show that the required assumptions to ensure convergence holds for the proposed algorithm. For simplicity, these assumptions are listed here:

1. The cost function $E(w^{(\ell)})$ is three-times differentiable with continuous derivatives. It is also bounded from below.

2. The usual conditions on the learning rates are fulfilled, i.e. $\sum \alpha_t = \infty$ and $\sum \alpha_t^2 < \infty$.

3. The second moment of the update term should not grow more than linearly with size of the weight vector. In other words,

$$E(w^{(\ell)}) \leq a + b\|w^{(\ell)}\|_2^2$$

for some constants a and b.

4. When the norm of the weight vector $w^{(\ell)}$ is larger than a certain horizon D, the opposite of the gradient $-\nabla E(w^{(\ell)})$ points towards the origin. Or in other words:

$$\inf \|w^{(\ell)}\|_2 > Dw \cdot \nabla E(w^{(\ell)}) > 0$$

5. When the norm of the weight vector is smaller than a second horizon F, with $F > D$, then the norm of the update term $\left(2y^{(\ell)}(t)x^{(\ell)}(t) + \beta\Gamma(w^{(\ell)}(t))\right)$ is bounded regardless of $x^{(\ell)}(t)$, where $x^{(\ell)}(t)$ is the subpattern of pattern $x(t) \in \mathcal{X}$ corresponding to cluster ℓ. This is usually a mild requirement, where for all patterns $x(t)$ in the dataset $\mathcal{X}$ we should have,

$$\sup_{\|w^{(\ell)}\|_2 \leq F} \| \left(2y^{(\ell)}(t)x^{(\ell)}(t) + \beta\Gamma(w^{(\ell)}(t))\right) \|_2 \leq K_0$$

Also recall that $A_x = x^{(\ell)}(x^{(\ell)})^\top$, and $A = \mathbb{E}\{A_x | x \in \mathcal{X}\}$ represent the correlation among patterns in the training set $\mathcal{X}$, so $E(w^{(\ell)}) = \sum_{x\in\mathcal{X}} |x^{(\ell)} \cdot w^{(\ell)}|^2 = (w^{(\ell)})^\top A w^{(\ell)}/\mathcal{C}$.

To start, assumption 1 holds trivially as the cost function is three-times differentiable, with continuous derivatives. Furthermore, $E(w^{(\ell)}) \geq 0$. Assumption 2 holds because of our choice of the step size α_t, as mentioned in the lemma description.

Assumption 3 ensures that the vector $w^{(\ell)}$ could not escape by becoming larger and larger. Due to the constraint $\|w^{(\ell)}\|_2 = 1$, this assumption holds as well.

Assumption 4 holds as well because:

$$\begin{aligned}
\mathbb{E}_x \left(2A_x w^{(\ell)} + \beta\Gamma(w^{(\ell)})\right)^2 &= 4(w^{(\ell)})^\top \mathbb{E}_x(A_x^2) w^{(\ell)} \\
&+ \beta^2 \|\Gamma(w^{(\ell)})\|_2^2 \\
&+ 4\beta (w^{(\ell)})^\top \mathbb{E}_x(A_x)\Gamma(w^{(\ell)}) \\
&\leq 4\|w^{(\ell)}\|_2^2 \zeta^2 + \beta^2 \|w^{(\ell)}\|_2^2 \\
&+ 4\beta\Upsilon \|w^{(\ell)}\|_2^2 \\
&= \|w^{(\ell)}\|_2^2 (4\zeta^2 + 4\beta\Upsilon + \beta^2)
\end{aligned}$$

Finally, assumption 5 holds because:

$$\begin{aligned}
\|2A_x w^{(\ell)} + \beta\Gamma(w^{(\ell)})\|_2^2 &= 4(w^{(\ell)})^\top A_x^2 w^{(\ell)} + \beta^2 \|\Gamma(w^{(\ell)})\|_2^2 \\
&+ 4\beta (w^{(\ell)})^\top A_x \Gamma(w^{(\ell)}) \\
&\leq \|w^{(\ell)}\|_2^2 (4\zeta^2 + 4\beta\zeta + \beta^2) \qquad (9)
\end{aligned}$$

Therefore, $\exists F > D$ such that as long as $\|w^{(\ell)}\|_2^2 < F$:

$$\sup_{\|w^{(\ell)}\|_2^2 < E} \|2A_x w^{(\ell)} + \beta\Gamma(w^{(\ell)})\|_2^2 \leq (2\zeta + \beta)^2 F = \text{constant} \qquad (10)$$

Since all necessary assumptions hold for the learning algorithm 1, it converges to a local minimum where $\nabla E(\hat{w}^{(\ell)}) = 0$.

Next, we prove the desired result, i.e. the fact that in the local minimum, the resulting weight vector is orthogonal to the patterns, i.e. $Aw^{(\ell)} = 0$. Since $\nabla E(\hat{w}^{(\ell)}) = 2A\hat{w}^{(\ell)} + \beta\Gamma(\hat{w}^{(\ell)}) = 0$, we have:

$$\hat{w}^{(\ell)} \cdot \nabla E(\hat{w}^{(\ell)}) = 2(\hat{w}^{(\ell)})^\top A \hat{w}^{(\ell)} + \beta \hat{w}^{(\ell)} \cdot \Gamma(\hat{w}^{(\ell)}) \quad (11)$$

The first term is always greater than or equal to zero. Now as for the second term, we have that $|\Gamma(w_i^{(\ell)})| \leq |w_i^{(\ell)}|$ and $\text{sign}(w_i^{(\ell)}) = \text{sign}(\Gamma(w_i^{(\ell)}))$, where $w_i^{(\ell)}$ is the i^{th} entry of $w^{(\ell)}$. Therefore, $0 \leq \hat{w}^{(\ell)} \cdot \Gamma(\hat{w}^{(\ell)}) \leq \|\hat{w}^{(\ell)}\|_2^2$. Therefore, both terms on the right hand side of (11) are greater than or equal to zero. And since the left hand side is known to be equal to zero, we conclude that $(\hat{w}^{(\ell)})^T A \hat{w}^{(\ell)} = 0$ and $\Gamma(\hat{w}^{(\ell)}) = 0$. The former means $(\hat{w}^{(\ell)})^\top A \hat{w}^{(\ell)} = \sum_{x\in\mathcal{X}} (\hat{w}^{(\ell)} \cdot x^{(\ell)})^2 = 0$. Therefore, we must have $\hat{w}^{(\ell)} \cdot x = 0$, for all $x \in \mathcal{X}$. Which simply means the vector w^* is orthogonal to all the patterns in the training set.

Although in problem (2) we have the constraint $\|w^{(\ell)}\|_2 = 1$ to make sure that the algorithm does not converge to the trivial solution $w^{(\ell)} = 0$, due to approximations we made when developing the optimization algorithm, we should make sure to choose the parameters such that the all-zero solution is still avoided.

To this end, denote $w'^{(\ell)}(t) = w^{(\ell)}(t) - \alpha_t y^{(\ell)}(t) \left(x^{(\ell)}(t) - \frac{y^{(\ell)}(t) w^{(\ell)}(t)}{\|w^{(\ell)}(t)\|_2^2}\right)$ and consider the

following inequalities:

$$\begin{aligned}\|w^{(\ell)}(t+1)\|_2^2 &= \|w^{(\ell)}(t) - \alpha_t y^{(\ell)}(t)(x^{(\ell)}(t) \\ &- \frac{y^{(\ell)}(t)w^{(\ell)}(t)}{\|w^{(\ell)}(t)\|_2^2}) - \alpha_t\beta\Gamma(w^{(\ell)}(t))\|_2^2 \\ &= \|w^{'(\ell)}(t)\|^2 + \alpha_t^2\beta^2\|\Gamma(w^{(\ell)}(t))\|^2 \\ &- 2\alpha_t\beta\Gamma(w^{(\ell)}(t))\cdot w^{'(\ell)}(t) \\ &\geq \|w^{'(\ell)}(t)\|_2^2 - 2\alpha_t\beta\Gamma(w^{(\ell)}(t))\cdot w^{'(\ell)}(t)\end{aligned}$$

Now in order to have $\|w^{(\ell)}(t+1)\|_2^2 > 0$, we must have that $2\alpha_t\beta|\Gamma(w^{(\ell)}(t))^\top w^{'(\ell)}(t)| \leq \|w^{'(\ell)}(t)\|_2^2$. Given that, $|\Gamma(w^{(\ell)}(t))\cdot w^{'(\ell)}(t)| \leq \|w^{'(\ell)}(t)\|_2\|\Gamma(w^{(\ell)}(t))\|_2$, it is therefore sufficient to have $2\alpha_t\beta\|\Gamma(w^{(\ell)}(t))\|_2 \leq \|w^{'(\ell)}(t)\|_2$. On the other hand, we have:

$$\begin{aligned}\|w^{'(\ell)}(t)\|_2^2 &= \|w^{(\ell)}(t)\|_2^2 + \alpha_t^2 y^{(\ell)}(t)^2\|x^{(\ell)}(t) \\ &- \frac{y^{(\ell)}(t)w^{(\ell)}(t)}{\|w^{(\ell)}(t)\|_2^2}\|_2^2 \\ &\geq \|w^{(\ell)}(t)\|_2^2 \qquad (12)\end{aligned}$$

As a result, in order to have $\|w^{(\ell)}(t+1)\|_2^2 > 0$, it is sufficient to have $2\alpha_t\beta\|\Gamma(w^{(\ell)}(t))\|_2 \leq \|w^{(\ell)}(t)\|_2$. Finally, since we have $|\Gamma(w^{(\ell)}(t))| \leq |w^{(\ell)}(t)|$ (entry-wise), we know that $\|\Gamma(w^{(\ell)}(t))\|_2 \leq \|w^{(\ell)}(t)\|_2$. Therefore, having $2\alpha_t\beta < 1 \leq \|w^{(\ell)}(t)\|_2/\|\Gamma(w^{(\ell)}(t))\|_2$ ensures $\|w^{(\ell)}(t)\|_2 > 0$.

Proof of Theorem 2

In the case of a single error, we are sure that the corrupted node will always be updated towards the correct direction. For simplicity, let's assume the first pattern neuron of cluster ℓ is the noisy one. Furthermore, let $z = \{1, \ldots, 0\}$ be the noise vector. Denoting the i^{th} column of the weight matrix by $W_i^{(\ell)}$, we will have $y^{(\ell)} = \text{sign}(W_1^{(\ell)})$. Then in algorithm 2 $g_1 = 1 > \varphi$. This means that the noisy node gets updated towards the correct direction.

Therefore, the only source of error would be a correct node gets updated mistakenly. Let P_{x_i} denote the probability that a correct pattern neuron x_i gets updated. This happens if $|g_{x_i}| > \varphi$. For $\varphi = 1$, this is equivalent to having $W_i^{(\ell)}\cdot\text{sign}(z_1 W_1^{(\ell)}) = \|W_i^{(\ell)}\|_0$. Note that $W_i^{(\ell)}\cdot\text{sign}(W_1^{(\ell)}) < \|W_i^{(\ell)}\|_0$ in cases that the neighborhood of x_i is different from the neighborhood of x_1 among the constraint nodes. More specifically, in the case that $\mathcal{N}(x_i)\cap\mathcal{N}(x_1) \neq \mathcal{N}(x_i)$, there are non-zero entries in $W_i^{(\ell)}$ while $W_1^{(\ell)}$ is zero and vice-versa. Therefore, letting P'_{x_i} being the probability of $\mathcal{N}(x_i)\cap\mathcal{N}(x_1) \neq \mathcal{N}(x_i)$, we note that

$$P_{x_i} \leq P'_{x_i}$$

Therefore, to get an upper bound on P_{x_i}, we bound P'_{x_i}.

Let $\Lambda_i^{(l)}$ be the fraction of pattern neurons with degree i in cluster l, $d_{\text{avg}}^{(l)} = \sum_i i\Lambda_i^{(l)}$ be the average degree of pattern neurons and finally $d_{\text{min}}^{(l)}$ be the minimum degree of pattern neurons in cluster l. Then, we know that a noisy pattern neuron is connected to $d_{\text{avg}}^{(l)}$ constraint neurons on average. Therefore, the probability of x_i and x_1 share exactly the same neighborhood would be:

$$P'_{x_i} = \left(\frac{d_{\text{avg}}^{(l)}}{m}\right)^{d_{x_i}} \qquad (13)$$

Taking the average over the pattern neurons, we have

$$\begin{aligned}P'_e &= \Pr\{x \in C_t\}\mathbb{E}_{d_{x_i}}\{P'_{x_i}\} \\ &= (1 - \frac{1}{n_l})\Lambda(\frac{d_{\text{avg}}^{(l)}}{m}) \\ &= \Lambda^{(l)}(\frac{d_{\text{avg}}^{(l)}}{m}) \qquad (14)\end{aligned}$$

where C_t is the set of correct nodes at iteration t and $\Lambda^{(l)}(x) = \sum_i \Lambda_i^{(l)} x^i$.

Therefore, the probability of correcting one noisy input, $P_c = 1 - P_e \geq 1 - P'_e$ would be

$$\begin{aligned}P_c &\geq 1 - \Lambda^{(l)}(\frac{d_{\text{avg}}^{(l)}}{m}) \\ &\geq 1 - \left(\frac{d_{\text{avg}}^{(l)}}{m}\right)^{d_{\text{min}}^{(l)}} \qquad (15)\end{aligned}$$

Proof of Theorem 3

The proof is similar to Theorem 3.50 in (Richardson & Urbanke, 2008). Each cluster node receives an error message from its neighboring pattern nodes with probability z. Now consider a given *noisy* pattern neuron which is connected to a given cluster $v^{(\ell)}$. Let $\pi^{(\ell)}(t)$ be the probability that the cluster node $v^{(\ell)}$ with degree $\widetilde{d}_\ell$ sends an error message during iteration t of Algorithm 3. This event happens if the cluster node $v^{(\ell)}$ receives at least one error message from its other neighbors among pattern neurons along its input edges, i.e. if it is connected to more than one noisy pattern neuron. Therefore,

$$\pi^{(\ell)}(t) = 1 - (1 - z(t))^{\widetilde{d}_\ell - 1} \qquad (16)$$

As a result, if $\pi(t)$ shows the average probability that a cluster node sends a message declaring the violation

of at least one of its constraint neurons, we will have,

$$\pi(t) = \mathbb{E}_{\widetilde{d}_\ell}\{\pi^{(\ell)}(t)\} = \sum_i \widetilde{\rho}_i(1-(1-z(t))^{\widetilde{d}_\ell - 1}) = 1-\widetilde{\rho}(1-z(t)) \tag{17}$$

Now consider a given pattern neuron x_i with degree d_i. This node will remain noisy in iteration $t+1$ of Algorithm 3 if it was noisy in the first place and in iteration $t+1$ all of its neighbors among constraint neurons send a violation message. Therefore, the probability of this node being noisy will be $z(0)\pi(t)^{d_i}$. As a result, noting that $z(0) = p_e$, the average probability that a pattern neurons remains noisy will be

$$z(t+1) = p_e \cdot \sum_i \widetilde{\lambda}_i \pi(t)^i = p_e \cdot \widetilde{\lambda}(\pi(t)) = p_e \cdot \widetilde{\lambda}(1-\widetilde{\rho}(1-z(t))) \tag{18}$$

Therefore, the decoding operation will be successful if $z(t+1) < z(t),\ \forall t$. As a result, we must look for the maximum p_e such that we will have $p_e \cdot \widetilde{\lambda}(1-\widetilde{\rho}(1-z)) < z$ for $z \in [0, p_e]$.

Proof of Theorem 4

The proof is based on construction: we construct a data set $\mathcal{X}$ with the required properties such that it can be memorized by the proposed neural network.

To start, consider a matrix $G \in \mathbb{R}^{k \times n}$ with rank k and $k = rn$, with $0 < r < 1$. Let the entries of G be non-negative integers, between 0 and $\gamma - 1$, with $\gamma \geq 2$.

We start constructing the patterns in the data set as follows: consider a random vector $u \in \mathbb{R}^k$ with integer-valued-entries between 0 and $\upsilon - 1$, where $\upsilon \geq 2$. We set the pattern $x \in \mathcal{X}$ to be $x = u \cdot G$, *if* all the entries of x are between 0 and $S - 1$. Obviously, since both u and G have only non-negative entries, all entries in x are non-negative. Therefore, it is the $S - 1$ upper bound that we have to worry about.

The j^{th} entry in x is equal to $x_j = u \cdot g_j$, where g_j is the j^{th} column of G. Suppose g_j has d_j non-zero elements. Then, we have:

$$x_j = u \cdot g_j \leq d_j(\gamma - 1)(\upsilon - 1)$$

Therefore, denoting $d^* = \max_j d_j$, we could choose γ, υ and d^* such that

$$S - 1 \geq d^*(\gamma - 1)(\upsilon - 1) \tag{19}$$

to ensure all entries of x are less than S.

As a result, since there are υ^k vectors u with integer entries between 0 and $\upsilon - 1$, we will have $\upsilon^k = \upsilon^{rn}$ patterns forming $\mathcal{X}$. Which means $\mathcal{C} = \upsilon^{rn}$, which would be an exponential number in n if $\upsilon \geq 2$.

As an example, if G is selected to be a sparse 200×400 matrix with 0/1 entries (i.e. $\gamma = 2$) and $d^* = 10$, and u is also chosen to be a vector with 0/1 elements (i.e. $\upsilon = 2$), then it is sufficient to have $S \geq 11$, i.e. the maximum firing rate of neurons should be 11 to have a pattern retrieval capacity of $\mathcal{C} = 2^{rn}$.

Remark 1. *Note that the inequality (19) was obtained for the worst-case scenario and in fact is very loose. Therefore, even if it does not hold, we will still be able to memorize a very large number of patterns since a big portion of the generated vectors x will have entries less than S. These vectors correspond to the message vectors u that are "sparse" as well, i.e. do not have all entries greater than zero. The number of such vectors is a polynomial in n, the degree of which depends on the number of non-zero entries in u.*