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 (**??**) 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 $d_{\min}^{(l)}$ 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

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

Proof of Theorem 3

The proof is similar to Theorem 3.50 in (**?**). 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,

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

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 (**??**) 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.*