# A Simple Algorithm for Topographic ICA

Ewaldo Santana[1], Allan Kardec Barros[1], Christian Jutten[2], Eder Santana[1] and Luis Claudio Oliveira[1]
[1]*Federal University of Maranhao, Dept. Electrical Engineering, Sao Luis, Maranhao, Brazil*
[2]*GIPSA-lab, Department Images et Signal, Saint Martin d'Heres Cedex, France*

Keywords: Independent Component Analysis (ICA), Topographic ICA, Receptive Fields.

Abstract: A number of algorithms have been proposed which find structures that resembles that of the visual cortex. However, most of the works require sophisticated computations and lack a rule for how the structure arises. This work presents an unsupervised model for finding topographic organization with a very easy and local learning algorithm. Using a simple rule in the algorithm, we can anticipate which kind of structure will result. When applied to natural images, this model yields an efficient code for natural images and the emergence of simple-cell-like receptive fields. Moreover, we conclude that the local interactions in spatially distributed systems and local optimization with norm L2 are sufficient to create sparse basis, which normally requires higher order statistics.

## 1 INTRODUCTION

Perhaps the most productive set of self-organizing principles are information theoretic in nature. Shannon in his seminal work showed how to optimally design communication systems by manipulating the two fundamental descriptors of information: entropy and mutual information Shannon (1948). Inspired by this achievement, Linsker proposed the "infomax" principle, which shows that maximization of mutual information for Gaussian probability density functions comes to correlation minimization Linsker (1992). Using correlation as the basis of learning, some researchers have developed several algorithms to mimic the process of learning Oja (1989); Ray et al. (2013). Bell and Sejnowski applied this same principle for independent component analysis (ICA) with a very clever local learning rule taking advantage of the statistical properties of the input data Bell and Sejnowski (1995). Barlow hypothesized that the role of the visual system is exactly to minimize the redundancy in real world scenes Barlow (1989).

Finding the maximum entropy distribution, one of the principles of redundancy reduction when performed with the constraint of a fixed mean, provides probability density functions with sharp peaks at the mean and heavy tails, i.e. sparse distributions Simoncelli and Olshausen (2001). Thus, assuming that the goal of the primary visual (V1) cortex of mammals is to obtain a sparse representation of the natural scenes, Olshausen and Field derived a neural network that self-organizes into oriented, localized and bandpass filters reminiscent of receptive fields in V1 Olshausen and Field (1996). Exploiting V1, Vinje and Gallant Vinje and Gallant (2000) experimentally proved the sparse representations in natural vision. In modeling natural scenes, Hyvarinen et al Hyvarinen and Hoyer (2000) defined an Independent Component Analysis (ICA) generative model in which the components are not completely independent but have a dependency that is defined in relation to some topography. Components close to one another in this topography have greater co-dependence than those that are further away. Using a similar rationale, Osindero et al Osindero (2006) presented a hierarchical energy-based density model that is applicable to data-sets that have a sparse structure, or that can be well characterized by constraints that are often well-satisfied, but occasionally violated by a large amount. Moreover, there have been a number of studies on sparse coding and ICA that match rather well the receptive fields of the visual cortex, and also a number of algorithms which finds excellent organization that resembles that of the visual cortex, i.e., Olshausen and Field (1996).

One of the difficulties faced by all these studies is the sophisticated computation required to obtain sparse representations and the generative models which are generally derived with several constraints to conform with mathematical analysis. Analyzing sensory coding in a supervised framework, Barros et al Barros and Ohnishi (2002) proposed the minimization of a nonlinear version of the local discrepancy between the sensory input and a neuron internal reference signal which also yields wavelet like recep-



Santana, E., Barros, A., Jutten, C., Santana, E. and Oliveira, L.
A Simple Algorithm for Topographic ICA.
DOI: 10.5220/0005683501500155
In *Proceedings of the 9th International Joint Conference on Biomedical Engineering Systems and Technologies (BIOSTEC 2016) - Volume 4: BIOSIGNALS*, pages 150-155
ISBN: 978-989-758-170-0

tive fields when applied to natural images. In this work we propose an unsupervised model for finding topographic organization using convex nonlinearities with a very easy and local learning algorithm. Our development is based on the fact that most nonlinearities yield higher order statistical properties in modeling. In our context, we model the neuron as a system which performs non-linear stimuli filtering Barros and Ohnishi (2002), and use its output as input signals from neighboring neurons. This arrangement gives spatially coherent responses. This paper is organized as follows: In Section 2, we present the model and develop methods for obtaining the algorithm. Section 3 shows simulations and results. Discussions and conclusions are proposed in section 4.

## 2 METHODS

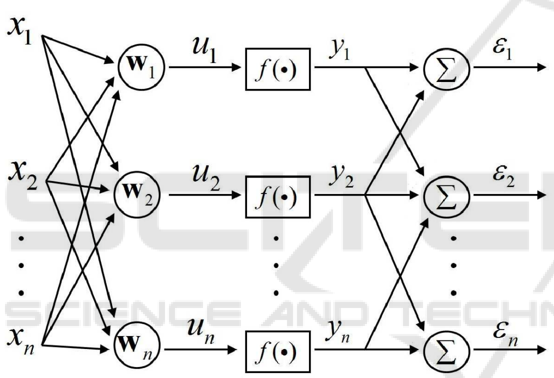



Figure 1: Proposed model, which works in *parallel* fashion. First layer: input vector **x**; second layer: the input vector are locally linearly combined by synaptic weights vector **w** yielding the signal **u**; Third layer: the signal **u** pass through a non-linearity, f(.) to get higher order characteristics of input signal; Fourth layer: the local discrepancy.

Input signals, $\mathbf{x}_k = [x_1(k), \ldots, x_n(k)]$, are assumed generated by a linear combination of real world source signals, $\mathbf{s}(k) = [s_1(k), \ldots, s_n(k)]^T$, $\mathbf{x} = \mathbf{A}\mathbf{s}$, where $\mathbf{A}$ is an unknown basis functions set.

In order to facilitate further development, we also assume that the input signals are preprocessed by whitening Hyvarinen and Hoyer (2000). In the demixing model each signal, $x_j$, at the input of synapse $j$ connected to neuron $k$ will be multiplied by the synaptic weight $w_{kj}$ and summed to produce the linear combined output signal of neuron $k$, which is an estimation of $s_k$,

$$u_k = \sum_{j=1}^{n} w_{kj} x_j, y_k = f(u_k), \tag{1}$$

where $f$ is an even symmetric non-linear function, which is chosen to exploit the higher order statistical information from the input signal Hyvarinen and Hoyer (2000). This signal, $y_k$, interacts with neighboring outputs, since in the nervous system the response of a neuron tends to be correlated with the response of its neighboring area Durbin and Mitchison (2007). So, $y_k$ is used to generate a signal to adapt $w_{kj}$ and make its output coherent with the response of its neighbors.

We define the neighborhood of neuron $k$, all neurons $i$ that are close to neuron $k$ in a matrix representation, and define the action of a neuron's neighborhood as a weighted sum of output signals from all neurons in the neighborhood,

$$v_k = \sum_i y_{ik} = \sum_i f(u_{ik}), \tag{2}$$

where $i$ are the indices of the neighboring outputs $k$ and $v_k$ interacts with $y_k$ to generate a self reference signal, $\varepsilon_k = y_k + \alpha v_k$, where $\alpha$ determines the neighborhood influence on neuron's activation. This self reference signal must be minimized in order to reduce the local redundancy. Thus, we define the following local cost,

$$\mathbf{J}_k = E[\varepsilon_k^2] = E[(y_k + \alpha v_k)^2], \tag{3}$$

which should be minimized with respect to all weights of neuron $k$. Taking the partial derivative of Eq.(3) with respect to the parameters and equating it to zero we obtain the following equation:

$$2E\left[f'(u_k)(y_k + \alpha v_k)\mathbf{x}\right] = 0, \tag{4}$$

where we used the fact that $f'(u_{ik})$, with respect to $\mathbf{w}_k$, is zero for $i \neq k$.

In order to avoid the trivial solution of zero weight, we utilize the Lagrange multipliers to minimize $\mathbf{J}_k$ under constraint $\|\mathbf{w}_k\| = 1$ and observe that the derivative of $\sum_i f'(u_{ik})$ with respect to $\mathbf{w}_k$ is zero. With this, we obtain the following two step algorithm,

$$\begin{aligned} \mathbf{w}_k &= E\left[f'(u_k)\varepsilon_k \mathbf{x}\right] \\ \mathbf{w}_k &= \frac{\mathbf{w}_k}{\|\mathbf{w}_k\|}. \end{aligned} \tag{5}$$

This procedure is accomplished in parallel to adjust all weights for each neuron in the network, where we include orthogonalization steps taken from Foldiák (1990). In this way a matrix $\mathbf{W}$ is obtained.

Moreover, we can show that the proposed algorithm converges as given in the theorem below.

**Theorem:** Assume that the observed stimuli follow the ICA model and that the whitened version of those signals are **x=VAs**, where **A** is an invertible matrix and **V** is an orthogonal matrix, and **s** is the source vector. Suppose that $\mathbf{J}(\mathbf{w}_k) = E\{[f(\mathbf{w}_k^T\mathbf{v}) + \alpha.v_k]^2\}$ is an appropriate smoothing even function. Moreover, suppose that any source signal $s_k$ is statistically independent from $s_j$, for all $i \neq j$. This way, the local minimum of $\mathbf{J}(\mathbf{w}_k)$, under restriction $\|\mathbf{w}_k\| = 1$ describes one line of the matrix $(\mathbf{VA})^{-1}$ and the correspondent separated signal obeys

$$\begin{aligned} & E[s_i^2(f'(s_k))^2 + s_i^2 f(s_k) f''(s_k) + \\ + \quad & s_i f(s_k) f'(s_k) - s_k f(s_k) f'(s_k)] + \\ & + E[(f'(s_k))^2 + f(s_k) f''(s_k) - \\ - \quad & s_k f(s_k) f'(s_k)] < 0, \forall i \in v_k, \end{aligned} \tag{6}$$

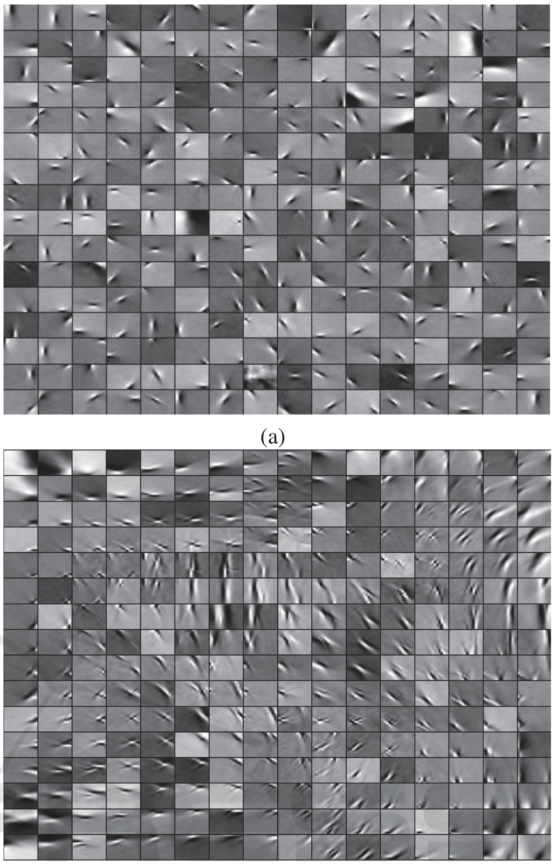
We show the proof in the appendix.

## 3 SIMULATIONS AND RESULTS

We performed simulations with the closest eight neurons neighbors, as defined in Appendix A. We chose the following even symmetric non-linearity: $f(u_k) = u_k.atan(0.5u_k) - ln(1+0.25u_k^2))$. This function has some advantages in terms of information transmission between input and outputButko and Triesch (2007). Moreover, this function will work on all statistics of the signal, yielding therefore a better approximation to the probability density function (pdf) of it Bell and Sejnowski (1995). This can be intuitively understood by expanding the sigmoidal function into a Taylor series, whereas one can easily notice that $f(u_k)$ can be written as a linear combination of $u_k^2$. That function also satisfies the requirement of a unique minimum as shown by Gersho Gersho (1969).

We applied the algorithm given by Equation (5) to encode natural images, in a way similar to that carried out in the literature Olshausen and Field (1996). We take random samples of 16 x 16 pixel image patches from a set of 13 wild life pictures,(http://www.naturalimagestatistics.net/) and put each sample into a column vector of dimension 256x1. We used 200,000 whitened versions of those samples as stimuli. For each set of 200,000 points the algorithm runs over 20,000 times in order to obtain a good estimation for matrix **A**. In this case we have also 16 x 16 units in the network. We varied $\alpha$ from 0.005 to 1.0 to highlight the influence of the neighborhood in the results.

In Fig.2, one can see the obtained basis vectors (columns of matrix **A**) for $\alpha$ = 0.1 and 0.9, respectively. The clustered topographical organization can be easily seen taking one little square (receptive field) and considering its eight closest neighbors.



(a)

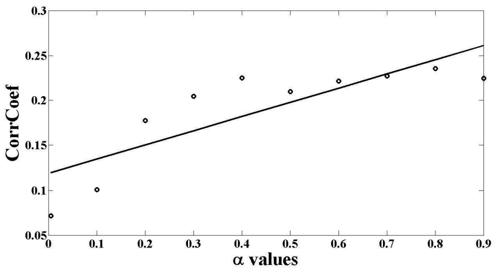
(b)

Figure 2: The topographic basis functions (columns of matrix **A**) estimated from natural image data when $f(u_k) = u_k \cdot atan(0.5u_k) - Ln(1+0.25u_k^2)$ was the non-linearity utilized and neighborhood consisting of the 8 closest neighbors.. (a) $\alpha = 0.005$; (b) $\alpha = 0.9$.

In these figures one can observe that basis vectors which are similar in location, orientation and frequency are close to each other. Moreover, we see the importance of neighbours contribution on organization: increasing $\alpha$, the filters became more correlated. In Figure 3 , one can see the $\alpha$ values and corresponding mean of correlation coefficients of all neurons in neighborhood.

## 4 DISCUSSIONS AND CONCLUSIONS

In this paper we have presented a model for unsupervised neural network adaptation. In (5) we have a very simple unsupervised rule to update the weights in a neural network. In topographic organization literature, as far as we know, this is the simplest ever




(a)

Figure 3: Mean of correlation coefficient between pair of basis functions as a function of α values.

suggested, when compared to others in the literature-Hyvarinen et al. (2000); Hyvarinen and Hoyer (2000); Osindero (2006). Besides, we have proven mathematically that the adaptation converges. One possible application of this method is to image/signal recognition, as for each $\alpha$ in (3), we can adjust the characteristics of the topological structure in order to fit more closely to the desired image.

There are at least two advantages in this model: it organizes the resulting filters topographically and it may be regarded as biologically plausible. Firstly, we can see in Fig. 2 (a) that a small $\alpha$ results in no organization at all, while in Fig. 2 (b), we can see that an $\alpha = 0.9$ causes the filters to appear organized. Moreover, we have found the correlation in neighborhood filters. By using equation $Cov(\mathbf{a}_i, \mathbf{a}_j)/(std(\mathbf{a}_i)std(\mathbf{a}_))$, where $\mathbf{a}_i$ and $\mathbf{a}_j$ are two basis functions in the neighborhood of one neuron, we can see that the correlation is directly proportional to the value of $\alpha$, as shown in Fig. 3. The correlation between $\alpha$ and $Cov(\mathbf{a}_i, \mathbf{a}_j)/(std(\mathbf{a}_i)std(\mathbf{a}_))$ was 0.88.

Secondly, it is well known that receptive fields in the mammalian visual cortex resemble Gabor filters Laughlin (1981), Linsker (1992). This way, the information about the visual world would excite different cells Laughlin (1981). In Figure 2 (a) and (b), we can see that the estimation of matrix **A** is a bank of localized, oriented and frequency selective Gabor-like filters. Each little square, in the figures above, represent one receptive field. By visual inspection of Fig. 2 (b), one can see that the orientation and location of each visual field mostly change smoothly as a function of position on the topographic grid. Thus, the emergence of organized Gabor-like receptive fields can be understood as a biologically plausibility of our model.

This model can be regarded also as biologically plausible as it works in an unsupervised fashion with local adaptation rules Olshausen and Field (1996). To do this, we had chosen some even symmetric nonlinearities which were applied upon a neuron internal signal.This signal interacts with similar ones from neighborhoods neurons to generate its output. This model mimics the V1 cells by an interaction between neighborhood signals, which is possible due to the fact that signals from neighborhood neurons are reference to the activation of one specific neuron Field (1987). In addition, our method extracts oriented, localized and bandpass filters as basis functions of natural scenes without restricting the probability density function of the network output to exponential ones, as the IP algorithm of Butko et al Butko and Triesch (2007). Moreover, Stork and WilsonStork and Wilson (1990) proposed an architecture very similar to the one proposed here, which is largely based on neural architecture.

## ACKNOWLEDGEMENTS

The authors would like to thank Brazilian Agency FAPEMA for continued support.

## REFERENCES

Barlow, H. B., Unsupervised learning, Neural Computation,v.01, pp. 295-311, 1989.

Barros, A. K. and Ohnishi, N., Wavelet like Receptive fields emerge by non-linear minimization of neuron error, International Journal of Neural Systems,v.13, n. 2, pp. 87-91, 2002.

Bell, Anthony J. and Sejnowski, Terrence J., An Information-Maximixation Approach to Blind Separation and Blind Deconvolution, Neural Computation, n. 7, pp. 61-72, 1995.

Butko, N. J. and Triesch, J., Learning sensory representations with intrinsic plasticity, Neurocomputing, v. 70, pp. 1130-1138, 2007.

Comon, Pierre. Independent component analysis, A new concept?. Signal Processing, v.36,p 287-314, 1994.

Durbin, R and Mitchison, G., A dimension reduction framework for understanding cortical maps, Nature, n.343, pp. 644-647, 2007.

Field, David J., Relation between the statistics of natural images and the response properties of cortical cells. J. Optical Society of America, v. 04, n.12, pp. 2379 - 2394, 1987.

Foldiák, P., Some Aspects of Linear Estimation wiht Non-Mean Squares error Criteria. Biological Cybernetics, v. 64, pp. 165 - 170, 1990.

Gersho, A., Forming sparse representations by local anti-Hebbian learning., Proc. Asilomar Ckts. and Systemas Conf. (1969).

Hyvarinen, Aapo; Hoyer, Patrik O. and Inki, Mika., Topographic ICA as a Model of V1 Receptive Fields. Proc. of International Joint Conference on Neural Networks, v.03, pp 83-88, 2000.

Hyvarinen, Aapo and Hoyer, Patrik O., Topographic Independent Component Analysis. Neural Computation, v.13, pp 1527-1558. (2000).

Kohonen, T., Self-Organizing Maps. Springer, 2000.

Linsker, R., Local Synaptic Rules Suffice to maximize Mutual Information in a Linear Network. Neural Computation, v.04, pp 691-702, 1992.

Laughlin, Simon., A Simple Coding Procedure Enhances a Neuron's Information Capacity. Z. Naturforsch, v.36, pp 910-912, 1981.

Oja, E., Neural Networks,Principal Components and Linear Neural Networks. Neural Networks, v.05, pp 27-935, 1989.

Olshausen, B. A. and Field, D. J., Emergence of Simple-Cell Receptive Field Properties by Learning a Sparse Coding for natural Images. Nature, v.381, pp 607-609, 1996.

Osindero, Simon; Welling, Max and Hinton, Geoffrey E., Topographic Products Models Applied to Natural Scene Statistics. Neural Computation, v.18, pp 381-414, 2006.

Ray, K. L., McKay, D. R., Fox, P. M., Riedel, M. C., Uecker, A. M., Beckmann, C. F., Smith, S. M., Fox, P. T., Laird, A. ICA model order selection of task co-activation networks. Front. Neurosci. 7:237; 2013

Simoncelli, E. P and Olshausen, B. A., Natural Image Statistics and Neural representations. Annu. Rev. Neuroscience, n. 24, pp 1193-1216, 2001.

Shannon C. E., A Mathematical Theory of Communication. Bell Systems Technical Journal, v. 27, pp 379-423, 1948.

Stork, D.G. and Wilson, H.R. Do Gabor functions provide appropriate descriptions of visual cortical receptive fields? J Opt Soc Am A7, 1362-1373, 1990.

Vinje, W. E. and Gallant, J. L., Sparse Coding and Decorrelation in Primary Visual Cortex During natural Vision. Vision Science, v. 287, pp 1273-1276, 2000.

# APPENDIX

Take an indexed set of N neurons, as one can see in table 1, and put them into a $\sqrt{N} \times \sqrt{N}$ matrix (let's suppose N=64,

Table 1: A set of N neurons.

| | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | $\cdots$ | N |

for example), as one can see in table 2

We define the neighborhood of neuron $k$, all neurons $i$ that are close to the neuron $k$ in the matrix representation. In other words, neurons $i$ are neighbors of neuron $k$ if the coordinates of neuron $i$ obeys the following relation

$$D(i,k) = \sqrt{(a_i - a_k)^2 + (b_i - b_k)^2} \leqslant T, \tag{7}$$

where $(a_k, b_k)$ and $(a_i, b_i)$ are the coordinates of neuron $k$ and neuron $i$, respectively, in the matrix representation, $T$ is a constant that defines the neighborhood size. For example, for $T = 1$, the neighborhood of neuron 28 as highlighted in the table below. In this case $v_{28} = \{19, 20, 21, 27, 29, 35, 36, 37\}$.

Table 2: One neighborhood of neuron 28.

| | | | | | | | |
|---|---|---|---|---|---|---|---|
| 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 |
| 9 | 10 | 11 | 12 | 13 | 14 | 15 | 16 |
| 17 | 18 | 19 | 20 | 21 | 22 | 23 | 24 |
| 25 | 26 | 27 | 28 | 29 | 30 | 31 | 32 |
| 33 | 34 | 35 | 36 | 37 | 38 | 39 | 40 |
| 41 | 42 | 43 | 44 | 45 | 46 | 47 | 48 |
| 49 | 50 | 51 | 52 | 53 | 54 | 55 | 56 |
| 57 | 58 | 59 | 60 | 61 | 62 | 63 | 64 |

With this definition of neighborhood, we are able to analyze the algorithm stability in the following steps:

Assume that signal $\mathbf{x}$ can be modeled as $\mathbf{x}$=$\mathbf{As}$, where $\mathbf{A}$ is an invertible matrix;

Let us make the following change in coordinates, $\mathbf{p}_k = [p_1, p_2, \cdots, p_n]^T = \mathbf{A}^T\mathbf{V}^T\mathbf{w}_k$, where $\mathbf{V}$ is the whitening matrix to obtain the signals $\mathbf{x}$;

In this new coordinates basis, we define the cost function as

$$\mathbf{J}(\mathbf{p}_k) = E[\varepsilon_k^2] = E\{[f(\mathbf{p}_k^t\mathbf{s}) + \alpha \cdot v_k]^2\}; \tag{8}$$

One can analyze, without loss of generality, the stability of $\mathbf{J}(\mathbf{p}_k)$ at point $\mathbf{p}_k = [0, \cdots, p_k, \cdots, 0]^T = \mathbf{e}_k$. In this case, $p_k = 1$ and than $\mathbf{w}_k$ is one row of the inverse of $\mathbf{VA}$. For assumption, $\mathbf{J}(\mathbf{p}_k)$ is an even function and this analysis applies for $p_k = -1$;

Let $\mathbf{d} = [d_1, \cdots, d_n]^T$ be a small perturbation on $\mathbf{e}_k$. Expressing $\mathbf{J}(\mathbf{e}_k + \mathbf{d})$ into a Taylor series expansion about $\mathbf{e}_k$, we obtain

$$\mathbf{J}(\mathbf{e}_k + \mathbf{d}) = \mathbf{J}(\mathbf{e}_k) + \mathbf{d}^T\frac{\partial \mathbf{J}(\mathbf{e}_k)}{\partial \mathbf{p}} + \frac{1}{2}\mathbf{d}^T\frac{\partial^2 \mathbf{J}(\mathbf{e}_k)}{\partial \mathbf{p}^2}\mathbf{d}^T + o(\|\mathbf{d}^2\|). \tag{9}$$

$$\frac{\partial \mathbf{J}(\mathbf{p})}{\partial p_k} = s_k f'(\mathbf{p}^T\mathbf{s})[f(\mathbf{p}^T\mathbf{s}) + v_k], \tag{10}$$

$$\frac{\partial \mathbf{J}(\mathbf{p})}{\partial p_k^2} = s_k^2 f''(\mathbf{p}^T\mathbf{s})[f(\mathbf{p}^T\mathbf{s}) + v_k] + s_k^2 f'(\mathbf{p}^T\mathbf{s}) f'(\mathbf{p}^T\mathbf{s}) \tag{11}$$

and

$$\frac{\partial \mathbf{J}(\mathbf{p})}{\partial p_k^2} = s_k s_j f''(\mathbf{p}^T\mathbf{s})[f(\mathbf{p}^T\mathbf{s}) + v_k] + s_k s_j f'(\mathbf{p}^T\mathbf{s}) f'(\mathbf{p}^T\mathbf{s}) \tag{12}$$

Taking in account that signals out of neighborhood are statistically independent, we have

$$\begin{aligned} \mathbf{d}^T\frac{\partial(\mathbf{e}_k)}{\partial \mathbf{p}} &= 2\{E\left[s_k f(s_k) f'(s_k)\right] d_k + \\ &+ \alpha \sum_{i \in v_k} E\left[s_i f(s_k) f'(s_k)\right] d_i\}, \end{aligned} \tag{13}$$

because in consequence of independency, $E[s_j f(s_k) f'(s_k)] = 0$, if $j \notin v_k$.

In the same way, using the assumption of independence, one can obtain

$$
\begin{aligned}
& \frac{1}{2}\mathbf{d}^T \frac{\partial^2 \mathbf{J}((e)_k)}{\partial \mathbf{p}^2}\mathbf{d} = \\
= \quad & E\left[s_k^2(f'(s_k))^2 + s_k^2 f(s_k) f''(s_k)\right] d_k^2 \\
+ \quad & \alpha \sum_{i \in v_k} \{E\left[s_i^2(f'(s_k))^2 f''(s_k)\right] d_i^2 \\
+ \quad & E\left[s_k s_i (f'(s_k))^2 + s_k s_i f(s_k) f''(s_k)\right] d_k d_i\}, \\
+ \quad & \sum_{i \notin v_k} E\left[(f'(s_k))^2 + f(s_k) f''(s_k)\right] d_j^2 + o(\|\mathbf{d}\|^2).
\end{aligned}
\tag{14}
$$

Supposing $\|\mathbf{w}\| = 1$ and **VA** being orthogonal, we have $\|\mathbf{p}\| = \|\mathbf{A}^T\mathbf{V}^T\mathbf{w}\| = 1$. Consequently, $\|\mathbf{e}_i + \mathbf{d}\| = 1$, which implies $d_k = \sqrt{1 - \sum_{\forall l \neq i} d_l^2} - 1$.

Remembering that $\sqrt{1-\beta} = 1 - \frac{\beta}{2} + o(\beta)$, one can discard the terms which contains $d_k^2$ and $d_k d_i$ in step 8 above, because they are $o\|d^2\|$ .

Now we can take the following approximation: $d_k \approx \sum_{\forall l \neq i} d_l^2 = -\sum_i d_i^2 - \sum_j d_j^2$.

Using the above approximation in step 11, and using equation 13 and equation 14, in equation 9, one get

$$
\begin{aligned}
& \mathbf{J}(\mathbf{e}_k + \mathbf{d}) = \mathbf{J}(\mathbf{e}_i) \\
+ \quad & \alpha \sum_i \{E[s_i^2(f'(s_k))^2 + s_i^2 f(s_k) f''(s_k) \\
+ \quad & s_i f(s_k) f'(s_k) - s_k f(s_k) f'(s_k)] d_i^2\} \\
+ \quad & \sum_j E[(f'(s_k))^2 + f(s_k) f''(s_k) \\
- \quad & s_k f(s_k) f'(s_k)] d_j^2.
\end{aligned}
\tag{15}
$$

Choosing properly a function $f$ so that $E[s_i^2(f'(s_k))^2 + s_i^2 f(s_k) f''(s_k) + s_i f(s_k) f'(s_k) - s_k f(s_k) f'(s_k)] + E[(f'(s_k))^2 + f(s_k) f''(s_k) - s_k f(s_k) f'(s_k)] < 0, \forall i \in v_k$, we obtain that $\mathbf{J}(\mathbf{e}_i + \mathbf{d})$ will be always small that $\mathbf{J}(\mathbf{e}_i)$ in equation 9.