Energy Landscape and Learning on a Neural Network for the Sum Problem

Leonardo Franco [1] *and Sergio A. Cannas* [2]
Facultad de Matemática, Astronomía y Física. Universidad Nacional de Córdoba.
Ciudad Universitaria. (5000). Córdoba. Argentina.
email: {franco,cannas}@fis.uncor.edu

Abstract. A feed-forward Neural Network for the Sum problem of two numbers is implemented. The Error function is studied exhaustively showing many interesting features such as local minima (that does not appear in the 1-bit problem), plateaus, flat minima, etc. Different optimization strategies are implemented for the learning problem.

Using the local minima structure of the error, we analyze how the selection of examples in the training set influences the learning perfomance.

We obtain an upper bound for the minimal number of examples needed for generalization for the general case of adding two numbers of N-bits length.

Introduction

In 1996 Sprinkhuizen-Kuyper & Boers showed in Ref.[1] that the simplest XOR function has no local Minima. We decide to study more complicated functions to see if local minima appear, trying to understand why they do appear, and with this information test different algorithms through the learning process.

We study the energy function of a network that performs the sum operation of two numbers of N-bit length, which reduces for the 1-bit case to the XOR function.

The Structure of the Network to sum two N-bit numbers.

The network is a perceptron with one hidden layer, that solves, with determinated weights, the sum problem. See [2] for a complete description of the structure. The input layer has 2N units, the hidden layer has N, and the output layer has N+1 neurons. We consider the symmetric case, where synapses connecting one neuron to input bits with the same significant values are equal.

The energy of the network, E, is defined as the difference between the output calculated by the network and the desired output (target), calculated over the training set:

$$E\{w_j, K_i\} = \sum_{i=1}^{K} (S_i\{w_j, K_i\} - O_i\{K_i\})^2, \tag{1}$$

where K is the number of examples in the training set, K_i is a particular example, and $\{w_j\}$ are the synaptic weights.

[1] Fellow of the Secretaría de Ciencia y Tecnología de la Universidad Nacional de Córdoba. Argentina

[2] Member of the National Research Council (CONICET). Argentina.

The generalization error is similar to the energy but takes into account all possible examples.

Existence of Local minima

We study the sum of two numbers of 2-bit in the symmetric case, restricting our study to the second output bit, because the study of the first output bit was exhaustively done in [1], and also because it is a particular case of the second output bit problem. The structure for computing this second output bit is shown in fig. 1.

The Output neuron, S, computes the following function:

$$S = \sigma[w_1 * (x_1 + x_2) + w_2 * \sigma(w_3 * (x_1 + x_2) + w_4 * (x_3 + x_4) - h_1) + w_5 * \sigma(w_6 * (x_3 + x_4) - h_2) - h_0]$$

where w_i are the synaptic weights, h_i are the thresholds, and $x_i, (i = 1, .., 4)$ are the input bits, where x_1, x_2 are the bits with significative value 2^1 corresponding to the two input numbers and x_3 and x_4 are the bits with value 2^0. The activation function, σ that neurons compute is a sigmoid transfer function of the form:

$$\sigma(x) = \frac{1}{1 + e^{-x}}.$$

The input neurons $(x_i, i = 1, .., 4)$ are set with binary values $\{0,1\}$ depending on the example to be evaluated. The desired outputs are binaries, but as the output of the network is continuous, we set the targets to fixed values, to avoid an excessive growing of the synapses, as follows:

- δ in the case corresponding to an output 0 (Neuron OFF).
- $1 - \delta$ in the case corresponding to an output 1 (Neuron ON).

We select δ to be 0.1 but any small positive number can be used.

For the case of adding two numbers of two-bit length (N=2), the number of different training patterns (examples) is 16, but just 9 of them are independent due to the simmetry imposed to the synapses.

To find the possible local minima of Eq.(1) is necessary to solve the 9 equations:

$$\frac{\partial E}{\partial w_i} = \frac{\partial E}{\partial h_i} = 0 \;\; \forall i$$

This procedure leads to a set of 9 equations with 9 variables too complicated to solve analytically. Our approach use optimization methods to find local minima and then try to show that they are indeed local minima. This procedure involves the calculus of the eigenvalues of the Hessian Matrix corresponding to Eq.(1), and to see that a point is a local minima all these eigenvalues should be positive definite. The elements of the hessian matrix coming from Eq.(1) has the form;

$$H_{ij} = \frac{\partial^2 E}{\partial^2 w_{ij}}$$

Numerical problems appear trying to demostrate the positiveness of the eigenvalues, but we have strong evidence that many local minima exist for this problem.

The Energy Landscape and the behaviour of the examples.

Through the use of the optimization algorithms we encounter different characteristic elements of the landscape, **Flat minima**, **Plateaus**, **High barriers**, etc. All these components turn the optimization procedure very complex, slowing down the rate of learning and the probability of success searching for the global minima. In figure (2) we show the energy along the direction between two minima one local and the other global.

The energy landscape function Eq.(1) is sum of many particular energies of the individual examples. We study these individual energies obtaining the following characteristics:

Around the global minima we can clearly distinguish two kinds of examples, according to its contribution to the energy function; those with output 0.1 (OFF) and the others with output 0.9 (ON). Also, at the global minima there are some examples whose contribution to the energy its greater than the others. See figure (3).

We analyze the different local minima classifying them in two groups: The first one, where all the examples have the same residual energy, as shown in figure (4), and the second group where we can clearly distinguish between examples that have been learnt and others that have not.

Data collected from many local minima show that examples involving bit-carry are the most difficult ones to be learnt.

Improving Generalization: Selection of Examples

It was shown by Kinzel & Ruján [2] among others, that selection of examples is a good way to improve generalization. With the information obtained from the study of the energy landscape we make a selection of examples for the sum problem. For simplicity, we consider a network with binaries neurons, instead of the continuos ones used before, and we study the case $N = 3$ because the previous one, with $(N = 2)$, has too few examples, 9, to test a selection. For the case $N = 3$, we analyze the 27 equations for every particular energy of the examples, finding that 10 selected examples are enough to obtain generalization.

This result can be generalized for **arbitrary** N. In this way, we find **analytically** an upper bound for the minimal number of examples needed for perfect generalization, N_m:

$$N_m \leq 2(2N - 1).$$

Comparing this number with the total number of synapses of the net for the whole sum problem, N_s:

$$N_s = \frac{1}{2}(N^2 + 5N - 2),$$

we have, for $N \gg 1$, that the relation between them behaves as:

$$\frac{N_m}{N_s} \sim \mathcal{O}(\frac{1}{N})$$

.

We also test the generalization properties through numerical simulations (simulated annealing) for $N = 3, 4, 5$, using **random** equiprobably chosen examples. We find for this case that the **average** minimal number of examples, $\langle N_m \rangle$, behaves as:

$$\langle N_m \rangle \sim \mathcal{O}(N^2)$$

.

Conclusions

We analyze the energy landscape of the sum problem of two N-bit numbers, founding that, for $N > 1$, it shows a very complex structure, with **local minima, flat minima** and **plateaus**. The analysis of this structure leads to the conclusion that some **classes** of examples strongly influence the energy landscape and therefore, they must be included in the selection in order to obtain good generalization.

We find **analytically** an upper bound to the minimal number of examples for optimal generalization, $N_m \sim \mathcal{O}(N)$, which compares very well to the average minimal number of randomly chosen examples, $\langle N_m \rangle \sim \mathcal{O}(N^2)$.

All these results can be used for develop **general** criteria for improving the generalization procedure by selection of examples. Works along these lines are in progress.

References

1. Sprinkhuizen-Kuyper, I.G., Boers E.J.W. 1996. The Error Surface of the Simplest XOR Network Has Only Global Minima. *Neural Computation.* **8** 1301-1320.

2. Cannas, S.A. 1995. Arithmetic Perceptrons, *Neural Computation* **7** (1) 173-181.

3. Kinzel, W., and Ruján, P. 1990. Improving a Network Generalization Ability by Selecting Examples. *Europhysics Letters* **13** (5) 473-477.

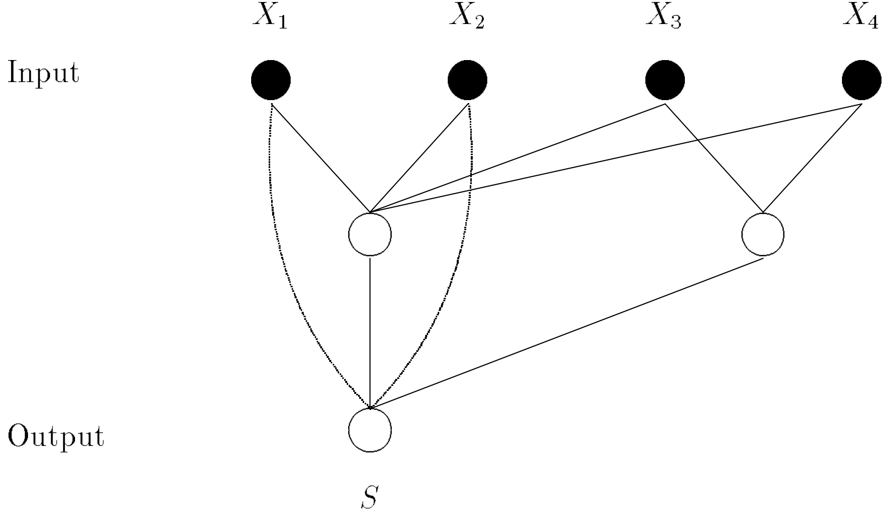
Captions for figures

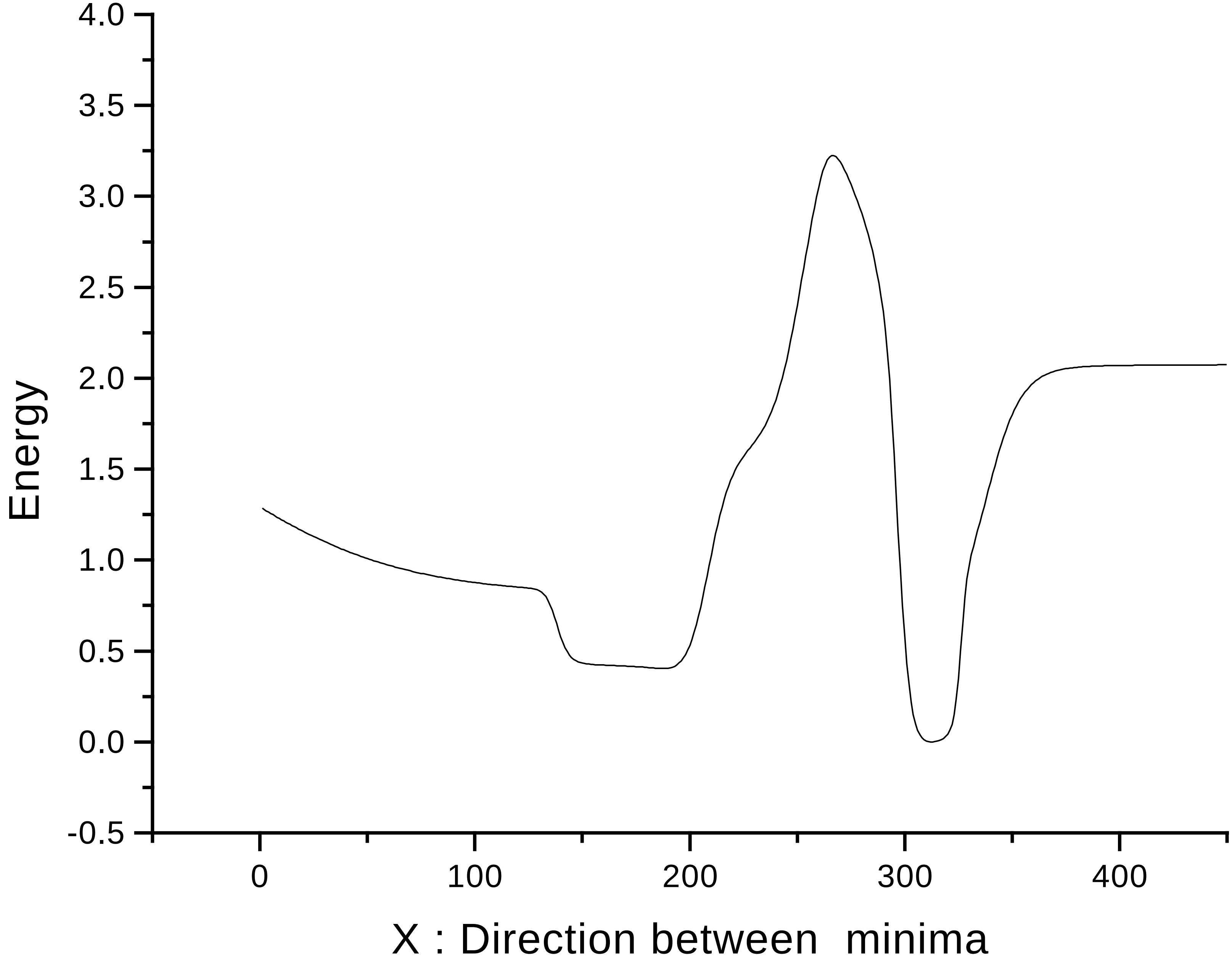
- Figure 1: Network Structure for the second Output bit for the Sum Problem of two numbers.

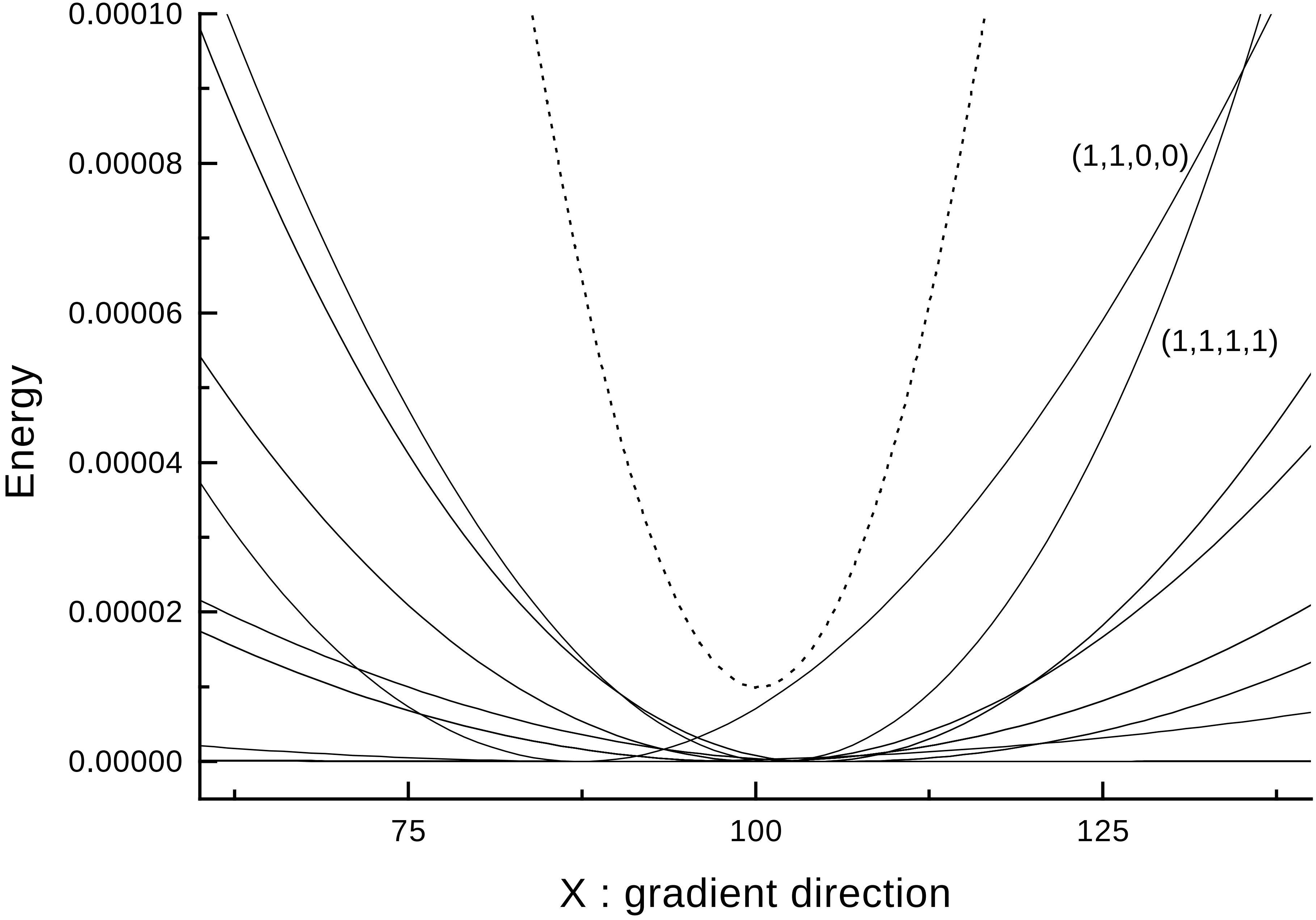
- Figure 2: Energy landscape along the direction between a global minimum and a local one.

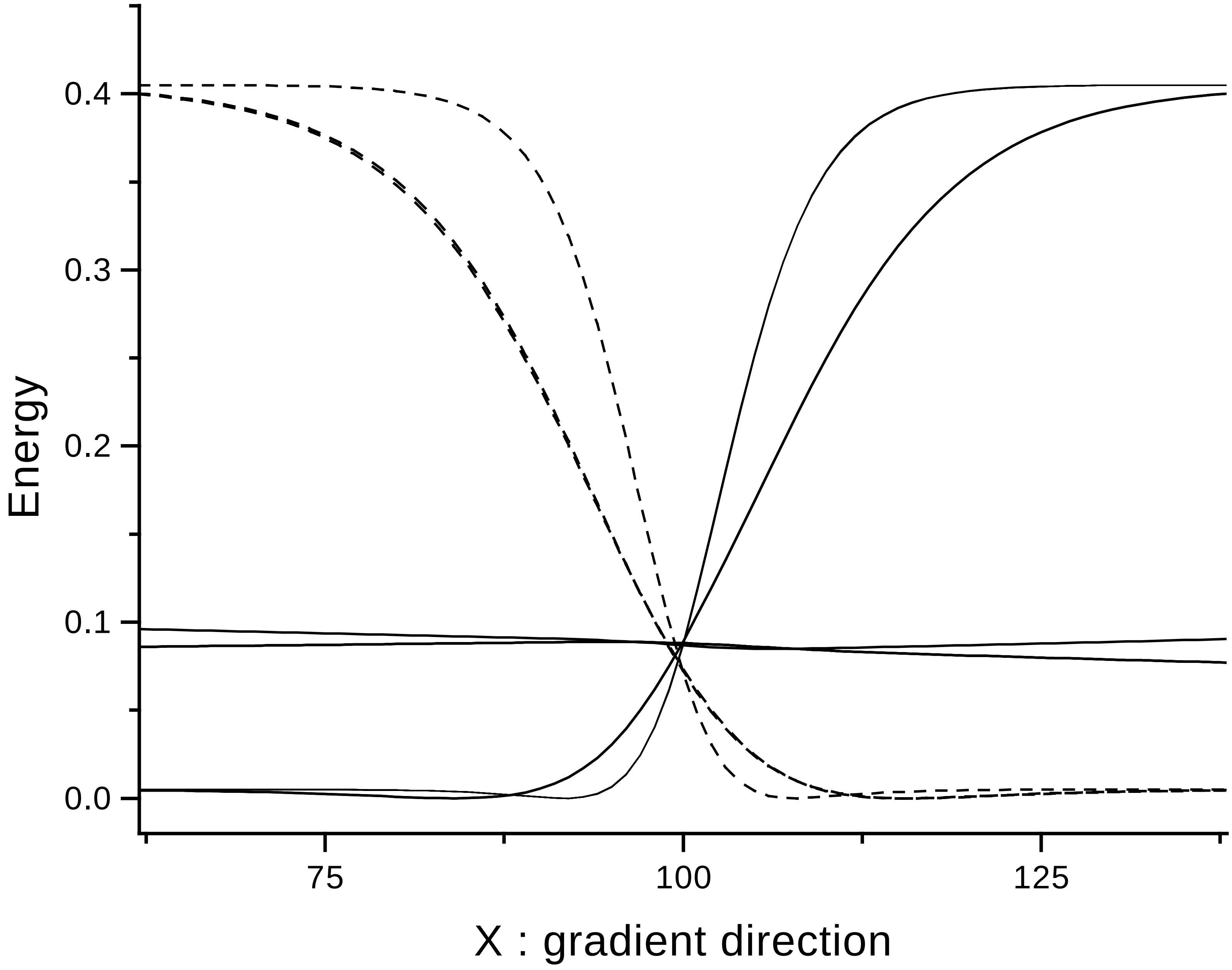
- Figure 3: Behaviour of the examples near a global minima. In dotted line the Total energy and in dashed lines the energy corresponding to the examples.

- Figure 4: Local minima where all the examples have the same non-zero energy equal to 0.08.


X_1
X_2
X_3
X_4
Input
Output
S


Energy
4.0
3.5
3.0
2.5
2.0
1.5
1.0
0.5
0.0
-0.5
0
100
200
300
400
X : Direction between minima


0.00010
0.00008
0.00006
0.00004
0.00002
0.00000
Energy
75
100
125
X : gradient direction
(1,1,0,0)
(1,1,1,1)


Energy
0.0
0.1
0.2
0.3
0.4
75
100
125
X : gradient direction