# Energy Functions for Efficient Nonlinear Dimensionality Reduction by Multi Layer Perceptrons

*Takashi TAKAHASHI†and Ryuji TOKUNAGA††*
*Email:ume@chaos.is.tsukuba.ac.jp*

†Doctoral Programs in Engineering,
University of Tsukuba, 305–8573, Japan
††Institute of Information Sciences and Electronics,
University of Tsukuba, 305–8573, Japan

## ABSTRACT

This paper investigates two simple energy functions which are valid for two different purposes of dimensionality reduction: feature extraction and data compression. These energy functions enable nonlinear perceptrons to organize data representations whose parameters, namely, outputs of the bottleneck layer units, are arranged in the order of their importance. The efficacy of these energy functions is shown by numerical experiments in comparison with conventional squared error functions and Principal Component Analysis.

**KEYWORDS: multi layer perceptron, dimensionality reduction, data compression**

## 1. Introduction

Dimensionality reduction is one of major topics in the field of neural networks and is widely applied to many engineering problems. Roughly speaking, the objectives of dimensionality reduction are

- feature extraction — Dimensionality reduction is applied to analyze the structure of data and to identify the intrinsic dimensionality, namely, the number of independent parameters which span the feature space.
- data compression — Dimensionality reduction is applied to represent each individual data with minimal number of effective parameters and to reduce its mean number.

In either case, parameters of the data representation are desired to be arranged in the order of their importance. This paper refers to such a data representation as **ordered representation**.

For instance, Principal Component Analysis (PCA) is one of the powerful tools for dimensionality reduction since it gives ordered representations suitable for both purposes. PCA transforms each data onto a coordinate system whose (principal) components are arranged in the decreasing order of their variance; the superior components representing the important features of the data. Hence, for feature extraction, feature space can be formed by extracting the significant elements of the principal components. For data compression, efficient quantization is feasible by allocating more bits to the dominant components and fewer bits to the rest.

On the other hand, Multi Layer Perceptron (MLP) can also perform dimensionality reduction if the number of units in the bottleneck layer is smaller than the dimensionality of the data. As for their capability of dimensionality reduction, it is shown that a three layer perceptron accomplishes equal reconstruction error as PCA[1, 2] and that nonlinear perceptrons with more (typically, five) layers attain smaller errors with the identical dimensionality[3, 4]. In these cases, outputs of the bottleneck layer units are the parameters of the data representation. However, they do not compose an ordered representation. Therefore, nonlinear MLP is not as useful as PCA for feature extraction or data compression.

In order to settle the above problem of MLP, this paper investigates energy functions which organize ordered representations. It also reports that two different energy functions can be derived with respect to feature extraction and data compression.

## 2. Energy Functions for Ordered Representation

Let us consider the dimensionality reduction of a data set

$$\left\{\boldsymbol{x}_t \in R^N\right\}_{t=1}^{T}. \tag{1}$$

assuming that $\mathrm{E}[\boldsymbol{x}_t] = \mathbf{0}$ for simplicity. The dimensionality reduction of (1) can be performed by MLP which consists of $N$ input, $N$ output, and $H$ bottleneck layer units. If such a perceptron is trained with a target set identical to the input data set, it tries to generate an identity mapping with the given data. To make differences between data representations of MLP and PCA clear, let $f_i$ $(i = 1, 2, \ldots, H)$ denote a function which approximates the data only with the first $i$ parameters of the data representation. In the case of PCA, $f_i$ is a linear transformation composed of the eigenvectors associated with the first $i$ principal components. In the case of MLP, $f_i$ is given by considering only the first $i$ units among all bottleneck layer units. This paper refers to the function $f_i$ of MLP as the $i$-th sub-perceptron (Fig. 1). The reconstruction error of $f_i$ associated with $\boldsymbol{x}_t$ is given by

$$e_{i,t} = ||\boldsymbol{x}_t - f_i(\boldsymbol{x}_t)||^2, \tag{2}$$

and the mean reconstruction error of $f_i$ is given by

$$E_i = \frac{1}{T}\sum_{t=1}^{T} e_{i,t} = \frac{1}{T}\sum_{t=1}^{T} ||\boldsymbol{x}_t - f_i(\boldsymbol{x}_t)||^2. \tag{3}$$

Note that $E_H$ is the energy function used for conventional error back propagation learning.

Figure 2 shows examples of $\{E_i\}_{i=1}^{H}$ obtained by PCA and a five layer nonlinear perceptron trained by $E_H$ with $(N, H) = (16, 4)$ (As for the target data, see Sec. 4). PCA obtains the ordered representation which minimizes all $E_i$s. As long as all units are used, MLP shows better performance than PCA. However, the error becomes larger when even one bottleneck layer unit is removed. Therefore, any subset of the parameters extracted from the data representation is not as effective as those of PCA.

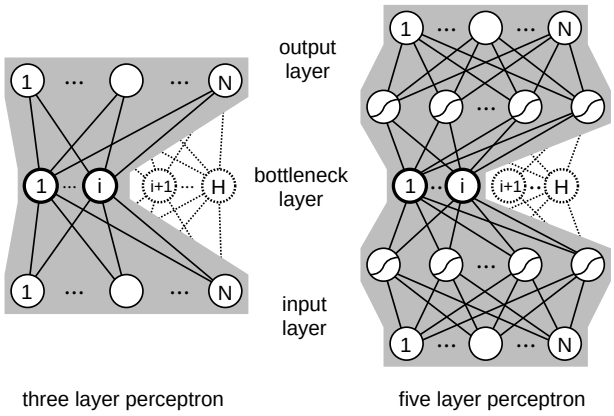


Figure 1: Perceptrons and sub-perceptrons : Shading regions represent the $i$-th sub-perceptrons.

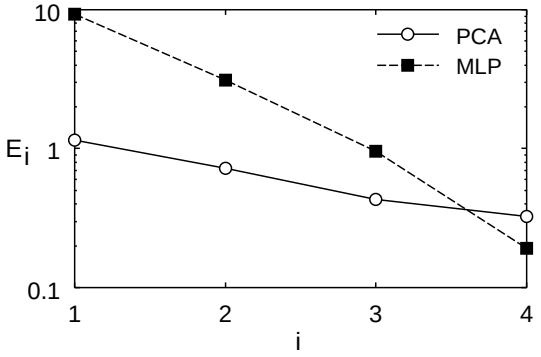


Figure 2: $i$–$E_i$ characteristics

Effective feature extraction requires the ordered representation of MLP to minimize all $E_i$s. One of the simplest methods is learning with a weighted sum of all $E_i$s

$$F = \sum_{i=1}^{H} \beta_i E_i, \tag{4}$$

where each $\beta_i$ is a positive constant. By using this energy function, the bottleneck layer units with smaller indices tend to extract more important features. Particularly, in the case of a three layer linear perceptron, it is proven that each bottleneck layer unit generates the corresponding principal component[5]. This paper refers to $F$ as **superposed energy function**[1].

[1]In this paper, the set of constants $\{\beta_i\}$ is given by

$$\beta_i \propto i^2 \qquad \text{provided that} \quad \sum_{i=1}^{H} \beta_i = 1.$$

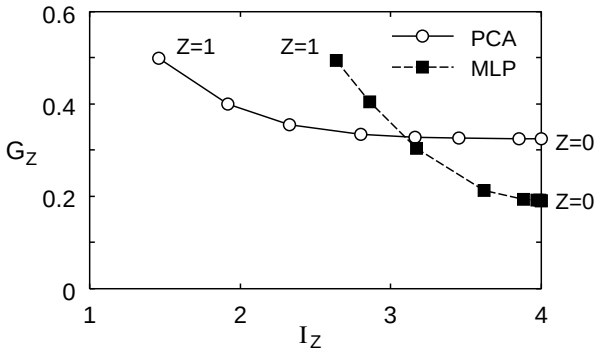


Figure 3: $I_Z$–$G_Z$ characteristics

On the other hand, to achieve an efficient data compression, each data must be represented with a minimal number of parameters which preserve the significant structure of the original data. In this view point, data compression capability of MLP and PCA can be measured with more accuracy in terms of **mean number of effective parameters** which is computed as follows.

1. Set a permissible error $Z \in R^1$.
2. Calculate the error $e_{i,t}$ $(i = 0, 1, \ldots, H)$ for each $t$ provided that $e_{0,t} \equiv ||\boldsymbol{x}_t - \mathrm{E}[\boldsymbol{x}_t]||^2$.
3. For each $t$, determine the minimal sub-perceptron or necessary number of effective parameters by

$$I(t, Z) = \begin{cases} H & (e_{i,t} \geq Z \text{ for all } i \in \{0, 1, \ldots, H\}) \\ \text{min. } i \text{ satisfying } e_{i,t} < Z & \text{(otherwise).} \end{cases} \tag{5}$$

4. Compute the mean number of effective parameters

$$I_Z = \frac{1}{T} \sum_{t=1}^{T} I(t, Z), \tag{6}$$

and the associated reconstruction error

$$G_Z = \frac{1}{T} \sum_{t=1}^{T} ||\boldsymbol{x}_t - f_{I(t,Z)}(\boldsymbol{x}_t)||^2. \tag{7}$$

Figure 3 shows $I_Z$–$G_Z$ characteristics with respect to the above example. It is seen that the ordered representation of PCA obtains better results compared with MLP. MLP cannot reduce the mean number of effective parameters without an increase in error.

In order to optimize the reconstruction error (7), let us consider it as another energy function for MLP

$$G = \frac{1}{T} \sum_{t=1}^{T} ||\boldsymbol{x}_t - f_{I(t,Z)}(\boldsymbol{x}_t)||^2. \tag{8}$$

Equation (8) is equal to $E_H$ provided that the permissible error $Z = 0$. If an appropriate value of $Z$ is given, the perceptron trained by this energy function approximates each

data using as few bottleneck layer units as possible. Hence energy function $G$ is expected to organize better data representation for data compression than energy function $F$. This paper refers to $G$ as **adaptively selected energy function**[2].

## 3. Experiment

To evaluate the validity of the proposed energy functions, nonlinear dimensionality reduction is performed using five layer nonlinear perceptrons. The SIDBA standard image 'Girl' ($256 \times 256$ pixels, 8 bit grey scale) is selected as the target. It is divided into image vectors $\{\boldsymbol{v}_t \in R^{64}\}_{t=1}^{1024}$ consisting of $8 \times 8$ pixels. The training data set $\{\boldsymbol{x}_t \in R^{16}\}_{t=1}^{1024}$ is generated by extracting the first 16 principal components from $\{\boldsymbol{v}_t\}_{t=1}^{1024}$. The perceptrons consist of

- 16 linear input units
- 32 sigmoidal hidden units
- $H = 4$ or 8 linear bottleneck layer units
- 32 sigmoidal hidden units
- 16 linear output units

and they are trained by the following three energy functions:

- conventional energy function $E_H$,
- superposed energy function $F$,
- adaptively selected energy function $G$.

Figure 4 (a) (resp. (b)) shows the reconstruction errors $\{E_i\}_{i=1}^{H}$ obtained by PCA and the perceptrons[3] with $H = 4$ (resp. 8). Any of the three energy functions achieve smaller errors than PCA if all the units are used. The errors $E_H$ obtained by superposed energy are a little larger than those of the other two energy functions. As for the errors of the sub-perceptrons, however, the best result is obtained by superposed energy function. Each sub-perceptron has smaller error than that of PCA for the same dimensionality. Therefore, superposed energy function acquires the most efficient data representation for feature extraction.

Next, the mean number of effective parameters $I_Z$ is computed. Figure 5 (a) (resp. (b)) plots the $I_Z$–$G_Z$ characteristics of PCA and the perceptrons with $H = 4$ (resp. 8). It is clearly seen that the mean number of effective parameters corresponding to the proposed energy functions are almost the half of the original. These perceptrons realize more effective data compression than PCA for a wide range of $Z$.

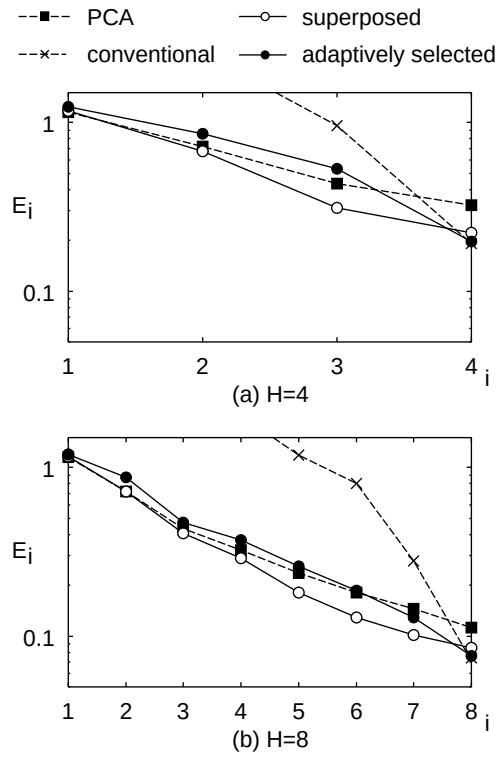


Figure 4: Feature extraction from the image data

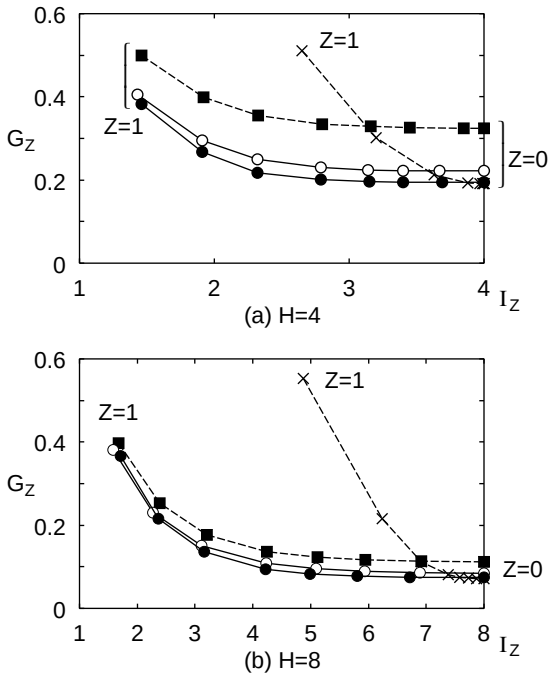


Figure 5: Compression of the image data

[2] In this paper, $Z$ is given by

$$Z = \frac{1}{T}\sum_{t=1}^{T} e_{H,t} = E_H$$

and the value is also floated while learning.

[3] Since the order of the indices $i$ has no meaning in the case of conventional energy function, they are sorted after learning by using the method described in [5].

## 4. Discussion

This paper reported that superposed and adaptively selected energy functions enable MLP to organize ordered representations which are suitable for feature extraction and data compression, respectively. Superposed energy function generates nonlinear MLP to extract feature components whose dimensionality is smaller than that of PCA. Adaptively selected energy function generates nonlinear MLP to reduce the mean number of parameters of a set of data[4].

Although superposed energy function is effective, it suffers from practical drawbacks. Since its learning procedure requires learning of $H$ different sub-perceptrons, its computational cost is extremely high compared with that of conventional energy function. On the other hand, the computational cost for adaptively selected energy is comparable to that of conventional energy function. Table 1 shows the number of multiplicative operations required for the forward procedure ( calculation of outputs) as well as the backward procedure (propagation of errors and modification of weights). They are estimated for the perceptron used in the above experiment ($H = 8$). It is seen that the computational cost for adaptively selected energy function is lower than that of conventional energy function.

Table 1: Comparison of the computational costs

| | conventional | superposed | adaptive |
|---|---|---|---|
| forward | 1600 (times) | 5152 | 1344 |
| backward | 2624 | 11808 | 2112 |
| total | 4224 | 16960 | 3456 |
| ratio | 1 | 4.02 | 0.82 |

## References


[1] Bourlard, H. and Kamp, Y., "Auto-association by multilayer perceptrons and singular value decomposition," Biol. Cybern., 1988.

[2] Baldi, P. and Hornik, K., "Neural networks and principal component analysis: Learning from examples without local minima," Neural Networks, vol.2, pp.53–58, 1989.

[3] Usui, S., Nakauchi, S. and Nakano, M., "Internal color representation acquired by a five-layer neural network," Proc. Int. Conf. Artificial Neural Networks, vol.1, pp.867–872, 1991.

[4] DeMers, D. and Cottrell, G., "Non-linear dimensionality reduction," in Advances in Neural Information Processing Systems 5, eds. S. J. Hanson, J. D. Cowan and C. L. Giles, pp.580–587, 1993.

[5] Takahashi, T. and Tokunaga, R., "Removing the redundancy of perceptrons in terms of a simple energy function," Proceedings of the ICONIP'97, vol.1, pp.271–274 (1997).


[4] If each data is expressed by $I(t, Z)$ parameters, the mean number of parameters is reduced to $I_Z$ without loss of the information. In this case, additional $\log_2(H+1)$ bits are required to specify the number of parameters. However, this amount is small in comparison with the number of bits to represent the parameters.