

---

# An Info-Equilibrium Model for Supervised Information-Theoretic Neural Networks

---

**Chih-Chung Kao**  
Information Science PhD Program  
University at Albany, SUNY  
Albany, NY 12222  
*ck7879@csc.albany.edu*

**George Berg**  
Computer Science Department  
University at Albany, SUNY  
Albany, NY 12222  
*berg@cs.albany.edu*

## Abstract

In this paper, we present an Info-Equilibrium (Info-Eq) model for learning in supervised, information-theoretic neural networks. Info-Eq uses a damped maximum entropy approach that stabilizes the mutual information between network inputs and outputs, which promotes generalization in network learning. We show the relationship of the Info-Eq model to other information-theoretic network models, and how it is derived from a pure maximum entropy method (MEM) algorithm, which we also introduce. In Info-Eq, a damping rate allows control over the balance between *infomax* and maximal entropy while learning. Using network simulations, the performance of the network is shown relative to other algorithms. Data from the simulations show that while pure MEM apparently approaches maximal *network* entropy, causing its performance to suffer, Info-Eq can find a good balance between representing the feature space of the problem and the maximal *feature space* entropy. In addition, because it only uses local information, Info-Eq's learning runs 20% more quickly per epoch than related information theoretic algorithms that must gather non-local information.

## 1 Introduction

Information-theoretic models for neural network learning stem from Shannon's *communication theory*, originally applied to implement efficient communication codes subject to the channel capacity, i.e. the maximal mutual information between the input and output of a communication channel. In 1988, Linsker proposed a principle of maximum information preservation, also called the *infomax principle*, of unsupervised learning that formalized the self-organization process of neural network learning to maximize the mutual information between the input and output layers [6]. During the last decade, many information-theoretic neural network studies have used this principle. Most of the studies attempted to maximize the mutual infor-

mation [1,2,5], in order to detect the feature space within input patterns. In other work [4], Deco et al. minimized mutual information to penalize network complexity in order to eliminate overtraining.

Information-theoretic approaches to neural networks require *a priori* knowledge about the probability density function of the network input or output. Previous approaches can be organized into two major categories: the discrete probability approach (Gibbs distribution), for example Bridle’s *softmax* model [2]; and the continuous probability approach (Jacobian transformation), for example Bell and Sejnowski’s Blind Separation and Blind Deconvolution [1]. The Gibbs distribution is frequently applied as the probability density function for its simplicity and localizability. On the other hand, the Jacobian transformation requires the computation of the inverse of a Jacobian matrix. The computation requires the matrix to be an invertible square matrix [7]. The applicability of this approach is limited in that most real world problems are ill-posed, requiring a rectangular matrix and thus unsuited to the Jacobian transformation.

Most of the discrete probability approaches use aggregated mutual information, i.e. the averaged mutual information across input patterns. This approach limits the network training to per-epoch updating. Moreover, due to the need to average mutual information across the input patterns, the feedforward process must complete two passes of computation before the backward processing [2]. This extra pass increases the complexity of the algorithm relative to error backpropagation. Clearly, reducing this source of complexity is desirable. Also, requiring only local information would allow a per-pattern updating algorithm [4].

## 2 Previous information theoretic approaches

### 2.1 Softmax model

In order to create a network whose outputs represent the conditional probability  $P[\Phi_j|\xi_a]$  the softmax model [2] applies a normalized exponential transformation, the Gibbs distribution:

$$y_j = P[\Phi_j|\xi_a] = \frac{e^{\Phi_j}}{\sum_k e^{\Phi_k}} \tag{1}$$

where  $\Phi_j$  is the un-normalized activation at  $j$ -th unit, and  $\xi_a$  is the  $a$ -th input pattern. If the output units’ activations are all positive, as is the case with a standard logistic output function, a simpler, normalizing version is often used [4,5]:

$$y_j = P[\Phi_j|\xi_a] = \frac{\Phi_j}{\sum_k \Phi_k} = \frac{\Phi_j}{S} \tag{2}$$

We will use  $S$  to represent the normalizing denominator  $\sum_k \Phi_k$ .

### 2.2 Mutual information

Using aggregated mutual information, Bridle et al. [3] derive an average mutual information across the input patterns of the training set as:

$$I(X, Y) = H(Y) - H(X|Y)$$

$$\begin{aligned}
&= -\sum_{j=1}^k \bar{y}_j \log \bar{y}_j + \frac{1}{N_a} \sum_{a=1}^{N_a} \sum_{j=1}^k y_j \log y_j \\
&= H(\bar{y}_j) - \overline{H(y_j)}
\end{aligned} \tag{3}$$

where  $I(X, Y)$  is the mutual information between the network input,  $X$ , and output,  $Y$ .  $H(Y)$  is the output entropy, and  $H(X|Y)$  is the conditional entropy. In the second line,  $y_j$  is the conditional probability  $P[\Phi_j|\xi_a]$  of the activation of unit  $j$  given the  $a$ -th training pattern  $\xi_a$  (cf Equation 2),  $\bar{y}_j$  is the average  $y_j$  across the training patterns,  $k$  is the number of units at the hidden layer, and  $N_a$  is the number of training patterns. Using softmax, it is not necessary to model the output entropy  $H(Y)$  and conditional entropy  $H(Y|X)$  (i.e.  $H(y_j)$ ). For the mutual information,  $I(X, Y)$ , only the conditional probability  $y_j$  is needed, and that is available locally at the hidden layer.

In a supervised information-theoretic network, the mutual information acts to regulate the error correction training rule of backpropagation as follows:

$$\Delta w = -\eta \left( \frac{\partial E}{\partial w} + \lambda \frac{\partial I(X, Y)}{\partial w} \right) \tag{4}$$

where the error correction term is adjusted by the mutual information using the information rate parameter,  $\lambda$ .

The rationale for maximizing mutual information is that it preserves mutual information, thus capturing the structure in the input patterns. However, it may not generalize if the sampling of input space is biased or corrupted by noise. This is one problem cited by Deco et al. [4], whose algorithm minimizes mutual information.

### 3 The maximum entropy method (MEM)

For the purpose of generalization, maximizing entropy may be more useful than minimizing mutual information. The Maximum Entropy Method (MEM) has been developed in order to address ill-posed problems due to sampling error and noise in the input space, which often cause overtraining in artificial neural networks. A maximum entropy model chooses from the set of actual solutions that one with the maximum entropy [8]. Therefore, we propose a supervised information-theoretic algorithm based on the maximum entropy method using the following cost function:

$$G = \frac{1}{2} \sum_t (T_t - O_t)^2 - \lambda \left[ -\sum_k y_k \log(y_k) \right] \tag{5}$$

The first term on the right hand side is the sum of the squared error between the unit target value  $T_t$  and network output  $O_t$  at each of the  $t$  outputs. The second term optimizes the network entropy at the  $k$  unit hidden layer. The information rate parameter,  $\lambda$ , controls the relaxation of the error correction during the course of training.

### 3.1 A per-pattern learning algorithm based on MEM

The training rule for the connection  $w_{tj}$  between the  $j$ -th hidden unit and the  $t$ -th output unit can be derived from Equation 5 as:

$$\Delta w_{tj} = \eta(T_t - O_t)\dot{\sigma}_t y_j \quad (6)$$

where  $\eta$  is the learning rate,  $\dot{\sigma}_t$  is the derivative of the logistic function at the  $t$ -th output unit, and  $y_j$  is the normalized activation value at the  $j$ -th hidden unit.

The training rule for the weight  $w_{ji}$  between the  $i$ -th input unit and the  $j$ -th hidden unit can be derived as:

$$\begin{aligned} \Delta w_{ji} = & \eta \xi_i \frac{\Phi'_j}{S} \left[ \sum_t (T_t - O_t) \dot{\sigma}(w_{tj} - O_t) \right] \\ & - \eta \lambda \xi_i \frac{\Phi'_j}{S} \left[ \log(y_j) - \sum_k y_k \log(y_k) \right] \end{aligned} \quad (7)$$

where  $\Phi'_j$  is the derivative of the logistic function at the  $j$ -th hidden layer unit.

The effects of the training rule are explicit in that the first term on the right hand side is similar to backpropagation except that the weight term,  $w_{tj}$ , is modified by the network output  $O_t$ . The second term on the right hand side of Equation 7 is the entropy regulator used to introduce “randomness” in the network. This is done by adjusting the weights in proportion to the difference between the individual entropy of unit  $j$ ,  $-\log(y_j)$ , and the averaged network entropy,  $-\sum_k y_k \log(y_k)$ , of the  $k$  units in the hidden layer, given the input pattern  $\xi_i$ . The second term dominates as the network training proceeds, when the rate of error correction significantly decreases.

If the information rate,  $\lambda$ , is negative, the network tends to maximize the mutual information between the network input and output. Therefore, Equation 7 can have one of two distinct effects: maximizing entropy or maximizing mutual information, depending on whether  $\lambda$  is positive or negative.

Using the terminology of Bridle et al. [3], MEM allows the network to relax the *firmness* (decisiveness) of the feature space in order to introduce the *fairness* (randomness) of the feature space. As a result, the trained neural network is capable of capturing the general structure of the sample population while avoiding overfitting to the bias and noise of the training set. However, without any limitation, MEM learning may approach the maximum network entropy instead of the maximum entropy of the feature space. Consequently, the network may become over-generalized and suffer decreased accuracy.

Figure 1 shows the test set performance of MEM using a New York State retail establishment data set (unpublished). The data is obtained from the census data set of New York State in 1992, which contains the retail establishments data of cities with population less than 20,000. The network learns the (normalized) number of retail establishments (E) in a city based on the population (P), the proportion of the workforce in a place who work outside that place (O) and the interaction between P and O, O\*P. Therefore, the network has three normalized inputs (i.e. P, O, and O\*P) and one output E. Three hidden units are used in this network learning.

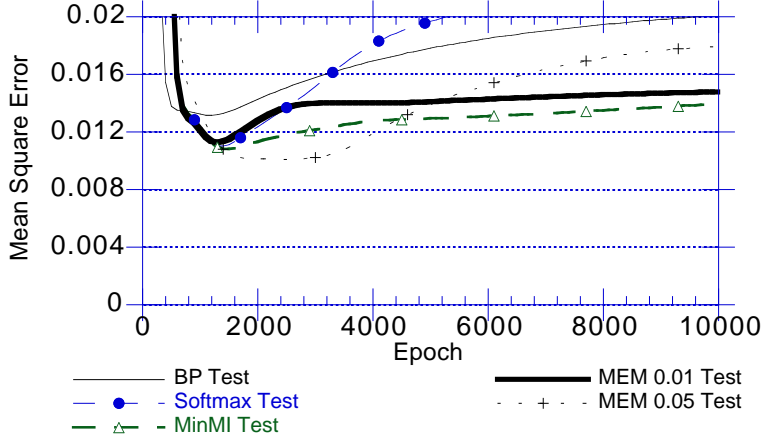


Figure 1: Test Set Error Rates for MEM using  $\lambda = 0.01$  and  $\lambda = 0.05$ . For comparison, results on the same task using backpropagation (BP), softmax and minimizing mutual information (MinMI) are given.

The 150 city data used here are randomly separated into 30 test patterns and 120 training patterns.

In Figure 1, the results of test performance between backpropagation, softmax, Deco’s minimizing mutual information (MinMI) algorithm [4] and MEM are compared. Backpropagation and softmax show overtraining starting from around epoch 1200. Although MEM, both with  $\lambda = 0.01$  and  $\lambda = 0.05$  reduces overtraining, it is not complete. The networks’ performance start to degrade around epochs 1400 and 3200, respectively. MinMI gives results comparable to MEM. These results imply that an equilibrium effect should be added to MEM to balance the entropy so that it is maximized enough to avoid overtraining but also keeps the entropy from approaching the maximal network entropy, thus becoming over-generalized.

#### 4 The Info-Equilibrium model (Info-Eq)

Since the distribution of the feature space in input patterns is not usually random in real world problems, the entropy of the feature space is not necessarily the maximum network entropy,  $-\log(1/k)$ , unless the density function of the feature space is uniform. Therefore, the optimal network entropy should be the one that effectively eliminates overtraining while not becoming over-generalized.

Here, we propose an “Info-Equilibrium” (Info-Eq) model that pushes the network entropy to a stable state by adding damping to the second term in Equation 7:

$$\Delta w_{ji} = \eta \xi_i \frac{\Phi'_j}{S} \left[ \sum_t (T_t - O_t) \dot{\sigma}(W_{tj} - O_t) \right] - \eta \lambda \xi_i \frac{\Phi'_j}{S} \left\{ \left[ \log(y_j) - \sum_k y_k \log(y_k) \right] - \gamma \left[ \log(y_j) + \log\left(\frac{1}{k}\right) \right] \right\} \quad (8)$$

This added term, controlled by a damping rate  $\gamma$ , is a mechanism to avoid MEM’s

tendency to over-generalize: MEM’s entropy eventually approaches the maximal network entropy rather than the maximal entropy of the problem feature space.

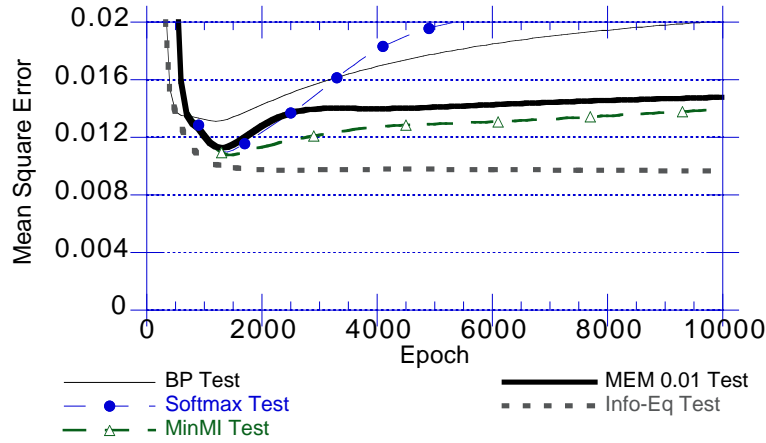


Figure 2: Test Set Error Rates for Info-Eq. For comparison, results on the same task using backpropagation (BP), softmax, MinMI and MEM are given.

Figure 2 shows that the test error of Info-Eq is lower than that achieved by backpropagation, softmax, MinMI or MEM. The figure also shows that Info-Eq’s better test set performance is not subsequently degraded by over-generalization. In addition, Info-Eq runs approximately 20% faster per epoch than the related algorithms that collect non-local information in an additional forward pass (data not shown).

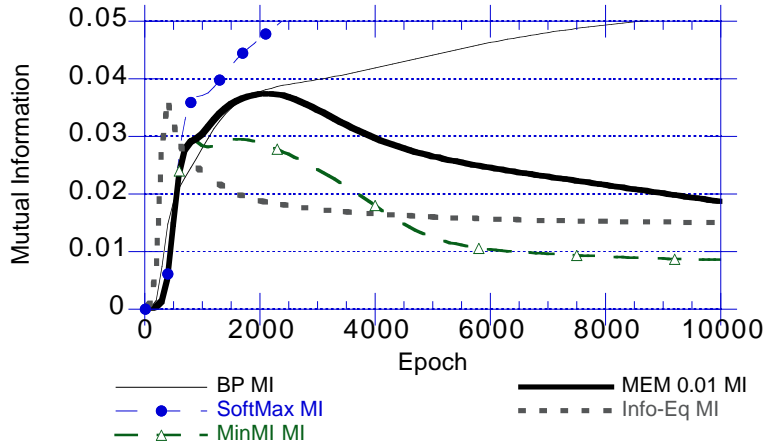


Figure 3: The Mutual Information of Info-Eq, MEM, MinMI, softmax and backpropagation. These results are from the runs shown in Figure 2.

Figure 3 shows the mutual information,  $I(X, Y)$  as the networks train. Backpropagation and softmax show mutual information increasing throughout training. MEM and MinMI both begin to reduce mutual information only after their networks show

overtraining (cf Figure 2). Info-Eq, using  $\lambda = 0.01$  and  $\gamma = 1$ , successfully damps the mutual information to an equilibrium state early in the training.

We tested these methods on another task, predicting the geographic distribution of students attending a state university. Info-Eq also shows stable generalization at a lower error level than the other algorithms on this task (data not shown).

## 5 Conclusion

We have introduced Info-Eq – a damped maximum entropy model of learning in artificial neural networks, and indicated that its stabilized mutual information can yield networks with better generalization properties. While infomax, MinMI and pure MEM have desirable properties, in order to get good, stable generalization it appears necessary to strike a balance between maximizing mutual information and entropy. With its adjustable damping, Info-Eq appears to do just that in our results to date.

We are currently working to apply Info-Eq to larger scale problems to further test its generalization properties empirically. We are also examining the theoretical underpinnings of the Info-Eq algorithm, in particular, how the damping term can be motivated, as well as exactly how it stabilizes the network's learning.

## Acknowledgments

The authors would like to thank Dr. G. Deco for his kindly and helpful assistance, and Thomas O'Connell for his valuable discussions about this work.

## References

- [1] Bell, A. & Sejnowski, T.J. (1995) "An information-maximization approach to blind separation and blind deconvolution." *Neural Computation*, 7(6):1129–1159.
- [2] Bridle, J.S. (1990) "Training stochastic model recognition algorithms as networks can lead to maximum mutual information estimation of parameters." In D.S. Touretzky (ed.), *Advances in Neural Information Processing Systems 2*, pp. 211–217. Cambridge, MA: MIT Press.
- [3] Bridle, J.S., Heading, A.J.R. & MacKay, D.J.C. (1992) "Unsupervised classifiers, mutual information and 'Phantom Targets'." In J.E. Moody, S.J. Hanson and R.P. Lippmann (eds.), *Advances in Neural Information Processing Systems 4*, pp. 1096–1101. Cambridge, MA: MIT Press.
- [4] Deco, G., Finnoff, W. & Zimmermann, H.G. (1995) "Unsupervised Mutual Information criterion for Elimination of Overstraining in Supervised Multilayer Network." *Neural Computation*, 7(1):86–107.
- [5] Kamimura, R. & Nakanishi, S. (1995) "Hidden information maximization for feature detection and rule discovery." *Network: Computation in Neural Systems*, 6:577–602.
- [6] Linsker, R. (1988). Self-organization in a Perceptual Network. *IEEE Computer*, 21(3):105–117.
- [7] Linsker, R. (1997). "A local learning rule that enables information maximization for arbitrary input distributions." *Neural Computation*, 9(8):1661-1665.
- [8] Wu N. (1997) *The Maximum Entropy Method*. New York: Springer.