IDIAP RESEARCH REPORT

Variational Information Maximization for Population Coding

Felix Agakov [a] David Barber [b]

IDIAP–RR 04-85

[a] University of Edinburgh, EH12QL, Edinburgh, UK
[b] IDIAP, Martigny, Switzerland

IDIAP Research Report 04-85

Variational Information Maximization for Population Coding

Felix Agakov David Barber

Abstract. The goal of neural processing assemblies is varied, and in many cases still rather unclear. However, a possibly reasonable *subgoal* is that sensory information may be encoded efficiently in a population of neurons. In this context, Mutual Information is a long studied measure of coding efficiency, and many attempts to apply this to *population coding* have been made. However, this is a numerically intractable task, and most previous studies redefine the criterion in forms of an approximation to Mutual Information, the Fisher Information being one such well-known approach. Here we describe a principled bound maximisation procedure for Mutual Information learning of population codes in a simple point neural model, and compare it with other approaches.

1 Introduction

Whilst the use of information theory cannot be justified as a central goal of neural processing – since arguably it is compression and decisions that are key to survival – nevertheless, much effort has been focussed on a possible *subgoal*, namely that of maximal information transmission from a stimulus to its representation as a set of spiking neurons[16, 3, 14].

The problem of encoding real-valued stimuli x by a population of neural spikes y may be addressed in many different ways. Essentially the framework is to, for a given set of patterns x, to adapt the parameters of any mapping $p(\mathsf{y}|\mathsf{x})$ to make a desirable population code. There are many possible desiderate.

One could be that any reconstruction based on the population should be accurate. This is typically handled by appealing to the Fisher Information which, with care, can be used bound mean square reconstruction error. Another approach is to bound the *probability* of a correct reconstruction (to the best of our knowledge, surprisingly, this has not yet been studied). The approach we examine here, is to consider maximizing the amount of information which the spiking patterns contain about the stimuli [6, 13]. The fundamental information theoretic measure in this context is the mutual information

$$I(\mathsf{x}, \mathsf{y}) \equiv H(\mathsf{x}) - H(\mathsf{x}|\mathsf{y}), \tag{1}$$

which indicates the decrease of uncertainty in x due to the knowledge of y. Here $H(\mathsf{x}) \equiv -\langle \log p(\mathsf{x}) \rangle_{p(\mathsf{x})}$ and $H(\mathsf{x}|\mathsf{y}) \equiv -\langle \log p(\mathsf{x}|\mathsf{y}) \rangle_{p(\mathsf{x},\mathsf{y})}$ are marginal and conditional entropies respectively, and the angled brackets represent averages over all variables contained within the brackets.

It is sometimes argued that MI is desirable over and above other approaches which require the explicit definition of a reconstruction procedure[16]. However, we will argue that the specification of a reconstruction/decoding procedure is essentially unavoidable.

The principled information theoretic approach to learning neural codes involves maximization of the objective (1) with respect to parameters of the encoder $p(\mathsf{y}|\mathsf{x})$. However, it is easy to see that in large-scale systems exact evaluation of $I(\mathsf{x}, \mathsf{y})$ is in general computationally intractable. The key difficulty lies in the computation of the conditional entropy $H(\mathsf{x}|\mathsf{y})$ for the posterior distribution $p(\mathsf{x}|\mathsf{y})$, which is tractable only in a few special cases. Standard techniques address the problem of optimizing (1) by assuming that $p(\mathsf{x}, \mathsf{y})$ is jointly Gaussian [11], the output spaces are very low-dimensional [12], or the channels are deterministic and invertible [4]. Other popular methods suggest alternative objective functions (e.g. approximations based on the *Fisher Information* criterion [6]), which, however, do not retain proper bounds on $I(\mathsf{x}, \mathsf{y})$.

Recently we described a simple variational approach to information maximization which optimizes a proper lower bound on the mutual information [2]. In this paper we investigate applicability of the method in the context of neural coding. First, we briefly review the lower bound on $I(\mathsf{x}, \mathsf{y})$. Then we analyze the learning rule obtained by maximizing the bound for sigmoidal networks. Finally, we analyze the relation between our approach and standard techniques for approximate information maximization, focusing specifically on a comparison with the Fisher Information criterion.

1.1 Variational Lower Bound on Mutual Information

A simple lower bound on the mutual information $I(\mathsf{x}, \mathsf{y})$ follows from non-negativity of the Kullback-Leibler divergence $KL(p(\mathsf{x}|\mathsf{y})||q(\mathsf{x}|\mathsf{y}))$ between the exact posterior $p(\mathsf{x}|\mathsf{y})$ and its variational approximation $q(\mathsf{x}|\mathsf{y})$. Clearly,

$$\langle \log p(\mathsf{x}|\mathsf{y}) \rangle_{p(\mathsf{x}|\mathsf{y})} - \langle \log q(\mathsf{x}|\mathsf{y}) \rangle_{p(\mathsf{x}|\mathsf{y})} \geq 0 \Rightarrow \langle \log p(\mathsf{x}|\mathsf{y}) \rangle_{p(\mathsf{x}|\mathsf{y})p(\mathsf{y})} \geq \langle \log q(\mathsf{x}|\mathsf{y}) \rangle_{p(\mathsf{x}|\mathsf{y})p(\mathsf{y})}. \tag{2}$$

This leads to the variational lower bound

$$I(\mathsf{x}, \mathsf{y}) \geq \tilde{I}(\mathsf{x}, \mathsf{y}) \stackrel{\text{def}}{=} H(\mathsf{x}) + \langle \log q(\mathsf{x}|\mathsf{y}) \rangle_{p(\mathsf{x},\mathsf{y})}, \tag{3}$$

where $q(\mathsf{x}|\mathsf{y})$ is an arbitrary variational distribution saturating the bound for $q(\mathsf{x}|\mathsf{y}) \equiv p(\mathsf{x}|\mathsf{y})$. Note that the objective (3) explicitly includes[1] both the encoder $p(\mathsf{y}|\mathsf{x})$ (distribution of neural spikes for a given stimulus) and decoder $q(\mathsf{x}|\mathsf{y})$ (reconstruction of the stimulus from a population of neural firings). It is possible to consider other lower bounds on the mutual information [9]. However, the flexibility of the choice of the decoder $q(\mathsf{x}|\mathsf{y})$ makes the bound (3) particularly computationally convenient. Moreover, (3) corresponds to a moment matching approximation of $p(\mathsf{x}|\mathsf{y})$ by $q(\mathsf{x}|\mathsf{y})$, which is particularly beneficial in terms of decoding [15, 2].

2 Variational Learning of Population Codes

To learn optimal stochastic representations of the continuous training patterns $\mathsf{x}_1, \ldots, \mathsf{x}_M$ according to the lower bound $\tilde{I}(\mathsf{x}, \mathsf{y})$, we need to choose a continuous density function for the decoder $q(\mathsf{x}|\mathsf{y})$. Computationally, it is convenient to assume that the decoder is given by the isotropic Gaussian $q(\mathsf{x}|\mathsf{y}) \sim \mathcal{N}(\mathsf{U}\mathsf{y}, \sigma^2\mathsf{I})$, where $\mathsf{U} \in \mathbb{R}^{|\mathsf{x}| \times |\mathsf{y}|}$. In this case, exact evaluation of the bound $\tilde{I}(\mathsf{x}, \mathsf{y})$ is straightforward, since it only involves computations of the second-order moments of y over the factorized distribution. Of course, other (correlated and non-linear) choices may be considered – however, for clarity we limit the discussion to isotropic linear Gaussian mappings. Effectively, this choice indicates that small changes in the post-synaptic firings do not significantly vary our guesses about the generating stimuli.

For the empirical distribution $p(\mathsf{x}) = \sum_{m=1}^{M} \delta(\mathsf{x} - \mathsf{x}_m)/M$ the bound (3) is then given by

$$\tilde{I}(\mathsf{x}, \mathsf{y}) \propto \sum_{m=1}^{M} \operatorname{tr}\left\{ \mathsf{U}\langle \mathsf{y} \rangle_{p(\mathsf{y}|\mathsf{x}_m)} \mathsf{x}_m^T - \frac{1}{2} \mathsf{U}^T \mathsf{U} \langle \mathsf{y}\mathsf{y}^T \rangle_{p(\mathsf{y}|\mathsf{x}_m)} \right\} + c \tag{4}$$

where c is a constant. Expressing $\tilde{I}(\mathsf{x}, \mathsf{y})$ as a function of the encoder $p(\mathsf{y}|\mathsf{x})$ alone, we get

$$\mathsf{U} = \langle \mathsf{x}\mathsf{y}^T \rangle \langle \mathsf{y}\mathsf{y}^T \rangle^{-1}, \quad \tilde{I}(\mathsf{x}, \mathsf{y}) \propto \operatorname{tr}\left\{ \langle \mathsf{x}\mathsf{y}^T \rangle \langle \mathsf{y}\mathsf{y}^T \rangle^{-1} \langle \mathsf{y}\mathsf{x}^T \rangle \right\} + c. \tag{5}$$

Note that the objective (5) is a proper bound for any choice of the stochastic mapping $p(\mathsf{y}|\mathsf{x})$. We may therefore[2] use it for optimizing a variety of channels with continuous source vectors.

Whilst making an objective without the the explicit appearance of the decoding weights U is theoretically attractive, the resulting learning rules for the encoding weights will become biologically unrealistic. However, the reader should bear in mind that in the sequel, local rules can be formed by incremental optimisation of (4) jointly with respect to the encoding and decoding weights.

2.1 Sigmoidal Activations

Here we consider the case of high-dimensional continuous patterns $\mathsf{x} \in \mathbb{R}^{|\mathsf{x}|}$ represented by stochastic firings of the post-synaptic neurons $\mathsf{y} \in \{-1, +1\}^{|\mathsf{y}|}$. For each neuron y_i we assume the logistic parameterization of the encoder $p(y_i|\mathsf{x})$, so that the probability of firing monotonically increases with an increase in the membrane potential. For conditionally independent activations, we obtain

$$p(\mathsf{y}|\mathsf{x}) = \prod_{i=1}^{|\mathsf{y}|} p(y_i|\mathsf{x}) \stackrel{\text{def}}{=} \prod_{i=1}^{|\mathsf{y}|} \sigma(y_i(\mathsf{w}_i^T \mathsf{x} + b_i)) \tag{6}$$

where $\mathsf{w}_i \in \mathbb{R}^{|\mathsf{x}|}$ is a vector of the synaptic weights for neuron y_i, b_i is the corresponding threshold, and $\sigma(a) \stackrel{\text{def}}{=} 1/(1 + e^{-a})$. If the decoder's weights U are unconstrained, we may optimize the bound

[1]The bound (3) corresponds to the criteria used by Blahut-Arimoto algorithms (e.g. [7]); however, we optimize it for both encoder and decoder, which is constrained to lie in a tractable family.

[2]From (5) it is clear that if $\langle \mathsf{y}\mathsf{y}^T \rangle$ is near-singular, the varying part of the objective $\tilde{I}(\mathsf{x}, \mathsf{y})$ may be infinitely large. However, if the mapping $\mathsf{x} \mapsto \mathsf{y}$ is probabilistic and the number of training stimuli M exceeds the dimensionality of the neural codes $|\mathsf{y}|$, the optimized criterion is typically positive and finite.

(5), which in this case is given by

$$\tilde{I}(\mathsf{x},\mathsf{y}) \propto \mathrm{tr}\left\{\langle \mathsf{x}\boldsymbol{\lambda}_x^T\rangle\langle \boldsymbol{\lambda}_x\boldsymbol{\lambda}_x^T - \mathsf{D}_{\lambda_x} + \mathsf{I}\rangle^{-1}\langle \boldsymbol{\lambda}_x \mathsf{x}^T\rangle\right\}. \tag{7}$$

Here $\boldsymbol{\lambda}_x \in [-1,1]^{|\mathsf{y}|}$ is a vector whose elements $\lambda_i(\mathsf{x}) \stackrel{\text{def}}{=} \langle y_i\rangle_{p(y_i|\mathsf{x})} = 2\sigma(\mathsf{w}_i^T\mathsf{x} + b_i) - 1$ correspond to expected firings of the i^{th} unit for a fixed stimulus x, and

$$\mathsf{D}_{\lambda_x} \stackrel{\text{def}}{=} \mathrm{diag}\left(\lambda_1^2(\mathsf{x}),\dots,\lambda_{|\mathsf{y}|}^2(\mathsf{x})\right) = \mathsf{I} - \mathrm{cov}(\mathsf{y}|\mathsf{x}) \in [0,1]^{|\mathsf{y}|\times|\mathsf{y}|} \tag{8}$$

is a measure of consistency of neural firings.

Since the lower bound (7) depends only on the thresholds and synaptic weights, the learning rule is easily obtained by differentiating (7) with respect to $\mathsf{b} \in \mathbb{R}^{|\mathsf{y}|}$ and $\mathsf{W} \in \mathbb{R}^{|\mathsf{y}|\times|\mathsf{x}|}$ (where rows of W are given by $\mathsf{w}_i^T \in \mathbb{R}^{1\times|\mathsf{x}|}$). This leads to

$$\Delta\mathsf{W} \propto \sum_{m=1}^{M}\left(\mathsf{I} - \mathsf{D}_{\lambda_{x_m}}\right)\left(\tilde{\mathsf{D}}\boldsymbol{\lambda}_{x_m} + \boldsymbol{\Sigma}_{yy}^{-1}\boldsymbol{\Sigma}_{yx}\left(\mathsf{x}_m - \boldsymbol{\Sigma}_{xy}\boldsymbol{\Sigma}_{yy}^{-1}\boldsymbol{\lambda}_{x_m}\right)\right)\mathsf{x}_m^T, \tag{9}$$

where $\boldsymbol{\Sigma}_{yy} \stackrel{\text{def}}{=} \langle \mathsf{y}\mathsf{y}^T\rangle$, $\boldsymbol{\Sigma}_{yx} \equiv \boldsymbol{\Sigma}_{xy}^T \stackrel{\text{def}}{=} \langle \mathsf{y}\mathsf{x}^T\rangle$ are the second-order moments, and $\tilde{\mathsf{D}}$ corresponds to the diagonal of $\boldsymbol{\Sigma}_{yy}^{-1}\boldsymbol{\Sigma}_{yx}\left(\boldsymbol{\Sigma}_{yy}^{-1}\boldsymbol{\Sigma}_{yx}\right)^T$. The update for the threshold $\Delta\mathsf{b}$ has the same form as (9) without the post-multiplication of each term by the training stimulus x_m^T.

From (9) it is clear that the magnitude of each weight update $\Delta\mathsf{w}_i \in \mathbb{R}^{|\mathsf{x}|}$ decreases with a decrease in the corresponding conditional variance $\mathrm{var}(y_i|\mathsf{x}_m)$. Effectively, this corresponds to a variable learning rate – as training continues and magnitudes of the synaptic weights increase, the firings become more deterministic, and learning slows down.

Although this update rule is biologically difficult to interpret, as we stated before, interpretable rules are available by optimising (4) with respect to both U and W.

3 Fisher Information and Mutual Information

Here we briefly review a popular class of approximations to the mutual information based on the Fisher Information criterion [6]. We also outline the corresponding learning rule for sigmoidal channels.

3.1 Fisher Approximation

Let $\hat{\mathsf{x}} \in \mathbb{R}^{|\mathsf{x}|}$ be a statistical estimator of the input stimulus x obtained from the stochastic neural firings y. It is easy to see that $\mathsf{x} \to \mathsf{y} \mapsto \hat{\mathsf{x}}$ forms a Markov chain with $p(\hat{\mathsf{x}}|\mathsf{y}) \sim \delta(\hat{\mathsf{x}} - \hat{\mathsf{x}}(\mathsf{y}))$. If $\hat{\mathsf{x}}$ is *efficient*, its covariance saturates the Cramer-Rao bound (see e.g. [7]), which results in an upper bound on the entropy of the conditional distribution $H\left(p(\hat{\mathsf{x}}|\mathsf{x})\right)$. From the data processing inequality, one may obtain a lower bound on the mutual information

$$I(\mathsf{x},\mathsf{y}) \geq H(\hat{\mathsf{x}}) + \frac{1}{2}\langle \log|\mathsf{F}_\mathsf{x}|\rangle_{p(\mathsf{x})} + c, \tag{10}$$

where $\mathsf{F}_\mathsf{x} = \{F_{ij}(\mathsf{x})\} \stackrel{\text{def}}{=} -\langle \partial^2 \log p(\mathsf{y}|\mathsf{x})/\partial x_i \partial x_j\rangle_{p(\mathsf{y}|\mathsf{x})}$ is the Fisher Information matrix and c is an irrelevant constant. Despite the fact that the mapping $\mathsf{y} \mapsto \hat{\mathsf{x}}$ is deterministic, exact computation of the entropy of statistical estimates $H(\hat{\mathsf{x}})$ in the objective (10) is in general computationally intractable. [6] show that under some assumptions $H(\hat{\mathsf{x}}) \approx H(\mathsf{x})$, leading to the approximation

$$I(\mathsf{x},\mathsf{y}) \gtrsim \tilde{I}_F(\mathsf{x},\mathsf{y}) \stackrel{\text{def}}{=} H(\mathsf{x}) + \frac{1}{2}\langle \log|\mathsf{F}_\mathsf{x}|\rangle_{p(\mathsf{x})} + c. \tag{11}$$

Note that since $H(\mathsf{x})$ is independent of the parameters of $p(\mathsf{y}|\mathsf{x})$, maximization of (11) is equivalent to maximization of (10) where the intractable entropic term is ignored.

3.2 Sigmoidal Activations

For sigmoidal activations (6), the criterion (11) is given by

$$\tilde{I}_F(\mathsf{x},\mathsf{y}) \propto \sum_{m=1}^{M} \log \left|\mathsf{W}^T \left(\mathsf{I} - \mathsf{D}_{\lambda_{x_m}}\right) \mathsf{W}\right| + c. \tag{12}$$

Again, $\mathsf{W} \in \mathbb{R}^{|\mathsf{y}|\times|\mathsf{x}|}$ is the matrix of synaptic weights, $\mathsf{I} - \mathsf{D}_{\lambda_{x_m}}$ is the conditional covariance of the stochastic spikes (see (8)), and c corresponds to the remaining terms which do not affect the optimization.

It is interesting to note that for the square model with $|\mathsf{x}| = |\mathsf{y}|$, optimization of (12) leads to

$$\Delta\mathsf{W} = 2\mathsf{W}^{-T} - \langle \boldsymbol{\lambda}_x \mathsf{x}^T \rangle, \tag{13}$$

which (apart from the coefficient at the inverse weight – *redundancy* term) has the same form as the learning rule of [4] derived for the invertible channel with $p(\mathsf{y}|\mathsf{x}) \sim \delta(\mathsf{y} - \sigma(\mathsf{x}))$. If $|\mathsf{x}| < |\mathsf{y}|$ and the conditional variances of the responses are invariant under the training stimuli, the redundancy term in (13) is replaced by the transposed pseudo-inverse $\mathsf{W}(\mathsf{W}^T\mathsf{W})^{-1} \in \mathbb{R}^{|\mathsf{y}|\times|\mathsf{x}|}$. Notably, the weight update (13) has no Hebbian terms.

More importantly, from (12) it is clear that as the variance of the stochastic firings decreases, the objective $\tilde{I}_F(\mathsf{x},\mathsf{y})$ may become infinitely loose. Since directions of low variation swamp the volume of the manifold, neural spikes generated by a fixed stimulus may often be inconsistent. It is also clear that optimization of the Fisher information-based objective (12) is limited to the cases when $\mathsf{W}^T\mathsf{W} \in \mathbb{R}^{|\mathsf{x}|\times|\mathsf{x}|}$ is full-rank, which complicates applicability of the method for a variety of tasks involving relatively low-dimensional encodings of very high-dimensional stimuli.

4 Variational Lower Bound *vs.* Fisher Approximation

Since $\tilde{I}_F(\mathsf{x},\mathsf{y})$ is in general not a proper lower bound on the mutual information, it is difficult to analyze its tightness or compare it with the variational bound (3). To illustrate a relation between the approaches, we may consider a Gaussian decoder $q(\mathsf{x}|\mathsf{y}) \sim \mathcal{N}_{\mathsf{x}}(\boldsymbol{\mu}_{\mathsf{y}}; \boldsymbol{\Sigma})$, which transforms the variational bound into

$$\tilde{I}(\mathsf{x},\mathsf{y}) = -\frac{1}{2}\left\langle \mathrm{tr}\left\{\boldsymbol{\Sigma}^{-1}(\mathsf{x}-\boldsymbol{\mu}_y)(\mathsf{x}-\boldsymbol{\mu}_y)^T\right\}\right\rangle_{p(\mathsf{x},\mathsf{y})} + \frac{1}{2}\log|\boldsymbol{\Sigma}^{-1}| + \acute{c}. \tag{14}$$

Here $\acute{c}$ incorporates $H(\mathsf{x})$ and other irrelevant constants, and $\boldsymbol{\Sigma} \in \mathbb{R}^{|\mathsf{x}|\times|\mathsf{x}|}$ is a function of parameters of the conditional $p(\mathsf{y}|\mathsf{x})$. Clearly, if the log eigenspectrum of the inverse covariance of the decoder is constrained to satisfy

$$\sum_{i=1}^{|\mathsf{x}|} \log l_i(\boldsymbol{\Sigma}^{-1}) = \sum_{i=1}^{|\mathsf{x}|} \langle \log l_i(\mathsf{F}_{\mathsf{x}})\rangle_{p(\mathsf{x})}, \tag{15}$$

where $\{l_i(\boldsymbol{\Sigma}^{-1})\}$ and $\{l_i(\mathsf{F}_{\mathsf{x}})\}$ are eigenvalues of $\boldsymbol{\Sigma}^{-1}$ and F_{x} respectively, then the lower bound (14) reduces to the objective (11) amended with the average quadratic reconstruction error

$$\tilde{I}(\mathsf{x},\mathsf{y}) = -\frac{1}{2}\underbrace{\left\langle \mathrm{tr}\left\{\boldsymbol{\Sigma}^{-1}(\mathsf{x}-\boldsymbol{\mu}_y)(\mathsf{x}-\boldsymbol{\mu}_y)^T\right\}\right\rangle_{p(\mathsf{x},\mathsf{y})}}_{\text{reconstruction error}} + \frac{1}{2}\underbrace{\langle \log|\mathsf{F}_{\mathsf{x}}|\rangle_{p(\mathsf{x})}}_{\text{Fisher criterion}} + \acute{c}. \tag{16}$$

Arguably, it is due to the subtraction of the non-negative quadratic term that (14) remains a general lower bound independently of the parameterization of the model and spectral properties of F_{x}.

Another principal advantage of the variational approach to information maximization is the flexibility in the choice of the decoder. Clearly, if the Fisher Information matrices are small or nearly

Table 1: Objective functions for approximate information maximization

1.	*Invertible channels*:	$I(\mathsf{x},\mathsf{y}) = \langle \log \lvert \mathsf{J}_\mathsf{x} \rvert \rangle_{p(\mathsf{x})}$
2.	*Fisher approximation*:	$\tilde{I}_F(\mathsf{x},\mathsf{y}) = \langle \log \lvert \mathsf{F}_\mathsf{x} \rvert \rangle_{p(\mathsf{x})}$
3.	*Variational lower bound*:	$\tilde{I}(\mathsf{x},\mathsf{y}) = \langle \log q(\mathsf{x}\lvert\mathsf{y}) \rangle_{p(\mathsf{x},\mathsf{y})}$

singular, both (11) and (16) are quite weak. However, by relaxing (15) and choosing other non-singular covariances (or completely different decoder types), the variational bound may be significantly strengthened.

Table 1 summarizes effective criteria optimized by [4], [6], and the variational approach of [2]. Here $\mathsf{J}_\mathsf{x} = \{J_{ij}(\mathsf{x})\} \stackrel{\text{def}}{=} \{\partial y_i(\mathsf{x})/\partial x_j\} \in \mathbb{R}^{|\mathsf{x}|\times|\mathsf{x}|}$ is the Jacobian of the deterministic invertible mapping $\mathsf{x} \mapsto \mathsf{y}$ (with $|\mathsf{y}| = |\mathsf{x}|$), $\mathsf{F}_\mathsf{x} \in \mathbb{R}^{|\mathsf{x}|\times|\mathsf{x}|}$ is the Fisher Information matrix, and $q(\mathsf{x}|\mathsf{y})$ is an approximate decoder lying in a tractable family. In contrast to the first two techniques, the suggested variational method optimizes a proper lower bound independently of the choice of the decoder, dimensionality of the input stimuli, number of post-synaptic neurons, or noise of the stochastic firings. This extends applicability of the variational approach to dimensionality reduction, compression, syndrome decoding, and generalizes applications to population coding.

5 Experiments

Variational IM *vs* Fisher criterion

In the first set of experiments we were interested to see if, by maximising our bound $\tilde{I}(\mathsf{x},\mathsf{y})$ on the MI, the parameters found at each iteration indeed increased the true MI. We compared this with how the value of the true MI changed as the parameters were updated by maximising the Fisher criterion $\tilde{I}_F(\mathsf{x},\mathsf{y})$. The dimensionality of the response variables $|\mathsf{y}|$ was set to be sufficiently small, so that the true mutual information $I(\mathsf{x},\mathsf{y})$ could be computed. Figure 1 illustrates changes in $I(\mathsf{x},\mathsf{y})$ with iterations of the variational and Fisher-based learning, where the variational decoder was chosen to be an isotropic linear Gaussian with the optimal weights (5). We found that for $|\mathsf{x}| \leq |\mathsf{y}|$ (Figure 1 (*left*)), both approaches tend to lead to consistent improvements in the true mutual information (however, the variational approach typically resulted in higher values of $I(\mathsf{x},\mathsf{y})$ after just a few iterations). For $|\mathsf{x}| > |\mathsf{y}|$ (Figure 1 (*right*)), optimization of the Fisher criterion was numerically unstable and lead to no visible improvements of $I(\mathsf{x},\mathsf{y})$ over its starting value at initialization. Further approximations aimed at handling singularities of the gradients could lead to slight improvements, though their dynamics was rather inconsistent.

Variational IM: stochastic representations of the digit data

Here we apply the variational learning to stochastic coding and reconstruction of visual patterns. In our experiments we used the simplest form of the linear Gaussian decoder discussed in Section 2.1. After numerical optimization with an explicit constraint on the channel noise, we performed reconstruction of 196-dimensional continuous visual stimuli from 10 spiking neurons. The training stimuli consisted of 30 instances of digits 1 and 8 (15 of each class). The source variables were reconstructed from 50 stochastic spikes at the mean of the optimal approximate decoder $q(\mathsf{x}|\mathsf{y})$. Note that since $|\mathsf{x}| > |\mathsf{y}|$, the problem cannot be efficiently addressed by optimization of the Fisher Information-based criterion (12). Clearly, the approach of [4] is not applicable either, due to its fundamental assumption of invertible mappings between the spikes and the visual stimuli. Figure 2 illustrates a subset of the original source signals, samples of the corresponding binary responses, and reconstructions of the source data.

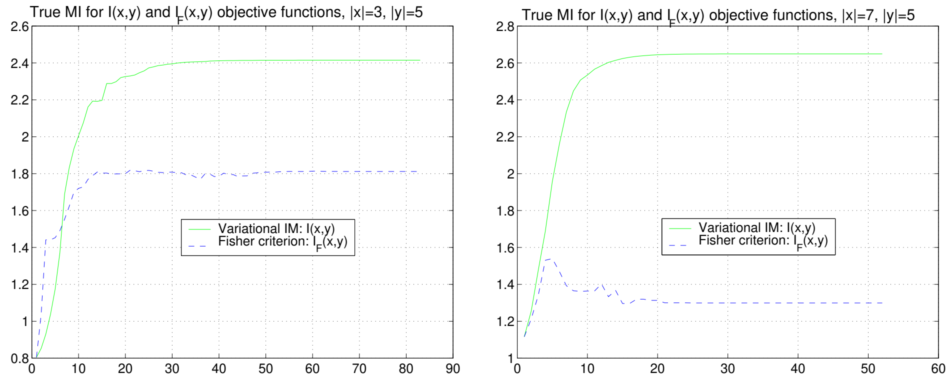

Figure 1: Changes in the exact mutual information $I(\mathsf{x},\mathsf{y})$ for parameters of the coder $p(\mathsf{y}|\mathsf{x})$ obtained by maximizing the variational lower bound and the Fisher information criterion for $M = 20$ training stimuli. *Left*: $|\mathsf{x}| = 3$, $|\mathsf{y}| = 5$ *Right*: $|\mathsf{x}| = 7$, $|\mathsf{y}| = 5$. Generally, using the parameters given at each iteration from the bounding procedure, *does* increase the MI. However, those parameters given by the Fisher criterion, do not generally increase the MI.

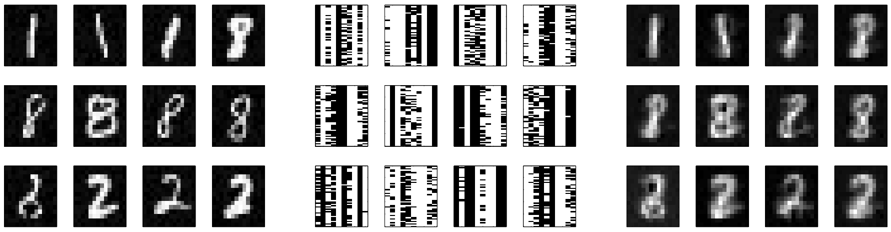

Figure 2: *Left*: a subset of the original visual stimuli. *Middle*: 20 samples of the corresponding spikes generated by each of the 10 neurons. *Right*: Reconstructions from 50 samples of neural spikes. Note that we numerically constrained slightly the weights so as to a solution being reached which would constitute deterministic firings.

6 Discussion

We described a variational approach to information maximization for the case when continuous source stimuli are represented by stochastic binary responses. Our results indicate that other approximate methods for information maximization [12, 6] may be viewed as approximations of our approach, which, however, do not always preserve a proper bound on the mutual information. We do not wish here to discredit generally the use of the Fisher Criterion, since this can be relevant for bounding reconstruction error. However, for the case considered here as a method for maximising information, we believe that our method is more attractive. An obvious extension of our work would be to more realistic spiking networks[1]. However, this case is difficult in the variational information maximisation framework, which cannot easily deal with correlations in the neural firings. Surrogate Mutual Information methods have been applied in this case[8], although we believe that an approach similar to a standard variational treatment of input-output Hidden Markov Models[5, 10], under a maximum probability of reconstruction measure would be relatively straightforward and arguably more desirable surrogate.

Whilst this paper is largely theoretical, we hope to have sustained the reader's interest at least to demonstrate that there is indeed a *principled* way to maximise information transmission to form efficient population codes. Its possible biological interpretation may interest others more expert than ourselves in that arena.

[1] D. Barber. Learning in Spiking Neural Assemblies. *NIPS*, 2002.

[2] D. Barber and F. V. Agakov. The IM Algorithm: A Variational Approach to Information Maximization. In *NIPS*. MIT Press, 2003.

[3] H. Barlow. Unsupervised Learning. *Neural Computation*, 1:295–311, 1989.

[4] A. J. Bell and T. J. Sejnowski. An information-maximization approach to blind separation and blind deconvolution. *Neural Computation*, 7(6):1129–1159, 1995.

[5] Y. Bengio and P. Frasconi. An input-output HMM architecture. *NIPS 7*, 1995.

[6] N. Brunel and J.-P. Nadal. Mutual Information, Fisher Information and Population Coding. *Neural Computation*, 10:1731–1757, 1998.

[7] T. M. Cover and J. A. Thomas. *Elements of Information Theory*. Wiley, New York, 1991.

[8] G. Deco and B. Schürmann. Spatiotemporal Coding in the Cortex: Information Flow-Based Learning in Spiking Neural Networks. *Neural Computation*, 11(4):919–934, 1999.

[9] T. S. Jaakkola and M. I. Jordan. Improving the Mean Field Approximation via the Use of Mixture Distributions. In M. I. Jordan, editor, *Learning in Graphical Models*. Kluwer Academic Publishers, 1998.

[10] M.I. Jordan, Z. Ghahramani, T.S. Jaakkola, and L.K. Saul. An Introduction to Variational Methods for Graphical Models. In M.J. Jordan, editor, *Learning in Graphical Models*, chapter 1. MIT Press, 1998.

[11] R. Linsker. An Application of the Principle of Maximum Information Preservation to Linear Systems. In David Touretzky, editor, *Advances in Neural Information Processing Systems*, volume 1. Morgan-Kaufmann, 1989.

[12] R. Linsker. How to generate ordered maps by maximizing the mutual information between input and output signals. *Neural Computation*, 1, 1989.

[13] J.-P. Nadal, N. Brunel, and N. Parga. Nonlinear feedforward networks with stochastic outputs: infomax implies redundancy reduction. *Network: Computation in Neural Systems*, 9(2):207–217, 1998.

[14] F. Rieke, D. Warland, R R. de Ruyter van Steveninck, and W. Bialek. *Spikes – Exploring the Neural Code*. MIT Press, Cambridge, MA, 1996.

[15] D. Saad and M. Opper. *Advanced Mean Field Methods Theory and Practice*. MIT Press, 2001.

[16] E. Schneiderman, W. Bialek, and M.J. Berry II. Synergy, Redundancy, and Independence in Population Codes. *the Journal of Neuroscience*, 23(37):11539–11553, 2003.