Variational Information Maximization for Neural Coding

Felix Agakov[1] and David Barber[2]

[1] University of Edinburgh, 5 Forrest Hill, EH1 2QL Edinburgh, UK
felixa@inf.ed.ac.uk, www.anc.ed.ac.uk
[2] IDIAP, Rue du Simplon 4, CH-1920 Martigny Switzerland,
www.idiap.ch

Abstract. Mutual Information (MI) is a long studied measure of coding efficiency, and many attempts to apply it to population coding have been made. However, this is a computationally intractable task, and most previous studies redefine the criterion in forms of approximations. Recently we described properties of a simple lower bound on MI [2]. Here we describe the bound optimization procedure for learning of population codes in a simple point neural model. We compare our approach with other techniques maximizing approximations of MI, focusing on a comparison with the Fisher Information criterion.

1 Introduction

The problem of encoding real-valued stimuli x by a population of neural spikes y may be addressed in many different ways. The goal is to adapt the parameters of any mapping $p(\mathsf{y}|\mathsf{x})$ to make a desirable population code for a given set of patterns $\{\mathsf{x}\}$. There are many possible desiderata. 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 to bound mean square reconstruction error. Another approach is to bound the probability of a correct reconstruction. Here we consider maximizing the amount of information which the spiking patterns y contain about the stimuli x (e.g. [7], [5]). 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.

The principled information theoretic approach to learning neural codes maximizes (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 $H(\mathsf{x}|\mathsf{y})$, which is tractable only in a few special cases. Standard techniques often assume that $p(\mathsf{x}, \mathsf{y})$ is jointly Gaussian, the output spaces are very low-D, or the channels are invertible [4]. Other methods suggest alternative objective functions (e.g. approximations based on the *Fisher Information* [5]), which, however, do not retain proper bounds on $I(\mathsf{x}, \mathsf{y})$. Here we analyze the relation between a simple variational lower bound on the mutual information [2] and standard approaches to 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 $q(\mathsf{x}|\mathsf{y})$, leading to

$$I(\mathsf{x}, \mathsf{y}) \geq \tilde{I}(\mathsf{x}, \mathsf{y}) \overset{\text{def}}{=} H(\mathsf{x}) + \langle \log q(\mathsf{x}|\mathsf{y}) \rangle_{p(\mathsf{x},\mathsf{y})}. \tag{2}$$

Here $q(\mathsf{x}|\mathsf{y})$ is an arbitrary distribution saturating the bound for $q(\mathsf{x}|\mathsf{y}) \equiv p(\mathsf{x}|\mathsf{y})$. The objective (2) 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). The flexibility of the choice of the decoder $q(\mathsf{x}|\mathsf{y})$ makes (2) particularly computationally convenient.

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 (2), 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 $\mathsf{U} \in \mathbb{R}^{|\mathsf{x}| \times |\mathsf{y}|}$. For simplicity, we limit the discussion to this case only (though other, e.g. correlated or nonlinear cases may also be considered). Then for the empirical distribution $p(\mathsf{x}) = \sum_{m=1}^{M} \delta(\mathsf{x} - \mathsf{x}_m)/M$ we may express the bound (2) as a function of the encoder $p(\mathsf{y}|\mathsf{x})$ alone

$$\tilde{I}(\mathsf{x}, \mathsf{y}) \propto \operatorname{tr}\left\{ \langle \mathsf{x}\mathsf{y}^T \rangle \langle \mathsf{y}\mathsf{y}^T \rangle^{-1} \langle \mathsf{y}\mathsf{x}^T \rangle \right\} + const. \tag{3}$$

Note that the objective (3) 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.

[1] The bound (2) corresponds to the criteria optimized by Blahut-Arimoto algorithms (e.g. [6]); however, we optimize it for both encoder and decoder subject to enforced tractability constraints.

[2] From (3) 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.

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 conditionally independent activations, we obtain

$$p(\mathsf{y}|\mathsf{x}) = \prod_{i=1,\dots,|\mathsf{y}|} p(y_i|\mathsf{x}) \overset{\text{def}}{=} \prod_{i=1,\dots,|\mathsf{y}|} \sigma(y_i(\mathsf{w}_i^T\mathsf{x} + b_i)) \tag{4}$$

where $\mathsf{w}_i \in \mathbb{R}^{|\mathsf{x}|}$ is a vector of the synaptic weights for neuron y_i, b_i is its threshold, and $\sigma(a) \overset{\text{def}}{=} 1/(1+e^{-a})$. Optimization of (3) for $\mathsf{W} \overset{\text{def}}{=} \{\mathsf{w}_1, \dots, \mathsf{w}_{|\mathsf{y}|}\} \in \mathbb{R}^{|\mathsf{x}|\times|\mathsf{y}|}$ readily gives

$$\mathit{\Delta}\mathsf{W} \propto \sum_{m=1,\dots,M} \mathrm{cov}(\mathsf{y}|\mathsf{x}_m) \left(\tilde{\mathsf{D}}\boldsymbol{\lambda}_{x_m} + \boldsymbol{\Sigma}_{yy}^{-1}\boldsymbol{\Sigma}_{yx}\left(\mathsf{x}_m - \boldsymbol{\Sigma}_{xy}\boldsymbol{\Sigma}_{yy}^{-1}\boldsymbol{\lambda}_{x_m}\right)\right)\mathsf{x}_m^T, \tag{5}$$

where $\boldsymbol{\Sigma}_{yy} \overset{\text{def}}{=} \langle \mathsf{y}\mathsf{y}^T\rangle$, $\boldsymbol{\Sigma}_{yx} \equiv \boldsymbol{\Sigma}_{xy}^T \overset{\text{def}}{=} \langle \mathsf{y}\mathsf{x}^T\rangle$ are the second-order moments, $\tilde{\mathsf{D}}$ corresponds to the diagonal of $\boldsymbol{\Sigma}_{yy}^{-1}\boldsymbol{\Sigma}_{yx}\left(\boldsymbol{\Sigma}_{yy}^{-1}\boldsymbol{\Sigma}_{yx}\right)^T$, and $\lambda_i(\mathsf{x}) \overset{\text{def}}{=} \langle y_i\rangle_{p(y_i|\mathsf{x})} = 2\sigma(\mathsf{w}_i^T\mathsf{x} + b_i) - 1$ is the expected conditional firing of y_i. The update for the threshold $\mathit{\Delta}\mathsf{b}$ has the same form as (5) without the post-multiplication of each term by the training stimulus x_m^T.

From (5) it is clear that the magnitude of each weight update $\mathit{\Delta}\mathsf{w}_i \in \mathbb{R}^{|\mathsf{x}|}$ 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. One may also obtain a stochastic rule

$$\mathit{\Delta}\mathsf{W} \propto \tilde{\mathsf{D}}\langle\boldsymbol{\lambda}_x\mathsf{x}^T\rangle + \boldsymbol{\Sigma}_{yy}^{-1}\langle\boldsymbol{\lambda}_x\mathsf{x}^T\rangle\left(\boldsymbol{\Sigma}_{xx} - \langle\mathsf{x}\boldsymbol{\lambda}_x^T\rangle\boldsymbol{\Sigma}_{yy}^{-1}\langle\boldsymbol{\lambda}_x\mathsf{x}^T\rangle\right) \tag{6}$$

where $\boldsymbol{\Sigma}_{xx} \overset{\text{def}}{=} \langle\mathsf{x}\mathsf{x}^T\rangle$. Clearly, (6) is decomposable as a combination of the stochastic Hebbian and anti-Hebbian terms, with the weighting coefficients determined by the second-order moments of the firings and input stimuli. Additionally, from (6) one may see that the *"as-if Gaussian"* approximations [7] are suboptimal under the variational lower bound (2) – see [1] for details.

3 Fisher Information and Mutual Information

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. [6]), 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

$$I(\mathsf{x}, \mathsf{y}) \geq H(\hat{\mathsf{x}}) + \langle \log|\mathsf{F}_{\mathsf{x}}|\rangle_{p(\mathsf{x})}/2 + \mathit{const}, \tag{7}$$

where $\mathsf{F}_\mathsf{x} = \{F_{ij}(\mathsf{x})\} \overset{\text{def}}{=} -\langle \partial^2 \log p(\mathsf{y}|\mathsf{x})/\partial x_i \partial x_j \rangle_{p(\mathsf{y}|\mathsf{x})}$ is the Fisher Information matrix. 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 (7) is in general computationally intractable. It was shown that under certain assumptions $H(\hat{\mathsf{x}}) \approx H(\mathsf{x})$ [5], leading to the approximation

$$I(\mathsf{x},\mathsf{y}) \gtrsim \tilde{I}_F(\mathsf{x},\mathsf{y}) \overset{\text{def}}{=} H(\mathsf{x}) + \langle \log |\mathsf{F}_\mathsf{x}| \rangle_{p(\mathsf{x})}/2 + const, \tag{8}$$

which is then used as an approximation of $I(\mathsf{x},\mathsf{y})$ independently of the bias of the estimator. Since $H(\mathsf{x})$ is independent of $p(\mathsf{y}|\mathsf{x})$, maximization of (8) is equivalent to maximization of (7) where the intractable entropic term is ignored.

For sigmoidal activations (4), the criterion (8) is given by

$$\tilde{I}_F(\mathsf{x},\mathsf{y}) \propto \sum_{m=1,\ldots,M} \log \left| \mathsf{W}^T \mathrm{cov}(\mathsf{y}|\mathsf{x}_m) \mathsf{W} \right| + const. \tag{9}$$

Interestingly, if $|\mathsf{x}| = |\mathsf{y}|$ then optimization of (9) leads to $\varDelta \mathsf{W} = 2\mathsf{W}^{-T} - \langle \boldsymbol{\lambda}_x \mathsf{x}^T \rangle$, which (apart from the coefficient at the inverse weight – *redundancy* term) has the same form as the learning rule of [4] derived for noisless invertible channels. Notably, the weight update has no Hebbian terms. Moreover, from (9) 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 $\tilde{I}_F(\mathsf{x},\mathsf{y})$ 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-D encodings of high-D 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 (2). 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

$$\tilde{I}(\mathsf{x},\mathsf{y}) = -\frac{1}{2} \left\langle \mathrm{tr}\left\{ \boldsymbol{\Sigma}^{-1} (\mathsf{x}-\boldsymbol{\mu}_y)(\mathsf{x}-\boldsymbol{\mu}_y)^T \right\} \right\rangle_{p(\mathsf{x},\mathsf{y})} + \frac{1}{2} \log |\boldsymbol{\Sigma}^{-1}| + const. \tag{10}$$

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

$$\sum_{i=1,\ldots,|\mathsf{x}|} \log l_i(\boldsymbol{\Sigma}^{-1}) = \sum_{i=1,\ldots,|\mathsf{x}|} \langle \log l_i(\mathsf{F}_\mathsf{x}) \rangle_{p(\mathsf{x})}, \tag{11}$$

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 (10) reduces to the objective (8) amended with the average quadratic reconstruction error

$$\tilde{I}(\mathsf{x},\mathsf{y}) = -\frac{1}{2} \underbrace{\left\langle \mathrm{tr}\left\{ \boldsymbol{\Sigma}^{-1} (\mathsf{x}-\boldsymbol{\mu}_y)(\mathsf{x}-\boldsymbol{\mu}_y)^T \right\} \right\rangle_{p(\mathsf{x},\mathsf{y})}}_{\text{reconstruction error}} + \frac{1}{2} \underbrace{\langle \log |\mathsf{F}_\mathsf{x}| \rangle_{p(\mathsf{x})}}_{\text{Fisher criterion}} + const. \tag{12}$$

Arguably, it is due to the subtraction of the non-negative squared error that (10) 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 [1].

5 Experiments

Variational Information Maximization *vs* Fisher criterion

In the first set of experiments we were interested to see how the value of the true MI changed as the parameters were updated by maximising the Fisher criterion $\tilde{I}_F(\mathsf{x}, \mathsf{y})$ and the variational bound $\tilde{I}(\mathsf{x}, \mathsf{y})$. The dimension $|\mathsf{y}|$ was set to be small, so that the true $I(\mathsf{x}, \mathsf{y})$ could be computed. Fig. 1 illustrates changes in $I(\mathsf{x}, \mathsf{y})$ with iterations of the variational and Fisher-based learning rules, where the variational decoder was chosen to be an isotropic linear Gaussian with the optimal weights (3). We found that for $|\mathsf{x}| \leq |\mathsf{y}|$ (Fig. 1 (*left*)), both approaches tend to increase $I(\mathsf{x}, \mathsf{y})$ (though 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.

Variational IM: stochastic representations of the digit data

Here we apply the simple linear isotropic Gaussian decoder to stochastic coding and reconstruction of visual patterns. After numerical optimization with an explicit constraint on the channel noise, we performed reconstruction of 196-dimensional continuous visual stimuli from 7 spiking neurons. The training stimuli consisted of 30 instances of digits 1, 2, and 8 (10 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 could not be efficiently addressed by optimization of the Fisher Information-based criterion (9). Clearly, the approach of [4] is not applicable either, due to its fundamental assumption of invertible mappings between the spikes and the visual stimuli. Fig. 2 illustrates a subset of the original source signals, samples of the corresponding binary responses, and reconstructions of the source data.

6 Discussion

We described a variational approach to information maximization for the case when continuous source stimuli are represented by stochastic binary responses. We showed that for this case maximization of the lower bound on the mutual information gives rise to a form of Hebbian learning, with additional factors depending on the source and channel noise. Our results indicate that other approximate methods for information maximization [7], [5] may be viewed as approximations of our approach, which, however, do not always preserve a proper

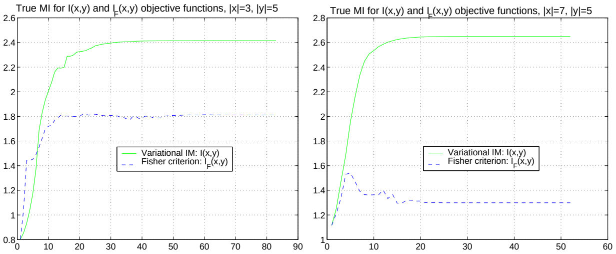

Fig. 1. Changes in the exact mutual information $I(\mathsf{x}, \mathsf{y})$ for parameters of the coder $p(\mathsf{y}|\mathsf{x})$ obtained by maximizing the variational lower bound and the Fisher information criterion for $M = 20$ training stimuli. *Left*: $|\mathsf{x}| = 3$, $|\mathsf{y}| = 5$ *Right*: $|\mathsf{x}| = 7$, $|\mathsf{y}| = 5$. In both cases, the global maximum is given by $I^{\star} = \log M \approx 3.0$.

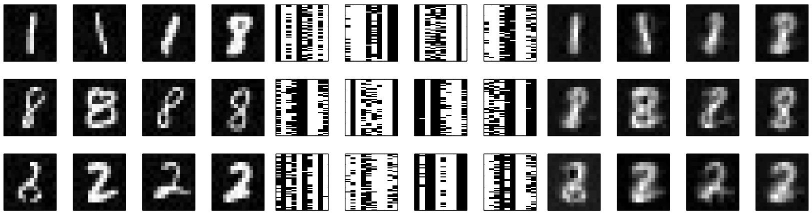

Fig. 2. *Left*: a subset of the original visual stimuli. *Middle*: 20 samples of the corresponding spikes generated by each of the 7 neurons. *Right*: Reconstructions from 50 samples of neural spikes (with soft constraints on the variances of firings).

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.

References

1. Agakov, F. V. and Barber, D (2004). Variational Information Maximization and Fisher Information. Technical report, UoE.
2. Barber, D. and Agakov, F. V. (2003). The IM Algorithm: A Variational Approach to Information Maximization. In *NIPS*.
3. Barlow, H. (1989). Unsupervised Learning. *Neural Computation*, 1:295–311.
4. Bell, A. J. and Sejnowski, T. J. (1995). An information-maximization approach to blind separation and blind deconvolution. *Neural Computation*, 7(6):1129–1159.
5. Brunel, N. and Nadal, J.-P. (1998). Mutual Information, Fisher Information and Population Coding. *Neural Computation*, 10:1731–1757.
6. Cover, T. M. and Thomas, J. A. (1991). *Elements of Information Theory*. Wiley.
7. Linsker, R. (1989). An Application of the Principle of Maximum Information to Linear Systems. In *NIPS*.