Compressive neural representation of sparse, high-dimensional probabilities

xaq pitkow
Department of Brain and Cognitive Sciences
University of Rochester
Rochester, NY 14607
xpitkow@bcs.rochester.edu

Abstract

This paper shows how sparse, high-dimensional probability distributions could be represented by neurons with exponential compression. The representation is a novel application of compressive sensing to sparse probability distributions rather than to the usual sparse signals. The compressive measurements correspond to expected values of nonlinear functions of the probabilistically distributed variables. When these expected values are estimated by sampling, the quality of the compressed representation is limited only by the quality of sampling. Since the compression preserves the geometric structure of the space of sparse probability distributions, probabilistic computation can be performed in the compressed domain. Interestingly, functions satisfying the requirements of compressive sensing can be implemented as simple perceptrons. If we use perceptrons as a simple model of feedforward computation by neurons, these results show that the mean activity of a relatively small number of neurons can accurately represent a high-dimensional joint distribution implicitly, even without accounting for any noise correlations. This comprises a novel hypothesis for how neurons could encode probabilities in the brain.

1 Introduction

Behavioral evidence shows that animal behaviors are often influenced not only by the content of sensory information but also by its uncertainty. Different theories have been proposed about how neuronal populations could represent this probabilistic information [1, 2]. Here we propose a new theory of how neurons could represent probability distributions, based on the burgeoning field of 'compressive sensing.'

An arbitrary probability distribution over multiple variables has a parameter count that is exponential in the number of variables. Representing these probabilities can therefore be prohibitively costly. One common approach is to use graphical models to parameterize the distribution in terms of a smaller number of interactions. Here I consider an alternative approach. In many cases of interest, only a few unknown states have high probabilities while the rest have neglible ones; such a distribution is called 'sparse'. I will show that sufficiently sparse distributions can be described by a number of parameters that is merely linear in the number of variables.

Until recently, it was generally thought that encoding of sparse signals required dense sampling at a rate greater than or equal to signal bandwidth. However, recent findings prove that it is possible to fully characterize a signal at a rate limited not by its bandwidth but by its information content [3, 4, 5, 6] which can be much smaller. Here I apply such compression to sparse probability distributions over binary variables, which are, after all, just signals with some particular properties.

In most applications of compressive sensing, the ultimate goal is to reconstruct the original signal efficiently. Here, we do not wish to reconstruct the signal at all. Instead, we use the guarantees that the signal *could* be reconstructed to ensure that the signal is accurately represented by its compressed version. Below, when we do reconstruct it is only to show that our method actually works in practice. We don't expect that the brain needs to explicitly reconstruct a probability distribution in some canonical mathematical representation in order to gain the advantages of probabilistic reasoning.

Traditional compressive sensing considers signals that lives in an N-dimensional space but have only S nonzero coordinates in some basis. We say that such a signal is S-sparse. If we were told the location of the nonzero entries, then we would need only S measurements to characterize their coefficients and thus the entire signal. But even if we don't know where those entries are, it still takes little more than S linear measurements to perfectly reconstruct the signal. Furthermore, those measurements can be fixed in advance without any knowledge of the structure of the signal. Under certain conditions, these excellent properties can be guaranteed [3, 4, 5].

The basic mathematical setup of compressive sensing is as follows. Assume that an N-dimensional signal $\boldsymbol{s}$ has S nonzero coefficients. We make M linear measurements $\boldsymbol{y}$ of this signal by applying the $M \times N$ matrix A:

$$\boldsymbol{y} = A\boldsymbol{s} \tag{1}$$

We would then like to recover the original signal $\boldsymbol{s}$ from these measurements. Under conditions on the measurement matrix A described below, the original can be found perfectly by computing the vector with minimal ℓ_1 norm that reproduces the measurements,

$$\hat{\boldsymbol{s}} = \underset{\boldsymbol{s}'}{\operatorname{argmin}} \|\boldsymbol{s}'\|_{\ell_1} \text{ such that } A\boldsymbol{s}' = \boldsymbol{y} = A\boldsymbol{s} \tag{2}$$

The ℓ_1 norm is usually used instead of ℓ_0 because (2) can be solved far more efficiently [3, 4, 5, 7].

Compressive sensing is generally robust to two deviations from this ideal setup. First, target signals may not be strictly S-sparse. However, they may be 'compressible' in the sense that they are well approximated by an S-sparse signal. Signals whose rank-ordered coefficients fall off at least as fast as rank^{-1} satisfy this property [4]. Second, measurements may be corrupted by noise with bounded amplitude ϵ. Under these conditions, the error of the ℓ_1-reconstructed signal $\hat{\boldsymbol{s}}$ is bounded by the error of the best S-sparse approximation $\boldsymbol{s}_S$ plus a term proportional to the measurement noise:

$$\|\hat{\boldsymbol{s}} - \boldsymbol{s}\|_{\ell_2} \leq C_0 \|\boldsymbol{s}_S - \boldsymbol{s}\|_{\ell_2} / \sqrt{S} + C_1 \epsilon \tag{3}$$

for some constants C_0 and C_1 [8].

Several conditions on A have been used in compressive sensing to guarantee good performance [4, 6, 9, 10, 11]. Modulo various nuances, they all essentially ensure that most or all relevant sparse signals lie sufficiently far from the null space of A: It would be impossible to recover signals in the null space since their measurements are all zero and cannot therefore be distinguished. The most commonly used condition is the Restricted Isometry Property (RIP), which says that A preserves ℓ_2 norms of all S-sparse vectors within a factor of $1 \pm \delta_S$ that depends on the sparsity,

$$(1 - \delta_S)\|\boldsymbol{s}\|_{\ell_2} \;\leq\; \|A\boldsymbol{s}\|_{\ell_2} \;\leq\; (1 + \delta_S)\|\boldsymbol{s}\|_{\ell_2} \tag{4}$$

If A satisfies the RIP with small enough δ_S, then ℓ_1 recovery is guaranteed to succeed. For random matrices whose elements are independent and identically distributed Gaussian or Bernoulli variates, the RIP holds as long as the number of measurements M satisfies

$$M \geq CS \log N/S \tag{5}$$

for some constant C that depends on δ_S [8]. No other recovery method, however intractable, can perform substantially better than this [8].

2 Compressing sparse probability distributions

Compressive sensing allows us to use far fewer resources to accurately represent high-dimensional objects if they are sufficiently sparse. Even if we don't ultimately intend to reconstruct the signal, the reconstruction theorem described above (3) ensures that we have implicitly represented all the relevant information. This compression proves to be extremely useful when representing multivariate joint probability distributions, whose size is exponentially large even for the simplest binary states.

Consider the signal to be a probability distribution over an n-dimensional binary vector $\boldsymbol{x} \in \{-1,+1\}^n$, which I will write sometimes as a function $p(\boldsymbol{x})$ and sometimes as a vector $\boldsymbol{p}$ indexed by the binary state $\boldsymbol{x}$. I assume $\boldsymbol{p}$ is sparse in the canonical basis of delta-functions on each state, $\delta_{\boldsymbol{x},\boldsymbol{x}'}$. The dimensionality of this signal is $N = 2^n$, which for even modest n can be so large it cannot be represented explicitly.

The measurement matrix A for probability vectors has size $M \times 2^n$. Each row corresponds to a different measurement, indexed by i. Each column corresponds to a different binary state $\boldsymbol{x}$. This column index $\boldsymbol{x}$ ranges over all possible binary vectors of length n, in some conventional sequence. For example, if $n = 3$ then the column index would take the 8 values

$$\boldsymbol{x} \in \{---\,;\ --+\,;\ -+-\,;\ -++\,;\ +--\,;\ +-+\,;\ ++-\,;\ +++\}$$

Each element of the measurement matrix, $A_i(\boldsymbol{x})$, can be viewed as a function applied to the binary state. When this matrix operates on a probability distribution $p(\boldsymbol{x})$, the result $\boldsymbol{y}$ is a vector of M expectation values of those functions, with elements

$$y_i = A_i \boldsymbol{p} = \sum_{\boldsymbol{x}} A_i(\boldsymbol{x}) p(\boldsymbol{x}) = \langle A_i(\boldsymbol{x}) \rangle_{p(\boldsymbol{x})} \tag{6}$$

For example, if $A_i(\boldsymbol{x}) = x_i$ then $y_i = \langle x_i \rangle_{p(\boldsymbol{x})}$ measures the mean of x_i drawn from $p(\boldsymbol{x})$.

For suitable measurement matrices A, we are guaranteed accurate reconstruction of S-sparse probability distributions as long as the number of measurements is

$$M \geq O(S \log N/S) = O(Sn - S \log S) \tag{7}$$

The exponential size of the probability vector, $N = 2^n$, is cancelled by the logarithm. For distributions with a fixed sparseness S, the required number of measurements per variable, M/n, is then independent of the number of variables.[1]

In many cases of interest it is impractical to calculate these expectation values directly: Recall that the probabilities may be too expensive to represent explicitly in the first place. One remedy is to draw T samples $\boldsymbol{x}_t$ from the distribution $p(\boldsymbol{x})$, and use a sum over these samples to approximate the expectation values,

$$y_i \approx \frac{1}{T} \sum_t A_i(\boldsymbol{x}_t) \qquad \boldsymbol{x}_t \sim p(\boldsymbol{x}) \tag{8}$$

The probability $\hat{p}(\boldsymbol{x})$ estimated from T samples has errors with variance $p(\boldsymbol{x})(1 - p(\boldsymbol{x}))/T$, which is bounded by $1/4T$. This allows us to use the performance limits from robust compressive sensing, which according to (3) creates an error in the reconstructed probabilities that is bounded by

$$\|\hat{\boldsymbol{p}} - \boldsymbol{p}\|_{\ell_2} \leq C_0 \|\boldsymbol{p}_S - \boldsymbol{p}\|_{\ell_2} + \frac{C_1}{\sqrt{T}} \tag{9}$$

where $\boldsymbol{p}_S$ is a vector with the top S probabilities preserved and the rest set to zero. Strictly speaking, (3) applies to bounded errors, whereas here we have a bounded variance but possibly large errors. To ensure accurate reconstruction, we can choose the constant C_1 large enough that errors larger than some threshold (say, 10 standard deviations) have a negligible probability.

2.1 Measurements by random perceptrons

In compressive sensing it is common to use a matrix with independent Bernoulli-distributed random values, $A_i(\boldsymbol{x}) \sim \mathcal{B}(\frac{1}{2})$, which guarantees A satisfies the RIP [12]. Each row of this matrix represents all possible outputs of an arbitrarily complicated Boolean function of the n binary variables $\boldsymbol{x}$.

Biological neural networks would have great difficulty computing such arbitrary functions in a simple manner. However, neurons can easily compute a large class of simpler boolean functions, the perceptrons. These are simple threshold functions of a weighted average of the input

$$A_i(\boldsymbol{x}) = \operatorname{sgn}\left(\sum\nolimits_j W_{ij} x_j - \theta_j\right) \tag{10}$$

[1] Depending on the problem, the number of significant nonzero entries S may grow with the number of variables. This growth may be fast (e.g. the number of possible patterns grows as e^n) or slow (e.g. the number of possible translations of a given pattern grows only as n).

where W is an $M \times n$ matrix. Here I take W to have elements drawn randomly from a standard normal distribution, $W_{ij} \sim \mathcal{N}(0,1)$, and call the resultant functions 'random perceptrons'. An example measurement matrix for random perceptrons is shown in Figure 1. These functions are readily implemented by individual neurons, where x_j is the instantaneous activity of neuron j, W_{ij} is the synaptic weight between neurons i and j, and the sgn function approximates a spiking threshold at θ.

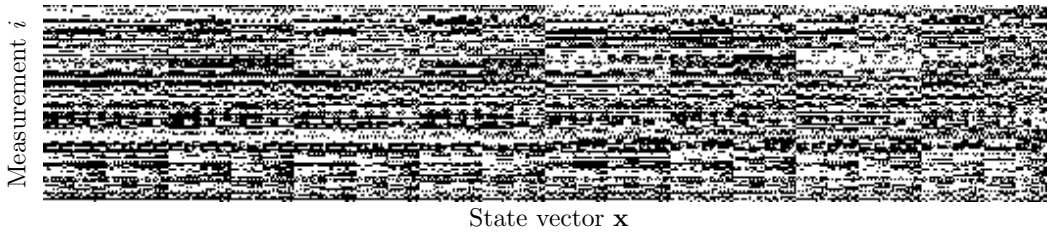

Figure 1: Example measurement matrix $A_i(\boldsymbol{x})$ for $M = 100$ random perceptrons applied to all 2^9 possible binary vectors of length $n = 9$.

The step nonlinearity sgn is not essential, but *some* type of nonlinearity is: Using a purely linear function of the states, $A = W\boldsymbol{x}$, would result in measurements $\boldsymbol{y} = A\boldsymbol{p} = W\langle \boldsymbol{x} \rangle$. This provides at most n linearly independent measurements of $p(\boldsymbol{x})$, even when $M > n$. In most cases this is not enough to adequately capture the full distribution. Nonlinear $A_i(\boldsymbol{x})$ allow a greater number of linearly independent measurements of $p(\boldsymbol{x})$. Although the dimensionality of W is merely $M \times n$, which is much smaller than the 2^n-dimensional space of probabilities, (10) can generate $O(2^{n^2})$ distinct perceptrons [13]. By including an appropriate threshold, a perceptron can assign any individual state $\boldsymbol{x}$ a positive response and assign a negative response to every other state. This shows that random perceptrons generate the canonical basis and can thus span the space of possible $p(\boldsymbol{x})$. In what follows, I assume that $\boldsymbol{\theta} = 0$ for simplicity.

In the Appendix I prove that random perceptrons with zero threshold satisfy the requirements for compressive sensing in the limit of large n. Present research is directed toward deriving the condition number of these measurement matrices for finite n, in order to provide rigorous bounds on the number of measurements required in practice. Below I present empirical evidence that even a small number of random perceptrons largely preserves the information about sparse distributions.

3 Experiments

3.1 Fidelity of compressed sparse distributions

To test random perceptrons in compressive sensing of probabilities, I generated sparse distributions using small Boltzmann machines [14], and compressed them using random perceptrons driven by samples from the Boltzmann machine. Performance was then judged by comparing ℓ_1 reconstructions to the true distributions, which are exactly calculable for modest n.

In a Boltzmann Machine, binary states $\boldsymbol{x}$ occur with probabilities given by the Boltzmann distribution with energy function $E(\boldsymbol{x})$,

$$p(\boldsymbol{x}) \propto e^{-E(\boldsymbol{x})} \qquad E(\boldsymbol{x}) = -\boldsymbol{b}^\top \boldsymbol{x} - \boldsymbol{x}^\top J \boldsymbol{x} \tag{11}$$

determined by biases $\boldsymbol{b}$ and pairwise couplings J. Sampling from this distribution can be accomplished by running Glauber dynamics [15], at each time step turning a unit on with probability $p(x_i = +1|\boldsymbol{x}_{\setminus i}) = 1/(1 + e^{-\Delta E})$, where $\Delta E = E(x_i = +1, \boldsymbol{x}_{\setminus i}) - E(x_i = -1, \boldsymbol{x}_{\setminus i})$. Here $\boldsymbol{x}_{\setminus i}$ is the vector of all components of $\boldsymbol{x}$ except the ith.

For simulations I distinguished between two types of units, hidden and visible, $\boldsymbol{x} = (\boldsymbol{h}, \boldsymbol{v})$. On each trial I first generated a sample of all units according to (11). I then fixed only the visible units and allowed the hidden units to fluctuate according to the conditional probability $p(\boldsymbol{h}|\boldsymbol{v})$ to be represented. This probability is given again by the Boltzmann distribution, now with energy function

$$E(\boldsymbol{h}|\boldsymbol{v}) = -(\boldsymbol{b}_h - J_{hv}\boldsymbol{v})^\top \boldsymbol{h} - \boldsymbol{h}^\top J_{hh} \boldsymbol{h} \tag{12}$$

All bias terms $\boldsymbol{b}$ were set to zero, and all pairwise couplings J were random draws from a zero-mean normal distribution, $J_{ij} \sim \mathcal{N}(0, \frac{1}{3})$. Experiments used n hidden and n visible units, with $n \in \{8, 10, 12\}$. This distribution of couplings produced sparse posterior distributions whose rank-ordered probabilities fell faster than rank^{-1} and were thus compressible [4].

The compression was accomplished by passing the hidden unit activities $\boldsymbol{h}$ through random perceptrons $\boldsymbol{a}$ with weights W, according to $\boldsymbol{a} = \mathrm{sgn}(W\boldsymbol{h})$. These perceptron activities fluctuate along with their inputs. The mean activity of these perceptron units compressively senses the probability distribution according to (8). This process of sampling and then compressing a Boltzmann distribution can be implemented by the simple neural network shown in Figure 2.

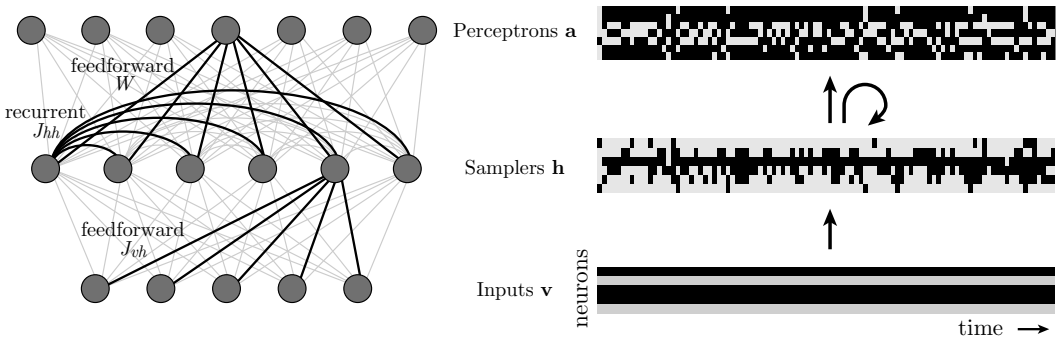

Figure 2: Compressive sensing of a probability distribution by model neurons. Left: a neural architecture for generating and then encoding a sparse, high-dimensional probability distribution. Right: activity of each population of neurons as a function of time. Sparse posterior probability distribution are generated by a Boltzmann Machine with visible units $\boldsymbol{v}$ (Inputs), hidden units $\boldsymbol{h}$ (Samplers), feedforward couplings J_{vh} from visible to hidden units, and recurrent connections between hidden units J_{hh}. The visible units' activities are fixed by an input. The hidden units are stochastic, and sample from a probability distribution $p(\boldsymbol{h}|\boldsymbol{v})$. The samples are recoded by feedforward weights W to random perceptrons $\boldsymbol{a}$. The mean activity $\boldsymbol{y}$ of the time-dependent perceptron responses captures the sparse joint distribution of the hidden units.

We are not ultimately interested in reconstruction of the large, sparse distribution, but rather the distribution's compressed representation. Nonetheless, reconstruction is useful to show that the information has been preserved. I reconstruct sparse probabilities using nonnegative ℓ_1 minimization with measurement constraints [16, 17], minimizing

$$\|\boldsymbol{p}\|_{\ell_1} + \lambda \|A\boldsymbol{p} - \boldsymbol{y}\|_{\ell_2}^2 \tag{13}$$

where λ is a regularization parameter that was set to $2T$ in all simulations. Reconstructions were quite good, as shown in Figure 3. Even with far fewer measurements than signal dimensions, reconstruction accuracy is limited only by the sampling of the posterior. Enough random perceptrons do not lose any available information.

In the context of probability distributions, ℓ_1 reconstruction has a serious flaw: All distributions have the same ℓ_1 norm: $\|\boldsymbol{p}\|_{\ell_1} = \sum_{\boldsymbol{x}} p(\boldsymbol{x}) = 1$! To minimize the ℓ_1 norm, therefore, the estimate will not be a probability distribution. Nonetheless, the individual probabilities of the most significant states are accurately reconstructed, and only the highly improbable states are set to zero. Figure 3B shows that the shortfall is small: ℓ_1 reconstruction recovers over 90% of the total probability mass.

3.2 Preserving computationally important relationships

There is value in being able to compactly represent these high-dimensional objects. However, it would be especially useful to perform probabilistic computations using these representations, such as marginalization and evidence integration. Since marginalization is a linear operation on the probability distribution, this is readily implementable in the linearly compressed domain. In contrast, evidence integration is a multiplicative process acting in the canonical basis, so this operation will be more complicated after the linear distortions of compressive measurement A. Nonetheless, such computations should be feasible as long as the informative relationships are preserved in the compressed space: Similar distributions should have similar compressive representations, and dissimilar

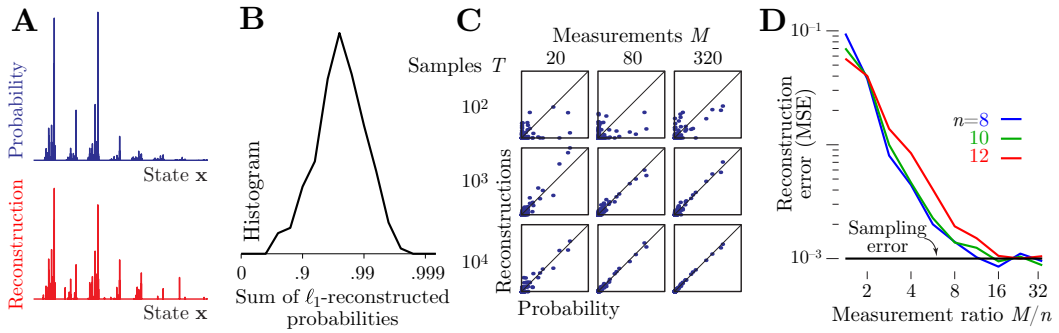

Figure 3: Reconstruction of sparse posteriors from random perceptron measurements. (**A**) A sparse posterior distribution over 10 nodes in a Boltzmann machine is sampled 1000 times, fed to 50 random perceptrons, and reconstructed by nonnegative ℓ_1 minimization. (**B**) A histogram of the sum of reconstructed probabilities reveals the small shortfall from a proper normalization of 1. (**C**) Scatter plots show reconstructions versus true probabilities. Each box uses different numbers of compressive measurements M and numbers of samples T. (**D**) With increasing numbers of compressive measurements, the mean squared reconstruction error falls to $1/T = 10^{-3}$, the limit imposed by finite sampling.

distributions should have dissimilar compressive representations. In fact, that is precisely the guarantee of compressive sensing: topological properties of the underlying space are preserved in the compressive domain [18]. Figure 4 illustrates how not only are individual sparse distributions recoverable despite significant compression, but the topology of the set of all such distributions is retained.

For this experiment, an input $\boldsymbol{x}$ is drawn from a dictionary of input patterns $\mathcal{X} \subset \{+1, -1\}^n$. Each pattern in $\mathcal{X}$ is a translation of a single binary template $\boldsymbol{x}^0$ whose elements are generated by thresholding a noisy sinusoid (Figure 4A): $x_j^0 = \mathrm{sgn}\left[4\sin\left(2\pi j/n\right) + \eta_j\right]$ with $\eta_j \sim \mathcal{N}(0, 1)$. On each trial, one of these possible patterns is drawn randomly with equal probability $1/|\mathcal{X}|$, and then is measured by a noisy process that randomly flips bits with a probability $\eta = 0.35$ to give a noisy pattern $\boldsymbol{r}$. This process induces a posterior distribution over the possible input patterns

$$p(\boldsymbol{x}|\boldsymbol{r}) = \frac{1}{Z}p(\boldsymbol{x})\prod_i p(r_i|x_i) = \frac{1}{Z}p(\boldsymbol{x})\eta^{N-h(\boldsymbol{x},\boldsymbol{r})}(1-\eta)^{h(\boldsymbol{x},\boldsymbol{r})} \tag{14}$$

where $h(\boldsymbol{x}, \boldsymbol{r})$ is the Hamming distance between $\boldsymbol{x}$ and $\boldsymbol{r}$. This posterior is nonzero for all patterns in the dictionary. The noise level and the similarities between the dictionary elements together control the sparseness.

1000 trials of this process generates samples from the set of all possible posterior distributions. Just as the underlying set of inputs has a translation symmetry, the set of all possible posterior distributions has a cyclic permutation symmetry. This symmetry can be revealed by a nonlinear embedding [19] of the set of posteriors into two dimensions (Figure 4B).

Compressive sensing of these posteriors by 10 random perceptrons produces a much lower-dimensional embedding that preserves this symmetry. Figure 4C shows that the same nonlinear embedding algorithm applied to the reduced representation, and one sees the same topological pattern. In compressive sensing, similarity is measured by Euclidean distance. When applied to probability distributions it will be interesting to examine instead how well information-geometric measures like the Kullback-Leibler divergence are preserved under this dimensionality reduction [20].

4 Discussion

Probabilistic inference appears to be essential for both animals and machines to perform well on complex tasks with natural levels of ambiguity, but it remains unclear how the brain represents and manipulates probability. Present population models of neural inference either struggle with high-dimensional distributions [1] or encode them by hard-to-measure high-order correlations [2]. Here I have proposed an alternative mechanism by which the brain could efficiently represent probabilities: random perceptrons. In this model, information about probabilities is compressed and distributed

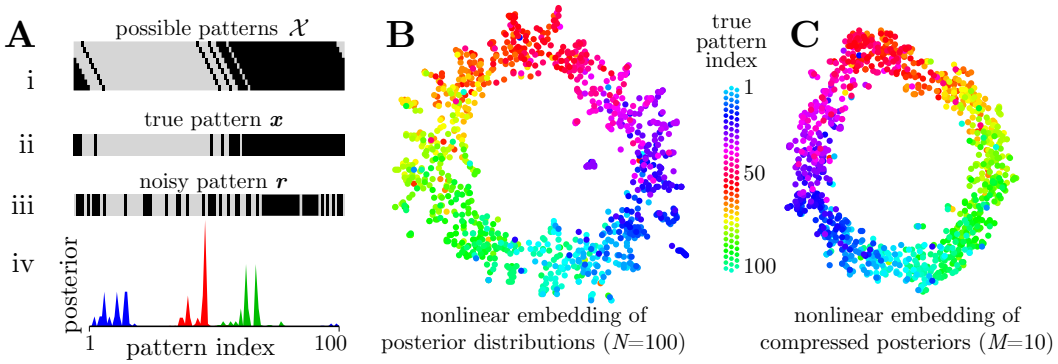

Figure 4: Nonlinear embeddings of a family of probability distributions with a translation symmetry. (**A**) The process of generating posterior distributions: (i) A set of 100 possible patterns is generated as cyclic translations of a binary pattern (only 9 shown). With uniform probability, one of these patterns is selected (ii), and a noisy version is obtained by randomly flipping bits with probability 0.35 (iii). From such noisy patterns, an observer can infer posterior probability distributions over possible inputs (iv). (**B**) The set of posteriors from 1000 iterations of this process is nonlinearly mapped [19] from 100 dimensions to 2 dimensions. Each point represents one posterior and is colored according to the actual pattern from which the noisy observations were made. The permutation symmetry of this process is revealed as a circle in this mapping. (**C**) This circular structure is retained even after each posterior is compressed into the mean output of 10 random perceptrons.

in neural population activity. Amazingly, the brain need not measure any correlations between the perceptron outputs to capture the joint statistics of the sparse input distribution. Only the mean activities are required. Figure 2 illustrates one network that implements this new representation, and many variations on this circuit are possible.

Successful encoding in this compressed representation requires that the input distribution be sparse. Posterior distributions over sensory stimuli like natural images are indeed expected to be highly sparse: the features are sparse [21], the prior over images is sparse [22], and the likelihood produced by sensory evidence is usually restrictive, so the posteriors should be even sparser. Still, it will be important to quantify just how sparse the relevant posteriors are under different conditions. This would permit us to predict how neural representations in a fixed population should degrade as sensory evidence becomes weaker.

Brains appear to have a mix of structure and randomness. The results presented here show that purely random connections are sufficient to ensure that a sparse probability distribution is properly encoded. Surprisingly, more structured connections cannot allow a network with the same computational elements to encode distributions with substantially fewer neurons, since compressive sensing is already nearly optimal [8]. On the other hand, some representational structure may make it easier to perform computations later. Note that unknown randomness is not an impediment to further processing, as reconstruction can be performed even without explicit knowledge of random perceptron measurement matrix [23].

Even in the most convenient representations, inference is generally intractable and requires approximation. Since compressive sensing preserves the essential geometric relationships of the signal space, learning and inference based on these relationships may be no harder after the compression, and could even be more efficient due to the reduced dimensionality. Biologically plausible mechanisms for implementing probabilistic computations in the compressed representation is important work for the future.

Appendix: Asymptotic orthogonality of random perceptron matrix

To evaluate the quality of the compressive sensing matrix A, we need to ensure that S-sparse vectors are not projected to zero by the action of A. Here I show that the random perceptrons are asymptotically well-conditioned: $\hat{A}^\top \hat{A} \to I$ for large n and M, where $\hat{A} = A/\sqrt{M}$. This ensures that distinct inputs yield distinct measurements.

First I compute the mean and variance of the mean inner product $\langle C_{\boldsymbol{x}\boldsymbol{x}'}\rangle_W$ between columns of $\hat{A}$ for a given pair of states $\boldsymbol{x} \neq \boldsymbol{x}'$. For compactness I will write $\boldsymbol{w}_i$ for the ith row of the perceptron weight matrix W. Angle brackets $\langle\,\rangle_W$ indicate averages over random perceptron weights $W_{ij} \sim \mathcal{N}(0,1)$. We find

$$\langle C_{\boldsymbol{x}\boldsymbol{x}'}\rangle_W = \left\langle \sum_i \hat{A}_i(\boldsymbol{x})\hat{A}_i(\boldsymbol{x}') \right\rangle_W = \frac{1}{M}\sum_i \langle \mathrm{sgn}\,(\boldsymbol{w}_i\cdot\boldsymbol{x})\,\mathrm{sgn}\,(\boldsymbol{w}_i\cdot\boldsymbol{x}')\rangle_W \tag{15}$$

and since the different $\boldsymbol{w}_i$ are independent, this implies that

$$\langle C_{\boldsymbol{x}\boldsymbol{x}'}\rangle_W = \langle \mathrm{sgn}\,(\boldsymbol{w}_i\cdot\boldsymbol{x})\,\mathrm{sgn}\,(\boldsymbol{w}_i\cdot\boldsymbol{x}')\rangle_W \tag{16}$$

The n-dimensional half-space in W where $\mathrm{sgn}\,(\boldsymbol{w}_i\cdot\boldsymbol{x}) = +1$ intersects with the corresponding half-space for $\boldsymbol{x}'$ in a wedge-shaped region with an angle of $\theta = \cos^{-1}(\boldsymbol{x}\cdot\boldsymbol{x}'/\|\boldsymbol{x}\|_{\ell_2}\|\boldsymbol{x}'\|_{\ell_2})$. This angle is related to the Hamming distance $h = h(\boldsymbol{x},\boldsymbol{x}')$:

$$\theta(h) = \cos^{-1}(\boldsymbol{x}\cdot\boldsymbol{x}'/n) = \cos^{-1}(1-2h/n) \tag{17}$$

The signs of $\boldsymbol{w}_i\cdot\boldsymbol{x}$ and $\boldsymbol{w}_i\cdot\boldsymbol{x}'$ agree within this wedge region and its reflection about $W=0$, and disagree in the supplementary wedges. The mean inner product is therefore

$$\langle C_{\boldsymbol{x}\boldsymbol{x}'}\rangle_W = P\left[\mathrm{sgn}\,(\boldsymbol{w}_i\cdot\boldsymbol{x}) = \mathrm{sgn}\,(\boldsymbol{w}_i\cdot\boldsymbol{x}')\right] - P\left[\mathrm{sgn}\,(\boldsymbol{w}_i\cdot\boldsymbol{x}) \neq \mathrm{sgn}\,(\boldsymbol{w}_i\cdot\boldsymbol{x}')\right] \tag{18}$$

$$= 1 - \tfrac{2}{\pi}\theta(h) \tag{19}$$

The variance of $C_{\boldsymbol{x}\boldsymbol{x}'}$ caused by variability in W is given by

$$V_{\boldsymbol{x}\boldsymbol{x}'} = \left\langle C^2_{\boldsymbol{x}\boldsymbol{x}'}\right\rangle_W - \langle C_{\boldsymbol{x}\boldsymbol{x}'}\rangle^2_W \tag{20}$$

$$= \sum_{i=j}\left\langle \hat{A}_i^2(\boldsymbol{x})\hat{A}_i^2(\boldsymbol{x}')\right\rangle_W + \sum_{i\neq j}\left\langle \hat{A}_i(\boldsymbol{x})\hat{A}_i(\boldsymbol{x}')\hat{A}_j(\boldsymbol{x})\hat{A}_j(\boldsymbol{x}')\right\rangle_W - \langle C_{\boldsymbol{x}\boldsymbol{x}'}\rangle^2_W \tag{21}$$

$$= \sum_i\left\langle \frac{\mathrm{sgn}\,(\boldsymbol{w}_i\cdot\boldsymbol{x})^2}{M}\frac{\mathrm{sgn}\,(\boldsymbol{w}_i\cdot\boldsymbol{x}')^2}{M}\right\rangle_W + \sum_{i\neq j}\left\langle \frac{\mathrm{sgn}\,(\boldsymbol{w}_i\cdot\boldsymbol{x})}{\sqrt{M}}\frac{\mathrm{sgn}\,(\boldsymbol{w}_i\cdot\boldsymbol{x}')}{\sqrt{M}}\right\rangle^2_W - \langle C_{\boldsymbol{x}\boldsymbol{x}'}\rangle^2_W \tag{22}$$

$$= \frac{1}{M} + \frac{M^2-M}{M^2}(1-2\theta(h)/\pi)^2 - \langle C_{\boldsymbol{x}\boldsymbol{x}'}\rangle^2_W \tag{23}$$

$$= \frac{1}{M}\left(1-\left(1-\tfrac{2}{\pi}\theta(h(\boldsymbol{x},\boldsymbol{x}'))\right)^2\right) \tag{24}$$

This variance falls with M, so for large numbers of measurements M the inner products between columns concentrates around the various state-dependent mean values (19).

Next I consider the diversity of inner products for different pairs $(\boldsymbol{x},\boldsymbol{x}')$ of binary state vectors. I take the limit of large M so that the diversity is dominated by variations over the particular pairs, rather than by variations over measurements. The mean inner product depends only on the Hamming distance h between $\boldsymbol{x}$ and $\boldsymbol{x}'$, which for sparse signals with random support has a binomial distribution, $p(h) = \binom{n}{h}2^{-n}$ with mean $n/2$ and variance $n/4$. Designating by an overbar the average over randomly chosen states $\boldsymbol{x}$ and $\boldsymbol{x}'$, the mean $\overline{C}$ and variance $\overline{\delta C^2}$ of the inner product are

$$\overline{C} = \overline{\langle C_{\boldsymbol{x}\boldsymbol{x}'}\rangle_W} = 1 - \tfrac{2}{\pi}\overline{\cos^{-1}(1-\tfrac{2h}{n})} = 0 \tag{25}$$

$$\overline{\delta C^2} = \overline{\delta h^2}\left(\frac{\partial C}{\partial h}\right)^2 = \frac{n}{4}\frac{16}{\pi^2 n^2} = \frac{4}{\pi^2 n} \tag{26}$$

This proves that in the limit of large n and M, different columns of the random perceptron measurement matrix have inner products that concentrate around 0. The matrix of inner products is thus orthonormal almost surely, $\hat{A}^\top\hat{A} \to I$. Consequently, with enough measurements the random perceptrons asymptotically provide an isometry. Future work will investigate how the measurement matrix behaves for finite n and M, which will determine the number of measurements required in practice to capture a signal of a given sparseness.

Acknowledgments

Thanks to Alex Pouget, Jeff Beck, Shannon Starr, and Carmelita Navasca for helpful conversations.

References

[1] Ma W, Beck J, Latham P, Pouget A (2006) Bayesian inference with probabilistic population codes. Nat Neurosci 9: 1432–8.

[2] Berkes P, Orbán G, Lengyel M, Fiser J (2011) Spontaneous cortical activity reveals hallmarks of an optimal internal model of the environment. Science 331: 83–7.

[3] Candès E, Romberg J, Tao T (2006) Robust uncertainty principles: Exact signal reconstruction from highly incomplete frequency information. IEEE Transactions on Information Theory 52: 489–509.

[4] Candès E, Tao T (2006) Near-optimal signal recovery from random projections: Universal encoding strategies? IEEE Transactions on Information Theory 52: 5406–5425.

[5] Donoho D (2006) Compressed sensing. IEEE Transactions on Information Theory 52: 1289–1306.

[6] Candès E, Plan Y (2011) A probabilistic and RIPless theory of compressed sensing. IEEE Transactions on Information Theory 57: 7235–7254.

[7] Donoho DL, Maleki A, Montanari A (2009) Message-passing algorithms for compressed sensing. Proc Natl Acad Sci USA 106: 18914–9.

[8] Candès E, Wakin M (2008) An introduction to compressive sampling. Signal Processing Magazine 25: 21–30.

[9] Kueng R, Gross D (2012) RIPless compressed sensing from anisotropic measurements. Arxiv preprint arXiv:12051423 .

[10] Calderbank R, Howard S, Jafarpour S (2010) Construction of a large class of deterministic sensing matrices that satisfy a statistical isometry property. Selected Topics in Signal Processing 4: 358–374.

[11] Gurevich S, Hadani R (2009) Statistical rip and semi-circle distribution of incoherent dictionaries. arXiv cs.IT.

[12] Mendelson S, Pajor A, Tomczak-Jaegermann N (2006) Uniform uncertainty principle for Bernoulli and subgaussian ensembles. arXiv math.ST.

[13] Irmatov A (2009) Bounds for the number of threshold functions. Discrete Mathematics and Applications 6: 569–583.

[14] Ackley D, Hinton G, Sejnowski T (1985) A learning algorithm for Boltzmann machines. Cognitive Science 9: 147–169.

[15] Glauber RJ (1963) Time-dependent statistics of the Ising model. Journal of Mathematical Physics 4: 294–307.

[16] Yang J, Zhang Y (2011) Alternating direction algorithms for L1 problems in compressive sensing. SIAM Journal on Scientific Computing 33: 250–278.

[17] Zhang Y, Yang J, Yin W (2010) YALL1: Your ALgorithms for L1. CAAM Technical Report : TR09-17.

[18] Baraniuk R, Cevher V, Wakin MB (2010) Low-dimensional models for dimensionality reduction and signal recovery: A geometric perspective. Proceedings of the IEEE 98: 959–971.

[19] der Maaten LV, Hinton G (2008) Visualizing high-dimensional data using t-SNE. Journal of Machine Learning Research 9: 2579–2605.

[20] Carter KM, Raich R, Finn WG, Hero AO (2011) Information-geometric dimensionality reduction. IEEE Signal Process Mag 28: 89–99.

[21] Olshausen BA, Field DJ (1996) Emergence of simple-cell receptive field properties by learning a sparse code for natural images. Nature 381: 607–9.

[22] Stephens GJ, Mora T, Tkacik G, Bialek W (2008) Thermodynamics of natural images. arXiv q-bio.NC.

[23] Isely G, Hillar CJ, Sommer FT (2010) Deciphering subsampled data: adaptive compressive sampling as a principle of brain communication. arXiv q-bio.NC.